ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ & ΠΛΗΡΟΦΟΡΙΚΗΣ

Διδακτορική Διατριβή

Τεχνικές και Συστήματα Διαχείρισης Γνώσης στο Διαδίκτυο

Πηνελόπη Θ. Μαρκέλλου

Επιβλέπων

Αθανάσιος Τσακαλίδης, Καθηγητής

Νοέμβριος 2005

*στους γονείς μου, Φάνη και Γιώτα που πάντα με στηρίζουν
στον αδερφό μου Κώστα που είναι πάντα αυτός που είναι
στο Βλάχη που είναι πάντα εκεί
και στη Θεοφανία που είναι τα πάντα*

Διδακτορική Διατριβή

Τεχνικές και Συστήματα Διαχείρισης Γνώσης στο Διαδίκτυο

Πηνελόπη Θ. Μαρκέλλου

Επταμελής Εξεταστική Επιτροπή

Αθανάσιος Τσακαλίδης, Καθηγητής Τμήματος Μηχανικών Ηλεκτρονικών
Υπολογιστών & Πληροφορικής (Επιβλέπων)

Σπυρίδων Λυκοθανάσης, Καθηγητής Τμήματος Μηχανικών Ηλεκτρονικών
Υπολογιστών & Πληροφορικής (Μέλος Τριμελούς Επιτροπής Παρακολούθησης)

Ιωάννης Γαροφαλάκης, Αναπληρωτής Καθηγητής Τμήματος Μηχανικών
Ηλεκτρονικών Υπολογιστών & Πληροφορικής (Μέλος Τριμελούς Επιτροπής
Παρακολούθησης)

Γεώργιος Παυλίδης, Καθηγητής Τμήματος Διοίκησης Επιχειρήσεων

Χρήστος Μακρής, Επίκουρος Καθηγητής Τμήματος Μηχανικών Ηλεκτρονικών
Υπολογιστών & Πληροφορικής

Ιωάννης Χατζηλυγερούδης, Επίκουρος Καθηγητής Τμήματος Μηχανικών
Ηλεκτρονικών Υπολογιστών & Πληροφορικής

Σωτήριος Νικολετσέας, Λέκτορας Τμήματος Μηχανικών Ηλεκτρονικών
Υπολογιστών & Πληροφορικής

Νοέμβριος 2005

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους ανθρώπους που με βοήθησαν και με στήριξαν σε όλη τη διάρκεια εκπόνησης αυτής της διδακτορικής διατριβής και όχι μόνο.

Πρώτα από όλους ευχαριστώ τον **καθηγητή κ. Αθανάσιο Τσακαλίδη**, επιβλέποντα της διατριβής μου, που μου έδωσε την ευκαιρία το 1995 να γίνω μέλος του Εργαστηρίου του. Η εμπιστοσύνη που μου έδειξε χωρίς να με γνωρίζει με συγκίνησε και θέλω να πιστεύω ότι όλα αυτά τα χρόνια στάθηκαν αντάξια της.

Τον **αναπληρωτή καθηγητή κ. Γιάννη Γαροφαλάκη** για τη βοήθεια του και την πολύτιμη υποστήριξή του.

Τον **καθηγητή κ. Σπύρο Λυκοθανάση**, μέλος της τριμελούς επιτροπής παρακολούθησης της διατριβής μου.

Τα μέλη της επταμελούς εξεταστικής επιτροπής μου, **καθηγητή κ. Γιώργο Παυλίδη**, **επίκουρο καθηγητή κ. Χρήστο Μακρή**, **επίκουρο καθηγητή κ. Γιάννη Χατζηλυγερούδη** και **λέκτορα κ. Σωτήρη Νικολετσέα**.

Το **Σπύρο Συρμακέση**, τον άνθρωπο που μέσα από τις περιεκτικές συζητήσεις μας και τις διαφωνίες μας μου έδειξε το δρόμο. Κυρίως όμως για την υποστήριξή του τις δύσκολες ώρες που έχανα την πίστη μου. Πέρα από την πολύτιμη βοήθεια και καθοδήγησή του, η φιλία του είναι πιο σημαντική για μένα.

Τη **Μαρία Ρήγκου**, συμπαρασάτη και συνοδοιπόρο όλα αυτά τα χρόνια. Πέρασαμε άπειρες ώρες δουλεύοντας μαζί, άγχη, προβληματισμούς, αλλά και χαρές που δύσκολα θα ξαναπεράσω με κάποιον άλλο άνθρωπο. Ακόμα και στις στιγμές αμφισβήτησης, θέλω να ξέρει ότι υπολογίζω στη φιλία της.

Τη **Γιάννα Μουσουρούλη**, που η βοήθεια της τα τελευταία δύο χρόνια ήταν ουσιαστική και ελπίζω σύντομα να την ανταποδώσω.

Στους **δικούς μου ανθρώπους** και σε όλους τους **φίλους** που μου στάθηκαν, πίστεψαν σε εμένα, με ενθάρρυναν και με έκαναν να νιώσω ότι οι προσπάθειες μου θα πιάσουν τελικά τόπο. Ένα μεγάλο ευχαριστώ γιατί είσαστε εσείς.

Περίληψη διδακτορικής διατριβής

Ο Παγκόσμιος Ιστός Πληροφοριών (Web) χαρακτηρίζεται σαν ένα περιβάλλον αχανές, ετερογενές, κατανεμημένο και πολύπλοκο με αποτέλεσμα να είναι δύσκολος ο αποδοτικός χειρισμός των δεδομένων των e-εφαρμογών με βάση παραδοσιακές μεθόδους και τεχνικές. Αυτό με τη σειρά του οδηγεί στην απαίτηση για σχεδιασμό, ανάπτυξη και υιοθέτηση «ευφυών» εργαλείων που θα επιλέξουν και θα εμφανίσουν στο χρήστη την κατάλληλη πληροφορία, στον κατάλληλο χρόνο και με την κατάλληλη μορφή. Η παρούσα διδακτορική διατριβή ασχολείται με το πρόβλημα της εξόρυξης «κρυμμένης» γνώσης από συστήματα και εφαρμογές ηλεκτρονικής μάθησης (e-learning), ηλεκτρονικού εμπορίου (e-commerce) και επιχειρηματικής ευφυΐας (business intelligence) με κύριο στόχο τη βελτίωση της ποιότητας και της απόδοσης των παρεχόμενων υπηρεσιών προς τους τελικούς χρήστες. Συγκεκριμένα, τα ερευνητικά αποτελέσματα επικεντρώνονται στα ακόλουθα:

- Μεθοδολογίες, τεχνικές και προτεινόμενοι αλγόριθμοι εξόρυξης «κρυμμένης» γνώσης από e-εφαρμογές λαμβάνοντας υπόψη τη σημασιολογία των δεδομένων.
- Παραγωγή εξατομικευμένων εκπαιδευτικών εμπειριών.
- Παραγωγή αποδοτικών συστάσεων για την αγορά online προϊόντων.
- Παραγωγή επιστημονικών και τεχνολογικών δεικτών από διπλώματα ευρεσιτεχνίας για την ανάδειξη του επιπέδου καινοτόμου δραστηριότητας μιας αγοράς.
- Προτάσεις για μελλοντικές ερευνητικές κατευθύνσεις που επεκτείνουν τις τεχνικές εξόρυξης γνώσης σε πιο σύνθετους τύπους εφαρμογών και αναδεικνύουν νέες ερευνητικές ευκαιρίες.

Στο πρώτο κεφάλαιο παρουσιάζεται μια προσέγγιση για την υποστήριξη εξατομικευμένου e-learning όπου η δομή και η σχέση των δεδομένων και των πληροφοριών παίζουν ουσιαστικό ρόλο. Ο προτεινόμενος αλγόριθμος βασίζεται σε μια οντολογία (ontology) η οποία βοηθά στη δόμηση και στη διαχείριση του περιεχομένου που σχετίζεται με μια δεδομένη σειρά μαθημάτων, ένα μάθημα ή ένα θεματικό. Η διαδικασία χωρίζεται σε δύο στάδια: στις offline ενέργειες προετοιμασίας των δεδομένων, δημιουργίας της οντολογίας και εξόρυξης από δεδομένα χρήσης (usage mining) και στην online παροχή της εξατομίκευσης. Το σύστημα βρίσκει σε πρώτη φάση ένα αρχικό σύνολο συστάσεων βασισμένο στην οντολογία του πεδίου και στη συνέχεια χρησιμοποιεί τα frequent itemsets (συχνά εμφανιζόμενα σύνολα στοιχείων) για να το εμπλουτίσει, λαμβάνοντας υπόψη την πλοήγηση άλλων παρόμοιων χρηστών (similar users). Με τον τρόπο αυτό, μειώνουμε το χρόνο που απαιτείται για την ανάλυση όλων των frequent itemsets και των κανόνων συσχέτισης. Εστιάζουμε μόνο σε εκείνα τα σύνολα που προέρχονται από το συνδυασμό της ενεργούς συνόδου (current session) του χρήστη και των συστάσεων της οντολογίας. Επιπλέον, αυτή η προσέγγιση ανακουφίζει και το πρόβλημα των μεγάλων χρόνων απόκρισης, το οποίο μπορεί στη συνέχεια να οδηγήσει στην εγκατάλειψη του e-learning συστήματος. Αν και η εξατομίκευση απαιτεί αρκετά βήματα επεξεργασίας και ανάλυσης, το εμπόδιο αυτό αποφεύγεται με την εκτέλεση σημαντικού μέρους της διαδικασίας offline.

Στο δεύτερο κεφάλαιο μελετάται το πρόβλημα της παραγωγής προτάσεων σε μια εφαρμογή e-commerce. Τα συστήματα συστάσεων (recommendations systems ή RSs) αποτελούν ίσως την πιο δημοφιλή μορφή εξατομίκευσης και τείνουν να μετατραπούν στις μέρες μας σε σημαντικά επιχειρησιακά εργαλεία. Η προτεινόμενη υβριδική προσέγγιση στοχεύει στην παραγωγή αποτελεσματικών συστάσεων για τους πελάτες ενός online καταστήματος που νοικιάζει κινηματογραφικές ταινίες. Η γνώση για τους πελάτες και τα προϊόντα προκύπτει από δεδομένα χρήσης και τη δομή της οντολογίας σε συνδυασμό με τις εκτιμήσεις-βαθμολογίες των πελατών για τις ταινίες καθώς και την εφαρμογή τεχνικών ταιριάσματος «όμοιων» πελατών. Όταν ένα ή περισσότερα κριτήρια ταιριάσματος ικανοποιούνται, τότε άλλες ταινίες μπορούν να προσδιοριστούν σύμφωνα με το οντολογικό σχήμα που έχουν παρόμοια χαρακτηριστικά με αυτές που ο πελάτης έχει ήδη νοικιάσει. Στην περίπτωση ενός νέου πελάτη όπου το ιστορικό του είναι κενό, πληροφορίες από την αίτηση εγγραφής του αναλύονται ώστε να ταξινομηθεί σε μια συγκεκριμένη κλάση πελατών και να παραχθούν προτάσεις με βάση το οντολογικό σχήμα. Αυτή η ολοκλήρωση παρέχει πρόσθετη γνώση για τις προτιμήσεις των πελατών και επιτρέπει την παραγωγή επιτυχημένων συστάσεων. Ακόμη και στην περίπτωση του «cold-start problem» όπου δεν είναι διαθέσιμη αρχική πληροφορία για τη συμπεριφορά του πελάτη, η προσέγγιση μπορεί να προβεί σε σχετικές συστάσεις.

Τέλος, στο τρίτο κεφάλαιο μελετάται το πρόβλημα της εξόρυξης γνώσης από καταχωρήσεις διπλωμάτων ευρεσιτεχνίας που καταδεικνύουν το επίπεδο της καινοτόμου δραστηριότητας μιας αγοράς. Η προτεινόμενη προσέγγιση αφορά στην εφαρμογή τεχνικών Text Mining σε διπλώματα ευρεσιτεχνίας που βρίσκονται καταχωρημένα σε βάσεις δεδομένων διαφόρων διεθνών οργανισμών διαχείρισής τους, με στόχο την παραγωγή επιστημονικών και τεχνολογικών δεικτών για την ανάδειξη του επιπέδου καινοτομίας μιας αγοράς και συνεπώς την επιχειρηματική ευφυΐα. Αρχικά τα δεδομένα καθαρίζονται προκειμένου να βελτιωθεί η ποιότητά τους πριν την επεξεργασία. Στη συνέχεια εφαρμόζονται δύο τύποι επεξεργασίας η απλή ανάλυση (simple analysis) και η στατιστική ανάλυση (statistical analysis). Στην πρώτη περίπτωση παράγονται γραφήματα που συσχετίζουν τις πληροφορίες π.χ. κύριοι τομείς ανάπτυξης σε μια χώρα. Στη δεύτερη περίπτωση αναλύονται γλωσσολογικά τα πεδία title και abstract των διπλωμάτων ευρεσιτεχνίας και ομαδοποιούνται τα λήμματα των λέξεων. Στη συνέχεια πάνω στα δεδομένα εφαρμόζονται τεχνικές correspondence και clustering analysis έτσι ώστε αυτά να ομαδοποιηθούν σύμφωνα με τις τεχνολογίες στις οποίες αναφέρονται. Τα clusters πλέον αυτά προβάλλονται όπως και στην απλή ανάλυση παρέχοντας στο χρήστη μια πιο λεπτομερή απεικόνιση της πληροφορίας των διπλωμάτων ευρεσιτεχνίας. Ο συνδυασμός των αναλύσεων που εφαρμόζονται με βάση την προτεινόμενη μεθοδολογία επιτρέπει την αποτύπωση των τεχνολογικών εξελίξεων και καινοτομιών. Οι δείκτες που παράγονται είναι πολύ σημαντικοί αφού μπορούν να ποσοτικοποιήσουν τις πληροφορίες που αφορούν σε συγκεκριμένες τεχνολογίες. Με αυτό τον τρόπο μπορούμε να παράγουμε δείκτες για τη δραστηριότητα συγκεκριμένων φορέων, εφευρετών, χωρών, κλπ. Τέλος, τεχνολογικοί δείκτες που υποδεικνύουν μελλοντικές ελπιδοφόρες τεχνολογίες καθώς και ποιοι φορείς θα είναι πρωτοπόροι σε αυτές μπορούν να εξαχθούν.

Dissertation summary

The World Wide Web (Web) has been characterized as a vast, heterogeneous, distributed and complicated environment resulting in difficulties for the efficient handling of e-applications' data with traditional methods and techniques. This leads to the requirement for designing, implementing and adopting "intelligent" tools, able to select and present to the user the suitable information, at the suitable time and in the suitable form. The present dissertation deals with the problem of mining "hidden" knowledge from systems and applications of electronic learning (e-learning), electronic commerce (e-commerce) and business intelligence (BI), aiming mainly at the improvement of quality and performance of the services provided to the end-users. Specifically, the results are focused on the following:

- Methodologies, techniques and proposed algorithms of mining "hidden" knowledge from e-applications taking into consideration the semantics of data.
- Production of personalized educational experiences.
- Generation of efficient recommendations for the online purchase of products.
- Discovery of scientific and technological indicators in patents that indicate the level of innovation activity of a market.
- Proposals for future research directions that extend the techniques of knowledge mining to more complex types of applications and indicate new research opportunities.

The first chapter presents an approach for the support of personalized e-learning in the cases where the structure and the relation of data and information play essential role. The proposed algorithm is based on an ontology which helps in structuring and managing the content that is related with a given course, a lesson or a topic. The process is decomposed into two stages: the offline phase of data preparation, ontology creation and data usage mining and the online phase of producing personalization. The system finds a initial set of recommendations based on the ontology of the domain and then identifies a set of frequent itemsets (sets of items observed often) in order to enrich the initial recommendations, taking into consideration the navigation of other similar users. In this way, we decrease the time required for the analysis of all the frequent itemsets and association rules, by focusing only on those sets that derive from the combination of the current active session of the user and the ontology recommendations. Moreover, this approach also alleviates the problem of long response times that can lead to the abandonment of the system. Even if the personalization requires considerable steps of preparation and analysis, this obstacle is avoided with the implementation of important part of the process offline.

The second chapter studies the problem of recommendations' production in an e-commerce application. Recommendation systems or RSs constitute perhaps the most popular form of personalization and they tend to become in our days an important business tool. The proposed hybrid approach aims in the production of effective recommendations for the

customers of an online shop that rents movies. The knowledge for the customers and the movies results from usage data and the structure of an ontology in combination with customer rates about the movies, as well as with the application of matching techniques for discovering "similar" customers. When one or more matching criteria are satisfied, then other movies can be specified according to the ontological schema that has similar characteristics with those that the customer already has rented. In the case of a new customer with no history information, data from his registration form are analyzed so that he is categorized in a specific group of customers and the recommendations are based on the ontology. This integration provides additional knowledge for the preferences of customers and allows the production of successful recommendations. Even in the case of "cold-start problem" where initial information on the customer's behavior is not available, the approach can produce qualitative and relatively precise recommendations.

Finally, the third chapter describes the problem of mining knowledge from patent registrations which indicate the level of innovation activity of a market. The proposed approach concerns the application of Text Mining techniques in patents retrieved from the databases of various national and international Patent Offices, aiming at the production of scientific and technological indicators of the innovation level of a market activity and consequently business intelligence. Initially, the data are cleaned in order to improve their quality before the analysis steps. Then two types of analysis are applied on the data: simple analysis and statistical analysis. In the first case, several charts are produced that connect the information e.g. main sectors of development in a country. In the second case, the title and abstract fields of the patents are linguistically analyzed and the lemmas of words are grouped. Then correspondence and clustering analysis are applied. The produced clusters are depicted as in the simple analysis providing the user with a detailed representation of patent information. The combination of analyses that are applied based on the proposed methodology allows the identification of technological evolutions and innovations. The produced indicators are very important since they can quantify the information that concerns specific technologies. In this way, we can produce indicators for the activity of specific institutions, inventors, countries, etc. Finally, technological indicators about the potential emerging technologies as well as the institutions that will be pioneers can be exported.

Περιεχόμενα

ΕΥΧΑΡΙΣΤΙΕΣ.....	7
ΠΕΡΙΛΗΨΗ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ.....	9
DISSERTATION SUMMARY	11
ΠΕΡΙΕΧΟΜΕΝΑ	13
ΛΙΣΤΑ ΣΧΗΜΑΤΩΝ	15
ΛΙΣΤΑ ΠΙΝΑΚΩΝ	17
ΠΡΟΛΟΓΟΣ	19
ΒΑΣΙΚΕΣ ΔΗΜΟΣΙΕΥΣΕΙΣ ΔΙΑΤΡΙΒΗΣ	29
ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ E-LEARNING.....	37
1.1 ΕΙΣΑΓΩΓΗ.....	37
1.1.1 Το πρόβλημα και η προτεινόμενη προσέγγιση.....	40
1.1.2 Η δομή του κεφαλαίου.....	41
1.2 Το SEMANTIC WEB.....	41
1.2.1 Η αρχιτεκτονική του Semantic Web.....	42
1.2.2 Ο ορισμός της οντολογίας.....	44
1.2.3 Ταξινόμηση των οντολογιών.....	45
1.3 ΑΠΟ ΤΟ WEB MINING ΣΤΟ SEMANTIC WEB MINING	47
1.3.1 Web Content Mining.....	47
1.3.2 Web Structure Mining	48
1.3.3 Web Usage Mining.....	48
1.4 SEMANTIC WEB MINING ΚΑΙ E-LEARNING.....	50
1.5 Η ΔΙΑΔΙΚΑΣΙΑ ΕΞΑΤΟΜΙΚΕΥΣΗΣ ΣΕ ΕΝΑ ΤΥΠΙΚΟ E-LEARNING ΣΥΣΤΗΜΑ	52
1.5.1 Συλλογή των δεδομένων.....	53
1.5.2 Καθαρισμός των δεδομένων και μοντελοποίηση του χρήστη.....	55
1.5.3 Ανάλυση των δεδομένων.....	56
1.5.4 Ολοκλήρωση των δεδομένων.....	57
1.5.5 Παραγωγή εξατομίκευσης.....	58
1.6 ΈΝΑ ΣΕΝΑΡΙΟ ΓΙΑ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΕΞΑΤΟΜΙΚΕΥΜΕΝΟ E-LEARNING.....	58
1.6.1 Ορισμός του προβλήματος παραγωγής εξατομίκευσης σε ένα e-learning σύστημα στο θεματικό Web Mining.....	58
1.6.2 Περιορισμοί και προβλήματα των Web Mining προσεγγίσεων.....	60
1.6.3 Σχετικές εργασίες.....	61
1.6.4 Η e-learning οντολογία.....	62
1.6.5 Προ-επεξεργασία των δεδομένων.....	66
1.6.6 Εξαγωγή των κανόνων συσχέτισης.....	68
1.7 Ο ΑΛΓΟΡΙΘΜΟΣ ΠΑΡΑΓΩΓΗΣ ΕΞΑΤΟΜΙΚΕΥΣΗΣ.....	73
1.8 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΑΝΟΙΚΤΑ ΘΕΜΑΤΑ	80
1.9 ΒΙΒΛΙΟΓΡΑΦΙΑ	82
ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ E-COMMERCE	93
2.1 ΕΙΣΑΓΩΓΗ.....	93

2.1.1	Το πρόβλημα και η προτεινόμενη προσέγγιση.....	95
2.1.2	Η δομή του κεφαλαίου.....	96
2.2	ΤΑ ΔΟΜΙΚΑ ΣΤΟΙΧΕΙΑ ΕΝΟΣ RS.....	97
2.2.1	Η είσοδος.....	97
2.2.2	Η έξοδος.....	98
2.2.3	Η μέθοδος.....	100
2.2.4	Ο τρόπος αποστολής.....	102
2.2.5	Ο βαθμός εξατομίκευσης.....	103
2.3	ΤΑΞΙΝΟΜΗΣΗ ΤΩΝ RSS.....	103
2.3.1	Content-based approaches.....	104
2.3.2	Collaborative filtering approaches.....	107
2.3.3	Hybrid approaches.....	110
2.4	ΤΟ ΜΟΝΤΕΛΟ ΤΟΥ E-CUSTOMER.....	112
2.5	ΈΝΑ RS ΜΟΝΤΕΛΟ ΠΟΥ ΒΑΣΙΖΕΤΑΙ ΣΕ ΔΕΔΟΜΕΝΑ ΧΡΗΣΗΣ ΙΣΤΟΥ ΚΑΙ ΣΤΗΝ ΟΝΤΟΛΟΓΙΑ ΤΟΥ E-SHOP.....	121
2.5.1	Ορισμός του προβλήματος παραγωγής προτάσεων σε ένα e-shop που νοικιάζει κινηματογραφικές ταινίες.....	121
2.5.2	Περιορισμοί και προβλήματα των RSs τεχνικών.....	123
2.5.3	Σχετικές εργασίες.....	125
2.5.4	Η οντολογία του e-shop.....	126
2.5.5	Συλλογή δεδομένων.....	130
2.5.6	Ανάλυση δεδομένων.....	131
2.5.7	Ταξινόμηση χρηστών με βάση τα δεδομένα εγγραφής.....	132
2.5.8	Ανάλυση των click-streams των χρηστών.....	133
2.6	ΑΛΓΟΡΙΘΜΟΣ ΠΑΡΑΓΩΓΗΣ RECOMMENDATIONS.....	134
2.7	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΑΝΟΙΚΤΑ ΘΕΜΑΤΑ.....	141
2.8	ΒΙΒΛΙΟΓΡΑΦΙΑ.....	142
	ΕΞΟΥΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ.....	153
3.1	ΕΙΣΑΓΩΓΗ.....	153
3.1.1	Το πρόβλημα.....	155
3.1.2	Η δομή του κεφαλαίου.....	156
3.2	Η ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ.....	157
3.3	ΕΞΟΥΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΒΑΣΕΙΣ ΔΙΠΛΩΜΑΤΩΝ ΕΥΡΕΣΤΕΧΝΙΑΣ ΓΙΑ ΤΗΝ ΠΑΡΑΓΩΓΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΚΑΙ ΕΠΙΣΤΗΜΟΝΙΚΩΝ ΔΕΙΚΤΩΝ.....	159
3.3.1	Το κίνητρο για την παραγωγή τεχνολογικών και επιστημονικών δεικτών.....	159
3.3.2	Περιγραφή μιας πατέντας.....	160
3.3.3	Η μεθοδολογική προσέγγιση.....	163
3.4	ΕΠΙΛΟΓΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ.....	164
3.5	ΓΛΩΣΣΟΛΟΓΙΚΗ ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ.....	166
3.5.1	Τα βήματα.....	167
3.5.2	Ο αλγόριθμος.....	170
3.6	CORRESPONDENCE ΑΝΑΛΥΣΗ.....	174
3.6.1	Ιστορική ανασκόπηση.....	174
3.6.2	Ο αλγόριθμος.....	175
3.7	CLUSTER ΑΝΑΛΥΣΗ.....	180
3.7.1	Ο αλγόριθμος.....	180
3.8	SIMPLE ΑΝΑΛΥΣΗ.....	190
3.8.1	Ιστογράμματα.....	190
3.8.2	Ραβδογράμματα.....	191
3.8.3	Διαγράμματα πιτών.....	191
3.9	Η ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ.....	193
3.10	CASE STUDIES.....	194
3.11	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΑΝΟΙΚΤΑ ΘΕΜΑΤΑ.....	200
3.12	ΒΙΒΛΙΟΓΡΑΦΙΑ.....	202
	ΓΛΩΣΣΑΡΙ ΟΡΩΝ.....	205
	ΕΥΡΕΤΗΡΙΟ ΟΡΩΝ.....	213

Λίστα σχημάτων

ΣΧΗΜΑ 1. ΤΑ ΕΠΙΠΕΔΑ ΤΗΣ ΑΡΧΙΤΕΚΤΟΝΙΚΗΣ ΤΟΥ SEMANTIC WEB.....	42
ΣΧΗΜΑ 2. ΤΑΞΙΝΟΜΗΣΗ ΟΝΤΟΛΟΓΙΩΝ ΣΥΜΦΩΝΑ ΜΕ ΤΟΝ GUARINO (1998).....	46
ΣΧΗΜΑ 3. ΟΙ ΚΑΤΗΓΟΡΙΕΣ ΤΟΥ WEB MINING.....	47
ΣΧΗΜΑ 4. ΣΥΣΧΕΤΙΣΗ ΠΡΟΤΥΠΩΝ E-LEARNING ΜΕΤΑΔΕΔΟΜΕΝΩΝ.....	52
ΣΧΗΜΑ 5. Η ΔΙΑΔΙΚΑΣΙΑ ΕΞΑΤΟΜΙΚΕΥΣΗΣ ΣΕ ΕΝΑ ΤΥΠΙΚΟ E-LEARNING ΣΥΣΤΗΜΑ.....	53
ΣΧΗΜΑ 6. ΠΑΡΑΔΕΙΓΜΑ LOG FILE.....	55
ΣΧΗΜΑ 7. ΤΜΗΜΑ ΤΗΣ ΟΝΤΟΛΟΓΙΑΣ ΓΙΑ ΤΗ ΣΕΙΡΑ ΜΑΘΗΜΑΤΩΝ «WEB MINING».....	64
ΣΧΗΜΑ 8. ΤΜΗΜΑ ΤΗΣ ΟΝΤΟΛΟΓΙΑΣ ΓΙΑ ΤΟ LEARNINGRESOURCE.....	65
ΣΧΗΜΑ 9. ΤΟ ΕΝΝΟΙΟΛΟΓΙΚΟ ΜΟΝΤΕΛΟ ΤΟΥ LEARNER.....	65
ΣΧΗΜΑ 10. ΔΙΚΤΥΩΤΟ ΠΛΕΓΜΑ ITEMSETS ΓΙΑ ΤΟ {A,B,C,D}.....	69
ΣΧΗΜΑ 11. ΥΠΟΣΥΝΟΛΑ ΤΟΥ {ACD}.....	70
ΣΧΗΜΑ 12. ΠΡΟΤΕΙΝΟΜΕΝΟ ΣΕΝΑΡΙΟ ΓΙΑ ΤΗΝ ΠΑΡΑΓΩΓΗ ΤΩΝ ΣΥΣΤΑΣΕΩΝ ΣΕ ΕΝΑ ΣΥΣΤΗΜΑ E-LEARNING.....	73
ΣΧΗΜΑ 13. Η ΑΡΧΙΚΗ ΣΕΛΙΔΑ ΤΟΥ ΠΙΛΟΤΙΚΟΥ E-LEARNING ΣΥΣΤΗΜΑΤΟΣ.....	78
ΣΧΗΜΑ 14. ΠΑΡΑΓΩΓΗ ΠΡΟΤΑΣΕΩΝ ΓΙΑ ΤΟΝ ΕΚΠΑΙΔΕΥΟΜΕΝΟ.....	79
ΣΧΗΜΑ 15. Η ΥΠΗΡΕΣΙΑ MYPAGE.....	80
ΣΧΗΜΑ 16. ΔΟΜΙΚΑ ΣΤΟΙΧΕΙΑ ΕΝΟΣ RS.....	97
ΣΧΗΜΑ 17. ΠΡΟΣΕΓΓΙΖΟΝΤΑΣ ΤΟ E-SHOP.....	113
ΣΧΗΜΑ 18: ΜΕΣΑ ΣΤΟ E-SHOP.....	116
ΣΧΗΜΑ 19: ΑΦΗΝΟΝΤΑΣ ΤΟ E-SHOP.....	120
ΣΧΗΜΑ 20. ΜΕΡΟΣ ΤΗΣ ΟΝΤΟΛΟΓΙΑΣ ΤΟΥ E-SHOP.....	127
ΣΧΗΜΑ 21. ΤΟ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ CATEGORY.....	128
ΣΧΗΜΑ 22. ΠΑΡΑΔΕΙΓΜΑ ΤΗΣ ΤΑΙΝΙΑΣ «LORD OF THE RINGS – THE RETURN OF THE KING».....	128
ΣΧΗΜΑ 23. ΣΤΙΓΜΙΟΤΥΠΟ ΤΗΣ ΟΝΤΟΛΟΓΙΑΣ.....	129
ΣΧΗΜΑ 24. Η ΠΡΟΤΕΙΝΟΜΕΝΗ ΠΡΟΣΕΓΓΙΣΗ ΓΙΑ ΤΟ RS.....	131
ΣΧΗΜΑ 25. ΕΙΣΑΓΩΓΗ ΕΝΟΣ ΧΡΗΣΤΗ ΣΤΟ ONLINE DVD CLUB.....	139
ΣΧΗΜΑ 26. ΠΑΡΑΓΩΓΗ ΠΡΟΤΑΣΕΩΝ ΓΙΑ ΚΙΝΗΜΑΤΟΓΡΑΦΙΚΕΣ ΤΑΙΝΙΕΣ.....	139
ΣΧΗΜΑ 27. ΤΟ ΙΣΤΟΡΙΚΟ ΠΑΡΑΓΕΛΙΩΝ ΕΝΟΣ ΠΕΛΑΤΗ.....	140
ΣΧΗΜΑ 28. ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ ΤΟΥ DVD CLUB.....	140
ΣΧΗΜΑ 29. Η ΔΙΑΔΙΚΑΣΙΑ KDD.....	157
ΣΧΗΜΑ 30. Η ΑΜΦΙΔΡΟΜΗ ΣΧΕΣΗ ΤΗΣ ΕΡΕΥΝΑΣ ΚΑΙ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΣΥΝΔΥΑΣΜΕΝΗ ΜΕΣΩ ΠΑΤΕΝΤΩΝ.....	160
ΣΧΗΜΑ 31. ΠΛΗΡΟΦΟΡΙΕΣ ΠΟΥ ΠΕΡΙΕΧΟΝΤΑΙ ΣΕ ΜΙΑ ΠΑΤΕΝΤΑ.....	161
ΣΧΗΜΑ 32. ΠΑΡΑΔΕΙΓΜΑ ΠΑΤΕΝΤΑΣ.....	162
ΣΧΗΜΑ 33. ΤΟ ΕΡΓΑΛΕΙΟ ΑΝΑΚΤΗΣΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΒΑΣΕΙΣ ΠΑΤΕΝΤΩΝ MIMOSA.....	165
ΣΧΗΜΑ 34. Η ΔΙΑΔΙΚΑΣΙΑ ΑΠΟΘΗΚΕΥΣΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΟ ΣΥΣΤΗΜΑ.....	165
ΣΧΗΜΑ 35. ΔΕΔΟΜΕΝΑ ΕΙΣΟΔΟΥ ΣΤΟ ΣΥΣΤΗΜΑ.....	166
ΣΧΗΜΑ 36. ΔΕΞΙΚΟΓΡΑΦΙΚΗ ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ.....	169
ΣΧΗΜΑ 37. ΓΛΩΣΣΟΛΟΓΙΚΗ ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ.....	172
ΣΧΗΜΑ 38. ΕΠΙΛΟΓΗ ΤΩΝ ΔΗΜΜΑΤΩΝ ΚΑΙ ΓΡΑΦΙΚΗ ΤΟΥΣ ΑΝΑΠΑΡΑΣΤΑΣΗ.....	172
ΣΧΗΜΑ 39. ΠΙΝΑΚΑΣ ΓΕΓΝΙΑΣΗΣ.....	173
ΣΧΗΜΑ 40. ΙΣΤΟΓΡΑΜΜΑ ΤΩΝ ΔΙΑΣΤΑΣΕΩΝ.....	173
ΣΧΗΜΑ 41. ΔΙΑΔΙΚΑΣΙΑ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΣΤΑΔΩΝ.....	181
ΣΧΗΜΑ 42. ΔΕΝΔΡΟΓΡΑΜΜΑ.....	188
ΣΧΗΜΑ 43. ΑΠΕΙΚΟΝΙΣΗ CLUSTERS.....	189
ΣΧΗΜΑ 44. ΑΠΕΙΚΟΝΙΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΑΝΑ CLUSTER.....	189
ΣΧΗΜΑ 45. ΙΣΤΟΓΡΑΜΜΑ ΓΙΑ ΕΝΑ ΣΥΝΟΛΟ ΑΠΟ 3.000 ΔΕΔΟΜΕΝΑ ΜΕ ΜΕΓΕΘΟΣ BINS=10.....	191
ΣΧΗΜΑ 46. ΙΣΤΟΓΡΑΜΜΑ ΓΙΑ ΤΟ ΙΔΙΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ ΑΛΛΑ ΜΕ ΜΕΓΕΘΟΣ BINS=20.....	191
ΣΧΗΜΑ 47. SIMPLE ANALYSIS ΚΑΙ ΓΡΑΦΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ.....	192
ΣΧΗΜΑ 48. ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΟΣ.....	194
ΣΧΗΜΑ 49. ΚΑΤΗΓΟΡΙΕΣ ΔΕΙΚΤΩΝ.....	195
ΣΧΗΜΑ 50. Η ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΤΩΝ ΧΩΡΩΝ (CASE STUDY G10).....	196
ΣΧΗΜΑ 51. Η ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΤΩΝ ΕΤΑΙΡΙΩΝ (CASE STUDY G10).....	197

ΣΧΗΜΑ 52. ΟΙΚΩΔΙΚΟΙ ΜC ΑΝΑ CLUSTER (CASE STUDY A23L).....	198
ΣΧΗΜΑ 53. ΟΙ ΠΙΟ ΕΝΕΡΓΟΙ ΕΦΕΥΡΕΤΕΣ (CASE STUDY F16F).....	199

Λίστα πινάκων

ΠΙΝΑΚΑΣ 1. E-LEARNING ΚΑΙ SEMANTIC WEB MINING.	51
ΠΙΝΑΚΑΣ 2. Ο ΑΛΓΟΡΙΘΜΟΣ ΠΟΥ ΔΗΜΙΟΥΡΓΕΙ ΕΝΑ ΥΠΕΡΣΥΝΟΛΟ FREQUENT ITEMSETS.	70
ΠΙΝΑΚΑΣ 3. Ο ΑΛΓΟΡΙΘΜΟΣ APRIORI.	71
ΠΙΝΑΚΑΣ 4. Ο ΑΛΓΟΡΙΘΜΟΣ ΠΟΥ ΕΞΑΓΕΙ ΤΟΥΣ ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ.	72
ΠΙΝΑΚΑΣ 5. ΠΡΟΤΕΙΝΟΜΕΝΟΣ ΑΛΓΟΡΙΘΜΟΣ ΓΙΑ ΠΑΡΑΓΩΓΗ ΣΥΣΤΑΣΕΩΝ ΣΤΟΥΣ ΧΡΗΣΤΕΣ.	75
ΠΙΝΑΚΑΣ 6. ΤΑΞΙΝΟΜΗΣΗ ΤΕΧΝΙΚΩΝ ΓΙΑ RSS ΚΑΙ ΑΝΤΙΠΡΟΣΩΠΕΥΤΙΚΑ ΕΡΕΥΝΗΤΙΚΑ ΠΑΡΑΔΕΙΓΜΑΤΑ.	104
ΠΙΝΑΚΑΣ 7. ΒΑΘΜΟΛΟΓΗΣΕΙΣ ΤΑΙΝΙΩΝ ΑΠΟ ΧΡΗΣΤΕΣ.	122
ΠΙΝΑΚΑΣ 8. ΕΞΑΓΩΓΗ ΣΥΜΠΕΡΑΣΜΑΤΩΝ ΑΠΟ ΤΙΣ ΒΑΘΜΟΛΟΓΙΕΣ ΤΩΝ ΧΡΗΣΤΕΣ ΓΙΑ ΤΙΣ ΤΑΙΝΙΕΣ.	130
ΠΙΝΑΚΑΣ 9. ΜΕΡΟΣ ΕΝΟΣ ΠΡΟΦΙΛ ΧΡΗΣΤΗ.	132
ΠΙΝΑΚΑΣ 10. ΣΤΙΓΜΙΟΤΥΠΟ ΤΗΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ.	132
ΠΙΝΑΚΑΣ 11. Ο ΑΛΓΟΡΙΘΜΟΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΕΙΤΑΙ ΓΙΑ ΤΗΝ ΠΑΡΑΓΩΓΗ ΤΩΝ RECOMMENDATIONS.	137
ΠΙΝΑΚΑΣ 12. ΈΝΑ ΠΑΡΑΔΕΙΓΜΑ ΛΕΞΙΚΟΥ ΚΑΙ ΛΕΞΙΚΟΓΡΑΦΙΚΟΥ ΠΙΝΑΚΑ.	168
ΠΙΝΑΚΑΣ 13. ΑΝΑΓΝΩΡΙΣΗ ΤΩΝ ΛΕΞΕΩΝ.	168
ΠΙΝΑΚΑΣ 14. ΑΦΑΙΡΕΣΗ ΤΩΝ ΛΕΞΕΩΝ ΧΩΡΙΣ ΣΗΜΑΣΙΑ.	170
ΠΙΝΑΚΑΣ 15. Η ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΤΩΝ ΧΩΡΩΝ (CASE STUDY G10).	196
ΠΙΝΑΚΑΣ 16. Η ΔΡΑΣΤΗΡΙΟΤΗΤΑ ΤΩΝ ΕΤΑΙΡΙΩΝ (CASE STUDY G10).	197
ΠΙΝΑΚΑΣ 17. ΟΙ ΚΩΔΙΚΟΙ MC ΑΝΑ CLUSTER (CASE STUDY A23L).	198
ΠΙΝΑΚΑΣ 18. ΟΙ ΠΙΟ ΕΝΕΡΓΟΙ ΕΦΕΥΡΕΤΕΣ (CASE STUDY F16F).	199

Πρόλογος

Η ευρεία χρήση και η ταχύτατη εξάπλωση του **Παγκόσμιου Ιστού Πληροφοριών (World Wide Web ή Web)** είχε σαν αποτέλεσμα τη μετατροπή του σε μια τεράστια αποθήκη πληροφοριών με δισεκατομμύρια σελίδες και περισσότερους από 800 εκατομμύρια χρήστες παγκοσμίως (Internet World Stats, 2005). Σήμερα θεωρείται ένα από τα πιο σημαντικά μέσα συλλογής, διαμοίρασης και διάδοσης πληροφοριών και υπηρεσιών. Την ίδια όμως στιγμή ο όγκος των διαθέσιμων πληροφοριών προκαλεί αρκετά προβλήματα που σχετίζονται με τη συνεχώς αυξανόμενη δυσκολία αναζήτησης, εύρεσης, οργάνωσης, πρόσβασης και συντήρησης της αιτούμενης πληροφορίας από τους χρήστες. Αυτό το γεγονός έχει σημαντικές επιπτώσεις στον τρόπο που σχεδιάζουμε και υλοποιούμε εφαρμογές στο Web και τα συστήματα **ηλεκτρονικής μάθησης (e-learning)**, **ηλεκτρονικού εμπορίου (e-commerce)** και **επιχειρηματικής ευφυΐας (business intelligence)** δεν θα μπορούσαν να αποτελέσουν εξαίρεση.

Μια λύση στο παραπάνω πρόβλημα έρχεται από την επιστημονική περιοχή του **Data Knowledge Mining (Εξόρυξη Γνώσης από Δεδομένα)** ή απλώς **Data Mining (Εξόρυξη Δεδομένων)** (Hand et al., 2001). Πράγματι η εξόρυξη γνώσης από μεγάλες αποθήκες δεδομένων έχει εξελιχθεί σε ένα από τα βασικότερα ερευνητικά ζητήματα. Μέχρι πρόσφατα η εξόρυξη γνώσης αφορούσε αποκλειστικά **δομημένα (structured) δεδομένα** δηλαδή δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων (databases). Τα τελευταία χρόνια το ενδιαφέρον στράφηκε και σε **μη δομημένα δεδομένα (semi-structured και unstructured)** π.χ. κείμενα, εικόνες, ιστοσελίδες (web pages), κλπ. Αυτό οδήγησε στη δημιουργία νέων κλάδων όπως το **Text Mining (Εξόρυξη Γνώσης από Κείμενα)** και το **Web Mining (Εξόρυξη Γνώσης από Δεδομένα του Web)**.

Η πρόσφατη εμφάνιση του **Semantic Web (Σημασιολογικός Παγκόσμιος Ιστός Πληροφοριών)** στοχεύει στην ενσωμάτωση ενός ακόμα επιπέδου **ευφυΐας (intelligence)** στο περιβάλλον των διαδικτυακών εφαρμογών. Σύμφωνα με τους Berners-Lee et al. (2001) *“the Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation”*. Ένας πιο επίσημος ορισμός από το W3C (2005) αναφέρει ότι *“the Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using eXtensible Markup Language (XML) for syntax and Uniform Resource Identifiers (URIs) for naming”*. Η ικανότητα του Semantic Web να προσθέτει νόημα στις πληροφορίες, αποθηκευμένο με τέτοιο τρόπο που να μπορεί να αναζητηθεί και να υποβληθεί σε επεξεργασία, παρέχει πολλές ευκαιρίες επέκτασης για τις web-based εφαρμογές (Anderson & Whitelock, 2004). Ο συνδυασμός του με το Web Mining έχει δημιουργήσει μια νέα και γρήγορα-αναδυόμενη ερευνητική περιοχή που είναι γνωστή ως **Semantic Web Mining**. Η ιδέα πίσω από τη χρησιμοποίησή του στη δημιουργία **εξατομικευμένων (personalized) web εμπειριών** είναι να βελτιωθεί το Web Mining με την εκμετάλλευση των νέων σημασιολογικών δομών (Markellou et al., 2004).

Με το συνδυασμό και την ολοκλήρωση τεχνικών και τεχνολογιών εξόρυξης γνώσης, οι e-εφαρμογές θα βελτιώσουν την ποιότητα των παρεχόμενων υπηρεσιών τους προς τους τελικούς χρήστες. Συγκεκριμένα, για να το επιτύχουν αυτό μελετούν τη συμπεριφορά των χρηστών και προσπαθούν να εξάγουν χρήσιμα συμπεράσματα από αυτή. Η απαίτηση για πρόβλεψη των αναγκών των χρηστών με στόχο την ευκολία χρήσης αλλά και τη διατήρηση των χρηστών ενός δικτυακού τόπου μπορεί να ικανοποιηθεί με την εφαρμογή της **εξατομίκευσης** ή **personalization**. Ορίζουμε σαν web personalization οποιαδήποτε ενέργεια που προσαρμόζει τις πληροφορίες ή τις υπηρεσίες που παρέχονται από ένα δικτυακό τόπο στις ανάγκες ενός συγκεκριμένου χρήστη ή ενός συνόλου χρηστών, χρησιμοποιώντας τη γνώση που πηγάζει από την ανάλυση της συμπεριφοράς πλοήγησης των χρηστών και από τα ενδιαφέροντά τους, σε συνδυασμό με τη δομή και το περιεχόμενο του δικτυακού τόπου. Στόχος είναι η παροχή στους χρήστες των πληροφοριών που θέλουν ή χρειάζονται χωρίς να απαιτεί από αυτούς να τις ζητήσουν με ρητό τρόπο. Η διαδικασία της εξατομίκευσης περιλαμβάνει την κατηγοριοποίηση και την προεπεξεργασία των δεδομένων, την εξαγωγή συσχετίσεων μεταξύ των δεδομένων και τον καθορισμό των ενεργειών που θα προταθούν.

Με στόχο την αντιμετώπιση των παραπάνω προβλημάτων, στα πλαίσια της παρούσας διδακτορικής διατριβής παρουσιάζονται τεχνικές και αλγόριθμοι που βελτιώνουν τις παρεχόμενες υπηρεσίες προς τον τελικό χρήστη. Αρχικά μελετάται **το πρόβλημα της εξόρυξης γνώσης από μια εφαρμογή e-learning με στόχο την παραγωγή εξατομίκευσης**. Ο όρος e-learning αφορά όλες τις μορφές εκπαίδευσης που βασίζονται στο Web και χρησιμοποιεί υπολογιστές και δίκτυα υπολογιστών για τη δημιουργία, την παράδοση, τη διαχείριση και την υποστήριξη online σειρών μαθημάτων. Τελευταία, χάρη στην προαναφερθείσα έκρηξη του Web, η έρευνα στο e-learning έχει κερδίσει όλο και περισσότερη προσοχή από τις επιστημονικές κοινότητες. Σημαντικοί εκπαιδευτικοί και εμπορικοί οργανισμοί καταδεικνύουν ένα συνεχές ενδιαφέρον για την περιοχή, κίνηση που αποτελεί μια πολύ ισχυρή κατευθυντήρια δύναμη εκφρασμένη μέσα από πολυάριθμη έρευνα και εμπορικές προσπάθειες. Η μεγάλη ποικιλία των διαθέσιμων συστημάτων και εφαρμογών e-learning αποτελεί άλλωστε μια ισχυρή ένδειξη της ωριμότητας στην περιοχή (Urdan & Weggen, 2000), (Wentling et al., 2000), (Fry, 2001).

Εντούτοις, στην πλειοψηφία των e-learning συστημάτων του παρελθόντος, οι σειρές των μαθημάτων και το εκπαιδευτικό υλικό δεν ήταν αρκετά δυναμικά, παρέχοντας ένα μάλλον περιορισμένο σύνολο χαρακτηριστικών, με πολύπλοκη δομή και δυσκολίες στην πλοήγηση με αποτέλεσμα να μη μπορούν να ανταποκριθούν αποδοτικά στις ανάγκες και στις απαιτήσεις των χρηστών, οδηγώντας σε «φτωχές» εκπαιδευτικές εμπειρίες. Γενικά, το εκπαιδευτικό υλικό με ενσωματωμένους **υπερσυνδέσμους (hyperlinks)** επιτρέπει στους εκπαιδευόμενους να ακολουθήσουν οποιοδήποτε μονοπάτι πλοήγησης επιθυμούν και να μη χρησιμοποιήσουν απαραίτητα τη δομή που έχουν καθορίσει οι σχεδιαστές των δικτυακών τόπων ή οι δημιουργοί του περιεχομένου (οι οποίοι έχουν ένα ορισμένο σχέδιο πλοήγησης στο μυαλό τους). Αυτή η ελευθερία μπορεί να αποδειχτεί αρνητικός παράγοντας, δεδομένου ότι σε πολλές περιπτώσεις οι εκπαιδευόμενοι δεν έχουν την απαραίτητη ωριμότητα και ικανότητα να ακολουθήσουν μια αποτελεσματική πορεία και συχνά να οδηγούνται σε θεματικά που είναι είτε πάρα πολύ δύσκολα, είτε πάρα πολύ εύκολα, ή ακόμα και άσχετα με τις προσωπικές τους εκπαιδευτικές ανάγκες (Markellos et al., 2004).

Μια λύση στο πρόβλημα αυτό αποτελεί η ενσωμάτωση **προσαρμοστικών χαρακτηριστικών (adaptive features)** που επιτρέπουν την παροχή **εξατομικευμένης μάθησης (personalized learning)** (Brusilovsky, 2001). Αυτά τα συστήματα είναι η απάντηση στην παραδοσιακή «one-size-fits-all» προσέγγιση που παρέχει το ίδιο στατικό υλικό εκπαίδευσης σε όλους τους χρήστες, χωρίς να λαμβάνει υπόψη τις ανάγκες και τις προτιμήσεις, τις εμπειρίες και τις γνώσεις τους, οι οποίες μπορεί να ποικίλουν εντυπωσιακά (De Bra et al., 1999). Με τον τρόπο αυτό οι εφαρμογές e-learning παρέχουν υψηλής ποιότητας περιεχόμενο, αποδοτική δομή, εύκολη πλοήγηση καθώς επίσης και πλήρη υποστήριξη για τις διαφορετικές εργασίες όλων των προφίλ χρηστών που συμμετέχουν σε ένα τυπικό σενάριο εκπαίδευσης από απόσταση (Mödritscher et al., 2004). Συγκεκριμένα, ανάλογα με το γνωστικό υπόβαθρο του εκπαιδευομένου, τις γνώσεις και τις αδυναμίες του, το εκπαιδευτικό στυλ που προτιμάει, την πρόοδο που έχει σημειώσει μέχρι τώρα, το σύστημα αποφασίζει τι εκπαιδευτικό υλικό θα του παρουσιάσει και με ποιον τρόπο. Πιθανές παράμετροι αποτελούν τα διαφορετικά εκπαιδευτικά μονοπάτια, οι διαφορετικοί τρόποι παρουσίασης του ίδιου περιεχομένου (π.χ. με ή χωρίς ήχο) ή η προσφορά ενός διαφορετικού συνόλου λειτουργιών που το **περιβάλλον διεπαφής (user interface)** του συστήματος εκπαίδευσης μπορεί να παρέχει για να μειώσει την πολυπλοκότητα (Dietinger, 2003).

Στην προσέγγιση που παρουσιάζεται στη διδακτορική διατριβή για την υποστήριξη του εξατομικευμένου e-learning, η δομή και η σχέση των δεδομένων και των πληροφοριών παίζουν ουσιαστικό ρόλο. Ο προτεινόμενος αλγόριθμος βασίζεται σε μια **οντολογία (ontology)** η οποία βοηθά στη δόμηση και στη διαχείριση του περιεχομένου που σχετίζεται με μια δεδομένη σειρά μαθημάτων, ένα μάθημα ή ένα θεματικό (Markellou et al., 2005a), (Markellou et al., 2006). Ειδικότερα, το πλαίσιο εξατομίκευσης λαμβάνει υπόψη τα προφίλ των χρηστών και την οντολογία του θεματικού, ενώ η διαδικασία χωρίζεται σε δύο στάδια: στις offline ενέργειες προετοιμασίας των δεδομένων, δημιουργίας της οντολογίας και **εξόρυξης από δεδομένα χρήσης (usage mining)** και στην online παροχή εξατομίκευσης.

Το σύστημα βρίσκεται σε πρώτη φάση ένα αρχικό σύνολο συστάσεων βασισμένο στην οντολογία του πεδίου και στη συνέχεια χρησιμοποιεί τα **frequent itemsets (συχνά εμφανιζόμενα σύνολα στοιχείων)** για να το εμπλουτίσει, λαμβάνοντας υπόψη την πλοήγηση άλλων **παρόμοιων χρηστών (similar users)**. Με τον τρόπο αυτό, μειώνουμε το χρόνο που απαιτείται για την ανάλυση όλων των frequent itemsets και των κανόνων συσχέτισης. Εστιάζουμε μόνο σε εκείνα τα σύνολα που προέρχονται από το συνδυασμό της **ενεργούς συνόδου (current session)** του χρήστη και των συστάσεων της οντολογίας. Η χρονική μείωση προκύπτει λόγω του γεγονότος ότι τα frequent itemsets φιλτράρονται με το σύνολο των συστάσεων της οντολογίας οδηγώντας σε ένα μικρότερο χώρο αναζήτησης. Επιπλέον, αυτή η προσέγγιση ανακουφίζει και το πρόβλημα των μεγάλων χρόνων απόκρισης, το οποίο μπορεί στη συνέχεια να οδηγήσει στην εγκατάλειψη του e-learning συστήματος. Αν και η εξατομίκευση απαιτεί αρκετά βήματα επεξεργασίας και ανάλυσης, το εμπόδιο αυτό αποφεύγεται με την εκτέλεση σημαντικού μέρους της διαδικασίας offline.

Σε ένα δεύτερο επίπεδο στα πλαίσια της διδακτορικής διατριβής μελετάται το **πρόβλημα της παραγωγής προτάσεων σε μια εφαρμογή e-εμπορίου**. Τα **συστήματα συστάσεων (recommendations systems ή RSs)** αποτελούν ίσως την πιο δημοφιλή μορφή εξατομίκευσης και τείνουν να μετατραπούν στις μέρες μας σε σημαντικά επιχειρησιακά εργαλεία (Μαρκέλλου, 2003). Αρχικά, εμφανίστηκαν στα μέσα της δεκαετίας του '90 και

χρησιμοποιούνταν από λίγα web sites. Σήμερα δεν νοούνται εφαρμογές e-εμπορίου που να μην ενσωματώνουν τεχνολογίες RSs προκειμένου να βοηθήσουν τους πελάτες στις αγορές τους π.χ. Amazon.com, το CDNow.com, Yahoo.com, eBay.com, κλπ. Με αυτό τον τρόπο, τα **e-shops (ηλεκτρονικά καταστήματα)** μπορούν να βελτιώσουν την απόδοσή τους καλύπτοντας τις ανάγκες και τις προτιμήσεις κάθε πελάτη ξεχωριστά, να αυξήσουν την ικανοποίησή τους, να προωθήσουν το e-loyalty και να καθιερώσουν ένα-προς-ένα σχέσεις (Markellou et al., 2005c).

Γενικά, τα RSs μπορούν να ταξινομηθούν σε τρεις βασικές κατηγορίες: **content based filtering (φιλτράρισμα με βάση το περιεχόμενο)**, **collaborative filtering (συνεργατικό φιλτράρισμα)** και **hybrid approaches (υβριδικές προσεγγίσεις)**. Το content based filtering λαμβάνει υπόψη του τις παραμέτρους χρήστη με τις πληροφορίες των προϊόντων (μεταδεδομένα, λέξεις κλειδιά, κλπ.) ενώ το collaborative filtering (CF) την «ομοιότητα» των χρηστών προκειμένου να παράγουν συστάσεις. Οι δύο αυτές μέθοδοι μπορούν να συνδυαστούν σε υβριδικές προσεγγίσεις. Συγκεκριμένα, ένα RS σύστημα εκμεταλλεύεται τις απόψεις των χρηστών ή/και των κοινοτήτων χρηστών προκειμένου να υποστηρίξει τον πελάτη να προσδιορίσει τις πληροφορίες ή τα προϊόντα για τα οποία ενδιαφέρεται και είναι κοντά στις ανάγκες και τις προτιμήσεις του. Οι συστάσεις μπορούν να παρουσιαστούν με πολλές μορφές π.χ. εξατομικευμένες προσφορές, τιμές, προϊόντα ή υπηρεσίες, προσθήκη ή παράλειψη θεματικών ενοτήτων ή παραγράφων, ταξινόμηση, κρύψιμο, εμφάνιση, προσθήκη, παράληψη ή έμφαση συνδέσμων, προαιρετικές εξηγήσεις ή αναλυτικές πληροφορίες, κλπ. (Μαρκέλλου, 2005).

Η προτεινόμενη υβριδική προσέγγιση (Markellou et al., 2005b) στοχεύει στην παραγωγή αποτελεσματικών συστάσεων για τους πελάτες ενός online καταστήματος που νοικιάζει κινηματογραφικές ταινίες. Η γνώση για τους πελάτες και τα προϊόντα προκύπτει από **δεδομένα χρήσης (usage data)** και τη **δομή της οντολογίας (ontology structure)** σε συνδυασμό με τις **εκτιμήσεις-βαθμολογίες (ratings)** των πελατών για τις ταινίες καθώς και με την εφαρμογή **τεχνικών ταιριάσματος (matching)** «όμοιων» (similar) πελατών. Αρχικά, το σχήμα της οντολογίας αποτυπώνει το πεδίο του e-shop με τον προσδιορισμό όλων εννοιών του, των σχέσεων μεταξύ τους και άλλων ιδιοτήτων, όρων ή κανονισμών. Το πιλοτικό e-shop αποτελείται από μια συλλογή ιστοσελίδων που περιέχουν πληροφορίες για τις ταινίες (τίτλος, κατηγορία, στούντιο, ηθοποιός, σκηνοθέτης, παραγωγός, βραβεία, έτος, διάρκεια, κοινό, περίληψη, κλπ.). Τα αρχικά στοιχεία ανακτήθηκαν από τη βάση δεδομένων Internet Movie Database (<http://www.imdb.com>) και στη συνέχεια οργανώθηκαν στο οντολογικό σχήμα προκειμένου να αποτελέσουν ένα πλούσιο σύνολο δεδομένων (superset) το οποίο χρησιμοποιήθηκε στη **διαδικασία ανακάλυψης γνώσης (knowledge discovery process)**. Συγκεκριμένα, αποτέλεσε την είσοδο στη φάση εξόρυξης γνώσης ώστε να συνδυαστούν και να μετασχηματιστούν τα υπονοούμενα/ρητά δεδομένα (κλάση χρηστών, ιστορικό, προφίλ, δομή και περιεχόμενο e-shop) σε νέες μορφές. Το RS λαμβάνει επίσης υπόψη τις εκτιμήσεις-βαθμολογίες των χρηστών για τις ταινίες σε μια κλίμακα από 1 έως 10, όπου 1 σημαίνει μια πολύ κακή ταινία και 10 μια πολύ καλή και παράγει ως έξοδο μια λίστα πιθανών προτάσεων για κάθε χρήστη ξεχωριστά.

Παράλληλα η οντολογία δημιουργεί συνδέσεις μεταξύ των ταινιών σύμφωνα με τα χαρακτηριστικά τους (attributes). Χρησιμοποιώντας το **ιστορικό των συναλλαγών (transactions history file)** του συγκεκριμένου πελάτη, οι προτιμήσεις του μπορούν να προβλεφτούν. Για παράδειγμα, μπορεί να βρεθεί εάν σε ένα πελάτη αρέσει ή όχι μια ταινία

μια συγκεκριμένης κατηγορίας, ενός συγκεκριμένου σκηνοθέτη, ενός συγκεκριμένου ηθοποιού ή συνδυασμών άλλων χαρακτηριστικών. Όταν ένα ή περισσότερα κριτήρια ταιριάσματος ικανοποιούνται, τότε άλλες ταινίες μπορούν να προσδιοριστούν σύμφωνα με το οντολογικό σχήμα που έχουν παρόμοια χαρακτηριστικά με αυτές που ο πελάτης έχει ήδη νοικιάσει. Στην περίπτωση ενός νέου πελάτη όπου το ιστορικό του είναι κενό, οι πληροφορίες από την **αίτηση εγγραφής (registration form)** του αναλύονται ώστε να ταξινομηθεί σε μια συγκεκριμένη **κλάση πελατών (user class)**, ενώ οι προτάσεις βασίζονται στο οντολογικό σχήμα. Αυτή η ολοκλήρωση παρέχει πρόσθετη γνώση για τις προτιμήσεις των πελατών και επιτρέπει την παραγωγή πιο επιτυχημένων συστάσεων. Ακόμη και στην περίπτωση του **«cold-start problem»** όπου δεν είναι διαθέσιμη αρχική πληροφορία για τη συμπεριφορά του πελάτη, η προσέγγιση μπορεί να προβεί σε λογικές και σχετικές συστάσεις. Τα πειραματικά αποτελέσματα έδειξαν ότι ο συνδυασμός τεχνικών από τις περιοχές του **Web Mining** και του **Semantic Web** και η ενσωμάτωσή τους σε ένα υβριδικό RS μπορεί να οδηγήσει σε ποιοτικότερες και πιο επιτυχημένες συστάσεις.

Τέλος, σε ένα τρίτο επίπεδο μελετάται το **πρόβλημα της εξόρυξης γνώσης από κείμενα διπλωμάτων ευρεσιτεχνίας με στόχο την παραγωγή επιστημονικών και τεχνολογικών δεικτών για την ανάδειξη του επιπέδου καινοτόμου δραστηριότητας μιας αγοράς και συνεπώς την επιχειρηματική ευφυΐα (business intelligence ή BI)**. Είναι γεγονός ότι τις τελευταίες δεκαετίες η συγκέντρωση τεράστιων ποσοτήτων δεδομένων από τις επιχειρήσεις και όχι μόνο έχει δημιουργήσει την ανάγκη για γρήγορη, συνεπή και αποδοτική διαχείρισή τους. Τα δεδομένα αυτά είναι αποθηκευμένα σε **μεγάλες βάσεις δεδομένων (databases)** ή **αποθήκες δεδομένων (data warehouses)** και μπορεί να προέρχονται από την καταγραφή συστημάτων δοσοληψιών που λειτουργούν επί πολλά χρόνια π.χ. τραπεζικές συναλλαγές, τηλεφωνικές κλήσεις, αγορές με πιστωτικές κάρτες, κλπ. ή διαδικασίες που συσσωρεύουν μεγάλα ψηφιακά αρχεία π.χ. δορυφορικές φωτογραφίες, πειραματικά δεδομένα, κλπ. Σε αυτά τα δεδομένα συνήθως εφαρμόζονται α) **σχεσιακές ερωτήσεις** με στόχο την επιλογή ενός μικρού υποσυνόλου τους που ικανοποιεί κριτήρια ορισμένα από το χρήστη ή/και β) **στατιστική επεξεργασία** με τη μορφή αθροιστικών συναρτήσεων που εφαρμόζονται σε πολλές διαστάσεις. Και στις δύο περιπτώσεις οδηγούμαστε στην παραγωγή πολύ συγκεκριμένων αποτελεσμάτων τόσο ως προς τη δομή όσο και ως προς τη σημασία τους (Βαζιργιάννης & Χαλκίδη, 2003).

Από την άλλη πλευρά είναι πολύ πιθανό ο τεράστιος αυτός όγκος πληροφοριών να κρύβει μέσα του **ανεξερεύνητη «γνώση»**. Η γνώση αυτή, αν και όχι γνωστή εκ των προτέρων μπορεί να αποδειχθεί εξαιρετικά χρήσιμη κυρίως για τις επιχειρήσεις. Συγκεκριμένα, συσχετίσεις μεταξύ των δεδομένων, εξαρτήσεις και αλληλουχίες γεγονότων, προβλέψεις για στόχους και πολλά άλλα είναι «θαμμένα» μέσα σε αυτά δεδομένα και περιμένουν με κάποιον τρόπο να έρθουν στην επιφάνεια. Το ερώτημα λοιπόν που κυριαρχεί είναι πως αυτή η γνώση θα ανακαλυφθεί. Μια απάντηση έρχεται από το χώρο της BI που ορίζεται σαν **«μια αμφίδρομη διαδικασία ανάλυσης και αξιοποίησης πληροφοριών από διάφορες πηγές που είναι αποθηκευμένες σε μεγάλες βάσεις δεδομένων με στόχο την εξαγωγή χρήσιμων συμπερασμάτων για τη λήψη αποφάσεων»** (McDonald et al., 2002). Με τον τρόπο αυτό η διοίκηση μιας επιχείρησης έχει τη δυνατότητα να παρακολουθεί αποτελεσματικότερα τις νέες τάσεις και τα πρότυπα, να μοιράζεται πληροφορίες με συνεργάτες ή πελάτες και να λαμβάνει καλύτερες επιχειρηματικές αποφάσεις. Σε αυτό το πλαίσιο μπορούμε να θεωρήσουμε ότι η BI αφορά σε **πελάτες, προϊόντα, υπηρεσίες και ανταγωνιστές**.

Γενικά η ΒΙ δεν είναι μια καινούργια έννοια στην περιοχή της Πληροφορικής αφού πρώιμες εκδοχές της συναντάμε στα τέλη της δεκαετίας του '70. Ένα διαδεδομένο παράδειγμα είναι η τοποθέτηση των προϊόντων στα ράφια των super-markets π.χ. οι σάλτσες κοντά στα ζυμαρικά, οι ξηροί καρποί κοντά στα ποτά, κλπ. Τα τελευταία όμως χρόνια με τη συσσώρευση των πληροφοριών η ανάγκη για την ανακάλυψη της γνώσης έγινε επιτακτική. Οι μέθοδοι που χρησιμοποιούνται κυρίως αναλύουν τα δεδομένα, εξαγουν συμπεράσματα και βασίζονται σε στατιστικούς κανόνες. Το **Data Mining (DM)** και το **OLAP (Online Analytical Processing)** είναι δύο από τις πιο γνωστές μεθόδους. Όπως έχει αναφερθεί οι τεχνικές DM είναι σχεδιασμένες να εφαρμόζονται σε **δομημένες βάσεις δεδομένων (structured databases)**, ενώ οι τεχνικές **Text Mining (TM)** έχουν αναπτυχθεί για την ανακάλυψη γνώσης μέσα από την επεξεργασία **αδόμητων (unstructured)** ή **ημιδομημένων εγγράφων (semi-structured documents)**. Συνεπώς, το TM αποτελεί μια επέκταση του παραδοσιακού DM και αφορά στην εξόρυξη υπονοούμενης γνώσης που περιλαμβάνεται σε μεγάλες συλλογές εγγράφων. Τα τελευταία χρόνια οι τεχνικές αυτές παρουσιάζουν ιδιαίτερο ενδιαφέρον αφού η πλειοψηφία των δεδομένων σήμερα σε ποσοστό μεγαλύτερο του 80% είναι μη δομημένα ή ημιδομημένα π.χ. κείμενα, εικόνες, έγγραφα, web pages, κλπ. Μάλιστα το ποσοστό αυτό αναμένεται να αυξηθεί στο μέλλον λόγω της ευρείας χρήσης του Internet και του e-mail.

Μια σημαντική πηγή γνώσεων στον τομέα της τεχνολογίας αποτελούν οι καταχωρήσεις των **διπλωμάτων ευρεσιτεχνίας ή πατέντες (patents)**. Ένα δίπλωμα ευρεσιτεχνίας είναι ένας νομικός τίτλος που χορηγεί στον κάτοχό του το αποκλειστικό δικαίωμα να χρησιμοποιήσει μια εφεύρεση σε συγκεκριμένη περιοχή και χρόνο, αποτρέποντας παράλληλα άλλους να τη χρησιμοποιήσουν ή πουλήσουν χωρίς έγκριση (EPO, 2005). Τα διπλώματα ευρεσιτεχνίας καταδεικνύουν το επίπεδο της καινοτόμου δραστηριότητας μιας αγοράς (Narin, 1995). Η εφαρμογή τεχνικών εξόρυξης γνώσης σε τέτοιου είδους δεδομένα μπορεί να οδηγήσει σε πολύ χρήσιμα συμπεράσματα σχετικά με την τεχνολογική ανάπτυξη, τις τάσεις της τεχνολογίας αλλά και τις καινοτομικές κατευθύνσεις (Chappelier et al., 2002). Η προτεινόμενη προσέγγιση αφορά στην εφαρμογή τεχνικών TM σε διπλώματα ευρεσιτεχνίας που βρίσκονται καταχωρημένα σε βάσεις δεδομένων διαφόρων διεθνών οργανισμών διαχείρισής τους, με **στόχο την παραγωγή επιστημονικών και τεχνολογικών δεικτών για την ανάδειξη του επιπέδου καινοτόμου δραστηριότητας μιας αγοράς**. Ένα δίπλωμα ευρεσιτεχνίας αποτελείται από διάφορα **πεδία (fields)**, κάθε ένα από τα οποία περιέχει συγκεκριμένες πληροφορίες. Ένας κωδικός (συχνά χρησιμοποιούνται περισσότεροι από ένας κωδικοί) χαρακτηρίζει την πατέντα δηλώνοντας τα τεχνικά χαρακτηριστικά της. Γενικά, οι πληροφορίες αυτές μπορούν να ταξινομηθούν στις ακόλουθες τέσσερις κατηγορίες α) **τεχνολογικά** χαρακτηριστικά γνωρίσματα της εφεύρεσης, β) χαρακτηριστικά σχετικά με την **εφαρμογή** της εφεύρεσης, γ) πληροφορίες για τις **χώρες** στις οποίες η εφεύρεση προστατεύεται και δ) πληροφορίες για την **προέλευση** της εφεύρεσης.

Σύμφωνα με την προτεινόμενη προσέγγιση (Markellos et al., 2002), (Markellos et al., 2003) αρχικά τα δεδομένα καθαρίζονται προκειμένου να βελτιωθεί η ποιότητά τους πριν την επεξεργασία. Αυτή η διαδικασία συχνά αναφέρεται σαν «**αποφυγή της Garbage in – Garbage out επίδρασης ή GIGO**». Στη συνέχεια εφαρμόζονται δύο τύποι επεξεργασίας στα δεδομένα η **απλή ανάλυση (simple analysis)** και η **στατιστική ανάλυση (statistical analysis)**. Στην πρώτη περίπτωση παράγονται γραφήματα που συσχετίζουν τις πληροφορίες π.χ. κύριοι τομείς ανάπτυξης σε μια χώρα. Στη δεύτερη περίπτωση αναλύονται **γλωσσολογικά** τα κείμενα των διπλωμάτων ευρεσιτεχνίας (τα πεδία title και abstract) και

ομαδοποιούνται τα λήμματα των λέξεων. Η **μορφολογική κανονικοποίηση** αναφέρεται σε μια ομάδα διαδικασιών φυσικής γλώσσας (NLP – Natural Language Processing) που στοχεύουν στην παραγωγή των κανονικών μορφών των λέξεων. Βασικός στόχος της παραπάνω διαδικασίας είναι η εξαγωγή της ρίζας των λέξεων από την κλιμένη ή παραγόμενη μορφή τους, με τη εφαρμογή ενός συνόλου κανόνων. Αυτός ο τύπος μορφολογικής κανονικοποίησης είναι ευρέως γνωστός σαν διαδικασία **stemming (ανάλυση γενεαλογικού δένδρου)** και εμπεριέχει διαδικασίες **part of speech tagging (δήλωση σε ποιο μέρος του λόγου ένας όρος ανήκει)** για την εξαγωγή πληροφορίας όσον αφορά στους τομείς με τους οποίους έχει σχέση το κάθε δίπλωμα ευρεσιτεχνίας.

Στη συνέχεια πάνω στα δεδομένα των διπλωμάτων ευρεσιτεχνίας εφαρμόζονται τεχνικές **correspondence** και **clustering analysis** έτσι ώστε αυτά να ομαδοποιηθούν σύμφωνα με τις τεχνολογίες στις οποίες αναφέρονται. Τα clusters πλέον αυτά προβάλλονται όπως και στην απλή ανάλυση παρέχοντας στο χρήστη μια πιο λεπτομερή απεικόνιση της πληροφορίας των διπλωμάτων ευρεσιτεχνίας. Ο συνδυασμός των αναλύσεων που εφαρμόζονται με βάση την προτεινόμενη μεθοδολογία επιτρέπει την αποτύπωση των τεχνολογικών εξελίξεων και καινοτομιών. Οι δείκτες που παράγονται για κάθε cluster είναι πολύ σημαντικοί αφού μπορούν να ποσοτικοποιήσουν τις πληροφορίες που αφορούν σε συγκεκριμένες τεχνολογίες. Με αυτό τον τρόπο μπορούμε να παράγουμε δείκτες για τη δραστηριότητα συγκεκριμένων φορέων, εφευρετών, χωρών, κλπ. Τέλος, τεχνολογικοί δείκτες που υποδεικνύουν μελλοντικές ελπιδοφόρες τεχνολογίες καθώς και ποιοι φορείς θα είναι πρωτοπόροι σε αυτές μπορούν να εξαχθούν.

Βιβλιογραφία

- Anderson, T., & Whitelock, D. (2004). The Educational Semantic Web: Visioning and Practicing the Future of Education. *Journal of Interactive Media in Education (JIME), Special Issue, Educational Semantic Web*, 1-15.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5): 34-43.
- Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modelling and User-Adapted Interaction*, 11: 87-110.
- Chappelier, J., Peristera, V., Rajman, M., & Seydoux, F. (2002). Evaluation of Statistical and Technological Innovation Using Statistical Analysis of Patents, *JADT*.
- De Bra, P., Brusilovsky, P., & Houben, G.J. (1999). Adaptive Hypermedia: from Systems to Framework. *ACM Computing Surveys*, 31(4es).
- Dietinger, Th. (2003). Aspects of e-Learning Environments. *PhD Thesis*, Institute for Information Processing and Computer Supported New Media (IICM), Graz University of Technology, Austria.
- EPO (2005). European Patent Office. Retrieved from: <http://www.european-patent-office.org/index.htm>.

Fry, K. (2001). E-learning Markets and Providers: Some Issues and Prospects. *Education and Training, Emerland*, 43(4): 233-239.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Internet World Stats (2005). Internet Usage Statistics - The Big Picture. World Internet Users and Population Stats. Retrieved from: <http://www.internetworldstats.com/stats.htm>.

Markellos, K., Markellou, P., Mayritsakis, G., Panagopoulou, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis (2003). STING: Evaluation of Scientific and Technological Innovation and Progress in Europe through Patents. In the book *Statistical Data Mining and Knowledge Discovery*, Bozdogan, H. (Ed.), Chapman and Hall/CRC Press, ISBN: 1584883448, 29th July. Retrieved from: <http://www.amazon.com/exec/obidos/ASIN/1584883448/qid%3D1066835737/sr%3D-3/ref%3Dsr%5F2%5F3/104-9162346-7852756>.

Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2002). Knowledge Discovery in Patent Databases. *Proceedings of the ACM CIKM 2002, 11th International Conference on Information and Knowledge Management*, Saic Headquarters, McLean, Virginia, USA, 4-9 November, 672-674. Retrieved from: <http://portal.acm.org/citation.cfm?id=584915&coll=GUIDE&dl=GUIDE&ret=1#FullText>.

Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004a). Web Personalization for Enhancing e-Learning Experience. *Proceedings of the 5th International Conference on Information Communication Technologies in Education, ICICTE*, Samos Island, Greece.

Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2006). Semantic Web Mining for E-Learning Personalization. *International Journal of Knowledge and Learning (IJKL)*, Inderscience Publishers, to appear.

Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005a). Using Semantic Web Mining Technologies for Personalized e-Learning Experiences. *Proceedings of the 4th IASTED International Conference on Web-Based Education (WBE-2005)*, Grindelward, Switzerland, 21-23 February, 522-527.

Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005b). Personalized E-commerce Recommendations. *Proceedings of the IEEE International Conference on e-Business Engineering (ICEBE 2005)*, Beijing, China, 18-20 October.

Markellou, P., Rigou, M., & Sirmakessis, S. (2005c). An Online Consumer Purchase Decision Cycle. In the book *Advances in Electronic Marketing: Volume 1*, Clarke, I., & Flaherty, T. (Eds.), James Madison University, USA, Idea Group Publishing Inc., 287-299.

Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004). Personalization in the Semantic Web Era: a Glance Ahead. *Proceedings of the 5th International Conference on Data Mining, Text Mining and their Business Applications, Data Mining 2004*, Zanas, A., Ebecken, N.F., & Brebbia, C.A. (Eds.), Wessex Institute of Technology (UK), Malaga, Spain, Southampton, Boston: PWIT Press, 3-11.

McDonald, K., Wilmsmeier, A., Dixon, D.C., Inmon, W.H. (2002). *Mastering the SAP Business Information Warehouse*. Wiley Publishers, ISBN 0471219711.

Mödrtscher, F., Barrios, V.M.G., & Gütl, C. (2004). The Past, the Present and the Future of Adaptive e-Learning. *Proceedings of the ICL2004, International Conference on Interactive Computer Aided Learning*, Villach, Austria.

Narin, F. (1995). Patents as Indicators for the Evaluation of Industrial Research Output. *Scientometrics*, 34(3): 489-496.

Urdan, T., & Weggen, C. (2000). *Corporate e-Learning: Exploring a New Frontier*, WR Hambrecht and Co.

W3C (2005). World Wide Web Consortium: Semantic Web. Retrieved from: <http://www.w3.org/2001/sw>.

Wentling, T., Waight, C., Gallaher, J., La Fleur, J., Wang, C., & Kanfer, A. (2000). E-learning: a Review Literature. *Knowledge and Learning Systems Group*, National Center for Supercomputing Applications, University of Illinois.

Μαρκέλλου, Π. (2003). Recommendation Συστήματα. Κεφάλαιο στις διδακτικές σημειώσεις *e-Επιχειρείν* για το προπτυχιακό μάθημα «e-Επιχειρείν», Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής της Πολυτεχνικής Σχολής του Πανεπιστημίου Πατρών, Εκδόσεις Πανεπιστημίου Πατρών.

Μαρκέλλου, Π. (2005). Ηλεκτρονικό Κατάστημα: Σχεδιασμός και Λειτουργίες. Κεφάλαιο 3 στο βιβλίο *e-Επιχειρηματικότητα: από τη Θεωρία στην Πράξη*, Εκδόσεις Ελληνικά Γράμματα – Ε.Α.Ι.Τ.Υ. CTI press, Σειρά: Τεχνολογία Πληροφορικής και Επικοινωνιών, 105-144, υπό έκδοση.

Βασικές δημοσιεύσεις διατριβής

Κεφάλαια σε Βιβλία

- [1] Markellou, P., Rigou, M., & Sirmakessis, S. (2005). An Online Consumer Purchase Decision Cycle. Chapter in the book *Advances in Electronic Marketing*, Clarke, I., & Flaherty, T. (Eds.), Idea Group Publishing Inc., 287-299. Retrieved from: <http://cob.jmu.edu/flahertb/emarketingbook>.
- [2] Markellou, P., Rigou, M., & Sirmakessis, S. (2005). Product Catalog and Shopping Cart Effective Design. Chapter in the book *Web Systems Design and Online Consumer Behavior*, Gao, Y. (Ed.), Idea Group Publishing Inc., 48-68.
- [3] Markellou, P., Rigou, M., & Sirmakessis, S. (2005). Web Personalization for E-Marketing Intelligence. Chapter in the book *Contemporary Research in E-Marketing: Volume 1*, Krishnamurthy, S. (Ed.), Idea Group Publishing Inc., 232-250.
- [4] Markellos, K., Markellou, P., Mayritsakis, G., Panagopoulou, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis (2003). STING: Evaluation of Scientific and Technological Innovation and Progress in Europe through Patents. Chapter in the book *Statistical Data Mining and Knowledge Discovery*, Bozdogan, H. (Ed.), Chapman and Hall/CRC Press, ISBN: 1584883448. Retrieved from: <http://www.amazon.com/exec/obidos/ASIN/1584883448/qid%3D1066835737/sr%3D-3/ref%3Dsr%5F2%5F3/104-9162346-7852756>.
- [5] Μαρκέλλου, Π. (2005). Ηλεκτρονικό Κατάστημα: Σχεδιασμός και Λειτουργίες. Κεφάλαιο στο βιβλίο *e-Επιχειρηματικότητα: από τη Θεωρία στην Πράξη*, Εκδόσεις Ελληνικά Γράμματα – CTI press, Σειρά: Τεχνολογία Πληροφορικής και Επικοινωνιών, 105-144, υπό έκδοση.
- [6] Μαρκέλλου, Π. (2003). Recommendation Συστήματα. Κεφάλαιο στις διδακτικές σημειώσεις *e-Επιχειρείν* για το προπτυχιακό μάθημα «e-Επιχειρείν», Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής της Πολυτεχνικής Σχολής του Πανεπιστημίου Πατρών, Εκδόσεις Πανεπιστημίου Πατρών.
- [7] Markellou, P., Panayiotaki, A., & Tsakalidis, A. (2006). Web Mining for Public e-Services Personalization. Article in the *Encyclopedia of Digital Government*, Anttiroiko, A., & Malkia, M. (Eds.), Idea Group Publishing Inc., in press.
- [8] Markellou, P., Rigou, M., & Sirmakessis, S. (2006). A Closer Look to the Online Consumer Behavior. Article in the *Encyclopedia of E-Commerce, E-Government and Mobile Commerce*, Khosrow-Pour, M. (Ed.), Idea Group Publishing Inc., in press.

- [9] Markellos, K., Markellou, P., Panayiotaki, A., & Tsakalidis, A. (2006). Semantic Web Mining for Personalized Public e-Services. Chapter in the book *Global E-Government: Theory, Applications and Benchmarking*, Al-Hakim, L., & Soliman, K., (Eds.), Idea Group Publishing Inc., to appear.

Διεθνή Επιστημονικά Περιοδικά

- [1] Markellou, P., Mousourouli, I., Spiros, S., & Tsakalidis, A. (2006). Semantic Web Mining for E-Learning Personalization. *International Journal of Knowledge and Learning (IJKL)*, Inderscience Publishers, to appear.
- [2] Markellou, P., Rigou, M., & Sirmakessis, S. (2006). Looking Inside Online Consumers Behavior. *International Journal of Social Behavior and Personality*, to appear.
- [3] Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2005). Shaping Online Learning Communities and the Way Adaptiveness Adds to the Picture. *International Journal of Knowledge and Learning (IJKL)*, Inderscience Publishers, 1(1/2): 80-95.
- [4] Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2000). A Web Adaptive Educational System for People with Hearing Difficulties. *Education and Information Technologies Journal (the Official Journal of the IFIP Technical Committee on Education)*, Kluwer Academic Publishers, 5(3): 189-200. Retrieved from: <http://www.kluweronline.com/oasis.htm/275677>.

Εθνικά Επιστημονικά Περιοδικά

- [1] Μάρκελλος, Κ., Μαρκέλλου, Π., Ρήγκου, Μ., Συρμακέσης, Σ., & Τσακαλίδης, Α. (2004). Εξατομικευμένη Ηλεκτρονική Μάθηση. *Περιοδικό Επιστήμες της Αγωγής, Ειδική Έκδοση «Δια βίου και Εξ Αποστάσεως Εκπαίδευση στην Κοινωνία της Πληροφορίας»*, Παιδαγωγικό Τμήμα Δ.Ε. Πανεπιστημίου Κρήτης, Σεπτέμβριος.

Διεθνή Επιστημονικά Συνέδρια

- [1] Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005). Personalized E-commerce Recommendations. *Proceedings of the IEEE International Conference on e-Business Engineering (ICEBE'2005)*, Beijing, China, 18-20 October.
- [2] Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005). Using Semantic Web Mining Technologies for Personalized e-Learning Experiences. *Proceedings of the 4th IASTED International Conference on Web-Based*

Education (WBE-2005), Web Based Teaching and Learning Technologies, Grindelward, Switzerland, 21-23 February, 522-527.

- [3] Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2002). Knowledge Discovery in Patent Databases. *Proceedings of the ACM CIKM 2002, 11th International Conference on Information and Knowledge Management*, Saic Headquarters, McLean, Virginia, USA, 4-9 November, 672-674. Retrieved from: <http://portal.acm.org/citation.cfm?id=584915&coll=GUIDE&dl=GUIDE&ret=1#FullText>.
- [4] Markellos, K., Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005). Semantic Web Usage Mining for E-commerce Personalization. *Proceedings of the 4th International Conference New Horizons in Industry, Business and Education (NHIBE'05)*, Corfu, Greece, 25-26 August.
- [5] Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2003). Designing and Developing an Effective Recommendation System for eCommerce Stores. *Proceedings of the WEBTEC Euromedia*, Plymouth, UK, 14-16 April, 50-54.
- [6] Markellos, K., Markellou, P., Sirmakessis, S., & Tsakalidis, A. (2001). Towards a Usable Business-to-Consumer E-store: Development of an Integrated Model for E-commerce Transactions. *Proceedings of the ETEC Euromedia*, Valencia, Spain, 18-20 April, 297-299.
- [7] Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2002). Who is Today's E-Customer? A Description of his Behavioral Model. *Proceedings of the eBusiness & eWork Conference and Exhibition*, Prague, Czech Republic, 16-18 October.
- [8] Markellou, P., Panayiotaki, A., Sirmakessis, S., Spinakis, A., & Tsakalidis, A. (2001). Using Hierarchy to Extract Innovation; The Use of Patents in Clarifying Innovation. *Proceedings of the ETK-NTTS Exchange of Technology and Know-How, New Techniques-Technologies for Statistics*, Crete, Greece, 18-22 June, 881-882. Retrieved from: http://webfarm.jrc.cec.eu.int/etk-ntts/poster_session_4.htm.
- [9] Markellos, K., Markellou, P., Mayritsakis, G., Panagopoulou, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2002). STING: Evaluation of Scientific and Technological Innovation and Progress in Europe through Patents. *Proceedings of the New Frontiers of Statistical Data Mining and Knowledge Discovery and E-Business Conference*, Knoxville, Tennessee, USA, 22-25 June.
- [10] Markellos, K., Markellou, P., & Tsakalidis, A. (2003). Mining Learners Profiles to Improve eLearning Experience. *Proceedings of the 3rd International Conference on New Horizons in Industry and Education*, Santorini, Greece, 28-29 August, 223-228.

- [11] Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004). Adaptive Hypermedia in Learning on the Web. *Proceedings of the iCEER, International Conference on Engineering Education and Research, Progress Through Partnership*, Olomouc and Bouzov Castle, Czech Republic, 27-30 June.
- [12] Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004). Web Personalization for Enhancing e-Learning Experience. *Proceedings of the 5th International Conference on Information Communication Technologies in Education, ICICTE*, Samos, Greece, 1-3 July.
- [13] Markellou, P., Rigou, M., & Sirmakessis, S. (2005). Knowledge Mining: A Quantitative Synthesis of Research Results and Findings. *Knowledge Mining, Springer-Verlag, Series: Studies in Fuzziness and Soft Computing*, Sirmakessis, S. (Ed.), in press.
- [14] Kondilis, T., Makris, C., Markellos, K., Markellou, P., & Tsakalidis, A. (2000). Helcom Project. *Proceedings of the 2nd SE European e-Commerce Conference*, Sofia, Bulgaria, 24-26 October, 121-126.
- [15] Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2001). HELCOM Project: an Integrated eCommerce System for SMEs in Greece. *Proceedings of the eBusiness and eWork Conference and Exhibition*, Venice, Italy, 17-19 October.

Εθνικά Επιστημονικά Συνέδρια

- [1] Markellos, K., Markellou, P., Rigou, M., & Sirmakessis, S. (2001). Modeling the Behaviour of e-Customers. *Proceedings of the PCHCI Panhellenic Conference with International Participation in Human-Computer Interaction*, Patras, Greece, 7-9 December, 333-338.
- [2] Markellos, K., Markellou, P., Rigou, M., & Sirmakessis, S., & Tsakalidis, A. (2001). Developing Variations of an Integrated eShop Model for SMEs. *Proceedings of the 8th Panhellenic Conference in Informatics with International Participation*, Nicosia, Cyprus, 8-10 November, 338-347.

Επιστημονικά Συμπόσια (Workshops)

- [1] Markellou, P. (2001). Designing a Usable Shopping Cart. *Proceedings of the Workshop W2 Transforming Web Surfers to E-Shoppers, PCHCI 1st Panhellenic Conference with International Participation in Human-Computer Interaction*, Patras, Greece, 7-9 December, 57-61. Retrieved from: <http://www.hci.gr/papers.htm>.

- [2] Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2003). A Multivariate Analysis and Visualization System for Analyzing Patent Data and Producing Technological and Scientific Indicators. *Abstract Proceedings of the International Workshop on Computational Management Science, Economics, Finance and Engineering*, Limassol, Cyprus, 28-30 March, p. 36.
- [3] Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2003). Mining Patent Databases for Monitoring Technological Trends. *Proceedings of the PANDA Workshop on Pattern-base Management Systems (PBMS)*, Como, Italy, 10th April, 83-92.
- [4] Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2001). STING: Evaluation of Scientific & Technological Innovation and Progress in Europe through Patents. *Proceedings of the 1st STING User-Focus Group Meeting*, Lausanne, Switzerland, 30 November, 23-29.
- [5] Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2001). STING: Functionalities of the System Prototype. *Proceedings of the 1st STING User-Focus Group Meeting*, Lausanne, Switzerland, 30 November, 51-58.

Τεχνικές Αναφορές

- [1] *TR.2000.02.02*, Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2002). An Internet Based Adaptive Educational System for the Greek Sign Language.
- [2] *TR.2000.04.02*, Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2002). Electronic Shops' Engineering Methodology.
- [3] *TR.2000.11.03*, Μαρκέλλου, Π., Ρήγκου, Μ., & Συρμακέσης, Σ. (2003). Μελέτη Αξιολόγησης Ευχρηστίας Ελληνικών Ηλεκτρονικών Βιβλιοπωλείων.

«All men by nature desire knowledge.»

Aristotle

Greek Metaphysics

(384 BC - 322 BC)

Κεφάλαιο
1**ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ E-LEARNING**

«The Semantic Web comes with new emerging standards based on evolving Web technologies that allow the reuse of material in different contexts, flexible solutions, as well as robust and scalable handling.» (Cristea, 2004).

1.1 Εισαγωγή

Η ευρεία χρήση και η ταχύτατη εξάπλωση του **Παγκοσμίου Ιστού Πληροφοριών (World Wide Web ή Web)** είχε σαν αποτέλεσμα τη μετατροπή του σε μια τεράστια αποθήκη πληροφοριών με δισεκατομμύρια σελίδες και περισσότερους από 800 εκατομμύρια χρήστες παγκοσμίως (Internet World Stats, 2005). Σήμερα θεωρείται ένα από τα πιο σημαντικά μέσα συλλογής, διαμοίρασης και διάδοσης πληροφοριών και υπηρεσιών. Ταυτόχρονα όμως ο όγκος των διαθέσιμων πληροφοριών προκαλεί αρκετά προβλήματα που σχετίζονται με τη συνεχώς αυξανόμενη δυσκολία αναζήτησης, εύρεσης, οργάνωσης, πρόσβασης και συντήρησης της αιτούμενης πληροφορίας από τους χρήστες. Αυτό το γεγονός έχει σημαντικές επιπτώσεις στον τρόπο που σχεδιάζουμε και υλοποιούμε εφαρμογές στο Web και τα συστήματα **e-learning (ηλεκτρονική μάθηση)** δεν θα μπορούσαν να αποτελέσουν εξαίρεση. Άλλωστε μεταξύ όλων των «e», το e-learning είναι η εφαρμογή εκείνη με τη γρηγορότερη ανάπτυξη και αποδοχή παγκοσμίως (Ευρωπαϊκή Επιτροπή, 2003).

Ο όρος e-learning αφορά σε όλες τις μορφές εκπαίδευσης που βασίζονται στο Web και χρησιμοποιεί υπολογιστές και δίκτυα υπολογιστών για τη δημιουργία, την παράδοση, τη διαχείριση και την υποστήριξη online σειρών μαθημάτων (Stockley, 2005). Τελευταία, χάρη στην προαναφερθείσα έκρηξη του Web αλλά και στην ευρεία χρήση των νέων **Τεχνολογιών Πληροφορίας και Επικοινωνιών (ΤΠΕ)**, η έρευνα στο e-learning έχει κερδίσει όλο και περισσότερη προσοχή από τις επιστημονικές και όχι μόνο κοινότητες (Urđan & Weggen, 2000). Διάφοροι εκπαιδευτικοί και εμπορικοί οργανισμοί καταδεικνύουν ένα συνεχές ενδιαφέρον για την περιοχή, κίνηση που αποτελεί μια πολύ ισχυρή κατευθυντήρια δύναμη, εκφρασμένη μέσα από πολυάριθμη έρευνα και εμπορικές προσπάθειες (Wentling et al., 2000), (Fry, 2001). Η μεγάλη ποικιλία των διαθέσιμων συστημάτων και εφαρμογών e-learning αποτελεί άλλωστε μια ισχυρή ένδειξη της ωριμότητας στην περιοχή. Τα πιο γνωστά εκπαιδευτικά περιβάλλοντα είναι: το **webCT** (<http://www.webct.com>), το **Lotus Learning Space** (<http://www.lotus.com>), το **Librarian** (<http://www.click2learn.com>), το **Blackboard** (<http://www.blackboard.net>), το **TopClass** (<http://www.wbtsystems.com>), το **Intralearn**

(<http://www.intralearn.com>), το **Ecollege** (<http://www.ecollege.com>), το **Embanet** (<http://www.embanet.com>), το **eduprise** (<http://www.eduprise.com>), κλπ.

Η πραγματική δύναμη του e-learning δεν είναι το γεγονός ότι επιτρέπει στον καθένα, οπουδήποτε και οποτεδήποτε να αποκτήσει πρόσβαση στη γνώση αλλά η ικανότητά του να παρέχει την **κατάλληλη πληροφορία, στους κατάλληλους ανθρώπους, την κατάλληλη χρονική στιγμή**. Στο μέλλον μάλιστα προβλέπεται ότι ο εκπαιδευόμενος θα μπορεί να ελέγχει και να προσαρμόζει σε πολύ μεγάλο βαθμό την προσωπική του εκπαιδευτική εμπειρία βελτιώνοντας παράλληλα και την απόδοσή του (Brusilovsky et al., 1998). Εντούτοις, στην πλειοψηφία των e-learning συστημάτων του παρελθόντος, οι σειρές των μαθημάτων και το εκπαιδευτικό υλικό δεν ήταν αρκετά δυναμικά, παρέχοντας ένα μάλλον περιορισμένο σύνολο χαρακτηριστικών, με πολύπλοκη δομή και δυσκολίες στην πλοήγηση με αποτέλεσμα να μη μπορούν να ανταποκριθούν αποδοτικά στις ανάγκες και στις απαιτήσεις των χρηστών, οδηγώντας σε «φτωχές» εκπαιδευτικές εμπειρίες. Γενικά, το εκπαιδευτικό υλικό με ενσωματωμένους **υπερσυνδέσμους (hyperlinks)** επιτρέπει στους εκπαιδευόμενους να ακολουθήσουν οποιοδήποτε μονοπάτι πλοήγησης επιθυμούν και να μη χρησιμοποιήσουν απαραίτητως τη δομή που έχουν καθορίσει οι σχεδιαστές των δικτυακών τόπων ή οι δημιουργοί του περιεχομένου (οι οποίοι έχουν ένα ορισμένο σχέδιο πλοήγησης στο μυαλό τους). Αυτή η ελευθερία μπορεί να αποδειχτεί αρνητικός παράγοντας, δεδομένου ότι σε πολλές περιπτώσεις οι εκπαιδευόμενοι δεν έχουν την απαραίτητη ωριμότητα και ικανότητα να ακολουθήσουν την κατάλληλη πορεία και συχνά οδηγούνται σε θεματικά που είναι είτε πάρα πολύ δύσκολα, είτε πάρα πολύ εύκολα, ή ακόμα και άσχετα με τις προσωπικές τους εκπαιδευτικές ανάγκες (Markellos et al., 2004a).

Μια λύση στο πρόβλημα αυτό αποτελεί η ενσωμάτωση **προσαρμοστικών χαρακτηριστικών (adaptive features)** που επιτρέπουν την παροχή **εξατομικευμένης μάθησης (personalized learning)** (Brusilovsky, 1996), (Brusilovsky, 1998). Αυτά τα συστήματα είναι η απάντηση στην παραδοσιακή **«one-size-fits-all»** προσέγγιση που παρέχει το ίδιο στατικό υλικό εκπαίδευσης σε όλους τους χρήστες, χωρίς να λαμβάνει υπόψη τις ανάγκες και τις προτιμήσεις τους, τις εμπειρίες και τις γνώσεις τους, οι οποίες μπορεί να ποικίλουν εντυπωσιακά (De Bra et al., 1999). Τα **προσαρμοστικά εκπαιδευτικά συστήματα που βασίζονται στο Web (adaptive web-based educational systems)** επιτρέπουν στις σειρές των μαθημάτων να δημιουργούνται, ικανοποιώντας τις ανάγκες των εκπαιδευόμενων στα διάφορα γνωστικά αντικείμενα χωρίς να σπαταλούν το χρόνο τους σε περιοχές που ήδη κατέχουν ή που δεν τους ενδιαφέρουν (Brusilovsky, 2003). Για να επιτευχθεί αυτός ο στόχος είναι απαραίτητη η γνώση του **προφίλ (profile)** του κάθε χρήστη που περιλαμβάνει πληροφορίες για την ταυτότητά του, τις προτιμήσεις του, τα ενδιαφέροντά του, τις δεξιότητές του, το εκπαιδευτικό υπόβαθρό του, κλπ. Συνήθως για τη συλλογή αυτών των πληροφοριών απαιτείται η συμπλήρωση εξαντλητικών ερωτηματολογίων ή φορμών καθώς και η καταγραφή της πλοηγητικής συμπεριφοράς των χρηστών κατά τη χρήση του e-learning συστήματος.

Με αυτό τον τρόπο οι εφαρμογές e-learning παρέχουν υψηλής ποιότητας περιεχόμενο, αποδοτική δομή, εύκολη πλοήγηση καθώς επίσης και πλήρη υποστήριξη για τις διαφορετικές εργασίες όλων των προφίλ χρηστών που συμμετέχουν σε ένα τυπικό σενάριο εκπαίδευσης από απόσταση (Mödritscher et al., 2004). Συγκεκριμένα, ανάλογα με το γνωστικό υπόβαθρο του εκπαιδευόμενου, τις γνώσεις και τις αδυναμίες του, το εκπαιδευτικό στυλ που προτιμά, την πρόοδο που έχει σημειώσει μέχρι τώρα, το σύστημα αποφασίζει τι εκπαιδευτικό υλικό θα

του παρουσιάσει και με ποιον τρόπο (Brusilovsky, 2001). Πιθανές παράμετροι αποτελούν τα διαφορετικά εκπαιδευτικά μονοπάτια, οι διαφορετικοί τρόποι παρουσίασης του ίδιου περιεχομένου (π.χ. με ή χωρίς ήχο) ή η προσφορά ενός διαφορετικού συνόλου λειτουργιών που το **περιβάλλον διεπαφής (user interface)** του συστήματος εκπαίδευσης μπορεί να παρέχει για να μειώσει την πολυπλοκότητα (Dietinger, 2003).

Τα τελευταία χρόνια, ένας σημαντικός αριθμός από ερευνητικά αλλά και εμπορικά e-learning περιβάλλοντα που ενσωματώνουν προσαρμοστικά χαρακτηριστικά έχουν σχεδιαστεί και αναπτυχθεί. Αυτά τα συστήματα συλλέγουν «γνώση» για κάθε χρήστη, την αναλύουν και έπειτα τη «μετατρέπουν» σε εξατομικευμένη εμπειρία μάθησης (Neumann & Zirvas, 1998), (Blom, 2000), (Zaoane, 2002). Παρέχουν δηλαδή **προσωποποιημένες (personalized)**, **αλληλεπιδραστικές (interactive)**, **ανά πάσα στιγμή (just-in-time)**, **άμεσες (current)** και **κεντροποιημένες προς το χρήστη (user-centric)** υπηρεσίες. Τα συστήματα αυτά περιλαμβάνουν όλες τις λειτουργικότητες μιας σειράς μαθημάτων και παράλληλα μπορούν και διαφοροποιούνται κατά τη διάρκεια χρήσης τους ώστε να προσαρμόζονται στις απαιτήσεις των χρηστών και να γίνονται περισσότερο αποδοτικά. Στα πιο αντιπροσωπευτικά ανήκουν το **InterBook** (Brusilovsky et al., 1998) το οποίο εστιάζει στην παροχή προσαρμοστικής πλοήγησης σε ένα e-learning σύστημα μέσω τεχνικών σήμανσης υπερσυνδέσμων, ενώ το **AHA! (Adaptive Hypermedia Architecture)** χρησιμοποιεί την τεχνική απόκρυψης υπερσυνδέσμων (De Bra & Calvi, 1998). Το **NetCoach** (Weber & Kuhl, 2001) προέκυψε από το **ELM-ART** (Weber & Specht, 1997), ένα από τα πρώτα web-based προσαρμοστικά e-learning συστήματα και επιτρέπει στους συγγραφείς να δημιουργούν τα εξατομικευμένα μαθήματα χωρίς να απαιτεί ιδιαίτερες προγραμματιστικές γνώσεις. Το **WebPersonalizer** (Mobasher et al., 2000) είναι ένα σύστημα γενικότερου σκοπού το οποίο παρέχει μια λίστα από προτεινόμενους υπερσυνδέσμους στο χρήστη καθώς αυτός πλοηγείται σε ένα web site. Τέλος, το **INSPIRE (INtelligent System for Personalized Instruction in a Remote Environment)** δημιουργεί μαθήματα με δυναμικό τρόπο τα οποία καλύπτουν σταδιακά το γνωστικό στόχο που έχει επιλέξει ο εκπαιδευόμενος ακολουθώντας το γνωστικό του επίπεδο, το στυλ μάθησής του αλλά και την εξέλιξή του (Papanikolaou et al., 2003).

Ένα από τα ανοικτά προβλήματα της περιοχής σχετίζεται με το πώς θα συνδυαστούν οι κατάλληλες πηγές στοιχείων, θα βρεθούν οι σωστές και χρήσιμες πληροφορίες κάθε φορά και θα γίνει κατανοητή η συμπεριφορά του χρήστη, ώστε να υποστηριχθεί μια ικανοποιητική και εξατομικευμένη ηλεκτρονική μάθηση. Αυτό είχε σαν αποτέλεσμα πολλοί ερευνητές να στραφούν στην επιστημονική περιοχή του **Data Knowledge Mining (Εξόρυξη Γνώσης από Δεδομένα)** ή απλώς **Data Mining (Εξόρυξη Δεδομένων)** (Hand et al., 2001). Πράγματι η εξόρυξη γνώσης από μεγάλες αποθήκες δεδομένων έχει εξελιχθεί σε ένα από τα βασικότερα ερευνητικά ζητήματα. Μέχρι πρόσφατα η εξόρυξη γνώσης αφορούσε αποκλειστικά **δομημένα (structured)** δεδομένα δηλαδή δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων (databases). Τα τελευταία χρόνια το ενδιαφέρον στράφηκε και σε **μη δομημένα δεδομένα (semi-structured και unstructured)** π.χ. κείμενα, εικόνες, ιστοσελίδες (web pages), κλπ. Αυτό οδήγησε στη δημιουργία νέων κλάδων όπως το **Text Mining (Εξόρυξη Γνώσης από Κείμενα)** και το **Web Mining (Εξόρυξη Γνώσης από Δεδομένα του Web)**. Ιδιαίτερα ο τελευταίος κλάδος αποτελεί ένα ελπιδοφόρο πεδίο ερευνητικής δραστηριότητας που θα μπορούσε να εκπληρώσει τους παραπάνω στόχους.

Η πρόσφατη εμφάνιση του **Semantic Web (Σημασιολογικός Παγκόσμιος Ιστός Πληροφοριών)** στοχεύει στην ενσωμάτωση ενός ακόμα επιπέδου **ευφυΐας (intelligence)** στο περιβάλλον των διαδικτυακών εφαρμογών. Σύμφωνα με τους Berners-Lee et al. (2001):

«the Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation».

Ένας πιο επίσημος ορισμός από το W3C (2005) αναφέρει ότι:

«the Semantic Web is the representation of data on the World Wide Web. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using eXtensible Markup Language (XML) for syntax and Uniform Resource Identifiers (URIs) for naming».

Η ικανότητα του Semantic Web να προσθέτει νόημα στις πληροφορίες, αποθηκευμένο με τέτοιο τρόπο που να μπορεί να αναζητηθεί και να υποβληθεί σε επεξεργασία, παρέχει πολλές ευκαιρίες επέκτασης για τις web-based εφαρμογές (Anderson & Whitelock, 2004). Ο συνδυασμός του με το Web Mining έχει δημιουργήσει μια νέα και γρήγορα-αναδυόμενη ερευνητική περιοχή που είναι γνωστή ως **Semantic Web Mining**. Η ιδέα πίσω από τη χρησιμοποίησή του στη δημιουργία **εξατομικευμένων (personalized)** web εμπειριών είναι να βελτιωθεί το Web Mining με την εκμετάλλευση των νέων σημασιολογικών δομών (Markellou et al., 2004b), (Mobasher et al., 2003), (Dai & Mobasher, 2004).

1.1.1 Το πρόβλημα και η προτεινόμενη προσέγγιση

Το πρόβλημα που μελετάται στα πλαίσια της διδακτορικής διατριβής είναι η **εξόρυξη γνώσης από μια εφαρμογή e-learning για την παραγωγή εξατομίκευσης**. Συγκεκριμένα, στην προτεινόμενη προσέγγιση η δομή και η σχέση των δεδομένων και των πληροφοριών διαδραματίζουν ουσιαστικό ρόλο. Ο προτεινόμενος αλγόριθμος βασίζεται σε μια **οντολογία (ontology)** η οποία βοηθά στη δόμηση και στη διαχείριση του περιεχομένου που σχετίζεται με μια δεδομένη σειρά μαθημάτων, ένα μάθημα ή ένα θεματικό (Markellou et al., 2005a), (Markellou et al., 2006). Ειδικότερα, το πλαίσιο εξατομίκευσης λαμβάνει υπόψη τα προφίλ των χρηστών και την οντολογία του θεματικού, ενώ η διαδικασία χωρίζεται σε δύο στάδια: στις offline ενέργειες προετοιμασίας των δεδομένων, δημιουργίας της οντολογίας και **εξόρυξης από δεδομένα χρήσης (usage mining)** και στην online παροχή εξατομίκευσης.

Το σύστημα βρίσκεται σε πρώτη φάση ένα αρχικό σύνολο συστάσεων βασισμένο στην οντολογία του πεδίου και στη συνέχεια χρησιμοποιεί τα **frequent itemsets (συχνά εμφανιζόμενα σύνολα στοιχείων)** για να το εμπλουτίσει, λαμβάνοντας υπόψη την πλοήγηση άλλων **παρόμοιων χρηστών (similar users)**. Με τον τρόπο αυτό, μειώνουμε το χρόνο που απαιτείται για την ανάλυση όλων των frequent itemsets και των κανόνων συσχέτισης. Εστιάζουμε μόνο σε εκείνα τα σύνολα που προέρχονται από το συνδυασμό της **ενεργούς συνόδου (current session)** του χρήστη και των συστάσεων της οντολογίας. Η χρονική μείωση προκύπτει λόγω του γεγονότος ότι τα frequent itemsets φιλτράρονται με το σύνολο των συστάσεων της οντολογίας οδηγώντας σε ένα μικρότερο χώρο αναζήτησης. Επιπλέον, αυτή η προσέγγιση ανακουφίζει και το πρόβλημα των μεγάλων χρόνων απόκρισης, το οποίο μπορεί στη συνέχεια να οδηγήσει στην εγκατάλειψη του e-learning

συστήματος. Αν και η εξατομίκευση απαιτεί αρκετά βήματα επεξεργασίας και ανάλυσης, το εμπόδιο αυτό αποφεύγεται με την εκτέλεση σημαντικού μέρους της διαδικασίας offline.

1.1.2 Η δομή του κεφαλαίου

Στο κεφάλαιο αυτό διερευνούμε πώς οι τεχνολογίες Semantic Web Mining και ειδικότερα οι οντολογίες μπορούν να ενσωματωθούν σε εφαρμογές e-learning. Συγκεκριμένα, η δομή του κεφαλαίου είναι η ακόλουθη. Στις επόμενες ενότητες γίνεται μια σύντομη παρουσίαση του Semantic Web και του Web Mining. Περιγράφεται το πως έχουμε οδηγηθεί στο Semantic Web Mining και με ποιο τρόπο αυτό μπορεί να αποτελέσει το μέσο για περισσότερο εύχρηστες web-based εφαρμογές. Στη συνέχεια μελετώνται οι οντολογίες και καταγράφεται μια σύντομη ταξινόμησή τους. Ακολουθεί η περιγραφή της προσέγγισής μας για την υποστήριξη εξατομίκευσης σε ένα δεδομένο e-learning σύστημα. Το κεφάλαιο ολοκληρώνεται με την παράθεση των σημαντικότερων συμπερασμάτων καθώς και την καταγραφή ανοικτών θεμάτων για μελλοντική έρευνα.

1.2 To Semantic Web

Το **Semantic Web** αποτελεί σήμερα μια από τις πιο σημαντικές προσπάθειες για την αυτόματη ενοποίηση των συστημάτων και των εφαρμογών ώστε να συνεργάζονται διαλειτουργικά σε παγκόσμιο επίπεδο. Ο Tim Berners-Lee που δημιούργησε το **Web** το 1989 οραματιζόταν ένα ιστό δεδομένων αυτόματα επεξεργάσιμων από τις εφαρμογές βάσει του νοήματος και όχι της μορφής της πληροφορίας. Αυτό θα έχει σαν αποτέλεσμα από το απλό κείμενο που το καταλαβαίνει πλήρως μόνο ο άνθρωπος να κινούμαστε προς τις **ημιδομημένες (semi-structured)** και **δομημένες (structured)** πληροφορίες που μπορούν να γίνουν κατανοητές από τις διαδικτυακές εφαρμογές όπως για παράδειγμα συμβαίνει στα συστήματα **ηλεκτρονικού εμπορίου (e-commerce)**, **ηλεκτρονικού επιχειρείν (e-business)** και **ηλεκτρονικής μάθησης (e-learning)**.

Το Semantic Web αποτελεί πρωτοβουλία της **Κοινοπραξίας του Παγκοσμίου Ιστού** ή **World Wide Web Consortium** ή **W3C** (<http://www.w3.org>). Η ιδέα στην οποία βασίζεται αφορά στην οργάνωση και στη διασύνδεση της διαθέσιμης πληροφορίας στο Διαδίκτυο ώστε να χρησιμοποιηθεί πιο αποτελεσματικά στην ανακάλυψη, στην αυτοματοποίηση, στην ομαδοποίηση και στην επαχρησιμοποίησή της από διαφορετικές μεταξύ τους διαδικτυακές εφαρμογές. Το Semantic Web στοχεύει δηλαδή στη μετεξέλιξη του σημερινού Web ώστε να μετατραπεί σε ένα περιβάλλον όπου οι πληροφορίες που υπάρχουν και διακινούνται να είναι επεξεργάσιμες από τους ηλεκτρονικούς υπολογιστές. Κινείται προς την κατεύθυνση της επέκτασης και της βελτίωσης της δόμησης της πληροφορίας για να είναι προσπελάσιμη από προγράμματα υπολογιστών με τελικό στόχο την αυτοματοποίηση πολλών λειτουργιών στο Web. Σε αυτό το πλαίσιο, τα σημερινά κείμενα στις σελίδες του Web θα αντικατασταθούν με δομημένα κείμενα και δεδομένα σε μορφή XML και RDF. Από αυτή την άποψη το Semantic Web είναι ένα πρόβλημα **αναπαράστασης γνώσης (knowledge representation)** από και για τους υπολογιστές και απαιτεί ένα μηχανισμό που θα επιτρέπει την επεξεργασία αυτής της γνώσης. Αυτός ο μηχανισμός θα πρέπει να υποστηρίζει τη δυνατότητα λογικής επεξεργασίας

των πληροφοριών με σκοπό την εξαγωγή συμπερασμάτων, τη δημιουργία νέας γνώσης, την υποστήριξη στη λήψη αποφάσεων και την αυτόματη εκτέλεση ενεργειών.

Η υλοποίηση του παραπάνω οράματος είναι πολύ δύσκολη προς το παρόν. Προς αυτή την κατεύθυνση η χρήση των οντολογιών μπορεί να αποδειχθεί ιδιαίτερα κρίσιμη παράμετρος. Μια **οντολογία (ontology)** είναι μια αυστηρή (formal) περιγραφή των αντικειμένων-πόρων και των σχέσεων μεταξύ τους (Gomez-Perez et al., 2004). Οι οντολογίες δίνουν τη δυνατότητα σε μια κοινότητα χρηστών να έχουν κοινή ονοματολογία και κοινή αντίληψη για τα αντικείμενα-πόρους τα οποία δηλώνουν ή χρησιμοποιούν. Επίσης, τους επιτρέπουν να περιγράψουν τον «κόσμο» που χρησιμοποιούν (που τους αφορά) με ένα τέτοιο τρόπο, ώστε οι υπολογιστές να μπορούν να επεξεργαστούν και να «συνδέσουν» τα δεδομένα από διαφορετικούς κόσμους. Η δυσκολία στις οντολογίες έγκειται στο γεγονός ότι οι κοινότητες των χρηστών θα πρέπει να συμφωνήσουν στην οντολογική περιγραφή του πεδίου ενδιαφέροντός τους.

1.2.1 Η αρχιτεκτονική του Semantic Web

Η βασική ιδέα του Semantic Web είναι να συλλαμβάνει και να εκφράζει έννοιες και σημασίες. Για το σκοπό αυτό στηρίζεται σε μια αρχιτεκτονική διαφορετικών επιπέδων (Berners-Lee, 2000), τα σημαντικότερα από τα οποία απεικονίζονται στο Σχήμα 1:

- **Unicode και URI επίπεδα:** διασφαλίζουν ότι διεθνή σύνολα χαρακτήρων και μέσα χρησιμοποιούνται για την αναγνώριση των αντικειμένων του Semantic Web.
- **XML επίπεδο:** αναπαριστά τη δομή των δεδομένων. Το επίπεδο αυτό μαζί με τους ορισμούς namespace (NS) και xmlschemas διασφαλίζουν ότι είναι δυνατή η ενσωμάτωση σημασιολογικών ορισμών με τα άλλα XML πρότυπα.
- **RDF επίπεδο:** αναπαριστά τη σημασιολογία των δεδομένων. Με RDF και rdfschema είναι δυνατό να γίνουν οι δηλώσεις για τα αντικείμενα με URIs και να καθοριστούν τα λεξιλόγια που μπορούν να αναφερθούν από URIs. Αυτό είναι το επίπεδο όπου μπορούν να δοθούν οι τύποι στους πόρους και τις συνδέσεις.

Σχήμα 1. Τα επίπεδα της αρχιτεκτονικής του Semantic Web.

- **Ontology επίπεδο:** αναπαριστά την επίσημη κοινή «συμφωνία» για το νόημα των δεδομένων. Το επίπεδο αυτό υποστηρίζει την εξέλιξη των λεξιλογίων δεδομένου ότι μπορεί να καθορίσει τις σχέσεις μεταξύ των διαφορετικών εννοιών.
- **Digital signature επίπεδο:** το επίπεδο αυτό στοχεύει στην ανίχνευση των αλλαγών των εγγράφων και μαζί με το ontology επίπεδο αποτελούν τα επίπεδα που τυποποιούνται αυτή την περίοδο από τις ομάδες εργασίας του W3C.
- **Logic, proof και trust επίπεδα:** αποτελούν τα ανώτερα επίπεδα της αρχιτεκτονικής τα οποία ερευνώνται αυτή την περίοδο, ενώ παράλληλα υλοποιούνται και απλές εφαρμογές επίδειξής τους. Το επίπεδο λογικής επιτρέπει το γράψιμο των κανόνων, ενώ το επίπεδο απόδειξης εκτελεί τους κανόνες και αξιολογεί μαζί με το μηχανισμό του επιπέδου εμπιστοσύνης για τις εφαρμογές αν θα εμπιστευθεί τη δεδομένη απόδειξη ή όχι.

Αξίζει να σημειωθεί ότι η πραγματική δύναμη του Semantic Web θα υλοποιηθεί όταν οι άνθρωποι δημιουργήσουν συστήματα τα οποία θα συλλέγουν τα περιεχόμενα του Web από διαφορετικές πηγές, θα επεξεργάζονται τις πληροφορίες και θα ανταλλάσσουν τα αποτελέσματα με άλλους ανθρώπους ή μηχανές. Με αυτό τον τρόπο, η αποτελεσματικότητα του Semantic Web θα βελτιωθεί δραστικά αφού μεγαλύτερο ποσοστό περιεχομένου Web θα είναι «αναγνώσιμο» από τους υπολογιστές και περισσότερες αυτοματοποιημένες υπηρεσίες θα είναι διαθέσιμες. Δύο σημαντικές τεχνολογίες για την ανάπτυξη του Semantic Web είναι ήδη σε ισχύ: η **XML (eXtensible Markup Language)** και το **RDF (Resource Description Framework)**.

Η XML (<http://www.w3.org/XML>) επιτρέπει στον καθένα να δημιουργήσει ετικέτες (tags) που σχολιάζουν τις ιστοσελίδες ή τμήματα κειμένου σε μια ιστοσελίδα. Τα προγράμματα που διαχειρίζονται XML μπορούν να αξιοποιήσουν την πληροφορία αυτών των ετικετών με πολύπλοκους τρόπους, αρκεί να ξέρουν τι αυτές σημαίνουν (πως δηλαδή ο συγγραφέας της ιστοσελίδας χρησιμοποιεί κάθε ετικέτα). Ουσιαστικά, η XML επιτρέπει μια στοιχειώδη και αυθαίρετη δόμηση των δεδομένων, αλλά όχι την προσθήκη σημασιολογίας σε αυτά σε ικανοποιητικό βαθμό (Erdmann & Studer, 2000). Η έννοια των XML εγγράφων είναι «διαισθητικά» σαφής λόγω της σήμανσης και των ετικετών. Εντούτοις, οι υπολογιστές δεν διαθέτουν αυτή τη «διαίσθηση». Επίσης, τόσο τα **DTDs (Data Definition Types)** όσο και το **XMLSchema** δεν λύνουν το πρόβλημα. Τα πρώτα αποτελούν μια δυνατότητα να «κτιστεί» το περιεχόμενο των εγγράφων, όμως δεν είναι κατάλληλα για να περιγράψουν τη σημασιολογία ενός XML εγγράφου. Το ίδιο ισχύει και για το XMLSchema (<http://www.w3.org/XML/Schema>) αφού μπορεί μόνο να καθορίσει τη δομή μέσω μιας πλουσιότερης γλώσσας. Άρα, η XML από μόνη της δεν προσφέρει πλήρως τη δυνατότητα διάδοσης σημασιολογίας μέσω του Semantic Web αλλά μπορεί να παίξει το ρόλο του ενδιαμέσου σαν ένα σχήμα δεδομένων εύκολα επεξεργάσιμου από ηλεκτρονικούς υπολογιστές.

Το RDF (<http://www.xml.com/xml/pub/98/06/rdf.html>) παρέχει πραγματική σημασιολογία στα δεδομένα, τα οποία περιγράφει μέσω **δηλώσεων (statements)**. Οι δηλώσεις αυτές είναι κατανοησιμες από τον υπολογιστή και οι μηχανές αναζήτησης ή οι πράκτορες μπορούν να χρησιμοποιήσουν αυτή την πληροφορία. Το RDF είναι ανεξάρτητο υλοποίησης και μπορεί η σύνταξή του να καθοριστεί σε XML. Η διαδικασία με την οποία σημασιολογικές πληροφορίες προστίθενται σε Web έγγραφα είναι γνωστή σαν **σημασιολογικός σχολιασμός (semantic**

annotation) (Handschuh et al., 2001). Το RDF σε συνδυασμό με το **RDFS** ή **RDFSchema** (<http://www.w3.org/TR/PRrdf-schema>) προσφέρει διαμορφώσεις οι οποίες μπορούν να επεκταθούν ανάλογα με τις ανάγκες. Οι βασικές ιεραρχίες των κλάσεων και οι σχέσεις μεταξύ αυτών και των αντικειμένων μπορούν να εκφραστούν σε RDFS. Γενικά, θα μπορούσαμε να πούμε ότι το μοντέλο RDF(S) δεν είναι «διαισθητικό», γιατί η σημασιολογία δεν δίνεται με τυπικό τρόπο και συχνά είναι δύσκολη η διερμηνεία της με αποτέλεσμα να αποτελεί μια επιρρεπή σε λάθη διαδικασία.

Μια λύση σε αυτό το πρόβλημα παρέχεται από το τρίτο βασικό συστατικό του Semantic Web, δηλαδή τις **οντολογίες**. Η χρήση του όρου είναι δανεισμένη από τη Φιλοσοφία όπου ασχολείται με την **ύπαρξη** και τις θεωρίες για τη **φύση** και τα διάφορα **είδη** της. Στην **Τεχνητή Νοημοσύνη (Artificial Intelligence)** και γενικότερα στην έρευνα στο Web ο όρος χρησιμοποιείται για να ορίσει μια επίσημη και κοινή «σύλληψη» μιας συγκεκριμένης περιοχής ενδιαφέροντος (Gruber 1993). Είναι δε κατάλληλη για να περιγράψει ετερογενείς, κατανεμημένες και ήμι-δομημένες πηγές πληροφοριών που μπορούν να βρεθούν στο Web. Με τον καθορισμό των κοινών περιοχών, οι οντολογίες βοηθούν τόσο τους ανθρώπους όσο και τους υπολογιστές να επικοινωνήσουν μεταξύ τους, ενισχύοντας την ανταλλαγή της σημασιολογίας και όχι μόνο της σύνταξης. Οι οντολογίες τυπικά αποτελούνται από τους ορισμούς των εννοιών που είναι σχετικές με την περιοχή, τις σχέσεις τους και τα αξιώματα για αυτές τις έννοιες και τις σχέσεις. Διάφορες γλώσσες και συστήματα αναπαράστασης υπάρχουν π.χ. η **OIL (Ontology Interchange Language)** (<http://www.ontoknowledge.org/oil>) (Fensel et al., 2001), η **DAML+OIL (DARPA Agent Markup Language + Ontology Interchange Language)** (<http://www.daml.org/2001/03/reference.html>) (Horrocks, 2002), η **OWL (Web Ontology Language)** (<http://www.w3.org/TR/owl-ref>) (McGuinness & van Harmelen, 2004), κλπ. Ενώ μεταξύ των πιο γνωστών εργαλείων είναι: το **KAON** (<http://kaon.semanticweb.org>), το **OilEd** (<http://oiled.man.ac.uk>), το **Ontolingua** (<http://www-ksl.stanford.edu>), το **OntoSaurus** (<http://www.isi.edu/isd/ontosaurus.html>), το **Protégé2000** (<http://protege.stanford.edu>), το **OntoEdit** (<http://ontoserver.aifb.unikarlsruhe.de/ontoeedit>), το **WebOntoform** (<http://kmi.open.ac.uk/projects/webonto>), το **WebODE** (<http://webode.dia.fi.upm.es/webODE>), κλπ. Στις επόμενες ενότητες οι οντολογίες μελετώνται πιο συστηματικά, δίνονται επίσης ορισμοί και περιγράφονται οι προτεινόμενες ταξινομήσεις.

1.2.2 Ο ορισμός της οντολογίας

Οι οντολογίες αποτελούν τον πυρήνα για την υλοποίηση του Semantic Web καθώς είναι το μέσο με το οποίο τα σημασιολογικά δεδομένα γίνονται αναγνώσιμα από τον υπολογιστή (Davies et al., 2004), (Ding, 2001), (Ding et al., 2002). Κατά καιρούς έχουν δοθεί διάφοροι ορισμοί για το τι είναι οντολογία. Ένας επίσημος κοινά αποδεκτός ορισμός είναι ο ακόλουθος:

«An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what 'exists' is that which can be represented.» (Gruber, 1993).

Ο όρος **conceptualization** αφορά σε ένα αφηρημένο μοντέλο που αποδίδει κάποιο φαινόμενο του κόσμου καθορίζοντας τις σχετικές έννοιες αυτού του φαινομένου. Ο όρος **explicit** σημαίνει ότι οι τύποι των εννοιών που χρησιμοποιούνται και οι περιορισμοί στη χρήση τους είναι επακριβώς καθορισμένοι. Αυτός ο ορισμός μπορεί να επεκταθεί:

«An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest.»

Σε αυτή την περίπτωση ο όρος **formal** σημαίνει ότι η οντολογία είναι αναγνώσιμη από τον υπολογιστή, **shared** ότι συλλαμβάνει προφορική γνώση η οποία δεν είναι ιδιωτική αλλά κοινή, ενώ το **domain of interest** δηλώνει ότι κάποιος δεν ενδιαφέρεται να αναπαραστήσει όλο τον κόσμο αλλά κάποιο συγκεκριμένο τμήμα του που παρουσιάζει ενδιαφέρον. Παράλληλα, ο Gruber (1993) εισάγει σαν μέρος του «**Frame Ontology**» τα ακόλουθα χαρακτηριστικά που χρησιμοποιούνται για να μοντελοποιήσουμε μια οντολογία:

- **classes** ή **κλάσεις** (συχνά αναφέρονται και σαν **concepts**). Οι classes μπορούν να οργανωθούν σε ταξινομίες μέσω των οποίων μηχανισμοί κληρονομικότητας μπορούν να εφαρμοστούν. Παραδείγματα κλάσεων από την περιοχή του e-learning μπορεί να είναι ο Learner, το Course, το Lesson, το Topic, κλπ.
- **relations** ή **σχέσεις** (συχνά, κυρίως στην RDF κοινότητα αναφέρονται σαν **properties**). Τα relations αναπαριστούν ένα τύπο αλληλεπίδρασης μεταξύ των concepts του domain. Ένα παράδειγμα σχέσης μπορεί να είναι το SubclassOf που ορίζει ότι μια κλάση αποτελεί υποκλάση μιας άλλης κλάσης π.χ. το Lesson είναι υποκλάση του Course, το Topic είναι υποκλάση του Lesson, κλπ.
- **axioms** ή **αξιώματα**. Τα axioms χρησιμοποιούνται για να μοντελοποιήσουν προτάσεις οι οποίες είναι πάντα αληθείς. Επίσης, χρησιμοποιούνται για να αναπαραστήσουν γνώση που δεν μπορεί να καθοριστεί από τα άλλα στοιχεία. Παράδειγμα αξιώματος μπορεί να είναι ότι ένας Learner για να απαντήσει ένα ερώτημα πρέπει να έχει διαβάσει την εκφώνησή του.
- **instances**. Τέλος, τα instances αποτελούν στιγμιότυπα των concepts και των relations. Ένα παράδειγμα της κλάσης Learner είναι ο εκπαιδευόμενος «Πέννυ Μαρκέλλου» που παρακολουθεί το Topic «Πως το Web Usage Mining μπορεί να βοηθήσει στην παροχή εξατομικευμένων εκπαιδευτικών εμπειριών» που ανήκει στο Lesson «Web Mining».

1.2.3 Ταξινόμηση των οντολογιών

Οι οντολογίες μπορούν να ταξινομηθούν με διάφορους τρόπους και σε διάφορες κατηγορίες. Αρχικά, οι Mizoguchi et al., (1995) πρότειναν τις ακόλουθες τέσσερις κατηγορίες οντολογιών:

- **Οντολογίες περιεχομένου για επαναχρησιμοποίηση γνώσης (content ontologies for reusing knowledge)**. Οι οντολογίες αυτής της κατηγορίας είναι δυνατόν να διακριθούν περαιτέρω στις υποκατηγορίες: **οντολογίες στόχου (task ontologies)**, **οντολογίες περιοχών (domain ontologies)** και **γενικές ή κοινές οντολογίες (general ή common ontologies)**.
- **Οντολογίες επικοινωνίας (λέω & ρωτάω) για διαμοιρασμό της γνώσης (communication (tell & ask) ontologies for sharing knowledge)**.
- **Οντολογίες ευρετηρίου για την ανάκτηση περιπτώσεων (indexing ontologies for case reasoning)**.

- **Μετα-οντολογίες (meta-ontologies)**, οι οποίες για τους συγγραφείς είναι ισοδύναμες με τις **οντολογίες αναπαράστασης γνώσης (knowledge representation ontologies)**.

Οι Van Heijst et al., (1997) ταξινόμησαν τις οντολογίες σύμφωνα με: α) το βαθμό και τον τύπο δόμησης του μοντέλου που αποδίδει το φαινόμενο και β) του θέματός του. Με βάση το (α) οι οντολογίες διακρίνονται σε τρεις κατηγορίες: τις **ονοματολογικές ή ορολογικές οντολογίες (terminological ontologies)** π.χ. λεξικά, τις **οντολογίες πληροφοριών (information ontologies)** π.χ. σχήματα βάσεων δεδομένων και τις **οντολογίες μοντελοποίησης της γνώσης (knowledge modeling ontologies)** που αποσαφηνίζουν τις αντιλήψεις της γνώσης. Με βάση το (β) προσδιορίζονται οι ακόλουθες τέσσερις κατηγορίες: **οντολογίες αναπαράστασης (representation ontologies)**, **γενικές οντολογίες (generic ontologies)**, **οντολογίες περιοχών (domain ontologies)** και **οντολογίες εφαρμογών (applications ontologies)**.

Ο Guarino (1998) ταξινόμησε τους τύπους των οντολογιών σύμφωνα με το επίπεδο εξάρτησής τους με κάποιο συγκεκριμένο στόχο ή άποψη. Σύμφωνα με αυτή την ταξινόμηση (Σχήμα 2) έχουμε:

- **Υψηλού επιπέδου οντολογίες (top-level ontologies)** που περιγράφουν πολύ γενικές έννοιες (όπως ο χώρος, ο χρόνος, το γεγονός) και οι οποίες είναι ανεξάρτητες από ένα συγκεκριμένο πρόβλημα ή περιοχή γνώσης. Είναι λογική η ύπαρξη ενοποιημένης οντολογίας υψηλού επιπέδου για μεγάλες κοινότητες χρηστών.
- **Οντολογίες περιοχών (domain ontologies)** που περιγράφουν ένα λεξιλόγιο σχετικό με ένα γενικό πεδίο γνώσης εξειδικεύοντας τις έννοιες που εισήχθησαν στην οντολογία υψηλού επιπέδου.
- **Οντολογίες εργασιών (tasks ontologies)** που περιγράφουν ένα λεξιλόγιο σχετικό με μια γενική δραστηριότητα ή εργασία εξειδικεύοντας την οντολογία υψηλού επιπέδου.
- **Οντολογίες εφαρμογών (application ontologies)** οι οποίες είναι οι πιο συγκεκριμένες οντολογίες. Οι έννοιες σε οντολογίες εφαρμογών συχνά ανταποκρίνονται σε ρόλους οντοτήτων της περιοχής γνώσης καθώς εκτελούν συγκεκριμένη δραστηριότητα.

Σχήμα 2. Ταξινόμηση οντολογιών σύμφωνα με τον Guarino (1998).

Τέλος, οι Lassila & McGuinness (2001) ταξινομήσαν τις οντολογίες σύμφωνα με τις πληροφορίες που αυτές χρειάζονται για να εκφράσουν τον πλούτο της εσωτερικής δομής τους. Συγκεκριμένα, όρισαν τις ακόλουθες κατηγορίες: **ελεγχόμενα λεξιλόγια (controlled vocabularies)**, **γλωσσάρια (glossaries)**, **θησαυρούς (thesauri)**, **ανεπίσημες is-a ιεραρχίες (informal is-a hierarchies)**, **επίσημες is-a ιεραρχίες (formal is-a hierarchies)**, **επίσημα στιγμιότυπα (formal instances)**, **πλαίσια (frames)**, **περιορισμένων τιμών (value restriction)** και **γενικών λογικών περιορισμών (general logical constraints)**.

1.3 Από το Web Mining στο Semantic Web Mining

Ο όρος **Web Mining** πρωτοεμφανίστηκε στην εργασία (Etzioni, 1996) και είναι η εφαρμογή τεχνικών **Data Mining** στο περιεχόμενο, τη δομή και τη χρήση δεδομένων Web με στόχο την ανακάλυψη «κρυμμένης» γνώσης (**patterns ή προτύπων**). Το Web Mining μπορεί να ωφεληθεί από τη συγκεκριμένη δομή των δεδομένων (π.χ. πίνακες των βάσεων δεδομένων) αλλά μπορεί επίσης να εφαρμοστεί και σε ημι-δομημένα ή αδόμητα δεδομένα π.χ. κείμενα e-mails. Μια ευρέως αποδεκτή ταξινόμηση του Web Mining το διακρίνει στις τρεις κατηγορίες που φαίνονται στο Σχήμα 3: το **Web Content Mining**, το **Web Structure Mining** και το **Web Usage Mining** (Kosala & Blockeel, 2000), (Srivastava et al., 2000), (Mandria et al., 1999), (Eirinaki & Vazirgiannis, 2003).

Σχήμα 3. Οι κατηγορίες του Web Mining.

1.3.1 Web Content Mining

Το Web Content Mining αφορά στην εξόρυξη γνώσης από δεδομένα περιεχομένου Web. Στην ουσία αποτελεί μια μορφή Text Mining (Chakrabarti, 2000), (Sebastiani, 2002). Εστιάζει στο κείμενο των ιστοσελίδων που περιέχεται στον HTML, script, κλπ. κώδικα και μπορεί να εκμεταλλευτεί την ημι-δομημένη φύση αυτών των τύπων κειμένων. Τα HTML tags καθώς και τα XML markup των web pages ενσωματώνουν πληροφορίες που αφορούν όχι μόνο στη διαμόρφωση (layout) αλλά και στη λογική δομή. Το Web Content Mining μπορεί να

χρησιμοποιηθεί για να ανιχνεύσει εμφανίσεις όρων στα κείμενα. Για παράδειγμα, μπορεί να βρεθεί ότι η λέξη «champion» συναντιέται πολύ συχνά με τη λέξη «Olympiacos» στην Ελλάδα και με τη λέξη «Real» στην Ισπανία. Επίσης, μπορούν να ανακαλυφθούν τάσεις στη διάρκεια του χρόνου υποδεικνύοντας μια άνοδο ή μια πτώση ενδιαφέροντος για ορισμένα θέματα π.χ. η γλώσσα προγραμματισμού Java. Ένας άλλος τομέας εφαρμογής αποτελεί η ανίχνευση γεγονότων δηλαδή ο προσδιορισμός ιστοριών στα συνεχιζόμενα δελτία ειδήσεων που αντιστοιχούν σε νέα ή προηγούμενα μη αναγνωρισμένα γεγονότα (Chang et al., 2001).

1.3.2 Web Structure Mining

Το Web Structure Mining αφορά στην εξόρυξη γνώσης από δεδομένα δόμησης Web. Η αρχική πηγή των δεδομένων από την οποία εξάγεται η γνώση είναι ένα σύνολο σελίδων και μπορεί να αφορά ένα απλό web site ή και ολόκληρο το Web. Στην ουσία εκμεταλλεύεται τις πρόσθετες πληροφορίες που περιλαμβάνονται (συχνά όχι με ρητό αλλά με υπονοούμενο τρόπο) στη δομή του υπερκειμένου (hypertext). Ένας σημαντικός τομέας εφαρμογής είναι ο προσδιορισμός της σχετικότητας των διαφορετικών σελίδων που εμφανίζονται εξίσου κατάλληλες όταν αναλύονται σε σχέση με το περιεχόμενό τους ή όταν αντιμετωπίζονται ξεχωριστά. Η δομή των συνδέσμων υπερκειμένου χρησιμοποιείται σε διάφορες προσεγγίσεις π.χ. ο Kleinberg (1999) αναλύει την τοπολογία των υπερσυνδέσμων και ορίζει τα hubs και τα authorities, ενώ η μηχανή αναζήτησης Google οφείλει την επιτυχία της στον αλγόριθμο PageRank (Page et al., 1998) που καθορίζει τη σχετικότητα μιας σελίδας με βάση άλλες σελίδες που «αναφέρεται» (η σχετικότητα μιας σελίδας αυξάνεται με τον αριθμό των συνδέσμων προς αυτή από άλλες σελίδες και ειδικότερα από άλλες σχετικές σελίδες).

Απλές σελίδες μπορούν επίσης να αναλυθούν όσον αφορά στη δομή τους, η οποία δίνει πληροφορίες για τη λειτουργία τους. Οι Cooley et al., (1999) διακρίνουν με βάση την εργασία (Rao et al., 1996) πέντε τύπους web pages α) «**head pages**» αποτελούν τα σημεία εισόδου στο site, β) «**navigation pages**» περιέχουν συνδέσμους και λίγη πληροφορία, γ) «**content pages**» περιέχουν μικρό αριθμό συνδέσεων και κυρίως οι χρήστες τις επισκέπτονται για το περιεχόμενό τους, δ) «**look-up pages**» έχουν αρκετές συνδέσεις εισόδου, λίγες εξερχόμενες συνδέσεις και κανένα σημαντικό περιεχόμενο, όπως οι σελίδες που χρησιμοποιούνται για να παρέχουν έναν ορισμό ή κάποιο ακρωνύμιο και (ε) «**personal pages**» έχουν πολύ διαφορετικά χαρακτηριστικά και όχι σημαντική κυκλοφορία.

Από μια άλλη σκοπιά, οι σύνδεσμοι αποτελούν μέρος του περιεχομένου μιας ιστοσελίδας. Αυτό ισχύει ιδιαίτερα για τις σελίδες στις οποίες οι σύνδεσμοι όπως και άλλα στοιχεία τους, είναι σημασιολογικά χαρακτηρισμένα. Για αυτό το λόγο σε πολλές περιπτώσεις, οι κατηγορίες Web Content Mining και Web Structure Mining αντιμετωπίζονται σαν μία. Συχνά χρησιμοποιούνται για να εξάγουν πληροφορίες από ιστοσελίδες, να καθορίσουν λέξεις κλειδιά περιγράφοντας το περιεχόμενο, να αναθέσουν ιστοσελίδες σε ένα μοντέλο, να ανιχνεύσουν τα γεγονότα ή τις διαδρομές σε πόρους του Web, κλπ.

1.3.3 Web Usage Mining

Το Web Usage Mining αφορά στην εξόρυξη γνώσης από δεδομένα χρήσης Web. Αυτό σημαίνει ότι κυρίως χρησιμοποιούνται τα αρχεία logs τα οποία καταγράφουν τα αιτήματα των

επισκεπτών ενός web site (Srivastava et al., 2000). Το περιεχόμενο και η δομή των web pages και ιδιαίτερα ενός web site απεικονίζουν τις προθέσεις των σχεδιαστών των σελίδων καθώς και την αρχιτεκτονική των πληροφοριών. Η πραγματική συμπεριφορά των χρηστών αυτών των σελίδων μπορεί να αποκαλύψει πρόσθετη γνώση αποκαλύπτοντας επιπλέον συσχετίσεις.

Αρχικά οι σχέσεις μπορούν να εξαχθούν από τα δεδομένα χρήσης χωρίς να έχει σχεδιαστεί κάποια ιδιαίτερη δομή ή οργάνωση. Για παράδειγμα, έστω ο ηλεκτρονικός κατάλογος ενός online καταστήματος ο οποίος δεν διαθέτει κάποια ιεραρχική οργάνωση με βάση την κατηγορία των προϊόντων, τον κατασκευαστή, κλπ. Η εφαρμογή εξόρυξης στα δεδομένα των επισκέψεων σε αυτό το site μπορεί να δείξει ότι οι περισσότεροι χρήστες π.χ. σε ποσοστό 80% που ενδιαφέρθηκαν για το προϊόν Α ενδιαφέρθηκαν επίσης για το προϊόν Β. Το «ενδιαφέρον» μπορεί να καθορίζεται από τα αιτήματα για τις σελίδες περιγραφής των προϊόντων, ή την τοποθέτηση του συγκεκριμένου προϊόντος στο καλάθι αγορών, κλπ. Αυτοί οι κανόνες αποτελούν βασικές στρατηγικές cross-selling και up-selling ενός site ηλεκτρονικού εμπορίου π.χ. όταν ένας νέος χρήστης ενδιαφέρεται για το προϊόν Α τότε θα λάβει μια σύσταση για το προϊόν Β (Mobasher et al., 2000), (Lin et al., 2002).

Κατά δεύτερον, σχέσεις μπορούν να εξαχθούν από δεδομένα χρήσης σε περιπτώσεις που διαφορετικές σχέσεις έχουν προδιαγραφεί. Για παράδειγμα, η εξόρυξη ακολουθιών μπορεί να δείξει ότι οι περισσότεροι από εκείνους τους χρήστες που επισκέφθηκαν τη σελίδα C πήγαν αργότερα στη σελίδα D (Kato et al., 2000), (Cooley, 2000). Αυτό μπορεί να ερμηνευθεί ότι οι επισκέπτες επιθυμούν να φθάσουν στη D από τη C, αλλά ότι αυτό δεν είχε προβλεφθεί στην αρχιτεκτονική και ως εκ τούτου δεν υπάρχει απευθείας σύνδεσμος από τη C στο D. Αυτή η γνώση μπορεί να χρησιμοποιηθεί για τη στατική βελτίωση του site για όλους τους χρήστες (προσθέτοντας μια σύνδεση από τη C στη D) ή για την παραγωγή δυναμικών εξατομικευμένων συστάσεων για το υποσύνολο των χρηστών που πηγαίνουν στη C π.χ. «μπορεί να σας ενδιαφέρει να δείτε επίσης τη D».

Είναι επίσης χρήσιμο να συνδυαστεί το Web Usage Mining με το Web Content Mining και το Web Structure Mining προκειμένου να «κατανοηθούν» τα πιο συχνά παρατηρημένα μονοπάτια και οι σελίδες που τα απαρτίζουν. Αυτό μπορεί να γίνει με διάφορες μεθόδους όπως: ταξινόμηση των σελίδων από την άποψη μιας προκαθορισμένης οντολογίας, εξαγωγή των λέξεων κλειδιών που περιέχονται σε αυτές τις σελίδες και εν συνεχεία ονομασία των ομάδων των λέξεων κλειδιών που αντιπροσωπεύουν τα συχνότερα μονοπάτια. Η οντολογία μπορεί να δημιουργηθεί με το χέρι ή να εξαχθεί (ήμι)αυτόματα, ενώ και η ταξινόμηση των σελίδων από την άποψη της οντολογίας μπορεί να γίνει με διάφορους τρόπους.

Στην πλειοψηφία των περιπτώσεων, οι εφαρμογές e-learning βασίζουν την εξατομίκευση στη τεχνική Web Usage Mining, η οποία αναλαμβάνει το ρόλο συλλογής και εξαγωγής όλων των δεδομένων που απαιτούνται για την κατασκευή και τη συντήρηση των προφίλ των εκπαιδευόμενων βασισμένη στη συμπεριφορά κάθε χρήστη όπως αυτή καταγράφεται στα server logs (Markellou et al., 2004a), (Pierrakos, et al., 2003). Από την άλλη πλευρά, ο συνδυασμός του Semantic Web και του Web Mining έχει δημιουργήσει μια νέα και γρήγορα αναδυόμενη ερευνητική περιοχή γνωστή ως **Semantic Web Mining**. Η ιδέα πίσω από τη χρησιμοποίηση του Semantic Web για τη δημιουργία προσωποποιημένων Web εμπειριών είναι να βελτιωθεί η διαδικασία του Web Mining με την εκμετάλλευση των νέων σημασιολογικών δομών (Markellou et al., 2004b). Συνδέοντας νοηματικά ένα σύνολο από

πόρους (κείμενα, ιστοσελίδες, εικόνες, ανθρώπους, έννοιες, κλπ.) είναι δυνατή η μετάβαση σε ένα πιο εκφραστικό και σημασιολογικά πλούσιο Web, όπου μπορούμε σταδιακά να ορίζουμε ένα σύνολο από νέες σχέσεις μεταξύ των δεδομένων που μέχρι τώρα μπορούσαμε μόνο έμμεσα να υπονοήσουμε (π.χ. HasLocation, WorksFor, IsAuthorOf, HasSubjectOf, DependsOn, κλπ.). Αυτό θα επιτρέψει την εφαρμογή εκλεπτυσμένων τεχνικών mining που απαιτούν δομημένη είσοδο για αποδοτική ολοκλήρωση των δεδομένων, καλύτερη διαχείρισή τους και θα οδηγήσουν σε πιο ευφυείς e-learning εφαρμογές που θα «αυτοπροσαρμόζονται» ώστε να ταιριάζουν καλύτερα σε κάθε ξεχωριστό χρήστη.

1.4 Semantic Web Mining και e-learning

Αν θεωρήσουμε ότι «*E-learning is just-in-time education integrated with high velocity value chains. It is the delivery of individualized, comprehensive, dynamic learning content in real time, aiding the development of communities of knowledge, linking learners and practitioners with experts*» (Drucker 2000) φαίνεται ότι το Semantic Web Mining ικανοποιεί τις βασικές απαιτήσεις του. Καταρχήν, το εκπαιδευτικό υλικό μπορεί να σχολιαστεί σημασιολογικά και να συνδεθεί με την οντολογία. Έτσι κάθε φορά που θα προκύπτει μια νέα ανάγκη μάθησης να είναι δυνατός και εύκολος ο συνδυασμός του υλικού αυτού σε μια νέα σειρά μαθημάτων. Θα δημιουργηθεί λοιπόν ένα εξατομικευμένο περιβάλλον για το χρήστη που θα σχετίζεται με τα ενδιαφέροντά του. Η δυνατότητα αυτή εξυπηρετεί την απαίτηση για διανομή πληροφορίας ατομικά στον εκπαιδευόμενο, με ύλη που εκείνος ορίζει και που αυτή ανταποκρίνεται στις εκάστοτε ανάγκες του.

Επιπλέον, οι διάφοροι web agents θα δίνουν τη δυνατότητα για επερωτήσεις που θα αφορούν σε ετερόκλητα τμήματα της πληροφορίας. Αυτό σημαίνει ότι το περιεχόμενο δεν παρέχεται με «γραμμικό» τρόπο σε συνέχειες, αλλά η πρόσβαση γίνεται σε διάφορα τμήματά του. Η δυνατότητα αυτή σε ένα περιβάλλον e-learning προσφέρει το πλεονέκτημα της γρήγορης και επικεντρωμένης στην ουσία μάθησης και δεδομένου ότι αυτή θα ανανεώνεται συνεχώς, ο εκπαιδευόμενος επιτυγχάνει συνεχή πρόοδο. Τα παραπάνω αναδεικνύουν την αλληλεπίδραση του χρήστη με το Web, ως βασική προϋπόθεση του e-learning, γεγονός που σημαίνει ότι ο τρόπος λειτουργίας δεν είναι κεντριοποιημένος αλλά κατανομημένος. Ο Πίνακας 1 παρουσιάζει τον τρόπο που το Semantic Web μπορεί να ικανοποιήσει τις βασικές απαιτήσεις του e-learning (Stojanovic et al., 2001).

Για να επιτευχθεί η **διαλειτουργικότητα**, η **προσβασιμότητα** και η **επαναχρησιμοποίηση** του εκπαιδευτικού περιεχομένου θα πρέπει να ακολουθηθούν κάποιοι κανόνες. Τα **πρότυπα e-learning μεταδεδομένων (metadata)** αποτελούν τυπικές περιγραφές των όρων που χρησιμοποιούνται για το σημασιολογικό σχολιασμό του εκπαιδευτικού υλικού. Ένα e-learning πρότυπο μπορεί να αναφέρεται στη **δομή ενός αντικειμένου μάθησης** (με τη χρήση μεταδεδομένων) ή στο **διαμοιρασμό του περιεχομένου μέσω της χρήσης μοντέλων δομής περιεχομένου (content structure models)**, ενώ υπάρχουν πρότυπα που αναφέρονται και στα δύο.

Απαιτήσεις	E-learning	Semantic Web Mining
Διανομή γνώσεων	Ο εκπαιδευόμενος αποφασίζει για την ύλη του.	Το υλικό συγκεντρώνεται κατόπιν επερωτήσεων.
Ανταπόκριση	Ανταποκρίνεται στις ανάγκες που εντοπίζει.	Το υλικό παρέχεται μέσω προφίλ χρήστη.
Πορεία	Μη γραμμική, επιτρέπει την απ' ευθείας πρόσβαση σε οποιοδήποτε κομμάτι γνώσης.	Πλοήγηση στις πληροφορίες του ενδιαφέροντος του χρήστη.
Πρόοδος	Συνεχής, η μάθηση δε σταματάει ποτέ.	Η βάση των πληροφοριών ανανεώνεται.
Αρχή	Κατανοητή, ο εκπαιδευτής και ο εκπαιδευόμενος συναποφασίζουν.	Αναγκαία η αλληλεπίδραση χρήστη-συστήματος.
Εξατομίκευση	Προσωποποιημένη, σχεδιασμένη για τις ανάγκες κάθε εκπαιδευόμενου.	Ο χρήστης αναζητεί πληροφορίες και η οντολογία συνδέει το πληροφοριακό υλικό με τις συγκεκριμένες αναζητήσεις του.
Προσαρμοστικότητα	Δυναμική, το περιεχόμενο της γνώσης αλλάζει σύμφωνα με την εμπειρία του εκπαιδευόμενου.	Όλα τα παραπάνω!

Πίνακας 1. E-learning και Semantic Web Mining.

Η διαρκώς αυξανόμενη ανάγκη για τον καθορισμό προτύπων οδήγησε πολλούς γνωστούς Οργανισμούς και Ιδρύματα να ασχοληθούν με το αντικείμενο και να καταθέσουν τις προτάσεις τους:

- **ADL/SCORM** – Advanced Distributed Learning/Sharable Content Object Reference Model (<http://www.adlnet.org>).
- **AICC** – Aviation Industry Computer Based Training Committee (<http://www.aicc.org>).
- **ARIADNE** – Alliance of Remote Instructional Authoring & Distribution Network for Europe (<http://www.ariadne-eu.org>).
- **DC** – Dublin Core (<http://dublincore.org>).
- **IEEE/LOM** – Institute of Electrical and Electronic Engineers/Learning Object Metadata (<http://ltsc.ieee.org/wg12>).
- **IMS** – Instructional Management System (<http://www.imspjproject.org>).

Τα πρότυπα αυτά δεν είναι ανεξάρτητα μεταξύ τους αλλά σχετίζονται στενά όπως φαίνεται και στο Σχήμα 4 (Stratakis et al., 2003). Η πλειονότητα των προτύπων που υπάρχουν αυτή τη στιγμή, χρησιμοποιούν ως γλώσσα κωδικοποίησης των περιγραφών (specifications) την XML (IEEE LOM, IMS, SCORM). Τον τελευταίο καιρό γίνεται μία προσπάθεια για να οριστούν RDF bindings για κάποια από τα πρότυπα αυτά (IEEE LOM, IMS). Αυτή η κίνηση οδηγείται από το γεγονός ότι η RDF παρουσιάζει σημαντικά πλεονεκτήματα σε σχέση με την XML. Καταρχήν, η χρήση της RDF ευνοεί τη διαλειτουργικότητα μεταξύ των προτύπων. Αυτό επιτυγχάνεται χάρη στο μοναδικό μοντέλο αποθήκευσης που χρησιμοποιείται για διαφορετικούς τύπους δεδομένων και σχημάτων. Επίσης, σύννητες πρόβλημα είναι ότι δεν υπάρχουν τυποποιημένοι τρόποι κωδικοποίησης και διανομής λεξιλογίων (vocabularies). Με την RDF λύνονται όσα προβλήματα σχετίζονται με τα λεξιλόγια, καθώς αυτά αποτελούν θεμελιώδες τμήμα της RDF Schema περιγραφής. Σημαντικό είναι και το ότι καθίσταται

απλούστερη η επαναχρησιμοποίηση των μεταδεδομένων. Ο ορισμός μεταμεταδεδομένων οποιουδήποτε βαθμού είναι εξίσου εύκολος με τον ορισμό απλών μεταδεδομένων.

Σχήμα 4. Συσχέτιση προτύπων e-learning μεταδεδομένων.

1.5 Η διαδικασία εξατομίκευσης σε ένα τυπικό e-learning σύστημα

Η ανάγκη για εξατομίκευση στο Web είναι αποτέλεσμα της τεράστιας αύξησης του αριθμού και της πολυπλοκότητας των πληροφοριών και των υπηρεσιών. Η δυνατότητα ενός web site και ειδικότερα ενός e-learning site να χειριστεί τους χρήστες του με διαφορετικό τρόπο και να τους καθοδηγεί επιτυχώς στην πλοήγησή τους αποτελεί κρίσιμο παράγοντα της επιτυχίας του και έχει μετατρέψει την εξατομίκευση σε ένα ιδιαίτερα χρήσιμο εργαλείο τόσο για τα ίδια τα sites όσο και για τους χρήστες τους. Σαν **web personalization** μπορούμε να θεωρήσουμε οποιαδήποτε ενέργεια που προσαρμόζει τις πληροφορίες ή τις υπηρεσίες που παρέχονται από ένα δικτυακό τόπο στις ανάγκες ενός συγκεκριμένου χρήστη ή ενός συνόλου χρηστών, χρησιμοποιώντας τη γνώση που πηγάζει από την ανάλυση της συμπεριφοράς πλοήγησης των χρηστών και από τα ενδιαφέροντά τους, σε συνδυασμό με τη δομή και το περιεχόμενο του δικτυακού τόπου. Στόχος είναι η παροχή στους χρήστες των πληροφοριών που θέλουν ή χρειάζονται χωρίς να απαιτεί από αυτούς να το ζητήσουν με ρητό τρόπο. Η διαδικασία της εξατομίκευσης περιλαμβάνει τη **μοντελοποίηση** των web αντικειμένων (π.χ. σελίδες, μαθήματα, προϊόντα, κλπ.) και των υποκειμένων (π.χ. χρήστες, εκπαιδευόμενοι, πελάτες, κλπ.), την **κατηγοριοποίηση** των αντικειμένων/υποκειμένων, το **ταίριασμα** μεταξύ των αντικειμένων ή/και των υποκειμένων και τον **προσδιορισμό** ενός συνόλου ενεργειών που θα προταθούν για εξατομίκευση.

Τα τελευταία χρόνια, το Web Usage Mining (Srivastava et al., 2000), (Mobasher et al., 2000) χρησιμοποιείται σαν η βασική τεχνική για την επίτευξη εξατομίκευσης στο πεδίο του e-learning (Σχήμα 5) (Μάρκελλος et al., 2004). Στόχος είναι:

- η **συλλογή πληροφοριών για το χρήστη** (στόχοι, γνωστικό επίπεδο, προτιμήσεις, ενδιαφέροντα) καθώς και άλλων πρόσθετων δεδομένων που αφορούν στο περιβάλλον χρήσης (υλικό, λογισμικό, σύνδεση, εκδόσεις), τα οποία θα χρησιμοποιηθούν για τη **δημιουργία και ενημέρωση του προφίλ του**,

- η **ανάλυση της συμπεριφοράς του χρήστη** με βάση τεχνικές όπως κανόνες συσχέτισης (association rules), ομαδοποίηση (clustering), κατηγοριοποίηση (classification), ανακάλυψη προτύπων (patterns discovery),
- η **εξαγωγή συμπερασμάτων**,
- και τέλος η **παραγωγή εξατομικευμένων διεπαφών χρήσης**.

Σχήμα 5. Η διαδικασία εξατομίκευσης σε ένα τυπικό e-learning σύστημα.

Παραδείγματα τέτοιων εξατομικευμένων αλληλεπιδράσεων των εκπαιδευόμενων με μια εφαρμογή ηλεκτρονικής μάθησης θα μπορούσαν να είναι τα ακόλουθα: ένας χρήστης επιθυμεί περισσότερο οπτικό υλικό από ότι κείμενο, ενώ κάποιος άλλος το αντίστροφο ή ένας χρήστης θέλει πρώτα να ολοκληρώσει μια ενότητα και μετά να προχωρήσει σε άλλες, ή προτείνεται σε ένα χρήστη να λύσει μια άσκηση μετά τη μελέτη μιας συγκεκριμένης ενότητας την οποία συνήθως λύνουν οι χρήστες που έχουν παρακολουθήσει την ίδια ενότητα (Brusilovsky, 2001), (Markellou et al., 2004a). Επίσης, θα μπορούσαν να αφορούν στο πλάνο μελέτης, σε προτάσεις για περαιτέρω μελέτη, σε σχετικούς υπερσυνδέσμους, στην εμφάνιση ή στην απόκρυψη υπερσυνδέσμων, κλπ. Για να επιτευχθούν τα ανωτέρω, τη βάση αποτελεί ένα καλό **μοντέλο χρηστών (user modeling)**, στο οποίο φυλάσσονται πάντα ενημερωμένες πληροφορίες για αυτούς δεδομένου ότι οι χρήστες δεν είναι στατικοί και αλλάζουν χαρακτηριστικά και συμπεριφορά με την πάροδο του χρόνου (Cooley et al., 1999), (Mobasher et al., 2002a), (Mobasher et al., 2002b).

1.5.1 Συλλογή των δεδομένων

Το πρώτο στάδιο στη διαδικασία εξόρυξης γνώσης από δεδομένα του Web είναι η **συλλογή των δεδομένων**, ή αλλιώς η καταγραφή και η καταχώρηση πληροφοριών που αφορούν στους χρήστες. Η συλλογή των δεδομένων αποτελεί τη βάση της διαδικασίας εξόρυξης γνώσης αφού τα τελικά αποτελέσματα εξαρτώνται τόσο από τη **διαθεσιμότητα** των

στοιχείων όσο και από την **ποιότητα** τους. Υπάρχουν πολλοί τρόποι για να συλλεχθούν στοιχεία για τους χρήστες μιας e-learning εφαρμογής. Τα στοιχεία αυτά μπορούν να παρασχεθούν είτε **ρητά** (ζητώντας τα από το χρήστη), είτε **έμμεσα** (καταγράφοντας τη συμπεριφορά και ακολουθώντας τα «ίχνη» του στο μονοπάτι πλοήγησης που επιλέγει). Συγκεκριμένα, συνήθως συγκεντρώνονται οι ακόλουθοι τύποι στοιχείων:

Δεδομένα που εισάγουν ρητά οι χρήστες

- **Πληροφορίες προφίλ.** Ο χρήστης εισάγει ρητά στο σύστημα δεδομένα που αφορούν στα δημογραφικά στοιχεία του, καθώς και πληροφορίες για το εκπαιδευτικό του υπόβαθρο, τις δεξιότητές του, τις ανάγκες και τις προτιμήσεις του αναφορικά με τη χρήση του συστήματος ή τα περιεχόμενα και τη δομή της εκπαιδευτικής διαδικασίας. Τα δεδομένα αυτής της κατηγορίας συγκεντρώνονται συνήθως χρησιμοποιώντας ένα ηλεκτρονικό ερωτηματολόγιο το οποίο εμφανίζεται σε κάποιο στάδιο της διαδικασίας εγγραφής του χρήστη στο σύστημα και αποτελεί την προσωπική του καρτέλα (ή εγγραφή/record). Είναι σκόπιμο, οι χρήστες να έχουν δυνατότητα πρόσβασης και επεξεργασίας των στοιχείων αυτών ώστε να είναι συνεχώς ενημερωμένο το σύστημα και να μπορεί να παίρνει τις κατάλληλες αποφάσεις κατά τη διαδικασία της εξατομίκευσης. Μέρος του προφίλ αποτελούν και τα δεδομένα που εισάγει ο χρήστης κατά τη διαδικασία πιστοποίησής του κάθε φορά που θέλει να χρησιμοποιήσει το σύστημα, δηλαδή το όνομα (login) και ο κωδικός πρόσβασής του (password). Με αυτό τον τρόπο αναγνωρίζεται μονοσήμαντα, καταγράφεται η συμπεριφορά του και το ιστορικό του στο σύστημα και ενημερώνεται το προφίλ του ώστε να εφαρμοστούν με μεγαλύτερη ακρίβεια οι τεχνικές εξατομίκευσης.
- **Δεδομένα αξιολόγησης του συστήματος.** Ορισμένα e-learning συστήματα που υποστηρίζουν εξατομίκευση περιλαμβάνουν στα δεδομένα που εισάγουν οι χρήστες και τις αξιολογήσεις τους για την εκπαιδευτική διαδικασία που εφαρμόζεται στο σύστημα ή για επιμέρους συστατικά της. Η μορφή τους μπορεί να είναι βαθμολόγηση σε κλίμακα Likert ή σχόλια ελεύθερου κειμένου σε forums και bulletin boards του συστήματος (στην τελευταία περίπτωση, οι πληροφορίες δεν είναι άμεσα χρησιμοποιήσιμες από το σύστημα και απαιτείται η παρέμβαση του ανθρώπινου παράγοντα ώστε να εξαχθούν ποιοτικά συμπεράσματα).

Δεδομένα που συγκεντρώνει το σύστημα για το χρήστη

- **Αρχεία logs.** Στα αρχεία αυτά, τα οποία κρατούνται στους εξυπηρετητές, καταγράφονται λεπτομέρειες που αφορούν σε κάθε αίτηση για εξυπηρέτηση όλων των ιστοσελίδων που βρίσκονται αποθηκευμένες στο συγκεκριμένο εξυπηρετητή. Συγκεκριμένα ένα αρχείο log περιέχει (Luotonen, 1995) τη διεύθυνση IP του χρήστη, τη χρονική στιγμή που παρέλαβε την αίτηση για μια συγκεκριμένη ιστοσελίδα, τη διάρκεια παραμονής του σε αυτή, το URL της ιστοσελίδας από όπου πραγματοποιήθηκε η αίτηση και ένα σύνολο από επιπλέον στοιχεία που ενδέχεται να είναι εξίσου χρήσιμα ανάλογα με το σχεδιασμό των αλγορίθμων εξατομίκευσης που θα εφαρμοστούν. Στο Σχήμα 6 παρουσιάζεται απόσπασμα ενός αρχείου log.
- **Cookies.** Τα αρχεία αυτά (η χρήση των οποίων συχνά συνδυάζεται με τα αρχεία log) είναι δυνατό να περιέχουν ποικίλες πληροφορίες, συμπεριλαμβανομένων των σελίδων που ο χρήστης έχει επισκεφθεί, των κωδικών πρόσβασης που έχει

εισαγάγει, κλπ. Ένα cookie αποθηκεύει τις πληροφορίες στον τοπικό δίσκο του χρήστη (κάτι που συνήθως συμβαίνει την πρώτη φορά που ζητάει εξυπηρέτηση ο χρήστης) και εκεί έχει πρόσβαση ο εξυπηρετητής σε κάθε μελλοντική αίτηση για εξυπηρέτηση που θα παραλάβει από το συγκεκριμένο χρήστη, επιτρέποντας με αυτό το «διαφανή» τρόπο την αυτόματη αναγνώριση του χρήστη από το σύστημα. Αξίζει όμως να σημειωθεί ότι είναι εύκολο οι χρήστες να απενεργοποιήσουν τα cookies από αντίστοιχη ρύθμιση του browser (φυλλομετρητή).

```
10:08:31 150.140.142.30 - GET /elearning/index.asp 200 16 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/styles/cuglobals.css 200 93 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/scripts/cuglobals.js 200 78 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/bg_training.gif 200 0 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/spacer.gif 404 0 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/osyl.gif 200 32 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/mathima.gif 200 15 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/spacer.gif 404 0 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/header_sbtc_2.gif 200 31 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/line.gif 200 0 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/epilogos.gif 200 15 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/cmos.jpg 200 15 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/ipt.gif 404 0 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/arrow_link_sm.gif 200 0 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:31 150.140.142.30 - GET /elearning/images/VertDot.gif 200 141 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:43 150.140.142.21 - GET /elearning/ 302 0 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
10:08:43 150.140.142.21 - GET /elearning/index.asp 200 16 Mozilla/4.0+(compatible;+MSIE+5.01;+windows+NT+5.0)
```

Σχήμα 6. Παράδειγμα log file.

Άλλα δεδομένα

Εκτός από τα δεδομένα που αφορούν στο χρήστη σε μια τυπική διαδικασία εξατομίκευσης χρησιμοποιούνται επίσης:

- **Το περιεχόμενο του site.** Τα δεδομένα αυτής της κατηγορίας αφορούν στο περιεχόμενο του site δηλαδή στο θεματικό του, στα μαθήματα από τα οποία αποτελείται, στις ασκήσεις και στα παραδείγματα, κλπ.
- **Η δομή του site.** Τα δεδομένα αυτής της κατηγορίας αφορούν στον τρόπο που είναι οργανωμένο το περιεχόμενο του site. Μπορεί να είναι είτε στοιχεία που χρησιμοποιούνται μέσα σε μια ιστοσελίδα (π.χ. επικέτες HTML ή XML), είτε στοιχεία που διασυνδέουν ένα site (π.χ. υπερσυνδέσμοι).

1.5.2 Καθαρισμός των δεδομένων και μοντελοποίηση του χρήστη

Μετά την ολοκλήρωση της συλλογής, ακολουθεί η φάση του **καθαρισμού των δεδομένων** κατά την οποία αφαιρούνται όλες οι περιττές πληροφορίες ώστε στη συνέχεια να χρησιμοποιηθούν τα «καθαρά» δεδομένα για την κατασκευή ή την ενημέρωση των **προφίλ των χρηστών (users' profiles)**. Το προφίλ κάθε χρήστη αναπαρίσταται εσωτερικά με βάση κάποιο **μοντέλο (model)** το οποίο επιλέγεται κατάλληλα ώστε να επιτρέψει την αποδοτικότερη εφαρμογή των κανόνων εξόρυξης γνώσης που θα χρησιμοποιηθούν σε τελικό στάδιο για την εξατομίκευση της ηλεκτρονικής μαθησιακής εμπειρίας (Fink et al., 1996). Η διαδικασία καθορισμού του μοντέλου αυτού είναι γνωστή σαν **μοντελοποίηση χρήστη ή user modeling** και αποτελεί μια ώριμη επιστημονική περιοχή που έχει να επιδείξει ένα σημαντικό σύνολο τεχνικών, οι οποίες έχουν αξιολογηθεί στην πράξη. Οι πιο γνωστές από αυτές τις τεχνικές είναι οι Bayesian ή πιθανοτικές, οι βασισμένες σε κανόνες, τα νευρωνικά δίκτυα, οι βασισμένες σε στερεότυπα και οι επαγωγικοί κανόνες (Fink & Kobsa, 2000). Τα προφίλ που προκύπτουν περιέχουν το σύνολο των πληροφοριών που έχουν

συλλεχθεί για κάθε χρήστη και κρατούνται σε μια βάση δεδομένων, η οποία και ενημερώνεται συνεχώς με τις νέες πληροφορίες που καταγράφονται κατά την παρακολούθηση των ενεργειών κάθε χρήστη στο σύστημα.

1.5.3 Ανάλυση των δεδομένων

Μετά από τη φάση της συλλογής των δεδομένων και του καθαρισμού τους, ακολουθεί η περαιτέρω επεξεργασία τους προκειμένου να εντοπιστούν **πρότυπα ή μοτίβα (patterns)** συμπεριφορών. Έτσι, το επόμενο βήμα είναι η **ανάλυση (analysis)** των διαθέσιμων στοιχείων και η **εφαρμογή τεχνικών Web Mining**. Στις πιο συχνά χρησιμοποιούμενες τεχνικές περιλαμβάνονται οι:

- **Κατηγοριοποίηση (classification)**. Εφαρμόζεται στο σύνολο των δεδομένων ώστε να εξαχθούν οι κλάσεις των χρηστών με βάση **προαποφασισμένες ετικέτες (class labels)**. Για παράδειγμα, έστω ότι θέλουμε να δημιουργήσουμε τις κλάσεις των χρηστών που παρακολουθούν τη θεματική ενότητα Microsoft Excel. Με τη βοήθεια **δέντρων αποφάσεων (decision trees)** μπορούμε να βρούμε ότι το προφίλ αυτών των χρηστών είναι: άτομα ηλικίας μικρότερης των 30 ετών, ανύπαντρα και χωρίς παιδιά, απόφοιτοι τριτοβάθμιας εκπαίδευσης που χρησιμοποιούν ηλεκτρονικό υπολογιστή στην εργασία τους. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν αργότερα για να προτείνουμε συγκεκριμένο εκπαιδευτικό πλάνο σε κάθε χρήστη. Η κατηγοριοποίηση ανήκει στις **προγνωστικές τεχνικές (predictive web mining)**. Για να λειτουργήσει απαιτείται ένα σύνολο δεδομένων εκπαίδευσης (training data set) ώστε να αναγνωριστούν τα ιδιαίτερα χαρακτηριστικά κάθε κλάσης και στη συνέχεια να ταξινομηθούν τα νέα αντικείμενα με βάση τις τιμές τους στα χαρακτηριστικά αυτά. Οι πιο γνωστοί αλγόριθμοι που χρησιμοποιούνται είναι: η απλοϊκή κατηγοριοποίηση κατά Bayes (naive Bayes), ο K-πλησιέστερος γείτονας (K-nearest neighbor ή KNN), τα δέντρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), κλπ. (Chakrabarti, 2003).
- **Συσταδοποίηση ή ομαδοποίηση (clustering)**. Η λογική της τεχνικής είναι ο διαχωρισμός των χρηστών σε ομάδες με βάση το ότι χρήστες που ανήκουν στην ίδια ομάδα είναι όσο το δυνατόν περισσότερο όμοιοι, ενώ χρήστες διαφορετικών ομάδων είναι όσο το δυνατόν πιο διαφορετικοί (Han & Kamber, 2001). Η ομοιότητα μετριέται χρησιμοποιώντας μια **συνάρτηση ομοιότητας (similarity function)** που εφαρμόζεται στα προφίλ των χρηστών. Αφού δημιουργηθούν οι ομάδες, είναι στη συνέχεια δυνατό να τους αποδοθούν ετικέτες ανάλογα με το είδος της κάθε ομάδας και να χρησιμοποιηθούν οι ετικέτες αυτές σαν βάση για την εφαρμογή κατηγοριοποίησης (με αυτό τον τρόπο χρησιμοποιούμε την ομαδοποίηση σαν ένα αρχικό βοηθητικό στάδιο πριν εφαρμόσουμε κατηγοριοποίηση, στις περιπτώσεις που δεν είναι εύκολο να προκαθορίσουμε τις κλάσεις). Εκτός από την ομαδοποίηση των χρηστών είναι δυνατή η εφαρμογή του clustering με στόχο το σχηματισμό ομάδων σελίδων. Αυτό σημαίνει ότι σελίδες που αντιμετωπίζονται με παρόμοιο τρόπο από τους χρήστες ή περιέχουν παρόμοιο περιεχόμενο ομαδοποιούνται στο ίδιο cluster. Πρέπει να σημειωθεί ότι είναι απαραίτητο να επιλέγονται τεχνικές που επιτρέπουν την αλληλοεπικάλυψη των clusters (overlapping), καθώς στις περισσότερες περιπτώσεις ένα χρήστης ή μια σελίδα ανήκουν σε περισσότερα από ένα clusters. Αρκετοί αλγόριθμοι ομαδοποίησης είναι διαθέσιμοι: η ιεραρχική

συσσωρευτική συσταδοποίηση (hierarchical agglomerative clustering ή HAC) (Rasmussen, 1992), (Willett, 1988), ο k-means (MacQueen, 1967), οι αυτό-οργανωμένοι χάρτες χαρακτηριστικών (self-organizing maps ή SOMs) με πιο γνωστούς τους χάρτες Kohonen (Kohonen, 1997), κλπ.

- **Κανόνες συσχέτισης (association rules).** Χρησιμοποιούνται για να εντοπίσουν ομοιότητες μεταξύ διαφορετικών τύπων δεδομένων και να επιτευχθεί καλύτερη κατανόηση της συμπεριφοράς των χρηστών. Οι κανόνες αυτοί συνδέουν ένα ή περισσότερα (αρχικά μη συσχετιζόμενα) γεγονότα και ανακαλύπτουν σχέσεις που δεν μπορούν (εύκολα) να προβλεφθούν (Agrawal et al., 1993), (Chen et al., 1996), (Chen et al., 1998). Ένας τέτοιος κανόνας έχει τη μορφή $A \Rightarrow B$ και υποδηλώνει μια σχέση εξάρτησης ανάμεσα σε δύο (ξένα) σύνολα στοιχείων A και B δηλαδή υπολογίζει την πιθανότητα να συμβεί το B με δεδομένο ότι έχει συμβεί το A . Για παράδειγμα, σε ένα e-learning σύστημα ένας κανόνας συσχέτισης θα μπορούσε να είναι ότι το «20% των χρηστών που μελετούν το μάθημα Microsoft Windows 2000 επιλέγουν στη συνέχεια να μελετήσουν το μάθημα Microsoft Word 2000», ή ότι το «50% των χρηστών που επισκέφθηκαν τις σελίδες της βοήθειας ανήκουν στην ηλικιακή ομάδα 25-30». Ο πιο γνωστός αλγόριθμος για την ανακάλυψη κανόνων συσχέτισης είναι ο Apriori (Agrawal & Srikant, 1994) και οι παραλλαγές του, ενώ υπάρχουν και προσεγγίσεις που χρησιμοποιούν Bayesian Networks (Ardissono & Torasso, 2000), (Schwarzkopf, 2001). Μια επέκταση των κανόνων συσχέτισης αποτελεί η τεχνική **ανακάλυψης ακολουθιακών μοτίβων ή sequential pattern discovery** η οποία ενσωματώνει την έννοια του χρόνου στα patterns. Ένα παράδειγμα θα μπορούσε να είναι της μορφής $(C)(A, B)$ όπου A , B , C είναι σελίδες ενός e-learning site και σημαίνει ότι οι εκπαιδευόμενοι που επισκέπτονται τη σελίδα C κάποια στιγμή στη συνέχεια ζητούν να δουν τις σελίδες A και B (είναι δυνατόν να παρεμβληθούν και άλλες σελίδες μεταξύ των συγκεκριμένων αιτήσεων). Οι περισσότερες υλοποιήσεις αυτής της τεχνικής βασίζονται σε αλγορίθμους που μοιάζουν με τον Apriori (Han & Kamber, 2001), καθώς και σε μοντέλα Markov (Zukerman et al, 1993), (Pierrakos et al., 2003).

1.5.4 Ολοκλήρωση των δεδομένων

Τα συμπεράσματα και η γνώση που εξορύσσεται με βάση μια ή περισσότερες από τις παραπάνω τεχνικές χρησιμοποιούνται σαν είσοδος στην επόμενη φάση της διαδικασίας εξατομίκευσης που είναι γνωστή σαν **ολοκλήρωση δεδομένων**, καθώς συνδυάζει το σύνολο των πληροφοριών που είτε συλλέχθηκαν είτε παράχθηκαν στην πορεία. Ανάλογα με την κατά περίπτωση τεχνική (ή τις τεχνικές) εξατομίκευσης που εφαρμόζει ένα e-learning σύστημα, ενδέχεται να χρησιμοποιεί σαν είσοδο και πληροφορίες που σχετίζονται με τη δομή και την τοπολογία του εκπαιδευτικού υλικού, το ακριβές περιεχόμενο ή το σκοπό μιας ιστοσελίδας, το πάτημα κάποιου κουμπιού, την επιλογή κάποιου συγκεκριμένου συνδέσμου, το χρόνο παραμονής σε μια ιστοσελίδα και γενικότερα δεδομένα που σχετίζονται με τη συμπεριφορά των χρηστών σε επίπεδο ακολουθίας κλικ (clickstream) σε συνδυασμό όμως με τη σημασιολογία κάθε ενέργειας στα πλαίσια του συγκεκριμένου συστήματος. Σε αυτές τις περιπτώσεις, εκτός από το λογικό επακόλουθο της αύξησης της πολυπλοκότητας του συστήματος, όσο και των απαιτήσεών του σε πόρους, για την εξαγωγή συμπερασμάτων, είναι απαραίτητη η συμβολή εκπαιδευτών οι οποίοι μπορούν να αντιληφθούν και να

επεξηγήσουν τη μαθησιακή συμπεριφορά των χρηστών και να προτείνουν το είδος της κατάλληλης εξατομικευμένης προσαρμογής.

1.5.5 Παραγωγή εξατομικεύσεως

Στη συνέχεια περνάμε στην τελική φάση στην οποία παράγουμε για κάθε χρήστη μια διαφορετική **εξατομικευμένη εμπειρία μάθησης** με βάση τις συγκεκριμένες προτιμήσεις και τις ανάγκες του, ώστε κάθε φορά που επιστρέφει ο χρήστης στο σύστημα, αυτό να προσαρμόζεται κατά το δυνατό στο δικό του προφίλ. Αυτό το γεγονός αυξάνει τη φιλικότητα του συστήματος προς το χρήστη κάνοντας την εκπαιδευτική διαδικασία αποδοτικότερη και διευκολύνοντας την απόκτηση νέων δεξιοτήτων, παρέχει αμεσότητα στην αλληλεπίδραση, αυξάνει την εμπιστοσύνη των χρηστών προς το σύστημα και βελτιώνει τη συνολική εμπειρία της ηλεκτρονικής μάθησης. Ενδεικτικές προσαρμογές που μπορούν να πραγματοποιηθούν σε ένα e-learning σύστημα ώστε να επιτευχθεί εξατομικεύση περιλαμβάνουν:

- **Διαφοροποιήσεις στο εκπαιδευτικό περιεχόμενο:** εισαγωγή/αφαίρεση θεματικών εννοιών, τμημάτων ή παραγράφων, προαιρετικές εξηγήσεις ή αναλυτικές πληροφορίες, εξατομικευμένες συστάσεις για μελέτη άλλων θεματικών εννοιών, κλπ.
- **Διαφοροποιήσεις στη δομή του εκπαιδευτικού υλικού:** τρόπος ταξινόμησης, κρύψιμο, προσθήκη, αφαίρεση ή σήμανση υπερσυνδέσμων, κλπ.
- **Διαφοροποιήσεις στον τρόπο παρουσίασης του εκπαιδευτικού περιεχομένου:** με περισσότερες ή λιγότερες εικόνες, με ή χωρίς ήχο, με ή χωρίς βίντεο, με μεγαλύτερη ή μικρότερη γραμματοσειρά, χρωματισμένο κείμενο ή όχι, κλπ.

1.6 Ένα σενάριο για σημασιολογικά εξατομικευμένο e-learning

1.6.1 Ορισμός του προβλήματος παραγωγής εξατομικεύσεως σε ένα e-learning σύστημα στο θεματικό Web Mining

Θεωρούμε ένα e-learning web site που παρέχει online μαθήματα στους χρήστες του. Το συγκεκριμένο θεματικό αντικείμενο που μπορούν να παρακολουθήσουν οι χρήστες σχετίζεται με το «Web Mining». Το πρόβλημα που μελετάται στις επόμενες ενότητες αφορά στην παρακολούθηση του εκπαιδευόμενου κατά την πλοήγησή του στο περιεχόμενο του μαθήματος ώστε να του προταθεί ποιο μονοπάτι θα ακολουθήσει στη συνέχεια. Συγκεκριμένα, το πρόβλημα μπορεί να ορισθεί ως εξής:

Έστω U το σύνολο όλων των χρηστών (users) και P το σύνολο όλων των πιθανών σελίδων (pages) του web site. Κάθε σελίδα προσδιορίζεται από το URI (Uniform Resource Identifier). Έστω το μάθημα C (course) του web site, όπου:

$$C = \{p_o, p_1, \dots, p_i\}$$

Κάθε web page $p \in P$ καθορίζεται ως εξής:

$$p = \{pageID, content, topic, level\}$$

όπου:

- $pageID$: είναι ένα ενιαίο (μοναδικό) στοιχείο με το οποίο αναγνωρίζεται η σελίδα.
- $contents$: είναι το περιεχόμενο της σελίδας p .
- $topic$: περιλαμβάνει το θέμα της σελίδας p .
- $level$: καθορίζει το επίπεδο της σελίδας p και είναι της μορφής αρχάριο επίπεδο, ενδιάμεσο επίπεδο, προχωρημένο επίπεδο, κλπ.

Η επίσκεψη ενός χρήστη $u \in U$ στο web site είναι ένα σύνολο από **διαδοχικά clicks του ποντικιού (clickstream)**. Είναι δηλαδή μια χρονικά διατεταγμένη ακολουθία ενεργειών του χρήστη που μπορούν σε συνδυασμό να αποτελέσουν τις **συνόδους (sessions)**. Ένα session ορίζεται σαν μια ακολουθία από **pageviews** που ζητήθηκαν (μέσω διαδοχικών αιτήσεων HTTP) από ένα και μοναδικό χρήστη και εξυπηρετήθηκαν από ένα και μοναδικό server, όπου pageview είναι μια σελίδα του site που ολοκληρώθηκε στην οθόνη του χρήστη. Αυτό σημαίνει ότι αυτό που καταλήγει να βλέπει ο χρήστης ενός web site στο παράθυρο του browser δεν είναι ένα και μόνο html αρχείο αλλά ένας συνδυασμός από τέτοια αρχεία και άλλα συστατικά (εικόνες, αρχεία βίντεο, ήχου, flash animations, κλπ.) που όταν ζητηθούν και καταφθάσουν όλα από τους αντίστοιχους web servers συμπληρώνουν τα περιεχόμενα του παραθύρου. Το σύνολο όλων αυτών των συστατικών αποτελούν ένα pageview.

Για το χρήστη $u \in U$ θέλουμε να επιλέξουμε εκείνη τη σελίδα $p \in C$ δεδομένου της **τρέχουσας πλοήγησής του (current session window)** στο μάθημα. Για παράδειγμα, αν χρησιμοποιεί ένα σταθερού-μεγέθους παράθυρο είναι δυνατόν να συλληφθεί η τρέχουσα ιστορία του χρήστη. Αυτό σημαίνει ότι για παράθυρο μεγέθους 3, αν το current session window είναι $\langle A, B, C \rangle$ και ο χρήστης πηγαίνει στη σελίδα D τότε το νέο current session window θα είναι το $\langle B, C, D \rangle$.

Για κάθε χρήστη $u \in U$ καθορίζεται το **προφίλ (profile)** του ως εξής:

$$Userprofile = \{userID, demographics, history, background\}$$

όπου:

- $userID$: είναι ένα ενιαίο (μοναδικό) στοιχείο με το οποίο αναγνωρίζεται η ταυτότητα του χρήστη.
- $demographics$: περιλαμβάνει τα δημογραφικά στοιχεία του χρήστη όπως το ονοματεπώνυμό του, η ηλικία του, το φύλο του, το επάγγελμά του, η οικογενειακή του κατάσταση, κλπ.
- $history$: περιλαμβάνει το ιστορικό πλοήγησής του χρήστη και είναι της μορφής $\{\langle p_1, r_1 \rangle, \dots, \langle p_n, r_n \rangle\}$ όπου p_i είναι η web page i που ο χρήστης έχει μελετήσει και r_i η προτίμησή του (ή απόδοσή του) για τη σελίδα p_i για $i = 1, 2, \dots, n$.

- *background*: καθορίζει το μορφωτικό επίπεδο του χρήστη και είναι της μορφής $\{<topic_{p_1}, l_1>, \dots, <topic_{p_n}, l_n>\}$ όπου $topic_{p_i}$ είναι το θέμα της p_i και l_i είναι το επίπεδο του χρήστη για το $topic_{p_i}$ π.χ. αρχάριος, ενδιάμεσος, προχωρημένος, κλπ. για $i = 1, 2, \dots, n$.

1.6.2 Περιορισμοί και προβλήματα των Web Mining προσεγγίσεων

Ο στόχος των Web Mining προσεγγίσεων είναι να συλληφθούν και να διαμορφωθούν τα πρότυπα συμπεριφοράς και τα προφίλ των χρηστών καθώς αυτοί αλληλεπιδρούν με ένα web site. Τα πρότυπα που προκύπτουν αντιπροσωπεύονται συνήθως σαν συλλογές σελίδων που προσεγγίζονται συχνά από ομάδες χρηστών με κοινές ανάγκες και ενδιαφέροντα και μπορούν να χρησιμοποιηθούν για να κατανοήσουν τα χαρακτηριστικά και τη συμπεριφορά ενός χρήστη ή μιας ομάδας χρηστών, να βελτιώσουν την οργάνωση και τη δομή του site και να παράγουν εξατομικευμένες εμπειρίες για αυτούς με την παροχή δυναμικών προτάσεων (Lui & Li, 2002), (Συρμακέσης et al., 2003). Αν και τα πλεονεκτήματα από τη χρήση των Web Mining τεχνικών είναι σημαντικά, τα ελλιπή δεδομένα που καταγράφονται από τους web servers στα logs οδηγούν στην ανάγκη για την υιοθέτηση πιο εξειδικευμένων τεχνικών ώστε να εμπλουτιστούν αυτά τα δεδομένα και να οδηγηθούμε σε αποδοτικότερες εξατομικευμένες εμπειρίες (Zaiane, 2001).

Σε ένα εννοιολογικό επίπεδο είναι δυνατόν να υπάρξουν πολλά διαφορετικά είδη αντικειμένων μέσα σε ένα site που μπορούν να προσπελάσουν οι χρήστες. Στο φυσικό επίπεδο, αυτά τα αντικείμενα μπορούν να αναπαρασταθούν σε μια ή περισσότερες σελίδες. Για παράδειγμα, σε ένα e-learning site είναι δυνατόν να περιέχονται σελίδες σχετικές με ένα μάθημα, τις ασκήσεις που αυτό το μάθημα περιλαμβάνει, τα παραδείγματα, τα πειράματα, κλπ. Εννοιολογικά, κάθε μια από αυτές τις οντότητες αντιπροσωπεύει ένα διαφορετικό τύπο σημασιολογικού αντικειμένου. Κατά τη διάρκεια μιας επίσκεψης σε αυτό το site, ένας χρήστης μπορεί να έχει πρόσβαση σε αρκετά από αυτά τα αντικείμενα μαζί κατά τη διάρκεια ενός session.

Τα προφίλ των χρηστών από την άλλη πλευρά συλλαμβάνουν μόνο τα κοινά πρότυπα χρήσης στο επίπεδο των σελίδων. Δεν απεικονίζουν τους μη προφανείς λόγους για τους οποίους ένας χρήστης ή μια ομάδα χρηστών ενδιαφέρεται για συγκεκριμένες σελίδες. Αυτό μπορεί να οδηγήσει σε σημαντικούς περιορισμούς των συστημάτων εξατομίκευσης που βασίζονται σε Web Mining τεχνικές. Για παράδειγμα, εάν ένα νέο μάθημα έχει προστεθεί πρόσφατα στο web site είναι πιθανό οι σελίδες του να μην αποτελούν μέρος οποιουδήποτε από τα ανακαλυπτόμενα πρότυπα λόγω έλλειψης ικανοποιητικών σε ποσότητα και ποιότητα δεδομένων χρήσης. Αυτό το πρόβλημα είναι γνωστό και σαν «**πρόβλημα νέου αντικειμένου**» (**new item problem**). Από την άλλη πλευρά, ένας χρήστης που ταιριάζει σε ένα προφίλ μπορεί πράγματι να ενδιαφερθεί για το νέο μάθημα, επειδή μη προφανή χαρακτηριστικά του μαθήματος μπορεί να σχετίζονται με εκείνα τα χαρακτηριστικά άλλων μαθημάτων που έχουν παρακολουθήσει άλλοι χρήστες.

Μια προσέγγιση για την επίλυση των παραπάνω προβλημάτων είναι να ενσωματωθούν τα χαρακτηριστικά περιεχομένου των σελίδων με τα πρότυπα χρήσης. Γενικά, είναι δυνατόν να εξαχθούν **λέξεις κλειδιά (keywords)** από το περιεχόμενο του web site και να χρησιμοποιηθούν είτε για την ευρετηρίαση των σελίδων του περιεχομένου είτε για την ταξινόμηση τους σε διάφορες κατηγορίες. Στα πλαίσια της εξατομίκευσης, αυτή η

προσέγγιση θα επέτρεπε στο σύστημα να συστήσει σελίδες σε ένα χρήστη, όχι μόνο βασισμένο στο ταίριασμα με το προφίλ χρήσης, αλλά και (ή εναλλακτικά) βασισμένο στην ομοιότητα του περιεχομένου αυτών των σελίδων με τις σελίδες που ο χρήστης έχει ήδη επισκεφτεί.

Εντούτοις, οι βασισμένες στα keywords προσεγγίσεις είναι ανίκανες να συλλάβουν τις πιο σύνθετες σχέσεις μεταξύ των αντικειμένων, δηλαδή σε βαθύτερο σημασιολογικό επίπεδο που αφορούν στις έμφυτες ιδιότητες που συνδέονται με αυτά τα αντικείμενα. Στο παράδειγμα του e-learning, η παραπάνω προσέγγιση επιτρέπει στο σύστημα να προτείνει άλλα μαθήματα ή αντικείμενα ενός μαθήματος με βάση την ομοιότητα του περιεχομένου ή άλλα χαρακτηριστικά. Όμως το σύστημα παρουσιάζει ιδιαίτερη δυσκολία να προτείνει σελίδες που έχουν π.χ. κοινά παραδείγματα. Για να είναι σε θέση το σύστημα να συστήνει διαφορετικούς τύπους σύνθετων αντικειμένων χρησιμοποιώντας τις μη προφανείς ιδιότητες και τα χαρακτηριστικά τους, θα πρέπει να βασίζεται στο χαρακτηρισμό των τμημάτων και των αντικειμένων όχι μόνο με λέξεις κλειδιά αλλά και σημασιολογικά χρησιμοποιώντας τις οντολογίες περιοχών για αυτά τα αντικείμενα.

Στη συνέχεια προτείνεται μια προσέγγιση, η οποία λαμβάνει υπόψη της τη δομή και τη σχέση των δεδομένων και των πληροφοριών μέσω της **οντολογίας** του e-learning site. Η οντολογία αυτή αποτελεί ένα χρήσιμο εργαλείο που συνεισφέρει ουσιαστικά στη δόμηση, στην οργάνωση και στη διαχείριση του περιεχομένου που σχετίζεται με μια δεδομένη σειρά μαθημάτων, ένα μάθημα ή ένα θεματικό. Η όλη διαδικασία χωρίζεται σε δύο στάδια. Στις offline ενέργειες προετοιμασίας των δεδομένων, δημιουργίας της οντολογίας και **εξόρυξης από δεδομένα χρήσης (usage mining)** και στην online παροχή της εξατομίκευσης.

Το σύστημα βρίσκει σε πρώτη φάση ένα πιθανό σύνολο συστάσεων βασισμένο στην οντολογία του πεδίου και στη συνέχεια χρησιμοποιεί τα **frequent itemsets (συχνά εμφανιζόμενα σύνολα στοιχείων)** για να το εμπλουτίσει, λαμβάνοντας υπόψη την πλοήγηση άλλων **παρόμοιων χρηστών (similar users)**. Με αυτό τον τρόπο, εστιάζουμε μόνο σε εκείνα τα σύνολα που προέρχονται από το συνδυασμό της **ενεργούς συνόδου (current session)** του χρήστη και των συστάσεων της οντολογίας. Αυτή η προσέγγιση ανακουφίζει το πρόβλημα των μεγάλων χρόνων απόκρισης, το οποίο μπορεί στη συνέχεια να οδηγήσει στην εγκατάλειψη του e-learning συστήματος, ενώ μειώνει το χρόνο που απαιτείται για την ανάλυση όλων των frequent itemsets και των κανόνων συσχέτισης οδηγώντας σε ένα μικρότερο χώρο αναζήτησης.

1.6.3 Σχετικές εργασίες

Πριν συνεχίσουμε με την παρουσίαση της προτεινόμενης προσέγγισης θεωρούμε σκόπιμο να καταγράψουμε σύντομα μερικές από τις πιο αντιπροσωπευτικές εργασίες του πεδίου. Είναι άλλωστε γεγονός ότι η τρέχουσα έρευνα στην περιοχή των οντολογιών αποδεικνύει τον ουσιαστικό ρόλο που μπορούν να παίξουν στην υλοποίηση e-learning συστημάτων. Συγκεκριμένα, οι Stojanovic et al., (2001) περιγράφουν πώς οι Semantic Web τεχνολογίες μπορούν να χρησιμοποιηθούν για την ανάπτυξη σύνθετων e-learning σεναρίων ώστε να βελτιώσουν τη διαχείριση των πόρων τους. Σε αυτή την περίπτωση οι οντολογίες χρησιμοποιούνται για την περιγραφή της σημασιολογίας και για τον καθορισμό του πλαισίου εκπαίδευσης του υλικού, καθώς επίσης και για τη δόμηση των σειρών των μαθημάτων.

Ένα πλαίσιο για εξατομικευμένο e-learning στην περιοχή του Semantic Web καθώς και ο τρόπος που τα σχήματα περιγραφής των πόρων μπορούν να χρησιμοποιηθούν για την αυτόματη παραγωγή των δομών υπερκειμένου από κατανεμημένα μετα-δεδομένα παρουσιάζεται από τους Henze et al., (2004). Ειδικότερα, διαφορετικές οντολογίες χρησιμοποιούνται για την περιγραφή των χαρακτηριστικών γνωρισμάτων των περιοχών, των χρηστών και των παρατηρήσεων. Η προσέγγισή τους βασίζεται στην TRIPLE μια γλώσσα ερωτήσεων με κανόνες.

Το Courseware Watchdog είναι ένα ολοκληρωμένο εργαλείο βασισμένο σε οντολογίες που επιτρέπει τη δημιουργία των περισσότερων πόρων e-learning που είναι διαθέσιμοι στο Web (Tane et al., 2004). Το εργαλείο καλύπτει τις διαφορετικές ανάγκες τόσο των εκπαιδευτών όσο και των εκπαιδευόμενων και οργανώνει το μαθησιακό υλικό σύμφωνα με αυτές.

Το ELENA πρόγραμμα (<http://www.elena-project.org>) έχει σαν στόχο την εξατομικευμένη πρόσβαση σε διανεμημένες e-learning αποθήκες εκμεταλλευόμενο τις σημασιολογικές τεχνολογίες και τα πρότυπα περιγραφής των μετα-δεδομένων. Οι ρητές περιγραφές των εκπαιδευτικών αντικειμένων χρησιμοποιούν συσχετισμούς RDF σε LOM και DC ενώ οι των χρηστών RDFSchemas των προτύπων PAPI και IMS LIP που επιτρέπουν δυνατότητες ερωτοαπαντήσεων στην P2P υποδομή του Edutella. Η προσέγγιση είναι βασισμένη σε κανόνες ταιριάσματος των αντικειμένων εκμάθησης και των περιγραφών των χρηστών με στόχο την παραγωγή προτάσεων υπηρεσιών ή κατανεμημένων αντικειμένων και την προσαρμοσμένη πρόσβαση και παράδοσή τους (Dolog & Nejd, 2003).

Οι Hatem et al., (2005) πρότειναν μια προσέγγιση που εξασφαλίζει τη δημιουργία μιας αποθήκης RDF η οποία απεικονίζει τα πραγματικά δεδομένα και τη σημασιολογία όλων των πόρων και των σειρών μαθημάτων καθώς και των υλικών της e-βιβλιοθήκης. Επίσης, επιτρέπει το σχολιασμό και τη σήμανση του εκπαιδευτικού υλικού. Οι σειρές των μαθημάτων προσπελαύνονται σύμφωνα με τα προφίλ των χρηστών και τα προνόμια που τους έχουν αποδοθεί. Το σενάριο αυτό υλοποιήθηκε στο Sultan Qaboos University (SQU).

Τέλος, μια επισκόπηση σχετικά με τη χρήση των οντολογιών και των μετα-δεδομένων στο πεδίο του e-learning, καθώς και καινοτόμες προσεγγίσεις και τεχνικές παρουσιάζονται από τους Brase & Nejd (2004). Οι συγγραφείς δίνουν έμφαση στα σχετικά πρότυπα μετα-δεδομένων, στις συνδέσεις, στα σχήματα και στα σχόλια, τις ταξινομήσεις για την περιγραφή του περιεχομένου ή του θέματος ενός πόρου, κλπ. Στη συνέχεια, εισάγουν διαφορετικές οντολογίες και παρουσιάζουν ένα RDF peer-to-peer δίκτυο για τους ψηφιακούς πόρους και για την ανταλλαγή των εκπαιδευτικών αντικειμένων και των υπηρεσιών.

1.6.4 Η e-learning οντολογία

Ας υποθέσουμε ένα απλό e-learning σενάριο σύμφωνα με το οποίο ένας φοιτητής ενός Τμήματος «Πληροφορικής» επιθυμεί να διαβάσει εκπαιδευτικό υλικό σχετικά με τις τεχνικές «Web Mining» για να μπορέσει να ετοιμάσει μια εργασία για το μάθημα «Web Data Mining for Business Intelligence». Για το σκοπό αυτό αποφασίζει να εγγραφεί στο πιλοτικό e-learning site που έχουμε κατασκευάσει ώστε να παρακολουθήσει το μάθημα και να πλοηγηθεί στο εκπαιδευτικό περιεχόμενό του. Ένας άλλος φοιτητής, ο οποίος ήδη γνωρίζει τα βασικά θέματα του «Web Mining» θέλει να μάθει για το συγκεκριμένο topic «Web Usage

Mining». Είναι σαφές ότι ο πρώτος φοιτητής θα ακολουθήσει αρχικά το πλάνο του μαθήματος όπως έχει καθοριστεί από το σχεδιασμό του. Στη συνέχεια, ανάλογα με την πορεία του, το πλάνο αυτό μπορεί να αλλάξει για να προσαρμοστεί καλύτερα στην πλοήγησή του. Ο δεύτερος φοιτητής δεν θα παρακολουθήσει από την αρχή το μάθημα αλλά θα στείλει ένα ερώτημα στο site με το οποίο θα ζητάει συγκεκριμένες σελίδες που περιέχουν εκπαιδευτικό υλικό για το «Web Usage Mining». Οι σελίδες που θα του επιστρέψει το σύστημα θα έχουν δείκτες που θα προσδιορίζουν τη σχετικότητα της σελίδας και θα βασίζονται αρχικά στα στοιχεία του προφίλ του π.χ. εκπαιδευτικό επίπεδο, προτιμήσεις, κλπ.

Για την υλοποίηση του παραπάνω σεναρίου απαιτείται πληροφορία τόσο για το εκπαιδευτικό υλικό όσο και για τον εκπαιδευόμενο. Για το σκοπό αυτό επιλέχθηκε η υιοθέτηση μιας οντολογίας η οποία χρησιμοποιείται για να δημιουργήσει μια αναπαράσταση της e-learning περιοχής με τον καθορισμό όλων των εννοιών, των πιθανών σχέσεων μεταξύ τους και άλλων ιδιοτήτων, όρων ή κανονισμών της περιοχής. Η ανάπτυξη της οντολογίας είναι συγγενής με τον καθορισμό ενός συνόλου στοιχείων και της δομής τους. Κατά αυτό τον τρόπο, η οντολογία μπορεί να θεωρηθεί σαν μια βάση γνώσεων που χρησιμοποιείται περαιτέρω για την εξαγωγή χρήσιμης «κρυμμένης» γνώσης και την παραγωγή εξατομικευμένων μορφών του e-learning συστήματος. Τα βήματα που ακολουθήθηκαν για την κατασκευή της οντολογίας του πιλοτικού e-learning συστήματος βασίζονται σε μια **middle-out προσέγγιση** η οποία ξεκινά από τον καθορισμό των πιο σημαντικών εννοιών και σταδιακά οδηγείται τόσο προς τη **γενίκευση** όσο και προς τη **συγκεκριμενοποίηση**. Τα βήματα αυτά παρουσιάζονται στη συνέχεια:

- Αρχικά καθορίστηκε το πεδίο και ο σκοπός της οντολογίας. Στην περίπτωση μας, το πεδίο εφαρμογής είναι μια σειρά online μαθημάτων για το θεματικό «Web Mining» και ο στόχος είναι η αναπαράσταση όλων των εννοιών που το απαρτίζουν. Πιθανές ερωτήσεις που θα πρέπει να απαντηθούν είναι της μορφής «μετά το θεματικό αντικείμενο Web Usage Mining τι πρέπει να μελετήσω;» ή «τι είδους σύνδεση υπάρχει μεταξύ δυο παραδειγμάτων;», κλπ.
- Στη συνέχεια καθορίστηκαν οι σημαντικοί όροι όπως: Course, Lesson, Teacher, Learner, LearningResource, Example, κλπ.
- Ακολούθησε ο ορισμός των κλάσεων, των υποκλάσεων και της ιεραρχίας τους.
- Το επόμενο βήμα ήταν ο καθορισμός των συσχετίσεων. Στο βήμα αυτό ορίστηκαν οι σχέσεις μεταξύ των κλάσεων της οντολογίας (επίπεδο πολλαπλότητας μιας συσχέτισης και το είδος της τιμής της).

Για λόγους απλότητας της υλοποίησης επιλέχθηκε η χρήση standards όπου αυτό ήταν εφικτό. Αυτά καθορίζουν για παράδειγμα τις ιδιότητες κάθε αντικειμένου και συνήθως τα ταξινομούν σε συγκεκριμένες κατηγορίες. Για το μάθημα «Web Mining» υιοθετήθηκε η ταξινόμια που προτάθηκε από τους Cooley et al., (1997) και Chakrabarti et al., (1999). Ένα τμήμα της απεικονίζεται στο Σχήμα 7. Τα συγκεκριμένα θεματικά αντικείμενα περιγράφονται από τα concepts τους και τις αμοιβαίες σχέσεις τους στην περιοχή. Τα concepts της περιοχής μπορεί να έχουν πολύπλοκες δομές. Στο παράδειγμα απεικονίζεται ένα τμήμα της γνώσης της περιοχής που καλύπτει το θεματικό του «Web Mining» και περιλαμβάνει τη σχέση isa (subConceptOf) μεταξύ των concepts. Συνεπώς, το concept «Web Mining» έχει subconcepts το «Web Content Mining», το «Web Structure Mining» και το «Web Usage

Mining». Τα «Agent-based Approach» και «Database Approach» αποτελούν subconcepts του «Web Content Mining», κ.ο.κ.

Σχήμα 7. Τμήμα της οντολογίας για τη σειρά μαθημάτων «Web Mining».

Επίσης, ένα τμήμα της οντολογίας που παριστάνει ένα εκπαιδευτικό πόρο φαίνεται στο Σχήμα 8. Συγκεκριμένα, η κλάση **LearningResource** καθορίζει τα χαρακτηριστικά που χρησιμοποιούνται για να περιγράψουν τους πόρους όπως αυτά έχουν υιοθετηθεί από συνδυασμό των Duplin Core ή DC (<http://duplincore.org>) και Learning Objects Metadata Standard ή LOM (<http://ltsc.ieee.org/wg12/index.html>) (Dolog et al., 2004). Οι υποκλάσεις που παρουσιάζονται στο Σχήμα 8 αφορούν σε ειδικούς τύπους εκπαιδευτικού υλικού:

- **Course:** είναι ένα LearningResource που έχει ένα Location και οδηγεί σε συγκεκριμένα Qualifications.
- **LearningUnit:** είναι ένα LearningResource που έχει ένα συγκεκριμένο Duration.
- **Examples:** εξηγούν το περιβάλλον μιας έννοιας ή ενός αντικειμένου που έχει διδαχθεί και αναπαρίσταται από Source.

Τα LearningResources και οι υποκατηγορίες τους δημιουργούν δομές (dcterms: hasPart σχέση) ή μπορεί να συνδεθούν με σχέση αναγκαιότητας (dcterms: requires).

Η κλάση **Concept** χρησιμοποιείται για να περιγράψει βασικές πληροφοριακές οντότητες από την περιοχή της γνώσης που συνδέονται με τα LearningResources. Concept και LearningResource συνδέονται με την ιδιότητα dc:subject. Οι έννοιες αυτές μπορούν να περιοριστούν από τη σχέση isPrerequisiteFor εάν οι προϋποθέσεις οριστούν στο επίπεδο του Concept ή στο επίπεδο του LearningResource.

Σχήμα 8. Τμήμα της οντολογίας για το LearningResource.

Επίσης, οι πληροφορίες σχετικά με τους εκπαιδευόμενους π.χ. τα ενδιαφέροντά τους, την απόδοσή τους σε διαφορετικές σειρές μαθημάτων, τους στόχους τους, τις προτιμήσεις τους, κλπ. θα πρέπει να οργανωθούν σε μια οντολογία. Το **εννοιολογικό μοντέλο (conceptual model)** για το προφίλ ενός Learner φαίνεται στο Σχήμα 9. Αυτό περιλαμβάνει την κλάση **Learner** η οποία συνδέεται με άλλες κλάσεις όπως, **Performance**, **Preference** και **History**. Η κλάση History περιλαμβάνει εκείνα τα LearningResources τα οποία έχουν προσπελαστεί και άρα μελετηθεί από τον εκπαιδευόμενο. Η κλάση Performance συνδέεται τόσο με την κλάση Learner όσο και με την κλάση History και απεικονίζει την απόδοση του εκπαιδευόμενου με βάση το υλικό που έχει μελετήσει, ενώ η κλάση Preference δηλώνει τις προτιμήσεις του.

Σχήμα 9. Το εννοιολογικό μοντέλο του Learner.

1.6.5 Προ-επεξεργασία των δεδομένων

Όπως έχει αναφερθεί προηγουμένως η όλη διαδικασία εξατομίκευσης διακρίνεται στις ενέργειες που θα γίνουν offline και αυτές που θα γίνουν online. Στις πρώτες εντάσσονται εκτός από την κατασκευή της e-learning οντολογίας, η προ-επεξεργασία των δεδομένων και οι ενέργειες εξόρυξης των δεδομένων χρήσης. Η προ-επεξεργασία των δεδομένων έχει σαν αποτέλεσμα τη δημιουργία δομών όπως είναι το ιστορικό με τις αλληλεπιδράσεις των χρηστών με το web site ώστε να υπολογισθούν σημαντικά σημασιολογικά στοιχεία της δραστηριότητας των χρηστών που θα χρησιμοποιηθούν στο στάδιο εξόρυξης. Έχοντας στη διάθεσή μας τα προ-επεξεργασμένα δεδομένα είναι δυνατή στη συνέχεια η εκτέλεση μιας σειράς από τεχνικές εξόρυξης δεδομένων. Στην προτεινόμενη προσέγγιση εστιάζουμε στην ανακάλυψη **κανόνων συσχέτισης (association rules mining)** και χρησιμοποιούμε τον αλγόριθμο **Apriori** (Agrawal & Sricant, 1994).

Η πιο σημαντική πηγή δεδομένων clickstreams είναι ο web server στον οποίο βρίσκεται το e-learning site. Ο web server καταγράφει σε πραγματικό χρόνο την αλληλεπίδραση των χρηστών με το browser με τη μορφή αρχείων server logs. Γενικά, ένα αρχείο log καταγράφει την κίνηση και τη συμπεριφορά πλοήγησης των χρηστών στις σελίδες ενός δικτυακού τόπου (Mobasher, 2005). Το συγκεκριμένο σύστημα χρησιμοποιεί τα logs αρχεία που περιγράφουν τη δραστηριότητα πλοήγησης των χρηστών και πιο συγκεκριμένα, κάθε φορά που ο web server εξυπηρετεί μια αίτηση, μια νέα εγγραφή προστίθεται στο αρχείο log με λεπτομερείς πληροφορίες σχετικά με το τι ζητήθηκε, πότε, από ποιόν, σε ποια σελίδα βρισκόταν ο χρήστης όταν το ζήτησε, το εάν η αίτηση εξυπηρετήθηκε με επιτυχία ή υπήρχαν προβλήματα από την πλευρά του server και ένα σύνολο από επιπλέον στοιχεία που εξαρτώνται από τις ρυθμίσεις του server και τη μορφοποίηση των αρχείων log που κρατάει (Pierrakos et al., 2003).

Στην ουσία, αυτά τα αρχεία συμπυκνώνουν όλες τις πληροφορίες που σχετίζονται με τη χρήση του e-learning site από τους χρήστες. Σε αυτό το στάδιο, τα server logs καθαρίζονται σύμφωνα με τα αρχεία του web site. Αυτή η εργασία περιλαμβάνει την αφαίρεση όλων των πλεονάζουσων αναφορών. Επίσης, απαιτούνται λεπτομερείς πληροφορίες σχετικά με τη δομή του site προκειμένου να καθοριστούν ποιες προσβάσεις αρχείων σελίδων αντιστοιχούν σε ένα μοναδικό παράθυρο φυλλομετρητή και πιο συγκεκριμένα ποιο περιεχόμενο αντιστοιχεί στο αίτημα κάθε χρήστη.

Οι βασικές ενέργειες για την προ-επεξεργασία των log δεδομένων είναι οι ακόλουθες:

- **Καθαρισμός των δεδομένων.** Σε αυτό το βήμα εξετάζονται τα διαθέσιμα δεδομένα και απομακρύνονται όλες οι άσχετες ή πλεονάζουσες εγγραφές. Αυτές συνήθως προέρχονται από το γεγονός ότι το HTTP πρωτόκολλο απαιτεί ξεχωριστά αιτήματα στο server για κάθε αρχείο εικόνας, βίντεο, ήχου, κλπ. που είναι ενσωματωμένο στη web page. Έτσι τα αιτήματα για αρχεία γραφικών, βίντεο, ήχου, CGI scripts, κλπ. εντοπίζονται σε αυτό το στάδιο και απομακρύνονται με βάση την κατάληξή τους π.χ. .gif, .jpg, .avi, κλπ. (Cooley et al., 1999). Επειδή όμως η απομάκρυνση αυτών των δεδομένων μπορεί να οδηγήσει σε απώλεια χρήσιμης πληροφορίας ιδιαίτερα σε sites με πολυμεσικό περιεχόμενο, τα αρχεία αυτά μπορεί να παραμείνουν στο σύνολο δεδομένων και να συμπεριληφθούν στα αντίστοιχα pageviews σε επόμενο βήμα της προ-επεξεργασίας. Επιπλέον, φιλτράρονται οι εγγραφές που αφορούν σε

αιτήματα που δεν έχουν ικανοποιηθεί π.χ. HTTP απαντήσεις λάθους, εγγραφές από spiders ή crawlers, κλπ. (Tan & Kumar, 2002).

- **Αναγνώριση των χρηστών.** Αυτό το βήμα υλοποιήθηκε εύκολα αφού στο συγκεκριμένο e-learning site οι χρήστες δεν είναι ανώνυμοι αλλά είναι απαραίτητη η χρήση login και password μέσω μηχανισμού εγγραφής (registration) για να μπορέσει ένας εκπαιδευόμενος να παρακολουθήσει την εκπαιδευτική διαδικασία.
- **Αναγνώριση των sessions.** Το επόμενο βήμα είναι να αναγνωριστούν τα sessions από τα clickstreams κάθε χρήστη. Γενικά αυτό το βήμα είναι αρκετά δύσκολο αφού πρέπει να αναγνωριστούν ακολουθίες από γεγονότα της μορφής $A \Rightarrow B \Rightarrow C \Rightarrow B \Rightarrow D \dots$ όπου $A, B, C, D \dots$, κλπ. είναι αιτήματα σελίδων ή scripts. Η πρόκληση είναι να αναγνωρίσουμε την αρχή και το τέλος ενός session. Αυτό σε ένα σύστημα e-learning είναι αρκετά δύσκολο αφού οι εκπαιδευόμενοι είναι δυνατόν να πλοηγούνται σε άλλα sites αναζητώντας και συλλέγοντας σχετικές πληροφορίες με το μάθημά τους, ενώ το session τους στο συγκεκριμένο e-learning site είναι ακόμα «ανοικτό». Επίσης, ένα learning session μπορεί να καλύπτει περισσότερες μέρες με διαφορετικές προσβάσεις, ενώ πολλές σελίδες του δημιουργούνται δυναμικά με αιτήματα scripts π.χ. σελίδες τεστ. Συνεπώς το heuristic που συνήθως χρησιμοποιείται και ορίζει ένα **χρονικό κατώφλι** γνωστό και ως **page viewing time** (Shahabi, et al., 1997) π.χ. στα 30 λεπτά (Cooley et al., 1999) δεν μπορεί να εφαρμοστεί για να διαχωρίσει το clickstream σε sessions. Για το σκοπό αυτό τα sessions αναγνωρίζονται με βάση τις σελίδες περιεχομένου που ο χρήστης επιλέγει να δει στο site (content pages) και τα μονοπάτια που ακολούθησε για να φτάσει σε αυτές τις σελίδες (navigational pages). Χρησιμοποιούμε την ακόλουθη μέθοδο για να αναγνωρίσουμε τα sessions. Το **reference length** που εξετάζει το χρόνο που ο εκπαιδευόμενος ξοδεύει στο site. Αν αυτός ο χρόνος είναι μεγαλύτερος από ένα συγκεκριμένο όριο (threshold) τότε θεωρείται ότι η σελίδα περιέχει χρήσιμο περιεχόμενο και ταξινομούνται στις content pages διαφορετικά ταξινομούνται στις navigational pages ενώ θεωρούμε ότι η τελευταία σελίδα που ο χρήστης επισκέπτεται είναι πάντα μια σελίδα περιεχομένου.
- **Αναγνώριση των pageviews.** Το βήμα αυτό εξαρτάται σημαντικά από τη δομή και την οργάνωση του e-learning site, καθώς επίσης και από το περιεχόμενο των σελίδων και τη γνώση του πεδίου όπως αυτό καθορίστηκε κατά τη σχεδίαση και την υλοποίησή του. Το συγκεκριμένο site αποτελείται από ένα frame οπότε η έννοια του pageview ταυτίζεται με αυτή της σελίδας (HTML αρχείο). Επίσης, η σημασία κάθε pageview είναι διαφορετική αφού αυτά που αντιστοιχούν σε content pages είναι σημαντικότερα από αυτά που αντιστοιχούν σε navigational pages. Για το λόγο αυτό καταγράφεται με κάθε pageview ένας αριθμός από χαρακτηριστικά όπως: το pageviewID (συνήθως το URL αντιπροσωπεύει μοναδικά το pageview), η διάρκεια, το είδος του pageview (content page, navigational page, κλπ.) και άλλα metadata όπως ιδιότητες ή χαρακτηριστικά του περιεχομένου.
- **Αντιστοίχιση των εγγραφών log σε συγκεκριμένες e-learning δραστηριότητες.** Αυτό το βήμα έχει σαν αποτέλεσμα μια ακολουθία από δραστηριότητες του εκπαιδευόμενου της μορφής:

Login $\Rightarrow$ TestList $\Rightarrow$ SubmissionTest1 $\Rightarrow$ TestList $\Rightarrow$ ViewMessage....

- **Συμπλήρωση των μονοπατιών.** Το βήμα αυτό αφορά στη συμπλήρωση των μονοπατιών που ο εκπαιδευόμενος ακολούθησε κατά την αλληλεπίδρασή του με το e-learning site. Συχνά παρατηρείται το φαινόμενο να λείπουν από τα αρχεία log εγγραφές για αιτήσεις σελίδων τις οποίες ο χρήστης ζήτησε να επισκεφθεί κατά την πλοήγησή του. Οι αιτήσεις αυτές δεν καταγράφονται στο server επειδή λόγω του caching εξυπηρετούνται από τους proxy servers ή από την τοπική cash στο browser του χρήστη. Το heuristic που χρησιμοποιούμε για τον εντοπισμό των «χαμένων» αιτήσεων ώστε να συμπληρωθεί το ακριβές μονοπάτι που ακολούθησε ο χρήστης κατά την πλοήγησή του επιλέγει κάθε φορά το συντομότερο μονοπάτι (δηλαδή αυτό που αποτελείται από τις λιγότερες μεταβάσεις μεταξύ των σελίδων).
- **Ομαδοποίηση των sessions πρόσβασης κάθε εκπαιδευομένου με στόχο τον καθορισμό των learning sessions.** Στο βήμα αυτό εντοπίζουμε και ομαδοποιούμε τα sessions πρόσβασης ενός χρήστη με στόχο τον καθορισμό των learning sessions που έχουν σημασιολογική αξία. Έτσι επιτυγχάνεται σε μεγαλύτερο βαθμό διάσπαση μέσα στο session και παράγονται δυναμικές ομαδοποιήσεις (clusters) με ιδιαίτερη σημασία.
- **Ολοκλήρωση με άλλα δεδομένα για τους εκπαιδευόμενους ή για ομάδες εκπαιδευόμενων.** Τέλος, στο βήμα αυτό συνδυάζουμε τα δεδομένα από τα καθαρισμένα clickstreams με υπάρχοντα δεδομένα για τους εκπαιδευόμενους. Αυτά μπορεί να προέρχονται από τα προφίλ τους, στοιχεία από την αξιολόγησή τους, κλπ. Για παράδειγμα, συνδυάζοντας τις βαθμολογίες που σχετίζονται με την ολοκλήρωση συγκεκριμένων δραστηριοτήτων με την ακολουθία των γεγονότων που οδήγησαν σε αυτές τις δραστηριότητες μπορούν να αναγνωριστούν χρήσιμα πρότυπα ώστε να αναγνωριστούν ακολουθίες δραστηριοτήτων που οδηγούν σε καλά αποτελέσματα και άλλες που οδηγούν σε ότι και τόσο καλά.

Οι εργασίες προ-επεξεργασίας οδηγούν σε ένα σύνολο:

$$n \text{ pageviews } P = \{p_1, p_2, \dots, p_n\}$$

όπου κάθε pageview αναπαρίσταται μοναδικά από το συσχετισμένο URL και ένα σύνολο:

$$m \text{ user transactions } T = \{t_1, t_2, \dots, t_m\}$$

όπου κάθε $t_i \in T$ είναι υποσύνολο του P .

Η διαδικασία της προ-επεξεργασίας των web logs είναι μια απαιτητική διαδικασία και σύμφωνα με την εργασία (Edelstein, 2001) απαιτεί περίπου το 80% έως 95% της προσπάθειας και των πόρων όλης της Web Usage Mining φάσης.

1.6.6 Εξαγωγή των κανόνων συσχέτισης

Έχοντας στη διάθεσή μας το σύνολο των αλληλεπιδράσεων T , το πρόβλημα είναι να ανακαλυφθούν όλοι εκείνοι οι κανόνες συσχέτισης που έχουν τόσο **υποστήριξη (support)** σ όσο και **εμπιστοσύνη (confidence)** α μεγαλύτερη από μια συγκεκριμένη **ελάχιστη τιμή υποστήριξης (αποκαλούμενη ως minsup)** και μια **ελάχιστη τιμή εμπιστοσύνης**

(**minconf**) αντίστοιχα. Ο αλγόριθμος που χρησιμοποιείται για την εύρεση όλων των κανόνων συσχέτισης είναι ο **Apriori** (Agrawal & Srikant, 1994). Ο αλγόριθμος αυτός είναι ο πιο γνωστός για την εύρεση κανόνων συσχέτισης και χρησιμοποιείται από πολλά εμπορικά προϊόντα. Βασίζεται στην ακόλουθη ιδιότητα που είναι γνωστή σαν ιδιότητα των **συχνών στοιχειοσυνόλων (frequent itemsets)**:

«Οποιοδήποτε υποσύνολο ενός *frequent itemset* πρέπει να είναι συχνό»

Τα frequent itemsets ονομάζονται επίσης και **κλειστά προς τα κάτω** επειδή εάν ένα itemset ικανοποιεί τις απαιτήσεις της ελάχιστης υποστήριξης, το ίδιο συμβαίνει και για όλα τα υποσύνολά του. Επιπλέον, εάν γνωρίζουμε ότι ένα itemset δεν είναι συχνό, δεν χρειάζεται να δημιουργήσουμε κανένα υπερσύνολό του, σαν υποψήφιο, αφού και αυτό αποκλείεται να είναι συχνό. Η έννοια της σημαντικότητας παρουσιάζεται στο Σχήμα 10 και στο Σχήμα 11.

Έστω ότι έχουμε τέσσερα στοιχεία $\{A, B, C, D\}$. Οι γραμμές στο πλέγμα (Σχήμα 10) αναπαριστούν τη σχέση του υποσυνόλου και έτσι η ιδιότητα του frequent itemset καθορίζει ότι οποιοδήποτε σύνολο στο μονοπάτι επάνω από ένα itemset πρέπει να είναι συχνό εάν το αρχικό itemset είναι συχνό. Στο Σχήμα 11 τα μη κενά υποσύνολα του $\{ACD\}$ φαίνεται να είναι τα $\{AC, AD, CD, A, C, D\}$. Εάν το $\{ACD\}$ είναι συχνό, το ίδιο ισχύει και για τα άλλα υποσύνολά του. Εάν κάποιο από αυτά τα υποσύνολα δεν είναι συχνό, τότε ούτε το $\{ACD\}$ είναι συχνό.

Η βασική ιδέα του αλγορίθμου είναι η δημιουργία των υποψηφίων itemsets ενός συγκεκριμένου μεγέθους και στη συνέχεια η σάρωση της βάσης δεδομένων για να υπολογίσουμε αν αυτά είναι συχνά. Κατά τη διάρκεια του i περάσματος καταμετρούνται τα υποψήφια itemsets μεγέθους i , C_i . Μόνο εκείνοι οι υποψήφιοι που είναι συχνοί χρησιμοποιούνται για τη δημιουργία υποψηφίων για το επόμενο πέρασμα. Αυτό σημαίνει ότι το L_i χρησιμοποιείται για τη δημιουργία του C_{i+1} . Ένα itemset θεωρείται σαν υποψήφιο μόνο όταν όλα του τα υποσύνολα είναι επίσης συχνά. Για τη δημιουργία υποψηφίων μεγέθους $i+1$ γίνονται συνενώσεις συχνών itemsets που βρίσκονται στο προηγούμενο πέρασμα.

Σχήμα 10. Δικτυωτό πλέγμα itemsets για το $\{A, B, C, D\}$.

Σχήμα 11. Υποσύνολα του {ACD}.

Ο Apriori εφαρμόζεται στο αρχείο των συναλλαγών που έχουν προκύψει από την προεπεξεργασία των δεδομένων, προκειμένου να ανακαλυφθούν ένα σύνολο κανόνων συσχέτισης που αντιστοιχούν στις συγκεκριμένες συναλλαγές. Ο αλγόριθμος βρίσκει αρχικά ομάδες αντικειμένων που εμφανίζονται συχνά μαζί σε μεγάλο αριθμό συναλλαγών. Τα αντικείμενα στην περίπτωση μας είναι τα URLs που εμφανίζονται στο προ-επεξεργασμένο αρχείο log. Αρχικά πρέπει να δημιουργηθούν τα υποψήφια frequent itemsets για κάθε πέρασμα μετά το πρώτο. Όλα τα items ενός itemset χρησιμοποιούνται σαν υποψήφια για το πρώτο πέρασμα. Εδώ, το σύνολο των συχνών itemsets του προηγούμενου περάσματος L_{i-1} , συνενώνεται με τον εαυτό του για να καθορίσει τους υποψηφίους. Μεμονωμένα itemsets πρέπει να έχουν κοινά όλα τα στοιχεία εκτός από ένα ώστε να συνδυαστούν. Μετά το πρώτο πέρασμα κάθε συχνό itemset συνδυάζεται με όλα τα άλλα συχνά itemsets. Ο Πίνακας 2 παρουσιάζει τον αλγόριθμο **AprioriGen** που θα δημιουργήσει το υπερσύνολο των frequent itemsets μεγέθους i , $C_i \supset L_i$, όταν η είσοδος είναι το L_{i-1} .

Είσοδος

L_{i-1} : large itemsets of size $i-1$

Αλγόριθμος AprioriGen

$C_i = \emptyset$

For each $I \in L_{i-1}$ do

For each $J \neq I \in L_{i-1}$ do

If $i-2$ of the elements in I and J are equal

Then $C_k = C_k \cup \{I \cup J\}$

Endif

EndFor

EndFor

Έξοδος

C_i : candidates of size i

Πίνακας 2. Ο αλγόριθμος που δημιουργεί ένα υπερσύνολο frequent itemsets.

Με δεδομένη την ιδιότητα των frequent itemsets που αναφέρθηκε προηγούμενα και τον Apriori-gen ο αλγόριθμος **Apriori** παρουσιάζεται στη συνέχεια (Πίνακας 3).

Είσοδος

I : itemsets

T : transactions

σ : support

Αλγόριθμος Apriori

$k = 0$ (k is used as the scan number)

$L = \emptyset$

$C_1 = I$ (initial candidates are set to be the items)

Repeat

$k = k + 1$

$L_k = \emptyset$

For each $I_i \in C_k$ do

$C_i = 0$ (initial counts for each itemset are 0)

EndFor

For each $t_j \in T$ do

For each $I_i \in C_k$ do

If $I_i \in t_j$ Then $C_i = C_i + 1$

EndIf

EndFor

EndFor

For each $I_i \in C_k$ do

If $C_i \geq (\sigma * |T|)$ Then $L_k = L_k \cup I_i$

EndIf

EndFor

$L = L \cup L_k$

$C_{k+1} = \text{AprioriGen}(L_k)$

Until $C_{k+1} = \emptyset$

Έξοδος

L : large itemsets

Πίνακας 3. Ο αλγόριθμος Apriori.

Δεδομένου ενός συνόλου $L = \{L_1, L_2, \dots, L_k\}$ από frequent itemsets, το support του L_i ορίζεται σαν:

$$\sigma(L_i) = \frac{|\{t \in T : L_i \subseteq t\}|}{|T|}$$

και αναπαριστά το βαθμό των συναλλαγών στο σύνολο T που έχουν το frequent itemset L_i (Mobasher et al., 2000).

Το **κατώτατο όριο υποστήριξης (minsup)** χρησιμοποιείται από τον αλγόριθμο για την περικοπή του διαστήματος αναζήτησης και γενικά καθορίζεται πριν από το βήμα εξόρυξης. Οι κανόνες συσχέτισης συλλαμβάνουν τις σχέσεις μεταξύ των αντικειμένων βασισμένοι στα πρότυπα επανεμφάνισής τους στις συναλλαγές. Στην περίπτωση των συναλλαγών στο Web, οι κανόνες συσχέτισης συλλαμβάνουν τις σχέσεις μεταξύ των αναφορών URLs βασισμένοι στα πρότυπα πλοήγησης των χρηστών.

Ένας κανόνας συσχέτισης r είναι μια έκφραση της μορφής:

$$X \Rightarrow Y (\sigma_r, \alpha_r)$$

όπου σ_r είναι το support του $X \cup Y$ και α_r είναι το confidence του κανόνα r και δίνεται από τη σχέση:

$$\frac{\sigma(X \cup Y)}{\sigma(X)}$$

Η εμπιστοσύνη του κανόνα r απεικονίζει το λόγο των συναλλαγών στο σύνολο συναλλαγών T , που περιέχουν το X και περιέχουν επίσης και το Y . Στην προσέγγισή μας χρησιμοποιούμε τα frequent itemsets και τους κανόνες συσχέτισης για να παρέχουμε συστάσεις στους εκπαιδευόμενους. Έχοντας βρει όλα τα frequent itemsets στη συνέχεια υπολογίζονται οι κανόνες συσχέτισης. Ο Πίνακας 4 παρουσιάζει τον αλγόριθμο **AssociationRulesGen** (Goethals, 2003).

Είσοδος

T : transactions
 I : items
 L : large itemsets
 σ : support
 α : confidence

Αλγόριθμος AssociationRulesGen

$R = \emptyset$
 For each $l \in L$ do
   For each $x \subset l$ such that $x \neq \emptyset$ do
     If $\frac{\sigma(l)}{\sigma(x)} \geq \alpha$ Then $R = R \cup \{x \Rightarrow (l - x)\}$
   Endif
 EndFor

Έξοδος

R : association rules satisfying σ and α

Πίνακας 4. Ο αλγόριθμος που εξαγει τους κανόνες συσχέτισης.

1.7 Ο αλγόριθμος παραγωγής εξατομικεύσης

Με στόχο την εξασφάλιση της όσο το δυνατόν καλύτερης παραγωγής προτάσεων, συνδυάζουμε την οντολογία περιεχομένου του e-learning web site με τη γνώση που προέρχεται από την πλοήγηση των χρηστών σε αυτό. Συγκεκριμένα, χρησιμοποιούμε τα δεδομένα χρήσης ώστε να εξαγάγουμε χρήσιμα συμπεράσματα σχετικά με τον τρόπο που οι εκπαιδευόμενοι μαθαίνουν ένα εκπαιδευτικό αντικείμενο. Οι προτάσεις που πρόκειται να παρουσιαστούν στο χρήστη εκτός από τα παραπάνω σχετίζονται και με την τρέχουσα θέση του μέσα στο σύστημα. Ο ρόλος της οντολογίας είναι να καθορίσει ποια εκπαιδευτικά υλικά είναι καταλληλότερα να συστηθούν στο χρήστη και σύμφωνα με τα frequent itemsets (πορείες πλοήγησης του χρήστη) ποιες από αυτές τις επιλογές έχουν το μέγιστο support. Η διαδικασία παρουσιάζεται στο Σχήμα 12.

Στην προσέγγισή μας αποφασίσαμε να χρησιμοποιήσουμε μια κοινή οντολογία για να εκφράσουμε το περιεχόμενο του e-learning συστήματος που αντιστοιχεί σε κάθε ένα από τα υπάρχοντα Courses και τα οποία αντιστοιχούν σε υπο-τμήματα της οντολογίας με την κατάλληλη ονοματολογία. Αυτό μας επιτρέπει να συγκρίνουμε εύκολα τη γνώση ενός χρήστη σε σχέση με κάθε ένα από τα εκπαιδευτικά αντικείμενα. Μια άλλη προσέγγιση θα ήταν να χωριστεί η οντολογία για κάθε ένα Course και να κατασκευαστεί η συνολική οντολογία με μέσω κατάλληλης χαρτογράφησης (ontology mapping). Όμως έτσι δεν δίνεται η δυνατότητα να αντιστοιχίσουμε τη γνώση κάθε χρήστη σε σχέση με τα Courses με αποτέλεσμα να είναι λιγότερο ευέλικτη.

Σχήμα 12. Προτεινόμενο σενάριο για την παραγωγή των συστάσεων σε ένα σύστημα e-learning.

Σε ότι αφορά το online τμήμα της μεθοδολογίας μας, το σύστημα παρακολουθεί την ενεργή σύνοδο του κάθε χρήστη, η οποία απεικονίζει τις πιο πρόσφατες επιλογές του. Σύμφωνα με την τρέχουσα κατάστασή του, η μηχανή συστάσεων του προτείνει την επόμενη πιο κατάλληλη σύνδεση. Αυτή η μηχανή δέχεται σαν είσοδο την ενεργό σύνοδο του χρήστη και λαμβάνει υπόψη επίσης την οντολογία της περιοχής και το σύνολο των κανόνων συσχέτισης που προήλθαν από τις συναλλαγές των χρηστών κατά τη διάρκεια της offline φάσης.

Ειδικότερα, βασιζόμαστε στα ακόλουθα για την εύρεση των πιο κατάλληλων συστάσεων που θα παρουσιαστούν στον εκπαιδευόμενο:

- **Η οντολογία των εγγράφων.** Τα έγγραφα έχουν σχολιαστεί με βάση ένα συνδυασμό των metadata standards Dublin Core και LOM.
- **Το αρχείο με τους κανόνες συσχέτισης.** Αυτοί οι κανόνες προκύπτουν από τις συναλλαγές των χρηστών κατά τη διάρκεια του βήματος της προ-επεξεργασίας.

Ο ρόλος της μηχανής είναι να υπολογίσει ένα σύνολο προτάσεων, το οποίο αποτελείται από συνδέσεις προς σελίδες που ο χρήστης μπορεί να επισκεφτεί. Αναπαριστά ουσιαστικά ένα σύνολο από ενδεχομένως χρήσιμους συνδέσμους με βάση τη δραστηριότητα πλοήγησης του χρήστη μέσα στο site. Αυτές οι συνδέσεις προστίθενται στην τελευταία σελίδα της συνόδου που δημιουργήθηκε από το χρήστη προτού αυτή η σελίδα να σταλεί στον φυλλομετρητή του. Χρησιμοποιούμε ένα παράθυρο σταθερού μεγέθους πάνω στην τρέχουσα ενεργό σύνοδο ώστε να συλλάβουμε την τρέχουσα πλοηγητική του συμπεριφορά.

Οι παράγοντες που λαμβάνουμε υπόψη μας στην όλη διαδικασία παραγωγής συστάσεων είναι:

- η οντολογία της περιοχής.
- τα κριτήρια ταιριάσματος με το frequent itemset.
- εάν τα υποψήφια για σύσταση URLs τα έχει επισκεφτεί ο χρήστης στην τρέχουσα ενεργό σύνοδο ή σε παλιότερες επισκέψεις του.
- το γράφο του web site.

Στη συνέχεια, υπολογίζεται το πιθανό σύνολο συστάσεων χρησιμοποιώντας την οντολογία του e-learning web site. Ο στόχος μας είναι να ανακαλύψουμε εκείνο το σύνολο συστάσεων που θα βρίσκεται σε συμφωνία με την οντολογία της περιοχής. Αυτό το σύνολο φιλτράρεται μέσω των frequent itemsets, τα οποία υπολογίστηκαν κατά τη διάρκεια του σταδίου προ-επεξεργασίας. Τα frequent itemsets απεικονίζουν τη γνώση που προέρχεται από τη δραστηριότητα πλοήγησης άλλων χρηστών που ενεργούν με παρόμοιο τρόπο με αυτό του τρέχοντα χρήστη.

Όπως έχουμε ήδη αναφέρει, στους παραπάνω παράγοντες έχουμε συμπεριλάβει το γράφο του web site. Ο γράφος αυτός πρόκειται να χρησιμοποιηθεί για τον υπολογισμό της απόστασης ενός υποψηφίου URL από την τρέχουσα θέση του χρήστη. Σαν απόσταση θεωρούμε τον αριθμό των clicks (clickstream) που ο χρήστης πρέπει να κάνει προκειμένου να πάει από την τρέχουσα θέση του στο συνιστώμενο URL. Ο Πίνακας 5 παρουσιάζει τον αλγόριθμο που χρησιμοποιείται για την παραγωγή των συστάσεων προς τους χρήστες.

Ο αλγόριθμος βρίσκει αρχικά ένα πιθανό σύνολο συστάσεων σύμφωνα με την οντολογία της περιοχής. Αυτό είναι το σύνολο w . Στη συνέχεια ο αλγόριθμος βρίσκει όλα τα frequent itemsets I_i συνδυάζοντας την ενεργό σύνοδο του χρήστη z με τις συστάσεις w που προήλθαν από την οντολογία. Κάθε frequent itemset επικυρώνεται εάν το support του είναι μεγαλύτερο ή ίσο με το ελάχιστο κατώτατο όριο σ . Με αυτό το βήμα ο αλγόριθμος φιλτράρει όλες τις συστάσεις που έχουν χαμηλή υποστήριξη. Εάν το frequent itemset περάσει αυτό το βήμα, τότε ο αλγόριθμος ελέγχει εάν το confidence του αντίστοιχου κανόνα συσχέτισης είναι μεγαλύτερο ή ίσο με το ελάχιστο κατώτατο όριο α . Εάν το frequent itemset έχει υποστήριξη μεγαλύτερη ή ίση με α τότε ακολουθεί το βήμα υπολογισμού (*).

Είσοδος

Active user session z (session window max size n)
Domain ontology
Frequent itemsets and association rules
Minsup threshold σ
Minconf threshold α
Site graph

Αλγόριθμος Recommendations

```
Recommendation set  $r = \emptyset$ .
Potential recommendation set  $w = \emptyset$ .
Potential recommendation set  $w$  according to the ontology
 $w = \{w_1, w_2, \dots, w_k\}$ 
For each set  $z + w_i, w_i \in W$ 
    For each frequent itemset  $I_i = z + w_i$ 
        If  $\text{sup}(I_i) \geq \sigma$  Then  $c = \text{conf}(z \Rightarrow w_i)$ 
            If  $c \geq \alpha$  Then
                 $w_i\_score = c * \text{click\_number} (*)$ 
                 $r = r \cup w_i$ 
            EndIf
        EndIf
    EndFor
EndFor
```

Έξοδος

Μια λίστα r από recommendations

Πίνακας 5. Προτεινόμενος αλγόριθμος για παραγωγή συστάσεων στους χρήστες.

Σε αυτό το στάδιο ο αλγόριθμος υπολογίζει το **score** των προτεινόμενων συνδέσεων w_i . Αυτό το score υπολογίζεται πολλαπλασιάζοντας την εμπιστοσύνη των frequent itemsets με τον αριθμό των clicks που πρέπει να κάνει ο χρήστης προκειμένου να πάει από την τρέχουσα θέση του στο συνιστώμενο URL, σύμφωνα με το γράφο του web site. Κατόπιν τα

συνιστώμενα URL w_i προστίθενται στο σύνολο προτάσεων r . Ας δούμε ένα παράδειγμα που παρουσιάζει το βήμα υπολογισμού (*).

Παράδειγμα

Υποθέτουμε ότι έχουμε δύο συνιστώμενα URLs w_i και w_j με confidence $c_i = 0,3$ και $c_j = 0,6$ και αριθμούς clicks $clicks\_n_i = 6$ και $clicks\_n_j = 3$ αντίστοιχα.

Τότε:

$$w_i\_score = c_i * clicks\_n_i = 1,8$$

$$w_j\_score = c_j * clicks\_n_j = 1,8$$

Από τους παραπάνω υπολογισμούς είναι φανερό ότι τα δύο URLs έχουν το ίδιο score. Αυτό σημαίνει ότι ο αλγόριθμος αντισταθμίζει τις διαφορές μεταξύ της απόστασης και της εμπιστοσύνης. Τα URLs με μεγάλη απόσταση και μικρή εμπιστοσύνη θα είναι στο ίδιο επίπεδο με αυτά που έχουν μικρή απόσταση αλλά μεγάλη εμπιστοσύνη. Συνεπώς, εάν υπάρχουν διάφοροι χρήστες που αποφασίζουν να επισκεφτούν ένα URL που είναι σε μεγάλη απόσταση από την τρέχουσα θέση τους, ο αλγόριθμος θα το βρει και θα δώσει σε αυτό το ίδιο βάρος με άλλα URLs με μεγαλύτερη εμπιστοσύνη και μικρότερη απόσταση. Επομένως, ο αλγόριθμος λαμβάνει ιδιαίτερα υπόψη του τις επιλογές που άλλοι χρήστες έχουν κάνει. Τελικά, ο προτεινόμενος αλγόριθμος βάζει σε μια σειρά τις συστάσεις που προέρχονται από την οντολογία σύμφωνα με το βήμα υπολογισμού (*) και αποκλείει αυτές τις συστάσεις που έχουν χαμηλή υποστήριξη και εμπιστοσύνη. Με αυτό τον τρόπο το αρχικό σύνολο συστάσεων φιλτράρεται και ενισχύεται μέσω των frequent itemsets και των κανόνων συσχέτισης.

Για την αξιολόγηση της προσέγγισής μας χρησιμοποιήσαμε τα δεδομένα 23 εκπαιδευομένων από τα αρχεία logs του πιλοτικού e-learning web site. Τα δεδομένα συλλέχθηκαν στο διάστημα Ιανουάριος-Απρίλιος 2005 και τα χωρίσαμε ανά μήνα. Τα δεδομένα από τους δυο πρώτους μήνες χρησιμοποιήθηκαν για το «κτίσιμο» του μοντέλου μας δηλαδή για να εκπαιδεύσουν το μηχανισμό παραγωγής προτάσεων. Τα δεδομένα των δυο τελευταίων μηνών χρησιμοποιήθηκαν για την αξιολόγηση. Η ιδέα ήταν ότι δεδομένου ενός session S από το μήνα m , αν το σύστημα βασισμένο στα δεδομένα του μήνα $m-1$ και μερικών prefix από το session S , μπορεί να προτείνει ένα σύνολο σελίδων p_i που περιέχουν μερικές από τις σελίδες στο suffix του session S , τότε θεωρείται επιτυχημένη. Επιπλέον, η απόσταση στον αριθμό των clicks ανάμεσα στο suffix του S και στην προτεινόμενη σελίδα p_i θεωρείται κέρδος. Συγκεκριμένα, χρησιμοποιήσαμε τις ακόλουθες μετρικές:

- **Recommendation accuracy** είναι ο λόγος των επιτυχημένων προτάσεων σε σχέση με όλες τις προτάσεις. Οι επιτυχημένες προτάσεις είναι αυτές που εμφανίζονται στο suffix του session για το οποίο το prefix προκαλεί την πρόταση. Σαν παράδειγμα, ας θεωρήσουμε ότι έχουμε S sessions στο αρχείο ελέγχου log. Για κάθε session S , παίρνουμε κάθε σελίδα p και δημιουργούμε τη λίστα $R(p)$. Το $R(p)$ στη συνέχεια συγκρίνεται με το υπόλοιπο τμήμα του S (το suffix του S).

Ορίζουμε αυτό το τμήμα σαν $T(p)$. Η μετρική recommendation accuracy για το δοσμένο session δίνεται από το πόσο συχνά το $T(p)$ και το $R(p)$ τέμνονται. Ο τύπος είναι ο ακόλουθος:

$$RecommendationAccuracy = \frac{\sum_s \frac{|\bigcup_p (T(p) \cap R(p))|}{\bigcup_p R(p)}}{S}.$$

- **Shortcut gain** είναι η μετρική που υπολογίζει πόσα clicks η πρόταση επιτρέπει στους χρήστες να γλιτώνουν αν η πρόταση ακολουθείται. Υποθέτουμε ότι έχουμε ένα session a, b, c, d, e και στη σελίδα b , το σύστημα προτείνει τη σελίδα e . Αν ακολουθήσουμε αυτή την πρόταση μπορούμε να κερδίσουμε δυο βήματα δηλαδή τις σελίδες c και d . Υπάρχει ένα θέμα όταν μετράμε το shortcut gain όταν η λίστα των προτάσεων περιλαμβάνει περισσότερες από μία σελίδες στο suffix του session. Θα πρέπει να θεωρήσουμε το μικρότερο ή το μεγαλύτερο shortcut gain; Για να λύσουμε αυτό το πρόβλημα, διακρίνουμε μεταξύ key pages και auxiliary pages. Μια key page είναι μια σελίδα που περιέχει σχετική πληροφορία με αυτή που ο χρήστης περνάει κάποιο χρόνο. Μια auxiliary page είναι μια ενδιαμέση σελίδα που χρησιμοποιείται σαν σύνδεσμος και στις οποίες ο χρήστης περνάει λίγο χρόνο. Στο πείραμα επιλέξαμε ένα threshold 30 δευτερολέπτων για να δείξουμε αυτή τη διάκριση. Δεδομένου αυτών των δύο τύπων σελίδων, το shortcut gain υπολογίζεται σαν να είναι το μικρότερο πηδύμα προς την key page που έχει προταθεί. Εάν μια key page δεν προτείνεται τότε είναι το μεγαλύτερο πηδύμα προς μια auxiliary page. Το σύνολο των σελίδων στο session που διερχόμαστε με τη βοήθεια του συστήματος ορίζεται σαν το μικρότερο session s' . Για το σύνολο των sessions S στο test log τότε το shortcut gain δίνεται από τον τύπο:

$$ShortcutGain = \frac{\sum_s \frac{|s| - |s'|}{|s|}}{S}.$$

Τα αποτελέσματα της αξιολόγησης του πιλοτικού e-learning συστήματος έδειξαν ότι η μετρική recommendation accuracy μπορεί να φτάσει ακόμα και το ποσοστό του 85%. Ένα άλλο ενδιαφέρον σημείο είναι ότι η recommendation accuracy είναι αντιστρόφως ανάλογη της μετρικής shortcut gain. Στην περίπτωση που η πρώτη φτάνει σε ποσοστό 70% τότε η δεύτερη είναι στο 33%, ενώ όταν η δεύτερη ανεβαίνει σε ποσοστό 60% τότε η πρώτη πέφτει στο 50%. Κάποια web sites (συνήθως αυτά με μεγάλη συχνότητα συνδέσεων) ενδιαφέρονται για υψηλή recommendation accuracy, ενώ κάποια άλλα για υψηλό shortcut gain. Γενικά το πώς οι δύο μετρικές θα ισορροπήσουν ώστε να επιτευχθούν τα καλύτερα αποτελέσματα παραμένει ένα ανοικτό θέμα.

The screenshot shows the homepage of a 'Web Mining' course. The header features a molecular structure graphic on the left, the title 'Web Mining' in the center, and 'Distance Learning Courses' on the right. A navigation bar includes links for 'Login', 'MyPage', 'Help', and 'Glossary'. The main content area is titled 'Welcome to the Web Mining Course' and contains a detailed introduction to web mining, its applications, and course objectives. A sidebar on the left lists various course sections under 'General Information', 'Theory', 'Practice', and 'Services'. A diagram on the right illustrates three interconnected network patterns labeled 'Pattern #1', 'Pattern #2', and 'Pattern #3'. The footer includes a secondary navigation bar with 'Home', 'Announcements', 'Contact', 'Help', 'Glossary', and 'Login', along with a copyright notice for 2005.

Web Mining

Distance Learning Courses

Login | MyPage | Help | Glossary

Welcome to the Web Mining Course

Web mining refers to the discovery of interesting and useful knowledge from the data associated with the usage, content, and the structure of Web resources. It has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in e-commerce, e-CRM, Web analytics, information retrieval/filtering, and Web information systems.

The primary focus of this course is on **Web usage mining** and its applications to e-learning, e-commerce and business intelligence. Specifically, we will consider techniques from machine learning, data mining, text mining, and databases to extract useful knowledge from Web data which could be used for business intelligence, site management, personalization, and user profiling.

The course will also provide a brief overview of other areas in Web mining, such as Web content mining and Web structure mining. The first half of the course will be focused on a detailed overview the data mining process and techniques, specifically those that are most relevant to Web mining. The second half will concentrate on the applications of these techniques to Web and e-commerce data, and their use in Web analytics, user profiling and personalization.

- If you are a registered user [click here](#) .
- If you are not yet a registered user, apply for membership [now](#) .

Home Announcements Contact Help Glossary Login

Copyright © 2005 All rights reserved.

General Information

- Instructor
- Objectives
- Prerequisites
- Grading Policy
- List of Topics

Theory

- Contents
- Introduction
- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Practice

- Examples
- Questions
- Tests

Services

- MyPage
- Announcements
- Chat Room
- Forums
- Shared Folders
- Newsletter
- Diary
- Links

Pattern #1
Pattern #2
Pattern #3

Σχήμα 13. Η αρχική σελίδα του πιλοτικού e-learning συστήματος.

Web Mining

Distance
Learning
Courses

[Login](#) | [MyPage](#) | [Help](#) | [Glossary](#)

General Information

- Instructor
- Objectives
- Prerequisites
- Grading Policy
- List of Topics

Theory

- Contents
- Introduction
- Web data**
- The three axes**
- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Practice

- Examples
- Questions
- Tests

Services

- MyPage
- Announcements
- Chat Room
- Forums
- Shared Folders
- Newsletter
- Diary
- Links

Theory : Introduction : The three axes user: *Penelope Markellou*

The three axes of Web Mining

The Web mining research is a converging area from several research communities, such as databases, IR, machine learning and NLP. Even though it is strongly related to data mining, it is not equivalent to it. Web mining processes Web data, which come in various categories and formats and this is what makes the combination of many techniques applied to the research areas mentioned before, essential.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. In the literature, three main axes of Web mining have been identified, according to the Web data used as input in the data mining process, namely Web structure, Web content and Web usage mining.

The goal of **Web structure mining** is to categorize the Web pages and generate information such as the similarity and relationship between them, taking advantage of their hyperlink topology. In the latter years, the area of Web structure mining focuses on the identification of authorities, i.e. pages that are considered as important sources of information from many people in the Web community.

Web content mining has to do with the retrieval of information (content) available on the Web into more structured forms as well as its indexing for easy tracking information locations. Web content may be unstructured (plain text), semi-structured (HTML documents), or structured (extracted from databases into dynamic Web pages). Such dynamic data cannot be indexed and consist what is called "the hidden Web". A research area closely related to content mining is text mining.

Web usage mining is the process of identifying browsing patterns by analyzing the user's navigational behaviour. This information takes as input the usage data, i.e. the data residing in the Web server logs, recording the visits of the users to a Web site. Extensive research in the area of Web usage mining led to the appearance of a related research area, that of Web personalization. Web personalization utilizes the results produced after performing Web usage mining, in order to dynamically provide recommendations to each user.

Even though there exist three distinct axes in the Web mining research, the distinctions between them are not clear-cut. Web content mining might utilize the text contained in the links, or even the text surrounding them. On the other hand, in Web structure mining the same information can be used for different purposes. Moreover, web usage mining can be enhanced if the content semantics are taken into consideration.

Most of the research efforts nowadays, propose systems or algorithms that combine methods from these Web mining categories. Therefore, Web mining moves to a more abstract level, where data representation is achieved using semantics. These semantics are defined using tools that emerged along with the Semantic Web vision, such as XML, RDF, and most importantly, ontologies.

You can study next:

- [Web content mining](#) > [Searching the Web](#)
- [Web structure mining](#) > [The Web is a Graph](#)
- [Web usage mining](#) > [Identifying navigational patterns](#)

[Back](#)
 [Next](#)

[Home](#)
[Announcements](#)
[Contact](#)
[Help](#)
[Glossary](#)
[Login](#)

Copyright © 2005 All rights reserved.

Σχήμα 14. Παραγωγή προτάσεων για τον εκπαιδευόμενο.

Σχήμα 15. Η υπηρεσία MyPage.

1.8 Συμπεράσματα και ανοικτά θέματα

Στο κεφάλαιο αυτό μελετήθηκε το πρόβλημα της εξόρυξης γνώσης από μια εφαρμογή e-learning για την παραγωγή εξατομικεύσης. Συγκεκριμένα, ερευνήθηκε το επιστημονικό πεδίο του Web Mining καθώς και η μετάβαση προς το Semantic Web Mining και παρουσιάστηκε το πλαίσιο για την ολοκλήρωση οντολογικών δεδομένων σε μια εφαρμογή e-learning. Στη συνέχεια προτάθηκε και αξιολογήθηκε ένα μοντέλο που βοηθάει τους χρήστες στην πλοήγησή τους μέσα στο πιλοτικό e-learning σύστημα. Τα πειραματικά αποτελέσματα έδειξαν πώς αυτή η προσέγγιση που συνδυάζει την οντολογία της περιοχής και τα frequent itemsets μπορεί να παράσχει καλύτερα πρότυπα και ποιοτικότερες υπηρεσίες εξατομικεύσης. Κάποια ενδιαφέροντα ερευνητικά ανοικτά προβλήματα για μελλοντική εργασία παρουσιάζονται στη συνέχεια.

- **Δημιουργία οντολογίας.** Η δημιουργία της οντολογίας δεν είναι μια τετριμμένη ή απλή διαδικασία. Απαιτεί τη συμμετοχή ειδικών, που έχουν αφαιρετική σκέψη και βαθιά γνώση της περιοχής που περιγράφεται, προκειμένου να παραχθεί μια κοινή

και λογική κατανόηση αυτής. Η επιλογή και ο καθορισμός των εννοιών και η σχέση μεταξύ τους, καθώς επίσης και το επίπεδο λεπτομέρειας είναι μια δαπανηρή και χρονοβόρα εργασία, δεδομένου ότι όσο αναλυτικότερη είναι η οντολογία τόσο μεγαλύτερη είναι η πολυπλοκότητα και η δυσκολία που επιβάλλει (Wilson, 2004). Συχνά συμβαίνει οι οντολογικές αποφάσεις που ενσωματώνονται στα αποθηκευμένα δεδομένα να μην αντιστοιχούν σε εκείνες του χρήστη. Αυτό συμβαίνει σε πολλά περιβάλλοντα e-learning, όπου οι διαφορές των εννοιών μπορούν συχνά να μην είναι ευδιάκριτες. Μια λύση θα μπορούσε να είναι η χρήση πολλαπλών οντολογιών, μια για κάθε προφίλ χρήστη, αντί μιας ενιαίας καθολικής οντολογίας (Markellou et al., 2006).

- **Ιδιωτικότητα εκπαιδευομένων.** Τα προσαρμοστικά συστήματα και η καταγραφή των προφίλ των χρηστών φέρνουν στην επιφάνεια τον κίνδυνο της παραβίασης της ιδιωτικότητάς τους που παραμένει ένα άλυτο πρόβλημα (Markellos et al., 2004b). Η έρευνα στον τομέα του Web Usage Mining συνοδεύεται από μεθόδους ασφάλειας για την αύξηση της εμπιστοσύνης των χρηστών που αλληλεπιδρούν με την online εφαρμογή. Η ανάγκη για ιδιωτικότητα έχει δημιουργήσει μια νέα αγορά αφιερωμένη στο σχεδιασμό και στην ανάπτυξη προϊόντων για την ιδιωτική προστασία των πληροφοριών. Οι μελλοντικές προκλήσεις και η έρευνα στην κατεύθυνση της παράδοσης προσαρμοστικών Web εμπειριών χωρίς τη διακινδύνευση (αλλά στην πραγματικότητα προστατεύοντας) την ιδιωτικότητα μπορούν να συνοψιστούν στο εξής (Kobsa, 2001): P3P (Platform for Privacy Preferences Project) υποστήριξη, νοητή κοινοποίηση των δεδομένων, κοινοποίηση των μεθόδων, παροχή οργανωτικών και τεχνικών μέσων για τους χρήστες για να τροποποιήσουν τις εισαγωγές των μοντέλων των χρηστών, κεντρικοί εξυπηρετητές μοντέλων χρηστών που υποστηρίζουν διάφορες μεθόδους ανωνυμίας και προσαρμογή των μεθόδων διαμόρφωσης χρηστών στις προτιμήσεις και τη νομοθεσία ιδιωτικότητας.
- **Χάσιμο του ελέγχου.** Παρά τις προβλεπόμενες προοπτικές της εξατομίκευσης και ενώ τεχνολογικά η περιοχή είναι ώριμη και ήδη πολλές εφαρμογές στο Web (κυρίως ηλεκτρονικής μάθησης και ηλεκτρονικού εμπορίου) χρησιμοποιούν τέτοιες δυνατότητες, υπάρχει ένα σύνολο θεμάτων που εκκρεμούν και απασχολούν τους επιστήμονες του χώρου. Πιο συγκεκριμένα, ένα πρώτο βασικό ζήτημα προκύπτει από το γεγονός ότι η εξατομίκευση στηρίζεται στην αυτόματη εφαρμογή ενός συνόλου προσαρμογών στις ιστοσελίδες μιας εφαρμογής Web πριν αυτές αποσταλούν στους χρήστες ώστε να ικανοποιούν τις ατομικές τους ανάγκες, προτιμήσεις και δεξιότητες. Αυτό όμως δίνει στο σύστημα αυξημένο βαθμό ελευθερίας και περιθώρια πρωτοβουλίας και απομακρύνει τον έλεγχο από το χρήστη, γεγονός που έρχεται σε αντίθεση με μια από τις βασικές αρχές της αλληλεπίδρασης ανθρώπου-υπολογιστή που υπαγορεύει ότι ο χρήστης πρέπει ανά πάσα στιγμή να νιώθει ότι έχει τον έλεγχο της εφαρμογής και ότι η εφαρμογή συμπεριφέρεται με τρόπο προβλέψιμο και ελεγχόμενο (Markellou et al., 2005b).
- **Έλλειψη δεδομένων.** Ένα άλλο θέμα που πρέπει να επιλυθεί αφορά στα περιορισμένα στοιχεία που είναι διαθέσιμα για τους χρήστες και προκύπτει από το ότι οι χρήστες του Web είναι αρνητικοί στο να αποκαλύψουν προσωπικά στοιχεία καθώς ανησυχούν για το κατά πόσο κάτι τέτοιο είναι ασφαλές ή αν παραβιάζεται το δικαίωμά τους στο ιδιωτικό απόρρητο. Η επιφυλακτικότητα αυτή είναι δικαιολογημένη λόγω του ότι τα συστήματα που στηρίζουν την εξατομίκευση με

τεχνικές Web Mining καταγράφουν αναλυτικά όλες τις ενέργειες των χρηστών, πολλές φορές χωρίς να τους ενημερώνουν ή να τους ζητούν την άδεια.

- **Απόκριση σε πραγματικό χρόνο.** Τέλος, σημαντική παράμετρος που θα πρέπει να λαμβάνεται υπόψη κατά το σχεδιασμό και την υλοποίηση τεχνικών εξατομίκευσης σε εφαρμογές Web είναι το ότι η απόκρισή τους θα πρέπει να είναι σε πραγματικό χρόνο. Σε πολλές περιπτώσεις το trade-off που μπορεί να περιορίσει αυτή την απαίτηση, είναι το μέρος της διαδικασίας που σχετίζεται με τον καθαρισμό των δεδομένων και την ανάλυσή τους να πραγματοποιείται offline σε τακτά χρονικά διαστήματα και το σύστημα να ενσωματώνει τα νέα στοιχεία στη διαδικασία λήψης αποφάσεων για προσαρμογές σε περιόδους μειωμένου φόρτου.

1.9 Βιβλιογραφία

Agrawal, R., & Sricant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 487-499.

Agrawal, R., Imielinski, & T., Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 207-216.

Anderson, T., & Whitelock, D. (2004). The Educational Semantic Web: Visioning and Practicing the Future of Education. *Journal of Interactive Media in Education (JIME), Special Issue, The Educational Semantic Web*, 1-15.

Ardissono, L., & Torasso, P. (2000). Dynamic User Modeling in a Web Store Shell. *Proceedings of the 14th Conference ECAI, Berlin*, Germany, 621-625.

Berners-Lee, T. (2000). What the Semantic Web Can Represent. <http://www.w3.org/DesignIssues/RDFnot.html>.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5): 34-43.

Blom, J. (2000). Personalization - A Taxonomy. *Proceedings of the CHI 2000 Workshop on Designing Interactive Systems for 1-to-1 Ecommerce*. New York, ACM Press, 313-314.

Brase, J., & Nejd, W. (2004). Ontologies and Metadata for eLearning. *Handbook on Ontologies*, Staab, S., & Studer, R. (Eds.), Springer-Verlag, 555-573.

Brusilovsky, P. (1996). Methods and Techniques of Adaptive Hypermedia. *User Modeling and User Adapted Interaction*, 6(2/3): 87-129.

Brusilovsky, P. (1998). Adaptive Educational Systems on the World-Wide-Web: A Review of Available Technologies. *Proceedings of Workshop WWW-Based Tutoring at Forth International Conference on Intelligent Tutoring Systems (ITS'98)*, San Antonio, TX.

- Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11: 87-110.
- Brusilovsky, P. (2003). Developing Adaptive Educational Hypermedia Systems: From Design Models to Authoring Tools. *Authoring Tools for Advanced Technology Learning Environment*, Murray, T., Blessing, S., & Ainsworth, S. (Eds.), Dordrecht: Kluwer Academic Publishers.
- Brusilovsky, P., Eklund, J., Schwarz, E. (1998). Web-based Education for All: a Tool for Developing Adaptive Courseware. *Computer Networks and ISDN Systems*, 30(1-7): 291-300.
- Chakrabarti, S. (2000). Data Mining for Hypertext: A Tutorial Survey. *SIGKDD Explorations*, 1(2):1-11.
- Chakrabarti, S. (2003). *Mining the Web: Discovering Knowledge from Hypertext Data*. Elsevier Science.
- Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Mining the Link Structure of the World Wide Web. *IEEE Computer*, 32(8): 60-67.
- Chang, G., Healey, M.J., McHugh, J.A.M., & Wang, J.T.L. (2001). *Mining the World Wide Web: An Information Search Approach*. Boston: Kluwer Academic Publishers.
- Chen, M., Park, J., & Yu, P. (1996). Data Mining for Path Traversal Patterns in a Web Environment. *Proceedings of the 16th International Conference on Distributed Computing Systems*, 385-392.
- Chen, M., Park, J., & Yu, P. (1998). Efficient Data Mining for Path Traversal Patterns. *IEEE Transactions Knowledge and Data Engineering*, 10(2): 209-221.
- Cooley, R. (2000). Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. *PhD Thesis*, University of Minnesota, Faculty of the Graduate School.
- Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, Los Alamitos, CA, USA, IEEE Press, 474-477.
- Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *Journal of Knowledge and Information Systems*, 1(1): 5-32.
- Cristea, A.I. (2004). What can the Semantic Web do for Adaptive Educational Hypermedia? *Educational Technology & Society*, 7(4): 40-58.
- Dai, H., & Mobasher, B. (2004). Integrating Semantic Knowledge with Web Usage Mining for Personalization. *Web Mining: Applications and Techniques*, A. Scime (Ed.), Hershey: Idea Group Publishing, 276-306.

- Davies, J., Fensel, D., & van Harmelen, F. (2004). Towards the Semantic Web – Ontology Driven Knowledge Management. John Wiley & Sons. ISBN: 0470848677, 1-9.
- De Bra, P., & Calvi, L. (1998). AHA! An Open Adaptive Hypermedia Architecture. *The New Review of Hypermedia and Multimedia*, 4: 115-139.
- De Bra, P., Brusilovsky, P., & Houben, G.J. (1999). Adaptive Hypermedia: from Systems to Framework. *ACM Computing Surveys*, 31(4es).
- Dietinger, Th. (2003). Aspects of E-learning Environments. *PhD Thesis*, Institute for Information Processing and Computer Supported New Media (IICM), Graz University of Technology, Austria.
- Ding, Y. (2001). Ontology: The Enabler for the Semantic Web. *Journal of Information Science*, 27(6).
- Ding, Y., Fensel, D., Klein, M., & Omelayenko, B. (2002). The Semantic Web: Yet Another Hip?. *Data and Knowledge Engineering*, 41(3): 205-227.
- Dolog, P., & Nejdl, W. (2003). Personalisation in Elena: How to Cope with Personalisation in Distributed eLearning Networks. *Proceedings of the International Conference on Worldwide Coherent Workforce, Satisfied Users - New Services For Scientific Information*, Oldenburg, Germany, September.
- Dolog, P., Henze, N., Nejdl, W., & Sintek, M., (2004). Personalization in Distributed e-Learning Environments. *Proceedings of the 13th International World Wide Web Conference*, NY, USA, 170-179.
- Drucker, P. (2000). Need to Know: Integrating e-Learning with High Velocity Value Chains. *A Delphi Group White*. Retrieved from: <http://www.delphigroup.com/pubs/whitepapers/20001213-e-learning-wp.pdf>.
- Edelstein, H.A. (2001). Pan for Gold in the Clickstream. *Informationweek*. Retrieved from: <http://www.informationweek.com/828/mining.htm>.
- Eirinaki, M., & Vazirgiannis, M. (2003). Web Mining for Web Personalization. *ACM Transactions on Internet Technology (TOIT)*, ACM Press New York, 3(1): 1-27.
- Erdmann, M., & Studer, R. (2000). How to Structure and Access XML Documents with Ontologies. *Data and Knowledge Engineering, Special Issue on Intelligent Information Integration DKE*, 36(3): 317-335.
- Etzioni, O. (1996). The World Wide Web: Quagmire or Gold Mine. *Communications of the ACM*, 39(11): 65-88.
- Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D.L., & Patel-Schneider, P.F. (2001). OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2):38-44.

Fink, J., & Kobsa, A. (2000). A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web. *User Modeling and User-Adapted Interaction*, 10(3-4): 209-249.

Fink, J., Kobsa, A., & Nill, A. (1996). User-oriented Adaptivity and Adaptability in the AVANTI Project. *Proceedings of the Conference Designing for the Web: Empirical Studies*, Microsoft Usability Group, Redmond, WA, 1996. Retrieved from: <http://fit.gmd.de/hci/projects/avanti/publications/ms96.html>.

Fry, K. (2001). E-learning Markets and Providers: Some Issues and Prospects. *Education and Training, Emerland*, 43(4): 233-239.

Goethals, B. (2003). Survey on Frequent Pattern Mining. HIIT Basic Research Unit Department of Computer Science University of Helsinki, Finland. Retrieved from: <http://www.adrem.ua.ac.be/~goethals/software/index.html>.

Gomez-Perez, A., Fernandez-Lopez, M., & Corcho, O. (2004). Theoretical Foundations of Ontologies (Chapter 1). *Ontological Engineering (with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web)*, Springer-Verlang, 1-45.

Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199-220.

Guarino, N. (1998). Formal Ontology in Information Systems. *Proceedings of the 1st International Conference on Formal Ontology in Information Systems (FOIS'98)*, Trento, Italy, IOS press, Amsterdam, 3-15.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco: Morgan-Kaufmann Publishers.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Handschuh, S., Staab, S., & Mädche, A. (2001). CREAM – Creating Relational Metadata with a Component-based, Ontology-driven Annotation Framework. *Proceedings of the 1st International Conference on Knowledge Capture, K-CAP*, October, Vancouver.

Hatem, M.S., Ramadan, H.A., & Neagu, D.C. (2005). E-learning Based on Context Oriented Semantic Web. *Journal of Computer Science*, 1(4): 500-504.

Henze, N., Dolog, P., & Nejdl, W. (2004). Reasoning and Ontologies for Personalized e-learning in the Semantic Web. *Educational Technology and Society*, 7(4): 82-97.

Horrocks, I. (2002). DAML+OIL: A Description Logic for the Semantic Web. *IEEE Data Engineering Bulletin*, 25(1): 4-9.

Internet World Stats (2005). Internet Usage Statistics - The Big Picture. *World Internet Users and Population Stats*. Retrieved from: <http://www.internetworldstats.com/stats.htm>.

Kato, H., Nakayama, T., & Yamane, Y. (2000). Navigation Analysis Tool Based on the Correlation Between Contents Distribution and Access Patterns. *Working Notes of the*

Workshop on Web Mining for E-Commerce - Challenges and Opportunities (WebKDD 2000) at the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 95-104.

Kleinberg, J.M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604-632.

Kobsa, A. (2001). Tailoring Privacy to Users' Needs. Invited Keynote in Bauer, M., Gmytrasiewicz, P.J., & Vassileva, J. (Eds.), *User Modeling 2001, 8th International Conference*, Berlin-Heidelberg, Springer Verlag, 303-313.

Kohonen, T. (1997). *Self-Organizing Maps*. New York: Springer-Verlag.

Kosala, R., & Blockeel, H. (2000). Web Mining Research: A Survey. *SIGKDD Explorations*, 2(1): 1-15.

Lassila, O., & McGuinness, D. (2001). The Role of Frame-based Representation on the Semantic Web. Technical Report KSL-01-02, Knowledge Systems Laboratory, Stanford University, California.

Lin, W., Alvarez, S.A., & Ruiz, C. (2002). Efficient Adaptive-support Association Rule Mining for Recommender Systems. *Data Mining and Knowledge Discovery*, 6: 83-105.

Lui, A., & Li, S.C. (2002). Patterns of Web Based Learning in the Semantic Web Era. *Proceedings of the 1st International Conference on Web-based Learning in China (ICWL)*, Hong Kong, 17-19 August, 135-146.

Luotonen, A. (1995). The Common Logfile Format. Logging Control in W3C. Retrieved from: <http://www.w3.org/pub/WWW/Daemon/User/Config/Logging.html>.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Le Cam, L.M., & Neyman, J. (Eds.), University of California Press, 1: 281-297.

Madria, S.K., Bhowmick, S.S., Ng, W.K., & Lim, E.P. (1999). Research Issues in Web Data Mining. *Proceedings of Data Waterhousing and Knowledge Discovery, First International Conference, DaWaK '99*, 303-312.

Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004a). Web Personalization for Enhancing E-learning Experience. *Proceedings of the 5th International Conference on Information Communication Technologies in Education (ICICTE)*. Samos Island, Greece.

Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004b). Web Personalization and the Privacy Concern. *Proceedings of the 7th ETHICOMP International Conf. on the Social and Ethical Impacts of Information and Communication Technologies, Challenges for the Citizen of the Information Society*, Syros, Greece.

- Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005a). Using Semantic Web Mining Technologies for Personalized e-Learning Experiences. *Proceedings of the 4th IASTED International Conference on Web-Based Education (WBE-2005)*, Grindelward, Switzerland, 21-23 February, 522-527.
- Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2006). Semantic Web Mining for E-Learning Personalization. *International Journal of Knowledge and Learning (IJKL)*, Published by Inderscience Publishers, to appear.
- Markellou, P., Rigou, M., & Sirmakessis, S. (2004a). Mining for Web Personalization. *Web Mining: Applications and Techniques*, A. Scime (Ed.), Hershey: Idea Group Publishing, 27-48.
- Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004b). Personalization in the Semantic Web Era: a Glance Ahead. *Proceedings of the 5th International Conference on Data Mining, Text Mining and their Business Applications, Data Mining*, A. Zanasi, N.F. Ebecken & C.A. Brebbia (Eds.), Wessex Institute of Technology (UK), Malaga, Spain, Southampton, Boston: PWIT Press, 3-11.
- Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2005b). Shaping Online Learning Communities and the Way Adaptiveness Adds to the Picture. *International Journal of Knowledge and Learning (IJKL)*, Inderscience Publishers, 1(1/2): 80-95.
- McGuinness, D.L., & van Harmelen, F. (2004). OWL Web Ontology Language Overview, W3C Recommendation 10 February. Retrieved from: <http://www.w3.org/TR/owl-features/>.
- Mizoguchi, R., Vanwikenhuysen, J., & Ikeda, M. (1995). Tasks Ontology for Reuse of Problem Solving Knowledge. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing (KBKS'95)*, Mars, N. (Ed.), University of Twente, Enschede, Netherlands, IOS Press, Amsterdam, Netherlands, 46-57.
- Mobasher, B. (2005). Web Usage Mining and Personalization. *Practical Handbook of Internet Computing*, Singh, M. (Ed.), CRC Press.
- Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic Personalization Based on Web Usage Mining. *Communications of the ACM*, 43(8): 142-151.
- Mobasher, B., Dai, H., Luo, T., & Nakagawa M. (2002a). Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery, Kluwer Publishing*, 6(1): 61-82.
- Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002b). Using Sequential and Non-Sequential Patterns in Predictive Web Usage Mining Tasks. *Proceedings of the IEEE International Conference on Data Mining (ICDM02)*, Maebashi City, Japan, December.
- Mobasher, B., Jin, X., & Zhou, Y. (2003). Semantically Enhanced Collaborative Filtering on the Web, *EWMF*, 57-76.

Mödritscher, F., Barrios, V.M.G., & Gütl, C. (2004). The Past, the Present and the Future of Adaptive E-Learning. *Proceedings of the International Conference on Interactive Computer Aided Learning (ICL)*, Villach, Austria.

Neumann, G., & Zirvas, J. (1998). SKILL – A Scallable Internet-based Teaching and Learning System. *Proceedings of WebNet 98, World Conference on WWW, Internet and Intranet*, Orlando, FL, 688-693.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 161-172.

Papanikolaou, K., Grigoriadou, M., Kornilakis, H., & Magoulas, G. (2003). Personalising the Interaction in a Web-based Educational Hypermedia System: the Case of INSPIRE. *User-Modeling and User-Adapted Interaction*, 13(3): 213-267.

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. (2003). Web Usage Mining as a Tool for Personalization: a Survey. *User Modeling and User-Adapted Interaction*, Kluwer Academic Publishers, Netherlands, 13: 311-372.

Rao, R., Pirolli, P., & Pitkow, J. (1996). Silk from a Sow's Ear: Extracting Usable Structures from the Web. *Proceedings of the ACM Conference Human Factors in Computing Systems*, CHI, New York, NY, ACM Press, 118-125.

Rasmussen, E. (1999). Clustering Algorithms. *Information Retrieval: Data Structure and Algorithms*, Chapter 16, Frakes, W.B., & Baeza-Yates, R. (Eds.), Prentice Hall.

Schwarzkopf, E. (2001). An Adaptive Web Site for the UM2001 Conference. *Proceedings of the UM2001 Workshop on Machine Learning for User Modeling*, 77-86.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1-47.

Shahabi, C., Zarkesh, A.M., Abidi, J., & Shah, V. (1997). Knowledge Discovery from User's Web-page Navigation. *Proceedings of the 7th IEEE International Workshop on Research Issues in Data Engineering (RIDE)*, 20-29.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining: Discovery and Application of Usage Patterns from Web Data. *SIGKDD Explorations*, 1(2):12-23.

Stockley, D. (2005). E-learning Definition and Explanation. E-learning, Online Training, Online Learning web site. Retrieved from: <http://derekstockley.com.au/elearning-definition.html>.

Stojanovic, L., Staab, S., & Studer, R. (2001). E-learning Based on the Semantic Web. *Proceedings of the WebNet2001, World Conference on the WWW and Internet*, Orlando, Florida, USA.

Stratakis, M., Christophides, V., Keenoy, K., & Magkanaraki, A. (2003). WP2 Deliverable 2.1: E-Learning Standards. SeLeNe – Self E-Learning Networks, IST Project IST-2001-39045. Retrieved from: <http://www.dcs.bbk.ac.uk/selene/reports/Del21.pdf>.

Tan, P.N., & Kumar, V. (2002). Discovery of Web Robot Sessions Based on Their Navigational Patterns. *Data Mining and Knowledge Discovery*, 6(1): 9-35.

Tane, J., Schmitz, C., & Stumme, G. (2004). Semantic Resource Management for the Web: an E-learning Application. *Proceedings of the 13th International WWW2004 Conference on Alternate Track Papers and Posters*, New York, USA, 1-10.

Urdan, T., & Weggen, C. (2000). Corporate E-learning: Exploring a New Frontier. WR Hambrecht and Co.

Van Haijst, G., Schreiber, A., & Wielinga, B. (1997). Using Explicit Ontologies in KBS Development. *International Journal of Human-Computer Studies*, 45: 183-292.

W3C (2005). World Wide Web Consortium: Semantic Web. Retrieved from: <http://www.w3.org/2001/sw>.

Weber, G., & Kuhl, H. (2001). Developing Adaptive Internet Based Courses with the Authoring System. Retrieved from: <http://www.weibelzahl.de/literatur/weber-ah2001.pdf>.

Weber, G., & Specht, M. (1997). User Modeling and Adaptive Navigation Support in WWW-based Tutoring Systems. *Proceedings of the Sixth International Conference on User Modeling (UM'97)*, Jameson, A., & Tasso, C. (Eds.), 289-300.

Wentling, T., Waight, C., Gallaher, J., La Fleur, J., Wang, C., & Kanfer, A. (2000). E-learning: a Review Literature. *Knowledge and Learning Systems Group*, National Center for Supercomputing Applications, University of Illinois.

Willett, P. (1988). Recent Trends in Hierarchic Document Clustering: A Critical Review. *Information Processing and Management*, 24(5).

Wilson, R. (2004). The Role of Ontologies in Teaching and Learning, *TechWatch*. Retrieved from: http://www.jisc.ac.uk/uploaded_documents/ACF11AA.doc.

Zaiane, O. (2001). Web Usage Mining for a Better Web-based Learning Environment. *Proceedings of Conference on Advanced Technology for Education*, Banff, Alberta, June 27-28, 60-64.

Zaoane, O. (2002). Building a Recommender Agent for e-Learning Systems. *Proceedings of the International Conference on Computers in Education*, Auckland, New Zealand, 55-59.

Zukerman, I., Albrecht, D.W., & Nicholson, A.E. (1999). Predicting Users' Requests on the WWW. *Proceedings of the 7th International Conference on User Modeling*, Banff, Canada, 275-284.

Ευρωπαϊκή Επιτροπή (2003). Γενική Διεύθυνση Εκπαίδευσης και Πολιτισμού. eLearning: Καλύτερη Ηλεκτρονική Μάθηση για την Ευρώπη. Retrieved from: (<http://europa.eu.int>).

Μάρκελλος, Κ., Μαρκέλλου, Π., Ρήγκου, Μ., Συρμακέσης, Σ., & Τσακαλίδης, Α. (2004). Εξατομικευμένη Ηλεκτρονική Μάθηση. *Περιοδικό Επιστήμες της Αγωγής, Ειδική Έκδοση «Δια βίου και Εξ Αποστάσεως Εκπαίδευση στην Κοινωνία της Πληροφορίας»*, Παιδαγωγικό Τμήμα Δ.Ε. Πανεπιστημίου Κρήτης, Σεπτέμβριος.

Συρμακέσης, Σ., Ρήγκου, Μ., Ρήγκου, Ε., Κατσής, Μ., & Τσακαλίδης, Α. (2003). Ηλεκτρονικές Κοινότητες Μάθησης: Αναγκαιότητα, Στόχοι, και Τεχνολογικές Δυνατότητες. *2ο Πανελλήνιο Συνέδριο στην «Ανοικτή και εξ Αποστάσεως Εκπαίδευση»*, Πάτρα, 28-30 Μαρτίου, 698-707.

«The important thing is not to stop questioning.»

Albert Einstein

US (German-born) physicist

(1879 - 1955)

Κεφάλαιο
2**ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ E-COMMERCE**

«Recommendation systems have become paramount to customise information access and reduce information overload. They serve multiple uses, ranging from suggesting products and artifacts, to bringing people together by the connections induced by (similar) reactions to products and services.» (Mirza, 2001).

2.1 Εισαγωγή

Τα ηλεκτρονικά καταστήματα (**electronic shops, e-shops ή online shops**) αποτελούν την κυριότερη εφαρμογή του ηλεκτρονικού εμπορίου (**electronic commerce, e-commerce, e-εμπόριο ή EC**), υλοποιούν το σενάριο **επιχείρηση-προς-καταναλωτή (business-to-consumer ή B2C)** και στοχεύουν στην προώθηση αλλά και στην πώληση προϊόντων και υπηρεσιών μέσω του Παγκοσμίου Ιστού Πληροφοριών (World Wide Web ή Web) (Adam et al., 1999). Είναι γεγονός ότι τα τελευταία χρόνια ένας μεγάλος αριθμός από τέτοια καταστήματα λειτουργούν, ενώ καθημερινά νέα e-shops ανοίγουν τις διαδικτυακές τους πύλες. Άλλωστε, το e-εμπόριο έδωσε τη δυνατότητα στους χρήστες-πελάτες να επιλέξουν προϊόντα και υπηρεσίες μέσα από μια μεγάλη ποικιλία. Υπάρχουν παραδείγματα καταστημάτων τα οποία λειτουργούν μόνο μέσω ηλεκτρονικών καναλιών όπως το βιβλιοπωλείο Amazon (<http://www.amazon.com>), ενώ άλλα συνδυάζουν τη φυσική τους παρουσία με την ηλεκτρονική όπως η εταιρία Microsoft (<http://www.microsoft.com>).

Είναι αναμφισβήτητο γεγονός ότι οι δυνατότητες που προσφέρει το Διαδίκτυο σε μια επιχείρηση είναι πολύ σημαντικές. Οι επιχειρήσεις που θα το κατανοήσουν αυτό νωρίς και θα ενσωματώσουν το “e” στην οργάνωση και τη λειτουργία τους είναι αυτές που θα αποκτήσουν συγκριτικό πλεονέκτημα στην ιδιαίτερα ανταγωνιστική ψηφιακή οικονομία. Μεταξύ των κυριότερων πλεονεκτημάτων περιλαμβάνονται: η παροχή πληροφοριών, προϊόντων και υπηρεσιών 24 ώρες το 24ώρο, 7 ημέρες την εβδομάδα, η παγκόσμια πρόσβαση ανεξαρτήτως γεωγραφικής περιοχής, η άμεση απόκριση, η απουσία χαρτιού, η μείωση του κόστους, οι εξατομικευμένες υπηρεσίες, κλπ. Ένα επιτυχημένο e-shop μπορεί να ενισχύσει τη δημόσια εικόνα της επιχείρησης, να αυτοματοποιήσει εσωτερικές και εξωτερικές διαδικασίες, να μειώσει το κόστος και το χρόνο εξυπηρέτησης και να την τροφοδοτεί συνεχώς με στοιχεία που αφορούν τόσο τους πελάτες όσο και τις συναλλαγές τους καθώς επίσης και με ερωτήσεις, σχόλια, παρατηρήσεις, συστάσεις, κλπ. Όλα αυτά αν

χρησιμοποιηθούν σωστά από την επιχείρηση θα οδηγήσουν στη βελτίωση των παρεχόμενων υπηρεσιών και προϊόντων της.

Από την άλλη πλευρά όμως, αντί για τα μερικές εκατοντάδες βιβλία που ο χρήστης μπορεί να βρει για παράδειγμα σε ένα παραδοσιακό βιβλιοπωλείο, έχει τώρα πρόσβαση σε χιλιάδες βιβλία που διαθέτουν τα online καταστήματα. Με την αύξηση όμως των προσφερόμενων επιλογών, αυξάνει παράλληλα και ο φόρτος που απαιτείται από το χρήστη για να επιλέξει τα προϊόντα που ταιριάζουν στις ανάγκες και στις απαιτήσεις του. Για να αντιμετωπισθεί αυτό το πρόβλημα της **πληθώρας των πληροφοριών** που είναι γνωστό και σαν **information overload problem** (Montebello, 1998) εφαρμόζονται συνήθως τεχνικές **εξατομίκευσης (personalization)** (Kim, 2002). Ένας ορισμός της εξατομίκευσης είναι ο ακόλουθος: *«εξατομίκευση είναι η διαδικασία προσαρμογής των σελίδων ενός web site στις προτιμήσεις και τα ιδιαίτερα χαρακτηριστικά κάθε χρήστη ξεχωριστά. Αποτελεί το μέσο που θα ικανοποιήσει τις ανάγκες ενός πελάτη πιο αποτελεσματικά, θα καταστήσει ευκολότερες και γρηγορότερες τις αλληλεπιδράσεις του και συνεπώς θα αυξήσει την ικανοποίησή του καθώς και την πιθανότητα να επισκεφτεί ξανά το site.»* (DTV, 2005).

Τα **συστήματα συστάσεων (recommendations systems, recommender systems ή RSs)** αποτελούν ίσως την πιο δημοφιλή μορφή εξατομίκευσης και τείνουν να μετατραπούν στις μέρες μας σε σημαντικά επιχειρησιακά εργαλεία (Μαρκέλλου, 2003), (Markellos et al., 2003). Αρχικά, εμφανίστηκαν στα μέσα της δεκαετίας του '90 και χρησιμοποιούνταν από λίγα web sites. Σήμερα δεν νοούνται εφαρμογές e-εμπορίου που να μην ενσωματώνουν τεχνολογίες RSs προκειμένου να βοηθήσουν τους πελάτες στις αγορές τους όπως:

- προτάσεις για βιβλία, CDs και άλλα προϊόντα στο **Amazon.com** (Linden et al., 2003),
- προτάσεις για κινηματογραφικές ταινίες στο **MovieLens** (Miller et al., 2003),
- ειδήσεις στο **VERSIFI Technologies**¹ (Billsus et al., 2002), κλπ.,

ενώ ακόμα και εταιρίες λογισμικού όπως η **Microsoft** ενσωμάτωσαν recommendation χαρακτηριστικά στα προϊόντα τους π.χ. στον Commerce Server (Peddy & Armentrout, 2003). Με αυτό τον τρόπο, τα e-shops μπορούν να βελτιώσουν την απόδοσή τους καλύπτοντας τις ανάγκες και τις προτιμήσεις κάθε πελάτη ξεχωριστά, να αυξήσουν την **ικανοποίησή του (satisfaction)**, να εδραιώσουν την **e-εμπιστοσύνη (e-trust)** και την **e-πίστη του (e-loyalty)** και να καθιερώσουν **ένα-προς-ένα σχέσεις (1-1 relationships)** (Markellou et al., 2005b).

Τα συστήματα αυτά χρησιμοποιούν γνώση που παρέχεται είτε από ειδικούς ή που εξάγεται από την παρακολούθηση της συμπεριφοράς των πελατών για να οδηγήσουν τους τελευταίους στον εντοπισμό των προϊόντων-υπηρεσιών που τους ενδιαφέρουν. Οι **προτάσεις μπορεί να βασίζονται**: στις μεγαλύτερες συνολικές πωλήσεις του site, στα δημογραφικά δεδομένα του πελάτη, στην ανάλυση της προηγούμενης αγοραστικής συμπεριφοράς του σαν μια πρόβλεψη της μελλοντικής συμπεριφοράς του, σε δεδομένα από κοινότητες χρηστών (user communities), κλπ. Οι **τύποι των προτάσεων περιλαμβάνουν**: εξατομικευμένες προσφορές, τιμές, προϊόντα ή υπηρεσίες, προσθήκη ή παράλειψη θεματικών ενότητων ή παραγράφων, ταξινόμηση, κρύψιμο, εμφάνιση, προσθήκη,

¹ Παλιότερα ήταν γνωστό σαν AdaptiveInfo.com.

παράληψη ή έμφαση συνδέσμων, προαιρετικές εξηγήσεις ή αναλυτικές πληροφορίες, σύνοψη απόψεων μιας κοινότητας, παροχή κριτικών μιας κοινότητας, κλπ. (Μαρκέλλου, 2005). Η όλη διαδικασία μπορεί να υλοποιηθεί **χειροκίνητα** π.χ. συνδυασμός προϊόντων (cross-selling) ή **αυτόματα** από τον ηλεκτρονικό υπολογιστή π.χ. συσχέτιση πελατών με παρόμοια αγοραστική συμπεριφορά.

Ένα RS ενισχύει τις πωλήσεις ενός καταστήματος e-εμπορίου με τους ακόλουθους τρόπους: **μετατρέπει τους επισκέπτες σε αγοραστές, αυξάνει το cross-selling και το up-selling και «χτίζει» την εμπιστοσύνη και την πίστη των πελατών**. Σε έναν κόσμο όπου οι ανταγωνιστές είναι μόνο ένα «κλικ» μακριά, το να κερδιθεί η εμπιστοσύνη των πελατών αποτελεί μια σημαντική επιχειρηματική στρατηγική. Τα RSs βελτιώνουν την εμπιστοσύνη αφού δημιουργούν μια ιδιαίτερη σχέση μεταξύ του πελάτη και του site. Τα sites που επενδύουν στο να μάθουν τους πελάτες τους, χρησιμοποιούν αυτά τα συστήματα ώστε να παρουσιάσουν interfaces που ταιριάζουν στις ανάγκες των χρηστών τους. Οι πελάτες από τη μεριά τους ανταποδίδουν αυτές τις ευκολίες με το να ξαναγυρίσουν στα sites που τους ικανοποιούν σε μεγαλύτερο βαθμό. Όσο περισσότερο ο πελάτης χρησιμοποιεί το σύστημα (το οποίο μαθαίνει τι αυτός χρειάζεται), τόσο μεγαλύτερη είναι και η εμπιστοσύνη του σε αυτό το site. Ακόμη, δίνοντας τη δυνατότητα δημιουργίας σχέσεων μεταξύ των πελατών π.χ. μέσω μιας κοινότητας χρηστών με κοινά χαρακτηριστικά και ενδιαφέροντα, αυξάνεται η πιθανότητα οι πελάτες να επιστρέψουν σε αυτό το site αφού το προτείνουν και άλλοι χρήστες με τους οποίους έχουν επικοινωνία (Markellou et al., 2005e).

Γενικά, τα RSs μπορούν να ταξινομηθούν σε τρεις βασικές κατηγορίες: **content based filtering (φιλτράρισμα με βάση το περιεχόμενο)**, **collaborative filtering (συνεργατικό φιλτράρισμα)** και **hybrid approaches (υβριδικές προσεγγίσεις)**. Το content based filtering λαμβάνει υπόψη του τις παραμέτρους του χρήστη με τις πληροφορίες των προϊόντων (μεταδεδομένα, λέξεις κλειδιά, κλπ.), ενώ το collaborative filtering (CF) την «ομοιότητα» των χρηστών προκειμένου να παράγουν συστάσεις. Οι δύο αυτές μέθοδοι μπορούν να συνδυαστούν σε υβριδικές προσεγγίσεις.

2.1.1 Το πρόβλημα και η προτεινόμενη προσέγγιση

Το πρόβλημα που μελετάται στα πλαίσια της διδακτορικής διατριβής είναι αυτό της **παραγωγής προτάσεων σε μια εφαρμογή e-εμπορίου**. Συγκεκριμένα, η προτεινόμενη υβριδική προσέγγιση (Markellou et al., 2005a) στοχεύει στην παραγωγή αποτελεσματικών συστάσεων για τους πελάτες ενός online καταστήματος που νοικιάζει κινηματογραφικές ταινίες², επεκτείνοντας τις παραδοσιακές τεχνικές και ενσωματώνοντας σημασιολογική γνώση για τα αντικείμενα προς σύσταση. Η γνώση για τους πελάτες και τα προϊόντα προκύπτει από **δεδομένα χρήσης (usage data)** και τη **δομή της οντολογίας (ontology structure)** σε συνδυασμό με τις **εκτιμήσεις-βαθμολογίες (ratings) των πελατών για τις ταινίες** καθώς και με την εφαρμογή **τεχνικών ταιριάσματος (matching) «όμοιων» (similar) πελατών**. Αρχικά, το σχήμα της οντολογίας αποτυπώνει το πεδίο του e-shop με τον προσδιορισμό όλων των εννοιών του, των σχέσεων μεταξύ τους καθώς και άλλων ιδιοτήτων, όρων ή κανονισμών. Το πιλοτικό e-shop αποτελείται από μια συλλογή ιστοσελίδων που περιέχουν πληροφορίες για τις ταινίες (τίτλος, κατηγορία, στούντιο, ηθοποιός, σκηνοθέτης, παραγωγός, βραβεία, έτος, διάρκεια, κοινό, περίληψη, κλπ.). Τα

² Η προσέγγιση μπορεί να εφαρμοστεί και σε άλλου τύπου e-shops π.χ. βιβλιοπωλεία, δισκοπωλεία, κλπ.

αρχικά στοιχεία ανακτήθηκαν από τη βάση δεδομένων Internet Movie Database (<http://www.imdb.com>) και στη συνέχεια οργανώθηκαν σε ένα οντολογικό σχήμα προκειμένου να αποτελέσουν ένα πλούσιο σύνολο δεδομένων (superset) το οποίο χρησιμοποιήθηκε στη **διαδικασία ανακάλυψης γνώσης (knowledge discovery process)**. Συγκεκριμένα, αποτέλεσε την είσοδο στη φάση εξόρυξης γνώσης ώστε να συνδυαστούν και να μετασχηματιστούν τα υπονοούμενα/ρητά δεδομένα (κλάση χρηστών, ιστορικό, προφίλ, δομή και περιεχόμενο e-shop) σε νέες μορφές. Το RS λαμβάνει επίσης υπόψη τις εκτιμήσεις-βαθμολογίες των χρηστών για τις ταινίες σε μια κλίμακα από 1 έως 10, όπου 1 σημαίνει μια πολύ κακή ταινία και 10 μια πολύ καλή και παράγει ως έξοδο μια λίστα πιθανών προτάσεων για κάθε χρήστη ξεχωριστά.

Παράλληλα η οντολογία δημιουργεί συνδέσεις μεταξύ των ταινιών σύμφωνα με τα χαρακτηριστικά τους (attributes). Χρησιμοποιώντας το **ιστορικό των συναλλαγών (transactions history file)** του συγκεκριμένου πελάτη, οι προτιμήσεις του μπορούν να προβλεφτούν. Για παράδειγμα, μπορεί να βρεθεί εάν σε ένα πελάτη αρέσει ή όχι μια ταινία μιας συγκεκριμένης κατηγορίας, ενός συγκεκριμένου σκηνοθέτη, ενός συγκεκριμένου ηθοποιού ή συνδυασμών άλλων χαρακτηριστικών. Όταν ένα ή περισσότερα κριτήρια ταιριάσματος ικανοποιούνται, τότε άλλες ταινίες μπορούν να προσδιοριστούν σύμφωνα με το οντολογικό σχήμα που έχουν παρόμοια χαρακτηριστικά με αυτές που ο πελάτης έχει ήδη νοικιάσει. Στην περίπτωση ενός νέου πελάτη όπου το ιστορικό του είναι κενό, οι πληροφορίες από την **αίτηση εγγραφής (registration form)** του στο e-shop αναλύονται ώστε να ταξινομηθεί σε μια συγκεκριμένη **κλάση πελατών (user class)**, ενώ οι προτάσεις παράγονται με βάση το οντολογικό σχήμα. Αυτή η ολοκλήρωση παρέχει πρόσθετη γνώση για τις προτιμήσεις των πελατών και επιτρέπει την παραγωγή πιο επιτυχημένων συστάσεων. Ακόμη και στην περίπτωση του **«cold-start problem»** όπου δεν είναι διαθέσιμη αρχική πληροφορία για τη συμπεριφορά του πελάτη, η προσέγγιση μπορεί να παράγει λογικές και σχετικές συστάσεις. Τα πειραματικά αποτελέσματα έδειξαν ότι ο συνδυασμός τεχνικών από τις περιοχές του **Web Mining** και του **Semantic Web** και η ενσωμάτωσή τους σε ένα υβριδικό RS μπορεί να οδηγήσει σε ποιοτικότερες και πιο επιτυχημένες συστάσεις.

2.1.2 Η δομή του κεφαλαίου

Στο κεφάλαιο αυτό παρουσιάζεται μια νέα προσέγγιση για την παραγωγή recommendations που επεκτείνει τις παραδοσιακές τεχνικές και ενσωματώνει σημασιολογική γνώση για τα αντικείμενα προς σύσταση. Με αυτό τον τρόπο επεκτείνουμε τις δυνατότητες των RSs και βελτιώνουμε την αποτελεσματικότητά τους. Συγκεκριμένα, η δομή του κεφαλαίου είναι η ακόλουθη. Στην επόμενη ενότητα δίνονται τα δομικά στοιχεία ενός RS. Στη συνέχεια μελετώνται οι βασικές κατηγορίες των RSs και καταγράφεται το state-of-the-art της περιοχής. Ακολουθεί η περιγραφή και η αξιολόγηση του προτεινόμενου υβριδικού RS μοντέλου. Το κεφάλαιο ολοκληρώνεται με την παράθεση των σημαντικότερων συμπερασμάτων καθώς και την καταγραφή ανοικτών θεμάτων για μελλοντική έρευνα.

2.2 Τα δομικά στοιχεία ενός RS

Τα βασικά δομικά στοιχεία ενός RS είναι η **είσοδος**, η **έξοδος** και η **μέθοδος παραγωγής των προτάσεων**. Επιπρόσθετα, μπορούμε να θεωρήσουμε τον **τρόπο αποστολής των προτάσεων** και το **βαθμό εξατομίκευσης** (Σχήμα 1). Αυτά τα στοιχεία δεν είναι ανεξάρτητα μεταξύ τους αφού για παράδειγμα συγκεκριμένες σχεδιαστικές επιλογές απαιτούν συγκεκριμένες εισόδους ή συγκεκριμένες έξοδοι μπορούν να παραχθούν μόνο με κάποιες και όχι με όλες τις μεθόδους προτάσεων.

2.2.1 Η είσοδος

Ένα RS δέχεται σαν είσοδο μια συλλογή από δεδομένα που μπορεί να περιλαμβάνουν μεταξύ άλλων τις προτιμήσεις των πελατών, τα χαρακτηριστικά των προϊόντων, τις συσχετίσεις των συναλλαγών, κλπ. ώστε να παράγει εξόδους-προτάσεις. Τα δεδομένα αυτά καλύπτουν ένα μεγάλο εύρος και μπορούν να διακριθούν σε σχέση με την προέλευσή τους σε α) **δεδομένα για τον πελάτη** στον οποίο θέλουμε να προτείνουμε προϊόντα και β) **γενικά δεδομένα από τις κοινότητες ή τη συγκεκριμένη κοινότητα που ανήκει ο πελάτης**.

Σχήμα 16. Δομικά στοιχεία ενός RS.

α) Προσανατολισμένα στον πελάτη δεδομένα εισόδου

Τα δεδομένα αυτά εισάγονται στο RS ώστε να παραχθούν προσανατολισμένες στον πελάτη προτάσεις. Ένα σύστημα που δεν χρησιμοποιεί πληροφορίες για κάθε πελάτη ξεχωριστά, μπορεί να εμφανίσει μόνο γενικές προτάσεις. Οι κυριότεροι τύποι δεδομένων αφορούν:

- **Υπονοούμενη πλοήγηση (implicit navigation).** Τα δεδομένα σε αυτή την περίπτωση εξάγονται από την πλοηγητική συμπεριφορά του πελάτη χωρίς αυτός να είναι ενήμερος για την παρουσία του RS. Μπορεί να περιλαμβάνουν τα

συγκεκριμένα προϊόντα που ο πελάτης βλέπει την τρέχουσα στιγμή ή αυτά που περιέχονται στο καλάθι αγορών του (shopping cart). Επίσης την κατηγορία ή το χαρακτηριστικό στο οποίο ο πελάτης πλοηγείται.

- **Σαφής πλοήγηση (explicit navigation).** Σε αντίθεση με την υπονοούμενη, στη σαφή πλοήγηση οι πελάτες εισάγουν τα δεδομένα με σκοπό να πληροφορήσουν το σύστημα για τις προτιμήσεις τους. Ο πιο απλός τρόπος είναι με τη χρήση ερωτηματολογίων ή φορμών που συγκεντρώνουν δημογραφικά ή άλλα στοιχεία.
- **Λέξεις κλειδιά (keywords) και χαρακτηριστικά των προϊόντων (attributes).** Συχνά τα δεδομένα από τον πελάτη δεν περιορίζονται σε μια απλή κατηγορία ή κάποιο προϊόν που τον ενδιαφέρει. Μπορεί επίσης να χρησιμοποιηθούν λέξεις κλειδιά ή χαρακτηριστικά των προϊόντων είτε με σαφή τρόπο μέσω π.χ. μιας μηχανής αναζήτησης, είτε με υπονοούμενο τρόπο π.χ. από τα προϊόντα που ο πελάτης βλέπει την τρέχουσα στιγμή. Σε κάθε περίπτωση, αυτά τα δεδομένα μεταφράζονται σαν είσοδος που υποδεικνύει τα τρέχοντα ενδιαφέροντα του πελάτη.
- **Βαθμολόγηση (rating).** Οι πελάτες παρέχουν τις πιο χρήσιμες και σαφείς πληροφορίες στο RS όταν αξιολογούν τα προϊόντα που έχουν ήδη αγοράσει. Η αξιολόγηση μπορεί να γίνει με βαθμούς π.χ. «βαθμολόγησε ένα προϊόν σε μια κλίμακα από 1-5» ή με απλή ερώτηση της μορφής «σου άρεσε αυτό;».
- **Ιστορικό συναλλαγών (transactions history).** Το ιστορικό μπορεί να χρησιμοποιηθεί σαν μια μορφή υπονοούμενης βαθμολόγησης. Με βάση αυτό μπορούν να συσχετισθούν προϊόντα και να εξαχθούν συμπεράσματα για τις προτιμήσεις των πελατών.

β) Δεδομένα από τις κοινότητες

Τα δεδομένα αυτά περιλαμβάνουν μια ποικιλία πληροφοριών για τον τρόπο που τα διάφορα μέλη της κοινότητας ή ολόκληρη η κοινότητα αντιλαμβάνεται τα προϊόντα. Δεδομένα εισόδου που αντανakλούν τις απόψεις μιας κοινότητας είναι εκείνα τα **χαρακτηριστικά των προϊόντων (item attributes)** στα οποία αποδίδονται ετικέτες ή κατηγορίες π.χ. το είδος ενός βιβλίου ή μιας ταινίας. Ομοίως, η **δημοτικότητα ενός προϊόντος (item popularity)** μπορεί να εκφράζει τις συνολικές πωλήσεις ή μια λίστα με best-sellers. Με τον ίδιο τρόπο που χρησιμοποιείται το ιστορικό συναλλαγών για να παραχθούν προτάσεις για έναν πελάτη, μπορεί να χρησιμοποιηθεί και το **ιστορικό συναλλαγών μιας κοινότητας (community transactions history)**. Τα παραπάνω μπορούν να συνδυασθούν για να παραχθεί η λίστα των καλύτερων προϊόντων ενός site ή να εξαχθούν συμπεράσματα σχετικά με την κίνηση των πωλήσεων ή τις ομοιότητες/διαφορές των προϊόντων. Παράλληλα, πολλά sites ενθαρρύνουν τα **σχόλια κειμένου (text comments)** από τους πελάτες τους, τα οποία αν και θεωρούνται πολύ χρήσιμα, εντούτοις απαιτούν αρκετό χρόνο επεξεργασίας, αφού κάποιος πρέπει να διαβάσει μια παράγραφο για να καταλάβει αν το σχόλιο είναι θετικό ή αρνητικό. Για να διευκολυνθεί αυτή η διαδικασία τα περισσότερα RSs παρέχουν δυνατότητες **βαθμολόγησης (rating)** των προϊόντων.

2.2.2 Η έξοδος

Η έξοδος ενός RS εξαρτάται από τον τύπο, την ποσότητα και την ποιότητα των δεδομένων που ανέλυσε. Ο πιο κοινός τύπος εξόδου είναι οι **προτάσεις (suggestions)**. Αυτές συνήθως

έχουν τη μορφή «δοκίμασε αυτό» ή «αυτό» στη σελίδα που ο χρήστης πλοηγείται. Ένας δεύτερος τύπος αφορά σε **προβλέψεις (predictions)** για τη βαθμολόγηση που θα έδινε ο πελάτης σε ένα προϊόν. Μπορούν να παρουσιάζονται σαν μια μεμονωμένη πρόταση ή σαν μια λίστα προτάσεων ή να εμφανίζονται στη σελίδα παρουσίασης του προϊόντος. Όταν οι κοινότητες είναι μικρές ή τα μέλη γνωρίζονται μεταξύ τους είναι χρήσιμο να παρουσιάζονται οι **βαθμολογίες (ratings)** συγκεκριμένων μελών ώστε κάθε χρήστης να μπορεί να εξαγάγει τα δικά του συμπεράσματα για τη σημασία της πρότασης. Αυτή η τεχνική είναι χρήσιμη όταν ο πελάτης μπορεί να επιλέξει γνωστά μέλη της κοινότητας ή όταν οι βαθμολογίες συνοδεύονται και από **κριτικές (reviews)**. Οι κριτικές είναι παραδείγματα προτάσεων που περιέχουν εκτιμήσεις τις οποίες δεν μπορεί να κατανοήσει πλήρως ένα υπολογιστικό σύστημα. Πράγματι, είναι δύσκολο να διακριθούν τα θετικά ή τα αρνητικά σχόλια για ένα προϊόν και για αυτό το λόγο συχνά συνοδεύονται και από αριθμητικές βαθμολογίες.

Μερικά αντιπροσωπευτικά παραδείγματα εξόδων από το RS του Amazon.com (<http://www.amazon.com>) είναι τα ακόλουθα:

- **Customers who bought.** Η επιλογή αυτή εμφανίζεται μέσα στη σελίδα παρουσίασης του βιβλίου και περιλαμβάνει δύο λίστες προτάσεων. Η πρώτη λίστα προτείνει βιβλία που αγοράζονται συχνά από τους πελάτες που έχουν αγοράσει το επιλεγμένο βιβλίο. Η δεύτερη λίστα προτείνει συγγραφείς των οποίων τα βιβλία αγοράζονται συχνά από πελάτες που αγοράζουν έργα του συγγραφέα του επιλεγμένου βιβλίου.

Customers who bought this book also bought: • The Annotated Huckleberry Finn by Mark Twain, et al • A Murder, A Mystery, and a Marriage by Mark Twain, et al • Theodore Roosevelt by Louis Auchincloss • Ken Burns's America by Gary R. Edgerton ► Explore similar items	Customers who bought titles by Geoffrey C. Ward also bought titles by these authors: • Bruce Catton • Mark Twain • Phillip Lopate • Katrina Kenison • Jonathan Franzen ► Explore similar authors

- **Your recommendations.** Το Amazon.com ενθαρρύνει την άμεση ανάδραση (feedback) από τους πελάτες για τα βιβλία που έχουν διαβάσει. Οι πελάτες βαθμολογούν τα βιβλία με μια κλίμακα από 1-5 από «δεν μου αρέσει καθόλου» έως «μου άρεσε πάρα πολύ».

[Rate this item](#) to get personal recommendations.
Average Customer Review: ★★★★★ Based on 2 reviews.
[Write a review.](#)

- **Eyes.** Αυτό το χαρακτηριστικό επιτρέπει στους πελάτες να ενημερωθούν μέσω e-mail για τα νέα βιβλία του καταλόγου. Οι πελάτες εισάγουν ερωτήματα με βάση το όνομα του συγγραφέα, τον τίτλο του βιβλίου, το θέμα, την ημερομηνία έκδοσης, κλπ., ενώ μπορούν να χρησιμοποιήσουν απλά ή πολύπλοκα κριτήρια (AND, OR).
- **Amazon.com delivers.** Αποτελεί μια παραλλαγή του Eyes. Οι πελάτες επιλέγουν με ✓ από μια λίστα συγκεκριμένων κατηγοριών ή ειδών (βιογραφίες, ελληνική λογοτεχνία, ξένη λογοτεχνία, αστυνομικά, κλπ.). Περιοδικά οι συντάκτες στέλνουν τις τελευταίες τους προτάσεις με e-mail στους συνδρομητές κάθε κατηγορίας.
- **Bookstore gift ideas.** Αυτό το χαρακτηριστικό επιτρέπει στους πελάτες να λάβουν προτάσεις από τους συντάκτες. Οι πελάτες επιλέγουν μια κατηγορία βιβλίων για την

οποία επιθυμούν να λάβουν προτάσεις. Αν πλοηγηθούν σε αυτή την κατηγορία του «Gift Department» θα μπορέσουν να δουν μια γενική λίστα από προτάσεις που δημιουργήθηκε από τους συντάκτες του Amazon.com. Επίσης, μπορούν να επιλέξουν να δουν προτάσεις από μια από τις ήδη διαμορφωμένες κατηγορίες π.χ. νέοι, επιχειρηματίες, κλπ. Από πολλές απόψεις αυτό το χαρακτηριστικό μοιάζει με την online έκδοση του Amazon.com «delivers» που περιγράφηκε προηγουμένως. Εντούτοις, οι πελάτες μπορούν να λάβουν προτάσεις ανώνυμα αφού δεν υπάρχει ανάγκη να εγγραφούν στο site όπως συμβαίνει στο «delivers».

- **Customer comments.** Αυτό το χαρακτηριστικό επιτρέπει στους πελάτες να λάβουν κείμενα-προτάσεις βασισμένα στις απόψεις άλλων πελατών. Σε κάθε σελίδα πληροφοριών κάθε βιβλίου υπάρχει μια λίστα από 1-5 αστέρια που το βαθμολογούν, καθώς και σχόλια πελατών που το έχουν διαβάσει και έχουν υποβάλλει μια κριτική. Οι πελάτες μπορούν να λάβουν υπόψη τους τις προτάσεις για να αποφασίσουν. Επίσης, μπορούν να βαθμολογήσουν τα σχόλια αφού σε κάθε σχόλιο υπάρχει μια ερώτηση «σε βοήθησε αυτό το σχόλιο;» όπου μπορούν να διαλέξουν «ναι» ή «όχι». Τα αποτελέσματα συνοψίζονται και παρουσιάζονται με τη μορφή «5 από 7 αναγνώστες βρήκαν την ακόλουθη κριτική χρήσιμη».

- **Purchase circles.** Αυτό το χαρακτηριστικό επιτρέπει στους πελάτες να δουν μια λίστα με τα «top-10» για μια συγκεκριμένη γεωγραφική περιοχή, εταιρία, εκπαιδευτικό ίδρυμα, κυβέρνηση ή οργανισμό. Για παράδειγμα, ένας πελάτης μπορεί να ζητήσει να δει ποια βιβλία είναι best-sellers για τους πελάτες της Oracle, του MIT, ή για τους κατοίκους της Ν. Υόρκης. Συνεπώς, οι Purchase circles παρέχουν μια μορφή προτάσεων που επιτρέπει στους πελάτες όχι μόνο να δουν τι διαβάζεται γενικότερα αλλά επίσης και να προσαρμόσουν τις προτάσεις αφού μπορούν να διαλέξουν με ποιους θέλουν να συναναστρέφονται.

2.2.3 Η μέθοδος

Στη συνέχεια παρουσιάζονται οι πιο γνωστές μέθοδοι που χρησιμοποιούν τα RSs. Κάθε μια αντιπροσωπεύει μια οικογένεια αλγορίθμων και τεχνικών, ενώ πρέπει να σημειωθεί ότι είναι δυνατός ο συνδυασμός περισσότερων της μιας μεθόδου για την παραγωγή των προτάσεων.

- **Raw retrieval.** Στη μέθοδο αυτή παρέχεται στους πελάτες ένα interface αναζήτησης μέσω του οποίου μπορούν να κάνουν ερωτήσεις στη βάση δεδομένων των

προϊόντων. Στη συνέχεια, το RS με απλή συντακτική διαδικασία επιστρέφει στον πελάτη ό,τι ζήτησε. Για παράδειγμα, όταν ο πελάτης ζητά από ένα online βιβλιοπωλείο ένα βιβλίο του Νίκου Καζαντζάκη, τότε το σύστημα επιστρέφει μια λίστα με όλα τα βιβλία του, δίνοντας τη δυνατότητα στον πελάτη να βρει ίσως κάποιο βιβλίο του συγγραφέα που δεν γνώριζε.

- **Manual selection.** Ορισμένα RSs παράγουν προτάσεις με βάση δεδομένα που έχουν συλλεχθεί χειρωνακτικά από συντάκτες, καλλιτέχνες, κριτικούς ή άλλους ειδικούς. Αυτές οι προτάσεις αντανακλούν τις προτιμήσεις, τα ενδιαφέροντα, τους υποκειμενικούς λόγους και σκοπούς των δημιουργών τους και μπορεί συχνά να συνοδεύονται από σχόλια κειμένου. Για παράδειγμα, αν ο χρήστης επιλέξει ένα συγκεκριμένο είδος ταινιών π.χ. θρίλερ, τότε εμφανίζεται μια λίστα ταινιών τις οποίες κάποιος κριτικός θεωρεί σαν τα καλύτερα θρίλερ όλων των εποχών. Η μέθοδος δεν κάνει υπολογιστική επεξεργασία, απλώς αναπαράγει ό,τι θα εμφανιζόταν σε μια στήλη περιοδικού κριτικής ταινιών.
- **Statistical summaries.** Αρκετά RSs παρέχουν στατιστικές αναλύσεις των απόψεων μιας κοινότητας. Αυτά τα στοιχεία μπορεί να περιλαμβάνουν μετρήσεις δημοτικότητας μέσα στην κοινότητα (π.χ. ποσοστό ανθρώπων που τους αρέσει ή αγόρασαν ένα συγκεκριμένο προϊόν) ή συνοπτικές βαθμολογίες (π.χ. ο αριθμός των ατόμων που προτείνουν ένα προϊόν ή η μέση βαθμολογία για ένα προϊόν). Επίσης, διάφορα συστήματα παρουσιάζουν μέσες τιμές βαθμολογιών αγοραστών αλλά και πωλητών. Οι πιθανοί πελάτες μπορούν να συμβουλευτούν τις μέσες τιμές και τις εκτιμήσεις αλλά δεν μπορούν να δουν βαθμολογίες από πελάτες με τους οποίους συμφωνούν. Με αυτό τον τρόπο παράγονται μόνο μη-εξατομικευμένες προτάσεις (δηλαδή ίδιες για όλους τους πελάτες).
- **Attribute-based technologies.** Αυτά τα RSs βασίζονται στα συντακτικά χαρακτηριστικά των προϊόντων. Αν και στην πιο απλή μορφή τους εκφυλίζονται στη μέθοδο raw retrieval, όσα συστήματα χρησιμοποιούν αυτό το μοντέλο προχωράνε ένα βήμα παραπέρα από το απλό ερώτημα. Για παράδειγμα, ένα πελάτης πλοηγείται στο τμήμα «κλασσική μουσική» ενός e-shop και έχει στο καλάθι του διάφορα CDs ειδικής προσφοράς. Τότε μπορεί να λάβει μια πρόταση για έκπτωση σε κάποιο CD κλασσικής μουσικής. Άλλα συστήματα για να παράγουν προτάσεις χρησιμοποιούν τα προφίλ των χρηστών που υποδεικνύουν τις προτιμήσεις τους π.χ. μπορεί ένας πελάτης να αγοράζει μόνο CDs που έχουν έκπτωση, ενώ κάποιος άλλος ποτέ δεν αγοράζει μουσική της δεκαετίας του '70.
- **Item-to-item correlation.** Πολλά RSs αναγνωρίζουν τα προϊόντα που εμφανίζονται πιο συχνά μαζί με τα προϊόντα για τα οποία ο χρήστης έχει εκφράσει κάποιο ενδιαφέρον. Η συσχέτιση μπορεί να βασίζεται σε δεδομένα αγορών, προτιμήσεων από κοινούς πελάτες, ή άλλες μετρικές. Στην πιο απλή μορφή της, μπορεί να χρησιμοποιηθεί για να αναγνωρίσει ταιριαστά είδη για ένα συγκεκριμένο προϊόν όπως άλλου είδους ρουχισμός που αγοράζεται συνήθως μαζί με ένα παντελόνι π.χ. πουκάμισο ή πουλόβερ. Πιο προηγμένα συστήματα συσχετίζουν ένα ολόκληρο σύνολο από προϊόντα π.χ. αυτά που έχει ο χρήστης στο καλάθι του για να παράγουν τις κατάλληλες προτάσεις. Συνήθως, χρησιμοποιούνται οι τρέχουσες πωλήσεις ή άλλα τρέχοντα ενδιαφέροντα αντί για τη μακροπρόθεσμη ιστορία του πελάτη γεγονός που τις κάνει πολύ χρήσιμες για προτάσεις δώρων. Τέλος, ο

πελάτης θα πρέπει να υποδείξει και κάποια πρόσθετα χαρακτηριστικά ή προϊόντα που μπορεί να αρέσουν σε αυτόν που θα λάβει το δώρο.

- **User-to-user correlation.** Πολλά RSs προτείνουν προϊόντα σε έναν πελάτη με βάση τη συσχέτισή του με άλλους πελάτες που έχουν αγοράσει προϊόντα από το e-shop. Αυτή η τεχνολογία είναι γνωστή και σαν collaborative filtering ή social filtering και προέρχεται από την τεχνική information filtering που χρησιμοποιεί απόψεις ομάδων για να προτείνει προϊόντα σε διάφορα άτομα. Το My CDNOW είναι ένα σύστημα που χρησιμοποιεί user-to-user correlation για να αναγνωρίσει μια κοινότητα από πελάτες που τείνουν να έχουν και να τους αρέσουν τα ίδια σύνολα από CDs. Η τεχνική στηρίζεται στην αρχή ότι εάν διάφορα μέλη της κοινότητας που ανήκει ο πελάτης έχουν και τους αρέσει το τελευταίο άλμπουμ του Sting τότε είναι αρκετά πιθανό να αρέσει και σε αυτόν.

Ένα σημαντικό θέμα που θα πρέπει να προσεχθεί ιδιαίτερα σχετικά με τη μέθοδο recommendation είναι αν ο υπολογισμός θα γίνει εξολοκλήρου **online**, όταν δηλαδή ο πελάτης αλληλεπιδρά με το e-shop, ή μέρη του να υπολογίζονται **offline**, κυρίως για λόγους απόδοσης. Οι online προτάσεις προτιμώνται γιατί απαντούν απευθείας στο χρήστη. Οι raw retrieval, manual selection, statistical summaries, attribute-based μέθοδοι εκτελούν απλούς υπολογισμούς που συνήθως γίνονται στη διάρκεια της αλληλεπίδρασης του πελάτη με το site. Οι item-to-item correlation και user-to-user correlation τεχνικές είναι περισσότερο πολύπλοκες υπολογιστικά και συχνά απαιτούν μια offline επεξεργασία η οποία θα προετοιμάσει το μοντέλο που θα εκτελεσθεί online. Μια πρόκληση στο σχεδιασμό του μοντέλου είναι να διασφαλισθεί ότι το online σύστημα θα παράγει αποτελέσματα που θα ανταποκρίνονται όσο το δυνατόν περισσότερο στην αλληλεπιδραστική είσοδο του χρήστη.

2.2.4 Ο τρόπος αποστολής

Ο **τρόπος αποστολής** των προτάσεων στους πελάτες αποτελεί μια κρίσιμη σχεδιαστική απόφαση των e-shops. Οι **push τεχνολογίες** έχουν το πλεονέκτημα ότι φτάνουν στον πελάτη όταν αυτός δεν αλληλεπιδρά με το κατάστημα. Το e-mail είναι η πιο διαδεδομένη μέθοδος αυτής της κατηγορίας. Η αποστολή προτάσεων και προσφορών μπορεί να οδηγήσει έναν πελάτη στο κατάστημα. Οι σημερινές τεχνολογίες επιτρέπουν στους πελάτες να κάνουν «κλικ» σε ένα link μέσα στο κείμενο του e-mail και να μεταφερθούν κατευθείαν στο προϊόν. Οι εφαρμογές που χρησιμοποιούν **pull τεχνολογίες** επιτρέπουν στον πελάτη να ελέγξει πότε θα του παρουσιασθούν οι προτάσεις. Ο πελάτης ενημερώνεται για το χρόνο που είναι ενεργές οι προτάσεις (π.χ. εμφανίζοντας ένα link) αλλά δεν παρουσιάζονται σε αυτόν πριν ο ίδιος να τις ζητήσει. Σήμερα, οι pull τεχνολογίες είναι μια επιλογή σχεδίασης για εφαρμογές όπου οι τύποι των recommendations είναι περιφερειακοί (π.χ. η λίστα των top 10 προτάσεων για δώρα) αντί να είναι ενσωματωμένοι στην εφαρμογή. Οι **passive τεχνολογίες** παρουσιάζουν τα recommendations στο φυσικό περιβάλλον του e-shop π.χ. προτάσεις που σχετίζονται με το τρέχον προϊόν, προτάσεις για προϊόντα που σχετίζονται με το θέμα ενός κειμένου, προτάσεις σχετικές με το περιεχόμενο της εξερεύνησης του χρήστη, κλπ. Έχουν το πλεονέκτημα ότι φτάνουν στον πελάτη όταν ο ίδιος είναι ήδη ανοικτός στην ιδέα να λάβει προτάσεις. Τα e-shops χρησιμοποιούν αυτή την τεχνολογία σαν ένα μέρος της διαδικασίας παραγγελίας, προτείνοντας καλύτερες επιλογές αποστολής των προϊόντων, ή την ώρα που ο πελάτης ολοκληρώνει την παραγγελία του γεγονός που την κάνει πιο αποδοτική. Γενικά, οι μέθοδοι αποστολής αλλάζουν με την πάροδο του χρόνου. Οι πρώιμες

εφαρμογές RSs εστίαζαν σε push και pull τεχνολογίες γιατί η επιθυμία να δείξουν ενδιαφέρον στους πελάτες ήταν σημαντική. Πρόσφατα, οι εφαρμογές κινούνται προς την κατεύθυνση των passive και push τεχνολογιών, με την πρώτη να εφαρμόζεται στα web sites και τη δεύτερη να χρησιμοποιείται για να φέρει τους πελάτες πίσω σε αυτά.

2.2.5 Ο βαθμός εξατομίκευσης

Τα RSs μπορούν να παράγουν προτάσεις με διαφορετικό **βαθμό εξατομίκευσης (personalization degree)**. Όταν το σύστημα προτείνει όμοια recommendations σε κάθε πελάτη τότε η εφαρμογή θεωρείται ότι είναι **μη-εξατομικευμένη (non-personalized)**. Τα συγκεκριμένα recommendations μπορεί να βασίζονται σε επιλογή με το χέρι, στατιστικά στοιχεία, ή άλλες τεχνικές. Οι μεγαλύτερες πωλήσεις ενός προϊόντος, οι επιλογές των κριτικών, οι μέσοι όροι, τα ακατέργαστα σχόλια, κλπ. παρουσιάζουν τις ίδιες προτάσεις σε όλους τους πελάτες του e-shop. Τα RSs που χρησιμοποιούν δεδομένα των πελατών για να προσαρμόσουν τις προτάσεις στα ενδιαφέροντά τους παρέχουν **εφήμερη εξατομίκευση (ephemeral personalization)**. Στην περίπτωση αυτή, οι προτάσεις ανταποκρίνονται στην πλοήγηση του πελάτη και στις επιλογές του. Αυτά τα συστήματα βασίζονται συνήθως σε item-to-item correlation, attribute-based recommendation ή και τα δυο μαζί. Τέλος, οι εφαρμογές που χρησιμοποιούν **επίμονη εξατομίκευση (persistent personalization)** μπορούν να δημιουργήσουν προτάσεις που είναι διαφορετικές για διαφορετικούς πελάτες ακόμα και αν αυτοί κοιτάνε τα ίδια προϊόντα. Συνήθως χρησιμοποιούν item-to-item correlation, user-to-user correlation ή attribute-based recommendation. Απαιτούν από τους πελάτες να συντηρούν μόνιμες **ταυτότητες (profiles)** αλλά τους επιβραβεύουν με μεγαλύτερο βαθμό εξατομίκευσης.

2.3 Ταξινόμηση των RSs

Όπως έχουμε ήδη αναφέρει τα RSs μπορούν να ταξινομηθούν σε τρεις βασικές κατηγορίες: **content-based filtering (φιλτράρισμα με βάση το περιεχόμενο)**, **collaborative filtering (συνεργατικό φιλτράρισμα)** και **hybrid approaches (υβριδικές προσεγγίσεις)**. Τα τελευταία χρόνια σημαντική έρευνα έχει γίνει στην περιοχή. Ο Πίνακας 6 παρουσιάζει μια ταξινόμηση των τεχνικών RSs καθώς και αντιπροσωπευτικά ερευνητικά παραδείγματα (Adomavicius & Tuzhilin, 2005).

Κατηγορίες τεχνικών RSs	Ευριστικές (heuristic-based)	Μοντέλου (model-based)
Φιλτράρισμα με βάση το περιεχόμενο ή content-based filtering	Τεχνικές: • TF-IDF (Information Retrieval) • Clustering Αντιπροσωπευτικά παραδείγματα: • Lang, 1995 • Balabanovic & Shoham, 1997 • Pazzani & Billsus, 1997	Τεχνικές: • Bayesian classifiers • Clustering • Decision trees • Artificial neural networks Αντιπροσωπευτικά παραδείγματα: • Pazzani & Billsus, 1997 • Mooney et al., 1998 • Mooney & Roy, 1999 • Billsus & Pazzani, 1999, 2000

		• Zhang et al., 2002
Συνεργατικό φιλτράρισμα ή collaborative filtering	Τεχνικές: • Nearest neighbor (cosine, correlation) • Clustering • Graph theory Αντιπροσωπευτικά παραδείγματα: • Resnick et al., 1994 • Hill et al., 1995 • Shardanand & Maes, 1995 • Breese et al., 1998 • Nakamura & Abe, 1998 • Aggarwal et al., 1999 • Delgado & Ishii, 1999 • Pennock & Horwitz, 1999 • Sarwar et al., 2001	Τεχνικές: • Bayesian networks • Clustering • Artificial neural networks • Linear regression • Probabilistic models Αντιπροσωπευτικά παραδείγματα: • Billsus & Pazzani, 1998 • Breese et al., 1998 • Ungar & Foster, 1998 • Chien & George, 1999 • Getoor & Sahami, 1999 • Pennock & Horwitz, 1999 • Goldberg et al., 2001 • Kumar et al., 2001 • Pavlov & Pennock, 2002 • Shani et al., 2002 • Yu et al., 2002, 2004 • Hofmann 2003, 2004 • Marlin, 2003 • Si & Jin, 2003
Υβριδικές προσεγγίσεις ή hybrid approaches	Τεχνικές: • Linear combination of predicted ratings • Various voting schemes • Incorporating one component as a part of the heuristic for the other Αντιπροσωπευτικά παραδείγματα: • Balabanovic & Shoham, 1997 • Craypool et al., 1999 • Good et al., 1999 • Pazzani, 1999 • Billsus & Pazzani, 2000 • Tran & Cohen, 2000 • Melville et al., 2002	Τεχνικές: • Incorporating one component as a part of the model for the other • Building one unifying model Αντιπροσωπευτικά παραδείγματα: • Basu et al., 1998 • Condliff et al., 1999 • Soboroff & Nicholas, 1999 • Ansari et al., 2000 • Popescul et al., 2001 • Schein et al., 2002

Πίνακας 6. Ταξινόμηση τεχνικών για RSs και αντιπροσωπευτικά ερευνητικά παραδείγματα.

2.3.1 Content-based approaches

Η content-based προσέγγιση για την παραγωγή recommendations έχει τις ρίζες της στις επιστημονικές περιοχές της **Ανάκτησης Πληροφοριών (Information Retrieval)** (Salton, 1989), (Baeza-Yates & Ribeiro-Neto, 1999) και του **Φιλτραρίσματος Πληροφοριών (Information Filtering)** (Belkin & Croft, 1992) και υιοθετεί τεχνικές που έχουν μελετηθεί εκτενώς στο παρελθόν. Η βελτίωση των παραδοσιακών τεχνικών προέρχεται από τη χρήση των προφίλ των χρηστών που περιέχουν πληροφορίες για την ταυτότητά τους, τις προτιμήσεις, τα ενδιαφέροντα και τις ανάγκες τους. Οι πληροφορίες αυτές μπορούν να εξαχθούν από τους χρήστες ρητά π.χ. μέσω ερωτηματολογίων, ή υπονοούμενα μέσω της παρακολούθησης της συμπεριφοράς και των συναλλαγών τους κατά τη διάρκεια του χρόνου.

Οι αλγόριθμοι αυτοί συνήθως χρησιμοποιούνται όταν οι συστάσεις αφορούν σε ιστοσελίδες, δημοσιεύσεις, αστεία, ειδήσεις, κλπ. Το σύστημα συλλέγει πληροφορίες για τις προτιμήσεις των χρηστών είτε μέσω της διαδικασίας εγγραφής στο site, είτε μέσω της εκτίμησης των εγγράφων. Οι συστάσεις παράγονται στη συνέχεια λαμβάνοντας υπόψη το περιεχόμενο των εγγράφων και φιλτράρονται ώστε να επιτευχθεί η καλύτερη αντιστοίχιση με τις προτιμήσεις του χρήστη όπως αυτές έχουν καταγραφεί στο προφίλ του. Για παράδειγμα, ένα RS κινηματογραφικών ταινιών προκειμένου να παράγει προτάσεις για ταινίες m_i για το χρήστη u , προσπαθεί να βρει μεταξύ των ταινιών m_j που ο u έχει βαθμολογήσει υψηλά στο παρελθόν τις ομοιότητες (π.χ. ηθοποιοί, σκηνοθέτες, είδος ταινιών, κλπ.). Έτσι, μόνο οι ταινίες που έχουν υψηλό βαθμό ομοιότητας με τις προτιμήσεις του χρήστη προτείνονται.

Έστω $Content(m)$ είναι το προφίλ ενός αντικειμένου m , δηλαδή ένα σύνολο από χαρακτηριστικά που αντιπροσωπεύουν το περιεχόμενο του m και το χαρακτηρίζουν. Το περιεχόμενο αυτό συνήθως περιγράφεται με **λέξεις-κλειδιά (keywords)**. Για παράδειγμα, το **σύστημα FAB** (Balabanovic & Shoham, 1997) προτείνει Web σελίδες στους χρήστες του και ορίζει τα περιεχόμενα της σελίδας με βάση τις 100 πιο σημαντικές λέξεις της. Η «σημαντικότητα» της λέξης k_i στο έγγραφο d_j καθορίζεται με μια **μετρική βάρους** w_{ij} που μπορεί να οριστεί με διάφορους τρόπους. Μια από τις πιο γνωστές μετρικές από την περιοχή του Information Retrieval είναι η **TF – IDF (term frequency/inverse document frequency)** που ορίζεται ως εξής: υποθέτουμε ότι N είναι ο συνολικός αριθμός των documents που μπορούν να προταθούν στους χρήστες και το keyword k_j εμφανίζεται σε n_i από αυτούς. Σαν f_{ij} ορίζουμε τον **αριθμό εμφάνισης (συχνότητα)** του keyword k_i στο document d_j . Τότε η TF_{ij} (term frequency) του keyword k_i στο document d_j ορίζεται σαν:

$$TF_{ij} = \frac{f_{ij}}{\max_z f_{zj}}$$

όπου το $\max$ υπολογίζεται σε όλες τις συχνότητες f_{zj} όλων των k_z που εμφανίζονται σε ένα d_j . Εντούτοις, τα keywords που εμφανίζονται σε πολλά έγγραφα δεν μπορούν πάντα να διακρίνουν αν ένα έγγραφο είναι σχετικό ή όχι. Έτσι, η μετρική IDF_i (**inverse document frequency**) χρησιμοποιείται σε συνδυασμό με την TF_{ij} και ορίζεται για το keyword k_i σαν:

$$IDF_i = \log \frac{N}{n_i}$$

Τότε, το **TF – IDF** βάρος w_{ij} του keyword k_i στο document d_j δίνεται από τη σχέση:

$$w_{ij} = TF_{ij} \times IDF_i$$

και το περιεχόμενο (content) του document d_j ορίζεται σαν:

$$Content(d_j) = (w_{1j}, \dots, w_{kj})$$

Έστω $ContentBasedProfile(u)$ είναι το προφίλ του χρήστη u που περιλαμβάνει τα χαρακτηριστικά, τα ενδιαφέροντα και τις προτιμήσεις του. Αυτό το προφίλ δημιουργείται από την επεξεργασία του περιεχομένου των αντικειμένων που έχει δει και βαθμολογήσει και βασίζεται με τεχνικές keywords ανάλυσης. Για παράδειγμα, μπορεί να ορισθεί σαν ο πίνακας των βαρών $(w_{u1}, \dots, w_{uk})$ όπου κάθε βάρος w_{ui} δηλώνει τη σημασία του keyword k_i για το χρήστη u και μπορεί να εκτιμηθεί από τις ατομικές βαθμολογίες με διάφορες τεχνικές π.χ. μέσος όρος. Αυτό που πρέπει να υπολογισθεί στη συνέχεια είναι το:

$$score(ContentBasedProfile(u), Contents(m))$$

Τόσο το $ContentBasedProfile(u)$ του χρήστη u όσο και το $Contents(m)$ του εγγράφου m μπορούν να αναπαρασταθούν σαν $TF-IDF$ πίνακες από βάρη $\vec{w}_u$ και $\vec{w}_m$ αντίστοιχα. Το $score$ μπορεί να είναι η μετρική **cosine similarity**:

$$\cos(\vec{w}_u, \vec{w}_m) = \frac{\vec{w}_u \cdot \vec{w}_m}{\|\vec{w}_u\|_2 \times \|\vec{w}_m\|_2} = \frac{\sum_{i=1}^K w_{ui} w_{im}}{\sqrt{\sum_{i=1}^K w_{ui}^2} \sqrt{\sum_{i=1}^K w_{im}^2}}$$

όπου K είναι ο συνολικός αριθμός των keywords στο σύστημα. Για παράδειγμα, αν ο χρήστης u διαβάζει πολλά online άρθρα για Βιοπληροφορική, τότε το RS θα του προτείνει και άλλα άρθρα Βιοπληροφορικής. Αυτό συμβαίνει γιατί τα άρθρα αυτά περιέχουν τους σχετικούς όρους π.χ. γωνιδίωμα, ακολουθία, κλπ. περισσότερο από ότι τα άρθρα με άλλα θεματικά. Οπότε το $ContentBasedProfile(u)$ όπως παριστάνεται από το $\vec{w}_u$ θα αναπαριστά του όρους k_i με υψηλά βάρη w_{ui} . Το RS χρησιμοποιώντας την cosine ή κάποια άλλη μετρική θα δώσει υψηλό $score$ σε αυτά τα άρθρα m που έχουν υψηλά βαθμολογημένους τους όρους Βιοπληροφορικής στο $\vec{w}_m$ και χαμηλό $score$ στα υπόλοιπα άρθρα που οι όροι αυτοί είναι χαμηλά βαθμολογούμενοι.

Άλλες τεχνικές της κατηγορίας που έχουν επίσης χρησιμοποιηθεί είναι οι **Bayesian Classifiers (Ταξινομητές)** (Pazzani & Billsus, 1997), (Mooney et al., 1998) οι **Machine Learning** τεχνικές συμπεριλαμβανομένων και τεχνικών **Clustering (Συσταδοποίησης)**, **Decision Trees (Δέντρων Απόφασης)** και **Artificial Neural Networks (Τεχνητών Νευρωνικών Δικτύων)** (Pazzani & Billsus, 1997). Αυτές οι τεχνικές διαφέρουν στο γεγονός ότι υπολογίζουν τις προβλέψεις όχι με ευρυστικό τρόπο π.χ. το μέτρο ομοιότητας συνημίτονου (cosine similarity measure), αλλά βασισμένες σε ένα μοντέλο που εκπαιδεύεται από συλλεγμένα στοιχεία χρησιμοποιώντας στατιστικές και Machine Learning τεχνικές. Για παράδειγμα, με βάση ένα σύνολο ιστοσελίδων που εκτιμήθηκαν ως «σχετικές» ή «άσχετες» από το χρήστη (Pazzani & Billsus, 1997) χρησιμοποιούμε τον naïve Bayesian Classifier (Duda et al., 2001) που κατηγοριοποιεί τις αταξινομήτες ιστοσελίδες. Συγκεκριμένα, ο ταξινομητής υπολογίζει την πιθανότητα ότι μια σελίδα p_j ανήκει σε μια συγκεκριμένη κλάση C_i (π.χ. σχετική ή άσχετη) δεδομένου ενός συνόλου από keywords $k_{1j}, \dots, k_{nj}$ της σελίδας: $P(C_i | k_{1j} \& \dots \& k_{nj})$. Χρησιμοποιώντας την υπόθεση ότι τα keywords είναι ανεξάρτητα

μεταξύ τους, τότε η προηγούμενη πιθανότητα είναι ανάλογη του $P(C_i) \prod_x P(k_{xj} | C_i)$.

Πειραματικά αποτελέσματα έδειξαν ότι ο naïve Bayesian Classifier παρουσιάζει υψηλή ακρίβεια στην ταξινόμηση. Επιπλέον, οι $P(k_{xj} | C_i)$ και $P(C_i)$ μπορούν να υπολογιστούν από δεδομένα εκπαίδευσης. Έτσι για κάθε σελίδα p_j , η πιθανότητα $P(C_i | k_{1j} \& \dots \& k_{nj})$ υπολογίζεται για κάθε κλάση C_i και η σελίδα p_j τοποθετείται στην κλάση C_i που έχει τη μεγαλύτερη πιθανότητα.

2.3.2 Collaborative filtering approaches

Σε αντίθεση με τις content-based τεχνικές, τα collaborative filtering συστήματα προσπαθούν να προβλέψουν τη χρησιμότητα των αντικειμένων για ένα συγκεκριμένο χρήστη λαμβάνοντας υπόψη τα αντικείμενα που έχουν βαθμολογηθεί προηγουμένως από άλλους χρήστες. Αυτό σημαίνει ότι η εκτίμηση του **αντικειμένου** m για το **χρήστη** u υπολογίζεται με βάση τις **εκτιμήσεις** που δίνονται στο m από εκείνους τους χρήστες u_j που είναι «παρόμοιοι» με τον u . Για παράδειγμα ένα CF-RS προκειμένου να προτείνει ταινίες στο χρήστη u προσπαθεί να βρει άλλους χρήστες με παρόμοιες προτιμήσεις στις ταινίες (έχουν εκτιμήσει τις ίδιες ταινίες με παρόμοιο τρόπο). Στη συνέχεια, θα προτείνει μόνο εκείνες τις ταινίες που άρεσαν περισσότερο στους **«γείτονες»** του u . Σύμφωνα με τους Breese et al., (1998), τα CF-RSs χωρίζονται σε δυο κατηγορίες: στα **memory-based** (ή **heuristic-based**) και στα **model-based**.

Memory-based τεχνικές

Οι memory-based αλγόριθμοι αποτελούν heuristics που κάνουν προβλέψεις βασισμένοι σε μια ολόκληρη συλλογή από προηγούμενα βαθμολογημένα αντικείμενα από τους χρήστες. Η τιμή μιας άγνωστης βαθμολογίας r_{um} για το χρήστη u και το αντικείμενο m συνήθως υπολογίζεται σαν η συνάθροιση των βαθμολογιών κάποιων άλλων χρηστών (συνήθως των N περισσότερο όμοιων) για το ίδιο αντικείμενο m : $r_{um} = \text{aggr}(r_{u'm})$ όπου το $\hat{U}$ είναι το σύνολο των N χρηστών που είναι πιο όμοιοι με τον u και έχουν βαθμολογήσει το αντικείμενο m (το N μπορεί να κυμαίνεται από το 1 έως τον αριθμό όλων των χρηστών). Μερικά παραδείγματα της **aggregation συνάρτησης** είναι:

$$r_{um} = \frac{1}{N} \sum_{u' \in \hat{U}} r_{u'm} \quad (\alpha),$$

$$r_{um} = k \sum_{u' \in \hat{U}} \text{sim}(u, u') \times r_{u'm} \quad (\beta)$$

$$r_{um} = \bar{r}_u + k \sum_{u' \in \hat{U}} \text{sim}(u, u') \times (r_{u'm} - \bar{r}_{u'}) \quad (\gamma)$$

όπου ο πολλαπλασιαστής k λειτουργεί σαν **παράγοντας κανονικοποίησης** και είναι:

$$k = \frac{1}{\sum_{u' \in \hat{U}} |sim(u, u')|}$$

ενώ η μέση εκτίμηση $\bar{r}_u$ του χρήστη u στην (γ) δίνεται από τον τύπο:

$$\bar{r}_u = \left(\frac{1}{|M_u|} \right) \sum_{m \in M_u} r_{um}$$

όπου $M_u = \{m \in M \mid r_{um} \neq \emptyset\}$ ($r_{um} = \emptyset$ σημαίνει ότι το αντικείμενο m δεν έχει βαθμολογηθεί από το χρήστη u). Στην πιο απλή περίπτωση η aggregation είναι ο μέσος όρος (α), ενώ σε πιο προχωρημένες περιπτώσεις είναι το σταθμισμένο (weighted) άθροισμα (β). Η συνάρτηση ομοιότητας μεταξύ των χρηστών u και u' $sim(u, u')$ είναι μια μετρική απόστασης. Στις περισσότερες προσεγγίσεις για τον υπολογισμό της **similarity**, η ομοιότητα μεταξύ δύο χρηστών είναι βασισμένη στις εκτιμήσεις των αντικειμένων και από τους δύο χρήστες. Οι πιο δημοφιλείς μετρικές είναι η **correlation** και η **cosine-based**. Έστω M_{XY} είναι το σύνολο όλων των αντικειμένων που και οι δύο χρήστες X και Y έχουν εκτιμήσει δηλαδή $M_{XY} = \{m \in M \mid r_{Xm} \neq \emptyset \ \& \ r_{Ym} \neq \emptyset\}$. Το M_{XY} είναι συνήθως ένα ενδιάμεσο αποτέλεσμα στη διαδικασία για τον υπολογισμό των «κοντινότερων γειτόνων» του χρήστη X και υπολογίζεται συχνά σαν η τομή των συνόλων M_X και M_Y . Εντούτοις, μερικές μέθοδοι, όπως η **graph-theoretic** (Aggarwal et al., 1999) μπορούν να καθορίσουν τους κοντινότερους γείτονες του X χωρίς τον υπολογισμό του M_{XY} για όλους τους χρήστες Y . Στην περίπτωση της correlation-based, ο **συντελεστής Pearson** χρησιμοποιείται για να μετρήσει την ομοιότητα (Resnick et al., 1994), (Shardanand & Maes, 1995).

$$sim(X, Y) = \frac{\sum_{m \in M_{XY}} (r_{Xm} - \bar{r}_X)(r_{Ym} - \bar{r}_Y)}{\sqrt{\sum_{m \in M_{XY}} (r_{Xm} - \bar{r}_X)^2 \sum_{m \in M_{XY}} (r_{Ym} - \bar{r}_Y)^2}}$$

Στην περίπτωση της cosine-based (Breese et al., 1998), (Sarwar et al., 2001) οι δύο χρήστες X και Y αντιμετωπίζονται σαν δύο διανύσματα στο μ -διαστάσεων χώρο όπου $\mu = |M_{XY}|$. Η ομοιότητα μεταξύ των δύο διανυσμάτων μπορεί να μετρηθεί με τον υπολογισμό του συνημίτονου της γωνίας μεταξύ τους:

$$sim(X, Y) = \cos(\bar{X}, \bar{Y}) = \frac{\bar{X} \cdot \bar{Y}}{\|\bar{X}\|_2 \times \|\bar{Y}\|_2} = \frac{\sum_{m \in M_{XY}} r_{Xm} r_{Ym}}{\sqrt{\sum_{m \in M_{XY}} r_{Xm}^2} \sqrt{\sum_{m \in M_{XY}} r_{Ym}^2}}$$

Μια άλλη προσέγγιση στη μέτρηση της ομοιότητας μεταξύ των χρηστών είναι η **mean squared difference μετρική** (Shardanand & Maes, 1995). Διαφορετικά RSs υιοθετούν

διαφορετικές μεθόδους προκειμένου να υπολογίσουν την ομοιότητα των χρηστών και να παράγουν εκτιμήσεις όσο το δυνατόν αποτελεσματικότερα. Μια μεθοδολογία είναι να υπολογιστούν όλες οι ομοιότητες $sim(X, Y)$ των χρηστών (συμπεριλαμβανομένου και του υπολογισμού M_{XY}) εκ των προτέρων και να τους υπολογίζει εκ νέου μόνο όταν απαιτείται (δεδομένου ότι δεν συμβαίνουν δραματικές αλλαγές σε μικρό χρονικό διάστημα). Κατόπιν, όποτε ο χρήστης ζητά μια σύσταση, οι εκτιμήσεις μπορούν να υπολογιστούν αποτελεσματικά χρησιμοποιώντας τις προϋπολογισμένες ομοιότητες. Πρέπει να σημειωθεί σε αυτό το σημείο ότι οι content-based και οι CF προσεγγίσεις χρησιμοποιούν τη μετρική cosine από την περιοχή του Information Retrieval. Στην πρώτη όμως περίπτωση χρησιμοποιείται για να μετρήσει την ομοιότητα μεταξύ των διανυσμάτων των $TF - IDF$ βαρών, ενώ στη δεύτερη την ομοιότητα μεταξύ των διανυσμάτων των πραγματικών εκτιμήσεων των χρηστών.

Τροποποιήσεις των παραπάνω έχουν προταθεί σε πολλές εργασίες με στόχο τη βελτίωση της απόδοσης π.χ. **default voting**, **inverse user frequency**, **case amplification** (Breese et al., 1998) και **weighted-majority prediction** (Delgado & Ishii, 1999), (Nakamura & Abe, 1998). Ενώ οι παραπάνω τεχνικές παραδοσιακά έχουν χρησιμοποιηθεί για να υπολογίσουν τις ομοιότητες μεταξύ των χρηστών, οι Sarwar et al., (2001) πρότειναν να υπολογιστούν οι ομοιότητες μεταξύ των αντικειμένων και να ληφθούν οι εκτιμήσεις από αυτές. Αυτή η ιδέα επεκτάθηκε περαιτέρω στην εργασία των Deshpande & Karypis (2004) για την παραγωγή των N καλύτερων προτάσεων. Τα εμπειρικά αποτελέσματά τους δείχνουν ότι έχουν καλύτερη υπολογιστική απόδοση από τις παραδοσιακές μεθόδους.

Model-based τεχνικές

Οι model-based μέθοδοι βασίζονται αρχικά σε ένα σύνολο δεδομένων για να εκπαιδεύσουν ένα μοντέλο, το οποίο χρησιμοποιούν στη συνέχεια για να κάνουν τις προβλέψεις (Billsus, & Pazzani, 1998), (Breese et al., 1998), (Getoor & Sahami, 1999), (Goldberg et al., 2001), (Hofmann, 2003), (Marlin, 2003), (Pavlov & Pennock, 2002), (Ungar & Foster, 1998). Στην εργασία (Breese et al., 1998) προτείνεται μια πιθανοτική CF προσέγγιση, όπου οι άγνωστες εκτιμήσεις υπολογίζονται ως εξής:

$$r_{um} = E(r_{um}) = \sum_{i=0}^n i \times \Pr(r_{um} = i | r_{um'}, m' \in M_u)$$

Οι τιμές εκτίμησης είναι ακέραιοι αριθμοί μεταξύ 0 και n και η έκφραση πιθανότητας είναι η πιθανότητα ότι ο χρήστης u θα κάνει μια ιδιαίτερη εκτίμηση για το στοιχείο m δεδομένων των προηγούμενων εκτιμήσεων του. Για να υπολογισθεί αυτή η εκτίμηση προτείνονται δύο εναλλακτικά πιθανοτικά μοντέλα: το **Cluster model** και τα **Bayesian δίκτυα**. Στην πρώτη περίπτωση «ομοιόδεατες» χρήστες ομαδοποιούνται σε **clusters (ομάδες)**. Λαμβάνοντας υπόψη την ιδιότητα μέλους κάθε ομάδας, οι εκτιμήσεις του χρήστη υποθέτουμε ότι ήταν ανεξάρτητες, δηλαδή η δομή του μοντέλου είναι αυτή του naïve Bayesian. Ο αριθμός των ομάδων και οι παράμετροι του μοντέλου εκπαιδεύονται από τα δεδομένα εκπαίδευσης. Το δεύτερο μοντέλο αντιπροσωπεύει κάθε αντικείμενο στην περιοχή σαν κόμβο σε ένα Bayesian δίκτυο, όπου οι καταστάσεις κάθε κόμβου αντιστοιχούν σε πιθανές εκτιμήσεις για κάθε αντικείμενο. Τόσο η δομή του δικτύου όσο και οι υπό συνθήκη πιθανότητες εκπαιδεύονται από τα δεδομένα. Ένας περιορισμός αυτής της προσέγγισης είναι ότι κάθε χρήστης μπορεί να τοποθετηθεί σε μια μόνο ομάδα, ενώ τα RSs μπορούν να ωφεληθούν

από τη δυνατότητα να τοποθετούνται οι χρήστες σε διαφορετικές ομάδες. Σε ένα RS που προτείνει βιβλία, ένας χρήστης μπορεί να ενδιαφερθεί για το θέμα «Προγραμματισμός» για εργασιακούς λόγους και για το τελείως διαφορετικό θέμα «Κηπουρική» για χόμπι.

Επιπλέον, οι Billsus & Pazzani, (1998) προτείνουν μια μέθοδο CF στην περιοχή του **Machine Learning** όπου διάφορες τεχνικές όπως τα **Artificial Neural Networks** μαζί με **τεχνικές εξαγωγής χαρακτηριστικών γνωρισμάτων** μπορούν να χρησιμοποιηθούν. Σε ένα άλλο στατιστικό μοντέλο που μελετήθηκε από τους (Ungar & Foster, 1998) συγκρίθηκαν διαφορετικοί αλγόριθμοι όπως ο **K-means Clustering** και ο **Gibbs sampling**. Άλλες μέθοδοι είναι το **Bayesian model** (Chien & George, 1999), το **πιθανοτικό relational model** (Getoor & Sahami, 1999), η **linear regression** τεχνική (Sarwar et al., 2001) και το **maximum entropy model** (Pavlov & Pennock, 2002). Πρόσφατα έχει γίνει προσπάθεια να μοντελοποιηθεί η διαδικασία παραγωγής προτάσεων χρησιμοποιώντας πιο σύνθετα πιθανοτικά πρότυπα. Οι Shani et al., (2002) θεωρούν αυτή τη διαδικασία σαν ένα ακολουθιακό πρόβλημα απόφασης και χρησιμοποιούν **Markov διαδικασίες** (γνωστή πιθανοτική τεχνική για τη μοντελοποίηση διαδοχικών αποφάσεων) για την παραγωγή των συστάσεων. Άλλες παρόμοιες τεχνικές προτείνονται στις εργασίες (Hofmann, 2003 & 2004), (Marlin, 2003), (Si & Jin, 2003), (Kumar et al., 2001). Στην εργασία (Pennock & Horvitz, 1999) συνδυάζονται memory-based και model-based προσεγγίσεις και αποδεικνύεται εμπειρικά ότι οδηγούν σε καλύτερες προτάσεις. Οι Yu et al., (2002) παρουσιάζουν μια διαφορετική προσέγγιση για τη βελτίωση της απόδοσης των CF αλγορίθμων, όπου το σύνολο δεδομένων που περιλαμβάνει τις καθορισμένες από το χρήστη εκτιμήσεις έχει επιλεγεί προσεκτικά με βάση διάφορες τεχνικές που αποκλείουν το θόρυβο, τον πλεονασμό και εκμεταλλεύονται την αραιότητα των εκτιμήσεων. Τα εμπειρικά αποτελέσματα έδειξαν βελτίωση της ακρίβειας και της αποδοτικότητας των model-based CF αλγορίθμων.

2.3.3 Hybrid approaches

Αρκετά RSs χρησιμοποιούν την **υβριδική προσέγγιση** συνδυάζοντας collaborative και content-based τεχνικές με στόχο να αποφύγουν τους περιορισμούς που ενέχει κάθε μία από αυτές (Balabanovic & Shoham, 1997), (Basu et al., 1998), (Ungar & Fosters, 1998), (Claypool et al., 1999), (Soboroff & Nicholas, 1999), (Pazzani, 1999), (Schein et al., 2002). Οι διαφορετικοί τρόποι που μπορούν να συνδυαστούν οι δύο τεχνικές είναι οι ακόλουθοι:

- Υλοποίηση των τεχνικών ξεχωριστά και συνδυασμός των προβλέψεων.
- Ενσωμάτωση content-based χαρακτηριστικών σε μια collaborative προσέγγιση.
- Ενσωμάτωση collaborative χαρακτηριστικών σε μια content-based προσέγγιση.
- Δημιουργία ενός γενικού μοντέλου το οποίο ενσωματώνει και content-based και collaborative χαρακτηριστικά.

Συνδυασμός ξεχωριστών RSs

Στην προσέγγιση αυτή υλοποιούνται ξεχωριστά ένα collaborative και ένα content-based RS. Στη συνέχεια μπορούν να εφαρμοστούν δύο σενάρια. Πρώτον μπορούν να συνδυαστούν τα αποτελέσματα (εκτιμήσεις) που λαμβάνονται από τα μεμονωμένα RSs σε μια τελική λίστα (Claypool et al., 1999), (Pazzani, 1999). Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε κάθε

δεδομένη στιγμή ένα από τα RSs, επιλέγοντας εκείνο που δίνει τα καλύτερα αποτελέσματα με βάση κάποια μετρικής «ποιότητας». Για παράδειγμα, το σύστημα DailyLearner (Billsus & Pazzani, 2000) επιλέγει το RS που μπορεί να δώσει τη σύσταση με το υψηλότερο επίπεδο εμπιστοσύνης (confidence), ενώ το (Tran & Cohen, 2000) επιλέγει αυτό το RS του οποίου οι προτάσεις είναι πιο σύμφωνες με τις προηγούμενες εκτιμήσεις του χρήστη.

Ενσωμάτωση content-based χαρακτηριστικών σε ένα collaborative μοντέλο

Πολλά υβριδικά RSs όπως στο σύστημα Fab (Balabanovic & Shoham, 1997) και σε αυτό που παρουσιάζεται στην εργασία (Pazzani, 1999) και αναφέρεται σαν «συνεργασία μέσω περιεχομένου» ο συνδυασμός των δύο τεχνικών γίνεται μέσω (α) της δημιουργίας και της διατήρησης των προφίλ των χρηστών με βάση την ανάλυση περιεχομένου και κάνοντας χρήση μεθόδων ανάκτησης πληροφοριών και άλλων content-based τεχνικών και (β) συγκρίνοντας άμεσα τα προφίλ που κατασκευάζονται ώστε να ευρεθούν όμοιοι χρήστες και να παραχθούν τα recommendations. Αυτό σημαίνει ότι στους χρήστες μπορούν να προταθούν αντικείμενα τα οποία είτε έχουν υψηλή βαθμολόγηση σε σχέση με το προφίλ του χρήστη, είτε έχουν βαθμολογηθεί υψηλά από έναν άλλο χρήστη με παρόμοιο προφίλ (Balabanovic & Shoham, 1997). Επιπλέον, επιτρέπει να ξεπεραστούν ορισμένα προβλήματα που οφείλονται στην αραιότητα των δεδομένων των CF-RSs αφού πολλά ζευγάρια χρηστών έχουν ένα σημαντικό αριθμό από κοινά βαθμολογημένα αντικείμενα. Οι Good et al., (1999) πρότειναν μια παρόμοια προσέγγιση που χρησιμοποιεί μια ποικιλία από διαφορετικούς **filterbots (ειδικοί content-based πράκτορες)** που λειτουργούν σαν επιπλέον συμμετέχοντες στη CF κοινότητα. Σαν αποτέλεσμα, οι χρήστες των οποίων οι βαθμολογίες συμφωνούν με τις εκτιμήσεις των filterbots μπορούν να λάβουν καλύτερες προτάσεις. Ομοίως, οι (Melville et al., 2002) χρησιμοποιούν την CF προσέγγιση όπου το παραδοσιακό διάλυμα με τις εκτιμήσεις των χρηστών εμπλουτίζεται με πρόσθετες εκτιμήσεις που υπολογίζονται χρησιμοποιώντας έναν **content-based predictor**.

Ενσωμάτωση collaborative χαρακτηριστικών σε ένα content-based μοντέλο

Η πιο δημοφιλής προσέγγιση αυτής της κατηγορίας είναι η χρήση της **dimensionality reduction τεχνικής** σε μια ομάδα από content-based προφίλ. Οι Soboroff & Nicholas (1999) χρησιμοποιούν **latent semantic indexing (LSI)** έτσι ώστε να δημιουργηθεί μια collaborative άποψη από μια συλλογή από προφίλ χρηστών, τα οποία αναπαρίστανται σαν διανύσματα όρων οδηγώντας σε βελτίωση της απόδοσης σε σύγκριση με τις content-based τεχνικές.

Δημιουργία ενός γενικού μοντέλου το οποίο ενσωματώνει και content-based και collaborative χαρακτηριστικά

Πολλοί ερευνητές ακολουθούν αυτή την προσέγγιση τα τελευταία χρόνια. Για παράδειγμα, οι Basu et al., (1998) προτείνουν τη χρήση content-based και collaborative χαρακτηριστικών (π.χ. ηλικία ή φύλο των χρηστών ή είδος των ταινιών) σε ένα ενιαίο **βασισμένο σε κανόνες (rule-based) ταξινομητή (classifier)**. Οι Popescul et al., (2001) και οι Schein et al., (2002) προτείνουν μια πιθανοτική μέθοδο που συνδυάζει collaborative και content-based συστάσεις και βασίζεται στη **latent semantic analysis (LSA)** (Hofmann, 1999). Μια άλλη προσέγγιση προτείνεται από τους (Condliff et al., 1999), (Ansari et al., 2000), όπου **Bayesian regression μοντέλα** χρησιμοποιούνται και τα οποία υιοθετούν μεθόδους **Markov chain Monte Carlo** για παραμετρική εκτίμηση και πρόβλεψη. Ειδικότερα στην (Ansari et al., 2000),

χρησιμοποιείται το προφίλ των χρηστών και των αντικειμένων σε ένα ενιαίο στατιστικό μοντέλο που υπολογίζει τις άγνωστες εκτιμήσεις r_{ij} για το χρήστη u και το αντικείμενο m :

$$r_{ij} = x_{ij}\mu + z_i\gamma_j + w_j\lambda_i + e_{ij}, \text{ με } e_{ij} \sim N(0, \sigma^2), \lambda_i \sim N(0, \Lambda), \gamma_j \sim N(0, \Gamma) \text{ και}$$

αφορούν στις τυχαίες μεταβλητές που λαμβάνουν υπόψη το θόρυβο επίδρασης, καθώς και απαραίτητες πηγές ετερογένειας των χρηστών και ετερογένειας των αντικειμένων αντίστοιχα, ενώ τα $u = 1, \dots, U$ και $m = 1, \dots, M$ αντιπροσωπεύουν τους χρήστες και τα αντικείμενα. Επίσης, x_{ij} είναι μια μήτρα που περιέχει τα χαρακτηριστικά των χρηστών και των αντικειμένων, z_i είναι ένα διάνυσμα με τα χαρακτηριστικά των χρηστών και w_j είναι ένα διάνυσμα με τα χαρακτηριστικά των αντικειμένων. Οι άγνωστες παράμετροι αυτού του μοντέλου μ , σ^2 , Λ και Γ υπολογίζονται από τα στοιχεία των ήδη γνωστών εκτιμήσεων χρησιμοποιώντας τις μεθόδους Markov chain Monte Carlo. Στην ουσία, χρησιμοποιούν τα χαρακτηριστικά των χρηστών $\{z_i\}$ (ένα τμήμα του προφίλ των χρηστών), τα χαρακτηριστικά των αντικειμένων $\{w_j\}$ (ένα τμήμα του προφίλ των αντικειμένων) και τις αλληλεπιδράσεις τους $\{x_{ij}\}$ για να υπολογίσουν την εκτίμηση για ένα αντικείμενο.

Τα υβριδικά RSs μπορούν επίσης να βελτιωθούν με χρήση τεχνικών βασισμένων στη γνώση όπως **case-based reasoning**, προκειμένου να αυξηθεί η ακρίβεια των συστάσεων και να ξεπεραστούν οι περιορισμοί των παραδοσιακών RSs π.χ. νέος χρήστης, νέα αντικείμενα. Το knowledge-based RS σύστημα Entree (Burke, 2000) χρησιμοποιεί γνώση για τα εστιατόρια, τις κουζίνες και τα φαγητά (π.χ. «τα θαλασσινά» δεν είναι «λαχανικά») για να προτείνει εστιατόρια στους χρήστες του. Το κύριο μειονέκτημα των βασισμένων στη γνώση εφαρμογών είναι η ανάγκη για την απόκτηση αυτής της γνώσης, που αποτελεί άλλωστε και περιορισμό πολλών εφαρμογών τεχνητής νοημοσύνης. Εντούτοις, τα βασισμένα στη γνώση RSs έχουν αναπτυχθεί για περιοχές εφαρμογής όπου η γνώση των περιοχών είναι εύκολα διαθέσιμη σε κάποια δομημένη και αναγνώσιμη από τους υπολογιστές μορφή π.χ. οντολογία. Τα συστήματα Quickstep και Foxtrot (Middleton et al., 2004) χρησιμοποιούν την οντολογία θέματος ερευνητικών εργασιών για να προτείνουν online ερευνητικά άρθρα. Τέλος, πολλές εργασίες όπως (Balabanovic & Shoham, 1997), (Melville et al., 2002), (Pazzani, 1999), (Soboroff & Nicholas, 1999) συγκρίνουν εμπειρικά την απόδοση των υβριδικών προσεγγίσεων με τις content-based και collaborative μεθόδους και αποδεικνύουν ότι οι πρώτες μπορούν να οδηγήσουν σε καλύτερες προτάσεις.

2.4 Το μοντέλο του e-customer

Ένα τυπικό RS «παρακολουθεί» τον e-customer και να καταγράφει όσες περισσότερες πληροφορίες για αυτόν μπορεί. Αυτές τις πληροφορίες, αφού τις επεξεργαστεί κατάλληλα, θα τις χρησιμοποιήσει στη φάση της ανάλυσης για την παραγωγή των recommendations. Στην προτεινόμενη υβριδική προσέγγιση υιοθετείται το **μοντέλο «Online Consumer's Model»** (Markellos et al., 2001), (Markellos et al., 2002), (Markellou et al., 2005b), (Markellou et al., 2006a), (Markellou et al., 2006a) το οποίο διακρίνει την αγοραστική διαδικασία στις **13 καταστάσεις (states)** που παρουσιάζονται στις επόμενες παραγράφους.

Κατάσταση 1: Πρόθεση αγορών online

Ένας πελάτης για να επιλέξει να αγοράσει προϊόντα από το Διαδίκτυο θα πρέπει αρχικά να έχει ικανοποιητικό κίνητρο όπως φαίνεται στην **κατάσταση 1** στο Σχήμα 17 που μοντελοποιεί όλα τα βήματα που αυτός ακολουθεί από τη στιγμή που συνειδητοποιεί την ανάγκη του για κάποιο προϊόν μέχρι να εισέλθει σε ένα e-shop. Το κίνητρο αυτό μπορεί να σχετίζεται με τη στάση του απέναντι στους υπολογιστές, στο Διαδίκτυο και στις νέες τεχνολογίες γενικότερα, την εμπιστοσύνη του σε αυτό το είδος συναλλαγών, την πληροφόρησή του για ηλεκτρονικά καταστήματα που εμπορεύονται τα προϊόντα που επιθυμεί, την προηγούμενη εμπειρία του από e-shops ή την εμπειρία φίλων και γνωστών (Blackwell et al., 2000).

Ειδικά η **εμπιστοσύνη (trust)** είναι μια έννοια που είναι παρούσα σε όλα τα βήματα της ηλεκτρονικής αγοραστικής διαδικασίας (Hertefelt, 2000), αποκτιέται σταδιακά μέσα από την αλληλεπίδραση του χρήστη με το κατάστημα και εύκολα μπορεί να χαθεί, όταν αυτό αποτύχει να εκπληρώσει τις προσδοκίες του. Άλλωστε οι ανταγωνιστές στο Διαδίκτυο είναι μόνο ένα «κλικ» μακριά! Γνωστές επωνυμίες, μάρκες και εμπορικά σήματα (brand names) δημιουργούν ως ένα βαθμό αυτόματη εμπιστοσύνη (Hurst, 1999). Άλλοι τρόποι για να καλλιεργηθεί ένα κλίμα εμπιστοσύνης στις σχέσεις ενός e-shop με τους πελάτες τους είναι: η παρουσίαση αναλυτικών πληροφοριών για την επιχείρηση, τα πολλαπλά κανάλια επικοινωνίας (π.χ. e-mail, chat rooms, bulleting boards, forums, κλπ.), οι πολιτικές ασφαλείας και διαφύλαξης της ιδιωτικότητας των χρηστών, τα πιστοποιητικά και οι σφραγίδες από διεθνείς οργανισμούς και φορείς, η στοχευόμενη διαφήμιση και το μάρκετινγκ, κλπ.

Σχήμα 17. Προσεγγίζοντας το e-shop.

Κατάσταση 2: Απόφαση αγορών online

Αφού ο χρήστης αποφασίσει να αγοράσει προϊόντα ή υπηρεσίες online (**κατάσταση 2**, Σχήμα 17), στο επόμενο βήμα προσπαθεί να εντοπίσει το κατάλληλο e-shop. Κάθε μέρα νέα καταστήματα «ανοίγουν» τις διαδικτυακές τους πύλες, εκατομμύρια προϊόντα είναι διαθέσιμα και ο ανταγωνισμός είναι πολύ σκληρός. Τα καταστήματα θα πρέπει να εκπληρώνουν σε ικανοποιητικό βαθμό τις ακόλουθες βασικές απαιτήσεις που θέτουν συνήθως οι πελάτες (Nielsen, 1993), (Cheskin Research & Studio Archetype Sapient, 1999):

- **αξιοπιστία** (σχετίζεται με την αναγνωρισιμότητα και την οικειότητα με την επωνυμία και το εμπορικό σήμα),

- **ευκολία στην πλοήγηση** (αφορά στη δυνατότητα του χρήστη να κατανοήσει, να χρησιμοποιήσει και να θυμηθεί στην επόμενη επίσκεψή του πώς να κινηθεί μέσα στο κατάστημα),
- **αισθητική και συνεπής παρουσίαση του περιεχομένου και των λειτουργιών,**
- **διαθεσιμότητα βασικών πληροφοριών για την επιχείρηση** πίσω από το e-shop,
- **ακρίβεια πληροφοριών και συνεχής ενημέρωσή τους,**
- **ασφάλεια συναλλαγών και προστασία προσωπικών δεδομένων.**

Οι καλύτερες τιμές, η ποιότητα των παρεχόμενων προϊόντων και υπηρεσιών, οι προσφορές, οι τεχνικές υποστήριξης, οι ευκολίες εξατομίκευσης, κλπ. είναι άλλοι παράγοντες που βοηθούν τους καταναλωτές στο να επιλέξουν ένα e-shop. Φυσικά δεν μπορούμε να αγνοήσουμε και το μικρό ποσοστό των χρηστών που επιλέγουν ένα κατάστημα κατά τύχη, συνήθως επειδή η μηχανή αναζήτησης που χρησιμοποιούν επιστρέφει την ηλεκτρονική του διεύθυνση με υψηλή προτεραιότητα στον κατάλογο των αποτελεσμάτων της.

Οι επιχειρήσεις από τη μεριά τους μπορούν να χρησιμοποιήσουν διάφορες τακτικές επικοινωνίας για να καταστήσουν τους χρήστες ενήμερους για το web site τους π.χ. καταχωρήσεις σε μηχανές αναζήτησης και δικτυακούς καταλόγους, μαζικές διαφημίσεις, «αναρτήσεις» ηλεκτρονικών πανό (banners), συνδέσεις (links) σε sites ή e-malls ή portals με σχετικά προϊόντα και διαφημιστικά e-mails. Το σημαντικότερο όμως ζήτημα για μια επιχείρηση είναι να προσδιορίσει μεταξύ των online χρηστών, την ομάδα πελατών στην οποία απευθύνεται και να την οδηγήσει στο site της. Ο στόχος είναι να τους δώσει το κίνητρο να επισκεφτούν το κατάστημα για πρώτη φορά. Για τους πελάτες με γνωστές διευθύνσεις (παραδοσιακό ή ηλεκτρονικό ταχυδρομείο) η άμεση επαφή είναι και η αποτελεσματικότερη. Διαφορετικά, όταν η επιχείρηση δε γνωρίζει ποιοι είναι οι πιθανοί πελάτες της, μια έρευνα για να προσδιορίσει τα χαρακτηριστικά τους είναι χρήσιμη και θα εξασφαλίσει ότι η τακτική διαφήμισης και μάρκετινγκ που θα επιλεγεί θα είναι η πλέον κατάλληλη και θα επιφέρει τα καλύτερα αποτελέσματα (Bellman et al., 1999), (Cuthbert, 2000), (Markellos et al., 2002).

Όλα τα παραπάνω ισχύουν κυρίως για τους νέους e-πελάτες. Για χρήστες με προγενέστερη εμπειρία στις ηλεκτρονικές αγορές, το ζήτημα της εμπιστοσύνης επεκτείνεται στην έννοια της **πίστης (loyalty)**. Σε αυτή την περίπτωση, το web site μπορεί να συγκεντρώσει και να αναλύσει, κάνοντας χρήση **τεχνικών Web Mining** (Kosala & Blockeel, 2000), (Eirinaki & Vazirgiannis, 2003), (Markellos et al., 2005c) όσο το δυνατόν περισσότερα στοιχεία για αυτούς, προκειμένου να εξατομικεύσει την εμπειρία τους όταν θα το ξαναεπισκεπτούν. Αυτό επιτρέπει στις επιχειρήσεις να προσαρμόσουν το περιεχόμενο και τις υπηρεσίες στα ενδιαφέροντα και στις ανάγκες των πελατών τους, καθώς και να υποστηρίξουν online κοινότητες χρηστών. Καταλήγοντας θα μπορούσαμε να πούμε ότι χωρίς σημαντικό κίνητρο και εμπιστοσύνη, ο εν δυνάμει πελάτης αποτυγχάνει να επιλέξει ένα e-shop και εγκαταλείπει την προσπάθεια (**τελική κατάσταση «Εγκατάλειψη»** του μοντέλου, Σχήμα 17).

Κατάσταση 3/8: Εγγραφή/Είσοδος

Η **εγγραφή (registration)** ενός νέου πελάτη μπορεί να πραγματοποιηθεί κατά την είσοδό του στο e-shop (**κατάσταση 3**, Σχήμα 18) ή ακόμα και πριν από τη διαδικασία αγοράς των προϊόντων (**κατάσταση 8** του ίδιου σχήματος). Η **πρώτη προσέγγιση** βοηθά το e-shop να συλλέξει χρήσιμες πληροφορίες για τον πελάτη που μπορεί να τις χρησιμοποιήσει με

διάφορους τρόπους για να παράγει εξατομίκευση π.χ. να τον καλωσορίσει στο κατάστημα, να του παρουσιάσει αρχικά όλα τα νέα προϊόντα που γνωρίζει ότι τον ενδιαφέρουν με βάση τις δηλωμένες προτιμήσεις και ανάγκες του, να επιταχύνει τη διαδικασία αγοράς με την αυτόματη συμπλήρωση πεδίων όπως το ονοματεπώνυμο, η διεύθυνση, κλπ. Το πλεονέκτημα αυτής της προσέγγισης είναι ότι κάθε πελάτης μπορεί να βλέπει διαφορετικό κατάστημα ως προς το περιεχόμενο, την παρουσίαση, την πλοήγηση, τις παρεχόμενες υπηρεσίες, κλπ., γεγονός που αυξάνει την ικανοποίηση και την εμπιστοσύνη του σε αυτό (Mobasher et al., 2000), (Mulvenna et al., 2000), (Kobsa et al., 2001).

Η **δεύτερη προσέγγιση**, όπου ο πελάτης εγγράφεται στο e-shop αμέσως πριν την αγορά των προϊόντων, του δίνει τη δυνατότητα να πλοηγείται ανώνυμα στις σελίδες του καταστήματος που επιθυμεί, να αναζητεί προϊόντα χρησιμοποιώντας τη μηχανή αναζήτησης και να τα τοποθετεί στο καλάθι αγορών του. Όταν ο πελάτης αποφασίσει να προχωρήσει στην ολοκλήρωση της παραγγελίας του καλείται να συμπληρώσει τις προσωπικές του πληροφορίες. Η τακτική αυτή δεν ενοχλεί τον πελάτη και του δημιουργεί ένα αίσθημα ελευθερίας, αλλά από την άλλη πλευρά δεν μπορεί το σύστημα να του παράσχει εξατομικευμένες υπηρεσίες αφού αν είναι νέος πελάτης δεν γνωρίζει τις ανάγκες του, ενώ αν είναι παλιός πελάτης δεν τον αναγνωρίζει. Κατά την εγγραφή του ένας νέος πελάτης πρέπει να συμπληρώσει συνήθως ένα online ερωτηματολόγιο με τα προσωπικά του στοιχεία. Οι ελάχιστες πληροφορίες που πρέπει να εισάγει αφορούν το ονοματεπώνυμο, το e-mail, το login, το password και την επιβεβαίωση του password του. Εκτός από αυτά τα στοιχεία το e-shop μπορεί επιπλέον να ζητάει από τον πελάτη να συμπληρώσει τη διεύθυνσή του, τη διεύθυνση αποστολής των προϊόντων (αν είναι διαφορετική), τα αναλυτικά στοιχεία απόδειξης ή τιμολόγησης και τα στοιχεία της πιστωτικής του κάρτας καθώς και άλλες δημογραφικές πληροφορίες όπως το φύλο του, την ηλικία του, το εισόδημά του, το επίπεδο εκπαίδευσής του, το επάγγελμά του, την οικογενειακή του κατάσταση, τις προτιμήσεις του, τα ενδιαφέροντά του, κλπ. Όσα από αυτά τα στοιχεία απαιτηθούν σε κάποιο επόμενο στάδιο της συναλλαγής συμπληρώνονται αυτόματα και δεν επιβαρύνεται ο πελάτης με την επαναπληκτρολόγησή τους. Πρέπει να σημειωθεί ότι οι πληροφορίες αυτές εξαρτώνται από το είδος των προϊόντων που εμπορεύεται το e-shop, τον τύπο των πελατών στους οποίους στοχεύει και τη στρατηγική που έχει υιοθετήσει.

Η **είσοδος (login)** είναι η διαδικασία κατά την οποία ο πελάτης αναγνωρίζεται από το σύστημα σε κάθε επόμενη επίσκεψή του (μετά δηλαδή από την εγγραφή του) και αποτελεί έναν αποτελεσματικό και ασφαλή τρόπο για να αντιμετωπισθεί το πρόβλημα του προσδιορισμού του ίδιου χρήστη που εισέρχεται στο site από διαφορετικά IPs ή να διακρίνει διαφορετικούς χρήστες που χρησιμοποιούν την ίδια IP. Η αναγνώριση του πελάτη γίνεται συνήθως με την εισαγωγή του login και του password του. Το σύστημα θα πρέπει να εξασφαλίζει μηχανισμούς άμεσης βοήθειας, όταν ο πελάτης ξεχάσει το login ή το password του π.χ. χρήση pass hint questions (δηλαδή ερωτημάτων για τα οποία ο χρήστης γνωρίζει την απάντηση και η απάντηση αυτή τον βοηθά να θυμηθεί το password). Γενικά ο στόχος στη φάση αυτή είναι οι πελάτες να συμπληρώνουν τις απαιτούμενες πληροφορίες με απλό, λειτουργικό και γρήγορο τρόπο π.χ. κάνοντας χρήση αναδιπλούμενων (drop-down) λιστών και προεπιλεγμένων απαντήσεων όπως η χώρα κατοικίας, κλπ. ώστε να υπάρξει η ελάχιστη δυνατή πληκτρολόγηση και να αποτραπούν πιθανά ορθογραφικά λάθη. Επιπλέον, λόγω του γεγονότος ότι οι πελάτες είναι απρόθυμοι να αποκαλύπτουν προσωπικές πληροφορίες και νιώθουν ανασφάλεια για το ποιος θα έχει πρόσβαση σε αυτά τα στοιχεία, τα e-shops πρέπει να παρέχουν σαφείς δηλώσεις στην πολιτική τους για τον τρόπο με τον οποίο

επεξεργάζονται τις συλλεχθείσες πληροφορίες και ποια μέτρα ασφάλειας έχουν λάβει (π.χ. με την τοποθέτηση σφραγίδων πιστοποίησης, αναλυτικές πληροφορίες για όλες τις σχετικές πολιτικές κοινοποίησης όπως ήδη αναφέραμε στην προηγούμενη ενότητα). Τέλος, για να πεισθεί ο πελάτης να εισάγει στο σύστημα τα προσωπικά του στοιχεία θα πρέπει το κατάστημα να τον πείσει ότι τα συγκριτικά πλεονεκτήματα θα είναι πολύ σημαντικά για αυτόν.

Σχήμα 18: Μέσα στο e-shop.

Κατάσταση 4: Κατάλογος

Μετά την είσοδό του στο e-shop ο πελάτης μπορεί να αναζητήσει τα προϊόντα που θα τοποθετήσει στο **καλάθι αγορών (shopping cart)** είτε πλοηγούμενος μέσα στον **κατάλογο (online catalog)** ή μέσω της διαδικασίας **αναζήτησης (searching)**. Η περιήγηση του πελάτη στον **online κατάλογο (κατάσταση 5, Σχήμα 18)** παρουσίασης των προϊόντων και των υπηρεσιών του e-shop αποτελεί βασικό συστατικό στοιχείο της αγοραστικής διαδικασίας αφού ουσιαστικά ο πελάτης καλείται να επιλέξει τι θα αγοράσει. Ο κατάλογος πρέπει να βρίσκεται σε περίοπτη θέση στο e-shop και να είναι εύκολα προσπελάσιμος από όλες τις σελίδες του, δίνοντας την ευκαιρία στους πελάτες να δουν τα προϊόντα και τις υπηρεσίες που μπορούν να επιλέξουν. Επίσης, πρέπει να επιτρέπει τη διαδραστική επαφή, ώστε ο πελάτης να μπορεί να ενημερωθεί για τα χαρακτηριστικά των προϊόντων που τον ενδιαφέρουν, δηλαδή με ένα απλό «κλικ» να μπορεί να δει περισσότερες πληροφορίες για το προϊόν (Koch & Turk, 1997). Οι βασικές απαιτήσεις που πρέπει να ικανοποιεί ο κατάλογος είναι: πλήθος διαθέσιμων προϊόντων, οργάνωση και κατηγοριοποίηση των προϊόντων, πληροφόρηση, πλοήγηση, παρουσίαση, αναζήτηση, προσθήκη στο καλάθι αγορών. Ιδιαίτερα, η παρουσίαση κάθε προϊόντος αποτελεί καθοριστικό παράγοντα για να μετατραπεί ένας επισκέπτης του e-shop σε πελάτη. Η επαρκής πληροφόρηση το βοηθάει να ενημερωθεί και να καταλήξει σε μια τεκμηριωμένη απόφαση. Επειδή δεν μπορεί να έχει άμεση επαφή με το αντικείμενο που του προτείνεται να αγοράσει, εκτός από τη γραπτή περιγραφή του, θα πρέπει επιπλέον να παρέχεται οπτική παρουσίαση του προϊόντος π.χ. εικόνα, διάγραμμα, βιντεοκλίπ, κλπ. Συγκεκριμένα, μια ολοκληρωμένη παρουσίαση ενός προϊόντος πρέπει να περιλαμβάνει: τον κωδικό, την κατηγορία στην οποία ανήκει, τα χαρακτηριστικά, την περιγραφή, την εικόνα, την τιμή, τη διαθεσιμότητα, το χρόνο αποστολής, κλπ. Άλλες πιο

προχωρημένες δυνατότητες ενός καταλόγου είναι: η εφαρμογή τεχνικών cross-selling και up-selling, οι κριτικές, τα σχόλια, οι βαθμολογίες, τα δημοφιλή προϊόντα, οι συγκρίσεις των προϊόντων, κλπ. (Markellou et al., 2005d).

Κατάσταση 5: Αναζήτηση

Γενικά οι χρήστες του Διαδικτύου είναι εξοικειωμένοι με την έννοια της αναζήτησης και τη λειτουργία των διαθέσιμων μηχανών (π.χ. Google, Altavista, Robby, κλπ.). Ανεξάρτητα όμως από το πόσο καλά δομημένος και αποδοτικός είναι ο κατάλογος των προϊόντων, ένα e-shop πρέπει επίσης να παρέχει λειτουργία **αναζήτησης (κατάσταση 4, Σχήμα 18)** διαθέσιμη και ορατή σε όλες τις σελίδες του καταστήματος, έτσι ώστε οι πελάτες να μπορούν να προσφύγουν σε αυτή όποτε την έχουν ανάγκη. Αποτελεί μάλιστα πλεονέκτημα του e-shop η παροχή εναλλακτικών τρόπων αναζήτησης όπως: γρήγορη αναζήτηση (fast search), απλή αναζήτηση (simple search) και εκτεταμένη αναζήτηση (advanced search). Για τα προϊόντα τα οποία δε βρίσκει ο χρήστης κατά την αναζήτηση είναι χρήσιμο να μπορεί να υποβάλλει σχετικές ερωτήσεις στο τμήμα υποστήριξης του e-shop ή μέσω e-mail, ενώ και η παρουσίαση οδηγιών για το πώς μπορεί να χρησιμοποιήσει τις διαθέσιμες μηχανές αναζήτησης του e-shop εκτιμάται θετικά. Μια άλλη γενική συμβουλή είναι ότι η αναζήτηση πρέπει να λειτουργεί με τον τρόπο που οι χρήστες αναμένουν. Αυτό σημαίνει ότι ο μηχανισμός πρέπει να είναι ανεκτικός σε ορθογραφικά λάθη, να επιτρέπει τα συνώνυμα και τις λέξεις κλειδιά και να επιστρέφει εύχρηστες σελίδες αποτελεσμάτων και συνδέσεις με τις σελίδες των προϊόντων. Όταν ο μηχανισμός δεν επιστρέφει αποτελέσματα ή αυτά δεν είναι ικανοποιητικά θα πρέπει να δίνεται η δυνατότητα στους χρήστες να επεξεργάζονται τις ερωτήσεις-αναζητήσεις τους μέχρι να εντοπίσουν τα προϊόντα που επιθυμούν (υπό την προϋπόθεση βέβαια ότι αυτά είναι διαθέσιμα στο e-shop).

Οι **καταστάσεις 4 και 5** είναι μεγάλης σπουδαιότητας, όχι μόνο επειδή η επιλογή των προϊόντων αποτελεί αναπόσπαστο τμήμα κάθε διαδικασίας αγοράς, αλλά και επειδή καθορίζει την εντύπωση του χρήστη για το e-shop και μια πρώτου-επιπέδου αξιολόγησή του. Οι χρήστες καταναλώνουν το μεγαλύτερο μέρος του χρόνου τους κατά τη διάρκεια αυτών των διαδικασιών και χρησιμοποιούν την ανατροφοδότηση για να αποφασίσουν εάν θα συνεχίσουν με την πληρωμή ή θα εγκαταλείψουν το καλάθι αγορών τους και θα εξέλθουν από το e-shop. Τα e-shops πάσχουν από αυτό το φαινόμενο, αφού το Διαδίκτυο είναι ένας απρόσωπος, απομακρυσμένος και εξαιρετικά ανταγωνιστικός κόσμος, ο οποίος μπορεί να είναι ευλογία ή κατάρα. Για να είναι το e-shop επιτυχημένο θα πρέπει οι πελάτες να ικανοποιηθούν στο μεγαλύτερο βαθμό. Αποτυχία να εντοπισθούν τα προϊόντα σημαίνει ένας πελάτης λιγότερος και είναι δύσκολο να υπολογισθούν οι επιπτώσεις και η έκταση της πιθανής δυσφήμισης για το e-shop.

Κατάσταση 6: Καλάθι αγορών

Το **καλάθι αγορών** που μπορεί να αναφέρεται και σαν **shopping cart, shopping basket, shopping bag, shopping sled** ή **wheelbarrow** αφορά σε μια ενιαία σελίδα του e-shop που εμφανίζει όλα τα προϊόντα που ο πελάτης έχει επιλέξει να αγοράσει (Markellou, 2001). Αποτελεί **μεταφορά (metaphor)** του καροτσιού που χρησιμοποιούμε στα παραδοσιακά σουπερ-μάρκετ και στοχεύει στη χρησιμοποίηση της γνώσης που έχουν αποκτήσει οι πελάτες από την καθημερινή τους ζωή, στη διαδικασία online αγορών (Bidigare, 2000), (Hudson, 2000). Με την προσθήκη ενός αντικειμένου στο **καλάθι αγορών (κατάσταση 6,**

Σχήμα 18), ο πελάτης συνήθως μεταφέρεται στη σελίδα με τα περιεχόμενά του. Εκεί έχει τη δυνατότητα να διαγράψει ένα ή περισσότερα προϊόντα, να αλλάξει τις ποσότητες τους και να υπολογίσει εκ νέου το συνολικό κόστος αυτόματα. Επιπλέον, πρέπει να υπάρχει δυνατότητα «επιστροφής στις αγορές» που θα μεταφέρει τον πελάτη στο σημείο που βρισκόταν πριν μεταφερθεί στα περιεχόμενα του καλαθιού. Αυτή η λογική απεικονίζει ακριβώς την πραγματικότητα π.χ. όταν περπατάμε στους διαδρόμους ενός σουπερ-μάρκετ επιλέγουμε προϊόντα, τα τοποθετούμε στο καλάθι μας, και συνεχίζουμε να περπατάμε έως ότου να βρούμε όλα τα προϊόντα που έχουμε καταγράψει στη λίστα μας. Μάλιστα, η πραγματική κατάσταση παρουσιάζει ένα μειονέκτημα: δεν μπορούμε εύκολα να υπολογίσουμε το συνολικό κόστος του περιεχομένου του καλαθιού μας όπως στα e-shops, όπου αυτό γίνεται με το πάτημα ενός κουμπιού.

Ένα καλά σχεδιασμένο καλάθι αγορών μπορεί να διευκολύνει τις συναλλαγές και να ενθαρρύνει τους πελάτες να αγοράσουν περισσότερο, διαφορετικά μπορεί να τους μπερδέψει με αποτέλεσμα να ματαιώσουν την ολοκλήρωση της παραγγελίας τους. Σύμφωνα με μια έρευνα, το 78% των online πελατών εγκαταλείπουν το καλάθι αγορών. Από αυτούς το 55% εγκαταλείπει τη συναλλαγή πριν να εισέλθει στην κατάσταση «Ταμείο», ενώ το 32% εγκαταλείπει τη συναλλαγή στο σημείο που πρέπει να συμπληρώσει τα στοιχεία παράδοσης και πληρωμής της παραγγελίας (Rohn, 1998). Οι λόγοι για τους οποίους οι πελάτες εγκαταλείπουν το καλάθι αγορών ποικίλλουν και περιλαμβάνουν μεταξύ άλλων: ανεπαρκείς πληροφορίες, φτωχή πλοήγηση, δυσκολία στο χειρισμό, περιορισμένες λειτουργίες, δυσνόητα κουμπιά ή εικονίδια, χρονοβόρα διαδικασία στο ταμείο, χαμηλή ποιότητα της διεπαφής χρήστη, κλπ. (Hill, 2001). Σύμφωνα με τον Rewick (2000), το κλειδί είναι η κατανόηση της ψυχολογίας των πελατών, των αναγκών, των απαιτήσεων και των ενδιαφερόντων τους και πώς αυτοί οι παράγοντες επηρεάζονται από το σχεδιασμό του e-shop. Εντούτοις, η εγκατάλειψη του καλαθιού αγορών δεν σημαίνει πάντα ύπαρξη σχεδιαστικών προβλημάτων, αφού συχνά εξαρτάται από τον τρόπο που οι πελάτες αγοράζουν online (Smith, 2000). Άλλωστε τα περιεχόμενα του καλαθιού, ακόμη και αν ο πελάτης το έχει εγκαταλείψει, δηλώνουν το ενδιαφέρον του για τα συγκεκριμένα προϊόντα.

Το καλάθι αγορών είναι λοιπόν ένα χρήσιμο εργαλείο για να τοποθετήσουν οι πελάτες τα προϊόντα που επιθυμούν να αγοράσουν. Στην ουσία τα επιλεγμένα προϊόντα εμφανίζονται συγκεντρωμένα σε μια σελίδα του e-shop, η οποία πρέπει να είναι εύκολα προσβάσιμη και να επιτρέπει αλλαγές στα περιεχόμενά της πριν ο πελάτης ολοκληρώσει και υποβάλλει την παραγγελία. Τα πλεονεκτήματά του είναι σημαντικά και σχετίζονται με: γρήγορες και εύκολες συναλλαγές, οικειότητα, απλότητα και σαφήνεια υπό τη μορφή σταδιακών βημάτων. Πολλά e-shops χρησιμοποιούν μια απλουστευμένη έκδοση του καλαθιού αγορών, γνωστή ως persistent shopping cart, το οποίο είναι ορατό από κάθε σελίδα του καταστήματος και παρέχει έναν γρήγορο τρόπο ελέγχου των περιεχομένων του (προϊόντα, συνολικό κόστος).

Η μεταφορά του καλαθιού αγορών βοηθά τους πελάτες διαισθητικά να καταλάβουν τη βασική λειτουργία του (Nielsen, 2000), υπό την έννοια ότι είναι εξοικειωμένοι με αυτή. Παρά την επιτυχία και την ευρεία αποδοχή του, η ίδια η μεταφορά μπορεί να προκαλέσει σύγχυση. Η λειτουργία αλλαγής των ποσοτήτων των προϊόντων και κυρίως του μηδενισμού της ποσότητας κάποιου προκειμένου να το απομακρύνουμε δεν είναι μια ρουτίνα με την οποία ο πελάτης είναι εξοικειωμένος στην καθημερινότητά του (π.χ. στο σούπερ-μάρκετ). Για να μην αναφέρουμε ότι στην πραγματικότητα για να πληρώσουμε τα προϊόντα του καλαθιού μας πρέπει στην ουσία να αδειάσουμε το καλάθι μας από τα προϊόντα του! Παρόλα αυτά τα

προβλήματα, το καλάθι αγορών χρησιμοποιείται ευρέως από τα e-shops αφού η υψηλή οικειότητα που έχουν οι πελάτες με αυτό ελαχιστοποιεί τα περιστατικά κακής χρήσης του.

Κατάσταση 7: Ταμείο

Όταν ο πελάτης επιλέξει τα προϊόντα που επιθυμεί να αγοράσει και τα τοποθετήσει στο καλάθι αγορών του (οι **καταστάσεις 4, 5 και 6** μπορούν να επαναληφθούν αρκετές φορές), τότε μπορεί να προχωρήσει στο **ταμείο (κατάσταση 7, Σχήμα 18)**. Στη φάση αυτή καλείται να συμπληρώσει τις πληροφορίες που σχετίζονται με την παράδοση των προϊόντων όπως τη διεύθυνσή του, την ημερομηνία παράδοσης, τις λεπτομέρειες συσκευασίας ή άλλες προαιρετικές δυνατότητες αποστολής, κλπ. Εάν ο χρήστης δεν έχει εγγραφεί ακόμα στο e-shop ή δεν έχει κάνει login, το κατάστημα του ζητά να συμπληρώσει τα πρόσθετα στοιχεία που διαφορετικά θα συμπληρώνονταν αυτόματα π.χ. το ονοματεπώνυμό του, τη διεύθυνση, κλπ. Είναι σημαντικό σε αυτό το σημείο η πληκτρολόγηση να περιοριστεί στο ελάχιστο και όλα τα προαιρετικά πεδία να σημειωθούν με σαφήνεια και συνέπεια. Αυτό ισχύει επίσης και για ορισμένες πληροφορίες που απαιτούνται κατά τη διάρκεια της φάσης **πληρωμής (κατάσταση 9)** που ακολουθεί την παρούσα φάση.

Κατάσταση 8: Πληρωμή

Η **πληρωμή (κατάσταση 9, Σχήμα 18)** αποτελεί το τελευταίο στάδιο της παραμονής του πελάτη μέσα στο e-shop και περιλαμβάνει την υποβολή των στοιχείων για τη χρέωσή του με το κόστος της παραγγελίας του (Pilioura, 1998). Ο πελάτης καλείται να επιλέξει μια από τις διαθέσιμες δυνατότητες πληρωμής (π.χ. με πιστωτική κάρτα, με αντικαταβολή, κλπ.) και να υποβάλλει όλες τις απαραίτητες πληροφορίες έτσι ώστε το κατάστημα να λάβει το αντίστοιχο χρηματικό ποσό. Η πληρωμή απαιτεί απλότητα, σαφήνεια, πιστοποίηση, διαβεβαίωση και τέλος αίσθηση επαγγελματισμού και ασφάλειας της όλης συναλλαγής και επικοινωνίας. Η ασφάλεια στις συναλλαγές δεν είναι διαπραγματεύσιμη. Πρέπει να υπάρχουν αναλυτικές πληροφορίες για τα μέτρα που λαμβάνονται για την εγγύησή της και να είναι διαθέσιμες σε κάθε σελίδα της διαδικασίας πληρωμής για όποιον πελάτη επιθυμεί να τις διαβάσει. Αυτή η φάση θέτει επίσης και το ζήτημα της e-εμπιστοσύνης για το οποίο έχουμε μιλήσει σε προηγούμενη ενότητα. Με την επιτυχή ολοκλήρωση της πληρωμής, ο πελάτης θα πρέπει να ενημερωθεί από το e-shop για αυτό, συνήθως μέσω μιας **σελίδας επιβεβαίωσης**. Η σελίδα αυτή θα πρέπει να περιέχει τα προϊόντα που παραγγέλθηκαν, ένα μοναδικό κωδικό αριθμό που προσδιορίζει την παραγγελία, καθώς και το χρόνο παράδοσής της. Χρησιμοποιώντας τον κωδικό αριθμό ο πελάτης μπορεί να ενημερωθεί για την τρέχουσα κατάστασή της οποιαδήποτε στιγμή μετά την υποβολή της και πριν να παραλάβει τα προϊόντα. Αυτές οι πληροφορίες μπορούν εναλλακτικά να σταλούν στους πελάτες και μέσω e-mail.

Σε πολλές περιπτώσεις, το e-shop επιτρέπει στους χρήστες να επιλέξουν τον τρόπο με τον οποίο θα λαμβάνουν επιβεβαίωση. Για παράδειγμα οι αεροπορικές εταιρίες συχνά ρωτούν τους πελάτες τους αν θέλουν να παραλάβουν το δρομολόγιο τους με fax, με e-mail ή με συμβατικό ταχυδρομείο. Συχνά η σελίδα επιβεβαίωσης προσφέρει στους πελάτες την επιλογή να εγγραφούν ως μέλη στο e-shop α) αν αυτό δεν είναι απαραίτητη προϋπόθεση για τη διενέργεια των συναλλαγών τους και β) αν φυσικά δεν είναι ήδη εγγεγραμμένοι. Προσοχή πρέπει να δοθεί στο γεγονός ότι στις περιπτώσεις αυτές οι χρήστες ήδη έχουν καταχωρήσει όλα τα προσωπικά στοιχεία τα οποία απαιτούνται. Έτσι, το μόνο που χρειάζεται να επιλέξουν είναι ταυτότητα και κωδικό πρόσβασης. Η αποθήκευση του αριθμού πιστωτικής

κάρτας θα πρέπει να είναι προαιρετική. Επίσης, πρέπει να σημειωθούν με σαφήνεια τα πλεονεκτήματα που προσφέρει η εγγραφή μέλους στο συγκεκριμένο e-shop και να εξασφαλισθεί σύνδεσμος προς τη σελίδα που περιγράφει την επικοινωνιακή πολιτική του.

Ένα άλλο χαρακτηριστικό που εκτιμάται ιδιαίτερα από τους πελάτες είναι η εξασφάλιση ενός μηχανισμού αλλαγής-ακύρωσης μιας ήδη ενεργής παραγγελίας. Οι πελάτες θα πρέπει να μπορούν να αλλάζουν ή ακόμα και να ακυρώνουν μια παραγγελία πριν φυσικά αυτή παραδοθεί. Με αυτό τον τρόπο προκαταβάλλεται θετικά ο πελάτης, ο οποίος νιώθει μεγαλύτερο βαθμό ελευθερίας στις συναλλαγές του με το e-shop. Για να ικανοποιηθεί αυτό θα πρέπει ο πελάτης να διαθέτει τον κωδικό επιβεβαίωσης της παραγγελίας και να παρέχεται η δυνατότητα αλλαγής της είτε μέσω του Web είτε μέσω τηλεφωνικής επικοινωνίας. Τέλος, είναι προφανές ότι οι παραγγελίες πρέπει να διεκπεραιωθούν το συντομότερο δυνατόν και φυσικά να μη ξεπερνούν το χρόνο παράδοσης που το κατάστημα έχει υποσχεθεί στον πελάτη. Έτσι ακόμα και ο πιο δύσπιστος πελάτης θα μείνει ικανοποιημένος. Είναι περίεργο το γεγονός ότι πολλές επιχειρήσεις ξοδεύουν σημαντικά ποσά για την προσέλκυση πελατών στο e-shop τους, χωρίς όμως να δείχνουν αντίστοιχο ενδιαφέρον για τον τρόπο με τον οποίο θα τους διατηρήσουν την πελατειακή τους βάση.

Κατάσταση 10: Παρακολούθηση παραγγελίας

Το τυπικό σενάριο μετά την υποβολή της παραγγελίας του πελάτη στο e-shop, είναι να φύγει από αυτό και να μπει σε μια κατάσταση αναμονής των προϊόντων. Υπάρχουν περιπτώσεις που η φύση των προϊόντων επιτρέπει την απευθείας παράδοσή τους (μέσω του Web) όταν π.χ. αφορούν σε κάποιο λογισμικό ή πρόσβαση σε κάποια υπηρεσία, κλπ. Σε διαφορετική περίπτωση, όπως αναφέρθηκε σε προηγούμενη ενότητα, το κατάστημα θα πρέπει να παρέχει στον πελάτη εκτίμηση του χρόνου παράδοσης, καθώς επίσης και έναν κωδικό αριθμό με τον οποίο θα μπορεί να παρακολουθεί σε ποιο στάδιο βρίσκεται η παραγγελία του. Στην περίπτωση που η παραγγελία δεν παραδίδεται μέσα στον προκαθορισμένο χρόνο, ο πελάτης μπορεί να **παρακολουθήσει τη διαδικασία (κατάσταση 10, Σχήμα 19)**. Υπάρχει επίσης περίπτωση ο πελάτης να μπει σε αυτή την κατάσταση απλώς για να επιβεβαιώσει ότι όλα προχωρούν σύμφωνα με τον αρχικό προγραμματισμό.

Σχήμα 19: Αφήνοντας το e-shop.

Καταστάσεις 11 & 12: Παραλαβή & αξιολόγηση προϊόντων

Μετά την **παραλαβή των προϊόντων (κατάσταση 11)** ο πελάτης προχωρά στην **αξιολόγησή τους (κατάσταση 12)** προκειμένου να ελέγξει την ποιότητά τους. Εάν το αποτέλεσμα της αξιολόγησης είναι αρνητικό, η εντύπωση αυτή συνδέεται απευθείας με το e-shop. Το κατάστημα χάνει αυτόν τον πελάτη και αυτό μεταδίδεται και σε όλους εκείνους με τους οποίους μοιράστηκε αυτή τη δυσάρεστη εμπειρία.

Κατάσταση 13: Υποστήριξη πελάτη

Υπάρχουν επίσης πολλές περιπτώσεις που τα προϊόντα απαιτούν **μεταγοραστική υποστήριξη** όπως το υλικό ή το λογισμικό των υπολογιστών (**κατάσταση 13**). Σε τέτοιες περιπτώσεις, η γενική εντύπωση των πελατών μπορεί να επαναξιολογηθεί στο μέλλον με βάση και τα αποτελέσματα αυτής της υποστήριξης. Η εμπειρία μάλιστα έχει δείξει ότι ακόμα και αν το ίδιο το προϊόν είναι ικανοποιητικό, η έλλειψη επαρκούς υποστήριξης οδηγεί σε απροθυμία να επαναληφθούν ανάλογες αγορές και τελικά σε αποτυχία των καταστημάτων να κρατήσουν τους πελάτες τους.

2.5 Ένα RS μοντέλο που βασίζεται σε δεδομένα χρήσης ιστού και στην οντολογία του e-shop

2.5.1 Ορισμός του προβλήματος παραγωγής προτάσεων σε ένα e-shop που νοικιάζει κινηματογραφικές ταινίες

Θεωρούμε ένα e-shop το οποίο νοικιάζει κινηματογραφικές ταινίες στους πελάτες του. Το πρόβλημα που μελετάται στις επόμενες ενότητες αφορά στην παραγωγή και στην παρουσίαση στους χρήστες εξατομικευμένων recommendations για ταινίες που δεν έχουν δει. Συγκεκριμένα το πρόβλημα μπορεί να ορισθεί ως εξής:

Έστω U το σύνολο όλων των χρηστών και M το σύνολο όλων των πιθανών αντικειμένων, στην περίπτωση μας ταινιών που μπορούν να προταθούν (το πρόβλημα εκτός από ταινίες μπορείς να αφορά και συστάσεις για βιβλία, CDs, εστιατόρια ή ακόμα και web pages). Τα σύνολα U (των πιθανών χρηστών) και M (των πιθανών ταινιών) μπορεί να είναι πολύ μεγάλα ανάλογα με το σύστημα και την εφαρμογή.

Ορίζουμε σαν f τη συνάρτηση που υπολογίζει τη «χρησιμότητα» μιας ταινίας m σε ένα χρήστη u . Τότε:

$$f: U \times M \rightarrow R$$

όπου R είναι το παραγόμενο σύνολο από ταξινομημένα recommendations. Η ταξινόμηση ακολουθεί τη λογική της παρουσίασης πρώτων των recommendations με τη μεγαλύτερη πιθανότητα να τα επιλέξει ο χρήστης.

Για κάθε χρήστη $u \in U$ θέλουμε να επιλέξουμε εκείνη την ταινία $m' \in M$ που μεγιστοποιεί τη συνάρτηση f :

$$\forall u \in U, m'_u = \arg \max_{m \in M} f(u, m)$$

Συνήθως, στα RSs η συνάρτηση f είναι μια **εκτίμηση (βαθμολογία)** που δηλώνει κατά πόσο ένας συγκεκριμένος χρήστης «εκτίμησε» ένα συγκεκριμένο αντικείμενο. Για παράδειγμα ο χρήστης «Πέννυ» έδωσε στην κινηματογραφική ταινία με τίτλο «Ο Άρχοντας το Δακτυλιδιών – Η Επιστροφή του Βασιλιά» το βαθμό 9 (σε μια κλίμακα από 1 έως 10, όπου το 10 είναι το άριστα). Σε άλλες περιπτώσεις η f μπορεί να είναι μια **συνάρτηση κέρδους (profit function)**, ενώ ανάλογα με την εφαρμογή, μπορεί να προσδιοριστεί από το χρήστη ή να υπολογισθεί από την ίδια την εφαρμογή.

Κάθε χρήστης, δηλαδή κάθε στοιχείο του συνόλου U καθορίζεται από το **προφίλ (profile)** του. Αυτό περιλαμβάνει τα ιδιαίτερα χαρακτηριστικά του όπως το ονοματεπώνυμό του, την ηλικία του, το φύλο του, το επάγγελμά του, το μορφωτικό του επίπεδο, την οικογενειακή του κατάσταση, κλπ. Στην απλούστερη περίπτωση, περιλαμβάνει μόνο το userID του χρήστη που αποτελεί ένα μοναδικό στοιχείο με το οποίο αναγνωρίζεται η ταυτότητά του.

Ομοίως, κάθε στοιχείο του συνόλου των ταινιών M καθορίζεται από μια σειρά χαρακτηριστικών. Κάθε ταινία αντιπροσωπεύεται από το monieID αλλά και από τον τίτλο της, την κατηγορία της, το σκηνοθέτη της, τον παραγωγό της, τη χρονολογία έκδοσής της, το κάστ της, τα βραβεία της, τη διάρκειά της, το κοινό της, την περίληψή της, κλπ.

Το πρόβλημα που παρουσιάζεται είναι ότι η f δεν καθορίζεται σε ολόκληρο το χώρο $U \times M$ αλλά σε κάποιο υποσύνολό του. Αυτό σημαίνει ότι θα πρέπει να συνάγουμε τις τιμές της f για ολόκληρο το χώρο $U \times M$. Στην περίπτωση μας, η f αντιπροσωπεύει βαθμολογίες και αρχικά καθορίζεται μόνο για τις ταινίες που έχουν ήδη βαθμολογηθεί από τους χρήστες. Ο Πίνακας 7 παρουσιάζει ένα παράδειγμα συσχετίσεων χρηστών-ταινιών, όπου οι χρήστες έχουν βαθμολογήσει ταινίες σε μια κλίμακα από 1 έως 10 (το $\emptyset$ σημαίνει ότι η συγκεκριμένη ταινία δεν έχει βαθμολογηθεί από το χρήστη).

	«Ο Άρχοντας των Δακτυλιδιών – Η Επιστροφή του Βασιλιά»	«Ο Χάρυ Πότερ και η Φιλοσοφική Λίθος»	«Ο Ιντιάνα Τζόουνς και η Χαμένη Κιβωτός»	«Ο Βράχος»
Πέννυ	9	8	10	9
Κώστας	9	6	8	7
Βλάσης	2	1	$\emptyset$	5
Θεοφανία	$\emptyset$	7	$\emptyset$	8

Πίνακας 7. Βαθμολογήσεις ταινιών από χρήστες.

Η μηχανή συστάσεων θα πρέπει συνεπώς να είναι σε θέση να υπολογίσει (να προβλέψει) τις βαθμολογίες για τις ταινίες που δεν έχουν εκτιμηθεί από τους χρήστες. Αυτός ο υπολογισμός μπορεί να γίνει με χρήση **ευριστικών μεθόδων (heuristics)** που καθορίζουν την f και εμπειρικά επικυρώνουν την απόδοσή της ή με **υπολογισμό** της ώστε να βελτιστοποιηθεί ένα ορισμένο κριτήριο απόδοσης π.χ. το μέσο τετραγωνικό λάθος. Όταν υπολογιστούν οι

άγνωστες βαθμολογίες, τότε μπορούν να παραχθούν recommendations ταινιών και στη συνέχεια να παρουσιαστούν σε ένα χρήστη. Από το σύνολο των παραγόμενων recommendations επιλέγεται για παρουσίαση στο χρήστη εκείνη η ταινία με την **υψηλότερη βαθμολογία**. Εναλλακτικά, μπορούμε να προτείνουμε τις **N καλύτερες ταινίες (top N)** σε έναν χρήστη ή να αντιστοιχίσουμε ένα σύνολο χρηστών σε κάποια ταινία.

2.5.2 Περιορισμοί και προβλήματα των RSs τεχνικών

Οι τεχνικές **content-based** και **collaborative-based** που χρησιμοποιούνται συνήθως για την παραγωγή των συστάσεων αντιμετωπίζουν σημαντικά προβλήματα και περιορισμούς που σχετίζονται με:

- **Ελλιπή δεδομένα.** Όταν τα δεδομένα που χρησιμοποιεί το RS είναι ελλιπή, τότε είναι αναγκαίο να γίνουν μια σειρά από υποθέσεις πριν την εφαρμογή των τεχνικών εξόρυξης γνώσης. Αυτό έχει σαν αποτέλεσμα ορισμένα από τα πρότυπα (patterns) που παράγονται να μην είναι κατάλληλα ή σωστά.
- **Περιορισμένα δεδομένα.** Το πρόβλημα αυτό σχετίζεται τόσο με τους χρήστες όσο και με τα αντικείμενα. Για να είναι επιτυχημένο ένα RS απαιτείται σημαντικός αριθμός χρηστών να έχουν βαθμολογήσει σημαντικό αριθμό αντικειμένων. Ένα παράδειγμα που φανερώνει το πρόβλημα είναι όταν πολλές ταινίες έχουν βαθμολογηθεί από λίγους χρήστες, τότε αυτές θα προτείνονται σπάνια, ακόμα και αν έχουν λάβει υψηλές βαθμολογίες.
- **Ανακριβή δεδομένα.** Το πρόβλημα αυτό σχετίζεται με τον ίδιο το χρήστη και τα δεδομένα που εισάγει στο RS είτε με ρητό είτε με υπονοούμενο τρόπο. Στην πρώτη περίπτωση έχουμε το παράδειγμα του χρήστη που συμπληρώνει τη φόρμα εγγραφής με ψεύτικα ή ανακριβή στοιχεία που οδηγούν σε αποτυχημένες μελλοντικές συστάσεις. Στη δεύτερη περίπτωση μπορεί το πρόβλημα να δημιουργείται χωρίς την υπαιτιότητα του χρήστη, αν και το αποτέλεσμα θα είναι το ίδιο, δηλαδή «άχρηστα» recommendations. Αυτό συμβαίνει όταν ο χρήστης χαθεί μέσα στο site, οπότε τα click streams που καταγράφονται στο αρχείο log είναι πλασματικά. Αυτό γίνεται περισσότερο προβληματικό όταν το web site είναι άστοχα σχεδιασμένο και οι χρήστες επισκέπτονται αυτόκλητες σελίδες κάνοντας τις να φαίνονται δημοφιλείς.
- **Νέες σελίδες.** Το πρόβλημα προκύπτει όταν νέες σελίδες που παρουσιάζουν νέα αντικείμενα προστίθενται στο web site. Επειδή αυτές τις σελίδες δεν τις έχουν ακόμα επισκεφτεί οι χρήστες και φυσικά δεν έχουν βαθμολογήσει τα προϊόντα τους, το RS δεν τα συνιστά, ακόμα και στην περίπτωση που αυτά σχετίζονται με το προφίλ του χρήστη και πιθανότατα θα τα έβρισκε ενδιαφέροντα. Παράλληλα, όσο περισσότερο μια σελίδα προτείνεται, τόσο περισσότερο την επισκέπτονται οι χρήστες αυξάνοντας την πιθανότητα να είναι υποψήφια για μελλοντικές συστάσεις.
- **Νέοι χρήστες.** Όταν ένας νέος χρήστης επισκέπτεται το web site, επειδή το ιστορικό του είναι κενό είναι δύσκολο να συσχετισθεί με άλλους χρήστες. Επίσης, σε πολλές περιπτώσεις πρέπει να βαθμολογήσει ένα σημαντικό αριθμό αντικειμένων για να μπορέσει το RS να κατανοήσει τις προτιμήσεις του και να του παρουσιάσει ένα αξιόπιστο σύνολο από recommendations.

- **«Cold start problem».** Το πρόβλημα αυτό έχει ονομαστεί και σαν **«day one»** λόγω του γεγονότος ότι την πρώτη μέρα της υπηρεσίας του, το σύστημα δεν έχει διαθέσιμα δεδομένα για να ξεκινήσει να παράγει προτάσεις. Ιδιαίτερα οι CF τεχνικές παρουσιάζουν φτωχά αποτελέσματα για τα νέα προϊόντα π.χ. προϊόντα που δεν έχουν βαθμολογηθεί ακόμα από κάποιον πελάτη. Από την άλλη μεριά όσο ο αριθμός των χρηστών που επισκέπτονται το e-shop μεγαλώνει, τόσο πληθαίνουν οι συγκρίσεις που πρέπει να γίνουν μεταξύ όλων των χρηστών, γεγονός που επηρεάζει σοβαρά το scalability του RS.
- **«Banana problem».** Τα RSs που εξάγουν συσχετίσεις από τις συναλλαγές των χρηστών είναι ιδιαίτερα «ευαίσθητα» σε σχέση με τη συχνότητα των βαθμολογιών συγκεκριμένων αντικειμένων. Για παράδειγμα, εφόσον οι μπανάνες αγοράζονται πολύ συχνά από τους πελάτες των super-markets, το RS που χρησιμοποιεί δεδομένα από το καλάθι αγορών τους για να εξάγει recommendations θα προτείνει πάντα τις μπανάνες. Ένα αντίστροφο πρόβλημα είναι όταν προτείνεται ένα προϊόν που δεν αγοράζεται συχνά και μόνο ένα κάθε φορά (π.χ. αυτοκίνητο). Επίσης, είναι δυνατόν να αναγνωριστεί κάποιος χρήστης που έχει ασυνήθιστες προτιμήσεις, με αποτέλεσμα να είναι δύσκολο να βρεθούν «παρόμοιοι» πελάτες ώστε να παραχθούν recommendations.
- **Υποκειμενικότητα των βαθμολογήσεων.** Το πρόβλημα αυτό προκύπτει από το γεγονός ότι οι χρήστες βαθμολογούν τα προϊόντα με βάση τα ξεχωριστά και ιδιαίτερα χαρακτηριστικά τους. Για παράδειγμα, ένας χρήστης είναι ελαστικός στις βαθμολογίες του, ενώ ένας άλλος είναι πιο αυστηρός, ακόμα και αν εκτιμούν τα ίδια προϊόντα με την ίδια σειρά.
- **Μονοδιάστατες προτάσεις.** Όταν ένα RS προτείνει αντικείμενα που έχουν υψηλή βαθμολογία λαμβάνοντας υπόψη μόνο το προφίλ του χρήστη, τότε ποτέ δεν θα του προτείνει αντικείμενα που είναι τελείως διαφορετικά από αυτά που ο χρήστης έχει ήδη βαθμολογήσει. Αυτό σημαίνει για παράδειγμα ότι ένας χρήστης που δεν έχει εμπειρία στην ελληνική κουζίνα, δε θα λάβει ποτέ recommendation για το καλύτερο ελληνικό εστιατόριο στην πόλη. Σε ειδικές περιπτώσεις το σύστημα δεν θα προτείνει αντικείμενα που είναι πολύ όμοια με αυτά που έχει δει ήδη ο χρήστης π.χ. διαφορετικά άρθρα που όμως περιγράφουν το ίδιο γεγονός. Ιδανικά, ο χρήστης πρέπει να λαμβάνει σειρά από επιλογές και όχι ένα ομοιογενές σύνολο από εναλλακτικές προτάσεις. Για παράδειγμα, δεν είναι απαραίτητα καλή ιδέα να προτείνονται σε ένα χρήστη όλες τις ταινίες του Woody Allen, ενώ αυτός έδειξε προτίμηση για μια μόνο από αυτές.

Τα τελευταία χρόνια η έρευνα στην περιοχή έχει στραφεί στο συνδυασμό τεχνικών και αλγορίθμων από διάφορες επιστημονικές περιοχές (π.χ. Web Usage Mining) και έχει οδηγήσει στα **hybrid RSs**, τα οποία ως ένα βαθμό αντιμετωπίζουν τα προβλήματα και τους περιορισμούς των πιο παραδοσιακών προσεγγίσεων. Εντούτοις, αυτά τα συστήματα και οι εφαρμογές, όπως άλλωστε και τα παραδοσιακά, δεν λαμβάνουν υπόψη τους τη **σημασιολογική γνώση (semantic knowledge)** για την περιοχή που εφαρμόζονται. Χωρίς αυτή τη γνώση, τα RSs δεν έχουν τη δυνατότητα να προτείνουν διαφορετικούς τύπους πολύπλοκων αντικειμένων με βάση τα υποκείμενα χαρακτηριστικά και τις ιδιότητές τους. Πράγματι, η ολοκλήρωση σημασιολογικής γνώσης αποτελεί την πρόκληση για την επόμενη γενιά των RSs.

Λαμβάνοντας υπόψη τα παραπάνω, στο κεφάλαιο αυτό προτείνεται μια προσέγγιση, η οποία ταξινομείται στην κατηγορία των hybrid RSs και ενσωματώνει χαρακτηριστικά από content-based και collaborative-based τεχνικές. Επιπλέον, προχωράει ένα βήμα πιο πέρα και χρησιμοποιεί τη **σημασιολογική γνώση** που εξάγει από την **οντολογία του e-shop** σε συνδυασμό με τα **δεδομένα χρήσης ιστού** ώστε να παράγει offline το μοντέλο πλοήγησης των χρηστών.

Το RS διακρίνει δύο περιπτώσεις: αυτή του **νέου** και αυτή του **έμπειρου (παλαιού)** χρήστη οι οποίες αντιμετωπίζονται διαφορετικά. Για κάθε χρήστη (νέο ή παλιό) αποθηκεύει και ενημερώνει το προφίλ με τα στοιχεία του καθώς και το ιστορικό των συναλλαγών του. Επιπλέον, χρησιμοποιεί τις **βαθμολογίες** του για τις ταινίες καθώς και την **ομοιότητά** του με άλλους χρήστες σε μια online φάση ώστε να ταξινομήσει τις συστάσεις. Με το συνδυασμό των διαφορετικών διαθέσιμων καναλιών πληροφοριών, το μοντέλο πετυχαίνει να ικανοποιήσει σε μεγαλύτερο βαθμό τις ανάγκες και τις προτιμήσεις των χρηστών.

Το πιλοτικό web site που σχεδιάστηκε και υλοποιήθηκε για τις ανάγκες της προσέγγισης και τη μετέπειτα αξιολόγησή της αποτελείται από μια συλλογή ιστοσελίδων που περιέχουν πληροφορίες για τις ταινίες. Τα αρχικά στοιχεία ανακτήθηκαν από τη βάση δεδομένων **Internet Movie Database** (<http://www.imdb.com>) και οργανώθηκαν σε ένα οντολογικό σχήμα προκειμένου να αποτελέσουν ένα πλούσιο σύνολο δεδομένων που στη συνέχεια χρησιμοποιήθηκε στη διαδικασία ανακάλυψης γνώσης. Στις επόμενες ενότητες του κεφαλαίου περιγράφεται αναλυτικά η προτεινόμενη προσέγγιση.

2.5.3 Σχετικές εργασίες

Πριν προχωρήσουμε στην παρουσίαση του RS μοντέλου θεωρούμε σκόπιμο να κάνουμε μια σύντομη ανασκόπηση σε επιλεγμένες ερευνητικές εργασίες του πεδίου, που αν και νέο έχει να επιδείξει σημαντικά αποτελέσματα. Άλλωστε η ανάγκη παροχής εξατομικευμένων πληροφοριών και υπηρεσιών μέσω του Web έχει αποτελέσει ισχυρή κατευθυντήρια δύναμη για έναν αξιόλογο αριθμό ερευνητών που συνδυάζουν αλγορίθμους και τεχνικές από διαφορετικά επιστημονικά πεδία όπως είναι το Web Usage Mining, το Semantic Web και οι οντολογίες.

Η Berendt (2001) προτείνει τις ιεραρχίες εννοιών (concept hierarchies) για την ανάλυση των σύνθετων δεδομένων χρήσης Web των χρηστών. Αυτές οι ιεραρχίες χρησιμοποιούνται σαν η βασική μέθοδος για τη συνάθροιση των ιστοσελίδων. Επιπλέον, interval-based coarsening και η inverse zooming χρησιμοποιούνται για την εξόρυξη δεδομένων χρήσης σε διαφορετικά επίπεδα αφαίρεσης. Βασικά και coarsened stratograms απεικονίζουν τα δεδομένα χρήσης σε διαφορετικούς βαθμούς λεπτομέρειας. Μια περιπτωσιολογική μελέτη για online αγορές με τη χρήση ενός πράκτορα με ανθρώπινη μορφή αποδεικνύει ότι αυτό το είδος της αφαίρεσης προσφέρει νέες δυνατότητες για σύνθετα μονοπάτια μέσω ενός ημιδομημένου, πλούσιου σε αλληλεπίδραση περιβάλλοντος.

Στην εργασία (Eirinaki et al., 2004) προτείνεται μια μέθοδος σύστασης που συνδυάζει τη σημασιολογία του περιεχομένου και τα δεδομένα πλοήγησης των χρηστών. Η τεχνική χρησιμοποιεί τους όρους μιας οντολογίας της περιοχής προκειμένου να χαρακτηριστούν ομοιόμορφα και το περιεχόμενο και τα δεδομένα πλοήγησης των χρηστών ώστε να

παραχθεί με αυτό τον τρόπο ένα σύνολο συστάσεων που συσχετίζονται σημασιολογικά με την τρέχουσα επίσκεψη του χρήστη. Η εργασία εξετάζει το σημαντικό πρόβλημα της πολυγλωσσίας, το οποίο προκύπτει όταν εμφανίζεται το περιεχόμενο ενός ιστοχώρου σε περισσότερες από μια γλώσσες και ενσωματώνεται στο σύστημα εξατομίκευσης SEWeP (Eirinaki et al., 2003).

Οι Middleton et al. (2004) περιγράφουν μια οντολογική προσέγγιση για user profiling μέσα σε recommendation συστήματα, για το πρόβλημα της σύστασης online ακαδημαϊκών ερευνητικών εργασιών. Τα έγγραφα αυτά ταξινομούνται χρησιμοποιώντας οντολογικές κατηγορίες. CF και content-based αλγόριθμοι σύστασης χρησιμοποιούνται για να συστήσουν σε ένα χρήστη εργασίες που έχουν δει παρόμοιοι με αυτόν χρήστες σε σχέση με τα τρέχοντα ενδιαφέροντά τους. Χρησιμοποιούν τις οντολογικές σχέσεις μεταξύ των θεμάτων ενδιαφέροντος για να προκύψουν άλλα θεματικά που ακόμα δεν έχουν πλοηγηθεί σε αυτά για να τα συστήσουν στους χρήστες.

Οι Mobasher et al. (2004) προτείνουν τη χρήση σημασιολογικής γνώσης για τα αντικείμενα (items) σε στόχο να βελτιώσουν το item-based CF. Σε αυτή την εργασία, αναπαριστούν τη σημασιολογική γνώση για ένα αντικείμενο σαν ένα διάνυσμα χαρακτηριστικών και υπολογίζουν την ομοιότητα του με άλλα αντικείμενα με βάση αυτές τις πληροφορίες. Αυτή η item-similarity στη συνέχεια συνδυάζεται με τη rating-similarity ώστε να υπολογισθεί μια συνολική μετρική (overall measure) η οποία χρησιμοποιείται για να προβλέψει τη βαθμολογία του χρήστη για αυτό το μη-βαθμολογημένο αντικείμενο. Παρέχουν πειραματικά αποτελέσματα για δύο διαφορετικά σύνολα αντικειμένων τα οποία δείχνουν ότι η προσέγγισή τους παράγει σημαντικά πλεονεκτήματα σε σχέση με α) το επίπεδο ακρίβειας, β) την εφαρμογή σε πολύ αραιά σύνολα αντικειμένων και γ) την προσθήκη νέων αντικειμένων.

Οι Cho & Kim (2004) προτείνουν μια μεθοδολογία σύστασης βασισμένη στην εξόρυξη δεδομένων χρήσης και σε μια ταξινόμηση των προϊόντων για να βελτιώσουν την ποιότητα των συστάσεων και την απόδοση των τρεχόντων CF συστημάτων. Η εξόρυξη δεδομένων χρήσης χρησιμοποιεί τη βάση δεδομένων των βαθμολογιών με την παρακολούθηση της αγοραστικής συμπεριφορά των πελατών στο Web ώστε να οδηγήσει σε ποιοτικότερες συστάσεις. Η ταξινόμηση των προϊόντων χρησιμοποιείται για να βελτιώσει την απόδοση της αναζήτησης των κοντινότερων γειτόνων μέσω της μείωσης των διαστάσεων της βάσης δεδομένων των βαθμολογιών.

2.5.4 Η οντολογία του e-shop

Πολλά online καταστήματα χρησιμοποιούν ένα βασικό σχήμα ταξινόμησης για τα προϊόντα που πωλούν. Αυτό το σχήμα θα μπορούσε να απεικονιστεί σαν ένα δέντρο, όπου η ρίζα αντιπροσωπεύει τη γενικότερη κλάση ή κατηγορία, οι εσωτερικοί κόμβοι όλες τις ενδιάμεσες κατηγορίες και τα φύλλα τα συγκεκριμένα προϊόντα. Ο ρόλος του είναι σημαντικός ειδικά στη διαδικασία ανακάλυψης γνώσης, δεδομένου ότι αντιπροσωπεύει την εξαρτώμενη γνώση του e-shop και μπορεί να επηρεάσει τα τελικά αποτελέσματα.

Σε αυτό το πλαίσιο, η προτεινόμενη προσέγγιση συστάσεων είναι βασισμένη σε ένα οντολογικό σχήμα που έχει ενσωματωθεί στο πιλοτικό e-shop που νοικιάζει κινηματογραφικές ταινίες στους πελάτες του. Συγκεκριμένα, ο δικτυακός τόπος αποτελείται

από μια συλλογή ιστοσελίδων οι οποίες περιέχουν πληροφορίες για τα χαρακτηριστικά των ταινιών όπως ο τίτλος, η κατηγορία, το στούντιο, ο ηθοποιός, ο σκηνοθέτης, ο παραγωγός, ο σεναριογράφος, ο μουσικοσυνθέτης, τα βραβεία, το έτος, η διάρκεια, το ακροατήριο, η περίληψη, η γλώσσα, κλπ. Ένας ηθοποιός ή ένας σκηνοθέτης με τη σειρά τους μπορεί να περιλαμβάνει το ονοματεπώνυμό του, τη φιλομορφία του (δηλ. ένα σύνολο από ταινίες), το φύλο του, την εθνικότητά του, την ημερομηνία γέννησης ή θανάτου του, κλπ. Επίσης ένα βραβείο απονέμεται είτε σε ένα πρόσωπο που συμμετέχει στην ταινία π.χ. σκηνοθέτη, ηθοποιό, κλπ. ή σχετίζεται με την ίδια την ταινία και μπορεί να έχει ιδιότητες όπως τον τύπο του (π.χ. πρώτο ανδρικό ρόλο), το έτος βράβευσης-απονομής, κλπ. Η οντολογία του site περιγράφει συνεπώς τις **κλάσεις (classes)** **Movie**, **Person**, **Award**, κλπ. Οι σελίδες του e-shop αποτελούν ένα σύνολο από ενσωματωμένα αντικείμενα τα οποία είναι **στιγμιότυπα (instances)** αυτών των κλάσεων. Ένα μέρος της οντολογίας του e-shop απεικονίζεται στο Σχήμα 20.

Σχήμα 20. Μέρος της οντολογίας του e-shop.

Η κλάση **Movie** έχει **χαρακτηριστικά (attributes)** όπως Title, Category, Actor (σχέση «acted by»), Director (σχέση «directed by»), Producer (σχέση «produced by»), Studio, Award, Year, Duration, Audience, Story, κλπ. Τα χαρακτηριστικά Actor, Director, Producer, έχουν τιμές που είναι άλλα αντικείμενα της οντολογίας, στιγμιότυπα της κλάσης **Person**. Στην ουσία αποτελούν υποκλάσεις της **Person** και χαρακτηρίζονται από τον τρόπο συμμετοχής τους στην ταινία. Το χαρακτηριστικό Year είναι παράδειγμα attribute του οποίου ο τύπος είναι θετικός ακέραιος αριθμός. Το χαρακτηριστικό Category έχει ένα σταθερό τύπο δεδομένων του οποίου οι τιμές είναι ένα σύνολο από ετικέτες π.χ. Comedy, Romance, Thriller, Western, Drama, κλπ. Ένα μέρος της ιεραρχίας αυτού του χαρακτηριστικού που βασίζεται στη σχέση «is-a» φαίνεται στο Σχήμα 21.

Σχήμα 21. Το χαρακτηριστικό Category.

Όπως έχουμε αναφέρει τα αρχικά δεδομένα έχουν ανακτηθεί από τη βάση δεδομένων κινηματογραφικών ταινιών Internet Movie Databases (<http://www.imdb.com>). Τα δεδομένα αυτά οργανώθηκαν στο οντολογικό σχήμα και ενισχύθηκαν προκειμένου να αποτελέσουν ένα πλούσιο υπερσύνολο (superset) το οποίο καλύπτει όλες τις αναγκαίες έννοιες, σχέσεις και συσχετίσεις. Το Σχήμα 22 δείχνει τη web page του e-shop όπου παρουσιάζεται η ταινία «Ο Άρχοντας των Δακτυλιδιών – Η επιστροφή του Βασιλιά», ενώ το Σχήμα 23 απεικονίζει το συγκεκριμένο στιγμιότυπο της κλάσης Movie καθώς και τα χαρακτηριστικά και τις σχέσεις που έχουν εξαχθεί από τη web page και την αντιστοίχιση με την οντολογία. Συγκεκριμένα, το χαρακτηριστικό Category περιλαμβάνει 3 αντικείμενα: «Action», «Adventure» και «Fantasy» και τα οποία αποτελούν στιγμιότυπα της κλάσης Category, ενώ το Actor έχει 6 αντικείμενα: «Elijah Wood», «Viggo Mortensen», «Orlando Bloom», «Ian Holm», «Ian McKellen» και «Liv Tyler» τα οποία αποτελούν στιγμιότυπα της κλάσης Actor (για λόγους απλότητας χρησιμοποιούμε το ονοματεπώνυμο των ηθοποιών για να ορίσουμε τα στιγμιότυπα της κλάσης Actor).

Σχήμα 22. Παράδειγμα της ταινίας «Lord of the Rings – The Return of the King».

Η οντολογία αποτελεί μια αναπαράσταση του e-shop με τον καθορισμό όλων των εννοιών του, των σχέσεων μεταξύ τους και άλλων ιδιοτήτων, των όρων, των συνθηκών και των κανονισμών. Επιτρέπει το σημασιολογικό σχολιασμό, τη σημασιολογική σημείωση και παρέχει τη δυνατότητα εκτέλεσης σημασιολογικής αναζήτησης και browsing βασισμένου στην οντολογία. Το «κτίσιμο» της οντολογίας ήταν μια σύνθετη και χρονοβόρα διαδικασία και

βασίστηκε στη διαίσθησή μας προκειμένου να απεικονιστούν όλες οι έννοιες του e-shop, να οργανωθούν οι ταξινομικές ιεραρχίες τους και να αναπαρασταθούν οι σχέσεις τους. Η ανάπτυξη ήταν παρόμοια με τον καθορισμό ενός συνόλου στοιχείων και της δομής τους. Με αυτό τον τρόπο, η οντολογία μπορεί να θεωρηθεί ως μια **βάση γνώσεων** που χρησιμοποιείται περαιτέρω για την εξαγωγή της «**χρήσιμης**» γνώσης. Συγκεκριμένα, ο ρόλος της είναι να χρησιμοποιηθεί σαν είσοδος στη φάση εξόρυξης ώστε να εξαχθεί, να συνδυαστεί και να μετασχηματιστεί η υπάρχουσα υπονοούμενη και ρητή γνώση (κλάση χρήστη, προφίλ χρήστη, περιεχόμενα και δομή e-shop) σε νέες μορφές. Η έξοδος αυτής της διαδικασίας είναι μια λίστα πιθανών recommendations.

Σχήμα 23. Στιγμιότυπο της οντολογίας.

Οι ερωτήσεις που μπορούν να γίνουν πάνω στο οντολογικό σχήμα της βάσης δεδομένων μπορεί να είναι της μορφής:

- **Βρες όλες τις ταινίες στις οποίες η γλώσσα είναι η ελληνική**
SELECT X, Y FROM {X} language{Y} WHERE Y = "Greek"
- **Βρες όλες τις ταινίες που έχουν έτος παραγωγής το 1993 ή το 1994 ή το 1995**
SELECT X, Y FROM {X} year{Y} WHERE Y in bag ("1993", "1994", "1995")
- **Βρες όλα τα περιεχόμενα της κλάσης Actor**
SELECT X FROM Actor{X}

Ειδικότερα, η οντολογία δημιουργεί τις συνδέσεις μεταξύ των κινηματογραφικών ταινιών σύμφωνα με τα διαφορετικά χαρακτηριστικά που τις χαρακτηρίζουν. Χρησιμοποιώντας το ιστορικό (history file) κάθε πελάτη, οι προτιμήσεις του μπορούν να εξαχθούν. Μπορούμε δηλαδή να βρούμε αν στον πελάτη αρέσουν ή όχι ταινίες:

- από μια συγκεκριμένη κατηγορία π.χ. «θράση», «θρόμα», «κωμωδία», «μούζικαλ», «εποχής», «θρίλερ», κλπ.
- από ένα συγκεκριμένο σκηνοθέτη π.χ. «Steven Spielberg», «Francis Ford Copolla», «Sydney Pollack», κλπ.
- με ένα συγκεκριμένο ηθοποιό π.χ. «Sean Penn», «Nicole Kidman», «Brad Pitt», «Julia Roberts», κλπ.

- ή συνδυασμό άλλων χαρακτηριστικών.

Όταν ένα ή περισσότερα κριτήρια ταιριάσματος ικανοποιούνται, τότε άλλες ταινίες που έχουν παρόμοια χαρακτηριστικά με εκείνες που ο πελάτης έχει ήδη νοικιάσει μπορεί να ανακαλυφθούν σύμφωνα με το οντολογικό σχήμα. Στην περίπτωση ενός νέου πελάτη (history file = $\emptyset$) οι πληροφορίες από την αίτηση εγγραφής του στο e-shop αναλύονται αρχικά και με βάση αυτές γίνεται η αναζήτηση στην οντολογία ώστε να παραχθούν οι συστάσεις.

	«Το Ημερολόγιο της Bridget Jones - 1»	«Chicago»	«Cold Mountain»	«Άγρυπνοι στο Seattle»	«Το Ημερολόγιο της Bridget Jones - 2»
Πέννυ	10	$\emptyset$	9	$\emptyset$	10
Κώστας	5	5	10	10	5
Βλάσης	9	5	$\emptyset$	10	10

Πίνακας 8. Εξαγωγή συμπερασμάτων από τις βαθμολογίες των χρηστών για τις ταινίες.

Ένα παράδειγμα παρουσιάζει ο Πίνακας 8. Έστω ότι θέλουμε να προτείνουμε μια ταινία στην «Πέννυ». Ένα κλασικό CF-RS θα έβρισκε ότι ο «Βλάσης» έχει παρόμοιες προτιμήσεις με την «Πέννυ» οπότε θα κατέληγε ότι μια καλή πρόταση είναι η ταινία «Άγρυπνος στο Seattle». Αν χρησιμοποιήσουμε τώρα το οντολογικό σχήμα μπορούμε να συνάγουμε ότι ο «Βλάσης» προτιμά ταινίες που ανήκουν στην κατηγορία «Romandy Comedy» και η «Πέννυ» προτιμά ταινίες με πρωταγωνίστρια τη «Renee Zellweger» ανεξαρτήτως κατηγορίας. Οπότε το «Chicago» στο οποίο πρωταγωνιστεί η «Renee Zellweger» είναι μια καλύτερη πρόταση για την «Πέννυ» από ότι το «Άγρυπνος στο Seattle». Αυτό είναι το αποτέλεσμα της επίδρασης των διαφορετικών χαρακτηριστικών των ταινιών στο οντολογικό σχήμα και στην παραγωγή της πρότασης η οποία απαιτεί την παρατήρηση της συμπεριφοράς των χρηστών.

2.5.5 Συλλογή δεδομένων

Ο στόχος της φάσης αυτής είναι ο προσδιορισμός και η συλλογή όλων των διαθέσιμων πηγών δεδομένων ενός πελάτη. Αυτή διαδικασία δεν είναι στατική αλλά εκτελείται συνεχώς και οδηγεί (μετά τον κατάλληλο καθαρισμό και μετασχηματισμό τους) στη δημιουργία των μοντέλων αναπαράστασης των πελατών και θα επιτρέψει την περαιτέρω επεξεργασία και την ενημέρωσή τους. Συγκεκριμένα, καλύπτει δεδομένα για:

- το **χρήστη** (π.χ. δημογραφικά, γνώσεις, επιδεξιότητες, ικανότητες, ενδιαφέροντα, προτιμήσεις, στόχοι, κλπ.),
- τη **χρήση** (π.χ. επιλεγμένες ενέργειες, βαθμολογήσεις, συχνότητα χρήσης, συσχετισμοί δράσης ή ακολουθίες, κλπ.),
- το **περιβάλλον χρήσης**, υλικό, λογισμικό και φυσικό (π.χ. έκδοση διαφυλλιστή, plugins, εύρος ζώνης, συσκευές επίδειξης/εισαγωγής, κλπ.).

Υπάρχουν δύο γενικές προσεγγίσεις για τη συλλογή των δεδομένων για το χρήστη: είτε ο χρήστης καλείται **ρητά (explicitly)** να τα παράσχει (χρησιμοποιώντας ερωτηματολόγια, συμπληρώνοντας φόρμες, κλπ.), ή το σύστημα συλλέγει **υπονοούμενα (implicitly)** τέτοιες πληροφορίες χωρίς την συμμετοχή του (χρησιμοποιώντας acquisition rules, plan recognition,

stereotype reasoning, κλπ.). Στην προσέγγισή μας χρησιμοποιούμε το μοντέλο που παρουσιάστηκε στην Ενότητα 2.4 και το οποίο μας βοηθά να κατανοήσουμε τον e-customer και τον τρόπο που αλληλεπιδρά με το e-shop.

2.5.6 Ανάλυση δεδομένων

Το επόμενο βήμα είναι η δημιουργία των προφίλ των χρηστών και η ανάλυση των δεδομένων. Αρχικά, μερικές δραστηριότητες προετοιμασίας πραγματοποιούνται προκειμένου να καθαριστούν τα δεδομένα και να διευκολυνθεί ο χειρισμός τους. Για παράδειγμα, καταχωρήσεις που δεν διαθέτουν πραγματικές πληροφορίες χρήσης απομακρύνονται και τα ελλιπή δεδομένα συμπληρώνονται. Κατόπιν ακολουθεί η εφαρμογή στατιστικών και τεχνικών εξόρυξης γνώσης από δεδομένα προκειμένου να ανιχνευθούν τα ενδιαφέροντα μοτίβα (patterns) στα προεπεξεργασμένα δεδομένα. Οι περισσότεροι γνωστές τεχνικές που χρησιμοποιούνται για την ανάλυση των δεδομένων περιλαμβάνουν το **Clustering** (Hand et al., 2001), το **Classification** (Chakrabarti, 2003) και το **Association Rules Mining** (Agrawal & Srikant, 1994).

Σχήμα 24. Η προτεινόμενη προσέγγιση για το RS.

Το συγκεκριμένο RS απεικονίζεται στο Σχήμα 24. Όπως έχουμε αναφέρει προηγουμένως, το πιλοτικό μας e-shop νοικιάζει ταινίες στους πελάτες βασισμένο σε τεχνικές εξόρυξης γνώσης από δεδομένα χρήσης Web και σημασιολογικά μεταδεδομένα, καθώς επίσης και σε βαθμολογήσεις των ταινιών από τους πελάτες και τεχνικές ταιριάσματος μεταξύ διαφορετικών χρηστών ώστε να παράγει recommendations. Η λειτουργία του e-shop είναι απλή. Στην περίπτωση ενός νέου πελάτη το e-shop τον παρακινεί να γίνει μέλος με να συμπληρώσει την αίτηση εγγραφής. Με αυτό τον τρόπο το κατάστημα συλλέγει τις απαραίτητες αρχικές πληροφορίες για τον πελάτη προκειμένου να τον υποστηρίξει σε όλη τη διάρκεια της πλοήγησής του στο site. Στην περίπτωση ενός παλαιού πελάτη, το κατάστημα τον αναγνωρίζει μέσω μιας διαδικασίας login/password και του παρέχει έναν εξατομικευμένο χαιρετισμό. Στη συνέχεια ο πελάτης μπορεί να πλοηγηθεί στο e-shop, να επιλέξει τις κινηματογραφικές ταινίες από τον online κατάλογο ή να χρησιμοποιήσει τη δυνατότητα αναζήτησης για να ενημερωθεί για τα χαρακτηριστικά τους π.χ. περίληψη, καστ, και να τις προσθέσει στο καλάθι αγορών του και να πληρώσει για το τελικό ποσό ενοικίασης. Επιπλέον, μπορεί να βαθμολογήσει μια συγκεκριμένη ταινία χρησιμοποιώντας βαθμούς από το 1 έως το 10, όπου 1 σημαίνει μια πολύ κακή ταινία και 10 μια πολύ καλή. Η προσέγγιση

ενσωματώνει την τεχνική της ταξινόμησης καθώς και κανόνες συσχέτισης. Στις επόμενες ενότητες περιγράφεται ο τρόπος με τον οποίο παράγονται οι συστάσεις.

2.5.7 Ταξινόμηση χρηστών με βάση τα δεδομένα εγγραφής

Ο κύριος στόχος αυτού του βήματος είναι να ταξινομήσουμε τους χρήστες σε ένα σύνολο προκαθορισμένων κλάσεων. Αυτές οι κλάσεις αντιπροσωπεύουν διαφορετικά προφίλ χρηστών και η ταξινόμηση εκτελείται χρησιμοποιώντας επιλεγμένα χαρακτηριστικά γνωρίσματα με υψηλή διακριτική δυνατότητα. Τα χαρακτηριστικά γνωρίσματα που καθορίζουν μια συγκεκριμένη κλάση μπορούν να περιλάβουν χαρακτηριστικά όπως το userID, την ηλικία, το φύλο, την εθνικότητα, το επάγγελμα, καθώς επίσης την εκπαίδευση και τις προτιμήσεις του. Για παράδειγμα, από την κλάση «νοικιάζω κωμωδίες», μπορούμε να δούμε ποιοι πελάτες προτιμούν να νοικιάζουν κωμωδίες και ποιοι όχι. Ο Πίνακας 9 απεικονίζει το προφίλ ενός ενεργού πελάτη. Έστω ότι έχουμε τα χαρακτηριστικά και τις τιμές όπως αυτές που παρουσιάζει ο Πίνακας 10 και θέλουμε να αποφασίσουμε εάν ο επόμενος άγνωστος χρήστης Χ ανήκει σε αυτή την κλάση ή όχι δηλαδή αν νοικιάζει κωμωδίες ή όχι:

$X = \{userID = 101, age = 22, sex = w, occupation = student, education = medium\}$

userID	"67"
age	"30"
sex	"male"
occupation	"seller"
education	"higher"

Πίνακας 9. Μέρος ενός προφίλ χρήστη.

Η ταξινόμηση απαιτεί ένα σύνολο στοιχείων εκπαίδευσης και τις προκαθορισμένες κλάσεις αφού ανήκει στην κατηγορία των supervised machine learning τεχνικών. Στη συνέχεια, παρακολουθώντας τις αναθέσεις των κλάσεων στο σύνολο εκπαίδευσης, ο ταξινομητής μαθαίνει να αναθέτει τα νέα στοιχεία στις κλάσεις. Συχνά το clustering εφαρμόζεται πριν από το classification για να καθορίσει το σύνολο των κλάσεων.

userID	age	sex	occupation	education	class: rent comedies
1	30	w	engineer	high	yes
2	25	m	student	medium	no
3	34	m	salesman	low	yes
4	18	m	student	low	yes
5	45	w	director	high	no
...	...	...	...	...	...
100	28	w	artist	high	yes

Πίνακας 10. Στιγμιότυπο της βάσης δεδομένων.

Έχουμε βασίσει το μοντέλο ταξινόμησης στο naïve Bayes classifier, όπου κάθε στιγμιότυπο X αποτελείται από ένα σύνολο ιδιοτήτων $x_1, x_2, \dots, x_n$. Καθορίζουμε τις m κλάσεις $C_1, C_2, \dots, C_m$. Δοθέντος ενός αγνώστου στιγμιότυπου X για το οποίο δεν ξέρουμε την κλάση του, ο

ταξινομητής προβλέπει ότι το X ανήκει στην κλάση με την υψηλότερη πιθανότητα. Ο ταξινομητής ορίζει το X στην κατηγορία C_i εάν:

$$P(C_i|X) > P(C_j|X) \text{ για } 1 \leq j \leq m, j \neq i$$

Σύμφωνα με το θεώρημα Bayes:

$$P(C_i|X) = P(X|C_i) * P(C_i) / P(X)$$

Το $P(X)$ είναι σταθερό για όλα τα στιγμιότυπα οπότε πρέπει να μεγιστοποιηθεί η έκφραση:

$$P(X|C_i) * P(C_i)$$

Για την ταξινόμηση του X , υπολογίζεται ότι η πιθανότητα $P(X|C_i) * P(C_i)$ για κάθε κατηγορία C_i . Το X θα τοποθετηθεί στην κατηγορία C_i εάν:

$$P(X|C_i) * P(C_i) > P(X|C_j) * P(C_j) \text{ για } 1 \leq j \leq m, j \neq i$$

Αυτό σημαίνει ότι το X τοποθετείται στην κλάση με τη μέγιστη πιθανότητα.

Το μοντέλο αυτό μπορεί να αναθεωρηθεί με το πέρασμα του χρόνου και να ενημερωθεί με βάση τα συλλεχθέντα στοιχεία από τις συναλλαγές. Η διαδικασία της τοποθέτησης των χρηστών στις κλάσεις μπορεί επίσης να χρησιμοποιηθεί για την ενίσχυση της συνεργασίας και της επικοινωνίας μεταξύ των χρηστών, καθώς και για επιτρέπει στο e-shop να κατανοήσει τη δυναμική των «**εικονικών κοινοτήτων**» (**virtual communities**). Για κάθε κλάση ένα recommendations είναι συνημμένο. Αυτές οι συστάσεις είναι λίστες με π.χ. τις καλύτερες και τις χειρότερες κινηματογραφικές ταινίες που άλλοι χρήστες «κοντινοί» στον τρέχοντα έχουν συνεισφέρει.

2.5.8 Ανάλυση των click-streams των χρηστών

Όπως έχουμε αναφέρει στο Κεφάλαιο 1 οι association rules συνδέουν ένα ή περισσότερα γεγονότα μεταξύ τους. Ο στόχος τους είναι να ανακαλυφθούν συσχετίσεις και ενώσεις ανάμεσα σε διαφορετικούς τύπους δεδομένων χωρίς προφανή σημασιολογική εξάρτηση. Παράλληλα, συσχετισμοί μεταξύ σελίδων μη άμεσα συνδεδεμένων και προηγούμενων άγνωστες ενώσεις μεταξύ ομάδων χρηστών με συγκεκριμένα ενδιαφέροντα μπορούν να έρθουν στην επιφάνεια. Το μοντέλο παρακολουθεί όλες τις ενέργειες των χρηστών και τις θεωρεί σαν διαδοχικά αιτήματα σελίδων p_i που καταγράφονται στα log files των εξυπηρετητών. Τα log files καθαρίζονται στη συνέχεια από όλες τις περιττές πληροφορίες (π.χ. δευτεροβάθμιες αυτόματες αιτήσεις για πολυμεσικά δεδομένα όπως γραφικά, εικόνες, ήχους, βίντεο, κλπ.). Συνδυάζοντας τα υπόλοιπα αιτήματα με τις πληροφορίες για το περιεχόμενο του web site και τη δομή του, το σύστημα φιλτράρει τις προσβάσεις των χρηστών στις σελίδες των κινηματογραφικών ταινιών. Το σύνολο των ταινιών που έχουν προσπελασθεί από ένα συγκεκριμένο χρήστη u κατά τη διάρκεια όλων των προηγούμενων επισκέψεων του στο e-shop αποθηκεύονται στο $profile(u)$ προφίλ του και εκεί είναι που το σύστημα προσπαθεί να ανακαλύψει κανόνες συσχέτισης.

Ένα παράδειγμα κανόνα συσχέτισης είναι το ακόλουθο:

$$\{movie_i, movie_j\} \rightarrow \{movie_x\} \text{ με } support=0.02 \text{ και } confidence=0.68$$

Αυτός ο κανόνας υποδηλώνει τη σχέση ότι οι χρήστες που προσπέρασαν τη $movie_i$ και τη $movie_j$ επίσης τείνουν (με confidence 68%) να ενδιαφερθούν για τη $movie_x$. Το support υποδηλώνει το γεγονός ότι το σύνολο $\{movie_i, movie_j, movie_x\}$ παρατηρείται στο 2% των συνόλων των κινηματογραφικών ταινιών που προσπελάζονται. Άλλα παραδείγματα κανόνων συσχέτισης που μπορούν να εξαχθούν είναι τα ακόλουθα:

- 60% των χρηστών που έχουν ενοικιάσει την ταινία «Lord of the Rings – The Fellowship» στη συνέχεια επέλεξαν την ταινία «Lord of the Rings – The Two Towers» και «Lord of the Rings – The Return of the King».
- 75% των χρηστών που επισκέφτηκαν τη σελίδα του site που περιλάμβανε την ταινία «Shrek» ανήκει στην ομάδα ηλικίας 15-20.
- 25% χρηστών που έχουν ενοικιάσει την ταινία «Harry Potter and the Philosopher's Stone» ανήκουν στην ηλικιακή ομάδα 20-30 και ζουν στην Αθήνα.

Οι κανόνες συσχέτισης μπορούν να χρησιμοποιήσουν σαν είσοδο είτε τα σύνολα των ταινιών που προσπελάζονται από όλους τους χρήστες, είτε μόνο αυτές που τις έχουν ενοικιάσει οι χρήστες που ανήκουν στην ίδια κλάση C με τον τρέχοντα χρήστη. Μια άλλη επιλογή είναι να χρησιμοποιηθούν και οι δυο προσεγγίσεις και στο χρήστη να προταθούν οι ταινίες που ανήκουν στην ένωση των δύο συνόλων. Αυτό το σενάριο είναι πολύ χρήσιμο όταν αποτυγχάνουν οι κανόνες συσχέτισης να παράγουν αξιόπιστα recommendations λόγω έλλειψης επαρκών και αξιόλογων δεδομένων εισόδου.

2.6 Αλγόριθμος παραγωγής recommendations

Στην ενότητα αυτή περιγράφεται λεπτομερώς ο αλγόριθμος που χρησιμοποιείται για την παραγωγή των recommendations για τις κινηματογραφικές ταινίες. Αυτές οι συστάσεις εμφανίζονται στον πελάτη με τη μορφή:

*«πελάτες που τους άρεσε ή που ενοικίασαν αυτή την ταινία
τους άρεσε ή ενοικίασαν επίσης...»*

Η διαδικασία φιλτραρίσματος βασίζεται σε διαφορετικούς τύπους στοιχείων. Πιο συγκεκριμένα, εξαρτώνται από την κλάση στην οποία ο χρήστης είναι ταξινομημένος, τα click-streams, το ιστορικό των συναλλαγών του και τα μεταδεδομένα της οντολογίας του e-shop. Η κύρια ιδέα είναι αρχικά να παραχθεί ένα πιθανό σύνολο recommendations συνδυάζοντας όλους τους προαναφερθέντες τύπους δεδομένων. Στη συνέχεια να μειώσουμε αυτό το σύνολο αφαιρώντας τις ταινίες που βαθμολογικά βρίσκονται κάτω από ένα προκαθορισμένο κατώφλι α . Τέλος, να ταξινομήσουμε τα υπόλοιπα στοιχεία του (πρώτα η ταινία με τη μεγαλύτερη προβλεπόμενη βαθμολογία).

Διακρίνουμε δύο περιπτώσεις ανάλογα με την ύπαρξη ή όχι ιστορικών στοιχείων για το χρήστη:

- **Περίπτωση 1: Όταν ο χρήστης είναι παλαιός και το σύστημα έχει παρακολουθήσει το ιστορικό του.**

Ορίζουμε σαν $M_{History}$ το σύνολο των κινηματογραφικών ταινιών που ο χρήστης έχει δει και έχει βαθμολογήσει. Σε αυτή την περίπτωση, το σύστημα σύμφωνα με $M_{History}$ βρίσκει «παρόμοιους» χρήστες με τον τρέχοντα χρήστη, δηλαδή αυτούς που έχουν «παρόμοιο» ιστορικό και βαθμολογήσεις. Ορίζουμε αυτό το σύνολο χρηστών ως U_L . Κάθε χρήστης U_i που ανήκει στο U_L συνδέεται με ένα αριθμό S_i που απεικονίζει την **ομοιότητα (similarity)** του U_i με τον τρέχοντα χρήστη. Επίσης ορίζουμε ως U_M το σύνολο χρηστών που ανήκουν στο U_L και έχουν δει την κινηματογραφική ταινία M . Επιπλέον, σύμφωνα με $M_{History}$ το μοντέλο χρησιμοποιεί την οντολογία για να ανακαλύψει τις συσχετίσεις μεταξύ των προϊόντων που ο χρήστης έχει βαθμολογήσει. Ορίζουμε ως $R_{Ontology}$ τη λίστα των recommendations που προέρχονται από την οντολογία με βάση το αρχείο ιστορίας του χρήστη.

- **Περίπτωση 2: Όταν ο χρήστης είναι νέος και το ιστορικό του είναι κενό.**

Σε αυτή την περίπτωση, το σύστημα χρησιμοποιεί τις πληροφορίες από την εγγραφή του χρήστη και συνεπώς την αρχική κλάση στην οποία έχει ταξινομηθεί. Ορίζουμε ως το U_K το σύνολο χρηστών της κλάσης που έχουν «παρόμοια» χαρακτηριστικά με το τρέχοντα χρήστη. Αυτά τα χαρακτηριστικά είναι οι πληροφορίες που ο χρήστης έδωσε στο σύστημα μέσω της διαδικασίας συμπλήρωσης της φόρμας εγγραφής. Για παράδειγμα μπορεί να περιλαμβάνουν την ηλικία του, το φύλο του, το επάγγελμα του, την εκπαίδευση του, τις προτιμήσεις του, κλπ. Κάθε χρήστης U_i που ανήκει στο U_K συνδέεται με έναν αριθμό S_i που δηλώνει την ομοιότητα του U_i με τον τρέχοντα χρήστη σύμφωνα με τα προαναφερθέντα χαρακτηριστικά. Ορίζουμε ως U_M το σύνολο χρηστών που ανήκουν στο U_K και έχουν δει την κινηματογραφική ταινία M . Σε κάθε κλάση και σε αυτή την περίπτωση αντιστοιχεί ένα σύνολο κινηματογραφικών ταινιών. Ορίζουμε ως $R_{Cluster}$ τις προτεινόμενες ταινίες σύμφωνα με την κατηγορία του χρήστη.

Στη συνέχεια, ορίζουμε ως $W_{Current}$ το τρέχον παράθυρο συνόδου που περιέχει τις ταινίες που ο χρήστης προσπελαίνει στην τρέχουσα συναλλαγή του. Τα recommendations θα παραχθούν σύμφωνα με το τρέχον click-stream του χρήστη με τη χρησιμοποίηση κανόνων συσχέτισης και της οντολογίας του web site. Ορίζουμε σαν $R_{Rules-current}$ τα recommendations που προέρχονται από το τρέχον παράθυρο του χρήστη και τους κανόνες συσχέτισης και σαν $R_{Ontology-current}$ εκείνα που προκύπτουν από το τρέχον παράθυρο συνόδου και την οντολογία.

Χρησιμοποιούμε δύο μεθόδους για την επεξεργασία των πληροφοριών της τρέχουσας συνόδου επειδή από το συνδυασμό τους επιτυγχάνουμε καλύτερα αποτελέσματα. Κάθε μέθοδος χρησιμοποιεί διαφορετικούς τρόπους να προβλεφθούν οι προτιμήσεις του χρήστη, δηλαδή οι κανόνες συσχέτισης ενσωματώνουν την πλοηγητική συμπεριφορά των χρηστών, ενώ η οντολογία απεικονίζει τις συνδέσεις μεταξύ των χαρακτηριστικών των ταινιών και

ανακαλύπτει τα πρότυπα που εξηγούν τις προτιμήσεις του χρήστη. Στη συνέχεια, ονομάζουμε $R_{Initial}$ το σύνολο με τα αρχικά recommendations το οποίο θα φιλτραρισθεί από τον αλγόριθμο και θα διαταχθεί. Υπολογίζεται από την ένωση του $R_{Ontology-current}$ και $R_{Rules-current}$ και στην πρώτη περίπτωση (όταν έχει ο χρήστης ιστορικό) χρησιμοποιούμε το $R_{Ontology}$, ενώ στη δεύτερη περίπτωση (όταν ο χρήστης είναι νέος) χρησιμοποιούμε το $R_{Cluster}$. Υπολογίζουμε $R_{Initial}$ για κάθε περίπτωση ως εξής:

■ **Περίπτωση 1:**

$$R_{Initial} = R_{Ontology-current} \cup R_{Rules-current} \cup R_{Ontology}$$

■ **Περίπτωση 2:**

$$R_{Initial} = R_{Ontology-current} \cup R_{Rules-current} \cup R_{Cluster}$$

Ο Πίνακας 11 παρουσιάζει τον αλγόριθμο που χρησιμοποιείται για την παραγωγή των recommendations στους πελάτες του e-shop. Όσον αφορά στην ομοιότητα μεταξύ των χρηστών, θεωρούμε ότι κάθε χρήστης έχει ένα διάνυσμα με τις ταινίες που έχει βαθμολογήσει (δηλαδή ένα διάνυσμα για το ιστορικό του). Για παράδειγμα, έστω X και Y δύο χρήστες-πελάτες του e-shop. Ορίζουμε ως X_R και Y_R τα διανύσματα των ιστορικών τους αντίστοιχα και R_X και R_Y τα σύνολα με τις βαθμολογημένες ταινίες. Θεωρούμε ότι C είναι το σύνολο με τις κοινά βαθμολογημένες ταινίες δηλαδή:

$$C = R_X \cap R_Y.$$

Εάν οι χρήστες δεν έχουν καμία κοινή ταινία τότε $C = \emptyset$. Επιπλέον, καθορίζουμε ως X_C και Y_C τα διανύσματα με τις ταινίες που και οι δύο χρήστες έχουν βαθμολογήσει (ταινίες που ανήκουν στο C). Ένας περιορισμός που προκύπτει σε αυτό το βήμα αφορά στο ποσοστό των κοινών βαθμολογημένων ταινιών σε σχέση με όλες τις βαθμολογημένες ταινίες. Για παράδειγμα, εάν ο χρήστης X έχει βαθμολογήσει 60 ταινίες, $R_X = 60$ και ο χρήστης Y έχει βαθμολογήσει 40 ταινίες, $R_Y = 40$ και οι κοινές ταινίες τους είναι μόνο 5, δηλαδή $C = 5$ τότε δεν μπορούμε να θεωρήσουμε ότι οι χρήστες αυτοί είναι παρόμοιοι ακόμα και αν οι βαθμολογίες που έχουν δώσει είναι παρόμοιες. Για να θεωρήσουμε δύο χρήστες ως παρόμοιους πρέπει να έχουν κοινές περισσότερες από το 50% των ταινιών αυτού με τις λιγότερες βαθμολογημένες ταινίες.

Η ομοιότητα μεταξύ των χρηστών X και Y είναι η απόσταση μεταξύ των διανυσμάτων X_C και Y_C και μπορεί να υπολογιστεί από τον ακόλουθο τύπο:

$$sim(X_C, Y_C) = \sqrt{\sum_{i=1}^N (R(X_C, M_i) - R(Y_C, M_i))^2}$$

Όταν ο χρήστης είναι νέος, υπολογίζουμε την ομοιότητα μεταξύ των χρηστών στην ίδια κλάση. Σε αυτή την περίπτωση, χρησιμοποιούμε τα χαρακτηριστικά που προήλθαν από την αίτηση εγγραφής μέλους. Αυτά τα χαρακτηριστικά έχουν μια ορισμένη θέση στο διάνυσμα που περιγράφει το προφίλ του χρήστη. Για παράδειγμα, εάν τα χαρακτηριστικά είναι το φύλο, η ηλικία, το επάγγελμα και η εκπαίδευση, μια περίπτωση αυτού του διανύσματος για το

χρήστη X θα μπορούσε να είναι $\{0, 22, 3, 2\}$. Αυτό σημαίνει: $\{\text{sex} = \text{male}, \text{age} = 22, \text{occupation} = \text{student}, \text{education} = \text{medium}\}$. Με άλλα λόγια, κάθε χαρακτηριστικό παίρνει τις τιμές που αντιστοιχούν σε ορισμένες πληροφορίες π.χ. $\text{occupation}=3$ δείχνει έναν σπουδαστή.

Είσοδος

U : όλοι οι χρήστες

U_i : ο χρήστης με το ID i

M : όλες οι ταινίες

M_j : η ταινία j

$M_{History}$: οι ταινίες που ένας χρήστης έχει βαθμολογήσει

$R_{Initial}$: αρχικό σύνολο recommendations

Αλγόριθμος

For each M_j in $R_{Initial}$

    If $M_{History} = \emptyset$ then

        Find U_M in U_K

    Else Find U_M in U_L

    For each U_i in U_M

$r_{M_j} = R_{U_i M_j} * S_i$

        if $r_{M_j} \succ \alpha$ then

$R = R \cup \{M_j\}$

    end

end

Έξοδος

Μια λίστα R από recommendations ταινιών σε διάταξη (πρώτη η ταινία με το μεγαλύτερο r) τις οποίες ο χρήστης U_i δεν έχει δει και με μεγάλη πιθανότητα θα επιθυμούσε να δει

Πίνακας 11. Ο αλγόριθμος που χρησιμοποιείται για την παραγωγή των recommendations.

Για την αξιολόγηση της προσέγγισής μας χρησιμοποιήσαμε δύο σύνολα δεδομένων. Το πρώτο περιλαμβάνει 100 στιγμιότυπα του οντολογικού σχήματος από το web site του monie e-shop. Ένας spider χρησιμοποιήθηκε για να εξάγει αυτά τα στιγμιότυπα από τις HTML σελίδες του web site και να τα τοποθετήσει στη βάση δεδομένων. Το δεύτερο σύνολο περιλαμβάνει δεδομένα από την πλοηγητική συμπεριφορά των χρηστών και συλλέχθηκαν από τα αρχεία logs. Το σύνολο αυτό για την ώρα περιλαμβάνει 25 χρήστες και 218 επισκέψεις. Συγκεκριμένα, ακολουθήσαμε τα παρακάτω βήματα:

1. Πήραμε δείγματα βαθμολογιών από τους χρήστες για τις ταινίες που έχουν δει.
2. Για κάθε χρήστη u υπολογίσαμε τη βαθμολογία r_{um} που ο χρήστης προβλέπουμε ότι θα βάλει στην ταινία m που δεν έχει δει.

2.1 Παρουσιάσαμε τη σελίδα με τα στοιχεία της ταινίας m με το μεγαλύτερο r_{um} στο χρήστη.

2.2 Ζητήσαμε από το χρήστη να βαθμολογήσει την ταινία m χωρίς να την έχει δει με βάση τα στοιχεία που παρουσιάζονται στην ιστοσελίδα.

3. Ζητήσαμε από το χρήστη να δει την ταινία και να την ξαναβαθμολογήσει.

Ορίσαμε σαν **εκτιμημένη-βαθμολογία** er_{um} τη βαθμολογία που προκύπτει από το βήμα 2.2 χωρίς ο χρήστης u να έχει δει την ταινία m . Επίσης ορίσαμε σαν **πραγματική-βαθμολογία** ar_{um} τη βαθμολογία που προκύπτει από το βήμα 3 αφού έχει δει την ταινία.

Χρησιμοποιήσαμε το **mean absolute error (MAE)** το οποίο υπολογίζει τη μέση απόλυτη απόκλιση μεταξύ μιας πρόβλεψης και της πραγματικής βαθμολογίας ενός χρήστη. Αρχικά υπολογίσαμε το **absolute error AE** κάθε πρόβλεψης και στη συνέχεια υπολογίσαμε το MAE σαν το άθροισμα των AE διαιρεμένο με τον αριθμό των προβλέψεων που έγιναν. Όσο μικρότερο είναι το MAE τόσο καλύτερη είναι και η αποτελεσματικότητα της τεχνικής. Συγκεκριμένα υπολογίσαμε το MAE για τη βαθμολογία του RS μοντέλου σε σχέση με την εκτιμώμενη ($\overline{E_{er}}$) και την πραγματική βαθμολογία ($\overline{E_{ar}}$) του χρήστη όσο και μεταξύ των δύο τελευταίων ($\overline{E_{aer}}$) όπως φαίνεται από τους ακόλουθους τύπους:

$$\overline{E_{er}} = \frac{\sum_{u \in U'} \sum_{m \in M'} |er_{um} - r_{um}|}{\# predictions} \quad (\alpha)$$

$$\overline{E_{ar}} = \frac{\sum_{u \in U'} \sum_{m \in M'} |ar_{um} - r_{um}|}{\# predictions} \quad (\beta)$$

$$\overline{E_{aer}} = \frac{\sum_{u \in U'} \sum_{m \in M'} |er_{um} - ar_{um}|}{\# predictions} \quad (\gamma)$$

για να μετρήσουμε την **ακρίβεια των recommendations (accuracy)** όπου U' είναι το σύνολο το χρηστών για τους οποίους έγιναν οι προβλέψεις από το RS και M' το σύνολο των ταινιών που δεν έχουν δει. Τα πειράματα έδειξαν καταρχήν ότι η εκτιμώμενη από την πραγματική βαθμολογία του χρήστη διαφέρει σε ποσοστό της τάξης 8,99%. Η προβλεπόμενη βαθμολογία από το RS σε σχέση με την εκτιμημένη βαθμολογία αποκλίνει σε ποσοστό 9,68% ενώ σε σχέση με την πραγματική βαθμολογία ποσοστό 10,38%. Τα παραπάνω ποσοστά δείχνουν ότι όλη η διαδικασία ενισχύεται από το συνδυασμό οντολογίας του site, των αρχείων ιστορικού των χρηστών, των δεδομένων εγγραφής τους και των κανόνων συσχέτισης, οι οποίοι ενσωματώνουν όλη την απαραίτητη γνώση για την πλοηγητική συμπεριφορά των χρηστών.

DVD Club

Contact | About

Παρασκευή 18 Νοεμβρίου 2005

Home

DVD Club
Entrance
Site Administrator
Administrator

Meet the Fockers

Fill in the fields for entrance to DVD Club

Login:
Password:
Login

Need an account?
The subscription for a new account takes only few minutes.

Forgot your password?
If you forgot your login and password we will send it back to you.

November 2005

M T W T F S S
- 1 2 3 4 5 6
7 8 9 10 11 12 13
14 15 16 17 18 19 20
21 22 23 24 25 26 27
28 29 30 - - -
- - - - -

Michael Moore
Fahrenheit 9/11

Copyright 2005 DVD Club

Σχήμα 25. Εισαγωγή ενός χρήστη στο online DVD Club.

DVD Club

Welcome user: markel@cid.upatras.gr

Παρασκευή 18 Νοεμβρίου 2005

Movies | My Purchases | My Shopping Cart | Logout

Shopping Cart

Item	Price	Quantity	Total	Delete
Analyze That	1 Euros	1	1 Euros	Delete
Gone with the Wind	1 Euros	1	1 Euros	Delete
Pirates of the Caribbean: The Curse of the Black Pearl	1 Euros	1	1 Euros	Delete
The Godfather	1 Euros	1	1 Euros	Delete
4 Items			4 Euros	

Save Changes

Proceed to checkout

DVD Club Recommendations

The Godfather: Part II
Price: 1 euros
Add to Shopping Cart

The Godfather: Part III
Price: 1 euros
Add to Shopping Cart

Total price: 2 euros
Add to Shopping Cart

Copyright 2005 DVD Club

Σχήμα 26. Παραγωγή προτάσεων για κινηματογραφικές ταινίες.

DVD Club

Welcome user: market@ceid.upatras.gr
Δευτέρα 14 Νοεμβρίου 2005

[Movies](#) | [My Purchases](#) | [My Shopping Cart](#) | [Logout](#)

My Purchases

		Date	Total
1.	View Order	2005-10-13 22:00:09	2 Euros
2.	View Order	2005-10-12 16:07:59	5 Euros

2 purchases found

Order Number: 2	Date: 2005-10-12 16:07:59	
Movie Name	Quantity	Total
1. Bridget Jones: The Edge of Reason	1	1 Euros
2. Dracula	1	1 Euros
3. Love Actually	1	1 Euros
4. Pirates of the Caribbean: The Curse of the Black Pearl	1	1 Euros
5. Showtime	1	1 Euros
	5 Items	5 Euros

This order contains 5 items

Σχήμα 27. Το ιστορικό παραγγελιών ενός πελάτη.

DVD Club

Welcome administrator
Παρασκευή 18 Νοεμβρίου 2005

[Movies](#) | [Statistics](#) | [Personalization](#) | [Logout](#)

[General](#) | [Genre](#) | [Age](#) | [Region](#) | [Background](#) | [Category](#)

Statistics Page - General

There are **22** users in our database

12 out of 22 users are **females**. Percentage: **55%**
10 out of 22 users are **males**. Percentage: **45%**

9 out of 22 are from region **Ammochostos**. Percentage: **41%**
4 out of 22 are from region **Lefkosia**. Percentage: **18%**
6 out of 22 are from region **Kerinia**. Percentage: **27%**
3 out of 22 are from region **Larnaka**. Percentage: **14%**

5 out of 22 are with background **very high**. Percentage: **23%**
5 out of 22 are with background **medium**. Percentage: **23%**
9 out of 22 are with background **uneducated**. Percentage: **41%**
3 out of 22 are with background **high**. Percentage: **14%**

12 out of 22 prefer movies from category **Action**. Percentage: **55%**
14 out of 22 prefer movies from category **Romance**. Percentage: **64%**
9 out of 22 prefer movies from category **Adventure**. Percentage: **41%**
9 out of 22 prefer movies from category **Crime**. Percentage: **41%**
8 out of 22 prefer movies from category **Documentary**. Percentage: **36%**
4 out of 22 prefer movies from category **Fantasy**. Percentage: **18%**
4 out of 22 prefer movies from category **Thriller**. Percentage: **18%**
3 out of 22 prefer movies from category **Mystery**. Percentage: **14%**
5 out of 22 prefer movies from category **Comedy**. Percentage: **23%**
4 out of 22 prefer movies from category **Drama**. Percentage: **18%**
2 out of 22 prefer movies from category **Horror**. Percentage: **9%**

1 out of 22 are with age **teen - 17 years and less**. Percentage: **5%**
5 out of 22 are with age **young - 18 to 30 years**. Percentage: **23%**
8 out of 22 are with age **middle-aged - 30 to 60 years**. Percentage: **36%**
8 out of 22 are with age **old - more than 60 years**. Percentage: **36%**

Σχήμα 28. Στατιστικά στοιχεία του DVD Club.

2.7 Συμπεράσματα και ανοικτά θέματα

Στο κεφάλαιο αυτό μελετήθηκε το πρόβλημα της ενσωμάτωσης σημασιολογικής γνώσης στη διαδικασία παραγωγής εξατομικευμένων προτάσεων σε ένα ηλεκτρονικό κατάστημα χρησιμοποιώντας δεδομένα χρήσης Web. Συγκεκριμένα, ερευνήθηκε το επιστημονικό πεδίο, παρουσιάστηκε το πλαίσιο για την ολοκλήρωση οντολογικών δεδομένων και προτάθηκε ένα RS που αξιολογήθηκε σε ένα πιλοτικό web site που νοικιάζει κινηματογραφικές ταινίες στους πελάτες του. Τα πειραματικά αποτελέσματα έδειξαν πώς αυτή η προσέγγιση μπορεί να παράσχει καλύτερα πρότυπα και ποιοτικότερες υπηρεσίες εξατομικεύσης. Κάποια ενδιαφέροντα ερευνητικά ανοικτά προβλήματα για μελλοντική εργασία αφορούν:

- **Κατασκευή/ενημέρωση της οντολογίας.** Η διαδικασία κατασκευής, συντήρησης και εμπλουτισμού μιας οντολογίας είναι γνωστή και σαν «ontology engineering». Για μικρά web sites με στατικές ιστοσελίδες, είναι εφικτή η δημιουργία της χειρωνακτικά. Αυτός ο τρόπος απαιτεί σημαντική προσπάθεια από την πλευρά των μηχανικών για τα μεγάλα sites ή για αυτά που δυναμικά παράγουν το περιεχόμενό τους. Στην τελευταία περίπτωση, τα πρότυπα των ιστοσελίδων (templates) είναι συνήθως βασισμένα σε δομημένες ερωτήσεις που εκτελούνται στις βάσεις δεδομένων. Τότε το σχήμα των βάσεων δεδομένων μπορεί να χρησιμοποιηθεί άμεσα ώστε να εξαχθούν οι οντολογικές πληροφορίες. Μερικοί Web εξυπηρετητές στέλνουν δομημένα τα αρχεία των στοιχείων (π.χ. αρχεία XML) στους χρήστες και αφήνουν στους μηχανισμούς μορφοποίησης από την πλευρά του χρήστη (π.χ. CSS αρχεία) να επιλύσουν την τελική μορφή στους agents των χρηστών. Οπότε είναι γενικά δυνατό να προκύψει το σχήμα από τα δομημένα αρχεία των στοιχείων. Όταν δεν υπάρχει καμία άμεση πηγή για την απόκτηση των οντολογιών, machine learning και text mining τεχνικές πρέπει να υιοθετηθούν για την εξόρυξη γνώσης από το περιεχόμενο και τη δομή των συνδέσμων υπερκειμένου των ιστοσελίδων. Αρκετές εργασίες προσπαθούν να παράγουν αυτόματα κατανοητές οντολογίες από τον υπολογιστή στην περιοχή των web sites (Clerkin et al., 2001), (Stumme et al., 2000), (Maedche & Staab, 2000). Το αποτέλεσμα αυτής της φάσης είναι ένα σύνολο τυπικά καθορισμένων οντολογιών που αντιπροσωπεύουν πλήρως το web site. Η επιλογή της γλώσσας αντιπροσώπευσης της οντολογίας έχει άμεση επίδραση στην ευελιξία της φάσης εξόρυξης των δεδομένων. Οι κοινές προσεγγίσεις αντιπροσώπευσης είναι το vector-space model (Loh et al., 2000), οι περιγραφικές λογικές (όπως DAML+OIL) (Giugno & Lukasiewicz, 2002), (Horrocks, 2002), η λογική πρώτης κατάταξης (Craven et al., 2000), τα σχεσιακά μοντέλα (Dai & Mobasher, 2002), τα πιθανοτικά σχεσιακά μοντέλα (Getoor et al., 2001) και τα πιθανοτικά Markov μοντέλα (Anterson et al., 2002).
- **Μετρικές ομοιότητας.** Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι μετρικές που υπολογίζουν αποτελεσματικά την ομοιότητα μεταξύ των αντικειμένων μια οντολογίας καθώς και των παραγόμενων προτύπων.
- **Χρήση μετρικών αξιολόγησης του μοντέλου.** Ένα άλλο σημαντικό ζήτημα που χρήζει επιπλέον μελέτης αφορά στη χρήση περισσότερων μετρικών προκειμένου να αξιολογηθεί εκτενώς το μοντέλο όπως η κάλυψη (coverage) που μετρά τη δυνατότητα του συστήματος να προτείνει όλες τις σελίδες που είναι πιθανό να προσπελασθούν από τους χρήστες.

- **Διαφύλαξη προσωπικών δεδομένων του χρήστη.** Η διαδικασία παραγωγής εξατομίκευσης προσπαθεί για αναγνώριση του χρήστη, να καταγράψει την online συμπεριφορά του με όσο το δυνατόν περισσότερες λεπτομέρειες καθώς και τις ανάγκες και τις προτιμήσεις του με έναν τρόπο που ο χρήστης δεν μπορεί να παρατηρήσει, να καταλάβει ή να ελέγξει (Markellos et al., 2004b). Μεγάλη πρόκληση παραμένει η έλλειψη εμπιστοσύνης από την πλευρά των χρηστών και η έλλειψη δεδομένων για αυτούς από την πλευρά του web site για να βασίσει τις αποφάσεις εξατομίκευσης.

2.8 Βιβλιογραφία

- Adam, N., Dogramaci, O., Gangopadhyay, A., & Yesha, Y. (1999). *Electronic Commerce, Technical, Business, and Legal Issues*. Prentice Hall Inc, ISBN: 0139490825.
- Adomavicius, G., & Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: a Survey of the State-of-the-art and Possible Extensions. *IEEE Transactions on Knowledge Management and Data Engineering*, 17(6): 734-749.
- Aggarwal, C.C., Wolf, J.L., Wu, K.-L., & Yu, P.S. (1999). Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB*, J. Bocca, M. Jarke, & C., Zaniolo, (Eds.), Morgan Kaufmann, 487-499.
- Anderson, C., Domingos, P., & Weld, D. (2002). Relational Markov Models and their Application to Adaptive Web Navigation. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*. Edmonton, Alberta, Canada.
- Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet Recommendations Systems. *Journal of Marketing Research*, 363-375.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.
- Balabanovic, M., & Shoham, Y. (1997). Fab: Content-Based, Collaborative Recommendation. *Communication of ACM*, 40(3): 66-72.
- Basu, C., Hirsh, H., & Cohen, W. (1998). Recommendation as Classification: Using Social and Content-Based Information in Recommendation. *Recommender Systems Workshop*, Technical Report WS-98-08, AAAI Press.
- Belkin, N., & Croft, B. (1992). Information Filtering and Information Retrieval. *Communications of ACM*, 35(12): 29-37.
- Bellman, S., Lohse, G., & Johnson, E. (1999). Predictors of Online Buying Behavior. *Communications of the ACM*, 42(12): 32-38.
- Berendt, B. (2001). Understanding Web Usage at Different Levels of Abstraction: Coarsening and Visualizing Sequences. *Proceedings of the Mining Log Data Across All Customer TouchPoints*, San Francisco, USA.

- Bidigare, S. (2000). Information Architecture of the Shopping Cart. Best Practices for the Information Architectures of E-Commerce Ordering Systems. Argus Associates, Inc. Retrieved from: http://argus-acia.com/white_papers/shopping_cart_ia.pdf.
- Billsus, D., & Pazzani, M. (1998). Learning Collaborative Information Filters. *Proceedings of the International Conference on Machine Learning*.
- Billsus, D., & Pazzani, M. (1999). A Personal News Agent that Talks, Learns and Explains. *Proceeding of the 3rd Annual Conference on Autonomous Agents*.
- Billsus, D., & Pazzani, M. (2000). User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction*, 10(2-3): 147-180.
- Billsus, D., Brunk, C.A., Evans, C., Gladish, B., & Pazzani, M. (2002). Adaptive Interfaces for Ubiquitous Web Access. *Communications of ACM*, 45(5): 34-38.
- Blackwell, R., Miniard, P.W., & Engel, J. (2000). *Consumer Behavior*. South-Western College Pub.
- Breese, J.S., Heckerman, D., & Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*.
- Burke, R. (2000). Knowledge-Based Recommender Systems. *Encyclopedia of Library and Information Systems*, Kent, A. (Ed.), 69(32).
- Chakrabarti, S. (2003). *Mining Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann.
- Cheskin Research, & Studio Archetype Sapient (1999). *eCommerce Trust Study*. Retrieved from: <http://www.cheskin.com/p/ar.asp?mliid=7&arid=40&art=0>.
- Chien, Y.-H., & George, E.I. (1999). A Bayesian Model for Collaborative Filtering. *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*.
- Cho, Y.H., & Kim, J.K. (2004). Application of Web Usage Mining and Product Taxonomy to Collaborative Recommendations in E-commerce. *Expert Systems with Applications*, 26: 233-246.
- Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999). Combining Content-Based and Collaborative Filters in an Online Newspaper. *Proceedings of the ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*.
- Clerkin, P., Cunningham, P., & Hayes, C. (2001). Ontology Discovery for the Semantic Web Using Hierarchical Clustering. *Semantic Web Mining Workshop at ECML/PKDD*, Freiburg, Germany.
- Condliff, M., Lewis, D., Madigan, D., & Posse, C. (1999). Bayesian Mixed-Effects Models for Recommender Systems. *Proceedings of the ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*.
- Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. (2000). Learning to Construct Knowledge Bases from the World Wide Web. *Artificial Intelligence*, 118(1-2): 69-113.

- Cuthbert, M. (2000). The Six Basic Types of E-Shoppers. *E-Commerce Times*. Retrieved from: <http://www.ecommercetimes.com/story/4430.html>.
- Dai, H., & Mobasher, B. (2002). Using Ontologies to Discover Domain-level Web Usage Profiles. *Proceedings of the 2nd Semantic Web Mining Workshop at ECML/PKDD*, Helsinki, Finland.
- Delgado, J., & Ishii, N. (1999). Memory-Based Weighted-Majority Prediction for Recommender Systems. *Proceedings of the ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*.
- Deshpande, M., & Karypis, G. (2004). Item-Based Top-N Recommendation Algorithms. *ACM Transactions of Information Systems*, 22(1): 143-177.
- DTV (2005). Digital Dictionary. Retrieved from: http://www.iptv.org/digital/dictionary_internet.cfm.
- Duda, R.O., Hart, P.E., & Stork, D.G. (2001). *Pattern Classification*. John Wiley & Sons.
- Eirinaki, M., & Vazirgiannis, M. (2003). Web Mining for Web Personalization. *ACM Transactions on Internet Technology (TOIT)*, ACM Press New York, 3(1): 1-27.
- Eirinaki, M., Lampos, C., Paulakis, S., & Vazirgiannis, M. (2004). Web Personalization Integrating Content Semantics and Navigational Patterns. *Proceedings of 6th ACM International Workshop on Web Information and Data Management*, Washington, USA, 12-13 November 2004, 72-79.
- Eirinaki, M., Vazirgiannis, M., & Varlamis, I. (2003). SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process. *Proceedings of 9th ACM Conference on Knowledge Discovery and Data Mining*, Washington, USA, 99-108.
- Getoor, L., & Sahami, M. (1999). Using Probabilistic Relational Models for Collaborative Filtering. *Proceedings of the Workshop Web Usage Analysis and User Profiling (WEBKDD'99)*.
- Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2001). Learning Probabilistic Models of Relational Structure. *Proceedings of the 18th International Conference on Machine Learning*.
- Giugno, R., & Lukasiewicz, T. (2002). *P-SHOQ(D)*: A Probabilistic Extension of SHOQ(D) for Probabilistic Ontologies in the Semantic Web. *Proceedings of the 8th European Conference on Logics in Artificial Intelligence*, Cosenza, Italy.
- Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval Journal*, 4(2): 133-151.
- Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J.L., & Riedl, J. (1999). Combining Collaborative Filtering with Personal Agents for Better Recommendations. *Proceedings of the Conference of American Association of Artificial Intelligence (AAAI'99)*, 439-446.
- Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. MIT Press.
- Hertefelt, S. (2000). Trust and the Perception of Security. Interaction Architect. Retrieved from: <http://www.interactionarchitect.com/research/report20000103shd.htm>.

- Hill, A. (2001). Top 5 Reasons Your Customers Abandon Their Shopping Carts (and What You Can Do About It). Ziff Davis Smart Business. Retrieved from: http://www.findarticles.com/p/articles/mi_zdzb/is_200103/ai_ziff8142.
- Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and Evaluating Choices in a Virtual Community of Use. *Proceedings of the Conference on Human Factors in Computing Systems*.
- Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 289-296.
- Hofmann, T. (2003). Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. *Proceedings of the 26th Annual Intelligent ACM SIGIR Conference*.
- Hofmann, T. (2004). Latent Semantic Models for Collaborative Filtering. *ACM Transactions on Information Systems*, 22(1): 89-115.
- Horrocks, I. (2002). DAML+OIL: A Description Logic for the Semantic Web. *IEEE Data Engineering Bulletin*, 25(1): 4-9.
- Hudson, W. (2000). Metaphor: a Double-Edged Sword. *Interactions*, 7(3): 11-15.
- Hurst, M. (1999). White Paper: Building a Great Experience to Develop Brand, Increase Loyalty and Grow Revenues. ZD Internet.inc, Leadership Strategies for Transformation.
- Kim, W. (2002). Personalization: Definition, Status, and Challenges Ahead. *Journal of Object Technology*, 1(1): 29-40. Retrieved from: http://www.jot.fm/issues/issue_2002_05/column3.
- Kobsa, A., Koenemann, J., & Pohl, W. (2001). Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *The Knowledge Engineering Review*, 16(2): 111-155.
- Koch, N., & Turk, A. (1997). Towards a Methodical Development of Electronic Catalogues. *Electronic Markets*, 7(3): 28-31.
- Kosala, R., & Blockeel, H., (2000). Web Mining Research: a Survey. *SIGKDD Explorations*, 2(1), 1-15.
- Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (2001). Recommendation Systems: A Probabilistic Analysis. *Journal of Computer and System Sciences*, 63(1): p. 42-61.
- Lang, K. (1995). Newsweeder: Learning to Filter Netnews. *Proceedings of the 12th International Conference on Machine Learning*.
- Linden, G., Smith, B., & York, J. (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1): 76-80. Retrieved from: http://hugo.csie.ntu.edu.tw/~yjhsu/courses/u2010/papers/Amazon%20Recommendation_s.pdf.
- Loh, S., Wives, L.K., & de Oliveira, J.P. (2000). Concept-based Knowledge Discovery in Texts Extracted from the Web. *SIGKDD Explorations*, 2(1): 29-39.

Maedche, A., & Staab, S. (2000). Discovering Conceptual Relations from Text. *Proceedings of the European Conference on Artificial Intelligence (ECAI00)*. Berlin.

Markellos, K., Markellou, P., Rigou, M., & Sirmakessis, S. (2001). Modeling the Behaviour of e-Customers., *Proceedings of the PCHCI 2001 Panhellenic Conference with International Participation in Human-Computer Interaction*, Patras, Greece, 7-9 December, 333-338.

Markellos, K., Markellou, P., Rigou, M., & Sirmakessis, S. (2004a). Web Mining: Past, Present and Future. In Sirmakessis, S. (Ed.), *Proceedings of the 1st International Workshop on Text Mining & its Applications, Studies in Fuzziness and Soft Computing*, Springer-Verlag Berlin Heidelberg, 25-35.

Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2003). Designing and Developing an Effective Recommendation System for eCommerce Stores. *Proceedings of the WEBTEC Euromedia 2003*, Plymouth, UK, 50-54.

Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2002). Who is Today's E-Customer? A Description of his Behavioral Model. *Proceedings of the eBusiness and eWork 2002 Conference and Exhibition*, Prague, Czech Republic.

Markellos, K., Markellou, P., Rigou, M., Sirmakessis, S., & Tsakalidis, A. (2004b). Web Personalization and the Privacy Concern. *Proceedings of the 7th ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technologies, Challenges for the Citizen of the Information Society*, Syros, Greece, 14-16 April.

Markellou, P. (2001). Designing a Usable Shopping Cart. *Proceedings of the Workshop W2 Transforming Web Surfers to E-Shoppers, PCHCI 2001 Panhellenic Conference with International Participation in Human-Computer Interaction*, Patras, Greece, 57-61. Retrieved from: <http://www.hci.gr/files/pchci/Markellou.pdf>.

Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005a). Personalized E-commerce Recommendations. *IEEE International Conference on e-Business Engineering (ICEBE 2005)*, Beijing, China, 18-20 October 2005.

Markellou, P., Rigou, M., & Sirmakessis, S. (2005b). An Online Consumer Purchase Decision Cycle. In Clarke, I., & Flaherty, Th. (Eds.), *Advances in Electronic Marketing*, Idea Group Publishing Inc., 287-299.

Markellou, P., Rigou, M., & Sirmakessis, S. (2005c). Mining for Web Personalization. In Scime, A. (Ed.), *Web Mining: Applications and Techniques*, Idea Group Publishing Inc, ISBN: 1-59140-414-2, 27-48.

Markellou, P., Rigou, M., & Sirmakessis, S. (2005d). Product Catalog and Shopping Cart Effective Design. In Gao, Y. (Ed.), *Web Systems Design and Online Consumer Behavior*, Idea Group Publishing Inc, ISBN: 1-59140-328-6, 48-68.

Markellou, P., Rigou, M., & Sirmakessis, S. (2005e). Web Personalization for e-Marketing Intelligence. In Krishnamurthy, S. (Ed.), *Contemporary Research in E-Marketing*, 1, Idea Group Publishing Inc., ISBN: 1-59140-372-3, 232-250.

Markellou, P., Rigou, M., & Sirmakessis, S. (2006a). A Closer Look to the Online Consumer Behavior. *Encyclopedia of E-Commerce, E-Government and Mobile Commerce*, Khosrow-Pour, M. (Ed.), Idea Group Publishing Inc., to appear Spring 2006.

- Markellou, P., Rigou, M., & Sirmakessis, S. (2006b). Looking Inside Online Consumers Behavior. *International Journal of Social Behavior and Personality*, to appear.
- Marlin, B. (2003). Modeling User Rating Profiles for Collaborative Filtering. *Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS'03)*.
- Melville, P., Mooney, R.J., & Nagarajan, R. (2002). Content-Boosted Collaborative Filtering for Improved Recommendations. *Proceedings of the 18th National Conference on Artificial Intelligence*.
- Middleton, S.E., Shadbolt, N.R., & De Roure, D.C. (2004). Ontological User Profiling in Recommender Systems. *ACM Transactions on Information Systems*, 22(1): 54-88.
- Miller, B.N., Albert, I., Lam, S.K., Konstan, J.A., & Riedl, J. (2003). MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System. *Proceedings of the ACM Conference of Intelligent User Interfaces*, Chapel Hill, NC, USA.
- Mirza, B. (2001). Jumping Connections: A Graph-Theoretic Model for Recommender Systems. MSc Thesis, Virginia Polytechnic Institute and State University.
- Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic Personalization Based on Web Usage Mining. *Communications of the ACM*, 43(8): 142-151.
- Mobasher, B., Jin, X., & Zhou, Y. (2004). Semantically Enhanced Collaborative Filtering on the Web. *Proceedings of the European Web Mining Forum*, B. Berendt et al. (Eds.), Lecture Notes in Artificial Intelligence, 3209, Springer-Verlag.
- Montebello, M. (1998). Information Overload - An IR Problem?. *Proceedings of String Processing and Information Retrieval: A South American Symposium (SPIRE)*, 65. Retrieved from: <http://doi.ieeecomputersociety.org/10.1109/SPIRE.1998.712984>.
- Mooney, R.J., & Roy, L. (1999). Content-Based Book Recommending Using Learning for Text Categorization. *Proceedings of the ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*.
- Mooney, R.J., Bennett, P.N., & Roy, L. (1998). Book Recommending Using Text Categorization with Extracted Information. *Proceedings of the Recommender Systems Workshop*, Technical Report WS-98-08.
- Mulvenna, M., Anand, S., & Bchner, A. (2000). Personalization on the Net Using Web Mining. *Communications of the ACM*, 43(8): 122-125.
- Nakamura, A., & Abe, N. (1998). Collaborative Filtering Using Weighted Majority Prediction Algorithms. *Proceedings of the 15th International Conference on Machine Learning*.
- Nielsen, J., (1993). *Usability Engineering*. Morgan Kaufmann Publishers, Inc.
- Nielsen, J., (2000) E-commerce User Experience: Design Guidelines for Shopping Carts, Checkout, and Registration, Unpublished report, Nielsen Norman Group.
- Pavlov, D., & Pennock, D. (2002). A Maximum Entropy Approach to Collaborative Filtering in Dynamic, Sparse, High-Dimensional Domains. *Proceedings of the 16th Annual Conference on Neural Information Processing Systems (NIPS'02)*.

Pazzani, M. (1999). A Framework for Collaborative, Content-Based, and Demographic Filtering. *Artificial Intelligence Review*, 393-408.

Pazzani, M., & Billsus, D. (1997). Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning Journal*, 27: 313-331.

Peddy, C., & Armentrout, D., (2003). Building Solutions with Microsoft Commerce Server 2002. Microsoft Press.

Pennock, D.M., & Horvitz, E. (1999). Collaborative Filtering by Personality Diagnosis: A Hybrid Memory And Model-Based Approach. *Proceedings of the International Joint Conference on Artificial Intelligence Workshop: Machine Learning for Information Filtering*.

Pilioura, T. (1998). Electronic Payment Systems on Open Computer Networks: a Survey, Working Paper.

Popescul, A., Ungar, L.H., Pennock, D.M., & Lawrence, S. (2001). Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments. *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*.

Resnick, P., Iakovou, N., Sushak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings of the Computer Supported Cooperative Work Conference*.

Rewick, J. (2000). E-Tailers Try to Keep Shoppers from Bolting at Checkout Point. *The Wall Street Journal*.

Rohn, J. (1998). Creating Usable E-Commerce Sites. *StandardView*, 6(3): 110-115.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-Based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th International WWW Conference*.

Schein, A.I., Popescul, A., Ungar, L.H., & Pennock, D.M. (2002). Methods and Metrics for Cold-Start Recommendations. *Proceedings of the 25th Annual International ACM SIGIR Conference*.

Shani, G., Brafman, R., & Heckerman, D. (2002). An MDP-Based Recommender System. *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*.

Shardanand, U, & Maes, P. (1995). Social Information Filtering: Algorithms for Automating 'Word of Mouth'. *Proceedings of the Conference on Human Factors in Computing Systems*.

Si, L., & Jin, R. (2003). Flexible Mixture Model for Collaborative Filtering. *Proceedings of the 20th International Conference on Machine Learning*.

Smith, E.R. (2000). *e-Loyalty: How to Keep Customers Coming Back to Your Website*. HarperBusiness Edition.

Soboroff, I., & Nicholas, C. (1999). Combining Content and Collaboration in Text Filtering. *Proceedings of the International Joint Conference on Artificial Intelligence Workshop: Machine Learning for Information Filtering*.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2000). Fast Computation of Concept Lattices Using Data Mining Techniques. *Proceedings of the Knowledge Representation Meets Databases Conference (KRDB00)*, Berlin.

Tran, T., & Cohen, R. (2000). Hybrid Recommender Systems for Electronic Commerce. *Proceedings of the Knowledge-Based Electronic Markets, AAAI Workshop*, Technical Report WS-00-04, AAAI Press.

Ungar, L.H., & Foster, D.P. (1998). Clustering Methods for Collaborative Filtering. *Proceedings of the Recommender Systems Workshop*, Technical Report WS-98-08.

Yu, K., Schwaighofer, A., Tresp, V., Xu, X., & Kriegel, H.-P. (2004). Probabilistic Memory-Based Collaborative Filtering. *IEEE Transactions on Knowledge and Data Engineering*, 16(1): 56-69.

Yu, K., Xu, X., Tao, J., Ester, M., & Kriegel, H.-P. (2002). Instance Selection Techniques for Memory-Based Collaborative Filtering. *Proceedings of the 2nd SIAM International Conference on Data Mining (SDM'02)*.

Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and Redundancy Detection in Adaptive Filtering. *Proceedings of the 25th Annual International ACM SIGIR Conference*, 81-88.

Μαρκέλλου, Π. (2003). Recommendation Συστήματα. Κεφάλαιο στις διδακτικές σημειώσεις «e-Επιχειρείν» για το προπτυχιακό μάθημα «e-Επιχειρείν» που διδάσκεται στο Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής της Πολυτεχνικής Σχολής του Πανεπιστημίου Πατρών, Εκδόσεις Πανεπιστημίου Πατρών.

Μαρκέλλου, Π. (2005). Ηλεκτρονικό Κατάστημα: Σχεδιασμός και Λειτουργίες. Κεφάλαιο 3 στο βιβλίο *e-Επιχειρηματικότητα: από τη Θεωρία στην Πράξη*, Εκδόσεις Ελληνικά Γράμματα – Ε.Α.Ι.Τ.Υ. CTI press, Σειρά: Τεχνολογία Πληροφορικής και Επικοινωνιών, pp. 105-144, υπό έκδοση.

«The beginning of knowledge is the discovery of something we do not understand.»

Frank Herbert
US Science Fiction Novelist
(1920 - 1986)

Κεφάλαιο 3

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ

«Business Intelligence can be defined as the process of turning data into information and then into knowledge.» (Golfarelli et al., 2004).

3.1 Εισαγωγή

Τις τελευταίες δεκαετίες η συγκέντρωση τεράστιων ποσοτήτων δεδομένων από τις επιχειρήσεις, τα πανεπιστήμια και άλλους φορείς έχει δημιουργήσει την ανάγκη για γρήγορη, συνεπή και αποδοτική διαχείρισή τους. Τα δεδομένα αυτά είναι αποθηκευμένα σε μεγάλες **βάσεις δεδομένων (databases)** ή **αποθήκες δεδομένων (data warehouses)** και μπορεί να προέρχονται από την καταγραφή συστημάτων δοσοληψιών που λειτουργούν επί πολλά χρόνια π.χ. τραπεζικές συναλλαγές, τηλεφωνικές κλήσεις, αγορές με πιστωτικές κάρτες, κλπ. ή διαδικασίες που συσσωρεύουν μεγάλα ψηφιακά αρχεία π.χ. δορυφορικές φωτογραφίες, πειραματικά δεδομένα, κλπ. Σε αυτά τα δεδομένα συνήθως εφαρμόζονται α) **σχεσιακές ερωτήσεις** με στόχο την επιλογή ενός μικρού υποσυνόλου τους που ικανοποιεί κριτήρια ορισμένα από το χρήστη ή/και β) **στατιστική επεξεργασία** με τη μορφή αθροιστικών συναρτήσεων που εφαρμόζονται σε πολλές διαστάσεις. Και στις δύο περιπτώσεις οδηγούμαστε στην παραγωγή πολύ συγκεκριμένων αποτελεσμάτων τόσο ως προς τη δομή όσο και ως προς τη σημασία τους (Βαζιργιάννης & Χαλκίδη, 2003).

Από την άλλη πλευρά είναι πολύ πιθανό ο τεράστιος αυτός όγκος πληροφοριών να κρύβει μέσα του **ανεξερεύνητη «γνώση»**. Η γνώση αυτή, αν και όχι γνωστή εκ των προτέρων, μπορεί να αποδειχθεί εξαιρετικά χρήσιμη κυρίως για τις επιχειρήσεις. Συγκεκριμένα, συσχετίσεις μεταξύ των δεδομένων, εξαρτήσεις και αλληλουχίες γεγονότων, προβλέψεις για στόχους και πολλά άλλα είναι «θαμμένα» μέσα σε αυτά δεδομένα και περιμένουν με κάποιον τρόπο να έρθουν στην επιφάνεια. Το ερώτημα λοιπόν που κυριαρχεί είναι:

«Πώς αυτή η γνώση θα ανακαλυφθεί;».

Μια απάντηση στο ερώτημα αυτό μπορεί να έρθει από την περιοχή της **Επιχειρηματικής Ευφυΐας (Νοημοσύνης) ή Business Intelligence (BI)**. Η επιχειρηματική ευφυΐα ορίζεται ως:

«μια αμφίδρομη διαδικασία ανάλυσης και αξιοποίησης πληροφοριών από διάφορες πηγές που είναι αποθηκευμένες σε βάσεις δεδομένων με στόχο την εξαγωγή χρήσιμων συμπερασμάτων για τη λήψη αποφάσεων.» (McDonald et al., 2002).

Με τον τρόπο αυτό η διοίκηση μια επιχείρησης έχει τη δυνατότητα να παρακολουθεί αποτελεσματικότερα τις νέες τάσεις και τα πρότυπα, να μοιράζεται πληροφορίες με συνεργάτες ή πελάτες και να λαμβάνει καλύτερες επιχειρηματικές αποφάσεις. Συνεπώς, η BI αφορά **πελάτες, προϊόντα, υπηρεσίες και ανταγωνιστές**.

Η BI δεν είναι μια καινούργια έννοια στην περιοχή της Πληροφορικής αφού πρώιμες εκδοχές της συναντάμε στα τέλη της δεκαετίας του '70. Ένα πολύ γνωστό παράδειγμά της είναι η τοποθέτηση των προϊόντων στα ράφια των super-markets π.χ. οι σάλτσες κοντά στα ζυμαρικά, οι ξηροί καρποί κοντά στα ποτά, κλπ. Αυτές οι επιλογές δεν γίνονται τυχαία αλλά βασίζονται στην ανάλυση των καταναλωτικών συνηθειών των πελατών με διάφορες μεθόδους BI. Προχωρώντας λίγο παραπέρα, η BI μπορεί να δώσει ώθηση σε πολλές άλλες εφαρμογές από τον καθορισμό της ώρας προβολής των διαφημιστικών μηνυμάτων στην τηλεόραση έως τις παροχές των ασφαλιστήριων συμβολαίων.

Με το πέρασμα όμως των χρόνων οι επιχειρηματικές ανάγκες και απαιτήσεις αυξήθηκαν και μαζί με αυτές και η συσσώρευση του όγκου των πληροφοριών κάνοντας την ανακάλυψη της κρυμμένης γνώσης πιο επιτακτική από ποτέ. Ενώ αρχικά οι επιχειρήσεις περιορίζονταν στην οργάνωση και στην απλή επεξεργασία των αποθηκευμένων πληροφοριών σε δομές όπως οι βάσεις δεδομένων και τα συστήματα διαχείρισης και προγραμματισμού επιχειρηματικών πόρων (Enterprise Resource Planning ή ERP) στη συνέχεια με την αύξηση των μεγεθών οδηγηθήκαμε σε δομές όπως οι αποθήκες δεδομένων οι οποίες πλαισιώνονται και με ειδικές εφαρμογές διαχείρισης που μπορούν να αναλύσουν τις πληροφορίες πολυδιάστατα, παρέχοντας τη δυνατότητα για εξελιγμένες επιτελικές αναφορές (reporting).

Η γνώση είναι κρυμμένη γιατί οι παραπάνω δομές εξυπηρετούν συγκεκριμένους σκοπούς. Οι μέθοδοι που χρησιμοποιούνται για να έρθει στην επιφάνεια αναλύουν κυρίως τα δεδομένα, εξαγουν συμπεράσματα και βασίζονται σε στατιστικούς κανόνες. Το Data Mining είναι από τις πιο γνωστές μεθόδους. **Data Mining (DM)** ή **Εξόρυξη Γνώσης** είναι ένα βήμα στη διαδικασία **Ανακάλυψης Γνώσης από Βάσεις Δεδομένων** ή **Knowledge Discovery in Databases** ή **KDD**, η οποία *«χρησιμοποιεί συγκεκριμένους αλγόριθμους και τεχνικές με στόχο την αναγνώριση έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων (patterns) στα δεδομένα»* (Piatetsky-Shapiro et al., 1996). Η KDD διαδικασία περιλαμβάνει πολλά βήματα από την προετοιμασία των δεδομένων, την αναζήτηση προτύπων έως την αξιολόγηση της αποκτημένης γνώσης (Fayyad et al., 1996). Τα πρότυπα αυτά είναι πλούσια σε σημασιολογία αφού εκμεταλλεύονται τις πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων, ενώ δεν είναι δυνατόν να εξαχθούν με γυμνό μάτι από τον άνθρωπο-χρήστη της βάσης δεδομένων (Han & Kamber, 2000).

Το DM συνδυάζει διαφορετικές περιοχές όπως είναι: η Τεχνητή Νοημοσύνη (Artificial Intelligence ή AI), η Μηχανική Μάθηση (Machine Learning), η Διαχείριση Βάσεων Δεδομένων (Database Management), η Οπτικοποίηση Δεδομένων (Data Visualization), οι Μαθηματικοί Αλγόριθμοι (Mathematic Algorithms), η Στατιστική (Statistics), κλπ. (Liao, 2003). Οι τεχνικές DM είναι σχεδιασμένες να εφαρμόζονται σε **δομημένες βάσεις δεδομένων (structured databases)**. Οι τεχνικές **Text Mining (TM)** ή **Εξόρυξη Γνώσης από Κείμενα**

έχουν αναπτυχθεί για την ανακάλυψη γνώσης μέσα από την επεξεργασία **αδόμητων (unstructured) ή ημιδομημένων εγγράφων (semi-structured documents)**. Κατά συνέπεια, το TM αποτελεί μια επέκταση του παραδοσιακού DM και αφορά στην εξόρυξη υπονοούμενης γνώσης που περιλαμβάνεται σε μεγάλες συλλογές εγγράφων (Tan, 1999). Τα τελευταία χρόνια οι τεχνικές αυτές παρουσιάζουν ιδιαίτερο ενδιαφέρον αφού η πλειοψηφία των δεδομένων σήμερα σε ποσοστό μεγαλύτερο του 80% είναι μη δομημένα ή ημιδομημένα π.χ. κείμενα, εικόνες, έγγραφα, web pages, κλπ. Μάλιστα το ποσοστό αυτό αναμένεται να αυξηθεί στο μέλλον λόγω της ευρείας χρήσης του Internet και του e-mail.

Τα δεδομένα αυτά συνήθως είναι αποθηκευμένα σε μια βάση δεδομένων κειμένου και έχουν ημιδομημένη μορφή. Αυτό σημαίνει ότι μπορούν να περιλαμβάνουν κάποια δομημένα πεδία. Για παράδειγμα στην περίπτωση της περιγραφής ενός βιβλίου μπορεί να έχουμε τα δομημένα πεδία: τίτλος βιβλίου, συγγραφέας, εκδότης, ημερομηνία έκδοσης, κλπ. καθώς και ένα μεγάλο τμήμα μη δομημένου κειμένου π.χ. η περίληψη του βιβλίου, τα περιεχόμενά του, κλπ. Οι πρώτες προσπάθειες εξόρυξης γνώσης από κείμενα σχετίζονταν με τη χειροκίνητη κατασκευή ενός συνόλου κανόνων π.χ. αν η ιστοσελίδα παρουσίασης ενός βιβλίου περιλάμβανε τη φράση «δημιουργία με HTML» τότε το βιβλίο ταξινομούταν στην κατηγορία «βιβλία πληροφορικής». Αυτό απαιτούσε πολύ καλή γνώση του πεδίου και σημαντικό κόπο και χρόνο και πλέον είναι αδύνατο, αν λάβουμε υπόψη μας το τεράστιο όγκο δεδομένων που είναι διαθέσιμα σήμερα (El Wakil, 2002).

Οι πιο σημαντικές προκλήσεις που πρέπει να αντιμετωπίσει το TM είναι οι ακόλουθες:

- Οι συλλογές των δεδομένων είναι ελεύθερα κείμενα. Αυτό σημαίνει ότι τα δεδομένα δεν έχουν κάποια συγκεκριμένη οργάνωση και είναι ημιδομημένα ή αδόμητα.
- Τα κείμενα της καθημερινής γλώσσας περιλαμβάνουν ασάφειες σε πολλά επίπεδα, λεκτικές, συντακτικές, σημασιολογικές, κλπ.
- Οι τεχνικές για την επεξεργασία των κειμένων απαιτούν εκπαίδευση με δεδομένα.

3.1.1 Το πρόβλημα

Μια σημαντική πηγή γνώσεων στον τομέα της τεχνολογίας αποτελούν οι καταχωρήσεις των **διπλωμάτων ευρεσιτεχνίας ή πατέντες (patents)**. Ένα δίπλωμα ευρεσιτεχνίας είναι ένας νομικός τίτλος που χορηγεί στον κάτοχό του το αποκλειστικό δικαίωμα να χρησιμοποιήσει μια εφεύρεση σε συγκεκριμένη περιοχή και χρόνο, αποτρέποντας παράλληλα άλλους να τη χρησιμοποιήσουν ή να την πουλήσουν χωρίς έγκριση (EPO, 2005). Τα διπλώματα ευρεσιτεχνίας καταδεικνύουν το επίπεδο της καινοτόμου δραστηριότητας μιας αγοράς (Narin, 1995). Η εφαρμογή τεχνικών εξόρυξης γνώσης σε τέτοιου είδους δεδομένα μπορεί να οδηγήσει σε πολύ χρήσιμα συμπεράσματα σχετικά με την τεχνολογική ανάπτυξη, τις τάσεις της τεχνολογίας αλλά και τις καινοτομικές κατευθύνσεις (Chappelier et al., 2002). Η προτεινόμενη προσέγγιση αφορά στην εφαρμογή τεχνικών Text Mining σε διπλώματα ευρεσιτεχνίας που βρίσκονται καταχωρημένα σε βάσεις δεδομένων διαφόρων διεθνών οργανισμών διαχείρισής τους, με στόχο την παραγωγή επιστημονικών και τεχνολογικών δεικτών για την ανάδειξη του επιπέδου καινοτόμου δραστηριότητας μιας αγοράς.

Σύμφωνα με την προτεινόμενη προσέγγιση (Markellou et al., 2001), (Markellos et al., 2002a), (Markellos et al., 2002b), (Markellos et al., 2003a) αρχικά τα δεδομένα καθαρίζονται

προκειμένου να βελτιωθεί η ποιότητά τους πριν την επεξεργασία. Αυτή η διαδικασία συχνά αναφέρεται σαν **«αποφυγή της Garbage in – Garbage out επίδρασης ή GIGO»**. Στη συνέχεια, εφαρμόζονται δύο τύποι επεξεργασίας στα δεδομένα η **απλή ανάλυση (simple analysis)** και η **στατιστική ανάλυση (statistical analysis)**. Στην πρώτη περίπτωση παράγονται γραφήματα που συσχετίζουν τις πληροφορίες π.χ. κύριοι τομείς ανάπτυξης σε κάθε χώρα. Στη δεύτερη περίπτωση αναλύονται συντακτικά τα κείμενα των διπλωμάτων ευρεσιτεχνίας (τα πεδία abstract και title) και ομαδοποιούνται τα λήμματα των λέξεων. Η μορφολογική κανονικοποίηση αναφέρεται σε μια ομάδα διαδικασιών φυσικής γλώσσας (NLP – Natural Language Processing) για την παραγωγή κανονικών μορφών των λέξεων. Βασικός στόχος της παραπάνω διαδικασίας είναι η εξαγωγή της ρίζας των λέξεων από την κλιμένη ή παραγόμενη μορφή τους, με την εφαρμογή ενός συνόλου κανόνων. Αυτός ο τύπος μορφολογικής κανονικοποίησης είναι ευρέως γνωστός σαν διαδικασία **stemming (ανάλυση γενεαλογικού δένδρου)** και εμπεριέχει διαδικασίες **part of speech tagging (δήλωση σε ποιο μέρος του λόγου ένας όρος ανήκει)** για την εξαγωγή πληροφορίας όσον αφορά στους τομείς με τους οποίους έχει σχέση το κάθε δίπλωμα ευρεσιτεχνίας.

Στη συνέχεια πάνω στα δεδομένα των διπλωμάτων ευρεσιτεχνίας εφαρμόζονται τεχνικές **correspondence** και **clustering analysis** έτσι ώστε αυτά να ομαδοποιηθούν σύμφωνα με τις τεχνολογίες στις οποίες αναφέρονται. Τα clusters πλέον αυτά προβάλλονται όπως και στην απλή ανάλυση παρέχοντας στο χρήστη μια πιο λεπτομερή απεικόνιση της πληροφορίας των διπλωμάτων ευρεσιτεχνίας. Ο συνδυασμός των αναλύσεων που εφαρμόζονται με βάση την προτεινόμενη μεθοδολογία επιτρέπει την αποτύπωση των τεχνολογικών εξελίξεων και καινοτομιών. Οι δείκτες που παράγονται για κάθε cluster είναι πολύ σημαντικοί αφού μπορούν να ποσοτικοποιήσουν τις πληροφορίες που αφορούν σε συγκεκριμένες τεχνολογίες. Με αυτό τον τρόπο μπορούμε να παράγουμε δείκτες για τη δραστηριότητα συγκεκριμένων φορέων, εφευρετών, χωρών, κλπ. Τέλος, τεχνολογικοί δείκτες που υποδεικνύουν μελλοντικές ελπιδοφόρες τεχνολογίες καθώς και ποιοι φορείς θα είναι πρωτοπόροι σε αυτές μπορούν να εξαχθούν.

3.1.2 Η δομή του κεφαλαίου

Στο κεφάλαιο αυτό διερευνούμε τον τρόπο που οι τεχνικές Text Mining μπορούν να εφαρμοστούν σε δεδομένα από διπλώματα ευρεσιτεχνίας με στόχο την παραγωγή επιστημονικών και τεχνολογικών δεικτών για την ανάδειξη του επιπέδου καινοτόμου δραστηριότητας μιας αγοράς. Συγκεκριμένα, η δομή του κεφαλαίου είναι η ακόλουθη. Στην επόμενη ενότητα γίνεται μια σύντομη παρουσίαση της διαδικασίας ανακάλυψης γνώσης. Στη συνέχεια μελετάται το πρόβλημα της εξόρυξης γνώσης από διπλώματα ευρεσιτεχνίας, καταγράφεται το κίνητρο για την παραγωγή τεχνολογικών και επιστημονικών δεικτών, περιγράφονται οι πληροφορίες που περιέχει ένα δίπλωμα ευρεσιτεχνίας και παρουσιάζεται η μεθοδολογική προσέγγισή μας για την υποστήριξη της επιχειρηματικής ευφυΐας. Ακολουθεί η περιγραφή της διαδικασίας επιλογής των δεδομένων, της γλωσσολογικής προ-επεξεργασίας τους και των τεχνικών ανάλυσης που χρησιμοποιούνται, ενώ δίνονται το αρχιτεκτονικό σχήμα και ενδεικτικά αποτελέσματα των υλοποιημένων case studies. Το κεφάλαιο ολοκληρώνεται με την παράθεση των σημαντικότερων συμπερασμάτων καθώς και την καταγραφή ανοικτών θεμάτων για μελλοντική έρευνα.

3.2 Η διαδικασία ανακάλυψης γνώσης

Η διαδικασία ανακάλυψης γνώσης περιλαμβάνει τα ακόλουθα **βασικά στάδια** (Σχήμα 29) (Bao, 2004):

- Το **πρώτο βήμα** αφορά στην κατανόηση της περιοχής της εφαρμογής και στη διαπίστωση του προβλήματος. Αυτό το βήμα αποτελεί προϋπόθεση για την εξαγωγή χρήσιμης γνώσης και για την επιλογή των κατάλληλων μεθόδων εξόρυξης δεδομένων (τρίτο βήμα) σύμφωνα με το σκοπό της εφαρμογής, το στόχο του τελικού χρήστη και τη φύση των δεδομένων. Η KDD διαδικασία διακρίνεται με βάση το σκοπό της σε α) **επαλήθευση** όταν το σύστημα καλείται να επιβεβαιώσει την υπόθεση που έχει κάνει ο χρήστης και β) **ανακάλυψη** όταν το σύστημα βρίσκει αυτόνομα νέα πρότυπα. Η ανακάλυψη μπορεί να χωριστεί περαιτέρω σε **πρόβλεψη** όταν το σύστημα βρίσκει πρότυπα με σκοπό την πρόβλεψη της μελλοντικής συμπεριφοράς κάποιων οντοτήτων και σε **περιγραφή** όταν το σύστημα βρίσκει πρότυπα με σκοπό την παρουσίασή τους σε κάποιο χρήστη σε κατανοητή και απτή μορφή.
- Το **δεύτερο βήμα** αφορά στη συλλογή και στην προ-επεξεργασία των δεδομένων, συμπεριλαμβανομένων και των πηγών τους, στην απομάκρυνση του θορύβου και οποιονδήποτε άλλων περιττών στοιχείων, στην επεξεργασία των χαμένων δεδομένων, στο μετασχηματισμό και στη μείωση των δεδομένων, κλπ. Αυτό το βήμα απαιτεί συνήθως τον περισσότερο χρόνο της διαδικασίας KDD.

Σχήμα 29. Η διαδικασία KDD.

- Το **τρίτο βήμα** αφορά στην επιλογή της κατάλληλης μεθόδου εξόρυξης (στόχοι, απαιτήσεις και γενικά κριτήρια KDD διαδικασίας) και στην εφαρμογή της για την εξόρυξη των δεδομένων και την ανακάλυψη των **προτύπων (patterns) γνώσης** ή/και των **μοντέλων (models)** που κρύβονται στα δεδομένα. Ένα πρότυπο είναι «μια τοπική δομή, που συνήθως σχετίζεται με μια ομάδα μεταβλητών και περιπτωσιακές μελέτες», ενώ ένα μοντέλο μπορεί να αντιμετωπισθεί ως «μια

σφαιρική αναπαράσταση μιας δομής που συνοψίζει τα συστηματικά συστατικά που κρύβονται κάτω από τα δεδομένα ή που περιγράφει πώς τα δεδομένα μπορεί να έχουν προκύψει». Ένα σημαντικό πλήθος μεθόδων εξόρυξης δεδομένων έχουν προταθεί για να ικανοποιήσουν τις απαιτήσεις των διαφορετικών εφαρμογών:

- **Clustering** (ομαδοποίηση ή συσταδοποίηση).
- **Classification** (κατηγοριοποίηση ή ταξινόμηση).
- **Association rules** (κανόνες συσχέτισης).
- **Time series** (χρονολογικές σειρές).
- **Regression** (παλινδρόμηση).
- **Summarization** (περιληπτική παρουσίαση της πληροφορίας).

Η απόδοση και τα αποτελέσματα της επιλεγμένης μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα βήματα.

- Το **τέταρτο βήμα** είναι να ερμηνευθεί και να αξιολογηθεί η εξορυγμένη γνώση (μετα-επεξεργασία). Τα πειράματα δείχνουν ότι τα εξαγόμενα πρότυπα/μοντέλα δεν είναι πάντα ενδιαφέροντα ή άμεσης χρήσης. Για το σκοπό αυτό αξιολογούνται με τη χρήση καθορισμένων μετρικών ώστε να προσδιοριστούν αυτά που αντιπροσωπεύουν τη γνώση. Ένα πρότυπο είναι **ενδιαφέρον** όταν είναι εύκολα κατανοητό, πιθανότατα χρήσιμο, καινούριο ή επικυρώνει μια υπόθεση την οποία θέλει να επιβεβαιώσει ο χρήστης. Ο βαθμός ενδιαφέροντος ενός προτύπου μπορεί να διαχωριστεί σε **αντικειμενικό** και **υποκειμενικό**, ανάλογα με το αν βασίζεται σε στατιστικά δεδομένα ή στην κρίση του χρήστη. Για παράδειγμα στη βάση δεδομένων ενός ηλεκτρονικού καταστήματος που νοικιάζει κινηματογραφικές ταινίες στους πελάτες του μπορεί να εξαχθούν τα ακόλουθα δύο πρότυπα:

«Όταν Πόλη=Πάτρα τότε και Νομός=Αχαΐα»

και

«Όταν Ταινία=Σικάγο τότε και Ταινία=Μουλέν Ρουζ».

Το πρώτο πρότυπο είναι χωρίς ενδιαφέρον αφού είναι προφανές και δεν προσφέρει καινούργια γνώση. Το δεύτερο όμως έχει ενδιαφέρον, είναι καινούριο και χρήσιμο για το μηχανισμό συστάσεων του καταστήματος αφού δείχνει τη μη προφανή συσχέτιση δύο διαφορετικών ταινιών. Επίσης, πρέπει να σημειωθεί ότι η KDD είναι μια **επαναληπτική** διαδικασία. Ένας τυποποιημένος τρόπος για να αξιολογηθούν τα πρότυπα είναι να διαιρεθούν τα δεδομένα σε δύο σύνολα: το πρώτο σύνολο να χρησιμοποιηθεί για την εκπαίδευση και το δεύτερο σύνολο για τον έλεγχο. Η διαδικασία μπορεί να επαναληφθεί αρκετές φορές και στο τέλος της να υπολογισθούν οι μέσες τιμές των αποτελεσμάτων και ακολούθως η απόδοσή τους.

- Το **πέμπτο και τελικό βήμα** είναι να τεθεί η εξορυγμένη γνώση σε **πρακτική χρήση**. Αυτό μπορεί να σημαίνει την ενσωμάτωσή της σε κάποιο σύστημα ή απλώς την παρουσίαση της στον τελικό χρήστη. Επίσης, στο βήμα αυτό ελέγχονται και επιλύονται τυχόν συγκρούσεις με προηγούμενη εξορυγμένη γνώση. Η πρακτική χρήση των αποτελεσμάτων αποτελεί ίσως το σημαντικότερο στόχο της διαδικασίας KDD.

3.3 Εξόρυξη γνώσης από βάσεις διπλωμάτων ευρεσιτεχνίας για την παραγωγή τεχνολογικών και επιστημονικών δεικτών

3.3.1 Το κίνητρο για την παραγωγή τεχνολογικών και επιστημονικών δεικτών

Στις μέρες μας προκειμένου να παρακολουθήσουμε και να αναλύσουμε τις οικονομικές, τεχνολογικές και επιστημονικές δραστηριότητες είναι απαραίτητο να λάβουμε υπόψη μας τις ροές των πληροφοριών που σχετίζονται με αυτές. Συνήθως αυτές οι πληροφορίες είναι αποθηκευμένες σε μεγάλες βάσεις δεδομένων και διακρίνονται σε διάφορες κατηγορίες. Για παράδειγμα οι πληροφορίες που σχετίζονται με την έρευνα είναι αποθηκευμένες στις ερευνητικές δημοσιεύσεις, στα επιστημονικά περιοδικά, κλπ. Οι πληροφορίες που σχετίζονται με την ανάπτυξη και την παραγωγή είναι κυρίως αποθηκευμένες στα **διπλώματα ευρεσιτεχνίας** ή **πατέντες**. Οι πατέντες συνδέονται στενά με τις τεχνολογικές και επιστημονικές δραστηριότητες δηλαδή με το **Technology Watch**. Θα μπορούσαμε να ορίσουμε το Technology Watch σαν «*την παρακολούθηση της ανάπτυξης των νέων τεχνολογιών, των νέων προϊόντων και των τάσεων της τεχνολογίας και τη μέτρηση της επίδρασής τους στους ανθρώπους, στις επιχειρήσεις, στους οργανισμούς, αλλά και στις τεχνολογικές εξελίξεις γενικότερα*». Συνεπώς, σχετίζεται άμεσα με την **καινοτομία (innovation)**. Η στατιστική εκμετάλλευση αυτών των πληροφοριών μπορεί να οδηγήσει στην εξαγωγή χρήσιμων συμπερασμάτων για την τεχνολογική ανάπτυξη καθώς και τις μελλοντικές κατευθύνσεις και τάσεις τόσο της τεχνολογίας όσο και της καινοτομίας.

Η πρόσβαση στην επιστημονική και τεχνολογική γνώση και η δυνατότητα εκμετάλλευσής της αποτελούν κρίσιμους παράγοντες για την απόδοση των επιχειρήσεων, των οργανισμών, των ιδιωτών, ακόμα και των χωρών. Παράλληλα, η μέτρηση της καινοτομίας παρουσιάζει ενδιαφέρον για πολλά πεδία της οικονομικής και τεχνολογικής δραστηριότητας. Για το λόγο αυτό οι σχεδιαστές πολιτικών έρευνας και τεχνολογίας, οι επιχειρηματικοί αναλυτές, οι αναλυτές πατεντών, τα εθνικά και διεθνή γραφεία διαχείρισης των πατεντών, οι οικονομικοί οργανισμοί, τα εθνικά στατιστικά γραφεία, καθώς και άλλοι βιομηχανικοί φορείς και οργανισμοί με σημαντική επιστημονική και τεχνολογική δραστηριότητα ενδιαφέρονται για την ανακάλυψη πληροφοριών «κρυμμένων» στις βάσεις δεδομένων πατεντών. Οι πατέντες είναι μια πηγή πληροφοριών που όχι μόνο υποστηρίζει την έρευνα που σχετίζεται με τις τεχνολογικές εξελίξεις αλλά επηρεάζει επίσης και τις τεχνολογικές τάσεις αφού η απουσία ή η κυριαρχία μιας συγκεκριμένης καινοτομίας μπορεί να εξαχθεί μέσω της ανάλυσής τους. Αυτή η σχέση μεταξύ έρευνας και καινοτομίας απεικονίζεται στο Σχήμα 30.

Η ανάλυση των πληροφοριών που «κρύβονται» στις βάσεις δεδομένων πατεντών, μπορεί να δώσει μια πολύ σαφή άποψη των σύγχρονων τάσεων σχετικά με την τεχνολογική και την επιστημονική καινοτομία και μπορεί να παρέχει μετρήσιμα και συγκρίσιμα αποτελέσματα (Comanor & Scherer, 1969), (Narin, 1995). Παραδοσιακά, οι εμπειρογνώμονες βασίζονταν στη χειρωνακτική ανάλυση για την εξόρυξη της γνώσης. Σήμερα αυτό είναι αδύνατο καθώς οι βάσεις δεδομένων πατεντών περιέχουν τεράστιο όγκο δεδομένων ο οποίος αυξάνεται εκθετικά. Αυτό το γεγονός οδηγεί στην απαίτηση επέκτασης των χρησιμοποιούμενων μεθοδολογιών και εργαλείων ή ανάπτυξης νέων ώστε να διασφαλιστούν εκείνες οι συνθήκες

που θα επιτρέψουν την παρακολούθηση και την αποτίμηση των συνεχιζόμενων αλλαγών και των αποτελεσμάτων τους καθώς επίσης και την καλύτερη απεικόνισή τους.

Σχήμα 30. Η αμφίδρομη σχέση της έρευνας και της τεχνολογίας συνδυασμένη μέσω πατεντών.

3.3.2 Περιγραφή μιας πατέντας

Γενικά «μια **πατέντα** είναι ένας νομικός τίτλος που παρέχει στον κάτοχό του το αποκλειστικό δικαίωμα να χρησιμοποιήσει μια εφεύρεση για μια περιορισμένη περιοχή και χρόνο εμποδίζοντας άλλους από το να την υλοποιήσουν, να τη χρησιμοποιήσουν ή να την πουλήσουν χωρίς έγκριση» (EPO, 2005). Ο δικαιούχος για την κατοχύρωση της πατέντας πρέπει να παρέχει μια λεπτομερή τεχνική περιγραφή της εφεύρεσής του αλλά και να αναφέρει τα σημεία που την καθιστούν ως μια αυθεντική εφαρμογή με καινοτόμα στοιχεία. Μία εφεύρεση κρίνεται νέα αν δεν ανήκει στη στάθμη της τεχνικής και εμπεριέχει εφευρετική δραστηριότητα, αν δηλαδή κατά την κρίση ειδικού, δεν προκύπτει με προφανή τρόπο από την υπάρχουσα στάθμη της τεχνικής. Επιδεκτική βιομηχανικής εφαρμογής είναι εφόσον μπορεί να παραχθεί και να χρησιμοποιηθεί σε οποιοδήποτε τομέα της παραγωγικής δραστηριότητας. Η **διαδικασία χορήγησης** περιλαμβάνει τα ακόλουθα βήματα:

- την κατάθεση της αίτησης.
- την προθεσμία 4 μηνών για τυχόν διορθώσεις ή συμπλήρωση ελλείψεων.
- τον έλεγχο για το «νέο» και την εφευρετική δραστηριότητα (σύνταξη έκθεσης έρευνας).
- την προθεσμία 3 μηνών για παρατηρήσεις του καταθέτη στην έκθεση έρευνας.
- τη σύνταξη τελικής έκθεσης έρευνας.
- την έκδοση της πατέντας.

Τα δικαιώματα που παρέχονται με την πατέντα αποκτώνται και διατηρούνται σε ισχύ με την καταβολή αντίστοιχων τελών στον **ΟΒΙ (Οργανισμός Βιομηχανικής Ιδιοκτησίας)** που είναι ο φορέας που χορηγεί τις πατέντες στην Ελλάδα (ΟΒΙ, 2005).

Μια πατέντα αποτελείται από διάφορα **πεδία**, κάθε ένα από τα οποία περιέχει συγκεκριμένες πληροφορίες. Ένας **κωδικός** (συνχνά χρησιμοποιούνται περισσότεροι από ένας κωδικοί) χαρακτηρίζει την πατέντα δηλώνοντας τα τεχνικά χαρακτηριστικά της. Γενικά, οι πληροφορίες που περιλαμβάνονται στις πατέντες μπορούν να ταξινομηθούν σε τέσσερις μεγάλες κατηγορίες όπως φαίνεται και στο Σχήμα 31. Η πρώτη κατηγορία αφορά στα **τεχνολογικά**

χαρακτηριστικά γνωρίσματα της εφεύρεσης, ενώ η δεύτερη αναφέρεται σε όλα τα χαρακτηριστικά που σχετίζονται με την **εφαρμογή** της. Επιπλέον, παρέχονται πληροφορίες για τις **χώρες** στις οποίες μια εφεύρεση προστατεύεται καθώς επίσης και πληροφορίες για την **προέλευση** της εφεύρεσης. Όπως είναι προφανές υπάρχουν περιπτώσεις πατεντών για τις οποίες δεν είναι διαθέσιμες όλες αυτές οι πληροφορίες και αυτό μπορεί να οφείλεται είτε σε λάθος ή σε άλλους παράγοντες.

Σχήμα 31. Πληροφορίες που περιέχονται σε μια πατέντα.

Μεταξύ όλων των πεδίων, ιδιαίτερη έμφαση δίνεται στον κωδικό που χαρακτηρίζει την πατέντα δεδομένου ότι συσχετίζεται συνήθως με τα τεχνικά χαρακτηριστικά της εφεύρεσης και επομένως μπορεί να κατανέμει τις εφευρέσεις σε συγκεκριμένους τομείς της τεχνολογίας. Αυτοί οι κωδικοί δίνονται στις πατέντες με βάση το **International Patents Classification System (IPC)** ή κάποιο άλλο σύστημα. Το IPC έχει υιοθετηθεί από 52 χώρες και 4 διεθνείς οργανισμούς. Σύμφωνα με αυτό το σύστημα οι πατέντες ταξινομούνται με βάση την τεχνολογία που σχετίζεται με την εφεύρεση. Το IPC έχει ιεραρχική δομή και ταξινομείται σε sections, τα οποία με τη σειρά τους υποδιαιρούνται σε subsections, τα subsections σε classes, οι classes σε subclasses, οι subclasses σε groups, τα groups σε subgroups, κ.ο.κ. Κάθε επίπεδο σε αυτή την ιεραρχία αναπαρίσταται από μια κωδικοποίηση. Αν και το IPC είναι ένα διεθνές σύστημα υπάρχουν κάποια **γραφεία πατεντών (patent offices)** που χρησιμοποιούν τα δικά τους συστήματα π.χ. το **United States Patent Office** χρησιμοποιεί το **USPOC (United States Patent Office Classification)**.

Ένας σημαντικός αριθμός από βάσεις δεδομένων που περιέχουν πληροφορίες για πατέντες είναι διαθέσιμος. Η ποιότητα των παραγόμενων αποτελεσμάτων εξαρτάται σε πολύ μεγάλο βαθμό από την επιλογή των πληροφοριών που συλλέγονται από αυτές τις βάσεις και οι περισσότερες προσφέρουν τη δυνατότητα ανάκτησης πληροφοριών με τη χρήση διαφόρων κριτηρίων. Διακρίνονται δε στις **online βάσεις δεδομένων** π.χ. WPI(L), EPAT, US, CLAIMS, κλπ. και αυτές που διατίθενται σε **CD-ROM** π.χ. ESPACE ACCESS, ESPACE-FIRST, ESPACE BULLETIN, κλπ. Οι τελευταίες έχουν παραχθεί από το EPO και περιλαμβάνουν πληροφορίες για τις ευρωπαϊκές πατέντες. Ειδικά η **ESPACE ACCESS** αποτελεί ένα εργαλείο αναζήτησης που προσφέρει ολοκληρωμένες βιβλιογραφικές πληροφορίες και περιλήψεις (abstracts) στην αγγλική γλώσσα όλων των ευρωπαϊκών πατεντών και όλων των διεθνών εφαρμογών από την ίδρυση του EPO το 1978 και έπειτα.

Στη συνέχεια παρουσιάζονται τα πεδία που περιγράφουν μια πατέντα στη βάση δεδομένων ESPACE ACCESS (Σχήμα 32):

- **PN:** Priority Number (ο αριθμός της πατέντας).
- **AN:** Application Number (ο αριθμός της αίτησης).
- **PR:** Priority Year (το έτος έναρξης ισχύος της πατέντας).
- **DS:** Designated States (οι χώρες ισχύος της πατέντας).
- **MC:** Main Classification Codes (η κύρια κατηγοριοποίηση).
- **IC:** All Classification (η δευτερεύουσα κατηγοριοποίηση).
- **ET:** English Title (ο αγγλικός τίτλος).
- **FT:** French Title (ο γαλλικός τίτλος).
- **IN:** Inventor (ο εφευρέτης).
- **PA:** Applicant (η επωνυμία της εταιρίας κατάθεσης της πατέντας).
- **AB:** English Abstract (η αγγλική περίληψη).
- **AF:** French Abstract (η γαλλική περίληψη).

Patent number 2	
2	PN: EP 0965753 A1 19991222
2	AN: EP 99304418 19990608
2	PR: ZA 9805172 19980615
2	DS: AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
2	MC: F03D- 1/06
2	IC: F03D- 3/06, F03B- 1/00, F03B- 17/00
2	ET: Fluid rotor with spherical vanes
2	FT: Rotor entra nι par un fluide avec des pales sphiriques
2	GT: Flóssigkeitsgetriebener rotor mit Sphórischen Schaufeln
2	IN: Patel, Dinesh
2	PA: Patel, Dinesh
2	ND: ESPACE 99/094 FIRST 99/006
2	AB: "The invention provides an improved vane system comprising of two or more vanes (10) being part of a .....
2	HIT _LIST_CDREF: ESPACE 99/094 FIRST 99/006

Σχήμα 32. Παράδειγμα πατέντας.

Κάθε πεδίο που συνδέεται με την πατέντα παρέχει πληροφορίες για τη συγκεκριμένη εφεύρεση και η στατιστική ανάλυση είναι δυνατόν να χρησιμοποιήσει διάφορα από αυτά τα πεδία. Στην προτεινόμενη προσέγγιση η ανάλυση βασίζεται είτε στα πεδία των **κωδίκων (IPC)** είτε στα πεδία των **τίτλων/περιλήψεων (titles/abstracts)** προκειμένου να δημιουργηθούν για παράδειγμα ομοιογενείς **ομάδες (clusters)** τεχνολογίας. Εντούτοις, συμπληρωματικές πληροφορίες που περιγράφονται από άλλα πεδία π.χ. ο εφευρέτης χρησιμοποιούνται για την παροχή όλων των πληροφοριών των αποτελεσμάτων που προκύπτουν από την ανάλυση των βασικών πεδίων των πατεντών.

Είναι προφανές ότι οι βάσεις δεδομένων λόγω της πολυπλοκότητας της δομής των πατεντών παρέχουν μια πηγή πληροφοριών με πολλές μεταβλητές. Επομένως, η

εκμετάλλευση των χρήσιμων πληροφοριών μπορεί να είναι μια δύσκολη διαδικασία όχι μόνο για τα άτομα με λίγη εμπειρία στην περιοχή αλλά και για τους πιο έμπειρους που πρέπει να ξοδέψουν πολύ χρόνο και να καταβάλουν μεγάλη προσπάθεια για την παραγωγή χρήσιμων αποτελεσμάτων. Η πολυπλοκότητα του προβλήματος καθιστά προφανές ότι η ανάλυση των πατεντών δεν είναι εύκολη και επιπλέον η ανάπτυξη ενός συστήματος εξειδικευμένου στην ανάλυση πατεντών απαιτεί το συνδυασμό γνώσεων από διαφορετικές περιοχές όπως η στατιστική, η ανακάλυψη γνώσης από βάσεις δεδομένων, η ανάλυση πατεντών, οι στατιστικοί δείκτες καθώς επίσης και η E&A και οι δείκτες τεχνολογίας.

3.3.3 Η μεθοδολογική προσέγγιση

Οι βασικοί στόχοι της προτεινόμενης μεθοδολογικής προσέγγισης είναι α) να υποστηριχθούν πολυδιάστατες συγκρίσεις μεταξύ χωρών, τομέων ή επιχειρήσεων, β) να προσδιοριστεί ο ανταγωνισμός σε ένα δεδομένο τομέα και γ) να συλληφθούν οι αλληλεπιδράσεις που μπορεί να υπάρχουν μεταξύ των περιοχών της τεχνολογικής δραστηριότητας και των πόλων καινοτομίας μέσα σε αυτές τις περιοχές (Markellos et al., 2001a). Για να επιτευχθούν αυτοί οι στόχοι υιοθετήθηκαν μια σειρά από τεχνικές. Συγκεκριμένα:

- **Τεχνικές προ-επεξεργασίας κειμένων.**
- **Τεχνικές στατιστικής ανάλυσης πολλών μεταβλητών: Correspondence Analysis και Cluster Analysis.**
- **Τεχνικές απλής ανάλυσης δεδομένων: Simple Analysis.**

Όπως έχει ήδη αναφερθεί μια πατέντα είναι μια συλλογή των διαφορετικών πεδίων που τη χαρακτηρίζουν. Ανάλογα με τις πληροφορίες για τις οποίες κάποιος ενδιαφέρεται, μερικά από αυτά τα πεδία ή όλα μπορούν να χρησιμοποιηθούν στην ανάλυση. Στη διαδικασία του clustering είναι απαραίτητο να χρησιμοποιηθούν τα πεδία που περιγράφουν τα τεχνικά χαρακτηριστικά μιας πατέντας προκειμένου να ταξινομηθούν αυτές κατά ομάδες που εκφράζουν παρόμοιες τεχνολογίες. Συνεπώς, είναι απαραίτητο σε αυτή την περίπτωση να χρησιμοποιηθούν οι IPC κωδικοί και οι τίτλοι ή οι περιλήψεις. Αυτά τα πεδία θεωρούνται ως οι κύριες μεταβλητές, ενώ όλα τα άλλα πεδία είναι οι συμπληρωματικές μεταβλητές. Είναι κοινή αίσθηση ότι οι συμπληρωματικές μεταβλητές χρησιμοποιούνται για την παροχή μιας πλήρους περιγραφής των κύριων μεταβλητών και ακολούθως των πατεντών.

Μόλις επιλεγούν τα δεδομένα, μια σειρά από διαφορετικού τύπου αναλύσεις είναι διαθέσιμες στο χρήστη. Αρχικά, εκτελείται η γλωσσολογική προ-επεξεργασία των δεδομένων. Αυτή με τη σειρά της μπορεί να υποδιαιρεθεί σε διαφορετικές ενέργειες όπως η ληματοποίηση των όρων, η δήλωση του μέρους του λόγου που ανήκει ένας όρος, η σημασιολογική ανάλυση κλπ. μέσω της χρήσης συγκεκριμένων λεξικών. Κατόπιν εφαρμόζεται η cluster ανάλυση στα δεδομένα. Τα δεδομένα εισόδου είναι οι κωδικοί IPC και οι λέξεις που προέκυψαν από τη ληματοποίηση των δεδομένων των τίτλων και των περιλήψεων. Τελικά εκτελείται μια ιεραρχική ανάλυση των συστάδων στους παραγοντικούς άξονες που λαμβάνονται από την correspondence ανάλυση. Στην περίπτωση που ο αριθμός των πατεντών που συγκεντρώνονται είναι μεγαλύτερος από έναν προκαθορισμένο αριθμό (π.χ. 500 πατέντες) τότε εφαρμόζεται k-means clustering πριν από το ιεραρχικό clustering. Για την ανάλυση αυτή ο αριθμός των K συστάδων είναι ήδη γνωστός. Κατόπιν το ιεραρχικό clustering παίρνει ως είσοδο τις K συστάδες που λαμβάνονται από τη διαδικασία k-means (Alderferer &

Blashfield, 1986). Μέσω της ανάλυσης των συστάδων οι πατέντες ομαδοποιούνται σε συστάδες που έχουν παρόμοια χαρακτηριστικά, με άλλα λόγια σε συστάδες που περιέχουν πατέντες που περιγράφουν παρόμοιες τεχνολογίες. Διαφορετικές γραφικές παραστάσεις και πίνακες μπορούν να ληφθούν για την περιγραφή των συστάδων. Ιδιαίτερο ενδιαφέρον παρουσιάζει ο χάρτης σχέσεων που απεικονίζει τις σχέσεις μεταξύ των συστάδων σύμφωνα με το βαθμό ομοιότητάς τους. Αναφέρουμε επίσης ότι ο αριθμός των συστάδων επιλέγεται μέσω μιας αυτόματης διαδικασίας. Εντούτοις, ο χρήστης έχει τη δυνατότητα να επιλέξει εάν το επιθυμεί τον αριθμό των συστάδων που θεωρεί κατάλληλο για την εκτέλεση της ανάλυσης. Αυτό γίνεται είτε υποκειμενικά είτε με τη χρήση ενός εμπειρικού κανόνα βασισμένου στο δένδρογραμμα (dendrogram).

Εκτός από τη διαδικασία clustering υπάρχει η δυνατότητα να χρησιμοποιηθούν στατιστικές μέθοδοι όπως η correspondence analysis (Benzecri, 1992) για την περιγραφή των πατεντών μέσω ενός λεξιλογίου που είναι χαρακτηριστικό για τις πατέντες και «χτισμένο» βάσει των διάφορων περιγραφικών μεταβλητών που συνδέονται με αυτές. Κάποιος μπορεί να χρησιμοποιήσει στην ανάλυση διαφορετικά πεδία όπως οι εφευρέτες, οι χώρες, κλπ. Η χρήση παραγοντικών γραφικών παραστάσεων επιτρέπει την απεικόνιση σε μια δυσδιάστατη γραφική παράσταση των διαφορετικών μεταβλητών. Κατά συνέπεια, ο προσδιορισμός της θέσης των διαφορετικών μεταβλητών στους άξονες επιτρέπει την ερμηνεία της τρέχουσας τεχνολογίας. Επιπλέον, ο χρήστης έχει τη δυνατότητα να επέμβει στην παραγωγή των αποτελεσμάτων με τον καθορισμό διαφορετικών παραμέτρων σχετικών με την επιλογή των παραγοντικών αξόνων, τις συνεισφορές κάθε μονάδας για τη δημιουργία των αξόνων, κλπ. Τέλος, η simple analysis είναι διαθέσιμη, για τη δημιουργία διαφορετικών γραφικών παραστάσεων και πινάκων προκειμένου να αναπαρασταθούν αποτελεσματικά οι επιθυμητές πληροφορίες (Markellos et al., 2001b).

3.4 Επιλογή των δεδομένων

Οι βάσεις δεδομένων που χρησιμοποιούνται είναι η **SPACE ACCESS** και η **SPACE Bulletin**. Τα δεδομένα από αυτές τις βάσεις ανακτήθηκαν με τη χρήση του **MIMOSA** (2005) ενός λογισμικού που χρησιμοποιείται για την αναζήτηση των CD-ROMs του ΕΠΟ καθώς και άλλων οργανισμών διαχείρισης πατεντών (Σχήμα 33). Οι μεταβλητές **MC**, **IC**, **ET**, **AB** που αφορούν στους κύριους και δευτερεύοντες κωδικούς καθώς επίσης και στον αγγλικό τίτλο και στην περίληψη είναι οι μεταβλητές που περιλαμβάνονται στην clustering ανάλυση και στην παραγωγή των δεικτών και συνεπώς θεωρούνται οι βασικές μεταβλητές. Τα άλλα πεδία που περιγράφουν ένα δίπλωμα ευρεσιτεχνίας θεωρούνται συμπληρωματικές μεταβλητές και χρησιμοποιούνται για την περιγραφή των κύριων μεταβλητών. Κάθε ένα από τα παραπάνω πεδία αποθηκεύεται σε διαφορετικό αρχείο με προαποφασισμένη μορφοποίηση. Ο κωδικός προσδιορισμού κάθε πατέντας π.χ. patent 1, patent 2, κλπ. αποθηκεύεται επίσης σε κάθε ένα από τα αρχεία. Αυτό θα επιτρέψει να χρησιμοποιηθούν αποτελεσματικότερα οι πληροφορίες κάθε αρχείου δεδομένου ότι καθιστούν το περιεχόμενο κάθε τους συγκρίσιμο. Επιπλέον, αυτό είναι απαραίτητο για τη διαδικασία clustering δεδομένου ότι επιτρέπει να γνωρίζουμε ποια διπλώματα ευρεσιτεχνίας ανήκουν σε κάθε μια ομάδα και αφετέρου εκμεταλλεύονται τις συμπληρωματικές πληροφορίες για την περιγραφή των ομάδων. Η διαδικασία μεταφοράς και αποθήκευσης των δεδομένων στο σύστημα απεικονίζεται στο

Σχήμα 34. Η συγκεκριμένη δομή δίνει στο χρήστη τη δυνατότητα να έχει μια συνολική άποψη των πατεντών ώστε να εκτελέσει το είδος της ανάλυσης που επιθυμεί (Σχήμα 35).

The screenshot shows the MIMOSA software window. It has a menu bar (File, Edit, View, Options, Tools, Help) and a toolbar with various icons. Below the toolbar is a 'Form mask' section with a table of search criteria. The table has columns: Index, Op, Search terms, Search results, and Con. The search results for 'MOBILE AND TELEPHONY' show '17 doc(s) found'. At the bottom, there are buttons for 'Submit query', 'View mode', 'Clear all queries', and a 'Main result' field showing '59 hit(s) contained in 17 doc(s) found'.

Index	Op	Search terms	Search results	Con.
Q1	N° EP Publication number	=		OR
Q2	N° WO Publication number	=		OR
Q3	N° Application number	=		OR
Q4	Priority number	=		OR
Q5	Publication date	=		OR
Q6	Application date	=		OR
Q7	Priority date	=		OR
Q8	Designated states	=		OR
Q9	Main classification	=		OR
Q10	All classification	=		OR
Q11	Document kind	=		OR
Q12	WO number for EuroPCT	=		OR
Q13	German title	=		OR
Q14	French title	=		OR
Q15	English title	=		OR
Q16	Inventor	=		OR
Q17	Applicant	=		OR
Q18	Representative	=		OR
Q19	English Abstract	=	MOBILE AND TELEPHONY	17 doc(s) found
Q20	French Abstract	=		OR
Q21	Available data	=		

Σχήμα 33. Το εργαλείο ανάκτησης δεδομένων από βάσεις πατεντών MIMOSA.

Σχήμα 34. Η διαδικασία αποθήκευσης των δεδομένων στο σύστημα.

Σχήμα 35. Δεδομένα εισόδου στο σύστημα.

3.5 Γλωσσολογική προ-επεξεργασία των δεδομένων

Διάφορες μέθοδοι γλωσσολογικής (linguistic) ή λεξικομετρικής (lexicometric) προ-επεξεργασίας υπάρχουν οι οποίες βοηθούν στον προσδιορισμό και στην αρίθμηση των συγκεκριμένων στοιχείων ενός κειμένου (Beaugrande & Dressler, 1981), (Rajman & Lebart, 1998). Προκειμένου να επιλεχθούν τα στοιχεία αυτά οι ακόλουθες εργασίες μπορούν να εφαρμοστούν:

- κατάτμηση (segmentation),
- προσδιορισμός (identification),
- λημματοποίηση (lemmatization),
- αποσαφήνιση (disambiguation), κλπ.

Στη προσέγγισή μας εστιάζουμε στη **lemmatization** και στο **part-of-speech tagging**.

Η lemmatization αφορά στον περιορισμό των μορφολογικών παραλλαγών των στοιχείων των κειμένων με τη μείωση των διαφορετικών κλίσεων των λέξεων σε μια μοναδική κανονικοποιημένη μορφή. Σε ορισμένες περιπτώσεις προκειμένου να υπάρξει συστηματικός προσδιορισμός του λήμματος μιας λέξης του κειμένου απαιτείται να έχει εφαρμοστεί προηγουμένως αποσαφήνιση. Η ανάγκη αυτή προκύπτει όταν: μια λέξη αντιστοιχεί σε

κλισμένες μορφές διαφορετικών λημμάτων, όταν υπάρχουν λέξεις από την ίδια ετυμολογική πηγή, κλπ. Σε άλλες περιπτώσεις οι ασάφειες σχετικά με τη συντακτική λειτουργία μιας λέξης πρέπει να αφαιρεθούν, απαιτώντας γραμματική ανάλυση της πρότασης που την περιέχει. Η διαδικασία αυτή βοηθά στον προσδιορισμό των μορφο-συντακτικών κατηγοριών (π.χ. ουσιαστικό, ρήμα, επίθετο, κλπ.) των λέξεων στα κείμενα ο οποίος με τη σειρά του υποστηρίζει το φιλτράρισμα των ασήμαντων λέξεων.

Για να μειωθεί περαιτέρω το μέγεθος του λεξικού περιορίζεται η ανάλυση σε τρεις κατηγορίες λέξεων που φέρουν το μεγαλύτερο μέρος του σημασιολογικού περιεχομένου δηλαδή τα ουσιαστικά, τα ρήματα και τα επίθετα. Έτσι μπορούμε να αφαιρέσουμε τις πολύ συχνές λέξεις όπως *be*, *have*, *can*, κλπ. Σε μια πολύ γενική μορφή τα κύρια βήματα της ληματοποίησης μπορούν να μειωθούν στα εξής:

- τα ρήματα δίνονται στο απαρέμφατο.
- τα ουσιαστικά τίθενται στον ενικό αριθμό.

Αυτή η διαδικασία της γλωσσολογικής προ-επεξεργασίας είναι αυτόματη στο σύστημά μας, αλλά παράλληλα δίνει τη δυνατότητα στο χρήστη να επέμβει με τον καθορισμό διαφορετικών παραμέτρων. Για παράδειγμα, ο χρήστης μπορεί να θέσει ένα κατώτατο όριο στη συχνότητα των χρησιμοποιούμενων λέξεων ή να λάβει υπόψη του τις λέξεις με τη μεγαλύτερη ή τη μικρότερη συχνότητα από ένα προκαθορισμένο όριο. Έτσι μπορεί ο χρήστης να περιλάβει τις συχνές ή τις σπάνιες λέξεις ανάλογα με τις πληροφορίες που θέλει να εξαγάγει από την ανάλυση καθώς και να καθορίσει τα συνώνυμα και τις ισοδύναμες λέξεις.

Η σημασία της γλωσσολογικής προ-επεξεργασίας είναι πολύ μεγάλη δεδομένου ότι αποτελεί απαραίτητο βήμα πριν την εκτέλεση της στατιστικής ανάλυσης ή της *cluster* ανάλυσης. Τα αποτελέσματά της δίνουν χρήσιμες πληροφορίες για τα δεδομένα των κειμένων με τη απεικόνισή τους σε διαφορετικούς πίνακες συχνότητας. Συγκεκριμένα, ο χρήστης μπορεί να επιλέξει να δει τον πίνακα ταξινομημένο με βάση τις συχνότητες των λέξεων, με αλφαβητική σειρά ή και στις δύο μορφές. Στη συνέχεια, περιγράφονται τα βήματα και το προ-αλγοριθμικό σχήμα της γλωσσολογικής προ-επεξεργασίας των δεδομένων.

3.5.1 Τα βήματα

Ο στόχος είναι να μετατρέψουμε τα ακατέργαστα δεδομένα των κειμένων των πατεντών σε μια μορφή κατάλληλη για την εφαρμογή των στατιστικών αναλύσεων π.χ. σε έναν **πίνακα γειτνίασης (contingency table)** δηλαδή έναν αριθμητικό πίνακα που συνδέει δύο μεταβλητές.

- **Είσοδος:** ένα σύνολο από ακατέργαστα δεδομένα κειμένων.
- **Έξοδος:** ένα **λεξικό (lexicon)** και ένας **λεξικολογικός πίνακας (lexical table)** όπου κάθε σειρά αντιστοιχεί στο περιεχόμενο των κειμένων που δίνονται σαν είσοδος. Ο πίνακας αυτός είναι μια ειδική μορφή του contingency πίνακα όπου κάθε κελί περιέχει τη συχνότητα κάθε στοιχείου στο κείμενο. Το κελί (i, j) περιέχει τον αριθμό των εμφανίσεων του j στο περιεχόμενο i .
- **Απαιτήσεις:** Τα δεδομένα εισόδου πρέπει να έχουν καθαριστεί κατάλληλα.

- **Παρατηρήσεις:** είναι προφανές ότι ο πίνακας αυτός είναι αραιός, οπότε η έξοδος θα πρέπει να είναι της μορφής `</patent no = 000000/>`, `</title/>` (1,1) (2,2) (3,1) (4,1) (5,1), `</abstract/>` (2,3) (3,1) (4,2) (5,2), ... δηλαδή κωδικοποίηση του πίνακα με ένα ζευγάρι της μορφής (δείκτης, τιμή).

Ο Πίνακας 12 παρουσιάζει ένα μικρό παράδειγμα δεδομένων εισόδου και εξόδου.

Είσοδος	Έξοδος
<code></patent no = 000000/></code> <code></title/></code> AA ab ac ab, ad ae <code></abstract/></code> Ab ac ae, ad ad. Ae ab ac ac ab <code></patent no = xxxxxxx/></code> <code></title/></code> Ad ae AF AG <code></abstract/></code> AF AG aa ab; ac ab ab af ag	- <i>Λεξικό:</i> aa ab ac ad ae af ag - <i>Λεξικογραφικός πίνακας:</i> <code></patent no = 000000/></code> <code></title/></code> [1 2 1 1 1 0 0] <code></abstract/></code> [0 3 3 2 2 0 0] <code></patent no = xxxxxxx/></code> <code></title/></code> [0 0 0 1 1 1 1] <code></abstract/></code> [1 3 1 0 0 2 2]

Πίνακας 12. Ένα παράδειγμα λεξικού και λεξικογραφικού πίνακα.

Γενικά οι στατιστικές μέθοδοι βασίζονται στις μετρήσεις και στους υπολογισμούς των αντικειμένων που συγκρίνονται. Τα βήματα που ακολουθούμε στη λεξικογραφική ανάλυση είναι τα ακόλουθα (Σχήμα 36):

- **αναγνώριση των λέξεων (words)** στα ακατέργαστα δεδομένα.
- **αφαίρεση των μη σημαντικών λέξεων** π.χ. άρθρα, προθέσεις, κλπ.
- **ενοποίηση των ταυτόσημων λέξεων** που εμφανίζονται με διαφορετική γραφική μορφή.
- **μείωση του μεγέθους του λεξικού.**
- **κωδικοποίηση των δεδομένων κειμένου σε ένα λεξικογραφικό πίνακα.**

Αναγνώριση των λέξεων

Το πρώτο στάδιο είναι να χωρίσουμε τα δεδομένα εισόδου σε στοιχεία που ονομάζονται **tokens** και μπορεί να είναι λέξεις, αριθμοί, σημεία στίξης, κλπ. Αυτή η διαδικασία είναι γνωστή σαν **tokenization**. Ο στόχος είναι να ορίσουμε ποιες είναι οι λέξεις του κειμένου και να φιλτράρουμε τα σημεία στίξης και τα σύμβολα όπως οι παρενθέσεις, τα εισαγωγικά, κλπ. Ο Πίνακας 13 παρουσιάζει ένα παράδειγμα.

the utterance: A device for trimming vegetation, comprising a motor-housing(11a), [...]
μπορεί να χωριστεί ως ακολούθως:
A / device / for / trimming / vegetation / , / comprising / a / motor-housing / (/ 11a /) / ,
ή πιο περιληπτικά:
A / device / for / trimming / vegetation / comprising / a / motor-housing /

Πίνακας 13. Αναγνώριση των λέξεων.

Αν και το βήμα αυτό φαίνεται εύκολο, στην πράξη υπάρχουν πολλά θέματα που πρέπει να προσεχθούν ιδιαίτερα. Για παράδειγμα, οι περισσότερες προτάσεις τελειώνουν με κάποιο σημείο στίξης π.χ. τελεία. Υπάρχουν όμως και κάποιες που τελειώνουν με κάποια συντομογραφία της μορφής «etc.». Σε αυτές τις περιπτώσεις είναι καλή ιδέα να κρατάμε το σημείο στίξης μαζί με τη λέξη ώστε να μπορέσουμε να τη διακρίνουμε από άλλες λέξεις που έχουν την ίδια γραφική μορφή π.χ. η συντομογραφία Wash. για τη λέξη Washington και το ρήμα wash όταν εμφανίζεται στο τέλος μιας πρότασης. Επίσης και η μονή απόστροφος μπορεί να δημιουργήσει προβλήματα αφού στην αγγλική γλώσσα μπορεί να σημαίνει γραμματική έκθλιψη (π.χ. I'll ή isn't), κλισμένη μορφή (π.χ. the house I rented yesterday's garden is really big), δείκτης πληθυντικού αριθμού (π.χ. boys' toys), κλπ.

Σχήμα 36. Λεξικογραφική προ-επεξεργασία των δεδομένων.

Αφαίρεση των μη σημαντικών λέξεων

Η αφαίρεση του μεγαλύτερου μέρους των μη σημαντικών λέξεων που εμφανίζονται στα περιεχόμενα των κειμένων μπορεί να γίνει με την επιλογή των λέξεων που θα κρατήσουμε ή θα απορρίψουμε με βάση τη λεξικογραφική τους ετικέτα. Αυτό το φιλτράρισμα μπορεί να γίνει με τη βοήθεια του **POS-tagger (Part-Of-Speech tagger)**. Συγκεκριμένα, **tagging** είναι η διαδικασία της τοποθέτησης μιας ετικέτας σε κάθε λέξη της πρότασης που δηλώνει σε ποιο μέρος το λόγου (part of speech) ανήκει π.χ. ουσιαστικό, ρήμα, επίρρημα, κλπ. Με αυτό τον τρόπο είναι εύκολη η αφαίρεση των μη σημαντικών λέξεων (Πίνακας 14). Όπως αναφέραμε και προηγουμένως θα αφαιρέσουμε το μεγαλύτερο μέρος αυτών των λέξεων και θα κρατήσουμε ορισμένες για την ανάλυσή μας. Αυτή η διαδικασία είναι ευαίσθητη και δεν μπορεί να γίνει με αυτόματο τρόπο αφού εξαρτάται από το περιεχόμενο που εφαρμόζεται κάθε φορά. Για αυτό ο χρήστης έχει τη δυνατότητα να χρησιμοποιήσει μια απλή τεχνική φιλτραρίσματος τις **stop-lists**. Μια stop-list είναι ένα σύνολο από λέξεις φιλτραρίσματος. Ένα αρχικό σύνολο από τέτοιες λέξεις μπορεί να καταγραφούν σε μια γενική stop-list την οποία ο χρήστης μπορεί να επεξεργαστεί κατάλληλα.

the utterance: A / device / for / trimming / vegetation / such / as grass / and / the / like / comprising / a / motor / [...]

αν αφαιρέσουμε όλα τα επιρρήματα, τα συνδετικά, τα προσδιοριστικά, τις προθέσεις, κλπ. μπορούμε να μειώσουμε την πρόταση ώστε να περικλείει μόνο τις σημαντικές λέξεις:

device / trimming / vegetation / grass / like / motor /

οι λέξεις που αφαιρέσαμε είναι:

A / for / such / as / and / the / comprising

μετά το tagging προκύπτει:

A|D / device|Ncs / for|S / trimming|As / vegetation|Nc

Πίνακας 14. Αφαίρεση των λέξεων χωρίς σημασία.

Ενοποίηση των ταυτόσημων λέξεων

Η αναγνώριση των λέξεων που θεωρούνται ταυτόσημες δεν είναι μια εύκολη εργασία. Η ενοποίηση των διαφορετικών κλίσεων των λέξεων π.χ. sit, sits, sat, κλπ. μπορεί να γίνει με βάση το λήμμα τους π.χ. τα ρήματα μετατρέπονται σε απαρέμφατα, τα ουσιαστικά στην ονομαστική, οι εκθλίψεις αφαιρούνται, κλπ. Αυτές οι τεχνικές επιτρέπουν την ενοποίηση των μορφών των λέξεων που προέρχονται από το ίδιο λήμμα όχι όμως και την αναγνώριση των συνωνύμων. Για τη λύση αυτού του προβλήματος χρησιμοποιούνται λεξικά συνωνύμων τα οποία όμως στην περίπτωση των πατεντών όπου το λεξιλόγιο είναι πολύ συγκεκριμένο και εξαρτάται από το περιεχόμενο πρέπει να αφορούν κάθε πεδίο τεχνολογικής δραστηριότητας. Παράλληλα δίνεται η δυνατότητα στο χρήστη να ορίσει συνώνυμα αν το επιθυμεί στα πλαίσια της ανάλυσής του.

Μείωση του μεγέθους του λεξικού

Αυτό το βήμα έχει σαν στόχο τον περιορισμό του μεγέθους του λεξικού ώστε να είναι δυνατή η ανάλυση σε λογικούς χρόνους. Χρησιμοποιούμε ένα heuristic που λειτουργεί με βάση ένα ελάχιστο και ένα μέγιστο όριο στις συχνότητες εμφάνισης των όρων. Αυτό γίνεται με ένα ημιαυτόματο τρόπο όπου ο χρήστης ορίζει το μέγεθος του λεξικού που επιθυμεί να χρησιμοποιήσει.

Κωδικοποίηση των δεδομένων σε ένα λεξικογραφικό πίνακα

Σε αυτό το βήμα υπολογίζεται για κάθε όρο του λεξικού η συχνότητα εμφάνισής του στα δεδομένα. Το πλεονέκτημα αυτής της κωδικοποίησης σε ένα πίνακα συχνότητων είναι η ευκολία τροποποίησής του π.χ. σε έναν πίνακα παρουσίας/απουσίας και η δυνατότητα συνάθροισης των πατεντών.

3.5.2 Ο αλγόριθμος

Δημιουργία της εσωτερικής βάσης δεδομένων

Βήμα 1: Εισαγωγή των δεδομένων.

Βήμα 2: Καθαρισμός των δεδομένων και αφαίρεση των HTML markup.

Επιλογή των δεδομένων

Βήμα 1: Εξαγωγή από τα καθαρισμένα δεδομένα των ελάχιστων απαιτούμενων μεταβλητών για την cluster ανάλυση:

- Αγγλικός τίτλος (πεδίο ET) και αγγλική περίληψη (πεδίο AB).
- Βασικός IPC κωδικός (πεδίο MC).

Βήμα 2: Εξαγωγή από τα καθαρισμένα δεδομένα των συμπληρωματικών μεταβλητών:

- Χώρα (2 πρώτα ψηφία του πεδίου PR).
- Εφευρέτης (πεδίο IN).
- Επωνυμία εταιρίας κατάθεσης της πατέντας (πεδίο PA).
- Έτος έναρξης ισχύος της πατέντας (4 τελευταία ψηφία του πεδίου PR).
- Δευτερεύοντες IPC κωδικοί (πεδίο IC).
- Άλλα πεδία.

Προ-επεξεργασία των επιλεγμένων δεδομένων

Βήμα 1: Μετατροπή των HTML χαρακτήρων της μορφής «&, ...;» σε εκφράσεις συμβατές με το part-of-speech-tagging (POS-tagger), δηλαδή εκφράσεις που δεν περιλαμβάνουν σημεία στίξης.

Βήμα 2: Εκτέλεση του POS-tagger στο περιεχόμενο των τίτλων και των περιλήψεων των πατεντών.

Βήμα 3: Εξαγωγή του λεξικού λημμάτων από το tagged κείμενο (με συχνότητες) (Σχήμα 37).

Βήμα 4: Επιλογή των γλωσσολογικών όρων που συμμετέχουν στη στατιστική ανάλυση με βάση τους ακόλουθους κανόνες (Σχήμα 38):

- το κατάλληλο υποσύνολο των POS tags (π.χ. ουσιαστικό, επίθετο, ρήμα, κλπ.).
- τον καθορισμό άνω και κάτω ορίων στη συχνότητα.
- τον καθορισμό stop list.

Οπτικοποίηση

Βήμα 1: Πίνακας λέξεων/Πίνακας λημμάτων (ταξινόμηση με βάση τη συχνότητα).

Βήμα 2: Πίνακας λέξεων/Πίνακας λημμάτων (αλφαβητική ταξινόμηση) (Σχήμα 39).

Βήμα 3: Ιστόγραμμα των λέξεων (Σχήμα 40).

Βήμα 4: Περιβάλλον εμφάνισης κάθε συγκεκριμένου όρου.

Σχήμα 37. Γλωσσολογική προ-επεξεργασία.

Σχήμα 38. Επιλογή των λεγμάτων και γραφική τους αναπαράσταση.

Σχήμα 39. Πίνακας γεννιάσης.

Σχήμα 40. Ιστογράμμο των διαστάσεων.

3.6 Correspondence ανάλυση

Η **correspondence analysis (ανάλυση απλών αντιστοιχιών)** είναι μια μέθοδος αναγωγής πολυδιάστατων δεδομένων σε χώρους μικρότερης διάστασης με ελάχιστη απώλεια πληροφορίας (Benzecri, 1992). Με τη μέθοδο αυτή μπορούμε κυρίως να ανακαλύψουμε χαρακτηριστικά γνωρίσματα στα δεδομένα χωρίς να υποθέσουμε οποιοσδήποτε θεμελιώδεις κατανομές τους. Υιοθετήθηκε κυρίως από τους πίνακες γειννίας που περιέχουν κάποιο μέτρο αντιστοιχίας μεταξύ των σειρών και των στηλών προκειμένου να μελετηθούν οι ενδεχόμενες σχέσεις των ενσωματωμένων μεταβλητών. Συγκεκριμένα, αυτή η στατιστική μέθοδος επιτρέπει την ανάλυση και τη γραφική και συνθετική περιγραφή μεγάλων πινάκων γειννίας.

Από γεωμετρική άποψη η correspondence analysis παρέχει μια μέθοδο για την αναπαράσταση των δεδομένων σε ένα Ευκλείδειο χώρο έτσι ώστε να μπορεί να εξεταστεί οπτικά η δομή των αποτελεσμάτων. Για δεδομένα σε έναν τυπικό διπλής κατεύθυνσης πίνακα γειννίας οι μεταβλητές σειρών και οι μεταβλητές στηλών αναπαριστούνται στο ίδιο διάστημα. Αυτό σημαίνει ότι κάποιος μπορεί να εξετάσει τις σχέσεις όχι μόνο μεταξύ των μεταβλητών των σειρών ή των στηλών αλλά και μεταξύ των μεταβλητών και των σειρών και των στηλών. Λαμβάνοντας υπόψη έναν πίνακα συχνότητων, οι σειρές/πατέντες μπορούν να αντιμετωπισθούν ως n σημεία σε έναν χώρο p διαστάσεων. Όταν τα σημεία είναι γειτονικά, αυτό σημαίνει ότι οι δύο πατέντες με τα αντίστοιχα προφίλ σειράς είναι παρόμοιες. Επίσης, για το προφίλ στήλης/λέξης έχουμε μια αναπαράσταση p σημείων σε έναν n διαστάσεων χώρο. Προκειμένου να προσδιοριστούν τα σημεία που είναι κοντά το ένα στο άλλο πρέπει να υπολογίσουμε τις αποστάσεις μεταξύ τους. Επειδή είναι αρκετά δύσκολο να πραγματοποιηθεί οποιαδήποτε απεικόνιση σε περισσότερες από δύο διαστάσεις, είναι πολύ σημαντικό να αποφασιστεί πώς οι διαστάσεις (δηλ. οι παράγοντες) του προκύπτοντος διαστήματος παραγόντων πρέπει να ομαδοποιηθούν.

3.6.1 Ιστορική ανασκόπηση

Η correspondence analysis προήλθε από τη Γαλλία (analyse des correspondences), όπου είναι αρκετά δημοφιλής από τη δεκαετία του '60 από τον J.P. Benzecri και τους συνεργάτες του. Ο Benzecri πρέπει να θεωρηθεί ως ο πρώτος που δημοσιοποίησε τη γραφική αναπαράστασή της. Αυτή η μέθοδος αναπτύχθηκε αρχικά προκειμένου να παραχθεί μια μαθηματική ανάλυση συνόλων δεδομένων γειννίας στη γλωσσολογία. Η αλγεβρική διατύπωση της μεθόδου χρονολογείται από τον Hirschfeld (1935) αργότερα γνωστός και ως H.O. Hartley, στην εργασία του για το «συσχετισμό» μεταξύ των σειρών και των στηλών ενός πίνακα γειννίας. Παράλληλα, ο Horst (1935) πρότεινε παρόμοιες ιδέες στην ψυχομετρική λογοτεχνία, χρησιμοποιώντας τον όρο «μέθοδος των αμοιβαίων μέσων όρων». Ο Fisher (1940) παρήγαγε την ίδια θεωρία υπό τη μορφή διακρίνοντας ανάλυσης σε έναν πίνακα γειννίας. Εν τω μεταξύ, ο Guttman (1941) παρήγαγε ανεξάρτητα μια μέθοδο για την εξαγωγή κλιμάκων σε δεδομένα κατηγοριοποίησης. Ο Guttman ασχολήθηκε με τη γενική περίπτωση περισσότερων από δύο ποιοτικών μεταβλητών, ως εκ τούτου οι ιδέες του μπορούν να αντιμετωπιστούν ως η προέλευση, αυτού που αποκαλούμε, πολλαπλή correspondence analysis.

Δεδομένου ότι η εργασία του Hirschfeld δεν αναφέρθηκε από τον Fisher (1940) και οι Fisher και Guttman παρουσίασαν ουσιαστικά την ίδια θεωρία στα βιομετρικά και ψυχομετρικά περιβάλλοντα αντίστοιχα, ο Fisher έχει θεωρηθεί ως ο εφευρέτης της correspondence analysis για τους ερευνητές της βιομετρικής, ενώ ο Guttman για τους ερευνητές της ψυχομετρικής (Greenacre, 1984). Από τη δεκαετία του '50, αυτή η μέθοδος παρέμεινε ελάχιστα γνωστή έξω από τον ψυχομετρικό κόσμο μέχρι που εμφανίζονται οι εργασίες του Hill (1974), που τη διέδωσαν με το όνομα correspondence analysis. Ο Hayashi (1956) ανέφερε μεθόδους ποσοτικοποίησης, ενώ ο Nishisato (1980) για διπλή βαθμολόγηση. Ο Gifi (1990) ερευνά την ομογενοποιημένη ανάλυση για να προσδιορίσει την πολλαπλή correspondence analysis, ενώ ο Fielding (1992) παρέχει μια χρήσιμη αναθεώρηση των σχέσεων μεταξύ των διαφορετικών μεταβλητών.

3.6.2 Ο αλγόριθμος

Στόχος του προβλήματος της correspondence analysis είναι να βρεθεί μια χαμηλών διαστάσεων προσέγγιση των αρχικών δεδομένων.

Επιλογή των μεταβλητών που περιλαμβάνονται στην ανάλυση

Βήμα 1: Εάν αυτό είναι μέρος της διαδικασίας clustering τότε πήγαινε στο Βήμα 1.1 αλλιώς πήγαινε στο Βήμα 1.2.

Βήμα 1.1: Η μεταβλητή για τη συνάθροιση των πατεντών είναι το πεδίο IPC ή ο αριθμός της πατέντας και η μεταβλητή της περιγραφής της πατέντας είναι το λήμμα. Με άλλα λόγια δημιούργησε τον πίνακα γειννίας: οι σειρές του οποίου αντιστοιχούν στις μορφές των αθροιζόμενων μεταβλητών (δηλαδή τις πατέντες), οι στήλες είναι οι μορφές της περιγραφικής μεταβλητής (εδώ τα λήμματα) και το κελί (i, j) αποθηκεύει τον αριθμό εμφάνισης του λήμματος j στις πατέντες που καταχωρούνται στην πατέντα i .

Βήμα 1.2: (α) Επιλέξτε τη μεταβλητή για τη συνάθροιση των πατεντών.

(β) Επιλέξτε τη μεταβλητή για την περιγραφή των πατεντών.

(γ) Δημιούργησε τον πίνακα γειννίας.

Βήμα 2: Επιλέξτε τις συμπληρωματικές μεταβλητές εάν υπάρχουν.

Τα δεδομένα

Ο πίνακας δεδομένων K είναι ένας πίνακας γειννίας με n σειρές και p στήλες που

$$\text{δηλώνεται ως } K_{n \times p} \text{ και έχει τη μορφή } K_{n \times p} = \begin{matrix} & k_{11} & \dots & k_{1j} & \dots & k_{1p} \\ & \vdots & & \vdots & & \vdots \\ & k_{i1} & \dots & k_{ij} & \dots & k_{ip} \\ & \vdots & & \vdots & & \vdots \\ & k_{n1} & \dots & k_{nj} & \dots & k_{np} \end{matrix}$$

Θεωρήστε ότι $k_{i,j}$ είναι ο γενικός όρος της διατομής της i^{th} σειράς και της j^{th} στήλης, $k_{i.}$ είναι το άθροισμα των στοιχείων της σειράς i και $k_{.j}$ είναι το άθροισμα των στοιχείων της στήλης j . Με άλλα λόγια: $k_{i.} = \sum_j k_{ij}$, $k_{.j} = \sum_i k_{ij}$ και $k = \sum_{i,j} k_{ij}$.

Βήμα 1: Υπολογίστε τον πίνακα των σχετικών συχνοτήτων $F_{n,p}$ όπου:

$$F_{n \times p} = \begin{matrix} & \begin{matrix} f_{11} & \cdots & f_{1j} & \cdots & f_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \end{matrix} \\ \begin{matrix} f_{i1} & \cdots & f_{ij} & \cdots & f_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \end{matrix} & \\ \begin{matrix} f_{n1} & \cdots & f_{nj} & \cdots & f_{np} \end{matrix} & \end{matrix}$$

$$\text{με } f_{ij} = \frac{k_{ij}}{k}, f_{i.} = \sum_j f_{ij}, f_{.j} = \sum_i f_{ij} \text{ και } \sum_{i,j} f_{ij} = 1.$$

Μετασχηματισμός του πίνακα γεινίασης

Βήμα 1: Υπολογίστε τους διαγώνιους πίνακες D_n και D_p όπου:

$$D_n = \begin{matrix} \begin{matrix} f_{1.} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \end{matrix} & \begin{matrix} f_{.1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \end{matrix} \end{matrix} \text{ και } D_p = \begin{matrix} \begin{matrix} f_{.1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \end{matrix} & \begin{matrix} f_{1.} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \end{matrix} \end{matrix}$$

Βήμα 2: Υπολογίστε τους αντίστροφους πίνακες: D_n^{-1} και D_p^{-1} .

Βήμα 3: Υπολογίστε τους πίνακες των προφίλ των σειρών και των προφίλ των στηλών που δίνονται από τους τύπους:

$$\text{Matrix of Row Profile} = D_n^{-1} * F_{n \times p}$$

και

$$\text{Matrix of Column Profile} = D_p^{-1} * F_{n \times p}.$$

Τα προφίλ, που είναι σειρές n ή p αριθμών ανάλογα με το εάν αναφερόμαστε στις σειρές ή τις στήλες αντίστοιχα, χρησιμοποιούνται για να καθορίσουν σημεία σε χώρους p ή n διαστάσεων.

Υπολογισμός των αποστάσεων μεταξύ των στοιχείων του πίνακα

Η απόσταση που δίνεται παρακάτω επιτρέπει να ορίσουμε μία μετρική για «την εύρεση» του πόσο κοντά βρίσκονται τα διαφορετικά σημεία.

Βήμα 1: Υπολογίστε την χ^2 απόσταση μεταξύ δύο σημείων-σειράς i και i' :

$$M = d^2(i, i') = \sum_{j=1}^p \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Βήμα 2: Υπολογίστε την χ^2 απόσταση μεταξύ δύο σημείων-στήλης j και j' :

$$M = d^2(j, j') = \sum_{i=1}^n \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

Ταυτόχρονη αναπαράσταση των σημείων στους παραγοντικούς άξονες

Στόχος είναι η γραφική αναπαράσταση της **εγγύτητας (proximity)** μεταξύ των προφίλ σειρών και στηλών. Δεν είναι δυνατό να ερμηνευθούν τα διαγώνια-proximities μεταξύ ενός σημείου-σειράς και ενός σημείου-στήλης, δεδομένου ότι αυτά τα δύο σημεία δεν προέρχονται από το ίδιο αρχικό διάστημα. Εντούτοις, είναι δυνατό να ερμηνευθεί η θέση ενός μοναδικού σημείου-σειράς σε σχέση με το σύνολο των σημείων-στήλης ή ενός μοναδικού σημείου-στήλης σε σχέση με το σύνολο των σημείων-σειράς. Η ανάλυση προχωρά είτε σχετικά με το κέντρο προέλευσης είτε με το κέντρο βάρους. Στην πρώτη περίπτωση ο παραγοντικός άξονας συνδέει την προέλευση με το κέντρο βάρους.

Αναπαράσταση σε R^p

Σε αυτό το διάστημα υπάρχουν n σημεία που αντιπροσωπεύονται από τις n σειρές $D_n^{-1} * F_{n \times p}$. Έστω ότι u είναι ένα μοναδιαίο διάνυσμα για την απόσταση στο R_p έτσι ώστε: $u'D_p^{-1}u = 1$.

Βήμα 1: Διαγωνοποίησε τον πίνακα $S = F'D_n^{-1}FD_p^{-1}$ με το γενικό όρο

$$s_{ij} = \sum_{i=1}^n \frac{f_{ij}f_{ij'}}{f_{i.}f_{.j}}$$

προκειμένου να ληφθούν οι $p-1$ χαρακτηριστικές ρίζες και τα αντίστοιχα χαρακτηριστικά διανύσματα και συνεπώς οι $p-1$ παραγοντικοί άξονες.

Βήμα 2: Πάρε τις χαρακτηριστικές ρίζες λ_α που λαμβάνονται από τη λύση του συστήματος: $Su_\alpha = \lambda_\alpha u_\alpha$.

Βήμα 3: Αγνόησε την πρώτη χαρακτηριστική ρίζα που είναι ίση με 1 και το αντίστοιχο χαρακτηριστικό διάνυσμα.

Βήμα 4: Οι παραγοντικές συντεταγμένες λαμβάνονται από την εξίσωση:

$$\psi_a = D_n^{-1} F D_p^{-1} u_{\alpha} \text{ ή } \psi_{ai} = \sum_{j=1}^p \frac{f_{ij}}{f_{i.} f_{.j}} u_{aj} \text{ ή } \psi_{ai} = \frac{1}{\sqrt{\lambda_{\alpha}}} \sum_{j=1}^p \frac{f_{ij}}{f_{i.}} \varphi_{aj}.$$

Αναπαράσταση στο R^n

Υπάρχουν σημεία των οποίων οι p συντεταγμένες αντιστοιχούν στις p στήλες του $F_{n \times p} * D_p^{-1}$. Έστω ότι v είναι ένα μοναδιαίο διάνυσμα για την απόσταση R^n έτσι ώστε: $v' D_n^{-1} v = 1$.

Βήμα 5: Διαγωνοποίησε τον πίνακα $T = F D_p^{-1} F' D_n^{-1}$.

Βήμα 6: Πάρε τις χαρακτηριστικές ρίζες που λαμβάνονται από τη λύση του συστήματος: $T v_{\alpha} = \lambda_{\alpha} v_{\alpha}$. Είναι γνωστό ότι οι μη μηδενικές χαρακτηριστικές ρίζες είναι ίδιες και για τα δύο διαστήματα (σειρές και στήλες).

Βήμα 7: Αγνόησε την πρώτη χαρακτηριστική ρίζα που είναι ίση με 1 και το αντίστοιχο χαρακτηριστικό διάνυσμα.

Βήμα 8: Οι παραγοντικές συντεταγμένες λαμβάνονται από την εξίσωση:

$$\varphi_a = D_p^{-1} F D_n^{-1} v_{\alpha} \text{ ή } \varphi_{ai} = \sum_{j=1}^n \frac{f_{ij}}{f_{i.} f_{.j}} v_{aj} \text{ ή } \varphi_{ai} = \frac{1}{\sqrt{\lambda_{\alpha}}} \sum_{j=1}^n \frac{f_{ij}}{f_{.j}} \psi_{aj}.$$

Βήμα 9: Παραγάγετε μια γραφική παράσταση λαμβάνοντας υπόψη τις συντεταγμένες που δίνονται στα βήματα 3 και 6.

Κανόνες για την ερμηνεία των αποτελεσμάτων

Βήμα 1: Μετρήστε τη διακύμανση σε κάθε κύριο άξονα μέσω των χαρακτηριστικών ριζών λ_{α} .

Βήμα 2: Υπολογίστε το ίχνος t , που είναι το σύνολο όλων των χαρακτηριστικών ριζών και αντιπροσωπεύει τη συνολική διακύμανση της διαμόρφωσης των σημείων.

Βήμα 3: Υπολογίστε τη συμβολή του σημείου-σειρά i στη δημιουργία του άξονα a

από τον τύπο $Cr_a(i) = \frac{f_{i.} \psi_{ai}^2}{\lambda_a}.$

Βήμα 4: Υπολογίστε τη συμβολή του σημείου-στήλη j στη δημιουργία του άξονα a

από τον τύπο $Cr_a(j) = \frac{f_{.j} \varphi_{ja}^2}{\lambda_a}.$

Οι contributions είναι μετρήσεις της συμβολής κάθε στοιχείου στη συνολική διακύμανση κάθε άξονα. Με άλλα λόγια αυτή η μέτρηση επιτρέπει να γνωρίζουμε την αναλογία στην οποία ένα σημείο i συμβάλλει στη δημιουργία του άξονα a .

Βήμα 5: Υπολογίστε την τετραγωνική συσχέτιση για το i^{th} σημείο του R^p :

$$Cos_a^2(i) = \frac{d_a^2(i, G)}{d^2(i, G)} = \frac{\varphi_{ai}^2}{d^2(i, G)} \quad \text{όπου} \quad d^2(i, G) = \sum_{j=1}^p \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{.i}} - f_{.j} \right)^2 \quad \text{και} \quad G \quad \text{είναι}$$

το κέντρο βάρους.

Ο υπολογισμός των τετραγωνικών συσχετίσεων μας δίνει πληροφορίες για το εάν ένα σημείο αντιπροσωπεύεται καλά σε συγκεκριμένο παραγοντικό υποδιάστημα.

Κανόνες για την επιλογή των παραγοντικών αξόνων

Αυτές οι μετρήσεις χρησιμοποιούνται στη συνέχεια για την απόφαση του ποιοι είναι οι σημαντικοί παράγοντες που θα περιληφθούν στην ανάλυση. Η διαδικασία που περιγράφεται παρακάτω επιτρέπει να καθοριστούν οι παραγοντικοί άξονες που θα χρησιμοποιηθούν και να ληφθεί για κάθε πατέντα ένα σύνολο στατιστικών όπως οι συντεταγμένες τους, οι συνεισφορές και οι τετραγωνικές συσχετίσεις τους. Επιπλέον, για τους επιλεγμένους άξονες, λαμβάνουμε τις συντεταγμένες, τις συνεισφορές και τις τετραγωνικές συσχετίσεις των λημμάτων.

Βήμα 1: Για όλα τα στοιχεία στις σειρές (δηλαδή στις πατέντες) υπολογίστε τη συνολική συνεισφορά δηλαδή το άθροισμα του γινομένου της σημαντικότητας των παραγοντικών αξόνων με την απόλυτη συμβολή της πατέντας.

Βήμα 2: Βάλτε κατά φθίνουσα σειρά τη συνολική συνεισφορά.

Βήμα 3: Για κάθε μία από τις πατέντες εξετάστε τις τιμές των τετραγωνικών συσχετίσεων που συνδέονται με τους παραγοντικούς άξονες και επιλέξτε αυτές που συνδέονται με τιμές τετραγωνικών συσχετίσεων που είναι επάνω από ένα κατώτατο όριο.

Βήμα 4: Διατάξτε τον κατάλογο των λημμάτων σύμφωνα με το άθροισμα των τιμών των τετραγωνικών συσχετίσεων με τους άξονες που επιλέχθηκαν στο προηγούμενο βήμα.

Βήμα 5: Επιλέξτε είτε τα N πρώτα λήμματα (το N καθορίζεται από το χρήστη) είτε σύμφωνα με έναν αυτοματοποιημένο αλγόριθμο που βρίσκει την πρώτη σημαντική ασυνέχεια.

Ένας άλλος εμπειρικός κανόνας καθορίζεται παρακάτω:

Βήμα 1: Διαίρεστε τη συνεισφορά κάθε στοιχείου με το συνολικό αριθμό πατεντών.

Βήμα 2: Εάν η τιμή της συνεισφοράς είναι μεγαλύτερη από την ποσότητα που καθορίζεται στο βήμα 1 τότε χρησιμοποιήστε την ως ενεργό μεταβλητή, διαφορετικά χρησιμοποιήστε την ως συμπληρωματική μεταβλητή.

Ενεργές και συμπληρωματικές μεταβλητές

Η correspondence analysis χρησιμοποιείται για την εύρεση υποδιαστημάτων που αναπαριστούν την εγγύτητα μεταξύ των σειρών και των στηλών του πίνακα γειννίας. Μπορεί επίσης να χρησιμοποιηθεί για τον προσδιορισμό των συμπληρωματικών σειρών και στηλών του πίνακα δεδομένων σε αυτό το υποδιάστημα. Τα συμπληρωματικά στοιχεία δεν παρεμβαίνουν στην κατασκευή των αξόνων. Επομένως, αυτά τα στοιχεία έχουν μηδενική συνεισφορά. Σε σύγκριση με τα ενεργά στοιχεία, το άθροισμα των τετραγωνικών συσχετίσεων ενός στοιχείου σε σχέση με το σύνολο των αξόνων μπορεί να είναι λιγότερο από 1, ενώ για τα ενεργά στοιχεία είναι ακριβώς 1. Οι συντεταγμένες αυτών των στοιχείων υπολογίζονται από τους ίδιους τύπους που χρησιμοποιούνται για τα ενεργά στοιχεία.

3.7 Cluster ανάλυση

Η **cluster analysis (ανάλυση συστάδων)** είναι μια πολλών μεταβλητών διαδικασία για την εύρεση φυσικών ομάδων σε δεδομένα. Η ανάλυση αυτή είναι βασισμένη στην τοποθέτηση των αντικειμένων σε λιγότερο ή περισσότερο ομοιογενείς ομάδες, με έναν τρόπο έτσι ώστε η σχέση μεταξύ των ομάδων να αποκαλύπτεται (Alderfer & Blashfield, 1986). Η ιδιαιτερότητα αυτής της διαδικασίας ομαδοποίησης συνίσταται στην εφαρμογή μιας ιεραρχικής ανάλυσης συστάδων στο σύνολο των στοιχείων, τα οποία εντούτοις χαρακτηρίζονται από τις παραγοντικές συντεταγμένες που δημιουργούνται από την παραγοντική ανάλυση που έχει εκτελεστεί προηγουμένως (Lewis, 1991).

3.7.1 Ο αλγόριθμος

Πριν παρουσιάσουμε τον αλγόριθμο συνοψίζουμε την κύρια ιδέα της διαδικασίας ομαδοποίησης, δίνοντας μια σύντομη περιγραφή των βασικών βημάτων (Σχήμα 41):

Βήμα 1: Σε περίπτωση που έχουμε περισσότερα από N^3 στοιχεία πήγαινε στο βήμα 2.1. Διαφορετικά πήγαινε στο βήμα 2.2.

Βήμα 2.1: Χωρίστε τα αρχικά αντικείμενα σε έναν προκαθορισμένο αριθμό k συστάδων χρησιμοποιώντας την K-means clustering δηλαδή ορίζοντας κάθε στοιχείο στη συστάδα που έχει το κοντινότερο centroid. Κατόπιν εκτελέστε ιεραρχικό clustering στις k συστάδες.

Βήμα 2.2: Εκτελέστε ιεραρχικό clustering στα αρχικά αντικείμενα.

Βήμα 3: Χρησιμοποιήστε μια αυτόματη διαδικασία για την επιλογή του βέλτιστου κοψίματος του ιεραρχικού δέντρου για τον καθορισμό του αριθμού των συστάδων.

Βήμα 4: Περιγράψτε τις συστάδες χρησιμοποιώντας διαφορετικά στατιστικά μέτρα όπως ο έλεγχος τιμών (valuer-test), τα χαρακτηριστικά στοιχεία κλπ.

³ Στη φάση της υλοποίησης του αλγορίθμου, το N υπολογίστηκε εμπειρικά ότι αντιστοιχεί σε περίπου 500 στοιχεία.

Βήμα 5: Βρείτε τις σχέσεις μεταξύ των συστάδων και αναπαράστήστε τα γραφικά.

Σχήμα 41. Διαδικασία της ανάλυσης συστάδων.

Τα δεδομένα

Έχουμε n στοιχεία για ομαδοποίηση που αναπαριστάνονται ως σημεία σε ένα Ευκλείδειο διάστημα με p διαστάσεις. Κάθε στοιχείο αναπαριστάται από ένα σύνολο συντεταγμένων που αντιστοιχούν στους χρησιμοποιούμενους παραγοντικούς άξονες. Κατά συνέπεια κάθε σημείο x είναι ένα διάνυσμα με p τμήματα και είναι ισοδύναμο να εκτελέσει μια ομαδοποίηση των αντικειμένων είτε για τις p μεταβλητές είτε για τους p παραγοντικούς άξονες.

	f.1	f.j	f.p
pat.1	x_{11}	x_{1j}	x_{1p}
pat.i	x_{i1}	x_{ij}	x_{ip}
pat.n	x_{n1}	x_{nj}	x_{np}

Τα δεδομένα συνοψίζονται στον ακόλουθο πίνακα: $X =$

όπου κάθε σειρά αντιπροσωπεύει τις συντεταγμένες του στοιχείου κατά μήκος των παραγοντικών αξόνων p . Με άλλα λόγια κάθε σειρά αντιπροσωπεύει ένα στοιχείο που περιλαμβάνεται στην ανάλυση ενώ οι στήλες αντιπροσωπεύουν τους παραγοντικούς άξονες.

Βήμα 1: Ελέγξτε τον αριθμό των στοιχείων που πρέπει να ομαδοποιηθούν. Εάν είναι μεγαλύτερος από N στοιχεία (δηλαδή πατέντες) τότε πηγαίνετε στο μέρος Β, διαφορετικά πηγαίνετε στο μέρος Α.

Μέρος Α: Ιεραρχικό clustering

Βήμα 1: Ξεκινήστε με την ανάθεση κάθε σημείου σε κάθε συστάδα έτσι ώστε εάν έχετε τα n σημεία τώρα να έχετε n συστάδες.

Βήμα 2: Υπολογίστε τις αποστάσεις μεταξύ των σημείων. Για να υπολογίσετε τις $\frac{n(n-1)}{2}$ αποστάσεις μεταξύ των n σημείων: υπολογίστε το κέντρο βάρους μ^j και τη σταθερή απόκλιση σ^j κάθε παράγοντα:

$$\mu^j = \frac{1}{n} \sum_{i=1}^n x_{ij} \text{ και } \sigma^j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \mu^j)^2}{n}}.$$

Αντικαταστήστε τον πρώτο πίνακα T από ένα άλλο T' όπου κάθε όρος t_{ij} αντικαθίσταται από έναν όρο $t'_{ij} : x'_{ij} = \frac{x_{ij} - \mu^j}{\sigma^j}$.

Βήμα 3: Βρείτε το πιο κοινό (ή το περισσότερο παρόμοιο) ζευγάρι των σημείων και συγχωνεύστε τα σε ένα μοναδικό cluster στο οποίο έχουμε τώρα $(n-1)$ σημεία.

Βήμα 4: Επαναλάβετε το βήμα 3 έως ότου ανασυγκροτήσουμε όλα τα σημεία σε ένα μοναδικό partition.

Υπολογισμός της απόστασης μεταξύ των συστάδων

Εάν το A είναι μια συστάδα αποτελούμενη από n_a στοιχεία και B είναι μια συστάδα αποτελούμενη από n_b στοιχεία, τότε:

Βήμα 1: Το βάρος w_a της συστάδας A είναι: $w_a = \frac{n_a}{n}$.

Βήμα 2: Εάν g_a και g_b είναι τα κέντρα βάρους του A και του B αντίστοιχα:

$$\begin{aligned} g_a &= (g_a^1, g_a^2, \dots, g_a^j, \dots, g_a^p) \\ g_b &= (g_b^1, g_b^2, \dots, g_b^j, \dots, g_b^p) \end{aligned} \quad \text{όπου:} \quad \begin{aligned} g_a^j &= \frac{1}{w_a} \sum_{i \in A} x_{ij}' = \frac{n}{n_a} \sum_{i \in A} x_{ij}' \\ g_b^j &= \frac{1}{w_b} \sum_{i \in B} x_{ij}' = \frac{n}{n_b} \sum_{i \in B} x_{ij}' \end{aligned}$$

τότε το κέντρο βάρους g_{ab} των δύο ομάδων A και B είναι:

$$g_{ab} = \frac{w_a g_a + w_b g_b}{w_a + w_b} = \frac{n_a g_a + n_b g_b}{n_a + n_b}.$$

Η απόσταση μεταξύ των δύο συστάδων A και B είναι τότε:

$$\Delta(A, B) = \frac{w_a w_b}{w_a + w_b} d^2(g_a, g_b) = \frac{n_a n_b}{n(n_a + n_b)} d^2(g_a, g_b)$$

$$\text{και } d^2(g_a, g_b) = \sum_{j=1}^p (g_a^j - g_b^j)^2$$

Βήμα 4: Εάν A , B , και C είναι τρεις συστάδες με n_a , n_b και n_c στοιχεία αντίστοιχα. Εάν το A και το B συγχωνεύονται σε μια συστάδα Θ , δηλαδή: $\Theta = \{A, B\}$, τότε η απόσταση μεταξύ του C και Θ είναι:

$$\Delta(\Theta, C) = \frac{(w_a + w_c)\Delta(A, C) + (w_b + w_c)\Delta(B, C) - w_c\Delta(A, B)}{w_a + w_b + w_c}.$$

Μέρος Β: k-means clustering

Βήμα 1: Καθορίστε τον αριθμό των συστάδων K .

Βήμα 2: Επιλέξτε τυχαία, K αυθαίρετα σημεία X_i . Χρησιμοποιήστε το X_i ως το αρχικό centroid A_j .

- X_i όπου $i = 1, \dots, n$ είναι οι παρατηρήσεις σε μια συλλογή των μεταβλητών. Ο πίνακας των δεδομένων, έχει n σειρές ως παρατηρήσεις και p στήλες για τις μεταβλητές.
- A_j όπου $j = 1, \dots, k$ είναι οι ομάδες των centroids, για τις K συστάδες. Ο A_j είναι ένα m-διαστάσεων διάνυσμα $(a_{j1}, a_{j2}, \dots, a_{jl}, \dots, a_{jp})$ όπου κάθε

στοιχείο a_{jl} , είναι ο μέσος όρος της l^{th} μεταβλητής για τα X_i στοιχεία που ανήκουν στο j^{th} cluster.

Βήμα 3: Υπολογίστε το W (κριτήριο σύγκλισης) από τον τύπο:

$$W = \sum_{j=1}^k \sum_{x \in P_j} D(A_j, X).$$

D είναι οι Ευκλείδειες αποστάσεις που ορίζονται ως: η άμεση

απόσταση μεταξύ δύο αυθαίρετων σημείων X_1 και X_2 με συντεταγμένες $X_1 = (x_{11}, \dots, x_{1j}, \dots, x_{1p})$ και $X_2 = (x_{21}, \dots, x_{2j}, \dots, x_{2p})$ δίνεται από τον τύπο:

$$D(X_1, X_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \dots + (x_{1p} - x_{2p})^2}.$$

Σε μια στατιστική κατάσταση οι συντεταγμένες (x_{1i}, x_{2i}) είναι οι τιμές των στοιχείων X_1 και X_2 στους m παράγοντες.

Βήμα 4: Για κάθε X_i και P_j (όπου P_j με $j = 1, \dots, k$ είναι η μορφή των στοιχείων στα k clusters και κάθε P_j έχει ένα A_j) υπολογίστε το $w = \sum_{j=1}^k \sum_{x \in P_j} D(A_j, X)$.

Βήμα 5: Για κάθε X_i με $i = 1, \dots, n$ εάν υπάρχει τουλάχιστον ένα $D(A_j, X_i)$ με $X_i \notin P_j$ ώστε $D(A_j, X_i) < D(A_j, X_i)$ για $X_i \in P_j$ ανάθεσε το X_i στο P_j που αντιστοιχεί στο ελάχιστο $D(A_j, X_i)$ για $X_i \notin P_j$ αλλιώς άφησε $X_i \in P_j$.

Βήμα 6: Υπολογίστε τα k centroids A_j των k συστάδων, χρησιμοποιώντας το νέο χώρισμα δεδομένων, ως αποτέλεσμα του βήματος 5.

Βήμα 7: Υπολογίστε το W_{n+1} και: $\frac{|W_{n+1} - W_n|}{W_{n+1}}$ εάν είναι μικρότερο από ε τότε σταματήστε, αλλιώς πηγαίνετε στο βήμα 6. Όπου $\varepsilon > 0$ είναι η τιμή που καθορίζει πότε θα σταματήσει ο αλγόριθμος στην επανάληψη $n+1$. Σε κάθε βήμα υπολογίζουμε τη συνάρτηση $\frac{|W_{n+1} - W_n|}{W_{n+1}}$ μέχρι να γίνει $< \varepsilon$.

Βήμα 8: Παρακάτω, πρόκειται να χρησιμοποιήσουμε το τελικό centroid k των k συστάδων ώστε να εκτελέσουμε το ιεραρχικό clustering. Ως εκ τούτου, είναι σημαντικό να γίνουν τα εξής:

- Πάρτε τα τελικά centroids k από το βήμα 6.
- Τώρα, το σύνολο δεδομένων μπορεί να συνοψισθεί σε έναν πίνακα της μορφής:

$$A_j = T = \begin{matrix} pat.1 \\ pat.2 \\ \vdots \\ pat.n \end{matrix} \begin{bmatrix} \bar{t}_{11} & \bar{t}_{12} \dots & \bar{t}_{1l} \dots & \bar{t}_{1p} \\ \bar{t}_{21} & \bar{t}_{22} \dots & \bar{t}_{2l} \dots & \bar{t}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{t}_{k1} & \bar{t}_{k2} \dots & \bar{t}_{kl} & \bar{t}_{kp} \end{bmatrix}$$

- Κάθε σειρά, T είναι ένα m διαστάσεων διάνυσμα $(t_{j1}, t_{j2}, \dots, t_{jl}, \dots, t_{jp})$ όπου κάθε στοιχείο t_{jl} είναι ο μέσος όρος της i^{th} μεταβλητής για τα X_i στοιχεία που ανήκουν στη j^{th} συστάδα.

Βήμα 9: Πηγαίνετε στο μέρος A, όπου τα αρχικά στοιχεία είναι οι K συστάδες.

Βέλτιστη επιλογή συστάδων

Για να λάβει ο χρήστης μια απόφαση για τον αριθμό των πιθανών συστάδων έχει δύο διαφορετικές επιλογές:

- **Αυτόματη απόφαση:** σε αυτή την περίπτωση ο αλγόριθμος δίνει αυτόματα τον αριθμό των clusters.
- **Χειρωνακτική απόφαση:** σε αυτή την περίπτωση ο χρήστης πρέπει να αποφασίσει τον αριθμό των clusters.

Αυτόματη απόφαση

Έστω Cl είναι ο αριθμός των συστάδων, F_I : το πρώτο στοιχείο (ή κόμβος), S_I : το δεύτερο στοιχείο (κόμβος), N_I : ο αριθμός των στοιχείων στη συστάδα, I_L : το επίπεδο δεικτών (που δείχνει την απόσταση μεταξύ δύο στοιχείων), $I\_Lhistogram$: το ιστόγραμμα επιπέδων δεικτών. Υπενθυμίζουμε ότι n είναι ο αριθμός των στοιχείων που πρέπει να ταξινομηθούν και $I\_L(i)$ είναι το επίπεδο δεικτών των κόμβων $(2n - i)$.

Βήμα 1: Για $(i = 2 \text{ ως } (n - 2))$ υπολογίστε:

$$\tau_i = I\_L(i + 1) - 2 \times I\_L(i) + I\_L(i - 1).$$

Βήμα 2: Βρείτε i_o τέτοιο ώστε: $\tau_{i_o} = \text{Max}(\tau_i)$ για $i = 2, 3, \dots, (n - 2)$. Ο αριθμός συστάδων είναι: i_o και το

δενδρόγραμμα (Σχήμα 42) πρέπει να κοπεί μεταξύ του κόμβου $(2n - i_o)$ και του κόμβου $(2n - (i_o - 1))$.

Χειρωνακτική απόφαση

Αν και ο χρήστης μπορεί να δώσει τον αριθμό των συστάδων που θεωρεί ιδανικό, υπάρχει ένας πρακτικός κανόνας που τον βοηθά να αποφασίσει για τον αριθμό των clusters. Αυτός περιγράφεται παρακάτω:

Βήμα 1: Παράγετε το δενδρόγραμμα.

Βήμα 2: Αριθμήστε προς τα επάνω από το κατώτατο σημείο έως ότου ένα σημαντικό σπάσιμο στην κλίση (αριθμοί) εμφανιστεί.

Περιγραφή των συστάδων

Μέτρο της σπουδαιότητας κάθε λέξης που ανήκει σε μια συστάδα

Προκειμένου να αξιολογηθεί η σημασία κάθε λέξης στη δημιουργία μιας συστάδας πρέπει να καθορίσουμε ένα συγκεκριμένο στατιστικό μέτρο που να μας επιτρέπει να συλλαμβάνουμε τη σημασία κάθε λέξης ανά συστάδα. Στη βιβλιογραφία αυτό επιτυγχάνεται μέσω των συγκρίσεων των μέσων ή των ποσοστών στο εσωτερικό των συστάδων με τα αντίστοιχα μέσα ή ποσοστά για το σύνολο των στοιχείων που πρέπει να ταξινομηθούν. Στην περίπτωση μας ενδιαφερόμαστε για την αξιολόγηση της σημασίας κάθε λέξης για τη δημιουργία των συστάδων στις οποίες ανήκει. Επομένως, προκειμένου να χαρακτηριστεί μια συστάδα από τις λέξεις που την περιγράφουν, συγκρίνουμε τις μέσες τιμές μεταξύ των λέξεων της συστάδας και των συνολικών μέσων τιμών. Πιο συγκεκριμένα αυτή η διαδικασία περιγράφεται στα παρακάτω βήματα. Στην πραγματικότητα, μια μεταβλητή θεωρείται σημαντική για μια συστάδα εάν η παρουσία της στη συστάδα θεωρείται σημαντικότερη από την παρουσία της στο συνολικό πληθυσμό. Δύο στατιστικά μέτρα θα παρουσιαστούν στη συνέχεια. Το πρώτο είναι το **valuer-test** που υπολογίζει την απόσταση μεταξύ του μέσου όρου της συστάδας και του γενικού μέσου όρου στο συνολικό πληθυσμό. Το δεύτερο αναφέρεται στην πιθανότητα ότι η μεταβλητή δεν είναι μεγαλύτερη από την απόλυτη τιμή της παρατηρηθείσας διαφοράς.

Βήμα 1: Για τη συστάδα K , υπολογίστε για τη λέξη i της συστάδας την ακόλουθη ποσότητα: $t^i_K(x) = \frac{\bar{X}^i_K - \bar{X}}{s^{i^2}_K(X)}$ με $s^2_K(X) = \frac{n - n^i_K}{n - 1} * \frac{S^2(X)}{n^i_K}$ όπου $\bar{X}^i_K$ είναι

κάθε φορά ο μέσος όρος της λέξης υπό έρευνα για τη συγκεκριμένη συστάδα, $\bar{X}$ είναι ο γενικός μέσος όρος, $s^{i^2}_K(X)$ είναι η διαφορά της μεταβλητής $\bar{X}^i_K$ και $S^2(X)$ είναι η εμπειρική διαφορά του μεταβλητού X .

Βήμα 2: Επαναλάβετε τον υπολογισμό της ανωτέρω ποσότητας για κάθε λέξη της συστάδας.

Βήμα 3: Ταξινομήστε τις λέξεις της συστάδας με κατιούσα σειρά σύμφωνα με τις τιμές της ποσότητας $t^i_K(x)$.

Βήμα 4: Οι λέξεις για τις οποίες η ποσότητα $t^i_K(x)$ είναι μεγαλύτερη από 2, είναι σημαντικές για τη συστάδα.

Βήμα 5: Επαναλάβετε αυτή τη διαδικασία για κάθε συστάδα.

Ένα πρόσθετο κριτήριο για τον προσδιορισμό των χαρακτηριστικών μεταβλητών μιας συστάδας είναι η πιθανότητα για τη δοκιμή της άκυρης υπόθεσης ότι αυτές οι τιμές μιας μεταβλητής που χρησιμοποιείται στη συστάδα λαμβάνονται τυχαία. Αυτό το κριτήριο παρέχει παρόμοιες πληροφορίες όπως και το valuer-test, με τη διαφορά ότι όσο πιο μικρές είναι οι

τιμές της πιθανότητας τόσο περισσότερο ενισχύεται η υπόθεση ότι αυτές οι τιμές έχουν επιλεχτεί τυχαία.

Έστω ότι n_{kj} είναι ο αριθμός των πατεντών που χαρακτηρίζονται από μία μεταβλητή (λέξη) j μεταξύ των n_k πατεντών της συστάδας K . Ενώ, n_j είναι ο αριθμός των πατεντών που χαρακτηρίζονται από το μεταβλητή j και n ο συνολικός αριθμός των πατεντών. Έστω N είναι ο αριθμός των πατεντών.

Βήμα 1: Υπολογίστε την πληθώρα της μεταβλητής j συγκρίνοντας το ποσοστό της στη συστάδα K που ορίζεται από τον τύπο: $\frac{n_{kj}}{n_k}$.

Βήμα 2: Λαμβάνοντας υπόψη την υπόθεση της ανεξαρτησίας, ο αριθμός N των πατεντών της συστάδας K με τη μεταβλητή j ακολουθεί το υπεργεωμετρικό όριο με παραμέτρους n_{kj} , n_j και n_k . Υπολογίστε την πιθανότητα να βρεθεί μια τιμή N μεγαλύτερη από n_{kj} , δηλαδή την πιθανότητα $p_k(j) = \text{Pr ob}(N \geq n_{kj})$. Εάν ονομάσουμε C_a^b τον αριθμό των διακριτών χωρισμάτων των b στοιχείων τα οποία μπορούμε να εξαγάγουμε από ένα σύνολο a στοιχείων τότε:

$$\text{Pr ob}(N = x) = \frac{C_{n_j}^x C_{n-n_j}^{n_k-x}}{C_n^{n_k}}.$$

$$\text{Συνεπώς } p_k(j) = \text{Pr ob}(N \geq n_{kj}) = \sum_{x=n_{kj}}^{x=n_k} \text{Pr ob}(N = x).$$

Χαρακτηρισμός κάθε συστάδας από τις λέξεις

Βήμα 1: Υπολογίστε τον αριθμό εμφάνισης μιας λέξης σε μια συστάδα.

Βήμα 2: Υπολογίστε τον αριθμό εμφάνισης μιας λέξης.

Βήμα 3: Υπολογίστε τον αριθμό εμφάνισης μιας λέξης σε μια συστάδα προς τον αριθμό εμφάνισης αυτής της λέξης.

Βήμα 4: Υπολογίστε τον αριθμό εμφάνισης μιας λέξης σε μια συστάδα προς το συνολικό αριθμό των λέξεων για αυτή τη συστάδα.

Βήμα 5: Υπολογίστε τον αριθμό εμφάνισης μιας λέξης στη συστάδα προς το συνολικό αριθμό των λέξεων.

Οι δείκτες τεχνολογίας και η απεικόνιση των πληροφοριών

Ένα από τα σημαντικά χαρακτηριστικά αυτού του συστήματος είναι η παραγωγή των δεικτών τεχνολογίας μέσω της ανάλυσης των συστάδων. Η γραφική απεικόνισή τους (Σχήμα 43) θα παρουσιαστεί στο τμήμα της απλής ανάλυσης (simple analysis) που ακολουθεί.

Όπως ήδη έχει αναφερθεί αυτοί οι δείκτες είναι βασισμένοι σε απλούς στατιστικούς υπολογισμούς όπως τα ποσοστά και άλλα στατιστικά μέτρα. Επιπλέον, μπορούν να ταξινομηθούν σε διαφορετικούς τύπους δεικτών ανάλογα με τον πεδίο των πατεντών που είναι βασισμένοι. Επομένως, μπορούν να παραχθούν (Σχήμα 44):

- δείκτες στο επίπεδο χωρών/ηπείρων,
- δείκτες στο επίπεδο εφευρετών/εταιριών,
- δείκτες στο επίπεδο του τομέα τεχνολογίας (μέσω των κωδίκων IPC),
- χρονικοί δείκτες, κλπ.

Οι δείκτες αυτοί μπορούν επίσης να απεικονιστούν γραφικά μέσω διαφορετικών τύπων γραφημάτων. Η ερμηνεία αυτών των δεικτών τεχνολογίας επιτρέπει τη σύλληψη της τρέχουσας τεχνολογίας και τον προσδιορισμό των τάσεων, της εξέλιξη κατά τη διάρκεια του χρόνου ή άλλων φαινομένων σχετικά με την τεχνολογική δραστηριότητα.

Σχήμα 42. Δενδρόγραμμα.

Σχήμα 43. Απεικόνιση clusters.

Σχήμα 44. Απεικόνιση των αποτελεσμάτων ανά cluster.

3.8 Simple ανάλυση

Η **simple analysis (απλή ανάλυση)** χρησιμοποιείται για την περιγραφή των πατεντών μέσω απλών στατιστικών και γραφικών παραστάσεων είτε για τα clusters που έχουν παραχθεί είτε για το σύνολο των δεδομένων. Στη συνέχεια δίνονται οι αλγόριθμοι για την παραγωγή των πιο σημαντικών γραφικών παραστάσεων που χρησιμοποιούνται από το σύστημα (Σχήμα 47).

3.8.1 Ιστογράμματα

Τα **ιστογράμματα (histograms)** είναι οι πιο γνωστές μορφές γραφικών παραστάσεων και απεικονίζουν τη συχνότητα κάθε τιμής (ο ένας άξονας αφορά στις τιμές και ο άλλος στις συχνότητες). Για την κατασκευή τους τα βήματα είναι τα ακόλουθα:

Βήμα 1: Πρέπει να επιλέξουμε τον αριθμό των bins που θα χρησιμοποιήσουμε για το ιστόγραμμα. Αυτό σημαίνει ότι πρέπει να επιλέξουμε τον αριθμό k για να κατασκευαστούν τα $k+1$ ξένα διαστήματα $A_j = (a_{j-1}, a_j)$ με $j = 1, \dots, k+1$, όπου a_j είναι τα σημεία τέλους κάθε διαστήματος με τη σύμβαση ότι $a_{-1} = -\infty$ και $a_{k+1} = \infty$ με στόχο την ανάλυση της σειράς των τιμών μας. Αυτό είναι μια μάλλον δύσκολη εργασία αφού δεν υπάρχει μια σαφής απάντηση για τον αριθμό bins που πρέπει να χρησιμοποιήσουμε. Ο βέλτιστος αριθμός εξαρτάται από τη φύση των δεδομένων, το μέγεθος των δειγμάτων και την άγνωστη ελλοχεύουσα διασπορά από την οποία οι παρατηρήσεις προκύπτουν. Ως εκ τούτου μπορούμε είτε να επιλέξουμε έναν αυθαίρετο αριθμό k είτε να χρησιμοποιήσουμε κάποιο κριτήριο. Επιλέγοντας τον αριθμό των bins μπορούμε να επιλέξουμε το πλάτος τους έστω d . Κατόπιν τα σημεία τέλους κάθε διαστήματος μπορούν να υπολογιστούν από τον τύπο $a_j = a_{\infty} + jd$.

Βήμα 2: Μέτρηση του αριθμού των παρατηρήσεων σε κάθε διάστημα. Καθορίζοντας τα διαστήματα, κάθε παρατήρηση ανήκει σε ένα διάστημα και συνεπώς μετράμε απλά τον αριθμό των παρατηρήσεων σε κάθε διάστημα.

Βήμα 3: Σχηματισμός ενός box στο μεσαίο σημείο κάθε διαστήματος με το ύψος ίσο με τη συχνότητα του διαστήματος.

Οι παρατηρήσεις που βρίσκονται στο σημείο τέλους ενός διαστήματος ανήκουν στο μεγαλύτερο διάστημα. Στην περίπτωση που τα διαστήματα δεν έχουν το ίδιο πλάτος θα προσαρμόσουμε την περιοχή για αυτό το διάστημα να είναι ανάλογη προς τη συχνότητά της. Για το ίδιο σύνολο δεδομένων μπορούμε να λάβουμε πολύ διαφορετικές εικόνες με τη χρησιμοποίηση ενός διαφορετικού αριθμού bins. Για παράδειγμα, στο Σχήμα 45 μπορούμε να δούμε ένα ιστόγραμμα για ένα σύνολο στοιχείων από 3.000 παρατηρήσεις. Το πλάτος των bins είναι ίσο με 10. Στο Σχήμα 46 μπορούμε να δούμε ένα ιστόγραμμα για τα ίδια δεδομένα αλλά με πλάτος bins ίσο με 20.

Σχήμα 45. Ιστόγραμμα για ένα σύνολο από 3.000 δεδομένα με μέγεθος bins=10.

Σχήμα 46. Ιστόγραμμα για το ίδιο σύνολο δεδομένων αλλά με μέγεθος bins=20.

3.8.2 Ραβδογράμματα

Τα **ραβδογράμματα (barcharts)** είναι γραφικές παραστάσεις που αντιπροσωπεύουν οπτικά (χρησιμοποιώντας ράβδους) τα δεδομένα σε μεμονωμένο επίπεδο ή στις κατανομές συχνότητας και ποσοστού. Ακριβέστερα, είναι ιδιαίτερα χρήσιμα στην περίπτωση που εξετάζουμε τις ποσοτικές (συνεχείς) μεταβλητές, που έχουν αθροιστεί, δηλαδή δεν μας δίνονται οι αναλυτικές τιμές αυτών των μεταβλητών αλλά μόνο οι περιλήψεις για τις μεγάλες ομάδες περιπτώσεων. Κάθε ράβδος αντιπροσωπεύει την τιμή μιας μεμονωμένης παρατήρησης, όπου αυτή η παρατήρηση αναφέρεται σε μια συχνότητα, ένα ευρετήριο ή ένα ποσοστό. Επίσης χρησιμοποιούνται όταν θέλουμε να εξετάσουμε τα στοιχεία στην ονομαστική ή τακτική κλίμακα. Σε αυτή την περίπτωση, κάθε ράβδος (το ύψος της ειδικότερα) αντιπροσωπεύει τη συχνότητα (πόσες φορές κάθε ευδιάκριτη τιμή έχει εμφανιστεί στο σύνολο των στοιχείων), ή το ποσοστό κάθε ευδιάκριτης τιμής.

Τα βήματα για να κατασκευαστεί ένα ραβδόγραμμα είναι τα ακόλουθα:

Βήμα 1: Διαβάστε τις n τιμές της μεταβλητής X δηλαδή $X = (x_1, x_2, \dots, x_n)$.

Βήμα 2: Δημιουργήστε ένα σύστημα συντεταγμένων στο οποίο: ο κάθετος άξονας παίρνει τιμές από 0 έως τη μέγιστη τιμή της X και ο οριζόντιος δεν έχει καμία μέτρηση.

Βήμα 3: Δημιουργήστε n ράβδους (ορθογώνια κουτιά) που θα είναι τοποθετημένα το ένα δίπλα στο άλλο στον οριζόντιο άξονα του βήματος 2 και που θα έχουν τα ακόλουθα χαρακτηριστικά: α) το ύψος κάθε ράβδου H_i είναι ίσο με την τιμή της παρατήρησης i^{th} δηλαδή $H_i = x_i$, $i = 1, \dots, n$, β) όλες οι ράβδοι έχουν απαραίτητα το ίδιο πλάτος και γ) οι ράβδοι ισαπέχουν.

3.8.3 Διαγράμματα πιτών

Τα **διαγράμματα πιτών (pie charts)** είναι ιδιαίτερα αποδοτικά σε περιπτώσεις που ο βασικός στόχος είναι η παρουσίαση μιας ολότητας κατά τέτοιο τρόπο ώστε τα σχετικά μεγέθη να είναι διευκρινισμένα. Σε αντίθεση με τα άλλα είδη των γραφικών παραστάσεων, τα

διαγράμματα ποτών μπορούν να απεικονίσουν μια μεταβλητή κάθε φορά. Σε ένα διάγραμμα πίτας, το σύνολο παρουσιάζεται σαν κύκλος και το ποσοστό των περιπτώσεων που ανήκουν σε μια συγκεκριμένη κατηγορία εμφανίζεται ως τμήμα του κύκλου. Κατά συνέπεια:

αριθμός περιπτώσεων στην κατηγορία A / συνολικός αριθμός περιπτώσεων =

τομέας της κατηγορίας A / συνολικός τομέας του κύκλου =

κεντρική γωνία της κατηγορίας A / συνολική γωνία γύρω από τον κύκλο.

Η συνηθισμένη χρήση των διαγραμμάτων πίτας είναι η απεικόνιση των συχνοτήτων μιας ποιοτικής μεταβλητής. Μπορούν επίσης να χρησιμοποιηθούν στην περίπτωση που έχουμε ένα ποιοτικά αθροισμένα δεδομένα. Σε αυτή την περίπτωση, κάθε φέτα της πίτας αντιπροσωπεύει το σχετικό μέγεθος κάθε μεμονωμένης περίπτωσης, παρατήρησης, ομάδας, κλπ.

Σχήμα 47. Simple analysis και γραφική αναπαράσταση.

Τα βήματα για να κατασκευαστεί ένα διάγραμμα πίτας είναι τα ακόλουθα:

Βήμα 1: Διαβάστε τη μεταβλητή X με τις n παρατηρήσεις (ομάδες), $X = (x_1, x_2, \dots, x_n)$.

Βήμα 2: Υπολογίστε τις ποσότητες $p_i = \frac{x_i}{\sum_{j=1}^n x_j}$ για $i = 1, \dots, n$. Με αυτό τον τρόπο,

παίρνουμε τις σχετικές τιμές των παρατηρήσεών μας.

Βήμα 3: Δημιουργήστε έναν κύκλο που όπως είναι γνωστό έχει τη γωνία 360° .

Βήμα 4: Χωρίστε τον κύκλο σε n κομμάτια (φέτες) όπου το πρώτο κομμάτι καλύπτει $p_1 * 360^\circ$, το δεύτερο έχει την κεντρική γωνία $p_2 * 360^\circ, \dots$, και το τελευταίο κομμάτι έχει την κεντρική γωνία $p_n * 360^\circ$.

3.9 Η αρχιτεκτονική του συστήματος

Το σύστημα αποτελεί ένα ευέλικτο εργαλείο εξειδικευμένο να αναλύει δεδομένα πατεντών (Markellos et al., 2003b), (Markellos et al., 2003c). Συνδυάζει διαφορετικά χαρακτηριστικά και τεχνολογίες καθώς και τη στατιστική μεθοδολογία που παρουσιάστηκε στις προηγούμενες ενότητες του κεφαλαίου με στόχο την επεξεργασία μεγάλου όγκου δεδομένων κειμένων. Χαρακτηρίζεται από τη δυνατότητα να παράγει αποτελέσματα με ένα γρήγορο και ακριβή τρόπο. Επίσης παρέχει διαφορετικούς τρόπους οπτικοποίησης αφού ο χρήστης έχει τη δυνατότητα να αποτυπώνει τα αποτελέσματα μέσω γραφημάτων, πινάκων και αναφορών.

Το σύστημα είναι παραθυρικό με ένα εύχρηστο περιβάλλον διεπαφής. Η αρθρωτή του αρχιτεκτονική δίνει τη δυνατότητα εφαρμογής διαφορετικών τύπων ανάλυσης. Συγκεκριμένα, αποτελείται από τρία **modules (τμήματα)** κάθε ένα από τα οποία αντιστοιχεί σε μια βασική διαδικασία του συστήματος (Σχήμα 48):

- **Database Manager Module:** εκτελεί την εισαγωγή των δεδομένων, το καθάρισμά τους και την προ-επεξεργασία τους. Ειδικότερα, το module αυτό διαβάζει τα δεδομένα των πατεντών από επιλεγμένες βάσεις δεδομένων και τα μετατρέπει στην κατάλληλη μορφή ώστε να είναι έτοιμα για περαιτέρω επεξεργασία. Το σύστημα δέχεται δεδομένα κειμένου και αριθμητικά από τα πεδία των πατεντών. Ένας parser χρησιμοποιείται για το διάβασμα των δεδομένων και στη συνέχεια εφαρμόζεται λεξικολογική επεξεργασία. Η χρήση ενός λεξικού και μια γραμματικής είναι απαραίτητα.
- **Statistical Analysis Module:** εκτελεί απλή στατιστική ανάλυση καθώς επίσης και τις αναλύσεις της προτεινόμενης μεθοδολογίας. Συγκεκριμένα, εφαρμόζει μεθόδους ανάλυσης κειμένων στα προ-επεξεργασμένα δεδομένα ώστε να εξαχθούν χρήσιμες πληροφορίες και να δημιουργηθούν οι πρώτες ομάδες των πατεντών. Εάν ο χρήστης επιλέξει την correspondence analysis θα πρέπει προηγουμένως να έχει εφαρμοστεί λεξικογραφική ανάλυση στα δεδομένα. Η cluster analysis αποτελεί κύριο χαρακτηριστικό του συστήματος και τη βάση για την παραγωγή των τεχνολογικών δεικτών. Εξαρτάται από τον αριθμό των δεδομένων και διακρίνεται σε hierarchical clustering και στο k-means clustering.

- **Results Presentation Module:** εκτελεί τη γραφική απεικόνιση των αποτελεσμάτων. Αποτελεί πολύ σημαντικό module για τον τελικό χρήστη ώστε να μπορέσει να κατανοήσει τη σημασία τους. Πολλές επιλογές είναι διαθέσιμες όπως γραφήματα, πίνακες, αναφορές, κλπ.

Σχήμα 48. Αρχιτεκτονική συστήματος.

3.10 Case studies

Οι δείκτες τεχνολογίας και καινοτομίας μπορούν να ορισθούν σαν «στατιστικές, οι οποίες μετρούν τις ποσοτικά προσδιορίσιμες πτυχές της δημιουργίας, της διάδοσης και της εφαρμογής της επιστήμης και της τεχνολογίας. Οι δείκτες πρέπει να βοηθήσουν στην περιγραφή ενός συστήματος επιστήμης και τεχνολογίας, επιτρέποντας την καλύτερη κατανόηση της δομής του, των επιρροών των πολιτικών και των προγραμμάτων σε αυτό και της επίδρασης της επιστήμης και της τεχνολογίας στην κοινωνία και την οικονομία». Από όλους του δείκτες οι πατέντες είναι αυτές που χρησιμοποιούνται πιο συχνά. Αυτό συμβαίνει γιατί οι πατέντες σχετίζονται με την καινοτομία και αποτελούν ένδειξη για τη δομή και την εξέλιξη των δραστηριοτήτων στις χώρες, τις περιοχές και τις βιομηχανίες. Είναι φανερό ότι οι δείκτες αυτοί εξαρτώνται από τον τύπο των πατεντών που έχουν επιλεχθεί για την ανάλυση. Επίσης πρέπει να σημειωθεί ότι δεν υπάρχει μια συγκεκριμένη μεθοδολογία για τον

υπολογισμό των δεικτών από δεδομένα πατεντών. Διαφορετικές στατιστικές μέθοδοι και τεχνικές μπορούν να εφαρμοστούν από απλές αναλύσεις μέχρι πολύπλοκες.

Η προτεινόμενη προσέγγιση μπορεί να παράγει δείκτες είτε συνολικά είτε μέσα στα clusters, ενώ μπορούν να κατηγοριοποιηθούν περαιτέρω σύμφωνα με το πεδίο των πατεντών με το οποίο σχετίζονται. Συγκεκριμένα έχουμε (Σχήμα 49):

- Δείκτες στο επίπεδο της περιοχής της τεχνολογίας.
- Δείκτες στο επίπεδο των ηπείρων/χωρών.
- Δείκτες στο επίπεδο των εφευρετών/εταιριών.
- Δείκτες σε σχέση με το χρόνο.

Σχήμα 49. Κατηγορίες δεικτών.

Για την αξιολόγηση της προσέγγισης υλοποιήθηκαν **4 περιπτώσιακές μελέτες (case studies)** στους ακόλουθους βιομηχανικούς τομείς με βάση τον κωδικό MC που περιγράφει συγκεκριμένες τεχνολογίες:

- **A23L:** FOODS, FOODSTUFFS, OR NON-ALCOHOLIC BEVERAGES, NOT COVERED BY SUBCLASSES A23B TO A23J; THEIR PREPARATION OR TREATMENT, e.g. COOKING, MODIFICATION OF NUTRITIVE QUALITIES, PHYSICAL TREATMENT (shaping or working, not fully covered by this subclass, A23P); PRESERVATION OF FOODS OR FOODSTUFFS, IN GENERAL.
- **F16F:** SPRING; SHOCK-ABSORBERS; MEANS FOR DAMPING VIBRATION.
- **G03:** PHOTOGRAPHY; CINEMATOGRAPHY; ANALOGOUS TECHNIQUES USING WAVES OTHER THAN OPTICAL WAVES; ELECTROGRAPHY; HOLOGRAPHY (reproduction of pictures or patterns by scanning and converting into electrical signals H04N).
- **G10:** MUSICAL INSTRUMENTS; ACOUSTICS.

Τα δεδομένα ανακτήθηκαν από το Volume 2000/006 της βάσης δεδομένων ESPACE ACCESS και περιλαμβάνουν πατέντες που αντιστοιχούν:

- WO-00/00001-00/10370 06 Jan. 2000-24 Feb. 2000.

- EP-A-0967853-0981268 05 Jan 2000-23 Feb. 2000.

Από την ανάλυση προκύπτουν διάφοροι δείκτες όπως:

- Η χώρα US είναι πιο ενεργή στον τεχνολογικό τομέα G10 και ακολουθεί η Japan (Πίνακας 15 και Σχήμα 50).
- Η εταιρία με το μεγαλύτερο αριθμό πατεντών στον τεχνολογικό τομέα G03 είναι η Canon Kabushiki Kaisha και ακολουθεί η Kodak (Πίνακας 16 και Σχήμα 51).
- Ο Πίνακας 17 και το Σχήμα 52 παρουσιάζουν την κατανομή των MC κωδικών σε κάθε cluster (ταξινόμηση με βάση τις συχνότητες των κωδικών) του τεχνολογικού τομέα A23L και δίνει μια άποψη για την ομοιογένεια αυτών των clusters. Η ύπαρξη μεγάλων συχνοτήτων υποδηλώνει την ομοιογένεια κάθε cluster. Ένα cluster μπορεί να θεωρηθεί πολύ ομοιογενές αν ο μεγαλύτερος κωδικός που ανήκει σε αυτό έχει ποσοστό μεγαλύτερο το 70% στο σύνολο. Τα clusters όπου ο μεγαλύτερος κωδικός υπάρχει σε ποσοστό μεταξύ 55% και 75% θεωρούνται σχετικά ομοιογενή. Τέλος, μη ομοιογενή θεωρούνται τα clusters στα οποία ο μεγαλύτερος κωδικός εμφανίζεται με ποσοστό από 35% έως 45% στο σύνολο. Από τα παραπάνω προκύπτει ότι το cluster 9 είναι το πιο ομοιογενές ενώ τα clusters 1, 5 και 7 είναι σχετικά ομοιογενή στον τομέα A23L.
- Ο Πίνακας 18 και το Σχήμα 53 παρουσιάζουν τους πιο ενεργούς εφευρέτες στον τεχνολογικό τομέα F16F.

Πίνακας 15. Η δραστηριότητα των χωρών (case study G10).

Σχήμα 50. Η δραστηριότητα των χωρών (case study G10).

Εταιρία	Αριθμός πατεντών	Ποσοστό πατεντών
Canon Kabushiki Kaisha	25	11.68%
Kodak	21	9.81%
Agfa	12	5.61%
Fuji Photo	7	3.27%
Minnesota Mining And Manufacturing Company	7	3.27%
Clariant International Ltd.	5	2.34%
Asm Lithography B.V.	4	1.87%
Hewlett-Packard Company	4	1.87%
Mita Industrial Co. Ltd.	4	1.87%
Seiko Epson Corporation	4	1.87%
Carl Zeiss	3	1.40%
Ciba Specialty Chemicals Holding Inc.	3	1.40%
Copyer Co., Ltd.	3	1.40%
Imation Corp.	3	1.40%
Oci Printing Systems GmbH	3	1.40%
Xerox Corporation	3	1.40%

Πίνακας 16. Η δραστηριότητα των εταιριών (case study G10).

Σχήμα 51. Η δραστηριότητα των εταιριών (case study G10).

ΚΕΦΑΛΑΙΟ 3
ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ

Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6		Cluster 7		Cluster 8		Cluster 9		Cluster 10	
A23L-1/00	50.00%	A23L-1/236	12.50%	A23L-1/0522	12.50%	A23L-1/30	15.38%	A23L-1/40	33.33%	A23L-1/00	7.69%	A23L-1/308	50.00%	A23L-1/308	18.18%	A23L-1/0532	100.00%	A23L-1/16	25.00%
A23L-1/01	25.00%	A23L-1/22	8.33%	A23L-1/0524	12.50%	A23L-3/00	15.38%	A23L-1/0522	16.67%	A23L-1/056	7.69%	A23L-1/317	50.00%	A23L-1/0522	9.09%			A23L-1/325	25.00%
A23L-1/052	25.00%	A23L-1/30	8.33%	A23L-	6.25%	A23L-1/00	7.69%	A23L-1/0528	16.67%	A23L-1/09	7.69%			A23L-1/09	9.09%			A23L-3/01	25.00%
		A23L-1/302	8.33%	A23L-1/05	6.25%	A23L-1/0562	7.69%	A23L-1/302	16.67%	A23L-1/272	7.69%			A23L-1/16	9.09%			A23L-3/3418	25.00%
		A23L-1/304	8.33%	A23L-1/056	6.25%	A23L-1/10	7.69%	A23L-2/385	16.67%	A23L-1/29	7.69%			A23L-1/29	9.09%				
		A23L-1/16	4.17%	A23L-1/09	6.25%	A23L-1/212	7.69%			A23L-1/304	7.69%			A23L-1/30	9.09%				
		A23L-1/20	4.17%	A23L-1/22	6.25%	A23L-1/217	7.69%			A23L-1/305	7.69%			A23L-1/305	9.09%				
		A23L-1/212	4.17%	A23L-1/221	6.25%	A23L-1/275	7.69%							A23L-2/02	9.09%				
		A23L-1/217	4.17%	A23L-1/226	6.25%	A23L-1/31	7.69%							A23L-2/66	9.09%				
		A23L-1/226	4.17%	A23L-1/31	6.25%	A23L-3/18	7.69%							A23L-3/015	9.09%				
		A23L-1/24	4.17%	A23L-1/325	6.25%	A23L-3/34	7.69%												
		A23L-1/325	4.17%	A23L-1/39	6.25%														
		A23L-1/333	4.17%	A23L-2/00	6.25%														
		A23L-2/02	4.17%	A23L-2/78	6.25%														
		A23L-3/00	4.17%																
		A23L-3/3463	4.17%																
		A23L-3/3562	4.17%																
		A23L-3/375	4.17%																

A23L- 1/00	A23L- 1/24	A23L- 2/52
A23L- 1/05	A23L- 1/275	A23L- 2/78
A23L- 1/0522	A23L- 1/30	A23L- 3/01
A23L- 1/0528	A23L- 1/304	A23L- 3/16
A23L- 1/056	A23L- 1/308	A23L- 3/34
A23L- 1/09	A23L- 1/317	A23L- 3/3463
A23L- 1/16	A23L- 1/328	A23L- 3/365
A23L- 1/212	A23L- 1/39	A23L-3 00
A23L- 1/22	A23L- 2/00	
A23L- 1/226	A23L- 2/04	

Πίνακας 17. Οι κωδικοί MC ανά cluster (case study A23L).

Σχήμα 52. Οι κωδικοί MC ανά cluster (case study A23L).

Χώρα	Αριθμός πατεντών	Ποσοστό πατεντών
Bösing, Klaus	2	1.69%
Chauvet, Ludovic	2	1.69%
Cooke, Richard, David, Maitland	2	1.69%
Curtis, Anthony, John	2	1.69%
Eckel, Hans-Gerd, Dr.	2	1.69%
Hilburger, Mark D.	2	1.69%
Lambert, Adam, Johnson	2	1.69%
Leonard, Daniel J.	2	1.69%
Pohl, Andreas	2	1.69%
Rosenfeldt, Horst	2	1.69%
Taghizadch, Mohamad	2	1.69%
Tournier, Pascal	2	1.69%
Wendt, Eckhardt	2	1.69%

Πίνακας 18. Οι πιο ενεργοί εφευρέτες (case study F16F).

Σχήμα 53. Οι πιο ενεργοί εφευρέτες (case study F16F).

3.11 Συμπεράσματα και ανοικτά θέματα

Η προτεινόμενη μεθοδολογική προσέγγιση αναλύει δεδομένα διπλωμάτων ευρεσιτεχνίας με βάση πολυδιάστατες τεχνικές. Με αυτό τον τρόπο διευκολύνει τη σύλληψη πληροφοριών όχι μόνο στο επίπεδο ενός τεχνολογικού τομέα αλλά και στο επίπεδο μιας χώρας ή ενός συνόλου από τομείς επιτρέποντας την εύκολη σύγκριση των αποτελεσμάτων και οδηγώντας στην παραγωγή δεικτών για τις τεχνολογικές τάσεις και την καινοτομία. Στη συνέχεια παρουσιάζουμε τα καινοτόμα στοιχεία της αναπτυγμένης προσέγγισης τα οποία σχετίζονται κυρίως με τη χρήση και το συνδυασμό των στατιστικών τεχνικών που επιτρέπουν την αποτελεσματικότερη ανάλυση των πατεντών.

- **Δεδομένα.** Τα δεδομένα που χρησιμοποιούνται από το σύστημα λαμβάνουν υπόψη τους όλες τις πληροφορίες που περιγράφουν μια πατέντα. Σε σύγκριση με τις υπάρχουσες προσεγγίσεις (όπως της EUROSTAT), που εκμεταλλεύεται μόνο τα πρώτα ψηφία των κωδίκων IPC και συνεπώς χάνει κάποιες πληροφορίες, στο σύστημά μας οι αναλύσεις είναι βασισμένες στην εκμετάλλευση όλων των πεδίων. Πιο συγκεκριμένα, λαμβάνουμε υπόψη όχι μόνο τα πρώτα ψηφία των κωδίκων IPC αλλά όλο τον κωδικό καθώς επίσης και τους δευτερεύοντες κωδικούς. Παράλληλα, η εκμετάλλευση των κειμένων που περιγράφουν μια πατέντα όπως ο τίτλος και η περίληψη επιτρέπει να έχουμε πιο ολοκληρωμένη περιγραφή των τεχνικών χαρακτηριστικών της. Τέλος, η χρήση άλλων πεδίων όπως οι χώρες, οι επιχειρήσεις, τα έτη κλπ. βοηθούν σημαντικά στην εκτέλεση διαφορετικών τύπων αναλύσεων και στην απεικόνιση των πληροφοριών μέσω διαφορετικών τρόπων.
- **Επίπεδα ανάλυσης.** Μέσω της προσέγγισής μας υπάρχει η δυνατότητα να αναλυθούν οι επιστημονικές και τεχνολογικές καινοτομίες και η πρόοδος στην Ευρώπη σε τρία διαφορετικά επίπεδα. Έτσι, αρχικά η ανάλυση μπορεί να εκτελεστεί στο επίπεδο ενός τομέα. Αυτό σημαίνει ότι ένας συγκεκριμένος τομέας μπορεί να απομονωθεί προκειμένου να αναλυθεί και να προσδιοριστεί συνεπώς η τεχνολογική του εξέλιξη, η οποία δεν είναι εμφανής. Συγκεκριμένα, θα χρησιμοποιούμε για την ανάλυση τους κωδικούς IPC ή/και το κείμενο που περιγράφει την πατέντα προκειμένου να δημιουργήσουμε ομοιογενείς ομάδες από τις οποίες μπορούν να προσδιοριστούν τεχνολογικές τάσεις και να αντληθούν χρήσιμες πληροφορίες. Οι μεταβλητές που δεν περιλαμβάνονται στο κύριο μέρος της ανάλυσης όπως οι επιχειρήσεις που υποβάλλουν τις πατέντες, οι εφευρέτες, οι χώρες στις οποίες υποβλήθηκαν οι πατέντες κλπ. εμπλουτίζουν περαιτέρω το αποτέλεσμα της ομαδοποίησης και χρησιμοποιούνται ως συμπληρωματικές μεταβλητές ανάλυσης. Η ίδια ανάλυση μπορεί να εκτελεστεί στο επίπεδο όλων των τομέων. Αυτό μας επιτρέπει να πάρουμε πληροφορίες για την επιστημονική και την τεχνολογική δραστηριότητα ανά επιλεγμένο τομέα σχετικά με το σύνολο των επιχειρήσεων που υποβάλλουν πατέντες σε αυτό το συγκεκριμένο τομέα, για ένα συγκεκριμένο έτος ή για αρκετά έτη. Επιπλέον, η ανάλυση στο επίπεδο μιας χώρας είναι διαθέσιμη. Μέσω αυτής μπορούν να διαμορφωθούν ομοιογενείς ομάδες χωρών, οι οποίες παρουσιάζουν την πρόοδο της επιστημονικής και της τεχνολογικής εξέλιξης στις διαφορετικές χώρες της Ευρώπης.

- **Στατιστικές μέθοδοι.** Το καινοτόμο στοιχείο από την άποψη των χρησιμοποιούμενων στατιστικών μεθόδων δεν σχετίζεται τόσο με τις ίδιες τις μεθόδους αλλά με το συνδυασμό τους για την ανάλυση των πατεντών. Η ενσωμάτωση τεχνικών ανάλυσης κειμένου για την επεξεργασία του περιεχομένου των πατεντών (τίτλοι και περιλήψεις) είναι ένα από τα σημαντικά χαρακτηριστικά της προτεινόμενης μεθοδολογίας. Η ιδέα είναι ότι το λεξιλόγιο που είναι χαρακτηριστικό για τις κατηγορίες πατεντών βασισμένο σε διάφορες περιγραφικές μεταβλητές που συνδέονται με αυτές (όπως η χώρα ή η ημερομηνία της αίτησης), παρέχει πρόσθετες ενδιαφέρουσες ιδέες για την ανάλυση των ίδιων των πατεντών. Ανταγωνιστικοί δείκτες σχετικά με το ανταγωνιστικό επίπεδο κάθε χώρας μπορούν να εξαχθούν, καθώς επίσης και χώρες που είναι ενεργές στη συγκεκριμένη τεχνολογική περιοχή, κλπ. Ο συνδυασμός correspondence και cluster ανάλυσης παρέχει ένα ισχυρό εργαλείο για την αποδοτική εκμετάλλευση των πατεντών. Τέλος, η παραγωγή των δεικτών τεχνολογίας μέσω της χρήσης της ανάλυσης συστάδων είναι ένα από τα σημαντικά χαρακτηριστικά γνωρίσματα της προτεινόμενης μεθοδολογίας.

Τέλος, στα ανοικτά προβλήματα για μελλοντική εργασία περιλαμβάνονται τα ακόλουθα:

- **Η ενσωμάτωση σημασιολογίας.** Η σημασιολογική ανάλυση των δεδομένων των πατεντών θα βοηθήσει στη δημιουργία μιας πιο πλούσιας αναπαράστασής τους και στη σύλληψη των μεταξύ τους σχέσεων και συσχετίσεων. Επίσης, μπορεί να παίξει σημαντικό ρόλο στο στάδιο της εξόρυξης της γνώσης ώστε να οδηγήσει σε καλύτερα αποτελέσματα. Αν και οι σημασιολογικές μέθοδοι αυτές απαιτούν σημαντικό υπολογιστικό χρόνο αποτελεί ερευνητική πρόκληση η διερεύνηση του τρόπου που θα μπορέσουν να γίνουν πολύ περισσότερο αποδοτικές και scalable ιδιαίτερα για μεγάλα σύνολα δεδομένων.
- **Το πρόβλημα της πολυγλωσσίας.** Οι τεχνικές TM που υιοθετήθηκαν από την προτεινόμενη μεθοδολογία εφαρμόστηκαν στα αγγλικά κείμενα των πατεντών. Αποτελεί σημαντικό ζήτημα που απαιτεί επιπλέον μελέτη ο σχεδιασμός και η ανάπτυξη αλγορίθμων που θα επεξεργάζονται πολύγλωσσα κείμενα και θα παράγουν ανεξάρτητες της γλώσσας ενδιάμεσες μορφές τους. Η εξόρυξη γνώσης από αυτές θα επιτρέψει την πρόσβαση σε νέες πληροφορίες και θα ανοίξει νέες δυνατότητες εκμετάλλευσής τους.
- **Παροχή εξατομικευμένων δυνατοτήτων.** Τα τρέχοντα συστήματα και εργαλεία που εφαρμόζουν TM τεχνικές απευθύνονται σε ειδικευμένους χρήστες και σε εμπειρογνώμονες. Οι μελλοντικές εφαρμογές θα ήταν χρήσιμο να ενσωματώνουν μεγαλύτερο βαθμό ευχρηστίας ώστε να αφορούν περισσότερες κατηγορίες χρηστών π.χ. με την αυτόματη μετάφραση φυσικής γλώσσας ερωτημάτων που αυτόματα θα εκτελούν τις κατάλληλες λειτουργίες εξόρυξης. Επίσης τα εργαλεία αυτά θα μπορούσαν να έχουν την μορφή έξυπνων προσωπικών πρακτόρων (intelligent personal agents) οι οποίοι θα μαθαίνουν το προφίλ του κάθε χρήστη και με βάση αυτό θα εκτελούν τις TM λειτουργίες αυτόματα χωρίς να είναι απαραίτητη η ρητή απαίτησή τους από αυτόν.
- **Εφαρμογή της μεθοδολογίας σε άλλο πεδίο.** Ενδιαφέρον θα παρουσίαζε η εφαρμογή της προτεινόμενης μεθοδολογικής προσέγγισης σε διαφορετικών τύπων ημι-δομημένα ή αδόμητα δεδομένα όπως κείμενα ή ιστοσελίδες. Ιδιαίτερα η

εφαρμογή της σε δεδομένα Web π.χ. για το πρόβλημα της ταξινόμησης του περιεχομένου ενός web site ή για την πρόβλεψη των σελίδων που θα έχουν τη μεγαλύτερη ζήτηση πιστεύουμε ότι θα οδηγήσει σε χρήσιμα συμπεράσματα.

3.12 Βιβλιογραφία

Alderferer, M.S., & Blashfield, R.K. (1986). *Cluster Analysis*. Beverly Hills, CA., Sage Publications, Inc.

Bao, H. (2004). Introduction to Knowledge Discovery and Data Mining. Institute of Information Technology National Center for Natural Science and Technology. Retrieved from: <http://www.netnam.vn/unescocourse/knownlegde/knownlegd.htm>.

Beaugrande, R., & Dressler, W. (1981). *Introduction to Text Linguistics*. London Longman.

Benzecri, J.P. (1992). *Correspondence Analysis Handbook*. New York, Marcel Dekker.

Chappelier, J., Peristera, V., Rajman, M., Seydoux, F., & Spinakis, A. (2002). Evaluation of Statistical and Technological Innovation Using Statistical Analysis of Patents, Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT), Saint-Malo (France), March.

Comanor, W.S., & Scherer, F.M. (1969). Patent Statistics as a Measure of Technical Change. *Journal of Political Economy*, 392-398.

El Wakil, M. (2002). Introducing Text Mining. Proceedings of the 9th Scientific Conference for Information Systems and Information Technology (ISIT02). Retrieved from: http://myjavasever.com/~elwakil/mohamed/SCISIT02_EIWakil.pdf.

EPO (2005). European Patent Office. Retrieved from: <http://www.european-patent-office.org/index.htm>.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39(11): 27-34.

Fielding, A. (1992). Axiomatic Approaches to Scoring Ordered Classifications. University of Birmingham, Department of Economics, Discussion Paper 92-06.

Fisher, R.A. (1940). The Precision of Discriminant Functions. *Annals of Eugenics*, London, 10: 422-429.

Gifi, A. (1990). *Non Linear Multivariate Analysis*. J. Wiley.

Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond Data Warehousing: What's Next in Business Intelligence?. *Proceedings of the 7th International Workshop on Data Warehousing and OLAP (DOLAP 2004)*, November 12-13, Washington, DC, USA, 1-6.

- Greenacre, M. (1984). *Theory and Applications of Correspondence*. Analysis. London, Academic Press.
- Guttman, L. (1941). The Quantification of a Class of Attributes: A Theory and a Method of Scale Construction. *The Prediction of Personal Adjustment*, Horst et al. (Eds.), New York, Social Science Research Council.
- Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Hayashi, C. (1956). Theory and Examples of Quantification (II). *Proceedings of Institute of Statistical Mathematics*, 4(2):19-30.
- Hill, M.O. (1974). Correspondence Analysis: a Neglected Multivariate Technique. *Applied Statistics*, 23: 340-354.
- Hirschfeld, H.O. (1935). A Connection between Correlations and Contingency. *Proceedings of the Cambridge Philosophical Society*, 31: 520-524.
- Horst, P. (1935). Measuring Complex Attitudes. *Journal of Social Psychology*, 6: 369-374.
- Lewis, S. (1991). Cluster Analysis as a Technique to Guide Interface Design. *International Journal of Man-Machine Studies*, 35: 251-265.
- Liao, S. (2003). Knowledge Management Technologies and Applications – Literature Review from 1995 to 2002. *Expert Systems with Applications*, Elsevier Science Ltd, 25: 155-164.
- Markellos, K., Markellou, P., Mayritsakis, G., Panagopoulou, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2003a). STING: Evaluation of Scientific and Technological Innovation and Progress in Europe through Patents. In the book *Statistical Data Mining and Knowledge Discovery*, edited by Hamparsum Bozdogan, Publisher: Chapman & Hall/CRC Press, ISBN: 1584883448. Retrieved from: http://www.amazon.com/exec/obidos/ASIN/1584883448/qid%3D1066835737/sr%3D2-3/ref%3Dsr%5F2%5F3/104_-9162346-7852756.
- Markellos, K., Markellou, P., Mayritsakis, G., Panagopoulou, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A., (2002a). STING: Evaluation of Scientific and Technological Innovation and Progress in Europe through Patents. *Proceedings of the New Frontiers of Statistical Data Mining and Knowledge Discovery and E-Business Conference*, Knoxville, Tennessee, USA, 22-25 June 2002.
- Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A., (2002b). Knowledge Discovery in Patent Databases. *Proceedings of the ACM CIKM 2002, 11th International Conference on Information and Knowledge Management*, Saic Headquarters, McLean, Virginia, USA, 4-9 November, 672-674. Retrieved from: <http://portal.acm.org/citation.cfm?id=584915&coll=GUIDE&dl=GUIDE&ret=1#FullText>.
- Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2001a). STING: Evaluation of Scientific & Technological Innovation and Progress in Europe through Patents. *Proceedings of the 1st STING User-Focus Group Meeting*, Lausanne, Switzerland, 30 November 2001, 23-29.

Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2001b). STING: Functionalities of the System Prototype. *Proceedings of the 1st STING User-Focus Group Meeting*, Lausanne, Switzerland, 30 November 2001, 51-58.

Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2003b). A Multivariate Analysis and Visualization System for Analyzing Patent Data and Producing Technological and Scientific Indicators. *Abstract Proceedings of the International Workshop on Computational Management Science, Economics, Finance and Engineering*, , Limassol, Cyprus, 28-30 March 2003, p. 36.

Markellos, K., Markellou, P., Mayritsakis, G., Perdikouri, K., Sirmakessis, S., & Tsakalidis, A. (2003c). Mining Patent Databases for Monitoring Technological Trends. *Proceedings of the PANDA Workshop on Pattern-base Management Systems (PBMS)*, Como, Italy, 10th April 2003, 83-92.

Markellou, P., Panayiotaki, A., Sirmakessis, S., Spinakis, A., & Tsakalidis, A. (2001). Using Hierarchy to Extract Innovation; The Use of Patents in Clarifying Innovation. *Pre-Proceedings of the ETK-NTTS 2001 Exchange of Technology and Know-How, New Techniques-Technologies for Statistics*, Poster Session, Hersonissos, Crete, Greece, 18-22 June 2001, 881-882. Retrieved from: http://webfarm.jrc.cec.eu.int/etk-ntts/poster_session_4.htm.

McDonald, K., Wilmsmeier, A., Dixon, D.C., Inmon, W.H. (2002). *Mastering the SAP Business Information Warehouse*. Wiley Publishers, ISBN 0471219711.

MIMOSA (2005). MIMOSA Software. Retrieved from: http://db1.european-patent-office.org/www3/dwld/software/update_5-0-5-3.exe.

Narin, F. (1995). Patents as Indicators for the Evaluation of Industrial Research Output. *Scientometrics*, 34(3): 489-496.

Nishisato, S. (1980). Analysis of Categorical Data. *Dual Scaling and Its Application*, University of Toronto Press.

Piatetsky-Shapiro, G., Fayyad, U., & Smith P. (1996). From Data Mining to Knowledge Discovery: An Overview. In U.M. Fayyad, et al. (Eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1-35.

Rajman, M., & Lebart, L. (1998). Similarities for Textual Data. *Proceedings of the 4th International Conference on Statistical Analysis of Textual Data (JADT'98)*, Nice.

Tan, A. (1999). Text Mining: The State of the Art and the Challenges. *Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD'99)*, Workshop on Knowledge Discovery from Advanced Databases, 65-70. Retrieved from: http://www.ewa-strategist.com/papers/text_mining_kdad99.pdf.

Βαζιργιάννης, Μ., & Χαλκίδη, Μ. (2003). *Εξόρυξη Γνώσεις από Βάσεις Δεδομένων*. Εκδόσεις Τυπωθήτω, Γιώργος Δάρδανος, Αθήνα.

OBI (2005). Οργανισμός Βιομηχανικής Ιδιοκτησίας. Retrieved from: <http://www.obι.gr/online>.

Γλωσσάρι όρων

ADL/SCORM	Advanced Distributed Learning/Sharable Content Object Reference Model. Πρότυπο e-learning μεταδεδομένων.
AICC	Aviation Industry Computer Based Training Committee. Πρότυπο e-learning μεταδεδομένων.
ARIADNE	Alliance of Remote Instructional Authoring & Distribution Network for Europe. Πρότυπο e-learning μεταδεδομένων.
Association rules	Κανόνες συσχέτισης. Χρησιμοποιούνται για να εντοπίσουν ομοιότητες μεταξύ διαφορετικών τύπων δεδομένων και να επιτευχθεί καλύτερη κατανόηση της συμπεριφοράς των χρηστών. Οι κανόνες αυτοί συνδέουν ένα ή περισσότερα (αρχικά μη συσχετιζόμενα) γεγονότα και ανακαλύπτουν σχέσεις που δεν μπορούν (εύκολα) να προβλεφθούν. Ένας τέτοιος κανόνας έχει τη μορφή $A \Rightarrow B$ και υποδηλώνει μια σχέση εξάρτησης ανάμεσα σε δύο (ξένα) σύνολα στοιχείων A και B δηλαδή υπολογίζει την πιθανότητα να συμβεί το B με δεδομένο ότι έχει συμβεί το A.
Business intelligence (BI)	Επιχειρηματική Ευφυΐα/Νοημοσύνη. Η αμφίδρομη διαδικασία ανάλυσης και αξιοποίησης δομημένων πληροφοριών (που συχνά αποθηκεύονται σε data warehouses), με τη βοήθεια των οποίων η διοίκηση μιας επιχείρησης παρακολουθεί αποτελεσματικότερα νέες τάσεις και πρότυπα, εξάγει χρήσιμα συμπεράσματα και ως εκ τούτου λαμβάνει ορθότερες αποφάσεις. Η Επιχειρηματική Ευφυΐα αφορά σε πελάτες, προϊόντα, υπηρεσίες ή ανταγωνιστές.
Business-to-consumer (B2C)	Εμπορικές σχέσεις μεταξύ επιχειρήσεων και πελατών. Υπηρεσίες που ξεκινούν από επιχειρήσεις και απευθύνονται στο ευρύτερο καταναλωτικό κοινό.
Caching	Οι τεχνικές caching χρησιμοποιούνται πολύ συχνά από τα RSs με στόχο τη βελτίωση της απόδοσης τους. Με τη χρήση content-caching οι σελίδες που ζητούνται συχνά δεν χρειάζεται να ανακτούνται από απόσταση, ή να αναπαράγονται στο server για κάθε μια πρόσβαση. Αυτό μπορεί να μειώσει την καθυστέρηση (latency) στην παραλαβή των ιστοσελίδων, όπως επίσης να μειώσει το φορτίο στο server και το δίκτυο. Στο περιβάλλον του Web οι ιστοσελίδες που εμφανίζουν υψηλά ποσοστά προσβάσεων μπορούν να γίνουν cached στον browser του client, στους ενδιάμεσους proxy servers ή και στις caches του server.
CGI	Πρόκειται για το ακρωνύμιο του Common Gateway Interface, μιας μεθόδου για την εκτέλεση προγραμμάτων στο web server με βάση την είσοδο που λαμβάνει από κάποιο browser ή αλλιώς ενός προτύπου για το πέρασμα πληροφοριών από ένα browser σε ένα server. Τα CGI scripts επιτρέπουν στους επισκέπτες ενός site να «αλληλεπιδρούν» με τις σελίδες του αναζητώντας π.χ. ένα στοιχείο σε μια βάση δεδομένων, παρέχοντας σχόλια σχετικά με το υλικό στο site, ή συμπληρώνοντας μια φόρμα.
Classification	Κατηγοριοποίηση. Έχει σαν στόχο την εξαγωγή των κλάσεων των αντικειμένων με βάση προαποφασισμένες ετικέτες (class labels). Ανήκει στις προγνωστικές τεχνικές εξόρυξης δεδομένων (predictive

	web mining) και για να λειτουργήσει απαιτείται ένα σύνολο δεδομένων εκπαίδευσης (training data set) ώστε να αναγνωριστούν τα ιδιαίτερα χαρακτηριστικά κάθε κλάσης και στη συνέχεια να ταξινομηθούν τα νέα αντικείμενα με βάση τις τιμές τους στα χαρακτηριστικά αυτά. Οι πιο γνωστοί αλγόριθμοι που χρησιμοποιούνται είναι: η απλοϊκή κατηγοριοποίηση κατά Bayes (naive Bayes), ο K-πλησιέστερος γείτονας (K-nearest neighbor ή KNN), τα δέντρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), κλπ.
Click stream	Η πορεία που ακολουθεί ένας επισκέπτης ενώ πλοηγείται από site σε site, ή από σελίδα σε σελίδα εντός του ίδιου δικτυακού τόπου. Η πληροφορία αυτή είναι χρήσιμη στους σχεδιαστές των εκάστοτε δικτυακών τόπων, καθώς τους βοηθά να βελτιώσουν τη λειτουργικότητά τους με κριτήριο τον ίδιο το χρήστη.
Clustering	Συσταδοποίηση ή ομαδοποίηση. Η τεχνική αυτή στοχεύει στο διαχωρισμό των αντικείμενων σε ομάδες με βάση το ότι τα αντικείμενα που ανήκουν στην ίδια ομάδα είναι όσο το δυνατόν περισσότερο όμοια, ενώ τα αντικείμενα διαφορετικών ομάδων είναι όσο το δυνατόν πιο διαφορετικοί. Η ομοιότητα μετριέται χρησιμοποιώντας μια συνάρτηση ομοιότητας (similarity function). Αφού δημιουργηθούν οι ομάδες, είναι στη συνέχεια δυνατό να τους αποδοθούν ετικέτες ανάλογα με το είδος της κάθε ομάδας και να χρησιμοποιηθούν οι ετικέτες αυτές σαν βάση για την εφαρμογή κατηγοριοποίησης (με αυτό τον τρόπο χρησιμοποιούμε την ομαδοποίηση σαν ένα αρχικό βοηθητικό στάδιο πριν εφαρμόσουμε κατηγοριοποίηση, στις περιπτώσεις που δεν είναι εύκολο να προκαθορίσουμε τις κλάσεις). Αρκετοί αλγόριθμοι ομαδοποίησης είναι διαθέσιμοι: η ιεραρχική συσσωρευτική συσταδοποίηση (hierarchical agglomerative clustering ή HAC), ο k-means, οι αυτό-οργανωμένοι χάρτες χαρακτηριστικών (self-organizing maps ή SOMs) με πιο γνωστούς τους χάρτες Kohonen, κλπ.
Collaborative filtering (CF)	Συνεργατικό φιλτράρισμα. Βασίζεται στην «ομοιότητα» των χρηστών προκειμένου να παράγει προτάσεις.
Community	Συνεχώς μεταβαλλόμενη κοινότητα χρηστών που συνεργάζονται και ανταλλάσσουν ιδέες σε ένα ηλεκτρονικό δίκτυο (π.χ. στο Διαδίκτυο).
Consumer behavior	Η διαδικασία με την οποία οι καταναλωτές λαμβάνουν αποφάσεις για την αγορά αγαθών και υπηρεσιών.
Content-based filtering	Φιλτράρισμα με βάση το περιεχόμενο. Λαμβάνει υπόψη του τα χαρακτηριστικά του χρήστη και τις πληροφορίες των προϊόντων (μεταδεδομένα, λέξεις κλειδιά, κλπ.) προκειμένου να παράγει προτάσεις.
Cookie	Ένα μικρό αρχείο (συνήθως με τη μορφή .txt) το οποίο στέλνεται από ένα web site προς τους επισκέπτες του και αποθηκεύεται στο σκληρό τους δίσκο. Η αποστολή του cookie γίνεται, ώστε να υπάρχει ένα σημείο επαφής, για να αναγνωριστεί ο ίδιος χρήστης την επόμενη φορά που θα επισκεφτεί το ίδιο web site. Τα cookies παρέχουν πολύτιμες πληροφορίες στους ιδιοκτήτες των web sites, στους ερευνητές αγοράς για το χρήστη (π.χ. ένα cookie παρέχει πληροφορίες σχετικά με τη διάρκεια επίσκεψης στο web site καθώς και τις μετακινήσεις που έγιναν εντός του web site). Ο χρήστης μπορεί να αρνηθεί τα εισερχόμενα cookies αν θεωρεί ότι παρακολουθείται, απλώς ρυθμίζοντας κατάλληλα το web browser. Σε τέτοια περίπτωση όμως οι δικτυακοί τόποι δεν αναγνωρίζουν την επίσκεψη του ίδιου χρήστη τις επόμενες φορές και ίσως δεν μπορούν να παρέχουν κάποιες ειδικές υπηρεσίες (π.χ.

	personalization).
Data mining (DM)	Εξόρυξη δεδομένων. Ένα σύνολο αναλυτικών εφαρμογών που ψάχνουν για κρυμμένη πληροφορία σε μία βάση δεδομένων. Το DM είναι η διαδικασία αναζήτησης, μέσα σε μεγάλες ποσότητες δεδομένων προκειμένου να βρεθούν οι σχέσεις των δεδομένων. Είναι επίσης διαδικασία γνώση και ως Data Surfing. Εργαλεία τέτοιων λειτουργιών περιλαμβάνουν ποικιλία τεχνικών, όπως case-based reasoning, οπτικοποίηση δεδομένων (data visualization), ασαφή ανάλυση και νευρωνικά δίκτυα. Τα εργαλεία της τεχνικής «case-based reasoning» παρέχουν ένα μέσο εύρεσης εγγραφών παρόμοιων με μία εγγραφή αναφοράς. Αυτά τα εργαλεία επιτρέπουν στο χρήστη να ορίζει την «ομοιότητα» των ανακτηθέντων εγγραφών. Τα εργαλεία οπτικοποίησης επιτρέπουν στο χρήστη εύκολα και γρήγορα να δει με γραφικό τρόπο την πληροφορία από διαφορετικές όψεις.
Data Warehouse	Αποθήκη δεδομένων. Μια βάση δεδομένων φτιαγμένη για την υποστήριξη αποφάσεων. Ενημερώνεται τακτικά και είναι δομημένη, ώστε να υποστηρίζει γρήγορα καθημερινές ανάγκες για πληροφορία και διευθυντικές αναφορές. Οι αποθήκες δεδομένων εμπεριέχουν μεγάλη ποσότητα δεδομένων. Μία αποθήκη δεδομένων είναι μία ευμετάβλητη, αντικειμενοστραφής, ολοκληρωμένη συλλογή γνώσης προς υποστήριξη της λήψης μίας διοικητικής απόφασης.
DC	Dublin Core. Πρότυπο e-learning μεταδεδομένων.
Direct marketing	Διαδικασία κατά την οποία μια επιχείρηση αποστέλλει το διαφημιστικό μήνυμα απευθείας στους δυνητικούς της πελάτες. Οι πιο δημοφιλείς μορφές direct marketing είναι μέσω τηλεφώνου ή ταχυδρομείου, παραδοσιακού ή ηλεκτρονικού.
E-commerce (EC)	Ηλεκτρονικό εμπόριο. Πρόκειται για κάθε είδους εμπορικής συναλλαγής μεταξύ προσώπων (φυσικών και μη) που πραγματοποιείται με ηλεκτρονικά μέσα. Είναι η διάθεση και αγοραπωλησία προϊόντων, πληροφοριών και υπηρεσιών ηλεκτρονικά, καθώς και η διεκπεραίωση εμπορικών λειτουργιών και συναλλαγών χωρίς τη χρήση χαρτιού μέσα από δίκτυα ηλεκτρονικών υπολογιστών.
Electronic catalog (e-catalog)	Δυναμικός μηχανισμός που παρουσιάζει τα προς πώληση προϊόντα ή/και τις υπηρεσίες μιας επιχείρησης και επιτρέπει στους χρήστες την αγορά αγαθών ή υπηρεσιών σε ηλεκτρονικά καταστήματα.
Electronic shop (e-shop)	Ηλεκτρονικό κατάστημα. Ένα κανονικό κατάστημα πώλησης ειδών το οποίο λειτουργεί μέσω του Διαδικτύου με ολοκληρωμένη διαδικασία πώλησης, πληρωμή online, κλπ.
HTML	HyperText Markup Language. Γλώσσα σήμανσης υπερκειμένου που χρησιμοποιείται στη συγγραφή σελίδων για το Web. Δίνει τη δυνατότητα ορισμού γραμματοσειρών, καθορισμού μορφοποιήσεων, ενσωμάτωσης γραφικών και συνδέσεων υπερκειμένου. Οι ιστοσελίδες αποθηκεύονται σε αρχεία που συνήθως έχουν προέκταση .htm ή .html.
Hypermedia	Υπερμέσα. Συνδυασμός διάφορων τύπων μέσων όπως κείμενα, γραφικά και οπτικοακουστικό υλικό.
Hypertext	Υπερκείμενο. Μία προσέγγιση χειρισμού κειμένων και άλλων πληροφοριών που επιτρέπει στους χρήστες να μεταπηδούν από

	ένα δεδομένο θέμα, όποτε αυτοί θελήσουν, σε άλλα σχετικά θέματα. Μια τεχνική διαχείρισης γνώσης, όπου η γνώση απεικονίζεται σε συνδεδεμένα κείμενα και εξελίσσεται με τέτοιο τρόπο, ώστε να επιτρέπει σε ένα χρήστη να διαλέξει ένα link στη σελίδα που έχει επισκεφθεί, προκειμένου να έχει πρόσβαση σε μια άλλη σελίδα σχετική με το θέμα για το οποίο ενδιαφέρεται και στο οποίο οδηγεί το link.
IEEE/LOM	IEEE/LOM – Institute of Electrical and Electronic Engineers/Learning Object Metadata. Πρότυπο e-learning μεταδεδομένων.
IMS	Instructional Management System. Πρότυπο e-learning μεταδεδομένων.
Item-to-item correlation	Συσχέτιση αντικειμένου-με-αντικείμενο.
Keyword	Λέξη κλειδί η οποία χρησιμοποιείται για κάποιο σκοπό στο Internet. Παράδειγμα: keyword search (όταν κάποιος ψάχνει σε μία μηχανή αναζήτησης με λέξεις-κλειδιά), keyword targeting (όταν μία εκστρατεία τρέχει με κριτήρια στόχευσης τα keywords που γράφουν οι χρήστες).
Knowledge management (KM)	Διαχείριση γνώσης. Επιχειρηματική διαδικασία που μορφοποιεί τη διαχείριση και μεγιστοποιεί την αξία του γνωστικού κεφαλαίου μιας εταιρίας. Το KM ουσιαστικά προωθεί τη συνεργατική προσέγγιση και ολοκλήρωση όσον αφορά στη δημιουργία, σύλληψη, οργάνωση, πρόσβαση και χρήση του πληροφοριακού υλικού της επιχείρησης.
Log file αρχείο (ημερολόγιο)	Αρχείο στο οποίο πρωτοκολλούνται όλες οι κινήσεις που γίνονται σε ένα web site. Από την ανάλυση των log files, με τη χρήση ειδικών προγραμμάτων (log analyzers), προκύπτουν πληροφορίες, όπως οι συνολικοί επισκέπτες του web site, οι διευθύνσεις IP των επισκεπτών, ο τόπος προέλευσης, ο web browser τους, ο χρόνος παραμονής τους, η ημέρα και ώρα επίσκεψης, κλπ. Συνεπώς, τα log files αποτελούν τη βάση για την αξιολόγηση της επισκεψιμότητας και της συμπεριφοράς των χρηστών.
Logoff	Η αποσύνδεση/τερματισμός της σύνδεσης από κάποιον web server/web site.
Logon	Η διαδικασία σύνδεσης με έναν web server/web site.
Metadata	Μεταδεδομένα. Αφορούν δεδομένα σε μία αποθήκη δεδομένων. Παρέχουν ένα κατάλογο για να βοηθήσει το Decision Support System (DSS) να εντοπίσει τα περιεχόμενα της αποθήκης δεδομένων: είναι δηλαδή ένας οδηγός για τη χαρτογράφηση δεδομένων καθώς μεταβάλλεται από το λειτουργικό περιβάλλον στο περιβάλλον αποθήκης δεδομένων και χρησιμεύει σαν οδηγός στους αλγόριθμους που χρησιμοποιούνται για τη συγκέντρωση των πρόσφατων λεπτομερών δεδομένων. Τα metadata είναι σημαντική πληροφορία συνδεδεμένη με μια συγκεκριμένη μεταβλητή. Πρέπει να περιλαμβάνουν επιχειρηματικούς ορισμούς των δεδομένων και καθαρές, ακριβείς περιγραφές του είδους των δεδομένων, των ενδεχόμενων τιμών, κλπ. Τα metadata ορίζουν και περιγράφουν επιχειρησιακά δεδομένα. Παραδείγματα metadata περιλαμβάνουν περιγραφές στοιχείων των δεδομένων, περιγραφές του είδους των δεδομένων, περιγραφές χαρακτηριστικών/ιδιοτήτων, περιγραφές εύρους/κυριότητας και περιγραφές επεξεργασίας/μεθόδου. Το περιβάλλον της αποθήκης περιλαμβάνει όλους τους συλλογικούς πόρους metadata: κατάλογοι βάσεων δεδομένων, λεξικά

	δεδομένων και υπηρεσίες πλοήγησης. Επίσης, περιλαμβάνει στοιχεία όπως το όνομα, το μήκος, σωστές αξίες και περιγραφή ενός στοιχείου δεδομένων. Είναι αποθηκευμένα σε ένα λεξικό και σε μία αποθήκη δεδομένων.
OLAP	Online Analytical Processing. Λογισμικό για τη διαχείριση πολυδιάστατων δεδομένων από ποικιλία πηγών που έχουν αποθηκευτεί σε μια αποθήκη δεδομένων. Το λογισμικό μπορεί να δημιουργήσει διάφορες απόψεις και αναπαραστάσεις των δεδομένων. Το OLAP λογισμικό παρέχει γρήγορη, σταθερή και αλληλοεπιδρούσα πρόσβαση σε διαμοιρασμένα, πολυδιάστατα δεδομένα.
Ontology	Οντολογία. Είναι μια αυστηρή (formal) περιγραφή των αντικειμένων-πόρων και των σχέσεων μεταξύ τους. Οι οντολογίες δίνουν τη δυνατότητα σε μια κοινότητα χρηστών να έχουν κοινή ονοματολογία και κοινή αντίληψη για τα αντικείμενα-πόρους τα οποία δηλώνουν ή χρησιμοποιούν. Επίσης, τους επιτρέπουν να περιγράψουν τον «κόσμο» που χρησιμοποιούν (που τους αφορά) με ένα τέτοιο τρόπο, ώστε οι υπολογιστές να μπορούν να επεξεργαστούν και να «συνδέσουν» τα δεδομένα από διαφορετικούς κόσμους. Η δυσκολία στις οντολογίες έγκειται στο γεγονός ότι οι κοινότητες των χρηστών θα πρέπει να συμφωνήσουν στην οντολογική περιγραφή του πεδίου ενδιαφέροντός τους.
P3P	Platform for Privacy Preferences Project. Το πρότυπο αυτό δημιουργήθηκε από το World Wide Web Consortium με στόχο να λειτουργήσει σαν έναν απλό και αυτόματο μέσο ώστε οι χρήστες να αποκτήσουν μεγαλύτερο έλεγχο στα προσωπικά τους δεδομένα όταν επισκέπτονται διάφορα web sites. Στο βασικό του επίπεδο, το P3P περιλαμβάνει ένα σύνολο από συγκεκριμένες ερωτήσεις που καλύπτουν τα κύρια ζητήματα των πολιτικών ιδιωτικότητας των sites και παρουσιάζει τον τρόπο που αυτά χειρίζονται τις προσωπικές πληροφορίες των χρηστών. Το P3P διευκολύνει τα sites να κρατούν αυτά τα δεδομένα σε μια κατάλληλη μορφή ώστε οι browsers να τα διαβάζουν αυτόματα και να τα συγκρίνουν με το σύνολο των προτιμήσεων των χρηστών.
PAPI	Public and Private Information for Learners. Πρότυπο που περιγράφει τα αρχεία των εκπαιδευόμενων σε εφαρμογές e-learning.
Password	Μοναδική και απόρρητη λέξη κλειδί με την οποία ο χρήστης, σε συνδυασμό με το username, αποδεικνύει την ταυτότητά του όταν εισέρχεται σε περιορισμένης πρόσβασης σελίδες ή εφαρμογές.
Personalization	Εξατομίκευση του περιεχομένου μιας online υπηρεσίας βάσει του προφίλ του εκάστοτε χρήστη/πελάτη.
Privacy	Το δικαίωμα απομόνωσης και ελευθερίας ενός χρήστη από αδικαιολόγητες παρεμβάσεις και «εισβολές» στην ιδιωτική ζωή του.
Privacy policy	Δήλωση εμπιστευτικότητας. Η πολιτική μιας επιχείρησης που διαθέτει web site ως προς τη συλλογή και τη χρήση των προσωπικών δεδομένων των επισκεπτών/πελατών της.
Proxy server	Ένας server στον οποίο πραγματοποιείται ενδιάμεση αποθήκευση πληροφοριών του Web με σκοπό την εύκολη ανάκλησή τους. Για παράδειγμα, εκεί αποθηκεύονται εικόνες ή σελίδες που ζητάει ένας χρήστης από κάποιο web site έτσι ώστε να υπάρχουν πρόχειρες για τον επόμενο χρήστη που θα τις ξαναζητήσει. Το πλεονέκτημα του είναι η βελτιστοποίηση στην αναμετάδοση δεδομένων ενώ

	παράλληλα μειώνεται ο φόρτος διακίνησης στα δίκτυα. Το μειονέκτημα μπορεί να είναι ότι αν κάποιος χρήστης βρίσκεται πίσω από κάποιον proxy server (π.χ. σε εταιρίες) τότε πιθανόν δεν θα δει την τελευταία έκδοση του web site (ή κάποιας διαφήμισης) που επισκέπτεται αλλά αυτή που είχε αποθηκευτεί στον proxy server την τελευταία φορά που κάποιος άλλος επισκέφτηκε το ίδιο web site.
Pull marketing	Το pull-marketing βασίζεται στην ιδέα ότι οι χρήστες του Internet δείχνουν μεγάλο ενδιαφέρον για την αλληλεπίδραση. Συνεπώς, αυτό που έχει σημασία είναι η προσέλκυση χρηστών σε αντίστοιχους δικτυακούς τόπους και η συνεχής ανανέωση των περιεχομένων τους.
Push marketing	Με τις τεχνολογίες push που ουσιαστικά βασίζεται στο Internet broadcast, παρέχονται στην οθόνη του χρήστη έτοιμες ειδήσεις και πληροφορίες. Οι τεχνολογίες αυτές έφτασαν στο αποκορύφωμά τους στα μέσα της δεκαετίας του '90 (τεχνολογία point cast), αλλά πλέον δεν χρησιμοποιούνται σχεδόν καθόλου.
RDF	Resource Description Framework. Πρότυπο του Σημασιολογικού Ιστού που παρέχει ένα πλαίσιο διαχείρισης του δυναμικού, επιχειρησιακή ένταξη, καθώς και κοινή χρήση και επαναχρησιμοποίηση δεδομένων στον Παγκόσμιο Ιστό. Η μορφή αυτή επεκτείνεται σε εύρος εφαρμογών, επιχειρήσεις, καθώς και κοινότητες αφού όλες οι διαφορετικές κατηγορίες χρηστών μπορούν να μοιραστούν την ίδια πληροφορία, ακόμα και αν δε χρησιμοποιούν το ίδιο λογισμικό.
Recommendation system (RS)	Σύστημα παραγωγής προτάσεων.
Robot	Πρόγραμμα σε web sites/μηχανές αναζήτησης, το οποίο κάνει αναζήτηση σε κείμενα άλλων web sites ή συγκρίνει τις τιμές σε e-shops του Internet. Συνώνυμο του crawler και spider.
Search engine	Πρόγραμμα ή web site στο οποίο μπορεί κάποιος να αναζητήσει πληροφορίες από άλλους δικτυακούς τόπους, χρησιμοποιώντας keywords. Σημαντικές μηχανές αναζήτησης είναι, μεταξύ άλλων, οι Altavista και Lycos, κλπ.
Semantic usage mining	Σημασιολογική εξόρυξη γνώσης από δεδομένα χρήσης Παγκοσμίου Ιστού.
Semantic Web	Σημασιολογικός Παγκόσμιος Ιστός Πληροφοριών.
Server	Κεντρικός υπολογιστής ο οποίος εξυπηρετεί άλλους υπολογιστές (clients) και προμηθεύει στους χρήστες αυτών των υπολογιστών με το υλικό που του ζητήθηκε το οποίο είναι αποθηκευμένο στον σκληρό του δίσκο. Ένας web server π.χ. παρέχει σε έναν χρήστη τη δυνατότητα, να δει τις αποθηκευμένες πληροφορίες που υπάρχουν μέσα στο δίσκο του, υπό την μορφή HTML.
Session	Μία συνολική επιτυχημένη επίσκεψη ενός χρήστη (ουσιαστικά ενός υπολογιστή) σε ένα web site. Το session περιγράφει όλη τη διαδρομή του επισκέπτη από την πρώτη σελίδα έως τις υπόλοιπες, για όση ώρα αυτός βρισκόταν στο web site.
Shopping cart	Ένα εικονικό καλάθι αγορών, που επιτρέπει στους καταναλωτές να συγκεντρώσουν τεμάχια, «σερφάροντας» σε ένα web site (δικτυακό τόπο) ηλεκτρονικών πωλήσεων, και να τα «κρατήσουν» έως ότου είναι έτοιμοι να τα αγοράσουν.

Unicode	Προέκταση του ASCII που προσπαθεί να συμπεριλάβει τους χαρακτήρες όλων των ζωντανών γλωσσών.
URI	Uniform Resource Identifier ή Καθολικό Αναγνωριστικό Πόρων. Το URI είναι μια απλή ακολουθία κειμένου, αναφέρεται σε πόρους του Internet π.χ. έγγραφα, πόρους, ανθρώπους και έμμεσα σε οτιδήποτε και αποτελεί κάτι γενικότερο από το γνωστό μας URL του Διαδικτύου. Δεν μας λέει κατ' ανάγκη πως θα προσπελάσουμε κάτι: ακόμη περισσότερο, δεν δηλώνει απαραίτητα κάτι προσπελάσιμο στο διαδίκτυο. Είναι απλώς ένα παγκόσμιο, μοναδικό και μόνιμο αναγνωριστικό-ταυτότητα του πόρου που περιγράφει.
URL	Uniform Resource Locator. Η διεύθυνση που συνολικά καθορίζει ένα web site. Όπως σε μια ατζέντα διευθύνσεων προσδιορίζεται ακριβώς η θέση του κωδικού πόλης και του αριθμού, έτσι και στο URL ορίζεται η ακριβής δομή των στοιχείων μιας διεύθυνσης. Αρχικά αναφέρεται η μέθοδος (πρωτόκολλο) μετάδοσης των δεδομένων, όπως http ή ftp, η οποία ακολουθείται από διπλή κάθετο. Στη συνέχεια ακολουθεί η διεύθυνση του web server (domain name), π.χ. http://www.uom.gr και ενδεχομένως ο προσδιορισμός της τοποθεσίας μιας συγκεκριμένης σελίδας στο web site πχ http://www.uom.gr/news.html. Το URL μπορεί να πληκτρολογηθεί στο web browser με ή χωρίς το http://.
URN	Uniform Resource Name ή ομοιόμορφο όνομα πόρου. Ένα όνομα ιστοσελίδας που δεν αλλάζει όταν η σελίδα μεταφερθεί σε διαφορετικό υπολογιστή. Προτείνεται ως λύση στο πρόβλημα της κομμένης σύνδεσης (broken - link).
User identification	Πιστοποίηση/έλεγχος της ταυτότητας ή του δικαιώματος πρόσβασης, π.χ. μέσω ονόματος ή κωδικού πρόσβασης.
User tracking	Όρος που περιγράφει γενικά τη διαδικασία παρακολούθησης/αναγνώρισης ενός χρήστη και καταγραφής της συμπεριφοράς του κατά την πλοήγησή του στο Web. Tracking γίνεται είτε για την κατανόηση των αναγκών του χρήστη είτε για την προβολή διαφημιστικών banner στην οθόνη του ανάλογα με τις προτιμήσεις του. Αυτό επιτυγχάνεται με τη βοήθεια των cookies.
User-to-user correlation	Συσχέτιση χρήστη-με-χρήστη.
W3C (World Wide Web Consortium)	Μία κοινοπραξία /ένωση συμφερόντων που ιδρύθηκε το 1994 από διάφορες εταιρίες, στενά συνδεδεμένες με το Internet, με στόχο την άσκηση επιρροής στην εξέλιξη του World Wide Web. Στόχος είναι η δημιουργία ανοιχτών (μη πατενταρισμένων προτύπων), π.χ. HTTP και HTML.
Web usage mining	Εξόρυξη γνώσης από δεδομένα χρήσης Παγκοσμίου Ιστού.
XML (Extensible Markup Language)	Συμπλήρωση της HTML και πιθανότατα διάδοχός της. Επιτρέπει το διαχωρισμό προγραμματισμού για τη διαχείριση του περιεχομένου μία σελίδας από τον προγραμματισμό για τη μορφοποίηση και τη διάταξη.

Ευρετήριο όρων

A

Absolute error, 138
Accuracy, 138
Adaptive web-based educational systems, 38
AE. *See* Absolute error
Apriori, 66, 70
Association rules, 53, 57, 66, 131, 133, 158
Attribute-based technologies, 101
Axioms, 45

B

B2C. *See* Business-to-consumer
Banana problem, 124
Barchart, 191
BI. *See* Business intelligence
Business intelligence, 153
Business-to-consumer, 93

C

Classes, 45
Classification, 56
Clickstream, 59
Clustering, 56, 156, 180
Cold start problem, 124
Collaborative filtering, 95, 103, 107
    Memory-based approaches, 107
    Model-based approaches, 109
Community transactions history, 98
Confidence, 68
Content pages, 48
Content-based filtering, 95, 103, 104
Contingency table, 167
Cookies, 54
Correspondence analysis, 156, 174
Coverage, 141
Current session window, 59

D

Data definition types, 43
Data mining, 154
Database manager module, 193
Digital signature, 43
Disambiguation, 166
DTDs. *See* Data definition types

E

E-commerce. *See* Electronic commerce
E-learning, 20, 37, 38, 50
E-learning metadata, 50

ADL/SCORM, 51
AICC, 51
ARIADNE, 51
DC, 51
IEEE/LOM, 51
IMS, 51
Electronic commerce, 93
ESPACE ACCESS, 161, 164
Explicit navigation, 98, 130

F

Frequent itemsets, 69

G

Garbage in - garbage out, 156
GIGO. *See* Garbage in - garbage out

H

Head pages, 48
Histogram, 190
Hybrid approaches, 95, 103, 110
Hyperlink, 38

I

Identification, 166
Implicit navigation, 97, 130
Information overload problem, 94
Innovation, 159
Instances, 45
International Patents Classification, 161
IPC. *See* International Patents Classification
Item attributes, 98
Item popularity, 98
Item-to-item correlation, 101

K

KDD. *See* Knowledge Discovery in Databases
K-means clustering, 183
Knowledge Discovery in Databases, 154

L

Lemmatization, 166
Lexical table, 167
Linguistic pre-processing, 166
Log files, 54
Logic, 43
Look-up pages, 48

M

MAE. *See* Mean absolute error
 Manual selection, 101
 Mean absolute error, 138
 Metadata, 50
 MIMOSA, 164
 Model, 157

N

Navigation pages, 48
 New item problem, 60

O

Ontology, 42, 44

P

Page viewing time, 67
 Pageview, 59, 67
 Part-of-speech tagger, 169
 Part-of-speech-tagging, 156, 166
 Passive τεχνολογίες, 102
 Patent. *See* Πατέντα
 Patent Office, 161
 Pattern, 154, 157
 Personal pages, 48
 Personalization, 94
 Personalization degree, 103
     Ephemeral personalization, 103
     Non personalization, 103
     Persistent personalization, 103
 Personalized e-learning, 58
 Pie chart, 191
 POS-tagger. *See* Part-of-speech tagger
 Prediction, 99
 Privacy, 81
 Profit function, 122
 Proof, 43
 Proximity, 177
 Pull τεχνολογίες, 102
 Push τεχνολογίες, 102

R

Rating, 98, 99
 Raw retrieval, 100
 RDF, 42, 43
 RDFS. *See* RDFSschema
 RDFSschema, 44
 Recommendations systems, 94
 Relations, 45
 Results presentation module, 194
 Review, 99

S

Segmentation, 166
 Semantic annotation, 44
 Semantic Web, 40, 41, 42
 Semantic Web Mining, 40, 49

Sequential Pattern Discovery, 57
 Session, 59, 67
 Similarity function, 56
 Simple analysis, 156, 190
 Statistical analysis, 156
 Statistical analysis module, 193
 Statistical summaries, 101
 Stemming, 156
 Stop-list, 169
 Suggestion, 98
 Support, 68

T

Technology Watch, 159
 Text comments, 98
 Text mining, 154
 TM. *See* Text mining
 Tokenization, 168
 Transactions history, 98
 Trust, 43

U

Unicode, 42
 United States Patent Office Classification, 161
 URI, 42, 58
 User communities, 94
 User modeling, 53
 User profile, 38, 54, 59
 User-to-user correlation, 102
 USPOC. *See* United States Patent Office Classification

W

W3C, 40, 41
 Web Mining, 47
     Web Content Mining, 47
     Web Structure Mining, 47, 48
     Web Usage Mining, 47, 48
 Web Personalization, 52

X

XML, 42, 43
 XMLSchema, 43

A

Ανακάλυψη ακολουθιακών μοτίβων, 57

Γ

Γλωσσολογική επεξεργασία, 166

Δ

Δείκτης τεχνολογίας και καινοτομίας, 194
 Διάγραμμα πίτας, 191
 Δίπλωμα ευρεσιτεχνίας, 155

Ε

Εμπιστοσύνη, 68
 Εξατομικευμένη ηλεκτρονική μάθηση, 58
 Εξατομίκευση, 94
 Εξόρυξη γνώσης από κείμενα, 154
 Επιχειρηματική Ευφυΐα, 153

Η

Ηλεκτρονική μάθηση, 37
 Ηλεκτρονικό εμπόριο, 93

Ι

Ιεραρχικό clustering, 182
 Ιστόγραμμα, 190

Κ

Κανόνες συσχέτισης, 57
 Κατηγοριοποίηση, 56
 Κοινότητες χρηστών, 94

Λ

Λεξιλογικός πίνακας, 167

Μ

Μοντέλο, 157
 Μοντέλο χτηστών, 53

Ο

OBI. *See* Οργανισμός Βιομηχανικής Ιδιοκτησίας
 Ομαδοποίηση, 56
 Οντολογία, 42, 43, 44
 Οργανισμός Βιομηχανικής Ιδιοκτησίας, 160

Π

Πατέντα, 155, 160
 Πίνακας γειτνίασης, 167, 176
 Προσαρμοστικά εκπαιδευτικά συστήματα που
   βασίζονται στο Web, 38
 Πρότυπο, 154, 157
 Προφίλ χρήστη, 38, 54, 59

Ρ

Ραβδόγραμμα, 191

Σ

Σαφής πλοήγηση, 98
 Σημασιολογικός Παγκόσμιος Ιστός Πληροφοριών, 40
 Σημασιολογικός σχολιασμός, 44
 Σύναρτηση κέρδους, 122
 Σύνοδος, 59
 Συσταδοποίηση, 56
 Συστήματα συστάσεων, 94

Τ

Ταξινόμηση οντολογιών, 45
   Μετα-οντολογίες, 46
   Οντολογίες αναπαράστασης γνώσης, 46
   Οντολογίες επικοινωνίας (λέω & ρωτάω) για
     διαμοιρασμό της γνώσης, 45
   Οντολογίες ευρετηρίου για την ανάκτηση
     περιπτώσεων, 45
   Οντολογίες περιεχομένου για επαναχρησιμοποίηση
     γνώσης, 45
 Τεχνολογίες Πληροφορίας και Επικοινωνιών, 37
 ΤΠΕ. *See* Τεχνολογίες Πληροφορίας και Επικοινωνιών

Υ

Υπερσύνδεσμος, 38
 Υπονοούμενη πλοήγηση, 97
 Υποστήριξη, 68