Εθνικό Μετσόβιο Πολυτεχνείο

Σχολη Ηλεκτρολογων Μηχανικων
Και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης
και Υπολογιστων

Ευφυή συστήματα βασισμένα στη γνώση σε αβέβαια περιβάλλοντα

Intelligent knowledge-based systems in uncertain environments

Διδακτορική Διατριβή

του

ΓΟΥΑΛΛΕΣ Θ. ΕΜΜΑΝΟΥΗΛ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2001)

Αθήνα, Οκτώβριος 2005

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης
και Υπολογιστων

Ευφυή συστήματα βασισμένα στη γνώση σε αβέβαια περιβάλλοντα

Intelligent knowledge-based systems in uncertain environments

Διδακτορική Διατριβή

του

ΓΟΥΑΛΛΕΣ Θ. ΕΜΜΑΝΟΥΗΛ

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π. (2001)

Συμβουλευτική Επιτροπή: Στέφανος Κόλλιας
Ανδρέας–Γεώργιος Σταφυλλοπάτης
Παναγιώτης Τσανάκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 25^{η} Οκτωβρίου 2005.

...	...	...
Σ. Κόλλιας	Α.–Γ. Σταφυλοπάτης	Π. Τσανάκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
...	...	...
Τ. Σελλής	Π. Μαραγκός	Γ. Καραγιάννης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

...
Α. Ντελόπουλος
Επ. Καθηγητής Α.Π.Θ.

Αθήνα, Οκτώβριος 2005

...

ΓΟΥΑΛΛΕΣ Θ. ΕΜΜΑΝΟΥΗΛ
Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Γουάλλες Θ. Εμμανουήλ, 2005.
Με επιφύλαξη παντός δικαιώματος. All rights reserved

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πήγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Η παρούσα διδακτορική διατριβή αποτελεί υποέργο του προγράμματος: "Ηράκλειτος: Υποτροφίες έρευνας με προτεραιότητα στην βασική έρευνα"

Το πρόγραμμα "ΗΡΑΚΛΕΙΤΟΣ" συγχρηματοδοτείται από το Ευρωπαϊκό Κοινοτικό Ταμείο (75%) και από Εθνικούς Πόρους (25%).

The Project "HRAKLEITOS" is co-funded by the European Social Fund (75%) and National Resources (25%).

στη μητέρα μου

Περιεχόμενα

Περιεχόμενα **iii**

Σχήματα **vii**

Πίνακες **ix**

Πρόλογος **xi**

Περίληψη **xiii**

Abstract **xv**

Κατάλογος Απόδοσης Όρων **xvii**

1 Εισαγωγή **1**
1.1 Δομή της διατριβής 1
1.2 Ερευνητικές συνεισφορές 2

2 Σχεσιακή αναπαράσταση γνώσης και πλαίσιο γνώσης **5**
2.1 Εισαγωγή 5
2.2 Ασαφή σύνολα και ασαφείς σχέσεις 5
2.2.1 Ασαφή σύνολα 5
2.2.2 Ασαφείς σχέσεις 7
2.3 Οντολογίες 8
2.3.1 WordNet 9
2.4 Ασαφής σχεσιακή αναπαράσταση γνώσης 9
2.5 Πλαίσιο γνώσης 12
2.6 Πειραματικά αποτελέσματα 13

3 Αραιές σχέσεις και μεταβατικότητα **17**
3.1 Εισαγωγή 17
3.2 Υποθέσεις για την αραιότητα 18
3.3 Μοντέλο αραιής αναπαράστασης 18
3.4 Συμβατική μεθοδολογία μεταβατικού κλεισίματος 20
3.4.1 Συγκριτική μελέτη 23
3.5 Αλγόριθμος σταδιακής ενημέρωσης ITU 23
3.5.1 Αριθμητικό παράδειγμα 28
3.5.2 Συγκριτική μελέτη 30
3.6 Πειραματικά αποτελέσματα 30
3.6.1 Πειραματικά δεδομένα 30

3.6.2 Πλήρης αναπαράσταση 31
3.6.3 Αραιή σχέση και αραιή αναπαράσταση 33
3.6.4 Πυκνή σχέση και αραιή αναπαράσταση 34

4 sup $-t$ μεταβατικό κλείσιμο ασαφών σχέσεων 37
4.1 Εισαγωγή 37
4.2 Αλγόριθμος μεταβατικού κλεισίματος ITC 37
4.3 Αριθμητικό παράδειγμα 38
4.4 Συγκριτική μελέτη 39
4.5 Πειραματικά αποτελέσματα 43

5 Συστήματα ανάκτησης πληροφορίας και επέκταση ερωτήματος 47
5.1 Εισαγωγή 47
5.2 Συστήματα ανάκτησης πληροφορίας 47
5.2.1 Δομή συστημάτων ανάκτησης πληροφορίας 47
5.2.2 Μοντέλα συστημάτων αναζήτησης 49
5.2.3 Προβλήματα των συστημάτων αναζήτησης πληροφορίας 50
5.2.4 Επέκταση ερωτήματος 51
5.3 Ευφυής σημασιολογική επέκταση ερωτήματος 52
5.4 Πειραματικά αποτελέσματα 54

6 Ανάλυση και θεματική κατηγοριοποίηση εγγράφων 55
6.1 Εισαγωγή 55
6.2 Ασαφής ιεραρχική ομαδοποίηση οντοτήτων 57
6.3 Εξαγωγή θεματικών κατηγοριών 59
6.4 Πειραματικά αποτελέσματα 60
6.4.1 Συνθετικά δεδομένα 62
6.4.2 Εφαρμογή σε πραγματικά έγγραφα 63

7 Αποτίμηση ασαφών κανόνων 65
7.1 Εισαγωγή 65
7.2 Ασαφείς κανόνες και προαιρετικοί όροι 66
7.3 Δυνατοτική αποτίμηση 68
7.3.1 Πιθανοτική αποτίμηση απαραίτητων όρων 69
7.3.2 Πιθανοτική αποτίμηση προαιρετικών όρων 70
7.3.3 Ο δυνατοτικός χαρακτήρας 71
7.4 Πειραματικά αποτελέσματα 72

8 Ιεραρχική ομαδοποίηση 77
8.1 Εισαγωγή 77
8.2 Γενική δομή αλγορίθμου 78
8.3 Ιεραρχική ομαδοποίηση σε υψηλές διαστάσεις 79
8.4 Βελτίωση ομαδοποίησης και αξιολόγηση επίδοσης μέσω αναταξινόμησης 82
8.5 Πειραματικά αποτελέσματα 84
8.5.1 Συνθετικά δεδομένα 84
8.5.2 Βάση δεδομένων ίριδας 85
8.5.3 Βάση δεδομένων καρκίνου του μαστού 85
8.5.4 Βάση δεδομένων Ιονόσφαιρας 86

9 Αρχικοποίηση νευρωνικών δικτύων **89**
9.1 Εισαγωγή 89
9.2 Δομή δικτύου 89
9.3 Αρχικοποίηση με βάση την ομαδοποίηση 92
9.4 Πειραματικά αποτελέσματα 92

10 Εξαγωγή ασαφών κανόνων από νευρωνικά δίκτυα **95**
10.1 Εισαγωγή 95
10.2 Εξαγωγή κανόνων 96
10.3 Γενετική απλοποίηση κανόνων 97
10.3.1 Δομή αλγορίθμου 99
10.3.2 Υλοποίηση γενετικών τελεστών 99
10.4 Πειραματικά αποτελέσματα 100

11 Συμπεράσματα – Επεκτάσεις **103**

12 Βιογραφικό Σημείωμα **105**

13 Κατάλογος δημοσιεύσεων **109**
13.1 Περιοδικά 109
13.1.1 Δημοσιευμένα 109
13.1.2 Υποβεβλημένα προς κρίση 110
13.2 Editorials 110
13.3 Βιβλία 110
13.4 Κεφάλαια σε βιβλία 110
13.5 Συνέδρια 111
13.6 Τεχνικές Αναφορές 114

Βιβλιογραφία **115**

Σχήματα

2.1 Παράδειγμα οντολογικής ταξινομίας. 8
2.2 Μια απλή σχέση ταξινομίας. 15

3.1 Το δέντρο που διατάσσεται με βάση το δείκτη γραμμής i. 20
3.2 Το δέντρο που διατάσσεται με βάση το δείκτη στήλης j. 20
3.3 Γραφική αναπαράσταση της σταδιακής ενημέρωσης της σχέσης. 24
3.4 Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και πλήρη αναπαράσταση 32
3.5 Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n και αραιή αναπαράσταση. 34
3.6 Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και αραιή αναπαράσταση. 34

4.1 The sample fuzzy relation 38
4.2 Προσθήκη στοιχείου (#1,#2,0.95) 39
4.3 Προσθήκη στοιχείου (#2,#3,0.95) 39
4.4 Προσθήκη στοιχείου (#2,#8,0.95) 39
4.5 Προσθήκη στοιχείου (#3,#4,0.95) 40
4.6 Προσθήκη στοιχείου (#4,#1,0.95) 40
4.7 Προσθήκη στοιχείου (#4,#9,0.95) 40
4.8 Προσθήκη στοιχείου (#5,#3,0.95) 40
4.9 Προσθήκη στοιχείου (#6,#7,0.95) 41
4.10 Προσθήκη στοιχείου (#7,#8,0.95) 41
4.11 Προσθήκη στοιχείου (#8,#7,0.95) 41
4.12 Προσθήκη στοιχείου (#9,#6,0.95) 42
4.13 Χρόνοι εκτέλεσης για τον ITC με είσοδο R_n και R_n^t. 43
4.14 Χρόνοι εκτέλεσης σύνθεσης με είσοδο R_{90000}. 44
4.15 Χρόνοι εκτέλεσης του ITC με είσοδο R_n και R_n^t. 46

5.1 Η γενική μορφή ενός συστήματος αναζήτησης πληροφορίας. 48

6.1 Ανίχνευση θεματικών κατηγοριών σε πραγματικά έγγραφα 64

7.1 Η λεκτική μεταβλητή high_temp. 68
7.2 Το καρέ 39308. 72
7.3 Το καρέ 39308 μετά από επεξεργασία. 73

8.1 Το συνθετικό σύνολο δεδομένων. 84

9.1 Το νευρωνικό δίκτυο. 90

9.2 Το τετραγωνικό σφάλμα σαν συνάρτηση των εποχών εκπαίδευσης 93

10.1 Η αρχιτεκτονική του δικτύου μαζί με τις παραμέτρους του. 98

Πίνακες

2.1 Οι ασαφείς σημασιολογικές σχέσεις 10

3.1 Παράδειγμα αραιής σχέσης 19
3.2 Σύνοψη υπολογιστικής πολυπλοκότητας για συνθεση και μεταβατικό κλείσιμο με την κλασική μέθοδο 23
3.3 Αρχική μεταβατική σχέση R_{input} 29
3.4 Αποτέλεσμα R_{output} του ITU μετά την προσθήκη του στοιχείου (#9,#6,0.95) 29
3.5 Σύνοψη υπολογιστικής πολυπλοκότητας αλγορίθμων ανάκτησης μεταβατικότητας 29
3.6 Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και πλήρη αναπαράσταση 32
3.7 Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n και αραιή αναπαράσταση. 33
3.8 Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και αραιή αναπαράσταση. 35

4.1 Σύνοψη υπολογιστικής πολυπλοκότητας αλγορίθμων μεταβατικού κλεισίματος 42
4.2 Χρόνοι εκτέλεσης για τον ITC με είσοδο R_n και R_n^t. 44
4.3 Χρόνοι εκτέλεσης σύνθεσης με είσοδο R_{90000}. 44

5.1 Επέκταση όρου engine 53
5.2 Επέκταση όρου airplane 53
5.3 Επέκταση όρου propeller 53

6.1 Ονόματα σημασιολογικών οντοτήτων 61
6.2 Η σχέση T της θεματικής κατηγοριοποίησης 61
6.3 Η ασαφής σχέση δεικτοδότησης 61
6.4 Το αποτέλεσμα της μεθόδου 62

7.1 Οι τιμές των λεκτικών μεταβλητών στόματος και σαγονιού για το καρέ 39308. 73
7.2 Οι τιμές των λεκτικών μεταβλητών ματιών και φρυδιών για το καρέ 39308. 74
7.3 Συνολική έξοδος συστήματος ασαφών κανόνων. 74

8.1 Οι παράμετροι για την παραγωγή του συνόλου συνθετικών δεδομένων 84
8.2 Οι ομάδες που παράγονται, για το συνθετικό σύνολο δεδομένων. 85
8.3 Ποσοστά ταξινόμησης για δεδομένα ίριδας ($\kappa = \lambda = 2$) 85

8.4 Ποσοστό ταξινόμησης για δεδομένα Wisconsin ($\kappa = \lambda = 2$) 86
8.5 Ποσοστά ταξινόμησης για δεδομένα Wisconsin ($\kappa = \lambda = 5$) 86
8.6 Ποσοστά ταξινόμησης για δεδομένα Ιονόσφαιρας ($\kappa = \lambda = 2$) 86

9.1 Βαθμοί ταξινόμησης και πλήθη κρυμμένων κόμβων: 92

10.1 Ο πίνακας παραμέτρων M 101
10.2 Ο πίνακας παραμέτρων Σ 101
10.3 Ο πίνακας παραμέτρων W 101
10.4 Ο πίνακας παραμέτρων A 101

ΠΡΟΛΟΓΟΣ

Φτάνοντας στο τέλος μιας πορείας μερικών ετών για να υποστηρίξει κανείς μια διδακτορική διατριβή γνωρίζει πως το ζητούμενο είναι να αποδείξει όχι απλά πως η εργασία που έχει συντελεστεί είναι αξιόλογη και επαρκής, αλλά και ότι αυτή η εργασία ανήκει στον ίδιο αποκλειστικά. Και όσο προφανές δείχνει αυτό κατά τη διάρκεια της διαδρομής, άλλο τόσο οξύμωρο φαντάζει όταν έχεις πλέον φτάσει στο τέρμα. Γιατί πολύ απλά, φτάνοντας στο τέλος ο απολογισμός δεν μπορεί παρά να δείξει το μέγεθος της συνεισφοράς που δέχτηκες σε τόσο πολλά επίπεδα και από τόσο πολλούς ανθρώπους για να φτάσεις ως εκεί.

Πριν από οποιονδήποτε άλλο, θα ήθελα να θυμηθώ τη βοήθεια, τη συμπαράσταση και την εμπιστοσύνη που έχει δείξει στο πρόσωπό μου ο επιβλέπων την εργασία Καθ. Στέφανος Κόλλιας. Από την πρώτη μας συζήτηση σχετικά με την πιθανή μου ένταξη στο Εργαστήριο Εικόνας, Βίντεο, και Συστημάτων Πολυμέσων και σε κάθε μας αλληλεπίδραση ως σήμερα υπήρξε πάντα όχι απλά δίκαιος, αλλά πολύ περισσότερο έντιμος, δείχνοτας πάντα ειλικρινές ενδιαφέρον και προσφέροντας άδολες, σοφές και φιλαλληλες συμβουλές. Εκτιμώ ιδιαίτερα την ελευθερία που μου προσέφερε σε αυτή μου την πορεία γιατί είναι αυτή ακριβώς η ελευθερία που μου επιτρέπει να γνωρίζω πώς το αντικείμενο που πραγματεύομαι σε αυτή την εργασία είναι πραγματικά και αυτό που με ενδιαφέρει.

Βέβαια, δεν θα μπορούσα να μην αναφερθώ στα μέλη του Εργαστηρίου Εικόνας, Βίντεο, και Συστημάτων Πολυμέσων, νυν και τέως, που υπήρξαν για εμένα όλα αυτά τα χρόνια ένας ιδιότυπος μικρόκοσμος, ένα μικρό χωριό από φίλους μέσα στο κέντρο του Πολυτεχνείου. Έχω συνεργαστεί με τις περισσότερες και τους περισσότερους από αυτούς, έχω κερδίσει από τις εμπειρίες τους και έχω μάθει από τις συμβουλές τους. Αυτή την ώρα, όμως, περισσότερο θυμάμαι τη στήριξη που έλαβα από όλους σαν ομάδα, αλλά και από καθέναν χωριστά, σε κάθε δύσκολη στιμή. Δεν αναφέρομαι σε καθεναν ονομαστικά όχι γιατί φοβάμαι μήπως ξεχάσω κάποιον (γιατί πραγματικά δεν πιστεύω πως αυτό είναι δυνατό), αλλά γιατί δεν πιστεύω πως μπορώ να βάλω τα ονόματα σε μια σειρά, δίνοντάς τους έστω και έμμεσα σειρά στην εκτίμησή μου.

Τέλος, δεν ξεχνώ πως όσο και αν εγώ χανόμουν στη μελέτη, δεν χάνονταν από τη ζωή μου οι πραγματικοί φίλοι και η οικογένειά μου. Αναγνωρίζω πως από εκεί πρώτα και περισσότερο από ό,τι από οπουδήποτε αλλού αντλούσα τη σιγουριά και τη δύναμη να συνεχίζω. Τους ευχαριστώ όλους για αυτό, και κυρίως τη μητέρα μου, στην οποία και αφιερώνω αυτή τη διατριβή.

Γουάλλες Εμμανουήλ
Αθήνα, Οκτώβριος 2005

ΠΕΡΙΛΗΨΗ

Η αβεβαιότητα έχει αποκτήσει σταδιακά την αποδοχή και το ρόλο της στην επιστημονική έρευνα και την επιστημονική θεώρηση του κόσμου. Όσον αφορά στα ευφυή συστήματα που βασίζονται στη γνώση, σε όποιο επίπεδο και αν εξετάσουμε τη λειτουργία τους η αβεβαιότητα είναι παρούσα και ο ρόλος της καθοριστικός. Έτσι προτείναμε μια σειρά από λύσεις, η οποίες με τη σειρά τους ανοίγουν μια νέα σειρά από δρόμους.

Στο πρώτο τμήμα της διατριβής, που είναι και το πιο εκτενές, η έμφαση είναι στο σημασιολογικό επίπεδο. Σε αυτό το επίπεδο τα βασικά προβλήματα που πρέπει να αντιμετωπίσει κανείς είναι η μοντελοποίηση των εννοιών του πραγματικού κόσμου, καθώς και η πρακτική αξιοποίηση αυτής της γνώσης δεδομενου του μεγέθους της. Προς αυτή την κατεύθυνση, το κεφάλαιο 2 προτείνει τη χρήση ασαφών σχέσεων για την αναπαράσταση της γνώσης και εξηγεί πώς αυτή η γνώση μπορεί να χρησιμοποιηθεί για την αυτόματη εκτίμηση του πλαισίου γνώσης. Τα κεφάλαια 3 και 4 εστιάζουν στο μέγεθος της γνώσης και προτείνουν υπολογιστικά μοντέλα για τον αποδοτικό χειρισμό της. Τα κεφάλαια 5 και 6 εστιάζουν στην ευφυή αξιοποίηση αυτής της γνώσης από συστήματα ανάκτηση πληροφορίας.

Στο δεύτερο τμήμα της διατριβής περνάμε σε ένα επίπεδο ανάμεσα στις έννοιες και τα αριθμητικά δεδομένα. Έτσι, το κεφάλαιο 7 εξηγεί πώς λεκτική γνώση υψηλού επιπέδου μπορεί να χρησιμοποιηθεί στην πράξη για το χειρισμό αβέβαιων αριθμητικών δεδομένων χαμηλού επιπέδου. Έμφαση δίνεται τόσο στην αβεβαιότητα που χαρακτηρίζει τα δεδομένα χαμηλού επιπέδου, όσο και στην ευελιξία που χρειάζεται ώστε τα δεδομένα υψηλού επιπέδου να επιτρέπουν μια επαρκή αναπαράσταση του πραγματικού κόσμου.

Στο τρίτο και τελευταίο τμήμα της διατριβής εργαζόμαστε αποκλειστικά με αριθμητικά δεδομένα χαμηλού επιπέδου. Τα κεφάλαια 8 και 9 πραγματεύονται την αυτόματη επεξεργασία δεδομένων χαμηλού επιπέδου με σκοπό τη δημιουργία νευρωνικών μοντέλων ικανών να αποτυπώσουν τη δομή των δεδομένων, ενώ το κεφάλαιο 10 προχωρά στην επεξεργασία αυτών των μοντέλων με τελικό στόχο την αυτόματη εξαγωγή γνώσης υψηλότερου επιπέδου από τα διαθέσιμα αριθμητικά δεδομένα.

Το κεφάλαιο 11 συνοψίζει τα συμπεράσματα της διατριβής και αναφέρεται σε πιθανές μελλοντικές ερευνητικές κατευθύνσεις που πηγάζουν από την παρούσα εργασία.

ABSTRACT

Uncertainty has gradually attained acceptance and a very distinct role in scientific thought as well as in the scientific view of the world. As far as intelligent knowledge based systems are concenred, uncertainty is present at all levels of their operation and its role is determinant of their effectiveness. In this thesis we propose a series of solutions to uncertainty related problems. In their turn, these solutions provide for further thought and progress in a series of directions.

In the first part of the thesis, which is also the lengthiest, the emphasis is on the semantics. In this framework, the important problems to consider are those of modelling real world concepts thus constructing a formal knowledge base and of exploiting the information contained in this knowledge base in practical applications, given its size. In this direction, chapter 2 proposes the utilization of fuzzy relations for the representation of knowledge and explains how this knowledge can be used in order to automatically extract the context. Chapters 3 and 4 focus on the size of this knowledge and provide computational models for its efficient handling. Chapters 5 and 6 deal with the intelligent utilization of such knowledge in the framework if information retrieval.

In the second part of the thesis we move on to a level between concepts and numeric data. Thus, chapter 7 explains how we can use high level linguistic information in order to handle uncertain low level numerical data. Focus is both on the uncertainty within the low level data and on the flexibility required in order for the high level information to provide for an adequate description of the real world.

In the third and last part of the thesis we work solely with numerical data. Chapters 8 and 9 deal with the automated analysis of data for the generation of neural models that are able to map the structure of the data, while chapter 10 moves on to the processing of these models in order to automatically extract higher level information from the available numerical data.

Chapter 11 summarizes conclusions drawn from this thesis and refers to directions of possible further work that come out of this work.

Κατάλογος Απόδοσης Όρων

belief	:	πίστη
browsing	:	πλοήγηση
city block	:	δομική απόσταση
classification	:	ταξινόμηση
classification rate	:	βαθμός (σωστής) ταξινόμησης
clustering	:	ομαδοποίηση
centroid	:	εικονικό κέντρο ομάδας
competitive learning	:	ανταγωνιστική μάθηση
context	:	πλαίσιο, πλαίσιο γνώσης
crossover	:	γενετικός συνδυασμός
dimensionality curse	:	κατάρα των υψηλών διαστάσεων
dual triple	:	δυϊκή τριάδα
expert user	:	έμπειρος χρήστης
feature	:	μετρήσιμο χαρακτηριστικό
fitness function	:	συνάρτηση καταλληλότητας
gradient descent	:	μέθοδος καθόδου κλίσης βαθμίδας
index	:	ευρετήριο
indexing	:	δεικτοδότηση
information retrieval system (IRS)	:	σύστημα αναζήτησης/ανάκτησης πληροφορίας
metadata	:	μεταδεδομένα
modifier / linguistic hedge	:	ασαφής τροποποιητής
NLP	:	επεξεργασία φυσικής γλώσσας
partitioning clustering	:	διαμεριστική μέθοδος ομαδοποίησης
pattern	:	πρότυπο
plausibility	:	εφικτότητα
possibilistic	:	δυνατοτικός
precision	:	ακρίβεια
quantization	:	κβαντοποίηση / κβαντισμός
query by example (QbE)	:	ερώτημα μέσω παραδείγματος

recall : ανάκληση

radial basis function networks : νευρωνικά δίκτυα συναρτήσεων ακτινικής βάσης

relevance feedback : συσχετιστική ανάδραση

scalar cardinality : βαθμωτός πληθικός αριθμός, βαθμωτή πληθικότητα

supervised : εποπτευμένος / με επίβλεψη

thematic categorization : θεματική κατηγοριοποίηση

thesaurus : λεξικό συνωνύμων / θησαυρός

training data : δεδομένα εκπαίδευσης

unsupervised : ανεπίβλεπτος / χωρίς επίβλεψη

□

Κεφάλαιο 1

Εισαγωγή

Ανάμεσα στις πολλές κατακτήσεις της επιστημονικής κοινότητας κατά τη διάρκεια του περασμένου αιώνα, από τις πιο ιδιαίτερες είναι η αποδοχή της αδυναμίας μας να μετρήσουμε ή να περιγράψουμε με ακρίβεια και βεβαιότητα τον πραγματικό κόσμο. Αυτό δεν αναφέρεται μόνο στην αδυναμία μας να προσδιορίσουμε αντικειμενικά φυσικά μεγέθη όπως είναι η θέση και η ταχύτητα ενός σωματιδίου, αλλά πολύ περισσότερο σε υποκειμενικά μεγέθη όπως είναι οι αφηρημένες λεκτικές έννοιες.

Έτσι η αβεβαιότητα έχει αποκτήσει σταδιακά την αποδοχή και το ρόλο της στην επιστημονική έρευνα και την επιστημονική θεώρηση του κόσμου. Τα ασαφή μαθηματικά, καθώς είναι για πολλούς το καταλληλότερο μαθηματικό εργαλείο για την ποσολόγηση και το χειρισμό της αβεβαιότητας, κερδίζουν συνεχώς τόσο σε συχνότητα όσο και σε εύρος χρήσης.

Από την άλλη πλευρά, ο νέος αιώνας μας φέρνει μια εποχή στην οποία ο ρόλος του υπολογιστή αναθεωρείται. Από απλό εργαλείο εκτέλεσης αυστηρών εντολών ο υπολογιστής καλείται να να αποκτήσει "κατανόηση" σε διάφορα επίπεδα, ώστε να είναι σε θέση να βοηθήσει ή και να υποκαταστήσει τον άνθρωπο σε περισσότερες εργασίες. Προς αυτή την κατεύθυνση παρατηρούμε την ανάπτυξη μοντέλων για την αποθήκευση της ανθρώπινης γνώσης σε ψηφιακή μορφή και συστημάτων που χρησιμοποιούν την αποθηκευμένη γνώση για να μιμηθούν την ανθρώπινη κριτική σκέψη και να δώσουν λύση σε πρακτικά προβλήματα.

Είναι πλέον σαφές πώς και αυτή η προσπάθεια, καθώς βασίζεται σε γνώση γύρω από τον αβέβαιο πραγματικό κόσμο, θα πρέπει να δείνει έμφαση στην ανακρίβεια και την αβεβαιότητα στη δημιουργία, αναπαράσταση και χρήση της αποθηκευμένης γνώσης. Αυτό είναι και το αντικείμενο της παρούσας διδακτορικής διατριβής.

1.1 Δομή της διατριβής

Η διατριβή χωρίζεται σε τρία τμήματα, καθένα από τα οποία πραγματεύεται το συνδυασμό γνώσης και αβεβαιότητας σε διαφορετικό επίπεδο.

Στο πρώτο τμήμα της διατριβής, που είναι και το πιο εκτενές, η έμφαση είναι στο σημασιολογικό επίπεδο. Σε αυτό το επίπεδο τα βασικά προβλήματα που πρέπει να αντιμετωπίσει κανείς είναι η μοντελοποίηση των εννοιών του πραγματικού κόσμου, καθώς και η πρακτική αξιοποίηση αυτής της γνώσης δεδομενου του μεγέθους της. Προς αυτή την κατεύθυνση, το κεφάλαιο 2 προτείνει τη χρήση ασαφών σχέσεων για την αναπαράσταση της γνώσης και εξηγεί πώς αυτή η γνώση μπορεί να χρησιμοποιηθεί για την αυτόματη εκτίμηση του πλαισίου γνώσης. Τα κεφάλαια 3 και 4 εστιάζουν στο

μέγεθος της γνώσης και προτείνουν υπολογιστικά μοντέλα για τον αποδοτικό χειρισμό της. Τα κεφάλαια 5 και 6 εστιάζουν στην ευφυή αξιοποίηση αυτής της γνώσης από συστήματα ανάκτηση πληροφορίας.

Στο δεύτερο τμήμα της διατριβής περνάμε σε ένα επίπεδο ανάμεσα στις έννοιες και τα αριθμητικά δεδομένα. Έτσι, το κεφάλαιο 7 εξηγεί πώς λεκτική γνώση υψηλού επιπέδου μπορεί να χρησιμοποιηθεί στην πράξη για το χειρισμό αβέβαιων αριθμητικών δεδομένων χαμηλού επιπέδου. Έμφαση δίνεται τόσο στην αβεβαιότητα που χαρακτηρίζει τα δεδομένα χαμηλού επιπέδου, όσο και στην ευελιξία που χρειάζεται ώστε τα δεδομένα υψηλού επιπέδου να επιτρέπουν μια επαρκή αναπαράσταση του πραγματικού κόσμου.

Στο τρίτο και τελευταίο τμήμα της διατριβής εργαζόμαστε αποκλειστικά με αριθμητικά δεδομένα χαμηλού επιπέδου. Τα κεφάλαια 8 και 9 πραγματεύονται την αυτόματη επεξεργασία δεδομένων χαμηλού επιπέδου με σκοπό τη δημιουργία νευρωνικών μοντέλων ικανών να αποτυπώσουν τη δομή των δεδομένων, ενώ το κεφάλαιο 10 προχωρά στην επεξεργασία αυτών των μοντέλων με τελικό στόχο την αυτόματη εξαγωγή γνώσης υψηλότερου επιπέδου από τα διαθέσιμα αριθμητικά δεδομένα.

Το κεφάλαιο 11 συνοψίζει τα συμπεράσματα της διατριβής και αναφέρεται σε πιθανές μελλοντικές ερευνητικές κατευθύνσεις που πηγάζουν από την παρούσα εργασία.

1.2 Ερευνητικές συνεισφορές

Όπως σε κάθε έγγραφο ερευνητικού χαρακτήρα που πρέπει να κριθεί, έτσι και σε μια διατριβή σημαντικό ερώτημα είναι ο ακριβής διαχωρισμός της πρωτότυπης εργασίας από τις απλές αναφορές σε προϋπάρχουσα θεωρία. Για το σκοπό αυτό επιχειρούμε σε αυτή την παράγραφο να συνοψίσουμε τις βασικές ερευνητικές συνεισφορές της διατριβής.

Αυτές χωρίζονται παρακάτω σε κύριες και δευτερεύουσες συνεισφορές. Ο διαχωρισμός αυτός δεν αναφέρεται τόσο στην επιστημονική αξία τους, αλλά κυρίως στο αν τα συγκεκριμένα αποτελέσματα έχουν επιτευχθεί αποκλειστικά από τον συγγραφέα της διατριβής ή από μια ερευνητική ομάδα μέλος της οποίας ήταν και ο συγγραφέας.

Στις κύριες συνεισφορές συγκαταλέγονται μεταξύ άλλων:

- Ο αλγόριθμος ITU για το μεταβατικό κλείσιμο σχέσεων που είναι αρχικά μεταβατικές αλλά η μεταβατικότητά τους διαταράσσεται τοπικά. Ο αλγόριθμος ITU έχει πολυπλοκότητα κάτω της γραμμικής για την τυπική αραιή σχέση, που ξεπερνά κατά πολύ την προηγούμενη καλύτερη γνωστή μέθοδο.

- Ο αλγόριθμος ITC για το μεταβατικό κλείσιμο σχέσεων. Ο αλγόριθμος ITC ξεπερνά κατά πολύ σε πολυπλοκότητα κάθε γνωστή μέθοδο όταν εφαρμόζεται στην τυπική αραιή σχέση, ενώ, αντίθετα με άλλους υπολογιστικά ικανούς αλγορίθμους μεταβατικού κλεισίματος, δεν περιορίζεται σε συμμετρικές σχέσεις ή σε $\max - \min$ μεταβατικότητα.

- Ο αλγόριθμος DTC για την ανάλυση εγγράφων κειμένου. Δίνεται μια αλγοριθμική λύση για την πρακτική αξιοποίηση ταξινομικής γνώσης με στόχο την ανάλυση περιεχομένου. Με τον τρόπο αυτό επιτυγχάνεται η θεματική κατηγοριοποίηση με θεώρηση του πλαισίου γνώσης και περιορίζονται τα λάθη που συνοδεύουν τις αυστηρά στατιστικές μεθόδους.

- Η μεθοδολογία πιθανοτικής αποτίμησης ασαφών κανόνων. Η προτεινόμενη μεθοδολογία επιτρέπει τη χρήση συστημάτων ασαφών κανόνων ακόμη και στην περίπτωση που οι γλωσσικές μεταβλητές που χρησιμοποιούνται από τους κανόνες δεν μπορούν να αποτιμηθούν με βεβαιότητα.

- Ο εύρωστος αλγόριθμος ιεραρχικής ομαδοποίησης. Αν και ιδαιίτερα χρήσιμοι στην εξόρυξη πληροφορίας λόγω της εκτέλεσής τους χωρίς χειρονακτική αρχικοποίηση, οι ιεραρχικοί αλγόριθμοι έχουν σημαντικά μειονεκτήματα που σχετίζονται με την υψηλή πολυπλοκότητα και τον ατελή διαχωρισμό των ομάδων. Στην προτεινόμενη εργασία βελτιώνεται σημαντικά ο βαθμός διαχωρισμού των ομάδων, ενώ προτείνεται και τρόπος περιορισμού των υπολογιστικών αναγκών τις διαδικασίας.

Στις δευτερεύουσες συνεισφορές συγκαταλέγονται μεταξύ άλλων:

- Η ασαφής σχεσιακή αναπαράσταση γνώσης, σε συνεργασία με τους Δρ. Γ. Ακρίβα και Δρ. Γ. Στάμου. Η προτεινόμενη αναπαράσταση ακολουθεί την ταξινομική μορφή που απαντάται και στις οντολογίες, αλλά με τη χρήση ασαφών βαθμών αποκτά μεγαλύτερη περιγραφική δύναμη.

- Η αραιή αναπαράσταση ασαφών σχέσεων, σε συνεργασία με τον Δρ. Ι. Αβρίθη. Το προτεινόμενο μοντέλο, εκμεταλλευόμενο τα άριστα υπολογιστικά χαρακτηριστικά των δέντρων AVL, συνδυάζει συμπαγή αποθήκευση με ταχύτατη πρόσβαση στα δεδομένα.

- Ο ορισμός του πλαισίου γνώσης, σε συνεργασία με τον Δρ. Γ. Ακρίβα. Ο προτεινόμενος ορισμός προσφέρει άμεση αλγοριθμική λύση στο πρόβλημα της εκτίμησης του πλαισίου γνώσης, γεγονός που τον κάνει να βρίσκει σημαντικές εφαρμογές στο πεδίο της ανάλυσης και αναζήτησης εγγράφων.

- Η εξαγωγή κανόνων από νευρωνικά δίκτυα, σε συνεργασία με τον Δρ. Ν. Τσαπατσούλη. Με τη χρήση γενετικών αλγορίθμων επιτυγχάνεται η απλοποίηση των εξαγόμενων κανόνων χωρίς βλάβη της ποιότητάς τους.

□

Κεφάλαιο 2

Σχεσιακή αναπαράσταση γνώσης και πλαίσιο γνώσης

2.1 Εισαγωγή

Η αβεβαιότητα, η ατέλεια ή η ασάφεια είναι εγγενή στοιχεία στον πραγματικό κόσμο. Συνεπώς, η ασαφής άλγεβρα, καθώς είναι ικανή να αναπαραστήσει και να χειριστεί τέτοιου είδους πληροφορία αποτελεσματικά, είναι ιδιαίτερα χρήσιμο εργαλείο για την αναπαράσταση γνώσης σχετικά με τον πραγματικό κόσμο. Σε αυτό το κεφάλαιο, μετά από μια σύντομη παρουσίαση των βασικών όρων ασαφών συνόλων και ασαφών σχέσεων που θα χρησιμοποιηθούν στη συνέχεια, εξηγούμε πώς οι ασαφείς σχέσεις μπορούν να χρησιμοποιηθούν για την αναπαράσταση γνώσης. Επιπρόσθετα, εξηγούμε πώς η γνώση αυτή μπορεί να χρησιμοποιηθεί για τον ορισμό και την εκτίμηση του πλαισίου γνώσης.

2.2 Ασαφή σύνολα και ασαφείς σχέσεις

Στην ενότητα αυτή γίνεται μία εισαγωγή στις κυριότερες έννοιες της θεωρίας ασαφών συνόλων [143]. Η βασική πηγή για τη θεωρία είναι το [73].

2.2.1 Ασαφή σύνολα

Όπως στην κλασική θεωρία συνόλων, αφετηρία είναι ένα καθολικό σύνολο S. Ένα ασαφές σύνολο A επί του S, ή ισοδύναμα ένα ασαφές υποσύνολο του S είναι μια αντιστοίχηση

$$\mu_A : S \rightarrow [0, 1] \tag{2.1}$$

Η αντιστοίχιση αυτή ονομάζεται και συνάρτηση συμμετοχής. Συχνά το σύμβολο της συνάρτησης συμμετοχής είναι το ίδιο το σύμβολο A. Για $s \in S$, το $A(s)$ συμβολίζει το βαθμό, στον οποίο το s ανήκει στο A, ή ισοδύναμα την τιμή $\mu_A(s)$ στην οποία η συνάρτηση συμμετοχής αντιστοιχεί το s.

Η μέγιστη τιμή της συνάρτησης συμμετοχής για ένα ασαφές σύνολο ονομάζεται ύψος του ασαφούς συνόλου και συμβολίζεται ως

$$h(A) = sup_{s \in S} A(s) \tag{2.2}$$

Ένα ασαφές σύνολο A καλείται κανονικό αν το ύψος του $h(A)$ είναι 1. Άλλες έννοιες της κλασικής θεωρίας συνόλων, που γενικεύονται από τη θεωρία ασαφών συνόλων είναι ο βαθμωτός πληθικός αριθμός $|A|$:

$$|A| = \sum_{s \in S} A(s) \tag{2.3}$$

το υποσύνολο $A \subseteq B$:

$$A \subseteq B \Leftrightarrow A(s) \leq B(s), \forall s \in S \tag{2.4}$$

και οι έννοιες της τομής t

$$t(x, 1) = x \tag{2.5}$$

$$x_1 \leq x_2 \Leftrightarrow t(x, x_1) \leq t(x, x_2) \tag{2.6}$$

$$t(x_1, x_2) = t(x_2, x_1) \tag{2.7}$$

$$t(x_1, t(x_2, x_3)) = t(t(x_1, x_2), x_3)) \tag{2.8}$$

ένωσης u

$$u(x, 0) = x \tag{2.9}$$

$$x_1 \leq x_2 \Leftrightarrow u(x, x_1) \leq u(x, x_2) \tag{2.10}$$

$$u(x_1, x_2) = u(x_2, x_1) \tag{2.11}$$

$$u(x_1, u(x_2, x_3)) = u(u(x_1, x_2), x_3)) \tag{2.12}$$

και του συμπληρώματος c

$$c(0) = 1 \tag{2.13}$$

$$c(1) = 0 \tag{2.14}$$

$$x_1 \leq x_2 \Leftrightarrow c(x_1) \geq c(x_2) \tag{2.15}$$

Αντίθετα με την κλασική θεωρία συνόλων, υπάρχουν πολλές επιλογές για συναρτήσεις τομής, ένωσης και συμπληρώματος. Οι συνηθέστερες (κλασσικές) επιλογές είναι για τομή η

$$t(x_1, x_2) = min(x_1, x_2) \tag{2.16}$$

για ένωση η

$$u(x_1, x_2) = max(x_1, x_2) \tag{2.17}$$

και για συμπλήρωμα η

$$c(x) = 1 - x \tag{2.18}$$

Αν μια νόρμα τομής t (t-νόρμα) είναι συνεχής και ισχύει

$$t(x, x) < x, \ \forall x \in (0, 1) \tag{2.19}$$

τότε η νόρμα λέγεται Αρχιμήδεια.

Αν μια νόρμα τομής t, μια νόρμα ένωσης u (s-νόρμα) και ένα συμπλήρωμα c ικανοποιούν τις σχέσεις

$$c(t(x_1, x_2)) = u(c(x_1), c(x_2)) \tag{2.20}$$

$$c(u(x_1, x_2)) = t(c(x_1), c(x_2)) \tag{2.21}$$

(δηλαδή αν ικανοποιούν το νόμο De Morgan) τότε η τριάδα (t,u,c) ονομάζεται δυική τριάδα. Σημαντικές δυικές τριάδες είναι οι:

$$(min(x_1, x_2), max(x_1, x_2), 1 - x) \tag{2.22}$$

$$(x_1 x_2, x_1 + x_2 - x_1 x_2, 1 - x) \tag{2.23}$$

$$(max(0, x_1 + x_2 - 1), min(1, x_1 + x_2), 1 - x) \tag{2.24}$$

2.2.2 Ασαφείς σχέσεις

Μία ασαφής R σχέση από το (κλασικό) σύνολο A στο (κλασικό) B είναι ένα ασαφές σύνολο επί του $A \times B$. Αντίστοιχα λοιπόν με την περίπτωση των ασαφών συνόλων ορίζονται η τομή και η ένωση:

$$[R_1 \cup R_2](s_2, s_2) = u(R_1(s_1, s_2), R_2(s_1, s_2)) \tag{2.25}$$

$$[R_1 \cap R_2](s_2, s_2) = t(R_1(s_1, s_2), R_2(s_1, s_2)) \tag{2.26}$$

Η αντίστροφη σχέση ορίζεται ως

$$R^{-1}(s_1, s_2) \doteq R(s_2, s_1) \tag{2.27}$$

Η σύνθεση $\sup -t$ ορίζεται ως

$$\left[A \overset{t}{\circ} B\right](s_1, s_2) \doteq \sup_{s \in S}(t(A(s_1, s), B(s, s_2))) \tag{2.28}$$

ενώ για την αντίστροφη της σύνθεσης $(P \circ Q)^{-1}$ ισχύει

$$(P \circ Q)^{-1} = Q^{-1} \circ P^{-1} \tag{2.29}$$

Η ταυτοτική σχέση I

$$I(s_1, s_2) = 1, s_1 = s_2 \tag{2.30}$$

$$I(s_1, s_2) = 0, s_1 \neq s_2 \tag{2.31}$$

είναι το ουδέτερο στοιχείο της σύνθεσης

$$R \circ I = I \circ R = R, \forall R \tag{2.32}$$

Ιδιαίτερα χρήσιμοι ορισμοί είναι αυτοί της ανακλαστικότητας

$$R \supseteq I \tag{2.33}$$

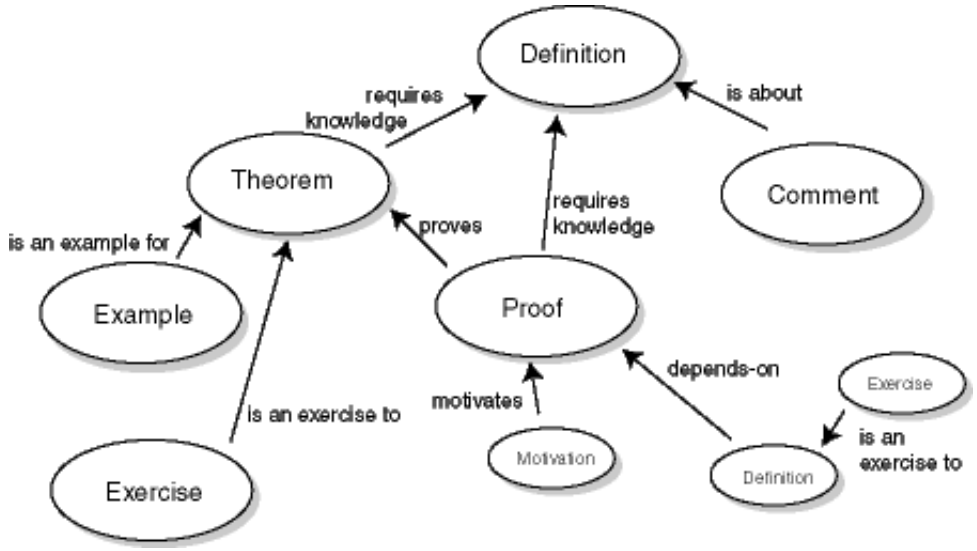
συμμετρικότητας

$$R = R^{-1} \tag{2.34}$$

και $\sup -t$ μεταβατικότητας

$$R \overset{t}{\circ} R \subseteq R \tag{2.35}$$

Επίσης, οι ιδιότητες της ισοδυναμίας (ανακλαστική, συμμετρική και μεταβατική), συμβατότητας (ανακλαστική και συμμετρική), διάταξης (αντισυμμετρική και μεταβατική) και ημιδιάταξης (μεταβατική και ανακλαστική).



Σχήμα 2.1: *Παράδειγμα οντολογικής ταξινομίας.*

2.3 Οντολογίες

Για την ευφυή αλληλεπίδραση με το χρήστη και την εκτέλεση ενεργειών στο σημασιολογικό επίπεδο ένα σύστημα χρειάζεται να χρησιμοποιεί κάποια μορφή συμβολικής γνώσης. Η αναπαράσταση που θα χρησιμοποιηθεί για τη γνώση καθορίζει τόσο την περιγραφική της δύναμη, όσο και τον τύπο των αλγορίθμων που μπορεί να αναπτυχθούν για να την αξιοποιήσουν.

Αναπτύσσοντας συστήματα που βασίζονται στη γνώση επιχειρούμε στην ουσία να πραγματοποιήσουμε με αυτόματο υπολογιστικό τρόπο μια λειτουργία που ο άνθρωπος εκτελεί σχετικά εύκολα μόνος του. Είναι σαφές πως ο λόγος που η ανθρώπινη λύση συνεχίζει να υπερτερεί σε μια πλειάδα περιπτώσεων δεν είναι η υπολογιστική υπεροχή. Σιγά - σιγά αρχίζει να διαφαίνεται πως η υπεροχή δεν είναι ούτε αλγοριθμική. Το μεγάλο πλεονέκτημα του ανθρώπου είναι πως έχει τη δυνατότητα να εκμεταλλεύεται συμβολική γνώση και να τη συνδυάζει με μη συμβολικά δεδομένα.

Έτσι, μεγάλη έμφαση δίνεται πλέον στην αυστηρή κωδικοποίηση της κοινής γνώσης που μοιράζονται οι άνθρωποι, ώστε να μπορεί να αξιοποιηθεί από αυτοματοποιημένες διεργασίες. Αυστηρά δοσμένος, ο όρος "γνώση" αναφέρεται σε αποθηκευμένη πληροφορία ή μοντέλα που χρησιμοποιούνται από έναν άνθρωπο ή μια μηχανή για να εξηγήσουν, να προβλέψουν και να αντιδράσουν κατάλληλα στον εξωτερικό κόσμο [57].

Ο σχετικός ερευνητικός χώρος είναι ο χώρος των οντολογιών. Μια οντολογία είναι αυστηρή περιγραφή ενός πεδίου. Εν γένει αποτελείται από τρία στοιχεία [89]:

$$O = \{S, R, I\} \tag{2.36}$$

όπου S είναι το σύνολο των σημασιολογικών οντοτήτων, R το σύνολο των σχέσεων στο S και I ένα σύνολο κανόνων που χρησιμοποιούνται για την αναπαράσταση πληροφοριών που δεν μπορούν να κωδικοποιηθούν με τη χρήση απλών δυαδικών σχέσεων.

Γενικά, αν και η εξέλιξη στο χώρο της αναπαράστασης απλών δεδομένων, δηλαδή στη δημιουργία των S και R, έχει προχωρήσει αρκετά, δεν έχουν γίνει ακόμη ικανά βήματα προς την αυτόματη εξαγωγή πιο σύνθετων συμπερασμάτων με χρήση του I. Έτσι, αν και έχουν αναπτυχθεί πλούσιες ως προς την πληρότητα της περιγραφής οντολογίες, λείπουν οι αλγόριθμοι εκείνοι που θα εκμεταλλευτούν αυτή την πληροφορία

στην πράξη. Στο κεφάλαιο αυτό χρησιμοποιούμε μια ασαφή οντολογική αναπαράσταση και παρουσιάζουμε μια αλγοριθμική λύση στο πρόβλημα της αξιοποίησης της πληροφορίας με στόχο την εκτίμηση του πλαισίου γνώσης. Σε ακόλουθα κεφάλαια θα εξετάσουμε επίσης πώς αυτή η γνώση μπορεί να χρησιμοποιηθεί για να δώσει λύση και σε άλλα πρακτικά προβλήματα, όπως είναι η επέκταση ερωτήματος και η ανάλυση εγγράφων.

2.3.1 WordNet

Το WordNet είναι ένα on line σύστημα λεξικολογικής αναφοράς για την αγγλική γλώσσα [56]. Αναφέρει ουσιαστικά, ρήματα, επίθετα και επιρρήματα, οργανωμένα σε ομάδες συνωνύμων. Διάφορες δυαδικές σχέσεις ανάμεσα στις ομάδες συνωνύμων περιέχονται στο WordNet.

Αυτή η διάταξη είναι αρκετά όμοια με την αναπαράσταση γνώσης που παρουσιάσαμε. Οδηγείται εύκολα, λοιπόν, κανείς στο συμπέρασμα πως το WordNet μπορεί να αποτελέσει την πηγή για την αυτόματη δημιουργία της γνώσης. Πραγματικά, μια πληθώρα αναφορών στη βιβλιογραφία επιχειρούν να εκμεταλλευτούν την πληροφορία που υπάρχει στο WordNet για πρακτικές εφαρμογές [155].

Οι σχέσεις που περιέχει το WordNet, όμως, είναι λεξικολογικές και όχι σημασιολογικές. Έτσι, δεν περιέχουν ικανή πληροφορία σχετικά με την ανθρώπινη κοινή γνώση, ώστε να συμβάλουν αποτελεσματικά στη δημιουργία ευφυών πληροφοριακών συστημάτων. Έχει χρησιμοποιηθεί εκτενώς ωστόσο στο παρελθόν σαν βάση γνώσης για συστήματα ανάκτησης πληροφορίας, κυρίως διότι περιέχει μεγάλο αριθμό λέξεων και ήταν διαθέσιμο πολύ πριν υπάρξουν οι πρώτες οντολογίες.

2.4 Ασαφής σχεσιακή αναπαράσταση γνώσης

Οι ασαφείς σχέσεις και οι ιδιοτητές τους έχουν ένα σημαντικό ρόλο στη μοντελοποίηση της πληροφορίας σε πληθώρα θεωρητικών και εφαρμοσμένων πεδίων. Με περιγραφική ικανότητα που εκτείνεται από απλή αναπαράσταση πληροφορίας [144] έως αναπαράσταση οντολογικής πληροφορίας [89] και πολυμεσικών δομών [4][110], τα πεδία εφαρμογών είναι πρακτικά απεριόριστα. Ένα πεδίο στο οποίο ο ρόλος τους έχει αναδειχθεί σε ιδιαίτερα σημαντικό είναι αυτό της αναζήτησης πληροφορίας [20][35][134]. Σε αυτό το πλαίσιο, η ασαφής σχεσιακή αναπαράσταση γνώσης μπορεί να συνεισφέρει σε διαδικασίες όπως η επέκταση ερωτήματος [5], η ανάλυση εγγράφων [6][132], η ανάλυση πολυμεσικής πληροφορίας [131], η εξαγωγή και χρήση προφίλ [133] κλπ.

Είναι γνωστό πως τα συστήματα πληροφορίας που βασίζονται σε όρους πάσχουν από την προβληματική αντιστοίχιση ανάμεσα σε όρους και έννοιες. Για να ξεπεράσουμε προβλήματα αυτού του είδους εργαζόμαστε απ' ευθείας με έννοιες αντί για όρους. Αναφερόμαστε σε αυτές τις έννοιες ως σημασιολογικές οντότητες [150]. Το σύνολο των οντοτήτων που είναι γνωστές είναι $S = \{s_1, s_2, \ldots, s_n\}$. Οι ορισμοί αυτών των εννοιών, μαζί με τις λεκτικές τους περιγραφές, δηλαδή τους αντίστοιχους όρους, περιέχονται στη σημασιολογική εγκυκλοπαίδεια. Η εγκυκλοπαίδεια περιέχει επίσης αναφορές στις συσχετίσεις ανάμεσα στις σημασιολογικές οντότητες [152][148].

Πρόσφατα, αρκετή προσοχή έχει δοθεί στο σχεδιασμό και τη δημιουργία τέτοιων σχέσεων, οδηγώντας στη δημιουργία των οντολογιών, όπου το πλαίσιο μπορεί προσδιορίζει την ακριβή έννοια ενός όρου. Γενικά, μια οντολογία περιγράφεται ως:

Πίνακας 2.1: *Οι ασαφείς σημασιολογικές σχέσεις*

Σύμβολο	Σχέση
Sp	Εξειδίκευση
Ct	Πλαίσιο
Ins	Όργανο
P	Τοποθεσία
Pat	Λήπτης ενέργειας
Loc	Τοποθεσία
Pr	Ιδιότητα

$$O = \{S, \{R_i\}\}, i = 1 \dots n \tag{2.37}$$

$$R_i : S \times S \rightarrow \{0, 1\}, i = 1 \dots n \tag{2.38}$$

όπου O είναι η οντολογία και R_i η i-στή σημασιολογική δυαδική σχέση ανάμεσα στις σημασιολογικές οντότητες. Εδώ έχουμε παραλείψει το τμήμα εξαγωγής συμπερασμάτων, καθώς η έρευνα σε αυτό είναι ακόμη πρώιμη.

Αν και σχέσεις όλων των τύπων υποστηρίζονται από τον ορισμό, οι δύο βασικοί τύποι είναι ταξινομικές σχέσεις (σχέσεις διάταξης) και σχέσεις συμβατότητας (συμμετρικές σχέσεις). Οι σχέσεις συμβατότητας έχουν παραδοσιακά χρησιμοποιηθεί για την επέκταση του ερωτήματος σε συστήματα ανάκτησης πληροφορίας. Αποτυγχάνουν όμως να αξιοποιήσουν σε αυτή τη διαδικασία την πληροφορία του πλαισίου γνώσης που χαρακτηρίζει τη συγκεκριμένη περίπτωση. Έτσι, μια πρόκληση είναι η αξιοποίηση των ταξινομικών σχέσεων για την εκτίμηση και εκμετάλλευση του πλαισίου γνώσης.

Είναι σαφές πως οι σχέσεις ανάμεσα σε πραγματικές οντότητες δεν είναι αυστηρές. Συνηθέστερα ισχύουν ή δεν ισχύουν σε κάποιο βαθμό, και για αυτό το λόγο μοντελοποιούνται καλύτερα με τη χρήση ασαφών δυαδικών σχέσεων. Οι οντολογίες, όμως, δεν περιέχουν στην πράξη ασαφείς σχέσεις και περιορίζονται σε κάποια α-τομή της κάθε σχέσης. Αυτό είναι ένα σημαντικό μειονέκτημα, που τις κάνει ανεπαρκείς για την υπηρεσία της ευφυούς πρόσβασης σε πληροφορία.

Για να ξεπεράσουμε τέτοιες δυσκολίες καταφεύγουμε στη χρήση ασαφών σημασιολογικών ανακλαστικών σχέσεων διάταξης για τη μοντελοποίηση και αναπαράσταση των σχέσεων του πραγματικού κόσμου. Η σχέση εξειδίκευσης Sp, για παράδειγμα, είναι μια ασαφής σχέση μερικής διάταξης ορισμένη στο S^2. $Sp(a, b) > 0$ σημαίνει πως το νόημα του a "περιλαμβάνει" το νόημα του b. Έτσι, αν ένα έγγραφο αναφέρεται στην έννοια b, τότε σχετίζεται και με την έννοια a, ενώ το αντίστροφο δεν ισχύει απαραίτητα. Προφανώς, η σχέση εξειδίκευσης περιέχει σημαντική πληροφορία που δεν είναι δυνατό να περιγραφεί με τη χρήση μιας σχέσης συμβατότητας. Το σύνολο των σχέσεων που χρησιμοποιούμε παρουσιάζονται στον πίνακα 2.1.

Η ασάφεια των σχέσεων έχει το παρακάτω νόημα: υψηλές τιμές της σχέσης $Sp(a, b)$ υπονοούν πως η έννοια b πλησιάζει το νόημα a. Από την άλλη πλευρά, καθώς η $Sp(a, b)$ ελαττώνεται, το νόημα b γίνεται πολύ πιο "στενό" από το a. Αντίστοιχα και για τις άλλες σημασιολογικές σχέσεις.

Αξίζει να σημειωθεί πως

$$a \neq b \rightarrow Sp(a, b) < 1 \tag{2.39}$$

ή ισοδύναμα

$$Sp(a,b) = 1 \rightarrow a = b \tag{2.40}$$

Τέλος, σημαντικό ρόλο έχει και η μεταβατικότητα των σχέσεων. Καθώς είναι σχέσεις διάταξης, είναι προφανές πως διατηρούν την ιδιότητα της μεταβατικότητας. Πιο παραστατικά, αν η έννοια b είναι εξειδίκευση της a και η c είναι εξειδίκευση της b, τότε και η c είναι εξειδίκευση της a. Αντίστοιχα επιχειρήματα μπορούν να γίνουν και για τις άλλες ταξινομικές σχέσεις.

Η μεταβατικότητα, όμως, δεν ορίζεται μονοσήμαντα, καθώς οι σχέσεις είναι ασαφείς. Η μορφή της μεταβατικότητας δεν μπορεί πάντως να είναι $\sup - \min$, γιατί αυτό δεν είναι συμβατό με την ερμηνεία που έχουμε δώσει στους βαθμούς των σχέσεων. Έτσι, απαιτούμε να ισχύει $\sup -t$ μεταβατικότητα, όπου t μια Αρχιμήδεια νόρμα.

Πιο αυστηρά, το μοντέλο γνώσης που παρουσιάσαμε μπορεί να περιγραφεί ως:

$$O_{\mathcal{F}} = \{S, \{r_i\}\}, i = 1 \ldots n \tag{2.41}$$

$$r_i : S \times S \rightarrow [0,1], i = 1 \ldots n \tag{2.42}$$

όπου r_i, $, i = 1 \ldots n$ μια $\sup -t$ μεταβατική σχέση και t μια Αρχιμήδεια νόρμα.

Η ύπαρξη πολλών σχέσεων στην εγκυκλοπαίδεια οδηγεί στο "μοίρασμα" της γνώσης ανάμεσά τους, ώστε η χρήση μιας μόνο από αυτές να μην επαρκεί για την αυτόματη εκτέλεση ενεργειών στο σημασιολογικό επίπεδο. Οδηγούμαστε, λοιπόν, στο συνδυασμό τους για τη δημιουργία μιας νέας σχέσης T που να συνδυάζει πληροφορία από όλες τις διαθέσιμες σχέσεις. Βασιζόμενοι στις σχέσεις r_i κατασκευάζουμε τη σχέση:

$$T = Tr^t(\bigcup_i r_i^{p_i}), p_i \in \{-1, 0, 1\}, i \in 1 \ldots n \tag{2.43}$$

όπου $Tr^t(A)$ είναι το $\sup -t$ μεταβατικό κλείσιμο της σχέσης A. Η μεταβατικότητα της σχέσης T δεν ήταν εξαρχής δεδομένη καθώς η ένωση μεταβατικών σχέσεων δεν είναι απαραίτητα μεταβατική. Επιπρόσθετα, δεν υπαρχει εξασφάλιση πως η T διατηρεί την ιδιότητα της διάταξης. Από την άλλη πλευρά, ˜το σημασιολογικό περιεχόμενο των r_i είναι τέτοιο που η T δεν απέχει πολύ από το να είναι σχέση διάταξης. Έτσι, αναφερόμαστε στην T ως σχέση ημιταξινομίας.

Μια πιθανή σχέση T δημιουργείται με τη χρήση των παρακάτω σχέσεων:

- Εξειδίκευση Sp.
- Πλαίσιο Ct, αντεστραμμένη. $Ct(a,b) > 0$ σημαίνει πως το b είναι το πλαίσιο γνώσης στο οποίο εμφανίζεται η έννοια a.
- Μέρος P, αντεστραμμένη. $P(a,b) > 0$ σημαίνει πως το b είναι μέρος του a.
- Όργανο Ins. $(a,b) > 0$ δείχνει πως το b είναι όργανο του a. Για παράδειγμα a θα μπορούσε να είναι το ποδόσφαιρο και b η μπαλα.
- Τοποθεσία L, αντεστραμμένη. $L(a,b) > 0$ δείχνει πως το b είναι η τοποθεσία του a.
- Λήπτης ενέργειας Pat. $(a,b) > 0$ δείχνει πως το b είναι ο λήπτης της ενέργεια του a.Για παράδειγμα a θα μπορούσε να είναι το μάθημα και b ο μαθητής.

- Ιδιότητα Pr, αντεστραμμένη. $(a, b) > 0$ δείχνει πως το b είναι μια ιδιότητα του a.

Έτσι έχουμε:

$$T = Tr^t(Sp \cup Ct^{-1} \cup Ins \cup P \cup Pat \cup Loc \cup Pr) \tag{2.44}$$

Η σχέση αυτή μπορεί να χρησιμοποιηθεί, όπως θα εξηγηθεί στο κεφάλαιο 6, για την ανάλυση και εκτίμηση του περιεχομένου ενός εγγράφου και την εξαγωγή των θεματικών κατηγοριών στις οποίες αναφέρεται. Αξίζει να σημειωθεί πως το σύνολο TC των θεματικών κατηγοριών είναι υποσύνολο των σημασιολογικών οντοτήτων.

$$TC \subseteq S \tag{2.45}$$

Όμοια, η σχέση

$$T = Tr^t(Sp \cup P^{-1}) \tag{2.46}$$

μπορεί να χρησιμοποιηθεί για επέκταση του ερωτήματος, όπως θα εξηγηθεί στο κεφάλαιο 5, κλπ.

2.5 Πλαίσιο γνώσης

Σύμφωνα με την ερμηνεία που δόθηκε στη σχέση T, ορίζουμε τα πλαίσιο μιας οντότητας k ως το σύνολο $T(k)$ των οντοτήτων που "περιλαμβάνει" σύμφωνα με τη σχέση ημιταξινομίας, δηλ. το σύνολο των απογόνων της. Όταν σε ένα σύνολο, όπως για παράδειγμα σε ένα ερώτημα, περιέχονται δύο οντότητες, πρέπει κανείς να συνδυάσει το πλαίσιό τους, δηλαδή να εντοπίσει το κοινό τους νόημα, για να μπορέσει να εκτιμήσει σωστά το σημασιολογικό περιεχόμενο του συνόλου.

Στο παράδειγμα όπου το σύνολο των οντοτήτων είναι ένα ερώτημα, όταν οι οντότητες έχουν σχετική μεταξύ τους σημασία, η παρουσία τους έχει σκοπό να αποσαφηνίσει το νόημά τους. Αν, αντίθετα, δεν υπάρχει κοινή σημασία, τότε η παρουσία τους έχει σκοπό την ανάκτηση απλώς των εγγράφων που περιέχουν τους όρους. Υποθέτοντας ότι η ερώτηση q δεν είναι ασαφής (δεν έχει βαθμούς), τότε το πλαίσιο $K(q)$ της q, που είναι ένα ασαφές σύνολο όρων, μπορεί να οριστεί απλώς ως το σύνολο των κοινών απογόνων τους:

$$K(q) = \bigcap_{k_i \in q} T(k_i) \tag{2.47}$$

Για παράδειγμα, το πλαίσιο της έννοιας λάστιχο θα περιλαμβάνει το αυτοκίνητο και το αεροπλάνο. Παρομοίως, το πλαίσιο της έννοιας έλικα θα περιλαμβάνει το αεροπλάνο και το πλοίο. Όμως το πλαίσιο του ερωτήματος λάστιχο και έλικα θα περιλαμβάνει μόνο το αεροπλάνο.

Προφανώς,

$$q_1 \subseteq q_2 \implies K(q_1) \supseteq K(q_2) \tag{2.48}$$

δηλαδή η παρουσία επιπλέον όρων θα κάνει το πλαίσιο "στενότερο".

Η δυνατότητα για ασαφείς ερωτήσεις δίνει στο χρήστη τη δυνατότητα να ελέγχει κατά πόσο ένας όρος επηρεάζει το πλαίσιο της ερώτησης. Ένας όρος με βάρος μονάδα δε θα επιτρέπει στο πλαίσιο της ερώτησης να είναι ευρύτερο από το δικό του πλαίσιο.

Αντίθετα, ένας όρος με βάρος κοντά στο μηδέν δεν επηρεάζει σχεδόν καθόλου το πλαίσιο.

Λαμβάνοντας υπόψη τα παραπάνω, απαιτούμε, όταν το σύνολο q είναι κανονικό και ασαφές, το σταθμισμένο πλαίσιο $\mathcal{K}(k_i)$ του k_i, δηλαδή το πλαίσιο λαμβάνοντας υπόψη το βάρος του k_i στο σύνολο q, γίνεται χαμηλό όταν το πλαίσιο $T(k_i)$ είναι χαμηλό και όταν ο βαθμός συμμετοχής w_i του k_i στο q είναι υψηλός. Επομένως:

$$c(\mathcal{K}(k_i)) = c(T(k_i)) \cap w_i S \tag{2.49}$$

όπου S το σύνολο των σημασιολογικών οντοτήτων. Εφαρμόζοντας το νόμο de Morgan, έχουμε:

$$\mathcal{K}(k_i) = T(k_i) \cup c(w_i S) \tag{2.50}$$

Η νόρμα του φραγμένου αθροίσματος διατηρεί καλύτερα το σημασιολογικό περιεχόμενο από ότι η πρότυπη νόρμα max στη σχέση 2.50. Όπως και στη σχέση 2.47, το πλαίσιο του συνόλου είναι η ασαφής τομή των επιμέρους σταθμισμένων πλαισίων:

$$K(q) = \bigcap_{k_i \in q} \mathcal{K}(k_i) \tag{2.51}$$

Η τομή συνεπάγεται ότι όσο μικρότερο είναι το σταθμισμένο πλαίσιο ενός όρου, τόσο μικρότερο θα είναι το συνολικό πλαίσιο.

Στο παράδειγμα του ερωτήματος και πάλι, όταν οι όροι της ερώτησης είναι υψηλά συσχετισμένοι μέσω της T, τότε το πλαίσιο ερώτησης θα έχει υψηλές τιμές. Χρησιμοποιούμε τον όρο ισχύς του πλαισίου για την υψηλότερη από αυτές, δηλαδή για το ύψος $h_q = h(K(q))$ του πλαισίου. Όταν δύο πλαίσια έχουν υψηλή συσχέτιση, θα τέμνονται. Έτσι, η ισχύς της τομής τους είναι ένα μέτρο της σημασιολογικής τους συσχέτισης:

$$sim(q_1, q_2) = h_{q_1 \cup h_{q_2}} = h(K(q_1 \cup q_2)) \tag{2.52}$$

2.6 Πειραματικά αποτελέσματα

Στο σχήμα 2.2 παρουσιάζεται μια απλή ταξινομική σχέση T που έχει φτιαχτεί σύμφωνα με τη σχέση 2.46. Με λεπτές γραμμές παρουσιάζονται τα μέλη της σχέσης Sp ενώ με πιο έντονες γραμμές τα μέλη της σχέσης P. Συνδέσεις που υπονοούνται από τη μεταβατικότητα δεν παρουσιάζονται για λόγους απλότητας. Υπονοείται η χρήση $sup - t$ μεταβατικότητας, όπου t το αλγεβρικό γινόμενο.

Υπολογίζουμε το πλαίσιο για καθένα από τα ακόλουθα ερωτήματα:

$$q_1 = Airplane/1 \tag{2.53}$$

$$q_2 = Airplane/1 + Propeller/0.7 \tag{2.54}$$

$$q_3 = Airplane/1 + Propeller/1 \tag{2.55}$$

Έχουμε λοιπόν:

$$\mathcal{K}(Airplane) = jet/0.9 + prop\ plane/0.9 \tag{2.56}$$

και για καθένα από τα τρία ερωτήματα

$$\mathcal{K}_1(Propeller) = S \tag{2.57}$$

$$\mathcal{K}_2(Propeller) = prop\ plane/1 + 0.3 \cdot S \tag{2.58}$$

$$\mathcal{K}_3(Propeller) = prop\ plane/0.9 \tag{2.59}$$

αντίστοιχα. Συνολικά το πλαίσιο που υπολογίζουμε για καθένα από τα ερωτήματα ακολουθεί:

$$K(q_1) = jet/0.9 + prop\ plane/0.9 \tag{2.60}$$

$$K(q_2) = jet/0.3 + prop\ plane/0.9 \tag{2.61}$$

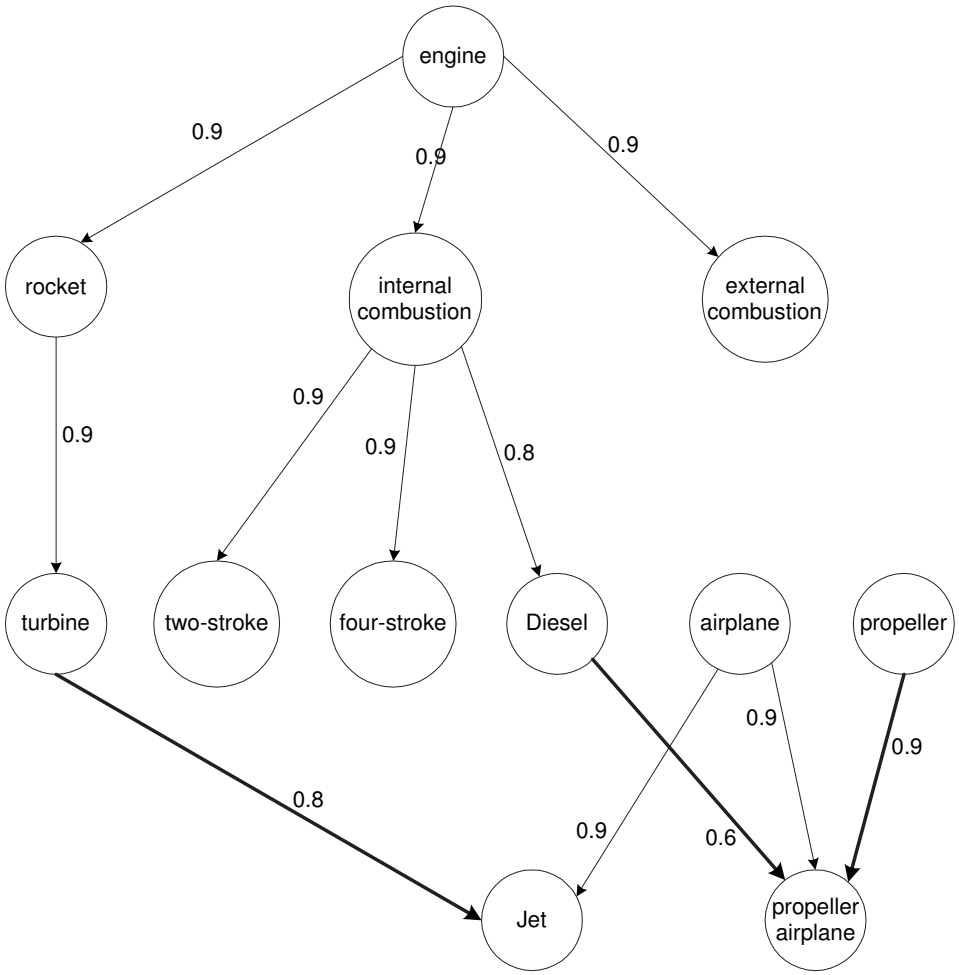
$$K(q_3) = prop\ plane/0.9 \tag{2.62}$$

Τα τρία ερωτήματα δεν ήταν ανεξάρτητα. Είναι και τα τρία της μορφής

$$q = Airplane/1 + Propeller/w \tag{2.63}$$

όπου το w παίρνει τις τιμές $0, 0.7, 1$. Η σταδιακή μεταβολή του πλαισίου με την αλλαγή της τιμής του w είναι προφανής, δείχνοντας καθαρά πως ο προτεινόμενος ορισμός του πλαισίου γνώσης επιτυγχάνει να αποτυπώσει το σημασιολογικό περιεχόμενο του ερωτήματος.

□



Σχήμα 2.2: *Μια απλή σχέση ταξινομίας.*

Κεφάλαιο 3

Αραιές σχέσεις και μεταβατικότητα

3.1 Εισαγωγή

Στις περισσότερες περιπτώσεις η ιδιότητα των ασαφών σχέσεων που είναι πιο σημαντική για την αναπαράσταση γνώσης σχετικής με τον πραγματικό κόσμο είναι αυτή της μεταβατικότητας, καθώς η συνεχής Αρχιμήδεια μεταβατικότητα περιγράφει τη διάδοση της πληροφορίας με ένα ίδιαίτερα φυσικό τρόπο. Έτσι, είναι αναμενόμενα να υπάρχουν μια πλειάδα αναφορών στη βιβλιογραφία που συζητούν την αναπαράσταση [58][82][105] και τις θεωρητικές ιδιότητες των μεταβατικών σχέσεων [30][40][43][44][45][50][62][120][142].

Η μεταβατική ιδιότητα, εξαιτίας της φυσικής της σημασίας, έχει συνδεθεί στενά με τη μελέτη των γράφων. Σε αυτό το πλαίσιο, το μεταβατικό κλείσιμο αντιστοιχεί στην εύρεση των ζευγών κόμβων που συνδέονται είτε άμεσα είτε μέσω κάποιου μονοπατιού. Έτσι, η πλειοψηφία της υπάρχουσας βιβλιογραφίας για μεταβατικό κλείσιμο εστιάζει κυρίως στην περίπτωση γράφων χωρίς βάρη και κατευθύνσεις [115] και γράφων απλάχωρίς βάρη [13][107][123][137], με τις περισσότερες εργασίες να βασίζονται στη δουλειά του Warshall [138]. Στο [126], επιπρόσθετα με την υπολογιστική πολυπλοκότητα της διαδικασίας του μεταβατικού κλεισίματος εξετάζεται και η πολυπλοκότητα εισόδου/εξόδου (I/O), αλλά και σε αυτή την περίπτωση η ανάλυση περιορίζεται στην περίπτωση που δεν υπάρχουν βάρη.

Το μεταβατικό κλείσιμο ασαφών σχέσεων έχει επίσης μελετηθεί, όπως για παράδειγμα στα [42][86][93]. Στα τελεταία δύο η εντυπωσιακή πολυπλοκότητα $O(n^2)$ επιτυγχάνεται. Το ίδιο αποτέλεσμα αναφέρεται και πιο νωρίς [52] άκολουθώντας όμως διαφορετική μεθοδολογία. Η εφαρμογή, όμως, αυτών των αλγορίθμων περιορίζεται στην περίπτωση σχεσεων ομοιότητας, δηλαδή σχέσεων συμμετρικών και ανακλαστικών, και μόνο για την περίπτωση της max-min μεταβατικότητας, δηλαδή για μία μόνη και μη Αρχιμήδεια περίπτωση. Γενικά δεν έχει δοθεί έμφαση στο γενικό $\sup -t$ μεταβατικο κλείσιμο ασαφών σχέσεων και η προσοχή δίνεται σχεδόν πάντα στη max-min μεταβατικότητα. (Τα [53][55] αποτελούν εξαιρέσεις.)

Αυτό που είναι ακόμη πιο σημαντικό είναι πως δεν έχουν προταθεί αλγόριθμοι που να εστιάζουν στην περίπτωση των αραιών σχέσεων. Καθώς οι σχέσεις που χρησιμοποιούνται από τα πεδίο των οντολογιών και της αναζήτησης πληροφορίας είναι αραιές και μεταβατικές, ο χειρισμός τέτοιων σχέσεων έχει πλέον αναδειχθεί σε ένα θέμα με αυξανόμενη σημασία. Σε αυτό το κεφάλαιο, αφού αναλύσουμε το μοντέλο αραιής αναπαράστασης που θα χρησιμοποιήσουμε, και αφού συζητήσουμε τη συμβατική μεθοδολογία μεταβατικού κλεισίματος, παρουσιάζουμε έναν αλγορίθμο που εστιάζεται

στην ανάκτηση της μεταβατικότητας όταν υπάρχει μια τοπική διαταραχή. Η συζήτηση ολοκληρώνεται στο επόμενο κεφάλαιο, όπου παρουσιάζουμε έναν αλγόριθμο για το γρήγορο μεταβατικό κλείσιμο μιας σχέσης, όταν αυτή είναι αραιή.

3.2 Υποθέσεις για την αραιότητα

Οι σχέσεις μπορούν να χρησιμοποιηθούν για να μοντελοποιήσουν διάφορες πλευρές του πραγματικού κόσμου. Ανάλογα με την περίπτωση, οι χρησιμοποιούμενες σχέσεις μπορεί να διαθέτουν διάφορες τυπικές μαθηματικές ιδιότητες όπως ανακλαστικότητα, συμμετρία, μεταβατικότητα της μιας ή της άλλης μορφής κλπ. Ο προσδιορισμός κάθε μίας από αυτές τις ιδιότητες είναι μια αντικειμενική διαδικασία, γεγονός που μας επιτρέπει να διακρίνουμε τους διαφορετικούς τύπους σχέσεων και να τις χειριζόμαστε με διαφορετικούς τρόπους.

Από την άλλη πλευρά, κάποιες υποκειμενικές ιδιότητες έχουν κάποιες φορές σημασία για τις πραγματικές σχέσεις. Η πιο σημαντική, ίσως, από αυτές είναι η αραιότητα. Οι αραιές σχέσεις, ακόμη και όταν έχουν όμοιες αντικειμενικές μαθηματικές ιδιότητες με τις αντίστοιχες πυκνές σχέσεις, χειρίζονται καλύτερα με διαφορετικές μεθοδολογίες. Βέβαια, λέγοντας "καλύτερα" δεν αναφερόμαστε στην εγκυρότητα των αποτελεσμάτων, καθώς αυτά είναι πανομοιότυπα, αλλά στους υπολογιστικούς και αποθηκευτικούς πόρους που απαιτούνται.

Καθώς οι αραιές σχέσεις έχουν σημαντικό ρόλο σε μια σειρά από πεδία, εξειδικευμένες δομές δεδομένων και αντίστοιχοι αλγόιθμοι αναπτύσσονται ειδικά για την περίπτωσή τους. Ένα από τα πιο σημαντικά προβλήματα με αυτές τις δομές και αυτούς τους αλγορίθμους είναι η αξιολόγησή τους. Παραδοσιακά, η αποτελεσματικότητα μιας δομής δεδομένων, καθώς και ενός αλγορίθμου, μετρώνται με βάση τις πολυπλοκότητες (χώρου και χρόνου). Ο ορισμός, όμως, της πολυπλοκότητας συνδέεται με τη θεώρηση της χειρότερης πιθανής περίπτωσης, κατί που είναι αντιφατικό με την έννοια της αραιότητας.

Έτσι, για να γίνει η αξιολόγηση δυνατή, συχνά αναφερόμαστε στη συνήθη ή τυπική περίπτωση, αντί για τη χειρότερη περίπτωση. Με τη σειρά του, αυτό δημιουργεί ανάγκη για αυστηρό ορισμό της έννοιας της τυπικής περίπτωσης. Εδώ, οδηγούμενοι από τα στατιστικά χαρακτηριστικά των σχέσεων που συνηθέστερα απαντώνται στους χώρους των οντολογιών και της αναζήτησης πληροφορίας, ορίζουμε την τυπική περίπτωση ως εξής:

Έστω $n = |S|$ η πληθυκότητα του καθολικού συνόλου. Ένα μικρό και σταθερό ποσοστό p_r των γραμμών και p_c των στηλών μιας τυπικής αραιής σχέσης είναι μη μηδενικές. Έτσι έχουμε $O(n)$ μη μηδενικές γραμμές και $O(n)$ μη μηδενικές στήλες. Επιπρόσθετα, το πλήθος των μη μηδενικών στοιχείων που περιέχονται σε μια μη μηδενική γραμμή ή στήλη είναι ανάλογο του λογαρίθμου του συνολικού αριθμού των στοιχείων. Έτσι έχουμε $O(\log n)$ μη μηδενικά στοιχεία σε κάθε γραμμή ή στήλη. Συνολικά, υπάρχουν $O(n \log n)$ μη μηδενικά στοιχεία στη σχέση.

3.3 Μοντέλο αραιής αναπαράστασης

Στην πρακτική εφαρμογή των παραπάνω κεντρικό πρόβλημα είναι το μέγεθος της σχέσης T, καθώς το πλήθος n των σημασιολογικών οντοτήτων είναι μεγάλο, με αποτέλεσμα να μην είναι διαθέσιμος χώρος για την αποθήκευση των n^2 στοιχείων που

Πίνακας 3.1: *Παράδειγμα αραιής σχέσης*

$$\begin{bmatrix} & (1,2) & & (1,5) & (1,6) \\ (2,1) & & (2,4) & (2,5) & \\ & (3,2) & & & \\ & & (4,4) & & (4,6) \\ (5,1) & & (5,4) & & \end{bmatrix}$$

περιέχει η σχέση. Από την άλλη πλευρά, οι οντολογικές σχέσεις είναι από τη φύση τους αραιές, οπότε ένα μικρό μόνο μέρος ων στοιχείων είναι μη μηδενικό. Μπορούμε, λοιπόν, να ξεπεράσουμε το πρόβλημα του χώρου αποθήκευσης χρησιμοποιώντας μια αραιή αναπαράσταση. Ακόμη, όμως, και έτσι, παραμένουν τα ακόλουθα προβλήματα που σχετίζονται με τη μεταβατική ιδιότητα της σχέσης:

- Ο αλγόριθμος εκτίμησης του πλαισίου γνώσης, καθώς και αλγόριθμοι που θα παρουσιαστούν σε επόμενες ενότητες, υποθέτουν πως η σχέση T είναι μεταβατική. Έτσι, για να είναι δυνατή η εφαρμογή τους πρέπει να υπολογιστεί το μεταβατικό κλείσιμο της $n \times n$ σχέσης.

- Η σχέση T στην πράξη "διορθώνεται" ακολουθώντας μια στρατηγική δοκιμής και λάθους, γεγονός που σημαίνει πως συχνά γίνονται μικρές αλλαγές που διαταράσσουν τοπικά τη μεταβατικότητα. Μετά από κάθε τροποποίηση το μεταβατικό κλείσιμο πρέπει να υπολογιστεί ξανά. Καθώς αυτό είναι υπολογιστικά ασύμφορο, η εφαρμογή της στρατηγικής δοκιμής και λάθους δεν είναι δυνατή

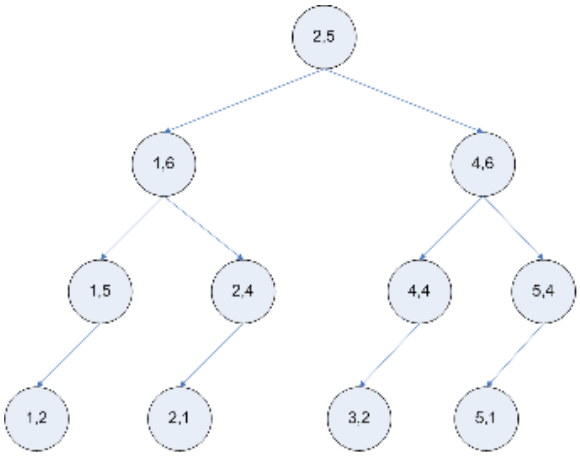
Η κλασική υλοποίηση αραιών πινάκων χρησιμοποιεί συνδεδεμένες λίστες για την αναπαράσταση των μη μηδενικών στοιχείων, ανεβάζοντας έτσι την πολυπλοκότητα της πρόσβασης σε ένα στοιχείο από $O(1)$ σε $O(n)$. Αν και οι απαιτήσεις σε χώρο αποθήκευσης περιορίζονται σημαντικά, το μοντέλο παραμένει μη εφαρμόσιμο για την αναπαράσταση και πρακτική αξιοποίηση οντολογικών σχέσεων, καθώς απαιτείται πλήθος αλγεβρικών πράξεων με τη χρήση των σχέσεων πριν το σύστημα που τις χρησιμοποιεί αποκριθεί.

Για να υπερβούμε αυτές τις δυσκολίες προτείνουμε το ακόλουθο μοντέλο: μια σχέση αναπαρίσταται χρησιμοποιώντας δύο δέντρα AVL. Το δέντρο AVL είναι ένα δυϊκό, ισοσκελισμένο και διαταγμένο δέντρο στο οποίο η πρόσβαση, η προσθήκη και η διαγραφή ενός κόμβου επιτυγχάνονται σε χρόνο $O(\log m)$, όπου m είναι το πλήθος των κόμβων του δέντρου [2]. Αν $n \log n$ κόμβοι υπάρχουν στο δέντρο, όπως θα συμβαίνει για την τυπική αραιή σχέση, τότε η πολυπλοκότητα για την πρόσβαση, την προσθήκη και τη διαγραφή παραμένουν $O(\log n)$ καθώς $n < n \log n < n^2 \Rightarrow O(\log n) \leq O(\log(n \log n)) \leq O(\log n^2) = O(\log n)$.

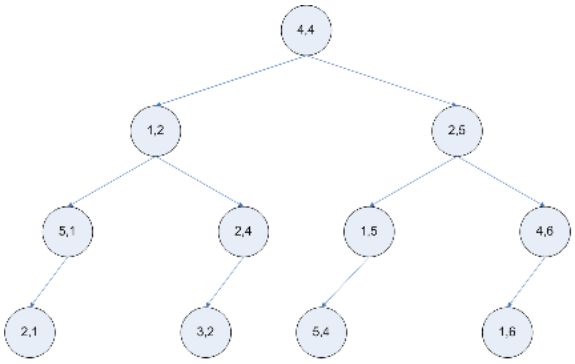
Και στα δύο δέντρα ο δείκτης γραμμής i και ο δείκτης στήλης j χρησιμοποιούνται για να διατάξουν λεξικογραφικά τους κόμβους. Στο πρώτο δέντρο ο δείκτης γραμμής χρησιμοποιείται πρώτα και σε περίπτωση κοινής γραμμής χρησιμοποιείται ο δείκτης στήλης και αντίστροφα για το δεύτερο δέντρο. Θα χρησιμοποιήσουμε τον πίνακα 3.1 για να διευκρινίσουμε τη διαδικασία.

Η ακολουθούμενη λεξικογραφική διάταξη παράγει τα ακόλουθα διανύσματα.

$[(1,2)(1,5)(1,6)(2,1)(2,4)(2,5)(3,2)(4,4)(4,6)(5,1)(5,4)]^T$



Σχήμα 3.1: *Το δέντρο που διατάσσεται με βάση το δείκτη γραμμής i.*



Σχήμα 3.2: *Το δέντρο που διατάσσεται με βάση το δείκτη στήλης j.*

$[(2,1)(5,1)(1,2)(3,2)(2,4)(4,4)(5,4)(1,5)(2,5)(1,6)(4,6)]^T$
και αυτά με τη σειρά τους αντιστοιχούν στα AVL δέντρα των σχημάτων 3.1 και 3.2. Βέβαια τα δέντρα δεν είναι μοναδικά, καθώς πολλαπλά ισοσκελισμένα δέντρα μπορούν να δημιουργηθούν όταν το πλήθος των στοιχείων δεν είναι ίσο με $(2^k - 1)$ για κάποιο $k \in \mathcal{N}$.

Αυτή η αναπαράσταση διατηρεί τα πλεονεκτήματα χώρου της κλασικής μεθοδολογίας αραιών πικάκων με τις συνδεδεμένες λίστες. Επιπρόσθετα, επιτρέπει την πρόσβαση σε ένα στοιχείο, μια στήλη ή μια γραμμή σε χρόνο $O(\log n)$, που είναι σημαντικά μικρότερος από αυτόν της γραμμικής πολυπλοκότητας των λιστών. Τέλος, η πολυπλοκότητα της εισαγωγής και της διαγραφής είναι επίσης $O(\log n)$.

3.4 Συμβατική μεθοδολογία μεταβατικού κλεισίματος

Το μεταβατικό κλείσιμο μια πλήρους σχέσης υπολογίζεται με την ακόλουθη μέθοδο [73]:

Στη γενική περίπτωση, το μεταβατικό κλείσιμο $Tr^t(r)$ της σχέσης r, με δεδομένη την τ-νόρμα t, υπολογίζεται ως:

$$Tr^t(r) = \bigcup_{f=1}^{\infty} r^f \tag{3.1}$$

$$r^{f+1} = r^f \circ^t r \tag{3.2}$$

$$r^1 = r \tag{3.3}$$

Στη σχέση 3.1, καθώς και στις υπόλοιπες σχέσεις της παραγράφου, χρησιμοποιείται η κλασσική ένωση max. Υποθέτοντας πως το καθολικό σύνολο S είναι φραγμένο ($|S| = n$), η σχέση 3.1 γράφεται ως:

$$Tr^t(r) = \bigcup_{f=1}^{n-1} r^f \tag{3.4}$$

Επιπλέον, μπορούμε να παραλείψουμε ορισμένα από τα βήματα της μεθόδου, μειώνοντας έτσι την υπολογιστική της πολυπλοκότητα, αν αντικαταστήσουμε τη σχέση 3.2 με την ακόλουθη [73]:

$$r^{2f} = r^f \circ^t r^f \tag{3.5}$$

Τέλος, για να αποφύγουμε της εκτέλεση του υπολογιστικά ακριβού βήματος της σύνθεσης πιο πολλές φορές από όσες είναι απαραίτητο, μπορούμε να σταματήσουμε τη διαδικασία πριν από το σημείο όπου $f \geq n-1$, αν βρεθεί πως $r^{2f} = r^f$, καθώς τότε εύκολα μπορούμε να δείξουμε πως $Tr^t(r) = r^f$.

Μια ειδική περίπτωση είναι η σχέση r να είναι ανακλαστική. Σε αυτή την περίπτωση αποδεικνύεται πως το μεταβατικό κλείσιμο δίνεται από τη σχέση

$$Tr^t(r) = r^{n-1} \tag{3.6}$$

Ο υπολογισμός της σύνθεσης σχέσεων έχει πολυπλοκότητα $O(n^3)$. Έτσι, το μεταβατικό κλείσιμο έχει πολυπλοκότητα $O(n^3 logn)$, τόσο για ανακλαστικές, όσο και για μη ανακλαστικές σχέσεις.

Το $\sup -t$ μεταβατικό κλείσιμο ασαφών σχέσεων μπορεί να υπολογιστεί με πολυπλοκότητα $O(n^3)$ χρησιμοποιώντας την πιο αποδοτική μεθοδολογία του [101]:

Αλγόριθμος $O(n^3)$ $\sup -t$**:**
Παράμετροι: R
Έξοδος: R

1. Για $i = 1 \ldots n$

   (α) Για $j = 1 \ldots n$

      i. Για $k = 1 \ldots n$

$$R(j,k) \leftarrow \sup(R(j,k), R(j,i) \wedge_t R(i,k))$$

Η αρχική διατύπωση αυτής της προσέγγισης βρίσκεται στο [55] και βασίζεται στις εργασίες [10], [8] και [9]. Αντίθετα με την μεθοδολογία του Dunn που παρουσιάσαμε παραπάνω, αυτός ο αλγόριθμος δεν μπορεί να προσαρμοστεί για εφαρμογή με το αραιό μοντέλο αναπαράστασης, ώστε να χρησιμοποιηθεί για συγκρίσεις με τη μεθοδολογία που θα παρουσιαστεί στις επόμενες ενότητες. Έτσι, στη συνέχεια με τον όρο κλασική μέθοδος θα αναφερόμαστε στη μεθοδολογία του Dunn, την οποία και θα χρησιμοποιήσουμε σε κάθε συγκριτική μελέτη.

Θεώρημα

Το μεταβατικό κλείσιμο με τη χρήση της προτεινόμενης αναπαράστασης και της κλασικής μεθόδου γίνεται με υπολογιστική πολυπλοκότητα $O(n^2log^2n)$ και $O(n^3log^2n)$ στην τυπική και τη χειρότερη περίπτωση αραιής σχέσης, αντίστοιχα.

Απόδειξη

Αν η γραμμή i και η στήλη j υπάρχουν στη σχέση, δηλαδή αν περιέχουν το λιγότερο από ένα μη μηδενικό στοιχείο, τότε εντοπίζονται στην αναπαράσταση της σχέσης. Όπως έχουμε ήδη αναφέρει, αυτό έχει πολυπλοκότητα $O(logn)$. Στη συνέχεια, το στοιχείο $r_{ij}^{(2)}$ της σύνθεσης υπολογίζεται ως

$$r_{ij}^{(2)} = \bigcup_{r_{ic} \in row \vee r_{cj} \in col} r_{ic} \cap_t r_{cj} \tag{3.7}$$

Καθώς τόσο η γραμμή, όσο και η στήλη, είναι διαθέσιμες σε ταξινομημένη μορφή, η παραπάνω πράξη έχει πολυπλοκότητα $O(k_i^{row} + k_j^{col})$, όπου k_i^{row} το πλήθος των μη μηδενικών στοιχείων της γραμμής i της σχέσης και k_j^{col} το πλήθος των μη μηδενικών στοιχείων της στήλης j. Για τον υπολογισμό της σύνθεσης οι πράξεις αυτής της μορφής που θα πρέπει να γίνουν είναι $a \cdot b$, όπου a το πλήθος των μη μηδενικών γραμμών και b το πλήθος των μη μηδενικών στηλών της σχέσης.

$$2 \cdot O(logn) + O(ab) \cdot O(max(k_i^{row}, k_j^{col})) \tag{3.8}$$

Στην τυπική περίπτωση, ένα ποσοστό των γραμμών και των στηλών θα είναι μηδενικό. Αυτό δεν επηρεάζει την πολυπλοκότητα καθώς

$$O(ab) = O((p_r n)(p_c n)) = O(n^2) \tag{3.9}$$

Όσο αφορά στο πλήθος που περιέχονται σε μια μη μηδενική γραμμή ή σε μια μη μηδενική στήλη, αυτό είναι ανάλογο του λογαρίθμου του πλήθους των στοιχείων του καθολικού συνόλου. Έτσι η $O(k_i^{row} + k_j^{col})$ αντικαθίσταται από $O(logn)$. Συνολικά έχουμε

$$2 \cdot O(logn) + O(n^2) \cdot O(logn) = O(n^2logn) \tag{3.10}$$

για τη σύνθεση της σχέσης με τον εαυτό της.

Στη χειρότερη περίπτωση η σχέση είναι πλήρης, οπότε $a = b = k_i^{row} = k_j^{col} = n$ και η πολυπλοκότητα της σύνθεσης είναι

$$2 \cdot O(logn) + O(n^2) \cdot O(n) = O(n^3) \tag{3.11}$$

Λαμβάνοντας υπόψη πως $O(logn)$ συνθέσεις χρειάζονται για το μεταβατικό κλείσιμο, είναι εύκολο να δούμε πως στην τυπική περίπτωση η πολυπλοκότητα του μεταβατικού κλεισίματος είναι $O(n^2log^2n)$ και στη χειρότερη περίπτωση $O(n^3logn)$.

✓

Πίνακας 3.2: *Σύνοψη υπολογιστικής πολυπλοκότητας για συνθεση και μεταβατικό κλείσιμο με την κλασική μέθοδο*

Πράξη	Μοντέλο	Αραιή σχέση	Πυκνη σχέση
Σύνθεση	Πλήρες	n^3	n^3
Σύνθεση	Αραιό	$n^2 \log n$	n^3
Μεταβατικό κλείσιμο	Πλήρες	$n^3 \log n$	$n^3 \log n$
Μεταβατικό κλείσιμο	Αραιό	$n^2 \log^2 n$	$n^3 \log n$

3.4.1 Συγκριτική μελέτη

Όσο αφορά στις υπολογιστικές απαιτήσεις για το μεταβατικό κλείσιμο, απαιτείται $O(n^3 \log n)$ χρόνος για την πλήρη αναπαράσταση και $O(n^2 \log^2 n)$ ή $O(n^3 \log n)$ για την αραιή αναπαράσταση, στην τυπική και τη χειρότερη περίπτωση αντίστοιχα. Βλεπουμε δηλαδή πως η προτεινόμενη αναπαράσταση πετυχαίνει βελτιωμένη απόδοση στην αραιή περίπτωση χωρίς να υστερεί σε απόδοση στη χειρότερη περίπτωση.

Όσο αφορά στις απαιτήσεις χώρου, η ύπαρξη δύο αντιγράφων της σχέσης κατά της εκτέλεση του αλγορίθμου απαιτείται και στις δύο περιπτώσεις αναπαράστασης (πλήρη και αραιή). Οι ακριβείς απαιτήσεις χώρου, όμως, διαφέρουν σημαντικά όταν η σχέση είναι αραιή:

1. Στην πυκνή αναπαράσταση n^2 στοιχεία αποθηκεύονται για κάθε αντίγραφο της σχέσης, οδηγόντας σε συνολικό χώρο για $2 \cdot n^2$ στοιχεία. Δηλαδή χώρο $O(n^2)$

2. Με την προτεινόμενη αραιή αναπαράσταση αποθηκεύονται δύο δέντρα, καθένα από τα οποία περιέχει τόσα στοιχεία, όσα και η εξεταζόμενη σχέση. Στην τυπική αραιή περίπτωση αυτό αντιστοιχεί σε $2 \cdot O(n \log n)$ κόμβους για μια σχέση και $4 \cdot O(n \log n) = O(n \log n)$ και για τα δύο αντίγραφα της σχέσης. Στη χειρότερη περίπτωση απαιτείται χώρος $O(n^2)$, όπως και στην πυκνή αναπαράσταση.

Ο πίνακας 3.2 συνοψίζει τα συμπεράσματα σχετικά με την υπολογιστική πολυπλοκότητα για τη σύνθεση και το μεταβατικό κλείσιμο σχέσεων με χρήση της κλασικής μεθόδου, για την πυκνή και την προτεινόμενη αραιή αναπαράσταση. Βλέπει κανείς πως η προτεινόμενη αναπαράσταση οδηγεί σε βελτιωμένες επιδόσεις, ακόμη και όταν συνδυάζεται με αλγορίθμους που δεν έχουν σχεδιαστεί ειδικά για την αραιή περίπτωση.

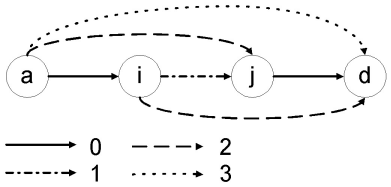
3.5 Αλγόριθμος σταδιακής ενημέρωσης ITU

Βασικό μειονέκτημα της κλασικής προσέγγισης στο μεταβατικό κλείσιμο είναι το γεγονός πως κάθε φορά που η σχέση μεταβάλλεται, πρέπει να εκτελείται μια πράξη μεγάλης πολυπλοκότητας για να εξασφαλίζεται η διατήρηση της μεταβατικότητας.

Θεώρημα

Όταν ένα στοιχείο προστίθεται στη σχέση, ή όταν η τιμή ενός στοιχείο μεγαλώνει, απαιτείται μια πράξη πολυπλοκότητας $O(n^3)$ στη χειρότερη περίπτωση και $O(n^2 logn)$ στην τυπική περίπτωση για τη διατήρηση της μεταβατικότητας.

Σχήμα 3.3: *Γραφική αναπαράσταση της σταδιακής ενημέρωσης της σχέσης.*



Απόδειξη

Έστω r μια μεταβατική σχέση. Στο σχήμα 3.3 η σχέση r παρουσιάζεται με γραμμές τύπου 0. Υποθέτουμε τώρα πως το στοιχείο r_{ij} προστίθεται στη σχέση, ή η τιμή του μεγαλώνει. Στο σχήμα 3.3 η αλλαγή παρουσιάζεται με γραμμή τύπου 1. Βέβαια, μετά από αυτή την αλλαγή δεν μπορούμε να υποθέτουμε πως η σχέση r παραμένει μεταβατική.

Μετά από μια σύνθεση της r με τον εαυτό της, ο πρόγονος a του στοιχείου i συνδέεται με το στοιχείο j, ενώ το στοιχείο i συνδέεται με τον απόγονο d του στοιχείου j. Στο σχήμα 3.3 αυτό παρουσιάζεται με γραμμές τύπου 2. Τέλος, μετά από άλλη μια σύνθεση, το στοιχείο a συνδέεται με το στοιχείο b. Στο σχήμα 3.3 αυτό παρουσιάζεται με τη γραμμή τύπου 3.

Καθώς χρειαζόμαστε ακριβώς δύο συνθέσεις, η συνολική πολυπλοκότητα είναι ίση με την πολυπλοκότητα της σύνθεσης.

✓

Έχοντας παρατηρήσει στο σχήμα 3.3 τον τρόπο με τον οποίο η μεταβατικότητα επαναφέρεται μετά από την αλλαγή σε ένα μόνο στοιχείο της σχέσης, μπορούμε να σχεδιάσουμε έναν αλγόριθμο για τη σταδιακή ενημέρωση μεταβατικών σχέσεων που θα έχει μικρότερη υπολογιστική πολυπλοκότητα καθώς θα εστιάζει στις αλλαγές που πρέπει να γίνουν ώστε η μεταβατικότητα να είναι σίγουρη.

Για οποιαδήποτε σχέση r αυτό επιτυγχάνεται με τα ακόλουθα βήματα:

Αλγόριθμος ITU (Incremental Transitive Update):

Παράμετροι: $R\ i\ j$

Έξοδος: R

1. Αναγνώριση του ασαφούς συνόλου A των προγόνων του στοιχείου i. Οι βαθμοί στο σύνολο A καθορίζονται ως

$$A(s) = r(s, i), s \in S \tag{3.12}$$

2. Αναγνώριση του ασαφούς συνόλου D των απογόνων του στοιχείου j. Οι βαθμοί στο σύνολο A καθορίζονται ως

$$A(s) = r(j, s), s \in S \tag{3.13}$$

3. Για κάθε στοιχείο s που εμφανίζεται στο A θέτουμε

$$r(s, j) \leftarrow r(s, j) \cup (r(s, i) \cap_t r(i, j)) \tag{3.14}$$

4. Για κάθε στοιχείο s που εμφανίζεται στο D θέτουε

$$r(i,s) \leftarrow r(i,s) \cup (r(i,j) \cap_t r(j,s)) \tag{3.15}$$

5. Για κάθε στοιχείο s_1 που εμφανίζεται στο A και για κάθε στοιχείο s_2 που εμφανίζεται στο D θέτουμε

$$r(s_1,s_2) \leftarrow r(s_1,s_2) \cup (r(s_1,j) \cap_t r(i,j) \cap_t r(j,s_2)) \tag{3.16}$$

Αν η σχέση r είναι ανακλαστική, τότε $A(i) = 1$ και $D(j) = 1$. Επομένως η παραπάνω διαδικασία απλοποιείται με την παράλειψη των βημάτων 3 και 4.

Θεώρημα

Η υπολογιστική πολυπλοκότητα του αλγορίθμου σταδιακής ενημέρωσης είναι $O(n^2 logn)$ στη χειρότερη περίπτωση και $O(log^3 n)$ στην τυπική περίπτωση, τόσο για ανακλαστικές, όσο και για μη ανακλαστικές σχέσεις.

Απόδειξη

Έχει ήδη εξηγηθεί πως η πολυπλοκότητα των βημάτων 1 και 2 είναι $O(logn)$, όση δηλαδή η πολυπλοκότητα πρόσβασης σε μια γραμμή ή μια στήλη μέσω των δέντρων AVL. Η πολυπλοκότητα των βημάτων 3 και 4 είναι $O(k_i^{row}) \cdot O(logn)$ και $O(k_j^{col}) \cdot O(logn)$, αντίστοιχα, και η πολυπλοκότητα του βήματος 5 είναι $O(k_i^{row} \cdot k_j^{col}) \cdot O(logn)$.

Στην τυπική περίπτωση $k_i^{row} = k_j^{col} = O(logn)$, και έτσι η συνολική πολυπλοκότητα είναι

$$2 \cdot O(logn) + 2 \cdot O(log^2 n) + O(log^3 n) = O(log^3 n) \tag{3.17}$$

Στη χειρότερη περίπτωση $k_i^{row} = k_j^{col} = n$, και έτσι η συνολική πολυπλοκότητα είναι

$$2 \cdot O(logn) + 2 \cdot O(nlogn) + O(n^2 logn) = O(n^2 logn) \tag{3.18}$$

Αγνοώντας τα βήματα 3 και 4 στους παραπάνω υπολογισμούς δεν επηρεάζουμε την πολυπλοκότητα, καθώς σε κάθε περίπτωση κάποιο άλλο βήμα συνεισφέρει περισσότερο σε αυτή. Έτσι, η ίδιο πολυπλοκότητα ισχύει και για την περίπτωση που η σχέση r είναι ανακλαστική.

✓

Είναι ακόμη δυνατό να επεκταθεί η παραπάνω μεθοδολογία ώστε να μπορεί να εφαρμοστεί και όταν χρησιμοποιείται το πλήρες μοντέλο αναπαράστασης:

Αλγόριθμος ITU για πλήρη αναπαράσταση:
Παράμετροι: r i j
Έξοδος: r

1. Για κάθε στοιχείο s_1 στη στήλη i

   (α) Για κάθε στοιχείο s_2 στη γραμμή j
   Ανάθεσε:

$$r(s_1,s_2) \leftarrow \sup(r(s_1,s_2), r(s_1,i) \wedge_t r(i,j) \wedge_t r(j,s_2)) \tag{3.19}$$

Θεώρημα

Η υπολογιστική πολυπλοκότητα του αλγορίθμου ITU είναι $O(n^2)$, θεωρώντας πλήρες μοντέλο αναπαράστασης.

Απόδειξη

Με το πλήρες μοντέλο αναπαράστασης η σχέση αποθηκεύεται σαν ένας $n \times n$ πίνακας. Έτσι, υπαρχουν n στοιχεία σε κάθε στήλη και σε κάθε γραμμή. Καθώς η σχέση 3.19 περιγράφει απλά δύο τομές, υπολογίζεται σε χρόνο $O(1)$. Συνολικά έχουμε μια πολυπλοκότητα:

$$O(n) \cdot O(n) \cdot O(1) = O(n^2)$$

✓

Μένει να δείξει κανείς πως η έξοδος του αλγορίθμου ITU είναι πραγματικά μια μεταβατική σχέση:

Λήμμα

Αν η σχέση r είναι $\sup -t$ μεταβατική στο S, τότε είναι επίσης $\sup -t$ μεταβατική στο $S' \supseteq S$.

Απόδειξη

Μια ασαφής σχέση r ορισμένη στο S καλείται $\sup -t$ μεταβατική όταν

$$r(x,z) \geq \sup_{y \in S}\{t(r(x,y), r(y,z))\}, \forall (x,z) \in S^2$$

Αν $x \in (S' - S)$, τότε $r(x,z) = 0$. Σε αυτή την περίπτωση

$$\sup_{y \in S'}\{t(r(x,y), r(y,z))\} = \sup_{y \in S'}\{0, r(y,z))\} = 0$$

και έτσι η μεταβατικότητα ισχύει και στο S'. Όμοια, αν $z \in (S' - S)$.

Αν $(x,z) \in S^2$, τότε

$$\sup_{y \in S'}\{t(r(x,y), r(y,z))\} =$$

$$= \sup(\sup_{y \in S}\{t(r(x,y), r(y,z))\}, \sup_{y \in (S'-S)}\{t(r(x,y), r(y,z))\})$$

$$= \sup(\sup_{y \in S}\{t(r(x,y), r(y,z))\}, 0) = \sup_{y \in S}\{t(r(x,y), r(y,z))\}$$

Καθώς η r είναι μεταβατική στο S,

$$r(x,z) \geq \sup_{y \in S}\{t(r(x,y), r(y,z))\} = \sup_{y \in S'}\{t(r(x,y), r(y,z))\}$$

και έτσι η μεταβατικότητα ισχύει και σε αυτή την περίπτωση.

✓

Με αυτό το λήμμα έχουμε δείξει πως η επέκταση του συνόλου ορισμού με την προσθήκη νέων στοιχείων, όπως συμβαίνει σε κάποιες περιπτώσεις κατά την εκτέλεση του αλγορίθμου ITU, δεν βλάπτει τη μεταβατικότητα, όταν η σχέση είναι ήδη μεταβατική.

Για τη συνέχεια της απόδειξης χωρίζουμε τη λειτουργία του ITU σε δύο βήματα:

1. προσθήκη νέων στοιχείων στο σύνολο ορισμού της σχέσης (αν χρειάζεται)
2. αύξηση της τιμής της σχέσης ανάμεσα σε στοιχεία που ήδη υπάρχουν

Σύμφωνα με το προηγούμενο λήμμα, μόνο το δεύτερο από αυτά τα βήμαρα επηρεάζει τη μεταβατικότητα και πρέπει να εξεταστεί. Έτσι περιορίζουμε την ανάλυσή μας στην περίπτωση που μια υπάρχουσα τιμή στη σχέση αυξάνεται (η αρχική τιμή μπορεί να είναι οποιαδήποτε στο πεδίο [0 1]). Με το επόμενο λήμμα ολοκληρώνεται η απόδειξη δείχνοντας πως ο ITU επιτυχώς επανακτά τη μεταβατικότητα όταν μια υπάρχουσα τιμή αυξάνεται.

Λήμα

Έστω r μια μεταβατική σχέση στο S. Έστω επίσης $i, j \in S$. Έστω q τυχαίος αλλά σταθερός αριθμός στο [0 1]. Τέλος, έστω η σχέση r_1, ορισμένη ως

$$r_1(x,z) = \begin{cases} \max(r(i,j),q), & \text{x=i και z=j;} \\ r(x,z), & \text{αλλιώς.} \end{cases}$$

Τότε η σχέση r', υπολογισμένη ως $r' = ITU(r_1, i, j)$, είναι μεταβατική.

Απόδειξη

Αν $r(i,j) \geq q$ από κατασκευή $r' = r$ και έτσι η μεταβατικότητα ισχύει. Αν $r(i,j) < q \Rightarrow r'(i,j) > r(i,j)$. Τότε πρέπει να δείξουμε πως

$$r'(x,z) \geq \sup_{y \in S} t(r'(x,y), r'(y,z)), \forall (x,z) \in S^2 \tag{3.20}$$

Για y τέτοια ώστε $r'(x,y) = r(x,y)$, εύκολα

$$\begin{aligned} &\sup_{y: r'(x,y) = r(x,y)} t(r'(x,y), r'(y,z)) = \\ &= \sup_{y: r'(x,y) = r(x,y)} t(r(x,y), r'(y,z)) \\ &\leq r(x,y) \\ &\leq r'(x,y) \end{aligned}$$

Εστιάζοντας στις υπόλοιπες περιπτώσεις, με y τέτοια ώστε $r'(x,y) > r(x,y)$, η απόδειξη βασίζεται στην παρατήρηση πως ο αλγόριθμος ITU επηρεάζει μόνο συγκεκριμένα στοιχεία της σχέσης.

Ας είναι $X = {}_{0+}A$. Αυτή είναι η ισχυρή 0-τομή του ασαφούς συνόλου A των προγόνων του i και περιέχει τα στοιχεία του S που συμμετέχουν στο σύνολο A σε μη μηδενικό βαθμό. Όμοια, ας είναι $Z = {}_{0+}D$.

Αν $x \in X \cup \{i\}$ και $z \in Z \cup \{j\}$, τότε η σχέση eq. 3.20 ισχύει από την κατασκευή.

Αν $x \notin X$ και $z \notin Z$, τότε έχουμε από κατασκευή $r'(x,y) = r(x,y)$ και $r'(y,z) = r(y,z)$. Έτσι

$$\sup_{y:r'(x,y)>r(x,y)} t(r'(x,y), r'(y,z)) =$$

$$= \sup_{y:r'(x,y)>r(x,y)} t(r(x,y), r(y,z))$$

$$\leq r(x,z)$$

$$= r'(x,z)$$

Αν $x \in X \cup \{i\}$ και $z \notin Z$, τότε έχουμε από κατασκευή ότι $r'(y,z) = r(y,z)$ και ότι $sup_{y:r'(x,y)>r(x,y)} t(r(j,y), r(y,z)) = 0$. Αν $r'(x,y) = r(x,y)$ η εξίσωση 3.20 προφανώς ισχύει. Αν $r'(x,y) > r(x,y)$, τότε έχουμε από κατασκευή $r'(x,y) = t(r(x,i), r_1(i,j), r(j,y))$. Τότε

$$\sup_{y:r'(x,y)>r(x,y)} t(r'(x,y), r'(y,z)) =$$

$$= \sup_{y:r'(x,y)>r(x,y)} t(r(x,i), r_1(i,j), r(j,y), r(y,z))$$

$$\leq \sup_{y:r'(x,y)>r(x,y)} t(r(j,y), r(y,z))$$

$$= 0$$

διότι $z \notin Z$. Έτσι η εξίσωση 3.20 ισχύει. Όμοια αν $x \notin X$ και $z \in Z$. Αν $x = i$ και $z \notin Z$ η απόδειξη ακολουθεί τα ίδια βήματα όπως παραπάνω. Όμοια αν $x \notin X$ και $z = j$, περίπτωση με την οποία ολοκληρώνονται όλα τα ενδεχόμενα. Έτσι η R' είναι μεταβατική.

✓

3.5.1 Αριθμητικό παράδειγμα

Σε αυτή την ενότητα παρουσιάζουμε ένα αριθμητικό παράδειγμα εφαρμογής του αλγορίθμου ITU, για να εξηγήσουμε καλύτερα τη λειτουργία του. Ως αρχική σχέση θεωρείται η R_{input} που παρουσιάζεται στον πίνακα 3.3. Η θεωρούμενη t-νόρμα για την $\sup - t$ μεταβατικότητα είναι η φραγμένη διαφορά. Το στοιχείο που προστίθεται στη σχέση είναι το (#9,#6,0.95). Παρακάτω εξηγούμε τη λειτουργία καθενός από τα βήματα του αλγορίθμου στην αρχική σχέση R_{input} ώστε να ληφθεί η τελική R_{output} ως αποτέλεσμα.

1. Το ασαφές σύνολο A είναι η στήλη #9.
2. Το ασαφές σύνολο D είναι η γραμμή #6.
3. Η στήλη #6 δημιουγείται
4. Η γραμμή #9 δημιουργείται.
5. Τα στοιχεία #3 → #7, #4 → #7 και #5 → #7ενημερώνονται.

Πίνακας 3.3: *Αρχική μεταβατική σχέση* R_{input}

	#1	#2	#3	#4	#7	#8	#9
#1	0.80	0.95	0.90	0.85	0.85	0.90	0.80
#2	0.85	0.80	0.95	0.90	0.90	0.95	0.85
#3	0.90	0.85	0.80	0.95	0.75	0.80	0.90
#4	0.95	0.90	0.85	0.80	0.80	0.85	0.95
#5	0.85	0.80	0.95	0.90	0.70	0.75	0.85
#6					0.95	0.90	
#7					0.90	0.95	
#8					0.95	0.90	

Πίνακας 3.4: *Αποτέλεσμα* R_{output} *του ITU μετά την προσθήκη του στοιχείου (#9,#6,0.95)*

	#1	#2	#3	#4	#6	#7	#8	#9
#1	0.80	0.95	0.90	0.85	0.75	0.85	0.90	0.80
#2	0.85	0.80	0.95	0.90	0.80	0.90	0.95	0.85
#3	0.90	0.85	0.80	0.95	0.85	0.80	0.80	0.90
#4	0.95	0.90	0.85	0.80	0.90	0.85	0.85	0.95
#5	0.85	0.80	0.95	0.90	0.80	0.75	0.75	0.85
#6						0.95	0.90	
#7						0.90	0.95	
#8						0.95	0.90	
#9					0.95	0.90	0.80	

Πίνακας 3.5: *Σύνοψη υπολογιστικής πολυπλοκότητας αλγορίθμων ανάκτησης μεταβατικότητας*

Αλγόριθμος	Μοντέλο	Αραιή σχέση	Πυκνή σχέση
Κλασσικός	Πλήρες	n^3	n^3
Κλασσικός	Αραιό	$n^2 \log n$	n^3
ITU	Πλήρες	n^2	n^2
ITU	Αραιό	$\log^3 n$	$n^2 \log n$

3.5.2 Συγκριτική μελέτη

Στον πίνακα 3.5 παρουσιάζουμε μια σύνοψη των υπολογιστικών πολυπλοκοτήτων των μεθόδων ανάκτησης της μεταβατικότητας που έχουμε αναφέρει.

Στην περίπτωση της τυπικής αραιής σχέσης, ο προτεινόμενος ITU αλγόριθμος ξεπερνά σε επίδοση τον κλασικό, και για τα δύο μοντέλα αναπαράστασης. Αξίζει να αναφερθεί, ωστόσο, πως μόνο με το συνδυασμό του προτεινόμενου αραιού μοντέλου και του προτεινόμενου αλγορίθμου φτάνουμε σε πολυπλοκότητες μικρότερες της γραμμικής, ενώ η κλασική μέθοδος με το πλήρες μοντέλο έχει πολυπλοκότητα $O(n^3)$. Στην περίπτωση πυκνών σχέσεων το προτεινόμενο ζεύγος μοντέλου και αλγορίθμου πετυχαίνουν πολυπλοκότητα $O(n^2 \log n)$, όταν η κλασική μεθοδολογία μένει στο $O(n^3)$.

Ακόμη και όταν ακολουθείται το πλήρες μοντέλο αναπαράστασης, ο αλγόριθμος ITU ξεπερνά τον κλασικό, έχοντας πολυπλοκότητα $O(n^2)$, σε σύγκριση με $O(n^3)$.

Τέλος ο προτεινόμενος αλγόριθμος είναι πιο αποδοτικός ως προς το χώρο αποθήκευσης, ανεξαρτήτως μοντέλου αναπαράστασης, καθώς δεν χρειάζεται την ταυτόχρονη ύπαρξη δύο αντιγράφων της σχέσης.

3.6 Πειραματικά αποτελέσματα

Σε αυτή την ενότητα παρουσιάζουμε αποτελέσματα από την εφαρμογή του αλγορίθμου ITU τόσο σε συνθετικά δεδομένα, όσο και σε δεδομένα από ένα πραγματικό σύνολο δεδομένων από το χώρο της αναζήτησης πληροφορίας [7][134]. Η υλοποίηση του προτινόμενου μοντέλου αναπαράστασης ασαφών σχέσεων, καθώς και του αλγορίθμου ITU, έχουν γίνει με τη χρήση του περιβάλοντος Java, και τα πειράματα έχουν εκτελεστεί σε PC (Centrino 1.6GHz, 256MB RAM) με λειτουργικό σύστημα Windows XP. Ο κώδικας που αντιστοιχεί σε αυτά τα πειράματα είναι ελεύθερα διαθέσιμος [156].

3.6.1 Πειραματικά δεδομένα

Στην αναζήτηση πληροφορίας με βάση τη γνώση, το σύστημα αναζήτησης πρέπει να εκμεταλλεύεται όλη, ή μέρος της γνώσης, όταν επεξεργάζεται για παράδειγμα ένα ερώτημα, ένα προφίλ ή ένα έγγραφο, πριν παραδώσει μια απάντηση. Στις περιπτώσεις που χρησιμοποιείται κάποιο ασαφές σχεσιακό μοντέλο γνώσης, το να διατηρεί κανείς τη γνώση αυτή διαθέσιμη σε μεταβατική μορφή συνήθως απομακρύνει την ανάγκη για αναδρομή, μειώνοντας έτσι σημαντικότατα το χρόνο επεξεργασίας.

Έτσι θα χρησιμοποιήσουμε μια τέτοια σχέση για να δοκιμάσουμε πειραματικά τον αλγόριθμο ITU, καθώς και τον αλγόριθμο ITC που θα παρουσιαστεί στο επόμενο κεφάλαιο. Συγκεκριμένα, το καηολικό σύνολο $S_{90,000}$ είναι το σύνολο των εννοιών. Βασιζόμενοι στην ένα προς ένα σχέση ανάμεσα σε αυτό το σύνολο και στο σύνολο των synsets ρημάτων και ουσιαστικών που έχουν οριστεί στο WordNet για την αγγλική γλώσσα, χρησιμοποιούμε το τελευταίο για την αυτόματα παραγωγή της σχέσης [56]. Η πληθυκότητα του συνόλου ξεπερνά τα 90,000 στοιχεία.

Μια ασαφής σχέση στο $S_{90,000}$ δημιουργείται αυτόματα, και πάλι με βάση το WordNet. Δύο από τις λεξικολογικές σχέσεις(hyponym και part meronym), χρησιμοποιούνται για να προσδιοριστούν τα συνεδεμένα ζεύγη. Καθώς οι σχέσεις στο WordNet δέν έχουν βαθμούς, ο βαθμός 0.9 αποδίδεται σε όλα τα ζεύγη. Επιπρόσθετα, η

σχέση γίνεται ανακλαστική. Συνολικά περίπου 110,000 ζεύγη συνδέονται σε βαθμό 0.9 και 90,000 ακόμε σε βαθμό 1 εξαιτίας της ανακλαστικότητας. Έτσι δημιουργείται η σχέση $R_{90,000}$.

Με τυχαία επιλογή στοιχείων από το σύνολο $S_{90,000}$ διαμορφώνουμε το σύνολο $S_{50,000}$ με 50,000 στοιχεία. Περιορίζοντας τη σχέση $R_{90,000}$ σε αυτό το σύνολο, δηλαδή κρατώντας μόνο τις γραμμές και τις στήλες που αντιστοιχούν σε στοιχεία του $S_{50,000}$, δημιουργούμε τη σχέση $R_{50,000}$. Αυτή είναι όμοια σε δομή και περιεχόμενο με τη σχέση $R_{90,000}$, αλλά μικρότερη σε μέγεθος. Όμοια, η σχέση $S_{20,000}$ δημιουργείται από την $S_{50,000}$, η $S_{10,000}$ από την $S_{20,000}$ κλπ, δημιουργόντας έτσι ένα ευρύ σύνολο: R_n, $n \in \{90000, 50000, 20000, 10000, 5000, 3000, 2000, 1000, 500\}$. Με την εφαρμογή των αλγορίθμων σε δεδομένα με τέτοια σχέση δομής και διαστάσεων μπορούμε να έχουμε μια καλή πειραματική εικόνα της πολυπλοκότητάς τους.

Η ασαφής σχέση R^t_{90000}, που λαμβάνεται με χρήση του αλγορίθμου ITC που θα παρουσιαστεί στο επόμενο κεφάλαιο, είναι το $\sup -t$ μεταβατικό κλείσιμο της $R_{90,000}$, όπου t είναι η t-νόρμα φραγμένης διαφοράς. Η σχέση R^t_{90000} περιέσει περίπου 760,000 μη μηδενικά στοιχεία; 90,000 στοιχεία με βαθμό 1 λόγω ανακλαστικότητας, 110,000 στοιχεία με βαθμό 0.9 που επίσης υπήρχαν στην αρχική $R_{90,000}$ και 560,000 στοιχεία με μικρότερους βαθμούς που προκύπτουν από το μεταβατικό κλείσιμο. Όμοια, έχουμε κατασκευάσει μεταβατικές σχέσεις R^t_n, όλες σύμφωνες με το

$$R^t_n = Tr^t(R_n) \tag{3.21}$$

Η χρήση του πλήρους μοντέλου αναπαράστασης για την $R_{90,000}$ ή και για τις άλλες σχέσεις με σημαντικά μεγάλο πεδίο ορισμού δεν είναι πρακτικά εφικτή. Στην περίπτωση της σχέσης $R_{90,000}$, για παράδειγμα, η διάσταση $90,000 \times 90,000$ απαιτεί χώρο αποθήκευσης για περίπου 8 δισεκατομύρια αριθμούς. Υποθέτοντας αριθμούς διπλής ακρίβειας, με 8 bytes να απαιτούνται για κάθε τέτοιο αριθμό, ο απαιτούμενος χώρος κύριας μνήμης προσεγγίζει τα 65Gb μόνο για ένα αντίγραφο της σχέσης. Έτσι μόνο το αραιό μοντέλο αναπαράστασης χρησιμοποιείται παρακάτω για τα πραγματικά δεδομένα.

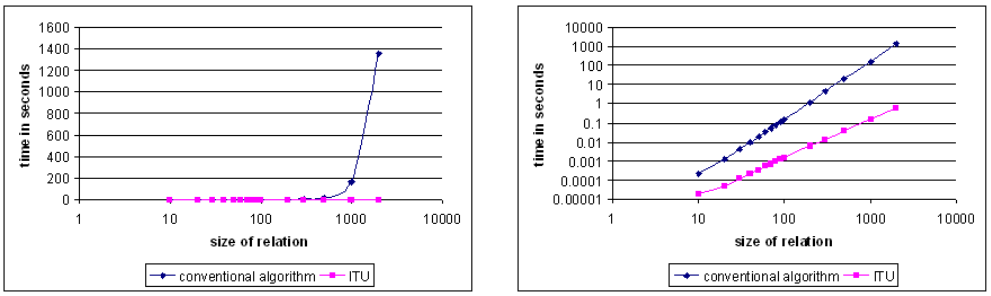
Για να εξεταστεί πειραματικά η απόδοση των προτεινόμενων μεθοδολογιών και στην περίπτωση των πυκνών σχέσεων, καθώς και στην περίπτωση χρήσης του πλήρους μοντέλου αναπαράστασης, δημιουργήσαμε συνθετικά μια κατάλληλη σειρά δεδομένων. Συγκεκριμένα, δημιουργήσαμε σχέσεις διάστασης $n \times n$ για διάφορες τιμές του n και με τυχαίες τιμές για κάθε ζεύγος στοιχείων, και εφαρμόσαμε τον αλγόριθμο μεταβατικού κλεισίματος. Οι διαστάσεις n των σχέσεων είναι $n \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 500, 1000, 2000\}$. Τα ονόματα αυτών των σχέσεων στη συνέχεια θα είναι R^d_n.

3.6.2 Πλήρης αναπαράσταση

Ο αλγόριθμος ITU έχει πολυπλοκότητα $O(n^2)$, σε σχέση με $O(n^3)$ της κλασικής μεθόδου, χρησιμοποιώντας το πλήρες μοντέλο. Για να πετύχουμε την πειραματική επιβεβαίωση αυτού εφαρμόσαμε και τις δύο μεθοδολογίες στο σύνολο R^d_n. Στις περιπτώσεις που το n ήταν πολύ μικρό για να επιτρέπεται ασφαλής μέτρηση του χρόνου, οι πράξεις εκτελέστηκαν 10, 100, 1,000 ή και 1,000,000 φορές, και ο συνολικός χρόνος διαιρέθηκε με το πλήθος των επαναλήψεων. Ο πίνακας 3.6 συνοψίζει τα ευρήματα.

Πίνακας 3.6: *Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και πλήρη αναπαράσταση*

Μέγεθος n	κλασικός αλγίριθμος (2 συνθέσεις)	ITU
10	0.000231s	0.00002s
20	0.001272s	0.00005s
30	0.004196s	0.00013s
40	0.009724s	0.000231s
50	0.018867s	0.00036s
60	0.032887s	0.000531s
70	0.052175s	0.000711s
80	0.077862s	0.000931s
90	0.11046s	0.001201s
100	0.15402s	0.001493s
200	1.20353s	0.00588s
300	4.496s	0.0136s
500	21.29s	0.039297s
1000	164.616s	0.1512s
2000	1353.796s	0.6039s



A B

Σχήμα 3.4: *Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και πλήρη αναπαράσταση*

Πίνακας 3.7: *Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n και αραιή αναπαράσταση.*

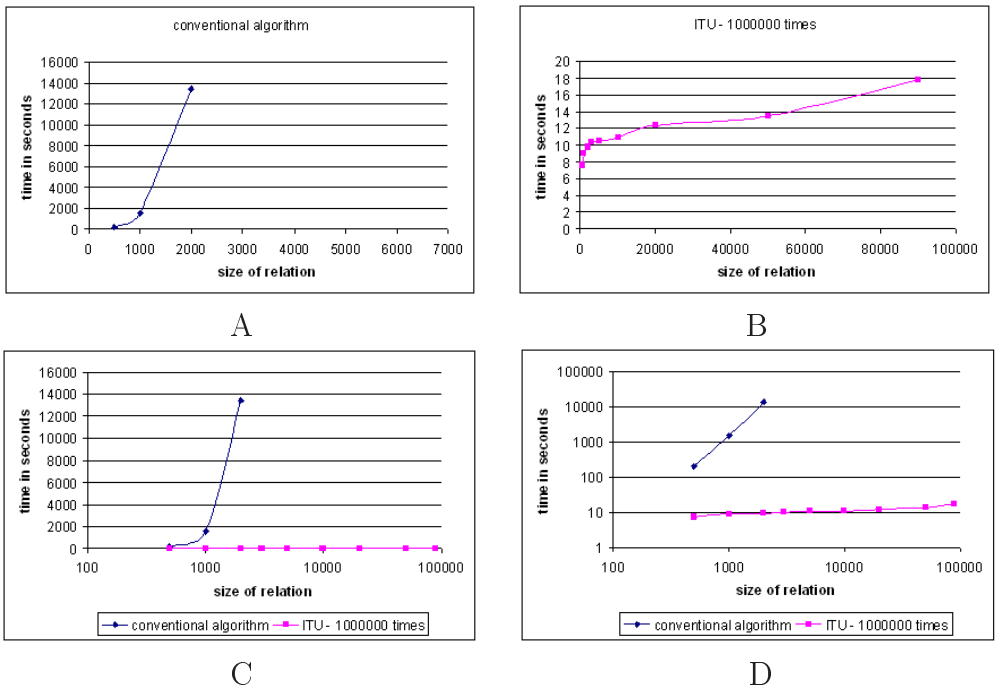
Μέγεθος n	κλασικός αλγόριθμος (2 συνθέσεις)	ITU (1,000,000 επαναλήψεις)
500	201.564s	7.64s
1000	1543.842s	8.95s
2000	13436s	9.74s
3000		10.29s
5000		10.45s
10000		10.95s
20000		12.40s
50000		13.53s
90000		17.80s

Είναι εύκολο να δει κανείς πως ο αλγόριθμος ITU χρειάζεται σημαντικά λιγότερο χρόνο για να πετύχει το ίδιο αποτέλεσμα με την κλασική μεθοδολογία, για κάθε μέγεθος σχέσης. Για παράδειγμα, όταν το μέγεθος της σχέσης είναι 2000×2000 η κλασική μέθοδος χρειάζεται 2,200 φορές περισότερο χρόνο από τον αλγόριθμο ITU για να ανακτήσει τη μεταβατικότητα. Πιο σημαντικά, όπως μπορεί να φανεί πιο καθαρά στο σχήμα 3.4, όπου η κλίση του γραφήματος για τον αλγόριθμο ITU είναι μικρότερη από αυτή για την κλασική μεθοδολογία όταν και οι δύο αποτυπώνονται σε λογαριθμική κλίμακα, ο αλγόριθμος ITU είναι πολύ πιο αποδοτικός από άποψη scaling (Στο σχήμα 3.4.A ο άξονας του χρόνου είναι γραμμικός, ενώ στο σχήμα 3.4.B είναι λογαριθμικός).

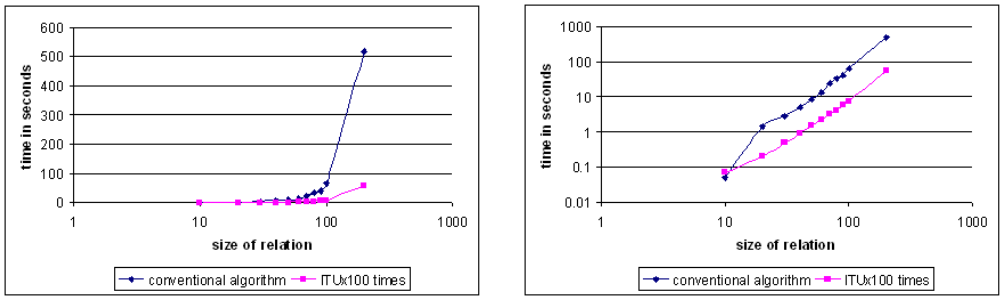
3.6.3 Αραιή σχέση και αραιή αναπαράσταση

Η πολυπλοκότητα του αλγορίθμου ITU είναι κάτω της γραμμικής, σε σύγκριση με $O(n^2 \log n)$ της κλασικής μεθόδου, υποθέτοντας αραιότητα και χρησιμοποιώντας το αραιό μοντέλο αναπαράστασης. Για να το επιβεβαιώσουμε πειραματικά εφαρμόζουμε και τις δύο μεθόδους στο σύνολο R_n. Τα αποτελέσματα συνοψίζονται στον πίνακα 3.7. Σε ορισμένες περιπτώσεις ο χρόνος εκτέλεσης ήταν απαγορευτικά μεγάλος και το πείραμα δεν μπορούσε να ολοκληρωθεί. Έτσι, κάποιες τιμές λείπουν από τον πίνακα. Το σχήμα 3.5 παρουσιάζει τα αποτελέσματα γραφικά. Τόσο στον πίνακα όσο και στο σχήμα οι χρόνοι που αναφέρονται για τον αλγόριθμο ITU αντιστοιχούν σε 1,000,000 επαναλήψεις, ενώ οι χρόνοι που αναφέρονται για την κλασική μέθοδο αντιστοιχούν σε μία μόνο επανάληψη.

Στο σχήμα 3.5.A παρατηρούμε πως η θεωρητική πολυπλοκότητα $O(n^2 \log n)$ επιβεβαιώνεται πειραματικά. Όμοια, το σχήμα 3.5.B επιβεβαιώνει την μικρότερη από γραμμική πολυπλοκότητα του αλγορίθμου ITU. Τα σχήματα 3.5.C και 3.5.D παρουσιάζουν μετρήσεις και από τα δύο πειράματα για να διευκολύνουν τη συγκριτική μελέτη. Μπορούμε εύκολα να διαπιστώσουμε την ανωτερότητα του αλγορίθμου ITU, τόσο σε χρόνο εκτέλεσης, όσο και σε scaling ως προς το μέγεθος της σχέσης.



Σχήμα 3.5: *Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n και αραιή αναπαράσταση.*



Σχήμα 3.6: *Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και αραιή αναπαράσταση.*

3.6.4 Πυκνή σχέση και αραιή αναπαράσταση

Ο αλγόριθμος ITU έχει πολυπλοκότητα $n^2 \log n$, σε σχέση με $O(n^3)$ της κλασικής μεθόδου, όταν η σχέση είναι πυκνή και η δομή δεδομένων το προτεινόμενο μοντέλο αραιής αναπαράστασης. Για πειραματική επιβεβαίωση χρησιμοποιούμε το σύνολο R_n^d. Ο πίνακας 3.8 συνοψίζει τα αποτελέσματα που παρουσιάζονται γραφικά και στο σχήμα 3.6. Αξίζει να παρατηρήσει κανείς πως τόσο στον πίνακα όσο και στο σχήμα οι τιμές που αναφέρονται για τον ανγόριθμο ITU αντιστοιχούν σε 100 επαναλήψεις, ενώ αυτές που αναφέρονται στην κλασική μέθοδο μόνο σε μία.

Για μια ακόμη φορά, η ανωτερότητα του αλγορίθμου ITU επιβεβαιώνεται και πειραματικά με εμφατικό τρόπο.

□

Πίνακας 3.8: *Χρόνοι εκτέλεσης για 2 συνθέσεις και για τον ITU με είσοδο R_n^d και αραιή αναπαράσταση.*

Μέγεθος n	κλασικός αλγόριθμος (2 συνθέσεις)	ITU (100 επαναλήψεις)
10	0.05s	0.07s
20	1.462s	0.211s
30	2.854s	0.481s
40	5.067s	0.911s
50	8.292s	1.522s
60	12.948s	2.283s
70	24.425s	3.205s
80	33.998s	4.286s
90	40.438s	5.618s
100	64.663s	7.181s
200	516.473s	55.83s

Κεφάλαιο 4

sup $-t$ μεταβατικό κλείσιμο ασαφών σχέσεων

4.1 Εισαγωγή

Αυτό το κεφάλαιο αποτελεί στην ουσία άμεση συνέχεια του προηγούμενου κεφαλαίου. Στο κεφάλαιο 3 είδαμε πώς μια αραιή σχέση μπορεί να αποθεκευτεί σε ένα ζεύγος δέντρων AVL ώστε να είναι δυνατή η γρήγορη πρόσβαση στα στοιχεία της, καθώς και πώς μπορεί να αποκατασταθεί η μεταβατικότητα μιας σχέσης όταν αυτή διαταραχτεί τοπικά. Σε αυτό το κεφάλαιο, χρησιμοποιούμε σαν βάση τόσο το αραιό μοντέλο αναπαράστασης, όσο και τον αλγόριθμο ITU, και αναπτύσσουμε έναν ιδαιέτερα αποδοτικό αλγόριθμο για το sup $-t$ μεταβατικό κλεισιμο ασαφών σχέσεων.

4.2 Αλγόριθμος μεταβατικού κλεισίματος ITC

Ο αλγόριθμος της σταδιακής ενημέρωσης ασαφών μεταβατικών σχέσεων εύκολα οδηγεί στο σχεδιασμό ενός αλγορίθμου και για τον συνολικό υπολογισμό του μεταβατικού κλεισίματος της σχέσης r. Τα βήματα του αλγορίθμου ακολουθούν:

1. Δημιουργούμε μια κενή σχέση r'.

2. Για κάθε μη μηδενικό στοιχείο r_{ij} της αρχικής σχέσης r:

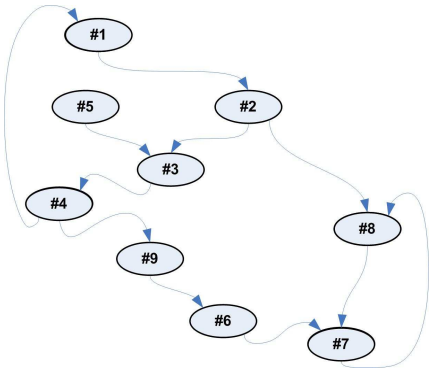
   (α) Θέτουμε

   $$r'(i,j) \leftarrow r'(i,j) \cup r(i,j) \tag{4.1}$$

   (β) Εκτελούμε τον αλγόριθμο σταδιακής ενημέρωσης για τη σχέση r' με παραμέτρους i και j

   $$r' \leftarrow ITU(r', i, j) \tag{4.2}$$

όταν ο αλγόριθμος τερματίζει έχουμε $Tr^t(r) = r'$.



Σχήμα 4.1: *The sample fuzzy relation*

Θεώρημα

Η υπολογιστική πολυπλοκότητα του αλγορίθμου συνολικού μεταβατικού κλεισίματος με χρήση της σταδιακής ενημέρωσης είναι $O(n^4 logn)$ στη χειρότερη περίπτωση και $O(nlog^4 n)$ στην τυπική περίπτωση.

Απόδειξη

Το βήμα 1 ολοκληρώνεται με μια διαδικασία πολυπλοκότητας $O(1)$. Το βήμα 2α έχει πολυπλοκότητα $O(logn)$, ενώ η πολυπλοκότητα του βήματος 2β δίνεται από το προηγούμενο κεφάλαιο. Το βήμα 2 εκτελείται τόσες φορές, όσα και τα μη μηδενικά στοιχεία της σχέσης r.

Στη χειρότερη περίπτωση η σχέση r είναι πλήρης, έχει δηλαδή n^2 στοιχεία, και το βήμα 2β έχει πολυπλοκότητα $O(n^2 logn)$, οπότε έχουμε συνολικά

$$O(1) + O(n^2) \cdot (O(logn) + O(n^2 logn)) = O(n^4 logn) \tag{4.3}$$

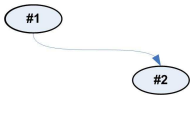
Στην τυπική περίπτωση έχουμε $O(n)$ μη μηδενικές γραμμές στη σχέση r, και καθεμιά από αυτές περιέχει $O(logn)$ μη μηδενικά στοιχεία, οπότε συνολικά $O(nlogn)$ μη μηδενικά στοιχεία στη σχέση r, και η πολυπλοκότητα του βήματος 2β είναι $O(log^3 n)$. Έτσι, έχουμε συνολικά

$$O(1) + O(nlogn) \cdot (O(logn) + O(log^3 n)) = O(nlog^4 n) \tag{4.4}$$

✓

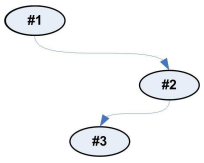
4.3 Αριθμητικό παράδειγμα

Σε αυτή την ενότητα εξηγούμε τη λειτουργία του αλγορίθμου ITC μέσω μιας βήμα-βήμα παρουσίασης. Η σχέση στην οποία θα εφαρμοστεί ο αλγόριθμος παρουσιάζεται στο σχήμα 4.1, όπου όλοι οι σύνδεσμοι που παρουσιάζονται με τη μορφή γραμμών θεωρείται πως έχουν βάρος 0.95, ενώ όλοι οι άλλοι σύνδεσμοι έχουν βάρος 0. Αξίζει να σημειωθεί πως στη σχέση υπάρχουν κύκλοι ($\#1 \rightarrow \#2 \rightarrow \#3 \rightarrow \#4 \rightarrow \#1$ όπως και $\#7 \rightarrow \#8 \rightarrow \#1$). Επιπλέον, η σχέση δεν είναι ανακλαστική, κάτι που είναι απαγορευτικό για τους περισσότερους γνωστούς αλγορίθμους μεταβατικού κλεισίματος. Τέλος, η t-νόρμα της φραγμένης διαφοράς επιλέγεται για την ιδιότητα της μεταβατικότητας.



	#2
#1	0.95

Σχήμα 4.2: *Προσθήκη στοιχείου (#1,#2,0.95)*



	#2	#3
#1	0.95	0.90
#2		0.95

Σχήμα 4.3: *Προσθήκη στοιχείου (#2,#3,0.95)*

Σύμφωνα με τον αλγόριθμο, για να υπολογίσουμε το μεταβατικό κλείσιμο της σχέσης ξεκινάμε από μια κενή σχέση την οποία συμπηρώνουμε σταδιακά χρησιμοποιώντας τον ελγόριθμο ITU. Στο πρώτο βήμα προσθέτουμε το σύνδεσμο ανάμεσα στα στοιχεία #1 and #2 (σχήμα 4.2). Στα σχήματα της ενότητας, στα αριστερά παρουσιάζουμε τους συνδέσμους που έχει ήδη επεξεργαστεί ο αλγόριθμος ITC, ενώ στα δεξιά την προσωρινή σχέση r'. Όταν ο αλγόριθμος τερματίσει η σχέση r' θα είναι το μεταβατικό κλείσιμο της αρχικής.

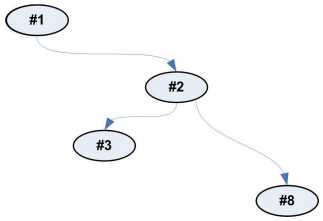
Στο επόμενο βήμα προστίθεται ο σύνδεσμος ανάμεσα στα #2 and #3 (σχήμα 4.3), ενώ ο ITU προσθέτει επίσης το σύνδεσμο ανάμεσα στα #1 και #3. Ο αναγνώστης προσκαλείται να παρακολουθήσει την εφαρμογή του ITC διατρέχοντας τα σχήματα 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11 και 4.12. Άξιο αναφοράς είναι το βήμα στο οποίο προστίθεται ο σύνδεσμος ανάμεσα στα στοιχεία #4 και #1, όπου βλέπουμε πως ο αλγόριθμος χειρίζεται με επιτυχία τον κύκλο που δημιουργείται.

4.4 Συγκριτική μελέτη

Ο πίνακας 4.1 συνοψίζει τις υπολογιστικές πολυπλοκότητες των αλγορίθμων μεταβατικού κλεισίματος. Στη χειρότερη περίπτωση η κλασική μεθοδολογία υπερτερεί, έχοντας πολυπλοκότητα $O(n^3 logn)$, σε σχέση με $O(n^4)$ της προτεινόμενης μεθόδου. Στην τυπική περίπτωση, όμως, η προτεινόμενη μεθοδολογία υπερτερεί κατά πολύ, καθώς η πολυπλοκότητά της $(O(nlog^3 n))$ είναι μικρότερη από τετραγωνική. Το να αποδείξει κανείς πως

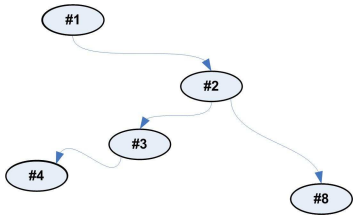
$$O(n^2 log^2 n) > O(n^2) > O(nlog^3 n) \tag{4.5}$$

είναι στοιχειώδες.



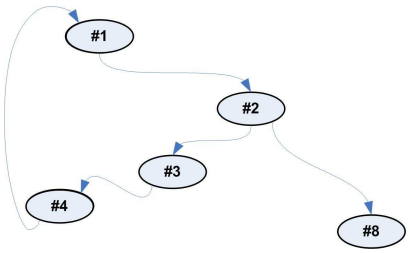
	#2	#3	#8
#1	0.95	0.90	0.90
#2		0.95	0.95

Σχήμα 4.4: *Προσθήκη στοιχείου (#2,#8,0.95)*



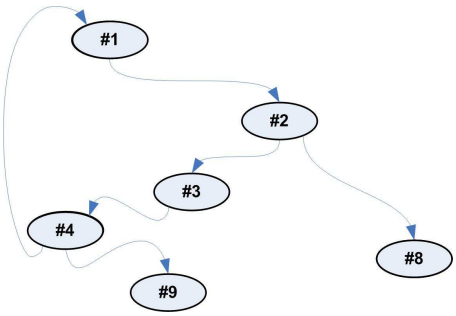
	#2	#3	#4	#8
#1	0.95	0.90	0.85	0.90
#2		0.95	0.90	0.95
#3			0.90	

Σχήμα 4.5: *Προσθήκη στοιχείου (#3,#4,0.95)*



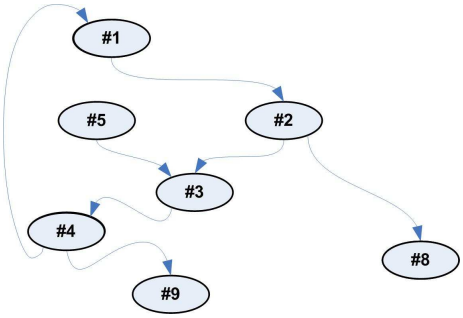
	#1	#2	#3	#4	#8
#1	0.80	0.95	0.90	0.85	0.90
#2	0.85	0.80	0.95	0.90	0.95
#3	0.90	0.85	0.80	0.95	0.80
#4	0.95	0.90	0.85	0.80	0.85

Σχήμα 4.6: *Προσθήκη στοιχείου (#4,#1,0.95)*



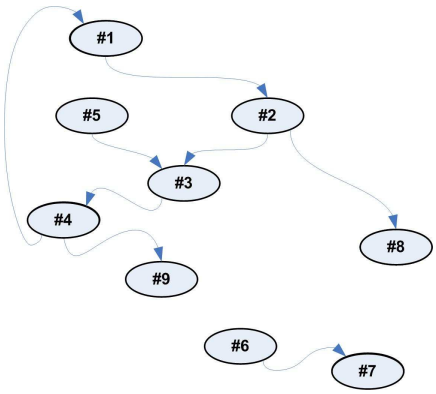
	#1	#2	#3	#4	#8	#9
#1	0.80	0.95	0.90	0.85	0.90	0.80
#2	0.85	0.80	0.95	0.90	0.95	0.85
#3	0.90	0.85	0.80	0.95	0.80	0.90
#4	0.95	0.90	0.85	0.80	0.85	0.95

Σχήμα 4.7: *Προσθήκη στοιχείου (#4,#9,0.95)*



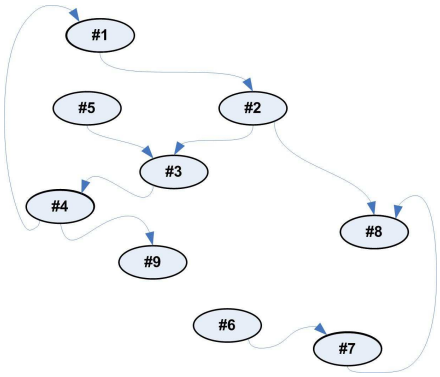
	#1	#2	#3	#4	#8	#9
#1	0.80	0.95	0.90	0.85	0.90	0.80
#2	0.85	0.80	0.95	0.90	0.95	0.85
#3	0.90	0.85	0.80	0.95	0.80	0.90
#4	0.95	0.90	0.85	0.80	0.85	0.95
#5	0.85	0.80	0.95	0.90	0.75	0.85

Σχήμα 4.8: *Προσθήκη στοιχείου (#5,#3,0.95)*



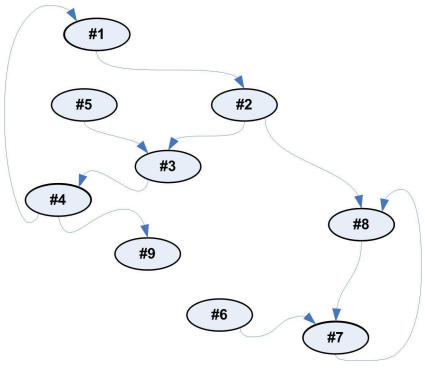
	#1	#2	#3	#4	#7	#8	#9
#1	0.80	0.95	0.90	0.85		0.90	0.80
#2	0.85	0.80	0.95	0.90		0.95	0.85
#3	0.90	0.85	0.80	0.95		0.80	0.90
#4	0.95	0.90	0.85	0.80		0.85	0.95
#5	0.85	0.80	0.95	0.90		0.75	0.85
#6					0.95		

Σχήμα 4.9: *Προσθήκη στοιχείου (#6,#7,0.95)*



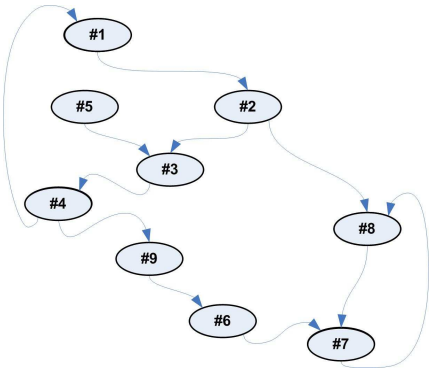
	#1	#2	#3	#4	#7	#8	#9
#1	0.80	0.95	0.90	0.85		0.90	0.80
#2	0.85	0.80	0.95	0.90		0.95	0.85
#3	0.90	0.85	0.80	0.95		0.80	0.90
#4	0.95	0.90	0.85	0.80		0.85	0.95
#5	0.85	0.80	0.95	0.90		0.75	0.85
#6					0.95	0.90	
#7						0.95	

Σχήμα 4.10: *Προσθήκη στοιχείου (#7,#8,0.95)*



	#1	#2	#3	#4	#7	#8	#9
#1	0.80	0.95	0.90	0.85	0.85	0.90	0.80
#2	0.85	0.80	0.95	0.90	0.90	0.95	0.85
#3	0.90	0.85	0.80	0.95	0.75	0.80	0.90
#4	0.95	0.90	0.85	0.80	0.80	0.85	0.95
#5	0.85	0.80	0.95	0.90	0.70	0.75	0.85
#6					0.95	0.90	
#7					0.90	0.95	
#8					0.95	0.90	

Σχήμα 4.11: *Προσθήκη στοιχείου (#8,#7,0.95)*



	#1	#2	#3	#4	#6	#7	#8	#9
#1	0.80	0.95	0.90	0.85	0.75	0.85	0.90	0.80
#2	0.85	0.80	0.95	0.90	0.80	0.90	0.95	0.85
#3	0.90	0.85	0.80	0.95	0.85	0.80	0.80	0.90
#4	0.95	0.90	0.85	0.80	0.90	0.85	0.85	0.95
#5	0.85	0.80	0.95	0.90	0.80	0.75	0.75	0.85
#6						0.95	0.90	
#7						0.90	0.95	
#8						0.95	0.90	
#9					0.95	0.90	0.80	

Σχήμα 4.12: *Προσθήκη στοιχείου (#9,#6,0.95)*

Πίνακας 4.1: *Σύνοψη υπολογιστικής πολυπλοκότητας αλγορίθμων μεταβατικού κλεισίματος*

Αλγόριθμος	Μοντέλο	Αραιή σχέση	Πυκνή σχέση
Κλασικός	Πλήρες	$n^3 \log n$	$n^3 \log n$
Κλασικός	Αραιό	$n^2 \log^2 n$	$n^3 \log n$
ITC	Πλήρες	-	-
ITC	Αραιό	$n \log^4 n$	$n^4 \log n$

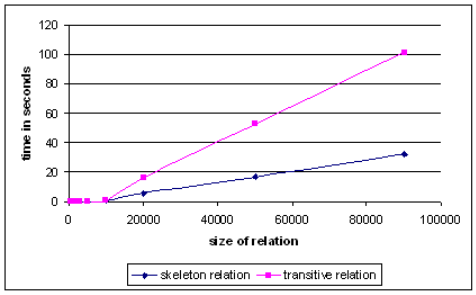
Έτσι, η σύγκριση των δύο μεθόδων εξαρτάται από την εγκυρότητα της υπόθεσης που κάναμε σχετικά με τα χαρακτηριστικά της τυπικής περίπτωσης αραιής σχέσης. Η βασική υπόθεση είναι πως πολύ μικρότερος αριθμός μη μηδενικών στοιχείων, και συγκεκριμένα $O(nlogn)$ αντί για $O(n^2)$ περιέχονται στη σχέση. Αυτή η υπόθεση όχι μόνο επιβεβαιώνεται από τις υπάρχουσες οντολογικές σχέσεις, που θα είναι και το πεδίο εφαρμογής της μεθοδολογίας, αλλά και δεν θα μπορούσε να μην ισχύει: αν η σχέση r περιείχε $O(n^2)$ μη μηδενικά στοιχεία, τότε θα ήταν αδύνατο να αποθηκευτεί στην κεντρική μνήμη κάποιου υπολογιστή, καθώς το n είναι πολύ μεγάλος αριθμός.

Όσο αφορά στις απαιτήσεις χώρου, η κλασσική προσέγγιση απαιτεί την ταυτόχρονη ύπαρξη δύο αντιγράφων της σχέσης στη μνήμη, διπλασιάζοντας έτσι τις ανάγκες χώρου. Από την άλλη, η προτεινόμενη μεθοδολογία χρειάζεται χώρο μόνο για ένα αντίγραφο της σχέσης. Στην πραγματικότητα δύο σχέσεις αποθηκεύονται, αλλά όπως προστίθενται στοιχεία στη μια αφαιρούνται από την άλλη.

Εκτός από τα προτερήματα σε σχέση με τους απαιτούμενους πόρους για την εκτέλεση, η προτεινόμενη μέθοδος μεταβατικού κλεισίματος έχει και τα ακόλουθα πλεονεκτήματα:

- Στην κλασική προσέγγιση χρειάζεται μια πλήρης σύνθεση, δηλαδή μια πράξη πολυπλοκότητας $O(n^2logn)$ στην τυπική περίπτωση, για να διαπιστωθεί πως ο αλγόριθμος θα μπορούσε να είχε τερματίσει. Αυτό ισχύει ακόμη και στην περίπτωση που η σχέση είναι αρχικά μεταβατική. Στην προτεινόμενη προσέγγιση και για την ίδια περίπτωση ο αλγόριθμος θα είχε πολυπλοκότητα $O(nlog^4n)$.

- Στην κλασική προσέγγιση η σχέση δεν είναι μεταβατική παρά μόνο όταν ο



Σχήμα 4.13: *Χρόνοι εκτέλεσης για τον ITC με είσοδο R_n και R_n^t.*

αλγόριθμος έχει τερματίσει. Έτσι, στην περίπτωση που το n είναι μεγάλο, χρειάζεται πολύς χρόνος πριν η σχέση γίνει αξιοποιήσιμη από αλγορίθμους που υποθέτουν μεταβατικότητα. Αντίθετα, η σχέση r' είναι μεταβατική μετά από κάθε επανάληψη, οπότε οι σχετικοί αλγόριθμοι μπορούν να εφαρμόζονται σε αυτή πριν την ολοκλήρωση της διαδικασίας.

- Εξαιτίας της αναδρομικής μορφής της κλασικής προσέγγισης, το σημείο τερματισμού εξαρτάται βαθμό από το πλήθος των κόμβων στο μέγιστο μονοπάτι που εμφανίζεται στο γράφο της σχέσης. Αντίθετα, η προτεινόμενη μέθοδος εξασφαλίζει τη μεταβατικότητα σε ένα πέρασμα και το πλήθος των βημάτων που θα χρειαστούν είναι γνωστό εκ των προτέρων. Έτσι, είναι δυνατό εάν το επιθυμούμε να έχουμε ενημέρωση για την πρόοδο της διαδικασίας στη μορφή ποσοστού ολοκλήρωσης, εάν το επιθυμούμε.

- Η προτεινόμενη μεθοδολογία δεν επηρεάζεται από την ύπαρξη κυκλικών γράφων. Αυτό το χαρακτηριστικό δεν το μοιράζονται όλες οι μεθοδολογίες μεταβατικού κλεισίματος [126].

- Η πολυπλοκότητα εκτέλεσης του αλγορίθμου ITC είναι πολύ κοντά στην πολυπλοκότητα φόρτωσης της σχέσης από τη δευτερεύουσα στην κύρια μνήμη ($O(nlog^4n)$ σε σχέση με $O(nlogn)$). Έτσι, μας δίνει την επιλογή να αποθηκεύουμε τη σχέση στην μη μεταβατική μορφή της (απαιτώντας λιγότερο χρόνο) και να επιτυγχάνουμε τη μεταβατικότητα κατά τη μεταφορά της στην κύρια μνήμη. Η δομή του αλγορίθμου είναι μάλιστα τέτοια που επιτρέπει την on-line εκτέλεση του αλγορίθμου κατά τη διαδικασία της φόρτωσης.

4.5 Πειραματικά αποτελέσματα

Για την πειραματική επιβεβαίωση των θεωρητικών αποτελεσμάτων σε σχέση με τον αλγόριθμο ITC χρησιμοποιούμε τα ίδια πειραματικά δεδομένα, συνθετικά και πραγματικά, όπως και στην περίπτωση του αλγορίθμου ITU. Η παρουσίαση των πειραματικών δεδομένων βρίσκεται στην ενότητα 3.6.1.

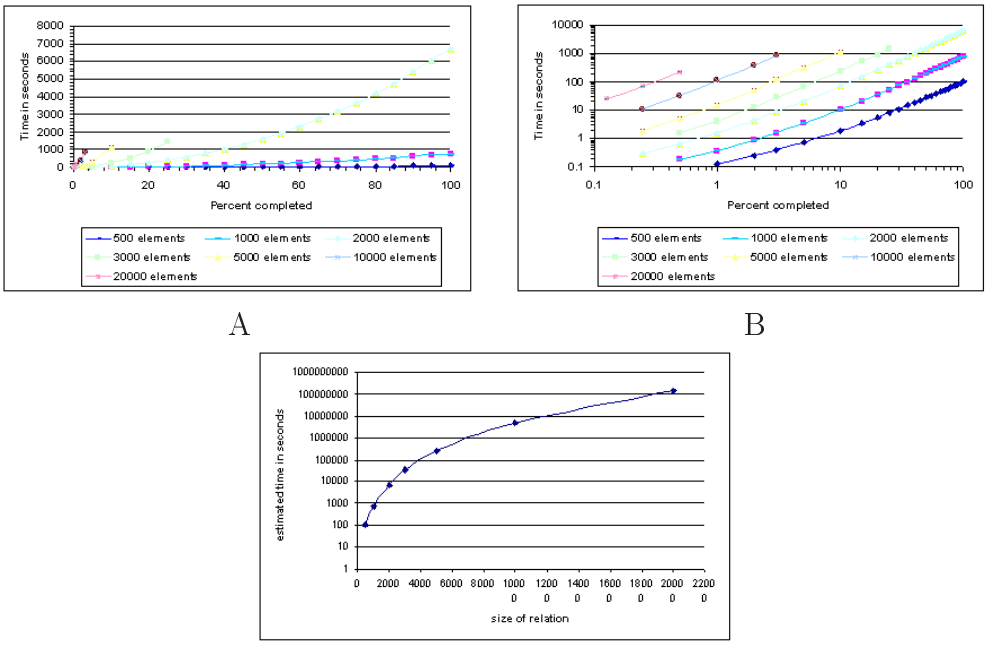
Ο αλγόριθμος ITC έχει πολυπλοκότητα $n \log^4 n$, σε σχέση με $n^2 \log^2 n$ της κλασικής μεθόδου, όταν αναφερόμαστε σε αραιές σχέσεις και χρησιμοποιούμε αραιό μοντέλο αναπαράστασης. Για να επιβεβαιώσουμε την απόδοση του αλγορίθμου πειραματικά τον εφαρμόζουμε στα δεδομένα R_n και R_n^t. Τα αποτελέσματα παρουσιάζονται στον πίνακα 4.2 και στο σχήμα 4.13.

Πίνακας 4.2: *Χρόνοι εκτέλεσης για τον ITC με είσοδο R_n και R_n^t.*

Μέγεθος n	R_n δεδομένα	R_n^t δεδομένα
500	0.03s	0.03s
1000	0.05s	0.04s
2000	0.07s	0.06s
3000	0.101s	0.1s
5000	0.17s	0.16s
10000	0.591s	0.761s
20000	5.518s	15.832s
50000	16.473s	52.745s
90000	32.106s	101.396s

Πίνακας 4.3: *Χρόνοι εκτέλεσης σύνθεσης με είσοδο R_{90000}.*

Γραμμή	χρόνος για αυτό το βήμα	συνολικός χρόνος
1	7.262s	7.262s
2	7.461s	14.723s
3	7.661s	22.384
4	7.871s	30.255s
5	8.202s	38.457s
6	8.182s	46.639s
7	8.301s	54.94s
8	9.134s	64.074s



A B C

Σχήμα 4.14: *Χρόνοι εκτέλεσης σύνθεσης με είσοδο R_{90000}.*

Μπορούμε να δούμε πως ο αλγόριθμος εξελίσσεται σχεδόν γραμμικά και για τα δύο σετ δεδομένων. Επιπρόσθετα, η πυκνότητα της σχέσης έχει επίδραση στο συνολικό χρόνο εκτέλεσης. Ωστόσο, όπως φαίνεται και στο σχήμα, αυτό επιδρά στην κλίση του γραφήματος αλλά όχι και στη μορφή του. Με άλλα λόγια, η πολυπλοκότητα παραμένει κοντά στη γραμμική και σαφώς κάτω από $O(n^2)$ όταν εφαρμόζεται σε αραιές σχέσεις που είναι ήδη μεταβατικές και έτσι περιέχουν περισσότερα μη μηδενικά στοιχεία.

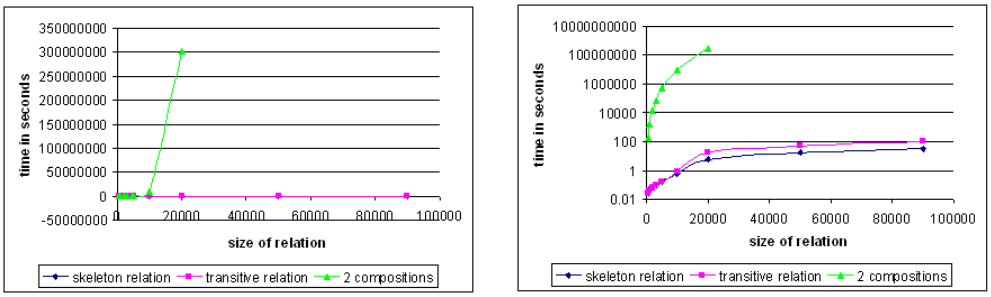
Αυτό που είναι πιο σημαντικό είναι να συγκρίνει κανείς την απόδοση του αλγορίθμου ITC με αυτή της κλασικής διαδικασίας. Όπως φάνηκε καθαρά και από τα στοιχεία που έλειπαν από τον πίνακα 3.7 και τη μορφή του γραφήματος στο σχήμα 3.5, ο χρόνος εκτέλεσης ακόμη και μίας μόνο σύνθεσης είναι απαγορευτικός για την εφαρμογή της κλασικής μεθοδολογίας στα σετ δεδομένων R_n και R_n^t. Έτσι έχουμε προβεί σε εκτιμήσεις του χρόνου που θα χρειαζόταν για την εφαρμογή της κλασικής μεθοδολογίας στο σύνολο R_n:

Αναπτύξαμε το λογισμικό που είναι υπεύθυνο για τη σύνθεση έτσι ώστε να δείχνει σε κάθε στιγμή το ποσοστό ολοκλήρωσης της πράξης καθώς και τον χρόνο που έχει περάσει. Η έξοδος των πρώτων βημάτων της σύνθεσης της σχέσης με τον εαυτό της από την εφαρμογή στα δεδομένα R_{90000} παρουσιάζεται στον πίνακα 4.3.

Το σχήμα 4.14.A παρουσιάζει τους χρόνους γραφικά. Εκεί βλέπουμε καθαρά πως ο αλγόριθμος γίνεται πιο αργός όσο η διαδικασία προχωρά και η σχέση εξόδου σταδιακά μεγαλώνει. Σχεδιάζοντας τα ίδια σημεία σε λογαριθμική κλίμακα, όπως στο σχήμα 4.14.B, αποκτούμε πολύ πιο απλές καμπύλες, με βάση τις οποίες μπορούμε να εκτιμήσουμε χονδρικά πιθανούς χρόνους εκτέλεσης για το 100% της σύνθεσης με έναν οπτικό τρόπο. Ακολουθώντας την ίδια προσέγγιση για διάφορες τιμές του n προσέγγιση καταλήγουμε στις εκτιμήσεις που παρουσιάζονται γραφικά στο σχήμα 4.14.C. Σε αυτό το σχήμα τα πρώτα τρία σημεία είναι πραγματικές μετρημένες τιμές χρόνου και τα υπόλοιπα είναι εκτιμήσεις. Οι πρώτες εκτιμήσεις έγιναν με μεγαλύτερο μέρος της διαδικασίας ολοκληρωμένο, και έτσι με μεγαλύτερη ακρίβεια και βεβαιότητα, ενώ οι τελευταίες με πολύ μικρότερο μέρος ολοκληρωμένο και έτσι πιο ασαφώς. Για να αντισταθμιστεί η έλλειψη ακρίβειας για μεγαλύτερες σχέσεις οι εκτιμήσεις για μεγάλες τιμές του n είναι πιο συντηρητικές. Μια παρατήρηση που εύκολα μπορεί να γίνει με βάση σχήμα είναι πως η πολυπλοκότητα της σύνθεσης απέχει πολύ από τη γραμμική.

Βέβαια, ακόμη και αν υποθέσουμε πως οι εκτιμώμενοι χρόνοι εκτέλεσης για μια σύνθεση είναι σωστοί, και πάλι δεν θα είμαστε σε θέση να υπολογίσουμε το συνολικό χρόνο για το μεταβατικό κλείσιμο χρησιμοποιώντας την κλασική μέθοδο, καθώς αυτός εξαρτάται από το "βάθος" της σχέσης. Για τη σχέση R_{90000}, για παράδειγμα, σύμφωνα με τη θεωρία η διαδικασία θα μπορούσε να χρειάζεται από 2 έως 17 συνθέσεις. Για να γίνουν κάποιες συγκρίσεις ακολουθούμε το ιδανικό σενάριο για την κλασική μέθοδο υποθέτοντας πως μόνο δύο συνθέσεις αρκούν για την επίτευξη της μεταβατικότητας. Το σχήμα 4.15 παρουσιάζει τη σύγκριση με βάση τις υποθέσεις που έχουμε κάνει. Μπορούμε εύκολα να δούμε, όπως και ήταν αναμενόμενο, πως η διαφορά ανάμεσα στις δύο μεθοδολογίες είναι τεράστια, τόσο όσον αφορά στους χρόνους εκτέλεσης για ίδιες διαστάσεις σχέσεων, όσο και όσον αφορά στον τρόπο που αυτοί οι χρόνοι εξελίσσονται όταν το μέγεθος των σχέσεων μεγαλώνει.

□



A B

Σχήμα 4.15: *Χρόνοι εκτέλεσης του ITC με είσοδο R_n και R_n^t.*

Κεφάλαιο 5

Συστήματα ανάκτησης πληροφορίας και επέκταση ερωτήματος

5.1 Εισαγωγή

Τα συστήματα ανάκτησης πληροφορίας επιχειρούν να εντοπίζουν εκείνα τα στοιχεία από το διαθέσιμο υλικό που ικανοποιούν τις επιθυμίες του χρήστη. Η διαδικασία αυτή έχει διάφορα εγγενή προβλήματα που εστιάζονται συνήθως στην αδυναμία πλήρους περιγραφής των επιθυμιών του χρήστη και του περιεχομένου των διαθεσίμων στοιχείων. Η έρευνα που αφορά τα συστήματα ανάκτησης πληροφορίας επιχειρεί να λύσει αυτά τα προβλήματα με τη χρήση πιο αλληλεπιδραστικών μεθόδων οργάνωσης και αναζήτησης και με την εκμετάλλευση προϋπάρχουσας γνώσης.

5.2 Συστήματα ανάκτησης πληροφορίας

5.2.1 Δομή συστημάτων ανάκτησης πληροφορίας

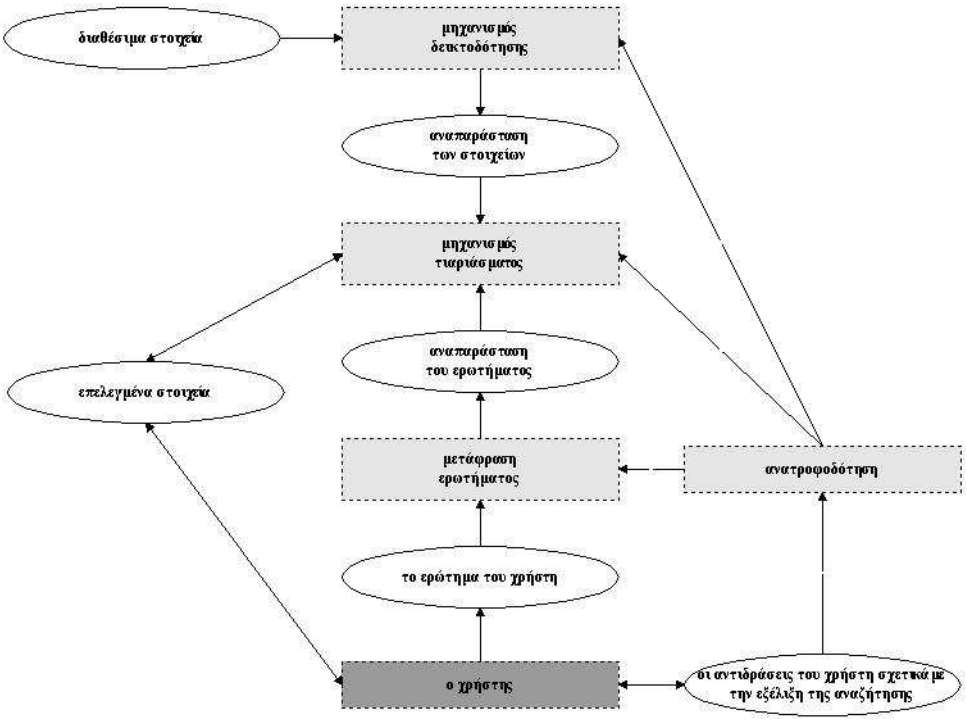
Δεδομένου ενός συνόλου διαθεσίμων στοιχείων, ένα σύστημα ανάκτησης πληροφορίας αναλαμβάνει την αποθήκευση των στοιχείων αυτών με τη χρήση κατάλληλης αναπαράστασής τους και την πρόσβαση σε αυτά. Αναλαμβάνει επίσης την επαφή με τους τελικούς χρήστες, την εξαγωγή πληροφορίας σχετικά με τις επιθυμίες τους και την ανάκληση των στοιχείων που ικανοποιούν τις επιθυμίες αυτές.

Η γενική δομή ενός συστήματος ανάκτησης πληροφορίας παρουσιάζεται στο Σχήμα 5.1. Στο σχήμα φαίνεται πως μέσω της διαδικασίας δεικτοδότησης κάθε στοιχείο $\tilde{d}$ του συνόλου $\tilde{D} = \{\tilde{d_1}, \tilde{d_2}, \tilde{d_3}, ..\}$ διαθεσίμων στοιχείων αντιστοιχίζεται σε κάποιο στοιχείο d του συνόλου $D = \{d_1, d_2, d_3, ..\}$ των πιθανών αναπαραστάσεων. Η αντιστοίχιση είναι μια συνάρτηση f_1 ορισμένη στο σύνολο $\tilde{D}$, η οποία δεν είναι απαραίτητα ένα προς ένα. Πιο αυστηρά

$$f_1 : \tilde{D} \rightarrow D \tag{5.1}$$

Τα διαθέσιμα στοιχεία μπορεί να έχουν οποιαδήποτε μορφή, ηλεκτρονική ή μη, ενώ η διαδικασία δεικτοδότησης μπορεί να γίνεται αυτόματα από το IRS ή με τη συμβολή χρηστών που θεωρούνται κατάλληλοι για τη δεικτοδότηση των συγκεκριμένων κειμένων. Οι χρήστες αυτοί αναφέρονται ως έμπειροι χρήστες.

Αντίστοιχα, η διαδικασία της μετάφρασης του ερωτήματος του χρήστη αναπαριστά την επιθυμία του χρήστη σε μορφή συμβατή με το σύστημα ανάκτησης πληροφορίας.



Σχήμα 5.1: *Η γενική μορφή ενός συστήματος αναζήτησης πληροφορίας.*

Έτσι το ερώτημα $\tilde{q}$ του χρήστη αντιστοιχίζεται μέσω της συνάρτησης f_2 στο στοιχείο q του συνόλου $Q = \{q_1, q_2, q_3, ..\}$ των δυνατών αναπαραστάσεων του ερωτήματος του χρήστη. Πιο αυστηρά

$$f_2 : \tilde{Q} \rightarrow Q \tag{5.2}$$

Είναι προφανές πως ούτε η μετάφραση του ερωτήματος είναι απαραίτητα σχέση ένα προς ένα.

Ο μηχανισμός ταιριάσματος επιλέγει από τα διαθέσιμα στοιχεία ένα υποσύνολο $\tilde{D}_{\alpha\pi} \subseteq D$ που περιέχει τα στοιχεία που ταιριάζουν περισσότερο με το ερώτημα $\tilde{q}$ του χρήστη. Ανάλογα με την περίπτωση, η επιλογή μπορεί να είναι δυαδική ή να γίνεται με τη βοήθεια βαθμών ταιριάσματος των αναπαραστάσεων των διαθεσίμων στοιχείων με την αναπαράσταση του ερωτήματος.

Για το σκοπό αυτό ο μηχανισμός ταιριάσματος επεξεργάζεται την αναπαράσταση q του ερωτήματος και τις αναπαραστάσεις D των διαθεσίμων στοιχείων και όχι τα ίδια τα $\tilde{q}$ και $\tilde{D}$. Έτσι, ο μηχανισμός ταιριάσματος μπορεί να οριστεί ως η διαδικασία που επιλέγει το υποσύνολο $D_{\alpha\pi} \subseteq D$ των διαθεσίμων στοιχείων που είναι σχετικά με το ερώτημα q.

$$D_{\alpha\pi} = \{d \in D : d \text{ σχετικό με } q\} \tag{5.3}$$

Μαθηματικά η διαδικασία αυτή μπορεί να περιγραφεί από τη συνάρτηση f_3 που αντιστοιχίζει κάθε ερώτημα σε ένα υποσύνολο του συνόλου των διαθέσιμων στοιχείων.

$$f_3 : Q \rightarrow \mathcal{P}(D) \tag{5.4}$$

Έτσι, όταν ένας χρήστης θέτει το ερώτημα $\tilde{q}$, η απάντηση του συστήματος δίδεται από τη σχέση

$$\tilde{D}_{\alpha\pi} = (f_2 \circ f_3 \circ f_1^{-1})(\tilde{q}) \tag{5.5}$$

Είναι προφανές πως ο μηχανισμός δεικτοδότησης είναι υπεύθυνος για την αναπαράσταση της πληροφορίας, ο μηχανισμός μετάφρασης του ερωτήματος για την κατανόηση της επιθυμίας του χρήστη και ο μηχανισμός ταιριάσματος για την επιλογή των στοιχείων που ικανοποιούν αυτή την επιθυμία. Αυτοί οι τρεις μηχανισμοί μπορούν να θεωρούνται ως βασικοί για ένα σύστημα ανάκτησης πληροφορίας.

Έστω τώρα $R(\tilde{q})$ το σύνολο των στοιχείων που ικανοποιούν την επιθυμία ενός χρήστη που θέτει το ερώτημα $\tilde{q}$. Τα δύο κλασσικά μέτρα αποτίμησης της ανάκτησης, η ανάκληση και η ακρίβεια, ορίζονται ως εξής:

$$\begin{aligned} r(q) &= \frac{|\tilde{D}_{\alpha\pi}(q) \cap R(\tilde{q})|}{|R(\tilde{q})|} \\ p(q) &= \frac{|\tilde{D}_{\alpha\pi}(q) \cap R(\tilde{q})|}{\tilde{D}_{\alpha\pi}(q)} \end{aligned}$$

5.2.2 Μοντέλα συστημάτων αναζήτησης

Τα δυαδικά μοντέλα αναζήτησης βασίζονται στη θεωρία συνόλων, με την αναπαράσταση κάθε διαθέσιμου στοιχείου να είναι ένα σύνολο από όρους.

$$d = \{t_1, t_2, t_3., , , \} \tag{5.6}$$

Το ερώτημα q του χρήστη μεταφράζεται σε μια δυαδική παράσταση όρων

$$q = (t_1 \wedge t_2 \wedge t_3 \wedge ...) \vee (t_4 \wedge t_5 \wedge t_6 ... \wedge) \vee ... \tag{5.7}$$

και το ταίριασμα γίνεται με την απαίτηση της ικανοποίησης της παράστασης από τη δεικτοδότηση των διαθέσιμων στοιχείων.

Το δυαδικό μοντέλο αναζήτησης, συγκρινόμενο με άλλα μοντέλα, έχει σημαντικά μικρότερες απαιτήσεις σε χώρο αποθήκευσης των αναπαραστάσεων των στοιχείων. Καθώς το πρόβλημα του χώρου είναι το κυριότερο που αντιμετωπίζουν τα συστήματα που χειρίζονται μεγάλους αριθμούς στοιχείων, τα περισσότερα εμπορικά συστήματα αναζήτησης βασίζονται στο δυαδικό μοντέλο. Από την άλλη, ερευνητικά έχουν προ πολλού εγκαταλειφθεί, εξαιτίας της αδυναμίας τους να περιγράψουν την ανακρίβεια και την υποκειμενικότητα που χαρακτηρίζει τη διαδικασία αναζήτησης πληροφορίας.

Το διανυσματικό μοντέλο περιγράφεται στο [112]. Βασική του παραδοχή είναι πως για τη δεικτοδότηση όλων των διαθέσιμων στοιχείων και τη μετάφραση όλων των δυνατών ερωτημάτων του χρήστη αρκεί ένα πεπερασμένο πλήθος όρων. Αν T είναι το σύνολο των όρων, τότε η δεικτοδότηση αντιστοιχίζει σε κάθε διαθέσιμο στοιχείο $\tilde{d}$ ένα διάνυσμα d του χώρου $R^{|T|}$. Η τιμή στη θέση i του διανύσματος δείχνει το βαθμό στον οποίο το συγκεκριμένο στοιχείο σχετίζεται με τον όρο t_i.

Όμοια, το ερώτημα $\tilde{q}$ του χρήστη αντιστοιχίζεται σε ένα διάνυσμα q του χώρου $R^{|T|}$. Στο διανυσματικό μοντέλο, δηλαδή, η αναπαράσταση του ερωτήματος συμπίπτει με την αναπαράσταση του ιδανικού για το χρήστη στοιχείου.

Η διαδικασία ταιριάσματος υπολογίζει τις αποστάσεις ανάμεσα στις αναπαραστάσεις των διαθεσίμων στοιχείων και την αναπαράσταση του ερωτήματος του χρήστη. Οποιοδήποτε μετρικό ορισμένο στο χώρο $R^{|T|}$ μπορεί να χρησιμοποιηθεί για το ταίριασμα, αλλά το εσωτερικό γινόμενο των αναπαραστάσεων είναι το πλέον σύνηθες.

$$sim(q,d) = \sum_{i \in N_{|T|}} q_i \cdot d_i \tag{5.8}$$

Η δεικτοδότηση των διανυσματικών μοντέλων τους επιτρέπει να περιγράφουν πολύ καλύτερα τους διάφορους βαθμούς συσχέτισης των όρων με τα διαθέσιμα στοιχεία. Έχουν όμως το σημαντικό μειονέκτημα της απαίτησης T θέσεων αποθήκευσης δεδομένων για κάθε διαθέσιμο στοιχείο. Όταν τα διαθέσιμα στοιχεία είναι πολλά και ποικίλου περιεχομένου ο αριθμός $|T|$ των όρων που χρησιμοποιεί το σύστημα γίνεται μεγάλος και αποτρέπει τη χρησιμοποίηση του διανυσματικού μοντέλου.

Τα πιθανοτικά μοντέλα παρουσιάζονται στα [59], [128]. Βασίζονται στην παραδοχή ότι σε ένα IRS η σχετικότητα ενός στοιχείου με το ερώτημα του χρήστη μπορεί μόνο να υποτεθεί. Στο [128] προτείνεται ένας αλγόριθμος που επιχειρεί μέσω δοκιμαστικών αναζητήσεων να βελτιώσει την αξιοπιστία της υπόθεσης που γίνεται γύρω από τη σχετικότητα ερωτήματος και στοιχείου. Η αναπαράσταση των στοιχείων γίνεται μέσω δυαδικών διανυσμάτων και το ταίριασμα είναι μια στοχαστική διαδικασία που βασίζεται στο θεώρημα του Bayes. Συνήθως το ταίριασμα γίνεται για κάθε όρο χωριστά, οπότε χρειάζεται να υποθέσουμε στατιστική ανεξαρτησία των όρων για να συνδυάσουμε τα επιμέρους αποτελέσματα σε μια συνολική εκτίμηση για την πιθανότητα να είναι ένα στοιχείο σχετικό με το ερώτημα του χρήστη

Καθώς τα δυαδικά μοντέλα είναι τα μόνα που μπορούν εύκολα να χειριστούν μεγάλες συλλογές διαθεσίμων στοιχείων, έχουν γίνει προσπάθειες να εμπλουτιστούν με μερικά από τα πλεονεκτήματα των άλλων μοντέλων. Στη βιβλιογραφία υπάρχουν πολλές αναφορές σχετικά με μοντέλα που προκύπτουν από την επέκταση των δυαδικών μοντέλων [22] [39] [24], [23], [28], [77], [80], [29]. Συνήθως οι γενικεύσεις επιτρέπουν την εισαγωγή της έννοιας του βάρους στην αντιστοίχιση όρων σε στοιχεία, ή στη μετάφραση του ερωτήματος. Οι προτάσεις αυτές μοντελοποιούνται μέσω των ασαφών συνόλων. Άλλες προτάσεις διατηρούν τη δυαδική δεικτοδότηση αλλά μετά από το ταίριασμα κατατάσσουν τα επιλεγμένα στοιχεία σε σειρά προτεραιότητας χρησιμοποιώντας συναρτήσεις Dempster-Shafer [21].

Στα [97], [104] παρουσιάζεται η οικογένεια των μοντέλων αναζήτησης που βασίζονται σε νευρωνικά δίκτυα. Τα διαθέσιμα στοιχεία αποθηκεύονται στα νευρωνικά δίκτυα με τη μορφή προτύπων μέσω των διαδικασιών εκμάθησης. Το ερώτημα του χρήστη μεταφράζεται σε κατάλληλη είσοδο για το δίκτυο και η διαδικασία ταιριάσματος δίνει το ερώτημα στην είσοδο του δικτύου και λαμβάνει ως απάντηση στην έξοδό του τα πρότυπα που ενεργοποιήθηκαν, δηλαδή τα στοιχεία των οποίων οι αναπαραστάσεις είναι συμβατές με την αναπαράσταση του ερωτήματος.

5.2.3 Προβλήματα των συστημάτων αναζήτησης πληροφορίας

Τα συστήματα ανάκτησης πληροφορίας δεν αποδίδουν τέλεια, ακόμη και όταν οι αναζητήσεις γίνονται σε απλά αρχεία κειμένου. Αυτό είναι αναμενόμενο καθώς ο ρόλος τους είναι ο εντοπισμός στοιχείων που είναι σχετικά με αυτό που ο χρήστης επιθυμεί. Κανένα αυτόματο σύστημα δεν μπορεί να αποδώσει ιδανικά κυρίως για τους παρακάτω λόγους:

1. Συνήθως τα ερωτήματα των χρηστών προς τις μηχανές αναζήτησης αποτελούνται από δύο μόνο λέξεις [38], [31]. Προφανώς αυτές δεν αρκούν για να χαρακτηριστούν πλήρως οι περίπλοκες έννοιες που σχετίζονται με την επιθυμία του χρήστη.

2. Πολύ συχνά οι τελικοί χρήστες χρησιμοποιούν διαφορετικούς όρους από τους έμπειρους χρήστες που έχουν χαρακτηρίσει το αναζητήσιμο υλικό [88].

3. Οι χρήστες δεν ξέρουν πάντα από την αρχή τί ακριβώς ψάχνουν [19]. Αντίθετα, τα αποτελέσματα της αναζήτησής τους συχνά επηρεάζουν τις επιθυμίες τους.

4. Η αβεβαιότητα είναι εγγενής στον ορισμό της σχετικότητας εγγράφων με την επιθυμία του χρήστη [79].

Το γεγονός ότι οι κυριότερες πηγές αστοχίας των IRSs έχουν ήδη εντοπιστεί και επισημανθεί διευκολύνει σημαντικά την προσπάθεια για τη βελτίωσή τους και σε μεγάλο βαθμό κατευθύνει τη σχετική έρευνα. Έτσι από τα παραπάνω, για παράδειγμα, προκύπτουν αβίαστα τα ακόλουθα συμπεράσματα.

1. Καθώς το μέσο μήκος του ερωτήματος κρίνεται ανεπαρκές θα πρέπει να αναζητηθούν μέθοδοι επέκτασής του, ώστε να γίνει πληρέστερο.

2. Η ύπαρξη πολλαπλών όρων που περιγράφουν κοινές έννοιες, όπως και η ύπαρξη όρων που περιγράφουν πολλαπλές έννοιες οδηγεί στη χρήση λεξικών συνωνύμων κατά τη διαδικασία της αναζήτησης.

3. Καθώς τα αποτελέσματα της αναζήτησης επηρεάζουν την επιθυμία του χρήστη είναι λογικό να αναζητηθούν τρόποι ώστε η διαδικασία να αποκτήσει περισσότερο αλληλεπιδραστικό χαρακτήρα. Έτσι προκύπτει η ανάγκη για αξιοποίηση της ανατροφοδότησης από το χρήστη. Αυτή μπορεί να επιλύσει και προβλήματα που σχετίζονται με τις πολλαπλές έννοιες των όρων.

4. Η αδυναμία των όρων να περιγράψουν επαρκώς το ερώτημα ή τα διαθέσιμα στοιχεία οδηγεί στη διαπίστωση ότι τα IRSs θα έπρεπε να είναι σε θέση να χειρίζονται έννοιες αντί για όρους.

5. Το γεγονός ότι η αβεβαιότητα συνδέεται στενά με τις διαδικασίες που σχετίζονται με την αναζήτηση οδηγεί στη μοντελοποίηση των συστημάτων ανάκτησης της πληροφορίας με λιγότερο αυστηρό τρόπο. Έτσι όλο και περισσότερες μέθοδοι αναζήτησης βασίζονται σε ασαφή σύνολα και ασαφείς σχέσεις.

5.2.4 Επέκταση ερωτήματος

Η επέκταση ερωτήματος, δηλαδή ο εμπλουτισμός του με όρους που δεν έδωσε ο χρήστης αλλά είναι σχετικοί με αυτό, είναι μία τεχνική βελτίωσης της ανάκλησης (του πλήθους των σωστών απαντήσεων) των συστημάτων ανάκτησης πληροφορίας. Αναφέρονται τεχνικές τοπικής επέκτασης, όπου το σύνολο $\tilde{D}_{\alpha\nu}$ των ανακτημένων εγγράφων αναλύεται και τα χαρακτηριστικά που εξάγονται ρησιμοποιούνται για τον εμπλουτισμό της ερώτησης, και καθολικής επέκτασης, όπου το σύνολο $\tilde{D}$ των διαθέσιμων εγγράφων του συστήματος αναλύεται και το αποτέλεσμα της ανάλυσης είναι ένας θησαυρός, που είναι μία ασαφής ανακλαστική και συμμετρική σχέση.

Οι τοπικές μέθοδοι έχουν το μειονέκτημα ότι εξαρτώνται από το σύνολο των ανακτημένων εγγράφων. Αν δεν έχουν βρεθεί έγγραφα που να αντιστοιχούν σε έναν ή περισσότερους όρους της ερώτησης, τα χαρακτηριστικά αυτών των όρων δεν θα επηρεάσουν την επέκταση. Από την άλλη πλευρά, μία επέκταση βασιζόμενη σε θησαυρό παρουσιάζει το μειονέκτημα ότι ο θησαυρός είναι μία σχέση συμμετρική. Επομένως,

αν ένα ζευγάρι όρων t_1, t_2 συνδέονται μέσω του θησαυρού, η παρουσία του όρου t_1 στην ερώτηση θα οδηγήσει στην επέκτασή της με τον όρο t_2, στον ίδιο βαθμό που η παρουσία του όρου t_2 οδηγεί στην επέκταση της ερώτησης με τον όρο t_1. Αυτού του είδους η επέκταση έρχεται σε αντίθεση με το νόημα των όρων.

5.3 Ευφυής σημασιολογική επέκταση ερωτήματος

Όπως ειπώθηκε ανωτέρω, μία επέκταση ερωτήματος βασιζόμενη σε θησαυρό παρουσιάζει το μειονέκτημα ότι ο θησαυρός είναι μία σχέση συμμετρική. Αν ένα ζευγάρι όρων k_1, k_2 συνδέονται μέσω του θησαυρού, η παρουσία του όρου k_1 στην ερώτηση θα οδηγήσει στην επέκτασή της με τον όρο k_2, στον ίδιο βαθμό που η παρουσία του όρου k_2 οδηγεί στην επέκταση της ερώτησης με τον όρο k_1. Αυτού του είδους η επέκταση έρχεται σε αντίθεση με το νόημα των όρων.

Για παράδειγμα, αν δύο όροι συνδέονται ταξινομικά, δηλαδή αν $k_1 < k_2$, τότε η παρουσία του όρου k_2 στο ερώτημα θα έπρεπε να οδηγήσει στην επέκτασή του με τον πιο ειδικό όρο κ_1. Το αντίστροφο, αν ισχύει, θα ισχύει σε μικρότερο βαθμό. Αντίστοιχοι συλλογισμοί μπορούν να γίνουν και για την περίπτωση που οι δύο όροι σχετίζονται μερονομικά. Τέλος, σε πολλές περιπτώσεις, μόνο ορισμένοι από τους όρους που σχετίζονται με κάποιο όρο του ερωτήματος σχετίζονται και με το σημασιολογικό περιεχόμενό του. Αυτή η διάκριση, όμως, δεν είναι δυνατό να γίνει αυτόματα χωρίς τη χρήση προϋπάρχουσας γνώσης και τη θεώρηση του πλαισίου.

Βγαίνουν επομένως τα εξής συμπεράσματα:

1. Η επέκταση πρέπει να γίνεται από μία μη συμμετρική σχέση

2. Η πληροφορία του πλαισίου πρέπει να λαμβάνεται υπόψη

3. Η οντολογική γνώση είναι χρήσιμη στην κατασκευή της σχέσης επέκτασης

Στην ενότητα αυτή δίνουμε μία μέθοδο επέκτασης του ερωτήματος με βάση τη σχέση ημιταξινομίας T που δίνεται από τη σχέση 2.46 και του πλαισίου του ερωτήματος. Με $X(k_i)$ θα συμβολίζεται το ασαφές σύνολο των οντοτήτων που επεκτείνουν την οντότητα k_i, ενώ με x_{ij} θα συμβολίζεται ο βαθμός στον οποίο η οντότητα k_j συμμετέχει στο $X(k_i)$.

Δεδομένου του ορισμού που δόθηκε για τη σχέση T, οι οντότητες που μετέχουν στο πλαίσιο ενός όρου μπορεί να χρησιμοποιηθούν για να τον επεκτείνουν. Οι βαθμοί συμμετοχής αντιστοιχούν στη σημασία των όρων, ως προς την ικανοποίηση του χρήστη, ή αλλιώς, στην εκτίμηση που μπορούμε να κάνουμε σχετικά με τα έγγραφα που περιέχουν τους αντίστοιχους όρους. Αν μια οντότητα συμμετέχει στο πλαίσιο σε υψηλό βαθμό, τότε ένα έγγραφο που την περιέχει θα ικανοποιήσει το χρήστη. Αν, αντίθετα, ένα έγγραφο περιέχει μόνο οντότητες που συμμετέχουν στο πλαίσιο σε χαμηλό βαθμό, τότε πιθανότατα δεν συμβαδίζει με την επιθυμία του χρήστη.

Όταν η ισχύς του πλαισίου είναι χαμηλή, τότε οι όροι της ερώτησης έχουν μικρό κοινό νόημα. Στην περίπτωση αυτή η παρουσία των επιπλέον όρων δεν προτίθεται να καθορίσει το νόημά τους, αλλά απλώς να ανακτήσει έγγραφα που τους περιέχουν. Επομένως, πρέπει η επέκταση του ερωτήματος να γίνει χωρίς να ληφθεί υπόψη το πλαίσιο.

Σε μία επέκταση που δε λαμβάνει υπόψη το πλαίσιο, η τιμή x_{ij} είναι ανάλογη του βάρους w_i και του βαθμού συμμετοχής $T(k_i, k_j)$.

Πίνακας 5.1: *Επέκταση όρου engine*

Όρος	Χωρίς πλαίσιο	q_1	q_2	q_3
engine	1	1	1	1
ext-combustion	0.9	0.38	0.51	0.51
int-combustion	0.9	0.77	0.9	0.9
4-stroke	0.81	0.34	0.46	0.46
2-stroke	0.81	0.34	0.46	0.46
rocket	0.8	0.8	0.7	0.46
diesel	0.72	0.61	0.72	0.72
turbine	0.72	0.72	0.63	0.41
jet	0.58	0.58	0.51	0.33
prop plane	0.43	0.37	0.43	0.43

Πίνακας 5.2: *Επέκταση όρου airplane*

Όρος	Χωρίς πλαίσιο	q_1	q_2	q_3
airplane	1	1	1	1
prop plane	0.9	0.77	0.9	0.9
jet	0.9	0.9	0.78	0.51

$$x_{ij} \propto w_{ij} = w_i T(s_i, s_j) \tag{5.9}$$

Σε μία επέκταση που λαμβάνει υπόψη το πλαίσιο, η τιμή x_{ij} είναι αύξουσα συνάρτηση του βαθμού στον οποίο το πλαίσιο του k_j είναι σχετικό με το πλαίσιο του ερωτήματος. Η ισχύς του κοινού πλαισίου μπορεί να χρησιμοποιηθεί ως εκτίμηση αυτής της συνάφειας.

$$h_j = \max\left(\frac{h(T(k_j) \cap K(q))}{h_q}, c(h_q)\right) \tag{5.10}$$

όπου c είναι ένα ασαφές συμπλήρωμα.

Ένα μη γραμμικό συμπλήρωμα είναι απαραίτητο, για να έχει υψηλότερη διακριτική ικανότητα. Το συμπλήρωμα Yager, με w περίπου 0.5 δίνει καλά αποτελέσματα. Έχοντας ορίσει τη συνάφεια με το πλαίσιο, ορίζουμε το βαθμό επέκτασης ως:

$$x_{ij} = h_j w_{ij} \tag{5.11}$$

Έτσι, όταν ένας όρος δεν σχετίζεται με το πλαίσιο του ερωτήματος, εξαλείφεται.

Πίνακας 5.3: *Επέκταση όρου propeller*

Όρος	Χωρίς πλαίσιο	q_1	q_2	q_3
propeller	1	-	0.7	1
prop plane	0.9	-	0.63	0.9

5.4 Πειραματικά αποτελέσματα

Τα αποτελέσματα που παρουσιάζονται εδώ βασίζονται και πάλι στη σχέση του σχήματος 2.2. Θα δώσουμε την επέκταση των όρων των παρακάτω ερωτημάτων:

$$q_1 = Engine/1 + Airplane/1 \tag{5.12}$$

$$q_2 = Engine/1 + Airplane/1 + Propeller/0.7 \tag{5.13}$$

$$q_3 = Engine/1 + Airplane/1 + Propeller/1 \tag{5.14}$$

Η στήλη "χωρίς πλαίσιο" στον Πίνακα 5.1 δείχνει το βαθμό στον οποίο οι διάφορες έννοιες μετέχουν το $T(Engine)$. Αυτός θα ήταν και ο βαθμός τους στο $X(Engine)$, εάν δεν λαμβάναμε υπόψη το πλαίσιο.

Οι πίνακες 5.2 και 5.3 έχουν τα αντίστοιχα δεδομένα για τις σημασιολογικές οντότητες Airplane και Propeller. Παρατηρούμε τα ακόλουθα:

Η χρήση του πλαισίου στην επέκταση του όρου Engine στο ερώτημα q_1 έχει σαν αποτέλεσμα τη δραστική μείωση της συμμετοχής των όρων four-stroke, two-stroke και external combustion στην επέκταση. Αυτό ήταν επιθυμητό, καθώς εύκολα διαπιστώνει κανείς πως αυτοί οι όροι δεν σχετίζονταν με το συνολικό νόημα του ερωτήματος. Αντίθετα, όροι που σχετίζονται με το κοινό νόημα δεν απομακρύνονται. Συνολικά, μπορούμε να πούμε πως η διαδικασία της επέκτασης με βάση το πλαίσιο "μετακινεί" το ερώτημα προς την κατεύθυνση του κοινού νοήματος των όρων του.

Και σε αυτό το παράδειγμα, όπως και σε αυτό της ενότητας 2.6, τα τρία ερωτήματα δεν ήταν ανεξάρτητα. Είναι και τα τρία της μορφής

$$q = Engine/1 + Airplane/1 + Propeller/w \tag{5.15}$$

όπου το w παίρνει τις τιμές 0, 0.7, 1. Η σταδιακή μεταβολή του πλαισίου με την αλλαγή της τιμής του w είναι προφανής. Εύκολα βλέπει κανείς πως οι τιμές στην επέκταση των όρων του q_2 είναι πάντα ανάμεσα στις αντίστοιχες τιμές για τα q_1 και q_3. Συνεπώς, η μετάβαση από $w = 0$ σε $w = 1$ είναι σταδιακή, δείχνοντας έτσι πως η εξαγωγή πλαισίου από ασαφή ερωτήματα και η χρήση του για την επέκταση των ερωτημάτων είναι λογική.

□

Κεφάλαιο 6

Ανάλυση και θεματική κατηγοριοποίηση εγγράφων

6.1 Εισαγωγή

Ο ορισμός θεματικών κατηγοριών είναι στην ουσία μια ταξινόμηση κειμένων σε σημασιολογικό επίπεδο. Στο πλαίσιο της πλοήγησης σε συλλογές κειμένων, η έννοια της θεματικής κατηγοριοποίησης χρησιμοποιείται σαν ένα πρώτο επίπεδο οργάνωσης των διαθέσιμων κειμένων και διαχωρισμού τους σε υποχώρους αναζήτησης. Στο πλαίσιο της δεικτοδότησης, η θεματική κατηγοριοποίηση μπορεί να συμβάλει στον καθορισμό του πλαισίου του περιεχομένου ενός κειμένου για τη σωστή ερμηνεία των λέξεων που περιέχει. Στο πλαίσιο της αναζήτησης, οι χρήστες συχνά βρίσκουν πιο εύκολη την περιγραφή της θεματικής κατηγορίας που επιθυμούν παρά του ακριβούς περιεχομένου. Άλλες χρήσεις της θεματικής κατηγοριοποίησης περιλαμβάνουν τον προσδιορισμό του προφίλ του χρήστη, την συμπαγή αναπαράσταση του περιεχομένου, τη συσχετιστική ανάδραση υψηλού επιπέδου ή τις αυτόματες προτάσεις κειμένων που σχετίζονται με το πλαίσιο της επαφής με το χρήστη.

Στην περίπτωση των λεκτικών κειμένων, η θεματική κατηγοριοποίηση είναι συχνά άμεσα προσδιορίσιμη. Σε πολλές περιπτώσεις, όμως, πρέπει να εκτιμηθεί από το περιεχόμενο του κειμένου, δηλαδή από τις λέξεις και φράσεις που περιέχει. Για να επιτευχθεί αυτό, μια βάση γνώσης σχετική με τις έννοιες του κειμένου πρέπει να είναι διαθέσιμη. Το πρόβλημα της θεματικής κατηγοριοποίησης για πολυμεσικά έγγραφα είναι αρκετά διαφορετικό και πιο δύσκολο. Καταρχήν, οι οντότητες της δεικτοδότησης δεν απαντώνται στο έγγραφο στη λεκτική τους μορφή. Αναγνωρίσιμα χαρακτηριστικά πρέπει να εξαχθούν και να ταιριαστούν με αυτά που περιγράφουν τις οντότητες στη βάση γνώσης. Επιπρόσθετα, τα πολυμεσικά έγγραφα περιέχουν αντικείμενα και γεγονότα που σχετίζονται χωροχρονικά, όχι απλώς γραμματικά. Τέλος, έννοιες όπως αθλητισμός ή τέχνη δεν απαντώνται άμεσα στα πολυμεσικά έγγραφα και πρέπει να εξαχθούν με βάση μη αφηρημένα αντικείμενα που μπορούν να εντοπιστούν [99][100].

Αυτές οι δυσκολίες οδήγησαν στο μέσο της προηγούμενης δεκαετίας το MPE Group να αναγνωρίσει πως η περιγραφή των πολυμεσικών εγγράφων είναι σε πολλά σημεία αρκετά διαφορετική από εκείνη των λεκτικών κειμένων. Αυτό είχε ως συνέπεια την ανάπτυξη του προτύπου περιγραφής πολυμεσικού περιεχομένου MPEG-7, η πρώτη έκδοση του οποίου έγινε διαθέσιμη τον Οκτώβριο του 2001 [149]. Στο πρότυπο τα χαρακτηριστικά των εγγράφων περιγράφονται μέσω μιας πλούσιας συλλογής περιγραφέων: Τα χαρακτηριστικά χαμηλού επιπέδου μέσω περιγραφέων δομής και τα

χαρακτηριστικά υψηλού επιπέδου μέσω σημασιολογικών περιγραφέων, όπως είναι η σημασιολογική οντότητα. Τέλος, η θεματική κατηγοριοποίηση υποστηρίζεται μέσω των περιγραφέων Format Classification και Genre Classification.

Όσο αφορά στη σημασιολογική δεικτοδότηση των εγγράφων, αξίζει να σημειωθεί πως δεν είναι λογικό να απαιτούμε όλες οι οντότητες να συσχετίζονται με ένα έγγραφο στον ίδιο βαθμό. Μια πιο ευέλικτη δομή είναι απαραίτητη, ώστε κάθε σύνδεση οντότητας με έγγραφο να συνοδεύεται και από το βαθμό της σημαντικότητας της οντότητας για το συγκεκριμένο έγγραφο. Σε λεκτικά κείμενα, μια αυστηρά σημασιολογική εκτίμηση του βαθμού αυτού απαιτεί επεξεργασία φυσικής γλώσσας που σαν ερευνητικό πρόβλημα θεωρείται ακόμη ανοικτό. Πιο απλές τεχνικές περιλαμβάνουν τη στατιστική ανάλυση της εμφάνισης των όρων. Από την άλλη πλευρά, ένα πολυμεσικό έγγραφο περιέχει κάποια κανάλια πληροφορίας που μπορεί να χρησιμοποιηθούν για την εκτίμηση της σημασίας των αντικειμένων. Τέτοια μπορεί να είναι το βάθος (συνήθως τα αντικείμενα στο προσκήνιο είναι πιο σημαντικά από αυτά που βρίσκονται το βάθος/φόντο), η κίνηση (το αντικείμενο που κινείται είναι συνήθως το επίκεντρο ενδιαφέροντος της σκηνής) κλπ.

Μια άλλη πηγή βαθμών είναι η αβεβαιότητα [79],[81]. Αν και αυτή έχει μικρό ρόλο στη δεικτοδότηση κειμένου, η παρουσία της είναι κεντρική στην περίπτωση των πολυμεσικών εγγράφων, καθώς η ασφαλής εξαγωγή σημασιολογικών αντικειμένων από πολυμεσική πληροφορία παραμένει ένα πρόβλημα ανοικτό και η περιγραφή αντικειμένων μέσω οπτικών περιγραφέων είναι ατελής. Τέλος, βαθμούς μπορεί να προσφέρει το πλαίσιο γνώσης, καθώς μπορεί να καθορίσει ποια αντικείμενα σχετίζονται πραγματικά με το περιεχόμενο του εγγράφου.

Αξίζει πάντως να σημειωθεί πως, αν και η σημασία και η αβεβαιότητα είναι έννοιες διαφορετικές [51], οι βαθμοί τους στο πρόβλημα της δεικτοδότησης πολυμεσικών κειμένων είναι αλληλένδετοι. Έτσι, χρησιμοποιούμε έναν αριθμό για την αναπαράσταση και των δύο. Πιο αυστηρά, ένα πολυμεσικό έγγραφο δεικτοδοτείται ως ένα ασαφές σύνολο από σημασιολογικές οντότητες.

Ακόμη και αν λυθεί το δύσκολο πρόβλημα του εντοπισμού απλών αντικειμένων και γεγονότων σε ένα πολυμεσικό έγγραφο και την αντιστοίχισής τους με σημασιολογικές οντότητες της εγκυκλοπαίδειας, η επιλογή των αφηρημένων εννοιών με τις οποίες το έγγραφο σχετίζεται είναι δύσκολη και βασίζεται σε μεγάλο βαθμό στην αποθηκευμένη γνώση.

Τα βασικά προβλήματα που πρέπει να ξεπεράσουμε είναι πως

1. αν και μια σημασιολογκή οντότητα μπορεί να σχετίζεται με πολλές θεματικές κατηγορίες, λίγες μόνο από αυτές σχετίζονται με το έγγραφο.

2. ένα έγγραφο μπορεί να σχετίζεται με πολλές διαφορετικές μεταξύ τους θεματικές κατηγορίες.

3. η παρουσία μερικών οντοτήτων είναι τυχαία και δεν συνεπάγεται σχέση με τις αντίστοιχες θεματικές κατηγορίες.

4. η παρουσία μερικών οντοτήτων ίσως οφείλεται σε λάθη της διαδικασίας αναγνώρισης αντικειμένων.

Θέτοντας το πρόβλημα σε μια πιο αυστηρή βάση, η διαδικασία της θεματικής κατηγοριοποίησης λαμβάνει ως είσοδο τη σημασιολογική δεικτοδότηση των εγγράφων. Σε αυτή, κάθε έγγραφο αναπαρίσταται ως ένα ασαφές σύνολο d από σημασιολογικές

οντότητες. Με βάση αυτό το σύνολο, και τη γνώση που περιέχεται στη σημασιολογική εγκυκλοπαίδεια, η διαδικασία επιχειρεί να ανιχνεύσει το βαθμό στον οποίο το έγγραφο d σχετίζεται με τη θεματική κατηγορία $t \in TC$. Ορίζουμε αυτό το βαθμό ως $R_{\mathrm{TC}}(t, d)$.

Με άλλα λόγια, η διαδικασία που θέλουμε να ορίσουμε θα υπολογίζει τη σχέση

$$R_{\mathrm{TC}} : TC \times D \rightarrow [0, 1] \tag{6.1}$$

όπου D είναι το σύνολο των εγγράφων.

Σύμφωνα με το πρόβλημα 1, μια σημασιολογική οντότητα μπορεί να σχετίζεται με πολλαπλές, ασυσχέτιστες θεματικές κατηγορίες. Έτσι, είναι απαραίτητο ο αλγόριθμός μας να μπορεί να καθορίζει ποια ή ποιες από αυτές πραγματικά σχετίζονται με το έγγραφο. Για να γίνει αυτό με τρόπο λογικό, το νόημα των υπολοίπων οντοτήτων του εγγράφου θα πρέπει να συνυπολογιστεί.

Από την άλλη, όταν ένα έγγραφο συνδέεται με διαφορετικές θεματικές κατηγορίες σύμφωνα με το πρόβλημα 2, δεν πρέπει να υποθέτουμε πως κάθε οντότητα θα σχετίζεται με όλες τις κατηγορίες. Αντίθετα, είναι πιο λογικό να περιμένουμε πως η πλειονότητα των οντοτήτων θα συνδέεται με μια μόνο ομάδα συσχετισμένων θεματικών κατηγοριών. Έτσι, μια ομαδοποίηση των οντοτήτων είναι απαραίτητη πριν από την εξαγωγή των θεματικών κατηγοριών που σχετίζονται με αυτές.

Από αυτή τη διαδικασία πρέπει να εξαιρεθούν όροι των οποίων η παρουσία ίσως είναι παραπλανητική, όπως είναι οι όροι που περιγράφονται στα προβλήματα 3 και 4.

Η προτεινόμενη προσέγγιση περιέχει τα παρακάτω βήματα:

- Καθορισμός του πλήθους των διακριτών θεμάτων με τα οποία σχετίζεται ένα έγγραφο, χρησιμοποιώντας ιεραρχική ομαδοποίηση των οντοτήτων που το έγγραφο περιέχει με βάση το κοινό τους νόημα.
- Ασαφοποίηση της διαμέρισης επιτρέποντας επικάλυψη των ομάδων και ασαφείς βαθμούς συμμετοχής.
- Απομάκρυνση παραπλανητικών οντοτήτων και εξαγωγή των θεματικών κατηγοριών κάθε ομάδας.
- Συνάθροιση των αποτελεσμάτων των επιμέρους ομάδων για τον καθορισμό της θεματικής κατηγοριοποίησης του εγγράφου.

6.2 Ασαφής ιεραρχική ομαδοποίηση οντοτήτων

Η ιεραρχική ομαδοποίηση δεν μπορεί να εφαρμοστεί σε ένα ασαφές σύνολο. Έτσι ξεκινάμε παίρνοντας το σύνολο

$${}^{0+}d = \{s \in S : d(s) > 0\} \tag{6.2}$$

Καθώς δεν είναι γνωστό εκ των προτέρων το πλήθος των διακριτών θεμάτων με τα οποία σχετίζεται το έγγραφο d, ένας ιεραρχικός αλγόριθμος ομαδοποίησης πρέπει να εφαρμοστεί [95]. Τα δύο στοιχεία που πρέπει να οριστούν για να έχει καθοριστεί πλήρως ο αλγόριθμος ομαδοποίησης είναι το μετρικό για την εκτίμηση της απόστασης ανάμεσα σε δύο ομάδες και το κριτήριο τερματισμού.

Όταν ομαδοποιούμε σημασιολογικές οντότητες, το ιδανικό μετρικό είναι ένα που να εκτιμά τη σημασιολογική τους συσχέτιση. Έτσι, χρησιμοποιούμε την ισχύ του κοινού πλαισίου, όπως έχει οριστεί στη σχέση 2.52. Για τον ορισμό του πλαισίου χρησιμοποιούμε την αντίστροφη της ημιταξινομικής ασαφούς σχέσης T που δημιουργείται σύμφωνα με την εξίσωση 2.44. Η διαδικασία πρέπει να τερματίσει όταν οι οντότητες έχουν μοιραστεί σε ομάδες ανάλογα με το νόημά τους. Αυτό το γεγονός μπορεί να αναγνωριστεί από τη χαμηλή τιμή του κοινού πλαισίου για όλα τα ζεύγη των ομάδων, καθώς αυτό δείχνει πως οι ομάδες είναι σημασιολογικά διακριτές. Έτσι, το κριτήριο τερματισμού είναι ένα κατώφλι στην τιμή του μετρικού που χρησιμοποιείται και για την επιλογή του ζεύγους που θα ενωθεί σε κάθε βήμα.

Καθώς η προσέγγισή μας είναι ιεραρχική, θα βρει το πλήθος των ομάδων οντοτήτων που υπάρχουν στο σύνολο ${}^{0+}d$. Υστερεί, όμως, σε σχέση με άλλες προσεγγίσεις στα παρακάτω:

- Δημιουργεί αυστηρές διαμερίσεις, μην επιτρέποντας την επικάλυψη ανάμεσα στις ομάδες.
- Δεν επιτρέπει την εκτίμηση ασαφών βαθμών συμμετοχής.

Και τα δύο αυτά στοιχεία είναι σημαντικές αδυναμίες για το πρόβλημα που επιχειρούμε να λύσουμε, καθώς δεν είναι συμβατά με τη σημασιολογική ερμηνεία της ομαδοποίησης των οντοτήτων. Στον πραγματικό κόσμο, μια έννοια μπορεί να σχετίζεται με ένα θέμα σε βαθμό διαφορετικό από 100%, και μπορεί επίσης να σχετίζεται με περισσότερες από μία έννοιες. Για να ξεπεράσουμε αυτά τα προβλήματα, συνεχίζουμε με την ασαφοποίηση των αποτελεσμάτων της ομαδοποίησης. Με αυτό τον τρόπο "διορθώνονται" και οι τιμές της πληθικότητας των ομάδων, ώστε να μπορούν να χρησιμοποιηθούν στη συνέχεια για την εκτίμηση της εγκυρότητάς τους.

Κάθε ομάδα περιγράφεται από το σύνολο $c \subseteq^{0+} d$ των οντοτήτων που περιέχει. Με αυτές, μπορούμε να σχεδιάσουμε έναν ασαφή ταξινομητή, δηλαδή μια συνάρτηση

$$C_c : S \to [0, 1] \tag{6.3}$$

που να μετρά το βαθμό συσχέτισης μιας οντότητας s με την ομάδα. Προφανώς, ο βαθμός $C_c(c, s)$ θα πρέπει να είναι μεγάλος όταν η οντότητα σχετίζεται σημασιολογικά με την ομάδα, δηλαδή όταν το κοινό τους πλαίσιο είναι ισχυρό. Έτσι, η τιμή

$$Cor_1(c, s) = h(K(c \cup \{s\})) \tag{6.4}$$

είναι ένα λογικό μέτρο. Από την άλλη πλευρά, δεν είναι όλες οι ομάδες το ίδιο "συμπαγείς" ως προς το σημασιολογικό τους περιεχόμενο. Άλλες περιέχουν οντότητες που αναφέρονται σε ένα πολύ συγκεκριμένο θέμα και άλλες περιγράφουν κάτι πιο ευρύ. Μια εκτίμηση για τη σημασιολογική ευρύτητα μιας ομάδας μας δίδεται από την ισχύ του πλαισίου που τη χαρακτηρίζει. Έτσι, διορθώνουμε την τιμή $Cor_1(c, s)$ ως εξής:

$$C_c(s) = \frac{Cor_1(c, s)}{h(K(c))} = \frac{h(K(c \cup \{s\}))}{h_c} \tag{6.5}$$

Χρησιμοποιώντας τέτοιους ταξινομητές μπορούμε να επεκτείνουμε τις μη ασαφείς διαμερίσεις ώστε να περιλάβουν περισσότερες οντότητες, καθεμιά στο βαθμό της. Συγκεκριμένα, η διαμέριση c αντικαθίσταται από την ασαφή ομάδα

$$c_{fuzzy} = \sum_{s \in {}^{0+}d} s/C_c(s) \tag{6.6}$$

Προφανώς

$$c_{fuzzy} \supseteq c \tag{6.7}$$

6.3 Εξαγωγή θεματικών κατηγοριών

Η διαδικασία της ομαδοποίησης έχει βασιστεί στο σύνολο ${}^{0+}d$, αγνοώντας έτσι την ύπαρξη βαθμών στη δεικτοδότηση του εγγράφου d. Για να εισαχθεί αυτή η πληροφορία στις ομάδες υπολογίζουμε τους διορθωμένους βαθμούς συμμετοχής των οντοτήτων ως εξής:

$$c_{final}(s) = t(c_{fuzzy}(s), d(s)) \tag{6.8}$$

όπου t μια τριγωνική νόρμα. Η σημασιολογική έννοια αυτής της πράξης επιβάλει η νόρμα να είναι Αρχιμήδεια.

Σύμφωνα με τη σχέση 2.45 το σύνολο TC των θεματικών κατηγοριών είναι υποσύνολο του S στο οποίο έχει οριστεί η σχέση T. Αυτό απλοποιεί τη διαδικασία της εξαγωγής της θεματικής κατηγοριοποίησης από τις ομάδες σημασιολογικών οντοτήτων με τη βοήθεια του πλαισίου. Συγκεκριμένα, με κάθε ομάδα σχετίζονται εκείνες οι θεματικές κατηγορίες που ανήκουν στο πλαίσιο της ομάδας. Πρέπει όμως να ληφθούν υπόψη τόσο το πλαίσιο της ομάδας όσο και η πληθικότητά της.

Το πλαίσιο έχει οριστεί μόνο για κανονικά ασαφή σύνολα. Μετά το βήμα της αξιοποίησης των βαθμών της δεικτοδότησης, όμως, δεν είναι βέβαιο πως οι ομάδες παραμένουν κανονικές. Ξεκινάμε, λοιπόν, με κανονικοποίηση των βαθμών συμμετοχής σε κάθε ομάδα

$$c_{normal}(s) = \frac{c_{final}(s)}{h(c_{final})} \ , \forall s \in {}^{0+} d \tag{6.9}$$

Προφανώς, οι θεματικές κατηγορίες που δεν ανήκουν στο πλαίσιο του ασαφούς συνόλου c_{normal} δεν σχετίζονται με την αντίστοιχη ομάδα. Έτσι

$$R_{\mathrm{TC}}(c_{final}) \subseteq R^1_{\mathrm{TC}}(c_{normal}) = w(K(c_{normal}) \cap T) \tag{6.10}$$

όπου w ένας ασθενής ασαφής τροποποιητής [73].

Όταν όλες οι οντότητες που δεικτοδοτούν το έγγραφο d περιλαμβάνονται στην ίδιο ομάδα c_{final}, τότε λογικά

$$R_{\mathrm{TC}}(d) = R^1_{\mathrm{TC}}(c_{normal}) \tag{6.11}$$

Όταν, όμως, οι σημασιολογικές οντότητες μοιράζονται σε διαφορετικές ομάδες, τότε είναι απαραίτητο η πληθικότητα της κάθε ομάδας να ληφθεί υπόψη. Ομάδες μικρής πληθικότητας είναι πολύ πιθανό να περιέχουν μόνο παραπλανητικές οντότητες, οπότε καλό είναι να αγνοηθούν κατά την εκτίμηση της σχέσης $R_{\mathrm{TC}}(d)$. Αντίθετα, ομάδες μεγάλης πληθικότητας σχεδόν βέβαια αντιστοιχούν σε κάποιο από τα διακριτά θέματα που αφορούν το d. Η έννοια της "μεγάλης" πληθικότητας μοντελοποιείται με τη χρήση ενός ασαφούς αριθμού $L(\cdot)$. $L(a)$ είναι η τιμή αληθείας της πρότασης "Η τιμή a είναι μεγάλη", οπότε $L(|b|)$ είναι η τιμή αληθείας της πρότασης "η πληθικότητα της ομάδας b είναι μεγάλη".

Συνολικά, η σχέση $R_{\mathrm{TC}}(d)$ υπολογίζεται από τη συνένωση των αποτελεσμάτων για κάθε ομάδα, αφού πρώτα ληφθεί υπόψη η πληθικότητα:

$$R_{\mathrm{TC}}(d) = \bigcup_{c_{final} \in G} (R_{\mathrm{TC}}(c^{final})) \tag{6.12}$$

$$R_{\mathrm{TC}}(c_{final}) = R^1_{\mathrm{TC}}(c_{normal}) \cdot L(|c_{final}|) \tag{6.13}$$

όπου $\cup$ μια σ-νόρμα και G το σύνολο των ασαφών ομάδων σημασιολογικών οντοτήτων που ανιχνεύθηκαν στο έγγραφο d.

Έτσι, η σχέση $R_{\mathrm{TC}}(d,t)$ θα είναι μεγάλη αν μια ομάδα c_{final}, της οποίας το πλαίσιο περιέχει τη θεματική κατηγορία t, ανιχνευθεί στο d, και επιπρόσθετα η πληθικότητα της c_{final} είναι μεγάλη και ο βαθμός συμμετοχής της t στο πλαίσιο της ομάδας είναι μεγάλος. Πρακτικά αυτό σημαίνει πως η σχέση $R_{\mathrm{TC}}(d,t)$ είναι μεγάλη όταν η t ανήκει σε μεγάλο βαθμό σε μια ομάδα και η ομάδα δεν αποτελείται από παραπλανητικές οντότητες.

Οι καλύτερες επιδόσεις έχουν παρατηρηθεί με χρήση των παρακάτω:

- Η τριγωνική νόρμα που χρησιμοποιείται για το μεταβατικό κλείσιμο της σχέσης T είναι η νόρμα Yager με τιμή παραμέτρου ίση με 3.

- Στή σχέση 6.8, η τριγωνική νόρμα t που χρησιμοποιούμε είναι το αλγεβρικό γινόμενο.

- Στή σχέση 6.10, ο ασθενής ασαφής τροποποιητής είναι

$$w(a) = \sqrt{a} \tag{6.14}$$

- Στή σχέση 6.12, χρησιμοποιούμε την κλασσική ένωση max.

- Το κατώφλι για τον τερματισμό της διαδικασίας ιεραρχικής ομαδοποίησης είναι 0.3.

- Στη σχέση 6.13, ο μεγάλος ασαφής αριθμός L ορίζεται ως ο τριγωνικός αριθμός $(1.3, 3, \infty)$:

$$L(a) = \begin{cases} a \leq 1.3, & 0 \\ 1.3 < a < 1, & \frac{1}{1.7}a - \frac{1.3}{1.7} \\ a >\geq 1, & 1 \end{cases} \tag{6.15}$$

6.4 Πειραματικά αποτελέσματα

Στις ακόλουθες παραγράφους ξεκινάμε δείχνοντας τη λειτουργία της μεθόδου σε μια σειρά συνθετικών δεδομένων και συνεχίζουμε παρουσιάζοντας αντίστοιχα αποτελέσματα από εφαρμογή σε πραγματικά έγγραφα.

Πίνακας 6.1: *Ονόματα σημασιολογικών οντοτήτων*

Οντότητα	Μνημονικό	Οντότητα	Μνημονικό
arts	**art**	army or police uniform	unf
tank	tnk	lawn	lwn
missile	msl	goal	gol
scene	scn	shoot	sht
war	**war**	tier	tir
cinema	**cnm**	river	riv
performer	prf	speak	spk
sitting person	spr	F16	f16
explosion	exp	football player	fpl
launch of missile	lms	goalkeeper	glk
screen	scr	**theater**	**thr**
football	**fbl**	fighter airplane	far
curtain	crn	seat	sit

Πίνακας 6.2: *Η σχέση T της θεματικής κατηγοριοποίησης*

s_1	s_2	$T(s_1, s_2)$	s_1	s_2	$T(s_1, s_2)$
war	unf	0.90	war	exp	0.60
war	far	0.80	fbl	gol	0.80
war	tnk	0.80	fbl	sit	0.60
war	msl	0.80	cnm	sit	0.60
thr	scn	0.90	fbl	sht	0.90
thr	prf	0.90	fbl	tir	0.80
thr	spr	0.80	fbl	fpl	0.90
war	lms	0.70	fbl	lwn	0.90
cnm	scr	0.90	cnm	spr	0.80
fbl	spr	0.60	thr	sit	0.60
thr	crn	0.70	art	thr	0.80
far	f16	1.00	art	cnm	0.80
fpl	glk	1.00			

Πίνακας 6.3: *Η ασαφής σχέση δεικτοδότησης*

s	$d_1(s)$	s	$d_2(s)$	s	$d_3(s)$	s	$d_4(s)$
prf	0,9	spr	0.9	spr	0.8	spr	0.2
spr	0.9	spk	0.8	unf	0.9	unf	0.3
spk	0.6	sit	0.9	lwn	0.6	lwn	0.4
sit	0.7	scr	1.00	gol	0.9	gol	0.3
crn	0.8	tnk	0.4	tir	0.7	tir	0.4
scn	0.9			spk	0.9	spk	0.2
tnk	0.7			glk	0.6	glk	0.3
				sht	0.5	sht	0.4

Πίνακας 6.4: *Το αποτέλεσμα της μεθόδου*

	$R_{TC}(d_1)$	$R_{TC}(d_2)$	$R_{TC}(d_3)$	$R_{TC}(d_4)$	$R_{TC}(d_5)$
arts	0.84	0.73			0.85
cinema		0.74			0.86
theater	0.89				0.33
football			0.84	0.37	0.77
war					0.77

6.4.1 Συνθετικά δεδομένα

Οι σημασιολογικές οντότητες του παραδείγματος παρουσιάζονται στον Πίνακα 6.1 (οι θεματικές κατηγορίες εμφανίζονται με έντονη γραφή). Τα μηδενικά στοιχεία της σχέσης, καθώς και όσα υπονοούνται από τη μεταβατικότητα της σχέσης T, έχουν παραλειφθεί για λόγους απλότητας στον Πίνακα 6.2. Η δεικτοδότηση των εγγράφων παρουσιάζεται στον πίνακα 6.3 και τα αποτελέσματα της εφαρμογής του αλγορίθμου στον πίνακα 6.4.

Το έγγραφο d_1 περιέχει μια σκηνή από ένα θέατρο. Η παράσταση στο θέατρο είναι σχετική με πόλεμο. Στη δεικτοδότηση τα αντικείμενα και τα γεγονότα έχουν εντοπισθεί με μικρούς βαθμούς βεβαιότητας. Επίσης ορισμένες οντότητες, όπως η speak, δεν σχετίζονται άμεσα με το συνολικό νόημα της σκηνής. Ο αλγόριθμος σωστά αγνοεί τις οντότητες tank και speak.

Το έγγραφο d_2 περιέχει μια σκηνή από ένα σινεμά. Η ταινία στο σινεμά είναι και πάλι σχετική με πόλεμο. Αν και αρκετές οντότητες είναι κοινές ανάμεσα σε $d1$ και $d2$, ο αλγόριθμος σωστά ανιχνεύει πως αυτή τη φορά το συνολικό θέμα αφορά το σινεμά και όχι το θέατρο. Αυτό γίνεται εξαιτίας της επίδρασης της οντότητας screen στον καθορισμό του πλαισίου.

Τα έγγραφα d_3 και d_4 σχετίζονται με το ποδόσφαιρο. Περιέχουν τις ίδιες οντότητες, αλλά η βεβαιότητα με την οποία αυτές έχουν ανιχνευθεί διαφέρει. Ο αλγόριθμος επιτυχώς χρησιμοποιεί τους βαθμούς αβεβαιότητας στη διαμόρφωση της τελικής εξόδου.

Τέλος, ας θεωρήσουμε το έγγραφο d_5, που έχει μια ακολουθία σκηνών από ένα δελτίο ειδήσεων. Καθώς περιέχει αναφορές σε διάφορα θέματα, οι οντότητες που περιέχονται σε αυτό είναι πολλές και δεν σχετίζονται όλες μεταξύ τους:

d_5 = spr/0.9 + unf/0.8 + lwn/0.5 + gol/0.9 + tir/0.7 + spk/0.9 + glk/0.8 + sht/0.5 + prf/0.7 + sit/0.9 + crn/0.7 + scn/0.8 + tnk/0.9 + msl/0.8 + exp/0.9 + riv/1

Ο αλγόριθμος δημιουργεί τις ακόλουθες ασαφείς ομάδες:

c_1 = spk/0.9
c_2 = riv/1.0
c_3 = spr/0.9 + prf/0.7 + sit/0.77 + crn/0.7 + scn/0.8
c_4 = spr/0.9 + lwn/0.5 + gol/0.9 + tir/0.7 + glk/0.8 + sht/0.5 + sit/0.9
c_5 = unf/0.8 + tnk/0.9 + msl/0.8 + exp/0.9

Βλέπουμε πως ο αλγόριθμος αναγνωρίζει την ύπαρξη διαφορετικών θεμάτων στο έγγραφο. Επιπλέον, οντότητες όπως seat και sitting-person μετέχουν σε περισσότερα

από ένα από αυτά τα θέματα, καθώς σχετίζονται με περισσότερα από ένα από τα πλαίσια του κειμένου. Στα επόμενα βήματα του αλγορίθμου οι πρώτες δύο ομάδες παραλείπονται εξαιτίας της μικρής τους πληθικότητας.

6.4.2 Εφαρμογή σε πραγματικά έγγραφα

Για την εφαρμογή του αλγορίθμου σε πραγματικά δεδομένα χρησιμοποιήθηκε το WordNet για τη δημιουργία του καθολικού συνόλου σημασιολογικών οντοτήτων. Στους ορισμούς πολλών οντοτήτων προστέθηκαν με τη χρήση κατάλληλων εργαλείων μεταφράσεις στα ελληνικά ώστε να ανιχνεύονται αυτόματα σε ελληνικά κείμενα.

Επιπρόσθετα, χρησιμοποιήθηκε η μεθοδολογία σταδιακής ενημέρωσης μεταβατικής σχέσης για τη σταδιακή δημιουργία μιας σχέσης T ικανής να περιγράψει το περιεχόμενο των κειμένων μιας σειράς μεταδεδομένων από το αρχείο της ΕΡΤ.

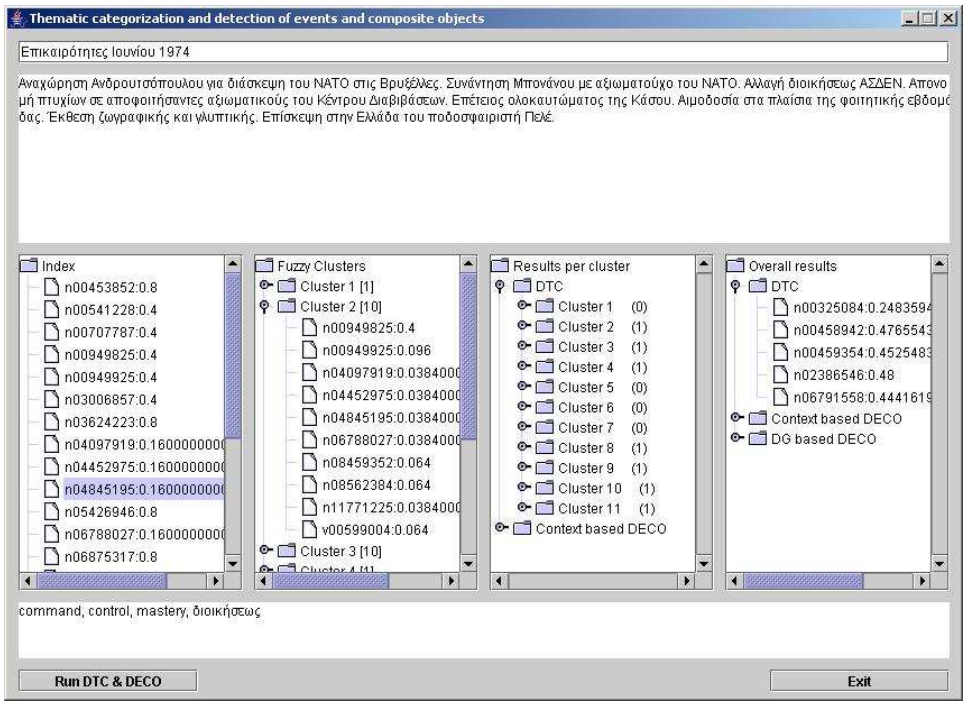
Με αυτά σαν είσοδο, ο αλγόριθμος χρησιμοποιήθηκε για τον αυτόματο καθορισμό της θεματικής κατηγοριοποίησης του υλικού. Για παράδειγμα, για το έγγραφο με τίτλο και περιγραφή:

> Επικαιρότητες Αυγούστου 1974
>
> 39η Διεθνής Έκθεση Θεσσαλονίκης. Σύσκεψη για την στρατιωτική κατάσταση της Ελλάδας. Ανάληψη καθηκόντων νέας στρατιωτικής ηγεσίας. Ανάληψη καθηκόντων νέου Υφυπουργού Εθνικής Αμύνης. Συνάντηση ΑΓΕΕΘΑ με Αρχιεπίσκοπο Αθηνών. Επίσκεψη Γενικού Γραμματέα ΟΗΕ. Επίσκεψη Υπουργού Κοινωνικών Υπηρεσιών στο νοσοκομείο Άγιος Σάββας. Ορκωμοσία εφέδρων αξιωματικών 74Α ΕΣΟ. Ορκωμοσία οπλιτών 74Δ ΕΣΟ. Ορκωμοσία οπλιτών 74Δ ΕΣΟ.

η έξοδος του αλγορίθμου είναι στρατιωτικά/0.71 + ηγέτες/0.25 + υγεία/0.25. Όμοια, για το έγγραφο με τίτλο και περιγραφή:

> Επικαιρότητες Ιουνίου 1974
>
> Αναχώρηση Ανδρουτσόπουλου για διάσκεψη του ΝΑΤΟ στις Βρυξέλλες. Συνάντηση Μπονάνου με αξιωματούχο του ΝΑΤΟ. Αλλαγή διοικήσεως ΑΣΔΕΝ. Απονομή πτυχίων σε αποφοιτήσαντες αξιωματικούς του Κέντρου Διαβιβάσεων. Επέτειος ολοκαυτώματος της Κάσου. Αιμοδοσία στα πλαίσια της φοιτητικής εβδομάδας. Έκθεση ζωγραφικής και γλυπτικής. Επίσκεψη στην Ελλάδα του ποδοσφαιριστή Πελέ. Ποδόσφαιρικός αγώνας κυπέλου Ολυμπιακού-ΠΑΟΚ.

η έξοδος του αλγορίθμου είναι αθλητικά/0.73 + εκπαιδευση/0.48 + τεχνες/0.48 + πολιτική/0.44 + στρατιωτικά/0.44.



Σχήμα 6.1: *Ανίχνευση θεματικών κατηγοριών σε πραγματικά έγγραφα*

Κεφάλαιο 7

Αποτίμηση ασαφών κανόνων

7.1 Εισαγωγή

Οι ασαφείς κανόνες και τα συστήματα που βασίζονται σε αυτούς έχουν χρησιμοποιηθεί εκτενώς στο παρελθόν για την ανάπτυξη έμπειρων συστημάτων, καθώς προσφέρουν μια αναπαράσταση γνώσης που βοηθά πολύ τους έμπειρους χρήστες να τυποποιούν τη γνώση που καθοδηγεί ένα ευφυές σύστημα. Το αποτέλεσμα είναι συστήματα ικανά να επεξεργαστούν περίπλοκα δεδομένα σε πολύ λίγο χρόνο και να αντιδρούν κατάλληλα.

Όταν ο χρόνος απόκρισης είναι ένα κρίσιμο στοιχείο, τότε τα συστήματα που βασίζονται σε ασαφείς κανόνες θεωρούνται συχνά ως η προφανής ή η μόνη επιλογή. Μετρήσεις που λαμβάνονται από αισθητήρες αποτυπώνονται σε λεκτικές μεταβλητές υψηλού επιπέδου, οι οποίες με τη σειρά τους χρησιμοποιούνται για να επιτευχθεί μια πολύ γρήγορη προσέγγιση της ιδανικής απόκρισης. Οι μόνες απαιτήσεις είναι

- η ιδανική απόκριση να είναι συνεχής συνάρτηση των παραμέτρων εισόδου
- η γνώση που περιέχεται στους κανόνες να είναι σωστή
- οι είσοδοι να είναι διαθέσιμες

Αν και αυτές ίσως μοιάζουν με χαλαρές απαιτήσεις που εύκολα ικανοποιούνται, υπάρχουν πλέον περιπτώσεις στις οποίες δύο από αυτές τις υποθέσεις δεν ισχύουν απαραίτητα, αλλά η χρήση συστημάτων ασαφών σχέσεων είναι επιθυμητή:

- Άν κάποιοι όροι είναι προαιρετικοί, τα κλασικά συστήματα είτε αγνοούν την προαιρετικότητα ή αγνοούν παντελώς αυτούς τους όρους
- Όταν οι μεταβλητές εισόδου δεν προέρχονται από κάποιο αισθητήρα αλλά από την έξοδο κάποιου άλλου αυτοματοποιημένου συστήματος, τότε η διαθεσιμότητά τους εξαρτάται από τη λειτουργία αυτού του συστήματος. Αν, για παράδειγμα, το σύστημα αυτό επεξεργάζεται ακολουθίες βίντεο, τότε είναι πιθανό για ορισμένα frames είτε να μην έχει έξοδο ή αυτή η έξοδος να μην είναι ασφαλής.

Σε αυτό το κεφάλαιο προσπαθούμε να δώσουμε μια λύση και στα δύο αυτά προβλήματα. Για την αναπαράσταση προαιρετικών όρων επεκτείνουμε το μοντέλο των ασαφών κανόνων ενώ για το ενδεχόμενο ελλειπούς ή αβέβαιας εισόδου προτείνουμε ένα δυνατοτικό μοντέλο αποτίμησης ασαφών κανόνων.

7.2 Ασαφείς κανόνες και προαιρετικοί όροι

Τα έμπειρα συστήματα είναι αυτοματοποιημένα συστήματα βασισμένα στη γνώση που επιχειρούν να μιμηθούν τη διαδικασία με την οποία ο άνθρωπος λαμβάνει αποφάσεις. Ένα κεντρικό στοιχείο για το σχεδιασμό και τη λειτουργία τους είναι η αναπαράσταση της γνώσης που περιέχουν. Αυτή πρέπει να είναι εύκολα κατανοητή από τον άνθρωπο, ώστε οι έμπειροι χρήστες να μπορούν εύκολα να την ελέγξουν ή να τη διορθώσουν αν έχει εξαχθεί αυτόματα από το σύστημα μέσω μια διαδικασίας μάθησης, εποπτευμένης ή μη.

Τα συστήματα ασαφών κανόνων είναι έμπειρα συστήματα που αποθηκεύουν τη γνώση στη μορφή κανόνων όπως ο ακόλουθος:

$$IF\ x_1, x_2 \ldots x_n\ THEN\ y \tag{7.1}$$

όπου y είναι η έξοδος του κανόνα και x_i οι όροι εισόδου. Οι όροι εισόδου συνήθως ακολουθούν τη μορφή:

$$x_i : f_i\ IS\ X_i \tag{7.2}$$

όπου f_i είναι ένα μετρήσιμο χαρακτηριστικό και X_i ένας ασαφής αριθμός που ποσολογεί τη μέτρηση. Όμοια, η έξοδος ακολουθεί τη μορφή

$$y : o\ IS\ Y \tag{7.3}$$

όπου o η μεταβλητή που ελέγχεται από την έξοδο του κανόνα και Y ο ασαφής αριθμός που ποσολογεί την τιμή της. Για παράδειγμα, ο επόμενος είναι ένας ασαφής κανόνας που θα μπορούσε να χρησιμοποιείται από ένα σύστημα:

$$\begin{aligned} &IF\ temprature\ IS\ high\_temp \\ &AND\ humidity\ IS\ high\_hum \\ &THEN\ it\_feels\_hot\ IS\ true \end{aligned} \tag{7.4}$$

Σε αυτό το παράδειγμα έχουμε $n = 2$ για το πλήθος των όρων του κανόνα, *temprature* και *humidity* τα μετρήσιμα χαρακτηριστικά, *high_temp* και *high_hum* τις λεκτικές μεταβλητές που τα ποσολογούν, η μεταβλητή εξόδου είναι it_feels_hot και η ποσολόγηση της μεταβλητής εξόδου γίνεται από το *true*.

Κανόνες σαν και αυτόν αναφέρονται στη βιβλιογραφία ως ασαφείς διότι τα μετρήσιμα χαρακτηριστικά ποσολογούνται με τη χρήση ασαφών αριθμών. Για παράδειγμα, όπως φαίνεται στο σχήμα 7.1 μια μέτρηση θερμοκρασίας ανάμεσα στους είκοσι και τους τριάντα βαθμούς θα ποσολογηθεί σε κάποιο βαθμό διάφορο της μονάδας και του μηδενός.

Η ίδια η γνώση, από την άλλη πλευρά, όπως περιγράφεται από τον κανόνα, δεν είναι καθόλου ασαφής. Ο κανόνας περιγράφει πως υψηλή θερμοκρασία και υψηλή υγρασία έχουν σαν αποτέλεσμα αίσθηση ζέστης. Δεν μπορεί όμως να περιγράψει, για παράδειγμα, πως και μόνο η υψηλή θερμοκρασία μπορεί να φέρει το ίδιο αποτέλεσμα, ενώ η υψηλή υγρασία είναι ένας προαιρετικός όρος που απλά κάνει το φαινόμενο πιο έντονο. Ένα πρώτο βήμα προς την αντιμετώπιση αυτού του προβλήματος γίνεται στο [34], όπου βαθμοί σημαντικότητας ανατίθενται στους όρους του κανόνα. Αυτή η λύση, όμως, δεν λύνει συνολικά το πρόβλημα, καθώς όταν λείπει ένας προαιρετικός όρος είναι αδύνατο να ενεργοποιηθεί ένας κανόνας σε βαθμό μονάδα. Αυτό συμβαίνει

μόνο όταν η σημασία του όρου μειωθεί στο μηδέν, δηλαδή όταν αφαιρεθεί τελείως από τον κανόνα και δεν επηρεάζει την ενεργοποίησή του ακόμη και όταν υπάρχει.

Στη βιβλιογραφία συναντάμε και τον όρο "σταθμισμένος ασαφής κανόνας", αλλά ο όρος δεν αναφέρεται σε σταθμισμένους όρους στον κανόνα, που θα μπορούσε κανείς να θεωρήσει σαν μια προσπάθεια λύσης του προβλήματος. Αντίθετα, αναφέρεται σε βαθμούς που συνοδεύουν τους κανόνες συνολικά και προσδιορίζουν τον βαθμό αξιοπιστίας του κανόνα ή το μέρος των δεδομένων εκπαίδευσης που υποστηρίζουν τον κανόνα.

Βέβαια είναι δυνατή η αντιμετώπιση του προβλήματος και με τους κλασικούς ασαφείς κανόνες, οι λύσεις αυτές όμως αντιμετωπίζουν σοβαρά προβλήματα:

- Μπορούμε να αφαιρέσουμε παντελώς το προαιρετικό κομμάτι από τον κανόνα. Στο παραπάνω παράδειγμα, ο κανόνας θα ορίζει πως υψηλή θερμοκρασία συνεπάγεται και αίσθημα ζέστης. Από την άλλη πλευρά, όταν η μέτρηση της υγρασίας θα είναι διαθέσιμη, αυτός ο κανόνας δεν θα είναι σε θέση να την αξιοποιήσει.

- Μπορούμε να ορίσουμε δύο διαφορετικούς κανόνες. Ο ένας θα περιέχει τον προαιρετικό όρο και ο δεύτερος όχι. Αυτό όμως έχει τα μειονεκτήματα πως:

  1. Η βάση γνώσης γίνεται πιο περίπλοκη. Έτσι είναι πιο δύσκολη στο χειρισμό της και σε περίπτωση εκπαίδευσης χρειάζεται πολύ περισσότερα δεδομένα.
  2. Οι δύο κανόνες μαθαίνονται χωριστά και ίσως και διαφορετικά. Στη διαδικασία εκμάθησης θα είναι ανταγωνιστικοί, αν και αντιστοιχούν στην ίδια κατάσταση.
  3. Είναι πιο δύσκολο για έναν έμπειρο χρήστη να προσδιορίσει μια πλήρη και συνεπή βάση γνώσης χρησιμοποιώντας τέτοιους κανόνες

Το πρόβλημα των προαιρετικών όρων μπορεί να λυθεί αποτελεσματικά και με τρόπο πολύ φυσικό χρησιμοποιώντας κανόνες της μορφής:

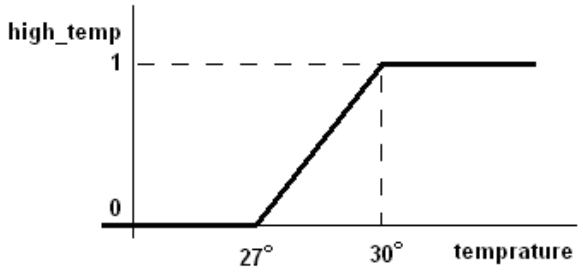
$$\begin{aligned} &IF\ temprature\ IS\ high\_temp \\ &THEN\ it\_feels\_hot\ IS\ true \\ &MORE\ SO\ IF\ humidity\ IS\ high\_hum \end{aligned} \tag{7.5}$$

Έτσι, επεκτείνουμε την κλασική δομή κανόνα της σχέσης 7.1 ως εξής:

$$IF\ x_1, x_2 \dots x_n\ THEN\ y\ MORE\ SO\ IF\ x_{n+1}, x_{n+2} \dots x_m \tag{7.6}$$

όπου τα $x_{n+1}, x_{n+2} \dots x_m$ είναι οι όροι που αντιστοιχούν στο προαιρετικό μέρος του κανόναι και, όμοια με τους όρους $x_1, x_2 \dots x_n$, ακολουθούν τη μορφή:

$$x_i : f_i\ IS\ X_i \tag{7.7}$$



Σχήμα 7.1: *Η λεκτική μεταβλητή high_temp.*

7.3 Δυνατοτική αποτίμηση

Στο παράδειγμα με τη θερμοκρασία, θεωρούμε πως η μέτρηση της θερμοκρασίας είναι διαθέσιμη με απόλυτη βεβαιότητα και ακρίβεια, ώστε με τη χρήση του σχήματος 7.1 να μπορούμε να της ποσολογήσουμε. Αν και αυτό είναι λογική υπόθεση για εφαρμογές που οδηγούνται από αισθητήρες, όταν η εφαρμογή οδηγείται από την έξοδο κάποιου άλλου περίπλοκου, ασαφούσ και αβέβαιου συστήματος είναι πιθανό να καταρρέει.

Σε πολλές πραγματικές εφαρμογές, όπως είναι η ανάλυση προσώπου που θα παρουσιάσουμε στην ενότητα με τα πειραματικά αποτελέσματα, μια σειρά από ευαίσθητα θέματα πρέπει να εξεταστούν, όπως:

- κάποιοι όροι είτε δεν μπορούν να εκτιμηθούν ή η εκτίμησή τους γίνεται με μεγάλη αβεβαιότητα
- είναι πιθανό να ενεργοποιούνται αντιφατικοί κανόνες

Μια τυπική μεθοδολογία για την αποτίμηση ασαφών κανόνων της μορφής της εξίσωσης 7.1 είναι η

$$y = t(x_1, x_2 \dots x_n) \tag{7.8}$$

όπου t είναι μια ασαφής t-νόρμα, όπως η minimum

$$t(x_1, x_2 \dots x_n) = min(x_1, x_2 \dots x_n) \tag{7.9}$$

το αλγεβραϊκό γινόμενο

$$t(x_1, x_2 \dots x_n) = x_1 \cdot x_2 \cdot \dots \cdot x_n \tag{7.10}$$

η φραγμένη διαφορά

$$t(x_1, x_2 \dots x_n) = x_1 + x_2 + \dots + x_n + 1 - n \tag{7.11}$$

κλπ. Άλλη μια δημοφιλής μεθοδολογία αποτίμησης ασαφών κανόνων παρουσιάζεται στο [87] και χρησιμοποιεί τη σταθμισμένη μέση τιμή αντί για μια t-νόρμα για να συνδυάσει την πληροφορία των διαφορετικών όρων του κανόνα.

$$y = w_1 x_1 + w_2 x_2 + \dots + w_n x_n \tag{7.12}$$

Και οι δύο μεθοδολογίες έχουν μελετηθεί εκτενώς στο χώρο του ασαφούς αυτομάτου ελέγχου. Παρόλαυτά δεν επαρκούν για προβλήματα όπως η ανάλυση προσώπου για την εκτίμηση τής έκφρασης και του συναισθήματος. Το βασικό τους μειονέκτημα

είναι πως υποθέτουν πως όλοι οι όροι είναι γνωστοί, δηλαδή πως όλα τα χαρακτηριστικά μετρώνται με επιτυχία και ακρίβεια. Στο παράδειγμα του προσώπου, σημεία μπορεί να εντοπίζονται με μικρή βεβαιότητα ή να μην εντοπίζονται καθόλου. Έτσι απαιτείται μια πιο ευέλικτη μεθοδολογία αποτίμησης κανόνων που θα μπορεί να χειριστεί αυτή την αβεβαιότητα.

Επιπρόσθετα, η μεθοδολογία της σχέσης 7.12, εξαιτίας της αθροιστικής μορφής, μπορεί να έχει υψηλές ενεργοποιήσεις στην έξοδό της ακόμη και όταν ένας απαραίτητος όρος λείπει από την είσοδο. Προφανώς αυτό δεν είναι αποδεκτό για τον τύπο των προβλημάτων που θεωρούμε σε αυτό το κεφάλαιο. Για παράδειγμα, η δεδομένη μη ενεργοποίηση ενός συγκεκριμένου χαρακτηριστικού του προσώπου μπορεί να είναι αρκετή για να αποκλείσει μια σειρά από πιθανές εκφράσεις με μεγάλη βεβαιότητα. Έτσι, το ιδανικό μοντέλο αποτίμησης, το οποίο και θα επεκτείνουμε, είναι αυτό της σχέσης 7.8.

7.3.1 Πιθανοτική αποτίμηση απαραίτητων όρων

Στην πράξη τομής της σχέσης 7.8, οι όροι που έχουν πιο μικρές τιμές επηρεάζουν περισσότερο την τιμή του αποτελέσματος, ενώ οι όροι με τιμή κοντά στη μονάδα την επηρεάζουν λιγότερο ή καθόλου. Έχοντας αυτό στο μυαλό μας, μπορούμε να απαιτήσουμε μόνο όροι που είναι γνωστοί με μεγάλη βεβαιότητα να επιτρέπεται να έχουν μικρή τιμή σε αυτή την πράξη. Πιο αυστηρά, απαιτούμε ο βαθμός $k(x)$ με τον οποίο ο όρος x συμμετέχει στην πράξη να είναι μικρός μόνο όταν η τιμή του x είναι μικρή και συγχρόνως η βεβαιότητα x^c με την οποία γνωρίζουμε αυτή την τιμή είναι μεγάλη. Ανάλογα με την εφαρμογή, ο βαθμός εμπιστοσύνης x^c μπορεί να προσφέρεται απευθείας από έναν αισθητήρα ή, πιθανότατα, να υπολογίζεται από ένα βήμα ελέγχου της εγκυρότητας των αποτελεσμάτων μέσω αντιπαραβολής με κατάλληλα κριτήρια [135]. Αυτό μπορεί να εκφραστεί ως:

$$c(k(x)) = t(x^c, c(x)) \tag{7.13}$$

όπου c ένα ασαφές συμπλήρωμα. Εφαρμόζοντας τον κανόνα του de Morgan έχουμε πως ο βαθμός στον οποίο ο όρος θα πρέπει να θεωρηθεί είναι:

$$k(x) = s(c(x^c), x) \tag{7.14}$$

όπου s μια s-νόρμα.

Εύκολα βλέπουμε πως η σχέση 7.14 ικανοποιεί τις επιθυμητές οριακές συνθήκες:

- Όταν $x^c \to 1$, τότε $k(x) \to x$, δηλαδή ο όρος χρησιμοποιείται κανονικά.
- Όταν $x^c \to 0$, τότε $k(x) \to 1$, δηλαδή ο όρος δεν επηρεάζει το αποτέλεσμα.

Σε αυτή τη συζήτηση θεωρούμε πως η αποτίμηση του κανόνα γίνεται με βάση τη σχέση

$$y = t(k(x_1), k(x_2) \dots k(x_n)) \tag{7.15}$$

Έυκολα βλέπουμε πως αν όλοι οι όροι είναι γνωστοί με βεβαιότητα 1, ο κανόνας θα αποτιμηθεί με τον ίδιο ακριβώς τρόπο όπως και με την κλασική μεθοδολογία. Όταν, όμως, όλοι οι όροι είναι άγνωστοι, ή ισοδύναμα γνωστοί με βεβαιότητα 0, ο κανόνας θα αποτιμηθεί σε βαθμό 1. Προφανώς λοιπόν η έννοια της ενεργοποίησης

είναι διαφορετική στη μέθοδό μας. Μπορεί να ερμηνευτεί με ένα δυνατοτικό τρόπο, δηλαδή ως μια εκτίμηση της δυνατότητας η αντίστοιχη έξοδος να ισχύει. Αυτό το δυνατοτικό μέτρο είναι γνωστό ως εφικτότητα [73].

Όσο αφορά στην εμπιστοσύνη στην έξοδο του συστήματος, στην κλασική μέθοδο, όπου υποθέτουμε απόλυτη εμπιστοσύνη στην είσοδο, θεωρούμε απλά πως οι βαθμοί ενεργοποίησης είναι σωστοί. Στην επέκταση που παρουσιάζουμε εδώ, όπου δεχόμαστε το ενδεχόμενο η είσοδος να είναι ατελής και αβέβαιη, δεν είναι λογικό να υποθέτουμε πως αυτή η αβεβαιότητα δεν διαδίδεται και στην έξοδο. Αντίθετα, το εκτιμώμενο επίπεδο ενεργοποίησης y είναι πληροφοριακά πλήρες μόνο όταν συνοδεύεται και από τον αντίστοιχο βαθμό βεβαιότητας y^c.

Η εμπιστοσύνη στην έξοδο υπολογίζεται με βάση την εμπιστοσύνη στους διάφορους όρους της εισόδου ως εξής:

$$y^c = \frac{x_1^c + x_2^c + \cdots + x_n^c}{n} \tag{7.16}$$

Ο υπολογισμός του y^c με αυτό τον τρόπο έχει το επιθυμητό αποτέλεσμα η τιμή $y^c = 0$ να αντιστοιχεί στην απόλυτη έλλειψη πληροφορίας, κάτι είναι βασικό στη δυνατοτική συλογιστική.

7.3.2 Πιθανοτική αποτίμηση προαιρετικών όρων

Η παραπάνω μεθοδολογία είναι ικανή να χειριστεί κανόνες της μορφής της σχέσης 7.1, αλλά και το πρώτο τμήμα κανόνων της μορφής που παρουσιάζεται στη σχέση 7.6. Όσο αφορά στην αποτίμηση της συνεισφοράς του προαιρετικού τμήματος του κανόνα, αυτή:

- Δεν θα πρέπει να μπορεί να επηρεάσει την εμπιστοσύνη στην έξοδο του κανόνα η οποία καθορίζεται από τους απαραίτητους όρους.
- Δεν θα πρέπει να είναι σε θέση να μειώσει το βαθμό ενεργοποίησης του κανόνα σε σχέση με το ενδεχόμενο ο προαιρετικός όρος να μην ήταν καθόλου διαθέσιμος.

Έτσι οι προαιρετικοί όροι επιδρούν σαν ασθενείς modifiers [73], μόνο αυξάνοντας το επίπεδο ενεργοποίησης. Βέβαια, ένας προαιρετικός όρος μπορεί να επηρεάσει σημαντικά το βαθμό ενεργοποίησης του κανόνα μόνο αν και η δική του τιμή είναι υψηλή και είναι γνωστός με μεγάλη βεβαιότητα. Πιο αυστηρά, ο βαθμός $l(x)$ στον οποίο ο προαιρετικός όρος x συμμετέχει στον υπολογισμό της τελικής εξόδου y' υπολογίζεται ως:

$$l(x) = t(x, x^c) \tag{7.17}$$

Ένας προαιρετικός όρος x, λοιπόν, αυξάνει το βαθμό ενεργοποίησης y που υπολογίζεται με βάση τους απαραίτητους όρους ως εξής:

$$y' = H(y, l(x)) \tag{7.18}$$

όπου H είναι ένας παραμετροποιημένος ασθενής modifier. Ο modifier εφαρμόζεται στην τιμή y, ενώ η τιμή $l(x)$ είναι η παράμετρος που χρησιμοποιείται ώστε να μεγιστοποιεί την επίδραση της πράξης όταν παίρνει μεγάλες τιμές και να την ακυρώνει όταν

παίρνει την τιμή μηδέν. Για παράδειγμα, εδώ χρησιμοποιούμε τον ακόλουθο modifier:

$$y' = H(a, b) = a^{\frac{1}{1+b}} \tag{7.19}$$

Για τη θεώρηση πολλών προαιρετικών όρων οι αντίστοιχοι modifiers εφαρμόζονται διαδοχικά στην υπολογισμένη έξοδο.

Συνοψίζοντας, ο βαθμός ενεργοποίησης y και ο βαθμός εμπιστοσύνης y^c στην αποτίμηση ενός κανόνα της μορφής της σχέσης 7.6 (θεωρώντας βαθμούς αβεβαιότητας για τους όρους εισόδου και τροποποιώντας ελαφρά τους συμβολισμούς) δίνονται από τις σχέσεις:

$$y = y_m \tag{7.20}$$

$$y_i = H(y_{i-1}, l(x_i)), i \in m, m-1, \ldots (n+1) \tag{7.21}$$

$$H(a, b) = a^{\frac{1}{1+b}} \tag{7.22}$$

$$l(x_i) = t(x_i, x_i^c), i \in m, m-1, \ldots (n+1) \tag{7.23}$$

$$y_n = t(k(x_1), k(x_2) \ldots k(x_n)) \tag{7.24}$$

$$k(x_i) = s(c(x_i^c), x_i), i \in 1, 2, \ldots n \tag{7.25}$$

$$y^c = \frac{x_1^c + x_2^c + \cdots + x_n^c}{n} \tag{7.26}$$

7.3.3 Ο δυνατοτικός χαρακτήρας

Όταν χρησιμοποιούμε τη μέθοδο αποτίμησης της σχέσης 7.8 συχνά απομακρύνουμε την ασάφεια απλά επιλέγοντας την έξοδο του κανόνα με την υψηλότερη ενεργοποίηση και αγνοώντας τους υπόλοιπους κανόνες. Αυτό σημαίνει πως σιωπηρά δίνουμε στους βαθμούς ενεργοποίησης μια πιθανοτική χρειά. Αυτό φαίνεται και από το ότι:

- Ο κανόνες με την υψηλότερη ενεργοποίηση θεωρείται πιο πιθανός από τους άλλους.
- Οι περιπτώσεις στις οποίες δύο διαφορετικοί κανόνες ενεργοποιούνται σε υψηλά και παρόμοια επίπεδα δεν θεωρούνται ξεκάθαρες και οι δύο κανόνες θεωρούνται σχεδόν εξίσου πιθανοί.

Από την άλλη πλευρά αξίζει να σημειωθεί πως η έξοδος ενός συσήματος κανόνων, αν και έχει μια πιθανοτική χρειά, δεν μπορεί να χρησιμοποιηθεί σαν πιθανότητα, καθώς δεν ικανοποιεί τον αξιωματικό ορισμό της πιθανότητας. Για παράδειγμα, το άθροισμα των ενεργοποιήσεων μη συμβατών κανόνων δύναται να ξεπερνά τη μονάδα. Όμοια, δεν μπορούμε να ισχυριστούμε πως η προτεινόμενη μεθοδολογία αποτίμησης δίνει στην έξοδό της δυνατοτικά μέτρα. Έχει όμως ξακάθαρα μια δυνατοτική χροιά.

Η ενεργοποίηση y του κανόνα έχει μια δυνατοτική ερμηνεία που αντιστοιχεί στην εφικτότητα του κανόνα. Για να έχουμε πλήρη δυνατοτική αναπαράσταση ενός ενδεχομένου όπως είναι η ισχύς ενός κανόνα, μαζί με την εφικτότητα πρέπει να έχουμε και την αντίστοιχη πίστη, δηλαδή το βαθμό στον οποίο τα δεδομένα εισόδου υποστηρίζουν την ενεργοποίηση του κανόνα.

Το μέτρου της πίστης θα πρέπει να είναι υψηλό όταν υπάρχει επαρκής πληροφορία και αυτή η πληροφορία συντείνει στην ενεργοποίηση του κανόνα. Η ποσότητα

Σχήμα 7.2: *Το καρέ 39308.*

της πληροφορίας που είναι διαθέσιμη ποσολογείται από την εμπιστοσύνη y^c στον κανόνα, ενώ ο βαθμός στον οποίο αυτή η πληροφορία υποστηρίζει τον κανόνα από την ενεργοποίηση y. Έτσι, η πλήρης δυνατοτική έξοδος λαμβάνεται ως:

$$Bel = t(y, y^c) \tag{7.27}$$

$$Pl = y \tag{7.28}$$

Η πιθανότητα ενεργοποίησης πολλαπλών και αντιφατικών κανόνων του συστήματος δεν είναι προβληματική για αυτή την προσέγγιση της αποτίμησης ασαφών κανόνων. Σε αυτή την περίπτωση αναμένεται οι τιμές εμπιστοσύνης να έιναι χαμηλές, και έτσι η εφικτότητα υψηλή, δείχνοντας πως πολλαπλοί κανόνες δεν μπορούν να αποκλειστούν με βάση τη διαθέσιμη πληροφορία. Αντίθετα, οι τιμές Bel, όπως αναφέρονται από τη σχέση 7.27, αναμενεται να είναι χαμηλές, δείχνοντας πως η διαθέσιμη πληροφορία δεν αρκεί για να θεωρήσουμε πως κάποιος από τους κανόνες ενεργοποιείται με βεβαιότητα.

7.4 Πειραματικά αποτελέσματα

Η μεθοδολογία που παρουσιάσαμε σε αυτό το κεφάλαιο εφαρμόστηκε στην πράξη και δοκιμάστηκε στο πλαίσιο του IST προγράμματος ERMIS. Ο στόχος του προγράμματος ERMIS είναι η ανάπτυξη ενός συστήματος για αλληλεπίδραση μεξαξύ υπολογιστή και ανθρώπου που να λαμβάνει υπόψη τη συναισθηματική κατάσταση του χρήστη, μέσω της εξέτασης των εκφράσεων του προσώπου και της στάσης του σώματος.

Η εκτίμηση των εκφράσεων του προσώπου γίνεται μέσω της αποτίμησης ασαφών κανόνων, με όρους εισόδου που προέρχονται από την αυτόματη ανάλυση των εικόνων

Σχήμα 7.3: *Το καρέ 39308 μετά από επεξεργασία.*

Πίνακας 7.1: *Οι τιμές των λεκτικών μεταβλητών στόματος και σαγονιού για το καρέ 39308.*

λεκτική μεταβλητή	τιμή	βεβαιότητα
open_jaw_low	0	0.72
open_jaw_medium	0.58	0.72
open_jaw_high	0.42	0.72
lower_top_midlip_low	0	0.72
lower_top_midlip_medium	1	0.72
raise_bottom_midlip_verylow	1	0.72
raise_bottom_midlip_low	0.34	0.72
raise_bottom_midlip_high	0	0.72
widening_mouth_low	0.5	0.72
widening_mouth_medium	0.43	0.72
widening_mouth_high	0	0.72
raise_left_outer_cornerlip_low	1	0.72
raise_left_outer_cornerlip_medium	0.27 0.72	
raise_left_outer_cornerlip_high	0	0.72
raise_right_outer_cornerlip_low	1	0.724562
raise_right_outer_cornerlip_medium	0	0.724562
raise_right_outer_cornerlip_high	0	0.724562

Πίνακας 7.2: *Οι τιμές των λεκτικών μεταβλητών ματιών και φρυδιών για το καρέ 39308.*

λεκτική μεταβλητή	τιμή	βεβαιότητα
raise_right_medium_eyebrow_medium	0	0.87
raise_right_medium_eyebrow_low	0	0.87
raise_right_medium_eyebrow_high	1	0.87
raise_left_outer_eyebrow_low	0	0.7
raise_left_outer_eyebrow_medium	0	0.71
raise_left_outer_eyebrow_high	1	0.7
raise_right_outer_eyebrow_low	0	0.87
raise_right_outer_eyebrow_medium	0	0.87
raise_right_outer_eyebrow_high	1	0.87
squeeze_left_eyebrow_low	1	0.7
squeeze_left_eyebrow_medium	0	0.7
close_left_eye_low	0	0.55
squeeze_left_eyebrow_high	0	0.7
close_left_eye_high	1	0.55
squeeze_right_eyebrow_low	1	0.87
close_right_eye_low	0	0.33
squeeze_right_eyebrow_medium	0	0.87
close_right_eye_high	0.89	0.33
squeeze_right_eyebrow_high	0	0.87
raise_left_inner_eyebrow_low	0	0.7
wrinkles_between_eyebrows_low	1	0.7
raise_left_inner_eyebrow_medium	0.086	0.7
wrinkles_between_eyebrows_medium	0	0.7
raise_left_inner_eyebrow_high	1	0.7
wrinkles_between_eyebrows_high	0	0.7
raise_right_inner_eyebrow_low	0	0.87
raise_right_inner_eyebrow_medium	0 0.87	
raise_right_inner_eyebrow_high	1	0.87
raise_left_medium_eyebrow_low	0	0.7
raise_left_medium_eyebrow_medium	0.27	0.7
raise_left_medium_eyebrow_high	0.85	0.7

Πίνακας 7.3: *Συνολική έξοδος συστήματος ασαφών κανόνων.*

Τεταρτημόριο	Ground truth	Κλασική μέθοδος	Bel	Pl
1	1	0	0.21608	0.3015
2	0	0	0.06160	0.09135
3	0	0	0.00238	0.00352
Ουδέτερο	0	0	<0.00001	<0.00001

προσώπου. Καθώς οι εικόνες είναι συχνά χαμηλής ποιότητας, υπάρχει θόρυβος, ο φωτισμός δεν είναι πάντα σταθερός, υπάρχει κάποιες φορές απόκρυψη των χαρακτηριστικών που αναζητάμε και οι όροι εισόδου δεν είναι πάντα διαθέσιμοι με ακρίβεια και βεβαιότητα. Έτσι, το σύστημα αναγνώρισης έκφρασης είναι ένα ιδανικό πεδίο για τον πειραματικό έλεγχο της προτεινόμενης μεθοδολογίας δυνατοτικής αποτίμησης ασαφών κανόνων.

Όλοι οι κανόνες του συστήματος περιέχουν μεγάλο αριθμό όρων εισόδου, με αποτέλεσμα η πιθανότητα να λείψει κάποιο χαρακτηριστικό και οι κανόνες να μην ενεργοποιούνται είναι μεγάλη. Στο σύστημα έχουν δοκιμαστεί τόσο η κλασική όσο και η προτεινόμενη μεθοδολογία αποτίμησης των κανόνων, και τα αποτελέσματα έχουν αξιολογηθεί με βάση ground truth που έχει δημιουργηθεί από ανθρώπους που παρατηρώντας τις ίδιες εικόνες προσώπου έδωσαν τις δικές τους εκτιμήσεις για τη συναισθηματική κατάσταση του παρατηρούμενου ανθρώπου.

Σε ένα σύνολο 30000 χαρακτηριστικών καρέ, η κλασική μεθοδολογία αποτίμησης κατηγοριοποιεί σωστά τις εκφράσεις σε ποσοστό 65.1% ενώ η προτεινόμενη δυνατοτική αποτίμηση σε ποσοστό 78.4%.

Για να δείξουμε καλύτερα τον τρόπο με τον οποίο λειτουργεί η μεθοδολογία σε αβέβαιο περιβάλλον παρουσιάζουμε τα αποτελέσματα από ένα καρέ αναλυτικά. Είναι το καρέ 39308 της ακολουθίας RD, που παρουσιάζεται στο σχήμα 7.2. Η συγκεκριμένη εικόνα προέρχεται από τη βάση δεδομένων ERMIS [154]. Μετά από την επεξεργασία της εικόνας τα σημεία που παρουσιάζονται στο σχήμα 7.3 επιλέγονται ως τα οριακά σημεία των βασικών χαρακτηριστικών του προσώπου, και με βάση αυτά θα πρέπει να εκτιμηθεί η διάθεση του ατόμου. Για αυτό το καρέ οι αξιολογητές που δημιούργησαν το ground truth εκτίμησαν πως η διάθεση ανήκει στο πρώτο τεταρτημόριο του τροχού των συναισθημάτων, δηλαδή πως η διάθεση είναι γενικά θετική.

Μπορούμε να παρατηρήσουμε στο σχήμα 7.3 πως αν και η αρχική εικόνα είναι διαθέσιμη σε υψηλή ανάλυση, χωρίς θόρυβο και αποκρύψεις χαρακτηριστικών, ένα από τα σημεία για τα μάτια δεν έχει εντοπιστεί σωστά. Οι τιμές των λεκτικών μεταβλητών που χρησιμοποιούνται στον ορισμό των κανόνων παρουσιάζονται στους πίνακες 7.1 και 7.2.

Καθώς όλοι οι κανόνες του συστήματος είναι συμμετρικοί, αυτό έχει σαν αποτέλεσμα κανένας από τους κανόνες να μην ενεργοποιείται όταν ακολουθείται η κλασική μέθοδος αποτίμησης. Με την προτεινόμενη μέθοδο ενεργοποιούνται 3 κανόνες. Αυτοί οι κανόνες είναι:

IF close_left_eye IS close_left_eye_low
AND close_right_eye IS close_right_eye_low
AND raise_left_inner_eyebrow IS raise_left_inner_eyebrow_high
AND raise_right_inner_eyebrow IS raise_right_inner_eyebrow_high
AND raise_left_medium_eyebrow IS raise_left_medium_eyebrow_high
AND raise_right_medium_eyebrow IS raise_right_medium_eyebrow_high
AND raise_left_outer_eyebrow IS raise_left_outer_eyebrow_high
AND raise_right_outer_eyebrow IS raise_right_outer_eyebrow_high
AND squeeze_left_eyebrow IS squeeze_left_eyebrow_low
AND squeeze_right_eyebrow IS squeeze_right_eyebrow_low
THEN output IS quadrant_1
με $y = 0.3015$ και $y^c = 0.714883$

IF close_left_eye IS close_left_eye_low
AND close_right_eye IS close_right_eye_low
AND raise_left_inner_eyebrow IS raise_left_inner_eyebrow_high
AND raise_right_inner_eyebrow IS raise_right_inner_eyebrow_high
AND raise_left_medium_eyebrow IS raise_left_medium_eyebrow_high
AND raise_right_medium_eyebrow IS raise_right_medium_eyebrow_high
AND raise_left_outer_eyebrow IS raise_left_outer_eyebrow_high
AND raise_right_outer_eyebrow IS raise_right_outer_eyebrow_high
THEN output IS quadrant_1
με $y = 0.3015$ και $y^c = 0.716673$

IF close_left_eye IS close_left_eye_low
AND close_right_eye IS close_right_eye_low
AND raise_left_inner_eyebrow IS raise_left_inner_eyebrow_high
AND raise_right_inner_eyebrow IS raise_right_inner_eyebrow_high
AND raise_left_medium_eyebrow IS raise_left_medium_eyebrow_high
AND raise_right_medium_eyebrow IS raise_right_medium_eyebrow_high
AND raise_left_outer_eyebrow IS raise_left_outer_eyebrow_high
AND raise_right_outer_eyebrow IS raise_right_outer_eyebrow_high
AND squeeze_left_eyebrow IS squeeze_left_eyebrow_low
AND squeeze_right_eyebrow IS squeeze_right_eyebrow_low
AND wrinkles_between_eyebrows IS wrinkles_between_eyebrows_low
THEN output IS quadrant_1
με $y = 0.3015$ και $y^c = 0.69938$

Τα αποτελέσματα των δύο μεθόδων για αυτό το καρέ συνοψίζονται στον πίνακα 7.3. Βλέπουμε αμέσως πρώτα απ'όλα πως η προτεινόμενη μεθοδολογία λειτουργεί ενώ για τα ίδια δεδομένα η κλασική προσέγγιση δεν δίνει καμία έξοδο. Επιπρόσθετα, αν και η αβεβαιότητα στην είσοδο είναι δεδομένη, το σύστημα σωστά αναγνωρίζει πως η συναισθηματική κατάσταση που παρατηρούμε ανήκει στο πρώτο τεταρτημόριο, καθώς μόνο για αυτό το τεταρτημόριο έχουμε υψηλές τιμές για τα Bel και Pl.

□

Κεφάλαιο 8

Ιεραρχική ομαδοποίηση

8.1 Εισαγωγή

Η ομαδοποίηση των δεδομένων είναι ένα πρόβλημα που συσχετίζεται με πολυάριθμους ερευνητικούς και εφαρμοσμένους τομείς [66]. Αν και ερευνητές στον τομέα της εξόρυξης δεδομένων έχουν εργαστεί σε αυτήν την κατεύθυνση για μεγάλο χρονικό διάστημα, και πολυάριθμα σχετικά κείμενα υπάρχουν στη βιβλιογραφία, αυτό θεωρείται ακόμα ένα ανοικτό ζήτημα, το οποίο είναι δύσκολο να αντιμετωπιστεί, ιδιαίτερα στις περιπτώσεις που τα δεδομένα χαρακτηρίζονται από πολυάριθμα μετρήσιμα χαρακτηριστικά. Αυτό αναφέρεται συχνά ως κατάρα των υψηλών διαστάσεων.

Εργασίες στον τομέα της ταξινόμησης εστιάζουν στη χρήση χαρακτηρισμένων δεδομένων, γνωστών επίσης και ως δεδομένων εκπαίδευσης, για την αυτόματη παραγωγή συστημάτων που είναι σε θέση να ταξινομήσουν (χαρακτηρίσουν) μελλοντικά δεδομένα. Αυτή η διαδικασία στηρίζεται στην ομοιότητα των εισερχόμενων δεδομένων με τα δεδομένα εκπαίδευσης.

Τυπικά, προκειμένου να επιτευχθεί ένας τέτοιος στόχος, χρειάζεται πρώτα κανείς να ανιχνεύσει τα πρότυπα που υποβόσκουν στα δεδομένα, και έπειτα να μελετήσει τον τρόπο με τον οποίο τα πρότυπα αυτά σχετίζονται με τις υπάρχουσες κλάσεις. Ακόμα και χρησιμοποιώντας αυτοεκπαιδευόμενα συστήματα, όπως νευρωνικά δίκτυα ανακατανομής πόρων, τα οποία είναι σε θέση να προσαρμοστούν δυναμικά στα δεδομένα εκπαίδευσης, καλά αποτελέσματα μπορούν μόνο να επιτευχθούν όταν τα πρότυπα είναι γνωστά από πριν, έτσι ώστε να μπορούν να χρησιμοποιηθούν για κατάλληλη αρχικοποίηση [65].

Αν και οι στόχοι της ταξινόμησης και της ομαδοποίησης είναι πολύ συγγενικοί, υπάρχει μια σημαντική διαφορά ανάμεσά τους: ενώ κατά την ταξινόμηση βασικός σκοπός είναι η διάκριση μεταξύ των κατηγοριών, δηλ. η ανίχνευση των ορίων της κατηγορίας, κατά την ομαδοποίηση βασικός σκοπός είναι συνήθως ο προσδιορισμός των χαρακτηριστικών των ομάδων. Το τελευταίο αντιμετωπίζεται συνήθως μέσω επιλογής αντιπροσωπευτικών δειγμάτων ή εικονικών κέντρων των ομάδων, ή μέσω της εξαγωγής (ασαφών) κανόνων [66].

Αποδοτικές λύσεις έχουν προταθεί στη βιβλιογραφία και για τους δύο στόχους, για την περίπτωση στην οποία καθορίζεται ένα μοναδικό μέτρο ομοιότητας ή απόστασης μεταξύ των δεδομένων [122]. Όταν, αντίθετα, πολλαπλά και ανεξάρτητα χαρακτηριστικά χαρακτηρίζουν τα δεδομένα, και έτσι μπορούν να καθοριστούν περισσότερα από ένα μέτρα ομοιότητας ή απόστασης, η επίλυση και των δύο προβλημάτων γίνεται πιο δύσκολη. Μια συνήθης προσέγγιση στο πρόβλημα είναι η μείωση των διαστάσεων ει-

σόδου, που μπορεί να επιτευχθεί αγνοώντας μερικά από τα διαθέσιμα χαρακτηριστικά (επιλογή χαρακτηριστικών) ή με την εφαρμογή κάποιου χωρικού μετασχηματισμού.

Στην περίπτωση που τα χαρακτηριστικά εισόδου δεν είναι ανεξάρτητα μεταξύ τους, μια μείωση των διαστάσεων είναι πολύ χρήσιμη. Από την άλλη μεριά, όταν τα χαρακτηριστικά εισόδου είναι ανεξάρτητα, ή όταν η σχέση μεταξύ τους δεν είναι γνωστή εκ των προτέρων, όπως συχνά συμβαίνει με τα πραγματικά δεδομένα, μείωση των διαστάσεων του χώρου δεν μπορεί να επιτευχθεί χωρίς απώλεια πληροφορίας. Επομένως, εάν η σχέση μεταξύ των χαρακτηριστικών δεν είναι γνωστή από πριν, και ο στόχος είναι να ανιχνευθούν τα πρότυπα που υπάρχουν στα δεδομένα, η μείωση των διαστάσεων δεν είναι δυνατή. Επιπλέον, οι διαφορές στην κλίμακα μέτρησης μεταξύ των διαφορετικών χαρακτηριστικών δυσχεραίνει τη διαδικασία.

8.2 Γενική δομή αλγορίθμου

Οι περισσότερες μέθοδοι ομαδοποίησης ανήκουν σε μία από τις δύο γενικές κατηγορίες, τις διαμεριστικές ή τις ιεραρχικές μεθόδους. Οι διαμεριστικές μέθοδοι δημιουργούν μια σαφή ή ασαφή ομαδοποίηση ενός δεδομένου συνόλου δεδομένων, αλλά απαιτούν τον αριθμό των ομάδων ως είσοδο. Όταν το πλήθος των προτύπων που υπάρχουν στα δεδομένα δεν είναι γνωστό εκ των προτέρων, οι διαμεριστικές μέθοδοι ομαδοποίησης είναι μη εφαρμόσιμες. Σε αυτές τις περιπτώσεις χρειάζεται να εφαρμοστεί ένας ιεραρχικός αλγόριθμος ομαδοποίησης. Η γενική δομή τους είναι η ακόλουθη [95]:

1. Μετέτρεψε κάθε στοιχείο εισόδου σε μια ομάδα ενός μόνο στοιχείου.
2. Για κάθε ζευγάρι ομάδων c_1, c_2 υπολόγισε έναν μια απόσταση $d(c_1, c_2)$.
3. Συγχώνευσε το ζευγάρι των ομάδων που έχουν τη μικρότερη απόσταση.
4. Συνέχισε στο βήμα 2, μέχρι την ικανοποίηση του κριτηρίου τερματισμού. Το κριτήριο τερματισμού που χρησιμοποιείται πιο συχνά είναι ο καθορισμός ενός κατωφλίου για την τιμή της απόστασης μεταξύ των ομάδων, ενώ καλά αποτελέσματα έχουν επιτευχθεί και μέσω της χρήσης κατωφλίου για το ρυθμό αύξησης της απόστασης.

Τα δύο βασικά σημεία που διαφοροποιούν τις ιεραρχικές μεθόδους μεταξύ τους, και καθορίζουν την αποδοτικότητά τους, είναι ο ορισμός της απόστασης μεταξύ των ομάδων και το κριτήριο τερματισμού που χρησιμοποιείται. Σημαντικά μειονεκτήματα των ιεραρχικών μεθόδων είναι η υψηλή πολυπλοκότητα, η ευαισθησία στα λάθη στα αρχικά βήματα, τα οποία διαδίδονται μέχρι το τελικό αποτέλεσμα, και η ευαισθησία στη σειρά με την οποία παρουσιάζονται τα δεδομένα.

Ο πυρήνας του ανωτέρω γενικού αλγορίθμου είναι η δυνατότητα να καθοριστεί ένα μοναδικό μέτρο απόστασης μεταξύ οποιουδήποτε ζευγαριού ομάδων. Επομένως, όταν ο χώρος εισόδου έχει περισσότερες από μία διαστάσεις, μια αθροιστική συνάρτηση απόστασης χρησιμοποιείται συνήθως ως d [139].

$$d(c_1, c_2) = \sqrt[\kappa]{\frac{\sum_{a \in c_1, b \in c_2} r(a, b)^\kappa}{|c_1||c_2|}} \tag{8.1}$$

όπου r είναι ένα μετρικό στο χώρο των δεδομένων. Μετρικά που χρησιμοποιούνται συχνά είναι η Ευκλείδεια απόσταση

$$r_{Ευκλ}(a,b) = \sqrt{\sum_{i \in N_F} (a_i - b_i)^2} \tag{8.2}$$

όπου F είναι η διάσταση του χώρου των δεδομένων, η δομική απόσταση [16]

$$r_{cityblock}(a,b) = \sum_{i \in N_F} |a_i - b_i| \tag{8.3}$$

η απόσταση minimax, ή αλλιώς απόσταση Chebyshev

$$r_{minimax}(a,b) = max_{i \in N_F} |a_i - b_i| \tag{8.4}$$

η γωνία διανυσμάτων

$$r_{\gamma\omega\nu}(a,b) \frac{\sum_{i \in N_F} a_i \cdot b_i}{\sqrt{\sum_{i \in N_F} a_i^2} \cdot \sqrt{\sum_{i \in N_F} b_i^2}} \tag{8.5}$$

κλπ.

8.3 Ιεραρχική ομαδοποίηση σε υψηλές διαστάσεις

Ο στόχος είναι να αντιμετωπίσουμε την ανίχνευση των προτύπων σε πολυδιάστατα δεδομένα που δεν έχουν προεπεξεργαστεί, όταν το πλήθος των διακριτών προτύπων στα δεδομένα και η σχέση μεταξύ των χαρακτηριστικών εισόδου είναι άγνωστα. Ο προτεινόμενος αλγόριθμος είναι μια επέκταση της ιεραρχικής ομαδοποίησης και είναι βασισμένος σε μια ασαφή επιλογή χαρακτηριστικών που λαμβάνονται υπόψη κατά τη σύγκριση των δεδομένων. Τα αποτελέσματα αυτής της αρχικής ομαδοποίησης βελτιώνονται μέσω ενός βήματος επαναταξινόμησης. Αυτό το βήμα, αν και ανεπίβλεπτο, είναι βασισμένο στις αρχές του ταξινομητή του Bayes. Αυτό το βήμα συμβάλλει επίσης στην πειραματική αξιολόγηση της αποδοτικότητας της μεθόδου.

Καθώς το πλήθος των προτύπων, και συνεπώς το αναμενόμενο πλήθος των διαφορετικών ομάδων δεδομένων, δεν είναι γνωστό εκ των πρότερων, είναι απαραίτητο να εφαρμοστεί μια ιεραρχική μέθοδος ομαδοποίησης [95]. Οι ιεραρχικές μέθοδοι ομαδοποίησης βασίζονται στον ορισμό μιας "λογικής" απόστασης ανάμεσα στα δεδομένα, και με βάση αυτή στον ορισμό αποστάσεων ανάμεσα στις ομάδες δεδομένων. Όταν ο συνολικός αριθμός χαρακτηριστικών είναι υψηλός, μικρές αποστάσεις σε ένα μικρό υποσύνολο τους έχουν ελάχιστη επίδραση στη συνολική απόσταση, όταν χρησιμοποιείται μια συνάθροιση των αποστάσεων σε όλα τα χαρακτηριστικά. Κατά συνέπεια, μόνο όταν εξετάζεται το σωστό υποσύνολο των χαρακτηριστικών, μπορούν να συγκριθούν σωστά τα στοιχεία [119],[48].

Αυτό, φυσικά, δεν έχει πάντα νόημα. Υπάρχουν περιπτώσεις, στις οποίες το πλαίσιο μπορεί να αλλάξει το μέτρο ομοιότητας ή ανομοιότητας που χρησιμοποιείται [136]. Παραδείγματος χάριν, δύο ταινίες μπορούν να συγκριθούν με βάση το θέμα τους, ή τους σκηνοθέτες τους. Σε τέτοιες περιπτώσεις, πριν από τον υπολογισμό των αποστάσεων ανάμεσα στις ομάδες, είναι απαραίτητη μια επιλογή των σημαντικών, και

λογικών, χαρακτηριστικών. Στο παράδειγμα των ταινιών, δύο ταινίες μπορούν να είναι όμοιες η μία με την άλλη όσον αφορά το περιεχόμενό τους ενώ δύο άλλες ταινίες μπορούν να είναι όμοιες όσον αφορά το καστ τους. Με άλλα λόγια, μπορεί να μην είναι δυνατό να επιλεχθεί μια ενιαία μετρική απόσταση, που θα ισχύει σε όλες τις περιπτώσεις.

Επιπλέον, ένα χαρακτηριστικό μπορεί να είναι περισσότερο σημαντικό από άλλα, ενώ όλα τα χαρακτηριστικά είναι χρήσιμα, κάθε ένα στο βαθμό του. Με άλλα λόγια, δεν είναι πάντα δυνατή η επιλογή απόλυτων (σαφών) χαρακτηριστικών. Σε αυτή την εργασία αντιμετωπίζουμε την επιλογή χαρακτηριστικών σύμφωνα με την ακόλουθη αρχή: ενώ περιμένουμε τα στοιχεία ενός δεδομένου συνόλου να έχουν τυχαίες αποστάσεις μεταξύ τους σύμφωνα με τα περισσότερα χαρακτηριστικά, περιμένουμε να έχουν μικρές αποστάσεις σύμφωνα με τα χαρακτηριστικά που τα συσχετίζουν. Στηριζόμενοι σε αυτήν την διαφορά στην κατανομή των τιμών απόστασης μπορούμε να προσδιορίσουμε το πλαίσιο μιας ομάδας, δηλαδή τον υποχώρο στον οποίο η ομάδα καθορίζεται καλύτερα.

Τυπικότερα, έστω c_1 και c_2 δύο ομάδες στοιχείων. Έστω επίσης r_i, $i \in N_F$ η μετρική που συγκρίνει το i-στο χαρακτηριστικό, και F το πλήθος των χαρακτηριστικών, δηλαδή η διάσταση του χώρου εισόδου. Ένα μέτρο απόστασης μεταξύ των δύο ομάδων, θεωρώντας μόνο το i-στο χαρακτηριστικό, δίδεται από την

$$f_i(c_1, c_2) = \sqrt[\kappa]{\frac{\sum_{a \in c_1, b \in c_2} r_i(a_i, b_i)^\kappa}{|c_1||c_2|}} \tag{8.6}$$

όπου a_i, b_i είναι το i-στο χαρακτηριστικό των στοιχείων a, b, αντίστοιχα, $|c|$ είναι η πληθικότητα της ομάδας c και $\kappa \in R$ είναι μια σταθερά.

Το πλαίσιο είναι μια ασαφής επιλογή από χαρακτηριστικά που πρέπει να λάβουμε υπόψη όταν υπολογίζουμε μια συνολική τιμή απόστασης. Μπορεί να οριστεί ως ένα ασαφές σύνολο x, ορισμένο στο N_F, με βαθμωτή πληθικότητα ίση με ένα. Τότε η συνολική απόσταση μεταξύ c_1 και c_2 υπολογίζεται ως

$$d(c_1, c_2) = \sum_{i \in N_F} x_i(c_1, c_2)^\lambda \cdot f_i(c_1, c_2) \tag{8.7}$$

όπου x_i είναι ο βαθμός στον οποίο το i, και έτσι και το f_i, περιέχεται στο πλαίσιο, $i \in N_F$ και $\lambda \in R$ είναι μία σταθερά.

Σύμφωνα με την αρχή στην οποία βασιζόμαστε, τα χαρακτηριστικά που σχετίζουν c_1 και c_2 είναι αυτά που παράγουν τις μικρότερες αποστάσεις f_i. Επομένως, το "σωστό" πλαίσιο μπορεί να υπολογιστεί μέσω της λύσης ενός προβλήματος βελτιστοποίησης, ως το πλαίσιο που παράγει τη μικρότερη συνολική απόσταση.

Όταν $\lambda = 1$ η λύση είναι τετριμμένη: το χαρακτηριστικό που παράγει την μικρότερη απόσταση είναι το μόνο το οποίο επιλέγεται. Ο βαθμός στον οποίο επιλέγεται είναι 1. Αν περισσότερα του ενός χαρακτηριστικά παράγουν την καλύτερη απόσταση, εφαρμόζοντας την αρχή της μέγιστης αβεβαιότητας επιλέγονται εξίσου, καθώς δεν υπάρχει κάποια πληροφορία για το ποιο θα πρέπει να προτιμηθεί.

Όταν $\lambda \neq 1$ και $\exists i \in N_F : f_i(c_1, c_2) = 0$, τότε τα χαρακτηριστικά για τα οποία $f_i(c_1, c_2) = 0$ είναι εκείνα τα οποία επιλέγονται (εξίσου).

Όταν $\lambda \neq 1$ και $f_i(c_1, c_2) \neq 0 \forall i \in N_F$, τότε το πρόβλημα βελτιστοποίησης δεν είναι τετριμμένο και πρέπει να επιλυθεί. Σύμφωνα με το ακόλουθο θεώρημα, μπορεί να επιλυθεί αναλυτικά, γεγονός που σημαίνει ότι το πρόβλημα βελτιστοποίησης δεν επηρεάζει την αλγοριθμική πολυπλοκότητα της διαδικασίας:

Θεώρημα

Όταν $\lambda \neq 1$ και $f_i(c_1, c_2) \neq 0 \forall i \in N_F$, τότε το καλύτερο πλαίσιο x, και ισοδύναμα η μικρότερη απόσταση $d(c_1, c_2)$, δίδονται από τις σχέσεις:

$$x_F(c_1, c_2) = \frac{1}{\sum_{i \in N_F} [\frac{f_F(c_1,c_2)}{f_i(c_1,c_2)}]^{\frac{1}{\lambda-1}}} \tag{8.8}$$

$$x_i(c_1, c_2) = x_F(c_1, c_2) \cdot [\frac{f_F(c_1, c_2)}{f_i(c_1, c_2)}]^{\frac{1}{\lambda-1}} \tag{8.9}$$

Απόδειξη

Απαιτήσαμε η βαθμωτή πληθικότητα του πλαισίου να είναι ένα, έτσι το πρόβλημα βελτιστοποίησης με το οποίο έχουμε να καταπιαστούμε είναι περιορισμένο. $|x| = 1$ είναι ισοδύναμο με το $\sum_{i \in N_F} x_i = 1$. Έτσι, αντικαθιστώντας

$$x_F = 1 - \sum_{i \in N_{F-1}} x_i = 1 \tag{8.10}$$

η ελαχιστοποίηση του

$$d(c_1, c_2) = x_F(c_1, c_2)^\lambda \cdot f_F(c_1, c_2) + \sum_{i \in N_{F-1}} x_i(c_1, c_2)^\lambda \cdot f_i(c_1, c_2) \tag{8.11}$$

μετατρέπεται σε ένα πρόβλημα βελτιστοποίησης χωρίς περιορισμούς. Από την 8.10 έχουμε

$$\frac{\partial x_F(c_1, c_2)}{\partial x_i(c_1, c_2} = -1 \forall i \in N_{F-1} \tag{8.12}$$

και έτσι

$$\frac{\partial \{x_F(c_1, c_2)^\lambda \cdot f_F(c_1, c_2)\}}{\partial x_i(c_1, c_2)} = -\lambda \cdot x_F(c_1, c_2)^{\lambda-1} \forall i \in N_{F-1} \tag{8.13}$$

Εύκολα τώρα, απαιτώντας ότι

$$\frac{\partial d(c_1, c_2)}{\partial x_i(c_1, c_2)} = 0 \forall i \in N_F \tag{8.14}$$

έχουμε

$$x_i(c_1, c_2) = x_F(c_1, c_2) \cdot [\frac{f_N(c_1, c_2)}{f_i(c_1, c_2)}]^{1-\lambda} \forall i \in N_{F-1} \tag{8.15}$$

Συνδυάζοντας την 8.15 με την 8.10 έχουμε επίσης

$$x_F(c_1, c_2) = \frac{1}{\sum_{i \in N_F} [\frac{f_F(c_1,c_2)}{f_i(c_1,c_2)}]^{\frac{1}{\lambda-1}}} \tag{8.16}$$

✓

Καθώς το λ αυξάνει αποδίδονται μεγαλύτερες αποστάσεις στα ζευγάρια από ομάδες που σχετίζονται με λιγότερα χαρακτηριστικά, και συνεπώς έχουν μεγαλύτερες τιμές στο πλαίσιό τους. Για να είναι οι αποστάσεις χρησιμοποιήσιμες στη διαδικασία ιεραρχικής ομαδοποίησης πρέπει να είναι απρόσβλητες από την κατεύθυνση των ομάδων σε σχέση με τους άξονες. Κατά συνέπεια, είναι επιτακτικό να μετασχηματίζονται, ώστε να γίνουν άμεσα συγκρίσιμες μεταξύ τους. Χρησιμοποιείται το ακόλουθο, σταθμισμένο μέτρο απόστασης:

$$d_\lambda(c_1, c_2) = \frac{d(c_1, c_2)}{x_\lambda(c_1, c_2)} \tag{8.17}$$

$$x_\lambda(c_1, c_2) = \sum_{i \in N_F} [x_i(c_1, c_2)]^\lambda \tag{8.18}$$

Όταν τα χαρακτηριστικά κβαντίζονται σε ένα μικρό σύνολο επιπέδων, όπως είναι συχνά η περίπτωση με τα ψηφιακά δεδομένα, περιπτώσεις για τις οποίες $f_i(c_1, c_2) = 0$ δεν είναι σπάνιες. Ειδικά στα πρώτα βήματα της ιεραρχικής ομαδοποίησης, όταν οι ομάδες είναι μικρού μεγέθους, οι καλύτερες αποστάσεις είναι σχεδόν πάντα μηδενικές. Καθώς λάθη στα αρχικά βήματα της ιεραρχικής ομαδοποίησης διαδίδονται μέχρι την τελική έξοδο, είναι σημαντικό να γίνεται πάντα η καλύτερη πιθανή επιλογή για το ζευγάρι των ομάδων προς συγχώνευση. Επομένως, ειδικά για την περίπτωση αποστάσεων που είναι μηδενικές, εισάγουμε ένα ακόμα κριτήριο: από όλα τα ζευγάρια για τα οποία $d_\lambda(c_1, c_2) = 0$, επιλέγεται αυτό που έχει μηδενικές αποστάσεις f_i για τα περισσότερα χαρακτηριστικά. Με άλλα λόγια, από όλα τα ζευγάρια παρόμοιων ομάδων, επιλέγονται αυτά που είναι παρόμοια σύμφωνα με το μέγιστο αριθμό χαρακτηριστικών.

Όσο αφορά στο κριτήριο τερματισμού, μπορεί να χρησιμοποιηθεί ένα κατώφλι στην τιμή της απόστασης $d_\lambda(c_1, c_2) = 0$, καθώς η ελάχιστη τιμή της απόστασης d_λ δεν ελαττώνεται καθώς προχωρούμε από το ένα βήμα στο επόμενο.

Με τον τρόπο αυτό, ο αλγόριθμος ομαδοποιεί προοδευτικά τα στοιχεία, σύμφωνα με τις ομοιότητές τους. Για κάθε ομάδα μπορεί να θεωρηθεί ένα διαφορετικό ασαφές υποσύνολο από χαρακτηριστικά για τον υπολογισμό των ομοιοτήτων. Αυτή η ασαφής επιλογή χαρακτηριστικών μπορεί επίσης να γίνει αντιληπτή ως μία επανακλιμάκωση των χαρακτηριστικών, αντικαθιστώντας έτσι το βήμα της προεπεξεργασίας των δεδομένων που παραλείψαμε.

8.4 Βελτίωση ομαδοποίησης και αξιολόγηση επίδοσης μέσω αναταξινόμησης

Ο κύριος στόχος των ιεραρχικών αλγορίθμων ομαδοποίησης δεν είναι τόσο η σωστή ταξινόμηση των δεδομένων, αλλά κυρίως η αναγνώριση των προτύπων που υποβόσκουν σε αυτά. Γι' αυτό, "λανθασμένα" στοιχεία σε ομάδες μπορούν να είναι αποδεκτά, εφ' όσον η συνολική ομάδα περιγράφει σωστά ένα υπάρχον πρότυπο. Αυτό υπονοεί ότι μετρώντας τον ρυθμό ταξινόμησης σε δεδομένα που είναι ήδη σωστά χαρακτηρισμένα μπορεί να μην έχουμε κάνει αρκετά για την αξιολόγηση της πραγματικής αποδοτικότητας ενός αλγορίθμου ομαδοποίησης.

Προκειμένου να εκτιμηθεί πραγματικά ένας αλγόριθμος ομαδοποίησης, πρέπει να εξαχθούν και να εξεταστούν τα πρότυπα που περιγράφονται από τις ανιχνευμένες ομάδες. Εδώ εξετάζουμε κατά πόσον τα ανιχνευμένα πρότυπα έχουν νόημα, εκτιμώντας

ένα ταξινομητή ο οποίος δημιουργείται από αυτά. Από τα πολυάριθμα σχήματα ταξινόμησης που υπάρχουν στην βιβλιογραφία, επιλέγουμε να εργαστούμε με τον Μπεϋζιανό ταξινομητή [108], αν και άλλοι θα μπορούσαν επίσης να έχουν επιλεγεί [83].

Συγκεκριμένα, κάθε ομάδα θεωρείται ότι περιγράφει μια διακριτή κλάση. Επιπλέον, υποθέτουμε ότι όλα τα χαρακτηριστικά των μελών μιας κλάσης ακολουθούν μια κανονική κατανομή. Κατά συνέπεια, χρησιμοποιώντας το εικονικό κέντρο και τις τυπικές αποκλίσεις κάθε ομάδας, μπορούμε να σχεδιάσουμε τη μίξη των γκαουσιανών που περιγράφει την κλάση. Η Μπεϋζιανή προσέγγιση ταξινόμησης υπολογίζει για κάθε στοιχείο εισόδου a τις πιθανότητες $P(p_i/a)$, $i \in N_T$, όπου T είναι το πλήθος των ανιχνευμένων προτύπων, και ταξινομεί το a στο πρότυπο, για το οποίο έχει την μεγαλύτερη πιθανότητα. Οι πιθανότητες υπολογίζονται εύκολα εφαρμόζοντας τον μετασχηματισμό:

$$P(p_i/a) = P(a/p_i)\frac{P(p_i)}{P(a)} \tag{8.19}$$

καθώς η $P(p_i)$ είναι ίση με την σχετική πληθικότητα της ομάδας c_i, δηλαδή

$$P(p_i) = \frac{|c_i|}{\sum_{j \in N_P} |c_j|} \tag{8.20}$$

η $P(a/p_i)$ δίδεται από την τιμή του a στην μίξη των γκαουσσιανών που περιγράφουν το πρότυπο i, δηλαδή

$$P(a/p_i) = \prod_{j \in N_F} \frac{1}{\sqrt{2\pi}s_{ij}} e^{-(\frac{a_j - m_{ij}}{2s_{ij}})^2} \tag{8.21}$$

όπου m_{ij} και s_{ij} είναι η κεντρική τιμή και τυπική απόκλιση για το j-στο χαρακτηριστικό του προτύπου i και a_j είναι το j-στο χαρακτηριστικό του στοιχείου a και αφού δεν υπάρχου a priori γνώση για τις πιθανές τιμές των χαρακτηριστικών $P(a) = \frac{1}{p_0}$, όπου p_0 σταθερά.

$$P(p_i/a) = p_0 P(a/p_i) P(p_i) \tag{8.22}$$

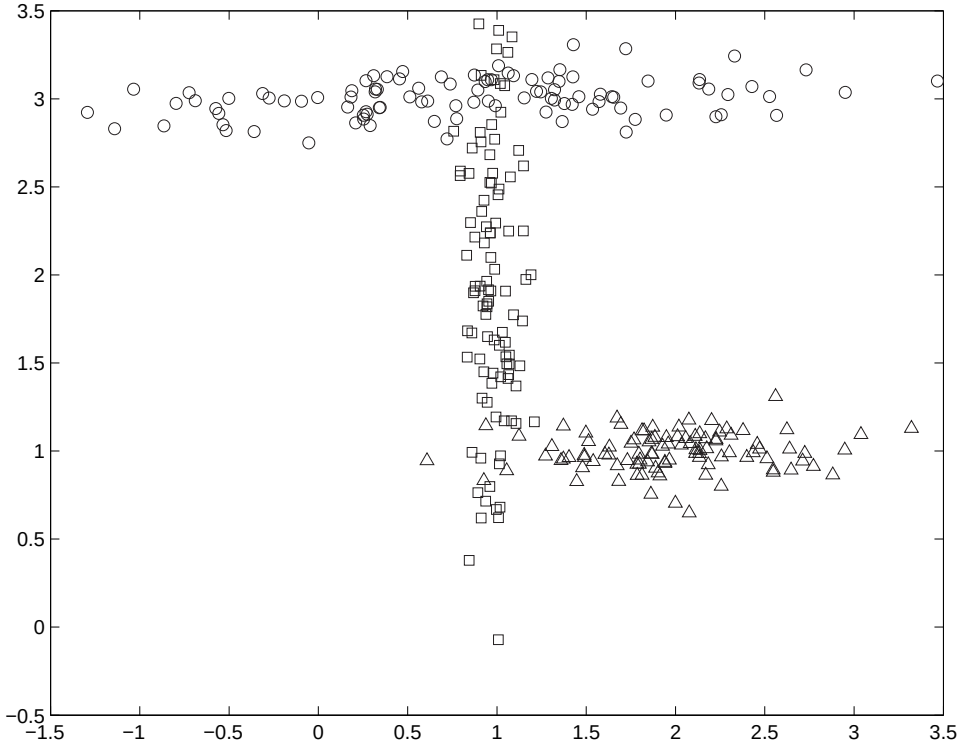
Χρησιμοποιώντας αυτό το σχήμα, μπορούμε να αναταξινομήσουμε όλα τα δεδομένα που χρησιμοποιήθηκαν για ομαδοποίηση. Εάν η ομαδοποίηση ήταν επιτυχής, δηλ. εάν τα ανιχνευμένα πρότυπα έχουν νόημα, τότε αυτή η διαδικασία θα βελτιώσει το ποσοστό ταξινόμησης απομακρύνοντας μερικά από τα μέλη των ομάδων, τα οποία ήταν αποτέλεσμα λαθών στα αρχικά βήματα. Κατά συνέπεια, αυτή η διαδικασία προσφέρει μια ένδειξη της αληθινής απόδοσης της αρχικής ομαδοποίησης. Επιπλέον, καθιστά το συνολικό αλγόριθμο πιο σταθερό, σε αντίθεση προς την απλή ιεραρχική ομαδοποίηση, δεδομένου ότι είναι πιο ανθεκτικός στα λάθη στα αρχικά βήματα.

Εναλλακτικά, μπορούμε να εφαρμόσουμε το αρχικό βήμα της ομαδοποίησης σε ένα μόνο μέρος των δεδομένων. Η αναταξινόμηση εφαρμόζεται στη συνέχεια για όλα τα δεδομένα, γενικεύοντας έτσι τα αποτελέσματα. Λαμβάνοντας υπόψη πως η διαδικασία της ιεραρχικής ομαδοποίησης έχει ιδιαίτερα υψηλή πολυπλοκότητα, ενώ αντίθετα η αναταξινόμηση έχει γραμμική πολυπλοκότητα, μπορούμε με αυτό τον τρόπο να μειώσουμε σημαντικά τις υπολογιστικές απαιτήσεις του συνολικού αλγορίθμου.

Τέλος, τα πειραματικά αποτελέσματα έχουν δείξει πως στις περισσότερες περιπτώσεις και ανεξάρτητα από το πιο μέρος των δεδομένων χρησιμοποιήθηκε για την αρχική ομαδοποίηση, αναδρομική εφαρμογή του βήματος αναταξινόμησης οδηγεί στο ίδιο σημείο ισορροπίας. Συνεπώς, ο συνολικός αλγόριθμος έχει μειωμένη ευαισθησία στη σειρά με την οποία τα δεδομένα εμφανίζονται στην είσοδό του.

Πίνακας 8.1: *Οι παράμετροι για την παραγωγή του συνόλου συνθετικών δεδομένων*

κλάση	m_1	s_1	m_2	s_2	στοιχεία
A	2	0.5	1	0.1	100
B	1	0.9	3	0.1	100
C	1	0.1	2	0.7	100



Σχήμα 8.1: *Το συνθετικό σύνολο δεδομένων.*

8.5 Πειραματικά αποτελέσματα

Σε αυτό το τμήμα παραθέτονται μερικά ενδεικτικά πειραματικά αποτελέσματα της προτεινόμενης μεθοδολογίας. Αρχίζουμε με αποτελέσματα εφαρμογής σε ένα απλό συνθετικό σύνολο δεδομένων, το οποίο διευκολύνει την παρουσίαση της λειτουργίας του αλγορίθμου. Στη συνέχεια, απαριθμούνται αποτελέσματα από την εφαρμογή σε πραγματικά σύνολα δεδομένων από τη βάση δεδομένων μηχανικής μάθησης του UCI [146].

8.5.1 Συνθετικά δεδομένα

Για να είναι η απεικόνιση των συνθετικών δεδομένων εφικτή, τα έχουμε περιορίσει στις δύο διαστάσεις. Δημιουργήθηκαν τρεις κλάσεις δεδομένων, χρησιμοποιώντας μια γκαουσιανή γεννήτρια. Οι παράμετροι των γκαουσιανών κατανομών που χρησιμοποιούνται για την παραγωγή του συνόλου δεδομένων παρουσιάζονται στον πίνακα 8.1.

Όπως μπορεί να φανεί από τον Πίνακα, καθώς επίσης και από το διάγραμμα, οι τρεις κλάσεις δεν διακρίνονται σαφώς η μια από την άλλη, και οι υποχώροι που χαρακτηρίζουν καλύτερα κάθε κλάση διαφέρουν σε έναν μεγάλο βαθμό. Αυτό κάνει τις προσεγγίσεις που βασίζονται στην συνάθροιση απόστασης ανεπαρκείς, πράγμα που

Πίνακας 8.2: *Οι ομάδες που παράγονται, για το συνθετικό σύνολο δεδομένων.*

Μέθοδος	ομάδα 1	ομάδα 2	ομάδα 3	% ταξινόμησης
Ευκλείδια ομαδ.	(0,0,13)	(28,0,87)	(72,100,0)	66.7%
Αρχική ομαδ.	(4,0,86)	(92,5,8)	(4,95,6)	91%
Μπεϋζ. επαναταξ.	(6,0,98)	(94,5,2)	(0,95,0)	95.7%

Πίνακας 8.3: *Ποσοστά ταξινόμησης για δεδομένα ίριδας ($\kappa = \lambda = 2$)*

Μέθοδος	ομάδα 1	ομάδα 2	ομάδα 3	% ταξινόμησης
Αρχικ. ομαδ.	(36,4,12)	(13,0,38)	(1,46,0)	80%
Μπεϋζιαν. επαναταξ. 1	(33,0,2)	(17,0,48)	(0,50,0)	87.3%
Μπεϋζιαν. επαναταξ. 2	(35,0,0)	(15,0,48)	(0,50,0)	90%

επιβεβαιώνεται στον Πίνακα 8.2. Στον πίνακα, η μορφή με την οποία παρουσιάζονται τα αποτελέσματα είναι (κλάση 1, κλάση 2, κλάση 3).

Το αρχικό βήμα ταξινόμησης παράγει ένα ποσοστό ταξινόμησης 91% (κατατάσσοντας κάθε ομάδα στην κλάση που κυριαρχεί), και η επαναταξινόμηση το επαναπροσδιορίζει σε 95.7%, δείχνοντας ότι το αρχικό βήμα, αν και έχει ένα μικρότερο ποσοστό, έχει σωστά προσδιορίσει τα υπάρχοντα πρότυπα. Η κλασσική Ευκλείδεια προσέγγιση, από την άλλη, έχει πολύ χαμηλότερες επιδόσεις.

8.5.2 Βάση δεδομένων ίριδας

Το σύνολο δεδομένων ίριδας περιλαμβάνει 150 στοιχεία, που χαρακτηρίζονται από 4 χαρακτηριστικά και ανήκουν σε τρεις κλάσεις. Δύο από αυτές τις κλάσεις δεν είναι γραμμικά διαχωρίσιμες. Πρόκειται για ένα σχετικά εύκολο σύνολο δεδομένων, καθώς ο αριθμός των ομάδων στα δεδομένα είναι ίσος με τον αριθμό των κλάσεων. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 8.3.

Η σημαντική βελτίωση που προσφέρει ένα και μόνο βήμα Μπεϋζιανής επαναταξινόμησης είναι ενδεικτική της εγκυρότητας των ανιχνευμένων ομάδων. Αυτή η παρατήρηση ισχυροποιείται ακόμη περισσότερο από το γεγονός ότι επαναλαμβανόμενη εφαρμογή του βήματος επαναταξινόμησης βελτιώνει ακόμη περισσότερο τα αποτελέσματα, ακόμα κι αν αυτό το βήμα είναι ανεπίβλεπτο.

8.5.3 Βάση δεδομένων καρκίνου του μαστού

Η βάση δεδομένων καρκίνου του μαστού του Wisconsin περιλαμβάνει 699 στοιχεία, τα οποία χαρακτηρίζονται από 9 ιδιότητες. Όλες αυτές οι ιδιότητες λαμβάνουν τιμές ακέραιων αριθμών στο $[1, 10]$. Τα στοιχεία συνοδεύονται επίσης από μία ταυτότητα, και πληροφορίες κλάσης, με πιθανές κλάσεις την καλοήθη και την κακοήθη. 65.5% των στοιχείων ανήκουν στη καλοήθη κλάση και 34.5% στην κακοήθη κλάση. 16 στοιχεία είναι ελλιπή (λείπει η τιμή μίας ιδιότητας) και έχουν αποκλειστεί από τη βάση δεδομένων για την εφαρμογή του αλγορίθμου.

Αυτό το σύνολο δεδομένων, που έχει ένα σχετικά μεγάλο αριθμό χαρακτηριστικών, είναι δυσκολότερο από το σύνολο δεδομένων των ιρίδων. Λεπτομερή αποτελέσματα που ελήφθησαν χρησιμοποιώντας την προτεινόμενη μεθοδολογία παρουσιάζονται στους Πίνακες 8.4 και 8.5. Αξίζει να σημειωθεί ότι αν και το ποσοσστό ταξινόμησης της αρχικής διαδικασίας ομαδοποίησης δεν είναι ιδιαίτερα υψηλό, το βήμα επα-

Πίνακας 8.4: *Ποσοστό ταξινόμησης για δεδομένα Wisconsin ($\kappa = \lambda = 2$)*

Μέθοδος	ομάδα 1	ομάδα 2	ομάδα 3	% ταξινόμησης
Αρχ. ομαδ.	(31,42)	(3,136)	(410,61)	86.1%
Μπεϋζιαν. επαναταξ.	(5,18)	(2,177)	(437,44)	92.5%

Πίνακας 8.5: *Ποσοστά ταξινόμησης για δεδομένα Wisconsin ($\kappa = \lambda = 5$)*

Μέθοδος	ομάδα 1	ομάδα 2	ομάδα 3	% ταξινόμησης
Αρχ. ομαδ.	(192,56)	(3,154)	(249,29)	87.1%
Μπεϋζιαν. επαναταξ.	(0,0)	(10,218)	(434,21)	95.5%

ναταξινόμησης το επαναπροσδιορίζει σημαντικά. Επιπλέον, για την περίπτωση όπου $\kappa = \lambda = 5$, το βήμα επαναταξινόμησης ταξινομεί κάθε στοιχείο σε μία από ακριβώς δύο ομάδες, με καθεμιά να κυριαρχείται σχεδόν αποκλειστικά από μια κλάση.

Αυτή η απόδοση δεν απέχει πολύ από αυτήν των εκπαιδευμένων συστημάτων ταξινόμησης, που χρησιμοποιούν το ίδιο σύνολο δεδομένων; ποσοστό ταξινόμησης 97% αναφέρεται στο [12]. Αυτό είναι ενδεικτικό της αποδοτικότητας της μεθόδου, λαμβάνοντας υπόψη ότι αναφερόμαστε στην σύγκριση μιας ανεπίβλεπτης μεθόδου με μια εποπτευμένη.

8.5.4 Βάση δεδομένων Ιονόσφαιρας

Αυτά τα δεδομένα από ραντάρ συλλέχθηκαν από ένα σύστημα στον κόλπο των χήνων στο Λαμπραντόρ. Οι στόχοι ήταν ελεύθερα ηλεκτρόνια στην ιονόσφαιρα. "Καλές" επιστροφές ραντάρ είναι οι επιστροφές εκείνες που παρουσιάζουν στοιχεία κάποιου τύπου δομής μέσα στην ιονόσφαιρα. "Κακές" επιστροφές είναι εκείνες που δεν παρουσιάζουν κάτι τέτοιο. Τα στοιχεία του συνόλου δεδομένων χαρακτηρίζονται από 34 χαρακτηριστικά και ταξινομούνται είτε ως καλά, είτε ως κακά. Αποτελέσματα από την εφαρμογή της προτεινόμενης μεθοδολογίας εμφανίζονται στον Πίνακα 8.6.

Λαμβάνοντας υπόψη ότι οι εποπτευμένοι αλγόριθμοι ταξινόμησης αναφέρουν ένα ποσοστό ταξινόμησης περίπου 90%, είναι εύκολο να συναχθεί το συμπέρασμα ότι η αρχική ομαδοποίηση είναι εξαιρετικά αποδοτική. Εάν λάβουμε υπόψη και ότι οι ανεπίβλεπτες μέθοδοι ομαδοποίησης δεν υπερβαίνουν ένα ποσοστό ταξινόμησης 80% για 10 ομάδες [3], τότε μπορούμε να καταλήξουμε στο συμπέρασμα ότι η ανίχνευση δύο ομάδων με ένα ποσοστό 87.2% είναι εξαιρετικά επιτυχής.

Το βήμα της Μπεϋζιανής επαναταξινόμησης, βελτιώνει την ομαδοποίηση, ώστε να επιτευχθεί ένα ποσοστό ταξινόμησης 91.2 για 25 ομάδες. Επιπρόσθετα, αυτό το βήμα διακρίνει την έξοδο του αρχικού βήματος σε έχουσα σημασία και σε τυχαία. Συγκεκριμένα, με τη βοήθεια αυτού του βήματος είναι εύκολο να δει κανείς ότι αν

Πίνακας 8.6: *Ποσοστά ταξινόμησης για δεδομένα Ιονόσφαιρας ($\kappa = \lambda = 2$)*

Αριθμός από ομάδες	Αρχική ομαδοπ.	Μπεϋζιανή. επαναταξ.
2	87.2%	80%
3	87.2%	80.1%
10	87.2%	84.9%
15	87.2%	87.2%
20	87.2%	87.7%
25	87.2%	91.2%

και αναφέρεται το ίδιο ποσοστό ταξινόμησης, μια ομαδοποίηση σε λιγότερες από 10 ομάδες δεν είναι ικανή να παρέχει αποδοτική αρχικοποίηση σε ένα ταξινομητή, και συνεπώς ότι δεν περιγράφει επαρκώς τα υπάρχοντα πρότυπα, ενώ ένας χωρισμός σε 25 ομάδες είναι αποτελεσματικότερος.

□

Κεφάλαιο 9

Αρχικοποίηση νευρωνικών δικτύων

9.1 Εισαγωγή

Στην περίπτωση δικτύων ανακατανομής πόρων που δημιουργούν κόμβους κατά περίπτωση η αρχικοποίηση με λίγους κόμβους δεν είναι απαγορευτική. Στην περίπτωση δικτύων που αφαιρούν περιττούς κόμβους η αρχικοποίηση με πολλούς κόμβους είναι επίσης αποδεκτή. Έτσι, θα μπορούσαμε, θεωρητικά, να ξεκινήσουμε στην πρώτη περίπτωση από ένα μόνο κόμβο, ενώ στη δεύτερη δημιουργώντας ένα ξεχωριστό κόμβο για καθένα από τα δεδομένα εισόδου. Σε κάθε περίπτωση, φαίνεται το πλήθος των κόμβων στην αρχική κατάσταση να μην είναι τόσο κρίσιμο, καθώς η διαδικασία της εκπαίδευσης που ακολουθεί επιδρά και στη δομή του δικτύου.

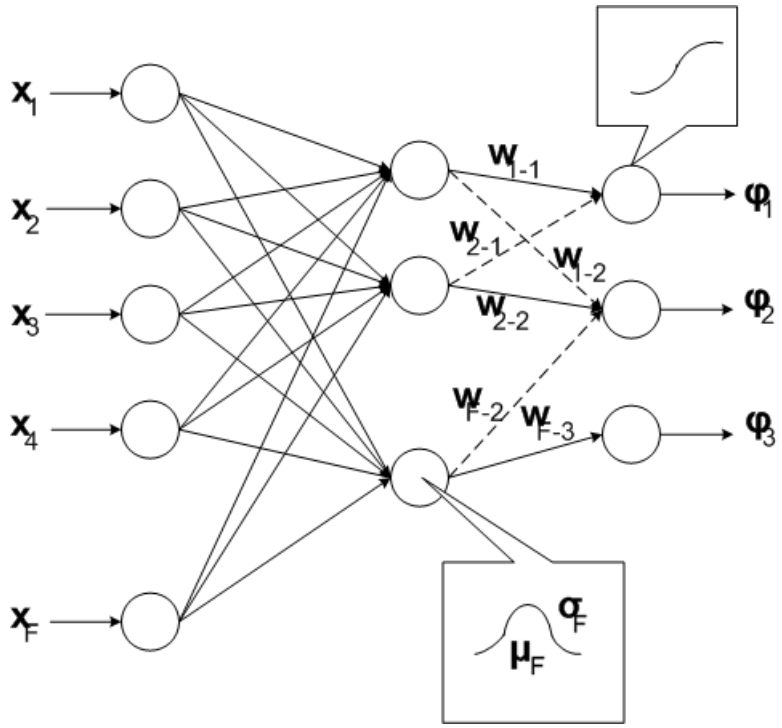
Στην πράξη, όμως, τα πράγματα είναι αρκετά διαφορετικά. Όταν η αρχικοποίηση του δικτύου απέχει πολύ από την επιθυμητή θέση ισορροπίας, τότε η διαδικασία εκπαίδευσης αντιμετωπίζει μια σειρά από προβλήματα, όπως τα παρακάτω:

- Η σύγκλιση αργεί πολύ να έρθει. Μεγαλώνει, δηλαδή, ο απαιτούμενος χρόνος εκπαίδευσης.
- Υπάρχει κίνδυνος το δίκτυο να εγκλωβιστεί σε κάποιο τοπικό ελάχιστο μακριά από την βέλτιστη και επιθυμητή λύση.
- Η διαδικασία ανακατανομής πόρων είναι ιδιαίτερα πιθανό να δημιουργήσει πληθώρα κόμβων. Αυτό θα έχει σαν αποτέλεσμα το δίκτυο να χάσει τη ικανότητα να γενικεύει τη γνώση που περιέχει ώστε να κατηγοριοποιεί σωστά μελλοντικά δεδομένα.

Έτσι δημιουργείται το πρόβλημα της αυτόματης επιλογής της αρχικής δομής του δικτύου, καθώς και της αυτόματης επιλογής των αρχικών τιμών των παραμέτρων του. Εδώ εξάγουμε και τα δύο με βάση τα αποτελέσματα της μεθόδου ομαδοποίησης του προηγούμενου κεφαλαίου.

9.2 Δομή δικτύου

Η αρχιτεκτονική που χρησιμοποιούμε περιέχει τρία επίπεδα: το επίπεδο εισόδου που περιέχει n κόμβους, μέσω του οποίου ένα διάνυσμα $\underline{x} \in \mathcal{R}^n$ δίδεται στο δίκτυο ως είσοδος, ένα κρυμμένο επίπεδο που περιέχει $q(t)$ κόμβους ακτινικής βάσης (στη χρονική στιγμή t), και το επίπεδο εξόδου, που περιέχει p σιγμοειδείς κόμβους [86].



Σχήμα 9.1: *Το νευρωνικό δίκτυο.*

Η μάθηση του δικτύου βασίζεται στη μέθοδο της καθόδου κλίσης βαθμίδας. Για την εκτίμηση της επίδοσης του δικτύου χρησιμοποιούμε ένα κριτήριο τετραγωνικού σφάλματος. Το τετραγωνικό σφάλμα $e(t)$ στη χρονική στιγμή t υπολογίζεται με το συνήθη τρόπο:

$$e(t) = \frac{1}{2}\sum_{k=1}^{p}(d_k(t) - y_k(t))^2 \tag{9.1}$$

όπου $d_k(t)$ είναι η επιθυμητή έξοδος και $y_k(t)$ είναι η πραγματική έξοδος του κόμβου k. Η πραγματική έξοδος δίνεται από τη σχέση:

$$y_k = \frac{1 - e^{2z_k}}{1 + e^{2z_k}}, \quad z_k = \underline{w}_k^T \cdot \underline{\phi} \tag{9.2}$$

όπου $\underline{w}_k = [w_{k1}\ w_{k2} \dots w_{kq(t)}]^T$, $k =\in \mathcal{N}_p$, είναι τα βάρη που συνδέουν τους κόμβους του κρυμμένου επιπέδου με τους κόμβους του επιπέδου εξόδου και $\underline{\phi}$ είναι η έξοδος του κρυμμένου επιπέδου.

Κάθε κόμβος του κρυμμένου επιπέδου αναπαριστά μια ξεχωριστή περιοχή του χώρου εισόδου και υπολογίζει μια συνάρτηση της εισόδου $\underline{x}$ σύμφωνα με τη σχέση:

$$\phi_j(\underline{x}) = exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - \mu_{ji}}{\sigma_{ji}})^2\} \tag{9.3}$$

όπου $\mu_j = [\mu_{j1}, \mu_{j2} \dots \mu_{jn}]$ και $\sigma_j = [\sigma_{j1}, \sigma_{j2} \dots \sigma_{jn}]$ είναι το κέντρο και οι διασπορές του j-στού κρυμμένου κόμβου, αντίστοιχα. Η συνολική έξοδος του κρυμμένου επιπέδου δίνεται προφανώς από τη σχέση

$$\underline{\phi} = [\phi_1, \phi_2 \dots \phi_n] \tag{9.4}$$

Οι τρεις διαφορετικοί τύποι παραμέτρων του δικτύου (μ_j, σ_j, $j \in \mathcal{N}_{q(t)}$ και w_k, $k \in \mathcal{N}_p$) ενημερώνονται με τη χρήση των ακόλουθων σχέσεων:

$$w_{kj}(t+1) = w_{kj}(t) - \eta(t) \cdot a_j(t) \frac{\vartheta e(t)}{\vartheta w_{kj}(t)} \tag{9.5}$$

$$\mu_{ji}(t+1) = \mu_{ji}(t) - \eta(t) \cdot a_j(t) \frac{\vartheta e(t)}{\vartheta \mu_{ji}(t)} \tag{9.6}$$

$$\sigma_{ji}(t+1) = \sigma_{ji}(t) - \eta(t) \cdot a_j(t) \frac{\vartheta e(t)}{\vartheta \sigma_{ji}(t)} \tag{9.7}$$

$\eta(t)$ είναι ο ρυθμός μάθησης που υπολογίζεται δυναμικά ώστε να εκφράζει μεγάλους ρυθμούς μάθησης στα πρώτα βήματα και σταδιακή σταθεροποίηση της λειτουργίας στα επόμενα.

Η παράμετρος $a_j(t)$ στις σχέσεις 9.5, 9.6 και 9.7 σχετίζεται με τον j-στό κρυμμένο κόμβο και συμβάλει σε μια διαδικασία ανταγωνιστικής μάθησης. Συγκεκριμένα, ο όρος $a_j(t)$ δείχνει την ομοιότητα μεταξύ του j-στού κόμβου και της εισόδου $\underline{x}(t)$. Υπολογίζεται με τη σχέση:

$$a_j(t) = 1 - \frac{\|\underline{x}(t) - \underline{\mu}_j\| - \|\underline{x}(t) - \underline{\mu}_{nearest}\|}{\|\underline{x}(t) - \underline{\mu}_{farthest}\| - \|\underline{x}(t) - \underline{\mu}_{nearest}\|} \tag{9.8}$$

όπου $\underline{\mu}_{farthest}$ και $\underline{\mu}_{nearest}$ είναι τα κέντρα του πιο κοντινού και του πιο μακρινού κόμβου από το $\underline{x}(t)$, αντίστοιχα, και $\|\cdot\|$ είναι η Ευκλείδεια απόσταση.

Τα δεδομένα εκπαίδευσης προσφέρονται στο δίκτυο στη μορφή ζευγαριών $(\underline{x}(t), \underline{d}(t))$ από διανύσματα εισόδου και αντίστοιχες επιθυμητές εξόδους. Άν μια νέα είσοδος $\underline{x}(t)$ δεν ενεργοποιεί σημαντικά κανένα από τους κόμβους του κρυμμένου επιπέδου και το τετραγωνικό σφάλμα $e(t)$ είναι μικρό, τότε ένας νέος κόμβος δημιουργείται στο κρυμμένο επίπεδο σύμφωνα με τις σχέσεις:

$$q(t) = q(t-1) + 1 \tag{9.9}$$

$$\underline{\mu}_{q(t)} = \underline{x}(t) \tag{9.10}$$

$$\underline{\sigma}_{q(t)} = k \cdot \|\underline{x}(t) - \underline{\mu}_{nearest}\| \tag{9.11}$$

$$w_{kq(t)} = d_k(t) - y_k(t), \quad k \in \mathcal{N}_p \tag{9.12}$$

όπου k είναι μια σταθερά.

Άν είτε η νέα είσοδος $\underline{x}(t)$ ενεργοποιεί σημαντικά ένα τουλάχιστον κόμβο ή το τετραγωνικό σφάλμα είναι μικρό, οι παράμετροι του δικτύου ενημερώνονται με βάση τις σχέσεις 9.5, 9.6, 9.7, χρησιμοποιώντας τις ακόλουθες τιμές:

$$\frac{\vartheta e(t)}{\vartheta w_{kj}(t)} = \phi_j(\underline{x}(t))\{d_k(t) - y_k(t)\}\{1 - (y_k(t))^2\} \tag{9.13}$$

$$\frac{\vartheta e(t)}{\vartheta \mu_{ji}(t)} = \phi_j(\underline{x}(t)) \frac{\{x_i(t) - \mu_{ji}(t)\}}{\sigma_{ji}^2(t)} \sum_{k=1}^{p} (w_{kj}(t)\{d_k(t) - y_k(t)\}\{1 - (y_k(t))^2\}) \tag{9.14}$$

$$\frac{\vartheta e(t)}{\vartheta \sigma_{ji}(t)} = \phi_j(\underline{x}(t)) \frac{\{x_i(t) - \mu_{ji}(t)\}}{\sigma_{ji}^3(t)} \sum_{k=1}^{p} (w_{kj}(t)\{d_k(t) - y_k(t)\}\{1 - (y_k(t))^2\}) \tag{9.15}$$

Πίνακας 9.1: *Βαθμοί ταξινόμησης και πλήθη κρυμμένων κόμβων:*

	Χωρίς εκπ.	Τυχαία	Bayes	Προτεινόμενη
Πλήθος κόμβων	3	6	5	3
Βαθμός ταξινόμησης	87.33%	96%	97.3	98%

9.3 Αρχικοποίηση με βάση την ομαδοποίηση

Οι μέσες τιμές των χαρακτηριστικών για κάθε ομάδα διαμορφώνουν το εικονικό κέντρο, δηλ. ένα εικονικό στοιχείο που βρίσκεται στο κέντρο της ομάδας, όταν όλα τα στοιχεία του τοποθετούνται στον F-διάστατο χώρο. Η θέση του μπορεί να θεωρηθεί ως μια γενική περιγραφή των τιμών των χαρακτηριστικών του προτύπου, στο οποίο αντιστοιχεί αυτή η ομάδα. Οι διακυμάνσεις των τιμών για κάθε χαρακτηριστικό δείχνουν την σημασία κάθε χαρακτηριστικού για τον καθορισμό της ομάδας; αυτό μπορεί να θεωρηθεί ως εκτίμηση της ακτίνας της ομάδας στον άξονα του κάθε χαρακτηριστικού.

Υποθέτοντας ότι η ομαδοποίηση έχει δημιουργήσει λογικές ομάδες από στοιχεία, τα τελευταία μπορούν να χρησιμοποιηθούν για την αρχικοποίηση ενός προσαρμοζόμενου νευρωνικού ταξινομητή. Το γεγονός ότι οι ομάδες περιγράφονται μέσω συνδυασμών κέντρων και διασπορών κάνει την έξοδο ιδανική για την αρχικοποίηση νευρωνικών δικτύων ακτινικής βάσης [124]. Τα βάρη με τα οποία συνδέονται οι κόμβοι του κρυμμένου επιπέδου με το επίπεδο εξόδου μπορούν επίσης να εξαχθούν από τα αποτελέσματα της ομαδοποίησης.

Η αρχικοποίηση του δικτύου επιτυγχάνεται θέτοντας τις τιμές των $q(0)$, $\underline{\mu}_j$, $\underline{\sigma}_j$, $j \in \mathcal{N}_{q(t)}$ και $\underline{w}_k$, $k \in \mathcal{N}_p$ σύμφωνα με τα αποτελέσματα της ιεραρχικής ομαδοποίησης. Έτσι, το πλήθος $q(0)$ των κόμβων στο κρυμμένο επίπεδο τίθεται ίσο με το πλήθος των ανιχνευμένων ομάδων. Τα κέντρα $\underline{\mu}_j$ των κρυμμένων κόμβων λαμβάνονται από τα εικονικά κέντρα των αντίστοιχων ομάδων:

$$\mu_{ji} = m_j \tag{9.16}$$

όμοια και για τις διασπορές:

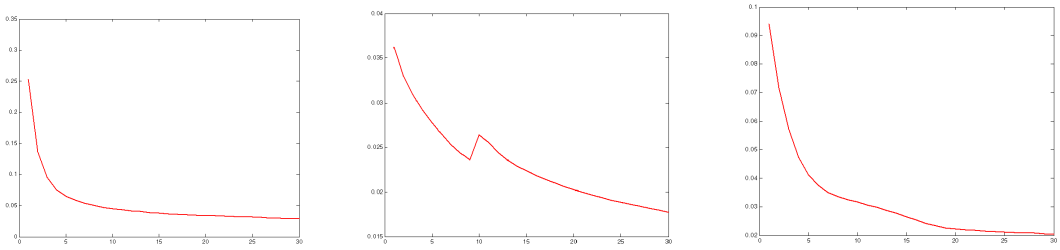
$$\sigma_{ji} = \sqrt{\frac{1}{N_j}\sum_{k=1}^{N_j}(\nu_{ji}^k - m_j)^2} \tag{9.17}$$

όπου ν_{ji}^k είναι το i-στό στοιχείο του k-στού διανύσματος της j-στής ομάδας, $\underline{m}_j$ είναι το εικονικό κέντρο της ομάδας και N_j είναι το πλήθος των διανυσμάτων που περιέχονται στην ομάδα. Τα βάρη $\underline{w}_k$ καθορίζονται από τον τρόπο με τον οποίο οι ομάδες αντιστοιχούν στις πραγματικές κλάσεις. Συγκεκριμένα, αν $per\%$ των διανυσμάτων της ομάδας j ανήκουν στην κλάση k, τότε ο κρυμμένος κόμβος j συνδέεται με τον κόμβο εξόδου k με βαθμό

$$w_{kj} = \frac{per}{100} \tag{9.18}$$

9.4 Πειραματικά αποτελέσματα

Για τον έλεγχο της επίδοσης της προτεινόμενης μεθοδολογίας την εφαρμόζουμε σε μια σειρά συνόλων δεδομένων. Τέσσερα διαφορετικά πειράματα έγιναν με το σύνολο των δεδομένων ίριδας:



Σχήμα 9.2: *Το τετραγωνικό σφάλμα σαν συνάρτηση των εποχών εκπαίδευσης*

- Το δίκτυο αρχικοποιήθηκε με την προτεινόμενη μεθοδολογία και δεν εκπαιδεύτηκε.
- Το δίκτυο αρχικοποιήθηκε με τρεις κόμβους. Οι υπόλοιπες παράμετροι ήταν τυχαίες. Ακολούθησε εκπαίδευση.
- Ακολουθήθηκε μια πιθανοτική προσέγγιση: τα κέντρα και οι αποκλίσεις των γνωστών κλάσεων χρησιμοποιήθηκαν για αρχικοποίηση και ακολούθησε εκπαίδευση.
- Η προτεινόμενη μεθοδολογία εφαρμόστηκε για αρχικοποίηση και ακολούθησε εκπαίδευση.

Στο Σχήμα 9.2 παρουσιάζεται το σφάλμα σαν συνάρτηση της χρονικής στιγμής t για τις τρεις δοκιμές που περιέλαβαν στάδιο εκπαίδευσης. Το πρώτο σχήμα αντιστοιχεί στην τυχαία αρχικοποίηση, το δεύτερο στην πιθανοτική και το τελευταίο στην προτεινόμενη μεθοδολογία. Η τυχαία αρχικοποίηση γρήγορα συναντά ένα άνω όριο απόδοσης και σταματά να μειώνει το σφάλμα, έχοντας προφανώς συναντήσει ένα τοπικό ελάχιστο. Ο προσεγγίσεις που έχουν "λογική" αρχικοποίηση προχωρούν πολύ καλύτερα. Στα πρώτα βήματα η πιθανοτική αρχικοποίηση προχωρά πιο γρήγορα, αλλά αυτό σύντομα αλλάζει και η προτεινόμενη μέθοδος πετυχαίνει ίδια επίδοση ως προς το τετραγωνικό σφάλμα.

Στον Πίνακα 9.1 παρουσιάζονται οι επιδόσεις των δικτύων ως προς το βαθμό σωστής ταξινόμησης. Αν και το τετραγωνικό σφάλμα είναι το κριτήριο που χρησιμοποιείται κατά το στάδιο της εκπαίδευσης, τελικά είναι ο βαθμός ταξινόμησης που κρίνει την επιτυχία του δικτύου. Το πρώτο σχόλιο είναι πως το δίκτυο που δεν εκπαιδεύτηκε είχε πολύ χειρότερη επίδοση από τα άλλα, δείχνοντας έτσι πως η εκπαίδευση είναι ένα στάδιο που δεν μπορεί να αποφευχθεί.

Όσο αφορά στο βαθμό ταξινόμησης σε σχέση με το πλήθος των κρυμμένων κόμβων, η προτεινόμενη μεθοδολογία ισορροπεί σε ένα δίκτυο με τρεις μόνο κόμβους, όσες και οι πραγματικές κλάσεις, χωρίς να υστερεί στις επιδόσεις ταξινόμησης. Εκτός από τις δοκιμές που παρουσιάζονται εδώ, η μέθοδος ξεπερνά και άλλες στη βιβλιογραφία (7 κόμβοι, 96.7% [103]) (17 κόμβοι, 95.3% [70]) (9 κανόνες, 95.3% [71]) (7 κανόνες, 96% [63]).

Παρόμοια αποτελέσματα παρατηρούνται και σε άλλα σύνολα δεδομένων. Ειδικά στα δεδομένα της ιονόσφαιρας, που χαρακτηρίζονται από το μεγάλο αριθμό διαστάσεων, πιο απλές τεχνικές αρχικοποίησης, αντίθετα με την προτεινόμενη μεθοδολογία, αποτυγχάνουν παντελώς.

□

Κεφάλαιο 10

Εξαγωγή ασαφών κανόνων από νευρωνικά δίκτυα

10.1 Εισαγωγή

Υπάρχουν αρκετά υβριδικά μοντέλα που λειτουργούν με γνώση προερχόμενη από τα δεδομένα, μοντελοποιημένη με τη μορφή ασαφών κανόνων, και αναπραιστώμενη στη δομή νευρωνικών δικτύων. Μερικοί από τους λόγους που αυτή η μεθοδολογία είναι τόσο δημοφιλής αναφέρονται παρακάτω:

- Η εξαγωγή κανόνων προσδίδει στα νευρωνικά δίκτυα τη δυνατότητα επεξήγησης της λειτουργίας τους, επιτρέποντας έτσι στο χρήστη να ελέγχει την εσωτερική λογική του συστήματος.
- Η εξαγωγή κανόνων συχνά συνδράμει στον εντοπισμό εξαρτήσεων που δεν είναι ήδη γνωστές, συμβάλλοντας έτσι στην επέκταση της κατανόησης του συστήματος.
- Γενικά πιστεύεται πως ένα σύστημα που είναι εύκολα κατανοητό έχει και μεγαλύτερες δυνατότητες γενίκευσης.
- Η προσθήκη γνώσης σε ένα δίκτυο επιταχύνει τη διαδικασία εκπαίδευσης.

Μια συνήθης τακτική για την εξαγωγή γνώσης από αριθμητικά δεδομένα είναι η χρήση τεχνικών ομαδοποίησης. Όπως αναφέραμε και στο κεφάλαιο 8, όταν το πλήθος των διαστάσεων αυξάνεται αυτή η προσέγγιση γίνεται πιο δύσκολη στην εφαρμογή και λιγότερο αξιόπιστη ως προς τα αποτελέσματά της.

Η ιδέα να χρησιμοποιηθούν νευρωνικά δίκτυα ακτινικής βάσης για την εξαγωγή των κανόνων δεν είναι τόσο καινούρια. Ο τρόπος λειτουργίας των κόμβων του κρυμμένου επιπέδου σε αυτά τα δίκτυα, καθώς και ο τρόπος με τον οποίο ομαδοποιούν τα δεδομένα εκπαίδευσης, τα κάνει ιδανικά για τον ορισμό κανόνων if-then με βάση τα αποτελέσματα της εκπαίδευσης του δικτύου. Ωστόσο, για να εξαχθούν πραγματικά χρήσιμοι κανόνες να επιλεγεί προσεκτικά το τμήμα if πρέπει, δηλαδή ο κατάλληλος υποχώρος για κάθε ομάδα.

Για την εξαγωγή κανόνων της μορφής if-then από νευρωνικά δίκτυα ακτινικής βάσης βασιζόμαστε στα παρακάτω:

- Οι κρυμμένοι κόμβοι του δικτύου συνδυάζουν τις εισόδους με μια πράξη τύπου ΚΑΙ, δημιουργώντας έτσι το if κομμάτι του κανόνα.

- Οι κόμβοι εξόδου συνδυάζουν τις εξόδους των κρυμμένων κόμβων με μια πράξη τύπου Ή. Έτσι οι παράμετροι ένωσης του κρυμμένου επιπέδου με το επίπεδο εξόδου δημιουργούν το then κομμάτι του κανόνα.

- Η γνώση στη μορφή if-then προέρχεται από την ομαδοποίηση των διαθεσίμων δεδομένων, δηλαδή από την εκπαίδευση του δικτύου.

- Κατά τη λειτουργία του δικτύου η ασάφεια υποστηρίζεται τόσο από το κρυμμένο επίπεδο όσο και από τον περιορισμό της τελικής εξόδου στο [0 1], ταιριάζοντας έτσι απόλυτα στη λογική των συστημάτων ασαφών κανόνων.

Το παραπάνω πλαίσιο είναι βέβαια λογικό, στην πράξη όμως αποδεικνύεται ανεπαρκές. Οι κανόνες που εξάγονται με αυτόν τον τρόπο πραγματικά αντιστοιχούν στα δεδομένα εκπαίδευσης και όταν χρησιμοποιούνται πετυχαίνουν υψηλά ποσοστά κατηγοριοποίησης, αλλά δεν είναι εύκολοι στη χρήση και κατανόηση στη λεκτική μορφή τους καθώς όλες οι πιθανές είσοδοι συμμετέχουν σε κάθε κανόνα. Αυτό γίνεται πιο έντονο όταν οι διαστάσεις των δεδομένων εισόδου είναι πολλές.

Σε αυτό το κεφάλαιο χρησιμοποιούμε γενετικούς αλγορίθμους για να αντιμετωπίσουμε αυτό το πρόβλημα. Η μέθοδός μας είναι πολύ διαφορετική από το συνήθη συνδυασμό δικτύων ακτινικής βάσης και γενετικών, καθώς εκεί η έμφαση δίνεται σε επιλογή χαρακτηριστικών για όλο το δίκτυο [140][1], ενώ εδώ επιλέγουμε τις κατάλληλες εισόδους για κάθε κανόνα χωριστά.

Στην επόμενη ενότητα συζητάμε γενικά το θέμα της εξαγωγής κανόνων και παρουσιάζουμε τους βασικούς μαθηματικούς συμβολισμούς. Στη συνέχεια, βασιζόμενοι στη θεωρία των κεφαλαίων 8 και 9, εξηγούμε πώς μπορούμε με τη βοήθεια γενετικών αλγορίθμων να εξάγουμε απλοποιημένους αλλά αποδοτικούς κανόνες από αριθμητικά δεδομένα.

10.2 Εξαγωγή κανόνων

Γενικά, η μέθοδος που ακολουθούμε για την εξαγωγή κανόνων από αριθμητικά δεδομένα συνοψίζεται στα παρακάτω:

Εφαρμόζουμε μια εποπτευμένη διαδικασία μάθησης για ένα χώρο D. Η διαδικασίακμάθησης έχει στη διάθεσή της ζεύγη από στοιχεία $\underline{x} \in \mathcal{R}^n$ του χώρου εισόδου E και αντίστοιχων τιμών $\underline{d}(\underline{x}) \in \mathcal{R}^m$ στο χώρο D. Όταν ολοκληρωθεί η διαδικασία μάθησης ένα σύνολο παραμέτρων G, συνήθως αποθηκευμένων με τη μορφή πινακων, αποτυπώνουν τη σχέση

$$g : \mathcal{R}^n \rightarrow \mathcal{R}^m \tag{10.1}$$

έτσι ώστε

$$\|G(\underline{x}) - g(\underline{x})\| < \epsilon \tag{10.2}$$

για κάποιο $\epsilon > 0$.

Στη συγκεκριμένη μεθοδολογία που ακολουθούμε σε αυτό το κεφάλαιο, το σύνολο G των παραμέτρων συνίσταται από 4 πίνακες που αντιστοιχούν στα μέσα διανύσματα (πίνακας M) και τις διασπορές (πίνακας Σ) των κόμβων του κρυμμένου επιπέδου, στις συσχετίσεις ανάμεσα στο κρυμμένο επίπεδο και το επίπεδο εξόδου (πίνακας W) και στις συσχετίσεις ανάμεσα στο επίπεδο εισόδου και το επίπεδο εξόδου (πίνακας A).

Θεωρώντας πως υπάρχουν q κόμβοι στο κρυμμένο επίπεδο, ο ακόλουθος συμβολισμός ακολουθείται:

$$M \in \mathcal{R}^{n \times q}, M = |\underline{\mu_1}\ \underline{\mu_2} \dots \underline{\mu_q}| \tag{10.3}$$

όπου $\underline{\mu_i} \in \mathcal{R}^n$ το μέσο διάνυσμα για τον i-στό κόμβο του κρυμμένου επιπέδου,

$$\Sigma \in \mathcal{R}^{n \times q}, \Sigma = |\underline{\sigma_1}\ \underline{\sigma_2} \dots \underline{\sigma_q}| \tag{10.4}$$

όπου $\underline{\sigma_i} \in \mathcal{R}^n$ το διάνυσμα διασπορών για τον i-στό κόμβο του κρυμμένου επιπέδου (επιτρέπουμε διαροφερική διασπρά ανά διάσταση και όχι μια γενική διασπορά όπως στη συνήθη προσέγγιση),

$$W \in \mathcal{R}^{q \times p}, W = |\underline{w_1}\ \underline{w_2} \dots \underline{w_p}| \tag{10.5}$$

όπου $\underline{w_i} \in \{\prime, \infty\}^q$ ένα διαδικό διάνυσμα που δείχνει ποιοί κρυμμένοι κόμβοι συνεισφέρουν στην i-οστή κλάση εξόδου και

$$A \in \mathcal{R}^{n \times q}, A = |\underline{\alpha_1}\ \underline{\alpha_2} \dots \underline{\alpha_q}| \tag{10.6}$$

όπου $\underline{\alpha_i} \in \{\prime, \infty\}^n$ ένα διαδικό διάνυσμα που δείχνει ποιοί κόμβοι εισόδου συνεισφέρουν στον i-οστό κόμβο του κρυμμένου επιπέδου.

Τα παραπάνω παρουσιάζονται και γραφικά στο σχήμα 10.1, όπου παρουσιάζουμε παράδειγμα δικτύου ακτινικής βάσης με παραμέτρους $n = 4$, $q = 5$, $p = 2$.

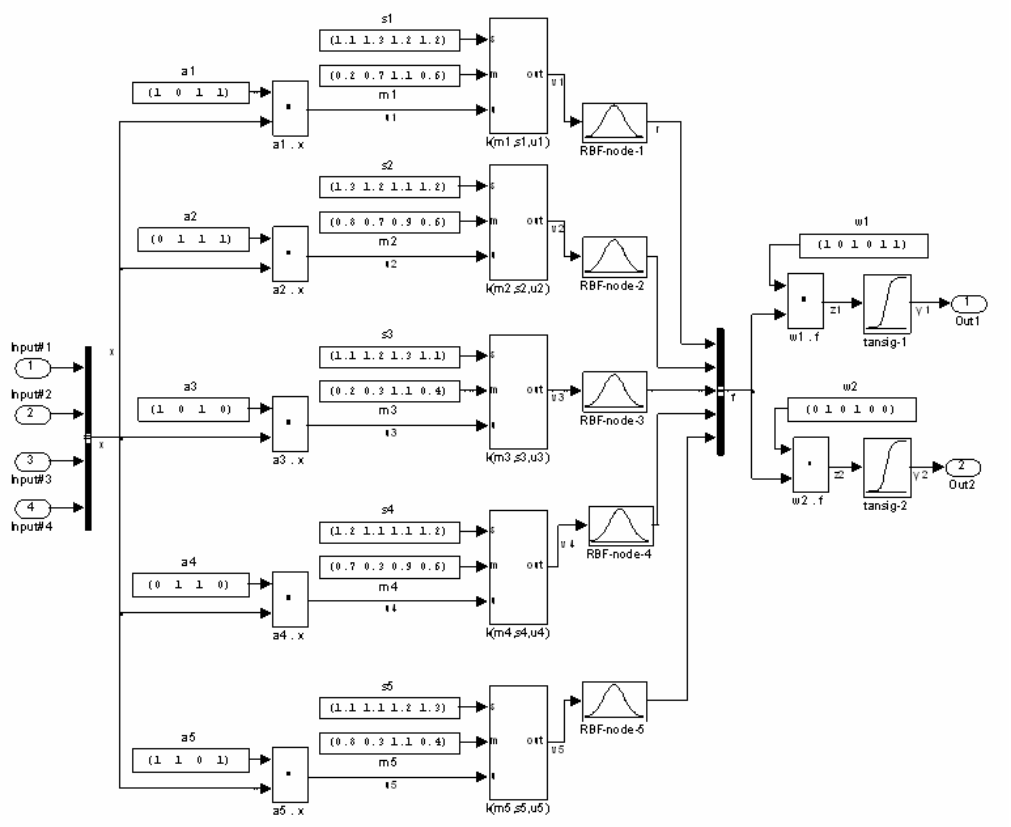
Οι τιμές των πινάκων M, Σ και W εξάγονται απευθείας από την εκπαίδευση του δικτύου (βλέπε κεφάλαιο 9), ενώ ο πίνακας A βελτιστοποιείται με χρήση γενετικών αλγορίθμων, όπως εξηγείται στην ακόλουθη ενότητα.

10.3 Γενετική απλοποίηση κανόνων

Το τελευταίο, και πιθανότατα και πιο σημαντικό, βήμα στην εξαγωγή κανόνων είναι η εξαγωγή του if τμήματος των κανόνων, δηλαδή ο υπολογισμός του πίνακα A. Με την αρχικοποίηση του δικτύου με βάση τα αποτελέσματα της διαδικασίας ομαδοποίησης που παρουσιάζεται κεφάλαιο 8, οι κόμβοι κρυμμένου επιπέδου συνδέονται στην ουσία με υποχώρους εισόδου. Μετά την εκπαίδευση του δικτύου όμως, όπως αυτή παρουσιάζεται στο κεφάλαιο 9, όλοι οι κόμβοι συνδέονται με όλες τις μεταβλητές εισόδου, και έτσι όλες οι τιμές του πίνακα A είναι ίσες με τη μονάδα. Για να εξαχθούν πρακτικά αξιοποιήσιμοι κανόνες από το δίκτυο πολλές από αυτές τις τιμές πρέπει να γίνουν ίσες με το μηδέν, δηλαδή πολλές συνδέσεις ανάμεσα στο επίπεδο εισόδου και το κρυμμένο επίπεδο πρέπει να καταργηθούν. Ακόμη σημαντικότερα, οι συνδέσεις που πρέπει να καταργηθούν δεν είναι ίδιες για κάθε ομάδα, οπότε δεν είναι ίδιες και για κάθε κανόνα. Για να ικανοποιήσουμε τα παραπάνω, εφαρμόζουμε τη γενετική διαδικασία που περιγράφεται παρακάτω.

Έστω ϕ_i η ενεργοποίηση του i-στού κόμβου του κρυμμένου επιπέδου. Αυτή δίνεται από τη σχέση

$$\phi_i(\underline{x}) = exp\{-\frac{1}{2}\sum_{j=1}^{n}(\frac{x_j - \mu_{ij}}{\sigma_{ij}})^2\} \tag{10.7}$$



Σχήμα 10.1: *Η αρχιτεκτονική του δικτύου μαζί με τις παραμέτρους του.*

όπου $\mu_i = [\mu_{i1}, \mu_{i2} \ldots \mu_{in}]$ και $\sigma_i = [\sigma_{i1}, \sigma_{i2} \ldots \sigma_{in}]$ είναι το κέντρο και οι διασπορές του i-στού κρυμμένου κόμβου, αντίστοιχα. Έστω επίσης S_i το υποσύνολο εκείνο των δεδομένων εισόδου x που ενεργοποιούν τον κόμβο i περισσότερο από τους άλλους κόμβους:

$$S_i = \{x/\phi_i(x) = max_j(\phi_j(x))\} \tag{10.8}$$

Ο στόχος της γενετικής βελτιστοποίησης για κάθε κόμβο είναι να βρεθούν οι συνδέσεις εκείνες ανέμεσα στο επίπεδο και στο κρυμμένο επίπεδο που βελτιστοποιούν τη συνάρτηση ϕ_i στο πεδίο S_i.

Καθώς η επιλογή χαρακτηριστικών είναι από τους χώρους του οι γενετικές μέθοδοι έχουν δείξει ιδιαίτερα καλές επιδόσεις, αναμένουμε η προσέγγισή μας να εξάγει τους κανόνες με ιδανικό τρόπο. Τα σημαντικότερα πλεονεκτήματα των γενετικών αλγορίθμων είναι πως δεν απαιτούν ο χώρος των παραμέτρων που βελτιστοποιούν να είναι συνεχής καθώς και ότι μπορούν να βελτιστοποιούν συγχρόνως μεγάλο αριθμό παραμέτρων, δηλαδή να αναζητούν σε χώρους πολλών διαστάσεων. Η αναζήτηση ξεκινά με ένα πληθυσμό από πιθανές λύσεις, που στην περίπτωσή μας είναι μια δυαδική ακολουθία με τις τιμές του πίνακα A. Μέσω της γενετικής διαδικασίας ο πληθυσμός εξελίσσεται συνεχώς προς καλύτερες περιοχές του χώρου αναζήτησης, δηλαδή σε περιοχές με καλύτερες τιμές για τις παραμέτρους.

10.3.1 Δομή αλγορίθμου

Οι τυπικοί τελεστές της διαδικασίας είναι η επιλογή, η μετάλλαξη και ο συνδυασμός. Η επιλογή επιβραβεύει τις καλύτερες ακολουθίες (όπως αυτές επιλέγονται σύμφωνα με μια συνάρτηση αξιολόγησης) με το να τις χρησιμοποιεί πιο συχνά για την αναπαραγωγή του πληθυσμού. Έτσι οι νεώτεροι πληθυσμοί βασίζονται στις ακολουθίες που βρίσκονται στις καλύτερες περιοχές του χώρου αναζήτησης. Ο γενετικός συνδυασμός επιτρέπει την ανάμιξη δύο ακολουθιών, με το στόχο κάποιος από τους απογόνους να συνδυάζει τα καλύτερα στοιχεία καθενός από τους προγόνους. Η μετάλλαξη, τέλος, επιδρά τυχαία στις ακολουθίες, επιτρέποντας έτσι την εξερεύνηση και νέων περιοχών του χώρου των παραμέτρων.

Η διαδικασία αρχίζει με την τυχαία δημιουργία του αρχικού πληθυσμού. Στην περίπτωσή μας αυτό αντιστοιχεί στην τυχαία δημιουργία δυαδικών ακολουθιών με μήκος ίσο με το πλήθος των στοιχείων στον πίνακα A. Στη συνέχεια εφαρμόζεται αναδρομικά η ακόλουθη διαδικασία έως ότου αποφασιστεί ο τερματισμός του αλγορίθμου:

1. Κάθε ακολουθία του πληθυσμού αξιολογείται ως προς την ποιότητά του με βάση μια συνάρτηση καταλληλότητας.

2. Γενετικοί τελεστές εφαρμόζονται στον πληθυσμό για τη δημιουργία ενός νέου πληθυσμού. Ο τελεστής της επιλογής φροντίζει ο νέος πληθυσμός να είναι στατιστικά καλύτερος από τον προηγούμενο, ο τελεστής συνδυασμού να αναμιγνύεται η γενετική πληροφορία και ο τελεστής μετάλλαξης να εξερευνώνται νέες περιοχές.

3. Ο νέος πληθυσμός αντικαθιστά τον παλιό.

Σαν κριτήριο τερματισμού έχουμε επιλέξει την εκτέλεση ενός προκαθορισμένου αριθμού επαναλλήψεων.

Το πλέον καθοριστικό στοιχείο για την απόδοση ενός γενετικού αλγορίθμου είναι συνήθως η συνάρτηση καταλληλότητας. Καθώς ο στόχος μας είναι να βελτιστοποιήσουμε τον πίνακα A, η συνάρτηση καταλληλότητας πρέπει να αποτυπώνει αυτό που θεωρούμε ιδανικό για τις παραμέτρους που ο πίνακες περιέχει. Συγκεκριμένα, θα επιθυμούσαμε οι κανόνες να είναι κατά το δυνατό απλοί, δηλαδή ο πίνακας A να περιέχει πολλά μηδενικά, αλλά και ικανοί να δώσουν υψηλά επίπεδα ενεργοποίησης για τα δεδομένα εισόδου που αντιστοιχούν στην περιοχή τους. Έτσι καταλήγουμε στην ακόλουθη συνάρτηση καταλληλότητας:

$$F(\alpha_i) = \frac{\sum_{\underline{x} \in S_i} \phi_i(\underline{x})}{|\alpha_i|} \tag{10.9}$$

όπου $|\alpha_i|$ είναι η πληθικότητα του διανύσματος, δηλαδή ο αριθμός των μη μηδενικών στοιχείων που περιέχει. Μηδενικά διανύσματα δεν επιτρέπονται, καθώς αντιστοιχούν σε κανόνες που δεν δέχονται καμία είσοδο. Ο στόχος είναι η εύρεση ακολουθιών που μεγιστοποιούν τη συνάρτηση F.

10.3.2 Υλοποίηση γενετικών τελεστών

Η υλοποίηση των τριών γενετικών τελεστών που χρησιμοποιούνται είναι αρκετά απλή και παρουσιάζεται παρακάτω:

Επιλογή. Η συνάρτηση καταλληλότητας υπολογίζεται για όλες τις ακολουθίες του πληθισμού. Καθώς υψηλές τιμές αυτής της συνάρτησης πρέπει να αντιστοιχούν σε μεγάλη πιθανότητα επιλογής ακολουθούμε την ακόλουθη διαδικασία:

1. Τα μέλη του πληθυσμού διατάσσονται, αποκτώντας έτσι ένα δείκτη $z \in \{1, 2 \dots g\}$, όπου g το πλήθος των ακολουθιών στον πληθυσμό.

2. Υπολογίζεται το άθοισμα όλων των συναρτήσεων καταλληλότητας:

$$SF_i^{sum} = \sum_{k=1}^{g} F(\alpha_i^k) \tag{10.10}$$

   όπου α_i^k το k-στό διάνυσμα στον πληθυσμό που επεξεργαζόμαστε για την απλοποίηση του κανόνα i.

3. Το διάστημα $[0\ F_i^{sum}]$ χωρίζεται σε g υποδιαστήματα της μορφής $[SF_i^{j-1}\ SF_i^j]$ όπου

$$SF_i^j = \sum_{k=1}^{j} F(\alpha_i^k), j = 1, 2, \dots g \tag{10.11}$$

   και

$$SF_i^0 = 0 \tag{10.12}$$

4. Ένας αριθμός $R_0 \in [0\ F_i^{sum}]$ επιλέγεται τυχαία.

5. Η ακολουθία α_i^j που αντιστοιχεί στο διάστημα $[SF_i^{j-1}\ SF_i^j]$ στο οποίο ανήκει ο αριθμός R_0 επιλέγεται για τον ενδιάμεσο πληθυσμό, στον οποίο θα εφαρμοστούν οι υπόλοιποι τελεστές για να προκύψει ο επόμενος πληθυσμός. Καθώς το εύρος του διαστήματος που αντιστοιχεί σε μια ακολούθία α_i^j είναι $[SF_i^j - SF_i^{j-1}\ ] = F(\alpha_i^j$ πετυχαίνουμε την φυσική επιλογή που επιθυμούμε.

6. Η διαδικασία συνεχίζεται με τα βήματα 4 και 5 έως ότου επιλεγούν g ακολουθίες. Η επιλογή της ίδιας ακολουθίες περισσότερες από μια φορές είναι επιτρεπτή.

Συνδυασμός. Δεδομένων δύο ακολουθιών, επιλέγεται τυχαία ένα σημείο τους και οι τιμές τους από αυτό το σημείο και μετά ανταλλάζονται.

Μετάλλαξη. Με μια μικρή πιθανότητα επιλέγεται τυχαία κάποιο σημείο μια ακολουθίας και μετατρέπεται από μηδέν σε ένα ή αντίστροφα.

10.4 Πειραματικά αποτελέσματα

Για τον πειραματικό έλεγχο της διαδικασίας χρησιμοποιούμε και πάλι, όπως και σε προηγούμενα κεφάλαια, τα δεδομένα ίριδας.

Αρχικά ομαδοποιούμε τα δεδομένα και χρησιμοποιούμε τα αποτελέσματα για να αρχικοποιήσουμε ένα νευρωνικό δίκτυο ακτινικής βάσης και το εκπαιδεύουμε. Οι διαδικασίες παρουσιάζονται αναλυτικά στα κεφάλαια 8 και 9. Το εκπαιδευμένο δίκτυο μπορεί να κατηγοριοποιεί τα δεδομένα με επιτυχία 98%.

Στη συνέχεια εφαρμόζουμε τη γενετική μέθοδο που παρουσιάσαμε σε αυτό το κεφάλαιο για να απλοποιήσουμε το δίκτυο, ώστε οι εξαγώμενοι κανόνες να είναι πιο εύκολα κατανοητοί και αξιοποιήσιμοι.

Πίνακας 10.1: *Ο πίνακας παραμέτρων M*

$$\begin{bmatrix} 0.66 & 0.59 & 0.50 \\ 0.30 & 0.28 & 0.34 \\ 0.56 & 0.43 & 0.15 \\ 0.20 & 0.13 & 0.03 \end{bmatrix}$$

Πίνακας 10.2: *Ο πίνακας παραμέτρων Σ*

$$\begin{bmatrix} 0.13 & 0.10 & 0.07 \\ 0.06 & 0.06 & 0.08 \\ 0.11 & 0.10 & 0.04 \\ 0.05 & 0.04 & 0.02 \end{bmatrix}$$

Πίνακας 10.3: *Ο πίνακας παραμέτρων W*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Πίνακας 10.4: *Ο πίνακας παραμέτρων A*

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Οι παράμετροι που λαμβάνονται μετά από το τέλος και αυτής της διαδικασίας παρουσιάζονται στους πίνακες 10.1, 10.2, 10.3 και 10.4. Βλέπουμε πως ο πίνακας A έχει απλοποιηθεί αρκετά καθώς οι μισές τιμές του είναι τώρα ίσες με το μηδέν. Έτσι και οι εξαγώμενοι κανόνες έχουν αντίστοιχα τις μισές μεταβλητές στο τμήμα if. Αξίζει να σημειωθεί πως η ποιότητα των κανόνων δεν έχει ελαττωθεί, καθώς το δίκτυο του σχάματος 10.1 κατηγοριοποιεί τα δεδομένα ίριδας με επιτυχία 98%, δηλαδή ίδια με αυτή του εκπαιδευμένου δικτύου, όταν αρχικοποιείται με τις παραμέτρους των πινάκων 10.1, 10.2, 10.3 και 10.4.

□

Κεφάλαιο 11

Συμπεράσματα – Επεκτάσεις

Με την παρούσα διδακτορική διατριβή επιχειρήσαμε να συγκεντρώσουμε και να επεκτείνουμε τη θεωρία και την πρακτική στο χώρο της ευφυούς λειτουργίας σε καθεστώς αβεβαιότητας. Μέσα τόσο από τη βιβλιογραφική έρευνα, όσο και από τα πειραματικά μας αποτελέσματα, έγινε σαφές πως η αβεβαιότητα έχει ένα σημαντικό ρόλο στα σύγχρονα ευφυή συστήματα. Τέλος, όπως θα εξηγήσουμε παρακάτω, μέσα από τη θεωρία που αναπτύχθηκε ανοίγονται μια σειρά από νέους δρόμους για περαιτέρω διερεύνηση.

Στο πρώτο τμήμα της διατριβής εστιάσαμε στο σημασιολογικό επίπεδο. Εκεί, είδαμε πως η αβεβαιότητα είναι εγγενής στον πραγματικό κόσμο και συνεπώς αντιστοίχως ευέλικτα μοντέλα γνώσης χρειάζεται να αναπτυχθούν για την αναπαράστασή του. Προς αυτή την κατεύθυνση προτείναμε την ασαφή σχεσιακή αναπαράσταση της γνώσης. Παράλληλα με την εξέλιξη της παρούσας διατριβής ο χώρος των οντολογιών έχει σημειώσει μεγάλη πρόοδο στην τυποποίηση των σχέσεων που οφείλει κανείς να χρησιμοποιεί για να περιγράφει τον πραγματικό κόσμο, χωρίς όμως αυτές να είναι ασαφής. Θα είνα ενδιαφέρον να εξατάσει κανείς πώς οι δύο αυτές συνεισφορές και προσεγγίσεις μπορούν να συνδυαστούν, ώστε να προκείψει ένα ακόμη πιο εκφραστικό μοντέλο αναπαράστασης της γνώσης.

Ασχοληθήκαμε επίσης με υπολογιστικά θέματα, προτείνοντας ένα αποδοτικό μοντέλο αραιής αναπαράστασης και δύο ιδιαίτερα ταχείς αλγορίθμους μεταβατικού κλεισίματος ειδικά για την περίπτωση αραιών σχέσεων. Κατά τη φάση της βιβλιογραφικής επισκόπησης για αυτή την ενότητα έγινε καθαρό πως πολλοί λίγοι αλγόριθμοι έχουν αναπτυχθεί ειδικά για σχέσεις μεγάλες και αραιές, καθώς δεν υπήρχε στο παρελθόν η ανάλογη πρακτική ανάγκη. Καθώς το πεδίο των οντολογιών κερδίζει συνεχώς σε έδαφος, και οι σχέσεις που εμφανίζονται σε αυτό είναι σχεδόν πάντα μεγάλες και αραιές, είναι ενδιαφέρον να δει κανείς ποιοί άλλοι αλγόριθμοι μπορούν να αναθεωρηθούν ειδικά για μεγάλες και αραιές σχέσεις. Επίσης, αξίζει να εξεταστεί η εφαρμογή των προτεινόμενων υπολογιστικών μοντέλων και σε άλλους χώρους με συναφείς ανάγκες. Για παράδειγμα, οι αλγόριθμοι μεταβατικού κλεισίματος ίσως βρουν εφαρμογή στο πεδίο της δρομολόγησης πακέτων στα δίκτυα δεδομένων.

Στο δεύτερο τμήμα της διατριβής ασχοληθήκαμε με την αποτίμηση ασαφών κανόνων. Είδαμε πως οι συνήθεις μεθοδολογίες αποτίμησης ασαφών κανόνων δεν είναι ιδανικές για τις περιπτώσεις που οι είσοδοι δεν προέρχονται από σένσορες αλλά από την έξοδο άλλων περίπλοκων υπολογιστικών συστημάτων. Για αυτή την περίπτωση προτείναμε μια νέα μεθοδολογία αποτίμησης ασαφών κανόνων που έχει δυνατοτικό χαρακτήρα. Η προτεινόμενη μεθοδολογία είναι σε θέση τόσο να λειτουγεί ικανοποιητικά ακόμη και όταν το περιβάλλον της είναι αβέβαιο, όσο και να προσδιορίζει το

ποσοστό αβεβαιότητας που αντιστοιχεί στην κάθε έξοδό της. Αυτό που έχει ιδιαίτερο ενδιαφέρον είναι να μελετήσει κανείς πώς παρόμοιες μέθοδοι αποτίμησης μπορούν να συνδυαστούν με νευροασαφή δίκτυα, ώστε να διαθέτουν και την ιδιότητα της αυτόματης μάθησης.

Ενώ τα δύο πρώτα τμήματα της διατριβής πραγματεύτηκαν την αξιοποίηση γνώσης σε πρακτικά προβλήματα, το τρίτο και τελευταίο μέρος της διατριβής πραγματεύεται την αυτόματα εξαγωγή της γνώσης από αριθμητικά δεδομένα. Σε αυτή τη διαδικασία παρατηρούμε πως υπάρχουν μια πληθώρα από μεθοδολογίες που μπορεί κανείς να χρησιμοποιήσει για να αναλύσει αριθμητικά δεδομένα, αλλά δεν υπάρχουν σαφείς και ασφαλείς τρόποι για να επιλέξει κανείς την ιδανική μεθοδολογία ανάλυσης για την κάθε περίπτωση. Έτσι, γίνεται εμφανές πως η ανάγκη είναι να αναπτυχθούν μεθοδολογίες που κατά το δυνατόν δεν θα χρειάζονται αρχικοποίηση με βάση προϋπάρχουσα γνώση, ή σημαντική παραμετροποίηση από το χρήστη.

Προς αυτή την κατεύθυνση παρουσιάσαμε μια μη εποπτευμένη διαδικασία για την ομαδοποίηση δεδομένων η οποία είναι σε θέση αυτόματα να διακρίνει τους υποχώρους που ορίζουν μια ομάδα και να χρησιμοποιεί αυτή την πληροφορία στα βήματα που ακολουθούν. Αφού χρησιμοποιήσουμε τα αποτελέσματα της ομαδοποίησης για να αρχικοποιήσουμε και να εκπαιδεύσουμε νερωνικά δίκτυα ώστε να ανιχνευθούν τα ελλοχεύοντα πρότυπα με ακόμη μεγαλύτερη ακρίβεια, χρησιμοποιούμε μια γενετική μεθοδολογία για την αυτόματη εξαγωγή ασαφών κανόνων από τα δεδομένα εισόδου. Θα ήταν ενδιαφέρον αυτή η μεθοδολογία να συνδυαστεί με τη μεθοδολογία δυνατοτικής αποτίμησης που προτάθηκε νωρίτερα στη διατριβή, ώστε να προκύψει ένα ολοκληρωμένο, αυτοεκπαιδευόμενο ευφυές σύστημα που θα χρειάζεται ελάχιστη ή και μηδενική παρεμβολή από το χρήστη για να αρχίσει να λειτουργεί και να λαμβάνει αποφάσεις.

Συνολικά, είδαμε πως η αβεβαιότητα είναι παρούσα σε όποιο επίπεδο και αν εξετάσουμε τη λειτουργία των ευφυών συστημάτων και ο ρόλος της είναι καθοριστικός. Έτσι προτείναμε μια σειρά από λύσεις, η οποίες με τη σειρά τους ανοίγουν μια νέα σειρά από δρόμους. Ελπίζουμε σύντομα και αυτοί οι δρόμοι να περπατηθούν, προσφέροντάς μας μεγαλύτερη κατανόηση της αβεβαιότητας καθώς και πιό ισχυρά εργαλεία για το χειρισμό της.

□

Κεφάλαιο 12

Βιογραφικό Σημείωμα

Σπουδές

2001–	Υποψήφιος διδάκτορας και μέλος του Εργαστηρίου Εικόνας, Βίντεο και Συστημάτων Πολυμέσων, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο.
1996–2001	Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο. Τίτλος διπλωματικής εργασίας: "Μελέτη και Ανάπτυξη Πλατφόρμας για τη Δημιουργία Διαδικτυακής Εφαρμογής Δημοσίων Σχέσεων"
1995–1996	Σχολή Μηχανολόγων Μηχανικών, Εθνικό Μετσόβιο Πολυτεχνείο.

Επαγγελματική και ακαδημαϊκή εμπειρία

2004–	Πρόεδρος Τμήματος Πληφορικής, University of Indianapolis, Athens Campus.
2001–	Λέκτορας Πληφορικής, University of Indianapolis, Athens Campus.
2001–2004	Επικουρικό διδακτικό έργο, Ψηφιακή επεξεργασία εικόνας, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών, Εθνικό Μετσόβιο Πολυτεχνείο.
2001	Επικουρικό διδακτικό έργο, Εισαγωγή στον Προγραμματισμό, Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, Εθνικό Μετσόβιο Πολυτεχνείο.
2001–2003	Σεμινάρια πληροφορικής. Διάφοροι οργανισμοί διοργάνωσης σεμιναρίων με τη στήριξη του ΟΑΕΔ.
2000–2001	Μηχανικός λογισμικού και υπεύθυνος ομάδας ανάπτυξης έργου, Innovart A.E.

Συμμετοχή σε ερευνητικά προγράμματα

2002–2005	ΗΡΑΚΛΕΙΤΟΣ. Ευφυής Εννοιολογική Πρόσβαση σε Πολυμεσική Πληροφορία. Συγχρηματοδοτουέμενο από το Ευρωπαϊκό Κοινοτικό Ταμείο (75%) και από Εθνικούς Πόρους (25%).
2004–2005	ERMIS. Emotinally Rich Man-machine Intelligent Sustem. Χρηματοδοτούμενο από την Ευρωπαϊκή Ένωση.
2001–2004	FAETHON. Unified Intelligent Access to Heterogeneous Audiovisual Content. Χρηματοδοτούμενο από την Ευρωπαϊκή Ένωση.
2004	ΣΥΝΕΝΝΟΗΣΗ. Σύστημα για την αναλλαγή πληροφορίας και την επικοινωνία ατόμων με προβλήματα ακοής. Χρηματοδοτούμενο απο το Υπουργείο Μεταφορών και Επικοινωνιών.
2002–2003	ORESTEIA - Modular Hybrid Artefacts with Adaptive Functionality. Χρηματοδοτούμενο από την Ευρωπαϊκή Ένωση.

Βραβεία – Υποτροφίες

2002–2005	ΗΡΑΚΛΕΙΤΟΣ: υποτροφία με έμφαση στη βασική έρευνα για την εκπόνηση διατριβής στο πεδίο "Ευφυής Εννοιολογική Πρόσβαση σε Πολυμεσική Πληροφορία".
2004	Travel Grant για την εργασία "Computationally efficient incremental transitive closure of sparse fuzzy binary relations", IEEE Neural Networks Society.

Ξένες γλώσσες

Αγλικά	Cambridge Proficiency
Γαλλικά	Diplôme d'études en langue française
Ισπανικά	Básico

Συμμετοχή σε συλλογικούς φορείς

2002-	Institute of Electrical and Electronics Engineers → IEEE Systems, Man and Cybernetics Society → IEEE Computer Society → IEEE Computational Intelligence Society
2001–	Τεχνικό επιμελητήριο Ελλάδας
2003–	Soft Computing in Image Processing Working Group
2001–	Πανελλήνιος Σύλογος Διπλωματούχων Μηχανολόγων και Ηλεκτρολόγων

Εθελοντικοί επιστημονικοί ρόλοι

Πρόεδρος οργαν. επιτροπής	3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI) 2006
Προετ. ειδικού τεύχους	→ Intelligent Image and Video Processing and Applications: The Role of Uncertainty στο περιοδικό International Journal of Intelligent Systems Technologies and Applications. → Image, Signal and Distributed Data Processing for Networked eHealth Applications στο περιοδικό IEEE Engineering in Medicine and Biology Magazine.
Διοργ. ειδικής συνεδρίας	→ Uncertainty in Image and Video Processing στο συνέδριο IEEE International Conference on Fuzzy Systems, Reno, Nevada, May 2005. → Image, Signal and Distributed Data Processing for Networked eHealth Applications στο συνέδριο International Network Conference, Samos, Greece, July 2005.
Πρόεδρος συνεδρίας	Data Mining στο συνέδριο International Fuzzy Systems Association World Congress (IFSA), Istanbul, Turkey, June-July 2003.
Κριτής περιοδικού	→ IEEE Transactions on System, Man and Cybernetics - Part B → IEEE Transactions on Neural Networks → IEEE Transactions on System, Man and Cybernetics - Part A → IEEE Transactions on System, Man and Cybernetics - Part C → Fuzzy Sets and Systems → IEEE Transactions on Circuits and Systems for Video Technology → Neural Networks → Multimedia Tools and Applications → International Journal of Modeling and Simulation → IEEE Transactions on Education → Educational Technology and Society
Κριτής κεφαλαίων βιβλίου	Ma Z. (ed) Soft Computing in Ontologies and Semantic Web, Springer,2006
Κριτής συνεδρίου	→ IEEE International Conference on Fuzzy Systems 2005,2004,2003 → International Conference on Artificial Neural Networks 2003 → International Conference on Education and Information Systems 2005, 2004 → ASEE/IEEE Frontiers in Education 2004 → World Multi-Conference on Systemics 2005, 2004 → IEEE International Symposium on Industrial Electronics 2005

□

Κεφάλαιο 13

Κατάλογος δημοσιεύσεων

13.1 Περιοδικά

13.1.1 Δημοσιευμένα

1. **Wallace M.**, Avrithis Y., Kollias S., "Computationally efficient sup-t transitive closure for sparse fuzzy binary relations" Fuzzy Sets and Systems, accepted for publication.

2. **Wallace M.**, Athanasiadis T., Avrithis Y., Delopoulos A., Kollias S. "Integrating Multimedia Archives: The Architecture and the Content Layer" IEEE Trans. Systems, Man, and Cybernetics, Part A: Systems and Humans, accepted for publication.

3. **Wallace M.**, Tsapatsoulis N., Kollias S. "Intelligent Initialization of Resource Allocating RBF Networks" Neural Networks 18(2), pp. 117-122, 2005.

4. **Wallace M.**, Ioannou S., Karpouzis K., Kollias S., "Dealing with Feature Uncertainty in Facial Expression Recognition", International Journal of Intelligent Systems Technologies and Applications, accepted for publication.

5. Athanasiadis T., **Wallace M.**, Karpouzis K., Nikolakopoulos Y., Kollias S., "Utilization of evidence theory in the detection of salient regions in successive CT images", International Journal of Oncology, accepted for publication.

6. **Wallace M.**, Karpouzis K., Stefanou M., Maglogiannis I., Kollias S. "Electronic Roads in Historical Documents: a Student Oriented Approach" Education and Information Technologies 9(3), pp. 271ὐ289, 2004.

7. **Wallace M.**, Maglogiannis. I., Karpouzis, K., Kormentzas, G., Kollias, S., "Intelligent One-Stop-Shop Travel Recommendations Using an Adaptive Neural Network and Clustering of History" Information Technology & Tourism 6(3), 2004.

8. **Wallace M.**, Karpouzis, K., Stamou, G. Moschovitis, G. Kollias, S., Schizas, C., "The Electronic Road: Personalised Content Browsing", IEEE Multimedia 10(4), pp. 49-59, 2003.

9. Avrithis Y., Stamou G., **Wallace M.**, Marques F., Salembier P., Giro X., Haas W., Vallant H., Zufferey M., "Unified Access to Heterogeneous Audiovisual Archives" Journal of Universal Computer Science 9(6), pp. 510-519, 2003

13.1.2 Υποβεβλημένα προς κρίση

10. Ioannou S., **Wallace M.**, Karpouzis K., Kollias S., "Facial Expression Recognition Using Possibilistic Fuzzy Rule Evaluation", submitted to IEEE Transactions on SMC, Part B in August 2005.

11. Mylonas P., Athanasiadis T., **Wallace M.**, Avrithis Y., Kollias S., "Semantic Representation of Multimedia Content - Part I: Data Models, Analysis & Indexing" submitted to IEEE Transactions on Multimedia in July 2005.

12. **Wallace M.**, Kollias, S., "Two algorithms for fast incremental transitive closure of sparse fuzzy binary relations, submitted to the International Journal of Computational Methods in December 2004.

13. Ioannou S., **Wallace M.**, Karpouzis K., Kollias S., "Robust Facial Analysis For Expression Recognition" submitted to Behavior Research Methods in October 2005.

14. **Wallace M.**, Georgiou N., "Determining whether web based tools are suitable for a class" submitted to Educational Technology & Society in July 2005.

13.2 Editorials

15. **Wallace M.**, Kollias S., "Intelligent Image and Video Processing and Applications: The Role of Uncertainty" International Journal of Intelligent Systems Technologies and Applications, 2006.

16. Maglogiannis I., Karpouzis K., **Wallace M.**, "Image, Signal and Distributed Data Processing for Networked eHealth Applications, IEEE Engineering in Medicine and Biology Society 2006.

13.3 Βιβλία

17. Maglogiannis I., Karpouzis K., **Wallace M.**, "Image, Signal and Distributed Data Processing for Networked eHealth Applications" Morgan & Claypool Publishers, 2006.

13.4 Κεφάλαια σε βιβλία

18. **Wallace M.**, Avrithis Y., Stamou G., Kollias S., "Knowledge-based Multimedia Content Indexing and Retrieval" in Stamou G., Kollias S. (eds) Multimedia Content and Semantic Web: Methods, Standards and Tools, Wiley, 2005.

19. **Wallace M.**, Mylonas P., Akrivas G., Avrithis Y., Kollias S., "Automatic thematic categorization of multimedia documents using ontological information and fuzzy algebra",in Ma Z. (ed) Soft Computing in Ontologies and Semantic Web, Springer, 2006.

20. **Wallace M.**, Avrithis Y., Kollias S., "Multimedia archives and mediators" in Furht B. (ed) Encyclopedia of Multimedia, Springer, 2006.

13.5 Συνέδρια

21. **Wallace M.**, Kollias S., "Possibilistic Evaluation of Extended Fuzzy Rules in the Presence of Uncertainty", Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Reno, Nevada, May 2005.

22. **Wallace M.**, Tsapatsoulis N. "Combining GAs and RBF Neural Networks for Fuzzy Rule Extraction from Numerical Data" International Conference on Artificial Neural Networks, Warsaw Poland, September 2005.

23. Falelakis M., Diou C., **Wallace M.**, Delopoulos A. "Minimizing Uncertainty in Semantic Identification when Computing Resources are Limited" International Conference on Artificial Neural Networks, Warsaw Poland, September 2005.

24. Ioannou S., **Wallace M.**, Karpouzis K., Kollias S. "A Robust Scheme for Facial Analysis and Expression Recognition" Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands, 30 August - 2 September, 2005.

25. Ioannou S., **Wallace M.**, Karpouzis K., Raouzaiou A., Kollias S. "Combination of Multiple Extraction Algorithms in the Detection of Facial Features" Proceedings of the IEEE International Conference on Image Processing (ICIP), Genova, Italy, September 2005.

26. Cowie R., Douglas-Cowie E., Taylor J., Ioannou S., **Wallace M.**, Kollias S. "An Intelligent System for Facial Emotion Recognition" IEEE International Conference on Multimedia & Expo, Amsterdam, The Netherlands, July 6-8, 2005.

27. Ioannou S., **Wallace M.**, Karpouzis K., Raouzaiou A., Kollias S., "Confidence-Based Fusion of Multiple Feature Cues for Facial Expression Recognition", Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Reno, Nevada, May 2005.

28. Stefanou H., Kakouros S., Cavouras D., **Wallace M.**, "Wavelet-based Mammographic Enhancement" Proceedings of the Fifth International Network Conference (INC), July 2005, Samos, Greece.

29. **Wallace M.**, Kollias S., "Computationally efficient incremental transitive closure of sparse fuzzy binary relations", Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Budapest, Hungary, July 2004.

30. **Wallace M.**, Avrithis Y., "Fuzzy Relational Knowledge Representation and Context in the Service of Semantic Information Retrieval", Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Budapest, Hungary, July 2004.

31. **Wallace M.**, Raouzaiou A., Tsapatsoulis N., Kollias S., "Facial Expression Classification based on MPEG-4 FAPs: The use of evidence and prior knowledge for uncertainty removal", Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Budapest, Hungary, July 2004.

32. **Wallace M.**, Kollias S., "Robust, Generalized, Quick and Efficient Agglomerative Clustering" Proceedings of 6th International Conference on Enterprise Information Systems (ICEIS), Porto, Portugal, April 2004.

33. **Wallace M.**, Mylonas, P., Kollias, S., "Automatic Extraction of Semantic Preferences from Multimedia Documents" Proceedings of 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), April 21-23, 2004, Lisboa, Portugal, April 2004.

34. **Wallace M.**, Athanasiadis T., Avrithis Y., "Knowledge Assisted Analysis & Categorization for Semantic Video Retrieval" Proceedings of 3rd International Conference on Image and Video Retrieval (CIVR), July 2004, Dublin, Ireland.

35. **Wallace M.**, Stefanou H., Kollias S., "Iterated Function Systems as Human Perceivable Spectral Tests of Randomness" Proceedings of the International Conference on Non-Linear Analysis, Non-Linear Systems and Chaos (NOLASC), December 2004, Vouliagmeni, Greece.

36. **Wallace M.**, Karpouzis K., "Personalized Content Browsing Based on Notions of Context" Proceedings of the International Conference on Multi-platform e-Publishing, November 2004, Athens, Greece.

37. **Wallace M.**, Athanasiadis T., Avrithis Y., Stamou G., Kollias S., "A mediator system for hetero-lingual audiovisual content" Proceedings of the International Conference on Multi-platform e-Publishing, November 2004, Athens, Greece.

38. Mylonas P., **Wallace M.**, Kollias S., "Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering" Proceedings of 3rd Hellenic Conference on Artificial Intelligence (SETN), Samos, Greece, May 2004.

39. Andreou G., Mylonas P., **Wallace M.**, Kollias S., "Offering Access to Personalized Interactive Video" Proceedings of the International Conference on Mathematical Methods and Computational Techniques in Electrical Engineering (MMACTEE), December 2004, Vouliagmeni, Greece.

40. **Wallace M.**, Akrivas, G. and Stamou, G., "Automatic Thematic Categorization of Documents Using a Fuzzy Taxonomy and Fuzzy Hierarchical Clustering", Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), St. Louis, MO, USA, May 2003.

41. **Wallace M.** and Kollias, S., "Soft Attribute Selection for Hierarchical Clustering in High Dimensions", Proceedings of the International Fuzzy Systems Association World Congress(IFSA), Istanbul, Turkey, June-July 2003.

42. **Wallace M.**, Mylonas, P. and Kollias, S., "Detecting and Verifying Dissimilar Patterns in Unlabelled Data" 8th Online World Conference on Soft Computing in Industrial Applications (WSC8), September - October 2003.

43. **Wallace M.**, Akrivas, G., Mylonas, P., Avrithis, Y., Kollias, S. "Using context and fuzzy relations to interpret multimedia content", Proceedings of the Third International Workshop on Content-Based Multimedia Indexing (CBMI), IRISA, Rennes, France, September 2003.

44. Tsapatsoulis, N., **Wallace M.** and Kasderidis, S., "Improving the Performance of Resource Allocation Networks through Hierarchical Clustering of High – Dimensional Data", Proceedings of the International Conference on Artificial Neural Networks (ICANN), Istanbul, Turkey, June 2003.

45. Avrithis, Y., Stamou, G, **Wallace M.**, Marques, F., Salembier, P., Giro, X., Haas, W., Vallant, H. and Zufferey, M., "Unified Access to Heterogeneous Audiovisual Archives", Proceedings of the 3rd International Conference on Knowledge Management (I-KNOW), Graz, Austria, July 2003.

46. **Wallace M.**, Stefanou, M., Karpouzis, K. and Kollias, S., "Towards supporting the teaching of history using an intelligent information system that relies on the electronic road metaphor", Proceedings of the 3rd IEEE International Conference on Advanced Learning Technologies (ICALT), Athens, Greece, July 2003.

47. **Wallace M.** and Stamou, G., "Towards a Context Aware Mining of User Interests for Consumption of Multimedia Documents", Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Lausanne, Switzerland, August 2002.

48. **Wallace M.**, Akrivas, G., Stamou, G. and Kollias, S., "Representation of user preferences and adaptation to context in multimedia content – based retrieval", Proceedings of the Workshop on Multimedia Semantics, SOFSEM 2002: Theory and Practice of Informatics, Milovy, Czech Republic, November 2002.

49. Akrivas, G., **Wallace M.**, Stamou, G. and Kollias, S., "Context - Sensitive Query Expansion Based on Fuzzy Clustering of Index Terms", Proceedings of the Fifth International Conference on Flexible Query Answering Systems (FQAS), Copenhagen, Denmark, October 2002.

50. Akrivas, G., **Wallace M.**, Andreou, G., Stamou, G. and Kollias, S., "Context - Sensitive Semantic Query Expansion", Proceedings of the IEEE International Conference on Artificial Intelligence Systems (ICAIS), Divnomorskoe, Russia, September 2002.

13.6 Τεχνικές Αναφορές

51. **Wallace M.** "Συσχετιστική Ανάδραση και Συστήματα Ανάκτησης Πληροφορίας", Τεχνική Αναφορά 2001_01, Εργαστήριο Ψηφιακής Επεξεργασίας Εικόνας, Βίντεο και Συστημάτων Πολυμέσων, Ιούνιος 2001.

□

Βιβλιογραφία

[1] Addison D., Wermter S., Arevian G. (2003) A Comparison of Feature Extraction and Selection Techniques. ICANN:212–215.

[2] Adelson-Velskii G.M., Landis E.M. (1962) An algorithm for the organization of information. Doklady Akademia Nauk SSSR 146:263–266. Also as Adelson-Velskii G.M., Landis E.M. (1962) An algorithm for the organization of information. English translation in Soviet Math 3:1259-1263.

[3] Aggarwal, C.C., Yu, P.S. (2002) Redefining clustering for High–Dimensional Applications. IEEE Transactions on Knowledge and Data Engineering 14(2):210–225.

[4] Akrivas G. and Stamou G. (2001) Fuzzy Semantic Association of Audiovisual Document Descriptions, International Workshop on Very Low Bitrate Video Coding (VLBV), Athens, Greece.

[5] Akrivas G., Wallace M., Andreou G., Stamou G. and Kollias S. (2002) Context - Sensitive Semantic Query Expansion. ICAIS, Divnomorskoe, Russia.

[6] Athanasiadis T., Avrithis Y. (2004) Adding Semantics to Audiovisual Content: The FAETHON Project. in Enser P., Kompatsiaris Y., OŏConnor N.E., et al. (Eds) Image and Video Retrieval, LNCS 3115:665–673.

[7] Avrithis Y., Stamou G., Wallace M., Marques F., Salembier P., Giro X., Haas W., Vallant H.,Zufferey M. (2003) Unified Access to Heterogeneous Audiovisual Archives .Journal of Universal Computer Science 9(6):510-519.

[8] Backhouse R.C. (1992) Calculating the Floyd-Warshall path algorithm. Eindhoven University of Technology.

[9] Backhouse R.C., van den Eijnde J.P.H.W., van Gasteren A.J.M. (1994) Calculating path algorithms. Sci. Comput. Programming 22(3):3–19.

[10] Backhouse R.C., van Gasteren A.J.M. (1993) Calculating a Path Algorithm. Lecture Notes in Computer Science 669:32–44.

[11] Baeza-Yates, R.A., Ribeiro-Neto, B.A. (1999) Modern Information Retrieval. ACM Press / Addison-Wesley.

[12] Bagui, S.C., Bagui, S., Pal, K., Pal, N.R. (2003) Breast cancer detection using rank nearest neighbor classification rules. Pattern Recognition 36:25–34.

[13] Baker, J.J. (1962) A note on multiplying Boolean matrices. Comm. ACM 5(2):102.

[14] Balabanovic, M., Shoham, Y. (1997) Fab: content - based collaborative recommendation. Communications of the ACM 40(3):66–72.

[15] Barry, C. (1994) User-defined relevance criteria: An explanatory study. Journal of the American Society for Information Science 45(3):149–159.

[16] Batchelor, B.G. (1978) Classification and data analysis in vector spaces in Pattern Recognition. Batchelor, B.G. (editor) Plenum Press.

[17] Battista, S., Casalino, F., Lande C. (1999) MPEG-4: A Multimedia Standard for the Third Millenium, Part 1. IEEE Multimedia 6(4):74–83.

[18] Battista, S., Casalino, F., Lande, C. (2000) MPEG-4: A Multimedia Standard for the Third Millenium, Part 2. IEEE Multimedia 7(1):76–84.

[19] Belkin, N. J., Cool, C., Stein, A., Thiel, U. (1995) Cases, Scripts, and Information-Seeking Strategies: On the Design of Interactive Information Retrieval Systems. Expert Systems with Applications 9(3):379–395.

[20] Bedek J.C., Biswas G. and Huang L.Y., (1986) Transitive closures of fuzzy thesauri for information retrieval systems. International Journal of Man-Machine Studies 25:343–356.

[21] Biswas, G., Bezdek, J. C., Marques, M., Subramanian, V. (1987) Knowledge-assisted document retrieval. Journal of the American Society for Information Science 38(2):83–110.

[22] Bordogna, G., Carrara, P., Pasi, G. (1995) Fuzzy approaches to extend Boolean Information retrieval. Fuzziness in Database Management Systems, Studies in Fuzziness series 5:231–274.

[23] Bordogna, G., Carrara, P., Pasi, G. (1991) Query term weights as constraints in fuzzy information retrieval. Information Processing and Management 27(1):15–26.

[24] Bordogna, G., Pasi, G. (1993) A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation. Journal of the American Society for Information Science 44(2):70–82.

[25] Bowman C.M., Danzing P.B., Manber U., Schwartz F. (1994) Scalable Internet resources discovery: research problems and approaches. Communications of the ACM 37:98–107.

[26] Brunzell, H., Eriksson, J. (2000) Feature reduction for classification of multidimensional data. Pattern Recognition 33: 1741–1748.

[27] Buckley, C., Allan, J., Salton, G. (1995) Automatic Routing and Retrieval Using Smart: TREC-2. Information Processing and Management 31(3):315–326.

[28] Buel, D. A. (1985) A problem in information retrieval with fuzzy sets. Journal of the American Society for Information Science 36(6):398–401.

[29] Buel, D. A. (1982) An analysis of some fuzzy subset applications to information retrieval systems. Fuzzy Sets and Systems 7(1):35–42.

[30] Boixader D., Jacas J. and Recasens J. (2001) Transitive closure and betweenness relations. Fuzzy Sets & Systems 120:415–422.

[31] Chang, C. H., Hsu, C. C. (1998) Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval. Computer Networks and ISDN Systems 30(1-7):621–623.

[32] Chen, J., Mikulcic, A. and Kraft, D. H. (1998) An Integrated Approach to Information Retrieval with Fuzzy Clustering and Fuzzy Inferencing. Pons, O.,Ampara Vila, M., Kacprzyk, J. (editors) Knowledge Management in Fuzzy Databases. Physica Verlag, Heidelberg, Germany.

[33] Chen, P. M. and Kuo, F. C. (2000) An information retrieval system based on a user profile. Journal of Systems and Software 54(1):3–8.

[34] Chen S.M. (1994) A weighted fuzzy reasoning algorithm for medical diagnosis. Decision Support Systems 11:37–43.

[35] Chen S.-M., Horng Y.-J. and Lee C.-H. (2003) Fuzzy information retrieval based on multi-relationship fuzzy concept networks. Fuzzy Sets & Systems 140:183–205.

[36] Ciocca G. και Schettini R. (1999) A relevance feedback mechanism for content-based image retrieval. Information Processing and Management 35(5):605–632.

[37] Constantinou, C., Kakas, A., Katsikides, S., Papadopoulos, G., Pattichis, C., Pitsillides, A., Schizas, C. (1999) Constructing Electronic Roads in the Information Society: A Case Study in Cyprus. Proceedings of Neties99, Krems, Austria.

[38] Croft, B., Cook, R. and Wilder, D. (1995) Government Information on the Internet: Experiences with TOMAS. Proceedings of Digital Libraries (DL'95).

[39] Cross, V. (1994) Fuzzy information retrieval. Journal of Intelligent Information systems 3(1):29–56.

[40] Dasgupta M. and Deb R. (2001) Factoring fuzzy transitivity. Fuzzy Sets & Systems 118:489–502.

[41] Dash, M., Liu, H., Scheuermann, P. and Tan, K.L. (2003) Fast hierarchical clustering and its validation. Data and Knowledge Engineering 44(1):109–138.

[42] Dawyndt P., De Meyer H., De Baets B. (in press) The complete linkage clustering algorithm revisited. Soft Computing.

[43] De Baets B., De Meyer H. (2003) On the existence and construction of T-transitive closures. Inform. Sc. 152:167–1793.

[44] De Baets B., De Meyer H. (2003) Transitive approximation of fuzzy relations by alternating closures and openings. Soft Computing:210–219.

[45] De Baets B., De Meyer H., Naessens H. (in press) A top-down algorithm for generating the Hasse tree of a fuzzy preorder closure. IEEE Trans. Fuzzy Systems.

[46] Del Bimbo, A. (1999) Visual Image Retrieval. Morgan Kaufmann, San Francisco.

[47] Dimitrova, N., Zhang, H.-J., Shahraray, B., Sezan, I., Huang, T., Zakhor, A. (2002) Applications of video-content analysis and retrieval. IEEE Multimedia 9(3):42–55.

[48] Dong M., Ravi Kothari, R. (2003) Feature subset selection using a new definition of classifiability. Pattern Recognition Letters 24:1215–1225

[49] Doulamis, N., Doulamis, A., Kollias, S. (2000) On-line Retrainable Neural Networks: Improving Performance of Neural Networks in Image Analysis. IEEE Transactions on Neural Networks 11(1):1–20.

[50] Duan J.-S. (2004) The transitive closure, convergence of powers and adjoint of generalized fuzzy matrices. Fuzzy Sets & Systems 145:301–311.

[51] Dubois D. and Prade H. (1997) The three semantics of fuzzy sets. Fuzzy Sets and Systems 90:142–150.

[52] Dunn J.C. (1974) A graph-theoretical analysis of pattern classification via Tamura's fuzzy relation. IEEE Trans. SMC 5:310–313.

[53] Dunn J.C. (1974) Some recent investigations of a new fuzzy partitioning algorithm and its applications to pattern classification problems. Journal of Cybernetics 4:1–15.

[54] El-Sonbaty, Y., Ismail, M. A., (1998) On-line hierarchical clustering. Pattern Recognition Letters 19:1285–1291.

[55] Feijs L.M.G., van Ommering R.C. (1997) Abstract derivation of transitive closure algorithms. Information Processing Letters 63:159–164.

[56] Fellbaum, F. (ed) (1998) WordNet: An Electronic Lexical Database. MIT Press.

[57] Fischler, R.A., Firchein, O., (1997) Intelligence: The eye, the brain and the computer. Reading, MA, Assison-Wesley.

[58] Fodor, J. and Roubens M. (1995) Structure of transitive valued binary relations. Mathematical Social Sciences 30:71–94.

[59] Fuhr, N. (1989) Models for retrieval with probabilistic indexing. Information Processing and Management 25(2):55–72.

[60] Ganter, B., Wille, R. (1999) Formal Concept Analysis, Springer.

[61] Gosh, J., Nag, A. (2000) An overview of Radial Basis Function Networks. Howlerr, J., Jain, L.C. (editors) Radial Basis Function Neural Network Theory and Applications, Physica-Verlag.

[62] Guu S.-M., Chen H.-H. and Pang C.-T. (2001) Convergence of products of fuzzy matrices. Fuzzy Sets & Systems 121:203–207.

[63] Halgamuge, S., Glesner, M. (1994) Neural Networks in designing fuzzy systems for real world applications. Fuzzy Sets and Systems 65:1-12.

[64] Harman, D. (19992) Relevance feedback revisited. Proceedings of ACM SIGIR 1992, International Conference on Research and Development in Information Retrieval.

[65] Haykin, S. (1999) Neural Networks: A Comprehensive Foundation, 2nd edition. Prentice Hall

[66] Hirota, K., Pedrycz, W. (1999) Fuzzy computing for data mining. Proceedings of the IEEE 87:1575–1600.

[67] Hu, Y.C., Chen, R.S. and Tzeng, G.H, (2003) Finding fuzzy classification rules using data mining techniques, Pattern Recognition Letters 24(1-3)509–519.

[68] Ide, E. (1971) New experiments in relevance feedback. The SMART system - experiments in automatic document processing, Prentice Hall.

[69] Judd, S. (1988) On the complexity of loading shallow neural networks. Journal of complexity 4.

[70] Kasabov, N., Woodford, B. (1999) Rule insertion and rule extraction from evolving fuzzy neural networks: Algorithms and applications for building adaptive, intelligent, expert systems. Proceedingss of FUZZ-IEEE99.

[71] Kasabov, N. (1996) Learning fuzzy rules and approximate reasoning in fuzzy neural networks and hybrid systems. Fuzzy Sets and Systems 82:135–149.

[72] Kerschen, G. and Golinval, J.C. (2002) Non-Linear Generalization of Principal Component Analysis: from a Global to a Local Approach, Journal of Sound and Vibration 254(5):867–876.

[73] Klir, G., Yuan, B. (1995) Fuzzy Sets and Fuzzy Logic, Theory and Applications. Prentice Hall.

[74] Koenen R. (2002) Overview of the MPEG-4 Standard. ISO/IEC JTC 1/SC 29/WG 11/N4668.

[75] Kohavi, R., Sommerfield, D. (1995) Feature Subset Selection Using the Wrapper Model: Overfitting and Dynamic Search Space Topology. Proceedings of KDD-95

[76] Koza, J. (1990) Genetic programming: a paradigm for genetically breeding populations of computer programs to solve problems. Report No. STAN-CS-90-1314, Department of Computer Science, Stanford University.

[77] Kraft, D. H. (1985) Advances in Information Retrieval: Where is that /# × %@^ Record?. Advances in computers 24:277–318.

[78] Kraft, D. H., Bordogna, G., Pasi, G. (1999) Fuzzy Set Techniques in Information Retrieval. Bezdek, J.C., Dubois, D., Prade, H. (editors.) Fuzzy Sets in Approximate Reasoning and Information Systems 3, The Handbook of Fuzzy Sets Series.

[79] Kraft, D. H., Bordogna, G., Pasi, G. (1998) Information Retrieval Systems: Where is the fuzz?. IEEE World Congress on Computational Intelligence; IEEE International Conference on Fuzzy Systems.

[80] Kraft, D. H. και Buel, D. A. (1983) Fuzzy sets and generalized Boolean retrieval systems. International Journal of Man-Machine Studies:45–56.

[81] Kraft, D. H., Petry, F. E. (1997) Fuzzy information systems: managing uncertainty in databases and information retrieval systems. Fuzzy Sets and Systems 90(2):183–191.

[82] Kundu S. (2000) A representation theorem for min-transitive fuzzy relations. Fuzzy Sets & Systems 109:453–457.

[83] Lim, T.-S., Loh, W.-Y., Shih, Y.-S. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning 40:203–229

[84] Laaksonen, J., Koskela, M., Laakso, S., Oja, E. (2000) PicSOM ? content-based image retrieval with self-organizing maps. Pattern Recognition Letters 21(13-14):1199–1207.

[85] Lee, J. H. (1998) Combining the evidence of different relevance feedback methods for information retrieval. Information Processing and Management 34(6):681–691.

[86] Lee, K., Street, W. N. (2001) Intelligent Image Analysis using adaptive resource allocating networks. Proceedings of IEEE International Workshop on Neural Networks for Signal Processing.

[87] Lee M.A., Takagi H. (1993) Integrating design stages of fuzzy systems using genetic algorithms. IEEE International Conference on Fuzzy Systems.

[88] Li, W. S. και Agrawal, D. (2000) Supporting web query expansion efficiently using multi-granularity indexing and query processing. Data and Knowledge Engineering 35(3):239–257.

[89] Maedche, A.,Motik, B., Silva, N., Volz, R. (2002) MAFRA - An Ontology MApping FRAmework in the Context of the SemanticWeb. Proceedings of the Workshop on Ontology Transformation at ECAI2002.

[90] Manjunath B. S., Salembier P., Sikora T. (2002) Introduction to MPEG-7, John Wiley and Sons.

[91] Maruyama, M. (1968) Mutual causality in general systems. Positive Feedback: a General Systems Approach to Positive/Negative Feedback and Mutual Casuality:80-100, Pergamon Press.

[92] Maruyama, M. (1963) The second cybernetics: deviation-amplifying mutual causal processes. American Scientist 51:164–179.

[93] De Meyer H., Naessens H., De Baets B. (2004) Algorithms for computing the min-transitive closure and associated partition tree of a symmetric fuzzy relation. European J. Oper. Res. 155:226–238.

[94] Mitra, S., De, R.K., Pal, S.K. (1997) Knowledge-based fuzzy MLP for classification and rule generation. IEEE Transactions on Neural Networks 8:1338–1350.

[95] Miyamoto, S. (1990) Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer Academic Publishers

[96] Mood, A. M., Graybill, F. A., Boes, D. C. (1998) Introduzione alla statica, McGraw-Hill.

[97] Mori, H., Chung, C. L., Kinoe, Y., Hayashi, Y. (1990) An adaptive document retrieval system using a neural network. International journal of Human-Computer Interaction 2(3):267–280.

[98] Moukas, A., Maes, P., (1998) Amalthaea: evolving multi-agent information filtering and discovery systems for the WWW. Autonomous agents and multi-agent systems 1:59–88.

[99] Nack, F., Lindsay, A. (1999) Everything You Wanted to Know About MPEG-7: Part 1. IEEE Multimedia 6(3):65–77.

[100] Nack, F., Lindsay, A. (1999) Everything You Wanted to Know About MPEG-7: Part 2. IEEE Multimedia 6(4):64-73.

[101] Naessens H., De Meyer H., De Baets B. (2002) Algorithms for the computation of T-transitive closures. IEEE Trans. Fuzzy Systems 10:541–551.

[102] Naphade, M.R., Huang, T.S. (2002) Extracting semantics from audio-visual content: the final frontier in multimedia retrieval. IEEE Transactions on Neural Networks 13(4):793–810.

[103] Nauk, D., Kruse, R. (1997) A neuro-fuzzy method to learn fuzzy classification rules from data. Fuzzy sets and Systems 8:277–288.

[104] Oakes, M. P., Reid, D., McEnery, T. (1991) Some practical applications of neural networks in information retrieval. British Computer Society 13th Information Retrieval Colloquium.

[105] Ovchinnikov S. (2002) Numerical representation of transitive fuzzy relations. Fuzzy Sets & Systems 126:225–232.

[106] Platt, J. (1991) A resource-allocating network for function interpolation. Neural Computing 3:213–225.

[107] Purdom, P. (1970) A transitive closure algorithm., BIT 10:76-94.

[108] Rish, I. (2001) An Empirical Study of the Naive Bayes Classifier. Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence, Seattle, Washington, USA.

[109] Rocchio, J. J. Jr. (1971) Relevance feedback in Information Retrieval. The SMART system - experiments in automatic document processing, Prentice Hall.

[110] Stamou G., Avrithis Y., Kollias S., Marques F., Salembier P. (2003) Semantic Unification of Heterogenous Multimedia Archives. Proc. of 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), London, UK, April 9–11.

[111] Salton, G. (1971) Relevance Feedback and the optimization of retrieval effectiveness. The SMART system - experiments in automatic document processing, Prentice Hall.

[112] Salton, G.,McGill, M. J. (1983) Introduction to modern information retrieval, McGraw-Hill.

[113] Santini, S. (2001)Exploratory Image Databases: Content-based Retrieval. Academic, New York.

[114] Schuyten, G., Dekeyser, H., Goeminne, K. (1999) Towards an Electronic Independent Learning Environment for Statistics in Higher Education. Education and Information Technologies 4:409–424.

[115] Seidel, R. (1992) On the All-Pairs-Shortest-Path problem in unweighted undirected graphs. J. Computer & System Sciences 51:400-403.

[116] Shapira Y. and Gath, I. (1999) Feature selection for multiple binary classification problems, Pattern Recognition Letters 20(8):823–832.

[117] Soltysiak S. J., Crabtree I. B. (1998) Automatic learning of user profiles – towards the personalisation of agent services, BT Technology Journal 16(3).

[118] Spink, A.,Saracevic, T. (1998) Human-computer interaction in information retrieval: nature and manifestations of feedback. Interacting with Computers 10(3):249–267.

[119] Swiniarski, R.W., Skowron, A. (2003) Rough set methods in feature selection and recognition. Pattern Recognition Letters 24:833–849

[120] Tan Y.-J. (2002) On compositions of lattice matrices. Fuzzy Sets & Systems 129:19–28.

[121] Tao, C.W. (2002)Unsupervised fuzzy clustering with multi-center clusters, Fuzzy Sets and Systems 128(3):305–322.

[122] Theodoridis, S, Koutroumbas, K. (1998) Pattern Recognition, Academic Press

[123] Thorelli, L.-E. (1966) An algorithm for computing all paths in a graph. BIT 6:347-349.

[124] Tsapatsoulis, N., Wallace, M. and Kasderidis, S. (2003) Improving the Performance of Resource Allocation Networks through Hierarchical Clustering of High – Dimensional Data. Proceedings of the International Conference on Artificial Neural Networks (ICANN), Istanbul, Turkey

[125] Turtle, H. R., Croft, W. B. (1992) A comparison of text retrieval models. The computer Journal 35(3):279–290.

[126] Ullman, J.D., Yannakakis, M., (1991) The Input/Output Complexity of Transitive Closure. Annals of Mathematics and Artificial Intelligence, 331-360.

[127] Vapnik, V. (1998) Statistical Learning Theory. John Willey and sons.

[128] van Rijsbergen, C. J. (1979) Information Retrieval. Butterworths and Co.

[129] Vertsetis, A.B. (1998) Teaching of History, Athens.

[130] Voorhees, E. M., Harman, D. (2000) Overview of the Sixth Text Retrieval Conference (TREC-6). Information Processing and Management 36(1):3–35.

[131] Wallace M., Akrivas G., Mylonas P., Avrithis Y., Kollias S. (2003) Using context and fuzzy relations to interpret multimedia content. CBMI, IRISA, Rennes, France.

[132] Wallace, M., Akrivas, G., Stamou, G. (2003) Automatic Thematic Categorization of Documents Using a Fuzzy Taxonomy and Fuzzy Hierarchical Clustering, Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), St. Louis, MO, USA.

[133] Wallace, M., Akrivas, G., Stamou, G., Kollias, S. (2002) Representation of user preferences and adaptation to context in multimedia content – based retrieval. Proceedings of the Workshop on Multimedia Semantics at SOFSEM2002, Milovy, Czech Republic.

[134] Wallace M., Avrithis Y., Stamou G., Kollias S. (2004) Knowledge-based Multimedia Content Indexing and Retrieval. Multimedia Content and Semantic Web: Methods, Standards and Tools, Stamou G., Kollias S. (Editors), Wiley.

[135] Wallace M., Ioannou S., Karpouzis K., Kollias S. (in press) Dealing with Feature Uncertainty in Facial Expression Recognition. International Journal of Intelligent Systems Technologies and Applications, in press.

[136] Wallace, M., Stamou, G. (2002) Towards a Context Aware Mining of User Interests for Consumption of Multimedia Documents. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Lausanne, Switzerland.

[137] Warren, H.S. (1975) A modification of Warshall's algorithm for the transitive closure of binary relations. Comm. ACM 18(4):218-220.

[138] Warshall, S. (1962) A theorem on Boolean matrices. J. ACM 9(1):11-12.

[139] Yager, R.R. (2000) Intelligent control of the hierarchical agglomerative clustering process. IEEE Transactions on Systems, Man and Cybernetics, Part B 30(6):835–845.

[140] Yang J., Honavar V. (1998) Feature Subset Selection Using A Genetic Algorithm. Intelligent Systems and Their Applications 13(2):44–49.

[141] Yong, R., Huang, T.S., Ortega, M. and Mehrotra, S. (1998) Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology 8:644–655.

[142] Yoshida Y. (2000) A limit theorem in dynamic fuzzy systems with transitive fuzzy relations. Fuzzy Sets & Systems 109:371–378.

[143] Zadeh, L.A. (1965) Fuzzy Sets. Information and Control 8:228-353.

[144] Zadeh L.A. (1971) Similarity relations and fuzzy orderings. Information Sciences 3:177–200.

[145] Zhao, R., W.I. Grosky (2002) Narrowing the Semantic Gap-Improved Text-Based Web Document Retrieval Using Visual Features. IEEE Transactions on Multimedia 4(2).

[146] UCI Machine Learning Repository. `http://www.ics.uci.edu/~mlearn/MLRepository.html`

[147] Principal components analysis. `http://obelia.jde.aca.mmu.ac.uk/multivar/pca.htm`

[148] ISO/IEC JTC1/SC29/WG11 (2000) Text of ISO/IEC CD 15938-2 Information technology – Multimedia content description interface – Part 2: Description definition language.

[149] ISO/IEC JTC 1/SC 29 M4242 (2001) Text of 15938-5 FDIS Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes.

[150] ISO/IEC JTC1/SC29/WG11, "MPEG-7 Context, Objectives and Technical Roadmap, (v.12)", Doc. N2861, July 1999.

[151] ISO/IEC JTC1/SC29/WG1 N1646R (2000) JPEG 2000 Part I. Final Committee Draft Version 1.0.

[152] XML Schema Part 0: Primer, W3C Working Draft, Sept 2000 `http://www.w3.org/TR/xmlschema-0`

[153] IST-1999-20502 Project FAETHON: Unified Intelligent Access to Heterogeneous Audiovisual Content, 2001-2003. `http://www.image.ece.ntua.gr/faethon/`

[154] [IST Project: Emotionally Rich Man-Machine Interaction Systems (ERMIS), 2001-2003. `http://www.image.ntua.gr/ermis/`

[155] List of WordNet publications `http://engr.smu.edu/~rada/wnb/`

[156] Java implementation of the sparse relation model and the ITU/ITC algorithms `http://image.ntua.gr/~wallace/java/transitive`

□