Πολυτεχνείο Κρήτης

Σχεδιασμός και Υλοποίηση Συστήματος Λογικών Ερωτήσεων και Ερωτήσεων Ομοιότητας σε Συστήματα Διαχείρισης Σχεσιακών Βάσεων Δεδομένων

Ανέστης Γιώργος

Διπλωματική Εργασία

που παρουσιάστηκε στο

Τμήμα Ηλεκτρονικής και Μηχανικών Υπολογιστών

Σεπτέμβριος 1997

Περίληψη

Η εργασία αυτή πραγματεύεται την εκτέλεση ερωτήσεων ομοιότητας (similarity queries) και ερωτήσεων δίτιμης λογικής (boolean queries) πάνω σε σχεσιακές βάσεις δεδομένων. Αποσκοπεί στην δημιουργία ενός συστήματος ανάκτησης πληροφορίας (information retrieval system) βασιζόμενο σε μια σχεσιακή βάση δεδομένων.

Το σύστημα παρέχει τη δυνατότητα να ορίζεται πάνω σε μια σχεσιακή βάση δεδομένων το σχήμα ενός συστήματος ανάκτησης πληροφορίας. Δίνει δηλαδή τη δυνατότητα να ορίζεται το σύνολο των αντικειμένων τα οποία θα ανακτώνται με τις ερωτήσεις ομοιότητας καθώς και το σύνολο των όρων οι οποίοι από τη μια συμβάλλουν στην περιγραφή των αντικειμένων και από την άλλη στο σχηματισμό των ερωτήσεων ομοιότητας.

Για την εκτέλεση των ερωτήσεων ομοιότητας παρουσιάζεται μια συγκεκριμένη μεθοδολογία με βάση την οποία η εκτέλεση μιας ερώτησης ομοιότητας ανάγεται τελικά σε εκτέλεση μιας σειράς ερωτήσεων εκφρασμένων με τη γλώσσα *SQL*. Η μεθοδολογία αυτή υποστηρίχθηκε με την υλοποίηση κατάλληλων εργαλείων.

Η απόκριση του συστήματος στην υποβολή των ερωτήσεων είναι το διατεταγμένο σύνολο των αντικειμένων τα οποία ικανοποιούν την ερώτηση. Η διάταξη των αντικειμένων γίνεται με βάση το βαθμό ομοιότητας του καθενός ως προς την υποβληθείσα ερώτηση. Ο βαθμός ομοιότητας μεταξύ μιας ερώτησης και ενός αντικειμένου μπορεί να υπολογιστεί με τη βοήθεια διαφόρων μοντέλων υπολογισμού ομοιότητας, και το σύστημα δίνει στο χρήστη τη δυνατότητα να επιλέξει το μοντέλο που επιθυμεί.

Η παρουσίαση των αντικειμένων που ικανοποιούν την ερώτηση γίνεται με φθίνουσα σειρά, ως προς το βαθμό ομοιότητάς τους με την ερώτηση, συμβάλλοντας έτσι στη μείωση του χρόνου που ξοδεύει ο χρήστης για να βρει απάντηση στην ερώτησή του, διότι τα αντικείμενα με το μεγαλύτερο βαθμό ομοιότητας παρουσιάζουν μεγαλύτερη πιθανότητα να περιέχουν την προς αναζήτηση πληροφορία.

Η καινοτομία του συστήματος έγκειται στο ότι ομοιόμορφα υποστηρίζει την τεχνολογία και τις δυνατότητες των συστημάτων ανάκτησης πληροφορίας με τις δυνατότητες και τα πλεονεκτήματα των συστημάτων διαχείρισης βάσεων δεδομένων που υποστηρίζουν μόνο ερωτήσεις δίτιμης λογικής. Η προσπάθεια επικεντρώνεται στην υλοποίηση ερωτήσεων ομοιότητας, βασισμένες σε πολλά διαφορετικά μοντέλα υπολογισμού ομοιότητας, μέσα από τις δυνατότητες που παρέχουν τα συστήματα διαχείρισης βάσεων δεδομένων, χωρίς να γίνεται προσφυγή σε ειδικές μεθόδους πρόσβασης ανάλογες με αυτές που χρησιμοποιούνται στα κλασικά συστήματα ανάκτησης πληροφορίας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Σταύρο Χριστοδουλάκη για την επίβλεψή του στην ολοκλήρωση αυτής της διπλωματικής εργασίας καθώς και για την ευκαιρία που μου πρόσφερε να εργαστώ στο εργαστήριο Διανεμημένων Πληροφοριακών Συστημάτων και Εφαρμογών, από όπου αποκόμισα πολύτιμες εμπειρίες.

Θα ήθελα να ευχαριστήσω προκαταβολικά τους καθηγητές κ. Βασίλη Διγαλάκη και κ. Παναγιώτη Τριανταφύλλου για το χρόνο που διέθεσαν για την ανάγνωση της παρούσας διπλωματικής εργασίας και τις τυχόν παρατηρήσεις τους.

Ιδιαίτερα θα ήθελα να ευχαριστήσω το μεταπτυχιακό φοιτητή Νεκτάριο Μουμουτζή για τη βοήθειά του στο σχεδιασμό της εργασίας αυτής και για τη συμπαράστασή του καθόλη τη διάρκειά της.

Ανέστης Γιώργος

Χανιά, Σεπτέμβριος 97

1. Εισαγωγή

Η σημερινή εποχή συχνά χαρακτηρίζεται ως εποχή της πληροφορίας διότι παράγεται περισσότερη πληροφορία από ποτέ άλλοτε. Η ανάπτυξη της τεχνολογίας έδωσε τη δυνατότητα μετάβασης από τα παραδοσιακά αποθηκευτικά μέσα, όπως το χαρτί για παράδειγμα, σε πιο εξελιγμένα, ψηφιακά μέσα αποθήκευσης, με πολλαπλάσια χωρητικότητα. Η εξέλιξη αυτή είχε ως αποτέλεσμα να δημιουργηθούν τεράστιες τράπεζες πληροφοριών.

Μετά την επίλυση του προβλήματος της αποθήκευσης με ψηφιακό τρόπο, παρουσιάστηκε ένα ακόμα πιο δύσκολο πρόβλημα. Η ανάκτηση της επιθυμητής πληροφορίας, σε σύντομο χρονικό διάστημα αν όχι σε πραγματικό χρόνο, μέσα από το σύνολο της αποθηκευμένης πληροφορίας ήταν εξαιρετικά δύσκολη αν όχι αδύνατη. Αυτό είχε ως αποτέλεσμα την επινόηση μηχανισμών για την αυτόματη εύρεση της επιθυμητής πληροφορίας. Οι μηχανισμοί αυτοί στηρίζονται στο γεγονός ότι οι δυνατότητες για γρήγορη ανάκτηση της πληροφορίας εξαρτώνται σε πολύ μεγάλο βαθμό από τις μεθόδους που χρησιμοποιήθηκαν για την αποθήκευσή της.

Η υιοθέτηση αποδοτικών μεθόδων αποθήκευσης της πληροφορίας επιταχύνει την ανάκτησή της, δεν λύνει όμως το πρόβλημα της εύρεσής της. Υπάρχουν δύο τομείς που ασχολούνται με την αποθήκευση και εύρεση πληροφορίας, ο τομέας των **Βάσεων Δεδομένων** και ο τομέας των **Συστημάτων Ανάκτησης Πληροφορίας**.

Τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων ακολουθούν ένα δομημένο τρόπο αποθήκευσης της πληροφορίας. Προσφέρουν ένα απλό μοντέλο

αναπαράστασης των δεδομένων, κατανοητό στους χρήστες. Παρέχουν επίσης έναν απλό και αποδοτικό τρόπο έκφρασης σύνθετων και πολύπλοκων ερωτήσεων, ο οποίος είναι πολύ κοντά στον τρόπο αναπαράστασης των δεδομένων. Όλα αυτά έχουν συμβάλλει στην ευρεία αποδοχή και στη μεγάλη εμπορική επιτυχία αυτών των συστημάτων.

Η θεμελιώδης οντότητα σ'αυτά τα συστήματα είναι η σχέση (relation) ή πίνακας (table). Ο χρήστης αντιλαμβάνεται τα δεδομένα σαν ένα σύνολο από πίνακες οι οποίοι οργανώνονται σε γραμμές και στήλες. Οι λειτουργίες που εκτελούνται πάνω στους πίνακες έχουν σαν αποτέλεσμα νέους πίνακες. Η αλληλεπίδραση των χρηστών με αυτά τα συστήματα γίνεται μέσω της Γλώσσας Δομημένων Ερωτήσεων (**S**tructured **Q**uery **L**anguage *SQL*). Η απάντηση σε μια ερώτηση είναι το σύνολο των οντοτήτων για τις οποίες η λογική έκφραση που περιγράφει την ερώτηση είναι αληθής. Οι ερωτήσεις που εκτελούνται κατ' αυτό τον τρόπο καλούνται ερωτήσεις δίτιμης λογικής (boolean queries).

Στα συστήματα ανάκτησης πληροφορίας και συγκεκριμένα στα συστήματα ανάκτησης κειμένου βασική οντότητα είναι το έγγραφο (document) το οποίο περιέχει μη δομημένη πληροφορία. Ένα τέτοιο σύστημα ανάκτησης πληροφορίας επεξεργάζεται, και γενικότερα εκτελεί το σύνολο των λειτουργιών του σε μια συλλογή (collection) από έγγραφα. Η απόκριση, στις αιτήσεις για ανάκτηση πληροφορίας που δέχεται, είναι ένα υποσύνολο του συνόλου των εγγράφων της συλλογής. Η απόφαση για την ανάκτηση ενός υποσυνόλου εγγράφων σε σχέση με μια ερώτηση βασίζεται στην "ομοιότητα" που υπάρχει μεταξύ των εγγράφων και της ερώτησης. Η ομοιότητα αυτή, υπολογίζεται συγκρίνοντας τις τιμές συγκεκριμένων χαρακτηριστικών, που αντιστοιχούν σε κάθε έγγραφο, με τις τιμές των χαρακτηριστικών που σχηματίζουν την ερώτηση. Σε κάθε έγγραφο που συμμετέχει στην απάντηση αντιστοιχεί διαφορετικός βαθμός ομοιότητας με την ερώτηση διότι τα έγγραφα της συλλογής δεν είναι πανομοιότυπα. Οι ερωτήσεις που εκτελούνται κατ' αυτό τον τρόπο καλούνται ερωτήσεις ομοιότητας (similarity queries).

Αν και αυτοί οι δύο τομείς αναπτύσσονται ανεξάρτητα ο ένας από τον άλλο είναι φανερό ότι με τον κατάλληλο συνδυασμό τους θα μπορούσαν να προκύψουν συστήματα με πολύ ισχυρές δυνατότητες. Ένας τέτοιος συνδυασμός θα εκμεταλλεύται το απλό μοντέλο αναπαράστασης των δεδομένων και την πρότυπη γλώσσα ερωτήσεων των συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων από τη μια και από την άλλη, τη δυνατότητα των συστημάτων ανάκτησης πληροφορίας

για εκτέλεση πολύπλοκων ερωτήσεων που είναι πιο κοντά στη σημασιολογία που αποδίδει ο χρήστης στα δεδομένα.

Ο σχεδιασμός και η υλοποίηση ενός συστήματος το οποίο θα συνδυάζει τις δύο αυτές τεχνολογίας αποτελεί το αντικείμενο αυτής της εργασίας. Το αποτέλεσμα είναι ένα σύστημα το οποίο παρέχει την δυνατότητα εκτέλεσης ερωτήσεων ομοιότητας, οι οποίες απαντώνται σε συστήματα ανάκτησης πληροφορίας, πάνω σε σχεσιακές βάσεις δεδομένων. Επιπλέον παρέχεται η δυνανότητα να συνδυάζονται ερωτήσεις ομοιότητας και ερωτήσεις δίτιμης λογικής μέσα στην ίδια αίτηση ανάκτησης πληροφορίας.

1.1 Αναγκαιότητα – Στόχοι

Εδώ και χρόνια οι τομείς των βάσεων δεδομένων και των συστημάτων ανάκτησης πληροφοριών αναπτύσσονται ανεξάρτητα ο ένας από τον άλλο, ενώ είναι προφανές ότι με τον κατάλληλο συνδυασμό τους θα μπορούσαν να προκύψουν συστήματα με πολύ ισχυρές δυνατότητες. Τα τελευταία χρόνια έχει παρατηρηθεί μια ιδιαίτερη κινητικότητα σε ερευνητικό επίπεδο σε μια προσπάθεια συνδυασμού των δύο αυτών τεχνολογιών. Οι άξονες στους οποίους κινείται αυτή η προσπάθεια είναι:

- Ο εμπλουτισμός υπαρχόντων συστημάτων διαχείρισης βάσεων δεδομένων με δυνατότητες αναζήτησης με βάση σχέσεις ομοιότητας (για παράδειγμα ο τελεστής LIKE της γλώσσας *SQL*).
- Ο συνδυασμός συστημάτων διαχείρισης βάσεων δεδομένων με συστήματα ανάκτησης πληροφορίας με τεχνικές από την τεχνολογία των κατανεμημένων συστημάτων και των ετερογενών συστημάτων.
- Ο εμπλουτισμός συστημάτων ανάκτησης πληροφορίας με δομημένα χαρακτηριστικά που περιγράφουν τα έγγραφα και μπορούν να χρησιμοποιηθούν σε ερωτήσεις δίτιμης λογικής (για παράδειγμα οι συγγραφείς εγγράφων, ο εκδοτικός οίκος, η χρονολογία έκδοσης).

Παρά το γεγονός ότι πολλές από τις ερευνητικές προσπάθειες προς τις κατευθύνσεις που αναφέρθηκαν έχουν επιτύχει και μάλιστα έχουν οδηγήσει και σε προϊόντα, εξακολουθεί να είναι ζητούμενο η ουσιαστική ολοκλήρωση των δύο αυτών τεχνολογιών. Σε μια τέτοια ολοκλήρωση αποσκοπεί η παρούσα εργασία. Στόχος

συνεπώς της εργασίας είναι να δημιουργήσει τους μηχανισμούς για την υποστήριξη πάνω σε μια σχεσιακή βάση δεδομένων με έναν διάφανο τρόπο των εξής:

- Τη δυνατότητα να εκφραστούν σύνθετες ερωτήσεις ομοιότητας (με χρήση λογικών τελεστών, *AND, OR, NOT*). Οι ερωτήσεις αυτές πρέπει να υποστηρίζουν και την ύπαρξη βαρών σπουδαιότητας στους όρους που τις σχηματίζουν (άτομα). Τα βάρη αυτά εκφράζουν τη σπουδαιότητα που έχει κάθε όρος της ερώτησης για το χρήστη που σχηματίζει την ερώτηση.
- Την ανάπτυξη μεθοδολογίας για την υποστήριξη ερωτήσεων ομοιότητας σε συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων η οποία να είναι αρκετά γενική ώστε να μπορεί να εφαρμοστεί σε οποιοδήποτε σχεσιακό σύστημα.
- Τη δυνατότητα να μπορούν να οριστούν νέα χαρακτηριστικά, τα οποία εκφράζουν επιπλέον σημασιολογική πληροφορία πάνω σε ένα υπάρχον σύστημα και υπολογίζονται με βάση τις τιμές κάποιων ήδη υπαρχόντων χαρακτηριστικών. Τα χρακτηριστικά αυτά χρησιμοποιούνται - κατάλληλα μετασχηματισμένα φυσικά - τόσο σε ερωτήσεις δίτιμης λογικής όσο και σε ερωτήσεις ομοιότητας.
- Τη δυνατότητα χρήσης διαφορετικών μοντέλων αξιολόγησης για την αποτίμηση των ερωτήσεων ομοιότητας, δεδομένου ότι, όπως έχει τεκμηριωθεί από σειρά πειραμάτων που περιγράφονται στη βιβλιογραφία, δεν υπάρχει ένα και μοναδικό μοντέλο κατάλληλο για όλα τα πεδία εφαρμογών ερωτήσεων ομοιότητας.
- Τη δυνατότητα να συνδυάζονται ερωτήσεις δίτιμης λογικής και ερωτήσεις ομοιότητας μέσα στην ίδια ερώτηση ανάκτησης, όπου η ερώτηση δίτιμης λογικής λειτουργεί σαν φίλτρο αποκλεισμού κάποιων αντικειμένων και η ερώτηση ομοιότητας ως συνάρτηση αξιολόγησης των αντικειμένων που "πέρασαν" από το φίλτρο αποκλεισμού.

Η τεχνολογική βάση στην οποία στηρίζεται το σύστημα είναι τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων, ενώ η ικανοποίηση των αιτήσεων ανάκτησης πληροφορίας ανάγεται στην εκτέλεση μιας σειράς ερωτήσεων εκφρασμένων με τη γλώσσα *SQL*. Η επιλογή των συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων ως τεχνολογική βάση για την ανάπτυξη του συστήματος έγινε προκειμένου να εκμεταλλευτούμε τις ισχυρές δυνατότητες που μας προσφέρουν, όπως έλεγχος ταυτοχρονισμού (concurrency control), κατανεμημένες ή πελάτη-

εξυπηρετητή αρχιτεκτονικές, ευρεία αποδοχή, άρα και μεγάλη αγορά κ.λ.π. Επιπλέον δίνεται η δυνατότητα να εμπλουτιστούν υπάρχουσες βάσεις δεδομένων με νέα λειτουργικότητα προσδίδοντας έτσι μεγαλύτερη αξία σε αυτές.

Προκειμένου όλα τα παραπάνω να γίνουν περισσότερο κατανοητά αλλά και για να παρουσιαστεί μια πιο πρακτική προσέγγιση, ας θεωρήσουμε το ακόλουθο παράδειγμα.

Έστω ότι μια σχεσιακή βάση δεδομένων περιέχει πληροφορίες για ξενοδοχεία. Οι πληροφορίες αυτές αφορούν για παράδειγμα το όνομα του ξενοδοχείου, το όνομα της ευρύτερης γεωγραφικής περιοχής στην οποία ανήκει ένα ξενοδοχείο, τις διευκολύνσεις (facilities) που αυτό παρέχει κ.λ.π.

Μια κλασσική ερώτηση δίτιμης λογικής θα ήταν η εξής:

"Ποια είναι τα ξενοδοχεία που έχουν Indoor Theatre ή Video και Air Condition ή Swimming pool και βρίσκονται στην Κρήτη."

Η απάντηση σ' αυτή την ερώτηση θα ήταν όλα τα ξενοδοχεία τα οποία βρίσκονται στη Κρήτη και παρέχουν τις ζητούμενες διευκολύνσεις. Είναι φανερό ότι αν ανάλογες ερωτήσεις εκτελεστούν πάνω σε μια μεγάλη βάση δεδομένων η ποσότητα της πληροφορίας που θα επιστρέφεται σαν απάντηση θα είναι πολύ μεγάλη. Στην προκειμένη περίπτωση ο αριθμός των ξενοδοχείων που θα επιστρέφονταν σαν απάντηση θα ήταν πολύ μεγάλος.

Ας υποθέσουμε τώρα ότι σε κάθε διευκόλυνση που παρέχεται από ένα ξενοδοχείο αποδίδεται (και έχει αποθηκευτεί στη βάση δεδομένων) και ένας βαθμός ο οποίος είναι μεταξύ του 0 και του 10. Ο βαθμός αυτός εκφράζει το πόσο καλά παρέχεται μια διευκόλυνση από ένα ξενοδοχείο.Το 10 θεωρείται το άριστα ενώ το 0 σημαίνει ότι η διευκόλυνση δεν παρέχεται από το ξενοδοχείο. Αν για παράδειγμα η διευκόλυνση *Swimming pool* που παρέχεται από το ξενοδοχείο *Καψής* βαθμολογείται με 9 σημαίνει ότι το ξενοδοχείο αυτό έχει μια πάρα πολύ καλή πισίνα

Με δεδομένη λοιπόν αυτή την κατάσταση η απάντηση στην προηγούμενη ερώτηση θα μπορούσε να είναι τα ξενοδοχεία που παρέχουν τις ζητούμενες διευκολύνσεις και βρίσκονται στη Κρήτη συνοδευόμενα όμως με ένα βαθμό ο οποίος θα εκφράζει το πόσα καλά παρέχονται οι ζητούμενες διευκολύνσεις σε κάθε ένα από τα ξενοδοχεία που παρέχουν αυτές τις διευκολύνσεις. Επίσης τα ξενοδοχεία που συγκέντρωσαν τον καλύτερο βαθμό, ως προς τη συγκεκριμένη ερώτηση, θα

παρουσιάζονται πρώτα διότι παρουσιάζουν μεγαλύτερη πιθανότητα να ικανοποιούν τις απαιτήσεις του χρήση που υπέβαλλε την ερώτηση. Συνεπώς γίνεται μια ταξινόμηση των ξενοδοχείων, με βάση το βαθμό του κάθε ξενοδοχείου ως προς την ερώτηση, κατά φθίνουσα σειρά, έτσι ώστε τα "καλύτερα" ξενοδοχεία να παρουσιάζονται πρώτα. Με τον τρόπο αυτό μειώνεται και ο χρόνος που δαπανά κάποιος χρήστης προκειμένου να βρει τα ξενοδοχεία της αρεσκείας του.

Τα προαναφερθέντα όμως δεν μπορούν πραγματοποιηθούν με την εκτέλεση μιας ερώτησης δίτιμης λογικής αλλά με την εκτέλεση μια ερώτησης ομοιότητας η οποία, θα υπολογίζει την "ομοιότητα" του κάθε ξενοδοχείου ως προς μια ερώτηση και θα βαθμολογεί τα ξενοδοχεία με βάση αυτή την ομοιότητα. Έτσι τα ξενοδοχεία με τη μεγαλύτερη ομοιότητα θα παίρνουν καλύτερο βαθμό.

Ένας χρήστης όμως μπορεί να ενδιαφέρεται περισσότερο για μια διευκόλυνση απ' ότι για μια άλλη. Στην προηγούμενη ερώτηση θα μπορούσε για παράδειγμα να ενδιαφέρεται περισσότερο για την διευκόλυνση *Video* απ' ότι για τη διευκόλυνση *Indoor Theatre.* Θα πρέπει λοιπόν να δίνεται στο χρήστη η δυνατότητα όταν σχηματίζει μια ερώτηση να εκφράζει τη σπουδαιότα που έχει γι' αυτόν κάθε ένας από τους όρους που σχηματίζουν την ερώτηση. Αυτή η σπουδαιότητα θα πρέπει να λαμβάνεται υπόψιν όταν εκτελείται η ερώτηση και υπολογίζεται ο βαθμός του κάθε αντικειμένου που ανακτάται (στη συγκεκριμένη περίπτωση του κάθε ξενοδοχείου που ανακτάται). Τα παραπάνω μπορούν να υποστηριχτούν μόνο με την εκτέλεση ερωτήσεων ομοιότητας.

Μια ακόμα πιο ενδιαφέρουσα ερώτηση θα ήταν η εξής:

"Ποια είναι τα ξενοδοχεία που έχουν Indoor Theatre ή Video και Air Condition ή Swimming pool και δεν είναι ακριβά."

Η απαίτηση του χρήστη να ανακτηθούν τα ξενοδοχεία τα οποία παρέχουν κάποιες συγκεκριμένες διευκολύνσεις και δεν είναι ακριβά καθιστά αναγκαίο τον ορισμό ενός "ακριβού ξενοδοχείου". Για παράδειγμα ένα ακριβό ξενοδοχείο θα μπορούσε να ήταν αυτό στο οποίο ένα μονόκλινο δωμάτιο κοστίζει πάνω από 30000 δραχμές. Αυτός ο ορισμός όμως δεν είναι απόλυτα σωστός διότι το αν ένα ξενοδοχείο είναι ακριβό δεν εξαρτάται μόνο από την τιμή ενός μονόκλινου δωματίου αλλά εξαρτάται και, για παράδειγμα, από την κατηγορία στην οποία ανήκει το ξενοδοχείο. Έτσι για παράδειγμα ένα ξενοδοχείου πρώτης κατηγορίας με τιμή δωματίου 30000

δραχμές δεν μπορεί να χαρακτηριστεί ακριβό ενώ ένα ξενοδοχείο τρίτης κατηγορίας με τιμή δωματίου 20000 δραχμές μπορεί να χαρακτηριστεί ως ακριβό. Θα μπορούσε συνεπώς να οριστεί το χαρακτηριστικό *"ακριβό"* για κάθε ξενοδοχείο το οποίο θα υπολογίζεται με βάση την τιμή του δωματίου και την κατηγορία ενός ξενοδοχείου.

Το παραπάνω όμως δεν μπορεί να γίνει στα παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων. Στα πλαίσια αυτής της εργασίας υλοποιήθηκαν οι μηχανισμοί για την υποστήριξη ανάλογων λειτουργιών.

1.2 Γενική Περιγραφή – Λειτουργικότητα

Αρχικά πρέπει να επισημανθεί ότι οι χρήστες του συστήματος χωρίζονται σε δύο κατηγορίες. Η πρώτη κατηγορία περιλαμβάνει το χρήστη-διαχειριστή του συστήματος, ο οποίος αναλύει τις απαιτήσεις που εμφανίζονται κάθε φορά, προβαίνει στον αναγκαίο σχεδιασμό και κατασκευάζει, κάνοντας χρήση του συστήματος, τις αναγκαίες δομές. Η δεύτερη κατηγορία περιλαμβάνει τον απλό χρήστη ο οποίος το μόνο που κάνει είναι να υποβάλλει, χρησιμοποιώντας το σύστημα, αιτήσεις ανάκτησης πληροφορίας.

Η εκτέλεση ερωτήσεων ομοιότητας πάνω σε σχεσιακές βάσεις δεδομένων αποτελεί το βασικό στόχο αυτής της εργασίας. Οι ερωτήσεις ομοιότητας όμως, απαντώνται μόνο σε συστήματα ανάκτησης πληροφορίας, ενώ αντίθετα, στα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων υπάρχει η δυνατότητα εκτέλεσης μόνο ερωτήσεων δίτιμης λογικής (boolean queries). Η προσέγγιση που ακολουθείται στα πλαίσια αυτής της εργασίας ***είναι η υλοποίηση ενός υψηλότερου επιπέδου μοντέλου, πάνω από το σχεσιακό μοντέλο***, το οποίο επιτρέπει αφενός μεν τον προσδιορισμό επιπλέον σημασιολογικής πληροφορίας πάνω από το σχήμα μιας σχεσιακής βάσης δεδομένων, αφετέρου δε αυτόματους μηχανισμούς μετάφρασης των ερωτήσεων ομοιότητας σε ερωτήσεις εκφρασμένες στη γλώσσα *SQL* που επενεργούν πάνω στους πίνακες της βάσης δεδομένων.

Στα συστήματα ανάκτησης πληροφοριών, υπάρχει καταρχάς ένα σύνολο αντικειμένων τα οποία και θα αποτελούν τη ***συλλογή (collection)*** του συστήματος ανάκτησης πληροφορίας, δηλαδή το σύνολο των αντικειμένων τα οποία θα ανακτώνται με τις ερωτήσεις ομοιότητας. Επιπλέον, υπάρχει ένα σύνολο όρων οι οποίοι από τη μια συμβάλλουν στην αναπαράσταση των αντικειμένων της συλλογής και από την άλλη στο σχηματισμό των ερωτήσεων ομοιότητας. Οι όροι αυτοί

καλούνται ***όροι δεικτοδότησης (index terms)***. Στην προσέγγιση που ακολουθείται στην εργασία αυτή, οι συλλογές αντικειμένων και τα σύνολα όρων δεικτοδότησης μοντελοποιούνται με την έννοια του ***συνόλου οντοτήτων (entity set)*** και η μεταξύ τους συσχέτιση με την έννοια του ***συνόλου συσχετίσεων (relationship set)***. Οι έννοιες αυτές είναι δανεισμένες από το δημοφιλές μοντέλο οντοτήτων-συσχετίσεων (entity relationship model) και ο λόγος που οδήγησε σε αυτή την προσέγγιση είναι ακριβώς η ευρεία αποδοχή του συγκεκριμένου μοντέλου, η ευκολία στην κατανόηση των εννοιών αυτών και το γεγονός ότι βρίσκεται σε υψηλότερο επίπεδο από το σχεσιακό μοντέλο.

Φυσικά, οι έννοιες αυτές έχουν επεκταθεί κατάλληλα για να καλύψουν τις ανάγκες έκφρασης και εκτέλεσης ερωτήσεων ομοιότητας. Έτσι, η έννοια του συνόλου οντοτήτων το οποίο έχει (στο παραδοσιακό μοντέλο οντοτήτων συσχετίσεων) ένα σύνολο από χαρακτηριστικά, εδώ έχει ένα σύνολο από ***ασαφή χαρακτηριστικά (fuzzy attributes)*** τα οποία ισοδυναμούν με ιδιότητες των οντοτήτων οι οποίες, για κάθε οντότητα έχουν μια τιμή σπουδαιότητας (ή σημασίας) από το 0 (μηδενική σπουδαιότητα) μέχρι το 1 (μέγιστη σπουδαιότητα). Για παράδειγμα, ένα ξενοδοχείο με βάση την τιμή του μονόκλινου δωματίου και την κατηγορία στην οποία ανήκει μπορεί να χαρακτηριστεί λιγότερο ή περισσότερο "ακριβό" (Expensive). Ορίζεται συνεπώς ένα ασαφές χαρακτηριστικό "*ακριβό ξενοδοχείο"* το οποίο υπολογίζεται με βάση την τιμή του μονόκλινου δωματίου και την κατηγορία στην οποία ανήκει και η τιμή του κυμαίνεται μεταξύ 0 και 1.

Τα προαναφερθέντα υποστηρίζονται μέσω ενός τμήματος του συστήματος το οποίο ονομάζεται **Εκδότης Σχήματος (Schema Editor)**. Ο **Ε**κδότης **Σ**χήματος (ΕΣ) μπορεί να θεωρηθεί ως το μέσο αλληλεπίδρασης με τη βάση δεδομένων προκειμένου να οριστούν σύνολα οντοτήτων και συσχετίσεων καθώς και ασαφή χαρακτηριστικά, προκειμένου δηλαδή, να οριστεί το σχήμα ενός συστήματος ανάκτησης πληροφορίας. Το αποτέλεσμα της αλληλεπίδρασης αυτής είναι το σχήμα του συστήματος ανάκτησης πληροφορίας που κατασκευάστηκε.

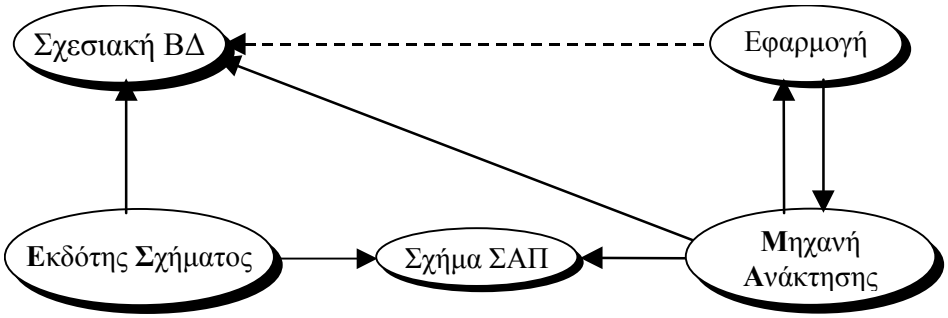
Προκειμένου να απαντηθούν ερωτήσεις οι οποίες είναι συνδυασμός ερωτήσεων ομοιότητας και ερωτήσεων δίτιμης λογικής ορίστηκε μια συγκεκριμένη μεθοδολογία με βάση την οποία η εκτέλεση μιας ερώτησης ομοιότητας ανάγεται στην εκτέλεση μιας σειράς *SQL* ερωτήσεων. Για την υποστήριξη αυτής της μεθοδολογίας κατασκευάστηκε η **Μηχανή Ανάκτησης (Retrieval Engine).**

Το επόμενο λοιπόν βασικό τμήμα του συστήματος είναι η **Μ**ηχανή **Α**νάκτησης (ΜΑ). Η ΜΑ κατασκευάστηκε με βάση το μοντέλο πελάτη-εξυπηρετητή και είναι το τμήμα του συστήματος, το οποίο αναλαμβάνει την εκτέλεση των υποβαλλόμενων ερωτήσεων και την ανάκτηση της αντίστοιχης πληροφορίας. Η ΜΑ κατά την εκκίνησή της λαμβάνει γνώση του σχήματος του συστήματος ανάκτησης πληροφορίας, το οποίο ορίστηκε με τον ΕΣ, και στη συνέχεια δέχεται σαν είσοδο αιτήσεις (requests) για ανάκτηση πληροφορίας. Η έξοδός της είναι το διατεταγμένο, κατά φθίνουσα σειρά, σύνολο των αντικειμένων που ικανοποιούν την εκάστοτε αίτηση. Η διάταξη των αντικειμένων γίνεται με βάση το βαθμό ομοιότητας τους ως προς την ερώτηση. Ο βαθμός του κάθε αντικειμένου συνοδεύει το αντικείμενο στην έξοδο της ΜΑ.

Ένα άλλο μέρος του συστήματος είναι η εφαρμογή η οποία παρέχει στον απλό χρήστη το επίπεδο επαφής-ζεύξης (interface) για το σχηματισμό των υποβαλλόμενων ερωτήσεων. Η εφαρμογή είναι εντελώς ανεξάρτητη από τα υπόλοιπα τμήματα του συστήματος. Αυτό το σημείο χρήζει ιδιαίτερης προσοχής διότι αναδεικνύει ένα από τα σημαντικότερα χαρακτηριστικά του συστήματος, πού είναι, η ανεξαρτησία της εφαρμογής από το υπόλοιπο σύστημα. Το μόνο το οποίο πρέπει να γνωρίζει η εφαρμογή είναι το πρωτόκολλο επικοινωνίας με την ΜΑ. Έτσι, κατασκευάζεται κάθε φορά ένα καινούργιο επίπεδο επαφής-ζεύξης, ανάλογα με τις ανάγκες της εκάστοτε εφαρμογής και εκτελούνται ερωτήσεις χωρίς να χρειαστεί να αλλάξει τίποτα στη ΜΑ. Η εφαρμογή κατά συνέπεια μπορεί να αναπτύσσεται από ανεξάρτητους κατασκευαστές λογισμικού οι οποίοι θα φροντίζουν να ακολουθούν το πρωτόκολλο επικοινωνίας με τη ΜΑ.

Στα πλαίσια αυτής της εργασίας, προς πιστοποίηση της σωστής λειτουργίας του συστήματος, κατασκευάστηκε η **Εφαρμογή Ψηφιακών Βιβλιοθηκών**.

Η γραφική απαράσταση του συστήματος που κατασκευάστηκε δίνεται στο επόμενο σχήμα:



Σχήμα 1-1: Γραφική αναπαράσταση του συστήματος που κατασκευάστηκε όπου διακρίνονται, η ΒΔ πάνω στην οποία χτίζεται το σύστημα ανάκτησης πληροφορίας,ο ΕΣ με τον οποίο ορίζονται σύνολα οντοτήτων, συσχετίσεων και τα χαρακτηριστικά, καθώς και το αρχείο που προκύπτει σαν έξοδός του. Επίσης φαίνεται η εφαρμογή με την οποία σχηματίζονται οι υποβαλλόμενες ερωτήσεις και η ΜΑ η οποία απάντα στις ερωτήσεις που δέχεται από την εφαρμογή

Η διαδικασία που ακολουθούμε προκειμένου να κάνουμε χρήση του συστήματος αποτελείται από τα εξής βήματα:

- Με τη χρήση του ΕΣ, ορίζουμε πάνω στη βάση δεδομένων το σχήμα του συστήματος ανάκτησης πληροφορίας, δηλαδή τα σύνολα οντοτήτων, συσχετίσεων και τα ασαφή χαρακτηριστικά.
- Θέτουμε σε λειτουργία τη ΜΑ δίνοντάς της σαν είσοδο το σχήμα του συστήματος ανάκτησης πληροφορίας που προέκυψε στο προηγούμενο βήμα από τον ΕΣ.
- Κατασκευάζουμε την εφαρμογή η οποία παρέχει το περιβάλλον για την κατασκευή και υποβολή των ερωτήσεων. Η εφαρμογή μπορεί να επικοινωνεί με τη ΜΑ και να στέλνει ερωτήσεις σ’ αυτήν. Επίσης αναλαμβάνει να παρουσιάσει τα αποτελέσματα της ερώτησης στο χρήστη.

Στο σημείο αυτό ολοκληρώνεται η σύντομη περιγραφή του συστήματος. Στα επόμενα κεφάλαια θα γίνει εκτενής αναφορά σε κάθε ένα από τα τμήματά του και θα δοθούν αναλυτικά παραδείγματα.

1.3 Δομή του κειμένου

Στο επόμενο κεφάλαιο γίνεται μια σύντομη αναφορά σε σχετική εργασία, δηλαδή στα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων, στα συστήματα ανάκτησης πληροφορίας, και σε κάποια από τα μοντέλα που έχουν προταθεί για τον υπολογισμό ομοιότητας. Στόχος δεν είναι να γίνει μια λεπτομερής περιγραφή των συστημάτων αυτών, αλλά να δοθεί ένα γενικό πλαίσιο έτσι ώστε να οριοθετηθεί η συγκεκριμένη εργασία και τα πλαίσια στα οποία κινείται.

Στο 3ο κεφάλαιο θα παρουσιαστεί πολύ σύντομα το μαθηματικό μοντέλο για τον υπολογισμό ερωτήσεων ομοιότητας.

Στο 4ο κεφάλαιο θα παρουσιαστεί η ***μεθοδολογία για την εκτέλεση ερωτήσεων ομοιότητας πάνω σε σχεσιακές βάσεις δεδομένων***. Η μεθοδολογία αυτή αποτελεί τη μεταφορά των αλγεβρικών μεθόδων που περιγράφονται το 3ο κεφάλαιο στο σχεσιακό σύστημα με τη χρήση της *SQL* και αποτελείται από τέσσερις φάσεις.

Στο 5ο κεφάλαιο γίνεται λεπτομερής περιγραφή του **Εκδότη Σχήματος**. Ο ΕΣ υποστηρίζει μια από τις φάσεις της μεθοδολογίας για την εκτέλεση ερωτήσεων ομοιότητας πάνω σε σχεσιακές βάσεις δεδομένων. Περιγράφονται με λεπτομέρεια οι λειτουργίες που αυτός υποστηρίζει αλλά γίνεται και εκτενής αναφορά στη σχεδίαση όπως και στην υλοποίησή του. Επίσης δίνονται και παραδείγματα χρήσης του ΕΣ προς διευκόλυνση της κατανόησης της χρήσης του.

Στο 6ο κεφάλαιο περιγράφεται η **Μηχανή Ανάκτησης**. Η ΜΑ υποστηρίζει κάποιες φάσεις της μεθοδολογίας εκτέλεσης ερωτήσεων ομοιότητας Γίνεται περιγραφή της αλληλεπίδρασης μεταξύ της ΜΑ και της εφαρμογής πελάτη η οποία στέλνει ερωτήσεις στη ΜΑ και παρουσιάζει τις απαντήσεις που προκύπτουν. Αναφέρονται επίσης κάποια ιδιαίτερα χαρακτηριστικά της ΜΑ.

Στο 7ο κεφάλαιο περιγράφεται η **Εφαρμογή Ψηφιακών Βιβλιοθηκών** που κατασκευάστηκε προς επίδειξη των δυνατοτήτων του συστήματος. Η εφαρμογή αυτή είναι ένα Java Applet, προκειμένου να δοθεί η δυνατότητα εκτέλεσης ερωτήσεων ομοιότητας μέσω κάποιων σελίδων υπερκειμένου (HTML pages) οι οποίες παρουσιάζονται στον Παγκόσμιο Ιστό (WWW). Η εφαρμογή αποτελεί ένα εύχρηστο γραφικό περιβάλλον το οποίο παρέχει τη δυνατότητα σχηματισμού ερωτήσεων ομοιότητας, την ανάθεση βαρών στους όρους της ερώτησης, την επιλογή του

μοντέλου υπολογισμού της ομοιότητας μεταξύ αντικειμένου - ερώτησης και την παρουσίαση των αποτελεσμάτων της ερώτησης.

Τέλος, στο 8ο κεφάλαιο γίνεται μια σύντομη ανακεφαλαίωση της παρούσας εργασίας και αναφέρονται μερικές ιδέες οι οποίες και θα μπορούσαν να αποτελέσουν μελλοντικές επεκτάσεις του συστήματος.

2. Σχετική εργασία

2.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιαστούν μερικά βασικά στοιχεία για τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων και για τα συστήματα ανάκτησης πληροφορίας. Επιχειρείται να δοθεί κατ' αυτό τον τρόπο το γενικό πλαίσιο μέσα στο οποίο κινείται η παρούσα εργασία.

Στο επόμενο τμήμα θα αναφερθούν κάποια ιδιαίτερα γνωρίσματα των συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων στα οποία στηρίχτηκε η παρούσα εργασία για την υλοποίηση ερωτήσεων ομοιότητας ενώ, στο τμήμα 2.3 θα γίνει μια σύντομη εισαγωγή στα συστήματα ανάκτησης πληροφορίας. Στο τμήμα 2.4 θα περιγραφούν δύο μοντέλα για την αναπαράσταση συστημάτων ανάκτησης πληροφορίας. Τα μοντέλα αυτά είναι *το μοντέλο διανυσματικού χώρου* και *το επεκταμένο πλαίσιο εργασίας δίτιμης ανάκτησης*. Αν και στα πλαίσια αυτής της εργασίας δεν ακολουθήθηκε πιστά κάποιο από αυτά τα μοντέλα, ωστόσο υιοθετήθηκαν κάποια στοιχεία και από τα δύο. Στο τμήμα 2.5 θα παρουσιαστούν τέσσερα μοντέλα υπολογισμού ομοιότητας, *το μοντέλο ασαφούς συνόλου, το μοντέλο Waller-Kraft, το επεκταμένο μοντέλο δίτιμης λογικής ή p-norm* και *το μοντέλο Infinite-One*. Τα μοντέλα αυτά υποστηρίζονται από το σύστημα που κατασκευάστηκε. Τέλος στο τμήμα 2.6 θα δοθεί μια σύντομη περίληψη του κεφαλαίου.

2.2 Τα Συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων

Τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων ακολουθούν ένα δομημένο τρόπο αποθήκευσης των δεδομένων. Η θεμελιώδης οντότητα σ'αυτά είναι η σχέση (relation) ή πίνακας (table). Ο χρήστης αντιλαμβάνεται τα δεδομένα σαν ένα σύνολο από πίνακες και όλες οι λειτουργίες εκτελούνται πάνω σε πίνακες για να έχουν σαν αποτέλεσμα νέους πίνακες. Αποδίδοντας με μαθηματικούς όρους την ιδιότητα αυτή θα λέγαμε ότι το σύνολο των σχέσεων σε μια σχεσιακή βάση δεδομένων είναι κλειστό ως προς όλες τις πράξεις που ορίζονται σ'αυτό.

Η δομική μονάδα των πινάκων είναι η πλειάδα (tuple) και συνεπώς ένας πίνακας μπορεί να θεωρηθεί σαν μια συλλογή από πλειάδες. Η δομική μονάδα των πλειάδων αποτελεί και τη μικρότερη δομική μονάδα αυτών των συστημάτων και ονομάζεται χαρακτηριστικό (attribute) ή πεδίο (field). Οι πίνακες, οι πλειάδες και τα χαρακτηριστικά αποτελούν διαφορετικές μορφές απεικόνισης των δεδομένων που βρίσκονται αποθηκευμένα σε ένα τέτοιο σύστημα, αλλά το βασικό χαρακτηριστικό τους είναι ότι όλα μπορούν να αντιμετωπισθούν, όπως και γίνεται, σαν πίνακες.

Τα συστήματα αυτά παρέχουν κάποιο συγκεκριμένο και καλά ορισμένο τρόπο περιγραφής και ορισμού των προς αποθήκευση δεδομένων όπως και μια σαφή διαδικασία για την ανάκτησή τους. Συγκεκριμένα η αλληλεπίδραση των χρηστών με αυτά τα συστήματα γίνεται μέσω της γλώσσας *SQL*. Η γλώσσα αυτή παρέχει την δυνατότητα ορισμού των δεδομένων (Data Definition Language) όπως και την δυνατότητα διαχείρισης και ανάκτησης των δεδομένων (Data Manipulation Language).

Μπορούμε να θεωρήσουμε ότι τα συστήματα αυτά περιέχουν ομοιογενείς δομικές μονάδες αναπαράστασης των δεδομένων, όπως οι πλειάδες, οι οποίες συνίστανται από ένα καθορισμένο και σταθερό αριθμό χαρακτηριστικών (attributes). Οι τιμές αυτών των χαρακτηριστικών περιγράφουν κάθε πλειάδα με τρόπο σαφή, κατηγορηματικό και **μοναδικό.** Εξαιτίας αυτών ακριβώς των ιδιοτήτων είναι εννοιολογικά εύκολο να καθορίσουμε το υποσύνολο των δεδομένων το οποίο ανταποκρίνεται σε μια συγκεκριμένη αίτηση αναζήτησης του χρήστη. Στις περιπτώσεις αυτές όμως η ίδια, η σχηματισθείσα από το χρήστη ερώτηση, πρέπει να περιέχει ένα υποσύνολο - έστω και ελάχιστο - του συνόλου των τιμών που μπορεί να

πάρει ένα χαρακτηριστικό. Κάτω από αυτές τις προϋποθέσεις, η ανάκτηση της πληροφορίας βασίζεται πάνω στο ακριβές ταίριασμα (exact match) μεταξύ των τιμών των χαρακτηριστικών που χρησιμοποιήθηκαν στον σχηματισμό της ερώτησης και των αντίστοιχων χαρακτηριστικών των αποθηκευμένων πλειάδων. Κάθε ένα από τα ανακτηθέντα αντικείμενα θα περιέχει την ακριβή τιμή του χαρακτηριστικού που καθορίστηκε στην ερώτηση -και πιθανόν τιμές άλλων χαρακτηριστικών που δεν καθορίστηκαν στην ερώτηση. Από την άλλη για κάθε ένα από τα αντικείμενα που δεν ανακτήθηκαν θα υπάρχει τουλάχιστον μια διαφορά (mismatch) μεταξύ της τιμής του χαρακτηριστικού στην ερώτηση και της τιμής του αντίστοιχου χαρακτηριστικού στο απορριφθέν αποθηκευμένο αντικείμενο.

Συμπερασματικά μπορούμε να πούμε ότι τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων παρέχουν ένα καλά ορισμένο τρόπο περιγραφής των δεδομένων και των προς υποβολή ερωτήσεων, αλλά προϋποθέτουν γνώση, από πλευράς του χρήστη, μέρους της αποθηκευμένης πληροφορίας ή ακριβής και αυστηρή περιγραφή των κριτηρίων που πρέπει να ικανοποιούν τα προς αναζήτηση αντικείμενα. Η απάντηση στην ερώτηση είναι το σύνολο των αντικειμένων για τα οποία η λογική έκφραση που περιγράφει την ερώτηση είναι αληθής. Η ανάπτυξη τέτοιων συστημάτων τα τελευταία 25 χρόνια και η παρουσία στην αγορά ισχυρών πακέτων διαχείρισης σχεσιακών βάσεων δεδομένων καθίστα την χρήση τους αρκετά ελκυστική.

2.3 Τα συστήματα ανάκτησης πληροφορίας

Οι μορφές αναπαράστασης της πληροφορίας μπορούν να διαχωριστούν σε τρεις κατηγορίες, τα γραπτά κείμενα, τον προφορικό λόγο και τις εικόνες, φωτογραφίες, σκίτσα ή σχήματα. Ο γραπτός λόγος θεωρείται ως ο βασικός τρόπος τυπικής επικοινωνίας μεταξύ των ανθρώπων και αποτελείται από σημειώσεις, μηνύματα, γράμματα, βιβλία, εφημερίδες, περιοδικά και πολλά άλλα, Ο προφορικός λόγος είναι λιγότερο τυπικός και αντίθετα με τον γραπτό μπορεί να χρησιμοποιηθεί και από ανθρώπους οι οποίοι δεν μπορούν να διαβάσουν ή να γράψουν. Τέλος, οι εικόνες και τα σχήματα μπορεί να συνοδεύουν τα γραπτά κείμενα, αλλά μπορούν επίσης και να χρησιμοποιούνται αυτόνομα όπως για παράδειγμα στις ταινίες ή στη ζωγραφική.

Τα υπάρχοντα συστήματα ανάκτησης πληροφορίας έχουν να κάνουν με την ανάκτηση πληροφορίας η οποία αναπαρίσταται με γραπτά κείμενα. Τα κείμενα μπορούν να θεωρηθούν σαν μονοδιάστατες σειρές από χαρακτήρες σε αντίθεση με το προφορικό λόγο και τις εικόνες που είναι δισδιάστατες. Πράγματι, η ομιλία μπορεί να αναλυθεί και να αναπαρασταθεί με δισδιάστατες κυματομορφές, οι οποίες μετασχηματίζονται σε ψηφιακή μορφή έτσι ώστε να αποθηκευτούν σε ψηφιακά μέσα αποθήκευσης και να καταστεί δυνατή η επεξεργασία τους με ηλεκτρονικούς υπολογιστές. Ομοίως, οι εικόνες είναι δισδιάστατες και θεωρούμε ότι συνίστανται από πολύ μικρά δομικά στοιχεία, τα εικονοστοιχεία (pixels), έτσι ώστε να καθίσταται δυνατή η ψηφιακή τους αποθήκευση και επεξεργασία.

Η μονοδιάστατη φύση των κειμένων αυξάνει εκθετικά τη δυνατότητα ψηφιακής τους αποθήκευσης και επεξεργασίας σε σχέση με την ομιλία και τις εικόνες. Ενδεικτικά αναφέρεται ότι ο χώρος που απαιτείται για να αποθηκεύσουμε μια τυπωμένη σελίδα υπολογίζεται περίπου σε 3.000 bytes[1] , ενώ αν η σελίδα αυτή διαβαζόταν και αποθηκευόταν ως ομιλία θα απαιτούνταν περίπου 300.000 bytes[2], δηλαδή, δύο τάξεις μεγέθους μεγαλύτερος χώρος σε σχέση με το γραπτό κείμενο. Η κατάσταση είναι ακόμα χειρότερη αν θεωρήσουμε την σελίδα σαν μια υψηλής ανάλυσης εικόνα (graphic data) και προσπαθήσουμε να την αποθηκεύσουμε με αυτή την μορφή. Θα χρειαζόμασταν περίπου 3.000.000 bytes[3], δηλαδή τρεις τάξεις μεγέθους μεγαλύτερο χώρο σε σχέση με το γραπτό κείμενο. Δεδομένης της απλούστερης αναπαράστασης του κειμένου, σε σχέση με την ομιλία και τις εικόνες, καθώς και της αποδοτικότερης αποθήκευσης του δεν προκαλεί έκπληξη το γεγονός ότι η επεξεργασία, με τη βοήθεια υπολογιστών, του κειμένου είναι πολύ ευκολότερη από την επεξεργασία του προφορικού λόγου και των γραφικών. Αυτοί είναι και οι λόγοι οι οποίοι συνέβαλαν στην ανάπτυξη συστημάτων ανάκτησης πληροφορίας στα οποία η πληροφορία θεωρείται ότι βρίσκεται σε μορφή κειμένου και για αυτά τα συστήματα θα μιλήσουμε στο επόμενο τμήμα.

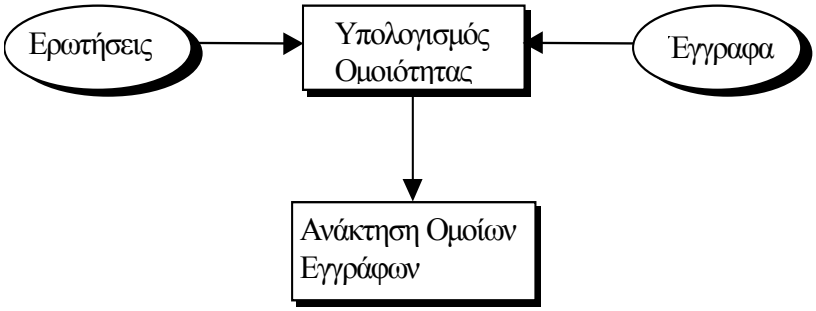
[1] Υποθέτουμε, 500 λέξεις / σελίδα, 6 χαρακτήρες ή bytes / λέξη.

[2] Υποθέτουμε, 4 λεπτά ομιλίας -240 δευτερόλεπτα- για την ανάγνωση μιας σελίδας, 9600 bps ως ρυθμό δειγματοληψίας της ομιλίας, 8 bits / byte.

[3] Υποθέτουμε, 200 bits / εκατοστό ,40.000 bits / τετραγωνικό εκατοστό, 600 τετραγωνικά εκατοστά / σελίδα.

2.3.1 Παραδοσιακά συστήματα ανάκτησης κειμένου

Βασική οντότητα αυτών των συστημάτων είναι το έγγραφο (document). Ένα τέτοιο σύστημα ανάκτησης πληροφορίας επεξεργάζεται, και γενικότερα εκτελεί το σύνολο των λειτουργιών του σε μια συλλογή (collection) από έγγραφα. Η απόκριση, στις αιτήσεις για ανάκτηση πληροφορίας που δέχεται, είναι ένα υποσύνολο του συνόλου των εγγράφων της συλλογής. Η απόφαση για την ανάκτηση ενός εγγράφου σε σχέση με μια ερώτηση βασίζεται στην "ομοιότητα" που υπάρχει μεταξύ του εγγράφου και της ερώτησης. Η ομοιότητα αυτή με τη σειρά της υπολογίζεται συγκρίνοντας τις τιμές συγκεκριμένων χαρακτηριστικών, που επισυνάπτονται σε κάθε έγγραφο, με τις τιμές των χαρακτηριστικών που σχηματίζουν την ερώτηση. Η γραφική αναπαράσταση ενός τέτοιου συστήματος φαίνεται στο επόμενο σχήμα.



Σχήμα 2-1: Γραφική απεικόνιση ενός τυπικού συστήματος ανάκτησης πληροφορίας. Η λήψη της απόφασης για την ανάκτηση ενός εγγράφου σε σχέση με την υποβαλόμενη ερώτηση βασίζεται στην "ομοιότητα" που υπάρχει μεταξύ του εγγράφου και της ερώτησης.

Τα χαρακτηριστικά τα οποία επισυνάπτονται σε κάθε έγγραφο της συλλογής είναι γνωστά ως λέξεις κλειδιά (keywords), όροι δεικτοδότησης (index terms) ή περιγραφικοί όροι (descriptor terms). Κάθε όρος δεικτοδότησης θεωρείται ότι περιγράφει το περιεχόμενο ενός εγγράφου σε κάποιο βαθμό αλλά όχι ολοκληρωτικά, ούτε και με απόλυτα σαφή και κατηγορηματικό τρόπο. Συχνά ένας μεγάλος αριθμός από διαφορετικούς όρους δεικτοδότησης επισυνάπτεται σε κάθε έγγραφο της συλλογής.

Υπό αυτές τις συνθήκες θα ήταν ίσως ανώφελο να επιζητήσουμε και να επιμείνουμε στο ακριβές ταίριασμα των όρων που σχηματίζουν μια ερώτηση και αυτών που επισυνάπτονται σε κάθε έγγραφο. Πράγματι, στα συστήματα αυτά η

ανάκτηση ενός εγγράφου βασίζεται στον επαρκή βαθμό σύμπτωσης του εγγράφου με τους όρους μιας ερώτησης. Ο βαθμός της σύμπτωσης αυτής υπολογίζεται με προσεγγιστικές μεθόδους.

Τα πλεονεκτήματα της δημιουργίας βαθμών ομοιότητας μεταξύ εγγράφων και ερωτήσεων στα περιβάλλοντα ανάκτησης πληροφορίας είναι:

- Τα ανακτηθέντα έγγραφα μπορούν να ταξινομηθούν κατά φθίνουσα σειρά, με βάση το βαθμό ομοιότητάς τους με την εκάστοτε ερώτηση, καθιστώντας έτσι δυνατή την παρουσίαση πρώτα των εγγράφων με τον καλύτερο βαθμό. Το γεγονός αυτό μπορεί να συμβάλει στην μείωση του χρόνου που δαπανά κάποιος χρήστης προκειμένου να αποκτήσει την πληροφορία που επιθυμεί.
- Το μέγεθος του συνόλου των ανακτηθέντων εγγράφων μπορεί να προσαρμοστεί στις απαιτήσεις και τις γνώσεις του κάθε χρήστη. Έτσι για απλούς χρήστες δίνεται η δυνατότητα να ανακτώνται τα πρώτα N "καλύτερα", με βάση τον βαθμό τους, έγγραφα, ενώ για τους πιο ειδικούς χρήστες μπορεί να ανακτάται ένα ποιο διευρυμένο σύνολο.
- Τα έγγραφα τα οποία ανακτώνται πρώτα, και συνεπώς παρουσιάζουν τη μεγαλύτερη ομοιότητα με την ερώτηση, μπορούν να χρησιμοποιηθούν για να "βελτιωθεί" η σχηματισθείσα ερώτηση με τη χρήση μηχανισμών ανάδρασης.

Στα συστήματα αυτά πρωτεύοντα ρόλο διαδραματίζει η διαδικασία εύρεσης των όρων δεικτοδότησης, οι οποίοι και περιγράφουν τα αποθηκευμένα έγγραφα αλλά και οι μέθοδοι που ακολουθούνται για την σύγκριση, και τον υπολογισμό της ομοιότητας, μεταξύ μιας ερώτησης και ενός εγγράφου. Σ' αυτό το σημείο αξίζει να επισημανθεί και η διαφορά που εμφανίζεται σε σχέση με τα συστήματα διαχείρισης βάσεων δεδομένων στα οποία δεν εμφανίζεται η ανάγκη ανάλυσης του περιεχομένου (content analysis) της προς αποθήκευση πληροφορίας αλλά υπεισέρχεται η ανάγκη κατασκευής αφαιρετικών μοντέλων για την αναπαράσταση των δεδομένων και η παροχή γλωσσών υψηλού επιπέδου για τον ορισμό και τη διαχείριση των δεδομένων.

2.3.2 Μέθοδοι πρόσβασης στα συστήματα ανάκτησης κειμένου

Όπως προαναφέραμε, στα συστήματα ανάκτησης πληροφορίας τα έγγραφα αναπαρίστανται από σύνολα όρων δεικτοδότησης τα οποία συχνά λέγονται και διανύσματα όρων (term vectors). Συνήθως οι όροι δεικτοδότησης δεν φέρουν βάρος σπουδαιότητας (importance weight) αν και σε ορισμένα συστήματα οι όροι είναι δυνατόν να φέρουν κάποιο βάρος. Στις περιπτώσεις αυτές το βάρος αντανακλά τη σχετική σπουδαιότητα του κάθε όρου για κάθε έγγραφο. Συνήθως οι ερωτήσεις εκφράζονται ως υποσύνολα όρων δεικτοδότησης οι οποίοι πρέπει να εμφανίζονται σε ένα έγγραφο για να ανακτηθεί. Είναι δυνατόν σε αρκετά συστήματα να εκφραστεί και η διαφορετική σημασία που αποδίδεται σε καθέναν από τους όρους δεικτοδότησης που συμμετέχει στον σχηματισμό μιας ερώτησης με την απόδοση διαφορετικού βάρους. Σε πιο ισχυρά συστήματα είναι εφικτός ο σχηματισμός ερωτήσεων με τη δημιουργία λογικών εκφράσεων των οποίων τα άτομα είναι όροι δεικτοδότησης και χρησιμοποιούνται οι λογικοί τελεστές *AND, OR, NOT*.

Εφ' όσον οι αιτήσεις για ανάκτηση πληροφορίας εκφράζονται ως σύνολα όρων δεικτοδότησης ή ως λογικοί συνδυασμοί αυτών των όρων είναι φανερό ότι και οι λειτουργίες αναζήτησης πρέπει να είναι σε θέση να ανακαλύπτουν εκείνα τα έγγραφα τα οποία περιέχουν ένα συγκεκριμένο συνδυασμό όρων. Εξαιτίας αυτών ακριβώς των απαιτήσεων και για λόγους καλύτερης απόδοσης έχουν υιοθετηθεί ειδικές μέθοδοι προσπέλασης (access methods) στην αποθηκευμένη πληροφορία. Ενδεικτικά αναφέρονται τα αρχεία υπογραφών (signatures files) και τα αντεστραμμένα αρχεία (inverted files). Θα παρουσιαστεί εν' συντομία η μέθοδος των αντεστραμμένων αρχείων.

Σύμφωνα με αυτή την μέθοδο, για κάθε όρο κατασκευάζεται ένα ξεχωριστός δείκτης, ο οποίος αποθηκεύει τις ταυτότητες των εγγράφων (document identifiers) τα οποία περιγράφονται από τον συγκεκριμένο όρο. Προκειμένου να απαντηθεί μια ερώτηση ανακτώνται οι ταυτότητες των εγγράφων οι οποίες περιέχονται στο αντεστραμμένο αρχείο κάθε όρου της ερώτησης και συνδυάζονται κατάλληλα. Ας θεωρήσουμε, για παράδειγμα, τον παρακάτω πίνακα εγγράφων-όρων, όπου τέσσερα έγγραφα καθορίζονται από τέσσερις πιθανούς όρους.

	Όρος 1	Όρος 2	Όρος 3	Όρος 4
Έγγραφο 1	1	1	0	1
Έγγραφο 2	0	1	1	1
Έγγραφο 3	1	0	1	1
Έγγραφο 4	0	0	1	1

Πίνακας 2-1: Πίνακας εγγράφων - όρων. Η ύπαρξη 1 στη τομή της i γραμμής με την j στήλη υποδηλώνει ότι ο όρος j επισυνάπτεται στο έγγραφο i. Αντίθετα η ύπαρξη 0 δηλώνει την απουσία του όρου από το έγγραφο.

Το αντεστραμμένο αρχείο που προκύπτει από τον παραπάνω πίνακα εγγράφων-όρων θα είχε ως εξής:

	Λίστα εγγράφων
Όρος 1	1,3
Όρος 2	1,2
Όρος 3	2,3,4
Όρος 4	1,2,3,4

Πίνακας 2-2: Παράδειγμα αντεστραμμένου αρχείου όπως προκύπτει από τον πίνακα 1-1. Κάθε γραμμή περιγράφει τη λίστα των εγγράφων τα οποία αντιστοιχούν σε ένα συγκεκριμένο όρο.

2.4 Μοντέλα αναπαράστασης συστημάτων ανάκτησης πληροφορίας

Έχουν προταθεί διάφορα μαθηματικά μοντέλα για την αναπαράσταση των συστημάτων ανάκτησης πληροφορίας και των λειτουργιών τους. Δύο από αυτά είναι το μοντέλο διανυσματικού χώρου (Vector Space Model) και το επεκταμένο πλαίσιο εργασίας δίτιμης ανάκτησης (extended boolean retrieval framework).

2.4.1 Το μοντέλο διανυσματικού χώρου

Το μοντέλο διανυσματικού χώρου θεωρεί ότι ένα διαθέσιμο σύνολο όρων χρησιμοποιείται τόσο για την περιγραφή των αποθηκευμένων εγγράφων όσο και για τον σχηματισμό των ερωτήσεων για ανάκτηση πληροφορίας [1,2]. Τα έγγραφα αλλά

και οι ερωτήσεις μπορούν να αναπαρασταθούν ως διανύσματα όρων της ακόλουθης μορφής:

$$D_i = \left(a_{i1}, a_{i2}, \ldots, a_{it}\right) \text{ και } Q_j = \left(q_{j1}, q_{j2}, \ldots, q_{jt}\right)$$

όπου, οι συντελεστές a_{ik} και q_{jk} αναπαριστούν τις τιμές του όρου k στο έγγραφο D_i ή στην ερώτηση Q_j αντίστοιχα. Τυπικά οι τιμές των όρων a_{ik} ή q_{jk} είναι 1 όταν ο όρος k εμφανίζεται στο έγγραφο D_i ή στην ερώτηση Q_j και 0 στην αντίθετη περίπτωση. Εναλλακτικά, οι συντελεστές του διανύσματος μπορούν να πάρουν και αριθμητικές τιμές των οποίων το μέγεθος αναπαριστά την σπουδαιότητα του όρου μέσα στο αντίστοιχο έγγραφο ή ερώτηση.

Ας υποτεθεί τώρα ότι t διακριτοί όροι είναι διαθέσιμοι για να χαρακτηρίσουν και να αναπαραστήσουν το περιεχόμενο κάθε εγγράφου. Κάθε ένας από τους t όρους μπορεί να προσδιοριστεί με ένα διάνυσμα όρων T, και ένας *διανυσματικός χώρος* ορίζεται αν τα T διανύσματα είναι γραμμικώς ανεξάρτητα. Αλλά είναι γνωστό ότι κάθε στοιχείο ενός διανυσματικού χώρου μπορεί να οριστεί σαν γραμμικός συνδυασμός των διανυσμάτων της βάσης του. Συνεπώς κάθε έγγραφο D_r της συλλογής μπορεί να γραφεί ως:

$$D_r = \sum_{i=1}^{t} a_{ri} T_i$$

όπου τα a_{ri} ερμηνεύονται σαν τα στοιχεία του εγγράφου D_r κατά μήκος του διανύσματος T_i. Ομοίως κάθε ερώτηση Q_s μπορεί να αναπαρασταθεί ως:

$$Q_s = \sum_{j=1}^{t} a_{sj} T_j$$

Σε ένα διανυσματικό χώρο, μπορεί να θεωρηθεί ως μέτρο ομοιότητας, μεταξύ δύο διανυσμάτων του, x και y το εσωτερικό γινόμενο αυτών, δηλαδή το γινόμενο:

$$x \cdot y = |x| \cdot |y| \cdot \cos(\alpha)$$

όπου $|x|$ είναι το μέτρο του διανύσματος x και α είναι η γωνία μεταξύ των δύο διανυσμάτων. Κατ' επέκταση, δεδομένου ενός εγγράφου D_r και μιας ερώτησης Q_s, ο βαθμός ομοιότητας εγγράφου - ερώτησης μπορεί να υπολογισθεί ως:

$$D_r \cdot Q_s = \sum_{i,j=1}^{t} a_{ri} q_{sj} T_i \cdot T_j$$

Ο υπολογισμός της ομοιότητας με βάση την προηγούμενη σχέση απαιτεί τον καθορισμό των στοιχείων τα οποία σχηματίζουν την ερώτηση, των όρων που αναπαριστούν το έγγραφο όπως επίσης και γνώση της συσχέτισης $T_i \cdot T_j$ για όλα τα ζευγάρια των όρων.

Ο καθορισμός των διανυσμάτων για τα έγγραφα γίνεται μέσω κάποιας διαδικασίας δεικτοδότησης, από την οποία προκύπτει ο παρακάτω πίνακας όρων-εγγράφων για μια συλλογή N εγγράφων:

$$A = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_N \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \cdots & T_t \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1t} \\ a_{21} & a_{22} & & a_{2t} \\ \vdots & & & \\ a_{N1} & a_{N2} & & a_{Nt} \end{bmatrix} \end{array}$$

Πίνακας Όρων-Εγγράφων για συλλογή N εγγράφων και t όρων δεικτοδότησης

Ο προσδιορισμός της συσχέτισης των όρων όμως δεν είναι εκ' των προτέρων γνωστός, ενώ και ο υπολογισμός της δεν μπορεί να προκύψει εύκολα.

Για την επίλυση αυτού του προβλήματος, στην πράξη, συχνά θεωρείται ότι οι όροι δεν παρουσιάζουν καμιά συσχέτιση μεταξύ τους. Εκφράζοντας μαθηματικά αυτή την υπόθεση λέμε ότι τα διανύσματα των όρων είναι ορθοκανονικά μεταξύ τους, δηλαδή, ότι $T_i \cdot T_j = 0, i \neq j$ και $T_i \cdot T_j = 1, i = j$. Όταν τα διανύσματα των t όρων είναι ορθογώνια, η γραμμική ανεξαρτησία έπεται ως άμεσο επακόλουθο και συνεπώς τα διανύσματα των t όρων αποτελούν βάση του διανυσματικού χώρου. Κάτω από αυτή την υπόθεση, η σχέση για τον υπολογισμό της ομοιότητας εγγράφου-ερώτησης περιορίζεται σε ένα απλό άθροισμα γινομένων και γίνεται:

$$Sim(D_r, Q_s) = \sum_{i,j=1}^{t} a_{ri} q_{sj}$$

Παρουσιάζεται παράδειγμα υπολογισμού ομοιότητας εγγράφου - ερώτησης με βάση το μοντέλο διανυσματικού χώρου [3].

$$D_1 = 2T_1 + 3T_2 + 5T_3$$
$$D_2 = 3T_1 + 7T_2 + 1T_3$$
$$Q = 0T_1 + 0T_2 + 2T_3$$

Δείγματα εγγράφων και ερώτησης

	T_1	T_2	T_3
T_1	1	0,5	0
T_2	0,5	1	$-0,2$
T_3	0	$-0,2$	1

Υποτιθέμενοι βαθμοί συσχέτισης των όρων

$$Sim(D_1, Q) = 2 \cdot 0 + 3 \cdot 0 + 5 \cdot 2 = 10$$
$$Sim(D_2, Q) = 3 \cdot 0 + 7 \cdot 0 + 1 \cdot 2 = 2$$

Υπολογισμοί ομοιότητας για ασυσχέτιστους όρους

$$Sim(D_1, Q) = (2T_1 + 3T_2 + 5T_3) \cdot (2T_3) = 4T_1 \cdot T_3 + 6T_2 \cdot T_3 + 10T_3 \cdot T_3 = -6 \cdot 0,2 + 10 \cdot 1 = 8,8$$

$$Sim(D_2, Q) = (3T_1 + 7T_2 + 1T_3) \cdot (2T_2) = 6T_1 \cdot T_3 + 14T_2 \cdot T_3 + 2T_3 \cdot T3 = -14 \cdot 0,2 + 2 \cdot 1 = -0,8$$

Υπολογισμός ομοιότητας για σχετιζόμενους όρους

Κλείνοντας την αναφορά στο μοντέλο διανυσματικού χώρου αναφέρονται μερικά από τα πλεονεκτήματα και τα μειονεκτήματα του συγκεκριμένου μοντέλου.

Μειονέκτημα μπορεί να θεωρηθεί η υπόθεση ορθογωνιότητας, και επομένως η ανεξαρτησία μεταξύ των όρων δεικτοδότησης, όπως και η έλλειψη θεωρητικού υποβάθρου για κάποιες από τις διαδικασίες χειρισμού των διανυσμάτων. Για παράδειγμα, η επιλογή ενός συγκεκριμένου μέτρου ομοιότητας μεταξύ διανυσμάτων για κάποια εφαρμογή δεν περιγράφεται ούτε καθορίζεται σε κάποιο βαθμό από το μοντέλο.

Στα πλεονεκτήματα του μοντέλου μπορεί να συμπεριληφθεί η απλότητά του, η δυνατότητα που παρέχει να θεωρηθούν όροι με κάποιο βάρος και όχι απλώς 0 και 1, καθώς και η βαθμολογημένη και ταξινομημένη έξοδος που παρέχει. Το τελευταίο αυτό πλεονέκτημα μπορεί να χρησιμοποιηθεί για τη δημιουργία μηχανισμών ανάδρασης για τους οποίους θα γίνει μια πολύ σύντομη αναφορά στο αμέσως επόμενο τμήμα.

2.4.1.1 Τροποποίηση Διανύσματος

Στα συστήματα ανάκτησης πληροφορίας είναι πολύ σημαντικό, αλλά και εξίσου δύσκολο, να δημιουργήθουν ερωτήσεις οι οποίες θα προκαλούν την ακριβή ανάκτηση της επιθυμητής πληροφορίας. Για να μπορούσαν να σχηματιστούν τέτοιες ερωτήσεις θα έπρεπε να υπήρχε γνώση σχετικά με τη σύνθεση της συλλογής των εγγράφων αλλά κάτι τέτοιο δεν ισχύει στην πράξη για το ευρύτερο κοινό. Μπροστά σ' αυτή την κατάσταση η συνήθης διαδικασία που ακολουθείται είναι η επαναλαμβανόμενη εκτέλεση ερωτήσεων. Οι πρώτες ερωτήσεις είναι συνήθως αρκετά γενικές και αποφέρουν ένα ευρύ σύνολο εγγράφων σαν απάντηση. Στη συνέχεια γίνεται προσπάθεια να βελτιωθεί η ερώτηση με βάση τις απαντήσεις που ελήφθησαν από τις προηγούμενες ερωτήσεις. Μια μέθοδος για την αυτόματη δημιουργία βελτιωμένων ερωτήσεων είναι γνωστή σαν **ανάδραση σχετικότητας (relevance-feedback)**.

Η βασική υπόθεση στην οποία βασίζεται αυτή η μέθοδος είναι ότι τα έγγραφα τα οποία είναι σχετικά με μια συγκεκριμένη ερώτηση θα πρέπει να "μοιάζουν" μεταξύ τους, με την έννοια ότι θα πρέπει να αναπαρίστανται από παρόμοια διανύσματα. Το γεγονός αυτό όμως υπονοεί ότι αν ένα ανακτηθέν έγγραφο έχει χαρακτηρισθεί ως σχετικό με την εκτελεσθείσα ερώτηση τότε η ερώτηση μπορεί να βελτιωθεί αν μετατραπεί, έτσι ώστε, να μοιάζει με το δεδομένο αυτό έγγραφο. Η τροποποιημένη ερώτηση αναμένεται να ανακτήσει επιπρόσθετα έγγραφα τα οποία θα είναι παρόμοια με το πρωτότυπο ανακτηθέν έγγραφο.

Η προσπάθεια συνεπώς εστιάζεται στο να βρεθεί ένας αποδοτικός τρόπος για να "μετακινηθεί" μια δοσμένη ερώτηση προς τα σχετικά έγγραφα και να "απομακρυνθεί" από τα μη σχετικά. Μια ιδανική ερώτηση Q_{opt} λοιπόν, θα μεγιστοποιεί τη μέση ομοιότητα εγγράφου - ερώτησης για τα σχετικά έγγραφα και ταυτόχρονα θα ελαχιστοποιεί τη μέση ομοιότητα εγγράφου - ερώτησης για τα μη σχετικά έγγραφα. Κάτω από αυτές τις προϋποθέσεις μια τέτοια ερώτηση θα έχει την εξής μορφή:

$$Q_{opt} = k \cdot \left(\frac{1}{R} \cdot \sum_{\mathrm{Re}\,l} \frac{D_i}{|D_i|} - \frac{1}{N-R} \cdot \sum_{Nonrel} \frac{D_i}{|D_i|} \right)$$

όπου, k είναι σταθερά, R και $N-R$ είναι ο υποτιθέμενος αριθμός των σχετικών και μη σχετικών εγγράφων, ενώ τα δύο αθροίσματα εκτείνονται στο σύνολο των κανονικοποιημένων σχετικών και μη σχετικών εγγράφων αντίστοιχα [4].

Είναι προφανές ότι η προηγούμενη έκφραση δεν μπορεί να χρησιμοποιηθεί άμεσα για την κατασκευή ερωτήσεων διότι το σύνολο των σχετικών και μη σχετικών εγγράφων σε σχέση με τις ερωτήσεις δεν είναι γνωστά εκ' των προτέρων. Η εύρεσή τους θα απαιτούσε εξαντλητικό ψάξιμο (exhaustive search) της συλλογής. Ωστόσο, η έκφραση αυτή μπορεί να προσεγγιστεί με την εξής διαδικασία. Μετά την εκτέλεση μιας αρχικής ερώτησης μπορεί να ζητηθεί από το χρήστη να καθορίσει ποια από τα έγγραφα που ανακτήθηκαν, από αυτή την αρχική ερώτηση, είναι σχετικά με αυτά που πραγματικά επιθυμεί. Με αυτή τη διαδικασία προκύπτει ένα υποσύνολο R' του R των σχετικών εγγράφων όπως και ένα υποσύνολο N' του $N-R$ των μη σχετικών εγγράφων. Μπορεί έτσι να προκύψει μια προσέγγιση της βέλτιστης ερώτησης $Q^{(i+1)}$ προσθέτοντας σε μια δεδομένη ερώτηση $Q^{(i)}$ τα διανύσματα όρων των εγγράφων τα οποία χαρακτηρίστηκαν ως σχετικά και ταυτόχρονα αφαιρώντας τα διανύσματα των εγγράφων που χαρακτηρίστηκαν ως μη σχετικά. Έτσι προκύπτει:

$$Q^{(i+1)} = Q^{(i)} + \frac{1}{|R'|} \cdot \sum_{D_i \in R'} D_i - \frac{1}{|N'|} \cdot \sum_{D_i \in N'} D_i$$

ή

$$Q^{(i+1)} = Q^{(i)} + \alpha \cdot \sum_{D_i \in R'} D_i - \beta \cdot \sum_{D_i \in N'} D_i$$

για κατάλληλους συντελεστές α και β.

Όπως είναι φανερό από τις προηγούμενες σχέσεις μπορούν να εκτελεστούν διαδοχικές επαναλήψεις, προκειμένου να προσεγγιστεί η βέλτιστη ερώτηση, καθώς όλο και περισσότερα έγγραφα ανακτώνται.

Στο σημείο αυτό ολοκληρώνεται η παρουσίαση του μοντέλου διανυσματικού χώρου. Στο επόμενο τμήμα θα παρουσιαστεί το εκτεταμένο πλαίσιο εργασίας δίτιμης ανάκτησης (Extended Boolean Retrieval Framework) το οποίο είναι και πλησιέστερο προς αυτή την εργασία.

2.4.2 Επεκταμένο πλαίσιο εργασίας δίτιμης ανάκτησης (extended boolean retrieval framework)

Ας θεωρηθεί ένα τυπικό σύστημα ανάκτησης πληροφορίας όπου για την ανάκτησή της σχηματίζονται ερωτήσεις δίτιμης λογικής (boolean queries). Η διαδικασία απάντησης τέτοιου είδους ερωτήσεων συνίσταται στην σύγκριση των όρων δεικτοδότησης οι οποίοι επισυνάπτονται σε κάθε έγγραφο με τους όρους που σχηματίζουν την ερώτηση. Όπως προαναφέρθηκε, στο περιβάλλον των συστημάτων ανάκτησης πληροφορίας έχουν προταθεί διάφοροι τρόποι για τον υπολογισμό της ομοιότητας - ή της απόστασης- μεταξύ εγγράφου-ερώτησης. Αυτοί οι τρόποι ωστόσο δεν μπορούν να χρησιμοποιηθούν για ερωτήσεις οι οποίες σχηματίζονται βασιζόμενες στις αρχές της δίτιμης λογικής διότι γι' αυτές τις ερωτήσεις μπορούν να χρησιμοποιηθούν μόνο δύο τιμές για να εκφράσουν την ομοιότητα εγγράφου-ερώτησης: η τιμή 1 όταν οι όροι δεικτοδότησης ταιριάζουν ακριβώς με τους όρους της ερώτησης και η τιμή 0 στην αντίθετη περίπτωση. Είναι λοιπόν φανερό ότι κάτω από αυτές τις συνθήκες δεν μπορεί να δημιουργηθεί βαθμολογημένη έξοδος και συνεπώς ταξινομημένη παρουσίαση των εγγράφων.

Εξαιτίας των πρακτικών δυνατοτήτων και πλεονεκτημάτων που παρουσιάζει η τεχνολογία των ερωτήσεων δίτιμης λογικής επιχειρείται μια επέκταση αυτής της τεχνολογίας έτσι ώστε να βρει εφαρμογή και στα συστήματα ανάκτησης πληροφορίας. Το αποτέλεσμα αυτής της προσπάθειας ήταν το εκτεταμένο πλαίσιο εργασίας δίτιμης ανάκτησης (extended boolean retrieval framework) [5]. Η μαθηματική περιγραφή αυτού του πλαισίου έχει ως εξής.

Ένα σύστημα ανάκτησης πληροφορίας βασιζόμενο στο εκτεταμένο πλαίσιο δίτιμης ανάκτησης ορίζεται ως η ακόλουθη τετράδα:

$$\langle T, Q, D, F \rangle$$

Όπου,

- T είναι το σύνολο των όρων δεικτοδότησης (index terms) οι οποίοι χρησιμοποιούνται για να αναπαραστήσουν τα αποθηκευμένα έγγραφα άλλα και για τον σχηματισμό των υποβαλλόμενων ερωτήσεων.
- Q είναι το σύνολο των ερωτήσεων οι οποίες μπορούν να αναγνωριστούν και να γίνουν αποδεκτές από το σύστημα. Κάθε ερώτηση $q \in Q$ είναι μια λογική

έκφραση απαρτιζόμενη από όρους δεικτοδότησης και από τους λογικούς τελεστές *AND*, *OR*, και *NOT*.

- D είναι το σύνολο των εγγράφων της συλλογής. Κάθε έγγραφο $d \in D$ αναπαρίσταται ως ένα σύνολο $\{(t_1, w_{d1}),(t_2, w_{d2}),\ldots,(t_n, w_{dn})\}$ όπου w_{di} είναι το βάρος του όρου t_i στο έγγραφο d και μπορεί να παίρνει τιμές μεταξύ μηδέν και ένα, δηλαδή $0 \le w_{di} \le 1$.
- F είναι μια συνάρτηση ανάκτησης $F: D \times Q \to [0,1]$ η οποία αντιστοιχίζει σε κάθε ζεύγος (d,q) έναν αριθμό ο οποίος ανήκει στο κλειστό διάστημα $[0,1]$. Αυτός ο αριθμός είναι το μέτρο ομοιότητας μεταξύ του εγγράφου d και της ερώτησης q και καλείται τιμή του εγγράφου d σε σχέση με την ερώτηση q. Η συνάρτηση ανάκτησης (retrieval function) $F(d,q)$ ορίζεται ως ακολούθως:

    - Για κάθε όρο (term) t_i, ο οποίος συμμετέχει στην ερώτηση, η συνάρτηση $F(d,t_i)$ ορίζεται ως το βάρος του όρου t_i πάνω στο έγγραφο d, δηλαδή w_{di}.
    - Οι λογικοί τελεστές - *AND, OR, NOT* - υπολογίζονται τότε χρησιμοποιώντας την αντίστοιχη νόρμα. Για παράδειγμα μπορούν να χρησιμοποιηθούν οι τελεστές που προτείνει το μοντέλο του ασαφούς συνόλου (fuzzy set) ή αυτοί που προτείνει το εκτεταμένο μοντέλο δίτιμης λογικής όπως θα δειχθεί παρακάτω. Για λογικές ερωτήσεις οι οποίες περιέχουν περισσότερους από ένα λογικούς τελεστές, η διαδικασία υπολογισμού τους αναπτύσεται αναδρομικά από την εσωτερικότερη πρόταση (clause).

Είναι αξιοσημείωτο το γεγονός ότι το παραπάνω πλαίσιο συμπεριλαμβάνει και το μοντέλο δίτιμης λογικής. Πράγματι, για μια ερώτηση q της μορφής $((t_1 \quad OR \quad t_2)\, AND \quad NOT \quad t_3)$ και για ένα έγγραφο d που περιέχει το όρο t_1 αλλά όχι τον όρο t_2 ή t_3 η συνάρτηση $F(d,q)$ ορίζεται ως: $F(d,t_1) = 1$ και $F(d,t_2) = F(d,t_3) = 0$. Η πρώτη πρόταση $F(d, t_1 \quad OR \quad t_2)$ υπολογίζεται τότε στο

1 και ολόκληρη η έκφραση υπολογίζεται ως: $1 \cdot (1-0) = 1$, δείχνοντας ότι η τιμή του εγγράφου d ως προς την ερώτηση q είναι 1.

Στο επόμενο τμήμα θα αναφερθούν κάποια από τα μοντέλα τα οποία έχουν προταθεί για τον υπολογισμό της συνάρτησης ανάκτησης $F(d,q)$ και υποστηρίζονται από το σύστημα που κατασκευάστηκε.

2.5 Μοντέλα υπογισμού ομοιότητας στα συστήματα ανάκτησης πληροφορίας

Το μοντέλο που χρησιμοποιείται για τον υπολογισμό της ομοιότητας μεταξύ αντικειμένων και όρων δεικτοδότησης διαδραματίζει σημαντικό ρόλο στα συστήματα ανάκτησης πληροφορίας αφού καθορίζει σε σημαντικό βαθμό την ορθότητα η μη των αποτελεσμάτων που προκύπτουν. Έχει όμως τεκμηριωθεί, από σειρά πειραμάτων που περιγράφονται στη βιβλιογραφία, ότι δεν υπάρχει ένα και μοναδικό μοντέλο κατάλληλο για όλα τα πεδία εφαρμογών ερωτήσεων ομοιότητας. Στο επόμενα τμήματα περιγράφονται κάποια από τα μοντέλα που έχουν προταθεί για τον υπολογισμό ομοιότητας και τα οποία υλοποιήθηκαν στα πλαίσια αυτής της εργασίας.

2.5.1 Το μοντέλο του ασαφούς συνόλου (fuzzy set model)

Η ανάπτυξη της θεωρίας των ασαφών συνόλων επέτρεψε την εφαρμογή των εννοιών της θεωρίας συνόλων σε σύνολα αντικειμένων των οποίων η τιμή της σχέσης ενός αντικειμένου με το σύνολο (membership value) δεν είναι 0 ή 1 αλλά μπορεί να πάρει τιμές στο διάστημα $[0,1]$. Το παραπάνω επέτρεψε τον ορισμό νέων τελεστών για τους αντίστοιχους τελεστές της κλασικής θεωρίας συνόλων. Συνήθως υπάρχουν περισσότεροι από ένας ασαφείς τελεστές που αντιστοιχούν σε ένα δεδομένο κλασικό τελεστή, και οι διαφορετικοί αυτοί τελεστές έχουν διαφορετικά λειτουργικά χαρακτηριστικά.

Μια ποικιλία από ασαφείς τελεστές έχουν προταθεί για τους λογικούς τελεστές *AND* και *OR* στην βιβλιογραφία[4]. Δύο από αυτούς είναι οι *MIN* και *MAX*, όπου ο *MIN* αντιστοιχίζεται στο λογικό *AND* και ο *MAX* στο λογικό *OR*. Δηλαδή η συνάρτηση ανάκτησης θα υπολογίζεται ως εξής:

$$F(d, t_1 \quad AND \quad t_2) = MIN(w_{d1}, w_{d2})$$

$$F(d, t_1 \quad OR \quad t_2) = MAX(w_{d1}, w_{d2})$$

$$F(d, NOT \quad t_1) = 1 - w_{d1}$$

όπου w_{d1} είναι το βάρος του όρου t_1 στο έγγραφο d , w_{d2} είναι το βάρος του όρου t_2 στο έγγραφο d και παίρνουν τιμές μεταξύ 0 και 1.

Με τη χρήση αυτών των τελεστών στον υπολογισμό της συνάρτησης $F(d,q)$ μπορούμε να έχουμε βάρη (weights) που κυμαίνονται μεταξύ 0 και 1 στους όρους που περιγράφουν ένα έγγραφο. Οι τιμές των βαρών αυτών απεικονίζουν την σπουδαιότητα του κάθε όρου στο κάθε έγγραφο. Στα πλαίσια αυτά αναπτύσσεται το μοντέλο ασαφούς συνόλου [6,7,8]. Είναι προφανές, ότι αν οι τιμές των βαρών περιοριστούν στις τιμές 0 και 1 τότε προκύπτει το μοντέλο δίτιμης λογικής.

Ας θεωρήσουμε το παράδειγμα της προηγούμενης ερώτησης $((t_1 \quad OR \quad t_2)\, AND \quad NOT \quad t_3)$ και ας υποθέσουμε ότι $F(d,t_1) = 0{,}7$, $F(d,t_2) = 0{,}2$ και $F(d,t_3) = 0{,}1$. Τότε η έκφραση $F(d, t_1 \quad OR \quad t_2)$ υπολογίζεται ως $\max(F(d,t_1), F(d,t_2)) = 0{,}7$ και ολόκληρη η τιμή του εγγράφου ως προς την συγκεκριμένη ερώτηση υπολογίζεται ως $\min\{0.7, (1-0.1)\} = 0.7$.

Το μοντέλο ασαφούς συνόλου είναι λιγότερο "αυστηρό" και περιοριστικό από το καθαρό μοντέλο δίτιμης λογικής διότι παρέχει βαθμολογημένα έγγραφα στην έξοδό του και επιτρέπει προφανώς την ταξινόμησή τους, ωστόσο, δεν είναι αρκετά ευέλικτο και παρουσιάζει αρκετά σημαντικά προβλήματα τα οποία και θα αναφερθούν αμέσως.

Η χρήση του μοντέλου ασαφούς συνόλου για τον υπολογισμό της συνάρτησης ανάκτησης F προκαλεί σε αρκετές περιπτώσεις λανθασμένη βαθμολόγηση των εγγράφων, βαθμολόγηση η οποία δεν συμφωνεί με την ανθρώπινη διαίσθηση. Αυτό

[4] Οι τελεστές αυτοί χωρίζονται σε δύο κατηγορίες, στους Τ-τελεστές (T-operators) και στους μέσους τελεστές (averaging operators) [9,18].

συμβαίνει διότι το αποτέλεσμα που προκύπτει από την χρήση των τελεστών *MIN* και *MAX* βασίζεται πάνω στον ένα μόνο, από τους δύο τελεστέους, χωρίς να λαμβάνεται καθόλου υπ'όψιν ο άλλος. Το πρόβλημα αυτό φαίνεται ξεκάθαρα στο ακόλουθο παράδειγμα.

Έστω δύο έγγραφα d_1 και d_2 και η ερώτηση q όπως φαίνεται παρακάτω:

$$d_1 = \{(Information, 0.40), (\mathrm{Re}trieval, 0.40)\}$$

$$d_2 = \{(Information, 0.99), (\mathrm{Re}trieval, 0.39)\}$$

$$q = Information \quad AND \quad \mathrm{Re}trieval$$

Όταν ο τελεστής *MIN* χρησιμοποιηθεί για τον υπολογισμό του *AND* η τιμή του εγγράφου d_1 για την ερώτηση q θα είναι 0.40 και του εγγράφου d_2 θα είναι 0.39.Δηλαδή το d_1 ανακτάται έχοντας υψηλότερο βαθμό από το d_2. Οι περισσότεροι άνθρωποι όμως θα έλεγαν ότι το έγγραφο d_2 παρά το έγγραφο d_1 είναι περισσότερο όμοιο με την ερώτηση q.

Αν και έχουν προταθεί αρκετοί ασαφείς τελεστές για τον προσδιορισμό των τελεστών δίτιμης λογικής δεν θα γίνει αναφορά σ'αυτούς μιας και δεν υποστηρίζονται στο σύστημα αλλά και διότι όλοι αυτοί οι τελεστές οδηγούν, σε κάποιες περιπτώσεις, σε λανθασμένη βαθμολόγηση.

2.5.2 Το μοντέλο Waller-Kraft

Το μοντέλο Waller-Kraft [10] χρησιμοποιεί επίσης τους τελεστές *MIN* και *MAX* αλλά προκειμένου να αποφύγει τα προβλήματα που παρουσιάζονται στο μοντέλο του ασαφούς συνόλου υιοθετεί μια πολύπλοκη έκφραση για τον υπολογισμό των *AND* και *OR* τελεστών. Έτσι το μοντέλο αυτό προτείνει τις εξής συναρτήσεις ανάκτησης:

$$F(d, t_1 AND \ldots \; AND \; t_n) = (1-\gamma)\cdot MIN(w_{d1}, \; \ldots \; , w_{dn}) + \gamma \cdot MAX(w_{d1}, \; \ldots \; , w_{dn}), 0 \le \gamma \le 0.5$$

$$F(d, t_1 OR \ldots \; OR \; t_n) = (1-\gamma')\cdot MIN(w_{d1}, \; \ldots \; , w_{dn}) + \gamma' \cdot MAX(w_{d1}, \; \ldots \; , w_{dn}), 0.5 \le \gamma' \le 1$$

Γίνεται εύκολα αντιληπτό ότι, με την χρήση αυτών των συναρτήσεων ανάκτησης αποφεύγονται τα προβλήματα τα οποία παρουσιάζονται από την ίδια τη φύση των τελεστών *MIN* και *MAX* ,δηλαδή το ότι το αποτέλεσμα προκύπτει μόνο από

τον ένα τελεστέο, ωστόσο το μοντέλο αυτό λαμβάνει υπ' όψιν μόνο δύο τελεστέους, δηλαδή τον ελάχιστο και τον μέγιστο τελεστέο, για τον υπολογισμό της ομοιότητας εγγράφου - ερώτησης. Για παράδειγμα, έστω τα έγγραφα d_1 και d_2 και η ερώτηση q όπως φαίνεται παρακάτω:

$$d_1 = \{(t_1,0),(t_2,0.9),(t_3,0.9),\ldots,(t_{99},0.9),(t_{100},1)\}$$

$$d_2 = \{(t_1,0),(t_2,0.1),(t_3,0.1),\ldots,(t_{99},0.1),(t_{100},1)\}$$

$$q = t_1 \quad AND \quad t_2 \quad AND \quad \ldots \quad AND \quad t_{100}$$

Χρησιμοποιώντας το μοντέλο Waller-Kraft προκύπτει ότι και τα δύο έγγραφα έχουν τον ίδιο βαθμό ομοιότητας με την ερώτηση q, διότι το αποτέλεσμα των *MIN* και *MAX* είναι 0 και 1 αντίστοιχα και για τα δύο έγγραφα, ωστόσο, οι περισσότεροι άνθρωποι θα αποφάσιζαν ότι το έγγραφο d_1 είναι περισσότερο όμοιο με την ερώτηση q από ότι το έγγραφο d_2. Το πρόβλημα προέκυψε διότι το μοντέλο χρησιμοποιεί μόνο δύο από τους τελεστέους που σχηματίζουν την ερώτηση, τον ελάχιστο και τον μέγιστο.

2.5.3 Το επεκταμένο μοντέλο δίτιμης λογικής (extended boolean model)

Στα προηγούμενα τμήματα παρουσιάστηκαν μοντέλα τα οποία πρότειναν μεθόδους υπολογισμού της συνάρτησης ανάκτησης πληροφορίας χωρίς να λαμβάνουν υπ' όψιν το ενδεχόμενο ότι θα ήταν ίσως επιθυμητό οι όροι της ερώτησης να φέρουν και αυτοί κάποιο βάρος. Το βάρος ενός όρου μιας ερώτησης υποδηλώνει τη σπουδαιότητα αυτού του όρου στη συγκεκριμένη ερώτηση.

Είναι γνωστό ότι η αποδοτικότητα της ανάκτησης μπορεί να βελτιωθεί θέτοντας συντελεστές σπουδαιότητας ή βάρη στους όρους και στις προτάσεις που σχηματίζουν μια ερώτηση. Για παράδειγμα μια τέτοια, με βάρη, ερώτηση q μπορεί να έχει τη μορφή: $q = \left(\left((t_1, w_{q1})OR(t_2, w_{q2})\right), w_{q1ORq2}\right) AND(t_3, w_{q3})$ όπου w_{qi} είναι το βάρος του όρου t_i μέσα στη ερώτηση q και $w_{q1 \quad OR \quad q2}$ είναι το βάρος της πρότασης $t_1 \quad OR \quad t_2$. Τα βάρη των όρων της ερώτησης, όπως και τα βάρη των προτάσεών της (clauses), μπορούν να παίρνουν τιμές μεταξύ 0 και 1, δηλαδή

$0 \le w_{qi}, w_{q1 \quad OR \quad q2} \le 1$. Μια ερώτηση χωρίς βάρη είναι ισοδύναμη με μια ερώτηση με βάρη στην οποία τα βάρη όλων των όρων της είναι ίσα με 1.

Το επεκταμένο μοντέλο δίτιμης λογικής [11] είναι ένα από τα μοντέλα τα οποία υποστηρίζουν τη δυνατότητα να υπάρχουν βάρη και στους όρους της ερώτησης. Η συνάρτηση ανάκτησης $F(d,q)$ υπολογίζεται βασιζόμενη σε γενικευμένες συναρτήσεις υπολογισμού απόστασης μεταξύ των διανυσμάτων της ερώτησης και του εγγράφου. Βασικό χαρακτηριστικό του μοντέλου είναι ότι χειρίζεται ενιαία τόσο τα βάρη των όρων της ερώτησης όσο και τα βάρη των όρων δεικτοδότησης, οι οποίοι αναπαριστούν τα έγγραφα της συλλογής. Επίσης ο υπολογισμός της ομοιότητας εγγράφου-ερώτησης προσαρμόζεται μέσω του ελέγχου μιας ειδικής παραμέτρου γνωστή ως τιμή-p (p-value). Έτσι για διαφορετικές τιμές του p παίρνουμε διαφορετικούς βαθμούς ομοιότητας μεταξύ εγγράφων-ερώτησης.

Η συνάρτηση ανάκτησης πληροφορίας F μπορεί να υπολογιστεί σύμφωνα με τις παρακάτω εκφράσεις:

$$F\left(d,\left(t_1,w_{q1}\right)AND\ldots \quad AND\left(t_n,w_{qn}\right)\right)=1-\left(\frac{\sum_{i=1}^{n}\left(\left(1-w_{di}\right)^p\cdot w_{qi}^p\right)}{\sum_{i=1}^{n}w_{qi}^p}\right)^{\frac{1}{p}} \quad , \quad 1\le p\le\infty$$

$$F\left(d,\left(t_1,w_{q1}\right)OR\ldots \quad OR\left(t_n,w_{qn}\right)\right)=\left(\frac{\sum_{i=1}^{n}\left(w_{di}^p\cdot w_{qi}^p\right)}{\sum_{i=1}^{n}w_{qi}^p}\right)^{\frac{1}{p}} \quad , \quad 1\le p\le\infty$$

Ιδιαίτερο ενδιαφέρον παρουσιάζουν κάποιες ειδικές περιπτώσεις όπως για $p=\infty$, $p=1$ και $1<p<\infty$. Παρουσιάζονται κάθε μια από αυτές ξεχωριστά θεωρώντας μόνο δύο όρους.

1. Για $p=\infty$,

$$F\left(d,\left(t_1,w_{q1}\right)AND\left(t_2,w_{q2}\right)\right)=1-\frac{\max\left[w_{q1}\cdot\left(1-w_{d1}\right),w_{q2}\cdot\left(1-w_{d2}\right)\right]}{\max\left(w_{q1},w_{q2}\right)}$$

Όταν χρησιμοποιούνται μόνο οι τιμές 0 και 1 για τα βάρη, δηλαδή όταν $w_{q1} = w_{q2} = 1$, τότε η προηγούμενη έκφραση περιορίζεται στην:

$$F\left(d,\left(t_1,w_{q1}\right)AND\left(t_2,w_{q2}\right)\right) = \min\left(w_{d1},w_{d2}\right)$$

Ομοίως προκύπτει ότι:

$$F\left(d,\left(t_1,w_{q1}\right)OR\left(t_2,w_{q2}\right)\right) = \frac{\max\left(w_{q1}\cdot w_{d1},w_{q2}\cdot w_{d2}\right)}{\max\left(w_{q1},w_{q2}\right)}$$

επίσης θεωρώντας ότι $w_{q1} = w_{q2} = 1$ προκύπτει:

$$F\left(d,\left(t_1,w_{q1}\right)OR\left(t_2,w_{q2}\right)\right) = \max\left(w_{d1},w_{d2}\right)$$

Από τα προηγούμενα γίνεται φανερό ότι για $p = \infty$ και θεωρώντας τα βάρη των όρων της ερώτησης ίσα με 1 το επεκταμένο μοντέλο δίτιμης λογικής περιορίζεται στο μοντέλο ασαφούς συνόλου και για την περίπτωση που τα βάρη είναι 0 και 1 στο μοντέλο δίτιμης λογικής.

2. Για $p = 1$,

$$F\left(d,\left(t_1,w_{q1}\right)AND\left(t_2,w_{q2}\right)\right) = \frac{w_{q1}\cdot w_{d1} + w_{q2}\cdot w_{d2}}{w_{q1} + w_{q2}}$$

$$F\left(d,\left(t_1,w_{q1}\right)OR\left(t_2,w_{q2}\right)\right) = \frac{w_{q1}\cdot w_{d1} + w_{q2}\cdot w_{d2}}{w_{q1} + w_{q2}}$$

δηλαδή

$$F\left(d,\left(t_1,w_{q1}\right)AND\left(t_2,w_{q2}\right)\right) = F\left(d,\left(t_1,w_{q1}\right)OR\left(t_2,w_{q2}\right)\right)$$

Συνεπώς όταν $p = 1$ προκύπτουν όμοια αποτελέσματα για *AND* και *OR* ερωτήσεις. Επίσης ο βαθμός που αποδίδεται σε κάθε έγγραφο, σε σχέση με μια ερώτηση, είναι ο ίδιος με αυτόν που θα προέκυπτε από ένα σύστημα το οποίο θα χρησιμοποιούσε το μοντέλο διανυσματικού χώρου στο οποίο ο βαθμός ομοιότητας προκύπτει ως άθροισμα γινομένων $\left(\sum q_i \cdot d_i\right)$.

3. Όταν $1 < p < \infty$ δημιουργείται ένα "χαλαρό" σύστημα δίτιμης λογικής στο οποίο και η ερμηνεία των λογικών τελεστών είναι επίσης "χαλαρή". Στο σημείο αυτό αξίζει να γίνουν δύο παρατηρήσεις σχετικά με τη φύση του μοντέλου:

- Όταν υποβάλλονται *AND* ερωτήσεις, όπως *(A AND B)*, τότε προτιμούνται έγγραφα τα οποία περιέχουν και τους δύο όρους παρά μόνο τον ένα, αυτό όμως δεν σημαίνει ότι τα έγγραφα τα οποία περιέχουν τον ένα μόνο από τους δύο όρους πρέπει να απορριφθούν.
- Εντελώς ανάλογα όταν εκτελούνται *OR* ερωτήσεις όπως *(A OR B)* τα έγγραφα τα οποία περιέχουν οποιονδήποτε από τους όρους *A* και *B* θα ανακτηθούν, ωστόσο, τα έγγραφα τα οποία περιέχουν και τους δύο όρους είναι γενικά περισσότερο κατάλληλα σε σχέση με αυτά που περιέχουν μόνο τον ένα από τους δύο.

Η αναφορά στο επεκταμένο μοντέλο δίτιμης λογικής θα ολοκληρωθεί παρουσιάζοντας μερικά από τα πλεονεκτήματα αλλά και τα μειονεκτήματά του.

Το επεκταμένο μοντέλο δίτιμης λογικής παρέχει τα πλεονεκτήματα που προκύπτουν από την χρήση των δομημένων ερωτήσεων δίτιμης λογικής (boolean queries) και δίνει τη δυνατότητα να υπάρχουν βαθμολογημένα έγγραφα στην έξοδο. Η "χαλαρή" ερμηνεία των λογικών τελεστών -*AND, OR, NOT*- εμφανίζεται περισσότερο κατάλληλη για το περιβάλλον των συστημάτων ανάκτησης πληροφορίας. Τα έγγραφα τα οποία περιέχουν μόνο έναν από τους όρους μιας *AND*-πρότασης μπορούν να ανακτηθούν αλλά με χαμηλότερο βαθμό, σε σχέση με αυτά που περίεχουν δύο όρους ή και όλους τους όρους, ενώ, τα έγγραφα που περιέχουν περισσότερους από έναν όρους μιας *OR* - πρότασης θα έχουν καλύτερο βαθμό από αυτά που περιέχουν μόνο έναν.

Στα μειονεκτήματα του μοντέλου μπορεί να συμπεριληφθεί το γεγονός ότι οι περισσότερες από τις ιδιότητες των λογικών τελεστών, όπως για παράδειγμα η προσεταιριστικότητα, δεν ισχύουν. Αυτό σημαίνει ότι λογικά ισοδύναμες ερωτήσεις όπως οι: *(A AND (B AND C))* και (*(A AND B) AND C)* μπορεί να δώσουν διαφορετικούς βαθμούς στα έγγραφα της εξόδου αν και οι διαφορές είναι μικρές και δεν φαίνεται να επηρεάζουν την ποιότητα των αποτελεσμάτων.

Ένα δεύτερο, και περισσότερο σοβαρό πρόβλημα, είναι η πολυπλοκότητα την οποία περιέχουν οι υπολογισμοί για τη λήψη των βαθμών των εγγράφων στην έξοδο. Συγκεκριμένα η αναδρομική εφαρμογή της σχέσης υπολογισμού της συνάρτησης ανάκτησης *F* παράγει περίπλοκες εκφράσεις, για λογικές ερωτήσεις με μεγάλο αριθμό τελεστών.

Στο σημείο αυτό ολοκληρώνεται η αναφορά στο επεκταμένο μοντέλο δίτιμης λογικής.

2.5.4 Το μοντέλο Infinite-One

Ένα ακόμα μοντέλο το οποίο υποστηρίζει την ύπαρξη βαρών στους όρους μιας ερώτησης είναι το μοντέλο Infinite-One [10]. Στο μοντέλο αυτό γίνεται χρήση των γνωστών τελεστών *MIN* και *MAX*. Συγκεκριμένα το μοντέλο προτείνει τις εξής σχέσεις για τον υπολογισμό της συνάρτησης ανάκτησης F :

$$F\left(d,\left(t_1,w_{q1}\right)AND\ldots\quad AND\left(t_n,w_{qn}\right)\right)=$$

$$\gamma\cdot\left(1-\frac{MAX\left(\left(1-w_{d1}\right)\cdot w_{q1},\ldots,\left(1-w_{dn}\right)\cdot w_{qn}\right)}{MAX\left(w_{q1},\ldots,w_{qn}\right)}\right)+\left(1-\gamma\right)\cdot\frac{\sum_{i=1}^{n}w_{di}\cdot w_{qi}}{\sum_{i=1}^{n}w_{qi}},\quad 0\le\gamma\le 1$$

$$F\left(d,\left(t_1,w_{q1}\right)OR\ldots\quad OR\left(t_n,w_{qn}\right)\right)=$$

$$\gamma\cdot\frac{MAX\left(w_{d1}\cdot w_{q1},\ldots,w_{dn}\cdot w_{qn}\right)}{MAX\left(w_{q1},\ldots,w_{qn}\right)}+\left(1-\gamma\right)\cdot\frac{\sum_{i=1}^{n}w_{di}\cdot w_{qi}}{\sum_{i=1}^{n}w_{qi}},\quad 0\le\gamma\le 1$$

Το μοντέλο Infinite-One λαμβάνει υπ'όψιν όλους τους τελεστέους για τον υπολογισμό της τιμής του εγγράφου, ωστόσο, υπολογίζει αυτή την τιμή βασιζόμενο κυρίως στο ελάχιστο και στο μέγιστο βάρος για το *AND* και το *OR* αντίστοιχα. Ας θεωρηθεί το επόμενο παράδειγμα, όπου υπάρχουν δύο έγγραφα d_1 και d_2 καθώς και η ερώτηση q της οποίας οι όροι θεωρείται ότι έχουν όλοι βάρος 1 όπως φαίνεται παρακάτω:

$$d_1 = \{(t_1,0),(t_2,1),(t_3,1),\ldots,(t_{100},1)\}$$

$$d_2 = \{(t_1 0.4),(t_2,0.6),(t_3,0.6),\ldots,(t_{100},0.6)\}$$

$$q = t_1 \quad AND \quad t_2 \quad AND \quad \ldots \quad AND \quad t_{100}$$

Για $\gamma = 0{,}5$ το μοντέλο δίνει τον ίδιο βαθμό ομοιότητας μεταξύ των εγγράφων d_1 και d_2 ως προς την ερώτηση q, ωστόστο διαισθητικά φαίνεται ότι το έγγραφο d_1 είναι περισσότερο όμοιο με την ερώτηση q απ' ότι το έγγραφο d_2

Στο σημείο αυτό ολοκληρώνεται η αναφορά στις ερωτήσεις ομοιότητας και στα μαθηματικά μοντέλα που χρησιμοποιούνται για να αναπαραστήσουν τα συστήματα ανάκτησης πληροφορίας. Το σύστημα που κατασκευάστηκε υποστηρίζει, για τον υπολογισμό της συνάρτησης ανάκτησης F και τα τέσσερα μοντέλα, ασαφούς συνόλου, Waller-Kraft, εκτεταμένης δίτιμης λογικής (ή P-Norm) και Infinite-One. Το σπουδαιότερο όμως είναι, όπως θα δειχθεί παρακάτω, ότι όλα αυτά τα μοντέλα αντιμετωπίζονται με ενιαίο και παράλληλο τρόπο.

2.6 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκαν βασικά στοιχεία για τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων και για τα συστήματα ανάκτησης πληροφορίας. Έτσι αναφέρθηκαν κάποια ιδιαίτερα γνωρίσματα των συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων και έγινε μια σύντομη εισαγωγή στα συστήματα ανάκτησης πληροφορίας. Παρουσιάστηκαν δύο μοντέλα για την αναπαράσταση συστημάτων ανάκτησης πληροφορίας, το μοντέλο διανυσματικού χώρου και το επεκταμένο πλαίσιο εργασίας δίτιμης ανάκτησης. Στοιχεία των δύο αυτών μοντέλων υιοθετήθηκαν στα πλαίσια αυτής της εργασίας .

Παρουσιάστηκαν επίσης τα τέσσερα μοντέλα υπολογισμού ομοιότητας τα οποία υποστηρίζονται από την παρούσα εργασία. Έγινε εκτενής αναφορά τόσο στις εκφράσεις που προτείνει το κάθε μοντέλο για τον υπολογισμό των λογικών τελεστών *(AND, OR)* όσο και στα πλεονεκτήματα και μειονεκτήματα του κάθε μοντέλου.

Στο επόμενο κεφάλαιο θα παρουσιαστεί θα παρουσιαστεί ένα μαθηματικό μοντέλο για την αναπαράσταση και την εκτέλεση ερωτήσεων ομοιότητας.

3. Το μοντέλο εκτέλεσης ερωτήσεων ομοιότητας

3.1 Εισαγωγή

Μετά την ανασκόπηση της σχετικής εργασίας στο προηγούμενο κεφάλαιο στο σημείο αυτό θα παρουσιαστεί το θεωρητικό μοντέλο για την αναπαράσταση και την εκτέλεση ερωτήσεων ομοιότητας. Το μοντέλο αναπτύχθηκε στα πλαίσια της εργασίας [14] και όλες οι σχέσεις που θα αναφερθούν έχουν αποδειχθεί στα πλαίσια αυτής της εργασίας.

Το μοντέλο που θα παρουσιαστεί βασίζεται στη μαθηματική έννοια της ***σχέσης***. Ξεκινώντας από το μοντέλο της δίτιμης λογικής ορίζεται η έννοια αυτή και στη συνέχεια με κατάλληλες τροποποιήσεις προκύπτει ο ορισμός της ***σχέσης αξιολόγησης***. Με βάση αυτή τη νέα έννοια καθορίζεται ένας αλγεβρικός μηχανισμός εκτέλεσης ερωτήσεων ομοιότητας που μπορεί να προσαρμοστεί πολύ εύκολα σε διαφορετικά μοντέλα υπολογισμού ομοιότητας. Θα δοθούν πρώτα μερικοί ορισμοί και στη συνέχεια θα περιγραφεί η μεθοδολογία που ακολουθείται για την εκτέλεση ερωτήσεων ομοιότητας. Η μεθοδολογία αυτή βασίζεται πάνω στην πράξη της ***σύνθεσης σχέσεων***.

Στο επόμενο τμήμα δίνονται οι ορισμοί της σχέσης, της σχέσης αξιολόγησης και της σύνθεσης σχέσεων. Στο τμήμα 3.3 περιγράφεται ο τρόπος με τον οποίο

υποστηρίζονται διάφορα μοντέλα υπολογισμού ομοιότητας ενώ στο τμήμα 3.4 γίνεται λόγος για την υποστήριξη ανάδρασης σχετικότητας (τμήμα 2.4.1.1). Τέλος στο τμήμα 3.5 δίνεται μια σύντομη περίληψη του κεφαλαίου.

3.2 Ορισμοί

3.2.1 Η έννοια της σχέσης στη δίτιμη λογική

Μια σχέση *r* μεταξύ δύο συνόλων *V* και *W*, όπως είναι γνωστό, είναι ένα υποσύνολο του καρτεσιανού γινομένου $V \times W$. Δηλαδή, $r \subseteq V \times W$. Είναι επίσης γνωστό ότι ένα υποσύνολο *A* ενός συνόλου *S* μπορεί να ορισθεί ισοδύναμα με την χαρακτηριστική του συνάρτηση, δηλαδή μια συνάρτηση από το *S* στο $\{0,1\}$ η οποία αποδίδει τιμή 1 σε κάθε στοιχείο του *S* το οποίο ανήκει στο *A* και την τιμή 0 σε κάθε στοιχείο του *S* το οποίο δεν ανήκει στο *A*. Συνεπώς μια **σχέση** $\boldsymbol{r \subseteq V \times W}$ ***(3.1)*** μπορεί να δοθεί από τη χαρακτηριστική της συνάρτηση f_r η οποία ορίζεται, βάσει της σχέσης *r* ως εξής:

$$f_r : V \times W \to \{0,1\},\ f_r(v,w) = \begin{cases} 1, \langle v,w \rangle \in r \\ 0, \langle v,w \rangle \notin r \end{cases}$$

Και αντίστροφα, αν είναι γνωστή η χαρακτηριστική συνάρτηση $f_r : V \times W \to \{0,1\}$ η σχέση *r* προκύπτει μονοσήμαντα ως εξής:

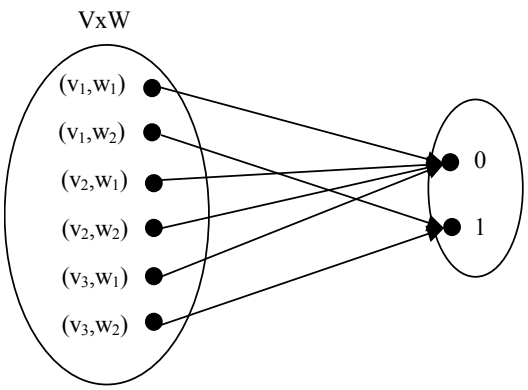
$$r = f_r^{-1}[\{1\}] = \left\{ \langle v,w \rangle \in V \times W \middle| f_r(v,w) = 1 \right\}$$

Ένα παράδειγμα σχέσης έχει ως εξής. Έστω τα σύνολα $V = \{v_1, v_2, v_3\}$ και $W = \{w_1, w_2\}$. Το καρτεσιανό τους γινόμενο είναι το $V \times W = \{(v_1,w_1),(v_1,w_2),(v_2,w_1),(v_2,w_2),(v_3,w_1),(v_3,w_2)\}$.

Ορίζουμε για παράδειγμα τη σχέση $r \subseteq V \times W$ ως $r = \{(v_1,w_2),(v_3,w_2)\}$. Κατά συνέπεια θα ισχύει :

$f_r(v_1,w_2) = f_r(v_3,w_2) = 1$ και $f_r(v_1,w_1) = f_r(v_2,w_1) = f_r(v_2,w_2) = f_r(v_3,w_1) = 0$.

Σχηματικά η σχέση r θα απεικονιζόταν ως εξής:



Σχήμα 3-1:Γραφική απεικόνιση της σχέσης r του προηγούμενου παραδείγματος

Η σχέση r του παραδείγματος μπορεί να αναπαρασταθεί και από τον εξής πίνακα:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \end{array} \begin{array}{c} \begin{array}{cc} w_1 & w_2 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

Αναπαράσταση της σχέσης r του παραδείγματος με τη χρήση πίνακα

Το σύνολο, συνεπώς, των σχέσεων μεταξύ των συνόλων V και W είναι το σύνολο των συναρτήσεων με πεδίο ορισμού το καρτεσιανό γινόμενο $V \times W$ και πεδίο τιμών το $\{0,1\}$. Το σύνολο αυτό συμβολίζεται με $\{0,1\}^{V \times W}$. Το γεγονός ότι η r είναι μια σχέση μεταξύ των συνόλων V και W θα συμβολίζεται στο εξής ως $r{:}V \times W \rightarrow \{0,1\}$ ή $r \in \{0,1\}^{V \times W}$. Δηλαδή θα ταυτίζεται η σχέση με τη χαρακτηριστική της συνάρτηση.

Οι βασικές πράξεις που ορίζονται σε σχέσεις είναι οι γνωστές συνολοθεωρητικές πράξεις (ένωση, τομή διαφορά, συμπλήρωμα) καθώς και η ***σύνθεση σχέσεων***. Για τις πράξεις ένωση, τομή, διαφορά και συμπλήρωμα οι ορισμοί είναι γνωστοί, για την πράξη της σύνθεσης όμως τα πράγματα δεν είναι τόσο προφανή. Ο κλασικός ορισμός της σύνθεσης για δύο σχέσεις $r \subseteq V \times W$ και $s \subseteq W \times Z$ είναι ο εξής:

$$r \circ s = \left\{ \langle v,z \rangle \in V \times Z \middle| \exists w \in W{:} \langle v,w \rangle \in r AND \langle w,z \rangle \in s \right\}$$

Ο ***ορισμός*** αυτός της σύνθεσης μπορεί να τροποποιηθεί ως εξής:

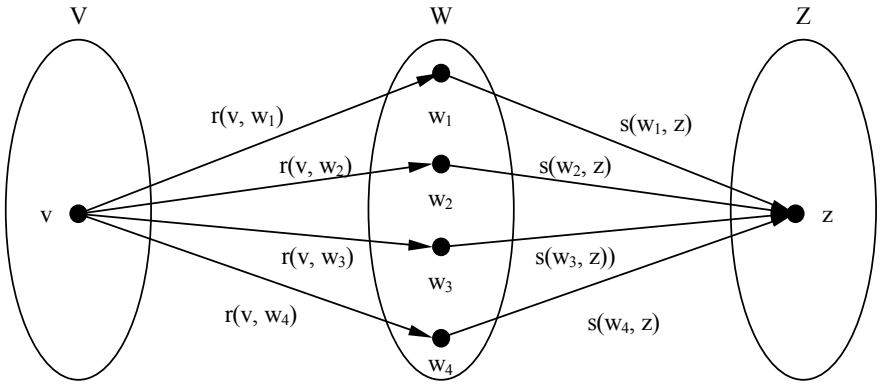
$$\boldsymbol{r \circ s} = \left\{ \langle \boldsymbol{v,z} \rangle \in \boldsymbol{V \times Z} \,\middle|\, \underset{\boldsymbol{w \in W}}{\boldsymbol{OR}} \left(\langle \boldsymbol{v,w} \rangle \in \boldsymbol{r AND} \langle \boldsymbol{w,z} \rangle \in \boldsymbol{s} \right) \right\} \qquad (3.2)$$

Οι δυο ορισμοί είναι τελείως ισοδύναμοι [18]. Όμως, ο δεύτερος, προσφέρει τη δυνατότητα να ορισθεί η σύνθεση με τρόπο παρόμοιο με τον πολλαπλασιασμό πινάκων.

Μια σχέση $r \in \{0,1\}^{V \times W}$ μπορεί να αναπαρασταθεί με έναν δισδιάστατο πίνακα $|V| \times |W|$, του οποίου κάθε στοιχείο είναι η αντίστοιχη τιμή της συνάρτησης *r*. Φυσικά, για να είναι αυτό εφικτό πρέπει τα σύνολα *V* και *W* να είναι πεπερασμένα και να οριστεί μια διάταξη των στοιχείων τους. Έτσι, η ***τιμή της σύνθεσης*** $\boldsymbol{r \circ s}$ για οποιοδήποτε $\langle v,z \rangle \in V \times Z$ δίνεται από τον επόμενο τύπο:

$$(\boldsymbol{r \circ s})(\boldsymbol{v,z}) = \underset{\boldsymbol{w \in W}}{\boldsymbol{OR}} \left(\boldsymbol{r}(\boldsymbol{v,w})\, \boldsymbol{AND} \quad \boldsymbol{s}(\boldsymbol{w,z}) \right) \qquad (3.3)$$

Σχηματικά, ο υπολογισμός αυτός δίνεται στο επόμενο σχήμα:



$r \circ s(v, z) = r(v, w_1) \wedge s(w_1, z) \vee r(v, w_2) \wedge s(w_2, z) \vee r(v, w_3) \wedge s(w_3, z) \vee r(v, w_4) \wedge s(w_4, z)$

Σχήμα 3-2: Αναπαράσταση του υπολογισμού στοιχείου της σύνθεσης δύο σχέσων r, s

Από τον τελευταίο τύπο είναι προφανές ότι ***ο ορισμός της σύνθεσης σχέσεων είναι πανομοιότυπος με την πράξη του πολλαπλασιασμού πινάκων***. Η λογική σύζευξη (*AND*) παίζει το ρόλο του πολλαπλασιασμού αριθμών και η λογική διάζευξη (*OR*) το ρόλο της πρόσθεσης αριθμών. Στην συνέχεια θα μελετηθεί πως οι σχέσεις όπως ορίστηκαν τελικά, ως πίνακες, μπορούν να χρησιμοποιηθούν για να απαντηθούν ερωτήσεις σε συστήματα ανάκτησης πληροφορίας.

3.2.2 Από τη σχέση στη σχέση με αξιολόγηση

Στην προηγούμενη ενότητα περιγράφηκε η έννοια της σχέσης στο μοντέλο δίτιμης λογικής και ταυτίστηκε η σχέση με τη χαρακτηριστική της συνάρτηση. Μέσα από αυτή την αμφιμονοσήμαντη αντιστοιχία σχέσεων και χαρακτηριστικών συναρτήσεων προκύπτει η έννοια της σχέσης με αξιολόγηση με κατάλληλη γενίκευση της χαρακτηριστικής συνάρτησης όπως θα δοθεί στη συνέχεια.

Έστω μια σχέση $r \subseteq V \times W$. Στα μοντέλα ασαφούς λογικής, η αποτίμηση μιας πρότασης δεν είναι κατ' ανάγκη μόνο *αληθής* (1) ή *ψευδής* (0), αλλά μπορεί να πάρει οποιαδήποτε τιμή στο διάστημα $[0,1]$. Με βάση αυτή την προσέγγιση η αποτίμηση της πρότασης $\langle v, w \rangle \in r$ που γίνεται με βάση τη χαρακτηριστική συνάρτηση f_r της σχέσης *r*, σε ένα μοντέλο ασαφούς λογικής, θα δίνεται και πάλι με τη χρήση χαρακτηριστικής συνάρτησης, της οποίας όμως, το πεδίο τιμών θα πρέπει να είναι το διάστημα $[0,1]$ και όχι μόνο το σύνολο $\{0,1\}$. Με βάση αυτές τις παρατηρήσεις, ο ορισμός της σχέσης με αξιολόγηση είναι ο εξής:

Μια σχέση με αξιολόγηση r από το σύνολο V στο σύνολο W είναι μια συνάρτηση $r{:}V \times W \to [0,1]$***. Το σύνολο των σχέσεων με αξιολόγηση από το σύνολο V στο σύνολο W θα συμβολίζεται με*** $[0,1]^{V \times W}$***.***

Ισοδύναμα μια τέτοια σχέση μπορεί να αναπαρασταθεί με έναν δισδιάστατο πίνακα $|V| \times |W|$, του οποίου κάθε στοχείο είναι η αντίστοιχη τιμής της *r*. Φυσικά, για να είναι αυτό εφικτό πρέπει τα σύνολα *V* και *W* να είναι πεπερασμένα και να οριστεί μια διάταξη των στοιχείων τους. Η αναπαράσταση αυτή απλοποιεί τη μελέτη των συναρτήσεων αυτών και επιτρέπει την άμεση εφαρμογή συμπερασμάτων από τη γραμμική άλγεβρα. Στη συνέχεια, λοιπόν, μια σχέση με αξιολόγηση *r* θα παριστάνεται με έναν πίνακα $|V| \times |W|$.

Για παράδειγμα αν θεωρήσουμε τα σύνολα $V = \{v_1, v_2, v_3\}$ και $W = \{w_1, w_2\}$ μπορούμε να ορίσουμε για παράδειγμα τη σχέση $r \subseteq V \times W$ με βάση τη χαρακτηριστική της συνάρτηση f_r ως εξής:

$f_r(v_1, w_1) = f_r(v_2, w_1) = f_r(v_3, w_1) = 0$ και $f_r(v_1, w_2) = 0.2, f_r(v_2, w_2) = 1, f_r(v_3, w_2) = 0.1$

Η σχέση r του παραδείγματος μπορεί να αναρασταθεί από τον εξής πίνακα:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \end{array} \begin{array}{c} \begin{array}{cc} w_1 & w_2 \end{array} \\ \begin{bmatrix} 0 & 0.2 \\ 0 & 1 \\ 0 & 0.1 \end{bmatrix} \end{array}$$

Με r^T συμβολίζουμε την αντίστροφη σχέση της r και με $\bar{r}$ την αντίθετη. Δηλαδή,

$$r^T(w,v) = r(v,w) \tag{3.4}$$

$$\bar{r}(v,w) = 1 - r(v,w) \tag{3.5}$$

Πριν περιγραφεί ο τρόπος με τον οποίο απαντώνται ερωτήσεις ομοιότητας σε ένα σύστημα ανάκτησης πληροφορίας χρησιμοποιώντας σχέσεις με αξιολόγηση, είναι απαραίτητο να μελετηθεί περισσότερο η πράξη της σύνθεσης, ως πράξη μεταξύ σχέσεων με αξιολόγηση πλέον. Γενικεύοντας θα οριστεί ως ***σύνθεση σχέσεων αξιολόγησης*** μια πράξη $\circ$ μεταξύ σχέσεων αξιολόγησης που απεικονίζει τις σχέσεις $r: V \times W \to [0,1]$ και $s: W \times Z \to [0,1]$ στη σχέση $(r \circ s): V \times Z \to [0,1]$ βάσει μιας πράξης πολλαπλασιασμού $\otimes: [0,1] \times [0,1] \to A$ και μιας πράξης πρόσθεσης $\oplus: A \times A \to [0,1]$ ως εξής:

$$(\boldsymbol{r} \circ \boldsymbol{s})(\boldsymbol{v}, \boldsymbol{z}) = \bigoplus_{\boldsymbol{w} \in \boldsymbol{W}} \left(\boldsymbol{r}(\boldsymbol{v}, \boldsymbol{w}) \otimes \boldsymbol{s}(\boldsymbol{w}, \boldsymbol{z}) \right) \tag{3.6}$$

Το A είναι ένα σύνολο ενδιάμεσων τιμών που δεν είναι πάντα το διάστημα $[0,1]$ - αν και συνήθως είναι αυτό - όπως θα φανεί στη συνέχεια. Είναι προφανές ότι για κάθε διαφορετικό ζευγάρι πράξεων $\oplus$ και $\otimes$ προκύπτει διαφορετική πράξη σύνθεσης.

Το ζητούμενο είναι να ορισθούν κατάλληλα οι πράξεις $\oplus$ και $\otimes$ ώστε με τη βοήθεια πράξεων σύνθεσης να είναι εφικτός ο υπολογισμός της απάντησης σε μια ερώτηση σε ένα σύστημα ανάκτησης πληροφορίας.

Προκειμένου τα παραπάνω να γίνουν περισσότερο κατανοητά ας θεωρήσουμε το εξής παράδειγμα. Έστω οι σχέσεις αξιολόγησης $R: A \times B \to [0,1]$ και $S: B \times C \to [0,1]$. Η κάθε μια από αυτές τις σχέσεις μπορεί να αναπαρασταθεί με έναν δισδιάστατο πίνακα του οποίου κάθε στοιχείο είναι η αντίστοιχη τιμής τους. Έστω λοιπόν ότι αναπαρίστανται από τους παρακάτω πίνακες:

$$R:|A|\times|B| = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array}\begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.4 \\ 0.2 & 0.7 \\ 0.3 & 0.5 \end{bmatrix} \end{array} \qquad S:|B|\times|C| = \begin{array}{c} \\ b_1 \\ b_2 \end{array}\begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.7 \end{bmatrix} \end{array}$$

Από τους παραπάνω πίνακες προκύπτει, για παράδειγμα, για τη σχέση αξιολόγησης R ότι το στοιχείο a_1 σχετίζεται με το στοιχείο b_1 με ένα βάρος σπουδαιότητας 0.1.

Η σύνθεση $R \circ S$ ανάγεται στον πολλαπλασιασμό των παραπάνω πινάκων (τμήμα 3.2.2), δηλαδή:

$$(R\circ S)(a,c) = \oplus\left(R(a,b)\otimes S(b,c)\right) = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array}\begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.4 \\ 0.2 & 0.7 \\ 0.3 & 0.5 \end{bmatrix} \end{array} \circ \begin{array}{c} \\ b_1 \\ b_2 \end{array}\begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.7 \end{bmatrix} \end{array} =$$

$$= \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array}\begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{bmatrix} 0.1\otimes 0.1\oplus 0.4\otimes 0.2 & 0.1\otimes 0.3\oplus 0.4\otimes 0.7 \\ 0.2\otimes 0.1\oplus 0.7\otimes 0.2 & 0.2\otimes 0.3\oplus 0.7\otimes 0.7 \\ 0.3\otimes 0.1\oplus 0.5\otimes 0.2 & 0.3\otimes 0.3\oplus 0.5\otimes 0.7 \end{bmatrix} \end{array}$$

Ένα σύστημα ανάκτησης πληροφορίας μπορεί να οριστεί, όπως αναφέρθηκε στο 2ο κεφάλαιο ως η ακόλουθη τετράδα:

$$\langle T, Q, D, F\rangle$$

Ας θεωρηθεί μια σχέση αξιολόγησης $r: D\times T \to [0,1]$. Η σχέση αυτή εκφράζει τη συσχέτιση των αποθηκευμένων έγγραφων με τους όρους δεικτοδότησης. Έστω ότι οι ερωτήσεις που αναγνωρίζονται από το σύστημα είναι απλές διαζευκτικές ερωτήσεις χωρίς βάρη στα άτομα της ερωτήσης και έχουν την μορφή $q = \underset{j=1}{\overset{n}{OR}}\, t_j$. Στην περίπτωση αυτή, μια οποιαδήποτε ερώτηση $q = \underset{j=1}{\overset{n}{OR}}\, t_j$ μπορεί να θεωρηθεί ως το υποσύνολο $\{t_1, t_2, \ldots, t_n\} \subseteq T$. Συνεπώς το σύνολο Q των ερωτήσεων που αναγνωρίζονται από το σύστημα μπορεί να θεωρηθεί ως υποσύνολο του δυναμοσυνόλου του T. Δηλαδή $Q \subseteq \wp(T)$.

Η σχέση

$$s: Q\times T \to \{0,1\} \text{ με } s(q,t) = \begin{cases} 1, t \in q \\ 0, t \notin q \end{cases}$$

εκφράζει τη συσχέτιση όρων δεικτοδότησης με διαζευκτικές ερωτήσεις. Αν υπάρχει μια πράξη $\circ$ που ορίζεται βάσει των πράξεων $\oplus$ και $\otimes$, τα στοιχεία της σύνθεσης $s \circ r$ δίνονται από την επόμενη σχέση:

$$(s \circ r)(q,d) = \bigoplus_{t \in T} \left(s(q,t) \otimes r(t,d) \right)$$

Είναι δυνατόν να αποδειχθεί ότι τα στοιχεία της σύνθεσης αντιστοιχούν στις τιμές της συνάρτησης ανάκτησης *F*, δηλαδή ισχύει για κάθε $\langle q,d \rangle \in Q \times D$:

$$F = s \circ r \Leftrightarrow F(q,d) = (s \circ r)(q,d) \Leftrightarrow f_{OR}\left(F(t_1,d),\ldots,F(t_n,d) \right) = \bigoplus_{t \in T} \left(s(q,t) \otimes r(t,d) \right)$$

Αν *M* είναι ένα μοντέλο υπολογισμού ομοιότητας, τότε θα συμβολίζεται με $\circ_M^{OR}$ η πράξη της σύνθεσης που αντιστοιχεί στην διάζευξη (OR) και με $\oplus_M^{OR}$, $\otimes_M^{OR}$ οι πράξεις βάσει των οποίων ορίζεται η σύνθεση αυτή. Συνεπώς στην περίπτωση των απλών διαζευκτικών ερωτήσεων προκύπτει:

$$F = s \circ_M^{OR} r$$

Έστω τώρα, ότι οι ερωτήσεις που αναγνωρίζονται από το σύστημα είναι απλές συζευκτικές ερωτήσεις της μορφής $q = \underset{j=1}{\overset{n}{AND}} t_j$. Μπορεί επίσης να δειχθεί ότι:

$$F = s \circ_M^{AND} r$$

Αποδεικνύεται ότι όλες οι ερωτήσεις μπορούν να τεθούν σε κανονική (συζευκτική ή διαζευκτική) μορφή. Οι ερωτήσεις σε κανονική μορφή περιέχουν και άτομα της μορφής *NOT p*. Αν και το μοντέλο που αναπτύχτηκε στα πλαίσια της εργασίας [14] υποστηρίζει την ύπαρξη όρων της μορφής *NOT p* στις ερωτήσεις το σύστημα που κατασκευάστηκε δέχεται ερωτήσεις οι οποίες δεν περιέχουν τον τελεστή *NOT*.

Ας υποτεθεί λοιπόν ότι οι ερωτήσεις που αναγνωρίζονται από το σύστημα είναι σε κανονική συζευκτική μορφή, δηλαδή $q = \underset{i-1}{\overset{n}{AND}} \left(\underset{j=1}{\overset{m}{OR} t_j} \right)$. Έστω Q_{OR} το σύνολο όλων των διαζευκτικών που εμφανίζονται στις ερωτήσεις που αναγνωρίζονται από το σύστημα. Και πάλι κάθε διαζευκτικό μπορεί να θεωρηθεί σαν ένα υποσύνολο του *T*, οπότε μπορεί να οριστεί μια σχέση *s* ως εξής:

$$s: Q_{OR} \times T \to \{0,1\} \text{ με } s(q_{OR}, t) = \begin{cases} 1, t \in q_{OR} \\ 0, t \notin q_{OR} \end{cases}$$

Είναι προφανές, σύμφωνα με όσα έχουν διατυπωθεί για διαζευκτικές ερωτήσεις, ότι η συνάρτηση ανάκτησης F_{OR} των διαζευκτικών δίνεται από τη σύνθεση:

$$F_{OR} = s \circ_M^{OR} r$$

Τελείως ανάλογα με τη θεώρηση των διαζευκτικών ως υποσυνόλων του συνόλου *T*, μια συζευκτική ερώτηση μπρορεί να θεωρηθεί ως ένα υποσύνολο του συνόλου Q_{OR} όλων των διαζευκτικών. Έτσι, τελείως ανάλογα μπορεί να οριστεί η σχέση *t* ως εξής:

$$t: Q \times Q_{OR} \to \{0,1\} \text{ με } t(q, q_{OR}) = \begin{cases} 1, q_{OR} \in q \\ 0, q_{OR} \notin q \end{cases}$$

Τελείως αντίστοιχα με την περίπτωση των ερωτήσεων σε κανονική συζευκτική μορφή, είναι δυνατόν να δειχθεί [14] ότι:

$$\boldsymbol{F = t \circ_M^{AND} \left(s \circ_M^{OR} r \right)} \qquad (3.7)$$

Στην περίπτωση ερωτήσεων σε κανονική διαζευκτική μορφή θα είναι:

$$\boldsymbol{F = t \circ_M^{OR} \left(s \circ_M^{AND} r \right)} \qquad (3.8)$$

Αποδεικνύεται ότι, οι παραπάνω σχέσεις ισχύουν και στην περίπτωση που τα μοντέλα υπολογισμού ομοιότητας υποστηρίζουν βάρη στα άτομα της ερώτησης.

3.3 Ερωτήσεις με διάφορα μοντέλα υπολογισμού ομοιότητας

Σύμφωνα με όσα αναφέρθηκαν προηγουμένως, για να οριστεί μια πράξη σύνθεσης $\circ$ χρειαζεται μια πράξη $\oplus$ που να παίζει το ρόλο της πρόσθεσης και μια πράξη $\otimes$ που να παίζει το ρόλο του πολλαπλασιασμού. Για κάθε μοντέλο υπολογισμού ομοιότητας *M* πρέπει να οριστούν δύο διαφορετικές πράξεις σύνθεσης, μια από τις οποίες απαντά σε διαζευκτικές και μια σε συζευκτικές ερωτήσεις. Έστω ότι αυτές οι πράξεις συμβολίζονται με $\circ_M^{OR}$ και $\circ_M^{AND}$ αντίστοιχα. Φυσικά για κάθε μια από αυτές πρέπει να οριστεί ο "πολλαπλασιασμός" και η "πρόσθεση". Ας συμβολίσουμε με $\oplus_M^{OR}$, $\otimes_M^{OR}$, $\oplus_M^{AND}$, $\otimes_M^{AND}$ τις πράξεις αυτές. Είναι λοιπόν αναγκαίο να ορισθούν οι τέσσερις αυτές πράξεις για καθένα από τα μοντέλα

υπολογισμού ομοιότητας ώστε να ισχύει η ανάλυση που έγινε και αφορά την απάντηση ερωτήσεων ομοιότητας με χρήση σχέσεων αξιολόγησης.

Αν και στα πλαίσια αυτής της εργασίας δεν έχει βρεθεί συγκεκριμένη μεθοδολογία που ακολουθείται προκειμένου να οριστούν οι πράξεις $\oplus_M^{OR}$, $\otimes_M^{OR}$, $\oplus_M^{AND}$, $\otimes_M^{AND}$ για κάποιο μοντέλο M το θέμα αυτό, αν και δύσκολο, είναι εξαιρετικά ενδιαφέρον και μπορεί να αποτέλει αντικείμενο έρευνας και μελλοντικής επέκτασης του συστήματος.

Στη συνέχεια δίνονται οι πράξεις αυτές για καθένα από τα μοντέλα πλειότιμης λογικής που έχουμε υιοθετήσει και αναφέρθηκαν στο δεύτερο κεφάλαιο. Στα μοντέλα τα οποία δεν προσδιορίζουν βάρη στους όρους μιας ερώτησης το βάρος είναι ίσο με 1.

P-Norm

Τροποποιούμε τα βάρη τα οποία εμφανίζονται στους όρους της ερωτήσης και θέτουμε:

$$w_i^{'} = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

Οπότε οι πράξεις ορίζονται ως εξής:

$$a \otimes_{PN}^{OR} b = (a \cdot b)^p$$

$$a_1 \oplus_{PN}^{OR} a_2 \oplus_{PN}^{OR} \ldots \oplus_{PN}^{OR} a_n = \left(\sum_{i=1}^{n} a_i \right)^{1/p}$$

$$a \otimes_{PN}^{AND} b = \left((1-a) \cdot b\right)^p$$

$$a_1 \oplus_{PN}^{AND} a_2 \oplus_{PN}^{AND} \ldots \oplus_{PN}^{AND} a_n = 1 - \left(\sum_{i=1}^{n} a_i \right)^{1/p}$$

Infinite-One

Τροποποιούμε τα βάρη τα οποία εμφανίζονται στους όρους της ερωτήσης και θέτουμε:

$$w_i' = \left\langle \frac{w_i}{MAX\{w_1, w_2, \ldots w_n\}}, \frac{w_i}{\sum_{i=1}^{n} w_i} \right\rangle$$

Οπότε οι πράξεις ορίζονται ως εξής:

$$a \otimes_{IO}^{OR} \langle b, c \rangle = \langle a \cdot b, a \cdot c \rangle$$

$$\langle b_1, c_1 \rangle \oplus_{IO}^{OR} \langle b_2, c_2 \rangle \oplus_{IO}^{OR} \ldots \oplus_{IO}^{OR} \langle b_n, c_n \rangle = \gamma \cdot MAX\{b_1, b_2, \ldots, b_n\} + (1-\gamma)\left(\sum_{i=1}^{n} c_i\right)$$

$$a \otimes_{IO}^{AND} \langle b, c \rangle = \langle (1-a) \cdot b, a \cdot c \rangle$$

$$\langle b_1, c_1 \rangle \oplus_{IO}^{AND} \langle b_2, c_2 \rangle \oplus_{IO}^{AND} \ldots \oplus_{IO}^{AND} \langle b_n, c_n \rangle = \gamma \cdot \left(1 - MAX\{b_1, b_2, \ldots, b_n\}\right) + (1-\gamma)\left(\sum_{i=1}^{n} c_i\right)$$

Fuzzy Set

Οι πράξεις ορίζονται ως εξής:

$$a \otimes_{FS}^{OR} b = a \cdot b$$

$$a_1 \oplus_{FS}^{OR} a_2 \oplus_{FS}^{OR} \ldots \oplus_{FS}^{OR} a_n = MAX\{a_{1,}, a_2, \ldots, a_n\}$$

$$a \otimes_{FS}^{AND} b = MAX\{1-a, b\}$$

$$a_1 \oplus_{FS}^{AND} a_2 \oplus_{FS}^{AND} \ldots \oplus_{FS}^{AND} a_n = MIN\{a_{1,}, a_2, \ldots, a_n\}$$

Waller-Kraft

Οι πράξεις ορίζονται ως εξής:

$$a \otimes_{WK}^{OR} b = MAX\{1-a, b\}$$

$$a_1 \oplus_{WK}^{OR} a_2 \oplus_{WK}^{OR} \ldots \oplus_{WK}^{OR} a_n = (1-\gamma) \cdot MIN\{a_{1,}, a_2, \ldots, a_n\} + \gamma \cdot MAX\{a_{1,}, a_2, \ldots, a_n\}$$

$$a \otimes_{WK}^{AND} b = MAX\{1-a, b\}$$

$$a_1 \oplus_{WK}^{AND} a_2 \oplus_{WK}^{AND} \ldots \oplus_{WK}^{AND} a_n = (1-\gamma) \cdot MIN\{a_{1,}, a_2, \ldots, a_n\} + \gamma \cdot MAX\{a_{1,}, a_2, \ldots, a_n\}$$

3.4 Εκτέλεση ερωτήσεων κάνοντας χρήση της ανάδρασης σχετικότητας (relevance feedback)

Σε συστήματα ανάκτησης πληροφορίας έχει παρατηρηθεί μια αδυναμία των χρηστών ως προς τη δυνατότητά τους να διατυπώσουν μια ερώτηση η οποία να περιγράφει πραγματικά την πληροφορία που αναζητούν. Σε μια προσπάθεια αντιμετώπισης αυτής της αδυναμίας υπάρχει συχνά η δυνατότητα για τροποποίηση μιας ερώτησης με βάση τη σχετικότητα των αντικειμένων που επιστράφηκαν με τις πραγματικές ανάγκες του χρήστη. Αυτό ονομάζεται ανάδραση σχετικότητας (τμήμα 2.4.1.1)

Ο ίδιος ο ορισμός μιας σχέσης με αξιολόγηση μας επιτρέπει τον ισχυρισμό ότι τα μέλη των δύο συνόλων μιας σχέσης αξιολόγησης μπορούν να έχουν διπλό ρόλο. Μπορούν δηλαδή να είναι και τα προς ανάκτηση αντικείμενα όπως μπορούν να είναι και οι όροι δεικτοδότησης.

Έστω ότι στο παράδειγμα των ξενοδοχείων ορίζεται μια σχέση με αξιολόγηση *R* από το σύνολο *HotelsSet* των ξενοδοχείων στο σύνολο *FacilitiesSet* των διευκολύνσεων.

Έστω ότι τα προς ανάκτηση αντικείμενα είναι τα ξενοδοχεία, τότε οι όροι δεικτοδότησης και συνεπώς οι όροι που σχηματίζουν τις ερωτήσεις θα είναι κάποιες διευκολύνσεις. Μπορεί όμως επίσης να υποτεθεί ότι και τα ξενοδοχεία μπορούν να είναι όροι δεικτοδότησης και σχηματισμού ερωτήσεων, συνεπώς, στην περίπτωση αυτή τα αντικείμενα ανάκτησης θα είναι οι διευκολύνσεις.

Στο παράδειγμα των ξενοδοχείων αν διατυπωθεί μια ερώτηση χρησιμοποιώντας ως όρους δεικτοδότησης τις διευκολύνσεις σημαίνει ότι αναζητώνται τα ξενοδοχεία τα οποία ικανοποιούν τους όρους αυτούς (δηλαδή τις διευκολύνσεις) καλύτερα. Αντίστροφα, αν χρησιμοποιηθούν τα ξενοδοχεία ως όροι δεικτοδότησης, τότε αναζητώνται οι καλύτερες διευκολύνσεις των ξενοδοχείων που περιέχει η ερώτηση. Αυτή η ερώτηση θα ενδιέφερε για παράδειγμα έναν τουριστικό πράκτορα ο οποίος συνεργάζεται με αυτά τα ξενοδοχεία και ενδιαφέρεται να βρει ποιες είναι οι καλύτερες διευκολύνσεις που προσφέρουν, έτσι ώστε, να προσεγγίσει πελάτες που ενδιαφέρονται για τέτοιες διευκολύνσεις.

Στο συγκεκριμένο παράδειγμα η συνήθης περίπτωση είναι η πρώτη, δηλαδή τα αντίκειμενα ανάκτησης να είναι τα ξενοδοχεία. Ας θεωρηθεί λοιπόν ότι οι

ερωτήσεις αποσκοπούν στη ανάκτηση ξενοδοχείων και ότι, οι όροι που χρησιμοποιούνται για το σχηματισμό των ερωτήσεων είναι οι διευκολύνσεις. Έστω το εξής σενάριο:

1. Ο χρήστης προσδιορίζει μια αρχική ερώτηση ομοιότητας έτσι ώστε να ανακτήσει τα ξενοδοχεία που τον ενδιαφέρουν.
2. Η Μηχανή Ανάκτησης επεξεργάζεται την ερώτηση και επιστρέφει ένα σύνολο ξενοδοχείων που ικανοποιούν την ερώτηση συνοδευόμενα με τον βαθμό ομοιότητάς τους με την ερώτηση.
3. Ο χρήστης μελετά τα ξενοδοχεία και προσδιορίζει αυτά που πραγματικά τον ενδιαφέρουν. Το σύστημα έχει με αυτό τον τρόπο στη διάθεσή του ένα σύνολο ξενοδοχείων τα οποία, σύμφωνα με το χρήστη, είναι αυτά για τα οποία ενδιαφερόταν.
4. Το σύστημα μπορεί τώρα χρησιμοποιώντας αυτά τα ξενοδοχεία να κατασκευάσει μια "αντίστροφη" ερώτηση, έτσι ώστε, να ανακτήσει τις διευκολύνσεις αυτών των ξενοδοχείων και μάλιστα να ανακτήσει και το βαθμό της κάθε διευκόλυνσης. Μπορεί συνεπώς να διαλέξει τις δευκολύνσεις με το μεγαλύτερο βαθμό.
5. Με τον τρόπο αυτό το σύστημα κατάφερε να ανακαλύψει κάποιους όρους δεικτοδότησης - κάποιες διευκολύνσεις στην συγκεκριμένη περίπτωση - οι οποίοι χαρακτηρίζουν τα επιθυμητά για το χρήστη ξενοδοχεία. Μπορεί συνεπώς, χρησιμοποιώντας αυτούς τους όρους, να τροποποιήσει την αρχική ερώτηση κάνοντας χρήση οποιουδήποτε μοντέλου ανάδρασης επιθυμεί.

Η γενίκευση του παραπάνω σεναρίου προκύπτει εύκολα, το σύστημα επιτρέπει, έχοντας ανακτήσει ένα υποσύνολο αντικειμένων το οποίο ικανοποιεί τις απαιτήσεις της ερώτησης, να ανακαλυφθούν οι όροι δεικτοδότησης οι οποίοι περιγράφουν αυτά τα αντικείμενα καθώς και το βάρος τους. Αυτό γίνεται εκτελώντας "αντίστροφες" ερωτήσεις -οι οποίες μπορούν να ακολουθούν οποιοδήποτε από τα μοντέλα υπολογισμού ομοιότητας υποστηρίζεται- το αποτέλεσμα των οποίων είναι μια "αξιολόγηση" των όρων δεικτοδότησης των ανακτηθέντων αντικειμένων. Από αυτούς τους όρους δεικτοδότησης επιλέγονται οι "καλύτεροι" και χρησιμοποιούνται για να τροποποιηθεί η αρχική ερώτηση. Το συστημά λοιπόν, μπορεί να υποστηρίξει

και μηχανισμούς ανάδρασης σχετικότητας το βασικό σχήμα των οποίων περιγράφεται στο επόμενο τμήμα

3.4.1 Βασικό σχήμα για την υποστήριξη ανάδρασης σχετικότητας

Όπως αναφέρθηκε (τμήμα 2.4.2) ένα σύστημα ανάκτησης πληροφορίας ορίζεται ως η ακόλουθη τετράδα:

$$\langle T, Q, D, F \rangle$$

Στο προηγούμενο τμήμα αναφέρθηκε ότι η χρήση σχέσεων αξιολόγησης δίνει τη δυνατότητα να υποστηριχτούν ερωτήσεις τόσο στα αντικείμενα μιας συλλογής D όσο και στα στοιχεία T βάσει των οποίων περιγράφονται τα αντικείμενα, δηλαδή στους όρους δεικτοδότησης.

Πιο συγκεκριμένα, αν R είναι μια σχέση αξιολόγησης μεταξύ των αντικειμένων του D και των στοιχείων του T δηλαδή $R: D \times T \rightarrow [0,1]$, τότε μια ερώτηση που περιγράφεται σε κανονική συζευκτική μορφή με άτομα από το σύνολο T μπορεί να εκφραστεί με δυο σχέσεις με αξιολόγηση *Queries* και *Terms*. Η απάντηση στην ερώτηση αυτή δίνεται με βάση τη σχέση (3.7), δηλαδή:

$$Queries \circ_{M}^{AND} \left(Terms \circ_{M}^{OR} R \right)$$

Η σχέση που προκύπτει σαν αποτέλεσμα είναι μια σχέση ανάμεσα στο σύνολο Q των ερωτήσεων και το σύνολο D των αντικειμένων.

Για να γίνει κατανοητός ο μηχανισμός που προτείνεται για την υποστήριξη ανάδρασης σχετικότητας ας σημειωθεί ότι ο χαρακτηρισμός ενός αντικειμένου ως σχετικού ή μη σχετικού ως προς μια ερώτηση μπορεί να αποτυπωθεί ως μια νέα σχέση αξιολόγησης $F: Q \times D \rightarrow [0,1]$, όπου όσο πιο σχετικό είναι ένα αντικείμενο τόσο πιο κοντά στη μονάδα είναι η αξιολόγησή του, ενώ όσο πιο μη σχετικό είναι τόσο πιο κοντά στο μηδέν.

Το ζητούμενο τώρα είναι να προσδιοριστεί μια νέα ερώτηση, δηλαδή, καινούργιες σχέσεις *Queries'* και *Terms'* βάσει των αρχικών *Queries* και *Terms* που να λαμβάνει υπόψη τη σχέση ανάδρασης *F*. Ο μηχανισμός που χρησιμοποιείται για

να οριστούν οι *Queries'* και *Terms'* δεν είναι μονοσήμαντα ορισμένος. Για παράδειγμα μια πιθανή εκδοχή είναι η ακόλουθη:

$$Queries' = \left(F \circ_M^{AND} R^T\right) \circ_M^{OR} Terms^T$$

$$Terms' = \left(Queries^T \circ_M^{AND} F\right) \circ_M^{OR} R^T$$

Ας σημειωθεί ότι διαφορετικά σχήματα για υποστήριξη ανάδρασης σχετικότητας μπορούν να εφαρμοστούν με βάση:

- Διαφορετικά μοντέλα αξιολόγησης
- Διαφορετικούς τύπους υπολογισμού των σχέσεων *Queries'* και *Terms'*.

Μια απλή και αποτελεσματική προσέγγιση είναι η ακόλουθη. Θεωρούμε τη σχέση με αξιολόγηση $F' = F \circ_M^{AND} R^T \rightarrow [0,1]$. Η σχέση αυτή εκφράζει τη σχετικότητα των όρων δεικτοδότησης βάσει της σχετικότητας των αντικειμένων που εκφράζει η σχέση *F*. Με βάση τη σχέση αυτή προσδιορίζονται οι καινούργιες σχέσεις *Queries'* και *Terms'* ως εξής:

$$Queries' = F' \circ_M^{OR} Terms^T$$

$$Terms' = Q^T \circ_M^{OR} F'$$

Με τελείως ανάλογο τρόπο ορίζεται ένα παραπλήσιο σχήμα με :

$$F' = F \circ_M^{OR} R^T \rightarrow [0,1]$$

Μια επιπλέον δυνατότητα είναι αφού υπολογιστεί η σχέση *F'* να δοθεί στο χρήστη η δυνατότητα να επιλέξει τα στοιχεία που θεωρεί περισσότερο σχετικά ώστε να μειωθεί το πλήθος των στοιχείων που περιέχουν οι όροι της ερώτησης και να μειωθεί το πλήθος των αντικειμένων που την ικανοποιούν.

3.5 Σύνοψη

Στο κεφάλαιο αυτό περιγράφηκε το θεωρητικό μοντέλο για την εκτέλεση ερωτήσεων ομοιότητας. Αρχικά δόθηκαν οι αναγκαίοι μαθηματικοί ορισμοι της έννοιας της σχέσης και της σχέσης με αξιολόγηση.

Ακολούθως ορίστηκε η πράξη της σύνθεσης σχέσεων και της σύνθεσης σχέσεων αξιολόγησης. Επίσης περιγράφηκε και ο τρόπος με τον οποίο το μοντέλο μπορεί να υποστηρίξει και μηχανισμούς ανάδρασης σχετικότητας.

Στο επόμενο κεφάλαιο θα παρουσιαστεί η μεθοδολογία με βάση την οποία εκτελούνται ερωτήσεις ομοιότητας. Η μεθοδολογία αυτή έγκειται στον μετασχηματισμό των αλγεβρικών μεθόδων που παρουσιάστηκαν σε αυτό το κεφάλαιο σε μεθόδους που θα επιτρέπουν την εκτέλεση ερωτήσεων ομοιότητας πάνω σε ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων.

4. Μεθοδολογία εκτέλεσης ερωτήσεων ομοιότητας

4.1 Εισαγωγή

Μετά την παρουσίαση του μαθηματικού μοντέλου για την εκτέλεση ερωτήσεων ομοιότητας, στο προηγούμενο κεφάλαιο, θα περιγραφεί η μεθοδολογία που αναπτύχθηκε και υλοποίηθηκε για την εκτέλεση ερωτήσεων ομοιότητας πάνω σε μια σχεσιακή βάση δεδομένων. Με την μεθοδολογία αυτή πραγματοποιείται η μετάβαση από τον αλγεβρικό τρόπο υπολογισμού ερωτήσεων ομοιότητας, ο οποίος περιγράφηκε στο προηγούμενο κεφάλαιο, στον υπολογισμό των ερωτήσεων ομοιότητας με χρήση της γλώσσας *SQL* πάνω σε ένα σχεσιακό σύστημα.

Η υποστήριξη ερωτήσεων ομοιότητας πάνω σε σχεσιακές βάσεις δεδομένων αποτελεί το βασικό στόχο αυτής της εργασίας. Οι ερωτήσεις ομοιότητας όμως, απαντώνται μόνο σε συστήματα ανάκτησης πληροφορίας, ενώ αντίθετα, στα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων υπάρχει η δυνατότητα εκτέλεσης μόνο ερωτήσεων δίτιμης λογικής (boolean queries). Η προσέγγιση που ακολουθείται στα πλαίσια αυτής της εργασίας, και θα παρουσιαστεί σε αυτό το κεφάλαιο, είναι η υλοποίηση ενός υψηλότερου επιπέδου μοντέλου, πάνω από το σχεσιακό μοντέλο. Το μοντέλο αυτό επιτρέπει τον προσδιορισμό επιπλέον σημασιολογικής πληροφορίας πάνω από το σχήμα μιας σχεσιακής βάσης δεδομένων

ορίζοντας νέα χαρακτηριστικά τα οποία θα υπολογίζονται με βάση τις τιμές κάποιων ήδη υπαρχόντων χαρακτηριστικών. Σε δεύτερη φάση παρέχονται οι μηχανισμοί μετάφρασης από το μοντέλο υψηλότερου επιπέδου στο σχεσιακό μοντέλο και σε *SQL* εκφράσεις.

Στο επόμενο τμήμα θα παρουσιαστούν οι βασικές έννοιες του μοντέλου ασαφών οντοτήτων-συσχετίσεων το οποίο βρίσκεται πάνω από το σχεσιακό μοντέλο. Στο τμήμα 4.3 θα δοθεί η υλοποίηση σχέσεων αξιολόγησης καθώς της σύνθεσης σχέσεων αξιολόγησης πάνω σε σχεσιακή βάση δεδομένων. Η υλοποίηση αυτή αποτελεί το υλοποιηστικό ανάλογο, σε *SQL*, των αλγεβρικών μηχανισμών που παρουσιάστηκαν στο προηγούμενο κεφάλαιο ενώ, στο τμήμα 4.4 θα παρουσιαστεί ο τρόπος αναπαράστασης των ερωτήσεων που δέχεται το σύστημα σε σχεσιακή βάση δεδομένων. Το τμήμα 4.5 περιγράφει τη διαδικασία εκτέλεσης ερωτήσεων ομοιότητας. Στο τμήμα 4.6 θα περιγραφεί ο τρόπος διαμόρφωσης των *SQL* ερωτήσεων με βάση το μοντέλου υπολογισμού ομοιότητας, ενώ στο επόμενο τμήμα θα δοθεί η μεθοδολογία υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων. Τέλος στο τμήμα 4.8 θα δοθεί μια περίληψη του κεφαλαίου.

4.2 Το μοντέλο ασαφών οντοτήτων-συσχετίσεων (fuzzy entity-relationship model)

Οι ερωτήσεις ομοιότητας, όπως προαναφέρθηκε, απαντώνται σε συστήματα ανάκτησης πληροφορίας. Προκειμένου να καταστεί δυνατή η εκτέλεση ερωτήσεων ομοιότητας πάνω σε μια σχεσιακή βάση δεδομένων, ορίζεται ένα μοντέλο υψηλότερου επιπέδου, πάνω από το σχεσιακό μοντέλο.

Στα συστήματα ανάκτησης πληροφοριών, υπάρχει καταρχάς ένα σύνολο αντικειμένων τα οποία και θα αποτελούν τη ***συλλογή (collection)*** του συστήματος ανάκτησης πληροφορίας, δηλαδή το σύνολο των αντικειμένων τα οποία θα ανακτώνται με τις ερωτήσεις ομοιότητας. Επιπλέον, υπάρχει ένα σύνολο όρων οι οποίοι από τη μια συμβάλλουν στην αναπαράσταση των αντικειμένων της συλλογής και από την άλλη στο σχηματισμό των ερωτήσεων ομοιότητας. Οι όροι αυτοί καλούνται ***όροι δεικτοδότησης (index terms)***. Στην προσέγγιση που ακολουθείται στην εργασία αυτή, οι συλλογές αντικειμένων και τα σύνολα όρων δεικτοδότησης μοντελοποιούνται με την έννοια του ***συνόλου οντοτήτων (entity set)*** και η μεταξύ τους συσχέτιση με την έννοια του ***συνόλου συσχετίσεων (relationship set)***. Οι έννοιες

αυτές είναι δανεισμένες από το δημοφιλές μοντέλο οντοτήτων συσχετίσεων (entity relationship model) και ο λόγος που οδήγησε σε αυτή την προσέγγιση είναι ακριβώς η ευρεία αποδοχή του συγκεκριμένου μοντέλου, η ευκολία στην κατανόηση των εννοιών αυτών και το γεγονός ότι βρίσκεται σε υψηλότερο επίπεδο από το σχεσιακό μοντέλο.

Φυσικά, οι έννοιες αυτές έχουν επεκταθεί κατάλληλα για να καλύψουν τις ανάγκες έκφρασης και εκτέλεσης ερωτήσεων ομοιότητας. Έτσι, η έννοια του συνόλου οντοτήτων το οποίο έχει (στο παραδοσιακό μοντέλο οντοτήτων συσχετίσεων) ένα σύνολο από χαρακτηριστικά, εδώ έχει ένα σύνολο από ***ασαφή χαρακτηριστικά (fuzzy attributes)*** τα οποία ισοδυναμούν με ιδιότητες των οντοτήτων, οι οποίες για κάθε οντότητα έχουν μια τιμή σπουδαιότητας (ή σημασίας) από το 0 (μηδενική σπουδαιότητα) μέχρι το 1 (μέγιστη σπουδαιότητα). Για παράδειγμα, ένα ξενοδοχείο με βάση την απόστασή του από το κέντρο της πόλης, μπορεί να χαρακτηριστεί λιγότερο ή περισσότερο απομονωμένο. Δηλαδή ορίζεται ένα ασαφές χαρακτηριστικό "*κεντρικό ξενοδοχείο*" το οποίο υπολογίζεται με βάση την απόσταση κάθε ξενοδοχείου από το κέντρο της πόλης και παίρνει την τιμή 1 για τα ξενοδοχεία που βρίσκονται στο κέντρο και όσο η απόσταση από το κέντρο μεγαλώνει η σπουδαιότητα ως προς αυτό το ασαφές χαρακτηριστικό φθίνει και τείνει στο μηδέν.

Πέρα από την επέκταση της έννοιας του συνόλου οντοτήτων για να περιλάβει ασαφή χαρακτηριστικά, επεκτείνεται και η έννοια του συνόλου συσχετίσεων για να καταστεί δυνατή η αποτίμηση κάθε συσχέτισης με τιμές στο διάστημα [0,1] και όχι μόνο 0 ή 1 όπως συμβαίνει με συσχετίσεις που ακολουθούν το μοντέλο της δίτιμης λογικής. Για παράδειγμα η συσχέτιση ανάμεσα σε ένα ξενοδοχείο και μια διευκόλυνση που παρέχει συνδέεται με μια τιμή $w \in [0,1]$ που εκφράζει τη σπουδαιότητα της διευκόλυνσης στο ξενοδοχείο, δηλαδή, πόσο καλά παρέχεται η διευκόλυνση αυτή στο συγκεκριμένο ξενοδοχείο.

Το μεγάλο πλεονέκτημα της προσέγγισης αυτής που ακολουθείται είναι ότι ένα σύνολο οντοτήτων μπορεί να λειτουργήσει είτε ως συλλογή αντικειμένων προς ανάκτηση[1] είτε ως ένα σύνολο όρων δεικτοδότησης[2]. Το ίδιο μπορεί να συμβεί με τα

[1] Όταν κάποιο άλλο σύνολο οντοτήτων που συνδέεται με το πρώτο μέσω ενός συνόλου συσχετίσεων λειτουργεί ως σύνολο όρων δεικτοδότησης.

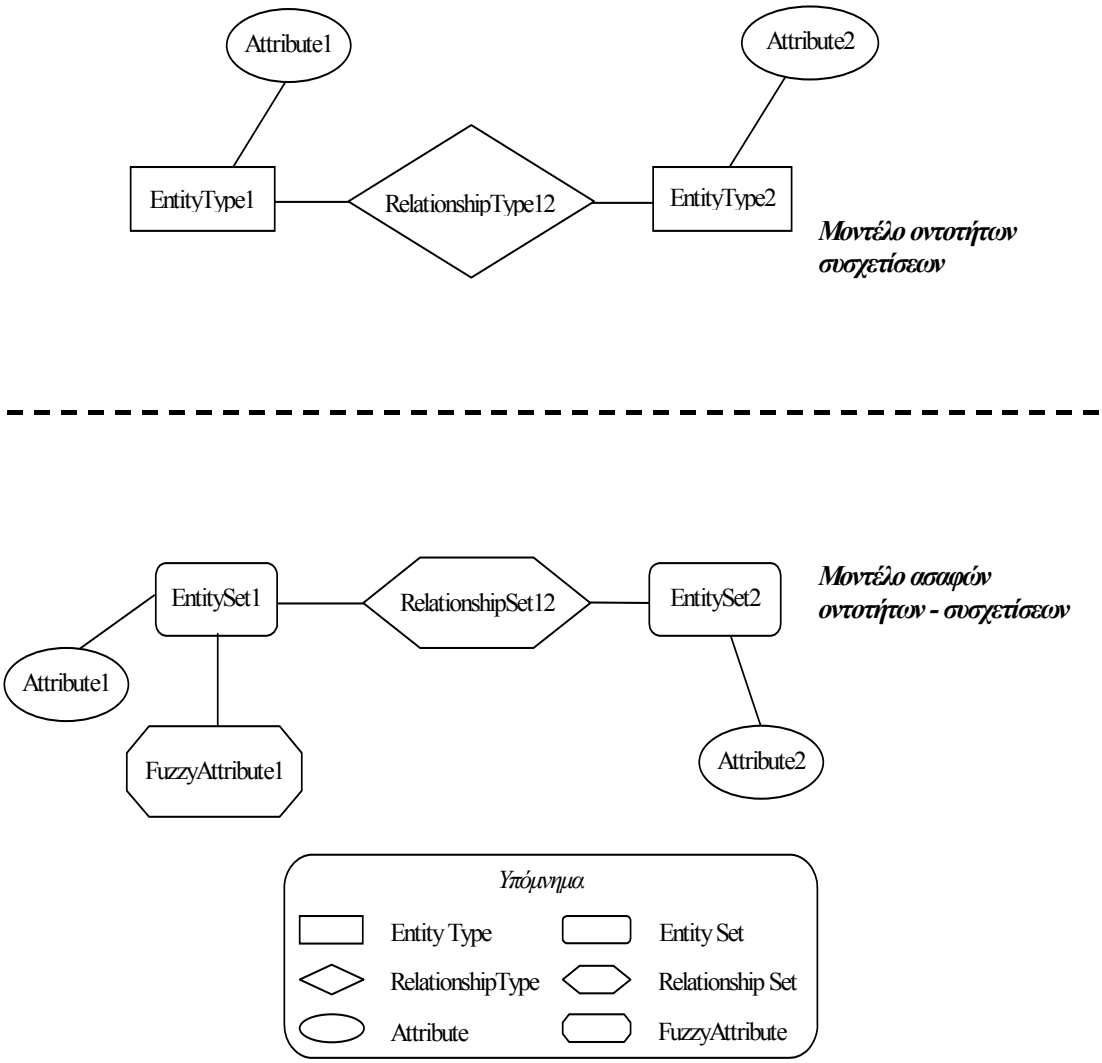
[2] όταν κάποιο άλλο σύνολο οντοτήτων που συνδέεται με το πρώτο μέσω ενός συνόλου συσχετίσεων λειτουργεί ως συλλογή αντικειμένων.

ασαφή χαρακτηριστικά κάθε συνόλου οντοτήτων τα οποία μπορούν να χρησιμοποιηθούν είτε ως όροι δεικτοδότησης για την ανάκτηση οντοτήτων του αντίστοιχου συνόλου οντοτήτων είτε και αντίστροφα ως αντικείμενα ανάκτησης αν οι οντότητες θεωρηθούν ως όροι δεικτοδότησης.

Στο παράδειγμα των ξενοδοχείων, το σύνολο των ξενοδοχείων και το σύνολο των διευκολύνσεων μπορούν να θεωρηθούν ως σύνολα οντοτήτων και η μεταξύ τους σχέση ως ένα σύνολο συσχετίσεων. Σε αυτή την περίπτωση αν τα προς ανάκτηση αντικείμενα είναι τα ξενοδοχεία, τότε οι όροι δεικτοδότησης και συνεπώς οι όροι που σχηματίζουν τις ερωτήσεις ομοιότητας θα είναι κάποιες διευκολύνσεις. Αν τα ξενοδοχεία είναι όροι δεικτοδότησης που σχηματίζουν ερωτήσεις τότε τα αντικείμενα ανάκτησης θα είναι οι διευκολύνσεις.

Το μοντέλο ασαφών οντοτήτων-συσχετίσεων χρησιμοποιείται για την περιγραφή του σχήματος του συστήματος ανάκτησης πληροφορίας που κατασκευάζεται. Συνεπώς το μοντέλο ασαφών οντοτήτων-συσχετίσεων είναι ορατό στο χρήστη-διαχειριστή που ορίζει το σχήμα του συστήματος ανάκτησης πληροφορίας.

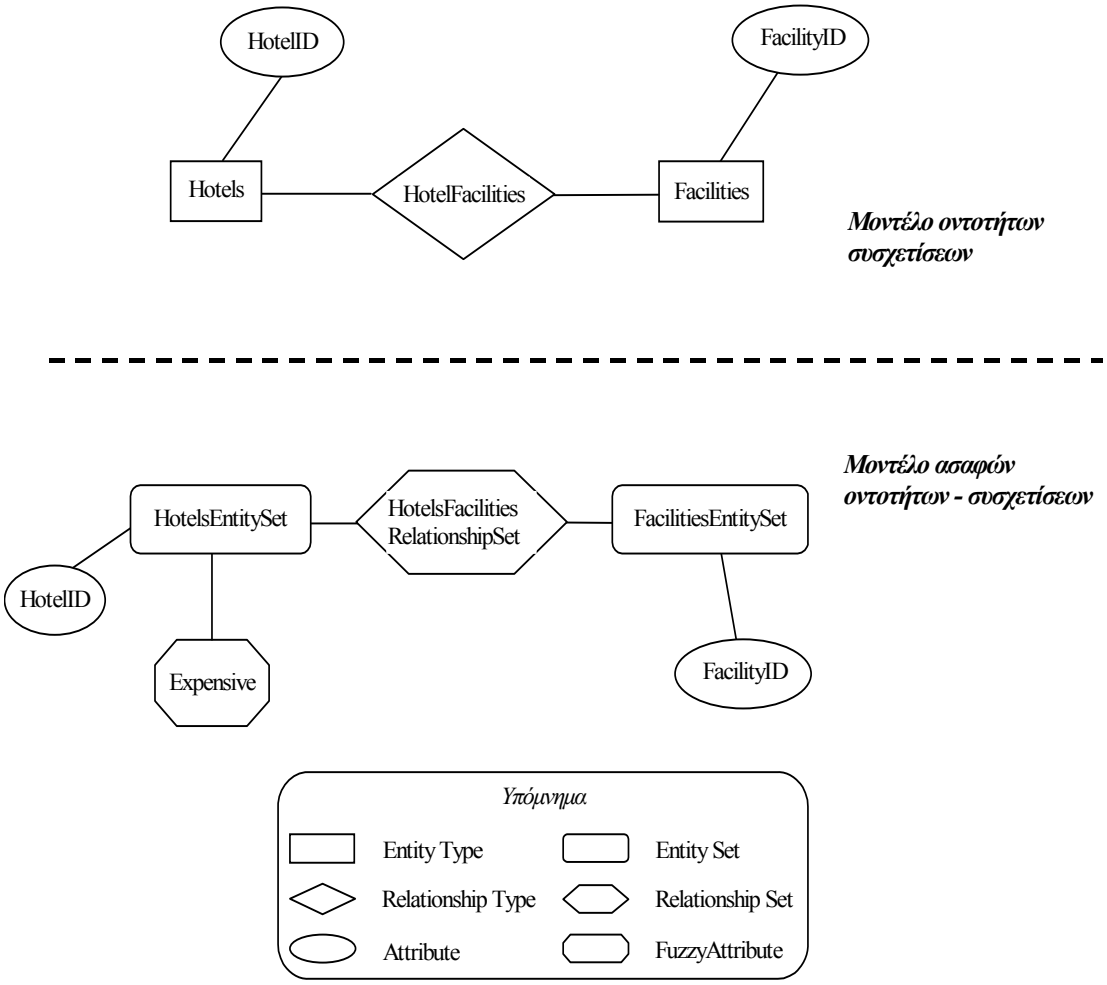
Μια γραφική απεικόνιση των παραπάνω παρουσιάζεται στο επόμενο σχήμα:



Σχήμα 4-1:Γραφική αναπαράσταση της αντιστοίχισης μεταξύ του μοντέλου ασαφών οντοτήτων-συσχετίσεων και του κλασσικού μοντέλου οντοτήτων συσχετίσεων. Ένας τύπος οντότητας μπορεί να αντιστοιχηθεί με ένα σύνολο οντοτήτων ενώ ένας τύπος συσχέτισης μεταξύ δύο τύπων οντοτήτων μπορεί να εκφραστεί μέσω ενός συνόλου συσχετίσεων. Σε ένα σύνολο οντοτήτων μπορούν να οριστούν ασαφή χαρακτηριστικά τα οποία υπολογίζονται βάση ήδη υπαρχόντων χαρακτηριστικών

Στο παράδειγμα των ξενοδοχείων ορίζεται το σύνολο οντοτήτων των ξενοδοχείων, για παράδειγμα *HotelsEntitySet* πάνω στην οντότητα *Hotels*, όπως και το σύνολο οντοτήτων των διευκολύνσεων, *FacilitiesEntitySet* πάνω στην οντότητα

Facilities. Επίσης ορίζεται και το σύνολο συσχετίσεων *HotelFacilitiesRelationshipSet* το οποίο εκφράζει τη σύνδεση μεταξύ των μελών των δύο συνόλων οντοτήτων. Καθίσταται επίσης δυνατός ο ορισμός νέων ασαφών χαρακτηριστικών όπως το χαρακτηριστικό *Expensive* το οποίο εκφράζει το πόσο ακριβό είναι ένα ξενοδοχείο. Η γραφική αναπαράσταση των οριζόμενων συνόλων οντοτήτων, συσχετίσεων και χαρακτηριστικών έχει όπως φαίνεται στο επόμενο σχήμα.



Σχήμα 4-2: Γραφική απεικόνιση των συνόλων οντοτήτων και συσχετίσεων καθώς και των ασαφών χαρακτηριστικών στο παράδειγμα των ξενοδοχείων.

Στα επόμενα τμήματα θα παρουσιαστούν αναλυτικότερα οι προαναφερθείσες έννοιες.

4.2.1 Η έννοια του συνόλου οντοτήτων (entity set)

Όπως προαναφέρθηκε στα συστήματα ανάκτησης πληροφορίας υπάρχει ένα σύνολο αντικειμένων το οποία αποτελούν τη ***συλλογή (collection)*** του συστήματος. Οι υποβαλλόμενες ερωτήσεις αποσκοπούν στην ανάκτηση αυτών των αντικειμένων. Επιπλέον υπάρχει ένα σύνολο όρων οι οποία από τη μια συμβάλλουν στην αναπαράσταση των αντικειμένων της συλλογής και από την άλλη στο σχηματισμό των υποβαλλόμενων ερωτήσεων. Οι όροι αυτοί ονομάζονται ***όροι δεικτοδότησης (index terms)***.

Για υποστηριχτούν ερωτήσεις ομοιότητας είναι αναγκαίο να καθοριστούν τόσο τα προς ανάκτηση αντικείμενα όσο και οι όροι δεικτοδότησης, δηλαδή να οριστεί το σχήμα του συστήματος ανάκτησης πληροφορίας. Αυτή ακριβώς την αναγκή ικανοποιεί η έννοια του συνόλου οντοτήτων. ***Ένα σύνολο οντοτήτων (entity set) αναπαριστά είτε ένα σύνολο αντικειμένων της συλλογής είτε ένα σύνολο όρων δεικτοδότης.***

Δεδομένου ότι η τεχνολογική βάση στην οποία στηρίζεται το σύστημα είναι τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων η έννοια του συνόλου οντοτήτων αποτελεί επέκταση της έννοιας της οντότητας στο γνωστό μοντέλο οντοτήτων συσχετίσεων το οποίο βρίσκεται πάνω από το σχεσιακό μοντέλο.

Επιπλέον η έννοια του συνόλου οντοτήτων παρέχει τη δυνατότητα ορισμού νέων ασαφών χαρακτηριστικών τα οποία βασίζονται πάνω σε ήδη υπάρχοντα χαρακτηριστικά και ισοδυναμούν με ιδιότητες των οντοτήτων. Τα χαρακτηριστικά αυτά θα παρουσιαστούν στο επόμενο τμήμα.

4.2.2 Η έννοια του ασαφούς χαρακτηριστικού (fuzzy attribute)

Τα ***ασαφή χαρακτηριστικά (fuzzy attributes)*** όπως αναφέρθηκε ***ισοδυναμούν με ιδιότητες των οντοτήτων οι οποίες για κάθε οντότητα έχουν μια τιμή σπουδαιότητας (ή σημασίας) από το 0 (μηδενική σπουδαιότητα) μέχρι το 1 (μέγιστη σπουδαιότητα)***. Τα ασαφή χαρακτηριστικά υπολογίζονται με βάση κάποια πεδία του πίνακα πάνω στον οποίο ορίζεται ένα σύνολο οντοτήτων και με μια συνάρτηση αξιολόγησης.

Συνεπώς ***ενα χαρακτηριστικό ορίζεται ως η ακόλουθη τριάδα:***

$$\langle \textbf{\textit{EntitySet, EvaluationExpression, RankingFunction}} \rangle$$

όπου:

- *EntitySet* είναι το σύνολο οντοτήτων πάνω στο οποίο ορίζεται το ασαφές χαρακτηριστικό.
- *EvaluationExpression* είναι η έκφραση η οποία περιγράφει τον υπολογισμό του με βάση χαρακτηριστικά του σύνολου οντοτήτων *EntitySet*.
- *RankingFunction* είναι η συνάρτηση αξιολόγησης που χρησιμοποιείται γι' αυτό το χαρακτηριστικό. Η συνάρτηση αυτή προσδιορίζει μια αντιστοίχιση, της τιμής που προκύπτει από την έκφραση υπολογισμού *EvaluationExpression* σε μια τιμή στο κλειστό διάστημα $[0,1]$.

Η έκφραση υπολογισμού (EvaluationExpression) του χαρακτηριστικού συγκροτείται από τους μαθηματικούς τελεστές που υποστηρίζει η γλώσσα *SQL* (+,-,*,/…) και από κάποια χαρακτηριστικά του συνόλου οντοτήτων (EntitySet) στο οποίο βασίζεται το χαρακτηριστικού. Είναι προφανές ότι, ο τύπος αυτών των χαρακτηριστικών πρέπει να είναι αριθμητικός και γενικά συμβατός με τους χρησιμοποιούμενους αριθμητικούς τελεστές, έτσι ώστε, να προκύπτει μια έγκυρη μαθηματική έκφραση. Μπορεί κατ' αυτό τον τρόπο να συνδυαστούν δύο ή περισσότερα χαρακτηριστικά και να δημιουργηθεί ένα νέο υπολογιζόμενο ασαφές χαρακτηριστικό.

Από τη στιγμή που θα οριστεί κάποιο χαρακτηριστικό πάνω σε ένα σύνολο οντοτήτων κάθε μέλος του συνόλου έχει μια τιμή ως προς αυτό το χαρακτηριστικό.

4.2.2.1 Ειδικές συναρτήσεις μετασχηματισμού χαρακτηριστικού

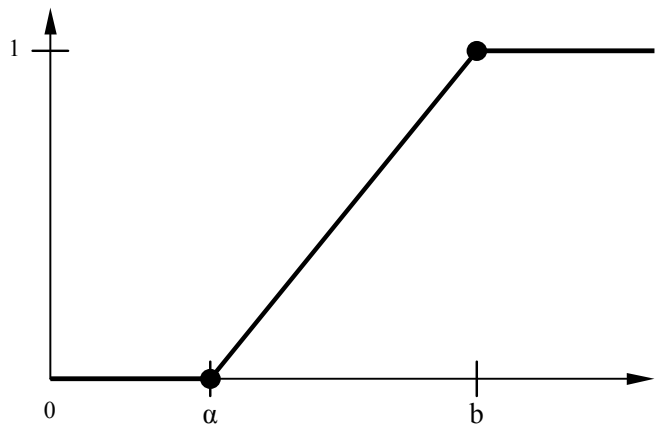
Οι συναρτήσεις μετασχηματισμού που χρησιμοποιούνται στο σύστημα είναι γνωστές συναρτήσεις από τη θεωρία των ασαφών συνόλων και δίνοται στη συνέχεια.

Γραμμική

Η απλή γραμμική συνάρτηση, έστω $\mu_{a,b}$, σε ένα διάστημα $[a,b]$ αυξάνει γραμμικά από το 0 στο 1, ενώ για τιμές μικρότερες του *α* είναι 0 και για τιμές μεγαλύτερες του *b* είναι 1:

$$\mu_{a,b} = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

Η γραφική αναπαράσταση αυτής της συνάρτησης δίνεται στο επόμενο σχήμα:



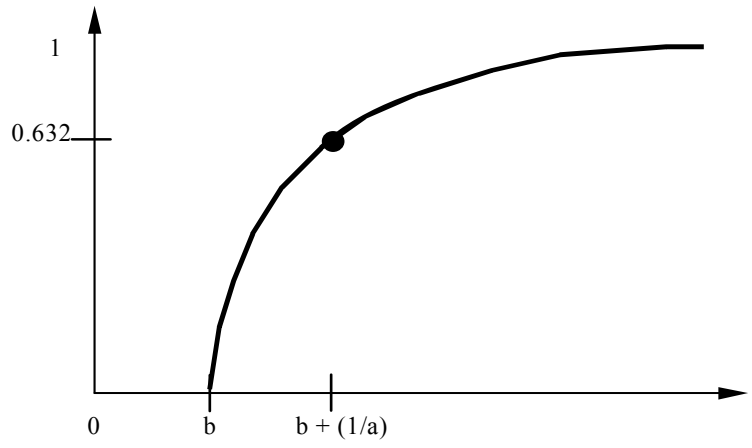
Γραφική απεικόνιση απλής γραμμικής συνάρτησης μετασχηματισμού

Εκθετική

Στην περίπτωση που η μετάβαση από το 0 στο 1 δεν είναι επιθυμητό να γίνει γραμμικά αλλά εκθετικά, ορίζεται η παρακάτω συνάρτηση:

$$\mu_{a,b} = \begin{cases} 0, & x < b \\ 1 - e^{(-a(x-b))}, & x \ge b \end{cases}$$

Είναι φανερό ότι η συνάρτηση αυτή τείνει στο 1 καθώς το x αυξάνει ενώ ο ρυθμός με τον οποίο τείνει στο 1 καθορίζεται από την παράμετρο *a*. Για τιμές μικρότερες του *b* είναι 0. Στο επόμενο σχήμα δίνεται η γραφική αναπαράσταση της συνάρτησης αυτής:



Γραφική απεικόνιση εκθετικής συνάρτησης μετασχηματισμού

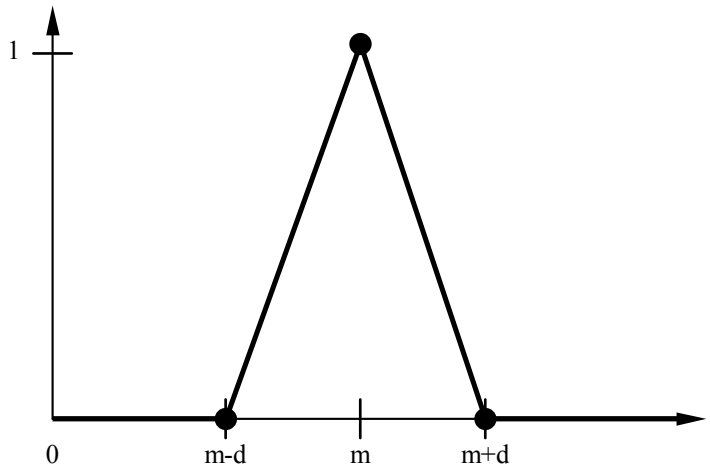
Η εκθετική συνάρτηση, θεωρητικά, ποτέ δεν θα δώσει την τιμή 1 καθώς τείνει μεν προς τη μονάδα, χωρίς όμως ποτέ να γίνεται μονάδα. Η πρακτική μετάφραση αυτής της ιδιότητας είναι ότι το γνώρισμα αξιολόγησης δεν είναι ποτέ απόλυτα βέβαιο ότι ικανοποιείται.

Τριγωνική

Η συνάρτηση αυτή χρησιμοποιείται όταν θέλουμε να δηλώσουμε ότι κάποιο γνώρισμα αξιολογείται με 1 για κάποια συγκεκριμένη τιμή, ενώ καθώς απομακρυνόμαστε από αυτή την τιμή η αξιολόγηση ελαττώνεται γραμμικά και μηδενίζεται μετά από ένα ορισμένο σημείο. Η συνάρτηση αυτή θα μπορούσε να χρησιμοποιηθεί από ένα χαρακτηριστικό το οποίο θα περιγραφόταν ως "…περίπου ίσο με *m*…" και ορίζεται ως:

$$\mu_{m,d} = \begin{cases} 0, & x < m - d \\ 1 - \left|\dfrac{m-x}{d}\right|, & m - d \le x \le m + d \\ 0, & x > m + d \end{cases}$$

Η γραφική παράσταση της συνάρτησης αυτής έχει ως ακολούθως:



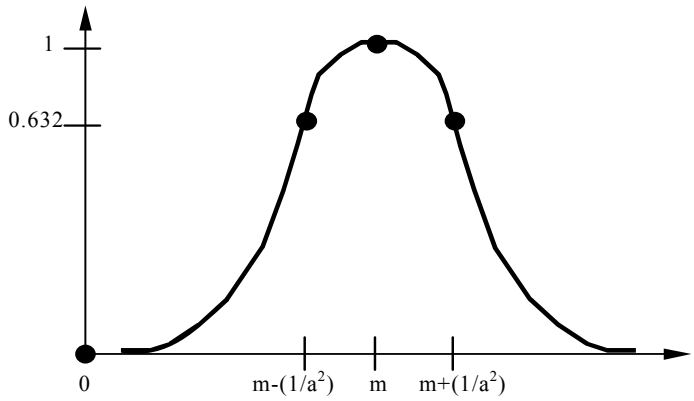
Γραφική απεικόνιση τριγωνικής συνάρτησης μετασχηματισμού

Gaussian

Αν το χαρακτηριστικό είναι του τύπου που περιγράφηκε στην τριγωνική συνάρτηση, αλλά θέλουμε η συνάρτηση αξιολόγησης να έχει μια περισσότερο ομαλή συμπεριφορά μπορούμε να χρησιμοποιήσουμε μια συνάρτηση παρόμοια με την συνάρτηση πυκνότητας της κατανομής Gauss όπως περιγράφεται από την επόμενη εξίσωση:

$$\mu_{a,m} = e^{\left(-a(x-m)^2\right)}$$

όπου m είναι η μέση τιμή, δηλαδή η τιμή γύρω από την οποία το χαρακτηριστικό αξιολογείται με 1 και a είναι η διασπορά, η οποία ουσιαστικά προσδιορίζει το ρυθμό μείωσης της τιμής της συνάρτησης καθώς απομακρυνόμαστε από την τιμή m. Στο επόμενο σχήμα δίνεται η γραφική παράσταση της συνάρτησης αυτής.



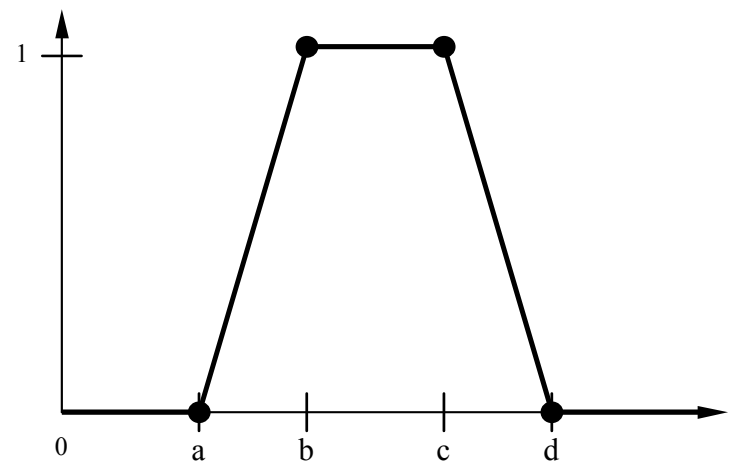
Γραφικη απεικόνιση Gauss συνάρτησης ματασχηματισμού

Τραπεζοειδής

Η συνάρτηση αυτή χρησιμοποιείται για να περιγράψει χαρακτηριστικά του τύπου "κατά προσέγγιση ανάμεσα στο *b* και *c*". Η τραπεζοειδής συνάρτηση ορίζεται ως εξής:

$$\mu_{a,b,c,d} = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a}, & a \le x < b \\ 1, & b \le x < c \\ \dfrac{x-d}{c-d}, & c \le x < d \\ 0 & d \le x \end{cases}$$

Η γραφική αναπαράσταση της τραπεζοειδούς συνάρτησης δίνεται στο επόμενο σχήμα:



Γραφική απεικόνιση τραπεζοειδούς συνάρτησης μετασχηματισμού

4.2.3 Η έννοια του συνόλου συσχετίσεων (relationship set)

Η έννοια του συνόλου συσχετίσεων ***εκφράζει την αντιστοίχιση μεταξύ των αντικειμένων μιας συλλογής - δηλαδή των αντικειμένων τα οποία θα ανακτώνται με τις ερωτήσεις ομοιότητας - και των όρων δεικτοδότησης, δηλαδή των όρων οι οποίοι αναπαριστούν τα αντικείμενα της συλλογής και χρησιμοποιούνται για το σχηματισμό των ερωτήσεων***. Η έννοια του συνόλου συσχετίσεων αποτελεί επέκταση της έννοιας της συσχέτισης στο κλασσικό μοντέλο οντοτήτων συσχετίσεων. Η επέκταση αυτή επιτρέπει την αποτίμηση κάθε συσχέτισης με τιμές στο διάστημα [0,1]

και όχι μόνο με τις τιμές 0 ή 1 όπως συμβαίνει με συσχετίσεις που ακολουθούν το μοντέλο δίτιμης λογικής.

Στο κλασσικό μοντέλο οντοτήτων συσχετίσεων η συσχέτιση μεταξύ δύο οντοτήτων εκφράζεται με 0 ή 1, όπου το 0 δηλώνει την απουσία και το 1 την ύπαρξη συσχέτισης. Στο επεκταμένο μοντέλο οντοτήτων συσχετίσεων η συσχέτιση μεταξύ δύο συνόλων οντοτήτων αξιολογείται με μια τιμή σπουδαιότητας (ή σημασίας) στο διάστημα [0,1] όπου το 0 εκφράζει μηδενική σπουδαιότητα και το 1 μέγιστη σπουδαιότητα.

4.2.4 Η έννοια της έκφρασης αξιολόγησης (ranking expression)

Η έννοια της έκφρασης αξιολόγησης παρέχει επιπλέον δυνατότητες στον καθορισμό συνόλων συσχετίσεων. Έρχεται να αντιμετωπίσει προβλήματα τα οποία είναι δυνατόν να υπάρξουν κατά τον ορισμό του σχήματος του συστήματος ανάκτησης πληροφορίας, εξαιτίας του σχήματος της βάσης δεδομένων, αλλά και να προσδώσει στο σύστημα επιπλέον λειτουργικότητα. Αλλά ας δούμε τι ακριβώς αναπαριστά μια έκφραση αξιολόγησης.

Ένα σύνολο συσχετίσεων εκφράζει τον τρόπο με τον οποίο συνδέονται τα μέλη δύο συνόλων οντοτήτων καθώς και το βάρος σπουδαιότητας κάθε μέλους του ενός συνόλου ως προς το αντίστοιχο μέλος του άλλου συνόλου. Το βάρος αυτό όμως στη γενική περίπτωση, και ανάλογα με το τι εκφράζει πρακτικά κάθε φορά, μπορεί να παίρνει οποιαδήποτε τιμή. Με άλλα λόγια το πεδίο ορισμού του πεδίου το οποίο αναπαριστά το βάρος είναι το σύνολο των πραγματικών αριθμών.

Στο πεδίο αυτό μπορεί να προσδωθεί η έννοια της ασάφιας με την αντιστοίχιση των τιμών του στο διάστημα [0,1] έτσι ώστε η τιμή 0 να αντιστοιχεί σε μηδενική σπουδαιότητα και η τιμή 1 στην μέγιστη σπουδαιότητα. Υπενθυμίζεται στο σημείο αυτό ότι κάτι παρόμοιο υπήρχε και στην περίπτωση των χαρακτηριστικών (τμήμα 4.2.2).

Προκειμένου να επιτευχθεί η μετατροπή του πεδίο που αναπαριστά το βάρος ενός όρου δεικτοδότησης, σε ασαφές (fuzzy) χρησιμοποιούνται (όπως ακριβώς και στην περίπτωση των χαρακτηριστικών) ειδικές συναρτήσεις μετασχηματισμού. Οι συναρτήσεις μετασχηματισμού που χρησιμοποιούνται στις εκφράσεις αξιολόγησης

είναι ακριβώς οι ίδιες με αυτές που χρησιμοποιήθηκαν στα χαρακτηριστικά (τμήμα 4.2.2.1).

Αλλά ας δοθεί ο ακριβής ορισμός μιας έκφρασης αξιολόγησης. ***Μια έκφραση αξιολόγησης ορίζεται σαν μια τριάδα***:

$$\langle \textbf{\textit{RelationshipSet, EvaluationExpression, RankingFunction}} \rangle$$

όπου:

- *RelationshipSet* είναι το σύνολο συσχετίσεων πάνω στο οποίο ορίζεται η έκφραση αξιολόγησης.
- *EvaluationExpression* είναι η έκφραση η οποία περιγράφει τον υπολογισμό της έκφρασης αξιολόγησης. Η έκφραση αυτή είναι μια έγκυρη μαθηματική έκφραση η οποία εμπεριέχει το πεδίο που αναπαριστά το βάρος (Weight).
- *RankingFunction* είναι η συνάρτηση αξιολόγησης που χρησιμοποιείται γι' αυτή την έκφραση αξιολόγησης. Η συνάρτηση αυτή προσδιορίζει κατά κάποιο τρόπο μια αντιστοίχιση, της τιμής του βάρους με μια τιμή στο κλειστό διάστημα $[0,1]$.

Στο παράδειγμα των ξενοδοχείων και πάλι έχει οριστεί το σύνολο οντοτήτων *HotelsEntitySet* πάνω στην οντότητα *Hotels* το σύνολο *FacilitiesEntitySet* πάνω στην οντότητα *Facilities* καθώς και το σύνολο συσχετίσεων *HotelsFacilitiesRelationshipSet* πάνω στη συσχέτιση *HotelFacilitiesRelationshipSet* το οποία εκφράζει τη συσχέτιση μεταξύ των συνόλων οντοτήτων *HotelsEntitySet* και *FacilitiesEntitySet*.

Ας θεωρηθεί ότι το πεδίο ορισμού του πεδίου *Weight* της συσχέτισης *HotelFacilities* είναι το κλειστό διάστημα $[1,100]$. Το πεδίο αυτό περιγράφει στην πραγματικότητα πόσο καλά παρέχεται η κάθε διευκόλυνση σε κάθε ξενοδοχείο.
Μπορεί να οριστεί η έκφραση αξιολόγησης *HotelFacilitiesRankingExpression* πάνω στο σύνολο συσχετίσεων *HotelsFacilitiesRelationshipSet.* Η έκφραση υπολογισμού της θα είναι τότε η εξής:

EvaluationExpression : Weight / 100

ενώ ως συνάρτηση αξιολόγησης θα μπορούσε να θεωρηθεί η απλή γραμμική (τμήμα 4.2.2.1) με παραμέτρους για παράδειγμα α=0.375 και b=0.8.

4.2.5 Συσχετίσεις μεταξύ του μοντέλου ασαφών οντοτήτων-συσχετίσεων και του μαθηματικού μοντέλου εκτέλεσης ερωτήσεων ομοιότητας

Στο σημείο αυτό θα παρουσιαστούν οι συσχετίσεις μεταξύ του μοντέλου ασαφών οντοτήτων-συσχετίσεων και του μαθηματικού μοντέλου εκτέλεσης ερωτήσεων ομοιότητας που παρουσιάστηκε στο τρίτο κεφάλαιο. Οι αντιστοιχήσεις αυτές είναι απολύτως αναγκαίες προκειμένου να μεταβούμε από την αλγεβρική μεθοδολογία εκτέλεσης ερωτήσεων ομοιότητας στη μεθοδολογία εκτέλεσης ερωτήσεων ομοιότητας στα πλαίσια του μοντέλου ασαφών οντοτήτων-συσχετίσεων.

Η έννοια του συνόλου οντοτήτων (entity set) αποτελεί τη μεταφορά της κλασικής μαθηματικής έννοιας του συνόλου στο μοντέλο ασαφών οντοτήτων συσχετίσεων(fuzzy entity-relationship model).

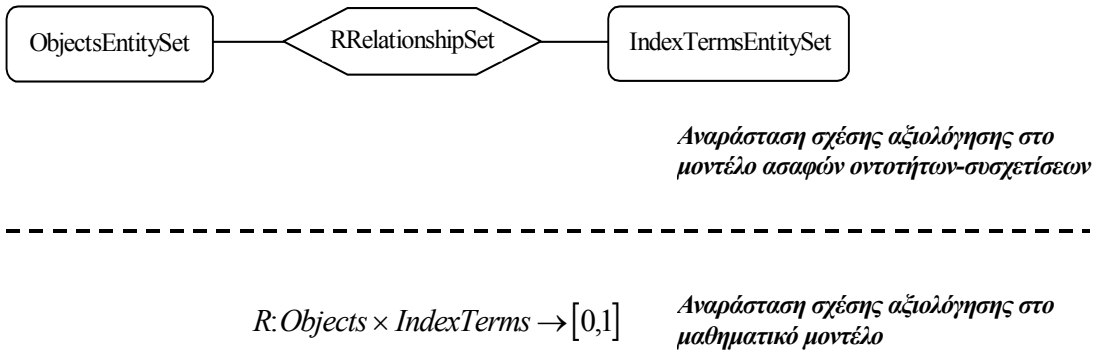
Επίσης, ***η έννοια του συνόλου συσχετίσεων (relationship set) αποτελεί τη μεταφορά της έννοιας της σχέσης αξιολόγησης στο μοντέλο ασαφών οντοτήτων συσχετίσεων***. Η συσχέτιση, όπως και η σχέση με αξιολόγηση, ορίζεται μεταξύ δύο συνόλων και συσχετίζει τα μέλη του ενός συνόλου με τα μέλη του άλλου αποδίδοντας ταυτόχρονα και κάποιο βάρος σπουδαιότητας. Το βάρος αυτό εκφράζει το βαθμό συσχέτισης μεταξύ των αντίστοιχων μελών των δύο συνόλων και αυξάνεται αναλογικά, δηλαδή όσο μεγαλύτερο είναι το βάρος τόσο μεγαλύτερη είναι και η συσχέτιση των αντίστοιχων μελών των δύο συνόλων.

Έστω, για παράδειγμα, ένα σύνολο (με τη μαθηματική έννοια του όρου) *Objects* και ένα σύνολο *IndexTerms.* Ας υποθέσουμε ότι ορίζεται η σχέση αξιολόγησης *R* από το σύνολο *Objects* στο σύνολο *IndexTerms*, δηλαδή:

$$R: Objects \times IndexTerms \rightarrow [0,1]$$

Εκφράζοντας τα παραπάνω στα πλαίσια του μοντέλου ασαφών οντοτήτων-συσχετίσεων ορίζουμε το σύνολο οντοτήτων (entity set) *ObjectsEntitySet* το οποίο αντισχοιχεί στο σύνολο *Objects* και το σύνολο οντοτήτων *IndexTermsEntitySet* το οποίο αντιστοιχεί στο σύνολο *IndexTerms*. Ορίζεται επίσης το σύνολο συσχετίσεων *R_RelationshipSet* το οποίο αντισχοιχεί στη σχέση αξιολόγησης *R* και συσχετίζει μέλη του συνόλου οντοτήτων *ObjectsEntitySet* με μέλη του συνόλου οντοτήτων

IndexTermsEntitySet αποδίδοντας ταυτόχρονα και κάποιο βάρος σπουδαιότητας. Οι προαναφερθείσες αντισχοιχίσεις απεικονίζονται στο ακόλουθο σχήμα:



Σχήμα 4-3:Αντιστοίχηση μεταξύ των εννοιών του μοντέλου ασαφών οντοτήτων-συσχετίσεων και των εννοιών του μαθηματικού μοντέλου εκτέλεσης ερωτήσεων ομοιότητας

Μετά τον καθορισμό των συσχετίσεων μεταξύ των εννοιών του μαθηματικού μοντέλου εκτέλεσης ερωτήσεων ομοιότητας και του μοντέλου ασαφών οντοτήτων-συσχετίσεων θα χρησιμοποιείται η ορολογία του μαθητικού μοντέλου προς διευκόλυνση ενώ, η μεταφορά των λεγομένων στα πλαίσια του μοντέλου ασαφών οντοτήτων-συσχετίσεων θα θεωρείται δεδομένη και αυτονόητη.

4.2.6 Αναπαράσταση μιας συλλογής με τη χρήση της σχέσης με αξιολόγηση

Στο τμήμα αυτό θα παρουσιαστεί ο τρόπος με τον οποίο αναπαρίστανται τα δεδομένα ενός συστήματος με βάση το μοντέλο ασαφών οντοτήτων-συσχετίσεων. Χρησιμοποιώντας το μοντέλο ασαφών οντοτήτων-συσχετίσεων ορίζεται το σχήμα ενός συστήματος ανάκτησης πληροφορίας. Ο ορισμός του σχήματος αποτελεί την πρώτη φάση της μεθοδολογίας εκτέλεσης ερωτήσεων ομοιότητας. Στις επόμενες παραγράφους θα παρουσιαστεί ο τρόπος με τον οποίο περιγράφεται το σχήμα ενός συστήματος ανάκτησης πληροφορίας στα πλάισια του επεκταμένου μοντέλου οντοτήτων-συσχετίσεων.

Η έννοια της σχέσης με αξιολόγηση (τμήμα 3.2.2) βρίσκει πρακτική εφαρμογή στην αναπαράσταση των αντικειμένων μιας συλλογής και στη σύνδεσή τους με τους όρους που περιγράφουν κάθε αντικείμενο, δηλαδή τους όρους δεικτοδότησης (τμήμα 4.2.1). Υπενθυμίζεται ότι μια συλλογή αποτελείται από ένα

σύνολο αντικειμένων, και σε κάθε αντικείμενο επισυνάπτεται ένα σύνολο όρων οι οποίοι περιγράφουν το αντικείμενο. Αυτοί οι όροι λέγονται όροι δεικτοδότησης. Τα σύνολα των αντικειμένων και των όρων δεικτοδότησης μοντελοποιούνται με την έννοια του συνόλου οντοτήτων ενώ, η συσχέτιση μεταξύ αντικειμένων και όρων δεικτοδότησης με την έννοια του συνόλου συσχετίσεων.

Τα παραπάνω μπορούν να περιγραφούν με έναν πιο ξεκάθαρο τρόπο με τη βοήθεια της σχέσης με αξιολόγηση. Έτσι ορίζεται μια σχέση αξιολόγησης R από το σύνολο *Objects* των αντικειμένων μιας συλλογής στο σύνολο *IndexTerms* των όρων δεικτοδότησης της συλλογής, δηλαδή :

$$R: Objects \times IndexTerms \rightarrow [0,1]$$

Έτσι γίνεται εφικτή όχι μόνο η σύνδεση κάθε αντικειμένου της συλλογής με τους όρους δεικτοδότησης του αλλά επιπλέον σε κάθε ζεύγος $(Object, IndexTerm) \in R$, όπου $Object \in Objects$ και $IndexTerm \in IndexTerms$, αντιστοιχίζεται ένας αριθμός στο κλειστό διάστημα [0,1] ο οποίος εκφράζει το πόσο καλά περιγράφεται το συγκεκριμένο αντικείμενο από αυτόν τον όρο δεικτοδότησης, αποδίδεται δηλαδή σε κάθε όρο δεικτοδότησης ένα ***βάρος σπουδαιότητας (Importance Weight)*** ως προς το αντικείμενο που περιγράφει.

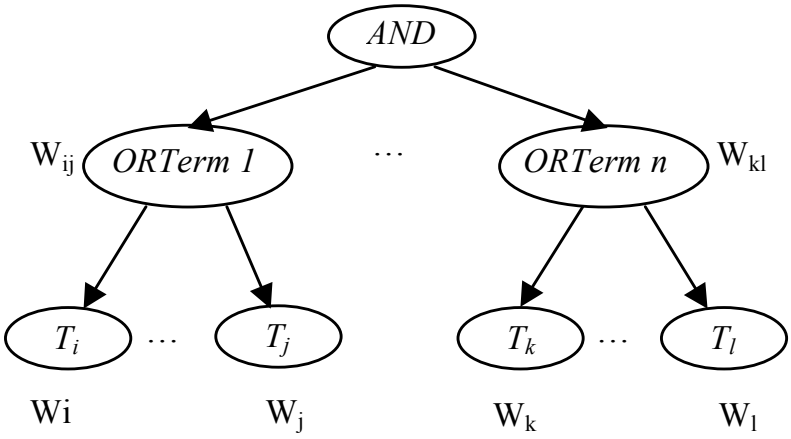
Συνοψίζοντας, αυτό που προκύπτει είναι ότι η συλλογή ενός συστήματος ανάκτησης πληροφορίας μπορεί να αναπαρασταθεί με τη βοήθεια σχέσεων αξιολόγησης. Στο επόμενο τμήμα θα παρουσιαστεί ο τρόπος με τον οποίο οι υποβαλλόμενες ερωτήσεις αναπαρίστανται ως σχέσεις με αξιολόγηση.

4.2.7 Αναπαράσταση των υποβαλλόμενων ερωτήσεων ως σχέσεις αξιολόγησης

Στο τμήμα αυτό θα παρουσιαστεί ο τρόπος με τον οποίο αναπαρίστανται οι ερωτήσεις στα πλαίσια του μοντέλου ασαφών οντοτήτων-συσχετίσεων. ***Οι ερωτήσεις στο μοντέλο ασαφών οντοτήτων-συσχετίσεων θεωρείται ότι βρίσκονται σε κανονική συζευκτική μορφή***, δηλαδή μια ερώτηση ομοιότητας Q θα έχει την ακόλουθη γενική μορφή:

$Q = ((T_i , W_i) \ OR \ ... OR \ (T_j, W_j), W_{ij}) \ AND \ ... \ AND \ ((T_k , W_k) OR \ ... OR \ (T_l , W_l), W_{kl})$

Η γραφική απεικόνιση αυτής της ερώτησης θα είχε ως εξής:



Σχήμα 4-4: Γραφική απεικόνιση γενικής ερώτησης σε κανονική συζευκτική μορφή

Μια ερώτηση σε κανονική συζευκτική μορφή, μπορεί να περιγραφεί με δύο σχέσεις αξιολόγησης. Προκειμένου να γίνει αυτό ορίζεται το σύνολο *ORTerms* των διαζευκτικών όρων της ερώτησης καθώς και το σύνολο *Q* των ερωτήσεων. Υπενθυμίζεται ότι έχει ήδη οριστεί, στο προηγούμενο τμήμα, το σύνολο *IndexTerms* των όρων δεικτοδότησης. Μια ερώτηση ορίζεται ως ένα υποσύνολο όρων δεικτοδότησης οι οποίοι πρέπει να επισυνάπτονται σε ένα έγγραφο για να ανακτηθεί.

Ορίζεται η σχέση αξιολόγησης Queries από το σύνολο Q των ερωτήσεων στο σύνολο ORTerms των διαζευκτικών όρων της ερώτησης. Δηλαδή:

$$Queries: Q \times ORTerms \rightarrow [0,1]$$

Κάθε ζεύγος $(q, or_t) \in Queries$ με $q \in Q$ και $or_t \in ORTerms$ σημαίνει ότι ο διαζευκτικός όρος *or_t* λαμβάνει μέρος στο σχηματισμό της ερώτησης *q* και μάλιστα έχει και κάποιο βάρος σπουδαιότητας το οποίο εκφράζει τη σπουδαιότητα του διαζευκτικού όρου (*OR-Term*) στην ερώτηση *q* ενώ παίρνει τιμές στο διάστημα [0,1]. Στο παράδειγμα των ξενοδοχείων ας θεωρήσουμε την εξής ερώτηση *q*:

"Τα ξενοδοχεία που έχουν (Indoor Theatre(0.2) ή Video(0.3)) (0.4) και (Air Condition(0.4) ή Swimming pool(0.6)) (0.7)."

Η λογική έκφραση αυτής της ερώτησης έχει ως εξής:

((Indoor Theatre, 0.2 OR Video, 0.3), 0.4) AND
((Air Condition, 0.4 OR Swimming pool, 0.6),0.7)

Η ερώτηση *q* έχει δύο διαζευκτικούς όρους. Ο ένας όρος είναι ο *(Indoor Theatre, 0.2 OR Video, 0.3)* και έχει βάρος σπουδαιότητας 0.4 και ο άλλος είναι ο *(Air Condition, 0.4 OR Swimming pool, 0.6)* και έχει βάρος σπουδαιότητας 0.7. Στον πρώτο διαξευκτικό όρο αποδίδεται η ταυτότητα (id) 1 και στο δεύτερο η ταυτότητα 2. Συνεπώς θα υπάρχουν τα εξής δύο ζεύγη:

(q, (Indoor Theatre, 0.2 OR Video, 0.3))=(q, 1) $\rightarrow$ 0.4
(q, (Air Condition, 0.4 OR Swimming pool, 0.6) =(q, 2) $\rightarrow$ 0.7

Ορίζεται η σχέση αξιολόγησης Terms από το σύνολο ORTerms των διαζευκτικών όρων στο σύνολο IndexTerms των όρων δεικτοδότησης. Δηλαδή:

$$Terms: ORTerms \times IndexTerms \rightarrow [0,1]$$

Κάθε ζέυγος $(or_t, IndexT) \in Terms$ με $or_t \in T$ και $IndexT \in IndexTerms$ σημαίνει ότι ο όρος δεικτοδότησης $IndextT$ λαμβάνει μέρος στο σχηματισμό του διαζευκτικού όρου or_t και μάλιστα φέρει κάποιο βάρος σπουδαιότητας το οποίο εκφράζει τη σπουδαιότητα του όρου στην ερώτηση, ενώ παίρνει τιμές στο διάστημα [0,1].
Ας θεωρήσουμε και πάλι την ερώτηση *q* του προηγούμενου παραδείγματος:

"Τα ξενοδοχεία που έχουν (Indoor Theatre(0.2) ή Video(0.3)) (0.4) και (Air Condition(0.4) ή Swimming pool(0.6)) (0.7)."

Η ερώτηση *q* έχει δύο διαζευκτικούς όρους. Ο πρώτος όρος είναι ο *(Indoor Theatre, 0.2 OR Video, 0.3)* έχει βάρος σπουδαιότητας 0.4 και του αποδίδεται η ταυτότητα (id) 1. Ο δεύτερος είναι ο *(Air Condition, 0.4 OR Swimming pool, 0.6)* έχει βάρος σπουδαιότητας 0.7 και του αποδίδεται η ταυτότητα 2.

Συνεπώς θα υπάρχουν τα εξής ζεύγη:

$$(1, \textit{Indoor Theatre}) \rightarrow 0.2$$
$$(1, \textit{Video}) \rightarrow 0.3$$
$$(2, \textit{Air Conditioned}) \rightarrow 0.4$$
$$(2, \text{Swimming pool}) \rightarrow 0.6$$

Συμπερασματικά, προκύπτει ότι μια ερώτηση σε κανονική συζευκτική μορφή μπορεί να αναπαρασταθεί με τη βοήθεια δύο σχέσεων αξιολόγησης. Η μία σχέση ορίζεται από το σύνολο Q των ερωτήσεων στο σύνολο *ORTerms* των διαζευκτικών όρων μιας ερώτησης και αναπαριστά τη συσχέτιση μιας ερώτησης με τους διαζευκτικούς της όρους και η άλλη ορίζεται από το σύνολο *ORTerms* των διαζευκτικών όρων στο σύνολο *IndexTerms* των όρων δεικτοδότησης και αναπαριστά τη συσχέτιση των διαζευκτικών όρων μιας ερώτησης με τους όρους δεικτοδότησης που σχηματίζουν την ερώτηση.

4.3 Υλοποίηση σχέσεων αξιολόγησης και σύνθεσής τους σε σχεσιακή βάση δεδομένων

Στο προηγούμενο κεφάλαιο (τμήμα 3.2.2) ορίστηκε η έννοια της σχέσης με αξιολόγηση. Υπενθυμίζεται ότι μια σχέση με αξιολόγηση r από το σύνολο V στο σύνολο W είναι μια συνάρτηση $r{:}V \times W \rightarrow [0,1]$. Επίσης ορίστηκε η σύνθεση σχέσεων αξιολόγησης ως μια πράξη $\circ$ μεταξύ σχέσεων αξιολόγησης που απεικονίζει τις σχέσεις $r{:}V \times W \rightarrow [0,1]$ και $s{:}W \times Z \rightarrow [0,1]$ στη σχέση $(r \circ s){:}V \times Z \rightarrow [0,1]$ βάσει μιας πράξης πολλαπλασιασμού $\otimes{:}[0,1] \times [0,1] \rightarrow A$ και μιας πράξης πρόσθεσης $\oplus{:}\, A \times A \rightarrow [0,1]$ ως εξής:

$$(r \circ s)(v,z) = \bigoplus_{w \in W} \left(r(v,w) \otimes s(w,z) \right)$$

Το A είναι ένα σύνολο ενδιάμεσων τιμών που δεν είναι πάντα το διάστημα $[0,1]$ - αν και συνήθως είναι αυτό - όπως θα φανεί στη συνέχεια. Για κάθε διαφορετικό ζευγάρι πράξεων $\oplus$ και $\otimes$ προκύπτει διαφορετική πράξη σύνθεσης.

Το ζητούμενο είναι ο μετασχηματισμός των προηγούμενων αλγεβρικών εκφράσεων σε *SQL* εκφράσεις οι οποίες θα χρησιμοποιούνται για την εκτέλεση

ερωτήσεων ομοιότητας πάνω σε μια σχεσιακή βάση δεδομένων. Ο μετασχηματισμός αυτός επιτυγχάνεται μέσω δύο βημάτων τα οποία περιγράφονται στα επόμενα δύο τμήματα.

Βήμα 1 *(υλοποίηση σχέσεων αξιολόγησης σε σχεσιακή βάση δεδομένων)*

Έχοντας θεωρήσει ως τεχνολογική βάση για την ανάπτυξη του συστήματος τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων το πρώτο βήμα είναι η υλοποίηση μιας σχέσης αξιολόγησης πάνω σε μια σχεσιακή βάση δεδομένων.

Η αποθήκευση μιας σχέσης με αξιολόγηση μπορεί να γίνει σε ένα πίνακα μιας βάσης δεδομένων. Ας θεωρήσουμε μια σχέση με αξιολόγηση *R* η οποία ορίζεται από το σύνολο *Objects* στο σύνολο *IndexTerms,* δηλαδή:

$$R: Objects \times IndexTerms \rightarrow [0,1]$$

Το σύνολο *Objects* αποθηκεύεται στον πίνακα *Objects(ObjectID,…)* ενώ το σύνολο *IndexTerms* στον πίνακα *IndexTerms(IndexTermID,…).* Η σχέση με αξιολόγηση *R* αποθηκεύεται στον πίνακα *ObjectsIndexTerms(ObjectID, IndexTermID, Weight,…).*

Κάθε στοιχείο *[ObjectID, IndexTermID, Weight]* του πίνακα *ObjectsIndexTerms* σημαίνει ότι το στοιχείο *ObjectID* του *Objects* συσχετίζεται με το στοιχείο *IndexTermID* του *IndexTerms* με βάρος *Weight*.

Βασιζόμενοι στην αρχή του closed world assumption αποθηκεύονται μόνο τα στοιχεία των σχέσεων με μη μηδενικό βάρος. Σύμφωνα με την αρχή αυτή ότι δεν ορίζεται είναι ψευδές, οπότε, σε μια βάση δεδομένων είναι δυνατόν να αποθηκεύονται μόνο τα μη μηδενικά στοιχεία, ενώ τα μηδενικά στοιχεία θεωρείται ότι δεν ορίζονται και συνεπώς είναι ψευδείς. Με τον τρόπο αυτό μειώνεται ο όγκος της πληροφορίας που αποθηκεύεται δεδομένου ότι το υποσύνολο των μη μηδενικών στοιχείων είναι συνήθως πολύ μικρότερο από το υποσύνολο των μηδενικών.

Στο παράδειγμα των ξενοδοχείων μπορεί να θεωρηθεί ότι το σύνολο *Objects* είναι το σύνολο των ξενοδοχείων και αποθηκεύεται σε ένα πίνακα *Hotels(HotelID,…)* ενώ, το σύνολο *IndexTerms* είναι το σύνολο των διευκολύνσεων και αποθηκεύεται σε ένα πίνακα *Facilities(FacilityID…)*. Η σχέση με αξιολόγηση *R* ορίζεται τότε από το

σύνολο των ξενοδοχείων στο σύνολο των διευκολύνσεων, αντιστοιχίζει δηλαδή σε κάθε ξενοδοχείο τις διευκολύνσεις που αυτό παρέχει και μάλιστα αποδίδει και κάποιο βάρος σπουδαιότητας σε κάθε διευκόλυνση ως προς το ξενοδοχείο που την παρέχει. Το βάρος αυτό εκφράζει το πόσο καλά προσφέρεται η διευκόλυνση από το ξενοδοχείο. Η σχέση αυτή αποθηκεύεται σε ένα πίνακα *HotelFacilities(HotelID,FacilityID,Weight,…)*.

Προκύπτει συνεπώς ότι μια σχέση με αξιολόγηση καθώς και τα σύνολα που την απαρτίζουν μπορούν να αποθηκευτούν σε πίνακες μιας βάσης δεδομένων.

Βήμα 2 *(υλοποίηση σύνθεσης σχέσεων αξιολόγησης σε σχεσιακή βάση δεδομένων)*

Το δεύτερο βήμα είναι η υλοποίηση της πράξης της σύνθεσης σχέσεων αξιολόγησης με τη χρήση *SQL*. Θα μελετηθεί αρχικά πως υλοποιείται μια γενική πράξη σύνθεσης $\circ$ που ορίζεται βάσει της πράξης $\oplus$ που παίζει το ρόλο της πρόσθεσης και της πράξης $\otimes$ που παίζει το ρόλο του πολλαπλασιασμού.

Η πράξη της σύνθεσης των σχέσεων αξιολόγησης αντιστοιχεί στην πράξη της φυσικής σύνδεσης (natural join) πινάκων όπου το γνώρισμα ως προς το οποίο γίνεται η σύνδεση είναι αυτό που αντιστοιχεί στο κοινό σύνολο των δύο σχέσεων αξιολόγησης.

Έστω μια σχέση αξιολόγησης $R: A \times B \to [0,1]$ και μια σχέση $S: B \times C \to [0,1]$. Σε μια σχεσιακή βάση δεδομένων όπου τα σύνολα *A, B, C* έχουν αποθηκευτεί αντίστοιχα στους πίνακες $A = (IDA,\ldots)$, $B = (IDB\ldots)$, $C = (IDC,\ldots)$, οι σχέσεις αυτές θεωρείται ότι έχουν αποθηκευτεί στους πίνακες $R = (IDA, IDB, WR,\ldots)$, $S = (IDB, IDC, WS,\ldots)$.

Για τις σχέσεις αξιολόγησης $R: A \times B \to [0,1]$ και $S: B \times C \to [0,1]$, η φυσική σύνδεση θα γίνει πάνω στο γνώρισμα *B*. Για να γίνει αυτό κατανοητό, ας θεωρηθεί αρχικά η παρακάτω ερώτηση στην *ANSI SQL*:

```
SELECT R.IDA AS IDA,R.IDB AS IDB, S.IDC AS IDC, ⊗(R.WR, S.WS)
FROM R JOIN S ON R.IDB=S.IDB
```

Η ερώτηση αυτή υπολογίζει τους (μη μηδενικούς) όρους που εμφανίζονται στην έκφραση της σύνδεσης των δύο στοιχείων. Για να υπολογιστούν, όμως, τα στοιχεία της σχέσης θα πρέπει να ομαδοποιηθούν οι όροι αυτοί ως προς τα γνωρίσματα *R.IDA* και *S.IDC* και για κάθε ομάδα να υπολογιστεί η διάξευξη των όρων στους οποίους αντιστοιχεί. Αυτό επιτυγχάνεται με την επόμενη ερώτηση που είναι μια επέκταση της προηγούμενης, σύμφωνα με όσα ειπώθηκαν, και είναι η γενική μορφή της ερώτησης σε *SQL* που υλοποιεί την πράξη της σύνθεσης:

Γενική μορφή ερώτησης σε *SQL* που υλοποιεί την πράξη της σύνθεσης

```
SELECT IDA, IDC, ⊕(W1, W2,…, Wn) AS W
FROM (
      SELECT R.IDA AS IDA, S.IDC AS IDC,
      ⊗(R.WR, S.WS) AS W1, W2,…,Wn
      FROM R JOIN S ON R.IDB=S.IDB
)
GROYP BY IDA, IDC
```

Όπου η πράξη $\oplus$ παίζει το ρόλο της πρόσθεσης και η πράξη $\otimes$ παίζει το ρόλο του πολλαπλασιασμού. Το εσωτερικό τμήμα της παραπάνω ερώτησης υπολογίζει τον πολλαπλασιασμό $\otimes$ ενώ, το εξωτερικό τμήμα υπολογίζει, χρησιμοποιώντας φυσικά το εσωτερικό, την πρόσθεση $\oplus$.

Προκειμένου τα παραπάνω να γίνουν περισσότερο κατανοητά ας θεωρήσουμε το εξής παράδειγμα. Έστω οι σχέσεις αξιολόγησης $R: A \times B \to [0,1]$ και $S: B \times C \to [0,1]$. Η κάθε μια από αυτές τις σχέσεις μπορεί να αναπαρασταθεί με έναν δισδιάστατο πίνακα του οποίου κάθε στοιχείο είναι η αντίστοιχη τιμής τους (τμήμα 3.2.2). Έστω λοιπόν ότι αναπαρίστανται από τους παρακάτω πίνακες:

$$R:|A| \times |B| = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.4 \\ 0.2 & 0.7 \\ 0.3 & 0.5 \end{bmatrix} \end{array} \qquad S:|B| \times |C| = \begin{array}{c} \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.7 \end{bmatrix} \end{array}$$

Από τους παραπάνω πίνακες προκύπτει, για παράδειγμα, για τη σχέση αξιολόγησης *R* ότι το στοιχείο α_1 σχετίζεται με το στοιχείο b_1 με ένα βάρος σπουδαιότητας 0.1.

Η σύνθεση $R \circ S$ ανάγεται στον πολλαπλασιασμό των παραπάνω πινάκων (τμημα 3.2.2), δηλαδή:

$$(R \circ S)(a,c) = \oplus\left(R(a,b) \otimes S(b,c)\right) = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{cc} b_1 & b_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.4 \\ 0.2 & 0.7 \\ 0.3 & 0.5 \end{bmatrix} \end{array} \circ \begin{array}{c} \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.7 \end{bmatrix} \end{array} =$$

$$= \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{bmatrix} 0.1 \otimes 0.1 \oplus 0.4 \otimes 0.2 & 0.1 \otimes 0.3 \oplus 0.4 \otimes 0.7 \\ 0.2 \otimes 0.1 \oplus 0.7 \otimes 0.2 & 0.2 \otimes 0.3 \oplus 0.7 \otimes 0.7 \\ 0.3 \otimes 0.1 \oplus 0.5 \otimes 0.2 & 0.3 \otimes 0.3 \oplus 0.5 \otimes 0.7 \end{bmatrix} \end{array} \qquad (4.1)$$

Αν υποθέσουμε τώρα ότι οι σχέσεις αξιολόγησης $R: A \times B \to [0,1]$ και $S: B \times C \to [0,1]$ αποθηκεύονται στους πίνακες R και S μια βάσης δεδομένων οι πίνακες αυτοί θα έχουν (τμήμα 4.3) ως εξής:

Πίνακας R

AID	BID	W
α_1	b_1	0.1
α_1	b_2	0.4
α_2	b_1	0.2
α_2	b_2	0.7
α_3	b_1	0.3
α_3	b_2	0.5

Μορφή πίνακας όπου αποθηκεύεται η σχέση αξιολόγησης R σε μια βάση δεδομένων

Πίνακας S

BID	CID	W
b_1	c_1	0.1
b_1	c_2	0.3
b_2	c_1	0.2
b_2	c_2	0.7

Μορφή πίνακας όπου αποθηκεύεται η σχέση αξιολόγησης S σε μια βάση δεδομένων

Ας υποθέσουμε τώρα ότι εκτελείται το εσωτερικό τμήμα της ερώτησης που υλοποιεί την πράξη της σύνθεσης και υπολογίζει τον πολλαπλασιασμό $\otimes$, δηλαδή το τμήμα:

```
SELECT R.IDA AS IDA, S.IDC AS IDC,
⊗(R.WR, S.WS) AS W1
FROM R JOIN S ON R.IDB=S.IDB
```

Ο πίνακας που προκύπτει από την εκτέλεση της προηγούμενης έκφρασης θα έχει ως εξής:

AID	CID	W1
α_1	c_1	$0.1 \otimes 0.1$
α_1	c_2	$0.1 \otimes 0.3$
α_1	c_1	$0.4 \otimes 0.2$
α_1	c_2	$0.4 \otimes 0.7$
α_2	c_1	$0.2 \otimes 0.1$
α_2	c_2	$0.2 \otimes 0.3$
α_2	c_1	$0.7 \otimes 0.2$
α_2	c_2	$0.7 \otimes 0.7$
α_3	c_1	$0.3 \otimes 0.1$
α_3	c_2	$0.3 \otimes 0.3$
α_3	c_1	$0.5 \otimes 0.2$
α_3	c_2	$0.5 \otimes 0.7$

Πίνακας που προκύπτει μετά την εκτέλεση του εσωτερικού τμήματος της ερώτησης που υλοποιεί την πράξη της σύνθεσης.

Το εξωτερικο τμήμα της ερώτησης που υλοποιεί την πράξη της σύνθεσης και υπολογίζει την πρόσθεση $\oplus$, εκτελείται στην ουσία πάνω στον προηγούμενο πίνακα. Ο τελικός πίνακας που προκύπτει θα έχει ως εξής:

AID	CID	W
α_1	c_1	$0.1 \otimes 0.1 \oplus 0.4 \otimes 0.2$
α_1	c_2	$0.1 \otimes 0.3 \oplus 0.4 \otimes 0.7$
α_2	c_1	$0.2 \otimes 0.1 \oplus 0.7 \otimes 0.2$
α_2	c_2	$0.2 \otimes 0.3 \oplus 0.7 \otimes 0.7$
α_3	c_1	$0.3 \otimes 0.1 \oplus 0.5 \otimes 0.2$
α_3	c_2	$0.3 \otimes 0.3 \oplus 0.5 \otimes 0.7$

Τελικός πίνακας που προκύπτει μετά την εκτέλεση του εξωτερικού τμήματος της ερώτησης που υλοποιεί την πράξη της σύνθεσης.

Είναι φανερό ότι ο πίνακας αυτός αποθηκεύει το αποτέλεσμα της σύνθεσης $R \circ S$ όπως αυτό εκφράζεται από τη *σχέση (4.1)*.

Αυτό που έχουμε καταφέρει μέχρι τώρα είναι να υπολογίσουμε τη σύνθεση σχέσεων με αξιολόγηση με την εκτέλεση μιας *SQL* έκφρασης.

4.4 Υλοποίηση των υποβαλλόμενων ερωτήσεων σε σχεσιακή βάση δεδομένων

Όπως προαναφέρθηκε (τμήμα 4.2.7) οι ερωτήσεις στο μοντέλο ασαφών οντοτήτων συσχετίσεων βρίσκονται σε κανονική συζευκτική μορφή. Μια ερώτηση σε κανονική συζευκτική μορφή μπορεί να αναπαρασταθεί με τη βοήθεια δύο σχέσεων αξιολόγησης. Η σχέση αξιολόγησης *Queries* ορίζεται από το σύνολο *Q* των ερωτήσεων στο σύνολο *ORTerms* των διαζευκτικών όρων μιας ερώτησης και αναπαριστά τη συσχέτιση μιας ερώτησης με τους διαζευκτικούς της όρους. Η σχέση αξιολόγησης *Terms* ορίζεται από το σύνολο *ORTerms* των διαζευκτικών όρων στο σύνολο *IndexTerms* των όρων δεικτοδότησης που σχηματίζουν την ερώτηση.

Η σχέση αξιολόγησης *Queries* μεταξύ των συνόλων *Q* και *ORTerms*, με βάση την ερώτηση που απεικονίζεται στο σχήμα 4.4, αποθηκεύεται σε ένα πίνακα της βάσης δεδομένων του οποίου η γενική μορφή θα έχει ως εξής:

Πίνακας *Queries*

QueryID	ORTermID	Weight
Q_{id}	*OR Term 1*	W_{ij}
…	…	…
Q_{id}	*OR Term n*	W_{kl}

Πίνακας ο οποίος χρησιμοποιείται για την αποθήκευση της σχέσης αξιολόγησης μεταξύ του συνόλου Q των ερωτήσεων και του συνόλου ORTerms των διαζευκτικών όρων μια ερώτησης. Το πρωτεύων κλειδί του πίνακα είναι ο συνδυασμός των QueryID και ORTermID

Όπου,

- *QueryID* είναι η ταυτότητα κάθε ερώτησης η οποία αποδίδεται αυτόματα από το σύστημα για κάθε ερώτηση.
- *ORTermID* είναι η ταυτότητα του κάθε διαζευκτικού όρου *(OR-Term)* της ερώτησης και αποδίδεται επίσης αυτόματα από το σύστημα για κάθε διαζευκτικό όρο.
- *Weight* είναι το βάρος σπουδαιότητας που μπορεί να έχει κάθε διαζευκτικός όρος της ερώτησης εφ' όσον φυσικά οι διαζευκτικοί όροι έχουν κάποια βάρη.
- Το πρωτεύων κλειδί του παραπάνω πίνακα είναι ο συνδυασμός των *QueryID* και *ORTermID.*

Η σχέση αξιολόγησης *Terms* μεταξύ των συνόλων *ORTerms* και *IndexTerms*, με βάση την ερώτηση που απεικονίζεται στο σχήμα 4.4, αποθηκεύεται σε ένα πίνακα της βάσης δεδομένων του οποίου η γενική μορφή θα έχει ως εξής:

Πίνακας *Terms*

ORTermID	IndexTermID	Weight
OR Term 1	*Ti*	W_I
OR Term 1	T_j	*Wj*
…	…	…
OR Term n	T_k	W_k
OR Term n	T_l	W_l

Πίνακας ο οποίος χρησιμοποιείται για την αποθήκευση της σχέσης μεταξύ του συνόλου ORTerms των διαζευκτικών όρων και του συνόλου IndexTerms των όρων δεικτοδότησης. Το πρωτεύων κλειδί του πίνακα είναι ο συνδυασμός των ORTermID και IndexTermID

Όπου,

- *IndexTermID* είναι η ταυτότητα του όρου δεικτοδότησης που συμμετέχει στο σχηματισμό κάποιου διαζευκτικού όρου της ερώτησης.
- *ORTermID* είναι η ταυτότητα του κάθε διαζευκτικού όρου της ερώτησης η οποία δίνεται αυτόματα από το σύστημα.
- *Weight* είναι το βάρος που μπορεί να έχει κάθε όρος της ερώτησης εφ' όσον φυσικά οι όροι της ερώτησης έχουν κάποια βάρη.
- Το πρωτεύων κλειδί του παραπάνω πίνακα είναι ο συνδυασμός των *IndexTermID* και *ORTermID*.

Στο παράδειγμα των ξενοδοχείων ας θεωρήσουμε την εξής ερώτηση:

"Τα ξενοδοχεία που έχουν (Indoor Theatre(0.2) ή Video(0.3)) (0.4) και (Air Condition(0.4) ή Swimming pool(0.6)) (0.7)."

Η λογική έκφραση αυτής της ερώτησης έχει ως εξής:

((Indoor Theatre, 0.2 OR Video, 0.3), 0.4) AND
((Air Condition, 0.4 OR Swimming pool, 0.6),0.7)

Για τη συγκεκριμένη ερώτηση θεωρώντας ότι της έχει αποδοθεί η ταυτότητα (id) 1 ο πίνακας *Queries* θα έχει ως εξής:

Πίνακας *Queries*

QueryID	ORTermID	Weight
1	1	0.4
1	2	0.7

Πίνακας ο οποίος αποθηκεύει τους διαζευκτικούς όρους για την ερώτηση του προηγούμενου παραδείγματος

Κάθε εγγραφή στον πίνακα αυτό, παριστάνει την παρουσία ενός συγκεκριμένου συζευτικού όρου σε μια συγκεκριμένη ερώτηση και με ενα συγκεκριμένο βάρος.
Επίσης ο πίνακας ο *Terms* θα έχει ως εξής:

Πίνακας *Terms*

ORTermID	IndexTermID[3]	Weight
1	Indoor Theatre	0.2
1	Video	0.3
2	Air Conditioned	0.4
2	Swimming pool	0.6

Πίνακας ο οποίος αποθηκεύει τους όρους δεικτοδότησης που σχηματίζουν τους διαζευκτικούς όρους της ερώτησης του προηγούμενου παραδείγματος.

Συμπερασματικά, προκύπτει ότι μια ερώτηση σε κανονική συζευκτική μορφή μπορεί να αναπαρασταθεί με τη βοήθεια δύο σχέσεων αξιολόγησης. οι οποίες αποθηκεύονται σε κατάλληλους πίνακες στη βάση δεδομένων πάνω στην οποία χτίζεται το όλο σύστημα.

Μέχρι αυτό το σημείο λοιπόν έχει καταστεί δυνατό να αναπαρασταθεί και η συλλογή του συστήματος ανάκτησης πληροφορίας αλλά και οι υποβαλλόμενες ερωτήσεις ομοιότητας με κάποιες σχέσεις αξιολόγησης οι οποίες αποθηκεύονται σε πίνακες στη βάση δεδομένων του συστήματος.

[3] Στην πραγματικότητα το *IndexTermID* είναι ένας αριθμός, αλλά εδώ δίνεται το όνομα του για ευκολότερη ανάγνωση.

4.5 Εκτέλεση ερωτήσεων ομοιότητας

Στο τμήμα 3.2.2 ορίστηκε η πράξη της σύνθεσης μεταξύ σχέσεων με αξιολόγηση και αποδείχθηκε ότι η απάντηση σε μια ερώτηση εκφρασμένη σε κανονική συζευκτική ή διαζευκτική μορφή μπορεί να προκύψει ως σύνθεση σχέσεων αξιολόγησης.

Στα προηγούμενα τμήματα παρουσιάστηκε ο τρόπος με τον οποίο μπορούν να αναπαρασταθούν η συλλογή ενός συστήματος ανάκτησης πληροφορίας (τμήμα 4.2.6) καθώς και οι υποβαλλόμενες ερωτήσεις (τμήμα 4.2.7) ως σχέσεις με αξιολόγηση

Επίσης δείχτηκε ότι ο υπολογισμός της σύνθεσης σχέσεων αξιολόγησης ανάγεται τελικά στη φυσική σύνδεση (natural join) μεταξύ των πινάκων που αποθηκεύουν τις σχέσεις αξιολόγησης.. Συνδυάζοντας όλα αυτά το τελικό συμπέρασμα είναι ότι η απάντηση μιας ερώτησης ομοιότητας σε κανονική συζευκτική μορφή θα προκύπτει ως:

$$Queries \circ_{M}^{AND} \left(Terms \circ_{M}^{OR} R \right) \qquad (4.2)$$

για κατάλληλα ορισμένες πράξεις $\oplus_{M}^{OR}$, $\otimes_{M}^{OR}$, $\oplus_{M}^{AND}$, $\otimes_{M}^{AND}$.

Όπου, υπενθυμίζεται,

$R: Objects \times IndexTerms \rightarrow [0,1]$,

$Queries: Q \times ORTerms \rightarrow [0,1]$,

$Terms: ORTerms \times IndexTerms \rightarrow [0,1]$

και

- *Objects* το σύνολο των αντικειμένων της συλλογής.
- *IndexTerms* το σύνολο των όρων δεικτοδότησης της συλλογής.
- *Q* το σύνολο των ερωτήσεων.
- *ORTerms* το σύνολο των διαζευκτικών όρων των ερωτήσεων.

Αναπαριστώντας την παραπάνω σχέση χρησιμοποιώντας την πράξη της φυσικής σύνδεσης προκύπτει:

(ObjectsIndexTerms JOIN Terms) *JOIN Queries*

ON IndexTermID *ON ORTermID*

Όπου, υπενθυμίζεται,

- *ObjectsIndexTerms(ObjectID, IndexTermID, Weight,…)* είναι ο πίνακας στον οποίο αποθηκεύεται η συσχέτιση αντικειμένων – όρων δεικτοδότησης.
- *Queries(QueryID, ORTermID, Weight)* είναι ο πίνακας στον οποίο αποθηκεύονται τα διαζευκτικά μέρη μιας ερώτησης. Αποθηκεύεται δηλαδή η σχέση αξιολόγησης *Queries*.
- *Terms(ORTermID, IndexTermID, Weight)* είναι ο πίνακας στον οποίο αποθηκεύονται οι όροι δεικτοδότησης που σχηματίζουν τους διαζευκτικούς όρους. Αποθηκεύεται δηλαδή η σχέση αξιολόγησης *Terms*.

Με βάση την υλοποίηση της σύνθεσης σχέσεων αξιολόγησης με *SQL*, στο προηγούμενο τμήμα η εκτέλεση μιας ερώτησης ομοιότητας ανάγεται στην εκτέλεση μιας σειράς *SQL* ερωτήσεων. Το γενικό σχήμα της *SQL* ερώτησης που υλοποιεί την πράξη της σύνθεσης (τμήμα 4.3) χωρίζεται, προς διευκόλυνση της κατανόησης, σε δύο τμήματα. Το ένα τμήμα, το εσωτερικό, υπολογίζει τον πολλαπλασιασμό $\otimes$ και το άλλο, το εξωτερικό, υπογίζει την πρόσθεση $\oplus$.

Οι ερωτήσεις **OR1, OR2,** που δίνονται παρακάτω, υπολογίζουν τα ***OR*-τμήματα** δηλαδή τα διαζευκτικά μέρη της ερώτησης και αντιστοιχούν στο τμήμα $\left(Terms \circ_M^{OR} R\right)$ της σχέση *(4.2)***,** ενώ οι **AND1, AND2** υπολογίζουν το ***AND*-τμήμα**, το συζευκτικό και αντιστοιχούν στο **AND** τμήμα ($\circ_M^{AND}$) της σχέση *(4.2)***.** Η ερώτηση **OR1 (**$\otimes_M^{OR}$ **)** υπολογίζει τον πολλαπλασιασμό του διαζευκτικού τμήματος και η **OR2 (**$\oplus_M^{OR}$ **)** υπολογίζει, χρησιμοποιώντας όπως φαίνεται την **OR1**, την πρόσθεση του διαζευκτικού τμήματος. Οι ερωτήσεις **AND1 (**$\otimes_M^{AND}$ **)** και **AND2 (**$\oplus_M^{AND}$ **)** υπολογίζουν τον πολλαπλασιασμό και την πρόσθεση, αντίστοιχα, του συζευκτικού τμήματος.

Οι ερωτήσεις αυτές υπολογίζουν τη σύνθεση σχέσεων αξιολόγησης. Οι ερωτήσεις **OR1, OR2** υπολογίζουν τη σύνθεση $A = Terms \circ R$ και οι **AND1, AND2** τη σύνθεση $Queries \circ A$, συνεπώς η μορφή τους προκύπτει με βάση τη γενική μορφή της *SQL* ερώτησης που υλοποιεί την πράξη της σύνθεσης (τμήμα 4.3).

Συγκεκριμένα η εκτέλεση μιας ερώτησης ομοιότητας πραγματοποιείται με την εκτέλεση των ακόλουθων τεσσάρων *SQL* ερωτήσεων γενικής μορφής:

OR1 ($\otimes_M^{OR}$)

```
SELECT ObjectsIndexTerms.ObjectID, ORTermID,
F1_OR(Terms.Weight, ObjectsIndexTerms.Weight) AS W1
FROM Terms JOIN ObjectsIndexTerms ON
Terms.IndexTermID = ObjectsIndexTerms.IndexTermID
```

OR2 ($\oplus_M^{OR}$)

```
SELECT ObjectID ,ORTermID, F2_OR(OR.W1) AS W
FROM OR1
GROUP BY ObjectID ,ORTermID
```

AND1 ($\otimes_M^{AND}$)

```
SELECT QueryID, ObjectID,
F1_AND( Queries.Weight, OR2.W ) AS W1
FROM Queries JOIN OR2 ON Queries.ORTermID = OR2.ORTermID
```

AND2 ($\oplus_M^{AND}$)

```
SELECT QueryID, ObjectID, F2_AND(AND1.W1) AS W
FROM AND1
GROUP BY QueryID, ObjectID
```

Οι συναρτήσεις *F1_OR, F2_OR, F1_AND, F2_AND* χρησιμοποιούνται για τον υπολογισμό βαρών και τον τελικό βαθμό ομοιότητας του κάθε ανακτηθέντος αντικειμένου ως προς μια ερώτηση. Η διαμόρφωσή τους γίνεται σύμφωνα με το μοντέλο υπολογισμού ομοιότητας που χρησιμοποιείται κάθε φορά και ο ακριβής καθορισμός τους θα είναι το αντικείμενο του επόμενου τμήματος.

Στο παράδειγμα των ξενοδοχείων το σύνολο *Objects* είναι το σύνολο των ξενοδοχείων και αποθηκεύεται στον πίνακα *Hotels(HotelID,…)* ενώ, το σύνολο *IndexTerms* είναι το σύνολο των διευκολύνσεων και αποθηκεύεται στον πίνακα *Facilities(FacilityID,…)*. Η σχέση με αξιολόγηση *R* ορίζεται τότε από το σύνολο των ξενοδοχείων στο σύνολο των διευκολύνσεων, αντιστοιχίζει δηλαδή σε κάθε ξενοδοχείο τις διευκολύνσεις που αυτό παρέχει και μάλιστα αποδίδει και κάποιο

βάρος σπουδαιότητας σε κάθε διευκόλυνση ως προς το ξενοδοχείο που την παρέχει. Το βάρος αυτό εκφράζει το πόσο καλά προσφέρεται η διευκόλυνση από το ξενοδοχείο. Η σχέση αυτή αποθηκεύεται στον πίνακα *HotelFacilities(HotelID, FaicilityID, Weight,…)*. Για το παράδειγμα αυτό λοιπόν υπάρχουν οι εξής αντιστοιχίσεις:

Objects → Hotels
ObjectID → HotelID
IndexTerm → Facilities
IndextTermID → FacilityID
ObjectsIndextTerms → HotelFacilities

Οι πίνακες *Queries* και *Terms* θα έχουν ως εξής:

Πίνακας *Queries*

Όνομα πεδιού	Τύπος
QueryID	Integer
ORTermID	Integer
Weight	Double

Πίνακας *Terms*

Όνομα πεδίου	Τύπος
ORTermID	Integer
FacilityID	Integer
Weight	Double

Ας υποθέσουμε τώρα ότι υποβάλλεται η εξής ερώτηση :

"Τα ξενοδοχεία που έχουν (Indoor Theatre(0.2) ή Video(0.3)) (0.4) και (Air Condition(0.4) ή Swimming pool(0.6)) (0.7)."

Η λογική έκφραση αυτής της ερώτησης έχει ως εξής:

((Indoor Theatre, 0.2 OR Video, 0.3), 0.4) AND

((Air Condition, 0.4 OR Swimming pool, 0.6),0.7)

Στη συγκεκριμένη περίπτωση οι πίνακες *Queries* και *Terms* θα έχουν ως εξής :

Πίνακας *Queries*

QueryID	ORTermID	Weight
1	1	0.4
1	2	0.7

Πίνακας *Terms*

ORTermID	FacilityID[4]	Weight
1	Indoor Theatre	0,2
1	Video	0,3
2	Air Condition	0,4
2	Swimming Pool	0,6

Οι προαναφερθείσες *SQL* ερωτήσεις γίνονται ως εξής:

OR1 ($\otimes_M^{OR}$)

```
SELECT HotelFacilties.HotelID, ORTermID,
F1_OR(Terms.Weight, HotelFacilties.Weight) AS W1
FROM Terms JOIN HotelFacilties ON
Terms.FacilityID = HotelFacilties.FacilityID
```

OR2 ($\oplus_M^{OR}$)

```
SELECT HotelID, ORTermID, F2_OR(OR.W1) AS W
FROM OR1
GROUP BY HotelID, ORTermID
```

AND1 ($\otimes_M^{AND}$)

```
SELECT QueryID, HotelID,
F1_AND( Queries.Weight, OR2.W ) AS W1
FROM Queries JOIN OR2 ON Queries.ORTermID = OR2.ORTermID
```

[4] Στην πραγματικότητα το *FacilityID* είναι ένας αριθμός, αλλά εδώ δίνεται το όνομα του Facility για ευκολότερη ανάγνωση.

AND2 ($\oplus_{M}^{AND}$)

```
SELECT QueryID, HotelID, F2_AND(AND1.W1) AS W
FROM AND1
GROUP BY QueryID, HotelID
```

4.6 Διαμόρφωση ερωτήσεων με βάση το μοντέλο υπολογισμού ομοιότητας

Όταν δόθηκε η γενική περιγραφή του συστήματος, υποστηρίχτηκε ότι η εκτέλεση των ερωτήσεων ομοιότητας θα αντιμετωπιστεί με ενιαίο τρόπο, για όλα τα μοντέλα τα οποία υπολογίζουν την ομοιότητα μεταξύ αντικειμένου-ερώτησης (Fuzzy set, Waller-Kraft, P-Norm, Infinite-One). Στο τμήμα αυτό θα δειχθεί πως επιτεύχθηκε αυτό.

Αναφέρθηκε ότι η εκτέλεση μιας ερώτησης ομοιότητας μπορεί να πραγματοποιηθεί με την εκτέλεση τεσσάρων *SQL* ερωτήσεων. Οι συναρτήσεις *F1_OR, F2_OR, F1_AND, F2_AND* χρησιμοποιούνται για τον υπολογισμό των βαρών των διαζευκτικών και συζευκτικών τμημάτων και συνεπώς για τον υπολογισμό του τελικού βαθμού ομοιότητας του κάθε ανακτηθέντος αντικειμένου. Ο υπολογισμός όμως αυτός εξαρτάται από το μοντέλο υπολογισμού ομοιότητας αντικειμένου-ερώτησης και συνεπώς, για κάθε μοντέλο οι συναρτήσεις αυτές θα έχουν και διαφορετικό τύπο. Αρκεί λοιπόν, να επιλέγεται κάθε φορά η αντίστοιχη για το κάθε μοντέλο συνάρτηση και να εκτελείται κανονικά η σειρά των τεσσάρων *SQL* ερωτήσεων, προς απάντηση της κάθε ερώτησης ομοιότητας.

Παρουσιάζεται το πως διαμορφώνοναι οι προηγούμενες τέσσερις *SQL* ερωτήσεις για τα μοντέλα υπολογισμού ομοιότητας που υποστηρίζονται στο σύστημα. Όλες οι παρακάτω *SQL* ερωτήσεις μπορούν αν προσαρμοστούν στο παράδειγμα των ξενοδοχείων με βάση τις ακόλουθες αντιστοιχίσεις:

Objects ⟶ Hotels
ObjectID ⟶ HotelID
IndexTerm ⟶ Facilities
IndextTermID ⟶ FacilityID
ObjectsIndextTerms ⟶ HotelFacilities

Μοντέλα που δεν έχουν βάρη στους όρους μια ερώτησης

Μορφή SQL ερωτήσεων για το μοντέλο ασαφούς τελεστών (Fuzzy set)

OR1

```
SELECT ObjectsIndexTerms.ObjectID ,ORTermID,
ObjectsIndexTerms.Weight AS W1
FROM ObjectsIndexTerms JOIN Terms ON
ObjectsIndexTerms.IndexTermID = Terms.IndexTermID
```

Υπολογισμός πολλαπλασιασμού του διαζευκτικού τμήματος (OR - part)

OR2

```
SELECT ObjectID, ORTermID, MAX(W1) AS W
FROM OR1
GROUP BY ObjectID, ORTermID
```

Υπολογισμός πρόσθεσης του διαζευκτικού τμήματος (OR - part)

AND1

```
SELECT QueryID, ObjectID, OR2.W AS W1
FROM Queries JOIN OR2 ON Queries.ORTermID = OR2.ORTermID
```

Υπολογισμός πολλαπλασιασμού του συζευκτικού τμήματος (AND - part)

AND2

```
SELECT QueryID, ObjectID, MIN(AND1.W1) AS W
FROM AND1
GROUP BY QueryID, ObjectID
```

Υπολογισμός πρόσθεσης του συζευκτικού τμήματος (AND - part)

Μορφή SQL ερωτήσεων για το μοντέλο Waller-Kraft

OR1

```
SELECT ObjectsIndexTerms.ObjectID, ORTermID,
ObjectsIndexTerms Weight AS W1
FROM ObjectsIndexTerms JOIN Terms ON
ObjectsIndexTerms.IndexTermID = Terms.IndexTermID
```

Υπολογισμός πολλαπλασιασμού του διαζευκτικού τμήματος (OR - part)

OR2

```
SELECT ObjectID, ORTermID,
(1-γ)*MIN(W1) + γ*MAX(W1)) AS W
FROM OR1
GROUP BY ObjectID, ORTermID
```

Υπολογισμός πρόσθεσης του διαζευκτικού τμήματος (OR - part)

AND1

```
SELECT QueryID, ObjectID, OR2.W AS W1
FROM Queries JOIN OR2 ON Queries.ORTermID = OR2.ORTermID
```

Υπολογισμός πολλαπλασιασμού του συζευκτικού τμήματος (AND - part)

AND2

```
SELECT QueryID, ObjectID,
(1-γ)*MIN(W1) + γ*MAX(W1) AS W
FROM AND1
GROUP BY QueryID, ObjectID
```

Υπολογισμός πρόσθεσης του συζευκτικού τμήματος (AND - part)

Μοντέλα που έχουν βάρη στους όρους μια ερώτησης

Μορφή SQL ερωτήσεων για το μοντέλο P-Norm

OR1

```
SELECT ObjectsIndexTerms.ObjectID, ORTermID,
exp((log(Terms.Weight)+log(IndexTable.Weight))*p) AS W1
FROM ObjectsIndexTerms JOIN Terms ON
ObjectsIndexTerms.IndexTermID = Terms.IndexTermID
```

Υπολογισμός πολλαπλασιασμού του διαζευκτικού τμήματος (OR - part)

OR2

```
SELECT ObjectID, ORTermID, exp(log(SUM(W1))/p) AS W
FROM OR1
GROUP BY ObjectID, ORTermID
```

Υπολογισμός πρόσθεσης του διαζευκτικού τμήματος (OR - part)

AND1

```
SELECT QueryID, ObjectID,
exp((log(1-OR2.W)+log(Queries.Weight))*p) AS W1
FROM Queries JOIN OR2 ON Queries.ORTermID = OR2.ORTermID
```

Υπολογισμός πολλαπλασιασμού του συζευκτικού τμήματος (AND - part)

AND2

```
SELECT QueryID, ObjectID,
1-exp(log(SUM(W1))/p) AS W
FROM AND1
```

```
GROUP BY QueryID, ObjectID
```

Υπολογισμός πρόσθεσης του συζευκτικού τμήματος (AND - part)

Μορφή SQL ερωτήσεων για το μοντέλο Infinite-One

OR1

```
SELECT ObjectsIndexTerms.ObjectID, ORTermID,
Terms.Weight1 * ObjectsIndexTerms.Weight AS W1,
Terms.Weight2 * ObjectsIndexTerms.Weight AS W2
FROM ObjectsIndexTerms JOIN Terms ON
ObjectsIndexTerms.IndexTermID = Terms.IndexTermID
```

Υπολογισμός πολλαπλασιασμού του διαζευκτικού τμήματος (OR - part)

OR2

```
SELECT ObjectID, ORTermID,
γ*MAX(W1) + (1-γ)*SUM(W2) AS W
FROM OR1
GROUP BY ObjectID, ORTermID
```

Υπολογισμός πολλαπλασιασμού του διαζευκτικού τμήματος (OR - part)

AND1

```
SELECT QueryID, ObjectID,
(1-IndexTable.Weight)*Terms.Weight1 AS W1,
IndexTable.Weight*Terms.Weight2 AS W2,
FROM Queries JOIN OR2 ON Queries.ORTermID = OR2.ORTermID
```

Υπολογισμός πολλαπλασιασμού του συζευκτικού τμήματος (AND - part)

AND2

```
SELECT QueryID, ObjectID,
γ*(1-MAX(W1)) + (1-γ)*SUM(W2) AS W
FROM AND1
GROUP BY QueryID, ObjectID
```

Υπολογισμός πρόσθεσης του συζευκτικού τμήματος (AND - part)

Πρέπει να επισημανθεί ότι οι τύποι που χρησιμοποιούνται στις *SQL* ερωτήσεις, για τον υπολογισμό των βαρών, δεν είναι όπως ακριβώς προτείνονται στα αντίστοιχα μοντέλα, αλλά, έχουν μετασχηματιστεί έτσι ώστε να διευκολυνθεί ο υπολογισμός τους με χρήση της γλώσσας *SQL*.

4.7 Μεθοδολογία υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων

Στο τμήμα 4.2 παρουσιάστηκε το επεκταμένο μοντέλο οντοτήτων συσχετίσεων. Παρουσίαστηκε ο τρόπος αναπαράστασης της συλλογής ενός συστήματος ανάκτησης πληροφορίας ως σχέσεις αξιολόγησης (τμήμα 4.2.6) καθώς και ο τρόπος αναπαράστασης των υποβαλλόμενων ερωτήσεων ως σχέσεις αξιολόγησης (4.2.7). Στο τμήμα 4.3 παρουσιάστηκε η υλοποίηση των σχέσεων αξιολόγησης και της σύνθεσης σχέσεων αξιολόγησης πάνω σε σχεσιακή βάση δεδομένων ενώ, στο τμήμα 4.4 περιγράφηκε η υλοποίηση των υποβαλλόμενων ερωτήσεων σε σχεσιακή βάση δεδομένων. Σαν άμεση συνέπεια δόθηκε, στο τμήμα 4.5, ο τρόπος εκτέλεσης ερωτήσεων ομοιότητας. Αμέσως μετά στο τμήμα 4.6 περιγράφηκε η διαμόρφωση των ερωτήσεων με βάση το μοντέλο υπολογισμού ομοιότητας.

Όλα αυτά τα τμήματα στην πραγματικότητα περιγράφουν τη μεθοδολογία εκτέλεσης ερωτήσεων ομοιότητας πάνω σε σχεσιακή βάση δεδομένων. Τα τμήματα αυτά περιγράφουν μία οι περισσότερες φάσεις της μεθοδολογίας η οποία έχει ως εξής:

Φάση 1: Ορισμός σχήματος ΣΑΠ

- *Βήμα 1*

  Ορισμός του σχήματος του συστήματος ανάκτησης πληροφορίας μέσω του ορισμού συνόλων οντοτήτων, συνόλων συσχετίσεων και χαρακτηριστικών από το χρήστη-διαχειριστή του συστήματος (τμήμα 4.2.6)

- *Βήμα 2*

  Αυτόματη κατασκευή, από το σύστημα, των αναγκαίων πινάκων για την αποθήκευση των ερωτήσεων.

Φάση 2: Αποκωδικοποίηση υποβαλλόμενων ερωτήσεων

Για κάθε υποβαλλόμενη ερώτηση:

- *Βήμα 1*

  Αποκωδικοποίηση ερώτησης από το σύστημα και:

  1. Καθορισμός των διαζευκτικών και συζευκτικών τμήματων της. ερώτησης (τμήμα 4.2.7).
  2. Καθορισμός του μοντέλου υπολογισμού ομοιότητας
  3. Καθορισμός του συνόλου οντοτήτων μέλη του οποίου στοχεύει να ανακτήσει η ερώτηση (τμήμα 4.2.6).
  4. Καθορισμός του συνόλου συσχετίσεων που εκφράζει την σύνδεση των προς ανάκτηση αντικειμένων με τους όρους δεικτοδότησης (τμήμα 4.2.6).

- *Βήμα 2*

  Αποθήκευση, από το σύστημα, των διαζευκτικών και συζευκτικών τμημάτων της ερώτησης στους κατάλληλους πίνακες οι οποίοι κατασκευάστηκαν στη **Φάση 1** (τμήμα 4.4)

Φάση 3: Κατασκευή *SQL* εκφράσεων

Με βάση το μοντέλο υπολογισμού ομοιότητας που προέκυψε από τη **Φάση 2,** *Βήμα 2***:**

- Αυτόματη κατασκευή *SQL* εκφράσεων για τον υπολογισμό των διαζευκτικών τμημάτων (**OR1** και **OR2** όπως περιγραφηκαν στο τμήμα 4.6)
- Αυτόματη κατασκευή *SQL* εκφράσεων για τον υπολογισμό των συζευκτικών τμημάτων (**AND1** και **AND2** όπως περιγράφηκαν στο τμήμα 4.6)

Φάση 4: Εκτέλεση *SQL* εκφράσεων

- *Βημα 1*

  Εκτέλεση, από το σύστημα, των ερωτήσεων που κατασκευάστηκαν στη **Φάση 3**
- *Βημα 2*

  Κατάλληλη μορφοποίηση, από το σύστημα, των αποτελεσμάτων που προέκυψαν και αποστολή τους στην εφαρμογή παρουσίασης των αποτελεσμάτων.

Στα επόμενα κεφάλαια θα παρουσιαστούν τα τμήματα του συστήματος που υλοποιούν κάθε μια από τις παραπάνω φάσεις

Σ' αυτό το σημείο θα δοθεί ένα παράδειγμα χρήσης της μεθοδολογίας. Ας θεωρήσουμε και πάλι το παράδειγμα των ξενοδοχείων. Στόχος είναι η ανάκτηση ξενοδοχείων με βάση τις διευκολύνσεις που παρέχουν. Έστω ότι σε μια βάση δεδομένων υπάρχει ο πίνακας *Hotels(HotelID,…)* ο οποίος περιέχει ξενοδοχεία, ο πίνακας *Facilities(FacilityID,…)* ο οποίος περιέχει διευκολύνσεις και ο πίνακας *HotelFacilities(HotelID, FacilityID, Weight,…)* ο οποίος για κάθε ξενοδοχείο περιέχει τις διευκολύνσεις που αυτό παρέχει καθώς και ένα βάρος σπουδαιότητας που εκφράζει πόσο καλά παρέχεται μια διευκόλυνση από ένα ξενοδοχείο. Με βάση την προαναφερθήσα μεθοδολογία θα έχουμε:

Φάση 1: Ορισμός σχήματος ΣΑΠ

- *Βήμα 1*

  Ο χρήστης-διαχειριστής του συστήματος ορίζει το σύνολο οντοτήτων *HotelsEntitySet* και το σύνολο οντοτήτων *FacilitiesEntitySet.* Ορίζει επίσης και το σύνολο συσχετίσεων *HotelFacilitiesRelSet* για τη συσχέτιση του κάθε ξενοδοχείου με τις διευκολύνσεις που παρέχει. Επίσης ορίζει ασαφή χαρακτηριστικά όπως για παράδειγμα το "ακριβό" ("Expensive") ξενοδοχείο.

- *Βήμα 2*

  Κατασκευάζονται αυτόματα από το σύστημα οι πίνακες *Queries*, *Terms* όπως προαναφέρθηκε (τμήμα 4.4).

Φάση 2: Αποκωδικοποίηση υποβαλλόμενων ερωτήσεων

Έστω ότι υπαβάλλεται, από έναν απλό χρήστη, η εξής ερώτηση:

"Τα ξενοδοχεία που έχουν (Indoor Theatre(0.2) ή Video(0.3)) (0.4) και (Air Condition(0.4) ή Swimming pool(0.6)) (0.7)."

Ο χρήστης επιλέγει επίσης το μοντέλο υπολογισμού ομοιότητας ενώ, η υποβαλλόμενη ερώτηση συνοδεύεται και από επιπρόσθετη πληροφορία η οποία στην συγκεκριμένη περίπτωση είναι, το όνομα του συνόλου οντοτήτων των ξενοδοχείων (*HotelsEntitySet*) και το όνομα του συνόλου συσχετίσεων (*HotelFacilitiesRelSet*). Η πληροφορία αυτή στέλνεται αυτόματα, από την εφαρμογή υποβολής ερωτήσεων μαζί με την ερώτηση.

- *Βήμα 1*

  Αποκωδικοποίηση ερώτησης:

  1. Καθορισμός των διαζευκτικών και συζευκτικών τμήματων της. ερώτησης.
  2. Καθορισμός του μοντέλου υπολογισμού ομοιότητας.
  3. Καθορισμός του συνόλου οντοτήτων μέλη του οποίου στοχεύει να ανακτήσει η ερώτηση.
  4. Καθορισμός του συνόλου συσχετίσεων που εκφράζει την σύνδεση των προς ανάκτηση αντικειμένων με τους όρους δεικτοδότησης.

- *Βήμα 2*

Αποθήκευση, από το σύστημα, των διαζευκτικών και συζευκτικών τμημάτων της ερώτησης στους κατάλληλους πίνακες οι οποίοι κατασκευάστηκαν στη **Φάση 1**. Για την συγκεκριμένη ερώτηση οι πίνακες αυτοί θα έχουν ως εξής:

Πίνακας *Queries*

QueryID	ORTermID	Weight
1	1	0.4
1	2	0.7

Πίνακας ο οποίος αποθηκεύει τους διαζευκτικούς όρους για την ερώτηση του παραδείγματος

Πίνακας *Terms*

ORTermID	IndexTermID[5]	Weight
1	Indoor Theatre	0.2
1	Video	0.3
2	Air Conditioned	0.4
2	Swimming pool	0.6

Πίνακας ο οποίος αποθηκεύει τους όρους δεικτοδότησης που σχηματίζουν τους διαζευκτικούς όρους της ερώτησης του παραδείγματος.

Φάση 3: Κατασκευή *SQL* εκφράσεων

Με βάση το μοντέλο υπολογισμού ομοιότητας που προέκυψε από τη **Φάση 2,** *Βήμα 2***:**

- Αυτόματη κατασκευή *SQL* εκφράσεων για τον υπολογισμό των διαζευκτικών τμημάτων (**OR1** και **OR2** όπως περιγραφηκαν στο τμήμα 4.6). Αν για παράδειγμα το μοντέλο υπολογισμού ομοιότητας είναι το *Fuzzy Set* θα έχουμε:

OR1

```
SELECT HotelFacilties.HotelID ,ORTermID,
HotelFacilties.Weight AS W1
```

[5] Στην πραγματικότητα το *IndexTermID* είναι ένας αριθμός, αλλά εδώ δίνεται το όνομα του για ευκολότερη ανάγνωση.

```
FROM HotelFacilties JOIN Terms ON
HotelFacilties.FacilityID = Terms.FacilityID
```

Υπολογισμός πολλαπλασιασμού του διαζευκτικού τμήματος (OR - part)

OR2

```
SELECT HotelID, ORTermID, MAX(W1) AS W
FROM OR1
GROUP BY HotelID, ORTermID
```

Υπολογισμός πρόσθεσης του διαζευκτικού τμήματος (OR - part)

- Αυτόματη κατασκευή *SQL* εκφράσεων για τον υπολογισμό των συζευκτικών τμημάτων (**AND1** και **AND2** όπως περιγράφηκαν στο τμήμα 4.6). Αν για παράδειγμα το μοντέλο υπολογισμού ομοιότητας είναι το *Fuzzy Set* θα έχουμε:

AND1

```
SELECT QueryID, HotelID, OR2.W AS W1
FROM Queries JOIN OR2 ON Queries.ORTermID =
OR2.ORTermID
```

Υπολογισμός πολλαπλασιασμού του συζευκτικού τμήματος (AND - part)

AND2

```
SELECT QueryID, HotelID, MIN(AND1.W1) AS W
FROM AND1
GROUP BY QueryID, HotelID
```

Υπολογισμός πρόσθεσης του συζευκτικού τμήματος (AND - part)

Φάση 4: Εκτέλεση *SQL* εκφράσεων

- *Βημα 1*

  Εκτέλεση, από το σύστημα, των ερωτήσεων που κατασκευάστηκαν στη **Φάση 3.** Οι ερωτήσεις αυτές στέλνονται για εκτέλεση στη βάση δεδομένων.

- *Βημα 2*

  Κατάλληλη μορφοποίηση, από το σύστημα, των αποτελεσμάτων που προέκυψαν και αποστολή τους στην εφαρμογή παρουσίασης των αποτελεσμάτων. Όλα τα ζεύγη *(HotelID, Rank)*, όπου *Rank* είναι ο βαθμός ομοιότητας του κάθε ξενοδοχείου με τη συγκεκριμένη ερώτηση, διαμορφώνονται σε ένα string και αποστέλονται, μέσω socket, στην εφαρμογή παρουσίασης των αποτελεσμάτων.

4.8 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκε η μεθοδολογία εκτέλεσης ερωτήσεων ομοιότητας. Αρχικά περιγράφηκε το επεκταμένο μοντέλο οντοτήτων συσχετίσεων και δόθηκαν οι έννοιες του συνόλου οντοτήτων, του συνόλου συσχετίσεων και των ασαφών χαρακτηριστικών. Στη συνέχεια περιγράφηκε η αναπαράσταση των συνόλων οντοτήτων και συσχετίσεων αλλά και των υποβαλλόμενων ερωτήσεων ως σχέσεις με αξιολόγηση.

Παρουσιάστηκε επίσης η υλοποίηση των σχέσεων αξιολόγησης και της σύνθεσης σχέσεων αξιολόγησης σε σχεσιακή βάση δεδομένων. Περιγράφηκε η εκτέλεση των ερωτήσεων ομοιότητας καθώς και ο τρόπος διαμόρφωσης των ερωτήσεων με βάση το μοντέλο υπολογισμού ομοιότητας. Τα παραπάνω αποτελούν τη μεταφορά σε σχεσιακή βάση δεδομένων των αλγεβρικών εκφράσεων για τον υπολογισμό ερωτήσεων ομοιότητας που παρουσιάστηκαν στο τρίτο κεφάλαιο. Τέλος παρουσιάστηκαν οι συγκεκριμένες φάσεις της μεθοδολογίας εκτέλεσης ερωτήσεων ομοιότητας.

Στο επόμενο κεφάλαιο θα παρουσιαστεί ο **Ε**κδότης **Σ**χήματος (ΕΣ). Ο ΕΣ είναι το εργαλείο το οποίο δίνει τη δυνατότητα ορισμού πάνω σε μια σχεσιακή βάση δεδομένων συνόλων οντοτήτων και συσχετίσεων καθώς και ασαφών χαρακτηριστικών, περιγράφοντας κατ' αυτό τον τρόπο το σχήμα ενός συστήματος ανάκτηση πληροφορίας. Είναι συνεπώς το εργαλείο για την υποστήριξη του επεκταμένου μοντέλου οντοτήτων σχέσεων και υλοποιεί την **Φάση 1** της μεθοδολογίας υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων. Οι

υπόλοιπες φάσης της μεθοδολογίας υλοποιούνται από τη **Μηχανή Ανάκτησης** η οποία θα παρουσιαστεί στο 6ο κεφάλαιο.

5. Ο Εκδότης Σχήματος

5.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο (τμήμα 4.2) περιγράφηκε το επτεκταμένο μοντέλο οντοτήτων συσχετίσεων. Το μοντέλο αυτό αποτελεί μια επέκταση του μοντέλου οντοτήτων συσχετίσεων (Entity Relationship model). Οι βασικές έννοιες σε αυτό είναι η έννοια του ***συνόλου οντοτήτων (entity set)*** η έννοια του ***συνόλου συσχετίσεων (relationship set)*** και έννοια των ***ασαφών χαρακτηριστικών (fuzzy attributes)*** όπως αυτές περιγραφήκαν στο τμήμα 4.2. Δόθηκαν επίσης οι έννοιες της ***συλλογής*** και των ***όρων δεικτοδοτησης***. Παρουσιάστηκε επίσης μια υλοποίηση αυτού του μοντέλου πάνω στο σχεσιακό μοντέλο έτσι ώστε να υποστηριχθεί από υπάρχουσες βάσεις δεδομένων.

*Το εργαλείο με το οποίο ορίζονται πάνω σε μια σχεσιακή βάση δεδομένων σύνολα οντοτήτων, σύνολα συσχετίσεων και ασαφή χαρακτηριστικά είναι ο **Εκδότης Σχήματος** (ΕΣ).* Τα σύνολα οντοτήτων, συσχετίσεων και τα ασαφή χαρακτηριστικά τα οποία ορίζονται με τη χρήση του ΕΣ αποτελούν την περιγραφή του κατασκευαζόμενου συστήματος ανάκτησης πληροφορίας και βρίσκονται πάνω από τη σχεσιακή βάση δεδομένων στη οποία δομείται το σύστημα. Με τη χρήση του ΕΣ υλοποιείται η **Φάση 1** της μεθοδολογίας υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων, γίνεται δηλαδή τόσο ο ορισμός του σχήματος του

συστήματος ανάκτησης πληροφορίας μέσω του ορισμού συνόλων οντοτήτων, συνόλων συσχετίσεων και χαρακτηριστικών (*Βήμα 1* της **Φασης 1**) όσο και. Η κατασκευή αναγκαίων πινάκων για την αποθήκευση των ερωτήσεων (*Βήμα 2* της **Φάσης 1**).

Το αποτέλεσμα της όλης διαδικασίας πρέπει να εκφράζεται με τρόπο απλό και κατανοητό έτσι ώστε το τμήμα αυτό του συστήματος να αποτελεί μια ανεξάρτητη μονάδα (module). Με γνώμονα τα παραπάνω, το αποτέλεσμα της όλης διαδικασίας, η έξοδος δηλαδή του τμήματος αυτού του συστήματος, αποθηκεύεται σε πίνακες της βάσης δεδομένων πάνω στην οποία οικοδομείται το σύστημα ανάκτησης πληροφορίας. Οι πίνακες αυτοί περιγράφουν το κατασκευασθέν σύστημα ανάκτησης πληροφορίας και θα χρησιμοποιηθούν από κάποιο άλλο τμήμα του συστήματος, συγκεκριμένα τη Μηχανή Ανάκτησης.

Έτσι, στο επόμενο τμήμα θα παρουσιαστεί η έννοια του συνόλου οντοτήτων η οποία είναι θεμελιώδης στο σύστημα. Στο τμήμα 5.3 θα παρουσιαστεί η επίσης σημαντική έννοια του χαρακτηριστικού, ενώ στο επόμενο τμήμα (5.4) θα δοθεί η έννοια της συσχέτισης και της έκφρασης αξιολόγησης. Στο επομένο τμήμα (5.5) θα αναφερθούν θέματα υλοποίησης. Μια σύντομη περίληψη θα δοθεί στο τμήμα 5.6.

5.2 Σύνολα οντοτήτων (entity sets)

Ένα σύνολο οντοτήτων ορίζεται πάνω σε ένα πίνακα της σχεσιακής βάσης δεδομένων στην οποία θα δομηθεί το σύστημα ανάκτησης πληροφορίας. Τα μέλη του συνόλου είναι οι πλειάδες (tuples) του πίνακα και συνεπώς, η ταυτότητα κάθε μέλους του συνόλου θα είναι το πρωτεύων κλειδί (primary key) του πίνακα. Το γεγονός ότι ορίζεται ένα σύνολο οντοτήτων πάνω σε ένα πίνακα της βάσης δεδομένων, προσφέρει τη δυνατότητα εκτέλεσης ερωτήσεων ομοιότητας ως προς τα δεδομένα που περιέχει αυτός ο πίνακας.

Στο παράδειγμα των ξενοδοχείων θεωρούμε ότι υπάρχει ένας πίνακας *Hotels(HotelID,…)* πάνω στον οποίο ορίζεται το σύνολο οντοτήτων *HotelsEntitySet* διότι οι ερωτήσεις θα αποσκοπούν στην ανάκτηση ξενοδοχείων τα οποία ικανοποιούν κάποιες προυποθέσεις. Επίσης ορίζεται το σύνολο οντοτήτων *FacilitiesEntitySet* πάνω στον πίνακα *Facilities(FacilityID,…)* διότι οι διευκολύνσεις που παρέχονται από κάθε ξενοδοχείο "περιγράφουν" κατά κάποιο τρόπο το ξενοδοχείο και συνεπώς

μπορούν να θεωρηθούν όροι δεικτοδότησης. Βεβαίως, είναι δυνατόν οι ρόλοι να αντιστραφούν και οι διευκολύνσεις να ανακτώνται βάσει ξενοδοχείων.

Μια χαρακτηριστική ερώτηση ομοιότητας με όρους δεικτοδότησης διευκολύνσεις είναι για παράδειγμα η εξής:

"Ποια είναι τα ξενοδοχεία που έχουν Color T.V (0.2) και Swimming pool (0.8)"

Και ένα παράδειγμα αντίστροφης ερώτησης είναι:

"Ποιες είναι οι καλύτερες διευκολύνσεις των ξενοδοχείων Πανόραμα (0.5) και Σαμαριά (0.5)"

5.2.1 Ορισμός συνόλου οντοτήτων πάνω σε άποψη

Σε μια προσπάθεια να δοθεί μεγαλύτερη ευελιξία στον ορισμό ενός συνόλου οντοτήτων, παρέχεται η δυνατότητα ορισμού ενός συνόλου οντοτήτων πάνω σε μια προκαθορισμένη, και αποθηκευμένη στη βάση δεδομένων, ερώτηση. Ουσιαστικά ορίζεται πάνω σε μια άποψη (view) της βάσης δεδομένων. Το σημείο αυτό χρήζει ιδιαίτερης προσοχής για δύο λόγους.

Ο πρώτος είναι ότι η έννοια της άποψης θα χρησιμοποιηθεί κατά κόρον, τόσο σε αυτό το κεφάλαιο, όσο και στα επόμενα. Τα περισσότερα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων παρέχουν την δυνατότητα ορισμού απόψεων και τη χρήση τους σε μια ερώτηση σαν να ήταν κανονικοί πίνακες της βάσης δεδομένων.

Ο δεύτερος λόγος είναι ακόμα σπουδαιότερος. Η δυνατότητα να ορίζεται ένα νέο σύνολο οντοτήτων πάνω σε μια άποψη πολλαπλασιάζει τις δυνατότητες του ΕΣ και δίνει μεγάλη ευελιξία στην κατασκευή του συστήματος ανάκτησης πληροφορίας. Μπορεί να δημιουργηθεί μια άποψη η οποία συνδυάζει περισσότερους του ενός πίνακες, και κατά συνέπεια συνδυάζει πληροφορία η οποία βρίσκεται διασκορπισμένη σε διάφορους πίνακες της βάσης δεδομένων.

5.2.2 Άμεσες συνέπειες του ορισμού ενός συνόλου οντοτήτων

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο (τμήμα 3.5) οι ερωτήσεις που αναγνωρίζονται από το σύστημα βρίσκονται σε κανονική συζευκτική μορφή και μπορούν να αναπαρασταθούν ως δύο σχέσεις με αξιολόγηση. Ορίζεται δηλαδή η σχέση αξιολόγησης *Queries* από το σύνολο των ερωτήσεων στο σύνολο των διαζευκτικών όρων των ερωτήσεων καθώς και η σχέση αξιολόγησης *Terms* από το σύνολο των διαζευκτικών όρων των ερωτήσεων στο σύνολο των όρων δεικτοδότησης, οι οποίοι χρησιμοποιούνται στο σχηματισμό των ερωτήσεων.

Απαιτείται συνεπώς ο ορισμός ενός πίνακα ο οποίος αποθηκεύει τη σχέση αξιολόγησης *Queries* ανάμεσα στις ερωτήσεις και στα διαζευκτικά των ερωτήσεων καθώς και ένας πίνακας ο οποίο θα αποθήκευει για κάθε ερώτηση τη σχέση αξιολόγησης *Terms* ανάμεσα στα διαζευκτικά της ερώτησης και στους όρους που σχηματίζουν αυτά τα διαζευκτικά και οι οποίοι θα είναι φυσικά όροι δεικτοδότησης.

Ο πίνακας στον οποίο θα αποθηκεύεται η σχέση αξιολόγησης *Queries* όνομαζεται *SYS_Queries*, είναι κοινός για όλες τις υποβαλλόμενες ερωτήσεις και κατασκευάζεται αυτόματα από τον ΕΣ. Το σχήμα αυτού του πίνακα έχει ως εξής:

Πίνακας *SYS_Queries*

Όνομα πεδίου	Τύπος
QueryID	Integer
ORTermID	Integer
Weight	Double

Πίνακας ο οποίος χρησιμοποιείται για την αποθήκευση της σχέσης αξιολόγησης μεταξύ του συνόλου των ερωτήσεων και του συνόλου των διαζευκτικών όρων μια ερώτησης. Το πρωτεύων κλειδί του πίνακα είναι ο συνδυασμός των QueryID και ORTermID

Όπου,

- *QueryID* είναι η ταυτότητα κάθε ερώτησης η οποία αποδίδεται αυτόματα από το σύστημα για κάθε ερώτηση.

- *ORTermID* είναι η ταυτότητα του κάθε διαζευκτικού όρου *(OR-Term)* της ερώτησης και αποδίδεται επίσησης αυτόματα από το σύστημα για κάθε διαζευκτικό όρο.
- *Weight* είναι το βάρος σπουδαιότητας που μπορεί να έχει κάθε διαζευκτικός όρος της ερώτησης εφ' όσον φυσικά οι διαζευκτικοί όροι έχουν κάποια βάρη.
- Το πρωτεύων κλειδί του παραπάνω πίνακα είναι ο συνδυασμός των *QueryID* και *ORTermID*.

Ο πίνακας όμως που αποθηκεύει τη σχέση ανάμεσα στα διαζευκτικά μιας ερώτησης και τους όρους δεικτοδότησης που σχηματίζουν την ερώτηση δεν μπορεί να είναι κοινός διότι συνήθως υπάρχουν περισσότερα του ενός υποσύνολα όρων δεικτοδότησης, τα οποία περιγράφουν αντίστοιχα υποσύνολα αντικειμένων προς ανάκτηση. Για κάθε τέτοιο υποσύνολο ορίζεται και ο αντίστοιχος πίνακας. Στο παράδειγμα των ξενοδοχείων υπάρχει το υποσύνολο των διευκολύνσεων που χρησιμοποιούνται ως όροι δεικτοδότησης του υποσυνόλου των ξενοδοχείων. Όταν όμως οι διευκολύνσεις θεωρηθούν αντικείμενα προς ανάκτηση τότε ως υποσύνολο όρων δεικτοδότησης θεωρείται το σύνολο των ξενοδοχείων.

Ο ορισμός συνεπώς ενός συνόλου οντοτήτων, το οποίο στην πραγματικότητα αντιστοιχεί είτε σε ένα υποσύνολο του συνόλου των όρων δεικτοδότησης είτε σε ένα υποσύνολο του συνόλου των προς ανάκτηση αντικειμένων, έχει ως επιπρόσθετο αποτέλεσμα τη δημιουργία ενός νέου πίνακα στη βάση δεδομένων. Ο πίνακας αυτός θα χρησιμοποιηθεί αργότερα από τη Μηχανή Ανάκτησης προκειμένου να απαντηθεί μια ερώτηση ομοιότητας.

Η γενική μορφή του πίνακα θα έχει ως εξής:

Πίνακας *SYS_EntitySetName*:

Όνομα πεδίου	Τύπος
ORTermID	Integer
A1	Type1
…	…
An	Typen
Weight	Double

Πίνακας ο οποίος δημιουργείται από τον ΕΣ όταν ορίζεται ένα νέο σύνολο οντοτήτων και χρησιμοποιείται για να αποθηκεύονται οι διαζευκτικοί όροι μιας ερώτησης.Το πρωτεύων κλειδί του πίνακα είναι ο συνδυασμός των ORTerm και A1,…, An.

Όπου:

- Το πεδίο *ORTermID* είναι η ταυτότητα του κάθε διαξευκτικού όρου μιας ερώτησης και αποδίδεται αυτόματα από το σύστημα.
- Τα πεδία A1,…,An είναι τα πεδία του πρωτεύοντος κλειδιού του πίνακα πάνω στον οποίο ορίζεται το σύνολο οντοτήτων.
- Το πεδίο *Weight* περιέχει το βάρος σπουδαιότητας του κάθε όρου της ερώτησης σε σχέση με τους υπόλοιπους όρους της ερώτησης.
- Το πρωτεύων κλειδί του πίνακα είναι ο συνδυασμός των *ORTerm* και *A1,…, An*.

Όλα τα παραπάνω γίνονται αυτόματα, κατά τον ορισμό ενός συνόλου οντοτήτων, και δεν είναι ορατά από τον χρήστη-διαχειριστή. Ο χρήστης-διαχειριστής δίνει μόνο το όνομα του συνόλου οντοτήτων και επιλέγει τον πίνακα, ή την ερώτηση (άποψη), πάνω στον (-ην) οποίο (-α) ορίζεται το σύνολο αυτό.

Ο πίνακας αυτός όπως και ο πίνακας *SYS_Queries* αποτελεί τμήμα του συστήματος και ανήκει σε ένα σύνολο πινάκων και απόψεων οι οποίοι λέγονται πίνακες και απόψεις συστήματος, αντίστοιχα. Όλοι οι πίνακες και οι απόψεις συστήματος έχουν στο όνομα τους το πρόθεμα *"SYS_"* και θα χρησιμοποιηθούν, αργότερα, από την Μηχανή Ανάκτησης στη διαδικασία απάντησης των ερωτήσεων. Πρέπει συνεπώς να γίνει σαφής διάκριση ανάμεσα στους πίνακες και τις απόψεις του συστήματος, οι οποίοι κατασκευάζονται από τον ΕΣ, και στους πίνακες και τις

ερωτήσεις της βάσης δεδομένων πάνω στην οποία χτίζεται το σύστημα ανάκτησης πληροφορίας και οι οποίοι, αφενός προϋπάρχουν στη βάση δεδομένων και αφετέρου περιέχουν τα δεδομένα καθευατά για τα οποία γίνεται προσπάθεια ανάκτησής τους.

Προκειμένου τα παραπάνω να γίνουν περισσότερο κατανοητά στο παράδειγμα των ξενοδοχείων τα δύο σύνολα οντοτήτων (ξενοδοχεία και διευκολύνσεις) οδηγούν στον ορισμό των πινάκων που δίνονται στη συνέχεια.

Έστω λοιπόν ότι ορίζεται το σύνολο οντοτήτων *HotelsEntitySet* πάνω στον πίνακα *Hotels(HotelID,…)* και το σύνολο οντοτήτων *FacilitiesEntitySet* πάνω στον πίνακα *Facilities(FacilityID,…)*. Τότε οι πίνακες που κατασκευάζονται θα έχουν ως εξής :

Πίνακας *SYS_HotelsEntitySet*:

Όνομα πεδίου	**Τύπος**
ORTermID	Integer
HotelID	Integer
Weight	Double

Πίνακας *SYS_FacilitiesEntitySet*:

Όνομα πεδίου	**Τύπος**
ORTermID	Integer
FacilityID	Integer
Weight	Double

Επίσης κατασκευάζεται και ο επόμενος πίνακας:

Πίνακας *SYS_Queries*

Όνομα πεδίου	**Τύπος**
QueryID	Integer
ORTermID	Integer
Weight	Double

Ας υποθέσουμε τώρα ότι υποβάλλεται η εξής ερώτηση :

"Τα ξενοδοχεία που έχουν (Indoor Theatre(0.2) ή Video(0.3)) (0.4) και (Air Condition(0.4) ή Swimming pool(0.6)) (0.7)."

Η λογική έκφραση αυτής της ερώτησης έχει ως εξής:

((Indoor Theatre, 0.2 OR Video, 0.3), 0.4) AND
((Air Condition, 0.4 OR Swimming pool, 0.6),0.7)

Στη συγκεκριμένη περίπτωση οι πίνακες *SYS_Queries* και *SYS_FacilitiesEntitySet* θα έχουν ως εξής:

Πίνακας *SYS_Queries*

QueryID	ORTermID	Weight
1	1	0.4
1	2	0.7

Πίνακας *SYS_FacilitiesEntity*

ORTermID	FacilityID[1]	Weight
1	Indoor Theatre	0,2
1	Video	0,3
2	Air Condition	0,4
2	Swimming Pool	0,6

Ας υποθέσουμε τώρα ότι υποβάλλεται η ερώτηση:

"Ποιες είναι οι καλύτερες διευκολύνσεις των ξενοδοχείων (Πανόραμα (0.5) ή Ομαλός (0,6)) (0.3) και (Σαμαριά (0.5)) (0.8)"

Η λογική έκφραση αυτής της ερώτησης έχει ως εξής :

((Πανόραμα, 0.5 OR Ομαλός, 0.6), 0.3) AND
((Σαμαριά, 0,5), 0.8)

[1] Στην πραγματικότητα το *FacilityID* είναι ένας αριθμός, αλλά εδώ δίνεται το όνομα του Facility για ευκολότερη ανάγνωση.

Στη συγκεκριμένη περίπτωση οι πίνακες *SYS_Queries* και *SYS_HotelsEntitySet* θα έχουν ως εξής:

Πίνακας *SYS_Queries*

QueryID	ORTermID	Weight
1	1	0.4
1	2	0.7
2	3	0.3
2	4	0.8

Πίνακας *SYS_HotelsEntity*

ORTermID	HotelID[2]	Weight
3	Πανόραμα	0.5
3	Ομαλός	0.6
4	Σαμαριά	0.5

5.3 Ασαφή χαρακτηριστικά (fuzzy attributes)

Τα ασαφή χαρακτηριστικά (fuzzy attributes*)* όπως αναφέρθηκε (τμήμα 4.2.2) ισοδυναμούν με ιδιότητες των οντοτήτων οι οποίες για κάθε οντότητα έχουν μια τιμή σπουδαιότητας (ή σημασίας) από το 0 (μηδενική σπουδαιότητα) μέχρι το 1 (μέγιστη σπουδαιότητα). Τα ασαφή χαρακτηριστικά υπολογίζονται με βάση κάποια έκφραση υπολογισμού (evaluation expression), η οποία συγκροτείται από πεδία του πίνακα πάνω στον οποίο ορίζεται ένα σύνολο οντοτήτων, και κάποια συνάρτηση αξιολόγησης (ranking function)

Από τη στιγμή που θα οριστεί κάποιο χαρακτηριστικό πάνω σε ένα σύνολο οντοτήτων κάθε μέλος του συνόλου έχει μια τιμή ως προς αυτό το χαρακτηριστικό. Αυτό το νέο χαρακτηριστικό μπορεί να θεωρηθεί ως ένα καινούργιο πεδίο του πίνακα πάνω στον οποίο ορίζεται το σύνολο οντοτήτων του χαρακτηριστικού.

[2] Στην πραγματικότητα το *HotelID* είναι ένας αριθμός, αλλά εδώ δίνεται το όνομα του ξενδοχείου για ευκολότερη ανάγνωση.

5.3.1 Υπολογισμός των ασαφών χαρακτηριστικών μιας οντότητας

Όπως προαναφέρθηκε, για κάθε μέλος ενός συνόλου οντοτήτων υπάρχει η αντίστοιχη τιμή για κάθε ασαφές χαρακτηριστικό οριζόμενο πάνω σ' αυτό το σύνολο. Ο υπολογισμός της τιμής του κάθε ασαφούς χαρακτηριστικού για κάθε μέλος του συνόλου οντοτήτων ενσωματώνεται (ως έκφραση) σε μια άποψη που δημιουργεί αυτόματα ο ΕΣ πάνω στον πίνακα που αντιστοιχεί στο σύνολο οντοτήτων, λαμβάνοντας φυσικά υπ' όψιν, τόσο την έκφραση υπολογισμού του ασαφούς χαρακτηριστικού, όσο και την συνάρτηση αξιολόγησής του. Η άποψη αυτή χρησιμοποιείται αργότερα από την Μηχανή Ανάκτησης.

Στο παράδειγμα των ξενοδοχείων έστω ότι ορίζεται το σύνολο οντοτήτων *HotelsEntitySet* πάνω στον πίνακα *Hotels* ο οποίος έχει το ακόλουθο σχήμα:

Πίνακας *Hotels*

Όνομα πεδίου	Τύπος	Περιγφραφή
HotelID	Integer	Το πρωτεύων κλειδί του πίνακα
Name	Text	Το όνομα του ξενοδοχείο
SingleRoomPrice	Integer	Η τιμή του μονόκλινου δωματίου
Category	Byte	Η κατηγορία του ξενοδοχείου. Μπορεί να πάρει μια από τις εξής τιμές 1,2,3,4

Μια ενδιαφέρουσα ερώτηση θα ήταν η εξής:

"Ποια είναι τα ξενοδοχεία που δεν είναι ακριβά"

Η απαίτηση του χρήστη να ανακτηθούν τα ξενοδοχεία που δεν είναι ακριβά καθιστά αναγκαίο τον ορισμό ενός "ακριβού ξενοδοχείου". Για παράδειγμα ένα ακριβό ξενοδοχείο θα μπορούσε να ήταν αυτό στο οποίο ένα μονόκλινο δωμάτιο κοστίζει πάνω από 30000 δραχμές. Αυτός ο ορισμός όμως δεν είναι απόλυτα σωστός διότι το αν ένα ξενοδοχείο είναι ακριβό δεν εξαρτάται μόνο από την τιμή ενός μονόκλινου δωματίου αλλά εξαρτάται και, για παράδειγμα, από την κατηγορία στην οποία ανήκει το ξενοδοχείο. Έτσι για παράδειγμα ένα ξενοδοχείου πρώτης κατηγορίας με τιμή δωματίου 30000 δραχμές δεν μπορεί να χαρακτηριστεί ακριβό

ενώ ένα ξενοδοχείο τρίτης κατηγορίας με τιμή δωματίου 20000 δραχμές μπορεί να χαρακτηριστεί ως ακριβό.

Μπορούμε συνεπώς να ορίσουμε ένα νέο χαρακτηριστικό (attribute), το "ακριβό" ("Expensive") το οποίο θα υπολογίζεται ως εξής:

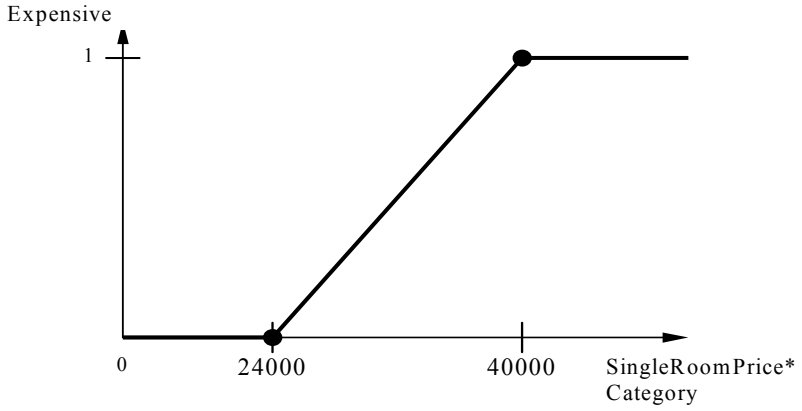
*Expensive: SingleRoomPrice * Category*

Τα SingleRoomPrice, Category είναι πεδία του πίνακα Hotels που προαναφέρθηκε. Συνέπως το νέο χαρακτηριστικό Expensive ορίζεται πάνω στον πίνακα Hotels και υπολογίζεται με βάση τα πεδία SingleRoomPrice και Category του πίνακα αυτού. Το χαρακτηριστικό Expensive μπορεί να χαρακτηριστεί σαν υπολογιζόμενο χαρακτηριστικό (computed attribute) του οποίου η έκφραση υπολογισμού *EvaluationExpression* το γινόμενο της τιμής του δωματίου με την κατηγορία του ξενοδοχείου.

Με τον ορισμό του χαρακτηριστικού Expensive ένα ξενοδοχείο θα μπορούσε να θεωρηθεί ακριβό όταν η τιμή του χαρακτηριστικού Expensive είναι, για παράδειγμα, μεγαλύτερη του 40000. Όμως υπάρχει περίπτωση να θεωρηθεί ακριβό ένα ξενοδοχείο και όταν η τιμή του χαρακτηριστικού Expensive είναι μικρότερη του 40000, γεγονός που εξαρτάται από την περιοχή που βρίσκεται, τις διευκολύνσεις που παρέχει κ.λ.π. Με άλλα λόγια ένα ξενοδοχείο για το οποίο η τιμή του χαρακτηριστικού Expensive ειναι μεγαλύτερη του 40000 θεωρείται σίγουρα ακριβό, ενώ για τιμές μικρότερες του 40000 μπορεί να θεωρηθεί ακριβό αλλό όχι με απόλυτη βεβαιότητα. Τέλος στο παράδειγμά μας ας θεωρήσουμε ακόμη ότι αν η τιμή του χαρακτηριστικού Expensive ειναι μικρότερη από 24000, τότε το ξενοδοχείο δεν είναι σίγουρα ακριβό.

Η αποτύπωση αυτών των δεδομένων σε μια συνάρτηση αξιολόγησης των ξενοδοχείων που βασίζεται στην τιμή του χαρακτηριστικού Expensive - το οποίο με τη σειρά του βασίζεται στα πεδία SingleRoomPrice και Category του πίνακα Hotels - μπορεί να δωθεί με μια συνάρτηση η οποία σε κάθε τιμή του χαρακτηριστικού Expensive αντιστοιχεί έναν αριθμό στο διάστημα [0,1]. Το 0 σημαίνει ότι ένα ξενοδοχείο σίγουρα δεν είναι ακριβό ενώ το 1 ότι σίγουρα είναι. Άρα για τιμές του χαρακτηριστικού Expensive στο διάστημα [0,24000] η συνάρτηση αυτή πρέπει να δίνει την τιμή 0 ενώ για τιμές στο διάστημα $[40000, \infty]$ πρέπει να δίνει 1. Στο διάστημα [24000, 40000] μπορεί να θεωρηθεί, για παράδειγμα, ότι η συνάρτηση

αυξάνει γραμμικά από το 0 στο 1. Μια γραφική απεικόνιση της συνάρτησης αξιολόγησης του χαρακτηριστικού Expensive έχει όπως φαίνεται στο επόμενο σχήμα.



Γραφική απεικόνιση της συνάρτησης αξιολόγησης του χαρακτηριστικού Expensive που ορίζεται χρησιμοποιώντας τα πεδία SingleRoomPrice και Category του πίνακα Hotels και αξιολογείται με βάση την απλή γραμμική συνάρτηση με παραμέτρους α=24000 και.b=40000

Το χαρακτηριστικό Expensive είναι ένα ασαφές χαρακτηριστικό. Επίσης η έκφραση υπολογισμού του *(EvaluationExpression)* είναι το γινόμενο της τιμής του δωματίου με την κατηγορία του ξενοδοχείου ενώ, η συνάρτηση αξιολόγησης (*RankingFunction)* που χρησιμοποιείται γι' αυτό είναι η γραμμική συνάρτηση. Το χαρακτηριστικό διαδραματίζει το ρόλο όρου δεικτοδότησης και μπορούν να χρησιμοποιηθεί για το σχηματισμό ερωτήσεων.

5.4 Σύνολα συσχετίσεων (relationship set)

Ένα σύνολο συσχετίσεων συνδέει δύο σύνολα οντοτήτων και ορίζεται είτε πάνω σε ένα πίνακα είτε πάνω σε μια άποψη. Στον πίνακα ή στην άποψη αποθηκεύεται ο τρόπος με τον οποίο συνδέονται τα μέλη των δύο συνόλων οντοτήτων μεταξύ των οποίων ορίζεται το σύνολο συσχετίσεων.

5.4.1 Ορισμός συνόλου συσχετίσεων πάνω σε πίνακα

Ένα σύνολο συσχετίσεων ορίζεται για να δηλωθεί η ύπαρξη μιας σύνδεσης μεταξύ των μελών δύο συνόλων οντοτήτων. Η σύνδεση αυτή εκφράζει στην πραγματικότητα μια αντιστοίχιση αντικειμένων με τους όρους δεικτοδότησής τους συμπεριλαμβάνοντας επίσης και την ύπαρξη βαρών που δηλώνουν τη σπουδαιότητα κάθε όρου ως προς κάθε αντικείμενο. Η αντιστοίχιση αυτή αποθηκεύεται στον πίνακα πάνω στον οποίο ορίζεται η συσχέτιση.

Στο παράδειγμα των ξενοδοχείων έστω ότι ορίζεται το σύνολο οντότητων *HotelsEntitySet* πάνω στον πίνακα *Hotels(HotelID,…)* και το σύνολο οντοτήτων *FacilitiesEntitySet* πάνω στον πίνακα *Facilities(FacilityID,…)*. Ορίζεται τότε το συσχετίσεων *HotelFacilitiesRelationshipSet* πάνω στο πίνακα *HotelFacilities(HotelID, FacilityID,Weight,…)*. Το σύνολο συσχετίσεων αυτό συνδέει κάθε ξενοδοχείο με τις διευκολύνσεις που αυτό παρέχει και μάλιστα βαθμολογεί την κάθε παρεχόμενη από το ξενοδοχείο διευκόλυνση.

5.4.2 Ορισμός συνόλου συσχετίσεων πάνω σε άποψη

Σε μια προσπάθεια να δοθεί μεγαλύτερη ευελιξία στον ορισμό ενός συνόλου συσχετίσεων παρέχεται η δυνατότητα ορισμού του πάνω σε μια προκαθορισμένη και αποθηκευμένη στη βάση δεδομένων ερώτηση, ουσιαστικά ορίζεται πάνω σε μια άποψη (view) της βάσης δεδομένων (όπως συμβαίνει και για τα σύνολα οντοτήτων).

Η πρακτική σημασία του γεγονότος αυτού είναι ότι ο πίνακας στο οποίο αποθηκεύεται ο τρόπος με τον οποίο συσχετίζονται τα μέλη δύο συνόλων οντοτήτων δεν είναι αναγκαίο να βρίσκεται ήδη ορισμένος στη βάση δεδομένων. Ο πίνακας αυτός μπορεί να προκύψει σαν το αποτέλεσμα της εκτέλεσης μιας *SQL* ερώτησης.

Αποθηκεύοντας την ερώτηση αυτή είναι δυνατόν στην συνέχεια να οριστεί

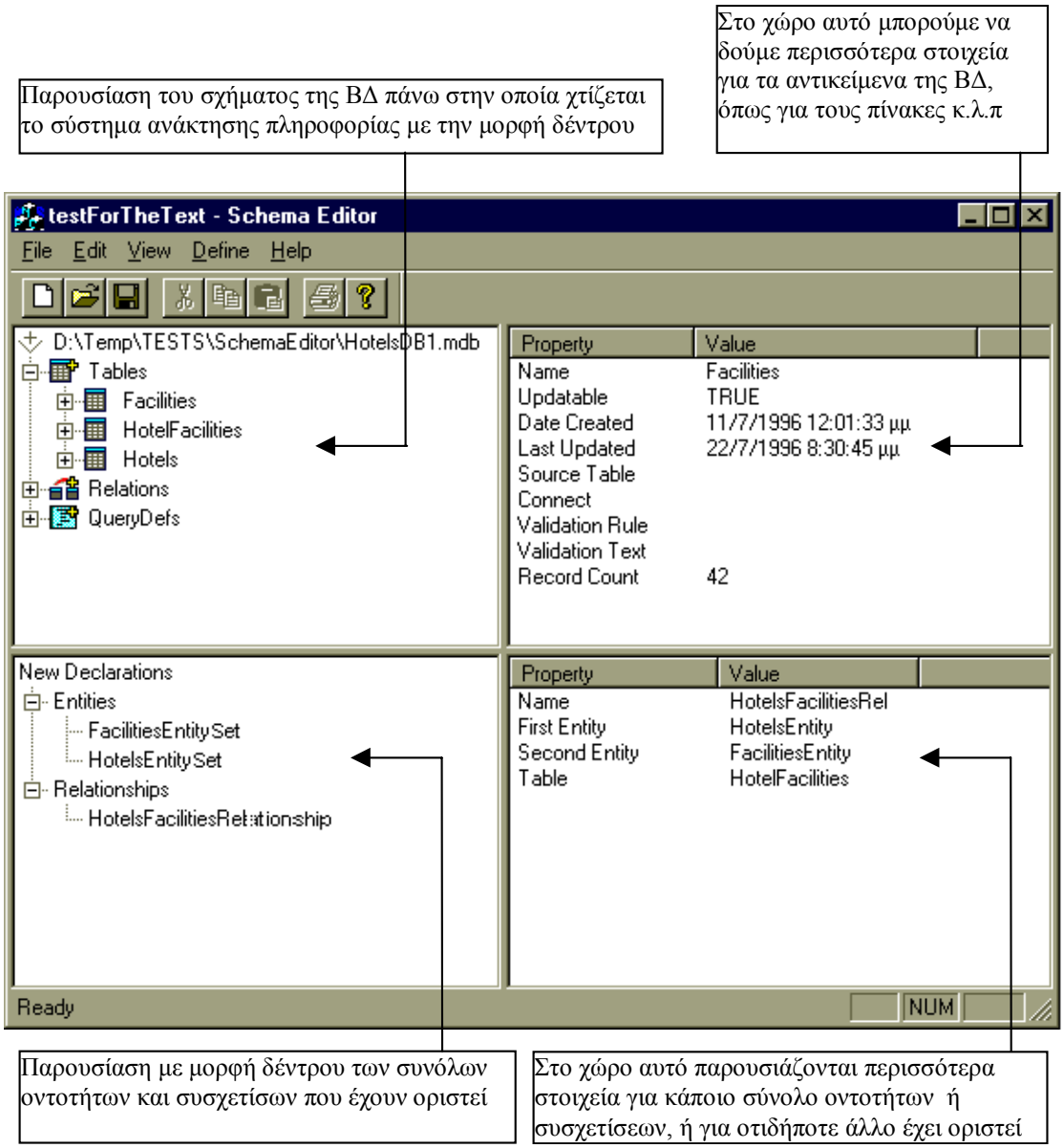
πάνω της ένα σύνολο συσχετίσεων, όπως ακριβώς ορίζεται πάνω σ' ένα πίνακα. Με τον τρόπο αυτό αντιμετωπίζονται τυχόν περιορισμοί που μπορεί να προέρχονται από το σχήμα της βάσης δεδομένων πάνω στην οποία οικοδομείται το σύστημα ανάκτησης πληροφορίας.

5.5 Θέματα υλοποίησης

Κατ' αρχήν, η βάση δεδομένων πάνω στην οποία οικοδομήθηκε το όλο σύστημα είναι κατασκευασμένη με το πρόγραμμα Microsoft Access 97. Το πρόγραμμα αυτό είναι εύχρηστο και καλύπτει πλήρως τις απαιτήσεις που τέθηκαν ως προς την υλοποίηση.

Ο εκδότης σχήματος κατασκευάστηκε με τη χρήση του πακέτου Microsoft Visual C++ 5.0. Το πακέτο αυτό παρέχει ένα σύνολο κλάσεων για την κατασκευή επιπέδων επαφής-ζεύξης (interfaces) καθώς και για την υποστήριξη λειτουργιών σε βάσεις δεδομένων. Ειδικότερα για την εργασία με βάσεις δεδομένων παρέχει την τεχνολογία Data Access Object (DAO) [12, 13], η οποία παρέχει πολλές διευκολύνσεις όσον αφορά την αλληλεπίδραση με βάσεις δεδομένων και επιπλέον, συνεργάζεται άψογα με την Access. Δεν θα αναφερθούν περισσότερα διότι τότε, θα χρειαζόταν να παρουσιαστούν λεπτομέρειες οι οποίες είναι μεν πολύ σημαντικές αλλά δεν μπορούν να συμπεριληφθούν σε ένα τέτοιο κείμενο.

Ένα στιγμιότυπο του ΕΣ βασιζόμενο πάνω στο προηγούμενο παράδειγμα ακολουθεί στην επόμενη σελίδα :



Εικόνα 5.1: Στιγμιότυπο του εκδότη σχήματος όπου φαίνονται: το σχήμα της ΒΔ καθώς και τα σύνολα οντοτήτων και συσχετίσεων τα οποία αναφέρθηκαν στο τελευταίο παράδειγμα.

5.6 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκε ο **Ε**κδότης **Σ**χήματος. Ο ΕΣ είναι το τμήμα εκείνο του συστήματος το οποίο χρησιμοποιείται για τον προσδιορισμό του συνόλου των προς ανάκτηση αντικειμένων καθώς και του συνόλου των όρων δεικτοδότησης του υπό κατασκευή συστήματος ανάκτησης πληροφορίας. Ο ΕΣ προβαίνει επίσης και στην κατασκευή μιας σειράς πινάκων συστήματος οι οποίοι χρησιμεύουν στην εκτέλεση των ερωτήσεων ομοιότητας. Ο ΕΣ χρησιμοποιείται για την υλοποίηση της **Φάσης 1** της μεθοδολογίας υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων (τμήμα 4.8).

Μια από τις βασικές έννοιες που παρουσιάστηκαν ήταν αυτή του συνόλου οντοτήτων, το οποίο αναπαριστά τα προς ανάκτηση αντικείμενα και τους όρους δεικτοδότησης. Επίσης δόθηκε η έννοια των ασαφούς χαρακτηριστικών (fuzzy attributes) τα οποία ισοδυναμούν με ιδιότητες των οντοτήτων οι οποίες, για κάθε οντότητα έχουν μια τιμή σπουδαιότητας (ή σημασίας) από το 0 (μηδενική σπουδαιότητα) μέχρι το 1 (μέγιστη σπουδαιότητα)

Τέλος παρουσιάστηκε η έννοια του συνόλου συσχετίσεων (relationship set) και της έκφρασης αξιολόγησης (ranking expression). Ένα σύνολο συσχετίσεων ορίζεται πάνω σε ένα πίνακα ή σε μια άποψη και συνδέει δύο σύνολα οντοτήτων. Η συσχέτιση αποτελή την μεταφορά της έννοιας της σχέσης με αξιολόγηση στο επεκταμένο μοντέλο οντοτήτων σχέσεων και πρακτικά αναπαριστά την σύνδεση αντικειμένων-όρων δεικτοδότησης ενώ, η έκφραση αξιολόγησης συμβάλλει στην μετατροπή του πεδίου το οποίο αναπαριστά το βάρος σπουδαιότητας των όρων δεικτοδότησης ως προς τα αντικείμενα που περιγράφουν σε ασαφές (fuzzy).

Στο επόμενο κεφάλαιο θα παρουσιαστεί η **Μ**ηχανή **Α**νάκτησης. Η ΜΑ είναι το τμήμα εκείνο του συστήματος το οποίο αναλαμβάνει την εκτέλεση των ερωτήσεων. Στην ουσία λοιπόν η ΜΑ υλοποιεί τη **Φάση 2**, τη **Φάση 3** και τη **Φάση 4** της μεθοδολογίας υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων (τμήμα 4.8)

6. Η Μηχανή Ανάκτησης

6.1 Εισαγωγή

Μετά τον ορισμό τον ορισμό των συνόλων οντοτήτων και συσχετίσεων καθώς και των ασαφών χαρακτηριστικών, αλλά, και την κατασκευή των αναγκαίων πινάκων με τη χρήση του **Ε**κδότη **Σ**χήματος, όπως περιγράφηκε στο προηγούμενο κεφάλαιο το επόμενο βήμα είναι η υποστήριση της εκτέλεσης των ερωτήσεων ομοιότητας.

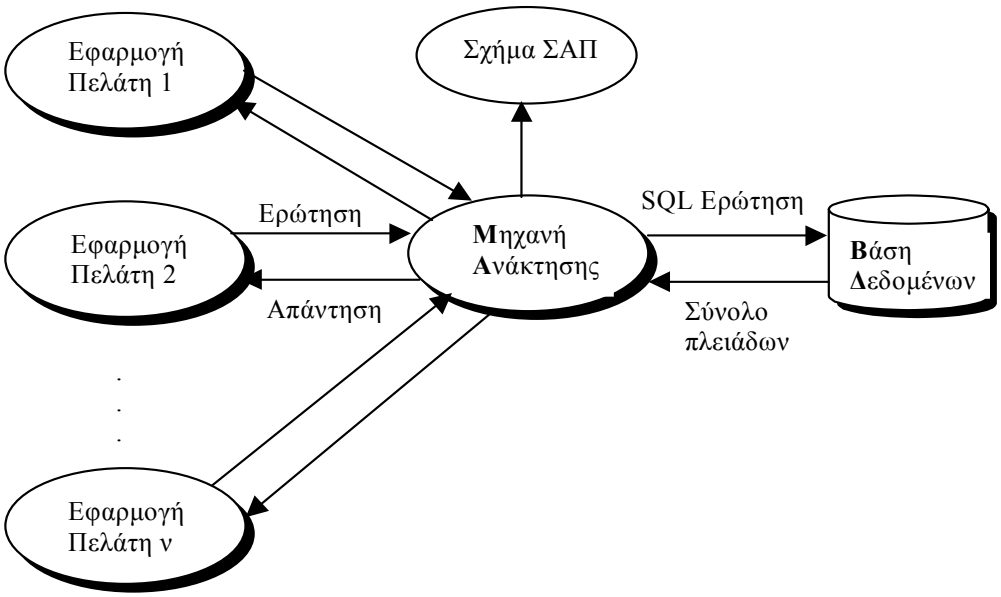
Στο κεφάλαιο αυτό λοιπόν θα παρουσιαστεί η **Μ**ηχανή **Α**νάκτησης (ΜΑ). Η ΜΑ είναι το τμήμα εκείνο του συστήματος το οποίο αναλαμβάνει την εκτέλεση των ερωτήσεων. Αυτό σημαίνει ότι δέχεται αιτήσεις (requests) από διάφορους χρήστες, οι οποίες είναι στην πραγματικότητα ερωτήσεις ομοιότητας, και στέλνει απαντήσεις (responses) οι οποίες αποτελούνται από την ταυτότητα των αντικειμένων, τα οποία ικανοποιούν την ερώτηση, και το βαθμό ομοιότητάς τους ως προς την ερώτηση.

Η ΜΑ βασίζεται στη μεθοδολογία που περιγράφηκε στο τέταρτο κεφάλαιο (τμήμα 4.7). Σύμφωνα με αυτή τη μεθοδολογία η εκτέλεση μιας ερώτησης ομοιότητας ανάγεται σε σύνθεση μεταξύ σχέσεων αξιολόγησης. Η πράξη της σύνθεσης όμως αντιστοιχεί στην πράξη της φυσικής σύνδεσης (natural join) μεταξύ πινάκων. Με βάση την υλοποίηση της σύνθεσης σχέσεων αξιολόγησης με *SQL* η εκτέλεση μιας ερώτησης ομοιότητας ανάγεται στην εκτέλεση μιας σειράς *SQL* ερωτήσεων. Η ΜΑ υλοποιεί τη **Φάση 2**, τη **Φάση 3** και τη **Φάση 4** της μεθοδολογίας υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων (τμήμα 4.7)

Στο επόμενο τμήμα θα περιγραφεί η διαδικασία αλληλεπίδρασης των διάφορων χρηστών, οι οποίοι υπαβάλλουν ερωτήσεις, με τη ΜΑ καθώς και η διακασία που ακολουθείται από τη ΜΑ για την απάντηση μιας ερώτησης. Στο τμήμα 6.3 θα παρουσιαστούν κάποια ιδιαίτερα χαρακτηριστικά της ΜΑ τα οποία την καθιστούν ιδιαίτερα ισχυρή ενώ, στο τμήμα 6.4 θα αναφερθούν μερικά θέματα υλοποίησης της ΜΑ. Τέλος στο τμήμα 6.5 θα δοθεί μια σύντομη περίληψη του κεφαλαίου.

6.2 Αλληλεπίδραση πελατών με τη Μηχανή Ανάκτησης και διαδικασία εκτέλεσης ερωτήσεων

Η ΜΑ κατασκευάστηκε με βάση το μοντέλο πελάτη-εξυπηρετητή (client-server) ώστε να εκμεταλλευτούμε τις δυνατότητες που παρέχει αυτό το μοντέλο. Κατά συνέπεια πολλοί πελάτες μπορούν να συνδέονται ταυτόχρονα στη ΜΑ προκειμένου να υποβάλλουν αιτήσεις ανάκτησης πληροφορίας. Στο επόμενο σχήμα παρουσιάζεται η διαδικασία αλληλεπίδρασης μεταξύ πελατών και **Μηχανής Ανάκτησης**.



*Σχήμα 6-1:Περιγραφή αλληλεπίδρασης μεταξύ πελατών και ΜΑ καθώς και μεταξύ ΜΑ και ΒΔ.Η ΜΑ κατά την εκκίνησή της διαβάζει από ένα αρχείο το σχήμα του κατασκευασθέντος **Σ**υστήματος **Α**νάκτησης **Π**ληροφορίας. Οι πελάτες στέλνουν ερωτήσεις στη ΜΑ και λαμβάνουν τις αντίστοιχες απαντήσεις.Η ΜΑ εκτελεί μια σειρά SQL ερωτήσεων στη ΒΔ προκειμένου να ικανοποιήσει μια αίτηση ανάκτησης πληροφορίας.*

Ένας από τους βασικούς σκοπούς του κατασκευασθέντος συστήματος είναι η ικανοποίηση αιτήσεων ανάκτησης πληροφορίας. Συγκρεκριμένα, δίνεται η δυνατότητα στο χρήστη να υποβάλλει ερωτήσεις προκειμένου να ανακτήσει την πληροφορία που αναζητεί. Η εφαρμογή πελάτη παρέχει στο χρήστη το περιβάλλον για την κατασκευή των υποβαλλόμενων ερωτήσεων και αναλαμβάνει την παρουσίαση των αποτελεσμάτων. Αφού ο χρήστης κατασκευάσει μια ερώτηση η εφαρμογή επικοινωνεί με τη ΜΑ και της υποβάλλει την ερώτηση. Στη συνέχεια λαμβάνει την απάντηση από τη ΜΑ και την παρουσιάζει στον χρήστη.

Η ΜΑ κατά την εκκίνησή της διαβάζει το κατάλληλο - ανάλογα με τη συγκεκριμένη εφαρμογή - αρχείο το οποίο έχει προκύψει από τον ΕΣ. Το αρχείο αυτό, όπως προαναφέρθηκε, περιγράφει το σχήμα του κατασκευασθέντος **Σ**υστήματος **Α**νάκτησης **Π**ληροφορίας (ΣΑΠ). Συνεπώς η μηχανή γνωρίζει τα σύνολα οντοτήτων, τα σύνολα συσχετίσεων, τα ασαφή χαρακτηριστικά και τις εκφράσεις αξιολόγησης, πράγμα που είναι απαραίτητο για την απάντηση των ερωτήσεων ομοιότητας που δέχεται.

Προκειμένου να απαντηθεί μια ερώτηση ομοιότητας υλοποιείται η **Φάση 2**, η **Φάση 3** και η **Φάση 4** της μεθοδολογίας υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων (τμήμα 4.7). Έτσι όταν η ΜΑ δέχεται μια αίτηση καταρχήν "αποκωδικοποιεί" την ερώτηση (**Φάση 2**, *Βήμα 1*) και προσδιορίζεται ποια είναι τα διαζευκτικά και συζευκτικά τμήματα της ερώτησης, ποιο είναι το μοντέλου υπολογισμού ομοιότητας, ποιο είναι το συνόλου οντοτήτων μέλη του οποίου στοχεύει να ανακτήσει η ερώτηση και ποιο είναι το συνόλου συσχετίσεων που εκφράζει την σύνδεση των προς ανάκτηση αντικειμένων με τους όρους δεικτοδότησης. Στη συνέχεια τα διαζευκτικά και τα συζευκτικά τμημάτα της ερώτησης αποθηκεύονται στους κατάλληλους πίνακες (**Φάση 2**, *Βήμα 2*).

Το επόμενο στάδιο είναι, με βάση το μοντέλο υπολογισμού ομοιότητας που προέκυψε από τη **Φάση 2**, να κατασκευστούν οι *SQL* εκφράσεις για τον υπολογισμό των διαζευκτικών τμημάτων (**OR1** και **OR2** όπως περιγραφηκαν στο τμήμα 4.6) και οι *SQL* εκφράσεις για τον υπολογισμό των συζευκτικών τμημάτων (**AND1** και **AND2** όπως περιγράφηκαν στο τμήμα 4.6) (**Φάση 3**).

Τέλος γίνεται εκτέλεση των ερωτήσεων που κατασκευάστηκαν **(Φάση 4,** *Βήμα 1*) καθώς και κατάλληλη μορφοποίηση των αποτελεσμάτων που προέκυψαν και αποστολή τους στην εφαρμογή πελάτη για την παρουσίαση τους (**Φάση 4,** *Βήμα 2*).

6.3 Ιδιαίτερα χαρακτηριστικά της Μηχανής Ανάκτησης

Το σημαντικότερο χαρακτηριστικό της ΜΑ είναι ότι είναι εντελώς ανεξάρτηση από την εφαρμογή του πελάτη. Η ΜΑ ανάκτησης, όπως προαναφέρθηκε, δέχεται, κατά την εκκίνηση, σαν είσοδο ένα αρχείο το οποίο περιγράφει το σχήμα του κατασκευασθέντος συστήματος ανάκτησης πληροφορίας. Στη συνέχεια περιμένει αιτήσεις ανάκτησης.

Η επικοινωνία της εφαρμογής πελάτη με τη ΜΑ γίνεται μέσω του μηχανισμού sockets. Η εφαρμογή κατά την εκκίνησή της ανοίγει ένα socket με την ΜΑ και μέσω αυτού επικοινωνεί με αυτή, δηλαδή, στέλνει ερωτήσεις και λαμβάνει τις αντίστοιχες απαντήσεις. Κατά συνέπεια είναι δυνατόν πολλοί πελάτες να είναι συνδεδεμένοι ταυτόχρονα στη ΜΑ η οποία αναλαμβάνει να τους ικανοποιήσει όλους.

Το γεγονός ότι η ΜΑ λειτουργεί εντελώς αυτόνομα δίνει τη δυνατότητα της αυτονομίας και στην εφαρμογή πελάτη. Η εφαρμογή πελάτη, όπως προαφέρθηκε, παρέχει στο χρήστη ένα περιβάλλον για τον σχηματισμό των ερωτήσεων και την παρουσίαση των αποτελεσμάτων. Είναι δυνατόν να κατασκευαστούν διαφορετικές εφαρμογές πελάτη, ανάλογα με τις εκάστοτε ανάγκες, χωρίς να χρειαστεί να αλλάξει τίποτα στη ΜΑ. Οι εφαρμογές πελάτη το μόνο που χρειάζεται να γνωρίζουν είναι το πρωτόκολλο επικοινωνίας με τη μηχανή ανάκτησης δηλαδή, να γνωρίζουν το σχήμα το οποίο πρέπει να έχουν τα δεδομένα που στέλνονται ώστε να είναι αναγνωρίσιμα από τη ΜΑ καθώς και το σχήμα που έχουν τα δεδομένα που επιστρέφονται από τη μηχανή ώστε να μπορούν να τα παρουσιάσουν στο χρήστη. Επιπλέον είναι δυνατόν να παρέχεται ένα σύνολο συναρτήσεων ή κλάσεων το οποίο θα μπορεί να χρησιμοποιείται από τους κατασκευαστές εφαρμογών πελάτη για την επικοινωνία με τη ΜΑ.

6.4 Θέματα υλοποίησης

Η Μηχανή Ανάκτησης κατασκευάστηκε με τη χρήση του πακέτου Microsoft Visual C++ 5.0. Χρησιμοποιήθηκαν οι κλάσεις που παρέχει για την υποστήριξη των λειτουργιών σε βάσεις δεδομένων. Ειδικότερα, για την εργασία με βάσεις δεδομένων

παρέχει την τεχνολογία Data Access Object (DAO) [12,13] η οποία παρέχει πολλές διευκολύνσεις όσον αφορά την αλληλεπίδραση με βάσεις δεδομένων.

Η ΜΑ δέχεται διάφορους πελάτες (clients) οι οποίοι συνδέονται μαζί της έτσι ώστε να μπορούν να της στέλνουν τις αιτήσεις τους. Η επικοινωνία της μηχανής με τους πελάτες γίνεται με sockets και για την υποστήριξη αυτής της επικοινωνίας χρησιμοποιήθηκαν οι κλάσεις που παρέχονται από την MFC (Microsoft Foundation Classes). Οι κλάσεις αυτές αναλαμβάνουν όλες τις εργασίες χαμηλού επιπέδου που απαιτούνται για την πραγματοποίηση μιας τέτοιας επικοινωνίας και παρέχουν ένα σύνολο συναρτήσεων (member functions) υψηλού επιπέδου. Έτσι δεν χρειάστηκε να αντιμετωπιστούν κάποια δύσκολα θέματα επικοινωνίας μεταξύ μηχανής και πελάτη.

Δεν θα αναφερθούν περισσότερα διότι τότε, θα χρειαζόταν να παρουσιαστούν λεπτομέρειες οι οποίες είναι μεν πολύ σημαντικές αλλά δεν μπορούν να συμπεριληφθούν σε ένα τέτοιο κείμενο.

6.5 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκε η **Μ**ηχανή **Α**νάκτησης. Η ΜΑ είναι το τμήμα του συστήματος το οποίο είναι υπεύθυνο για την εκτέλεση των ερωτήσεων ομοιότητας. Δέχεται συνεπώς ερωτήσεις και σαν έξοδο δίνει μια ταξινομημένη λίστα με τις ταυτότητες των αντικειμένων που ικανοποιούν την εκάστοτε ερώτηση και το βαθμό ομοιότητας του κάθε αντικειμένου με την ερώτηση.

Βασικό χαρακτηριστικό της ΜΑ είναι η απόλυτη ανεξαρτησία της από την εφαρμογή πελάτη. Αυτό σημαίνει ότι είναι δυνατόν να κατασκευάζονται διαφορετικές εφαρμογές πελάτη χωρίς να χρειάζεται να αλλάζει τίποτα σε σχέση με τη ΜΑ. Η εφαρμογή πελάτη το μόνο που χρειάζεται να γνωρίζει είναι το πρωτόκολλο επικοινωνίας με την ΜΑ.

Στο επόμενο κεφάλαιο θα περιγραφεί μια εφαρμογή ψηφιακών βιβλιοθηκών και θα παρουσιαστεί ένα ολοκληρωμένο σενάριο χρήσης του συστήματος.

7. Εφαρμογή ψηφιακών βιβλιοθηκών

7.1 Εισαγωγή

Μέχρι τώρα έχει περιγραφεί η διαδικασία κασκευής ενός σύστημα ανάκτησης πληροφορίας πάνω σε σχεσιακή βάση δεδομένων, με τη χρήση του Εκδότης Σχήματος, καθώς και ο τρόπος με τον οποίο θα απαντώνται οι ερωτήσεις ομοιότητας από την Μηχανή Ανάκτησης.

Στο κεφάλαιο αυτό θα παρουσιαστεί η εφαρμογή που κατασκευάστηκε για την υποβολή ερωτήσεων ομοιότητας. Ουσιαστικά θα παρουσιαστεί ένα ολοκληρωμένο σενάριο χρησιμοποίησης του συστήματός. Η εφαρμογή **δεν** αποτελεί μέρος του συστήματος, διότι εξαρχής, στόχος ήταν, το σύστημά να είναι αυτόνομο και κατά συνέπεια να μπορεί να χρησιμοποιηθεί από πολλές και διαφορετικές εφαρμογές. Η κατασκευή της συγκεκριμένης εφαρμογής αποσκοπεί στο να παρουσιάσει ένα πιθανό σενάριο χρησιμοποίησης του συστήματος αλλά και στο να πιστοποιήσει τη σωστή λειτουργία του. Η εφαρμογή εντάσσεται στο πεδίο των **ψηφιακών βιβλιοθηκών (digital libraries)** και των **ηλεκτρονικών εκδόσεων (electronic publications)** και έγινε στα πλαίσια το προγράμματος *VENIVA (ESPRIT n# 20638)* στο οποίο συμμετέχει το Εργαστήριο Κατανεμημένων Πληροφοριακών Συστημάτων Πολλαπλών Μέσων.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η κατασκευή ενός επιπέδου επαφής-ζεύξης (interface) για την κατασκευή ερωτήσεων. Το ενδιαφέρον αυτό μάλιστα, γίνεται ακόμα μεγαλύτερο αν αναλογιστούμε ότι αναφερόμαστε σε ερωτήσεις ομοιότητας των οποίων οι όροι μπορούν να φέρουν βάρη. Η κατασκευή ενός τέτοιου επιπέδου επαφής-ζεύξης (interface) παρουσιάζει ιδιαίτερη δυσκολία διότι πρέπει από τη μια να είναι εύχρηστο, απλό και κατανοητό για τον απλό χρήστη αλλά, από την άλλη, πρέπει να παρέχει την δυνατότητα κατασκευής πολύπλοκων ερωτήσεων ώστε να ικανοποιούνται οι απαιτήσεις και των πιο έμπειρων χρηστών. Επίσης πρέπει να θεωρηθεί ότι ένας τυπικός χρήστης δεν γνωρίζει τη σημασία των παρενθέσεων και τη σειρά των υπολογισμών στις σχηματιζόμενες ερωτήσεις οπότε, αυτές οι έννοιες θα πρέπει να υποστηριχθούν με διαδικασίες προσιτές για τον απλό χρήστη. Τα παραπάνω αποτέλεσαν τις απαιτήσεις σχεδιασμού και λειτουργικότητας της κατασκευασθείσας εφαρμογής.

Η εφαρμογή υποβολής ερωτήσεων είναι μια εφαρμογή του παγκόσμιου ιστού (web application) και υλοποιήθηκε με την γλώσσα προγραμματισμού Java. Η Java είναι μια γλώσσα η οποία, αν και είναι εξαιρετικά καινούργια, έχει κερδίσει περιόπτη θέση στο χώρο των γλωσσών προγραμματισμού και τα πεδία χρήσης της συνεχώς αυξάνονται.

Τα θέματα τα οποία θα αναφερθούν έχουν ως εξής. Στο επόμενο τμήμα θα παρουσιαστεί η βάση δεδομένων πάνω στην οποία οικοδομήθηκε η όλη εφαρμογή. Στο τμήμα 7.3 θα γίνει επακριβής καθορισμός των απαιτήσεων τις οποίες πρέπει να ικανοποιήσει η εφαρμογή, ενώ στο τμήμα 7.4 θα περιγραφούν τα οριζόμενα σύνολα οντοτήτων, οι συσχετίσεις και τα χαρακτηριστικά, προσδιορίζοντας έτσι την συλλογή του συστήματος ανάκτησης πληροφορίας. Στο τμήμα 7.5 θα αναφερθούν οι μέθοδοι που προτείνονται και υλοποιούνται για το σχηματισμό ερωτήσεων ομοιότητας. Τέλος το τμήμα 7.6 αποτελεί μια σύνοψη του κεφαλαίου.

7.2 Γενική περιγραφή

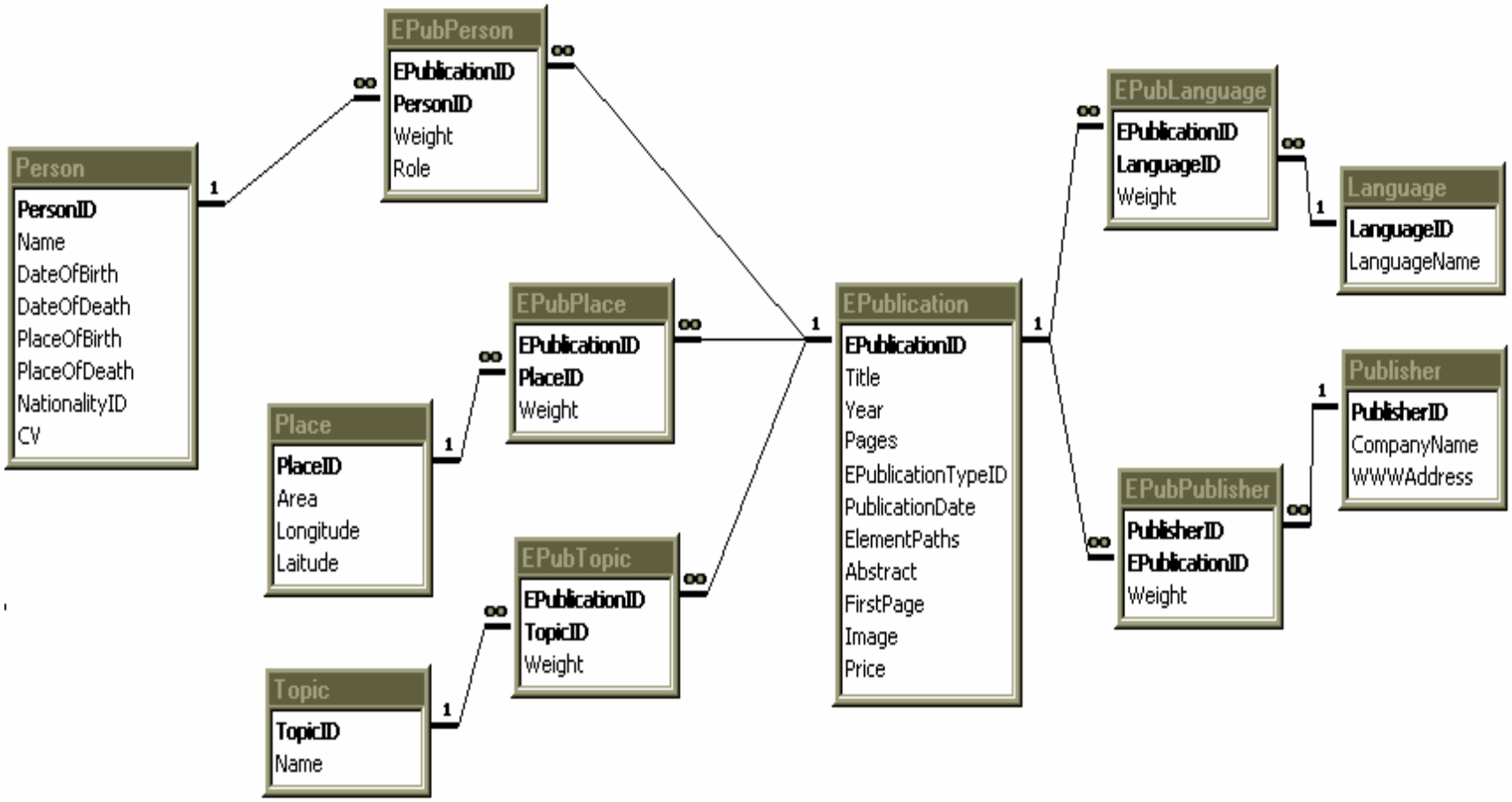
Η συγκεκριμένη εφαρμογή αποσκοπεί στην ανάκτηση πληροφορίας σχετικά με κάποιες ηλεκτρονικές εκδόσεις. Σε μια βάση δεδομένων έχουν αποθηκευτεί στοιχεία για μια σειρά από ηλεκτρονικές εκδόσεις τα οποία περιέχουν πληροφορία σχετικά με αυτές.

Τα στοιχεία αυτά είναι για παράδειγμα τα πρόσωπα και οι τοποθεσίες οι οποίες σχετίζονται με την κάθε έκδοση είτε έμμεσα είτε άμεσα, η γενικότερη θεματική ενότητα στην οποία εντάσσεται η έκδοση, ο εκδοτικός της οίκος και άλλα. Κάποιος χρήστης έχει την δυνατότητα να σχηματίσει και να υποβάλλει ερωτήσεις ομοιότητας προκειμένου να ανακτήσει τις ηλεκτρονικές εκδόσεις για τις οποίες ενδιαφέρεται. Σε πρώτη φάση ο χρήστης λαμβάνει σαν απάντηση τον τίτλο της κάθε έκδοσης που ικανοποιεί την ερωτήσή του και το βαθμό ομοιότητάς της ως προς την ερώτηση. Στην συνέχεια μπορεί επιλέγοντας κάποια από τις ανακτηθείσες εκδόσεις να πάρει περισσότερες πληροφορίες γι' αυτή, όπως, και την ίδια την έκδοση.

Η παραπάνω διαδικασία γίνεται μέσω ενός browser αφού πρόκειται για μια εφαρμογή του Παγκόσμιου Ιστού (Web application). Αυτό σημαίνει ότι κάποιος, ο οποίος έχει πρόσβαση στο Internet, μπορεί με τη βοήθεια ενός browser να επισκεφτεί τη συγκεμένη διεύθυνση στην οποία βρίσκεται η εφαρμογή και να κάνει τις όποιες ερωτήσεις του. Στο επόμενο τμήμα θα παρουσιαστεί η βάση δεδομένων του συστήματος.

7.3 Η βάση δεδομένων του συστήματος

Σ' αυτό το τμήμα θα παρουσιαστεί η βάση δεδομένων πάνω στην οποία θα οικοδομηθεί το σύστημα ανάκτησης πληροφορίας. Το σχήμα της βάσης δεδομένων διαδραματίζει αρκετά σημαντικό ρόλο στην κατασκευή του όλου συστήματος, γι' αυτό το λόγο, θα γίνει μια σύντομη περιγραφή των πινάκων που έχουν άμεση σχέση με την όλη διαδικασία κατασκευής του συστήματος. Στην επόμενη σελίδα δίνεται μια γραφική απεικόνιση της βάσης δεδομένων και των σχέσεων που υπάρχουν μεταξύ των πινάκων της.



Εικόνα 7-1 :Γραφική απεικόνιση της βάσης δεδομένων σύμφωνη με τις συμβάσεις της MS Access. Τα πεδία με πιο έντονο χρώμα είναι τα πρωτεύοντα κλειδιά των πινάκων ενώ οι γραμμές σύνδεσης μεταξύ των πινάκων δηλώνουν την ύπαρξη εξωτερικών κλειδιών .

Ο πίνακας *EPublication* περιέχει τα στοιχεία όλων των ηλεκτρονικών εκδόσεων που υπάρχουν στη βάση και είναι ο σημαντικότερος πίνακας όσον αφορά την εργασία μας.

Ο πίνακας *Person* περιέχει τα στοιχεία των προσώπων που έχουν σχέση με κάποια έκδοση. Τα πρόσωπα που συνδέονται με μια έκδοση μπορούν να είναι είτε ο συγγραφέας του κειμένου είτε ο εικονογράφος (illustrator) είτε ακόμα και κάποια πρόσωπα, πραγματικα ή φανταστικα, για τα οποία γίνεται αναφορά στο κείμενο.

Ο πίνακας *EPubPerson* συνδέει κάθε έκδοση με τα πρόσωπα που έχουν σχέση με αυτή. Το πεδίο *Weight* αυτού του πίνακα εκφράζει τη σπουδαιότητα του κάθε προσώπου σε σχέση με την έκδοση με την οποία συνδέεται και κυμαίνεται μεταξύ 0 και 1. Ο βαθμός αυτός θα μπορούσε για παράδειγμα να είναι 1 αν το πρόσωπο είναι ο συγγραφέας, 0,4 αν είναι ο εικονογράφος ενώ αν είναι κάποιο πρόσωπο για το οποίο γίνεται λόγος στο κείμενο τότε, ο βαθμός θα πρέπει να προκύπτει με κάποια ειδική μέθοδο αξιολόγησης. Ο βαθμός σπουδαιότητας μπορεί να εισάγεται μαζί με τα υπόλοιπα δεδομένα αλλά είναι δυνατόν να έχει προκύψει, αν αυτό είναι εφικτό, με αυτόματες μεθόδους βαθμολόγησης.

Ο πίνακας *Place* περιέχει τους τόπους οι οποίοι σχετίζονται με μια ηλεκτρονική έκδοση. Η σχέση που μπορεί να υπάρχει ανάμεσα σε μια έκδοση και ένα τόπο είναι πολύπλευρη και δεν υπάρχει κάποιος συγκεκριμένος κανόνας για τον καθορισμό της. Έτσι, για παράδειμα, ένα κείμενο το οποίο αναφέρεται στην Κρήτη και μιλάει για τα Χανιά και το Ηράκλειο μπορεί να ειπωθεί ότι σχετίζεται με αυτές τις δύο πόλεις. Μπορεί επίσης να θεωρηθεί ότι μια έκδοση σχετίζεται και με τον τόπο καταγωγής ή εργασίας του συγγραφέα της. Έτσι, ένα επιστημονικό κείμενο το οποίο έχει συνταχθεί από κάποιο καθηγητή του Πολυτεχνείου Κρήτης μπορεί να θεωρηθεί ότι συνδέεται έμμεσα με τα Χανιά. Ο πίνακας που εκφράζει τη σύνδεση έκδοσης-τόπου είναι ο *EPubPlace* και το πεδίο *Weight* που περιέχει εκφράζει την σπουδαιότητα της κάθε τοποθεσίας ως προς μια έκδοση και το αντίστροφο.

Ο πίνακας *Topic* περιέχει τις θεματικές ενότητες στις οποίες κατατάσσονται οι εκδόσεις. Έτσι μπορεί να υπάρχει για παράδειγμα μια θεματική ενότητα που αφορά τα δίκτυα υπολογιστών μια άλλη που αφορά τα πολυμέσα και μια τρίτη που περιλαμβάνει τις βάσεις δεδομένων. Κάθε έκδοση εντάσσεται έτσι σε μια ή περισσότερες θεματικές ενότητες. Η αντιστοίχιση έκδοσης-ενότητας γίνεται μέσω του πίνακα *EPubTopic*. Το πεδίο *Weight* εκφράζει το βαθμό συσχέτισης μιας έκδοσης με κάποια θεματική ενότητα. Έτσι, για παράδειγμα, ένα κείμενο που αναφέρεται στα

ιδιαίτερα χαρακτηριστικά των βάσεων δεδομένων που χρησιμοποιούνται στα πολυμέσα, θα συνδέοταν και με την θεματική ενότητα των βάσεων δεδομένων και με αυτή των πολυμέσων, ενώ το βάρος στη μια και στην άλλη περίπτωση θα ήταν διαφορετικό (ή και το ίδιο) και θα προέκυπτε από την λεπτομερέστερη ανάλυση του κειμένου.

Ο πίνακας *Language* περιέχει τις γλώσσες στις οποίες είναι γραμμένα τα κείμενα ενώ ο πίνακας *EPubLanguage* εκφράζει την σύνδεση κάθε κειμένου με την γλώσσα στην οποία είναι γραμμένο. Το πεδίο *Weight* σ' αυτόν τον πίνακα αποδίδει και πάλι κάποιο βαθμό σπουδαιότητας. Μια πιθανή κατάσταση στην οποία η ύπαρξη του βάρους έχει νόημα είναι η περίπτωση στην οποία ένα κείμενο μεταφράζεται σε πολλές γλώσσες και η μετάφραση από την γλώσσα την οποία χρησιμοποίησε ο συγγραφέας σε μια άλλη γλώσσα δεν είναι καλή. Συνεπώς μπορεί να αποδωθεί βαθμός 1 για την έκδοση η οποία βρίσκεται στην γλώσσα του συγγραφέα ενώ για τις μεταφράσεις της μπορεί να αποδωθεί ένας βαθμός μεταξύ 0-1 ανάλογα με το πόσο καλή είναι η μετάφραση. Ένα άλλο σενάριο είναι να υπάρχει μια έκδοση στη γλώσσα του συγγραφέα και να υπάρχουν περιλήψεις της έκδοσης σε άλλες γλώσσες. Οπότε στη πρώτη περίπτωση αποδίδεται βαθμός 1 ενώ στην δεύτερη ένας βαθμός μεταξύ 0-1 ανάλογα με το πόσο εκτενής και καλή είναι η περίληψη. Αν αυτά τα σενάρια θεωρηθούν κάπως προχωρήμενα ή δεν μπορούν να εφαρμοστούν ο βαθμός μπορεί να είναι ο ίδιος για όλες τις αντιστοιχίσεις γλώσσας-έκδοσης, αναγνωρίζοντας φυσικά, τη σπατάλη χώρου λόγο της μη χρησιμοποίησης, στην πράξη, του πεδίου *Weight*.

Τέλος ο πίνακας *Publisher* περιέχει τους εκδοτικούς οίκους των οποίων οι εκδόσεις περιέχονται στη βάση, ενώ ο πίνακας *EPubPublisher* παρέχει την σύνδεση εκδοτικού οίκου-έκδοσης. Η ύπαρξη του βάρους βρίσκει εφαρμογή σε κάποιες ειδικές περιπτώσεις όπως για παράδειγμα όταν μια έκδοση έχει προκύψει με την συνεργασία δύο ή περισσότερων εκδοτικών οίκων, οπότε, ανάλογα με την συνεισφορά του κάθε οίκου μπορεί να αποδωθεί διαφορετικός βαθμός. Μπορεί και πάλι να τεθεί η ίδια τιμή βάρους για όλες τις αντιστοιχήσεις εκδοτικού οίκου-έκδοσης αγνοώντας στην ουσία το πεδίο *Weight*.

7.4 Καθορισμός απαιτήσεων και περιγραφή των δυνατοτήτων του συστήματος

Οι απαιτήσεις και κατά συνέπεια, οι δυνατότητες που θα παρέχονται από το σύστημα έχουν άμεση σχέση με τη βάση δεδομένων πάνω στην οποία κατασκευάζεται το σύστημα ανάκτησης πληροφορίας. Αυτό είναι απολύτως λογικό, αφού ο καθορισμός της μορφής των υποβαλλόμενων ερωτήσεων δεν είναι ανεξάρτητος της συλλογής των αντικείμενων τα οποία πρέπει να ανακτηθούν. Η συλλογή με την σειρά της, κατασκευάζεται με βάση τα δεδομένα που υπάρχουν, αλλά, και λαμβάντας υπ' όψιν τον τρόπο με τον οποίο έχουν αποθηκευτεί αυτά τα δεδομένα.

Στην συγκεκριμένη περίπτωση, αντικειμενικός στόχος των υποβαλλόμενων ερωτήσεων είναι η ανάκτηση ηλεκτρονικών εκδόσεων. Οι ερωτήσεις που μπορούν να υποβληθούν απ' ευθείας στην βάση δεδομένων είναι ερωτήσεις δίτιμης λογικής (boolean queries). Έχει γίνει ήδη αναφορά στον περιοριστικό χαρακτήρα αυτών των ερωτήσεων. Συνοπτικά αναφέρεται ότι, οι ερωτήσεις δίτιμης λογικής ενδείκνυνται σε περιπτώσεις όπου υπάρχει επαρκής πληροφορία γι' αυτό για το οποίο αναζητείται, πιο απλά, ξέρουμε ακριβώς τι ζητάμε. Συνήθως όμως δεν βρισκόμαστε σε αυτή την ευχάριστη θέση και θα θέλαμε να εκτελέσουμε πιο γενικές ερωτήσεις, ερωτήσεις που θα αντιμετώπιζαν με επιτυχία το γεγονός ότι δεν γνωρίζουμε τι ακριβώς ζητάμε αλλά μπορούμε να το περιγράψουμε σε κάποιο βαθμό. Σ' αυτή ακριβώς την κατάσταση, οι ερωτήσεις ομοιότητας μπορούν να ανταποκριθούν με αρκετή επιτυχία.

Είναι λογικό να θέλουμε να ρωτάμε για μια έκδοση με βάση τα πρόσωπα που σχετίζονται με αυτή, είτε πρόκειται για το συγγραφέα είτε για ένα πρόσωπο για το οποίο γίνεται λόγος σε κάποιο κείμενο. Θα μπορούσαμε έτσι, αν είχαμε μια βάση δεδομένων με ιστορικά κείμενα, να ζητήσουμε όλα τα κείμενα τα οποία σχετίζονται με τον Ηρόδοτο. Συνεπώς θα παίρναμε τα κείμενα τα οποία έγγραψε ο Ηρόδοτος καθώς και εκείνα στα οποία γίνεται αναφορά στον Ηρόδοτο. Συνεπώς θα πρέπει να μπορούμε να κάνουμε ερωτήσεις με βάση τα πρόσωπα.

Ένα άλλο κριτήριο με βάση το οποίο θα θέλαμε ίσως να ανακτήσουμε κάποιες εκδόσεις είναι οι τοποθεσίες στις οποίες αυτές αναφέρονται. Θα μπορούσαμε έτσι να ζητήσουμε τις εκδόσεις οι οποίες σχετίζονται με το Ηράκλειο, οπότε και θα παίρναμε τις εκδόσεις στις οποίες γίνεται λόγος για το Ηράκλειο αλλά ίσως και κάποιες

εκδόσεις των οποίων συγγραφέας είναι για παράδειγμα ο Καζαντζάκης ο οποίος κατάγεται από το Ηράκλειο.

Επίσης συχνά θέλουμε να ρωτήσουμε με βάση την γενικότερη θεματική ενότητα στην οποία εντάσσονται κάποιες εκδόσεις. Ένας ιστορικός για παράδειγμα θα μπορούσε να ζητήσει όλες τις εκδόσεις που σχετίζονται με την περίοδο του Διαφωτισμού και να πάρει σαν απάντηση κείμενα τα οποία, είτε αναφέρονται στο Διαφωτισμό είτε γράφτηκαν από σπουδαίους Διαφωτιστές είτε ακόμα αναφέρονται σε περιοχές που συνδέθηκαν ιδιαίτερα με το κίνημα του Διαφωτισμού.

Οι εκδοτικοί οίκοι συνδέονται φυσικά άμεσα με τις εκδόσεις τους. Έτσι κάποιος που ενδιαφέρεται να μελετήσει τις εκδόσεις ενός συγκεκριμένου οίκου θα μπορούσε να ανακτήσει κάποιες εκδόσεις με βάση τον εκδοτικό τους οίκο. Πρέπει συνεπώς να παρέχουμε και αυτή την δυνατότητα.

Η γλώσσα στην οποία έχει εκδοθεί ένα κείμενο είναι ένας ακόμα σημαντικός παράγοντας που χαρακτηρίζει το κείμενο, ιδιαίτερα στην εποχή μας. Θα πρέπει συνεπώς ο χρήστης να μπορεί να απευθύνει ερωτήσεις στις οποίες θα καθορίζει την γλώσσα ή τις γλώσσες για τις οποίες ενδιαφέρεται. Έτσι θα ήταν άσκοπο να επιστρέψουμε αναφορές για εκδόσεις που είναι στα Γερμανικά ή στα Αγγλικά σε κάποιον που είναι γνώστης μόνο της ελληνικής γλώσσας ενώ θα ήταν επιθυμητό να παρουσιάσουμε αυτές τις εκδόσεις σε κάποιο γλωσσομαθή. Συμπερασματικά, η δυνατότητα επιλογής γλώσσας έχει γίνει πλέον απαραίτητη και πρέπει να παρέχεται στον χρήση.

Τα προαναφερθέντα γίνονται ίσως περισσότερα κατανοητά, αλλά και ενδιαφέροντα, όταν τα δούμε μέσα σε ένα γενικότερο πλαίσιο υποβολής ερωτήσεων στο οποίο γίνεται συνδυασμός όλων τα παραπάνω δυνατοτήτων. Έτσι μια ερώτηση στην οποία θα ζητούσαμε τις εκδόσεις οι οποίες αναφέρονται στην αρχαία Αθήνα και έχουν σχέση με τον Πλάτωνα ή το Σωκράτη και εντάσσονται στην ενότητα φιλοσοφία και είναι εκδόσεις του Πανεπιστημίου Αθηνών ή του Πανεπιστημίου Θεσσαλονίκης θα μπορούσε να απαντηθεί βασιζόμενη στις δυνατότητες που προαναφέραμε.

Στο επόμενο τμήμα θα παρουσιαστεί η κατασκευή της συλλογής του συστήματος και θα προσδιορισμούν οι όροι δεικτοδότησης έτσι ώστε να ικανοποιούνται όλες οι προηγούμενες απαιτήσεις.

7.5 Προσδιορισμός της συλλογής του συστήματος ανάκτησης πληροφορίας

Λαμβάνοντας υπ' όψιν τις προαναφερθείσες απαιτήσεις αλλά και το σχήμα της βάσης δεδομένων θα παρουσιαστούν τα σύνολα οντοτήτων, τα σύνολα συσχετίσεων και τα χαρακτηριστικά που πρέπει να οριστούν έτσι ώστε να ικανοποιηθούν οι προηγούμενες απαιτήσεις. Κατ' αυτό τον τρόπο προσδιορίζεται η συλλογή του κατασκευαζόμενου συστήματος ανάκτησης πληροφορίας. Ο ορισμός του σχήματος του συστήματος ανάκτησης πληροφορίας από το χρήστη-διαχειριστή είναι η ***Φάση 1: Ορισμός σχήματος ΣΑΠ*** (τμήμα 4.7) της μεθοδολογίας υλοποίησης ερωτήσεων ομοιότητας σε σχεσιακή βάση δεδομένων. Ο ορισμός του σχήματος γίνεται με τον **Εκδότη Σχήματος** ο οποίος παρουσιάστηκε στο 5ο κεφάλαιο.

Με δεδομένο το γεγονός ότι σκοπός είναι η ανάκτηση εκδόσεων είναι φανερό ότι θα πρέπει να οριστεί ένα σύνολο οντοτήτων πάνω στον πίνακα *EPublication*. Ας ονομάσουμε αυτή την οντότητα *EPubEntitySet*.

Από τη στιγμή που επιθυμείται η υποβολή ερωτήσεων για την ανάκτηση εκδόσεων με βάση τα πρόσωπα με τα οποία αυτές σχετίζονται προκύπτει το συμπέρασμα ότι, αφενός ένα πρόσωπο "περιγράφει" σε κάποιο βαθμό μια έκδοση, άρα μπορεί να θεωρηθεί ως όρος δεικτοδότησής της (index term) και αφετέρου ότι, ένα πρόσωπο μπορεί να χρησιμοποιηθεί για τον σχηματισμό μιας ερώτησης. Υπενθυμίζεται ότι όταν αναφέρθηκε η έννοια του συνόλου οντοτήτων (στο 4ο κεφάλαιο) δηλώθηκε ότι ένα κριτήριο για το ποια σύνολα οντοτήτων θα οριστούν και πάνω σε ποιους πίνακες είναι οι όροι δεικτοδότησης. Συνεπώς θα πρέπει να οριστεί ένα σύνολο οντοτήτων πάνω στο πίνακα *Persons* το οποίο και ονομάζουμε *PersonsEntitySet*. Με το ίδιο ακριβώς σκεπτικό γίνεται ο ορισμός συνόλων οντοτήτων πάνω στους πίνακες *Places, Topics, Publishers, Languages*.

Μια πλήρης περιγραφή όλων των οριζόμενων συνόλων οντοτήτων δίνεται στον ακόλουθο πίνακα:

Ονόματα Συνόλων Οντότητας	**Πίνακας στον οποίο ορίζεται το συνόλο**
EPubEntitySet	*Epublications*
PersonsEntitySet	*Persons*
PlacesEntitySet	*Places*
TopicsEntitySet	*Topics*
PublishersEntitySet	*Publishers*
LanguagesEntitySet	*Languages*

Πίνακας 7-1: Τα σύνολα οντοτήτων που ορίζονται για τη συγκεκριμένη εφαρμογή

Με τον καθορισμό των συνόλων οντοτήτων έχει προσδιοριστεί το σύνολο των προς ανάκτηση αντικειμένων της συλλογής όπως και οι όροι δεικτοδότησης που χρησιμοποιούνται. Είναι όμως αναγκαίο να συνδεθούν τα προς ανάκτηση αντικείμενα με τους όρους δεικτοδότησης τους. Η αναγκαιότητα αυτή ίσως να μην είναι τόσο εμφανής στην συγκεκριμένη περίπτωση διότι έχουμε μόνο μια κατηγορία αντικειμένων ανάκτησης, δηλαδή τις εκδόσεις, και το σύνολο των όρων δεικτοδότησης χρησιμοποιείται για να περιγράψει αυτή τη μια κατηγορία. Στην γενική περίπτωση όμως, όπου υπάρχουν πολλές κατηγορίες αντικειμένων προς ανάκτηση και χρησιμοποιούνται διαφορετικοί όροι δεικτοδότησης για κάθε κατηγορία, η σύνδεση αντικειμένων προς ανάκτηση και όρων δεικτοδότησης είναι επιβεβλημένη.

Η σύνδεση λοιπόν αυτή πραγματοποιείται με τον ορισμό μια σειράς συσχετίσεων. Υπενθυμίζεται ότι μια συσχέτιση ορίζεται πάνω σε ένα πίνακα και δηλώνει τη συσχέτιση δύο συνόλων οντοτήτων. Έτσι για παράδειγμα, προκειμένου να δηλωθεί ότι το σύνολο οντοτήτων των προσώπων συνδέεται με το σύνολο οντοτήτων των ηλεκτρονικών εκδόσεων ορίζεται η συσχέτιση *EPubPersonsRel* πάνω στο πίνακα *EPubPersons.*

Με το ίδιο σκεπτικό γίνεται ο ορισμός και των υπόλοιπων συσχετίσεων όπως φαίνεται στον επόμενο πίνακα.

Όνομα σχέσης	Πρώτη οντότητα	Δεύτερη οντότητα	Πίνακας στον οποίο ορίζεται
EpubPersonsRel	*EPubEntitySet*	*PersonsEntitySet*	*EPubPersons*
EpubPlacesRel	*EPubEntitySet*	*PlacesEntitySet*	*EPubPlaces*
EpubTopicsRel	*EPubEntitySet*	*TopicsEntitySet*	*EPubTopics*
EpubPublishersRel	*EPubEntitySet*	*PublishersEntitySet*	*EPubPublishers*
EpubLanguagesRel	*EPubEntitySet*	*LanguagesEntitySet*	*EPubLanguages*

Πίνακας 7-2: Οι συσχετίσεις που ορίζονται μεταξύ των συνόλων οντοτήτων για τη συγκεκριμένη εφαρμογή. Οι συσχετίσεις αυτές στην ουσία συνδέουν τα προς ανάκτηση αντικείμενα με τους όρους δεικτοδότησής τους.

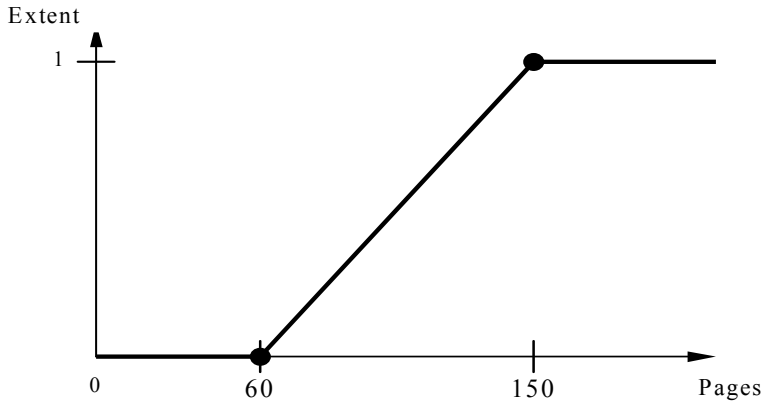
Η δομή του πίνακα *EPublication* επιτρέπει τον ορισμό δύο χαρακτηριστικών τα οποία αναφέρονται στο μέγεθος, ως αριθμός σελίδων, μιας έκδοσης και στο χρονικό σημείο κατά το οποίο παρουσιάστηκε η έκδοση στην αγορά.

Ορίζεται έτσι πάνω στο σύνολο οντοτήτων *EPubEntitySet* το χαρακτηριστικό *Extent*. Το χαρακτηριστικό αυτό είναι ασαφές (fuzzy attribute) και η τιμή του υπολογίζεται για κάθε έκδοση με βάση τον αριθμό των σελίδων της. Η έκφραση υπολογισμού του απαρτίζεται μόνο από το πεδίο *Pages* του πίνακα *EPublication* ενώ η συνάρτηση αξιολόγησής του είναι η γραμμική με παραμέτρους έστω, για παράδειγμα, $A = 60$ και $B = 150$. Αυτό σημαίνει ότι για τις εκδόσεις οι οποίες έχουν κάτω από 60 σελίδες το χαρακτηριστικό *Extent* έχει τιμή 0 ενώ για αυτές που έχουν πάνω απο 150 έχει τιμή 1. Το χαρακτηριστικό θα παίρνει τιμές μεταξύ 0 και 1 για εκδόσεις που έχουν περισσότερες από 60 σελίδες και λιγότερες από 150. Έτσι θεωρείται ότι, οι εκδόσεις για τις οποίες το χαρακτηριστικό *Extent* έχει τιμή 0 δεν είναι σίγουρα εκτενείς ενώ αυτές για τις οποίες έχει τιμή 1 είναι με βεβαιότητα εκτενείς. Για τιμές όμως του χαρακτηριστικού μεταξύ 0 και 1 δεν υπάρχει βεβαιότητα για το αν η έκδοση είναι εκτενής η όχι και η εκτίμηση αυξάνει γραμμικά από το 0 στο 1. Έτσι μπορεί κάποιος για παράδειγμα να ζητήσει τις εκδόσεις που αναφέρονται στα Χανιά και είναι εκτενείς

Η πλήρης περιγραφή του χαρακτηριστικού δίνεται στον επόμενο πίνακα και η γραφική του αναπαράσταση στο αμέσως επόμενο σχήμα:

Όνομα Χαρακτηριστικού	Όνομα Συνόλου Οντοτήτων	Έκφραση Υπολογισμού	Συνάρτηση Αξιολόγησης
Extent	*EPubEntitySet*	*Pages*	*Γραμμική α=60, b=150*

Πίνακας 6-3: Πλήρης περιγραφή του χαρακτηριστικού Extent



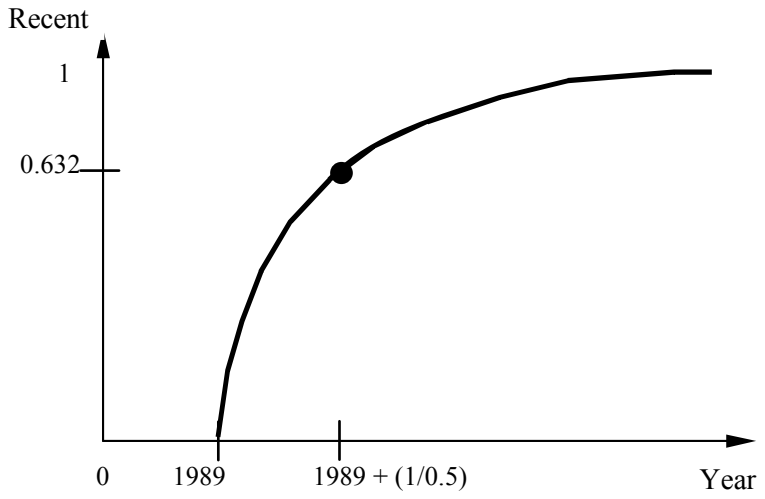
Γραφική απεικόνιση της συνάρτησης αξιολόγησης του ασαφούς χαρακτηριστικού Extent

Ορίζεται επίσης πάνω στο σύνολο οντοτήτων *EPubEntitySet* και ένα δεύτερο ασαφές χαρακτηριστικό το *Recent*. Το χαρακτηριστικό αυτό έχει να κάνει με το χρονικό σημείο κατά το οποίο δημοσιεύτηκε μια συγκεκριμένη έκδοση. Η λογική είναι ότι όσο πιο πρόσφατη είναι μια έκδοση τόσο πιο πολύ η τιμή του *Recent* θα πλησιάζει το 1 ενώ για εκδόσεις οι οποίες έχουν κυκλοφορίσει πριν από ένα ορισμένο χρονικό σημείο η τιμή του χαρακτηριστικού θα γίνεται 0. Έτσι η έκφραση υπολογισμού του χαρακτηριστικού *Recent* αποτελείται μόνο από το πεδίο *Year* του πίνακα *EPublication* ενώ, ως συνάρτηση αξιολόγησής του επιλέχθηκε η εκθετική με παράμετρο *A = 0,5* και *B = 1989*. Έτσι για εκδόσεις οι οποίες κυκλοφόρησαν πριν το 1989 η τιμή του *Recent* είναι 0 ενώ για αυτές που κυκλοφόρησαν απο το 1989 και μετά τη τιμή του χαρακτηριστικού αυξάνεται εκθετικά με ρυθμό 0.5. Μπορούν συνεπώς να ζητηθούν, για παράδειγμα, οι πρόσφατες εκδόσεις του οίκου Addison-Wesley με θέμα τα συστήματα ανάκτησης πληροφορίας.

Η πλήρης περιγραφή του χαρακτηριστικού *Recent* δίνεται στον επόμενο πίνακα και η γραφική του απεικόνιση στο αμέσως επόμενο σχήμα:

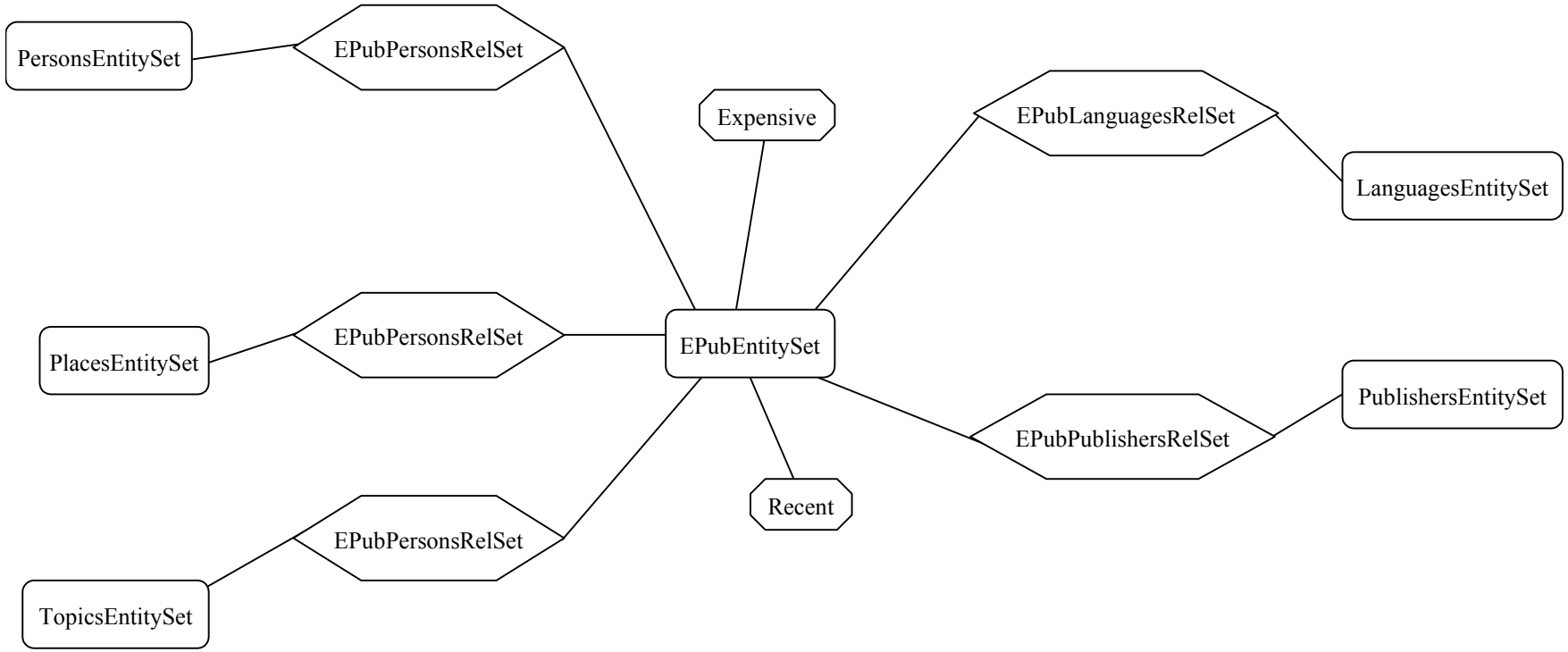
Όνομα Χαρακτηριστικού	Όνομα Συνόλου Οντοτήτων	Έκφραση Υπολογισμού	Συνάρτηση Αξιολόγησης
Recent	*EPubEntitySet*	*Year*	*Εκθετική α=0.5, b=1989*

Πίνακας 7-4: Πλήρης περιγραφή του χαρακτηριστικού Recent



Γραφική απεικόνιση της συνάρτησης αξιολόγησης του ασαφούς χαρακτηριστικού Recent

Μια γραφική αναπαράσταση των οριζόμενων συνόλων οντοτήτων, συνόλων συσχετίσεων και των ασαφών χαρακτηριστικών δίνεται στο επόμενο σχήμα:



Γραφική αναπαράσταση των οριζόμενων συνόλων οντοτήτων, συνόλων συσχετίσεων και χαρακτηριστικών. Οι ορισμοί αυτοί γίνονται από το χρήστη-διαχειριστή του συστήματος χρησιμοποιώντας τον ***Εκδότη Σχήματος***

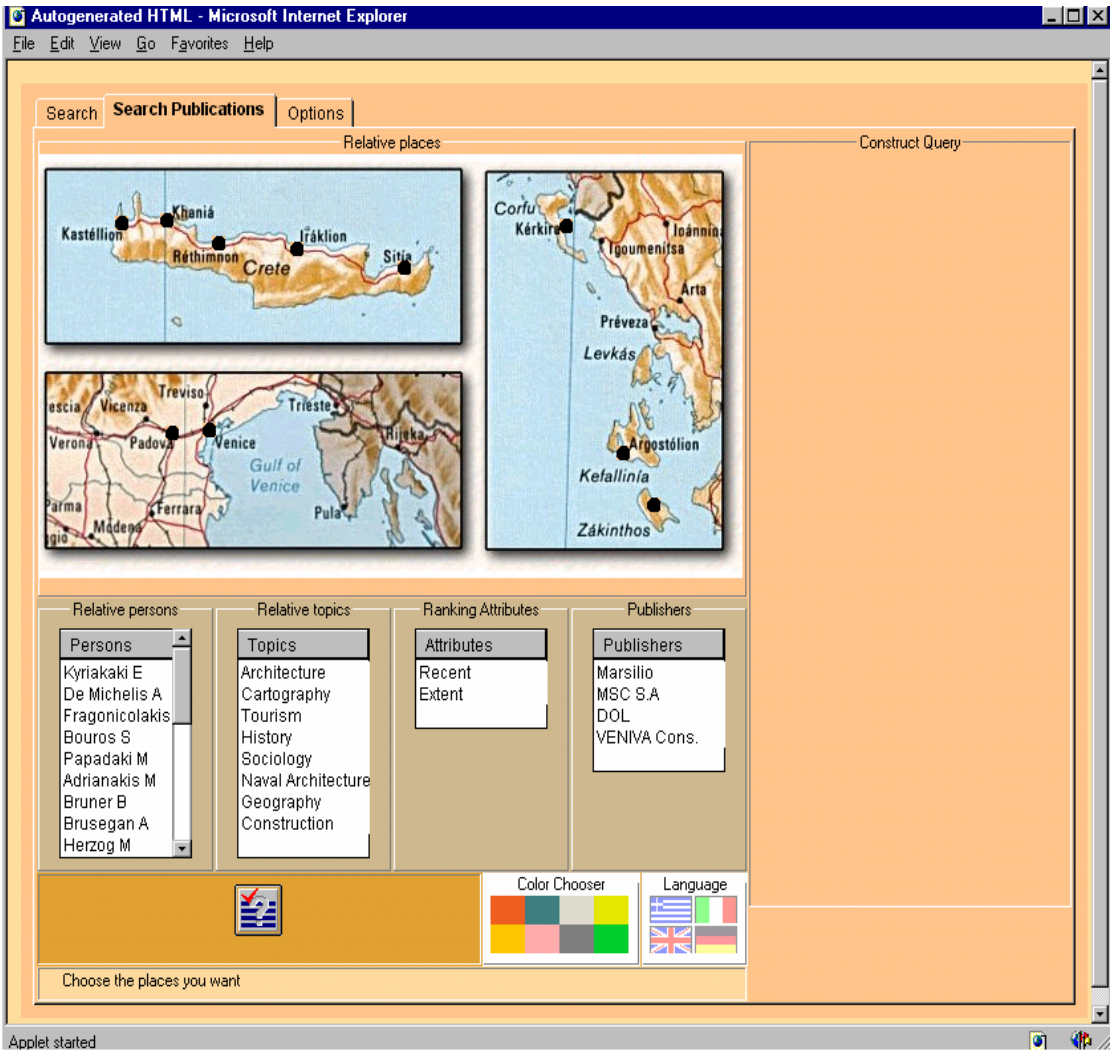
7.6 Το περιβάλλον για τον σχηματισμό των ερωτήσεων

Όπως αναφέρθηκε και στην εισαγωγή, υπάρχουν αρκετές δυσκολίες στην κατασκευή ενός γραφικού περιβάλλοντος για το σχηματισμό και την υποβολή λογικών ερωτήσεων - είτε ερωτήσεων δίτιμης λογικής είτε ομοιότητας - σε κάποιο πληροφοριακό σύστημα. Η δυσκολία προκύπτει από το γεγονός ότι θα πρέπει να δοθεί στον χρήστη η δυνατότητα να κάνει πολύπλοκα πράγματα -εν προκειμένω να σχηματίσει πολύπλοκες ερωτήσεις- με απλό τρόπο, κάτι το οποίο ακούγεται αντιφατικό.

Θα μπορούσε να υιοθετηθεί μια συγκεκριμένη μεθοδολογία, όπου με μια σειρά διαδοχικών βημάτων θα δινόταν στον χρήστη η δυνατότητα να κατασκευάσει μια πολύπλοκη έκφραση. Η τμηματοποίηση της διαδικασίας και ο διαχωρισμός της σε επίπεδα, για να αντιμετωπιστεί το πρόβλημα των παρενθέσεων και της σειράς των υπολογισμών, δίνει την απατηλή εικόνα της απλοποίησης της όλης διαδικασίας. Το να παρέχεται δηλαδή στο χρήστη (και η αναφορά γίνεται πάντα στον απλό χρήστη) η δυνατότητα να ορίζει διάφορα επίπεδα προτεραιότητας των υπολογισμών και να δημιουργεί κατ' αυτό τον τρόπο μια δενδροειδή δομή ερώτησης δεν ελαττώνει την πολυπλοκότητα της διαδικασίας.

Η μέθοδος που ακολουθήθηκε για την αντιμετώπιση του προβλήματος βασίζεται στην εξής σκέψη. Αφού είναι δύσκολο να προταθούν απλοί τρόποι για την κατασκευή πολύπλοκων ερωτήσεων ας μετασχηματιστούν οι ερωτήσεις σε μια ισοδύναμη αλλά πιο απλή μορφή. Υπάρχει όμως ένας τέτοιος μετασχηματισμός; Είναι γνωστό ότι όλες οι λογικές ερωτήσεις μπορούν να αναχθούν σε ισοδύναμες ερωτήσεις σε κανονική συζευκτική ή διαζευκτική μορφή, αν και αυτό δεν ισχύει για την περίπτωση των ερωτήσεων ομοιότητας δίοτι δεν ισχύουν κάποιες βασικές ιδιότητες όπως είναι η προσεταιριστική, η αντιμεταθετική κ.λ.π. Συνεπώς αν δοθεί στο χρήστη, με έναν εύκολο και απλό τρόπο, η δυνατότητα να κατασκευάζει ερωτήσεις σε κανονική συζευκτική ή διαζευκτική μορφή το πρόβλημα έχει αντιμετωπιστεί. Επιπλέον μια τέτοια μεθοδολογία βοηθάει τον χρήστη στο να δομήσει και να εκφράσει τις σκέψεις του. Έτσι, το περιβάλλον για τον σχηματισμό ερωτήσεων που παρέχεται δίνει την δυνατότητα για τον σχηματισμό ερωτήσεων σε κανονική συζευκτική μορφή.

Οι όροι που χρησιμοποιούνται για τον σχηματισμό των ερωτήσεων είναι τα πρόσωπα, οι τοποθεσίες, οι θεματικές ενότητες, οι εκδοτικοί οίκοι και οι γλώσσες που σχετίζονται με τις αποθηκευμένες στην βάση ηλεκτρονικές εκδόσεις. Στην επόμενη εικόνα εμφανίζεται ένα στιγμιότυπο της εφαρμογής:



Εικόνα: 7-2: Στιγμιότυπο της εφαρμογής υποβολής ερωτήσεων

Δίνοντας μια σύντομη περιγραφή αυτού του στιγμιότυπου διακρίνονται:

- Η περιοχή *Relative places*, όπου παρουσιάζονται πάνω σε ένα χάρτη οι πόλεις οι οποίες σχετίζονται με τις αποθηκευμένες εκδόσεις. Οι πόλεις αυτές σημειώνονται με τις μαύρες κουκίδες.
- Η περιοχή *Relative persons,* όπου δίνεται μια λίστα με τα πρόσωπα τα οποία συνδέονται με κάποια έκδοση.

- Η περιοχή *Relative topics*, όπου παρουσιάζονται οι θεματικές ενότητες στις οποίες εντάσσονται οι αποθηκευμένες εκδόσεις.
- Η περιοχή *Ranking Attributes*, όπου δίνονται τα χαρακτηριστικά τα οποία περιγράφουν το μέγεθος (*Extent*) και το χρόνο κυκλοφορίας (*Recent*) μιας έκδοσης.
- Η περιοχή *Publishers* όπου δίνονται οι εκδοτικοί οίκοι οι οποίοι έθεσαν σε κυκλοφορία τις συγκεκριμένες εκδόσεις.
- Η περιοχή *Language* όπου δείχνονται οι γλώσσες στις οποίες είναι γραμμένα τα αποθηκευμένα κείμενα.

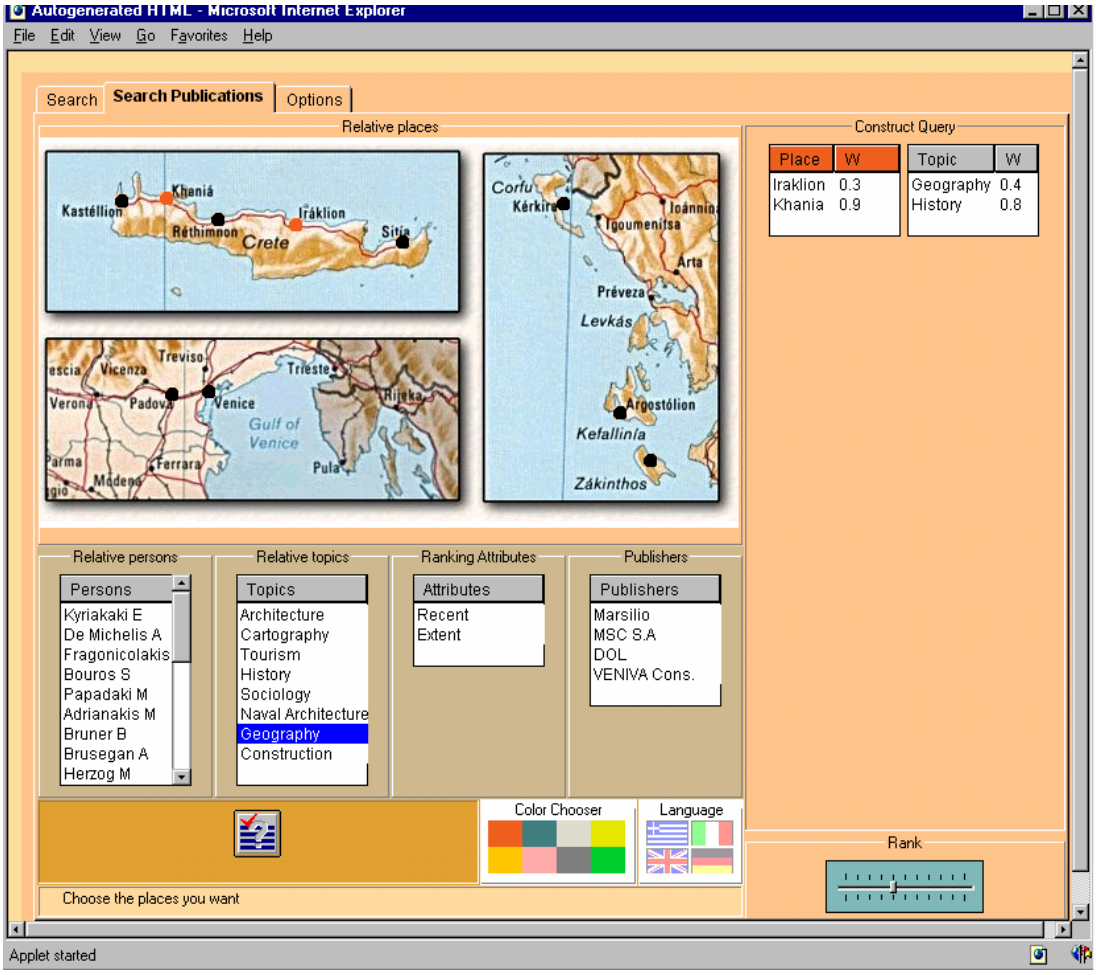
Η περιοχή *Construct Query* έχει έναν ιδιαίτερο και σημαντικό ρόλο, είναι η περιοχή στην οποίο κατασκευάζονται οι υποβαλλόμενες ερωτήσεις. Ας περιγραφεί όμως ο τρόπος με τον οποίο σχηματίζεται μια ερώτηση.

Η βασική ιδέα είναι ότι επιλέγεται ο επιθυμητός όρος και "τραβιέται" (drag and drop) μέσα στην περιοχή *Construct Query*. Όταν "αφεθεί" ο όρος μέσα σ' αυτή την περιοχή δημιουργείται μια μικρή λίστα η οποία περιέχει τον όρο που επιλέχτηκε. Η λίστα αυτή θα περιέχει ομοειδής όρους οι οποίοι συνδέονται μεταξύ τους με διάζευξη. Ο όρος ομοειδής σημαίνει ότι θα περιέχονται όροι του ίδιο τύπου, για παράδειγμα, θα περιέχονται μόνο πρόσωπα ή τοποθεσίες. Αν συνεπώς επιλεχθεί ένας δεύτερος όρος, του ίδιου τύπου και "τραβηχτεί" (drag) μέσα στην περιοχή *Construct Query* υπάρχουν δύο ενδεχόμενα. Το πρώτο ενδεχόμενο είναι ο όρος να αφεθεί μέσα σε μια υπάρχουσα λίστα ομοιειδών όρων, ενώ το δεύτερο είναι να αφεθεί μέσα στην περιοχή *Construct Query* αλλά όχι μέσα σε μια υπάρχουσα λίστα. Σ' αυτή την περίπτωση δημιουργείται μια νέα λίστα η οποία περιέχει τον επιλεγόμενο όρο. Κάθε λίστα μέσα στην περιοχή *Construct Query* αναπαριστά έναν *OR*-όρο (*OR*-λίστα) , δηλαδή οι όροι της συνδέονται μεταξύ τους με *OR*. Οι λίστες αυτές όμως συνδέονται μεταξύ τους με *AND*. Δημιουργείται κατ' αυτό τον τρόπο μια ερώτηση σε κανονική συζευκτική μορφή.

Έστω ότι υποβάλλεται η εξής ερώτηση:

"Οι εκδόσεις που σχετίζονται με τα Χανιά ή το Ηράκλειο και εντάσσονται στην ενότητα Ιστορία ή Γεωγραφία."

Ο σχηματισμός αυτής της ερώτησης θα παρουσιαζόταν όπως φαίνεται στην επόμενη εικόνα:



Εικόνα 7-3: Στιγμιότυπο της εφαρμογής υποβολής ερωτήσεων στο οποίο έχουμε σχηματίσει την ερώτηση : ((Khania, 0.9) OR (Hraklion, 0.3)) AND ((Geography, 0.4) OR(History, 0.8))

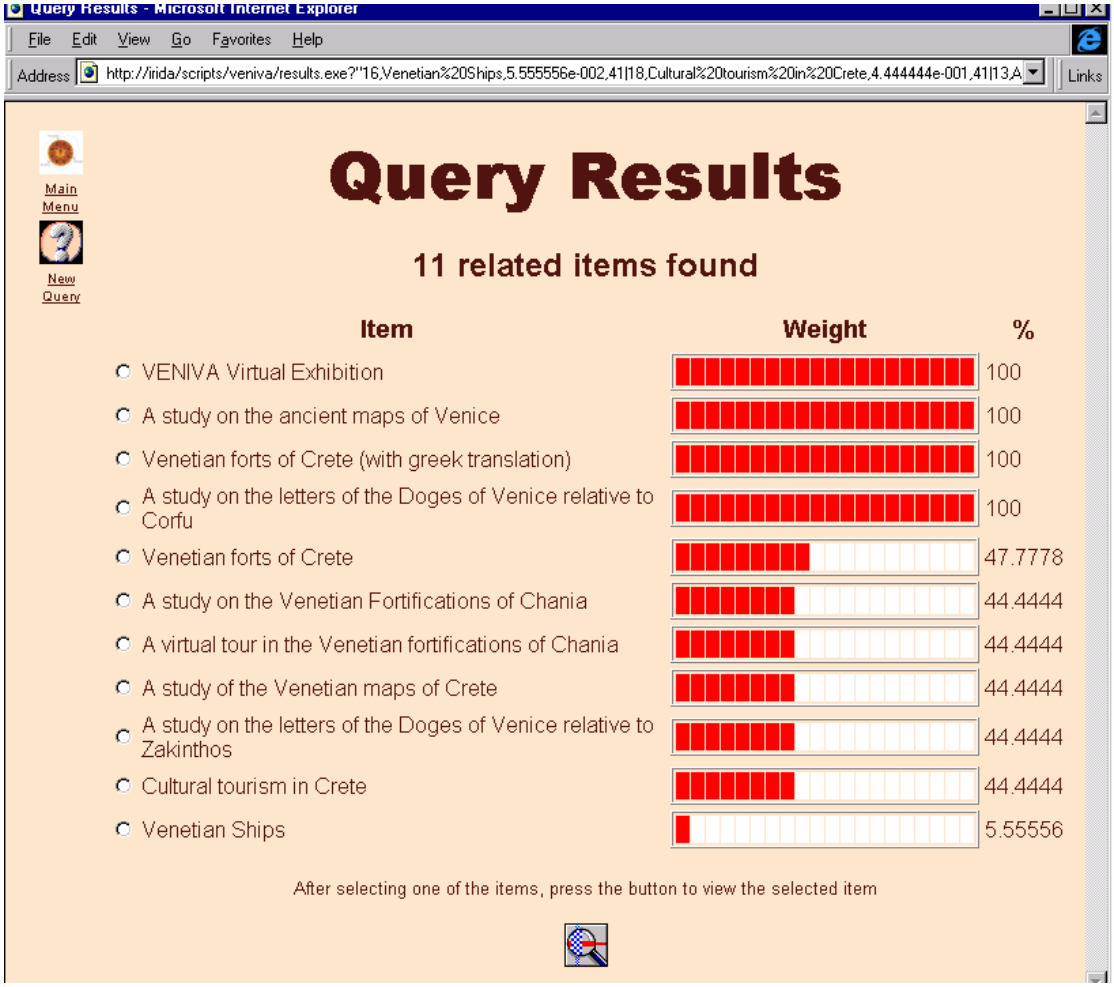
Παρέχεται επίσης η δυνατότητα να τοποθετηθεί και κάποιο βάρος στους όρους που σχηματίζουν την ερώτηση. Έτσι μπορεί να επιλεχθεί κάποιος όρος μιας *OR*-λίστας και με τη χρήση του ολισθητή (slider) στην περιοχή Rank να επιδοθεί σ' αυτόν κάποιο βάρος μεταξύ 0 και 1.

Ιδιαίτερη αναφορά θα πρέπει να γίνει στον τρόπο με το οποίο επιλέγονται οι τοποθεσίες. Αρχικά επιλέγεται κάποιο χρώμα από την περιοχή *Color Chooser*. Στην συνέχεια επιλέγονται πάνω στο χάρτη οι πόλεις οι οποίες πρέπει να συνδεθούν μεταξύ τους με *OR*. Οι επιλεγόμενες πόλεις τοποθετούνται αυτόματα μέσα σε μια *OR*-λίστα στην περιοχή *Construct Query*. Αν πρέπει να δημιουργηθεί μια νέα *OR*-

λίστα με πόλεις επιλέγεται ένα διαφορετικό χρώμα και στην συνέχεια γίνεται η επιλογή των πόλεων πάνω στο χάρτη.

Πρέπει να σημειωθεί ότι η εφαρμογή παρέχει στον χρήστη την δυνατότητα να επιλέξει το μοντέλο αξιολόγησης που επιθυμεί. Έτσι μπορεί για παράδειγμα, για μια ερώτηση να επιλέξει το μοντέλο *P-Norm* και για μια άλλη το *Waller-Kraft* ή, να εκτελέσει την ίδια ερώτηση χρησιμοποιώντας διαφορετικά μοντέλα αξιολόγησης και να μελετήσει στην συνέχεια τις ανακτηθείσες εκδόσεις και τους βαθμούς τους, εξάγοντας έτσι χρήσιμα συμπεράσματα για την συμπεριφορά των μοντέλων αξιολόγησης.

Οι απαντήσεις που παίρνονται θα έχουν όπως φαίνεται στην επόμενη εικόνα:



Εικόνα 7-4: Στιγμιότυπο της εφαρμογής υποβολής ερωτήσεων όπου παρουσιάζονται τα αποτελέσματα από την εκτέλεση μιας ερώτησης ομοιότητας

7.7 Θέματα υλοποίησης

Όπως αναφέρθηκε και στην εισαγωγή, η εφαρμογή υποβολής ερωτήσεων είναι εφαρμογή του παγκόσμιου ιστού και κατασκευάστηκε με την γλώσσα προγραμματισμού Java [15,16,17] . Η Java είναι μια καινούργια αντικειμενοστραφής γλώσσα προγραμματισμού η οποία προσφέρεται ιδιαίτερα για την ανάπτυξη εφαρμογών στο Internet.

Το βασικότερο χαρακτηριστικό της Java είναι ότι τα προγράμματά της μπορούν να εκτελεστούν σε διάφορους υπολογιστές (machine independent). Το γεγονός αυτό είναι ιδιαίτερα επιθυμητό για τις εφαρμογές στο Internet όπου υπάρχουν συνδεδεμένοι υπολογιστές με διαφορετική αρχιτεκτονική και διαφορετικά λειτουργικά συστήματα. Ένα πρόγραμμα σε Java μπορεί να είναι είτε μια αυτόνομη εφαρμογή (stand-alone application) είτε ένα Java Applet. Ένα Applet είναι στην πραγματικότητα ένα πρόγραμμα γραμμένο σε Java το οποίο όμως προβλέπεται να εκτελεστεί μέσω ενός browser. Έτσι ένα Applet μπορεί να τοποθετηθεί μέσα σε μια σελίδα υπερκειμένου (HTML) χρησιμοποιώντας τη ετικέτα (tag) <APPLET>. Τοποθετώντας Java Applets μέσα σε σελίδες υπερκειμένου αυξάνεται κατακόρυφα η λειτουργικότητά τους και οι δυνατότητες αλληλεπίδρασης με το χρήστη.

Η εφαρμογή υποβολής ερωτήσεων είναι ένα Java Applet το οποίο τοποθετείται μέσα σε μια σελίδα υπερκειμένου. Κατά την εκκίνησή του συνδέεται με την Μηχανή Ανάκτησης μέσω ενός διαύλου (socket). Έτσι οι χρήστες έχουν την δυνατότητα μέσω ενός browser να βλέπουν τη σελίδα υπερκειμένου μέσα στην οποία έχει τοποθετηθεί το Applet, να σχηματίζουν και να υποβάλλουν τις ερωτήσεις τους, χρησιμοποιώντας τις δυνατότητες που παρέχει το Applet, και φυσικά να βλέπουν τις απαντήσεις.

7.8 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκε η εφαρμογή υποβολής ερωτήσεων. Η εφαρμογή αυτή κατασκευάστηκε προκειμένου να δοθεί ένα πλήρες σενάριο χρησιμοποίησης του κατασκευασθέντος συστήματος αλλά και για να πιστοποιήσει την σωστή λειτουργία του.

Παρουσιάστηκε έτσι το σχήμα της βάσης δεδομένων πάνω στην οποία

οικοδομήθηκε το όλο σύστημα και δόθηκε μια σύντομη περιγραφή των πινάκων της. Στη συνέχεια έγινε μια ανάλυση των απαιτήσεων που θα έπρεπε, αλλά και θα μπορούσαν, να ικανοποιηθούν και πραγματοποιήθηκε ο ορισμός των συνόλων οντοτήτων, των συσχετίσεων και των χαρακτηριστικών, προς ικανοποίηση αυτών το απαιτήσεων. Προσδιορίστηκαν έτσι η συλλογή και οι όροι δεικτοδότησης του κατασκευασθέντος συστήματος ανάκτησης πληροφορίας.

Στη συνέχεια περιγράφηκε το περιβάλλον για τον σχηματισμό των ερωτήσεων. Αιτιολογήθηκε η απόφασα να δοθεί η δυνατότητα σχηματισμού ερωτήσεων σε κανονική συζευκτική μορφή και περιγράφηκε με αρκετή λεπτομέρεια ο τρόπος σχηματισμού των ερωτήσεων.

Τέλος έγινε μια σύντομη αναφορά σε θέματα σχετικά με την υλοποίηση. Αναφέρθηκε ότι εφαρμογή είναι ένα Java Applet και εξηγήθηκε εν' συντομία γιατί επιλέχτηκε η Java σαν γλώσσα υλοποίησης.

Στο επόμενο κεφάλαιο θα αναφερθούν η συνεισφορά και οι μελλοντικές επεκτάσεις της εργασίας.

8. Συνεισφορά και μελλοντικές επεκτάσεις

8.1 Εισαγωγή

Στο κεφάλαιο αυτό θα παρουσιαστεί η τελική συνεισφορά αυτής της εργασίας καθώς και κάποια γενικά συμπεράσματα. Ο σχεδιασμός αλλά και η υλοποίηση του συστήματος είναι τέτοια ώστε να καθιστά εφικτές τις όποιες μελλοντικές επεκτάσεις. Θα παρουσιαστούν έτσι κάποιες ιδέες για τη μελλοντική επέκταση και εξέλιξη του συστήματος.

Στο επόμενο λοιπόν τμήμα θα γίνει μια σύντομη ανακεφαλαίωση, ενώ, στο τμήμα 8.3 θα παρουσιαστεί η συνεισφορά της εργασίας αλλά και κάποια γενικά συμπεράσματα. Τέλος θα αναφερθούν και κάποιες πιθανές μελλοντικές επεκτάσεις.

8.2 Ανακεφαλαίωση

Η εργασία αυτή αποσκοπούσε στο σχεδιασμό και την υλοποίηση ενός συστήματος και μιας μεθοδολογίας που επιτρέπει την εκτέλεση ερωτήσεων ομοιότητας -εναλλακτικών μοντέλων υπολογισμού ομοιότητας- και ερωτήσεων δίτιμης λογικής πάνω σε σχεσιακές βάσεις δεδομένων χρησιμοποιώντας ένα υψηλό περιβάλλον ερωτήσεων (high level query interface) κατάλληλο για απλούς χρήστες. Οι ερωτήσεις ομοιότητας συναντώνται σε συστήματα ανάκτησης πληροφορίας ενώ,

στα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων εκτελούνται ερωτήσεις δίτιμης λογικής εκφρασμένες με τη γλώσσα *SQL*.

Το γεγονός ότι οι ερωτήσεις ομοιότητας εμφανίζονται στα συστήματα ανάκτησης πληροφορίας έκανε φανερή την ανάγκη κατασκευής ενός τέτοιου συστήματος το οποίο όμως θα βασίζεται πάνω σε μια σχεσιακή βάση δεδομένων. Αυτό σημαίνει ότι το σύνολο των προς ανάκτηση αντικειμένων, δηλαδή το σύνολο των αντικειμένων τα οποία θα ανακτώνται με τις ερωτήσεις ομοιότητας, αλλά και το σύνολο των όρων οι οποίοι, από τη μια θα συμβάλλουν στην αναπαράσταση των αντικειμένων και από την άλλη στο σχηματισμό των ερωτήσεων ομοιότητας, θα αποθηκεύονται σε πίνακες της βάσης δεδομένων.

Η προσέγγιση που ακολουθείται στα πλαίσια αυτής της εργασίας είναι η υλοποίηση ενός υψηλότερου επιπέδου μοντέλου, πάνω από το σχεσιακό μοντέλο, το οποίο επιτρέπει αφενός μεν τον προσδιορισμό επιπλέον σημασιολογικής πληροφορίας πάνω από το σχήμα μιας σχεσιακής βάσης δεδομένων, αφετέρου δε αυτόματους μηχανισμούς μετάφρασης των ερωτήσεων ομοιότητας σε ερωτήσεις εκφρασμένες στη γλώσσα *SQL* που επενεργούν πάνω στους πίνακες της βάσης δεδομένων. Το μοντέλο αυτό είναι μια επέκταση του γνωστού μοντέλου οντοτήτων-συσχετίσεων και χρησιμοποιείται για την περιγραφή του σχήματος του συστήματος ανάκτησης πληροφορίας που κατασκευάζεται.

Προκειμένου λοιπόν να υποστηριχτεί το μοντέλο αυτό και να οριστεί το σχήμα του κατασκευασθέντος συστήματος ανάκτησης πληροφορίας, κατασκευάστηκε ο **Ε**κδότης **Σ**χήματος (ΕΣ). Ο ΕΣ είναι ένα εργαλείο το οποίο παρέχει, μέσω ενός γραφικού περιβάλλοντος, τη δυνατότητα να περιγραφεί το σχήμα του συστήματος ανάκτησης πληροφορίας ορίζοντας σύνολα οντοτήτων και συσχετίσεις μεταξύ αυτών. Επίσης μπορούν να οριστούν και νέα ασαφή χαρακτηριστικά τα οποία υπολογίζονται μέσω αριθμητικών εκφράσεων οι οποίες συγκροτούνται από τους μαθηματικούς τελεστές που υποστηρίζει η γλώσσα *SQL* (+,-,*, /, κ.λ.π) και τα πεδία ενός πίνακα.

Προκειμένου να απαντηθούν ερωτήσεις οι οποίες είναι συνδυασμός ερωτήσεων ομοιότητας και ερωτήσεων δίτιμης λογικής, παρουσιάστηκε μια συγκεκριμένη μεθοδολογία, με βάση την οποία η εκτέλεση μιας ερώτησης ομοιότητας ανάγεται στην εκτέλεση μια σειράς *SQL* ερωτήσεων. Με βάση αυτή τη μεθολογία κάθε υποβαλλόμενη ερώτηση ομοιότητας σε πρώτη φάση "αποκωδικοποιείται" και αποθηκεύεται σε κατάλληλους πίνακες στη βάση

δεδομένων. Στην συνέχεια με την εκτέλεση τεσσάρων διαδοχικών *SQL* ερωτήσεων προκύπτει η απάντηση.

Ιδιαίτερο χαρακτηριστικό της μεθοδολογίας είναι ότι καταφέρνει να αντιμετωπίσει με ενιαίο τρόπο τις ερωτήσεις ανεξάρτητα από το μοντέλο υπολογισμού ομοιότητας που χρησιμοποιείται. Επίσης, παρέχει τη δυνατότητα να συνδυάζονται ερωτήσεις δίτιμης λογικής και ερωτήσεις ομοιότητας μέσα στην ίδια αίτηση ανάκτησης, όπου η ερώτηση δίτιμης λογικής λειτουργεί σαν φίλτρο αποκλεισμού κάποιων αντικειμένων και η ερώτηση ομοιότητας ως συνάρτηση αξιολόγησης των αντικειμένων που "πέρασαν" από το φίλτρο αποκλεισμού. Στο σύστημα υποστηρίζονται τέσσερα μοντέλα υπολογισμού ομοιότητας τα οποία ο χρήστης μπορεί να επιλέγει κάθε φορά αλλά ο αριθμός των μοντέλων που μπορούν να χρησιμοποιηθούν είναι απεριόριστος.

Για την υποστήριξη αυτής της μεθοδολογίας κατασκευάστηκε η **Μηχανή Ανάκτησης (Retrieval Engine).** Η **Μ**ηχανή **Α**νάκτησης κατασκευάστηκε με βάση το μοντέλο πελάτη – εξυπηρετητή. Δέχεται αιτήσεις, οι οποίες είναι στην πραγματικότητα ερωτήσεις ομοιότητας, από διάφορους πελάτες και στέλνει σαν απάντηση μια ταξινομημένη λίστα από τις ταυτότητες των αντικειμένων τα οποία ικανοποιούν την εκάστοτε ερώτηση μαζί με το βαθμό ομοιότητας του κάθε αντικειμένου ως προς την ερώτηση.

Αξιοσημείωτο είναι το γεγονός ότι με την προτεινόμενη μεθοδολογία μπορούν να υποστηριχθούν και διαδικασίες ανάδρασης. Αυτό επιτυγχάνεται εκτελώντας "αντίστροφες" ερωτήσεις οι οποίες δίνουν τους όρους δεικτοδότησης των ανακτηθέντων αντικειμένων τα οποία έχουν χαρακτηριστεί από το χρήστη ως καλά. Από αυτούς τους όρους μπορούν να επιλεχθούν οι καλύτεροι και να τροποποιηθεί κατάλληλα η υποβαλλόμενη ερώτηση.

Σε μια προσπάθεια να δοθεί ένα πλήρες σενάριο χρησιμοποίησης του συστήματος και να πιστοποιηθεί η σωστή λειτουργία του, κατασκευάστηκε μια εφαρμογή ψηφιακών βιβλιοθηκών. Σκοπός της εφαρμογής είναι η ανάκτηση ηλεκτρονικών εκδόσεων οι οποίες βρίσκονται αποθηκευμένες σε μια βάση δεδομένων και ικανοποιούν κάποιες απαιτήσεις του χρήστη. Έτσι, κατασκευάστηκε ένα γραφικό περιβάλλον το οποίο παρέχει στο χρήστη τη δυνατότητα να σχηματίσει και να υποβάλλει ερωτήσεις ομοιότητας με τρόπο απλό και κατανοητό. Οι εκδόσεις που ικανοποιούν μια ερώτηση παρουσιάζονται στο χρήστη μαζί με τον βαθμό ομοιότητάς τους. Ο χρήστης στην συνέχεια μπορεί να ζητήσει περισσότερες

πληροφορίες για κάθε μια από αυτές. Η εφαρμογή είναι ένα Java Applet και κατασκευάστηκε με τη γλώσσα προγραμματισμού Java η οποία είναι μια καινούργια και πολλά υποσχόμενη γλώσσα. Στο επόμενο τμήμα θα αναφερθεί η συνεισφορά της εργασίας.

8.3 Συνεισφορά

Η εργασία αυτή είναι μια προσπάθεια συνδυασμού της τεχνολογίας των συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων και των συστημάτων ανάκτησης πληροφορίας. Έγινε προσπάθεια να βρεθεί ένας τρόπος, έτσι ώστε, να συνδυαστούν οι δύο τεχνολογίες αξιοποιώντας τα πλεονεκτήματα τα οποία παρέχουν η κάθε μια ξεχωριστά. Αν και οι δύο αυτές τεχνολογίες στηρίζονται σε διαφορετικές φιλοσοφίες, και πολλές φορές παρουσιάζονται και σαν συγκρουόμενες μπορεί να θεωρηθεί ότι και οι δύο στοχεύουν στην αποθήκευση και στην ανάκτηση πληροφορίας.

Θεωρώντας λοιπόν ως τεχνολογική βάση ανάπτυξης του συστήματος τα συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων προτάθηκε η αποθήκευση της συλλογής ενός συστήματος ανάκτησης πληροφορίας στους πίνακες μιας βάσης δεδομένων. Συνήθως για την αποθήκευση της συλλογής αυτών των συστημάτων χρησιμοποιούνται ειδικές δομές όπως, για παράδειγμα, τα αντεστραμμένα αρχεία (inverted files).

Η αποθήκευση της συλλογής σε μια βάση δεδομένων αποφέρει σημαντικά οφέλη. Κατ΄αρχήν δεν χρειάζεται να κατασκευαστούν ειδικές δόμες για την αποθήκευση της συλλογής χάνοντας χρόνο και καταβάλλοντας σημαντική προσπάθεια. Επιπλέον μπορεί να γίνει πλήρης εκμεταλλευση των δυνατότητων που παρέχουν τα συστήματα διαχείρισης βάσεων δεδομένων για την αποθήκευση και την διαχείριση των δεδομένων. Αποθηκεύοντας τη συλλογή με τη χρήση πινάκων αντιμετωπίζονται εύκολα τα θέματα της εισαγωγής νεών δεδομένων και της διαγραφής και ενημέρωσης των ήδη υπαρχόντων, σε αντίθεση, με την περίπτωση που χρησιμοποιούνται ειδικές δομές αποθήκευσης. Επίσης καθίσταται δυνατός ο εμπλουτισμός υπαρχόντων βάσεων δεδομένων με νέα λειτουργικότητα. Αυτό γίνεται με την παροχή της δυνατότητας προσδιορισμού επιπλέον σημασιολογικής πληροφορίας πάνω από το σχήμα μιας σχεσιακής βάσης δεδομένων. Για την υποστήριξης των προαναφερθέντων κατασκευάστηκε ένα εργαλείο, ο Εκδότης

Σχήματος, το οποίο παρέχει τη δυνατότητα ορισμού του σχήματος ενός συστήματος ανάκτησης πληροφορίας πάνω σε μια σχεσιακή βάση δεδομένων.

Έχοντας ορίσει το σχήμα του συστήματος ανάκτησης πληροφορίας πάνω σε μια βάση δεδομένων προτάθηκε μια συγκεκριμένη μεθοδολογία, τεσσάρων φάσεων, για την εκτέλεση ερωτήσεων ομοιότητας πάνω σε σχεσιακή βάση δεδομένων. Ακολουθώντας την μεθοδολογία αυτή η εκτέλεση μιας ερώτησης ομοιότητας ανάγεται στην εκτέλεση τεσσάρων διαδοχικών *SQL* ερωτήσεων οι οποίες φυσικά κατασκευάζονται και εκτελούνται αυτόματα. Το γεγονός αυτό, από μόνο του, επιφέρει κέρδος αφού η γλώσσα *SQL* είναι απλή και σχετικά εύχρηστη. Τα ακόλουθα δύο βασικά χαρακτηριστικά αυτής της μεθολογίας είναι που την καθιστούν ακόμα ποιο ελκυστική.

Το πρώτο είναι ότι κατέστει δυνατόν να αντιμετωπιστούν τα μοντέλα υπολογισμού ομοιότητας ερώτησης-αντικειμένου με ενιαίο τρόπο. Αυτό σημαίνει ότι για να εκτελεστεί μια ερώτηση ομοιότητας, με διαφορετικά κάθε φορά μοντέλα υπολογισμού ομοιότητας, δεν πρέπει παρά να αλλαχτεί μόνο ένα τμήμα των τεσσάρων *SQL* ερωτήσεων. Κατά συνέπεια μπορούν να προστεθούν σε ένα υπάρχον σύστημα καινούρια μοντέλα γνωστοποιώντας του μόνο το γενικό σχήμα των *SQL* ερωτήσεων που θα πρέπει να εκτελεστούν προκειμένου να απαντηθεί μια ερώτηση ομοιότητας. Η προγραμματιστική δουλειά που απαιτείται είναι μικρή σε σχέση με το γεγονός ότι θα υποστηριχτεί ένα καινούργιο μοντέλο υπολογισμού ομοιότητας. Επιπλέον παρέχεται η δυνατότητα να συνδυάζονται ερωτήσεις δίτιμης λογικής και ερωτήσεις ομοιότητας μέσα στην ίδια αίτηση ανάκτησης.

Το δεύτερο χαρακτηριστικό είναι ότι μπορούν εύκολα να υποστηριχτούν διαδικασίες ανάδρασης σχετικότητας. Αυτό γίνεται εκτελώντας κατ' ουσίαν "αντίστροφες" ερωτήσεις οι οποιές δίνουν τους όρους δεικτοδότησης των ανακτηθέντων αντικειμένων τα οποία έχουν χαρακτηριστεί, από το χρήστη, ως "καλά". Από αυτούς τους όρους μπορούν να επιλεγούν οι "καλύτεροι", αυτοί δηλαδή με το μεγαλύτερο βαθμό σπουδαιότητας, για να τροποποιηθεί κατάλληλα η αρχική ερώτηση του χρήστη.

Όλα τα παράπανω υποστηρίχθηκαν με την κατασκευή της Μηχανής Ανάκτησης η οποία δέχεται ερωτήσεις και ακολουθεί τη συγκεκριμένη μεθολογία τεσσάρων φάσεων που προτάθηκε για την απάντησή τους.

Η προσπαθειά για συνδυασμό των δύο τεχνολογιών έδειξε, αν μη τι άλλο, ότι η συνεργασία τους είναι όχι μονό εφικτή αλλά μπορεί να αποφέρει και εντυπωσιακά αποτελέσματα.

8.4 Μελλοντικές επεκτάσεις

Στα πλαίσια των μελλοντικών επεκτάσεων θα μπορούσε να ενταχθεί η μελέτη της απόδοσης του συστήματος. Η στενότητα χρόνου δεν επέτρεψε να δοκιμαστεί το σύστημά πάνω σε μια μεγάλη βάση δεδομένων έτσι ώστε να μελετηθεί ο χρόνος εκτέλεσης των υποβαλλόμενων ερωτήσεων. Το γεγονός ότι πραγματοποιούνται δύο συνδέσεις (joins) θα μπορούσε να επιβαρύνει την απόδοση του συστήματος αφού η σύνδεση θεωρείται, και είναι, μια από τις πιο χρονοβόρες λειτουργίες στα συστήματα διαχείρισης βάσεων δεδομένων. Το γεγονός, ωστόσο, ότι οι συνδέσεις αυτές πραγματοποιούναι μεταξύ δύο πίνακων εκ των οποίων ο ένας θα είναι σίγουρα μικρός (είναι ο πίνακας με τους όρους που σχηματίζουν μια ερώτηση) επιτρέπει την ύπαρξη αισιοδοξίας για την απόδοση του συστήματος. Η περαιτέρω μελέτη της απόδοσης ωστόσο εξακολουθεί να είναι ένα αρκετά σημαντικό θέμα.

Μια άλλη ενδιαφέρουσα επέκταση θα ήταν να διερευνηθεί το κατά πόσον θα ήταν εφικτό να μπορεί ο χρήστης της εφαρμογής υποβολής ερωτήσεων να ορίζει δικά του ασαφή χαρακτηριστικά. Στο σύστημα που κατασκευάσαμε ο ορισμός των νέων χαρακτηριστικών γίνεται στο επίπεδο του εκδότη σχήματος. Αυτό έγινε διότι ο ορισμός ενός νέου χαρακτηριστικού έχει άμεση επίδραση στην βάση δεδομένων, εξαιτίας του ότι, δημιουργούνται κάποιες *SQL* ερωτήσεις για τον υπολογισμό της τιμής του χαρακτηριστικού. Το να δοθεί αυτή η δυνατότητα στο χρήστη της εφαρμογής υποβολής ερωτήσεων δημιουργεί τα εξής προβλήματα. Πρώτον θα πρέπει να βρεθεί ένας κοινός τρόπος ορισμού αυτών των χαρακτηριστικών για όλες τις εφαρμογές υποβολής ερωτήσεων. Αυτό είναι εξαιρετικά δύσκολο διότι οι διάφορες εφαρμογές είναι εντελώς ανεξάρτητες μεταξύ τους και διαμορφώνονται κάθε φορά με βάση συγκεκριμένες απαιτήσεις. Γι' αυτό το λόγο άλλωστε θεωρήθηκε ότι δεν αποτελούν τμήμα του συστήματος. Επιπλέον θα πρέπει να δημιουργηθούν και μηχανισμοί ελέγχου έτσι ώστε να προστατεύεται η βάση από κακόβουλους χρήστες. Παρά τις όποιες δυσκολίες όμως θα είχε ενδιαφέρον να εξεταστεί αυτή η προοπτική.

Μια επέκταση που θα μπορούσε να γίνει στην Μηχανή Ανάκτησης είναι η υποστήριξη profiles για τους χρήστες. Αυτό θα απαιτούσε αρκετή δουλειά αλλά θα

προσέδιδε σημαντική δυναμική στο σύστημα. Επίσης, η Μηχανή Ανάκτησης αυτή τη στιγμή απαντά μόνο σε ερωτήσεις. Θα μπορούσαν να προστεθούν και άλλες δυνατότητες όπως το να παρέχεται πρόσβαση μόνο σε συγκεκριμένους χρήστες ή να υπάρχουν διάφορα επίπεδα παροχής υπηρεσιών προς τους χρήστες. Επίσης θα μπορούσαν να προστέθουν λειτουργίες για την παρακολούθηση ανά πάσα στιγμή των συνδεδεμένων χρηστών και της χρήσης των όποιων πόρων του συστήματος, καθώς ακόμα και την καταγραφή στατιστικών στοιχείων .

Βιβλιογραφία

[1] G. Salton, *"Automatic Information Organization and Retrieval",* McGraw-Hill Book Co., NY, 1968, Chapter 2, 21-65.

[2] G.Salton, A. Wong, and C.S. Yang, "*A vector Space Model for Automatic Indexing*", Communications of the ACM, 18:11, November 1975, 613-620.

[3] V.V. Raghavan and S.K.M. Yong, "*A Critical Analysis of the Vector Space Model for Information Retrieval*", Journal of the American Society for Information Science, September 1986

[4] J.J. Rocchio, Jr., "*Relevance Feedback in Information Retrieval*", in The Smart System – "*Experiments in Automatic Document Processing*", G.Salton editor, Prentice-Hall Inc., Englewood Cliffs, NJ, 1971, 337-357.

[5] Lee JH, Kim WY, Kim MH, Lee YJ. "*On the evaluation of boolean operators in the extended boolean retrieval framework*", in Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1993, pp 291-297

[6] Buell DA. "*A general model of query processing in information retrieval system*", in Information Processing & Management 1981

[7] Radecki T. "*Fuzzy set theoretical approach to document retrieval*", in Information Processing & Management 1979

[8] Sachs WM. "*An approach to associative retrieval through the theory of fuzzy sets*", Journal of the American Society for Information Science, Vol. 27, pp.85-86, 1976.

[9] Zimmerman HJ. "*Fuzzy set theory and its applications*", 2nd edition, Kluwer Academic Publishers, 1991

[10] Lee JH, "*Properties of Extended Boolean Models in Information Retrieval*", SIGIR '94, pp 182-190.

[11] G. Salton, *"Automatic Text Processing: the transformation, analysis, and retrieval of information by computer"* , Addison-Weslay, 1989.

[12] David J. Kruglinski, *"Inside Visual C++"*, Microsoft Press, 1994.

[13] P. Litwin, K. Getz, "*Microsoft Access 95 Developer's Handbook*", Sybex, 1996

[14] Μουμουτζής Ν.,"*Σχεδιασμός Συστήματος Ανάπτυξης Αλληλεπιδραστικών Γεωγραφικών Εφαρμογών*", Διατριβή Μεταπτυχιακής Ειδίκευσης, Τμήμα Ηλεκτρονικής και Μηχανικών Ηλεκτρονικών Υπολογιστών, Πολυτεχνείο Κρήτης, 1997.

[15] Tim Ritchey, "*Java!*", New Riders, 1995

[16] Arthur van Hoff, Sami Shaio, and Orca Starbuck, "*Hooked on Java*", Addison-Wesley, 1996.

[17] Luke Cassady-Dorion, "*Industrial Strength Java*", New Riders, 1997

[18] R.Kruse, J. Gebhardt, F. Klawonn, *"Foundations of Fuzzy Systems",* John Wiley & Sons, Inc. 1994.

[19] Jamie Jaworski, *"Java Developer's Guide",* First Edition, Sams, 1996