ΑΠΟΔΟΤΙΚΗ ΟΡΓΑΝΩΣΗ ΚΑΙ
ΔΙΑΧΕΙΡΙΣΗ ΠΟΛΥΔΙΑΣΤΑΤΩΝ
ΑΝΤΙΚΕΙΜΕΝΩΝ ΓΙΑ ΤΗΝ ΑΝΑΚΑΛΥΨΗ
ΓΝΩΣΗΣ

(Effective Organization and Management of Multi-
Dimensional Objects for Knowledge Discovery Achievement)

Διδακτορική Διατριβή

ΑΙΚΑΤΕΡΙΝΗ Ν. ΚΡΟΤΟΠΟΥΛΟΥ

Τριμελής Συμβουλευτική Επιτροπή:

Καθηγητής	Αθανάσιος	Τσακαλίδης	(επιβλέπων)
Καθηγητής	Πάυλος	Σπυράκης	(συνεπιβλέπων)
Καθηγητής	Θεόδωρος	Παπαθεοδώρου	(συνεπιβλέπων)

Σχολή: Πολυτεχνική Σχολή, Πανεπιστήμιο Πατρών
Τμήμα: Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής
Τομέας: Λογικού των Υπολογιστών

Σε Σας Τους Δύο

Στον παπα-Γιάννη Συρεγγέλα

Στη μαμά μου

Ευκαιρία μπορεί να είναι το ψευδώνυμο του Θεού

Anatole France

Στον πατέρα μου οφείλω το ότι ζω,
αλλά στον δάσκαλό μου ότι έμαθα να ζω καλά

Αλέξανδρος ο Μέγας

Πρόλογος – Ευχαριστίες

Πριν αρχίσω να αναφέρομαι στην έρευνα και τα αποτελέσματα αυτής της διατριβής, θα ήθελα να ευχαριστήσω πρώτα απ' όλα το Θεό που μου Έδωσε τη δυνατότητα τη δυνατότητα για κάτι που δεν είχα βάλει ποτέ στο μυαλό μου, αλλά ήρθε τυχαία στο δρόμο μου. Τόσο τυχαία που δεν έφυγε ποτέ. Συγκεκριμένα πολλές φορές ήρθε αντιμέτωπο με την εγκατάλειψη αλλά τελικά το επεδίωξα γιατί είχε καταστεί ανάγκη μου και 'ανάγκα και θεοί πείθονται'. Ή και το αντίθετο..

Ευχαριστώ πολύ τη μητέρα μου που με τον τρόπο της μου εμφύσησε την αγάπη για τα γράμματα, την τέχνη και ότι πάει τον άνθρωπο ένα βήμα πιο πέρα, τον πατέρα μου που μου κληροδότησε την αγάπη για τα μαθηματικά και με έμαθε να πολεμάω και τα αδέρφια μου που χωρίς ίσως να το ξέρουν με στηρίζουν πάντα με την αγάπη, τη ζωντάνια και την εφευρετικότητά τους. Την αδερφή μου μάλιστα πρέπει να την ευχαριστήσω διπλά γιατί μου πρόσφερε και 'τεχνική' βοήθεια στη χρήση της Αγγλικής γλώσσας.

Δεν ξεχνώ ποτέ τις δυσκολίες που πέρασα και αποτέλεσαν έναν από τους λόγους της αργοπορίας αυτής της διατριβής. Σ' αυτές τις δυσκολίες πάντα έπαιρνα απλόχερη ενθάρρυνση και ευλογία από τον πνευματικό μου, πατέρα Ιωάννη Συρεγγέλα, τον οποίο και ευχαριστώ.

Με όλα αυτά σαν βάση, κρυφή ή φανερή, πριν από πολλά χρόνια συνάντησα μία καλή μου τύχη, την οποία και ευχαριστώ. Και εμφανίστηκε αρχικά στο πρόσωπο του καθηγητή μου κ. Παύλου Σπυράκη και της συνεργάτιδας στην αρχή και στη συνέχεια φίλης από καρδιάς, Δήμητρας Τέρπου. Τον κ. Σπυράκη τον ευχαριστώ για τους πνευματικούς δρόμους που μου άνοιξε. Τα εμπόδια ήταν πολλά και μερικές φορές σκληρά για την ηλικία και την εποχή εκείνη. Όμως πάντα θα ευγνωμονώ εκείνες τις ερευνητικές συναντήσεις από τις οποίες έφευγα 'φωτισμένη' ήταν με τον κ. Σπυράκη στην αρχή και μετά και με τον καθηγητή μου και επιβλέποντα κ. Θανάση Τσακαλίδη, τους οποίους ευχαριστώ για το διανοητικό, ανθρώπινο και οικονομικό (μέσω προσφοράς εργασίας) οξυγόνο που απλόχερα μου προσέφεραν. Δε θέλω να αξιολογήσω τι είμαι σήμερα. Αυτό που μπορώ σίγουρα να πω είναι ότι χωρίς αυτούς τους δύο ανθρώπους θα ήμουν κάτι λιγότερο σε ότι αφορά τον τρόπο διαχείρισης της σκέψης μου.

Για τη Δήμητρα τι να πω. Σε σχέση με τη διατριβή, νομίζω πως αν δεν ήταν εκείνη δε θα είχα το σθένος ούτε να αρχίσω ούτε να συνεχίσω εκείνα τα δύσκολα χρόνια. Ήμασταν πολύ δεμένο και αυθεντικό μαθητικό δίδυμο. Δεν είναι τυχαίο πως πολλοί μπέρδευαν τα ονόματά μας και είχαμε αποκτήσει ένα κοινό όνομα, 'τα κορίτσια', όπως μπορεί να θυμούνται φίλοι και συνεργάτες από εκείνη την εποχή. Το σπουδαίο είναι πως ακόμη και τώρα που τα κορίτσια έχουν μεγαλώσει 'λίγο', αυτό που έχω ανάγκη εδώ, είναι να αντιστρέψω το πνεύμα της ευχαριστίας. Δεν είναι τόσο ότι ευχαριστώ τη Δήμητρα για τη συμβολή και τη συνεργασία της στη διδακτορική μου πορεία, αλλά ότι ευχαριστώ το διδακτορικό που έγινε αφορμή να αποκτήσω μία τέτοια φίλη.

Στην πορεία των χρόνων που ζούσα ακόμα στην Πάτρα και δραστηριοποιούμουν στο Πανεπιστήμιο, απέκτησα φίλους τους οποίους θέλω να ευχαριστήσω για την επιστημονική και ανθρώπινη παρουσία τους στη ζωή μου. Ευχαριστώ το Δημήτρη το Φωτάκη και τον Σπύρο τον Κοντογιάννη όχι τόσο για κείνα που μοιραστήκαμε αλλά για τη σπάνια καθαρότητα και αγνότητα της επικοινωνίας που μπορούσαν να προσφέρουν. Επίσης ευχαριστώ την Τζίνα την Παναγοπούλου και το Σπύρο τον Συρμακέση για τη συνεργασία, την παρέα και τη φιλία τους, την καλή γειτονία στο Πανεπιστήμιο, αλλά – ας μου επιτρέψουν, είναι προσωπική μου προτεραιότητα - κυρίως για τα γέλια και τα τραπέζια που απολαμβάναμε μαζί.

Όλα αυτά αναφέρονται στην εποχή αμέσως μετά το πτυχίο, μέχρι το 1997. Από τότε και μετά, έκανα μία αναγκαστική ίσως στροφή, στην οποία συνάντησα νέες καταστάσεις και ανθρώπους που με επηρέασαν. Κάποια χρόνια τα έζησα αρχικά στην πολυαγαπημένη μου Ζάκυνθο – στην οποία ζω και τώρα – η οποία αποτέλεσε μεγάλη πηγή εσωτερικής ευφορίας και έμπνευσης στη ζωή μου. Την ευχαριστώ και ευχαριστώ και την ωραία σύμπτωση της γειτνιάσής της με την Πάτρα, η οποία στάθηκε πολύ βολική τα τελευταία χρόνια που μου δόθηκε η ευκαιρία να συνεχίσω το διδακτορικό μου. Εκεί δέθηκα και με τον Αγαπημένο μου Άγιο, τον Οποίο ευχαριστώ για όλα.

Την ενεργή μου επιστροφή στην έρευνα την οφείλω στην αργοπορημένη εκπαιδευτική άδεια που έλαβα από τη δευτεροβάθμια εκπαίδευση. Θέλω να εκφράσω την ευγνωμοσύνη μου γενικά και ειδικά, για αυτή τη δυνατότητα και την άνεση που μου δόθηκε. Αυτή η επιστροφή αρχικά ήταν ιδιαίτερα δύσκολη και χρειάστηκα κάποιες επιπλέον κατευθύνσεις. Μου της έδωσαν πολύ απλόχερα και γενναιόδωρα οι επίκουροι καθηγητές Χρήστος Μακρής, Σπύρος Σιούτας και Γιάννης Τζήμας, ο λέκτορας Δημήτρης Φωτάκης, ο συμφοιτητής μου αναπληρωτής καθηγητής Βασίλης Μεγαλοοικονόμου, ο αναπληρωτής καθηγητής Σπύρος Συρμακέσης καθώς και διδάκτορας Δημήτρης Σοφοτάσιος. Πολλές φορές μάλιστα μου έδιναν ιδέες με δική τους πρωτοβουλία και ενδιαφέρον. Το σπουδαίο είναι πως δεν είχαν καμία τέτοια θεσμική υποχρέωση. Τους ευχαριστώ πολύ γι' αυτό και το ξέρουν.

Σ' αυτό το τελευταίο στάδιο του διδακτορικού μου, σημαντική στάθηκε η συνδρομή και η συμπαράσταση του Τάκη. Αν και στην αρχή κάποια πράγματα τα θεωρούσε υπερβολικά, στην πορεία όχι μόνο έδειξε κατανόηση στις ιδιαιτερότητες και στις δυσκολίες που προκαλούσε αυτή η ενασχόληση στην καθημερινή μας ζωή, αλλά μπορώ να πω πως συνέπασχε μαζί μου στις αγωνίες και στις ανησυχίες μου ενώ παράλληλα πολύτιμη ήταν και η άμεση τεχνική βοήθεια που μου προσέφερε σε ό,τι προέκυπτε με το υπολογιστικό σύστημα που χρησιμοποιούσα. Όσο κι αν κάποια πράγματα είναι αυτονόητα μεταξύ των ανθρώπων που ενώνουν τις ζωές τους, θα κλείσω τις ευχαριστίες μου με ένα ευχαριστώ στον Τάκη.

Ζάκυνθος 8/8/2010

Κατερίνα Κροτοπούλου - Καμπίτση

ΠΕΡΙΕΧΟΜΕΝΑ

Πρόλογος.....	1
Πίνακας Περιεχομένων.....	3
Κατάλογος Σχημάτων.....	6
Κατάλογος Πινάκων.....	8
Περίληψη	9
Δημοσιεύσεις	22
Αναφορές.....	24

**Μέρος Α': Οργάνωση και Διαχείριση Δεδομένων και Γνώσης – Βάσεις
Δεδομένων – Γνώσης**

**Κεφάλαιο 1: Ορισμός Νέων Μοντέλων και Μεθόδων Ταξινόμησης
Τριδιάστατων Δεδομένων Πολύπλοκης Δομής και
Ταξινόμησης Γνώσης (BDB, 3D-IFO και MITOS)**

1.1. Το πρόβλημα: Χαρτογράφηση του Ανθρώπινου Εγκεφάλου.....	30
1.1.1. Αρχικός Σκοπός	34
1.2. Brain Data Base	35
1.3. 3D-IFO Μοντέλο	37
1.3.1. IFO Model.....	37
1.3.1.1. Αντικείμενα IFO (IFO Objects).....	37
1.3.2. 3D-IFO Τοπολογικές Επεκτάσεις.....	39
1.3.3. 3D-IFO Επέκταση Μηχανισμών Κατασκευής (Constructor), Εισαγωγή Πεδίων (attributes) και Σχετικών Τελεστών	39
1.3.4. Παράδειγμα χρήσης	42
1.4. Αντικειμενοστραφές Μοντέλο MITOS (Model for the Intelligent Support of Three-Dimensional Objects – Μοντέλο για την Ευφυή Υποστήριξη Τριδιάστατων Αντικειμένων)	42
1.4.1. Εισαγωγή Νέου Χαρακτηριστικού στην Έννοια του Αντικειμένου.....	43
1.4.1.1. Εισαγωγή του Χαρακτηριστικού στην Τυπική Έκφραση της Κλάσης	44
1.4.2. Εισαγωγή της Διάταξης και της Έκδοσης στον Ορισμό της Κλάσης	45
1.4.2.1. Ορισμός Διάταξης και Έκδοσης	45
1.4.2.2. Ορισμός Ενσωματωμένων (Built in) Μεθόδων Διαχείρισης Διάταξης και Έκδοσης	46
1.4.3. Νέος ορισμός των Κανόνων	47
1.5. Ανακεφαλαίωση – Συνεισφορά	48
1.6. Δημοσιεύσεις του Κεφαλαίου.....	50
1.7. Αναφορές σε Δημοσιεύσεις μας	51
1.8. Αναφορές του Κεφαλαίου	51

Μέρος Β': Ανεύρεση Ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα (Similarity Search among Multidimensional Objects)

Κεφάλαιο 2: Δύο Μέθοδοι για Ακριβέστερη Τμηματική Αναζήτηση (Lui k-n-match και INTESIS)

2.1. Βασικές Έννοιες.....	57
2.1.1. Μέθοδοι <i>k-n-match</i> και <i>frequent k-n-match</i> , Προβλήματα- Ανεπάρκειες.....	58
2.2.	
2.2.1. Μέθοδος Lui k-n-match.....	60
2.2.2. Αξιολόγηση – Συμπεράσματα	64
2.3. Μέθοδος INTESIS	
2.3.1. Τμηματική Ομοιότητα	65
2.3.2. Ολοκλήρωση Τμηματικής και Πλήρους Ομοιότητας.....	70
2.3.3. Αξιολόγηση – Συμπεράσματα	70
2.4. Δημοσιεύσεις του Κεφαλαίου.....	73
2.5. Αναφορές του Κεφαλαίου.....	73
2.7. Παράρτημα 2 ^ο Κεφαλαίου – R*-trees	75

Κεφάλαιο 3: Υβριδική Μέθοδος Αναζήτησης Ομοιοτήτων Ανάμεσα σε Αντικείμενα, Ανεξάρτητη από το Πλήθος των Διατάσεων τους (Μέθοδος DIGENIS)

3.1. Το Πρόβλημα (dimensionality curse)- Υπαρκτές Μέθοδοι (State of the Art) και Σύγκρισή τους.....	79
3.2. Μέθοδος DIGENIS	80
3.2.1. Δέντρο DIGENIS (DIGENIS-Tree).....	81
3.2.1.1. Αναζήτηση Πλησιέστερων Γειτόνων (k-NN Queries) στο Δέντρο Digenis ..	85
3.2.1.2. Ενημέρωση Δέντρου Digenis	87
3.2.2. Αρχείο DIGENIS (DIGENIS-File).....	88
3.2.2.1 . Τοποθέτηση στο Αρχείο Digenis	88
3.2.2.2 Ενημέρωση του Αρχείου Digenis	90
3.2.3.Αξιολόγηση της Μεθόδου	
3.2.3.1. Θεωρητική Αξιολόγηση.....	90
3.2.3.2. Πειραματική Αξιολόγηση	92
3.3. Ανακεφαλαίωση – Ανοικτά θέματα.....	93
3.4. Δημοσιεύσεις του Κεφαλαίου.....	95
3.5. Αναφορές του Κεφαλαίου.....	95
3.6. Παράρτημα 3 ^ο Κεφαλαίου – VA-Files	97

Κεφάλαιο 4: Μέθοδος Αναζήτησης Ομοιοτήτων Ανάμεσα σε Χαρακτηριστικά Αντικειμένων (Δέντρο Chiron)

4.1. Εισαγωγή – Πλαίσιο Αναφοράς – Ορισμός του Προβλήματος.....	100
4.2. Τυποποίηση του Προβλήματος.....	102
4.2.1. Ορισμοί.....	102
4.3. Δέντρο Chiron.....	104
4.3.1. Θεώρημα Chiron.....	106
4.3.2. Χρήση του Δέντρου Chiron.....	108
4.3.2.1. Εντοπισμός Θέσης (Positioning)	108
4.3.2.2. Εντοπισμός των Όμοια Εξελισσόμενων Χαρακτηριστικών (Ομοιότητα Chiron)	110
4.4. Πειραματική Μελέτη – Αξιολόγηση	112
4.5. Ανοικτά Θέματα – Μελλοντικές Κατευθύνσεις	114
4.6. Ανακεφαλαίωση – Συμπεράσματα	115
4.7. Δημοσιεύσεις του Κεφαλαίου.....	117
4.8. Αναφορές του Κεφαλαίου.....	117

Μέρος Γ': Επίλογος

Κεφάλαιο 5: Επίλογος / Ερωτήματα - Παρατηρήσεις – Συμπεράσματα Ολοκλήρωση

5.1. Χαρτογράφηση Εγκεφάλου. Ποια είναι η συνολική προσφορά.....	121
5.2. Τμηματική Ομοιότητα: Lui k-n-match ή INTESIS;	122
5.3. Επέκταση INTESIS βάσει άλλων προτάσεων.	123
5.4. Γιατί παρόλο που τα δέντρα Digenis και Chiron μοιάζουν, οι αλγόριθμοι εντοπισμού θέσης σ' αυτά διαφέρουν τόσο;	124

Κατάλογος Σχημάτων

Σχήμα 1.1.: Έξω Όψη Ανθρώπινου Εγκεφάλου	31
Σχήμα 1.2.: Τομή κατά μέτωπο στο ύψος του μεσοκοιλιακού τμήματος, της χοάνης και του αμυγδαλοειδούς σώματος.....	32
Σχήμα 1.3.: Τομή κατά μέτωπο του βασικού τμήματος του πρόσθιου εγκεφάλου στο ύψος του θαλάμου και του οπίσθιου άκρου του κελύφους.....	33
Σχήμα 1.4.: Εγκάρσια τομή του εγκεφαλικού στελέχους στο ύψος του πυρήνα του απαγωγού νεύρου, της άνω ελαίας και του τραπεζοειδούς σώματος.	33
Σχήμα 1.5.: Παράδειγμα Μαγνητικής Τομογραφίας Εγκεφάλου (MRI).....	34
Σχήμα 1.6: Παράδειγμα συμβολισμού ιεραρχικής διάταξης λοβών.....	36
Σχήμα 1.7.: Οι τρεις ατομικοί τύποι IFO	38
Σχήμα 1.8.: Ομαδοποιημένοι (constructed) Τύποι Αντικειμένων	38
Σχήμα 1.9: Τυπικός συμβολισμός του ordering	40
Σχήμα 1.10.: Παράδειγμα χρήσης του constructor ordering για πληροφορία του τύπου: ‘Η περιοχή Brodman X κείται βορειοδυτικά του έλικα Y’.	40
Σχήμα 1.11.: Εκφράζοντας την έννοια της χρονικής έκδοσης	41
Σχήμα 1.12.: Listing βασισμένο στην πιθανότητα.....	41
Σχήμα 1.13.: Ορισμός Σχέσης με το Περιβάλλον.....	44
Σχήμα 1.14.: Γενικό παράδειγμα διάταξης αντικειμένων κλάσης.....	45
Σχήμα 1.15.: Ο τυπικός ορισμός της ιεραρχικής διάταξης αντικειμένων κλάσης και της έκδοσης.....	46
Σχήμα 1.16. Τυπικός Ορισμός Κεφαλίδας Κανόνα.....	48
Σχήμα 2.1.: Προσανατολισμός (Π)R*-tree	68
Σχήμα 2.2.: Μέγεθος (M)R*-tree.....	69
Σχήμα 2.3.: Χρώμα (X)R*-tree.....	69
Σχήμα 2.4.: Παράδειγμα ενός R-tree για 2D δεδομένα (ορθογώνια)	75
Σχήμα 3.1.: Παράδειγμα Γράφου με $Lt=4$. Οι ακμές αντιπροσωπεύουν τις αποστάσεις ανάμεσα στις οικογένειες. Για λόγους απλότητας και πρακτικότητας, έχουν παραλειφθεί οι πολύ μεγάλες αποστάσεις (μεγαλύτερες από 3, στη συγκεκριμένη περίπτωση) αφού δεν οδηγούν σε συνδέσεις στο παρακάτω δένδρο (με $ch=3$).	83
Σχήμα 3.2.: Το δένδρο Digenis, με βάση το γράφο του σχήματος 1, με $ch=3$	83
Σχήμα 3.3.: Γράφος με αραιές περιοχές, με $ch=3$ και $Lt=3,5$ (και εδώ έχουν παραληφθεί οι αποστάσεις μεγαλύτερες του 3).....	83
Σχήμα 3.4.: Τα δένδρα Digenis, βασισμένα στο γράφο του σχήματος 3, με $ch=3$ για το πρώτο δένδρο και $ch=4$ για το δεύτερο.	84
Σχήμα 3.5. : Παράδειγμα δένδρου DIGENIS-Tree με $ch=3$ και δοθέν αντικείμενο q.	86
Σχήμα 3.6. : Αξιολόγηση σύμφωνα με το πλήθος των διαστάσεων.....	93
Σχήμα 3.7. : Κατασκευή του VA-File	97
Σχήμα 4.1.: Πέντε εμφανίσεις 6-διάστατης βάσης δεδομένων.....	103
Σχήμα 4.2.: Αντικείμενα Πορείας (Course Objects) για κάθε διάσταση των	

εμφανίσεων αντικειμένων του Σχήματος 1	103
Σχήμα 4.3.: Παράδειγμα αναπαράστασης γράφου διαφορών ανάμεσα σε ομάδες (groups).....	105
Σχήμα 4.4.: Ένα Δένδρο Chiron, βασισμένο στο Σχήμα 3.....	106
Σχήμα 4.5.: Παράδειγμα για την απόδειξη του Θεωρήματος.....	107
Σχήμα 4.6.: Απόδοση σύμφωνα με τον αριθμό των εμπλεκόμενων χαρακτηριστικών...	113
Σχήμα 4.7.: Απόδοση σύμφωνα με τον αριθμό των εμπλεκόμενων αντικειμένων	113
 Σχήμα 5.1. : Δέντρο Digenis, με $ch=3$	124
Σχήμα 5.2. : Δένδρο Chiron.....	124

Κατάλογος Πινάκων

Πίνακας 1.1.: Ορισμός Ενσωματωμένων(Built in) Μεθόδων Διαχείρισης Διάταξης / Έκδοσης	47
Πίνακας 2.1. Πίνακας Συμβολισμών	56
Πίνακας 2.2.: Παράδειγμα όπου δεν ανιχνεύεται η τμηματική ομοιότητα	58
Πίνακας 2.3.: Πίνακας των διατεταγμένων σε αύξουσα σειρά n-match διαφορές για κάθε αντικείμενο	62
Πίνακας 2.4.: Τα ενδιάμεσα αποτελέσματα (k-n-matches) της frequent k-n-match, για k=2	63
Πίνακας 2.5.: Ακρίβεια αποτελεσμάτων ανά Σύνολο Δεδομένων	64
Πίνακας 2.6.: Αντικείμενο A	66
Πίνακας 2.7.: Προκαθορισμένα υποσύνολα διαστάσεων	66
Πίνακας 2.8.: Τα Υπο-Αντικείμενα του A	66
Πίνακας 2.9.: Το Βοηθητικό Αρχείο Συσχέτισης Κωδικών	68
Πίνακας 2.10.: Πίνακας αποτελεσμάτων	69
Πίνακας 3.1.: Πίνακας Συμβολισμών	78
Πίνακας 3.2.: Σύγκριση της απόδοσης του Αρχείου Digenis με άλλες μεθόδους	93
Πίνακας 4.1.: Πίνακας Συμβολισμών	99

Αποδοτική Οργάνωση και Διαχείριση Πολυδιάστατων Αντικειμένων για την Ανακάλυψη Γνώσης

Αικατερίνη Ν. Κροτοπούλου

Επιβλέπων: Καθηγητής Αθανάσιος Κ. Τσακαλίδης

Περιοχές Έρευνας: Βάσεις Δεδομένων, Δομές Δεδομένων, Εξόρυξη Δεδομένων

Λέξεις Κλειδιά: Πολυμεσικές Βάσεις Δεδομένων
Χωρικές Βάσεις Δεδομένων
Ιατρική Πληροφορική
Πολύπλοκα Δικτυακά Αντικείμενα
Ανεύρεση Ομοιοτήτων
Πλήρης Ομοιότητα (Full Match)
Τμηματική Ομοιότητα (Partial Match)
k-n-match
'Dimensionality Curse' ('Η κατάρα που ακολουθεί την
αύξηση των διαστάσεων')
Χρονικές Πορείες Χαρακτηριστικών

Περίληψη

Ο σκοπός αυτής της διατριβής είναι η ανεύρεση μεθόδων αποδοτικής οργάνωσης και διαχείρισης πολυδιάστατων αντικειμένων (multi-dimensional objects) προκειμένου να ανακαλυφθεί χρήσιμη γνώση. Αρχική αφορμή για αυτή τη μελέτη αποτέλεσαν οι ανάγκες μιας απαιτητικής εφαρμογής με σκοπό τη *χαρτογράφηση του ανθρώπινου εγκεφάλου* προκειμένου να εντοπιστούν επιληπτικές εστίες¹. Οι απαιτήσεις *Αναπαράστασης και Διαχείρισης των Δεδομένων του Εγκεφάλου*, έφεραν στην επιφάνεια δύο κεντρικά ερευνητικά προβλήματα:

- ✓ *Τις ιδιαιτερότητες των πολύπλοκων, μη-ομοιογενών, δικτυακών μερικές φορές, τρισδιάστατων αντικειμένων (τμημάτων του εγκεφάλου – brain objects).*

---

¹ Στα πλαίσια του ευρωπαϊκού προγράμματος CEC-Program AIM, Project A2020 Magnobrain.

- ✓ Την ανάγκη για αποτελεσματική διαχείριση-χρήση γνωστών αλλά και παραγόμενων εξαρτήσεων δεδομένων και γνώσης (*data and knowledge dependencies*), η οποία μπορεί να αναβαθμίσει την απόδοση και τη δυναμική της εφαρμογής. Το μεγαλύτερο μέρος της μελέτης που αφορούσε αυτό το πρόβλημα, οδήγησε σε :
- ✓ διερεύνηση θεμάτων ανεύρεσης ομοιοτήτων (*similarity search*). Καθώς η συγκεκριμένη περιοχή διαθέτει μεγάλο εύρος εφαρμογών αλλά και ανοικτών προβλημάτων, αποτέλεσε τελικά μεγάλο μέρος της παρούσας διατριβής.

Δεδομένου ότι πολλά από τα γεωμετρικά χαρακτηριστικά των δεδομένων αλλά και από τις εξαρτήσεις γνώσης που αφορούν τον ανθρώπινο εγκέφαλο, συναντώνται – καθ'ολοκληρία ή τμηματικά – σε πλήθος σύγχρονων πολυμεσικών (*multimedia*) εφαρμογών, τα παραπάνω προβλήματα εντάσσονται στα βασικά προβλήματα της έρευνας του τομέα των Βάσεων Δεδομένων.

Επικεντρώνοντας την έρευνά στα παραπάνω προβλήματα, καταλήξαμε:

- στον ορισμό νέων ευέλικτων τύπων δεδομένων, εννοιών και μοντέλων καθώς και εργαλείων και μεθόδων ταξινόμησης δεδομένων και γνώσης (βάση δεδομένων *BDB*[Σ2,Σ4,Σ9] και μοντέλα *3D-IFO*[Σ5] και *MITOS*[Σ12,Σ13]) οι οποίες οργανώνουν πιο ευέλικτα και αποδοτικά τα δεδομένα μας, με τρόπους που όχι μόνο κάνουν την πρόσβασή τους ευκολότερη αλλά αξιοποιούν παράλληλα τις ‘κρυμμένες’ μεταξύ τους σχέσεις για την άντληση επιπλέον γνώσης.
- στον ορισμό νέων μεθόδων και δέντρων αναζήτησης, για :
  - τον αποδοτικό εντοπισμό τμηματικών ομοιοτήτων (*partial similarity*) ανάμεσα σε πολυδιάστατα αντικείμενα (*Lui k-n-match* [Σ18] και *INTESIS* [Σ16])
  - την εξάλειψη της μεγάλης πτώσης της απόδοσης των δέντρων με την αύξηση των διαστάσεων των αντικειμένων [Σ15](‘*dimensionality curse*’)(*δομή Digenis*).
  - την ανεύρεση χαρακτηριστικών/διαστάσεων με παρόμοια εξέλιξη στην πορεία του χρόνου – για πολυδιάστατα κυρίως αντικείμενα – με σκοπό τη μελέτη πιθανής αλληλεπίδρασής τους [Σ17].

Γενικά, η παρούσα μελέτη αποτελείται από δύο βασικά μέρη, τα οποία αναφέρονται σε δύο περιοχές με μεγάλη αλληλεπίδραση:

- Τη Μοντελοποίηση σε Πολυμεσικές Βάσεις Δεδομένων
- Την Αναζήτηση Ομοιοτήτων ανάμεσα σε Πολυδιάστατα Αντικείμενα

Στο πρώτο κεφάλαιο αρχικά παρουσιάζεται το πρόβλημα της χαρτογράφησης του ανθρώπινου εγκεφάλου [A8, A14] για τον εντοπισμό επιληπτικών εστιών, απ’όπου εγείρονται τα πρώτα προβλήματα αναπαράστασης και οργάνωσης τριδιάστατων αντικειμένων πολύπλοκης δομής και λειτουργικών σχέσεων και εξαρτήσεων μεταξύ τους [Σ1,Σ10,Σ11,Π1]. Σε μια πρώτη προσέγγιση προτείνεται το λογικό μοντέλο *BDB* (*Brain Data Base*) [Σ2,Σ3,Σ4,Σ9] όπου εισάγονται νέοι τύποι οντοτήτων [A1, A2, A3, A7, A12, A16, A17, A18]. Εδώ, ιδιαίτερο ενδιαφέρον παρουσιάζει η προσθήκη της ιεραρχικής διάταξης στο *Σχεσιακό Μοντέλο*, προκειμένου οι περιοχές του εγκεφάλου να οργανωθούν με βάση την πιθανότητα εμφάνισης επιληπτικής εστίας έτσι ώστε να βελτιώνονται στατιστικά οι χρόνοι ανάκτησής τους. Στη συνέχεια, η μελέτη επεκτείνεται σε άλλα – επόμενης γενιάς - είδη μοντέλων [A5, A19, A20]. Πιο συγκεκριμένα, οι ανάγκες της εφαρμογής μελετώνται με βάση ένα Σημαντικό (*semantic model*) - το μοντέλο *IFO* [A6] - και ένα

Αντικειμενοστραφές Μοντέλο (object oriented model), με αποτέλεσμα τη δημιουργία των μοντέλων 3D-IFO [Σ5] και MITOS[Σ12,Σ13] αντίστοιχα. Στο 3D-IFO εισήχθησαν νέοι τύποι δεδομένων προκειμένου να υποστηριχθούν αποδοτικά τα ιδιαίτερα δεδομένα μας καθώς και νέοι τελεστές για την καλύτερη διαχείριση των σύνθετων δεδομένων. Επιπλέον, εισήχθη ένας νέος *constructor* και ένα κατάλληλο πεδίο για την υποστήριξή του, προκειμένου να υποστηριχτεί η αναπαράσταση της διάταξης των μερών του εγκεφάλου με βάση κάποιο κριτήριο έτσι ώστε να διευκολυνθεί η μελλοντική απλή και συνδυαστική ανάκτηση πληροφορίας. Τέλος το αντικειμενοστραφές μοντέλο MITOS, εισάγει πάλι ένα νέο μοντέλο δεδομένων (MITOS Data Model - MDM) το οποίο συνεργάζεται με μία νέα γλώσσα ερωτημάτων (MITOS Query Language - MQL). Το μοντέλο MITOS εισάγει διάφορες καινοτομίες οι οποίες εξυπηρετούν μια περισσότερο εκφραστική και έξυπνη αναπαράσταση και διαχείριση πολυδιάστατων δεδομένων και γνώσης. Η μία από αυτές τις καινοτομίες είναι ο ορισμός ενός ακόμη βασικού χαρακτηριστικού των αντικειμένων (object characteristic), της σχέσης τους με το περιβάλλον, απεγκλωβίζοντάς την από την κατάσταση ή τη συμπεριφορά [A20], όπου αποδυναμώνεται σαν έννοια. Η δεύτερη καινοτομία του MITOS η οποία αφορά την MQL σχετίζεται με την εισαγωγή 'κλειδιού' στους κανόνες (rules). Η διερεύνηση αυτής της δυνατότητας – η ιδέα προέρχεται από το χώρο των Βάσεων Δεδομένων [A10] – οδηγεί πράγματι σε ένα είδος κλειδιού, κατά την έννοια που θα μπορούσε να έχει στις Βάσεις Γνώσης και η οποία δεν μπορεί να είναι ακριβώς ίδια με την αντίστοιχη των Βάσεων Δεδομένων, λόγω των ειδοποιών διαφορών των δύο Βάσεων.

Στο δεύτερο κεφάλαιο μελετάται η αναζήτηση ενός ελάχιστα διερευνημένου είδους ομοιότητας ανάμεσα σε πολυδιάστατα κυρίως αντικείμενα, της *τμηματικής ομοιότητας* (*partial similarity*). Η τμηματική ομοιότητα σε αντίθεση με τον ιδιαίτερα διερευνημένο τύπο της πλήρους ομοιότητας (*full similarity*), αναφέρεται σε πραγματικές ομοιότητες οι οποίες δεν είναι πλήρεις. Κι αυτό συμβαίνει γιατί ένα πολύ συνηθισμένο σενάριο κατά τη διερεύνηση ομοιοτήτων είναι το ακόλουθο: Συνήθως η ανεύρεση πλήρους ομοιότητας βασίζεται σε υπολογισμό αποστάσεων, όπως η Ευκλείδεια απόσταση, οι οποίες είναι συνάρτηση όλων των διαστάσεων των εμπλεκόμενων αντικειμένων. Όταν λοιπόν υπάρχουν διαστάσεις με μεγάλες διαφορές, ακόμη κι αν είναι λίγες, αυξάνουν αρκετά την υπολογιζόμενη απόσταση έτσι ώστε οι αποστάσεις τέτοιων αντικειμένων που στην πραγματικότητα μπορεί να είναι όμοια, να καταλήγουν να έχουν μεγάλες τιμές και συνεπώς να μην ανιχνεύεται η ομοιότητά τους (π.χ. όμοια αντικείμενα με πολύ διαφορετικό χρώμα). Από την άλλη πλευρά, για αντικείμενα τα οποία διαφέρουν λίγο σε κάθε διάσταση (π.χ. λίγο διαφορετικό χρώμα, σχήμα, προσανατολισμό κ.λ.π.) και καταλήγουν να είναι στην πραγματικότητα συνολικά πολύ διαφορετικά, η υπολογιζόμενη μεταξύ τους απόσταση έχει μικρή τιμή, οπότε ανιχνεύονται σαν όμοια, χωρίς να είναι.

Οι περισσότερες εργασίες οι οποίες έχουν μελετήσει την τμηματική ομοιότητα, έχουν εστιάσει σε γεωμετρικά δεδομένα [A62, A63, A64]. Η εργασία που επεκτείνεται σε πολυδιάστατα αντικείμενα γενικά είναι η [A39] και έχει οδηγήσει σε αξιόλογα αποτελέσματα στο θέμα της τμηματικής ομοιότητας. Εισάγει τις αποδοτικές μεθόδους k-n-match και frequent k-n-match, οι οποίες επιστρέφουν k αντικείμενα, όμοια με τα δοθέντα όχι σε όλες αλλά σε n διαστάσεις, αποφεύγοντας έτσι εκείνες τις λίγες διαστάσεις με τις μεγάλες διαφορές, οι οποίες οδηγούν σε παραπλανητικά αποτελέσματα. Παρόλ' αυτά αυτές οι μέθοδοι κρύβουν κάποιες αδυναμίες οι οποίες τελικά οδηγούν είτε σε ανεύρεση πλήρους ομοιότητας (όταν τελικά ληφθούν υπ' όψιν όλα τα n), είτε σε μία κατά περίπτωση μόνο (και σχεδόν τυχαία) ανίχνευση τμηματικής ομοιότητας (με τα κατάλληλα n's τα οποία δεν πρέπει να είναι ούτε πολύ μεγάλα ούτε πολύ μικρά, αλλά δεν ορίζονται από κάποιο τύπο ή μέθοδο). Βασιζόμενοι σ' αυτές τις μεθόδους, προτείνουμε δύο νέες τεχνικές οι οποίες όπως

αποδεικνύεται μπορούν να εντοπίσουν πραγματικές τμηματικές ομοιότητες. Η πρώτη, η *Lui k-n-match* [Σ18], επιτυγχάνει τον κατά προσέγγιση εντοπισμό των κατάλληλων n 's για τα k - n -matches, με τη βοήθεια της αλληλεπίδρασης με το χρήστη και του ελέγχου των αποδεκτών προτάσεων των k - n -matches. Πιο συγκεκριμένα, μέσω της μεθόδου k - n -match, προτείνεται για κάθε n ένα σύνολο αντικειμένων πιθανά όμοιων με το δεδομένο αντικείμενο του ερωτήματος (query object)². Ο χρήστης φιλτράρει αυτό το σύνολο, επιλέγοντας εκείνα τα αντικείμενα που θεωρεί πραγματικά όμοια με το δεδομένο. Αυτή η διαδικασία συνεχίζεται μέχρι αφού το n γίνει μεγαλύτερο από το ήμισυ των διαστάσεων των αντικειμένων, υπάρξει σύνολο προτεινόμενων αντικειμένων από το οποίο ο χρήστης δεν επιλέγει κανένα ως όμοιο³. Μ'αυτόν τον τρόπο επιτυγχάνεται μεγαλύτερη εγκυρότητα των αποτελεσμάτων (λόγω της εμπλοκής του χρήστη) με περιορισμένο ταυτόχρονα αριθμό εκτελούμενων k - n -matches.

Η δεύτερη μέθοδος (*INTESIS*) [Σ16] βασίζεται στην εξής παρατήρηση: στην ουσία όταν δύο αντικείμενα μοιάζουν αυτό συνήθως σημαίνει ότι μοιάζουν στα περισσότερα χαρακτηριστικά τους, καθένα από τα οποία αναπαριστάται και αντιπροσωπεύεται από ένα σύνολο (μικρό συνήθως) διαστάσεων-πεδίων του αντικειμένου. Εάν λοιπόν οριστεί από τους ειδικούς κάθε εφαρμογής αυτή η αντιστοιχία χαρακτηριστικών και διαστάσεων - δημιουργώντας υποσύνολα διαστάσεων - τότε μπορούν να συμβούν διαδοχικά τα παρακάτω:

α) Να γίνει έλεγχος πλήρους ομοιότητας σε κάθε τέτοιο υποσύνολο διαστάσεων

β) Να οργανωθούν αυτά τα υποσύνολα σε ισάριθμα ιεραρχικά δέντρα για την εύκολη και αποδοτική διαχείρισή τους. Η επιπλέον απλούστευση αυτής της επιλογής έγκειται στο ότι δεδομένου ότι τα εν λόγω υποσύνολα διαστάσεων θα είναι μικρά, είναι πολύ εύκολη η επιλογή δέντρου γι' αυτά, αφού σχεδόν όλα τα ιεραρχικά δέντρα έχουν μεγάλη απόδοση όταν πρόκειται για μικρό αριθμό διαστάσεων. Συνεπώς ο αναλυτής της κάθε εφαρμογής μπορεί να χρησιμοποιήσει όποιο τέτοιο δέντρο κρίνει εκείνος σαν καλύτερο (Το R^* -tree είναι η δική μας πρόταση).

Τελικά, για να ολοκληρωθεί η διαδικασία πρέπει να έχει οριστεί ένας ελάχιστος αριθμός απαιτούμενων όμοιων χαρακτηριστικών προκειμένου να θεωρηθούν δύο αντικείμενα όμοια.

Για την αξιολόγηση αυτής της μεθόδου, πρέπει αρχικά να σημειωθεί ότι αναφέρεται σε συνολικό αριθμό διαστάσεων μικρότερο του 100 και συνεπώς σε σχετικά μικρό αριθμό δέντρων. Όπως είναι φανερό, σε μονο-επεξεργαστικό σύστημα οι τελικοί χρόνοι απόκρισης είναι το άθροισμα των χρόνων κάθε δέντρου. Λαμβάνοντας υπ' όψιν το ότι τα δέντρα λόγω του μικρού αριθμού διαστάσεων που αντιστοιχούν στο καθένα έχουν πολύ καλές αποδόσεις, βγαίνει εύκολα το συμπέρασμα ότι ο εκάστοτε τελικός χρόνος απόκρισης της μεθόδου - όντας ένα μικρό πολλαπλάσιο των πολύ μικρών χρόνων προσπέλασης των δέντρων - είναι αρκετά χαμηλός. Με δεδομένο ότι η χρήση κάθε δέντρου δεν προϋποθέτει την χρήση κάποιου άλλου πριν ή μετά, οι αναζητήσεις σε κάθε δέντρο μπορούν να γίνονται παράλληλα. Συνεπώς σε πολυεπεξεργαστικό σύστημα, ο συνολικός χρόνος απόδοσης μπορεί να μειωθεί σημαντικά, φτάνοντας μέχρι και το χρόνο που απαιτείται μόνο για αναζήτηση σε ένα δέντρο (όταν υπάρχουν τόσοι επεξεργαστές όσα και δέντρα). Φυσικά, εάν λάβει κανείς υπ' όψιν το ότι η τμηματική ομοιότητα αποτελεί ένα ιδιαίτερα απαιτητικό είδος τότε όχι μόνο οι χρόνοι απόκρισης σε πολυεπεξεργαστικό σύστημα αλλά και εκείνοι του συστήματος ενός επεξεργαστή, αποτελούν ικανοποιητικές αποδόσεις.

² Το σύνολο δεν περιλαμβάνει αντικείμενα που έχουν επιλεγεί για προηγούμενο (μικρότερο) n .

³ Αυτό σημαίνει ότι τα τμηματικά όμοια αντικείμενα έχουν ήδη επιλεγεί και από εδώ και πέρα τα επόμενα προτεινόμενα αντικείμενα θα είναι μάλλον πλήρως όμοια με το δοθέν.

Το τρίτο κεφάλαιο μελετά τη δυνατότητα δημιουργίας μιας νέας δομής η οποία δε θα 'υποφέρει' από τη μεγάλη πτώση της απόδοσης των δέντρων με την αύξηση των διαστάσεων των αντικειμένων ('dimensionality curse') ενώ ταυτόχρονα θα εξασφαλίζει καλή απόδοση και σε μικρό αριθμό διαστάσεων. Οι μέχρι τώρα μελέτες έχουν καταλήξει στο εξής συμπέρασμα: Τα γνωστά διαδεδομένα δέντρα αναζήτησης (είτε πρόκειται για δέντρα οργανωμένα βάση κατανομής χώρου (space partitioning) [A37] είτε για δέντρα βάση κατανομής δεδομένων (data partitioning) [A22, A33, A38, A44, A45, A49, A50, A52, A56, A58, A60, A61]) αποδίδουν πολύ καλύτερα σε μικρό αριθμό διαστάσεων ενώ όσο αυτός ο αριθμός αυξάνει - ειδικά από 10 και πάνω - η απόδοση χειροτερεύει δραματικά [A42]. Το VA-File (σχήμα προσέγγισης διανύσματος) από την άλλη πλευρά - το οποίο είναι ένας απλός πίνακας-αρχείο γεωμετρικών προσεγγίσεων των αντικειμένων - με την αύξηση των διαστάσεων αποδίδει καλύτερα στην αναζήτηση ομοιοτήτων αλλά παρουσιάζει χαμηλή απόδοση σε μικρό αριθμό διαστάσεων [A42].

Προκειμένου να ξεπεραστεί αυτή η καθοριστική εξάρτηση της απόδοσης από το πλήθος των διαστάσεων των προς διαχείριση αντικειμένων, προτείνουμε τη νέα υβριδική δομή *Digenis* [Σ15], η οποία παντρεύει τη λογική των δέντρων αναζήτησης με κείνη των VA αρχείων. Πιο συγκεκριμένα, ορίζεται και χρησιμοποιείται ένα στατικό⁴ παραμετροποιημένο δέντρο (δέντρο *Digenis*) σε εννοιολογικό επίπεδο ενώ σε φυσικό επίπεδο χρησιμοποιείται το αρχείο *Digenis* το οποίο κατασκευάζεται με βάση το δέντρο. Με αυτή τη συσχέτιση επιτυγχάνεται αναζήτηση σε μικρό μόνο μέρος του αρχείου κατά τη διαδικασία ανεύρεσης ομοιοτήτων ανάμεσα σε αντικείμενα πολλών αλλά και λίγων διαστάσεων, γεγονός που δίνει γενικότητα και ευελιξία στη μέθοδο.

Πιο συγκεκριμένα, για το σχηματισμό του δέντρου, αρχικά ορίζονται οι οικογένειες αντικειμένων, οι οποίες αποτελούνται από αντικείμενα με μικρή απόσταση (βάση ενός προκαθορισμένου από τον εκάστοτε αναλυτή ορίου fl) και αντιπροσωπεύονται από το 'μέσο' αντικείμενο της οικογένειας (εάν δεν υπάρχει δημιουργείται για αυτό το ρόλο και μόνο). Κάθε κόμβος του δέντρου αντιπροσωπεύει-φιλοξενεί μία τέτοια οικογένεια. Το είδος των αποστάσεων που χρησιμοποιείται είναι η πλέον διαδεδομένη απόσταση, η Ευκλείδεια απόσταση, για την οποία ισχύει και η τριγωνική ανισότητα στην οποία θα βασιστεί μεγάλο μέρος της μεθόδου. Επίσης ένα δεύτερο όριο απόστασης (Lt) ορίζεται - από τον αναλυτή πάλι - σαν όριο με βάση το οποίο δύο αντικείμενα μπορούν να θεωρηθούν όμοια. Το δέντρο *Digenis* τελικά χτίζεται έχοντας ρίζα την πιο 'κεντρική' οικογένεια της περιοχής των αντικειμένων και κόμβους-παιδιά της τις ch πιο γειτονικές της οικογένειες, κάθε μία από αυτές έχει παιδιά της τις ch πιο γειτονικές της οικογένειες κ.ο.κ. Η δεδομένη ισχύ της τριγωνικής ανισότητας ανάμεσα στις Ευκλείδειες αποστάσεις των αντικειμένων-οικογενειών, αποδεικνύεται ένα χρήσιμο θεώρημα βάση του οποίου καθιστάται εφικτή η ασφαλής εξαίρεση μεγάλου μέρους του δέντρου⁵ από τους ελέγχους ομοιότητας, κατευθύνοντας τον τελικό έλεγχο σε μία μικρή περιοχή του. Αυτή η ανάλυση της

⁴ Η στατικότητα σχετικά με τον αριθμό των παιδιών (ch) σε κάθε επίπεδο είναι απαραίτητη λόγω του μετασχηματισμού του δέντρου σε αρχείο όπου επιδιώκεται γρήγορος αποκλεισμός των περισσότερων ανόμοιων με το δεδομένο αντικείμενο περιοχών.

⁵ Σύμφωνα με την τριγωνική ανισότητα, εάν α, β, γ είναι οι πλευρές ενός τριγώνου, ισχύει γενικά: $\alpha - \beta < \gamma < \alpha + \beta$. Συνεπώς, θεωρώντας ότι το δεδομένο (query) σημείο q , η ρίζα r του δέντρου και ένας κόμβος παιδί της ρίζας (b) του δένδρου μας ορίζουν ένα τρίγωνο, η απόσταση του q από το b (d_{qb}) είναι σίγουρα μεγαλύτερη από την απόλυτη τιμή της διαφοράς των αποστάσεων του q (d_q) και του b (d_b) από τη ρίζα ($d_{qb} > |d_q - d_b|$). Με δεδομένες αυτές τις δύο αποστάσεις και το Lt της εφαρμογής, εύκολα εκτιμάται εάν το d_{qb} είναι μεγαλύτερο του Lt χωρίς να χρειάζεται απ'ευθείας υπολογισμός απόστασης. Πιο συγκεκριμένα, εάν $d_{qb} \geq |d_q - d_b|$ και $|d_q - d_b| \geq Lt$ τότε (μεταβατική ιδιότητα) $d_{qb} \geq Lt$.

Με βάση το παραπάνω, μπορούμε να κινηθούμε ανάλογα και για τα επόμενα επίπεδα του δέντρου.

αναζήτησης μέσα στο δέντρο είναι πολύ χρήσιμη σε ό,τι αφορά τη χρήση του αρχείου Digenis, όπου εκεί πραγματοποιείται η πραγματική αναζήτηση (φυσικό επίπεδο).

Το αντίστοιχο αρχείο Digenis στο φυσικό επίπεδο σχηματίζεται εάν αντιστοιχίσουμε σε κάθε του εγγραφή έναν κόμβο του δέντρου, ξεκινώντας από τη ρίζα του δέντρου και περνώντας από κάθε επίπεδο, από αριστερά προς τα δεξιά. Με αυτή την αντιστοίχιση, μπορούν πολύ εύκολα να χρησιμοποιηθούν οι τεκμηριωμένες τεχνικές⁶ εύκολου, ασφαλούς και γρήγορου αποκλεισμού περιοχών.

Ο απολογισμός της μεθόδου (θεωρητικά αλλά και πειραματικά) περιλαμβάνει θετικές και αρνητικές όψεις.

Θετικές όψεις:

- Το αρχείο έχει πολύ καλή απόδοση όταν διαχειριζόμαστε αντικείμενα πολλών διαστάσεων. Αυτό ήταν αναμενόμενο αφού το αρχείο λειτουργήσε σαν ένα είδος VA αρχείου, όπου το ζητούμενο ήταν η δημιουργία συμπαγών γεωμετρικών προσεγγίσεων. Κι αυτό γιατί και η χρήση των οικογενειών επέφερε μία πρώτη ‘συμπίεση’ των δεδομένων αλλά και η προ-τακτοποίηση των αντικειμένων μέσω της εννοιολογικής χρήσης του δέντρου οδήγησε σε ένα είδος ομαδοποίησης γειτονικών αντικειμένων σε γειτονικές περιοχές.
- Το αρχείο έχει επίσης καλές επιδόσεις και όταν διαχειριζόμαστε αντικείμενα λίγων διαστάσεων. Αυτό συμβαίνει γιατί σε σχέση με το αρχείο VA είναι αναμενόμενα καλύτερο αφού βασίζεται σε δενδρική διάταξη, ενώ για τον ίδιο λόγο είναι ανταγωνιστικό και των παραδοσιακών ιεραρχικών δέντρων.

Αρνητικές όψεις:

- Η στατικότητα στον ορισμό του αριθμού(ch) των παιδιών ανά κόμβο του δέντρου, δημιουργεί προβλήματα στην κατασκευή του, γιατί συνήθως οι πραγματικά όμοιες οικογένειες μπορεί είναι περισσότερες ή λιγότερες από ch.

Αντιμετώπιση: Αν είναι περισσότερες, τοποθετούνται στο σύνολο των παιδιών οι ch κοντινότερες (με μικρότερες αποστάσεις από τον γονέα). Αν είναι λιγότερες, τότε ορίζεται ένα σχετικό όριο παιδιών και γεμάτων κόμβων στο δέντρο, πάνω από το οποίο τα παιδιά τοποθετούνται κανονικά στο δέντρο και οι υπόλοιποι κόμβοι μέχρι να συμπληρωθεί ο αριθμός παιδιών ch, συμπληρώνεται με κενούς κόμβους. Όταν όμως ο αριθμός των παιδιών μιας οικογένειας και οι υπόλοιποι γεμάτοι κόμβοι στο δέντρο είναι κάτω από αυτό το όριο, το αντίστοιχο προς δημιουργία δέντρο αποκόπτεται και δημιουργείται νέο μικρότερο δέντρο - με μικρότερο ch – ενώ το αρχικό δέντρο αναδιατάσσεται. Συνεπώς η τελική εφαρμογή μπορεί να περιλαμβάνει περισσότερα του ενός αρχεία Digenis, τα οποία κατά την αναζήτηση προσπελούνται από το μεγαλύτερο προς το μικρότερο, μέχρι να βρεθεί ομοιότητα (εάν υπάρχει).

- Μπορεί να υπάρχουν απομακρυσμένες οικογένειες – να μη συνδέονται με καμία άλλη – οι οποίες δεν μπορούν να ενταχθούν σε κανένα δέντρο.

Αντιμετώπιση: Δημιουργείται ένα Αρχείο Απομακρυσμένων (‘remote’ αρχείο) στο οποίο τοποθετούνται σειριακά οι απομακρυσμένες οικογένειες. Κατά την αναζήτηση αυτό το αρχείο προσπελώνεται πρώτο, γιατί εφόσον εν γένει θα φιλοξενεί λίγες οικογένειες, η αναζήτηση σ’ αυτό θα είναι γρήγορη. Εάν υπάρχει ομοιότητα μεταξύ

---

⁶ Με βάση τις τριγωνικές ανισότητες, όπως αποδεικνύονται για το δέντρο Digenis.

του αντικειμένου του ερωτήματος (query) και κάποιας οικογένειας του αρχείου, τότε έχει αποφευχθεί όλη η αναζήτηση στα δέντρα ενώ εάν πάλι δεν υπάρχει τέτοια ομοιότητα, λόγω του μικρού μεγέθους του αρχείου, η χρονική επιβάρυνση είναι σχεδόν αμελητέα.

Στο τελευταίο κεφάλαιο εξετάζεται ένα είδος δυναμικής αναζήτησης ομοιότητας, το οποίο ασχολείται με τις *χρονικές ακολουθίες* όχι των ίδιων των αντικειμένων [A48, A65, A66, A67, A68] αλλά των *πεδίων (χαρακτηριστικών) τους*. Δηλαδή αυτό που ανιχνεύεται είναι το κατά πόσο μοιάζει η εξέλιξη δύο χαρακτηριστικών στο χρόνο, πληροφορία που μπορεί να σταθεί πολύ χρήσιμη σε πολλά είδη εφαρμογών (ιατρικές, οικονομικές, επιστημονικές γενικά, κλπ). Χρησιμοποιώντας ένα παράδειγμα ιατρικών δεδομένων που αφορούν ορμόνες, με τη βοήθεια της προτεινόμενης μεθόδου (*Chiron*) [Σ17] εντοπίζονται με αποδοτικό τρόπο όμοια εξελισσόμενες τιμές ορμονών. Πιο συγκεκριμένα, ορίζονται νέα αντικείμενα (*property course objects* ή *Chiron objects*) τα οποία κωδικοποιούν τις μεταβολές κάθε χαρακτηριστικού, σε συγκεκριμένα χρονικά διαστήματα και στη συνέχεια αυτά τα αντικείμενα οργανώνονται σε ένα δέντρο (*Chiron tree*). Ο τρόπος που ορίζονται αυτά τα αντικείμενα, οι διαφορές τους καθώς και το δέντρο *Chiron*, καθιστούν την πλοήγηση στο δέντρο και τον εντοπισμό όμοιων *Chiron* αντικειμένων - και συνεπώς όμοια εξελισσόμενων χαρακτηριστικών - μια γρήγορη και εύκολη διαδικασία. Αυτό επιτυγχάνεται μέσω της κατανομής των *Chiron* αντικειμένων στο δέντρο *Chiron* με βάση τον αριθμό των διαφορετικών μεταξύ τους ψηφίων⁷. Μ' αυτόν τον τρόπο, δεδομένου ενός *Chiron* αντικειμένου, κατά την αναζήτηση όμοιων του στο δέντρο, μπορεί να ακολουθηθεί ένας απλός και συμπαγής αλγόριθμος μέσω του οποίου αποφεύγεται ο έλεγχος σε κείνα τα επίπεδα του δέντρου τα οποία φιλοξενούν *Chiron* αντικείμενα που σίγουρα είναι ανόμοια προς το δεδομένο αντικείμενο *Chiron*. Γενικά η μέθοδος κρίνεται πολλά υποσχόμενη, γιατί εκτός των άλλων θέτει νέα ανοικτά θέματα προς διερεύνηση, όπως η στατιστική ανάλυση των αποτελεσμάτων, η αναζήτηση αντίθετα εξελισσόμενων χαρακτηριστικών, η διαχείριση των μετατοπίσεων χρόνου (*time shifts*) ανάμεσα στα αντικείμενα πορείας χαρακτηριστικών καθώς και η βελτιστοποίηση του δέντρου *Chiron*.

⁷ Όπως συνοπτικά αναφέρθηκε παραπάνω, τα ψηφία των *Chiron* αντικειμένων κωδικοποιούν τις αυξομειώσεις κάθε χαρακτηριστικού σε συγκεκριμένη χρονική περίοδο.

Effective Organization and Management of Multi-Dimensional Objects for Knowledge Discovery Achievement

Aikaterini N. Krotopoulou

Abstract

The subject of this dissertation is the invention of methods which assure effective organization and management of multi-dimensional objects in order to achieve knowledge discovery. The initial target behind this study was the needs of a demanding application intending to *map the human brain* in order to help the localization of epileptic foci⁸. During the corresponding research, *the Representation and Management needs of human brain data* raised two core research problems:

- ✓ *The representation peculiarity of the composite, non-uniform, network structured three-dimensional objects (brain objects), and*
- ✓ *The needs for effective management-use of known and derived data and knowledge dependencies, which can upgrade the application performance and dynamics.* The most important part of our relative research, leaded to the:
  - *Investigation of similarity search aspects.* As this research area has great application and open problem width, it constitutes a great part of this dissertation.

Taking into account that the certain geometrical and knowledge dependency features of human brain data are common – all or part of them - in many modern multimedia applications, the above problems are included in the basic Data Base research problems.

Focusing our research in the above problems, we lead up to the:

- *Definition of new flexible data types, concepts, models, tools and data and knowledge ordering methods* (Data Base *BDB* [Σ2, Σ4, Σ9] and models *3D-IFO*[Σ5] and *MITOS*[Σ12, Σ13]) which organize our data more flexibly and effectively, using methods that not only assure easier data access but also exploit their ‘hidden’ relationships and dependencies for more knowledge discovery.
- *Definition of new search trees and methods for:*
  - *Effective detection of partial similarity among multi-dimensional objects* (*Lui k-n-match* [Σ18] και *INTESIS* [Σ16]).
  - *Obliteration of the high performance fall which occurs in similarity trees as dimensionality increases* (‘dimensionality curse’) (*Digenis structure* [Σ15]).
  - *Detection of object features/attributes/properties (dimensions) which have similar course in the time course – for multi-dimensional objects mostly –*

---

⁸ Within the framework of the CEC-Program AIM, Project A2020 Magnobrain.

---

aiming at the study and detection of possible interaction among them (Chiron proposal [Σ17]).

Generally, this dissertation consists of two basic parts, which refer to two research areas with great interaction:

- The Multi-Dimensional Data Base Modelling
- The Similarity Search among Multi-Dimensional objects.

In the first chapter, the problem of human brain mapping [A8, A14] for the localization of epileptic foci is discussed. This problem raises issues related to the peculiarities of the representation and the organization of three dimensional objects with complex structures/shapes and functional dependencies and relationships among them (brain objects) [Σ1, Σ10, Σ11, Π1]. In the beginning, the logical model *BDB (Brain Data Base)* [Σ2, Σ3, Σ4, Σ9] is proposed as a first approach, *introducing new entity types* [A1, A2, A3, A7, A12, A16, A17, A18]. In the corresponding study, a very interesting proposal is *the introduction of hierarchical ordering in the Relational Model* in order to organize the brain areas according to their frequencies of epileptic foci presence, improving statistically the corresponding response times. In the following, the needs of the application are studied in the basis of a Semantic – IFO model [A6] - and of an Object-oriented Model, resulting in the definition of the *3D-IFO* [Σ5] and the *MITOS (Model for the Intelligent Three-dimensional Object Support)* [Σ12, Σ13] model, respectively. In the framework of 3D-IFO model, *new data types and new operators have been introduced*, in order to achieve effective representation and better management of the complex brain objects. Additionally, *a new constructor and the suitable attribute for its support have been introduced*, in order to effectively represent the ordering among brain parts, based on a certain criterion, thus facilitating combined data retrieval. In the end, the object-oriented model MITOS, introduces a new data model (*MITOS Data Model – MDM*) which cooperates with an intelligent knowledge base approach (*MITOS Query Language – MQL*). MITOS model *introduces many novelties which serve a more expressive and intelligent representation and management of multi-dimensional data and knowledge*. One of these novelties constitutes *the definition of one more basic object characteristic* (in object-oriented theory), *the relationship with the environment*, releasing it from the situation or the behaviour [A20], where its concept and representation weakens. The second MITOS novelty concerns MQL and is related to the *introduction of the concept of 'key' in the rules area*. The extension of this potentiality – the idea comes from Data Base area [A10] – leads in fact to a kind of a key, with a meaning that it could have in Knowledge Bases and can not be exactly the same with that in Data Bases, because of the specific distinctions of these two Bases.

The subject of the second chapter is the detection of a least investigated similarity kind among multi-dimensional objects, *the partial similarity*. Partial similarity refers to similarities which are not full⁹ but they really exist. It is difficult to capture them using common techniques based on similarity functions (e.g. Euclidian distance) because these functions are affected by the whole set of object dimensions. Thus, when the objects are similar but 'very different' in few dimensions (e.g. very different colour and size) then the corresponding calculated functions (distances) will have very high values because of these few high dissimilarities and the similarity result will be negative while the objects will actually be similar. On the other hand, when between two objects there are low dissimilarities

---

⁹ Full similarity is the mostly related investigated similarity kind.

in most dimensions, they are actually dissimilar but the resultant function will have low value, so the dissimilar objects will be discerned as similar. In both cases, the common full similarity detection methods are not reliable.

The few studies that have investigated partial similarity, have mostly focused on geometric data [A62, A63, A64]. The study which is extended to multi-dimensional objects in general and has led to significant results in partial similarity, is [A39]. [A39] introduces the effective methods *k-n-match* and *frequent k-n-match*, which result in *k* objects being similar to the given ones not in all their dimensions but at least in *n* ones, avoiding in this way those few very dissimilar dimensions –if any– which lead to false results. Nonetheless, these methods have some weaknesses which finally result either in full similarity (when finally, in *frequent k-n-match*, all *n*'s are taken into account) or in an occasional partial similarity detection (with the suitable *n*'s, which should not be very high or very low, without having however any type or method to calculate the 'best' *n*'s). Based on these methods, we propose two techniques which can provably detect real partial similarities. The first of them, *Lui k-n-match* [Σ18], succeeds in the approximate specification of the suitable *n*'s for the *k-n-matches*, based on human-computer interaction and on the suitable checks of the similar objects that *k-n-matches* propose. More precisely, using *k-n-match*, for each *n* a set with objects possibly similar to the given one (query object), is proposed¹⁰. The user filters this set and decides which objects of the proposed set are really similar to the given one. This procedure continues until the point where, while *n* has become larger than $d/2^*$, the user does not select any object as similar¹¹ from the proposed object set. In this way, *the results are more reliable and valid (because of human-computer interaction) while in parallel the number of the executed k-n-matches are remarkably reduced.*

The second partial similarity detection method (*INTESIS*) [Σ16] is based on the following observation: when two objects are similar, it usually means that they are similar in most of their characteristics. In data bases, each of object characteristic is represented by a set (usually small) of features-attributes(dimensions). Thus, if this correspondence between a characteristic and a set of attributes is defined by the developer of each application - creating dimension subsets – then the following can be successively done:

- a) A full similarity detection for each dimension subset
- b) Organization of these subsets in the corresponding hierarchical trees for their easy and effective management. The additional simplification of this choice derives from the fact that as long as the dimension subsets are small, the selection of the corresponding tree will be a very easy task, while almost all hierarchical trees have high performance for low dimensionalities. Consequently, the developer of each application can use the hierarchical tree that he/she considers as best (our proposition is *R*-tree*).

Finally, in order to complete the procedure, the application developer has to define which is the minimum number of the requisite similar characteristics that indicate partial similarity, for the particular application.

For the evaluation of the method, first of all, it is necessary to mention that it refers to a total number of dimensions less than 100 and consequently to a relatively small number of trees. As it is obvious, the final response time in a uniprocessor system is the sum of the response

¹⁰ The set does not include objects selected in previous stages (i.e. for previous (small) *n*'s)

* *d*=dimensionality of the related objects

¹¹ It means that the partial similar objects have already been selected and from now on, the next proposed objects will be rather fully similar to the given one.

times of each tree. Taking into account that the number of dimensions which correspond to each tree is small, these trees have very good response times and consequently the total response time is low enough. While the use of each tree does not presuppose the use of another tree before or after it, the search in each tree can be performed in parallel. Therefore, *in a multi-processing system, the total response time can be considerably reduced, achieving to reach the time needed for only one tree* (when the number of processors is equal to the number of trees). Furthermore, *bearing in mind that partial similarity forms a very demanding similarity search kind, not only the response times in multi-processing systems but those times in a uniprocessor system constitute satisfying performances.*

The third chapter studies the potentiality of *defining a new structure which does not ‘suffer’ from ‘dimensional curse’, while it assures good performance for low dimensionalities too.* The latest studies have resulted in the following: Although the known similarity trees (either based on space partitioning [A37] or on data partitioning [A22, A33, A38, A44, A45, A49, A50, A52, A56, A58, A60, A61] perform effectively in cases of low dimensionality, their performance generally degrades as dimensionality increases (especially for more than 10 dimensions) [A42]. On the other hand, VA-File constitutes a simple approximate method (it is a simple array-file of object geometric approximations) which manages to outperform any other similarity search method at high dimensionality but it has low performance for low dimensionality [A42]. In order to overcome this determinant dependence between the performance and the dimensionality of a data-object set, we propose the new hybrid structure called *Digenis*¹² [Σ15], which marries the logic of similarity trees with VA-Files logic. More precisely, a static¹³ parametric tree (*Digenis tree*) is defined in conceptual level while the *Digenis file*, based on Digenis tree, is used in physical level. Using this correlation, a) the similarity search procedure is located in a small part of the file, excluding most dissimilar objects from the search and b) the method is used effectively for both low and high dimensional objects, preserving generality and flexibility.

The first necessary definition for Digenis proposal is related to the *object families*. They consist of objects having a small distance among them (based on a certain limit *fl* defined from the analyst, in each case) and they are represented by the ‘mean’ object of the family (if it does not exist, it is created just for this role). Each object family is hosted in a node of Digenis tree. The distance which is used is the most spread one, the *Euclidian distance*, for which the triangle inequality – where the method is mainly based - stands. Additionally, a second distance limit (*Lt*) is defined – from the analyst- which forms the limit used to conclude if two objects are similar or not. Finally, the root of the Digenis tree is the most ‘centered’ family in the total object area and the nodes being the children of it are its *ch* nearest families-nodes. The children of each of them are its *ch* nearest families, and so on. The triangle inequality which stands among the Euclidian distances of the object-families, is proved to be a very useful Theorem for the safe check exclusion of a great part of the tree¹⁴,

¹² The name selected is Digenis, the name of the famous brave guard of the borders of the Byzantine empire. ‘Digenis’ means bi-descent, while his father was a Syrian and his mother was Greek. Behind the selection of the name lies the hybrid character of the proposal.

¹³ The static number of children (*ch*) in each level, is necessary in order to serve the transformation of the tree to the corresponding Digenis file.

¹⁴ According to triangle inequality, if *a, b, c* are the edges of the triangle, then the following stands: $a - b < c < a + b$. Consequently, considering that the data query object *q*, the Digenis tree root *r* and the node *b* being child of the root, define a triangle, the distance between *q* and *b* (d_{qb}) is definitely longer than the absolute value of the difference between the distance of *q* (d_q) and *b* (d_b) from the root ($d_{qb} > |d_q - d_b|$). While these two distances and *Lt* are known, it is easy to estimate if d_{qb} is longer than *Lt* without any distance calculation. More precisely, if $d_{qb} \geq |d_q - d_b| \wedge |d_q - d_b| \geq Lt$ then (transitional property) $d_{qb} \geq Lt$ stands.

leading to a final check in a small tree area. The search analysis of the tree is very helpful for the use of Digenis file, where the real search is performed (physical level).

The corresponding Digenis file in the physical level is created if each tree node composes a record of the file, beginning from the tree root and passing from each level, from left to right. Using this correspondence, the proved Digenis tree techniques¹⁵ of easy, safe and quick exclusion of Digenis record areas can be used.

The (theoretical and experimental) evaluation of the method results in the detection of certain advantages and disadvantages of it.

Advantages:

- *The file has very good performance for high dimensionalities.* This was expected because the file works as a kind of VA-File, where the records are compact geometric approximations. This matters because both the use of object families achieves a first data ‘compression’ and the pre-arrangement of the objects via the conceptual use of the tree lead to a kind of grouping of neighboring objects in neighboring areas.
- *The file has also good performance for low dimensionality,* because in comparison to VA-File, it is expectably better while it is based on a tree structure. For the same reason, Digenis file is competitive to the classic hierarchical similarity trees.

Drawbacks:

- The fact that the number of children for each node is statically defined as ch in each application is a disadvantage for the construction of the tree, because usually the really similar families may be more or less than ch .

Confrontation: If the similar families of a node are more than ch , then only the ch closest to the family are placed as its children, in the next level. If they are less than ch , then a limit of children and full nodes in the tree is defined. When this limit is overcome, the nodes-children are normally placed in the tree and the rest nodes – until ch -th one – remain empty. When however the number of the children of a family and of the full nodes in the tree, are less than this limit, the corresponding subtree is separated, creating a new smaller tree – with smaller ch – while the initial tree is reorganized. Consequently, the final application may include more than one Digenis tree, which are accessed from the bigger to the smaller, until the similarity is found (if any).

- Perhaps there are remote areas of object families – without any connection with other families – which can not be included in any other tree.

Confrontation: A file including sequentially the remote families (called ‘remote’ file’) is created. During the similarity search, this file is the first which is accessed because while it usually hosts a few families, the search will be quick enough. If a similarity is detected (among the query object and a family in the file), then the search in the trees will be avoided while if no similarity exists, the time overhead of the file search is almost negligible, because of its size.

In the last chapter, a new kind of dynamic similarity search is investigated. It is related with the *time streams* not of the objects themselves [A48, A65, A66, A67, A68] but of *their*

Based on the above analysis, analogous conclusions can be extracted for the next levels.

¹⁵ Based on the triangle inequalities.

---

properties/attributes/dimensions. In other words, what is detected is whether the courses of two or more properties resemble. This kind of information can be very useful for several kinds of applications (medical, financial, scientific in general, e.t.c). Using medical data related to hormonal tests as an example, we prove that, based on our method *Chiron*¹⁶ [Σ17], the hormones which are developed in the same way are accurately and effectively detected. More precisely, new objects (*property course objects* or *Chiron objects*) which encode the variations of each property in certain time intervals, are defined and organized in a tree (*Chiron tree*). The way these objects are defined, their differences and the Chiron tree itself make its navigation and the detection of similar Chiron objects – and consequently of properties which are developed in a similar way - a quick and easy procedure. This is achieved via the distribution of the Chiron objects in the Chiron tree according to the number of the different digits¹⁷ that exist among them. In this way, when we search in the Chiron tree for objects similar to a given one, a simple and compact algorithm is used, which avoids a vast amount of useless checks among very different objects. Generally, the method is promising enough because it *poses new problems for investigation*, like *the statistical analysis of its results*, *the search for objects that are developed in a reverse way*, *the management of time shifts among the property course objects* and *the Chiron tree optimization*.

---

¹⁶ The whole proposal is named Chiron, the name of the famous wise Centaur, which was a race of creature composed of part human and part horse. That ‘difference’ – indicating generally the concept of difference – was a key means of the wisdom hunting. We borrow his name, to indicate our focus on the *difference in order to extract new data dependency knowledge*.

¹⁷ As concisely referred above, the Chiron object digits encode the fluctuations of each property in certain time intervals.

ΔΗΜΟΣΙΕΥΣΕΙΣ♦ **ΣΥΝΕΛΠΙΑ**

- [Σ1] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "Mapping the brain for the localization of epileptic foci", Satellite Symposium on Neuroscience and Technology 14th International Conference IEEE, Lyon, France, 1992.
- [Σ2] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "BDB (Brain Data Base), DEXA 93 4th International Conference on Database and Expert Systems Applications, Prague, Czech Republic.
- [Σ3] Γ. Ανωγειανάκης, Α. Κροτοπούλου, Π. Σπυράκης, Δ. Τέρπου, Α. Τσακαλίδης, «Ένα Πληροφοριακό Σύστημα για τη Χαρτογράφηση του Ανθρώπινου Εγκεφάλου», Ε.Π.Υ., 4ο Πανελλήνιο Συνέδριο Πληροφορικής ΕΠΥ, 16-18 Δεκεμβρίου 1993, Πάτρα.
- [Σ4] Γ. Ανωγειανάκης, Α. Κροτοπούλου, Π. Σπυράκης, Δ. Τέρπου, Α. Τσακαλίδης, «BDB Σύστημα» : Ένα Σύστημα Χαρτογράφησης του Ανθρώπινου Εγκεφάλου για τον εντοπισμό των επιληπτικών εστιών», 3ο Πανελλήνιο Συνέδριο Ιατρικής Πληροφορικής, 13-15 Μαΐου 1994, Θεσσαλονίκη.
- [Σ5] A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "A Conceptual DataBase Approach for Modelling 3D Objects of Irregular Geometry", DEXA 94, 5th International Conference on Database and Expert Systems Applications, Athens, Greece, September 1994.
- [Σ6] Krotopoulou A., Lafazanis M., Panagopoulou G., Sirmakessis S., Spirakis P., Tampakas V., Terpou D., Tsakalidis A., Vlacholoulos P., "The Design of a Text DataBase for Biomedical Information (FINE DataBase)", 1st International Conference on Medical Physics and Biomedical Engineering, Nicosia, Cyprus, 1994.
- [Σ7] Krotopoulou A., Lafazanis M., Panagopoulou G., Sirmakessis S., Spirakis P., Tampakas V., Terpou D., Tsakalidis A., Pallikarakis N., "Fine Database: A DataBase for Information Exchange Among Clinical Engineers", 12th International Congress on Medical Informatics, Lisboa, 1994.
- [Σ8] Krotopoulou A., Lafazanis M., Panagopoulou G., Sirmakessis S., Spirakis P., Tampakas V., Terpou D., Tsakalidis A., Pallikarakis N., "The Development of a Database (FDB) for Information Exchange Among Clinical Engineers", Fourth International Systems Development, (ISD '94) Bled, Slovenia.
- [Σ9] G. Anogianakis, Krotopoulou, G. Nendidis, P. Spirakis, D. Terpou, A. Tsakalidis, M. Psaltikidou. MAGNOBRAIN II: The Brain Database (BDB). 8th Congress of the International Society of the Greek Neuroscientists, Rhodes, September 1994.
- [Σ10] G. Anogianakis, M. Apostolakis, S. Bountzioukas, K. Krotopoulou, G. Nendidis, P. Spirakis, D. Terpou, A. Tsakalidis. Mapping the Brain in a structural way helps in Localizing Brain Function. Topographic Representation of the Visual Cortex Symposium, Porto Carras, 1994.
- [Σ11] G. Anogianakis, M. Apostolakis, G.F.A Harding, K. Krotopoulou, G. Nendidis, M. Psaltikidou, P. Spirakis, D. Terpou, A. Tsakalidis, 1995. MAGNOBRAIN: Integrating Radiological, Anatomical and Electrophysiological Information about the Brain for Source Localization. In Interactive Technology and the New Paradigm for Healthcare. Washington, DC:IOS Press. pp. 4-14.

- [Σ12]Krotopoulou A., Spirakis P., Terpou D, Tsakalidis A., «Mitos: A Model for the Intelligent Support of Three-Dimensional Objects», Joint European Conference and Exhibition on Geographical Information, (JEC'97), 16-18 April 1997, Vienna, Austria.
- [Σ13]Krotopoulou A., Spirakis P., Terpou D, Tsakalidis A., "An Object Oriented Data Model for the Intelligent Support of Three-Dimensional Objects", DEXA 97, 8th International Conference on Database and Expert Systems Applications, 1-5 September 1997, Toulouse, France.
- [Σ14]Krotopoulou A., Spirakis P., Terpou D, Tsakalidis A., "Similarity Queries in the Presence of Scaling, Translation and Rotation in Large DataBases", RSDMGrC'98, Durham, U.S.A.
- [Σ15]Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, "Digenis: An Hybridic Proposal for (Dis)Similarity Search", IKE 2009 (ranking:0,80).
- [Σ16]Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, "Integrated Similarity Search Among Multi-Dimensional Objects (InteSiS)", IKE 2009 (ranking:0,80).
- [Σ17] Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, "CHIRON: A Property Based Proposal for Discovery", IKE 2009 (ranking:0,80).
- [Σ18]Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, "Improving the Results of Partial Match", short paper, DMIN 2009 (ranking:0,79).

♦ ΠΕΡΙΟΔΙΚΑ

- [Π1] G.Anogianakis, M.Apostolakis, S.Bountzioukas, K.Krotopoulou, G.Nendides, P. Spirakis, D. Terpou, A. Tsakalidis, "*Mapping the Brain in a Structural Way Helps in Localizing Brain-Function*", International Journal of Psychophysiology, Vol. 18, Iss 2, pp.91-91 1994 (36^ο σε σύνολο 482 Journals στην κατηγορία του).
- [Π2] Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, *TIF (Tree Index File) : A Flexible Proposal For Similarity Search*, SIGMOD Record, υπό κρίση.

ΤΕΧΝΙΚΕΣ ΑΝΑΦΟΡΕΣ

- [1] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "The logical design of the Brain Data Base", Technical Report MB1.1., April 1992.
- [2] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "Design and Implementation issues of the Brain Data Base (BDB)", Technical Report MB1.2, September 1992.
- [3] V. Tambakas, S. Sirmakessis, G. Panagopoulou, K. Krotopoulou, D. Terpou, M. Lafazanis, "Design of the S/W components of the FINE System and System Configuration", Chapter 5, Deliverable #13, December 1992.
- [4] V. Tambakas, S. Sirmakessis, G. Panagopoulou, K. Krotopoulou, D. Terpou, M. Lafazanis, "FINE: The Pilot Phase Demonstrator", Chapter 5, Deliverable #21, October 1993.
- [5] V. Tambakas, S. Sirmakessis, G. Panagopoulou, K. Krotopoulou, D. Terpou, M. Lafazanis, "FINE Integration, Evaluation and System Status", Deliverable #30.

ΑΝΑΦΟΡΕΣ

- [A1] Laird AR, Lancaster JL, Fox PT. (2005). BrainMap: The social evolution of a functional neuroimaging database. *Neuroinformatics* 3, 65-78.
- [A2] Fox PT, Lancaster JL. Mapping context and content: The BrainMap model. *Nature Rev Neurosci* 3, 319-321, 2002.
- [A3] Batrancourt B., Levy R., Lehericy S., Hasboun D., Samson Y., Lavallee I., Lamure M., Dubois B.: Base de données anatomo-fonctionnelle sur le cerveau. An anatomofunctional brain database, *Comptes Rendus Biologies*, Volume 325, Number 4, April 2002 , pp. 439-455(17).
- [A4] Bjaalie JG, Grillner S, Usui S.: *Neuroinformatics: databases, tools, and computational modeling for studying the nervous system*. Neural Netw. 2008; 21:1045-6.
- [A5] Nenonen, H. – M., Rajala and T. Katila.: Biomagnetic localization and 3D modelling, Report TKK-F-A689, Helsinki University of Technology, Department of Technical Physics, Otaniemi 1991.
- [A6] Critical reviews in biomagnetic engineering, vol. 14, issue 2.
- [A7] Franzblau, D.S., Kleitman, D.J.: An Algorithm for Constructing Regions with Rectangles, Independence and Minimum generating sets for collections of Intervals. In 16th ACM Symposium Theory Computing, pp. 167-174, 1984.
- [A8] Willard, D.: New Data Structures for orthogonal range queries. In *SIAM Journal on Computing*, Vol. 14, pp.232-253, 1985.
- [A9] Preparata, F., Shamos, M.: *Computational Geometry. An Introduction*. Springer-Verlag 1986.
- [A10] Korth, E., Silberschatz, A.: *Database system concepts*. McGrawhill, University of Atlas at Austin, 1986.
- [A11] Rubeinstein, W.B.: *A Database Design for Musical Information*. Computer Science Division, University of California, Berkeley, ACM SIGMOD 1987.
- [A12] Abiteboul, S., Hull, R.: IFO: A Formal Semantic Database Model. In *ACM Transactions on Database Systems*, 1987.
- [A13] Overmars, M.H.: Efficient Data Structures for Range Searching on a Grid. In *Journal of Algorithms* 9, pp.254-275, 1988.
- [A14] Talairach, J., Tournoux, P.: *Co-Planar stereotaxic atlas of the brain*. Georg Thieme Verlag Stuttgart, New York, Thieme Medical Publishers, Inc. New York, 1988.
- [A15] Hull, R., Yap, C.K.: The Format Model: A theory of database organization. *J.ACM* 31, 3, 1984.
- [A16] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufmann, 1989.
- [A17] Parsaye, K., Chignell, M., Khoshafian, S., Wong, H.: *Intelligent Databases*. John Wiley and Sons Inc, 1989.
- [A18] Mehlhorn, K., Tsakalidis, A.: Chapter 6: Data Structures. In *Handbook of Theoretical Computer Science*, Editor J. V. Leeuwen, Elsevier Science Publishers, co-published by MIT-Press, 1990.

-
- [A19] Egemhofer, M., Herring, J.: A Mathematical Framework for the Definitions of Topological Relationships. In the 4th International Symposium on Spatial Data Handling 1990.
- [A20] Booch G.: Object-Oriented Design with applications. Benjamin 1991.
- [A21] Jacobson I.: Object-Oriented Software Engineering – A use case Driven Approach. Addison-Wesley 1992.
- [A22] Beckmann, N., Kriegel, H-P., Schneider, R., Seeger, B.: The R*-tree: An Efficient and Robust Access method for Points and Rectangles. In SIGMOD 1990.
- [A23] Theodoridis Y., Sellis T., A Model for the Prediction of R-tree Performance, PODS 1996, pp. 161-171
- [A24] Chakrabarti K., Mehrotra S.: Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces, VLDB 2000.
- [A25] Pechura, C., Martin, J.: Mapping the Brain and its functions. National Academy Press, Washington, D.C., 1991.
- [A26] Hiroshi, I. et al.: The Model, Language and Implementation of an Object-Oriented Multimedia Knowledge Base Management System. ACM Transactions on Database Systems, 1993.
- [A27] Bistiolas, V., Sofotassios, D., Tsakalidis, A.: Computing Rectangle Enclosures. In Computational Geometry: Theory and Applications, 2,2 1993.
- [A28] Kapelios, V., Papamichail, G.: Intersection of orthocanonic regions and computational geometry algorithms. Diploma Thesis, Department of Computer Engineering and Informatics, University of Patras, Greece, 1993.
- [A29] Margioulas, V., Pikrakis, A.: Algorithms for compression of files with geometric data. Diploma Thesis, Department of Computer Engineering and Informatics, University of Patras, Greece, 1993.
- [A30] Novak, M., Gardarin, G., Valduriez, P.: Flora: A Functional-Style Language for Object and Relational Algebra. In DEXA 1994.
- [A31] Hohenstein, U., Lauffer, R., Weikert, P.: Object-Oriented Database Systems: How Much SQL Do They Understand?. In DEXA 1994.
- [A32] Chen, Q., Chen, L., Lian, X., Liu, Y., Xu, Yu.: Indexable PLA for Efficient Similarity Search. In Very Large Data Bases (VLDB) 2007.
- [A33] Ciaccia, P., Patella P., Zezula P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In Very Large Data Bases (VLDB) 1997.
- [A34] Hellerstein, J.M., Jeffrey F. Naughton, J.F., Pfeffer, A.: Generalized Search Trees for Database Systems. In Very Large Data Bases (VLDB) 1995.
- [A35] Hu H., Lee D.L., Lee V.S.: Distance Indexing on Road Networks. In Very Large Data Bases (VLDB) 2006.
- [A36] Kibriya, A., Frank, Eibe.: An Empirical Comparison of Exact Nearest Neighbour Algorithms. In PKDD 2007.
- [A37] Nievergelt, J., Hinterberger, H., Servik, K.: The grid file: An adaptable symmetric multikey file structure. ACM Transactions on Database Systems, 1984.
-

-
- [A38] Sakurai Y., Yoshikawa M., Uemura S., Kojima H.: The A-tree: An Index Structure for High-Dimensional Spaces Using Relative Approximation. In Very Large Data Bases (VLDB) 2000.
  - [A39] Tung, A.K.H, Zhang, R., Koudas, N., Ooi, B.G.: Similarity Search: A Matching Based Approach. In Very Large Data Bases (VLDB) 2006.
  - [A40] Beyer, K., Goldstein, J., Shaft, U.: When is Nearest Neighbors Meaningful? In ICDT 1999.
  - [A41] Hinneburg, A., Aggarwal, C., Keim, D.: What is the nearest neighbor in high dimensional spaces? In VLDB 2000
  - [A42] Weber, R., Schek, H.J, Blott, S.: A Quantitive Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In Very Large Data Bases VLDB) 1998.
  - [A43] Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search - The Metric Space Approach, Springer, 2006.
  - [A44] Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Searching. In SIGMOD 1984.
  - [A45] Seeger, B., Kriegel, H.P.: The Buddy Tree: An Efficient and Robust Access Method for Spatial Data Base Systems. In Very Large Data Bases (VLDB) 1990.
  - [A46] Korn P., Sidiropoulos N., Faloutsos C., Siegel E., Protopapas Z.: Fast and Effective Retrieval of Medical Tumor Shapes. IEEE Transactions on Knowledge and Data Engineering, Vol 10, No 6, 1998.
  - [A47] Arya S., D.M. Mount, N.S. Netanyahu R. Silverman, and A. Y. Wu. An Optimal Algorithm for Approximate Nearest Neighbor Searching in Fixed Dimensions. Journal of the ACM, vol. 45, no. 6, pp.891-923
  - [A48] Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases. In SIGMOD 1994.
  - [A49] Kamel, I., Faloutsos, C.: Hilbert R-tree: An Improved R-tree Using Fractals. In Very Large Data Bases (VLDB) 1994.
  - [A50] Ciaccia, P., Patella, M.: The M2-tree: Processing Complex Multi-Feature Queries with Just One Index. In 1st DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland, 2000.
  - [A51] Aggarwal, C.C.: Towards meaningful high-dimensional nearest neighbor search by human-computer interaction. In ICDE 2002.
  - [A52] Arge, L., Mark de Berg, Haverkort, H.J., Yi K.: The Priority R-Tree: A Practically Efficient and Worst-Case Optimal R-Tree. In SIGMOD 2004.
  - [A53] Augsten, N., Bohlen, M., Gamper, J.: An Incrementally Maintainable Index for Approximate Lookups in Hierarchical Data. In Very Large Data Bases (VLDB) 2006.
  - [A54] Chen, Q., Chen, L., Lian, X., Liu, Y., Xu, Yu.: Indexable PLA for Efficient Similarity Search. In Very Large Data Bases (VLDB) 2007.
  - [A55] Wong, R.C.-W., Tao, Y., Fu, A.W.-C., Xiao, X.: On Efficient Spatial Matching. In Very Large Data Bases (VLDB) 2007.
-

-
- [A56] White, D.A., Jain, R.: Similarity indexing with the SS-tree. In 12th International Conference on Data Engineering (ICDE) 1996.
- [A57] Goh, K.-S., Li, B., Chang, E.: DynDex: A Dynamic and Non-metric Space Indexer. In Multimedia 2002.
- [A58] Qian, G., Zhu, Q., Xue, Q., Pramanik, S.: The ND-Tree: A Dynamic Indexing Technique for Multidimensional Non-ordered Discrete Data Spaces. In Very Large Data Bases (VLDB) 2003.
- [A59] Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A., Theodoridis, Y.: R-Trees: Theory and Applications, Springer 2005. ISBN 1-85233-977-2.
- [A60] Berchtold, S., Keim, D., Kriegel, H.-P.: The X-tree: An Index Structure for High-Dimensional Data. In Very Large Data Bases (VLDB) 1996.
- [A61] Katayama, N., Satoh, S.: The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. In SIGMOD 1997.
- [A62] Bronstein, A., Bronstein, M., Bruckstein, A., Kimmel, R.: Partial Similarity of Objects, or How to Compare a Centaur to a Horse, International Journal of Computer Vision, 2008.
- [A63] Bruckstein, A. M., Katzir, N., Lindenbaum, M., & Porat, M.: Similarity-invariant signatures for partially occluded planar shapes. *IJCV* 1992, 7(3), 271–285.
- [A64] Kim, C. R., & Chung, C. W.: XMage: An image retrieval method based on partial similarity. *Information Processing and Management*, 2006, 42, 484–502.
- [A65] Negi, T., Bansal V.: Time Series: Similarity Search and its Applications, ICSCI 2005.
- [A66] S.Zhang C. S. Perng, H. Wang and D. S. Parker. Landmark: A new technique for similarity based pattern querying in time series databases. In Proc ICDE, 2000.
- [A67] Xianping Ge, P. Smyth. Deformable markov model templates for time series pattern matching. Intl Conf. of SIGKDD, August 2000.
- [A68] Chan. K and A.W. Fu. Efficient time series matching by wavelets. In Proc. of the 15th Intl. Conf. on Data Engineering, pages 126–134, 1999.
-

Μέρος Α':

Οργάνωση και Διαχείριση Δεδομένων και Γνώσης –

Βάσεις Δεδομένων - Γνώσης

Η επιστήμη είναι δεδομένα. Ακριβώς όπως τα σπίτια χτίζονται με τούβλα, έτσι και η επιστήμη χτίζεται με δεδομένα. Αλλά όπως ένας σωρός τούβλα δεν κάνει ένα σπίτι έτσι και μια συλλογή δεδομένων δεν είναι απαραίτητα επιστήμη.

Ανρί Πουανκαρέ, 1854-1912, Γάλλος Μαθηματικός

Κεφάλαιο 1

Ορισμός Νέων Μοντέλων και Μεθόδων Ταξινόμησης Τρισδιάστατων Δεδομένων Πολύπλοκης Δομής και Ταξινόμησης Γνώσης (BDB, 3D-IFO και MITOS)

Περιεχόμενα Κεφαλαίου

- 1.1. Το πρόβλημα: Χαρτογράφηση του Ανθρώπινου Εγκεφάλου
  - 1.1.1. Αρχικός Σκοπός
- 1.2. Brain Data Base
- 1.3. 3D-IFO Μοντέλο
  - 1.3.1. IFO Model
    - 1.3.1.1. Αντικείμενα IFO (IFO Objects)
  - 1.3.2. 3D-IFO Τοπολογικές Επεκτάσεις
  - 1.3.3. 3D-IFO Επέκταση Μηχανισμών Κατασκευής (Constructor), Εισαγωγή Πεδίων (attributes) και Σχετικών Τελεστών
  - 1.3.4. Παράδειγμα χρήσης
- 1.4. Αντικειμενοστραφές Μοντέλο MITOS (Model for the Intelligent Support of Three-Dimensional Objects – Μοντέλο για την Ευφυή Υποστήριξη Τρισδιάστατων Αντικειμένων)
  - 1.4.1. Εισαγωγή Νέου Χαρακτηριστικού στην Έννοια του Αντικειμένου
    - 1.4.1.1. Εισαγωγή του Χαρακτηριστικού στην Τυπική Έκφραση της Κλάσης
  - 1.4.2. Εισαγωγή της Διάταξης και της Έκδοσης στον Ορισμό της Κλάσης
    - 1.4.2.1. Ορισμός Διάταξης και Έκδοσης
    - 1.4.2.2. Ορισμός Ενσωματωμένων (Built in) Μεθόδων Διαχείρισης Διάταξης και Έκδοσης
  - 1.4.3. Νέος ορισμός των Κανόνων
- 1.5. Ανακεφαλαίωση - Συνεισφορά
- 1.6. Δημοσιεύσεις του Κεφαλαίου
- 1.7. Αναφορές σε Δημοσιεύσεις μας
- 1.8. Αναφορές του Κεφαλαίου

Σύνοψη

Σε κάποιους τύπους επιληψίας, η αξιοποίηση των ευρημάτων συγκεκριμένων εξετάσεων για τον εντοπισμό επιληπτικών εστιών είναι αρχικά πρακτικά ανέφικτη λόγω του μεγάλου όγκου δεδομένων που πρέπει να συνδυαστούν με τα εργαστηριακά αποτελέσματα προκειμένου να προκύψουν ασφαλή συμπεράσματα. Αυτή η αδυναμία μπορεί να αρθεί με τη χρήση της κατάλληλης βάσης δεδομένων για την ηλεκτρονική χαρτογράφηση του εγκεφάλου [Σ1, Σ7, Σ8, Π1], μέσω της οποίας γίνεται εφικτή και εύκολη η οργάνωση και αποδοτική χρήση των δεδομένων του εγκεφάλου.

Ο εγκέφαλος αποτελεί ένα ιδιαίτερα σύνθετο και πολύπλοκο σύνολο δεδομένων προς αναπαράσταση και διαχείριση σε επίπεδο βάσεων δεδομένων, αφού αποτελείται από τρισδιάστατα πολύπλοκης, δικτυακής πολλές φορές δομής σχετιζόμενα αντικείμενα, για τα οποία υπάρχει πολλή και ποικίλη σχετική πληροφορία. Η έρευνά μας που σκοπό έχει τη διαχείριση αυτών των δεδομένων κινείται σε δύο κατευθύνσεις: α) *Την αποδοτική αναπαράσταση αυτών* (και γενικότερα τέτοιου είδους) *των αντικειμένων* και β) *Την ευέλικτη και αποτελεσματική οργάνωσή τους*, έτσι ώστε να επιτυγχάνεται ευκολότερη και γρηγορότερη ανάκτηση της κατάλληλης πληροφορίας, η οποία με τη χρήση νευροφυσιολογικής κυρίως γνώσης μπορεί να οδηγήσει σε ανακάλυψη νέας ιατρικής γνώσης.

Το αποτέλεσμα αυτής της έρευνας είναι οι ορισμοί τριών μοντέλων. Αρχικά ορίστηκε η σχεσιακή βάση δεδομένων Brain Data Base (BDB) [Σ2, Σ3, Σ4, Σ6] στα πλαίσια της οποίας εισήχθησαν νέοι τύποι δεδομένων για την υποστήριξη των απαιτήσεων των τμημάτων του εγκεφάλου καθώς και μία νέα έννοια (*ιεραρχική ταξινόμηση*) στο Σχεσιακό Μοντέλο. Η μελέτη νεώτερων μοντέλων και λογικών οδήγησε στο Σημαντικό Μοντέλο IFO [A12] το οποίο αποτέλεσε τη βάση για τον ορισμό του 3D-IFO Μοντέλου (επέκταση του IFO). Στο 3D-IFO [Σ5] εισήχθησαν οι απαιτούμενοι για τα δεδομένα του εγκεφάλου *νέοι τύποι* καθώς και *νέοι τελεστές* για τη διαχείρισή τους. Επίσης ορίστηκε ένας νέος μηχανισμός κατασκευής (constructor) για την απ'ευθείας αναπαράσταση της διάταξης των μερών του εγκεφάλου με βάση κάποιο κριτήριο.

Ερευνώντας ένα ακόμη βήμα πιο μπροστά και παρακολουθώντας την τρέχουσα εξέλιξη των μοντέλων βάσεων δεδομένων της εποχής (μέσα δεκαετίας του '90), μελετήθηκαν οι απαιτήσεις της αρχικής εφαρμογής – χαρτογράφηση εγκεφάλου – σε συνδυασμό με τις δυνατότητες του αντικειμενοστραφούς (object oriented) προγραμματισμού. Το προϊόν αυτής της έρευνας ήταν το αντικειμενοστραφές μοντέλο MITOS [Σ9, Σ10] το οποίο εισάγει ένα νέο μοντέλο δεδομένων (MDM) που συνεργάζεται με μία νέα γλώσσα ερωτημάτων (MQL). Το MITOS εισάγει κάποιες καινοτομίες προκειμένου να υποστηρίξει την αναπαράσταση και διαχείριση των πολύπλοκων δεδομένων και της γνώσης που αφορά τον ανθρώπινο εγκέφαλο. Η μία από αυτές αφορά την εισαγωγή ενός ακόμη βασικού χαρακτηριστικού των αντικειμένων, της *σχέσης τους με το περιβάλλον*, ενώ η άλλη βασική καινοτομία αφορά την υιοθέτηση της *έννοιας του κλειδιού και στους κανόνες των βάσεων γνώσης*.

Λέξεις Κλειδιά: Τρισδιάστατα δεδομένα, τοπολογικές συσχετίσεις, ιεραρχική διάταξη, σχέση αντικειμένου με το περιβάλλον, κλειδί κανόνα.

1.1. Το πρόβλημα: Χαρτογράφηση του Ανθρώπινου Εγκεφάλου [Σ1, Σ7, Σ8, Π1]

Ο ανθρώπινος εγκέφαλος αποτελεί ένα σύνθετο σύστημα δομικών (Σχήματα 1.1. έως 1.4.), χημικών και ηλεκτρικών συσχετίσεων και διασυνδέσεων, το οποίο εδώ και μερικές μόνο δεκαετίες έχει αρχίσει να 'αποκαλύπτεται' σταδιακά και τμηματικά από την ιατρική. Είναι το

πιο ανεξερευνήτο σήμερα ανθρώπινο όργανο, γεγονός που οφείλεται στην τεράστια πολυπλοκότητά του (ανατομική και λειτουργική, ποιοτικά και ποσοτικά). Αυτή η πολυπλοκότητα αποτέλεσε και την αρχική αφορμή συνεργασίας της νευροφυσιολογίας με τον τομέα των βάσεων δεδομένων, με σκοπό όχι μόνο την ευκολότερη και γρηγορότερη εξαγωγή αποτελεσμάτων αλλά και την επίτευξη του εφικτού εκεί που δεν υπήρχε. Πιο συγκεκριμένα, υπάρχουν κάποιες μέθοδοι τα αποτελέσματα των οποίων μπορούν να οδηγήσουν σε διάγνωση μόνο όταν συνδυαστούν με πολλά άλλα δεδομένα, το πλήθος και η πολυπλοκότητα των οποίων δεν επιτρέπει την έγκαιρη εξαγωγή αποτελεσμάτων, στις παραδοσιακές συνθήκες. Εκεί είναι που η κατάλληλη χρήση βάσεων δεδομένων μπορεί με την οργάνωση και τα εργαλεία που προσφέρει να οδηγήσει σε έγκαιρες και έγκυρες λύσεις.

Σχήμα 1.1. : Έξω Όψη Ανθρώπινου Εγκεφάλου

Σήμερα υπάρχουν αρκετές εξειδικευμένες εφαρμογές [A1, A2, A3, A4] που χειρίζονται δεδομένα του ανθρώπινου εγκεφάλου, γεγονός που συνδέεται με τη σπουδαιότητα αυτού του είδους χαρτογράφησης. Η δική μας σχετική έρευνα τοποθετείται γύρω στο 1992, εποχή που ακόμα και η σπουδαιότερη ίσως σήμερα βάση δεδομένων για τον εγκέφαλο, η BrainMap [A1,

A2], διάνυνε τα πρώτα στάδια ανάπτυξής της. Έτσι παρόλο που τα αποτελέσματά μας δεν παρουσιάζουν την πρωτοτυπία που είχαν τότε, αποτελούν πρωτότυπη δική μας εργασία και έθεσαν τις βάσεις για τη συνέχιση της έρευνάς μας, ενώ υπάρχει και σχετική αναφορά άλλων ερευνητών σε αυτή [C1, C2].

Σχήμα 1.2.: Τομή κατά μέτωπο στο ύψος του μεσοκοιλιακού τμήματος, της χοάνης και του αμυγδαλοειδούς σώματος

Σχήμα 1.3.: Τομή κατά μέτωπο του βασικού τμήματος του πρόσθιου εγκεφάλου στο ύψος του θαλάμου και του οπίσθιου άκρου του κελύφους.

Σχήμα 1.4.: Εγκάρσια τομή του εγκεφαλικού στελέχους στο ύψος του πυρήνα του απαγωγού νεύρου, της άνω ελαίας και του τραπεζοειδούς σώματος.

Σ' αυτό το κεφάλαιο λοιπόν, θα περιγραφούν οι αρχές, οι στόχοι καθώς και αποτελέσματα της έρευνάς μας για τη χαρτογράφηση εγκεφάλου. Όλα αυτά έχουν ακολουθήσει μία πορεία που ξεκίνησε από τη διερεύνηση μιας σχεσιακής βάσης δεδομένων για τον ανθρώπινο εγκέφαλο (τη Brain Data Base - BDB) με σκοπό την εξαγωγή συμπερασμάτων που αφορούσαν τον εντοπισμό επιληπτικών εστιών. Στη συνέχεια, τα συμπεράσματα χρησιμοποιήθηκαν για την επέκταση της βάσης και σε άλλα πιο εκφραστικά και μοντέρνα μοντέλα, όπως το Σημαντικό και το Αντικειμενοστραφές Μοντέλο, όπου προτάθηκαν συγκεκριμένες επεκτάσεις και αναθεωρήσεις, με πολλές εφαρμογές.

1.1.1. Αρχικός Σκοπός

Ο σκοπός της δημιουργίας μίας βάσης δεδομένων για τον ανθρώπινο εγκέφαλο ήταν ο εντοπισμός επιληπτικών εστιών σε συγκεκριμένες περιπτώσεις. Γενικά η ακριβής διάγνωση σχετικά με την επιληψία (στις αρχές της δεκαετίας του '90 αλλά ακόμα σήμερα) είναι ένα πολύ σύνθετο και πολλές φορές άλυτο πρόβλημα.

Σχήμα 1.5.: Παράδειγμα Μαγνητικής Τομογραφίας Εγκεφάλου (MRI)

Η επιληψία μπορεί να διακριθεί σε εστιακή, γενικευμένη και σε επιληψία με δευτερεύουσα γενίκευση (συνδυασμός των δύο πρώτων). Και οι τρεις τύποι επιληψίας έχουν μελετηθεί αξιοποιώντας τη σύγχρονη τεχνολογία στον τομέα των εξετάσεων (π.χ. αξονική, μαγνητική τομογραφία). Ένα από τα συμπεράσματα είναι πως μικρό ποσοστό του πρώτου τύπου επιδέχεται χειρουργική επέμβαση ενώ ο τελευταίος τύπος αφήνει πολλά περιθώρια επιπλέον μελέτης. Για παράδειγμα, έχει αποδειχτεί πως για ένα είδος επιληψίας με δευτερεύουσα γενίκευση, τη φωτοευαίσθητη επιληψία, είναι δυνατόν να συγκεντρωθεί τεράστια πληροφορία σχετικά με την πρωτεύουσα εστία της. Αυτό επιτυγχάνεται με την παρατήρηση μεταβολών και αντιδράσεων μετά από τη χρήση κάποιων συγκεκριμένων φαρμάκων και την παράλληλη χρήση του Βιομαγνητισμού (Σχήμα 1.5.) για τον εντοπισμό των ηλεκτρικά ενεργών πληθυσμών νευρώνων στο συγκεκριμένο σημείο του εγκεφάλου [A5, A6]. Πιο συγκεκριμένα, σε τελικό επίπεδο, αφού έχει μετρηθεί ταυτόχρονα (ή περίπου ταυτόχρονα) το μαγνητικό πεδίο σε έναν αριθμό σημείων, ζητείται η αρχική κατανομή του ρεύματος στα εγκεφαλικά σημεία. Αυτό συνθέτει το λεγόμενο 'αντίστροφο πρόβλημα' (inverse problem)¹ το οποίο δεν έχει μοναδική λύση και συνεπώς επιστρέφει σαν πιθανά σημεία επιληπτικής εστίας, παραπάνω από ένα. Κάτι τέτοιο δεν αποτελεί 'λύση' γιατί αποκλείει τη δυνατότητα στοχευμένης επέμβασης, πράγμα που ήταν και ο αρχικός σκοπός. Έτσι λοιπόν δεν είναι δυνατόν να βρεθεί η μία και μοναδική λύση κατευθείαν. Είναι δυνατόν όμως, με τη βοήθεια και τη χρήση σχετικής

¹ Αντίστοιχα το 'ευθύ πρόβλημα' (forward problem) υπολογίζει – μέσω των εξισώσεων Maxwell – το ηλεκτρικό δυναμικό και το μαγνητικό πεδίο σε εγκεφαλικά σημεία, με βάση την αναπαράσταση του εγκεφάλου με μαθηματικά μοντέλα.

πληροφορίας για τον εγκέφαλο (ανατομική κατανομή λειτουργιών, αντιδράσεις σε φάρμακα ανά πιθανή επιληπτική εστία, συμπτώματα, κλπ) να αποκλειστούν πολλές λύσεις – σημεία πιθανών επιληπτικών εστιών – γιατί τα υπόλοιπα συγκεκριμένα δεδομένα που τις αφορούν δεν συνάδουν με την επιληψία. Εδώ παρόλο που θεωρητικά ολοκληρώνεται η λύση του ‘αντίστροφου’ προβλήματος, πρακτικά – χωρίς τη χρήση τεχνολογίας - μένει ανέφικτη. Ο λόγος βρίσκεται στον τεράστιο όγκο και την πολυπλοκότητα της υπαρκτής σχετικής πληροφορίας. *Αυτό το γεγονός μπορεί να πάψει να είναι απαγορευτικό αν χρησιμοποιηθεί η κατάλληλη βάση δεδομένων.* Πρόκειται για ένα ακόμη παράδειγμα όπου η χρήση της τεχνολογίας δεν είναι πολυτέλεια αλλά ανάγκη.

Συνοπτικά λοιπόν, το πλαίσιο στο οποίο γεννήθηκε η ανάγκη της αρχικής μας έρευνας ήταν: *Χρησιμοποιώντας δεδομένα από βιομαγνητικές – και στην πορεία και από άλλες – εξετάσεις είναι δυνατόν να εντοπιστεί η πιθανή εστία συγκεκριμένων τύπων επιληψίας (π.χ. φωτοευαίσθητη επιληψία) με τη βοήθεια της οργάνωσης των υπερβολικά πολλών και πολύπλοκων ανατομικών, λειτουργικών, ηλεκτρικών και άλλων στοιχείων του εγκεφάλου σε μία βάση δεδομένων. Αυτή η βάση δεδομένων καλείται να απεικονίσει πολύπλοκα τρισδιάστατα με δικτυακή πολλές φορές δομή αντικείμενα και να τα οργανώσει με τρόπο αποδοτικό αλλά και ικανό να ανακαλυφθεί νέα σχετική γνώση.*

1.2. Brain Data Base (BDB)

[Σ2, Σ3, Σ4, Σ6]

Για τον σκοπό που περιγράφηκε παραπάνω, ορίστηκε αρχικά η σχεσιακή βάση δεδομένων BDB (Brain Data Base) με στόχο τη χαρτογράφηση του ανθρώπινου εγκεφάλου και πιο συγκεκριμένα για την οργάνωση των ανατομικών και λειτουργικών χαρακτηριστικών του. Το BDB αποτελεί ένα απλό λογικό μοντέλο, του οποίου η σημαντικότερη ίσως χρήση είναι το ότι μπορεί να αποτελέσει τη βάση και το αρχικό πλαίσιο για τον ορισμό ενός περισσότερο σύνθετου μοντέλου, όχι μόνο για την εφαρμογή μας, αλλά και για άλλες εφαρμογές που σχετίζονται με σύνθετα τοπολογικά δεδομένα.

Τα ιδιαίτερα σημεία του BDB αφορούν δύο θέματα:

- ο Την αποδοτική αναπαράσταση των πολύπλοκων τρισδιάστατων δικτυακής μορφής αντικειμένων (τμήματα εγκεφάλου)
- ο Την αποδοτική απεικόνιση, οργάνωση και αξιοποίηση των ιεραρχικών σχέσεων ανάμεσα στα αντικείμενα αυτά.

Στα πλαίσια του πρώτου θέματος, λόγω των γεωμετρικών ιδιοτεροτήτων, χρειάστηκε κατ’αρχήν να χωριστεί ο εγκέφαλος σε κύβους, τους ‘*βασικούς κύβους*’ (*basic cubes*), προκειμένου κάθε περιοχή του να αναπαρασταθεί με το σύνολο των βασικών κύβων από τους οποίους αποτελείται. Επίσης ορίστηκαν τα *ευέλικτα κουτιά* (*flexible boxes*) για την ομαδοποίηση των βασικών κύβων και προτάθηκαν τα *τμήματα δέσμης* (*fiber bundle*) για την ευκολότερη αναπαράσταση των πολύπλοκων δικτυακής δομής νευρικών ινών².

Το δεύτερο θέμα, χρειάστηκε αρκετή αρχική διερεύνηση και μελέτη γιατί δεν ήταν προφανής η ύπαρξη ιεραρχικών σχέσεων. Υπήρχε η παρατήρηση και η υποψία πως μπορεί να υπάρχουν και εάν συνέβαινε κάτι τέτοιο θα ήταν σημαντικό για την αποδοτικότητα του μοντέλου να απεικονιστούν αντιπροσωπευτικά. Τελικά βεβαιώθηκε πως η πιθανότητα εμφάνισης επιληπτικών εστιών είναι διερευνημένα διαφορετική ανάμεσα στα μέρη του εγκεφάλου.

---

² Η διαχείριση και των τριών τύπων δεδομένων επιτεύχθηκε με τη χρήση γνωστών [A7, A8, A9, A13, A18] αλλά και νέων μεθόδων Υπολογιστικής Γεωμετρίας [A21, A22, A23].

Δεδομένου ότι τέτοιου είδους ταξινομήσεις δεν μπορούν να εκφραστούν μέσω του Σχεσιακού Μοντέλου, προτείνουμε την έννοια της *ιεραρχικής ταξινόμησης*, η οποία αναπαρίσταται μέσω ενός απλού δενδρικού διαγράμματος [A10], όπου οι κόμβοι τοποθετούνται ανάλογα με την πιθανότητα εμφάνισης επιληπτικής εστίας ανά περιοχή (Σχήμα 1.6.). Για παράδειγμα, έστω ότι εφαρμόζεται ιεραρχική ταξινόμηση στις εμφανίσεις (instances) της σχέσης ‘Λοβοί-Τμήματα Λοβών’³, η οποία βασίζεται στο γεγονός ότι οι επιληπτικές εστίες μπορούν να εμφανίζονται σε κάθε λοβό με διαφορετική συχνότητα, ενώ κάθε λοβός χωρίζεται σε επιμέρους τμήματα με διαφορετικές συχνότητες παρουσίας επιληπτικής εστίας. Η διάταξη ανάμεσα στους λοβούς (ή στα τμήματά τους) είναι τέτοια ώστε ο λοβός (ή το τμήμα λοβού) με τη μεγαλύτερη συχνότητα να εμφανίζεται πρώτος στο διάγραμμα. Συνεπώς, η αναζήτηση επιληπτικής εστίας θα ξεκινήσει από την περιοχή όπου υπάρχει μεγαλύτερη πιθανότητα παρουσίας εστίας κερδίζοντας σημαντικό χρόνο αναζήτησης (συνολικά).

Σχήμα 1.6. : Παράδειγμα συμβολισμού ιεραρχικής διάταξης λοβών

Για την περισσότερο ευέλικτη ‘κίνηση’ σε μια τέτοια δομή, επεκτάθηκαν ειδικοί τελεστές ορισμένοι σε προηγούμενες μελέτες [A11] και προτάθηκαν νέοι. Συγκεκριμένα χρησιμοποιήθηκε ο τελεστής ‘under’⁴ και προτάθηκαν οι τελεστές ‘after(k)’, ‘before(k)’ και ‘close(k)’. Οι δύο επόμενοι τελεστές εφαρμόζονται στις περιπτώσεις όπου υπάρχει διάταξη ανάμεσα στα παιδιά ενός πατέρα. Ορίζοντας ως $ch(i)$, το i -στο παιδί σε μία τέτοια ταξινόμηση, δίνονται οι παρακάτω ορισμοί:

A) Τελεστής **after(k)**:

$$\forall (i,j) \in N: a=ch(i) \wedge b=ch(j) \wedge i=j+k \ (k \in N) \Rightarrow a \text{ after}(k) b = \text{TRUE},$$

Π.χ. τμήμα10 after(1) τμήμα9 = TRUE.

B) Τελεστής **before(k)**:

$$\forall (i,j) \in N: a=ch(i) \wedge b=ch(j) \wedge i=j-k \ (k \in N) \Rightarrow a \text{ before}(k) b = \text{TRUE}$$

Π.χ. τμήμα65 before(2) τμήμα67 = TRUE, ενώ τμήμα5 before(3) τμήμα10 = FALSE.

³ Ο εγκέφαλος χωρίζεται σε 5 λοβούς (μετωπιαίος, βρεγματικός, ινιακός, κροταφικός και νήσος του Reil), λευκή ουσία και βασικά γάγγλια.

⁴ Η έκφραση ‘c under d’ είναι αληθής όταν το ‘c’ είναι παιδί του ‘d’ στο διάγραμμα.

Και με τους δύο παραπάνω τελεστές ανιχνεύεται η απόσταση ανάμεσα σε κόμβους-αδέλφια και συγκεκριμένα η σχετική θέση τους.

Θεωρώντας ότι I είναι το σύνολο των εμφανίσεων (instances) σε μία δενδρική δομή ιεραρχικής ταξινόμησης, τότε ο τελεστής $close(k)$ ορίζεται αναδρομικά όπως παρακάτω:

Γ) Τελεστής **close(k)**:

$$\text{Av } (Q,P) \in I: (a \text{ under } P) \wedge (b \text{ first under } Q) \wedge ((Q \text{ after}(k) P = \text{true}) \vee (P \text{ close}(k) Q = \text{true}))$$

$$\Rightarrow a \text{ close}(k) b = \text{TRUE}.$$

Π.χ. $\text{tmhima65 close}(4) \text{tmhima94} = \text{TRUE}$, ενώ $\text{tmhima12 close}(2) \text{tmhima7} = \text{FALSE}$.

Ο τελεστής ‘ $\vee$ ’ πρέπει να εκτελεστεί από αριστερά προς τα δεξιά και στη βάση του ότι ο ‘ $\text{after}(k)$ ’ θα γίνει αληθής πρώτος. Η έκφραση ‘ $a \text{ first under } Q$ ’ σημαίνει ότι το ‘ a ’ είναι το πρώτο στη σειρά παιδί του ‘ Q ’ (κάτω από το Q).

1.3. 3D-IFO Μοντέλο

[Σ5]

Η BDB βασιζόμενη στο Σχηματικό Μοντέλο, με τα νέα στοιχεία που εισήγαγε, αποτέλεσε την αρχική προσέγγιση-πρότασή μας για τη χαρτογράφηση του εγκεφάλου. Τα δύο βασικά θέματα που αναδύθηκαν – δηλαδή οι ιδιαίτερες απαιτήσεις αναπαράστασης και διαχείρισης των σύνθετων τρισδιάστατων αντικειμένων που απαρτίζουν τον ανθρώπινο εγκέφαλο και οι πολλές – κυρίως ιεραρχικού χαρακτήρα - λειτουργικές συσχετίσεις που υπάρχουν μεταξύ των αντικειμένων αυτών – οδήγησαν στην αναζήτηση άλλων νεώτερων μοντέλων τα οποία προσφέρουν περισσότερα και σχετικά με τις απαιτήσεις μας εκφραστικά εργαλεία. Ένα τέτοιο μοντέλο κρίναμε πως είναι το Σηματικό Μοντέλο (Semantic Model) IFO των Abiteboul και Hull [A12], το οποίο δίνει ιδιαίτερη έμφαση στην αναπαράσταση και στη διαχείριση (μέσω αυτόματων μηχανισμών και constructors) συσχετίσεων που ενυπάρχουν ανάμεσα στα δεδομένα μιας βάσης δεδομένων. Βασιζόμενοι σ’ αυτό το μοντέλο, προτείναμε μία επέκτασή του, το *μοντέλο 3D-IFO*, για την αναπαράσταση-προτυποποίηση (model) των δικών μας ιδιαίτερων δεδομένων και των συσχετίσεών τους.

1.3.1. IFO – Model

Το Μοντέλο IFO αποτελείται από τύπους (object types) οι οποίοι ομαδοποιούνται με τη βοήθεια μηχανισμών κατασκευής (constructors) και fragments, τα οποία σε συνδυασμό με τις συσχετίσεις ISA^5 συνδυάζονται για τη δημιουργία σχημάτων IFO (IFO schemas). Η δική μας επέκταση εφόσον αφορά το είδος των δεδομένων και τις συσχετίσεις τους, εστιάζεται στους τύπους και τους constructors, τα οποία και θα παρουσιαστούν περιληπτικά στην παρακάτω υποενότητα.

1.3.1.1. Αντικείμενα IFO (IFO Objects)

Η βάση οποιουδήποτε σχήματος IFO είναι η αναπαράσταση των διάφορων δομών αντικειμένων, οι οποίες καλούνται τύποι (object types). Υπάρχουν τρεις *ατομικοί(atomic) τύποι* IFO αντικειμένων:

- ♦ Ο *εκτυπώσιμος τύπος (printable)*, ο οποίος αντιστοιχεί σε προκαθορισμένους τύπους που χρησιμοποιούνται ως βάση για τα δεδομένα εισόδου και εξόδου, πχ. STRING, INTEGER, BOOLEAN, PICTURE κλπ. Για λόγους απλότητας σε όλη την παρακάτω

⁵ Π.χ. η συσχέτιση ‘ $MAΘHTΗΣ ISA ΠΡΟΣΩΠΟ$ ’ δείχνει πως κάθε αντικείμενο τύπου $MAΘHTΗΣ$ είναι και τύπου $ΠΡΟΣΩΠΟ$, δηλαδή ο $MAΘHTΗΣ$ κληρονομεί τα χαρακτηριστικά και τις λειτουργίες του $ΠΡΟΣΩΠΟ$.

ανάλυση θα θεωρήσουμε έναν μόνο τύπο, τον τύπο STRING. Ο αντίστοιχος συμβολισμός είναι αυτός που φαίνεται στο Σχήμα 1. 7.(α) (ένα τετράγωνο).

- ♦ Ο *αφηρημένος τύπος (abstract)*, ο οποίος τυπικά αντιστοιχεί σε αντικείμενα που δεν έχουν καμία εσωτερική δομή (σχετική τουλάχιστον με την οπτική του σχεδιαστή της βάσης δεδομένων. Π.χ. ο τύπος ΠΡΟΣΩΠΟ (Σχήμα 1.7.(β)) αντιμετωπίζεται από το σχεδιαστή της βάσης σαν να δεν έχει δομή παρόλο που μπορεί να έχει πολλά πεδία (attributes) και υποτύπους. Μία εμφάνιση (instance) αυτού του τύπου είναι ένα πεπερασμένο σύνολο ανθρώπων. Πρέπει να αναφερθεί πως το πεδίο τιμών ενός αφηρημένου τύπου, δεν είναι εκτυπώσιμο αλλά ο χρήστης μπορεί να αναφερθεί σ' αυτές απ'ευθείας, ως υπαρκτές οντότητες (όπως αναφερόμαστε στους ανθρώπους) ή έμμεσα μέσω εκτυπώσιμων τύπων (π.χ. για τα ονόματα των ανθρώπων).
- ♦ Ο *ελεύθερος τύπος (free)*, ο οποίος αντιστοιχεί σε οντότητες που λαμβάνονται μέσω συσχετίσεων ISA. Π.χ. σε μία βάση δεδομένων μαθητών, υπάρχουν τα αντικείμενα ΠΡΟΣΩΠΟ, ΟΝΟΜΑ-ΠΡΟΣΩΠΟΥ και ΜΑΘΗΤΗΣ. Τυπικά το ΠΡΟΣΩΠΟ αντιστοιχεί σε αφηρημένο τύπο, το ΟΝΟΜΑ-ΠΡΟΣΩΠΟΥ σε εκτυπώσιμο και ο ΜΑΘΗΤΗΣ σε ελεύθερο.

Σχήμα 1. 7. : Οι τρεις ατομικοί τύποι IFO

Οι μηχανισμοί κατασκευής (constructors) μη ατομικών τύπων από τους οποίους προκύπτουν σύνθετα αντικείμενα, είναι οι εξής:

α) Η *ομαδοποίηση αντικειμένων της ίδιας δομής (το collection – συλλογή - στο Μοντέλο Format[A15])*. Π.χ. στο Σχήμα 1.8.(α) ο constructor που σχηματίζει την Τάξη χρησιμοποιείται για την ομαδοποίηση του συνόλου των μαθητών της τάξης και συμβολίζεται $\otimes$ [ΜΑΘΗΤΗΣ].

β) Η *ομαδοποίηση-ταξινόμηση τύπων για το σχηματισμό ενός νέου (το Καρτεσιανό τους γινόμενο, αντίστοιχο του aggregation στο Μοντέλο Format[A15])*. Π.χ. στο Σχήμα 1. 8.(β) ο Constructor ΜΗΧΑΝΟΚΙΝΗΤΗ ΒΑΡΚΑ είναι το διατεταγμένο ζεύγος ΚΟΥΒΟΥΚΛΙΟΥ και ΜΗΧΑΝΗΣ και συμβολίζεται $\oplus$ [ΜΗΧΑΝΟΚΙΝΗΤΗ ΒΑΡΚΑ].

Οι δύο constructors μπορούν να εφαρμοστούν αναδρομικά με οποιαδήποτε σειρά.

Σχήμα 1. 8.: Ομαδοποιημένοι (constructed) Τύποι Αντικειμένων

1.3.2. 3D-IFO Τοπολογικές Επεκτάσεις

Τα ιδιαίτερα τοπολογικά χαρακτηριστικά των τμημάτων του ανθρώπινου εγκεφάλου οδήγησαν και σ' αυτό το μοντέλο στην εισαγωγή νέων τύπων καθώς και τελεστών για τη διαχείρισή τους. Πιο συγκεκριμένα ορίστηκαν:

- 4 νέοι εκτυπώσιμοι τύποι (συμπληρώνοντας τους κλασσικούς, real, string κλπ) που αφορούν τους βασικούς κύβους, τα ευέλικτα κουτιά, περιοχές του εγκεφάλου καθώς και τα τμήματα δέσμης.
- Νέες κατηγορίες τελεστών:
  - Γεωμετρικοί Τελεστές (Geometric Operators)
  - Τελεστές Μέτρησης (Measurement Operators)
  - Τοπολογικοί Τελεστές (Topological Operators)
  - Τελεστές Ταξινόμησης (Sorting Operators), οι οποίοι θα παρουσιαστούν στη συνέχεια, μαζί με τους constructors γιατί βασίζονται σ' αυτούς.

1.3.3. 3D-IFO Επέκταση Μηχανισμών Κατασκευής (Constructor), Εισαγωγή Πεδίων (attributes) και Σχετικών Τελεστών

Τα μέρη του εγκεφάλου συσχετίζονται, συνδέονται και αλληλεπιδρούν με πολλούς τρόπους, η κωδικοποίηση των οποίων αποτελεί αναγκαία προϋπόθεση για την αποδοτικότητα της προτεινόμενης βάσης δεδομένων. Ένα παράδειγμα τέτοιας συσχέτισης αφορά τις σχετικές θέσεις των μερών του εγκεφάλου μεταξύ τους. Για παράδειγμα έστω ότι ισχύουν οι παρακάτω τοπολογικές συσχετίσεις:

- Η περιοχή Brodmann X κείται βορειοδυτικά του έλικα Y.
- Ο έλικας X διασχίζει τις περιοχές Brodmann A, B και Z (μ' αυτή τη σειρά).

Παρατηρώντας αυτές τις εκφράσεις, μπορεί κανείς εύκολα να παρατηρήσει πως δεν έχει σημασία μόνο η συσχέτισή των περιοχών ('κείται βορειοδυτικά', 'διασχίζει') αλλά και η σειρά με την οποία εκείνες εμπλέκονται στη συσχέτιση. Για παράδειγμα, η έκφραση 'Ο έλικας Y κείται βορειοδυτικά από την περιοχή Brodmann X' δίνει τελείως διαφορετική πληροφορία (την αντίθετη) από το πρώτο παράδειγμα. Ένα άλλο θέμα που πρέπει να ληφθεί υπ' όψιν σ' αυτό το σημείο και προκύπτει από τη διάταξη ανάμεσα στα συσχετιζόμενα μέρη στο δεύτερο παράδειγμα, έχει σχέση με το εάν η συσχέτιση μπορεί να είναι 'πολλά με πολλά' (many to many). Όπως είναι φανερό, η σειρά 'περιοχή Brodmann A, περιοχή Brodmann B και περιοχή Brodmann Z', υποδηλώνει τη σειρά με την οποία ο έλικας X τις διασχίζει. Προκειμένου λοιπόν να αναπαρασταθούν τέτοιου είδους τοπολογικές συσχετίσεις, εισάγουμε τον constructor 'ordering':

***Ορισμός:** Ο μηχανισμός κατασκευής (constructor) ordering, ορίζει ένα πεπερασμένο σύνολο διατεταγμένων υποσυνόλων αντικειμένων του ίδιου ή διαφορετικού τύπου. Σε οποιοδήποτε υποσύνολο, η διάταξη καθώς και ο αριθμός των αντικειμένων κάθε τύπου, είναι προκαθορισμένος. Επιπλέον ορίζεται και ο (οι) τύπος (τύποι) που υποδηλώνει (ουν) την ακριβή συσχέτιση των συσχετιζόμενων αντικειμένων – σε κάθε υποσύνολο. Ο ordering συμβολίζεται με ένα κόμβο (O) ο οποίος περιέχει το σύμβολο ($\uparrow$) και συνδέεται με απλές γραμμές με τους τύπους αντικειμένων που συνδέει. Κάθε γραμμή εμπλουτίζεται με παραμέτρους που υποδεικνύουν τη διάταξη ανάμεσα στα αντικείμενα ($x_1, \dots, x_2, y_1, \dots, y_w$ στο Σχήμα 1.9.) και στον αριθμό αντικειμένων κάθε τύπου σε κάθε υποσύνολο ((z), (w) στο Σχήμα 1.9.). Ο τύπος*

αντικειμένου που υποδηλώνει τη συσχέτιση των σχετιζόμενων αντικειμένων συμβολίζεται (Σχήμα 1.9.) με ένα διακεκομμένων γραμμών ορθογώνιο παραλληλόγραμμο, γιατί δεν υπάρχει πάντα.

Για καλύτερη κατανόηση της χρήσης του νέου constructor, δίνεται στο Σχήμα 1.10. το 3D-IFO Σχήμα που αντιστοιχεί στο πρώτο παράδειγμα. Από το Σχήμα 1.10. φαίνεται πως κάθε πρόταση που εκφράζεται μέσω της ‘Γειτνίασης’ περιλαμβάνει πρώτα (είναι το 1) μία και μόνο μία (είναι το (1)) πάνω από τις περιοχές Brodmann) περιοχή Brodmann η οποία κείται όπου υποδηλώνει η ‘Σχετική Θέση’ σε σχέση με (είναι το 2) ένα και μόνο ένα έλικα (είναι το (1)) πάνω από τους ‘Έλικες’).

Συμπερασματικά, μπορεί κανείς με ασφάλεια να πει πως ο constructor *ordering* οδηγεί στον ορισμό συσχετίσεων μεταξύ ‘οντοτήτων’ της βάσης, των οποίων τα μέρη που συμμετέχουν σ’ αυτή είναι διατεταγμένα με κάποια λογική.

Σχήμα 1. 9.: Τυπικός συμβολισμός του ordering

Ένα άλλο θέμα που ενδιαφέρει τις μοντέρνες βάσεις δεδομένων αφορά τις αλλαγές που συμβαίνουν σε ένα αντικείμενο στην πορεία του χρόνου, γεγονός που εισάγει την ανάγκη προσδιορισμού και χειρισμού της έννοιας ‘έκδοση’. Στο μοντέλο IFO κάθε έκδοση αντικειμένου είναι ένα ξεχωριστό αντικείμενο που αποτελεί μέρος μιας λίστας (λίστα διαφορετικών εκδόσεων). Το κριτήριο ή καλύτερα η βάση στην οποία στηρίζεται μία έκδοση, δεν είναι αναγκαστικά ο χρόνος. Μπορεί – ανάλογα με το είδος των δεδομένων – να είναι ένα οποιοδήποτε άλλο κριτήριο. Π.χ. σε πληροφορία του τύπου ‘Στην περιοχή Brodmann Y βρίσκεται η βασική εστία της ακοής’ μπορεί να υπάρξουν πολλές διαφορετικές γνώμες ανά επιστημονική ομάδα. Έτσι αν η παραπάνω πρόταση είναι αποτέλεσμα των ερευνών κάποιου Πανεπιστήμιου A, μπορεί ένα Πανεπιστήμιο B να έχει καταλήξει στο συμπέρασμα: ‘ Στον έλικα Z βρίσκεται η βασική εστία της ακοής ‘. Πολλές τέτοιες διαφορετικές απόψεις πάνω στο ίδιο θέμα μπορούν και πρέπει να συσχετιστούν μεταξύ τους με κάποιο τρόπο. Γι’ αυτό το λόγο, προτείναμε τον συνδυασμό του constructor ‘listing’ με τον νέο τύπο πεδίου (attribute) ‘based_on’, όπως φαίνεται στα παραδείγματα των Σχημάτων 1.11. και 1.12.

Σχήμα 1. 10: Παράδειγμα χρήσης του constructor ordering για πληροφορία του τύπου: ‘Η περιοχή Brodmann X κείται βορειοδυτικά του έλικα Y’.

Στο Σχήμα 1.11. παρουσιάζεται η δημιουργία ενός συνόλου σύνθετων αντικειμένων διατεταγμένων με χρονική σειρά. Στο Σχήμα 1.12., από την άλλη πλευρά, αντιπροσωπεύεται μία άλλου είδους πληροφορία, την οποία είχαμε αναλύσει στην προηγούμενη ενότητα (BDB). Συγκεκριμένα, πρόκειται για την πολύ σημαντική πληροφορία σχετικά με τις διαφορετικές πιθανότητες εμφάνισης επιληπτικής εστίας στις διάφορες περιοχές Brodmann. Με τη βοήθεια του listing και του αντίστοιχου based_on, οι περιοχές Brodmann διατάσσονται με βάση την πιθανότητα εμφάνισης επιληπτικής εστίας. Αυτή η διάταξη μπορεί να χρησιμοποιηθεί για την αναζήτηση στις εν λόγω περιοχές με κριτήριο αυτές τις πιθανότητες, κερδίζοντας έτσι χρόνο αναζήτησης.

Με δεδομένο ότι η πιθανότητα μπορεί να χρειαστεί να εκφραστεί όχι μόνο με έναν απλό αριθμό αλλά και με σύνθετες εκφράσεις – ανάλογα με την περίπτωση - προκύπτει η ανάγκη ορισμού του εκτυπώσιμου τύπου 'Distribution', με πεδίο τιμών του ένα σύνολο συναρτησιακών συμβόλων.

Σχήμα 1. 11: Εκφράζοντας την έννοια της χρονικής έκδοσης

Σχήμα 1. 12.: Listing βασισμένο στην πιθανότητα

Οι παραπάνω επεκτάσεις δημιουργούν την ανάγκη ορισμού σχετικών τελεστών, προκειμένου να υποστηριχτεί η αποδοτική χρήση τους. Εισάγονται λοιπόν οι *Τελεστές Ταξινόμησης (Sorting Operators)* (ενότητα 1.3.2.), οι οποίοι ορίζονται όπως φαίνεται παρακάτω:

α) *Previous* (αντικείμενο, όνομα_διάταξης, x , z), ο οποίος επιστρέφει:

i) Αν $z=0$: Το x -στο αντικείμενο πριν από το 'αντικείμενο', στο listing ή στο ordering με το όνομα 'όνομα-διάταξης'.

ii) Αν $z=1$: Όλα τα αντικείμενα από το 'αντικείμενο' μέχρι το x -στο αντικείμενο πριν απ' αυτό, στο listing ή στο ordering με το όνομα 'όνομα-διάταξης'.

β) *Next* (αντικείμενο, όνομα_διάταξης, x , z), ο οποίος ορίζεται ανάλογα με τον previous, αλλά για τα επόμενα αντικείμενα.

γ) *Basis_of*(τύπος αντικειμένου), ο οποίος επιστρέφει τον/τους τύπο/τύπους που αποτελεί/αποτελούν το κριτήριο(α) για το listing ή το ordering που ανήκει ο 'τύπος_αντικειμένου'.

Οι δύο παρακάτω τελεστές μπορούν να χρησιμοποιηθούν κυρίως για εκείνα τα listings που υποδηλώνουν εκδόσεις, γι' αυτό δανειζόμαστε αυτό τον όρο στον ορισμό τους.

δ) *According_to*(αντικείμενο, όνομα_διάταξης), επιστρέφοντας την έκδοση στην οποία ανήκει το 'αντικείμενο', στο listing με το όνομα 'όνομα_διάταξης'. Για παράδειγμα, στην περίπτωση του Σχήματος 1.11., μία πιθανή έκφραση είναι: *according_to*(x,λειτουργίες)=1979.

ε) *Same_according*(όνομα_διάταξης, αντικείμενο₁, αντικείμενο₂,...,αντικείμενο_N) το οποίο επιστρέφει 'αληθές' ή 'ψευδές', υποδηλώνοντας εάν τα αναφερόμενα αντικείμενα ('αντικείμενο₁', 'αντικείμενο₂',..., 'αντικείμενο_N') είναι της ίδιας έκδοσης ή όχι, στο listing με το όνομα 'listing_name'.

1.3.4. Παράδειγμα χρήσης

Ο λόγος της επέκτασης του μοντέλου IFO σε 3D-IFO με την εισαγωγή νέων τύπων και τελεστών, είναι η εξασφάλιση δυνατότητας ορισμού τοπολογικών αντικειμένων καθώς και η έκφραση τοπολογικών ελέγχων ακεραιότητας (integrity constraints) και ερωτημάτων (queries). Ένα ενδεικτικό ερώτημα με έμφαση σε ό,τι παρουσιάστηκε παραπάνω είναι η εξής: 'Βρες τις περιοχές Brodmann, οι οποίες εμπλέκονται στην εγκεφαλική λειτουργία 'ακοή', σύμφωνα με τις περισσότερες εκτιμήσεις/θεωρήσεις'. Η αντίστοιχη 3D-IFO έκφραση, χρησιμοποιώντας την προσέγγιση του predicate calculus είναι:

$$\text{Retrieve name}(x) \text{ where } (x \in E_{\text{Brodmann}} \wedge (\exists B: \text{Basis\_of}(E_{\text{function}})=B \wedge (Mv \in B, \exists y: y \in E_{\text{function}} \wedge \text{Name}(y)='ακοή' \wedge \text{according\_to}(y,B)=v \wedge \text{Cross}(x,y)=\text{true})))$$

όπου:

- E_{Brodmann} , η συλλογή των περιοχών Brodmann
- E_{function} , η συλλογή των εγκεφαλικών λειτουργιών ανά περιοχή
- Name , εκτυπώσιμος τύπος που αφορά το όνομα εγκεφαλικής λειτουργίας
- $\text{Cross}(x,y)$, ένας από τους 3D-IFO τελεστές. Επιστρέφει 'αλήθεια' ή 'ψέμα', ανάλογα με το εάν οι περιοχές που αντιστοιχούν στις παραμέτρους του, τέμνονται /διασταυρώνονται ή όχι.

1.4. Αντικειμενοστραφές Μοντέλο MITOS (Model for the Intelligent Support of Three-Dimensional Objects – Μοντέλο για την Ευφυή Υποστήριξη Τρισδιάστατων Αντικειμένων) [Σ9, Σ10]

Παρακολουθώντας την εξέλιξη στην ανάπτυξη νέων μοντέλων για την αποδοτικότερη οργάνωση δεδομένων, στραφήκαμε επίσης στην κατηγορία των αντικειμενοστραφών (object oriented) μοντέλων προκειμένου να αλληλεπιδράσουμε με τις δυνατότητές τους έτσι ώστε να βελτιώσουμε και να επεκτείνουμε τα μέχρι τώρα αποτελέσματά μας. Το αποτέλεσμα ήταν ένα αντικειμενοστραφές μοντέλο, το μοντέλο *MITOS* (*Model for the Intelligent Support of Three-Dimensional Objects*), το οποίο αποτελείται από το αντίστοιχο μοντέλο δεδομένων (*MITOS Data Model – MDM*) και μια νέα γλώσσα ερωτημάτων (*MITOS Query Language – MQL*). Το MITOS γενικά εισάγει καινοτομίες οι οποίες εξυπηρετούν μια περισσότερο αντιπροσωπευτική εκφραστική και έξυπνη αναπαράσταση και διαχείριση πολυδιάστατων δεδομένων και γνώσης. Αυτές οι καινοτομίες αφορούν θέματα όπως:

- Νέο Ορισμό Αντικειμένων (objects)
- Ορισμό Διατάξεων Αντικειμένων

- Νέο Ορισμό Κανόνων (rules)
- Νέους Τύπους Δεδομένων για την αναπαράσταση πολύπλοκων πολυδιάστατων αντικειμένων
- Νέες Μέθοδοι Αντικειμένων
- Νέα Κατηγορήματα (predicates) (Κυρίως Τοπολογικά)
- Νέους Τύπους Περιορισμών - Ελέγχων Ακεραιότητας (constraints)

Στην παρούσα διατριβή θα παρουσιαστεί ό,τι αφορά τα τρία πρώτα θέματα ενώ τα υπόλοιπα θα προσεγγιστούν μόνο από τις πλευρές που αλληλεπιδρούν με τα πρώτα.

1.4.1. Εισαγωγή Νέου Χαρακτηριστικού στην Έννοια του Αντικειμένου

Αν και το αντικείμενο αποτελεί το βασικό δομικό στοιχείο της αντικειμενοστραφούς προσέγγισης (object oriented) δεν υπάρχει ένας κοινά αποδεκτός ορισμός για αυτό. Κάθε μεθοδολογία δίνει το δικό της ορισμό. Έτσι κατά τον Jacobson [A24], αντικείμενο είναι μία οντότητα που χαρακτηρίζεται από την ικανότητα να διατηρεί μία κατάσταση (πληροφορία) και προσφέρει ένα σύνολο λειτουργιών (συμπεριφορά) που παρέχουν τη δυνατότητα αφ' ενός μεν εξέτασης της κατάστασης αφ' ετέρου δε επίδρασης πάνω σ' αυτή. Δηλαδή, ένα αντικείμενο χαρακτηρίζεται από ένα σύνολο λειτουργιών και μια κατάσταση που θυμάται τα αποτελέσματα των λειτουργιών.

Κατά κάποια άλλη εκδοχή, το αντικείμενο είναι μια οντότητα της οποίας η συμπεριφορά χαρακτηρίζεται από τις ενέργειες που προσφέρει και που απαιτεί από τα άλλα αντικείμενα.

Κατά τον Booch [A20] πάλι – ο οποίος δίνει έναν πιο τυπικό ορισμό – ένα αντικείμενο έχει κατάσταση, συμπεριφορά και ταυτότητα. Πιο συγκεκριμένα, η δομή και η συμπεριφορά όμοιων αντικειμένων ορίζεται από μία κοινή κλάση, ενώ κάθε αντικείμενο χαρακτηρίζεται από τα παρακάτω:

1. *Ταυτότητα*: είναι εκείνη η ιδιότητα του αντικειμένου που το διακρίνει από όλα τα άλλα αντικείμενα.
2. *Συμπεριφορά*: είναι το πώς ένα αντικείμενο δρα και ενεργεί σε συνάρτηση των ανταλλασσόμενων μηνυμάτων και των αλλαγών καταστάσεων. Η συμπεριφορά του αντικειμένου ορίζεται πλήρως από τις αντιδράσεις του στα μηνύματα που αυτό δέχεται. Η συμπεριφορά αυτή επηρεάζεται από τη σειρά με την οποία κάποιος ενεργεί πάνω στο αντικείμενο.
3. *Κατάσταση*: η κατάσταση ενός αντικειμένου περικλείει όλες τις ιδιότητες του αντικειμένου και τις τρέχουσες τιμές κάθε ιδιότητας, είναι δε η αιτία για την εξαρτώμενη από το χρόνο συμπεριφορά του.

Η δυνατότητα αυτής της ποικιλίας υφίσταται γιατί υπάρχει μεγάλη ευελιξία στη χρήση των αντικειμένων. Πιο συγκεκριμένα, υπάρχει μία πολύ χρήσιμη σταθερά στη χρήση τους, πράγμα που σημαίνει πως ακόμη κι αν αλλάξει ο ορισμός και συνεπώς και η υλοποίηση, δεν είμαστε αναγκασμένοι να αλλάξουμε τον τρόπο χρήσης των αντικειμένων. Έτσι ανάλογα με τις ανάγκες που αναδεικνύονται, μπορούμε να εμβαθύνουμε και να επεκτείνουμε τον ορισμό του αντικειμένου για την αποδοτικότερη έκφρασή του χωρίς να διαταράξουμε τη χρήση του.

Το γεγονός ότι η έρευνα μας είχε στον πυρήνα της τα *τοπολογικά χαρακτηριστικά αντικειμένων και τις συσχετίσεις τους καθώς και την ανάγκη της μεταξύ τους διάταξης-ιεράρχησης*, οδήγησε στην παρακάτω διαπίστωση:

Η σχέση ενός αντικειμένου με το περιβάλλον του (είτε ως θέση είτε ως συνάφεια) αποτελεί ένα κομβικό χαρακτηριστικό των αντικειμένων, το οποίο εάν – με τα υπάρχοντα δεδομένα - εκφραστεί είτε ως κατάσταση είτε ως συμπεριφορά, αποδυναμώνεται εκφραστικά.

Με αυτό το σκεπτικό ορίσαμε ως MITOS αντικείμενο την αδιαίρετη ενότητα με τα τέσσερα παρακάτω χαρακτηριστικά:

1. *Ταυτότητα* (κατά Booch)
2. *Συμπεριφορά* (κατά Booch)
3. *Κατάσταση* (κατά Booch)
4. *Σχέση με το περιβάλλον*, ορισμένο όπως παραπάνω.

1.4.1.1. Εισαγωγή του Χαρακτηριστικού στην Τυπική Έκφραση της Κλάσης

Σε κάθε αντικειμενοστραφές μοντέλο υπάρχει μία συγκεκριμένη τυποποίηση του ορισμού της κλάσης αντικειμένου. Εδώ θα παρουσιαστεί ο ορισμός της ‘Σχέσης με το περιβάλλον’ στο μοντέλο MITOS, ο οποίος μπορεί να εύκολα να προσαρμοστεί και σε τυποποιήσεις άλλων ανάλογων – αντικειμενοστραφών – μοντέλων. Πιο συγκεκριμένα, στο μοντέλο MITOS η σχέση ενός αντικειμένου με το περιβάλλον του μοντελάρεται με τον ορισμό ‘Correlation’ (Σχήμα 1.13.). Κάθε συσχέτιση έχει ένα μοναδικό όνομα για κάθε κλάση και συσχετίζει την κλάση με μία ή περισσότερες κλάσεις. Η ακριβής συσχέτιση ανάμεσά τους (αν υπάρχει) ορίζεται από τα ‘correlation attributes’ (πεδία συσχέτισης). Ένα αντικείμενο μιας ορισμένης κλάσης μπορεί να συσχετιστεί με ένα ή περισσότερα αντικείμενα κάθε σχετιζόμενης κλάσης (κατηγορήμα Multi), ενώ κάθε πεδίο συσχέτισης μπορεί να έχει τα κατηγορήματα: Multi, Common και Value⁶. Επιπλέον, είναι δυνατόν να οριστούν κατηγορήματα περιορισμού (constraint predicate) και/ή το κατηγορήμα ‘defined’, το οποίο υποδηλώνει ότι οι τιμές που αντιστοιχούν στη συσχέτιση αποθηκεύονται στα αντίστοιχα αντικείμενα μίας από τις εμπλεκόμενες/συσχετιζόμενες κλάσεις. Ως εκ τούτου, τα δεδομένα ορίζονται μόνο στα αντικείμενα της μίας κλάσης, ενώ στις υπόλοιπες συσχετιζόμενες κλάσεις υπάρχει μόνο το κατηγορήμα ‘defined’ στην ανάλογη θέση.

Ένα σημαντικό είδος συσχέτισης αναφέρεται στις ποικίλες τοπολογικές συσχετίσεις ανάμεσα σε πολυδιάστατα (τρισδιάστατα στην περίπτωσή μας) αντικείμενα. Οι σημαντικότερες είναι η ‘γειτνίαση’ και η ‘τομή’ από τις οποίες μπορεί να εξαχθεί πολύ χρήσιμη πληροφορία. Οι τοπολογικές συσχετίσεις μπορούν βέβαια να εξαχθούν με βάση τις θέσεις των αντικειμένων, η απαιτούμενη όμως επεξεργασία είναι πολύ μεγάλη· αντίθετα η ποσότητα της πληροφορίας που χρειάζεται για την περιγραφή/αποθήκευση της τοπολογίας είναι πολύ μικρότερη.

Correlation: Correlation_name_1 **with** Class_name_1 [**Multi**], ..
 [Corel_attribute-1:Type-1,
   [**Multi**
     Common
     Value: Value]
   ...]
   Defined
   Constraint_Predicates: procedure {param1,...,param_k}]
 .....

Σχήμα 1. 13. : Ορισμός Σχέσης με το Περιβάλλον

Για να πάρουμε μία ιδέα αυτών των τοπολογικών συσχετίσεων που μπορούν να οριστούν, ας δούμε ένα παράδειγμα. Έστω η κλάση ‘Δέσμη Νεύρων’, τα αντικείμενα της οποίας συσχετίζονται τοπολογικά με την κλάση ‘Λοβοί’ με τον παρακάτω τρόπο: Κάθε δέσμη νευρών διαπερνά κάποιους λοβούς, γεγονός που επηρεάζει τη λειτουργία τους καθώς και τις αντίστοιχες εγκεφαλικές λειτουργίες. Αυτή η σχέση των δεσμών νευρών με το ‘περιβάλλον’

⁶ Multi: Το πεδίο μπορεί να έχει πολλές τιμές.

Common: Το πεδίο είναι κοινό σε όλες τις εμφανίσεις (instances) στην κλάση.

Value: Ορίζεται μία εξ’ ορισμού τιμή.

τους μπορεί να εκφραστεί στο μοντέλο MITOS στον ορισμό της κλάσης ‘Δέσμη Νευρών’, ως εξής:

Correlation: Διάσχιση **with** ‘Λοβοί [Multi]

[Είδος: String, (πχ. με τιμή : Βορειοδυτικά ή Διαγώνια από Βορρά κοκ)
[Multi]

ενώ ο αντίστοιχος ορισμός στην κλάση ‘Λοβοί’ θα είναι:

Correlation: Διάσχιση **with** ‘Δέσμη Νευρών’ [Multi]

Defined (γιατί οι σχετικοί ορισμοί έχουν γίνει στην κλάση ‘Δέσμη Νευρών’)

1.4.2. Εισαγωγή της Διάταξης και της Έκδοσης στον Ορισμό της Κλάσης

Στα μοντέλα των προηγούμενων εννοιών έγινε η εισαγωγή της *διάταξης* και κατ’ επέκταση της έννοιας της *έκδοσης*, με την πρόταση συγκεκριμένων δομών και εργαλείων για την υποστήριξή τους. Αυτές οι έννοιες προτυποποιήθηκαν (model) και στο MITOS, πράγμα που επιτεύχθηκε με την εισαγωγή των αντίστοιχων ορισμών σε επίπεδο *ορισμού αντικειμένων* αλλά και σε *επίπεδο μεθόδων* (για την ανάλογη διαχείριση διάταξης/έκδοσης).

1.4.2.1. Ορισμός Διάταξης και Έκδοσης

Τα αντικειμενοστραφή μοντέλα υποστηρίζουν μεν την έννοια της ιεραρχίας μεταξύ των κλάσεων αλλά όχι της ιεραρχικής διάταξης ανάμεσα στα αντικείμενα μιας κλάσης, πράγμα που θα βελτίωνε τη διαχείρισή τους [Σ2]. Στο Σχήμα 1.14. αποτυπώνεται μία τέτοια διάταξη, η οποία υποδεικνύεται με βέλη από το ένα αντικείμενο στο άλλο μέσα στη διάταξη, όπου κάθε αντικείμενο έχει μια συγκεκριμένη θέση.

Σχήμα 1. 14. : Γενικό παράδειγμα διάταξης αντικειμένων κλάσης

Τέτοιου είδους ιεραρχικές διατάξεις ορίζονται όπως φαίνεται παρακάτω (Σχήμα 1.15.), όπου το πεδίο *Order_attribute* υποδηλώνει το κριτήριο στο οποίο βασίζεται η διάταξη και είναι δυνατόν να είναι πολλαπλό. Για κάθε τέτοιο πεδίο είναι επίσης δυνατός ο ορισμός προσαρτημένων κατηγορημάτων (attached predicates), όπως τα *Value* και τα *Constraint Predicates* (κατηγορήματα περιορισμού).

Η έννοια της έκδοσης εκφράζει συγκεκριμένα δυναμικά χαρακτηριστικά των αντικειμένων τα οποία υποδηλώνουν την ιστορία ή/και την άποψη της διαφορετικής οπτικής-θεώρησης. Το πεδίο *Version_attribute* υποδηλώνει το κριτήριο της δημιουργίας της έκδοσης και είναι δυνατόν να είναι πολλαπλό. (Π.χ. κάποια κλάση μπορεί να προκύπτει κατά τη θεώρηση κάποιου επιστήμονα σε συγκεκριμένη έκδοση κατά έτος). Οποσδήποτε μπορούν να οριστούν και προσαρτημένα κατηγορήματα (attached predicates), όπως τα *Value* και τα *Constraint Predicates* (κατηγορήματα περιορισμού). Οι εμφανίσεις αυτής της κλάσης έχουν μία αντίστοιχη τιμή για κάθε ‘version attribute’. Κάποιες φορές η ύπαρξη εκδόσεων διαφοροποιεί τα αντικείμενα μιας κλάσης δημιουργώντας υποσύνολα με διαφορετικό είδος και πλήθος

πληροφορίας (λόγω των αναγκών της εκάστοτε έκδοσης). Αυτό υποστηρίζεται με τη δημιουργία νέων υποκλάσεων.

Order: Order_attribute_1: Type_1
           [Value: Value
           Constraint Predicates: procedure {param_1,...,param_k}]

Order_attribute_2: Type_2
           [Value: Value
           Constraint Predicates: procedure {param_1,...,param_k}]

...

Version: Version_attribute_1: Type_1
           [Value: Value
           Constraint Predicates: procedure {param_1,...,param_k}]

Version_attribute_1: Type_2
           [Value: Value
           Constraint Predicates: procedure {param_1,...,param_k}]

....

Σχήμα 1. 15. : Ο τυπικός ορισμός της ιεραρχικής διάταξης και της έκδοσης αντικειμένων κλάσης

1.4.2.2. Ορισμός Ενσωματωμένων (Built in) Μεθόδων Διαχείρισης Διάταξης και Έκδοσης

Οι αντίστοιχοι τελεστές του 3D-IFO, μεταφέρθηκαν ως ενσωματωμένες μέθοδοι στο αντικειμενοστραφές μοντέλο MITOS και αποτυπώνονται στον παρακάτω πίνακα:

Μέθοδος Διάταξης / Έκδοσης	Περιγραφή
Previous (object, order_attribute, x, z)	Αν z=0, τότε επιστρέφεται το x-στο αντικείμενο πριν από το 'object', στη διάταξη η οποία βασίζεται στο 'order_attribute'. Αν z=1, τότε επιστρέφονται όλα τα αντικείμενα που μεσολαβούν ανάμεσα στο 'object' και το x-στο αντικείμενο πριν από αυτό, στη διάταξη η οποία βασίζεται στο 'order_attribute'.
Next (object, order_attribute, x, z)	Αν z=0, τότε επιστρέφεται το x-στο αντικείμενο μετά από το 'object', στη διάταξη η οποία βασίζεται στο 'order_attribute'. Αν z=1, τότε επιστρέφονται όλα τα αντικείμενα που μεσολαβούν ανάμεσα στο 'object' και το x-στο αντικείμενο μετά από αυτό, στη διάταξη η οποία βασίζεται στο 'order_attribute'.
Basis_of (class_name)	Επιστρέφει το/τα πεδία διάταξης (order

	attributes) το/τα οποία αποτελούν κριτήρια για τη διάταξη/τις διατάξεις στην κλάση 'class_name'.
Version (class_name)	Επιστρέφει το/τα πεδίο/α έκδοσης που υποδεικνύουν τη βάση κάθε έκδοσης στην κλάσης 'class_name'.
According_to (object, version_attribute)	Επιστρέφει την έκδοση στην οποία ανήκει το 'object', σύμφωνα με το 'version_attribute'.

Πίνακας 1.1.: Ορισμός Ενσωματωμένων (Built in) Μεθόδων Διαχείρισης Διάταξης / Έκδοσης

1.4.3. Νέος ορισμός των Κανόνων

Οι κανόνες αποτελούν τη βασική έννοια-εργαλείο στη γλώσσα MITOS (MQL). Η τυπική τους μορφή είναι: $P(x_1, \dots, x_n) \leftarrow F$, όπου το P (κεφαλή) είναι εκείνο το κατηγορημα (predicate) που βάση των τοπικών του μεταβλητών, πρέπει να αποδειχτεί.

Όπως έχει αναφερθεί και προηγούμενα, στο μοντέλο MITOS τα θέματα βάσεων δεδομένων και βάσεων γνώσης, μελετώνται και αναπαριστώνται – σε ένα βαθμό – σε κοινή βάση. Αυτό δεν οφείλεται μόνο σε μία τάση ολοκλήρωσης και ομοιογένειας αλλά και στην πεποίθησή μας πως τα αποτελέσματα κάθε περιοχής μπορούν να εμπνεύσουν και να στηρίξουν βελτιώσεις άλλων περιοχών. Έτσι, λαμβάνοντας υπ' όψιν τα αποτελέσματα από το χώρο των βάσεων δεδομένων, όπως αυτά παρουσιάστηκαν στην προηγούμενη υποενότητα αλλά και γενικότερα, εκτιμούμε πως είναι πολύ αποδοτική και αυξάνει την εκφραστική δυνατότητα – άρα και την αποτελεσματικότητα – των κανόνων, η εισαγωγή των παρακάτω:

- i. *Ταξινόμηση* (classification) των κανόνων. Οι κανόνες μπορούν να ταξινομηθούν σε κλάσεις.
- ii. *Ορισμός εκδόσεων* των κανόνων
- iii. Ορισμός *περιορισμών ακεραιότητας* (integrity constraints), για τον έλεγχο της εγκυρότητας των κανόνων καθώς και για την αποφυγή αντιφάσεων (rule conflicts).
- iv. Ορισμός ενός είδους *κλειδιού κανόνα*. Πιο συγκεκριμένα, διερευνώντας 'τι θα μπορούσε να παίζει το ρόλο κλειδιού για ένα κανόνα', καταλήξαμε στο συμπέρασμα πως κανένα μέρος ενός κανόνα δεν μπορεί να παίζει αυτό το ρόλο. Παρόλα αυτά όμως, βασισμένοι σε μια ευρεία αντίληψη και θεώρηση της έννοιας του κλειδιού, κάναμε τη παρακάτω σκέψη: 'Ένας από τους βασικότερους λόγους ορισμού και χρήσης του κλειδιού στις βάσεις δεδομένων, είναι η αποφυγή προσπέλασης κάθε χαρακτηριστικού των αντικειμένων κατά τους ελέγχους των βάσεων δεδομένων και συνεπώς η μείωση του συνολικού αριθμού των εκτελούμενων ελέγχων. Σε μία βάση γνώσης (knowledge base), μια τέτοια μείωση μπορεί να επιτευχθεί εάν τα μέρη του κανόνα που βοηθούν στην ευκολότερη ματαίωσή του (abort), εμφανίζονται πρώτα στην έκφρασή του, λειτουργώντας έτσι ως ένα είδος κλειδιού – εφόσον οδηγεί σε μείωση των ελέγχων'. Καλούμε αυτό το τμήμα του κανόνα, 'σχετικό κλειδί' του κανόνα.

Για την υποστήριξη αυτών των δυνατοτήτων προτείνουμε τους ορισμούς:

- α. **Κεφαλίδα Κανόνα (Rule heading):** Εδώ ορίζονται τα πρόσθετα χαρακτηριστικά του κανόνα (Σχήμα 1. 16.).

Classification:	Όνομα Κλάσης
Version:	Όνομα Έκδοσης
Key_Order:	Τιμή (Αληθής / Ψευδής)

Σχήμα 1. 16. : Τυπικός Ορισμός Κεφαλίδας Κανόνα

- i. Το κατηγορημα Classification ορίζει την κλάση στην οποία ανήκει ο κανόνας.
- ii. Το κατηγορημα Version ορίζει την κλάση του (αν υπάρχει).
- iii. Το Key-Order, εκφράζει εάν η σειρά των τμημάτων του κανόνα (formulae) εκφράζει την ύπαρξη 'σχετικού κλειδιού'.

β. **Περιορισμοί Ακεραιότητας Κανόνα**, οι οποίοι μπορεί να είναι:

- a. *Έλεγχος βασισμένος στην Άλγεβρα Boole*, για την αποφυγή ορισμού συγκρουόμενων κανόνων (π.χ. ' $c \leftarrow a \wedge b$ ' και ' $c \leftarrow a \wedge (\text{not } b)$ '). Επίσης στην περίπτωση συγκρούσεων κανόνων όπως ' $c \leftarrow a \wedge b$ ' και ' $c \leftarrow \text{not } (\text{not}(a) \wedge \text{not } (b))$ ' (το οποίο είναι ισοδύναμο με ' $c \leftarrow a \vee b$ '), το σύστημα μπορεί να οριστεί :
  - i. *Αυστηρό*, οπότε θα δέχεται τον πρώτο κανόνα
  - ii. *Χαλαρό* και θα δέχεται τον δεύτερο.
- b. *Έλεγχος Σχετικού Κλειδιού*, για να γίνεται έλεγχος εάν δύο key-ordered κανόνες είναι ίδιοι εκτός από τη σειρά των μερών (formulae) του κανόνα, οπότε και να επιλέγεται από τον προγραμματιστή ο ένας από αυτούς, ως ορθός.
- c. *Έλεγχος Ορισμένος από τον προγραμματιστή ή χρήστη*, για τον ορισμό των 'έξωθεν' απαιτούμενων περιορισμών.

1.5. Ανακεφαλαίωση - Συνεισφορά

Ο αρχικός φιλόδοξος στόχος της συνολικής έρευνας που περιγράφεται και αναλύεται στην παρούσα διατριβή ήταν η *οργάνωση και διαχείριση δεδομένων και γνώσης που αφορούν τον ανθρώπινο εγκέφαλο* και είναι ικανά - σε συνεργασία με τα δεδομένα του Βιομαγνητισμού - να οδηγήσουν στον εντοπισμό επιληπτικών εστιών. Οι ανάγκες μιας τέτοιας χαρτογράφησης έθεσαν θέματα *αναπαράστασης πολύπλοκων τρισδιάστατων και δικτυακών αντικειμένων με συγκεκριμένες λογικές διατάξεις και ιεραρχίες μεταξύ τους*. Αυτά τα ζητήματα μελετήθηκαν σε ένα εύρος τύπων μοντέλων βάσεων δεδομένων προκειμένου ακολουθώντας τη χρονική εξέλιξη των μοντέλων να αξιοποιηθεί από καθένα τους τα κατάλληλα στοιχεία. Έτσι τελικά κινηθήκαμε ανάμεσα στο πολύ διαδεδομένο σχεσιακό μοντέλο το οποίο προσφέρει απλότητα και αρκετή εννοιολογική εκφραστική δύναμη, στο Σημαντικό μοντέλο IFO [A12] με τα ευέλικτα εργαλεία αναπαράστασης και στη σύγχρονη, (ειδικά για τη δεκαετία του '90) συμπαγούς και πλούσιας εκφραστικής ισχύος, αντικειμενοστραφή λογική.

Το αποτέλεσμα αυτής της εκτεταμένης και πολύπλευρης έρευνας ήταν ο ορισμός νέων τύπων δεδομένων για την αναπαράσταση των ιδιαίτερων εγκεφαλικών δεδομένων καθώς και η εισαγωγή νέων τελεστών, εννοιών και δομικών στοιχείων για τη διαχείριση των λογικών διατάξεων και των ιεραρχιών τους. Αυτό πραγματοποιήθηκε διαβαθμισμένα από μοντέλο σε μοντέλο, όπου οι ορισμοί σε καθένα από αυτά βασίζονταν στους αντίστοιχους του προηγούμενου, αποτελώντας επέκτασή τους, ανάλογα με τις δυνατότητες του μοντέλου. Έτσι ορίστηκε η σχεσιακή βάση δεδομένων *BDB (Brain Data Base)* [Σ2, Σ3, Σ4, Σ6], η επέκταση

του μοντέλου IFO, το *3D-IFO* [Σ5] καθώς και το αντικειμενοστραφές μοντέλο *MITOS* [Σ9, Σ10].

Η συνεισφορά της συγκεκριμένης έρευνας και πρότασης μοντέλων βάσεων δεδομένων, έγκειται στις τεράστιες δυνατότητες εφαρμογής και επέκτασης που διαθέτει. Αυτό κυρίως οφείλεται στην ιδιαίτερη πολυπλοκότητα των δεδομένων και της γνώσης που αφορούν τον ανθρώπινο εγκέφαλο - το πιο σύνθετο ανθρώπινο 'σύμπαν'. Πιο συγκεκριμένα, ο ορισμός των κατάλληλων τύπων δεδομένων και εργαλείων γνώσης για την οργάνωση και διαχείριση τόσο σύνθετων δεδομένων και γνώσης, μπορεί να επεκταθεί και σε άλλους ενδιαφέροντες χώρους με ανάλογα χαρακτηριστικά (π.χ. γεωγραφικά συστήματα, άλλες ιατρικές βάσεις δεδομένων) ή να αποτελέσει μαγιά για νέα πιο εξελιγμένα μοντέλα.

Εδώ ας σημειωθεί πως η έρευνα της βιβλιογραφίας [A8, A11, A12, A15, A17, A20, A25-A29] σχετικά με τον ορισμό και τις επεκτάσεις ειδικών ή γενικών μοντέλων δεδομένων, δείχνει πως συνήθως η προσφορά κάθε νέας πρότασης είναι η εξής: Εξυπηρετεί τις συγκεκριμένες ανάγκες για τις οποίες ορίστηκε, αποτελώντας κάποιες φορές τον πυρήνα δημιουργίας άλλων σχετικών μοντέλων. Αυτό όμως που από επιστημονικής/ερευνητικής πλευράς είναι το σημαντικότερο, είναι πως σχεδόν πάντα πυροδοτεί νέες ιδέες οι οποίες συνθέτουν κάτι εντελώς καινούργιο.

1.6. Δημοσιεύσεις του κεφαλαίου

♦ ΣΥΝΕΛΠΙΑ

- [Σ1] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "Mapping the brain for the localization of epileptic foci", Satellite Symposium on Neuroscience and Technology 14th International Conference IEEE, Lyon, France, 1992.
- [Σ2] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "BDB (Brain Data Base), DEXA 93 4th International Conference on Database and Expert Systems Applications, Prague, Czech Republic.
- [Σ3] Γ. Ανωγειανάκης, Α. Κροτοπούλου, Π. Σπυράκης, Δ. Τέρπου, Α. Τσακαλίδης, «Ένα Πληροφοριακό Σύστημα για τη Χαρτογράφηση του Ανθρώπινου Εγκεφάλου», Ε.Π.Υ., 4ο Πανελλήνιο Συνέδριο Πληροφορικής ΕΠΥ, 16-18 Δεκεμβρίου 1993, Πάτρα.
- [Σ4] Γ. Ανωγειανάκης, Α. Κροτοπούλου, Π. Σπυράκης, Δ. Τέρπου, Α. Τσακαλίδης, «BDB Σύστημα: Ένα Σύστημα Χαρτογράφησης του Ανθρώπινου Εγκεφάλου για τον εντοπισμό των επιληπτικών εστιών», 3ο Πανελλήνιο Συνέδριο Ιατρικής Πληροφορικής, 13-15 Μαΐου, Θεσσαλονίκη.
- [Σ5] A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "A Conceptual DataBase Approach for Modelling 3D Objects of Irregular Geometry", DEXA 94, 5th International Conference on Database and Expert Systems Applications, Athens, Greece, September 1994.
- [Σ6] G. Anogianakis, Krotopoulou, G. Nendidis, P. Spirakis, D. Terpou, A. Tsakalidis, M. Psaltikidou. MAGNOBRAIN II: The Brain Database (BDB). 8th Congress of the International Society of the Greek Neuroscientists, Rhodes, September 1994.
- [Σ7] G. Anogianakis, M. Apostolakis, S. Bountzioukas, K. Krotopoulou, G. Nendidis, P. Spirakis, D. Terpou, A. Tsakalidis. Mapping the Brain in a structural way helps in Localizing Brain Function. Topographic Representation of the Visual Cortex Symposium, Porto Carras, 1994.
- [Σ8] G. Anogianakis, M. Apostolakis, G.F.A Harding, K. Krotopoulou, G. Nendidis, M. Psaltikidou, P. Spirakis, D. Terpou, A. Tsakalidis, 1995. MAGNOBRAIN: Integrating Radiological, Anatomical and Electrophysiological Information about the Brain for Source Localization. In Interactive Technology and the New Paradigm for Healthcare. Washington, DC:IOS Press. pp. 4-14.
- [Σ9] Krotopoulou A., Spirakis P., Terpou D, Tsakalidis A., «Mitos: A Model for the Intelligent Support of Three-Dimensional Objects», Joint European Conference and Exhibition on Geographical Information, (JEC'97), 16-18 April 1997, Vienna, Austria.
- [Σ10] Krotopoulou A., Spirakis P., Terpou D, Tsakalidis A., "An Object Oriented Data Model for the Intelligent Support of Three-Dimensional Objects", DEXA 97, 8th International Conference on Database and Expert Systems Applications, 1-5 September 1997, Toulouse, France.

♦ ΠΕΡΙΟΔΙΚΑ

- [Π1] G.Anogianakis, M.Apostolakis, S.Bountzioukas, K.Krotopoulou, G.Nendides, P. Spirakis, D. Terpou, A. Tsakalidis, "*Mapping the Brain in a Structural Way Helps in Localizing Brain-Function*", International Journal of Psychophysiology, Vol. 18, Iss 2, pp.91-91 1994 (36th of a total of 482 Journals in its category).

♦ ΤΕΧΝΙΚΕΣ ΑΝΑΦΟΡΕΣ

- [1] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "The logical design of the Brain Data Base", Technical Report MB1.1., April 1992.
- [2] G. Anogianakis, A. Krotopoulou, P. Spirakis, D. Terpou, A. Tsakalidis, "Design and Implementation issues of the Brain Data Base (BDB)", Technical Report MB1.2, September 1992.

1.7. Αναφορές σε Δημοσιεύσεις μας

- [C1] Αναφορά στη δημοσίευση [Σ2]: Glaucia Faria, Claudia Bauzer Medeiros, Mario A. Nascimento, "An Extensible Framework for SpatioTemporal Database Applications", in: Proc. of 10th Int. Conf. on Scientific and Statistical Database Management (SSDBM), IEEE Computer Society, Capri, Italy, 1998, pp. 202--205.
- [C2] Αναφορά στις τεχνικές αναφορές [T1] και [T2]: Georgia Panagopoulou, Spiros Sirmakessis, John Tzimas, "Efficient Storage Compression for 3D Regions", in the Proceedings of the IEEE Data Compression Conference (DCC '97), pp 463, IEEE Computer Society Press, ISBN 0 8186-7761 9, 25-27 March 1997, Snowbird, Utah, US.

1.8. Αναφορές του Κεφαλαίου

- [A1] Laird AR, Lancaster JL, Fox PT. (2005). BrainMap: The social evolution of a functional neuroimaging database. *Neuroinformatics* 3, 65-78.
- [A2] Fox PT, Lancaster JL. Mapping context and content: The BrainMap model. *Nature Rev Neurosci* 3, 319-321, 2002.
- [A3] Batrancourt B., Levy R., Lehericy S., Hasboun D., Samson Y., Lavallee I., Lamure M., Dubois B.: Base de données anatomo-fonctionnelle sur le cerveau. An anatomofunctional brain database, *Comptes Rendus Biologies*, Volume 325, Number 4, April 2002 , pp. 439-455(17).
- [A4] Bjaalie JG, Grillner S, Usui S.: *Neuroinformatics: databases, tools, and computational modeling for studying the nervous system*. *Neural Netw.* 2008; 21:1045-6.
- [A5] Nenonen, H. – M., Rajala and T. Katila.: *Biomagnetic localization and 3D modelling*, Report TKK-F-A689, Helsinki University of Technology, Department of Technical Physics, Otaniemi 1991.
- [A6] *Critical reviews in biomagnetic engineering*, vol. 14, issue 2.
- [A7] Franzblau, D.S., Kleitman, D.J.: An Algorithm for Constructing Regions with Rectangles, Independence and Minimum generating sets for collections of Intervals. In 16th ACM Symposium Theory Computing, pp. 167-174, 1984.
- [A8] Willard, D.: New Data Structures for orthogonal range queries. In *SIAM Journal on Computing*, Vol. 14, pp.232-253, 1985.
- [A9] Preparata, F., Shamos, M.: *Computational Geometry. An Introduction*. Springer-Verlag 1986.

- [A10] Korth, E., Silberschatz, A.: Database system concepts. McGrawhill, University of Atlas at Austin, 1986.
- [A11] Rubeinstein, W.B.: A Database Design for Musical Information. Computer Science Division, University of California, Berkeley, ACM SIGMOD 1987.
- [A12] Abiteboul, S., Hull, R.: IFO: A Formal Semantic Database Model. In ACM Transactions on Database Systems, 1987.
- [A13] Overmars, M.H.: Efficient Data Structures for Range Searching on a Grid. In Journal of Algorithms 9, pp.254-275, 1988.
- [A14] Talairach, J., Tournoux, P.: Co-Planar stereotaxic atlas of the brain. Georg Thieme Verlag Stuttgart, New York, Thieme Medical Publishers, Inc. New York, 1988.
- [A15] Hull, R., Yap, C.K.: The Format Model: A theory of database organization. J.ACM 31, 3, 1984.
- [A16] Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan-Kaufmann, 1989.
- [A17] Parsaye, K., Chignell, M., Khoshafian, S., Wong, H.: Intelligent Databases. John Wiley and Sons Inc, 1989.
- [A18] Mehlhorn, K., Tsakalidis, A.: Chapter 6: Data Structures. In Handbook of Theoretical Computer Science, Editor J. V. Leeuwen, Elsevier Science Publishers, co-published by MIT-Press, 1990.
- [A19] Egemhofer, M., Herring, J.: A Mathematical Framework for the Definitions of Topological Relationships. In the 4th International Symposium on Spatial Data Handling 1990.
- [A20] Booch G.: Object-Oriented Design with applications. Benjamin 1991.
- [A21] Bistiolas, V., Sofotassios, D., Tsakalidis, A.: Computing Rectangle Enclosures. In Computational Geometry: Theory and Applications, 2,2 1993.
- [A22] Kapelios, V., Papamichail, G.: Intersection of orthocanonic regions and computational geometry algorithms. Diploma Thesis, Department of Computer Engineering and Informatics, University of Patras, Greece, 1993.
- [A23] Margioulas, V., Pikrakis, A.: Algorithms for compression of files with geometric data. Diploma Thesis, Department of Computer Engineering and Informatics, University of Patras, Greece, 1993.
- [A24] Jacobson I.: Object-Oriented Software Engineering – A use case Driven Approach. Addison-Wesley 1992.
- [A25] Carey , M. J., Dewitt, D. J., Vandenberg, S. L.: A data model and query language for EXODUS. ACM SIGMOD 1988.
- [A26] Fishman, D. H., et al.:IRIS: An object-oriented database management system. ACM Trans. Office Inf. Syst. 5,1, January 1987.
- [A27] Jagannathan, D., et al.: SIM: A database system based on the semantic data model. ACM SIGMOD 1988.
- [A28] Kuper, G., Vardi, M.: The Logical Data Model. ACM Transactions on Database Systems, September 1993.
- [A29] Ishikawa, H. et al.: The Model, Language and Implementation of an Object-Oriented Multimedia Knowledge Base Management System. ACM Transactions on Database Systems, March 1993.
- [A30] Becker, L., Guting, H. R.: Rule-Based optimization and Query Processing in an Extensible Geometric Database System. ACM Transactions on Database Systems, June 1992.

- [A31] Abel, D. J.: SIRO-DBMS: A Database Tool-kit for Geographical Information Systems. Int. Journal Geographical Information Systems, 3,2, 1989.
- [A32] Brodie, M. L.: On the Development of Data Models. Conceptual Modeling, 1982.

Η μάθηση σ' όσους ευτυχούν είναι στολίδι
και σ' όσους ατυχούν είναι καταφύγιο...

Δημόκριτος

Μέρος Β':

Ανέυρεση Ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα (Similarity Search among Multidimensional Objects)

Όπως έχει προαναφερθεί, η αρχική μας έρευνα είχε σαν αφετηρία και κίνητρο το χρηματοδοτούμενο έργο της χαρτογράφησης του ανθρώπινου εγκεφάλου για την ανακάλυψη ιδιαίτερα χρήσιμης ιατρικής γνώσης. Κατά την έρευνα αυτή ανέκυσαν δύο βασικές ερευνητικές κατηγορίες:

- 1) Η *Αναπαράσταση και Διαχείριση* πολύπλοκων, μη-ομοιογενών, δικτυακών μερικές φορές τρισδιάστατων αντικειμένων (τμημάτων του εγκεφάλου – *brain objects*).
- 2) Η *αποτελεσματική διαχείριση-χρήση* γνωστών αλλά και παραγόμενων εξαρτήσεων και συσχετίσεων δεδομένων και γνώσης (data and knowledge dependencies), η οποία μπορεί να αναβαθμίσει την απόδοση και τη δυναμική της εν λόγω χαρτογράφησης. Η μελέτη που αφορούσε αυτό το πρόβλημα, οδήγησε στη διαπίστωση πως είναι αναγκαία η:
  - ♦ *διερεύνηση θεμάτων ανεύρεσης ομοιοτήτων (similarity search) ανάμεσα σε πολυδιάστατα αντικείμενα.*

η οποία αποτελεί και το μεγαλύτερο (δεύτερο) μέρος της συνολικής έρευνας.

Μέχρι το πέρας της χρηματοδότησης, είχαν μελετηθεί προβλήματα που αφορούν την πρώτη από τις παραπάνω κατηγορίες (παρουσιάστηκαν στο πρώτο μέρος της παρούσας διατριβής), ενώ η έρευνα σχετικά με την ανίχνευση ομοιοτήτων ακολούθησε, ανεξάρτητη μεν αλλά εμπνευσμένη από το αρχικό έργο-εφαρμογή της χαρτογράφησης του ανθρώπινου εγκεφάλου. Γι' αυτό το λόγο, αν και οι ανάγκες της συγκεκριμένης εφαρμογής αποτελούν σημείο εκκίνησης και αναφοράς σ' αυτό το δεύτερο μέρος της έρευνας, δεν εστιάζουμε σε αυτή την εφαρμογή αλλά υπάρχει μία μεγαλύτερη επέκταση στα 'πεδία ορισμού' των μεθόδων καθώς

και των βασικών παραδειγμάτων εφαρμογής. Αυτό γίνεται κυρίως για λόγους *μεγαλύτερης γενίκευσης/εμβέλειας των αποτελεσμάτων και ευκολότερης κατανόησης των παραδειγμάτων* – αφού ο εγκέφαλος είναι ένα θαυμάσιο αλλά ιδιαίτερα πολύπλοκο και μη οικείο γενικά πεδίο.

Το δεύτερο λοιπόν μέρος της παρούσας διατριβής, πραγματεύεται ζητήματα που αφορούν την αναζήτηση ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα. Αυτό το θέμα αποτελεί ένα ιδιαίτερα ευρύ και πολύπλευρο πεδίο έρευνας το οποίο έχει διερευνηθεί από διάφορες πλευρές και οπτικές γωνίες. Παρόλα αυτά, πολλά είναι τα ανοικτά θέματα καθώς και οι δυνατότητες βελτιστοποίησης και διεύρυνσης υπαρκτών λύσεων. Στα πλαίσια της έρευνάς μας, διερευνώνται τρία διαφορετικά ζητήματα αναζήτησης ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα και παρουσιάζονται στα τρία επόμενα κεφάλαια με τα αντίστοιχα θέματα:

- Τμηματική Ομοιότητα (*ανοικτό θέμα*).
- Dimensional Curse (πτώση της απόδοσης των δέντρων ομοιοτήτων με την αύξηση των διαστάσεων των αντικειμένων) και αντιμετώπισή της (*βελτιστοποίηση – ανοικτό θέμα*).
- Αναζήτηση συγγένειας (κοινής πορείας) ανάμεσα σε χαρακτηριστικά πολυδιάστατων αντικειμένων (*διεύρυνση*).

Η επιστήμη πρέπει να προήλθε από την αίσθηση ότι κάτι δεν πήγαινε καλά.

Thomas Carlyle, 1795-1888, Άγγλος συγγραφέας

Κεφάλαιο 2

Δύο Μέθοδοι για Ακριβέστερη Αναζήτηση Τμηματικής Ομοιότητας (Lui k-n-match και INTESIS)

Περιεχόμενα Κεφαλαίου

- 2.1. Βασικές Έννοιες
  - 2.1.1. Μέθοδοι *k-n-match* και *frequent k-n-match*, Προβλήματα- Ανεπάρκειες
- 2.2.
  - 2.2.1. Μέθοδος Lui k-n-match
  - 2.2.2. Αξιολόγηση – Συμπεράσματα
- 2.3. Μέθοδος INTESIS
  - 2.3.1. Τμηματική Ομοιότητα
  - 2.3.2. Ολοκλήρωση Τμηματικής και Πλήρους Ομοιότητας
  - 2.3.3. Αξιολόγηση – Συμπεράσματα
- 2.4. Δημοσιεύσεις του Κεφαλαίου
- 2.5. Αναφορές του Κεφαλαίου
- 2.6. Παράρτημα 2^ο Κεφαλαίου – R*-trees

Σύμβολο	Έννοια - Σημασία
NN	Πρόβλημα του κοντινότερου γείτονα (nearest neighbor (NN))
D	Απόσταση μεταξύ του δοθέντος αντικειμένου(query object) και των αντικειμένων της εφαρμογής(data objects) - (Συνήθως Ευκλείδεια)
d_i	i-οστή διάσταση πολυδιάστατου αντικειμένου
Q, q	Αντικείμενο Ερώτησης (Query Object)
k	Ο αριθμός των αντικειμένων που ανακτώνται στο εκάστοτε k-n-match
n	Ο αριθμός των διαστάσεων με ομοιότητα
BΔ	Η Βάση Δεδομένων: Ένα σύνολο πολυδιάστατων σημείων-αντικειμένων
P	Ένα σημείο-αντικείμενο
p_i	Η συντεταγμένη του P στη i-στη διάσταση
q_i	Η συντεταγμένη του Q στη i-στη διάσταση
S	Σύνολο σημείων-αντικειμένων
Π_A	Υπο-Αντικείμενο του A, ως προς τον Προσανατολισμό
M_A	Υπο-Αντικείμενο του A, ως προς το Μέγεθος
X_A	Υπο-Αντικείμενο του A, ως προς το Χρώμα

Πίνακας 2.1. : Πίνακας Συμβολισμών

Σύνοψη

Το θέμα του δεύτερου κεφαλαίου είναι η επίλυση ενός ανοικτού θέματος, της *ανίχνευσης της τμηματικής ομοιότητας ανάμεσα σε πολυδιάστατα αντικείμενα και της δεικτοδότησης των αποτελεσμάτων της*. Συγκεκριμένα προτείνονται δύο μέθοδοι: η *Lui k-n-match* για την ανίχνευση και η *INTESIS* για την ανίχνευση και τη δεικτοδότηση των αποτελεσμάτων.

Η πρώτη μέθοδος χρησιμοποιεί αποτελέσματα της πιο σημαντικής πάνω στο θέμα μελέτης [A12], προτείνοντας μία παραλλαγή της όπου εισάγεται η συνεργασία με το χρήστη οδηγώντας σε σαφώς πιο ασφαλή αποτελέσματα. Παράλληλα λαμβάνεται μέριμνα για τη μείωση του απαιτούμενου χρόνου, εισάγοντας έναν ευρετικό (heuristic) τρόπο γρηγορότερου τερματισμού της μεθόδου.

Η δεύτερη μέθοδος βασίζεται σε μία παρατήρηση των συνθηκών όπου μπορεί να ισχύσει η μεταβατική ιδιότητα (βασική για τη χρήση των δέντρων), προτείνει μία ‘διαίρει και βασίλευε’ διαδικασία, η οποία δίνει τη δυνατότητα ανίχνευσης τμηματικής, πλήρους αλλά και *ατομικής ομοιότητας*¹ τα αποτελέσματα των οποίων μπορούν να δεικτοδοτηθούν. Η απόδοση της μεθόδου μπορεί να συναγωνιστεί εκείνη του R*-tree ενώ ακόμη και η χειρότερή της περίπτωση δεν εισάγει μεγάλες επιβαρύνσεις.

Λέξεις Κλειδιά: Εξόρυξη πολυδιάστατων δεδομένων, τμηματική ομοιότητα (partial similarity), k-n-match, αλληλεπίδραση ανθρώπου-υπολογιστή (human-computer interaction), ατομική ομοιότητα (individual similarity), πλήρης ομοιότητα (full similarity), διαίρεση προβλήματος (problem division), αλγόριθμος MultiTreeSimilarity.

2.1. Βασικές Έννοιες

Οι περισσότερες μελέτες σχετικά με την αναζήτηση ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα, έχουν μοντελοποιήσει (προτυποποιήσει) το πρόβλημα αυτού του είδους της ομοιότητας ως το *πρόβλημα του κοντινότερου γείτονα (nearest neighbor (NN) problem)*. Για τη λύση του NN χρησιμοποιείται η απόσταση D^* μεταξύ του δοθέντος αντικειμένου(query object) και των αντικειμένων της εφαρμογής(data objects), με βάση το σύνολο των χαρακτηριστικών(features) των αντικειμένων. Μια τέτοια προσέγγιση έχει ένα βασικό μειονέκτημα: όταν υπάρχουν διαστάσεις με μεγάλες διαφορές, ακόμη κι αν είναι λίγες, αυξάνουν αρκετά την υπολογιζόμενη απόσταση έτσι ώστε οι αποστάσεις αντικειμένων που στην πραγματικότητα είναι όμοια, να καταλήγουν να έχουν μεγάλες τιμές και συνεπώς να μην ανιχνεύεται η ομοιότητά τους (π.χ. όμοια αντικείμενα με πολύ διαφορετικό μέγεθος). Από την άλλη πλευρά, για αντικείμενα τα οποία διαφέρουν λίγο σε κάθε διάσταση (π.χ. λίγο διαφορετικό χρώμα, σχήμα, προσανατολισμό κ.λ.π.) και καταλήγουν να είναι στην πραγματικότητα συνολικά πολύ διαφορετικά, η υπολογιζόμενη μεταξύ τους απόσταση έχει μικρή τιμή, οπότε ανιχνεύονται σαν όμοια, χωρίς να είναι. Το τελικό αποτέλεσμα-μειονέκτημα είναι να μην εντοπίζονται τμηματικές(partial) ομοιότητες και συνεπώς τμηματικά όμοια αντικείμενα και τη θέση τους στο σύνολο των όμοιων αντικειμένων να παίρνουν αντικείμενα λιγότερο όμοια. Οι τμηματικές ομοιότητες, σε αντίθεση με τον ιδιαίτερα διερευνημένο τύπο της πλήρους ομοιότητας (full similarity), αναφέρονται σε πραγματικές ομοιότητες οι οποίες δεν είναι πλήρεις, αλλά καθιστούν τα αντικείμενα όμοια. Για παράδειγμα, όταν έχουμε ένα σύνολο αντικειμένων με χαρακτηριστικά όπως διαστάσεις, προσανατολισμό, θέση και χρώμα και κάποια αντικείμενα είναι πραγματικά όμοια ή ακόμη και ίδια στα τρία πρώτα

¹ Έννοια που εισάγεται μέσω της μεθόδου INTESIS.

* Συνήθως η Ευκλείδεια απόσταση

χαρακτηριστικά αλλά έχουν πολύ διαφορετικό χρώμα (π.χ. άσπρο-μαύρο), αυτά αποτελούν τμηματικά όμοια αντικείμενα (όπως είναι φανερό, σ' αυτή την περίπτωση η απόσταση D μεταξύ τους θα είναι αρκετά μεγάλη (λόγω του χρώματος) για να παραπέμψει στην υπαρκτή² ομοιότητα).

Οι περισσότερες εργασίες οι οποίες έχουν μελετήσει την τμηματική ομοιότητα, έχουν εστιάσει σε δεδομένα εικόνας [A33, A34, A35]. Η εργασία που επεκτείνεται σε πολυδιάστατα αντικείμενα γενικά, είναι η [A12] και έχει οδηγήσει σε αξιόλογα αποτελέσματα στο θέμα της τμηματικής ομοιότητας, εισάγοντας τις μεθόδους *k-n-match* και *frequent k-n-match*. Παρόλα αυτά παρουσιάζει κάποιες αδυναμίες οι οποίες τελικά οδηγούν είτε σε ανεύρεση πλήρους ομοιότητας είτε σε μία κατά περίπτωση μόνο (και σχεδόν τυχαία) ανίχνευση τμηματικής ομοιότητας (με τα κατάλληλα n 's, τα οποία όμως δεν μπορούν να προσδιοριστούν από κάποιο τύπο ή μέθοδο). Όλα αυτά θα παρουσιαστούν στην επόμενη ενότητα.

2.1.1. Μέθοδοι *k-n-match* και *frequent k-n-match*, Προβλήματα-Ανεπάρκειες

Προτού συζητηθούν οι δύο αυτές μέθοδοι, θα ασχοληθούμε με ένα αντιπροσωπευτικό παράδειγμα όπου αφενός οι τμηματικές ομοιότητες είναι φανερές 'με το μάτι' αλλά δεν ανιχνεύονται από τη χρήση της κλασσικής Ευκλείδειας απόστασης και αφετέρου αναδεικνύεται η ανωτερότητα – σε αρχικό επίπεδο – της μεθόδου *k-n-match*. Έστω λοιπόν τέσσερα ενδεικτικά αντικείμενα μίας βάσης δεδομένων δέκα (10) διαστάσεων [A12], όπως φαίνονται στον Πίνακα 2.2. και ένα αντικείμενο ερώτησης Q(query object) (1,1,1,1,1,1,1,1,1). Η αναζήτηση του κοντινότερου γείτονα του Q με τη χρήση της Ευκλείδειας απόστασης θα επιστρέψει σαν αποτέλεσμα το αντικείμενο 4, ενώ είναι εύκολο να διακρίνει κανείς πως τα τρία προηγούμενα αντικείμενα είναι στην πραγματικότητα περισσότερο όμοια με το Q αλλά έχει θεωρηθεί το αντίθετο λόγω της μεγάλης τους διαφοράς με το Q σε μία διάσταση (η διάσταση με την τιμή 100).

ID	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
1	1,05	100	1,2	1,6	1,6	1,05	1,2	1,2	1	1
2	1,4	1,4	1,4	15	100	1,4	1,2	1,2	1,05	1,1
3	1,1	1,1	1,1	1,1	1,1	1,1	2	100	2	2
4	1	1	20	20	20	20	20	20	20	20
5	1	1	1,05	1,05	50	50	50	50	50	100

Πίνακας 2.2.: Παράδειγμα όπου δεν ανιχνεύεται η τμηματική ομοιότητα

Αυτό το φαινόμενο γίνεται περισσότερο έντονο σε δεδομένα πολλών διαστάσεων όπου μεγαλώνει η πιθανότητα ύπαρξης κάποιων διαστάσεων με πολύ μεγάλες διαφορές.

² Δύο τέτοια αντικείμενα στη γενική περίπτωση – και αυτή θα εξετάσουμε - θεωρούνται όμοια. Σε πολύ ειδικές εφαρμογές είναι δυνατόν διαφορές σε κάποια χαρακτηριστικά να αποτελούν πράγματι συνολική διαφορά (π.χ. σε κάποιες ιατρικές εφαρμογές η διαφορά στο χρώμα ενός όγκου).

Χαρακτηριστικά παραδείγματα τέτοιων εφαρμογών έχουμε όταν υπάρχει μεταφορά δεδομένων όπου μπορεί να υπάρξει θόρυβος και συνεπώς αλλοίωση κάποιων δεδομένων.

Για να αποφευχθούν λαθεμένα αποτελέσματα όπως αυτό του παραδείγματος του Πίνακα 2.2., στην [A12] προτείνεται αρχικά η μέθοδος *k-n-match*, η οποία με δυναμικό τρόπο εστιάζει τον έλεγχο στις διαστάσεις όπου τα αντικείμενα είναι περισσότερο όμοια. Μ' άλλα λόγια, η *k-n-match* βασίζεται σε υπολογισμό διαφορών - ανάμεσα στα αντικείμενα - μικρού αριθμού διαστάσεων(των πιο όμοιων) και όχι όλων. Πιο συγκεκριμένα οι ορισμοί που οδηγούν στον τελικό ορισμό του *k-n-match* προβλήματος είναι:

Ορισμός 1: *N-match difference*

Δοθέντων δύο σημείων(αντικειμένων) $P(p_1, p_2, \dots, p_d)$ και $Q(q_1, q_2, \dots, q_d)$ με d διαστάσεις, έστω $\delta_i = |a_i - b_i|$, $i=1, \dots, d$. Ο πίνακας ταξινομείται σε αύξουσα σειρά και έστω $\{\delta'_1, \dots, \delta'_d\}$ ο ταξινομημένος πίνακας. Τότε η δ'_n είναι η διαφορά *n-match* του A από το B .

Ορισμός 2: *Πρόβλημα N-match*

Δοθείσης μίας βάσης δεδομένων BA d διαστάσεων, ενός σημείου (αντικειμένου) ερώτησης Q και ενός ακεραίου n ($1 \leq n \leq d$), βρίσκεται το σημείο (αντικείμενο) $P \in BA$ το οποίο έχει τη μικρότερη διαφορά *n-match* από το Q . Το P καλείται το *n-match* σημείο του Q .

Στο παράδειγμα του Πίνακα 2.2., το σημείο 3 είναι το 6-match ($\delta=0,1$) του Q , το σημείο 1 είναι το 7-Match($\delta=0,2$) και το σημείο 2 το 8-match ($\delta=0,4$).

Μετά από τους παραπάνω ορισμούς μπορεί να οριστεί το πρόβλημα *k-n-match*.

Ορισμός 3: *Πρόβλημα k-n-match*

Δοθείσης μίας βάσης δεδομένων BA , d διαστάσεων και c σημείων(αντικειμένων), ενός σημείου (αντικειμένου) ερώτησης Q , ενός ακεραίου n ($1 \leq n \leq d$) και ενός ακεραίου $k \leq c$, βρίσκεται ένα σύνολο S αποτελούμενο από k σημεία(αντικείμενα) της BA τέτοια ώστε για κάθε σημείο $P1 \in S$ και κάθε σημείο $P2 \in BA-S$, η διαφορά *n-match* μεταξύ των $P1$ και Q να είναι μικρότερη ή ίση με τη διαφορά *n-match* μεταξύ των $P2$ και Q . Το S είναι το *k-n-match* σύνολο του Q .

Όπως είναι φανερό, το *k-n-match* ερώτημα για το παράδειγμα του Πίνακα 2.2., θα επιστρέψει ως όμοια προς το Q (ανάλογα με την τιμή του n) κάποια από τα αντικείμενα(σημεία) 1,2 και 3, ενώ θα αποκλείσει το τέταρτο και το πέμπτο αντικείμενο τα οποία είναι φανερά ανόμοια με το Q . Το ποιά ακριβώς αντικείμενα θα επιστρέψει η μέθοδος, εξαρτάται από την τιμή του n (δηλαδή από το πόσες διαστάσεις θα ληφθούν υπ' όψιν στην ομοιότητα) γεγονός που παραμετροποιεί τη μέθοδο. Σε πειράματα που παρουσιάζονται στην [A12] φαίνεται πολύ καθαρά πως η ποιότητα και η ακρίβεια των αποτελεσμάτων εξαρτάται από την εκάστοτε τιμή του n , για την οποία όμως δεν υπάρχει κάποιος τρόπος να εντοπιστεί η καταλληλότερη. Πειραματικά αλλά και θεωρητικά έχει αποδειχτεί ότι ούτε οι μικρές αλλά ούτε και οι πολύ μεγάλες τιμές του n είναι ιδανικές για καλά/αξιόπιστα αποτελέσματα, περιορίζοντας έτσι το σύνολο τιμών του n χωρίς όμως να μπορεί να εντοπιστεί το ακριβές 'καλό' n . Για να αντιμετωπιστεί αυτό το πρόβλημα, στην [A12] ορίζεται η μέθοδος *Frequent K-N-Match*, κατά την οποία εντοπίζονται πρώτα τα *k-n-match* σύνολα απαντήσεων για ένα εύρος $[n_0, \dots, n_1]$ των τιμών του n και κατόπιν επιλέγονται ως απάντηση τα k αντικείμενα/σημεία που εμφανίζονται πιο συχνά στα σύνολα απαντήσεων.

Ορισμός 4: *Πρόβλημα Frequent K-N-Match*

Δοθείσης μίας βάσης δεδομένων BA , d διαστάσεων και c σημείων(αντικειμένων), ενός σημείου (αντικειμένου) ερώτησης Q , ενός διαστήματος ακεραίων $[n_0, \dots, n_1]$ εντός του $[1, \dots, d]$

και ενός ακεραίου $k \leq c$, έστω ότι τα $S_0, \dots, S_i$ είναι τα σύνολα απαντήσεων των k - n_0 -match, ..., k - n_i -match αντίστοιχα. Βρίσκεται ένα σύνολο T από k σημεία (αντικείμενα), τέτοιο ώστε για κάθε σημείο(αντικείμενο) $P_1 \in T$ και για κάθε σημείο(αντικείμενο) $P_2 \in B \Delta - T$, ο αριθμός εμφανίσεων του P_1 στα $S_0, \dots, S_i$, να είναι μεγαλύτερος ή ίσος με κείνον του P_2 .

Η λογική αυτής της μεθόδου έχει στηριχτεί στο ότι αφού δεν μπορεί να εντοπιστεί το καταλληλότερο n για το k - n -match, παράγουμε περισσότερα k - n -matches (για πολλά n) και διαλέγουμε τα αντικείμενα που εμφανίζονται πιο συχνά στις απαντήσεις. Εδώ όμως κρύβεται μια παγίδα: εφόσον λαμβάνονται υπ'όψιν πολλά n – και μάλιστα και μεγάλα n – αυτό σημαίνει πως τα αποτελέσματα επηρεάζονται από πολλές διαστάσεις – όλες ή σχεδόν όλες – οπότε καταλήγουμε σε εντοπισμό πλήρους ομοιότητας. Αυτό αποδεικνύεται στην [A12], όπου αποσαφηνίζεται ότι τελικά το frequent k - n -match είναι μέθοδος ανεύρεσης πλήρους ομοιότητας. Μετά από αυτό, εάν επιστέψουμε στο αρχικό πρόβλημα που είναι ο εντοπισμός τμηματικών ομοιοτήτων, συμπεραίνουμε πως το πρόβλημα ουσιαστικά δεν έχει αντιμετωπιστεί ούτε με το frequent k - n -match και συνεπώς παραμένει ανοικτό. Στόχος αυτού του κεφαλαίου είναι η παρουσίαση δύο νέων μεθόδων (*Lui k-n-match* και *INTESIS*) για τον εντοπισμό τμηματικών ομοιοτήτων αλλά και την επίτευξη μεγάλης εγκυρότητας και αξιοπιστίας των αποτελεσμάτων.

2.2.

2.2.1. Μέθοδος Lui (Limited User-Interactive) k-n-match

[Σ18]

Μετά την προσεκτική μελέτη των μεθόδων που προτάθηκαν στο [A12] εντοπίστηκαν τα βασικά σημεία αδυναμίας τους και η έρευνα στράφηκε στον ορισμό μίας νέας μεθόδου η οποία βασισμένη σε κείνες του [A12] θα άρει τις αδυναμίες αυτές δίνοντας μία αποδοτική λύση.

Πιο συγκεκριμένα, όσον αφορά την k - n -match το πρόβλημα της βρίσκεται στην αδυναμία εντοπισμού του κατάλληλου n . Από την άλλη μεριά, η frequent k - n -match προκειμένου να ξεπεράσει αυτή την αδυναμία καταλήγει να χάνει τον αρχικό στόχο και να εντοπίζει ομοιότητες όχι πια τμηματικές, αλλά πλήρεις. Στην πραγματικότητα οι περισσότερες τμηματικές ομοιότητες εντοπίζονται σε κάποια k - n -matches αλλά λόγω του ότι δεν εμπλέκουν το σύνολο των διαστάσεων συνήθως δεν εμφανίζονται σε τόσα σύνολα απαντήσεων ώστε να αποτελέσουν μέρος του τελικού συνόλου απάντησης. Αυτή η παρατήρηση μας δίνει την ιδέα εισαγωγής αλληλεπίδρασης ανθρώπου-υπολογιστή, έτσι ώστε το τελικό σύνολο απαντήσεων σε κάθε k - n -match να καθορίζεται από το φιλτράρισμα που θα κάνει ο χρήστης. Μ' αυτό τον τρόπο τα τμηματικά όμοια αντικείμενα δε θα 'χαθούν' στην πορεία φιλτραρίσματος των αποτελεσμάτων αλλά θα παραμείνουν στο τελικό σύνολο απάντησης λόγω της εμπλοκής του χρήστη. Πιο συγκεκριμένα μια αρχική βελτίωση του frequent k - n -match περιλαμβάνει την παρουσίαση στο χρήστη του συνόλου αποτελεσμάτων για κάθε k - n -match και την επιλογή από εκείνον των αντικειμένων που θεωρεί πως όντως είναι όμοια με το δοθέν αντικείμενο (Q). Αυτά τα επιλεγμένα αντικείμενα θα αποτελούν το τελικό σύνολο αποτελεσμάτων για το συγκεκριμένο k - n -match. Αν η τελική επιλογή των k αντικειμένων γίνει με τη λογική της frequent k - n -match, τότε παρόλο που τα τμηματικά όμοια αντικείμενα δε θα χαθούν στην πορεία των k - n -matches, πιθανώς να χαθούν κατά την τελική επιλογή των πιο συχνά εμφανιζόμενων στα σύνολα απαντήσεων, αφού ειδικά για τα

μεγάλα n δε θα αποτελούν μέρος των απαντήσεων³. Έτσι προτείνουμε και σ' αυτό το στάδιο να υπάρχει εμπλοκή του χρήστη, ο οποίος θα επιλέγει τα τελικά k όμοια αντικείμενα ανάμεσα στα αντικείμενα των επιμέρους συνόλων αποτελεσμάτων. Το αποτέλεσμα είναι μέχρι εδώ να έχει επιτευχθεί υψηλή εγκυρότητα αποτελεσμάτων και επιλογή των τμηματικά όμοιων αντικειμένων, εφόσον υπεισέρχεται το υπεύθυνο φιλτράρισμα του χρήστη. Το πρόβλημα που εγείρεται εδώ αφορά τη χρονική επιβάρυνση που δημιουργείται όχι μόνο από τη διεπαφή με το χρήστη αλλά και από τον υπολογισμό πολλών k - n -matches σημαντικό ποσοστό των οποίων – όπως εξηγήθηκε παραπάνω – είναι περιττά λόγω του ότι προτείνουν μη όμοια αντικείμενα. Η βασική μας λοιπόν σκέψη σχετικά με τη μείωση αυτής της επιβάρυνσης στράφηκε ακριβώς σε αυτό το δεύτερο παράγοντα καθυστέρησης. Λαμβάνοντας υπ' όψιν τη διαπίστωση της προηγούμενης ενότητας, σύμφωνα με την οποία ούτε οι μικρές αλλά ούτε και οι πολύ μεγάλες τιμές του n είναι ιδανικές για καλά/αξιόπιστα αποτελέσματα, προσπαθήσαμε να βρούμε κάποια όρια για τα n των k - n -matches. Είναι κατανοητό πως τα πολύ μικρά n εφόσον εμπλέκουν πολύ λίγες διαστάσεις, στη γενική περίπτωση δεν μπορούν να εντοπίσουν πολλές πραγματικές ομοιότητες. Παρόλα αυτά, όπως είναι θεωρητικά αναμενόμενο αλλά φαίνεται και από τα πειράματα της [A12], αρκετά πραγματικά όμοια αντικείμενα εντοπίζονται ήδη από k - n -matches για μικρά n . Αυτό συμβαίνει γιατί κάποια από τα όμοια αντικείμενα παρουσιάζουν μεγάλη ομοιότητα σε κάποιες διαστάσεις που εμπλέκονται στο συγκεκριμένο μικρό n και αυτές οι διαστάσεις είναι που διαμορφώνουν τις αντίστοιχες χαμηλές διαφορές n -match, οι οποίες καθορίζουν τα αποτελέσματα σε κάθε k - n -match σύνολο αποτελεσμάτων. Από την άλλη μεριά, για τα μεγάλα n , υπάρχει ένα n από το οποίο και μετά, τα σύνολα απαντήσεων των αντίστοιχων k - n -matches δεν περιλαμβάνουν κανένα αντικείμενο με πραγματική τμηματική ομοιότητα με το Q . Πιο συγκεκριμένα, όπως παρατηρήθηκε από πειράματα του [A12], όσο μεγαλύτερος είναι ο αριθμός των διαστάσεων του συνόλου δεδομένων, τόσο περισσότερα είναι τα n των οποίων τα k - n -matches δεν εντοπίζουν τμηματικές ομοιότητες. Αυτό συμβαίνει γιατί όσο μεγαλώνει ο αριθμός των διαστάσεων, οι υπολογιζόμενες διαφορές δεν μπορούν να αντιπροσωπεύσουν ομοιότητα επειδή η απόσταση του κοντινότερου στο Q αντικειμένου προσεγγίζει εκείνη του πιο απομακρυσμένου [A13, A14]. Παρόλο που αυτή η παρατήρηση έχει αποδειχτεί για άλλες μετρικές αποστάσεων (π.χ. Ευκλείδεια απόσταση) δεν περιορίζεται σ' αυτές, κι αυτό γιατί η βάση της απόδειξης είναι ότι οι αποστάσεις αντικειμένων πολλών διαστάσεων επιστρέφουν μία μέση απόσταση η οποία – καθώς αυξάνονται οι διαστάσεις – δεν μπορεί να αντιπροσωπεύσει τα συγκεκριμένα χαρακτηριστικά και τις διαφορές των αντικειμένων. Μ' άλλα λόγια δεν μπορούμε να ξέρουμε αν η υπολογισμένη απόσταση έχει επηρεαστεί κυρίως από πολλές μικρές ή μέτριες διαφορές ή από λίγες αλλά μεγάλες διαφορές ανάμεσα στις διαστάσεις. Στην πρώτη περίπτωση τα αντίστοιχα αντικείμενα είναι μάλλον ανόμοια ενώ στη δεύτερη συμβαίνει το αντίθετο.

Λαμβάνοντας υπ' όψιν αυτές τις παρατηρήσεις, βγάλαμε το συμπέρασμα πως ο περιορισμός των ανώφελων k - n -matches μπορεί να επιτευχθεί εστιάζοντας και ελέγχοντας τη συμπεριφορά των μεγάλων n . Χρησιμοποιώντας λοιπόν την αλληλεπίδραση με το χρήστη – όπως περιγράφηκε παραπάνω – καταφέραμε να ελαττώσουμε τον αριθμό των εκτελούμενων k - n -matches, ελέγχοντας για μεγάλα n , σε ποιο n δεν επιλέγονται από το χρήστη νέα όμοια αντικείμενα. Και αφού δεν υπάρχει κάποιος τρόπος ορισμού του ορίων για τον προσδιορισμό των μεγάλων και των μικρών n , ορίσαμε αυθαίρετα ως όριο το $d/2$ – το ήμισυ του αριθμού των διαστάσεων – προκειμένου να έχουμε ένα συγκεκριμένο και ασφαλές πλαίσιο.

³ Εφόσον για τα μεγάλα n στην ουσία ελέγχεται η πλήρης ομοιότητα, λόγω των πολλών διατάσεων που εμπλέκονται.

Το αποτέλεσμα αυτής της ανάλυσης και συλλογιστικής είναι ο ορισμός της παρακάτω μεθόδου, η οποία με ασφάλεια και εγκυρότητα εξασφαλίζει την ανεύρεση των τμηματικών ομοιοτήτων σε ένα πολυδιάστατο σύνολο αντικειμένων, με παράλληλη ελαχιστοποίηση της χρονικής επιβάρυνσης:

Lui k-n-match:

Δοθείσης μίας βάσης δεδομένων BD , d διαστάσεων και c σημείων(αντικειμένων), ενός σημείου (αντικειμένου) ερώτησης Q , ενός διαστήματος ακεραίων $[n_0, \dots, n_1]$ εντός του $[1, \dots, d]$ και ενός ακεραίου $k \leq c$, υπολογίζονται τα k -n-matches για ένα σύνολο από n $[n_{lui0}, n_{lui1}]$. Για κάθε n τα αποτελέσματα εμφανίζονται στο χρήστη ο οποίος αποφασίζει ποιά είναι τα k ή λιγότερα πραγματικά τμηματικά όμοια προς το Q αντικείμενα, δημιουργώντας το σύνολο k' το οποίο περιλαμβάνει εκείνα τα επιλεγμένα αντικείμενα τα οποία δεν έχουν επιλεγεί σε προηγούμενα k -n-matches (διαφορετικά (distinct) αντικείμενα). Η διαδικασία σταματά όταν για κάποιο συγκεκριμένο n μεγαλύτερο του $d/2$, το σύνολο k' παραμένει άδειο· αυτό το n ονομάζεται *stop n*. Τελικά, αν το συνολικό πλήθος των αντικειμένων που φιλοξενούνται στα σύνολα k' είναι μικρότερο ή ίσο με το k , τότε το σύνολο αυτών των αντικειμένων αποτελεί την απάντηση-έξοδο της μεθόδου. Διαφορετικά, όλα τα σύνολα k' παρουσιάζονται στο χρήστη ο οποίος επιλέγει τα k περισσότερα όμοια στο Q αντικείμενα και τα οποία αποτελούν το τελικό σύνολο-απάντηση.

Για να γίνει περισσότερο κατανοητή η μέθοδος, θα επεξεργαστούμε το παράδειγμα του Πίνακα 2.2. Θα δοκιμάσουμε τη frequent k-n-match και Lui k-n-match για αυτό το παράδειγμα και θα συγκρίνουμε τα αποτελέσματα και την απόδοσή τους. Προηγουμένως αξίζει να επισημάνουμε ποια είναι η πραγματική κατάσταση. Έτσι διακρίνουμε ότι τα αντικείμενα που όντως μοιάζουν με το δοθέν Q , είναι το 1 και το 3, ενώ τα αντικείμενα 4 και 5 είναι πολύ διαφορετικά από το Q , και το αντικείμενο 2 αν και δεν μπορεί να θεωρηθεί όμοιο με το Q δεν είναι τόσο ακραία διαφορετικό όσο το 4 και το 5.

Παρακάτω παραθέτουμε τον πίνακα των διατεταγμένων n-match διαφορών τους ο οποίος είναι απαραίτητος και για τις δύο μεθόδους (οι στήλες αντιστοιχούν στα n των αντίστοιχων n-matches):

	n – match διαφορές									
ID	1	2	3	4	5	6	7	8	9	10
1	0	0	0,05	0,05	0,2	0,2	0,2	0,6	0,6	99
2	0,05	0,1	0,2	0,2	0,4	0,4	0,4	0,4	14	99
3	0,1	0,1	0,1	0,1	0,1	0,1	1	1	1	99
4	0	0	19	19	19	19	19	19	19	19
5	0	0	0,05	0,05	49	49	49	49	49	99

Πίνακας 2.3.: Πίνακας των διατεταγμένων σε αύξουσα σειρά n-match διαφορών για κάθε αντικείμενο

Εάν ‘τρέξουμε’ τη frequent k-n-match για $k=2$ και n ίσο με 2,4,6,8 και 10, τότε όπως φαίνεται στον Πίνακα 2.4., τα αντικείμενα που εμφανίζονται συχνότερα ως όμοια με το Q είναι τα 1,4,5,3,2, μ’αυτή τη σειρά και συνεπώς η μέθοδος θα επιστρέψει ως απάντηση τα 1 και 4, αντί για τα 1 και 3 όπως θα ήταν το σωστό. Αυτό ήταν και θεωρητικά αναμενόμενο γιατί το αντικείμενο 1 είναι πολύ όμοιο με τα δοθέν ενώ από τα υπόλοιπα τέσσερα το τέταρτο είναι εκείνο που αν και πολύ ανόμοιο με το Q , λόγω του ότι τελικά λαμβάνονται υπ’όψιν όλες οι διαστάσεις – στις οποίες έχει μέτριες διαφορές με το Q – καταλήγει να προκύπτει ως όμοιο. Μ’ άλλα λόγια πάλι ανιχνεύεται ένα είδος πλήρους ομοιότητας που παρουσιάζει τα μειονεκτήματα που περιγράψαμε παραπάνω.

Frequent k-n-match						
n=	2	4	6	8	10	Αποτέλεσμα για $k=2$
αντικείμενα- απάντηση της μεθόδου	1,4	1,5	1,3	1,2	1,4	1,4

Πίνακας 2.4.: Τα ενδιάμεσα αποτελέσματα (k-n-matches) της frequent k-n-match, για $k=2$

Τρέχοντας τώρα τη μέθοδο Lui k-n-match για το ίδιο παράδειγμα θα συμβεί το εξής:

- ⇒ Για $n=2$ παρουσιάζονται τα αντικείμενα 1,4 στο χρήστη και εκείνος θα επιλέξει μόνο το αντικείμενο 1 γιατί όπως σημειώθηκε το αντικείμενο 4 δε μοιάζει ιδιαίτερα με το δοθέν. Συνεπώς το k' θα περιλαμβάνει μόνο το 1 ($k'_2 \rightarrow 1$).
- ⇒ Για $n=4$ παρουσιάζονται στο χρήστη τα αντικείμενα 1 και 5 και ο χρήστης επιλέγει ως όμοιο πάλι το 1 μόνο (αφού το 5 διαφέρει πολύ από το Q). Το k αυτού του βήματος δε θα περιλαμβάνει τίποτα, αφού εξ’ορισμού αποτελείται μόνο από νεοεπιλεγέντα όμοια αντικείμενα.
- ⇒ Για $n=6$ παρουσιάζονται στο χρήστη τα αντικείμενα 1 και 3, ο χρήστης επιλέγει και τα δύο ως όμοια και το αντίστοιχο k' περιλαμβάνει μόνο το 3 ($k'_4 \rightarrow 3$).
- ⇒ Για $n=8$ παρουσιάζονται στο χρήστη τα αντικείμενα 1 και 2, ο χρήστης επιλέγει μόνο το 1 γιατί το αντικείμενο 2 είναι ανόμοιο του Q και το αντίστοιχο σύνολο k' παραμένει άδειο. Δεδομένου όμως ότι εδώ το n είναι μεγαλύτερο από το $d/2$ ($8 > (10/2)$), σύμφωνα με τη μέθοδο η διαδικασία υπολογισμού των k-n-matches σταματά.
- ⇒ Ελέγχονται όλα τα επιμέρους k' και συγκεντρώνεται το περιεχόμενό τους σε ένα τελικό k' , το οποίο είναι $k' \rightarrow 1,3$. Εφόσον το αρχικό k είναι ίσο με 2 και τα αντικείμενα του k' είναι μόνο δύο, δε χρειάζεται άλλη διεπαφή με το χρήστη και ως όμοια του Q παρουσιάζονται τελικά τα αντικείμενα 1 και 3.

Όπως είναι φανερό η Lui k-n-match έχει εντοπίσει τα πραγματικά τμηματικά όμοια με το Q αντικείμενα, έχοντας επιπλέον αποφύγει το τελευταίο k-n-match (το 10-match) το οποίο αποτελεί το 20% του συνόλου των k-n-matches.

Το παράδειγμα ήταν ενδεικτικό και για αυτό το λόγο και ο αριθμός των διαστάσεων και των αντικειμένων δεν ήταν μεγάλος. Στην πραγματικότητα, όσο μεγαλύτερος είναι ο αριθμός των διαστάσεων τόσο μεγαλύτερο ποσοστό των k - n -matches αποφεύγεται, όπως έχει αποδειχτεί στα πειράματά μας, αλλά μπορεί κανείς και διαισθητικά να καταλάβει.

2.2.2. Αξιολόγηση – Συμπεράσματα

Με τη μέθοδο Lui k - n -match επιτυγχάνεται ανίχνευση τμηματικής ομοιότητας με υψηλή ακρίβεια λόγω των δύο φίλτρων στα οποία βασίζεται:

α) τα k - n -matches και

β) την αλληλεπίδραση με το χρήστη (*human-computer interaction*) με τη βοήθεια της οποίας ‘ξεκαθαρίζονται’ - φιλτράρονται τα αποτελέσματα του (α).

Το κύριο προσόν της μεθόδου είναι η ικανότητά της να αξιοποιεί - και να αλληλοσυμπληρώνει - τις δυνατότητες του ανθρώπου και της μηχανής στην ανεύρεση τμηματικών ομοιοτήτων. Είναι φανερό ότι η ακρίβεια της μεθόδου είναι υψηλότερη εκείνης της frequent k - n -match, η οποία με τη σειρά της ξεπερνά την ακρίβεια άλλων αποδοτικών μεθόδων αναζήτησης ομοιοτήτων [A12].

Η μέθοδος έχει ελεγχθεί σε δύο σύνολα δεδομένων. Το πρώτο(A) περιλαμβάνει 40 διαστάσεις ενώ το δεύτερο(B) 15. Χρησιμοποιήθηκαν 5 αντικείμενα για αναζήτηση ομοιότητας και όπως ήταν αναμενόμενο η μέση ακρίβεια αποτελεσμάτων που επιτεύχθηκε ήταν μεγαλύτερη για το δεύτερο σύνολο⁴ (Πίνακας 2.5.). Όπως αναφέρθηκε παραπάνω, η Lui k - n -match προτείνεται προκειμένου να επιτευχθεί ο εντοπισμός τμηματικών διαφορών, κάτι που δεν έχει καταφέρει ουσιαστικά και για οποιοδήποτε είδος δεδομένων κάποια άλλη μέθοδος. Συνεπώς δεν υπάρχει η δυνατότητα σύγκρισης των αποτελεσμάτων της μεθόδου με τα αποτελέσματα κάποιας άλλης αντίστοιχης. Παρόλα αυτά, προκειμένου να υπάρχει μία σφαιρική συγκριτική εικόνα της μεθόδου, ενδεικτικά χρησιμοποιούμε για την αξιολόγησή της την πιο σχετική μέθοδο, τη frequent k - n -match (Πίνακας 2.5.).

Σύνολα Δεδομένων	Ακρίβεια της Freq. k - n - match	Ακρίβεια της Lui k - n -match
A(40διαστάσεις)	83%	92%
B(15 διαστάσεις)	87%	96%

Πίνακας 2.5.: Ακρίβεια αποτελεσμάτων ανά Σύνολο Δεδομένων

Εδώ πρέπει να μην παραβλεφθεί το θέμα των καθυστερήσεων που υπεισέρχονται στο χρόνο απόκρισης της μεθόδου μας, λόγω της αλληλεπίδρασης με τον χρήστη. Αυτό αποτελεί ένα αναγκαίο κακό, όπως σε άλλες μεθόδους [A22] και είναι το τίμημα της ακρίβειας των αποτελεσμάτων. Παρόλα αυτά, η μέθοδος επιτυγχάνει μία σοβαρή ελάττωση αυτής της καθυστέρησης, χρησιμοποιώντας το stop n. Μ’ αυτό τον τρόπο – όπως αποδείχτηκε και πειραματικά – ένα ποσοστό σχεδόν 30% των k - n -matches το οποίο εκτελείται κατά τη frequent k - n -match, αποφεύγεται με ασφάλεια.

⁴ Ο βασικός λόγος γι’ αυτό είναι πως όσο μεγαλώνει ο αριθμός των διαστάσεων των αντικειμένων, τόσο η ομοιότητα γίνεται αμφίβολη και σχετική [A13, A14], γεγονός που επηρεάζει/δυσχεραίνει τη διαδικασία επιλογής από τον χρήστη.

Συμπερασματικά, αυτό που επιτυγχάνεται από τη Lui k-n-match είναι ο ιδιαίτερα ασφαλής εντοπισμός τμηματικών ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα, με σχετικά μικρή χρονική επιβάρυνση.

2.3. Μέθοδος INTESIS (INTEgrated SIMilarity Search among multi-dimensional objects)

[Σ16]

Η απλούστερη λύση είναι συνήθως και η σωστή.

Γουλιέλμος του Όκαμ, 1285-1345, Άγγλος θεολόγος

Η μέθοδος INTESIS είναι αποτέλεσμα της έρευνας μας για την ανεύρεση όχι μόνο κάποιας μεθόδου για τον εντοπισμό τμηματικών ομοιοτήτων αλλά και για την ευέλικτη οργάνωση-δεικτοδότηση των αποτελεσμάτων. Πιο συγκεκριμένα, μελετώντας τις αδυναμίες και τις δυνατότητες της πλήρους ομοιότητας [A1-A11, A13-A32] – την οποία αφορά το μεγαλύτερο μέρος της συνολικής έρευνας σχετικά με την αναζήτηση ομοιοτήτων – προέκυψαν τα παρακάτω ερωτήματα-προβληματισμοί:

- i. Εάν με κάποιο οποιοδήποτε τρόπο εντοπιστούν οι τμηματικές ομοιότητες, πώς θα μπορούσαν τα τμηματικά όμοια αντικείμενα να οργανωθούν/δεικτοδοτηθούν (indexing) προκειμένου να επιτυγχάνεται εύκολη και αποδοτική ανάκτηση;
- ii. Είναι δυνατόν οι μέθοδοι ανεύρεσης τμηματικής ομοιότητας να συνεργαστούν με κείνες της πλήρους ομοιότητας, εξασφαλίζοντας ευελιξία, ολοκλήρωση και αποδοτικότητα;

Έχοντας ήδη ορίσει την προηγούμενη μέθοδο (*Lui k-n-match*), η επινόηση κάποιας μεθόδου δεικτοδότησης των αποτελεσμάτων οδήγησε σε αδιέξοδα. Έτσι ακολουθήθηκε η ‘αντίθετη’ οδός. Επιχειρήσαμε να δεικτοδοτήσουμε τμηματικά όμοια αντικείμενα, χωρίς να λαμβάνουμε υπ’όψιν τη μέθοδο ανεύρεσής τους. Ξεκινώντας λοιπόν από την παραδοχή ‘Έστω πως υπάρχει κάποια μέθοδος ανεύρεσης τμηματικής ομοιότητας η οποία μπορεί να οδηγήσει σε δεικτοδότηση’ συνεχίσαμε αναζητώντας κάποια μέθοδο δεικτοδότησης. Αυτή η αναζήτηση μας επέστρεψε και στο αρχικό πρόβλημα αλλά και στο ζήτημα της ολοκλήρωσης των μεθόδων (ερώτημα ii), καταλήγοντας στον ορισμό μίας συνολικής μεθόδου (*INTESIS*). Η παρουσίαση της μεθόδου θα ακολουθήσει τη σειρά με την οποία επινοήθηκε, θα ξεκινήσει δηλαδή από τη δεικτοδότηση.

2.3.1. Τμηματική Ομοιότητα

Έστω πως υπάρχει κάποια – τυχαία - μέθοδος τμηματικής αναζήτησης η οποία ίσως να μπορεί να υποστηριχτεί από δεικτοδότηση. Εδώ ήδη εγείρεται ένα θεμελιώδες πρόβλημα: *Η κλασσική χρήση οποιουδήποτε δέντρου είναι ανέφικτη και λανθασμένη στην περίπτωση της τμηματικής ομοιότητας*. Κι αυτό γιατί όταν ένα αντικείμενο A μοιάζει τμηματικά με το B, δε σημαίνει ότι μοιάζει και με τα τμηματικά όμοια προς το B αντικείμενα. Με άλλα λόγια, η *μεταβατική ιδιότητα* δεν ισχύει για την τμηματική ομοιότητα. Πιθανώς το αντικείμενο A να είναι τμηματικά όμοιο με το B, λόγω ενός συνόλου όμοιων διατάσεων x (που αντιπροσωπεύουν π.χ. το μέγεθος και τον προσανατολισμό του αντικειμένου) και τμηματικά όμοιο με το αντικείμενο Γ λόγω του συνόλου διαστάσεων y (που π.χ. αντιπροσωπεύουν το χρώμα και το σχήμα). Στις περισσότερες περιπτώσεις, τα x και y είναι αρκετά διαφορετικά,

ούτως ώστε το B να μην είναι τμηματικά όμοιο με το Γ. Συνεπώς, αν χρησιμοποιηθεί ένα δέντρο, η γειτνίαση ανάμεσα στα αντικείμενα δε θα δίνει πραγματική πληροφορία για την ομοιότητα τους, καθιστώντας τη χρήση του δέντρου άσκοπη. Εάν λοιπόν κανείς σκοπεύει να δεικτοδοτήσει πολυδιάστατα αντικείμενα για την απεικόνιση και υποστήριξη των τμηματικών τους διαφορών, θα πρέπει πρώτα απ' όλα να εξαλείψει αυτό το πρόβλημα. Αρχικά αυτό δε φαίνεται εύκολα εφικτό, αλλά μετά από αρκετή μελέτη, εστίασαμε στην παρακάτω παρατήρηση η οποία είναι ικανή να οδηγήσει σε κάποια διέξοδο: *Οι τμηματικές ομοιότητες δεν επηρεάζονται/καθορίζονται από όλες τις διαστάσεις των αντικειμένων, αλλά από υποσύνολά τους. Αυτά τα σύνολα όμως δεν είναι τυχαία. Αντιστοιχούν σε διαστάσεις οι οποίες περιγράφουν ένα συγκεκριμένο χαρακτηριστικό του αντικειμένου. Συνεπώς, η μεταβατική ιδιότητα δεν ισχύει για την τμηματική ομοιότητα αλλά ισχύει για την ομοιότητα συγκεκριμένων συνόλων διατάσεων.*

Για να εφαρμοστεί αυτό το συμπέρασμα, πρέπει αρχικά να οριστούν τα υποσύνολα διαστάσεων που αντιστοιχούν σε κάθε χαρακτηριστικό των αντικειμένων. Για παράδειγμα, έστω το αντικείμενο A με:

d1	d2	d3	d4	d5	d6	d7	d8	d9
3	6	7	8	12	3	4	55	34

Πίνακας 2.6.: Αντικείμενο A

και τα προκαθορισμένα υποσύνολα διαστάσεων που αντιστοιχούν σε κάθε χαρακτηριστικό, όπως φαίνονται παρακάτω:

Προσανατολισμός (Π)	Μέγεθος (Μ)	Χρώμα (Χ)
d1- d3	d4-d7	d8-d9

Πίνακας 2.7.: Προκαθορισμένα υποσύνολα διαστάσεων

Με βάση τα παραπάνω, το αντικείμενο υποδιαιρείται σε τρία υποσύνολα (Π_A, M_A, X_A) τα οποία αντιστοιχούν στα προκαθορισμένα υποσύνολα διαστάσεων, σχηματίζοντας τα παρακάτω:

Π_A			M_A				X_A	
d1	d2	d3	d4	d5	d6	d7	d8	d9
3	6	7	8	12	3	4	55	34

Πίνακας 2.8.: Τα Υπο-Αντικείμενα του A

Αυτά τα Π_A , M_A και X_A συνιστούν ένα σύνολο νέων αντικειμένων (τα υπο-αντικείμενα), καθένα από τα οποία αντιπροσωπεύει ένα χαρακτηριστικό του A. Εάν έχουμε ένα σύνολο

αντικειμένων $A, B, C, \dots$ και N , η αναζήτηση ομοιότητας ανάμεσα στα πρώτα υπο-αντικείμενά τους (προσανατολισμού, $\Pi_A, \Pi_B, \Pi_C, \dots, \Pi_N$) συνιστά ένα πρόβλημα αναζήτησης πλήρους ομοιότητας, όπου ισχύει η μεταβατική ιδιότητα. Συνεπώς, μπορεί να χρησιμοποιηθεί ένα ιεραρχικό δέντρο για την οργάνωση αυτών των υπο-αντικειμένων και πιο συγκεκριμένα, ένα τέτοιο δέντρο για κάθε υποσύνολο διαστάσεων (Π , M και X). Λαμβάνοντας υπ'όψιν ότι ασχολούμαστε κυρίως με σύνολα δεδομένων με λιγότερες από 100 διαστάσεις, γίνεται φανερό πως αφού τα υπο-αντικείμενα αντιστοιχούν σε τμήματα των αρχικών αντικειμένων, ο αριθμός των διαστάσεων για κάθε υποσύνολο διαστάσεων δεν μπορεί να είναι μεγάλος. Από την άλλη μεριά, έχει αποδειχτεί [A15, A17] πως το R^* -tree (βλ. Παράρτημα κεφαλαίου) είναι ένα από τα καλύτερα ιεραρχικά δέντρα για χώρους λίγων διαστάσεων (low-dimensional), χωρίς να προϋποτίθενται περιορισμοί σχετικά με το είδος της συνάρτησης ομοιότητας που χρησιμοποιείται. (π.χ. μία συνάρτηση για την οποία να ισχύει η τριγωνική ανισότητα, όπως είναι η Ευκλείδεια). Με βάση αυτό το συλλογισμό, *προτείνουμε το R^* -tree για την υποστήριξη των παραγόμενων υπο-αντικειμένων (τα οποία στο παράδειγμά μας ονομάζονται $(\Pi)R^*$ -tree, $(M)R^*$ -tree και $(X)R^*$ -tree)*. Παρόλα αυτά, δεδομένου ότι υπάρχουν πολλά ιεραρχικά δέντρα και μέθοδοι [A15] που αποδίδουν πολύ καλά σε χώρους λίγων διαστάσεων, μπορεί να χρησιμοποιηθεί οποιοδήποτε από αυτά, ανάλογα με την προτίμηση και εκτίμηση που υπάρχει σχετικά με τις ανάγκες της εκάστοτε εφαρμογής. *Η πρόθεσή μας δεν είναι η επιλογή ή η πρόταση ενός συγκεκριμένου δέντρου αλλά η ανεύρεση και η πρόταση του πλαισίου το οποίο θα καταστήσει εφικτή τη χρήση δέντρων στο πρόβλημα αναζήτησης τμηματικής ομοιότητας.*

Η οργάνωση των υπο-αντικειμένων σε κάθε δέντρο, προϋποθέτει τη χρήση και εφαρμογή μίας μεθόδου ανεύρεσης πλήρους ομοιότητας, η οποία θα είναι κατάλληλη για το επιλεγμένο είδος δέντρου.

Εδώ ακριβώς βρίσκεται και το σημείο το οποίο μας επέστρεψε πίσω στο αρχικό πρόβλημα, το πρόβλημα ανεύρεσης τμηματικών ομοιοτήτων. Κι αυτό γιατί από την παραπάνω ανάλυση προκύπτει το εξής συμπέρασμα:

Η ανεύρεση τμηματικής ομοιότητας μπορεί να επιτευχθεί μέσω ενός συνόλου απλών αναζητήσεων πλήρους ομοιότητας ανάμεσα στα υπο-αντικείμενα τα οποία αντιστοιχούν σε κάθε συγκεκριμένο χαρακτηριστικού των αντικειμένων.

Προκειμένου να δημιουργηθεί μία καλύτερη ιδέα σχετικά με το πώς μπορεί να χρησιμοποιηθεί αυτό το συμπέρασμα, έστω το παρακάτω παράδειγμα: Έστω τα πολυδιάστατα αντικείμενα A, B, C, D, E, F, G και H τα οποία χωρίζονται (αποτελούνται από) στα υπο-αντικείμενα Π_x, M_x και X_x (x είναι ένα από τα $A, B, C, \dots, N$). Κάθε ένα από αυτά τα υπο-αντικείμενα έχει ένα συγκεκριμένο κωδικό, οπότε απαιτείται ένα βοηθητικό αρχείο όπου θα φυλάσσεται η πληροφορία συσχετισμού του κωδικού του αρχικού αντικειμένου με κείνους των υπο-αντικειμένων του (Πίνακας 2.9.). Τα υπο-αντικείμενα οργανώνονται στα αντίστοιχα R^* -trees ($(\Pi)R^*$ -tree [Σχήμα 2.1.], $(M)R^*$ -tree [Σχήμα 2.2.] και $(X)R^*$ -tree [Σχήμα 2.3.]) αντιπροσωπεύοντας και αντανακλώντας τις μεταξύ τους ομοιότητες. *Αυτά τα δέντρα μπορούν να χρησιμοποιηθούν ατομικά ή και σε συνδυασμό μεταξύ τους.* Πιο συγκεκριμένα:

- ♦ *Ατομικά:* Είναι φανερό πως πολλές φορές είναι δυνατόν να αναζητούνται αντικείμενα που παρουσιάζουν ομοιότητα σε ένα χαρακτηριστικό (π.χ. το χρώμα). Δεδομένου ότι έχει απομονωθεί κάθε χαρακτηριστικό σε ένα R^* -tree, η αναζήτηση αυτού του είδους καθίσταται ιδιαίτερα εύκολη, με τη χρήση του αντίστοιχου αλγόριθμου πλήρους ομοιότητας (ChooseSubtree)[A17]. Καλούμε αυτό το είδος ομοιότητας *ατομική ομοιότητα (individual similarity)*.

- ♦ Σε συνδυασμό μεταξύ τους: Με τη χρήση του αλγόριθμου *MultiTreeSimilarity* είναι δυνατό να βρεθούν τα αντικείμενα που είναι όμοια σε ένα δεδομένο αντικείμενο (q). Πριν από αυτό, ο υπεύθυνος της εκάστοτε εφαρμογής πρέπει να ορίσει το ποσοστό χαρακτηριστικών p , το οποίο πρέπει να παρουσιάζει ομοιότητα προκειμένου να θεωρηθούν δύο αντικείμενα όμοια.

	Κωδικός Αντικείμενο	Κωδικός Π-αντικειμένου	Κωδικός Μ-αντικειμένου	Κωδικός Χ-αντικειμένου
A	1	10	20	30
B	2	11	21	31
C	3	12	22	32
D	4	13	23	33
E	5	14	24	34
F	6	15	25	35
G	7	16	26	36
H	8	17	27	37

Πίνακας 2.9.: Το Βοηθητικό Αρχείο Συσχέτισης Κωδικών

Τα βασικά βήματα του αλγορίθμου *MultiTreeSimilarity* είναι:

- *Αναζήτηση στα δέντρα:* Χρησιμοποιώντας τον αλγόριθμο *ChooseSubtree*[A17] πραγματοποιείται αναζήτηση σε κάθε δέντρο, προκειμένου να ανακτηθεί ένα σύνολο υπο-αντικειμένων (Π, Μ και Χ στο παράδειγμά μας) τα οποία είναι όμοια με τα αντίστοιχα υπο-αντικείμενα του q .
- *Δημιουργία των συνόλων των αρχικών αντικειμένων:* Βασισμένοι στο βοηθητικό αρχείο συσχέτισης κωδικών, καθορίζονται τα αντίστοιχα σύνολα αρχικών αντικειμένων (Π_o , M_o και X_o).
- *Σχηματισμός της απάντησης,* με τη χρήση του προκαθορισμένου ποσοστού p και των αποτελεσμάτων του προηγούμενου βήματος.

Σχήμα 2.1.: Προσανατολισμός (Π)R*-tree

Σχήμα 2.2.: Μέγεθος (M)R*-tree

Σχήμα 2.3.: Χρώμα (X)R*-tree

Ας ‘τρέξουμε’ τώρα τον αλγόριθμο για το παράδειγμά μας, με τυχαία – αλλά λογική – p , ίση με 66%, που αντιστοιχεί σε δύο από τα τρία πολυδιάστατα χαρακτηριστικά του αντικειμένου. Έχοντας τρία αντικείμενα ερώτησης (query objects) – έστω τα q_1 , q_2 και q_3 – και θεωρώντας ότι τα σύνολα Π_o , M_o και X_o που επιστράφηκαν ως απάντηση από το δεύτερο βήμα του αλγορίθμου είναι τα επόμενα:

	q_1	q_2	q_3
Π	C,D	G,H	A,B
M	H	A,C	A,C
X	C,D	F	A,B

Πίνακας 2.10.: Πίνακας αποτελεσμάτων

τα συμπερασματικά αποτελέσματα είναι:

- ✓ Τα αντικείμενα C και D είναι όμοια με το q_1 , εφόσον 66% των χαρακτηριστικών τους (προσανατολισμός και χρώμα) είναι όμοια. Το H είναι απλά όμοιο με το q_1 (ατομική ομοιότητα) στο μέγεθος και θα μπορούσε να αποτελεί απάντηση εάν η εφαρμογή αναζητούσε ανάλογη ομοιότητα.

Μελετώντας το (M)R*-tree στο Σχήμα 2.2., μπορεί κανείς να παρατηρήσει ότι τα αντικείμενα D και C διαφέρουν αρκετά από το H, το οποίο είναι όμοιο στο μέγεθος με το q_1 , επομένως τα D και C διαφέρουν αρκετά από το q_1 στο μέγεθος. Συνεπώς,

εάν πραγματοποιούνταν μία κλασσική αναζήτηση ομοιότητας (π.χ. με υπολογισμό απόστασης) δε θα ήταν δυνατό να εντοπιστεί η τμηματική ομοιότητα ανάμεσα στα αντικείμενα C, D και q1, λόγω της μεγάλης τους διαφοράς στο μέγεθος.

- ✓ Το αντικείμενο q2 είναι όμοιο με τα αντικείμενα G και H στον προσανατολισμό, με τα αντικείμενα A και B στο μέγεθος και με το αντικείμενο F στο χρώμα, οπότε δεν υπάρχει καμία ‘στέρεη’ ομοιότητα. Κάθε ένα από αυτά τα αντικείμενα μοιάζει με το q2 σε ποσοστό 33%. Εδώ, όπως και στην προηγούμενη περίπτωση, η μέθοδος INTESIS θα επέστρεφε απάντηση ατομικής ομοιότητας, μόνο εάν η εφαρμογή το απαιτούσε.
  - ✓ Στην περίπτωση του αντικειμένου q3, παρατηρείται ότι τα αντικείμενα A και B είναι όμοια με το q3 κατά τουλάχιστον 66%, οπότε μπορούν να θεωρηθούν όμοιά του. Το αντικείμενο C παρουσιάζει μόνο ατομική ομοιότητα στο μέγεθος, οπότε το θέμα της απάντησης της μεθόδου θα είχε το χειρισμό που αναφέρθηκε και παραπάνω.
- Σε αυτή την τρίτη περίπτωση, παρατηρείται ένα ακόμη πολύ ενδιαφέρον σημείο: Το αντικείμενο A είναι όμοιο με q3 σε όλα τα χαρακτηριστικά του, πράγμα που παραπέμπει σε ‘ισχυρή’ πλήρη ομοιότητα και συνιστά την αφετηρία για την ανάλυση της επόμενης ενότητας.

2.3.2. Ολοκλήρωση Τμηματικής και Πλήρους Ομοιότητας

Όταν ο αλγόριθμος MultiTreeSimilarity επιστρέφει ένα ή περισσότερα αντικείμενα ως όμοια σε όλα τα χαρακτηριστικά με το δεδομένο αντικείμενο(α), είναι φανερό πως στην ουσία έχει ανιχνευτεί πλήρης ομοιότητα. Το γεγονός αυτό επιτρέπει την επέκταση της μεθόδου INTESIS σε μία ολοκληρωμένη μέθοδο, όπου μπορούν να ανιχνευτούν *ατομική, τμηματική καθώς και πλήρης ομοιότητα*. Σε αυτό το σημείο εγείρεται ένα εύλογο ερώτημα: Εφόσον τις τελευταίες δεκαετίες έχουν προταθεί πολλές αποδοτικές μέθοδοι ανεύρεσης και δεικτοδότησης πλήρους ομοιότητας, ποιά μπορεί να είναι η συνεισφορά της INTESIS σ’ αυτό το πεδίο; Γιατί να χρησιμοποιηθεί αυτή η μέθοδος για την πλήρη ομοιότητα;

Η απάντηση είναι πως η INTESIS δεν προτείνεται σαν λύση για την ανεύρεση πλήρους ομοιότητας και δεν εστιάζει σ’ αυτό το είδος. Αντίθετα ασχολείται κατά κύριο λόγο με τις τμηματικές και ατομικές ομοιότητες. Εφόσον όμως η έξοδος του αντίστοιχου αλγορίθμου MultiTreeSimilarity, μπορεί να αξιοποιηθεί και για τον εντοπισμό πλήρους ομοιότητας ταυτόχρονα και σχεδόν χωρίς καθόλου κόστος - απλά περιλαμβάνεται στα αποτελέσματα, όταν υπάρχει - αυτό αποτελεί ένα σπουδαίο παράπλευρο όφελος. Επιπλέον μ’ αυτό τον τρόπο, παρέχεται ένα ολοκληρωμένο πλαίσιο αναζήτησης και οργάνωσης ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα.

2.3.3. Αξιολόγηση – Συμπεράσματα

Στην κλασσική αναζήτηση πλήρους ομοιότητας ανάμεσα σε πολυδιάστατα αντικείμενα χρησιμοποιείται μόνο ένα δέντρο για την οργάνωση των αντικειμένων. Στον αντίποδα, η INTESIS απαιτεί τη δημιουργία και τη συντήρηση περισσότερων δέντρων. Εφόσον κάτι τέτοιο μπορεί να μειώσει πολύ την απόδοση, γεννιέται το ερώτημα: *πόσο αποδοτική μπορεί να είναι η μέθοδος;* Για να απαντηθεί αυτή η ερώτηση πρέπει να ληφθεί υπ’ όψιν το πλαίσιο στο οποίο αναφερόμαστε, το οποίο είναι η τμηματική ομοιότητα ανάμεσα σε πολυδιάστατα αντικείμενα, η οποία όπως περιγράφηκε παραπάνω είναι πολύ πιο απαιτητική από την αντίστοιχη πλήρη ομοιότητα. Αφού οι υπάρχουσες δενδρικές δομές και μέθοδοι δεν μπορούν να αποτελέσουν λύση και επιλογή για την οργάνωση των τμηματικά όμοιων αντικειμένων, είναι εύλογο και αναμενόμενο ότι εάν υπάρχει κάποια δομή για την υποστήριξη αυτού του

‘δύσκολου’ είδους, θα είναι πιο ‘ακριβή’ από τα κλασσικά ιεραρχικά δέντρα. Λαμβάνοντας υπ’ όψιν την ανάλυση των προηγούμενων ενοτήτων, υποστηρίζουμε πως η μέθοδος INTESIS είναι αρκετά αποδοτική, διότι:

⇒ Ο πολυδιάστατος χώρος μετατρέπεται σε ένα μικρό σύνολο χώρων λίγων διαστάσεων, οδηγώντας σε χαμηλούς χρόνους απόκρισης και στη διαδικασία εύρεσης πλήρων ομοιοτήτων αλλά και στην οργάνωση των R*-trees. Πιο συγκεκριμένα, όπως έχει αποδειχτεί από προηγούμενες μελέτες [A4, A15, A17], η απόδοση του R*-tree αυξάνεται όσο μειώνεται ο αριθμός των διαστάσεων, εξασφαλίζοντας υψηλές αποδόσεις για κάθε υπο-δέντρο. Εφόσον στο πλαίσιο μας τα υποσύνολα διαστάσεων περιλαμβάνουν μικρό αριθμό διαστάσεων και ο αριθμός τους – και συνεπώς και ο αριθμός των αντίστοιχων δέντρων – είναι μικρός, η συνολική απόδοση μπορεί να παραμείνει υψηλή.

Όπως έχει αποδειχτεί [A4], υπάρχει ένα μοντέλο για την πρόβλεψη της απόδοσης των R*-trees και των παραλλαγών/εκδοχών τους, το οποίο εισάγει ένα είδος ‘σταθερών’ στη διαδικασία αξιολόγησης της μεθόδου μας. Συνεπώς, αφού κατά την INTESIS έχουμε έναν (μικρό) αριθμό n από R*-trees και ένα μέσο χρόνο t για την αναζήτηση σε κάθε δέντρο, ο συνολικός απαιτούμενος χρόνος αναζήτησης για τον αλγόριθμο MultiTreeSimilarity είναι σχεδόν⁵ $n*t$.

Έχοντας ως δεδομένο ότι η αναζήτηση σε ένα δέντρο είναι ανεξάρτητη από την αναζήτηση σε ένα άλλο δέντρο, αυτές οι αναζητήσεις μπορούν να πραγματοποιηθούν ταυτόχρονα/παράλληλα. Κατά συνέπεια, σε πολυεπεξεργαστικά (multiprocessing) συστήματα ο χρόνος απόκρισης του βασικού αλγόριθμου μπορεί να μειωθεί σε t . Αφού αυτό το t είναι αρκετά μικρό – εφόσον τα δέντρα αναφέρονται σε χώρους λίγων διαστάσεων – η απόδοση της INTESIS είναι πολύ ικανοποιητική. Με άλλα λόγια, αυτό που επιτυγχάνεται σε ένα πολυεπεξεργαστικό σύστημα, είναι η πτώση του χρόνου απόκρισης ενός απαιτητικού είδους αναζήτησης ομοιότητας, σε τιμές που αναλογούν σε απλές αναζητήσεις πλήρους ομοιότητας σε χώρους λίγων διαστάσεων.

⇒ Αυτό το πλαίσιο ελεγχόμενης ‘διαίρεσης προβλήματος’ (problem division) έχει αποδειχτεί ιδιαίτερα αποδοτικό όχι μόνο από τη γνωστή θεωρία ‘διαίρει και βασίλευε’ (divide and conquer) αλλά και από πρόσφατα σχετικά πεδία και μελέτες [A36].

⇒ Η μέθοδος INTESIS ανακαλύπτει ομοιότητες διαφορετικού εύρους (ατομικές, τμηματικές και πλήρεις), χρησιμοποιώντας μόνο έναν απλό αλγόριθμο (MultiTreeSimilarity).

Για τον έλεγχο των παραπάνω εκτιμήσεων, πραγματοποιήθηκαν κάποια βασικά πειράματα, χρησιμοποιώντας δύο σύνολα δεδομένων, των 100 αντικειμένων το καθένα. Το A είχε 9 διαστάσεις ενώ το B 16. Το A διαιρέθηκε σε 3 υποσύνολα υπο-αντικειμένων ενώ το B σε 5, οδηγώντας σε 3 και 5 R*-trees αντίστοιχα.

Το σύστημα ήταν ενός επεξεργαστή και ο χρόνος απόκρισης ήταν πράγματι ένα απλό και μικρό πολλαπλάσιο (περίπου ίσο με τον αριθμό των δέντρων κάθε φορά) του αντίστοιχου χρόνου αναζήτησης σε ένα R*-tree. Από την άλλη μεριά ο απαιτούμενος χώρος δεν ακολούθησε τον πολλαπλασιαστικό κανόνα, γιατί αν και είχαμε τρία ή πέντε R*-trees αντίστοιχα, υπήρχε μόνο μία 100% και 270% (περίπου) επιβάρυνση, αντίστοιχα. Η

⁵ Σχεδόν, γιατί η αναζήτηση στο R*-tree (πρώτο βήμα) αποτελεί το κύριο μέρος του αλγορίθμου ενώ τα άλλα δύο βήματα απαιτούν ελάχιστο αναλογικά χρόνο.

θεωρητική εξήγηση αυτού του γεγονότος βρίσκεται στο ότι τα φύλλα των δέντρων που σχηματίζουν το πολυπληθέστερο επίπεδο δέντρου, αποτελούνται από τα υπο-αντικείμενα, όχι από τα ίδια τα αντικείμενα. Έτσι, εάν ενώσουμε κατάλληλα τα φύλλα όλων των R^* -trees, θα δημιουργηθεί το σύνολο των αντικειμένων. Με άλλα λόγια, η επιβάρυνση δημιουργείται από τα επίπεδα πριν από τα φύλλα. Παρόλα αυτά αυτή η επιβάρυνση δεν είναι ικανή να υποβιβάσει την αποτελεσματικότητα και αποδοτικότητα της μεθόδου μας γιατί:

- α) καμία άλλη μέθοδος δεν επιτυγχάνει εύρεση και δεικτοδότηση τμηματικών ομοιοτήτων, έτσι ώστε να υπάρχει δυνατότητα σύγκρισης
- β) στην εποχή μας τέτοιες επιβαρύνσεις χώρου δεν αποτελούν ‘πραγματική’ επιβάρυνση, λόγω του ότι υπάρχουν τεράστιες δυνατότητες χώρου με πολύ χαμηλό κόστος.

2.4. Δημοσιεύσεις του Κεφαλαίου

♦ ΣΥΝΕΛΠΙΑ

[Σ16] Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, “Integrated Similarity Search Among Multi-Dimensional Objects (InteSiS)”, IKE 2009 (ranking:0,80).

[Σ18] Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, “Improving the Results of Partial Match”, short paper, DMIN 2009 (ranking:0,79).

2.5. Αναφορές του Κεφαλαίου

- [A1] Overmars, M.H.: Efficient Data Structures for Range Searching on a Grid. In Journal of Algorithms 9, pp.254-275, 1988.
- [A2] Mehlhorn, K., Tsakalidis, A.: Chapter 6: Data Structures. In Handbook of Theoretical Computer Science, Editor J. V. Leeuwen, Elsevier Science Publishers, co-published by MIT-Press, 1990.
- [A3] Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Searching. In SIGMOD 1984.
- [A4] Theodoridis Y., Sellis T., A Model for the Prediction of R-tree Performance, PODS 1996, pp. 161-171.
- [A5] Chen, Q., Chen, L., Lian, X., Liu, Y., Xu, Yu.: Indexable PLA for Efficient Similarity Search. In Very Large Data Bases (VLDB) 2007.
- [A6] Ciaccia, P., Patella P., Zezula P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In Very Large Data Bases (VLDB) 1997.
- [A7] Hellerstein, J.M., Jeffrey F. Naughton, J.F., Pfeffer, A.: Generalized Search Trees for Database Systems. In Very Large Data Bases (VLDB) 1995.
- [A8] Hu H., Lee D.L., Lee V.S.: Distance Indexing on Road Networks. In Very Large Data Bases (VLDB) 2006.
- [A9] Kibriya, A., Frank, Eibe.: An Empirical Comparison of Exact Nearest Neighbour Algorithms. In PKDD 2007.
- [A10] Nievergelt, J., Hinterberger, H., Servik, K.: The grid file: An adaptable symmetric multikey file structure. ACM Transactions on Database Systems, 1984.
- [A11] Sakurai Y., Yoshikawa M., Uemura S., Kojima H.: The A-tree: An Index Structure for High-Dimensional Spaces Using Relative Approximation. In Very Large Data Bases (VLDB) 2000.
- [A12] Tung, A.K.H, Zhang, R., Koudas, N., Ooi, B.G.: Similarity Search: A Matching Based Approach. In Very Large Data Bases (VLDB) 2006.
- [A13] Beyer, K., Goldstein, J., Shaft, U.: When is Nearest Neighbors Meaningful? In ICDT 1999.
- [A14] Hinneburg, A., Aggarwal, C., Keim, D.: What is the nearest neighbor in high dimensional spaces? In VLDB 2000.
- [A15] Weber, R., Schek, H.J, Blott, S.: A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In Very Large Data Bases (VLDB) 1998.

- [A16] Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search - The Metric Space Approach, Springer, 2006.
- [A17] Beckmann, N., Kriegel, H.-P., Schneider, R., Seeger, B.: The R*-tree: An Efficient and Robust Access method for Points and Rectangles. In SIGMOD 1990.
- [A18] Seeger, B., Kriegel, H.-P.: The Buddy Tree: An Efficient and Robust Access Method for Spatial Data Base Systems. In Very Large Data Bases (VLDB) 1990.
- [A19] Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases. In SIGMOD 1994.
- [A20] Kamel, I., Faloutsos, C.: Hilbert R-tree: An Improved R-tree Using Fractals. In Very Large Data Bases (VLDB) 1994.
- [A21] Ciaccia, P., Patella, M.: The M2-tree: Processing Complex Multi-Feature Queries with Just One Index. In 1st DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland, 2000.
- [A22] Aggarwal, C.C.: Towards meaningful high-dimensional nearest neighbor search by human-computer interaction. In ICDE 2002.
- [A23] Arge, L., Mark de Berg, Haverkort, H.J., Yi K.: The Priority R-Tree: A Practically Efficient and Worst-Case Optimal R-Tree. In SIGMOD 2004.
- [A24] Augsten, N., Bohlen, M., Gamper, J.: An Incrementally Maintainable Index for Approximate Lookups in Hierarchical Data. In Very Large Data Bases (VLDB) 2006.
- [A25] Chen, Q., Chen, L., Lian, X., Liu, Y., Xu, Yu.: Indexable PLA for Efficient Similarity Search. In Very Large Data Bases (VLDB) 2007.
- [A26] Wong, R.C.-W., Tao, Y., Fu, A.W.-C., Xiao, X.: On Efficient Spatial Matching. In Very Large Data Bases (VLDB) 2007.
- [A27] White, D.A. Jain, R.: Similarity indexing with the SS-tree. In 12th International Conference on Data Engineering (ICDE) 1996.
- [A28] Goh, K.-S., Li, B., Chang, E.: DynDex: A Dynamic and Non-metric Space Indexer. In Multimedia 2002.
- [A29] Qian, G., Zhu, Q., Xue, Q., Pramanik, S.: The ND-Tree: A Dynamic Indexing Technique for Multidimensional Non-ordered Discrete Data Spaces. In Very Large Data Bases (VLDB) 2003.
- [A30] Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A., Theodoridis, Y.: R-Trees: Theory and Applications, Springer 2005. ISBN 1-85233-977-2.
- [A31] Berchtold, S., Keim, D., Kriegel, H.-P.: The X-tree: An Index Structure for High-Dimensional Data. In Very Large Data Bases (VLDB) 1996.
- [A32] Katayama, N., Satoh, S.: The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. In SIGMOD 1997.
- [A33] Bronstein, A., Bronstein, M., Bruckstein, A., Kimmel, R.: Partial Similarity of Objects, or How to Compare a Centaur to a Horse, International Journal of Computer Vision, 2008.
- [A34] Bruckstein, A. M., Katzir, N., Lindenbaum, M., & Porat, M.: Similarity-invariant signatures for partially occluded planar shapes. *IJCV* 1992, 7(3), 271–285.
- [A35] Kim, C. R., & Chung, C. W.: XMage: An image retrieval method based on partial similarity. *Information Processing and Management*, 2006, 42, 484–502.
- [A36] Chakrabarti K., Mehrotra S.: Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces, VLDB 2000.

2.7. Παράρτημα 2^ο Κεφαλαίου – R*-trees

Τα R-trees αποτελούν μία οικογένεια δέντρων τα οποία έχουν τα βασικά γνωρίσματα των κλασσικών ισοζυγισμένων δέντρων (δέντρα για τα οποία ισχύει ότι για κάθε κόμβο το αριστερό του υποδέντρο διαφέρει από το δεξί του υποδέντρο κατά το πολύ ένα κόμβο). Η διαφορά τους έγκειται στο ότι χρησιμοποιούνται για μεθόδους χωρικής προσπέλασης, π.χ. δεικτοδότηση πολυδιάστατων δεδομένων. Για παράδειγμα, έστω οι γεωγραφικές συντεταγμένες (X,Y). Μια κοινή πραγματική χρήση ενός R-tree θα μπορούσε να είναι: 'Βρες όλα τα θέατρα σε ακτίνα 1 χιλιομέτρου από το σημείο'.

Για να φιαχτεί ένα R-tree (Σχήμα 2.4.) ο χώρος χωρίζεται σε ιεραρχικά εμφωλευμένα και πιθανά επικαλυπτόμενα MBRs (Minimum Bounding Rectangles – ελάχιστα ορθογώνια οριοθέτησης⁶). Κάθε κόμβος ενός R-tree μπορεί να περιλαμβάνει ένα μεταβλητό αριθμό εγγραφών (μέχρι ένα προκαθορισμένο όριο), ενώ κάθε κόμβος (εκτός των φύλλων) περικλείει δύο τμήματα: μία κωδικοποίηση των παιδιών του κόμβου και έναν προσδιορισμό του ορθογωνίου οριοθέτησης στο οποίο αντιστοιχεί κάθε παιδί.

Ο σκοπός του δέντρου είναι η απεικόνιση της γειτνίασης των χωρικών δεδομένων, πράγμα που σημαίνει πως γειτονικά αντικείμενα-δεδομένα πρέπει να βρίσκονται στον ίδιο ή γειτονικό κόμβο (MBR). Αυτή η αρχή καθορίζει και τις λειτουργίες εισαγωγής και διαγραφής στο δέντρο. Η λειτουργία της αναζήτησης από την άλλη χρησιμοποιεί τα ορθογώνια οριοθέτησης προκειμένου να αποφασιστεί εάν υπάρχει ή όχι λόγος να γίνει αναζήτηση σε έναν κόμβο. Μ' αυτό τον τρόπο, οι περισσότεροι κόμβοι δεν 'αγγίζονται' καν κατά την αναζήτηση, κερδίζοντας έτσι πολύ χρόνο. Αυτό κάνει τα R-trees κατάλληλα για Βάσεις Δεδομένων, όπου οι κόμβοι μπορούν να ακολουθήσουν τη σελιδοποίηση της μνήμης.

Σχήμα 2.4.: Παράδειγμα ενός R-tree για 2D δεδομένα (ορθογώνια)

⁶ Στο γράμμα R του MBR οφείλεται το όνομα των R-trees.

Από το 1984 που ο Guttman παρουσίασε το R-tree[A3], έχουν προταθεί πολλές παραλλαγές του με σκοπό τη βελτίωσή του σε κάποιο χαρακτηριστικό το οποίο επηρεάζει την απόδοσή του. Το R*-tree [A17] αποτελεί μία από τις σημαντικότερες παραλλαγές του, με κύριο χαρακτηριστικό τη μείωση των επικαλύψεων των MRBs. Αυτό έχει σαν αποτέλεσμα τη βελτίωση της απόδοσης του δέντρου - και ιδιαίτερα σε 'δύσκολες' κατανομές των δεδομένων στο χώρο - με μικρή μόνο επιβάρυνση στο κόστος υλοποίησής του.

Σύμφωνα με την [A17] στην οποία γίνεται μία συγκριτική μελέτη ανάμεσα σε μεθόδους αναζήτησης ομοιότητας αντικειμένων σε πολυδιάστατους χώρους, το R*-tree έχει αξιολογηθεί ως ένα από τα αποδοτικότερα δέντρα(μεθόδους) αναζήτησης για χώρους μέχρι δέκα διαστάσεων.

Να κάνεις τα πράγματα όσο πιο απλά γίνονται,
αλλά όχι απλούστερα.

Άλμπερτ Αϊνστάιν

Κεφάλαιο 3

Υβριδική Μέθοδος Αναζήτησης Ομοιοτήτων Ανάμεσα σε Αντικείμενα, Ανεξάρτητη από το Πλήθος των Διατάσεών τους (Μέθοδος DIGENIS)

[Σ15]

Περιεχόμενα Κεφαλαίου

- 3.1. Το Πρόβλημα (dimensionality curse)- Υπαρκτές Μέθοδοι (State of the Art) και Σύγκρισή τους
- 3.2. Μέθοδος DIGENIS
  - 3.2.1. Δέντρο DIGENIS (DIGENIS-Tree)
    - 3.2.1.1. Αναζήτηση Πλησιέστερων Γειτόνων (k-NN Queries) στο Δέντρο Digenis
    - 3.2.1.2. Ενημέρωση Δέντρου Digenis
  - 3.2.2. Αρχείο DIGENIS (DIGENIS-File)
    - 3.2.2.1 Τοποθέτηση στο Αρχείο Digenis
    - 3.2.2.2 Ενημέρωση του Αρχείου Digenis
  - 3.2.3. Αξιολόγηση της Μεθόδου
    - 3.2.3.1. Θεωρητική Αξιολόγηση
    - 3.2.3.2. Πειραματική Αξιολόγηση
- 3.3. Ανακεφαλαίωση – Ανοικτά θέματα
- 3.4. Δημοσιεύσεις του Κεφαλαίου
- 3.5. Αναφορές του Κεφαλαίου
- 3.7. Παράρτημα 3^ο Κεφαλαίου – VA-Files

Σύμβολο ή Έννοια	Έννοια - Σημασία
fl (family limit)	Προκαθορισμένο επιτρεπτό όριο απόστασης για αντικείμενα της ίδιας οικογένειας
Object family	Οικογένεια Αντικειμένων
Lt (limit tolerance)	Προκαθορισμένο επιτρεπτό όριο απόστασης για συγγενικά (kindred) αντικείμενα
ft (family tolerance)	Η μέγιστη απόσταση ανάμεσα σε αντικείμενο της οικογένειας και τον αντιπρόσωπό της
ch	Πλήθος παιδιών για κάθε γονέα (εκτός φύλλων)
d(i,j)	Απόσταση (Ευκλείδεια) ανάμεσα στον κόμβο i και j
Fi	Οικογένεια i
q	Δοθέν αντικείμενο του ερωτήματος
dp	Απόσταση ενός παιδιού από το γονέα του
pf	Ο γονέας ή ο αντιπρόσωπος της οικογένειας στην οποία τελικά ανήκει το q
D	Η διαφορά του $d(d(\rho\acute{\iota}\zeta\alpha, q))$ από το άθροισμα όλων των ενδιάμεσων αποστάσεων ανάμεσα στους κόμβους που συνιστούν το μονοπάτι ανάμεσα στο pf και στη ρίζα.
lp	Το πρώτο επίπεδο για το οποίο ισχύει $D \leq Lt$ στο δέντρο
v	Σταθερή τιμή απόστασης ανάμεσα σε γονιό και παιδί (υποθετική και αδύνατη)
lev1	$d \text{ DIV } Lt$: η ασφαλέστερη τιμή ορίου επιπέδου δέντρου από το οποίο και μετά είναι δυνατόν να βρίσκεται η πιθανή θέση του q
c	Ο κόμβος για τον οποίο το q έχει τη μικρότερη απόσταση σε σχέση με τους υπόλοιπους του επιπέδου, απόσταση η οποία είναι μεγαλύτερη και από τα ft του επιπέδου και από το Lt.
reci	Η i-οστή θέση εγγραφής η οποία υπολογίζεται μέσω του αλγόριθμου τοποθέτησης στο αρχείο Digenis
tl	Υποθετικός αριθμός επιπέδων δέντρου
N	Αριθμός αντικειμένων εφαρμογής
M	Αριθμός οικογενειών στις οποίες ομαδοποιούνται οι N οικογένειες
DistCalc	Αριθμός υπολογισμένων αποστάσεων, για το επίπεδο lev1 (ή $\lceil tl/2 \rceil$ για τη μέση περίπτωση ή tl-1 για τη χειρότερη περίπτωση)
TotalDistCalc	Συνολικός αριθμός υπολογισμών αποστάσεων
f	Μέσος αριθμός αντικειμένων ανά οικογένεια
P	Ποσοστό υπολογισμών αποστάσεων

Πίνακας 3.1.: Πίνακας Συμβολισμών

Σύνοψη

Στο ερευνητικό πεδίο της ανεύρεσης ομοιοτήτων σε πολυδιάστατους χώρους έχουν προταθεί πολλές νέες μέθοδοι (κυρίως δέντρα ομοιοτήτων), οι περισσότερες από τις οποίες βασίζονται στην κατανομή δεδομένων και χώρου, με σοβαρότερο μειονέκτημα την καθοριστική εξάρτηση της απόδοσής τους από το πλήθος των διαστάσεων του εκάστοτε χώρου. Πιο συγκεκριμένα, ενώ για μικρό πλήθος διαστάσεων η απόδοσή τους είναι υψηλή, όσο αυτό το πλήθος αυξάνεται τόσο η απόδοσή τους πέφτει, ενώ από ένα όριο και πάνω (10 διαστάσεις) [A8] οι εν λόγω μέθοδοι υπολείπονται ακόμη και της σειριακής αναζήτησης. Αυτό το φαινόμενο ονομάζεται ‘κατάρα των πολλών διαστάσεων’ (*‘dimensional curse’*) και με τα σημερινά δεδομένα, έχει αντιμετωπιστεί κυρίως από το απλό διανυσματικό προσεγγιστικό σχήμα VA-File [A8], το οποίο χαρακτηρίζεται από αντίστροφη απόδοση σε σχέση με τις πρώτες μεθόδους, αφού η απόδοσή του αυξάνεται αναλογικά με το πλήθος των διαστάσεων του χώρου.

Έχοντας αυτές τις δύο κατηγορίες μεθόδων ως πραγματικότητα στο χώρο της αναζήτησης ομοιοτήτων σε πολυδιάστατους χώρους, εγείρεται το εξής θέμα: εφόσον και στις δύο περιπτώσεις η απόδοση εξαρτάται (με ευθύ ή αντίστροφο τρόπο) από το πλήθος των διαστάσεων των αντικειμένων, ούτε τα δέντρα αλλά ούτε και τα VA-Files μπορούν να χρησιμοποιηθούν για οποιοδήποτε πλήθος διαστάσεων, ως μέθοδοι πολλαπλού σκοπού (*multi-purpose*). Αυτή ακριβώς η παρατήρηση στάθηκε η αρχή της ερευνητικής πρόκλησης για τη δημιουργία μίας μεθόδου πολλαπλού σκοπού, η οποία θα αποτελεί έναν ευέλικτο συνδυασμό δέντρου και αρχείου τύπου VA-File, ‘παντρεύοντας’ την εκφραστική και οργανωτική ισχύ των δέντρων ομοιοτήτων με τα οφέλη των απαλλαγμένων από την ‘κατάρα των πολλών διαστάσεων’ VA-Files. Το αποτέλεσμα αυτής της μελέτης είναι η πρόταση *Digenis*¹, η οποία περιλαμβάνει ένα δέντρο *Digenis* για την οργάνωση των αντικειμένων σε εννοιολογικό επίπεδο (*conceptually*) και ένα αρχείο *Digenis* – βασισμένο στο δέντρο *Digenis* – για την τμηματική αναζήτηση σε φυσικό επίπεδο (*physical level*).

Λέξεις Κλειδιά: ‘Κατάρα των πολλών διαστάσεων’ (*‘dimensional curse’*), Δεικτοδότηση ανεξάρτητη από το πλήθος των διαστάσεων (*Dimension-Independent Indexing*), Δέντρο *Digenis* (*Digenis-tree*), Αρχείο *Digenis* (*Digenis-File*), Οικογένεια Αντικειμένων (*object family*), Αντιπρόσωπος Οικογένειας (*family representative*), Αρχείο Απομακρυσμένων (*remote file*).

3.1. Το Πρόβλημα (dimensionality curse)- Υπαρκτές Μέθοδοι (State of the Art) και Σύγκρισή τους

Μελετώντας κανείς το θέμα της ανεύρεσης ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα, διαπιστώνει πως κάποια – λίγα - συγκεκριμένα θέματα και μέθοδοι μονοπωλούν το ερευνητικό ενδιαφέρον. Πιο συγκεκριμένα, αυτά που κυρίως διερευνώνται είναι η πλήρης ομοιότητα (*full similarity*), το πρόβλημα του κοντινότερου γείτονα (*nearest neighbor – NN*) και τα ερωτήματα εύρους (*range queries*). Οι αντίστοιχες ομοιότητες ‘μετρώνται’ με συγκεκριμένες μετρικές αποστάσεων (κυρίως την Ευκλείδεια απόσταση), ενώ την οργάνωση των όμοιων αντικειμένων αναλαμβάνει κάποιο από τα δέντρα ομοιοτήτων ή τις μεθόδους δεικτοδότησης που έχουν προταθεί για αυτό το σκοπό [A2, A5, A9, A13, A17, A19, A20, A21, A22, A23, A24, A25]. Τα δέντρα αυτά βασίζονται είτε σε κατανομή δεδομένων (*R-tree* [A9, A20], *R⁺-tree* [A20], *R*-tree* [A20], *X-tree* [A21], *SR-*

¹ Η επιλογή του συγκεκριμένου ονόματος, οφείλεται στον υβριδικό χαρακτήρα της μεθόδου.

tree[A22], *M-tree*[A2], *TV-tree*[A23], *ND-tree*[A19], *SS-tree*[A17], *Priority-tree*[A13], *hB-tree*[A11]) είτε σε κατανομή χώρου (*grid-file*[A5], *K-D-B-tree*[A24], *quad-tree*[A25]) (*data-space partitioning*). Όπως όμως έχει αποδειχτεί [A8], και οι δύο κατηγορίες δέντρων παρουσιάζουν μεγάλη απόδοση για μικρό πλήθος διαστάσεων, ενώ όσο αυτό το πλήθος αυξάνεται τόσο η απόδοσή τους πέφτει (από ένα όριο και πάνω (10 διαστάσεις) [A8] οι εν λόγω μέθοδοι υπολείπονται ακόμη και της σειριακής αναζήτησης). Αυτό το φαινόμενο ονομάζεται ‘κατάρτα των πολλών διαστάσεων’ (dimensional curse) και με τα σημερινά δεδομένα, έχει αντιμετωπιστεί κυρίως από το απλό διανυσματικό προσεγγιστικό σχήμα *VA-File*[A8], το οποίο χαρακτηρίζεται από αντίστροφη απόδοση σε σχέση με τις πρώτες μεθόδους, αφού η απόδοσή του είναι μικρή για μικρό πλήθος διαστάσεων και αυξάνεται όσο αυτό το πλήθος μεγαλώνει (δηλ. αναλογικά με το πλήθος των διαστάσεων του χώρου).

Δεδομένου λοιπόν ότι ούτε τα δέντρα αλλά ούτε και τα *VA-Files* μπορούν να χρησιμοποιηθούν ανεξάρτητα από το πλήθος των διαστάσεων των εμπλεκόμενων αντικειμένων, επιχειρήσαμε να ορίσουμε μία νέα μέθοδο η οποία θα αποδίδει ανεξάρτητα από το πλήθος των διαστάσεων των δεδομένων.

3.2. Μέθοδος DIGENIS

Με στόχο τον ορισμό μίας μεθόδου αναζήτησης ομοιοτήτων ανεξάρτητης από το πλήθος των διαστάσεων των δεδομένων, κάναμε την αρχική σκέψη να συνδυάσουμε τα πλεονεκτήματα των δέντρων και των *VA-Files* για το σχηματισμό μίας νέας δομής. Ο συλλογισμός μας ήταν ο εξής: Το βασικό προσόν του δέντρου είναι η εκφραστική του δύναμη η οποία είναι κοντά στην αντίληψη του ανθρώπου. Από την άλλη πλευρά, το *VA-File* είναι μία πολύ απλή δομή η οποία οργανώνει τα δεδομένα με περισσότερο πρακτικό τρόπο, με βάση μία λογική ‘πρώτου επιπέδου’ (low level). Αν και αυτές οι δύο μέθοδοι φαίνονται αρκετά ασύμβατες, μπορούμε να τις ‘παντρέψουμε’, χρησιμοποιώντας:

- ♦ αφενός την εκφραστική δύναμη των δέντρων για τον αρχικό σχεδιασμό και την αναπαράσταση της εκάστοτε κατάστασης σε επίπεδο ομοιοτήτων των αντικειμένων και
- ♦ αφετέρου ένα αρχείο τύπου *VA-File* βασισμένο στην οργάνωση του δέντρου.

Μετά από πολύ μελέτη με αφετηρία αυτό το συλλογισμό, καταλήξαμε στην πρόταση *Digenis*, η οποία αποτελείται από ένα δέντρο, το δέντρο *Digenis* (*Digenis-tree*) και ένα αντίστοιχο αρχείο, το αρχείο *Digenis* (*Digenis-File*). Αρχικά, το δέντρο *Digenis* χρησιμοποιείται για την οργάνωση των πολυδιάστατων αντικειμένων, σύμφωνα με τις Ευκλείδειες αποστάσεις τους. Όπως θα φανεί κατά την ανάλυση παρακάτω, είναι πολύ αποτελεσματικό να ομαδοποιηθούν τα αντικείμενα σε ‘οικογένειες’ πολύ όμοιων αντικειμένων, οι οποίες οργανώνονται με ένα συγκεκριμένο τρόπο στο δέντρο *Digenis*. Αυτό το δέντρο είναι παραμετροποιημένο και εφόσον λειτουργεί ως δέντρο ομοιοτήτων, δεν αποφεύγει την ‘κατάρτα των πολλών διαστάσεων’. Για να ξεπεράσουμε αυτό το ελάττωμα, δε χρησιμοποιούμε τις κλασσικές σελιδοποιημένες και μη (paged, not-paged) υλοποιήσεις των δέντρων, αλλά προτείνουμε μία αντιμετώπιση τύπου *VA-File*, το αρχείο *Digenis*. Συγκεκριμένα, προτείνουμε τη δημιουργία αρχείων, των οποίων οι προσεγγίσεις και οι θέσεις θα βασίζονται στα δέντρα *Digenis*. Με άλλα λόγια, το συνολικό σενάριο στην πρότασή μας είναι το εξής:

Αρχικά οι οικογένειες αντικειμένων οργανώνονται στο δέντρο με ένα συγκεκριμένο τρόπο. Στη συνέχεια αυτή η οργάνωση, με την κατάλληλη επεξεργασία, αποτελεί ένα είδος εισόδου (input) για το αρχείο, το οποίο δομείται με τον κατάλληλο τρόπο ώστε να εξυπηρετεί μία έξυπνη προσέγγιση οργάνωσης και ‘συμπίεση’ των αντικειμένων.

Παράλληλα, έξυπνες τεχνικές αναλαμβάνουν την επιτάχυνση της όλης διαδικασίας και κυρίως της ανάκτησης όμοιων δεδομένων.

Ο σκοπός της μελέτης σε αυτό το στάδιο ορισμού της πρότασης Digenis ήταν η διερεύνηση του πώς θα μπορούσε να διενεργηθεί η αναζήτηση ομοιοτήτων σε ένα σκηνικό σαν αυτό που προτείνεται μέσω της Digenis. Συνεπώς, δε θα εστιάσουμε στους αλγόριθμους κατασκευής του δέντρου Digenis.

3.2.1. Δέντρο DIGENIS (DIGENIS-Tree)

Ο βασικός στόχος κατά το σχεδιασμό του δέντρου μας ήταν η επίτευξη μιας πρώτης ‘συμπίεσης’ και οργάνωσης των δεδομένων. Έτσι καταλήξαμε στον ορισμό δύο συνόλων ομοιοτήτων:

- Στα πολύ όμοια αντικείμενα (*very similar objects*) - με μεταξύ τους αποστάσεις μικρότερες από ένα χαμηλό προκαθορισμένο όριο (fl) – τα οποία μπορούν να σχηματίσουν μία οικογένεια (*object family*) που θα αντιπροσωπεύεται από έναν κόμβο του δέντρου.
- Στα υπόλοιπα όμοια αντικείμενα, οι συγγενικές οικογένειες (*kindred families*), τα οποία απέχουν μεταξύ τους κατά αποστάσεις μικρότερες από ένα προκαθορισμένο όριο (Lt). Αυτή η ομοιότητα μπορεί να υποδηλωθεί απλά από την κλασσική σχέση παιδιού-γονέα όπως καταγράφεται στα δέντρα.

Πιο συγκεκριμένα, δημιουργήθηκαν *οικογένειες αντικειμένων* (*object families*), οι οποίες είναι οι τυπικοί κάτοικοι των κόμβων του δέντρου. Εφόσον ο κόμβος της οικογένειας δεν μπορεί να φιλοξενεί όλα τα μέλη της, χρησιμοποιείται το *αντικείμενο-αντιπρόσωπος* (*representative object*). Αυτό το αντικείμενο θα είναι ένα μέσο αντικείμενο της οικογένειας, δηλ. το πιο ‘κεντραρισμένο’ μέλος της και εάν δεν υπάρχει τέτοιο, θα δημιουργείται ένα νέο το οποίο θα επιλέγεται ως συμβατικός αντιπρόσωπος της οικογένειας. Η μέγιστη απόσταση ανάμεσα στον αντιπρόσωπο και ένα αντικείμενο-μέλος της οικογένειας, συνιστά την *απόκλιση-ανοχή της οικογένειας* (*family tolerance*) η οποία αποτελεί σημαντική πληροφορία σχετικά με το ‘εύρος’ της οικογένειας, αξιοποιήσιμη σε επόμενα στάδια εκτίμησης ομοιότητας. Για να εισαχθεί-υπαχθεί ένα αντικείμενο σε μία οικογένεια, πρέπει η απόστασή του από τον αντιπρόσωπο της οικογένειας να είναι μικρότερη από το προκαθορισμένο όριο απόστασης fl , το οποίο θα ορίζεται από τον υπεύθυνο της εκάστοτε εφαρμογής. Κάθε οικογένεια έχει ένα σύνολο ch παιδιών, τις *συγγενικές οικογένειες* (*kindred families*) (Σχήμα 3.2.). Οι τελευταίες είναι οι ch οικογένειες (ch -πλησιέστερες, ch -nearest) των οποίων οι αντιπρόσωποι κείνται πλησιέστερα στον αντιπρόσωπο της οικογένειας-πατέρας, σε σχέση με τους άλλους αντιπροσώπους των υπόλοιπων οικογενειών. Η έννοια του ‘πλησίον’ (*close*) θα οριοθετείται πάντα από ένα προκαθορισμένο όριο Lt ($Lt > fl$), προκειμένου να εξασφαλίζεται ένας ρεαλισμός στην αντιπροσώπευση της γειτνίασης (*vicinity*). Αυτό το όριο μπορεί να οριστεί στατικά από τον υπεύθυνο της εκάστοτε εφαρμογής ή δυναμικά, χρησιμοποιώντας στατιστικά στοιχεία-αποτελέσματα από τις ίδιες τις εφαρμογές. Στην ειδική περίπτωση που ο αριθμός των οικογενειών με απόσταση μικρότερη από Lt είναι μικρότερος από ch , τότε τοποθετούνται στο δέντρο κενοί κόμβοι-οικογένειες (βλ. περισσότερα παρακάτω). Κάθε κόμβος-παιδί συσχετίζεται με το γονιό του με μία ακμή που περιλαμβάνει τη μεταξύ τους απόσταση. Κάθε κόμβος-οικογένεια περιλαμβάνει : i) το αντικείμενο αντιπρόσωπο της οικογένειας και ii) τη μέγιστη απόσταση (ανοχή οικογένειας, *family tolerance*, fl) ανάμεσα στον αντιπρόσωπο και ένα μέλος της οικογένειας.

Η βασική ιδέα στην οποία στηρίζεται ο ορισμός του δέντρου Digenis, είναι ότι υπάρχει ένα σύμπαν αντικειμένων με πολλά χαρακτηριστικά, τα οποία μπορούν να αντιμετωπιστούν ως

πολυδιάστατα και έχουν συγκεκριμένες αποστάσεις μεταξύ τους, οι οποίες υπολογίζονται βάση κάποιων από τις γνωστές συναρτήσεις ομοιότητας – την Ευκλείδεια απόσταση στην πρότασή μας. Εάν ‘απεικονίσουμε’ αυτή την κατάσταση, θα δημιουργήσουμε ένα γράφο κόμβων οι οποίοι συνδέονται με ακμές που περιλαμβάνουν τις αντίστοιχες αποστάσεις (Σχήμα 3.1.). Προκειμένου να ‘κινήθουμε’ σ’ αυτό το γράφο προσπαθώντας κάθε φορά να εντοπίσουμε αντικείμενα όμοια με κάποιο δεδομένο και να δημιουργήσουμε τελικά το δέντρο μας, θα χρησιμοποιήσουμε την ‘ακτινωτή’ λογική. Πιο συγκεκριμένα, αρχικά θα επιλεγεί μία ρίζα, η οποία θα είναι το αντικείμενο που βρίσκεται στο κέντρο της πολυδιάστατης περιοχής². Οι ch κοντινότεροι γείτονες (συγγενικές οικογένειες) της οικογένειας-ρίζας θα γίνουν τα ch παιδιά της στο δέντρο, αποτελώντας το δεύτερο δενδρικό επίπεδο. Η ρίζα ‘δείχνει’ τους κόμβους-οικογένειες-παιδιά της με ακμές οι οποίες περιλαμβάνουν τη μεταξύ τους απόσταση (d_i). Οι οικογένειες αδέρφια οργάνωνονται-τοποθετούνται στο δέντρο ιεραρχικά (με αύξουσα σειρά), σύμφωνα με την απόστασή τους από το γονέα. Για κάθε μία από αυτές τις οικογένειες, δημιουργείται το σύνολο των ch κόμβων-παιδιών (από τις συγγενικές οικογένειες του γράφου), κ.ο.κ. Κατά τη διάρκεια αυτής της διαδικασίας, πιθανώς να προκύπτει η απορία σχετικά με το τι συμβαίνει με τις οικογένειες που μπορούν να είναι παιδιά περισσότερων του ενός κόμβου-οικογένειας, στο ίδιο επίπεδο. Η απάντηση είναι ότι η επίμαχη οικογένεια τοποθετείται στο δέντρο ως παιδί της οικογένειας με την οποία έχει τη μικρότερη απόσταση (από τον αντιπρόσωπό της). Αυτή η επιλογή οδηγεί στο παρακάτω συμπέρασμα:

Συμπέρασμα: Κάθε οικογένεια η οποία έχει απόσταση dr από το γονέα της, δεν μπορεί να απέχει από τα αδέρφια του γονέα της, λιγότερο από dr .

Αυτό το συμπέρασμα θα αποβεί πολύ χρήσιμο στη διαχείριση ερωτημάτων εύρους (range queries) παρέχοντας ένα ισχυρό εργαλείο απ’ευθείας αποκλεισμών κατά την αναζήτηση ομοιοτήτων.

Σύμφωνα με τα παραπάνω, δεν μπορεί να υπάρχει καμία άλλη αμφιβολία σχετικά με το ποιος θα είναι ο πατέρας ενός κόμβου, αφού ξεκινώντας από την επιλεγμένη ρίζα, το δέντρο χτίζεται επίπεδο-επίπεδο. Συνεπώς εάν ένας κόμβος συνδέεται με περισσότερους από έναν στο γράφο (που είναι και το πιο συνηθισμένο) θα τοποθετηθεί στο δέντρο, σύμφωνα με το πέρασμα που γίνεται ανά επίπεδο. Για παράδειγμα, στο Σχήμα 3.1. η οικογένεια F7 συνδέεται με έξι κόμβους (συγγενικούς) ως ρίζα του δέντρου έχει επιλεγεί η F8. Η F7 γίνεται παιδί της ρίζας από το πρώτο ήδη πέρασμα, λόγω της απόστασης 1 που έχει από αυτή. Στη συνέχεια, αφού το ch είναι ίσο με 3, οι τρεις κοντινότεροι από τους ελεύθερους συγγενικούς της κόμβους γίνονται παιδιά της. Αν κανείς απορεί γιατί η F15 (με απόσταση 2) και όχι η F9 (με απόσταση 1,5) δεν εντάσσεται στο σύνολο παιδιών της F7 παρόλο που απέχει λιγότερο από την τελευταία, η απάντηση προκύπτει από τους κανόνες κατασκευής του δέντρου, αφού η F9 σύμφωνα με αυτούς έχει ήδη ενταχθεί στα παιδιά της ρίζας από το προηγούμενο πέρασμα. Όσο για την F2, δε συνδέεται άμεσα με την F7 γιατί ούτε γονιός της θα μπορούσε να είναι – αφού προηγείται σ’ αυτό η ρίζα – αλλά ούτε και παιδί της γιατί είναι η τέταρτη συγγενική της – κατά αύξουσα σειρά - οικογένεια και το ch είναι μόνο 3 για το συγκεκριμένο δέντρο.

² Με κάποιο τρόπο που θα συζητηθεί στα πλαίσια άλλης εργασίας.

Σχήμα 3.1: Παράδειγμα Γράφου με $Lt=3$. Οι ακμές αντιπροσωπεύουν τις αποστάσεις ανάμεσα στις οικογένειες. Για λόγους απλότητας και πρακτικότητας, έχουν παραλειφθεί οι πολύ μεγάλες αποστάσεις (μεγαλύτερες από 3, στη συγκεκριμένη περίπτωση) αφού δεν οδηγούν σε συνδέσεις στο παρακάτω δέντρο.

Σχήμα 3.2. : Το δέντρο Digenis, με βάση το γράφο του σχήματος 1, με $ch=3$.

Σχήμα 3.3. : Γράφος με αραιές περιοχές, με $Lt=3$ (και εδώ έχουν παραληφθεί οι αποστάσεις μεγαλύτερες του 3)

Σχήμα 3.4. Τα δέντρα Digenis, βασισμένα στο γράφο του σχήματος 3.3., με $ch=3$ για το πρώτο δέντρο και $ch=4$ για το δεύτερο.

Συνεπώς, προσπαθώντας να συνοψίσουμε τα παραπάνω σε ένα συμπαγές συμπέρασμα-κανόνα, καταλήγουμε στο ότι *μία οικογένεια B με απόσταση d από την οικογένεια A, με $d \leq Lt$, μπορεί να συμπεριλαμβάνεται στο σύνολο οικογενειών-παιδιών (συγγενών-kindred) της οικογένειας A, μόνο*

- ο *εάν δεν έχει καταστεί παιδί οικογένειας σε προηγούμενο επίπεδο και*
- ο *εάν καμία άλλη οικογένεια που βρίσκεται στο ίδιο επίπεδο με την A δεν έχει απόσταση από τη B μικρότερη από d και*
- ο *δεν υπάρχουν τουλάχιστον ch οικογένειες που έχουν απόσταση από την A μικρότερη από d και μπορούν να γίνουν παιδιά της.*

Κατά τη μελέτη του ορισμού του δέντρου Digenis, μας απασχόλησε ιδιαίτερα ο χειρισμός των ‘αραιών περιοχών’. Ειδικότερα, όταν υπάρχουν μεγάλα κενά – στην εικόνα του γράφου – ανάμεσα στα πολυδιάστατα αντικείμενα όπως φαίνεται στο γράφο του Σχήματος 3.3., τα αντικείμενα δεν μπορούν να αντιπροσωπευτούν αποδοτικά στο δέντρο, γιατί συγκεκριμένα σύνολα αντικειμένων-οικογενειών δεν μπορούν να συνδεθούν με οποιαδήποτε άλλη οικογένεια του δέντρου (π.χ. καμία από τις οικογένειες F12 έως F16 καθώς και η F17, δε γίνεται να συνδεθούν με κάποια από τις F1 έως F10). Εάν προσπαθούσαμε να συνδέσουμε αυτές τις ξεκάθαρα απομακρυσμένες οικογένειες, τότε το Lt θα έπρεπε να αυξηθεί υπερβολικά. Σε μια τέτοια περίπτωση, το Lt δε θα αντιπροσώπευε μία απόσταση που υποδηλώνει ομοιότητα, οπότε η όλη μέθοδος δε θα ήταν αποτελεσματική. Για να ξεπεραστούν αυτά τα προβλήματα, υιοθετήσαμε την εξής λύση:

Όταν υπάρχουν κενές περιοχές ανάμεσα σε πολυδιάστατα αντικείμενα (οικογένειες), δημιουργώντας απομονωμένες περιοχές αντικειμένων, χτίζουμε διαφορετικά δέντρα Digenis (με το δικό του ch το καθένα), ένα για κάθε απομονωμένη περιοχή. Έτσι, στο Σχήμα 3.4. έχουμε δημιουργήσει δύο δέντρα Digenis, ένα για κάθε περιοχή, με $ch=3$ και $ch=4$, αντίστοιχα. Μελετώντας αυτά τα δέντρα, διαπιστώνουμε πως το σύνολο παιδιών της F4 είναι αρκετά φτωχό, αφού δύο από τις τρεις οικογένειες-παιδιά είναι κενές. Γενικά, για να αποφύγουμε τη σπατάλη, αποφασίσαμε να ‘κόβουμε’ το αντίστοιχο υπο-δέντρο του δέντρου, δημιουργώντας ένα νέο μικρότερο και στη συνέχεια να ξαναοργανώνουμε το αρχικό δέντρο.

Παρόλα αυτά - όπως είναι φανερό στο παράδειγμα του Σχήματος 3.4. - αυτή η αντιμετώπιση δεν αξίζει τον κόπο σε κάθε περίπτωση κενών κόμβων-παιδιών. Αντίθετα είναι αποτελεσματική μόνο όταν το αφαιρούμενο υπο-δέντρο έχει ένα αξιόλογο μέγεθος³. Διαφορετικά, είναι προτιμότερο να διατηρούνται οι κενοί κόμβοι. Πρόκειται για μια μικρή επιβάρυνση χώρου, η οποία όμως - όπως θα φανεί παρακάτω - επιστρέφει ως αντάλλαγμα μελλοντική ευελιξία και μεγάλη διευκόλυνση στην εφαρμογή απλών και γρήγορων αλγορίθμων για την υποστήριξη της αποδοτικής αναζήτησης όμοιων πολυδιάστατων αντικειμένων.

Επιστρέφοντας στο παράδειγμα των κενών κόμβων του Σχήματος 3.4. - οι οποίοι βρίσκονται στο τελευταίο επίπεδο (φύλλα) - είναι φανερό πως δεν αξίζει να ανεξαρτητοποιήσουμε το φτωχό υπο-δέντρο της F4. Επιπλέον, στις περισσότερες περιπτώσεις ο κενός 'άδικα' δεσμευμένος χώρος των κενών κόμβων θα μειωθεί σταδιακά λόγω της εισαγωγής νέων οικογενειών στο δέντρο, κάποιοι από τους οποίους θα είναι συγγενικοί των γονέων των κενών κόμβων και συνεπώς θα γίνουν παιδιά τους, γεμίζοντας τους κενούς κόμβους. Αυτό αποτελεί ένα πρόσθετο πλεονέκτημα του δέντρου Digenis, αφού καθιστά τη συντήρησή του ευκολότερη και φθηνότερη.

Εκτός από τις παραπάνω περιπτώσεις, υπάρχει και η περίπτωση των απομακρυσμένων κόμβων-οικογενειών, όπως είναι η F17 στο Σχήμα 3.3. Γενικά, *ονομάζουμε απομακρυσμένες οικογένειες, τις οικογένειες που δεν μπορούν να είναι παιδιά καμίας άλλης*. Σε αυτές τις περιπτώσεις φυσικά και δε δημιουργείται κάποιο δέντρο για τις απομακρυσμένες οικογένειες, αλλά ορίζουμε τα *αρχεία απομακρυσμένων (remote files)*, τα οποία χρησιμοποιούνται συμπληρωματικά του ή των δέντρων Digenis και περιλαμβάνουν όλες τις απομακρυσμένες οικογένειες της εφαρμογής.

Συνοψίζοντας, με βάση τα παραπάνω, στην πρόταση Digenis, σε εννοιολογικό επίπεδο, *ένας αριθμός από διάφορα, μικρότερα ή μεγαλύτερα δέντρα Digenis θα αντιπροσωπεύει τις σχέσεις ομοιότητας που υπάρχουν ανάμεσα στα πολυδιάστατα αντικείμενα, ενώ ένα Αρχείο Απομακρυσμένων θα φιλοξενεί τα απομακρυσμένα αντικείμενα*.

3.2.1.1. Αναζήτηση Πλησιέστερων Γειτόνων (k-NN Queries) στο Δέντρο Digenis

Σε αυτή την ενότητα, θα μελετήσουμε την πρότασή μας σε σχέση με τον πιο κοινό τύπο αναζήτησης ομοιοτήτων, την αναζήτηση των k-πλησιέστερων γειτόνων (k-NN). Γενικά, όταν χρησιμοποιείται ένα δέντρο ομοιοτήτων, το πρόβλημα της αναζήτησης των k-πλησιέστερων γειτόνων για ένα δοθέν αντικείμενο q, αποτελείται από δύο υπο-προβλήματα: α) την *Τοποθέτηση (Positioning)*, δηλ. την ανεύρεση της θέσης στο δέντρο όπου το q θα μπορούσε να τοποθετηθεί, εάν έπρεπε να εισαχθεί σ' αυτό και β) την *k-NN Ανάκτηση (k-NN retrieval)*, την ανάκτηση δηλαδή των k πλησιέστερων γειτόνων του q. Αρχικά εστιάζουμε τη μελέτη μας στο πρώτο υπο-πρόβλημα. Το δεύτερο έχει μετατοπιστεί στα πλαίσια μεταγενέστερης μελέτης, χωρίς να αφήνει πραγματικά κενά στην παρούσα έρευνα, δεδομένου ότι η εάν λυθεί το υπο-πρόβλημα της τοποθέτησης, η k-NN ανάκτηση είναι σχετικά απλή. Σε μία πρώτη χονδρική προσέγγιση, οι k γείτονες του q - όταν το τελευταίο 'έχει βρει' τη θέση του στο δέντρο - είναι τα k-κοντινότερά σ' αυτό, μέλη της οικογένειας

³ Γενικά δεν είναι εύκολο να οριστεί ένας γενικός κανόνας σχετικά με το πότε τα υπο-δέντρα έχουν αξιόλογο μέγεθος ώστε να γίνουν ανεξάρτητα. Αυτό θα είναι μία επιλογή και απόφαση του υπεύθυνου της εκάστοτε εφαρμογής. Παρόλα αυτά, βασισμένοι στην εμπειρία και τα σχετικά μας πειράματα, προτείνουμε το παρακάτω: Όταν σε ένα υπο-δέντρο οι κενοί κόμβοι ξεπερνούν το 50% των κόμβων του και τα επίπεδά του (εξαιρουμένης της ρίζας) είναι περισσότερα από το 30% των επιπέδων του αρχικού δέντρου, τότε το υπο-δέντρο πρέπει να ανεξαρτητοποιηθεί (βλ. ενότητα 3.3.) - δημιουργώντας ένα νέο δέντρο Digenis - ενώ το αρχικό δέντρο να αναδιοργανωθεί.

στην οποία ανήκει. Εάν αυτό το k είναι μεγαλύτερο από το πλήθος των μελών της οικογένειάς του, ένας απλός αλγόριθμος θα ψάξει για τους υπόλοιπους γείτονες, στο γονιό, στα παιδιά και πιθανώς στα 'αδέρφια' και στα αδέρφια του γονέα της οικογένειας του q .

Με άλλα λόγια, το θέμα σ' αυτή την ενότητα είναι η διαδικασία *τοποθέτησης του q* , η οποία έχει ως αποτέλεσμα την ανεύρεση της κατάλληλης θέσης για το q στο δέντρο και όχι την εισαγωγή του σ' αυτή τη θέση. Αυτό σημαίνει πως σ' αυτό το στάδιο δεν ενδιαφερόμαστε για την ενημέρωση(update) του δέντρου.

Η λογική λοιπόν της διαδικασίας *τοποθέτησης για το δέντρο Digenis*, είναι η ακόλουθη:

Το δοθέν αντικείμενο q , συγκρίνεται με τη ρίζα του δέντρου, δηλαδή υπολογίζεται η απόσταση d ανάμεσα στον αντιπρόσωπο της οικογένειας της ρίζας και στο q . Εάν ισχύει ότι $d \leq fl$, τότε σημαίνει πως το αντικείμενο ανήκει στην οικογένεια της ρίζας και η διαδικασία σταματά εδώ. Αλλιώς, θα διεξαχθεί ένα απλό άλμα, ανάλογα με τα επίπεδα που θα εκτιμηθεί ότι μπορούν να υπερπηδηθούν με μια πρώτη ασφαλή εκτίμηση. Ειδικότερα, θα υπολογιστεί ένα άνω όριο αριθμού επιπέδων, με βάση το οποίο μπορούν να εξαιρεθούν τα επίπεδα και τα αντικείμενα του δέντρου που αναμφίβολα είναι αδύνατο να αποτελούν πιθανές θέσεις του q . Θα υπολογιστεί το $lev1 = (d \text{ DIV } Lt)$ το οποίο – όπως θα αποδειχτεί στο Θεώρημα – υποδεικνύει το πρώτο επίπεδο του δέντρου Digenis για το οποίο υπάρχει πιθανότητα να μπορεί να αποτελεί τη σωστή θέση του q .

Σχήμα 3.5. : Παράδειγμα δέντρου Digenis-Tree με $ch = 3$ και δοθέν αντικείμενο q .

Θεώρημα Digenis: Δοθέντος ενός αντικειμένου q με απόσταση d από τη ρίζα του δέντρου Digenis, κανένα επίπεδο πριν από το επίπεδο $lev1$ ($lev1 = (d \text{ DIV } Lt)$) δεν είναι δυνατόν να μπορεί να αποτελεί σωστή θέση για το q .

Απόδειξη: Το δέντρο Digenis έχει φτιαχτεί χρησιμοποιώντας τις Ευκλείδειες αποστάσεις όπου ισχύει η τριγωνική ανισότητα. Συνεπώς στο Σχήμα 3.5., όπου το q (έστω ότι) έχει βρεθεί να είναι παιδί του F15 και δεδομένου ότι ως $d(Fi, Fj)$ συμβολίζουμε την απόσταση μεταξύ των κόμβων-οικογενειών Fi και Fj , ισχύει το παρακάτω:

$$\begin{aligned} d(F15, q) &\geq |d(F4, q) - d(F4, F15)| \text{ (τριγωνική ανισότητα) } \wedge \\ d(F4, q) &\geq |d(F1, q) - d(F1, F4)| \text{ (τριγωνική ανισότητα) } \Rightarrow \\ d(F15, q) &\geq |d(F1, q) - d(F1, F4) - d(F4, F15)| \\ \text{και εφόσον } d(F1, q) &\geq |d(\text{ρίζα}, q) - d(\text{ρίζα}, F1)| \text{ (τριγωνική ανισότητα) } \Rightarrow \\ d(F15, q) &\geq |d(\text{ρίζα}, q) - d(\text{ρίζα}, F1) - d(F1, F4) - d(F4, F15)| = D \end{aligned}$$

Με άλλα λόγια, η απόσταση ανάμεσα στο δοθέν αντικείμενο q και στο γονιό ή στον αντιπρόσωπο της οικογένειας στην οποία θα πρέπει να υπάγεται (έστω pf) είναι μεγαλύτερη από την D που ισούται με : $d(d(\text{ρίζα}, q))$ πλην το άθροισμα όλων των ενδιάμεσων αποστάσεων ανάμεσα στους κόμβους που συνιστούν το μονοπάτι ανάμεσα στο pf και στη ρίζα. Διασχίζοντας το δέντρο Digenis, η D παίρνει διαφορετικές τιμές. Για τα επίπεδα όπου κάθε D είναι μεγαλύτερη από το Lt ($D > Lt$), ισχύει η ανισότητα $d(Fi, q) \geq D > Lt$ ($\forall i$ στο

επίπεδο), οπότε δεν υπάρχει καμία πιθανότητα ομοιότητας ανάμεσα στο q και σε κάποιο κόμβο αυτών των επιπέδων. Για το πρώτο επίπεδο (έστω lp) όπου η απόσταση D ανάμεσα σε κάποιο κόμβο του και στο q γίνεται μικρότερη από Lt , μπορούμε να είμαστε σίγουροι ότι το επίπεδο $lp-1$ είναι το τελευταίο επίπεδο όπου κανένας κόμβος δε θα μπορούσε να είναι η σωστή θέση του q (αφού το $lp-1$ είναι το τελευταίο επίπεδο όπου ισχύει η ανισότητα $d(Fi, q) \geq D > Lt$ ($\forall i$ στο επίπεδο)).

Όσο μεγαλύτερες είναι οι ενδιάμεσες αποστάσεις (π.χ. $d(F4, F15)$, $d(F1, F4)$), τόσο γρηγορότερα φτάνουμε στο επίπεδο με την υποψήφια θέση του q . Με άλλα λόγια, όσο μεγαλύτερες οι αποστάσεις, τόσο μικρότερο το lp . Εάν όλες οι ενδιάμεσες αποστάσεις γονέα-παιδιού ήταν ίσες με μια συγκεκριμένη τιμή (έστω v) τότε θα ίσχυε ότι $lp = d \text{ DIV } v$. Φυσικά, δεν είναι δυνατόν να υπάρχει ένα σταθερό συγκεκριμένο v σε πραγματικά δεδομένα αλλά εφόσον η Lt είναι η μέγιστη αποδεκτή απόσταση ανάμεσα στα αντικείμενα διαδοχικών επιπέδων του δέντρου Digenis, το $lev1$ ($d \text{ DIV } Lt$) μπορεί να πάρει τη θέση της χειρότερης περίπτωσης άνω ορίου επιπέδου⁴, το οποίο θέτει τα όρια ασφαλούς αποκλεισμού επιπέδων κατά την αναζήτηση.

Μετά τον ορισμό του επιπέδου $lev1$, εάν αυτό είναι μεγαλύτερο από τον συνολικό αριθμό των επιπέδων του δέντρου, η διαδικασία σταματά, πράγμα που σημαίνει πως δεν υπάρχει κατάλληλη θέση για το q στο δέντρο (απ'ευθείας αποκλεισμός – *direct exclusion*). Διαφορετικά, υπολογίζονται οι αποστάσεις μεταξύ του q και του αντιπροσώπου κάθε κόμβου στο επίπεδο $lev1$ ⁵. Εάν υπάρχει κάποια απόσταση μικρότερη από την αντίστοιχη fl του συγκεκριμένου κόμβου-οικογένειας, τότε αυτός ο κόμβος αποτελεί τη σωστή θέση του q στο δέντρο. Αλλιώς ολοκληρώνεται ο υπολογισμός των αποστάσεων και εάν κάποια από αυτές τις αποστάσεις είναι μικρότερη από το fl τότε η σωστή θέση του q είναι η αντίστοιχη οικογένεια. Διαφορετικά, εάν κάποια ή κάποιες αποστάσεις είναι μικρότερες από το Lt , τότε η σωστή θέση για το q είναι η θέση του παιδιού του κόμβου με τον αντιπρόσωπο του οποίου έχει τη μικρότερη απόσταση. Εάν δε συμβαίνει τίποτα από αυτά, τότε κρατάμε τον κόμβο με τον οποίο το q είχε τη μικρότερη απόσταση, έστω τον c , και θεωρώντας τον τη ρίζα του προς ανίχνευση υπο-δέντρου, γίνονται τα εξής:

- Υπολογίζονται οι Ευκλείδειες αποστάσεις ανάμεσα στο q και σε καθένα από τους ch κόμβους-παιδιά του c .
- Εάν κάποια από αυτές είναι μικρότερη από κάποιο αντίστοιχο fl και Lt , τότε προχωρούμε όπως παραπάνω. Εάν όχι, τότε ο κόμβος με τη μικρότερη απόσταση από το q αποτελεί τη ρίζα του νέου προς διερεύνηση υπο-δέντρου κ.ο.κ.
- Η διαδικασία σταματά όταν βρεθεί η κατάλληλη θέση για το q ή τελειώσει το δέντρο. Στη δεύτερη περίπτωση, δεν υπάρχει κατάλληλη θέση για το q στο δέντρο (όπως και όταν το $lev1$ είναι μεγαλύτερο του πλήθους των επιπέδων του δέντρου, όπως σημειώθηκε παραπάνω).

3.2.1.2. Ενημέρωση Δέντρου Digenis

Όπως θα φανεί και παρακάτω, το δέντρο Digenis ορίστηκε να έχει ένα σταθερό αριθμό ch παιδιών για κάθε κόμβο (εκτός από τα φύλλα), γιατί αυτό διευκόλυνε πολύ τη διαχείριση του αρχείου Digenis. Είναι φανερό πως όταν μία οικογένεια-κόμβος έχει ήδη ch παιδιά-κόμβους

⁴ Εφόσον η Lt παίρνει για παντού τη μέγιστη τιμή της, το $lev1 = d \text{ DIV } Lt$ παίρνει την ελάχιστη τιμή του, υποδεικνύοντας τον ελάχιστο δυνατό αριθμό επιπέδων που μπορούν να αποφευχθούν, γεγονός που αποτελεί πράγματι τη χειρότερη ασφαλή σχετική περίπτωση.

⁵ Εδώ θα μπορούσε να υιοθετηθεί η αντίστοιχη τεχνική που προτείνεται στο επόμενο κεφάλαιο για το δέντρο Chiron, αλλά κρίθηκε καλύτερη η παρούσα, για τους λόγους που θα συζητηθούν στο πέμπτο κεφάλαιο.

και χρειαστεί να εισαχθεί ένας ακόμη κόμβος-παιδί, τότε προκύπτει πρόβλημα υπερχειρίσης παιδιών (children overflow). Σε αυτή την περίπτωση, η οικογένεια με τη μεγαλύτερη απόσταση από τον κόμβο-γονιό απομακρύνεται από το σύνολο των παιδιών του γονιού και μία κατάλληλη διαδικασία αναζητά την επόμενη καλύτερη για εκείνη θέση στο δέντρο.

Η άλλη δύσκολη περίπτωση που σχετίζεται με την οργάνωση του δέντρου Digenis είναι η αντίθεση της υπερχειρίσης παιδιών και συμβαίνει όταν μία περιοχή αντικειμένων-οικογενειών είναι τόσο αραιή, που ο αριθμός των πραγματικών παιδιών κάποιας οικογένειας είναι πάντα πολύ μικρότερος από ch . Η αντιμετώπιση αυτής της περίπτωσης παρουσιάστηκε σε προηγούμενη ενότητα και συνοψίζεται στην κατά περίπτωση και υπό κάποιες συγκεκριμένες συνθήκες διάσπαση του δέντρου στα σημεία όπου υπάρχουν τα πολλά κενά, δημιουργώντας νέα μικρότερα δέντρα Digenis.

3.2.2. Αρχείο DIGENIS (DIGENIS-File)

Το διανυσματικό προσεγγιστικό σχήμα (VA-File)[A8] (βλ. Παράρτημα του Κεφαλαίου) είναι ένας απλός πίνακας συμπαγών γεωμετρικών προσεγγίσεων πολυδιάστατων αντικειμένων, με τη βοήθεια και χρήση του οποίου αποκλείεται σε πρώτη φάση, πολύ μεγάλος αριθμός αντικειμένων από τη διαδικασία αναζήτησης ομοιοτήτων. Το ιδιαίτερο πλεονέκτημα του VA-File, είναι το γεγονός ότι η απόδοσή του αυξάνεται όσο αυξάνεται ο αριθμός των διαστάσεων των εμπλεκόμενων αντικειμένων, σε αντίθεση με τα δέντρα ομοιοτήτων των οποίων η απόδοση φθίνει δραματικά με την αύξηση των διαστάσεων. Αυτός είναι και ο λόγος που – όπως προαναφέρθηκε – χρησιμοποιούμε την ιδέα των VA-Files στην πρότασή μας. Πιο συγκεκριμένα, ορίζουμε ένα αρχείο Digenis για τη δημιουργία και τη διαχείριση συμπαγών προσεγγίσεων των αντικειμένων, οι οποίες στην περίπτωσή μας, δεν είναι γεωμετρικές αλλά βασίζονται στην οργάνωση του δέντρου ομοιοτήτων Digenis. Έχοντας μία οργάνωση των όμοιων αντικειμένων η οποία αναπαριστάται από το δέντρο Digenis, το αρχείο Digenis δημιουργείται ως εξής: Κάθε εγγραφή (record) αποτελείται από έναν αντιπρόσωπο οικογένειας (τις συντεταγμένες του) και την ανοχή της οικογένειας (ft). Η ρίζα του αντίστοιχου δέντρου αντιστοιχεί στην πρώτη εγγραφή του αρχείου, περιλαμβάνοντας τον αντιπρόσωπό της και την αντίστοιχη ft . Οι κόμβοι-παιδιά της ρίζας αντιπροσωπεύονται ένα-ένα από τις επόμενες εγγραφές. Στη συνέχεια ακολουθούν τα στοιχεία των κόμβων του επόμενου επιπέδου κ.ο.κ.

3.2.2.1. Τοποθέτηση στο Αρχείο Digenis

Το πρώτο βήμα στην αναζήτηση της σωστής θέσης του q στη δομή (αρχείο), είναι η σειριακή αναζήτηση του Αρχείου Απομακρυσμένων αντικειμένων. Αυτό το αρχείο το οποίο περιλαμβάνει όλα τα απομακρυσμένα αντικείμενα-οικογένειες, ελέγχεται ξεχωριστά και πλήρως σειριακά, πρώτα από όλα, γιατί εάν το q βρίσκεται κοντά σε κάποια απομακρυσμένη οικογένεια, είναι ανώφελο και δαπανηρό να ελέγξουμε το ή τα δέντρα Digenis. Σε αυτή την περίπτωση, αποφεύγοντας την αναζήτηση στο ή στα δέντρα – όπου δε θα εντοπίζονταν καμία ομοιότητα, έτσι κι αλλιώς – βρίσκουμε τα όμοια αντικείμενα με λίγες μόνο συγκρίσεις στο Αρχείο Απομακρυσμένων. Από την άλλη μεριά, εάν η τελική θέση του q βρίσκεται όντως σε ένα δέντρο, η επιβάρυνση της ανώφελης τελικά αναζήτησης στο Αρχείο Απομακρυσμένων είναι πολύ μικρή λόγω του μικρού γενικά μεγέθους αυτού του αρχείου.

Ο έλεγχος αυτού του Αρχείου Απομακρυσμένων ξεκινά με το διαδοχικό υπολογισμό της απόστασης του q από κάθε αντικείμενο-οικογένεια-εγγραφή του αρχείου. Εάν βρεθεί κάποια απόσταση μικρότερη από το ft της συγκεκριμένης οικογένειας, τότε αυτή η εγγραφή-οικογένεια είναι και η σωστή θέση για το q . Διαφορετικά, ολοκληρώνεται ο υπολογισμός

των αποστάσεων και εάν κάποια απόσταση βρεθεί μικρότερη από το fl τότε η σωστή θέση του q είναι η αντίστοιχη οικογένεια. Αλλιώς, εάν κάποιες από τις αποστάσεις είναι μικρότερες του Lt , τότε το q είναι όμοιο με τα αντικείμενα της οικογένειας με την οποία έχει τη μικρότερη απόσταση. Διαφορετικά η αναζήτηση πρέπει να συνεχιστεί στα αρχεία Digenis. Λαμβάνοντας υπ'όψιν πως ίσως αρχικά έχουν οριστεί περισσότερα από ένα δέντρα Digenis, θεωρούμε πως έχουμε αντίστοιχα και περισσότερα (ίσο αριθμό με τα δέντρα) αρχεία Digenis. Η διαδικασία αναζήτησης ξεκινά από το μεγαλύτερο αρχείο Digenis γιατί αντιστοιχεί στην πιο πυκνή περιοχή αντικειμένων-οικογενειών οπότε η πιθανότητα εύρεσης της σωστής θέσης για το q είναι μεγαλύτερη. Εάν δε βρεθεί τέτοια θέση, η ίδια διαδικασία τρέχει για το επόμενο μεγαλύτερο αρχείο Digenis, κ.ο.κ. Η βασική διαδικασία αναζήτησης στο αρχείο Digenis – η οποία βασίζεται στην ανάλυση του δέντρου Digenis – είναι η παρακάτω:

Πρώτα απ'όλα υπολογίζεται η απόσταση ανάμεσα στην πρώτη εγγραφή του αρχείου Digenis και του q . Εάν είναι μικρότερη ή ίση του fl ($d \leq fl$) τότε το q θα μπορούσε να είναι μέλος της οικογένειας και η διαδικασία σταματά εδώ. Διαφορετικά, αξιοποιώντας τον τρόπο κατασκευής του αρχείου, πρέπει να εκτιμήσουμε την περιοχή του αρχείου όπου βρίσκονται οι υποψήφιες εγγραφές-θέσεις του q . Υπολογίζεται το $lev1 = d \text{ DIV } Lt$ και έτσι αρκετά απλά μπορεί να εκτιμηθεί η θέση της πρώτης εγγραφής για την οποία υπάρχει πιθανότητα να είναι η σωστή θέση του q . Αυτή είναι η πρώτη εγγραφή του επιπέδου $lev1$ και υπολογίζεται από τον παρακάτω τύπο:

$$rec1 = 1 + ch + ch^2 + \dots + ch^{lev1-1} + 1$$

Στη συνέχεια υπολογίζονται οι αποστάσεις του q από την $rec1$ και για τις επόμενες ch^{lev1-1} εγγραφές. Η λογική που ακολουθείται είναι η ίδια με παραπάνω. Δηλαδή, εάν βρεθεί κάποια απόσταση μικρότερη από το fl κάποιας οικογένειας-εγγραφής, τότε αυτή η εγγραφή είναι και η σωστή θέση του q . Αλλιώς, εάν κάποια απόσταση βρεθεί μικρότερη από το fl τότε η σωστή θέση του q είναι η οικογένεια-εγγραφή με την οποία έχει τη μικρότερη απόσταση. Αλλιώς, εάν κάποια ή κάποιες από τις αποστάσεις είναι μικρότερη του Lt , τότε η θέση του q είναι στο σύνολο παιδιών της οικογένειας-εγγραφής με την οποία έχει τη μικρότερη απόσταση. Διαφορετικά, επιλέγεται η οικογένεια c με τη μικρότερη απόσταση από το q (βλ. προηγούμενη ενότητα) - η οποία βρίσκεται έστω στη θέση i του επιπέδου – και ο έλεγχος μεταφέρεται στην εγγραφή $rec2$ η οποία αντιπροσωπεύει το πρώτο παιδί του κόμβου που αντιστοιχεί στην εγγραφή c . Η θέση του $rec2$ υπολογίζεται ως εξής:

$$rec2 = rec1 + ch^{lev1} - i + ((i-1) * ch)$$

Εκεί, για ch διαδοχικές εγγραφές υπολογίζεται η απόστασή τους από το q και η διαδικασία προχωρά όπως παραπάνω. Πιο συγκεκριμένα, εάν βρεθεί μία θέση για το q – με την παραπάνω λογική - η διαδικασία σταματά. Διαφορετικά, βασιζόμενοι στην οικογένεια-εγγραφή με την οποία υπάρχει η μικρότερη απόσταση (έστω στη θέση i) ο έλεγχος προχωρά στην εγγραφή:

$$rec3 = rec2 + ch^{lev1+1} - i + ((i-1) * ch), \text{ κ.ο.κ.}$$

$$(\text{Γενικά } rec_j = rec_{j-1} + ch^{lev1+j-2} - i + ((i-1) * ch))$$

Τελικά η όλη διαδικασία σταματά όταν είτε βρεθεί θέση για το q είτε τελειώσει το αρχείο.

Η συγκεκριμένη χρήση του αρχείου Digenis αναδεικνύει τα οφέλη της προετοιμασίας που ουσιαστικά γίνεται μέσω της ύπαρξης του δέντρου Digenis. Ειδικότερα, στην περίπτωση του αρχείου Digenis, η αποδεδειγμένη ανωτερότητα των VA-Files στους χώρους μεγάλων διαστάσεων ενισχύεται από i) τις προσεγγίσεις που βασίζονται σε ομοιότητες, και επιτυγχάνονται με τον ορισμό και τη χρήση συνόλου (οικογένειας) αντικειμένων και όχι ξεχωριστών και ii) την κατανομή των εγγραφών βάση του δέντρου Digenis, γεγονός που

αποτελεί κωδικοποίηση των διερευνημένων σχέσεων ομοιότητας των εμπλεκόμενων αντικειμένων.

Από την άλλη μεριά, εξαλείφονται τα μειονεκτήματα της οργάνωσης δέντρου ομοιοτήτων για τα πολυδιάστατα αντικείμενα, αφού η αναπαράστασή του ενδυναμώνεται από τα οφέλη της VA-File λογικής.

3.2.2.2. Ενημέρωση του Αρχείου Digenis

Μία λογική ερώτηση σ' αυτό το σημείο, θα αφορούσε το θέμα της συντήρησης του αρχείου Digenis, δεδομένου ότι ο ορισμός του είναι αρκετά στατικός. Πρώτα απ' όλα, όσο αφορά τις εισαγωγές και διαγραφές μελών από μία οικογένεια, αυτό είναι κάτι που μπορεί να διεξαχθεί παράλληλα με το βασικό αλγόριθμο. Πιο συγκεκριμένα, μία διαδικασία αναλαμβάνει να ενημερώνει τη βάση δεδομένων που περιλαμβάνει την πληροφορία σχετικά με την ταυτότητα των μελών μίας οικογένειας. Η άλλη περίπτωση εισαγωγών και διαγραφών αφορά τις εγγραφές του αρχείου, οι οποίες μπορούν να έχουν την κλασική διαχείριση των αρχείων δεικτών, χρησιμοποιώντας μαζική επεξεργασία(batch) ή επεξεργασία πραγματικού χρόνου (real time).

Πέραν όμως των παραπάνω κλασικών ενημερώσεων, πρέπει να ληφθεί μέριμνα για τις ενημερώσεις που αφορούν τις 'σχέσεις' των αρχείων Digenis και του Απομακρυσμένου Αρχείου. Ειδικότερα, κατά τακτά – αλλά όχι ιδιαίτερα κοντινά - χρονικά διαστήματα πρέπει να ελέγχεται εάν οι ανακατατάξεις των αρχείων έχουν δημιουργήσει γειτνιάσεις ανάμεσα στις εγγραφές-οικογένειες των αρχείων Digenis και στα Απομακρυσμένα Αρχεία. Σε αυτή την περίπτωση, θα πρέπει αυτές οι εγγραφές των Απομακρυσμένων Αρχείων να μεταφερθούν στο κατάλληλο αρχείο Digenis. Δεδομένου ότι αυτό δε θα είναι συχνό, ο απαιτούμενος χρόνος για τις αντίστοιχες μαζικές (batch) επεξεργασίες/διαδικασίες δε θα μειώνει συνολικά την απόδοση της μεθόδου.

3.2.3. Αξιολόγηση της Μεθόδου

3.2.3.1. Θεωρητική Αξιολόγηση

Η θεωρητική αξιολόγηση της μεθόδου μας εστιάζει σε δύο αντιπροσωπευτικές περιπτώσεις: τη μέση(mean) και τη χειρότερη(worst) περίπτωση (για λόγους απλότητας θα θεωρήσουμε αρχικά τις περιπτώσεις όπου υπάρχει μόνο ένα αρχείο Digenis).

⇒ Ως μέση περίπτωση θεωρούμε εκείνη όπου το lev1 είναι το μεσαίο επίπεδο του δέντρου και η ομοιότητα εντοπίζεται στο μέσον της υπόλοιπης αναζήτησης. Πιο συγκεκριμένα, έστω ότι έχουμε N αντικείμενα τα οποία έχουν ομαδοποιηθεί σε M ($M < N$) οικογένειες. Αυτές οι M οικογένειες έχουν τοποθετηθεί σε ένα δέντρο Digenis του οποίου το πλήθος επιπέδων εξαρτάται από το συγκεκριμένο ch του δέντρου. Εάν ο αριθμός των επιπέδων των δέντρων είναι έστω tl, το μεσαίο επίπεδο είναι $\lceil tl/2 \rceil$, οπότε οι υπολογισμένες αποστάσεις για το μεσαίο επίπεδο είναι:

$$DistCalc = ch^{\lceil tl/2 \rceil}$$

Ο υπόλοιπος έλεγχος θα χρειαστεί το πολύ $\lceil (tl/2)/2 \rceil * ch$ υπολογισμούς αφού σε κάθε ένα από τα υπόλοιπα $\lceil (tl/2)/2 \rceil$ επίπεδα, πρέπει να υλοποιηθούν ch υπολογισμοί. Συνεπώς, ο συνολικός αριθμός υπολογισμών είναι:

$$TotalDistCalc = ch^{\lceil tl/2 \rceil} + \lceil tl/2 \rceil * ch$$

Παράδειγμα: Εάν έχουμε $N=16000$ το οποίο έχει οδηγήσει σε $M=3800$ οικογένειες, με $ch=5$, τότε ο αριθμός επιπέδων του δέντρου θα είναι 5 και ο μέσος αριθμός υπολογισμών θα είναι ίσος με $5^3 + \lceil (5/2) \rceil * 5 = 125 + 2*5 = 130$, αριθμός που αντιστοιχεί σε $0,0081\%$ του N .

⇒ Η χειρότερη περίπτωση είναι εκείνη κατά την οποία θα απαιτηθεί ο μεγαλύτερος αριθμός υπολογισμών. Αυτό θα συμβεί όταν το αντικείμενο του ερωτήματος (q) τυχαίνει να είναι όμοιο με κάποιο από τα φύλλα του δέντρου. Τότε το $lev1$ θα είναι ίσο με $tl-1$, οπότε θα εκτελεστούν ch^{tl-1} υπολογισμοί για αυτό το επίπεδο και ch επιπλέον υπολογισμοί για το επόμενο και τελευταίο, έχοντας:

$$TotalDistCalc = ch^{tl-1} + ch.$$

Παράδειγμα: Για τα δεδομένα του προηγούμενου παραδείγματος, έχουμε $TotalDistCalc = 5^4 + 5 = 625 + 5 = 630$, που αντιστοιχεί σε $0,04\%$ του N .

Γενικά, στην περίπτωση ενός πλήρους δέντρου και θεωρώντας ως f το μέσο αριθμό αντικειμένων ανά οικογένεια, ισχύει το παρακάτω για το ποσοστό P των εκτελούμενων υπολογισμών:

$$P = (\text{χειρότερη περίπτωση υπολογισμών}) / N =$$

$$(ch^{\lceil tl-1 \rceil} + ch) / (f * (\text{αριθμό οικογενειών})) = (ch^{\lceil tl-1 \rceil} + ch) / (f * (ch^{tl+1} - 1)) \cong ch^{tl-1} / (f * ch^{tl+1}) = 1 / (f * ch^2)$$

παρέχοντας ένα γενικό τύπο μέτρησης/εκτίμησης για τη χειρότερη περίπτωση του ποσοστού υπολογισμών. Για παράδειγμα, για $f=ch=2$, η οποία είναι μία ασυνήθιστα ακραία χειρότερη περίπτωση, το αποτέλεσμα είναι $12,5\%$, ενώ για μεγαλύτερα (και πιο ‘φυσιολογικά’) f και ch , τα ποσοστά μειώνονται εντυπωσιακά (π.χ. για $f=4$ και $ch=5$, το P είναι 1%)

Προτού συγκρίνουμε την απόδοση της μεθόδου μας με τις αποδόσεις άλλων αντίστοιχων μεθόδων, πρέπει να λάβουμε υπ’όψιν το εξής: στον $TotalDistCalc$ πρέπει να προστεθεί ο χρόνος καθορισμού των συγκεκριμένων κάθε φορά k γειτόνων, όπως το απαιτεί το αρχικό ερώτημα. Όπως αναφέρθηκε και προηγούμενα, αυτός ο καθορισμός είναι αρκετά εύκολη διαδικασία, συνεπώς ο αντίστοιχος χρόνος απόκρισης θα είναι αρκετά χαμηλός. Συνεπώς, τα ποσοστά που εκτιμήθηκαν παραπάνω (π.χ. $0,04\%$ και $0,0081\%$) θα είναι στην πραγματικότητα - για τη συνολική αναζήτηση ομοιοτήτων χρησιμοποιώντας το αρχείο Digenis - λίγο μεγαλύτερα (δηλ. σχεδόν $0,04\%$ και $0,0081\%$).

Σύμφωνα με την [A8], τα VA-Files είναι πολύ καλύτερα σε απόδοση σε σχέση με τις μεθόδους κατανομής δεδομένων και χώρου για πολυδιάστατους χώρους, επιτυγχάνοντας ένα κόστος I/O χαμηλότερο πάντα από 20% . Στην πράξη αυτό το ποσοστό είναι συνήθως γύρω στο 1% , επιβεβαιώνοντας την ανωτερότητα της μεθόδου. Συγκρίνοντας αυτή την απόδοση με τη χειρότερη περίπτωση του παραδείγματός μας (σχεδόν $0,04\%$), μπορούμε να έχουμε μία πρώτη ιδέα σχετικά με τις βελτιώσεις που εισάγει το αρχείο Digenis.

Στις περιπτώσεις τώρα όπου χρησιμοποιούνται περισσότερα από ένα αρχεία Digenis, τότε i) για τα περισσότερα από αυτά, το $lev1$ θα είναι μεγαλύτερο από τον αριθμό των αντίστοιχων επιπέδων, οδηγώντας σε απ’ευθείας αποκλεισμούς και ii) ένας μικρός μόνο αριθμός αρχείων θα ελεγχθεί άδικα γιατί παρόλο που δεν ‘κρύβουν’ τη θέση του q , το $lev1$ τους είναι μικρότερο του αριθμού των επιπέδων τους. Το τελευταίο συμβαίνει όταν αυτά τα δέντρα κείνται ‘κοντά’ στο δέντρο όπου βρίσκεται η πραγματική θέση του q . Σ’ αυτές τις περιπτώσεις θα τρέξει μία από τις χειρότερες εκδοχές (worst case) της διαδικασίας, η οποία θα έχει ως αποτέλεσμα μία επιβάρυνση στην εκτιμώμενη απόδοση, η οποία θα είναι ένα μικρό πολλαπλάσιο της απόδοσης χειρότερης περίπτωσης. Επιπρόσθετα, εάν υπάρχει ένα Αρχείο Απομακρυσμένων και το q βρει τελικά τη θέση του εκεί, ο συνολικός χρόνος

αναζήτησης θα είναι μόνο ο χρόνος σειριακής αναζήτησης σ' αυτό, αφού δε θα ακολουθήσει κανένας έλεγχος στα αρχεία Digenis. Όπως προαναφέρθηκε, αυτό το αρχείο θα είναι γενικά μικρό, άρα και ο χρόνος για την αναζήτηση θα είναι ιδιαίτερα μικρός, οδηγώντας σε ένα συνολικό χρόνο αναζήτησης πολύ χαμηλό, πολύ μικρότερο από το χρόνο των TotalDistCalc που εκτιμήθηκε παραπάνω.

3.2.3.2. Πειραματική Αξιολόγηση

Εκτός από τη θεωρητική αξιολόγηση της πρότασης Digenis, πραγματοποιήθηκε και πειραματική αξιολόγηση της μεθόδου, βασισμένη σε συνθετικά ομοιόμορφα κατανεμημένα δεδομένα. Χρησιμοποιήθηκαν πέντε σύνολα δεδομένων, των 500 ($N=500$) αντικειμένων το καθένα, με πλήθος διαστάσεων 5, 8, 10, 14 και 28 αντίστοιχα (Σχήμα 3.6.) Τα πειράματα πραγματοποιήθηκαν σε προσωπικό υπολογιστή 1.1 GHz CPU και 512 RAM. Οι προς σύγκριση δομές και αλγόριθμοι ήταν η σειριακή, το R*-tree, το VA-File και το αρχείο DIGENIS. Αξιολογώντας τα αποτελέσματα των πειραμάτων είδαμε – όπως ήταν αναμενόμενο – ότι η απλή σειριακή μέθοδος χρειάζεται ένα τυχαίο αριθμό προσπελάσεων (μικρότερο από N) σε κάθε περίπτωση, ενώ καθώς αυξάνεται το πλήθος των διαστάσεων η R*-tree αναζήτηση εκφυλίζεται σε σειριακή αναζήτηση ανάμεσα στα φύλλα του δέντρου[A8]. Από την άλλη μεριά, η απόδοση του VA-File βελτιώνεται όσο αυξάνεται το πλήθος των διαστάσεων των εμπλεκόμενων αντικειμένων [A8], ενώ το αρχείο Digenis αποδίδει καλύτερα από όλες τις μεθόδους για όλες τις διαστάσεις. Πιο συγκεκριμένα, το αρχείο Digenis έχει σχεδόν την ίδια απόδοση με το R*-tree για μικρό πλήθος διαστάσεων (Πίνακας 3.2.) αλλά το ξεπερνά για διαστάσεις περισσότερες των 10, αφού είναι μία βελτιωμένη εκδοχή αρχείου τύπου VA-File. Από την άλλη πλευρά, το αρχείο Digenis αποδίδει όμοια με το VA-File ή καλύτερα (για μικρό πλήθος διαστάσεων, δεδομένου ότι βασίζεται σε οργάνωση δέντρου ομοιοτήτων (δέντρο Digenis) (Πίνακας 3.2.)).

Κάτι άλλο για το οποίο ελέγξαμε τα δεδομένα μας μέσω των πειραμάτων, ήταν το κατά πόσο και πώς η ομοιογενής ή όχι κατανομή των δεδομένων μπορεί να επηρεάσει την απόδοση της μεθόδου. Για αυτόν το σκοπό χρησιμοποιήσαμε δύο άλλα μη ομοιογενώς κατανεμημένα σύνολα δεδομένων, με $N=500$ και διαστάσεις 14 και 28 αντίστοιχα. Τα συμπεράσματα ήταν:

- Οι αποδόσεις ήταν ελαφρώς χειρότερες από εκείνες που αφορούσαν τα αντίστοιχα ομοιογενή δεδομένα. Ο λόγος για αυτή τη μικρή - και όχι μεγάλη - όπως μπορεί να περίμενε κανείς - επιβάρυνση είναι το ότι η ανομοιογένεια στην κατανομή των αντικειμένων οδηγεί στη δημιουργία πολλαπλών δέντρων Digenis γεγονός που επιβαρύνει χρονικά εκείνη τη φάση – τη φάση κατασκευής των δέντρων – και όχι τη φάση που εξετάζουμε.
- Μελετώντας τα δέντρα Digenis που δημιουργήθηκαν για αυτά τα δεδομένα, παρατηρήθηκε πως υπήρχαν πολλοί κενοί κόμβοι, κατάσταση που τροφοδότησε την έρευνά μας σχετικά με το πότε αξίζει τον κόπο να 'κόβονται' τα αντίστοιχα υπο-δέντρα, για τη δημιουργία νέων μικρότερων. Τα πειράματα έδειξαν πως - όπως αναφέρθηκε και προηγούμενα στην ενότητα 3.2.1.1.) – όταν οι κενοί κόμβοι ξεπερνούν το 50% των κόμβων του υπο-δέντρου και τα επίπεδά του είναι περισσότερα από το 30% των επιπέδων του αρχικού δέντρου, αξίζει να αποκοπεί το υπο-δέντρο, να αποτελέσει νέο αναμορφωμένο δέντρο και να αναδιοργανωθεί το αρχικό.

Σχήμα 3.6. Αξιολόγηση σύμφωνα με το πλήθος των διαστάσεων

	Σειριακή Αναζήτηση		R*-tree		VA-File	
	LD	HD	LD	HD	LD	HD
DIGENIS-File	↑	↑	–	↑	↑	↑
Αιτίες	1*		2*	3*	4*	5*

LD: Μικρό Πλήθος Διαστάσεων (Low Dimensionality),
HD: Μεγάλο Πλήθος Διαστάσεων (High Dimensionality),

Τα βέλη (και το μέγεθός τους) αντιπροσωπεύουν την υπεροχή του αρχείου Digenis, σε σχέση με την αντίστοιχη μέθοδο της στήλης του πίνακα.
Η παύλα (-) σημαίνει 'καμία αξιοσημείωτη διαφορά στην απόδοση'.

* Αιτίες

1*. Η Σειριακή Αναζήτηση έχει τυχαία απόδοση (μέση απόδοση= $N/2$).

2*. Το αρχείο DIGENIS βασίζεται σε δέντρο κατανομής δεδομένων γι' αυτό έχει παρόμοια απόδοση με το R*-tree για μικρό πλήθος διαστάσεων.

3*. Το αρχείο DIGENIS είναι μία βελτιωμένη εκδοχή αρχείου τύπου VA-File, η οποία αποδίδει πολύ καλύτερα από τα δέντρα για μεγάλο πλήθος διαστάσεων.

4*. Το αρχείο DIGENIS-File, βασίζεται σε δέντρο κατανομής δεδομένων το οποίο είναι πάντα πολύ καλύτερο από το VA-File για μικρό πλήθος διαστάσεων.

5*. Το αρχείο DIGENIS είναι μία βελτιωμένη εκδοχή αρχείου τύπου VA-File, οπότε αποδίδει λίγο καλύτερα από αυτό για μεγάλο πλήθος διαστάσεων, λόγω της βελτίωσης που εισάγει.

Πίνακας 3.2. Σύγκριση απόδοσης του Αρχείου Digenis με άλλες μεθόδους

3.3. Ανακεφαλαίωση – Ανοικτά θέματα

Όπως έχει αποδειχτεί [A8], τα γνωστά δέντρα κατανομής δεδομένων και χώρου είναι πολύ αποδοτικά για δεδομένα μικρού πλήθους διαστάσεων, ενώ οδηγούν σε διαδικασίες αργότερες της σειριακής αναζήτησης όταν το πλήθος των διαστάσεων αυξάνεται αρκετά (π.χ. πάνω από 10). Από την άλλη μεριά τα VA-Files τα οποία είναι απλοί πίνακες γεωμετρικών προσεγγίσεων των αντικειμένων, έχουν την ακριβώς αντίθετη απόδοση, αφού αποδίδουν πολύ καλύτερα όσο αυξάνεται το πλήθος των διαστάσεων.

Σ' αυτό το κεφάλαιο παρουσιάστηκε η υβριδική πρόταση Digenis, η οποία αξιοποιεί τις δυνατότητες των δέντρων ομοιοτήτων και των VA-Files[A8], προκειμένου να εξασφαλίσει μία ισορροπία στην απόδοση της αναζήτησης ομοιοτήτων ανάμεσα σε αντικείμενα, η οποία θα είναι σχετικά ανεξάρτητη από το πλήθος των διαστάσεων των αντικειμένων, παρέχοντας ευελιξία και γενικότητα. Η μέθοδος Digenis περιλαμβάνει:

- *Εννοιολογικά:* Τα δέντρα Digenis για εύκολη και ευέλικτη αναπαράσταση των υπαρχόντων ομοιοτήτων ανάμεσα στα αντικείμενα. Βασικό χαρακτηριστικό αυτών των δέντρων είναι πως η ομοιότητα των αντικειμένων ορίζεται σε δυο επίπεδα. Το πρώτο αφορά τα πολύ όμοια αντικείμενα τα οποία ομαδοποιούνται σε *οικογένειες* οι οποίες αποτελούν τους κόμβους του δέντρου και το δεύτερο τα απλώς όμοια αντικείμενα τα οποία οργανώνονται στο δέντρο με βάση τις σχέσεις γονέα-παιδιού.
- *Φυσικά:* Τα αρχεία Digenis τα οποία φιλοξενούν τα δεδομένα με τη σειρά και τη λογική που προκύπτει από τα αντίστοιχα δέντρα. Καταχρηστικά μπορούμε να πούμε πως τα αρχεία Digenis είναι ένα είδος VA-File, το οποίο επιτυγχάνει ακόμα μεγαλύτερη συμπίεση, προσέγγιση και ομαδοποίηση δεδομένων από το VA-File, μέσω των:
  - ✓ Οικογενειών αντικειμένων
  - ✓ Των δέντρων Digenis τα οποία προ-διατάσσουν τα αντικείμενα με βάση τις ομοιοτητές τους.

Με βάση αυτές τις δομές και με τη βοήθεια μιας διαδικασίας ασφαλούς αποκλεισμού μεγάλου μέρους των αντικειμένων από την αναζήτηση, γίνεται δυνατός ο εντοπισμός των όμοιων αντικειμένων σε χρόνους καλύτερους από τους αντίστοιχους των δέντρων ομοιοτήτων και των VA-Files, ανεξάρτητα από το πλήθος των διαστάσεων των εμπλεκόμενων αντικειμένων. Το τελευταίο χαρακτηριστικό οδηγεί σε ομογενοποίηση των χώρων λίγων διαστάσεων με κείνους των πολλών διαστάσεων, οι οποίοι αντιμετωπίζονται στην ίδια βάση.

Κλείνοντας, υπογραμμίζουμε πως η πρόταση Digenis θέτει νέα σχετικά ανοικτά προβλήματα: Υπάρχει κάποια τιμή του $ch(\text{αριθμού παιδιών})$ η οποία μπορεί αποδεδειγμένα να θεωρηθεί η πιο κατάλληλη για ένα δέντρο Digenis; Υπάρχουν κανόνες που καθορίζουν την αποτελεσματικότητα αυτής της τιμής; Και προχωρώντας ένα βήμα πιο πέρα, είναι πιθανό (και πώς) να αποφευχθεί η ύπαρξη ενός συγκεκριμένου αριθμού παιδιών (ch) για κάθε κόμβο κάθε δέντρου Digenis, χωρίς σημαντική απώλεια σε απόδοση;

3.4. Δημοσιεύσεις του Κεφαλαίου

♦ ΣΥΝΕΛΠΙΟ

- [Σ15] Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, “Digenis: An Hybridic Proposal for (Dis)Similarity Search”, IKE 2009 (ranking:0,80).

3.5. Αναφορές του Κεφαλαίου

- [A1] Chen, Q., Chen, L., Lian, X., Liu, Y., Xu, Yu.: Indexable PLA for Efficient Similarity Search. In Very Large Data Bases (VLDB) 2007.
- [A2] Ciaccia, P., Patella P., Zezula P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In Very Large Data Bases (VLDB) 1997.
- [A3] Hellerstein, J.M., Jeffrey F. Naughton, J.F., Pfeffer, A.: Generalized Search Trees for Database Systems. In Very Large Data Bases (VLDB) 1995.
- [A4] Kibriya, A., Frank, Eibe.: An Empirical Comparison of Exact Nearest Neighbour Algorithms. In PKDD 2007.
- [A5] Nievergelt, J., Hinterberger, H., Servik, K.: The grid file: An adaptable symmetric multikey file structure. ACM Transactions on Database Systems, 1984.
- [A6] Sakurai Y., Yoshikawa M., Uemura S., Kojima H.: The A-tree: An Index Structure for High-Dimensional Spaces Using Relative Approximation. In Very Large Data Bases (VLDB) 2000.
- [A7] Hinneburg, A., Aggarwal, C., Keim, D.: What is the nearest neighbor in high dimensional spaces? In VLDB 2000
- [A8] Weber, R., Schek, H.J, Blott, S.: A Quantitive Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In Very Large Data Bases (VLDB) 1998.
- [A9] Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Searching. In SIGMOD 1984.
- [A10] Seeger, B., Kriegel, H.P.: The Buddy Tree: An Efficient and Robust Access Method for Spatial Data Base Systems. In Very Large Data Bases (VLDB) 1990.
- [A11] Kamel, I., Faloutsos, C.: Hilbert R-tree: An Improved R-tree Using Fractals. In Very Large Data Bases (VLDB) 1994.
- [A12] Ciaccia, P., Patella, M.: The M2-tree: Processing Complex Multi-Feature Queries with Just One Index. In 1st DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland, 2000.
- [A13] Arge, L., Mark de Berg, Haverkort, H.J., Yi K.: The Priority R-Tree: A Practically Efficient and Worst-Case Optimal R-Tree. In SIGMOD 2004.
- [A14] Augsten, N., Bohlen, M., Gamper, J.: An Incrementally Maintainable Index for Approximate Lookups in Hierarchical Data. In Very Large Data Bases (VLDB) 2006.
- [A15] Chen, Q., Chen, L., Lian, X., Liu, Y., Xu, Yu.: Indexable PLA for Efficient Similarity Search. In Very Large Data Bases (VLDB) 2007.
- [A16] Wong, R.C.-W., Tao, Y., Fu, A.W.-C., Xiao, X.: On Efficient Spatial Matching. In Very Large Data Bases (VLDB) 2007.
- [A17] White, D.A. Jain, R.: Similarity indexing with the SS-tree. In 12th International Conference on Data Engineering (ICDE) 1996.
- [A18] Goh, K.-S., Li, B., Chang, E.: DynDex: A Dynamic and Non-metric Space Indexer. In Multimedia 2002.

- [A19] Qian, G., Zhu, Q., Xue, Q., Pramanik, S.: The ND-Tree: A Dynamic Indexing Technique for Multidimensional Non-ordered Discrete Data Spaces. In Very Large Data Bases (VLDB) 2003.
- [A20] Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A., Theodoridis, Y.: R-Trees: Theory and Applications, Springer 2005. ISBN 1-85233-977-2.
- [A21] Berchtold, S., Keim, D., Kriegel, H.-P.: The X-tree: An Index Structure for High-Dimensional Data. In Very Large Data Bases (VLDB) 1996.
- [A22] Katayama, N., Satoh, S.: The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. In SIGMOD 1997.
- [A23] Lin, K.-L., Jagadish, H., Faloutsos, C.: The TV-tree: An index structure for high-dimensional data. In VLDB Journal, 3(4): 517-549, Oct.1994.
- [A24] Robinson, J.: The k-d-b-tree: A search structure for large multidimensional dynamic indexes. In SIGMOD 1981.
- [A25] Finkel, R., Bentley, J.: Quad-trees: A data structure for retrieval on composite keys. In ACTA Informatica 4(1):1-9, 1974.

3.7. Παράρτημα 3^ο Κεφαλαίου – VA-Files

Το διανυσματικό προσεγγιστικό αρχείο (Vector Approximation File – VA- File) χωρίζει το χώρο των δεδομένων σε 2^b ορθογώνια κελιά όπου το b αντιστοιχεί σε έναν αριθμό ψηφίων (bits) προκαθορισμένο από τον χρήστη (κάποιος αριθμός ψηφίων ανά διάσταση). Αντί όμως να οργανωθούν αυτά τα κελιά ιεραρχικά όπως στα grid-files ή στα R-trees, το VA-File αντιστοιχίζει μία μοναδική ακολουθία ψηφίων (bit-string) μήκους b σε κάθε κελί και προσεγγίζει τα δεδομένα σημεία που βρίσκονται σ' αυτό το κελί, με αυτή την ακολουθία ψηφίων. Όπως είναι φανερό, το VA-File δεν είναι τίποτε άλλο από έναν απλό πίνακα αυτών των συμπαγών γεωμετρικών προσεγγίσεων. Τα ερωτήματα του κοντινότερου γείτονα (nearest neighbor, NN) υλοποιούνται με έλεγχο ολόκληρου του προσεγγιστικού αρχείου και μέσω του αποκλεισμού της μεγάλης πλειοψηφίας των διανυσμάτων από την αναζήτηση (filtering step), με βάση αυτές τις προσεγγίσεις.

Από πρακτικής πλευράς, τα VA-Files και η λειτουργία τους μπορούν να γίνουν καλύτερα αντιληπτά με τη βοήθεια του παρακάτω σχήματος:

Σχήμα 3.7. : Κατασκευή του VA-File

Κάθε σημείο-αντικείμενο προσεγγίζεται από μία ακολουθία ψηφίων, που υποδεικνύει το κελί στο οποίο βρίσκεται, μέσα στο πλέγμα. Έτσι κατά την αναζήτηση ελέγχεται πρώτα ο πίνακας των προσεγγίσεων, από όπου γίνεται ένα πρώτο μεγάλο φιλτράρισμα. Όταν εντοπιστεί το κελί στο οποίο ανήκει το σημείο-αντικείμενο, τότε πια ελέγχεται ειδικότερα το κελί και τα γειτονικά του, ανάλογα με τις απαιτήσεις ανεύρεσης ομοιοτήτων που έχει θέσει το εκάστοτε ερώτημα.

Ανάλογα με την ακρίβεια των αντικειμένων και τον αριθμό των ψηφίων που έχει επιλεγεί για την εκάστοτε προσέγγιση, το προσεγγιστικό αρχείο είναι 4 με 8 φορές μικρότερο από το αρχείο των δεδομένων και συνεπώς η αντίστοιχη επιβάρυνση χώρου (0,125-0,25 του αρχικού) είναι πολύ μικρή. Από την άλλη πλευρά, το γεγονός ότι το προσεγγιστικό αρχείο είναι κατά μεγάλο ποσοστό μικρότερο από το αρχικό αρχείο δεδομένων, συμβάλλει καθοριστικά στην αύξηση της απόδοσης της μεθόδου με την αύξηση του πλήθους των διαστάσεων των δεδομένων (αφού έτσι μεγαλώνει το αρχικό αρχείο και συνεπώς και η διαφορά μεταξύ των αρχείων).

Λίγη παρατήρηση και πολλή ανάλυση οδηγούν σε λάθη.
Πολλή παρατήρηση και λίγη ανάλυση οδηγεί στην αλήθεια.

Alexis Carrel, 1873-1944, Γάλλος ερευνητής ιατρικής

Κεφάλαιο 4

Μέθοδος Αναζήτησης Ομοιοτήτων Ανάμεσα σε Χαρακτηριστικά Αντικειμένων (Δέντρο Chiron)

[Σ17]

Περιεχόμενα Κεφαλαίου

- 4.1. Εισαγωγή – Πλαίσιο Αναφοράς – Ορισμός του Προβλήματος
- 4.2. Τυποποίηση του Προβλήματος
  - 4.2.1. Ορισμοί
- 4.3. Δέντρο Chiron
  - 4.3.1. Θεώρημα Chiron
  - 4.3.2. Χρήση του Δέντρου Chiron
    - 4.3.2.1. Εντοπισμός Θέσης (Positioning)
    - 4.3.2.2. Εντοπισμός των Όμοια Εξελισσόμενων Χαρακτηριστικών (Ομοιότητα Chiron)
- 4.4. Πειραματική Μελέτη - Αξιολόγηση
- 4.5. Ανοικτά Θέματα – Μελλοντικές Κατευθύνσεις
- 4.6. Ανακεφαλαίωση – Συμπεράσματα
- 4.7. Δημοσιεύσεις του Κεφαλαίου
- 4.8. Αναφορές του Κεφαλαίου

Σύμβολο	Έννοια - Σημασία
d	Dimensionality – Αριθμός διαστάσεων
At	Absolute tolerance – απόλυτη ανοχή (όριο διαφορών)
Rt	Relevant tolerance – σχετική ανοχή (όριο διαφορών)
Tolerance	Ανοχή (απόλυτη ή σχετική)
Iid	Instance identifier – Ταυτότητα εμφάνισης
CO _I	Course Object I – I-στο Αντικείμενο Πορείας Χαρακτηριστικού
cd	Course Difference – Διαφορά ψηφίων ανάμεσα σε δύο αντικείμενα πορείας (αριθμός διαφορετικών ψηφίων)
κωδ _i	Κωδικός χαρακτηριστικού με αντικείμενο πορείας το αντικείμενο της ομάδας του συγκεκριμένου κόμβου
cdr	Διαφορά πορείας ανάμεσα σε ένα κόμβο (αντικείμενο) και τη ρίζα ενός δένδρου Chiron
cd _p	Διαφορά πορείας ανάμεσα σε μία ομάδα(κόμβο) αντικειμένων πορείας και στον κόμβο(ομάδα) γονέα του
Gi	Group I – Η i-στη ομάδα αντικειμένων πορείας
t	Ανοχή Δένδρου Chiron (επιτρεπτή διαφορά ομάδων (κόμβων) γονέα-παιδιού)
MaxDif	$cd_a + cd_b$ – Μέγιστη Δυνατή Διαφορά δύο αντικειμένων πορείας
MinDif	$cd_a - cd_b$ – Ελάχιστη Δυνατή Διαφορά δύο αντικειμένων πορείας
ro1	Αντικείμενο ερωτήματος (query object)
ommin	Ομάδα αντικειμένων πορείας με τη μικρότερη διαφορά από μία άλλη δοθείσα ομάδα
qt	query tolerance – ανοχή που θέτει ένα ερώτημα βάσης δεδομένων
qg	query group – ομάδα αντικειμένων πορείας στην οποία 'υπάγεται' το αντικείμενο του ερωτήματος (query object)

Πίνακας 4.1.: Πίνακας Συμβολισμών

Σύνοψη

Ένα από τα σύγχρονα πεδία μελέτης σχετικά με την αναζήτηση ομοιοτήτων ανάμεσα σε αντικείμενα είναι εκείνο που αφορά τα *χρονικά δεδομένα (time-streams)*. Ο σκοπός αυτού του είδους της έρευνας είναι ο εντοπισμός των ομοιοτήτων στην εξέλιξη δεδομένων αντικειμένων, στην πορεία του χρόνου. Στο κεφάλαιο αυτό στρέφουμε την έρευνά μας ένα βήμα πιο πέρα από αυτού του είδους την αναζήτηση, στην *ανίχνευση ομοιοτήτων της πορείας όχι αντικειμένων πια αλλά των χαρακτηριστικών τους (ομοιότητα Chiron)*. Κάτι τέτοιο μπορεί να δώσει ιδιαίτερα χρήσιμα συμπεράσματα σε πολλά είδη εφαρμογών (π.χ. σε ιατρικές εφαρμογές), όπου η κοινή ή παρόμοια πορεία χαρακτηριστικών (properties) κάποιων αντικειμένων 'προδίδει' χρήσιμη γνώση σχετικά με αλληλεπιδράσεις, τις εξαρτήσεις και τις 'συγγένειες' ανάμεσα στα χαρακτηριστικά. Από την άλλη μεριά, η διαχείριση αυτού του τόσο χρήσιμου είδους αναζήτησης συνιστά ένα πρόβλημα ευκολότερο από εκείνο της κλασσικής αναζήτησης ομοιοτήτων, αφού τελικά δεν ενδιαφερόμαστε για τις πραγματικές τιμές αλλά μόνο για τις μεταβολές τους. *Με την επιλογή και τον ορισμό απλών συμβάσεων για την αναπαράσταση αυτών των μεταβολών καθώς και με την εισαγωγή ενός σχετικά απλού αλλά μελετημένου δένδρου ομοιοτήτων (Chiron tree) για τη διαχείριση αυτών των μεταβολών, εισάγουμε μία πρώτη προσέγγιση μιας συνολικής πρότασης για την αναζήτηση ομοιοτήτων Chiron.*

Λέξεις Κλειδιά: *Εξόρυξη Πολυδιάστατων και Χρονικών Δεδομένων, Αντικείμενο Πορείας Χαρακτηριστικού (Chiron Object), Σχεδόν Ίδια Αντικείμενα, Ομοιότητα Chiron, Δένδρο Chiron.*

4.1. Εισαγωγή – Πλαίσιο Αναφοράς – Ορισμός του Προβλήματος

Γενικά, διακρίνονται δύο βασικές κατηγορίες αναζήτησης ομοιοτήτων (similarity search): η στατική και η δυναμική. Στα πλαίσια της πρώτης επιχειρείται η εύρεση ομοιοτήτων σε ένα σύνολο αντικειμένων (πολυδιάστατων και μη) [A1, A5, A6, A7] ενώ η δεύτερη ασχολείται με το πόσο όμοια αλλάζουν συγκεκριμένα αντικείμενα στην πορεία του χρόνου [A3, A4, A21, A22, A23, A24, A25]. Αν και όπως είναι φανερό κάθε κατηγορία έχει τη σπουδαιότητά της ανάλογα με την εφαρμογή, η δυναμική αναζήτηση ομοιοτήτων αποτελεί περισσότερο ελκυστικό ερευνητικό προορισμό. Αυτό συμβαίνει και για ποσοτικούς αλλά και για ποιοτικούς λόγους. Πιο συγκεκριμένα:

- ⇒ *Ποσοτικά*, γιατί είναι γεγονός πως σχεδόν όλες πια οι εφαρμογές (ιατρικές, εμπορικές/οικονομικές, επιστημονικές κλπ) χειρίζονται δυναμικά μεταβαλλόμενα στο χρόνο δεδομένα, για τα οποία διατηρείται και αξιοποιείται κατάλληλα κάθε φορά, όλη η πορεία των τιμών τους. Αν σε αυτή την πραγματικότητα προσθέσει κανείς το ότι η δυναμική αναζήτηση ομοιοτήτων είναι πολύ λιγότερο διερευνημένη από τη στατική, το αποτέλεσμα που προκύπτει είναι μία τεράστια περιοχή έρευνας.
- ⇒ *Ποιοτικά*, γιατί ανιχνεύοντας τον τρόπο με τον οποίο μεταβάλλονται τα δεδομένα στην πορεία του χρόνου, είναι δυνατόν να εξαχθούν ιδιαίτερα χρήσιμα συμπεράσματα και προβλέψεις, τα οποία αποτελούν την αφετηρία για νέα ανακάλυψη γνώσης (*knowledge discovery*). Ένα αντιπροσωπευτικό παράδειγμα – από το χώρο της ιατρικής – είναι εκείνο της διερεύνησης ομοιοτήτων κατά την ανάπτυξη όγκων, προκειμένου να το εξαχθούν συμπεράσματα σχετικά με το εάν υπάρχει κάποια εξάρτηση ή συσχέτιση ανάμεσα σε συγκεκριμένους όγκους [A1].

Το τελευταίο παράδειγμα των όγκων ανήκει στο κλασικό γενικό είδος της δυναμικής αναζήτησης ομοιοτήτων ανάμεσα σε πολυδιάστατα (και μη) αντικείμενα. Δηλαδή, αυτό που ανιχνεύεται και μελετάται είναι η *‘πορεία’ και εξέλιξη των αντικειμένων στην πορεία του χρόνου και ο εντοπισμός ομοιοτήτων σ’αυτές*. Παρόλα αυτά, στην περίπτωση των πολυδιάστατων αντικειμένων, θα μπορούσε να εξαχθεί ιδιαίτερα χρήσιμη και αξιοποιήσιμη πληροφορία και από τη μελέτη των ομοιοτήτων στην *‘πορεία’ και εξέλιξη όχι μόνο των ίδιων των αντικειμένων αλλά και των χαρακτηριστικών/πεδίων τους (properties/attributes), στην πορεία του χρόνου (property course similarity search)*. Πιο συγκεκριμένα, παρακολουθώντας εάν δύο ή περισσότερα χαρακτηριστικά εξελίσσονται με τον ίδιο τρόπο στην πορεία του χρόνου – στα ίδια χρονικά διαστήματα – μπορεί κανείς να φτάσει σε πολύ σημαντικά συμπεράσματα σχετικά με το πώς το ένα χαρακτηριστικό επηρεάζει το άλλο. Σκοπός της μελέτης σε αυτό το κεφάλαιο είναι η αρχική πρόταση μιας μεθόδου (*Chiron*¹) για την ανίχνευση τέτοιων ομοιοτήτων. Αξίζει να σημειωθεί πως όπως φάνηκε και από τα παραπάνω, τεράστιος αριθμός εφαρμογών μπορεί να αξιοποιήσει τέτοιου είδους συμπεράσματα και αποτελέσματα – όπως φυσικά και η βασική μας αρχική εφαρμογή της χαρτογράφησης του ανθρώπινου εγκεφάλου, για την εξαγωγή συμπερασμάτων σχετικά με την πορεία εξετάσεων του εγκεφάλου, συσχέτισης συμπτωμάτων, σημείων και μεταβολών σε εργαστηριακά δεδομένα στην πορεία του χρόνου. Στην παρούσα μελέτη όμως θα χρησιμοποιηθεί ως πλαίσιο αναφοράς το προσιτό παράδειγμα της *εφαρμογής διαχείρισης ορμονικών εξετάσεων* για την ευκολότερη κατανόηση και παρακολούθηση της μελέτης.

Παρόλο που η προτεινόμενη μέθοδος *Chiron* αποτελεί μία δυναμική προοπτική στον τομέα της αναζήτησης ομοιοτήτων, τα απαιτούμενα εργαλεία δεν είναι σύνθετα ή δύσκολα στην υλοποίηση. Αυτό συμβαίνει γιατί αντιμετωπίζουμε το πρόβλημα της αναπαράστασης των μεταβολών των χαρακτηριστικών με αφαιρετικό τρόπο.

Η μέθοδος *Chiron* συνίσταται στην εισαγωγή νέων εννοιών και ενός νέου δένδρου. Πιο συγκεκριμένα, αρχικά διερευνώνται θεωρητικά θέματα ομοιοτήτων, τα οποία εστιάζονται σε ομοιότητα ανάμεσα στις πορείες (στο χρόνο) χαρακτηριστικών (properties) αντικειμένων. Οι ανάγκες που προκύπτουν οδηγούν στο δεύτερο βασικό άξονα της μελέτης, στον ορισμό μίας αποδοτικής μεθόδου για την υποστήριξή τους. Ειδικότερα, πολυδιάστατα αντικείμενα τα οποία αντιπροσωπεύουν την πορεία κάθε χαρακτηριστικού στην πορεία του χρόνου, οργανώνονται σε ένα δένδρο, το δένδρο *Chiron*. Αυτή η οργάνωση βασίζεται στις διαφορές ανάμεσα σ’αυτές τις πορείες, λαμβάνοντας υπ’όψιν ένα προκαθορισμένο άνω όριο ανοχής (*tolerance limit*). Η βασική ιδέα κατασκευής του δένδρου είναι ιδιαίτερα απλή και πρακτική, χωρίς αυτό να συνεπάγεται παράλληλη απώλεια σε επίπεδο ακρίβειας της αναπαριστώμενης πληροφορίας. Σε αυτά τα ιδιαίτερα χαρακτηριστικά του δένδρου οφείλεται και η δυνατότητα πολύ αποδοτικής αναζήτησης μέσα σ’ αυτό, πράγμα που αποτελεί ένα από τα βασικά ζητούμενα κατά τον ορισμό ενός δένδρου. Σε τελικό επίπεδο, χρησιμοποιώντας το δένδρο *Chiron* και τους αντίστοιχους προτεινόμενους αλγόριθμους, επιτυγχάνονται πολύ ικανοποιητικά αποτελέσματα αναζήτησης ομοιότητας ανάμεσα στις πορείες των χαρακτηριστικών των αντικειμένων. Επιπλέον το δένδρο *Chiron* δεν *‘πλήττεται’* από το μεγαλύτερο μειονέκτημα των δένδρων υποστήριξης πολυδιάστατων αντικειμένων, την *‘κατάρρα’* των πολλών διατάσεων (dimensional ‘curse’, βλ. προηγούμενο κεφάλαιο) [A26].

¹ Η όλη πρόταση ονομάστηκε *Chiron* (*Χείρων*) από το όνομα του γνωστού κενταύρου. Η διαφορά του μισού του σώματος από το άλλο μισό – υποδεικνύοντας γενικά την έννοια της διαφοράς – αποτελεί ένα συμβολισμό σε σχέση με την αναζήτηση της σοφίας. Δανειστήκαμε αυτό το όνομα, ακριβώς βασισμένοι σ’αυτό το συμβολισμό, για να υπογραμμίσουμε ότι εστιάζουμε στη διαφορά (στην πορεία του χρόνου) για την εξαγωγή νέας γνώσης συσχετίσεων και εξαρτήσεων δεδομένων (*data dependency knowledge*).

4.2. Τυποποίηση του Προβλήματος

Σ' αυτή την ενότητα θα οριστούν τα αντικείμενα πορείας χαρακτηριστικού (*property course objects*), οι ομάδες αντικειμένων πορείας (*course object groups*), η διαφορά πορείας (*course difference*) ανάμεσα σε χαρακτηριστικά αντικειμένων, η απόλυτη ανοχή (*absolute tolerance*) *at* και η σχετική ανοχή (*relative tolerance*) *rt*. Αυτοί οι ορισμοί έχουν νόημα και εφαρμογή σε βάσεις δεδομένων όπου ο χρόνος έχει διαιρεθεί σε συγκεκριμένα διαστήματα χρόνου (*time intervals*) τα οποία μπορεί να είναι σταθερά ή όχι, σύμφωνα με τις ανάγκες και τις αποφάσεις του υπεύθυνου κάθε εφαρμογής. Για κάθε αντικείμενο, λαμβάνεται μία εμφάνιση (*instance*) για κάθε διάστημα χρόνου. Με βάση αυτές τις εμφανίσεις, η ατομική πορεία κάθε χαρακτηριστικού (*property*) στο χρόνο, μπορεί να αναπαρασταθεί με ένα νέο απλό αντικείμενο, όπου αντανακλώνται όχι πια οι πραγματικές τιμές αλλά οι διακυμάνσεις τους (αυξομειώσεις). Πιο συγκεκριμένα, η τιμή του εκάστοτε χαρακτηριστικού σε κάθε διάσταση συγκρίνεται με την αντίστοιχη (χρονικά) επόμενη και εάν αυξάνεται, μειώνεται ή παραμένει σταθερή, τα ψηφία 2, 0 και 1 καταγράφονται αντίστοιχα στην ανάλογη θέση του αντικειμένου πορείας (*course object*), το οποίο αντικατοπτρίζει την πορεία του συγκεκριμένου χαρακτηριστικού. Συγκρίνοντας όλα τα αντικείμενα πορείας μπορούν να εξαχθούν ενδιαφέροντα συμπεράσματα σχετικά με το 'πώς ένα χαρακτηριστικό 'συγγενεύει' ή αλληλεπιδρά με το άλλο'. Με άλλα λόγια, απομονώνοντας και συγκρίνοντας τα αντικείμενα πορείας των χαρακτηριστικών εντοπίζονται εξαρτήσεις δεδομένων. Αυτός είναι και ο σκοπός αυτής της μελέτης: η ανακάλυψη 'όμοιων' χαρακτηριστικών, η οποία 'υποδοθεί' νέα έρευνα και αποτελέσματα σε πολλά επιστημονικά πεδία. Στο παράδειγμά μας οι περιοδικές ορμονικές εξετάσεις ενός συνόλου ατόμων συνιστούν τη βάση δεδομένων, από την οποία μπορεί κανείς να ανακαλύψει αλληλεπιδράσεις (εξαρτήσεις) ανάμεσα σε συγκεκριμένες ορμόνες, με βάση την πορεία των τιμών αυτών των ορμονών η οποία παρακολουθείται και ελέγχεται μέσω της σύγκρισης των αντικειμένων πορείας τους.

Στις ιδιαίτερες περιπτώσεις-εφαρμογές όπου έχει εκτιμηθεί ότι η αλληλεπίδραση των χαρακτηριστικών (π.χ. χημικές ουσίες) δεν είναι αναγκαστικά η ίδια (π.χ. εξατομίκευση περιστατικών), τότε η παραπάνω διαδικασία γίνεται ξεχωριστά για κάθε αντικείμενο της σχετικής κλάσης και στη συνέχεια πραγματοποιείται μία 'ολοκλήρωση' των αποτελεσμάτων-συμπερασμάτων. Το τελικό αποτέλεσμα θα είναι ένα είδος στατιστικής επεξεργασίας του συνόλου των διαφορετικών συμπερασμάτων, κάτι που βρίσκεται έξω από τους σκοπούς της αρχικής παρουσίασης της πρότασης Chiron και αποτελεί θέμα μελλοντικής έρευνας.

4.2.1. Ορισμοί

Έστω ένα αντικείμενο d διαστάσεων με n εμφανίσεις, με συγκεκριμένο παράδειγμα μία βάση δεδομένων αντικειμένων 6 διαστάσεων, με 5 εμφανίσεις, όπως φαίνεται στο Σχήμα 4.1. Τα αντίστοιχα αντικείμενα πορείας (*course objects*) είναι τα τετραδιάστατα αντικείμενα (ή 4 σημείων-ψηφίων) του Σχήματος 4.2., τα οποία έχουν δημιουργηθεί από την σύγκριση διαδοχικών τιμών πεδίων, συμβολίζοντας την αύξησης τιμής με το ψηφίο 2, τη μείωση με το ψηφίο 0 και την σταθερότητα με το ψηφίο 1. Όπως θα φανεί καλύτερα και στις επόμενες ενότητες, αυτά τα νέα αντικείμενα είναι ιδιαίτερα απλά και εύκολα διαχειρίσιμα, γεγονός που αποτελεί και το βασικό μετασχηματιστικό προσόν της προτεινόμενης μεθόδου.

Id	d1	d2	d3	d4	d5	d6

1	5	3	9	10	7	5
2	6	4	9	3	3	6
3	9	12	9	20	9	9
4	12	13	12	15	4	15
5	21	8	16	16	4	14

Σχήμα 4.1.: Πέντε εμφανίσεις 6-διάστατης βάσης δεδομένων

CO ₁	2222
CO ₂	2220
CO ₃	1122
CO ₄	0202
CO ₅	0201
CO ₆	2221

Σχήμα 4.2.: Αντικείμενα Πορείας (Course Objects) για κάθε διάσταση των εμφανίσεων αντικειμένων του Σχήματος 4.1.

Προκειμένου να υπάρχει ένα σταθερό πλαίσιο στο οποίο θα βασιστεί η ανάλυση της μεθόδου, δίνεται παρακάτω ο τυπικός ορισμός αυτών των αντικειμένων:

Ορισμός 1. Αντικείμενο Πορείας Χαρακτηριστικού (Property Course Object) – Chiron Object:

Δοθέντων n εμφανίσεων ενός αντικειμένου, το αντικείμενο πορείας ενός χαρακτηριστικού του, είναι ένα νέο αντικείμενο $n-1$ ψηφίων τα οποία παίρνουν τιμές από το σύνολο $\{2, 1, 0\}$, με δεδομένο ότι το πρώτο από αυτά τα ψηφία χρησιμοποιείται για να συμβολίσει την αύξηση, το δεύτερο τη σταθερότητα και το τρίτο τη μείωση τιμής του χαρακτηριστικού στο εκάστοτε χρονικό διάστημα.

Χρησιμοποιώντας τα αντικείμενα πορείας τα οποία έχουν δημιουργηθεί όπως περιγράφηκε παραπάνω, είναι φανερό πως είναι σχετικά εύκολο να βρεθεί πού ακριβώς διαφέρει η εξέλιξη δύο ή περισσότερων χαρακτηριστικών. Εφόσον κωδικοποιήθηκε η πορεία των χαρακτηριστικών (properties) με τη χρήση απλών εγγραφών/πλειάδων (tuples) όπου οι πραγματικές τιμές έχουν αντικατασταθεί από μικρές τιμές, το αποτέλεσμα είναι να μετατραπεί το αρχικό σύνθετο πρόβλημα αναζήτησης ομοιοτήτων σε ένα πολύ απλούστερο. Στο επόμενο βήμα, πρέπει να βρεθεί εάν δύο αντικείμενα πορείας είναι ίδια. Λαμβάνοντας όμως υπ'όψιν ότι δεν έχει αξία να γνωρίζουμε εάν δύο τέτοια αντικείμενα 'μυσομοιάζουν' αλλά εάν είναι ακριβώς ή περίπου ίδια (όμοια), συμπεραίνουμε πως αυτό που πραγματικά έχει νόημα να ξέρουμε και να είμαστε σε θέση να εκτιμούμε, είναι πόσο κοντά στην ταύτιση (identification) είναι οι συγκεκριμένες κάθε φορά πορείες χαρακτηριστικών (μέσω των αντικειμένων πορείας). Αυτό επιτυγχάνεται με τον παρακάτω τρόπο: Τα αντικείμενα πορείας συγκρίνονται (π.χ. μέσω αριθμητικών αφαιρέσεων) και υπολογίζεται ο αριθμός των διαφορετικών μεταξύ τους ψηφίων. Αυτός ο αριθμός καλείται διαφορά πορείας (course

difference-cd) δύο χαρακτηριστικών ενώ παράλληλα ορίζεται μία απόλυτη (*absolute*) ή σχετική (*relevant*) ανοχή (όριο) προκειμένου να αποφασίζεται εάν μία διαφορά πορείας είναι αρκετά μικρή για να θεμελιωθεί ομοιότητα πορείας ή όχι. Πιο τυπικά:

Ορισμός 2. Διαφορά Πορείας (Chiron (Course) Difference): Δοθέντων δύο αντικείμενων πορείας α και β , ο αριθμός των ψηφίων (αντανακλώντας τον αριθμό εμφανίσεων) όπου τα α και β διαφέρουν, καλείται *cd* (Chiron (Course) Difference).

Ορισμός 3. At (Απόλυτη Ανοχή – Absolute tolerance): At είναι ο αριθμός που αντανακλά την απόλυτη ανοχή η οποία είναι αποδεκτή ως μέγιστο όριο ($cd \leq At$) διαφοράς πορείας για την διασφάλιση-εγγύηση ομοιότητας πορείας.

Παραδείγματος χάριν, στην περίπτωση του Σχήματος 4.1., η At μπορεί να οριστεί ίση με 2, σημαίνοντας ότι εάν δύο αντικείμενα πορείας διαφέρουν το πολύ σε δύο ψηφία τότε είναι αρκετά όμοια ώστε πιθανώς να επηρεάζουν το ένα το άλλο ή να επηρεάζονται από τους ίδιους παράγοντες.

Ορισμός 4. Rt (Σχετική Ανοχή - Relevant Tolerance): Εάν τα αντικείμενα πορείας έχουν d διαστάσεις, η Rt είναι μία ποσοστιαία τιμή η οποία αναπαριστά την ανοχή η οποία είναι αποδεκτή ως μέγιστο όριο ($((cd * 100) / d) \leq Rt$) διαφοράς πορείας για την διασφάλιση-εγγύηση ομοιότητας πορείας.

Στο ίδιο πάλι παράδειγμα, λαμβάνοντας υπ' όψιν ότι ο συνολικός αριθμός εμφανίσεων είναι μόνο 5, ίσως ένα At ίσο με 2 να είναι αρκετά μεγάλο (40%) για να εγγυάται ομοιότητα. Αυτό κάνει φανερό ότι σε πολλές περιπτώσεις μία αναλογική ανοχή θα ήταν περισσότερο αποδεκτή και αποδοτική. Πάντως σε κάθε περίπτωση, απλά παρέχουμε το εργαλείο και από εκεί και πέρα είναι ευθύνη του υπεύθυνου της εκάστοτε εφαρμογής να επιλέξει τη βέλτιστη λογική για την κάθε περίπτωση.

Μετά τους παραπάνω ορισμούς, μπορούν να γίνουν οι βασικοί ορισμοί:

Ορισμός 5. Σχεδόν Ίδια Αντικείμενα (Almost Identical Objects): Αντικείμενα πορείας χαρακτηριστικών των οποίων η διαφορά πορείας είναι μη μηδενική και μικρότερη ή ίση με την προκαθορισμένη ανοχή (At ή Rt) θεωρούνται σχεδόν ίδια, αφού εξελίσσονται με παρόμοιο τρόπο.

Ορισμός 6. Ομάδες Αντικειμένων Πορείας Χαρακτηριστικών (Course Object Groups – Chiron Groups):

Όλα τα ίδια (*identical*) αντικείμενα πορείας χαρακτηριστικών, συνιστούν μία ομάδα αντικειμένων πορείας (*group*).

Ορισμός 7. Ομοιότητα Chiron (Chiron Similarity): Δύο ή περισσότερα χαρακτηριστικά αντικειμένων των οποίων τα αντικείμενα πορείας είναι ίδια ή σχεδόν ίδια (*almost identical*), ονομάζονται όμοια εξελισσόμενα χαρακτηριστικά Chiron (*Chiron similar properties*) και η ομοιότητά τους καλείται ομοιότητα Chiron (*Chiron similarity*).

4.3. Δένδρο Chiron (Chiron tree)

Όπως έχει γίνει φανερό, ο βασικός σκοπός της μελέτης που περιγράφεται σ' αυτό το κεφάλαιο είναι η αποδοτική εύρεση των χαρακτηριστικών των αντικειμένων, τα οποία (χαρακτηριστικά) αλληλεπιδρούν ή έχουν κάποια κοινή βάση. Το βασικό 'κλειδί' για

να επιτευχθεί κάτι τέτοιο είναι η παρακολούθηση της παράλληλης εξέλιξής τους. Μετά από τους ορισμούς της προηγούμενης ενότητας - οι οποίοι εξασφαλίζουν έναν αποδοτικό μετασχηματισμό της πορείας των χαρακτηριστικών των αντικειμένων σε απλά, λίγων ψηφίων αντικείμενα - εκείνο που αποτελεί το κύριο απαραίτητο εργαλείο για τον εντοπισμό των ομοιοτήτων Chiron, είναι ένα ειδικό δένδρο ομοιοτήτων για την αποδοτική οργάνωση και διαχείριση των αντικειμένων πορείας χαρακτηριστικών (property course objects).

Ξεκινώντας, τον ορισμό ενός τέτοιου δένδρου (του δένδρου Chiron) θεωρούμε ότι σε κάθε περίπτωση υπάρχει ένα πολυδιάστατο (άρα και πολλών χαρακτηριστικών (properties)) αντικείμενο με έναν επαρκή αριθμό εμφανίσεων (instances). Αρχικά υπολογίζονται και σχηματίζονται οι απαραίτητες ομάδες αντικειμένων πορείας των επιλεγμένων χαρακτηριστικών (Chiron groups). Οι υπεύθυνοι της εκάστοτε εφαρμογής ορίζουν το είδος (At και Rt) και την ακριβή τιμή της αποδεκτής ανοχής και στη συνέχεια υπολογίζονται (με αφαιρέσεις) οι διαφορές πορείας ανάμεσα στις ομάδες αντικειμένων. Με βάση αυτές τις διαφορές, μόνο τα δεδομένα που σχετίζονται με διαφορές πορείας μικρότερες ή ίσες από την προκαθορισμένη ανοχή (δηλ. μόνο τα almost identical objects) χρησιμοποιούνται στο επόμενο βήμα. Από αυτά τα αντικείμενα πορείας, ως ρίζα του προτεινόμενου δένδρου επιλέγεται το αντικείμενο με τα περισσότερα σχεδόν ίδια αντικείμενα (almost identical objects) (η ομάδα 8 (G8-group 8) στο παράδειγμα των Σχημάτων 3 και 4, γιατί έχει τέσσερα σχεδόν ίδια αντικείμενα, ενώ όλες οι υπόλοιπες ομάδες έχουν το πολύ δύο).

Σχήμα 4.3.: Παράδειγμα αναπαράστασης γράφου διαφορών ανάμεσα σε ομάδες (groups).

Συνεχίζοντας την κατασκευή του δένδρου Chiron, προχωρούμε με τη δημιουργία του επόμενου επιπέδου το οποίο αποτελείται από τις ομάδες αντικειμένων πορείας που είναι σχεδόν ίδια με την ομάδα της ρίζας του δένδρου. Η ρίζα συνδέεται με αυτές τις ομάδες με ακμές οι οποίες περιλαμβάνουν την αντίστοιχη διαφορά πορείας (labels). Αυτές οι ομάδες ταξινομούνται ιεραρχικά μεταξύ τους, με βάση τη διαφορά τους (σε αύξουσα σειρά) με την ομάδα-γονέα. Για καθεμία από αυτές τις ομάδες, δημιουργείται το αντίστοιχο υπο-δένδρο με τον ίδιο τρόπο, για κάθε ομάδα σε κάθε υπο-δένδρο δημιουργείται το αντίστοιχο υπο-δένδρο, κοκ.

Σχήμα 4.4. : Ένα Δένδρο Chiron, βασισμένο στο Σχήμα 4.3.

Κατά τη διαδικασία της κατασκευής αυτού του δένδρου, όταν υπάρχει κάποια σύγκρουση – δηλ. δημιουργώντας ένα επίπεδο, πολλές ομάδες αντικειμένων πορείας μπορεί να είναι υποψήφιοι γονείς κάποιας άλλης ομάδας – η ελάχιστη διαφορά πορείας καθορίζει τη σύνδεση. Δηλαδή:

Κανόνας Κατασκευής του Δένδρου Chiron: Η ομάδα με την αμφισβητούμενη θέση σε ένα επίπεδο του δένδρου, θα συνδέεται με την υποψήφια προς σύνδεση ομάδα του ανώτερου της επιπέδου, με την οποία έχει την ελάχιστη διαφορά.

Αυτός ο κανόνας θα σταθεί ιδιαίτερα χρήσιμος στο χειρισμό των ερωτημάτων εύρους (range queries), παρέχοντας ένα ισχυρό εργαλείο απ'ευθείας εξαιρέσεων μεγάλων τμημάτων του δένδρου από τη διαδικασία αναζήτησης ομοιότητας.

4.3.1. Θεώρημα Chiron

Η λογική με την οποία δομήθηκε το δέντρο μας είχε ως βασικό στόχο την ‘τοποθέτηση’ των ομάδων αντικειμένων πορείας έτσι ώστε να αντανakλώνται οι γειτνιάσεις ανάμεσά τους με τέτοιο τρόπο ώστε να είναι δυνατή η κατά προσέγγιση εκτίμηση της διαφοράς πορείας δύο ομάδων. Αυτό επιτυγχάνεται με τη χρήση των ενδιάμεσων διαφορών οι οποίες ‘συναντώνται’ στο μονοπάτι που ενώνει το εκάστοτε ζεύγος ομάδων, βάση του παρακάτω θεωρήματος²:

Θεώρημα Chiron: Σε ένα δένδρο Chiron, εάν η διαφορά πορείας ανάμεσα στην ομάδα αντικειμένων πορείας A και στην ομάδα αντικειμένων πορείας C είναι cd_A και η διαφορά πορείας ανάμεσα στην ομάδα αντικειμένων πορείας B και την ομάδα αντικειμένων πορείας C είναι cd_B , η διαφορά πορείας (cd) ανάμεσα στην A και B δεν μπορεί να είναι μεγαλύτερη από $cd_A + cd_B$ (MaxDif) και μικρότερη από $|cd_A - cd_B|$ (MinDif). [Με άλλα λόγια ισχύει, $|cd_A - cd_B| \leq cd \leq cd_A + cd_B$].

² Όπως είναι φανερό, η ανισότητα του θεωρήματος ταυτίζεται με την τριγωνική ανισότητα. Λόγω όμως του ότι οι διαφορές πορείας των ομάδων αντικειμένων δεν αντιστοιχούν σε Ευκλείδειες ή άλλες αποστάσεις όπου αποδεδειγμένα ισχύει η τριγωνική ανισότητα, η ισχύς της ανισότητας μελετήθηκε από την αρχή και αποδείχτηκε για τις δεδομένα της πρότασης Chiron.

Αν το $G10[k]$ αντιστοιχεί στο k -στο ψηφίο του $G10$ το οποίο είναι το ένα ψηφίο στο οποίο διαφέρει από την $G8$ (στο ψηφίο $G8[k]$) και $G12[i]$ και $G12[j]$ είναι τα δύο ψηφία της $G12$ τα οποία διαφέρουν από τα αντίστοιχα της $G8$ (στα ψηφία $G8[i]$ και $G8[j]$ αντίστοιχα) τότε:

<i>Εάν $k = i$ τότε</i> (ομοίως εάν $k=j$)	αν $G10[k]=G12[i]$ τότε ενώ $G10[j]=G8[j] \wedge G8[j] \neq G12[j] \Rightarrow$ $G10[j] \neq G12[j] \Rightarrow \underline{\text{διαφορές}=1}$ αν $G10[k] \neq G12[i]$ ($1^{\text{η}}$ διαφορά) τότε ενώ $G10[j]=G8[j] \wedge G8[j] \neq G12[j] \Rightarrow$ $G10[j] \neq G12[j]$ ($2^{\text{η}}$ διαφορά) $\Rightarrow \underline{\text{διαφορές}=2}$
<i>αλλιώς ($k \neq i, j$)</i>	$G12[k]=G8[k] \wedge G8[k] \neq G10[k] \Rightarrow$ $G12[k] \neq G10[k]$ ($1^{\text{η}}$ διαφορά) $G10[i]=G8[i] \wedge G8[i] \neq G12[i] \Rightarrow$ $G10[i] \neq G12[i]$ ($2^{\text{η}}$ διαφορά) $G10[j]=G8[j] \wedge G8[j] \neq G12[j] \Rightarrow$ $G10[j] \neq G12[j]$ ($3^{\text{η}}$ διαφορά) $\Rightarrow \underline{\text{διαφορές}=3}$

Σχήμα 4.5. : Παράδειγμα για την απόδειξη του Θεωρήματος

Απόδειξη:

Ας υποθέσουμε πως αναζητούμε τη διαφορά πορείας ανάμεσα στην $G10$ και στην $G12$, του δένδρου Chiron του Σχήματος 4.4. Εφόσον η $G10$ διαφέρει από την $G8$ μόνο σε ένα(1) ψηφίο (έστω στο k -στο ψηφίο) και η $G12$ σε δύο(2) ψηφία (έστω στο i -στό και στο j -στο ψηφίο), μόνο δύο περιπτώσεις είναι δυνατόν να ισχύουν (βλ. την περίπτωση 'Εάν-τότε' του Σχήματος 4.5.). Είτε ένα από αυτά τα δύο διαφορετικά ψηφία του $G12$ αντιστοιχεί στο ένα διαφορετικό ψηφίο του $G10$ (π.χ. $k=i$) ή όχι. Στην πρώτη περίπτωση, υπάρχουν πάλι δύο υποπεριπτώσεις: είτε $G10[k]^3=G12[i]$ ή όχι. Στην πρώτη πάλι περίπτωση, αυτό το ψηφίο εξαιρείται από την πιθανή τους διαφορά ενώ το άλλο (δηλ. το j -στο) παραμένει (ως διαφορά) σε κάθε περίπτωση, αφού στην ομάδα $G10$ είναι ίδιο με το αντίστοιχο της ομάδας $G8$ ενώ στην $G12$ αυτό το ψηφίο αποτελεί μία από τις διαφορές της ομάδας με την $G8$. Συνεπώς, η τελική διαφορά ανάμεσα στην ομάδα $G10$ και την ομάδα $G12$ ισούται με 1 ($|cd_A - cd_B|$ και συγκεκριμένα $cd_{12}-cd_{10}=2-1=1$). Εάν τώρα, ισχύει ότι $G10[k] \neq G12[i]$, τότε η διαφορά μεταξύ τους αυξάνεται σε 2.

Επιστρέφοντας τώρα αναδρομικά (recursively) στην περίπτωση όπου το k είναι διαφορετικό και από το i και από το j , τα αντίστοιχα k -στα, i -στα και j -στα ψηφία είναι διαφορετικά μεταξύ τους και για τις δύο ομάδες (βλ. την περίπτωση 'αλλιώς' του Σχήματος 4.5.). Συνεπώς, εδώ οι διαφορές προστίθενται, δίνοντας τελικά μία διαφορά ίση με τρία(3) (cd_A+cd_B και συγκεκριμένα $cd_{12}+cd_{10}=2+1=3$).

Με βάση το παραπάνω θεώρημα, είναι εύκολο κανείς να συμπεράνει πως παρόλο που η

³ Το $G10[k]$ αντιστοιχεί στο k -στο ψηφίο της ομάδας $G10$. Γενικά, το $Gx[i]$ αντιπροσωπεύει το i -στο ψηφίο της ομάδας Gx .

διαφορά πορείας cd ανάμεσα στις ομάδες G10 και G12 δεν αναπαριστάται απευθείας, είναι πολύ εύκολο να οριοθετηθεί και συνεπώς να εκτιμηθεί προσεγγιστικά.

Πριν κλείσει αυτή την ενότητα, δίνεται η περιγραφή του πραγματικού περιεχομένου του κόμβου του δένδρου Chiron:

Ομάδα Αντικειμένου Πορείας (Group), κωδ₁, κωδ₂, ..., κωδ_n, cd_p, cd_r

υποδεικνύοντας την ομάδα πορείας αντικειμένου που αντιστοιχεί στον κόμβο, τους κωδικούς των χαρακτηριστικών που αντιπροσωπεύονται από τη συγκεκριμένη ομάδα, τη διαφορά πορείας (cd_p) ανάμεσα στον κόμβο και στο γονέα του και τέλος το άθροισμα (cd_r) των διαφορών που απαντώνται στο μονοπάτι ανάμεσα στη ρίζα του δένδρου και τη συγκεκριμένη ομάδα. Η τελευταία αυτή τιμή θα σταθεί πολύ χρήσιμη στη συνέχεια, αφού είναι απαραίτητη στους περισσότερους υπολογισμούς. Όπως είναι φανερό η ύπαρξή της ως προϋπολογισμένης αποφέρει μεγάλο χρονικό όφελος απόδοσης.

4.3.2. Η χρήση του δένδρου Chiron

Τα ερωτήματα(queries) ομοιότητας διακρίνονται σε *ερωτήματα εύρους (range queries)* και *ερωτήματα k-κοντινότερου-γείτονα (k-nearest-neighbor)*. Σ' αυτή την ενότητα θα ασχοληθούμε με το πρώτο είδος, ενώ ο χειρισμός του δεύτερου είδους μπορεί να βασιστεί στα αποτελέσματα του πρώτου.

Στην περίπτωση του δένδρου Chiron, τα ερωτήματα εύρους αφορούν την αναζήτηση αντικειμένων πορείας χαρακτηριστικού τα οποία είναι σχεδόν ίδια (με μία ανοχή του ερωτήματος, qt (query tolerance)), με ένα δεδομένο αντικείμενο πορείας χαρακτηριστικού (roI). Τα βήματα που ακολουθούνται για την επιλογή των αντικειμένων που αντιστοιχούν στα χαρακτηριστικά (properties) τα οποία συνιστούν το αποτέλεσμα (output) ενός τέτοιου ερωτήματος, είναι:

- α) *Εντοπισμός Θέσης (positioning)*. Εντοπισμός της θέσης στο δένδρο Chiron, η οποία αντιστοιχεί στο δεδομένο αντικείμενο roI , εάν αυτό έπρεπε να ενταχθεί στο δένδρο. Η διαδικασία εντοπισμού θέσης αξιοποιεί το θεώρημα Chiron προκειμένου να αποφευχθεί μεγάλο μέρος άσκοπων ελέγχων ομοιότητας σε περιοχές του δένδρου.
- β) *Εντοπισμός των Σχεδόν Ίδιων με το roI Αντικειμένων Πορείας καθώς και των αντίστοιχων χαρακτηριστικών*

4.3.2.1. Εντοπισμός Θέσης (Positioning)

Έστω ότι αναζητείται η θέση του δοθέντος αντικειμένου πορείας roI , σε ένα δένδρο Chiron με απόλυτη ανοχή δένδρου ίση με t . Πρώτα απ' όλα υπολογίζεται η διαφορά πορείας ανάμεσα στο roI και στην ομάδα της ρίζας του δένδρου (έστω cd_{r1}). Εάν είναι ίση με μηδέν ($cd_{r1}=0$), τότε ο κωδικός χαρακτηριστικού του roI προστίθεται στην ομάδα της ρίζας και η διαδικασία σταματά εκεί. Διαφορετικά, εάν αυτή η διαφορά είναι μικρότερη από την ανοχή t ($cd_{r1} \leq t$), τότε αυτό σημαίνει πως το roI θα πρέπει να βρίσκεται στα 'παιδιά' της ρίζας. Για αυτό το λόγο, αρχικά υπολογίζονται οι αντίστοιχες διαφορές ανάμεσα στο roI και σε κάθε κόμβο παιδί της ρίζας. Εάν υπάρχει κάποια ομάδα(κόμβος) από αυτές με διαφορά πορείας (cd_i) ίση με μηδέν ($cd_i=0$), τότε ο κωδικός χαρακτηριστικού του roI προστίθεται στην αντίστοιχη ομάδα (κόμβος) και η διαδικασία σταματά. Εάν δεν υπάρχει καμία τέτοια ομάδα τότε δημιουργείται ένας νέος κόμβος-ομάδα παιδί της ρίζας, σύμφωνα με τους κανόνες του

δένδρου Chiron.

Επιστρέφοντας τώρα στην πρώτη διαφορά πορείας που αφορούσε τη ρίζα του δένδρου, εάν η $cdr1$ είναι μεγαλύτερη της ανοχής t (δηλ. $cdr1 > t$) τότε το $ro1$ αποκλείεται να βρίσκεται στα ‘παιδιά’ της ρίζας αλλά κάπου πιο κάτω στο δένδρο. Εδώ ακριβώς μπορεί να αξιοποιηθεί το θεώρημα Chiron έτσι ώστε να γίνει ένας προσεγγιστικός υπολογισμός/εκτίμηση της περιοχής του δένδρου όπου είναι δυνατόν να βρίσκεται η σωστή θέση για το $ro1$. Πιο συγκεκριμένα, μπορούμε να βασιστούμε στο θεώρημα Digenis⁴ και ο έλεγχος να μεταφερθεί με ασφάλεια στο επίπεδο $lev1 = cdr1 \text{ DIV } t$. Εάν αυτό είναι μεγαλύτερο από τον συνολικό αριθμό των επιπέδου του δέντρου, η διαδικασία σταματά, πράγμα που σημαίνει πως δεν υπάρχει κατάλληλη θέση για το q στο δένδρο. Διαφορετικά, υπολογίζονται τα $MinDif = |cdr1 - cd_r|$ για κάθε κόμβο του επιπέδου.

Για όσους κόμβους ισχύει $MinDif \leq t^5$ (πρώτος έλεγχος) υπολογίζονται οι διαφορές τους με το $ro1$.

Εάν υπάρχει κάποια ομάδα με μηδενική διαφορά πορείας με το $ro1$, τότε ο κωδικός του χαρακτηριστικού που αντιστοιχεί στο $ro1$ προστίθεται στην αντίστοιχη ομάδα

Αλλιώς Εάν βρεθούν ομάδες με διαφορά πορείας μικρότερη από το t , τότε εντοπίζεται η ομάδα με τη μικρότερη διαφορά (έστω $ommin$) και στη συνέχεια υπολογίζονται οι αντίστοιχες διαφορές πορείας (cd_i) ανάμεσα στο $ro1$ και σε κάθε μία από τις ομάδες/κόμβους παιδιά αυτής της ομάδας. Εάν υπάρχει – όπως παραπάνω – μία ομάδα με μηδενική διαφορά πορείας ($cd_i=0$) τότε ο κωδικός του χαρακτηριστικού που αντιστοιχεί στο $ro1$ προστίθεται στην αντίστοιχη ομάδα. Διαφορετικά δημιουργείται μία νέα ομάδα/κόμβος και τοποθετείται ως παιδί της ομάδας $ommin$, σύμφωνα με τους κανόνες κατασκευής του δένδρου Chiron.

Αλλιώς (δηλ. εάν δε βρεθούν ομάδες με διαφορά πορείας μικρότερη από το t), η διαδικασία συνεχίζει με τον ίδιο τρόπο, με τον υπολογισμό των $MinDif$ για κάθε κόμβο του επόμενου επιπέδου και τους αναγκαίους υπολογισμούς και ελέγχους, με τη λογική που περιγράφηκε παραπάνω, μέχρι να βρεθεί η κατάλληλη στο δένδρο θέση για το $ro1$ (εάν υπάρχει) ή να τελειώσει το δένδρο (πράγμα που σημαίνει πως δεν υπάρχει κατάλληλη θέση για το $ro1$ σ’ αυτό).

Αλλιώς (εάν δεν υπάρχει $MinDif \leq t$ για κανένα κόμβο στον πρώτο έλεγχο) η διαδικασία θα συνεχίσει όπως προηγουμένως, με τον υπολογισμό των $MinDif$ του επόμενου επιπέδου, μέχρι να βρεθεί η κατάλληλη στο δένδρο θέση για το $ro1$ (εάν υπάρχει) ή να τελειώσει το δένδρο (πράγμα που σημαίνει πως δεν υπάρχει κατάλληλη θέση για το $ro1$ σ’ αυτό).

Εάν τελικά βρεθεί η κατάλληλη θέση (δηλ. η ομάδα αντικειμένων πορείας) – έστω qg – του $ro1$ σε ένα συγκεκριμένο επίπεδο – έστω ql – και το απαιτεί η εφαρμογή, τότε το $Ro1$ τοποθετείται στη θέση του. Αλλιώς (η πιο συνηθισμένη περίπτωση) η θέση αυτή γίνεται

⁴ **Θεώρημα Digenis**(προσαρμοσμένο στο δένδρο Chiron): Δοθέντος ενός αντικειμένου q με απόσταση d από τη ρίζα του δέντρου Digenis και μίας ανοχής δέντρου t , κανένα επίπεδο πριν από το επίπεδο $lev1$ ($lev1 = (cdr1 \text{ DIV } t)$) δεν είναι δυνατόν να μπορεί να αποτελεί σωστή θέση για το q .

⁵ Λόγω της κατασκευής του δέντρου, η $MinDif$ είναι η μικρότερη δυνατή τιμή διαφοράς πορείας στο επίπεδο. Εάν $MinDif > t$, τότε σίγουρα κάθε τιμή διαφοράς ($cdr_i > MinDif$) με το $ro1$ στο επίπεδο, θα είναι μεγαλύτερη από το t , συνεπώς το $ro1$ αποκλείεται να έχει θέση στο συγκεκριμένο επίπεδο του δένδρου. Στην αντίθετη συνθήκη (δηλ. $MinDif \leq t$) υπάρχει πιθανότητα οι διαφορές πορείας (οι οποίες θα είναι μεγαλύτερες ή ίσες του $MinDif$) να είναι μικρότερες του t , πράγμα που από εδώ και πέρα μόνο με ακριβείς υπολογισμούς μπορεί να εκτιμηθεί).

θεωρητικά γνωστή, προκειμένου να χρησιμοποιηθεί στο επόμενο βήμα αναζήτησης του αρχικού ερωτήματος εύρους.

Πριν περάσουμε στην περιγραφή αυτού του επόμενου βήματος, αξίζει να σημειωθεί πως ο *εντοπισμός θέσης (positioning)* συνιστά μία διαδικασία η οποία εκτός από την *εισαγωγή (insertion)* και την *αναζήτηση* – όπως παραπάνω – χρησιμοποιείται επίσης και για τη *διαγραφή (deletion)* και την *ενημέρωση (update)*.

4.3.2.2. Εντοπισμός των Όμοια Εξελισσόμενων Χαρακτηριστικών (Ομοιότητα Chiron)

Με βάση τα αποτελέσματα του παραπάνω βήματος, η διαδικασία εντοπισμού των σχεδόν ίδιων με το $ro1$ αντικείμενων πορείας (και των αντίστοιχων χαρακτηριστικών) περιλαμβάνει τις παρακάτω περιπτώσεις:

Α. Εάν η ανοχή t του δένδρου είναι ίση με την ανοχή qt του ερωτήματος ($qt=t$), τότε σχεδόν ίδια του qg^6 αντικείμενα πορείας είναι:

- α) Ο ‘γονέας’ του pg .
- β) Τα ‘παιδιά’ του pg .
- γ) Ένα υποσύνολο ομάδων αντικειμένων που βρίσκονται στο προηγούμενο, στο τρέχον (ql) και στο επόμενο επίπεδο και πιο συγκεκριμένα, τα ‘αδέλφια’, οι ‘θείοι’ του προηγούμενου επιπέδου, τα ‘εξαδέλφια’ του τρέχοντος επιπέδου και τα ‘ανίψια’ του επόμενου επιπέδου του ql , για τα οποία η διαφορά πορείας από το qg είναι μικρότερη ή ίση του qt .
- δ) Τίποτα άλλο δεν μπορεί να είναι σχεδόν ίδιο με το qg .

Απόδειξη:

Βασιζόμενοι στον ορισμό του δένδρου Chiron, οι περιπτώσεις α) και β) είναι αυταπόδεικτες. Για την περίπτωση γ) και δ) ο συλλογισμός είναι ο εξής: Εφόσον στις περισσότερες περιπτώσεις η κατάσταση είναι ανάλογη αυτής που εμφανίζεται στο Σχήμα 4.3., το δένδρο θα δημιουργηθεί από το γράφο σύμφωνα με τους κανόνες που ορίστηκαν. Αρχικά επιλέγεται η ρίζα και το δεύτερο επίπεδο εξαρτάται από αυτή την επιλογή, καθώς αποτελείται από τις ομάδες που συνδέονται με την ομάδα της ρίζας στο γράφο. Συνεπώς εάν αυτές οι ομάδες (του δεύτερου επιπέδου) συνδέονταν μεταξύ τους στο γράφο, αυτό δε φαίνεται στο δένδρο γιατί εάν υπήρχαν μεταξύ τους συνδέσεις –στο ίδιο επίπεδο – θα έπαυε η δομή να είναι δένδρο. Παρόλα αυτά εάν χρησιμοποιηθεί το θεώρημα Chiron, είναι δυνατή η ανίχνευση στο δένδρο αυτών των αρχικών συνδέσεων που υπήρχαν στο γράφο και υποδηλώνουν ομοιότητα Chiron. Γενικά, ας υποθέσουμε πως έχουμε δύο ομάδες αντικειμένων πορείας x και y , οι οποίες είναι συνδεδεμένες μεταξύ τους στον αρχικό γράφο, πράγμα που σημαίνει πως είναι σχεδόν ίδεις. Όταν η ομάδα x έχει μόλις τοποθετηθεί στο επίπεδο $x1$ – άρα το $x1$ είναι υπό κατασκευή – ενώ η ομάδα y δεν έχει τοποθετηθεί ακόμη στο δένδρο, τότε ένα από τα παρακάτω μπορεί να συμβεί:

ι) Η ομάδα y έχει μία ή περισσότερες συνδέσεις στο γράφο με ομάδες του επιπέδου $x1-1$ στο δένδρο, οπότε συνδέεται με την ομάδα(κόμβο) – έστω z – του επιπέδου $x1-1$, με την οποία έχει την ελάχιστη διαφορά πορείας. Εδώ υπάρχουν δύο υποπεριπτώσεις:

- α) Η ομάδα z είναι η ομάδα ‘γονέας’ της ομάδας x , οπότε οι ομάδες x και y είναι ‘αδέλφια’ (π.χ. οι ομάδες $G10$ και $G4$ στα Σχήματα 4.3. και 4.4.).
- β) Η ομάδα z δεν είναι η ομάδα πατέρας της ομάδας x , οπότε οι ομάδες x και y είναι

⁶ Και συνεπώς και του $ro1$.

‘εξαδέλφια’ του ίδιου επιπέδου (π.χ. οι ομάδες G18 και G20 στα Σχήματα 3 και 4).

ii) Η ομάδα y δεν έχει καμία σύνδεση - στο γράφο - με ομάδα που κείται στο επίπεδο $x1-1$ του δένδρου. Τότε:

α) εάν στο επίπεδο $x1$ η ελάχιστη διαφορά της ομάδας y είναι η διαφορά της με την ομάδα x , τότε συνδέεται ως ‘παιδί’ της ομάδας x (και η ομάδα x είναι ο ‘γονέας’ της y).

διαφορετικά β) η ομάδα y συνδέεται με μία άλλη ομάδα, έστω τη z , του επιπέδου $x1$ - εκείνη με την οποία έχει τη μικρότερη διαφορά πορείας. Εφόσον η ομάδα z βρίσκεται στο ίδιο επίπεδο με την ομάδα x , η ομάδα y είναι μία ‘ένα επίπεδο κάτω ανιψιά’ της ομάδας x ή - αντίστροφα - η ομάδα x είναι ‘ένα επίπεδο πάνω θεία’ της ομάδας y . (π.χ. οι ομάδες G18 και G1 στα Σχήματα 4.3. και 4.4.).

Με άλλα λόγια, καταλήγουμε στο παρακάτω συμπέρασμα:

Λήμμα 1: Στο δένδρο Chiron, τα αντικείμενα που είναι σχεδόν ίδια με ένα δοθέν αντικείμενο, βρίσκονται στο ίδιο, στο προηγούμενο και στο επόμενο επίπεδο και μόνο εκεί.

Λήμμα 2: Τα όμοια χαρακτηριστικά ενός αντικειμένου του δένδρου Chiron, είναι εκείνα που βρίσκονται στην ίδια ομάδα με αυτό και εκείνα που αντιστοιχούν στα σχεδόν όμοια αντικείμενά του.

Συμπερασματικά, η αντίστοιχη διαδικασία για τον εντοπισμό των σχεδόν ίδιων αντικειμένων πορείας πρέπει να βρει:

α) Το ‘γονέα’ και τα ‘παιδιά’ του qg , έργο ιδιαίτερα εύκολο, και

β) Εκείνα τα ‘αδέλφια’, τα ‘εξαδέλφια’ του ίδιου επιπέδου, τους ‘θείους’ του προηγούμενου επιπέδου και τα ‘ανίψια’ του επόμενου επιπέδου, των οποίων η διαφορά πορείας από την qg είναι μικρότερη ή ίση του qt . Ο αλγόριθμος που θα ακολουθηθεί για αυτό το σκοπό θα περιγραφεί εδώ πολύ συνοπτικά γιατί είναι ιδιαίτερα εύκολος. Και εδώ θα χρησιμοποιηθεί το θεώρημα Chiron προκειμένου να αποφευχθεί μεγάλος αριθμός υπολογισμών διαφορών ανάμεσα στην qg και στις αντίστοιχες ομάδες των εμπλεκόμενων επιπέδων. Πιο συγκεκριμένα, για κάθε ομάδα(g_i) στο προηγούμενο, τρέχον και επόμενο επίπεδο - εκτός από τις ομάδες ‘γονιός’ και ‘παιδιά’ τα οποία έτσι κι αλλιώς είναι σχεδόν ίδια με την qg - υπολογίζεται η αντίστοιχη $MinDif$ (δηλ. $|cdr_{g_i} - cdr_{qg}|$), για την αποφυγή - όπως και παραπάνω - άσκοπων υπολογισμών διαφορών πορείας. Για όποια ομάδα η αντίστοιχη $MinDif$ είναι μικρότερη ή ίση με την ανοχή qt , τότε υπολογίζεται η πραγματική διαφορά πορείας (cd_q) ανάμεσα στην ομάδα qg και την ομάδα. Για όσες ομάδες και η cd_q είναι μικρότερη ή ίση με το qt ($cd_q \leq qt$), συμπεραίνεται ότι είναι σχεδόν ίδια με την ομάδα qg και συνεπώς συμπεριλαμβάνεται στο αποτέλεσμα (output) που δίνεται ως απάντηση στο αρχικό ερώτημα εύρους.

Β. Εάν $qt < t$, τότε ακολουθούμε τον ίδιο αλγόριθμο όπως στο Α., αλλά όπου γίνεται αναφορά σε ‘παιδιά’ κάποιας ομάδας, δεν θα εννοούνται πια όλα αλλά μόνο εκείνα των οποίων η διαφορά από το ‘γονέα’ τους είναι μικρότερη ή ίση με qt .

Γ. Εάν $qt > t$, τότε το θέμα που προκύπτει είναι ιδιαίτερα σύνθετο, αφού η αναζήτηση πρέπει ανάλογα με τις εκάστοτε ανάγκες να μετακινηθεί σε επίπεδα περισσότερα του ενός, πάνω και κάτω από το $q1$, υποβαθμίζοντας αρκετά την απόδοση. Από την άλλη πλευρά, εκτιμούμε πως ακόμη και αν προτεινόνταν κάποια άλλη αποδοτικότερη για την περίπτωση τεχνική, είναι αμφίβολο το κατά πόσο θα άξιζε τον κόπο, αφού καθώς

απομακρυνόμαστε από το t το οποίο έχει προταθεί ως σχετική εγγύηση ομοιότητας από ειδικούς της εκάστοτε εφαρμογής, το πιο πιθανό είναι να αποκομίσουμε άχρηστη πληροφορία ομοιότητας. Σ' αυτή την περίπτωση, πιθανότατα τα 'ακριβά' στον εντοπισμό σχεδόν ίδια αντικείμενα πορείας να αντιπροσωπεύουν ανόμοια αντικείμενα. Συνεπώς αυτό που προτείνουμε για τέτοια ερωτήματα σε αυτή τη φάση είναι ένα αποτέλεσμα σχετικό με $qt=t$, πληροφορώντας το χρήστη πως οι προτεινόμενες ομάδες αντικειμένων πορείας (και τα αντίστοιχα χαρακτηριστικά) είναι σχεδόν ίδιες με το αντικείμενο του ερωτήματος ($ro1$), σύμφωνα με την τυποποίηση της εφαρμογής.

4.4. Πειραματική Μελέτη - Αξιολόγηση

Για τα πειράματα αξιολόγησης της προτεινόμενης μεθόδου Chiron, χρησιμοποιήθηκαν πλασματικά (synthetic) και πραγματικά δεδομένα προερχόμενα από τη βάση δεδομένων ορμονικών εξετάσεων, που αναφέρθηκε στην αρχή. Πιο συγκεκριμένα, για κάθε έναν από τους ασθενείς της βάσης υπήρχε ένα σύνολο ορμονικών εξετάσεων (τιμών), για τις οποίες επιχειρήθηκε η αποδοτική ανεύρεση αλληλεξαρτήσεών τους, μέσω της μεθόδου Chiron.

Η βάση δεδομένων περιελάμβανε 25 σύνολα εξετάσεων ανά ασθενή (σε ένα σύνολο 80 ασθενών) οπότε για κάθε ασθενή δημιουργήθηκε ένα δένδρο Chiron με 25 το πολύ κόμβους. Λόγω όμως του γεγονότος ότι τα αντικείμενα πορείας πολλών χαρακτηριστικών ήταν ίδια, δημιουργήθηκαν αρκετές ομάδες τέτοιων αντικειμένων, κατεβάζοντας τον αριθμό των κόμβων ανά δένδρο σε 4 έως 18 το πολύ και τα επίπεδα των δένδρων σε 8 το μέγιστο. Επομένως, παρόλο που τα απαιτούμενα δένδρα ήταν αρκετά (80), το ύψος τους ήταν χαμηλό καθιστώντας την αναζήτηση σ' αυτά εύκολη και γρήγορη. Ακόμη όμως κι αν - ανάλογα με την περίπτωση - ο αριθμός των εξετάσεων (δεδομένων) αυξηθεί πολύ, η απόδοση μπορεί να παραμείνει ικανοποιητική, αφού συνήθως τα ερωτήματα αφορούν ένα μικρό υποσύνολο των αρχικών αντικειμένων (π.χ. ασθενών στο παράδειγμά μας). Επιπλέον, η αναζήτηση ομοιοτήτων Chiron στο σύνολο των αντικειμένων (π.χ. όλων των ασθενών) θα είναι μία περιοδικά εκτελούμενη διεργασία, οπότε μπορεί να 'τρέχει' μαζικά (batch).

Όπως αναφέρθηκε και στην αρχή του κεφαλαίου, το παράδειγμά μας επιλέχθηκε με κριτήριο τις δυνατότητες επαλήθευσης των αποτελεσμάτων. Με δεδομένο ότι υπάρχει αρκετή γνώση για την αλληλεπίδραση των ορμονών, ήμασταν πράγματι σε θέση να συγκρίνουμε αυτή τη γνώση με τα συμπεράσματα που προέκυψαν από τη χρήση της μεθόδου Chiron. Το αποτέλεσμα ήταν πως υπήρξε πλήρης συμφωνία, γεγονός που αποδεικνύει *την εγκυρότητα και την ακρίβεια της μεθόδου*. Αναφέρουμε ένα απλό παράδειγμα που αφορά τη συμπεριφορά των διαστάσεων που αντιστοιχούν στις μετρήσεις του θυρεοειδούς. Οι αλληλεπιδρώσεις ορμόνες TSH (Thyroid Stimulating Hormone) και TRH (Thyrotropin-Releasing Hormone) (που αντιστοιχούσαν στην πρώτη και τη δεύτερη διάσταση του βασικού αντικειμένου της βάσης δεδομένων μας) βρέθηκαν να είναι όντως όμοια εξελισσόμενα χαρακτηριστικά Chiron. Το ίδιο συνέβη και με τις αλληλεπιδρώσεις ορμόνες T3 και T4, οι οποίες αντιστοιχούσαν στην τρίτη και την τέταρτη διάσταση.

Ο δεύτερος δείκτης ποιότητας της πρότασης Chiron είναι η *αποδοτικότητά της*. Τα ιδιαίτερα στοιχεία της αναζήτησης ομοιότητας που εισάγει η μέθοδος, οδηγούν στον ορισμό μιας απλοποιημένης αναπαράστασης κάθε πορείας χαρακτηριστικού ενός αντικειμένου (course object), στην οποία υποδεικνύεται μόνο η αύξηση, η μείωση και η σταθερότητα της τιμής του χαρακτηριστικού σε κάθε προκαθορισμένο χρονικό διάστημα. Αυτό αποτελεί και την πρώτη μείωση πολυπλοκότητας που επιτυγχάνει η μέθοδος. Η δεύτερη μείωση σχετίζεται με τον ορισμό του δένδρου Chiron, αφού οι αρχές κατασκευής του δένδρου εξυπηρετούν τη γρήγορη εκτίμηση διαφορών πορείας ανάμεσα σε ομάδες αντικειμένων πορείας, επιτυγχάνοντας την αποφυγή μεγάλου αριθμού άσκοπων υπολογισμών. Αυτή η εύκολη

εκτίμηση της περιοχής του δένδρου όπου είναι πιθανότερος ο εντοπισμός σχεδόν ίδιων ομάδων αντικειμένων πορείας, οδηγεί σε μεγάλη μείωση του χρόνου απόκρισης (10%-60%) σε σχέση με τη σειριακή αναζήτηση. Τα αντίστοιχα ποσοστά μείωσης αυξάνονται αναλογικά με τον αριθμό των χαρακτηριστικών αντικειμένων (εξετάσεων για το παράδειγμά μας) που διερευνώνται, πράγμα που εξηγείται από το παρακάτω:

Ο συνολικός χρόνος απόκρισης της μεθόδου Chiron περιλαμβάνει:

- α) Τον σχηματισμό των αντικειμένων πορείας, χρησιμοποιώντας τα ψηφία 0,1 και 2.
- β) Τη δημιουργία του δένδρου Chiron
- γ) Την αναζήτηση στο δένδρο.

Εάν τα χαρακτηριστικά για τα οποία γίνεται η αναζήτηση είναι λίγα, τότε η σειριακή αναζήτηση είναι σχετικά γρήγορη σε σχέση με τη μέθοδο Chiron, λόγω της χρονικής επιβάρυνσης από το α) και το β). Αντίθετα, καθώς ο αριθμός των εμπλεκόμενων χαρακτηριστικών αυξάνεται, η χρονική επιβάρυνση των α) και β) αξίζει τον κόπο γιατί οδηγεί σε μεγάλη οικονομία χρόνου λόγω του γ).

Σχήμα 4.6. : Απόδοση σύμφωνα με τον αριθμό των εμπλεκόμενων χαρακτηριστικών.

Σχήμα 4.7. : Απόδοση σύμφωνα με τον αριθμό των εμπλεκόμενων αντικειμένων.

Μία άλλη σχετική παρατήρηση είναι ότι η απόδοση της σειριακής αναζήτησης είναι εντελώς τυχαία, κυμαινόμενη μεταξύ 1 και n (n : αριθμός των χαρακτηριστικών), ενώ η αντίστοιχη απόδοση της μεθόδου Chiron διαφοροποιείται ανάλογα με το χρόνο αναζήτησης (scan time) στο δένδρο, ο οποίος καθορίζεται κυρίως από τα επίπεδα τα οποία πρέπει να ελεγχθούν για τον εντοπισμό του κατάλληλου κόμβου(ομάδας). Αυτά τα επίπεδα κυμαίνονται από 1 έως $\log n$, σχηματίζοντας ένα προφανώς μικρότερο πεδίο αναζήτησης από εκείνο της σειριακής μεθόδου.

Όλα αυτά αντανakλώνται στα αποτελέσματα των πειραμάτων μας (Σχήματα 4.6. και 4.7.), όπου ως 5-διάστατα και 15-διάστατα καθώς και ως βάσεις δεδομένων 40 και 80 αντικειμένων, χρησιμοποιήθηκαν τμήματα της αρχικής μας βάσης δεδομένων.

Τέλος πρέπει να επισημανθεί πως ένα σημαντικό χαρακτηριστικό του δένδρου Chiron – όπως φαίνεται και από τα πειράματα – είναι πως δεν ‘πλήττεται’ από την ‘κατάρα’ των πολλών διαστάσεων [A26], δηλαδή η απόδοσή του δεν υποβαθμίζεται όσο αυξάνεται ο αριθμός των εμπλεκόμενων διαστάσεων (χαρακτηριστικών). Αυτό ήταν αναμενόμενο, αφού η ‘κατάρα’ των πολλών διαστάσεων σχετίζεται με δένδρα ομοιότητας τα οποία βασίζονται σε συναρτήσεις υπολογισμού αποστάσεων (όπως η Ευκλείδεια απόσταση), ενώ το δένδρο Chiron χρησιμοποιεί απλές διαφορές ανάμεσα σε αντικείμενα.

4.5. Ανοικτά Θέματα – Μελλοντικές Κατευθύνσεις

Σε αυτή την ενότητα εισάγεται αρχικά η έννοια της ομοιότητας στις πορείες των χαρακτηριστικών των αντικειμένων και στη συνέχεια μία πρώτη προσέγγιση (μέθοδος Chiron) για τον αποτελεσματικό εντοπισμό και την οργάνωση αυτού του είδους της ομοιότητας. Η έρευνα αυτή βρίσκεται στην αρχή και αφήνει πολλά ζητήματα ανοικτά για διερεύνηση και επέκταση. Έτσι στην επόμενη φάση της έρευνάς μας θα εστιάσουμε στα παρακάτω (με σειρά σπουδαιότητας) θέματα:

- ♦ *Στατιστική μελέτη των αποτελεσμάτων.* Πιο συγκεκριμένα, εάν δύο χαρακτηριστικά αλλάζουν με παρόμοιο τρόπο στην πορεία του χρόνου για σημαντικό ποσοστό των αντικειμένων της συγκεκριμένης κλάσης της εκάστοτε εφαρμογής (π.χ. ασθενών στην εφαρμογή μας), αυτό σημαίνει ότι υπάρχει μία σταθερή εξάρτηση ανάμεσά τους. Συνεπώς εκτιμούμε πως με βάση τη μέθοδο Chiron, θα πρέπει να διεξαχθεί μία λεπτομερής μελέτη για τον προσδιορισμό αυτού του ποσοστού, με σκοπό την ολοκλήρωση των αποτελεσμάτων των αντίστοιχων εφαρμογών.
- ♦ *Επέκταση της μεθόδου Chiron για την υποστήριξη της ανίχνευσης των αντικειμένων πορείας χαρακτηριστικών τα οποία αναπτύσσονται με αντίστροφο τρόπο⁷.* Μία πρώτη μελέτη εκτίμηση των αναγκών μίας τέτοιας επέκτασης έχει ήδη πραγματοποιηθεί και το πρώτο συμπέρασμα είναι πως πρόκειται για ένα όχι ιδιαίτερα σύνθετο πρόβλημα, παρόλη τη σπουδαιότητά του.
- ♦ *Επέκταση της πρότασης Chiron, για τη διαχείριση των μετατοπίσεων χρόνου (time shifts) ανάμεσα στα αντικείμενα πορείας χαρακτηριστικών.* Ο σκοπός αυτής της επέκτασης είναι η διερεύνηση του κατά πόσο τα χαρακτηριστικά επηρεάζουν το ένα το άλλο με κάποια καθυστέρηση.
- ♦ *Επέκταση της μεθόδου Chiron, για χαρακτηριστικά περισσότερων της μίας(1) διάστασης.* Με μία πρώτη εκτίμηση φαίνεται πως θα χρειαστεί να αντληθούν ιδέες από τα παραδοσιακά δέντρα ομοιοτήτων και να συνδυαστούν με τη λογική του δένδρου Chiron.
- ♦ *Βελτιστοποίηση του δένδρου Chiron.* Το δένδρο Chiron αποτελεί μία πρώτη προσέγγιση ενός δένδρου το οποίο μπορεί να υποστηρίξει την όλη πρόταση, συνεπώς θέματα βελτιστοποίησης δένδρου δεν έχουν εξεταστεί ακόμη. Βασικό ζητούμενο σ’ αυτήν την κατεύθυνση είναι η μετατροπή του δένδρου Chiron σε ισοζυγισμένο δένδρο (balanced tree) για μεγαλύτερη αποτελεσματικότητα και απόδοση [A8].
- ♦ *Διερεύνηση και εισαγωγή νέων λιγότερο ή περισσότερο σύνθετων ορισμών αντικειμένων πορείας χαρακτηριστικών, τα οποία δεν εξαρτώνται μόνο από τις αυξομειώσεις των τιμών των χαρακτηριστικών.* Ειδικότερα, μπορούν να οριστούν συγκεκριμένες κατηγοριοποιήσεις τιμών δεδομένων οι οποίες θα αντιπροσωπεύονται από συγκεκριμένα επιλεγμένα ψηφία. Ένα πολύ απλό τέτοιο παράδειγμα είναι εκείνο όπου δε θα αντιπροσωπεύονται αυξήσεις και μειώσεις

⁷ Όπως στην περίπτωση των ορμονών TSH και T3, T4, για τις οποίες – όπως είναι γνωστό – ισχύει πως όταν αυξάνει η TSH μειώνονται οι T3 και T4.

στις τιμές των χαρακτηριστικών, αλλά μόνο η μεταβολή των τιμών. Σε αυτή την περίπτωση θα χρησιμοποιούνται μόνο τα ψηφία 0 και 1 (για σταθερότητα και μεταβολή αντίστοιχα) πράγμα που κάνει το όλο πρόβλημα αναπαράστασης ιδιαίτερα ευκολότερο και απλούστερο. Αυτή η απλότητα κληρονομείται και στο επίπεδο της διαχείρισης, αφού οδηγεί στη χρήση είτε μιας πολύ απλής παραλλαγής/έκδοσης του δένδρου Chiron ή σε μία απλή αντιμετώπιση κατακερματισμού (hash), με μεγάλα οφέλη απόδοσης.

4.6. Ανακεφαλαίωση – Συμπεράσματα

Η μέθοδος Chiron η οποία παρουσιάστηκε σε αυτό το κεφάλαιο, αποτελεί μία πρώτη προσέγγιση ως *πλαίσιο ορισμού, διαχείρισης δεδομένων και μετα-δεδομένων (metadata) σχετικά με την αναζήτηση ομοιοτήτων ανάμεσα στις πορείες χαρακτηριστικών των αντικειμένων*. Πιο συγκεκριμένα, η έρευνα ξεκίνησε με τη μελέτη κλασσικών και ειδικότερων ομοιοτήτων ανάμεσα σε πολυδιάστατα αντικείμενα καταλήγοντας στο παρακάτω συμπέρασμα: Σε πολλές περιπτώσεις η χρήσιμη πληροφορία ομοιότητας ανάμεσα σε πολυδιάστατα αντικείμενα, δεν αφορά τα ίδια τα αντικείμενα αλλά τα χαρακτηριστικά τους (εκφράζοντας τιμές, ενέργειες, φαινόμενα, κλπ) και κυρίως την εξέλιξη των χαρακτηριστικών τους στην πορεία συγκεκριμένου χρόνου. Η τελευταία μάλιστα πληροφορία μπορεί να ‘πει’ σε έναν ερευνητή ότι συγκεκριμένα χαρακτηριστικά αντικειμένων αλληλεπιδρούν ή έχουν κοινή βάση.

Για την υποστήριξη λοιπόν τέτοιου είδους ομοιοτήτων, ορίσαμε τις απαραίτητες έννοιες και εργαλεία, όπως το ‘*αντικείμενο πορείας χαρακτηριστικού*’ (*Chiron object*), τη ‘*διαφορά πορείας*’ (*Chiron difference*), τα ‘*σχεδόν ίδια αντικείμενα*’, την *ομοιότητα Chiron* και άλλες επιπλέον βοηθητικές έννοιες. Για τη δεικτοδότηση και άρα την ευέλικτη και αποδοτική διαχείριση των αντικειμένων πορείας χαρακτηριστικών, ορίσαμε το αντίστοιχο δένδρο ομοιότητας (*Chiron tree*), το οποίο – όπως αποδείχτηκε – είναι αρκετά αποτελεσματικό. Αυτό το δένδρο χρησιμοποιεί ως βάση της οργάνωσής του, προ-υπολογισμένες διαφορές ανάμεσα στα αντικείμενα πορείας χαρακτηριστικών, γεγονός που παρέχει *μεγάλη μείωση στο χρόνο υπολογισμού (run-time computation) και γρήγορη πλοήγηση (piloting) στο δένδρο*. Ο υπολογισμός τώρα αυτών των διαφορών είναι ιδιαίτερα εύκολος λόγω της ευέλικτης κωδικοποίησης των αλλαγών των τιμών των χαρακτηριστικών αλλά και του τρόπου ορισμού του δένδρου Chiron ο οποίος εξυπηρετεί τη *γρήγορη εκτίμηση* των διαφορών ανάμεσα στα αντικείμενα πορείας.

Όπως γίνεται φανερό από τα παραπάνω, η πρότασή μας μπορεί να αξιοποιηθεί σε πολλά είδη εφαρμογών (ιατρικές, εμπορικές, επιστημονικές κλπ) προάγοντας:

- ✓ *Τη Διάγνωση / Κατηγοριοποίηση*. Στην εφαρμογή-παράδειγμα που χρησιμοποιήσαμε, οι ειδικοί μπορούν να συμπεράνουν ότι κάποια χαρακτηριστικά (π.χ. ορμόνες) αλληλεπιδρούν για κάποιον ασθενή - βασισμένοι στην ομοιότητα ανάμεσα στις πορείες συγκεκριμένων χαρακτηριστικών. Εάν η ίδια ομοιότητα ανιχνευτεί σε πολλούς ή σε όλους τους ασθενείς της βάσης δεδομένων, τότε η αντίστοιχη αλληλεπίδραση μπορεί να αποτελέσει πιθανά κανόνα αξιοποιήσιμο από τη διαγνωστική διαδικασία/γνώση.
- ✓ *Πρόβλεψη*. Έχοντας μία στέρεη ιδέα σχετικά με το πώς αλληλεπιδρούν κάποια χαρακτηριστικά, είναι δυνατή μία αξιόπιστη πρόγνωση μελλοντικών εξελίξεων και κινδύνων καθώς και η διερεύνηση πιθανών αντιμετωπίσεων (π.χ. θεραπειών) τους. Μ’ αυτό τον τρόπο μπορεί να κερδηθεί πολύτιμος χρόνος και κόπος – πολλές φορές – αυξάνοντας την αποτελεσματικότητα της όποιας εφαρμογής σε κάθε

επίπεδο.

Κλείνοντας, αξίζει να υπογραμμίσουμε πως ίσως το σημαντικότερο πλεονέκτημα της μεθόδου-πρότασης Chiron είναι το γεγονός ότι θέτει τις βάσεις μίας πολλά υποσχόμενης προοπτικής επέκτασής της.

4.7. Δημοσιεύσεις του Κεφαλαίου

♦ ΣΥΝΕΛΠΙΟ

[Σ17] Aikaterini Krotopoulou, Dimitra Terpou, Athanasios Tsakalidis, “CHIRON: A Property Based Proposal for Discovery”, IKE 2009 (ranking:0,80).

4.8. Αναφορές του Κεφαλαίου

- [A1] Korn P., Sidiropoulos N., Faloutsos C., Siegel E., Protopapas Z.: Fast and Effective Retrieval of Medical Tumor Shapes. IEEE Transactions on Knowledge and Data Engineering, Vol 10, No 6, 1998.
- [A2] Arya S., D.M. Mount, N.S. Netanyahu R. Silverman, and A. Y. Wu. An Optimal Algorithm for Approximate Nearest Neighbor Searching in Fixed Dimensions. Journal of the ACM, vol. 45, no. 6, pp.891-923
- [A3] Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases. In SIGMOD 1994.
- [A4] Gibbs S., Breiteneder C., Tsichritzis D.: Data Modeling of Time-Based Media. SIGMOD 1994.
- [A5] Bouteldja N., Gouet V., Scholl M.: HiPer: Hierarchical Progressive Exact Retrieval in Multi-dimensional Spaces. SISAP 2008.
- [A6] Fagin R., Kumar R., Sivakumar D.: Efficient similarity search and classification via rank aggregation. SIGMOD 2003.
- [A7] Faloutsos C., Equitz W., Flickner M., Niblack W., Petkovic D., Barber R.: Efficient and effective querying by image content. Journal of Intelligent Information Systems, 3(3):231-262,1994.
- [A8] Hellerstein J.M., Jeffrey F., Naughton J.F., Pfeffer A.: Generalized Search Trees for Database Systems. VLDB 1995.
- [A6] Kamel, I., Faloutsos, C.: Hilbert R-tree: An Improved R-tree Using Fractals. In Very Large Data Bases (VLDB) 1994.
- [A7] Ciaccia, P., Patella, M.: The M2-tree: Processing Complex Multi-Feature Queries with Just One Index. In 1st DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland, 2000.
- [A8] Aggarwal, C.C.: Towards meaningful high-dimensional nearest neighbor search by human-computer interaction. In ICDE 2002.
- [A9] Arge, L., Mark de Berg, Haverkort, H.J., Yi K.: The Priority R-Tree: A Practically Efficient and Worst-Case Optimal R-Tree. In SIGMOD 2004.
- [A10] Augsten, N., Bohlen, M., Gamper, J.: An Incrementally Maintainable Index for Approximate Lookups in Hierarchical Data. In Very Large Data Bases (VLDB) 2006.
- [A11] Chen, Q., Chen, L., Lian, X., Liu, Y., Xu, Yu.: Indexable PLA for Efficient Similarity Search. In Very Large Data Bases (VLDB) 2007.
- [A12] Wong, R.C.-W., Tao, Y., Fu, A.W.-C., Xiao, X.: On Efficient Spatial Matching. In Very Large Data Bases (VLDB) 2007.
- [A13] White, D.A. Jain, R.: Similarity indexing with the SS-tree. In 12th International Conference on Data Engineering (ICDE) 1996.
- [A14] Goh, K.-S., Li, B., Chang, E.: DynDex: A Dynamic and Non-metric Space Indexer. In Multimedia 2002.

- [A15] Qian, G., Zhu, Q., Xue, Q., Pramanik, S.: The ND-Tree: A Dynamic Indexing Technique for Multidimensional Non-ordered Discrete Data Spaces. In *Very Large Data Bases (VLDB)* 2003.
- [A16] Manolopoulos, Y., Nanopoulos, A., Papadopoulos, A., Theodoridis, Y.: *R-Trees: Theory and Applications*, Springer 2005. ISBN 1-85233-977-2.
- [A17] Berchtold, S., Keim, D., Kriegel, H.-P.: The X-tree: An Index Structure for High-Dimensional Data. In *Very Large Data Bases (VLDB)* 1996.
- [A18] Katayama, N., Satoh, S.: The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. In *SIGMOD* 1997.
- [A19] Bronstein, A., Bronstein, M., Bruckstein, A., Kimmel, R.: Partial Similarity of Objects, or How to Compare a Centaur to a Horse, *International Journal of Computer Vision*, 2008.
- [A20] Bruckstein, A. M., Katzir, N., Lindenbaum, M., & Porat, M.: Similarity-invariant signatures for partially occluded planar shapes. *IJCV* 1992, 7(3), 271–285.
- [A21] Kim, C. R., & Chung, C. W.: XMage: An image retrieval method based on partial similarity. *Information Processing and Management*, 2006, 42, 484–502.
- [A22] Negi, T., Bansal V.: Time Series: Similarity Search and its Applications, *ICSCI* 2005.
- [A23] S.Zhang C. S. Perng, H. Wang and D. S. Parker. Landmark: A new technique for similarity based pattern querying in time series databases. In *Proc ICDE*, 2000.
- [A24] Xianping Ge, P. Smyth. Deformable markov model templates for time series pattern matching. *Intl Conf. of SIGKDD*, August 2000.
- [A25] Chan. K and A.W. Fu. Efficient time series matching by wavelets. In *Proc. of the 15th Intl. Conf. on Data Engineering*, pages 126–134, 1999.
- [A26] Weber, R., Schek, H.J, Blott, S.: A Quantitive Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *Very Large Data Bases VLDB*) 1998.

Μέρος Γ':

Επίλογος

Τα προβλήματα είναι ευκαιρίες ντυμένες με τα ρούχα της δουλειάς

Henry Kaiser

Κεφάλαιο 5

Επίλογος

Ερωτήματα - Παρατηρήσεις - Συμπεράσματα Ολοκλήρωση

Περιεχόμενα Κεφαλαίου

- 5.1. Χαρτογράφηση Εγκεφάλου. Ποια είναι η συνολική προσφορά.
- 5.2. Τμηματική Ομοιότητα: Lui k-n-match ή INTESIS;
- 5.3. Επέκταση INTESIS βάση άλλων προτάσεων
- 5.4. Γιατί παρόλο που τα δέντρα Digenis και Chiron μοιάζουν, οι αλγόριθμοι εντοπισμού θέσης σ' αυτά διαφέρουν τόσο;
- 5.5. Αναφορές του Κεφαλαίου

Σύνοψη

Μετά από την παρουσίαση της έρευνάς μας ανά ερευνητικό θέμα-αποτέλεσμα, ολοκληρώνουμε με μία 'ένα επίπεδο πιο πάνω' μελέτη και προβληματισμό πάνω στις προτάσεις μας. Αυτή η εποπτική θεώρηση σκοπό έχει τη συνολική αποτίμηση της έρευνας, αίροντας κάποια πιθανά στεγανά ανάμεσα στα επιμέρους θέματα, επιτρέποντας έτσι την καλύτερη ενοποίηση και την ολοκλήρωσή τους. Οι υπόγειες αλληλεπιδράσεις φωτίζονται, ενώ διερευνώνται, αξιολογούνται και ερμηνεύονται οι σχέσεις και οι αποκλίσεις των συγγενικών μας προτάσεων. Μέσα από αυτή τη διερεύνηση έρχονται επίσης στην επιφάνεια δυνατότητες αλληλοσυμπλήρωσης και πρακτικής συνεργασίας μεταξύ των αποτελεσμάτων μας, για τη βελτίωση και την επέκτασή τους.

5.1. Χαρτογράφηση Εγκεφάλου. Ποια είναι η συνολική προσφορά

Όπως έχει προαναφερθεί, η έρευνά μας σχετικά με τη χαρτογράφηση του ανθρώπινου εγκεφάλου χρονολογείται περίπου στα μέσα της δεκαετίας του 1990, εποχή που αυτό το θέμα δεν είχε τη σημερινή ανάπτυξη. Το αποτέλεσμα αυτής της έρευνας ήταν ο ορισμός σχετικών μοντέλων διαβαθμισμένου τύπου (*σχεσιακό, σημαντικό και αντικειμενοστραφές μοντέλο*), καθένα από τα οποία αποτελούσε εξέλιξη του προηγούμενου.

Οι παρατηρήσεις-πυλώνες για τη συγκεκριμένη πορεία και τα αποτελέσματά μας (αλλά και γενικότερα) ήταν οι εξής:

- Τα *εγκεφαλικά δεδομένα* όντας τρισδιάστατα, πολύπλοκα αντικείμενα, απαιτούν ιδιαίτερους ορισμούς για να αντιπροσωπευτούν αποδοτικά.
- Οι *εγκεφαλικές λειτουργίες και οι αλληλεπιδράσεις* τους αποτελούν εξίσου σύνθετο πεδίο, γι' αυτό χρειάζεται ιδιαίτερη μελέτη για τη δημιουργία δομών υποστήριξης και διερεύνησής τους.

Το αποτέλεσμα αυτών των παρατηρήσεων ήταν:

- α) *Ο ορισμός νέων ενέλικτων τύπων δεδομένων*
- β) *Ο ορισμός εννοιών/δομών για την υποστήριξη και την αξιοποίηση σχέσεων/διατάξεων:*
  - i) *σε επίπεδο δομής μοντέλου δεδομένων* (ιεραρχική ταξινόμηση, ordering, διάταξη/έκδοση με βάση κάποιο κριτήριο (π.χ. πιθανότητα εμφάνισης επιληπτικής εστίας))
  - ii) *σε επίπεδο δομής μοντέλου βάσης γνώσης* (εισαγωγή της έννοιας κλειδιού για τις Βάσεις Γνώσης)
  - iii) *σε επίπεδο ορισμού αρχών μοντέλου* (ορισμός της 'σχέσης με το περιβάλλον' ως βασικού χαρακτηριστικού του αντικειμένου στο αντικειμενοστραφές μοντέλο)

Αυτή η αρχική διαπίστωση της κομβικής αξίας που έχει η αποδοτική διαχείριση σχέσεων, αλληλεξαρτήσεων και εσωτερικών διατάξεων των δεδομένων, κατηύθυνε φανερά αλλά και διαισθητικά το σύνολο της έρευνάς μας. Ακόμη και τα επόμενα θέματα – αποδοτική αναζήτηση διάφορων τύπων ομοιοτήτων – επηρεάστηκαν καθοριστικά από αυτή τη βασική

παρατήρηση, αφού το ενδιαφέρον επεκτάθηκε πλέον και στην ανακάλυψη νέων σχέσεων και εξαρτήσεων που μπορούν να ‘προδοθούν’ από τις αρχικές σχέσεις/διατάξεις ή τα δεδομένα. Αξίζει επίσης να σημειωθεί πως μεταγενέστερες εφαρμογές χαρτογράφησης του ανθρώπινου εγκεφάλου [A1, A2, A3, A4] αλλά και άλλων παρεμφερών πεδίων (π.χ. γεωγραφικών συστημάτων [A5]), κινήθηκαν σε παρόμοιους άξονες - κυρίως στο θέμα των τύπων δεδομένων – γεγονός που προσδίδει μία ‘αντικειμενικότητα’ στις αντίστοιχες επιλογές και προτάσεις μας.

5.2. Τμηματική Ομοιότητα: Lui k-n-match ή INTESIS;

Στο δεύτερο κεφάλαιο προτάθηκαν δύο μέθοδοι για την αποδοτική ανεύρεση τμηματικών ομοιοτήτων. Πιθανώς ο αναγνώστης να αναρωτήθηκε ποια μέθοδος είναι η καλύτερη και εάν όντως κάποια υπερέχει, ποιος είναι ο λόγος που προτάθηκαν και οι δύο.

Για να δοθεί απάντηση σε αυτά τα ερωτήματα, πρέπει να ληφθούν υπ’ όψιν τα εξής:

- i. Η μέθοδος INTESIS έχει οριστεί για μεγάλο μεν αλλά ορισμένο πλήθος διαστάσεων (μέχρι 100), άρα αν και μπορεί να καλύψει τεράστιο αριθμό ανάλογων εφαρμογών, τυπικά δεν είναι απόλυτα γενική μέθοδος.
- ii. Η INTESIS απαιτεί υπεύθυνη προετοιμασία και αποφάσεις από τον υπεύθυνο της εφαρμογής (ορισμό διαστάσεων ανά χαρακτηριστικό, επιλογή κατάλληλου δέντρου ομοιοτήτων, ορισμό ποσοστού p , το οποίο ορίζει το ποσοστό των χαρακτηριστικών στα οποία εάν εντοπιστεί ομοιότητα, δεχόμαστε πως υπάρχει τμηματική ομοιότητα των αντικειμένων), γεγονός που σε κάποιες περιπτώσεις μπορεί να θεωρηθεί αποτρεπτικό για την επιλογή της μεθόδου. Ειδικότερα, υπεισέρχονται θέματα σωστής κρίσης του υπεύθυνου για την επιλογή των παραπάνω παραμέτρων, πράγμα που επηρεάζει και καθορίζει πολλές φορές την ακρίβεια των αποτελεσμάτων.
- iii. Η μέθοδος INTESIS περιλαμβάνει αλλά και βασίζεται σε δεικτοδότηση των αντικειμένων (συγκεκριμένα των υπο-αντικειμένων από τα οποία αποτελούνται τα αρχικά αντικείμενα). Αυτό αποτελεί ένα προσόν που οδηγεί σε διευκόλυνση της ανάκτησης τμηματικά όμοιων αντικειμένων.
- iv. Η μέθοδος Lui k-n-match παρέχει μεγάλη ακρίβεια αποτελεσμάτων λόγω της εμπλοκής του χρήστη στην επιλογή των όμοιων αντικειμένων, γεγονός που αποτελεί παραδοσιακά μεγάλη εγγύηση [A6] ακρίβειας των αποτελεσμάτων (με δεδομένη βέβαια την καταλληλότητα του χρήστη). Η εμπλοκή του χρήστη δεν μπορεί να συγκριθεί σε θέματα σωστής κρίσης με κείνη του υπεύθυνου της εφαρμογής για τη μέθοδο INTESIS, γιατί ο μεν πρώτος απλά επιλέγει τα όμοια αντικείμενα ανάμεσα σε αυτά που βλέπει, ο δε δεύτερος πρέπει να αξιολογήσει και να επιλέξει τις τιμές των παραμέτρων που θα καθορίσουν την αποτελεσματικότητα και την ακρίβεια της μεθόδου. Παρόλα αυτά, η χρονική επιβάρυνση που εισάγει η εμπλοκή του χρήστη στη μέθοδο Lui k-n-match, δεν είναι αξιοκαταφρόνητη εάν συγκριθεί με τους χρόνους απόδοσης της μεθόδου INTESIS σε πολυεπεξεργαστικό περιβάλλον.

Εάν λοιπόν σε κάποια εφαρμογή μας ενδιαφέρει η ανακάλυψη τμηματικής ομοιότητας ανάμεσα σε αντικείμενα 150 διαστάσεων, θα προτιμηθεί η Lui k-n-match ενώ για βάση δεδομένων 50-διαστάσεων αντικειμένων, η οποία μάλιστα ‘τρέχει’ σε πολυεπεξεργαστικό σύστημα, η μέθοδος INTESIS είναι η καταλληλότερη.

Συνοψίζοντας λοιπόν, μπορεί κανείς να καταλήξει στο συμπέρασμα πως οι δύο μέθοδοι αλληλοσυμπληρώνονται κατά κάποιο τρόπο, αφού η μεν μέθοδος Lui k-n-match παρέχει

γενικότητα στο εύρος ανάλογων εφαρμογών καθώς και μεγάλη αξιοπιστία και ακρίβεια ενώ η INTESIS μπορεί να φτάσει σε πολύ καλές χρονικές αποδόσεις (ειδικά σε πολυεπεξεργαστικά συστήματα) λόγω της δεικτοδότησης που υποστηρίζει.

5.3. Επέκταση INTESIS βάσει άλλων προτάσεων

Κατά την περιγραφή της μεθόδου INTESIS, τέθηκε ένα άνω όριο για τις διαστάσεις των δεδομένων (100). Αυτό το όριο απαιτήθηκε λόγω της χρήσης κάποιου αριθμού δέντρων ομοιοτήτων λίγων διαστάσεων (R^* -trees) για την υποστήριξη της μεθόδου. Πιο συγκεκριμένα, όσο το πλήθος των διαστάσεων των δεδομένων αυξάνεται, το όλο σύνολο δεδομένων γίνεται περισσότερο σύνθετο και συμβαίνουν δύο πράγματα. Πρώτον αυξάνεται ο αριθμός των απαιτούμενων δέντρων και δεύτερον αυξάνεται και η πιθανότητα κάποιο/κάποια από τα χαρακτηριστικά των αντικειμένων να αντιπροσωπεύονται από όχι μικρό πλήθος διαστάσεων. Έτσι στην πρώτη περίπτωση εάν ο αριθμός των δέντρων μεγαλώσει πολύ, η απόδοση της μεθόδου θα πέσει εξίσου πολύ καθιστώντας τη μη ικανοποιητική. Στη δεύτερη περίπτωση πάλι θα πέσει η συνολική απόδοση, γιατί για τα δέντρα που θα αντιστοιχούν σε χαρακτηριστικά πολλών διαστάσεων (>10) η απόδοσή τους θα έχει μειωθεί πολύ [A7]. Για αυτούς τους λόγους λοιπόν, έχει οριστεί αυτό το άνω όριο διαστάσεων έτσι ώστε να μη φτάσει ποτέ η μέθοδος σε χαμηλές αποδόσεις.

Εδώ τίθεται το εξής λογικό ερώτημα: Μήπως η μέθοδος INTESIS θα μπορούσε να αναβαθμιστεί με την επιλογή άλλου δέντρου υποστήριξης για τα χαρακτηριστικά πολλών διαστάσεων; Η απάντηση είναι πως όντως μπορεί να γίνει κάτι τέτοιο και μάλιστα με δύο εναλλακτικούς τρόπους: τμηματικά ή συνολικά. Ειδικότερα:

α) *Τμηματική Αντιμετώπιση*: Μπορεί να χρησιμοποιούνται δύο εναλλακτικές μέθοδοι για τη αναζήτηση ομοιοτήτων κάθε χαρακτηριστικού. Για τα χαρακτηριστικά λίγων διαστάσεων η κατάσταση να παραμένει ως έχει (χρήση R^* -tree), ενώ για τα πολυδιάστατα χαρακτηριστικά να χρησιμοποιείται το VA-File.

β) *Συνολική αντιμετώπιση*: Να χρησιμοποιείται η μέθοδος Digenis για την αναζήτηση ομοιότητας για κάθε χαρακτηριστικό (πολυδιάστατο ή όχι), έτσι ώστε να μην εμπλέκονται περισσότερες μέθοδοι.

Η αξιολόγηση των δύο αυτών αντιμετώπισεων με σκοπό την επιλογή της αποδοτικότερης γενικά ή κατά περίπτωση, μπορεί να αποτελέσει θέμα μελλοντικής έρευνας.

Αυτή η δυνατότητα ‘ανοίγματος’ και σε αντικείμενα με κάποια πολυδιάστατα χαρακτηριστικά, δίνει τη δυνατότητα επαναπροσδιορισμού του αρχικού άνω ορίου διαστάσεων της μεθόδου INTESIS. Αυτός ο επαναπροσδιορισμός όμως δεν μπορεί να είναι γενικός αλλά κατά περίπτωση μόνο και συγκεκριμένα για τις πολλές διατάσεις (πάνω από 100) όταν αυτές ομαδοποιούνται σε πολυδιάστατα χαρακτηριστικά, γιατί αυτή μόνο την περίπτωση αντιμετωπίσαμε με τις παραπάνω προτάσεις. Για παράδειγμα, σε ένα χώρο 150 διαστάσεων οι οποίες ομαδοποιούνται σε 10 χαρακτηριστικά από τα οποία τα 6 αποτελούνται από 15-20 διαστάσεις, μπορεί να εφαρμοστεί η μέθοδος INTESIS βάση των παραπάνω νέων τροποποιήσεων. Εάν όμως αυτές οι 150 διαστάσεις αφορούσαν 20 ολιγοδιάστατα επιμέρους χαρακτηριστικά, τότε καμία από τις εκδοχές του INTESIS δε θα μπορούσε να προταθεί για ικανοποιητική υποστήριξη αυτού του χώρου.

5.4. Γιατί παρόλο που τα δέντρα Digenis και Chiron μοιάζουν, οι αλγόριθμοι εντοπισμού θέσης σ' αυτά διαφέρουν τόσο;

Τα δέντρα Digenis και Chiron φαινομενικά μοιάζουν αρκετά - όπως εν μέρει φαίνεται στα σχήματα 5.1. και 5.2. - γι' αυτό και οι αντίστοιχοι αλγόριθμοι αναζήτησης σε αυτά έχουν κοινά σημεία, αλλά διαφέρουν καθοριστικά στο σημαντικότερο. Διαφέρουν στο σημείο όπου κανένας κόμβος του επιπέδου δεν έχει απόσταση-διαφορά από το q μικρότερη από την απαιτούμενη από το δέντρο για να γίνει το q μέλος ή παιδί του κόμβου. Σε αυτό το σημείο, για το μεν δέντρο Digenis εντοπίζεται ένα υπο-δέντρο στο οποίο 'κατεβαίνει' ο αλγόριθμος ψάχνοντας την σωστή θέση του q , για το δε Chiron η διαδικασία συνεχίζεται σε ολόκληρο το επόμενο επίπεδο με τον ίδιο ακριβώς τρόπο που λειτουργήσε στο τρέχον. Η αιτία αυτής της σημαντικής διαφοροποίησης των αλγορίθμων γίνεται φανερή από τον παρακάτω συλλογισμό:

Σχήμα 5.1. : Δέντρο Digenis, με $ch=3$.

Σχήμα 5.2. : Δέντρο Chiron.

Τα δέντρα Digenis και Chiron παρουσιάζουν κάποιες σημαντικές ομοιότητες, αλλά η πληροφορία την οποία οργανώνουν και κωδικοποιούν είναι πολύ διαφορετική. Πιο συγκεκριμένα, οι αποστάσεις που αφορούν το δέντρο Digenis είναι οι Ευκλείδειες και αντιπροσωπεύουν γειτνίαση των αντίστοιχων αντικειμένων στον πολυδιάστατο χώρο τους (που μεταφράζεται σε ομοιότητα μεταξύ τους). Αυτό δηλαδή που ορίζεται με κάθε οικογένεια, είναι ένα είδος ακτίνας χώρου στον οποίο εκείνη εκτείνεται (βρίσκονται τα μέλη της). Έτσι, όταν ένα αντικείμενο κείται κοντά σε κάποια οικογένεια αντικειμένων του δέντρου Digenis, δεν είναι πιθανό να είναι τελικά όμοιο ή ίδιο με μέλος κάποιας άλλης

γειτονικής οικογένειας. Συνεπώς όταν μετά από τον υπολογισμό των αποστάσεων ανάμεσα σε ένα δοθέν αντικείμενο q και κάθε κόμβο-αντιπρόσωπο σε ένα συγκεκριμένο επίπεδο του δέντρου, απομονωθεί η μικρότερη από αυτές τις αποστάσεις, μπορούμε να ξέρουμε πως ο κόμβος στον οποίο αντιστοιχεί, ορίζει το υπο-δέντρο όπου θα βρίσκεται η κατάλληλη για το δοθέν αντικείμενο θέση. Αντίθετα, εφόσον στο δέντρο Chiron οι αποστάσεις-διαφορές ανάμεσα στους κόμβους ορίζουν πλήθος διαφορετικών ψηφίων ανάμεσα στις αντικείμενα-κόμβους, δεν μπορεί να εννοηθεί κάτι τέτοιο για το σχετικό υπο-δέντρο. Με άλλα λόγια, εάν η διαφορά πορείας ανάμεσα στο δοθέν αντικείμενο q και κάποιο κόμβο-αντικείμενο είναι η ελάχιστη του επιπέδου, αυτό δε σημαίνει πως κάπου στο υπο-δέντρο αυτού του κόμβου βρίσκεται η σωστή θέση του q . Ας δούμε το παρακάτω παράδειγμα: Έστω ότι στο σχήμα 5.2. η διαφορά πορείας του q από το αντικείμενο G4 είναι 7 και από το G12 είναι 8. Εάν οι 2 διαφορές του G4 από το G1, αντιστοιχούν στα ψηφία όπου το G4 ήταν ίδιο με το q , τότε το G1 θα διαφέρει από το q και σε αυτά τα ψηφία (βλέπε απόδειξη θεωρήματος Chiron, στο κεφάλαιο 4). Άρα η διαφορά του G1 με το q θα ανέλθει σε $9(7+2)$. Αντίθετα, μπορεί τα 2 ψηφία του G12 που διαφέρουν από το G3, να είναι από εκείνα που διέφεραν με το q και τα αντίστοιχα του q να είναι ίδια με κείνα του G3, οπότε η διαφορά του q από το G3 να είναι 6 ($8-2$). Σ' αυτή την περίπτωση, εάν η ανοχή του δέντρου είναι 6, το q θα εστιαστεί στην περιοχή παιδιών του G3, οπότε καμία σχέση δε θα έχει με το υπο-δέντρο του G4 με το οποίο είχε μικρότερη διαφορά απ'ότι με το G12. Συνεπώς, δεν είναι δυνατόν να ακολουθηθεί η ίδια λογική στον αλγόριθμο του Chiron σ' αυτό το σημείο, γι' αυτό ο αλγόριθμος συνεχίζει σε ολόκληρο το επόμενο επίπεδο και δεν εστιάζει σε κάποιο υπο-δέντρο του.

5.5. Αναφορές του Κεφαλαίου

- [A1] Laird AR, Lancaster JL, Fox PT. (2005). BrainMap: The social evolution of a functional neuroimaging database. *Neuroinformatics* 3, 65-78.
- [A2] Fox PT, Lancaster JL. Mapping context and content: The BrainMap model. *Nature Rev Neurosci* 3, 319-321, 2002.
- [A3] Batrancourt B., Levy R., Lehericy S., Hasboun D., Samson Y., Lavallee I., Lamure M., Dubois B.: Base de données anatomo-fonctionnelle sur le cerveau. An anatomofunctional brain database, *Comptes Rendus Biologies*, Volume 325, Number 4, April 2002 , pp. 439-455(17).
- [A4] Bjaalie JG, Grillner S, Usui S.: *Neuroinformatics: databases, tools, and computational modeling for studying the nervous system*. *Neural Netw.* 2008; 21:1045-6.
- [A5] Chen T., Schneider M.: Data structures and intersection algorithms for 3D spatial data types, *ACM SIGSPATIAL* 2009.
- [A6] Aggarwal, C.C.: Towards meaningful high-dimensional nearest neighbor search by human-computer interaction. In *ICDE* 2002.
- [A7] Weber, R., Schek, H.J, Blott, S.: A Quantitive Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *Very Large Data Bases (VLDB)* 1998.