Πανεπιστημιο Πατρων
Πολυτεχνικη Σχολη
Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών
και Πληροφορικής

Διπλωματική Εργασία στα Πλαίσια του Μ.Δ.Ε.
«Επιστήμη και Τεχνολογία Υπολογιστών»

Τεχνικές Επαναδιατύπωσης Ερωτημάτων στον Παγκόσμιο Ιστό για Ανάκτηση Πληροφορίας Προσανατολισμένης στο Σκοπό Αναζήτησης

Νικόλαος Π. Κύρτσης
Α.Μ. 670

Επιβλέπων: Βασίλειος Μεγαλοοικονόμου, Αναπληρωτής Καθηγητής

Τριμελής Επιτροπή: Βασίλειος Μεγαλοοικονόμου, Αναπληρωτής Καθηγητής
Σοφία Στάμου, Λέκτορας
Δημήτριος Χριστοδουλάκης, Καθηγητής

Πάτρα, Φεβρουάριος 2012

Περίληψη

Οι αναζητήσεις στον Παγκόσμιο Ιστό ωθούνται από τις πληροφοριακές ανάγκες των χρηστών και στοχεύουν στην επίτευξη συγκεκριμένων εργασιών. Οι χρήστες στα πλαίσια της αναζήτησης διατυπώνουν ένα ερώτημα το οποίο εκφράζει το θέμα για το οποίο αναζητούν πληροφορίες και στη συνέχεια πραγματοποιούν μια ακολουθία ενεργειών οι οποίες καθορίζονται από την πρόθεση και τους στόχους της αναζήτησής τους. Κατά τη διαδικασία της ανάκτησης της ζητούμενης πληροφορίας ο βαθμός εμπειρίας του χρήστη επηρεάζει σε μεγάλο βαθμό την επιτυχία της αναζήτησης ή και το χρόνο που είναι απαραίτητος για την ολοκλήρωσή της. Η υποβοήθηση αναζήτησης για τους χρήστες αποτελεί ενεργό πεδίο έρευνας και έχουν προταθεί ποικίλες μέθοδοι στην βιβλιογραφία, πολλές από τις οποίες εφαρμόζονται ήδη από τα εμπορικά συστήματα ανάκτησης.

Οι περισσότεροι χρήστες δυσκολεύονται στο να διατυπώσουν με επιτυχία ένα ερώτημα, καθώς δεν έχουν πλήρη γνώση του περιβάλλοντος ανάκτησης αλλά και της συλλογής των κειμένων. Στην πραγματικότητα, όπως έχει παρατηρηθεί και με τις μηχανές αναζήτησης στον Παγκόσμιο Ιστό, οι χρήστες ενδεχομένως να χρειάζεται να ξοδέψουν πολύ χρόνο στην προσπάθεια επαναδιατύπωσης των ερωτημάτων τους ώστε να επιτύχουν αποτελεσματική ανάκτηση. Αυτή η δυσκολία υποδεικνύει ότι η πρώτη διατύπωση του ερωτήματος θα έπρεπε να αντιμετωπίζεται ως μια αρχική (απλοϊκή) προσπάθεια για την ανάκτηση σχετικής πληροφορίας. Ακολούθως, τα κείμενα που ανακτώνται πρώτα θα μπορούσαν να εξετασθούν ως προς την σχετικότητα ενώ νέες βελτιωμένες διατυπώσεις ερωτημάτων θα μπορούσαν να κατασκευασθούν με την ελπίδα της ανάκτησης επιπρόσθετων χρήσιμων κειμένων. Τέτοιες επαναδιατυπώσεις ερωτημάτων εμπεριέχουν δύο βασικά βήματα: τον εμπλουτισμό του αρχικού ερωτήματος με νέους όρους και την επαναζύγιση των όρων στο νέο ερώτημα.

Οι μηχανές αναζήτησης στον Παγκόσμιο Ιστό διαφέρουν κατά πολύ από την κλασσική αναζήτηση πληροφορίας από τα συστήματα ανάκτησης πριν την έλευση του Διαδικτύου. Η μέθοδος αλληλεπίδρασης έχει παραμείνει η ίδια (εισαγωγή ερωτήματος, ανάκτηση αποτε-

λεσμάτων, σάρωση αποτελεσμάτων, προβολή αποτελεσμάτων, επαναδιατύπωση ερωτήματος εάν χρειάζεται). Ο τρόπος ανάκτησης είναι παρόμοιος, παρόλο που πρόκειται πλέον για ένα περιβάλλον υπερμέσων. Από την άποψη του σκοπού αναζήτησης και του τύπου των πηγών πληροφόρησης ωστόσο, οι αλλαγές είναι δραματικές. Στην πραγματικότητα, οι διαφορετικοί σκοποί αναζήτησης και το εύρος των πηγών πληροφόρησης αποτελούν κλασσικά παραδείγματα του long tail φαινομένου στον Παγκόσμιο Ιστό. Ο Παγκόσμιος Ιστός, έχει δηλαδή επεκτείνει σημαντικά και το εύρος του σκοπού αναζήτησης αλλά και το εύρος των διαθέσιμων πηγών πληροφόρησης, καθώς αυτές δεν είναι απαραίτητα πλέον να είναι μόνο πληροφοριακές. Αναφερόμαστε στον τύπο της πηγής έτσι όπως γίνεται αντιληπτή από την έκφραση του χρήστη ως πρόθεση του χρήστη. Μέσα σε αυτήν την τεράστια ποικιλομορφία, οι μηχανές αναζήτησης στον Παγκόσμιο Ιστό μπορούν να βοηθήσουν τους ανθρώπους στο να βρουν τις πηγές που αναζητούν προσδιορίζοντας πιο ξεκάθαρα την πρόθεση πίσω από το ερώτημα.

Ο προσδιορισμός της υποκείμενης πρόθεσης του χρήστη αποτελεί ένα αναπτυσσόμενο πεδίο έρευνας, το οποίο έχει την δυναμική να βελτιώσει δραστικά την απόδοση του συστήματος μιας μηχανής αναζήτησης του Παγκόσμιου Ιστού, με αντίκτυπο στις περιοχές της Ανάκτησης Πληροφορίας, της Εξόρυξης Δεδομένων και του Ηλεκτρονικού Εμπορίου. Η έρευνα γύρω από την πρόθεση χρήστη διαιρείται σε τρεις υποπεριοχές: (1) εμπειρικές μελέτες και έρευνες στην χρήση της μηχανής αναζήτησης, (2) χειρωνακτική ανάλυση των αρχείων δοσοληψιών της μηχανής αναζήτησης και (3) αυτόματη κατηγοριοποίηση των αναζητήσεων στον Παγκόσμιο Ιστό.

Στα πλαίσια της παρούσας διπλωματικής εργασίας, ασχολούμαστε με την αυτόματη κατηγοριοποίηση των αποτελεσμάτων των αναζητήσεων στον Παγκόσμιο Ιστό. Αρχικά, ορίζουμε τα χαρακτηριστικά των σελίδων που είναι κατάλληλα για κατηγοριοποίηση με βάση την πρόθεση του χρήστη. Έπειτα, με χρήση μεθόδων μείωσης της διαστατικότητας επιλέγουμε τα πιο αντιπροσωπευτικά από τα χαρακτηριστικά αυτά και αξιολογούμε την απόδοση διάφορων αλγορίθμων κατηγοριοποίησης. Ακολούθως, επιλέγουμε τον αλγόριθμο κατηγοριοποίησης που βασίζεται στα επιλεγμένα χαρακτηριστικά και επιτυγχάνει την καλύτερη απόδοση. Εφαρμόζοντας τον αλγόριθμο, κατηγοριοποιούμε τα αποτελέσματα των αναζητήσεων στον Παγκόσμιο Ιστό. Τέλος, προτείνουμε μια μέθοδο εξαγωγής όρων από τα κατηγοριοποιημένα αποτελέσματα και επαναδιατύπωσης του ερωτήματος με βάση τον σκοπό αναζήτησης του χρήστη.

Abstract

Web searches are driven by informational needs and intend the accomplishment of specific tasks. When searching the web, users submit a query that expresses the topic that interests them and upon receiving the results they perform a number of tasks determined by the user intent and the goals of the search. During retrieval process the user experience level affects the web searches' successful conclusion and also has an impact on the time necessary for the searches' completion. Assisting the user in searching the web has been an active field of research and many methods has been proposed, several of which are already being implemented in commercial retrieval systems.

Without detailed knowledge of the collection make-up and of the retrieval environment, most users find it difficult to formulate queries which are well designed for retrieval purposes. In fact, as observed with Web search engines, the users might need to spend large amounts of time reformulating their queries to accomplish effective retrieval. This difficulty suggests that the first query formulation should be treated as an initial (naive) attempt to retrieve relevant information. Following that, the documents initially retrieved could be examined for relevance and new improved query formulations could then be constructed in the hope of retrieving additional useful documents. Such query reformulation involves two basic steps: expanding the original query with new terms and reweighing the terms in the expanded query.

Web search engines differ most from classic information search and pre-Web retrieval systems. The method of interaction has remained the same (i.e., enter query, retrieve results, scan results, view results, refine query as needed). The mode of retrieval is similar, albeit within a hypermedia environment. In terms of goals and type of resources, however, the changes are dramatic. In fact, the facets of goals and range of resources are classic examples of the long tail effect of the Web. Namely, the Web

has extended significantly both the range of search goals for people and the range of resources available, and these resources need not be informational. We refer to the type of resource desired in the user's expression to the system as user intent. Within this great diversity, Web search engines can better assist people in finding the resources they are looking for by more clearly identifying the intent behind the query.

Research aimed at discovering the intent of Web searchers is a growing field of Web focus. Determining the underlying intent of user searches has the potential to drastically improve system performance of Web search engine, with impact in the areas of information retrieval, data mining, and e-commerce. User intent research falls into three sub-areas, which are: (1) empirical studies and surveys of search engine use, (2) manual analysis of search engine transaction logs, and (3) automatic classification of Web searches.

In this thesis, we tackle the problem of automatic classification of search results in Web environment. First, we define web pages features that are convenient for classification based on the user's intent. Next, we use dimensionality reduction techniques to choose the most representative features and we evaluate different classification algorithms. We choose the most efficient classification algorithm based on chosen features and by using it, we classify the results retrieved from web searches. In the end, we propose a method to extract terms from the classified results and to reformulate the query based on user intent.

Ευχαριστίες

Ολοκληρώνοντας μια προσπάθεια που διήρκεσε ούτε λίγο ούτε πολύ 2,5 ολόκληρα χρόνια, δεν μένει τίποτα άλλο από το να ευχαριστήσω κάποιους ανθρώπους που βοήθησαν σε αυτή την κατεύθυνση.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα της παρούσας εργασίας, Αναπληρωτή Καθηγητή κ. Βασίλειο Μεγαλοοικονόμου για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον ερευνητικό θέμα. Η καθοδήγηση και η βοήθεια που μου πρόσφερε κατά την διάρκεια των τακτικών συναντήσεών μας ήταν καταλυτική για την ολοκλήρωση της εργασίας και την τελική μορφή που αυτή πήρε.

Έπειτα, θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της τριμελούς επιτροπής. Τον Καθηγητή κ. Δημήτριο Χριστοδουλάκη για την ευκαιρία που μου έδωσε να εκπονήσω την εργασία μου στους χώρους του Εργαστηρίου Βάσεων Δεδομένων και να γίνω μέλος της ερευνητικής του ομάδας, και την Λέκτορα Σοφία Στάμου τόσο για την ανάπτυξη της αρχικής ιδέας όσο και για το ότι μοιράστηκε μαζί μου τις ερευνητικές τις ανησυχίες σε περιοχές που δεν σχετίζονται με την παρούσα εργασία.

Η παρούσα εργασία δεν θα είχε την ίδια ποιότητα χωρίς την βοήθεια της υποψήφιας Διδάκτωρ Παρασκευής Τζέκου. Η συμβολή της διαπερνά όλη την έκταση της εργασίας, από την ανάπτυξη του εργαλείου επισημείωσης σελίδων μέχρι την τελική διόρθωση του κειμένου. Βιβή, σε ευχαριστώ που ήσουν δίπλα μου κάθε στιγμή και με συγχωρείς που έγινες ο αποδέκτης του άγχους και των νεύρων μου όλον αυτόν τον καιρό ☺.

Στη συνέχεια, θα ήθελα να ευχαριστήσω όλα τα παιδιά από το Εργαστήριο Βάσεων Δεδομένων για τον χώρο που μοιραζόμαστε τα τελευταία χρόνια καθώς και για το κλίμα συνεργασίας και αλληλοβοήθειας που έχει αναπτυχθεί μεταξύ μας.

Έχοντας παράλληλα κλείσει 7,5 χρόνια παραμονής στην πόλη της Πάτρας, δεν θα μπορούσα να μην ευχαριστήσω τους φίλους μου Φραγκίσκο, Τάκη, Μπάμπη, Βασίλη, Παναγιώτη, Σταμάτη, Βασίλη, για την στήριξή τους και τις όμορφες στιγμές που περάσαμε

μαζί όλα αυτά τα χρόνια.

Τα τελευταία 2,5 χρόνια έχω την χαρά να συγκατοικώ με τον αδερφό μου Περικλή, τον οποίο και ευχαριστώ για πολλούς και διάφορους λόγους, από το πλύσιμο των πιάτων μέχρι το σκούπισμα και το σφουγγάρισμα ☺. Γνωρίζω ότι δεν είμαι ο καλύτερος συγκάτοικος που θα μπορούσε να έχει, γι΄ αυτό και τον ευχαριστώ για την υπομονή του.

Τέλος, νιώθω την ανάγκη να ευχαριστήσω τους γονείς μου Πέτρο και Αναστασία που με στηρίζουν οικονομικά και ψυχολογικά όλα αυτά τα χρόνια. Χωρίς την δική τους στήριξη, δεν θα είχα καταφέρει όσα έχω κάνει μέχρι σήμερα. Σας ευχαριστώ για ότι έχετε κάνει για μένα.

Νίκος Κύρτσης

Πάτρα, 19 Φεβρουαρίου 2012

Αφιερώνεται …

Στην μνήμη της γιαγιάς μου Ελένης

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

"Hello, world!"

– Dennis Ritchie

Οι αναζητήσεις στον Παγκόσμιο Ιστό ωθούνται από τις πληροφοριακές ανάγκες των χρηστών και στοχεύουν στην επίτευξη συγκεκριμένων στόχων. Οι πληροφοριακές ανάγκες καθορίζονται από το θέμα των ερωτημάτων. για την ακρίβεια *τι* ψάχνουμε, ενώ οι στόχοι καθορίζονται από τα κίνητρα των χρηστών που τους οδήγησαν στην υποβολή του ερωτήματος, για την ακρίβεια, *γιατί* ψάχνουμε. Αν και υπάρχουν πολλές έρευνες που βοηθούν τους χρήστες να γράψουν ερωτήματα που εκφράζουν τις πληροφοριακές τους ανάγκες, λίγα έχουν γίνει που να βοηθούν τους χρήστες να γράψουν ερωτήματα που εκφράζουν τον σκοπό της αναζήτησής τους.

Με τον πολλαπλασιασμό του on-line περιεχομένου και των χρηστών, έχει γίνει μια μετατόπιση ενδιαφέροντος από την ανάκτηση σχετικών με το ερώτημα κειμένων στην ανάκτηση πληροφορίας η οποία σχετίζεται με τον σκοπό που ο χρήστης προσπαθεί να ικανοποιήσει μέσω της αναζήτησης. Αυτό συμβαίνει επειδή οι στόχοι των χρηστών που εμπλέκονται σε μια διαδικασία αναζήτησης επηρεάζουν την κρίση τους στην χρησιμότητα των ανακτηθέντων αποτελεσμάτων. Ωστόσο, τα περισσότερα ερωτήματα δεν εκφράζουν ρητά τον στόχο της αναζήτησης. Για τον σκοπό αυτό, οι μηχανές αναζήτησης έχουν εξοπλισθεί με υπηρεσίες υποβοήθησης των χρηστών στην υποβολή καλύτερων ερωτημάτων [KS11].

Αν και οι μηχανές αναζήτησης λειτουργούν άριστα στον εντοπισμό της πληροφορίας, εντούτοις προσφέρουν περιορισμένη δυνατότητα οργάνωσης των ιστοσελίδων (web pages). Οι χρήστες του Διαδικτύου έρχονται αντιμέτωποι με εκατομμύρια ιστοσελίδων που επιστρέφονται από μηχανές αναζήτησης με χρήση απλών λέξεων-κλειδιών (keywords). Η

αναζήτηση σε αυτές τις ιστοσελίδες τείνει από μόνη της να γίνει αδύνατη για τους χρήστες. Έτσι, έχει περισσότερο ενδιαφέρον να μπορούν οι μηχανές αναζήτησης να βοηθούν σε μια σχετική και γρήγορη επιλογή της πληροφορίας που αναζητούμε [CY05].

Μία από τις πιο υποσχόμενες προσεγγίσεις σε αυτήν την κατεύθυνση είναι η κατηγοριοποίηση ιστοσελίδων (web page classification). Η κατηγοριοποίηση διαδραματίζει έναν ρόλο ζωτικής σημασίας στην ανάκτηση της πληροφορίας (information retrieval). Στο Διαδίκτυο, η κατηγοριοποίηση του περιεχομένου των σελίδων βρίσκει εφαρμογή στην εστιασμένη προσκόμιση (focused crawling), στην κατασκευή διαδικτυακών καταλόγων (web directories), στην θεματική ανάλυση συνδέσμων (link analysis), στην προβολή συναφών διαφημίσεων (advertising) και στην ανάλυση της θεματικής δομής του Διαδικτύου. Όπως αναφέραμε και προηγουμένως, η κατηγοριοποίηση των ιστοσελίδων μπορεί επίσης να βοηθήσει στην βελτίωση της ποιότητας της αναζήτησης (web search) [QD09].

1.1 Περιγραφή Προβλήματος

Ένα κεντρικό αξίωμα της κλασσικής ανάκτησης πληροφορίας είναι ότι ο χρήστης οδηγείται από μία πληροφοριακή ανάγκη (information need). Οι Schneiderman et al. [SBC97] ορίζουν την πληροφοριακή ανάγκη ως «την αντιλαμβανόμενη ανάγκη για πληροφορία που οδηγεί κάποιον να χρησιμοποιήσει ένα σύστημα ανάκτησης πληροφορίας». Αλλά η πρόθεση (intent) πίσω από μία αναζήτηση στο Διαδίκτυο δεν είναι πάντα πληροφόρησης. Μπορεί να είναι πλοήγησης ή συνδιαλλαγής [Bro02].

Οι χρήστες δυσκολεύονται να εκφράσουν ερωτήματα που είναι αντιπροσωπευτικά της πληροφοριακής τους ανάγκης. Για τον σκοπό αυτό, οι μηχανές αναζήτησης υλοποιούν μεθόδους επαναδιατύπωσης ερωτημάτων ώστε να βοηθήσουν τους χρήστες να διατυπώσουν καλύτερα την πληροφοριακή τους ανάγκη. Μία από τις πιο διαδεδομένες τεχνικές επαναδιατύπωσης ερωτημάτων είναι και η τοπική ανάλυση, η οποία βασίζεται στην κατηγοριοποίηση των ιστοσελίδων που έχουν επιστραφεί ως απάντηση στο ερώτημα του χρήστη.

Η κατηγοριοποίηση ιστοσελίδων (web page classification ή web page categorization) είναι η διαδικασία ανάθεσης μιας ιστοσελίδας σε μία ή περισσότερες ετικέτες από προκαθορισμένες κατηγορίες. Το γενικό πρόβλημα της κατηγοριοποίησης ιστοσελίδων μπορεί να διαιρεθεί σε πιο συγκεκριμένα προβλήματα: θεματική κατηγοριοποίηση (topic classification ή subject classification), κατηγοριοποίηση λειτουργίας (style classification ή functional classification), κατηγοριοποίηση άποψης (sentiment classification), και άλλους τύπους κατηγοριοποίησης.

Ορίζουμε αρχικά το πρόβλημα κατηγοριοποίησης ιστοσελίδων με βάση τον σκοπό ανα-

ζήτησης του χρήστη. Γεννιούνται όμως έτσι ερωτήματα όπως ποια είναι τα χαρακτηριστικά που θα χρησιμοποιηθούν για την κατηγοριοποίηση, ποια είναι η απόδοση της κατηγοριοποίησης και πως μπορεί η συγκεκριμένη κατηγοριοποίηση να βοηθήσει στην διαδικασία αναζήτησης και συγκεκριμένα στην επαναδιατύπωση ερωτήματος. Τα ερωτήματα αυτά θα απαντήσουμε στην συνέχεια της παρούσας εργασίας.

1.2 Σκοπός και Συνεισφορά της Εργασίας

Στην παρούσα διπλωματική εργασία μελετάμε το πρόβλημα της επαναδιατύπωσης ερωτημάτων με βάση την πρόθεση του χρήστη. Δοθέντος ενός ερωτήματος q, στόχος μας είναι να το επαναδιατυπώσουμε σε ένα ερώτημα $q^{'}$ για κάθε έναν από τους σκοπούς αναζήτησης του χρήστη. Οι επαναδιατυπώσεις του ερωτήματος βασίζονται στην κατηγοριοποίηση και την επεξεργασία των ιστοσελίδων που επιστρέφονται ως απάντηση στο ερώτημα q από τις οποίες εξάγονται όροι αντιπροσωπευτικοί του σκοπού αναζήτησης. Στη συνέχεια, τα επαναδιατυπωμένα ερωτήματα προβάλλονται στον χρήστη με σκοπό να επιλέξει εκείνο που ταιριάζει καλύτερα στην υποκείμενη πληροφοριακή του ανάγκη. Η συνεισφορά της παρούσας εργασίας συνοψίζεται στα εξής σημεία:

- **Σκοπός αναζήτησης**: Δείχνουμε ότι είναι δυνατή η αποτύπωση του σκοπού αναζήτησης του χρήστη στο ερώτημα σε αντίθεση με τις έως τώρα προσεγγίσεις οι οποίες αποτυπώνουν στο ερώτημα την πληροφοριακή ανάγκη του χρήστη και όχι τον σκοπό αναζήτησής του.

- **Κατηγοριοποίηση ιστοσελίδων**: Δείχνουμε ότι ο σκοπός αναζήτησης που ικανοποιεί μια ιστοσελίδα σε ένα δεδομένο ερώτημα μπορεί να αναγνωριστεί με βάση χαρακτηριστικά που δεν εξάγονται από το κείμενο της ιστοσελίδας αλλά από δομικά στοιχεία της καθώς και στοιχεία που εξάγονται από το url της.

- **Επαναδιατύπωση ερωτημάτων**: Δείχνουμε ότι η επαναδιατύπωση ερωτημάτων ανάλογα με τον σκοπό αναζήτησης πρέπει να βασίζεται σε διαφορετικές πηγές πληροφορίας και συγκεκριμένα ότι, η επιλογή όρων από το κείμενο ευνοεί τις informational αναζητήσεις, η επιλογή όρων από το url βελτιώνει τις transactional αναζητήσεις ενώ οι navigational αναζητήσεις βελτιώνονται από προσθήκη συγκεκριμένων λέξεων στο ερώτημα.

1.3 Οργάνωση της Εργασίας

Το υπόλοιπο της εργασίας οργανώνεται ως εξής:

Στο Κεφάλαιο 2 επιχειρούμε μια ανασκόπηση της σχετικής βιβλιογραφίας στις περιοχές έρευνας που βασίζεται η παρούσα εργασία. Πιο συγκεκριμένα, παρουσιάζουμε την σχετική έρευνα στα πεδία της πρόθεσης ερωτήματος (query intent), της επέκταση ερωτήματος (query expansion) και της κατηγοριοποίηση ιστοσελίδων (web page classification/categorization).

Στο Κεφάλαιο 3 παρουσιάζουμε τις βασικές έννοιες που χρησιμοποιούνται στην συνέχεια της εργασίας. Αρχικά, αναφέρουμε τις διαφορετικές προσεγγίσεις που μπορεί να ακολουθήσεις κάποιος στην προσπάθεια επαναδιατύπωσης ενός ερωτήματος και επικεντρώνουμε στην μέθοδο που μας ενδιαφέρει. Έπειτα, δίνουμε έναν ορισμό της κατηγοριοποίησης ιστοσελίδων και παρουσιάζουμε τους διάφορους τύπους κατηγοριοποίησης που έχουν προταθεί στην βιβλιογραφία. Στη συνέχεια, αναφέρουμε τους τρόπους με τους οποίους μπορεί να αναπαρασταθεί μια ιστοσελίδα χρησιμοποιώντας την πληροφορία που η ίδια μεταφέρει. Ακόμη, παρουσιάζουμε μεθόδους μείωσης της διαστατικότητας του προβλήματος με σκοπό την βελτίωση της απόδοσης της κατηγοριοποίησης. Τέλος, αναφερόμαστε στην ίδια την κατηγοριοποίηση παρουσιάζοντας διάφορους τύπους αλγορίθμων κατηγοριοποίησης καθώς και έναν αλγόριθμο από κάθε κατηγορία.

Στο Κεφάλαιο 4 παρουσιάζουμε την διαδικασία κατασκευής του κατηγοριοποιητή ιστοσελίδων. Αρχικά, περιγράφουμε συνοπτικά το περιβάλλον εργασίας μας και την διαδικασία που ακολουθήσαμε για την δημιουργία του συνόλου δεδομένων μας. Έπειτα, παρουσιάζουμε τον τρόπο αναπαράστασης των ιστοσελίδων που επιλέξαμε και τα διαθέσιμα χαρακτηριστικά που αυτός υποδεικνύει. Στη συνέχεια, περιγράφουμε την διαδικασία επιλογής χαρακτηριστικών με σκοπό την μείωση της διαστατικότητας του προβλήματος κατηγοριοποίησης. Τέλος, παρουσιάζουμε την διαδικασία επιλογής αλγορίθμου κατηγοριοποίησης.

Στο Κεφάλαιο 5 παρουσιάζουμε την μέθοδο επαναδιατύπωσης ερωτημάτων. Αρχικά, παραθέτουμε τις διάφορες τακτικές επαναδιατύπωσης ανάλογα με την πρόθεση του χρήστη. Στη συνέχεια, παρουσιάζουμε μια μέθοδο αξιολόγησης του αλγορίθμου επαναδιατύπωσης καθώς και το σύνολο δεδομένων πάνω στο οποίο εφαρμόσθηκε. Τέλος, παρουσιάζουμε τα αποτελέσματα της πειραματικής αξιολόγησης του αλγορίθμου επαναδιατύπωσης.

Τέλος, στο Κεφάλαιο 6 παρουσιάζουμε τα συμπεράσματα της παρούσας εργασίας καθώς και κάποιες πιθανές μελλοντικές επεκτάσεις της.

Κεφάλαιο 2

Σχετική Έρευνα

"If we knew what we were doing it wouldn't be research."

– Albert Einstein

Στο παρόν κεφάλαιο επιχειρούμε μια ανασκόπηση της σχετικής βιβλιογραφίας στις περιοχές έρευνας που βασίζεται η παρούσα εργασία. Αρχικά παρουσιάζουμε τις περιοχές της πρόθεσης ερωτήματος (query intent) και της επέκτασης ερωτήματος (query expansion) και στη συνέχεια την περιοχή της κατηγοριοποίησης ιστοσελίδων (web page classification/categorization).

2.1 Πρόθεση Ερωτήματος

Στην κλασσική πλέον εργασία [Bro02], ο Broder προτείνει την ταξινόμηση των ερωτημάτων με βάση την πρόθεση του χρήστη. Η ταξινόμηση που προτείνεται εμπεριέχει τρεις τύπους ερωτημάτων:

- *πλοήγησης (navigational)* - όπου ο χρήστης προσπαθεί να «φτάσει» σε μια συγκεκριμένη ιστοσελίδα,
- *πληροφόρησης (informational)* - όπου ο χρήστης προσπαθεί να συλλέξει πληροφορίες από διαφορετικές ιστοσελίδες και
- *συνδιαλλαγής (transactional)* - όπου ο χρήστης προσπαθεί να διενεργήσει μια συναλλαγή μέσα από μια ιστοσελίδα.

Δύο χρόνια αργότερα, επεκτείνοντας την εργασία του Broder, οι Rose και Levinson πρότειναν μια ιεραρχία για την ταξινόμηση των ερωτημάτων σύμφωνα με τον σκοπό αναζήτησης (search goal) του χρήστη [RL04]. Η ιεραρχία αυτή διατηρεί τις κατηγορίες που πρότεινε ο Broder, με μία αλλαγή και μία σειρά από προσθήκες. Πιο συγκεκριμένα, η ετικέτα της κατηγορίας των transactional ερωτημάτων αντικαθίσταται από την ετικέτα *resource* (*πόρου*) καθώς θεωρείται ότι είναι πιο αντιπροσωπευτική των σκοπών αναζήτησης που περιγράφει. Επίσης, σε κάθε κατηγορία υψηλού επιπέδου (top level), έχουν προστεθεί μία σειρά από υποκατηγορίες οι οποίες περιγράφουν καλύτερα την αρχική κατηγορία.

Οι παραπάνω εργασίες είχαν μεγάλο αντίκτυπο στην ερευνητική κοινότητα και πυροδότησαν με την σειρά τους ένα πλήθος από νέες εργασίες. [KK03, KK04, Kan05, LLC05, BYCBGC06, JBS07, JBS08, DDLH08, SPK08, SKK09, KJHS10] Σκοπός των εργασιών αυτών ήταν πλέον η *αυτόματη* ταξινόμηση των ερωτημάτων στους τρεις τύπους που προτάθηκαν σύμφωνα πάντα με τον σκοπό αναζήτησης. Κοινός παρονομαστής των περισσοτέρων από αυτές τις εργασίες είναι η εκτεταμένη χρήση logs από εμπορικές μηχανές αναζήτησης. Τέλος, οι Brenes et al. [BGAPG09] επιχείρησαν μία σύντομη ανασκόπηση της βιβλιογραφίας αλλά ταυτόχρονα έθεσαν και ένα πλαίσιο κοινής αξιολόγησης των κυριότερων από τις μεθόδους που προτάθηκαν.

2.2 Επέκταση Ερωτήματος

Η διαδικασία της τροποποίησης του ερωτήματος συνήθως αναφέρεται στην βιβλιογραφία είτε ως *«ανατροφοδότηση σχετικότητας»* (*relevance feedback*), όταν ο χρήστης παρέχει άμεσα πληροφορία για τα σχετικά κείμενα ενός ερωτήματος, είτε ως *«επέκταση ερωτήματος»* (*query expansion*), όταν χρησιμοποιείται πληροφορία σχετική με το ερώτημα για την επέκτασή του. Στην παρούσα ενότητα, αναφερόμαστε και στις δύο περιπτώσεις ως μεθόδους ανατροφοδότησης.

Διακρίνουμε δύο βασικές προσεγγίσεις: (α) *άμεση ανατροφοδότηση* (*explicit feedback*), στην οποία η πληροφορία για την επαναδιατύπωση του ερωτήματος παρέχεται απευθείας από τους χρήστες και (β) *έμμεση ανατροφοδότηση* (*implicit feedback*), στην οποία η πληροφορία για την επαναδιατύπωση του ερωτήματος εξάγεται έμμεσα από το σύστημα.

Σε έναν κύκλο άμεσης ανατροφοδότησης σχετικότητας, η πληροφορία ανατροφοδότησης παρέχεται απευθείας από τους χρήστες ή από μια ομάδα από αξιολογητές. Όπως αρχικά προτάθηκε, οι χρήστες εξετάζουν τα κορυφαία κείμενα και υποδεικνύουν αυτά που είναι σχετικά με το ερώτημα. Τα πρώτα πειράματα με χρήση του συστήματος Smart [Sal71] αλλά και μετέπειτα πειράματα με χρήση του πιθανοτικού μοντέλου [RJ76] έδειξαν

καλές βελτιώσεις στην ακρίβεια (precision) για μικρές συλλογές όταν χρησιμοποιήθηκε η ανατροφοδότηση σχετικότητας.

Στο Διαδίκτυο, τα κλικ των χρηστών στα αποτελέσματα της αναζήτησης αποτελούν μία νέα πηγή πληροφορίας (άμεσης) ανατροφοδότησης. Ένα κλικ δεν υποδεικνύει απαραίτητα ένα κείμενο που είναι σχετικό με το ερώτημα, αλλά υποδεικνύει ένα κείμενο που είναι ενδιαφέρον για τον χρήστη στα πλαίσια του τρέχοντος ερωτήματος. Η άμεση *ανατροφοδότηση μέσω των κλικ* (*click feedback*) είναι ένα πρόσφατο πεδίο έρευνας που έχει οδηγηθεί από το συνεχές έργο του Joachims και των συνεργατών του [Joa02, JGP+05, JGP+07, RKJ08] και έχει παρακινήσει και άλλους ερευνητές να ασχοληθούν με το πεδίο [STZ05, WRJ05, ABD06].

Η έμμεση ανατροφοδότηση τώρα, διακρίνεται σε (α) *καθολική ανάλυση* (*global analysis*), όπου η πληροφορία ανατροφοδότησης αντλείται από εξωτερικές πηγές γνώσης και σε (β) *τοπική ανάλυση* (*local analysis*), όπου η πληροφορία ανατροφοδότησης αντλείται από τα κορυφαία αποτελέσματα της ανάκτησης.

Τα πρώτα αποτελέσματα της έρευνας έδωσαν την εντύπωση ότι η επέκταση ερωτήματος που βασίζεται στην *καθολική* ανάλυση είναι μία μη αποδοτική τεχνική. Ωστόσο, πιο πρόσφατα αποτελέσματα υποδεικνύουν ότι κάτι τέτοιο δεν συμβαίνει. Στην πραγματικότητα, τα αποτελέσματα που παρατηρήθηκαν από τους Vorhees [Voo86], Crouch και Yang [CY92] και Qiu και Frei [QF93] υποδεικνύουν ότι η επέκταση ερωτήματος που βασίζεται στην καθολική ανάλυση μπορεί με συνέπεια να αποφέρει βελτιωμένη απόδοση ανάκτησης. Στις μέρες μας, έχουν χρησιμοποιηθεί στην ίδια κατεύθυνση «διάσημες» εξωτερικές πηγές γνώσης όπως το WordNet[1] [KSR04, LLYM04] και η Wikipedia[2] [LLHC07, MWN07]. Τέλος, στην εργασία των Bhogal et al. [BMS07] ο αναγνώστης μπορεί να βρει μια ολοκληρωμένη επισκόπηση της βιβλιογραφίας σχετικά με επέκταση ερωτήματος βασισμένη σε οντολογίες.

Η έρευνα γύρω από την επέκταση ερωτήματος μέσω τοπικής συσταδοποίησης (local clustering) βασίζεται κυρίως στο πρώιμο έργο των Attar και Fraenkel [AF77]. Πιο πρόσφατα, παρουσιάστηκε η ιδέα της *τοπικής ανάλυσης συμφραζομένων* (*local context analysis*) από τους Xu και Croft [XC96, XC00], μια προσέγγιση που συνδυάζει καθολική και τοπική ανάλυση και βασίζεται στην χρήση ομάδων ουσιαστικών.

[1] http://wordnet.princeton.edu/
[2] http://www.wikipedia.org/

2.3 Κατηγοριοποίηση Ιστοσελίδων

Όπως αναφέραμε σύντομα και στην Εισαγωγή (Κεφάλαιο 1), η κατηγοριοποίηση των ιστοσελίδων μπορεί να διαιρεθεί σε υποπροβλήματα. Οι δύο βασικές κατευθύνσεις ωστόσο είναι (α) *θεματική κατηγοριοποίηση* (*topic/subject classification*) [KL00, Coh02, CCM+03, GA05, QD06], όπου οι ιστοσελίδες κατηγοριοποιούνται με βάση το περιεχόμενό τους και (β) *κατηγοριοποίηση λειτουργίας* (*style/functional classification*) [FÖ99, SM00, FÖ02, SLN02], όπου οι ιστοσελίδες κατηγοριοποιούνται με βάση τον τύπο του κειμένου που περιέχουν. Άλλες κατηγοριοποιήσεις μπορεί να είναι κατηγοριοποίηση ανεπιθύμητων σελίδων (spam webpage classification) [GGM05, CDG+07], κατηγοριοποίηση ύφους (genre classification) [MzES04], κοκ. Στην συνέχεια παρουσιάζουμε ορισμένες προσεγγίσεις που έχουν προταθεί στην βιβλιογραφία με σκοπό την βελτίωση της ποιότητας των αποτελεσμάτων κατά την αναζήτηση.

Οι Chekuri et al. [CGRU97] μελέτησαν την αυτόματη κατηγοριοποίηση ιστοσελίδων με σκοπό την αύξηση της ακρίβειας κατά την αναζήτηση. Ένας στατιστικός κατηγοριοποιητής, εκπαιδευμένος σε διαδικτυακούς καταλόγους, εφαρμόζεται σε καινούριες ιστοσελίδες και παράγει μια ταξινομημένη λίστα από κατηγορίες στις οποίες μπορεί να τοποθετηθεί μια νέα σελίδα. Κατά τον χρόνο εκτέλεσης του ερωτήματος ζητείται από τον χρήστη να καθορίσει μια ή περισσότερες επιθυμητές κατηγορίες έτσι ώστε να επιστρέφονται μόνο τα αποτελέσματα σε αυτές τις κατηγορίες, ή η μηχανή αναζήτησης να επιστρέφει μια λίστα από κατηγορίες στις οποίες θα ανήκουν οι νέες σελίδες. Αυτή η προσέγγιση δουλεύει όταν ο χρήστης αναζητά ένα γνωστό αντικείμενο. Σε μια τέτοια περίπτωση, δεν είναι δύσκολο να καθοριστούν οι επιθυμητές κατηγορίες. Ωστόσο, υπάρχουν περιπτώσεις στις οποίες ο χρήστης είναι λιγότερο βέβαιος για το ποια κείμενα θα ταιριάζουν, για τα οποία η παραπάνω προσέγγιση δεν βοηθάει ιδιαίτερα.

Τα αποτελέσματα της αναζήτησης συνήθως παρουσιάζονται σε μια βαθμολογημένη λίστα. Ωστόσο, τα αποτελέσματα μπορούν να είναι πιο χρήσιμα στους χρήστες αν παρουσιάζονται κατηγοριοποιημένα ή ομαδοποιημένα. Μια προσέγγιση που προτάθηκε από τους Chen et al. [CD00] κατηγοριοποιεί τα αποτελέσματα της αναζήτησης σε μια προκαθορισμένη ιεραρχική δομή και παρουσιάζει μια κατηγοριοποιημένη όψη των αποτελεσμάτων στους χρήστες. Η μελέτη χρηστών έδειξε ότι η όψη με τις κατηγορίες άρεσε περισσότερο στους χρήστες από την απλή λίστα αποτελεσμάτων και ότι είναι πιο αποδοτική για τους χρήστες στο να ανακαλύψουν την επιθυμητή πληροφορία.

Συγκρίνοντας με την προσέγγιση των Chekuri et al., η προσέγγιση των Chen et al. είναι λιγότερο αποδοτική κατά τον χρόνο εκτέλεσης του ερωτήματος επειδή κατηγοριο-

ποιεί ιστοσελίδες on-the-fly. Ωστόσο, δεν απαιτεί από τον χρήστη να ορίσει επιθυμητές κατηγορίες. Επομένως, είναι πιο χρήσιμη όταν ο χρήστης δεν γνωρίζει καλά τους όρους του ερωτήματος. Παρομοίως, ο Kaki [KÖ5] επίσης πρότεινε την παρουσίαση μιας κατηγοριοποιημένης όψης των αποτελεσμάτων της αναζήτησης στους χρήστες. Τα πειράματα έδειξαν ότι η κατηγοριοποιημένη όψη είναι πιο αποδοτική για τους χρήστες, ειδικά όταν η παραδοσιακή βαθμολόγηση των αποτελεσμάτων δεν είναι ικανοποιητική.

Οι Page et al. [PBMW98] ανέπτυξαν έναν αλγόριθμο βαθμολόγησης βασισμένο σε συνδέσμους και τον ονόμασαν PageRank. Ο PageRank υπολογίζει το «κύρος» των ιστοσελίδων βασισμένος σε ένα γράφημα που κατασκευάζεται από ιστοσελίδες και τους υπερσυνδέσμους τους χωρίς να λαμβάνει υπόψιν το θέμα κάθε σελίδας. Από τότε, έχει διεξαχθεί έρευνα για τη διαφοροποίηση του κύρους των διαφόρων θεμάτων. Ο Haveliwala [Hav03] πρότεινε τον Topic-Sensitive PageRank, ο οποίος εκτελεί πολλαπλούς PageRank υπολογισμούς, έναν για κάθε θέμα. Όταν υπολογίζεται η PageRank βαθμολογία για κάθε κατηγορία, ένας τυχαίος περιηγητής «πηδά» σε μία σελίδα της ίδιας κατηγορίας τυχαία, αντί μιας οποιαδήποτε ιστοσελίδας. Αυτό έχει σαν αποτέλεσμα την πόλωση του PageRank για αυτό το θέμα. Αυτή η προσέγγιση χρειάζεται ένα σύνολο από σελίδες που είναι κατηγοριοποιημένες με ακρίβεια. Οι Nie et al. [NDQ06] πρότειναν έναν άλλο αλγόριθμο βαθμολόγησης που λαμβάνει υπόψιν τα θέματα των σελίδων. Σε αυτή τη μέθοδο, η συνεισφορά που έχει κάθε κατηγορία στο κύρος της ιστοσελίδας διακρίνεται από τις μέσες τιμές μιας soft κατηγοριοποίησης, στην οποία για μια ιστοσελίδα που ανήκει σε κάθε κατηγορία δίνεται μια κατανομή πιθανότητας. Τέλος, προκειμένου να απαντήσουν στην ερώτηση «μέχρι ποιο βάθος θεματικών κατηγοριών ο υπολογισμός των πολωμένων page ranks έχει νόημα ;», οι Kohlschutter et al. [KCN07] διεξήγαν μια ανάλυση των ODP[3] (Open Directory Project) κατηγοριών και έδειξαν ότι η απόδοση της βαθμολόγησης αυξάνει με το επίπεδο του ODP μέχρι ένα συγκεκριμένο σημείο.

[3] http://www.dmoz.org/

Κεφάλαιο 3

Βασικές Έννοιες

"If the facts don't fit the theory, change the facts."

– Albert Einstein

Στο παρόν κεφάλαιο παρουσιάζουμε τις βασικές έννοιες που χρησιμοποιούνται στην συνέχεια της εργασίας. Αρχικά, αναφέρουμε τις διαφορετικές προσεγγίσεις που μπορεί να ακολουθήσεις κάποιος στην προσπάθεια επαναδιατύπωσης ενός ερωτήματος και επικεντρώνουμε στην μέθοδο που μας ενδιαφέρει. Έπειτα, δίνουμε έναν ορισμό της κατηγοριοποίησης ιστοσελίδων και παρουσιάζουμε τους διάφορους τύπους κατηγοριοποίησης που έχουν προταθεί στην βιβλιογραφία. Στη συνέχεια, αναφέρουμε τους τρόπους με τους οποίους μπορεί να αναπαρασταθεί μια ιστοσελίδα χρησιμοποιώντας την πληροφορία που η ίδια μεταφέρει. Ακόμη, παρουσιάζουμε μεθόδους μείωσης της διαστατικότητας του προβλήματος με σκοπό την βελτίωση της απόδοσης της κατηγοριοποίησης. Τέλος, αναφερόμαστε στην ίδια την κατηγοριοποίηση παρουσιάζοντας διάφορους τύπους αλγορίθμων κατηγοριοποίησης καθώς και έναν αλγόριθμο από κάθε κατηγορία.

3.1 Επαναδιατύπωση Ερωτημάτων

Στην παρούσα ενότητα εξετάζουμε διάφορες προσεγγίσεις για την βελτίωση της αρχικής διατύπωσης του ερωτήματος χρησιμοποιώντας πληροφορία σχετική με την πρόθεση του ερωτήματος (query intent). Με τον όρο «σχετική» εννοούμε πληροφορία που μπορεί να χρησιμοποιηθεί για την ανάκτηση κειμένων που είναι πιθανόν να θεωρηθούν σχετικά με το αρχικό ερώτημα. Η διαδικασία της τροποποίησης του ερωτήματος συνήθως αναφέρεται στην βιβλιογραφία είτε ως *«ανατροφοδότηση σχετικότητας»* (*relevance feedback*), όταν ο

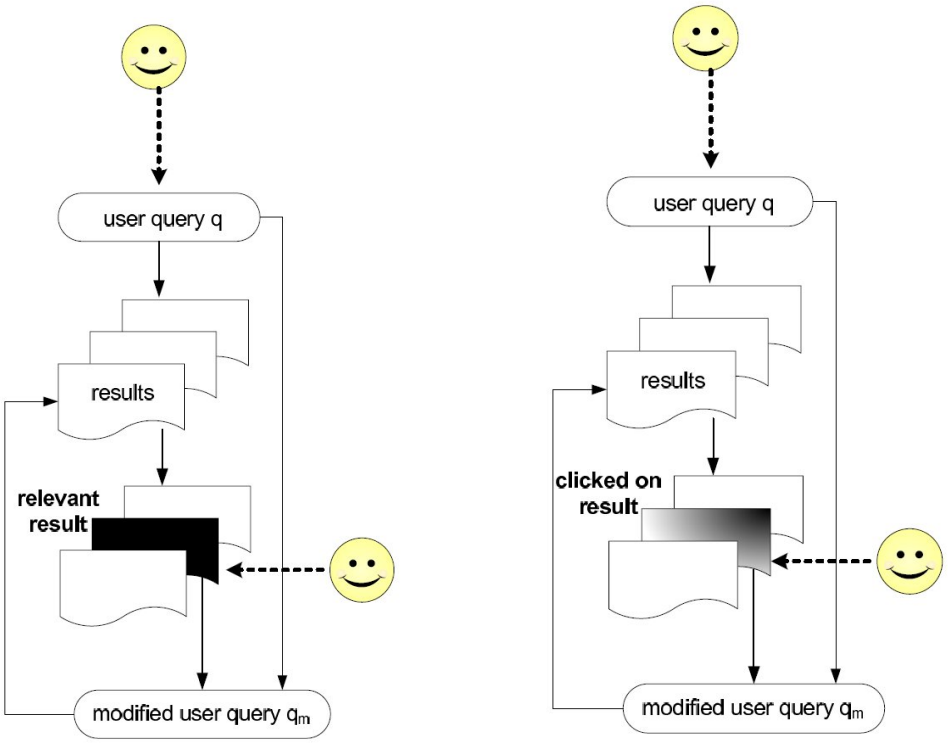
χρήστης παρέχει άμεσα πληροφορία για τα σχετικά κείμενα ενός ερωτήματος, είτε ως *«επέκταση ερωτήματος»* (*query expansion*), όταν χρησιμοποιείται πληροφορία σχετική με το ερώτημα για την επέκτασή του. Αναφερόμαστε και στις δύο περιπτώσεις ως μεθόδους ανατροφοδότησης.

Διακρίνουμε δύο βασικές προσεγγίσεις: (α) *άμεση ανατροφοδότηση* (*explicit feedback*), στην οποία η πληροφορία για την επαναδιατύπωση του ερωτήματος παρέχεται απευθείας από τους χρήστες και (β) *έμμεση ανατροφοδότηση* (*implicit feedback*), στην οποία η πληροφορία για την επαναδιατύπωση του ερωτήματος εξάγεται έμμεσα από το σύστημα. Παρουσιάζουμε στη συνέχεια μία κατηγοριοποίηση των μεθόδων ανατροφοδότησης.

Η *ανατροφοδότηση σχετικότητας* όπως αρχικά σχεδιάστηκε [Roc71], αναφέρεται σε έναν κύκλο ανατροφοδότησης στον οποίο, κείμενα που είναι γνωστό ότι είναι σχετικά με το τρέχων ερώτημα q χρησιμοποιούνται για το μετασχηματίσουν σε ένα τροποποιημένο ερώτημα q_m. Το σκεπτικό είναι ότι το ερώτημα q_m θα επιστρέψει έναν μεγαλύτερο αριθμό από κείμενα σχετικά με το q.

Ωστόσο, η λήψη πληροφορίας από κείμενα σχετικά με το τρέχων ερώτημα είναι ακριβή και απαιτεί την απευθείας παρέμβαση του χρήστη. Για να το εξηγήσουμε με ένα παράδειγμα, ενώ το σύστημα ανάκτησης θα μπορούσε να ρωτήσει τους χρήστες για το αν τα 10 πρώτα αποτελέσματα για ένα δοθέν ερώτημα είναι όντως σχετικά με αυτό, οι περισσότεροι χρήστες είναι απρόθυμοι να παρέχουν τέτοια πληροφορία, ειδικότερα δε στο Διαδίκτυο. Εξαιτίας αυτού του υψηλού κόστους, η ιδέα της ανατροφοδότησης σχετικότητας έχει «χαλαρώσει» με το πέρασμα του χρόνου ώστε να επιτρέπει την χρήση πληροφορίας που αναμένεται να είναι σχετική με το ερώτημα. Για παράδειγμα, αντί να ρωτήσουμε τους χρήστες για τα σχετικά κείμενα, μπορούμε να κοιτάξουμε τα κείμενα στα οποία έχουν κάνει κλικ ή να κοιτάξουμε τους όρους που ανήκουν στα κορυφαία κείμενα στο σύνολο αποτελεσμάτων. Και στις δύο περιπτώσεις, αν υποθέσουμε ότι η πληροφορία που συλλέγεται είναι σχετική με το αρχικό ερώτημα, αναμένουμε ότι ο κύκλος ανατροφοδότησης θα παράγει αποτελέσματα υψηλότερης ποιότητας.

Ένας κύκλος ανατροφοδότησης αποτελείται από δύο βασικά βήματα: (α) προσδιορισμό της πληροφορίας ανατροφοδότησης που είτε σχετίζεται είτε αναμένεται να σχετίζεται με το αρχικό ερώτημα q και (β) προσδιορισμό του πώς θα μετασχηματιστεί το ερώτημα q ώστε αυτή η πληροφορία να ληφθεί αποδοτικά υπόψιν. Το βήμα (α) μπορεί να επιτευχθεί με δύο ξεχωριστούς τρόπους: είτε λαμβάνοντας την πληροφορία ανατροφοδότησης άμεσα από τους χρήστες είτε λαμβάνοντας την πληροφορία ανατροφοδότησης έμμεσα από τα αποτελέσματα του ερωτήματος ή από εξωτερικές πηγές όπως ένας θησαυρός. Εφόσον ληφθεί η πληροφορία ανατροφοδότησης, το βήμα (β) μπορεί να εκτελεστεί μέσα από μία



(α΄) Ανατροφοδότηση σχετικότητας **(β΄)** Ανατροφοδότησης από κλικ

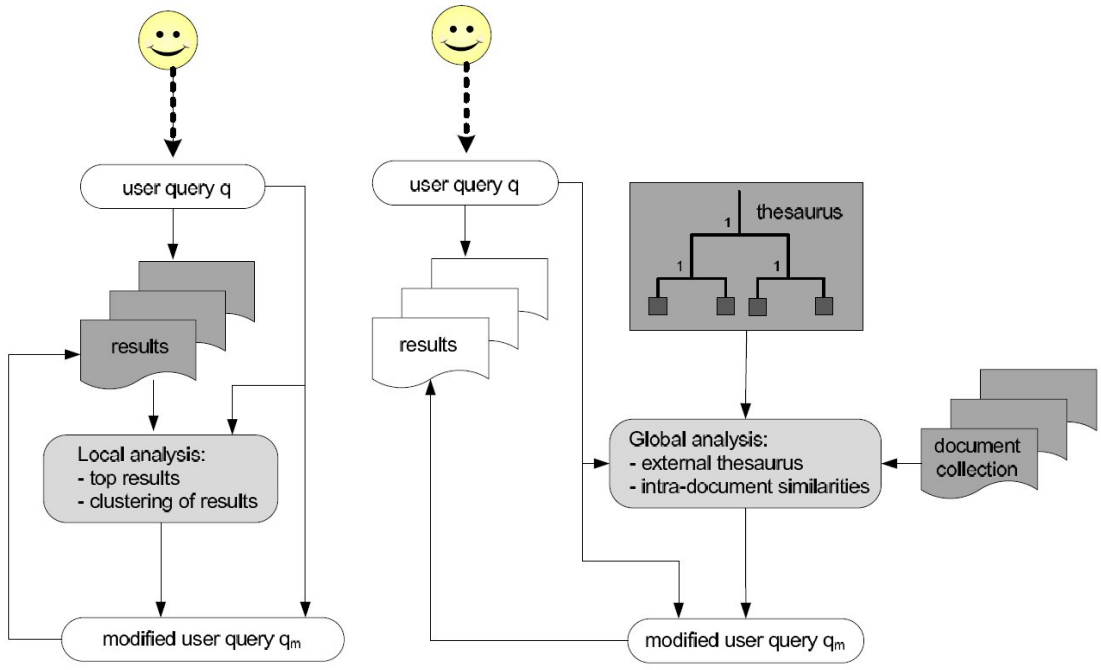
Σχήμα 3.1: Πληροφορία άμεσης ανατροφοδότησης [BYRN11].

ποικιλία μεθόδων. Η προσέγγιση που χρησιμοποιείται για την συλλογή της πληροφορίας ανατροφοδότησης, η οποία μπορεί να είναι άμεση ή έμμεση, είναι μία σημαντική διαφορά των μεθόδων ανατροφοδότησης σχετικότητας.

3.1.1 Άμεση Ανατροφοδότηση

Σε έναν *άμεσο* κύκλο ανατροφοδότησης σχετικότητας, η πληροφορία ανατροφοδότησης παρέχεται άμεσα από τους χρήστες ή από μία ομάδα από αξιολογητές. Στην αρχική της διατύπωση, οι χρήστες επιθεωρούν τα κορυφαία κείμενα και υποδεικνύουν αυτά που είναι πράγματι σχετικά με το ερώτημα. Για την ελαχιστοποίηση των παρερμηνειών, η πληροφορία ανατροφοδότησης συλλέγεται από διάφορους χρήστες και λαμβάνεται υπόψιν μόνο η πληροφορία που υποστηρίζεται από την πλειοψηφία των χρηστών. Επειδή οι χρήστες μπορεί να είναι απρόθυμοι να συνεργαστούν ή μπορεί να είναι αναξιόπιστοι στο να κρίνουν την σχετικότητα, μία εναλλακτική λύση είναι η χρήση μίας ομάδας από ειδικούς που θα κρίνουν την σχετικότητα. Σε κάθε περίπτωση, η συλλογή πληροφορίας ανατροφοδότησης είναι ακριβή και καταναλώνει χρόνο.

Στο Διαδίκτυο, τα κλικ των χρηστών στα αποτελέσματα της αναζήτησης αποτελούν



(α′) Τοπική ανάλυση **(β′)** Καθολική ανάλυση

Σχήμα 3.2: Πληροφορία έμμεσης ανατροφοδότησης [BYRN11].

μία νέα πηγή πληροφορίας ανατροφοδότησης. Ένα κλικ δεν υποδεικνύει απαραίτητα ένα κείμενο που είναι σχετικό με το ερώτημα, αλλά υποδεικνύει ένα κείμενο που είναι ενδιαφέρον για τον χρήστη στα πλαίσια του τρέχοντος ερωτήματος.

Η Εικόνα 3.1 απεικονίζει τους δύο τύπους του άμεσου κύκλου ανατροφοδότησης που διακρίνουμε: σχετικά αποτελέσματα που επιλέχθηκαν από τους χρήστες και αποτελέσματα που έγιναν κλικ από τους χρήστες. Και στις δύο περιπτώσεις, παρατηρούμε την απευθείας συμμετοχή του χρήστη στον κύκλο ανατροφοδότησης. Στην δεύτερη περίπτωση, ωστόσο, αυτή η συμμετοχή είναι πιο φυσική στους χρήστες, εφόσον δεν απαιτεί παρέκκλιση από τις τρέχουσες εργασίες τους. Επιπρόσθετα, η πληροφορία των κλικ μπορεί να συλλεχθεί σε μεγάλους όγκους χωρίς να αναστατώσει τους χρήστες, κάτι το οποίο δεν συμβαίνει με τα σχετικά αποτελέσματα.

3.1.2 Έμμεση Ανατροφοδότηση

Σε έναν *έμμεσο* κύκλο ανατροφοδότησης σχετικότητας, δεν υπάρχει συμμετοχή του χρήστη στην διαδικασία ανατροφοδότησης. Αντ' αυτού, η πληροφορία ανατροφοδότησης λαμβάνεται έμμεσα από το σύστημα. Υπάρχουν δύο βασικές προσεγγίσεις για την λήψη έμμεσης πληροφορίας ανατροφοδότησης: (α) άντληση της πληροφορίας ανατροφοδότησης από τα

κορυφαία κείμενα στο σύνολο αποτελεσμάτων, που αναφέρεται συνήθως ως *τοπική ανάλυση* (*local analysis*), ή (β) άντληση της πληροφορίας ανατροφοδότησης από εξωτερικές πηγές όπως ένας θησαυρός ή από σχέσεις όρων που έχουν εξαχθεί από μία συλλογή κειμένων, η οποία αναφέρεται συνήθως ως *καθολική ανάλυση* (*global analysis*).

Η Εικόνα 3.2 απεικονίζει τους δύο τύπους του έμμεσου κύκλου ανατροφοδότησης που διακρίνουμε: τοπική ανάλυση και καθολική ανάλυση. Και στις δύο περιπτώσεις, δεν υπάρχει άμεση συμμετοχή των χρηστών στον κύκλο ανατροφοδότησης. Προφανώς, η πληροφορία ανατροφοδότησης δεν σχετίζεται απαραίτητα με το τρέχων ερώτημα, κάτι το οποίο κάνει την χρήση της πιο δύσκολη από την πληροφορία που παρέχεται άμεσα από τους χρήστες. Παρά το γεγονός αυτό, εφόσον η έμμεση πληροφορία είναι άφθονη και μπορεί να συλλεχθεί με πολύ μικρό κόστος, υπάρχει ένα επίμονο ενδιαφέρον στην χρήση της έμμεσης πληροφορίας για την βελτίωση των αποτελεσμάτων του ερωτήματος.

Στη συνέχεια, επικεντρώνουμε το ενδιαφέρον μας στην τοπική ανάλυση.

Τοπική Ανάλυση

Σε μία στρατηγική *τοπικής* ανατροφοδότησης, τα κείμενα που ανακτώνται για ένα δοθέν ερώτημα q εξετάζονται κατά τον χρόνο υποβολής του ερωτήματος για τον προσδιορισμό των όρων που θα χρησιμοποιηθούν για την επέκταση του ερωτήματος, όπως παρουσιάζεται στην Εικόνα 3.2α′. Αυτή η διαδικασία είναι παρόμοια με τον κύκλο ανατροφοδότησης σχετικότητας αλλά γίνεται χωρίς την βοήθεια του χρήστη. Η τοπική στρατηγική που παρουσιάζεται εδώ ονομάζεται *τοπική συσταδοποίηση* (*local clustering*), βασίζεται στο πρώιμο έργο των Attar και Fraenkel [AF77] υποδεικνύει πολλές από τις θεμελιώδεις ιδέες και έννοιες σχετικά με την χρήση της συσταδοποίησης στην επέκταση ερωτημάτων.

Η υιοθέτηση τεχνικών συσταδοποίησης για επέκταση ερωτημάτων έχει γίνει η βασική προσέγγιση στην ανάκτηση πληροφορίας από τα πρώτα κιόλας χρόνια. Η καθιερωμένη διαδικασία είναι το χτίσιμο καθολικών δομών, όπως πίνακες συσχέτισης (association matrices), οι οποίοι ποσοτικοποιούν τις σχέσεις όρων και στη συνέχεια χρησιμοποιούν τους σχετιζόμενους όρους για την επέκταση του ερωτήματος. Το βασικό πρόβλημα είναι ότι οι καθολικές δομές δεν είναι πάντα αποδοτικές στην βελτίωση της ποιότητας της ανάκτησης με γενικές συλλογές. Ένας κύριος λόγος είναι ότι οι καθολικές δομές μπορεί να μην προσαρμόζονται καλά στο τοπικό πλαίσιο που ορίζεται από το τρέχων ερώτημα. Για να λυθεί αυτό το πρόβλημα, μπορεί να χρησιμοποιηθεί η *τοπική συσταδοποίηση* (*local clustering*) [AF77] όπως την παρουσιάζουμε ευθύς αμέσως.

Ορισμός 1. *Για ένα δοθέν ερώτημα q, το σύνολο D_l των κειμένων που ανακτώνται ουο-*

μάζεται τοπικό σύνολο κειμένων. Έστω N_l ο αριθμός κειμένων στο D_l. Επιπρόσθετα, το σύνολο V_l των διακριτών λέξεων στο τοπικό σύνολο κειμένων ονομάζεται τοπικό λεξιλόγιο. Η συχνότητα εμφάνισης του όρου k_i σε ένα κείμενο $d_j, d_j \in D_l$ αναφέρεται ως $f_{i,j}$. Έστω $\mathbf{M}_l = [m_{i,j}]$ το μητρώο όρων-κειμένου με V_l γραμμές και N_l στήλες όπου $m_{i,j} = f_{i,j}$. Δοθέντος ότι $\mathbf{M}_l^T$ είναι ο ανάστροφος του M_l, το μητρώο $\mathbf{C}_l = \mathbf{M}_l\mathbf{M}_l^T$ είναι ένα τοπικό μητρώο σχέσεων μεταξύ των όρων. Κάθε στοιχείο $c_{u,v} \in \mathbf{C}_l$ αναπαριστά μία σχέση μεταξύ των όρων k_u και k_v.

Το τοπικό μητρώο σχέσεων μεταξύ των όρων $\mathbf{C}_l$ αναπαριστά μία σχέση μεταξύ δύο οποιωνδήποτε όρων k_u και k_v, βασισμένο στις από κοινού συν-εμφανίσεις μέσα στα κείμενα που επιστρέφονται σαν απάντηση στο ερώτημα q. Όσο μεγαλύτερος ο αριθμός των κειμένων στα οποία συνεμφανίζονται οι δύο όροι, τόσο ισχυρότερη είναι η μεταξύ τους συσχέτιση.

Για να υπάρξει όφελος από παράγοντες όπως οι αποστάσεις μεταξύ όρων, μπορούν να οριστούν παράγοντες συσχέτισης με διαφορετικούς τρόπους. Εφόσον υπολογισθούν όλοι οι παράγοντες συσχέτισης μπορούν να χρησιμοποιηθούν για τον υπολογισμό τοπικών συστάδων από γειτονικούς όρους. Στη συνέχεια, όροι από την ίδια συστάδα μπορούν να χρησιμοποιηθούν για την επέκταση του ερωτήματος. Διάφοροι τύποι από συστάδες μπορούν να ληφθούν υπόψιν όπως *συστάδες συσχέτισης* (*association clusters*), *συστάδες απόστασης* (*metric clusters*) και *βαθμωτές συστάδες* (*scalar clusters*).

Οι συστάδες συσχέτισης βασίζονται στην συχνότητα συνεμφάνισης των όρων μέσα στα κείμενα αλλά δεν λαμβάνουν υπόψιν το πού εμφανίζονται οι όροι. Από την στιγμή που δύο όροι εμφανίζονται στην ίδια πρόταση τείνουν να συσχετίζονται περισσότερο από δύο όρους που εμφανίζονται μακρυά ο ένας από τον άλλον μέσα στο κείμενο, ίσως αξίζει τον κόπο να συνυπολογιστεί η απόσταση μεταξύ των δύο όρων στον υπολογισμό του παράγοντα συσχέτισης. Οι συστάδες απόστασης βασίζονται σε αυτήν την ιδέα.

Συστάδες Απόστασης

Ορισμός 2. *Μία συστάδα απόστασης υπολογίζεται από ένα μητρώο τοπικής συσχέτισης $\mathbf{C}_l$ επαναπροσδιορίζοντας τους παράγοντες συσχέτισης $c_{u,v}$ μεταξύ οποιουδήποτε ζεύγους όρων k_u και k_v ως συνάρτηση των αποστάσεών τους στα κείμενα της συλλογής. Έστω $k_u(n, j)$ μια συνάρτηση που επιστρέφει την n-οστή εμφάνιση του όρου k_u στο κείμενο d_j. Επιπρόσθετα, έστω $r(k_u(n, j), k_v(n, j))$ μια συνάρτηση που υπολογίζει την απόσταση μεταξύ την n-οστής εμφάνισης του όρου k_u και της m-οστής εμφάνισης του όρου k_v στο κείμενο d_j. Αυτή η απόσταση μπορεί να υπολογιστεί, για παράδειγμα, ως ο αριθμός των λέξεων μεταξύ των*

εμφανίσεων των όρων. Ορίζουμε,

$$c_{u,v} = \sum_{d_j \in D_l} \sum_{n} \sum_{m} \frac{1}{r(k_u(n,j), k_v(m,j))} \tag{3.1}$$

Σε αυτή τη περίπτωση το μητρώο συσχέτισης αναφέρεται ως μητρώο τοπικής απόστασης. Παρατηρούμε ότι αν οι k_u και k_v βρίσκονται σε διαφορετικά κείμενα, θεωρούμε την μεταξύ τους απόσταση άπειρη. Παραλλαγές της παραπάνω έκφρασης για τους $c_{u,v}$ έχουν προταθεί στη βιβλιογραφία, όπως $1/r^2(k_u(n,j), k_v(m,j))$.

Ο παράγοντας συσχέτισης απόστασης $c_{u,v}$ ποσοτικοποιεί τις απόλυτες αντίστροφες αποστάσεις και είναι μη κανονικοποιημένος. Έτσι, το μητρώο τοπικής απόστασης $\mathbf{C}_l$, είναι μη κανονικοποιημένο. Μία παραλλαγή είναι να κανονικοποιηθεί ο παράγοντας συσχέτισης. Για παράδειγμα,

$$c'_{u,v} = \frac{c_{u,v}}{\text{συνολικός αριθμός από } [k_u, k_v] \text{ ζευγάρια}} \tag{3.2}$$

Σε αυτή τη περίπτωση το μητρώο τοπικής απόστασης $\mathbf{C}_l$ είναι κανονικοποιημένο.

Δοθέντος ενός μητρώου τοπικής απόστασης $\mathbf{C}_l$, μπορούμε να το χρησιμοποιήσουμε για να κατασκευάσουμε συστάδες τοπικής απόστασης ως εξής.

Ορισμός 3. *Έστω $C_u(n)$ μια συνάρτηση που επιστρέφει τις n μεγαλύτερες $c_{u,v}$ τιμές σε ένα μητρώο τοπικής απόστασης $\mathbf{C}_l$, $v \neq u$. Τότε η $C_u(n)$ ορίζει μια συστάδα τοπικής απόστασης γύρω από τον όρο k_u. Αν ο $c_{u,v}$ δίνεται από την Εξίσωση (3.1), η συστάδα απόστασης είναι μη κανονικοποιημένη. Αν ο $c_{u,v}$ δίνεται από τη Εξίσωση (3.2), η συστάδα απόστασης είναι κανονικοποιημένη.*

Επέκταση Ερωτήματος με Γειτονικούς Όρους

Όροι που ανήκουν σε συστάδες που σχετίζονται με τους όρους του ερωτήματος μπορούν να χρησιμοποιηθούν για την επέκταση του αρχικού ερωτήματος. Τέτοιοι όροι αποκαλούνται «γείτονες» των όρων του ερωτήματος και χαρακτηρίζονται ως εξής.

Ένας όρος k_v που ανήκει σε μία συστάδα $C_u(n)$ και σχετίζεται με έναν άλλο όρο k_u λέγεται ότι είναι *γείτονας* του k_u. Κάποιες φορές, ο όρος k_v αποκαλείται *searchonym*[1] του k_u, αλλά εδώ υιοθετούμε την ορολογία «γείτονας». Ενώ υποστηρίζεται ότι γειτονικοί όροι έχουν σχέση συνωνυμίας, εντούτοις δεν είναι απαραίτητα συνώνυμοι με την γραμματική έννοια. Συχνά, γειτονικοί όροι αναπαριστούν διακριτές λέξεις-κλειδιά που σχετίζονται με το πλαίσιο του τρέχοντος ερωτήματος. Η τοπική πτυχή αυτής της σχέσης αντανακλάται στο

[1] Ο όρος δεν μεταφράζεται στην Ελληνική αλλά η ερμηνεία που του αποδίδεται είναι πως ο όρος k_v είναι ο αντίστοιχος του όρου k_u κατά την διαδικασία της αναζήτησης.

γεγονός ότι τα κείμενα και οι όροι που συμπεριλαμβάνονται στο μητρώο συσχέτισης είναι όλοι τοπικοί. Υπό την ευρεία της έννοια, οι γειτονικοί όροι είναι ένα σημαντικό παράγωγο της τοπικής συσταδοποίησης, από την στιγμή που μπορούν να χρησιμοποιηθούν για την επέκταση του ερωτήματος σε μια υποσχόμενη κατεύθυνση που αποσκοπούσε ο χρήστης αλλά δεν εκφράστηκε ρητά κατά την διατύπωση του ερωτήματος.

Έστω τώρα το πρόβλημα επέκτασης ενός δοθέντος ερωτήματος q με γειτονικούς όρους. Μία δυνατότητα είναι να επεκταθεί το ερώτημα ως εξής. Για κάθε όρο $k_u \in q$, επέλεξε m γειτονικούς όρους από την συστάδα $C_u(n)$ (που μπορεί να είναι τύπου συσχέτισης, απόστασης ή βαθμωτή) και πρόσθεσέ τους στο ερώτημα. Αυτό μπορεί να εκφραστεί ως εξής:

$$q_m = q \cup \{k_v | k_v \in C_u(n), k_u \in q\}$$

Ελπίζουμε πως οι επιπρόσθετοι γειτονικοί όροι k_v θα ανακτήσουν νέα σχετικά κείμενα. Για να καλύψουμε μια ευρύτερη γειτονιά, το σύνολο $C_u(n)$ μπορεί να συντίθεται από όρους που παρατηρήθηκαν χρησιμοποιώντας παράγοντες συσχέτισης κανονικοποιημένους και μη. Η ποιοτική ερμηνεία είναι ότι μια μη κανονικοποιημένη συστάδα τείνει να ομαδοποιεί όρους που σχετίζονται εξαιτίας των μεγάλων συχνοτήτων τους, ενώ μια κανονικοποιημένη συστάδα τείνει να ομαδοποιεί όρους που είναι πιο σπάνιοι. Έτσι, η ένωση των δύο συστάδων μπορεί να παρέχει μια καλύτερη αναπαράσταση των πιθανών συσχετίσεων.

Η επέκταση ερωτήματος είναι σημαντική επειδή τείνει να βελτιώνει την *ανάκληση* (*recall*), εφόσον ανακτάται ένας μεγαλύτερος αριθμός από κείμενα. Ωστόσο, ο μεγαλύτερος αριθμός των κειμένων που πρέπει να καταταχθούν επίσης τείνει να χαμηλώνει την *ακρίβεια* (*precision*), πιο συγκεκριμένα, καινούρια ανακτηθέντα κείμενα τα οποία δεν είναι σχετικά μπορεί να καταλάβουν υψηλότερη κατάταξη. Έτσι, η επέκταση ερωτήματος χρειάζεται να ασκηθεί με μεγάλη προσοχή και να τελειοποιηθεί για την συλλογή με το χέρι.

3.2 Κατηγοριοποίηση Ιστοσελίδων

Όπως αναφέραμε και στην Εισαγωγή (Κεφάλαιο 1), η κατηγοριοποίηση ιστοσελίδων (web page classification ή web page categorization) είναι η διαδικασία ανάθεσης μιας ιστοσελίδας σε μία ή περισσότερες ετικέτες από προκαθορισμένες κατηγορίες. Φορμαλιστικά, έχουμε τον παρακάτω ορισμό:

Ορισμός 4. *Έστω $C = \{c_1, c_2, ..., c_k\}$ ένα σύνολο από προκαθορισμένες κατηγορίες, $D = \{d_1, d_2, ..., d_N\}$ ένα σύνολο από ιστοσελίδες προς κατηγοριοποίηση, και $A = D \times C$ ένα μητρώο απόφασης (decision matrix) όπου, κάθε στοιχείο $a_{ij} (1 \leq i \leq N, 1 \leq j \leq K)$*

αναπαριστά το πότε μια ιστοσελίδα d_i ανήκει σε μια κατηγορία c_j ή όχι. Για κάθε $a_{ij} \in \{0, 1\}$ η μονάδα υποδεικνύει ότι η ιστοσελίδα d_i ανήκει στην κατηγορία c_j και το μηδέν ότι δεν ανήκει. Η διαδικασία της κατηγοριοποίησης ιστοσελίδων είναι η προσέγγιση της άγνωστης συνάρτησης $f : D \times C \rightarrow \{0, 1\}$ με την βοήθεια μιας συνάρτησης $f^{'} D \times C \rightarrow \{0, 1\}$, που καλείται κατηγοριοποιητής, μοντέλο ή υπόθεση, έτσι ώστε η $f^{'}$ να συμπίπτει με την f όσο το δυνατόν περισσότερο [Seb02].

Ανάλογα με τον αριθμό των κατηγοριών του προβλήματος, η κατηγοριοποίηση μπορεί να διαιρεθεί σε *δυαδική* (*binary*) κατηγοριοποίηση και κατηγοριοποίηση *πολλαπλών κατηγοριών* (*multiclass*). Στην δυαδική κατηγοριοποίηση τα στιγμιότυπα μιας κατηγορίας κατηγοριοποιούνται σε ακριβώς μία από δύο κατηγορίες (Εικόνα 3.3α'), ενώ η κατηγοριοποίηση πολλαπλών κατηγοριών έχει να κάνει με περισσότερες από δύο κατηγορίες.

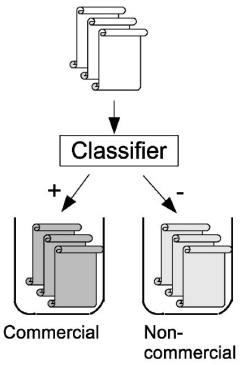
Ανάλογα με τον αριθμό των κατηγοριών που μπορούν να ανατεθούν σε ένα στιγμιότυπο, η κατηγοριοποίηση μπορεί να διαιρεθεί σε κατηγοριοποίηση *μοναδικής ετικέτας* (*single label*) και κατηγοριοποίηση *πολλαπλών ετικετών* (*multilabel*). Στην κατηγοριοποίηση μοναδικής ετικέτας, μία και μόνο μία ετικέτα κατηγορίας μπορεί να ανατεθεί σε ένα στιγμιότυπο (Εικόνα 3.3β'), ενώ στην κατηγοριοποίηση πολλαπλών ετικετών, μπορούν να ανατεθούν περισσότερες από μία κατηγορίες σε ένα στιγμιότυπο (Εικόνα 3.3γ').

Ανάλογα με τον τύπο της ανάθεσης της κατηγορίας, η κατηγοριοποίηση μπορεί να διαιρεθεί σε *hard* κατηγοριοποίηση και *soft* κατηγοριοποίηση. Στην hard κατηγοριοποίηση, ένα στιγμιότυπο μπορεί να ανήκει ή να μην ανήκει σε κάποια κατηγορία, χωρίς ενδιάμεση κατάσταση, ενώ στην soft κατηγοριοποίηση, ένα στιγμιότυπο μπορεί να προβλεφθεί να ανήκει σε κάποια κατηγορία με συγκεκριμένη πιθανότητα (συνήθως μια κατανομή πιθανότητας σε όλες τις κατηγορίες, όπως στην Εικόνα 3.3δ').

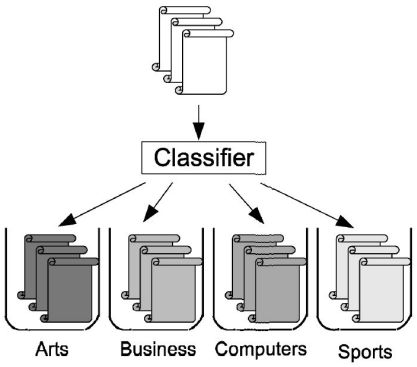
Ανάλογα με την οργάνωση των κατηγοριών, η κατηγοριοποίηση ιστοσελίδων μπορεί επίσης να διαιρεθεί σε *επίπεδη* (*flat*) και σε *ιεραρχική* (*hierarchical*) κατηγοριοποίηση. Στην επίπεδη κατηγοριοποίηση, οι κατηγορίες θεωρούνται παράλληλες, δηλαδή, μία κατηγορία δεν υπερισχύει μιας άλλης, ενώ στην ιεραρχική κατηγοριοποίηση, οι κατηγορίες οργανώνονται σε μια ιεραρχική δομή δέντρου, στην οποία κάθε κατηγορία μπορεί να έχει έναν αριθμό από υποκατηγορίες. Κάτι τέτοιο παρουσιάζεται στην Εικόνα 3.4.

3.3 Αναπαράσταση Ιστοσελίδων

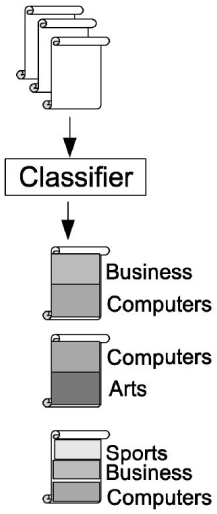
Το πρώτο βήμα στην κατηγοριοποίηση ιστοσελίδων είναι η αναπαράσταση της ιστοσελίδας, η οποία συνήθως αποτελείται από αλφαριθμητικά, υπερσυνδέσμους, εικόνες και ετικέτες



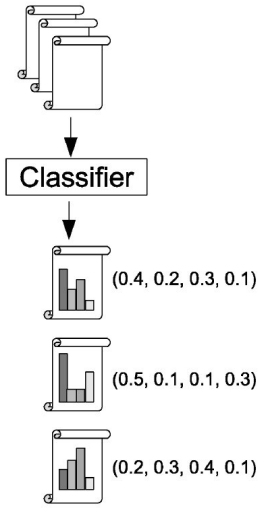
(α′) Δυαδική κατηγοριοποίηση



(β′) Multiclass, single-label, hard κατηγοριοποίηση



(γ′) Multiclass, multi-label, hard κατηγοριοποίηση

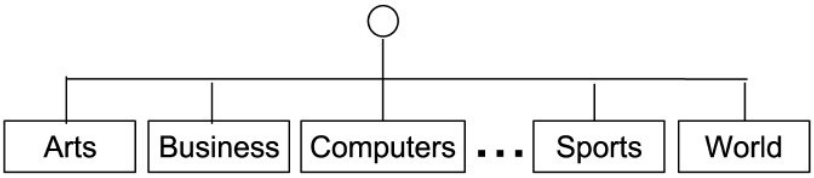


(δ′) Multiclass, soft κατηγοριοποίηση

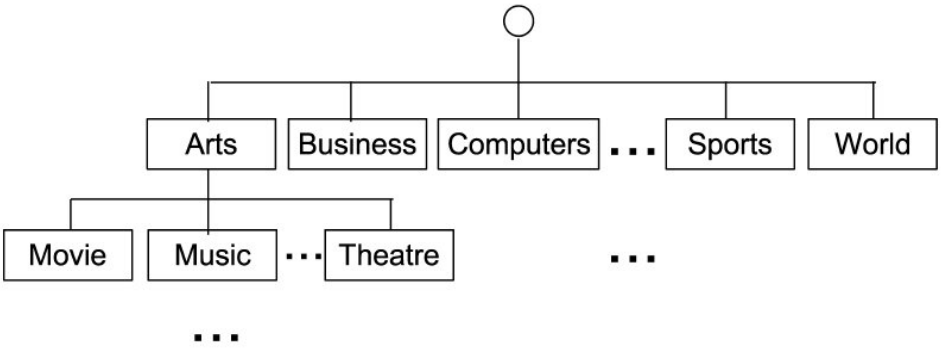
Σχήμα 3.3: Τύποι κατηγοριοποίησης [QD09].

HTML σε ένα διάνυσμα χαρακτηριστικών. Αυτή η διαδικασία χρησιμοποιείται για την απομάκρυνση της λιγότερο σημαντικής πληροφορία και την εξαγωγή των κυριότερων χαρακτηριστικών από τις ιστοσελίδες. Προφανώς, ανάλογα με τον τύπο της κατηγοριοποίησης, προτιμώνται διαφορετικά χαρακτηριστικά. Για παράδειγμα, η θεματική κατηγοριοποίηση προτιμά χαρακτηριστικά που αναπαριστούν το περιεχόμενο της ιστοσελίδας ενώ η κατηγοριοποίηση λειτουργίας προτιμά δομικά χαρακτηριστικά που φανερώνουν τον τύπο της ιστοσελίδας.

Το περιεχόμενο μιας ιστοσελίδας αναπαρίσταται κυρίως από το κείμενό της, για παράδειγμα, λέξεις, φράσεις, προτάσεις, κοκ. Για να ανακτηθούν σημαντικά χαρακτηριστικά του κειμένου, πρέπει πρώτα να προεπεξεργαστούν οι ιστοσελίδες για να απορριφθούν τα



(α′) Επίπεδη κατηγοριοποίηση



(β′) Ιεραρχική κατηγοριοποίηση

Σχήμα 3.4: Επίπεδη και ιεραρχική κατηγοριοποίηση [QD09].

λιγότερο σημαντικά δεδομένα. Η προεπεξεργασία αποτελείται από τα παρακάτω βήματα:

- *Απομάκρυνση HTML ετικετών*: Οι HTML ετικέτες υποδεικνύουν την μορφή μιας ιστοσελίδας. Για παράδειγμα, το περιεχόμενο μεταξύ των $< title >$ και $< /title >$ ζεύγους ετικετών είναι ο τίτλος μιας ιστοσελίδας ενώ το περιεχόμενο που περικλείεται μεταξύ των $< table >$ και $< /table >$ ζεύγους ετικετών είναι ένας πίνακας. Αυτές οι HTML ετικέτες μπορεί να υποδεικνύουν την σημαντικότητα των περιεχομένων τους και για αυτό μπορούν να βοηθήσουν στον καθορισμό βαρών για το περιεχόμενό τους. Οι ίδιες οι ετικέτες απομακρύνονται ύστερα από την διαδικασία καθορισμού βαρών.

- *Απομάκρυνση stop words*: Τα stop words είναι συχνές λέξεις που μεταφέρουν λίγη πληροφορία, όπως προθέσεις, αντωνυμίες και σύνδεσμοι. Απομακρύνονται από το κείμενο συγκρίνοντας το κείμενο εισόδου με μία λίστα από stop words.

- *Αποκατάληξη λέξεων*: Η διαδικασία επιτυγχάνεται με την ομαδοποίηση λέξεων που έχουν την ίδια ρίζα, όπως «υπολογιστής», «υπολογίζω» και «υπολογισμός». Ο αλγόριθμος αποκατάληξης του Porter είναι ένας γνωστός αλγόριθμος για την διαδικασία αυτή.

3.3.1 Αναπαράσταση Περιεχομένου

Στην πιο απλή της μορφή, μια ιστοσελίδα αναπαρίσταται από ένα διάνυσμα από M σταθμισμένες λέξεις. Αυτή η αναπαράσταση συνήθως αναφέρεται ως bag-of-words. Η βασική υπόθεση πίσω από αυτήν την αναπαράσταση είναι ότι κάθε λέξη στο κείμενο υποδεικνύει μια συγκεκριμένη έννοια του κειμένου. Φορμαλιστικά, έχουμε τον παρακάτω ορισμό:

Ορισμός 5. *Μια ιστοσελίδα αναπαρίσταται από ένα διάνυσμα d_i με τις λέξεις $t_1, t_2, ..., t_M$ ως χαρακτηριστικά, κάθε μια από τις οποίες σχετίζεται με ένα βάρος w_{ij}. Έτσι είναι*

$$d_i = (w_{i1}, w_{i2}, ..., w_{iM}) \tag{3.3}$$

όπου M είναι ο αριθμός των λέξεων και $w_{ij}(1 \leq j \leq M)$ είναι η σημαντικότητα (ή το βάρος) της λέξης t_j στην ιστοσελίδα d_i.

Από την στιγμή που μια φράση συνήθως περιέχει περισσότερη πληροφορία από ότι μια μοναδική λέξη, η bag-of-words αναπαράσταση μπορεί να εμπλουτιστεί προσθέτοντας νέα χαρακτηριστικά από ακολουθίες λέξεων που είναι γνωστές ως n-grams. Χρησιμοποιώντας τα n-grams μπορούμε να αναγνωρίσουμε κάποιους χαρακτηριστικούς συνδυασμούς λέξεων. Στο υπόλοιπο του κειμένου, αναφερόμαστε σε αυτήν την εμπλουτισμένη bag-of-words αναπαράσταση ως bag-of-terms, όπου ένας όρος μπορεί να είναι μια μοναδική λέξη ή οποιοδήποτε n-gram.

Ο απλούστερος τρόπος να ορίσουμε το βάρος w_j του όρου t_j μέσα σε μια ιστοσελίδα είναι να ληφθεί υπόψιν η εμφάνιση του όρου ως μια δυαδική τιμή ως εξής:

$$w_j = \begin{cases} 1 & \text{αν ο όρος } t_j \text{ βρίσκεται μέσα στην σελίδα} \\ 0 & \text{αλλιώς} \end{cases} \tag{3.4}$$

Μία από τους πιο επιτυχημένες μεθόδους ανάθεσης βάρους σε όρους είναι η TFIDF (Term Frequency Inverse Document Frequency) [SB88], η οποία προκύπτει από το γινόμενο της «τοπικής» σημαντικότητας του όρου (TF) και της «καθολικής» σημαντικότητας του όρου (IDF):

$$w_{ij} = TF(t_j, d_i \cdot IDF(t_j)) \tag{3.5}$$

όπου η συχνότητα όρου $TF(t_j, d_i)$ είναι ο αριθμός των εμφανίσεων του όρου t_j στο κείμενο d_i, και η συχνότητα κειμένου $DF(t_j)$ είναι ο αριθμός των κειμένων που εμφανίζεται ο όρος

t_j τουλάχιστον μία φορά:

$$DF(t_j) = \sum_{i=1}^{N} \begin{cases} 1 & \text{αν το } d_i \text{ περιέχει τον } t_j \\ 0 & \text{αλλιώς} \end{cases} \tag{3.6}$$

Η αντίστροφη συχνότητα κειμένου $IDF(t_j)$ μπορεί να υπολογιστεί από την συχνότητα κειμένου $DF(t_j)$:

$$IDF(t_j) = log\left(\frac{N}{DF(t_j)}\right) \tag{3.7}$$

όπου N είναι ο συνολικός αριθμός των κειμένων. Διαισθητικά, η αντίστροφη συχνότητα κειμένου ενός όρου είναι μικρή αν εμφανίζεται σε πολλά κείμενα και είναι μέγιστη όταν ο όρος εμφανίζεται μόνο σε ένα κείμενο. Επιπρόσθετα, το βάρος w_{ij} του όρου t_j σε ένα κείμενο d_i μπορεί να κανονικοποιηθεί με το μέγεθος του κειμένου, για παράδειγμα με το Ευκλείδειο μήκος (L2-νόρμα) του κειμένου:

$$w_{ij} = \frac{TF(t_j, d_i)IDF(t_j)}{\sqrt{\sum_{j=1}^{M}(TF(t_j, d_i)IDF(t_j))^2}} \tag{3.8}$$

όπου M είναι ο αριθμός των χαρακτηριστικών (μοναδικών όρων) σε όλες τις ιστοσελίδες εκπαίδευσης.

3.3.2 Αναπαράσταση Δομής

Η αναπαράσταση bag-of-terms δεν εκμεταλλεύεται την δομική πληροφορία του Διαδικτύου. Υπάρχουν τουλάχιστον δύο διαφορετικά είδη δομικής πληροφορίας που μπορούν να χρησιμοποιηθούν για να ενισχυθεί η απόδοση της κατηγοριοποίησης:

- Η δομή της HTML αναπαράστασης η οποία επιτρέπει τον εύκολο προσδιορισμό των σημαντικών σημείων μιας ιστοσελίδας όπως ο τίτλος και οι επικεφαλίδες.
- Η δομή του Διαδικτύου, όπου οι ιστοσελίδες συνδέονται η μια με την άλλη μέσω υπερσυνδέσμων.

Στις επόμενες παραγράφους παρουσιάζουμε τις δύο μεθόδους αναπαράστασης της δομής των ιστοσελίδων, που μόλις αναφέραμε.

HTML Δομή

Για την βελτίωση της αναπαράστασης των ιστοσελίδων, η εκμετάλλευση της HTML δομής θα μας βοηθήσει να προσδιορίσουμε που μπορεί να βρεθούν οι πιο αντιπροσωπευτικοί

όροι. Για παράδειγμα, μπορούμε να θεωρήσουμε ότι ένας όρος που εμπεριέχεται μέσα στις $< title >$ και $< /title >$ ετικέτες είναι γενικά πιο αντιπροσωπευτικός για το θέμα μιας ιστοσελίδας από ότι ένας όρος που εμπεριέχεται μέσα στις $< body >$ και $< /body >$ ετικέτες. Για παράδειγμα, κάποιες από αυτές τις ετικέτες μπορεί να είναι:

- **BODY**: Το κυρίως σώμα μιας ιστοσελίδας.
- **TITLE**: Ο τίτλος μιας ιστοσελίδας.
- **H1**-**H6**: Οι επικεφαλίδες των τμημάτων του κειμένου μιας ιστοσελίδας.
- **EM**: Περιεχόμενο που έχει τονισθεί.
- **URL**: Υπερσύνδεσμοι οι οποίοι μπορεί να περιέχουν περιγραφή για τις ιστοσελίδες που υποδεικνύουν. Αν μπορούν να εξαχθούν λέξεις κλειδιά από ένα URL, τότε αυτές οι λέξεις είναι ιδιαίτερα σημαντικές.
- **META**: Η meta περιγραφή μιας ιστοσελίδας. Η περιγραφή αυτή είναι αόρατη στους χρήστες που βλέπουν την ιστοσελίδα, αλλά μπορεί να παρέχει περιγραφή, λέξεις κλειδιά και ημερομηνία για την ιστοσελίδα.

Δομή Συνδέσμων

Σε αντίθεση με τις τυπικές βάσεις δεδομένων κειμένου, οι ιστοσελίδες μπορεί να μην περιέχουν προφανή κειμενικά χαρακτηριστικά σε σχέση με το θέμα τους. Για παράδειγμα, η αρχική σελίδα της `Google.com` δεν αναφέρει ρητά ότι πρόκειται για μηχανή αναζήτησης. Επίσης, κάποιες σελίδες είναι πολύ μικρές και παρέχουν ελάχιστη κειμενική πληροφορία ενώ άλλες μπορεί να μην βασίζονται σε κείμενο αλλά σε εικόνες, βίντεο, ήχο ή τεχνολογία flash. Εκτός από την ίδια την ανάλυση της ιστοσελίδας, ένας εφικτός τρόπος αναπαράστασης της ιστοσελίδας είναι η χρήση υπερσυνδέσμων που υποδεικνύουν άλλες σχετικές ιστοσελίδες. Η βασική υπόθεση που κάνει η ανάλυση των συνδέσμων είναι ότι ένας σύνδεσμος συχνά δημιουργείται εξαιτίας μιας υποκείμενης σύνδεσης μεταξύ της αρχικής ιστοσελίδας και της σελίδας που υποδεικνύεται.

Οι υπερσύνδεσμοι μιας ιστοσελίδας μπορούν να αναλυθούν έτσι ώστε να εξαχθεί επιπρόσθετη πληροφορία για τις ιστοσελίδες. Ένας υπερσύνδεσμος σε μια ιστοσελίδα A που υποδεικνύει μια ιστοσελίδα B ονομάζεται *in-link* (*incoming link - εισερχόμενος σύνδεσμος*) της ιστοσελίδας B. Ταυτόχρονα, ο ίδιος υπερσύνδεσμος είναι επίσης *out-link* (*ουτγοινγ λινκ - εξερχόμενος σύνδεσμος*) της ιστοσελίδας A. Ένας υπερσύνδεσμος σχετίζεται με το *κείμενο αγκύρωσης* (*anchortext*) που περιγράφει τον σύνδεσμο. Για παράδειγμα,

ο δημιουργός μιας ιστοσελίδας μπορεί να δημιουργήσει έναν σύνδεσμο που δείχνει στην `Google.com`, και να ορίσει το σχετικό anchortext ως «η αγαπημένη μου μηχανή αναζήτησης». Το anchortext, από την στιγμή που επιλέγεται από ανθρώπους που ενδιαφέρονται στο να υποδείξουν μια ιστοσελίδα, μπορεί να συνοψίζει καλύτερα το θέμα της σελίδας που υποδεικνύεται.

3.4 Μείωση της Διαστατικότητας

Στην αναπαράσταση περιεχομένου, μια ιστοσελίδα μπορεί να περιέχει εκατοντάδες μοναδικούς όρους και ολόκληρη η συλλογή από ιστοσελίδες μπορεί να περιέχει έναν συνολικό αριθμό από εκατοντάδες χιλιάδες μοναδικούς όρους. Αν όλοι οι μοναδικοί όροι χρησιμοποιηθούν σαν χαρακτηριστικά για την αναπαράσταση των σελίδων, η διάσταση του διανύσματος των χαρακτηριστικών μπορεί να είναι εκατοντάδες χιλιάδες. Παρομοίως, ο συνολικός αριθμός των δομικών χαρακτηριστικών στην αναπαράσταση δομής μπορεί επίσης να είναι πολύ μεγάλος για κατηγοριοποίηση. Η υψηλή διαστατικότητα του χώρου χαρακτηριστικών μπορεί να είναι προβληματική και ακριβή στον υπολογισμό. Από την άλλη, αναμένεται πάντα ότι μπορούμε να εκτελέσουμε την κατηγοριοποίηση σε έναν μικρότερο χώρο χαρακτηριστικών με σκοπό να μειώσουμε την υπολογιστική πολυπλοκότητα. Έτσι, θα πρέπει να χρησιμοποιηθούν τεχνικές *μείωσης της διαστατικότητας* (*dimensionality reduction*), των οποίων η ευθύνη είναι η μείωση της διαστατικότητας του χώρου χαρακτηριστικών με την εγγύηση όμως ότι αυτή η διαδικασία δεν μειώνει την ακρίβεια της κατηγοριοποίησης.

Η μείωση της διαστατικότητας είναι επίσης χρήσιμη επειδή τείνει να μειώνει το πρόβλημα του *υπερβολικού ταιριάσματος* (*over fitting*), δηλαδή του φαινομένου κατά το οποίο ο κατηγοριοποιητής ταιριάζει υπερβολικά στα δεδομένα εκπαίδευσης, αντί να γενικευθεί στα γενικά χαρακτηριστικά των δεδομένων εκπαίδευσης. Οι κατηγοριοποιητές που ταιριάζουν υπερβολικά στα δεδομένα εκπαίδευσης τείνουν να είναι πολύ καλοί στην κατηγοριοποίηση των δεδομένων εκπαίδευσης, αλλά είναι σημαντικά χειρότεροι στο να κατηγοριοποιούν άλλα δεδομένα. Μερικά πειράματα προτείνουν ότι για να αποφευχθεί το υπερβολικό ταίριασμα, χρειάζεται ένας αριθμός από δεδομένα εκπαίδευσης ανάλογα του αριθμού των χαρακτηριστικών. Αυτό σημαίνει ότι, ύστερα από την εφαρμογή της μείωσης της διαστατικότητας, το υπερβολικό ταίριασμα μπορεί να αποφευχθεί χρησιμοποιώντας έναν μικρότερο αριθμό από δεδομένα εκπαίδευσης.

Στην βιβλιογραφία έχουν προταθεί και τα σχετικά τους κέρδη έχουν αξιολογηθεί πειραματικά, πολυάριθμες μέθοδοι μείωσης της διαστατικότητας, είτε από το πεδίο της *Θεωρίας*

Πληροφορίας (*Information Theory*) είτε από το πεδίο της *Γραμμικής Άλγεβρας* (*Linear Algebra*). Αυτές οι μέθοδοι μπορούν να διαιρεθούν, ανάλογα με το τι είδους χαρακτηριστικά επιλέγουν, σε μεθόδους *επιλογής χαρακτηριστικών* (*feature selection*) και σε μεθόδους *εξαγωγής χαρακτηριστικών* (*feature extraction*), όπως περιγράφουμε παρακάτω:

- **Μέθοδοι Επιλογής Χαρακτηριστικών**: Οι μέθοδοι επιλογής χαρακτηριστικών επιλέγουν ένα υποσύνολο του αρχικού χώρου χαρακτηριστικών με βάση κάποια κριτήρια. Στην βιβλιογραφία έχουν προταθεί δύο κυρίαρχες προσεγγίσεις για επιλογή χαρακτηριστικών: η προσέγγιση του wrapper και η προσέγγιση του filter. Η προσέγγιση του wrapper χρησιμοποιεί μια αναζήτηση διαμέσου του χώρου των υποσυνόλων των χαρακτηριστικών. Επίσης, χρησιμοποιεί μια εκτιμώμενη ακρίβεια για έναν αλγόριθμο μάθησης ως μέτρο καταλληλότητας για ένα συγκεκριμένο υποσύνολο χαρακτηριστικών. Για παράδειγμα, για έναν αλγόριθμο νευρωνικών δικτύων, η προσέγγιση του wrapper επιλέγει ένα αρχικό υποσύνολο χαρακτηριστικών και μετράει την απόδοση του δικτύου. Έπειτα, παράγει ένα βελτιωμένο σύνολο χαρακτηριστικών και μετράει ξανά την απόδοση του δικτύου. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί μια συνθήκη τερματισμού (είτε μια ελάχιστη τιμή σφάλματος ή ένας αριθμός επαναλήψεων). Ενώ κάποιες μέθοδοι βασισμένες σε wrapper έχουν σημειώσει επιτυχία σε εργασίες κατηγοριοποίησης, είναι συχνά απαγορευτικά ακριβές για να εκτελεσθούν και μπορούν να καταρρεύσουν όταν χρησιμοποιούν έναν πολύ μεγάλο αριθμό από χαρακτηριστικά. Για την προσέγγιση του filter, η επιλογή χαρακτηριστικών εκτελείται σαν ένα βήμα προεπεξεργασίας πριν την εφαρμογή των μεθόδων μηχανικής μάθησης. Έτσι, η μέθοδος επιλογής χαρακτηριστικών είναι ανεξάρτητη του αλγορίθμου μάθησης. Ο αλγόριθμος filter δεν επιβαρύνει το κόστος υψηλής διαστατικότητας και χρησιμοποιείται ευρέως σε συστήματα κατηγοριοποίησης ακόμα και σε χώρους χαρακτηριστικών πολύ υψηλής διαστατικότητας. Επειδή τα πλεονεκτήματα της προσέγγισης με filter σε σχέση με την προσέγγιση με wrapper και λόγω της απαίτησης της μείωσης των χαρακτηριστικών σε έναν χώρο χαρακτηριστικών υψηλής διαστατικότητας, θα παραλειφθούν οι προσεγγίσεις βασισμένες σε wrapper και θα παρουσιασθούν στη συνέχεια διάφορες προσεγγίσεις που βασίζονται σε filter.

- **Μέθοδοι Εξαγωγής Χαρακτηριστικών**: Στην εξαγωγή χαρακτηριστικών, τα τελικά χαρακτηριστικά δεν είναι απαραίτητα υποσύνολο των αρχικών χαρακτηριστικών. Προκύπτουν από συνδυασμούς ή μετασχηματισμούς του αρχικού χώρου χαρακτηριστικών.

3.4.1 Επιλογή Χαρακτηριστικών

Η αρχή πίσω από την μείωση της διαστατικότητας με επιλογή χαρακτηριστικών είναι ότι πρέπει να επιλέγονται εκείνα τα χαρακτηριστικά που είναι πιο αντιπροσωπευτικά για κάποιες κατηγορίες σε σχέση με κάποιες άλλες. Η όλη διαδικασία της επιλογής χαρακτηριστικών απλοποιείται με την υπόθεση της ανεξαρτησίας των χαρακτηριστικών. Όλα τα χαρακτηριστικά αξιολογούνται ανεξάρτητα με βάση κάποια κριτήρια και ανατίθεται μια βαθμολογία σε κάθε ένα από αυτά. Έπειτα, ένα προκαθορισμένο κατώφλι βοηθά στην επιλογή των καλύτερων χαρακτηριστικών.

Έχουν χρησιμοποιηθεί πολλά κριτήρια για τον προσδιορισμό της αντιπροσωπευτικής ικανότητας των χαρακτηριστικών. Υπό μια ευρεία άποψη, τα κριτήρια επιλογής χαρακτηριστικών μπορούν να διαιρεθούν σε δύο σύνολα: Ένα σύνολο που υπολογίζει μόνο την τιμή που υποδηλώνει ένα χαρακτηριστικό που εμφανίζεται σε ένα δείγμα, όπως η επιλογή χαρακτηριστικών με την χρήση *Αμοιβαίας Πληροφορίας* (*Mutual Information*) και της *Cross Εντροπίας* (*Cross Entropy*). Το άλλο σύνολο υπολογίζει όλες τις πιθανές τιμές ενός χαρακτηριστικού συμπεριλαμβανομένου τωρινές και παρελθοντικές καταστάσεις, όπως η επιλογή χαρακτηριστικών με χρήση *Κέρδους Πληροφορίας* (*Information Gain*) και της χ^2 *Στατιστικής* (*Chi-square Statistic*). Οι μέθοδοι που μόλις αναφέραμε, περιγράφονται αναλυτικότερα στις επόμενες παραγράφους.

Αμοιβαία Πληροφορία

Η Αμοιβαία Πληροφορία [CH89, YP97] χρησιμοποιείται ευρέως για συσχετίσεις όρων και μπορεί να χρησιμοποιηθεί για επιλογή χαρακτηριστικών. Σε μια συλλογή από ιστοσελίδες, η Αμοιβαία Πληροφορία υπολογίζεται για κάθε χαρακτηριστικό και απομακρύνονται εκείνα τα χαρακτηριστικά των οποίων η αμοιβαία πληροφορία είναι μικρότερη από κάποιο προκαθορισμένο κατώφλι.

Η αμοιβαία πληροφορία μεταξύ ενός χαρακτηριστικού f και μιας κατηγορίας c_i ορίζεται ως

$$MI(f, c_i) = log\frac{Pr(f \wedge c_i)}{Pr(f)Pr(c_i)} \tag{3.9}$$

όπου η $Pr(f)$ είναι η αναλογία των δειγμάτων που περιέχουν ένα χαρακτηριστικό f σε σχέση με όλα τα δείγματα εκπαίδευσης, $Pr(c_i)$ είναι η αναλογία των δειγμάτων μιας κατηγορίας c_i σε σχέση με όλα τα δείγματα εκπαίδευσης και $Pr(f \wedge c_i)$ είναι η από κοινού πιθανότητα του χαρακτηριστικού f και της κατηγορίας c_i, η οποία είναι ίση με την αναλογία των δειγμάτων στα οποία εμφανίζονται από κοινού και το χαρακτηριστικό f και

η κατηγορία c_i. Η εξίσωση (3.9) μπορεί να μετασχηματισθεί σε

$$MI(f, c_i) = log\frac{Pr(c_i|f)}{Pr(c_i)} \quad \text{ή} \quad MI(f, c_i) = log\frac{Pr(f|c_i)}{Pr(f)} \tag{3.10}$$

όπου η $Pr(c_i|f)$ είναι η υπό συνθήκη πιθανότητα της κατηγορίας c_i δοθέντος ενός χαρακτηριστικού f και $Pr(f|c_i)$ είναι η υπό συνθήκη πιθανότητα του χαρακτηριστικού f δοθέντος της κατηγορίας c_i.

Η μέση και η μέγιστη τιμή της αμοιβαίας πληροφορίας ενός χαρακτηριστικού f υπολογίζεται πάνω σε όλες τις κατηγορίες ως εξής:

$$MI_{avg}(f) = \sum_{i=1}^{K} Pr(c_i)MI(f, c_i) \tag{3.11}$$

$$MI_{max}(f) = max_{i=1}^{K}\{MI(f, c_i)\} \tag{3.12}$$

όπου K είναι ο συνολικός αριθμός των κατηγοριών. Η χρονική πολυπλοκότητα υπολογισμού της αμοιβαίας πληροφορίας είναι $\mathcal{O}(MK)$, όπου M είναι το μέγεθος του χώρου χαρακτηριστικών και K είναι ο αριθμός των κατηγοριών.

Μια αδυναμία της αμοιβαίας πληροφορίας είναι ότι ευνοεί τα σπάνια χαρακτηριστικά. Από την (3.10), για χαρακτηριστικά με ίση υπό συνθήκη πιθανότητα $Pr(f|c_i)$, τα σπάνια χαρακτηριστικά θα έχουν υψηλότερη βαθμολογία από τα κοινά χαρακτηριστικά από την στιγμή που τα σπάνια χαρακτηριστικά έχουν μικρές τιμές $Pr(f)$ στον παρονομαστή.

Cross Εντροπία

Η cross εντροπία έχει χρησιμοποιηθεί στην κατηγοριοποίηση κειμένων [KS96, KJ97, MG99]. Χρησιμοποιεί μετρικές από την Θεωρία Πληροφορίας για τον προσδιορισμό ενός υποσυνόλου χαρακτηριστικών από τον αρχικό χώρο χαρακτηριστικών. Μπορεί επίσης να χρησιμοποιηθεί ως μέθοδος επιλογής χαρακτηριστικών για την απομάκρυνση περιττών χαρακτηριστικών. Φορμαλιστικά, έστω μ και σ δύο κατανομές πάνω σε κάποιον χώρο πιθανότητας Ω. Η cross πιθανότητα του μ και του σ ορίζεται ως εξής:

$$D(\mu, \sigma) = \sum_{x \in \Omega} \mu(x) \cdot log\frac{\mu(x)}{\sigma(x)} \tag{3.13}$$

Η (3.13) μας παρέχει μια έννοια της «απόστασης» μεταξύ μ και σ. Μπορούμε να μετασχηματίσουμε την παραπάνω εξίσωση ώστε να μπορεί να χρησιμοποιηθεί για επιλογή χαρακτηριστικών. Για κάθε χαρακτηριστικό f και ένα σύνολο κατηγοριών $C =$

$\{c_1, c_2, ..., c_K\}$, η $Pr(C|f)$ αντικαθιστά την μ και η $Pr(C)$ την σ. Έτσι, η αναμενόμενη cross εντροπία $CE(f)$ του χαρακτηριστικού f ορίζεται ως:

$$\begin{aligned} CE(f) &= Pr(f) \cdot D(Pr(C|f), Pr(C)) \\ &= Pr(f) \cdot \sum_{i=1}^{K} Pr(c_i|f) \cdot log\frac{Pr(c_i|f)}{Pr(c_i)} \end{aligned} \tag{3.14}$$

όπου $Pr(f)$ είναι ένας παράγοντας κανονικοποίησης. Τα χαρακτηριστικά των οποίων η $CE(f)$ είναι μικρότερη από ένα συγκεκριμένο προκαθορισμένο κατώφλι απομακρύνονται. Η χρονική πολυπλοκότητα για τον υπολογισμό της cross εντροπίας είναι $\mathcal{O}(MK)$, που είναι ίδια με αυτήν της αμοιβαίας πληροφορίας. Η cross εντροπία υπερνικά την αδυναμία της αμοιβαίας πληροφορίας ευνοώντας κοινά χαρακτηριστικά αντί των σπάνιων.

Κέρδος Πληροφορίας

Το κέρδος πληροφορίας χρησιμοποιείται συχνά ως κριτήριο καταλληλότητας ενός χαρακτηριστικού στην *Μηχανική Μάθηση* (*Machine Learning*) [Mit97]. Το κέρδος πληροφορίας ενός χαρακτηριστικού μετράει την αναμενόμενη μείωση της εντροπίας που προκαλείται από τον διαχωρισμό των δειγμάτων εκπαίδευσης σε σχέση με το χαρακτηριστικό. Η εντροπία χαρακτηρίζει την ετερογένεια μιας συλλογής από δεδομένα εκπαίδευσης. Το κέρδος πληροφορίας καλείται επίσης και *απώλεια εντροπίας* [GTL+02]. Φορμαλιστικά, το κέρδος πληροφορίας ενός χαρακτηριστικού f ορίζεται ως:

$$IG(f) \equiv Entropy(D) - \sum_{f,\overline{f}} \frac{D_v}{|D|} Entropy(D_v) \tag{3.15}$$

όπου D είναι μια συλλογή από δεδομένα εκπαίδευσης και D_v είναι ένα υποσύνολο του D το οποίο ορίζεται από τη δυαδική τιμή v ενός χαρακτηριστικού. Για παράδειγμα, D_f είναι ένα υποσύνολο του D στο οποίο κάθε δείγμα περιέχει το χαρακτηριστικό f και $D_{\overline{f}}$ είναι ένα υποσύνολο του D στο οποίο κάθε δείγμα δεν περιέχει το χαρακτηριστικό f. Η $Entropy(D)$ ορίζεται ως:

$$Entropy(D) \equiv -\sum_{i=1}^{K} Pr(c_i) log PR(c_i) \tag{3.16}$$

όπου K είναι ο συνολικός αριθμός των κατηγοριών στην συλλογή D και $Pr(c_i)$ είναι η αναλογία των δειγμάτων στην κατηγορία c_i σε σχέση με τον συνολικό αριθμό των παρα-

δειγμάτων. Αντικαθιστώντας την (3.16) στην (3.15), το κέρδος πληροφορίας του χαρακτηριστικού f είναι:

$$\begin{aligned} IG(f) = &- \sum_{i=1}^{K} Pr(c_i) log PR(c_i) \\ &+ Pr(f) \sum_{i=1}^{K} Pr(c_i|f) log Pr(ci|f) \\ &+ Pr(\overline{f}) \sum_{i=1}^{K} Pr(c_i|\overline{f}) log Pr(ci|\overline{f}) \end{aligned} \tag{3.17}$$

το οποίο είναι είναι ισοδύναμο με:

$$\begin{aligned} IG(f) = &Pr(f) \cdot \sum_{i=1}^{K} Pr(c_i|f) log \frac{Pr(c_i|f)}{Pr(c_i)} \\ &+ Pr(\overline{f}) \cdot \sum_{i=1}^{K} Pr(c_i|\overline{f}) log \frac{Pr(c_i|\overline{f})}{Pr(c_i)} \end{aligned} \tag{3.18}$$

όπου η $Pr(f)$ είναι η αναλογία των δειγμάτων στα οποία είναι παρών το χαρακτηριστικό f, $Pr(\overline{f})$ είναι η αναλογία των δειγμάτων στα οποία είναι απών το χαρακτηριστικό f, $Pr(c_i|f)$ είναι η υπό συνθήκη πιθανότητα της κατηγορίας c_i δοθέντος ενός χαρακτηριστικού f και $Pr(c_i|\overline{f})$ είναι η υπό συνθήκη πιθανότητα της κατηγορίας c_i δοθέντος ενός απόντος χαρακτηριστικού f.

Υπολογίζεται το κέρδος πληροφορίας κάθε χαρακτηριστικού και απομακρύνονται τα χαρακτηριστικά των οποίων το κέρδος πληροφορίας είναι μικρότερο από ένα προκαθορισμένο κατώφλι. Ο υπολογισμός περιλαμβάνει την εκτίμηση των υπό συνθήκη πιθανοτήτων μιας κατηγορίας δοθέντος ενός χαρακτηριστικού και των υπολογισμών της εντροπίας. Η εκτίμηση της πιθανότητας έχει χρονική πολυπλοκότητα $\mathcal{O}(N)$ όπου N είναι ο αριθμός των δειγμάτων εκπαίδευσης. Ο υπολογισμός της εντροπίας έχει χρονική πολυπλοκότητα $\mathcal{O}(MK)$ όπου M το μέγεθος του χώρου χαρακτηριστικών και K ο αριθμός των κατηγοριών.

χ^2 Στατιστική

Η χ^2 στατιστική μετράει την έλλειψη της ανεξαρτησίας μεταξύ του χαρακτηριστικού f και της κατηγορίας c_i. Η μετρική καταλληλότητας του χαρακτηριστικού με την χ^2 στατιστική

ορίζεται ως [YP97, GSS00]:

$$\chi^2(f, c_i) = \frac{N[Pr(f \wedge c_i)Pr(\overline{f} \wedge \overline{c_i}) - Pr(f \wedge \overline{c_i})Pr(\overline{f} \wedge c_i)]^2}{Pr(f)Pr(\overline{f})Pr(c_i)Pr(\overline{c_i})} \tag{3.19}$$

όπου N είναι ο συνολικός αριθμός των δειγμάτων εκπαίδευσης. Τα χαρακτηριστικά με υψηλές τιμές $\chi^2(f, c_i)$ σχετίζονται περισσότερο με την κατηγορία c_i και επιλέγονται ως το αποτέλεσμα της διαδικασίας επιλογής χαρακτηριστικών.

Η μέση και η μέγιστη τιμή της χ^2 στατιστικής ενός χαρακτηριστικού f υπολογίζονται πάνω σε όλες τις κατηγορίες ως εξής:

$$\chi^2_{avg}(f) = \sum_{i=1}^{K} Pr(c_i)\chi^2(f, c_i) \tag{3.20}$$

$$\chi^2_{max}(f) = max_{i=1}^{K}\{\chi^2(f, c_i)\} \tag{3.21}$$

Ο υπολογισμός της χ^2 στατιστικής έχει χρονική πολυπλοκότητα $\mathcal{O}(MK)$, όπου M είναι το μέγεθος του χώρου χαρακτηριστικών και K ο αριθμός των κατηγοριών. Τα χαρακτηριστικά με χαμηλές τιμές χ^2 στατιστικής απομακρύνονται. Η χρήση $\chi^2_{max}(f)$ είναι καλύτερη από την χρήση $\chi^2_{avg}(f)$ όπως υποδεικνύεται στις εργασίες.

Τα σπάνια χαρακτηριστικά ευνοούνται από την χ^2 στην μορφή της $Pr(f)$ στον παρονομαστή. Για την αποφυγή αυτής της κατάστασης, στην εργασία [GSS00] προτάθηκε μια απλοποιημένη χ^2:

$$\chi^2(f, c_i) = Pr(f \wedge c_i)Pr(\overline{f} \wedge \overline{c_i} - Pr(f \wedge \overline{c_i}))Pr(\overline{f} \wedge c_i) \tag{3.22}$$

Η μορφή αυτή ευνοεί την θετική συσχέτιση (όροι $Pr(f \wedge c_i)$ και $Pr(\overline{f} \wedge \overline{c_i})$ μεταξύ f και c_i και υποβαθμίζει την αρνητική συσχέτιση (όροι $Pr(f \wedge \overline{c_i})$) και $Pr(\overline{f} \wedge c_i)$). Στην εργασία [GSS00] έδειξαν ότι η απλοποιημένη χ^2 είναι καλύτερη από την αρχική χ^2 όταν απαιτείται μείωση των χαρακτηριστικών παραπάνω από 95%. Ωστόσο, όταν απαιτείται μείωση των χαρακτηριστικών μικρότερη από 95%, η απλοποιημένη εκδοχή είναι ελαφρά κατώτερη από την αρχική.

3.4.2 Εξαγωγή Χαρακτηριστικών

Η εξαγωγή χαρακτηριστικών συνθέτει ένα σύνολο από νέα χαρακτηριστικά από το σύνολο των αρχικών χαρακτηριστικών όπου ο αριθμός των νέων χαρακτηριστικών είναι πολύ μικρότερος από τον αριθμό των αρχικών χαρακτηριστικών. Η λογική για την χρήση συν-

θετικών αντί φυσικά εμφανιζόμενων χαρακτηριστικών, είναι ότι τα αρχικά χαρακτηριστικά μπορεί να μην σχηματίζουν την βέλτιστη διάσταση για την αναπαράσταση των ιστοσελίδων. Μέθοδοι για εξαγωγή χαρακτηριστικών στοχεύουν στην δημιουργία τεχνητών χαρακτηριστικών που δεν υποφέρουν από προβλήματα όπως η πολυσημία, η ομωνυμία και η συνωνυμία που είναι παρούσες στα αρχικά χαρακτηριστικά. Στη βιβλιογραφία, έχουν προταθεί και δοκιμασθεί διάφορες προσεγγίσεις. Στη συνέχεια, περιγράφουμε μία από τις πιο διαδεδομένες τεχνικές εξαγωγής χαρακτηριστικών, την *Λανθάνουσα Σημασιολογική Δεικτοδότηση* (*Latent Semantic Indexing - LSI*).

Latent Semantic Indexing

Η τεχνική latent semantic indexing (LSI) βασίζεται στην υπόθεση ότι υπάρχει μια υποκείμενη ή λανθάνουσα σημασιολογική δομή στο πρότυπο των χαρακτηριστικών που χρησιμοποιούνται στο σώμα κειμένων των ιστοσελίδων και ότι κάποιες στατιστικές τεχνικές μπορούν να χρησιμοποιηθούν για την εκτίμηση της δομής [DDL+90, BDO95]. Η LSI χρησιμοποιεί την τεχνική *Παραγοντοποίηση Ιδιαζουσών Τιμών (Singular Value Decomposition - SVD)*, η οποία είναι μια τεχνική παρόμοια με την παραγοντοποίηση ιδιοδιανυσμάτων και την ανάλυση παραγόντων.

Η βασική ιδέα στην LSI είναι η άμεση μοντελοποίηση των αλληλεξαρτήσεων μεταξύ των χαρακτηριστικών με την χρήση της SVD και την εκμετάλλευση αυτής για την βελτίωση της κατηγοριοποίησης. Η διαδικασία ξεκινάει κατασκευάζοντας ένα μητρώο M χαρακτηριστικών επί N κειμένων που το ονομάζουμε A, όπου κάθε στοιχείο του αναπαριστά το βάρος ενός χαρακτηριστικού στο κείμενο, για την ακρίβεια:

$$A = (a_{ij}) \tag{3.23}$$

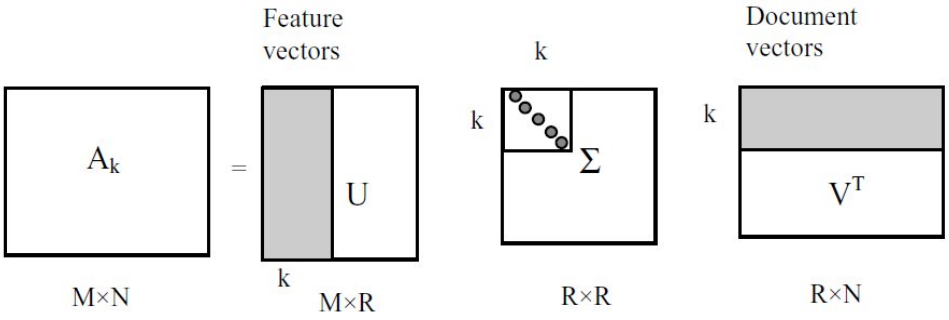
όπου, a_{ij} είναι το βάρος των χαρακτηριστικού $i(1 \leq i \leq M)$ στο κείμενο $j(1 \leq j \leq N)$. Από την στιγμή που δεν εμφανίζεται κάθε χαρακτηριστικό σε κάθε κείμενο, το μητρώο A είναι συνήθως *αραιό* (*sparse*). Η SVD ενός μητρώου A δίνεται από:

$$A_{M\times N} = U_{M\times R} \sum_{R\times R} V^T_{R\times N} \tag{3.24}$$

όπου R είναι η τάξη του $A(R \leq min(M, N))$, τα U και V έχουν ορθογώνιες μοναδιαίες στήλες ($UTU = I$ και $VTV = I$) και το Σ είναι το διαγώνιο μητρώο των ιδιαζουσών τιμών του $A(\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_R))$ οι οποίες είναι οι μη αρνητικές ρίζες των ιδιοτιμών του AAT. Ο Πίνακας 3.1 περιγράφει τους ορισμούς των όρων.

A: μητρώο M χαρακτηριστικών επί N κειμένων	M: αριθμός χαρακτηριστικών
U: διάνυσμα χαρακτηριστικών	N: αριθμός κειμένων
Σ: διάνυσμα ιδιαζουσών τιμών	R: τάξη του A
V: διάνυσμα κειμένων	k: αριθμός παραγόντων (k υψηλότερες ιδιάζουσες τιμές)

Πίνακας 3.1: Πίνακας συμβόλων.



Σχήμα 3.5: Γραφική απεικόνιση του μητρώου A_k [CY05].

Αν οι ιδιάζουσες τιμές στον Σ είναι διατεταγμένες με το μέγεθος, μπορούν να κρατηθούν οι k μεγαλύτερες και οι υπόλοιπες μικρότερες μπορούν να μηδενιστούν, για την ακρίβεια, $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_k, ..., \sigma_R)$, όπου $\sigma_i > 0$ για $1 \leq i \leq k$ και $\sigma_i = 0$ για $i > k$. Το γινόμενο των μητρώων που προκύπτουν είναι ένα μητρώο A_k το οποίο είναι μια προσέγγιση του A με τάξη k, για την ακρίβεια:

$$A_k = U_k \sum_k V_k^T \tag{3.25}$$

όπου $k \ll M$, το Σ_k προκύπτει από την διαγραφή των μηδενικών γραμμών και στηλών του Σ και τα U_k και V_k προκύπτουν από την διαγραφή των αντίστοιχων γραμμών και στηλών των U και V, όπως δείχνει η Εικόνα 3.5.

Το μητρώο A_k που προκύπτει εμπεριέχει την περισσότερη από την υποκείμενη δομή των σχέσεων μεταξύ των χαρακτηριστικών και των κειμένων στον A. Τα τρία μητρώα U_k, Σ_k και V_k αντανακλούν μια ανάλυση των αρχικών σχέσεων χαρακτηριστικών-κειμένων σε γραμμικώς ανεξάρτητα διανύσματα ή παράγοντες. Η χρήση των k παραγόντων ή των k μεγαλύτερων ιδιαζουσών τριπλετών είναι ισοδύναμη με την προσέγγιση του αρχικού μητρώου. Επιπρόσθετα, ένα νέο κείμενο d μπορεί να αναπαρασταθεί σαν ένα διάνυσμα στον k-διάστατο χώρο ως:

$$\tilde{d} = d^T U_k \Sigma_k^{-1} \tag{3.26}$$

όπου $d^T U_k$ αντανακλά το άθροισμα των k-διάστατων διανυσμάτων χαρακτηριστικών και

$\sum_k^{-1}$ βαρών σε ξεχωριστές διαστάσεις.

Είναι δύσκολο να ερμηνεύσουμε τον νέο μικρότερο k-διάστατο χώρο παρόλο που υποτίθεται ότι δουλεύει καλά στο να αιχμαλωτίσει την υποκείμενη δομή του μητρώου χαρακτηριστικών κειμένων. Ένα παράδειγμα που δίνεται στην εργασία [BDO95] μπορεί να μας βοηθήσει να κατανοήσουμε τον νέο χώρο: έστω οι λέξεις «αυτοκίνητο», «όχημα», «οδηγός» και «ελέφαντας». Οι λέξεις «αυτοκίνητο» και «όχημα» είναι συνώνυμες, ο «οδηγός» είναι σχετική έννοια και ο «ελέφαντας» δεν είναι σχετικός καθόλου. Οι λέξεις «αυτοκίνητο» και «όχημα» θα εμφανίζονται με πολλές από παρόμοιες λέξεις όπως «κινητήρας», «μοντέλο», «αμάξωμα» και «μηχανή». Έτσι, θα έχουν παρόμοιες αναπαραστάσεις στον k-διάστατο χώρο. Το πλαίσιο του «οδηγού» θα επικαλύπτεται σε μικρότερο βαθμό και το αντίστοιχο του «ελέφαντα» θα είναι πολύ ανόμοιο. Αυτό σχετίζεται με το γεγονός ότι χαρακτηριστικά που προκύπτουν σε παρόμοια κείμενα θα είναι κοντά το ένα στο άλλο στον k-διάστατο χώρο ακόμα και αν αυτά τα χαρακτηριστικά δεν συνεμφανίζονται στο ίδιο κείμενο. Αυτό περαιτέρω σημαίνει ότι δύο κείμενα μπορεί να είναι παρόμοια ακόμα και αν δεν μοιράζονται τις ίδιες λέξεις κλειδιά.

3.5 Αλγόριθμοι Κατηγοριοποίησης

Εφόσον επιλεγούν τα χαρακτηριστικά από τις προς εκπαίδευση ιστοσελίδες για τον σχηματισμό ακριβών αναπαραστάσεων των ιστοσελίδων, μπορούν να εφαρμοσθούν διάφορες μέθοδοι μηχανικής μάθησης και αλγόριθμοι κατηγοριοποίησης για την επαγωγή της συνάρτησης κατηγοριοποίησης f' όπως αυτή ορίστηκε στην αρχή του κεφαλαίου, ή για την επαγωγή των αναπαραστάσεων των κατηγοριών από τις αναπαραστάσεις των προς εκπαίδευση ιστοσελίδων. Όταν είναι να κατηγοριοποιηθεί μια νέα ιστοσελίδα, οι κατηγοριοποιητές χρησιμοποιούν την συνάρτηση για να αναθέσουν την ιστοσελίδα στις κατηγορίες.

Στις επόμενες παραγράφους παρουσιάζουμε κάποιους από τους state of the art κατηγοριοποιητές στην κατηγοριοποίηση ιστοσελίδων. Χωρίζουμε τους κατηγοριοποιητές που εμφανίζονται στην βιβλιογραφία σε τέσσερις κατηγορίες: *βασισμένους σε προφίλ*, (*profile based*), *βασισμένους σε κανόνα* (*rule learning based*), *βασισμένους σε δείγμα* (*direct example based*) και *βασισμένους σε παράμετρο* (*parameter based*), όπου οι τρεις πρώτοι ονομάζονται *μη παραμετρικές* (*non-parametric*) προσεγγίσεις ενώ η τελευταία *παραμετρική parametric*. Παρουσιάζουμε πρώτα κάποιους ορισμούς και κάποιους γενικούς συμβολισμούς. Μια ιστοσελίδα αναπαριστάται συνήθως από ένα διάνυσμα $d_i = \{w_1, w_2, ..., w_M\}$, όπου κάθε w_i είναι το βάρος ενός χαρακτηριστικού για την σελίδα και M είναι το μέγεθος του χώρου χαρακτηριστικών. Οι προκαθορισμένες κατηγορίες συμβολίζονται με ένα

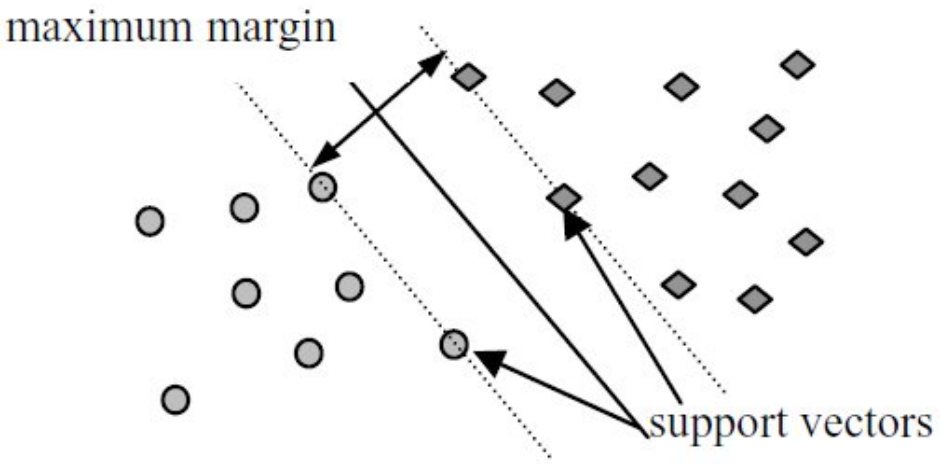
σύνολο $C = \{c_1, c_2, ..., c_K\}$, όπου κάθε c_i είναι μια ετικέτα κατηγορίας και υπάρχουν K κατηγορίες. Τα δείγματα εκπαίδευσης αποτελούνται από N ιστοσελίδες που αναπαριστώνται από διανύσματα $d_1, d_2, ..., d_N$, τα οποία είναι επισημειωμένα με τις σωστές ετικέτες κατηγοριών $y_1, y_2, ..., y_N$ αντίστοιχα. Έστω N_j ο αριθμός των σελίδων εκπαίδευσης των οποίων η σωστή ετικέτα κατηγορίας είναι c_j. Γενικότερα, η διαδικασία κατηγοριοποίησης αποτελείται από μια φάση εκπαίδευσης και από μια φάση ελέγχου. Κατά την διάρκεια της φάσης εκπαίδευσης, τα δείγματα εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση των κατηγοριοποιητών. Κατά την διάρκεια της φάσης ελέγχου, εφαρμόζονται οι κατηγοριοποιητές για την κατηγοριοποίηση των ιστοσελίδων. Μερικοί rule learning κατηγοριοποιητές αποτελούνται επίσης και από μία φάση επικύρωσης στην οποία βελτιστοποιούνται οι κανόνες.

3.5.1 Profile based Κατηγοριοποιητές

Για τους profile based κατηγοριοποιητές, εξάγεται ένα προφίλ (ή μια αναπαράσταση) για κάθε κατηγορία από ένα σύνολο από σελίδες εκπαίδευσης οι οποίες έχουν προκαθορισθεί ως αντιπροσωπευτικά δείγματα της κατηγορίας. Ύστερα από την εκπαίδευση όλων των κατηγοριών, οι κατηγοριοποιητές χρησιμοποιούνται για την κατηγοριοποίηση των νέων ιστοσελίδων. Όταν είναι να κατηγοριοποιηθεί μια νέα σελίδα, πρώτα αναπαρίσταται στην μορφή ενός διανύσματος χαρακτηριστικών. Το διάνυσμα αυτό συγκρίνεται με τα προφίλ όλων των κατηγοριών και βαθμολογείται. Στη γενική περίπτωση, η νέα σελίδα μπορεί να ανατεθεί σε περισσότερες από μία κατηγορίες ανάλογα με το κατώφλι των βαθμολογιών μεταξύ σελίδας και κατηγορίας. Οι μέθοδοι κατωφλίωσης μπορούν να επηρεάσουν σημαντικά τα αποτελέσματα της κατηγοριοποίησης. Στην περίπτωση που σε κάθε σελίδα αντιστοιχεί μία και μόνο μία κατηγορία, η νέα σελίδα ανατίθεται στην κατηγορία στην οποία πετυχαίνει την υψηλότερη βαθμολογία. Παραδείγματα κατηγοριοποιητών που χρησιμοποιούν αυτή την προσέγγιση είναι ο *Rocchio* κατηγοριοποιητής, τα *Support Vector Machines* και τα *Νευρωνικά Δίκτυα* (*Neural Networks*). Στη συνέχεια, παρουσιάζουμε αναλυτικότερα τα Support Vector Machines.

Support Vector Machines

Τα Support Vector Machines (SVMs) έχουν δείξει ότι επιτυγχάνουν καλή απόδοση σε ένα μεγάλο εύρος από προβλήματα κατηγοριοποίησης και πιο πρόσφατα στο πρόβλημα της κατηγοριοποίησης κειμένων [OFG97, Joa98, YL99]. Βασίζονται στην αρχή Structural Risk Minimization από την θεωρία υπολογιστικής μάθησης [CV95, Vap95]. Η ιδέα της



Σχήμα 3.6: Γραμμικό support vector machine. Η εικόνα δείχνει ένα παράδειγμα ενός απλού 2-διάστατου προβλήματος που είναι γραμμικώς διαχωρίσιμο. Οι ρόμβοι στην εικόνα αναπαριστούν θετικά δείγματα και οι κύκλοι αρνητικά. Το SVM ανακαλύπτει το υπερεπίπεδο h (δηλώνεται από την γραμμή), το οποίο διαχωρίζει τα θετικά και τα αρνητικά δείγματα εκπαίδευσης με το μέγιστο όριο (δηλαδή την απόσταση μεταξύ των δύο διακεκομμένων γραμμών). Τα δείγματα πιο κοντά στο υπερεπίπεδο καλούνται Support Vectors (υποδεικνύονται με βέλη). Με άλλα λόγια, το SVM ανακαλύπτει το h που μεγιστοποιεί την απόσταση των Support Vectors [CY05].

structural risk minimization είναι η εύρεση μιας υπόθεσης h που ορίζεται ως η συνάρτηση απόστασης με το μεγαλύτερο όριο μεταξύ των διανυσμάτων των θετικών και των αρνητικών δειγμάτων, όπως φαίνεται στην Εικόνα 3.6. Έχει δειχθεί ότι αν το σύνολο εκπαίδευσης είναι διαχωρίσιμο χωρίς λάθη από την h, η αναμενόμενη τιμή της πιθανότητας εύρεσης λάθους σε ένα δείγμα εκπαίδευσης φράσσεται από έναν πολύ μικρό αριθμό.

Στην πιο απλή τους (γραμμική) μορφή, ένα SVM είναι ένα υπερεπίπεδο το οποίο διαχωρίζει ένα σύνολο θετικών δειγμάτων από ένα σύνολο αρνητικών δειγμάτων με το μεγαλύτερο όριο (Εικόνα 3.6). Έστω $D = \{(y_i, d_i)\}$ συμβολίζεται το σύνολο εκπαίδευσης και $y_i \in \{+1, -1\}$ η κατηγοριοποίηση ενός διανύσματος κειμένου $\vec{d_i}$, όπου $+1$ υποδηλώνει ένα θετικό δείγμα και -1 υποδηλώνει ένα αρνητικό δείγμα μιας κατηγορίας. Το SVM εκπαιδεύει γραμμικές συναρτήσεις κατωφλίου του τύπου:

$$h(\vec{d}) = sign(\vec{w} \cdot \vec{d} + b) = \begin{cases} +1 & \text{αν } \vec{w} \cdot \vec{d} + b > 0 \\ -1 & \text{αλλιώς} \end{cases} \tag{3.27}$$

όπου $h(\vec{d})$ αναπαριστά μια δοθείσα υπόθεση και $\vec{w}$ αναπαριστά ένα διάνυσμα βαρών, ενώ το b είναι ένας βαθμωτός που δίνεται από την σχέση (3.32). Η εύρεση του υπερεπιπέδου που

έχει το μεγαλύτερο όριο μπορεί να μεταφραστεί στο ακόλουθο πρόβλημα βελτιστοποίησης:

$$\text{Ελαχιστοποίησε:} \quad \|\vec{w}\| \tag{3.28}$$

έτσι ώστε:

$$\forall i : y_i[\vec{w} \cdot \vec{d_i} + b] \geq 1 \tag{3.29}$$

όπου $\|\vec{w}\|$ συμβολίζει το Ευκλείδειο μήκος του διανύσματος βαρών $\vec{w}$. Ο περιορισμός που εκφράζεται στην (3.29) απαιτεί ότι όλα τα δείγματα εκπαίδευσης κατηγοριοποιούνται σωστά. Προκειμένου να λυθεί το παραπάνω πρόβλημα βελτιστοποίησης, χρησιμοποιούνται πολλαπλασιαστές Lagrange για την μετατροπή του προβλήματος σε ένα ισοδύναμο τετραγωνικό πρόβλημα βελτιστοποίησης:

$$\text{Ελαχιστοποίησε:} \quad -\sum_{i=1}^{N} a_i + \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j \vec{d_i} \cdot \vec{d_j} \tag{3.30}$$

έτσι ώστε:

$$\sum_{i=1}^{N} a_i y_i = 0 \quad \text{και} \quad \forall i : a_i \geq 0 \tag{3.31}$$

Για αυτό το τετραγωνικό πρόβλημα βελτιστοποίησης, μπορούν να χρησιμοποιηθούν αποδοτικοί αλγόριθμοι ώστε να βρεθεί το ολικό βέλτιστο. Το αποτέλεσμα της διαδικασίας βελτιστοποίησης είναι ένα σύνολο από συντελεστές a_i^* για τους οποίους η (3.30) ελαχιστοποιείται. Αυτοί οι συντελεστές μπορούν να χρησιμοποιηθούν για την κατασκευή του υπερεπιπέδου ως εξής:

$$\vec{w} \cdot \vec{d} = \left(\sum_{i=1}^{N} a_i^* y_i \vec{d_i} \right) \quad \text{και} \quad b = \frac{1}{2}(\vec{w} \cdot \vec{d_+} + \vec{w} \cdot \vec{d_-}) \tag{3.32}$$

Από την παραπάνω εξίσωση, μπορούμε να δούμε ότι το διάνυσμα βαρών $\vec{w}$ που προκύπτει, είναι κατασκευασμένο ως γραμμικός συνδυασμός των δειγμάτων εκπαίδευσης. Μόνο τα διανύσματα εκπαίδευσης, για τα οποία ο συντελεστής a_i είναι μεγαλύτερος από το μηδέν, συνεισφέρουν στον συνδυασμό. Αυτά τα διανύσματα καλούνται *Support Vectors*, όπως φαίνεται και στην Εικόνα 3.6. Για τον υπολογισμό του b, χρησιμοποιούνται οποιοδήποτε support vector $\vec{d_+}$ από τα θετικά δείγματα και οποιοδήποτε $\vec{d_-}$ από τα αρνητικά δείγματα.

Εφόσον καθορισθούν το διάνυσμα βαρών για κάθε μία από τις δοθείσες κατηγορίες, ένα νέο κείμενο d μπορεί να κατηγοριοποιηθεί από τον υπολογισμό $\vec{w} \cdot \vec{d} + b$ στην (3.27),

όπου $\vec{w}$ είναι το διάνυσμα βαρών που προέκυψε για μια δοθείσα κατηγορία και $\vec{d}$ είναι ένα διάνυσμα που αναπαριστά το νέο κείμενο. Αν μια τιμή είναι μεγαλύτερη του μηδενός, τότε το νέο κείμενο ανατίθεται σε αυτήν την κατηγορία.

3.5.2 Rule Learning based Κατηγοριοποιητές

Μία από τις πιο επεξηγηματικές και κατανοητές από τους ανθρώπους αναπαραστάσεις για τα μοντέλα είναι τα σύνολα από *Εάν-Τότε* (*If-Then*) κανόνες. Η πιο σημαντική ιδιότητα των αλγορίθμων επαγωγής κανόνων είναι ότι επιτρέπουν στις αλληλεξαρτήσεις των χαρακτηριστικών να επηρεάζουν το αποτέλεσμα της κατηγοριοποίησης, σε αντίθεση με άλλα σχήματα κατηγοριοποίησης όπως για παράδειγμα ο Naive Bayes, που υποθέτουν ότι τα χαρακτηριστικά είναι ανεξάρτητες συνιστώσες.

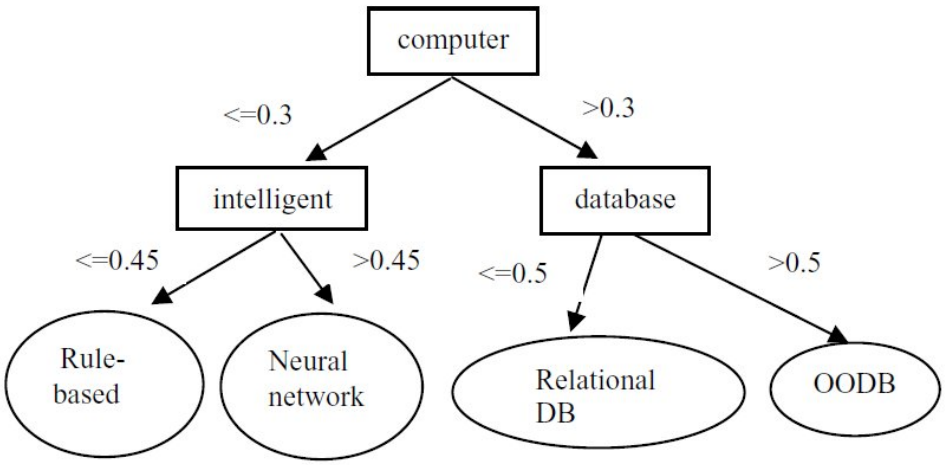
Στη γενική περίπτωση, για τους rule learning based κατηγοριοποιητές, οι σελίδες εκπαίδευσης για μια κατηγορία χρησιμοποιούνται για να εξάγουν ένα σύνολο κανόνων που περιγράφουν την κατηγορία. Μια ιστοσελίδα που είναι να κατηγοριοποιηθεί χρησιμοποιείται για να ταιριάξει τις συνθήκες των κανόνων. Οι κανόνες που ταιριάζουν προβλέπουν την κατηγορία που ανήκει η ιστοσελίδα βασισμένοι στην συνέπεια των κανόνων. Μία από τις πιο επιτυχημένες αναπαραστάσεις των rule learning based κατηγοριοποιητών είναι τα *Δέντρα Απόφασης* (*Decision Trees*) τα οποία και περιγράφουμε στις επόμενες παραγράφους.

Δέντρα Απόφασης

Οι κατηγοριοποιητές που χρησιμοποιούν δέντρα απόφασης θεωρούνται rule learning based κατηγοριοποιητές από την στιγμή που ένα δέντρο απόφασης μπορεί να μετατραπεί σε ένα σύνολο διαχωριστικών ή συνδυαστικών κανόνων. Τα δέντρα απόφασης έχουν χρησιμοποιηθεί εκτενώς στην κατηγοριοποίηση κειμένων, ωστόσο, αυτή η προσέγγιση δεν εμπεριέχει κανέναν ιδιαίτερο μηχανισμό για τον χειρισμό των μεγάλων συνόλων χαρακτηριστικών που απαντώνται στην κατηγοριοποίηση κειμένων και πιθανόν αυτός είναι ο λόγος για την σχετικά φτωχή τους απόδοση.

Για την κατηγοριοποίηση ενός κειμένου d με την χρήση ενός δέντρου απόφασης, το διάνυσμα κειμένου $\vec{d}$ εξετάζεται πάνω στο δέντρο απόφασης για τον προσδιορισμό της κατηγορίας στην οποία το κείμενο d ανήκει [BFOS84, Qui93, Mit97, AE99]. Το δέντρο απόφασης κατασκευάζεται από κείμενα εκπαίδευσης. Μια δημοφιλής προσέγγιση είναι ο αλγόριθμος CART [BFOS84].

Η προσέγγιση του CART όπως περιγράφεται στην [AE99], κατασκευάζει ένα δυαδικό



Σχήμα 3.7: Ένα δέντρο απόφασης. Κάθε κόμβος, εκτός από τα φύλλα, αναπαριστά ένα χαρακτηριστικό, κάθε κλαδί που ξεκινά από κάποιον κόμβο αντιστοιχεί σε μια από τις πιθανές τιμές του χαρακτηριστικού (πχ η TFIDF ενός όρου) και κάθε φύλλο είναι μια ετικέτα κατηγορίας. Ένα νέο κείμενο δοκιμής κατηγοριοποιείται με την διάτρεξη του δέντρου καταλήγοντας στο κατάλληλο φύλλο και επιστρέφοντας την κατηγορία που σχετίζεται με αυτό το φύλλο [CY05].

δέντρο απόφασης (πχ Εικόνα 3.7) από ένα σύνολο κειμένων εκπαίδευσης τα οποία αναπαριστώνται ως διανύσματα χαρακτηριστικών. Σε κάθε βήμα, επιλέγεται ένα χαρακτηριστικό από το σύνολο των διανυσμάτων χαρακτηριστικών και χρησιμοποιείται για τον διαχωρισμό του συνόλου των διανυσμάτων χαρακτηριστικών σε δύο υποσύνολα. Μια μετρική που καλείται *ποικιλότητα* (*diversity*) χρησιμοποιείται για τον προσδιορισμό των χαρακτηριστικών που θα επιλεγούν. Η καλύτερη επιλογή γίνεται με την μεγιστοποίηση:

$$diversity(beforesplit) - [diversity(leftchild) + diversity(rightchild)] \tag{3.33}$$

Μία από τις πιο κοινά χρησιμοποιούμενες μετρικές είναι η Εντροπία (Εξίσωση (3.16)) και η *Gini Εντροπία* (*GE*) που χρησιμοποιείται από τον CART. Η Gini Εντροπία ενός κόμβου t ορίζεται ως:

$$GE(t) = 1 - \sum_{j=1}^{K} Pr(c_j|t) \tag{3.34}$$

όπου K είναι ο αριθμός των κατηγοριών και $Pr(c_j|t)$ είναι η πιθανότητα ενός δείγματος εκπαίδευσης που ανήκει στην κατηγορία c_j που «πέφτει» σε έναν κόμβο t. $Pr(c_j|t)$ μπορεί να εκτιμηθεί από την σχέση:

$$Pr(c_j|t) = \frac{N_j(t)}{N(t)} \tag{3.35}$$

όπου $N_j(t)$ είναι ο αριθμός των δειγμάτων εκπαίδευσης της κλάσης c_j του κόμβου t και

$N(t)$ είναι ο συνολικός αριθμός από δείγματα εκπαίδευσης στον κόμβο t.

Για την επιλογή ενός χαρακτηριστικού για έναν κόμβο (πχ Εικόνα 3.7), αξιολογείται κάθε χαρακτηριστικό σε όλα τα διανύσματα εκπαίδευσης με χρήση των συναρτήσεων (3.33) και (3.34). Επιλέγεται το χαρακτηριστικό που προκύπτει με την μέγιστη τιμή στην σχέση (3.33) και χρησιμοποιείται για τον διαχωρισμό του συνόλου των διανυσμάτων εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται μέχρι τα διανύσματα εκπαίδευσης να μην μπορούν να χωριστούν περαιτέρω. Κάθε διάνυσμα εκπαίδευσης μπορεί στη συνέχεια να χρησιμοποιηθεί για να διατρεχθεί το δυαδικό δέντρο που προκύπτει από την ρίζα σε ένα φύλλο. Στο φύλλο που προκύπτει ανατίθεται η ετικέτα κατηγορίας του διανύσματος εκπαίδευσης.

Μετά την κατασκευή του αρχικού δυαδικού δέντρου με την χρήση του παραπάνω αλγορίθμου, το δέντρο που προκύπτει συνήθως ταιριάζει υπερβολικά στα κείμενα εκπαίδευσης και δεν είναι αποδοτικό στην κατηγοριοποίηση νέων κειμένων. Έτσι, το αρχικό δέντρο κλαδεύεται με την απομάκρυνση των κλαδιών που παρέχουν την ελάχιστη επιπρόσθετη πληροφορία για κάθε φύλλο. Το αποτέλεσμα του κλαδέματος είναι ένα νέο δέντρο. Το τελικό βήμα είναι η επιλογή ενός δέντρου το οποίο θα κατηγοριοποιεί βέλτιστα τα νέα κείμενα. Για αυτόν τον σκοπό, χρησιμοποιείται ένα νέο σύνολο από επισημειωμένα κείμενα ως σύνολο επικύρωσης. Κάθε ένα από τα υποψήφια δέντρα χρησιμοποιείται για την κατηγοριοποίηση των κειμένων του συνόλου επικύρωσης. Το δέντρο που επιτυγχάνει την υψηλότερη ακρίβεια επιλέγεται ως το τελικό δέντρο.

Ένας άλλος πολύ γνωστός αλγόριθμος δέντρου απόφασης είναι ο C4.5 [Qui93]. Διαφέρει από τον CART στο ότι παράγει ένα δέντρο με διαφορετικό αριθμό από κλαδιά ανά κόμβο ενώ ο CART παράγει δυαδικό δέντρο. Επίσης, χρησιμοποιεί μια διαφορετική προσέγγιση για το κλάδεμα του δέντρου μετατρέποντας το δέντρο σε ένα ισοδύναμο σύνολο κανόνων, οι οποίοι είναι το αποτέλεσμα της δημιουργίας ενός κανόνα για κάθε μονοπάτι από την ρίζα σε ένα φύλλο. Κάθε χαρακτηριστικό κατά μήκος του μονοπατιού γίνεται συνθήκη ενώ η ετικέτα κατηγορίας στο φύλλο γίνεται το αποτέλεσμα. Για παράδειγμα, το πιο αριστερό μονοπάτι στην Εικόνα 3.7 μεταφράζεται στον κανόνα: $IF\ (computer <= 0.3)\ and\ (intelligent <= 0.45)\ THEN\ class\ label\ is\ "Rule-based"$. Στη συνέχεια, κάθε τέτοιος κανόνας κλαδεύεται με την αφαίρεση οποιασδήποτε συνθήκης, της οποίας η απουσία δεν θα χειροτερέψει την εκτιμώμενη ακρίβεια.

3.5.3 Direct Example based Κατηγοριοποιητές

Για έναν direct example based κατηγοριοποιητή, μια ιστοσελίδα που είναι να κατηγοριοποιηθεί χρησιμοποιείται ως ερώτημα απευθείας πάνω σε ένα σύνολο των δειγμάτων που προσδιορίζουν τις κατηγορίες. Η ιστοσελίδα ανατίθεται σε μια κατηγορία της οποίας το σύνολο των δειγμάτων έχει την μεγαλύτερη ομοιότητα με την ιστοσελίδα. Αυτοί οι κατηγοριοποιητές ονομάζονται *τεμπέληδες* (*lazy*) και ένας χαρακτηριστικός αντιπρόσωπος αυτής της κατηγορίας αλγορίθμων είναι ο k *Κοντινότεροι Γείτονες* (*k Nearest Neighbors*).

k Nearest Neighbors

Σε αντίθεση με τους «πρόθυμους» κατηγοριοποιητές (πχ Rocchio) οι οποίοι έχουν μια φάση εκπαίδευσης πριν την κατηγοριοποίηση νέων κειμένων, ο k Nearest Neighbors (kNN) αποτελεί μια τεμπέλικη μέθοδο η οποία καθυστερεί την διαδικασία εκμάθησης μέχρι να βρεθεί ένα κείμενο προς κατηγοριοποίηση. Ο kNN έχει χρησιμοποιηθεί με επιτυχία στην κατηγοριοποίηση κειμένων [Yan99, YL99].

Ο kNN συγκρίνει ένα νέο κείμενο απευθείας με τα δοθέν κείμενα εκπαίδευσης. Χρησιμοποιεί την μετρική συνημιτόνου για τον υπολογισμό της ομοιότητας μεταξύ δύο διανυσμάτων κειμένου. Βαθμολογεί σε φθίνουσα κλίμακα τα κείμενα εκπαίδευσης με βάση τις ομοιότητες με το καινούριο κείμενο. Τα καλύτερα k κείμενα εκπαίδευσης αποτελούν τους k κοντινότερους γείτονες του καινούριου κειμένου και χρησιμοποιούνται για την πρόβλεψη της κατηγορίας του καινούριου κειμένου. Στην εργασία [Yan99] έδειξαν ότι η απόδοση του kNN είναι σχετικά σταθερή για ένα μεγάλο εύρος τιμών του k.

Η βαθμολογία ομοιότητας καθενός κειμένου γείτονα χρησιμοποιείται σαν βάρος για την σχετιζόμενη κατηγορία. Για την κατηγοριοποίηση ενός νέου κειμένου, η βαθμολογία πιθανότητας μιας κατηγορίας μπορεί να υπολογισθεί όπως στην [YL99]:

$$y(\vec{d'}, c_j) = \sum_{\vec{d_i} \in KNN} sim(\vec{d'}, \vec{d_i}) y(\vec{d_i}, c_j) - b_j \tag{3.36}$$

όπου $\vec{d_i}$ είναι ένας από τους k κοντινότερους γείτονες του καινούριου κειμένου $\vec{d'}$, $y(\vec{d_i}, c_j) \in \{0, 1\}$ είναι η κατηγοριοποίηση για τον γείτονα $\vec{d_i}$ σε σχέση με την κατηγορία c_j και $sim(\vec{d'}, \vec{d_i})$ είναι η ομοιότητα (πχ ομοιότητα συνημιτόνου) μεταξύ του νέου κειμένου $\vec{d'}$ και του γείτονά του $\vec{d_i}$ και b_j είναι το κατώφλι της συγκεκριμένης κατηγορίας. Το b_j μαθαίνεται αυτόματα με την χρήση ενός συνόλου κειμένων επικύρωσης. Δηλαδή, ο αλγόριθμος kNN μαθαίνει το βέλτιστο κατώφλι b_j για την κατηγορία c_j από το ότι παράγει την καλύτερη απόδοση στα κείμενα επικύρωσης. Το νέο κείμενο ανατίθεται σε αυτές τις κατηγορίες που

έχουν βαθμολογία πιθανότητας μεγαλύτερο από ένα προκαθορισμένο κατώφλι.

3.5.4 Parameter based Κατηγοριοποιητές

Για τους parameter based κατηγοριοποιητές, τα δείγματα εκπαίδευσης χρησιμοποιούνται για την εκτίμηση των παραμέτρων μιας κατανομής πιθανότητας. Ο αλγόριθμος Naive Bayes είναι ένα παράδειγμα αυτής της κατηγορίας.

Naive Bayes

Ο αλγόριθμος κατηγοριοποίησης naive Bayes, όπως περιγράφεται στις [Mit97, Joa97], εκπαιδεύεται με την χρήση δειγμάτων εκπαίδευσης για την εκτίμηση της πιθανότητας κάθε κατηγορίας δοθέντος ενός νέου κειμένου $d^{'}$, η οποία γράφεται:

$$Pr(c_j|d^{'}) = \frac{Pr(c_j) \cdot Pr(d^{'}|c_j)}{Pr(d^{'})} \tag{3.37}$$

Ο παρονομαστής στην παραπάνω εξίσωση δεν διαφέρει για τις διάφορες κατηγορίες και μπορεί να παραλειφθεί. Ο naive Bayes κάνει την υπόθεση της ανεξαρτησίας των χαρακτηριστικών σε σκοπό να εκτιμήσει την πιθανότητα:

$$Pr(c_j|d^{'}) = Pr(c_j) \prod_{f_i \in d^{'}} Pr(f_i|c_j) \tag{3.38}$$

όπου $Pr(c_j)$ είναι η αναλογία των δειγμάτων εκπαίδευσης στην κατηγορία c_j και f_i είναι ένα χαρακτηριστικό που βρίσκεται στο κείμενο $d^{'}$. Μαι εκτίμηση της $Pr(f_i|c_j)$ δίνεται από την σχέση [AE99]:

$$\tilde{Pr}(f_i|c_j) = \frac{1 + N_{ij}}{M + \sum_{k=1} M N_{kj}} \tag{3.39}$$

όπου N_{ij} είναι ο αριθμός των εμφανίσεων του χαρακτηριστικού f_i μέσα στα κείμενα της κατηγορίας c_j και M είναι ο συνολικός αριθμός των χαρακτηριστικών στο σύνολο εκπαίδευσης. Η κατηγορία με την μέγιστη τιμή της $P(c_j|d^{'})$ είναι η επιθυμητή κατηγορία για το κείμενο $d^{'}$. Διάφορες εργασίες κάνουν χρήση του αλγορίθμου naive Bayes για κατηγοριοποίηση κειμένων όπως οι [Joa98, Yan99] αλλά και άλλες. Επειδή όμως η υπόθεση που κάνει περί της ανεξαρτησίας χαρακτηριστικών δεν είναι ρεαλιστική για κείμενα, ο naive Bayes έχει δείξει κακή απόδοση σε μερικές από αυτές όπως για παράδειγμα στην [Yan99].

Κεφάλαιο 4

Σχεδιασμός και Ανάπτυξη Μεθόδου Κατηγοριοποίησης

"Within every difficulty lies opportunity."

– Ανώνυμος

Στο παρόν κεφάλαιο παρουσιάζουμε την διαδικασία κατασκευής του κατηγοριοποιητή ιστοσελίδων. Αρχικά, περιγράφουμε συνοπτικά το περιβάλλον εργασίας μας και την διαδικασία που ακολουθήσαμε για την δημιουργία του συνόλου δεδομένων μας. Έπειτα, παρουσιάζουμε τον τρόπο αναπαράστασης των ιστοσελίδων που επιλέξαμε και τα διαθέσιμα χαρακτηριστικά που αυτός υποδεικνύει. Στη συνέχεια, περιγράφουμε την διαδικασία επιλογής χαρακτηριστικών με σκοπό την μείωση της διαστατικότητας του προβλήματος κατηγοριοποίησης. Τέλος, παρουσιάζουμε την διαδικασία επιλογής αλγορίθμου κατηγοριοποίησης.

4.1 Περιβάλλον Εργασίας

Από τα διάφορα πακέτα λογισμικού εξόρυξης δεδομένων (data mining) που είναι διαθέσιμα, επιλέξαμε ως περιβάλλον εργασίας μας το WEKA[1] (Waikato Environment for Knowledge Analysis). Το WEKA παρέχει μια συλλογή από αλγορίθμους μηχανικής μάθησης οι οποίοι μπορούν να εφαρμοστούν απευθείας πάνω σε ένα σύνολο δεδομένων ή να κληθούν από κώδικα Java. Πιο συγκεκριμένα, το WEKA περιέχει εργαλεία για προεπεξεργασία δεδομένων (data pre-processing), κατηγοριοποίηση (classification), παλινδρόμηση (regres-

[1] http://www.cs.waikato.ac.nz/ml/weka/

sion), συσταδοποίηση (clustering), κανόνες συσχέτισης (association rules), και οπτικοποίηση (visualization). Επίσης είναι κατάλληλο για ανάπτυξη νέων αλγορίθμων μηχανικής μάθησης.

4.2 Σύνολο Δεδομένων

Για την εκπαίδευση και τον έλεγχο του κατηγοριοποιητή ιστοσελίδων απαιτείται ένα σύνολο από σελίδες επισημειωμένες με τον σκοπό αναζήτησης που ικανοποιούν. Επειδή ένα τέτοιο σύνολο δεδομένων δεν είναι διαθέσιμο, για την δημιουργία του, επιλέξαμε τα 100 πιο δημοφιλή ερωτήματα τον χρόνο που μας πέρασε[2] και τα υποβάλαμε σε μια εμπορική μηχανή αναζήτησης (Google). Από τα αποτελέσματα του κάθε ερωτήματος κρατήσαμε τα 20 πρώτα, δημιουργώντας έτσι ένα αρχικό σύνολο δεδομένων ~ 2000 σελίδων[3] τις οποίες και επισημειώσαμε χειρωνακτικά ως informational, navigational και transactional με την βοήθεια λογισμικού[4] που αναπτύχθηκε για τον σκοπό αυτό.

Το παραπάνω σύνολο δεδομένων περιέχει σε διαφορετική αναλογία τους τρεις τύπους σελίδων, κάτι που δεν είναι επιθυμητό διότι τα αποτελέσματα της κατηγοριοποίησης θα είναι πολωμένα στην κυρίαρχη κατηγορία. Για να αποφύγουμε αυτό το φαινόμενο, πραγματοποιήσαμε *δειγματοληψία* (*sampling*) στα δεδομένα έτσι ώστε να δημιουργήσουμε ένα καινούριο σύνολο σελίδων, το οποίο θα περιέχει δεδομένα και από τις τρεις κατηγορίες στην ίδια αναλογία. Πιο συγκεκριμένα, επιλέξαμε 340 στιγμιότυπα από κάθε κατηγορία (informational, navigational και transactional) καταλήγοντας έτσι σε ένα σύνολο δεδομένων 1020 στιγμιοτύπων.

4.3 Χαρακτηριστικά

Όπως παρουσιάστηκε και στο προηγούμενο κεφάλαιο (Κεφάλαιο 3), τα χαρακτηριστικά πάνω στα οποία βασίζεται ένας κατηγοριοποιητής ιστοσελίδων αντλούνται από τις διάφορες αναπαραστάσεις της σελίδας. Στην παρούσα εργασία, δεν βασιζόμαστε αποκλειστικά σε μια συγκεκριμένη αναπαράσταση, αλλά προσπαθούμε να εκμεταλλευτούμε όλες τις διαθέσιμες πηγές πληροφορίας της σελίδας και συγκεκριμένα το κείμενό της, την HTML δομή της, τους συνδέσμους της και το URL της.

[2] http://www.google.com/intl/en/press/zeitgeist2010/
[3] Πολύτιμο εργαλείο σε αυτήν την προσπάθεια η βιβλιοθήκη xgoogle για python του P. Krumins [Kru09].
[4] Πολύτιμη συνεισφορά της υποψήφιας Διδάκτωρ Π. Τζέκου

Χαρακτηριστικά Κειμένου

Η bag-of-terms αναπαράσταση είναι επιθυμητή στην θεματική κατηγοριοποίηση καθώς οι φράσεις ενός κειμένου μπορούν να αποκαλύψουν το θέμα του. Επειδή όμως οι φράσεις αυτές ποικίλουν ανάλογα με το θέμα, εισάγουν υψηλή διαστατικότητα στο πρόβλημα, καθώς πρέπει να δοθεί σαν είσοδος στον κατηγοριοποιητή ολόκληρο το διάνυσμα του κειμένου προκειμένου να μην παραληφθεί κάποια από αυτές τις φράσεις.

Από την άλλη, το κείμενο αυτό καθαυτό δεν είναι αντιπροσωπευτικό του σκοπού αναζήτησης, καθώς αυτός δεν περιγράφεται από φράσεις του κειμένου. Συγκεκριμένες λέξεις όμως που εμφανίζονται με μεγάλη συχνότητα σε ένα κείμενο μπορούν να αποκαλύψουν τον σκοπό αναζήτησης που εξυπηρετεί μια σελίδα. Για παράδειγμα οι λέξεις $blog$, $forum$, $homepage$, $site$, $webpage$, $website$, $welcome$ είναι ενδεικτικές για σελίδες που εξυπηρετούν navigational σκοπούς. Αντίστοιχα, οι λέξεις $availability$, $basket$, $cart$, $download$, $products$, $quantity$, $shipping$ είναι ενδεικτικές για σελίδες που εξυπηρετούν transactional σκοπούς.

Έτσι, ορίζουμε αρχικά ως χαρακτηριστικά του κειμένου τις συχνότητες εμφάνισης των παραπάνω λέξεων. Στο σημείο αυτό, πρέπει να αναφερθεί πως δεν υπάρχουν λέξεις που χαρακτηρίζουν σελίδες με informational σκοπό καθώς αυτές μπορούν να έχουν κείμενο με οποιοδήποτε περιεχόμενο. Η έκταση του κειμένου μιας σελίδας όμως μπορεί να είναι χρήσιμη στην διάκριση μεταξύ informational και των υπόλοιπων σελίδων καθώς οι πρώτες τείνουν να έχουν εκτενέστερο κείμενο. Έτσι, ορίζουμε τα εξής χαρακτηριστικά:

- το μέγεθος του κειμένου της σελίδας σε χαρακτήρες ($text\_length$)
- το πλήθος των λέξεων της σελίδας (num_of_words)
- το πλήθος των προτάσεων της σελίδας ($num\_of\_sentences$)

Δομικά Χαρακτηριστικά

Η δομή HTML μιας σελίδας μπορεί να βοηθήσει στην κατηγοριοποίηση με βάση τον σκοπό αναζήτησης, λαμβάνοντας υπόψιν όχι τους όρους που περικλείονται ανάμεσα στις HTML ετικέτες, αλλά τις ετικέτες αυτές καθαυτές. Πιο συγκεκριμένα, η δομή HTML μιας σελίδας είναι χρήσιμη στην διάκριση μεταξύ σελίδων με transactional σκοπό και των υπόλοιπων σελίδων, καθώς οι πρώτες τείνουν να περιέχουν σε μεγαλύτερη ποσότητα multimedia περιεχόμενο (εικόνες, video, κτλ), φόρμες, κουμπιά και άλλα παρόμοια χαρακτηριστικά. Τα χαρακτηριστικά αυτά εντοπίζονται εύκολα από τις HTML ετικέτες μιας σελίδας και για τον λόγο αυτό, ορίζουμε ως δομικά χαρακτηριστικά τις απόλυτες και σχετικές συχνότητες

Ετικέτα	**Ορισμός**
a	anchor
applet	embedded applet
audio	ήχος στην HTML 5
embed	embedded object
iframe	πλαίσιο που περιέχει κάποιο έγγραφο
img	εικόνα
input	πλαίσιο εισόδου χαρακτήρων
object	αντικείμενα όπως εικόνες, ήχος, βίντεο, Java applets, ActiveX, PDF, και Flash
video	βίντεο στην HTML 5
button	κουμπί
form	HTML φόρμα για εισαγωγή χαρακτήρων
map	χάρτης στην πλευρά του χρήστη που μπορεί να δέχεται clicks
select	drop-down λίστα επιλογής
textarea	πλαίσιο κειμένου

Πίνακας 4.1: Ορισμοί ετικετών HTML.

με τις οποίες εμφανίζονται οι αντίστοιχες HTML ετικέτες μέσα σε μία σελίδα. Οι ετικέτες αυτές παρουσιάζονται στον Πίνακα 4.1.

Χαρακτηριστικά των Συνδέσμων

Το anchortext των σελίδων του συνόλου δεδομένων μας δεν είναι διαθέσιμο εξαιτίας του τρόπου με τον οποίο συλλέχθηκαν οι σελίδες. Δεν είναι όμως μόνο το anchortext η αποκλειστική πληροφορία που μπορούν να συνεισφέρουν οι σύνδεσμοι μιας σελίδας. Σελίδες με transactional σκοπό τείνουν να έχουν μεγαλύτερο ποσοστό συνδέσμων από τις υπόλοιπες και για να εκμεταλλευτούμε αυτήν τους την ιδιότητα, ορίζουμε τις απόλυτες και σχετικές συχνότητες των συνδέσμων που δείχνουν στην ίδια (in_links και in_links_rel) ή σε άλλες σελίδες (out_links και out_links_rel).

Χαρακτηριστικά του URL

Το url δίνει πολλές σημαντικές πληροφορίες για τη σελίδα καθώς στην πλειονότητα των περιπτώσεων είναι μια εξαιρετικά συνοπτική και συμπαγής περιγραφή του σκοπού της. Επίσης διαισθητικά οι χρήστες τείνουν να αντλούν πληροφορίες από το url, πχ για την αξιοπιστία ή το περιεχόμενο και τη μορφή μιας ιστοσελίδας, πριν επιλέξουν το αποτέλεσμα. Για να επιβεβαιωθεί η διαίσθηση αυτή από τον αλγόριθμο κατηγοριοποίησης ορίζουμε τα εξής χαρακτηριστικά:

- το βάθος του url (*url_depth*)
- το μέγεθος του url (*url_length*)
- την εμφάνιση των όρων του ερωτήματος στο domain (*occ_in_domain*)
- την εμφάνιση των όρων του ερωτήματος στο path του url (*occ_in_url_path*)

4.4 Επιλογή Χαρακτηριστικών

Τα χαρακτηριστικά είναι 53 στο σύνολο και απαιτείται επιλογή κάποιων από αυτά. Όπως είδαμε και στο προηγούμενο κεφάλαιο (Κεφάλαιο 3), η μείωση του αριθμού των χαρακτηριστικών μπορεί όχι μόνο να βοηθήσει στην επιτάχυνση του χρόνου εκτέλεσης ενός αλγορίθμου, αλλά επίσης μπορεί να βοηθήσει να αποφευχθεί το «θάψιμο» ενός αλγορίθμου σε ένα πλήθος από χαρακτηριστικά, όταν μόλις μερικά από αυτά είναι απαραίτητα για την κατασκευή ενός καλού μοντέλου.

Η διαδικασία της επιλογής χαρακτηριστικών περιλαμβάνει την εξέταση όλων των πιθανών συνδυασμών χαρακτηριστικών στα δεδομένα ώστε να βρεθεί εκείνο το υποσύνολο των χαρακτηριστικών που θα είναι βέλτιστο για πρόβλεψη (prediction). Στο WEKA, αυτό επιτυγχάνεται ορίζοντας έναν «αξιολογητή» γνωρισμάτων[5] (attribute evaluator) και μία μέθοδο αναζήτησης (search method). Ο evaluator καθορίζει ποια μέθοδος χρησιμοποιείται για την ανάθεση μιας «αξίας» σε κάθε υποσύνολο γνωρισμάτων ενώ η μέθοδος αναζήτησης καθορίζει το είδος της αναζήτησης που εκτελείται.

Υπάρχουν δύο διαφορετικοί τύποι evaluators στο WEKA [BFH+09]:

- evaluators ενός γνωρίσματος (single attribute evaluators): εφαρμόζονται πάνω σε ένα γνώρισμα και χρησιμοποιούνται αποκλειστικά σε συνδυασμό με την μέθοδο αναζήτησης Ranker. Σε αυτήν την κατηγορία αλγορίθμων συμπεριλαμβάνονται οι εξής μέθοδοι: ChiSquared, CostSensitive, Filtered, GainRatio, InfoGain, LatentSemanticAnalysis, OneR, PrincipalComponents, ReliefF, SVM και SymmetricalUncert.
- evaluators υποσυνόλου γνωρισμάτων (attribute subset evaluators): εφαρμόζονται πάνω σε υποσύνολα των γνωρισμάτων του συνόλου δεδομένων και μπορούν να χρησιμοποιηθούν σε συνδυασμό με διάφορες μεθόδους αναζήτησης. Σε αυτήν την κατηγορία αλγορίθμων συμπεριλαμβάνονται οι εξής μέθοδοι: Cfs, Classifier, Consistency, CostSensitive, Filtered και Wrapper.

[5] Οι όροι χαρακτηριστικό (feature) και γνώρισμα (attribute) χρησιμοποιούνται εναλλάξ στο κείμενο.

Επιπρόσθετα, οι μέθοδοι αναζήτησης που παρέχει το WEKA είναι οι εξής: BestFirst, ExhaustiveSearch, GeneticSearch, GreedyStepwise, LinearForwardSelection, RaceSearch, RandomSearch, Ranker, RankSearch, ScatterSearchV1 και SubsetSizeForwardSelection.

Για να εξετάσουμε την επίδραση του αλγορίθμου επιλογής χαρακτηριστικών στα αποτελέσματα της κατηγοριοποίησης, διεξήγαμε το παρακάτω πείραμα: δοκιμάσαμε διάφορους συνδυασμούς από attribute evaluators και μεθόδους αναζήτησης υπολογίζοντας την απόδοση κατηγοριοποίησης με συγκεκριμένο κατηγοριοποιητή (J48). Για να διασφαλιστεί η εγκυρότητα των αποτελεσμάτων χρησιμοποιήσαμε την τεχνική του cross-validation.

Στην τεχνική του cross-validation, το αρχικό σύνολο δεδομένων διαμερίζεται τυχαία σε n υποσύνολα. Από τα n υποσύνολα, ένα μοναδικό υποσύνολο χρησιμοποιείται ως σύνολο δοκιμής (test set) για την εξέταση του μοντέλου και τα εναπομείναντα $n-1$ υποσύνολα χρησιμοποιούνται ως δεδομένα εκπαίδευσης (training data). Η διαδικασία του cross-validation επαναλαμβάνεται n φορές (folds), με καθένα από τα n υποσύνολα να χρησιμοποιούνται ακριβώς μία φορά ως δεδομένα δοκιμής. Τα n αποτελέσματα από τις διαφορετικές εκτελέσεις συνδυάζονται (συνήθως ως μέσος όρος) για να παράγουν μια μοναδική εκτίμηση. Το πλεονέκτημα της μεθόδου είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται ταυτόχρονα και για εκπαίδευση αλλά και για επικύρωση. Η πιο κοινή τιμή για το n είναι 10 την οποία και υιοθετήσαμε.

Η απόδοση της κατηγοριοποίησης για τους διαφορετικούς συνδυασμούς αξιολογητών γνωρισμάτων και μεθόδων αναζήτησης που δοκιμάσαμε παρουσιάζεται στον Πίνακα 4.2.

Αρχικά, παρατίθεται η απόδοση της κατηγοριοποίησης χωρίς καμία επιλογή χαρακτηριστικών ώστε να γίνει φανερή η ανάγκη μείωσής τους. Στο δεύτερο τμήμα του Πίνακα 4.2 παρουσιάζεται η απόδοση της κατηγοριοποίησης για evaluators ενός γνωρίσματος. Παρατηρούμε ότι, με εξαίρεση τις μεθόδους LatentSemanticAnalysis και PrincipalComponents, οι οποίες είναι μέθοδοι εξαγωγής χαρακτηριστικών όπως αυτές ορίστηκαν στο Κεφάλαιο 3, η απόδοση της κατηγοριοποίησης των υπολοίπων μεθόδων είναι μεγαλύτερη του 80% και πολύ κοντά η μία στην άλλη. Για να το επιτύχουν αυτό όμως, δεν χρησιμοποιούν όλες οι μέθοδοι τον ίδιο αριθμό από γνωρίσματα. Έτσι, οι μέθοδοι ChiSquared, InfoGain και OneR χρησιμοποιώντας μόλις 5 χαρακτηριστικά σε σχέση με την ReliefF η οποία χρησιμοποιεί 14.

Στα υπόλοιπα τρία τμήματα του Πίνακα 4.2 παρουσιάζεται η απόδοση της κατηγοριοποίησης για evaluators υποσυνόλου γνωρισμάτων.

Αρχικά, παρατίθεται η απόδοση μεθόδων βασισμένων σε κατηγοριοποιητή (επιλέξαμε ξανά τον J48). Παρατηρούμε ότι η απόδοση των μεθόδων αυτών δεν είναι ικανοποιητική

Αξιολογητής Γνωρισμάτων	**Μέθοδος Αναζήτησης**	**Γνωρίσματα**	**Απόδοση**
χωρίς	χωρίς	53	82.06%
ChiSquared	Ranker	5	83.33%
GainRatio	Ranker	10	83.14%
InfoGain	Ranker	5	83.33%
LatentSemanticAnalysis	Ranker	1	48.33%
OneR	Ranker	5	83.33%
PrincipalComponents	Ranker	17	70.68%
ReliefF	Ranker	14	83.92%
SymmetricalUncert	Ranker	9	83.53%
Classifier (J48)	RankSearch	48	82.06%
Wrapper (J48)	RankSearch	36	82.16%
CFS	BestFirst	11	83.73%
CFS	GeneticSearch	15	85.10%
CFS	GreedyStepwise	11	83.73%
CFS	LinearForwardSelection	11	83.73%
CFS	RankSearch	11	83.73%
CFS	ScatterSearch	11	83.73%
Consistency	BestFirst	17	82.35%
Consistency	GeneticSearch	22	80.69%
Consistency	GreedyStepwise	17	82.35%
Consistency	LinearForwardSelection	17	82.84%
Consistency	RankSearch	37	82.26%
Consistency	ScatterSearch	14	83.04%

Πίνακας 4.2: Απόδοση κατηγοριοποίησης J48 για αλγορίθμους επιλογής χαρακτηριστικών.

καθώς από την μία αποκλείουν ελάχιστα χαρακτηριστικά και από την άλλη δεν βελτιώνουν την απόδοση κατηγοριοποίησης.

Έπειτα, παρουσιάζεται η απόδοση της μεθόδου CFS (Correlation-based Feature Selection) σε συνδυασμό με διάφορες μεθόδους αναζήτησης. Παρατηρούμε ότι η απόδοση της μεθόδου είναι σταθερή και ανεξάρτητη της μεθόδου αναζήτησης με εξαίρεση την περίπτωση που χρησιμοποιείται σε συνδυασμό με την μέθοδο GreedyStepwise. Σε αυτήν την περίπτωση ο αλγόριθμος CFS επιτυγχάνει την καλύτερη απόδοση η οποία είναι και η μεγαλύτερη απόδοση που επιτεύχθηκε κατά την διενέργεια του πειράματος.

Τέλος, στο τελευταίο τμήμα του Πίνακα 4.2 παρουσιάζεται η απόδοση της μεθόδου Consistency σε συνδυασμό με διάφορες μεθόδους αναζήτησης. Παρατηρούμε ότι η μέθοδος αυτή παρουσιάζει ελάχιστη διακύμανση στην απόδοσή της ανάλογα με την μέθοδο αναζήτησης που χρησιμοποιείται αλλά σε κάθε περίπτωση, αυτή είναι μικρότερη της αντίστοιχης απόδοση του αλγορίθμου CFS και απαιτεί μεγαλύτερο αριθμό χαρακτηριστικών.

url_depth	*url_length*	*occ_in_domain*	*occ_in_url_path*
link_tags	*button_tags*	*out_link_tags*	*link_tags_rel*
embeded_tags_rel	*iframe_tags_rel*	*image_tags_rel*	*form_tags_rel*
num_of_sentences	*cart*	*site*	

Πίνακας 4.3: Χαρακτηριστικά που επιλέχθηκαν.

Όπως γίνεται φανερό, καμία μέθοδος δεν υπερτερεί απόλυτα και στην απόδοση και στο πλήθος χαρακτηριστικών που χρησιμοποιεί για να την επιτύχει. Η τελική επιλογή συνδυασμού αξιολογητή γνωρισμάτων και μεθόδου αναζήτησης μπορεί να γίνει με βάση διάφορα κριτήρια όπως μέγιστης απόδοσης ή χρήσης ελάχιστων χαρακτηριστικών. Επειδή ο αλγόριθμός μας απαιτεί όσο το δυνατόν μεγαλύτερη ακρίβεια χωρίς να ενδιαφέρεται για το πλήθος των χαρακτηριστικών, επιλέγουμε τον συνδυασμό μεθόδων CFS και GeneticSearch.

Ο συγκεκριμένος αξιολογητής χαρακτηριστικών επέλεξε τα χαρακτηριστικά που παρουσιάζονται στον Πίνακα 4.3. Παρατηρούμε ότι την πρώτη γραμμή του Πίνακα 4.3 καταλαμβάνουν χαρακτηριστικά που βασίζονται στο url της σελίδας, τις επόμενες δύο γραμμές χαρακτηριστικά που βασίζονται στην HTML δομή της, ενώ την τελευταία γραμμή καταλαμβάνουν κειμενικά χαρακτηριστικά.

Από τον Πίνακα 4.3, γίνεται φανερό ότι επελέχθησαν κυρίως χαρακτηριστικά σχετικά με τη δομή των ιστοσελίδων και όχι με τα κειμενικά δεδομένα. Αυτό είναι αναμενόμενο ως ένα βαθμό καθώς οι δείκτες που εξάγονται από τα κειμενικά δεδομένα τείνουν να εκφράζουν το θέμα της ιστοσελίδας, ενώ τα στοιχεία της δομής μπορούν να περιγράψουν καλύτερα το σκοπό που αυτή εξυπηρετεί. Επίσης, παρατηρούμε ότι τα χαρακτηριστικά που εντοπίζονται στο url της σελίδας έχουν σημαντική θέση ανάμεσα στα επιλεγμένα. Αυτό επιβεβαιώνει τη διαίσθησή που διατυπώθηκε προηγουμένως, ότι το url της ιστοσελίδας είναι ισχυρή ένδειξη για το αν μπορεί να ικανοποιήσει τον σκοπό αναζήτησης του χρήστη.

4.5 Επιλογή Κατηγοριοποιητή

Έχοντας επιλέξει τα χαρακτηριστικά πάνω στα οποία θα βασιστεί η κατηγοριοποίηση, μένει να επιλέξουμε τον ίδιο τον αλγόριθμο κατηγοριοποίησης. Για τον σκοπό αυτό, εκτελέσαμε μια σειρά πειραμάτων με σκοπό να διερευνήσουμε ποιος αλγόριθμος κατηγοριοποίησης παρουσιάζει την μεγαλύτερη ακρίβεια. Πειραματιστήκαμε με διάφορους αλγορίθμους κατηγοριοποίησης και συγκεκριμένα:

Κατηγοριοποιητής	**Ακρίβεια Κατηγοριοποίησης**
BayesNet	81.18%
NaiveBayes	77.65%
SimpleLogistic	83.04%
SMO	79.80%
IBk	82.16%
KStar	77.65%
JRip	82.06%
OneR	77.55%
PART	82.75%
Ridor	80.20%
FT	81.27%
J48	85.10%
RandomForest	86.08%
RandomTree	82.06%
SimpleCart	82.94%

Πίνακας 4.4: Ακρίβεια κατηγοριοποίησης για διάφορους αλγορίθμους.

- bayesian κατηγοριοποιητές: BayesNet και NaiveBayes
- κατηγοριοποιητές βασισμένους σε συναρτήσεις απόστασης: SimpleLogistic και SMO (Sequential Minimal Optimization)
- lazy κατηγοριοποιητές: IBk και KStar
- κατηγοριοποιητές βασισμένους σε κανόνες: JRip, OneR, PART και Ridor
- δέντρα απόφασης: FT, J48, RandomForest, RandomTree και SimpleCart

Η ακρίβεια της κατηγοριοποίησης για τους διάφορους αλγορίθμους που δοκιμάστηκαν παρουσιάζεται στον Πίνακα 4.4. Παρατηρούμε ότι χειρότερη επίδοση επιτυγχάνει ο αλγόριθμος OneR με ακρίβεια κατηγοριοποίησης 77.55% ενώ την καλύτερη επίδοση επιτυγχάνει ο αλγόριθμος RandomForest με ακρίβεια κατηγοριοποίησης 86.08%. Επίσης, παρατηρούμε ότι με εξαίρεση τα δέντρα απόφασης, υπάρχουν μεγάλες αποκλίσεις στην απόδοση των υπολοίπων αλγορίθμων με βάση την κατηγορία στην οποία ανήκουν. Η καλή απόδοση των δέντρων απόφασης και η μικρή απόκλιση μεταξύ τους οφείλεται ίσως στο γεγονός ότι κατά την επιλογή χαρακτηριστικών, η ακρίβεια κατηγοριοποίησης μετρήθηκε με βάση τον αλγόριθμο J48 που ανήκει στα δέντρα απόφασης. Εν τέλει, επιλέγουμε τον αλγόριθμο με την μεγαλύτερη απόδοση (RandomForest).

	inf	nav	trans
inf	265	4	71
nav	1	336	3
trans	68	5	267

Πίνακας 4.5: Confusion Matrix.

Στον Πίνακα 4.5 παρουσιάζεται το μητρώο «σύγχυσης» (confusion matrix) το οποίο μπορεί να δώσει περισσότερες πληροφορίες για την κατηγοριοποίηση. Όπως γίνεται φανερό από το confusion matrix, επιτυγχάνεται σχεδόν τέλεια ακρίβεια κατηγοριοποίησης για τις navigational σελίδες ($\sim 99\%$) κάτι που δεν συμβαίνει στην περίπτωση των informational και transactional σελίδων. Ένα σχετικά μεγάλο ποσοστό ($\sim 22\%$) των transactional σελίδων κατηγοριοποιούνται ως informational και το αντίστροφο. Συμπεραίνουμε λοιπόν ότι το 2-class πρόβλημα κατηγοριοποίησης μεταξύ informational και transactional σελίδων επιδέχεται μεγαλύτερης διερεύνησης.

Κεφάλαιο 5

Πειραματική Εφαρμογή και Αξιολόγηση Επαναδιατύπωσης Ερωτημάτων

"All life is an experiment. The more experiments you make the better."

– Ralph Waldo Emerson

Στο παρόν κεφάλαιο παρουσιάζουμε την μέθοδο επαναδιατύπωσης ερωτημάτων. Αρχικά, παραθέτουμε τις διάφορες τακτικές επαναδιατύπωσης ανάλογα με την πρόθεση του χρήστη. Στη συνέχεια, παρουσιάζουμε μια μέθοδο αξιολόγησης του αλγορίθμου επαναδιατύπωσης καθώς και το σύνολο δεδομένων πάνω στο οποίο εφαρμόσθηκε. Τέλος, παρουσιάζουμε τα αποτελέσματα της πειραματικής αξιολόγησης του αλγορίθμου επαναδιατύπωσης.

5.1 Επαναδιατύπωση Ερωτημάτων

Όπως είδαμε στο Κεφάλαιο 3, υπάρχουν δύο βασικές προσεγγίσεις στην επαναδιατύπωση ερωτημάτων: άμεση και έμμεση ανατροφοδότηση. Από την στιγμή που οι χρήστες δεν λαμβάνουν μέρος στην διαδικασία ανατροφοδότησης, έχουμε να κάνουμε με έμμεση ανατροφοδότηση η οποία όπως είδαμε χωρίζεται σε τοπική και καθολική ανάλυση. Η μέθοδος για την επαναδιατύπωση ερωτημάτων βασισμένη στον σκοπό αναζήτησης που παρουσιάζουμε εδώ ανήκει εξολοκλήρου στην τοπική ανάλυση.

Όπως φαίνεται και στην Εικόνα 3.2α', κατά την τοπική ανάλυση, ο χρήστης υποβάλει σε

μια μηχανή αναζήτησης ένα ερώτημα q για το οποίο λαμβάνει ένα σύνολο αποτελεσμάτων. Στη συνέχεια, επιλέγονται τα κορυφαία k αποτελέσματα (τυπικές τιμές του k είναι 10 και 20) τα οποία και ομαδοποιούνται. Από την ομαδοποίηση των κορυφαίων αποτελεσμάτων έπειτα, εξάγονται όροι για την επαναδιατύπωση του ερωτήματος.

Η μέθοδος της τοπικής ανάλυσης όπως και οι υπόλοιπες μέθοδοι επαναδιατύπωσης ερωτημάτων, αναπτύχθηκαν αρχικά για μεγάλες και στατικές συλλογές κειμένων. Στη συνέχεια, με την εμφάνιση του Παγκόσμιου Ιστού, εφαρμόστηκαν με μεγάλη επιτυχία σε αυτό το καινούριο περιβάλλον όπου κυριαρχούσαν τα κείμενα. Οι μέθοδοι αυτοί όπως επίσης είδαμε βελτιώνουν την ανάκληση (recall) καθώς επιστρέφουν επιπρόσθετα κείμενα στα αποτελέσματα ύστερα από την υποβολή του επαναδιατυπωμένου ερωτήματος.

Η συμπεριφορά των μεθόδων επαναδιατύπωσης και ιδιαίτερα της τοπικής ανάλυσης όπως μόλις παρουσιάστηκε, αναμένεται να είναι χρήσιμη στην επαναδιατύπωση ερωτημάτων με informational σκοπό. Καθώς ο χρήστης ζητάει πληροφορία η οποία βρίσκεται σε περισσότερες από μία σελίδες και δει κείμενα, η μέθοδος τοπικής ανάλυσης των Attar και Fraenkel [AF77] που παρουσιάσαμε στο Κεφάλαιο 3 παρουσιάζεται ιδιαίτερα ελκυστική.

Από την άλλη, ο Παγκόσμιος Ιστός στις μέρες μας δεν απαρτίζεται αποκλειστικά από κείμενα αλλά και από συνδέσμους, εικόνες, ήχο, βίντεο, αρχεία κτλ. Επίσης, ο σκοπός του χρήστη δεν είναι πάντα informational αλλά navigational ή transactional. Πρέπει λοιπόν να διερευνηθεί πως οι διαθέσιμες πηγές πληροφορίας μπορούν να αξιοποιηθούν κατά την διαδικασία της αναζήτησης και με ποιον σκοπό αναζήτησης σχετίζεται η καθεμία.

Στην κατεύθυνση ανάπτυξης μεθόδων επαναδιατύπωσης ερωτημάτων με βάση την navigational και την transactional πρόθεση του χρήστη, μπορούν να χρησιμοποιηθούν όροι που δεν προέρχονται από το κείμενο μιας σελίδας. Έχουμε ήδη αναφέρει ότι οι παραδοσιακές μέθοδοι επαναδιατύπωσης ερωτημάτων που βασίζονται σε κείμενα αναμένεται να είναι κατάλληλες για informational επαναδιατύπωση, διότι οι λέξεις του κειμένου αντικατοπτρίζουν το θέμα του. Αντίθετα, στην περίπτωση ερωτημάτων με navigational και transactional σκοπό μπορούν για την επαναδιατύπωση να δοκιμαστούν όροι που εξάγονται από το url των σελίδων, καθώς οι όροι αυτοί αναμένεται να σχετίζονται περισσότερο με τον σκοπό αναζήτησης. Άλλωστε, έχει παρατηρηθεί ότι οι χρήστες μπορούν να κρίνουν αν μια σελίδα ικανοποιεί τον σκοπό αναζήτησής τους από το url και μόνο.

Έτσι, στην περίπτωση των navigational ερωτημάτων μπορούν αρχικά να δοκιμαστούν όροι που εξάγονται από το domain του url μιας σελίδας, το ίδιο το domain, καθώς και οι καταλήξεις των url (.com, .net, κτλ). Αντίστοιχα, στην περίπτωση transactional ερωτημάτων μπορούν να δοκιμαστούν όροι που εξάγονται από ολόκληρο το url της σελίδας.

5.2 Μέθοδος Αξιολόγησης

Οι μέθοδοι επαναδιατύπωσης αξιολογούνται συνήθως με κλασσικές μετρικές ανάκτησης πληροφορίας όπως η ανάκληση (recall) και η ακρίβεια (precision). Για την μέτρηση των μεγεθών αυτών όμως, είναι απαραίτητη η ύπαρξη κάποιας συλλογής κειμένων, ένα σύνολο από ερωτήματα δοκιμής (test queries) καθώς και οι αντίστοιχες *αποφάσεις σχετικότητας* (*relevance judgements*) οι οποίες υποδεικνύουν ποια κείμενα είναι σχετικά με το κάθε ερώτημα. Δυστυχώς, μια τέτοια συλλογή κειμένων δεν είναι διαθέσιμη, ενώ η κατασκευή της είναι ιδιαίτερα ακριβή και επίπονη διαδικασία.

Για τον λόγο αυτό, σχεδιάσαμε το παρακάτω πείραμα: ορίσαμε ένα σύνολο από 54 ερωτήματα δοκιμής τα οποία και υποβάλαμε σε μια εμπορική μηχανή αναζήτησης (Google). Τα ερωτήματα αυτά είναι τα πιο δημοφιλή της φετινής χρονιάς[1] και ταυτόχρονα είναι διαφορετικά από εκείνα της περσινής που απαρτίζουν το σύνολο δεδομένων του προηγούμενου κεφαλαίου. Από τα αποτελέσματα του κάθε ερωτήματος, επαναδιατυπώσαμε τα ερωτήματα για κάθε έναν από τους τρεις σκοπούς αναζήτησης. Στη συνέχεια, υποβάλαμε τα επαναδιατυπωμένα ερωτήματα στην μηχανή αναζήτησης και κατηγοριοποιήσαμε τα δέκα κορυφαία αποτελέσματα με την ίδια μέθοδο κατηγοριοποίησης. Τέλος, μετρήσαμε την διαφορά μεταξύ των αποτελεσμάτων του αρχικού και του επαναδιατυπωμένου ερωτήματος.

Ο σκοπός του παραπάνω πειράματος εξηγείται διαισθητικά ως εξής: γνωρίζουμε ότι οι τεχνικές επαναδιατύπωσης ερωτημάτων αυξάνουν την ανάκληση, ένα μέγεθος το οποίο όμως δεν μπορούμε να μετρήσουμε άμεσα. Μπορούμε όμως να μετρήσουμε πόσα νέα αποτελέσματα που ικανοποιούν τον ίδιο σκοπό αναζήτησης εμφανίζονται ανάμεσα στα κορυφαία. Έτσι, προσεγγίζουμε υπό μια έννοια την ανάκληση μετρώντας τα νέα αποτελέσματα που ικανοποιούν τον ίδιο σκοπό αναζήτησης. Στην επόμενη ενότητα, παρουσιάζουμε τα αποτελέσματα του πειράματος.

5.3 Πειραματικά Αποτελέσματα

Αρχικά, εστιάζουμε στην περίπτωση των ερωτημάτων με informational σκοπό. Για να διερευνήσουμε αν όντως οι παραδοσιακές μέθοδοι επαναδιατύπωσης ερωτημάτων είναι κατάλληλες για την informational επαναδιατύπωση, πραγματοποιούμε το ακόλουθο πείραμα. Από τα αποτελέσματα της αναζήτησης ενός ερωτήματος, επαναδιατυπώνουμε το ερώτημα με την μέθοδο των Attar και Fraenkel [AF77] τρεις φορές, αντλώντας κάθε φορά

[1] http://www.googlezeitgeist.com/en/

inform.	$+21.76\%$
navig.	-27.31%
trans.	$+10.19\%$

(α′) Informational επαναδιατύπωση από informational σελίδες

inform.	$+20.83\%$
navig.	-22.69%
trans.	$+5.55\%$

(β′) Informational επαναδιατύπωση από navigational σελίδες

inform.	$+12.96\%$
navig.	-20.37%
trans.	$+11.57\%$

(γ′) Informational επαναδιατύπωση από transactional σελίδες

Πίνακας 5.1: Μεταβολή στα αποτελέσματα της κατηγοριοποίησης των αποτελεσμάτων ύστερα από την υποβολή του επαναδιατυπωμένου ερωτήματος με την μέθοδο των Attar και Fraenkel [AF77] όταν αυτή βασίζεται σε informational σελίδες, σε navigational και σε transactional.

inform.	-7.87%
navig.	-2.77%
trans.	$+10.65\%$

(α′) Επαναδιατύπωση με keywords που εξάγονται από τα domains και καταλήξεις

inform.	-2.77%
navig.	-20.37%
trans.	$+21.30\%$

(β′) Επαναδιατύπωση με keywords που εξάγονται από τα domains

inform.	-0.93%
navig.	-13.43%
trans.	$+18.06\%$

(γ′) Επαναδιατύπωση με ολόκληρα domains

inform.	-14.81%
navig.	$+0.46\%$
trans.	$+14.35\%$

(δ′) Επαναδιατύπωση με ολόκληρα domains και καταλήξεις

inform.	-7.41%
navig.	$+8.80\%$
trans.	-0.46%

(ε′) Επαναδιατύπωση με την λέξη *site*

Πίνακας 5.2: Μεταβολή στα αποτελέσματα της κατηγοριοποίησης των αποτελεσμάτων ύστερα από την υποβολή του navigational επαναδιατυπωμένου ερωτήματος.

όρους από μια κατηγορία ιστοσελίδων που ικανοποιούν ένα συγκεκριμένο σκοπό αναζήτησης. Στη συνέχεια, υποβάλουμε το επαναδιατυπωμένο ερώτημα στην μηχανή αναζήτησης και μετράμε την μεταβολή στα αποτελέσματα της αναζήτησης. Τα αποτελέσματα του πειράματος παρουσιάζονται στον Πίνακα 5.1.

Όπως γίνεται φανερό, η μέθοδος των Attar και Fraenkel [AF77] επιστρέφει περισσότερες informational σελίδες σε όλες τις περιπτώσεις. Ιδιαίτερα όταν αυτή αντλεί όρους από informational (Πίνακας 5.1α′) και navigational σελίδες (Πίνακας 5.1β′) η βελτίωση στα αποτελέσματα είναι της τάξης του 20%. Τα αποτελέσματα αυτά επιβεβαιώνουν την αρχική υπόθεση, ότι οι παραδοσιακές μέθοδοι επαναδιατύπωσης ερωτημάτων ευνοούν την informational επαναδιατύπωση καθώς αυτή βασίζεται στο κείμενο των σελίδων.

Στον Πίνακα 5.2 παρουσιάζεται η μεταβολή στα αποτελέσματα της κατηγοριοποίησης των αποτελεσμάτων ύστερα από την υποβολή του navigational επαναδιατυπωμένου ερωτήματος με διάφορους τρόπους. Όπως γίνεται φανερό από τους τέσσερις πρώτους πίνακες,

inform.	-12.96%
navig.	-14.81%
trans.	$+27.31\%$

(α′) Επαναδιατύπωση με keywords που εξάγονται από ολόκληρο το url και καταλήξεις

inform.	-7.87%
navig.	-21.76%
trans.	$+32.41\%$

(β′) Επαναδιατύπωση με keywords που εξάγονται από ολόκληρο το url

inform.	-23.15%
navig.	-19.44%
trans.	$+41.67\%$

(γ′) Επαναδιατύπωση με tokens από το url

inform.	-14.35%
navig.	-11.11%
trans.	$+24.07\%$

(δ′) Επαναδιατύπωση με tokens από το url και καταλήξεις

Πίνακας 5.3: Μεταβολή στα αποτελέσματα της κατηγοριοποίησης των αποτελεσμάτων ύστερα από την υποβολή του transactional επαναδιατυπωμένου ερωτήματος.

καμία μέθοδος που βασίζεται στο domain δεν βοηθάει στην ανάκτηση νέων navigational σελίδων, αντιθέτως, υπάρχει μια αύξηση των transactional σελίδων που στην περίπτωση της επαναδιατύπωσης με keywords που εξάγονται από τα domains, ξεπερνάει το 20% (Πίνακας 5.2β′). Τα αποτελέσματα αυτά προκαλούν έκπληξη και έρχονται σε αντίθεση με αυτά που αναμενόταν. Σύμφωνα με τα χαρακτηριστικά της κατηγοριοποίησης, keywords που εξάγονται από το domain αναμενόταν να βοηθήσουν στην ανάκτηση επιπλέον navigational σελίδων. Μένει να εξακριβωθεί αν η ίδια τακτική επιφέρει τα ίδια αποτελέσματα στην περίπτωση που οι όροι εξάγονται από transactional ιστοσελίδες.

Στον Πίνακα 5.2ε′ παρουσιάζονται τα αντίστοιχα αποτελέσματα όταν χρησιμοποιείται ρητά επιπλέον του ερωτήματος ο όρος *site*. Ο όρος αυτός δεν επιλέχθηκε τυχαία αλλά προέκυψε ύστερα από επισκόπηση των χαρακτηριστικών της κατηγοριοποίησης. Ο αλγόριθμος επιλογής χαρακτηριστικών επέλεξε τον όρο *site* τον οποίο συμπεριλάβαμε στα αρχικά χαρακτηριστικά λόγω της διακριτικής του ικανότητας για τις navigational σελίδες. Όπως φαίνεται και από τον Πίνακα 5.2ε′ τα αποτελέσματα είναι ενθαρρυντικά καθώς ο όρος βοηθάει στην ανάκτηση επιπρόσθετων navigational σελίδων. Στην ίδια κατεύθυνση, θα μπορούσαν να χρησιμοποιηθούν στο μέλλον συνώνυμοι όροι που αντλούνται από έναν θησαυρό ή searchonyms του όρου.

Στον Πίνακα 5.3 παρουσιάζεται η μεταβολή στα αποτελέσματα της κατηγοριοποίησης των αποτελεσμάτων ύστερα από την υποβολή του transactional επαναδιατυπωμένου ερωτήματος με διάφορους τρόπους, μόνο που οι όροι δεν εξάγονται πλέον από το domain αλλά από ολόκληρο το url μιας σελίδας. Όπως γίνεται φανερό και από τους τέσσερις πίνακες,

τα αποτελέσματα είναι όλα θετικά και μάλιστα η βελτίωση που επιτυγχάνεται ξεπερνά το 40% (Πίνακας 5.3γ'). Η βελτίωση αυτή οφείλεται στο γεγονός ότι όροι που εξάγονται από το url είναι περιγραφικοί του transactional σκοπού. Δυστυχώς, δεν μπορέσαμε να δοκιμάσουμε κάποια τακτική που βασίζεται στο anchortext των σελίδων καθώς αυτό δεν ήταν διαθέσιμο εξαιτίας του τρόπου με τον οποίον συλλέχθηκαν οι σελίδες.

Κεφάλαιο 6

Συμπεράσματα

"A conclusion is simply the place where you got tired of thinking."

– Ανώνυμος

Στην παρούσα διπλωματική εργασία μελετήσαμε το πρόβλημα της επαναδιατύπωσης ερωτημάτων με βάση την πρόθεση του χρήστη. Η προσέγγιση που ακολουθήσαμε βασίζεται στην μέθοδο τοπικής ανάλυσης των αποτελεσμάτων της αναζήτησης καθώς και στην κατηγοριοποίηση ιστοσελίδων με βάση τον σκοπό αναζήτησης του χρήστη. Η συνεισφορά της παρούσας εργασίας συνοψίζεται στα εξής σημεία:

- *Σκοπός αναζήτησης*: Δείξαμε ότι είναι δυνατή η αποτύπωση του σκοπού αναζήτησης του χρήστη στο ερώτημα σε αντίθεση με τις έως τώρα προσεγγίσεις οι οποίες αποτυπώνουν στο ερώτημα την πληροφοριακή ανάγκη του χρήστη και όχι τον σκοπό αναζήτησής του.

- *Κατηγοριοποίηση ιστοσελίδων*: Δείξαμε ότι ο σκοπός αναζήτησης που ικανοποιεί μια ιστοσελίδα σε ένα δεδομένο ερώτημα μπορεί να αναγνωριστεί με βάση χαρακτηριστικά που δεν εξάγονται από το κείμενο της ιστοσελίδας αλλά από δομικά στοιχεία της καθώς και στοιχεία που εξάγονται από το url της.

- *Επαναδιατύπωση ερωτημάτων*: Δείξαμε ότι η επαναδιατύπωση ερωτημάτων ανάλογα με τον σκοπό αναζήτησης πρέπει να βασίζεται σε διαφορετικές πηγές πληροφορίας και συγκεκριμένα ότι, η επιλογή όρων από το κείμενο ευνοεί τις informational αναζητήσεις, η επιλογή όρων από το url βελτιώνει τις transactional αναζητήσεις ενώ οι navigational αναζητήσεις βελτιώνονται από προσθήκη συγκεκριμένων λέξεων στο ερώτημα.

Οι μελλοντικές επεκτάσεις τις παρούσας εργασίας μπορούν να κινηθούν σε δύο βασικές κατευθύνσεις. Η μια κατεύθυνση αφορά την κατηγοριοποίηση και η άλλη την επαναδιατύπωση ερωτήματος. Για την κατηγοριοποίηση, αρχικά μπορούν να εξετασθούν επιπλέον χαρακτηριστικά με σκοπό το χτίσιμο ενός αποδοτικότερου μοντέλου. Ένα τέτοιο χαρακτηριστικό θα μπορούσε να είναι για παράδειγμα το ζύγισμα των όρων ανάλογα με την ετικέτα της HTML στην οποία περιέχονται [Rib02]. Επίσης, χρειάζεται επιπλέον πειραματισμός με ένα μεγαλύτερο και πιο ετερογενές σύνολο δεδομένων και αν είναι δυνατόν, με κάποιο gold standard όπως τα σύνολα δεδομένων που χρησιμοποιήθηκαν στα TREC web tracks[1] 2010 και 2011. Όσον αφορά την επαναδιατύπωση ερωτημάτων, μπορούν να εξετασθούν διάφορες άλλες πηγές πληροφορίας για την εξαγωγή όρων ιδιαίτερα για τις navigational και τις transactional σελίδες, όπως για παράδειγμα το anchortext των σελίδων καθώς και το κείμενο που υπάρχει γύρω από αυτό. Επίσης, και αυτή η μεθοδολογία πρέπει να εφαρμοσθεί σε μια μεγάλη και στατική συλλογή όπου μπορούν να μετρηθούν μεγέθη όπως precision και recall.

[1] http://trec.nist.gov/data/webmain.html

Βιβλιογραφία

[ABD06] E. Agichtein, E. Brill, and S. Dumais. *Improving web search ranking by incorporating user behavior information.* In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 19–26. ACM, 2006.

[AE99] K. Aas and L. Eikvil. *Text categorization: A survey.* Technical report, Norwegian Computing Center, 1999.

[AF77] R. Attar and A. S. Fraenkel. *Local feedback in full-text retrieval systems.* Journal of the ACM, 24(3):398–417, 1977.

[BDO95] M. W. Berry, S. T. Dumais, and G.W. O'Brien. *Using linear algebra for intelligent information retrieval.* SIAM Review, 37:573–595, 1995.

[BFH+09] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse. *Weka manual (3.7.1)*, 2009.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Chapman & Hall, New York, NY, 1984.

[BGAPG09] D. J. Brenes, D. Gayo-Avello, and K. Pérez-González. *Survey and evaluation of query intent detection methods.* In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 1–7. ACM, 2009.

[BMS07] J. Bhogal, A. Macfarlane, and P. Smith. *A review of ontology based query expansion.* Information Processing and Management, 43:866–886, 2007.

[Bro02] A. Broder. *A taxonomy of web search.* SIGIR Forum, 36:3–10, 2002.

[BYCBGC06] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. *The intention behind web queries*. In *String Processing and Information Retrieval*, Lecture Notes in Computer Science, pages 98–109. Springer Berlin, 2006.

[BYRN11] R. Baeza-Yates and B. Ribeiro-Neto. *Relevance feedback and query expansion*. In *Modern Information Retrieval: The Concepts and Technology Behind Search*, pages 177–202. Pearson Higher Education, 2011.

[CCM+03] P. Calado, M. Cristo, E. Moura, N. Ziviani, B. Ribeiro-Neto, and M. A. Gonçalves. *Combining link-based and content-based methods for web document classification*. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, CIKM '03, pages 394–401. ACM, 2003.

[CD00] H. Chen and S. Dumais. *Bringing order to the web: automatically categorizing search results*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 145–152. ACM, 2000.

[CDG+07] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. *Know your neighbors: web spam detection using the web topology*. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 423–430. ACM, 2007.

[CGRU97] C. Chekuri, M. H. Goldwasser, P. Raghavan, and E. Upfal. *Web search using automatic classification*. In *In 6th International Conference on the World Wide Web*, 1997.

[CH89] K. W. Church and P. Hanks. *Word association norms, mutual information, and lexicography*. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*, ACL '89, pages 76–83. Association for Computational Linguistics, 1989.

[Coh02] W. W. Cohen. *Improving a page classifier with anchor extraction and link analysis*. In *Advances in Neural Information Processing Systems 15*, 2002.

[CV95] C. Cortes and V. Vapnik. *Support-vector networks*. Machine Learning, 20:273–297, 1995.

[CY92] C. J. Crouch and B. Yang. *Experiments in automatic statistical thesaurus construction*. In *Proceedings of the 15th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR '92, pages 77–88. ACM, 1992.

[CY05] B. Choi and Z. Yao. *Web page classification*. In *Foundations and Advances in Data Mining*, Studies in Fuzziness and Soft Computing, pages 221–274. Springer, 2005.

[DDL+90] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. *Indexing by latent semantic analysis*. Journal of the American Society of Information Science, 41:391–407, 1990.

[DDLH08] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. *Understanding the relationship between searchers' queries and information goals*. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 449–458. ACM, 2008.

[FÖ99] J. Fürnkranz. *Exploiting structural information for text classification on the www*. In *Proceedings of the 3rd International Symposium on Advances in Intelligent Data Analysis*, IDA '99, pages 487–498. Springer-Verlag, 1999.

[FÖ02] J. Fürnkranz. *Hyperlink ensembles: a case study in hypertext classification*. Information Fusion, 3(4):299–312, 2002.

[GA05] K. Golub and A. Ardö. *Importance of html structural elements and metadata in automated subject classification*. In *ECDL*, Lecture Notes in Computer Science, pages 368–378. Springer, 2005.

[GGM05] Z. Gyöngyi and H. Garcia-Molina. *Web spam taxonomy*. In *AIRWeb, 1st International Workshop on Adversarial Information Retrieval on the Web*, pages 39–47, 2005.

[GSS00] L. Galavotti, F. Sebastiani, and M. Simi. *Experiments on the use of feature selection and negative evidence in automated text categorization*. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '00, pages 59–68. Springer-Verlag, 2000.

[GTL+02] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, and Gary W. F.. *Using web structure for classifying and describing web pages*. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 562–569. ACM, 2002.

[Hav03] T. H. Haveliwala. *Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search*. IEEE Transanctions on Knowledge and Data Engineering, 15:784–796, 2003.

[JBS07] B. J. Jansen, D. L. Booth, and A. Spink. *Determining the user intent of web search engine queries*. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1149–1150. ACM, 2007.

[JBS08] B. J. Jansen, D. L. Booth, and A. Spink. *Determining the informational, navigational, and transactional intent of web queries*. Information Processing and Management, 44:1251–1266, 2008.

[JGP+05] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. *Accurately interpreting clickthrough data as implicit feedback*. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 154–161. ACM, 2005.

[JGP+07] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. *Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search*. ACM Transactions on Information Systems, 25, 2007.

[Joa97] T. Joachims. *A probabilistic analysis of the rocchio algorithm with tfidf for text categorization*. In *Proceedings of the 14th International Conference on Machine Learning*, ICML '97, pages 143–151. Morgan Kaufmann Publishers Inc., 1997.

[Joa98] T. Joachims. *Text categorization with suport vector machines: Learning with many relevant features*. In *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142. Springer-Verlag, 1998.

[Joa02] T. Joachims. *Optimizing search engines using clickthrough data*. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142. ACM, 2002.

[KÖ5] M. Käki. *Findex: search result categories help users when document ranking fails*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 131–140. ACM, 2005.

[Kan05] I. H. Kang. *Transactional query identification in web search*. In *Information Retrieval Technology*, Lecture Notes in Computer Science, pages 221–232. Springer Berlin / Heidelberg, 2005.

[KCN07] C. Kohlschütter, P. Chirita, and W. Nejdl. *Utility analysis for topically biased pagerank*. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 1211–1212. ACM, 2007.

[KJ97] R. Kohavi and G. H. John. *Wrappers for feature subset selection*. Artificial Intelligence, 97:273–324, 1997.

[KJHS10] A. Kathuria, B. J. Jansen, C. Hafernik, and A. Spink. *Classifying the user intent of web queries using k-means clustering*. Internet Research, 20(5):563–581, 2010.

[KK03] I.H. Kang and G. C. Kim. *Query type classification for web document retrieval*. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 64–71. ACM, 2003.

[KK04] I. H. Kang and G. C. Kim. *Integration of multiple evidences based on a query type for web search*. Information Processing and Management, 40:459–478, 2004.

[KL00] O. W. Kwon and J. H.k Lee. *Web page classification based on k-nearest neighbor approach*. In *Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages*, IRAL '00, pages 9–15. ACM, 2000.

[Kru09] P. Krumins. *Python library for google search*. website, March 2009. http://www.catonmat.net/blog/python-library-for-google-search/.

[ΚΣ96] Δ. Κολλερ και Μ. Σαημι. *Τοωαρδ Οπτιμαλ Φεατυρε Σελεςτιον*. Στο *Προςεεδινγς οφ τηε 13τη Ιντερνατιοναλ ὃνφερενςε ον Μαςηινε Λεαρνινγ (ΓΜΛ '96)*, σελίδες 284–292. Μοργαν Καυφμαννν, 1996.

[KS11] N. Kirtsis and S. Stamou. *Query reformulation for task-oriented web searches*. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on, 3:289–292, 2011.

[KSR04] S. B. Kim, H. C. Seo, and H. C. Rim. *Information retrieval using word senses: root sense tagging approach*. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 258–265. ACM, 2004.

[LLC05] U. Lee, Z. Liu, and J. Cho. *Automatic identification of user goals in web search*. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 391–400. ACM, 2005.

[LLHC07] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. *Improving weak ad-hoc queries using wikipedia asexternal corpus*. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 797–798. ACM, 2007.

[LLYM04] S. Liu, F. Liu, C. Yu, and W. Meng. *An effective approach to document retrieval via utilizing wordnet and recognizing phrases*. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 266–272. ACM, 2004.

[MG99] D. Mladenic and M. Grobelnik. *Feature selection for unbalanced class distribution and naive bayes*. In *Proceedings of the 16th International Conference on Machine Learning*, ICML '99, pages 258–267. Morgan Kaufmann Publishers Inc., 1999.

[Mit97] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., 1997.

[MWN07] D. N. Milne, I. H. Witten, and D. M. Nichols. *A knowledge-based search engine powered by wikipedia*. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, CIKM '07, pages 445–454. ACM, 2007.

[MzES04] S. Meyer zu Eißen and B. Stein. *Genre classification of web pages: User study and feasibility analysis*. In *KI 2004: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, pages 256–269. Springer, 2004.

[NDQ06] L. Nie, B. D. Davison, and X. Qi. *Topical link analysis for web search*. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 91–98. ACM, 2006.

[OFG97] Edgar Osuna, Robert Freund, and Federico Girosi. *Support vector machines: Training and applications*. Technical report, 1997.

[PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. *The pagerank citation ranking: Bringing order to the web*. Technical report, Stanford University, 1998.

[QD06] X. Qi and B. D. Davison. *Knowing a web page by the company it keeps*. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 228–237. ACM, 2006.

[QD09] X. Qi and B. D. Davison. *Web page classification: Features and algorithms*. ACM Computing Surveys, 41:12:1–12:31, 2009.

[QF93] Y. Qiu and H. P. Frei. *Concept based query expansion*. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 160–169. ACM, 1993.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

[Rib02] D. Riboni. *Feature selection for web page classification*, 2002.

[RJ76] S. E. Robertson and S. K. Jones. *Relevance weighting of search terms*. Journal of the American Society for Information Science, 27(3):129–146, 1976.

[RKJ08] F. Radlinski, M. Kurup, and T. Joachims. *How does clickthrough data reflect retrieval quality?* In *Proceeding of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 43–52. ACM, 2008.

[RL04] D. E. Rose and D. Levinson. *Understanding user goals in web search*. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 13–19. ACM, 2004.

[Roc71] J. Rocchio. *Relevance feedback in information retrieval*. In *The SMART Retrieval System*, pages 313–323. 1971.

[Sal71] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., 1971.

[SB88] G. Salton and C. Buckley. *Term-weighting approaches in automatic text retrieval*. Information Processing and Management: an International Journal, 24:513–523, 1988.

[SBC97] B. Shneiderman, D. Byrd, and B. W. Croft. *Clarifying search: A user-interface framework for text searches*. D-Lib Magazine, 1997.

[Seb02] F. Sebastiani. *Machine learning in automated text categorization*. ACM Computing Surveys, 34:1–47, 2002.

[SKK09] M. Strohmaier, M. Kröll, and C. Körner. *Intentional query suggestion: making user goals more explicit during search*. In *Proceedings of the 2009 Workshop on Web Search Click Data*, WSCD '09, pages 68–74. ACM, 2009.

[SLN02] A. Sun, E. P. Lim, and W. K. Ng. *Web classification using support vector machine*. In *Proceedings of the 4th International Workshop on Web Information and Data Management*, WIDM '02, pages 96–99. ACM, 2002.

[SM00] S. Slattery and T. M. Mitchell. *Discovering test set regularities in relational domains*. In *Proceedings of the 17th International Conference on Machine Learning*, ICML '00, pages 895–902. Morgan Kaufmann Publishers Inc., 2000.

[SPK08] M. Strohmaier, P. Prettenhofer, and M. Kröll. *Acquiring explicit user goals from search query logs*. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 602–605. IEEE Computer Society, 2008.

[STZ05] X. Shen, B. Tan, and C. Zhai. *Context-sensitive information retrieval using implicit feedback*. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 43–50. ACM, 2005.

[Vap95] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[Voo86] E. M. Voorhees. *The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. PhD thesis, 1986.

[WRJ05] R. W. White, I. Ruthven, and J. M. Jose. *A study of factors affecting the utility of implicit relevance feedback*. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 35–42. ACM, 2005.

[XC96] J. Xu and W. B. Croft. *Query expansion using local and global document analysis*. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 4–11. ACM, 1996.

[XC00] J. Xu and W. B. Croft. *Improving the effectiveness of information retrieval with local context analysis*. ACM Transactions on Information Systems, 18:79–112, 2000.

[Yan99] Y. Yang. *An evaluation of statistical approaches to text categorization*. Information Retrieval, 1:69–90, 1999.

[YL99] Y. Yang and X. Liu. *A re-examination of text categorization methods*. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 42–49. ACM, 1999.

[YP97] Y. Yang and J. O. Pedersen. *A comparative study on feature selection in text categorization*. In *Proceedings of the 14th International Conference on Machine Learning*, ICML '97, pages 412–420. Morgan Kaufmann Publishers Inc., 1997.