Κεφάλαιο 1. Εισαγωγή[1]

1.1 Η ανάγκη για Ανάκτηση Πληροφορίας

Η επιστήμη της *Ανάκτησης Πληροφορίας* (ΑΠ στο εξής), ασχολείται με την αναπαράσταση, την αποθήκευση, την οργάνωση και την πρόσβαση σε πληροφοριακά αντικείμενα. Η αναπαράσταση και η οργάνωση των πληροφοριακών αντικειμένων πρέπει να γίνονται με τρόπο, ώστε να παρέχουν στον εκάστοτε χρήστη, εύκολη πρόσβαση στην πληροφορία που τον ενδιαφέρει. Δυστυχώς ο καθορισμός της *πληροφοριακής ανάγκης* του χρήστη, είναι ένα δύσκολο πρόβλημα.

Το παραπάνω πρόβλημα αντικατοπτρίζεται, για παράδειγμα, στην έκφραση της ακόλουθης πληροφοριακής ανάγκης στο χώρο του Διαδικτύου.

> *Ανάκτησε όλες τις σελίδες που περιέχουν πληροφορίες για κινηματογραφικές ταινίες στις οποίες: (1) Πρωταγωνιστεί ο Κλίντ Ήστγουντ, (2) είναι γουέστερν, (3) υπάρχουν σε DVD.* Σελίδες σχετικές με το παραπάνω ερώτημα θα πρέπει να περιέχουν πληροφορίες, για τους συντελεστές της ταινίας, κριτικές καθώς και περίληψη του σεναρίου.

Είναι εμφανής η δυσκολία έκφρασης της παραπάνω πληροφοριακής ανάγκης με πληρότητα, χρησιμοποιώντας το περιβάλλον διεπαφής μιας *Διαδικτυακής Μηχανής Αναζήτησης*. Συνεπώς ο χρήστης πρέπει να είναι σε θέση να επαναδιατυπώσει την πληροφοριακή ανάγκη, σε μορφή *ερωτήματος* (query), το οποίο να μπορεί να γίνεται αντικείμενο επεξεργασίας από την μηχανή αναζήτησης (ή το σύστημα ΑΠ).

Η μετατροπή αυτή συνήθως γίνεται με τη χρήση, ενός συνόλου λέξεων κλειδιών (keywords) ή ισοδύναμα όρων δεικτοδότησης (index terms), που συνοψίζουν την περιγραφή της πληροφοριακής ανάγκης του χρήστη. Δοθέντος του ερωτήματος του χρήστη, το ζητούμενο από ένα σύστημα ΑΠ είναι να *ανακτήσει πληροφορία*, η οποία μπορεί να είναι χρήσιμη ή σχετική προς την πληροφοριακή ανάγκη. Έμφαση δίνεται στην ανάκτηση πληροφορίας σε αντίθεση με την ανάκτηση δεδομένων τη διαφορά των οποίων θα εξετάσουμε αμέσως.

1.1.1 Ανάκτηση Πληροφορίας και όχι Δεδομένων

Η ανάκτηση δεδομένων σε ένα περιβάλλον ΑΠ, συνίσταται στην εύρεση όλων των κειμένων τα οποία περιέχουν κάποιες από τις λέξεις κλειδιά που εμφανίζονται σε ένα ερώτημα προς το σύστημα. Αυτή η προσέγγιση δίνει συχνά κάτι διαφορετικό από αυτό που πραγματικά θέλει ο χρήστης. Στην πράξη, αυτό που περισσότερο ενδιαφέρει τον χρήστη ενός συστήματος ΑΠ, είναι να *ανακτήσει πληροφορίες* για ένα συγκεκριμένο θέμα, παρά η ανάκτηση δεδομένων σχετικών με κάποιο ερώτημα. Μια γλώσσα ανάκτησης δεδομένων, στοχεύει στην ανάκτηση όλων των αντικειμένων, που ικανοποιούν ένα σύνολο καλά ορισμένων συνθηκών, που διατυπώνονται με μια κανονική έκφραση ή με χρήση των εργαλείων της σχεσιακής άλγεβρας. Επίσης σε ένα σύστημα ανάκτησης δεδομένων (βλ. μια σχεσιακή βάση δεδομένων), τα δεδομένα είναι οργανωμένα σε μία καλά ορισμένη δομή και έχουν συγκεκριμένη σημασιολογία. Έτσι σε ένα σύστημα ανάκτησης δεδομένων, η ανάκτηση ενός και μόνο λανθασμένου αποτελέσματος, θεωρείται ένδειξη εσφαλμένης λειτουργίας του μηχανισμού

[1] Βασική πηγή για τα κεφάλαια 1-3 είναι η αναφορά [BR99]

ανάκτησης. Αντίθετα στα συστήματα ΑΠ, τα ανακτόμενα αποτελέσματα μπορεί να είναι ανακριβή και η εμφάνιση κάποιων λαθών στα αποτελέσματα, περνά συχνά απαρατήρητη. Ο λόγος αυτής της διαφοροποίησης είναι ότι το σύστημα ΑΠ, διαχειρίζεται κείμενα γραμμένα σε φυσική γλώσσα, τα οποία δεν είναι πάντα επαρκώς δομημένα και είναι συχνά αμφίσημα. Μην ξεχνάμε άλλωστε και την δυσκολία της διατύπωσης της ακριβούς πληροφοριακής ανάγκης με τη χρήση λέξεων κλειδιών.

Έτσι ενώ η ανάκτηση δεδομένων δίνει λύσεις στο χρήστη ενός συστήματος βάσης δεδομένων, δεν λύνει το πρόβλημα της ανάκτησης πληροφορίας, σχετικής με κάποιο θέμα. Για να μπορέσει ένα σύστημα ΑΠ να ανταποκριθεί στην πληροφοριακή ανάγκη του χρήστη, θα πρέπει να είναι σε θέση, να 'διερμηνεύσει' με κάποιον τρόπο το σημασιολογικό περιεχόμενο το αντικειμένων (κείμενα) που διαχειρίζεται, και να τα διατάξει σύμφωνα με το βαθμό σχετικότητάς τους προς το ερώτημα του χρήστη. Η διαδικασία της 'διερμηνείας' συνίσταται στην εξαγωγή συντακτικής και σημασιολογικής πληροφορίας από τα κείμενα, η οποία θα χρησιμοποιηθεί για να ανταποκριθεί το σύστημα στην πληροφοριακή ανάγκη του χρήστη. Το πρόβλημα δεν εντοπίζεται μόνο στην εξαγωγή της παραπάνω πληροφορίας. Επιπλέον θα πρέπει να είναι εφικτή η χρήση της εξαγόμενης πληροφορίας για να αποφασιστεί η σχετικότητα προς κάποιο ερώτημα. Ο κύριος στόχος άλλωστε ενός συστήματος ΑΠ, είναι να μπορεί να επιστρέψει όλα τα κείμενα που είναι σχετικά προς κάποιο ερώτημα, ανακτώντας παράλληλα και όσο το δυνατόν λιγότερα μη σχετικά κείμενα. Γι' αυτό το λόγο η έννοια της *σχετικότητας*, διαδραματίζει κυρίαρχο ρόλο στην ανάκτηση πληροφορίας.

1.1.2 Η Ανάκτηση Πληροφορίας στο κέντρο του ενδιαφέροντος

Η αρχική ανάγκη για ανάπτυξη της ανάκτησης πληροφορίας ήταν η αυτοματοποιημένη δεικτοδότηση κειμένων και η ανάπτυξη μεθόδων για την αναζήτηση χρήσιμων κειμένων σε μια συλλογή. Στις ημέρες μας η έρευνα έχει επεκταθεί σε πολλούς παραπάνω τομείς, συμπεριλαμβάνοντας την μοντελοποίηση, την ταξινόμηση και κατηγοριοποίηση κειμένων, την οπτικοποίηση δεδομένων, τις διεπαφές προς τον χρήστη μηχανές ψαξίματος στο Παγκόσμιο Ιστό, συστήματα φιλτραρίσματος πληροφορίας, συστήματα προσαρμοστικών υπερμέσων, συστήματα εκπαιδευτικού λογισμικού, Βιοπληροφορική. Η άποψη που επικρατούσε μέχρι στις αρχές τις δεκαετίας του 90, ήταν ότι η ανάκτηση πληροφορίας απευθυνόταν μόνο σε εφαρμογές βιβλιοθηκονομίας. Όλα τα παραπάνω άλλαξαν δραματικά τα τελευταία χρόνια και κυρίως μετά την έλευση του Παγκοσμίου Ιστού.

Ο Παγκόσμιος Ιστός γίνεται μια ολοένα και μεγαλύτερη παρακαταθήκη ανθρώπινης γνώσης, που επιτρέπει την ανταλλαγή πληροφορίας και ιδεών σε έκταση πολύ μεγαλύτερη από ότι είχαμε δει μέχρι τώρα. Η επιτυχία του Ιστού συνίσταται στην ευκολία που παρέχει στο χρήστη να δημιουργήσει τις δικές του Ιστοσελίδες, όντας έτσι ένα εύκολα προσβάσιμο και σχετικά φθηνό μέσο προσωπικής έκφρασης. Επιπλέον η ύπαρξη του Ιστού, θέτει νέους τρόπους επικοινωνίας επανορίζοντας τις έννοιες απόσταση και χρόνος. Τέλος οι τρέχουσες εξελίξεις στην ολοκλήρωση διαφορετικών υπηρεσιών γύρω από τον Ιστό, έχουν αλλάξει τον τρόπο που ο άνθρωπος βλέπει τον υπολογιστή. Έννοιες όπως Ηλεκτρονικό Εμπόριο και Ψηφιακές Βιβλιοθήκες είναι δημοφιλείς και δημιουργούν νέες και πολλά υποσχόμενες αγορές.

Παρά την επιτυχημένη διάδοση του Παγκοσμίου Ιστού, η εύρεση χρήσιμης πληροφορίας στις Ιστοσελίδες, γίνεται μια ολοένα και πιο δύσκολη και επίπονη διαδικασία. Μια προσέγγιση εδώ είναι ο χρήστης να περιπλανιέται στον Κυβερνοχώρο, ακολουθώντας συνδέσμους που οδηγούν από σελίδα σε σελίδα, και να προσπαθεί να εντοπίσει την πληροφορία που καλύπτει την πληροφοριακή του ανάγκη. Η παραπάνω διαδικασία περιπλάνησης, είναι συχνά αναποτελεσματική, λόγω του μεγέθους του Παγκοσμίου Ιστού και γιατί τις περισσότερες φορές ο χρήστης δεν γνωρίζει ένα καλό 'σημείο εκκίνησης'. Για τους άπειρους χρήστες, το πρόβλημα της αναζήτησης γίνεται πολύ πιο δύσκολο, συχνά οδηγώντας τους σε απογοητευτικά αποτελέσματα. Το κύριο εμπόδιο εδώ, είναι η απουσία ενός καλά ορισμένου μοντέλου

δεδομένων για των Παγκόσμιο Ιστό, το οποίο σημαίνει ότι ο ορισμός και η δόμηση της πληροφορίας είναι ελλιπείς. Αυτές οι δυσκολίες έστρεψαν το ενδιαφέρον στον τομέα της ΑΠ και οδήγησαν στην υιοθέτηση των τεχνικών που χρησιμοποιούνται στο πεδίο της ΑΠ, ως πολλά υποσχόμενες λύσεις.

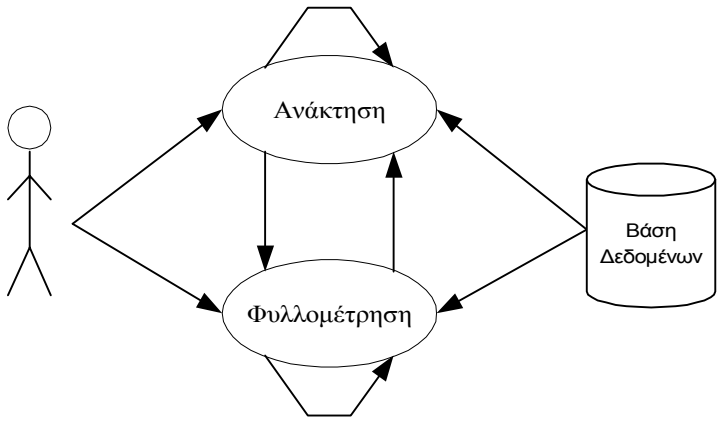
1.2 Βασικές έννοιες

Η αποδοτική ανάκτηση κειμένων αποτελεί συνάρτηση τόσο της *διαδικασίας χρήστη* όσο και της *λογικής αναπαράστασης* των κειμένων, όπως αυτή υιοθετείται από το σύστημα. Τις δύο αυτές παραμέτρους θα συζητήσουμε ευθύς αμέσως

1.2.1 Η διαδικασία χρήστη

Σε ένα σύστημα ανάκτησης, ο χρήστης πρέπει να μετατρέψει την πληροφοριακή του ανάγκη σε μορφή ερωτήματος σύμφωνα με την γλώσσα που του παρέχεται από το σύστημα. Σε ένα σύστημα ανάκτησης πληροφορίας, η παραπάνω διαδικασία ανάγεται στην επιλογή από τον χρήστη, ενός καταλλήλου συνόλου λέξεων, αντιπροσωπευτικές για τη σημασιολογία της πληροφοριακής του ανάγκης. Σε ένα σύστημα ανάκτησης δεδομένων, η διατύπωση ενός ερωτήματος, για παράδειγμα με τη χρήση μιας κανονικής έκφρασης συνίσταται στον καθορισμό του συνόλου των περιορισμών που θα πρέπει να ικανοποιεί το σύνολο της απάντησης. Και στις δύο περιπτώσεις, λέμε πως ο χρήστης αναζητά χρήσιμη πληροφορία και κατά συνέπεια εκτελεί μια διαδικασία *ανάκτησης*.

Έχοντας περιγράψει σε γενικές γραμμές την διαδικασία της αναζήτησης ας εξετάσουμε μια δεύτερη διαδικασία ανάκτησης, τη *φυλλομέτρηση* (browsing). Έστω ότι το ενδιαφέρον του χρήστη είτε δεν είναι καλά ορισμένο είτε καλύπτει ένα αρκετά ευρύ φάσμα πληροφοριών. Για παράδειγμα ο χρήστης μπορεί να ενδιαφέρεται για κείμενα σχετικά με αγώνες αυτοκινήτου. Σ' αυτή την περίπτωση θα μπορούσε ο χρήστης απλά να διαβάζει κείμενα από μια συλλογή για αγώνες αυτοκινήτου. Θα μπορούσε, για παράδειγμα, να βρει ενδιαφέροντα κείμενα σχετικά με αγώνες Φόρμουλα Ένα, κατασκευαστές αυτοκινήτων ή ακόμα και για τον αγώνα '24 ωρών του Λε Μαν'. Την ώρα που θα διαβάζει για τις '24 ώρες του Λε Μαν', μπορεί να στρέψει την προσοχή του σε μια παραπομπή για οδηγίες πρόσβασης στο σιρκουί του Λε Μαν και από 'κει για τον τουρισμό στη Γαλλία. Σ' αυτή την περίπτωση λέμε ότι ο χρήστης δεν ψάχνει τη συλλογή αλλά *φυλλομετρά* (browses), τα κείμενά της. Η φυλλομέτρηση είναι κι αυτή μια διαδικασία ανάκτησης πληροφορίας, της οποίας όμως οι σκοποί δεν είναι ξεκάθαρα προσδιορισμένοι τη στιγμή της εκκίνησης και που μπορεί να μεταβληθούν κατά τη διάρκεια της αλληλεπίδρασης με το σύστημα.



Εικόνα 1-1: Αλληλεπίδραση του χρήστη με το σύστημα ΑΠ

Η διαδικασία χρήστη σε ένα σύστημα ανάκτησης μπορεί να λαμβάνει δύο διακριτές μορφές: *ανάκτηση* δεδομένων ή πληροφορίας και *φυλλομέτρηση*. Τα κλασσικά συστήματα ανάκτησης πληροφορίας παρέχουν συνήθως μόνο τη δυνατότητα ανάκτησης. Για παράδειγμα στο σύστημα μιας βιβλιοθήκης, παρέχεται απλά η δυνατότητα ανάκτησης της βιβλιογραφίας που αντιστοιχεί για παράδειγμα σε ένα συγγραφέα. Στη συγκεκριμένη περίπτωση όμως η πληροφοριακή ανάγκη είναι πολύ συγκεκριμένη, ένας συγγραφέας. Τα συστήματα Υπερκειμένου (Hypertext), είναι συνήθως κατασκευασμένα με γνώμονα την εύκολη φυλλομέτρηση. Στις μοντέρνες Ψηφιακές Βιβλιοθήκες όμως καθώς και στις Μηχανές Αναζήτησης στο Παγκόσμιο Ιστό, υπάρχει προσπάθεια να συνδυαστούν οι δύο παραπάνω μορφές για την βελτίωση των δυνατοτήτων ανάκτησης.

Η Εικόνα 1-1 δείχνει την αλληλεπίδραση με το χρήστη μέσα από τις διαφορετικές μορφές διαδικασίας χρήστη που αναφέραμε. Αξίζει να σημειωθεί ότι οι μορφές διαδικασίας χρήστη μπορούν να εναλλάσσονται. Τα περισσότερα σύγχρονα συστήματα ΑΠ, παρέχουν τη δυνατότητα ανάκτησης δεδομένων και πληροφορίας. Επίσης τα περισσότερα από αυτά συνήθως παρέχουν και κάποιες στοιχειώδεις μορφές φυλλομέτρησης (συνήθως οδηγώντας μέσω υπερσυνδέσμου σε κάποια σελίδα που επιστράφηκε ως αποτέλεσμα μιας ερώτησης).

Στην ορολογία που χρησιμοποιείται στον Παγκόσμιο Ιστό, τόσο η διαδικασία της αναζήτησης όσο και η αντίστοιχη της φυλλομέτρησης είναι διαδικασίες '*ανάσυρσης*' (pulling). Αυτό σημαίνει ότι οι χρήστες διατυπώνουν μια απαίτηση πληροφορίας και ανασύρουν από τη συλλογή σχετικά κείμενα. Τη συγκεκριμένη διαδικασία θα συναντήσουμε στο Κεφάλαιο 3 και ως *ad hoc* ανάκτηση. Η αντίστροφη διαδικασία λέγεται '*προώθηση*' (pushing). Αυτή συνίσταται στον καθορισμό μιας σταθερής πληροφοριακής ανάγκης από τον χρήστη και της αποστολής ενός *πράκτορα λογισμικού* (software agent), ο οποίος εξετάζει την συνολική παρεχόμενη πληροφορία και 'προωθεί' τα σχετικά κείμενα προς το χρήστη. Για παράδειγμα κάποιος χρήστης θα ήθελε να παρακολουθεί από μια λίστα συζητήσεων, μόνο τα μηνύματα που τον *ενδιαφέρουν*. Η παραπάνω διαδικασία λέγεται και *φιλτράρισμα πληροφορίας* (information filtering) και θα την εξετάσουμε με συντομία στο Κεφάλαιο 3.

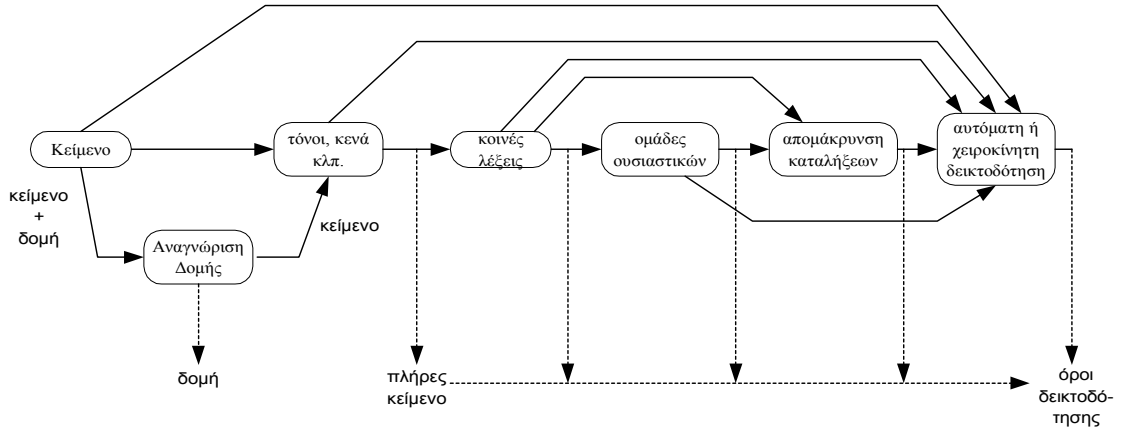
1.2.2 Λογική αναπαράσταση των κειμένων

Για ιστορικούς κυρίως λόγους, τα κείμενα μιας συλλογής αναπαρίστανται συνήθως μέσω ενός συνόλου από όρους δεικτοδότησης (index terms) ή λέξεις-κλειδιά (keywords). Τέτοιες λέξεις-κλειδιά μπορεί να εξάγονται αυτόματα ή να παρέχονται από τον ανθρώπινο παράγοντα (όπως συνηθίζεται σε επιστημονικές δημοσιεύσεις). Ανεξάρτητα με το αν αυτές οι λέξεις κλειδιά παράγονται από κάποιον ειδικό ή εξάγονται αυτόματα, μας παρέχουν μια λογική αναπαράσταση των κειμένων.

Με τους σύγχρονους υπολογιστές μας παρέχεται η δυνατότητα να αναπαραστήσουμε το πλήρες σύνολο όρων που αποτελούν ένα κείμενο. Σ' αυτήν την περίπτωση λέμε ότι έχουμε αναπαράσταση *πλήρους κειμένου*. Σε πολύ μεγάλες συλλογές κειμένων όμως (βλ. Μηχανή Αναζήτησης), οι ανάγκες για αποθήκευση είναι τεράστιες, οπότε ακόμα και με τους σημερινούς υπολογιστές, χρειάζεται να μειώσουμε το μέγεθος της αναπαράστασης. Έτσι καταφεύγουμε σε λύσεις όπως, απομάκρυνση των πιο *κοινών λέξεων* (άρθρα και σύνδεσμοι που καταλαμβάνουν το 40% περίπου των κειμένων), *απομάκρυνση καταλήξεων* (κρατάμε μόνο τη γραμματική ρίζα των λέξεων) και την αναγνώριση ομάδων από ουσιαστικά (απομακρύνοντας ρήματα, επίθετα, επιρρήματα). Τέλος μπορεί να εφαρμοστεί και συμπίεση. Όλες οι παραπάνω ενέργειες ονομάζονται *πράξεις σε κείμενο*. Σκοπός αυτών των πράξεων είναι να μειώσουν την πολυπλοκότητα αναπαράστασης των κειμένων και να μας οδηγήσουν από την αναπαράσταση πλήρους κειμένου, σε αυτή των *όρων δεικτοδότησης*.

Το πλήρες κείμενο είναι σίγουρα η πιο ολοκληρωμένη λογική αναπαράσταση ενός κειμένου αλλά η χρήση της συνήθως δεν είναι υπολογιστικά αποδοτική. Ένα μικρό σύνολο από

σημασιολογικές κατηγορίες, που παράγονται από εξειδικευμένο ανθρώπινο παράγοντα, είναι η πιο σύντομη και περιεκτική μορφή αναπαράστασης αλλά η χρήση της μπορεί να οδηγήσει σε χαμηλή ποιότητα ανάκτησης. Μεταξύ των δύο αυτών αναπαραστασιακών άκρων, βρίσκονται διάφορα επίπεδα λογικής αναπαράστασης, που μπορούν να χρησιμοποιηθούν για την λογική αναπαράσταση, όπως φαίνεται στην Εικόνα 1-2. Εκτός από τα ενδιάμεσα αυτά στάδια αναπαράστασης, το σύστημα είναι ίσως δυνατόν να αναγνωρίζει και κάποια δομικά στοιχεία, που συνήθως εμφανίζονται σε ένα κείμενο (π.χ. κεφάλαια, ενότητες, παράγραφοι κλπ.). Αυτή η πληροφορία μπορεί να είναι χρήσιμη και κυρίως όσον αφορά μοντέλα ανάκτησης δομημένου κειμένου, τα οποία όμως δεν αναπτύσσουμε σε αυτές τις σημειώσεις.



Εικόνα 1-2: Λογική Αναπαράσταση κειμένου, από το πλήρες κείμενο στους όρους δεικτοδότησης

1.3 Η διαδικασία της ανάκτησης

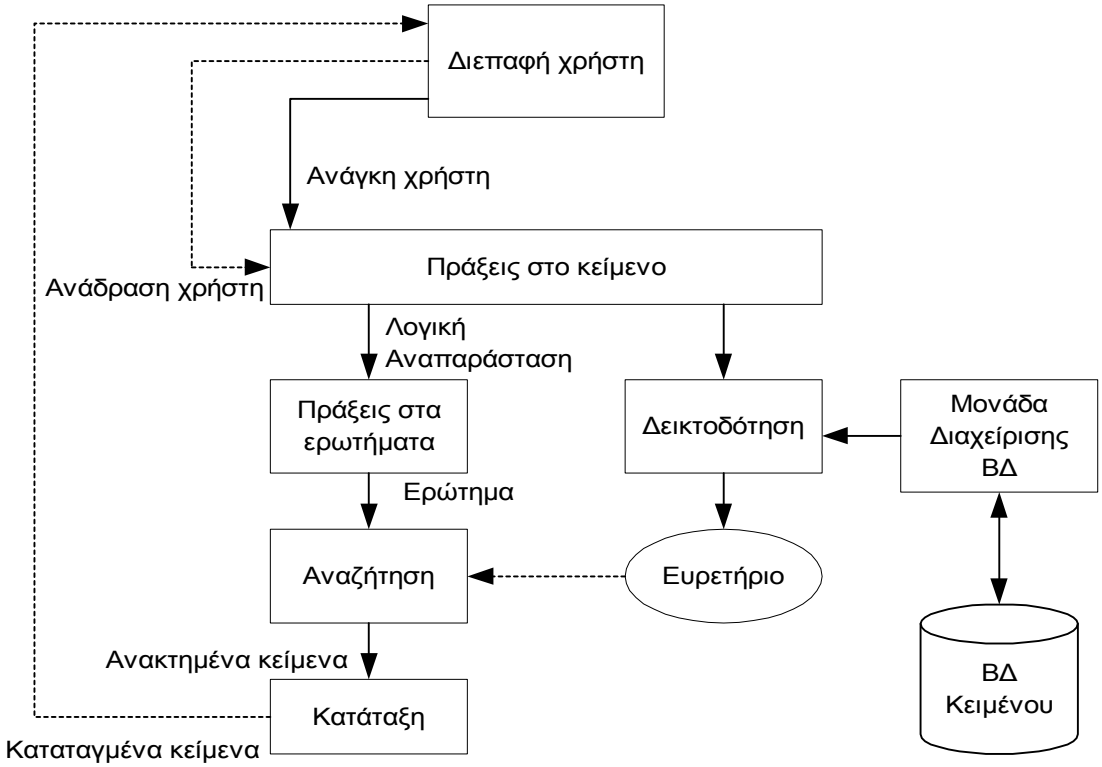
Για να περιγράψουμε τη διαδικασία της ανάκτησης, χρησιμοποιούμε μια απλή και γενικευμένη αρχιτεκτονική λογισμικού, όπως αυτή που φαίνεται στην Εικόνα 1-3. Πρώτ' απ' όλα πριν αρχικοποιηθεί η διαδικασία ανάκτησης, πρέπει να οριστεί η βάση δεδομένων των κειμένων. Αυτό συνήθως γίνεται από τον υπεύθυνο της βάσης δεδομένων, ο οποίος ορίζει τα εξής: α) τα κείμενα που θα χρησιμοποιηθούν, β) τις πράξεις που θα εφαρμοστούν στα κείμενα γ) το μοντέλο των κειμένων (δηλ. τη δομή των κειμένων και ποια είναι τα ανακτόμενα στοιχεία)

Από τη στιγμή που καθορίζεται η λογική αναπαράσταση των κειμένων, ο υπεύθυνος της ΒΔ, κατασκευάζει χρησιμοποιώντας τη Μονάδα Διαχείρισης ΒΔ, το ευρετήριο (index) των κειμένων. Το ευρετήριο είναι μια πολύ κρίσιμη δομή δεδομένων, γιατί επιτρέπει αποδοτική αναζήτηση σε μεγάλο όγκο δεδομένων. Μπορεί να χρησιμοποιηθεί μεγάλη ποικιλία δομών δεικτοδότησης αλλά η πιο δημοφιλής δομή είναι αυτή των *ανεστραμμένων αρχείων*. Τα έξοδα σε χώρο και χρόνο για τον καθορισμό της βάσης δεδομένων και την κατασκευή του ανεστραμμένου αρχείου, κατανέμονται εκτελώντας πολλά ερωτήματα πάνω στη βάση.

Δεδομένου του ότι έχουμε κατασκευάσει ευρετήριο για τη βάση δεδομένων, η διαδικασία της ανάκτησης μπορεί να ξεκινήσει. Ο χρήστης αρχικά καθορίζει μια *ανάγκη χρήστη*, η οποία αναλύεται συντακτικά και στην οποία εφαρμόζονται όλες οι πράξεις που εφαρμόζονται και στα κείμενα της βάσης. Στη συνέχεια πρέπει να εφαρμοστούν οι λεγόμενες *πράξεις στο ερώτημα* (query operations), για να προκύψει το πραγματικό ερώτημα, το οποίο αποτελεί αναπαράσταση, σε επίπεδο συστήματος, της ανάγκης χρήστη. Κατόπιν το ερώτημα επεξεργάζεται για να προκύψουν τα *ανακτημένα κείμενα.* Η επεξεργασία του ερωτήματος γίνεται γρήγορα, χάρη στο ευρετήριο, που χτίστηκε στο προηγούμενο βήμα.

Πριν παρουσιαστούν τα αποτελέσματα στο χρήστη, τα ανακτημένα κείμενα κατατάσσονται με βάση μια εκτίμηση για σχετικότητά τους. Στη συνέχεια, ο χρήστης εξετάζει το σύνολο των καταταγμένων κειμένων για να εντοπίσει χρήσιμη πληροφορία. Σ' αυτό το σημείο μπορεί να καταδείξει μια σειρά από κείμενα που είναι βέβαιο ότι ικανοποιούν την πληροφοριακή του ανάγκη και να ξεκινήσει έτσι έναν κύκλο *ανάδρασης χρήστη* (user feedback). Κατά τη διάρκεια ενός τέτοιου κύκλου, το σύστημα χρησιμοποιεί τα κείμενα που επιλέχθηκαν από τον χρήστη για να επαναδιατυπώσει το ερώτημα, με την ελπίδα ότι το επαναδιατυπωμένο ερώτημα είναι καλύτερη αναπαράσταση της πραγματικής ανάγκης χρήστη.

Δεδομένων των διαθέσιμων διεπαφών χρήστη που είναι διαθέσιμες στα σύγχρονα συστήματα ΑΠ (Μηχανές Αναζήτησης και Web browsers), εύκολα διαπιστώνει κανείς ότι ο χρήστης δεν διατυπώνει σχεδόν ποτέ την πραγματική του πληροφοριακή ανάγκη. Αυτό που στην πράξη συμβαίνει, είναι ο χρήστης να καλείται να παρέχει μια διατύπωση του ερωτήματος που θα επεξεργαστεί το σύστημα. Καθώς οι περισσότεροι χρήστες δεν έχουν γνώση των πράξεων που εφαρμόζονται στο κείμενο και στα ερωτήματα, το ερώτημα που παρέχουν είναι συχνά ανεπαρκώς διατυπωμένο. Γι' αυτό δεν ξενίζει το γεγονός ότι ελλιπώς διατυπωμένα ερωτήματα, οδηγούν σε κακή ανάκτηση πληροφορίας (όπως συμβαίνει συχνά στο Διαδίκτυο).



Εικόνα 1-3: Η Διαδικασία της Ανάκτησης Πληροφορίας

Κεφάλαιο 2. Μετρικές Εκτίμησης Απόδοσης Ανάκτησης

2.1 Εισαγωγή

Σε ένα σύστημα το οποίο είναι σχεδιασμένο για Ανάκτηση Δεδομένων, ο *χρόνος απόκρισης* και ο απαιτούμενος *χώρος* είναι συνήθως οι μετρικές απόδοσης που παρουσιάζουν το μεγαλύτερο ενδιαφέρον. Στην περίπτωση αυτή αντικείμενο των μετρικών είναι ο έλεγχος της απόδοσης των δομών δεικτοδότησης (που χρησιμοποιούνται για την επιτάχυνση του

ψαξίματος), η επικοινωνία με το λειτουργικό σύστημα, οι καθυστερήσεις στους διαύλους επικοινωνίας και οι επιβαρύνσεις που εισάγονται από τα πολλά επίπεδα λογισμικού που παρεμβάλλονται. Η χρήση τέτοιων μετρικών αναφέρεται απλώς ως *Εκτίμηση Απόδοσης Συστήματος*.

Σε ένα σύστημα το οποίο είναι σχεδιασμένο για *Ανάκτηση Πληροφορίας*, υπάρχουν και άλλες μετρικές που είναι ενδιαφέρουσες. Στην πραγματικότητα, εφόσον οι ερωτήσεις που υποβάλλουν οι χρήστες είναι εγγενώς ασαφείς, τα ανακτόμενα κείμενα δεν είναι ακριβείς απαντήσεις και θα πρέπει να διαταχθούν με βάση τη σχετικότητά τους με το υποβαλλόμενο ερώτημα. Αυτή η διάταξη σχετικότητας εισάγει ένα συστατικό στη λειτουργία του συστήματος που δεν υπάρχει σε συστήματα Ανάκτησης Δεδομένων και που παίζει ουσιώδη ρόλο στην Ανάκτηση Πληροφορίας. Συνεπώς είναι απαραίτητη η ύπαρξη μετρικών που θα ελέγχουν την ποιότητα (με βάση τη σχετικότητα) του ανακτόμενου συνόλου δεδομένων. Η χρήση αυτών των μετρικών αποτελεί την καλούμενη *Εκτίμηση Απόδοσης Ανάκτησης*.

Σε αυτό το κεφάλαιο θα συζητήσουμε για τεχνικές *Εκτίμησης Απόδοσης Ανάκτησης* σε συστήματα Ανάκτησης Πληροφορίας. Αυτή η εκτίμηση απόδοσης βασίζεται στην ύπαρξη μίας συλλογής κειμένων αναφοράς (test reference collection) και μίας μετρικής εκτίμησης απόδοσης ανάκτησης. Η συλλογή κειμένων αναφοράς αποτελείται από μία συλλογή κειμένων, ένα σύνολο προτύπων πληροφοριακών αναγκών (ερωτημάτων) και ένα σύνολο σχετικών κειμένων (όπως παρέχεται από ειδικούς) για κάθε πληροφοριακή ανάγκη. Παραδείγματα τέτοιων συλλογών κειμένων αναφοράς είναι οι TIPSTER/TREC, CACM, CISI και Cystic Fibrosis. Δοθείσης μίας νέας στρατηγικής ανάκτησης πληροφορίας *S*, η μετρική απόδοσης καθορίζει (για κάθε πληροφοριακή ανάγκη) την *ομοιότητα* ανάμεσα στο σύνολο των κειμένων που ανακτήθηκαν από την *S* και στο σύνολο των σχετικών κειμένων όπως έχει προκαθοριστεί από τους ειδικούς.

Στη συνέχεια θα καλύψουμε τις δύο πιο βασικές μετρικές εκτίμησης απόδοσης ανάκτησης, την *ακρίβεια* (precision) και την *ανάκληση* (recall) και θα αναφερθούμε σε άλλες εναλλακτικές μετρικές εκτίμησης απόδοσης ανάκτησης όπως η *E μετρική*, ο *αρμονικός μέσος όρος*, κ.λ.π.

2.2 Ανάκληση και Ακρίβεια

Έστω *I* μία πρότυπη πληροφοριακή ανάγκη (σε μία συλλογή κειμένων αναφοράς) και *R* το σύνολο των σχετικών της κειμένων. Έστω επίσης |*R*| ο αριθμός των κειμένων στο σύνολο *R*. Υποθέστε ότι μία δοσμένη στρατηγική ανάκτησης (της οποίας η απόδοση εκτιμάται) επεξεργάζεται την πληροφοριακή ανάγκη *I* και παράγει ένα σύνολο κειμένων απάντησης *A*. Έστω |A| ο αριθμός των κειμένων στο σύνολο A και έστω |*Rα*| ο αριθμός των κειμένων που είναι κοινά στα σύνολα *R* και *A*.

Τότε οι μετρικές *ανάκληση* (*recall*) και *ακρίβεια* (*precision*) ορίζονται ως εξής:

- *Ανάκληση* (*Recall)* είναι το ποσοστό των σχετικών κειμένων (σύνολο R) που έχει ανακτηθεί, δηλαδή,

$$Ανάκληση = \frac{|Ra|}{|R|}.$$

- *Ακρίβεια (Precision)* είναι το ποσοστό των ανακτηθέντων κειμένων (σύνολο A) που είναι σχετικό, δηλαδή,

$$Ακρίβεια = \frac{|Ra|}{|A|}$$

Η ακρίβεια και η ανάκληση, όπως έχουν οριστεί, υποθέτουν ότι όλα τα κείμενα στο σύνολο απάντησης *A* έχουν εξετασθεί από τον χρήστη. Εντούτοις, ο χρήστης συνήθως δεν βλέπει όλα τα κείμενα του συνόλου απάντησης *A* αμέσως, αλλά αντίθετα τα κείμενα του *A* εμφανίζονται σε αυτόν ένα προς ένα διατεταγμένα με βάση το βαθμό σχετικότητας τους με την πληροφοριακή ανάγκη *I* (η διάταξη και ο βαθμός σχετικότητας παράγονται από τον αλγόριθμο ανάκτησης, άρα αποτελούν και αυτά αντικείμενα προς εκτίμηση απόδοσης). Στην περίπτωση αυτή οι μετρικές ανάκλησης και ακρίβειας μεταβάλλονται καθώς ο χρήστης εξετάζει τα διάφορα κείμενα της ανακτώμενης συλλογής (από τα περισσότερο σχετικά προς τα λιγότερο σχετικά). Συνεπώς πλήρης εκτίμηση απόδοσης απαιτεί την σχεδίαση ενός διαγράμματος ακρίβειας/ανάκλησης όπως θα περιγραφεί άμεσα στη συνέχεια.

Όπως προηγουμένως λοιπόν, ας θεωρήσουμε μία συλλογή κειμένων αναφοράς, ένα σύνολο προτύπων πληροφοριακών αναγκών, ένα ερώτημα *q* το οποίο ανήκει στη συλλογή των προτύπων πληροφοριακών αναγκών και έστω R_q το σύνολο των σχετικών κειμένων για το ερώτημα q όπως έχει καθοριστεί από ειδικούς. Για παράδειγμα ας υποθέσουμε ότι το σύνολο *Rq* περιέχει τα ακόλουθα κείμενα $R_q=\{d_1, d_3, d_5, d_7, d_9, d_{13}, d_{21}, d_{41}, d_{43}, d_{45}\}$.

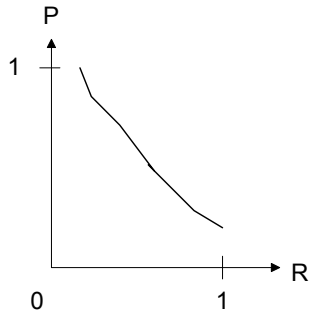
Θεωρείστε ένα νέο αλγόριθμο ανάκτησης που μόλις έχει σχεδιαστεί, και υποθέστε ότι ο αλγόριθμος αυτός επιστρέφει την ακόλουθη συλλογή κειμένων (η διάταξη σχετικότητας που παρέχει ο αλγόριθμος δηλώνεται από τους αριθμούς δίπλα σε κάθε κείμενο ενώ με έντονη σκίαση παρουσιάζονται τα κείμενα που ανήκουν στο σύνολο R_q).

1.	$\boldsymbol{d_7}$	6.	$\boldsymbol{d_5}$	11.	d_4
2.	d_2	7.	d_{28}	12.	d_{40}
3.	$\boldsymbol{d_3}$	8.	d_{12}	13.	d_{10}
4.	d_6	9.	d_{22}	14.	d_{36}
5.	d_8	10.	$\boldsymbol{d_{13}}$	15.	$\boldsymbol{d_1}$

Εάν εξετάσουμε την ανωτέρω διάταξη ξεκινώντας από το κείμενο που βρίσκεται στη θέση 1 παρατηρούμε τα εξής. Αρχικά το κείμενο d_7 που βρίσκεται στην θέση 1 είναι σχετικό, και αντιστοιχεί στο 10% του συνόλου των σχετικών κειμένων (το σύνολο R_q). Συνεπώς λέμε ότι έχουμε ακρίβεια 100% και ανάκληση 10%. Στη συνέχεια το κείμενο $\boldsymbol{d_3}$ που βρίσκεται στη θέση 3 είναι το επόμενο σχετικό κείμενο. Στο σημείο αυτό έχουμε ακρίβεια περίπου 66% (δύο στα τρία κείμενα είναι σχετικά) και ανάκληση 20% (δύο στα δέκα από τα σχετικά κείμενα έχουν ειδωθεί). Συνεχίζοντας με τον τρόπο παίρνουμε ένα σύνολο ζευγαριών (τιμή ακρίβειας, τιμή ανάκλησης) που μπορούμε να παραστήσουμε σε ένα διάγραμμα το καλούμενο *διάγραμμα ακρίβειας/ανάκλησης*. Ένα τέτοιο παράδειγμα διαγράμματος φαίνεται στην Εικόνα 2-1. Συνήθως το διάγραμμα αυτό βασίζεται σε 11 πρότυπα επίπεδα ανάκλησης τα 0%, 10%, ..., 100%, όπου σε κάθε επίπεδο η ακρίβεια υπολογίζεται με χρήση μίας διεργασίας παρεμβολής (interpolation) της ακόλουθης μορφής: έστω r_j, $j \in \{0,1,2,\ldots,10\}$ το *j*-οστό επίπεδο ανάκλησης (π.χ. το r_5 είναι το επίπεδο ανάκλησης 50%), τότε:

$$P(r_j)=max\ _{r_j \le r \le r_{j+1}} P(r)$$

Με άλλα λόγια η παρεμβεβλημένη ακρίβεια στο *j*-οστό πρότυπο επίπεδο ανάκλησης, είναι η μέγιστη γνωστή ακρίβεια σε οποιοδήποτε γνωστό επίπεδο ανάκλησης μεταξύ του *j*-οστού και του *j*+1-οστού επιπέδου.



Εικόνα 2-1: Τυπικό διάγραμμα Ακρίβειας/Ανάκλησης

Στο ανωτέρω παράδειγμα το *διάγραμμα* έχει υπολογιστεί για ένα απλό ερώτημα *q*. Συνήθως όμως η εκτίμηση απόδοσης ανάκτησης συστημάτων ανάκτησης πληροφορίας γίνεται με την πραγματοποίηση διάφορων ερωτήσεων από το σύνολο των προτύπων πληροφοριακών αναγκών την σχεδίαση ατομικών διαγραμμάτων για κάθε ερώτημα και στη συνέχεια τη δημιουργία ενός συνολικού διαγράμματος όπου οι συντεταγμένες κάθε σημείου θα προκύπτουν ως ο μέσος όρος των αντίστοιχων σημείων στα ατομικά διαγράμματα για κάθε ερώτημα.

Τα διαγράμματα *ακρίβειας/ανάκλησης* θεωρούνται ως μία από τις κλασσικές στρατηγικές εκτίμησης της απόδοσης ανάκλησης ενός συστήματος ανάκτησης πληροφορίας και χρησιμοποιούνται εκτεταμένα στην βιβλιογραφία των συστημάτων ανάκτησης. Τα διαγράμματα αυτά είναι χρήσιμα επειδή μας επιτρέπουν να εκτιμήσουμε ποσοτικά τόσο την ποιότητα του ανακτώμενου συνόλου κειμένων όσο και το εύρος του αλγορίθμου ανάκτησης. Επιπλέον είναι απλά στην κατανόηση και μπορούν να συνοψιστούν και εύκολα με τη χρήση μίας απλής αριθμητικής τιμής.

2.3 Σύνοψη Διαγραμμάτων με χρήση μίας αριθμητικής τιμής

Τα διαγράμματα *ακρίβειας/ανάκλησης* είναι χρήσιμα για την σύγκριση της απόδοσης ανάκτησης διακριτών αλγορίθμων ανάκτησης σε ένα σύνολο από πρότυπες πληροφοριακές ανάγκες. Εντούτοις υπάρχουν περιπτώσεις στις οποίες θα θέλαμε να συγκρίνουμε την απόδοση αλγορίθμων ανάκτησης για ατομικές πληροφοριακές ανάγκες. Οι λόγοι που θα θέλαμε να το κάνουμε αυτό είναι δύο: (1) η χρήση μέσων τιμών που προκύπτουν από την εκτέλεση διαφόρων ερωτημάτων μπορεί να αποκρύπτει σημαντικές ανωμαλίες στον αλγόριθμο ανάκτησης που εξετάζεται, (2) όταν συγκρίνουμε δύο αλγορίθμους μπορεί να θέλουμε να μελετήσουμε αν ο ένας είναι καλύτερος από τον άλλο για κάθε μία από τις πρότυπες πληροφοριακές ανάγκες

Στις περιπτώσεις αυτές **μία** μόνο τιμή ακρίβειας υπολογίζεται για κάθε ερώτημα που θα μπορούσε να θεωρηθεί ως σύνοψη του συνολικού διαγράμματος ακρίβειας ανάκλησης. Συνήθως αυτή η τιμή είναι η ακρίβεια σε κάποιο προκαθορισμένο επίπεδο ανάκλησης. Για παράδειγμα, η τιμή αυτή θα μπορούσε να είναι η ακρίβεια που υπάρχει όταν ο χρήστης συναντήσει στη λίστα των ανακτώμενων κειμένων το πρώτο κείμενο που είναι σχετικό. Άλλες δυνατές προσεγγίσεις είναι οι ακόλουθες:

2.3.1 Μέση ακρίβεια για κάθε σχετικό κείμενο που ανακτάται

Η ιδέα εδώ είναι να παραχθεί μία τιμή σύνοψης που υπολογίζεται ως η μέση τιμή της ακρίβειας για τις διάφορες τιμές ακρίβειας που εμφανίζονται όταν κάθε σχετικό κείμενο εμφανίζεται στη διάταξη του αλγορίθμου ανάκτησης.

Αυτή η μετρική ευνοεί συστήματα που ανακτούν τα σχετικά κείμενα, γρήγορα (στη λίστα διάταξης). Φυσικά είναι δυνατόν ένας αλγόριθμος να έχει καλή μέση ακρίβεια αλλά συνολικά να εμφανίζει άσχημη απόδοση σε όρους συνολικής ανάκλησης.

2.3.2 R-Ακρίβεια

Η ιδέα εδώ είναι να παραχθεί μία τιμή σύνοψης που υπολογίζεται ως η ακρίβεια στη R-οστή θέση διάταξης, όπου R είναι ο συνολικός αριθμός των σχετικών κειμένων για την τρέχουσα ερώτηση (δηλαδή ο αριθμός των κειμένων στο σύνολο R_q).

Η μετρική αυτή είναι μία χρήσιμη παράμετρος για την παρατήρηση της συμπεριφοράς του αλγορίθμου για κάθε ατομικό ερώτημα ενός πειράματος. Επιπλέον, είναι δυνατόν ο υπολογισμός ενός διαγράμματος που παρουσιάζει την τιμή της R-ακρίβειας για κάθε ερώτημα.

Εντούτοις η χρήση ενός μόνο αριθμού για να συνοψιστεί η πλήρης συμπεριφορά ενός αλγορίθμου ανάκτησης μπορεί να είναι αρκετά ανακριβής.

2.3.3 Ιστογράμματα Ακρίβειας

Η μετρική R-ακρίβειας για διάφορα πρότυπα ερωτήματα μπορεί να χρησιμοποιηθεί για να συγκρίνει την ποιότητα δύο αλγορίθμων ανάκτησης ως εξής. Έστω $RP_A(i)$ και $RP_B(i)$ οι τιμές της R-ακρίβειας για δύο αλγόριθμους ανάκτησης A,B για το i-οστό ερώτημα. Ορίζουμε την ακόλουθη διαφορά:

$RP_{A/B}(i)=RP_A(i)-RP_B(i)$.

Μια τιμή του $RP_{A/B}(i)$ ίση με το 0 σημαίνει ότι και οι δύο αλγόριθμοι έχουν ισοδύναμη απόδοση (σε όρους της R-ακρίβειας) για το i-οστό ερώτημα. Μία θετική τιμή του $RP_{A/B}$(i) δείχνει ότι ο αλγόριθμος A έχει καλύτερη απόδοση από τον B (για το i-οστό ερώτημα) ενώ μία αρνητική τιμή δείχνει μία καλύτερη απόδοση για τον B.

Οι τιμές $RP_{A/B}(i)$ μπορούν να αναπαρασταθούν σε ένα ειδικό διάγραμμα (όπου ο άξονας x θα παριστάνει τα διάφορα ερωτήματα και ο άξονας y τις διάφορες $RP_{A/B}(i)$ τιμές), το οποίο ονομάζεται ιστόγραμμα ακρίβειας και το οποίο μας επιτρέπει να συγκρίνουμε γρήγορα την ποιότητα ανάκτησης δύο αλγορίθμων ανάκτησης.

2.4 Καταλληλότητα Ακρίβειας και Ανάκλησης

Οι μετρικές ακρίβειας και ανάκλησης έχουν χρησιμοποιηθεί κατά κόρον για την εκτίμηση της απόδοσης ανάκτησης αλγορίθμων ανάκτησης. Εντούτοις η χρήση των δύο μετρικών παρουσιάζει ορισμένα εγγενή προβλήματα τα κυριότερα εκ των οποίων είναι τα ακόλουθα: (1) η κατάλληλη εκτίμηση της μέγιστης ανάκλησης για ένα ερώτημα απαιτεί λεπτομερή γνώση όλων των κειμένων της συλλογής, Σε μεγάλες συλλογές αυτή η γνώση δεν είναι διαθέσιμη κάτι που συνεπάγεται ότι η ανάκληση δεν μπορεί να εκτιμηθεί, (2) η ανάκληση και η ακρίβεια είναι σχετιζόμενες μετρικές που καλύπτουν διαφορετικά θέματα η καθεμία ενός συνόλου ανακτώμενων κειμένων. Σε πολλές περιπτώσεις η χρήση μίας μόνο μετρικής που να μπορεί να συνδυάζει ανάκληση και ακρίβεια μπορεί να θεωρηθεί πιο κατάλληλη, (3) η ανάκληση και η ακρίβεια μετράνε την αποτελεσματικότητα στην επεξεργασία ενός συνόλου ερωτημάτων, τα οποία επεξεργαζόμαστε χωρίς να υπάρχει αλληλεπίδραση με τον χρήστη. Εντούτοις σε μοντέρνα συστήματα η διεπαφή και η αλληλεπίδραση με τον χρήστη αποτελούν σημείο κλειδί στην επεξεργασία ενός ερωτήματος, κάτι που καθιστά επιτακτική την υιοθέτηση μετρικών που καλύπτουν αυτόν τον τρόπο λειτουργίας του συστήματος, (4) οι μετρικές ανάκλησης και ακρίβειας είναι κατάλληλες όταν υπάρχει μία γραμμική διάταξη στα

ανακτώμενα κείμενα. Για συστήματα που δεν το υποστηρίζουν αυτό η ανάκληση και η ακρίβεια μπορεί να είναι ανακριβείς.

2.5 Εναλλακτικές Μετρικές

Εφόσον η ανάκληση και η ακρίβεια, παρά τη συχνή χρήση τους δεν είναι πάντα οι πιο κατάλληλες μετρικές για την εκτίμηση της απόδοσης ανάκτησης, έχουν προταθεί διάφορες εναλλακτικές μετρικές, μία σύνοψη των οποίων παρατίθεται στη συνέχεια.

2.5.1 Αρμονικός Μέσος Όρος

Όπως έχει αναφερθεί πριν, μία μετρική που να μπορεί να συνδυάζει και ανάκληση και ακρίβεια είναι συχνά επιθυμητή. Μία τέτοια μετρική είναι ο αρμονικός μέσος όρος *F* ανάκλησης και ακρίβειας που ορίζεται ως εξής:

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

όπου *r(j)* είναι η ανάκληση για το *j*-οστό κείμενο στη διάταξη, *P(j)* είναι η ακρίβεια για το *j*-οστό κείμενο στη διάταξη και *F(j)* είναι ο αρμονικός μέσος όρος των *r(j), P(j)*. Η συνάρτηση *F* παίρνει τιμές στο διάστημα [0,1], όπου η τιμή 0 σημαίνει ότι κανένα σχετικό κείμενο δεν έχει ανακτηθεί και η τιμή 1 ότι όλα τα κείμενα που έχουν ανακτηθεί είναι σχετικά. Επιπλέον ο αρμονικός μέσος όρος παίρνει μεγάλες τιμές όταν τόσο η ακρίβεια όσο και η ανάκληση έχουν υψηλές τιμές. Συνεπώς ο προσδιορισμός της μέγιστης τιμής για την *F,* μπορεί να μεταφραστεί ως προσπάθεια προσδιορισμού του καλύτερου συνδυασμού των μετρικών ανάκλησης και ακρίβειας.

2.5.2 Η Μετρική E

Μία άλλη μετρική που συνδυάζει ακρίβεια και ανάκληση είναι η *μετρική E*. Το βασικό πλεονέκτημα της μετρικής αυτής είναι ότι επιτρέπει στο χρήστη να προσδιορίσει αν τον ενδιαφέρει περισσότερο η ανάκληση ή η ακρίβεια. Η μετρική *E* ορίζεται ως εξής:

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

Όπου *r(j)* είναι η ανάκληση για το *j*-οστό κείμενο στη διάταξη, *P(j)* είναι η ακρίβεια για το *j*-οστό κείμενο στη διάταξη, *E(j)* είναι η μετρική *E* των *r(j), P(j)* και *b* είναι μία παράμετρος που προσδιορίζεται από τον χρήστη και αντανακλά τη σχετική σημαντικότητα της ακρίβειας και της ανάκλησης. Όταν *b*=1 η μετρική είναι το συμπλήρωμα του αρμονικού μέσου όρου, τιμές του *b* μεγαλύτερες από 1 υποδηλώνουν ότι ο χρήστης ενδιαφέρεται περισσότερο για την ακρίβεια παρά για την ανάκληση, ενώ τιμές του *b* μικρότερες από 1 υποδηλώνουν ότι ο χρήστης ενδιαφέρεται περισσότερο για την ανάκληση παρά για την ακρίβεια.

2.5.3 Μετρικές προσανατολισμένες προς τον χρήστη

Οι μετρικές ακρίβειας και ανάκλησης βασίζονται στην υπόθεση ότι το σύνολο των σχετικών κειμένων είναι το ίδιο ανεξάρτητα από τον χρήστη. Εντούτοις διαφορετικοί χρήστες

μπορεί να βλέπουν με διαφορετικό τρόπο ποια κείμενα είναι σχετικά και ποια όχι. Για την αντιμετώπιση του προβλήματος αυτού έχουν προταθεί μετρικές προσανατολισμένες προς τον χρήστη όπως *βαθμός κάλυψης* (*coverage ratio*), *βαθμός καινοτομίας* (*novelty ratio*), *σχετική ανάκληση* (*relative recall*) και *κόστος ανάκλησης* (*recall effort*).

Όπως και στην προηγούμενη ενότητα θεωρούμε ότι έχουμε μία συλλογή κειμένων αναφοράς, μία πρότυπη πληροφοριακή ανάγκη I και μία στρατηγική ανάκτησης που θα πρέπει να εκτιμηθεί. Έστω R το σύνολο των σχετικών κειμένων για την πληροφοριακή ανάγκη I, A το σύνολο των κειμένων που έχει ανακτηθεί και $U \subset R$ το σύνολο των κειμένων που είναι γνωστό στο χρήστη ότι είναι σχετικά προς το ερώτημα του. Ο αριθμός των κειμένων στο σύνολο U συμβολίζεται με $|U|$. Η τομή των συνόλων A και U μας παρέχει το σύνολο των κειμένων που είναι γνωστό στο χρήστη ότι είναι σχετικά προς το ερώτημα του και που έχουν ανακτηθεί. Έστω $|R_k|$ ο αριθμός των κειμένων στο σύνολο αυτό. Επιπλέον έστω $|R_u|$ ο αριθμός των σχετικών κειμένων, που δεν γνώριζε πριν ο χρήστης και τα οποία έχουν ανακτηθεί.

Ορίζουμε ως *βαθμό κάλυψης* (*coverage ratio*) το ποσοστό των γνωστών (ως προς την σχετικότητα) κειμένων του χρήστη που έχουν ανακτηθεί, δηλαδή:

$$\textit{Βαθμός κάλυψης} = \frac{|R_k|}{|U|}.$$

Ορίζουμε ως *βαθμό καινοτομίας* (*novelty ratio*) το ποσοστό των σχετικών κειμένων που έχουν ανακτηθεί και που ήταν πριν άγνωστα στον χρήστη, δηλαδή:

$$\textit{Βαθμός καινοτομίας} = \frac{|R_u|}{|R_u| + |R_k|}$$

Υψηλός βαθμός κάλυψης σημαίνει ότι το σύστημα μπορεί να εντοπίσει τα περισσότερα από τα σχετικά κείμενα που ο χρήστης περιμένει να ανακτήσει, ενώ υψηλός βαθμός καινοτομίας σημαίνει ότι το σύστημα αποκαλύπτει στον χρήστη πολλά νέα σχετικά κείμενα που πριν ήταν άγνωστα.

Οι δύο άλλες μετρικές μπορούν να οριστούν ως εξής: η *σχετική ανάκληση* (*relative recall*) ορίζεται ως το πηλίκο ανάμεσα στον αριθμό των σχετικών κειμένων που έχουν ανακτηθεί και των σχετικών κειμένων που ο χρήστης περιμένει να ανακτηθούν. Όταν ο χρήστης εντοπίσει τα κείμενα που αναμένει τότε σταματά το ψάξιμο και η *σχετική ανάκληση* γίνεται ίση με τη μονάδα. Τέλος ως *κόστος ανάκλησης (recall effort)* ορίζουμε το πηλίκο ανάμεσα στα σχετικά κείμενα που ο χρήστης αναμένει να εντοπίσει και τα κείμενα που εξετάζει μέχρις ότου εντοπίσει αυτά που αναμένει.

Κεφάλαιο 3. Μοντελοποίηση

3.1 Εισαγωγή

Όπως είδαμε και στο Κεφάλαιο 1, η πιο συνηθισμένη πρακτική για την δεικτοδότηση και την ανάκτηση κειμένων είναι η χρήση των *όρων δεικτοδότησης* (index terms). Ένας όρος δεικτοδότησης είναι μια λέξη κλειδί ή μια ομάδα εννοιολογικά συσχετιζόμενων λέξεων, η εμφάνιση των οποίων λαμβάνει από μόνη της μια αυτόνομη έννοια (π.χ. computer algorithm). Κατά μια πιο απλοποιημένη εκδοχή, ένας όρος δεικτοδότησης είναι απλά μια λέξη που εμφανίζεται σε ένα κείμενο της συλλογής. Η ανάκτηση που βασίζεται στο ταίριασμα όρων δεικτοδότησης ερωτήματος και κειμένων της συλλογής, είναι πολύ απλή αλλά εισάγει ένα σύνολο προβληματισμών για την αποτελεσματικότητά της. Για παράδειγμα, η βασική υπόθεση

που εισάγει η παραπάνω στρατηγική, είναι ότι η σημασιολογία τόσο των κειμένων όσο και της πληροφοριακής ανάγκης του χρήστη, μπορεί να εκφραστεί με φυσικό τρόπο, μέσα από ένα σύνολο λέξεων. Στην πράξη ένα σημαντικό κομμάτι από τη σημασιολογία του κειμένου χάνεται κατά τη μεταφορά στο χώρο του ευρετηρίου.

Ο λόγος γι' αυτήν την απώλεια είναι ότι οι λέξεις αποκτούν την ερμηνεία τους ανάλογα με το πλαίσιο συμφραζομένων στο οποίο εμφανίζονται. Από αυτή την παρατήρηση πηγάζουν δυο φαινόμενα, η *πολυσημία* και η *συνωνυμία*. Στην πολυσημία, έχουμε το φαινόμενο ο ίδιος όρος να λαμβάνει διαφορετικές έννοιες ανάλογα με τα συμφραζόμενα που συνοδεύουν την εμφάνισή του. Για παράδειγμα ο όρος 'spider' μπορεί να χρησιμοποιείται για να δηλώσει ένα 'web spider' αν το κείμενο μιλάει για το Διαδίκτυο ή το έντομο σε άλλες περιπτώσεις. Στην συνωνυμία, διαφορετικοί όροι μπορούν να περιγράφουν την ίδια έννοια γιατί εμφανίζονται στα ίδια πλαίσια συμφραζομένων. Για παράδειγμα η έννοια 'αυτοκίνητο', μπορεί να περιγράφεται ισοδύναμα από τις λέξεις: 'αυτοκίνητο', 'αμάξι', 'τετράτροχο', 'όχημα'. Η συνωνυμία και η πολυσημία, αποτελούν κλασσικά προβλήματα που συνδέονται με τον τρόπο λογικής αναπαράστασης των κειμένων μέσω ευρετηρίου.

Έχοντας υπόψη μας τα παραπάνω προβλήματα και με δεδομένο ότι η διαδικασία της αντιστοίχισης του ερωτήματος στη συλλογή των κειμένων, γίνεται στο χώρο του ευρετηρίου, μπορούμε να κατανοήσουμε γιατί συχνά τα αποτελέσματα μιας ερώτησης διατυπωμένης με λέξεις-κλειδιά δεν είναι τα αναμενόμενα. Αν μάλιστα λάβουμε υπόψη μας και το γεγονός ότι πολλοί χρήστες δεν είναι σε θέση να επιλέξουν τις κατάλληλες λέξεις-κλειδιά για τον σχηματισμό ερωτήσεων, το πρόβλημα μεγαλώνει. Ένα καλό παράδειγμα του παραπάνω προβλήματος είναι τα απογοητευτικά αποτελέσματα σε πολλά από τα ερωτήματα που υποβάλλονται σε μια Μηχανή Αναζήτησης στο Παγκόσμιο Ιστό (όπου και μεγάλο μέρος των χρηστών είναι χωρίς μεγάλη εμπειρία στο σχηματισμό ερωτήσεων). Η πρόκληση για ένα μοντέλο για ΑΠ, είναι να δημιουργήσει το υπόβαθρο, ώστε να υπάρξει ταίριασμα της πληροφοριακής ανάγκης χρήστη με τα κείμενα της συλλογής, παρά την ανακριβή αναπαράσταση και με όσο το δυνατόν μικρότερες αποκλίσεις.

Στο πνεύμα της ανάκτησης πληροφορίας, ταίριασμα σημαίνει εκτίμηση από το σύστημα, της σχετικότητας των κειμένων ως προς το δοθέν ερώτημα. Μια τέτοια εκτίμηση επιτυγχάνεται με την χρήση ενός αλγορίθμου κατάταξης (ranking), με βάση τον οποίο, γίνεται μια απλή διάταξη των κειμένων. Τα κείμενα που εμφανίζονται στις πρώτες θέσεις αυτής της διάταξης, θεωρούνται ως το πιο πιθανό να είναι σχετικά με την ερώτηση, με την πιθανότητα να φθίνει, όσο εξετάζουμε τη διάταξη προς τις χαμηλότερες θέσεις. Οι αλγόριθμοι κατάταξης έχουν ζωτική σημασία σε ένα σύστημα ΑΠ. Συνεπώς μία βασική λειτουργία του μοντέλου είναι να παρέχει έναν αλγόριθμο κατάταξης για κάθε ερώτημα που υποβάλλεται.

Ο τρόπος θεώρησης της λογικής αναπαράστασης των κειμένων και η συσχέτισή του με τον αλγόριθμο κατάταξης, είναι το βασικό χαρακτηριστικό που διαφοροποιεί τα μοντέλα ΑΠ. Στο κεφάλαιο αυτό εξετάζουμε μια κατηγοριοποίηση των μοντέλων, κάποιους τυπικούς ορισμούς και τέλος παρουσιάζουμε τα κυριότερα μοντέλα ΑΠ.

3.2 Ταξινόμηση των Μοντέλων για Ανάκτηση Πληροφορίας

Τα τρία κλασσικά μοντέλα στην Ανάκτηση Πληροφορίας είναι το *Boolean Μοντέλο (Μοντέλο Δυαδικής Λογικής)*, το *Vector Space Μοντέλο (Μοντέλο Διανυσματικού Χώρου)* και το *Πιθανοτικό Μοντέλο*. Στο μοντέλο Boolean, τόσο τα κείμενα όσο και τα ερωτήματα αντιμετωπίζονται ως ένα *σύνολο* από όρους δεικτοδότησης. Κατά συνέπεια το μοντέλο μπορεί να θεωρηθεί ως *συνολοθεωρητικό*. Στο Vector Space μοντέλο, τα κείμενα και τα ερωτήματα

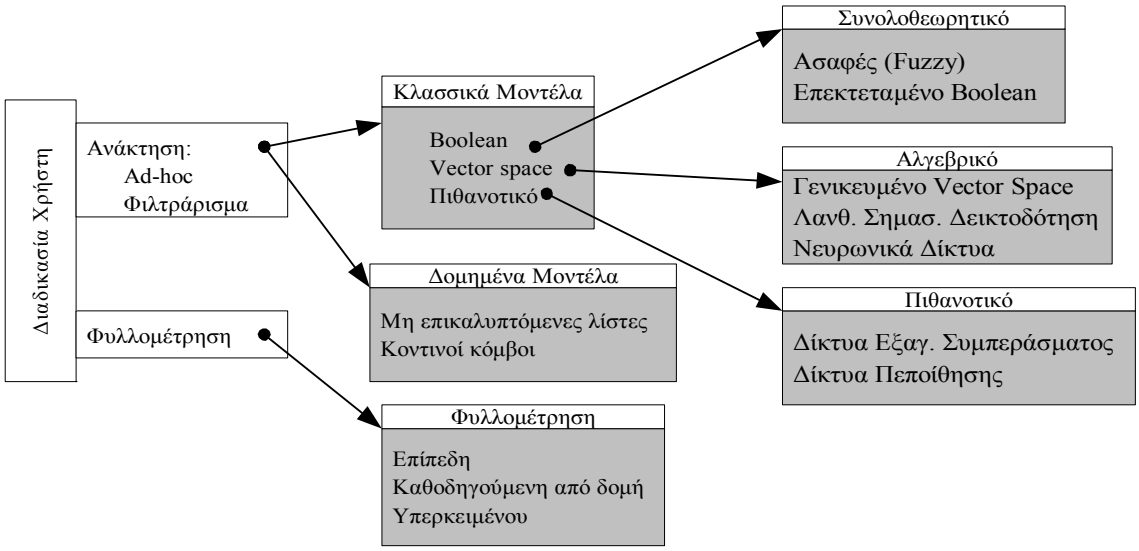
αναπαρίστανται ως διανύσματα σε έναν *t*-διάστατο[2] χώρο. Έτσι λέμε ότι το μοντέλο είναι *αλγεβρικό*. Το Πιθανοτικό μοντέλο εισάγει έναν τρόπο αναπαράστασης, ο οποίος βασίζεται στην πιθανοθεωρία. Κατά συνέπεια το μοντέλο είναι πιθανοτικού χαρακτήρα.

Με τον καιρό προτάθηκαν διάφορες νέες προσεγγίσεις σε καθεμία από τις κατηγορίες βασικών μοντέλων. Έτσι έχουμε στο συνολοθεωρητικό πεδίο τα μοντέλα: *ασαφές* (fuzzy) Boolean και *επεκτεταμένο* Boolean. Στα αλγεβρικά μοντέλα έχουμε το *γενικευμένο* Vector Space, την *λανθάνουσα σημασιολογική δεικτοδότηση* (latent semantic indexing, LSI) και το μοντέλο των *νευρωνικών δικτύων*. Στον πιθανοτικό τομέα εμφανίστηκαν τα *δίκτυα εξαγωγής συμπεράσματος* (inference networks) και τα *δίκτυα πεποίθησης* (belief networks). Η Εικόνα 3-1 δίνει σχηματικά την κατηγοριοποίηση αυτή.

Εκτός από την χρήση του περιεχομένου των κειμένων, ορισμένα μοντέλα εκμεταλλεύονται και την εσωτερική δομή που φυσιολογικά υπάρχει στο γραπτό λόγο. Σε αυτή την περίπτωση λέμε ότι έχουμε ένα δομημένο μοντέλο. Για τη δομημένη ανάκτηση κειμένου, συναντούμε δύο μοντέλα, τις *μη επικαλυπτόμενες λίστες* (non-overlapping lists) και τους *κοντινούς κόμβους* (proximal nodes).

Όπως είδαμε στο Κεφάλαιο 1, η διαδικασία του χρήστη μπορεί εκτός από αναζήτηση να έχει μορφή *φυλλομέτρησης*. Σε αυτή την κατηγορία εντοπίζουμε τρία μοντέλα για φυλλομέτρηση: *επίπεδη* (flat), *καθοδηγούμενη από τη δομή* (structure guided), *φυλλομέτρηση υπερκειμένου* (hypertext browsing).

Στο κεφάλαιο αυτό αναπτύσσουμε μόνο τα συνολοθεωρητικά και αλγεβρικά μοντέλα (εκτός των νευρωνικών δικτύων) και το βασικό πιθανοτικό μοντέλο. Τα υπόλοιπα μοντέλα τα αναφέρουμε απλώς για πληρότητα και ο ενδιαφερόμενος θα πρέπει να ανατρέξει στο [BR99] για μια πλήρη επισκόπηση όλων των μοντέλων. Η λανθάνουσα σημασιολογική δεικτοδότηση έχει προσεγγίσει σημαντικό ερευνητικό ενδιαφέρον τα τελευταία χρόνια και γι' αυτό το λόγο εξετάζεται ξεχωριστά και με πληρότητα στο Κεφάλαιο 4.



Εικόνα 3-1: Ταξινόμηση των μοντέλων Ανάκτησης Πληροφορίας

[2] Η σημασία του *t* θα αναλυθεί αργότερα σ' αυτό το κεφάλαιο

3.3 Ad hoc ανάκτηση και φιλτράρισμα

Στα περισσότερα συστήματα ΑΠ, η συλλογή των κειμένων παραμένει σχεδόν στατική (ακόμα και αν αλλάζει π.χ. μια φορά τη μέρα θεωρείται στατική), ενώ συνέχεια υποβάλλονται καινούρια ερωτήματα. Αυτός ο τρόπος λειτουργίας έχει ονομαστεί *ad hoc* ανάκτηση πληροφορίας και είναι ο πιο κοινή μορφή διαδικασίας χρήστη. Μια δεύτερη παρόμοιας μορφής διαδικασία χρήστη είναι το *φιλτράρισμα πληροφορίας* (information filtering). Σ' αυτή τη διαδικασία, τα ερωτήματα παραμένουν σχεδόν σταθερά, ενώ η συλλογή των κειμένων μεταβάλλεται με καινούρια κείμενα να φτάνουν συνεχώς στο σύστημα. Παράδειγμα της τελευταίας διαδικασίας είναι η λίστα αλληλογραφίας ή μία υπηρεσία πληροφόρησης για το χρηματιστήριο.

Στο φιλτράρισμα κατασκευάζεται ένα *προφίλ χρήστη*, το οποίο περιγράφει τις προτιμήσεις του. Το προφίλ συγκρίνεται με κάθε εισερχόμενο κείμενο για να αποφασίσει το σύστημα αν είναι σχετικό ή όχι το κείμενο με τις προτιμήσεις του χρήστη. Με άλλα λόγια το προφίλ είναι μια εναλλακτική μορφή ερωτήματος προς το σύστημα. Μια πιθανή εφαρμογή είναι το φιλτράρισμα ειδήσεων που φθάνουν κατά δεκάδες, on-line, έτσι ώστε να παρέχονται στο χρήστη αυτές που πιθανόν τον ενδιαφέρουν.

Συνήθως η διαδικασία φιλτραρίσματος, παρέχει τα κείμενα που είναι πιθανόν να ενδιαφέρουν το χρήστη χωρίς καμία κατάταξη σχετικότητας. Ο χρήστης καλείται από την παρεχόμενη ακολουθία κειμένων να επιλέξει αυτά που πραγματικά τον αφορούν και να αγνοήσει τα υπόλοιπα. Μερικές φορές παρουσιάζονται στοιχεία κατάταξης στο χρήστη, με τη λογική ότι αυτός θα εξετάσει τα κείμενα με την υψηλότερη κατάταξη, εξετάζοντας έτσι έναν ακόμα μικρότερο αριθμό κειμένων. Αυτή η διαδικασία λέγεται *δρομολόγηση* (routing).

Παρότι ο χρήστης μπορεί να μην έχει εικόνα για την κατάταξη σχετικότητας, μια τέτοια κατάταξη υπολογίζεται εσωτερικά στο σύστημα. Σκοπός του υπολογισμού αυτού είναι να γίνει διαχωρισμός των κειμένων σε σχετικά και μη σχετικά, ως προς το προφίλ. Ο διαχωρισμός γίνεται με τον υπολογισμό της σχετικότητας από τον αλγόριθμο που παρέχει το μοντέλο που χρησιμοποιείται (βλ. Ενότητα 3.1) και τη σύγκριση με κάποιο προκαθορισμένο κατώφλι. Τα κείμενα που κατατάσσονται πάνω από αυτό το κατώφλι εκτιμώνται ως σχετικά. Για τη διαδικασία κατάταξης μπορεί να χρησιμοποιηθεί οποιοδήποτε μοντέλο ΑΠ, συνήθως όμως για λόγους απλότητας χρησιμοποιείται το μοντέλο Vector Space.

Το κύριο ζήτημα όμως στη διαδικασία του φιλτραρίσματος δεν είναι το πώς γίνεται η κατάταξη αλλά με ποιον τρόπο μπορεί να κατασκευαστεί το προφίλ χρήστη. Η συνήθης προσέγγιση είναι το προφίλ να αποτελείται από ένα σύνολο από λέξεις-κλειδιά, τα οποία παρέχει ο χρήστης. Το βάρος εδώ πέφτει στη μεριά του χρήστη, ο οποίος θεωρείται ότι έχει τη δυνατότητα να εκφράσει ικανοποιητικά το ενδιαφέρον του με τη βοήθεια ενός συνόλου λέξεων. Αυτή η προσέγγιση είναι πιο απλή αλλά μπορεί να δυσκολεύει χρήστες που δεν είναι ιδιαίτερα εξοικειωμένοι με το σύστημα.

Μια πιο καλή στρατηγική είναι κατασκευάζεται ένα αρχικό προφίλ από τον ίδιο το χρήστη με τη χρήση λέξεων κλειδιών. Στη συνέχεια ο χρήστης καλείται να αξιολογήσει τη σχετικότητα των κειμένων που του παρουσιάζονται από το σύστημα ως πιθανόν ενδιαφέροντα. Τα κείμενα που αξιολογήθηκαν ως σχετικά αλλά και τα μη σχετικά, χρησιμοποιούνται από το σύστημα για να επαναδιατυπώσουν το προφίλ του χρήστη. Ο χρήστης άρα, εισάγεται σε έναν κύκλο ανάδρασης, ο οποίος δημιουργεί ένα συνεχώς μεταβαλλόμενο προφίλ χρήστη. Πιθανότατα η παραπάνω διαδικασία θα συγκλίνει σε ένα σχεδόν αμετάβλητο προφίλ, από τη στιγμή και μετά που η πληροφορία που έρχεται από την ανάδραση χρήστη, έχει ήδη χρησιμοποιηθεί σε κάποια προηγούμενη στιγμή.

Ανακεφαλαιώνοντας, το φιλτράρισμα μπορεί να θεωρηθεί ως μια διαδικασία ΑΠ, όπου τα ερωτήματα (προφίλ χρήστη) παραμένουν σταθερά αλλά τα κείμενα αλλάζουν συνεχώς. Το φιλτράρισμα όπως και η αναζήτηση πληροφορίας είναι δύο διαφορετικές μορφές διαδικασίας

σε επίπεδο χρήστη. Συνεπώς μπορούμε να χρησιμοποιήσουμε για το φιλτράρισμα οποιοδήποτε μοντέλο χρησιμοποιείται για ανάκτηση πληροφορίας. Η δυσκολία συνήθως στο φιλτράρισμα είναι στον καθορισμό του προφίλ. Μια προσέγγιση απαιτεί τον καθορισμό του προφίλ από τον ίδιο το χρήστη με την διατύπωση κάποιων λέξεων κλειδιών. Η πιο συνηθισμένη πρακτική είναι όμως, η συλλογή ενός αρχικού συνόλου από το χρήστη και η αξιοποίηση των προτιμήσεών του για τη δυναμική ενημέρωση του προφίλ του.

3.4 Τυπικός ορισμός των μοντέλων ΑΠ

Πριν προχωρήσουμε στην εξέταση των επί μέρους μοντέλων θα δώσουμε έναν τυπικό και ακριβή ορισμό για το τι είναι ένα μοντέλο ΑΠ.

Ορισμός *Ένα μοντέλο ανάκτησης πληροφορίας είναι μία τετράδα* [**D, Q,** *F, R*(q_i, d_j)] *όπου:*

1) **D** *είναι ένα σύνολο από λογικές αναπαραστάσεις για τα κείμενα της συλλογής*
2) **Q** *είναι ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές ανάγκες του χρήστη. Αυτές οι αναπαραστάσεις καλούνται ερωτήματα*
3) **F** *είναι ένα υπόβαθρο για την μοντελοποίηση της αναπαράστασης των κειμένων, των ερωτημάτων και των σχέσεων μεταξύ τους*
4) **R**(q_i, d_j) είναι μια συνάρτηση κατάταξης, η οποία συνδέει έναν πραγματικό αριθμό με ένα ερώτημα $q_i \in$ **Q** και μια αναπαράσταση κειμένου $d_j \in$ **D**. Μια τέτοια κατάταξη ορίζει μια διάταξη πάνω στα κείμενα πάντα με βάση το ερώτημα. q_i.

Διαισθητικά ο παραπάνω ορισμός περιγράφει τη διαδικασία καθορισμού ενός μοντέλου ΑΠ. Η διαδικασία ορισμού ενός μοντέλου είναι η ακόλουθη. Αρχικά επινοείται ένας τρόπος αναπαράστασης για τα κείμενα και την πληροφοριακή ανάγκη του χρήστη. Έπειτα καθορίζεται ένα υπόβαθρο στο οποίο θα μπορούν αυτές οι αναπαραστάσεις να μοντελοποιηθούν. Το υπόβαθρο αυτό, θα πρέπει να παρέχει και τον μηχανισμό κατάταξης. Για παράδειγμα στο Boolean μοντέλο, το υπόβαθρο αυτό αποτελείται από τις αναπαραστάσεις των κειμένων και των ερωτήσεων ως σύνολα, και τις κλασσικές πράξεις πάνω στα σύνολα. Αντίστοιχα στο μοντέλο διανυσματικού χώρου, το υπόβαθρο αποτελείται από τις διανυσματικές αναπαραστάσεις κειμένων στον *t*-διάστατο διανυσματικό χώρο και τις επιτρεπτές αλγεβρικές πράξεις πάνω σε διανύσματα.

3.5 Κλασσικά μοντέλα ΑΠ

Σ' αυτή την ενότητα παρουσιάζουμε εν συντομία τα μοντέλα Boolean, το μοντέλο Vector Space καθώς και το πιθανοτικό.

3.5.1 Βασικές υποθέσεις

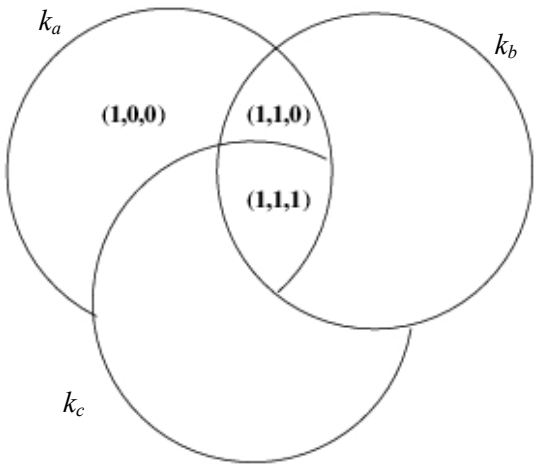
Τα κλασσικά μοντέλα στην ανάκτηση πληροφορίας θεωρούν ότι κάθε κείμενο περιγράφεται από ένα σύνολο από αντιπροσωπευτικές λέξεις κλειδιά, που ονομάζονται όροι δεικτοδότησης. Ένας *όρος δεικτοδότησης* (index term), είναι μια λέξη το σημασιολογικό περιεχόμενο της οποίας, περικλείει ένα μέρος του θέματος με το οποίο ασχολείται το κείμενο. Έτσι τα κείμενα μπορούν να αναπαρασταθούν ως σύνολα όρων, που συνοψίζουν το περιεχόμενό τους. Γενικά οι όροι δεικτοδότησης είναι συνήθως ουσιαστικά γιατί τα ουσιαστικά αναπαριστούν μια έννοια χωρίς την ανάγκη να εμφανίζονται δίπλα σε άλλο μέρος του λόγου και η σημασιολογία τους είναι εύκολα αντιληπτή. Σύνδεσμοι και επιρρήματα, θεωρούνται ότι έχουν κυρίως συμπληρωματικό χαρακτήρα. Συχνά όμως χρειάζεται να χρησιμοποιούμε και αυτά τα μέρη του λόγου στο ευρετήριο, όπως για παράδειγμα στις Μηχανές Αναζήτησης.

Με τη δεδομένη αναπαράσταση των κειμένων ως συλλογές όρων, μπορεί κάποιος να παρατηρήσει ότι δεν έχουν όλοι οι όροι την ίδια ισχύ ως προς την περιγραφή ενός κειμένου. Με άλλα λόγια η ερμηνεία ενός όρου συχνά μπορεί να δίνει μια γενικευμένη ή και ασαφή περιγραφή ενός κειμένου. Τέτοιοι όροι είναι αυτοί που εμφανίζονται με μεγάλη συχνότητα στην πλειονότητα των κειμένων μιας συλλογής. Έστω για παράδειγμα μία συλλογή κειμένων γύρω από υπολογιστές. Η λέξη 'computer' σε μια τέτοια συλλογή εμφανίζεται με μεγάλη βεβαιότητα σχεδόν σε κάθε κείμενο και αν και περιγράφει κάτι αρκετά συγκεκριμένο, δεν είναι αντιπροσωπευτική του κειμένου στο οποίο εμφανίζεται. Αντίθετα αν μια λέξη εμφανίζεται σε μικρό εύρος κειμένων, τότε είναι σχεδόν σίγουρο ότι έχει μεγαλύτερη *βαρύτητα* στην περιγραφή ενός κειμένου. Στο προηγούμενο παράδειγμα η λέξη 'inheritance', εμφανίζεται σίγουρα σε πολύ λιγότερα κείμενα απ' ότι η λέξη 'computer'. Η εμφάνισή της ως όρος δεικτοδότησης για κάποιο κείμενο, μας καθοδηγεί αμέσως ότι το συγκεκριμένο κείμενο συζητά για κληρονομικότητα σε αντικειμενοστραφή προγραμματισμό. Για να προσομοιώσουμε το γεγονός ότι διαφορετικοί όροι μπορεί να έχουν διαφορετική βαρύτητα ως προς στην δεικτοδότηση των κειμένων, σε κάθε όρο δεικτοδότησης αναθέτουμε και ένα αριθμητικό *βάρος*.

Συγκεκριμένα έστω k_i ένας όρος δεικτοδότησης, και d_j ένα κείμενο. Ο αριθμός $w_{i,j} \geq 0$ είναι το *βάρος*, που αντιστοιχεί στο ζεύγος (k_i, d_j) και αντιστοιχεί στο πόσο αντιπροσωπευτικός είναι ο k_i για το κείμενο d_j.

Ορισμός *Έστω t ο αριθμός των όρων δεικτοδότησης στο σύστημα και k_i ένας γενικός όρος δεικτοδότησης. Το σύνολο $K = \{k_1, k_2, \ldots, k_t\}$ περιέχει όλους τους όρους δεικτοδότησης. Ένα βάρος $w_{i,j} > 0$ συνδέεται με κάθε όρο k_i, που εμφανίζεται στο κείμενο d_j. Για κάποιον όρο δεικτοδότησης που δεν εμφανίζεται στο κείμενο, $w_{i,j} = 0$. Κάθε κείμενο d_j έχει ένα αντιπροσωπευτικό διάνυσμα $\vec{d}_j$, το οποίο αναπαρίσταται ως $\vec{d}_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$. Επιπλέον έστω g_i, μια συνάρτηση που επιστρέφει το βάρος που συνδέεται με τον όρο, σε κάθε t-διάστατο διάνυσμα* (δηλ. $g_i(\vec{d}_j) = w_{i,j}$).

Τα παραπάνω βάρη υποθέτουμε ότι είναι μεταξύ τους ανεξάρτητα. Αυτό σημαίνει ότι, για παράδειγμα, η τιμή του $w_{i,j}$ δεν επηρεάζει την τιμή του $w_{i+1,j}$. Αυτή η υπόθεση είναι υπερπλουστευτική δεδομένου ότι συχνά έχουμε συμπλέγματα όρων που εμφανίζονται μαζί.



Εικόνα 3-2: Οι συζευκτικές συνιστώσες του ερωτήματος $[q = k_a \wedge (k_b \vee \neg k_c)]$

Ένα τέτοιο παράδειγμα είναι οι όροι *δίκτυο* και *υπολογιστής*. Σε μια συλλογή με θέμα τα δίκτυα υπολογιστών, αναμένεται αυτοί οι δύο όροι να έχουν παρόμοιες συχνότητες εμφάνισης. Κατά συνέπεια οι δύο αυτοί όροι είναι συσχετισμένοι μεταξύ τους και ο υπολογισμός της ανάθεσης

βαρών θα πρέπει να λαμβάνει υπόψη του αυτή τη συσχέτιση. Λαμβάνοντας υπόψη μας τις συσχετίσεις των όρων μεταξύ τους, η πολυπλοκότητα υπολογισμού των βαρών αυξάνει, συμπαρασύροντας και τον υπολογισμό της κατάταξης. Γι' αυτό στο εξής θα θεωρούμε ότι οι διακριτοί όροι δεικτοδότησης είναι μεταξύ τους ανεξάρτητοι.

3.5.2 Το Boolean μοντέλο

Το Boolean μοντέλο, είναι ένα απλό μοντέλο ανάκτησης πληροφορίας που βασίζεται στη θεωρία συνόλων και στην άλγεβρα Boole. Το υπόβαθρό του είναι εύληπτο και ταυτόχρονα κομψό και καλά ορισμένο στη βάση της άλγεβρας συνόλων. Τα ερωτήματα μπορούν να αναπαρασταθούν με σαφή τρόπο, με χρήση άλγεβρας Boole.

Συγκεκριμένα στο Boolean μοντέλο, κάθε όρος δεικτοδότησης θεωρείται ότι ανήκει εξ' ολοκλήρου ή δεν ανήκει σε ένα κείμενο. Κατά συνέπεια τα βάρη θεωρούνται δυαδικά, δηλ. $w_{i,j} \in \{0,1\}$. Το κάθε ερώτημα θεωρείται ότι αποτελείται από όρους δεικτοδότησης οι οποίοι συνδέονται με έναν από τους τελεστές *and, or, not.* Δηλαδή κάθε ερώτημα είναι μια Boolean έκφραση που μπορεί να γραφεί σε διαζευκτική κανονική μορφή (Disjunctive Normal Form, DNF). Για παράδειγμα το ερώτημα $[q = k_a \wedge (k_b \vee \neg k_c)]$ μπορεί να γραφεί σε DNF ως $[q_{dnf} = (k_a \wedge k_b) \vee (k_a \wedge \neg k_c)]$. Έστω τώρα ένα διάνυσμα με δυαδικά βάρη που αντιστοιχεί σε ανάθεση αλήθειας (truth assignment) σε συζευκτικές εκφράσεις της τριάδας (k_a, k_b, k_c). Για παράδειγμα στην έκφραση $k_a \wedge k_b$ μια ανάθεση αλήθειας είναι η (1,1,0). Άρα το αρχικό ερώτημα μπορεί να αναλυθεί σε διάζευξη τέτοιων διανυσμάτων ως εξής, $[\vec{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)]$. Ο λόγος για την εισαγωγή των δυαδικών αυτών διανυσμάτων είναι γιατί υπάρχει απευθείας αντιστοιχία του ερωτήματος $\vec{q}_{dnf}$ στο διάγραμμα που φαίνεται στην Εικόνα 3-2

Ορισμός *Στο Boolean μοντέλο τα βάρη που ανατίθενται στους όρους δεικτοδότησης είναι δυαδικά δηλαδή, $w_{i,j} \in \{0,1\}$. Ένα ερώτημα q είναι μια συνήθης Boolean έκφραση. Έστω $\vec{q}_{dnf}$ η διαζευκτική κανονική μορφή του ερωτήματος και $\vec{q}_{cc}$ καθεμία από τις συζευκτικές συνιστώσες (conjunctive components) του $\vec{q}_{dnf}$ (τα δυαδικά διανύσματα που προαναφέραμε). Η ομοιότητα του κειμένου d_j προς το ερώτημα q ορίζεται ως εξής:*

$$sim(d_j, q) = \begin{cases} 1, \text{ αν } \exists \vec{q}_{cc} \text{ τέτοιο ώστε } (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall k_i, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0, \text{ διαφορετικά} \end{cases}$$

Αν $sim(d_j, q) = 1$, τότε το Boolean μοντέλο προβλέπει ότι το κείμενο d_j είναι σχετικό με το ερώτημα q (μπορεί και να μην είναι). Διαφορετικά η πρόβλεψη είναι ότι το κείμενο είναι άσχετο.

Το Boolean μοντέλο προβλέπει ότι κάθε κείμενο είτε είναι σχετικό είτε όχι, και δεν υπάρχει η έννοια της *μερικής ικανοποίησης* των συνθηκών του ερωτήματος. Για παράδειγμα έστω d_j τέτοιο ώστε να είναι $\vec{d}_j = (0,1,0)$. Το κείμενο d_j περιέχει τον όρο k_b αλλά θεωρείται άσχετο ως προς το ερώτημα $[q = k_a \wedge (k_b \vee \neg k_c)]$. Λόγω αυτής της έλλειψης το Boolean μοντέλο, στην ουσία εκτελεί περισσότερο *ανάκτηση δεδομένων* (data retrieval) παρά πληροφορίας.

Τα κύρια πλεονεκτήματα του Boolean μοντέλου είναι ο φορμαλισμός του και η απλότητά του. Το κύριο μειονέκτημά του είναι ότι δεν υπάρχει διαβάθμιση σχετικότητας ως προς το ερώτημα κάτι που μπορεί να οδηγήσει σε χαμηλής ποιότητας ανάκτηση πληροφορίας. Ένα δεύτερο μειονέκτημά του είναι ότι συχνά δεν είναι εύκολη η έκφραση της πληροφοριακής ανάγκης του χρήστη με τον φορμαλισμό που επιβάλλει το Boolean μοντέλο (με Boolean

άλγεβρα). Η πληροφοριακή ανάγκη μπορεί να έχει τόσο συγκεκριμένη μορφή, όταν για παράδειγμα ψάχνουμε σε μια βιβλιοθήκη για ένα περιοδικό. Τότε αρκεί εισάγουμε τον τίτλο του και να ανακτήσουμε τις ανάλογες εγγραφές. Λόγω αυτών των χαρακτηριστικών του, το Boolean μοντέλο έχει βρει εφαρμογή σε κυρίως εμπορικά συστήματα βιβλιοθηκών.

3.5.3 Το μοντέλο Vector Space

Το μοντέλο Vector Space [SL68, S71], αντιμετωπίζει την ανεπάρκεια της ανάθεσης δυαδικών βαρών και εισάγει ένα υπόβαθρο στο οποίο επιτρέπεται προσεγγιστικό ταίριασμα. Τα βάρη που ανατίθενται στους όρους δεικτοδότησης, τόσο για τα κείμενα όσο και για τα ερωτήματα είναι μη δυαδικά και χρησιμοποιούνται για τον υπολογισμό του *βαθμού ομοιότητας* μεταξύ του ερωτήματος και κάθε αποθηκευμένου κειμένου. Κατόπιν τα κείμενα διατάσσονται με φθίνουσα σειρά, με κριτήριο τον βαθμό ομοιότητάς τους με το ερώτημα του χρήστη. Έτσι στο μοντέλο Vector Space λαμβάνονται υπόψη και κείμενα που ικανοποιούν μερικώς τις συνθήκες του ερωτήματος και το τελικό αποτέλεσμα είναι πολύ πιο ακριβές σε σχέση με την Boolean ανάκτηση.

Ορισμός *Στο μοντέλο Vector Space το βάρος $w_{i,j}$ που αντιστοιχεί στο ζεύγος (k_i, d_j) είναι θετικό και όχι δυαδικό. Επιπλέον ανατίθενται βάρη και στους όρους δεικτοδότησης του ερωτήματος. Έστω $w_{i,q}$ το βάρος που αντιστοιχεί στο ζεύγος [k_i, q], όπου $w_{i,q} \geq 0$. Τότε το διάνυσμα του ερωτήματος ορίζεται ως $\vec{q}$ = ($w_{1,q}$, $w_{2,q}$, …, $w_{t,q}$) όπου* t *είναι ο συνολικός αριθμός των όρων δεικτοδότησης στο σύστημα. Όπως και πριν το διάνυσμα του $\vec{d}_j$, είναι $\vec{d}_j$ = ($w_{1,j}$, $w_{2,j}$, …, $w_{t,j}$).*

Μ' αυτόν τον τρόπο το κείμενο d_j και το ερώτημα χρήστη q αναπαρίστανται σαν διανύσματα διάστασης t. Στο μοντέλο Vector Space προτείνεται ο βαθμός της ομοιότητας μεταξύ του κειμένου d_j και του ερωτήματος q να υπολογιστεί ως ο βαθμός συσχέτισης (correlation) μεταξύ των δύο διανυσμάτων. Μέτρο του βαθμού συσχέτισης αποτελεί το *συνημίτονο της εμπεριεχόμενης γωνίας* των δύο διανυσμάτων. Συγκεκριμένα:

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

όπου $\left|\vec{d}_j\right|$ και $\left|\vec{q}\right|$ είναι οι νόρμες των διανυσμάτων.

Εφόσον $w_{i,j} \geq 0$ και $w_{i,q} \geq 0$, το $sim(d_j, q)$ παίρνει τιμές από 0 (καμία ομοιότητα) ως +1 (τέλειο ταίριασμα). Έτσι το μοντέλο Vector Space, αντί να προσπαθήσει να προσδιορίσει αν ένα κείμενο είναι ή όχι σχετικό, διατάσσει τα κείμενα με κριτήριο τον *βαθμό ομοιότητάς* τους προς το ερώτημα. Με αυτή τη στρατηγική ένα κείμενο μπορεί να ανακτηθεί ακόμα και αν ταιριάζει *κατά προσέγγιση* με το ερώτημα. Επειδή δεν θέλουμε να ανακτήσουμε όλα τα κείμενα που έχουν μη μηδενικό βαθμό σχετικότητας με το ερώτημα, αλλά αυτά που ταιριάζουν περισσότερο, θέτουμε ένα κατώφλι ελέγχου για το $sim(d_j, q)$. Κείμενα με βαθμό ομοιότητας μεγαλύτερο απ' αυτό το κατώφλι επιστρέφονται στο χρήστη. Πριν όμως υπολογίσουμε την κατάταξη των κειμένων πρέπει να εξετάσουμε τον τρόπο υπολογισμού των βαρών.

Το πρόβλημα υπολογισμού των βαρών ανάγεται στο εξής πρόβλημα ομαδοποίησης (clustering). Έστω μια συλλογή κειμένων C και ένα σύνολο A από κείμενα της συλλογής. Στο πρόβλημα της ΑΠ, το A είναι το σύνολο εκείνο των κειμένων που απαντούν σε μια

πληροφοριακή ανάγκη. Η διατύπωση της πληροφοριακής ανάγκης που καθορίζει το *A*, μπορεί να είναι σχετικά ασαφής, οπότε τα θέματα που πρέπει να αντιμετωπιστούν είναι δυο ειδών. Πρώτον, πρέπει να καθοριστεί ποια χαρακτηριστικά χαρακτηρίζουν τα κείμενα του *A*. Και δεύτερον πρέπει να καθοριστεί ποια χαρακτηριστικά διαχωρίζουν τα κείμενα του συνόλου *A* από τα κείμενα του *C*. Η εξισορρόπηση της επίδρασης αυτών των δύο ομάδων χαρακτηριστικών είναι το αντικείμενο ενός καλού σχήματος ανάθεσης βαρών.

Ένα καλό μέτρο για τον χαρακτηρισμό των στοιχείων εντός του συνόλου *A* είναι η συχνότητα εμφάνισης του όρου k_i σε κάθε κείμενο d_j. Διαισθητικά όσο πιο συχνά εμφανίζεται ένας όρος k_i σε ένα κείμενο d_j, τόσο πιο καλή περιγραφή του d_j αποτελεί ο k_i. Η συχνότητα εμφάνισης του όρου, ονομάζεται συχνά και *παράγοντας tf* (tf = term frequency). Επίσης ένα μέτρο για τον διαχωρισμό των συνόλων *A* και *C*, αποτελεί η αντίστροφη συχνότητα εμφάνισης του k_i στα κείμενα της συλλογής. Διαισθητικά αν ο k_i έχει μεγάλη συχνότητα εμφάνισης στη συλλογή, δεν είναι πολύ χρήσιμος για να χαρακτηρίσει ένα κείμενο και άρα να διαχωρίσει μια ομάδα κειμένων μες στη συλλογή. Η αντίστροφη συχνότητα εμφάνισης αναφέρεται συνήθως ως *παράγοντας idf* (idf = inverse document frequency). Το καλύτερο σχήμα υπολογισμού βαρών, πρέπει να προκύψει από τον κατάλληλο συνδυασμό αυτών των δύο παραγόντων.

Ορισμός *Έστω N ο συνολικός αριθμός των κειμένων και n_i ο αριθμός των κειμένων στα οποία εμφανίζεται ο όρος k_i. Έστω $freq_{i,j}$ η συχνότητα εμφάνισης του k_i στο d_j. Τότε η κανονικοποιημένη συχνότητα $f_{i,j}$ του όρου k_i στο d_j δίνεται από τη σχέση*

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \qquad (2.1)$$

όπου το max υπολογίζεται πάνω σε κάθε όρο που αναφέρεται στο κείμενο d_j. Αν ο k_i δεν εμφανίζεται στο d_j τότε $f_{i,j} = 0$. Επιπλέον, έστω idf_i η αντίστροφη συχνότητα εμφάνισης για τον k_i, που δίνεται από τον τύπο

$$idf_i = \log \frac{N}{n_i} \qquad (2.2)$$

Τότε το καλύτερο γνωστό σχήμα υπολογισμού βαρών δίνεται από το γινόμενο

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \qquad (2.3)$$

ή από παραλλαγή του παραπάνω. Τέτοια σχήματα υπολογισμού βαρών λέγονται σχήματα tf-idf.

Για τα βάρη των όρων στο ερώτημα οι Salton και Buckley [SB88], πρότειναν

$$w_{i,q} = \left(0.5 + \frac{0.5 freq_{i,q}}{\max_l freq_{l,q}} \right) \times \log \frac{N}{n_i} \qquad (2.4)$$

όπου $freq_{i,q}$ είναι η συχνότητα εμφάνισης του όρου k_i στο κείμενο που αντιπροσωπεύει την πληροφοριακή ανάγκη *q*. Ο αθροιστικός παράγοντας 0.5, έχει προκύψει πειραματικά πως εξισορροπεί το γεγονός ότι το ερώτημα απαρτίζεται συνήθως από πολύ λίγους όρους.

Τα κύρια πλεονεκτήματα του μοντέλου Vector Space, είναι τα εξής: 1) το σχήμα υπολογισμού βαρών που χρησιμοποιεί, βελτιώνει την απόδοση της ανάκτησης, 2) η στρατηγική προσεγγιστικού ταιριάσματος επιτρέπει την ανάκτηση κειμένων που προσεγγίζουν τις συνθήκες του ερωτήματος, 3) ο τρόπος του υπολογισμού της κατάταξης με βάση το συνημίτονο αφενός μεν επιτρέπει την ταξινόμηση των κειμένων βάσει του βαθμού ομοιότητάς τους με την ερώτηση αφετέρου δε υλοποιείται εύκολα με τις υπάρχουσες δομές δεικτοδότησης. Ένα μειονέκτημα είναι ότι οι όροι δεικτοδότησης θεωρούνται ανεξάρτητοι μεταξύ τους. Στην πράξη όμως, αυτό

μπορεί να είναι και πλεονέκτημα. Αν λαμβάναμε υπόψη τις συσχετίσεις των όρων μεταξύ τους για μία ολόκληρη μεγάλη συλλογή κειμένων τότε, λόγω της σχετικής τοπικότητας που αυτές τείνουν να έχουν, ενδεχομένως να εξάγαμε λανθασμένα συμπεράσματα και να βλάπταμε την συνολική απόδοση.

Τελικά το μοντέλο Vector Space, παρά την απλότητα της σύλληψής και της υλοποίησής του είναι ένα στιβαρό μοντέλο. Η δυνατότητα της εφαρμογής προσεγγιστικού ταιριάσματος, δίνει αποτελέσματα που είναι δύσκολο να βελτιωθούν χωρίς επέκταση του ερωτήματος ή εφαρμογή ανάδρασης χρήστη. Τα αλγεβρικά μοντέλα που ακολούθησαν το Vector Space αν και έχουν κατά σημεία καλύτερη απόδοση, είναι πιο δύσκολα στην υλοποίησή τους. Πάντως το Vector Space δεν αντιμετωπίζει επαρκώς τα προβλήματα της *Συνωνυμίας* και της *Πολυσημίας*. Λόγω όμως της ευκολίας στην υλοποίησή του, παραμένει το πιο δημοφιλές μοντέλο ΑΠ.

3.5.4 Το πιθανοτικό μοντέλο

Σ' αυτή την ενότητα παρουσιάζουμε το κλασσικό πιθανοτικό μοντέλο που πρωτοπαρουσιάστηκε από τους Robertson και Sparck Jones [RS76], το οποίο αργότερα έγινε γνωστό και ως μοντέλο *ανάκτησης δυαδικής ανεξαρτησίας* (binary independence retrieval – BIR). Η συζήτηση του μοντέλου είναι σύντομη και σκοπό έχει να τονίσει τα χαρακτηριστικά του μοντέλου και να δώσει την διαίσθηση πίσω από αυτό.

Το πιθανοτικό μοντέλο επιχειρεί να αντιμετωπίσει το πρόβλημα της ΑΠ παρέχοντας ένα πιθανοτικό υπόβαθρο. Η βασική ιδέα είναι η εξής. Δεδομένου ενός ερωτήματος χρήστη, υπάρχει ένα σύνολο κειμένων που αποτελείται ακριβώς από τα σχετικά κείμενα και μόνο απ' αυτά. Σ' αυτό το σύνολο θα αναφερόμαστε με τον όρο *ιδανικό* σύνολο απάντησης. Δεδομένης της περιγραφής του ιδανικού συνόλου απάντησης, δεν θα είχαμε κανένα πρόβλημα να ανακτήσουμε τα κείμενα που το αποτελούν. Συνεπώς μπορούμε να θεωρήσουμε ότι η διατύπωση ενός ερωτήματος ταυτίζεται με τη διαδικασία καθορισμού των ιδιοτήτων του ιδανικού συνόλου απάντησης (όπως όταν θέλαμε να περιγράψουμε το σύνολο A στο Vector Space). Το πρόβλημά μας είναι ότι δεν γνωρίζουμε ποιες ακριβώς είναι αυτές οι ιδιότητες. Το μόνο που έχουμε στη διάθεσή μας είναι μια ομάδα από όρους δεικτοδότησης, η σημασιολογία των οποίων μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει αυτές τις ιδιότητες. Αυτές οι ιδιότητες δεν είναι γνωστές τη στιγμή της διατύπωσης του ερωτήματος, οπότε πρέπει να γίνει μια αρχική προσπάθεια να προσδιοριστούν. Η αρχική αυτή εκτίμηση μας επιτρέπει να δημιουργήσουμε μια αρχική *πιθανοτική* περιγραφή του ιδανικού συνόλου απάντησης, η οποία θα χρησιμοποιηθεί για την ανάκτηση ενός πρώτου συνόλου κειμένων. Ακολουθεί αλληλεπίδραση με το χρήστη με σκοπό τη βελτίωση της περιγραφής του ιδανικού συνόλου απάντησης.

Ο χρήστης εξετάζει το αρχικό σύνολο των επιστρεφόμενων κειμένων και αποφασίζει ποια κείμενα είναι σχετικά και ποια όχι (στην πράξη αρκεί η εξέταση λίγων αρχικών κειμένων). Κατόπιν το σύστημα αξιοποιεί αυτή την πληροφορία για να βελτιώσει την περιγραφή του συνόλου απάντησης. Επαναλαμβάνοντας αυτή τη διαδικασία αρκετές φορές, αναμένεται ότι η περιγραφή θα συγκλίνει προς την ιδανική περιγραφή του συνόλου απάντησης. Έτσι πάντα θα πρέπει να έχουμε υπόψη μας την αρχική περιγραφή του ιδανικού συνόλου απάντησης. Επιπλέον πρέπει να γίνει προσπάθεια να περιγραφεί η παραπάνω διαδικασία, πιθανοτικά.

Το πιθανοτικό μοντέλο βασίζεται στην ακόλουθη θεμελιώδη υπόθεση.

Υπόθεση (Πιθανοτική Αρχή) Δοθέντος ενός ερωτήματος q και ενός κειμένου d_j της συλλογής, το πιθανοτικό μοντέλο προσπαθεί να εκτιμήσει την πιθανότητα ο χρήστης να βρει ενδιαφέρον το κείμενο d_j (δηλ. σχετικό προς το ερώτημα q). Υπόθεση του μοντέλου είναι ότι η πιθανότητα της σχετικότητας εξαρτάται από την αναπαράσταση του ερωτήματος και του κειμένου και μόνο. Επιπλέον γίνεται η υπόθεση ότι υπάρχει ένα υποσύνολο όλων των κειμένων,

το οποίο ο χρήστης προτιμά ως απάντηση στο ερώτημα *q*. Ένα τέτοιο *ιδανικό* σύνολο απάντησης, ονομάζεται *R* και θα πρέπει να μεγιστοποιεί τη συνολική πιθανότητα σχετικότητας προς την πληροφοριακή ανάγκη του χρήστη. Τα κείμενα στο *R* προβλέπεται ότι είναι *σχετικά* προς το ερώτημα. Τα κείμενα που δεν ανήκουν σ' αυτό το σύνολο προβλέπεται ότι είναι *μη-σχετικά*.

Μια τέτοια υπόθεση είναι κάπως προβληματική γιατί δεν παρέχει ένα μηχανισμό για τον υπολογισμό των πιθανοτήτων σχετικότητας. Επιπλέον ούτε καν προκύπτει ο δειγματοχώρος για τον υπολογισμό αυτών των πιθανοτήτων.

Δοθέντος λοιπόν ενός ερωτήματος *q*, το πιθανοτικό μοντέλο αναθέτει σε κάθε κείμενο d_j, την πιθανότητα να είναι σχετικό προς το ερώτημα. Η πιθανότητα αυτή δίνεται από το λόγο *P(d_j σχετικό με το q)/P(d_j μη σχετικό με το q)*. Λαμβάνοντας τον λόγο αυτό ως την συνάρτηση κατάταξης, ελαχιστοποιείται η πιθανότητα λανθασμένης κρίσης.

Ορισμός *Στο πιθανοτικό μοντέλο όλα τα βάρη των όρων δεικτοδότησης έχουν δυαδική μορφή δηλ,* $w_{i,j} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$. *Ένα ερώτημα q είναι ένα υποσύνολο των όρων δεικτοδότησης. Έστω R το σύνολο των κειμένων για το οποία υπάρχει η γνώση (ή αρχικά η εκτίμηση) ότι είναι σχετικά. Έστω* $\overline{R}$ *το συμπλήρωμα του R (δηλ. το σύνολο των μη σχετικών κειμένων). Έστω* $P(R \mid \vec{d}_j)$ *η πιθανότητα το κείμενο* d_j *να είναι σχετικό προς το ερώτημα q και* $P(\overline{R} \mid \vec{d}_j)$ *η πιθανότητα το κείμενο* d_j *να μην είναι σχετικό προς το ερώτημα q. Η ομοιότητα sim(d_j, q) του κειμένου* d_j *προς το ερώτημα q ορίζεται ως ο λόγος:*

$$sim(d_j, q) = \frac{P(R \mid \vec{d}_j)}{P(\overline{R} \mid \vec{d}_j)}$$

Από τον κανόνα του *Bayes*,

$$sim(d_j, q) = \frac{P(\vec{d}_j \mid R) \times P(R)}{P(\vec{d}_j \mid \overline{R}) \times P(\overline{R})}$$

όπου $P(\vec{d}_j \mid R)$ είναι η πιθανότητα το d_j να επιλέχθηκε τυχαία από το σύνολο *R,* δηλαδή να είναι σχετικό. Επιπλέον *P(R)* είναι η πιθανότητα το κείμενο που επιλέξαμε με τυχαίο τρόπο από ολόκληρη τη συλλογή, να είναι τυχαίο. Οι ερμηνείες των ποσοτήτων $P(\vec{d}_j \mid \overline{R})$ και $P(\overline{R})$ είναι ανάλογες των παραπάνω.

Καθώς οι ποσότητες *P(R)* και $P(\overline{R})$ είναι ίδιες για όλα τα κείμενα της συλλογής, μπορούμε να γράψουμε

$$sim(d_j, q) \approx \frac{P(\vec{d}_j \mid R)}{P(\vec{d}_j \mid \overline{R})}$$

Λόγω του ότι υποθέσαμε στοχαστική ανεξαρτησία στους όρους μπορούμε να γράψουμε την παραπάνω σχέση ως,

$$sim(d_j,q) \approx \frac{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i \mid R)\right) \times \left(\prod_{g_i(\vec{d}_j)=0} P(\bar{k}_i \mid R)\right)}{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i \mid \bar{R})\right) \times \left(\prod_{g_i(\vec{d}_j)=0} P(\bar{k}_i \mid \bar{R})\right)}$$

όπου $P(k_i \mid R)$ είναι η πιθανότητα ο όρος δεικτοδότησης k_i να εμφανίζεται σε ένα κείμενο το οποίο επιλέχθηκε τυχαία από το σύνολο *R*. Ο όρος $P(\bar{k}_i \mid R)$ δίνει την πιθανότητα ο όρος k_i να μην εμφανίζεται σε ένα κείμενο το οποίο επιλέχθηκε τυχαία από το σύνολο *R*. Οι πιθανότητες που σχετίζονται με το σύνολο $\bar{R}$ έχουν ανάλογη σημασία.

Λογαριθμίζοντας και λαμβάνοντας υπόψη μας ότι $P(k_i \mid R) + P(\bar{k}_i \mid R) = 1$, και αγνοώντας τους παράγοντες που είναι σταθεροί για όλα τα κείμενα για συγκεκριμένο ερώτημα, μπορούμε να γράψουμε,

$$sim(d_j,q) \approx \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \bar{R})}{P(k_i \mid \bar{R})}\right)$$

ο οποίος είναι ουσιαστικά ο τύπος με τον οποίο υπολογίζουμε την κατάταξη των κειμένων στο πιθανοτικό μοντέλο.

Καθώς δεν γνωρίζουμε το σύνολο *R* εξ' αρχής, είναι απαραίτητο να ορίσουμε μια μέθοδο υπολογισμού για τις πιθανότητες $P(k_i \mid R)$ και $P(\bar{k}_i \mid R)$. Παρακάτω θα δούμε με ποιον τρόπο μπορεί να γίνει ο υπολογισμός αυτός.

Αμέσως μετά την διατύπωση του ερωτήματος, δεν υπάρχουν ακόμα ανακτημένα κείμενα. Έτσι πρέπει να κάνουμε απλοποιητικές υποθέσεις σε ότι αφορά τις πιθανότητες. Αυτές οι υποθέσεις είναι: α) υποθέτουμε ότι η $P(k_i \mid R)$ είναι σταθερή για όλους τους όρους k_i (τυπικά 0.5) και β) υποθέτουμε ότι η κατανομή των όρων δεικτοδότησης στα μη σχετικά κείμενα μπορεί να προσεγγιστεί από την κατανομή των όρων δεικτοδότησης στο σύνολο των κειμένων (με άλλα λόγια, το μέγεθος του συνόλου των μη σχετικών κειμένων $\bar{R}$, είναι πολύ μεγαλύτερο από το μέγεθος *R*). Οι δύο παραπάνω υποθέσεις μας δίνουν,

$$P(k_i \mid R) = 0.5$$

$$P(k_i \mid \bar{R}) = \frac{n_i}{N}$$

όπου, όπως ορίστηκε προηγουμένως n_i είναι ο αριθμός των κειμένων που περιέχουν τον όρο k_i και *N* είναι ο συνολικός αριθμός των κειμένων της συλλογής. Έχοντας την αρχική εκτίμηση, μπορούμε να ανακτήσουμε ένα αρχικό σύνολο κειμένων που περιέχουν όρους που εμφανίζονται στο ερώτημα και να παρέχουμε μια πιθανοτική κατάταξη γι' αυτά. Κατόπιν βελτιώνουμε την αρχική κατάταξη με τον τρόπο που αναφέρουμε στη συνέχεια.

Έστω *V* ένα υποσύνολο των κειμένων που ανακτήθηκαν αρχικά και στα οποία δόθηκε μια κατάταξη από το πιθανοτικό μοντέλο. Για παράδειγμα το παραπάνω σύνολο θα μπορούσε να είναι τα κορυφαία *r* κείμενα, όπου το *r* είναι ένα προκαθορισμένο κατώφλι. Έστω επίσης V_i ένα υποσύνολο του *V* το οποίο αποτελείται από τα κείμενα που περιέχουν τον όρο k_i. Για λόγους απλότητας, θα χρησιμοποιούμε τους όρους *V* και V_i για να αναφερόμαστε στους πληθαρίθμους των αντιστοίχων συνόλων. Το πότε θα αναφερόμαστε στο ίδιο το σύνολο ή στο μέγεθός του θα είναι ξεκάθαρο από τα συμφραζόμενα. Για να βελτιώσουμε την πιθανοτική

κατάταξη, πρέπει να βελτιώσουμε τις εκτιμήσεις για τα $P(k_i|R)$ και $P(k_i \mid \overline{R})$. Αυτό επιτυγχάνεται με τις ακόλουθες υποθέσεις: α) μπορούμε να προσεγγίσουμε την $P(k_i|R)$ με την κατανομή του όρου k_i στα κείμενα που ανακτήθηκαν μέχρι στιγμής (σύνολο V), β) μπορούμε να προσεγγίσουμε την $P(k_i \mid \overline{R})$, αν θεωρήσουμε όλα τα μη ανακτημένα κείμενα ως μη-σχετικά. Έτσι μπορούμε να γράψουμε,

$$P(k_i \mid R) = \frac{V_i}{V}$$

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i}{N - V}$$

Αυτή η διαδικασία μπορεί να επαναληφθεί αναδρομικά, υπολογίζοντας κάθε φορά νέα V και V_i. Έτσι είμαστε σε θέση να βελτιώσουμε τις εκτιμήσεις μας για τα $P(k_i \mid R)$ και $P(k_i \mid \overline{R})$ χωρίς καμία εμπλοκή του ανθρωπίνου παράγοντα. Ενδεχομένως όμως να είναι απαραίτητη η ανθρώπινη παρέμβαση στην κατασκευή του συνόλου V.

Οι τελευταίοι δυο τύποι για τα $P(k_i \mid R)$ και $P(k_i \mid \overline{R})$ παρουσιάζουν προβλήματα για μικρές τιμές των V και V_i που εμφανίζονται στην πράξη (όπως π.χ. $V = 1$ και $V_i = 0$). Για να αντιμετωπιστούν αυτά τα προβλήματα, συνήθως εισάγεται ένας προσθετικός παράγοντας οπότε οι παραπάνω τύποι γράφονται ως:

$$P(k_i \mid R) = \frac{V_i + 0.5}{V + 1}$$

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i + 0.5}{N - V + 1}$$

Συχνά ένας σταθερός προσθετικός παράγοντας, όπως το 0.5, δεν είναι επαρκής. Μια εναλλακτική λύση είναι να θεωρηθεί ως προσθετικός παράγοντας η ποσότητα n_i/N, που μας δίνει,

$$P(k_i \mid R) = \frac{V_i + \frac{n_i}{N}}{V + 1}$$

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1}$$

Το κύριο *πλεονέκτημα* του πιθανοτικού μοντέλου είναι ότι τα κείμενα κατάσσονται σε φθίνουσα σειρά με βάση την *πιθανότητα* να είναι σχετικά με το αρχικό ερώτημα. Τα *μειονεκτήματα* είναι ότι 1) χρειάζεται μια αρχική *εκτίμηση* για τον διαχωρισμό της συλλογής των κειμένων σε σχετικά και μη, 2) δεν λαμβάνεται υπόψη η συχνότητα εμφάνισης του όρου μέσα σε ένα κείμενο (όλα τα βάρη είναι 0 ή 1), 3) η υιοθέτηση της άποψης ότι οι όροι είναι μεταξύ τους ανεξάρτητοι.

3.6 Εναλλακτικά συνολοθεωρητικά μοντέλα

Σ' αυτή την ενότητα θα συζητήσουμε δύο εναλλακτικά συνολοθεωρητικά μοντέλα, το μοντέλο ασαφών συνόλων (fuzzy set model) και το επεκτεταμένο Boolean μοντέλο (extended Boolean model).

3.6.1 Μοντέλο ασαφών συνόλων

Όπως είδαμε η αναπαράσταση των κειμένων και των ερωτημάτων με τη χρήση ενός συνόλου από λέξεις-κλειδιά δίνει μια προσεγγιστική περιγραφή του σημασιολογικού περιεχομένου τους. Κατά συνέπεια και το ταίριασμα των κειμένων με τις συνθήκες του ερωτήματος είναι επίσης προσεγγιστικό. Αυτή η διαδικασία μπορεί να μοντελοποιηθεί θεωρώντας ότι κάθε όρος που εμφανίζεται στο ερώτημα, ορίζει ένα *ασαφές* (fuzzy) σύνολο και κάθε κείμενο έχει έναν *βαθμό συμμετοχής* (membership degree) σ' αυτό, ο οποίος παίρνει τιμές από 0 (βεβαιότητα ότι δεν ανήκει στο σύνολο) μέχρι 1 (βεβαιότητα ότι ανήκει). Η ερμηνεία της διαδικασίας ΑΠ, με χρήση των εννοιών της *ασαφούς θεωρίας* (fuzzy theory) ορίζει το μοντέλο που θα εξετάσουμε ευθύς αμέσως. Πριν όμως θα συζητήσουμε κάποια στοιχεία της ασαφούς θεωρίας συνόλων.

Ασαφής θεωρία συνόλων

Η *ασαφής θεωρία συνόλων* (fuzzy set theory) [Z93], ασχολείται με την αναπαράσταση των συνόλων, των οποίων τα όρια δεν είναι καλά ορισμένα. Η βασική ιδέα είναι να αναθέσουμε σε κάθε στοιχείο του συνόλου, μια συνάρτηση που δείχνει το βαθμό συμμετοχής του στο σύνολο. Αυτή η συνάρτηση παίρνει τιμές στο διάστημα [0,1] με το 0 να αντιστοιχεί σε μη συμμετοχή στο σύνολο και το 1, σε πλήρη συμμετοχή. Τιμές του βαθμού συμμετοχής, στο ενδιάμεσο διάστημα, προσδιορίζουν τα *περιθωριακά στοιχεία* του συνόλου, δηλαδή στοιχεία για τα οποία υπάρχει αβεβαιότητα για το αν ανήκουν στο σύνολο. Η αβεβαιότητα μεγαλώνει όσο πιο μικρός είναι ο βαθμός συμμετοχής στο σύνολο. Με άλλα λόγια η συμμετοχή σε ένα ασαφές σύνολο έχει διάφορες *διαβαθμίσεις* και όχι διακριτό χαρακτήρα όπως στη Boolean λογική.

Ορισμός *Ένα ασαφές υποσύνολο A του σύμπαντος των κειμένων U χαρακτηρίζεται από μια συνάρτηση βαθμού συμμετοχής* $\mu_A : U \rightarrow [0,1]$, *η οποία συσχετίζει με κάθε στοιχείο u του U έναν αριθμό* $\mu_A(u)$ *ο οποίος παίρνει τιμές στο διάστημα* [0,1].

Οι τρεις πιο κοινότυπες πράξεις σε ασαφή σύνολα είναι: το *συμπλήρωμα* (complement) ενός ασαφούς συνόλου, η *ένωση* (union) δυο ή περισσοτέρων ασαφών συνόλων και η *τομή* (intersection) δυο ή περισσοτέρων ασαφών συνόλων. Ακολουθούν οι ορισμοί.

Ορισμός *Έστω U το σύμπαν των κειμένων και A και B δυο ασαφή υποσύνολα του U και* $\overline{A}$ *το συμπλήρωμα του A. Έστω επίσης u, ένα στοιχείο του U. Τότε,*

$$\mu_{\overline{A}}(u) = 1 - \mu_A(u)$$

$$\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$$

$$\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$$

Τα ασαφή σύνολα είναι χρήσιμα για την αναπαράσταση της αβεβαιότητας και της ανακρίβειας και έχουν χρησιμοποιηθεί σε πολλά πεδία, μεταξύ αυτών και της ανάκτησης πληροφορίας.

Ασαφής Ανάκτηση Πληροφορίας

Μια εναλλακτική μοντελοποίηση της διαδικασίας ανάκτησης πληροφορίας αποκτούμε με τη χρήση ενός *θησαυρού* (thesaurus). Ο θησαυρός είναι μια βοηθητική δομή που περιέχει τις συσχετίσεις των όρων μεταξύ τους, παρέχοντας στην ουσία εναλλακτικούς ορισμούς για κάθε όρο που χρησιμοποιείται. Η ιδέα εδώ είναι να επεκταθεί το σύνολο των όρων δεικτοδότησης που χρησιμοποιούνται στο ερώτημα με όρους από το θησαυρό και έτσι να ανακτηθούν επιπλέον σχετικά με το ερώτημα κείμενα. Ο θησαυρός μπορεί επίσης να χρησιμοποιηθεί για να μοντελοποιήσει την ανάκτηση πληροφορίας μέσω ασαφών συνόλων.

Ο θησαυρός μπορεί να κατασκευαστεί με τη δημιουργία ενός πίνακα $\vec{c}$, *συσχέτισης των όρων* (term-term correlation), γραμμές και στήλες του οποίου συνδέονται με τους όρους ευρετηρίου της συλλογής. Στον πίνακα $\vec{c}$, μπορεί να οριστεί ένας κανονικοποιημένος συντελεστής συσχέτισης $c_{i,l}$ μεταξύ των όρων k_i και k_l ως εξής,

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

όπου n_i είναι ο αριθμός των κειμένων που περιέχουν τον όρο k_i, n_l είναι ο αριθμός των κειμένων που περιέχουν τον όρο k_l, και $n_{i,l}$ είναι ο αριθμός των κειμένων που περιέχουν και τους δυο όρους.

Μπορούμε να χρησιμοποιήσουμε τον πίνακα $\vec{c}$ για να ορίσουμε ένα ασαφές σύνολο για κάθε όρο δεικτοδότησης k_i. Σε αυτό το ασαφές σύνολο το κείμενο d_j έχει βαθμό συμμετοχής $\mu_{i,j}$ που υπολογίζεται από τη σχέση,

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} \left(1 - c_{i,l}\right)$$

η οποία υπολογίζει το αλγεβρικό άθροισμα (εκπεφρασμένο ως συμπλήρωμα του αλγεβρικού γινομένου, π.χ. $A+B = 1 – (1 – A)(1 – B)$) πάνω σε όλους του όρους που ανήκουν στο κείμενο d_j. Ένα κείμενο d_j ανήκει στο ασαφές σύνολο που ορίζει ο όρος k_i αν οι όροι που περιέχει σχετίζονται με τον k_i. Στην περίπτωση που υπάρχει τουλάχιστον ένας όρος k_l του d_j που έχει μεγάλη συσχέτιση με τον όρο k_i (δηλ, $c_{i,l} \sim 1$), τότε $\mu_{i,j} \sim 1$ και τότε ο k_i είναι κατάλληλος όρος ασαφούς δεικτοδότησης (fuzzy index) για το d_j. Στην περίπτωση που όλοι οι όροι δεικτοδότησης του d_j είναι ασθενώς συσχετισμένοι με τον k_i, τότε ο k_i δεν είναι κατάλληλος όρος ασαφούς δεικτοδότησης (fuzzy index) για το d_j (δηλ. $\mu_{i,j} \sim 0$). Η υιοθέτηση του αλγεβρικού αθροίσματος όλων των όρων του κειμένου d_j, αντί για τη συνάρτηση *max*, επιτρέπει μια πιο ομαλή μετάβαση για τις τιμές του παράγοντα $\mu_{i,j}$.

Όπως ακριβώς και στο Boolean μοντέλο, ο χρήστης εκφράζει την πληροφοριακή του ανάγκη με την διατύπωση ενός Boolean ερωτήματος. Το ερώτημα αυτό μετατρέπεται σε διαζευκτική κανονική μορφή (DNF). Για παράδειγμα το ερώτημα $[q = k_a \wedge (k_b \vee \neg k_c)]$ μπορεί να γραφεί σε DNF ως $[\vec{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)]$, όπου κάθε όρος που συμμετέχει στη διάζευξη του ερωτήματος $\vec{q}_{dnf}$, είναι το διάνυσμα με δυαδικά βάρη για την τριάδα (k_a, k_b, k_c), που αναλύσαμε στην ενότητα 3.5.2. Υπενθυμίζουμε ότι το καθένα απ’ αυτά τα διανύσματα είναι συζευκτική συνιστώσα (οι όροι εμφανίζονται με *and*). Έστω cc_i ένας δείκτης στην *i*–στή τέτοια συνιστώσα. Τότε,

$$\vec{q}_{dnf} = cc_1 \vee cc_2 \vee \ldots \vee cc_p$$

όπου p είναι ο αριθμός των συζευκτικών συνιστωσών του $\vec{q}_{dnf}$. Η διαδικασία υπολογισμού των κειμένων που είναι σχετικά προς ένα ερώτημα, είναι ανάλογη με αυτή που χρησιμοποιείται στο κλασσικό Boolean μοντέλο. Η μόνη διαφορά είναι ότι χειριζόμαστε ασαφή, αντί για καλά ορισμένα σύνολα. Ας δούμε ένα παράδειγμα.

Έστω πάλι το ερώτημα $[q = k_a \wedge (k_b \vee \neg k_c)]$. Έστω D_a το ασαφές σύνολο των κειμένων που ορίζονται από τον όρο k_a. Το σύνολο αυτό αποτελείται για παράδειγμα από όλα τα κείμενα d_j που έχουν βαθμό συμμετοχής $\mu_{a,j}$ στο σύνολο, που υπερβαίνει ένα προκαθορισμένο κατώφλι K. Έστω επίσης το σύνολο $\overline{D_a}$, το συμπλήρωμα του D_a. Το ασαφές σύνολο, ορίζεται από τον όρο $\overline{k}_a$, την άρνηση του όρου k_a. Ανάλογα ορίζουμε τα σύνολα D_b

και D_c, για τους όρους k_b και k_c. Μια και όλα τα σύνολα είναι ασαφή, το κείμενο d_j μπορεί να ανήκει για παράδειγμα στο σύνολο D_a, ακόμα και αν το κείμενο του δεν αναφέρει τον όρο k_a.

Το ασαφές σύνολο D_q για το ερώτημα είναι η ένωση των ασαφών συνόλων που ορίζονται από τις τρεις συζευκτικές συνιστώσες του $\vec{q}_{dnf}$ (στις οποίες αναφερόμαστε ως cc_1, cc_2, και cc_3 αντίστοιχα). Ο βαθμός συμμετοχής του κειμένου d_j στο ασαφές σύνολο D_q υπολογίζεται ως εξής.

$$\begin{aligned}\mu_{q,j} &= \mu_{(cc_1+cc_2+cc_3),j} \\ &= 1-\prod_{i=1}^{3}(1-\mu_{cc_i,j}) \\ &= 1-(1-\mu_{a,j}\mu_{b,j}\mu_{c,j})\times \\ &\quad (1-\mu_{a,j}\mu_{b,j}(1-\mu_{c,j}))\times(1-\mu_{a,j}(1-\mu_{b,j})(1-\mu_{c,j}))\end{aligned}$$

όπου $\mu_{i,j}$, $i\in\{a, b, c\}$ είναι ο βαθμός συμμετοχής του d_j στο ασαφές σύνολο που ορίζεται από τον k_i.

Όπως προείπαμε, ο βαθμός συμμετοχής σε ένα διαζευκτικό ασαφές σύνολο υπολογίζεται με τη χρήση *αλγεβρικού αθροίσματος* και όχι με τη συνάρτηση *max*. Επιπλέον ο βαθμός συμμετοχής σε ένα διαζευκτικό ασαφές σύνολο, υπολογίζεται χρησιμοποιώντας *αλγεβρικό γινόμενο* και όχι με τη χρήση της συνάρτησης *min*. Αυτή η χρήση των αλγεβρικών συνολοπράξεων δίνει βαθμούς συμμετοχής οι οποίοι μεταβάλλονται ομαλότερα, από τους αντίστοιχους που υπολογίζονται με τη χρήση των συναρτήσεων *min* και *max*, και θεωρούνται καταλληλότερες για ένα σύστημα ΑΠ.

Το προηγούμενο παράδειγμα δείχνει πως υπολογίζεται το ασαφές σύνολο για τους όρους δεικτοδότησης δεδομένου ερωτήματος και πως κατατάσσονται τα κείμενα σύμφωνα με τον βαθμό συμμετοχής τους σ' αυτό το σύνολο. Το μοντέλο χρησιμοποιεί τον πίνακα συσχέτισης των όρων μεταξύ τους, για τον υπολογισμό των συσχετίσεων μεταξύ του κειμένου d_j και των όρων που το αποτελούν. Επιπλέον χρησιμοποιούνται αλγεβρικά αθροίσματα και γινόμενα για να υπολογιστεί ο συνολικός βαθμός συμμετοχής του d_j στο ασαφές σύνολο που ορίζεται από το ερώτημα χρήστη.

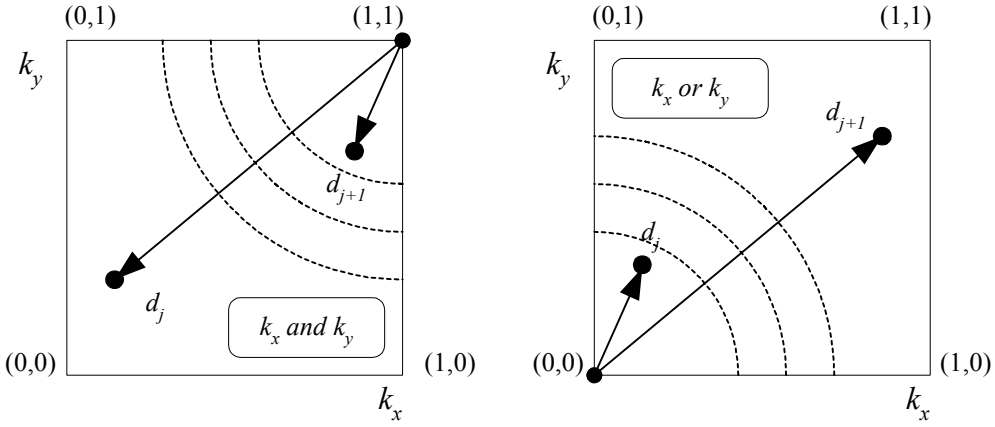
Γενικά το μοντέλο ασαφών συνόλων συζητήθηκε κυρίως στα πλαίσια μελέτης των ασαφών συνόλων και δεν είναι ιδιαίτερα δημοφιλές στην επιστημονική κοινότητα της ΑΠ.

3.6.2 Επεκτεταμένο Boolean μοντέλο

Το κλασσικό Boolean μοντέλο, παρά την απλότητα και την αυστηρότητα της θεμελίωσής του, έχει το μειονέκτημα ότι δεν προβλέπει κατάταξη σπουδαιότητας των κειμένων. Αυτό έχει σαν αποτέλεσμα το σύνολο της απάντησης να είναι είτε πολύ μικρό είτε πολύ μεγάλο και η ποιότητα της ανάκτησης μέτρια. Γι' αυτό το λόγο συνήθως στα περισσότερα συστήματα χρησιμοποιείται το μοντέλο Vector Space, το οποίο παρέχει μηχανισμούς κατάταξης των κειμένων (έχοντας έτσι καλύτερη απόδοση), ενώ παράλληλα είναι σχετικά απλό στην υλοποίησή του. Δεδομένων των πλεονεκτημάτων αυτών του Vector Space, έχουν γίνει προσπάθειες να επεκταθεί το Boolean μοντέλο με χαρακτηριστικά του Vector Space. Μια από αυτές τις προσεγγίσεις, το *επεκτεταμένο Boolean* (extended Boolean) μοντέλο, μελετήθηκε από τους Salton, Fox και Wu [SFW83] και είναι αντικείμενο αυτής της υποενότητας.

Το επεκτεταμένο Boolean μοντέλο βασίζεται σε μια κριτική στη βασική Boolean λογική. Έστω το εξής συζευκτικό ερώτημα, $q = k_x \wedge k_y$. Στο Boolean μοντέλο κάθε κείμενο περιέχει μόνο τον k_x ή τον k_y είναι το εξίσου άσχετο με το ερώτημα όσο και ένα κείμενο που δεν

περιέχει κανέναν από τους δυο. Αυτή η στρατηγική δυαδικής απόφασης είναι βέβαια έξω και από την καθημερινή λογική. Ο ίδιος συλλογισμός ισχύει και για καθαρά διαζευκτικά ερωτήματα.



Εικόνα 3-3: Επέκταση της Boolean λογικής, θεωρώντας μόνο το χώρο που αποτελείται από τους δύο όρους k_x και k_y.

Όταν τα ερωτήματα περιέχουν μόνο δυο όρους, μπορούμε να αναπαραστήσουμε τα κείμενα και τα ερωτήματα στις δύο διαστάσεις όπως φαίνεται στην Εικόνα 3-3. Ένα κείμενο d_j τοποθετείται σ' αυτό το χώρο με τη χρήση των βαρών $w_{x,j}$ και $w_{y,j}$ για τα ζεύγη $[k_x, d_j]$ και $[k_y, d_j]$, αντίστοιχα. Υποθέτουμε ότι αυτά τα βάρη είναι κανονικοποιημένα και άρα οι τιμές τους βρίσκονται μεταξύ 0 και 1. Για παράδειγμα τα βάρη αυτά μπορούν να υπολογιστούν ως κανονικοποιημένοι tf-idf παράγοντες, ως εξής,

$$w_{x,j} = f_{x,j} \times \frac{idf_x}{\max_i idf_i}$$

όπου όπως ορίστηκε στην Εξίσωση 2.1, $f_{x,j}$ είναι η κανονικοποιημένη συχνότητα του όρου k_x στο κείμενο d_j και idf_i είναι η αντίστροφη συχνότητα εμφάνισης για κάποιον όρο k_i. Για λόγους απλότητας στη συνέχεια θα αναφερόμαστε στο βάρος $w_{x,j}$ ως x, στο βάρος $w_{y,j}$ ως y, και στο διάνυσμα του κειμένου $\vec{d}_j = (w_{x,j}, w_{y,j})$ ως το σημείο $d_j = (x,y)$. Παρατηρώντας την Εικόνα 3-3, επισημαίνουμε δυο ιδιαιτερότητες. Πρώτον για το διαζευκτικό ερώτημα $q_{or} = k_x \vee k_y$, το σημείο (0, 0) είναι το σημείο που πρέπει να αποφευχθεί με την έννοια ότι, τα σχετικά προς το ερώτημα κείμενα, πρέπει να είναι όσο το δυνατόν πιο απομακρυσμένα στο χώρο, απ' αυτό το σημείο. Με άλλα λόγια η απόσταση από το σημείο (0, 0), θα πρέπει να ληφθεί ως μέτρο της ομοιότητας προς το ερώτημα q_{or}. Κατά δεύτερον στο συζευκτικό ερώτημα, $q_{and} = k_x \wedge k_y$, το επιθυμητό σημείο είναι το (1, 1). Αυτό σημαίνει ότι το σαν μέτρο ομοιότητας με το ερώτημα q_{and}. Επιπλέον μπορούμε να κανονικοποιήσουμε αυτές τις αποστάσεις και έτσι έχουμε,

$$sim(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

$$sim(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

Αν όλα τα βάρη είναι δυαδικά (δηλ. $w_{x,j} \in \{0, 1\}$), τα κείμενα τοποθετούνται σε μια από τις τέσσερις γωνίες (δηλ. (0, 0), (0, 1), (1, 0) ή (1, 1)) και οι τιμές του $sim(q_{or}, d)$ είναι 0, $1/\sqrt{2}$ και 1. Ανάλογα το $sim(q_{and}, d)$ παίρνει τις τιμές 0, 1 - $1/\sqrt{2}$ και 1.

Δεδομένου ότι ο πληθάριθμος των όρων δεικτοδότησης της συλλογής είναι *t*, τότε το παραπάνω Boolean μοντέλο μπορεί να επεκταθεί για τον υπολογισμό Ευκλείδειων αποστάσεων στις *t* διαστάσεις. Πάντως μπορούμε να μεταβούμε σε μια πιο εκτενή γενίκευση, με τη χρήση των *νορμών* διανυσμάτων.

Το μοντέλο *p*–νόρμας γενικεύει την έννοια της απόστασης για να περιέχει όχι μόνο Ευκλείδειες αποστάσεις αλλά και *p*–αποστάσεις, όπου $1 \leq p \leq \infty$, είναι μια καινούρια παράμετρος της οποίας η τιμή πρέπει να γίνει γνωστή τη στιγμή υποβολής της ερώτησης. Το γενικευμένο διαζευκτικό ερώτημα, θεωρώντας *p*–αποστάσεις στις *t* διαστάσεις, μπορεί τώρα να γραφεί ως

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m$$

Ανάλογα το γενικευμένο συζευκτικό ερώτημα μπορεί να γραφεί ως,

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$$

Οι αντίστοιχες ομοιότητες ερωτήματος-κειμένου δίνονται τώρα από τους τύπους,

$$sim(q_{or}, d) = \left(\frac{x_1^p + x_2^p + \ldots + x_m^p}{m} \right)^{\frac{1}{p}}$$

$$sim(q_{and}, d) = 1 - \left(\frac{(1 - x_1)^p + (1 - x_2)^p + \ldots + (1 - x_m)^p}{m} \right)^{\frac{1}{p}}$$

όπου κάθε x_i αντιστοιχεί στο βάρος $w_{i,j}$ που συνδέεται με το ζεύγος $[k_i, d_j]$.

Ο παραπάνω ορισμός της *p*–νόρμας, έχει διάφορες ενδιαφέρουσες ιδιότητες, όπως θα δούμε στη συνέχεια. Πρώτον αν $p = 1$, παρατηρούμε ότι

$$sim(q_{or}, d_j) = sim(q_{and}, d_j) = \frac{x_1 + x_2 + \ldots + x_m}{m}$$

Δεύτερον αν $p = \infty$, προκύπτει ότι,

$$sim(q_{or}, d_j) = \max(x_i)$$

$$sim(q_{and}, d_j) = \min(x_i)$$

Συμπερασματικά, όταν $p = 1$, το κριτήριο υπολογισμού ομοιότητας για συζευκτικά και διαζευκτικά ερωτήματα, είναι παρόμοιο μ' αυτό που χρησιμοποιείται στο μοντέλο Vector Space, όπου υπολογίζεται έμμεσα το άθροισμα με τη μορφή εσωτερικού γινομένου. Όταν πάλι $p = \infty$, τότε τα ερωτήματα αναπαρίστανται με βάση την ασαφή λογική, την οποία είδαμε στην προηγούμενη υποενότητα, ως γενίκευση της Boolean λογικής. Χρησιμοποιώντας τιμές για το *p*, μεταξύ 1 και ∞, το σχήμα κατάταξης με την χρήση *p*–νόρμας, προσαρμόζεται σε μια συμπεριφορά ενδιάμεσα σ' αυτή του Vector Space και σ' αυτή της Boolean λογικής. Η ευέλικτη αυτή συμπεριφορά αποτελεί το δυνατό σημείο του επεκτεταμένου Boolean μοντέλου.

Η επεξεργασία περισσότερο γενικών και ανάμικτων ερωτημάτων, γίνεται με την ομαδοποίηση των τελεστών με κάποια προκαθορισμένη σειρά. Για παράδειγμα στο ερώτημα

$q = (k_1 \wedge^p k_2) \vee^p k_3$, η ομοιότητα $sim(q, d_j)$ μεταξύ του κειμένου d_j και του ερωτήματος αυτού, μπορεί να υπολογιστεί από τον τύπο,

$$sim(q,d) = \left(\frac{\left(1 - \left(\frac{(1-x_1)^p + (1-x_2)^p}{2} \right)^{\frac{1}{p}} \right)^p + x_3^p}{2} \right)^{\frac{1}{p}}$$

Η παραπάνω διαδικασία μπορεί να επαναληφθεί αναδρομικά για οποιοδήποτε αριθμό τελεστών AND/OR.

Ένα επιπλέον ενδιαφέρον χαρακτηριστικό του επεκτεταμένου Boolean μοντέλου, είναι η δυνατότητα χρήσης συνδυασμού διαφορετικών τιμών της παραμέτρου p μέσα στο ίδιο ερώτημα. Για παράδειγμα θα μπορούσαμε να χρησιμοποιήσουμε το ερώτημα

$$(k_1 \vee^2 k_2) \wedge^\infty k_3$$

για να δηλώσουμε ότι, στα κείμενα του αποτελέσματος δεν είναι αναγκαίο να περιέχονται είτε ο k_1 είτε ο k_2, αρκεί τα κείμενα να κριθούν σχετικά από το σύστημα (όπως ακριβώς λειτουργεί το μοντέλο Vector Space), αλλά η ύπαρξη του όρου k_3 στα κείμενα του αποτελέσματος είναι απαραίτητη, (όπως ακριβώς θα λειτουργούσε ένα Boolean κριτήριο). Αυτή η δυνατότητα που μας δίνεται δεν είναι ξεκάθαρο αν έχει κάποια πρακτική χρησιμότητα, αλλά είναι θετικό ότι μας δίνεται από το μοντέλο χωρίς την ανάγκη για περίεργες επεκτάσεις για τον χειρισμό ειδικών περιπτώσεων.

Πρέπει επίσης να παρατηρήσουμε ότι το επεκτεταμένο Boolean μοντέλο, χειρίζεται λιγότερο αυστηρά την άλγεβρα Boole, με το να χειρίζεται τις Boolean πράξεις, με τη βοήθεια αλγεβρικών αποστάσεων. Υπ' αυτή την έννοια, πρόκειται για ένα μάλλον υβριδικό μοντέλο, που συνδυάζει ιδιότητες από τα συνολοθεωρητικά και τα αλγεβρικά μοντέλα. Το επεκτεταμένο Boolean μοντέλο δεν έχει βρει μεγάλη χρήση αλλά το κομψό θεωρητικό υπόβαθρο που παρέχει μπορεί να αποδειχτεί χρήσιμο στο μέλλον.

Βιβλιογραφία

[BR99] R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley Longman Inc., 1999

[RS76] S. E. Robertson, K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Sciences,* 27(3): 129-146, 1976

[S71] G. Salton. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall Inc., 1971

[SB88] G. Salton, C. Buckley. Term-weighting approaches in automatic retrieval. *Information Processing & Management*, 24(5): 513-523, 1988

[SFW83] G. Salton, E.A. Fox, H. Wu. Extended Boolean information retrieval. *Communications of the ACM,* 26(11): 1022-1036, November 1983

[SL68] G. Salton, M. E. Lesk. Computer evaluation of indexing and text processing. *Journal of the ACM,* 15(1): 8-36, January 1968

[Z93] L. A. Zadeh. Fuzzy sets. In D. Dubois, H. Prade and R. R. Yager editors, *Readings in fuzzy sets for intelligent systems*. Morgan Kaufmann, 1993