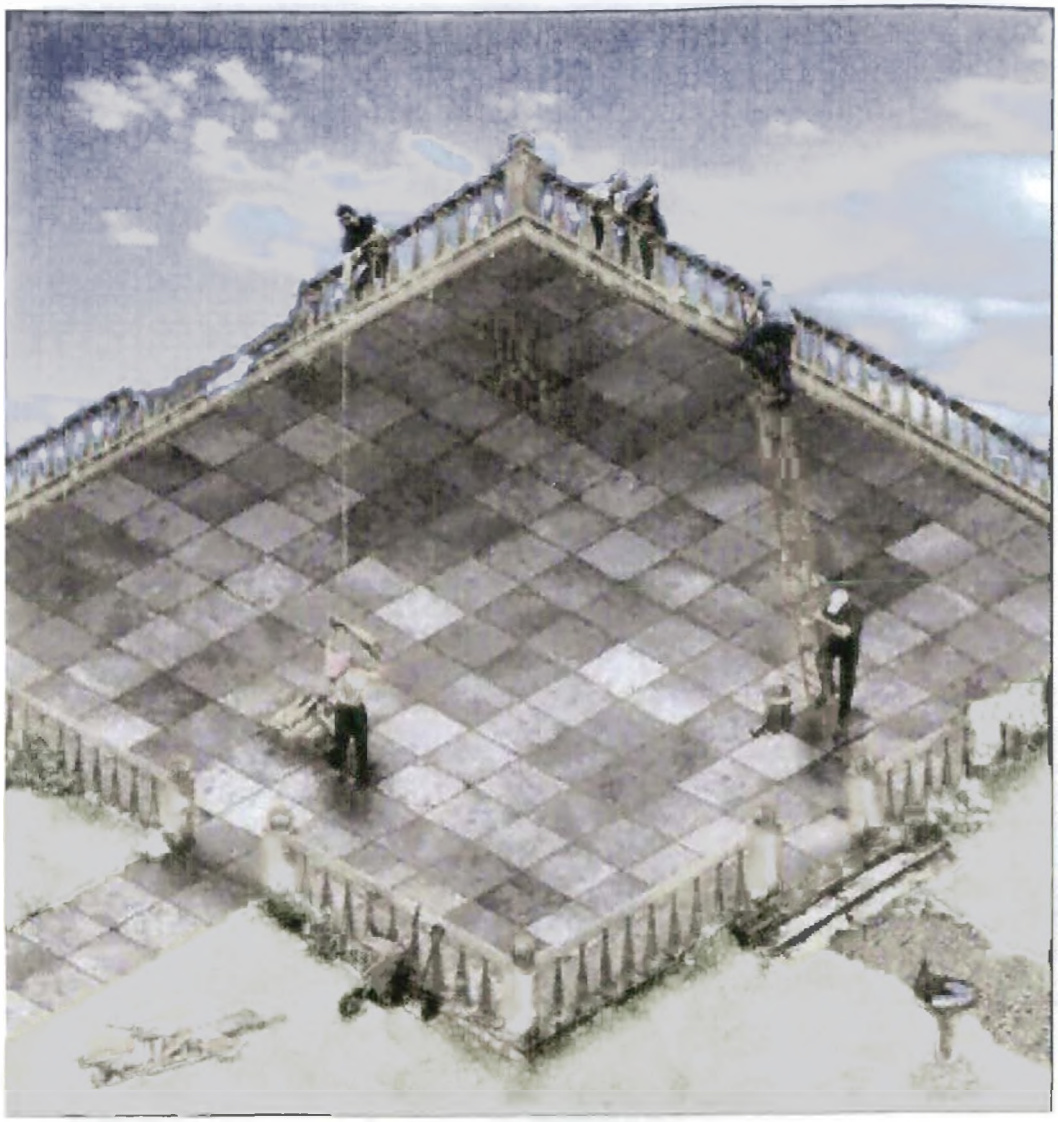
ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ (ΤΕΙ) ΜΕΣΟΛΟΓΓΙΟΥ
ΤΜΗΜΑ Ε.Π.Δ.Ο.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΜΕ ΘΕΜΑ:

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ



ΤΡΙΑΝΤΑΦΥΛΛΟΥ ΧΡΙΣΤΟΦΟΡΟΣ

ΚΩΣΤΑΡΑ ΣΠΥΡΙΔΟΥΛΑ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΚΩΝΣΤΑΝΤΙΝΟΣ ΙΩΑΝΝΟΥ

• ΜΕΣΟΛΟΓΓΙ 2006 •

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε όλους τους καθηγητές που στάθηκαν δίπλα μας και μας δώσανε όλες τις απαραίτητες γνώσεις. Για την πραγματοποίηση αυτής της πτυχιακής εργασίας ευχαριστούμε πάρα πολύ τον επιβλέπων καθηγητή μας κύριο Κώστα Ιωάννου για την πολύτιμη βοήθεια του και την άριστη συνεργασία μας.

Αφιερώνεται σε όλους αυτούς που αγωνίζονται προσφέροντας την επιστημονική τους γνώση, για την καλυτέρευση της ζωής των ανθρώπων. Σε όλους αυτούς που πιστεύουν ότι η επιστήμη πρέπει να είναι προς όφελος όλων των λαών της γης.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1 6

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ 6
Εισαγωγή 6
Ανάκτηση Πληροφοριών 7
Η διαδικασία χρήστη 9
Ένα σύστημα ανάκτησης πληροφορίας 10
Η προοπτική της ανάκτησης πληροφορίας 11
Αποτελεσματικότητα και αποδοτικότητα 12

ΚΕΦΑΛΑΙΟ 2 12

ΑΥΤΟΜΑΤΗ ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΩΝ 12
Ιδέες του Luhn 13
Παράγωγοι αντιπροσωπευτικών εγγράφων 14
Δεικτοδότηση 16
Στάθμιση όρου δεικτών 17
Πιθανολογική δεικτοδότηση 19
Διάκριση και/ή αντιπροσώπευση 22
Αυτόματη ταξινόμηση λέξεων κλειδιών 23
Κανονικοποίηση 24

ΚΕΦΑΛΑΙΟ 3 25

ΑΥΤΟΜΑΤΗ ΤΑΞΙΝΟΜΗΣΗ 25
Μέτρα συσχέτισης 26
Μέθοδοι ταξινόμησης 30
Η υπόθεση ομαδοποίησης 32
Η χρήση της ομαδοποίησης στην ανάκληση πληροφοριών 34
Single-Link 41
Η καταλληλότητα των διαστρωματωμένων ιεραρχικών μεθόδων cluster 43
Single-link και το ελάχιστο spanning δέντρο 43

ΚΕΦΑΛΑΙΟ 4 45

ΛΟΓΙΚΗ Ή ΦΥΣΙΚΗ ΟΡΓΑΝΩΣΗ ΚΑΙ ΑΝΕΞΑΡΤΗΣΙΑ ΔΕΔΟΜΕΝΩΝ 45
Μία γλώσσα για την περιγραφή των δομών αρχείου 46
Βασική ορολογία 46
Διαδοχικά αρχεία 48
Αντιστρεφόμενα αρχεία 49
Ακολουθιακά αρχεία δεικτών 50
Πολλαπλές λίστες 52
Κυψελικές πολλαπλές λίστες 53
Δομές δακτυλίων 54

Νηματοειδείς λίστες 56
Δέντρα 59

ΚΕΦΑΛΑΙΟ 5 62

ΣΤΡΑΤΗΓΙΚΕΣ ΑΝΑΖΗΤΗΣΗΣ 62
Λογική αναζήτηση 62
Συναρτήσεις ταυτοποίησης 64
Σειριακή αναζήτηση 65
Αναπαραστάσεις ομάδων 65
Ανάκτηση βασισμένη σε ομαδοποίηση 69
Διατύπωση αμφίδρομης αναζήτησης 70
Ανατροφοδότηση 70

ΚΕΦΑΛΑΙΟ 6 74

ΠΙΘΑΝΟΛΟΓΙΚΗ ΑΝΑΚΤΗΣΗ 74
Εκτίμηση ή υπολογισμός της σχετικότητας 75
Βασικό πιθανολογικό πρότυπο 76
Μορφή λειτουργίας ανάκτησης 79
Οι όροι δεικτών δεν είναι ανεξάρτητοι 82
Επιλογή των καλύτερων δέντρων εξάρτησης 85
Εκτίμηση των παραμέτρων 87

ΚΕΦΑΛΑΙΟ 7 89

ΑΞΙΟΛΟΓΗΣΗ 89
Σχετικότητα 91
Ακρίβεια και ανάκληση 92
Υπολογισμός μέσου όρου των τεχνικών 94
Παρεμβολή 97
Σύνθετα μέτρα 98
Το πρότυπο Swets 99
Το πρότυπο Robertson – ο λογικός μετασχηματισμός 101
Το πρότυπο Cooper – αναμενόμενο μήκος αναζήτησης 102
Τα μέτρα του Smart 106
Μια κανονικοποιημένη συμμετρική διαφορά 109
Το πρότυπο 110
Παρουσίαση των πειραματικών αποτελεσμάτων 113
Δοκιμές σημασίας 115

ΚΕΦΑΛΑΙΟ 8 117

ΤΟ ΜΕΛΛΟΝ 117
Μελλοντική έρευνα 117
Αυτόματη ταξινόμηση 117
Δομές αρχείων 118
Στρατηγικές αναζήτησης 119
Προσομοίωση 119

Αξιολόγηση 119
Ανάλυση περιεχομένου 120
Μελλοντικές εξελίξεις 121

ΒΙΒΛΙΟΓΡΑΦΙΑ 123

1. ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

ΕΙΣΑΓΩΓΗ

Η ανάκτηση πληροφορίας είναι ένας ευρύς και συχνά αόριστα καθορισμένος όρος αλλά παρακάτω θα ασχοληθούμε μόνο με τα αυτόματα συστήματα ανάκτησης πληροφορίας. Το αυτόματα ως αντίθετο του χειρονακτικά και η πληροφορία ως αντίθετο του δεδομένου ή του γεγονότος. Δυστυχώς η λέξη πληροφορία μπορεί να γίνει πολύ παραπλανητική. Στο ευρύτερο πλαίσιο της ανάκτησης πληροφορίας, η πληροφορία, με την τεχνική έννοια, δεν είναι εύκολα μετρήσιμη. Μάλιστα, σε πολλές περιπτώσεις κάποιος μπορεί να περιγράψει επαρκώς το είδος της ανάκτησης αντικαθιστώντας απλά τον όρο «πληροφορία» με τον όρο «έγγραφο». Ένα σύστημα ανάκτησης πληροφορίας δεν ενημερώνει τον χρήστη πάνω στο θέμα της έρευνάς του. Απλά πληροφορεί για την ύπαρξη (ή την ανυπαρξία) και την τοποθεσία εγγράφων που σχετίζονται με το αίτημά του. Αυτό ειδικά αποκλείει τα συστήματα Ερώτησης-Απάντησης. Επίσης, αποκλείει τα συστήματα ανάκτησης δεδομένων. Για να κάνουμε σαφή τη διαφορά μεταξύ της ανάκτησης δεδομένων και της ανάκτησης πληροφοριών παρουσιάζουμε στον πίνακα 1 μερικές από τις πιο χαρακτηριστικές ιδιότητες της ανάκτησης δεδομένων και πληροφοριών.

Πίνακας 1: ΑΝΑΚΤΗΣΗ ΔΕΔΟΜΕΝΩΝ Ή ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ

ΚΡΙΤΗΡΙΟ	ΑΝΑΚΤΗΣΗ ΔΕΔΟΜ.	ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦ.
Είδος ταιριάσματος	Ακριβές ταίριασμα	Μερικό ταίριασμα
Πόρισμα	Συμπερασματικό	Επαγωγικό
Πρότυπο	Ντετερμινιστικό	Πιθανοκρατικό
Ταξινόμηση	Μονοθεματική	Πολυθεματική
Γλώσσα ερωτήματος	Εξειδικευμένη	Φυσική
Ορισμός ερωτήματος	Πλήρης	Ατελής
Στοιχεία που ζητούνται	Ταιριαστά	Σχετικά
Ανταπόκριση λάθους	Ευαίσθητη	Όχι ευαίσθητη

Κάποιος μπορεί να θελήσει να κατακρίνει αυτή τη διχοτόμηση με βάση το ότι η διαχωριστική γραμμή μεταξύ των δύο είναι ασαφής. Και έτσι είναι, αλλά είναι χρήσιμο να διευκρινιστεί το εύρος της περιπλοκής σε σχέση με τον κάθε ένα τρόπο ανάκτησης. Ας πάρουμε κάθε στοιχείο του πίνακα και ας το εξετάσουμε πιο προσεκτικά. Στην ανάκτηση δεδομένων κανονικά ψάχνουμε για ένα ακριβές ταίριασμα, δηλαδή, ελέγχουμε για να δούμε εάν ένα στοιχείο εμφανίζεται ή όχι στο αρχείο. Στην ανάκτηση πληροφοριών αυτό μπορεί μερικές φορές να ενδιαφέρει αλλά γενικότερα θέλουμε να βρούμε εκείνα τα στοιχεία τα οποία μερικώς ταιριάζουν με το αίτημα και έπειτα να διαλέξουμε από αυτά εκείνα που ταιριάζουν καλύτερα. Το πόρισμα που χρησιμοποιείται στην ανάκτηση δεδομένων είναι του απλού

συμπερασματικού είδους, δηλαδή aRb και bRc τότε aRc. Στην ανάκτηση πληροφοριών είναι πιο συνηθισμένη η χρήση επαγωγικού πορίσματος, που σημαίνει ότι οι σχέσεις ορίζονται με ένα βαθμό βεβαιότητας ή αβεβαιότητας και γι' αυτό το λόγο η εμπιστοσύνη μας όσο αφορά το πόρισμα είναι μεταβλητή. Αυτή η αντίθεση οδηγεί κάποιον στο να χαρακτηρίσει την ανάκτηση δεδομένων ως αιτιοκρατική (ντετερμινιστική) και την ανάκτηση πληροφοριών ως πιθανοκρατική.
Συχνά στην ανάκτηση πληροφοριών απαιτείται το θεώρημα του Bayes για να εξαγάγει τα συμπεράσματα, αλλά στην ανάκτηση δεδομένων οι πιθανότητες δεν εισάγονται στην επεξεργασία. Άλλος ένας διαχωρισμός μπορεί να γίνει στους όρους ταξινομήσεων οι οποίες πιθανόν να αποδειχθούν χρήσιμες. Στην ανάκτηση δεδομένων ενδιαφερόμαστε περισσότερο για μία **μονοθετική** ταξινόμηση, δηλαδή, τα αντικείμενα θα πρέπει να πληρούν όλες τις αναγκαίες ιδιότητες που καθορίζονται από τα στοιχεία μιας κατηγορίας για να ανήκουν σε αυτήν. Στην ανάκτηση πληροφοριών μία τέτοιου είδους ταξινόμηση δεν είναι καθόλου χρήσιμη και στην πραγματικότητα αυτό που πιο συχνά απαιτείται είναι μία **πολυθετική** ταξινόμηση. Σε μία τέτοια ταξινόμηση κάθε στοιχείο μιας κατηγορίας θα πληροί μόνο ένα μέρος του συνόλου των ιδιοτήτων όλων των μελών της κατηγορίας αυτής. Γι' αυτό το λόγο καμία ιδιότητα δεν είναι αναγκαία ή αρκετή για συμμετοχή σε μία κατηγορία. Η γλώσσα υποβολής ερωτημάτων για την ανάκτηση δεδομένων θα είναι γενικά εξεζητημένης φύσεως, με περιορισμένο συντακτικό και λεξιλόγιο, ενώ στην ανάκτηση πληροφοριών προτιμούμε να χρησιμοποιούμε φυσική γλώσσα αν και υπάρχουν μερικές αξιοσημείωτες εξαιρέσεις. Στην ανάκτηση δεδομένων το ερώτημα είναι γενικά ένας πλήρης προσδιορισμός του τι ζητείται, ενώ στην ανάκτηση πληροφοριών είναι αμετάβλητα ατελής. Αυτή η τελευταία διαφορά προκύπτει μερικώς από το γεγονός ότι στην ανάκτηση πληροφοριών ψάχνουμε για σχετικά έγγραφα και όχι για στοιχεία που να ταιριάζουν απόλυτα. Ο βαθμός ταιριάσματος στην ανάκτηση πληροφοριών υποδεικνύει την πιθανότητα σχετικότητας των στοιχείων. Μία απλή επίπτωση αυτής της διαφοράς είναι ότι η ανάκτηση δεδομένων είναι πιο ευαίσθητη σε λάθη, με την έννοια ότι, εάν βρεθεί ένα λάθος στο ταίριασμα δεν θα ανακτήσει το ζητούμενο στοιχείο γεγονός το οποίο συνεπάγεται την ολοκληρωτική αποτυχία του συστήματος. Στην ανάκτηση πληροφοριών μικρά λάθη στο ταίριασμα γενικά δεν επηρεάζουν ουσιαστικά την απόδοση του συστήματος. **Πολλά αυτόματα συστήματα ανάκτησης πληροφοριών είναι πειραματικά.** Η πειραματική ανάκτηση πληροφοριών βασικά διεκπεραιώνεται σε ένα «εργαστήριο» ενώ τα λειτουργικά συστήματα είναι εμπορικά συστήματα τα οποία χρεώνουν για τις υπηρεσίες που προσφέρουν. Φυσικά τα δύο συστήματα αξιολογούνται διαφορετικά. Τα εμπορικά συστήματα ανάκτησης πληροφοριών αξιολογούνται βάση της ικανοποίησης του χρήστη και της τιμής που ο χρήστης είναι πρόθυμος να πληρώσει για τις υπηρεσίες τους. Τα πειραματικά συστήματα ανάκτησης πληροφοριών αξιολογούνται συγκρίνοντας τα πειράματα ανάκτησης με πρότυπα ειδικά κατασκευασμένα γι' αυτό το σκοπό.

Ανάκτηση πληροφοριών

Από τη δεκαετία του 1940 το πρόβλημα της αποθήκευσης και ανάκτησης πληροφοριών έχει τύχει αυξανόμενης προσοχής. Μία επίδραση αυτού είναι ότι οι σχετικές πληροφορίες αγνοούνται καθώς ποτέ δεν αποκαλύπτονται, γεγονός που με τη

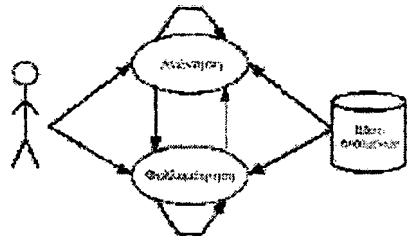
σειρά του οδηγεί στον διπλασιασμό μελέτης και προσπάθειας. Είναι απλώς μία κατάσταση: έχουμε τεράστια σύνολα πληροφοριών στα οποία η ακρίβεια και η ταχύτητα προσπέλασης γίνεται όλο και πιο δύσκολη. Με την εμφάνιση των υπολογιστών, δόθηκε μεγάλο βάρος στην ιδέα χρησιμοποίησής τους ώστε να παρέχουν γρήγορα και έξυπνα συστήματα ανάκτησης. Στις βιβλιοθήκες, πολλές από τις οποίες σαφώς έχουν πρόβλημα αποθήκευσης και ανάκτησης πληροφοριών, μερικές από τις πιο κοινότυπες εργασίες, όπως η κατηγοριοποίηση και η γενική διαχείριση, διεκπεραιώνονται επιτυχώς μέσω υπολογιστών. .Ωστόσο, το πρόβλημα της αποτελεσματικής ανάκτησης παραμένει σε μεγάλο βαθμό άλυτο. Σε γενικές γραμμές, η αποθήκευση και η ανάκτηση πληροφοριών είναι απλή. Υποθέστε ότι υπάρχει ένας σωρός από έγγραφα και ότι ένα άτομο (χρήστης του σωρού) διατυπώνει μία ερώτηση στην οποία η απάντηση είναι ένα σύνολο εγγράφων που ικανοποιεί τις πληροφορίες που χρειάζεται να διατυπωθούν από την ερώτησή του. Μπορεί επίσης, να αποκτήσει αυτό το σύνολο διαβάζοντας όλα τα έγγραφα του σωρού, κρατώντας τα σχετικά έγγραφα και απορρίπτοντας όλα τα άλλα. Κατά μία άποψη, αυτός ο τρόπος συνιστά «τέλεια» ανάκτηση. Αλλά αυτή η λύση είναι προφανώς ανεφάρμοστη. Και αυτό γιατί, ένας χρήστης είτε δεν έχει χρόνο είτε δεν επιθυμεί να περάσει το χρόνο του διαβάζοντας όλη τη συλλογή εγγράφων, πέρα από το γεγονός ότι μπορεί να είναι φυσικά αδύνατο γι' αυτόν να το κάνει. Όταν οι υπολογιστές υψηλής ταχύτητας έγιναν διαθέσιμοι όχι μόνο για υπολογιστική εργασία, πολλοί ήταν αυτοί που σκέφτηκαν ότι ένας υπολογιστής μπορεί να είναι ικανός να «διαβάσει» μία ολόκληρη συλλογή από έγγραφα και να αποσπάσει τα σχετικά έγγραφα. Σύντομα έγινε φανερό ότι η χρησιμοποίηση φυσικής γλώσσας στο κείμενο των εγγράφων δεν προκαλούσε μόνο προβλήματα εισαγωγής και αποθήκευσης (κάτι που ισχύει και σήμερα) αλλά επίσης αφήνει άλυτο το πρόβλημα του χαρακτηρισμού του περιεχομένου του εγγράφου. Είναι κατανοητό ότι οι μελλοντικές εξελίξεις στο υλικό μπορεί να κάνουν την είσοδο και την αποθήκευση της φυσικής γλώσσας πιο εφικτή. Αλλά ο αυτόματος χαρακτηρισμός μέσω του οποίου το λογισμικό επιχειρεί να αντιγράψει την ανθρώπινη διαδικασία «ανάγνωσης» είναι ωστόσο ένα πολύ δύσκολο πρόβλημα. Πιο συγκεκριμένα, αυτή η ανάγνωση εμπεριέχει προσπάθεια απόσπασης, συντακτικών και σημασιολογικών, πληροφοριών από το κείμενο και χρησιμοποίησή τους για να αποφανθούμε εάν κάθε έγγραφο είναι σχετικό ή όχι με ένα συγκεκριμένο αίτημα. Η δυσκολία λοιπόν, δεν είναι μόνο η γνώση στο πώς να αποσπάσουμε την πληροφορία αλλά επίσης και στο πώς να τη χρησιμοποιήσουμε για να αποφανθούμε για τη σχετικότητα. Η συγκριτικά αργή εξέλιξη της σύγχρονης γλωσσολογίας στο σημασιολογικό μέρος και η ευδιάκριτη αποτυχία της μηχανής μετάφρασης δείχνουν ότι αυτά τα προβλήματα είναι σε μεγάλο βαθμό άλυτα. Ήδη θα έχουμε παρατηρήσει ότι έχει εμπλακεί η ιδέα της «σχετικότητας». Αυτή η ιδέα βρίσκεται στο επίκεντρο της ανάκτησης πληροφοριών. Ο σκοπός μίας στρατηγικής αυτόματης ανάκτησης είναι να ανακτήσει όλα τα σχετικά έγγραφα ανακτώντας ταυτόχρονα όσο το δυνατόν λιγότερα μη σχετικά. Η στιγμή που διεκπεραιώνεται ο χαρακτηρισμός ενός εγγράφου, πρέπει να είναι τέτοια ώστε όταν το έγγραφο εμφανίζεται να είναι σχετικό σε ένα ερώτημα, να είναι εφικτή η ανάκτηση του εγγράφου ως απάντηση σε εκείνο το ερώτημα. Διανοητικώς είναι δυνατό για έναν άνθρωπο να αποδείξει τη σχετικότητα ενός εγγράφου ως προς ένα ερώτημα. Για να το κάνει αυτό ένας υπολογιστής χρειάζεται να κατασκευάσουμε ένα πρότυπο μέσα στο οποίο οι αποφάσεις σχετικότητας μπορούν να εκφραστούν με ποσοτικό τρόπο.

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 ανάκτησης πληροφορίας να ανταποκριθεί στην πληροφοριακή ανάγκη του χρήστη, θα πρέπει να είναι σε θέση, να «διερμηνεύσει» με κάποιον τρόπο το σημασιολογικό περιεχόμενο το αντικειμένων 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 μπορούμε να χρησιμοποιήσουμε την εξαγόμενη πληροφορία για να αποφασίσουμε τη σχετικότητα προς κάποιο ερώτημα. **Ο κύριος στόχος άλλωστε ενός συστήματος ανάκτησης πληροφορίας είναι να μπορεί να επιστρέψει όλα τα κείμενα που είναι σχετικά προς κάποιο ερώτημα, επιστρέφοντας παράλληλα και όσο το δυνατόν λιγότερα μη σχετικά κείμενα.** Γι' αυτό το λόγο η έννοια της **σχετικότητας,** διαδραματίζει κυρίαρχο ρόλο στην ανάκτηση πληροφορίας.

Η διαδικασία χρήστη

Η αποδοτική ανάκτηση κειμένων επηρεάζεται τόσο από την διαδικασία χρήστη όσο και από την λογική αναπαράσταση των κειμένων, όπως αυτή υιοθετείται από το σύστημα. Τις δύο αυτές παραμέτρους θα συζητήσουμε ευθύς αμέσως.
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, είναι ο καθορισμός 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 δει σε προηγούμενη ενότητα την διαδικασία της αναζήτησης ας εξετάσουμε μία δεύτερη διαδικασία ανάκτησης, τη ***φυλλομέτρηση*** (browsing). Έστω ότι το ενδιαφέρον του χρήστη είτε δεν είναι καλά ορισμένο είτε καλύπτει ένα αρκετά ευρύ φάσμα πληροφορίας. 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 μία 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 μία διαδικασία ανάκτησης πληροφορίας, της οποίας όμως οι σκοποί δεν είναι ξεκάθαρα προσδιορισμένοι τη στιγμή της εκκίνησης και που

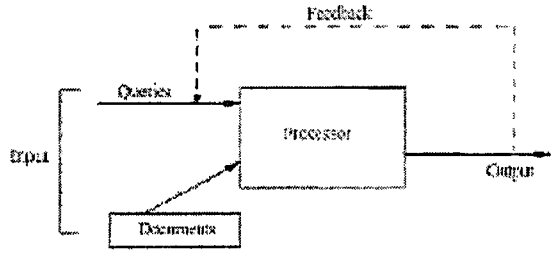
μπορεί να μεταβληθούν κατά τη διάρκεια της αλληλεπίδρασης με το σύστημα.



Σχήμα 1: Αλληλεπίδραση του χρήστη με το σύστημα ανάκτησης πληροφορίας

Ένα σύστημα ανάκτησης πληροφορίας

Το παρακάτω διάγραμμα δείχνει τρία συστατικά: μία είσοδο, έναν επεξεργαστή και μία έξοδο.



Σχήμα 2: Ένα σύστημα ανάκτησης πληροφορίας

Μία τέτοιου είδους διχοτομία μπορεί να φαίνεται λίγο κοινότυπη, άλλα το κάθε συστατικό της μας δίνει πρόσφορο έδαφος για συζήτηση. Ας ξεκινήσουμε από την είσοδο. Εδώ το κύριο πρόβλημα είναι να εξασφαλίσουμε μία απεικόνιση από κάθε έγγραφο και ερώτημα κατάλληλη για επεξεργασία από έναν υπολογιστή. Στο σημείο αυτό αξίζει να σημειωθεί ότι τα περισσότερα συστήματα ανάκτησης που βασίζονται σε υπολογιστή αποθηκεύουν μόνο μία απεικόνιση του εγγράφου (ή του ερωτήματος) πράγμα το οποίο σημαίνει ότι το κείμενο ενός εγγράφου χάνεται από τη στιγμή που θα επεξεργαστεί με σκοπό τη δημιουργία της απεικόνισής του. Παραδείγματος χάριν, η απεικόνιση ενός εγγράφου μπορεί να είναι μία λίστα απασπασματικών λέξεων που θεωρήθηκαν σημαντικές. Αντί να πρέπει ο υπολογιστής να επεξεργαστεί τη φυσική γλώσσα, μία εναλλακτική προσέγγιση, είναι να πρέπει να επεξεργαστεί μία εξεζητημένη γλώσσα στην οποία όλα τα ερωτήματα και τα έγγραφα μπορούν τα τυποποιηθούν. Βέβαια απαραίτητη προϋπόθεση θεωρείται ότι ο χρήστης είναι πρόθυμος να μάθει να εκφράζει την πληροφορία που χρειάζεται σε αυτήν τη γλώσσα. Όταν το σύστημα ανάκτησης είναι on-line, είναι δυνατό για τον χρήστη να αλλάξει το αίτημά του κατά τη διάρκεια μιας περιόδου αναζήτησης με στόχο ένα δείγμα ανάκτησης, έτσι προσδοκάται, βελτίωση της επόμενης εκτέλεσης ανάκτησης. Μία

τέτοια διαδικασία συχνά καλείται ανατροφοδότηση. Ένα παράδειγμα ενός εξεζητημένου on-line συστήματος ανάκτησης είναι το MEDLINE. Δεύτερο συστατικό, ο επεξεργαστής, αυτό το τμήμα του συστήματος ανάκτησης εμπλέκεται στη διαδικασία ανάκτησης. Η διαδικασία αυτή μπορεί να εμπεριέχει διάρθρωση της πληροφορίας με έναν κατάλληλο τρόπο, όπως είναι η ταξινόμηση. Μπορεί, επίσης, να εμπεριέχει εκτέλεση της πραγματικής λειτουργίας, η οποία είναι η εκτέλεση της στρατηγικής αναζήτησης ως απάντηση σε ένα ερώτημα. Στο διάγραμμα, τα έγγραφα έχουν τοποθετηθεί σε ένα ξεχωριστό πλαίσιο για να δοθεί έμφασηστο γεγονός ότι δεν είναι απλή είσοδος αλλά μπορούν να χρησιμοποιηθούν κατά τη διάρκεια της διαδικασίας ανάκτησης με την έννοια ότι η δομή τους θεωρείται πιο σωστή καθώς παίρνουν μέρος στη διαδικασία ανάκτησης. Τέλος, ερχόμαστε στην έξοδο, η οποία συνήθως είναι ένα σύνολο από παραπομπές και έγγραφα. Σε ένα λειτουργικό σύστημα η ιστορία τελειώνει εδώ. Ωστόσο, σε ένα πειραματικό σύστημα απομένει να γίνει η αξιολόγηση.

Η προοπτική της Ανάκτησης Πληροφορίας (IR in perspective)

Αν και η ανάκτηση πληροφοριών μπορεί να υποδιαιρεθεί με πολλούς τρόπους, φαίνεται ότι υπάρχουν **τρεις βασικοί τομείς έρευνας** μεταξύ των οποίων επιτυγχάνεται ένα αξιοσημείωτο τμήμα του θέματος. Οι τομείς αυτοί είναι οι εξής: **ανάλυση περιεχομένου, δομή πληροφορίας και αξιολόγηση**. Περιληπτικά, ο πρώτος τομέας ασχολείται με την απεικόνιση των περιεχομένων των εγγράφων σε μία μορφή που να είναι κατάλληλη για επεξεργασία από υπολογιστή, ο δεύτερος τομέας ασχολείται με την αξιοποίηση των σχέσεων μεταξύ των εγγράφων για να βελτιώσει την απόδοση και την αποτελεσματικότητα των στρατηγικών ανάκτησης και τέλος, ο τρίτος τομέας ασχολείται με τη μέτρηση της απόδοσης της ανάκτησης.
Σ΄ αυτό το σημείο, ίσως να είναι βολικό να περιγράψουμε λεπτομερώς τη χρήση της **λέξης κλειδί**. Έχει γίνει κοινή συνήθεια στη βιβλιογραφία της ανάκτησης πληροφοριών να αναφερόμαστε σε περιγραφικά στοιχεία που εξάγονται από κείμενο ως λέξεις κλειδιά ή όροι. Συχνά, τέτοια στοιχεία είναι το αποτέλεσμα κάποιας διαδικασίας όπως, για παράδειγμα, η ταυτόχρονη συγκέντρωση των διαφορετικών μορφολογικών παραλλαγών της ίδιας λέξης. Ο όρος δομή πληροφορίας (προς επιζήτηση καλύτερων λέξεων) καλύπτει συγκεκριμένα μία λογική οργάνωση της πληροφορίας, όπως τα αντιπροσωπευτικά έγγραφα, με σκοπό την ανάκτηση πληροφορίας. Η ανάπτυξη των δομών πληροφορίας είναι σχετικά πρόσφατη. Ο κύριος λόγος για την αργή ανάπτυξη σε αυτόν τον τομέα της ανάκτησης πληροφοριών είναι ότι για μεγάλο χρονικό διάστημα κανείς δεν συνειδητοποίησε ότι οι υπολογιστές δεν θα έδιναν έναν αποδεκτό χρόνο ανάκτησης με ένα μεγάλο σύνολο εγγράφων αν δεν επιβαλλόταν μία λογική δομή σε αυτό. Στην πραγματικότητα, οι ιδιοκτήτες μεγάλων βάσεων δεδομένων είναι ακόμη απρόθυμοι να δοκιμάσουν νέες τεχνικές οργάνωσης που υπόσχονται γρηγορότερη και καλύτερη ανάκτηση. Η καθυστέρηση στην αναγνώριση και υιοθέτηση νέων τεχνικών οφείλεται κυρίως στην έλλειψη πειραματικών αποδείξεων που να τις υποστηρίζουν. Προηγούμενα πειράματα με συστήματα ανάκτησης εγγράφων συχνά υιοθετούσαν μία σειριακή οργάνωση αρχείου η οποία, αν και ήταν αποτελεσματική όταν εκτελούνταν ένας αρκετά μεγάλος αριθμός από ερωτήματα ταυτόχρονα μέσα σε μία λειτουργία ομαδικής επεξεργασίας,

αποδείχτηκε ανεπαρκής όταν το κάθε ερώτημα απαιτούσε απάντηση σε σύντομο πραγματικό χρόνο. Η δημοφιλής οργάνωση που υιοθετήθηκε στη θέση της ήταν αυτή του αντιστρεφόμενου αρχείου (inverted file). Αυτή η οργάνωση θεωρήθηκε από πολλούς περιοριστική. Πιο πρόσφατα πειράματα προσπάθησαν να αποδείξουν την ανωτερότητα των ομαδοποιημένων (clustered) αρχείων στην on-line ανάκτηση. Η οργάνωση αυτών των αρχείων δημιουργείται από μία μέθοδο αυτόματης ταξινόμησης. Η αξιολόγηση των συστημάτων ανάκτησης έχει αποδειχθεί εξαιρετικά δύσκολη. Παρά τα πολυάριθμα μεγέθη αποδοτικότητας που έχουν προταθεί για μία γενική θεωρία αξιολόγησης αυτή δεν προέκυψε. Σήμερα η αποδοτικότητα ανάκτησης ακόμα επί το πλείστον μετριέται με όρους ακρίβειας και ανάκλησης ή με μεγέθη που βασίζονται σε αυτούς. Ακόμα δεν υπάρχει επαρκής στατιστική συμπεριφορά που να δείχνει πως μπορούν να χρησιμοποιηθούν κατάλληλα τεστ σημαντικότητας. Έτσι, έπειτα από μερικές δεκαετίες έρευνας σε αυτόν τον τομέα βασικά έχουμε μόνο ακρίβεια και ανάκληση και έναν αριθμό υποθετικών λειτουργιών.

Αποτελεσματικότητα και αποδοτικότητα

Μεγάλο ποσοστό της έρευνας και της ανάπτυξης στην ανάκτηση πληροφορίας σκοπεύει στη βελτίωση της αποτελεσματικότητας και της αποδοτικότητας της ανάκτησης. Η αποδοτικότητα μετριέται συνήθως από την άποψη των πόρων υπολογιστών που χρησιμοποιούνται όπως είναι ο πυρήνας, η βοηθητική μνήμη και ο χρόνος C.P.U.. Είναι δύσκολο να μετρηθεί η αποδοτικότητα με έναν ανεξάρτητο μηχανικό τρόπο. Σε οποιαδήποτε περίπτωση, η αποτελεσματικότητα πρέπει να μετριέται σε σχέση με την αποδοτικότητα για να έχουμε κάποια ιδέα για το όφελος από την άποψη της μονάδας κόστους. Σημειώνουμε ότι ακρίβεια είναι ο λόγος του αριθμού των σχετικών εγγράφων που ανακτήθηκαν προς τον συνολικό αριθμό εγγράφων που ανακτήθηκαν και ανάκληση είναι ο λόγος του αριθμού των σχετικών εγγράφων που ανακτήθηκαν προς τον συνολικό αριθμό των σχετικών εγγράφων (αυτών που ανακτήθηκαν και αυτών που δεν ανακτήθηκαν).

2. Αυτόματη Ανάλυση Κειμένων

Πριν το σύστημα ανάκτησης αυτοματοποιημένων πληροφοριών μπορέσει πραγματικά να λειτουργήσει για να ανακτήσει κάποια πληροφορία, αυτή η πληροφορία πρέπει να έχει ήδη καταχωρηθεί μέσα στον υπολογιστή. Αρχικά θα μπορούσε να είναι σε μορφή εγγράφου. Ο υπολογιστής, ωστόσο, δεν είναι πιθανό να έχει καταχωρήσει το πλήρες κείμενο κάθε εγγράφου στη φυσική γλώσσα στην οποία είχε γραφεί. Αντίθετα θα έχει, ένα αντιπροσωπευτικό έγγραφο, που μπορεί να έχει παραχθεί από τα έγγραφα είτε χειροκίνητα είτε αυτόματα.
Η αφετηρία της διαδικασίας ανάλυσης κειμένων μπορεί να είναι το πλήρες κείμενο του εγγράφου, μια περίληψη, ο τίτλος μόνο, ή ίσως μόνο ένας κατάλογος λέξεων. Από αυτό, η διαδικασία πρέπει να παράγει ένα έγγραφο αντιπροσωπευτικό σε μια μορφή που ο υπολογιστής μπορεί να χειριστεί.
Οι εξελίξεις και οι πρόοδοι στη διαδικασία της αντιπροσώπευσης ανασκοπούνται κάθε έτος από τα σχετικά κεφάλαια του Annual Review of Information Science and Technology του Cuadra. Σε αυτό το κεφάλαιο δίνεται έμφαση στην στατιστική παρά

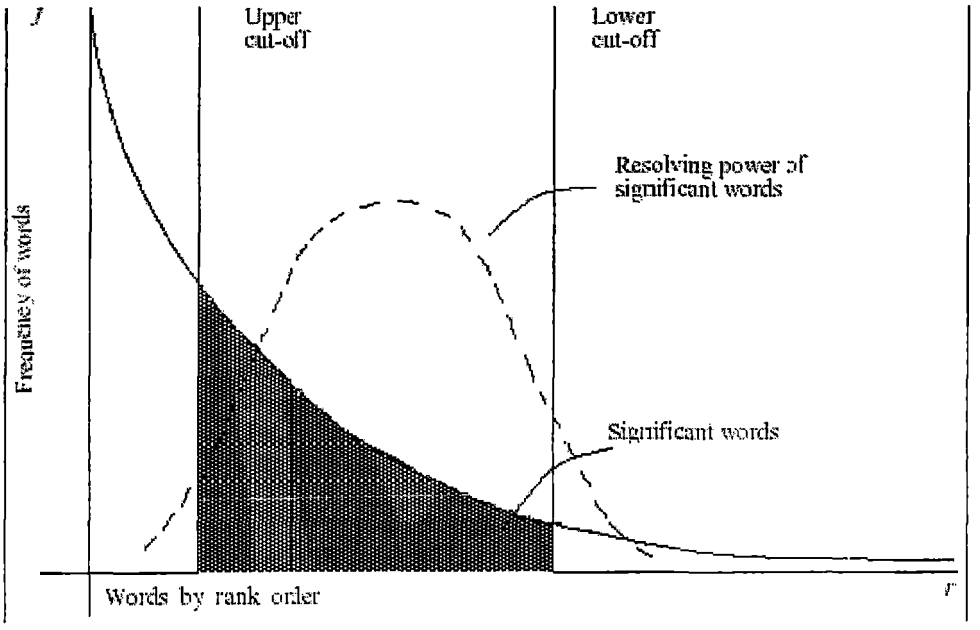
στις γλωσσικές προσεγγίσεις στην αυτόματη ανάλυση κειμένων. Οι λόγοι για αυτήν την έμφαση είναι ποικίλοι. Αναμφισβήτητα, μια θεωρία της γλώσσας θα είναι ακραίας σπουδαιότητας στην ανάπτυξη των ευφυών συστημάτων ανάκτησης πληροφορίας. Αλλά, μέχρι σήμερα, καμιά τέτοια θεωρία δεν έχει αναπτυχθεί αρκετά για να μπορεί να εφαρμοσθεί επιτυχώς στην ανάκτηση πληροφορίας. Εν πάση περιπτώσει, τα συστήματα ανάκτησης εγγράφων μπορούν να κατασκευαστούν ικανοποιητικά χωρίς μια τέτοια θεωρία. Η στατιστική προσέγγιση έχει εξεταστεί και έχει βρεθεί να είναι σχετικά επιτυχής.
Το κεφάλαιο ξεκινά με τις αρχικές ιδέες του Luhn στις οποίες ένα μεγάλο μέρος της αυτόματης ανάλυσης κειμένων έχει στηριχτεί και συνεχίζει έπειτα με την περιγραφή ενός συγκεκριμένου τρόπου παραγωγής αντιπροσώπων εγγράφων. Επιπλέον, αναφέρονται τρόποι αξιοποίησης και βελτίωσης των αντιπροσώπων εγγράφων μέσω της στάθμισης ή της ταξινόμησης των λέξεων κλειδιών. Κάποια στοιχεία παρουσιάζονται για την αυτόματη ευρετηρίαση.

Ιδέες του Luhn

Σε ένα από τα αρχικά έγγραφα δηλώνει: "Εδώ προτείνεται ότι η συχνότητα μιας λέξης σε ένα άρθρο παρέχει μια χρήσιμη διάσταση της σημασίας της λέξης. Περαιτέρω προτείνεται ότι η σχετική θέση των λέξεων, μέσα σε μια πρόταση, που έχουν τιμές σημασίας, παρέχει μια χρήσιμη μέτρηση για τον καθορισμό της σημασίας των προτάσεων. Ο παράγοντας σημασίας της πρότασης επομένως θα βασιστεί σε έναν συνδυασμό αυτών των δύο μετρήσεων."
Αυτό το απόσπασμα συνοψίζει την συνεισφορά του Luhn1 στην αυτόματη ανάλυση κειμένων. Η υπόθεσή του είναι ότι συχνά τα δεδομένα συχνότητας μπορούν να χρησιμοποιηθούν για να εξαγάγουν λέξεις και προτάσεις για να αναπαραστήσουν ένα έγγραφο.

Ας υποθέσουμε ότι f είναι η συχνότητα εμφάνισης των διάφορων τύπων λέξης σε μια δεδομένη θέση του κειμένου και r είναι η κατάταξη(rank order), δηλαδή η κατάταξη της συχνότητας εμφάνισής τους. Μια γραφική παράσταση που σχετίζει το f με το r δίνει μια καμπύλη παρόμοια με την υπερβολή στο σχήμα 3. Αυτή είναι στην πραγματικότητα μια καμπύλη που περιγράφει το νόμο του Zipf2, ο οποίος δηλώνει ότι το προϊόν της συχνότητας της χρήσης των λέξεων και της κατάταξης είναι περίπου σταθερό. Ο Zipf έχει αποδείξει το νόμο του για τα αμερικάνικα αγγλικά. Ο Luhn το χρησιμοποίησε ως άκυρη υπόθεση για να προσδιορίσει δύο οριακά σημεία, το ανώτερο και το κατώτερο (βλέπε σχήμα 3), κατά συνέπεια αποκλείοντας τις ασήμαντες λέξεις. Οι λέξεις που υπερβαίνουν το ανώτερο οριακό σημείο θεωρήθηκαν κοινές και εκείνες που ήταν χαμηλότερα από το κατώτερο οριακό σημείο σπάνιες και επομένως δεν συμβάλλουν σημαντικά στο περιεχόμενο του άρθρου. Επινόησε έτσι μια ολοκληρωμένη τεχνική αρίθμησης για την εύρεση σημαντικών λέξεων. Σύμφωνα με αυτή υπέθεσε ότι η αναλυτική δύναμη των σημαντικών λέξεων, εννοώντας την ικανότητα των λέξεων να διακρίνουν ως προς το περιεχόμενο, έφθασε σε μια αιχμή σε μια κατάταξη θέσης στη μέση μεταξύ των δύο οριακών σημείων και από την αιχμή μειωνόταν προς κάθε κατεύθυνση σε σχεδόν μηδέν στα οριακά σημεία. Μια ορισμένη αυθαιρεσία περιλαμβάνεται στον καθορισμό των οριακών σημείων. Πρέπει να

καθιερωθούν μέσω της δοκιμής και του σφάλματος.
Είναι ενδιαφέρον ότι αυτές οι ιδέες είναι πραγματικά βασικές σε ένα μεγάλο μέρος της πιο πρόσφατης εργασίας στην ανάκτηση πληροφορίας, ο Luhn ο ίδιος τις χρησιμοποίησε για να επινοήσει μια μέθοδο αυτόματης περίληψης.
Κατόπιν προσχώρησε στην ανάπτυξη ενός αριθμητικού μέτρου σημασίας για τις προτάσεις, βασισμένο στον αριθμό σημαντικών και ασήμαντων λέξεων σε κάθε τμήμα της πρότασης. Οι προτάσεις ταξινομούνταν σύμφωνα με το αριθμητικό αποτάλεσμα τους και η υψηλότερη ταξινόμηση περιλήφθηκε στην περίληψη. Οι Edmundson και Wyllys3 έχουν συνεχίσει με τη γενίκευση τμήματος από την εργασία του Luhn με την κανονικοποίηση των μετρήσεών του σε σχέση με τη συχνότητα εμφάνισης των λέξεων Δεν υπάρχει κανένας λόγος για τον οποίο μια τέτοια ανάλυση πρέπει να περιοριστεί μόνο στις λέξεις. Θα μπορούσε εξίσου καλά να εφαρμοστεί στις ρίζες των λέξεων (ή των φράσεων) και στην πραγματικότητα αυτό γίνεται συχνά.



Σχήμα 3: Μια γραφική παράσταση της υπερβολής που σχετίζει το f, συχνότητα εμφάνισης και το r, κατάταξη.

Παραγωγή αντιπροσωπευτικών εγγράφων – conflation

Τελικά θα ήταν επιθυμητή η ανάπτυξη ενός συστήματος επεξεργασίας κειμένου, το οποίο, με τη χρήση υπολογίσιμων μεθόδων και με το ελάχιστο της ανθρώπινης επέμβασης θα παραγάγει από ένα κείμενο εισόδου (πλήρης κείμενο, περίληψη ή τίτλος) ένα αντιπροσωπευτικό έγγραφο επαρκές για χρήση σε ένα αυτόματο σύστημα ανάκτησης. Αυτό είναι κάτι δύσκολο και μπορεί μόνο να πραγματοποιηθεί μερικώς. Το αντιπροσωπευτικό έγγραφο στο οποίο στοχεύουμε, αποτελείται απλά από έναν κατάλογο ονομάτων κλάσης, κάθε όνομα που αντιπροσωπεύει μια κλάση των λέξεων που εμφανίζονται στο συνολικό κείμενο εισόδου.

Ένα έγγραφο θα δεικτοδοτηθεί από ένα όνομα εάν μια από τις σημαντικές λέξεις της εμφανιστεί ως μέλος εκείνης της κλάσης.
Ένα τέτοιο σύστημα θα αποτελείται από τρία μέρη: (1) αφαίρεση των λέξεων υψηλής συχνότητας, (2) αφαίρεση παραγωγικών καταλήξεων, (3) ανίχνευση ισοδύναμων ριζών.
Η αφαίρεση των λέξεων υψηλής συχνότητας, λέξεων «stop» ή λέξεων «fluff»είναι ένας τρόπος να προσεγγιστεί το ανώτερο οριακό σημείο του Luhn. Αυτό γίνεται κανονικά με τη σύγκριση του κειμένου εισόδου με έναν «κατάλογο stop» των λέξεων που πρόκειται να αφαιρεθούν.

Ο πίνακας 2 παραθέτει ένα τμήμα ενός τέτοιου καταλόγου και καταδεικνύει το είδος λέξεων που περιλαμβάνονται. Τα πλεονεκτήματα της διαδικασίας είναι ότι οι ασήμαντες λέξεις αφαιρούνται και επομένως δεν θα παρέμβουν κατά τη διάρκεια της ανάκτησης, αλλά και ότι το μέγεθος του συνολικού αρχείου εγγράφων μπορεί να μειωθεί μεταξύ 30 και 50 τοις εκατό.
Το δεύτερο στάδιο, η αφαίρεση παραγωγικών καταλήξεων, είναι πιο περίπλοκο. Μια πρότυπη προσέγγιση είναι να υπάρξει ένας πλήρης κατάλογος από παραγωγικές καταλήξεις και να αφαιρεθεί η πιθανή μεγαλύτερη.

Ο πίνακας 3 περιέχει μερικές παραγωγικές καταλήξεις. Δυστυχώς, η αφαίρεση χωρίς να ληφθεί υπόψη το περιεχόμενο οδηγεί σε ένα σημαντικό ποσοστό σφάλματος. Παραδείγματος χάριν, έστω ότι θέλουμε το UAL να αφαιρεθεί από το FACTUAL αλλά όχι από το EQUAL. Για να αποφευχθεί η λάθος αφαίρεση παραγωγικών καταλήξεων, κανόνες που λαμβάνουν υπόψη το περιεχόμενο, επινοούνται έτσι ώστε μια παραγωγική κατάληξη να αφαιρεθεί μόνο εάν το περιεχόμενο είναι σωστό. "Σωστό" μπορεί να σημαίνει διάφορα πράγματα:

(1) το μήκος της υπόλοιπης ρίζας υπερβαίνει έναν δεδομένο αριθμό, συνήθως το 2.

(2) η κατάληξη της ρίζας πληροί κάποιους όρους, π.χ. δεν τελειώνει με Q.

Πολλές λέξεις που είναι ισοδύναμες με την ανώτερη έννοια, χαρτογραφούν σε μια Μορφολογική μορφή με την αφαίρεση των παραγωγικών τους καταλήξεων. Κάποιες άλλες, αν και είναι ισοδύναμες δεν το κάνουν. Αυτές απαιτούν ειδική επεξεργασία. Πιθανώς η απλούστερη μέθοδο να αντιμετωπιστεί αυτό, είναι να κατασκευαστεί ένας κατάλογος ισοδύναμων καταλήξεων ρίζας. Για να είναι δύο ρίζες ισοδύναμες πρέπει να ταιριάξουν εκτός από τις καταλήξεις τους, οι οποίες πρέπει να εμφανιστούν στον κατάλογο ως ισοδύναμες. Παραδείγματος χάριν, ρίζες όπως **ABSORB** - και **ABSORPT** - συγχωνεύονται επειδή υπάρχει μια είσοδος στον κατάλογο καθορίζοντας το **B** και **PT** ως ισοδύναμες καταλήξεις ριζών εάν οι προηγούμενοι χαρακτήρες είναι ίδιοι.

Η υπόθεση (στα πλαίσια του IR) είναι ότι εάν δύο λέξεις έχουν την ίδια θεμελιώδη

ρίζα αναφέρονται στην ίδια έννοια και πρέπει να δεικτοδοτηθούν υπό αυτήν τη μορφή. Αυτό είναι προφανώς μια υπεραπλούστευση εφόσον οι λέξεις με την ίδια ρίζα, όπως το NEUTRON και NEUTRALISE, πρέπει μερικές φορές να διαχωριστούν. Ακόμη και οι λέξεις που είναι ουσιαστικά ισοδύναμες μπορούν να σημαίνουν διαφορετικά πράγματα σε διαφορετικό περιεχόμενο. Δεδομένου ότι δεν υπάρχει κανένας φτηνός τρόπος για αυτές τις λεπτές διακρίσεις ανεχόμαστε ένα ορισμένο ποσοστό σφαλμάτων και υποθέτουμε (σωστά) ότι δεν θα υποβιβάσουν την αποτελεσματικότητα ανάκτησης πάρα πολύ.
Είναι αναπόφευκτο ότι ένα σύστημα επεξεργασίας όπως αυτό θα παραγάγει σφάλματα. Ευτυχώς τα πειράματα έχουν δείξει ότι το ποσοστό σφάλματος τείνει να είναι της τάξης του 5 τοις εκατό.

Δεικτοδότηση (Indexing)

Μια γλώσσα δεικτοδότησης είναι η γλώσσα που χρησιμοποιείται για να περιγράψει τα έγγραφα και τα αιτήματα. Τα στοιχεία της γλώσσας δεικτοδότησης είναι όροι δεικτών, που μπορούν να προέλθουν από το κείμενο του εγγράφου που θα περιγραφεί, ή μπορούν να προσεγγιστούν ανεξάρτητα. Οι γλώσσες δεικτοδότησης μπορούν να διαχωριστούν στις προ-συντεταγμένες και μετά-συντεταγμένες. Η πρώτη δείχνει ότι οι όροι συντονίζονται κατά την διάρκεια της δεικτοδότησης και της δεύτερης κατά την διάρκεια της αναζήτησης. Πιο συγκεκριμένα, στην προ-συντεταγμένη που συντάσσει

ευρετήριο, ένας λογικός συνδυασμός οποιωνδήποτε όρων δεικτοδότησης μπορεί να χρησιμοποιηθεί ως ετικέτα για να προσδιορίσει μια κλάση των εγγράφων, ενώ στην μετά-συντεταγμένη που συντάσσει ευρετήριο η ίδια κλάση θα προσδιοριζόταν στο χρόνο αναζήτησης με το συνδυασμό των κλάσεων των εγγράφων, επονομαζόμενων με τους μεμονωμένους όρους δεικτών.
Μια τελευταία διάκριση είναι ότι το λεξιλόγιο μιας γλώσσας δεικτοδότησης μπορεί να είναι ελεγχόμενο ή μη ελεγχόμενο. Το πρώτο αναφέρεται σε έναν κατάλογο εγκεκριμένων όρων δεικτοδότησης που ένας καταχωρητής μπορεί να χρησιμοποιήσει, όπως για παράδειγμα χρησιμοποιούνται στο MEDLARS. Οι έλεγχοι στη γλώσσα μπορούν επίσης να περιλάβουν τις ιεραρχικές σχέσεις μεταξύ των όρων δεικτοδότησης. Ή, κάποιος μπορεί να επιμείνει ότι ορισμένοι όροι μπορούν μόνο να χρησιμοποιηθούν ως επίθετα (ή χαρακτηριστικά). Δεν υπάρχει πραγματικά κανένα όριο στο είδος συντακτικών ελέγχων που κάποιος μπορεί να εφαρμόσει σε μια γλώσσα.
Η γλώσσα δεικτοδότησης που βγαίνει από τον αλγόριθμο conflation στο προηγούμενο Τμήμα μπορεί να περιγραφεί ως μη ελεγχόμενη, μετά-συντεταγμένη και εξαγόμενη. Το λεξιλόγιο των όρων ευρετηρίου σε οποιοδήποτε στάδιο στην εξέλιξη της συλλογής εγγράφων είναι ακριβώς το σύνολο όλων των ονομάτων της κλάσης conflation.
Υπάρχει πολλή διαμάχη για το είδος γλώσσας δεικτοδότησης που είναι καλύτερη για την ανάκτηση εγγράφων. Η βασική συζήτηση είναι για το εάν η αυτόματη δεικτοδότηση είναι τόσο καλή, ή καλύτερη από τη χειρωνακτική δεικτοδότηση. Κάθε μία μπορεί να γίνει σε διάφορα επίπεδα πολυπλοκότητας. Ωστόσο, φαίνεται να υπάρχουν αυξανόμενες αποδείξεις ότι και στις δυο περιπτώσεις, χειρωνακτική και αυτόματα δεικτοδότηση, αυξάνοντας την πολυπλοκότητα, με την μορφή πιο

περίπλοκων ελέγχων από τη στάθμιση των όρων δεικτοδότησης, δεν κερδίζουμε τίποτα. Το μήνυμα είναι ότι τα μη ελεγχόμενα λεξιλόγια βασισμένα στη φυσική γλώσσα επιτυγχάνουν αποτελεσματικότητα ανάκτησης, συγκρίσιμη με τα λεξιλόγια με επιμελημένους ελέγχους. Αυτό είναι εξαιρετικά ενθαρρυντικό, δεδομένου ότι η απλή γλώσσα δεικτοδότησης είναι η ευκολότερη για αυτοματοποίηση.
Πιθανώς τα ουσιαστικότερα στοιχεία για την αυτόματη δεικτοδότηση έχουν βγει από το πρόγραμμα SMART (1966). Η αυτόματη ανάλυση κειμένων πρέπει να χρησιμοποιήσει τους σταθμισμένους όρους που προέρχονται από τα αποσπάσματα εγγράφων των οποίων το μήκος είναι τουλάχιστον αυτό μιας περίληψης εγγράφων. Οι αντιπρόσωποι εγγράφων που χρησιμοποιούνται από το πρόγραμμα SMART είναι πιο περίπλοκοι από τους καταλόγους ριζών που εξάγονται από το conflation. .εν υπάρχει καμία αμφιβολία ότι οι ρίζες, παρά οι συνηθισμένες μορφές λέξεων είναι αποτελεσματικότερες. Συν τοις άλλοις το πρόγραμμα SMART προσθέτει τη στάθμιση όρου δεικτών, όπου ένας όρος δεικτών μπορεί να είναι μία ρίζα ή κάποια κατηγορία έννοιας που προσεγγίζεται μέσω της χρήσης των διάφορων λεξικών.
Παρακάτω θα συζητηθεί το είδος της συχνότητας πληροφοριών που μπορεί να χρησιμοποιηθεί για την στάθμιση των περιγραφητών εγγράφων και να εξηγήσει τη χρήση των αυτόματα κατασκευασμένων κατηγοριών όρου στην ανάκτηση πληροφορίας .

Στάθμιση όρου δεικτών

Οι δύο σημαντικότεροι παράγοντες που ελέγχουν την αποτελεσματικότητα μιας γλώσσας δεικτών είναι η διεξοδικότητα της δεικτοδότησης και η ειδικότητα της γλώσσας δεικτών. Για οποιοδήποτε έγγραφο, η διεξοδικότητα δεικτοδότησης ορίζεται ως ο αριθμός διαφορετικών θεμάτων που δεικτοδοτούνται και η γλωσσική ειδικότητα δεικτών είναι η δυνατότητα της γλώσσας δεικτών να περιγράφει τα θέματα ακριβώς. Περαιτέρω καθορίζεται η ιδιομορφία δεικτοδότησης ως επίπεδο ακρίβειας με το οποίο ένα έγγραφο συντάσσεται πραγματικά. Είναι πολύ δύσκολο να ποσοτικοποιηθούν αυτοί οι παράγοντες. Οι καταχωρητές είναι σε θέση να ταξινομήσουν την δεικτοδότηση τους περίπου κατά σειρά αυφανόμενης διεξοδικότητας ή ειδικότητας. Το ίδιο δεν είναι εύκολο στην αυτόματη δεικτοδότηση.
Έχει αναγνωριστεί ότι ένα υψηλό επίπεδο διεξοδικότητας της ευρετηρίασης οδηγεί στην υψηλή ανάκληση και τη χαμηλή ακρίβεια. Αντιθέτως, ένα χαμηλό επίπεδο διεξοδικότητας οδηγεί στη χαμηλή ανάκληση και την υψηλή ακρίβεια. Το αντίστροφο ισχύει για τα επίπεδα ειδικότητας δεικτοδότησης, υψηλή ειδικότητα οδηγεί σε υψηλή ακρίβεια και χαμηλή ανάκληση, κ.λ.π.... Φαίνεται επομένως, ότι υπάρχει ένα βέλτιστο επίπεδο διεξοδικότητας και ειδικότητας της δεικτοδότησης για έναν δεδομένο πληθυσμό χρηστών.
Πολλοί ερευνητές (Sparck Jones4,5, Salton και Yang6), έχουν προσπαθήσει να συσχετίσουν αυτούς τους δύο παράγοντες για να τεκμηριώσουν τη στατιστική συλλογή. Παραδείγματος χάριν, η διεξοδικότητα μπορεί να θεωρηθεί ότι αφορά τον αριθμό όρων δεικτών που εκχωρούνται σε ένα δεδομένο έγγραφο και η ειδικότητα αφορά των αριθμό εγγράφων στον οποίο ένας δεδομένος όρος εκχωρείται σε μια δεδομένη συλλογή. Η σημασία αυτής της μάλλον ασαφούς σχέσης είναι ότι οι δύο

παράγοντες συσχετίζονται με τη διανομή των όρων δεικτών στη συλλογή. Οι σχέσεις που τίθενται ως αίτημα είναι σύμφωνες με την παρατηρηθείσα ανταλλαγή μεταξύ της ακρίβειας και της ανάκλησης που προαναφέρθηκε. Οι αλλαγές στον αριθμό όρων δεικτών ανά έγγραφο οδηγούν στις αντίστοιχες αλλαγές στον αριθμό εγγράφων ανά όρο και αντίστροφα.
Υποστηρίζεται ότι με τη χρησιμοποίηση των κατανεμημένων πληροφοριών για τους όρους δεικτών για την παροχή, για παράδειγμα, της στάθμισης όρου δεικτών, στην πραγματικότητα ασχολούμαστε με το παλαιό πρόβλημα του ελέγχου της διεξοδικότητας και της ειδικότητας.
Με βάση τις αρχικές ιδέες του Luhn , είχε τεθεί ως αίτημα μια ποικίλη ισχύ διάκρισης για τους όρους δεικτών ως λειτουργία κατάταξης της συχνότητας εμφάνισης και η υψηλότερη ισχύς διάκρισης συσχετιζόταν με τις μέσες συχνότητες. Το πρότυπό του προτάθηκε για την επιλογή των σημαντικών όρων από ένα έγγραφο. Εντούτοις, οι ίδιες αρίθμησης συχνότητας μπορούν να χρησιμοποιηθούν για να παρέχουν ένα σχέδιο στάθμισης για τους μεμονωμένους όρους σε ένα έγγραφο. Στην πραγματικότητα, υπάρχει ένα κοινό σχέδιο στάθμισης σε χρήση που δίνει σε κάθε όρο δεικτών ένα βάρος άμεσα ανάλογο προς τη συχνότητα εμφάνισής της στο έγγραφο. Αρχικά, αυτό το σχέδιο φαίνεται να είναι ασυμβίβαστο με την υπόθεση Luhn ότι η ισχύς διάκρισης μειώνεται σε υψηλότερες συχνότητες. Εντούτοις, αναφέροντας στο σχήμα 3, το σχέδιο θα ήταν συνεπές εάν το ανώτερο οριακό σημείο μετακινηθεί προς το σημείο όπου εμφανίζεται η αιχμή. Αυτό φαίνεται ότι έχει συμβεί στα πειράματα που χρησιμοποιείται αυτή η ιδιαίτερη μορφή στάθμισης.
Έχουν γίνει προσπάθειες να εφαρμοστεί η στάθμιση βασισμένη στον τρόπο που οι όροι δεικτών κατανέμονται σε ολόκληρη συλλογή. Το λεξιλόγιο όρου δεικτών μιας συλλογής εγγράφων έχει συχνά μια κατανομή Zipfian, δηλαδή εάν μετράμε τον αριθμό εγγράφων στον οποίο κάθε όρος δεικτών εμφανίζεται και κάνουμε την γραφική τους παράσταση σύμφωνα με την κατάταξη, θα λάβουμε τη συνηθισμένη υπερβολή(καμπύλη).
Ο Sparck Jones22 απέδειξε πειραματικά ότι εάν υπάρχουν N έγγραφα και ένας όρος δεικτών εμφανίζεται σε n από αυτούς, τότε μια στάθμιση του log(N/n) + 1 οδηγεί σε αποτελεσματικότερη ανάκτηση από ότι θα συνέβαινε αν ο όρος που χρησιμοποιείται είναι αστάθμητος. Εάν η ειδικότητα δεικτοδότησης θεωρηθεί ότι είναι αντιστρόφως ανάλογη προς τον αριθμό των εγγράφων στα οποία ένας όρος δεικτών εμφανίζεται, τότε με την στάθμιση αποδίδεται μεγαλύτερη σημασία στους πιο συγκεκριμένους όρους.
Η διαφορά μεταξύ του τελευταίου τρόπου στάθμισης και του προηγούμενου μπορεί να συνοψιστεί λέγοντας ότι η στάθμιση συχνότητας εγγράφων δίνει έμφαση στην περιγραφή περιεχομένου ενώ σταθμίζοντας μέσω της ειδικότητας προσπαθεί να υπογραμμίσει τη δυνατότητα των όρων να διακρίνει ένα έγγραφο από άλλο.
Οι Salton και Yang6 έχουν πρόσφατα προσπαθήσει να συνδυάσουν και τις δύο Μεθόδους στάθμισης εξετάζοντας τις εντός των εγγράφων συχνότητες και τις μεταξύ των εγγράφων συχνότητες. Τα συμπεράσματά τους είναι ουσιαστικά, μια επέκταση αυτών του Luhn. Με την εξέταση της συνολικής συχνότητας της εμφάνισης ενός όρου και της κατανομής του στα έγγραφα, δηλαδή πόσες φορές εμφανίζεται σε κάθε έγγραφο, ήταν σε θέση να συναγάγουν διάφορα συμπεράσματα. Ένας όρος με υψηλή συνολική συχνότητα εμφάνισης δεν είναι πολύ χρήσιμος στην ανάκτηση ανεξάρτητα από τη κατανομή του. Οι μέσοι όροι συχνότητας είναι οι πιο χρήσιμοι ιδιαίτερα εάν η

κατανομή είναι ασύμμετρη. Οι σπάνιοι όροι με μια ασύμμετρη κατανομή είναι πιθανό να είναι χρήσιμοι αλλά λιγότερο από της μέσης συχνότητας. Οι πολύ σπάνιοι όροι είναι επίσης αρκετά χρήσιμοι αλλά τελευταίοι στον κατάλογο εκτός από αυτούς με υψηλή συνολική συχνότητα. Ο Salton και οι συνάδελφοί του έχουν αναπτύξει ένα ενδιαφέρον εργαλείο για να αποφανθούν εάν ένας δείκτης είναι "καλός" ή "κακός". Υποθέτουν ότι ένας καλός όρος δεικτών είναι ένας που, όταν εκχωρείται σε μια συλλογή των εγγράφων, καθιστά τα έγγραφα όσο το δυνατόν πιο ανόμοια, ενώ ένας κακός όρος είναι ένας που καθιστά τα έγγραφα πιο παρόμοια. Αυτό ποσοτικοποιείται μέσω μιας αξίας διάκρισης όρου που για συγκεκριμένο όρο μετρά την αύξηση ή την μείωση στη μέση ανομοιότητα μεταξύ των εγγράφων με την άρση εκείνου του όρου. Επομένως, ένας καλός όρος είναι εκείνος ο οποίος με την αφαίρεσή του από τη συλλογή των εγγράφων, οδηγεί σε μια μείωση στη μέση ανομοιότητα, ενώ ένας κακός όρος είναι εκείνος που οδηγεί με την αφαίρεσή του σε μια αύξηση. Η ιδέα είναι ότι ένας μεγαλύτερος διαχωρισμός μεταξύ των εγγράφων θα ενισχύσει την αποτελεσματικότητα ανάκτησης αλλά ότι μικρότερος διαχωρισμός θα καταστείλει την αποτελεσματικότητα ανάκτησης. Αν και επιφανειακά αυτό εμφανίζεται λογικό, αυτό που πραγματικά απαιτείται είναι τα σχετικά έγγραφα να έχουν μικρότερο διαχωρισμό (μεταξύ τους) σε σχέση με τα μη σχετικά. Έχουν αναφερθεί πειράματα που χρησιμοποιούν το πρότυπο διάκρισης όρου7,8. Μια σύνδεση μεταξύ της διάκρισης όρου και της συχνότητας μεταξύ εγγράφων έχει γίνει επίσης ενισχύοντας αποτελέσματα που αναφέρονται από τους Salton, Wong και Yang9. Παραδείγματος χάριν, το σχέδιο στάθμισης αντίστροφης συχνότητας εγγράφων που περιγράφεται παραπάνω, το οποίο εκχωρεί στάθμιση ανάλογη προς τον log(N/n) + 1 φαίνεται να είναι πιο αποτελεσματικό. Φυσικά, για να επιτύχουν μια απόδειξη αυτού του είδους πρέπει να γίνουν μερικές συγκεκριμένες υποθέσεις για το πώς να μετρήσουν την αποτελεσματικότητα και πώς να ταιριάξουν τα έγγραφα με τις ερωτήσεις. Καθιερώνουν επίσης την αποτελεσματικότητα μιας τεχνικής που χρησιμοποιείται για να συγχωνεύσει τους όρους χαμηλής συχνότητας, που αυξάνει την ανάκληση και μιας τεχνικής που χρησιμοποιείται για να συνδυάσει τους όρους υψηλής συχνότητας στις φράσεις, η οποία αυξάνει την ακρίβεια.

Πιθανολογική δεικτοδότηση (Probabilistic indexing)

Στο παρελθόν, ένα λεπτομερές ποσοτικό πρότυπο για την αυτόματη ευρετηρίαση βασισμένη σε στατιστικές υποθέσεις για τη κατανομή των λέξεων στο κείμενο, έχει προταθεί από τους Bookstein, Swanson, και Harter10, 11, 12. Η διαφορά μεταξύ των όρων τύπος και δείγμα είναι ουσιαστική για την κατανόηση του προτύπου τους. Ως εκ τούτου "η συχνότητα του περιστατικού της λέξης W σε ένα έγγραφο" σημαίνει τον αριθμό των δειγμάτων που εμφανίζονται σε εκείνο το έγγραφο που αντιστοιχεί σε έναν μοναδικό τύπο. Το πιστοποιητικό τύπος/ δείγμα μιας λέξης θα παραλείπεται όποτε το περιεχόμενο καθιστά σαφές τι σημαίνει όταν αναφερόμαστε σε μια λέξη.
Στο πρότυπό τους εξετάζουν τη διαφορά στην κατανεμημένη συμπεριφορά των λέξεων ως οδηγό για το εάν μια λέξη πρέπει να οριστεί ως όρος δεικτών. Βασίστηκαν στα παλαιότερα αποτελέσματα των Stone και Rubinoff13, Damerau14, και Dennis15, οι οποίοι έδειξαν ότι η στατιστική συμπεριφορά των λέξεων "ειδικότητας" ήταν

διαφορετική από αυτήν των λέξεων "λειτουργίας". Διαπίστωσαν ότι οι λέξεις λειτουργίας ακολουθούσαν μια κατανομή Poisson σε όλα τα έγγραφα ενώ οι λέξεις ειδικότητας δεν ακολούθησαν την κατανομή Poisson. Συγκεκριμένα, εάν εξετάζουμε την κατανομή μιας λέξης λειτουργίας W σε ένα σύνολο κειμένων, τότε η πιθανότητα f (n), ότι η λέξη λειτουργίας W θα εμφανίζεται n φορές σε ένα κείμενο δίνεται από τη σχέση

$$f(n) = \frac{e^{-\lambda} \lambda^{n}}{n!}$$

Γενικά η παράμετρος **x** θα ποικίλει από λέξη σε λέξη και για μια δεδομένη λέξη πρέπει να είναι ανάλογη προς το μήκος του κειμένου. Ερμηνεύουμε επίσης το **x** ως μέσο όρο των εμφανίσεων του **w** στο σύνολο κειμένων.
Το πρότυπο Bookstein-Swanson-Harter υποθέτει ότι οι λέξεις ειδικότητας είναι "φορείς-περιεχομένου" ενώ οι λέξεις λειτουργίας δεν είναι. Αυτό σημαίνει ότι μια λέξη που κατανέμεται τυχαία σύμφωνα με μια κατανομή Poisson δεν είναι πληροφοριακή για το έγγραφο στο οποίο εμφανίζεται. Συγχρόνως, το γεγονός ότι μια λέξη δεν ακολουθεί μια κατανομή Poisson υποθέτουμε ότι δείχνει ότι μεταβιβάζει πληροφορίες σχετικά με αυτό που το έγγραφο αφορά. Αυτό δεν είναι μια αδικαιολόγητη άποψη: ξέροντας ότι οι λέξη ειδικότητας WAR εμφανίζεται στη συλλογή, θα περιμέναμε να εμφανίζεται μόνο στα σχετικά λίγα έγγραφα που αναφέρονται στο WAR.
Το πρότυπο επίσης υποθέτει ότι ένα έγγραφο μπορεί να είναι για μια λέξη μέχρι ενός ορισμένου βαθμού. Αυτό υπονοεί ότι γενικά μια συλλογή εγγράφων μπορεί να χωριστεί σε υποσύνολα - κάθε υποσύνολο αποτελείται από έγγραφα που είναι για μια δεδομένη λέξη στον ίδιο βαθμό. Η θεμελιώδης υπόθεση που γίνεται τώρα είναι ότι μια λέξη φορέας-περιεχομένου είναι μια λέξη που διακρίνει περισσότερες από μια κατηγορίες εγγράφων όσον αφορά το βαθμό στον οποίο το θέμα που αναφέρεται από τη λέξη αντιμετωπίζεται στα έγγραφα σε κάθε κατηγορία. Είναι ακριβώς αυτές οι λέξεις που είναι οι υποψήφιοι για τους όρους δεικτών. Αυτές οι λέξεις φορείς-περιεχομένου μπορούν να ανιχνευθούν μηχανικά με τη μέτρηση του βαθμού στον οποίο οι κατανομές τους παρεκκλίνουν από αυτήν που αναμένεται μέσω μιας διαδικασίας Poisson. Σε αυτό το πρότυπο η θέση μιας από αυτές τις λέξεις περιεχομένου μέσα σε ένα υποσύνολο των εγγράφων του ίδιου "aboutness" είναι μια θέση μη φορέα περιεχομένου, δηλαδή μέσα στο δεδομένο υποσύνολο που δεν κάνει διακρίσεις μεταξύ περαιτέρω υποσυνόλων.
Ο Harter12 έχει διατυπώσει δύο υποθέσεις επάνω στις οποίες οι παραπάνω ιδέες μπορούν να χρησιμοποιηθούν για να παρέχουν μια μέθοδο αυτόματης δεικτοδότησης. Ο στόχος είναι να διευκρινιστεί ένας κανόνας που για οποιοδήποτε δεδομένο έγγραφο θα προσδιορίσει όρους δεικτών που επιλέγονται από τον κατάλογο υποψηφίων. Οι υποθέσεις είναι:

(1) Η πιθανότητα ότι ένα έγγραφο θα βρεθεί σχετικό με ένα αίτημα για τις πληροφορίες σε ένα θέμα είναι μια συνάρτηση του σχετικού βαθμού στον οποίο το θέμα διαπραγματεύεται στο έγγραφο.
(2) Ο αριθμός δειγμάτων σε ένα έγγραφο είναι μια συνάρτηση του βαθμού στον οποίο το θέμα που αναφέρεται από τη λέξη διαπραγματεύεται στο έγγραφο.
Σε αυτές τις υποθέσεις ένα "θέμα" προσδιορίζεται με το "θέμα του αιτήματος" και με το "θέμα που αναφέρεται από τη λέξη".
Επίσης, μόνο τα αιτήματα μεμονωμένης λέξης εξετάζονται, αν και οι Bookstein και ο Kraft16 σε μια πιο πρόσφατη εργασία προσπάθησαν μια επέκταση στα πολλών λέξεων αιτήματα. Ο κανόνας δεικτοδότησης βασισμένος σε αυτές τις υποθέσεις δεικτοδοτεί ένα έγγραφο με τη λέξη w εάν και μόνο εάν η πιθανότητα το έγγραφο να βρεθεί σχετικό με ένα αίτημα για πληροφορίες για το w, υπερβαίνει κάποια συνάρτηση κόστους. Για να υπολογίσουμε την απαραίτητη πιθανότητα της σχετικότητας για μια λέξη φορέα περιεχομένου πρέπει να ορίσουμε την εικόνα που θα είχε η κατανομή της.

Ξέρουμε ότι δεν μπορεί να είναι μια ενιαία κατανομή Poisson και ότι είναι ουσιαστικό μια λέξη φορέας περιεχομένου να διακρίνει μεταξύ των υποσυνόλων των εγγράφων που διαφέρουν στο βαθμό στον οποίο διαπραγματεύονται το θέμα που διευκρινίζεται από τη λέξη. Από την υπόθεση (2), μέσα σε ένα από αυτά τα υποσύνολα η κατανομή ενός φορέα περιεχομένου μπορεί εντούτοις να περιγραφεί με μια διαδικασία Poisson. Επομένως, εάν υπάρχουν μόνο δύο τέτοια υποσύνολα που διαφέρουν στο βαθμό στον οποίο είναι για μια λέξη w έπειτα η κατανομή του w μπορεί να περιγραφούν από ένα μίγμα δύο κατανομών Poisson.

Συγκεκριμένα, με το ίδιο σύστημα χαρακτήρων και συμβόλων όπως πριν έχουμε:

$$f(k) = \frac{p_1 e^{-x_1} x_1^k}{k!} + \frac{(1-p_1) e^{-x_2} x_2^k}{k!}$$

εδώ **p1** είναι η πιθανότητα ενός τυχαίου εγγράφου να ανήκει σε ένα από τα υποσύνολα και **x1** και **x2** είναι οι μέσες εμφανίσεις στις δύο κατηγορίες. Αυτή η εξίσωση εξηγεί και τον λόγο για τον οποίο το πρότυπο αυτό καλείται μερικές φορές πρότυπο 2-Poisson. Είναι σημαντικό να σημειωθεί ότι περιγράφει τη στατιστική συμπεριφορά μιας λέξης φορέα περιεχομένου σε δύο κατηγορίες που είναι "για" εκείνη την λέξη σε διαφορετικούς βαθμούς, αυτές οι κατηγορίες δεν είναι απαραιτήτως τα σχετικά και μη σχετικά έγγραφα αν και από την υπόθεση (1) μπορούμε να υπολογίσουμε την πιθανότητα της σχετικότητας για οποιοδήποτε έγγραφο από μια από αυτές τις κατηγορίες. Είναι η αναλογία

$$\frac{p_1 e^{-x_1} x_1^{\ k}}{p_1 e^{-x_1} x_1^{\ k} + (1-p_1) e^{-x_2} x_2^{\ k}}$$

που χρησιμοποιείται για να λάβει την απόφαση εάν θα ορίσει έναν όρο δεικτών w που εμφανίζεται k φορές σε ένα έγγραφο. Αυτή η αναλογία είναι στην πραγματικότητα η πιθανότητα ότι το συγκεκριμένο έγγραφο ανήκει στην κατηγορία που διαπραγματεύεται το w σε ένα μέσο βαθμό x1 δεδομένου ότι περιέχει ακριβώς k εμφανίσεις του w. Αυτή η αναλογία συγκρίνεται με κάποια συνάρτηση κόστους βασισμένη στο κόστος που ένας χρήστης είναι έτοιμος να αποδώσει στα λάθη που το σύστημα κάνει στην ανάκτηση.

Τέλος, αν και οι δοκιμές έχουν δείξει ότι αυτό το πρότυπο ορίζει "λογικούς" όρους δεικτών, δεν έχει εξεταστεί από την άποψη της αποτελεσματικότητάς του στην ανάκτηση. Τελικά αυτό θα καθορίσει εάν είναι αποδεκτό ως πρότυπο για την αυτόματη ευρετηρίαση.

Διάκριση και /ή αντιπροσώπευση

Υπάρχουν δύο συγκρουόμενοι τρόποι παρατήρησης του προβλήματος της χαρακτήρισης των εγγράφων για την ανάκτηση. Ο πρώτος είναι να χαρακτηριστεί ένα έγγραφο μέσω μιας αντιπροσώπευσης του περιεχομένου του, ανεξάρτητα από τον τρόπο με τον οποίο άλλα έγγραφα μπορεί να περιγράφονται και αυτή είναι η αντιπροσώπευση χωρίς διάκριση. Ο άλλος τρόπος είναι να επιμείνει ότι κατά τον χαρακτηρισμό ενός εγγράφου το διαχωρίζουμε από όλα, ή ενδεχομένως όλα, τα άλλα έγγραφα στη συλλογή, αυτό καλείται διάκριση χωρίς αντιπροσώπευση. Φυσικά, καμία από αυτές τις ακραίες θέσεις δεν εφαρμόζεται στην πράξη, αν και ο προσδιορισμός και των δύο είναι χρήσιμος όταν αναλύεται το πρόβλημα του χαρακτηρισμού.
Στην πράξη, επιδιώκεται κάποιο είδος βέλτιστης ανταλλαγής μεταξύ της αντιπροσώπευσης και της διάκρισης. Παραδοσιακά αυτό έχει προσπαθηθεί μέσω της εξισορρόπησης της διεξοδικής δεικτοδότησης σε αντίθεση με την ειδικότητα. Οι περισσότερες αυτόματες μέθοδοι δεικτοδότησης μπορούν να θεωρηθούν ένα μίγμα της αντιπροσώπευσης εναντίον της διάκρισης. Στην απλή περίπτωση της αφαίρεσης λέξεις υψηλής συχνότητας με τη βοήθεια ενός καταλόγου λέξεων "stop" προσπαθούμε να αυξήσουμε το επίπεδο διάκρισης μεταξύ του εγγράφου. Οι μέθοδοι του Salton βασισμένες στην αξία διάκρισης προσπαθούν το ίδιο πράγμα. Εντούτοις, πρέπει να είναι σαφές ότι κατά την αφαίρεση των πιθανών όρων δεικτών πρέπει να υπάρξει ένα στάδιο όπου οι εναπομείναντες δεν μπορούν πλέον να αντιπροσωπεύσουν επαρκώς το περιεχόμενο των εγγράφων. Το συμβατικό πρότυπο των Bookstein-Swanson-Harter μπορεί να θεωρηθεί ως ένα στο οποίο η σημασία ενός όρου στην αντιπροσώπευση του περιεχομένου ενός εγγράφου, είναι ισορροπημένη ενάντια στη σημασία της ως descriminator.
Η έμφαση στην αντιπροσώπευση οδηγεί σε προσανατολισμό -εγγράφου: δηλαδή σε ένα συνολικό προβληματισμό για το τι είναι ένα έγγραφο. Αυτή η προσέγγιση πιθανώς να εφαρμοστεί στην έρευνα για την τεχνητή νοημοσύνη, ιδιαίτερα με αυτήν που ασχολείται με την κατασκευή των προτύπων υπολογιστών του περιεχομένου οποιουδήποτε δεδομένου κομματιού του κειμένου φυσικής γλώσσας. Η σχετικότητα

αυτής της εργασίας στην τεχνητή νοημοσύνη, καθώς επίσης και άλλων εργασιών, έχουν συνοψιστεί από τον Smith17.
Αυτή η άποψη υιοθετείται επίσης από εκείνους που ασχολούνται με τον καθορισμό μιας έννοιας "των πληροφοριών", υποθέτουν ότι μόλις εξηγηθεί κατάλληλα αυτή η έννοια, ένα έγγραφο θα μπορεί να αντιπροσωπευθεί από τις "πληροφορίες" που περιέχει18.
Η έμφαση στη διάκριση οδηγεί σε έναν προσανατολισμό -ερώτησης. Από αυτήν την οπτική γωνία προϋποθέτουμε ότι μπορεί να προβλεφθεί ο πληθυσμός των ερωτήσεων που πιθανώς να υποβληθούν στο σύστημα ανάκτησης πληροφορίας. Λαμβάνοντας υπόψη τα στοιχεία για αυτόν τον πληθυσμό των ερωτήσεων, μπορούμε να δοκιμάσουμε και να χαρακτηρίσουμε τα έγγραφα στη βέλτιστη κατάσταση.

Αυτόματη ταξινόμηση λέξεων κλειδιών

Πολλά αυτόματα συστήματα ανάκτησης στηρίζονται στα λεξικά συνωνύμων για να τροποποιήσουν τις ερωτήσεις και τους αντιπροσώπους εγγράφων για να βελτιώσουν την πιθανότητα ανάκλησης σχετικών εγγράφων. Ο Salton40 έχει πειραματιστεί με πολλά διαφορετικά είδη λεξικά συνωνύμων και έχει καταλήξει στο συμπέρασμα ότι πολλά από αυτά δικαιολογούν την χρήση τους από την άποψη της βελτιωμένης αποτελεσματικότητας ανάκτησης.
Στην πράξη πολλά από τα λεξικά συνωνύμων κατασκευάζονται με το χέρι. Έχουν κατασκευαστεί κυρίως με δύο τρόπους:

(1) οι λέξεις που θεωρείται ότι αναφέρονται στο ίδιο θέμα, συνδέονται

(2) οι λέξεις που θεωρείται ότι αναφέρονται σε σχετικά πράγματα, συνδέονται.

Το πρώτο είδος λεξικών συνωνύμων συνδέει τις λέξεις που είναι αλληλοαντικαταστόμενες, δηλαδή τις βάζει σε κατηγορίες ισοδυναμίας. Κατόπιν μια λέξη θα μπορούσε να επιλεχτεί για να αντιπροσωπεύσει κάθε κατηγορία και ένας κατάλογος αυτών των λέξεων θα μπορούσε να χρησιμοποιηθεί για να διαμορφώσει ένα ελεγχόμενο λεξιλόγιο. Από αυτό ένας καταχωρητής θα μπορούσε να καθοδηγηθεί για να επιλέξει τις λέξεις για να δεικτοδοτήσει ένα έγγραφο, ή ο χρήστης θα μπορούσε να καθοδηγηθεί για να επιλέξει τις λέξεις με τις οποίες θα εκφράσει την ερώτησή του. Τα ίδια λεξικά συνωνύμων θα μπορούσαν να χρησιμοποιηθούν με έναν αυτόματο τρόπο για να προσδιορίσουν τις λέξεις μιας ερώτησης με σκοπό την ανάκτηση.
Το δεύτερο είδος λεξικών συνωνύμων χρησιμοποιεί σημασιολογικές συνδέσεις μεταξύ των λέξεων, για παράδειγμα, τις συνδέει ιεραρχικά.
Εντούτοις, μέθοδοι έχουν προταθεί για την κατασκευή λεξικών συνωνύμων αυτόματα. Αν και τα λεξικά συνωνύμων που φτιάχνονται με το χέρι είναι βασισμένα σε σημασιολογικές σχέσεις (π.χ. αναγνωρίζουν τα συνώνυμα, τις γενικότερες, ή πιο συγκεκριμένες σχέσεις) τα αυτόματα λεξικά συνωνύμων τείνουν να είναι βασισμένα σε συντακτικές και στατιστικές σχέσεις. Πάλι η χρήση της σύνταξης έχει αποδειχθεί μικρής αξίας, γι' αυτό θα επικεντρωθούμε στις στατιστικές μεθόδους. Αυτές βασίζονται κυρίως στα μοτίβα της συνεμφάνισης των λέξεων στα έγγραφα. Αυτές οι

"λέξεις" είναι συχνά τα περιγραφικά στοιχεία που εισήχθησαν νωρίτερα ως όροι ή λέξεις κλειδιά.
Η βασική σχέση που κρύβεται κάτω από την αυτόματη κατασκευή των κατηγοριών λέξεων κλειδιών είναι η ακόλουθη:
Εάν οι λέξεις κλειδιά a και b μπορούν να αντικατασταθούν η μία από την άλλη, υπό την έννοια ότι μπορούμε να δεχτούμε ένα έγγραφο που περιέχει την μία ως απάντηση σε ένα αίτημα που περιέχει την άλλη, αυτό γίνεται επειδή έχουν την ίδια έννοια ή αναφέρονται σε ένα κοινό θέμα ή ένα θέμα. Ένας τρόπος να βρούμε εάν δύο λέξεις κλειδιά σχετίζονται, είναι με την εξέταση των εγγράφων στα οποία εμφανίζονται. Εάν τείνουν να συνεμφανίζονται στα ίδια έγγραφα, είναι πολύ πιθανόν ότι έχουν να κάνουν με το ίδιο θέμα και έτσι μπορούν να αντικαταστήσουν η μία την άλλη.
Δεν είναι δύσκολο να καταλάβουμε ότι, βασιζόμενοι σε αυτήν την αρχή, μια ταξινόμηση των λέξεων κλειδιών μπορεί να κατασκευαστεί αυτόματα, μέσα στην οποία οι κατηγορίες χρησιμοποιούνται ανάλογα με εκείνες των λεξικών συνωνύμων που φτιάχνονται με το χέρι και αναφέρθηκαν πριν. Πιο συγκεκριμένα μπορούμε να προσδιορίσουμε δύο κύριες προσεγγίσεις στη χρήση των ταξινομήσεων λέξεων κλειδιών:

(1) αντικατάσταση κάθε λέξης κλειδί σε έναν αντιπρόσωπο εγγράφων (και ερώτησης) από το όνομα της κατηγορίας στην οποία εμφανίζεται

(2) αντικατάσταση κάθε λέξης κλειδί από όλες τις λέξεις κλειδιά που εμφανίζονται στην κατηγορία στην οποία ανήκει.

Ο Jones αναφέρει έναν μεγάλο αριθμό πειραμάτων, στα οποία χρησιμοποίησε τις Αυτόματες ταξινομήσεις λέξεων κλειδιών και διαπίστωσε ότι σε γενικές γραμμές είχε καλύτερη απόδοση ανάκτησης με την βοήθεια της αυτόματης ταξινόμησης λέξεων κλειδιών από ότι με τις αταξινόμητες λέξεις κλειδιά.
Δυστυχώς, ακόμη και εδώ το αποδεικτικό στοιχείο δεν είναι σημαντικό. Η εργασία από τον Minker19 και τους συνεργάτες του δεν έχει επιβεβαιώσει τα συμπεράσματα του Jones και στην πραγματικότητα έχουν δείξει ότι σε μερικές περιπτώσεις η ταξινόμηση λέξεων κλειδιών μπορεί να είναι επιζήμια στην αποτελεσματικότητα ανάκτησης.

Κανονικοποίηση (Normalisation)

Είναι πιθανώς χρήσιμό σε αυτή τη φάση να ανακεφαλαιωθεί και να επιδειχθεί ο τρόπος με τον οποίο διάφορα επίπεδα κανονοκοποίησης του κειμένου περιλαμβάνονται στην παραγωγή των αντιπροσώπων εγγράφων. Στο χαμηλότερο επίπεδο έχουμε το έγγραφο που περιγράφεται μόνο από μια σειρά των λέξεων. Το πρώτο βήμα στην κανονικοποίηση είναι να αφαιρεθούν οι λέξεις "fluff". Τώρα έχουμε αυτό που καλούμε "λέξεις κλειδιά". Το επόμενο στάδιο μπορεί να είναι η συγχώνευση αυτών των λέξεων σε κατηγορίες και να περιγραφούν τα έγγραφα από τα σύνολα ονομάτων κατηγορίας που στη σύγχρονη ορολογία είναι οι λέξεις κλειδιά ή οι όροι δεικτών. Το επόμενο επίπεδο είναι η κατασκευή των κατηγοριών λέξεων κλειδιών από την αυτόματη

ταξινόμηση. Για να κυριολεκτήσουμε εκεί σταματά η κανονικοποίηση.
Η στάθμιση όρου δεικτών μπορεί επίσης να θεωρηθεί ως διαδικασία της κανονικοποίησης, εάν το σχέδιο στάθμισης λαμβάνει υπόψη τον αριθμό διαφορετικών όρων δεικτών ανά έγγραφο. Παραδείγματος χάριν, μπορεί να θέλουμε να εξασφαλίσουμε ότι μια αντιστοιχία σε έναν όρο μεταξύ δέκα, έχει μεγαλύτερο βάρος από ότι ένας μεταξύ είκοσι. Ομοίως, η διαδικασία στάθμισης μέσω της συχνότητας του περιστατικού στη συνολική συλλογή εγγράφων είναι μια προσπάθεια να ομαλοποιηθούν οι αντιπρόσωποι εγγράφων όσον αφορά τις αναμενόμενες κατανομές συχνότητας.

3. Αυτόματη Ταξινόμηση (Automatic classification)

Στο κεφάλαιο αυτό θα προσπαθήσουμε να παρουσιάσουμε έναν συνεπή απολογισμό της ταξινόμησης κατά τέτοιο τρόπο ώστε οι σχετικές αρχές να γίνουν κατανοητές για οποιονδήποτε επιθυμήσει να χρησιμοποιήσει τεχνικές ταξινόμησης στην ανάκτηση πληροφοριών χωρίς ιδιαίτερη δυσκολία. Έμφαση θα δοθεί στην εφαρμογή τους στην ομαδοποίηση εγγράφων, αν και πολλές από αυτές τις ιδέες εφαρμόζονται στην αναγνώριση προτύπων, την αυτόματη ιατρική διάγνωση και την ομαδοποίηση λέξεων κλειδιών.
Ας σημειωθεί πως στις ενότητες που ακολουθούν δεν θα δοθεί ένας επίσημος ορισμός της ταξινόμησης, καθώς για το σκοπό μας είναι αρκετό να θεωρήσουμε την ταξινόμηση, ως περιγραφή της διαδικασίας από την οποία ένα ταξινομητικό σύστημα κατασκευάζεται. Η λέξη **"ταξινόμηση"** χρησιμοποιείται επίσης για να περιγράψει το αποτέλεσμα μιας τέτοιας διαδικασίας. Αν και η δεικτοδότηση θεωρείται συχνά ως "ταξινόμηση" αποκλείουμε αυτήν την έννοια. Μια περαιτέρω διάκριση που θα γίνει είναι μεταξύ "της ταξινόμησης" και "της διάγνωσης". Η καθημερινή γλώσσα είναι πολύ διφορούμενη σε αυτό το σημείο:
"Πώς θα ταξινομούσατε (προσδιορίζουμε) αυτό;"
"Πώς ταξινομούνται αυτά καλύτερα (ομαδοποίηση);"
Το πρώτο παράδειγμα αναφέρεται στη **διάγνωση** ενώ το δεύτερο μιλά για **ταξινόμηση.** Αυτές οι διακρίσεις έχουν ήδη γίνει στην βιβλιογραφία από τους Kendall20 και Jardine και Sibson21.
Στα πλαίσια της ανάκτησης πληροφοριών, μια ταξινόμηση απαιτείται για έναν σκοπό. Εδώ θα ακολουθήσουμε τον Macnaughton-Smith22 που δηλώνει: «Όλες οι ταξινομήσεις ακόμα και οι γενικότερες πραγματοποιούνται για κάποιον περισσότερο ή λιγότερο σαφή «ειδικό σκοπό» ή το σύνολο των σκοπών που θα επηρεάσουν την επιλογή της μεθόδου ταξινόμησης και τα αποτελέσματα που θα πάρουμε.» Ο σκοπός μπορεί να είναι να ομαδοποιηθούν τα έγγραφα κατά τέτοιο τρόπο ώστε η ανάκτηση να είναι πιο γρήγορη ή εναλλακτικά να κατασκευαστεί ένα λεξικό συνωνύμων αυτόματα. Όποιος και να είναι ο σκοπός η ποιότητα της ταξινόμησης μπορεί τελικά να μετρηθεί μόνο από την απόδοσή της κατά τη διάρκεια της ανάκτησης. Υπάρχουν δύο κύριοι τομείς της εφαρμογής των μεθόδων ταξινόμησης στο θεματικό πεδίο της ανάκτησης πληροφοριών:

(1) ομαδοποίηση λέξεων κλειδιών
(2) ομαδοποίηση εγγράφων.

Η ομαδοποίηση των εγγράφων, αν και συστήνεται από τον Salton και τους συνεργάτες του, έχει ασκήσει πολύ λίγη επίδραση. Ένας πιθανός λόγος είναι ότι οι λεπτομέρειες της εργασίας του Salton για την ομαδοποίηση των εγγράφων υποτιμήθηκαν μέσα στη σύγχυση των πειραμάτων που έγιναν στο σύστημα του SMART. Ένας άλλος πιθανός λόγος είναι, ότι καθώς ο ενθουσιασμός για την Ομαδοποίηση μειώθηκε, έγινε κατανοητό ότι σημαντικά πειράματα σε αυτήν την περιοχή απαιτούν μεγάλο όγκο δεδομένων και τεράστιο υπολογιστικό χρόνο.
Ο Good23 και ο Fairthorne24 ήταν από τους πρώτους που υπέδειξαν ότι η αυτόματη Ταξινόμηση μπορεί να αποδειχθεί χρήσιμη στην ανάκτηση εγγράφων. Μια σαφής δήλωση του τι υπονοείται από την ομαδοποίηση εγγράφων, έγινε αρχικά από τον P. M. Hayes25 : «Ορίζουμε την οργάνωση ως ομαδοποίηση στοιχείων (π.χ. έγγραφα, αντιπροσωπεύσεις εγγράφων) τα οποία αντιμετωπίζονται έπειτα ως μονάδα και χάνουν σε αυτό το βαθμό τις μεμονωμένες ταυτότητές τους. Με άλλα λόγια, η ταξινόμηση ενός εγγράφου σε μια θέση ταξινόμησης, προσδιορίζει για κάθε χρήση το έγγραφο με εκείνη την θέση. Έκτοτε, αυτό και άλλα έγγραφα σε αυτή την θέση αντιμετωπίζονται ως όμοια έως ότου εξεταστούν χωριστά. Φαίνεται επομένως, ότι τα έγγραφα ομαδοποιούνται επειδή είναι υπό κάποια έννοια σχετικά το ένα με το άλλο αλλά κυρίως ομαδοποιούνται, επειδή είναι πιθανό να τα θέλουμε μαζί και η λογική σχέση είναι ο τρόπος μέτρησης αυτής της πιθανότητας.» Η "λογική οργάνωση" έχει επιτευχθεί με δύο διαφορετικούς τρόπους. Αρχικά, μέσω της άμεσης ταξινόμησης των εγγράφων και αφετέρου μέσω του ενδιάμεσου υπολογισμού ενός μέτρου της σχετικότητας μεταξύ των εγγράφων. Η πρώτη προσέγγιση έχει αποδειχθεί θεωρητικά δυσεπίλυτη έτσι ώστε οποιαδήποτε πειραματικά αποτελέσματα της δοκιμής δεν μπορούν να θεωρηθούν αξιόπιστα. Η δεύτερη προσέγγιση της ταξινόμησης είναι πλέον αρκετά καλά τεκμηριωμένη και προ πάντων, υπάρχουν κάποια ισχυρά επιχειρήματα συστήνοντας την σε μα συγκεκριμένη μορφή. Σε αυτή την προσέγγιση θα δοθεί έμφαση εδώ.
Η αποδοτικότητα της ομαδοποίησης των εγγράφων έχει υπογραμμιστεί από τον Salton26, που λέει: "Σαφώς στην πράξη δεν είναι δυνατό να ταιριάξουμε κάθε έγγραφο που έχει αναλυθεί με κάθε αίτημα αναζήτησης που έχει αναλυθεί, επειδή ο χρόνος που θα καταναλωθεί από μια τέτοια λειτουργία θα ήταν υπερβολικός. Διάφορες λύσεις έχουν προταθεί για να μειώσουν τον αριθμό αναγκαίων συγκρίσεων μεταξύ των στοιχείων πληροφοριών και των αιτημάτων. Μια ιδιαίτερα ελπιδοφόρα λύση παράγει ομάδες σχετικών εγγράφων, χρησιμοποιώντας μια αυτόματη διαδικασία ταυτοποίησης εγγράφων. Ένα αντιπροσωπευτικό διάνυσμα ομάδας εγγράφων επιλέγεται έπειτα για κάθε ομάδα εγγράφων και ένα αίτημα αναζήτησης ελέγχεται αρχικά, μόνο σε σχέση με όλα τα διανύσματα ομάδας. Μετά, το αίτημα ελέγχεται σε σχέση μόνο με εκείνα τα μεμονωμένα έγγραφα όπου τα διανύσματα ομάδας παρουσιάζουν υψηλό αποτέλεσμα με το αίτημα. "Ο Salton θεωρεί ότι αν και η ομαδοποίηση των εγγράφων κερδίζει χρόνο, μειώνει την αποτελεσματικότητα ενός συστήματος ανάκτησης.

Μέτρα συσχέτισης (Measures of association)

Μερικές μέθοδοι ταξινόμησης βασίζονται σε μια δυαδική σχέση μεταξύ των αντικειμένων. Βάσει αυτής της σχέσης μια μέθοδος ταξινόμησης μπορεί να

κατασκευάσει ένα σύστημα ομάδων. Η σχέση περιγράφεται ποικιλοτρόπως ως "ομοιότητα", "συσχέτιση" και "ανομοιότητα". Αγνοώντας την ανομοιότητα προς το παρόν δεδομένου ότι θα καθοριστεί από μαθηματική άποψη αργότερα, οι άλλοι δύο όροι σημαίνουν σχεδόν τον ίδιο εκτός από το ότι η "συσχέτιση" θα χρησιμοποιηθεί για την ομοιότητα μεταξύ των αντικειμένων που χαρακτηρίζονται από διακριτές καταστάσεις ιδιοτήτων (discrete-state attributes).
Το μέτρο της ομοιότητας έχει ως σκοπό να ποσοτικοποιήσει την ομοιότητα μεταξύ των αντικειμένων έτσι ώστε εάν υποθέσουμε ότι είναι δυνατό να ομαδοποιηθούν τα αντικείμενα κατά τέτοιο τρόπο ώστε ένα αντικείμενο σε μια ομάδα μοιάζει πιο πολύ με τα άλλα μέλη της ομάδας, από ότι με οποιοδήποτε αντικείμενο έξω από την ομάδα, τότε μια μέθοδος ομαδοποίησης επιτρέπει σε μια τέτοια δομή ομάδας να ανακαλυφθεί. Ο Lerman έχει ερευνήσει τη μαθηματική σχέση μεταξύ πολλών από τα μέτρα και έχει δείξει ότι πολλά είναι μονότονα το ένα όσον αφορά το άλλο. Συνεπάγεται ότι μια μέθοδος ομαδοποίησης που εξαρτάται μόνο από την ταξινόμηση των τιμών συσχέτισης θα δώσει πανομοιότυπες ομάδες για όλα αυτά τα μέτρα.
Υπάρχουν πέντε μέτρα συσχέτισης που συνήθως χρησιμοποιούνται στην ανάκτηση πληροφοριών. Δεδομένου ότι, στην ανάκτηση πληροφοριών τα έγγραφα και τα αιτήματα συνήθως αντιπροσωπεύονται από καταλόγους όρων ή λέξεων κλειδιών, θα απλοποιήσουμε τα πράγματα υποθέτοντας ότι ένα αντικείμενο αντιπροσωπεύεται από ένα σύνολο λέξεων κλειδιών και ότι το μέτρο μέτρησης | . | δίνει το μέγεθος του συνόλου. Μπορούμε εύκολα να γενικεύσουμε στην περίπτωση όπου οι λέξεις κλειδιά έχουν σταθμιστεί, απλά επιλέγοντας ένα κατάλληλο μέτρο. Το πιο απλό μέτρο συσχέτισης είναι

$|X \cap Y|$ Συντελεστής απλού ταιριάσματος

το οποίο είναι ο αριθμός των κοινών όρων δεικτών. Αυτός ο συντελεστής δεν λαμβάνει υπόψη τα μεγέθη του X και του Y. Οι ακόλουθοι συντελεστές που έχουν χρησιμοποιηθεί στην ανάκτηση εγγράφων λαμβάνουν υπόψη τις πληροφορίες που παρέχονται από τα μεγέθη του X και του Y.

$2\frac{|X \cap Y|}{|X| + |Y|}$ *Dice's coefficient*

$\frac{|X \cap Y|}{|X \cup Y|}$ *Jaccard's coefficient*

$\frac{|X \cap Y|}{|X|^{1/2} \times |Y|^{1/2}}$ *Cosine coefficient*

$\frac{|X \cap Y|}{\min(|X|, |Y|)}$ *Overlap coefficient*

Αυτοί μπορούν όλοι να θεωρηθούν κανονικοποιημένες εκδόσεις του απλού συντελεστή ταυτοποίησης. Η αποτυχία κανονικοποίησης οδηγεί σε αντίθετα διαισθητικά αποτελέσματα όπως φαίνεται στο ακόλουθο παράδειγμα:

If $S_1(X, Y) = |X \cap Y|$ $S_2(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$

then $|X_1| = 1\ |Y_1| = 1\ |X_1 \cap Y_2| = 1 \Rightarrow S_1 = 1 S_2 = 1$

$|X_2| = 10\ |Y_2| = 10\ |X_2 \cap Y_2| = 1 \Rightarrow S_1 = 1 S_2 =$

$= 1/10$

S1 (X1, Y1) = S1 (X2, Y2) που είναι σαφώς παράλογο δεδομένου ότι X1 και Y1 είναι ίδιοι αντιπρόσωποι ενώ X2 και Y2 είναι ριζικά διαφορετικά. Η κανονικοποίηση για το S2 το διαβαθμίζει μεταξύ του 0 και του 1, το 1 δίνει τη μέγιστη ομοιότητα.
Επιστρέφουμε τώρα στον μαθηματικό καθορισμό της ανομοιότητας. Οι λόγοι για την προτίμηση της "ανομοιότητας" είναι κυρίως τεχνικοί και δεν θα επεκταθούμε εδώ.
Πρέπει να σημειωθεί, ότι οποιαδήποτε συνάρτηση ανομοιότητας μπορεί να μετασχηματιστεί σε μια συνάρτηση ομοιότητας με έναν απλό μετασχηματισμό της μορφής s=(1+d)_1 αλλά το αντίστροφο δεν ισχύει πάντα.
Εάν P είναι το σύνολο των αντικειμένων που ομαδοποιούνται, ένας συντελεστής D ανομοιότητας ζευγών, είναι μια συνάρτηση από P X P στο σύνολο των μη αρνητικών πραγματικών αριθμών. Το D γενικά, ικανοποιεί τις ακόλουθες συνθήκες:

D1 D(X,Y) >= 0 για όλα τα X,Y € P
D2 D(X,X) = 0 για όλα τα X,Y € P

Ένας συντελεστής ανομοιότητας θα μπορούσε να είναι ένα είδος συνάρτησης απόστασης. Στην πραγματικότητα, πολλοί από τους συντελεστές ανομοιότητας ικανοποιούν την τριγωνική ανισότητα:

D4 D(X, Y) +< D(X, Z) + D (Y, Z)

Ένα από τα πιο σημαντικά θεωρήματα της Ευκλείδειας Γεωμετρίας, που δηλώνει ότι το άθροισμα των μηκών δύο πλευρών ενός τριγώνου είναι πάντα μεγαλύτερο από το μήκος της τρίτης πλευράς.
Ένα παράδειγμα ενός συντελεστή ανομοιότητας που ικανοποιεί D1 - D4 είναι

$$\frac{|X \Delta Y|}{|X| + |Y|}$$

όπου $(X \Delta Y) = (X \cup Y) - (X \cap Y)$ είναι η συμμετρική διαφορά των συνόλων X και Y
Συσχετίζεται απλώ με τον συντελεστή Dice σε

$$1 - \frac{2|X \cap Y|}{|X| + |Y|} = \frac{|X \Delta Y|}{|X| + |Y|}$$

και είναι μονότονη όσον αφορά το συντελεστή Jaccard που αφαιρείται από 1. Για να ολοκληρώσουμε την εικόνα, θα εκφράσουμε τον συντελεστή του Dice με μια

διαφορετική μορφή. Αντί της αντιπροσώπευσης κάθε εγγράφου από ένα σύνολο λέξεων κλειδιών, το αντιπροσωπεύουμε από μια δυαδική σειρά όπου η απουσία ή η παρουσία της λέξης κλειδιού ith υποδεικνύεται από ένα μηδέν ή ένα στη θέση ith(ιοστή) αντίστοιχα. Σε αυτή την περίπτωση

$$D = \frac{\sum x_i(1-y_i) + \sum y_i(1-x_i)}{\sum x_i + \sum y_i}$$

όπου το άθροισμα είναι για το συνολικό αριθμό διαφορετικών λέξεων κλειδιών στη συλλογή εγγράφων. Ο Salton θεώρησε τους αντιπροσώπους εγγράφων ως δυαδικά διανύσματα που ενσωματώθηκαν σε ένα *ν* - διαστατικό Ευκλείδειο διάστημα, όπου *το ν* είναι ο συνολικός αριθμός όρων δεικτών.

$$B = \frac{|X \cap Y|}{|X|^{1/2}|Y|^{1/2}}$$

Αυτή η σχέση μπορεί τώρα να ερμηνευθεί ως συνημίτονο του γωνιακού χωρισμού των δύο δυαδικών διανυσμάτων X και Y. Αυτό γενικεύεται εύκολα στην περίπτωση όπου το X και το Y είναι αυθαίρετα πραγματικά διανύσματα (δηλ. σταθμισμένοι κατάλογοι λέξης κλειδιού) στην οποία περίπτωση γράφουμε όπου (X, Y) είναι το εσωτερικό γινόμενο και το || . || είναι το μήκος ενός διανύσματος.
Εάν το διάστημα είναι Ευκλείδειο τότε για X = (x1..., xn) και Y = (y1..., yn)
Παίρνουμε

$$\frac{\sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n} x_i^2\right)^{1/2}\left(\sum_{i=1}^{n} y_i^2\right)^{1/2}}$$

Μερικοί συγγραφείς έχουν προσπαθήσει να βασίσουν ένα μέτρο συσχέτισης σε ένα πιθανολογικό πρότυπο27. Μετρούν την συσχέτιση μεταξύ δύο αντικειμένων από το βαθμό στον οποίο οι κατανομές τους παρεκκλίνουν από την πιθανολογική ανεξαρτησία. Αργότερα θα χρησιμοποιείται ως το αναμενόμενο αμοιβαίο μέτρο πληροφοριών για να μετρηθεί η συσχέτιση. Για δύο διακριτές κατανομές πιθανοτήτων P(x i) και P(x j) μπορεί να οριστεί ως εξής:

$$I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

Όταν xi και xj είναι ανεξάρτητα τότε P(xi)P(xj) = P(xi,xj) και έτσι I(xi,xj) =0.
Επίσης I(xi, xj) = 0, Επίσης I(xi, xj) = I(xj, xi) που δείχνει ότι είναι συμμετρικό.
Έχει επίσης την ιδιότητα του να είναι αμετάβλητο κάτω από τους έναν προς έναν

μετασχηματισμούς των συντεταγμένων. Άλλες ενδιαφέρουσες ιδιότητες αυτού του μέτρου μπορούν να βρεθούν στην εργασία των Osteyee και Good28. Ο Rajski29 δείχνει πώς το I(xixj) μπορεί να μετασχηματιστεί απλά σε μια συνάρτηση απόστασης σε διακριτές κατανομές πιθανοτήτων. Το I(xixj) ερμηνεύεται συχνά ως μέτρο των στατιστικών πληροφοριών που περιλαμβάνονται στο xi για το xj (ή αντίστροφα). Όταν εφαρμόζουμε αυτήν την συνάρτηση για να μετρήσουμε την συσχέτιση μεταξύ δύο όρων δεικτών, για παράδειγμα i και j, τότε τα xi και xj είναι δυαδικές μεταβλητές. Κατά συνέπεια P(xi = 1) θα είναι η πιθανότητα της εμφάνισης του όρου i και ομοίως P(xi = 0) θα είναι η πιθανότητα της μη εμφάνισής του. Ο βαθμός στον οποίο οι δύο όροι δεικτών i και j συσχετίζονται, μετριέται έπειτα από το i (xi xj) που μετρά το βαθμό στον οποίο οι κατανομές τους παρεκκλίνουν από την πιθανολογική ανεξαρτησία.
Μια συνάρτηση όμοια με το αναμενόμενο αμοιβαίο μέτρο πληροφοριών προτάθηκε από τους Jardine και Sibson21 για να μετρήσει συγκεκριμένα την ανομοιότητα μεταξύ δύο κατηγοριών αντικειμένων. Για παράδειγμα, είμαστε σε θέση να κάνουμε διακρίνουμε δύο κατηγορίες, βάσει των κατανομών πιθανότητάς τους, για ένα απλό διάστημα δύο σημείων {1,0}. Κατά συνέπεια έστω P1(1), P1(0) και P2(1), P2(0) είναι οι κατανομές πιθανοτήτων που συσχετίζονται με την κατηγορία I και II αντίστοιχα. Τώρα βάσει της διαφοράς μεταξύ τους μετράμε την ανομοιότητα μεταξύ του I και II από αυτό που από τους Jardine και Sibson καλείται ακτίνα πληροφοριών, η οποία είναι

$$uP_1(1)\log\frac{P_1(1)}{uP_1(1)+vP_2(1)} + vP_2(1)\log\frac{P_2(1)}{uP_1(1)+vP_2(1)} +$$

$$uP_1(0)\log\frac{P_1(0)}{uP_1(0)+vP_2(0)} + vP_2(0)\log\frac{P_2(0)}{uP_1(0)+vP_2(0)}$$

Εδώ το u και το v είναι θετικά βάρη προσθέτοντας στη μονάδα. Αυτή η συνάρτηση γενικεύεται εύκολα για την περίπτωση περισσότερων κατηγοριών, ή για συνεχή κατανομή. Επίσης εύκολα αποδεικνύεται, ότι κάτω από κάποια ερμηνεία το αναμενόμενο αμοιβαίο μέτρο πληροφοριών είναι μια ειδική περίπτωση της ακτίνας πληροφοριών.

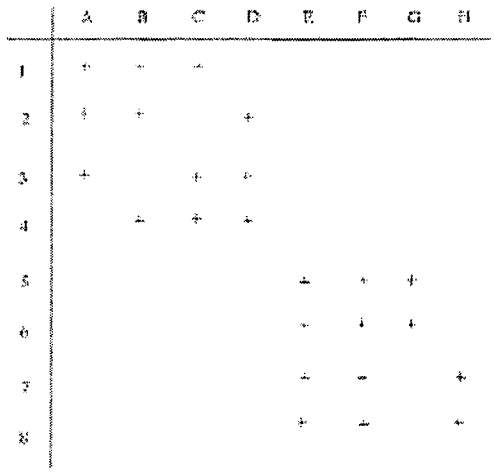
Μέθοδοι ταξινόμησης

Θα αρχίσουμε με μια περιγραφή του είδους δεδομένων για το οποίο οι μέθοδοι ταξινόμησης είναι κατάλληλες. Τα δεδομένα αποτελούνται από τα αντικείμενα και τις αντίστοιχες περιγραφές τους. Τα αντικείμενα μπορούν να είναι έγγραφα, λέξεις κλειδιά, χειρόγραφοι χαρακτήρες, ή είδη (στην τελευταία περίπτωση τα αντικείμενα τα ίδια είναι κατηγορίες σε αντιδιαστολή με τα μεμονωμένα). Οι περιγραφείς τους υπάρχουν με διάφορα ονόματα ανάλογα με τη δομή τους:
(1) ιδιότητες πολλαπλών καταστάσεων (π.χ. χρώμα)
(2) ιδιότητες δυαδικής κατάστασης (π.χ. λέξεις κλειδιά)
(3) αριθμητικές ιδιότητες (π.χ. σταθμισμένες λέξεις κλειδιά)
(4) κατανομές πιθανοτήτων.

Η τέταρτη κατηγορία περιγραφών ισχύει όταν τα αντικείμενα είναι κατηγορίες. Παραδείγματος χάριν, το πλάτος φύλλων ενός είδους φυτών μπορεί να περιγραφεί από μια κανονική κατανομή ορισμένου μέσου όρου και διακύμανσης. Οι μέθοδοι ταξινόμησης χρησιμοποιούνται για να συνοψιστεί και να απλοποιηθεί αυτό το είδος δεδομένων.
Ο Sparck Jones24 παρείχε μια πολύ σαφή ταξινόμηση των μεθόδων ταξινόμησης, βάση κάποιων γενικών χαρακτηριστικών του προκύπτοντος ταξινομητικού συστήματος. Αναφέρουμε:
(1) Σχέση μεταξύ των ιδιοτήτων και των κατηγοριών
(α) μονοθετικές
(β) πολυθετικές
(2) Σχέση μεταξύ των αντικειμένων και των κατηγοριών
(α) αποκλειστική
(β) επικαλύπτουσα
(3) Σχέση μεταξύ των κατηγοριών
(α) διατεταγμένη
(β) μη διατεταγμένη
Η πρώτη κατηγορία έχει εξερευνηθεί λεπτομερώς από τους αριθμητικούς ταξινόμους. Μια αρχική δήλωση της διάκρισης μεταξύ μονοθετικής και πολυθετικής δίνεται από τον Beckner30: "Μια κατηγορία καθορίζεται συνήθως από την αναφορά σε ένα σύνολο ιδιοτήτων που είναι και απαραίτητες και ικανοποιητικές (από τον όρο) για την ιδιότητα μέλους στην κατηγορία. Είναι δυνατό, ωστόσο, να καθοριστεί μια ομάδα Κ όσον αφορά ένα καθορισμένο σύνολο G των ιδιοτήτων f1, f2,...,fn κατά διαφορετικό τρόπο.

Υποθέστε ότι έχουμε ένα σύνολο των ατόμων (δεν θα τα ονομάσουμε ακόμα κατηγορία) έτσι ώστε
(1) καθένας κατέχει έναν μεγάλο (αλλά απροσδιόριστο) αριθμό των ιδιοτήτων στο G
(2) κάθε f στο G κατέχεται από μεγάλο αριθμό αυτών των ατόμων και
(3) κανένα f στο G δεν κατέχεται από κάθε άτομο στο σύνολο."

Ή πρώτη πρόταση της δήλωσης Beckner αναφέρεται στον κλασσικό αριστοτελικό Καθορισμό μιας κατηγορίας, η οποία καλείται τώρα μονοθετική. Το δεύτερο μέρος ορίζει την πολυθετική.
Για να επεξηγήσουμε τη βασική διάκριση θα δούμε το ακόλουθο παράδειγμα (σχήμα 4) 8 ατόμων (1-8) και 8 ιδιοτήτων (A-H). Η κατοχή ενός χαρακτηριστικού γνωρίσματος, υποδεικνύεται από το σύμβολο +. Τα άτομα 1-4 αποτελούν μια πολυθετική ομάδα, κάθε άτομο που κατέχει τρία από τα τέσσερα χαρακτηριστικά γνωρίσματα A, B, C, D. Τα άλλα 4 άτομα μπορούν να χωριστούν σε δύο μονοθετικές κατηγορίες {5, 6} και {7, 8} . Η διάκριση μεταξύ μονοθετικής και πολυθετικής είναι αρκετή, εφόσον τα χαρακτηριστικά γνωρίσματα είναι απλού είδους, π.χ. ιδιότητες δυαδικών καταστάσεων. Όταν τα χαρακτηριστικά γνωρίσματα είναι πιο σύνθετα, οι ορισμοί είναι δυσκολότερο να εφαρμοστούν και εν πάση περιπτώσει είναι μάλλον αυθαίρετοι.



Σχήμα 4: Μία επεξήγηση της διαφοράς μεταξύ monothetic και polythetic

Η διάκριση μεταξύ επικάλυψης και αποκλειστικότητας είναι σημαντική και τόσο από θεωρητική όσο και πρακτική άποψη. Πολλές μέθοδοι ταξινόμησης μπορούν να αντιμετωπισθούν ως μέθοδοι απλοποίησης δεδομένων. Στη διαδικασία της ταξινόμησης, πληροφορίες απορρίπτονται, έτσι ώστε τα μέλη μιας κατηγορίας να είναι όμοια.. Οι επικαλύπτουσες κατηγορίες επιτρέπονται για να ελαχιστοποιηθεί το ποσό πληροφοριών που απορρίπτουμε, ή με άλλα λόγια, για να έχουμε μια ταξινόμηση που είναι, υπό κάποια έννοια, πιο κοντά στα αρχικά δεδομένα. Δυστυχώς αυτό αλλάζει τα πάντα όσον αφορά την αποδοτικότητα της εφαρμογής. Ένας συμβιβασμός μπορεί να υιοθετηθεί στον οποίο οι μέθοδοι ταξινόμησης παράγουν επικαλύπτουσες κατηγορίες σε πρώτο στάδιο και τελικά "τακτοποιούνται" για να δώσουν τις αποκλειστικές

κατηγορίες.
Ένα παράδειγμα μιας διατεταγμένης ταξινόμησης, είναι μια ιεραρχία. Οι κατηγορίες ταξινομούνται με προσμέτρηση, π.χ. οι κατηγορίες σε ένα επίπεδο τοποθετούνται στις κατηγορίες στο επόμενο επίπεδο. Είναι δύσκολο να δοθεί ένα απλό παράδειγμα της μη διατεταγμένης ταξινόμησης. Μη διατεταγμένες κατηγορίες γενικά εμφανίζονται κατά την αυτόματη κατασκευή λεξικών συνωνύμων. Οι κατηγορίες που επιδιώκονται για ένα λεξικό συνωνύμων είναι εκείνες που ικανοποιούν ορισμένους όρους ομοιογένειας και απομόνωσης, αλλά γενικά δεν μπορούν απλά να σχετίζονται η μία με την άλλη. Για ορισμένες εφαρμογές ταξινόμησης είναι άσχετο, ενώ για κάποιες άλλες όπως η ομαδοποίηση εγγράφων είναι ζωτικής σημασίας. Η διάταξη επιτρέπει την επινόηση αποδοτικών στρατηγικών αναζήτησης.

Η υπόθεση ομαδοποίησης (clustering)

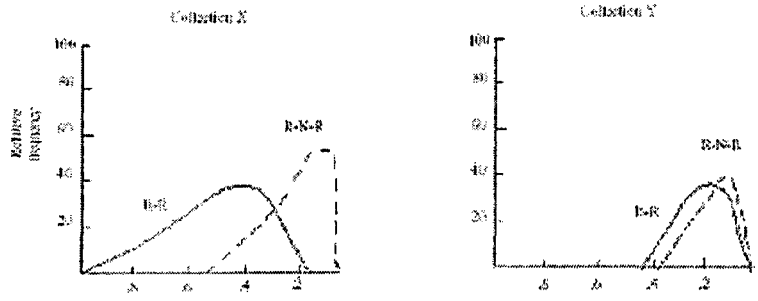
Πριν περιγράψουμε την πληθώρα των μεθόδων ταξινόμησης που χρησιμοποιούνται τώρα στην ανάκτηση πληροφοριών, πρέπει να συζητήσουμε την ισχύουσα υπόθεση για τη χρήση της ομαδοποίησης εγγράφων. Αυτή η υπόθεση μπορεί να εκφραστεί απλά ως εξής: έγγραφα που είναι στενά συσχετισμένα τείνουν να είναι σχετικά με τα

ίδια αιτήματα. Θα αναφερθούμε σε αυτήν την υπόθεση **ως υπόθεση ομαδοποίησης (clustering).**
Μία βασική υπόθεση σε συστήματα ανάκτησης είναι ότι έγγραφα σχετικά με ένα αίτημα διαχωρίζονται από εκείνα που δεν είναι σχετικά, δηλ. τα σχετικά έγγραφα μοιάζουν πιο πολύ το ένα με το άλλο από ότι με τα μη σχετικά έγγραφα. Εάν αυτό ισχύει για μια συλλογή μπορεί να εξεταστεί ως εξής. Υπολογίζουμε την συσχέτιση μεταξύ όλων των ζευγαριών των εγγράφων:

(α) και τα δύο είναι σχετικά με ένα αίτημα, και
(β) ένα από τα δύο είναι σχετικό και το άλλο μη σχετικό.

Το άθροισμα ενός συνόλου αιτημάτων δίνει τη σχετική κατανομή των συσχετίσεων των ζευγών σχετικό-σχετικό (R-R) και σχετικό -μη σχετικό (R-N-R) μιας συλλογής. Σχεδιάζοντας τη γραφική παράσταση της σχετικής συχνότητας προς την ένταση της συσχέτισης για δύο υποθετικές συλλογές X και Y μπορεί να πάρουμε τις κατανομές που φαίνονται στο σχήμα 5.
Από αυτά είναι προφανές:

(α) ότι ο διαχωρισμός για τη συλλογή X είναι καλός ενώ για τη Y δεν είναι και
(β) ότι η ένταση της συσχέτισης μεταξύ των σχετικών εγγράφων είναι μεγαλύτερη για τη X παρά για τη Y.



Σχήμα 5: R-R είναι η κατανομή των συσχετίσεων των ζευγών σχετικό-σχετικό και R-N-
R είναι η κατανομή των συσχετίσεων των ζευγών σχετικό -μη σχετικό.

Αυτό τον διαχωρισμό μεταξύ των κατανομών προσπαθούμε να εκμεταλλευτούμε την ομαδοποίηση των εγγράφων. Βάσει αυτού του διαχωρισμού μπορούμε να υποστηρίξουμε ότι η ομαδοποίηση των εγγράφων μπορεί να οδηγήσει σε αποτελεσματικότερη ανάκτηση από ότι μια γραμμική αναζήτηση. Μια γραμμική αναζήτηση αγνοεί τη σχέση που υπάρχει μεταξύ των εγγράφων. Εάν η υπόθεση ικανοποιείται για μια συγκεκριμένη συλλογή, τότε είναι σαφές ότι η δόμηση της συλλογής κατά τέτοιο τρόπο ώστε τα πολύ σχετικά έγγραφα εμφανίζονται σε μια κατηγορία, όχι μόνο θα επιταχύνει την ανάκτηση αλλά μπορεί επίσης να την καταστήσει αποτελεσματικότερη, δεδομένου ότι όταν βρίσκεται μια κατηγορία, θα τείνει να περιέχει μόνο σχετικά και όχι μη σχετικά έγγραφα.

Πρέπει να προσθέσουμε ότι αυτά τα συμπεράσματα μπορούν να επιβεβαιωθούν, μόνο από την πειραματική εργασία για έναν μεγάλο αριθμό συλλογών. Ένας λόγος για αυτό είναι ότι, αν και είναι δυνατό να δομηθεί μια συλλογή εγγράφων με τέτοιο τρόπο ώστε τα σχετικά έγγραφα να συγκεντρώνονται, δεν υπάρχει καμία εγγύηση ότι μια στρατηγική αναζήτησης θα βρει, χωρίς σφάλματα, την κατηγορία εγγράφων που περιέχει τα σχετικά έγγραφα. Η ικανότητα σχεδιασμού στρατηγικών αναζήτησης που θα κάνουν τη συγκεκριμένη εργασία, είναι ένα θέμα για πειραματισμό.
Σημειώστε ότι η υπόθεση των ομάδων αναφέρεται στις δεδομένες περιγραφές εγγράφων. Ο σκοπός της παραγωγής μόνιμων ή προσωρινών αλλαγών σε μια περιγραφή με τεχνικές όπως οι ταξινομήσεις λέξεων κλειδιών, μπορεί να εκφραστεί ως μία προσπάθεια να αυξήσουμε την απόσταση μεταξύ των δύο κατανομών R-R και RN-R. Δηλαδή θέλουμε να αυξήσουμε τις πιθανότητες ανάκτησης σχετικών εγγράφων και να μειώσουμε τις πιθανότητες ανάκτησης μη σχετικών εγγράφων. Όπως φαίνεται, η υπόθεση ομαδοποίησης είναι ένας βολικός τρόπος να εκφράσουμε το στόχο διαδικασιών όπως η ομαδοποίηση εγγράφων. Φυσικά, δεν δίνει πληροφορίες για την εκμετάλλευση του διαχωρισμού.

Η χρήση της ομαδοποίησης στην ανάκτηση πληροφοριών

Θα προσπαθήσουμε να απομονώσουμε τα ουσιαστικά χαρακτηριστικά των διάφορων μεθόδων ομαδοποίησης.
Στην επιλογή μιας μεθόδου ομαδοποίησης για χρήση στη πειραματική ανάκτηση πληροφοριών, χρησιμοποιούνται συχνά δύο αλληλοσυγκρουόμενα κριτήρια.

Το πρώτο από αυτά και πιθανώς το σημαντικότερο σ' αυτό το στάδιο ανάπτυξης του θέματος, είναι η θεωρητική πληρότητα της μεθόδου. Με αυτό εννοούμε ότι η μέθοδος πρέπει να ικανοποιεί συγκεκριμένα κριτήρια ορθότητας. Μερικά από το σημαντικότερα είναι:

(1) η μέθοδος παράγει ομαδοποίηση που είναι απίθανο να αλλαχτεί δραματικά όταν συμπεριληφθούν περαιτέρω αντικείμενα, δηλ. είναι σταθερό υπό αύξηση.

(2) η μέθοδος είναι σταθερή υπό την έννοια ότι μικρά λάθη στην περιγραφή των αντικειμένων οδηγούν σε μικρές αλλαγές στην ομαδοποίηση.

(3) η μέθοδος είναι ανεξάρτητη από την αρχική διάταξης των αντικειμένων.

Αυτό σημαίνει ότι οποιαδήποτε μέθοδος ομαδοποίησης που δεν ικανοποιεί αυτούς τους όρους είναι απίθανο να παραγάγει οποιαδήποτε σημαντικά πειραματικά αποτελέσματα. Πολλές μέθοδοι ομαδοποίησης δεν ικανοποιούν αυτά τα κριτήρια, πιθανώς επειδή οι αλγόριθμοι για την εφαρμογή τους τείνουν να είναι λιγότερο αποδοτικοί από τους ad hoc αλγόριθμους ομαδοποίησης.
Το δεύτερο κριτήριο για την επιλογή είναι η αποδοτικότητα της διαδικασίας ομαδοποίησης όσον αφορά τις απαιτήσεις ταχύτητας και αποθήκευσης. Σε κάποια πειραματική εργασία αυτό είναι η πρωταρχική μέριμνα. Αλλά πριν ακόμα μάθουμε αρκετά για τη συμπεριφορά των συγκεντρωμένων αρχείων όσον αφορά την

αποτελεσματικότητα της ανάκτησης (δηλ. η δυνατότητα ανάκτησης επιθυμητών και συγκράτησης ανεπιθύμητων εγγράφων) δεν μπορούμε να επιμείνουμε στην αποδοτικότητα. Εν πάση περιπτώσει, πολλές από τις "καλές" θεωρητικές μεθόδους (αυτές που είναι πιθανό να παραγάγουν σημαντικά πειραματικά αποτελέσματα) μπορούν να τροποποιηθούν για να αυξήσουν την αποδοτικότητα της διαδικασίας ομαδοποίησης.
Η αποδοτικότητα είναι στην πραγματικότητα ένα χαρακτηριστικό γνώρισμα του αλγόριθμου που εφαρμόζει τη μέθοδο ομαδοποίησης. Είναι μερικές φορές χρήσιμο να διακριθεί η μέθοδος ομαδοποίησης από τον αλγόριθμό της, αλλά στα πλαίσια της ανάκτησης πληροφορίας αυτή η διάκριση γίνεται λιγότερο χρήσιμη, δεδομένου ότι πολλές μέθοδοι ομαδοποίησης καθορίζονται από τον αλγόριθμό τους, έτσι δεν υπάρχει καμία ρητή μαθηματική διατύπωση.

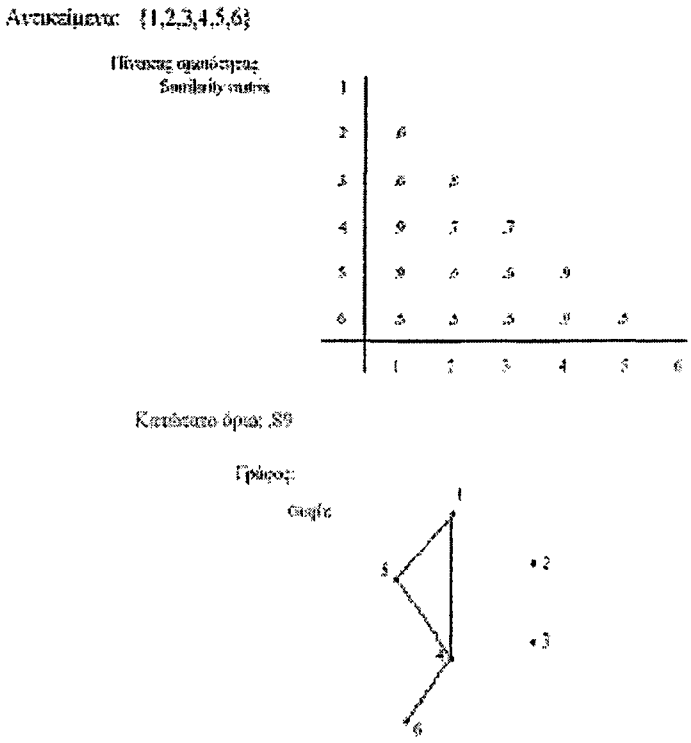
Δύο ευδιάκριτες προσεγγίσεις ομαδοποίησης μπορούν να προσδιοριστούν:

(1) ομαδοποίηση είναι βασισμένη σε ένα μέτρο της ομοιότητας μεταξύ των αντικειμένων που ομαδοποιούνται.

(2) η μέθοδος ομαδοποίησης προχωρά άμεσα από τις περιγραφές αντικειμένου.

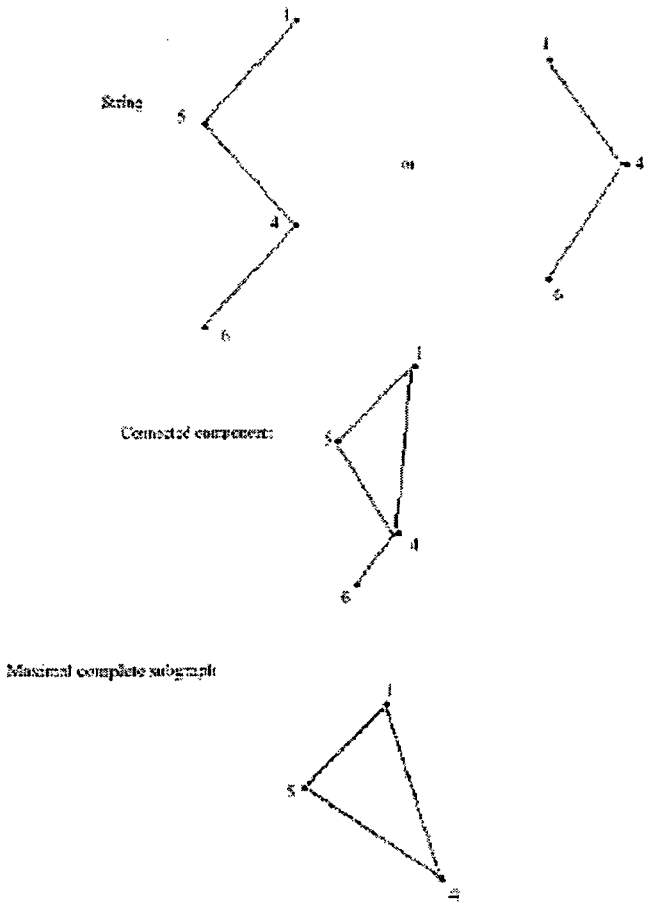
Τα προφανέστερα παραδείγματα της πρώτης προσέγγισης είναι οι θεωρητικές μέθοδοι γραφικών παραστάσεων που καθορίζουν τις ομάδες όσον αφορά μια γραφική παράσταση που προέρχεται από το μέτρο της ομοιότητας. Αυτή η προσέγγιση εξηγείται καλύτερα με ένα παράδειγμα (δείτε το σχήμα 6). Θεωρήστε ένα σύνολο αντικειμένων που θα ομαδοποιηθεί. Υπολογίζουμε μια αριθμητική τιμή για κάθε ζεύγος των αντικειμένων που δείχνει την ομοιότητά τους. Μια γραφική παράσταση που αντιστοιχεί σε αυτό το σύνολο τιμών ομοιότητας λαμβάνεται ως εξής:

Καθορίζεται μια τιμή κατώτατου ορίου (threshold) και δύο αντικείμενα θεωρούνται συνδεμένα εάν η τιμή ομοιότητάς τους είναι μεγαλύτερη από το κατώτατο όριο. Ο ορισμός της ομάδας γίνεται μέσω της γραφικής αντιπροσώπευσης.



Σχήμα 6: Ένας συντελεστής ομοιότητας για έξι αντικείμενα και η γραφική παράσταση
που παράγεται από αυτόν όταν θέτουμε κατώτατο όριο (threshold)

Μια σειρά (string) είναι μια συνδεδεμένη ακολουθία αντικειμένων από κάποια αφετηρία.
Ένα συνδεδεμένο συστατικό (connected component) είναι ένα σύνολο αντικειμένων έτσι ώστε κάθε αντικείμενο συνδέεται με τουλάχιστον ένα άλλο μέλος του συνόλου και το σύνολο είναι μέγιστο όσον αφορά αυτό το χαρακτηριστικό γνώρισμα.
Μέγιστος πλήρης υπογράφος (maximal complete subgraph) είναι ένας υπογράφος έτσι ώστε κάθε κόμβος συνδέεται με τους άλλους κόμβους υπογράφων και το σύνολο είναι μέγιστο όσον αφορά αυτό το χαρακτηριστικό γνώρισμα, δηλ. εάν ένας ακόμα κόμβος συμπεριλήφθηκε οπουδήποτε, ο όρος πληρότητας θα παραβιαζόταν. Ένα παράδειγμα για το κάθε ένα δίνεται στο σχήμα 7. Αυτές οι μέθοδοι έχουν χρησιμοποιηθεί εκτενώς στη ομαδοποίηση λέξεων κλειδιών από τους Sparck Jones και Jakson31, Augustson και Minker32 και Vaswani και τον Cameron33.



Σχήμα 7: Μερικοί πιθανοί ορισμοί ομάδων (clusters) με βάση υπογράφου (subgraph)

Μια μεγάλη κατηγορία ιεραρχικών μεθόδων ομαδοποίησης είναι βασισμένη στην αρχική μέτρηση της ομοιότητας. Η σημαντικότερη από αυτές είναι η single-link η οποία είναι η μοναδική που χρησιμοποιείται εκτενώς στην ανάκτηση εγγράφων. Ικανοποιεί όλα τα προαναφερθέντα κριτήρια της ορθότητας. Στην πραγματικότητα, οι Jardine και Sibson2 έχουν δείξει ότι κάτω από ένα συγκεκριμένο αριθμό λογικών όρων η single-link είναι η μόνη ιεραρχική μέθοδος που ικανοποιεί αυτά τα σημαντικά κριτήρια.
Μια περαιτέρω κατηγορία ομαδοποίησης μεθόδων βασισμένων στη μέτρηση της ομοιότητας είναι η κατηγορία τον ονομαζόμενων μεθόδων clump. Λειτουργούν με την αναζήτηση συνόλων που ικανοποιούν ορισμένους όρους συνοχής και απομόνωσης που καθορίζονται από το μέτρο ομοιότητας. **Οι υπολογιστικές δυσκολίες αυτής της μεθόδου έχουν οδηγήσει σε μεγάλο βαθμό στην εγκατάλειψή της.** Μια προσπάθεια να παραχθεί μια ιεραρχία clump, έγινε από τον van Rijsbergen34 αλλά, όπως ήταν αναμενόμενο, ο καθορισμός ομάδας ήταν τόσο αυστηρός που πολύ λίγα σύνολα μπορούσαν να τον ικανοποιήσουν.

Η αποδοτικότητα είναι η πρωταρχική μέριμνα στον καθορισμό των αλγοριθμικά καθορισμένων μεθόδων ομαδοποίησης που χρησιμοποιούνται στην ανάκτηση πληροφοριών. Για αυτόν τον λόγο οι περισσότερες από αυτές τις μεθόδους τείνουν να προχωρούν άμεσα από την περιγραφή του αντικειμένου στην τελική ταξινόμηση χωρίς έναν ενδιάμεσο υπολογισμό ενός μέτρου ομοιότητας. Ένα άλλο ιδιαίτερο χαρακτηριστικό αυτών των μεθόδων είναι ότι δεν επιδιώκουν μια υφιστάμενη δομή στα δεδομένα, αλλά προσπαθούν να επιβάλουν μια κατάλληλη δομή σε αυτά. Αυτό επιτυγχάνεται με τον περιορισμό του αριθμού ομάδων και με την οριοθέτηση του μεγέθους κάθε ομάδας.
Παρά να δοθεί μια λεπτομερής περιγραφή όλων των ευρετικών αλγορίθμων, θα συζητήσουμε αντ' αυτού μερικούς από τους κύριους τύπους. Πριν προχωρήσουμε, πρέπει να καθορίσουμε μερικές από τις έννοιες που χρησιμοποιούνται στο σχεδιασμό αυτών των αλγορίθμων.
Η σημαντικότερη έννοια είναι αυτή της αντιπροσωπευτικής συστάδας, που ονομάζεται προφίλ ομάδας, διάνυσμα ταξινόμησης, ή κεντροειδές (centroid). Είναι απλά ένα αντικείμενο που συνοψίζει και αντιπροσωπεύει τα αντικείμενα στην ομάδα. Ιδανικά πρέπει να είναι, με κάποιο τρόπο, κοντά σε κάθε αντικείμενο στην ομάδα, γι' αυτό χρησιμοποιείται ο όρος κεντροειδές (centroid). Η ομοιότητα των αντικειμένων με τον αντιπρόσωπο μετριέται από μια συνάρτηση ταυτοποίησης (μερικές φορές αποκαλούμενη συνάρτηση ομοιότητα ή συσχετισμού). Οι αλγόριθμοι χρησιμοποιούν επίσης διάφορες εμπειρικά καθορισμένες παραμέτρους όπως:

(1) επιθυμητό αριθμό ομάδων

(2) ένα ελάχιστο και μέγιστο μέγεθος για κάθε ομάδα

(3) μια τιμή κατώτατου ορίου στην συνάρτηση ταυτοποίησης, κάτω από την οποία ένα αντικείμενο δεν θα περιληφθεί σε μια ομάδα

(4) ο έλεγχος της επικάλυψης μεταξύ των ομάδων

(5) μια αυθαίρετα επιλεγμένη αντικειμενική συνάρτηση που βελτιστοποιείται.

Σχεδόν όλοι οι αλγόριθμοι είναι επαναληπτικοί, δηλ. η τελική ταξινόμηση επιτυγχάνεται με την επαναληπτική βελτίωση μιας ενδιάμεσης ταξινόμησης. Αν και οι περισσότεροι αλγόριθμοι έχουν καθοριστεί μόνο για την ταξινόμηση ενός επιπέδου, μπορούν προφανώς να επεκταθούν στην ταξινόμηση πολλαπλών επιπέδων, απλώς θεωρώντας τις ομάδες σε ένα επίπεδο, ως αντικείμενα που θα ταξινομηθούν στο επόμενο επίπεδο.
Πιθανώς ο σημαντικότερος αυτού του είδους αλγορίθμου είναι ο αλγόριθμος35 ομαδοποίησης του Rocchio, που αναπτύχθηκε στο πρόγραμμα SMART. Λειτουργεί σε τρία στάδια. Στο πρώτο στάδιο επιλέγει (με κάποιο κριτήριο) διάφορα αντικείμενα ως κέντρα ομαδοποίησης. Βάσει της αρχικής ανάθεσης οι αντιπρόσωποι ομάδας υπολογίζονται και όλα τα αντικείμενα ορίζονται ακόμα μια φορά στις ομάδες. Οι

κανόνες ανάθεσης καθορίζονται ρητά όσον αφορά τα κατώτατα όρια σε μια συνάρτηση ταυτοποίησης. Οι τελικές ομάδες μπορεί να επικαλύπτονται (δηλ. ένα αντικείμενο μπορεί να οριστεί σε περισσότερες από μια ομάδες). Το δεύτερο στάδιο είναι ουσιαστικά ένα επαναληπτικό βήμα για να επιτρέψει στις διάφορες παραμέτρους εισαγωγής να ρυθμιστούν, έτσι ώστε η προκύπτουσα ταξινόμηση να ανταποκρίνεται καλύτερα στις προδιαγραφές πραγμάτων, όπως το μέγεθος της ομάδας, κ.λπ.. Το τρίτο στάδιο είναι για "την τακτοποίηση". Τα αντικείμενα που δεν έχουν οριστεί, ορίζονται, και η επικάλυψη μεταξύ των ομάδων μειώνεται.
Οι περισσότεροι από αυτούς τους αλγορίθμους στοχεύουν στη μείωση του αριθμού περασμάτων που πρέπει να γίνουν από το αρχείο περιγραφών αντικειμένου. Υπάρχει ένας μικρός αριθμός αλγορίθμων ομαδοποίησης που απαιτούν μόνο ένα πέρασμα του αρχείου των περιγραφών αντικειμένου. Για αυτό το λόγο δίνεται το όνομα **αλγόριθμος απλού περάσματος** (Single-Pass Algorithm) σε μερικούς από αυτούς. Βασικά λειτουργούν ως εξής:

(1) οι περιγραφές αντικειμένου υποβάλλονται σε επεξεργασία σειριακά

(2) το πρώτο αντικείμενο γίνεται ο αντιπρόσωπος της πρώτης ομάδας

(3) κάθε επόμενο αντικείμενο αντιστοιχείται προς όλους τους αντιπροσώπους ομάδας που υπάρχουν στο χρόνο επεξεργασίας του

(4) ένα δεδομένο αντικείμενο ορίζεται σε μια ομάδα (ή περισσότερο εάν η επικάλυψη επιτρέπεται) σύμφωνα με κάποιο όρο στην συνάρτηση ταυτοποίησης

(5) όταν ένα αντικείμενο ορίζεται σε μια ομάδα ο αντιπρόσωπος για εκείνη την ομάδα υπολογίζεται ξανά

(6) εάν ένα αντικείμενο αποτυγχάνει μια ορισμένη δοκιμή γίνεται ο αντιπρόσωπος μιας νέας ομάδας.

Ακόμα μια φορά η τελική ταξινόμηση εξαρτάται από τις παραμέτρους εισαγωγής που μπορούν να καθοριστούν μόνο εμπειρικά (και που πιθανώς είναι διαφορετικές για διαφορετικά σύνολα αντικειμένων) και πρέπει να διευκρινιστούν εκ των προτέρων.
Η απλούστερη έκδοση αυτού του είδους αλγορίθμου οφείλεται στον Hill36. Στη συνέχεια, παράχθηκαν πολλές παραλλαγές, κυρίως ως αποτέλεσμα αλλαγών στους κανόνες ανάθεσης και στον καθορισμό των αντιπροσώπων ομάδας. (Δείτε για παράδειγμα Rieber και Marathe37, Johnson και Lafuente38 και Etzweiler και Martin39).
Ο αλγόριθμος MacQueen40, που αρχίζει με ένα αυθαίρετο αρχικό διαχωρισμό των αντικειμένων, αφορά την προσέγγιση του απλού περάσματος. Οι αντιπρόσωποι ομάδας υπολογίζονται για τα μέλη (σύνολα) του χωρίσματος, και τα αντικείμενα αναδιανέμονται στον κοντινότερο αντιπρόσωπο ομάδας.
Ένας τρίτος τύπος αλγόριθμου αντιπροσωπεύεται από την εργασία του Dattola41. Ο αλγόριθμός του, βασίζεται σε έναν προηγούμενο αλγόριθμο από τον Doyle. Όπως στην περίπτωση του MacQueen, ξεκινά με ένα αρχικά αυθαίρετο διαχωρισμό και ένα σύνολο αντιπροσώπων ομάδας. Μετά από κάθε αναδιανομή, ο αντιπρόσωπος ομάδας

υπολογίζεται ξανά, αλλά ο νέος αντιπρόσωπος ομάδας θα αντικαταστήσει τον παλιό, μόνο εάν ο νέος αντιπρόσωπος αποδειχθεί κοντινότερος υπό κάποια έννοια στα αντικείμενα στη νέα ομάδα από τον παλαιό αντιπρόσωπο. Ο αλγόριθμος του Dattola έχει χρησιμοποιηθεί εκτενώς από τον Murray42 για την παραγωγή ιεραρχικών ταξινομήσεων. Από την άλλη πλευρά η προσέγγιση του Crouch43 σχετίζεται με την προσέγγιση του Dattola. Ο Crouch αφιερώνει περισσότερο χρόνο στην δημιουργία του αρχικού χωρίσματος (τα ονομάζει κατηγορίες) και στους αντίστοιχους αντιπροσώπους ομάδας. Η αρχική φάση καλείται "στάδιο κατηγοριοποίησης", το οποίο ακολουθείται από το "στάδιο ταξινόμησης". Το δεύτερο στάδιο προχωρά να αναδιανείμει τα αντικείμενα με τον κανονικό τρόπο. Η εργασία του, είναι αρκετά ενδιαφέρουσα κυρίως λόγω των εκτενών συγκρίσεων που έκανε μεταξύ των αλγορίθμων Rocchio, Rieber και Marathe, Bonner και του δικού του.

Ένας ακόμα αλγόριθμος που πρέπει να αναφερθεί εδώ είναι αυτός του Litofsky. Ο αλγόριθμός του έχει σχεδιαστεί για να λειτουργεί μόνο για τα αντικείμενα που περιγράφονται από τις ιδιότητες δυαδικής κατάστασης. Χρησιμοποιεί τους αντιπροσώπους ομάδας και τις συναρτήσεις ταυτοποίησης με έναν εξ ολοκλήρου διαφορετικό τρόπο. Ο αλγόριθμος μεταθέτει τα αντικείμενα σε μία προσπάθεια να ελαχιστοποιηθεί ο μέσος αριθμός διαφορετικών ιδιοτήτων που υπάρχει στα μέλη κάθε ομάδας. Οι ομάδες χαρακτηρίζονται από τα σύνολα τιμών ιδιοτήτων όπου το κάθε σύνολο είναι το σύνολο των κοινών ιδιοτήτων για όλα τα μέλη της ομάδας. Η τελική ταξινόμηση είναι ιεραρχική. (για περαιτέρω λεπτομέρειες δείτε Lefkovitz44.)

Τέλος, ο αλγόριθμος Bonner45 πρέπει να αναφερθεί. Είναι ένα υβρίδιο των προσεγγίσεων γραφημάτων και ευρετικών (heuristic) προσεγγίσεων. Οι αρχικές ομάδες προσδιορίζονται με τις μεθόδους γραφημάτων (βασισμένες σε ένα μέτρο συσχέτισης), και έπειτα τα αντικείμενα αναδιανέμονται σύμφωνα με τους όρους στην συνάρτηση ταυτοποίησης.

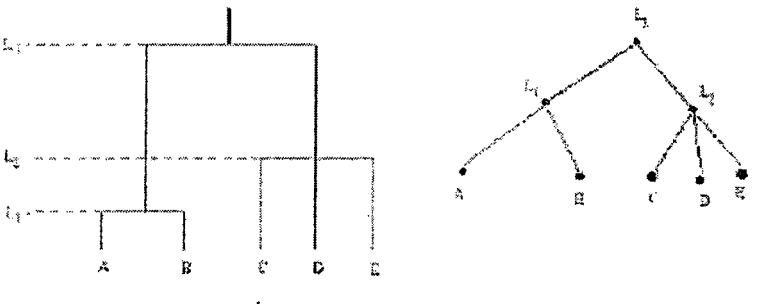
Το σημαντικότερο πλεονέκτημα των αλγοριθμικά καθορισμένων μεθόδων ομάδας είναι η ταχύτητά τους: O(nlogn) (όπου n είναι ο αριθμός αντικειμένων που θα ομαδοποιηθούν) έναντι του O(n2) για τις μεθόδους βασισμένες στα μέτρα συσχέτισης. Εντούτοις, έχουν μειονεκτήματα. Η τελική ταξινόμηση εξαρτάται από την σειρά με την οποία τα αντικείμενα εισάγονται στον αλγόριθμο αμάδας. Επιπρόσθετα τα αποτελέσματα των λαθών στις περιγραφές αντικειμένου είναι απρόβλεπτα.

Μια προφανής παράλειψη από τον κατάλογο μεθόδων ομαδοποίησης είναι η ομαδοποίηση των μεθόδων που βασίζονται στα μαθηματικά ή στη στατιστική όπως η ανάλυση παράγοντα (factor analysis) και η πιο πρόσφατη ανάλυση κατηγορίας (class analysis). Αν και, και οι δύο μέθοδοι χρησιμοποιήθηκαν αρχικά στο IR (δείτε Borko και Bernick46, Baker47) τώρα έχουν ουσιαστικά εκτοπιστεί από τις μεθόδους ομαδοποίησης που περιγράφονται πιο πάνω.

Η μέθοδος single-link αποφεύγει τα μειονεκτήματα που μόλις αναφέρθηκαν. Η καταλληλότητά της για την ομαδοποίηση των εγγράφων αναλύεται εδώ.

Single-link

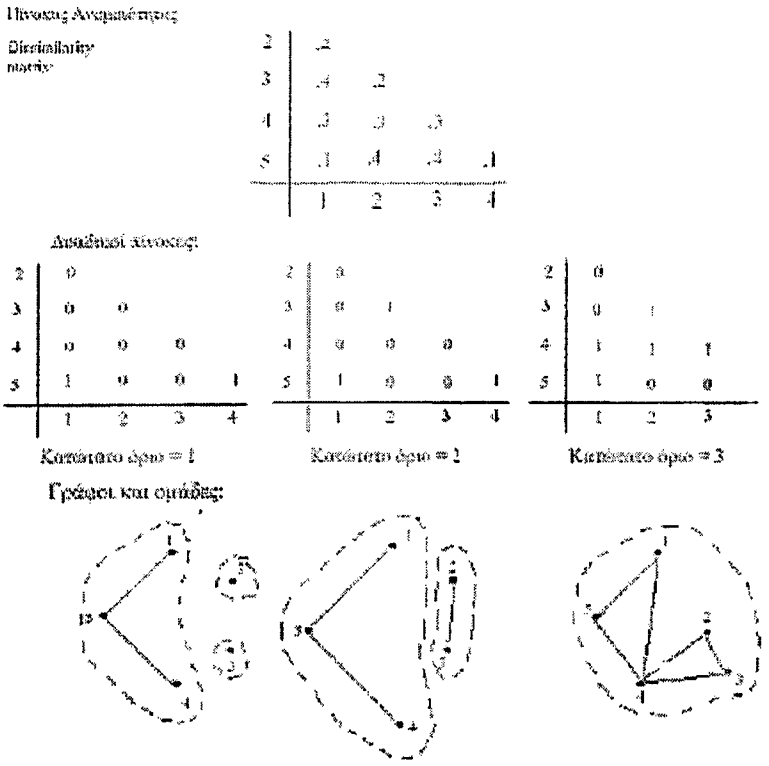
Ο συντελεστής ανομοιότητας είναι η βασική εισαγωγή σε έναν αλγόριθμο singlelink clustering. Το αποτέλεσμα είναι μια ιεραρχία με συσχετισμένα αριθμητικά επίπεδα, αποκαλούμενα δεντροδιαγράμματα. Συχνά η ιεραρχία αντιπροσωπεύεται από μια δομή δέντρων έτσι ώστε κάθε κόμβος να αντιπροσωπεύει μια ομάδα. Οι δύο αντιπροσωπεύσεις παρουσιάζονται δίπλα-δίπλα στο σχήμα 8 για το ίδιο σύνολο αντικειμένων A, B, C, D, E . Οι ομάδες είναι: {A,B} ,{C} ,{D} , {E} σε επίπεδο L1, {A,B}, {C, D, E} σε επίπεδο L2, και {A, B, C, D, E} σε επίπεδο L3. Σε κάθε επίπεδο της ιεραρχίας μπορεί να προσδιοριστεί ένα σύνολο κατηγοριών και καθώς προχωράμε στην ιεραρχία, οι κατηγορίες των χαμηλότερων επιπέδων τοποθετούνται στις κατηγορίες στα πιο υψηλά επίπεδα.



Σχήμα 8: Ένα δενδροδιάγραμμα με το αντίστοιχο δέντρο.

Για να γίνει καλύτερη κατανοητή μια ταξινόμηση single-link, υπάρχει ένα έτοιμο παράδειγμα (δείτε το σχήμα 9). Ένας συντελεστής ανομοιότητας, μπορεί να χαρακτηριστεί από ένα σύνολο γραφικών παραστάσεων, μια για κάθε τιμή που λαμβάνεται από τον συντελεστή ανομοιότητας. Οι διάφορες τιμές που λαμβάνονται από τον συντελεστή ανομοιότητας στο παράδειγμα είναι L = .1, .2, .3, .4. Η γραφική παράσταση σε κάθε επίπεδο δίνεται από ένα σύνολο κορυφών (vertices) που αντιστοιχούν στα αντικείμενα που θα ομαδοποιηθούν και οποιεσδήποτε δύο κορυφές (vertices) συνδέονται εάν η ανομοιότητά τους είναι το πολύ ίση με την τιμή του επιπέδου L. Πρέπει να είναι σαφές ότι αυτές οι γραφικές παραστάσεις χαρακτηρίζουν τον συντελεστή ανομοιότητας πλήρως. Λαμβάνοντας υπόψη τις γραφικές παραστάσεις και την ερμηνεία τους ένας συντελεστής ανομοιότητας μπορεί να ανακτηθεί, και αντίστροφα. Οι γραφικές παραστάσεις σε τιμές εκτός από εκείνες που λαμβάνονται από τον συντελεστή ανομοιότητας είναι απλά οι ίδιες όπως στην επόμενη μικρότερη τιμή που λαμβάνεται από τον συντελεστή ανομοιότητας, παραδείγματος χάριν, συγκρίνετε τις γραφικές παραστάσεις L = 1.5 και για L = .1.
Τώρα είναι απλό να καθοριστεί η single-link με βάση αυτές τις γραφικές παραστάσεις σε οποιοδήποτε επίπεδο. Ένα single-link cluster είναι ακριβώς το σύνολο των κορυφών (vertices) ενός συνδεδεμένου συστατικού της γραφικής παράστασης σε εκείνο το επίπεδο. Στο διάγραμμα φαίνεται κάθε ομάδα με μια μη συνεχή γραμμή.

Σημειώστε ότι ενώ οι γραφικές παραστάσεις σε οποιεσδήποτε δύο διακριτές τιμές που λαμβάνονται από τον συντελεστή ανομοιότητας θα είναι διαφορετικές, αυτό δεν συμβαίνει απαραίτητα για τις αντίστοιχες ομάδες σε εκείνα τα επίπεδα. Πιθανά, με την αύξηση του επιπέδου, οι συνδέσεις που εισάγαμε μεταξύ των κορυφών (vertices) δεν αλλάζουν το συνολικό αριθμό συνδεδεμένων κορυφών (vertices) σε ένα συστατικό.



Σχήμα 9: Πώς τα single-link clusters μπορούν να προέρχονται από τον συντελεστή ανομοιότητας, εφαρμόζοντας το κατώτατο όριο.

Για παράδειγμα, οι ομάδες στα επίπεδα .3 και .4 είναι τα ίδια. Η ιεραρχία επιτυγχάνεται μεταβάλλοντας το επίπεδο από τη χαμηλότερη πιθανή τιμή, αυξάνοντάς την μέσω διαδοχικών τιμών του συντελεστή ανομοιότητας, έως ότου περιλαμβάνονται όλα τα αντικείμενα σε μια ομάδα. Ο λόγος για την ονομασία single-link είναι τώρα προφανής: για να ανήκει ένα αντικείμενο σε μια ομάδα πρέπει να συνδεθεί μόνο με ένα άλλο μέλος της ομάδας.
Αυτή η περιγραφή οδηγεί αμέσως σε έναν ανεπαρκή αλγόριθμο για την παραγωγή των κατηγοριών single-link. Αποτελείται απλά από ορισμό του κατώτατου ορίου (thresholding) του συντελεστή ανομοιότητας σε αυξανόμενα επίπεδα ανομοιότητας. Ακολούθως, τα πλέγματα δυαδικής σύνδεσης (binary connection matrices) υπολογίζονται σε κάθε επίπεδο κατώτατων ορίων, από τα οποία τα συνδεδεμένα συστατικά μπορούν εύκολα να εξαχθούν. Αυτό είναι η βάση για πολλούς δημοσιευμένους αλγόριθμους single-link. Όσον αφορά την ανάκτηση πληροφορίας, όπου προσπαθούμε να κατασκευάσουμε ένα ερευνήσιμο (searchable) δέντρο είναι ανεπαρκείς (βλέπε van Rijsbergen48).

Η καταλληλότητα των διαστρωματωμένων ιεραρχικών μεθόδων cluster

Υπάρχουν πολλές άλλες ιεραρχικές μέθοδοι ομάδας, μερικές από τις οποίες είναι: πλήρης-σύνδεσης (complete-link), μέσης-σύνδεσης (average-link), κ.λπ. Εδώ θα δείξουμε την καταλληλότητά τους για την ανάκτηση εγγράφων. Χρειάζεται να συνειδητοποιήσουμε ότι το είδος ανάκτησης που θα κάνουμε είναι ένα στο οποίο ολόκληρη ομάδα ανακτάται χωρίς περαιτέρω επεξεργασία των εγγράφων στην ομάδα. Αυτό είναι αντίθετο με τις μεθόδους που προτείνονται από τους Rocchio, Litofsky, και Crouch οι οποίοι χρησιμοποιούν την ομαδοποίηση για να βοηθήσουν στον περιορισμό της έκτασης μιας γραμμικής αναζήτησης.
Τα διαστρωματωμένα συστήματα των ομάδων, είναι κατάλληλα επειδή το επίπεδο Μιας ομάδας μπορεί να χρησιμοποιηθεί σε στρατηγικές ανάκτησης ως παράμετρος ανάλογη της διάταξης (rank order) ή του κατώτατου ορίου της συνάρτησης ταυτοποίησης σε μια γραμμική αναζήτηση. Η ανάκτηση μιας ομάδας που είναι μια καλή αντιστοιχία για ένα αίτημα σε χαμηλό επίπεδο στην ιεραρχία, τείνει να δίνει υψηλή ακρίβεια αλλά χαμηλή ανάκληση, όπως το κατώτατο όριο σε μια χαμηλή θέση κατάταξης (rank position) σε μια γραμμική αναζήτηση τείνει να δίνει την υψηλή ακρίβεια αλλά τη χαμηλή ανάκληση. Ομοίως, η ανάκτηση ενός cluster που είναι μια καλή αντιστοιχία για ένα αίτημα σε ένα υψηλό επίπεδο στην ιεραρχία τείνει να δίνει υψηλή ανάκληση αλλά χαμηλή ακρίβεια. Τα ιεραρχικά συστήματα των ομάδων είναι κατάλληλα για τρεις λόγους. Κατ' αρχάς, πολύ αποδοτικές στρατηγικές μπορούν να επινοηθούν για να ψάξουν μία ιεραρχική ομαδοποίηση. Αφετέρου, η κατασκευή ιεραρχικών συστημάτων είναι πολύ γρηγορότερη από την κατασκευή μη ιεραρχικών (δηλαδή διαστρωματωμένων αλλά επικαλυπτόμενων) συστημάτων ομάδων. Τρίτον, οι απαιτήσεις αποθήκευσης για μια ιεραρχική δομή είναι αρκετά λιγότερες απ' ό,τι για μια μη ιεραρχική δομή, ιδιαίτερα κατά τη διάρκεια της φάσης ταξινόμησης.
Δεδομένου ότι οι ιεραρχικές μέθοδοι είναι κατάλληλες για την ομαδοποίηση των εγγράφων προκύπτει η ερώτηση: "Ποια μέθοδος;" Η απάντηση είναι ότι υπό ορισμένους όρους, η μόνη αποδεκτή διαστρωματωμένη ιεραρχική μέθοδος ομάδας είναι η single-link.

Single-link και το ελάχιστο spanning δέντρο

Το δέντρο single-link (όπως αυτό που παρουσιάζεται στο σχήμα 7) συσχετίζεται με ένα άλλο είδος δέντρου: το ελάχιστο spanning δέντρο, ή MST, που προέρχεται επίσης από έναν συντελεστή ανομοιότητας (Gower και Ross49). Αυτό το δεύτερο δέντρο είναι αρκετά διαφορετικό από το πρώτο, οι κόμβοι αντί να αντιπροσωπεύουν ομάδες, αντιπροσωπεύουν τα μεμονωμένα αντικείμενα θα ομαδοποιηθούν. Το MST, είναι το δέντρο του ελάχιστου μήκους που συνδέει τα αντικείμενα, όπου "μήκος " εννοούμε το άθροισμα των βαρών των connecting links στο δέντρο. Ομοίως μπορούμε να καθορίσουμε ένα μέγιστο spanning δέντρο ως ένα με μέγιστο μήκος. Εάν ενδιαφερόμαστε για ένα ελάχιστο ή μέγιστο spanning δέντρο εξαρτάται εξ ολοκλήρου από την εφαρμογή που εμείς έχουμε κατά νου. Για ευκολία θα επικεντρωθούμε στο ελάχιστο spanning δέντρο δεδομένου ότι προέρχεται φυσικά από έναν συντελεστή

αναμοιότητας και είναι πιο συχνό. Λαμβάνοντας υπόψη το ελάχιστο spanning δέντρο τότε οι single-link clusters αποκτούνται με τη διαγραφή των συνδέσεων από το MST κατά φθίνον μήκος, τα συνδεδεμένα σύνολα μετά από κάθε διαγραφή είναι τα singlelink clusters. Η σειρά της διαγραφής και η δομή του MST εξασφαλίζουν ότι τα clusters θα τοποθετηθούν σε μια ιεραρχία.

Το MST περιέχει περισσότερες πληροφορίες από την ιεραρχία single-link και μόνο έμμεσα πληροφορίες για τα single-link clusters. Κατά συνέπεια, αν και μπορούμε να εξάγουμε την ιεραρχία τα single-link από αυτό με μια απλή διαδικασία κατώτατου ορίου(thresholding), δεν μπορούμε να αντιστρέψουμε αυτή την διαδικασία και να εξάγουμε μεμονωμένα το MST από την ιεραρχία single-link. Με βάση αυτό, θα ήταν ενδιαφέρον να δούμε εάν το MST θα ήταν καταλληλότερο για το clustering των εγγράφων, από την ιεραρχία single-link. Δυστυχώς, δεν φαίνεται να είναι δυνατόν να ενημερωθεί ένα spanning δέντρο δυναμικά. Το να προστεθεί ένα νέο αντικείμενο σε μια single-link ιεραρχία είναι σχετικά απλό αλλά το να προστεθεί σε ένα MST είναι πιο περίπλοκο.

Η αντιπροσώπευση της ιεραρχίας single-link μέσω ενός MST έχει αποδειχθεί πολύ χρήσιμη στη σύνδεση του single-link με άλλες τεχνικές ομαδοποίησης50. Παραδείγματος χάριν, οι Boulton και Wallace51 έχουν δείξει, χρησιμοποιώντας την αντιπροσώπευση MST, ότι κάτω από ορισμένες υποθέσεις η ιεραρχία single-link μπορεί να ελαχιστοποιήσει το μέτρο πληροφορίας της ταξινόμησής τους. Είναι ενδιαφέρον ότι το MST, ανεξάρτητα από την εργασία τους, έχει χρησιμοποιηθεί για να μειώσει την αποθήκευση, όταν αποθηκεύουμε περιγραφές αντικειμένων.

4. Λογική ή Φυσική Οργάνωση και Ανεξαρτησία Δεδομένων

Υπάρχει ένας σημαντικός διαχωρισμός μεταξύ των δομών αρχείου. Αυτός ο διαχωρισμός είναι η διαφορά μεταξύ της λογικής και φυσικής οργάνωσης των δεδομένων. **Η λογική δομή των δεδομένων είναι οι σχέσεις που θα υπάρξουν μεταξύ των στοιχείων δεδομένων ανεξάρτητα από τον τρόπο με τον οποίο αυτές οι σχέσεις μπορούν πραγματικά να γίνουν κατανοητές στο εσωτερικό οποιουδήποτε υπολογιστή. Η φυσική οργάνωση ενδιαφέρεται περισσότερο για τη βελτιστοποίηση της χρήσης του μέσου αποθήκευσης δεδομένων όταν μία συγκεκριμένη λογική δομή αποθηκεύεται πάνω ή μέσα σε αυτό.** Χαρακτηριστικά για κάθε μονάδα φυσικής αποθήκευσης θα υπάρχει ένας αριθμός μονάδων λογικής δομής (πιθανώς εγγραφές) για να αποθηκευτούν σε αυτή. Παραδείγματος χάριν, εάν επρόκειτο να αποθηκεύσουμε μια δομή δέντρου σε ένα μαγνητικό δίσκο, η φυσική οργάνωση θα ενδιαφερόταν για τον καλύτερο τρόπο πακεταρίσματος των κόμβων του δέντρου πάνω στο δίσκο δεδομένων των χαρακτηριστικών προσπέλασης του δίσκου. Η εργασία πάνω στις βάσεις δεδομένων έχει μεγάλη σχέση με μια έννοια που αποκαλείται **ανεξαρτησία δεδομένων**. Ο στόχος αυτής της εργασίας είναι να επιτρέπει στα προγράμματα να γράφονται ανεξάρτητα από τη λογική δομή των δεδομένων με τα οποία θα αλληλεπιδρούσαν. Αν μια δομή αρχείου αλλάξει και από ανατραμμένο γίνει σειριακό αρχείο, τότε το πρόγραμμα πρέπει να παραμείνει το ίδιο. Αυτή η ανεξαρτησία επιτυγχάνεται παρεμβάλλοντας ένα **μοντέλο δεδομένων** μεταξύ του χρήστη και της βάσης δεδομένων. Έτσι, ο χρήστης βλέπει το μοντέλο δεδομένων αντί της βάσης δεδομένων και ολόκληρο το πρόγραμμά του επικοινωνεί με το μοντέλο. Για το λόγο αυτό ο χρήστης δεν ενδιαφέρεται για τη δομή του αρχείου.

Παρ' όλ' αυτά, αξίζει να ληφθεί σοβαρά υπ' όψιν η τάση απομάκρυνσης των χρηστών από τη γνώση των δομών αρχείων, μια τάση που έχει υποκινηθεί αρκετά από τις προσπάθειες κατασκευής μιας θεωρίας δεδομένων. Υπάρχουν διάφορες προτάσεις για ανταλλαγή δεδομένων σε αφηρημένο επίπεδο. Η πιο γνωστή από αυτές είναι αυτή που αποκαλείται **σχεσιακό μοντέλο**. Σε αυτήν τα δεδομένα περιγράφονται από n – αόριστους αριθμούς των τιμών ιδιοτήτων. Πιο τυπικά, εάν τα δεδομένα περιγράφονται από σχέσεις, μία σχέση σε ένα σύνολο από πεδία $D_1,..., D_n$ μπορεί να αντιπροσωπευθεί από ένα σύνολο n διαταγμένων αόριστων αριθμών κάθε ένας με τη μορφή $(d_1,..., d_n)$ όπου $d_i \varepsilon D_i$.

Μία δεύτερη προσέγγιση είναι **το ιεραρχικό μοντέλο** η οποία χρησιμοποιείται σε πολλά υπάρχοντα συστήματα βάσης δεδομένων. Αυτή η προσέγγιση λειτουργεί όπως κάποιος θα περίμενε: τα δεδομένα παριστάνονται με τη μορφή ιεραρχιών. Αν και αυτή η προσέγγιση είναι πιο περιοριστική από την σχεσιακή προσέγγιση συχνά φαίνεται να είναι ο φυσικός τρόπος για να προχωρήσει. Μπορεί να υποστηριχτεί ότι σε πολλές εφαρμογές μια ιεραρχική δομή είναι μια καλή προσέγγιση για τη φυσική δομή των δεδομένων και ότι η συνεπαγόμενη απώλεια στην ακρίβεια της αναπαράστασης αξίζει το κέρδος στην αποδοτικότητα και την απλότητα της αναπαράστασης.

Η τρίτη προσέγγιση είναι το **δικτυακό μοντέλο**. Εδώ τα στοιχεία δεδομένων είναι συνδεδεμένα σε ένα δίκτυο στο οποίο κάθε δεδομένη σύνδεση μεταξύ δύο στοιχείων υπάρχει γιατί ικανοποιεί κάποιο όρο όσο αφορά τα χαρακτηριστικά αυτών των στοιχείων, για παράδειγμα, μοιράζονται ένα χαρακτηριστικό γνώρισμα. Είναι πιο γενική από την ιεραρχική προσέγγιση από την άποψη ότι ένας κόμβος μπορεί να έχει

οποιοδήποτε αριθμό από άμεσους ανώτερους. Επίσης είναι ισοδύναμη με τη σχεσιακή προσέγγιση στην ικανότητα απεικόνισης.
Τα πλεονεκτήματα και τα μειονεκτήματα κάθε προσέγγισης συζητούνται πολύ λεπτομερώς. Υπάρχει επίσης μια πρόσφατη έρευνα υπολογισμού που αναθεωρεί την τρέχουσα κατάσταση προόδου. Έχουν υπάρξει μερικοί πρώτοι υπερασπιστές της συγγενικής προσέγγισης στο IR, από το 1967 οι Maron53 και Levien54 ζήτησαν το σχέδιο και την εφαρμογή ενός συστήματος IR μέσω των σχέσεων, είτε πρόκειται για δυαδικοί. Επίσης οι Prywes και Smith στο κεφάλαιο αναθεώρησής τους στην ετήσια αναθεώρηση της επιστήμης και της τεχνολογίας των πληροφοριών σύστησαν πιο πρόσφατα τις προτάσεις DBTG ως τρόπους για τα συστήματα ανάκτησης πληροφορίας.
Να κρυφτεί στο υπόβαθρο οποιασδήποτε συζήτησης των δομών αρχείων είναι σήμερα πάντα η ερώτηση εάν η τεχνολογία βάσεων δεδομένων θα προσπεράσει όλων. Κατά συνέπεια μπορεί να είναι ότι οποιαδήποτε εφαρμογή στον τομέα της ανάκτησης αυτοματοποίησης και πληροφοριών βιβλιοθηκών θα εφαρμοστεί μέσω της χρήσης κάποιας κατάλληλης συσκευασίας βάσεων δεδομένων. Αυτό είναι βεβαίως μια δυνατότητα αλλά μη πιθανό να συμβεί στο εγγύς μέλλον. Υπάρχουν διάφοροι λόγοι. Κάποιος είναι ότι τα συστήματα βάσεων δεδομένων είναι συστήματα γενικού σκοπού ενώ τα αυτοματοποιημένα συστήματα βιβλιοθηκών και ανάκτησης είναι ειδικός σκοπός. Κανονικά κάποιος καταβάλλει μια τιμή για τη γενικότητα και σε αυτήν την περίπτωση είναι ακόμα πάρα πολύ μεγάλο. Αφετέρου, τώρα υπάρχει μια ιδιαίτερη επένδυση στην παροχή των ειδικών συστημάτων σκοπού (παραδείγματος χάριν, MARC) και αυτό δεν γράφεται από πολύ εύκολα. Εντούτοις μια τάση προς την αύξηση της χρήσης της τεχνολογίας βάσεων δεδομένων υπάρχει και εμφανίζεται καλά στην αυξανόμενη προεξοχή που δίνεται σε το στην ετήσια αναθεώρηση της επιστήμης και της τεχνολογίας των πληροφοριών.

Μια Γλώσσα Για Την Περιγραφή Των Δομών Αρχείου

Όπως όλα τα θέματα στην πληροφορική έτσι και η ορολογία των δομών αρχείου έχει εξελιχθεί με ακατάστατο τρόπο χωρίς πολύ ενδιαφέρον για τη συνέπεια, την ασάφεια, ή το εάν ήταν δυνατό να γίνει το είδος των διακρίσεων που ήταν σημαντικές. Η ανάγκη για μία σαφώς προσδιορισμένη και αμφιλεγόμενη γλώσσα που να περιγράφει τις δομές των αρχείων άργησε κατά πολύ να γίνει φανερή. Ειδικότερα, προέκυψε μια ανάγκη για διαβίβαση ιδεών σχετικά με τις δομές αρχείου χωρίς όμως να πρέπει να εμπλακούν μελέτες υλικού.

Βασική Ορολογία

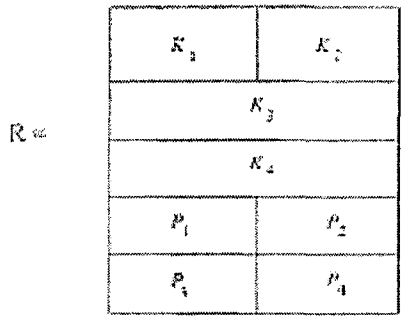
Δεδομένων ενός συνόλου «ιδιοτήτων» A και ενός συνόλου «τιμών» V, τότε μία εγγραφή R είναι ένα υποσύνολο του καρτεσιανού γινομένου A x V στο οποίο κάθε ιδιότητα έχει μία και μόνο μία τιμή.
Έτσι R είναι ένα σύνολο από ταξινομημένα ζεύγη της μορφής (ιδιότητα, τιμή ιδιότητας). Σε μαθηματική περιγραφή, μια εγγραφή ορίζεται ως

R = {(K1, x1), (K2, x2), . . . (Km, xm)}

Στον παραπάνω συμβολισμό τα Ki είναι λέξεις κλειδιά, που λειτουργούν σαν χαρακτηριστικά και η τιμή xi μπορεί να θεωρηθεί αριθμητική αξία. Συχνά έγγραφα χαρακτηρίζονται απλά από την παρουσία ή απουσία λέξεων κλειδιών. Σε αυτή την περίπτωση γράφουμε:
R = {Kt1, Kt2, . . . , Kti}

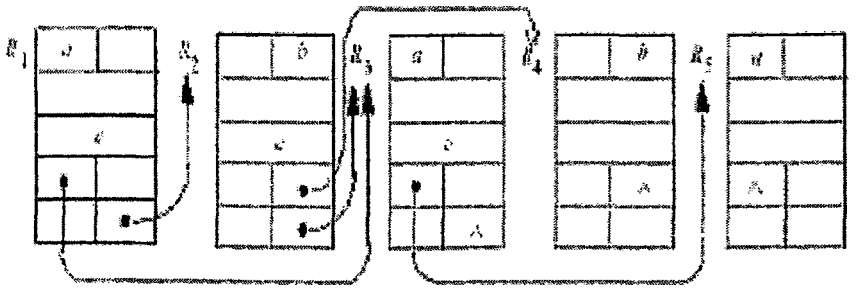
Όπου Kt1 υπάρχει όταν xt1=1 , αλλιώς δεν υπάρχει

Οι εγγραφές συγκεντρώνονται σε λογικές μονάδες που καλούνται αρχεία. Μπορεί κάποιος να αναφερθεί σε ένα σύνολο εγγραφών ονομαστικά, με το όνομα του αρχείου. Οι εγγραφές μέσα σε ένα αρχείο συχνά οργανώνονται σύμφωνα με τις μεταξύ τους σχέσεις. Αυτή η λογική οργάνωση έχει γίνει γνωστή ως δομή αρχείου (ή δομή δεδομένων).



Σχήμα 10: Ένα παράδειγμα μιας εγγραφής με συσχετιομένα πεδία.

Τα πεδία δεν είναι απαραίτητα σταθερά σε μήκος. Για να βρεθεί η τιμή του χαρακτηριστικού K4, πρώτα βρίσκεται η διεύθυνση της εγγραφής R (η οποία στην πραγματικότητα είναι η διεύθυνση της έναρξης της εγγραφής) και διαβάζονται τα δεδομένα στο τέταρτο πεδίο.
Στο ίδιο σχήμα φαίνονται κάποια πεδία με το όνομα Pi. Είναι διευθύνσεις άλλων εγγραφών και συχνά καλούνται δείκτες. Έχουμε επεκτείνει τον ορισμό μιας εγγραφής σε ένα σύνολο ζευγών χαρακτηριστικών – τιμών και δεικτών. Κάθε δείκτης συνήθως συσχετίζεται με ένα συγκεκριμένο ζεύγος χαρακτηριστικού - τιμής. Για παράδειγμα, δείκτες θα μπορούσαν να χρησιμοποιηθούν για να συνδέσουν όλες τις εγγραφές για τις οποίες η τιμή x1 (του χαρακτηριστικού K1) είναι α, παρομοίως για x2 ίσον με b, κλπ. Η κατάσταση αυτή απεικονίζεται στο σχήμα 10.



Σχήμα 11: Μία αναπαράσταση της χρήσης δεικτών για την σύνδεση εγγραφών.

Για να δείξουμε ότι μια εγγραφή είναι η τελευταία εγγραφή που δείχνουμε σε μια λίστα εγγραφών, χρησιμοποιούμε το μηδενικό δείκτη. Ο δείκτης που σχετίζεται με το χαρακτηριστικό K στην εγγραφή R θα ονομάζεται K δείκτης. Ένα χαρακτηριστικό (λέξη κλειδί) που χρησιμοποιείται με αυτόν τον τρόπο για να οργανώσει ένα αρχείο, ονομάζεται κλειδί.
Βασιζόμενοι στους Hsiao και Harary, ορίζουμε μια λίστα εγγραφών L, με βάση μια λέξη κλειδί K, ή για συντομία μια K λίστα, ως μια ομάδα εγγραφών που περιέχουν το K έτσι ώστε:
(1) οι K δείκτες είναι διακριτοί
(2) κάθε μη μηδενικός K δείκτης στην L δίνει την διεύθυνση μιας εγγραφής στην L
(3) υπάρχει ένα μοναδικό αρχείο στην L, που καμία εγγραφή που περιέχει το K δεν το δείχνει, ονομάζεται αρχή της λίστας και
(4) υπάρχει ένα μοναδικό αρχείο στην L που περιέχει τον μηδενικό K δείκτη, είναι το τέλος της λίστας.

Από το προηγούμενο παράδειγμά μας:

K_1 - list : R_1, R_2, R_5
K_2 - list: R_2, R_4
K_4 - list : R_1, R_2, R_3

Τέλος, χρειαζόμαστε τον ορισμό του καταλόγου ενός αρχείου. Έστω ότι F είναι ένα αρχείο του οποίου οι εγγραφές περιέχουν ακριβώς m διαφορετικές λέξεις κλειδιά K1, K2, . . . , Km. Έστω, επίσης, ότι ni είναι ο αριθμός των εγγραφών που περιέχουν τη λέξη κλειδί Ki και ότι hi είναι ο αριθμός των Ki-λιστών μέσα στο F. Επιπροσθέτως, με aij συμβολίζουμε την αρχική διεύθυνση της j-οστής Ki-λίστας. Τότε ο κατάλογος είναι το σύνολο των ακολουθιών
(Ki, ni, hi, ai1, ai2, . . . aihi) i = 1, 2, . . . m

Διαδοχικά Αρχεία (Sequential files)

Ένα διαδοχικό αρχείο είναι το πιο πρωτόγονο από όλες τις δομές αρχείου. Δεν έχει ούτε κατάλογο ούτε δείκτες σύνδεσης. Οι εγγραφές γενικά οργανώνονται με λεξικογραφική σειρά στην τιμή κάποιου κλειδιού. Με άλλα λόγια, επιλέγεται μια συγκεκριμένη ιδιότητα της οποίας η τιμή θα καθορίσει τη σειρά των εγγραφών. Μερικές φορές όταν η τιμή της ιδιότητας είναι σταθερή για έναν μεγάλο αριθμό εγγραφών επιλέγεται ένα δεύτερο κλειδί για ταξινόμηση όταν το πρώτο κλειδί αποτυγχάνει. Η εφαρμογή αυτής της δομής αρχείου απαιτεί τη χρήση μιας ρουτίνας ταξινόμησης.

Τα κύρια πλεονεκτήματα αυτής της δομής είναι:

(1) είναι εύκολο να χρησιμοποιηθεί,
(2) παρέχει γρήγορη πρόσβαση στην επόμενη εγγραφή χρησιμοποιώντας λεξικογραφική σειρά.

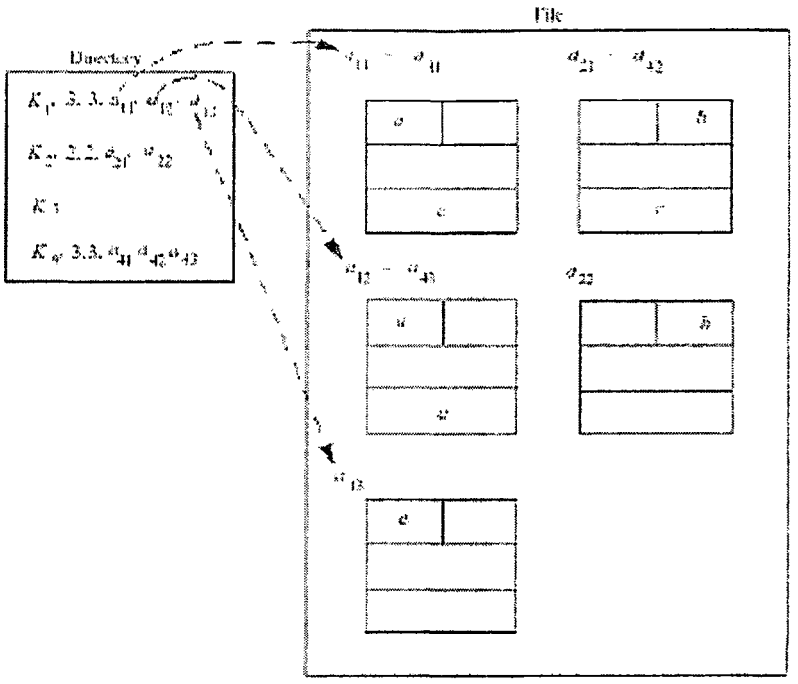
Τα μειονεκτήματα αυτής της δομής είναι:

(1) η ενημέρωση είναι δύσκολη – η εισαγωγή μιας νέας εγγραφής μπορεί να απαιτεί τη μετακίνηση ενός μεγάλου μέρους του αρχείου,
(2) η τυχαία προσπέλαση είναι εξαιρετικά αργή.

Μερικές φορές ένα αρχείο θεωρείται ότι πρέπει να οργανωθεί διαδοχικά παρά το γεγονός ότι δεν ταξινομείται σύμφωνα με κάθε κλειδί. Ίσως να θεωρείται ότι η ημερομηνία απόκτησης είναι η τιμή του κλειδιού και ότι οι νέες εισαγωγές προστίθενται στο τέλος του αρχείου οπότε η ανανέωση δεν είναι δύσκολη.

Αντιστρεφόμενα Αρχεία (inverted files)

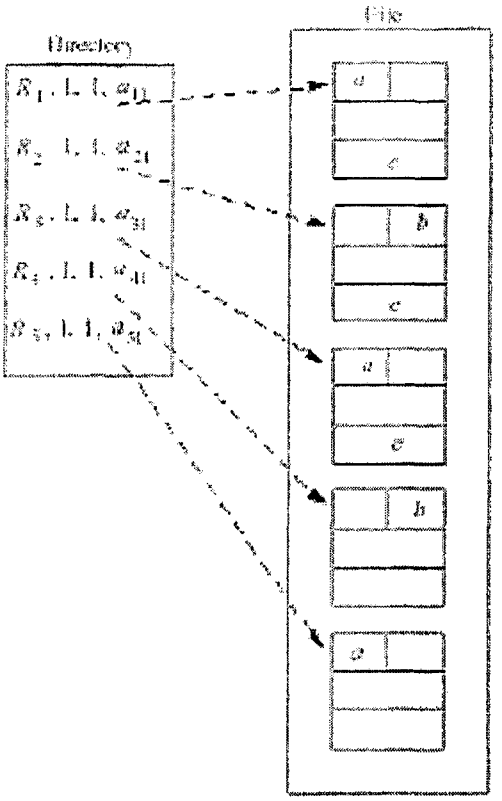
Ένα αντιστρεφόμενο αρχείο (inverted file) είναι μία δομή αρχείου στην οποία κάθε λίστα περιέχει μόνο μία εγγραφή. Υπενθυμίζουμε ότι μία λίστα προσδιορίζεται σε σχέση με μία λέξη κλειδί K, έτσι κάθε K-λίστα περιέχει μόνο μία εγγραφή. Αυτό συνεπάγεται ότι ο κατάλογος θα είναι τέτοιος ώστε ni = hi για όλα τα i, που σημαίνει ότι, ο αριθμός των εγγραφών που περιέχουν την Ki θα είναι ίσος με τον αριθμό των Ki-λιστών. Έτσι ο κατάλογος θα έχει μία διεύθυνση για κάθε εγγραφή που περιέχει την Ki. Για την ανάκτηση εγγράφων αυτό σημαίνει ότι δεδομένης μιας λέξης κλειδί Μπορούμε αμέσως να εντοπίσουμε τις διευθύνσεις όλων των εγγράφων που περιέχουν αυτή τη λέξη κλειδί.
Όσον αφορά το προηγούμενο παράδειγμα, ας υποθέσουμε ότι μία non-black εισαγωγή στο πεδίο που αντιστοιχεί σε ένα χαρακτηριστικό, δηλώνει την παρουσία μιας λέξης κλειδί και μία black εισαγωγή, την απουσία της. Τότε ο κατάλογος αρχείων θα δείχνει στο αρχείο την κατεύθυνση που φαίνεται στο σχήμα 12.
Ο ορισμός ενός αντιστρεφόμενου αρχείου δεν απαιτεί οι διευθύνσεις μέσα στον κατάλογο να είναι ταξινομημένες. Ωστόσο, για να διευκολύνονται πράξεις όπως η ένωση («and») και η διάζευξη («or») πάνω σε δύο οποιεσδήποτε αντιστρεφόμενες λίστες, οι διευθύνσεις συνήθως διατηρούν τη σειρά εγγραφής. Αυτό σημαίνει ότι οι πράξεις «and» και «or» μπορούν να εκτελεστούν με ένα πέρασμα και στις δύο λίστες. Βέβαια το τίμημα που πληρώνουμε είναι ότι η ανανέωση του αντιστρεφόμενου αρχείου γίνεται πιο αργή.



Σχήμα 12: Ένα αντιστρεφόμενο αρχείο

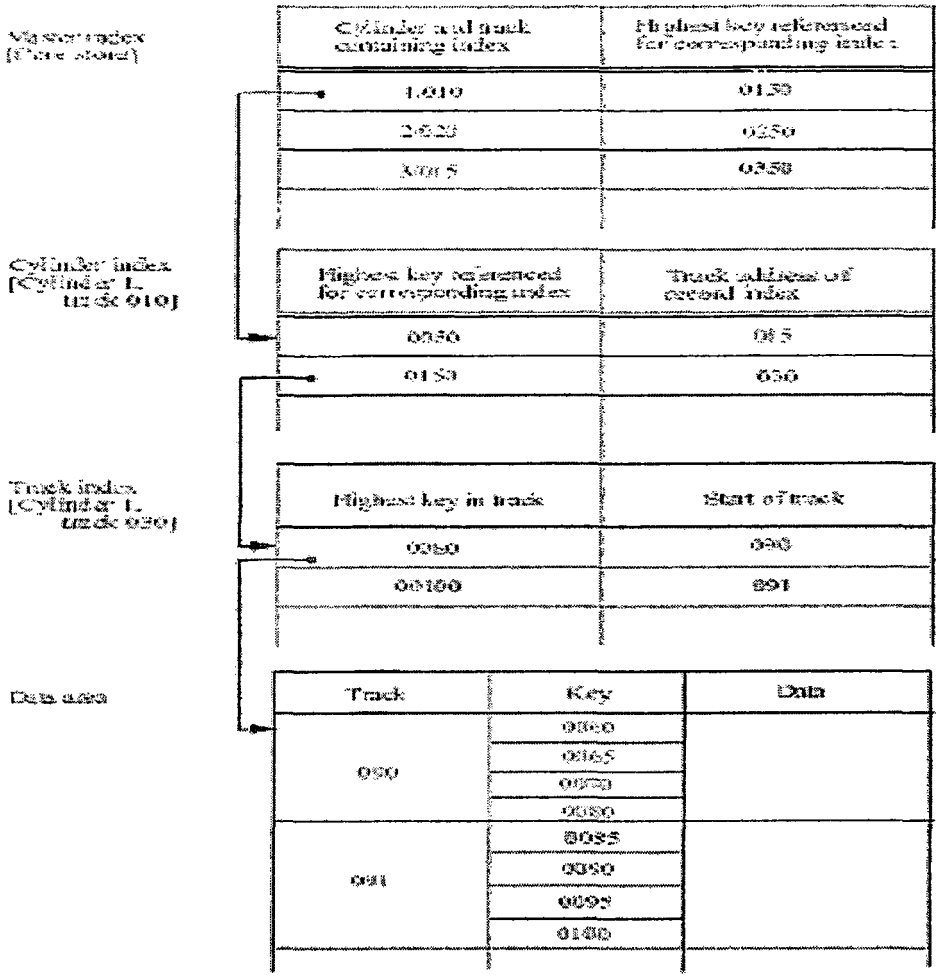
Ακολουθιακά αρχεία δεικτών (Index-sequential files)

Ένα ακολουθιακό αρχείο δεικτών (index-sequential file) είναι ένα αντεστραμμένο αρχείο, στο οποίο για κάθε λέξη κλειδί Ki έχουμε ni = hi = 1 και a11 <a21 . . . <am1. Αυτό θα συμβεί μόνο εάν κάθε εγγραφή έχει μόνο μία μοναδική λέξη κλειδί, ή ένα μοναδικό χαρακτηριστικό – τιμή (attribute-value). Στην πραγματικότητα, λοιπόν, αυτή η ομάδα αρχείων μπορεί, να ταξινομηθεί διαδοχικά με ένα κλειδί. Κάθε τιμή κλειδιού εμφανίζεται στον κατάλογο αρχείων με τη συσχετισμένη διεύθυνση της εγγραφής του (key value). Μια προφανής ερμηνεία ενός κλειδιού αυτού του είδους θα ήταν ο αριθμός της εγγραφής. Στο παράδειγμά μας καμία από τις ιδιότητες δεν θα έκανε αυτήν την εργασία εκτός από τον αριθμό εγγραφής. Διαγραμματικά το ακολουθιακό αρχείο δεικτών επομένως θα εμφανιζόταν όπως φαίνεται στο σχήμα 13. Έχουμε βάλει σκόπιμα Ri αντί Ki για να υπογραμμιστεί η φύση του κλειδιού.



Σχήμα 13 :Ένα ακολουθιακό αρχείο δεικτών

Στη βιβλιογραφία ένα αρχείο συχνά θεωρείται ως ένα διαδοχικό αρχείο με μία ιεραρχία καταλόγων. Αυτό δεν έρχεται σε αντίθεση με τον προηγούμενο ορισμό, απλά περιγράφει τον τρόπο με τον οποίο ο κατάλογος αρχείων χρησιμοποιείται. Δεν είναι περίεργο επομένως ότι οι δείκτες είναι συχνά προσανατολισμένοι στα χαρακτηριστικά του μέσου αποθήκευσης. Παραδείγματος χάριν (δείτε το σχήμα 14) θα μπορούσαν να **υπάρχουν τρία επίπεδα ευρετηριοποίησης: δείκτης ίχνους, δείκτης κυλίνδρου και κύριος δείκτης**. Κάθε εισαγωγή στο δείκτη διαδρομής θα περιέχει αρκετές πληροφορίες για να εντοπίσει την έναρξη της διαδρομής, και το κλειδί της τελευταίας εγγραφής στη διαδρομή, το οποίο είναι επίσης η υψηλότερη τιμή σε αυτή τη διαδρομή. Υπάρχει ένας δείκτης διαδρομής για κάθε κύλινδρο. Κάθε εισαγωγή στο δείκτη κυλίνδρων δίνει την τελευταία εγγραφή σε κάθε κύλινδρο και τη διεύθυνση του δείκτη διαδρομής για εκείνο τον κύλινδρο. Εάν ο ίδιος ο δείκτης κυλίνδρων αποθηκεύεται στις διαδρομές, τότε ο κύριος δείκτης θα δώσει το υψηλότερο κλειδί που παραπέμπεται για κάθε διαδρομή του δείκτη κυλίνδρων και την αρχική διεύθυνση εκείνης της διαδρομής.



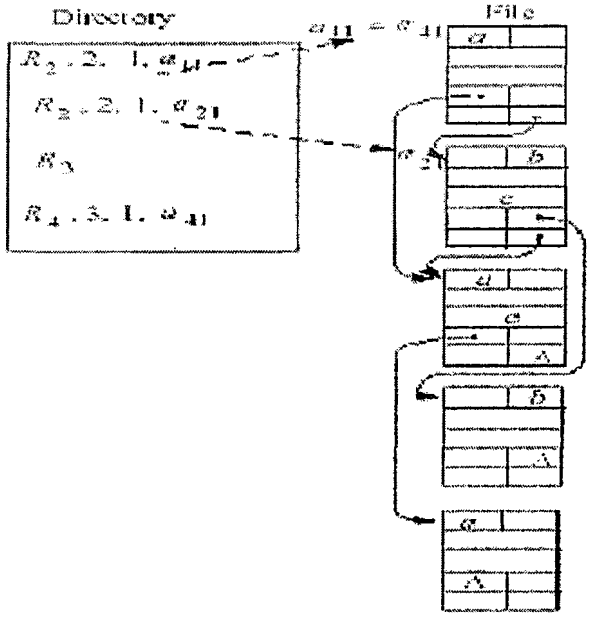
Σχήμα 14: Ένα παράδειγμα μιας εφαρμογής ενός ακολουθιακού αρχείου δεικτών

Καμία αναφορά δεν έχει γίνει στην πιθανότητα υπερχείλισης κατά την διάρκεια της διαδικασίας ενημέρωσης. Κανονικά μέτρα λαμβάνονται στον κατάλογο αρχείων για την διαχείριση μιας περιοχής υπερχείλισης. Αυτό αυξάνει φυσικά τον αριθμό καταχωρήσεων book-keeping σε κάθε εισαγωγή του δείκτη.

Πολλαπλές λίστες (Multi-lists)

Μία πολλαπλή λίστα (multi-list) είναι στην πραγματικότητα μία μικρή τροποποίηση του αντιστρεφόμενου αρχείου. Υπάρχει μία λίστα για κάθε λέξη κλειδί, i.e. hi=1. Τα αρχεία που περιέχουν μια συγκεκριμένη λέξη κλειδί Ki συσχετίζονται μεταξύ τους για να διαμορφώσουν τον κατάλογο Ki και η ένωρξη του Ki καταλόγου δίνεται στον κατάλογο των αρχείων όπως φαίνεται στο σχήμα 15. Δεδομένου ότι δεν υπάρχει K3 κατάλογος, το πεδίο που διατηρήθηκε για το δείκτη του θα μπορούσε να έχει παραλειφθεί. Οποιοσδήποτε κενός τομέας δεικτών, εφ' όσον δεν προκύπτει καμία

ασάφεια ως προς ποιο δείκτη ανήκει σε ποια λέξη κλειδί θα μπορούσε επίσης να παραληφθεί. Ένας τρόπος να εξασφαλιστεί αυτό, ιδιαίτερα εάν οι τιμές των στοιχείων (χαρακτηριστικό - τιμές) χαρακτηρίζονται από μια συγκεκριμένου τύπου διαμόρφωση (fixed format), είναι ο δείκτης να μην δείχνει την αρχή του αρχείου αλλά να δείχνει τη θέση του επομένου δείκτη στην αλυσίδα.
Η πολλαπλή λίστα σχεδιάστηκε για να ξεπεράσει τις δυσκολίες ενημέρωσης ενός αντιστρεφόμενου αρχείου. Σε ένα αντιστρεφόμενο αρχείο οι διευθύνσεις συνήθως διατηρούν τη σειρά του αριθμού εγγραφής. Αλλά, όταν έρχεται η στιγμή μία νέα εγγραφή να προστεθεί στο αρχείο, αυτή η σειρά πρέπει να διατηρηθεί και η εισαγωγή της νέας διεύθυνσης μπορεί να είναι δαπανηρή. Τέτοιο πρόβλημα δεν προκύπτει με την πολλαπλή λίστα, καθώς ενημερώνουμε τις κατάλληλες Κ-λίστες συνδέοντας αυτές στη νέα εγγραφή. Το τίμημα που πληρώνουμε γι' αυτό είναι βέβαια η αύξηση του χρόνου αναζήτησης. Στην πραγματικότητα όμως, αυτό είναι χαρακτηριστικό όλων των δομών αρχείου. Είναι σύμφυτη στο σχεδιασμό τους η εξισωτική σχέση μεταξύ του χρόνου αναζήτησης και του χρόνου ενημέρωσης.



Σχήμα 15: Μία πολλαπλή λίστα (multi-list)

Κυψελικές πολλαπλές λίστες (Cellular multi-lists)

Μία περαιτέρω τροποποίηση της πολλαπλής λίστας είναι εμπνευσμένη από το γεγονός ότι πολλά αποθηκευτικά μέσα διαιρούνται σε σελίδες, οι οποίες μπορούν να ανακτηθούν μία κάθε φορά. Μία Κ-λίστα μπορεί να διασχίσει διάφορα όρια σελίδων που σημαίνει ότι πρέπει να προσπελαθούν αρκετές σελίδες για να ανακτηθεί μία εγγραφή. Μία τροποποιημένη δομή πολλαπλής λίστας που επιτρέπει κάτι τέτοιο καλείται κυψελική πολλαπλή λίστα. Οι Κ-λίστες είναι περιορισμένου μεγέθους και γι' αυτό δεν θα διαπεράσουν τα όρια της σελίδας.
Ο κατάλογος για μία κυψελική πολλαπλή λίστα θα είναι ένα σύνολο των ακολουθιών

(Ki, ni, hi, aii, . . . aihi) i = 1, 2, . . . , m

όπου το hi έχει επιλεγεί τέτοιο ώστε να εξασφαλίζει ότι μία Ki-λίστα δεν θα διαπεράσει το όριο μιας σελίδας. Σε μια εφαρμογή όπως ακριβώς στην εφαρμογή ενός ακολουθιακού αρχείου δεικτών (index-sequential file), περαιτέρω πληροφορίες θα αποθηκευθούν με κάθε διεύθυνση, έτσι ώστε να επιτραπεί ο εντοπισμός της σωστής σελίδας για κάθε τιμή κλειδιού.

Δομές Δακτυλίων (Ring structures)

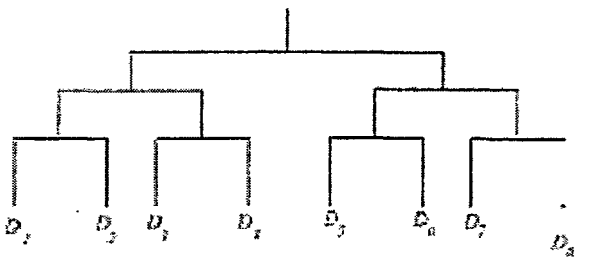
Ένας δακτύλιος είναι απλά μία γραμμική λίστα που κλείνει πάνω στην ίδια. Σύμφωνα με τον ορισμό μιας Κ-λίστας, η αρχή και το τέλος της λίστας είναι η ίδια εγγραφή. Αυτή η δομή δεδομένων είναι ιδιαίτερα χρήσιμη στην παρουσίαση της ταξινόμησης των δεδομένων.

Ας υποθέσουμε ότι μια ομάδα εγγράφων
{*D*1, *D*2, *D*3, *D*4, *D*5, *D*6, *D*7, *D*8}

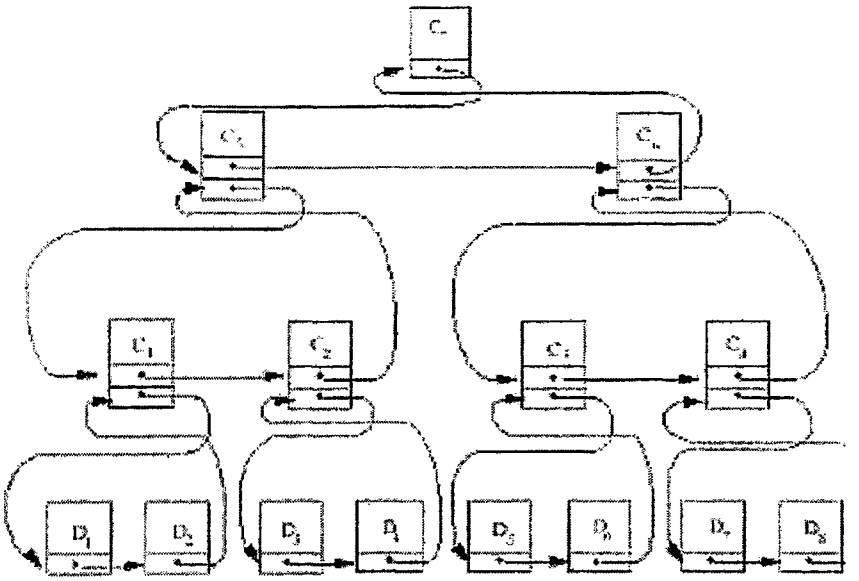
έχει κατηγοριοποιηθεί σε 4 ομάδες, δηλαδή
{(*D*1, *D*2), (*D*3, *D*4), (*D*5, *D*6), (*D*7, *D*8)}

Επιπλέον αυτές οι ομάδες έχουν ομαδοποιηθεί σε 2 άλλες ομάδες
{((*D*1, *D*2), (*D*3, *D*4)), ((*D*5, *D*6), (*D*7, *D*8))}.

Το δεντροδιάγραμμα για αυτή τη δομή δίνεται στο σχήμα 9. Η αναπαράσταση αυτού σε αποθήκευση μέσω δομών δακτυλίου είναι τώρα απλή (βλέπε σχ.16).



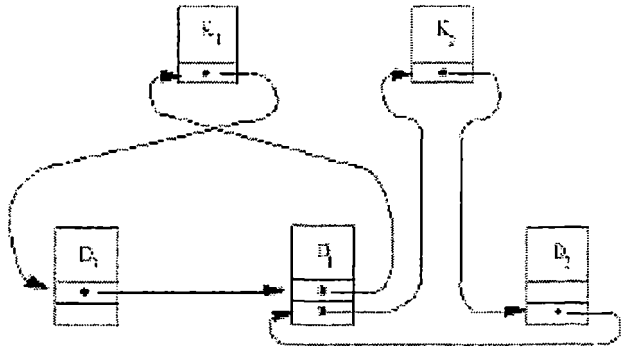
Σχήμα 16: Δενδροδιάγραμμα



Σχήμα 17: Η εφαρμογή ενός δενδροδιαγράμματος μέσω δομών δακτυλίων

Το Di δείχνει μια περιγραφή ενός εγγράφου.
Παρατηρήστε πώς οι δακτύλιοι σε χαμηλότερο επίπεδο εμπεριέχονται σε εκείνα με πιο υψηλό επίπεδο. Το πεδίο Ci περιέχει κανονικά κάποιες αναγνωριστικές πληροφορίες όσον αφορά το δακτύλιο που περιέχει. Παραδείγματος χάριν, το C1 με κάποιο τρόπο προσδιορίζει την κατηγορία εγγράφων {D1, D2} .
Εάν ομαδοποιούσαμε τα έγγραφα σύμφωνα με τις λέξεις κλειδιά που μοιράζονται, τότε για κάθε λέξη κλειδί θα είχαμε μια ομάδα εγγράφων, δηλαδή, αυτά που έχουν την λέξη κλειδί από κοινού. Το Ci θα ήταν έπειτα το πεδίο που περιέχει τη λέξη κλειδί που ενώνει εκείνη την συγκεκριμένη ομάδα. Οι δακτύλιοι θα επικαλύπτονταν (σχήμα 17), όπως σε αυτό το παράδειγμα:

D1 = {K1, K2}
D2 = {K2, K3}
D3 = {K1, K4}

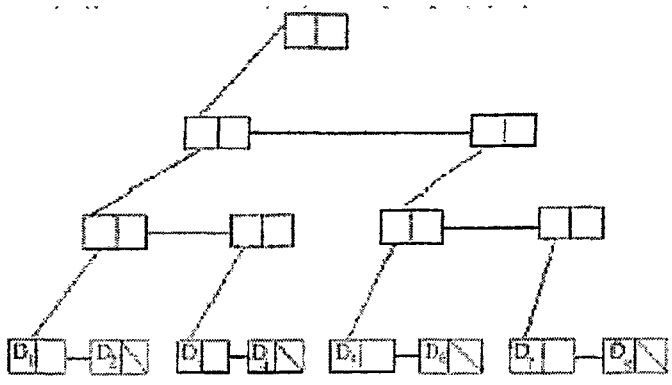


Σχήμα 18: Δύο επικαλυπτόμενοι δακτύλιοι

Η χρησιμότητα αυτού του είδους δομής θα γίνει προφανέστερη όταν συζητηθεί η αναζήτηση των ταξινομήσεων. Εάν κάθε δαχτυλίδι έχει συσχετισμένη μία εγγραφή η οποία περιέχει αναγνωριστικές πληροφορίες για τα μέλη του, τότε μια στρατηγική αναζήτησης που αναζητά μια δομή όπως αυτή, θα εξετάσει αρχικά το Ci (ή Ki στο δεύτερο παράδειγμα)για να καθορίσει εάν θα συνεχίσει ή εάν θα εγκαταλείψει την αναζήτηση.

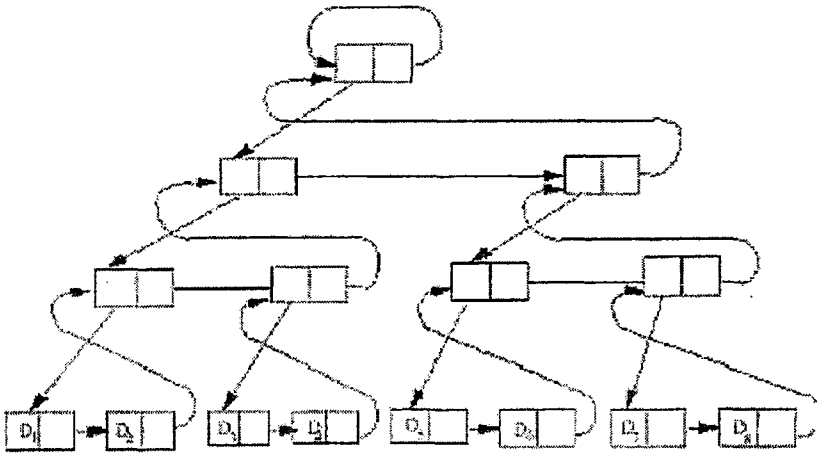
Νηματοειδής Λίστες (Threaded lists)

Μια απλή αντιπροσώπευση της ταξινόμησης
((D1, D2), (D3, D4)), ((D5, D6), (D7, D8)) δίνεται στο σχήμα 19.
Κάθε υπολίστα σε αυτήν την δομή έχει συσχετισμένη μια εγγραφή που περιέχει μόνο δύο δείκτες. (Μπορούμε να υποθέσουμε ότι το Di είναι στην πραγματικότητα ένας δείκτης του εγγράφου Di.). Η λειτουργία των δεικτών πρέπει να είναι ξεκάθαρη από το διάγραμμα. Το κύριο σημείο εντούτοις, είναι ότι το αρχείο που συνδέεται με έναν κατάλογο δεν περιέχει καθόλου αναγνωριστικές πληροφορίες.



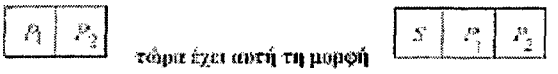
Σχήμα 19: Μια εφαρμογή δομής λίστας μιας ιεραρχικής ταξινόμησης

Μια τροποποίηση της εφαρμογής μιας δομής καταλόγων, όπως αυτή, που την κάνει να μοιάζει με ένα σύνολο δομών δακτυλίων, είναι να μετατραπεί ο δεξιός δείκτης του τελευταίου στοιχείου μιας υπολίστας σε επικεφαλίδα. Κάθε υπολίστα έχει γίνει αποτελεσματικά μια δομή δακτυλίου. Η δομή που δημιουργείται με τον τρόπο αυτό καλείται συνήθως νηματοειδής λίστα ((threaded list), δείτε το σχήμα 20). Η συγκεκριμένη αναπαράσταση είναι μια μικρή υπεραπλούστευση, δεδομένου ότι πρέπει να εντοπίσουμε ποια στοιχεία είναι στοιχεία δεδομένων (δίνουν πρόσβαση στα έγγραφα Di) και ποια στοιχεία είναι μόνο στοιχεία δεικτών. Το σημαντικότερο πλεονέκτημα που συνδέεται με μία περασμένη λίστα είναι ότι μπορεί να διατρεχθεί χωρίς την βοήθεια μιας στοίβας. Διαπερνώντας μια συμβατική δομή καταλόγων οι διευθύνσεις επιστροφής συσσωρεύονται, ενώ στην περασμένη λίστα έχουν ενσωματωθεί στη δομή δεδομένων.



Σχήμα 20: Μια εφαρμογή νηματοειδής λίστας μιας ιεραρχικής ταξινόμησης

Ένα μειονέκτημα που σχετίζεται με τη χρήση λιστών και δομών δαχτυλιδιού για την αντιπροσώπευση των ταξινομήσεων είναι ότι μπορούν μόνο να εισαχθούν στη "κορυφή" της λίστας. Ένας πρόσθετος δείκτης που επιτρέπει την είσοδο στη δομή σε κάθε ένα από τα στοιχεία δεδομένων αυξάνει την ταχύτητα ενημέρωσης αρκετά. Μια άλλη προσαρμογή της απλής αναπαράστασης καταλόγου έχει μελετηθεί εκτεταμένα από τους Stanfel54,55 και Patt56.Τα μεμονωμένα στοιχεία (ή κελιά) της δομής καταλόγου τροποποιούνται για να ενσωματώσουν ένα παραπάνω πεδίο, έτσι ώστε κάθε στοιχείο που έχει την μορφή



όπου τα Pi είναι δείκτες και το S είναι σύμβολο. Καμία άλλη ουσιαστική αλλαγή δεν έχει γίνει στην απλή αναπαράσταση. Αυτή η δομή έχει γίνει γνωστή ως Doubly Chained Tree.

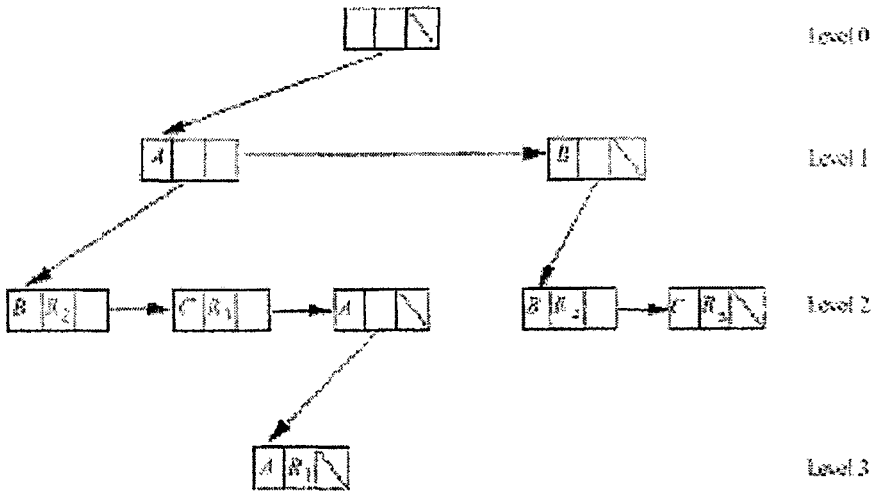
Οι ιδιότητές της έχουν ερευνηθεί κυρίως για την αποθήκευση κλειδιών μεταβλητού μήκους, όπου κάθε κλειδί κατασκευάζεται με επιλογή των συμβόλων από ένα πεπερασμένο (συνήθως μικρό) αλφάβητο. Παραδείγματος χάριν, ας υποθέσουμε ότι {A, B, C} είναι το σύνολο βασικών συμβόλων και έστω ότι R1, R2, R3, R4, R5 είναι πέντε αρχεία που θα αποθηκευτούν. Ας ορίσουμε τα κλειδιά, χρησιμοποιώντας 3 σύμβολα, στο αρχείο ως εξής:

AAA R1
AB R2
AC R3
BB R4
BC R5

Ένα παράδειγμα ενός διπλά ενοποιημένου δέντρου (doubly chained tree) που περιέχει τα κλειδιά και που δίνει την πρόσβαση στα αρχεία δίνεται στο σχήμα 21. Το κορυφαίο στοιχείο δεν περιέχει κανένα σύμβολο, λειτουργεί μόνο ως έναρξη της δομής. Λαμβάνοντας υπόψη ένα αυθαίρετο κλειδί η παρουσία ή η απουσία της ανιχνεύεται με το ταίριασμα του ενάντια στα κλειδιά στη δομή. Η ταυτοποίηση προχωρά επίπεδο επίπεδο, μόλις βρεθεί ένα σύμβολο ταυτοποίησης σε ένα επίπεδο, ο δείκτης P1 ακολουθείται στο σύνολο των εναλλακτικών συμβόλων του επόμενου επιπέδου. Η ταυτοποίηση θα ολοκληρωθεί είτε:

(1) όταν το κλειδί εξαντλείται, δηλαδή δεν υπάρχουν άλλα σύμβολα κλειδιά για να ταυτοποιηθούν ή

(2) όταν κανένα σύμβολο ταυτοποίησης δεν βρίσκεται στο τρέχον επίπεδο.



Σχήμα 21: Ένα παράδειγμα ενός διπλά ενοποιημένου δέντρου

Για την περίπτωση (1) έχουμε:

(α) το κλειδί υπάρχει εάν ο δείκτης P1, στο ίδιο κελί όπως στο προηγούμενο σύμβολο ταυτοποίησης, τώρα δείχνει σε ένα αρχείο
(β) ο P1 δείχνει σε ένα περαιτέρω σύμβολο, δηλαδή το κλειδί δεν ανταποκρίνεται στη λειτουργία του και επομένως δεν συμπεριλαμβάνεται μέσα στη δομή.

Για την περίπτωση (2), επίσης έχουμε ότι το κλειδί δεν είναι στη δομή, αλλά τώρα υπάρχει ένας κακός συνδυασμός.

Δέντρα (Trees)

Αν και οι επιστήμονες υπολογιστών έχουν υιοθετήσει τα δέντρα ως δομές αρχείων, οι ιδιότητές τους ερευνήθηκαν αρχικά από τους μαθηματικούς. Στην πραγματικότητα, ένα ουσιαστικό μέρος της θεωρίας των γραφικών παραστάσεων αφιερώνεται στη μελέτη των δέντρων. Τα καλύτερα βιβλία στις μαθηματικές πτυχές των δέντρων (και των γραφικών παραστάσεων) έχουν γραφτεί από τους Berge57, Harary58 και Ore59. Το βιβλίο του Harary περιέχει επίσης ένα χρήσιμο γλωσσάριο των εννοιών στη θεωρία γραφικών παραστάσεων. Επιπλέον οι Bertziss και Knuth60 συζητούν τα θέματα στη θεωρία γραφικών παραστάσεων με τις εφαρμογές στην επεξεργασία πληροφοριών. Υπάρχουν πολυάριθμοι ορισμοί των δέντρων. Έχουμε επιλέξει ιδιαίτερα έναν απλό από τον Berge. Εάν διαπραγματευτούμε μια γραφική παράσταση ως σύνολο κόμβων και σύνολο γραμμών έτσι ώστε κάθε γραμμή συνδέει ακριβώς δύο κόμβους, κατόπιν ένα δέντρο ορίζεται για να είναι μια πεπερασμένη συνδεδεμένη γραφική παράσταση

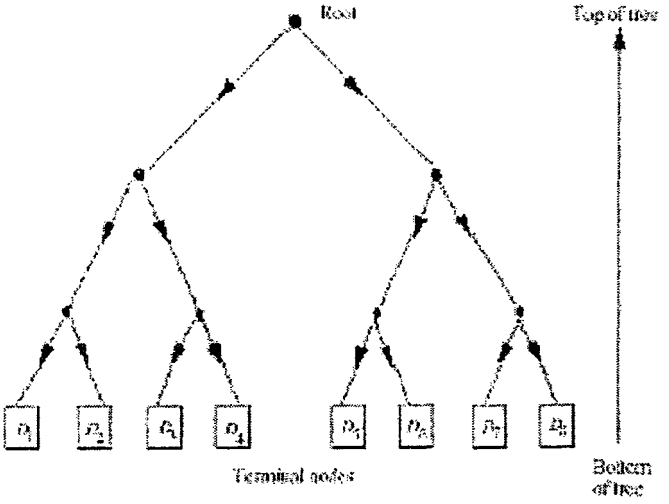
χωρίς τους κύκλους, και την κατοχή τουλάχιστον δύο κόμβων. Για να ορίσουμε έναν κύκλο ορίζουμε αρχικά μια αλυσίδα. Αντιπροσωπεύουμε τη γραμμή uk που ενώνει δύο

κόμβους x και y από uk = [x,y]. Μια αλυσίδα είναι μια ακολουθία γραμμών, στην οποία κάθε γραμμή uk έχει έναν κόμβο από κοινού με την προηγούμενη γραμμή uk-1, και άλλο vertex από κοινού με την πετυχαίνοντας γραμμή uk + 1. Ένα παράδειγμα μιας αλυσίδας είναι [a, x1], [x1, x2], [x2, x3], [x3, b]. Ένας κύκλος είναι μια πεπερασμένη αλυσίδα που αρχίζει σε έναν κόμβο και ολοκληρώνει στον ίδιο κόμβο (δηλ. στο παράδειγμα a = b).
Ο Berge δίνει το ακόλουθο θεώρημα στο οποίο παρουσιάζει πολλούς ισοδύναμους χαρακτηρισμούς των δέντρων.

Θεώρημα. Ας υποθέσουμε ότι H είναι μια γραφική παράσταση με τουλάχιστον n κόμβους, όπου n > 1 οποιαδήποτε από τις ακόλουθες ισοδύναμες ιδιότητες χαρακτηρίζουν ένα δέντρο.
(1) Το H συνδέεται και δεν περιέχει κανέναν κύκλο.
(2) Το H δεν περιέχει κανέναν κύκλο και έχει n – 1 γραμμές.
(3) Το H συνδέεται και έχει n -l γραμμές .
(4) Το H συνδέεται αλλά χάνει αυτήν την ιδιότητα εάν οποιαδήποτε γραμμή διαγραφεί.
(5) Κάθε ζευγάρι κόμβων συνδέεται με μια και μόνο μια αλυσίδα.

Αυτό που έχει παρατηρηθεί μέχρι τώρα, είναι ότι καμία αναφορά δεν έχει γίνει όσον αφορά την κατεύθυνση που σχετίζεται με μια γραμμή. Στις περισσότερες εφαρμογές στην πληροφορική (και το IR) ένας κόμβος επιλέγεται, και καλείται ρίζα του δέντρου. Για να φτάσουμε σε οποιοδήποτε άλλο κόμβο στο δέντρο, πρέπει να ξεκινήσουμε από τη ρίζα και να προχωρήσουμε κατά μήκος μιας αλυσίδας γραμμών έως ότου φτάσουμε στον κόμβο που αναζητούμε. Επομένως, μια κατεύθυνση συνδέεται με κάθε γραμμή. Στην πραγματικότητα, όταν πρέπει να αντιπροσωπεύσουμε ένα δέντρο μέσα σε έναν υπολογιστή μέσω μιας δομής λίστας, συχνά οι διευθύνσεις αποθηκεύονται με έναν τρόπο που επιτρέπει την κίνηση μόνο προς μια κατεύθυνση. Είναι βολικό να αντιμετωπίσουμε ένα δέντρο ως κατευθυνόμενη γραφική παράσταση με έναν συγκεκριμένο κόμβο ως ρίζα του δέντρου. Φυσικά, εάν έχουμε μια ρίζα τότε κάθε διαδρομή (κατευθυνόμενη αλυσίδα) που αρχίζει από τη ρίζα θα ολοκληρωθεί τελικά σε έναν συγκεκριμένο κόμβο από τον οποίο δεν θα προκύψει κανένας περαιτέρω κλάδος. Αυτοί οι κόμβοι ονομάζονται τελικοί κόμβοι του δέντρου.

Έχει γίνει κατανοητό ότι όταν μιλούσαμε για τις δομές δακτυλίων και τις νηματοειδείς λίστες σε μερικά από τα παραδείγματά μας ουσιαστικά δείχναμε πώς εφαρμόζεται μία δομή δέντρων. Το δενδροδιάγραμμα στο σχήμα 16 μπορεί εύκολα να αντιπροσωπευθεί ως δέντρο (σχήμα 22). Τα έγγραφα αποθηκεύονται στους τελικούς κόμβους και κάθε κόμβος αντιπροσωπεύει μια κατηγορία (ομάδα) εγγράφων. Η αναζήτηση ενός ιδιαίτερου συνόλου εγγράφων θα άρχιζε στη ρίζα και θα προχωρούσε κατά μήκος των βελών έως ότου βρεθεί η ζητούμενη κατηγορία.



Σχήμα 22: Ένα αντιπροσωπευτικό δέντρο ενός δενδροδιαγράμματος.

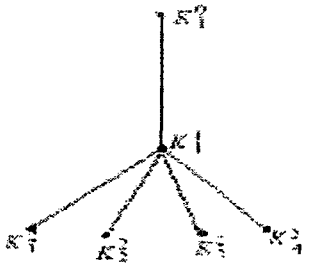
Ένα άλλο παράδειγμα μιας δομής δέντρων είναι ο κατάλογος που σχετίζεται με ένα ακολουθιακό αρχείο δεικτών. Περιγράφηκε ως ιεραρχία των δεικτών, αλλά θα μπορούσε εξίσου καλά να έχει περιγραφεί ως δομή δέντρου.
Η χρήση των δομών δέντρου στην πληροφορική χρονολογείται από τις αρχές της δεκαετίας του '50 όταν συνειδητοποιήθηκε ότι η αποκαλούμενη δυαδική αναζήτηση θα μπορούσε εύκολα να αντιπροσωπευθεί από ένα **δυαδικό δέντρο**. Ένα δυαδικό δέντρο είναι ένα δέντρο στο οποίο από κάθε κόμβο (εκτός από τους τελικούς κόμβους) ξεκινάνε δύο κλάδοι. Μια δυαδική αναζήτηση είναι μια αποδοτική μέθοδος για τον εντοπισμό της παρουσίας ή απουσίας μιας τιμής κλειδιού μεταξύ ενός συνόλου κλειδιών. Προϋποθέτει ότι τα κλειδιά έχουν ταξινομηθεί. Προχωρά με διαδοχικό χωρισμό του συνόλου, σε διαχωρισμό απορρίπτεται το μισό του συνόλου ως μη περιέχων το ζητούμενο κλειδί. Όταν το σύνολο περιέχει N ταξινομημένα κλειδιά ο χρόνος αναζήτησης είναι της τάξης log2N. Είναι προφανές πώς αυτή η διαδικασία μπορεί να αντιπροσωπευθεί απλά από ένα δυαδικό δέντρο.
Δυστυχώς, σε πολλές εφαρμογές θέλουμε να έχουμε τη δυνατότητα να εισάγουμε ένα κλειδί που λείπει. Εάν τα κλειδιά αποθηκεύονται διαδοχικά τότε ο χρόνος που χρειάζεται για τη λειτουργία εισαγωγής μπορεί να είναι της τάξης N. Εάν εντούτοις, τα κλειδιά αποθηκεύονται σε ένα δυαδικό δέντρο, αυτός ο μεγάλος χρόνος εισαγωγής μπορεί να ξεπεραστεί, τόσο ο χρόνος αναζήτησης όσο και ο χρόνος εισαγωγής θα είναι της τάξης log2N. Τα κλειδιά αποθηκεύονται στους κόμβους, σε κάθε κόμβο ο αριστερός κλάδος θα οδηγεί σε "μικρότερα" κλειδιά, ο δεξιός κλάδος θα οδηγεί στα "μεγαλύτερα" κλειδιά. Μια αναζήτηση που ολοκληρώνεται σε έναν τελικό κόμβο θα δείχνει ότι το κλειδί δεν υπάρχει και θα πρέπει να εισαχθεί.

5. Στρατηγικές Αναζήτησης

Στην περίπτωση της ανάκτησης πληροφορίας, η πληροφορία είναι το υποσύνολο των εγγράφων τα οποία κρίνονται ότι είναι σχετικά με το ερώτημα. Το είδος της αναζήτησης που μας ενδιαφέρει εδώ, δεν είναι το συνηθισμένο είδος όπου το αποτέλεσμα της αναζήτησης είναι ευδιάκριτο, δηλαδή είτε εάν το στοιχείο υπάρχει, είτε όχι εάν το στοιχείο είναι ανύπαρκτο. Όλες οι στρατηγικές αναζήτησης βασίζονται στη σύγκριση μεταξύ του ερωτήματος και των αποθηκευμένων εγγράφων. Μερικές φορές αυτή η σύγκριση επιτυγχάνεται μόνο έμμεσα όταν το ερώτημα συγκρίνεται με clusters (ή ακριβέστερα με τα profiles που αντιπροσωπεύουν τα clusters).
Οι διαχωρισμοί που γίνονται μεταξύ των διαφορετικών ειδών των στρατηγικών αναζήτησης μπορούν μερικές φορές να γίνουν κατανοητοί κοιτάζοντας την γλώσσα του ερωτήματος, που είναι η γλώσσα στην οποία χρειάζεται να εκφράζεται η πληροφορία. **Η φύση του ερωτήματος συχνά υπαγορεύει τη φύση της στρατηγικής αναζήτησης.** Για παράδειγμα, μία γλώσσα ερωτήματος η οποία επιτρέπει να εκφραστούν δηλώσεις ερωτήματος με όρους λογικών συνδυασμών από λέξεις κλειδιά κανονικά υπαγορεύει μία λογική αναζήτηση. Αυτή είναι μία αναζήτηση η οποία επιτυγχάνει τα αποτελέσματά της με λογικές (παρά με αριθμητικές) συγκρίσεις του ερωτήματος με τα έγγραφα.

Λογική Αναζήτηση (Boolean search)

Μία στρατηγική λογική αναζήτησης (boolean search) ανακτά εκείνα τα έγγραφα τα οποία το ερώτημα επιστρέφει αληθή τιμή. Αυτή η διατύπωση έχει νόημα μόνο εάν τα ερωτήματα εκφράζονται από την άποψη των όρων δεικτών (ή των λέξεων κλειδιών) και συνδυάζονται με τους συνήθης λογικούς τελεστές AND, OR, και NOT. Για παράδειγμα, εάν έχουμε το ερώτημα Q = (K1 AND K2) OR (K3 AND (NOT K4)) τότε η λογική αναζήτηση θα ανακτήσει όλα τα έγγραφα που δεικτοδοτούνται από την K1 και την K2, όπως επίσης και όλα τα έγγραφα που δεικτοδοτούνται από την K3 αλλά όχι από την K4.
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.



Σχήμα 23: Μια ομάδα από ιεραρχικά συσχετισμένες λέξεις-κλειδιά.

Ένας προφανής τρόπος για να εφαρμόσουμε την λογική αναζήτηση είναι μέσω του αντιστρεφόμενου αρχείου. Αποθηκεύουμε μία λίστα για την κάθε λέξη κλειδί μέσα στο λεξιλόγιο και σε κάθε λίστα τοποθετούμε τις διευθύνσεις (ή αριθμούς) των εγγράφων που περιέχουν τη συγκεκριμένη λέξη. Τώρα, για να ικανοποιήσουμε το ερώτημα εκτελούμε το σύνολο των πράξεων, που αντιστοιχούν στους λογικούς συνδέσμους, πάνω στη λίστα Ki. Για παράδειγμα, εάν

K_1 -list : D_1, D_2, D_3, D_4
K_2 -list : D_1, D_2
K_3 -list : D_1, D_2, D_3
K_4 -list : D_1

και $Q = (K_1 \text{ AND } K_2) \text{ OR } (K_3 \text{ AND } (\text{NOT } K_4))$

τότε για να ικανοποιηθεί το μέρος του ερωτήματος (K1 AND K2) τέμνουμε τις λίστες K1 και K2, για να ικανοποιηθεί το μέρος του ερωτήματος (K3 AND (NOT K4)) αφαιρούμε τη λίστα K4 από τη λίστα K3. Το OR ικανοποιείται παίρνοντας την ένωση των δύο συνόλων που αποκτήθηκαν από τα παραπάνω μέρη του ερωτήματος. Το αποτέλεσμα είναι το σύνολο των εγγράφων {D1, D2, D3} τα οποία ικανοποιούν το ερώτημα και κάθε έγγραφο ένα από αυτά τα έγγραφα είναι «true» για το ερώτημα.

Μία μικρή τροποποίηση της λογικής αναζήτησης είναι μία η οποία επιτρέπει μόνο το λογικό AND αλλά λαμβάνει υπόψη τον πραγματικό αριθμό των όρων που το ερώτημα έχει από κοινού με ένα έγγραφο. Αυτός ο αριθμός έχει γίνει γνωστός ως επίπεδο συντονισμού (co-ordination level). Αυτή η στρατηγική αναζήτησης συχνά **καλείται απλό ταίριασμα (simple matching)**. Επειδή σε οποιοδήποτε επίπεδο μπορούμε να έχουμε περισσότερα από ένα έγγραφα, τα έγγραφα θεωρούνται μερικώς ταξινομημένα από τα επίπεδα συντονισμού. Για το ίδιο παράδειγμα όπως πριν με το ερώτημα Q = K1 AND K2 AND K3 αποκτούμε την ακόλουθη ταξινόμηση:

επίπεδο συντονισμού

3 D1, D2
2 D3
1 D4

Στην πραγματικότητα το απλό ταίριασμα μπορεί να αντιμετωπισθεί ως Χρησιμοποίηση μιας πρωτόγονης συνάρτησης ταιριάσματος. Για κάθε έγγραφο D υπολογίζουμε το $|D \cap Q|$, που είναι το μέγεθος επικάλυψης μετξύ των D και Q, όπου το κάθε ένα αναπαρίσταται ως ένα σύνολο από λέξεις κλειδιά. Αυτό είναι ο συντελεστής απλού ταιριάσματος.

Συναρτήσεις ταυτοποίησης (matching functions)

Πολλές από τις πιο περίπλοκες στρατηγικές αναζήτησης εφαρμόζονται με τη βοήθεια μιας συνάρτησης ταυτοποίησης. Αυτή είναι μια συνάρτηση παρόμοια με ένα μέτρο σχέσης, αλλά διαφέρει στο ότι μια συνάρτηση ταυτοποίησης μετρά τη σχέση μεταξύ ενός ερωτήματος και ενός σχεδιαγράμματος εγγράφου ή ομάδας (cluster), ενώ ένα μέτρο σχέσης εφαρμόζεται σε αντικείμενα του ίδιου είδους. Από μαθηματικής άποψης οι δύο συναρτήσεις έχουν τις ίδιες ιδιότητες και η μόνη διαφορά τους είναι στην ερμηνεία τους.
Ίσως η απλούστερη είναι αυτή που συνδέεται με την απλή ταυτοποίηση με στρατηγική αναζήτησης.
Εάν M είναι η συνάρτηση ταυτοποίησης, D το σύνολο λέξεων κλειδιών που αντιπροσωπεύουν το έγγραφο και Q το σύνολο που αντιπροσωπεύει την ερώτηση, τότε:

$$M = \frac{2|D \cap Q}{|D| + |Q|}$$

είναι ένα άλλο παράδειγμα μιας συνάρτησης ταυτοποίησης. Ο δημοφιλής τρόπος που χρησιμοποιείται από το πρόγραμμα του SMART, καλεί το συσχετισμό συνημίτονου, υποθέτει ότι το έγγραφο και η ερώτηση αντιπροσωπεύονται ως αριθμητικά διανύσματα. Το οποίο είναι Q = (q1, q2, ...,qt) και D = (d1, d2, ..., dt) όπου qi και τα di είναι αριθμητικά βάρη που συνδέονται με τη λέξη κλειδί i. Ο συσχετισμός συνημίτονου είναι τώρα απλά

$$r = \frac{\sum_{i=1}^{t} q_i d_i}{\left(\sum_{i=1}^{t} (q_i)^2 \sum_{i=1}^{t} (d_i)^2\right)^{1/2}}$$

ή εναλλακτικά

$$r = \frac{(Q, D)}{\|Q\| \|D\|} = \cos\theta$$

όπου θ είναι η γωνία μεταξύ των διανυσμάτων Q και D.

Σειριακή Αναζήτηση

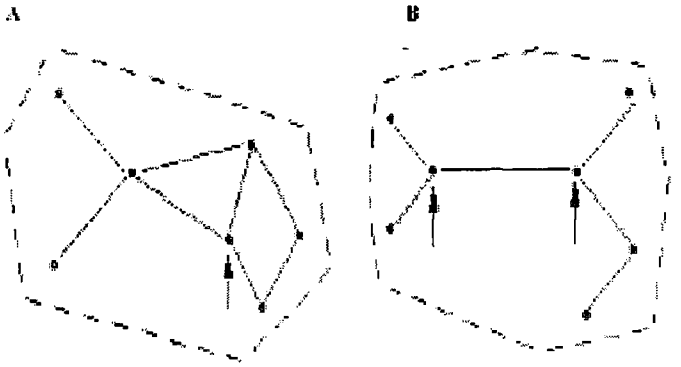
Αν και οι σειριακές αναζητήσεις είναι γενικά αργές, ωστόσο χρησιμοποιούνται συχνά ως μέρος μεγαλύτερων συστημάτων. Επίσης παρέχουν μία βολική επίδειξη της χρήσης των συναρτήσεων ταυτοποίησης.
Υποθέτουμε ότι υπάρχουν N έγγραφα Di στο σύστημα, κατόπιν οι τμηματικές εισπράξεις αναζήτησης με τον υπολογισμό των N τιμών M(Q, Di) το σύνολο εγγράφων που ανακτώνται καθορίζονται. Υπάρχουν δύο τρόποι:

(1) στην συνάρτηση ταυτοποίησης δίνεται ένα κατάλληλο κατώτατο όριο, που ανακτά τα έγγραφα επάνω από το κατώτατο όριο και που απορρίπτει τα παρακάτω. Εάν το T είναι το κατώτατο όριο, κατόπιν το ανακτημένο καθορισμένο B είναι το σύνολο { Di | M (Q, Di) > T } .
(2) τα έγγραφα ταξινομούνται κατά αυξανόμενη σειρά ως προς την τιμή της συνάρτησης ταυτοποίησης. Μια κατάταξη θέσης (rank position) R επιλέγεται ως διακοπή και όλα τα έγγραφα κάτω από την τάξη ανακτώνται έτσι ώστε B = { Di | r(i) < R } όπου το r(i) είναι η κατάταξη θέσης που ορίζεται στο Di. Η ελπίδα σε κάθε περίπτωση είναι ότι τα σχετικά έγγραφα περιλαμβάνονται στο ανακτημένο σύνολο.
Η κύρια δυσκολία με αυτό το είδος της στρατηγικής αναζήτησης είναι ο προσδιορισμός του κατώτατου ορίου ή της διακοπής. Θα είναι πάντα αυθαίρετο δεδομένου ότι δεν υπάρχει κανένας τρόπος που να καταδεικνύει εκ των προτέρων ποια τιμή για κάθε ερώτημα θα προκαλέσει την καλύτερη ανάκτηση.

Αναπαραστάσεις ομάδων (Cluster representatives)

Ενώ στη σειριακή αναζήτηση πρέπει να είμαστε σε θέση να ταιριάξουμε τα ερωτήματα με κάθε έγγραφο στο αρχείο, σε μια αναζήτηση ενός ομαδοποιημένου αρχείου πρέπει να είμαστε σε θέση να ταιριάξουμε τα ερωτήματα με τις ομάδες. Για το σκοπό αυτό οι ομάδες αναπαρίστανται με κάποιο είδος profile, το οποίο αποκαλείται αναπαράσταση. Μία αναπαράσταση ομάδας πρέπει να είναι τέτοια ώστε σε εισερχόμενο ερώτημα να εντοπιστεί στις ομάδες που περιέχει τα έγγραφα τα σχετικά με το ερώτημα. **Με άλλα λόγια αναμένουμε την αναπαράσταση της ομάδας για να διακρίνουμε τα σχετικά από τα μη σχετικά έγγραφα, όταν συνδυάζονται έναντι οποιουδήποτε ερωτήματος.** Δυστυχώς δεν υπάρχει καμία θεωρία που να επιτρέπει σε κάποιον να επιλέξει το σωστό είδος αναπαράστασης ομάδας. Υπάρχουν διάφοροι «λογικοί» τρόποι χαρακτηρισμού των ομάδων αλλά παραμένει έπειτα το θέμα της πειραματικής δοκιμής για να αποφασίσει ποιος από αυτούς είναι ο αποτελεσματικότερος.

Εάν υποθέτουμε ότι οι ομάδες προέρχονται από μια μέθοδο ομάδων βασισμένη σε ένα μέτρο ανομοιότητας, κατόπιν μπορούμε να αντιπροσωπεύουμε κάθε ομάδα σε κάποιο επίπεδο ανομοιότητας από μια γραφική παράσταση (δείτε το σχήμα 24). Εδώ το **A** και το **B** είναι δύο ομάδες. Οι κόμβοι αντιπροσωπεύουν τα έγγραφα και η γραμμή μεταξύ οποιωνδήποτε δύο κόμβων δείχνουν ότι τα αντίστοιχα έγγραφά τους είναι λιγότερο ανόμοια από κάποιο διευκρινισμένο επίπεδο ανομοιότητας.



Σχήμα 24: Παραδείγματα από μέγιστες συνδέσεις εγγράφων ως αντιπρόσωπους ομάδας

Τώρα, ένας τρόπος για να επιλεχτεί μια ομάδα είναι ένα χαρακτηριστικό μέλος από την ομάδα. Ένας απλός τρόπος είναι να βρεθεί εκείνο το έγγραφο που συνδέεται με το μέγιστο αριθμό άλλων εγγράφων στην ομάδα. Ένα κατάλληλο όνομα για αυτό το είδος αντιπροσώπου ομάδων είναι το ανώτατα συνδεμένο έγγραφο. Στις ομάδες **A** και **B** που διευκρινίζονται, υπάρχουν δείκτες στους υποψηφίους. Όπως κάποιος θα ανέμενε σε μερικές περιπτώσεις ο αντιπρόσωπος δεν είναι μοναδικός. Παραδείγματος χάριν, στη συστάδα **B** έχουμε δύο υποψηφίους.
Εξετάζεται τώρα ένας άλλος τρόπος για τις ομάδες. Επιδιώκεται μια μέθοδο αντιπροσώπευσης που με κάποιο τρόπο "υπολογίζει κατά μέσο όρο" τις περιγραφές των μελών των συστάδων. Η μέθοδος που αναπηδά αμέσως για να απασχολήσει είναι μια στην οποία το ένα υπολογίζει κεντροειδές (ή το κέντρο βάρους) της ομάδας. Εάν D1, D2,...,Dn είναι τα έγγραφα στην ομάδα και κάθε Di αντιπροσωπεύεται από ένα αριθμητικό διάνυσμα (d1, d2, ..., dt) έπειτα το κεντροειδές C της ομάδας δίνεται από τη σχέση.

$$C = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\|D_i\|}$$

όπου || *το Di* || είναι συνήθως η Ευκλείδεια απόσταση, δηλ.

$$\|D_i\| = \sqrt{d_1^2 + d_2^2 + \ldots + d_t^2}$$

Τις περισσότερες φορές τα έγγραφα δεν αντιπροσωπεύονται από τα αριθμητικά διανύσματα αλλά από τα δυαδικά διανύσματα (ή ισοδύναμα, σύνολα λέξεων κλειδιών). Σε εκείνη την περίπτωση μπορούμε ακόμα να χρησιμοποιήσουμε έναν κεντροειδή τύπο αντιπροσώπου ομάδων αλλά η κανονικοποίηση αντικαθίσταται με μια διαδικασία που κατώτατα όρια τα συστατικά του ποσού [[Σ]] Di. Για να είναι ακριβέστερος, αφήστε το Di ένα τώρα να είναι δυαδικό διάνυσμα, έτσι ώστε ένα 1 στη θέση jth (νιοστή) δείχνει την παρουσία της λέξης κλειδιού jth (j-οστό) στο έγγραφο και

0 δείχνουν το αντίθετο. Ο αντιπρόσωπος ομάδων προέρχεται τώρα από το διάνυσμα ποσού

$$S = \sum_{i=1}^{n} D_i$$

Αφήνουμε C = (c1, c2, ..., ct) να είναι ο αντιπρόσωπος ομάδων και [Di]j το jth (j-οστό) τμήμα του δυαδικού διανυσματικού Di, κατόπιν δύο μέθοδοι είναι:

(1) $$c_j = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} [D_i]_j > 1 \\ 0 \text{ otherwise} \end{cases}$$

(2) $$c_j = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} [D_i]_j > \log_2 n \\ 0 \text{ otherwise} \end{cases}$$

Έτσι τελικά λαμβάνουμε ως αντιπρόσωπο ομάδων ένα δυαδικό διάνυσμα C. Και στις δύο περιπτώσεις η άποψη είναι ότι οι λέξεις κλειδιά που εμφανίζονται μόνο μια φορά στην ομάδα πρέπει να αγνοηθούν. Στη δεύτερη περίπτωση ομαλοποιούμε το μέγεθος n της ομάδας.
Υπάρχουν κάποια στοιχεία για να δειχτεί ότι και οι δύο αυτές μέθοδοι αντιπροσώπευσης είναι αποτελεσματικές όταν χρησιμοποιούνται από κοινού με τις κατάλληλες στρατηγικές αναζήτησης (βλέπε για παράδειγμα van Rijsbergen61 και Murray62. Προφανώς υπάρχουν περαιτέρω παραλλαγές στη λήψη των αντιπροσώπων ομάδων. Είναι πιθανότερο ότι ο τρόπος που το στοιχείο στον αντιπρόσωπο ομάδων χρησιμοποιείται από τη στρατηγική αναζήτησης θα έχει μια μεγαλύτερη επίδραση στο μέλλον.
Υπάρχει ένας άλλος θεωρητικός τρόπος την κατασκευή των αντιπροσώπων ομάδων και αυτός είναι μέσω της έννοιας ενός μέγιστου συντελεστή πρόβλεψης για μια ομάδα 63. Λαμβάνοντας υπόψη ότι, όπως πριν, τα έγγραφα Di σε μια ομάδα είναι δυαδικά διανύσματα, μια δυαδική ομάδα αντιπροσωπευτική για αυτήν την ομάδα είναι μέτρο πρόβλεψης υπό την έννοια ότι κάθε συστατικό (ci) προβλέπει ότι είναι η πλέον πιθανή αξία εκείνης της οντότητας στα έγγραφα μελών. Είναι μέγιστο εάν οι σωστές προβλέψεις της είναι όσο το δυνατόν πιο πολυάριθμες. Εάν το ένα υποθέτει ότι κάθε μέλος μιας ομάδας των εγγράφων D1,..., Dn είναι εξίσου πιθανό έπειτα ο αναμενόμενος συνολικός αριθμός, οι ανακριβείς προβλεφθείσες οντότητες (ή απλά ο αναμενόμενος συνολικός αριθμός κακών συνδυασμών μεταξύ των εγγράφων αντιπροσώπων και μελών ομάδων από όλα στο δυαδικό) είναι,

$$S = \sum_{i=1}^{n} \sum_{j=1}^{t} ([D_i]_j - c_j)^2$$

Αυτό μπορεί να ξαναγραφεί ως

$$S = \sum_{i=1}^{n} \sum_{j=1}^{t} (|D_i|_j - D_j)^2 = n \sum_{j=1}^{t} (|D_i|_j - c_j)$$

όπου

$$D_j = \frac{1}{n} \sum_{i=1}^{n} |D_i|_j$$

Η έκφραση θα ελαχιστοποιηθεί, μεγιστοποιώντας κατά συνέπεια τον αριθμό σωστών προβλέψεων, όταν c= (c1,....., ct) επιλέγεται κατά τέτοιο τρόπο ώστε το

$$\sum_{j=1}^{t} (|D_i|_j - c_j)^2$$

να είναι ένα ελάχιστο. Αυτό επιτυγχάνεται με τη σχέση (1)

$$c_j = \begin{cases} 1 & \text{if } D_j > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Έτσι με άλλα λόγια μια λέξη κλειδί θα οριστεί σε έναν αντιπρόσωπο συστάδων εάν εμφανιστεί περισσότερο στα μισά από τα έγγραφα μελών. Αυτό μεταχειρίζεται τα λάθη της πρόβλεψης που προκαλούνται από την απουσία ή την παρουσία λέξεων κλειδιών σε ίση βάση. Ο Croft63 έχει δείξει ότι είναι λογικότερο να διαφοροποιηθούν οι δύο τύποι λαθών στις εφαρμογές ανάκτησης πληροφοριών. Έδειξε ότι για να προβλέψει ψευδώς 0 (cj = 0) είναι δαπανηρότερο από να προβλέψει ψευδώς ένα (cj = 1). Σε αυτές τις περιπτώσεις η 1/2 αξία που εμφανίζεται στην (1) είναι μια σταθερά από λιγότερο από 1/2. Η ακριβής αξία της που αφορά την ανάλογη σημασία αποδίδεται και στους δύο τύπους λαθών πρόβλεψης.
Αν και ο κύριος λόγος για αυτούς τους αντιπροσώπους συστάδων είναι να οδηγηθεί μια στρατηγική αναζήτησης στα σχετικά έγγραφα, πρέπει να είναι σαφές ότι μπορούν επίσης να χρησιμοποιηθούν για να καθοδηγήσουν μια αναζήτηση στα έγγραφα που ικανοποιούν κάποιο όρο στην ταιριάζοντας με λειτουργία. Παραδείγματος χάριν, μπορούμε να θελήσουμε να ανακτήσουμε όλα τα έγγραφα Di που ταιριάζουν με το Q καλύτερα από το T, δηλ.

$$\{ D_i \mid M(Q, D_i) > T \}$$

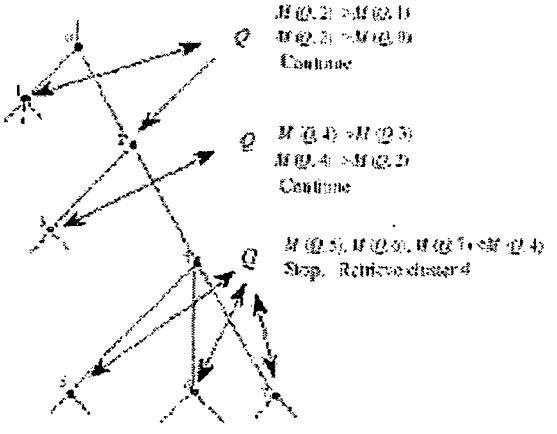
Ένα σημαντικό ελάττωμα στην περισσότερη εργασία για τους αντιπροσώπους συστάδων είναι ότι μεταχειρίζεται τη διανομή των λέξεων κλειδιών στις συστάδες σαν ανεξάρτητες. Αυτό δεν είναι πολύ ρεαλιστικό.
Τέλος, πρέπει να σημειωθεί ότι οι μέθοδοι συστάδων που προχωρούν άμεσα από τις περιγραφές εγγράφων στην ταξινόμηση χωρίς πρώτα να υπολογίσουν τον ενδιάμεσο συντελεστή ανομοιότητας, θα πρέπει να κάνουν μια επιλογή του αντιπροσώπου συστάδων από την αρχή. Αυτοί οι αντιπρόσωποι συστάδων βελτιώνονται έπειτα ως

αλγόριθμος, ρυθμίζοντας την ταξινόμηση σύμφωνα με κάποια αντικειμενική λειτουργία, πιθανώς με βήματα επανάληψης.

Ανάκτηση Βασισμένη σε ομαδοποίηση

Η ανάκτηση βασισμένη σε ομαδοποίηση έχει ως βάση της την υπόθεση ομάδας, η οποία δηλώνει ότι τα πολύ σχετικά έγγραφα τείνουν να είναι σχετικά με τα ίδια αιτήματα. Η συγκέντρωση διαλέγει τα πολύ σχετικά έγγραφα και τα συγκεντρώνει σε μια ομάδα
Υποθέτουμε ότι έχουμε μια ιεραρχική ταξινόμηση των εγγράφων έπειτα μια απλή στρατηγική αναζήτησης πηγαίνει ως εξής (σχήμα 25). Η αναζήτηση αρχίζει στη ρίζα του δέντρου, κόμβος 0 στο παράδειγμα. Προχωράει να αξιολογήσει μια συνάρτηση ταυτοποίησης στον απόγονο κόμβων αμέσως από τον κόμβο 0, στο παράδειγμα οι κόμβοι 1 και 2. Αυτό το σχέδιο επαναλαμβάνεται κάτω από το δέντρο. Η αναζήτηση κατευθύνεται από έναν κανόνα απόφασης, ο οποίος βάσει της σύγκρισης των τιμών της συνάρτησης ταυτοποίησης σε κάθε στάδιο αποφασίζει ποιο κόμβο να επεκταθεί περαιτέρω. Επίσης, είναι απαραίτητο να υπάρξει ένας οριακός κανόνας που ολοκληρώνει την αναζήτηση και αναγκάζει να γίνει μια ανάκτηση. Στο σχήμα 25 ο κανόνας απόφασης είναι: επεκτείνετε τον κόμβο που αντιστοιχεί στη μέγιστη τιμή της συνάρτησης ταυτοποίησης που επιτυγχάνεται μέσα σε ένα σύνολο. Ο οριακός κανόνας είναι: σταματήστε εάν το τρέχον μέγιστο είναι λιγότερο από το προηγούμενο μέγιστο. Μερικές παρατηρήσεις για αυτήν την στρατηγική είναι :

(1) υποθέτουμε ότι η αποτελεσματική ανάκτηση μπορεί να επιτευχθεί με την εύρεση μόνο μιας ομάδας
(2) υποθέτουμε ότι κάθε ομάδα μπορεί να αντιπροσωπευθεί επαρκώς με σκοπό την εντόπιση της ομάδας που περιέχει τα σχετικά έγγραφα
(3) εάν το μέγιστο της συνάρτησης ταυτοποίησης δεν είναι μοναδικό κάποια συγκεκριμένη δράση, θα πρέπει να ληφθεί
(4) η αναζήτηση ολοκληρώνει πάντα και θα ανακτήσει τουλάχιστον ένα έγγραφο.



Σχήμα 25: Ένα δέντρο αναζήτησης και οι κατάλληλες τιμές της συνάρτησης ταυτοποίησης, δείχνουν την δράση ενός κανόνα απόφασης.

Εάν εγκαταλείψουμε τώρα την ιδέα της κατοχής μιας πολλαπλής στάθμης συγκέντρωσης και δεχτούμε μια μονοεπίπεδη συγκέντρωση, καταλήγουμε με την προσέγγιση στη συγκέντρωση εγγράφων. Η κατάλληλη μέθοδος ομάδων απεικονίζεται από τον αλγόριθμο Rocchio που περιγράφεται στο κεφάλαιο 3. Η στρατηγική αναζήτησης είναι εν μέρει μια τμηματική αναζήτηση. Προχωράει πρώτα να βρει την καλύτερη (ή η πλησιέστερη) ομάδα (-ες) και έπειτα να κοιτάξει μέσα σε αυτούς. Το δεύτερο στάδιο επιτυγχάνεται με το να γίνει μια τμηματική αναζήτηση των εγγράφων στην επιλεγμένη ομάδα (-ες). Το αρχείο εξόδου είναι συχνά μια ταξινόμηση των εγγράφων που ανακτώνται έτσι.

Διατύπωση Αμφίδρομης Αναζήτησης

Ένας χρήστης που έρχεται αντιμέτωπος με ένα αυτόματο σύστημα ανάκτησης είναι δύσκολο να είναι σε θέση να εκφράσει την ανάγκη του για πληροφορία με σωστό τρόπο. Είναι πιθανότερο να θελήσει να επιτρέψει μια διαδικασία δοκιμής — και –λάθους (trial-and-error process) στην οποία διατυπώνει το ερώτημά του λαμβάνοντας υπόψη όλα όσα μπορεί να του πει το σύστημα. Το είδος της πληροφορίας που πιθανόν να θελήσει να χρησιμοποιήσει για την επαναδιατύπωση του ερωτήματός του είναι:

(1) η συχνότητα ύπαρξης στη βάση δεδομένων των όρων αναζήτησής του
(2) ο αριθμός εγγράφων που πιθανών να ανακτηθούν από το ερώτημά του
(3) οι εναλλακτικοί και σχετικοί όροι να είναι αυτοί που χρησιμοποιούνται στην ένα μικρό δείγμα των παραπομπών που πιθανών να ανακτηθούν και
(4) οι όροι που χρησιμοποιούνται για να ευρετηριοποιήσουν τις παραπομπές στο (4).
Όλο αυτό μπορεί να παρασχεθεί ως εξυπηρέτηση σε έναν χρήστη κατά τη διάρκεια της περιόδου αναζήτησής του από ένα σύστημα αμφίδρομης ανάκτησης. Εάν ανακαλύψει ότι ένας από τους όρους αναζήτησής του εμφανίζεται πολύ συχνά μπορεί να επιθυμήσει να τον καταστήσει πιο συγκεκριμένο συμβουλευόμενος ένα ιεραρχικό λεξικό που θα τον ενημερώσει για το ποιες είναι οι επιλογές του. Παρόμοια, εάν το ερώτημά του είναι πιθανό να ανακτήσει πάρα πολλά έγγραφα μπορεί να το κάνει πιο συγκεκριμένο.
Επίσης, το δείγμα των παραπομπών και η δεικτοδότησης τους θα του δώσουν κάποια ιδέα για το τι είδους έγγραφα είναι πιθανό να ανακτηθούν και συνεπώς κάποια ιδέα για το πόσο αποτελεσματικοί είναι οι όροι αναζήτησής του στην έκφραση της ανάγκης του για πληροφορία. Έτσι, ο χρήστης, μπορεί να τροποποιήσει το ερώτημά του λαμβάνοντας υπόψη το δείγμα ανάκτησης. Αυτή η διαδικασία στην οποία ο χρήστης τροποποιεί το ερώτημά του βασισμένος στα πραγματικά αποτελέσματα της αναζήτησης θα μπορούσε να περιγραφεί ως μία μορφή ανατροφοδότησης.

Ανατροφοδότηση (Feedback)

Η λέξη feedback (ανατροφοδότηση) συνήθως χρησιμοποιείται για να περιγράψει το μηχανισμό με τον οποίο ένα σύστημα μπορεί να βελτιώσει την απόδοση πάνω σε μία εργασία λαμβάνοντας υπόψη την προηγούμενη απόδοση. Με άλλα λόγια ένα απλό σύστημα εισόδου-εξόδου ανατροφοδοτεί τις πληροφορίες από την έξοδο έτσι ώστε

αυτές να μπορούν να χρησιμοποιηθούν για να βελτιώσουν την απόδοση στην επόμενη εισαγωγή.

Εξετάζουμε τώρα μια στρατηγική ανάκτησης που έχει εφαρμοστεί με τη βοήθεια μιας συνάρτησης ταυτοποίησης M. Επιπλέον, υποθέτουμε ότι και οι αντιπρόσωποι d ερώτησης q και εγγράφων είναι t - διαστατικά διανύσματα με τα πραγματικά συστατικά όπου το t είναι ο αριθμός όρων δεικτών.
Ο στόχος κάθε στρατηγικής ανάκτησης είναι να ανακτηθούν τα σχετικά έγγραφα A και να παρακρατηθούν τα μη-σχετικά έγγραφα A. Δυστυχώς η σχετικότητα καθορίζεται όσον αφορά τη σημασιολογική ερμηνεία του χρήστη στην ερώτησή του. Από την άποψη του συστήματος ανάκτησης η διατύπωση του μπορεί να μην είναι ιδανική. Μια ιδανική διατύπωση θα ήταν μία που ανάκτησε μόνο τα σχετικά έγγραφα. Στην περίπτωση μιας τμηματικής αναζήτησης το σύστημα θα ανακτήσει όλο το d για το οποίο M (Q,D) > T και να μην ανακτήσει οποιοδήποτε d για το οποίο M (Q,D) ≤ T, όπου το T είναι ένα διευκρινισμένο κατώτατο όριο. Συμβαίνει αυτό στην περίπτωση όπου το M είναι η λειτουργία συσχετισμού συνημίτονου, δηλ.

$$M(Q,D) = \frac{(Q,D)}{|Q|\,|D|} \qquad \frac{t}{|Q|\,|D|} \times (w_1 q_1 \cdot q_2 d_2 \cdots q_t d_t)$$

η διαδικασία απόφασης

$$M(Q,D) - T > 0$$

αντιστοιχεί σε μια γραμμική διακρίνουσα λειτουργία που χρησιμοποιείται για να χωρίσει γραμμικά δύο σύνολα A και A σε R[t]. Υποθέτουμε προς το παρόν ότι το A(σύνολο σχετικών εγγραφών) και A (σύνολο μη-σχετικών εγγραφών) είναι γνωστά εκ των προτέρων, κατόπιν η σωστή διατύπωση ερώτησης Q0 θα ήταν μια για την οποία

$M(Q_0,D) > T$ όταν $D \in A$
και
$M(Q_0,D) \leq T$ όταν $D \in \bar{A}$

Το ενδιαφέρον είναι ότι αρχίζοντας από οποιοδήποτε Q μπορούμε να το ρυθμίσουμε χρησιμοποιώντας επαναληπτικά τις πληροφορίες ανατροφοδότησης έτσι ώστε να συγκλίνει Q0. Υπάρχει ένα θεώρημα (Nilsson64 σελ. 81) που δηλώνει ότι η παροχή Q0 υπάρχει όταν μέσω μιας επαναληπτικής διαδικασίας θα εξασφαλίσει ότι το Q θα συγκλίνει σε κάποια τιμή Q0 σε έναν πεπερασμένο αριθμό βημάτων.
Η επαναληπτική διαδικασία καλείται διαδικασία διόρθωσης σφάλματος με σταθερό βήμααναζήτησης (fixed-increment error correction).
Θα ισχύουν τότε τα εξής:

$$Q_i = Q_{i-1} + cD \quad \text{αν} \quad M(Q_{i-1}, D) - T \le 0 \quad \text{και} \quad D \subseteq A$$
$$Q_i = Q_{i-1} - cD \quad \text{αν} \quad M(Q_{i-1}, D) - T > 0 \quad \text{και} \quad D \subseteq \bar{A}$$

καμία αλλαγή δεν θα γίνει σε Qi -1 εάν ο εντοπισμός γίνει σωστά. Το c είναι η αύξηση διορθώσεων, η αξία της είναι αυθαίρετη και συνήθως την δίνουμε τιμή στη μονάδα. Στην πράξη μπορεί να είναι απαραίτητο στον κύκλο μέσω του συνόλου εγγράφων αρκετές φορές να επιτευχθεί το σωστό σύνολο βαρών, δηλαδή εκείνοι που θα χωρίσουν το A και Ā γραμμικά (αυτό παρέχει πάντα ότι μια λύση υπάρχει).
Η κατάσταση στην πραγματική ανάκτηση δεν είναι τόσο απλή. Δεν ξέρουμε τα σύνολα A και Ā εκ των προτέρων, στην πραγματικότητα A είναι το σύνολο που ελπίζουμε να ανακτήσουμε. Ωστόσο, λαμβάνοντας υπόψη μια διατύπωσης ερωτήματος Q και τα ανακτημένα από αυτή έγγραφα, μπορούμε να ζητήσουμε από το χρήστη να μας ενημερώσει το σύστημα ποια από τα έγγραφα που ανακτήθηκαν ήταν σχετικά και ποια μη σχετικά. Το σύστημα μπορεί τότε να προσαρμόσει αυτόματα το Q έτσι ώστε θα μπορεί να επιλέξει σωστά εκείνα τα έγγραφα πυυ ο χρήστης έχει δει. Η υπόθεση είναι ότι μ' αυτόν τον τρόπο θα βελτιώσουμε την ανάκτηση την επόμενη φορά βασισμένοι στο γεγονός ότι η απόδοσή του είναι καλύτερη σε ένα δείγμα.
Είναι συχνά δύσκολο να καθοριστεί το κατώτατο όριο T εκ των προτέρων, έτσι άντ' αυτού τα έγγραφα ταξινομούνται με φθίνουσα τιμή ταυτοποίησης στην έξοδο. Είναι τώρα δυσκολότερο να καθοριστεί η ιδανική διατύπωση ερώτησης. Ο Rocchio15 στη διατριβή του καθόρισε τη βέλτιστη ερώτηση Q0 ως μια που όταν μεγιστοποιηθεί δίνει:

$$\Phi = \frac{1}{|A|}\sum_{D \in A} M(Q, D) - \frac{1}{|\bar{A}|}\sum_{D \in \bar{A}} M(Q, D)$$

Εάν το M λαμβάνεται για να είναι η λειτουργία συνημίτονου (Q, D) / | |Q| | | |D| |
έπειτα είναι εύκολο να δειχτεί ότι το Φ μεγιστοποιείται όταν

$$Q_0 = c\left(\frac{1}{|A|}\sum_{D \in A}\frac{D}{\|D\|} - \frac{1}{|\bar{A}|}\sum_{D \in \bar{A}}\frac{D}{\|D\|}\right)$$

όπου c είναι μια αυθαίρετη σταθερά αναλογίας.
Εάν τα αθροίσματα αντί για τα A και Ā γίνουν τώρα για τα **A.Bi και Ā.Bi** όπου Bi είναι το σύνολο ανακτημένων εγγράφων της ith (νιοστής) επανάληψης, κατόπιν έχουμε την διατύπωση ενός ερωτήματος που είναι το καλύτερο για το Bi, ένα υποσύνολο της συλλογής εγγράφων. Κατά αναλογία με τον γραμμικό ταξινομητή που χρησιμοποιήθηκε πριν, προσθέτουμε τώρα αυτό το διάνυσμα στη διατύπωση ερώτησης στο ith (νιοστό) βήμα, για να πάρουμε:

$$Q_{i+1} = w_1 Q_i - w_2 \left[\frac{1}{|A \cap B_i|} \sum_{D \in A \cap B_i} \frac{D}{|D|} - \frac{1}{|B|} \frac{1}{|A \cap B_i|} \sum_{D \in A \cap B_i} \frac{D}{|D|} \right]$$

όπου w1 και w2 είναι συντελεστές στάθμισης. Η σημαντικότερη διαφορά που είναι ότι υπάρχει μια επιλογή να παραχθεί Qi+1 από Qi, ή Q, η αρχική ερώτηση.
Πειράματα έχουν δείξει ότι η σχετική ανατροφοδότηση μπορεί να είναι πολύ αποτελεσματική. Δυστυχώς το μέγεθος της αποτελεσματικότητας είναι μάλλον δύσκολο να μετρηθεί, δεδομένου ότι είναι μάλλον δύσκολο να διαχωρίσουμε τη συμβολή στην αυξανόμενη αποτελεσματικότητα ανάκτησης που παράγεται όταν μεμονωμένα έγγραφα μετακινούνται σε ανώτερη θέση από τη συμβολή που παράγεται όταν ανακτώνται νέα έγγραφα.
Τέλος, θα λέγαμε ότι η εφαρμογή της ανατροφοδότησης σε μία λειτουργική βάση μπορεί να είναι πιο προβληματική. Δεν είναι σαφές πώς πρόκειται οι χρήστες να αξιολογήσουν τη σχετικότητα, ή τη μη-σχετικότητα ενός εγγράφου από τέτοια ανεπαρκή στοιχεία όπως είναι οι παραπομπές.

6. Πιθανολογική Ανάκτηση (Probabilistic Retrieval)

Εχουμε κάνει πολύ λίγη χρήση της θεωρίας πιθανότητας στη διαμόρφωση οποιουδήποτε υποσυστήματος στην ανάκτηση πληροφορίας. Ο λόγος για αυτό είναι απλά ότι ο όγκος της εργασίας στην ανάκτηση πληροφορίας είναι μη-πιθανολογικός και σχετικά πρόσφατα έγινε κάποια σημαντική πρόοδος με τις πιθανολογικές μεθόδους. Η χρήση των πιθανολογικών μεθόδων ξεκινά στις αρχές της δεκαετίας του εξήντα αλλά για κάποιους λόγους οι ιδέες αυτές ποτέ δεν εφαρμόστηκαν. Εδώ θα ασχοληθούμε με τη κατανομή των όρων δεικτών σε ένα σύνολο εγγράφων που αποτελεί μια συλλογή ή ένα αρχείο. Θα στηριχτούμε στη γνωστή υπόθεση ότι η κατανομή των όρων δεικτών σε όλη τη συλλογή, ή μέσα σε κάποιο υποσύνολο της, θα μας πει κάτι για την πιθανή σχετικότητα οποιουδήποτε δεδομένου εγγράφου.
Αν και η εργασία για το μοντέλο που θα συζητηθεί εδώ είναι μάλλον πρόσφατη και έτσι μερικοί μπορούν να θεωρήσουν ότι δεν έχει δοκιμαστεί, αντιπροσωπεύει πιθανώς τη σημαντικότεροι ανακάλυψη στην ανάκτηση πληροφορίας στα τελευταία έτη. Επομένως αυτό το κεφάλαιο είναι θεωρητικό, δεδομένου ότι η θεωρία πρέπει να γίνει κατανοητή λεπτομερώς πριν σημειωθεί περαιτέρω πρόοδος. Υπάρχουν διάφοροι ισοδύναμοι τρόποι παρουσίασης της βασικής θεωρίας.
Το θεμελιώδες μαθηματικό εργαλείο για αυτό το κεφάλαιο είναι το θεώρημα Bayes: οι περισσότερες από τις εξισώσεις προέρχονται άμεσα από αυτό. Αν και τα θεμελιώδη μαθηματικά μπορεί αρχικά να φαίνονται πολύπλοκα, η ερμηνεία είναι μάλλον απλή.
Ας θυμηθούμε ότι το βασικό όργανο που έχουμε για την προσπάθεια διαχωρισμού σχετικών και μη σχετικών εγγράφων είναι μια συνάρτηση ταιριάσματος, εάν βρισκόμαστε σε ομαδοποιημένο περιβάλλον ή σε μη δομημένο. Οι λόγοι για την επιλογή οποιασδήποτε συνάρτησης ταιριάσματος δεν είναι ξεκάθαροι. Τώρα θα προσπαθήσουμε να χρησιμοποιήσουμε την απλή θεωρία πιθανοτήτων για να δούμε πώς θα έπρεπε να φαίνεται μια συνάρτηση ταιριάσματος και πώς πρέπει να χρησιμοποιηθεί. Τα επιχειρήματα είναι κυρίως θεωρητικά. Η μόνη αμφιβολία είναι για την αποδοχή των υποθέσεων. Τα στοιχεία που χρησιμοποιούνται για τον προσδιορισμό μιας τέτοιας συνάρτησης ταιριάσματος προκύπτουν από τη γνώση της κατανομής των όρων δεικτών σε όλη τη συλλογή ή κάποιου υποσυνόλου της. Εάν ορίζεται σε κάποιο υποσύνολο εγγράφων, τότε αυτό το υποσύνολο μπορεί να οριστεί από ποικίλες τεχνικές: δειγματοληψία, ομαδοποίηση, ή δοκιμαστική ανάκτηση. Τα στοιχεία που συγκεντρώνονται έτσι χρησιμοποιούνται για να θέσουν τις τιμές ορισμένων παραμέτρων που σχετίζονται με την συνάρτηση ταιριάσματος. Σαφώς, εάν τα στοιχεία περιέχουν σχετικές πληροφορίες τότε η διαδικασία ορισμού της συνάρτησης ταιριάσματος μπορεί να επαναλαμβάνεται συνεχώς από κάποιο μηχανισμό ανατροφοδότησης, παρόμοιο με αυτόν του Rocchio. Κατ' αυτό τον τρόπο οι παράμετροι της συνάρτησης ταιριάσματος μπορούν "να μαθευτούν". Θα συγκεντρωθούμε. στις συναρτήσεις ταιριάσματος που προέρχονται από σχετικές πληροφορίες.
Στη συνέχεια υποθέτουμε ότι τα έγγραφα περιγράφονται από τις ιδιότητες της δυαδικής κατάστασης, δηλαδή απουσία ή παρουσία όρων δεικτών.

Εκτίμηση ή υπολογισμός της σχετικότητας

Όταν ψάχνουμε μια συλλογή εγγράφων, προσπαθούμε να ανακτήσουμε τα σχετικά έγγραφα χωρίς ανάκτηση μη σχετικών. Δεδομένου ότι δεν έχουμε κανέναν «μάντη» που θα μας πει με βεβαιότητα ποια έγγραφα είναι σχετικά και ποια δεν είναι σχετικά, πρέπει να χρησιμοποιήσουμε την ατελή γνώση για να μαντέψουμε για οποιοδήποτε δεδομένο έγγραφο εάν είναι σχετικό ή μη σχετικό. Θα υποθέσουμε ότι μπορούμε να μαντέψουμε τη σχετικότητα μόνο μέσω των περιληπτικών στοιχείων για το έγγραφο και των σχέσεών του με άλλα έγγραφα. Αυτό δεν είναι μια αδικαιολόγητη υπόθεση ιδιαίτερα εάν θεωρούμε ότι ο μόνος τρόπος που μπορεί να καθοριστεί η σχετικότητα είναι να διαβάσει ο χρήστης το πλήρες κείμενο. Επομένως, ένας λογικός τρόπος να υπολογίσουμε την εικασία μας είναι να δοκιμάσουμε να υπολογίσουμε για οποιοδήποτε έγγραφο την πιθανότητά σχετικότητας του

$$PQ (R=\text{σχετικότητα} / D=\text{έγγραφο})$$

όπου το Q χρησιμοποιείται για να υπογραμμίσει ότι αφορά μια συγκεκριμένη ερώτηση. .εν είναι καθόλου σαφές τι είδους πιθανότητα είναι αυτή (βλέπε Good65 για μια περίληψη των διαφορετικών ειδών), αλλά εάν πρόκειται να την κατανοήσουμε με έναν υπολογιστή και τα πρωτόγονα στοιχεία που έχουμε, σίγουρα πρέπει να είναι βασισμένη σε μετρήσεις συχνότητας. Κατά συνέπεια η πιθανότητα σχετικότητας είναι μια στατιστική έννοια παρά μια σημασιολογική, αλλά πιστεύουμε ότι ο βαθμός σχετικότητας που υπολογίζεται μέσω της στατιστικής ανάλυσης, θα τείνει να είναι παρόμοιος με αυτόν που υπολογίζεται σημασιολογικά.
Ακριβώς όπως μια συνάρτηση ταιριάσματος συνδέει ένα αριθμητικό αποτέλεσμα με κάθε έγγραφο και θα ποικίλει από έγγραφο σε έγγραφο, έτσι η πιθανότητα για μερικά,

θα είναι μεγαλύτερη από ό,τι για άλλα και φυσικά θα εξαρτηθεί από την ερώτηση. Η διαφοροποίηση μεταξύ των ερωτήσεων θα αγνοηθεί προς το παρόν, γίνεται σημαντική μόνο στο στάδιο αξιολόγησης. Έτσι θα υποθέσουμε ότι μόνο μια ερώτηση έχει υποβληθεί στο σύστημα και ενδιαφερόμαστε για την πιθανότητα

$$P (R=\text{σχετικότητα}/ D=\text{έγγραφο})$$

Γώρα ας υποθέσουμε (ακολουθώντας τον Robertson66) ότι:

1) Η σχετικότητα ενός εγγράφου σε ένα αίτημα είναι ανεξάρτητη από άλλα έγγραφα στη συλλογή.

Με αυτήν την υπόθεση μπορούμε τώρα να διατυπώσουμε μια αρχή, από την άποψη της πιθανότητας της σχετικότητας, η οποία δείχνει ότι οι πιθανολογικές πληροφορίες μπορούν να χρησιμοποιηθούν κατά τρόπο βέλτιστο στην ανάκτηση.
2) Η αρχή ταξινόμησης της πιθανότητας. Εάν η απάντηση ενός συστήματος ανάκτησης αναφορών σε κάθε αίτημα, είναι μια ταξινόμηση των εγγράφων στη συλλογή, κατά φθίνουσα πιθανότητα σχετικότητας, όσον αφορά τον χρήστη που

υπέβαλε το αίτημα, τότε οι πιθανότητες υπολογίζονται με την μεγαλύτερη δυνατή ακρίβεια, με βάσει οποιαδήποτε δεδομένα είναι στην διάθεση του συστήματος για αυτόν το λόγο.
Φυσικά αυτή η αρχή δημιουργεί πολλά ερωτήματα ως προς την αποδοχή των υποθέσεων. Παραδείγματος χάριν, η Υπόθεση Συστάδων, ότι τα πολύ σχετικά έγγραφα τείνουν να είναι σχετικά με τα ίδια αιτήματα, στηρίζει το αντίθετο της υπόθεσης (1). Ο Goffman67 επίσης, στην εργασία του έχει κάνει μια σαφή υπόθεση της εξάρτησης. Αναφέρουμε: «Έτσι, εάν ένα έγγραφο x έχει αξιολογηθεί ως σχετικό με μια ερώτηση s, η σχετικότητα των άλλων εγγράφων στο αρχείο X μπορεί να επηρεαστεί, δεδομένου ότι η αξία των πληροφοριών που μεταβιβάζονται από αυτά τα έγγραφα, μπορεί είτε να αυξηθεί είτε να μειωθεί, ως αποτέλεσμα των πληροφοριών που μεταβιβάζονται από το έγγραφο x.» Κατόπιν υπάρχει το θέμα του τρόπου με τον οποίο η γενική αποτελεσματικότητα πρόκειται να μετρηθεί. Ο Robertson στην εργασία του παρουσιάζει την αρχή που ταξινομεί την πιθανότητα, για να καθορίσει εάν μετράμε την αποτελεσματικότητα βάσει της ανάκλησης και του fallout (= το ποσοστό των μη-σχετικών εγγράφων που ανακτώνται).
Προειδοποιητικά η αρχή της ταξινόμησης της πιθανότητας μπορεί να αποδειχθεί ότι ισχύει μόνο για μια ερώτηση. Δεν λέει ότι η απόδοση σε μια σειρά ερωτήσεων θα βελτιστοποιηθεί, για να καθιερωθεί ένα αποτέλεσμα αυτού του είδους θα πρέπει να είμαστε συγκεκριμένοι για το πώς θα υπολογίζαμε τον μέσο όρο της απόδοσης σε όλες τις ερωτήσεις.
Η αρχή της ταξινόμησης της πιθανότητας υποθέτει ότι μπορούμε να υπολογίσουμε το P(R/D) και ότι μπορούμε να το κάνουμε με ακρίβεια. Αυτή είναι μια εξαιρετικά προβληματική υπόθεση και θα μας απασχολήσει περισσότερο. Το πρόβλημα είναι ότι δεν ξέρουμε πια είναι τα σχετικά έγγραφα, ούτε πόσα υπάρχουν, έτσι δεν έχουμε κανέναν τρόπο υπολογισμού του P (R/D). Αλλά μπορούμε, από τη δοκιμαστική ανάκτηση, να μαντέψουμε το P (R/D) και να βελτιώσουμε ενδεχομένως την εικασία μας από την επανάληψη. Για να απλοποιήσουμε τα πράγματα, θα υποθέσουμε ότι τα στατιστικά στοιχεία, που σχετίζονται με τα σχετικά και μη σχετικά έγγραφα, είναι διαθέσιμα και θα τα χρησιμοποιήσουμε για να φτιάξουμε τις σχετικές εξισώσεις.

Το άμεσο πρόβλημα είναι, να υπολογίσουμε, ή να εκτιμήσουμε το P(R/D). Για αυτό χρησιμοποιούμε το θεώρημα Bayes, το οποίο σχετίζει τη μεταγενέστερη πιθανότητα της σχετικότητας με την προγενέστερη πιθανότητα της σχετικότητας και την πιθανότητα της σχετικότητας μετά την παρατήρηση ενός εγγράφου. Πρέπει να δούμε μερικά σύμβολα, που θα καταστήσουν τα πράγματα λίγο ευκολότερα καθώς προχωράμε.

Βασικό πιθανολογικό πρότυπο

Δεδομένου ότι υποθέτουμε ότι κάθε έγγραφο περιγράφεται από την παρουσία/ απουσία όρων δεικτών οποιοδήποτε έγγραφο μπορεί να αντιπροσωπευθεί από ένα δυαδικό διάνυσμα **X** = (x1, x2. .., xn) όπου xi = 0 ή 1 δείχνει την απουσία ή την παρουσία του όρου δεικτών του i. Επίσης υποθέτουμε ότι υπάρχουν δύο αμοιβαία αποκλειόμενα γεγονότα,

w1 = το έγγραφο είναι σχετικό

w2 = το έγγραφο είναι μη σχετικό.

Με βάση αυτά τα σύμβολα, επιθυμούμε να υπολογίσουμε για κάθε έγγραφο το P(w1 / **x**) και ίσως το P(w2 / **x**) έτσι ώστε μπορούμε να αποφασίσουμε ποιο είναι σχετικό και ποιο είναι μη σχετικό. Αυτό είναι μια μικρή αλλαγή στο στόχο, αντί απλά να παραγάγει μια ταξινόμηση, επιθυμούμε επίσης η θεωρία να μας πει πώς να διακόψουμε την ταξινόμηση. Επομένως διατυπώνουμε το πρόβλημα ως πρόβλημα απόφασης. Φυσικά δεν μπορούμε να υπολογίσουμε το P(wi / **x**) άμεσα έτσι πρέπει να βρούμε έναν τρόπο υπολογισμού του με βάση τις ποσότητες, για τις οποίες ξέρουμε κάτι. Το Θεώρημα Bayes για διακριτές κατανομές δίνει:

$$P(w_i|x) = \frac{P(x/w_i)P(w_i)}{P(x)} \qquad i = 1, 2$$

Εδώ το P(wi) είναι η προγενέστερη πιθανότητα σχετικότητας (i = 1) ή μη-σχετικότητας (i = 2), P(**x** / wi) είναι ανάλογη προς την πιθανότητα σχετικότητας ή μη σχετικότητας, δοθέντος του **x**, σε περίπτωση συνέχειας αυτή θα ήταν μια συνάρτηση πυκνότητας και θα γράφαμε p(**x** / wi). Τέλος,

$$P(x) = \sum_{i=1}^{2} P(x/w_i)\,P(w_i),$$

η οποία είναι η πιθανότητα της παρατήρησης του x σε τυχαία βάση δεδομένου ότι μπορεί να είναι είτε σχετικό είτε μη-σχετικό. Πάλι αυτό θα γραφόταν ως συνάρτηση πυκνότητας p(x) στη συνεχή περίπτωση. Αν και το P (**x**) (ή p (**x**)) θα εμφανιστεί συνήθως όπως ένας παράγοντας κανονικοποίησης (δηλ. εξασφαλίζοντας ότι το **P (w1 / x) + P (w2 / x) = 1**) αυτό είναι υπό μία έννοια η συνάρτηση για την οποία ξέρουμε τα περισσότερα, δεν απαιτεί γνώση της σχετικότητας για να προσδιοριστεί. Πριν συζητηθεί πώς φτάνουμε στον υπολογισμό της δεξιάς πλευράς του θεωρήματος Bayes θα δούμε πώς λαμβάνεται η απόφαση υπέρ ή κατά της σχετικότητας.
Ο κανόνας απόφασης που χρησιμοποιούμε είναι ευρέως γνωστός ως ο κανόνας απόφασης του Bayes. Συγκεκριμένα είναι

$$[P(w_1/x) > P(w_2/x) \rightarrow \text{το } x \text{ είναι σχετικό, αλλιώς το } x \text{ είναι μη-σχετικό}] \qquad (D1)$$

Η έκφραση D1 είναι μια συντομογραφία για τα εξής: συγκρίνουμε το P (w1 / **x**) με το P(w2 / **x**) εάν η πρώτη είναι μεγαλύτερη από την δεύτερη, τότε αποφασίζουμε ότι το **x** είναι σχετικό διαφορετικά αποφασίζουμε ότι το **x** είναι μη σχετικό. Η υπόθεση P (w1 /

x) = P(w2 / **x**) αντιμετωπίζεται αυθαίρετα με την απόφαση μη-σχετικότητας. Η βάση για τον κανόνα D1 είναι απλά ότι ελαχιστοποιεί τη μέση πιθανότητα του λάθους, το λάθος του ορισμού ενός σχετικού εγγράφου ως μη σχετικό ή αντίστροφα για οποιοδήποτε **x** η πιθανότητα του λάθους είναι

$$P[\text{error}|x] = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1 \end{cases}$$

Με άλλα λόγια μόλις αποφασίσουμε μία κατάσταση (π.χ. σχετικό) τότε η πιθανότητα να κάνουμε λάθος δίνεται από την πιθανότητα της αντίθετης κατάστασης να ισχύει (π.χ. μη σχετικό). Έτσι για να κατακτήσουμε αυτό το λάθος όσο το δυνατόν μικρότερο για οποιοδήποτε δεδομένο **x** πρέπει πάντα να διαλέγουμε αυτό το wi για το οποίο το p(w1 / **x**) είναι το μεγαλύτερο και συνεπώς αυτό για το οποίο η πιθανότητα λάθους είναι η μικρότερη. Για να ελαχιστοποιήσουμε τη μέση πιθανότητα του λάθους πρέπει να ελαχιστοποιήσουμε την ποσότητα

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)\, dx = \int_{-\infty}^{\infty} P(\text{error}|x)\, P(x)\, dx$$

Αυτό το άθροισμα θα ελαχιστοποιηθεί κάνοντας το P (λάθος /**x**) όσο το δυνατόν μικρότερο για κάθε **x** δεδομένου ότι το P (λάθος /**x**) και P (**x**) είναι πάντα θετικό. Αυτό κατορθώνεται με τον κανόνα απόφασης D1 που τώρα δικαιολογείται.
Φυσικά το μέσο λάθος δεν είναι η μόνη λογική ποσότητα που αξίζει να ελαχιστοποιηθεί. Εάν σχετίσουμε με κάθε τύπο λάθους ένα κόστος μπορούμε να παραγάγουμε έναν κανόνα απόφασης που θα ελαχιστοποιήσει το γενικό κίνδυνο. Το

συνολικό σφάλμα είναι ένας μέσος όρος των υπό όρους κινδύνων **R(wi / x)** ο οποίος στη συνέχεια καθορίζεται από μια συνάρτηση κόστους **li j**. Πιο συγκεκριμένα **li j** είναι η απώλεια που υφίσταται για την απόφαση του **wi** όταν το **wj** ισχύει. Τώρα η σχετική αναμενόμενη απώλεια, όταν αποφασίζεται το **wi,** λέγεται υπό όρους κίνδυνος και δίνεται από τη σχέση

$$R(w_i / x) = l_{i1} P(w_1 / x) + l_{i2} P(w_2 / x) \quad i = 1, 2$$

Το συνολικό λάθος είναι ένα άθροισμα, όπως ήταν η μέση πιθανότητα του λάθους. Το R (wi / x) τώρα παίζει το ρόλο του P (wi / x). Το συνολικό λάθος ελαχιστοποιείται σε [R (w1/x) < R (w2/x) > x είναι σχετικό, αλλιώς το x είναι μη-σχετικό] (D2)

D1 και D2 μπορεί να αποδειχθεί ότι είναι ισοδύναμα υπό ορισμένους όρους. Πρώτα ξαναγράφουμε το D1, χρησιμοποιώντας το θεώρημα Bayes, σε μια μορφή στην οποία θα χρησιμοποιηθεί στη συνέχεια, δηλαδή:

[P(x / w1) P (w1) > P(x / w2) P(w2) > το x είναι σχετικό, αλλιώς το x είναι μη-σχετικό] (D3)

Το P(x) έχει εξαφανιστεί από την εξίσωση δεδομένου ότι δεν έχει επιπτώσεις στην έκβαση της απόφασης. Τώρα, χρησιμοποιώντας τον ορισμό R(wi / x) είναι εύκολο να δειχτεί ότι

$$[R(w_1 / x) < R(w_2 / x)] \equiv [(l_{21} - l_{11}) P(x / w_1) P(w_1) > (l_{12} - l_{22}) P(x / w_2) P(w_2)]$$

Όταν μια ειδική συνάρτηση απώλειας επιλέγεται, δηλαδή

$$l_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

η οποία υπονοεί ότι καμία απώλεια δεν ορίζεται σε μια σωστή απόφαση (αρκετά λογικό) και απώλεια μονάδας σε οποιοδήποτε λάθος (όχι τόσο λογικό), κατόπιν έχουμε

$$[R(w_1 / x) < R(w_2 / x)] \equiv [P(x / w_1) P(w_1) > P(x / w_2) P(w_2)]$$

η οποία παρουσιάζει την ισοδυναμία D2 και D3, και ως εκ τούτου D1 και D2 κάτω από μια συνάρτηση δυαδικής απώλειας.
Αυτό ολοκληρώνει την παραγωγή του κανόνα απόφασης που χρησιμοποιείται για να αποφασίσει τη σχετικότητα ή την μη-σχετικότητα, ή για να το θέσουμε διαφορετικά για να ανακτήσει ή για να μην ανακτήσει. Μέχρι τώρα κανένας περιορισμός δεν έχει τεθεί στη μορφή του P(**x** / w1), επομένως ο κανόνας απόφασης είναι αρκετά γενικός. Το πρόβλημα έχει οργανωθεί ως πρόβλημα απόφασης μεταξύ δύο κατηγοριών, αγνοώντας το πρόβλημα της ταξινόμησης προς το παρόν. Ένας λόγος για αυτό είναι ότι η ανάλυση είναι πιο απλή, ένας άλλος είναι ότι θέλουμε η ανάλυση να δίνει όσο το δυνατόν περισσότερες πληροφορίες για την τιμή του κατωφλίου αποκοπής. Κατά την

ταξινόμηση, η τιμή αυτή αφήνεται συνήθως στο χρήστη, μέσα στο πρότυπο ως τώρα κάποιος μπορεί ακόμα να ταξινομήσει, αλλά η τιμή του κατωφλίου αποκοπής θα ερμηνεύεται με βάση τις προγενέστερες πιθανότητες και τις συναρτήσεις κόστους. Η βελτιστότητα της πιθανότητας που ταξινομεί τον κανόνα προκύπτει αμέσως από βελτιστότητα του κανόνα απόφασης σε οποιαδήποτε διακοπή. Τώρα θα προχωρήσουμε στην ακριβή μορφή των συναρτήσεων πιθανότητας στον κανόνα απόφασης.

Μορφή λειτουργίας ανάκτησης

Αν και είναι λογικό να θελήσουμε να υπολογίσουμε το P (σχετικότητα / έγγραφο), δεν είναι καθόλου σαφής ο τρόπος με τον οποίο αυτό πρέπει να γίνει ή εάν η αντιστροφή μέσω του θεωρήματος Bayes είναι ο καλύτερος τρόπος για αυτό. Εντούτοις, θα προχωρήσουμε υποθέτοντας ότι το P(x / wi) είναι η κατάλληλη συνάρτηση για να εκτιμηθεί. Αυτή η συνάρτηση είναι φυσικά μια συνδυαστική συνάρτηση πιθανότητας και η αλληλεπίδραση μεταξύ των συνιστωσών του x μπορεί να είναι αυθαίρετα σύνθετη. Για να παραχθεί ένας εφαρμόσιμος κανόνας απόφασης, θα πρέπει να γίνει μια απλοποιημένη υπόθεση για το P(x / wi). Ο συμβατικός μαθηματικός βολικός τρόπος να απλοποιηθεί το P(x / wi) είναι να υποθέσουμε ότι οι συντεταγμένες xi του x

να είναι ανεξάρτητες. Αυτό σημαίνει ότι η κύρια υπόθεση γίνεται

$$P(\mathbf{x} / w_i) = P(x_1 / w_i)\, P(x_2 / w_i) \ldots P(x_n / w_i) \qquad A1$$

Πιο κάτω θα δείξουμε πώς αυτήν η αυστηρή υπόθεση μπορεί να γίνει λιγότερη αυστηρή. Επίσης, αγνοούμε το γεγονός ότι υποθέτοντας υπό όρους ανεξαρτησία στα w1 και w2 χωριστά έχει επιπτώσεις στην υπό όρους εξάρτηση του w1 και w2. Παίρνουμε τώρα την απλουστευμένη μορφή *του* P(**x** / wi) και υπολογίζουμε την μορφή που θα έχει ο κανόνας απόφασης. Πρώτα καθορίζουμε μερικές μεταβλητές

$$p_i = \text{Prob}\,(x_i = 1 / w_1)$$

$$q_i = \text{Prob}\,(x_i = 1 / w_2).$$

Το pi (qi) είναι η πιθανότητα ότι εάν το έγγραφο είναι σχετικό (μη σχετικό) τότε ο όρος δεικτών i-th θα είναι παρών. Οι αντίστοιχες πιθανότητες για την απουσία υπολογίζονται αφαιρώντας από το 1, δηλ. 1 – pi = Prob (xi = 0 / w1). Οι συναρτήσεις πιθανότητας που εισάγονται στο D3 τώρα έχουν την μορφή

$$P(x / w_1) = \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{1-x_i}$$

$$P(x / w_2) = \prod_{i=1}^{n} q_i^{x_i} (1 - q_i)^{1-x_i}$$

Για να εκτιμήσουμε πώς αυτές οι εκφράσεις λειτουργούν, θα ελεγξουμε το P((0,1,1,0,0,1) / w1) = (1 – p1) p2 p3 (1 – p4)(1 – p5) p6.
Αντικαθιστώντας για το P(**x** / wi) στο D3 και παίρνοντας τους λογαρίθμους , ο κανόνας απόφασης θα μετασχηματιστεί σε μια γραμμική διακρίνουσα συνάρτηση.

$$g(\mathbf{x}) = \sum_{i=1}^{n} (a_i x_i + b_i (1 - x_i)) + c$$

$$= \sum_{i=1}^{n} c_i x_i + c$$

όπου οι σταθερές ai,bi και το c είναι προφανείς.

$$c_i = \log \frac{p_i (1 - q_i)}{q_i (1 - p_i)}$$

και

$$C = \sum_{i=1}^{n} \log \frac{(1-p_i)}{(1-q_i)} + \log \frac{f(w_1)}{f(w_2)} + \log \frac{l_{21} - l_{11}}{l_{12} - l_{22}}$$

Η σημασία της γραφής της με αυτόν τον τρόπο, εκτός από την απλότητά της, είναι ότι για κάθε έγγραφο **x** για να υπολογίσουμε το g(**x**) προσθέτουμε απλά τους συντελεστές ci για εκείνους τους όρους δεικτών που είναι παρόντες, δηλ. για εκείνο το ci για το οποίο xi = 1. Τα ci συχνά θεωρούνται **ως βάρη**. Οι Robertson και Sparck Jones1 **καλούν το ci ένα βάρος σχετικότητας**, και ο Salton καλεί το exp(ci) **σχετικότητα όρου**. Θα αναφερθούμε σε αυτό απλά ως συντελεστή ή βάρος. Ως εκ τούτου το g (**x**) ονομάζεται συνάρτηση στάθμισης.
Το σταθερό C που υποθέτουμε ότι είναι το ίδιο για όλα τα έγγραφα **x** θα ποικίλει φυσικά από ερώτηση σε ερώτηση, αλλά μπορεί να ερμηνευθεί ως διακοπή που εφαρμόζεται στη συνάρτηση ανάκτησης. Το μόνο μέρος το οποίο μπορεί να ποικίλει όσον αφορά μια δεδομένη ερώτηση είναι η συνάρτηση κόστους και είναι αυτή η ποικιλία που θα μας επιτρέψει να ανακτήσουμε λίγα ή πολλά έγγραφα. Για να γίνει κατανοητό, ας υποθέσουμε ότι l11 = l22 = 0 και ότι έχουμε κάποια επιλογή στον καθορισμό της αναλογίας l21 / l11 με την επιλογή μιας τιμής για τη σχετική σημασία που αποδίδουμε στην απώλεια ενός σχετικού εγγράφου έναντι της ανάκτησης ενός μη σχετικού. Κατ' αυτό τον τρόπο μπορούμε να παραγάγουμε μια ταξινόμηση, κάθε θέση

της οποίας αντιστοιχεί σε μια διαφορετική αναλογία l21 / l12 .
Επιστρέφουμε στο άλλο μέρος της συνάρτησης g (**x**), δηλαδή ci και θα δοκιμάσουμε να το ερμηνεύσουμε μέσω του συμβατικού πίνακα "πιθανότητας".

	Σχετικά	**Μη-Σχετικά**	
$x_i = 1$	r	n - r	n
$x_i = 0$	**R - r**	**N - n - R + r**	**N - n**
	R	N - R	N

Θα υπάρξει ένας τέτοιος πίνακας για κάθε όρο δεικτών. Εδώ παρουσιάζεται για τον όρο δεικτών i αν και ο δείκτης i δεν έχει χρησιμοποιηθεί στα κελιά. Εάν έχουμε *πλήρεις* πληροφορίες για τα σχετικά και μη σχετικά έγγραφα στη συλλογή, τότε μπορούμε να υπολογίσουμε το pi από r/R και qi από (n - r) / (N - R). Επομένως η συνάρτηση g (**x**) μπορεί να ξαναγραφεί ως εξής:

$$g(x) = \sum_{i=1}^{n} x_i \log \frac{r(R-r)}{(n-r)(N-n-R+r)} + C$$

Αυτό είναι στην πραγματικότητα ο τύπος στάθμισης F4 που χρησιμοποιείται από τους Robertson και Sparck Jones στα αποκαλούμενα αναδρομικά πειράματά τους. Για ευκολία θέτουμε:

$$K(N, r, n, R) = \log \frac{r(R-r)}{(n-r)(N-n-R+r)}$$

Υπάρχουν διάφοροι τρόποι να εξεταστεί το Ki. Η πιο ενδιαφέρουσα ερμηνεία του Ki είναι να ειπωθεί ότι μετρά το βαθμό στον οποίο ο ith (ιοστός) όρος μπορεί να διακρίνει μεταξύ των σχετικών και μη σχετικών εγγράφων.

Χαρακτηριστικά το "βάρος" Ki (N, r, n, R) υπολογίζεται από έναν πίνακα πιθανότητας στον οποίο το N δεν είναι ο συνολικός αριθμός των εγγράφων του συστήματος αλλά αντ' αυτού είναι κάποιο υποσύνολο που επιλέγεται συγκεκριμένα για να επιτρέψει τον υπολογισμό του Ki. Πιο μετά θα χρησιμοποιηθεί η ανωτέρω ερμηνεία του Ki για να εξηγήσουμε μια άλλη συνάρτηση, παρόμοια με το Ki, για να μετρήσει τη διακριτική δύναμη ενός όρου δεικτών.

Οι όροι δεικτών δεν είναι ανεξάρτητοι

Αν και μπορεί να είναι από μαθηματική άποψη κατάλληλο να υποτεθεί ότι οι όροι δεικτών είναι ανεξάρτητοι, δεν σημαίνει ότι είναι. Η αντίρρηση στην ανεξαρτησία δεν είναι νέα, το 1964 ο H.H Williams68 το εξέφρασε με αυτόν τον τρόπο: "Η υπόθεση της ανεξαρτησίας των λέξεων σε ένα έγγραφο συνήθως τίθεται για λόγους μαθηματικής ευκολίας. Χωρίς την υπόθεση, πολλές από τις επόμενες μαθηματικές σχέσεις δεν θα μπορούσαν να εκφραστούν. Πολλά από τα συμπεράσματα θα έπρεπε να γίνουν αποδεκτά με πολύ μεγάλη προσοχή." Επειδή τα μαθηματικά γίνονται δυσεπίλυτα εάν υποθέσουμε εξάρτηση γι' αυτόν το λόγο υποθέτουμε την ανεξαρτησία. Αλλά, "η εξάρτηση είναι ο κανόνας παρά το αντίθετο" λέει ο διάσημος θεωρητικός των πιθανοτήτων De Finetti69. Επομένως η σωστή διαδικασία είναι να υποτεθεί η εξάρτηση και να επιτραπεί η ανάλυση για να απλοποιήσει στην ανεξάρτητη περίπτωση εάν η ανεξαρτησία ισχύει. Όταν μιλάμε για εξάρτηση εδώ εννοούμε την πιθανολογική εξάρτηση. Για τα δεδομένα της ανάκτησης πληροφορίας, η πιθανολογική εξάρτηση μετριέται απλά από μια συνάρτηση συσχέτισης ή με κάποιο άλλο ισοδύναμο τρόπο. Η υπόθεση της εξάρτησης θα μπορούσε να είναι κρίσιμη όταν προσπαθούμε να υπολογίσουμε *το P*(σχετικότητα / έγγραφο) με βάση *το P*(***x*** / *wi*) δεδομένου ότι η ακρίβεια με την οποία αυτή η τελευταία πιθανότητα υπολογίζεται, αναμφισβήτητα έχει επιπτώσεις στην απόδοση ανάκτησης. Έτσι ο άμεσος στόχος μας είναι να χρησιμοποιήσουμε την εξάρτηση (συσχετισμός) μεταξύ των όρων δεικτών για να βελτιώσουμε την εκτίμηση του P(***x*** / *wi*) στην οποία στηρίζεται ο κανόνας απόφασής.

Γενικά η εξάρτηση μπορεί να είναι αυθαίρετα σύνθετη όπως φαίνεται παρακάτω

$$P(\mathbf{x}) = P(x_1)P(x_2/x_1)P(x_3/x_1,x_2) \ldots P(x_n/x_1,x_2, \ldots, x_{n-1})$$

Επομένως, για να συμπεριλάβουμε όλα τα στοιχεία εξάρτησης θα πρέπει να προσαρμόσουμε κάθε μεταβλητή με τη σειρά της σε ένα σταθερά αυξανόμενο σύνολο άλλων μεταβλητών. Αν και σε γενικές γραμμές αυτό μπορεί να είναι δυνατό, είναι πιθανό να είναι υπολογιστικά ανεπαρκές και αδύνατο σε μερικές περιπτώσεις όπου υπάρχουνε ανεπαρκή στοιχεία για να υπολογιστούν οι εξαρτήσεις ανωτέρας τάξεως. Αντ' αυτού υιοθετούμε μια μέθοδο προσέγγισης για να υπολογίσουμε το P(x) που συμπεριλαμβάνει τις σημαντικές πληροφορίες εξάρτησης. Αυτό μπορεί να περιγραφεί ως μια μέθοδος που εξετάζει κάθε παράγοντα στην ανωτέρω επέκταση και επιλέγει από τις προσαρμοσμένες μεταβλητές μια συγκεκριμένη μεταβλητή, που ευθύνεται για το μεγαλύτερο μέρος της σχέσης εξάρτησης. Με άλλα λόγια επιδιώκουμε μια προσέγγιση της μορφής

$$P_t(x) = \prod_{i=1}^{n} P(x_{m_i} / x_{m_{j(i)}}) \qquad 0 \le j(i) < i \qquad \text{A2}$$

όπου (m1, m2,..., mn) είναι εναλλαγές των ακέραιων αριθμών 1,2,..., n και το j(.) είναι μια συνάρτηση που απεικονίζει το i σε ακέραιους αριθμούς μικρότερους από i και το

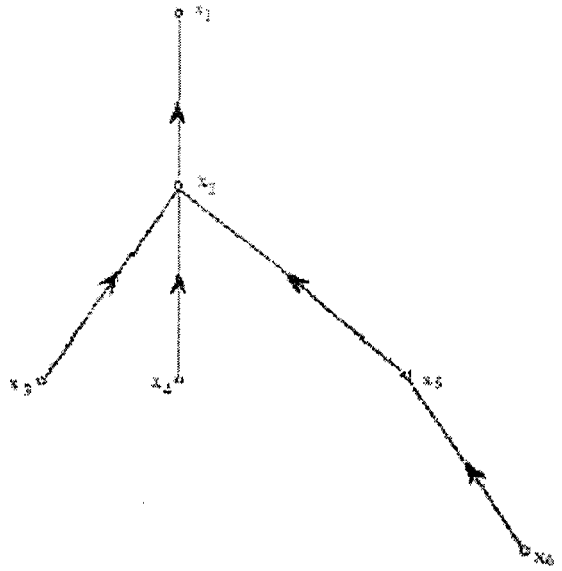
P(xi / xm0) είναι P (xi). Ένα παράδειγμα για ένα διάνυσμα έξι συστατικών
x = (x1,..., x6) μπορεί να είναι

$$P_t(x) = P(x_3)P(x_2/x_3)P(x_5/x_2)P(x_4/x_2)P(x_1/x_3)P(x_6/x_3)$$

Προσέξτε πόσο παρόμοια είναι η A2 υπόθεση, στην υπόθεση ανεξαρτησίας A1, η μόνη διαφορά είναι ότι στην A2 κάθε παράγοντας έχει μια συσχετισμένη προσαρμοσμένη μεταβλητή. Στο παράδειγμα η εναλλαγή (m1, m2, ..., m6) είναι (1,2,..., 6) που είναι απλά η φυσική διάταξη, φυσικά ο λόγος για την επέκταση Pt (**x**) με τον τρόπο που έγινε στο A2 είναι ότι πρέπει να αναζητηθεί εναλλαγή (1,2,...,6) το οποίο δίνει μια καλή προσέγγιση. Μόλις βρεθεί αυτή η εναλλαγή οι μεταβλητές θα μπορούσαν να ονομαστούν εκ νέου ώστε να υπάρξει φυσική διάταξη πάλι.

Η εναλλαγή και η συνάρτηση j(.) μαζί καθορίζουν ένα δέντρο εξάρτησης και το αντίστοιχο Pt(**x**) καλείται κατανομή πιθανότητας της εξάρτησης δέντρων πρώτης τάξεως(first-order). Το δέντρο αντιστοιχεί στο παράδειγμα των έξι μεταβλητών που φαίνεται στο σχήμα 26. Το δέντρο δείχνει ποιες μεταβλητές εμφανίζονται σε κάθε πλευρά του P (./.). Αν και έχουμε επιλέξει να γράψουμε τη συνάρτηση Pt(**x**) με xi ως την άνευ όρων μεταβλητή και ως εκ τούτου η ρίζα του δέντρου και όλες οι άλλες προσάρμοσαν με συνέπεια, κάθε μία στον κόμβο γονέων της, στην πραγματικότητα οποιοσδήποτε από τους κόμβους του δέντρου θα μπορούσε να επιλεχτεί ως ρίζα, εφ' όσον γίνεται με συνέπεια η προσαρμογή, όσον αφορά το νέο κόμβο ρίζας. (Στο σχήμα 26 η "κατεύθυνση" της προσαρμογής σημειώνεται από την κατεύθυνση που σχετίζεται

με μια άκρη). Η Pt(x) θα είναι η ίδια όπως μπορεί εύκολα να παρουσιαστεί χρησιμοποιώντας το γεγονός ότι



Σχήμα 26

$$P(x_{\omega_i} / x_{\omega_{j(i)}}) = P(x_{\omega_{j(i)}} / x_{\omega_i}) P(x_{\omega_i}) / P(x_{\omega_{j(i)}})$$

Η εφαρμογή αυτού στη σύνδεση μεταξύ του κόμβου ρίζας x1 και του άμεσου απογόνου της x2 στο παράδειγμα θα μετατοπίσει τη ρίζα x2 και θα αλλάξει την επέκταση σε

$$P_t(x_1, x_2, \ldots x_6) = P(x_2)P(x_1/x_2)P(x_3/x_2)P(x_4/x_2)P(x_5/x_2)P(x_6/x_5)$$

Φυσικά, για να τηρήσουμε τον κανόνα της εκ νέου ονομασίας θα ανταλλάσσαμε τα ονόματα "1" και "2". Όλες οι επεκτάσεις που μετασχηματίζονται κατ' αυτό τον τρόπο είναι ισοδύναμες όσον αφορά την αποδοτικότητα της προσέγγισης με το P (x). Είναι επομένως το δέντρο που αντιπροσωπεύει την κατηγορία ισοδύναμων επεκτάσεων. Σαφώς υπάρχει ένας μεγάλος αριθμός πιθανών δέντρων εξάρτησης, το πρόβλημα προσέγγισης που έχουμε είναι να βρούμε το καλύτερο, το οποίο ουσιαστικά σημαίνει την εύρεση της καλύτερης εναλλαγής και την απεικόνιση του j(.).
Από εδώ και πέρα υποθέτουμε ότι η εκ νέου ονομασία έχει γίνει και ότι xmi = xi.

Επιλογή των καλύτερων δέντρων εξάρτησης

Το πρόβλημά μας τώρα είναι να βρούμε μια συνάρτηση πιθανότητας του τύπου Pt(**x**) σε ένα σύνολο εγγράφων η οποία είναι η καλύτερη προσέγγιση στην πραγματική συσχετιστική συνάρτηση πιθανότητας P(x), και φυσικά μια καλύτερη προσέγγιση από αυτή που δίνεται κάνοντας την υπόθεση Α1. Το σύνολο στο οποίο η προσέγγιση καθορίζεται μπορεί να είναι αυθαίρετο, να είναι η ολόκληρη συλλογή, τα σχετικά έγγραφα (w1), ή τα μη σχετικά έγγραφα (w2). Προς το παρόν θα αφήσουμε το σύνολο απροσδιόριστο και τα τρία είναι σημαντικά. Εντούτοις, κατά την κατασκευή ενός κανόνα απόφασης παρόμοιο με το D4 θα πρέπει να προσεγγίσουμε το P(x/w1) και P(x/w2).
Η ποιότητα της προσέγγισης μετριέται από μια καλά γνωστή συνάρτηση (παραδείγματος χάριν, Kullback70) εάν το P(**x**) και το Pa(**x**) είναι δύο διακριτές κατανομές πιθανοτήτων τότε

$$I(P,P_a) = \sum_x P(x) \log \frac{P(x)}{P_a(x)}$$

είναι ένα μέτρο του βαθμού στον οποίο το Pa(x) προσεγγίζει το P(x). Όσον αφορά αυτήν την συνάρτηση θέλουμε να βρούμε μια κατανομή της εξάρτησης δέντρου Pt(x) τέτοια ώστε το I (P, Pt) να είναι ελάχιστο. Ή για να το θέσουμε διαφορετικά να βρούμε το δέντρο εξάρτησης μεταξύ όλων των δέντρων εξάρτησης που θα καταστήσουν το I(P, Pt) όσο το δυνατόν μικρότερο.
Εάν ο βαθμός στον οποίο δύο όροι δεικτών i και j παρεκκλίνουν από την ανεξαρτησία μετριέται από το αναμενόμενο αμοιβαίο μέτρο πληροφοριών (EMIM), τότε

$$I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

Κατόπιν η καλύτερη προσέγγιση Pt(**x**), από την άποψη της ελαχιστοποίησης του I(P, Pt), δίνεται από το μέγιστο spanning δέντρο (MST), σχετικά με τις μεταβλητές x1, x2,..., xn. Το spanning δέντρο προέρχεται από τη γραφική παράσταση της οποίας οι κόμβοι είναι οι όροι δεικτών 1,2,...,n και του οποίου τα άκρα είναι σταθμισμένα με το I(xi, xj). Το MST είναι απλά το δέντρο που καλύπτει τους κόμβους για τους οποίους το συνολικό βάρος

$$\sum_{i=1}^{n} I(x_i, x_{j(i)})$$

είναι ένα μέγιστο. Μια λεπτομερής απόδειξη της διαδικασίας βελτιστοποίησης δίνεται από τους Chow και Liu71. Εδώ ενδιαφερόμαστε κυρίως για την εφαρμογή της δομής δέντρων.
Το MST ενσωματώνει την σημαντικότερη των εξαρτήσεων μεταξύ των μεταβλητών, οι οποίες υπόκεινται στον περιορισμό ότι το άθροισμά τους πρέπει να είναι το μέγιστο. Παραδείγματος χάριν, στο σχήμα 26 οι συνδέσεις μεταξύ των μεταβλητών (κόμβοι,

x1,..., x6) δίνονται ακριβώς επειδή το άθροισμα

$$I(x_1,x_2) + I(x_2,x_3) + I(x_2,x_4) + I(x_2,x_5) + I(x_5/x_6)$$

είναι μέγιστο. Οποιοδήποτε άλλο άθροισμα θα είναι μικρότερο ή ίσο προς αυτό το άθροισμα. Προσέξτε ότι αυτό δεν σημαίνει ότι οποιοδήποτε μεμονωμένο βάρος που συνδέεται με μια άκρη στο δέντρο, θα είναι μεγαλύτερο από ένα που δεν είναι στο δέντρο, αν και αυτό συμβαίνει συνήθως.
Μόλις βρεθεί το δέντρο η κατανομή προσέγγισης μπορεί να γραφεί με την μορφή Α2. Από αυτό μπορούμε να παράγουμε μια συνάρτηση διάκρισης ακριβώς όπως κάναμε στην ανεξάρτητη περίπτωση.

$$t_i = \text{Prob}\,(x_i = 1/x_{j(i)} = 1)$$
$$r_i = \text{Prob}\,(x_i = 1/x_{j(i)} = 0) \quad \text{και} \quad r_i = \text{Prob}\,(x_i = 1)$$
$$P(x_i/x_{j(i)}) = [t_i^{x_i}(1-t_i)^{1-x_i}]^{x_{j(i)}}\,[r_i^{x_i}(1-r_i)^{1-x_i}]^{1-x_{j(i)}}$$

τότε

$$\log P(x) = \sum_{i=1}^{n}[x_i \log r_i + (1-x_i)\log(1-r_i)] +$$
$$+ \sum_{i=1}^{n}\left[x_{j(i)} \log\frac{1-t_i}{1-r_i} + x_i x_{j(i)} \log\frac{t_i(1-r_i)}{(1-t_i)r_i}\right] + \text{σταθερά}$$

Αυτή είναι μια μη γραμμική συνάρτηση στάθμισης που θα απλοποιήσει αυτή που προέρχεται από το Α1 όταν υποθέτουμε ότι οι μεταβλητές είναι ανεξάρτητες, δηλαδή όταν ti = ri. Η σταθερά έχει την ίδια ερμηνεία όσον αφορά τις προγενέστερες πιθανότητες και την συνάρτηση απώλειας. Η πλήρης συνάρτηση απόφασης είναι
g(x) = log P(x/w1) - log P(x/w2)

η οποία περιλαμβάνει τον υπολογισμό (ή την εκτίμηση) διπλάσιων παραμέτρων απ' ό,τι στη γραμμική περίπτωση. Είναι μόνο το άθροισμα που περιλαμβάνει τα xj(i), που καθιστά αυτήν την συνάρτηση στάθμισης διαφορετική από τη γραμμική και είναι αυτό το μέρος που επιτρέπει σε μια στρατηγική ανάκτησης να λάβει υπόψη το γεγονός ότι το xi εξαρτάται από το xj(i). Κατά τη χρησιμοποίηση της συνάρτησης στάθμισης ένα έγγραφο που περιέχει τα xj(i), ή τα xi και τα xj(i), θα λάβει μια συμβολή από εκείνο το μέρος της συνάρτησης στάθμισης.

Είναι ευκολότερο να δούμε πώς το g(x) συνδυάζει διαφορετικά βάρη για διαφορετικούς όρους εάν εξετάζουμε τα βάρη που συμβάλλουν στο g(x) για ένα δεδομένο έγγραφο x για τις διαφορετικές παραμέτρους ενός ζευγαριού μεταβλητών xi, xj(i). Όταν xi = 1 και xj(i) = 0 τα βάρη που συμβάλλουν είναι

$$\log \frac{Prob\,(x_i = 1/(x_{j(i)} = 0) \wedge \omega_1)}{Prob\,(x_i = 1/(x_{j(i)} = 0) \wedge \omega_2)}$$

και ομοίως για τις άλλες τρεις παραμέτρους των xi και xj(i).

Αυτό δείχνει πόσο απλή είναι η μη γραμμική συνάρτηση στάθμισης.

Παραδείγματος χάριν, λαμβάνοντας υπόψη ένα έγγραφο στο οποίο το i εμφανίζεται αλλά το j(i) όχι, τότε το συμβαλλόμενο βάρος στο g(**x**) είναι βασισμένο στην αναλογία δύο πιθανοτήτων. Η πρώτη είναι η πιθανότητα της εμφάνισης του i στο σύνολο σχετικών εγγράφων δεδομένου ότι το j(i) δεν εμφανίζεται. Η δεύτερη είναι η ανάλογη πιθανότητα που υπολογίζεται στα μη-σχετικά έγγραφα. Βάσει αυτής της αναλογίας αποφασίζουμε πόσα στοιχεία υπάρχουν για την ανάθεση του x στα σχετικά ή μη σχετικά έγγραφα. Είναι σημαντικό να αναφερθεί σε αυτό το σημείο ότι τα στοιχεία για την ανάθεση είναι συνήθως βασισμένα σε μια εκτίμηση του ζευγαριού των πιθανοτήτων.

Εκτίμηση των παραμέτρων

Η χρήση μιας συνάρτησης στάθμισης του είδους που παράχθηκε προηγουμένως απαιτεί σε μια πραγματική ανάκτηση την εκτίμηση των σχετικών παραμέτρων. Εδώ θα εξεταστεί η εκτίμηση του ti και ri για τη μη γραμμική περίπτωση, προφανώς η γραμμική περίπτωση θα ακολουθήσει αναλογικά. Για να παρουσιάσουμε τι περιλαμβάνεται, θα δώσουμε ένα παράδειγμα της διαδικασίας εκτίμησης χρησιμοποιώντας τις απλές εκτιμήσεις μέγιστης πιθανότητας. Η βάση για τις εκτιμήσεις μας είναι ο ακόλουθος 2 επί 2 πίνακας.

	$x_i = 1$	$x_i = 0$	
$x_{j(i)} = 1$	[1]	[2]	[7]
$x_{j(i)} = 0$	[3]	[4]	[8]
	[5]	[6]	[9]

Εδώ έχει υιοθετηθεί ένα σχέδιο ονομασίας για τα κελιά στα οποία το [x] σημαίνει τον Αριθμό εμφανίσεων στο κελί που ονομάζεται x. Αγνοώντας προς το παρόν τη φύση του συνόλου στο οποίο αυτός ο πίνακας είναι βασισμένος, οι εκτιμήσεις μας είναι οι ακόλουθες:

$$P(x_i = 1 | x_{j(i)} = 1) = t_i \sim \frac{[1]}{[7]}$$
$$P(x_i = 1 | x_{j(i)} = 1) = r_i \sim \frac{[3]}{[8]}$$

Γενικά θα είχαμε δύο πίνακες αυτού του είδους κατά την δημιουργία της συνάρτησής μας g(**x**), έναν για τον υπολογισμό των παραμέτρων που συνδέονται με το P(**x** / w1) και έναν για το P (**x** / w2). Εάν υπολογίζαμε τις εκτιμήσεις για αυτήν την περιοριστική περίπτωση, αυτό θα ήταν χρήσιμο μόνο στην παρουσίαση του ανώτερου ορίου της ανάκτησή μας κάτω από αυτό το συγκεκριμένο πρότυπο. Πιο ρεαλιστικά, θα είχαμε ένα δείγμα εγγράφων, πιθανώς μικρό (όχι απαραίτητα τυχαίο), για το οποίο η κατάσταση σχετικότητας κάθε εγγράφου θα ήταν γνωστή. Αυτό το μικρό σύνολο θα

ήταν έπειτα η πηγή πληροφοριών για οποιονδήποτε 2 επί 2 πίνακα που θα θέλαμε να κατασκευάσουμε.
Οι εκτιμήσεις που παρουσιάστηκαν είναι παραδείγματα των εκτιμήσεων σημείου. Υπάρχουν διάφοροι τρόποι για έναν κατάλληλο κανόνα εκτίμησης σημείου. Δυστυχώς η καλύτερη μορφή κανόνα εκτίμησης είναι ακόμα ένα ανοικτό πρόβλημα. Στην πραγματικότητα, μερικοί στατιστικολόγοι θεωρούν ότι η εκτίμηση σημείου δεν πρέπει να χρησιμοποιηθεί καθόλου. Εντούτοις στα πλαίσια της ανάκτησης πληροφορίας είναι δύσκολο να δούμε πώς μπορούμε να αποφύγουμε τις εκτιμήσεις σημείου. Μια σημαντική αντίρρηση σε οποιοδήποτε κανόνα εκτίμησης σημείου είναι ότι στην παραγωγή του γίνονται μερικές "αυθαίρετες" υποθέσεις. Ευτυχώς στην ανάκτηση πληροφορίας υπάρχει κάποια πιθανότητα δικαιολόγησης αυτών των υποθέσεων με την υπόδειξη πειραματικών στοιχείων, που συγκεντρώνονται από τα συστήματα ανάκτησης, μειώνοντας έτσι την αυθαιρεσία.

Δύο βασικές υποθέσεις που γίνονται στην παραγωγή οποιουδήποτε κανόνα εκτίμησης μέσω της θεωρίας απόφασης Baye's, αφορούν τα επόμενα δύο μεγέθη:
(1) Τη μορφή της προγενέστερης κατανομής στο διάστημα της παραμέτρου, δηλ. στην περίπτωσή μας η υποτιθέμενη κατανομή πιθανότητας στις πιθανές τιμές της δυωνυμικής παραμέτρου και
(2) Τη μορφή της συνάρτησης απώλειας που χρησιμοποιείται για να μετρήσει το λάθος που γίνεται στον υπολογισμό της παραμέτρου.

Μόλις γίνουν ρητές αυτές οι δύο υποθέσεις, με τον καθορισμό της μορφής της κατανομής και της συνάρτησης απώλειας, τότε μαζί με την αρχή Baye's που επιδιώκει να ελαχιστοποιήσει τη μεταγενέστερη υπό όρους αναμενόμενη απώλεια δεδομένων των παρατηρήσεων, μπορούμε να παραγάγουμε διαφορετικούς κανόνες εκτίμησης. Η βιβλιογραφία της στατιστικής δεν δίνει πολλή βοήθεια στην επιλογή του κανόνα. Οι σημαντικοί κανόνες εκτίμησης μιας αναλογίας p είναι της μορφής

$$\hat{p} = \frac{x + a}{n + a + b}$$

όπου το x είναι ο αριθμός επιτυχιών στις n δοκιμές , και a και b είναι παράμετροι που υπαγορεύονται από τον ιδιαίτερο συνδυασμό κατανομής και συνάρτησης απώλειας.

Κατά συνέπεια έχουμε μια ολόκληρη κατηγορία κανόνων εκτίμησης. Παραδείγματος χάριν όταν a = b = 0 έχουμε τη συνήθη εκτίμηση x/n, και όταν a = b = 1/2 έχουμε έναν κανόνα, που αποδίδεται, από τον Good, στον Sir Harold Jeffreys. Αυτός ο τελευταίος κανόνας είναι στην πραγματικότητα ο κανόνας που χρησιμοποιείται από τους Robertson και Sparck Jones1 στις εκτιμήσεις τους. Κάθε τιμή που αποδίδεται στις παραμέτρους ανάκτησης μπορεί να δικαιολογηθεί από την άποψη της λογικής της προκύπτουσας προγενέστερης κατανομής. Αφού ότι θεωρείται λογικό από ένα άτομο δεν είναι απαραίτητα λογικό για κάποιον άλλο, η τελευταία επιλογή πρέπει να στηριχτεί στην απόδοση σε μια πειραματική δοκιμή. Ευτυχώς στην ανάκτηση πληροφορίας είμαστε σε μοναδική θέση να κάνουμε αυτό το είδος δοκιμής.
Ένας σημαντικός λόγος για την ύπαρξη κανόνων εκτίμησης διαφορετικών από το απλό x/n, είναι ότι αυτό είναι μάλλον μη ρεαλιστικό για τα μικρά δείγματα. Ας

εξετάσουμε την περίπτωση ενός δείγματος (n = 1) και το δοκιμαστικό αποτέλεσμα x = 0 (ή x = 1) που θα οδηγούσε στην εκτίμηση για το p ως p = 0 (ή p = 1). Αυτό είναι σαφώς γελοίο, δεδομένου ότι στις περισσότερες περιπτώσεις θα ξέραμε ήδη με υψηλή πιθανότητα ότι 0 < p < 1. Για να υπερνικήσουμε αυτήν την δυσκολία θα μπορούσαμε να δοκιμάσουμε να ενσωματώσουμε αυτήν την προγενέστερη γνώση σε μια κατανομή των πιθανών τιμών της παραμέτρου που προσπαθούμε να υπολογίσουμε. Μόλις δεχτούμε τη δυνατότητα πραγματοποίησής του και διευκρινίσουμε τον τρόπο με τον οποίο το λάθος εκτίμησης θα μετρηθεί, η αρχή Baye's (ή κάποια άλλη αρχή) θα οδηγήσει σε έναν κανόνα διαφορετικό από x/n.

7. Αξιολόγηση (Evaluation)

Εισαγωγή

Πολλές προσπάθειες και έρευνες έχουν γίνει για την επίλυση του προβλήματος της αξιολόγησης των συστημάτων ανάκτησης πληροφοριών. Εντούτοις, είναι δίκαιο να ειπωθεί ότι οι περισσότεροι άνθρωποι ενεργοί στον τομέα της αποθήκευσης πληροφοριών και της ανάκτησης θεωρούν ότι το πρόβλημα είναι ακόμα μακριά από την επίλυση . Θα προσπαθήσουμε να εξηγήσουμε την συμβατική, τη συνήθη μέθοδο αξιολόγησης, που ακολουθείται από μια έρευνα για τις πιο ελπιδοφόρες προσπάθειες

βελτίωσης των μεθόδων αξιολόγησης.
Για να θέσουμε το πρόβλημα της αξιολόγησης υποβάλλουμε τρεις ερωτήσεις: (1) Γιατί να αξιολογήσουμε; (2) Τι να αξιολογήσουμε; (3) Με ποιους τρόπους να αξιολογήσουμε ; Οι απαντήσεις σε αυτές τις ερωτήσεις καλύπτουν αρκετά καλά ολόκληρο τον τομέα της αξιολόγησης.
Η απάντηση στην πρώτη ερώτηση είναι κυρίως κοινωνική και οικονομική. Το κοινωνικό μέρος αφορά κυρίως την επιθυμία να τεθεί ένα μέτρο σχετικά με τα οφέλη (ή τα μειονεκτήματα) που αποκτώνται από τα συστήματα ανάκτησης πληροφοριών. Χρησιμοποιούμε τη λέξη "όφελος" υπό μια ευρύτερη έννοια . Παραδείγματος χάριν, πως θα επωφεληθούν οι χρήστες (ή ποια ζημιά θα γίνει) με την αντικατάσταση των παραδοσιακών πηγών πληροφοριών από ένα πλήρως αυτόματο και αλληλεπιδρών σύστημα ανάκτησης. Οι μελέτες για να μετρηθεί αυτό επιτυγχάνεται αλλά τα αποτελέσματα είναι δύσκολο να ερμηνευθούν. Για μερικά είδη συστημάτων ανάκτησης το όφελος μπορεί να μετρηθεί ευκολότερα απ' ό,τι για άλλα. Τα οικονομικά ποσά απάντησης σε μια δήλωση πόσο πρόκειται να κοστίσει συνδέεται με την ερώτηση: "είναι αυτή η αξία του;". Ακόμη και μια απλή δήλωση του κόστους είναι δύσκολο να παρθεί. Οι δαπάνες υπολογιστών μπορούν να είναι εύκολο να υπολογιστούν, αλλά οι δαπάνες από την άποψη της προσωπικής προσπάθειας είναι πολύ πιο δύσκολες να εξακριβωθούν. Κατόπιν εάν αξίζει ή όχι εξαρτάται από το μεμονωμένο χρήστη.
Πρέπει να είναι προφανές πως στην αξιολόγηση ενός συστήματος ανάκτησης πληροφοριών ενδιαφερόμαστε κυρίως για την παροχή των στοιχείων έτσι ώστε οι χρήστες να μπορούν να λάβουν μια απόφαση (1) εάν θέλουν ένα τέτοιο σύστημα και (2) εάν θα αξίζει. Επιπλέον, αυτές οι μέθοδοι αξιολόγησης χρησιμοποιούνται με έναν

συγκριτικό τρόπο να μετρήσουν εάν ορισμένες αλλαγές θα οδηγήσουν σε βελτίωση της απόδοσης. Με άλλα λόγια, όταν γίνεται μια αξίωση για μια ιδιαίτερη στρατηγική αναζήτησης, το κριτήριο της αξιολόγησης μπορεί να εφαρμοστεί για να καθορίσει εάν η αξίωση είναι έγκυρη.
Η δεύτερη ερώτηση (τι να αξιολογήσει;) απεικονίζει τη δυνατότητα του συστήματος για να ικανοποιηθεί ο χρήστης. Στην πραγματικότητα, από το 1966, ο Cleverdon έδωσε μια απάντηση σε αυτό.

Απαρίθμησε έξι κύριες μετρήσιμες ποσότητες: (μετρικές απόδοσης συστημάτων ανάκτησης)
(1) Η κάλυψη της συλλογής, δηλαδή ο βαθμός στον οποίο το σύστημα περιλαμβάνει το σχετικό θέμα
(2) η χρονική καθυστέρηση, δηλαδή το μέσο διάστημα μεταξύ του χρόνου που το αίτημα αναζήτησης υποβάλλεται και ο χρόνος που δίνεται μια απάντηση
(3) η μορφή παρουσίασης του αρχείου εξόδου
(4) η προσπάθεια που περιλαμβάνεται εκ μέρους του χρήστη στη λήψη των απαντήσεων στα αιτήματα αναζήτησής του
(5) η ανάκληση του συστήματος, δηλαδή το ποσοστό του σχετικού υλικού που ανακτάται πραγματικά σε απάντηση ενός αιτήματος αναζήτησης
(6) η ακρίβεια του συστήματος, δηλαδή το ποσοστό του ανακτημένου υλικού που είναι πραγματικά σχετικό.

Υποστηρίζεται ότι οι (1)-(4) αξιολογούνται εύκολα., η ανάκληση και η ακρίβεια είναι αυτές που προσπαθούν να υπολογίσουν την αποτελεσματικότητα του συστήματος ανάκτησης. Με άλλα λόγια είναι ένα μέτρο της δυνατότητας του συστήματος να ανακτηθούν τα σχετικά έγγραφα συγχρόνως συγκρατώντας το μη-σχετικό. Υποτίθεται ότι το αποτελεσματικότερο σύστημα θα θεωρηθεί αυτό που θα ικανοποιήσει περισσότερο το χρήστη. Επίσης η ακρίβεια και η ανάκληση είναι ικανοποιητικές για τη μέτρηση της αποτελεσματικότητας.
Έχει συζητηθεί πολύ στο παρελθόν αν η ακρίβεια και η ανάκληση είναι στην πραγματικότητα οι κατάλληλες ποσότητες που χρησιμοποιούνται ως μέτρα της αποτελεσματικότητας. Μια δημοφιλής εναλλακτική λύση είναι ανάκληση και του fallout (το ποσοστό των μη-σχετικών εγγράφων που ανακτώνται). Εντούτοις, όλες οι εναλλακτικές λύσεις απαιτούν ακόμα τον προσδιορισμό με κάποιο σχετικό τρόπο. Τα πλεονεκτήματα στην ακρίβεια και την ανάκληση είναι ότι:

(1) πρόκειται για το συνηθέστερα χρησιμοποιημένο ζευγάρι
(2) είναι αρκετά καλά κατανοητές ποσότητες.

Η τελική ερώτηση (πώς να αξιολογήσει;) έχει μια μεγάλη τεχνική απάντηση. Είναι ενδιαφέρον να σημειωθεί ότι η τεχνική της αποτελεσματικής ανάκτησης έχει επηρεαστεί κατά ένα μεγάλο μέρος από την ιδιαίτερη στρατηγική ανάκτησης που υιοθετούνται και τη μορφή του αρχείου εξόδου. Παραδείγματος χάριν, όταν το αρχείο εξόδου είναι μια ταξινόμηση των εγγράφων μια προφανής παράμετρος, είναι η θέση κατάταξης η οποία διατίθεται για τον έλεγχο. Η χρησιμοποίηση της θέσης κατάταξης ως οριακή συνθήκη, θα υπολογίζει μία τιμή για κάθε μία της οριακής συνθήκης. Τα

αποτελέσματα θα μπορούσαν έπειτα να συνοψιστούν υπό μορφή συνόλου σημείων που ενώθηκαν από μια ομαλή καμπύλη. Η πορεία κατά μήκος της καμπύλης θα είχε έπειτα την άμεση ερμηνεία της ποικίλης αποτελεσματικότητας των σημείων μαζί με τα σημεία της οριακής συνθήκης . Δυστυχώς, αυτή η μορφή αξιολόγησης δεν απαντά σε ερωτήσεις της μορφής: πόσα ερωτήματα δούλεψαν καλύτερα από το μέσο όρο και πόσα χειρότερα; Εντούτοις, θα πρέπει να αφιερώσουμε περισσότερο χρόνο στην εξήγηση αυτής της προσέγγισης στη μέτρηση της αποτελεσματικότητας, δεδομένου ότι είναι η πιο κοινή προσέγγιση και πρέπει να γίνει κατανοητή.

Πριν προχωρήσουμε στις τεχνικές λεπτομέρειες σχετικά με τη μέτρηση της αποτελεσματικότητας πρόκειται επίσης να εξετάσουμε περισσότερο την σχετική έννοια που κρύβεται .

Σχετικότητα (Relevance)

Η σχετικότητα είναι μια υποκειμενική έννοια. Οι διαφορετικοί χρήστες σε δοθέντες ερωτήσεις μπορούν να διαφέρουν ως προς τη σχετικότητα ή τη μη-σχετικότητα των ιδιαίτερων εγγράφων. Εντούτοις, σε πολλές συλλογές εγγράφων οι ερωτήσεις εξέτασης με τις αντίστοιχες αξιολογήσεις της σχετικότητας είναι διαθέσιμες. Αυτές οι ερωτήσεις αποσπούνται συνήθως από τους αξιόπιστους χρήστες, δηλαδή χρήστες με μια ιδιαίτερη κατάταξη που έχουν μια ανάγκη πληροφοριών. Οι αξιολογήσεις της σχετικότητας γίνονται από μια ομάδα ειδικών για την συγκεκριμένη κατάταξη. Έτσι έχουμε τώρα την κατάσταση όπου διάφορες ερωτήσεις υπάρχουν και για τις οποίες οι "σωστές" απαντήσεις είναι γνωστές. Είναι μια γενική υπόθεση στον τομέα της ανάκτησης πληροφορίας που πρέπει μια τιμή στρατηγικής ανάκτησης κάτω από έναν μεγάλο αριθμό πειραματικών όρων να αποδώσει καλά σε μια λειτουργική κατάσταση όπου η σχετικότητα δεν είναι γνωστή εκ των προτέρων.

Υπάρχει μια έννοια σχετική που μπορεί να ειπωθεί για να είναι αντικειμενική και που αξίζει την αναφορά ως ενδιαφέρουσα πηγή μελέτης . Καλείται "λογική σχετικότητα". Η χρησιμότητά της στα παρόντα συστήματα ανάκτησης είναι περιορισμένη. Εντούτοις, μπορεί να αποδειχθεί ότι είναι κάποιας σπουδαιότητας όσον αφορά την ανάπτυξη ερωταπαντήσεων (question-answering) των συστημάτων.
Η λογική σχετικότητα εξηγείται εύκολα εάν οι ερωτήσεις είναι περιορισμένες στο τύπο yes-no. Καθορίζεται από την άποψη της λογικής συνέπειας για να καταστήσει πιθανό μια ερώτηση που αντιπροσωπεύεται από ένα σύνολο προτάσεων. Στην περίπτωση μιας yes-no ερώτησης που αντιπροσωπεύεται από δύο επίσημες δηλώσεις του τύπου "p" και "not-p". Παραδείγματος χάριν, εάν η ερώτηση ήταν "το υδρογόνο είναι ένα στοιχείο αλογόνου;", το μέρος των δηλώσεων θα ήταν ισοδύναμα "το υδρογόνο είναι ένα στοιχείο αλογόνου" και "το υδρογόνο δεν είναι ένα στοιχείο αλογόνου".
Εάν οι δύο δηλώσεις που αντιπροσωπεύουν την ερώτηση καλούνται συστατικές δηλώσεις, τότε το υποσύνολο του συνόλου αποθηκευμένων προτάσεων είναι προϋπόθεση που τίθενται για μια συστατική δήλωση εάν και μόνο εάν η δήλωση είναι μια λογική συνέπεια εκείνου του υποσυνόλου.
Οι ελάχιστες προϋποθέσεις που τίθενται για μια συστατική δήλωση αφορούν στο ότι

εάν οποιαδήποτε από τα μέλη του διαγράφηκαν, η συστατική δήλωση δεν θα ήταν πλέον λογική συνέπεια του προκύπτοντος συνόλου.
Η λογική σχετικότητα ορίζεται τώρα ως μια σχέση δύο-θέσεων μεταξύ των αποθηκευμένων προτάσεων και την απαιτούμενη πληροφορία για αντιπροσώπευση (δηλαδή η ερώτηση που αντιπροσωπεύεται ως συστατικές δηλώσεις). Ο τελικός ορισμός είναι ο ακόλουθος: Μια αποθηκευμένη πρόταση είναι λογικά σχετική (μια αντιπροσώπευση) εάν και μόνο εάν μια ανάγκη πληροφοριών είναι μέλος κάποιου ελάχιστου συνόλου προϋποθέσεων, αποθηκευμένων προτάσεων, για κάποια συστατική δήλωση.
Αν και η λογική σχετικότητα αρχικά καθορίζεται μόνο μεταξύ των προτάσεων μπορεί εύκολα να επεκταθεί για να ισχύσει και για τα αποθηκευμένα έγγραφα. Ένα έγγραφο είναι σχετικό με μια ανάγκη πληροφοριών εάν και μόνο εάν περιέχει τουλάχιστον μια πρόταση που είναι σχετική με εκείνη την ανάγκη.
Νωρίτερα αναφέρθηκε ότι αυτή η έννοια ήταν σχετική και μόνο περιορισμένης χρήσης. Ο κύριος λόγος γι' αυτό είναι ότι το είδος συστήματος που θα απαιτούσε να εφαρμόσει μια στρατηγική ανάκτησης που θα ανακτούσε μόνο τα λογικά σχετικά έγγραφα δεν έχει δημιουργηθεί ακόμα. Εντούτοις, τα συστατικά ενός τέτοιου συστήματος υπάρχουν μέχρι ένα σημείο. Σε γενικές γραμμές αυτό το σύστημα θα μπορούσε να επεκταθεί για να κατασκευάσει έναν κόσμο εγγράφων, δηλαδή το περιεχόμενο ενός εγγράφου αναλύεται και ενσωματώνεται στον κόσμο των "κατανοητών" εγγράφων. Ενδεχομένως όμως η κλίμακα ενός συστήματος αυτού του είδους να είναι πάρα πολύ μεγάλη για τους παρόντες υπολογιστές .

Ακρίβεια και ανάκληση (position and recall)

Η σχετικότητα για άλλη μια φορά θα θεωρηθεί ότι έχει την ευρύτερη έννοιά της "καταλληλότητας", δηλαδή ένα έγγραφο καθορίζεται τελικά να είναι σχετικό ή όχι από το χρήστη. Η αποτελεσματικότητα είναι καθαρά ένα μέτρο της δυνατότητας του συστήματος να ικανοποιηθεί ο χρήστης από την άποψη της σχετικότητας των εγγράφων που ανακτώνται. Αρχικά, θα επικεντρωθούμε στη μέτρηση της αποτελεσματικότητας από την ακρίβεια και την ανάκληση. Μια παρόμοια ανάλυση θα μπορούσε να δοθεί για οποιοδήποτε ζευγάρι των ισοδύναμων μέτρων.

Είναι χρήσιμο σε αυτό το σημείο να εισάγούμε τον γνωστό πίνακα "πιθανότητας".

	ΣΧΕΤΙΚΟ	ΜΗ ΣΧΕΤΙΚΟ	
ΑΝΑΚΤΗΜΕΝΟ	$A \cap B$	$\bar{A} \cap B$	B
ΜΗ ΑΝΑΚΤΗΜΕΝΟ	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	N

N= ο αριθμός των εγγράφων στο σύστημα

Ένας μεγάλος αριθμός μέτρων της αποτελεσματικότητας μπορεί να προκύψει από αυτόν τον πίνακα. Στον κατάλογο υπάρχουν μερικοί, όπως:

$$\text{PRECISION} = \frac{A \cap B}{B}$$

$$\text{RECALL} = \frac{A \cap B}{A}$$

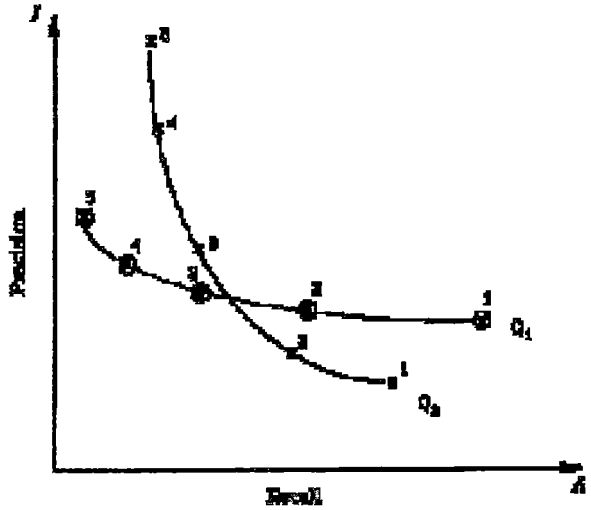
$$\text{FALLOUT} = \frac{\bar{A} \cap B}{\bar{A}}$$

(| | είναι το κριτήριο μέτρησης)

$$P = \frac{R \times G}{(R \times G) + F(1 - G)} \qquad \text{where } G = \frac{A}{N}$$

Για κάθε αίτημα που υποβάλλεται σε ένα σύστημα ανάκτησης ένας από αυτούς τους πίνακες μπορεί να κατασκευαστεί. Με βάση αυτούς τους πίνακες μπορεί να υπολογιστεί μια τιμή ακρίβειας-ανάκλησης. Εάν η έξοδος της στρατηγικής ανάκτησης εξαρτάται από μια παράμετρο, όπως η θέση κατάταξης ή του συντονισμού (ο αριθμός

όρων μιας ερώτησης που έχει κοινά με ένα έγγραφο), μπορεί να δώσει έναν διαφορετικό πίνακα για κάθε τιμή της παραμέτρου και ως εκ τούτου μια διαφορετική τιμή ακρίβειας-ανάκλησης. **Εάν λ είναι η παράμετρος, τότε Rλ δείχνει την ακρίβεια, Pλ την ανάκληση και μια τιμή ακρίβειας-ανάκλησης θα δοθεί από το διαταγμένο ζευγάρι (Rλ, Pλ).** Το σύνολο διαταγμένων ζευγαριών αποτελεί τη γραφική παράσταση ακρίβειας-ανάκλησης. Γεωμετρικά όταν τα σημεία ενωθούν αποτελούν την καμπύλη ακρίβειας-ανάκλησης, η οποία δίνει και την απόδοση κάθε αιτήματος. Για να μετρήσει τη γενική απόδοση ενός συστήματος, ένα για κάθε αίτημα, συνδυάζεται με κάποιο τρόπο για να παραχθεί μια μέση καμπύλη. Στο παρακάτω σχήμα δίνεται η γραφική παράσταση για δύο ερωτήματα, Q1 και Q2.



Σχήμα 27: Οι καμπύλες ακρίβειας-ανάκτησης για δύο ερωτήματα.

Υπολογισμός μέσου όρου των τεχνικών

Η μέθοδος του μέσου όρου των καμπυλών ακρίβειας-ανάκλησης φαίνεται να εξαρτάται κατά ένα μεγάλο μέρος από τη στρατηγική ανάκτησης που υιοθετείται. Υποτίθεται ότι η ανάκτηση γίνεται από το επίπεδο συντονισμού. Εάν s είναι το σύνολο αιτημάτων, τότε:

$$| \vec{A} | = \sum_{s \in S} | A_s |$$

όπου As είναι το σύνολο των σχετικών εγγράφων με το αίτημα s. Εάν λ είναι το επίπεδο συντονισμού, τότε:

$$| \vec{B} | = \sum_{s \in S} | B_{\lambda s} |$$

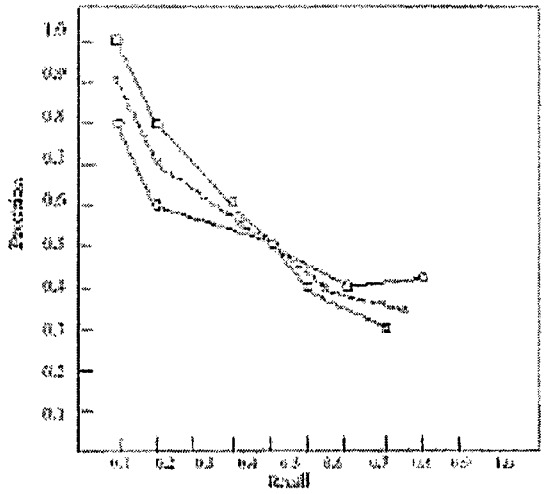
όπου Βλs είναι το σύνολο εγγράφων που ανακτώνται στο επίπεδο συντονισμού λ ή πιο πάνω. Τα σημεία (Rλ, Pλ) υπολογίζεται τώρα ως εξής:

$$R_\lambda = \sum_{s \in S} \frac{| A_s \cap B_{\lambda s}}{\vec{A}}$$

$$P_\lambda = \sum_{s \in S} \frac{A_s \cap B_{\lambda s}|}{\vec{A}|}$$

Το σχήμα 28 παρουσιάζει γραφικά τι συμβαίνει όταν δύο καμπύλες ακρίβειας-ανάκλησης συνδυάζονται κατ' αυτό τον τρόπο. Τα ακατέργαστα στοιχεία δίνονται στον πίνακα 27.

Μια εναλλακτική μέθοδος είναι η macro-evaluation. Η μέση καμπύλη λαμβάνεται με τη διευκρίνιση ενός συνόλου τυποποιημένων τιμών ανάκλησης για τις οποίες οι μέσες τιμές ακρίβειας υπολογίζονται με τον υπολογισμό μέσου όρου σε όλες τις ερωτήσεις των μεμονωμένων τιμών ακρίβειας που αντιστοιχούν στις τυποποιημένες υιμές ανάκλησης. Συχνά καμία μοναδική τιμή ακρίβειας δεν αντιστοιχεί ακριβώς έτσι .



Σχήμα 28: Ένα παράδειγμα υπολογισμού μέσου όρου εκτίμησης ακρίβειας-ανάκλησης.

Έστω I μια πρότυπη πληροφοριακή ανάγκη (σε μια συλλογή κειμένων αναφοράς) και R το σύνολο των σχετικών της κειμένων. Έστω επίσης |R| ο αριθμός των κειμένων στο σύνολο Μία δοσμένη στρατηγική ανάκτησης επεξεργάζεται την πληροφοριακή ανάγκη I και παράγει ένα σύνολο κειμένων απάντησης A. Έστω |A| ο αριθμός των κειμένων στο σύνολο A και έστω |Ra| ο αριθμός των κειμένων που είναι κοινά στα σύνολα R και A. Οι μετρικές ανάκληση και ακρίβεια θα οριστούν ως εξής :

Ανάκληση είναι το ποσοστό των σχετικών κειμένων στο σύνολο R που έχουν ανακτηθεί
Ανάκληση=| Ra| / |R |

Ακρίβεια είναι το ποσοστό των ανακτηθέντων κειμένων στο σύνολο A που είναι σχετικό
Ακρίβεια=| Ra| / |A |

Η ακρίβεια και η ανάκληση υποθέτουν ότι όλα τα κείμενα στο σύνολο απάντησης A έχουν εξεταστεί από το 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 διαταγμένα με βάση το βαθμό σχετικότητας με την πληροφοριακή ανάγκη I. Στην περίπτωση αυτή οι μετρικές αναλύσεις και ακρίβειας μεταβάλλονται καθώς ο χρήστης εξετάζει τα διάφορα κείμενα της ανακτώμενης συλλογής, από τα περισσότερα σχετικά προς τα λιγότερα σχετικά. Συνεπώς πλήρης εκτίμηση απόδοσης απαιτεί την σχεδίαση ενός διαγράμματος

Ακρίβειας /ανάκλησης .
Ας θεωρήσουμε μία συλλογή κειμένων αναφοράς, ένα σύνολο προτύπων αναγκών, ένα ερώτημα q το οποίο ανήκει στη συλλογή των προτύπων αναγκών και έστω Rq το σύνολο των σχετικών κειμένων για το ερώτημα q όπως έχει καθοριστεί από ειδικούς. Για παράδειγμα, υποθέτουμε ότι το σύνολο Rq περιέχει τα ακόλουθα κείμενα
Rq={d1,d3,d5,d7,d9,d13,d21,d41,d43,d45}
Υποθέτουμε ότι ο αλγόριθμος επιστρέφει την ακόλουθη συλλογή κειμένων, όπου η διάταξη σχετικότητας δηλώνεται από τους αριθμούς δίπλα σε κάθε κείμενο ενώ με έντονη σκίαση παρουσιάζονται τα κείμενα που ανήκουν στο σύνολο Rq.

1. **d7**	6. **d5**	11. d4
2. d2	7. d28	12. d40
3. **d3**	8. d12	13. d10
4. d6	9. d22	14. d36
5. d8	10. **d13**	15. **d1**

Αρχικά το κείμενο d7 που βρίσκεται στη θέση 1 είναι σχετικό και αντιστοιχεί στο 10% του συνόλου των σχετικών κειμένων .Συνεπώς λέμε ότι έχουμε ακρίβεια 100% και ανάκληση 10%. Στη συνέχεια το κείμενο που βρίσκεται στη θέση 3 είναι το επόμενο σχετικό κείμενο. Στο σημείο αυτό έχουμε ακρίβεια περίπου 66%(3 στα 3 κείμενα είναι σχετικά) και ανάκληση 20% (2 στα 10 σχετικά κείμενα έχουν ειδωθεί). Συνεχίζοντας με αυτό το τρόπο παίρνουμε ένα σύνολο ζευγαριών (τιμή ακρίβειας/ τιμή ανάκλησης) που μπορούμε να το παραστήσουμε σε ένα διάγραμμα το καλούμενο διάγραμμα ακρίβειας/ ανάκλησης.

Πίνακας 4 : Τα ακατέργαστα στοιχεία για micro-evaluation του σχήματος 28

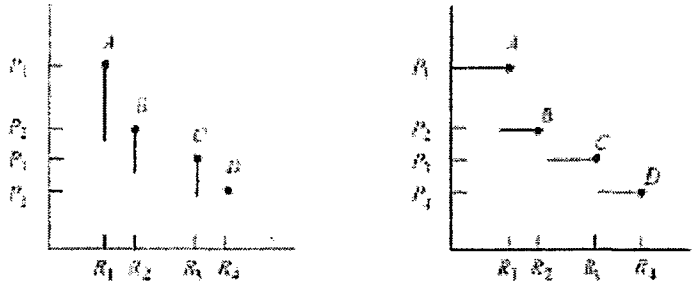
QUERY 1:	R	0.1	0.2	0.4	0.6	0.8	A_1 = 100
	P	1.0	0.8	0.6	0.2	0.3	

QUERY 2:	R	0.1	0.3	0.5	0.7	0.9	A_2 = 90
	P	0.8	0.6	0.5	0.4	0.1	

λ	$\|B_{2\lambda}\|$	$\|A_\lambda \cap B_{2\lambda}\|$	$\|B_{1\lambda}\|$	$\|A_\lambda \cap B_{1\lambda}\|$	R_λ	P_λ
1	10	10	10	8	0.1	0.8
2	25	20	30	24	0.24	0.68
3	66	40	80	40	0.40	0.55
4	150	60	140	56	0.62	0.40
5	266	80	190	72	0.80	0.34

Παρεμβολή

Πολλές τεχνικές παρεμβολής έχουν προταθεί . Βλέπε για παράδειγμα, Keen72



Σχήμα 29: Στο δεξιό σχήμα δείχνει τα σημεία παρεμβολής των σημείων του αριστερού σχήματος

Το σχήμα 29 παρουσιάζει χαρακτηριστική γραφική παράσταση P-R για μια ενιαία ερώτηση. Τα σημεία A, B, C και D καλούνται ως τα παρατηρηθέντα σημεία, δεδομένου ότι αυτά είναι τα μόνα σημεία που παρατηρούνται άμεσα κατά τη διάρκεια ενός πειράματος και άλλα μπορούν να προκύψουν από αυτά. Κατά συνέπεια δεδομένου ότι A = (R1, P1) έχει παρατηρηθεί, κατόπιν το επόμενο σημείο B είναι αυτό που αντιστοιχεί σε μια αύξηση στην ανάκληση, η οποία προκύπτει από μια αύξηση μονάδων στον αριθμό σχετικών εγγράφων που ανακτώνται. Μεταξύ δύο οποιωνδήποτε παρατηρηθέντων σημείων η ανάκληση παραμένει σταθερή, δεδομένου ότι τα άλλα μη-σχετικά έγγραφα ανακτώνται.

Είναι ένα πειραματικό γεγονός ότι οι μέσες γραφικές παραστάσεις ακρίβειας-ανάκλησης μειώνονται. Σύμφωνα με αυτό, μια γραμμική παρεμβολή υπολογίζει την καλύτερη δυνατή απόδοση μεταξύ οποιωνδήποτε δύο παρατηρηθέντων σημείων. Για να αποφύγουμε τα πειραματικά αποτελέσματα είναι πιθανώς καλύτερο να εκτελεσθεί μια πιο συντηρητική παρεμβολή ως εξής:

Έστω (Rλ, Pλ) είναι το σύνολο τιμών ακρίβειας-ανάκλησης που λαμβάνεται με κάποιας παραμέτρου λ. Για να λάβουμε το σύνολο παρατηρηθέντων σημείων διευκρινίζουμε ένα υποσύνολο των παραμέτρων λ. Κατά συνέπεια (Rθ, Pθ) είναι ένα παρατηρηθέν σημείο εάν θ αντιστοιχεί σε μια αξία λ στο οποίο μια αύξηση στην ανάκληση παράγεται. Έχουμε τώρα:

$$Gs = (R\theta s, P\theta s)$$

το σύνολο παρατηρηθέντων σημείων για ένα αίτημα. Για να παρεμβάλει μεταξύ οποιωνδήποτε δύο σημείων καθορίζουμε:

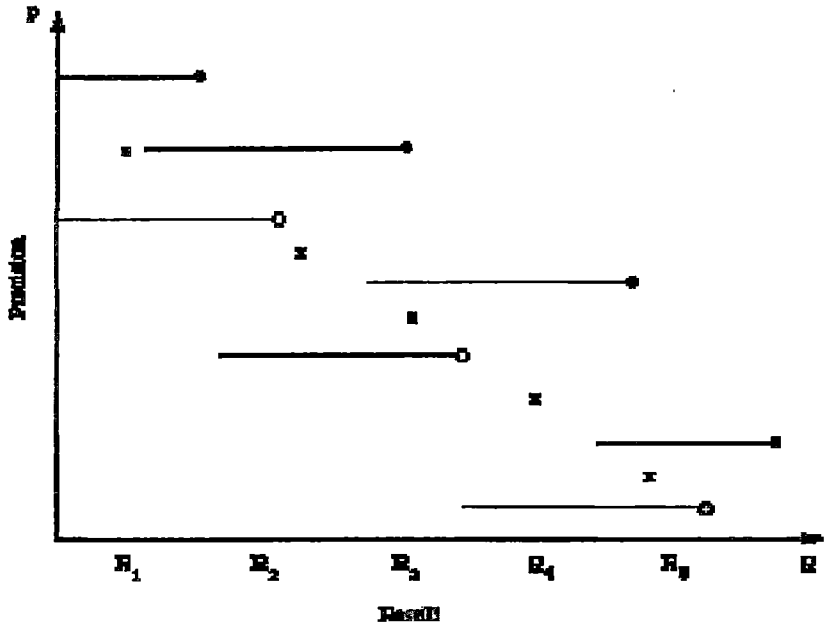
$$Ps(R) = \{\sup P : R' \geq R \text{ s.t. } (R', P) \in Gs\}$$

όπου το R είναι μια τυποποιημένη αξία ανάκλησης. Από αυτό λαμβάνουμε τη μέση

αξία ακρίβειας στην τυποποιημένη αξία R ανάκλησης από:

$$P(R) = \sum_{r \in S} \frac{P_r(R)}{|S|}$$

Το σύνολο παρατηρηθέντων σημείων είναι τέτοιο που η λειτουργία μειώνεται. Το σχήμα 30 παρουσιάζει την επίδραση της διαδικασίας παρεμβολής, ουσιαστικά μετατρέπει την καμπύλη ακρίβειας-ανάκλησης σε συνάρτηση σταθερού βήματος. Μια απαραίτητη συνέπεια της μονοτονίας της είναι ότι η μέση καμπύλη ακρίβειας-ανάκλησης θα μειώνεται μονοτικά. Είναι δυνατό να καθοριστεί το σύνολο παρατηρηθέντων σημείων κατά τέτοιο τρόπο ώστε η λειτουργία παρεμβολής να μειωθεί μονοτονικά.
Στο σχήμα 30 επεξηγούμε την παρεμβολή και υπολογισμός μέσου όρου της διαδικασίας.



Σχήμα 30: Ένα παράδειγμα macro-evaluation. Τα σημεία x στη γραφική παράσταση βρίσκονται ενδιάμεσα δύο οριζόντιων γραμμών και οι τετμημένες τους δίνονται από τις πρότυπες τιμές ανάκτησης Ri.

Σύνθετα μέτρα (composite measures)

Υπήρχε δυσαρέσκεια στο παρελθόν με τις μεθόδους αποτελεσματικότητας από ένα ζευγάρι αριθμών (π.χ. ακρίβεια και ανάκληση) που μπορεί με έναν αόριστο τρόπο να οδηγήσει σε προσπάθειες ώστε να δημιουργηθούν τα σύνθετα μέτρα. Αυτοί είναι ακόμα βασισμένοι στον πίνακα "πιθανότητας" αλλά συνδυάζουν τα μέρη σε ένα ενιαίο μέτρο αριθμού. Δυστυχώς πολλά από αυτά τα μέτρα είναι μάλλον ειδικά και δεν μπορούν να δικαιολογηθούν με οποιοδήποτε λογικό τρόπο. **Το απλούστερο**

παράδειγμα αυτού του είδους μέτρου είναι το ποσό της ακρίβειας και της ανάκλησης σαν άθροισμα:

$$S = P + R$$

Αυτό συσχετίζεται απλά με ένα μέτρο που προτείνεται από τον Borko

$$BK = P + R - 1$$

Οι πιο περίπλοκοι είναι οι

$$Q = \frac{R - F}{R + F - 2RF} \qquad (F = \text{Fallout})$$

$$E = 1 - \frac{1}{2\left(\frac{1}{P}\right) + 2\left(\frac{1}{R}\right) - 3}$$

Το μέτρο V του Vickery μπορεί να αποδειχθεί για να είναι μια ειδική περίπτωση ενός γενικού μέτρου που θα προκύψει κατωτέρω.
Μερικά μέτρα έχουν αρχεία εξόδου που μπορούν να δικαιολογηθούν κατά τρόπο λογικό.

Το πρότυπο Swets

Από το 1963 ο Swets73 εξέφρασε τη δυσαρέσκεια για τις υπάρχουσες μεθόδους στην αποτελεσματικότητα ανάκτησης. Το υπόβαθρό του στην ανίχνευση σημάτων τον οδήγησε στη διατύπωση ενός πρότυπου αξιολόγησης βασισμένο στη στατιστική θεωρία απόφασης. Το 1967 αξιολόγησε περίπου πενήντα διαφορετικές μεθόδους ανάκτησης έχοντας υπόψη του το πρότυπο74. Τα αποτελέσματα της αξιολόγησής του ήταν ενθαρρυντικά αλλά μη αποφασιστικά. Στη συνέχεια, ο Brookes75 πρότεινε μερικές λογικές τροποποιήσεις στο μέτρο Swets όσον αφορά την αποτελεσματικότητα και ο Robertson76 έδειξε ότι οι προτεινόμενες τροποποιήσεις στην πραγματικότητα απλά αφορούσαν ένα εναλλακτικό μέτρο που προτάθηκε ήδη από Swets. Ο Bookstein έχει επανεξετάσει πρόσφατα αυτό το πρότυπο που επιδεικνύει πώς ο Swets στηρίχθηκε σιωπηρά σε μια υπόθεση "ίσης διαφοράς".
Είναι ενδιαφέρον ότι αν και το πρότυπο Swets είναι θεωρητικά ελκυστικό και οι συνδέσεις στις μετρήσεις IR ανέπτυξαν τη στατιστική θεωρία, αυτό δεν έχει βρει τη γενική αποδοχή σε ευρεία αγορά.

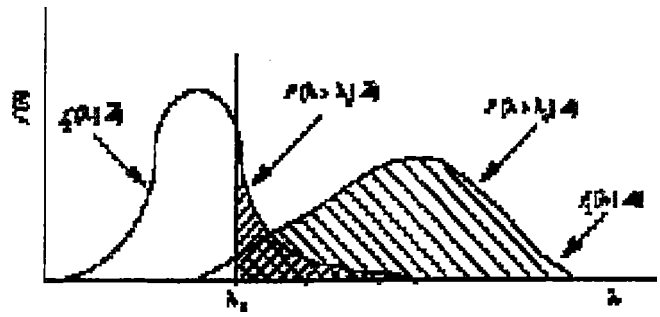
Πριν προχωρήσουμε σε μια εξήγηση του προτύπου Swets, πρόκειται επίσης να αναφέρουμε πλήρως τους όρους ότι το επιθυμητό μέτρο της αποτελεσματικότητας έχει σκοπό, να υπολογιστεί. Στην αρχή των εκθέσεων του ο Swets δήλωσε το1967 ότι:

Ένα επιθυμητό μέτρο της απόδοσης ανάκτησης θα είχε τις ακόλουθες ιδιότητες: Κατ' αρχάς, θα εκφράζει τη δυνατότητα ενός συστήματος ανάκτησης που διακρίνει μεταξύ των επιθυμητών και ανεπιθύμητων στοιχείων - δηλαδή θα είναι ένα μέτρο "της

αποτελεσματικότητας" και μόνο, ξεχωρίζοντας τους χωριστούς παράγοντες εκτίμησης σχετικά με το κόστος ή "την αποδοτικότητα". Δεύτερον, το επιθυμητό μέτρο δεν θα συγχεόταν από τη σχετική προθυμία του συστήματος να εξορύξει τα στοιχεία, θα εξέφραζε την διακριτική δύναμη ανεξάρτητα από οποιοδήποτε υιοθετούμενο "κριτήριο αποδοχής", ανεξάρτητα από το εάν το κριτήριο είναι χαρακτηριστικό του συστήματος ή ρυθμίζεται από το χρήστη. Τρίτον, το μέτρο θα ήταν ένας ενιαίος αριθμός , παραδείγματος χάριν, μια καμπύλη που αντιπροσωπεύει έναν πίνακα διάφορων ζευγαριών αριθμών - έτσι ώστε να μπορεί να δοθεί απλά και κατανοητά. Και τέταρτον το μέτρο θα επιτρέπει την πλήρη διάταξη των διαφορετικών αποδόσεων και θα αξιολογεί την απόδοση οποιουδήποτε συστήματος στους απόλυτους όρους - δηλαδή το μέτρο θα ήταν μια κλίμακα με μια μονάδα, ένα πραγματικό μηδέν και μια μέγιστη αξία. Λαμβάνοντας υπόψη ένα μέτρο με αυτές τις ιδιότητες, θα μπορούσαμε να είμαστε βέβαιοι της κατοχής ενός ευδιάκριτου και έγκυρου δείκτη για το πόσο καλό είναι ένα σύστημα ανάκτησης (ή μέθοδος) και εκτελεί λειτουργίες που έχουν ως σκοπό πρώτιστα να ολοκληρώσει και δεύτερον να μπορεί εύλογα να απαντάει σε θέματα της μορφής "θα πληρώσουμε X δολάρια για τις Y μονάδες της αποτελεσματικότητας? "
Ο Swets καθορίζει τις βασικές ιδιότητες, ακρίβειας, ανάκλησης και Fallout μεταβλητών στους πιθανολογικούς όρους.

Ανάκληση είναι το ποσοστό των σχετικών κειμένων στο σύνολο R που έχουν ανακτηθεί
Ακρίβεια είναι το ποσοστό των ανακτηθέντων κειμένων στο σύνολο A που είναι σχετικό
Fallout = το ποσοστό των μη-σχετικών εγγράφων που ανακτώνται.

Δέχεται ότι η μέτρηση αποτελεσματικότητας της ανάκτησης μπορεί να παραχθεί είτε από μια ακρίβεια-ανάκληση, είτε από μια ανάκληση-fallout καμπυλών η οποία παράγεται από παραλλαγή κάποιας μεταβλητής ελέγχου λ (π.χ. επίπεδο συντονισμού). Επιδιώκει να χαρακτηρίσει κάθε καμπύλη από έναν ενιαίο αριθμό. Απορρίπτει την ακρίβεια-ανάκληση και δέχεται την ανάκληση-fallout δεδομένου ότι το πρώτο δεν είναι ικανό αλλά αποτυγχάνει να παρουσιάσει την επιτυχία του δεύτερου.
Στην απλούστερη περίπτωση υποθέτει ότι η μεταβλητή λ διανέμεται κανονικά στο σύνολο σχετικών και μη εγγράφων. Οι δύο κατανομές δίνονται αντίστοιχα από τα N(μ1,σ1) και N(μ2,σ2). Οι συναρτήσεις πυκνότητας δίνονται από τα f1 (λ|A) και f2 (λ| A). Μπορούμε να απεικονίσουμε τη κατανομή όπως φαίνεται στο σχήμα 31.



Σχήμα 31: Δύο κανονικές κατανομές για το λ, η μία N(μ1,σ1), στην ομάδα σχετικών εγγράφων Α με πυκνότητα f1 (λ1|A) και η άλλη N(μ2,σ2). Στην ομάδα μη σχετικών εγγράφων Α με πυκνότητα f2 (λ| A). Το μέγεθος των γραμμοσκιασμένων περιοχών σε Β-Δ και Β-Α κατεύθυνση, αναπαριστούν την ανάκτηση και το fallout αντίστοιχα.

Η συνηθισμένη οργάνωση στην ανάκτηση πληροφοριών είναι τώρα να καθοριστεί ένας κανόνας απόφασης από την άποψη λ, για να καθορίσει ποια έγγραφα ανακτώνται (το κριτήριο αποδοχής). Με άλλα λόγια διευκρινίζουμε το λc έτσι ώστε ένα έγγραφο το οποίο συνδέεται με το λ και υπερβαίνει το λc, να ανακτάται. Μετράμε τώρα την αποτελεσματικότητα μιας στρατηγικής ανάκτησης με τη μέτρηση μερικών κατάλληλων μεταβλητών (όπως το R και το P, ή το R και το F) στις διάφορες τιμές λc. Προκύπτει ότι οι διαφορετικά σκιασμένες περιοχές κάτω από τις καμπύλες στο σχήμα 31 αντιστοιχούν στην ανάκληση και στο fallout.

Το πρότυπο Robertson - ο λογιστικός μετασχηματισμός

Ο Robertson σε συνεργασία με τον Teather έχουν αναπτύξει ένα πρότυπο για τον υπολογισμό των πιθανοτήτων που αντιστοιχούν στην ανάκληση και στο fallout77. Η διαδικασία εκτίμησης είναι ασυνήθιστη στην παραγωγή μιας εκτίμησης αυτών των πιθανοτήτων για μια ενιαία ερώτηση που λαμβάνει υπόψη δύο πράγματα: Πρώτο, το ποσό δεδομένων που χρησιμοποιείται για να υπολογιστούν οι εκτιμήσεις και δύο μέσους όρους των εκτιμήσεων σε όλες τις ερωτήσεις.

Χρησιμοποιώντας το λογαριθμικό μετασχηματισμό για τις πιθανότητες, ο οποίος είναι

$$\text{logit}\,\theta = \log\frac{\theta}{1-\theta} \qquad 0 < \theta < 1$$

το βασικό ποσοτικό πρότυπο για μια ενιαία ερώτηση j που προτείνεται είναι

$$\text{logit}\,\theta_{j1} = \alpha_j + \Delta_j$$
$$\text{logit}\,\theta_{j2} = \alpha_j - \Delta_j$$

Εδώ θj1 και θj2 είναι πιθανότητες που αντιστοιχούν στην ανάκληση και στο fallout αντίστοιχα όπως καθορίζονται στο προηγούμενο τμήμα. Οι παράμετροι αj και .j πρόκειται να ερμηνευθούν ως εξής:

αj : μετρά την ιδιομορφία της διατύπωσης ερώτησης
Δj : μετρά το χωρισμό από τα σχετικά και μη έγγραφα.

Για μια δεδομένη ερώτηση j εάν η ερώτηση i έχει διατυπωθεί με έναν πιο συγκεκριμένο τρόπο από το j, κάποιος θα ανέμενε την ανάκληση και το fallout να μειωθεί, δηλ.

$$\theta i1 < \theta j1 \text{ και } \theta i2 < \theta j2$$

Επίσης, εάν για την ερώτηση i το σύστημα είναι καλύτερο στο χωρισμό του μη-σχετικού από τα σχετικά έγγραφα για την ερώτηση j, κάποιος θα ανέμενε την ανάκληση να αυξηθεί και το fallout να μειωθεί, δηλ.

$$\theta i1 > \theta j1 \text{ και } \theta i2 < \theta j2$$

Η μέθοδος προσέγγισης υπολογισμού του αj και Δj παρουσιάζει μια τεχνική δυσκολία, απαιτούνται ορισμένες υποθέσεις για τα αj και Δj, η σημαντικότερη είναι το αj και το Δj να είναι ανεξάρτητα και να κατανέμονται κανονικά. Αυτές οι υποθέσεις είναι μάλλον δύσκολο να επικυρωθούν. Τα μόνα στοιχεία που προσκομίζονται μέχρι τώρα παράγουν τη κατανομή αj για ορισμένα στοιχεία δοκιμής. Στην περίπτωση Δj αυτό έχει διαπιστωθεί ότι είναι περίπου σταθερό στις ερωτήσεις έτσι ώστε ένα κοινό πρότυπο Δ δικαιολογείται:

$$\text{logit } \theta j1 = \alpha j1 + \Delta$$
$$\text{logit } \theta j2 = \alpha j2 - \Delta$$

Από ότι φαίνεται το Δ μπόρεσε να είναι υποψήφιο για ένα ενιαίο μέτρο αριθμού της αποτελεσματικότητας. Αυτές οι παράμετροι συσχετίζονται με την συμπεριφορά ενός συστήματος ανάκτησης πληροφοριών έτσι ώστε εάν πρόκειται να ορίσουμε το Δ ως μέτρο της αποτελεσματικότητας μπορούμε να απομακρύνουμε τις ζωτικής σημασίας πληροφορίες που απαιτήθηκαν για να κάνουν μια επέκταση και στην απόδοση άλλων συστημάτων.

Το πρότυπο Cooper – αναμενόμενο μήκος αναζήτησης

Το 1968, ο Cooper78 δήλωσε: "Η αρχική λειτουργία ενός συστήματος ανάκτησης ήταν να εστιάσει την προσοχή στην εξυπηρέτηση των χρηστών, σε μια τόσο μεγάλη έκταση εγγράφων και η εργασία της μελέτης και της απόρριψης των μη-σχετικών εγγράφων στην αναζήτησή των σχετικών". Είναι αυτή η "αποταμίευση" που μετριέται και θεωρείται ότι είναι ο ενιαίος δείκτης της αξίας για τα συστήματα ανάκτησης. Γενικά ο δείκτης ισχύει στα συστήματα ανάκτησης με τη διαταγμένη (ή ταξινομημένη) παραγωγή. Μετρά κατά προσέγγιση την προσπάθεια αναζήτησης που κάποιος θα ανέμενε να σώσει με τη χρησιμοποίηση του συστήματος ανάκτησης σε αντιδιαστολή να ψάχνει τη συλλογή τυχαία. Μια προσπάθεια γίνεται να λάβει υπόψη την ποικίλη δυσκολία για τα σχετικά έγγραφα στις διαφορετικές ερωτήσεις. Ο δείκτης υπολογίζεται για μια ερώτηση ενός ακριβώς διευκρινισμένου τύπου. Υποτίθεται ότι οι χρήστες είναι σε θέση να ποσοτικοποιήσουν την ανάγκη πληροφοριών τους σύμφωνα με έναν από τους ακόλουθους τύπους:

(α) μόνο ένα σχετικό έγγραφο είναι επιθυμητό

(β) κάποιος αυθαίρετος αριθμός ν είναι επιθυμητός

(γ) όλα τα σχετικά έγγραφα είναι επιθυμητά

(δ) ένα δεδομένο ποσοστό των σχετικών εγγράφων είναι επιθυμητό, κ.λπ.

Κατά συνέπεια, ο δείκτης είναι ένα μέτρο της απόδοσης για μια ερώτηση του δεδομένου τύπου.

Η παραγωγή μιας στρατηγικής αναζήτησης υποτίθεται ότι ήταν μια αδύνατη διάταξη των εγγράφων. Αρχικά εξετάζουμε μια ειδική περίπτωση, δηλαδή μια απλή διάταξη, η οποία είναι μια αδύνατη διάταξη έτσι ώστε για οποιαδήποτε δύο ευδιάκριτα στοιχεία e1 και e2 δεν υπάρχει περίπτωση το e1 R e2 και e2 R e1 (όπου το R είναι η σχέση διάταξης). Αυτό απλά σημαίνει ότι όλα τα έγγραφα στην έξοδο διατάσσονται γραμμικά χωρίς να υπάρχουν δύο ή περισσότερα έγγραφα στο ίδιο επίπεδο της διάταξης. Το μήκος αναζήτησης ορίζεται τώρα ως ο αριθμός μη-σχετικών εγγράφων που ένας χρήστης πρέπει να ανιχνεύσει προτού να ικανοποιηθεί η ανάγκη πληροφορίας (από την άποψη του προσδιορισμού της ποσότητας τύπων). Παραδείγματος χάριν, εξετάζουμε μια ταξινόμηση 20 εγγράφων στα οποία τα σχετικά κατανέμονται όπως στο σχήμα 32. Ένας τύπος 2 ερώτηση με n= 2 θα πρέπει να έχει το μήκος 2 αναζήτησης, με n= 6 αυτό θα είχαν το μήκος 3 αναζήτησης.

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Relevance	N	Y	N	Y	Y	Y	Y	N	Y	N	N	N	N	N	Y	N	N	N	N	N

Σχήμα 32: Απλή διάταξη για 20 ανακτημένα έγγραφα όπου η σειρά διάταξης είναι μοναδική. Τα Y αναπαριστούν τα σχετικά έγγραφα και τα N τα μη-σχετικά.

Δυστυχώς η ταξινόμηση που παράγεται από μια συνάρτηση ταυτοποίησης είναι σπάνια μια απλή διάταξη. Αυτό σημαίνει ότι σε οποιοδήποτε δεδομένο επίπεδο στην ταξινόμηση, υπάρχει τουλάχιστον ένα έγγραφο (πιθανώς περισσότερα) που καθιστά το μήκος αναζήτησης ακατάλληλο, δεδομένου ότι η διάταξη των εγγράφων μέσα σε ένα επίπεδο είναι τυχαία. Εάν η ανάγκη πληροφορίας ικανοποιείται σε ορισμένο επίπεδο στη διάταξη τότε ανάλογα με τη ρύθμιση των σχετικών εγγράφων μέσα στο επίπεδο θα πάρουμε διαφορετικά μήκη αναζήτησης. Εντούτοις μπορούμε να χρησιμοποιήσουμε μια ανάλογη ποσότητα που είναι το αναμενόμενο μήκος αναζήτησης. Για αυτό πρέπει να υπολογίσουμε την πιθανότητα κάθε πιθανού μήκους αναζήτησης με δοκιμές των σχετικών και μη-σχετικών εγγράφων στο επίπεδο, στο οποίο η ανάγκη των χρηστών ικανοποιείται.

Rank	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	4	4
Relevance	N	N	Y	Y	N	Y	Y	Y	N	Y	Y	N	N	N	N	N	N	N	Y	N

Σχήμα 33: Ένα παράδειγμα αδύναμης διάταξης 20 ανακτημένων εγγράφων με ισχυρή σειρά διάταξης.

Παραδείγματος χάριν, εξετάζουμε την αδύναμη διάταξη στο σχήμα 7.8. Εάν η ερώτηση είναι τύπου 2 με n= 6 έπειτα η ανάγκη ικανοποιείται σε επίπεδο 3. Τα πιθανά μήκη αναζήτησης είναι 3 ..4 ..5 ή 6 ανάλογα με πόσα μη-σχετικά έγγραφα προηγούνται του έκτου σχετικού εγγράφου. Μπορούμε να αγνοήσουμε τις πιθανές ρυθμίσεις μέσα στα επίπεδα 1 και 2 όπου οι συνεισφορές τους είναι πάντα οι ίδιες. Για να υπολογίσουμε το αναμενόμενο μήκος αναζήτησης χρειαζόμαστε την πιθανότητα κάθε μήκους αναζήτησης. Εξετάσουμε πρώτα τον αριθμό διαφορετικών τρόπων με τους οποίους δύο σχετικά έγγραφα θα μπορούσαν να κατανεμηθούν μεταξύ πέντε και είναι (5!/2!*(5-2)!) = 10. Από αυτά τα 4 θα οδηγούσε σε ένα μήκος αναζήτησης 3,3 σε ένα μήκος αναζήτησης 4,2 σε ένα μήκος αναζήτησης 5 και 1 σε ένα μήκος αναζήτησης 6. Οι αντίστοιχες πιθανότητές τους είναι επομένως, 4/10, 3/10, 2/10 και 1/10. Το αναμενόμενο μήκος αναζήτησης είναι τώρα:

(4/10) * 3 + (3/10) * 4 + (2/10) * 5 + (1/10) * 6 = 4

Η ανωτέρω διαδικασία οδηγεί αμέσως σε μια κατάλληλη δημιουργία τύπου για το αναμενόμενο μήκος αναζήτησης. Φαίνεται εύλογο ότι οι μέσοι όροι αποτελεσμάτων πολλών τυχαίων αναζητήσεων, μέσω του τελικού επιπέδου (επίπεδο στο οποίο η ανάγκη ικανοποιείται) θα είναι τα ίδια όπως για μια ενιαία αναζήτηση των σχετικών έγγραφων που χωρίζονται κατά διαστήματα "ομοιόμορφα" σε όλο εκείνο το επίπεδο.

Πρώτα απαριθμούμε τις μεταβλητές:

(α) το q είναι η ερώτηση του δεδομένου τύπου

(β) το j είναι ο συνολικός αριθμός των μη-σχετικών εγγράφων στο q σε όλα τα επίπεδα που προηγούνται.

(γ) το r είναι ο αριθμός των σχετικών εγγράφων στο τελικό επίπεδο

(δ) το i είναι ο αριθμός των μη-σχετικών εγγράφων στο τελικό επίπεδο

(ε) το s είναι ο αριθμός των σχετικών εγγράφων που απαιτούνται από το τελικό επίπεδο για να ικανοποιήσει την ανάγκη σύμφωνα με τον τύπο του.

Τώρα, για να κατανομήσουμε τα σχετικά έγγραφα r ομοιόμορφα μεταξύ των μη-σχετικών εγγράφων, χωρίζουμε τα μη-σχετικά έγγραφα r + υποσύνολα 1 κάθε περιέχον i / (r + 1) έγγραφα. Το αναμενόμενο μήκος αναζήτησης είναι τώρα:

$$ESL(q) = j + \frac{is}{r+1}$$

Δεδομένου ότι το μέτρο ESL αποτελεσματικότητας είναι ικανοποιητικό εάν οι ερωτήσεις συλλογής και δοκιμής εγγράφων καθορίζονται. Σε εκείνη την περίπτωση το γενικό μέτρο είναι το μέσο αναμενόμενο μήκος αναζήτησης

$$\overline{ESL} = \frac{1}{Q}\sum_{q \in Q} ESL(q)$$

όπου το Q είναι το σύνολο ερωτήσεων. Αυτή η στατιστική επιλέγεται πιο πολύ από οποιουδήποτε άλλο και αυτό γιατί η παράσταση ελαχιστοποιείται όταν το συνολικό μήκος αναζήτησης είναι αναμενόμενο (πολύ μεγάλο).

$$\sum_{q \in Q} ESL(q) \qquad \text{ελαχιστοποιείται}$$

Για να επεκτείνουμε τη δυνατότητα εφαρμογής του μέτρου, ώστε να μπορούμε να αντιμετωπίσουμε τις ποικίλες δοκιμαστικές ερωτήσεις και συλλογές εγγράφων, πρέπει να ικανοποιήσουμε το μέτρο ESL για να αντισταθμίσουμε με κάποιο τρόπο την προκατάληψη που εισάγεται, επειδή:

(1) οι ερωτήσεις ικανοποιούνται από τους διαφορετικούς αριθμούς εγγράφων σύμφωνα με τον τύπο της ερώτησης και επομένως μπορεί να αναμένουμε να έχουν διαφορετικά μήκη αναζήτησης.

(2) η πυκνότητα των σχετικών εγγράφων για μια ερώτηση σε μια συλλογή εγγράφων μπορεί να είναι σημαντικά διαφορετική από την πυκνότητα σε άλλη.

Η πρώτη πρόταση προτείνει ότι το μέτρο ESL ανά επιθυμητό σχετικό έγγραφο είναι Πραγματικά επιθυμητό ως δείκτης της αξίας. Η δεύτερη προτείνει το ESL σαν έναν παράγοντα ανάλογο προς τον αναμενόμενο αριθμό μη-σχετικών εγγράφων που συλλέγονται για κάθε ένα. Επιτυχώς προκύπτει ότι η διόρθωση της παραλλαγή στις ερωτήσεις δοκιμής και συλλογής εγγράφων μπορεί να γίνει με τη σύγκριση του ESL με το **αναμενόμενο τυχαίο μήκος αναζήτησης (ERSL)**. Αυτή η τελευταία ποσότητα μπορεί να προσεγγιστεί με τον υπολογισμό του αναμενόμενου μήκους αναζήτησης όταν ανακτάται ολόκληρη η συλλογή εγγράφων σε ένα επίπεδο. Το τελικό μέτρο είναι επομένως:

$$D = \frac{ERSL(q) - ESL(q)}{ERSL(q)}$$

το οποίο έχει οριστεί ως αναμενόμενο παράγοντα μείωσης μήκους αναζήτησης από τον Cooper. Κατά προσέγγιση μετρά τη βελτίωση πέρα από την τυχαία ανάκτηση. Η ρητή μορφή για ERSL δίνεται από τη σχέση:

$$\text{ERSL}(q) = \frac{S \cdot I}{R + 1}$$

όπου

(1) Το R είναι ο συνολικός αριθμός εγγράφων στη συλλογή σχετική με το q

(2) Το I είναι ο συνολικός αριθμός εγγράφων στη συλλογή μη-σχετική στο q

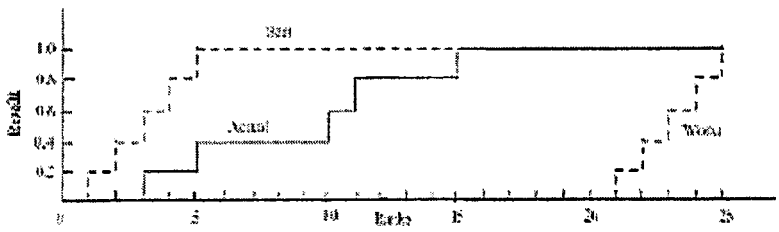
(3) Το S είναι ο συνολικός επιθυμητός αριθμός εγγράφων σχετικών με το q.
Τέλος, το γενικό μέτρο για ένα σύνολο ερωτήσεων Q είναι καθορισμένο, σύμφωνο με το μέσο ESL, για να είναι

$$\frac{ERSL - ESL}{ERSL}$$

ο όποιος είναι γνωστός ως μέσος αναμενόμενος παράγοντας μείωσης μήκους αναζήτησης. Η αποδοχή της ως μέτρο της αποτελεσματικότητας είναι ακόμα αμφισβητήσιμη. Αγνοεί συνολικά την πτυχή ανάκλησης της ανάκτησης, εκτός αν οι ερωτήσεις αξιολογούνται ώστε να εκφράζουν την ανάγκη για ένα ορισμένο ποσοστό των σχετικών εγγράφων στο σύστημα. Επομένως φαίνεται να είναι καλό υποκατάστατο της ακρίβειας, μια και λαμβάνει υπόψη τις ποσότητες ανάκτησης και ανάγκης των χρηστών.

Τα μέτρα του SMART

Το 1966, Rocchio έδωσε μια παραγωγή δύο γενικών δεικτών της τιμής βασισμένη στην ανάκληση και την ακρίβεια. Προτάθηκαν για την αξιολόγηση των συστημάτων ανάκτησης που ταξινόμησε τα έγγραφα, και είχαν ως σκοπό να είναι ανεξάρτητοι από την τιμή του κατωφλίου αποκοπής (cut-off).
Ο πρώτος αυτών των δεικτών είναι η ομαλοποιημένη ανάκληση. Μετρά κατά προσέγγιση την αποτελεσματικότητα της ταξινόμησης σε σχέση με την καλύτερη δυνατή και χειρότερη πιθανή ταξινόμηση. Η κατάσταση αυτή φαίνεται στο σχήμα 34 για 25 έγγραφα όπου σχεδιάζουμε στο Y - άξονας και οι τάξεις στο X - άξονας.



Σχήμα 34: Η κανονικοποιημένη καμπύλη ανάκλησης ορίζεται από τις καλύτερες και χειρότερες περιπτώσεις.

Η ομαλοποιημένη ανάκληση (Rnorm) είναι η περιοχή μεταξύ της πραγματικής περίπτωσης και του χειρότερου ως ποσοστό της περιοχής μεταξύ του καλύτερου και του χειρότερου. Εάν n είναι ο αριθμός σχετικών εγγράφων και ri η τάξη στην οποία το

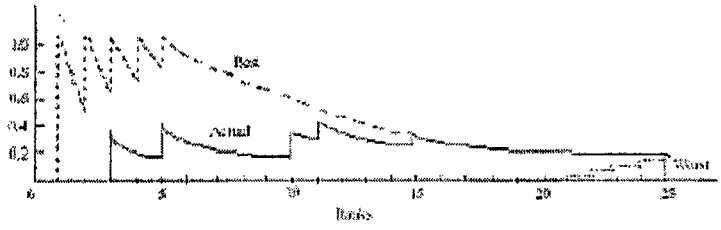
έγγραφο ith (ιοστό) ανακτάται, τότε η περιοχή μεταξύ της καλύτερης και πραγματικής περίπτωσης μπορεί να αποδειχθεί ότι είναι:

$$A_b - A_a = \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{a} i}{n}$$

Μια κατάλληλη ρητή μορφή ομαλοποιημένης ανάκλησης είναι:

$$R_{norm} = 1 - \frac{\Sigma r_i - \Sigma i}{n(N-n)}$$

όπου N είναι ο αριθμός εγγράφων στο σύστημα και το N - n η περιοχή μεταξύ καλύτερης και της χειρότερης περίπτωσης (για να δει αυτό το υποκατάστατο ri = N – I + 1 στον τύπο για το Ab - Aα). Η μορφή εξασφαλίζει ότι το Rnorm βρίσκεται μεταξύ 0 (για τη χειρότερη περίπτωση) και 1 (για την καλύτερη περίπτωση).



Σχήμα 35: Η κανονικοποιημένη καμπύλη ακρίβειας ορίζεται από τις καλύτερες και χειρότερες περιπτώσεις.

Κατά τρόπο ανάλογο η ομαλοποιημένη ακρίβεια βρίσκει λύση. Στο σχήμα 35 έχουμε ακόμα μια φορά τρεις καμπύλες που παρουσιάζουν (1) καλύτερη περίπτωση, (2) η πραγματική περίπτωση, και (3) η χειρότερη περίπτωση από την άποψη των τιμών ακρίβειας στις διαφορετικές θέσεις κατάταξης.
Ο υπολογισμός των περιοχών είναι λίγο πιο ακατάστατος αλλά γίνεται πιο απλός. Η περιοχή μεταξύ της πραγματικής και καλύτερης περίπτωσης δίνεται τώρα από τι τύπο:

$$A_b - A_a = \sum_{i=1}^{a} \log r_i - \sum_{i=1}^{a} \log i$$

Η λειτουργία λογάριθμων εμφανίζεται ως αποτέλεσμα να προσεγγίσει το Σ 1/r από το συνεχές ανάλογό του ∫ 1/rdr, το οποίο είναι log r + σταθερά.
Η περιοχή μεταξύ της χειρότερης και καλύτερης περίπτωσης λαμβάνεται με τον ίδιο τρόπο όπως πριν χρησιμοποιεί την ίδια αντικατάσταση, και είναι:

$$K = \log \frac{N!}{(N-n)!n!}$$

Η ρητή μορφή, με την κατάλληλη κανονικοποίηση, για την ομαλοποιημένη ακρίβεια είναι επομένως:

$$P_{norm} = 1 - \frac{\sum \log r_i - \sum \log i}{\log \left(\frac{N!}{(N-n)!\,n!} \right)}$$

Άλλη μια φορά ποικίλλει μεταξύ 0 (το χειρότερο) και 1 (το καλύτερο).

Μερικά σχόλια για αυτά τα μέτρα είναι για την κατάταξη. Αρχικά η συμπεριφορά τους είναι συνεπής υπό την έννοια ότι εάν ένας από τους είναι 0 (ή 1) έπειτα άλλη είναι 0 (ή 1). Με άλλα λόγια και οι δύο συμφωνούν σχετικά με την καλύτερη και χειρότερη απόδοση. Αφετέρου, διαφέρουν σε βάρη που ορίζονται στις αυθαίρετες θέσεις της καμπύλης ακρίβειας-ανάκλησης και αυτά τα βάρη μπορούν να διαφέρουν αρκετά από εκείνα που ο χρήστης θεωρεί ότι είναι σχετικά. Ισχύει ο κανόνας ότι: "το κανονικοποιημένο μέτρο ακρίβειας, ορίζει ένα πολύ μεγαλύτερο βάρος στις αρχικές βαθμίδες εγγράφων, απ' ό,τι στις πιο πρόσφατες, ενώ το κανονικοποιημένο μέτρο ανάκλησης, ορίζει ένα ομοιόμορφο βάρος σε όλα τα σχετικά έγγραφα". Δυστυχώς, η στάθμιση είναι αυθαίρετη και τρίτον, μπορεί να αποδειχθεί ότι οι ομαλοποιημένες, ανάκληση και η ακρίβεια, έχουν τις ερμηνείες ως προσεγγίσεις στις μέσες τιμές ανάκλησης και ακρίβειας για όλα τα πιθανά επίπεδα διακοπών. Δηλαδή εάν το R(i) είναι η ανάκληση στην πυκνή θέση i, και P(i) η αντίστοιχη τιμή ακρίβειας, τότε:

$$R_{norm} = \frac{1}{N} \sum_{i=1}^{N} R(i)$$

$$P_{norm} = \frac{1}{N} \sum_{i=1}^{N} P(i)$$

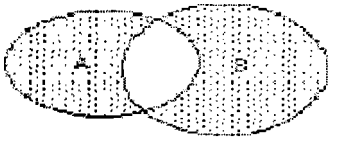
Τελικά η αναφορά πρέπει να αποτελεστεί από δύο παρόμοια αλλά απλούστερα μέτρα που χρησιμοποιούνται από το σύστημα SMART. Δηλαδή είναι :

$$\text{Rank Recall} = \frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i} \qquad \text{Log Precision} = \frac{\sum_{i=1}^{n} \ln i}{\sum_{i=1}^{n} \ln r_i}$$

μην λαμβάνοντας υπόψη το μέγεθος N της συλλογής, όπου n εδώ είναι ο αριθμός σχετικών εγγράφων για την ιδιαίτερη ερώτηση δοκιμής.

Μια κανονικοποιημένη συμμετρική διαφορά

Εξετάζουμε την κοινή κατάσταση όπου ένα σύνολο εγγράφων ανακτάται ως απάντηση σε μια ερώτηση, η πιθανή διάταξη αυτού του συνόλου αγνοείται. Ιδανικά το σύνολο πρέπει να αποτελεστεί μόνο από τα έγγραφα σχετικά με το αίτημα, το οποίο δίνει 100 τοις εκατό ακρίβειας και την ανάκληση 100 τοις εκατό (και επαγωγικώς 0 τοις εκατό fallout). Στην πράξη, εντούτοις, αυτό είναι σπάνια περίπτωση, και το ανακτημένο σύνολο αποτελείται και από τα σχετικά και μη-σχετικά έγγραφα. Η κατάσταση μπορεί επομένως να απεικονιστεί όπως φαίνεται στο σχήμα 36, όπου το Α είναι το σύνολο σχετικών εγγράφων, του Β το σύνολο ανακτημένων εγγράφων, και του Α ∩ Β το σύνολο ανακτημένων εγγράφων που είναι σχετικά.



Σχήμα 36: Ένα παράδειγμα της συμμετρικής διαφοράς μεταξύ δύο ομάδων Α και Β. Στην σκιασμένη περιοχή είναι $A \triangle B$.

Τώρα, ένας διαισθητικός τρόπος στην επάρκεια του ανακτημένου συνόλου είναι να μετρηθεί το μέγεθος της σκιασμένης περιοχής. Ή για να το θέσουμε διαφορετικά, μέχρι ποιο σημείο τα δύο σύνολα δεν ταιριάζουν . Η περιοχή είναι στην πραγματικότητα η συμμετρική διαφορά: Α . Β (ή Α □ Β – Α ∩ Β).
Δεδομένου ότι ενδιαφερόμαστε για το ποσοστό (παρά τον απόλυτο αριθμό) των σχετικών και μη-σχετικών εγγράφων που ανακτώνται, πρέπει να κανονοκοποιήσουμε αυτό το μέτρο. Μια απλή κανονικοποίηση δίνει

$$E = \frac{|A \triangle B|}{|A| + |B|}$$

In terms of P and R we have:

$$E = 1 - \frac{1}{\frac{1}{2}\left(\frac{1}{P}\right) + \frac{1}{2}\left(\frac{1}{R}\right)}$$

Το όποιο είναι ένα απλό σύνθετο μέτρο.

Το προηγούμενο επιχείρημα σε αυτό δεν είναι επαρκές για να δικαιολογήσει τη χρήση αυτού του ιδιαίτερου σύνθετου μέτρου. Εντούτοις, θα εισάγουμε τώρα ένα πλαίσιο μέσα στο οποίο ένα γενικό μέτρο μπορεί να προκύψει και έχει μεταξύ των άλλων το Ε ως μια από τις ειδικές περιπτώσεις του.

Το πρότυπο

Αρχίζουμε με την εξέταση της δομής υποθέτοντας λογικά, την μέτρηση της αποτελεσματικότητας. Με άλλα λόγια, εξετάζουμε τους όρους που περιμένουμε να ικανοποιήσουν οι παράγοντες που καθορίζουν την αποτελεσματικότητα. Περιορίζουμε τη συζήτηση εδώ σε δύο παράγοντες, δηλαδή την ακρίβεια και την ανάκληση, αν και αυτό δεν είναι κανένας περιορισμός, διαφορετικοί παράγοντες θα μπορούσαν να αναλυθούν και όπως θα υποδειχθεί αργότερα, περισσότεροι από δύο παράγοντες μπορούν να απλοποιήσουν την ανάλυση.
Εάν R είναι το σύνολο πιθανών τιμών ανάκλησης και P είναι το σύνολο πιθανών τιμών ακρίβειας τότε ενδιαφερόμαστε για τον καθορισμένο R x P. Θα αναφερθούμε σε αυτό ως συσχετική δομή και θα δείξουμε το < R x P, > >=όπου >= είναι η δυαδική σχέση R x P. Για οποιοδήποτε δεδομένο σημείο (R, P) είμαστε σε θέση να πούμε εάν δείχνει περισσότερη, λιγότερη ή ίση αποτελεσματικότητα από αυτή που υποδεικνύεται από κάποιο άλλο σημείο. Το είδος σχέσης διάταξης είναι μια αδύναμη διάταξη. Για να είναι ακριβής:

Ορισμός 1: Η συγγενικά δομή < R x P, >= > είναι μια αδύναμη διάταξη εάν και μόνο εάν για e1, e2, e3 € R x P τα ακόλουθα αξιώματα ικανοποιούνται:

(1) Συνδετικότητα: είτε e1 >= e2 είτε e2 >= e1

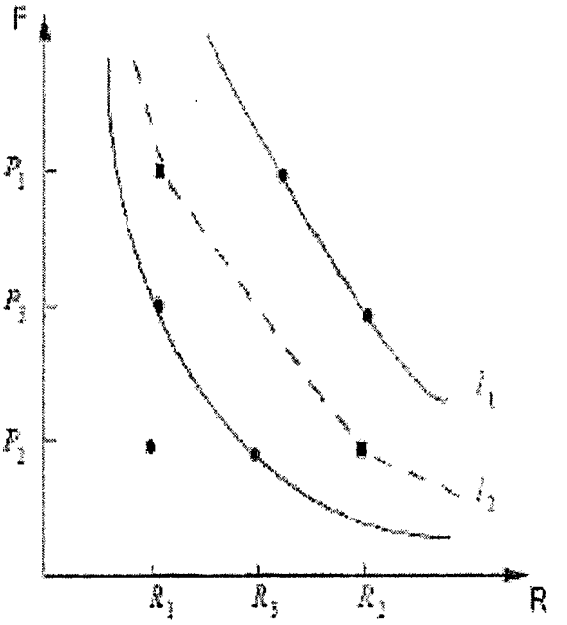
(2) Μεταβατικότητα: εάν e1>= e2 και e2>= e3 τότε e1 >= e3

Εάν δύο ζευγάρια μπορούν να διαταχτούν και με τους δύο τρόπους τότε (R1, P1) ~ (R2, P2), δηλ. άνισο όχι απαραιτήτως ίσο. Ο όρος μεταβατικότητα είναι προφανώς επιθυμητός.
Γυρίζουμε τώρα σε έναν δεύτερο όρο που καλείται **ανεξαρτησία.** Αυτή η έννοια συλλαμβάνει την ιδέα ότι τα δύο συστατικά συμβάλλουν τα αποτελέσματά τους ανεξάρτητα από την λειτουργικότητα του συστήματος.

Ορισμός 2: Μια σχέση >= για R x P είναι ανεξάρτητη εάν και μόνο εάν, για R1, R2 υπάρχει r, (R1, P) > = (R2, P) για κάποιο P € P συνεπάγεται (R1, P') >= (R2, P') για κάθε P' € P και για P1, P2 € P, (R, P1) >= (R, P2) για κάποιο R' € R συνεπάγεται (R', P1) >= (R', P2) για κάθε R' € R.

Δεδομένου ότι σε μια σταθερή ανάκληση (ή ακρίβεια) βρίσκουμε μια διαφορά στην αποτελεσματικότητα για δύο τιμές της ακρίβειας (ή ανάκλησης). Αυτή η διαφορά δεν μπορεί να αφαιρεθεί ή να αντιστραφεί με την αλλαγή της σταθερής τιμής.
Ερχόμαστε τώρα σε έναν όρο που δεν είναι αρκετά τόσο προφανής όσο τους προηγουμένους. Για να το καταστεθεί σημαντικότερο θα πρέπει να χρησιμοποιήσουμε ένα διάγραμμα, σχήμα 37, το οποίο αντιπροσωπεύει τη διάταξη που έχουμε μέχρι τώρα με τους ορισμούς 1 και 2. Οι γραμμές l1 και l2 είναι γραμμές ίσης αποτελεσματικότητας που είναι οποιαδήποτε δύο σημεία (R, P), (R', P')€ li τέτοιο που

(R, P) ~ (R', P') (όπου ~ δείχνει την ίση αποτελεσματικότητα). Τώρα υποθέτουμε ότι έχουμε τα σημεία 11 και 12 αλλά επιθυμούμε να συναγάγουμε τη συσχέτιση που διατάσει μεταξύ αυτών των δύο γραμμών. Κάποιος μπορεί να σκεφτεί αυτό ως διαδικασία παρεμβολής.



Σχήμα 37: Ένα παράδειγμα της συνθήκης Thomsen.

Ορισμός 3 (συνθήκη Thomsen): Για κάθε R1, R2, R3 € R και P1, P2, P3€ P, (R1, P3) ~ (R3, P2) και (R3, P1) ~ (R2, P3) συνεπάγεται (R1, P1) ~ (R2, P2).

Αυτό μπορεί να αιτιολογηθεί ως εξής. Τα διαστήματα R1 R3 και P2 P3 είναι ισοδύναμα από μια αύξηση στο R - παράγοντας από R1 R3 και μια αύξηση στο P - παράγοντα από το(R1, P3) μέχρι P2 P3 που οδηγεί στην ίδια αποτελεσματικότητα (σημεία 12). Επομένως ακολουθεί μια μείωση σε κάθε παράγοντα που αρχίζει από την ίση αποτελεσματικότητα. Σε αυτήν την περίπτωση τα δύο σημεία (R3, R1) και (R2, P3) στο 11, πρέπει να οδηγήσει σε ίση αποτελεσματικότητα.

Ο τέταρτος όρος έχει σχέση με τη συνοχή κάθε συστατικού. Κάνει ακριβώς ότι θα αναμέναμε κατά την εξέταση της ύπαρξης των ενδιάμεσων τιμών.

Ορισμός 4: (περιορισμένη επίλυση). Μια σχέση >= R x P ικανοποιεί την περιορισμένη επίλυση υπό τον όρο ότι:

(1) *R*, *R* , *R* € R και P, P' € P για το οποίο (R', P') >= (R, P) >= (R, P') τότε *R* € R s.t (R, P') ~ (R, P);

(2) ένας παρόμοιος όρος ισχύει στο δεύτερο συστατικό.

Με άλλα λόγια εξασφαλίζουμε ότι η εξίσωση (R', P') ~ (R, P) είναι επιλύσιμη για το R' υπό τον όρο ότι υπάρχει R' έτσι ώστε (`R, P')>= (R, P') >= (R, P'). Μια υπόθεση της συνοχής των παραγόντων ακρίβειας και ανάκλησης θα εξασφάλιζε αυτό.

Ο πέμπτος όρος δεν είναι περιοριστικός αλλά πρέπει να δηλωθεί. Απαιτεί, με έναν ακριβή τρόπο, ότι κάθε συστατικό είναι απαραίτητο.
Ορισμός 5. Το συστατικό R είναι ουσιαστικό εάν και μόνο εάν υπάρχει R1, R2 υπάρχει R και P1 υπάρχει P έτσι ώστε δεν είναι η περίπτωση που (R1, P1) ~ (R2, P1).
Έχουμε τώρα έξι όρους στη συσχετική δομή < R x P, >= >, που στη θεωρεία της μέτρησης είναι απαραίτητοι και επαρκείς όροι για να υπάρξει μια αθροιστική ενωμένη δομή. Αυτό είναι αρκετό για να δηλωθεί το κύριο θεώρημα αντιπροσώπευσης. Είναι ένα θεώρημα που βεβαιώνει ότι εάν μια δεδομένη συσχετική δομή ικανοποιεί ορισμένους όρους (αξιώματα), κατόπιν ένας ομοιομορφισμός στους πραγματικούς αριθμούς αναφέρεται συχνά ως κλίμακα. Η μέτρηση μπορεί επομένως να θεωρηθεί ως κατασκευή ομοιομορφισμών για τις εμπειρικές συσχετικές δομές στις αριθμητικές συσχετικές δομές που είναι χρήσιμες.
Στην περίπτωσή μας μπορούμε επομένως να αναμείνουμε να βρούμε τις real-valued λειτουργίες Φ1 στο R και Φ2 στο P και μια λειτουργία F από το Re x Re sto Re, 1:1 σε κάθε μεταβλητή, έτσι ώστε, για όλο το R, R' ανήκει στο R και P, P' ανήκει στο P έχουμε:

$$(R, P) \geq (R', P') \Leftrightarrow F[\Phi_1(R), \Phi_2(P)] \geq F[\Phi_1(R'), \Phi_2(P')]$$

(Σημειώνουμε ότι αν και το ίδιο σύμβολο >= χρησιμοποιείται, το πρώτο είναι μια δυαδική σχέση R x P, το δεύτερο είναι το σύνηθες στο Re, στο σύνολο των πραγματικών.)
Με άλλα λόγια υπάρχουν αριθμητικές κλίμακες Φi στα δύο συστατικά και ένας κανόνας F για το συνδυασμό τους έτσι ώστε το επακόλουθο μέτρο συντηρεί την ποιοτική διάταξη της αποτελεσματικότητας. Όταν μια τέτοια αντιπροσώπευση υπάρχει λέμε ότι η δομή είναι αποσυνθέσιμη (decomposable). Σε αυτήν την αντιπροσώπευση τα συστατικά (R και P) συμβάλλουν στο μέτρο αποτελεσματικότητας ανεξάρτητα. Δεν είναι αλήθεια ότι όλες οι συσχετικές δομές είναι αποσυνθέσιμο. Αυτό που είναι αληθινό, εντούτοις, είναι ότι οι μη- αποσυνθέσιμες δομές είναι εξαιρετικά δύσκολο να αναλυθούν.

Μια περαιτέρω απλοποίηση της λειτουργίας μέτρησης μπορεί να επιτευχθεί με την απαίτηση ενός ειδικού είδους μη αλληλεπίδρασης των συστατικών που έχει γίνει γνωστό ως πρόσθετη ανεξαρτησία. Αυτό απαιτεί ότι η εξίσωση για τις αποσυνθέσιμες δομές μειώνεται

$$(R, P) \geq (R', P') \Leftrightarrow \Phi_1(R) + \Phi_2(P) \geq \Phi_1(R') + \Phi_2(P')$$

όπου το F είναι απλά η λειτουργία προσθηκών. Ένα παράδειγμα μιας μη αποσυνθέσιμης δομής δίνεται από:

$$(R, P) \geq (R', P') \Leftrightarrow \Phi_1(R) + \Phi_2(P) + \Phi_1(R)\Phi_2(P) \geq \Phi_1(R') + \Phi_2(P') + \Phi_1(R')\Phi_2(P')$$

Μπορεί να αποδειχθεί ότι μια πρόσθετα ανεξάρτητη αντιπροσώπευση που οι ιδιότητες καθορίζονται σε 1 και 3. Οι δομικοί όροι 4 και 5 είναι ικανοποιητικοί.
Εδώ ο όρος Φ1 Φ2 αναφέρεται ως όρος αλληλεπίδρασης, οι απολογισμοί απουσίας του για την μη-αλληλεπίδραση στον προηγούμενο όρο.
Είμαστε τώρα σε θέση να δηλώσουμε το κύριο θεώρημα αντιπροσώπευσης.

Θεώρημα

Υποθέτουμε <R x P,>= > ότι είναι μια πρόσθετη ενωμένη δομή, κατόπιν υπάρχουν οι λειτουργίες, Φ1 από το R, και Φ2 από το P στους πραγματικούς αριθμούς έτσι ώστε, για όλο το R, R' ∈R και P, P' ∈ P:

$$(R, P) \geq (R', P') \Leftrightarrow \Phi_1(R) + \Phi_2(P) \geq \Phi_1(R') + \Phi_2(P')$$

Εάν Φi είναι δύο άλλες λειτουργίες με την ίδια ιδιότητα , κατόπιν υπάρχουν οι σταθερές θ > 0, γ1 και γ2 τέτοια ώστε

$$\Phi_1' = \theta\Phi_1 + \gamma_1 \qquad \Phi_2' = \theta\Phi_2 + \gamma_2$$

Μέχρι τώρα έχουμε συζητήσει τις ιδιότητες μιας πρόσθετης ενωμένης δομής και έχουμε δικαιολογήσει τη χρήση της για τη μέτρηση της αποτελεσματικότητας βασισμένη στην ακρίβεια και την ανάκληση. Επίσης έχουμε δείξει ότι μια πρόσθετα ανεξάρτητη αντιπροσώπευση (μοναδική μέχρι έναν γραμμικό μετασχηματισμό) υπάρχει για αυτό το είδος συσχετικής δομής. Για να καθοριστεί η μορφή του Φi πρέπει να εισαγάγουμε μερικές εκτιμήσεις. Αν και η αντιπροσώπευση F = Φ1 + Φ2, αυτό δεν είναι η καταλληλότερη μορφή για την έκφραση των περαιτέρω όρων που απαιτούμε από το F, ούτε για την ερμηνεία της. Έτσι, παρά το γεγονός ότι επιδιώκουμε μια πρόσθετα ανεξάρτητη αντιπροσώπευση εξετάζουμε τους όρους σε ένα γενικό F. Θα προκύψει ότι το F το οποίο είναι κατάλληλο μπορεί να μετασχηματιστεί απλά σε μια πρόσθετη αντιπροσώπευση. Ο μετασχηματισμός είναι $f(F) = -(F - 1)^{-1}$ που αυξάνεται γνησίως μονότονα στο $0 \leq F \leq 1$, το οποίο είναι σημείο ενδιαφέροντος. Εν πάση περιπτώσει, κατά τη μέτρηση της αποτελεσματικότητας ανάκτησης οποιοσδήποτε μετασχηματισμό της μονοτονίας του μέτρου θα γίνει εξίσου καλά.

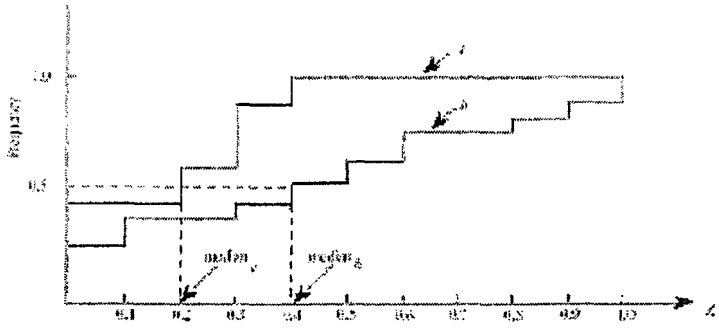
Παρουσίαση των πειραματικών αποτελεσμάτων

Διάφοροι τρόποι του μέτρου αποτελεσματικότητας στο σύνολο των ερωτήσεων προέκυψαν με έναν φυσικό τρόπο. Θέλουμε τώρα να εξετάσουμε τους τρόπους με τους οποίους μπορούμε να συνοψίσουμε τα αποτελέσματα ανάκτησής μας.
Σε αυτό το τμήμα η συζήτηση θα περιοριστεί στα ενιαία μέτρα αριθμού όπως μια κανονικοποιημένη συμμετρική διαφορά, κανονικοποιημένη ανάκληση, κ.λπ. θα χρησιμοποιήσουμε το Z για να δείξουμε οποιοδήποτε αυθαίρετο μέτρο. Οι ερωτήσεις δοκιμής θα είναι Qi και n σε αριθμό. Ο στόχος μας σε όλο αυτό είναι να κάνουμε τις

δηλώσεις για τις σχετικές τιμές της ανάκτησης υπό τους διαφορετικούς όρους a,b,c .. Από την άποψη του μέτρου της αποτελεσματικότητας Z. Οι "όροι" .. μπορεί να είναι διαφορετικής στρατηγικής αναζήτησης, ή δομές πληροφοριών, κ.λπ.... Με άλλα λόγια, έχουμε τη συνηθισμένη πειραματική οργάνωση όπου ελέγχουμε μια μεταβλητή και ένα μέτρο, το πώς δηλαδή η αλλαγή της επηρεάζει την αποτελεσματικότητα ανάκτησης. Προς το παρόν περιορίζουμε αυτές τις συγκρίσεις σε ένα σύνολο ερωτήσεων και της ίδιας συλλογής εγγράφων.
Οι μετρήσεις που έχουμε επομένως είναι:
{Za(Q1), Za(Q2), . . . }, {Zb(Q1), Zb(Q2), . . . },
{Zc(Q1), Zc(Q2), . . . }, . . .όπου Zx (Q1) είναι η τιμή του Z κατά τη μέτρησή της αποτελεσματικότητας της απάντησης σε Qi υπό τους όρους x. Εάν επιθυμούμε τώρα να κάνουμε μια γενική σύγκριση μεταξύ αυτών των συνόλων μετρήσεων θα μπορούσαμε να πάρουμε τα μέσα και να συγκρίνουμε αυτές. Το πρόβλημα με τα στοιχεία IR είναι ένα εμπόδιο από την αρχή του θέματος. Λόγω της μη παραμετρικής φύσης των στοιχείων δεν πρόκειται καλύτερα να αναφέρει μια ενιαία στατιστική αλλά άντ' αυτού να παρουσιάσει την παραλλαγή στην αποτελεσματικότητα με τη χάραξη των γραφικών παραστάσεων. Εάν είναι απαραίτητο να αναφερθούν τα "μέσα" αποτελέσματα είναι σημαντικό ότι αναφέρονται παράλληλα με τη κατανομή από την οποία προέρχονται.

Υπάρχουν διάφοροι τρόποι τα σύνολα του Z- τιμές να δοθούν γραφικά. Πιθανώς ο προφανέστερος είναι να χρησιμοποιηθεί ένα διάγραμμα διασποράς, όπου ο X - άξονας αναφέρεται για Za και ο Y - άξονας για Zb και κάθε ένα σχεδιασμένο σημείο είναι το ζεύγος (Za(Qi), Zb(Qi)).
Ο αριθμός σημείων που σχεδιάζονται θα είναι ίσος με τον αριθμό ερωτήσεων. Εάν σύρουμε τώρα μια γραμμή σε 45 μοίρες στο x - άξονας από την προέλευση θα είμαστε σε θέση να δούμε ποιο ποσοστό των ερωτήσεων λειτουργούν καλύτερα υπό συνθήκη a απ' ό,τι υπό συνθήκη b. Υπάρχουν δύο μειονεκτήματα σε αυτήν την μέθοδο αντιπροσώπευσης: η σύγκριση περιορίζεται σε δύο όρους, και είναι δύσκολο να αποκτηθεί μια ιδέα του βαθμού στον οποίο δύο όροι διαφέρουν.

Ένας καταλληλότερος τρόπος αποτελέσματος ανάκτησης αυτού του είδους είναι να σχεδιαστούν ως κατανομές μεμονωμένης συχνότητας, ή όπως καλούνται συχνά από τις εμπειρικές λειτουργίες κατανομής στατιστικών. Έστω {Z(Q1), Z(Q2), . . . , Z(Qn)} να είναι ένα σύνολο αποτελεσμάτων ανάκτησης, τότε η εμπειρική συνάρτηση κατανομής F (z) είναι μια συνάρτηση του z που είναι ίση με το ποσοστό του Z(Qi) που είναι λιγότερο ή ίσο προς z. Για να σχεδιάσουμε αυτήν την συνάρτηση διαιρούμε τη σειρά του z σε διαστήματα. Εάν υποθέτουμε ότι $0 \leq z \leq 1$, τότε ένα κατάλληλο σύνολο διαστημάτων είναι δέκα. Οι κατανομές θα πάρουν τη γενική μορφή όπως φαίνεται στο σχήμα 38. Όταν το μέτρο z είναι τέτοιο που όσο μικρότερη η τιμή του τόσο αποτελεσματικότερη η ανάκτηση, τότε η υψηλότερη καμπύλη είναι η καλύτερη. Παραδείγματος χάριν, για να βρούμε τη διάμεσο πρέπει μόνο να βρούμε το z (εκτιμούμε την αντιστοιχία σε 0,5 στον άξονα F (z)). Στα διαγράμματά μας είναι 0,2 και 0,4 αντίστοιχα για τους όρους a και b.



Σχήμα 38: Δύο αθροιστικές συχνότητες κατανομής που δείχνουνε την διαφορά της αποτελεσματικότητας κάτω από τις συνθήκες a και b

Έχουμε υπογραμμίσει τη μέτρηση της αποτελεσματικότητας από την άποψη του χρήστη. Εάν επιθυμούμε τώρα μπορούμε να συγκρίνουμε την ανάκτηση στις διαφορετικές συλλογές εγγράφων με τα διαφορετικά σύνολα ερωτήσεων τότε μπορούμε ακόμα να χρησιμοποιήσουμε αυτά τα μέτρα για να προσδιορίσουμε ποιο σύστημα ικανοποιεί το χρήστη περισσότερο. Αφ' ετέρου, δεν μπορούμε με αυτόν τον τρόπο να πιστοποιήσουμε ποιο σύστημα είναι αποτελεσματικότερο σε διαδικασίες ανάκτησής του. Μπορεί να είναι εκείνο το σύστημα Α που τα σύνολα σχετικών εγγράφων αποτελούν ένα μικρότερο ποσοστό του συνόλου των εγγράφων από αυτό που συμβαίνει στο σύστημα Β. Με άλλα λόγια, είναι πολύ πιο δύσκολο να βρεθούν τα σχετικά έγγραφα στο σύστημα Β απ 'ό,τι στο σύστημα Α. Έτσι, οποιαδήποτε άμεση σύγκριση πρέπει να σταθμιστεί από το μέτρο γενικότητας που δίνει τον αριθμό σχετικών εγγράφων ως ποσοστό του συνολικού αριθμού εγγράφων. Εναλλακτικά κάποιος θα μπορούσε να χρησιμοποιήσει το fallout που μετρά το ποσοστό των μη-σχετικών εγγράφων που ανακτώνται. Το σημαντικό σημείο πρόκειται εδώ να είναι σαφές για να μετρήσουμε την ικανοποίηση χρηστών ή την αποτελεσματικότητα συστημάτων.

Δοκιμές σημασίας

Μόλις έχουμε τους αριθμούς αποτελεσματικότητας ανάκτησής μας μπορούμε να καθορίσουμε ότι η διαφορά στην αποτελεσματικότητα υπό δύο όρους είναι στατιστικά σημαντική. Για αυτόν το λόγο πολλές στατιστικές δοκιμές έχουν σχεδιαστεί. Δυστυχώς δεν υπάρχει καμία γνωστή στατιστική δοκιμή εφαρμόσιμη στην ανάκτηση πληροφορίας. Αυτό μπορεί να ακουστεί ως μια πρόταση αποτυχίας αλλά προσθέτουμε ότι είναι δυνατό να επιλεχτεί μια δοκιμή που παραβιάζει μερικές από τις υποθέσεις που δεχτήκαμε. Οι παραμετρικές δοκιμές είναι ακατάλληλες επειδή δεν ξέρουμε τη μορφή της συνεπαγόμενης κατανομής. Σε αυτήν την κατηγορία πρέπει να περιλάβουμε τη γνωστή t- δοκιμή. Μια προφανής αποτυχία είναι ότι οι παρατηρήσεις δεν προέρχονται από τους κανονικά κατανεμημένους πληθυσμούς.
Υπάρχουν μερικές δοκιμές για την εξέταση της περίπτωσης των σχετικών δειγμάτων. Στην πειραματική οργάνωσή μας έχουμε ένα σύνολο ερωτήσεων που χρησιμοποιείται στα διαφορετικά περιβάλλοντα ανάκτησης. Επομένως, χωρίς ερώτηση εάν έχουμε τα τυχαία δείγματα, είναι σαφές ότι το δείγμα υπό τον όρο a συσχετίζεται με το δείγμα

υπό τον όρο b. Όταν σε αυτήν την κατάσταση μια κοινή δοκιμή που χρησιμοποιεί είναι η δοκιμή Wilcoxon Matched-Pairs. Δυστυχώς πάλι μερικές σημαντικές υποθέσεις δεν συναντιούνται. Η δοκιμή γίνεται στη διαφορά Di = Za(Qi) - Zb(Qi), αλλά υποτίθεται ότι το Di είναι συνεχές και ότι προέρχεται από μια συμμετρική κατανομή, καμία από την οποία δεν συναντιέται κανονικά στα δεδομένα ανάκτησης πληροφορίας. Φαίνεται επομένως ότι μερικές από τις περιπλοκότερες στατιστικές δοκιμές είναι ακατάλληλες. Υπάρχει, εντούτοις, μια απλή δοκιμή που κάνει πολύ λίγες υποθέσεις και σημειώνεται ότι μπορούν να χρησιμοποιηθούν παρέχοντας περιορισμούς . Ισχύει στην περίπτωση των σχετικών δειγμάτων. Υποθέτει ότι τα στοιχεία προκύπτουν από μία συνεχή μεταβλητή και ότι το Z (Qi) είναι στατιστικά ανεξάρτητα. Αυτοί οι δύο όροι είναι απίθανο να ικανοποιηθούν σε ένα πείραμα ανάκτησης. Εντούτοις, δεδομένου ότι μερικοί από τους όρους δεν ικανοποιούνται, μπορεί να χρησιμοποιηθεί περιοριστικά.
Ο τρόπος εργασίας είναι ο ακόλουθος: Αφήνουμε Za (Q1), Za (Q2)..., , Zb (Q1), Zb (Q2). .., να είναι δύο σύνολα μετρήσεών μας υπό τους όρους a και b αντίστοιχα. Μέσα σε κάθε ζευγάρι (Za (Qi), Zb (Qi)) γίνεται μια σύγκριση και κάθε ζευγάρι είναι ταξινομημένο όπως "+" εάν Za (Qi) > Zb (Qi), όπως "-" εάν Za (Qi) < Zb (Qi) ή "δεσμός" εάν Za (Qi) = Za (Qi). Τα ζευγάρια που είναι ταξινομημένα ως "δεσμός" αφαιρούνται από την ανάλυση με αυτόν τον τρόπο που μειώνει τον αποτελεσματικό αριθμό μετρήσεων. Η χωρίς τιμή (null) υπόθεση που επιθυμούμε να εξετάσουμε είναι :

P (Za > Zb) = P (Za < Zb) = ½

Σύμφωνα με αυτήν την υπόθεση αναμένουμε τον αριθμό ζευγαριών που έχουν Za > Zb να είναι ίσοι με τον αριθμό ζευγαριών που έχουν Za < Zb. Ένας άλλος τρόπος αφορά το πλήθος για το οποίο τα Za και Zb προέρχονται από την ίδια διάμεσο.
Στην ανάκτηση πληροφορίας μια εναλλακτική δοκιμή θα χρησιμοποιηθεί για να ορίσει την ανωτερότητα της ανάκτησης υπό τον όρο a πέρα από τον όρο b, ή αντίστροφα.

Ένας πίνακας για τα μικρά δείγματα όπου n <= 25 που δίνουν την πιθανότητα κάτω από την null υπόθεση για κάθε πιθανό συνδυασμό '+"s και '–"s και μπορεί να βρεθεί από τον Siegal.
Η χρήση της δοκιμής θίγει διάφορα ενδιαφέροντα σημεία. Ένας από αυτός είναι ότι αντίθετα από τη δοκιμή Wilcoxon υποθέτει ότι τα Z μετριούνται σχετικά με μια τακτική κλίμακα. Αυτό είναι ένα κατάλληλο χαρακτηριστικό γνώρισμα δεδομένου ότι συνήθως μόνο επιδιώκουμε να βρούμε ποια στρατηγική είναι καλύτερη υπό μια μέση έννοια και δεν επιθυμούμε το αποτέλεσμα για να επηρεαστούμε αδικαιολόγητα από την άριστη απόδοση ανάκτησης σε μια ερώτηση. Το δεύτερο σημείο είναι ότι κάποια προσοχή πρέπει να ληφθεί κατά το σύγκριση Za και Zb. Είναι επομένως σημαντικό να αποφασίσουμε εκ των προτέρων σε ποια τιμή ανήκει και να εξισώσουμε τα Za και Zb όταν | Za - Zb | <= ε.
Χρησιμοποιείται επίσης για να ανιχνεύσει μια σημαντική διαφορά μεταξύ των γραφικών παραστάσεων ακρίβειας-ανάκλησης. Ερμηνεύουμε τώρα τα Z ως τιμές ακρίβειας σε ένα σύνολο τυποποιημένων τιμών ανάκλησης. Αφήνουμε αυτό το σύνολο να είναι SR = {0.1, 0,2. .., 1,0} και έπειτα αντιστοιχίζοντας για κάθε R € SR έχουμε ένα ζευγάρι (Pa (R) Pb (R)). Το PA και τα PB αντιμετωπίζονται τώρα με τον ίδιο

τρόπο όπως του Za και του Zb. Κάνοντας την αξιολόγηση με αυτόν τον τρόπο, οι τιμές ακρίβειας-ανάκλησης ήδη θα έχει υπολογιστεί κατά μέσο όρο πέρα από το σύνολο ερωτήσεων από έναν από τους τρόπους που αναφέρθηκαν.

8. ΤΟ ΜΕΛΛΟΝ

Μελλοντική έρευνα

Στα προηγούμενα κεφάλαια έχει γίνει προσπάθεια να συγκεντρωθούν μερικά από τα πιο επιμελημένα εργαλεία που χρησιμοποιούνται κατά τη διάρκεια του σχεδίου ενός πειραματικού συστήματος ανάκτησης πληροφοριών. Πολλά από τα εργαλεία είναι μόνο στο πειραματικό στάδιο και η έρευνα απαιτείται ακόμα, όχι μόνο να αναπτύξει μια κατάλληλη κατανόηση τους, αλλά και για να επιλύσει τις επιπτώσεις τους στο παρόν και το μέλλον συστημάτων IR. Εν συντομία δείχνουμε μερικά από τα θέματα που προκαλούν την περαιτέρω έρευνα.

1. Αυτόματη ταξινόμηση

Τα ουσιαστικά στοιχεία είναι οι μεγάλες συλλογές εγγράφων οι οποίες μπορούν να αντιμετωπιστούν επιτυχώς με τη βοήθεια της αυτόματης ταξινόμησης . Αναμένεται να ωθήσουν το εμπορικό ενδιαφέρον και μαζί με αυτό να υποστηρίξουν την περαιτέρω ανάπτυξη.
Είναι επομένως κάποιας σπουδαιότητας να χρησιμοποιούμε το είδος στοιχείου , που χρησιμοποιεί τις περιγραφές εγγράφων από την άποψη των λέξεων κλειδιών και να καθορίσουμε ότι το έγγραφο που συγκεντρώνεται στις μεγάλες συλλογές εγγράφων μπορεί να είναι και αποτελεσματικό και αποδοτικό. Αυτό σημαίνει ότι περισσότερη

έρευνα απαιτείται για να επινοήσει τους τρόπους τους αλγορίθμους συγκέντρωσης χωρίς να θυσιάσουμε τη δομή στα στοιχεία. Μπορεί να είναι δυνατό να σχεδιαστούν οι πιθανολογικοί αλγόριθμοι, συγκεντρώνοντας τις διαδικασίες που υπολογίσουν μια ταξινόμηση στο μέσο όρο σε λιγότερο χρόνο από ότι μπορεί να απαιτήσει τη χειρότερη περίπτωση. Παραδείγματος χάριν, μπορεί να είναι δυνατό να περικοπεί το 0(n2) χρόνος υπολογισμού αναμενόμενο 0(nlogn), αν και για μερικές περιπτώσεις θα απαιτούσε ακόμα 0(n^2) Ένας άλλος τρόπος αυτό το πρόβλημα στη συγκέντρωση είναι να ψαχτεί τι κάποιο να καλέσει τις ταξινομήσεις. Μπορεί να είναι δυνατό να υπολογιστούν οι δομές ταξινόμησης που είναι κοντά στη θεωρητική δομή , αλλά είναι μόνο στενές προσεγγίσεις που μπορούν να υπολογιστούν αποτελεσματικότερα από το ιδανικό.

Μια μεγάλη ερώτηση, που δεν έχει λάβει ακόμα πολλή προσοχή, αφορά το βαθμό στον οποίο η αποτελεσματικότητα ανάκτησης περιορίζεται από τον τύπο περιγραφής των εγγράφων . Η χρήση των λέξεων κλειδιών για να περιγράψει τα έγγραφα έχει επηρεάσει τον τρόπο με τον οποίο το σχέδιο ενός αυτόματου συστήματος ταξινόμησης έχει προσεγγιστεί. Είναι δυνατό στο μέλλον, τα έγγραφα να αντιπροσωπευθούν από έναν υπολογιστή εξ ολοκλήρου διαφορετικά. Η ομαδοποίηση όμως των εγγράφων θα

είναι ακόμα ενδιαφέρουσα;
Η ταξινόμηση εγγράφων είναι μια ειδική περίπτωση μιας γενικότερης διαδικασίας που θα προσπαθούσε επίσης να εκμεταλλευτεί τις σχέσεις μεταξύ των εγγράφων. Έτσι συμβαίνει να χρησιμοποιούνται οι συντελεστές ανομοιότητας για να εκφράσουν μια τέτοια σχέση. Ποσολογικά η σχέση κατ' αυτό τον τρόπο εν μέρει έχει υπαγορευθεί από τη φύση της γλώσσας στην οποία τα έγγραφα περιγράφονται. Εντούτοις, ήταν αυτή η περίπτωση ότι τα έγγραφα αντιπροσωπεύθηκαν όχι από τις λέξεις κλειδιά αλλά με κάποιο άλλο τρόπο, ίσως με μια πιο σύνθετη γλώσσα. Συνεπώς, η δομή για να αντιπροσωπεύσει τις σχέσεις μπορεί να μην είναι μια απλή ιεραρχία, αλλά ίσως μια ειδική περίπτωση. Με άλλα λόγια, κάποιος πρέπει να πλησιάσει το έγγραφο που συγκεντρώνεται ως διαδικασία στη δομή των στοιχείων για να μπορεί να χρησιμοποιήσει και να καταστήσει την ανάκτηση αποτελεσματική και αποδοτική.

Οι μέθοδοι τις λέξεις κλειδιά, που έχουν αναπτυχθεί ήδη, θα απευθυνθούν επίσης στην αυτόματη κατασκευή των κατηγοριών "ικανοποιημένων μονάδων" που χρησιμοποιούνται κατά τη διάρκεια της ανάκτησης. Η ταξινόμηση λέξης κλειδί θα παραμείνει έπειτα ως ειδική περίπτωση.
Ένα ταξινομημένο σύστημα είναι ένα αποτελούμενος από τα υποσυστήματα για τα οποία οι αλληλεπιδράσεις μεταξύ των υποσυστημάτων είναι ενός διαφορετικού μεγέθους από αυτό των αλληλεπιδράσεων μέσα στα υποσυστήματα. Η αναλογία με μια ταξινόμηση είναι προφανής εάν κάποιος θεωρεί τις κατηγορίες ως υποσυστήματα. Οι σχετικές ιδιότητες σε ένα σχεδόν αποσυνθέσιμο σύστημα είναι, **(α) η βραχυπρόθεσμη συμπεριφορά που σε κάθε ένα από τα συστατικά υποσυστήματα είναι περίπου ανεξάρτητη από τη βραχυπρόθεσμη συμπεριφορά των άλλων συστατικών, και (β) η συμπεριφορά οποιοδήποτε από τα συστατικά εξαρτάται μόνο με έναν συνολικό τρόπο από τη συμπεριφορά των άλλων συστατικών.**

2. Δομές αρχείων

Στη δομή αρχείων που επιλέγονται και τον τρόπο που χρησιμοποιείται εξαρτάται η αποδοτικότητα ενός συστήματος ανάκτησης πληροφοριών. Τα αρχεία είναι μάλλον δημοφιλή στα συστήματα IR. Βεβαίως, στα συστήματα βασισμένα στις αζύγιστες λέξεις κλειδιά ειδικά όπου οι ερωτήσεις διατυπώνονται στις Boolean εκφράσεις, ένα αρχείο μπορεί να δώσει πολύ γρήγορη απάντηση. Δυστυχώς, δεν είναι δυνατό να επιτευχθεί μια αποδοτική προσαρμογή ενός αρχείου για να εξετάσει το ταίριασμα των πιο επιμελημένων περιγραφών εγγράφων και ερώτησης όπως οι σταθμισμένες λέξεις κλειδιά. Η έρευνα στις δομές αρχείων που θα μπορούσαν αποτελεσματικά να αντιμετωπίσει τις πιο περίπλοκες περιγραφές εγγράφων και ερώτησης απαιτείται ακόμα σε μεγαλύτερο βαθμό. Ο μόνος τρόπος σε αυτό μπορεί να είναι να αρχίσει με μια ταξινόμηση εγγράφων και να ερευνηθούν οι κατάλληλες δομές αρχείων. Σύμφωνα με αυτό καλό είναι να αποδειχθεί καρποφόρο η έρευνα στη σχέση μεταξύ της συγκέντρωσης εγγράφων και των συγγενικών βάσεων δεδομένων που οργανώνουν τα στοιχεία τους σύμφωνα με τις σχέσεις.

3. Στρατηγικές αναζήτησης

Μέχρι τώρα οι αρκετά απλές στρατηγικές αναζήτησης έχουν δοκιμαστεί. Στρατηγικές έχουν ποικίλει μεταξύ των απλών τμηματικών αναζητήσεων και των συστάδων . Κάθε στρατηγική που βασίζεται σε ομάδα είναι η μέθοδός της αντιπροσώπευσης συστάδων. Με την αλλαγή του αντιπροσώπου συστάδων, την απόφαση και την παύση των κανόνων των στρατηγικών αναζήτησης μπορεί συνήθως επίσης να αλλάξουν. Μια προσέγγιση που δεν φαίνεται να δοκιμάζεται θα περιελάμβανε την κατοχή διάφορων αντιπροσώπων κάθε μια συστάδας. Οι πιθανολογικές στρατηγικές αναζήτησης δεν έχουν ερευνηθεί πολύ ούτε αν και τέτοιες στρατηγικές έχουν δοκιμαστεί σε τομείς αναγνώρισης σχεδίων και αυτόματης ιατρικής διάγνωσης. Φυσικά, σε αυτούς τους τομείς οι περιγραφές αντικειμένου είναι πιο λεπτομερείς από ότι είναι οι περιγραφές εγγράφων στο IR, το οποίο μπορεί να σημαίνει ότι για αυτές τις στρατηγικές στην εργασία στο IR μπορούμε να απαιτήσουμε τις περιγραφές εγγράφων . Στο κεφάλαιο 5 αναφέραμε ότι οι από κάτω προς τα επάνω στρατηγικές αναζήτησης είναι προφανώς επιτυχέστερες από τις παραδοσιακότερες από επάνω προς τα κάτω αναζητήσεις. Αυτό οδηγεί για να σκεφτούμε ότι πολύ πιθανό η μέτρηση του δέντρου στα έγγραφα να είναι μια αποτελεσματική δομή για την καθοδήγηση μιας αναζήτησης των σχετικών εγγράφων. Μια στρατηγική αναζήτησης βασισμένη σε ένα εκτειμένος δέντρο για τα έγγραφα μπορεί να είναι σε θέση να χρησιμοποιήσει καλά τις πληροφορίες εξάρτησης που προέρχονται από το εκτειμένος δέντρο για όλους τους όρους δεικτών. Ένα ενδιαφέρον ερευνητικό πρόβλημα θα ήταν να δει κανείς εάν με την άδεια της κάποιας αλληλεπίδρασης μεταξύ των δύο δεικτών που εκτείνονται στα δέντρα θα μπορούσε να βελτιώσει την αποτελεσματικότητα ανάκτησης.

4. Προσομοίωση

Οι τρεις τομείς της έρευνας που συζητήθηκαν μέχρι τώρα θα μπορούσαν γόνιμα να εξερευνηθούν μέσω ενός προτύπου προσομοίωσης. Έχουμε τώρα αρκετά γνώση των λεπτομερειών για να επιτρέψουμε και να διευκρινίσουμε ένα λογικό πρότυπο προσομοίωσης ενός συστήματος ανάκτησης πληροφορίας. Παραδείγματος χάριν, η μορφή των διανομών των λέξεων κλειδιών σε όλη μια συλλογή εγγράφων είναι γνωστή για να επηρεάζει την αποτελεσματικότητα ανάκτησης. Με την ποικιλία αυτών των διανομών τι μπορεί να αναμείνει κανείς να συμβεί ώστε να τεκμηριώσει η να ταξινομήσει λέξεις κλειδιού; Μπορεί να είναι δυνατό να επινοηθούν οι αποδοτικότερες δομές αρχείων με τη μελέτη της απόδοσης των διάφορων δομών αρχείων μιμούμενος τις διαφορετικές διανομές λέξης κλειδιού.
Ένα σημαντικό ανοικτό πρόβλημα είναι ότι η προσομοίωση είναι σχετική.

5. Αξιολόγηση

Στη στήριξη μιας θεωρίας της αξιολόγησης στη θεωρία της μέτρησης, είναι δυνατό να επινοηθεί ένα μέτρο της αποτελεσματικότητας όχι αρχίζοντας από την ακρίβεια αλλά

με το σύνολο σχετικών εγγράφων και το σύνολο ανακτημένων εγγράφων; Σε αυτή την περίπτωση, μπορεί εμείς να γενικεύσουμε ένα τέτοιο μέτρο και να λάβουμε υπόψη ότι ο βαθμός είναι σχετικός; Μια εναλλακτική παραγωγή ενός μέτρου ε-τύπων θα μπορούσε να γίνει από την άποψη της ανάκλησης και του fallout . Υπάρχει οποιοδήποτε πλεονέκτημα να γίνει αυτό;
Μέχρι τώρα η μέτρηση της αποτελεσματικότητας έχει αποδειχθεί αρκετά ανυπάκουη στη στατιστική ανάλυση. Αυτό ήταν κυρίως επειδή κανένα λογικό στατιστικό πρότυπο δεν μπορούσε να βρεθεί Μπορεί να είναι "νόμοι" της ανάκτησης όπως η γνωστή ανταλλαγή μεταξύ της ακρίβειας και να υπάρξουν είτε εμπειρικά είτε από το θεωρητικό επιχείρημα. Έχει αποδειχθεί ότι η ανταλλαγή στην πραγματικότητα προκύπτει από περισσότερες βασικές υποθέσεις για το πρότυπο ανάκτησης. Τα παρόμοια επιχειρήματα απαιτούνται για να καθιερώσουν τα ανώτερα όρια στην ανάκτηση κάτω από ορισμένα πρότυπα .

6. Ανάλυση περιεχομένου

Υπάρχει μια ανάγκη για την εντατικότερη έρευνα στα προβλήματα αυτών που χρησιμοποιούν την αντιπροσώπευση του περιεχόμενου των εγγράφων σε έναν υπολογιστή.
Τα συστήματα ανάκτησης πληροφοριών, και λειτουργικά και πειραματικά βασίζονται στη λέξη κλειδί. Μερικοί έχουν γίνει αρκετά περίπλοκοι σε χρήση λέξεων κλειδιών τους, παραδείγματος χάριν, αυτοί μπορούν να περιλάβουν μια μορφή κανονικοποίησης και κάποιο είδος της στάθμισης. Μερικοί χρησιμοποιούν τις διανεμητικές πληροφορίες για να μετρήσουν τη δύναμη των σχέσεων μεταξύ των λέξεων κλειδιών ή μεταξύ των περιγραφών λέξης κλειδιού των εγγράφων. Το όριο της ευστροφίας μας με τις λέξεις κλειδιά φάνηκε να επιτυγχάνεται όταν καθορίστηκαν μερικές σημασιολογικές σχέσεις μεταξύ των λέξεων και χρησιμοποιήθηκαν.
Το μεγαλύτερο μέρος των πειραματικών στοιχείων κατά τη διάρκεια της τελευταίας δεκαετίας έχει δείξει την ανωτερότητα αυτής της προσέγγισης πέρα από τις πιθανές εναλλακτικές λύσεις. Εντούτοις υπάρχει περιθώριο για θεαματικότερες βελτιώσεις.

Φαίνεται ότι στη ρίζα της ανάκτησης η αποτελεσματικότητα βρίσκεται η επάρκεια (ή ανεπάρκεια) της απεικόνισης των εγγράφων στον υπολογιστή. Καμία αμφιβολία αυτό δεν αναγνωρίστηκε για να είναι αληθινή εκείνες τις ημέρες αλλά οι προσπάθειες εκείνη την περίοδο να απομακρυνθούν από την αντιπροσώπευση λέξης κλειδιού συναντήθηκαν με μικρή επιτυχία. Ο χρόνος είναι ώριμος για μια προσπάθεια στη χρησιμοποίηση της φυσικής γλώσσας για να αντιπροσωπεύσει τα έγγραφα μέσα σε έναν υπολογιστή. Υπάρχει λόγος αισιοδοξίας καθώς η φυσική γλώσσα προτιμάται περισσότερο για τη σύνταξη και τη σημασιολογία της γλώσσας. Στην τεχνητή νοημοσύνη, η εργασία έχει κατευθυνθεί προς τον προγραμματισμό ενός υπολογιστή για να καταλάβει τη φυσική γλώσσα. Επινοούνται μηχανικές διαδικασίες για τη φυσική γλώσσα επεξεργασίας (και κατανόησης). Ομοίως, στην ψυχογλωσσολογία ερευνάται ο μηχανισμός με τον οποίο ο ανθρώπινος εγκέφαλος καταλαβαίνει τη γλώσσα. Κατά γενική ομολογία ο τρόπος με τον οποίο οι εξελίξεις σε αυτούς τους τομείς μπορούν να εφαρμοστούν στην ανάκτηση πληροφορίας δεν είναι αμέσως προφανής, αλλά σαφώς είναι σχετικοί και επομένως αξίζουν την εκτίμηση. υποτίθεται

ότι ποτέ ένα σύστημα ανάκτησης δεν προσπάθησε "να καταλάβει" το περιεχόμενο ενός εγγράφου. Τα περισσότερα συστήματα ανάκτησης πληροφοριών προς το παρόν στοχεύουν μόνο σε μια βιβλιογραφική αναζήτηση. Τα έγγραφα κρίνονται για να είναι σχετικά βάσει μιας επιφανειακής περιγραφής. Δεν πρόκειται απλό θέμα να προγραμματιστεί ένας υπολογιστής για να καταλάβει τα έγγραφα. Αυτό που προτείνεται είναι ότι κάποια προσπάθεια πρέπει να γίνει χρησιμοποιώντας περισσότερο λέξεις κλειδιά και περιεχομένου κάθε εγγράφου στο σύστημα. Τα περιπλοκότερα question-answering συστήματα κάνουν κάτι πολύ παρόμοιο. Έχουν ένα *πρότυπο του κόσμου ομιλίας τους που μπορούν να απαντήσουν στις ερωτήσεις* και μπορούν να ενσωματώσουν τα νέα γεγονότα και τους κανόνες καθώς διατίθενται. Μια τέτοια προσέγγιση θα έκανε "την ανατροφοδότηση" ένα σημαντικό εργαλείο. Η ανατροφοδότηση, όπως χρησιμοποιείται αυτήν την περίοδο, είναι βασισμένη στην υπόθεση ότι ένας χρήστης θα είναι σε θέση να καθιερώσει τη σχετικότητα ενός εγγράφου βάσει των στοιχείων, όπως τον τίτλο του, της περίληψής του, ή/ και του καταλόγου όρων από τους οποίους έχει συνταχθεί. Αυτή η εργασία σε μια έκταση είναι ανεπαρκής. Εάν το περιεχόμενο του εγγράφου έγινε κατανοητό από τη μηχανή, η σχετικότητά της θα μπορούσε εύκολα να ανακαλυφθεί από το χρήστη.

Μελλοντικές εξελίξεις

Ένα μεγάλο μέρος της εργασίας στο IR πάσχει από τη δυσκολία στα αποτελέσματα ανάκτησης. Τα πειράματα έχουν γίνει σε μια μεγάλη ποικιλία συλλογών εγγράφων και σπάνια η ίδια συλλογή εγγράφων έχει χρησιμοποιηθεί αρκετά στην ίδια μορφή σε περισσότερα από ένα κομμάτι της έρευνας. Επομένως κάποιος μένει πάντα με την

υποψία ότι τα αποτελέσματα εργαζομένων Α μπορούν να είναι στοιχεία συγκεκριμένα και αυτό ήταν για να τους εξετάσει κατά την ημερομηνία εργαζομένων Β και έτσι δεν θα τους κρατούσαν.
Τα συστήματα ανάκτησης πληροφοριών είναι πιθανό να παίξουν έναν αυξανόμενο ρόλο στην κοινότητα. Είναι πιθανό να είναι σε απευθείας σύνδεση και διαλογικοί. Το υλικό αυτό για να ολοκληρωθεί θα πρέπει η καθολική εφαρμογή της να γίνει εμπορικά βιώσιμη. Μια σημαντική πρόσφατη ανάπτυξη είναι ότι οι υπολογιστές και οι βάσεις δεδομένων συνδέονται με δίκτυα. Είναι προβλέψιμο ότι τα άτομα θα έχουν πρόσβαση σε αυτά τα δίκτυα μέσω των ιδιωτικών τηλεφώνων τους και θα χρησιμοποιούν τα κανονικά τηλεοπτικά σύνολα ως συσκευές παραγωγής. Ο κύριος αντίκτυπος αυτού για τα συστήματα ανάκτησης πληροφοριών είναι ότι θα πρέπει να είναι απλοί ώστε να επικοινωνούν , τα μέσα θα πρέπει να χρησιμοποιήσουν τη συνηθισμένη γλώσσα και θα πρέπει να είναι ικανοί στη δυνατότητά τους να παρέχουν τις σχετικές πληροφορίες. Ακόμη και οι ειδικοί μπορούν καλά να επιθυμήσουν από ένα σύστημα ανάκτησης πληροφορίας περισσότερες παραπομπές από ότι αυτό ανακτά.
Για να φέρει όλο αυτό για το έγγραφο το σύστημα ανάκτησης θα πρέπει να διασυνδεθεί και να ενσωματωθεί με τα συστήματα ανάκτησης στοιχείων, για να δώσει την πρόσβαση στα σχετικά έγγραφα. Μια προφανής εφαρμογή βρίσκεται σε ένα χημικό ή ιατρικό σύστημα ανάκτησης. Υποθέτουμε ότι ένα πρόσωπο έχει ανακτήσει ένα σύνολο εγγράφων για μια συγκεκριμένη χημική ένωση και αυτή ίσως έδωσε μερικά φασματικά στοιχεία. Μπορεί να επιθυμήσει να συμβουλευθεί ένα σύστημα

ανάκτησης στοιχείων που του δίνει τις λεπτομέρειες για τις σχετικές ενώσεις. Ή μπορεί να θελήσει να κάνει on-line, για παράδειγμα, θα του παραθέτει έναν κατάλογο πιθανών ενώσεων σύμφωνων με τα φασματικά στοιχεία. Τέλως, μπορεί να επιθυμήσει να κάνει κάποια στατιστική ανάλυση των στοιχείων που περιλαμβάνονται στα έγγραφα. Για αυτό θα χρειαστεί την πρόσβαση σε ένα σύνολο στατιστικών προγραμμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. LUHN, H.P., 'The automatic creation of literature abstracts', IBM Journal of Research and Development, 2, 159-165 (1958).

2. ZIPF, H.P., Human Behaviour and the Principle of Least Effort, Addison-Wesley, Cambridge, Massachusetts (1949).

3. EDMONDSON, H.P. and WYLLYS, R.E., 'Automatic abstracting and indexing survey and recommendations', Communications of the ACM, 4, 226-234 (1961).

4. SPARCK JONES, K., 'A statistical interpretation of term specificity and its application in retrieval', Journal of Documentation, 28, 111-21 (1972).

5. SPARCK JONES, K., 'Does indexing exhaustivity matter?', Journal of the American Society for Information Science, 24, 313-316 (1973).

6. SALTON, G. and YANG, C.S., 'On the specification of term values in automatic indexing', Journal of Documentation, 29, 351-372 (1973).

7. SALTON, G., YANG, C.S. and YU, C.T., 'A theory of term importance in automatic text analysis', Journal of the American Society for Information Science, 26, 33-44 (1975).

8. SALTON, G., WONG, A. and YU, C.T., 'Automatic indexing using term discrimination and term precision measurements', Information Processing and Management, 12, 43-51 (1976).

9. SALTON, G., WONG, A. and YANG, S.S., 'A vector space model for automatic indexing', Communications of the ACM, 18, 613-620 (1975).

10. BOOKSTEIN, A. and SWANSON, D.R., 'Probabilistic models for automatic indexing', Journal of the American Society for Information Science, 25, 312-318 (1974).

11. BOOKSTEIN, A. and SWANSON, D.R., 'A decision theoretic foundation for indexing', Journal of the American Society for Information Science, 26, 45-50 (1975).

12. HARTER, S.P., 'A probabilistic approach to automatic keyword indexing, Part 1: On the distribution of speciality words in a technical literature, Part2: An algorithm for probabilistic indexing', Journal of the American Society for Information Science, 26, 197-206 and 280-289 (1975).

13. STONE, D.C. and RUBINOFF, M., 'Statistical generation of a technical

vocabulary', American Documentation, 19, 411-412 (1968).
14. DAMERAU, F.J., 'An experiment in automatic indexing', American Documentation, 16, 283-289 (1965).

15. DENNIS, S.F., 'The design and testing of a fully automatic indexing-search system for documents consisting of expository text', In: Information Retrieval: A Critical Review (Edited by G. Schecter),Thompson Book Co., Washington D.C., 67-94 (1967).

16. BOOKSTEIN, A. and KRAFT, D., 'Operations research applied to document indexing and retrieval decisions', Journal of the ACM, 24, 418-427 (1977).

17. SMITH, L.C., 'Artificial intelligence in information retrieval systems', Information Processing and Management, 12, 189-222 (1976).

18. BELKIN, N.J., 'Information concepts for information science', Journal of Documentation, 34, 55-85 (1978).

19. MINKER, J., WILSON, G.A. and ZIMMERMAN, B.H., 'An evaluation of query expansion by the addition of clustered terms for a document retrieval system', Information Storage and Retrieval, 8, 329-348 (1972).

20. KENDALL, M.G., In Multivariate Analysis (Edited by P.R. Krishnaiah), Academic Press, London and New York, 165-184 (1966)

21. JARDINE, N. and SIBSON, R., Mathematical Taxonomy, Wiley, London and New York (1971).

22. MACNAUGHTON-SMITH, P., Some Statistical and Other Numerical Techniques for Classifying Individuals, Studies in the causes of delinquency and the treatment of offenders.Report No. 6, HMSO, London (1965).

23. GOOD, I.J., Speculations Concerning Information Retrieval, Research Report PC-78, IBM Research Centre, Yorktown Heights, New York (1958).

24. FAIRTHORNE, R.A., 'The mathematics of classification'. Towards Information Retrieval, Butterworths, London, 1-10 (1961).

25. HAYES, R.M., 'Mathematical models in information retrieval'. In Natural Language and the Computer (Edited by P.L. Garvin), McGraw-Hill, New York, 287 (1963).

26. SALTON, G., Automatic Information Organization and Retrieval, McGraw-Hill, New York, 18 (1968).

27. MARON, M.E. and KUHNS, J.L., 'On relevance, probabilistic indexing and information retrieval', Journal of the ACM, 7, 216-244 (1960).

28. OSTEYEE, D.B. and GOOD, I.J., Information, Weight of Evidence, the

Singularity between Probability Measures and Signal Detection, Spring Verlag, Berlin (1974).

29. RAJSKI, C., 'A metric space of discrete probability distributions', Information and Control, 4, 371-377 (1961).

30. BECKNER, M., The Biological Way of Thought, Columbia University Press, New York, 22 (1959).

31. SPARCK JONES, K. and JACKSON, D.M., 'The use of automaticallyobtained keyword classifications for information retrieval', Information Storage and Retrieval, 5, 175-201 (1970).

32. AUGUSTSON, J.G. and MINKER, J., 'An analysis of some graph-theoretic cluster techniques', Journal of the ACM, 17, 571-588 (1970).

33. VASWANI, P.K.T. and CAMERON, J.B., The National Physical Laboratory Experiments in Statistical Word Associations and their use in Document Indexing and Retrieval, Publication 42, National Physical Laboratory, Division of Computer Science (1970).

34. Van RIJSBEGEN, C.J., 'A clustering algorithm', Computer Journal, 13, 113-115 (1970).

35. ROCCHIO, J.J., 'Document retrieval systems - optimization and evaluation', Ph.D. Thesis, Harvard University, Report ISR-10 to National Science Foundation, Harvard Computation Laboratory (1966).

36. HILL, D.R., 'A vector clustering technique', In Mechanised Information Storage, Retrieval and Dissemination, (Edited by Samuelson), North- Holland, Amsterdam (1968).

37. RIEBER, S. and MARATHE, U.P., 'The single pass clustering method', In Report ISR-16 to the National Science Foundation, Cornell University, Department of Computer Science (1969).

38. JOHNSON, D.B. and LAFUENTE, J.M., 'A controlled single pass classification algorithm with application to multilevel clustering', In Report ISR-18 to the National Science Foundation and the National Library of Medicine (1970).

39. ETZWEILER, L. and MARTIN, C., 'Binary cluster division and its application to a modified single pass clustering algorithm', In Report No. ISR-21 to the National Library of Medicine (1972).

40. Mac, J., 'Some methods for classification and analysis of multivariate observations', In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1965, University of California Press, 281-297 (1967).

41. DATTOLA, R.T., 'A fast algorithm for automatic classification', In Report ISR-14 to the National Science Foundation, Section V, Cornell University, Department of Computer Science (1968).

42. MURRAY, D.M. 'Document retrieval based on clustered files', Ph.D. Thesis, Cornell University, Report ISR-20 to the National Science Foundation and to the National Library of Medicine (1972).

43. CROUCH, D., 'A clustering algorithm for large and dynamic document collections', Ph.D. Thesis, Southern Methodist University (1972).

44. LEFKOVITZ, D., File Structures for On-line Systems, Spartan Books, New York (1969).

45. BONNER, R.E., 'On some clustering techniques', IBM Journal of Research and Development, 8, 22-32 (1964).

46. BORKO, H. and BERNICK, M., 'Automatic document classification', Journal of the ACM, 10, 151-162 (1963).

47. BAKER, F.B., 'Information retrieval based upon latest class analysis', Journal of the ACM, 9, 512-521 (1962).

48. van RIJSBERGEN, C.J., 'An algorithm for information structuring and retrieval', The Computer Journal, 14, 407-412 (1971).

49. GOWER, J.C. and ROSS, G.J.S., 'Minimum spanning trees and single- linkage cluster analysis', Applied Statistics, 18, 54-64 (1969).

50. ROHLF, J., 'Graphs implied by the Jardine-Sibson Overlapping clustering methods, Bk', Journal of the American Statistical Association, 69, 705-710 (1974).

51. BOULTON, D.M. and WALLACE, C.S., 'An information measure for single link classification', The Computer Journal, 18, 236-238 (1975).

52. CODD, E. E., 'A relational model of data for large shared data banks', Communications of the A CM, 13, 377-387 (1970)

53. MARON, M. E., 'Relational data file I: Design philosophy', In: Information Retrieval (Edited by Schecter 6), 211-223 (1967).

54. STANFEL, L. E., 'Practical aspect of doubly chained trees for retrieval', Journal of the ACM, 19, 425~36 (1972).

55. STANFEL, L. E., 'Optimal trees for a class of information retrieval problems', Information Storage and Retrieval, 9, 43-59 (1973).

56. PATT, Y. N., Minimum search tree structure for data partitioned into pages', IEEE Transations on Computers, C-21, 961-967 (1972).

57. BERGE, C., The Theory of Graphs and its Applications, Methuen, London (1966).

58. HARARY, F., NORMAN, R. Z. and CARTWRIGHT, D., Structural Models: An Introduction to the Theory of Directed Graphs, Wiley, New York (1966).

59. ORE, 0., Graphs and their Uses, Random House, New York (1963).

60. KNUTH, D. E., The Art of Computer Programming, Vol. 1, Fundamental Algorithms, Addison-Wesley, Reading, Massachusetts (1968).

61. van RIJSBERGEN, C.J., 'Further experiments with hierarchic clustering in document retrieval', Information Storage and Retrieval, **10**, 1-14 (1974).

62. MURRAY, D.M., 'Document retrieval based on clustered files', Ph.D. Thesis, Cornell University Report ISR-20 to National Science Foundation and to the National Library of Medicine (1972).

63. GOWER, J.C., 'Maximal predictive classification', Biometrics, **30**, 643-654 (1974).

64. CROFT, W.B., Organizing and Searching Large Files of Document Descriptions, Ph.D. Thesis, University of Cambridge (1979).

65. NILSSON, N.J., Learning Machines - Foundations of Trainable Classifying Systems, McGraw-Hill, New York (1965).

66. GOOD, I.J., Probability and the Weighting of Evidence, Charles Griffin and Co.Ltd., London (1950).

67. ROBERTSON, S.E., 'The probability ranking principle in IR', Journal of Documentation, 33, 294-304 (1977).

68. GOFFMAN, W., 'A searching procedure for information retrieval', Information Storage and Retrieval, 2, 294-304 (1977).

69. WILLIAMS, J.H., 'Results of classifying documents with multiple discriminant functions', In: Statistical Association Methods for Mechanized Documentation (Edited by Stevens et al.)National Bureau of Standards, Washington, 217-224 (1965).

70. DE FINETTI, B., Theory of Probability, Vol. 1, 146-161, Wiley, London(1974).

71. KULLBACK, S., Information Theory and Statistics, Dover, New York (1968).

72. CHOW, C.K. and LIU, C.N., 'Approximating discrete probability distributions

with dependence trees', IEEE Transactions on Information Theory, IT-14, 462-467 (1968).

73. KEEN, E.M., 'Evaluation parameters'. In Report ISR-13 to the National Science Foundation, Section IIK, Cornell University, Department of Computer Science (1967).

74. SWETS, J.A., 'Information retrieval systems', Science, 141, 245-250 (1963).

75. SWETS, J.A., Effectiveness of Information Retrieval Methods, Bolt, Beranekand Newman, Cambridge, Massachusetts (1967).

76. BROOKES, B.C., 'The measure of information retrieval effectiveness proposed by Swets', Journal of Documentation, 24, 41-54 (1968).

77. ROBERTSON, S.E., 'The parametric description of retrieval tests, Part 2: 'Overall measures', Journal of Documentation, 25, 93-107 (1969).

78. ROBERTSON, S.E. and TEATHER, D., 'A statistical analysis of retrieval tests: a Bayesian approach', Journal of Documentation, 30, 273-282 (1974).

79. COOPER, W.S., 'Expected search length: A single measure of retrieval effectiveness based on weakordering action of retrieval systems', Journal of the American Society for Information Science, 19, 30-41 (1968).