Πανεπιστήμιο Πατρών

Τμήμα Μαθηματικών και

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Διατμηματικό Μεταπτυχιακό Πρόγραμμα Ειδίκευσης

«Μαθηματικά των Υπολογιστών και των Αποφάσεων»

Διπλωματική Εργασία

Της Δέσποινας Α. Πλώτα

Τεχνικές Text Mining για την συγκριτική ανάλυση νοήματος κειμένου

Επιβλέπων Καθηγητής:

Βασίλειος Βουτσινάς

Πάτρα, Σεπτέμβριος 2010

Περίληψη

Τις τελευταίες δεκαετίες έχουν παραχθεί ασύλληπτα μεγάλες ποσότητες δεδομένων από διάφορες διεργασίες που έχουν οργανωθεί με χρήση υπολογιστικών συστημάτων.

Όλες αυτές οι τεράστιες βάσεις δεδομένων, που είτε έχουν να κάνουν με συστήματα δοσοληψιών (τραπεζικές συναλλαγές, αγορές με πιστωτικές κάρτες), είτε με διαδικασίες που συσσωρεύουν μεγάλα ψηφιακά αρχεία, είτε με επιχειρήσεις πώλησης προϊόντων που αποθηκεύουν καθημερινά τεράστιο όγκο δεδομένων (τα οποία προκύπτουν από τις συναλλαγές των διαφόρων πελατών τους), χρησιμοποιούνται για στατιστική επεξεργασία με την μορφή αθροιστικών συναρτήσεων που εφαρμόζονται σε πολλές διαστάσεις καθώς και σε σχεσιακού τύπου ερωτήσεις με στόχο την επιλογή μικρού συνόλου υποσυνόλου δεδομένων που ικανοποιούν συγκεκριμένα κριτήρια.

Στια παραπάνω περιπτώσεις ο χρήστης έχει γνώση του στόχου των ερωτήσεων και γνωρίζει την δομή και την σημασία των δεδομένων και των τιμών που εμφανίζονται.

Είναι πολύ πιθανόν όμως στις τεράστιες αυτές βάσεις δεδομένων να υπάρχει «κρυμμένη γνώση» που δεν είναι εκ των προτέρων γνωστή αλλά μπορεί να είναι χρήσιμη στους χρήστες. Αυτό που λείπει είναι η « επιβλεπόμενη ανάκτηση γνώσης», με άλλα λόγια η εφαρμογή αλγορίθμου στα δεδομένα μας για την ανακάλυψη και εξαγωγή της κρυμμένης γνώσης.

Ο χώρος της εξόρυξης δεδομένων καλύπτει αυτή ακριβώς της απαίτηση της περαιτέρω επεξεργασίας αυτών των αποθηκών δεδομένων.

Λέγοντας λοιπόν εξόρυξη δεδομένων - data mining, εννοούμε την διαδικασία ψαξίματος σε μεγάλο όγκο δεδομένων, με σκοπό την περισυλλογή πληροφορίας. Το data mining έχει περιγραφεί σαν: «η μη τετριμμένη εξαγωγή υπονοούμενης, ενδεχομένως χρήσιμης και μέχρι στιγμής άγνωστης πληροφορίας από διάφορα δεδομένα».

Το μεγαλύτερο βέβαια ποσό των δεδομένων βρίσκεται σε μορφή κειμένων και αυτός ο τύπος των μη δομημένων στοιχείων στερείται συνήθως «τα στοιχεία για τα στοιχεία». Η ανάγκη λοιπόν για την αυτοματοποιημένη εξαγωγή χρήσιμης γνώσης από τεράστια ποσά κειμενικών στοιχείων προκειμένου να βοηθηθεί η ανθρώπινη ανάλυση είναι προφανής.

Η εξόρυξη κειμένου (text mining) είναι ένας νέος ερευνητικός τομέας που προσπαθεί να επιλύσει το πρόβλημα της υπερφόρτωσης πληροφοριών με την χρησιμοποίηση των τεχνικών από την εξόρυξη από δεδομένα (data mining), την μηχανική μάθηση (machine learning), την επεξεργασία φυσικής γλώσσας (natural language processing), την ανάκτηση πληροφορίας (information retrieval), την εξαγωγή πληροφορίας (information extraction) και τη διαχείριση γνώσης (Knowledge management).

Βασιζόμενοι λοιπόν σε αυτήν την τεχνική εξόρυξης κειμένου παρουσιάζουμε σε αυτή την διπλωματική εργασία μια μεθοδολογία εξαγωγής γνώσης από κείμενο με απώτερο σκοπό την απόδοση της πατρότητας δυο έργων σε συγκεκριμένο συγγραφέα.

Το κύριο θέμα ενδιαφέροντος είναι το εξής: είναι η Ιλιάδα και Οδύσσεια έργα του ίδιου ποιητή;

Κατά την αρχαιότητα οι διάφοροι λόγιοι, σχεδόν ομόφωνα, διατηρούσαν την εξής Ουνιταριστική άποψη: και τα δυο έπη ήταν έργα του Ομήρου. Όμως στην συνέχεια, πολλοί κριτικοί αμφισβήτησαν αυτή την άποψη, αλλά δεν έτυχαν προσοχής. Συγκεκριμένα, αυτοί οι οποίοι, κατά το τέλος του δέκατου ένατου αιώνα, συμφωνούσαν ότι η Ιλιάδα και η Οδύσσεια δεν ήταν έργα του ίδιου ποιητή, στήριζαν την άποψή τους με τρία κύρια επιχειρήματα.

Πρώτον, ότι υπάρχει πλήθος από ασυνέπειες και στα δυο ποιήματα. Για παράδειγμα, κάποια πρόσωπα σκοτώνονται και εμφανίζονται ζωντανά αργότερα πάλι.

Δεύτερον, υπάρχουν λογοτεχνικές διαφορές. Για παράδειγμα η Οδύσσεια είναι λιγότερο ζωντανό έπος, ενώ η αντιμετώπιση των θνητών από τους θεούς είναι πολύ διαφορετική ανάμεσα στα δύο έπη.

Τρίτον, υπάρχουν γλωσσολογικές διαφορές ανάμεσά τους.

Αναπόφευκτα, μία πληθώρα αναλύσεων με χρήση Η/Υ έλαβε χώρα, έτσι ώστε να μπορέσει να αποσαφηνισθεί αν υπάρχουν αξιοσημείωτες διαφορές ανάμεσα στα δύο ποιήματα. Για παράδειγμα, πρόσφατα έγινε μια ανάλυση η οποία βασίστηκε στην εξέταση των συχνών και μη συχνών λέξεων στα δυο ποιήματα. Επιπλέον, η ίδια ανάλυση έγινε και σε έργα του Σοφοκλή, ώστε να υπάρχει ένα σημείο σύγκρισης. Ένα πρόγραμμα Η/Υ χρησιμοποιήθηκε για πινακοποίηση και για στατιστική ανάλυση του κειμένου και σαν αποτέλεσμα υπολογίστηκαν τρεις ομάδες λέξεων: πολύ συχνές, συχνές και μη συχνές λέξεις. Στη συνέχεια έγιναν δυο ανεξάρτητες αναλύσεις σε αυτές τις λέξεις (Ιλιάδα-Οδύσσεια, έργα του Σοφοκλή) έτσι ώστε να ορισθεί η ειδοποιός διαφορά στη χρήση των λέξεων ανάμεσα σε διαφορετικούς ποιητές. Η διακριτική ανάλυση βρήκε τον άξονα, κατά

μήκος του οποίου τα δυο έπη βρίσκονται σε μέγιστη απόσταση. Μετά τα τεστ που έγιναν ώστε να συγκριθούν τα δύο έργα σε κάθε λέξη, δεν βρέθηκαν σημαντικές ενδείξεις ότι ο 'Ομηρος έγραψε και τα δύο.

Αντιμετωπίσαμε το ίδιο ερώτημα χρησιμοποιώντας την τεχνική «εξόρυξης από δεδομένα», η οποία αποτελεί μια αναδυόμενη περιοχή έρευνας η οποία αναπτύσσει τεχνικές για ανεύρεση γνώσης σε μεγάλους όγκους δεδομένων. Μια κοινή χρήση της εξόρυξης από δεδομένα είναι η ανεύρεση προτύπων σε δεδομένα με την μορφή συσχετίσεων ανάμεσα σε έννοιες. Η μεθοδολογία μας βασίζεται στην ανάλυση του «σημαινόμενου» παρά του «σημαίνοντος» στην Ιλιάδα και στην Οδύσσεια.

Σε μία πρώτη φάση μετασχηματίζουμε τα δεδομένα: διατηρήθηκαν μόνο τα ουσιαστικά, τα ρήματα, τα επίθετα και τα επιρρήματα τα οποία οργανώθηκαν σε ομάδες συνωνύμων, όπου κάθε ομάδα αντιπροσωπεύει μία έννοια. Επιλέξαμε να κάνουμε ανάλυση των σχέσεων μεταξύ αυτών των εννοιών. 'Ετσι μετατρέψαμε όλες τις προτάσεις στο κείμενο, σε προτάσεις οι οποίες αποτελούνται μόνο από αυτές τις έννοιες, απαλείφοντας φυσικά τα διπλότυπα.

Στη συνέχεια μετασχηματίσαμε το κείμενο σε μια δομημένη μορφή, ώστε να μπορέσουμε να το αποθηκεύσουμε σε «εγγραφές» μιας βάσης δεδομένων. Συγκεκριμένα, θεωρήσαμε συνεχή τμήματα κειμένου σαν τέτοιες «εγγραφές». Πειραματιστήκαμε ορίζοντας είτε μία πρόταση είτε δύο συνεχόμενες ως «εγγραφή», χρησιμοποιώντας τον Apriori αλγόριθμο για να εξάγουμε «κανόνες συσχέτισης» της μορφής «90% των εγγραφών που περιέχουν την έννοια χ περιέχουν και την έννοια y». Εξάγαμε ένα μεγάλο αριθμό ισχυρών συσχετίσεων μεταξύ ίδιων εννοιών και στα δυο ποιήματα (π.χ. «γη»-«άνδρας»). Υπάρχουν επίσης συσχετίσεις μεταξύ διαφορετικών εννοιών (π.χ. «μάχη»-«άνδρας» μόνο στην Ιλιάδα) και διαφορετικές συσχετίσεις για την ίδια έννοια (π.χ. «ήρωας»-«μάχη» στην Ιλιάδα και «ήρωας»-«κατοικία» στην Οδύσσεια). 'Ομως, δεν βρήκαμε καμία αντίθεση. Αυτά τα αποτελέσματα ενδεχομένως να οδηγούν στο συμπέρασμα ότι ο 'Ομηρος έγραψε και τα δυο έπη.

Κεφάλαιο 1ο

1.0 Εισαγωγή

Είναι βέβαιο ότι ζούμε στην κοινωνία της πληροφορίας, όπου η μετατροπή των δεδομένων σε πληροφορία απαιτείται να οδηγεί στη μετατροπή της πληροφορίας σε γνώση. Μια από τις πιο προκλητικές εργασίες της εποχής μας είναι η ανακάλυψη προτύπων, τάσεων και ανωμαλιών σε τεράστια σύνολα δεδομένων, καθώς και η σύνοψή τους μέσω απλών και εύχρηστων μοντέλων. Τα προβλήματα της εξόρυξης δεδομένων ως τεχνικές έχουν προσεγγιστεί από ετερόκλητα επιστημονικά πεδία όπως της στατιστικής, της μηχανικής εκμάθησης, της θεωρίας της πληροφορίας και των υπολογιστικών διαδικασιών, έχει δημιουργήσει μια νέα επιστήμη με δυναμικά εργαλεία, η οποία καλείται «Εξόρυξη Δεδομένων» και είναι μέρος της διαδικασίας «Ανακάλυψης Γνώσης από Βάσεις Δεδομένων».

Η σύγκλιση της προόδου υπολογιστικών συστημάτων και της εξέλιξης στην επικοινωνία έχει οδηγήσει στην δημιουργία μιας κοινωνίας ικανής να παρέχει διαρκώς νέες πληροφορίες. Το υλικό που συγκεντρώνεται καταγράφεται διαρκώς, με αποτέλεσμα τη δημιουργία τεράστιων βάσεων δεδομένων. Το ζήτημα λοιπόν που προκύπτει, είναι εάν μπορούμε να διαχειριστούμε αυτές τις βάσεις δεδομένων.

Όλα αυτά τα θέματα προκάλεσαν το ενδιαφέρον και οδήγησαν στη διαδικασία της **Εξόρυξης Δεδομένων** (*Data Mining*). Πρόκειται για μία σειρά από τεχνικές που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς και ετερόκλητους κλάδους όπως οι: οικονομία, βιοστατιστική, δημογραφία, μετεωρολογία και γεωλογία. Υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων (ΕΔ).

Ωστόσο, ο ακόλουθος ορισμός [2], θεωρείται αξιόλογος:
«**Εξόρυξη Δεδομένων** είναι η ανάλυση – συνήθως τεράστιων – παρατηρούμενων συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων». Η δήλωση των σχέσεων και η σύνοψη των στοιχείων στην οποία αναφέρεται ο ορισμός αυτός, συχνά αναφέρεται ως **μοντέλο** ή **πρότυπο**.

Οι δυο βασικοί στόχοι της εξόρυξης δεδομένων είναι η εφαρμογή τεχνικών περιγραφής και πρόβλεψης σε μεγάλα σύνολα δεδομένων.

Η πρόβλεψη στοχεύει στον υπολογισμό της μελλοντικής αξίας ή στην πρόβλεψη της συμπεριφοράς κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον και οι οποίες βασίζονται στην συμπεριφορά άλλων μεταβλητών.

Η περιγραφή επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με ένα κατανοητό και αξιοποιήσιμο τρόπο.

Η σημαντικότητα της πρόβλεψης και της περιγραφής διαφέρει ανάλογα με τις εφαρμογές εξόρυξης δεδομένων. Ωστόσο ως προς την εξόρυξη γνώσης η περιγραφή τείνει να είναι περισσότερο σημαντική από την πρόβλεψη σε αντίθεση με την αναγνώριση προτύπων και την εφαρμογή μηχανικής μάθησης για τις οποίες η πρόβλεψη είναι πιο σημαντική.

Ένας αριθμός μεθόδων εξόρυξης δεδομένων έχουν προταθεί για να ικανοποιούν τις απαιτήσεις διαφορετικών εφαρμογών. Ωστόσο όλες επιτυγχάνουν μια ομάδα από διεργασίες εξόρυξης δεδομένων για να προσδιορίσουν και να περιγράψουν ενδιαφέροντα πρότυπα γνώσης που έχουν αντληθεί από ένα σύνολο δεδομένων. Θα αναφερθούμε στη συνέχεια της διπλωματικής σε αυτές τις μεθόδους αναλυτικά.

1.1 Data Mining

1.1.1 Ορισμοί

Επιστημονικός Ορισμός

Το **Data Mining** στη βιβλιογραφία έχει τον εξής ορισμό: "Η σύνθετη διαδικασία εξαγωγής συγκεκριμένης, προηγουμένως άγνωστης και δυνητικά ωφέλιμης, γνώσης από δεδομένα". [1]

Εναλλακτικά, συναντάται και ως "η επιστήμη της εξόρυξης χρήσιμης πληροφορίας από σύνολα ή βάσεις δεδομένων μεγάλου μεγέθους" [2]

Αναφορικά με τη διαχείριση επιχειρηματικών πόρων (ERP), το **Data Mining** θεωρείται ως η στατιστική και λογική ανάλυση εκτεταμένων συνόλων από δεδομένα συναλλαγών και εργασιών για τον εντοπισμό επαναλαμβανόμενων μοτίβων ή τάσεων που μπορούν να βοηθήσουν στη λήψη αποφάσεων.[3]

Όταν λέμε data sets εννοούμε τα πολύ μεγάλα σύνολα δεδομένων.

Σαν εξόρυξη δεδομένων θα μπορούσαμε να δώσουμε και τους ακόλουθους ορισμούς:

(1) η διαδικασία ανακάλυψης (discovery) προτύπων (patterns) που πριν δεν ήταν γνωστά, ισχύουν, είναι πιθανών χρήσιμα και είναι κατανοητά.

(2) η ανάλυση τους για να βρούμε μη αναμενόμενες σχέσεις ανάμεσα στα δεδομένα καθώς και να τα συνοψίσουμε με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στους χρήστες.

Τέτοια παραδείγματα είναι: αγορές από πολυκαταστήματα, προσπελάσεις ιστοσελίδων, πακέτα στο δίκτυο, αποτελέσματα επιστημονικών πειραμάτων, κίνηση μετοχών, βιολογικά δεδομένα κλπ

Η ανάγκη για την εφαρμογή αυτής της διαδικασίας προήλθε από το γεγονός ότι έπρεπε να αντιμετωπιστούν τα παρακάτω προβλήματα!
1. Το τεράστιο μέγεθος των δεδομένων
2. Ο μεγάλος αριθμός διαστάσεων
3. Η μη ομοιογενής και κατανεμημένη φύση των δεδομένων.

1.1.2 Χρήσεις Data Mining

Ο λόγος που χρησιμοποιούμε την Εξόρυξη Δεδομένων [19] είναι για να αναλύουμε βάσεις δεδομένων και να υποβοηθούμε στη λήψη αποφάσεων:

i. Ανάλυση αγοράς και διαχείριση:

Target marketing
Customer relation Management
Market basket analysis (supermarket)
Cross selling
Market segmentation

ii. Ανάλυση εταιρειών και διαχείριση ρίσκου:

Προβλέψεις
Διατήρηση πελατολογίου
Βελτιωμένη χρηματοδότηση (π.χ. τράπεζες)
Έλεγχος ποιότητας
Ανάλυση ανταγωνιστικότητας

iii. Εντοπισμός απάτης και διαχείριση:

Άλλες εφαρμογές που χρησιμοποιούν Εξόρυξη Δεδομένων:
Εξόρυξη κειμένου (newsgroup. Email, documents) and web analysis
Ευφυής απαντήσεις σε ερωτήματα

Πιο συγκεκριμένα για κάθε μια από τις παραπάνω περιπτώσεις ισχύει:

Ανάλυση αγοράς και διαχείριση.

π.χ. Η περίπτωση "Diapers and beer". Η παρατήρηση ότι πελάτες που αγοράζουν πάνες αγοράζουν και μπύρα επιτρέπουν στα καταστήματα να τοποθετούν αυτά τα είδη σχετικά κοντά, γνωρίζοντας ότι οι πελάτες θα κάνουν τη διαδρομή μεταξύ των ραφιών με τις πάνες και αυτών με τις μπύρες. Τοποθετώντας ανάμεσά τους και πατατάκια αυξάνουν τις πωλήσεις και στα τρία είδη.

Ανάλυση εταιρειών και διαχείριση ρίσκου.

π.χ. Κατασκευή δένδρων αποφάσεων από ιστορικά στοιχεία τραπεζικών δανείων για την παραγωγή αλγορίθμων, ώστε να αποφασίζεται αν πρέπει ή όχι να δοθεί ένα δάνειο σε έναν υποψήφιο πελάτη.

Εντοπισμός απάτης και διαχείριση ρίσκου.

π.χ. Άτομα. που σκηνοθετούν ατυχήματα για να εισπράξουν από τις ασφαλιστικές εταιρίες, ή κάποιοι που κάνουν ξέπλυμα «βρώμικου χρήματος» εντοπίζοντας ύποπτες μεταφορές χρημάτων ή κάποιοι που κλέβουν τους παρόχους τηλεπικοινωνιών και κάνουν τηλεφωνήματα που έχουν κάποια επαναλαμβανόμενα σχέδια είτε προς μια κλειστή ομάδα ατόμων (κινητά) είτε κάποια συγκεκριμένη ώρα της ημέρας κλπ.

π.χ. Εντοπισμός ακατάλληλων ιατρικών μεθόδων και θεραπειών.

Τα συστήματα Εξόρυξης δεδομένων είναι φτιαγμένα να διαχειρίζονται τεράστια πληροφορία, να μπορούν να έχουν και να ανατρέχουν σε ιστορικά δεδομένα, να χειρίζονται από υψηλόβαθμα στελέχη εταιριών έτσι ώστε σε μικρό χρονικό διάστημα να έχουν οπτική (διαγραμματική) αναπαράσταση πληροφοριών και επαναλαμβανόμενων Προτύπων έτσι ώστε να τους υποβοηθήσουν να πάρουν αποφάσεις.

Όταν χρησιμοποιούμε την έκφραση εξόρυξη δεδομένων- data mining, εννοούμε τις αποδοτικές τεχνικές που χρησιμοποιούνται για την ανάλυση πολύ μεγάλων συλλογών από δεδομένα και την εξαγωγή χρήσιμων πληροφοριών από αυτά. Αυτό συνήθως συμβαίνει γιατί συχνά υπάρχει πληροφορία «κρυμμένη» στα δεδομένα που δεν είναι προφανής και οι ανθρώπινοι αναλυτές μπορεί να χρειάζονται εβδομάδες για να ανακαλύψουν χρήσιμη πληροφορία. Πολλές είναι οι περιπτώσεις όπου πολλά δεδομένα δεν αναλύονται ποτέ.

Αναλύοντας την εμπορική πλευρά του θέματος καταλήγουμε ότι είναι απαραίτητη η εξόρυξη των δεδομένων αφού πολλά δεδομένα συγκεντρώνονται και εισάγονται σε αποθήκες δεδομένων όπως:

- Web δεδομένα, e-εμπόριο
- Αγορές σε πολυκαταστήματα/αλυσίδες
- Συναλλαγές με τράπεζες / πιστωτικές κάρτες

Και είναι δύσκολη η αντιμετώπιση τόσο μεγάλων βάσεων δεδομένων.

Επιπρόσθετα οι υπολογιστές γίνονται φτηνότεροι και πιο ισχυροί και υπάρχει μεγαλύτερος ανταγωνισμός, παροχή καλύτερων, προσωπικών υπηρεσιών σε κάποιο πεδίο (fraud detection, targeting marketing).

Αναλύοντας την επιστημονική πλευρά του θέματος καταλήγουμε ότι είναι απαραίτητη η εξόρυξη δεδομένων αφού τα δεδομένα συλλέγονται και αποθηκεύονται σε τρομερές ταχύτητες enormous speeds (GB/hour). Τέτοιες περιπτώσεις είναι και οι ακόλουθες:

- Απομακρυσμένοι αισθητήρες (remote sensors) σε δορυφόρους
- Τηλεσκόπια στον ουρανό
- Microarrays που παράγουν γονιδιακά δεδομένα
- Επιστημονικές προσομοιώσεις που παράγουν terabytes δεδομένων

Η εξόρυξη δεδομένων λοιπόν μπορεί να βοηθήσει τους επιστήμονες στους παρακάτω τομείς:

- Στην κατηγοριοποίηση και την τμηματοποίηση των δεδομένων
- Στην Διατύπωση Υποθέσεων

Παραδείγματα Δεδομένων

- Κυβερνητικά: IRS (εφορία), δημογραφικά δεδομένα, …
- Μεγάλες εταιρίες
  WALMART: 20M συναλλαγές την ημέρα
  MOBIL: 100 TB γεωλογικά σύνολα δεδομένων
  AT&T 300 M κλήσεις την ημέρα
  Εταιρίες πιστωτικών κρατών
- Επιστημονικά
  NASA, EOS project: 50 GB την ώρα
  Δεδομένα για το περιβάλλον
- «Κοινωνικά» - Ατομικά
  Νέα, ψηφιακές κάμερες, YouTube

1.1.3 Κύρια στοιχεία της διαδικασίας εξόρυξης Δεδομένων Είδη/Τεχνικές Εξόρυξης Δεδομένων

- **Ομαδοποίηση (Συσταδοποίηση) – clustering**
  Είναι η εργασία καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων. Αυτό που διαφοροποιεί την ομαδοποίηση από την κατηγοριοποίηση είναι ότι η ομαδοποίηση δεν βασίζεται σε προκαθορισμένες κατηγορίες. Στην κατηγοριοποίηση ο πληθυσμός διαιρείται σε κατηγορίες αναθέτοντας κάθε στοιχείο ή εγγραφή σε μια προκαθορισμένη κατηγορία με βάση ένα μοντέλο που αναπτύσσεται μέσα της εκπαίδευσης του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των π[προτέρων. Στην συσταδοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες. Οι εγγραφές ομαδοποιούνται σε σύνολα με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Εμείς καθορίζουμε την σημασία που θα έχει κάθε ομάδα από τις ομάδες που προκύπτουν.

- **Κανόνες συσχέτισης (Association rule mining)**
  Η εξαγωγή κανόνων συσχέτισης θεωρείται μια από τι σημαντικότερες διεργασίες εξόρυξης δεδομένων. Έχει προσελκύσει ιδιαίτερο ενδιαφέρον καθώς οι κανόνες συσχέτισης παρέχουν ένα συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου δεδομένων.
  Θα αναφερθούμε στο **Association rule mining** αργότερα σε άλλη παράγραφο.

- **Κατηγοριοποίηση (Classification)**
  Αποτελεί μια από τις βασικές εργασίες εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαρίστανται γενικά από τις εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις καθορισμένες κατηγορίες. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιεί δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί

- **Πρότυπα ακολουθιών (Sequential patterns)**

Είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικά με το χρόνο ή άλλες ακολουθίες. Οι περισσότερες μελέτες στα πρότυπα ακολουθιών επικεντρώνεται στα συμβολικά πρότυπα.

- **Ομοιότητα Χρονολογικών σειρών**
  Μια χρονολογική σειρά είναι μια ακολουθία ορισμών κάθε ένας από τους οποίους έχει ένα timestamp(ετικέτα χρόνου). Χαρακτηριστικά υποθέτουμε ότι οι διαδοχικοί αριθμοί χωρίζονται από ένα σταθερό χρονικό διάστημα και το πραγματικό timestamp παραλείπεται. Τα δεδομένα μιας χρονολογικής σειράς είναι πανταχού παρόντα. Διαφορετικές φυσικές διαδικασίες παράγουν δεδομένα υπό μορφή χρονολογικών σειρών οι οποίες εμφανίζονται μεταξύ άλλων στον οικονομικό τομέα, στον περιβαλλοντικό τομέα, στην ασφάλεια.

- **Παλινδρόμηση**
  Η παλινδρόμηση αναφέρεται στην εκμάθηση μιας λειτουργίας που εκχωρεί τα δεδομένα σε μια μεταβλητή πρόβλεψης η οποία παίρνει τιμές πραγματικές.

- **Περιληπτική παρουσίαση πληροφορίας**
  Περιλαμβάνει τη διαδικασία ανεύρεσης μιας συμπαγής περιγραφής για ένα σύνολο δεδομένων. Οι τεχνικές περιληπτικής παρουσίασης της πληροφορίας εφαρμόζεται συχνά στη διαλογική διερευνητική ανάλυση δεδομένων και την αυτοματοποιημένη παραγωγή εκθέσεων.

Είδη με βάση τα δεδομένα στα οποία γίνεται η εξόρυξη

- **Εξόρυξη στο διαδίκτυο**
  Μηχανές αναζήτησης – ενδιαφέρουσες (σημαντικές) σελίδες με βάση τους συνδέσμους.

1.1.4 Διαχωρισμός των μεθόδων εξόρυξης δεδομένων

Τα τέσσερα μέρη ή ομάδες εργασιών της ΕΔ **(*data mining tasks*)** είναι τα ακόλουθα [2]:

- Περιγραφική μοντελοποίηση (*descriptive modeling*)
- Η μοντελοποίηση πρόβλεψης (*predictive modeling*),
- Η ανάλυση συνάφειας (*association analysis*)
- Η ανίχνευση παρεκτροπών (*anomaly detection*).

Ο στόχος ενός **μοντέλου περιγραφής** είναι να γίνει περιγραφή όλου του συνόλου δεδομένων ή της διαδικασίας που παράγει τα δεδομένα. Η σημαντικότερη εφαρμογή των περιγραφικών μοντέλων είναι η **συσταδοποίηση**, η οποία επιχειρεί να βρει ομάδες παρατηρήσεων που είναι κοντά μεταξύ τους ως

προς τα χαρακτηριστικά που περιλαμβάνουν. Οι μέθοδοι περιγραφής και ειδικά η συσταδοποίηση είναι πολύ χρήσιμες σε πελατοκεντρικές επιχειρήσεις που βασίζονται στο CRM (*Customer Relationship Management*), καθώς έτσι μπορούν να εντοπιστούν ομάδες πελατών που αναμένεται να έχουν όμοια συμπεριφορά.

Η κατασκευή ενός **μοντέλου πρόβλεψης** στοχεύει στη δυνατότητα πρόγνωσης της τιμής μιας μεταβλητής (απόκριση) μέσα από τις τιμές άλλων μεταβλητών (επεξηγηματικές) που είναι γνωστές. Εάν η μεταβλητή απόκρισης είναι (ή μπορεί να θεωρηθεί) κατηγορική, τότε είμαστε σε θέση να εφαρμόσουμε μια **μέθοδο ταξινόμησης**. Όμως, αν έχουμε συνεχή απόκριση, τότε προχωράμε σε του καλαθιού αγοράς» (*market basket analysis*).

Άλλες εφαρμογές πραγματοποιούνται στην προώθηση προϊόντων ή στην τοποθέτησή τους στα ράφια καταστημάτων, στη διαχείριση αποθεμάτων κ.λπ. Στους κανόνες συνάφειας δίνεται και εκτίμηση για το πόσο πιθανό να συμβεί αυτή η σχέση αιτίας – αποτελέσματος.

Τέλος, στην **ανίχνευση παρεκτροπών** ανήκουν εργασίες εντοπισμού παρατηρήσεων των οποίων τα χαρακτηριστικά διαφέρουν σημαντικά από αυτά του υπόλοιπου συνόλου δεδομένων (έκτροπες παρατηρήσεις ή outliers). Στόχος είναι η υψηλού επιπέδου ανίχνευση πιθανών ανωμαλιών, διατηρώντας όμως χαμηλά ποσοστά λανθασμένης προειδοποίησης. Ως εφαρμογή μπορούμε να αναφέρουμε τον προσδιορισμό απειλής στην έγκριση δανείων ή πιστωτικών καρτών από μια τράπεζα.

Η ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΚΑΛΥΨΗΣ ΤΗΣ ΓΝΩΣΗΣ
Οι απαρχές της εξόρυξης δεδομένων

Στα πλαίσια της αναζήτησης περισσότερο αποτελεσματικών και δυναμικών εργαλείων διαχείρισης διαφορετικής φύσεως δεδομένων, ερευνητές από διάφορους επιστημονικούς κλάδους επιχείρησαν να ενώσουν τα αντικείμενα του ενδιαφέροντός τους. Η συνεργασία αυτή βρήκε πρόσφορο έδαφος στο πεδίο της ΕΔ, βασιζόμενη στην εφαρμογή μεθοδολογιών και αλγορίθμων που είχαν ήδη χρησιμοποιηθεί από τους ερευνητές.

Πιο συγκεκριμένα η ΕΔ [2], χρησιμοποιεί έννοιες όπως δειγματοληψία, εκτίμηση και έλεγχος υποθέσεων από τη Στατιστική, καθώς και εφαρμογές όπως αναζήτηση αλγορίθμων, τεχνικές δημιουργίας υποδειγμάτων, θεωρίες τεχνητής νοημοσύνης, αναγνώρισης προτύπων και μηχανικής εκμάθησης.

Επιπλέον, υπάρχουν αρκετοί άλλοι τομείς των επιστημών που στήριξαν την πρόοδο της ΕΔ, όπως για παράδειγμα, η τεχνολογία των βάσεων δεδομένων. Τέλος, τεχνικές υψηλής απόδοσης από υπολογιστικής πλευράς και σχετικές με την

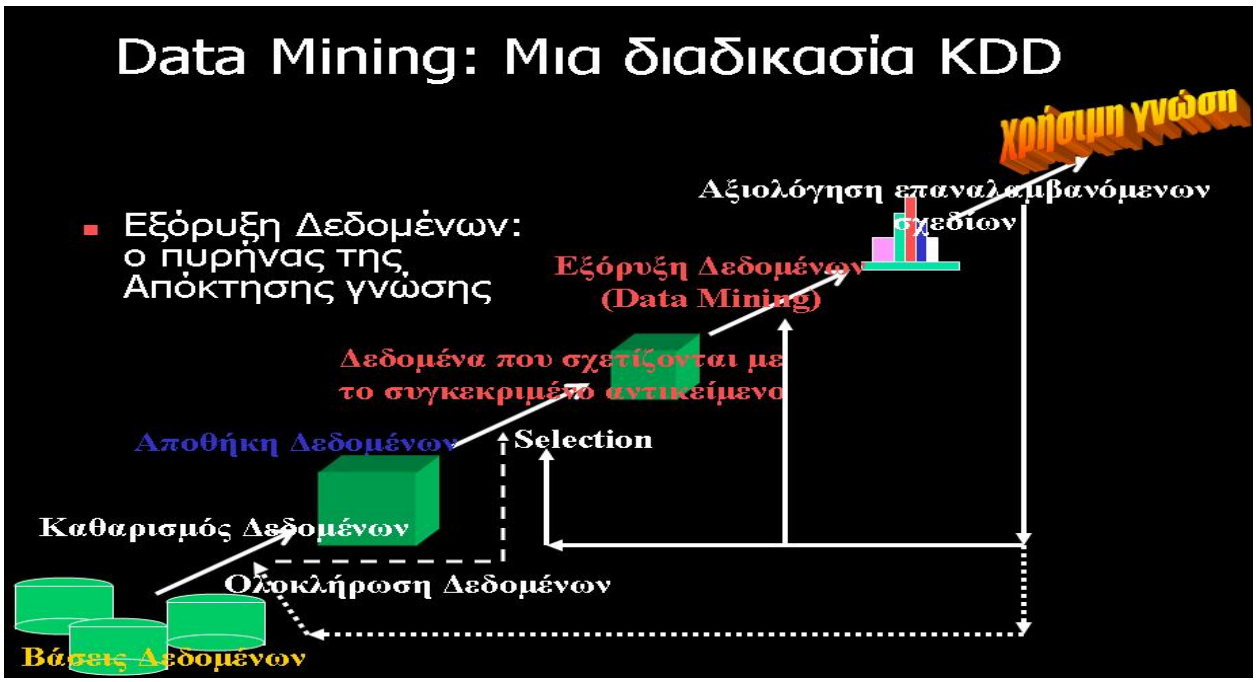
ταξινόμηση παρέχουν βοήθεια σε σχέση με τη διαχείριση του μεγέθους και της συλλογής των τεράστιων συνόλων δεδομένων.

1.1.5 Η διαδικασία KDD

Επεξεργαζόμενοι μια τεράστια βάση δεδομένων, είναι πιθανό να ανακαλύψουμε την ύπαρξη «κρυμμένης γνώσης». Δηλαδή, μπορεί να εντοπίσουμε συσχετίσεις, αλληλεξάρτηση ή ομαδοποιήσεις μεταξύ των δεδομένων, πράγματα τα οποία μπορεί να μην είναι άμεσα εμφανή. Το είδος αυτό της γνώσης θεωρείται ότι δεν είναι εκ των προτέρων διαθέσιμο, αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Υπό αυτές τις συνθήκες, κρίνεται απαραίτητη η «μη επιβλεπόμενη» ανάκτηση γνώσης, που υποστηρίζεται από την εφαρμογή αλγορίθμων. Αυτήν την ανάγκη έρχεται να καλύψει η ΕΔ, η οποία αποτελεί τον πυρήνα της γενικότερης μεθοδολογίας της **ανακάλυψης της γνώσης από βάσεις δεδομένων** (*Knowledge Discovery in Databases – KDD*).

Η KDD είναι μία αυτοματοποιημένη διαδικασία, μέσω της οποίας γίνεται προσπάθεια διερευνητικής ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Πρόκειται για μια συγκροτημένη μεθοδολογία αναγνώρισης έγκυρων και πρωτότυπων προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων, με στόχο τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά. Ένας γενικός ορισμός της διαδικασίας KDD που ερμηνεύει με σαφήνεια τον όρο αυτό [15] είναι:
«KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα».



Η ονομασία αυτή της KDD χρησιμοποιείται από το 1989 (πρώτο συνέδριο KDD) με στόχο να φανεί ότι η γνώση είναι το τελικό προϊόν μιας ανακάλυψης καθοδηγούμενης από τα δεδομένα [5]. Με βάση τη σχετική βιβλιογραφία, θα διαχωρίσουμε τη διαδικασία KDD σε εννέα βήματα, τα οποία είναι:

1) Την ανάπτυξη και κατανόηση της περιοχής της εφαρμογής, της σχετικά προγενέστερης γνώσης του προς εξέταση τομέα και τους στόχους του τελικού χρήστη.

2) Την επιλογή και δημιουργία ενός κατάλληλου συνόλου δεδομένων. Την ολοκλήρωση δηλαδή των δεδομένων. Υπάρχουν διαφορετικά είδη αποθηκών πληροφοριών που μπορούν να χρησιμοποιηθούν στη διαδικασία εξόρυξης γνώσης. Κατά συνέπεια, οι πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν καθορίζοντας το σύνολο στο οποίο τελικά η διαδικασία εξόρυξης θα εκτελεστεί.

3) Την δημιουργία στόχου – συνόλου δεδομένων. Επιλογή του συνόλου δεδομένων (δηλαδή μεταβλητές, δείγματα δεδομένων) στο οποίο η διαδικασία εξόρυξης πρόκειται να εκτελεσθεί.

4) Τον καθαρισμό και την προ-επεξεργασία δεδομένων. Αυτό το βήμα περιλαμβάνει βασικές διαδικασίες όπως η αφαίρεση θορύβου ή των outliers, η συλλογή των απαραιτήτων πληροφοριών για την διαμόρφωση ή τη μέτρηση του θορύβου, η απόφαση σχετικά με τις στρατηγικές διαχείρισης των ελλειπόντων πεδίων δεδομένων.

5) Τον μετασχηματισμό των δεδομένων. Τα δεδομένα μετασχηματίζονται ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη. Χρήση των μεθόδων μείωσης διαστάσεων ή μετασχηματισμού για τη μείωση του αριθμού των υπό εξέταση μεταβλητών ή την εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.

6) Την επιλογή των στόχων και των αλγορίθμων κατάλληλης μεθόδου εξόρυξης δεδομένων. Σε αυτό το βήμα αποφασίζουμε το στόχο της διαδικασίας KDD, επιλέγοντας τους στόχους εξόρυξης δεδομένων που θέλουμε να επιτύχουμε. Επίσης, επιλέγονται οι μέθοδοι που θα χρησιμοποιηθούν. Αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου και παραμέτρων. Επίσης η μέθοδος εξόρυξης δεδομένων πρέπει να αντιστοιχηθεί με τις απαιτήσεις και τα γενικά κριτήρια της διαδικασίας KDD.

7) Την εξόρυξη δεδομένων. Εφαρμόζουμε ευφυείς μεθόδους, ψάχνουμε για ενδιαφέροντα πρότυπα γνώσης. Τα πρότυπα θα μπορούσαν να είναι μιας συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου τέτοιων

αντιπροσωπεύσεων, όπως κανόνες κατηγοριοποίησης, δέντρα, παλινδρόμηση συσταδοποίηση κ.τ.λ. Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα βήματα.

8) Την αξιολόγηση των προτύπων. Τα εξαγόμενα πρότυπα αξιολογούνται με κάποια μέτρα, προκειμένου να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση, δηλαδή τα αληθινά ενδιαφέροντα πρότυπα.

9) Την σταθεροποίηση και την παρουσίαση της γνώσης. Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα ή απλά την απεικόνισής μας και κάποιες τεχνικές αντιπροσώπευσης γνώσης χρησιμοποιούνται για να παρουσιάσουν την εξορυγμένη γνώση στο χρήστη.

Η διαδικασία KDD θεωρείται διαλογική και επαναληπτική, δηλαδή μπορεί να απαιτηθεί η επιστροφή σε ένα προηγούμενο βήμα.

Ως εφαρμογές της KDD στον χώρο των επιχειρήσεων αναφέρουμε τις δραστηριότητες σε marketing, επενδύσεις, προσδιορισμό απειλών, βιομηχανική παραγωγή, τηλεπικοινωνίες, καθαρισμό δεδομένων. Προφανώς, η δράση υ964 της KDD σε αυτούς τους τομείς γίνεται μέσω της ΕΔ, δηλαδή η ΕΔ αποτελεί το **εργαλείο** της KDD.

Για να είναι σαφής η διαφορά μεταξύ διαδικασίας και εργαλείων, αναφέρουμε ότι ο όρος KDD χρησιμοποιείται για την περιγραφή ολόκληρης της διαδικασίας ανακάλυψης γνώσης από ένα σύνολο δεδομένων, ενώ ο όρος ΕΔ αναφέρεται στις τεχνικές που χρησιμοποιούνται για την ανακάλυψη της γνώσης.

Ο όρος ΕΔ αντιπροσωπεύει καλύτερα τη διαδικασία εύρεσης δομών γνώσης που περιγράφουν με ακρίβεια σύνολα πρωτογενών δεδομένων. Οι δομές αυτές αναδεικνύουν κρυμμένη γνώση (συνάφειες / κανόνες) που δεν είναι άμεσα ορατή και εκμεταλλεύονται πιθανές κοινές ιδιότητες των πρωτογενών δεδομένων.

1.2 . Market Basket Analysis

1.2.1 Εισαγωγή

Η τεχνική Market Basket Analysis βασίζεται στην θεωρία ότι αν κάποιος πελάτης αγοράσει κάποιο συγκεκριμένο προϊόν (ή σύνολο προϊόντων), τότε είναι πολύ πιθανό (ή αντίστοιχα ελάχιστα πιθανό) να αγοράσει και ένα άλλο προϊόν (ή σύνολο προϊόντων) [4]. Το σύνολο των προϊόντων που αγοράζει ένας πελάτης κατά την διάρκεια μιας συγκεκριμένης αγοράς του ονομάζεται itemset. Άρα, ένα itemset αποτελείται από κάποια προϊόντα. Η τεχνική market basket analysis έχει σαν κύριο στόχο την ανάλυση των δεδομένων που προκύπτουν από τις αγορές των πελατών, με σκοπό την ανακάλυψη συσχετίσεων μεταξύ των διαφόρων προϊόντων. Τυπικά, μια συσχέτιση μεταξύ δύο προϊόντων είναι της μορφής:

IF { προϊόν A } THEN { προϊόν B }

Η πιο πάνω συσχέτιση, δείχνει την σχέση μεταξύ των δύο προϊόντων. Φυσικά, μια συσχέτιση θα μπορούσε να περιλαμβάνει, αντί για μεμονωμένα προϊόντα και σύνολα προϊόντων. Μια τέτοια συσχέτιση θα μπορούσε να είναι η εξής:

IF { γάλα, ψωμί } THEN { βούτυρο, δημητριακά }

Με τέτοιες συσχετίσεις συνδέονται άμεσα στατιστικές μεταβλητές, οι οποίες ονομάζονται support και confidence.

Η πιθανότητα να αγοράσει ένας πελάτης γάλα και ψωμί καλείται support του συγκεκριμένου κανόνα (υποστήριξη) και η πιθανότητα- υπό όρους- να αγοράσει βούτυρο και δημητριακά καλείται confidence (εμπιστοσύνη).

Οι συσχετίσεις μεταξύ των προϊόντων ονομάζονται και κανόνες συσχέτισης, όπου οι μεταβλητές support και confidence δίνουν στατιστική πληροφορία που αφορά τους κανόνες αυτούς. Στην επόμενη παράγραφο δίνουμε έναν πιο τυπικό ορισμό της διαδικασίας market basket analysis.

Οι αλγόριθμοι οι οποίοι εφαρμόζονται σε αυτή την τεχνική είναι αρκετά απλοί. Τις περισσότερες φορές οι δυσκολίες παρουσιάζονται σε περιπτώσεις όπου απαιτείται ταξινόμηση προϊόντων όπως για παράδειγμα σε ένα supermarket όπου έχει στην αποθήκη 10000 ή και περισσότερα προϊόντα και θα πρέπει κάποιος να ασχοληθεί με μεγάλες συναλλαγές δεδομένων που υπάρχουν στην διάθεση.

Μια μεγάλη δυσκολία επίσης αποτελεί και ο μεγάλος αριθμός κανόνων ο οποίος μπορεί για τους οικείους με την εκάστοτε επιχείρηση, να φαίνεται ασήμαντος στην ουσία όμως είναι σαν να ψάχνεις κάποιος καρφίτσα σε αχυρώνα. Είναι

απαραίτητο να υπάρχει μέγιστο το ελάχιστο επίπεδο στήριξης - minimum support level καθώς και υψηλό διάστημα εμπιστοσύνης κινδύνου - confidence level risks.

1.2.2 Ορισμός Market Basket Analysis (MBA)

Ο όρος Market Basket Analysis (MBA), ή ανάλυση καλαθιού αγορών, αφορά την ανάλυση διαφόρων υποσυνόλων αντικειμένων (προϊόντων), τα οποία επιλέχθηκαν μέσα από κάποιον μεγαλύτερο πληθυσμό αντικειμένων [5].

Γνωρίζοντας και αναλύοντας τι προϊόντα αγοράζει σε συνδυασμό κάθε πελάτης είναι πολύ βοηθητικό σε όλων των ειδών τους πωλητές.

Για παράδειγμα ένα εμπορικό κέντρο μπορεί να χρησιμοποιήσει αυτήν την τεχνική για να οργανώνει και να διαθέσει τα προϊόντα τα οποία πωλούνται συχνότερα στην ίδια περιοχή. Οι μηχανές αναζήτησης στο διαδίκτυο επίσης μπορούν να χρησιμοποιήσουν την τεχνική αυτή για να καθορίσουν τη διάταξη του καταλόγου τους και την φόρμα παραγγελίας σε ένα δικτυακό τόπο.

Οι πωλητές οι οποίοι προωθούν νέα προϊόντα στην αγορά μπορούν να χρησιμοποιήσουν το MBA για να καθορίσουν ποια είναι τα νέα προϊόντα πρέπει να προσφέρουν στους πελάτες τους και να μπορέσουν να οργανώσουν συνδυασμούς προϊόντων προς πώληση.

Η εφαρμογή του MBA κατά κανόνα χρησιμοποιεί για την εφαρμογή της τα εργαλεία εξόρυξης δεδομένων. Ο πρωταρχικός στόχος του MBA είναι να βελτιώσει την αποτελεσματικότητα του μάρκετινγκ και των πωλήσεων αξιοποιώντας τα δεδομένα πωλήσεων που συσσωρεύονται σε μια επιχείρηση κατά την διάρκεια των συναλλαγών.

Επιπρόσθετα, η MBA τεχνική μπορεί να εφαρμοστεί, εκτός του εμπορικού τομέα, και στην απογραφή ενός καταστήματος βοηθώντας τους υπευθύνους να καταλάβουν ποια προϊόντα ήταν ευρείας κατανάλωσης καθώς και ποια προϊόντα πωλούνταν συνήθως μαζί.

Η τεχνική MBA λοιπόν συσχετίζεται με την ανάλυση σημαντικών υποσυνόλων από Items τα οποία λαμβάνονται από έναν πληθυσμό από Items. Τα υποσύνολα ή οι κανόνες, τα οποία λαμβάνονται υπόψιν είναι συνήθως αυτά τα οποία έχουν την ελάχιστη τιμή «εκτίμησης» και «εμπιστοσύνης».

Ένα παράδειγμα κανόνα είναι το A→B, όπου υποδηλώνει ότι: *«εάν το* αντικείμενο A υπάρχει στο καλάθι αγορών (market basket), τότε υπάρχει και το αντικείμενο B».

Το Α ονομάζεται προηγηθέν αντικείμενο (antecedent item), ενώ το Β συνεπακόλουθο (consequent).

Σε έναν κανόνα MBA, όπως τον Α→Β, μπορεί να έχουμε ότι ενώ ο κανόνας είναι αληθής (true), να ισχύει Α αληθές και Β μη αληθές, δηλαδή *A AND NOT B*. Έχοντας δηλαδή έναν κανόνα, μπορεί αυτός ενώ είναι αληθής, δηλαδή ενώ ισχύει, τα επιμέρους στοιχεία του να μην ισχύουν [2]. Αυτό συμβαίνει, επειδή οι κανόνες στο market basket analysis θεωρούνται να έχουν κάποιους βαθμούς συνέπειας άμεσα συσχετισμένους με αυτούς. Τέτοιοι είναι οι confidence και support στατιστικές.

Σαν καλάθι αγορών ονομάζουμε μια **συναλλαγή (transaction)** κάποιου πελάτη. Αν έχουμε για παράδειγμα μια υπεραγορά, τότε σαν καλάθι αγορών ονομάζουμε το σύνολο των προϊόντων που αγοράστηκαν από κάποιον πελάτη σε μια συγκεκριμένη συναλλαγή αυτού με το κατάστημα. Οι πιο πάνω κανόνες χρησιμοποιώντας τον όρο συναλλαγή αντί για καλάθι αγορών, θα γίνουν:

Το **support** κάποιου κανόνα Α→Β ορίζεται σαν:
Αν τα Α και Β ισχύουν μαζί για τουλάχιστον Χ% των καλαθιών αγορών, τότε το support του κανόνα είναι το Χ.

Το **confidence** κάποιου κανόνα Α→Β ορίζεται σαν:
Από όλα τα καλάθια αγορών που περιέχουν το Α, αν τουλάχιστον Χ% περιέχουν επίσης το Β, τότε το confidence του κανόνα είναι Χ.

Για σκοπούς marketing, το confidence διαβεβαιώνει ότι ο κανόνας ισχύει, δηλαδή είναι αληθής, μέχρι κάποιο συγκεκριμένο σημείο, ή αλλιώς με κάποια συγκεκριμένη πιθανότητα. Χρησιμοποιώντας το confidence, μπορεί κάποιος να διαβεβαιώσει ότι κάποιος κανόνας ισχύει αρκετά συχνά, ώστε να παίξει σημαντικό ρόλο κατά την λήψη αποφάσεων. Για παράδειγμα, αν σε κάποια υπεραγορά κάποιος κανόνας ισχύει αρκετά συχνά ή με αρκετά μεγάλη πιθανότητα, τότε λόγω του κανόνα αυτού μπορεί να αποφασιστεί διαρρύθμιση των προϊόντων που εμφανίζονται σε αυτόν.

Η χρήση του confidence από την άλλη, περιλαμβάνει και κάποιο ρίσκο. Αυτό συμβαίνει, επειδή όταν το συνεπακόλουθο αντικείμενο κάποιου κανόνα είναι δημοφιλές γενικότερα στις συναλλαγές, τότε το confidence του κανόνα μπορεί να είναι αρκετά μεγάλο, άσχετα με το αν τα δύο αντικείμενα (προηγηθέν και συνεπακόλουθο) δεν συσχετίζονται στην πραγματικότητα σε τόσο μεγάλο βαθμό. Βλέποντας το θέμα και διαισθητικά, από τον ορισμό του confidence, έχουμε ότι εάν ένας κανόνας Α→Β έχει confidence για παράδειγμα 95, τότε σημαίνει ότι από όλες τις συναλλαγές που περιέχουν το Α, τουλάχιστον 95% περιέχουν επίσης το Β. Στην περίπτωση όμως που το Β είναι πολύ δημοφιλές προϊόν στις συναλλαγές γενικότερα, τότε μπορεί μεν να εμφανίζεται σε πολύ υψηλό ποσοστό

(95%) στις ίδιες συναλλαγές με το Α, όμως στην πραγματικότητα δεν συσχετίζεται τόσο με το συγκεκριμένο προϊόν, όσο θα συσχετιζόταν αν το Β εμφανιζόταν κυρίως μόνο στις συναλλαγές που περιλαμβάνουν το Α (δηλαδή αν το Β δεν ήταν δημοφιλές).

Το support παρέχει ένα μέτρο για το πόσο συχνά ένας κανόνας συμβαίνει (σε πόσες συναλλαγές είναι αυτός αληθείς), στο σύνολο όλων γενικότερα των συναλλαγών. Χρησιμοποιώντας το support, ένας αναλυτής μπορεί να συμπεράνει κατά πόσο αξίζει την προσοχή του κάποιος κανόνας.

1.2.3 Άλλες περιοχές εφαρμογών του MBA.

Παρακάτω δίνουμε και κάποιες άλλες «περιοχές – τομείς» στους οποίου εφαρμόζεται η τεχνική Market Basket Analysis (MBA).[5]

• Ανάλυση των αγορών με πιστωτική κάρτα.

• Ανάλυση των τηλεφωνικών σχεδίων.

•Προσδιορισμός των πλαστών ιατρικών ασφαλιστικών απαιτήσεων.

• Ανάλυση των τηλεπικοινωνιακών υπηρεσιών και αγορών.

1.3 Κανόνες συσχέτισης - Association Rules

Οι κανόνες συσχέτισης, ή αλλιώς association rules[6], είναι κανόνες οι οποίοι εκφράζουν συσχετίσεις μεταξύ αντικειμένων. Πιο συχνά χρησιμοποιούνται σε συστήματα σημείων πώλησης προϊόντων, οπότε οι συσχετίσεις που εκφράζουν είναι μεταξύ των διαφόρων προϊόντων που αγοράζουν οι πελάτες. Οι κανόνες προκύπτουν με την διαδικασία εξόρυξης κανόνων συσχέτισης (association rule mining).

Οι κανόνες συσχέτισης είναι δημοφιλής στο data mining αλλά έχουν χρησιμοποιηθεί επίσης και στο text mining [33]. Ένας κανόνας συσχέτισης είναι μια απλή πιθανολογική ανάλυση σχετικά με την ομο-εμφάνιση ορισμένων γεγονότων σε μια βάση δεδομένων ή μια συλλογή κειμένων. Ο Apriori, θα αναφερθούμε εκτενέστερα σε επόμενο κεφάλαιο, είναι ο γνωστότερος αλγόριθμος εύρεσης κανόνων συσχέτισης.

Στα πλαίσια του text mining, οι κανόνες συσχέτισης έχουν χρησιμοποιηθεί για να ανακαλύψουν τις ενδεχομένως σημαντικές σχέσεις μεταξύ των εννοιών που ομο - εμφανίζονται στα κείμενα [35][34]. Λαμβάνοντας υπόψη έναν σύνολο εγγράφων, προσδιορίζονται οι σχέσεις μεταξύ των ιδιοτήτων (χαρακτηριστικά γνωρίσματα

που έχουν εξαχθεί από τα έγγραφα) όπως η παρουσία μιας λέξης ή ενός όρου να υπονοεί την ύπαρξη ενός άλλου όρου ή μιας λέξης.

Τα οφέλη με τους κανόνες συσχέτισης στο text mining προσδιορίζονται από το γεγονός ότι οι κανόνες που προκύπτουν από βάσεις δεδομένων είναι μια αρκετά εντατική διαδικασία. Αυτό κυρίως συμβαίνει λόγω της υψηλής διαστατικότητας του διαστήματος χαρακτηριστικών γνωρισμάτων. Ως εκ τούτου ο αριθμός των λέξεων που πρέπει να εξεταστούν κατά τη δημιουργία μεγάλων λιστών (Large itemsets) είναι μεγαλύτερα από τον αριθμό στοιχείων σε ένα σύνολο συναλλαγών.

Η εξόρυξη κανόνων συσχέτισης είναι μια πολύ διαδεδομένη διαδικασία, κατά την οποία γίνεται κατάλληλη ανάλυση των δεδομένων που είναι αποθηκευμένα σε βάσεις δεδομένων, για ανακάλυψη χρήσιμων συσχετίσεων μεταξύ προϊόντων, όπως για παράδειγμα η εύρεση προϊόντων τα οποία αγοράστηκαν μαζί.

Οι κανόνες συσχέτισης χρησιμοποιούνται σε εφαρμογές Market Basket Analysis. 'Οπως είδαμε σε προηγούμενη παράγραφο, με το όρο Market Basket Analysis εννοούμε την αναγνώριση διαφόρων ευκαιριών για πώληση προϊόντων που σχετίζονται μαζί (cross selling opportunities).

Για παράδειγμα:

1. Αναγνώριση ομάδων προϊόντων που αγοράζονται μαζί (product baskets)

2. Η πρόβλεψη άλλων προϊόντων που ενδεχομένως να μπορούν να αγοραστούν μαζί, δεδομένων των προϊόντων που έχουν είδη αγοραστεί μαζί.

Ας θεωρήσουμε λοιπόν ότι βρισκόμαστε σε ένα σούπερ-μάρκετ και στο οποίο υπάρχει μια μεγάλη συλλογή από αντικείμενα-items. Κάθε επιχειρηματική απόφαση που πρέπει να παρθεί με στόχο την σωστή διαχείριση της επιχείρησης βασίζεται στην σωστή μελέτη των προϊόντων που διατίθενται προς πώληση, στην σωστή τοποθέτηση των εμπορευμάτων στα ράφια και πολλά άλλα.

Η ανάλυση των πιο πρόσφατων συναλλαγών δεδομένων είναι μια ευρέως χρησιμοποιούμενη προσέγγιση προκειμένου να βελτιωθεί η ποιότητα των προηγούμενων αναφερθέντων αποφάσεων. Μέχρι πρόσφατα, ωστόσο, μόνο τα συνολικά στοιχεία σχετικά με τις συνολικές πωλήσεις σε συγκεκριμένο χρονικό διάστημα (μία μέρα, μία εβδομάδα, το μήνα, κ.λ.π.) ήταν διαθέσιμα και προσβάσιμα στον υπολογιστή. Με την πρόοδο της τεχνολογίας όμως και συγκεκριμένα με την bar - code τεχνολογία κατέστη δυνατή η αποθήκευση των λεγόμενων basket data όπου συγκεκριμένα αποθηκεύονται τα Items – αντικείμενα που αγοράστηκαν βασιζόμενοι σε κάθε συναλλαγή.

Οι συναλλαγές του τύπου basket data δεν περιλαμβάνουν απαραίτητα αντικείμενα τα οποία κατ 'ανάγκην αγοράστηκαν την ίδια χρονική περίοδο. Μπορεί να αποτελούνται και από αντικείμενα που αγοράζονται από κάποιον πελάτη κατά τη διάρκεια μιας χρονικής περιόδου. Τα παραδείγματα περιλαμβάνουν μηνιαίες αγορές από τα μέλη μιας λέσχης βιβλίου ή μιας επιχείρησης.

Πολλές οργανώσεις έχουν συλλέξει τεράστιες ποσότητες των δεδομένων αυτών. Αυτά τα σύνολα δεδομένων συνήθως αποθηκεύονται σε αποθήκες τριτογενούς μορφής και είναι πολύ δύσκολο να μεταφερθούν σε μεγάλα συστήματα βάσεων δεδομένων.

'Ενας από τους βασικούς λόγους για την περιορισμένη επιτυχία των συστημάτων των βάσεων δεδομένων στον τομέα αυτό είναι ότι τα τρέχοντα συστήματα διαχείρισης βάσεων δεδομένων, δεν παρέχουν την απαραίτητη λειτουργικότητα για ένα χρήστη που ενδιαφέρεται να επωφεληθείτε από αυτές τις πληροφορίες.

Το πρόβλημα το οποίο λοιπόν προκύπτει είναι το ακόλουθο: η εξόρυξη δεδομένων και πληροφοριών από μια μεγάλη συλλογή συναλλαγών basket data με σκοπό την εύρεση κανόνων συσχέτισης μεταξύ ενός συνόλου αντικειμένων.

'Ενα παράδειγμα ενός κανόνα συσχέτισης [6] αποτελεί μπορεί να δηλωθεί το εξής:
{ότι το 90% των συναλλαγών που αγοράζουν ψωμί και βούτυρο }$\Rightarrow${αγοράζουν γάλα}.

Στον κανόνα αυτό το ψωμί και το βούτυρο είναι τα προηγηθέντα προϊόντα, ενώ το γάλα το συνεπακόλουθο προϊόν. Επίσης, το confidence του κανόνα είναι 90 και εκφράζει την δύναμη του κανόνα. Στην περίπτωση αυτή, εκφράζει το ποσοστό των συναλλαγών που περιλαμβάνουν γάλα, δεδομένου ότι περιλαμβάνουν ψωμί και βούτυρο.

Οι Agrawal, Imielinski και Swami στο [5], χρησιμοποίησαν κανόνες συσχέτισης για την ανακάλυψη ομοιοτήτων μεταξύ προϊόντων σε μεγάλες βάσεις δεδομένων. Οι βάσεις δεδομένων αυτές περιλάμβαναν τεράστιους όγκους δεδομένων από συναλλαγές πελατών, σε διάφορα σημεία πώλησης προϊόντων όπως υπεραγορές (supermarkets).

Ακολουθεί ένα παράδειγμα του οποίου η δυναμική έγκειται στην βελτίωση των βάσεων δεδομένων και στην σωστή επεξεργασία των διαφόρων ερωτηματολογίων.

Για παράδειγμα λοιπόν βρισκόμαστε σε ένα σούπερ-μάρκετ και ακολουθούμε τα παρακάτω βήματα:

1. Βρείτε όλους τους κανόνες οι οποίοι περιέχουν τη λέξη «Diet Coke» σαν συνεπακόλουθο προϊόν.
   Αυτοί οι κανόνες θα μπορούσαν να προωθήσουν τις πωλήσεις του καταστήματος σε «Diet Coke»

2. Βρείτε όλους τους κανόνες που περιέχουν τη λέξη «bagels» σαν προηγηθέν προϊόν.
   Αυτοί οι κανόνες θα μπορέσουν να καθορίσουν οι πωλήσεις ποιων προϊόντων θα επηρεαστούν αν το κατάστημα ελαττώσει σε μεγάλο βαθμό τις πωλήσεις των «bagels».

3. Βρείτε όλους τους κανόνες που περιέχουν σαν προηγηθέν προϊόν τη λέξη «λουκάνικα» και σαν επακόλουθο προϊόν τη λέξη «μουστάρδα».
   Αυτού του είδους το ερώτημα μπορεί να διατυπωθεί εναλλακτικά σαν ένα ερώτημα για τα επιπρόσθετα προϊόντα τα οποία θα πρέπει να πωληθούν μαζί με τα λουκάνικα ώστε να είναι πολύ πιθανό ένα από αυτά να είναι και η μουστάρδα.

4. Βρείτε όλους τους κανόνες συσχέτισης των προϊόντων τα οποία βρίσκονται στα ράφια Α και Β στο κατάστημα.
   Οι κανόνες αυτοί μπορούν να βοηθήσουν στον καθορισμό του αν οι πωλήσεις του Item στο ράφι Α σχετίζονται με τις πωλήσεις του item στο ράφι Β.

5. Βρείτε *τους καλύτερους k κανόνες, που έχουν κάποιο προϊόν (π.χ. bagels) σαν συνεπακόλουθο προϊόν.* Ο όρος καλύτερος κανόνας αναφέρεται στον κανόνα με το μεγαλύτερο confidence factor ή την μεγαλύτερη συχνότητα εμφάνισης.

   Κάνουμε χρήση *μόνο* αυτών των κανόνων οι οποίοι είναι και οι πλέον κατάλληλοι για λήψη αποφάσεων, αφού οι κανόνες αυτοί είναι οι πιο «βάσιμοι» και χαρακτηρίζουν το μεγαλύτερο μέρος των συναλλαγών (εφόσον είναι οι πιο συχνά εμφανιζόμενοι).

1.3.1 Ορισμός των κανόνων συσχέτισης – Association Rules.

Δίνοντας πιο τυπικό ορισμό των κανόνων συσχέτισης, έχουμε:

A→B: Δεδομένης της αγοράς του προϊόντος Α, υπάρχει μεγάλη πιθανότητα να έχει αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για το προϊόν Β. Το προϊόν Α ονομάζεται προηγηθέν, ενώ το Β συνεπακόλουθο.

Από τα πιο πάνω, συμπεραίνουμε ότι οι κανόνες συσχέτισης μπορούν να προσφέρουν αξιοποιήσιμη πληροφορία που σχετίζεται με τα προϊόντα. Με χρήση της πληροφορίας αυτής, μπορεί να γίνει καλύτερη τιμολόγηση των προϊόντων, να αποφασιστούν ποια προϊόντα θα βγουν σε εκπτώσεις και προσφορές και ποια όχι, να μελετηθούν οι επιπτώσεις σε άλλα προϊόντα από τυχόν κατάργηση κάποιου προϊόντος, να αποφασιστεί η διαρρύθμιση των προϊόντων στα ράφια κ.α.

Επιπλέον, εκτός από το market basket analysis, οι κανόνες συσχέτισης χρησιμοποιούνται και σε άλλες εφαρμογές, όπως το Web usage mining, intrusion detection και βιοπληροφορική (bioinformatics).

1.4 Εξαγωγή κανόνων Συσχέτισης

1.4.1 Εισαγωγή

Ξεκινώντας ας δώσουμε τους ορισμούς που αφορούν στην Τεχνολογία της Επιστήμης, στην Εξόρυξη Δεδομένων καθώς και στην Εύρεση κανόνων συσχέτισης.[17]

Σαν **Επιστήμη της Τεχνολογίας** καλείται η μελέτη των θεωρητικών θεμελίων της πληροφορίας και των υπολογιστών καθώς και των πρακτικών τεχνικών που απαιτούνται για την εκτέλεση και εφαρμογή τους στα υπολογιστικά συστήματα. Συνήθως περιγράφεται σαν μια συστηματική μελέτη των διαδικασιών που ακολουθούν οι αλγόριθμοι οι οποίοι δημιουργούν, περιγράφουν και μετατρέπουν και ανακατασκευάζουν την πληροφορία.

Σαν **Εξόρυξη Δεδομένων** καλείται η διαδικασία της εξαγωγής κανόνων – προτύπων από τα δεδομένα. Η εξόρυξη Δεδομένων θεωρείται σαν ένα πολύ σημαντικό εργαλείο το οποίο χρησιμοποιείται από τις νεότερες επιχειρήσεις με σκοπό την μετατροπή των δεδομένων σε σημαντική πληροφορία.

Η **Διαδικασία Εύρεσης κανόνων συσχέτισης – Discovery Association Rules in Data Mining**, αποτελεί μια πολύ δημοφιλή και εμπλουτισμένη μέθοδος για την ανακάλυψη συσχετίσεων μεταξύ μεταβλητών σε μεγάλες βάσεις Δεδομένων. Οι Piatesky-Shapiro περιγράφουν την ανάλυση και την παρουσίαση των ισχυρών κανόνων οι οποίοι ανακαλύπτονται στις βάσεις δεδομένων χρησιμοποιώντας διαφορετικά μέτρα «ενδιαφέροντος». Βασιζόμενοι στην ιδέα των ισχυρών κανόνων ο Agrawal, παρουσίασε κανόνες συσχέτισης για την ανακάλυψη κανόνων μεταξύ προϊόντων σε μεγάλες τάξεις συναλλαγών οι οποίες καλούνται σαν point of sale συστήματα σε supermarket δηλαδή σε μεγάλους χώρους συναλλαγών.

Για παράδειγμα ο κανόνας {κρεμμύδια, πατάτες $\Rightarrow$βοδινό κρέας} το οποίο προκύπτει συνήθως σε αγορές στο supermarket θα μπορούσε να υποδηλώσει ότι αν κάποιος πελάτης αγοράζει ταυτόχρονα κρεμμύδια και πατάτες είναι πολύ πιθανό να αγοράζει και βοδινό κρέας. Μια τέτοιου είδους πληροφορία μπορεί να χρησιμοποιηθεί σαν τη βάση- θεμέλιο για την απόφαση κάποιων ενεργειών προώθησης προϊόντων όπως για παράδειγμα αλλαγές στην τιμολόγηση ή τοποθέτηση σε κατάλληλη- ιδανική θέση. Επιπρόσθετα οι κανόνες συσχέτισης έχουν εφαρμογή και σε άλλους τομείς όπως εξόρυξη στο Διαδίκτυο και στην Βιοπληροφορική.

Ο Apriori στην πληροφορική και στην διαδικασία εξόρυξης δεδομένων αποτελεί έναν αλγόριθμο ο οποίος αφορά στην γνώση και εύρεση των κανόνων συσχέτισης.

Έχει σχεδιαστεί με τέτοιο τρόπο ώστε να μπορεί να εφαρμοστεί σε βάσεις δεδομένων οι οποίες περιέχουν συναλλαγές(όπως για παράδειγμα σύνολα προϊόντων που αγοράστηκαν από πελάτες ή λεπτομέρειες για την συχνή επίσκεψη σε έναν δικτυακό τόπο) Άλλοι αλγόριθμοι είναι σχεδιασμένοι για να εφαρμόζονται, για την εύρεση κανόνων συσχέτισης, σε δεδομένα στα οποία δεν εμπλέκεται κανενός είδος συναλλαγή ή σε δεδομένα στα οποία δεν υπάρχει συγκεκριμένο χρονοδιάγραμμα για παράδειγμα στην DNA αλληλουχία.

Όπως είναι σύνηθες στην εξόρυξη κανόνων συσχέτισης, δίνεται ένα δεδομένο σύνολο από itemesets, (για παράδειγμα σύνολα από πωλήσεις λιανικής όπου σε καθεμία δημιουργείται μια μεμονωμένη λίστα με τα προϊόντα τα οποία αγοράστηκαν) και σε αυτό το σύνολο ο αλγόριθμος επιχειρεί να βρει όλα τα υποσύνολα τα οποία είναι κοινά σε τουλάχιστον ένα ελάχιστο αριθμό c των itemsets.

Ο Apriori αλγόριθμος χρησιμοποιεί τις τεχνικές breadth-first search και την tree τεχνική με στόχο να υπολογίσει τα υποψήφια Item sets σε ικανοποιητικό βαθμό. Παράγει υποψήφια item sets με μέγεθος k από Item sets με μέγεθος k-1.

Στην συνέχεια «κόβει» τα υποψήφια item sets τα οποία δεν έχουν συχνή εμφάνιση.

1.4.2 Συσχετίσεις – MBA

Η πιο διαδεδομένη μέθοδος για παραγωγή αλληλοσυσχετίσεων και αλληλεξαρτήσεων αποτελεί η εύρεση κανόνων συσχέτισης. Το πρόβλημα εξαγωγής κανόνων συσχέτισης παρουσιάστηκε αρχικά το 1993 ως μια προσπάθεια εξαγωγής χρήσιμων συσχετισμών μεταξύ των πεδίων μιας βάσης δεδομένων. [4], [5]

Όπως αναφέραμε σε προηγούμενο κεφάλαιο η πιο συνηθισμένη εφαρμογή της μεθόδου της συσχέτισης είναι η «Ανάλυση του καλαθιού της νοικοκυράς». (market basket analysis).

Σκοπός είναι αν αναγνωρισθούν τα αγαθά τα οποία αγοράστηκαν μαζί. Συγκεκριμένα ένας κανόνας συσχέτισης θα μπορούσε να πει ότι το γάλα πωλείται μαζί με το τυρί, με την προφανή αξιοποίηση της πληροφορίας που είναι η τοποθέτηση και των δυο αυτών προϊόντων στο ίδιο σημείο πώλησης.

Μερικές συνήθεις πρακτικές εφαρμογές τους είναι η εύρεση προϊόντων που πωλούνται μαζί σε μια συναλλαγή, η επεξεργασία ερωτηματολογίων, η εύρεση των προϊόντων που διακινούνται μαζί σε μια αποθήκη για πρόβλεψη προμήθειας καθώς και η εύρεση των λέξεων που συναντώνται μαζί σε ένα κείμενο.

¨Όπως βέβαια είπαμε και προηγούμενα, ένας κανόνας συσχέτισης είναι μια έκφραση της μορφής $X \Rightarrow Y$, όπου Χ και Υ είναι σύνολα τιμών των πεδίων, όπως για παράδειγμα σύνολα προϊόντων. Η σπουδαιότητα ενός κανόνα συσχέτισης καθορίζεται αναλογικά από το ποσοστό εφαρμογής του κανόνα επί του συνόλου εκπαίδευσης.

Συγκεκριμένα οι αλγόριθμοι συσχέτισης που έχουν προταθεί και εφαρμόζονται πρακτικά, εξάγουν κανόνες συσχέτισης της μορφής: «Το 98% των πελατών που αγοράζουν γάλα και κρέας αγοράζουν επίσης και τυρί. Αλλά στο 70% των αγορών έχουν αγορασθεί γάλα και τυρί και κρέας». Το πρώτο ποσοστό αναφέρεται ως **αξιοπιστία (confidence)** του κανόνα ενώ το δεύτερο ως **επιβεβαίωση (support).**

Η επιβεβαίωση αφορά στο ποσοστό που εμφανίζονται και τρία αγαθά μαζί επί του συνόλου εκπαίδευσης ενώ η αξιοπιστία αφορά στο ποσοστό που εμφανίζονται τα αγαθά επί του αριθμού αγορών που περιέχουν γάλα και κρέας. Το πρόβλημα εύρεσης κανόνων συσχέτισης εστιάζεται στην εύρεση όλων των κανόνων που έχουν μια καθορισμένη από τον χρήστη ελάχιστη τιμή επιβεβαίωσης και αξιοπιστίας.

1.5 Apriori αλγόριθμος

1.5.1 Εισαγωγή

Η παρουσίαση του αλγορίθμου Apriori, οι συμβολισμοί, οι ορισμοί καθώς και τα παραδείγματα που χρησιμοποιούμε προέρχονται από την πηγή [12], [13].

Ο αλγόριθμος Apriori δέχεται ως είσοδο ένα σύνολο αγορών (transactions) που αποτελεί και το σύνολο εκπαίδευσης. Κάθε αγορά είναι ουσιαστικά μια λίστα (itemset) από αγαθά (items) τα οποία αγοράστηκαν μαζί. [14]

1.5.2 Ορισμοί

- *Itemset*: Σύνολο από αντικείμενα – αγαθά (items).

- *k-itemset*: Σύνολο από k αντικείμενα- αγαθά

- *Σύνολο από k-itemsets*: Σύνολο από k -άδες items. Το σύνολο αυτό περιέχει υποσύνολα, όπου κάθε υποσύνολο περιέχει k - items το καθένα.
  π.χ. Ένα 3-itemset: {I1, I2, I3}
  Σύνολο από 3-itemsets: { {I1, I2, I3} {I1, I2, I5} {I2, I4, I5} }

- *minimum support threshold (min_sup)*: Κατώτερο όριο το οποίο πρέπει να ικανοποιούν τα itemsets για να είναι *frequent*. Ένα itemset ικανοποιεί το κατώφλι και είναι frequent, αν ο αριθμός εμφανίσεων του στην Βάση Δεδομένων είναι μεγαλύτερος ή ίσος από το κατώτερο όριο αυτό.

- *frequent itemset*: Ένα itemset I είναι frequent, αν ο αριθμός εμφανίσεων του στην Βάση Δεδομένων είναι μεγαλύτερος ή ίσος από το κατώτερο όριο του minimum support. (Αυτό δηλώνεται από το L_i για το i-στο itemset). $P(I) \geq min\_sup$

- *Σύνολο Lk:* σύνολο από frequent k-itemsets

- *Σύνολο Ck:* σύνολο από υποψήφια frequent k-itemsets

- *Apriori Property:* Κάθε υποσύνολο των frequent itemset πρέπει να είναι frequent.

- *Join Operation- λειτουργία Συνένωσης:* Εύρεση του L_k, του συνόλου δηλαδή των παραγώγων k-itemsets το οποίο προκύπτει από την συνένωση των L_{k-1} μεταξύ τους.

1.6 Μια γενική εισαγωγή στον αλγόριθμο Apriori

Πρόκειται για τον βασικό αλγόριθμο για εύρεση *frequent itemsets*. Τα frequent itemsets μας είναι χρήσιμα, επειδή από αυτά προκύπτουν με κατάλληλες μεθόδους οι κανόνες συσχέτισης. Επίσης, υπάρχουν και αλγόριθμοι οι οποίοι αποτελούν βελτιώσεις του Apriori. Στα πλαίσια της διπλωματικής αυτής, οι αλγόριθμοι αυτοί δεν μελετήθηκαν, αφού ο Apriori αλγόριθμος μπορεί να μας δώσει τα επιθυμητά αποτελέσματα. [16]

Το πρόβλημα της εύρεσης των κανόνων συσχέτισης που έχουν την επιθυμητή επιβεβαίωση και αξιοπιστία μπορεί να διαιρεθεί σε δυο υπο-προβλήματα:

- Εύρεση όλων των συνδυασμών των προϊόντων που έχουν επιβεβαίωση πάνω από την ελάχιστη επιβεβαίωση (minimum support). 'Ολοι αυτοί οι συνδυασμοί ονομάζονται μεγάλες λίστες από προϊόντα (large itemsets) και όλοι οι υπόλοιποι συνδυασμοί μικρές λίστες από προϊόντα(small itemsets).

- Χρήση όλων των μεγάλων λιστών από προϊόντα για εξόρυξη των κανόνων συσχέτισης που ικανοποιούν την ελάχιστη αξιοπιστία. Για παράδειγμα, έστω τα ABED και AB είναι μεγάλες λίστες από προϊόντα. Μπορούμε να καθορίσουμε αν ο κανόνας συσχέτισης $AB \Rightarrow CD$ ξεπερνά την ελάχιστη αξιοπιστία, υπολογίζοντας το λόγο:

  r= επιβεβαίωση (ABCD)/ επιβεβαίωση (AB)

Αν r $\geq$ελάχιστη αξιοπιστία, τότε ο κανόνας συσχέτισης γίνεται αποδεκτός.

Η εύρεση των μεγάλων λιστών από προϊόντα, για να αποφύγει κανείς ένα εξαντλητικό ψάξιμο όλων των δυνατών συνδυασμών, βασίζεται στο ότι: μια λίστα από προϊόντα είναι μεγάλη λίστα από προϊόντα αν κάθε υποσύνολό της είναι μεγάλη λίστα από προϊόντα.

Ο Apriori προτάθηκε από τους R .Agrawal και R .Srikant το 1994. Στόχος τους ήταν η εξόρυξη frequent itemsets για Boolean κανόνες συσχέτισης. Το όνομα βασίζεται στο γεγονός ότι ο αλγόριθμος χρησιμοποιεί «προηγούμενη γνώση» (prior knowledge) ιδιοτήτων των frequent itemsets. Επίσης, ο αλγόριθμος χρησιμοποιεί την προσέγγιση level - wise search, κατά την οποία k-itemsets χρησιμοποιούνται για την εύρεση k+1-itemsets.

Ο αλγόριθμος Apriori είναι ένας πολύ σημαντικός αλγόριθμος ο οποίος χρησιμοποιείται στην εξόρυξη των απλούστερων μορφών από frequent patterns – itemsets, με στόχο την εξαγωγή κανόνων συσχέτισης.

Ξεκινώντας λοιπόν καλό θα ήταν να κάνουμε μια πολύ σύντομη παρουσίαση στον τρόπο που λειτουργεί ο αλγόριθμος.

1. Βρίσκουμε τα αγαθά που εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση (minimum support), δηλαδή το σύνολο L_1=μεγάλες λίστες από 1 – αγαθό (large 1 item sets)
2. Από κ=2 και όσο L_{k-1} δεν είναι κενό κάνε:
   - Βρες το σύνολο C_k όλων των υποψήφιων μεγάλων λιστών από κ-αγαθά (candidate large k-itemsets) με βάση το L_{k-1} .

- Βρες ποια από αυτά εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση και φτιάξε το σύνολο L_k = μεγάλες λίστες από κ-αγαθά.

3. Για κάθε στοιχείο των L_1, L_2, L_3,..... L_n βρες ποια ικανοποιούν την ελάχιστη αξιοπιστία (minimum confidence).
4. Στην συνέχεια χρησιμοποιούμε αυτά τα frequent itemset για την παραγωγή κανόνων συσχέτισης.

1.6.1 Πιο συγκεκριμένα.....

Βρίσκουμε το σύνολο των frequent 1-itemsets (k=1) που ικανοποιούν τον περιορισμό του min support, με ψάξιμο (scanning) όλης της βάσης δεδομένων και μέτρηση των εμφανίσεων τους. Το αποτέλεσμα είναι το σύνολο L_1.

Βρίσκουμε το σύνολο των frequent 2-itemsets (k=2) χρησιμοποιώντας το σύνολο L_2. Το αποτέλεσμα το ονομάζουμε σύνολο L_2.

Ακολουθώντας την ίδια διαδικασία, βρίσκουμε τελικά το σύνολο $L_κ$ χρησιμοποιώντας το σύνολο $L_{κ-1}$.
Για την εύρεση κάθε L_n με n=1…k, χρειάζεται ένα πλήρες ψάξιμο (full scanning) όλης της Βάσης Δεδομένων.

Ο αλγόριθμος βασίζεται σε δύο βήματα: το join step (βήμα συνένωσης) και το pruning step (βήμα κλαδέματος). Το σύνολο $L_κ$ προκύπτει μετά από τις διαδικασίες join και pruning δύο συνόλων $L_{κ-1}$ και $L_{κ-1}$.

Βήμα 1: join step

Όπως είπαμε και προηγουμένως, το σύνολο $L_κ$ θα βρεθεί από το σύνολο $L_{κ-1}$. Για να βρεθεί το $L_κ$, πρέπει πρώτα να βρεθεί το σύνολο C_k των *υποψήφιων itemsets*. Το σύνολο C_k θα βρεθεί αφού γίνει joined το $L_κ$ με τον εαυτό του.
$L_{κ-1}$. *join* $L_{κ-1}$.: Γίνεται μόνο στα itemsets του $L_{κ-1}$. που είναι *joinable*.

Joinable: Δύο k-1-itemsets είναι *joinable*, εάν *τα items τους μέχρι και το k-2 είναι τα ίδια*.

Το k-itemset αυτό αποτελεί υποψήφιο k-itemset για το σύνολο $L_κ$ και θα μπει στο σύνολο $C_κ$ των υποψήφιων k-itemsets.

Παράδειγμα για k=4: Για να βρούμε το L_4 από το L_3 θα κάνουμε την πράξη L_3 join L_3: για κάθε itemset στο L_3 κοιτάμε αν είναι joinable με κάθε άλλο itemset στο L_3 (εκτός τον εαυτό του), δηλαδή αν έχουν τα ίδια items στις πρώτες 2 θέσεις. Αν δύο itemsets είναι joinable, τότε γίνονται joined και το 4-itemset που προκύπτει εισάγεται στο σύνολο C_4, αφού αποτελεί υποψήφιο 4-itemset για το L_4.

Βήμα 2: pruning step

Το σύνολο $C_κ$ από k-itemsets είναι υπερσύνολο του $L_κ$, το οποίο θέλουμε να βρούμε. Ισχύει ότι: τα k-itemsets-μέλη του Ck μπορεί *να είναι η να μην είναι frequent*, αλλά ***όλα*** *τα frequent k-itemsets συμπεριλαμβάνονται στο* $C_κ$. Το ποια από τα itemsets αυτά είναι frequent και ποια όχι, καθορίζεται με ένα ψάξιμο στην Βάση Δεδομένων για κάθε itemset και μέτρηση του πλήθους των εμφανίσεων του κάθε ενός από αυτά. *Frequent* είναι μόνο τα itemset που έχουν *αριθμό εμφανίσεων μεγαλύτερο ή ίσο του minimum support.*

Επειδή το C μπορεί να είναι τεράστιο, γίνεται μια διαδικασία ελαχιστοποίησης των itemsets του. Χρησιμοποιείται η *Apriori ιδιότητα*:

«Οποιοδήποτε k-1-itemset δεν είναι frequent, δεν μπορεί να είναι υποσύνολο κάποιου frequent k-itemset»

Έτσι, αν οποιοδήποτε από τα k-1-itemsets κάποιου υποψήφιου k-itemset στο $C_κ$ δεν υπάρχει στο $L_{κ-1}$ τότε το υποψήφιο k-itemset δεν μπορεί να είναι frequent και αφαιρείται από το $C_κ$. Με τον τρόπο αυτό, για κάθε τέτοιο k-itemset (που σίγουρα δεν είναι frequent), γλιτώνουμε την προσπέλαση στον πίνακα συναλλαγών και την μέτρηση του αριθμού των εμφανίσεων του, δηλαδή τις πράξεις που θα κάναμε για να δούμε αν θα ικανοποιούσε το minimum support.

Να τονίσουμε ότι με την pruning διαδικασία *δεν* προκύπτουν τα frequent k-itemsets, αλλά τα k-itemsets που *αποκλείεται* να είναι frequent. Έτσι, μετά το pruning, χρειάζεται η προσπέλαση της βάσης δεδομένων για κάθε ένα από τα εναπομείναντα itemsets στο Ck, ώστε να καθοριστεί αν ικανοποιούν το minimum support. Μόνο αν το ικανοποιούν θα είναι frequent itemsets.

1.6.2 Παράδειγμα εφαρμογής των κανόνων συσχέτισης.

Έστω ότι ο χρήστης θέτει ελάχιστη επιβεβαίωση ίση με 50% και ελάχιστη αξιοπιστία 70%.

Φτιάχνεται μια λίστα με τα προϊόντα τα οποία πωλήθηκαν σε μια αγοραπωλησία.
Έστω L_1={γάλα 100%, τυρί 75%, κρέας 50%}

Στην 1η επανάληψη του 2ου βήματος, οι υποψήφιες μεγάλες λίστες με δυο προϊόντα είναι: C_2={γάλα – τυρί 75%, γάλα – κρέας 50%, τυρί – κρέας 25%}.
Μετρώντας την επιβεβαίωση αυτή προκύπτει: L_2={ γάλα – τυρί, γάλα – κρέας}
Στην 2η επανάληψη του 2ου βήματος , οι υποψήφιες μεγάλες λίστες με τρία προϊόντα είναι: C_2={γάλα – τυρί – κρέας 25%}.

Από την επιβεβαίωση αυτής προκύπτει ότι L_3={}. Οπότε και οι επαναλήψεις σταματούν.

Στο τρίτο βήμα λοιπόν από όλους τους δυνατούς κανόνες που προκύπτουν από το L_2={ γάλα – τυρί, γάλα – κρέας} ο πρώτος έχει αξιοπιστία 25% και ο δεύτερος 50%.

'Αρα συνεπάγεται μόνο ο κανόνας συσχέτισης {γάλα - τυρί}.

1.6.3 Παράδειγμα εφαρμογής του αλγορίθμου

'Εστω ο πίνακας συναλλαγών D σε ένα κατάστημα (πίνακας 1) με 9 συναλλαγές, δηλαδή |D|=9. Ο πίνακας συναλλαγών δείχνει ποια προϊόντα αγόρασαν οι πελάτες στις διάφορες συναλλαγές τους. Για παράδειγμα, στην εκατοστή συναλλαγή T100, έχουν αγοραστεί τα αντικείμενα I1, I2 και I5.

Ορίζουμε το minimum support ίσο με 2 (min_sup=2)

Ακόμα θεωρούμε ότι το minimum confidence είναι 70%.

Θα πρέπει πρώτα να βρούμε τα frequent item set με τη βοήθεια του Apriori Αλγόριθμου.

Στη συνέχεια οι κανόνες συσχέτισης θα παραχθούν με τη βοήθεια του min.support και του min.confidence.

Πίνακας 1

TID	LIST OF ITEMS
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Τα βήματα του αλγορίθμου είναι τα εξής:

Βήμα 1: Βρίσκουμε τα 1-itemsets τα οποία θα αποτελέσουν το σύνολο των υποψήφιων frequent item sets C_1. Για να γίνει αυτό, θα χρειαστεί προσπέλαση

όλου του πίνακα D και μέτρηση του αριθμού των εμφανίσεων του κάθε item. Τα αποτελέσματα παρουσιάζονται στον πίνακα 2.

Πίνακας 2 – C_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Βήμα 2: Το σύνολο των frequent itemsets L_1 θα αποτελούν τα itemsets του C_1, τα οποία ικανοποιούν το min_sup, δηλαδή που έχουν αριθμό εμφανίσεων μεγαλύτερο ή ίσο του 2 (εφόσον min_sup=2). Στο συγκεκριμένο παράδειγμα, όλα τα itemsets ικανοποιούν το minimum support, άρα είναι frequent itemsets και ανήκουν στο L_1. (πίνακας 3)

Πίνακας 3 – L_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Βήμα 3: Για την εύρεση των frequent 2-itemsets L2, θα χρησιμοποιήσουμε την πράξη *join* του L_1 με τον εαυτό του, δηλαδή L_1 *join* L_1, για να προκύψει έτσι το σύνολο των υποψήφιων 2-itemsets C2. Τα αποτελέσματα του L_1 join L_1 είναι τα itemsets που φαίνονται στον ακόλουθο πίνακα. (πίνακας 4)

Κατά την διαδικασία pruning, κάθε υποσύνολο των itemsets του C2 πρέπει να ελεγχθεί για το αν υπάρχει επίσης στο σύνολο L_1 (δηλαδή αν είναι frequent). Εφόσον όλα τα υποσύνολα (στην περίπτωση αυτή items) είναι frequent, τότε κατά την διαδικασία pruning, κανένα υποψήφιο itemset δεν διαγράφεται από το C2.

Πίνακας 4 – C_2

itemeset
{I1,I2}
{I1,I3}
{I1,I4}
{I1,I5}
{I2,I3}
{I2,I4}

{I2,I5}
{I3,I4}
{I3,I5}
{I4,I5}

Βήμα 4: Στο βήμα αυτό, γίνεται προσπέλαση του πίνακα συναλλαγών D με σκοπό την μέτρηση του αριθμού των εμφανίσεων των 2-itemsets του C_2. Τα αποτελέσματα φαίνονται στον πίνακα 5.

Πίνακας 5 - C_2

item set	Sup.count
{I1,I2}	4
{I1,I3}	4
{I1,I4}	1
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2
{I3,I4}	0
{I3,I5}	1
{I4,I5}	0

Βήμα 5: Στην συνέχεια καθορίζεται το σύνολο των frequent 2-itemsets L2, το οποίο αποτελείται από τα 2-itemsets του C2 που ικανοποιούν το minimum support με αριθμό εμφανίσεων μεγαλύτερο ή ίσο του 2. Ο πίνακας 6 δείχνει το σύνολο L2. (πίνακας 6)

Πίνακας 6 – L_2

item set	Sup. Count
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

Βήμα 6: Το σύνολο των υποψήφιων 3-itemsets C3 θα καθοριστεί με την διαδικασία L_2 *join* L_2. Υπενθυμίζουμε ότι τα item sets που μπορούν να γίνουν joined είναι τα *joinable itemsets*, δηλαδή αυτά των οποίων τα items μέχρι και το k-2 είναι τα ίδια. Στο παράδειγμα μας, για τα 3-itemsets έχουμε k=3, άρα τα joinable itemsets είναι αυτά που έχουν το πρώτο τους item το ίδιο. Κατά την διαδικασία join θα έχουμε:

$C_3 = L_2$ join L_2 = {{I1, I2, I3} {I1, I2, I5} {I1, I3, I5} {I2, I3, I4} {I2, I3, I5} {I2, I4, I5}}

Στην συνέχεια εφαρμόζουμε την *pruning* διαδικασία χρησιμοποιώντας την *Apriori ιδιότητα* κατά την οποία όλα τα υποσύνολα k-1-itemsets κάποιου frequent k-itemset πρέπει επίσης να είναι frequent και ελέγχουμε τα 3-itemsets του C_3. Εδώ τα υποσύνολα είναι 2- itemsets.

1) 3-itemset {I1, I2, I3}: {I1, I2}: Μέλος του L2 => frequent 2-itemset
{I2, I3}: Μέλος του L2 => frequent 2-itemset
{I1, I3}: Μέλος του L2 => frequent 2-itemset
'Αρα {I1, I2, I3}: πιθανό frequent 3-itemset
2) 3-itemset {I1, I2, I5}: {I1, I2}: Μέλος του L2 => frequent 2-itemset
{I2, I5}: Μέλος του L2 => frequent 2-itemset
{I1, I5}: Μέλος του L2 => frequent 2-itemset
'Αρα {I1, I2, I5}: πιθανό frequent 3-itemset
3) 3-itemset {I1, I3, I5}: {I1, I3}: Μέλος του L2 => frequent 2-itemset
{I3, I5}: *'ΟΧΙ* Μέλος του L2 => *'ΟΧΙ* frequent 2-itemset
{I1, I5}: Μέλος του L2 => frequent 2-itemset
'Αρα {I1, I2, I3}: ΜΗ πιθανό frequent 3-itemset
4) 3-itemset {I2, I3, I4}: {I2, I3}: Μέλος του L2 => frequent 2-itemset
{I3, I4}: *'ΟΧΙ* Μέλος του L2 => *'ΟΧΙ* frequent 2-itemset
{I2, I4}: Μέλος του L2 => frequent 2-itemset
'Αρα {I2, I3, I4}: ΜΗ πιθανό frequent 3-itemset
5) 3-itemset {I2, I3, I5}: {I2, I3}: Μέλος του L2 => frequent 2-itemset
{I3, I5}: *'ΟΧΙ* Μέλος του L2 => *'ΟΧΙ* frequent 2-itemset
{I2, I5}: Μέλος του L2 => frequent 2-itemset
'Αρα {I2, I3, I5}: ΜΗ πιθανό frequent 3-itemset
6) 3-itemset {I2, I4, I5}: {I2, I4}: Μέλος του L2 => frequent 2-itemset
{I4, I5}: *'ΟΧΙ* Μέλος του L2 => *'ΟΧΙ* frequent 2-itemset
{I2, I5}: Μέλος του L2 => frequent 2-itemset
'Αρα {I2, I4, I5}: ΜΗ πιθανό frequent 3-itemset

Μετά το τέλος της pruning διαδικασίας, προκύπτει ότι μόνο τα δύο πρώτα itemsets του συνόλου C_3 είναι frequent, ενώ τα υπόλοιπα 4 δεν είναι, άρα τα αφαιρούμε από το C_3, όπως φαίνεται και στον πίνακα 7.

Πίνακας 7 – C_3

Itemset
{I1, I2, I3}
{I1, I2, I5}

Το πλεονέκτημα της pruning διαδικασίας είναι ότι *γλιτώνουμε την προσπέλαση στον πίνακα συναλλαγών D* για την μέτρηση του αριθμού των εμφανίσεων των itemsets, τα οποία *αποκλείεται να είναι frequent*. Έτσι στο συγκεκριμένο παράδειγμα, γλιτώσαμε τέσσερις προσπελάσεις του πίνακα συναλλαγών.

Βήμα 7: Γίνεται προσπέλαση του πίνακα συναλλαγών D για να μετρηθεί ο αριθμός των εμφανίσεων των 3-itemsets του C_3. Όσα ικανοποιούν το minimum support, δηλαδή εμφανίζονται σε περισσότερες ή ίσες των δύο συναλλαγών, θα εισαχθούν στο σύνολο L3. Τα αποτελέσματα παρατίθενται στον πίνακα 8.

Πίνακας 8 – C_3

Itemset	Sup.count
{I1, I2, I3}	2
{I1, I2, I5}	2

Βήμα 8: Αφού και τα δύο 3-itemsets του συνόλου C3 ικανοποιούν το minimum support, τότε και τα δύο τοποθετούνται στο σύνολο L3, όπως φαίνεται και στον πίνακα 9.

Πίνακας 8 – L_3

Itemset	Sup.count
{I1, I2, I3}	2
{I1, I2, I5}	2

Βήμα 9: Στην συνέχεια ο αλγόριθμος θα επιχειρήσει την πράξη L_3 *join* L_3, για την δημιουργία του συνόλου των υποψηφίων 4-itemsets C_4. Όπου: C4 = L3 join L3 = {{I1, I2, I3, I5}} .

Στην συνέχεια, εκτελείται η *pruning διαδικασία*, κατά την οποία θα ελεγχθούν όλα τα υποσύνολα 3-itemsets του πιο πάνω υποψήφιου 4-itemset.

Έχουμε λοιπόν:
itemset {I1, I2, I3, I5}: {I1, I2, I3}: Μέλος του L_3 => frequent 3-itemset
{I1, I2, I5}: Μέλος του L_3 => frequent 3-itemset
{I1, I3, I5}: *'ΟΧΙ* Μέλος του L_3 => *'ΟΧΙ* frequent 3-itemset
{I2, I3, I5}: *'ΟΧΙ* Μέλος του L_3 => *'ΟΧΙ* frequent 3-itemset
'Αρα {I1, I2, I3, I5}: Μη πιθανό frequent 4-itemset

Όπως παρατηρούμε, τα 3-itemset: {I1, I3, I5} και {I2, I3, I5} δεν περιλαμβάνονται στο σύνολο L_3, άρα δεν είναι frequent.

'Αρα το σύνολο C_4 παραμένει κενό C_4=0 και ο αλγόριθμος τερματίζει, έχοντας βρει όλα τα frequent itemsets L_1, L_2 και L_3.

1.6.4 Ο Ψευδοκώδικας του Apriori Αλγόριθμου

The Apriori Algorithm: Pseudo code [11], [12]

•Join Step: Ck is generated by joining Lk-1with itself

•Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

•Pseudo-code:
Ck: Candidate itemset of size k
Lk: frequent itemset of size k
L1= {frequent items};
For (*k*= 1; *Lk*!=□; *k*++) do begin
Ck+1= candidates generated from *Lk*;
For each transaction *t* in database do increment the count of all candidates in *Ck+1*that are contained in *t*
Lk+1= candidates in *Ck+1*with min_support
end
return□*kLk*;

1.7 Δημιουργία Κανόνων Συσχέτισης από Frequent Itemsets

1.7.1 Εισαγωγή

Η εξόρυξη κανόνων συσχέτισης ερευνά για σχέσεις μεταξύ item sets – αντικειμένων σε μια βάση δεδομένων. Κύριος στόχος είναι η εύρεση συσχετίσεων, συνδέσεων ή δομών συνάφειας ανάμεσα σε Items ή αντικείμενα σε μια βάση δεδομένων συναλλαγών, σχέσεων ή και άλλων «αποθηκών » πληροφορίας.

Ο κανόνας που χρησιμοποιείται είναι ο ακόλουθος:
$Body \rightarrow Head(\sup port, confidence)$
Για παράδειγμα: Κάποιος αγοράζει (x, ψωμί) →αγοράζει (x, γάλα) [0,6%, 65%]

1.7.2 Ορισμοί κανόνων και συνόλων

Βασικά Σύνολα των κανόνων συσχέτισης:

- Έστω Α μια συναλλαγή η οποία αποτελεί ένα σύνολο Items: $T = \{i_a, i_b, ..., i_t\}$. Το σύνολο Τα είναι ένα υποσύνολο του συνόλου I με $I = \{i_a, i_b, ..., i_n\}$.
- Το σύνολο D αποτελεί ένα σύνολο συναλλαγών.

Κανόνες συσχέτισης:

- $P \to Q \ \ με \ P \subset I, Q \subset I \ \ και \ \ P \cap Q = \varnothing$
- $P \to Q$ να ισχύει στο σύνολο συναλλαγών D με «support» s
- $P \to Q$ να ισχύει στο σύνολο συναλλαγών D με «confidence» c
- $\sup port(P \to Q) = πιθανότητα(P \cup Q)$
- $confidence(P \to Q) = πιθανότητα(Q / P)$

Item sets

- Ένα σύνολο items – κανόνων καλείται itemset.
- Κάθε Itemset αποτελείται από k-items και καλείται k-itemset.
- Ένα itemset μπορεί επίσης να θεωρηθεί και ένας συνδυασμός από αντικείμενα- items.

Support και Confidence (Υποστήριξη και εμπιστοσύνη)

- **Support** των $P = P_1 \cap P_2 \cap P_3 ..... \cap P_n$ στο σύνολο των συναλλαγών D.
  Το $\sigma(P / D)$ είναι το ποσοστό των συναλλαγών Τα που συμβαίνουν στο σύνολο D με ποσοστό P.

- **Confidence** του κανόνα $P \to Q$
  Το $\phi(P \to Q / D)$ είναι η αναλογία $\sigma((P \cap Q) / D)$ με support $\sigma(P / D)$.

Κατώτερα όρια

- **Minimum support σ**
- **Minimum Confidence φ**

1.7.3 Παράδειγμα – βασίζεται στις παραπάνω έννοιες

Για να απεικονίσουμε τις παραπάνω έννοιες θα χρησιμοποιήσουμε το παρακάτω απλό παράδειγμα βασιζόμενοι σε μια συναλλαγή σε ένα σούπερ μάρκετ.

Έστω το σύνολο των αντικειμένων I={γάλα, ψωμί, βούτυρο, μπύρα} και μια μικρή βάση δεδομένων η οποία περιέχει τα items (1 →παρουσία του αντικειμένου στην συναλλαγή και 0→απουσία του αντικείμενου).

Ένας κανόνας λοιπόν ο οποίος μπορεί αν προκύψει είναι ο ακόλουθος: {ψωμί, γάλα}→{βούτυρο}. Αυτό σημαίνει ότι αν ένας πελάτης αγοράσει ψωμί και γάλα ταυτόχρονα είναι πολύ πιθανό να αγοράσει και βούτυρο.

Ισχυροί Κανόνες

- Συχνοί (μεγάλοι) κανόνες – κατηγορήματα P σε ένα σύνολο δεδομένων D.

- Κανόνας $P \rightarrow Q \ \ (c\%)$. Είναι ένας ισχυρός κανόνας .
  Το κατηγόρημα $P \cap Q$ είναι συχνό(μεγάλο)
  Το c είναι μεγαλύτερο από το Minimum confidence.

1.8 Μεθοδολογία εξόρυξης κανόνων συσχέτισης.

Αρχικά θεωρούμε μια μεγάλη βάση δεδομένων ενός συνόλου συναλλαγών.
Κάθε συναλλαγή είναι μια λίστα από αντικείμενα – items (για παράδειγμα μια αγοραπωλησία ενός πελάτη).

Στη συνέχεια βρίσκουμε όλους τους κανόνες οι οποίοι συσχετίζουν τα δεδομένα που παρουσιάζονται στο ένα σύνολο αντικειμένων με τα δεδομένα που παρουσιάζονται στο άλλο σύνολο δεδομένων.

Δεν υπάρχουν περιορισμοί ως προς τον αριθμό των αντικειμένων στο support ή στο confidence ενός κανόνα.

Εξόρυξη των Frequent itemsets [17]

1. Βρίσκουμε τα frequent itemsets, δηλαδή τα σύνολα των αντικειμένων τα όποια έχουν minimum support. Ένα υποσύνολο ενός frequent itemset πρέπει επίσης να είναι frequent itemset. Για παράδειγμα ένα το {AB} είναι ένα frequent itemset, τότε πρέπει και το{A} και το {B} να αποτελούν από μόνα τους ένα frequent itemset.

2. Έχοντας βρει τα frequent itemsets των συναλλαγών μιας βάσης δεδομένων (ή πίνακα συναλλαγών), μπορούμε πλέον να βρούμε διάφορους κανόνες συσχέτισης μεταξύ των itemsets αυτών. Η διαδικασία για τον σκοπό αυτό περιγράφεται στο [11]. Οι κανόνες αυτοί λέγονται *strong association rules* επειδή ικανοποιούν και το *minimum support* και το *minimum confidence*.

Οι κανόνες συσχέτισης προκύπτουν με χρήση της εξίσωσης:
Confidence (A→B) = P (B|A) = support_count (A □ B) / support_count (A) (1)

Υπενθυμίζουμε τα εξής:

Το **confidence factor** είναι ένα μέτρο δύναμης του κανόνα που δείχνει σε *τι* ποσοστό ισχύει το συνεπακόλουθο itemset, εφόσον ισχύουν τα προηγηθέντα itemsets, ενώ το **support** προορίζεται για στατιστική χρήση και περιλαμβάνει το ποσοστό των συναλλαγών που ικανοποιούν τον κανόνα, στο σύνολο όλων των συναλλαγών.

A→B: Δεδομένης της αγοράς του προϊόντος Α, υπάρχει μεγάλη πιθανότητα να έχει αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για το προϊόν Β.

Η εξίσωση 1 δείχνει σε τι ποσοστό των συναλλαγών ισχύει το itemset B, εφόσον ισχύει το itemset A. Όπως βλέπουμε, αυτό ισοδυναμεί με την πιθανότητα να βρεθεί το itemset B σε μια συναλλαγή, δεδομένου ότι βρέθηκε το itemset A. Η πιθανότητα αυτή, εκφράζεται στην συνέχεια περιλαμβάνοντας στην εξίσωση το support count των itemsets, όπου support_count($A \cup B$) είναι το σύνολο των συναλλαγών που περιλαμβάνουν το itemset $A \cup B$, ενώ support_count(A) είναι το σύνολο των συναλλαγών που περιλαμβάνουν το itemset A.

Με βάση την προηγούμενη εξίσωση, οι κανόνες συσχέτισης προκύπτουν ως εξής:
Για κάθε frequent itemset l, δημιούργησε όλα τα μη κενά υποσύνολα του l.
Για κάθε μη κενό υποσύνολο s του l, παρήγαγε τον κανόνα:
s → (l – s) (2)
εάν support_count(l)/support_count(s) ≥ min_conf
όπου min_conf είναι το κατώτερο όριο minimum confidence. Εφόσον οι κανόνες έχουν δημιουργηθεί από frequent itemsets, τότε εξορισμού ικανοποιούν το κατώτερο όριο του minimum support . Με τον πιο πάνω περιορισμό, ικανοποιούν και το minimum confidence, οπότε έχουμε την παραγωγή strong association rules, οι οποίοι ικανοποιούν και το minimum support και το minimum confidence.

1.8.1 Εφαρμογή

Ας επιστρέψουμε στο προηγούμενο παράδειγμα των 9 συναλλαγών.

Τα frequent itemsets που προέκυψαν από την εφαρμογή του Apriori αλγορίθμου είναι τα L1, L2 και L3.

Πίνακας 3 – L_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Πίνακας 6 – L_2

itemeset	Sup. Count
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

Πίνακας 8 – L_3

Itemset	Sup.count
{I1, I2, I3}	2
{I1, I2, I5}	2

Θα εξάγουμε κανόνες συσχέτισης με βάση τα frequent itemsets αυτά.
Η διαδικασία είναι η ακόλουθη:

- Για κάθε frequent itemset l , να παραχθούν όλα τα με κενά υποσύνολα του l.
- Για κάθε μη κενό υποσύνολο s του l, να παραχθεί ο κανόνας $s \rightarrow l - s$ σε περίπτωση που το
  support_count (l) / support_count (s) = min_conf (σχέση 1)
  όπου το min_conf είναι το ελάχιστο κατώτερο όριο εμπιστοσύνης.

Είχαμε συνολικά το:
L = {{I1}, {I2}, {I3}, {I4}, {I5}, {I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I1,I2,I3}, {I1,I2,I5}}.
Θα χρησιμοποιήσουμε το frequent itemset **l = {I1, I2, I5}**.
Τα μη κενά υποσύνολα του L είναι: {I1, I2}, {I2, I5}, {I1,I5}, {I1}, {I2} και {I5}
Οι κανόνες συσχέτισης που προκύπτουν σύμφωνα με την σχέση 1 , είναι:

Για **s={I1, I2}** έχουμε: {I1, I2} → I5
Confidence = support_count (l) / support_count (s) = 2 / 4 = 50%

Για **s={I1, I5}** έχουμε: {I1, I5} → I2

Confidence = support_count (l) / support_count (s) = 2 / 2 = 100%

Για **s={I2, I5}** έχουμε: {I2, I5} → I1
Confidence = support_count (l) / support_count (s) = 2 / 2 = 100%

Για **s=I1** έχουμε: I1 → {I2, I5}
Confidence = support_count (l) / support_count (s) = 2 / 6 = 33%

Για **s=I2** έχουμε: I2 → { I1, I5}
Confidence = support_count (l) / support_count (s) = 2 / 7 = 29%

Για **s=I5** έχουμε: I5 → { I1, I2}
Confidence = support_count (l) / support_count (s) = 2 / 2 = 100%

Αν για παράδειγμα το κατώφλι **minimum confidence** είναι **70%,** τότε μόνο ο δεύτερος, τρίτος και τελευταίος κανόνας το ικανοποιούν, αφού έχουν confidence support πάνω από το κατώφλι, άρα μόνο αυτοί χαρακτηρίζονται ως strong association rules. Έτσι, το αποτέλεσμα της εφαρμογής του Apriori αλγορίθμου στο παράδειγμά μας είναι οι εξής κανόνες συσχέτισης:

1. {I1, I5} → I2

   Δεδομένης της αγοράς των προϊόντων I1 και I5, υπάρχει πιθανότητα 1 (confidence=100%) να έχει αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για το προϊόν I2.

{I2, I5} → I1

Δεδομένης της αγοράς των προϊόντων I2 και I5, υπάρχει πιθανότητα 1 (confidence=100%) να έχει αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για το προϊόν I1.

I5 → { I1, I2}

Δεδομένης της αγοράς του προϊόντος I5, υπάρχει πιθανότητα 1 (confidence=100%) να έχουν αγοραστεί ή να υπάρχει μεγάλο ενδιαφέρον για τα προϊόντα I1 και I2.

Ένας περιορισμός στην πρότυπη αυτή προσέγγιση συσχετίσεων είναι ότι ψάχνοντας ανάμεσα σε τεράστιο αριθμό συσχετίσεων για την εύρεση συνόλων αντικειμένων τα οποία πιθανόν να συσχετίζονται μεταξύ τους, υπάρχει μεγάλος κίνδυνος για την εύρεση πολλών λανθασμένων συσχετίσεων.

Τέτοιου είδους συσχετίσεις αποτελούν οι περιπτώσεις όπου οι συλλογές από αντικείμενα συνυπάρχουν με απροσδόκητη συχνότητα στα δεδομένα και αυτό συμβαίνει μόνο κατά τύχη.

Για παράδειγμα, ας υποθέσουμε ότι εξετάζουμε μια συλλογή από 10.000 αντικείμενα και ψάχνουμε για τους κανόνες που περιέχουν δύο στοιχεία στην

αριστερή πλευρά και 1 στοιχείο στη δεξιά πλευρά. Υπάρχουν περίπου 1.000.000.000.000 τέτοιοι κανόνες.

Αν εφαρμόσουμε ένα στατιστικό τεστ για την ανεξαρτησία, με επίπεδο σημαντικότητας της τάξης του 0,05 αυτό αυτόματα σημαίνει ότι υπάρχει μόνο μια 5% πιθανότητα αποδοχής του κανόνα, αν δεν υπάρχει συσχέτιση.

Αν υποθέσουμε ότι δεν υπάρχουν συσχετίσεις, θα πρέπει να περιμένουμε το λιγότερο 50.000.000.000 κανόνες.

Στατιστικά πάντα η εύρεση συσχετίσεων [15] [16] κοντρολάρει αυτό το ρίσκο και στις περισσότερες περιπτώσεις μειώνει τον κίνδυνο για πιθανή διαπίστωση λανθασμένων συσχετίσεων.

1.9 Άλλοι αλγόριθμοι για την εύρεση κανόνων συσχέτισης.

Πολλοί είναι οι Αλγόριθμοι οι οποίοι εμφανίστηκαν τα τελευταία χρόνια και οι οποίοι ασχολούνται με την παραγωγή κανόνων συσχέτισης. Κάποιοι πολλοί γνωστοί είναι οι Apriori, Eclat and FP-Growth, οι οποίοι βέβαια κάνουν το μισό της εργασίας που απαιτείται καθώς είναι αλγόριθμοι οι οποίοι απλώς εξάγουν frequent itemsets. Απαιτείται ακόμα ένα βήμα όπως αυτό της παραγωγής κανόνων από αυτά τα frequent itemsets τα οποία προκύπτουν από την βάση δεδομένων.

Ακολουθούν επιγραμματικά κάποιοι από αυτούς τους αλγόριθμους:

- **Apriori algorithm [12], [13], [14]**
- **Eclat algorithm**
- **FP-growth algorithm**
- **OPUS Search**
- **Zero-attribute-rule**
- **Lore**
- **GUHA procedure ASSOC**

Άλλοι τύποι εξαγωγής συσχετίσεων

- **Contrast set learning**
- **Weighted class learning**
- **Mining frequent sequences**
- **Generalized Association Rules**
- **Quantitiative Association Rules**

- **Interval Data Association Rules**
- **Maximal Association Rules**
- **Sequential Assocatiation Rules**

Κεφάλαιο 2ο

2.1 Εισαγωγή

Το Text mining [31] είναι ένα νεοσύστατο πεδίο που προσπαθεί να μαζέψει χρήσιμες πληροφορίες από το φυσικό γλωσσικό κείμενο. Γενικότερα μπορεί να θεωρηθεί ως διαδικασία ανάλυσης κειμένου με σκοπό την εξαγωγή πληροφοριών που είναι χρήσιμες για συγκεκριμένους σκοπούς.

Σε σύγκριση με το είδος των δεδομένων που αποθηκεύονται στις βάσεις δεδομένων, τα κείμενα είναι αδόμητα, άμορφα , και δύσκολο να ασχοληθεί κάποιος με αλγοριθμικές προσεγγίσεις.

Παρ 'όλα αυτά, στη σύγχρονη κουλτούρα, η γλώσσα αποτελεί το πιο κοινό μέσο για την επίσημη ανταλλαγή πληροφοριών. Στόχος του πεδίου αυτού, δηλαδή της «εξόρυξης δεδομένων από κείμενο», είναι να ασχοληθεί με κείμενα των οποίων η βασική λειτουργία είναι η ανακοίνωση των πραγματικών πληροφοριών ή απόψεων καθώς το κίνητρο για την προσπάθεια συλλογής πληροφοριών αυτομάτως είναι συναρπαστικό, ακόμα κι αν η επιτυχία είναι μόνο μερική.

Η φράση "text mining" χρησιμοποιείται γενικά για να υποδηλώσει κάθε σύστημα που αναλύει μεγάλες ποσότητες κειμένου φυσικής γλώσσας και ανιχνεύει λεξιλογικές ή γλωσσικές συνήθειες σε μια προσπάθεια για την εξαγωγή χρήσιμων πληροφοριών.

2.2 Εξόρυξη γνώσης από κείμενο

Ζούμε στην «Εποχή της πληροφορίας». Είναι πολύ γνωστή η ρήση του Συγγραφέα και μελλοντολόγου Gohn Naisbitt: «Πνιγόμαστε στις πληροφορίες αλλά διψάμε για γνώση». Το κύριο χαρακτηριστικό είναι η αύξηση των δεδομένων που παράγονται και αποθηκεύονται καθιστώντας πολύ δύσκολο να τα καταλάβουν οι άνθρωποι. Είναι γεγονός ότι οι δημόσιες και ιδιωτικές υπηρεσίες κατακλύζονται καθημερινά από πληροφορίες στις ογκώδεις βάσεις δεδομένων τους, οι οποίες έχουν έγγραφα ετερογενών μορφών και μπορεί να περιέχουν διαφορετικούς τύπους δεδομένων. (text, audio, image,video κτλ) και σε διαφορετικές γλώσσες. Έτσι και ο χρόνος και η προσπάθεια των υπαλλήλων σπαταλιούνται στις ατελείς αναζητήσεις μέσω των πολλαπλών πηγών πληροφοριών. Αυτό το πρόβλημα της μη υπερφόρτωσης πληροφοριών επιδεινώνεται περαιτέρω λόγω του μη δομημένου σχήματος και της διαφορετικής πλειοψηφίας των δεδομένων.

Ενώ το ποσό των κειμενικών στοιχείων που είναι διαθέσιμο σε μας αυξάνεται συνεχώς, η δυνατότητα μας να καταλάβουμε και να επεξεργαστούμε αυτές τις

πληροφορίες παραμένει σταθερή. Ένας ερευνητής μπορεί να αναγνωρίσει ότι ένα νέο γεγονός έχει εμφανιστεί μόνο αν ακολουθήσει άλλες κειμενικές πηγές. Αυτό είναι σαφώς ανεπαρκές για τον όγκο και την πολυπλοκότητα των σχετικών πληροφοριών. Η ανάγκη για την αυτοματοποιημένη εξαγωγή της χρήσιμης γνώσης από τεράστια ποσά κειμενικών στοιχείων προκειμένου να βοηθηθεί η ανθρώπινη ανάλυση είναι προφανής.

Η ανεύρεση γνώσης από κείμενο (knowledge discovery in text) και η άμεσα συνακόλουθη τεχνική εξόρυξη από κείμενο (text mining) [34] είναι οι πιο αυτοματοποιημένες τεχνικές που στοχεύουν στην ανακάλυψη πληροφοριών υψηλού επιπέδου στο τεράστιο ποσό κειμενικών στοιχείων και να τις παρουσιάσουν στον χρήστη.

Η ανεύρεση γνώσης από κείμενο (KDT) και το text mining (TM) είναι ένας νέος ερευνητικός τομέας ο οποίος προσπαθεί να επιλύσει το πρόβλημα της υπερφόρτωσης πληροφοριών με την χρησιμοποίηση των τεχνικών από την εξόρυξη από δεδομένα (data mining), την μηχανική μάθηση (machine learning), την επεξεργασία της φυσικής γλώσσας (natural language processing), την ανάκτηση πληροφορίας (Information retrieval), την εξαγωγή πληροφορίας (information extraction) και την διαχείριση γνώσης (Knowledge management).

Δεν υπάρχει παρόλα ταύτα κάποιο καθιερωμένο λεξιλόγιο για αυτές τι δυο έννοιες και μπορεί πολλές φορές να οδηγηθεί κάποιος στη σύγχυση σε κάποια προσπάθεια σύγκρισης αποτελεσμάτων. Θα κάνουμε μια προσπάθεια λοιπόν να ορίσουμε τις ακριβώς είναι το Text mining, που βρίσκει εφαρμογές, ποιες τεχνικές εμφανίζονται σε αυτό και ποια είναι τα χρήσιμα εργαλεία που χρησιμοποιεί αυτή η τεχνική.

2.3 Τι είναι το Text Mining

Αυτή την περίοδο το Text Mining απολαμβάνει ένα κύμα ενδιαφέροντος που τροφοδοτείται από την δημοτικότητα του Διαδικτύου, την επιτυχία της βιοπληροφορικής, και την αναγέννηση της Υπολογιστικής Γλωσσολογίας.

Το Text Mining είναι γνωστό επίσης και ως Text Data Mining [33] και ανεύρεση γνώσης σε κείμενο (Knowledge Discovery in Text) θα μπορούσαμε να το ορίσουμε ως μια διαδικασία εξαγωγής νέας – καινούριας πληροφορίας από μια συλλογή κειμένων. Με τον όρο νέες πληροφορίες εννοούμε τις συσχετίσεις, τις υποθέσεις ή τις τάσεις που δεν είναι ρητά παρούσες στην αρχική πηγή κειμένων που αναλύεται.

Βέβαια κατά καιρούς έχουν υπάρξει διαφορετικές προσεγγίσεις ως προς τον ορισμό του Text Mining.

Οι Karanikas, Theodoulidis [34] ορίζουν το Text Mining ως : «¨Ένα βήμα στην διαδικασία του KDT που αποτελείται από ιδιαίτερους αλγορίθμους του data mining και της επεξεργασίας της Φυσικής γλώσσας, που κάτω από μερικούς αποδεκτούς υπολογιστικούς περιορισμούς αποδοτικότητας, παράγουν έναν ιδιαίτερο αριθμό από υποδείγματα μέσα από ένα σύνολο μη δομημένων κειμενικών δεδομένων».

Οι Nahm και Mooney [35] περιγράφουν το Text Mining ως «την αναζήτηση των patterns σε μη δομημένο κείμενο», ενώ οι Besancon και Rajman [36] το θεωρούν ως μια επέκταση του γνωστού data mining σε μη δομημένα κείμενα και η οποία περιλαμβάνει διάφορους στόχους όπως εξαγωγή γνώσης και δημιουργία δομής βασισμένη στην ομοιότητα.

Σαν ορισμό του Text Mining θα μπορούσαμε να δώσουμε τον ακόλουθο ορισμό:
Το Text Mining είναι μια διαδικασία εξαγωγής νέας, έγκυρης και αγώγιμης γνώσης από διαφορετικούς γραπτούς πόρους, καθώς επίσης και όσο το δυνατόν καλύτερης οργάνωσης αυτής της νέας γνώσης της πληροφορίας για την όποια μελλοντική αναφορά.

Στόχος είναι να ανακαλυφθούν οι μέχρι τώρα άγνωστες και καλά κρυμμένες πληροφορίες, κάτι που κανένας ακόμα δεν ξέρει και δεν θα μπορούσε να γράψει κάτι για αυτό.

2.4 Text Mining & Data Mining

Ακριβώς όπως η εξόρυξη δεδομένων μπορεί να άνετα να περιγραφεί ως η αναζήτηση μοτίβων σε δεδομένα , η εξόρυξη κειμένου κατά αναλογία αναζητά για πρότυπα σε κείμενο. Ωστόσο, η επιφανειακή ομοιότητα μεταξύ των δύο κρύβει πραγματικές διαφορές. Η Εξόρυξη δεδομένων μπορεί πληρέστερα χαρακτηριστεί ως η εξόρυξη των «σιωπηρών», άγνωστων στο παρελθόν και ενδεχομένως χρήσιμων πληροφοριών από τα δεδομένα.[39],[40]

Η μέθοδος της εξόρυξης δεδομένων (data mining), δηλαδή της εύρεσης πολύτιμων μοτίβων/ προτύπων ανάμεσα στα δεδομένα, αποτελεί μια προφανή και πολύ καλή απάντηση για τη συλλογή και την αποθήκευση μεγάλου όγκου δεδομένων. Η εξόρυξη δεδομένων δεν είναι πλέον μια αναδυόμενη τεχνολογία εν αναμονή περαιτέρω ανάπτυξης. Παρόλο που η εφαρμογή της εξόρυξης δεδομένων δεν είναι καθολική και κοινώς αποδεκτή, οι τεχνικές εξόρυξης δεδομένων τυχαίνουν υψηλής ανάπτυξης και για κάποιες μορφές ανάλυσης εισέρχονται σε ώριμη φάση.

Οι πληροφορίες προκύπτουν έμμεσα από τα δεδομένα καθώς είναι καλά «κρυμμένες », άγνωστες , και δύσκολα θα μπορούσαν να εξαχθούν χωρίς την προσφυγή στις αυτόματες τεχνικές εξόρυξης δεδομένων.[30],[31],[32]. Σε αντίθεση η διαδικασία εξόρυξης δεδομένων σε κείμενο, είναι πιο προσιτή και εύκολη καθώς οι πληροφορίες που πρέπει να εξαχθούν αναφέρονται σαφώς και ρητώς στο κείμενο.

Το πρόβλημα, βέβαια, είναι ότι οι πληροφορίες δεν είναι διατυπωμένες κατά τέτοιο τρόπο ώστε να μπορούν να αποτελέσουν αντικείμενο αυτοματοποιημένης επεξεργασίας. Εδώ λοιπόν καλείται η διαδικασία εξόρυξης δεδομένων από κείμενο η οποία προσπαθεί φέρει το κείμενο στην κατάλληλη μορφή , για άμεση εφαρμογή των υπολογιστικών εφαρμογών.

Οι μέθοδοι της εξόρυξης δεδομένων βασίζονται και διδάσκονται από δείγματα της εμπειρίας του παρελθόντος. Αν μιλήσουμε σε ειδικούς προγνωστικής εξόρυξης δεδομένων, τα στοιχεία τους θα είναι σε αριθμητική μορφή. Αυτοί είναι γνωστοί ως και οι "'Ανθρωποι των Αριθμών". Οι "εξωρυκτές κειμένων" δεν αναμένουν μια εύρυθμη σειρά από αριθμούς, αλλά προτιμούν να κοιτάζουν συλλογές εγγράφων, όπου το περιεχόμενο είναι ευανάγνωστο και η σημασία τους είναι προφανής.

Αυτή είναι η πρώτη διάκριση / διαφορά μεταξύ της εξόρυξης δεδομένων και κειμένων: αριθμοί εναντίον κειμένου. Αυτό όμως δεν σημαίνει ότι πρόκειται για δύο διαφορετικές έννοιες. Και οι δύο βασίζονται σε δείγματα από παραδείγματα του παρελθόντος. Η σύνθεση των παραδειγμάτων είναι πολύ διαφορετική, αλλά πολλές από τις μεθόδους μάθησης είναι παρόμοιες. Αυτό συμβαίνει επειδή τα κείμενα υποβάλλονται σε επεξεργασία και μετατρέπονται σε μια αριθμητική εκπροσώπηση.

Οι παρουσιάσεις των δεδομένων για την κλασική εξόρυξη δεδομένων και την εξόρυξη κειμένων είναι αρκετά διαφορετικές. Ενώ οι μέθοδοι εξόρυξης δεδομένων προτιμούν τα δεδομένα να παρουσιάζονται σε μορφή υπολογιστικού φύλλου (Excel Spreadsheet), οι μέθοδοι εξόρυξης κειμένων προτιμούν τη μορφή κειμένου και κυρίως μια δημοφιλή εκδοχή κειμένων που ονομάζεται XML. 'Οπως είναι φυσικό, οι αριθμοί είναι πολύ διαφορετικοί από τα κείμενα. Παρόλα ταύτα, οι μέθοδοι που θα είναι παρεμφερείς με αυτούς της εξόρυξης δεδομένων. Οι μέθοδοι αυτές έχουν αποδειχθεί εξαιρετικά επιτυχείς, χωρίς την κατανόηση συγκεκριμένων ιδιοτήτων κειμένου όπως οι έννοιες της γραμματικής ή το νόημα των λέξεων. Πληροφορίες χαμηλής συχνότητας, όπως για παράδειγμα, το πόσες φορές εμφανίζεται μια λέξη σε ένα έγγραφο, χρησιμοποιούνται και στη συνέχεια εφαρμόζονται γνωστές μέθοδοι της μηχανικής μάθησης.

'Ενας από τους κύριους λόγους υποστήριξης της μεθόδου εξόρυξης κειμένου είναι η μετατροπή του κειμένου σε αριθμητικά δεδομένα, έτσι, αν και η αρχική

παρουσίαση είναι διαφορετική, σε κάποιο ενδιάμεσο στάδιο, τα αριθμητικά δεδομένα αποκωδικοποιούνται στη μορφή αυτών της εξόρυξης δεδομένων. Με αυτό τον τρόπο, τα αδόμητα δεδομένα μπορούν να διαρθρωθούν.

Οι μέθοδοι εξόρυξης κειμένων είναι παρόμοιες με κλασσικές μεθόδους εξόρυξης δεδομένων. Αυτές οι μέθοδοι θα μετατρέψουν τα στοιχεία από το κείμενο σε κοινές αριθμητικές μορφές. Για να λειτουργήσουν αυτές οι μέθοδοι, το κείμενο πρέπει να μετατραπεί σε μια τυποποιημένη μορφή λογιστικού φύλλου και όλα τα κενά του να συμπληρωθούν.

Οι γραμμές του λογιστικού φύλλου είναι παραδείγματα προηγούμενης εμπειρίας, άρα για το κείμενο μπορούμε να θεωρήσουμε ότι το έγγραφο είναι ένα ολοκληρωμένο παράδειγμα. Μια στήλη του λογιστικού φύλλου είναι μια ιδιότητα που μπορεί να μετρηθεί. Στο πιο θεμελιώδες μοντέλο ενός κειμένου, μπορούμε να θεωρήσουμε την παρουσία ή απουσία μιας λέξης να είναι ένα μετρήσιμο χαρακτηριστικό για κάθε έγγραφο. Έτσι, κάθε γραμμή αντιπροσωπεύει ένα έγγραφο και κάθε στήλη μια λέξη ή το αντίθετο.

Το μοντέλο λογιστικού φύλλου δεδομένων μας επιστρέφει στο οικείο έδαφος της κλασικής μεθόδου εξόρυξης δεδομένων. Παρ 'όλα αυτά, θα ήταν ανόητο να βιαστούμε να εφαρμόσουμε τις μεθόδους μάθησης στην αρχική τους μορφή και χωρίς να εκμεταλλευθούμε τα ιδιαίτερα χαρακτηριστικά του κειμένου. Το λογιστικό φύλλο παραμένει το εννοιολογικό μοντέλο, αλλά θα ήταν μη πρακτικό, μη αποδοτικό και ακόμα και αναποτελεσματικό μέχρι να καταλάβουμε κάποιες σημαντικές διαφορές του από τα κλασσικά αριθμητικά δεδομένα.

2.4.1 Μεθοδολογία

Ας θεωρήσουμε μια συλλογή εγγράφων. Το σύνολο των χαρακτηριστικών θα είναι το ολικό σύνολο των μοναδικών λέξεων σε αυτή τη συλλογή. Ονομάζουμε το σύνολο αυτών των λέξεων ως λεξικό. Τα παραδείγματα είναι τα ατομικά έγγραφα. Συνθέτουμε ένα λογιστικό φύλλο και συμπληρώνουμε τα κενά με τον αριθμό ένα για την παρουσία μια λέξεως και τον αριθμό μηδέν για την απουσία της. Η εφαρμογή μπορεί να έχει πολλές χιλιάδες ή ακόμα και εκατομμύρια έγγραφα. Το λεξικό θα συγκλίνει σε ένα μικρότερο αριθμό λέξεων από ότι ο αριθμός των εγγράφων, αλλά μπορεί εύκολα να φτάσει σε αριθμό αρκετές εκατοντάδες χιλιάδες. Εξειδικευμένα έγγραφα, όπως εγχειρίδια επισκευής με κωδικούς που είναι αλφαριθμητικοί, μπορεί να οδηγήσουν σε πολύ μεγάλα λεξικά. Φαίνεται ότι η μέθοδος του λογιστικού φύλλου είναι υπερβολικά επίπονη για να είναι πρακτική.

Κοιτάζοντας το λογιστικό φύλλο πιο κοντά, βλέπουμε σχεδόν παντού μηδενικά. Εκτός μεμονωμένων εγγράφων που είναι εκπληκτικά μεγάλα, σχεδόν στο μέγεθος ενός βιβλίου, ο λογιστικός πίνακας έχει πολλά «κελιά»: κάθε επιμέρους έγγραφο

θα χρησιμοποιεί μόνο ένα μικρό υποσύνολο όλων αυτών των λέξεων που βρίσκονται σε ένα λεξικό. Λόγω αυτού του ιδιαίτερου χαρακτηριστικού, το λογιστικό φύλλο παραμένει ένα λογικό εννοιολογικό μοντέλο δεδομένων. Οι μέθοδοι που επεξεργάζονται κείμενα θα αναμένουν αραιό-συμπληρωμένα λογιστικά φύλλα και θα αξιοποιήσουν αυτή την ιδιότητα, αποθηκεύοντας μόνο θετικές τιμές στα κενά.

Η αραιότητα δεν είναι η μόνη αντιπροσωπευτική διαφορά. Όλες οι τιμές σε ένα λογιστικό φύλλο εξόρυξης κειμένων είναι θετικές. Οι κλασσικές μέθοδοι εξόρυξης δεδομένων θα εξετάσουν όλες τις τιμές ενός χαρακτηριστικού, τόσο θετικές όσο και αρνητικές.

Το κριτήριο απόφασης θα μπορούσε εύκολα να είναι ότι "εάν η λέξη Χ έχει αξία μηδέν, τότε το συμπέρασμα είναι η έννοια Υ". Αντίθετα, οι μέθοδοι εξόρυξης κειμένων κυρίως επικεντρώνονται σε θετικό ταιριάσματα, χωρίς να απασχολούνται από το αν άλλες λέξεις απουσιάζουν από ένα έγγραφο. Αυτή η άποψη οδηγεί επίσης σε μεγάλη απλούστευση στον τομέα της μεταποίησης, συχνά επιτρέποντας προγράμματα εξόρυξης κειμένων να λειτουργούν σε περιπτώσεις με τεράστιες διαστάσεις δεδομένων για τις τακτικές εφαρμογής εξόρυξης δεδομένων.

Αν εστιαστούμε στις θετικές εμφανίσεις των λέξεων, έχουμε επίσης μια λύση σε ένα από τα bκte noires της εφαρμογής μεθόδων εξόρυξης δεδομένων: ελλείπουσες τιμές. Το μοντέλο λογιστικού φύλλου δεδομένων έχει ένα «κελί» για κάθε μετρήσιμη αξία σε κάθε παράδειγμα. Οι περισσότερες μέθοδοι αναμένουν κάθε «κελί» να έχει μια αξία.

Σε πρακτικές εφαρμογές στην καθημερινή ζωή, όπως όταν εξάγουμε πληροφορίες από μια βάση δεδομένων, πολλές πληροφορίες λείπουν και παραμένουν πολλά κενά. Ένα κενό δεν είναι το ίδιο με το ότι η απάντηση είναι μια προεπιλεγμένη τιμή, όπως, για παράδειγμα, λάθος για ένα δυαδικό αριθμό ή η μέση τιμή για μια πραγματική αξία. Πολλά συστήματα έχουν αναπτυχθεί για τη διαχείριση τέτοιων «ελλειπουσών» τιμών, σχεδόν όλα με εγγενή ελαττώματα και αδυναμίες.

Οι αδυναμίες αυτές εκδηλώνονται κυρίως όταν η πλειοψηφία των αξιών λείπουν. Για τα κείμενα, οι τιμές που λείπουν δεν είναι θέμα: οι λέξεις είναι είτε παρούσες είτε απούσες από ένα έγγραφο και μπορούμε να γεμίσουμε πλήρως το λογιστικό φύλλο και όλα τα «κελιά» του. Στον απλουστευμένο κόσμο της εξόρυξης κειμένου, έχουμε περιγράψει τα έγγραφα ως παραδείγματα και τις λέξεις ως χαρακτηριστικά σε ένα υπολογιστικό φύλλο. Αν και θα μπορούσε να υποστηριχθεί ότι αυτές είναι αρκετά σημαντικές απλουστεύσεις των αναγκών ενός κειμένου, είναι συνεπείς με το θέμα της μετατροπής λέξεων σε αριθμούς, έτσι ώστε να μπορούν να εφαρμοστούν γνωστές μέθοδοι εξόρυξης δεδομένων.

Έτσι, αν και η εξόρυξη κειμένου λειτουργεί σε πολύ υψηλές διαστάσεις, σε πολλές περιπτώσεις, η επεξεργασία είναι αποτελεσματική και αποδοτική λόγω της σποραδικότητας των χαρακτηριστικών των περισσότερων εγγράφων και των πιο πρακτικών εφαρμογών.

2.5 Εφαρμογές και περιορισμοί

2.5.1 Text mining vs web search

Το Text Mining είναι διαφορετικό από αυτό με το οποίο είμαστε εξοικειωμένοι ως web search. Στην αναζήτηση, ο χρήστης ψάχνει για κάτι που είναι ήδη γνωστό και έχει γραφτεί από κάποιον άλλον. Το πρόβλημα να βάλεις στην άκρη όλο το υλικό που δεν είναι σχετικό με τις ανάγκες σου προκειμένου να βρεθούν σχετικές πληροφορίες που ψάχνεις.

2.5.2 Text mining vs Information retrieval

Επίσης είναι διαφορετικό από αυτό που είναι γνωστό ως Ανάκτηση της Πληροφορίας ή καλύτερα αντιπροσωπεύει ένα σημαντικό βήμα προς τα εμπρός.
Στην ανάκτηση πληροφορίας γίνεται η εύρεση των κειμένων που περιέχουν ήδη τις απαντήσεις στις ερωτήσεις και όχι εύρεση νέας γνώσης. [33][37]
Γενικά στην ανάκτηση πληροφορίας γίνεται μια ερώτηση και στόχος είναι να εξαχθούν όλα τα έγγραφα που είναι πιο κοντά στην ερώτηση.

2.5.3 Text mining vs. Information Extraction

Υπάρχουν προγράμματα μου μπορούν με λογική ακρίβεια, να εξαγάγουν πληροφορίες από κείμενο με κάπως συστηματοποιημένη δομή. Για παράδειγμα τα προγράμματα που διαβάζουν περιλήψεις και εξάγουν ονόματα ανθρώπων, διευθύνσεις, δεξιότητες εργασίας και λοιπά, μπορούν να δώσουν ακρίβεια της τάξης του 80%. Δεν μπορεί όμως αυτό θεωρηθεί text mining σε καμία περίπτωση. Μάλλον προέρχεται από μια περιοχή αποκαλούμενη εξαγωγή πληροφοριών.

Η εξαγωγή χαρακτηριστικών γνωρισμάτων πληροφορίας δεν μπορεί να ταξινομηθεί ως text mining και αυτό διότι δεν περιλαμβάνει την έννοια της «καινούριας» πληροφορίας. Τα χαρακτηριστικά που εξάγονται είναι γνώση που είναι ήδη γνωστή.

Βέβαια η εξαγωγή πληροφορίας περιλαμβάνεται άμεσα σε μια διαδικασία text mining

2.5.4 Εφαρμογές

Ο πιο ενεργός τομές εφαρμογής για το Text Mining είναι στις βιο-επιστήμες. Το καλύτερα γνωστό παράδειγμα όπως αναφέρει η M.Hierst στο [38] είναι η εργασία του Dan Swanson για τις υποτιθέμενες αιτίες των σπάνιων ασθενειών με την έρευνα των έμμεσων συνδέσεων στα διαφορετικά υποσύνολα της λογοτεχνίας των βιοεπιστημών.

Ο Swanson (1988) άρθρωσε την ιδέα ότι η επιστημονική λογοτεχνία πρέπει να θεωρηθεί ως φυσικό φαινόμενο αντάξιο «της εξερεύνησης, του συσχετισμού και της σύνθεσης».

Σε ένα άλλο παράδειγμα, μια από τις τρέχουσες ερωτήσεις των επιστημόνων που ασχολούνται με τα γονίδιο είναι ποιες πρωτεΐνες αλληλεπιδρούν με ποιες άλλες. Έχει υπάρξει ξεχωριστή επιτυχία στην εξέταση ποιες λέξεις ομο - εμφανίζονται στα άρθρα που συζητούν για τις πρωτεΐνες προκειμένου να προβλεφθούν άλλες αλληλεπιδράσεις[39]. Το μεγαλύτερο μέρος της έρευνας φαίνεται πράγματι να εμφανίζεται να είναι στη βιοπληροφορική και στις βιοεπιστήμες. [37]

Η ανάπτυξη αποδοτικών text mining εργαλείων είναι κρίσιμη για τις τρέχουσες και μελλοντικές ανάγκες όσον αφορά τη συλλογή της πληροφορίας. Η απεραντοσύνη και ποικιλομορφία των διαθέσιμων κειμένων καθώς επίσης και η μη δομημένη φύση τους κάνουν την έρευνα σε αυτό τον τομέα συναρπαστική.

2.6 Στόχοι, μεθοδολογία και εργαλεία εξόρυξης κειμένου

Ο Sharp προτείνει ότι για να είναι μια διαδικασία εξόρυξης κειμένου ικανή και αποτελεσματική θα πρέπει να ακολουθεί ορισμένα κριτήρια.

Αρχικά θα πρέπει να λειτουργεί στις μεγάλες συλλογές κειμένων φυσικής γλώσσας. Στη συνέχεια θα πρέπει να χρησιμοποιεί περισσότερο αλγορίθμους από ότι τα ευρετικά και το χειρωνακτικό φιλτράρισμα. Θα πρέπει να εξάγει τις φαινομενολογικές μονάδες των πληροφοριών και τέλος θα πρέπει να ανακαλύπτει νέα γνώση.

2.6.1 Βήματα του Text Mining

Ουσιαστικά υπάρχουν τρία βήματα που θα πρέπει να ακολουθήσουμε στην διαδικασία του text mining.

- Συλλογή των σχετικών με το πρόβλημα εγγράφων – Document Collection

Το πρώτο βήμα είναι να προσδιοριστεί ποια έγγραφα πρόκειται να ανακτηθούν. Μόλις προσδιορίσουμε την πηγή εγγράφων μας θα πρέπει να ανακτήσουμε τα έγγραφα.

- Συλλογή των εγγράφων – (Preprocessing)

Αυτό το βήμα περιλαμβάνει οποιοδήποτε είδος διαδικασιών μετασχηματισμού των αρχικών εγγράφων που ανακτώνται. Αυτοί οι μετασχηματισμοί θα μπορούσαν να στοχεύσουν στη λήψη της επιθυμητής αντιπροσώπευσης των εγγράφων. Τα κείμενα τα οποία προκύπτουν, υποβάλλονται σε επεξεργασία για αν παρέχουν τις βασικές γλωσσικές πληροφορίες για το περιεχόμενο κάθε εγγράφου.

- Στόχοι – Λειτουργίες – (text mining operations)

Οι υψηλού επιπέδου πληροφορίες εξάγονται. Τα σχέδια και οι σχέσεις ανακαλύπτονται μέσα από τις αποσπασματικές πληροφορίες.

Στη συνέχεια θα πρέπει να αναφερθούμε στις τεχνικές που μπορούμε να εφαρμόσουμε σε κειμενικά δεδομένα χωρίς να αφαιρέσουμε όρους που να αφορούν τη σημασιολογική σημασία των κειμένων και των εγγράφων γενικότερα.

- **Γλωσσική επεξεργασία**
  Στην συγκεκριμένη περίπτωση διώχνουμε όλα τα περιττά σύμβολα. Πιθανόν και τους αριθμούς και όλα τα αλφαριθμητικά αν θεωρήσουμε ότι δεν προσθέτουν κάποια ιδιαίτερη σημασία στα έγγραφα. Η διαδικασία αυτή είναι γνωστή ως Tokenization. Επίσης θα μπορούσε κάποιος να περιλάβει και την ανάθεση των όρων στις συντακτικές του κατηγορίες (ουσιαστικό, ρήμα, επίρρημα κτλ) ή να κάνει Λημματοποίηση. Τέλος να κάνει case – folding. Το case – folding αποτελείται από μετατροπή όλων των χαρακτήρων των εγγράφων στο ίδιο σχήμα- format.

- **Αφαίρεση των stopwords**
  Τα stopwords είναι λέξεις που εμφανίζονται πολύ συχνά σε ένα έγγραφο. Περιλαμβάνουν χαρακτηριστικά τις προθέσεις , τα άρθρα, κλπ.
  Είναι λέξεις οι οποίες καλούνται ως μη περιγραφικές μέσα στο κείμενο. Δεδομένου ότι είναι πολύ κοινοί σε πολλά έγγραφα φέρνουν πολύ μικρή πληροφορία για το περιεχόμενο του εγγράφου στο οποίο εμφανίζονται. Είναι μια καλή ιδέα να αφαιρεθούν από την αντιπροσώπευση των εγγράφων.

- **Αφαίρεση των όρων με βάση το μήκος τους**
  Τεχνική συμπληρωματική της αφαίρεσης των stopwords. Με αυτόν τον τρόπο μπορούμε να αφαιρέσουμε μικρούς όρους που πιθανόν να αποτελούν συνδέσμους ή προθέσεις και πολύ μεγάλους όρους που είναι πιθανόν τα αποτελούν τυπογραφικά λάθη.

- **Αφαίρεση των όρων με βάση της συχνότητάς τους**
  Με τον τρόπο αυτό μπορούμε να αφαιρέσουμε σπάνια ή πολύ συχνά χρησιμοποιημένους όρους που πιθανόν δεν έχουν ιδιαίτερη αξία σαν όροι δεικτοδότησης.

- **Stemming**
  Είναι μια κοινή μορφή επεξεργασίας της γλώσσας των κειμένων στα περισσότερα συστήματα ανάκτηση πληροφορίας. Η ιδέα είναι να βελτιωθεί η ανάκληση με τον αυτόματο χειρισμό των καταλήξεων των λέξεων και τη μείωση αυτών στις ρίζες τους. Γίνεται συνήθως με αφαίρεση των οποιοδήποτε συνημμένων επιθεμάτων και των προθεμάτων από τους όρους.

- **N- grams**
  Αποτελεί μια εναλλακτική λύση για το stemming και την αφαίρεση των stopwords**.**

- **Στάθμιση**

2.6.2 Στόχοι του Text Mining

Ο κύριος στόχος του text mining είναι να βοηθήσει τους χρήστες να εξάγουν πληροφορίες από μεγάλους «κειμενικούς πόρους». Οι τεχνικές εξόρυξης δεδομένων, επεξεργασίας φυσικής γλώσσας, μηχανικής μάθησης και ανάκτησης πληροφορίας λειτουργούν μαζί για να ανακαλύψουν αυτόματα υποδείγματα στις πληροφορίες και στα δεδομένα που έχουν προκύψει από τα έγγραφα.

Θα αναφέρουμε επιγραμματικά κάποιους από τους βασικούς στόχους που χειρίζεται η διαδικασία εξόρυξης γνώσης από κείμενα:

- Εξαγωγή χαρακτηριστικών γνωρισμάτων
- Αναζήτηση και ανάκτηση
- Κατηγοριοποίηση
- Ομαδοποίηση
- Περιληπτική Παρουσίαση της πληροφορίας
- Σημασιολογική Ανάλυση
- Γλωσσικός Προσδιορισμός και απόδοση του κειμένου στον συγγραφέα
- Απεικόνιση
- Κατασκευή Οντολογιών

Στην συγκεκριμένη διπλωματική θα μας απασχολήσει ο γλωσσικός προσδιορισμός και η απόδοση του κειμένου στον συγγραφέα. Ένα εργαλείο γλωσσικού προσδιορισμού (language identification) μπορεί αυτόματα να ανακαλύψει τη γλώσσα στην οποία ένα έγγραφο γράφεται. Χρησιμοποιεί ενδείξεις

στο περιεχόμενο του εγγράφου για να προσδιορίσει τις γλώσσες, και εάν το έγγραφο γράφεται σε δυο γλώσσες. Ο προσδιορισμός είναι βασισμένο σε ένα σύνολο εγγράφων κατάρτισης στις γλώσσες.

Με την χρήση του data mining τεχνικών μπορούμε να αποδώσουμε ή όχι ένα κείμενο σε κάποιον συγγραφέα. Για παράδειγμα μπορεί να χρησιμοποιηθεί μια μεθοδολογία που να βασίζεται περισσότερο στην ανάλυση του περιεχομένου από ότι στην σύνταξη με χρήση μιας τεχνικής με κανόνες συσχέτισης. [40]

Κεφάλαιο 3ο

3.1 Το ομηρικό Πρόβλημα: Είναι η Ιλιάδα και η Οδύσσεια έργα ενός μόνο ποιητή;

Αυτό που συνήθως αποκαλείται το «Ομηρικό ερώτημα" είναι με διαφορά το παλαιότερο πρόβλημα απόδοσης συγγραφέα. Το Ομηρικό ερώτημα καλύπτει πραγματικά αρκετά θέματα, όπως π.χ. είναι η Ιλιάδα και η Οδύσσεια έργα ενός μόνο ποιητή; Στην παρούσα εργασία, προσπαθούμε να απαντήσουμε στην ερώτηση χρησιμοποιώντας μια τεχνική εξόρυξης δεδομένων. Η εξόρυξη δεδομένων είναι ένας αναδυόμενος τομέας έρευνας που αναπτύσσει τεχνικές για την ανακάλυψη της γνώσης μέσα σε τεράστιους όγκους δεδομένων. Μέθοδοι εξόρυξης δεδομένων έχουν εφαρμοστεί σε ένα ευρύ φάσμα τομέων, από την ανάλυση του καλαθιού της νοικοκυράς μέχρι την ανάλυση των δορυφορικών εικόνων και των ανθρώπινων γονιδιωμάτων.

Πιο συγκεκριμένα, σε αυτή την εργασία, παρουσιάζουμε μια εφαρμογή της εξόρυξης δεδομένων για να διαπιστωθεί κατά πόσο ένα έγγραφο αποδίδεται σε έναν συγγραφέα. Η μεθοδολογία μας βασίζεται στην ανάλυση του περιεχόμενου και όχι της σύνταξης. Πιο συγκεκριμένα, προτείνουμε μια τεχνική για την εξόρυξη κανόνων συσχέτισης, προκειμένου να αναλυθούν οι συσχετίσεις μεταξύ των εννοιών. Επίσης παρουσιάζουμε τα αποτελέσματα των αναλύσεων που έχουμε πραγματοποιήσει χρησιμοποιώντας αυτό τον αλγόριθμο.

3.1.1 Ειδική μεθοδολογία

Αυτό που συνήθως αποκαλείται «το Ομηρικό ερώτημα" είναι με διαφορά το παλαιότερο πρόβλημα απόδοσης συγγραφέα. Ωστόσο, γενικά έχει δοθεί ελάχιστη προσοχή τον τελευταίο καιρό σε αυτό το θέμα και σχεδόν καθόλου από την υπολογιστική στυλομετρία.

Το Ομηρικό ερώτημα καλύπτει πραγματικά διάφορα θέματα: είναι η Ιλιάδα και η Οδύσσεια έργα ενός μόνο ποιητή; Αν ναι, συνέθεσε ο ίδιος ποιητής και τα δύο έργα; Αν όχι, σε πόσα ξεχωριστά τμήματα θα πρέπει να χωρίσουμε αυτά τα ποιήματα;

Καθ 'όλη την αρχαιότητα, οι μελετητές σχεδόν ομόφωνα υποστήριξαν την ίδια άποψη: και τα δύο ποιήματα είναι έργα του Ομήρου. Με το πέρασμα των χρόνων, πολλοί άλλοι κριτικοί αμφισβήτησαν αυτή την άποψη, αλλά οι απόψεις τους είχαν αγνοηθεί. Εκείνοι που, μέχρι το τέλος του δέκατου ένατου αιώνα, συμφώνησαν ότι η Ιλιάδα και η Οδύσσεια δεν ήταν έργα ενός μόνο άτομου, βάσισαν την υπόθεσή τους σε τρεις κυρίως θεωρίες.

Πρώτον, υπάρχουν πολλές αντιφάσεις και στα δύο ποιήματα. Για παράδειγμα, οι χαρακτήρες χάνουν τη ζωή τους και στη συνέχεια εμφανίζονται ζωντανοί και πάλι. Δεύτερον, υπάρχουν διαφορές συγγραφικού ύφους. Για παράδειγμα, η Οδύσσεια είναι λιγότερο παραστατική και η στάση των θεών προς τους θνητούς είναι πολύ διαφορετική στα δύο ποιήματα.

Τρίτον, υπάρχουν διαφορές στη χρήση της γλώσσας μεταξύ των δυο ποιημάτων. Μέχρι τη δημοσίευση των Προλεγόμενων του Ομήρου, πιθανόν να υπήρχαν μόνο περίπου δέκα άνθρωποι σε ολόκληρο τον κόσμο που αμφέβαλαν για την ύπαρξη του Ομήρου. Οι Έλληνες και οι Ρωμαίοι ήταν πεπεισμένοι ότι ο Όμηρος ήταν μόνο ένα άτομο και έγραψε τα έργα που φέρουν το όνομά του. Ακόμη και ο Ηρόδοτος αν και είχε εξαιρετικά λίγα να πει για τον Όμηρο, φαινόταν απόλυτα πεπεισμένος ότι ο Όμηρος ήταν ένα άτομο και έγραψε τα δύο έπη, μεταξύ πολλών άλλων ποιημάτων που παρουσίαζαν τον Όμηρο ως συγγραφέα τους. Υπάρχουν επίσης μάζες αποδεικτικών στοιχείων που υποστηρίζουν ότι το όνομα Όμηρος ήταν ευρέως γνωστό από τον 6-5ος αιώνα. Μπορούμε λοιπόν απλά να αγνοήσουμε όλα τα αποδεικτικά στοιχεία που οδηγούν στο συμπέρασμα ότι ο Όμηρος ήταν ένα άτομο;

Ωστόσο, ακαδημαϊκοί που επιμένουν ότι υπήρχε μόνον ένας Όμηρος, έχουν ακόμα περισσότερα να πουν. Μια δημοφιλής θεωρία για την ύπαρξη Ομήρου είναι ότι υπήρχαν πολλά άτομα που συνέθεσαν τα ποιήματα και το όνομα του Ομήρου συνδέθηκε με τα έργα αργότερα. Το πιο δύσκολο να εξηγηθεί είναι το γεγονός ότι ακόμα και ο ίδιος ο Ηρόδοτος αμφιβάλλει ότι η Ιλιάδα και η Οδύσσεια ήταν πραγματικά γραμμένα από τον Όμηρο. Ο Ηρόδοτος μάλιστα μπορεί να ήταν το πρώτο πρόσωπο που αμφέβαλε ότι ο Όμηρος έγραψε όλα τα έργα που του αποδίδονται. Ωστόσο, το γεγονός αυτό και μόνο δεν αρκεί για να αποδείξει τίποτα.

Πολλοί θεωρητικοί του Ομήρου υποστηρίζουν ότι η Οδύσσεια γράφτηκε από τον Όμηρο μέχρι ένα σημείο και ότι το υπόλοιπο βιβλίο συντάχθηκε σε σταδιακά τμήματα από διαφορετικούς συγγραφείς. Αλλά δεν θα ήταν δυνατόν να υπήρχε ένας πυρήνας με όλα τα βασικά στοιχεία της ιστορίας που αργότερα κατασκευάστηκε από διαφορετικούς συντάκτες και κατέληξε στην σημερινή Ιλιάδα και Οδύσσεια;

Μέχρι σήμερα, αυτό παραμένει το ισχυρότερο επιχείρημα ότι ο Όμηρος δεν ήταν ένα μόνο πρόσωπο. Αν και αυτό είναι το ισχυρότερο επιχείρημα, δεν είναι όμως και η βάση της θεωρίας. Η όλη θεωρία στηρίζεται σε διαφορές μεταξύ της Ιλιάδας και της Οδύσσειας, καθώς και σε κάποιες ανωμαλίες που παρατηρούνται στα έργα.

Το πρώτο επιχείρημα είναι ότι τα έργα είναι φαινομενικά από διαφορετικά κοινωνικά περιβάλλοντα. Οι επικριτές επισημαίνουν επίσης τις τεράστιες

λεξιλογικές διαφορές μεταξύ των δύο επών, ωστόσο δεν μπορούν να εξηγήσουν τη χρήση παρόμοιου συγγραφικού ύφους. Υπάρχουν επίσης πολλές διαφορές μεταξύ των πλοκών των δύο επών.

Για παράδειγμα, στην Ιλιάδα, ο κυματοθραύστης θα έπρεπε να είχε ανεγερθεί στο 1ο έτος του πολέμου, αλλά και οι Τρώες δεν τον έχτισαν έως το 9ο έτος. Επίσης, ένας άνδρας που σκοτώθηκε στο βιβλίο V, παρουσιάζεται αργότερα στο βιβλίο XIII. Αυτό δεν σημαίνει όμως ότι δεν υπάρχουν τέτοια λάθη και στην Οδύσσεια. Τέτοιες διαφορές στην πλοκή είναι παρούσες και στα δύο έργα. Εντός της Οδύσσειας, φαίνεται να υπάρχουν πολλές μικρές αλλαγές ύφους και διαλέκτου που γίνονται πιο έντονες στο βιβλίο XXIV. Θα πρέπει επίσης να σημειωθεί ότι τόσο τα αρχαία όσο και τα μεταγενέστερα ελληνικά χρησιμοποιήθηκαν, γεγονός που υποδηλώνει ότι το έπος γραφόταν για παρατεταμένη χρονική περίοδο, και η Ελληνική γλώσσα είχε το χρόνο να εξελιχθεί και να αλλάξει.

Από όλα αυτά, η απλούστερη εξήγηση φαίνεται να είναι ότι υπήρχαν πολλά ποιήματα γραμμένα από πολλούς συγγραφείς που αργότερα ενώθηκαν μαζί, ίσως στην Αλεξάνδρεια.

Αυτά είναι βασικά επιχειρήματα που θα μπορούν πιθανώς να εξηγηθούν με μια θεωρία. Η θεωρία αυτή είναι ότι ο 'Ομηρος ήταν ένας βάρδος. Απάγγειλε τα ποιήματά του προφορικά, μπροστά σε ένα ακροατήριο, και δεν σκόπευε ποτέ τα έργα του να γραφούν και να διαβαστούν. Ο 'Ομηρος μάλλον δεν περίμενε καν τα έργα του να επιζήσουν περισσότερο από μερικές δεκαετίες μετά το θάνατό του. 'Ετσι, επέτρεψε στον εαυτό του κάποια ελευθερία κατά τη σύνθεση των έργων, δεδομένου ότι μόνο ένας κριτικός θα παρατηρούσε κάποια τέτοια λάθη και μόνο κατά την ανάγνωση αυτών των ποιημάτων και όχι κατά την ακρόαση, όπως ο 'Ομηρος σκόπευε.

3.2 Στυλομετρία και πατρότητα κειμένου

3.2.1. Εισαγωγή

Η Στυλομετρία είναι η εφαρμογή της μελέτης του γλωσσικού ύφους, συνήθως στη γραπτή γλώσσα. Κατά τα τελευταία χρόνια έχει εφαρμοστεί με επιτυχία επίσης και στη μουσική καθώς και στους art πίνακες ζωγραφικής.

Η στυλομετρία χρησιμοποιείται συχνά και σε περιπτώσεις όπου απαιτείται η απόδοση της πατρότητας σε ανώνυμα ή σε αμφισβητούμενα έγγραφα. 'Εχει νομικές καθώς και ακαδημαϊκές και λογοτεχνικές εφαρμογές, που κυμαίνονται για παράδειγμα από το ζήτημα της πατρότητας των έργων του Σαίξπηρ ως και σε εγκληματολογική γλωσσολογία.[8],[9].

Η στυλομετρία προέκυψε μετά από προηγούμενες τεχνικές ανάλυσης κειμένων για την απόδειξη της γνησιότητας, συγγραφικής ταυτότητας, καθώς άλλων παρόμοιων ζητημάτων.

Η σύγχρονη πρακτική της πειθαρχίας έλαβε σημαντική ώθηση από τη μελέτη των προβλημάτων πατρότητας των αγγλικών δραμάτων της Αναγέννησης. Οι ερευνητές και οι αναγνώστες παρατήρησαν ότι ορισμένοι συγγραφείς της εποχής είχαν προτιμήσεις σε συγκεκριμένα διακριτικά πρότυπα της γλώσσας και προσπάθησαν να χρησιμοποιήσουν αυτά τα πρότυπα για τον εντοπισμό δημιουργών ανάλογων κειμένων με αβέβαια ταυτότητα ή κειμένων που προέκυψαν με την συνεργασία διαφόρων δημιουργών.

Τα βασικά της στυλομετρίας περιγράφονται από την Πολωνή φιλόσοφο Wincenty Lutosławski στο βιβλίο "Principes de stylometrie" το 1890. Η Lutosławski χρησιμοποίησε αυτή την τεχνική για να δημιουργήσει ένα «χρονολόγιο» για τους διαλόγους του Πλάτωνα.

Η ανάπτυξη των υπολογιστών και των ικανοτήτων τους για την ανάλυση μεγάλων ποσοτήτων δεδομένων ενίσχυσαν αυτού του είδους προσπάθεια σε μεγάλο βαθμό.

Η μεγάλη χωρητικότητα των υπολογιστών για ανάλυση δεδομένων, ωστόσο, δεν εγγυάται την ποιότητα της παραγωγής. Στις αρχές της δεκαετίας του 1960 παράχθηκε μια ανάλυση του υπολογιστή για τις δεκατέσσερις Επιστολές της Καινής Διαθήκης που αποδίδονται στη μνήμη του Αγίου Παύλου, η οποία έδειξε ότι η εργασία αυτή γράφτηκε από έξι διαφορετικούς συγγραφείς. Με τον καιρό, όμως, και με την πρακτική, ερευνητές και επιστήμονες κατάφεραν να τελειοποιήσουν τις προσεγγίσεις και τις μεθόδους τους και να αποφέρουν καλύτερα αποτελέσματα.

Μια αξιοσημείωτη επιτυχημένη εφαρμογή της στυλομετρίας αποτελεί η υπο - αμφισβήτηση πατρότητα των δώδεκα Federalist Papers από Frederick Mosteller και David Wallace. [11] Αν και ερωτήσεις των αρχικών παραδοχών και της μεθοδολογίας ακόμη προκύπτουν (και, ίσως, θα είναι πάντα), λίγοι αμφισβητούν πλέον ότι η γλωσσολογική ανάλυση των γραπτών κειμένων αποτελεί βασική προϋπόθεση για την παραγωγή πολύτιμων πληροφοριών .

3.2.2 Σκοποί της στυλομετρίας

Επιγραμματικά αναφέρουμε κάποιους από τους στόχου της στυλομετρίας:

1. Ταξινόμηση Ύφους
2. Ιστορική μελέτη της αλλαγής γλώσσας
3. Ανάλυση λογοτεχνικών στοιχείων

4. Πατρότητα Κειμένου
5. Ρητορική γλωσσολογία

3.2.3 Μέθοδοι στυλομετρίας

Η σύγχρονη στυλομετρία στηρίζεται σε μεγάλο βαθμό στη βοήθεια των ηλεκτρονικών υπολογιστών οι οποίοι είναι πλέον οι ιδανικοί για τις στατιστικές αναλύσεις, για θέματα τεχνητής νοημοσύνης και πρόσβασης στο «αναπτυσσόμενο» σώμα κειμένων που διατίθενται μέσω του Διαδικτύου. [26],[27]

Ενώ στο παρελθόν, η στυλομετρία έδινε μεγαλύτερη έμφαση στα πιο σπάνια και πιο εντυπωσιακά στοιχεία , οι σύγχρονες τεχνικές μπορούν να απομονώσουν και να αναγνωρίσουν πρότυπα ακόμη και σε κοινά μέρη του λόγου.

Οι μέθοδοι της στυλομετρίας είναι οι ακόλουθοι:

1. Αναλλοίωτη συγγραφική ιδιότητα:

Η πρωταρχική στυλομετρική μέθοδος βασίζεται στην αναλλοίωτη συγγραφική ικανότητα- μια ιδιότητα ενός κειμένου που είναι αναλλοίωτη από το συντάκτη του. Ένα παράδειγμα αποτελεί η συχνότητα των λέξεων που χρησιμοποιεί κάποιος συγγραφέας.

Σε μία τέτοια μέθοδος, το κείμενο αναλύεται για να βρείτε τις 50 πιο συνηθισμένες λέξεις. Το κείμενο αποτελεί μια συνέχεια 5.000 κομμάτια – λέξεις και κάθε ένα από τα κομμάτια αναλύεται για να βρείτε τη συχνότητα αυτών των 50 λέξεων (των πιο συνηθισμένων) σε αυτό το κομμάτι. Αυτό δημιουργεί ένα μοναδικό αναγνωριστικό «50-αριθμός» για κάθε κομμάτι. Δημιουργείται με αυτόν τον τρόπο ένας 50-διάστατος χώρος. Αυτός ο 50-διάστατος χώρος είναι σαν πεπλατυσμένος χώρος όπου γίνεται χρήση των κύριων συνιστωσών ανάλυσης και τα αποτελέσματα αντιστοιχίζονται σε σημεία τα οποία καθορίζουν και το ύφος και το στυλ του συγγραφέα. Αν δύο λογοτεχνικά έργα που τίθενται επί του ιδίου επιπέδου, το αντίστοιχο πρότυπο μπορεί να δείξει εάν και τα δύο έργα ήταν του ίδιου συγγραφέα ή ανήκουν σε διαφορετικούς συγγραφείς.

2. Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα χρησιμοποιούνται κυρίως στην στυλομετρία για να αναλύσουν την πατρότητα του κειμένου.

3. Γενετικοί Αλγόριθμοι

Οι γενετικοί αλγόριθμοι είναι μια άλλη τεχνική της τεχνητής νοημοσύνη που χρησιμοποιείται στην στυλομετρία. Στην συγκεκριμένη μέθοδο ξεκινάμε θέτοντας ένα σύνολο κανόνων. Για παράδειγμα : "Εάν η λέξη but εμφανίζεται περισσότερες από 1,7 φορές σε κάθε χίλιες λέξεις, τότε το κείμενο αποδίδεται στον συγγραφέα Χ". Το πρόγραμμα παρουσιάζεται με κείμενο και χρησιμοποιεί τους κανόνες για τον προσδιορισμό του συντάκτη του. Οι κανόνες που δοκιμάζονται από ένα σύνολο γνωστών κειμένων και σε κάθε κανόνα δίνεται μια «βαθμολογία» . Οι 50 κανόνες με το χαμηλότερο σκορ απορρίπτονται ενώ οι υπόλοιποι 50 κανόνες που απομένουν επιδέχονται μικρές αλλαγές και 50 νέοι κανόνες εισάγονται. Αυτό επαναλαμβάνεται μέχρις οι κανόνες οι οποίοι προκύπτουν αποδίδουν σωστά το κείμενο.

4. Σπάνια Ζεύγη

Μια μέθοδος για τον εντοπισμό του στυλ και του ύφους ενός κειμένου ονομάζεται "σπάνια ζεύγη», και εξαρτάται από τις μεμονωμένες συνήθειες των συγγραφέων που έγραψαν το συγκεκριμένο κείμενο. Η χρήση ορισμένων λέξεων μπορεί, για ένα συγκεκριμένο συγγραφέα, συνεπάγεται ιδιοσυγκρασιακά την χρήση άλλων, προβλέψιμων λέξεων.

(όλα τα παραπάνω βασίζονται στις [7], [8], [9], [10], [11])

3.3 Πληροφορίες σχετικά με τα κείμενα Ιλιάδας και Οδύσσειας – Ιστορικά στοιχεία για τον επικό ποιητή 'Ομηρο.

Ο '**Ομηρος** φέρεται ως ο συγγραφέας των ποιητικών κειμένων της Ιλιάδας και της Οδύσσειας, από τα πρώτα κείμενα της Ιστορικής περιόδου της αρχαίας Ελλάδας, γνωστά ως «Ομηρικά Έπη». Για τη ζωή του υπάρχουν ελάχιστες πληροφορίες, και αυτές αντιφατικές, ενώ η φιλολογική επιστήμη των δύο τελευταίων αιώνων αμφισβήτησε ακόμη και την ύπαρξή του.

Στον 'Ομηρο κατά καιρούς αποδόθηκαν και άλλα έργα, τα οποία σήμερα είναι αποδεκτό ότι δεν είναι δικά του, αλλά ακόμη είναι αμφισβητήσιμο το αν τα δύο μεγάλα έπη είναι έργα του ίδιου ποιητή. Η *Ιλιάς* αποτελείται από περίπου 16.000 στίχους και αναφέρεται στις τελευταίες πενήντα μία (51), αποφασιστικής σημασίας ημέρες του πολέμου της Τροίας, ο οποίος συνολικά διήρκεσε, σύμφωνα με το μύθο, 10 χρόνια. Η *Οδύσσεια* αποτελείται από περίπου 12.000 στίχους και περιγράφει τις επίσης δεκαετούς διάρκειας περιπλανήσεις του βασιλιά της Ιθάκης, Οδυσσέα, κατά τη μετάβαση από την Τροία που είχε αλωθεί, στην πατρίδα του.

3.3.1 Αρχαίες μαρτυρίες για τη ζωή και το έργο του

Διαθέτουμε επτά βίους του Ομήρου που προέρχονται από την αρχαιότητα. Η καταγωγή του φαίνεται πως ήταν από την Ιωνία και θρυλείται ότι επτά πόλεις ερίζουν για την καταγωγή του, με επικρατέστερες τη Σμύρνη και τη Χίο. Ως γονείς του αναφέρονται ο Μαίων και η Κριθηίδα και λέγεται ότι το πραγματικό του όνομα ήταν Μελησιγένης, επειδή γεννήθηκε κοντά στον ποταμό Μέλητα της Σμύρνης και ότι πήρε αργότερα το όνομα «Όμηρος», είτε επειδή ήταν τυφλός, είτε επειδή ήταν όμηρος των Κολοφωνίων στον πόλεμο με τη Σμύρνη. Σύμφωνα με τους βίους του, περιόδευσε απαγγέλλοντας τα έργα του στις ελληνικές πόλεις, απέκτησε μεγάλη φήμη, αλλά σε ένα διαγωνισμό με τον Ησίοδο στη Χαλκίδα δεν πήρε βραβείο επειδή προτιμήθηκε ο Ησίοδος ως ποιητής που εξυμνούσε την ειρήνη. Ως τόπος θανάτου του παραδίδεται η Ίος..

Η σύγχρονη έρευνα, και ειδικότερα όσοι δέχονται ότι ο Όμηρος μπορεί να θεωρηθεί πραγματικό πρόσωπο, τοποθετεί τη ζωή του στον 8ο αι. π.Χ. και θεωρεί πιθανό ότι ήταν Ίωνας αοιδός, συνεχιστής μιας μακραίωνης παράδοσης **προφορικών ηρωικών αφηγήσεων,** που συνέθεσε την *Ιλιάδα* γύρω στο 740 π.Χ. και την *Οδύσσεια* (αν όντως συνέθεσε και τα δύο έργα) γύρω στα 710 π.Χ.

3.3.2 Το ομηρικό ζήτημα

Υπό τον όρο «ομηρικό ζήτημα» ομαδοποιούνται πολλά ερωτήματα που έχουν σχέση με την πατρότητα, τον τρόπο σύνθεσης και την καταγραφή της *Ιλιάδας* και της *Οδύσσειας*. Ειδικότερα, έχουν τεθεί τα θέματα:

- *Ήταν πραγματικό πρόσωπο ο Όμηρος*; Πότε έζησε, πώς συνέθεσε ή έγραψε τα έργα του και ποια είναι αυτά;
- *Τα κείμενα που έχουμε στη διάθεσή μας σήμερα είναι έργα του ίδιου ποιητή;* Κάποιες υφολογικές αλλά και πολιτισμικές διαφορές μεταξύ των δύο ποιημάτων καθιστούν πιθανό το γεγονός να μην γράφτηκαν από τον ίδιο συγγραφέα, χωρίς κάτι τέτοιο να μπορεί να αποδειχθεί με βεβαιότητα.
- *Τα κείμενα είναι ενιαίες ποιητικές συλλήψεις ή αποτελούνται από διάφορα στρώματα;* Αρκετοί έχουν υποστηρίξει ότι τα σημερινά κείμενα προέρχονται από συνένωση πολλών τμημάτων ή επέκταση παλαιοτέρων. Απέναντι σε αυτήν την «αναλυτική» θεωρία τάσσονται οι «ενωτικοί» που υποστηρίζουν ότι στο καθένα μπορεί να διακριθεί μία συνεπής λογοτεχνική σύλληψη και πραγμάτωση από ένα άτομο. Η σύγκριση με προφορικά έπη έδειξε ότι οι προφορικοί ποιητές, με τεχνικές που δεν είναι οικείες σε μια εγγράμματη κοινωνία, μπορούν να συνθέσουν και να απομνημονεύσουν ποιήματα μεγάλης έκτασης.
- Από την προφορική θεωρία, προκύπτει το ερώτημα *ποια ήταν η συμβολή της γραφής στη σύνθεση των ποιημάτων*: καταγράφηκαν την εποχή που

συνδέθηκαν κατά τη διάρκεια της απαγγελίας, υπαγορεύτηκαν από τον ποιητή ή επιβίωσαν προφορικά και καταγράφηκαν αργότερα;

3.3.3 Αναλυτική θεωρία

Η παρουσία κάποιων αντιφάσεων, λογικών κενών ή χασμάτων στο κείμενο της *Χιλιάδας* και της *Οδύσσειας* οδήγησε στην υπόθεση ότι τα σωζόμενα κείμενα δεν είναι ενιαίες ποιητικές συλλήψεις αλλά συνένωση περισσοτέρων έργων. Οι υποστηρικτές αυτής της θεωρίας ονομάστηκαν «αναλυτικοί» και οι απόψεις του μπορούν να διαιρεθούν σε επιμέρους τάσεις.

Για την *Ιλιάδα*, μία από τις αναλυτικές θεωρίες ήταν η **θεωρία της επέκτασης**, που υποστηρίχθηκε κυρίως από τον Gottfried Hermann (1772-1848): σύμφωνα με αυτή, υπήρχε ένα παλαιό κείμενο, μια αρχική Ιλιάδα, που σταδιακά επεκτάθηκε και πήρε τη σημερινή μορφή. Η **θεωρία των ασμάτων**, που αναπτύχθηκε από τον Karl Lachmann, θεωρούσε την Ιλιάδα συνένωση μικρότερων επικών ασμάτων (ο Lachmann εντόπιζε περίπου δεκαέξι άσματα). Συγγενική ήταν η **θεωρία της συγκόλλησης**, με βασικό εκπρόσωπο τον A. Kirchhoff, κατά τον οποίο η *Ιλιάδα* δημιουργήθηκε από συνένωση μικρότερων επών. Ο ίδιος επιχείρησε ανάλογη ανάλυση και για την *Οδύσσεια*, για την οποία διατυπώθηκε και μια άλλη άποψη, η **θεωρία του διασκευαστή**, δηλαδή η άποψη ότι υπήρχε μια αρχική Οδύσσεια που επεκτάθηκε στη συνέχεια με προσθήκες ενός διασκευαστή που υστερούσε σε ποιητική αξία από τον ποιητή του αρχικού έργου.

3.3.4 Προφορικότητα και γραφή

Διαφορετική κατεύθυνση δόθηκε στην ομηρική έρευνα από τη σύγκριση με τις τεχνικές της προφορικής ποίησης. Οι Milman Parry και Albert Lord, βασισμένοι στη διαπίστωση ότι τα δύο έπη εμφανίζουν στερεότυπες σκηνές και εκφράσεις, που συχνά επαναλαμβάνονται αυτούσιες, αξιοποίησαν τις έρευνές τους για την προφορική ηρωική ποίηση γης Γιουγκοσλαβίας για να φωτίσουν τον τρόπο σύνθεσης της *Ιλιάδας* και της *Οδύσσειας* και κατέληξαν στο συμπέρασμα ότι τα δύο κείμενα παρουσιάζουν ανάλογη τεχνική, αφού βασίζονται σε ένα σύνολο στερεότυπων μικρότερων ή μεγαλύτερων φράσεων λογότυπων και τυποποιημένων σκηνών.

Σήμερα είναι αποδεκτό το γεγονός ότι οι τεχνικές στις οποίες βασίστηκε η σύνθεση των δύο επών είναι οι τεχνικές της προφορικής ποίησης, όπως είχαν διαμορφωθεί τους προηγούμενους αιώνες. Η παράδοση τροφοδότησε τον ποιητή τους με μια ειδική τεχνητή διάλεκτο, με στοιχεία διαφόρων εποχών και περιοχών και πολλά συνώνυμα που μπορούν να χρησιμοποιηθούν σε διαφορετικές μετρικές θέσεις, ένα σύνολο λογοτύπων που αντιστοιχούν σε συγκεκριμένες θέσεις του στίχου, τυπικές σκηνές και τυποποιημένα ευρύτερα επεισόδια.

Η συμβολή της γραφής στη σύνθεση ή την καταγραφή της *Ιλιάδας* και της *Οδύσσειας* είναι δύσκολο να καθοριστεί και έχουν διατυπωθεί διάφορες υποθέσεις: μπορεί ο ποιητής να χρησιμοποίησε τη γραφή για να κάνει ένα σχέδιο της δομής και της σύνδεσης διαφόρων επεισοδίων, ή να υπαγόρευσε σε κάποιον το ποίημα. Είναι βέβαιο ότι και τους επόμενους αιώνες τα έπη είχαν συντηρηθεί στην προφορική παράδοση και απαγγέλλονταν, αλλά δεν γνωρίζουμε αν υπήρχε κάποιο παγιωμένο γραπτό κείμενο. Από τον 6ο αι. π.Χ. μαρτυρείτε και μια επαγγελματική ένωση ραψωδών που ονομάζονταν «Ομηρίδες», οι οποίοι απήγειλλαν κάποια εκδοχή των επών, αλλά δεν γνωρίζουμε αν είχαν στην κατοχή τους κάποιο γραπτό κείμενο.

Σημαντική θεωρείται στο θέμα της παγίωσης του ομηρικού κειμένου η συμβολή του Πεισίστρατου που λέγεται ότι καθιέρωσε απαγγελίες του Ομήρου στη γιορτή των Παναθήναιων με βάση ένα σταθερό κείμενο (η λεγόμενη «Πεισιστράτεια διόρθωση»).

3.3.5 Γλώσσα και μέτρο

Το μέτρο της *Ιλιάδας* και της *Οδύσσειας* είναι ο δακτυλικός εξάμετρος στίχος. Βάση του είναι ο δακτυλικός πους, δηλαδή μια μονάδα που αποτελείται από μία μακρόχρονη συλλαβή και δύο βραχύχρονες (που μπορεί να αντικατασταθούν από μία μακρόχρονη). Ο κάθε στίχος απαρτίζεται από έξι πόδες. Οι πέντε πρώτοι είναι δάκτυλοι και ο έκτος αποτελείται από δύο συλλαβές, την πρώτη υποχρεωτικά μακρόχρονη και τη δεύτερη αδιάφορη. Συνολικά ένας δακτυλικός στίχος μπορεί να αποτελείται από δώδεκα έως δεκαεπτά συλλαβές. Υπάρχει μια ισχυρή νοηματική παύση περίπου στο μέσον του στίχου, καθώς και άλλες μικρότερες που χωρίζουν τον στίχο έως και σε τέσσερις μικρές νοηματικές ενότητες.

Η γλώσσα του Ομήρου είναι μια τεχνητή γλώσσα, που ποτέ δεν μιλήθηκε, αλλά ήταν κατανοητή σε όλον τον ελληνόφωνο κόσμο. Το υλικό της προέρχεται από διάφορες διαλέκτους και χρονικές περιόδους. Βάση της είναι η ιωνική διάλεκτος όπως είχε διαμορφωθεί τον 8ο αι. π.Χ. στα παράλια της Μ. Ασίας. Υπάρχουν ακόμη πολλά αιολικά στοιχεία, αλλά και τύποι παλαιότεροι που ανάγονται στη μυκηναϊκή εποχή. Δωρικά στοιχεία δεν υπάρχουν, ενώ κάποιοι αττικοί τύποι ενδέχεται να είναι μεταγενέστερες προσθήκες. Η ύπαρξη πολλών συνώνυμων τύπων που προέρχονταν από ποικίλες διαλέκτους ή περιόδους παρείχε μετρικές ευκολίες στον ποιητή, αφού ανάλογα με τη θέση του στίχου μπορούσε να χρησιμοποιήσει μία από πολλές νοηματικά ισοδύναμες λέξεις.

3.3.6 Λογότυποι και τυπικές σκηνές

Από την προφορική επική παράδοση ο Όμηρος είχε κληρονομήσει ένα σύνολο στερεοτυπικού υλικού το οποίο προσάρμοζε ανάλογα με τις ανάγκες της κάθε

περίπτωσης. Η μικρότερη τυπική μονάδα είναι οι σύντομες φράσεις που αποτελούνται από ένα όνομα και επίθετο. Κάποιες φορές τα τυπικά επίθετα χρησιμοποιούνται ακόμη και όταν τα νοηματικά συμφραζόμενα δεν επιτρέπουν τη χρήση τους. Μεγαλύτερης έκτασης λογότυποι χρησιμοποιούνται για να δηλώσουν την αρχή και το τέλος μιας ομιλίας, την μετακίνηση ενός ήρωα ή τα γεγονότα των μαχών.

Εκτός από τις εκφραστικές στερεοτυπίες, υπήρχαν και τυποποιημένες ακολουθίες πράξεων για να περιγράψουν εκτενή γεγονότα όπως η θυσία, η ικεσία, η υποδοχή ενός φιλοξενούμενου, ένα γεύμα, μία μονομαχία. Για παράδειγμα η αφήγηση μιας αριστείας, δηλαδή μιας σειράς κατορθωμάτων ενός ήρωα, βασίζεται σε ένα συγκεκριμένο πρότυπο: πρώτα περιγράφεται ο εξοπλισμός του ήρωα. Όταν αρχίζει η μάχη, ο πρωταγωνιστής σκοτώνει κάποιους εχθρούς σε μονομαχίες και έπειτα επιτίθεται εναντίον του εχθρικού στρατού τον οποίο ωθεί σε φυγή. Η καταδίωξη διακόπτεται όταν αυτός πληγώνεται, αλλά με προσευχή σε ένα θεό θεραπεύεται, επιστρέφει στη μάχη και μονομαχεί με τον αρχηγό των εχθρών. Τον σκοτώνει και ακολουθεί μάχη των δύο παρατάξεων για το πτώμα, το οποίο αποκτούν τελικά οι φίλοι του νεκρού με θεϊκή παρέμβαση. Παρά την τυποποίηση, κάθε φορά που εμφανίζονται τυπικές σκηνές υπάρχουν διαφορές στις λεπτομέρειες ανάλογα με τις ανάγκες.

3.3.7 Εκτενείς παρομοιώσεις

Ένα ιδιαίτερο χαρακτηριστικό της ομηρικής, τεχνικής, που φαίνεται ότι δεν υπήρχε στα παλαιότερα έπη, είναι οι εκτενείς παρομοιώσεις. Αυτές οι παρομοιώσεις είναι επεκτάσεις των σύντομων παρομοιώσεων: για παράδειγμα η απλή παρομοίωση «επιτέθηκε σαν λιοντάρι» μπορεί να επεκταθεί στην εικόνα ενός πληγωμένου και αγριεμένου λιονταριού που επιτίθεται στους κυνηγούς του, στην εικόνα του λιονταριού που κατασπαράζει το πρόβατα ενός φοβισμένου βοσκού ή στην εικόνα του ζώου που προσπαθεί να προστατέψει τα νεογνά του από τους κυνηγούς.

Ο κόσμος των παρομοιώσεων είναι ο κόσμος της καθημερινής εμπειρίας, που είναι οικείος στον ακροατή / αναγνώστη, και έχει στόχο να βοηθήσει τη φαντασία του αναγνώστη να αντιληφθεί κάτι, συγκρίνοντάς το με μια εικόνα από την άμεση εμπειρία του. Τα θέματα από τα οποία προέρχονται οι παρομοιώσεις είναι τα φυσικά φαινόμενα (τρικυμίες, αστραπές και βροντές, πυρκαγιές), το κυνήγι, οι επιθέσεις άγριων ζώων και καθημερινές δραστηριότητες όπως η υφαντική, η ξυλουργική, η ποιμενική και αγροτική ζωή.

Οι εκτενείς παρομοιώσεις δεν είναι απλά διακοσμητικά στοιχεία. Ήδη οι αρχαίοι σχολιαστές είχαν επισημάνει ότι μπορεί να υπάρχουν πολλά σημεία επαφής μεταξύ των συγκρινόμενων όρων. Μπορεί πίσω από το άμεσο νόημα της σύγκρισης να κρύβεται και ένα δεύτερο, πιο σημαντικό: για παράδειγμα, όταν ο

Οδυσσέας, στο τέλος της ραψωδίας ε της Οδύσσειας φτάνει ναυαγός στη γη των Φαιάκων, ξαπλώνει κάτω από ένα σωρό φύλλων και παρομοιάζεται με το κρύψιμο ενός δαυλού στη στάχτη για να παραμείνει αναμμένος. Η άμεση σύγκριση είναι η κάλυψη του Οδυσσέα και του δαυλού, αλλά το βαθύτερο νόημα είναι ότι ο Οδυσσέας προσπαθεί να κρατήσει ζωντανή τη σπίθα της ζωής. Άλλοτε μια παρομοίωση μπορεί να προϊδεάζει για το μέλλον, όπως όταν οι Τρώες βλέπουν τον Αχιλλέα να σέρνει με το άρμα του τον νεκρό Έκτορα και ο θρήνος τους παρομοιάζεται με το θρήνο των κατοίκων μιας φλεγόμενης πόλης, όπως πράγματι έγινε αργότερα.

3.4 Παλαιότεροι τρόποι προσέγγισης του προβλήματος απόδοσης πατρότητας κειμένου σε συγγραφέα

3.4.1 Εισαγωγή

Συνοψίζοντας λοιπόν και όπως αναφέραμε και προηγούμενα, στην παρούσα εργασία προσπαθούμε να απαντήσουμε στο Ομηρικό ερώτημα χρησιμοποιώντας μια τεχνική εξόρυξης δεδομένων . Η εξόρυξη δεδομένων είναι ένας αναδυόμενος τομέας έρευνας που αναπτύσσει τεχνικές για την ανακάλυψη της γνώσης μέσα σε τεράστιους όγκους δεδομένων. Μια κοινή χρήση της εξόρυξης δεδομένων είναι να εκπροσωπεί συχνά πρότυπα δεδομένων με τη μορφή αλληλεξαρτήσεων μεταξύ των εννοιών και των ιδιοτήτων.[20],[24]

Η μεθοδολογία μας βασίζεται στην ανάλυση του περιεχόμενου και όχι της σύνταξης της Ιλιάδας και της Οδύσσειας.

Η πρώτη φάση αποτελέστηκε από μετατροπή των δεδομένων. Αυτό επιτεύχθηκε με την κράτηση μόνο ουσιαστικών, ρημάτων, επιθέτων και επιρρημάτων και την οργάνωση τους σε σύνολα συνωνύμων, με το κάθε ένα να εκπροσωπεί μια ξεχωριστή λεξιλογική έννοια.

Επιλέξαμε να αναλύσουμε τις σχέσεις μεταξύ των εννοιών αυτών. Έτσι, έχουμε μετατρέψει όλες τις φράσεις του κειμένου σε φράσεις που αποτελούνται μόνο από αυτές τις έννοιες, αφού έχουμε πρώτα αφαιρέσει οποιαδήποτε διπλότυπα. Τότε θα μπορέσουμε να μετατρέψουμε το κείμενο σε μια δομημένη μορφή, έτσι ώστε να αποθηκεύονται σε αρχεία της βάσης δεδομένων.

Πιο συγκεκριμένα, θα δούμε τα συνεχή τμήματα του κειμένου ως συνεχείς "συναλλαγές", όπου κάθε τέτοια «συναλλαγή» είναι αποθηκευμένη σε ένα αρχείο. Κάναμε δοκιμές και ορίσαμε κάθε "συναλλαγή" ως συνεχείς προτάσεις και ζεύγη προτάσεων.

Τέλος, εφαρμόσαμε τον αλγόριθμος εξόρυξης δεδομένων Apriori [17] με στόχο την εξαγωγή των κανόνων συσχέτισης. Θεωρήσαμε κανόνες συσχέτισης του τύπου ότι το **"90% των "συναλλαγών" που περιέχουν την έννοια Χ, επίσης περιέχουν την έννοια Υ".**

3.4.2 Το πρόβλημα – Τρόποι προσέγγισης.

Η ενότητα του Ομήρου και γενικά προβλήματα τα οποία αφορούν την απόδοση του κειμένου στο συγγραφέα, τα τελευταία χρόνια παρουσιάζουν μεγάλο ενδιαφέρον. Όπως γνωρίζουμε αυτή τη στιγμή, μόνο στυλομέτρες έχουν καταλήξει σε συμπεράσματα ως προς το ποιος θα μπορούσε να έχει γράψει ένα ποίημα ή ακόμα και ένα μυθιστόρημα και το έχουν κάνει χρησιμοποιώντας στατιστικές θεωρίες. (Στυλομετρία = Η στατιστική ανάλυση του λογοτεχνικού ύφους…αναφερόμαστε σε προηγούμενο κεφάλαιο).

Ο καθένας μπορεί να υποθέσει ότι το ύφος το οποίο υπαγορεύεται από το υποσυνείδητο του κάθε ανθρώπου, αποτελεί το γενετικό αποτύπωμα του έργου ενός συγγραφέα. Επομένως, είναι αδύνατον να συγκαλυφθεί το στυλ του καθενός. Η τεχνοτροπική ανάλυση του κειμένου της Ιλιάδας ίσως θα μπορούσε να αποκαλύψει ότι είναι στατιστικώς παρόμοια με την Οδύσσεια. Το πρόβλημα με αυτήν την υπόθεση ωστόσο είναι ότι δεν υπάρχουν ισχυρά αποδεικτικά στοιχεία για να υποστηρίξει κανείς την ιδέα ότι οι συγγραφείς έχουν μια υποσυνείδητη καθώς και μια συνειδητή πτυχή για το στυλ τους.

Οι δύο τρόποι αναγνώρισης δημιουργού, δηλαδή η Απόδοση Δημιουργού και οι Χρονολογικές μελέτες, έχουν πολλά αντιφατικά σημεία αναφοράς. Η μέθοδος απόδοσης δημιουργού υποστηρίζει ότι η υποσυνείδητη πλευρά του στυλ ενός συγγραφέα παραμένει σταθερή καθ 'όλη τη ζωή του, με άλλα λόγια, ο συγγραφέας θα αφήσει τα ίδια στιλιστικά δακτυλικά αποτυπώματα για κάθε ένα από τα έργα του. Ωστόσο, οι χρονολογικές μελέτες υποστηρίζουν ότι τα χαρακτηριστικά του ύφους ενός συγγραφέα αλλάζουν κατά τη διάρκεια της συγγραφικής του ζωής και έτσι κάνουν εύκολη τη χρονολόγησή τους.

Διάφορες στατιστικές μέθοδοι έχουν χρησιμοποιηθεί για τον προσδιορισμό των δημιουργών και των μετρήσεων του στυλ. Παραδείγματα τέτοιων μεθόδων είναι το μήκος των λέξεων, ο αριθμός των συλλαβών ανά λέξη και το μήκος της φράσης. Παρά το γεγονός ότι οι μέθοδοι που μπορεί να χρησιμοποιηθούν ποικίλουν, είναι εμφανές ότι τα στατιστικά μοντέλα παίζουν πολύ σημαντικό ρόλο για ένα καλό ερευνητικό αποτέλεσμα.

Στο σημείο αναφοράς [22], οι συγγραφείς αναφέρονται σε τρεις εφαρμογές των στατιστικών μεθόδων, μία από τις οποίες αναφέρεται στις διαφορές μεταξύ της Ιλιάδας και της Οδύσσειας, ενώ οι άλλες δυο προσπαθούν να διακρίνουν αν ένα μυθιστόρημα ανήκει σε ένα ή κάποιο άλλο συγγραφέα. Οι παραπάνω εφαρμογές

επικεντρώνονται στην κατανομή του λεξιλογίου ως ένα από τα κύρια διακριτικά του συγγραφικού ύφους.

Η ανάλυση ασχολείται κυρίως με:
(α) λέξεις που επαναλαμβάνονται με υψηλή συχνότητα και
(β) συνώνυμα, αφού αυτά θεωρούνται ως διακριτικά του συγγραφικού στυλ.

Για παράδειγμα, [23], η έρευνα έχει επικεντρωθεί κυρίως στη χρήση συνωνύμων όπως τα "ενώ" και "καθώς", ως διακριτικά του στυλ για να φτάσει στα συμπεράσματά της, ενώ το [18] επικεντρώνεται στην συχνή χρήση και επανάληψη λέξεων, ως σημαντικό στοιχείο του ύφους του συγγραφέα.

Στην περίπτωση του Ομηρικού ερωτήματος, εξετάσθηκε το πρότυπο της χρήσης των συχνών και σπάνιων λέξεων στην Ιλιάδα και την Οδύσσεια.

Σε αυτή την μεθοδολογία που ακολουθήθηκε, δύο ήταν τα κύρια προβλήματα:
(α) Η χρήση λέξεων που άλλαζαν με το πέρασμα του χρόνου και
(β) Οι διαφορές στη χρήση των λέξεων, ανάλογα με το κεντρικό θέμα του κάθε συγγραφικού κομματιού.

Για το λόγο αυτό, προκειμένου να υπάρξει μια κοινή βάση για την έρευνα, αναλύθηκαν κάποια παλαιά αλλά και νεότερα έργα του Σοφοκλή και του Ησίοδου, που διαφέρουν κατά πολύ στο θέμα από την Ιλιάδα και την Οδύσσεια.
(**Ησίοδος**: Θεογονία, Έργα και Ημέραι, Ασπίδα του Ηρακλή.)
(**Σοφοκλής**: Αντιγόνη, Οιδίπους Τύραννος, Ηλέκτρα, Φιλοκτήτης.)
(**Αισχύλος**: Αγαμέμνων. Ευριπίδης: Μήδεια).

Η Ιλιάδα και η Οδύσσεια είχαν ελεγχθεί για να εξασφαλιστεί ότι δεν διαφέρουν στην ορθογραφία. Το ίδιο έγινε και με τα άλλα κείμενα. Όλοι οι απόστροφοι, διακριτικά και σημεία στίξης αφαιρέθηκαν από όλα τα κείμενα. Αυτό έγινε για να είναι τα κείμενα όσο το δυνατόν περισσότερο συγκρίσιμα.

Τα είκοσι τέσσερα βιβλία κάθε έργου ήταν οι μονάδες ανάλυσης. Κάθε ένα από τα έργα του Ησίοδου χωρίστηκε σε τμήματα 100 γραμμών. Κάθε ένα από τα θεατρικά έργα χωρίστηκε σε έξι τυχαία τμήματα, με το καθένα να περιέχει τον ίδιο αριθμό γραμμών.

Η Ιλιάδα και η Οδύσσεια αποτελούνται από 198,863 λέξεις, τα ποιήματα του Ησίοδου από 16,134 λέξεις και τα θεατρικά έργα του Σοφοκλή από 34,096 λέξεις.

1° Βήμα
Για όλα τα έργα αυτά, υπολογίστηκε η συχνότητα των εξής τριών ειδών λέξεων:

1. **Λέξεις εξαιρετικής συχνότητας**: Οι 100 λέξεις που βρίσκονται πιο συχνά, εκτός από τις λέξεις που αναφέρονται σε ένα κείμενο και όχι σε κάποιο άλλο.

2. **Συχνές λέξεις**: Οι λέξεις που αναφέρονται με συχνότητα του βαθμού 101-200, με εξαίρεση τις προφανείς λέξεις σύνταξης (άρθρα, σύνδεσμοι κλπ).

3. **Σπάνιες λέξεις**: Οι λέξεις που εμφανίζονται με συχνότητα μικρότερη ή ίση με το 0,01% του συνολικού χρόνου του δημιουργήματος (24 ώρες ή λιγότερο για τον Όμηρο, 4 φορές ή λιγότερο για τον Σοφοκλή και 7 φορές ή λιγότερο για τις τραγωδίες).

Τέσσερις εντελώς ανεξάρτητες αναλύσεις πραγματοποιήθηκαν.

Πρώτον, η χρήση των λέξεων στην Ιλιάδα και την Οδύσσεια συγκρίθηκαν. Ο σκοπός ήταν φυσικά να δούμε αν η χρήση λέξεων στα δύο ποιήματα ήταν αρκετά διαφορετική σε σημείο που να δείχνει ότι τα δυο ποιήματα δεν προέρχονταν από τον ίδιο συγγραφέα.

Δεύτερον, έγινε μια σύγκριση στη χρήση συχνών όσο και σπάνιων λέξεων μεταξύ των παλαιών και νεότερων έργων του Σοφοκλή. Ο σκοπός ήταν να βγουν συμπεράσματα για το πόσο αλλάζει η χρήση των λέξεων κατά τη διάρκεια της συγγραφικής ζωής ενός δημιουργού.

Τρίτον, η χρήση λέξεων σε τρία Ησιόδεια ποιήματα συγκρίθηκε. Ο σκοπός ήταν να βγουν συμπεράσματα για το πόσες διαφορές αναμένονταν για ένα συγγραφέα που γράφει στο ίδιο είδος αλλά σε διαφορετικά θέματα.

Τέταρτον, συγκρίθηκε η χρήση λέξεων στα έργα του Αισχύλου, του Ευριπίδη και του Σοφοκλή. Ο σκοπός ήταν να βγουν συμπεράσματα για το αν **η μέθοδος της χρήσης συχνών λέξεων** στα αρχαία ελληνικά κείμενα μπορεί να θεωρηθεί ως μια γνήσια μέθοδος καταλογισμού πατρότητας ενός κειμένου και για το πόσο διαφορετική μπορεί να είναι η χρήση των λέξεων από διαφορετικούς συγγραφείς.

2° Βήμα

Το επόμενο βήμα ήταν να συγκρίνουν τις 100 πιο συχνές λέξεις από τα 48 βιβλία των Ομηρικών επών με όλα τα άλλα βιβλία. Αυτό θα έδειχνε πόσο παρεμφερή είναι όλα αυτά τα βιβλία, ως προς τη χρήση των λέξεων. Δύο βιβλία που θα χρησιμοποιούσαν κάθε μία από τις λέξεις με ακριβώς την ίδια συχνότητα θα είχαν συντελεστή συσχέτισης 1,00.

Η μέθοδος της κλίμακας πολλαπλών διαστάσεων (multidimensional scaling) χρησιμοποιήθηκε στον πίνακα των συντελεστών συσχέτισης. Αυτή η μέθοδος είναι παρόμοια με την εξόρυξη των κυρίων συστατικών από τον πίνακα αποτελεσμάτων. Στη μέθοδο της κλίμακας πολλαπλών διαστάσεων (multidimensional scaling), οι συσχετίσεις εκπροσωπούνται από αποστάσεις και όσο πιο παρεμφερή είναι δυο αντικείμενα, τόσο πιο κοντά είναι το ένα στο άλλο. Στο παράδειγμα αυτό, υπάρχει ένα σύνολο 48 βιβλίων, άρα μια τέλεια αναπαράσταση του συνόλου των συσχετισμών θα απαιτούσε ένα χώρο 47 διαστάσεων. Η απόσταση σε αυτό το χώρο που ορίζεται από αυτές τις διαστάσεις μεταξύ δύο βιβλίων, είναι μια ένδειξη για το πόσο ανόμοια είναι αυτά τα δυο βιβλία μεταξύ τους.

Η ίδια προσέγγιση χρησιμοποιήθηκε με τα άλλα δύο σύνολα κειμένων και δεδομένων. Αν ο στόχος τους ήταν να αποδειχτεί ότι η Ιλιάδα και η Οδύσσεια ήταν από διαφορετικούς συντάκτες, θα περίμενε κανείς πολύ σημαντικά αποτελέσματα για τα έργα αυτά και αλλά μικρά ή ασήμαντα αποτελέσματα για τον Ησίοδο και το Σοφοκλή. Ένα ασήμαντο αποτέλεσμα θα σήμαινε ότι, με βάση τις διαθέσιμες μεταβλητές και δεδομένα, δεν είναι δυνατόν να αναπτυχθεί μια λειτουργία η οποία διαφοροποιεί τα κείμενα από τον ίδιο συγγραφέα. Αυτό θα σήμαινε ότι το μοτίβο της χρήσης λέξεων είναι η ίδια σε όλα τα τμήματα των κειμένων του.

Το συμπέρασμα είναι ότι, η καλύτερη λύση είναι να δημιουργηθεί μια συγκεκριμένη διαδικασία για κείμενα γνωστής πατρότητας και να εφαρμόζεται στα κείμενα αγνώστων συγγραφέων. Ο σκοπός είναι να αναπτυχθεί μια εξίσωση σε ένα κείμενο και στη συνέχεια να επικυρωθεί σε άλλα κείμενα αγνώστων συγγραφέων.

Για παράδειγμα, στον Όμηρο μια τέτοια εξίσωση θα μπορούσε να δημιουργηθεί για τα μονά βιβλία της Ιλιάδας και της Οδύσσειας και να ελεγχθεί στα ζυγά βιβλία και το αντίθετο. Η μέθοδος της κλίμακας πολλαπλών διαστάσεων (multidimensional scaling) απέδωσε μια πεντοδιάσταση λύση.

Η Κανονική ανάλυση παράγει μια διακριτή λειτουργία η οποία συνιστά διάκριση των δύο ποιημάτων. Η έρευνα δείχνει ότι τέσσερα βιβλία της Ιλιάδας μπορούν να αποδοθούν στην Οδύσσεια, επειδή η πιθανότητα είναι μικρότερη από 0,5 ενώ, από την άλλη πλευρά, κανένα βιβλίο της Οδύσσειας δεν μπορεί να αποδοθεί στην Ιλιάδα. Αλλά το μόνο που μας λέει πραγματικά το αποτέλεσμα αυτό είναι ότι αυτά τα 4 βιβλία μοιάζουν περισσότερο στη χρήση του γλώσσας στα βιβλία της Οδύσσειας από ό,τι σε άλλα βιβλία της Ιλιάδας.

Στη συνέχεια, η παραπάνω ανάλυση έγινε στα ζυγά βιβλία και η εξίσωση που παρήχθη εφαρμόστηκε στα μονά βιβλία. Η ίδια μέθοδος της διακριτικής ανάλυσης εφαρμόστηκε και στα ποιήματα του Σοφοκλή και του Ησίοδου.

Η σύγκριση των τεσσάρων συνόλων των αποτελεσμάτων δείχνει ότι η Ιλιάδα και η Οδύσσεια διαφέρουν περισσότερο από ότι τα έργα του Αισχύλου και του Ευριπίδη. Αυτό φαίνεται να δείχνει ότι είναι από διαφορετικούς συγγραφείς.

Η έλλειψη διαφορών ανάμεσα στα πρώιμα και τα τελευταία έργα του Σοφοκλή, υποδεικνύει ότι είναι μάλλον απίθανο ότι η Ιλιάδα και η Οδύσσεια θα μπορούσαν να είναι τα πρώιμα και τελευταία έργα ενός συγγραφέα.

Η έλλειψη διαφορών μεταξύ των ποιημάτων του Ησίοδου καθιστά απίθανο ότι οι διαφορές μεταξύ της Ιλιάδας και της Οδύσσειας προκύπτουν απλώς από διαφορές στο θέμα των βιβλίων.

Η μέθοδος της κλίμακας πολλαπλών διαστάσεων (multidimensional scaling) επίσης εφαρμόστηκε σε λέξεις με συχνότητα 101-200 όπως και σε σπάνιες λέξεις. Το αποτέλεσμα ήταν το ίδιο.

Η Ιλιάδα και η Οδύσσεια είναι πολύ διαφορετικά μεταξύ τους και οι διαφορές στη χρήση λέξεων είναι ένα πολύ καλό παράδειγμα. Οι διαφορές που διαπιστώθηκαν ήταν τέτοιες που το συμπέρασμα είναι ότι, αν ο 'Ομηρος έγραψε την Ιλιάδα, τότε είναι απίθανο να συνέθεσε και την Οδύσσεια. Ασφαλώς όμως είναι της γνώμης ότι χρειάζεται περισσότερη έρευνα για πιο σίγουρα αποτελέσματα.

3.5 Προτεινόμενη Μεθοδολογία – Τι μεθοδολογία εφαρμόσαμε

Το πρόβλημα που αντιμετωπίζουμε μπορεί να προσεγγιστεί με διαφορετικό τρόπο και χωρίς να χρησιμοποιηθούν παραδοσιακές τεχνικές, γι 'αυτό και χρησιμοποιήθηκαν τεχνικές εξόρυξης δεδομένων και ειδικότερα κανόνες σύνδεσης. [41]

Η διαδικασία ξεκίνησε με τον έλεγχο των δύο ποιημάτων για να επιβεβαιωθεί ότι δε διαφέρουν στην ορθογραφία. Τότε όλες οι απόστροφοι, διακριτικά και σημεία στίξης αφαιρέθηκαν από όλα τα κείμενα. Αυτό έγινε για να είναι τα κείμενα συγκρίσιμα. Φυσικά, όλα τα αλφαβητικά σημεία και οι υποσημειώσεις διατηρήθηκαν και τα σημεία στίξεως διεγράφησαν για να γίνουν οι υπολογισμοί μας ευκολότεροι. Μετά από τις μορφολογικές αναλύσεις των κείμενων μας, όπως μπορούμε να ονομάσουμε ότι αναφέραμε πιο πάνω, καταλήξαμε σε δυο καθαρά κείμενα που είναι οι πηγές της ανάλυσης.

Καθένα από τα δύο κείμενα χωρίστηκε σε ενότητες. Αυτή η διαχώριση έγινε με τον εξής τρόπο:

1. **Τμήμα 1-παραγράφου** σύμφωνα με την κατανομή που είχε ήδη γίνει στο αρχικό κείμενο.
2. Για να καταλήξουμε σε πιο ακριβή αποτελέσματα, συνεχίσαμε τη διαίρεση σε **τμήματα 2-παραγράφων** και **τμήματα 1-πρότασης.**
Το σύνολο όλων των τμημάτων για κάθε έγγραφο, ορίζει το σύνολο δεδομένων.

Το επόμενο βήμα είναι να καθοριστεί ένα σύνολο κεντρικών εννοιών των λέξεων-κλειδιών. Αφού έγινε αυτό, χρησιμοποιώντας το Wordnet, μπορούμε να αναθέσουμε κάθε έννοια σε κάθε τμήμα. (αναλύουμε παρακάτω για ποιο λόγο επιλέχθηκαν οι συγκεκριμένες λέξεις).

Το **Wordnet** είναι ένα γλωσσικό εργαλείο που παρέχει μια λίστα σχετικών εννοιών για κάθε λέξη. Άρα, αν εισάγαμε στο Wordnet κάθε λέξη των καθαρών κειμένων που δημιουργήσαμε, τότε το παράγωγο του Wordnet θα μας έδειχνε αν κάποιο από τα τμήματα συσχετιζόταν με κάποια συγκεκριμένη έννοια.(Στο παράρτημα παραθέτουμε τα αποτελέσματα που προέκυψαν με την χρήση αυτού του εργαλείου)

Στη συνέχεια, στο τελικό στάδιο χρησιμοποιήθηκε ο αλγόριθμος Apriori για την εξαγωγή των κανόνων συσχέτισης για κάθε σύνολο δεδομένων.

Θεωρούμε κανόνες συσχέτισης με τη μορφή "90% των συναλλαγών που περιέχουν την έννοια Χ επίσης περιέχουν την έννοια Υ".(έχει αναλυθεί σε προηγούμενη παράγραφο).

Στο παρόν σενάριο, ο κανόνας συσχέτισης εξηγεί τη σύνδεση μεταξύ δύο ή περισσοτέρων αντικειμένων.

Ας δούμε μερικά παραδείγματα κανόνων συσχετίσεις που προέκυψαν από την ανάλυση.

Για παράδειγμα: Στο 80% των περιπτώσεων που ο Όμηρος χρησιμοποιεί τη λέξη γη, επίσης χρησιμοποιεί τη λέξη ηλικία. Αυτό μας δίνει πληροφορίες για τη σχέση μεταξύ γης και ηλικίας.

Εμείς το εκπροσωπούμε αυτό ως: **Γη => Ηλικία | 80%.**
Αυτό θα πρέπει να διατυπωθεί ως:
"Γη σημαίνει ή υπονοεί ηλικία, 80% των περιπτώσεων".
Σε αυτή την περίπτωση, ο " συντελεστής εμπιστοσύνης" για αυτό τον κανόνα είναι 80%.

Οι κανόνες σύνδεσης μπορούν να υπάρχουν για περισσότερα από 2 αντικείμενα.

Για παράδειγμα

Γη, Άντρας => Ηλικία | 60%

Γη => Άντρας, Ηλικία | 40%

Για κάθε κανόνα, μπορούμε να βρούμε εύκολα το συντελεστή εμπιστοσύνης του.

Για παράδειγμα, για το **"Γη, Άντρας => Ηλικία | 60%"** , μετράμε τον αριθμό των εγγραφών που περιέχουν τις λέξεις γη και τον άντρα, και το ονομάζουμε n_1.
Από αυτές, πόσες επίσης περιέχουν τη λέξη ηλικία?
Ας ονομάσουμε αυτό τον αριθμό n_2. Σε αυτή τη περίπτωση, ο συντελεστής εμπιστοσύνης είναι: n_2/n_1.

Με αυτή την διαδικασία λοιπόν πετύχαμε ένα μεγάλο αριθμό ισχυρών συσχετισμών μεταξύ ταυτόσημων εννοιών και από τα δύο ποιήματα (π.χ. μεταξύ των "γη" και "άντρας"). Υπάρχουν επίσης συσχετισμοί μεταξύ διαφορετικών εννοιών στα δύο ποιήματα (π.χ. μεταξύ των "πάλη" και "άντρας" μόνο στην Ιλιάδα), καθώς και διαφορετικοί συσχετισμοί για την ίδια έννοια (π.χ. μεταξύ των «ήρωας» και «μάχη» στην Ιλιάδα και "ήρωας» και «σπίτι» στην Οδύσσεια).
Αντιθέτως, δεν έχουμε βρει κάποιες αντιφάσεις.
Αυτά τα αποτελέσματα υποδεικνύουν ότι ο Όμηρος έγραψε και τα δύο αυτά ποιήματα.

3.6 Εμπειρικά αποτελέσματα

Πρώτα από όλα, ας δούμε τις κεντρικές έννοιες που εμείς επιλέξαμε να αναζητήσουμε σε αυτά τα δύο ποιήματα. Η επιλογή τους έγινε με βάση το θέμα και το περιεχόμενο των δύο ποιημάτων.

Πρόκειται για τις ακόλουθες έννοιες: Πλοία, πόλεμος, πάλη, σπίτι, άνθρωπος, θεοί, ήρωας, θάλασσα, θάνατος, άλογα, μητέρα, φωτιά, ήλιος, ταξίδι, γενναίος, σύντροφος, δύναμη, γη, τιμή, φόβος, ηλικία, ξένος.

Τότε, όπως έχουμε ήδη αναφέρει, με τη βοήθεια του WordNet βρήκαμε σε κάθε τμήμα τις σχετικές έννοιες. Στη συνέχεια χρησιμοποιήσαμε τον αλγόριθμο Apriori και βρήκαμε όλους τους πιθανούς κανόνες συσχέτισης σε κάθε ποίημα.

Με τη σύγκριση αυτών των αποτελεσμάτων Apriori μεταξύ των δύο ποιημάτων, καταλήγουμε στο συμπέρασμα ότι υπάρχουν πολλές ομοιότητες και ότι πολλές συσχετίσεις είναι κοινές και στα δύο ποιήματα. Για παράδειγμα, όταν οι έννοιες του ανθρώπου και της γης χρησιμοποιούνται μέσα στο κείμενο, τότε αναπόφευκτα, και στα δύο ποιήματα, η έννοια της ηλικίας χρησιμοποιείται

επίσης. Πιο κάτω παραθέτονται μερικά από τα αποτελέσματα που ήρθαν στο φως μέσα από την έρευνά μας:

3.6.1 Τα αποτελέσματα στην Ιλιάδα

Αν ήρωας τότε άνθρωπος. Εμπιστοσύνη = 97% και Υποστήριξη = 24%
Αν άνθρωπος τότε ηλικία. Εμπιστοσύνη = 97% και Υποστήριξη = 25%
Αν ήρωας τότε ηλικία. Εμπιστοσύνη = 97% και Υποστήριξη = 24%
Αν ήρωας, ηλικία τότε άνθρωπος. Εμπιστοσύνη = 97% και Υποστήριξη = 24%
Αν άνθρωπος, ήρωας, τότε ηλικία. Εμπιστοσύνη = 98% και Υποστήριξη = 24%
Αν άνθρωπος, γη, τότε ηλικία. Εμπιστοσύνη = 98% και Υποστήριξη = 24%

3.6.2 Τα αποτελέσματα στην Οδύσσεια

Αν σπίτι τότε ηλικία. Εμπιστοσύνη = 99% και Υποστήριξη = 22%
Αν άνθρωπος τότε ηλικία. Εμπιστοσύνη = 98% και Υποστήριξη = 24%
Αν ήρωας τότε άνθρωπος. Εμπιστοσύνη = 97% και Υποστήριξη = 23%
Αν γη τότε ηλικία. Εμπιστοσύνη = 97% και Υποστήριξη = 25%
Αν μητέρα τότε ήρωας. Εμπιστοσύνη = 98% και Υποστήριξη = 22%
Αν άνθρωπος, γη, τότε ηλικία. Εμπιστοσύνη = 98% και Υποστήριξη = 23%

Τα παραπάνω αποτελέσματα αποτελούν ένα δείγμα των κανόνων των συσχετίσεων που είχαν εξαχθεί. Ανάμεσά τους, καθώς υπήρχαν πολλά άλλα αποτελέσματα, παρουσιάζουμε τρεις κατηγορίες των κανόνων συσχέτισης.

3.7 Κατηγορίες Κανόνων συσχέτισης

Ισχυροί συσχετισμοί:
(υπάρχουν και στα δυο ποιήματα)

Ιλιάδα: Αν άνθρωπος τότε γη. Εμπιστοσύνη = 88% και Υποστήριξη = 22%
Οδύσσεια: Αν άνθρωπος τότε γη. Εμπιστοσύνη = 87% και Υποστήριξη = 29%
Ιλιάδα: Αν μητέρα τότε ήρωας. Εμπιστοσύνη = 97% και Υποστήριξη = 20%
Οδύσσεια: Αν μητέρα τότε ήρωας. Εμπιστοσύνη = 96% και Υποστήριξη = 26
Ιλιάδα: Αν άνθρωπος, ηλικία, τότε γη. Εμπιστοσύνη = 96% και Υποστήριξη = 17%
Οδύσσεια: Αν άνθρωπος, ηλικία, τότε γη. Εμπιστοσύνη = 94% και Υποστήριξη = 23%

Συσχετισμοί
(υπάρχουν μόνο σε ένα από τα δυο ποιήματα)

Αν ήρωας τότε μάχη. Εμπιστοσύνη = 92% και Υποστήριξη = 22%

Αν μάχη τότε γη. Εμπιστοσύνη = 95% και Υποστήριξη = 21% (βρίσκεται μόνο στην Ιλιάδα).
Αν θεοί τότε άνθρωπος. Εμπιστοσύνη = 94% και Υποστήριξη = 23% (βρίσκεται μόνο στην Οδύσσεια).

Συσχετισμοί που είναι διαφορετικοί για την ίδια έννοια:

Αν ήρωας τότε μάχη. Εμπιστοσύνη = 97% και Υποστήριξη = 24%, στην Ιλιάδα.
Αν ήρωας τότε σπίτι. Εμπιστοσύνη = 90% και Υποστήριξη = 26%, στην Οδύσσεια.

3.8 Συμπεράσματα

Παρουσιάζουμε μια μεθοδολογία για την αντιμετώπιση του προβλήματος ανάθεσης συγγραφέα, που στηρίζεται στην εξόρυξη δεδομένων. Αναλύουμε μια εφαρμογή αυτής της μεθοδολογίας στο "Ομηρικό ερώτημα". Επίσης, αποδεικνύουμε τα αποτελέσματα των αναλύσεων που πραγματοποιήσαμε.

Το ουσιώδες μέρος της έρευνάς μας είναι ότι δεν βρήκαμε αντιφάσεις μεταξύ των δύο ποιημάτων, κατά τη διάρκεια της σύμπραξης εννοιών. Ένας μεγάλος αριθμός ισχυρών συσχετίσεων μεταξύ κοινών εννοιών και από τα δυο ποιήματα, ήταν ένα από τα κυριότερα αποτελέσματα. Υπάρχουν επίσης συσχετισμοί μεταξύ διαφορετικών εννοιών στα δυο ποιήματα αλλά και διαφορετικοί συσχετισμοί για τις ίδιες έννοιες. Αντιθέτως, δεν έχουμε βρει κάποιες αντιφάσεις. Τα αποτελέσματα αυτά φαίνεται να υποδεικνύουν ότι ο Όμηρος έγραψε και τα δύο αυτά ποιήματα.

Κεφάλαιο 4ο

4.1 Άλλες πιθανές εφαρμογές της προτεινόμενης μεθοδολογίας

4.1.1 Αρχαιότερα προβλήματα

1. Όντως ο Όμηρος έχει γράψει τόσο την Ιλιάδα όσο και την Οδύσσεια; Και τα δυο έπη αποδίδονται σε έναν και μοναδικό συγγραφέα τον «Όμηρο», αλλά και τα δυο έργα προέρχονται από τα πολύ παλιά χρόνια σε προφορική παραδόσεις και αποδόσεις.
2. Ο Πλάτωνας ανέπτυξε την φιλοσοφία του με τη μορφή διαλόγων, διαμέσου στόματος του Σωκράτη ο οποίος ήταν και ο Δάσκαλός του. Διαπιστώνοντας τη σωστή χρονολογική σειρά των διαλόγων αυτών, θα μας βοηθήσει στην πορεία να κατανοήσουμε πώς ο Πλάτων ανέπτυξε τη φιλοσοφία του.
3. Ο Σαίξπηρ από την άλλη μεριά έχει γράψει όλα τα έργα του; Θεωρείται ότι διάφοροι συγγραφείς συμπεριλαμβανομένων και των Μπέικον και Marlowe, έγραψαν όλα ή ένα μέρος των έργων.

4.1.2 Νεότερα προβλήματα

Τα Federalist Papers είναι μια σειρά άρθρων που δημοσιεύθηκαν στο 1787-88, με στόχο την προώθηση της κύρωσης του νέου Συντάγματος των ΗΠΑ και είναι γραμμένα από τρεις συγγραφείς, Jay, Χάμιλτον και Madison, με το ψευδώνυμο "Πόπλιος". [11] Μερικά είναι γνωστά (και σε ορισμένες περιπτώσεις από κοινού) ως την προέλευσή τους αλλά άλλα έχουν αμφισβητηθεί.

Πρωτοποριακοί μέθοδοι της στυλομετρίας που χρησιμοποιήθηκαν από τους γνωστούς Mosteller και Wallace στις αρχές της δεκαετίας του 1960 επιχείρησαν να απαντήσουν στο ερώτημα αυτό και τα κατάφεραν. Το πρόβλημα λοιπόν που αφορούσε τα Federalist Papers παρουσίασε μεγάλη δυσκολία στην επίλυσή του, αλλά θεωρείται πλέον σαν σημείο αναφοράς για τη δοκιμή νέων ιδεών.

Άλλα θέματα τα οποία αφορούν το πρόβλημα πατρότητας κειμένου

4.1.3 Άλλες περιπτώσεις προβλημάτων πατρότητας κειμένου

4.1.3.1 ΤΟ ΕΥΑΓΓΕΛΙΟ ΤΟΥ ΙΩΑΝΝΗ

Τα προβλήματα απόδοσης κειμένου σε ένα συγγραφέα έχει σκοπό να αποδείξει αν ένα επίμαχο κείμενο είναι αρκετά κοντά στο ύφος με αυτό ενός αδιαμφισβήτητου συγγράμματος. Έχει αποδειχθεί, χρησιμοποιώντας αντιθέσεις ενός γραμμικού μοντέλου βασισμένου σε τυχαία αποτελέσματα, ότι τέτοιες καταστάσεις εμφανίζονται συχνά σε περιπτώσεις κλασικής άλλα αρχαίας γραφής.

Στο [28], η προσέγγιση αυτή χρησιμοποιήθηκε για να λύσει ένα πρόβλημα που προέρχεται από τη χριστιανική Αγία Γραφή. Στον τελευταίο αιώνα, έχουν υπάρξει πολλοί κριτικοί που υποστηρίζουν ότι υπάρχει κάποια άλλη άγνωστη πηγή για ένα σημαντικό ποσό των μερίδων αφήγησης του Ευαγγελίου του Ιωάννη.

Μια πρόταση για μια πιθανή πηγή των ιστοριών των θαυμάτων σε αυτό το ευαγγέλιο, είναι του Fortna's Gospel of Signs [29] (στο εξής FGOS). Η ανασυγκρότηση της κάπως σκοτεινής και φευγαλέας αυτής πηγής φαίνεται να απολαμβάνει στήριξη από πολλούς ειδικούς ερευνητές και το γεγονός ότι το κείμενό της FGOS είναι διαθέσιμο στα Ελληνικά, επιτρέπει τη λεπτομερή εξέταση των στατιστικών στοιχείων. Το Ευαγγέλιο του Ιωάννη χωρίστηκε σε τρία κύρια τμήματα, ίσου μεγέθους. Τα τρία τμήματα ήταν το F (που προέρχεται από το FGOS), το N (που προέρχεται από τα τμήματα αφήγησης του Ευαγγελίου) και το D (που προέρχεται από τα ομιλητικά τμήματα του ευαγγελίου). Κάθε κομμάτι άρχισε από τις αρχές μιας πρότασης, που για την παρούσα έρευνα ορίζεται ως μια σειρά λέξεων που μπορούσε να ολοκληρωθεί από τελείες, άνω τελείες ή ερωτηματικά. Προτάσεις στις οποίες περιέχονται αναγνωρίσιμα αποσπάσματα από τις Εβραϊκές Γραφές έχουν παραλειφθεί από την καταμέτρηση. Είκοσι τέσσερις μεταβλητές που είναι εύκολο να μετρηθούν χρησιμοποιήθηκαν για τη μέτρηση του ύφους.

Μια τροποποιημένη μορφή της συνηθισμένης μεθόδου στατιστικής pooled t χρησιμοποιήθηκε για να ελεγχθεί η υπόθεση ότι το ύφος του F (ανασυγκρότηση της Fortna) διαφέρει σημαντικά από τις άλλες δύο, αφήνοντας περιθώρια για το γεγονός ότι και οι άλλες δύο (Ν και Δ) μπορεί επίσης να διαφέρουν σημαντικά μεταξύ τους. Ο συγγραφέας συμπέρανε ότι η ισορροπία των πιθανοτήτων ήταν ότι το ύφος του F διέφερε από εκείνο του Ν και Δ περισσότερο από ότι θα μπορούσε να αναμένεται από την τύχη, ακόμη και όταν προσπαθούμε να λάβουμε υπόψη το είδος του συγγράμματος.

Ωστόσο, η απόφαση δεν ήταν τόσο απλή όσο θα περίμενε ή έλπιζε κανείς. Εάν υπήρχαν περισσότερα αναμφισβήτητα κείμενα (για παράδειγμα 8) που ήταν διαθέσιμα, το αποτέλεσμα θα μπορούσε κάλλιστα να ήταν αρκετά σημαντικό. Επίσης, αν είχαν χρησιμοποιηθεί πιο σύνθετες (γλωσσολογικά μιλώντας) μεταβλητές, ένα πιο αποφασιστικό αποτέλεσμα θα μπορούσε κάλλιστα να έχει επιτευχθεί.

4.1.3.2 ROMAIN GARY/EMILE AJAR

Ένα άλλο παράδειγμα προβλήματος απόδοσης κειμένου σε συγγραφέα έχει να κάνει με τη σύγκριση των δύο μυθιστορημάτων όσον αφορά το ύφος τους [8] (χωρίς να αποδίδονται τα έργα σε κάποιον συγκεκριμένο συγγραφέα).

Ο Γάλλος συγγραφέας Romain Gary, ο αποδέκτης του κορυφαίου Γαλλικού λογοτεχνικού βραβείου, εξέδωσε το βιβλίο Gros - Καλίν με το ψευδώνυμο Emile Ajar. Η πρόθεσή του ήταν να ξεκινήσει από την αρχή και να κριθεί το έργο του βάσει της ποιότητας του και όχι λόγω της φήμης του. Αυτό το μυθιστόρημα προσέλκυσε την προσοχή των κριτικών και των αναγνωστών και έγινε ένα ανάρπαστο best seller.

Όταν κάποιοι παρατήρησαν ομοιότητες μεταξύ του ύφους του Gary και του Ajar, άρχισαν να διαδίδονται φήμες ότι οι δυο συγγραφείς ήταν το ίδιο άτομο. Η αντίδραση του Gary ήταν να γράψει ένα άλλο μυθιστόρημα που ονομαζόταν La Vie devant soi και είχε διαφορετικό ύφος από το πρώτο για να παραπλανήσει τους κριτικούς και τους αναγνώστες. Στη συνέχεια ακολούθησαν αλλά δύο μυθιστορήματα υπό την επωνυμία του Ajar.

Αναπόφευκτα, ο Gary προσφέρει ένα εξαιρετικό παράδειγμα για μελέτη απόδοσης συγγραφέως ή stylometric μελέτη. Η απόδοση πατρότητας είναι μια μελέτη που χρησιμοποιεί τις υφολογικές ιδιαιτερότητες και ιδιοσυγκρασίες ενός συγγραφέα ως δείκτη της αυθεντικότητας και προσπαθεί να συλλάβει ποσοτικά την ουσία της χρήσης της γλώσσας από κάθε συγγραφέα.

Προκειμένου να ελεγχθεί αν πράγματι το κείμενο και ύφος του Gary είναι στατιστικά παρόμοια με το αυτά του Ajar, τέσσερα βιβλία εξετάσθηκαν, δύο από τον Romain Gary και δύο από τον Émile Ajar. Επειδή η λογοτεχνική καριέρα του Gary Romain διήρκησε σχεδόν τριάντα χρόνια, αυτά τα τέσσερα βιβλία επιλέχθηκαν γιατί γράφτηκαν μέσα σε μια περίοδο τεσσάρων ετών και έτσι θα μπορούσε να αποφευχθεί το πρόβλημα της χρονολόγησης.

Όταν δημιουργήθηκε μια γραφική παράσταση που κατέγραφε το πόσες φορές, ποσοστιαία, συγκεκριμένες λέξεις εμφανίζονταν στο κείμενο, κάποια συμπεράσματα ήταν πασιφανή. Το δεύτερο μυθιστόρημα του Ajar, La Vie devant soi, παρέκκλινε σημαντικά από τα άλλα μυθιστορήματα του Gary καθώς και όλα τα άλλα μυθιστορήματα. Τα στατιστικά αποτελέσματα που βρέθηκαν αποδεικνύουν πέραν πάσης αμφιβολίας, ότι ο συγγραφέας μπορεί πράγματι να χειραγωγήσει συνειδητά λέξεις υψηλής συχνότητας και ζεύγη συνωνύμων, τα οποία θεωρούνται, από πολλούς ως τα υποσυνείδητα στοιχεία του ύφους ενός συγγραφέα.

Η άποψη λοιπόν ότι οι κύριες λέξεις (και τα συνώνυμα τους) αποτελούν το γενετικό αποτύπωμα του ύφους ενός συγγραφέα, αμφισβητείται στην υπόθεση Romain Gary - Émile Ajar. Ενώ το Gros-Câ-lin, το πρώτο μυθιστόρημα του Ajar μοιάζει πολύ με τα δύο μυθιστορήματα του Gary, το La Vie devant soi είναι τόσο σημαντικά διαφορετικό από τα αλλά δυο μυθιστορήματα του Gary, που θα μπορούσε πολύ εύκολα να είχε γραφτεί από κάποιον άλλο συγγραφέα.

Βιβλιογραφία

[1] W. Frawley and G. Piatetsky-Shapiro and C. Matheus (Fall 1992). "Knowledge Discovery in Databases: An Overview". AI Magazine: pp. 213-228. ISSN 0738-4602.

[2] D. Hand, H. Mannila, P. Smyth (2001). Principles of Data Mining. MIT Press, Cambridge.MA

[3] Ellen Monk, Bret Wagner (2006). Concepts in Enterprise Resource Planning, Second Edition. Thomson Course Technology, Boston, MA. ISBN 0-619-21663-8. OCLC 224465825.

[4] Albion Research, Market Basket Analysis. http://www.albionresearch.com/data_mining/market_basket.php

[5] Gayle S., The Marriage of Market Basket Analysis to Predictive Modeling

[6] Agrawal R., Imielinski T., Swami A. *Mining Association Rules Between Sets of Items in Large Databases*

[7]D. Holmes "Authorship attribution" *Computers and the Humanities* 28 (1994), 87-106.

[8]D. Holmes "The Evolution of Stylometry in Humanities Scholarship" *Literary and Linguistic Computing* 13 (1998), 111-117. http://llc.oxfordjournals.org/cgi/reprint/13/3/111.pdf

[9]T. McEnery & M. Oates "Authorship identification and computational stylometry" in Dale et al (eds) *Handbook of Natural Language Processing*, New York (2000): Dekker, chapter 23.1

[10] http://en.wikipedia.org/wiki/Stylometry

[11] ^ F. Mosteller and D. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.

[12]http://www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf

[13]http://www.icaen.uiowa.edu/~comp/Public/Apriori.pdf

[14]http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput499/slides/Lect10/sld044.htm

[15] ^ A. Gionis, H. Mannila, T. Mielikainen, and P. Tsaparas, Assessing Data Mining Results via Swap Randomization, ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 1 , Issue 3 (December 2007) Article No. 14.

[16] ^ Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation. Data Mining and Knowledge Discovery 8:53-87, 2004.

[17] R. Agrawal, H. Mannila, R. Srikant, A.I. Verkamo, "Fast Discovery of Association Rules", in 3., pp. 307-328, 1996.

[18]J.F. Burrows, "Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method", Oxford: Clarendon P., 1987

[19] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "Advances in Knowledge Discovery and Data Mining", AAAI Press/MIT Press, 1996.

[20] R. Felton, "A New procedure for Author Attribution", Joint International Conference ALLC-ACH'96 Abstracts, pp. 74-75, 1996.

[21]R.T. Fortna, "The Gospel of Signs: A Reconstruction of the Narrative Source Underlying the Fourth Gospel", Cambridge University Press, 1970.

[22] C. Martindale, P. Tuffin, "If Homer is the Poet of the Iliad, then he may not be the Poet of the Odyssey", Literary and Linguistic Computing, 11(3), pp. 109-120, 1996.

[23]F. Mosteller, D.L. Wallace, "Inference and Disputed Authorship: The Federalist", Reading, MA: Addison-Wesley, 1964.

[24]V. Tirvengadum, "Two Methods of Author Identification: the Gary/Ajar case", Joint International Conference ALLC-ACH'96 Abstracts, 1996.

[25] D. Holmes "Authorship attribution" *Computers and the Humanities* 28 (1994), 87-106.

[26] D. Holmes "The Evolution of Stylometry in Humanities Scholarship" *Literary and Linguistic Computing* 13 (1998), 111-117. http://llc.oxfordjournals.org/cgi/reprint/13/3/111.pdf

[27] T. McEnery & M. Oates "Authorship identification and computational stylometry" in Dale et al (eds) *Handbook of Natural Language Processing*, New York (2000): Dekker, chapter 23.

[28] R. Felton, "A New procedure for Author Attribution", Joint International Conference ALLC-ACH'96 Abstracts, pp. 74-75, 1996.

[29] R.T. Fortna, "The Gospel of Signs: A Reconstruction of the Narrative Source Underlying the Fourth Gospel", Cambridge University Press, 1970.

[30] Ian H. Witten
Computer Science, University of Waikato, Hamilton, New Zealand
email ihw@cs.waikato.ac.nz

[31] *http://people.ischool.berkeley.edu/~hearst/text-mining.html*

[32]http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

[33] Hearst Marti A. "untangling Text Data Mining", Proceedings of ACL'99:the 37th Annual meeting of the association for computational Linguistics, University of Maryland, June 20-26,1999.

[34] Karanikas H., Theodoulidis B. "Knowledge Discovery in text and text mining software". Centre for Research in information management: November 2002.

[35] Mooney R.J and Nahm Un Yong, Text Mining with information extraction. Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium,22-23 September 2003.

[36] Besancon R & Rajman M (1998). Text Mining knowledge Extraction from unstructured textual data. 6th conference of international Federation Of classification Societies, Rome

[37] Sehgal, A.K Text Mining: The search for novelty in text Ph.D.Comprehensive Examination Report, Dept. of Computer Science, The university of Iowa, April 2004

[38] Hearst M. 2003: What is text Mining? http://www.sims.berkeley.edu/~heasrt/text -mining.html, October 2003.

[39] Sehgal A.K, Qui X.Y and Srinivasan P. Mining Medline metadata to explore genes and their connections. In proceedings of the SIGIR 2003. Workshop on Text Analysis and Search for Bioinformatics, July 2003.

[40] Sharp M, Text mining. Seminar in information Studies, Prof. Tefko Saracevic. 11 December 2001

[41] (D.Plotas,D.Tasoulis,B.Boutsinas,"Are the Iliad and Odyssey each work of a single poet? " 2nd World Congress"Ancient Greece and the Modern World", July, Ancient Olympia, Greece, 2002,pp. 617-623.)

Παράρτημα Α

Ακολουθεί το σύνολο των λέξεων που προέκυψαν από την χρήση του Wordnet. (Το **Wordnet** είναι ένα γλωσσικό εργαλείο που παρέχει μια λίστα σχετικών εννοιών για κάθε λέξη.) Άρα, αν εισάγαμε στο Wordnet κάθε λέξη των καθαρών κειμένων που δημιουργήσαμε, τότε το παράγωγο του Wordnet θα μας έδειχνε αν κάποιο από τα τμήματα συσχετιζόταν με κάποια συγκεκριμένη έννοια)

0) ships vessel
1) War battle warfare conflict
2) Fight arm weapon spear armour bow arrow
3) House region location
4) Man mortal person individual entity human man servant
5) Gods immortal leader spiritual being divinity deity
6) Hero leader mortal person human champion fighter defender
7) Sea ocean flow sail
8) Death kills
9) Horses soldiery
10) Mother parent mortal ancestor
11) Fire attack flame onslaught burning
12) Sun star sunshine
13) Voyage moralising ira wrath
14) Brave warrior
15) Comrade Brother friend
16) Power strength ability
17) land country state domain nation domain
18) Honour virtue award honour morality
19) Fear anxiety fright
20) Age long-time year's property era old age
21) Stranger traveller intruder

Παράρτημα Β

Προτάσεις _ 2_ Ιλιάδας 50/50

If man then fight .Confidence = 0,706356287002563 and Support = 789

If fight then man .Confidence = 0,822732031345367 and Support = 789

If hero then fight .Confidence = 0,711466193199158 and Support = 757

If fight then hero .Confidence = 0,789363920688629 and Support = 757

If land then fight .Confidence = 0,734307825565338 and Support = 854

If fight then land .Confidence = 0,890510976314545 and Support = 854

If hero then man .Confidence = 0,942669153213501 and Support = 1003

If man then hero .Confidence = 0,897940933704376 and Support = 1003

If land then man .Confidence = 0,844368040561676 and Support = 982

If man then land .Confidence = 0,879140555858612 and Support = 982

If age then man .Confidence = 0,853191494941711 and Support = 802

If man then age .Confidence = 0,717994630336761 and Support = 802

If mother then man .Confidence = 0,961661338806152 and Support = 903

If man then mother .Confidence = 0,808415412902832 and Support = 903

If voyage then man .Confidence = 0,953035533428192 and Support = 832

If man then voyage .Confidence = 0,744852304458618 and Support = 832

If land then hero .Confidence = 0,806534826755524 and Support = 938

If hero then land .Confidence = 0,881578922271729 and Support = 938

If age then hero .Confidence = 0,810638308525085 and Support = 762

If hero then age .Confidence = 0,71616542339325 and Support = 762

If mother then hero .Confidence = 0,966986179351807 and Support = 908

If hero then mother .Confidence = 0,853383481502533 and Support = 908

If voyage then hero .Confidence = 0,966781198978424 and Support = 844

If hero then voyage .Confidence = 0,793233096599579 and Support = 844

If age then land .Confidence = 0,943617045879364 and Support = 887

If land then age .Confidence = 0,762682735919952 and Support = 887

If voyage then mother .Confidence = 0,931271493434906 and Support = 813

If mother then voyage .Confidence = 0,865814685821533 and Support = 813

If hero, land then man .Confidence = 0,948827266693115 and Support = 890

If man, land then hero .Confidence = 0,90631365776062 and Support = 890

If man, hero then land .Confidence = 0,887337982654572 and Support = 890

If hero, mother then man .Confidence = 0,982378840446472 and Support = 892

If man, mother then hero .Confidence = 0,987818360328674 and Support = 892

If man, hero then mother .Confidence = 0,889331996440887 and Support = 892

If hero, voyage then man .Confidence = 0,976303339004517 and Support = 824

If man, voyage then hero .Confidence = 0,990384638309479 and Support = 824

If man, hero then voyage .Confidence = 0,821535408496857 and Support = 824

If land, age then man .Confidence = 0,865839898586273 and Support = 768

If man, age then land .Confidence = 0,957605957984924 and Support = 768

If man, land then age .Confidence = 0,782077372074127 and Support = 768

If mother, voyage then man .Confidence = 0,99138993024826 and Support = 806

If man, voyage then mother .Confidence = 0,96875 and Support = 806

If man, mother then voyage .Confidence = 0,892580270767212 and Support = 806

If mother, voyage then hero .Confidence = 0,9963099360466 and Support = 810

If hero, voyage then mother .Confidence = 0,959715664386749 and Support = 810

If hero, mother then voyage .Confidence = 0,892070472240448 and Support = 810

If hero, mother, voyage then man .Confidence = 0,995061755180359 and Support = 806

If man, mother, voyage then hero .Confidence = and Support = 806

If man, hero, voyage then mother .Confidence = 0,978155314922333 and Support = 806

If man, hero, mother then voyage .Confidence = 0,903587460517883 and Support = 806

Προτάσεις _2_ Οδύσσειας 50/50

If man then house .Confidence = 0,759036123752594 and Support = 882

If house then man .Confidence = 0,86982250213623 and Support = 882

If land then house .Confidence = 0,770520746707916 and Support = 873

If house then land .Confidence = 0,860946774482727 and Support = 873

If age then house .Confidence = 0,792569637298584 and Support = 768

If house then age .Confidence = 0,757396459579468 and Support = 768

If hero then house .Confidence = 0,755470991134644 and Support = 794

If house then hero .Confidence = 0,783037483692169 and Support = 794

If mother then house .Confidence = 0,76040607690812 and Support = 749

If house then mother .Confidence = 0,738658785820007 and Support = 749

If land then man .Confidence = 0,888790845870972 and Support = 1007

If man then land .Confidence = 0,866609275341034 and Support = 1007

If age then man .Confidence = 0,882352948188782 and Support = 855

If man then age .Confidence = 0,735800325870514 and Support = 855

If hero then man .Confidence = 0,96860134601593 and Support = 1018

If man then hero .Confidence = 0,876075744628906 and Support = 1018

If mother then man .Confidence = 0,964466989040375 and Support = 950

If man then mother .Confidence = 0,817555963993073 and Support = 950

If voyage then man .Confidence = 0,97042715549469 and Support = 886

If man then voyage .Confidence = 0,762478470802307 and Support = 886

If age then land .Confidence = 0,927760601043701 and Support = 899

If land then age .Confidence = 0,793468654155731 and Support = 899

If mother then hero .Confidence = 0,964466989040375 and Support = 950

If hero then mother .Confidence = 0,903901040554047 and Support = 950

If voyage then hero .Confidence = 0,960569560527802 and Support = 877

If hero then voyage .Confidence = 0,834443390369415 and Support = 877

If voyage then mother .Confidence = 0,934282600879669 and Support = 853

If mother then voyage .Confidence = 0,865989863872528 and Support = 853

If man, land then house .Confidence = 0,768619656562805 and Support = 774

If house, land then man .Confidence = 0,88659793138504 and Support = 774

If house, man then land .Confidence = 0,877551019191742 and Support = 774

If man, hero then house .Confidence = 0,756385087966919 and Support = 770

If house, hero then man .Confidence = 0,969773292541504 and Support = 770

If house, man then hero .Confidence = 0,873015880584717 and Support = 770

If land, age then man .Confidence = 0,892102360725403 and Support = 802

If man, age then land .Confidence = 0,938011705875397 and Support = 802

If man, land then age .Confidence = 0,796425044536591 and Support = 802

If hero, mother then man .Confidence = 0,982105255126953 and Support = 933

If man, mother then hero .Confidence = 0,982105255126953 and Support = 933

If man, hero then mother .Confidence = 0,916502952575684 and Support = 933

If hero, voyage then man .Confidence = 0,985176742076874 and Support = 864

If man, voyage then hero .Confidence = 0,97516930103302 and Support = 864

If man, hero then voyage .Confidence = 0,848722994327545 and Support = 864

If mother, voyage then man .Confidence = 0,991793692111969 and Support = 846

If man, voyage then mother .Confidence = 0,954853296279907 and Support = 846

If man, mother then voyage .Confidence = 0,890526294708252 and Support = 846

If mother, voyage then hero .Confidence = 0,991793692111969 and Support = 846

If hero, voyage then mother .Confidence = 0,964652240276337 and Support = 846

If hero, mother then voyage .Confidence = 0,890526294708252 and Support = 846

If hero, mother, voyage then man .Confidence = 0,995271861553192 and Support = 842

If man, mother, voyage then hero .Confidence = 0,995271861553192 and Support = 842

If man, hero, voyage then mother .Confidence = 0,974537014961243 and Support = 842

If man, hero, mother then voyage .Confidence = 0,902465164661407 and Support = 842

Προτάσεις _2 _Ιλιάδας _60/60

If hero then man .Confidence = 0,942669153213501 and Support = 1003

If man then hero .Confidence = 0,897940933704376 and Support = 1003

If land then man .Confidence = 0,844368040561676 and Support = 982

If man then land .Confidence = 0,879140555858612 and Support = 982

If mother then man .Confidence = 0,961661338806152 and Support = 903

If man then mother .Confidence = 0,808415412902832 and Support = 903

If land then hero .Confidence = 0,806534826755524 and Support = 938

If hero then land .Confidence = 0,881578922271729 and Support = 938

If mother then hero .Confidence = 0,966986179351807 and Support = 908

If hero then mother .Confidence = 0,853383481502533 and Support = 908

Προτάσεις_ 2_ Οδύσσεια_60/60

If man then house .Confidence = 0,759036123752594 and Support = 882

If house then man .Confidence = 0,86982250213623 and Support = 882

If land then house .Confidence = 0,770520746707916 and Support = 873

If house then land .Confidence = 0,860946774482727 and Support = 873

If land then man .Confidence = 0,888790845870972 and Support = 1007

If man then land .Confidence = 0,866609275341034 and Support = 1007

If hero then man .Confidence = 0,96860134601593 and Support = 1018

If man then hero .Confidence = 0,876075744628906 and Support = 1018

If mother then man .Confidence = 0,964466989040375 and Support = 950

If man then mother .Confidence = 0,817555963993073 and Support = 950

If voyage then man .Confidence = 0,97042715549469 and Support = 886

If man then voyage .Confidence = 0,762478470802307 and Support = 886

If age then land .Confidence = 0,927760601043701 and Support = 899

If land then age .Confidence = 0,793468654155731 and Support = 899

If mother then hero .Confidence = 0,964466989040375 and Support = 950

If hero then mother .Confidence = 0,903901040554047 and Support = 950

If voyage then hero .Confidence = 0,960569560527802 and Support = 877