ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής

ΤΕΧΝΙΚΕΣ ΟΔΗΓΗΣΗΣ & ΕΝΤΟΠΙΣΜΟΥ ΠΛΗΡΟΦΟΡΙΑΣ

ΣΤΟΝ ΚΥΒΕΡΝΟΧΩΡΟ, ΒΑΣΙΣΜΕΝΕΣ ΣΕ ΜΟΝΤΕΛΑ ΦΥΣΙΚΩΝ ΔΙΑΔΙΚΑΣΙΩΝ

ΚΟΥΖΑΣ Σ. ΓΕΩΡΓΙΟΣ

Διπλ. Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών ΕΜΠ

ΑΘΗΝΑ Μάιος 2007

Ευχαριστίες

Αν και αυτή η ενότητα είναι η πρώτη της διατριβής, αποτελεί τους τίτλους του τέλους, όχι μόνο για την ολοκλήρωση της διατριβής αλλά ενός κεφαλαίου της ζωής μας.

Θα ήθελα να ευχαριστήσω τους κ.κ. Καθηγητές Βασίλειο Λούμο και Ελευθέριο Καγιάφα για την αμέριστη βοήθεια και εμπιστοσύνη που μου επέδειξαν όλα αυτά τα χρόνια συνεργασίας. Υπήρξαν στιγμές που μου συμπαραστάθηκαν πολύ περισσότερο από ότι ίσως άξιζα. Θέλω να ευχαριστήσω επίσης τους Καθηγητές Ε. Πρωτονοτάριο και Μ. Θεολόγου για την συμμετοχή τους στην τριμελή επιτροπή κρίσης της έρευνάς μου.

Σε όποιο χώρο και να βρεθεί κανείς το μόνο σίγουρο είναι ότι γνωρίζει νέους ανθρώπους. Μερικές ως ελάχιστες φορές, με κάποιους μπορεί να γίνεις και φίλος. Σ' αυτή την κατηγορία ανήκει ο Γιάννης. Ένας φίλος δεν είναι ποτέ κόλακας, αλλά σου υπενθυμίζει τη σκληρή, μερικές φορές, πραγματικότητα. Τον ευχαριστώ για αυτό.

Στο χώρο του εργαστηρίου γνώρισα και την Έλενα, τον Άγγελο, το Χρήστο, αλλά και το Θανάση, αν και δεν ανήκει στο δυναμικό του εργαστηρίου. Θέλω να τους ευχαριστήσω γιατί κάθε φορά που τους χρειαζόμουνα, ήταν πρόθυμοι για βοήθεια.

Χαίρομαι δε ιδιαίτερα που γνώρισα τους Γιάννη Ψωρούλα, τον έτερο Γιώργο του εργαστηρίου και τον Δημήτρη. Πολύ φιλότιμοι και ευχάριστοι φίλοι. Ο κύκλος των αναμνήσεων στη σχολή κλείνει με τα υπόλοιπα παιδιά και το ευχάριστο κλίμα του εργαστηρίου αλά και ένα σύνολο ανθρώπων με τους οποίους συναναστράφηκα και με βοήθησαν.

Ένα μεγάλο ευχαριστώ στους παλιούς μου φίλους που ανέχτηκαν την γκρίνια ετών, άλλα και στη Σταυριάννα που θα ανεχτεί και αυτή με τη σειρά της, τη γκρίνια επόμενων ετών.

Κλείνοντας θα ήθελα να ευχαριστήσω τον Επικ. Καθηγητή κ. Η. Κουκούτση, και τον Λεκτ. Καθηγητή κ. Δ. Βέργαδο για την τιμή που μου προσέφεραν να συμμετάσχουν στην επταμελή επιτροπή κρίσης της διατριβής μου.

Ως συγγραφέας θέλω να αφιερώσω τη διατριβή μου στους γονείς μου Σωτήρη και Φωτεινή και φυσικά στην αδελφή μου Σοφία που με έχουν στηρίξει σε οποιαδήποτε απόφαση έχω πάρει στη ζωή μου.

Αθήνα, 01 Μαΐου 2007

Γεώργιος Σ. Κούζας

" Ἐν οἶδα, ὅτι οὐδὲν οἶδα "

Σωκράτης

ΕΙΣΑΓΩΓΗ

ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΣΑΓΩΓΙΚΟΥ ΚΕΦΑΛΑΙΟΥ

1 Εισαγωγή

Ξεκινώντας από μια ιδέα διασύνδεσης υπολογιστών το 1969 για στρατιωτικούς και ερευνητικούς σκοπούς ουσιαστικά δημιουργήθηκε ο σκελετός και η βάση του διαδικτύου. Στην πορεία η προσθήκη και άλλων ιδεών με σημαντικότερη αυτή του Tim Berners-Lee το 1989 από το ερευνητικό ινστιτούτο CERN της Ελβετίας, δημιούργησε τον παγκόσμιο ιστό. Έκτοτε ο Παγκόσμιος Ιστός παρουσιάζει ραγδαία ανάπτυξη μετρώντας 10.000 εξυπηρετητές και 10 εκατομμύρια χρήστες το 1994, 5 μόλις χρόνια μετά τη δημιουργία του. Η εκθετική εξάπλωση αποτυπώνεται στις μέρες μας όπου το αριθμός των συνδεδεμένων υπολογιστών ξεπερνά τα 300.000.000! Η μεγάλη επιτυχία του Ιστού έγκειται στην απλότητα των κανόνων που το διέπουν που συνοψίζονται στον εξής έναν "Δεν υπάρχει κανένας περιορισμός". Αυτό γίνεται εύκολα αντιληπτό αν παρατηρήσουμε τη δομή του. Επικρατεί πλήρης αναρχία.

Από την άλλη πλευρά, η συνεχώς αυξανόμενη ανάγκη των ανθρώπων για ανταλλαγή πληροφοριών βρήκε διέξοδο στην εξάπλωση του διαδικτύου. Το Διαδίκτυο καλύπτει πλέον ένα μεγάλο μέρος των πληροφοριακών μας αναγκών(ενημέρωση, έρευνα, ψυχαγωγία, επικοινωνία). Ωστόσο η άναρχη δομή του, και ο καταιγισμός πληροφορίας πολλές φορές προκαλούν το αντίθετο αποτέλεσμα. Οι χρήστες δεν μπορούν αρχικά να αναζητήσουν και στη συνέχεια να επεξεργαστούν τον μεγάλο όγκο πληροφορίας με τον οποίο έρχονται "αντιμέτωποι". Για να καλυφθούν οι ανάγκες των χρηστών, αναπτύχθηκαν διάφορες τεχνικές και εργαλεία αναζήτησης και επεξεργασίας πληροφορίας

Η παρούσα διατριβή προτείνει μια πρότυπη μεθοδολογία δρομολόγησης της διαδικτυακής αναζήτησης πληροφορίας, με τη χρήση έξυπνων αλγορίθμων που στηρίζονται στην αυτο-οργάνωση.

Βάση της προτεινόμενης μεθοδολογίας αποτελεί η ακόλουθη απλή αλλά σημαντική υπόθεση: "*Όταν σε έναν δικτυακό τόπο του παγκόσμιου ιστού υπάρχει μια πληροφορία σχετική με τις ανάγκες ενός χρήστη, τότε, με μεγάλη πιθανότητα, υπάρχει ένας άλλος δικτυακός τόπος, με πληροφορία αντίστοιχης σχετικότητας με τον αρχικό και με μικρό κόστος μετάβασης σε χρήση υπερσυνδέσμων*".

Η παραπάνω υπόθεση ουσιαστικά εκμεταλλεύεται τη θεμελιώδη χρήση των υπερσυνδέσμων στις ιστοσελίδες, που είναι η διασύνδεση των υπερκειμένων και κατ' επέκταση, της πληροφορίας. Περιληπτικά, η διατριβή προτείνει, μια μεθοδολογία αναζήτησης γύρω από μια αρχική πηγή σχετικής πληροφορίας, προσομοιώνοντάς την με την διεργασία αναζήτησης τροφής των μυρμηγκιών γύρω από την αποικία. Η βάση της τεχνικής αναζήτησης, είναι ο αλγόριθμος αποικίας μυρμηγκιών, οποίος αν και σχετικά πρόσφατος έχει ένα ευρύ πεδίο εφαρμογών όπως προβλήματα βελτιστοποίησης (ACO), προβλήματα βέλτιστης δρομολόγησης πακέτων (Ant-Net), προβλήματα δημιουργίας κανόνων ταξινόμησης (Ant-miner) ακόμη και προβλήματα σχεδιασμού βέλτιστου σχήματος σε φτερά αεροπλάνων.

2 Ανάλυση Κεφαλαίων

Η διατριβή είναι δομημένη σε 6 κεφάλαια. Στη συνέχεια αναφέρονται συνοπτικά τα περιεχόμενα των κεφαλαίων της διατριβής.

2.1 Κεφάλαιο 1

Στο κεφάλαιο 1 επιχειρείται μια σύντομη αναφορά στην ιστορία του Διαδικτύου, το οποίο στην σημερινή εποχή έχει αλλάξει ριζικά τις έννοιες της επικοινωνίας και ανταλλαγής πληροφορίας. Αναπτύσσονται επιπλέον, τα προβλήματα που αντιμετωπίζει ο χρήστης στην προσπάθεια του να εντοπίσει τις σχετικές πηγές που ικανοποιούν τις πληροφοριακές του ανάγκες, μέσα σε αυτόν τον άναρχα δομημένο πληροφοριακό χώρο. Πραγματοποιείται μια εκτεταμένη αναφορά στις Μηχανές Αναζήτησης (Μ.Α.), με την βοήθεια των οποίων οι χρήστες μπορούν να εντοπίσουν και να προσπελάσουν την πληροφορία από απομακρυσμένες πηγές.

2.2 Κεφάλαιο 2

Στο κεφάλαιο 2 αναλύεται η έννοια της επεξεργασίας της πληροφορίας, ενώ παράλληλα παρουσιάζονται οι υπάρχουσες παραλλαγές των τεχνικών της επεξεργασίας αυτής. Στη συνέχεια αναλύονται οι μέθοδοι αναπαράστασης κειμένου, οι τεχνικές επιλογής των κατάλληλων ιδιο-χαρακτηριστικών παράλληλα με τις τεχνικές μείωσης της διαστασιολόγησης, με στόχο την αναπαράσταση των εγγράφων ως διανύσματα. Επιπλέον παρουσιάζεται μια αναλυτική περιγραφή των μοντέλων ανάκτησης πληροφορίας και των μεθόδων που χρησιμοποιούνται για την αναπαράσταση και παρουσίαση της ανακτημένης πληροφορίας. Τέλος γίνεται μια εκτενής αναφορά στις μεθόδους συσταδοποίησης εγγράφων που χρησιμοποιούνται ευρέως για κατηγοριοποίηση εγγράφων σε δυναμικά περιβάλλοντα.

2.3 Κεφάλαιο 3

Στο κεφάλαιο 3 γίνεται μια αναφορά στο βασικό μοντέλο λειτουργίας αλγορίθμων οι οποίοι αντιγράφουν διεργασίες που συναντώνται στη φύση, και στηρίζονται στην αυτό – οργάνωση, όπως είναι οι γενετικοί αλγόριθμοι και οι αλγόριθμοι νοημοσύνης σμηνών καθώς και οι αλγόριθμοι αποικίας μυρμηγκιών. Στη συνέχεια αναλύεται η θεωρία λειτουργίας της οικογένειας αλγορίθμων αποικίας μυρμηγκιών. Το 3ο κεφάλαιο κλείνει με την παρουσίαση των εφαρμογών του αλγορίθμου αποικίας μυρμηγκιών σε διάφορα γνωστικά πεδία, δίνοντας βαρύτητα στην εφαρμογή του στο πεδίο εξόρυξης δεδομένων.

2.4 Κεφάλαιο 4

Στο κεφάλαιο 4 προτείνεται μια νέα μεθοδολογία διαδικτυακής αναζήτησης πληροφορίας στηριζόμενη στον αλγόριθμο αποικίας μυρμηγκιών. Πιο συγκεκριμένα, προτείνεται μια προσέγγιση του αλγορίθμου με δυνατότητα δρομολόγησης της αναζήτησης πληροφορίας σε δυναμικά περιβάλλοντα, όπως είναι ο παγκόσμιος ιστός. Παράλληλα με τη δρομολόγηση της αναζήτησης, γίνεται χρήση τεχνικών ανάκτησης πληροφοριών για τον εντοπισμό και την αξιολόγηση της πληροφορίας που βασίζονται στην ομοιότητα εγγράφων και επιπρόσθετα γίνεται χρήση μοντέλων συσταδοποίησης εγγράφων. Στη συνέχεια παρουσιάζονται πειραματικές μετρήσεις για την προτεινόμενη μεθοδολογία καθώς και ποιοτικά συμπεράσματα.

2.5 Κεφάλαιο 5

Στο κεφάλαιο 5 πραγματοποιείται ο συνολικός απολογισμός της προτεινόμενης μεθοδολογίας καθώς και προτάσεις για περαιτέρω ανάπτυξη και αξιοποίηση αυτής σε πραγματικά συστήματα αναζήτησης. Επιπλέον, γίνεται μια αναφορά στην αλματώδη ανάπτυξη του διαδικτύου και μελετάται η μελλοντική του δομή καθώς και πιθανά μοντέλα εκτίμησης και αξιοποίησης της παρεχόμενης πληροφορίας.

2.6 Κεφάλαιο 6

Στο κεφάλαιο 6 παρουσιάζεται το βιογραφικό σημείωμα και ο κατάλογος δημοσιεύσεων του συγγραφέα της παρούσας διατριβής.

ΚΕΦΑΛΑΙΟ

1

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΤΟ ΔΙΑΔΙΚΤΥΟ - ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

ΠΕΡΙΕΧΟΜΕΝΑ 1ου ΚΕΦΑΛΑΙΟΥ

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΤΟ ΔΙΑΔΙΚΤΥΟ - ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

1 Εισαγωγή

Στο κεφάλαιο αυτό επιχειρείται μια σύντομη αναφορά στην ιστορία του Διαδικτύου, το οποίο στην σημερινή εποχή έχει αλλάξει ριζικά τις έννοιες της επικοινωνίας και ανταλλαγής πληροφορίας. Αναπτύσσονται επιπλέον, τα προβλήματα που αντιμετωπίζει ο χρήστης στην προσπάθεια του να εντοπίσει τις σχετικές πηγές που ικανοποιούν τις πληροφοριακές του ανάγκες, μέσα σε αυτόν τον άναρχα δομημένο πληροφοριακό χώρο. Πραγματοποιείται μια εκτεταμένη αναφορά σε ειδικά εργαλεία λογισμικού γνωστά και ως Μηχανές Αναζήτησης (Μ.Α.), με την βοήθεια των οποίων οι χρήστες μπορούν να εντοπίσουν και να προσπελάσουν την πληροφορία από απομακρυσμένες πηγές. Παρουσιάζεται τέλος μια επισκόπηση και σύγκριση ορισμένων Μ.Α. με σκοπό την επισήμανση της ανομοιογένειας και της ασυμβατότητας στον τρόπο εντοπισμού και κάλυψης του συνολικού όγκου της πληροφορίας στο Διαδίκτυο.

2 Το Διαδίκτυο

2.1 Ιστορική αναδρομή

Όπως πολλές άλλες πετυχημένες ιδέες, το "δίκτυο των δικτύων" εξελίχθηκε από ένα έργο που ξεκίνησε με πολύ διαφορετικό σκοπό. Ένα δίκτυο που ονομαζόταν ARPANET, σχεδιασμένο και υλοποιημένο το 1969 από τους Bolt, Beranek και Newman, κατόπιν συμβολαίου με την υπηρεσία προωθημένων ερευνητικών έργων (Advanced Research Projects Agency – ARPA) του Υπουργείου Αμύνης των ΗΠΑ. Το ARPANET ήταν ένα δίκτυο που διασύνδεσε πανεπιστήμια και εταιρείες στρατιωτικών και αμυντικών έργων, δημιουργήθηκε για να βοηθήσει τους ερευνητές να μοιράζονται πληροφορίες και να μελετήσει τον τρόπο διατήρησης των επικοινωνιών σε περίπτωση πυρηνικής επίθεσης. Από το ξεκίνημα του οι ιδρυτές του ARPANET αρχικά επέτρεπαν μόνο σε ερευνητές να συνδέονται και να τρέχουν προγράμματα σε απομακρυσμένους υπολογιστές με αποτέλεσμα το δίκτυο να μεγαλώνει συνεχώς. Σύντομα πρόσθεσαν δυνατότητες μεταφοράς αρχείων, ηλεκτρονικού ταχυδρομείου και ταχυδρομικές λίστες για να κρατήσουν σε επικοινωνία ανθρώπους με κοινά ενδιαφέροντα.

Καθώς όμως το ARPANET μεγάλωνε, δημιουργήθηκαν παράλληλα και άλλα δίκτυα ενώ φαινόταν καθαρά ότι χρειαζόντουσαν νέες μέθοδοι επικοινωνίας. Από το 1973, μία δεκαετία πριν γίνει η επανάσταση των προσωπικών υπολογιστών, η ARPA με το νέο της όνομα DARPA (Defense Advanced Research Project Agency) ξεκίνησε ένα πρόγραμμα με το όνομα "Internetting Project" ή αλλιώς έργο διασύνδεσης δικτύων. Ο στόχος της ήταν να ερευνηθεί η διασύνδεση μεταξύ πολλών δικτύων. Κεντρικό σημείο αυτής της ιδέας ήταν η ανάγκη παράκαμψης των διαφορετικών μεθόδων που χρησιμοποιεί κάθε δίκτυο για την μεταφορά των πληροφοριών του. Όταν υλοποιηθούν με κατάλληλο τρόπο οι ονομαζόμενες πύλες επικοινωνίας, μπορούν να χρησιμοποιηθούν για την διασύνδεση δικτύων, μεταφέροντας πληροφορίες από το ένα στο άλλο με διαφανή τρόπο. Οι δημιουργοί του Διαδικτύου δεν είχαν την παραμικρή υποψία ότι θα εξελισσόταν σε ένα δίκτυο δημόσιας και ελεύθερης πρόσβασης.

2.2 *Η υπόσταση του Διαδικτύου*

Το πρώτο συστατικό στοιχείο είναι το φυσικό δίκτυο. Το Διαδίκτυο συμπεριφέρεται με τον ίδιο τρόπο όπως αν όλοι οι υπολογιστές των δικτύων που συμμετέχουν ήταν συνδεδεμένοι μεταξύ τους με ένα τεράστιο καλώδιο. Στην πραγματικότητα όλοι οι υπολογιστές του Διαδικτύου είναι συνδεδεμένοι με καλώδια διαφόρων τύπων. Η καλωδίωση, όμως, είναι μόνο η αφετηρία. Το καλώδιο υποστηρίζει τη διακίνηση των μηνυμάτων και των πληροφοριών μέσα στο δίκτυο, αλλά δεν μπορεί να δώσει κάποιο νόημα σε αυτή τη διακίνηση.

Έτσι ένα βήμα πιο πέρα από το απλό καλώδιο, είναι ένας αριθμός εξειδικευμένων γλωσσών για την διαβίβαση αυτών των μηνυμάτων, όπως είναι η διεύθυνση και ο ταχυδρομικός κώδικας που χρησιμοποιούνται από την ταχυδρομική υπηρεσία. Αυτές οι γλώσσες λέγονται πρωτόκολλα και διαιρούν το φυσικό δίκτυο σε διακεκριμένες περιοχές, επιτρέποντας την αποστολή μηνυμάτων από μία περιοχή σε μία άλλη. Το φυσικό δίκτυο του Διαδικτύου λειτουργεί σαν ένα τεράστιο, απλό κύκλωμα, μία γιγαντιαία γραμμή, μέσω της οποίας μεταφέρονται τα δεδομένα όλων των χρηστών. Όλοι οι υπολογιστές και το λογισμικό που απαρτίζουν το Διαδίκτυο είναι είτε πελάτες όπου λαμβάνουν και μεταφράζουν τα δεδομένα, είτε εξυπηρετητές όπου παρέχουν δεδομένα. Συνεπώς ένας απλώς χρήστης διαθέτοντας έναν υπολογιστή, μια συσκευή διαμόρφωσης-αποδιαμόρφωσης, μια σύνδεση σε κάποιον παροχέα Διαδικτύου καθώς και το κατάλληλο λογισμικό, μπορεί να προσπελάσει το Διαδίκτυο και να αποκτήσει πληροφορίες από αυτό. Στις παρακάτω ενότητες δεν θα γίνει περαιτέρω αναφορά σε τεχνικά στοιχεία και περιγραφές που αφορούν το Διαδίκτυο, εφ' όσον η παρούσα διατριβή ασχολείται με μεθόδους αναζήτησης, επεξεργασίας και ταξινόμησης της πληροφορίας θεωρώντας αποκλειστικά το Διαδίκτυο ως το μέσο μετάδοσης και διάχυσης αυτής.

3 Η πληροφορία στην σημερινή εποχή

Η πληροφορία ανέκαθεν ήταν σημαντική για τη ζωή των ανθρώπων. Με την τεχνολογική επανάσταση και την γενικότερη πρόοδο στις επιστήμες, η ποσότητα της πληροφορίας αφενός αυξήθηκε ραγδαία και αφετέρου η διάδοση της επεκτάθηκε από το χαρτί, που ακόμα αποτελεί ένα δημοφιλές μέσο διάδοσης, στα ηλεκτρονικά μέσα, όπως το ραδιόφωνο, την τηλεόραση και τα υπολογιστικά συστήματα. Έτσι σήμερα, με την χρήση των υπολογιστικών συστημάτων που συνδέονται σε δίκτυα υψηλών ταχυτήτων παρέχεται μία ολοένα και πιο ταχεία πρόσβαση σε μία τεράστια ποικιλία πηγών πληροφορίας.

Στη σημερινή εποχή της κοινωνίας της πληροφορίας, όπου η χρήση του Διαδικτύου έχει ήδη καθιερωθεί στην ζωή του ανθρώπου, η ποσότητα της διαθέσιμης πληροφορίας είναι τεράστια. Καθώς όμως μέσα σε αυτόν τον κυκεώνα πληροφοριών δεν είναι δυνατόν να είναι γνωστόν εκ των προτέρων ποιος παρέχει την ζητούμενη πληροφορία, ούτε τον αριθμό των πηγών πληροφοριών, το ενδιαφέρον του χρήστη εστιάζεται στην αναζήτηση της πληροφορίας. Προς αυτό συνέβαλαν τα διάφορα εργαλεία λογισμικού που αναπτύχθηκαν για να υποστηρίξουν την εύκολη και γρήγορη εύρεση πληροφοριών σχετικά με κάποιο θέμα. Παρόλα αυτά, η διαρκής ανάπτυξη του Διαδικτύου, ο συνεχής εμπλουτισμός των πηγών πληροφορίας και η προσθήκη νέων καθιστά τον εντοπισμό της πληροφορίας ολοένα και πιο δύσκολο σε έναν νοητό χώρο όπου είναι αποθηκευμένο ένα τεράστιο ακατέργαστο ή μη ποσό πληροφορίας που ονομάζεται παγκόσμιος ιστός .

3.1 Ο Παγκόσμιος Ιστός

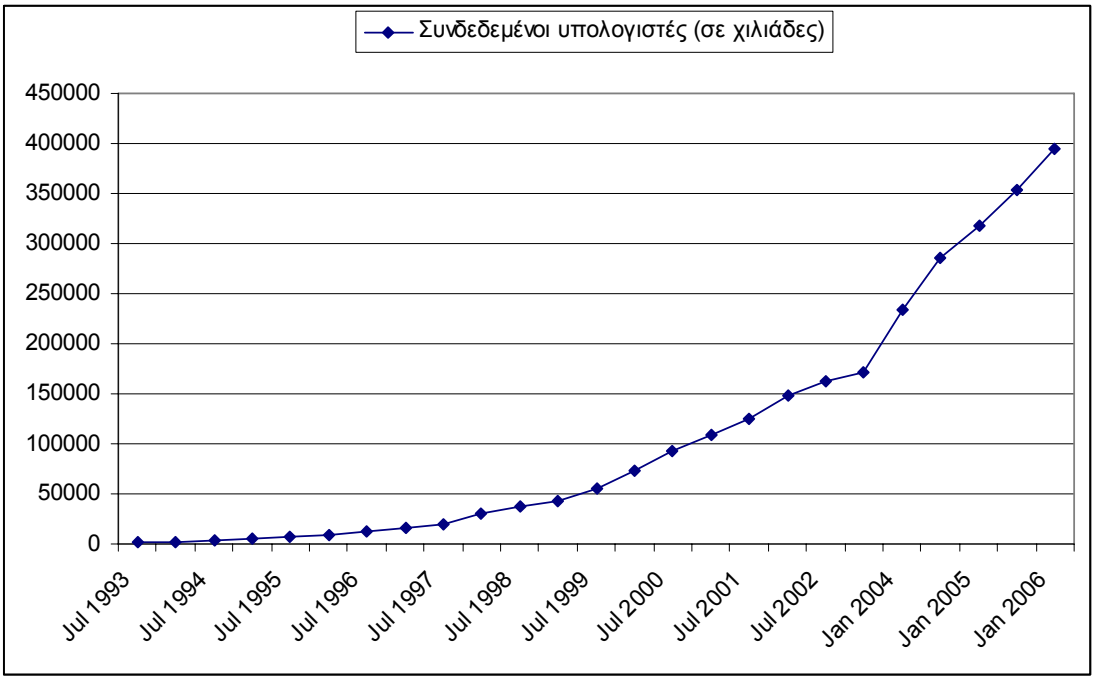
Η ιστορία του Παγκόσμιου Ιστού έχει ως σημείο εκκίνησης το έτος 1989 και υλοποιείται στο ερευνητικό ινστιτούτο CERN στην Ελβετία, από τον Tim Berners-Lee. Μέχρι τα τέλη του επόμενου χρόνου, πραγματοποιείται η επίδειξη του πρώτου λογισμικού που υλοποιεί τον πρώτο στοιχειώδη Παγκόσμιο Ιστό. Αρχικά γίνεται διαθέσιμη όλη η υπάρχουσα πληροφορία που βρίσκεται αποθηκευμένη στο υπολογιστικό σύστημα του ινστιτούτου η οποία έχει ως μέσο προσπέλασης τον επονομαζόμενο Πλοηγητή Παγκόσμιου Ιστού [CERN].

Μέχρι τα τέλη του 1994, ο Παγκόσμιος Ιστός παρουσιάζει ραγδαία ανάπτυξη μετρώντας 10,000 εξυπηρετητές και 10 εκατομμύρια χρήστες.. Η τεχνολογία του συνεχώς επεκτείνεται προκειμένου να καλύψει νέες ανάγκες, όπως ασφάλεια και εφαρμογές πολυμέσων. Το Σχήμα 1, παρουσιάζει τη ραγδαία ανάπτυξη του Διαδικτύου βάσει του αριθμού των συνδεδεμένων υπολογιστών (host computers) [ISC]. Αντίστοιχες τάσεις παρουσιάζει και το πλήθος των ιστοσελίδων στον Παγκόσμιο Ιστό.

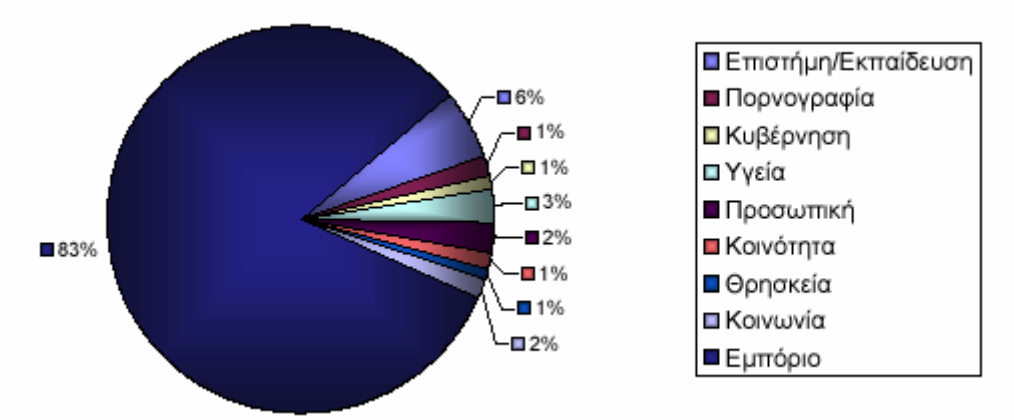
Η παραπάνω εξάπλωση του Παγκόσμιου Ιστού και του Διαδικτύου συνεπάγεται την διαθεσιμότητα ενός τεράστιου όγκου πληροφοριών και δεδομένων σε σημείο που ο εντοπισμός του καθίσταται εξαιρετικά δύσκολος έως αδύνατος, στην πληθώρα των περιπτώσεων. Μία ενδεικτική κατανομή της ποικιλίας της διαθέσιμης πληροφορίας παρουσιάζεται στο Σχήμα 2, όπου ενδιαφέρον παρουσιάζει το γεγονός ότι το 83% σχετίζεται με εμπορικές δραστηριότητες [Lawrence,99a], [Lawrence,99b].

Στο σημερινό χώρο του Διαδικτύου, όπου ο όγκος της πληροφορίας αυξάνεται εκθετικά κρίνεται ολοένα και πιο απαραίτητη η ύπαρξη των κατάλληλων υπηρεσιών διαμεσολάβησης ώστε οι χρήστες να καταβάλλουν την ελάχιστη δυνατή προσπάθεια να στόχο να ανακαλύψουν, συλλέξουν, συγκρίνουν, αναλύσουν και ταξινομήσουν πληροφορίες που εξυπηρετούν τις ανάγκες τους [Agichtein,00], [Budzik,00].

Η τεράστια ποσότητα της διαθέσιμης πληροφορίας στο Διαδίκτυο μπορεί να αποτελεί το μεγάλο του πλεονέκτημα αλλά είναι ταυτόχρονα και το αδύνατο σημείο του. Ο λόγος είναι ότι σε έναν τέτοιο όγκο πληροφορίας συχνά η αναζήτηση από την πλευρά του χρήστη καταλήγει να είναι δυσχερής ή μη πλήρης σε ορθά αποτελέσματα.



Σχήμα 1. Συνδεδεμένοι υπολογιστές στο Παγκόσμιο Ιστό (1993-2006)



Σχήμα 2. Κατανομή του είδους της πληροφορίας στον Παγκόσμιο Ιστό

Έτσι παρατηρείται, μία διαρκώς αυξανόμενη τάση προς εκτέλεση πολύπλοκων ερωτήσεων από τους χρήστες, ώστε να λάβουν τα δεδομένα που επιθυμούν. Όμως το να εντοπίσει κανείς τις σχετικές πηγές πληροφορίας θεωρείται μια εξαιρετικά δύσκολη εργασία. Αυτό γιατί υπάρχουν αναρίθμητες πηγές πληροφορίας που διαφέρουν στο είδος των στοιχείων πληροφορίας που περιέχουν, στις διαπροσωπείες που παρέχουν στους χρήστες καθώς επίσης και στην φυσική γεωγραφική κατανομή τους [Buyukkokten,99]. Πολλές πηγές περιέχουν πληροφορία σε μορφή κειμένου και υποστηρίζουν ερωτήσεις με χρήση λέξεων-κλειδιών ενώ άλλες πηγές περιέχουν δομημένα στοιχεία πληροφορίας και παρέχουν τη δυνατότητα χρήσης σχεσιακών γλωσσών [Adar,99]. Επιπλέον, συχνά οι χρήστες αναγκάζονται να συνδυάζουν οι ίδιοι τα επιμέρους αποτελέσματα της αναζήτησης τους για να έχουν τις επιθυμητές απαντήσεις. Το πιο δημοφιλές εργαλείο αναζήτησης πληροφοριών που αποτελεί ταυτόχρονα και μία λύση στα παραπάνω προβλήματα είναι οι διάφορες μηχανές αναζήτησης [Lawrence,98a].

4 Μηχανές αναζήτησης (Μ.Α.)

Οι μηχανές αναζήτησης (Μ.Α.) είναι ειδικά εργαλεία λογισμικού με την βοήθεια των οποίων οι χρήστες μπορούν να προσπελάσουν απομακρυσμένες πηγές πληροφορίας και δικτυακούς τόπους - ιστοχώρους με τη λιγότερη δυνατή προσπάθεια χωρίς να μετακινηθούν από τον προσωπικό τους χώρο. Η χρήση των υπερσυνδέσεων επιτρέπει τη μετακίνηση του χρήστη μεταξύ λογικά συνδεδεμένων πληροφοριών, ανεξάρτητα από την πραγματική και γεωγραφική τους απόσταση. Έτσι, οι χρήστες μεταβαίνουν σε δικτυακούς τόπους όπου υπάρχουν σχετικοί όροι με ότι αναζητείται.

Οι Μ.Α. ταξινομούνται κυρίως σε τρεις βασικές κατηγορίες. Στην πρώτη κατηγορία ανήκουν όσες Μ.Α. στηρίζονται σε ευρετήρια και επιτρέπουν την αναζήτηση βάσει λέξεων-κλειδιών που μπορούν να συνδυαστούν με λογικούς τελεστές. Αυτά τα ευρετήρια αποτελούνται από πληροφορίες που έχουν συγκεντρωθεί από ειδικά αυτόματα προγράμματα (αράχνες ή ρομπότ) τα οποία περιπλανώνται στο Διαδίκτυο μέσω των υπερσυνδέσεων των ιστοσελίδων και ελέγχουν το κείμενο που βρίσκεται σε κάθε δικτυακό τόπο που επισκέπτονται. Ο χρήστης παρέχει λέξεις-κλειδιά του επιθυμητού θέματος και η Μ.Α. αξιοποιεί τα ευρετήρια που είναι αποθηκευμένα σε βάσεις δεδομένων για να απαντήσει σχετικά. Οι Μ.Α. αυτού του είδους ονομάζονται αυτόματες Μ.Α..

Η δεύτερη κατηγορία μηχανών αναζήτησης βασίζεται σε καταλόγους που έχουν οργανωθεί με βάση το τύπο και το είδος της παρεχόμενης πληροφορίας στην οποία ο χρήστης καταλήγει με κατάλληλες διαδικασίες πλοήγησης. Οι υπηρεσίες αυτές ονομάζονται και Θεματικοί Κατάλογοι. Η τρίτη κατηγορία αποτελεί συνδυασμό των δύο παραπάνω. Έτσι αυτές οι υπηρεσίες αναζήτησης ονομάζονται υβριδικές Μ.Α.. Επιπρόσθετα, υπάρχουν και κάποια άλλα είδη Μ.Α. όπου υποστηρίζουν ότι πραγματοποιούν διαδικασίες επεξεργασίας φυσικής γλώσσας, στα ερωτήματα των χρηστών ή παρέχουν τα αποτελέσματά τους επί πληρωμή. Διαφέρουν δηλαδή στον τρόπο με τον οποίο παρουσιάζουν τα τελικά αποτελέσματα στον χρήστη, μπορούν όμως να ενταχθούν στις παραπάνω τρεις κατηγορίες.

Τέλος υπάρχουν και οι λεγόμενες μηχανές μετα-αναζήτησης ή πολύ-νηματικές Μ.Α., οι οποίες επιστρέφουν αποτελέσματα που προέρχονται από συνδυασμό αποτελεσμάτων άλλων υπηρεσιών αναζήτησης αντί από δικές τους βάσεις δεδομένων και ευρετήρια.

4.1 Είδη Μηχανών Αναζήτησης (Μ.Α.)

Οι Μηχανές Αναζήτησης διακρίνονται κυρίως σε τρία είδη, ανάλογα με τον τρόπο λειτουργία τους. Πιο συγκεκριμένα, η διάκριση αυτή γίνεται βάσει της διαδικασίας συλλογής, αποθήκευσης και επεξεργασίας της πληροφορίας που θα διανεμηθεί στον τελικό χρήστη. Έτσι υπάρχουν οι Αυτόματες Μ.Α. που συλλέγουν, αποθηκεύουν και επεξεργάζονται αυτόματα ένα τεράστιο ποσό ιστοσελίδων με την βοήθεια ευφυών προγραμμάτων που "σαρώνουν" τον κυβερνοχώρο και τις ιστοσελίδες. Εξ' αιτίας της λειτουργίας τους αυτής η οποία είναι ασταμάτητη, ακολουθώντας τις υπερσυνδέσεις μεταξύ των ιστοχώρων, τα προγράμματα αυτά ονομάζονται crawlers, "αράχνες" ή ρομπότ. Μια άλλη κατηγορία Μ.Α. στηρίζει τις διαδικασίες συλλογής, αποθήκευσης και επεξεργασίας της πληροφορίας σε ειδικούς συντάκτες οι οποίοι δημιουργούν "φακέλους" και κατηγορίες με σκοπό την ταξινόμηση των ιστοσελίδων σε ανάλογες περιοχές πληροφορίας. Βάσει της παρεχόμενης ταξινομημένης πληροφορίας σε κατηγορίες, αυτές οι Μ.Α. ονομάζονται και Θεματικοί Κατάλογοι. Τέλος υπάρχουν και οι λεγόμενες Μ.Α. υβριδικής μορφής οι οποίες χρησιμοποιούν ενίοτε και αυτόματους μηχανισμούς και ειδικούς συντάκτες για την συλλογή, αποθήκευση και επεξεργασία της πληροφορίας που βρίσκεται στο Διαδίκτυο [Anagnostopoulos,04].

4.2 Αυτόματες Μ.Α.

Οι μηχανές αυτές όπως προαναφέρθηκε συλλέγουν, αποθηκεύουν και επεξεργάζονται αυτόματα ένα τεράστιο ποσό ιστοσελίδων με την βοήθεια προγραμμάτων που ονομάζονται crawlers, "αράχνες" ή ρομπότ. Τα προγράμματα αυτά επισκέπτονται μια ιστοσελίδα, διαβάζουν το περιεχόμενό της και ακολουθούν τις υπερσυνδέσεις που βρίσκουν σε αυτό με σκοπό να επαναλάβουν την διαδικασία αυτή στις καινούργιες ιστοσελίδες [Cho,98]. Κατόπιν επιστρέφουν επεξεργασμένες πληροφορίες στη βάση δεδομένων της Μ.Α. ανά τακτά χρονικά διαστήματα [Gravano,99]. Φυσικά, όσο πιο συχνά επαναλαμβάνεται η διαδικασία αυτή, τόσο πιο ανανεωμένη είναι η πληροφορία που παρέχεται στον τελικό χρήστη.

Όποια πληροφορία βρίσκεται καταχωρείται στα λεγόμενα ευρετήρια της Μ.Α.. Τα ευρετήρια είναι στην ουσία τεράστιες βάσεις δεδομένων οι οποίες είναι υπεύθυνες για την αποθήκευση και την ανάκτηση των πληροφοριών. Τα αποτελέσματα που επιστρέφουν στους χρήστες εξαρτώνται άμεσα από τα ευρετήρια. Έτσι σε ¨άμεση συνάρτηση με τα παραπάνω, όσο πιο συχνά ανανεώνονται αυτά, όσο δηλαδή μεγαλώνει η συχνότητα επίσκεψης των αυτόματων προγραμμάτων αναζήτησης τόσο πιο ακριβή και σωστά είναι τα αποτελέσματα.

Οι crawlers, οι "αράχνες" ή τα ρομπότ συνήθως έχουν ως αφετηρία ιστοχώρους με μεγάλη επισκεψιμότητα από χρήστες ή ιστοχώρους που ταξινομούν πολλές περιοχές πληροφοριών που ονομάζονται και Πύλες. Βρίσκοντας ένα μεγάλο ποσό από υπερσυνδέσεις τα προγράμματα αναζήτησης εξαπλώνονται στον Κυβερνοχώρο με μεγάλο ρυθμό ανάπτυξης. Ένα παράδειγμα αποτελεί η ευρέως και πλέον σύγχρονη αυτόματη Μ.Α. Google, η οποία ξεκίνησε ως ένα πανεπιστημιακό ερευνητικό πρόγραμμα. Το σύστημα χρησιμοποιεί τρία αυτόματα προγράμματα κάθε φορά τα οποία με την σειρά τους είναι συνδεδεμένα με περίπου 300 συνδέσεις Εντοπιστών Ομοιόμορφων Πόρων σε ιστοσελίδες. Στην μέγιστη απόδοση λειτουργίας του συστήματος αυτού, χρησιμοποιούνται τέσσερα αυτόματα προγράμματα και "σαρώνονται" πάνω από 100 ιστοσελίδες το δευτερόλεπτο, δημιουργώντας έτσι γύρω στα 600 Kb δεδομένων το δευτερόλεπτο.

Για να επιτευχθούν τέτοιες ταχύτητες επεξεργασίας δεδομένων, το σύστημα τροφοδοτεί την απαραίτητη πληροφορία που χρειάζονται τα προγράμματα. Έτσι μια συνηθισμένη τακτική είναι κάθε Μ.Α. αυτού του τύπου να έχει το δικό της Εξυπηρετητή Ονομάτων Τομέα. Στις επόμενες παραγράφους αναλύονται τρεις βασικές λειτουργίες που συνιστούν την "σάρωση" της ιστοσελίδας, την καταχώρησή της στα ευρετήρια της και τους παρεχόμενους μηχανισμούς αναζήτησης στον χρήστη. Παραδείγματα αυτόματων Μ.Α. είναι οι AltaVista [AV], AllTheWeb [ATW], Excite [EXCITE], Lycos [LYCOS], Google [GOOGLE], Northern Light [NL], HotBot [HB] και MSN Search [MSN].

4.2.1 Ανάλυση περιεχομένου της ιστοσελίδας

Η διαδικασία αυτή ουσιαστικά ξεκινάει όταν "συλλαμβάνεται" η προς επεξεργασία ιστοσελίδα. Στην ουσία πραγματοποιείται ανάλυση του περιεχομένου του πηγαίου κώδικα της ιστοσελίδας, της οποίας ο τύπος είναι η Γλώσσα Υπερκείμενης Σήμανσης. Έτσι, όταν ένα αυτόματο πρόγραμμα αναλύει έναν τέτοιο κώδικα έχει ως στόχο να δημιουργήσει μια λίστα με τους ευρισκόμενους όρους και τους αντίστοιχους δείκτες που θα αντιπροσωπεύουν την ιστοσελίδα αυτή.

Οι όροι κατατάσσονται ανάλογα με το που βρίσκονται στον πηγαίο κώδικα της Γλώσσας Υπερκείμενης Σύνδεσης. Με άλλα λόγια λέξεις που βρίσκονται στην ετικέτα του τίτλου, υπό-τίτλων ή στα πεδία των μετα-ετικετών (meta-tags) και παρότι δεν ανήκουν στους όρους που φαίνονται κατά την περιήγηση στην ιστοσελίδα αξιολογούνται με μεγαλύτερη βαρύτητα όπως θα περιγραφεί παρακάτω. Επιπλέον, πρέπει να παραβλεφθούν όροι, σημεία στίξης και άλλα κειμενικά σύμβολα που δεν προσδίδουν χρήσιμη πληροφορία. Ακόμα, κατά την ανάλυση του περιεχομένου της εξεταζόμενης ιστοσελίδας μια ωφέλιμη τεχνική είναι η αποβολή των κοινών λέξεων που δεν έχουν καμία ουσιαστικά διακριτική ικανότητα και ισχύ

[Anagnostopoulos,04]. Μια άλλη τεχνική που εφαρμόζεται από πολλές αυτόματες Μ.Α. είναι η επιλεκτική συλλογή πληροφοριών από συγκεκριμένα πεδία και ετικέτες του πηγαίου κώδικα της Γλώσσας Υπερκείμενης Σήμανσης, όπως για παράδειγμα οι μετα-ετικέτες με παράλληλη συλλογή όρων από ένα συγκεκριμένο αριθμό γραμμών του διανεμημένου κειμένου.

Φυσικά, όπως είναι αναμενόμενο διαφορετικές προσεγγίσεις και μηχανισμοί σχετικά με την ανάλυση του περιεχομένου τις ιστοσελίδας επιφέρουν και διαφορετικά επιστρεφόμενα αποτελέσματα στο χρήστη, ακόμα και αν η συλλογή των εξεταζομένων ιστοσελίδων είναι ταυτόσημη [Anagnostopoulos,02]. Με άλλα λόγια μια αυτόματη Μ.Α. που αποβάλλει κατά την "σάρωση" μιας ιστοσελίδας τις κοινές λέξεις, ενώ παράλληλα λαμβάνει υπόψη τις πληροφορίες στα πεδία των μετα-ετικετών της, αποθηκεύει διαφορετικούς αντιπροσωπευτικούς όρους από μια αυτόματη Μ.Α. που υπολογίζει και τις κοινές λέξεις στην ίδια ιστοσελίδα. Αυτό είναι ένα κλασσικό παράδειγμα ανομοιογένειας στον τρόπο συλλογής της πληροφορίας που όπως γίνεται αντιληπτό, οδηγεί σε ανομοιογένεια στον τρόπο σύνταξης της ιστοσελίδας και φυσικά σε περαιτέρω ανομοιογένεια στα τελικά παρουσιαζόμενα αποτελέσματα στον χρήστη.

Όλα τα παραπάνω προϋποθέτουν ότι ο δημιουργός και κάτοχος του εκάστοτε ιστοχώρου επιθυμεί να συμπεριληφθούν οι ιστοσελίδες του στα ευρετήρια της αυτόματης Μ.Α.. Υπάρχουν και οι περιπτώσεις όπου δεν είναι επιθυμητή η προβολή μέσω μιας Μ.Α. ή ακόμα δεν είναι επιθυμητές οι διαδικασίες συλλογής και σύνταξης. Το αναφαίρετο δικαίωμα της μη-προβολής ή της "ανωνυμίας" μπορεί να επιτευχθεί μέσω ενός πρωτοκόλλου που απαγορεύει την "σάρωση" από τις αυτόματες Μ.Α.[WRP1]. Πρόκειται ουσιαστικά για μια εντολή που προστίθεται σε ένα ειδικά καθορισμένο πεδίο των μετα-ετικετών στην αρχή του πηγαίου κώδικα της Γλώσσας Υπερκείμενης Σήμανσης, που απαγορεύει οποιαδήποτε από τις παραπάνω περιγραφόμενες διαδικασίες. Στην περίπτωση λοιπόν που η ιστοσελίδα προστατεύεται από το πρωτόκολλο αυτό, η ύπαρξη της στον Κυβερνοχώρο αγνοείται από την αυτόματη Μ.Α., ούτε συνεχίζεται η διαδικασία ανεύρεσης άλλων υπερσυνδέσεων.

Μετα-Ετικέτες:

Οι μετα-ετικέτες επιτρέπουν στον δημιουργό της ιστοσελίδας να προσδιορίσει ποιες λέξεις κλειδιά, φράσεις και έννοιες περιγράφουν την εξεταζόμενη ιστοσελίδα. Αυτό είναι πολύ χρήσιμο στην περίπτωση που ορισμένοι όροι έχουν παραπάνω από μία σημασία. Βέβαια, υπάρχει ο κίνδυνος της μη σωστής αντιπροσώπευσης της ιστοσελίδας από τους όρους που βρίσκονται στα εν λόγω πεδία, εξ' αιτίας λάθους του δημιουργού της ιστοσελίδας. Ενδέχεται δηλαδή, το περιεχόμενο να μην εκπροσωπείται ορθά. Έτσι, σε περίπτωση μεταβολής μιας δυναμικής ιστοσελίδας δεν παρέχεται εγγύηση ότι θα μεταβληθούν και τα αντίστοιχα πεδία. Για την αποφυγή του φαινομένου αυτού πολλές αυτόματες Μ.Α. συσχετίζουν τα πεδία και τις περιγραφές των μετα-ετικετών με το απλό περιεχόμενο του κειμένου, απορρίπτοντας τους όρους που δεν ταυτίζονται.

4.2.2 Δημιουργία ευρετηρίων

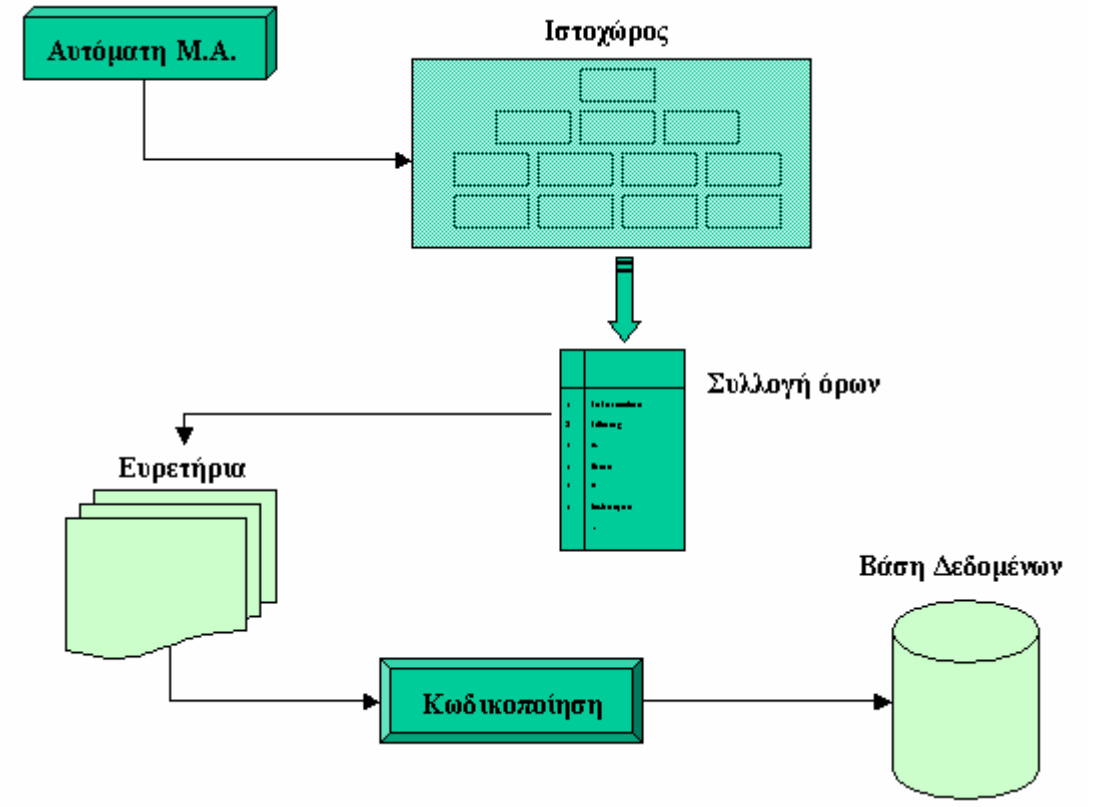
Λόγω της φύσης του Διαδικτύου, το οποίο συνεχώς μεγαλώνει και μεταβάλλεται οι λειτουργίες των αυτόματων προγραμμάτων δεν τερματίζονται ποτέ. Ολοένα και περισσότερες ιστοσελίδες δημιουργούνται, άλλες ανανεώνονται ενώ άλλες εξαφανίζονται. Όταν μια αυτόματη Μ.Α. τελειώσει τις διαδικασίες ανάλυσης κειμένου που αναφέρεται στην παραπάνω ενότητα, πρέπει να αποθηκεύσει τις πληροφορίες αυτές στη βάση δεδομένων της και να την συντάξει στα ευρετήριά της.

Στην πιο απλή περίπτωση, μια αυτόματη Μ.Α. μπορεί απλώς να αποθηκεύσει τον εκάστοτε όρο, και την διεύθυνση του Ομοιόμορφου Εντοπιστή Πόρου στην οποία εντοπίστηκε. Αυτό όμως θα περιόριζε τις δυνατότητες της και θα παρείχε λανθασμένα αποτελέσματα στον χρήστη στην περίπτωση όπου ο όρος αυτός δεν είχε κάποια σημαντική διακριτική ισχύ μέσα

στο περιεχόμενο της ιστοσελίδας. Επιπλέον δεν θα λαμβανόταν υπόψη η συχνότητα εμφάνισής της ή εάν η σελίδα περιλάμβανε υπερσυνδέσεις σε άλλες ιστοσελίδες που περιέχουν τον όρο αυτόν. Με άλλα λόγια, δεν θα μπορούσε σε καμία περίπτωση να παρουσιαστεί στον τελικό χρήστη ένας πίνακας με αποτελέσματα ταξινομημένα σύμφωνα με τη σχετικότητά τους.

Έτσι λοιπόν, μια αυτόματη Μ.Α. αποθηκεύει όχι μόνο τους όρους και διευθύνσεις Ομοιόμορφων Εντοπιστών Πόρων, αλλά και τον αριθμό εμφάνισης αυτών στην εξεταζόμενη ιστοσελίδα. Επιπλέον, ενδέχεται ο μηχανισμός να αναθέτει μεγαλύτερο βάρος σε κάθε όρο που εμφανίζεται κοντά στην κορυφή του κειμένου, στο πεδίο της ετικέτας του τίτλου, στις υπερσυνδέσεις ή στα πεδία των μετα-ετικετών. Όπως αναφέρθηκε και προηγουμένως οι Μ.Α. έχουν συνήθως διαφορετικό τρόπο ανάθεσης βαρών στους συντασσόμενους όρους και δημιουργούν διαφορετικά ευρετήρια και δείκτες ακόμα και για την ίδια συλλογή εξεταζόμενων ιστοσελίδων. Για το λόγο αυτό σχεδόν ποτέ τα επιστρεφόμενα αποτελέσματα δύο ή περισσοτέρων Μ.Α. δεν ταυτίζονται, ακόμα και στην περίπτωση που η ερώτηση του χρήστη είναι η ίδια.

Μετά από την στάθμιση των όρων και την σύνταξη τους στα ευρετήρια, ακολουθούν διαδικασίες συμπίεσης και κωδικοποίησης των δεδομένων με σκοπό την οικονομία του αποθηκευτικού χώρου. Για παράδειγμα η αυτόματη Μ.Α. Google, χρησιμοποιεί 2 δυφιοσυλλαβές των 8 δυφίων το καθένα για να αποθηκεύσει πληροφορίες σχετικά με την στάθμιση ενός όρου εάν αυτός είναι γραμμένος με κεφαλαία ή όχι, το φόντο του ή την θέση που κατέχει στο κείμενο. Κάθε μία από αυτές τις παραμέτρους μπορούν να λάβουν μέχρι 2 ή 3 δυφίων (bits). Ως αποτέλεσμα, ένα μεγάλο ποσό πληροφορίας μπορεί να αποθηκευτεί σε μια συμπαγής δομή. Μετά τη διαδικασία αυτή ακολουθεί η διαδικασία της σύνταξης των όρων.



Σχήμα 3. Μηχανισμοί ανάλυσης, σύνταξης, κωδικοποίησης και αποθήκευση μιας Αυτόματης Μ.Α.

Η διαδικασία αυτή αποσκοπεί κυρίως στη γρήγορη ανεύρεση της πληροφορίας. Υπάρχουν πολλοί τρόποι δημιουργίας ευρετηρίων, με πιο κοινό από όλους την δημιουργία πίνακα

κατατεμαχισμού. Στον μηχανισμό αυτό, εφαρμόζεται μια αριθμητική τιμή σε κάθε όρο και κάθε καταχώρηση διανέμεται σύμφωνα με έναν προκαθορισμένο αριθμό διαιρέσεων. Αυτή η αριθμητική κατανομή διαφέρει από την κατανομή των αλφαβητικών λέξεων, προσδίδοντας έτσι αποτελεσματικότητα στην μέθοδο αυτή [Salton,89]. Σε ευρετήρια λεξικών, υπάρχουν κάποια αρχικά γράμματα που συνιστούν περισσότερες λέξεις και όρους από κάποια άλλα. Στην Αγγλική για παράδειγμα, η ενότητα που περιέχει τις λέξεις που αρχίζουν με "a" είναι κατά πολύ περισσότερες από αυτές που αρχίζουν με "q". Αυτή η ανισότητα επιβαρύνει το σύστημα με μεγαλύτερους χρόνους απόκρισης στην περίπτωση εύρεσης μιας λέξης ή ενός όρου που αρχίζει με ένα πιο "δημοφιλές" γράμμα. Η διαδικασία του κατατεμαχισμού εξομαλύνει κάπως αυτή την διαφορά, ενώ παράλληλα μειώνει τον μέσο χρόνο ανεύρεσης μιας καταχώρησης. Ο πίνακας κατατεμαχισμού περιέχει τον αντίστοιχο αριθμό μαζί με έναν δείκτη που αντιστοιχεί στα δεδομένα, τα οποία είναι αποθηκευμένα με όποιο τρόπο επιτρέπεται για να αποθηκευτούν αποτελεσματικά. Ο συνδυασμός της αποτελεσματικής σύνταξης και αποθήκευσης επιτρέπει την γρήγορη πρόσβαση στα αποτελέσματα, ακόμα και στην περίπτωση όπου ο χρήστης υποβάλει ένα σύνθετο ερώτημα. Το Σχήμα 3 απεικονίζει τους δύο περιγραφόμενους μηχανισμούς, από την "σύλληψη" μιας ιστοσελίδας μέχρι την αποθήκευση των κωδικοποιημένων και συμπιεσμένων δεδομένων αυτής. Πιο συγκεκριμένα φαίνεται το αρχικό στάδιο "σάρωσης" ενός ιστοχώρου και των ιστοσελίδων που τον αποτελούν, ενώ παράλληλα πραγματοποιείται η συλλογή των όρων και η εύρεση των επόμενων ιστοχώρων μέσω των σχετικών υπερσυνδέσεων. Κατόπιν ακολουθεί η δημιουργία μιας λίστας με τους ευρισκόμενους όρους και τους αντίστοιχους δείκτες που δείχνουν σε αυτούς, ενώ ακολουθεί μετέπειτα η δημιουργία των ευρετηρίων βάσει της ακολουθούμενης μεθόδου στάθμισης. Τέλος, επιτελείται η κατάλληλη κωδικοποίηση των δεδομένων με σκοπό την πιο αποτελεσματική και οικονομική αποθήκευση των δεδομένων.

4.2.3 Μηχανισμοί ανάκτησης πληροφορίας

Μετά από την συλλογή, επεξεργασία, σύνταξη και αποθήκευση των δεδομένων μια αυτόματη Μ.Α. έχει μηχανισμούς που επιτρέπουν στο χρήστη την ανάκτηση πληροφορίας ανάλογα με τις απαιτήσεις του. Έτσι, παρέχεται σε αυτόν ένα γραφικό περιβάλλον, μέσα από το οποίο συντάσσει ερωτήσεις με σκοπό να καλύψει τις πληροφοριακές του ανάγκες. Αν και η πιο απλή ερώτηση μπορεί να είναι μια και μοναδική λέξη, συνήθως οι Μ.Α. παρέχουν στο χρήστη τη δυνατότητα να θέσει πιο συγκεκριμένα ερωτήματα, μειώνοντας έτσι τα άσχετα επιστρεφόμενα αποτελέσματα. Η δημιουργία πιο συγκεκριμένων ερωτημάτων γίνεται μέσω της λογικής Boole με τη χρήση αντίστοιχων συντελεστών "ΚΑΙ" – "AND", "Η" – "OR", "ΟΧΙ" – "NOT" καθώς και με άλλους συντελεστές που επιτρέπουν την εισαγωγή μιας ολοκληρωμένης φράσεως ή τον προσδιορισμό της απόστασης μεταξύ δύο όρων. Οι συντελεστές αυτοί μπορούν να χρησιμοποιηθούν είτε ξεχωριστά είτε σε συνδυασμό. Η σωστή χρήση αυτών βοηθάει το χρήστη στις αναζητήσεις του.

Λόγω του ότι οι μηχανισμοί ανάκτησης πληροφορίας είναι κοινοί για όλα τα είδη Μ.Α. (αυτόματες Μ.Α., Θεματικοί Κατάλογοι ή Υβριδικές Μ.Α.), θα γίνει εκτεταμένη αναφορά για αυτούς σε επόμενη ενότητα.

4.3 Θεματικοί Κατάλογοι

Οι Θεματικοί Κατάλογοι προσφέρουν πληροφορίες στους χρήστες οι οποίες έχουν προηγουμένως αναλυθεί, αξιολογηθεί και ταξινομηθεί από ειδικούς συντάκτες, σε αντιδιαστολή με την λειτουργία που επιτελούν οι μηχανισμοί συλλογής των αυτόματων Μ.Α.. Σε αυτό το είδος μηχανών αναζήτησης η δημιουργία των καταλόγων γίνεται με την συνδρομή του δημιουργού και κατόχου της ιστοσελίδας/ιστοχώρου. Πιο συγκεκριμένα ο δημιουργός της ιστοσελίδας αποστέλλει στην Μ.Α. μια σύντομη περιγραφή σχετικά με το τι παρουσιάζει ή τι υπηρεσίες προσφέρει ο ιστοχώρος του καθώς και την διεύθυνση του

Ομοιόμορφου Εντοπιστή Πόρου του. Εάν εγκριθεί η αίτηση αποδοχής, τότε ειδικοί συντάκτες κατατάσσουν τον ιστοχώρο στην κατάλληλη θεματική ενότητα ή κατηγορία με σκοπό να παρουσιαστεί ως ταξινομημένη πληροφορία στον χρήστη της Μ.Α..

Αυτό το είδος Μ.Α. προσφέρει συνήθως πιο συγκεντρωτικά αποτελέσματα σε σχέση με τις αυτόματες Μ.Α.. Αυτό γιατί η αναζήτηση και το ταίριασμα των αποτελεσμάτων πραγματοποιείται βάσει των περιγραφών που έχουν σταλεί και όχι βάσει του κειμένου που δημοσιεύεται σε μια ιστοσελίδα. Έτσι οι Θεματικοί Κατάλογοι πλεονεκτούν σε σχέση με τις αυτόματες Μ.Α. όταν το περιεχόμενο της ιστοσελίδας είναι δυναμικό ενώ η θεματική ενότητα παραμένει η ίδια. Η ανανέωση όσον αφορά την περιγραφή του ιστοχώρου γίνεται πάλι με μια αίτηση στον διαχειριστή της Μ.Α. Παραδείγματα Καταλόγων είναι οι Yahoo! [YAHOO] και DMOZ (Open Directory Project) [DMOZ].

4.4 Υβριδικές Μ.Α.

Οι Μ.Α. σήμερα γενικά κατατάσσονται είτε στις αυτόματες Μ.Α. είτε στους Θεματικούς Καταλόγους. Παρόλα αυτά κάποιες από αυτές παρουσιάζουν τα αποτελέσματα τους στους τελικούς χρήστες με διαφορετικό τρόπο από αυτό που συνήθως χρησιμοποιούν. Με άλλα λόγια ένας Θεματικός Κατάλογος ενδέχεται να επιστρέψει κάποια αποτελέσματα που βασίζονται σε τεχνικές απορρέουν από αυτόματα προγράμματα όπως είναι οι "αράχνες" και τα ρομπότ. Αυτό οφείλεται στο ότι πολλές Μ.Α. έστω και διαφορετικού τύπου συνεργάζονται μεταξύ τους, υποστηρίζοντας τα ευρετήριά τους εκατέρωθεν. Παράδειγμα αποτελεί ο Κατάλογος Yahoo! ο οποίος επιστρέφει και αποτελέσματα σε συνεργασία με την αυτόματη Μ.Α. Google. Ο χρήστης λοιπόν της Yahoo! ενδέχεται να λάβει αποτελέσματα που του προσφέρονται βάσει των μηχανισμών μιας αυτόματης Μ.Α. και όχι ταξινομημένα όπως θα περίμενε. Αυτό έχει παρατηρηθεί ότι συμβαίνει συνήθως για τα πιο δυσνόητα ερωτήματα που θα υποβληθούν.

4.5 Άλλες Μ.Α.

Το συντριπτικό ποσοστό των Μ.Α. που χρησιμοποιούνται σήμερα στο Διαδίκτυο κατατάσσονται στις παραπάνω τρεις κατηγορίες. Παρόλα αυτά υπάρχουν κάποιες Μ.Α. που λειτουργούν με διαφορετικό τρόπο όσον αφορά την επικοινωνία τους με τον τελικό χρήστη. Έτσι υπάρχουν Μ.Α. όπου υποστηρίζουν ότι πραγματοποιούν διαδικασίες επεξεργασίας φυσικής γλώσσας, στα ερωτήματα των χρηστών. Η επεξεργασία φυσικής γλώσσας είναι το κλειδί για την επόμενη γενιά των μηχανών αναζήτησης [Lawrence,99b]. Όμως, λόγω του τεράστιου όγκου πληροφορίας που πρέπει να επεξεργαστεί και του σύντομου χρονικού διαστήματος απόκρισης που πρέπει να πληροί μια σύγχρονη υπηρεσία πληροφορίας, τα επιστρεφόμενα αποτελέσματα υπολείπονται σε πληρότητα και ακρίβεια. Για αυτόν τον λόγο, δεν είναι ευρέως διαδεδομένες. Ένα παράδειγμα είναι η υπηρεσία αναζήτησης Ask Jeeves [AJ].

Επιπρόσθετα υπάρχουν οι Μ.Α. οι οποίες προσφέρουν τα επεξεργασμένα αποτελέσματά τους επί πληρωμής. Αυτές οι Μ.Α. συνήθως συντάσσουν ιστοσελίδες και κατ' επέκταση ιστοχώρους που προσφέρουν οικονομικές και διαφημιστικές υπηρεσίες. Ο κάτοχος του ιστοχώρου που συντάσσεται πληρώνει ανάλογα με την προώθηση που θα του παρέχεται από την υπηρεσία αναζήτησης. Ο χρήστης από την πλευρά του, πληρώνει βάσει του όγκου των πληροφοριών που λαμβάνει από τα επιστρεφόμενα αποτελέσματα, ενώ πρέπει να τονιστεί ότι τα αποτελέσματα που ταξινομούνται στις υψηλότερες θέσεις κοστίζουν περισσότερο.

Παράδειγμα μιας Μ.Α. όπου ο χρήστης πληρώνει βάσει του εύρους της αναζήτησής του είναι η Overture. Τέλος υπάρχουν και οι λεγόμενες μηχανές μετα-αναζήτησης, οι οποίες επιστρέφουν αποτελέσματα που προέρχονται από συνδυασμό αποτελεσμάτων άλλων Μ.Α.. Με άλλα λόγια οι μηχανές μετα-αναζήτησης αποστέλλουν την ερώτηση του χρήστη σε ένα

πλήθος από Μ.Α., συλλέγουν όλα ή ένα μέρος από τα ξεχωριστά αποτελέσματα και αφαιρώντας τα κοινά τα παρουσιάζουν πίσω στον τελικό χρήστη. Οι μηχανές μετα-αναζήτησης χρησιμοποιούνται όλο και περισσότερο σήμερα, αφού υπερτερούν στο τομέα της πληρότητας όσον αφορά τα πιο σχετικά αποτελέσματα. Γνωρίζοντας ότι ένας χρήστης μιας Μ.Α. συνήθως δεν αναζητά πληροφορίες σχετικές με ένα υποβαλλόμενο ερώτημα πέραν κάποιου βαθμού κατάταξης, οι μηχανές μετα-αναζήτησης επεξεργάζονται αυτό το καθορισμένο ποσό των πιο σχετικών αποτελεσμάτων ανά χρησιμοποιούμενη υπηρεσία αναζήτησης. Έτσι, όταν αφαιρεθούν τα διπλότυπα πεδία ο χρήστης δεν χάνει κανένα από τα μη κοινά και πιο σχετικά αποτελέσματα που δεν θα λάμβανε εάν δεν χρησιμοποιούσε από μόνος του όλες τις εμπλεκόμενες Μ.Α. Παραδείγματα μηχανών μετα-αναζήτησης είναι οι Copernic [COP], Ixquick [IXQ] και UMSE [Anagnostopoulos,02].

5 Χαρακτηριστικά των Μ.Α.

Στην ενότητα αυτή θα αναλυθούν βασικά χαρακτηριστικά που διέπουν την λειτουργία των Μ.Α. Είναι δε σκόπιμο τα χαρακτηριστικά αυτά να προσδιοριστούν και να ταξινομηθούν με τέτοιον τρόπο ώστε να γίνει αντιληπτή η ανομοιογένεια που υπάρχει στις λειτουργίες αυτές καθώς και να εξαχθούν χρήσιμα συμπεράσματα και τρόποι αντιμετώπισης των προβλημάτων που ανακύπτουν κατά την διάρκεια μιας αναζήτησης στο Διαδίκτυο. Επιπλέον, τα χαρακτηριστικά αυτά διαχωρίζονται σε δύο κατηγορίες. Στα εξωτερικά χαρακτηριστικά που περιγράφουν τις λειτουργίες μιας Μ.Α. όσον αφορά τα βήματα της "σύλληψης", της συλλογής και της σύνταξης των ιστοχώρων και των ιστοσελίδων που βρίσκονται στο Διαδίκτυο. Ακόμα αναλύονται τα εσωτερικά χαρακτηριστικά που περιγράφουν τις λειτουργίες και τους μηχανισμούς που προσφέρει μια υπηρεσία αναζήτησης για την ανάκτηση των πληροφοριών.

5.1 Εξωτερικά χαρακτηριστικά

Τα χαρακτηριστικά αυτά όπως προαναφέρθηκε αφορούν τις λειτουργίες της Μ.Α. όσον αφορά τα βήματα της "σύλληψης", της συλλογής και της σύνταξης των ιστοσελίδων. Παρότι ο χρήστης δεν κάνει χρήση αυτών των χαρακτηριστικών, επηρεάζουν την αναζήτησή του και τα επιστρεφόμενα αποτελέσματα που λαμβάνει. Από την άλλη πλευρά βέβαια η γνώση αυτών των χαρακτηριστικών είναι ιδιαίτερα χρήσιμη για τους υπεύθυνους και τους δημιουργούς των ιστοσελίδων. Αυτό γιατί γνωρίζοντας τον τρόπο με τον οποίο μια Μ.Α. επεξεργάζεται τις πληροφορίες, γίνεται γνωστός και ο τρόπος κατάταξης της ιστοσελίδας ανάλογα με τις ερωτήσεις που υποβάλλει ο χρήστης. Παρακάτω αναλύονται τα εξωτερικά χαρακτηριστικά των Μ.Α.

5.1.1 Χαρακτηριστικά αυτόματης αναζήτησης ιστοσελίδων

Τα παρακάτω χαρακτηριστικά προσδιορίζουν λειτουργίες αυτόματων και υβριδικών Μ.Α. των οποίων οι βασικές λειτουργίες και τα προγράμματα έχουν περιγραφεί παραπάνω.

Βαθιά αναζήτηση

Το χαρακτηριστικό αυτό αφορά συνήθως υβριδικές Μ.Α. στις οποίες αποστέλλονται οι διευθύνσεις των Ομοιόμορφων Εντοπιστών Πόρων και κατόπιν οι μηχανές αυτές αναζητούν τις ιστοσελίδες αυτής της διεύθυνσης με σκοπό να τις συντάξουν στα ευρετήριά τους.

Υποστήριξη πλαισίων

Το χαρακτηριστικό αυτό περιγράφει την δυνατότητα μιας Μ.Α. να επεκτείνει τις λειτουργίες της και μέσω των υπερσυνδέσεων που αντιστοιχούν σε πλαίσια της εξεταζόμενης ιστοσελίδας [Chakrabarti,99]. Οι υπηρεσίες αναζήτησης που υποστηρίζουν την λειτουργία αυτή πλεονεκτούν σε σχέση με τις άλλες γιατί ανιχνεύουν ένα επιπρόσθετο σημαντικό ποσοστό πληροφορίας. Η σημερινή άλλωστε μορφή των ιστοσελίδων κυρίως σε ιστοχώρους μεγάλης επισκεψιμότητας καθιστά πολύ χρήσιμη την λειτουργία αυτή.

Χαρτογράφηση εικόνων

Η λειτουργία αυτή είναι ιδιαιτέρως σημαντική όσον αφορά την σύνταξη των εικόνων και των εικονιδίων που συμπληρώνουν το περιεχόμενο μιας ιστοσελίδας. Ολοένα και περισσότερες Μ.Α. υποστηρίζουν αυτή την λειτουργία παρέχοντας έτσι την δυνατότητα στον χρήστη να ανακτήσει εικόνες και φωτογραφίες από το Διαδίκτυο.

Αποτροπή αυτόματης ανίχνευσης ιστοχώρου

Οι Μ.Α. που υποστηρίζουν την λειτουργία αυτή, επιτρέπουν στον δημιουργό ή τον κάτοχο του ιστοχώρου να εμποδίσει τις λειτουργίες της αυτόματης και της βαθιάς αναζήτησης όσον

αφορά τις αυτόματες και τις υβριδικές Μ.Α. αντίστοιχα. Το αναφαίρετο δικαίωμα της μη-προβολής ή της "ανωνυμίας" υποστηρίζεται από ένα πιστοποιημένο πρωτόκολλο [WRP1]. Η αποτροπή αυτή γίνεται με την εφαρμογή μιας εντολής που προστίθεται σε ένα ειδικά καθορισμένο πεδίο των μετα-ετικετών στην αρχή του πηγαίου κώδικα της Γλώσσας Υπερκείμενης Σήμανσης της ιστοσελίδας.

Αποτροπή αυτόματης ανίχνευσης ιστοσελίδας

Η λειτουργία αυτή είναι παρόμοια με την παραπάνω, με τη διαφορά όμως ότι η χρήση της αποτρέπει την ανίχνευση εκείνων των ιστοσελίδων που αναφέρονται σε ένα πιστοποιημένο πρωτόκολλο [WRP2]. Έτσι η χρήση αυτού του χαρακτηριστικού δίνει τη δυνατότητα στον δημιουργό ενός ιστοχώρου να επιτρέψει την πρόσβαση αυτόματων προγραμμάτων με σκοπό την ανεύρεση πληροφορίας, προστατεύοντας παράλληλα ορισμένες ιστοσελίδες από την λειτουργία αυτή.

Αναφορά από άλλες υπερσυνδέσεις

Το χαρακτηριστικό αυτό απαντάται στις πλέον σύγχρονες Μ.Α., όπου αποτελεί ταυτόχρονα και ένα μέτρο για τον αν κάποιες ιστοσελίδες θα περιληφθούν στους καταλόγους και τα ευρετήρια αναζήτησης.

Ανίχνευση ανανέωσης περιεχομένου

Η δυνατότητα ανίχνευσης της συχνότητας αλλαγής/ανανέωσης μιας ιστοσελίδας εντοπίζεται στην περίπτωση που κάποια Μ.Α. υποστηρίζει τη λειτουργία αυτή. Πρόκειται για μια πολύ σημαντική ικανότητα που μπορεί να προσδώσει και πληροφορίες σχετικά τον ρυθμό επισκεψιμότητας μιας ιστοσελίδας, αφού ο συχνός ρυθμός ανανέωσης σχετίζεται με το μέγεθος αυτό.

Ειδική προβολή με πληρωμή

Μερικές Μ.Α. υποστηρίζουν κάποια προγράμματα τα οποία προβάλλουν περισσότερο επί πληρωμή ορισμένες σχετικές ιστοσελίδες ανάλογα με ερωτήσεις χρηστών. Ουσιαστικά τα προγράμματα αυτά προωθούν και διαφημίζουν κάποια σχετικά αποτελέσματα στον τελικό χρήστη.

Έλεγχος διεύθυνσης Ομοιόμορφων Εντοπιστών Πόρων

Μέσω του χαρακτηριστικού αυτού γίνεται έλεγχος για την περίπτωση που μια συγκεκριμένη ιστοσελίδα έχει καταταχθεί στους καταλόγους μιας άλλης Μ.Α.

5.1.2 Χαρακτηριστικά σύνταξης ιστοσελίδων

Στην υπό-ενότητα αυτήν αναλύονται τα χαρακτηριστικά των Μ.Α. βάσει των οποίων συντάσσουν την επεξεργασμένη πληροφορία.

Σύνταξη "ορατού" κειμένου

Οι περισσότερες Μ.Α. υποστηρίζουν αυτή τη λειτουργία, συντάσσοντας όλο ή μερικό από το κείμενο που διανέμεται στο Διαδίκτυο.

Αποβολή κοινών λέξεων

Όταν μια Μ.Α. υποστηρίζει αυτό το χαρακτηριστικό, δεν συντάσσει στους καταλόγους και τα ευρετήριά της, κοινές λέξεις που δεν προσδίδουν κάποια χρήσιμη πληροφορία. Η αποβολή των λέξεων αυτών εκτός του ότι οδηγεί σε ταχύτερους ρυθμούς επεξεργασίας βοηθά στην εξοικονόμηση αποθηκευτικού χώρου.

Υποστήριξη πεδίων μετα-ετικετών

Η λειτουργία αυτή προσδιορίζει τη δυνατότητα μιας Μ.Α. να επεξεργάζεται τις πληροφορίες που βρίσκονται στα πεδία των μετα-ετικετών.

Δημιουργία παραγώγων λέξεων

Το χαρακτηριστικό αυτό επιτρέπει στις Μ.Α. που το υποστηρίζουν, να αναλύουν τις υποβαλλόμενες ερωτήσεις στις συνιστώμενες ρίζες λέξεων και να επεκτείνουν την αναζήτηση βάσει των παραγώγων που προέρχονται από αυτές.

5.1.3 Χαρακτηριστικά κατάταξης των αποτελεσμάτων

Τα χαρακτηριστικά αυτά προσδιορίζουν τις βασικές λειτουργίες που χρησιμοποιούν οι υπηρεσίες αναζήτησης προκειμένου να κατατάξουν σχετικά στο χρήστη τα επιστρεφόμενα αποτελέσματα που ανταποκρίνονται στις αναζητήσεις του. Συνήθως οι περισσότερες Μ.Α. υπολογίζουν την συχνότητα και τη θέση κάποιου σχετικού όρου στην εξεταζόμενη ιστοσελίδα. Παρόλα αυτά λαμβάνονται υπόψη και ορισμένα άλλα χαρακτηριστικά τα οποία αναφέρονται παρακάτω και που βάσει αυτών αποδίδεται περισσότερο ή μικρότερο βάρος στον όρο αυτό,.

Στάθμιση των πεδίων μετα-ετικετών

Ορισμένες Μ.Α. λαμβάνουν περισσότερο υπόψη το κείμενο που βρίσκεται στα πεδία των μετα-ετικετών. Τα πεδία αυτά περιέχουν λέξεις-κλειδιά, όρους και φράσεις που περιγράφουν και προσδιορίζουν με μεγαλύτερη σαφήνεια το περιεχόμενο του ιστοχώρου. Έτσι, προσδίδεται σε αυτούς μεγαλύτερο βάρος, το οποίο βρίσκεται συνήθως σε άμεση εξάρτηση με την συχνότητα εμφάνισης του ίδιου όρου στο υπόλοιπο κειμενικό περιεχόμενο της ιστοσελίδας.

Στάθμιση σε αναφορές από άλλες υπερσυνδέσεις

Όπως αναφέρθηκε και παραπάνω πολλές υπηρεσίες αναζήτησης μπορούν να προσδιορίσουν το πόσο δημοφιλής είναι μια ιστοσελίδα σε συνάρτηση με τον αριθμό των υπερσυνδέσεων που αναφέρονται σε αυτές. Έτσι, εκμεταλλεύονται το γεγονός αυτό αναθέτοντας τιμές στάθμισης με μεγαλύτερη βαρύτητα στους όρους που εμφανίζονται στις πιο δημοφιλείς ιστοσελίδες.

Στάθμιση ανάλογα με την επιλογή των αποτελεσμάτων

Η λειτουργία αυτή ουσιαστικά αφορά σε έναν μηχανισμό σχετικής ανατροφοδότησης όπου η Μ.Α. λαμβάνει υπόψη τα αποτελέσματα που επιλέγουν οι χρήστες σε προηγούμενες σχετικές ερωτήσεις. Η βαρύτητα σε αυτόν τον τρόπο στάθμισης δίνεται στα επιστρεφόμενα αποτελέσματα. Η πρόταση αυτή εφαρμόστηκε αρχικά με επιτυχία από την υπηρεσία αναζήτησης HotBot, αλλά παραμένει ακόμα σε ερευνητικά και πειραματικά επίπεδα [SER,99].

5.1.4 Χαρακτηριστικά αναγνώρισης και αντιμετώπιση τεχνικών Spam

Οι περισσότερες και σημαντικότερες Μ.Α. έχουν μηχανισμούς προστασίας που αποτρέπουν την αλλοίωση της ορθής κατάταξης των επιστρεφόμενων αποτελεσμάτων. Αυτό συμβαίνει γιατί οι περισσότεροι δημιουργοί ιστοσελίδων γνωρίζουν σε κάποιο ικανοποιητικό βαθμό τον τρόπο με τον οποίο κατατάσσει μια υπηρεσία αναζήτησης τα αποτελέσματά της. Έτσι, λοιπόν εάν είναι γνωστό ότι μια Μ.Α. χρησιμοποιεί κυρίως την συχνότητα εμφάνισης όρων ως κριτήριο για την στάθμιση των αποτελεσμάτων της, ο δημιουργός ενός ιστοχώρου ενδέχεται να συγκεντρώνει συγκεκριμένους όρους σε διάφορα πεδία ετικετών, προσπαθώντας έτσι να παραπλανήσει τις λειτουργίες μιας Μ.Α. Το φαινόμενο αυτό αντιμετωπίζεται με αρνητική στάθμιση των ιστοσελίδων ή και αποκλεισμό αυτών από τα επιστρεφόμενα αποτελέσματα.

Αντιμετώπιση "αόρατου" κειμένου

Ένας άλλος τρόπος αλλοίωσης των επιστρεφόμενων αποτελεσμάτων που συχνά χρησιμοποιούν οι δημιουργοί των ιστοσελίδων είναι η προσθήκη κειμενικής πληροφορίας στο ίδιο χρώμα με το φόντο της ιστοσελίδας όπως αυτή εκδίδεται στο Διαδίκτυο. Με τον τρόπο αυτό η περιττή πληροφορία δεν είναι ορατή στον τελικό χρήστη, από την άλλη όμως αλλοιώνει τους υπολογισμούς στάθμισης των όρων, εφόσον η πληροφορία αυτή επεξεργάζεται από τους μηχανισμούς συλλογής και σύνταξης. Έτσι πολλές Μ.Α. επεξεργάζονται την πληροφορία σε άμεση συνάρτηση με το χρώμα του φόντου και των γραμμάτων σε μια ιστοσελίδα.

Αντιμετώπιση κειμένου ελάχιστου μεγέθους

Συνδυάζοντας τις δύο παραπάνω τεχνικές, πολλοί δημιουργοί ιστοσελίδων δεν παραθέτουν αόρατη πληροφορία. Όμως προσπαθούν να αλλοιώσουν το σύστημα στάθμισης με κείμενα που αποτελούνται από γραμματοσειρές υπερβολικά μικρού μεγέθους. Το ανθρώπινο μάτι του χρήστη δεν αντιλαμβάνεται την περιττή πληροφορία, όχι όμως και το σύστημα στάθμισης. Έτσι για άλλη μια φορά οι Μ.Α. έχουν μηχανισμούς που αντιλαμβάνονται αυτήν την περιττή πληροφορία και αποκλείουν ή σταθμίζουν αρνητικά τις συγκεκριμένες ιστοσελίδες που εντοπίζονται.

5.2 Εσωτερικά χαρακτηριστικά ή λειτουργίες ανάκτησης πληροφορίας

Μετά από την συλλογή, επεξεργασία, σύνταξη και αποθήκευση των δεδομένων μια Μ.Α. ή ένας Κατάλογος έχει μηχανισμούς που επιτρέπουν στο χρήστη την ανάκτηση πληροφορίας ανάλογα με τις απαιτήσεις του. Οι μηχανισμοί αυτοί έχουν κάποια χαρακτηριστικά τα οποία δεν είναι κοινά για όλες τις Μ.Α.. Κρίνεται λοιπόν σκόπιμο να αναφερθούν οι διαφορετικές αυτές λειτουργίες ώστε να φανεί για άλλη μια φορά ο βαθμός της ανομοιογένειας και της ασυμβατότητας που χαρακτηρίζει τις υπηρεσίες αναζήτησης. Είναι φανερό ότι σε αντίθεση με τα εξωτερικά χαρακτηριστικά που ενδιαφέρουν τους δημιουργούς και κατόχους των ιστοσελίδων, τα εσωτερικά χαρακτηριστικά αφορούν αποκλειστικά τους τελικούς χρήστες των Μ.Α.. Ομοίως όπως στην προηγούμενη ενότητα, αναλύονται στα παρακάτω τα εσωτερικά χαρακτηριστικά των Μ.Α., ενώ στο Παράρτημα Β του κεφαλαίου αυτού, παρατίθεται ένας συγκεντρωτικός πίνακας αυτών ανά υπηρεσία αναζήτησης [searchengineshowdown]. Οι συγκρινόμενες Μ.Α. είναι οι AllTheWeb, AltaVista, Direct Hit, Excite, Google, Northern Light, Lycos, Inktomi, AOL Search, HotBot, MSN Search και Teoma.

5.2.1 Μαθηματικές Εντολές αναζήτησης – εντολές Boolean τύπου

Το Internet είναι μια αχανής βάση δεδομένων, γι αυτό το περιεχόμενο του πρέπει να ερευνάται βάσει των ίδιων κανόνων που διέπουν την έρευνα σε βάσεις δεδομένων, δηλαδή τη λογική των τελεστών Boolean τύπου. Η λογική των τελεστών Boole είναι μια μέθοδος για τη διατύπωση λογικών απόψεων με μαθηματικές εντολές. Οι κυριότεροι τελεστές Boole είναι οι εξής:

Τελεστής "AND" ή αλλιώς "+":

Ο τελεστής AND επιτρέπει στο χρήστη να προσθέσει λέξεις στις αναζητήσεις του, η εμφάνιση των οποίων είναι υποχρεωτική στα επιστρεφόμενα αποτελέσματα.. Η συνάρτηση δηλαδή "όρος$_1$" AND "όρος$_2$" είναι αληθής, όταν στο αποτέλεσμα περιέχονται οι όροι "όρος$_1$" και "όρος$_2$".

Τελεστής "NOT" ή αλλιώς "-":

Με τον τελεστή αυτόν ο χρήστης μπορεί να αποκλείσει όρους από τις αναζητήσεις του. Η συνάρτηση NOT "όρος$_1$" είναι αληθής όταν στο επιστρεφόμενο αποτέλεσμα ο όρος "όρος$_1$" δεν περιέχεται.

Τελεστής "OR":

Με τη χρήση του τελεστή αυτού ο χρήστης αναζητάει στα αποτελέσματα που λαμβάνει τουλάχιστον έναν από τους όρους που συνιστούν την ερώτησή του. Έτσι, η συνάρτηση "όρος$_1$" OR "όρος$_2$" είναι αληθής όταν τουλάχιστον ένας από τους όρους εμφανίζεται στα επιστρεφόμενα αποτελέσματα.

Ο τελεστής "ADJ" ή αλλιώς "":

Ο τελεστής ADJ προέρχεται από την Αγγλική λέξη Adjacency που σημαίνει γειτνίαση, έχει ταυτόσημη χρήση που έχουν και τα εισαγωγικά "". Χρησιμοποιώντας αυτόν τον τελεστή στις αναζητήσεις του, ο χρήστης εξασφαλίζει ότι οι εμπλεκόμενοι όροι θα περιέχονται διαδοχικά ο ένας μετά τον άλλο. Έτσι , η ερώτηση "όρος$_1$" ADJ "όρος$_2$" ADJ "όρος$_3$", θα επιστρέφει αποτελέσματα που θα περιέχουν και τους τρεις όρους με την ίδια ακολουθία. Ο τελεστής αυτός είναι ιδιαίτερα σημαντικός γιατί επιτρέπει την αναζήτηση ολοκληρωμένων προτάσεων ή φράσεων.

Ο τελεστής "NEAR":

Με τη χρήση του τελεστή NEAR ο χρήστης εξασφαλίζει ότι οι όροι που χωρίζονται από αυτόν τον τελεστή βρίσκονται σε κάποια συγκεκριμένη μέγιστη απόσταση μεταξύ τους στο περιεχόμενο της ιστοσελίδας. Η απόσταση μεταξύ αυτή μεταξύ των όρων διαφέρει ανάλογα με την Μ.Α., ενώ καθορίζεται αποκλειστικά από αυτήν.

Ο τελεστής "FAR":

Η χρήση του τελεστή αυτού έχει τα ακριβώς αντίθετα αποτελέσματα από αυτήν του NEAR.

Φώλιασμα ή Σύνθεση τελεστών:

Οι περισσότερες Μ.Α. επιτρέπουν την σύνθεση όλων των παραπάνω τελεστών, με σκοπό την σύνταξη πιο πολύπλοκων ερωτήσεων που βοηθούν το χρήστη να καθορίσει με μεγαλύτερη ακρίβεια τις πληροφοριακές απαιτήσεις του. Για παράδειγμα η ερώτηση "Ακριβής Φράση" AND ("όρος1" OR "όρος2") ή αλλιώς "Ακριβής" ADJ "Φράση" AND ("όρος1" OR "όρος2") είναι αληθής όταν τα επιστρεφόμενα αποτελέσματα περιέχουν την ακολουθία των όρων "Ακριβής Φράση" και τουλάχιστον έναν από τους όρους "όρος1" ή "όρος2".

5.2.2 Ενισχυμένες εντολές αναζήτησης

Ορισμένες Μ.Α. προσφέρουν ποικίλους τρόπους για να καθορίσουν και να ελέγξουν τις αναζητήσεις των χρηστών. Μερικές ειδικές εντολές μπορούν να τεθούν ως τμήμα της ερώτησης που θέτει ο χρήστης. Έτσι είναι δυνατή η αναζήτηση σχετικών αποτελεσμάτων βάσει παραδείγματος χάριν της διεύθυνσης Ομοιόμορφου Εντοπιστή Πόρων ή της διεύθυνσης ιστοχώρου. Παρακάτω παρατίθενται οι εντολές αυτές.

Αναζήτηση βάσει διεύθυνσης Ιστοχώρου

Πραγματοποιείται αναζήτηση σε δεδομένο από το χρήστη Δικτυακό Τόπο.

Αναζήτηση βάσει διεύθυνσης Ομοιόμορφου Εντοπιστή Πόρου

Πραγματοποιείται αναζήτηση σε δεδομένη από το χρήστη διεύθυνση Ομοιόμορφου Εντοπιστή Πόρου.

Αναζήτηση βάσει υπερσυνδέσμου

Πραγματοποιείται αναζήτηση σε δεδομένη από το χρήστη διεύθυνση υπερσυνδέσμου.

Χρήση χαρακτήρων Μπαλαντέρ ("?",""):*

Στις εντολές ενισχυμένης αναζήτησης συμπεριλαμβάνονται οι χαρακτήρες Μπαλαντέρ. Πρόκειται για ιδιαίτερα χρήσιμα σύμβολα που χρησιμοποιούνται για να αντικατασταθούν

κάποιοι χαρακτήρες με όλους τους πιθανούς συνδυασμούς τους. Έτσι, το αγγλικό ερωτηματικό "?" μπορεί να αντικαταστήσει ένα οποιοδήποτε γράμμα της αλφαβήτου, ενώ αστερίσκος "*" αντικαθιστά ολόκληρη σειρά γραμμάτων. Θέτοντας δηλαδή τον όρο "ca?s" η Μ.Α. αναζητά όρους όπως cars ή cats. Ο όρος όμως "ca*s" αν και αντιστοιχεί στους παραπάνω όρους cars ή cats, αντιστοιχεί ταυτόχρονα για παράδειγμα και στο όρο "cameras" "careers".

5.2.3 Χαρακτηριστικά αναζήτησης

Οι σημαντικότερες Μ.Α. χρησιμοποιούν ορισμένα χαρακτηριστικά αναζήτησης για να βοηθήσουν κυρίως τους αρχάριους χρήστες. Στην ενότητα αυτήν συνοψίζονται μερικά από τα σημαντικότερα χαρακτηριστικά γνωρίσματα αναζήτησης που είναι διαθέσιμα.

Σχετικές αναζητήσεις

Το χαρακτηριστικό αυτό παρέχεται για να βοηθήσει τους χρήστες ώστε να πραγματοποιήσουν πιο συγκεκριμένες αναζητήσεις ή να τους προτείνει παρεμφερείς ερωτήσεις άλλων χρηστών. Έτσι, οι Μ.Α. που το υποστηρίζουν εμφανίζουν συνήθως μια λίστα με υπερσυνδέσεις με σχετικές αναζητήσεις χρησιμοποιώντας γνωστούς όρους, οδηγώντας τον χρήστη συχνά σε καλύτερα αποτελέσματα.

Συγκέντρωση αποτελεσμάτων

Το χαρακτηριστικό αυτό αποτρέπει την ταυτόχρονη εμφάνιση πολλών ιστοσελίδων που ανήκουν στον ίδιο ιστοχώρο στα τελικά αποτελέσματα. Έτσι παρουσιάζεται ένα πιο συνοπτικό και αντιπροσωπευτικό δείγμα απαντήσεων ενώ ο χρήστης έχει μεγαλύτερη πιθανότητα να ανακτήσει μια ενδιαφέρουσα πληροφορία γρήγορα.

Εύρεση σχετικών ιστοσελίδων

Πρόκειται για μια λειτουργία παρόμοια με την λειτουργία που παρέχει το χαρακτηριστικό των Σχετικών Αναζητήσεων που αναφέρθηκε παραπάνω. Η Μ.Α. προτείνει στον χρήστη παρόμοιες ιστοσελίδες με αυτές που έχουν ανεβρεθεί. Ουσιαστικά η λειτουργία αυτή έχει συμβουλευτικό ρόλο όσον αφορά τις προτιμήσεις του χρήστη.

Δημιουργία παραγώγων λέξεων

Υποστηρίζοντας το χαρακτηριστικό αυτό μια Μ.Α. έχει τη δυνατότητα να αναζητήσει για τις παραλλαγές μιας λέξης που τίθεται από το χρήστη βάσει της ρίζας της.

Εσωτερική Αναζήτηση

Οι Μ.Α. που υποστηρίζουν αυτήν τη λειτουργία επιτρέπουν στους χρήστες τους να πραγματοποιήσουν επιπρόσθετες ερωτήσεις πάνω στο σύνολο των επιστρεφόμενων αποτελεσμάτων. Πρόκειται για μια πολύ χρήσιμη λειτουργία γιατί το σύνολο των αποτελεσμάτων παραμένει αναλλοίωτο, ενώ ταυτόχρονα “στενεύουν” οι αναζητήσεις του χρήστη.

Αναζήτηση κρυφών – αποθηκευμένων ιστοσελίδων

Ορισμένες φορές είναι χρήσιμο να φανεί η αρχική έκδοση μιας ιστοσελίδας που υπέστη επεξεργασία από ένα αυτόματο πρόγραμμα μιας Μ.Α.. Το χαρακτηριστικό αυτό επιτρέπει την ανάκτηση ιστοσελίδων όπως έχουν συνταχθεί πριν ανανεωθούν εκ νέου από τους μηχανισμούς μιας υπηρεσίας αναζήτησης. Έτσι, είναι δυνατή η παρουσίαση ιστοσελίδων που δεν είναι πλέον ενεργές.

Αναζήτηση ιστοσελίδων βάσει γλωσσικού περιεχομένου

Υπάρχουν αρκετές Μ.Α. που πραγματοποιούν αναζήτηση ιστοσελίδων βάσει του γλωσσικού της περιεχομένου. Πρόκειται ουσιαστικά για μια λειτουργία που ταξινομεί την πληροφορία που είναι καταγεγραμμένη σε διαφορετικές γλώσσες. Βέβαια, η ανάκτηση μιας τέτοιας

πληροφορίας δεν είναι πάντα ορθή. Οι Μ.Α. γενικά χρησιμοποιούν λεξικά με συγκεκριμένους όρους υψηλής συχνότητας για διαφορετικές γλώσσες, προκειμένου να προσδιορίσουν το γλωσσικό περιεχόμενο μιας ιστοσελίδας κατά τη διάρκεια της επεξεργασίας της. Έτσι σε περίπτωση όπου μια ιστοσελίδα δημοσιεύει το περιεχόμενό της πολύ-γλωσσικά, ενδέχεται να απορριφθεί λανθασμένα από τα επιστρεφόμενα τελικά αποτελέσματα.

Μετάφραση ιστοσελίδας

Το χαρακτηριστικό αυτό είναι πολύ χρήσιμο για τους χρήστες που δεν γνωρίζουν τη γλώσσα του περιεχομένου της ιστοσελίδας που τους ενδιαφέρει. Παρόλα αυτά πρέπει να τονιστεί, ότι ουσιαστικά πρόκειται για μια "a posteriori" διαδικασία που εφαρμόζεται μετά την εύρεση των σχετικών αποτελεσμάτων. Ορισμένες φορές είναι ιδιαίτερα χρονοβόρα, λόγω του ότι το αίτημα στέλνεται από τον χρήστη στον διακομιστή της Μ.Α., όπου πραγματοποιείται σε πραγματικό χρόνο η μετάφραση.

Έλεγχος και φιλτράρισμα "επικίνδυνου" περιεχομένου

Οι Μ.Α. που υποστηρίζουν αυτό το χαρακτηριστικό παρέχουν εάν ζητηθεί τη δυνατότητα να αποκλείσουν από τα επιστρεφόμενα αποτελέσματα ιστοσελίδες ή ιστοχώρους γενικά που δημοσιεύουν πορνογραφικό υλικό ή υλικό που προάγει τη βία και υποβαθμίζει την ανθρώπινη αξιοπρέπεια. Η ανίχνευση των ιστοσελίδων αυτών γίνεται με την ανίχνευση πορνογραφικών όρων που συντάσσονται από ομάδες ειδικών. Η λειτουργία αυτή δεν παρέχει τέλεια προστασία, όμως είναι ιδιαίτερα χρήσιμη εάν οι τελικοί χρήστες είναι ανήλικα παιδιά ελαχιστοποιώντας έτσι τον κίνδυνο προβολής επικίνδυνης, αντιπαιδαγωγικής και μη ωφέλιμης πληροφορίας.

5.2.4 Χαρακτηριστικά προσαρμογής απεικόνισης και προβολής

Μερικές Μ.Α. επιτρέπουν την προσαρμογή των επιστρεφόμενων αποτελεσμάτων σύμφωνα με επιλογές που ρυθμίζει ο χρήστης στην επιφάνεια διεπαφής που του προσφέρεται. Με αυτόν τον τρόπο οι Μ.Α. προσφέρουν στους χρήστες τους ένα όσο το δυνατόν φιλικό προς το χρήστη περιβάλλον. Παρακάτω αναφέρονται τα χαρακτηριστικά αυτά:

Ταξινόμηση αποτελεσμάτων ανά ημερομηνία

Πραγματοποιείται ταξινόμηση των ιστοσελίδων βάσει της ημερομηνίας που έχουν συνταχθεί. Ενδέχεται όμως κάποιες νεώτερες πηγές να συνταχθούν πιο αργά σε σχέση με παλιότερες ή ο ρυθμός ανανέωσης να διαφέρει από πηγή σε πηγή, με αποτέλεσμα η ταξινόμηση αυτή να μην είναι αξιόπιστη.

Ταξινόμηση αποτελεσμάτων σε καθορισμένο εύρος ημερομηνίας

Μερικές Μ.Α. περιορίζουν τα αποτελέσματά τους με τέτοιο τρόπο έτσι ώστε να προβάλλονται μόνο οι σελίδες που βρίσκονται μέσα σε μια συγκεκριμένο εύρος ημερομηνίας. Όπως έχει αναφερθεί και παραπάνω αυτό το χαρακτηριστικό γνώρισμα ενδέχεται να πάσχει από το γεγονός ότι οι ημερομηνίες σύνταξης της ιστοσελίδας δεν είναι αξιόπιστες.

Προβολή ημερομηνίας δημιουργίας ή μορφοποίησης της ιστοσελίδας

Μαζί με την περιγραφή των ιστοσελίδων, μερικές Μ.Α. παρουσιάζουν την ημερομηνία δημιουργίας, τροποποίησης ή ανανέωσής της. Για άλλη μια φορά τονίζεται ότι αυτές οι ημερομηνίες δεν είναι πάντα αξιόπιστες. Εντούτοις, παρέχουν μια χρήσιμη ένδειξη ως προς το πόσο ανανεωμένες ή ενημερωμένες είναι οι λίστες και τα ευρετήρια μιας Μ.Α..

Παροχή περιβάλλοντος ενισχυμένης αναζήτησης

Οι Μ.Α. που υποστηρίζουν αυτό το χαρακτηριστικό παρέχουν στο χρήστη ένα περιβάλλον όπου μπορούν να δημιουργήσουν πιο σύνθετες ερωτήσεις. Συνήθως στο περιβάλλον αυτό παρέχουν πληροφορίες για το πως συντάσσονται ερωτήσεις με τελεστές Boolean τύπου, πως

γίνεται το φώλιασμα σύνθετων ερωτήσεων και η χρήση των ειδικών χαρακτήρων Μπαλαντέρ.

Παροχή βοήθειας

Είναι επιπρόσθετες ιστοσελίδες όπου γίνεται αναλυτική αναφορά πάνω στη λειτουργίας της Μ.Α., επιπρόσθετες πληροφορίες σχετικά με την ιστορία της και λίστα με τις πιο συχνές ερωτήσεις και απορίες των χρηστών.

6 Επισκόπηση στο χώρο Μηχανών Αναζήτησης

Στην ενότητα αυτή θα γίνει μια σύντομη επισκόπηση σε συνολικά δέκα δημοφιλείς Μ.Α., εκ των οποίων οι οκτώ ανήκουν στην κατηγορία των Αυτόματων Μ.Α. και οι υπόλοιπες δύο στην κατηγορία των Θεματικών Καταλόγων. Επιπλέον γίνεται και μια αναφορά των λειτουργιών που τις ξεχωρίζουν μεταξύ τους, βάσει μιας μελέτης που έχει δημοσιεύεται στη [NAVNET]. Οι οκτώ Αυτόματες Μ.Α. είναι οι Google, Northern Light, AltaVista, Hotbot, Lycos, Direct Hit, Live search, AllTheWeb και Excite, ενώ οι δύο Θεματικοί Κατάλογοι είναι οι Yahoo! και DMOZ. Τέλος για την πληρότητα της διατριβής αυτής γίνεται και μια ανασκόπηση στις Ελληνικές υπηρεσίες αναζήτησης.

6.1 Επισκόπηση Αυτόματων Μηχανών Αναζήτησης

6.1.1 AllTheWeb

Η Μ.Α. AllTheWeb είναι από τις υπηρεσίες αναζήτησης που έχουν συντάξει ένα πολύ μεγάλο ποσό ιστοσελίδων του Παγκόσμιου Ιστού. Παρέχει ένα πολύ φιλικό περιβάλλον επικοινωνίας και δυνατότητα αναζήτησης αρχείων εικόνων και πολυμέσων. Είναι αρκετά γρήγορη και επιστρέφει συνήθως σχετικά αποτελέσματα, αλλά οι περιγραφές τους δεν είναι αρκετά βοηθητικές.

6.1.2 AltaVista

Η AltaVista [AV] είναι μια από τις πιο διάσημες Μ.Α. παγκοσμίως. Απαιτεί μεγάλη προσοχή στην διατύπωση της ερώτησης καθώς επιστρέφει πολλά αδιάφορα αποτελέσματα. Την αδυναμία της αυτή, έρχεται να αντιμετωπίσει η λειτουργία "Advance Search" που υποστηρίζει εκτεταμένη χρήση λογικών τελεστών, καθώς επίσης και η αναζήτηση ιστοσελίδων σε συγκεκριμένη γλώσσα, η μετάφραση, και οι έξυπνες τεχνικές φιλτραρίσματος των αποτελεσμάτων. Έχει δε, πολύ μεγάλη επιτυχία στο να επιστρέφει αποτελέσματα σχετικά με το χώρο της παγκόσμιας βιβλιογραφίας λόγω της συνεργασίας της με την γνωστή πύλη Amazon.

6.1.3 Direct Hit

Η υπηρεσία αναζήτησης Direct Hit προσφέρει στο χρήστη ομαδοποιημένα τα πιο σχετικά αποτελέσματα ενώ του επιτρέπει να παρέμβει άμεσα στο ποσό των αποτελεσμάτων αυτών. Ενδέχεται να χρησιμοποιήσει ευρετήρια άλλων Μ.Α. ή Καταλόγων, αλλά σε καμία περίπτωση δεν λειτουργεί ως Μηχανή Μετα-Αναζήτησης.

6.1.4 Excite

Στην σημερινή της μορφή η Μ.Α. Excite [EXCITE] έχει μετατραπεί σε μια Πύλη, η οποία χρησιμοποιείται συνήθως για υπηρεσίες ηλεκτρονικού εμπορίου. Είναι γρήγορη, φιλική προς το χρήστη ενώ η αρχική της σελίδα παραπέμπει σε μια υπηρεσία αναζήτησης με σαφή καταναλωτικό προσανατολισμό.

6.1.5 Google

Η μηχανή αναζήτησης Google [GOOGLE] αποτελεί την μετατροπή της εργασίας ενός μεταπτυχιακού φοιτητή του πανεπιστημίου του Stanford σε εμπορική εφαρμογή. Για την ανεύρεση απαντήσεων σε κάθε ερώτηση, η Μ.Α. αυτή χρησιμοποιεί δύο κριτήρια που δεν παρουσιάζονται σε άλλες μηχανές αναζήτησης. Το πρώτο κριτήριο χρησιμοποιεί το πλήθος των υπερσυνδέσμων που "δείχνουν" στην υπό επεξεργασία ιστοσελίδα, ενώ το δεύτερο την "σημαντικότητα" των παραπάνω υπερσυνδέσμων. Για παράδειγμα, το γεγονός ότι το ο όρος "AltaVista", αντιστοιχεί σε έναν δικτυακό τόπο, αυξάνει την βαρύτητά του πολύ περισσότερο από οποιονδήποτε άλλο υπερσύνδεσμο. Η Google, για την αξιολόγηση μίας ιστοσελίδας, χρησιμοποιεί επιπλέον μία πληθώρα κριτηρίων όπως το ποσοστό των λέξεων της ερώτησης που εμφανίζονται στον τίτλο και στο κείμενο της

ιστοσελίδας καθώς και την εγγύτητα των λέξεων μεταξύ τους. Η τεχνική αυτή είναι γνωστή και ως PageRank [Brin, 98]. Το περιβάλλον αυτής της Μ.Α. θεωρείται το πλέον φιλικό και εύχρηστο για τους χρήστες και αποτελεί ουσιαστικά την εμπορικότερη μηχανή αναζήτησης. Με την είσοδο μάλιστα στην αγορά δορυφορικών χαρτών αλλά και την ψηφιοποίηση βιβλιοθηκών την κατατάσσει ως την πιο ολοκληρωμένη μηχανή αναζήτησης.

6.1.6 Hotbot

Η Μ.Α. Hotbot [HB] παρουσιάζει αστάθεια στο συνολικό ποσό των συνταγμένων ιστοσελίδων που περιέχει που οφείλεται στη διαθεσιμότητα και τη λειτουργία των εξυπηρετητών του συστήματός της. Παρέχει μεγάλη ευελιξία στη διατύπωση ερωτήσεων και θεωρείται ότι μετά την Google είναι η καταλληλότερη υπηρεσία αναζήτησης για τους αρχάριους χρήστες.

6.1.7 Lycos

Η Μ.Α. Lycos [Lycos] είναι μια υπηρεσία αναζήτησης μεσαίου μεγέθους. Συνεργάζεται άμεσα με την Hotbot και ο δικτυακός χώρος της έχει την μορφή πύλης. Διαθέτει επιπλέον είδη ερωτήσεων για την ανεύρεση αρχείων εικόνας και ήχου.

6.1.8 Northern Light

Η Μ.Α. Northern Light [NL], εξειδικεύεται στην προσφορά αναζήτησης για περισσότερες από 5000 εφημερίδες, περιοδικά καθώς και διάφορα έγγραφα της Αμερικάνικης κυβέρνησης. Διαθέτει επίσης μηχανισμούς κατηγοριοποίησης των αποτελεσμάτων ώστε να μπορεί ο χρήστης να τα προσπελάσει ανάλογα με το θέμα τους, την "ηλικία" τους ή την πηγή, από την οποία προέρχονται. Δεν χρησιμοποιεί όμως ιδιαίτερα φίλτρα για την ανάκτηση των αποτελεσμάτων.

6.1.9 Live search

Η Live search είναι ο διάδοχος του MSN. και είναι η μηχανή αναζήτησης της Microsoft. Εγκαινιάστηκε τον Σεπτέμβριο του 2006, και χρησιμοποιεί τη δική του βάση δεδομένων. Η βάση δεδομένων εκτός των άλλων περιλαμβάνει εικόνες βίντεο αλλά και βιβλιοθήκες. Στα θετικά της είναι η πλήρης υποστήριξη λογικών τελεστών , αλλά δεν επιτρέπει ερωτήματα με περισσότερο από δέκα όρους.

6.2 Επισκόπηση Θεματικών Καταλόγων

6.2.1 DMOZ

Ο Θεματικός Κατάλογος DMOZ ή αλλιώς γνωστός και ως Open Directory Project [DMOZ], χρησιμοποιεί ως αξιολογητές και κριτές των πληροφοριών που παρέχει, εθελοντές χρήστες από οποιαδήποτε πλευρά του πλανήτη. Παρέχει δηλαδή στον απλό χρήστη την δυνατότητα να συμμετέχει ενεργά στο πρόβλημα της ταξινόμησης της πληροφορίας που διαχέεται στον Παγκόσμιο Ιστό, σε μια καθορισμένη θεματική κατηγορία, η οποία συνήθως αφορά το γνωστικό αντικείμενο ή την γεωγραφική θέση του. Επιπρόσθετα, εάν ένας δικτυακός χώρος ταξινομηθεί στον Κατάλογο DMOZ, αυτόματα θα σταλούν σχετικές πληροφορίες και σε άλλες συνεργαζόμενες υπηρεσίες αναζήτησης όπως η AOL Search, η AltaVista, η HotBot, η Google και η Lycos.

6.2.2 Yahoo!

Ο Θεματικός Κατάλογος Yahoo!, αποτελεί μια από τις πιο δημοφιλείς Μ.Α. παγκοσμίως. Είναι ιδιαίτερα διαδεδομένη στους αρχάριους χρήστες του Διαδικτύου που δεν την χρησιμοποιούν μόνο ως υπηρεσία αναζήτησης αλλά και ως Πύλη πληροφοριών καθώς προσφέρει και υπηρεσίες ηλεκτρονικού ταχυδρομείου ή συνομιλιών. Ως θεματικός κατάλογος δεν αξιολογεί το περιεχόμενο, αλλά οργανώνει θεματικά τους δικτυακούς τόπους που υποβάλλονται προς ένταξη στην υπηρεσία.

6.3 Υπηρεσίες αναζήτησης και Θεματικοί Κατάλογοι στον Ελληνικό κυβερνοχώρο

Ο Ελληνικός Κυβερνοχώρος, μετά την εμφάνιση των Πυλών, άρχισε να αναπτύσσεται και να εξελίσσεται με πολύ γρήγορο ρυθμό. Πρόσφατες στατιστικές αναφέρουν ότι στο Διαδίκτυο υπάρχουν πάνω από 2 εκατομμύρια ιστοσελίδες, με καθαρά ελληνικό περιεχόμενο. Σε πάρα πολλές από αυτές ο χρήστης μπορεί να βρει χρήσιμες πληροφορίες στην Ελληνική γλώσσα. Βέβαια, αρωγοί σε αυτήν την προσπάθεια είναι οι Ελληνικές Μ.Α., που στοχεύουν στην αρχειοθέτηση και στη ταξινόμηση των δικτυακών τόπων που αναφέρονται σε πληροφορίες Ελληνικού περιεχομένου. Να σημειωθεί εδώ ότι δεν υπάρχουν σαφείς διαχωριστικές γραμμές μεταξύ των θεματικών καταλόγων και των αυτόματων μηχανών στον Ελληνικό κυβερνοχώρο.

Πιο γνωστές Μ.Α. είναι οι GoGreece [GG], Greek Indexer [GI], Phantis [PHANTIS], Pathfinder [PF], Robby [ROBBY], Greek Web Index [GWI], Anazitisis [ANAZITISIS], Eseek [ESEEK], Thea [THEA] και in.gr [IN].

7 Μειονεκτήματα των Μηχανών Αναζήτησης

Οι Μ.Α. συχνά κρίνονται ανεπαρκείς παρουσιάζοντας μερικά σημαντικά μειονεκτήματα. Αυτά απορρέουν από τη δυναμική φύση τόσο της πληροφορίας όσο και του Διαδικτύου ως μέσου διάδοσης της πληροφορίας [Chu,96].

Συγκεκριμένα, είναι φανερό ότι οι πληροφορίες που είναι διαθέσιμες σήμερα στον Παγκόσμιο Ιστό μπορούν ανά πάσα στιγμή να τροποποιηθούν, να ανανεωθούν ή να διαγραφούν χωρίς καμία ενημέρωση. Στην περίπτωση αυτή υπάρχει ένα κενό διάστημα ως προς την ενημέρωση των Μ.Α. που περιορίζει την ακρίβειά τους.

Από την άλλη πλευρά, ο Παγκόσμιος Ιστός έχει μια άναρχη δομή που διαρκώς εξελίσσεται, ενώ οποιοσδήποτε μπορεί να το χρησιμοποιεί και να προσθέτει πληροφορίες. Συνεπώς δεν υπάρχει καμία συνολική εικόνα της ποσότητας και του είδους πληροφορίας που διαχέεται μέσα από αυτό. Αν επιπλέον ληφθεί υπόψη η ποικιλία της πληροφορίας μέσω των διαφορετικών μορφοποιήσεων είναι αντιληπτό ότι η αυτοματοποιημένη αναζήτηση και ανάκτηση της πληροφορίας καθίσταται αρκετές φορές δυσχερής [Tomaiuolo,96].

Υπάρχουν δε πολλές περιπτώσεις όπου τα αποτελέσματα αποκλίνουν αρκετά από τα επιθυμητά και ο χρήστης αναγκάζεται να καταβάλει μεγαλύτερη προσπάθεια για τον εντοπισμό της πληροφορίας που αναζητά. Αυτό σημαίνει ότι τα προγράμματα αναζήτησης ενεργούν παθητικά και έχουν ανάγκη από εξοικειωμένους και ευφυείς χρήστες για να προσφέρουν σωστά αποτελέσματα [Bressan,97].

Αναμφισβήτητα, οι Μ.Α. αποτελούν ένα εξαιρετικά σημαντικό και χρήσιμο μέσο βοήθειας στην ανάκτηση πληροφοριών ανάμεσα στην πληθώρα που διαθέτει ο Παγκόσμιος Ιστός. Για να είναι όμως και αποτελεσματικές, απαιτείται η συχνή και οργανωμένη καταγραφή καθώς και η χαρτογράφηση της διαθέσιμης πληροφορίας. Κάτι τέτοιο είναι απαραίτητο εξαιτίας των συνεχόμενων αλλαγών που αυτή υφίσταται. Εξαιτίας του όγκου αυτής της πληροφορίας, είναι αναμενόμενη και η μεγάλη δυσκολία στη συνεχή και πλήρη καταγραφή της.

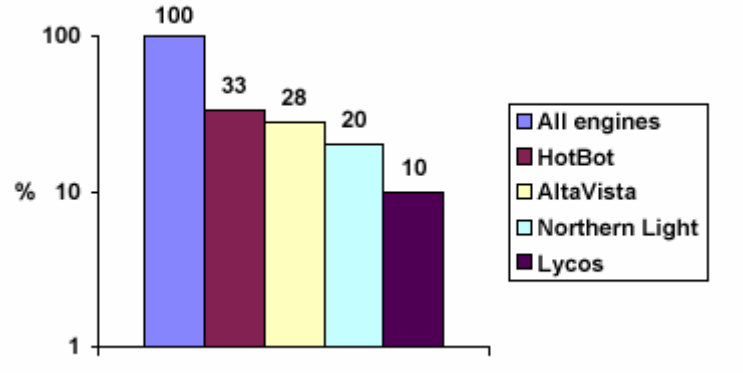
Ενδεικτικά μια πειραματική διαδικασία που πραγματοποιήθηκε τα τέλη του 1998, κατέληξε στο συμπέρασμα ότι καμία Μ.Α. (εκτός της Google που δεν συμπεριλήφθηκε στην μελέτη), δεν έχει καταφέρει να δεικτοδοτήσει παραπάνω από το 33% των συνολικών ιστοσελίδων του Παγκόσμιου Ιστού [Lawrence,98a], όπως χαρακτηριστικά απεικονίζεται στο Σχήμα 4 .

Σύμφωνα με την ίδια μελέτη, κάποιες από τις παραπάνω δεικτοδοτημένες ιστοσελίδες ενδέχεται να μην αντιστοιχούν στη διεύθυνση του Ομοιόμορφου Εντοπιστή Πόρου που η Μ.Α. είχε αρχικά εντοπίσει. Πρόκειται για ένα ποσοστό "άκυρων" ιστοσελίδων που η ύπαρξη τους δεν έχει πιστοποιηθεί εκ νέου και λανθασμένα είναι καταχωρημένη στα ευρετήρια της εκάστοτε Μ.Α., όπως φαίνεται στο Σχήμα 5.

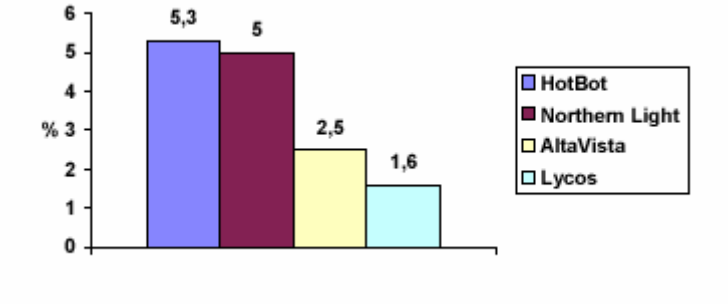
Μια άλλη μελέτη η οποία συμπεριλάμβανε και την Μ.Α. Google, παρουσίασε άκρως ενδιαφέροντα αποτελέσματα [Lawrence,99a]. Το Σχήμα 6, παρουσιάζει ξανά το ποσοστό του Παγκόσμιου Ιστού που έχει καλύψει καθεμία από τις εμπλεκόμενες Μ.Α., ενώ στο Σχήμα 7 απεικονίζεται το ποσοστό των ιστοσελίδων που προσδιόρισε κάθε μηχανή για μία συγκεκριμένη ερώτηση σε σχέση με το σύνολο των ιστοσελίδων από όλες τις Μ.Α. για την ίδια ερώτηση.

Είναι αξιοσημείωτο αλλά και εντυπωσιακό ότι σε μελέτες που παρουσιάζονται με διαφορά ενός μόνο έτους το ποσοστό κάλυψης του παγκόσμιου ιστού ανά εμπλεκόμενη Μ.Α. μειώνεται δραματικά [Lawrence,98a] [Lawrence,99a]. Έτσι συγκρίνοντας τα Σχήματα 4 και 6, συμπεραίνεται ότι η υπηρεσία αναζήτησης AltaVista καλύπτει το 15% του συνολικού όγκου του πληροφοριακού ιστού σε σχέση με το 28 % που πιστοποιήθηκε ότι κάλυπτε ένα χρόνο πριν. Αντίστοιχα, το ποσοστό της Northern Light μέσα σε ένα χρόνο περιορίζεται στο μισό, ενώ η HotBot από την πρώτη θέση που κατείχε στην πρώτη μελέτη με 33% εκπίπτει ένα χρόνο μετά στο 11%, παρουσιάζοντας μείωση κατά τρεις φορές όσον αφορά την κάλυψη της υπάρχουσας

συνολικής πληροφορίας. Τέλος, κατά τις ίδιες μελέτες η M.A. Lycos αδυνατεί να συντάξει τις νέες ιστοσελίδες σε πολύ μεγαλύτερο βαθμό αφού το ποσοστό κάλυψης του Παγκόσμιου Ιστού μειώνεται κατά πέντε φορές από 10% μόλις στο 2%.



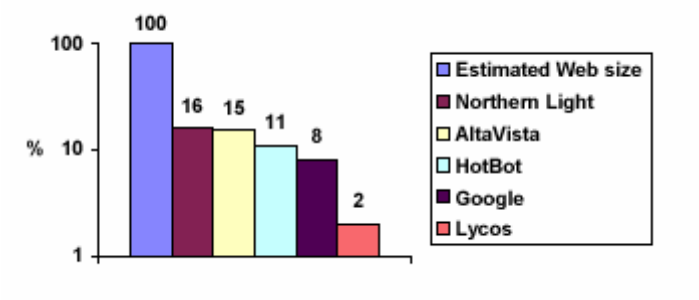
Σχήμα 4. Ποσοστό κάλυψης του Παγκόσμιου Ιστού ανά M.A. [Lawrence,98a]



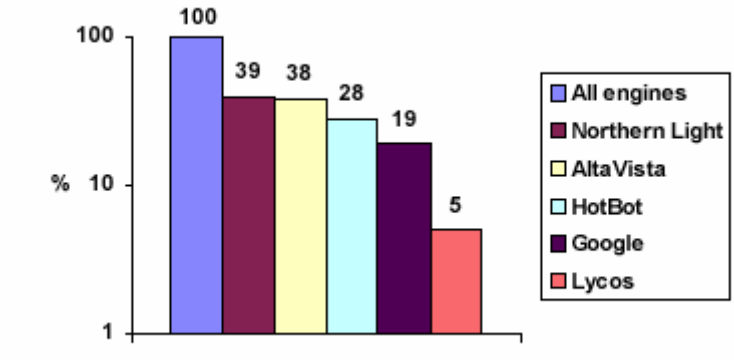
Σχήμα 5. Ποσοστό άκυρων ιστοσελίδων [Lawrence,98a]

Εκτός όμως της αδυναμίας κάλυψης του συνολικού όγκου της πληροφορίας που βρίσκεται στο Διαδίκτυο και της δυσκολίας ανανέωσης, παρακολούθησης και εντοπισμού των νέων πηγών που συνεχώς καταχωρούνται σε αυτό, προστίθεται και η διαφορετική απόκριση των M.A. σε σχέση με τις υποβαλλόμενες ερωτήσεις των χρηστών. Αυτό φυσικά οφείλεται στα διαφορετικά εξωτερικά ή εσωτερικά χαρακτηριστικά τους και στις διαφορετικές λειτουργίες που τις διέπουν. Έτσι λοιπόν όπως φαίνεται στο Σχήμα 7, παρά το γεγονός ότι εάν ένας χρήστης ανατρέξει στην υπηρεσία αναζήτησης Northern Light, θα λάβει τα περισσότερα αποτελέσματα, εντούτοις θα χάσει περισσότερη από τη μισή πληροφορία που υπάρχει στο Διαδίκτυο. Αυτό συμβαίνει διότι η Northern Light του παρέχει μόλις το 39% του συνολικού ποσού των συντεταγμένων ιστοσελίδων που επιστρέφουν οι υπόλοιπες M.A. για την ίδια πληροφοριακή ανάγκη, έχοντας εξαιρέσει τα διπλότυπα πεδία.

Από τα παραπάνω γίνεται φανερό ότι ο συνδυασμός όλων των M.A. δίνει το βέλτιστο δυνατό αποτέλεσμα. Λαμβάνοντας υπόψη την ένωση των επιστρεφόμενων αποτελεσμάτων από όλες τις M.A., επιτυγχάνεται βελτίωση όσον αφορά το βαθμό κάλυψης του Παγκόσμιου Ιστού, κατά 3,5 φορές περίπου από αυτή που θα λαμβάνονταν χρησιμοποιώντας μια μόνο M.A. για μία δεδομένη ερώτηση [Lawrence,98a]. Όμως παρ' όλα αυτά, η κάλυψη του Παγκόσμιου Ιστού από τις M.A παρουσιάζεται εξαιρετικά μικρή, ενώ όσο αυτός εξαπλώνεται, τόσο μειώνεται το ποσοστό της κάλυψής του, όπως αυτό φαίνεται από τα Σχήματα 4 και 6. Από τα παραπάνω φαίνεται ότι η διαδικασία της ανάκτησης χρήσιμων και νέων πληροφοριών γίνεται ολοένα πιο δύσκολη και περισσότερο χρονοβόρα [Selberg, 95].



Σχήμα 6. Ποσοστό κάλυψης του Παγκόσμιου Ιστού ανά Μ.Α. [Lawrence,99a]



Σχήμα 7. Ποσοστό ιστοσελίδων ανά Μ.Α. για μια συγκεκριμένη ερώτηση σε σχέση με το σύνολο των συνταγμένων ιστοσελίδων [Lawrence,99a]

8 Συμπεράσματα

Στο κεφάλαιο αυτό γίνεται μια εκτενής αναφορά στη λειτουργία, τις ιδιότητες και τα χαρακτηριστικά των ειδικών λογισμικών και υπηρεσιών αναζήτησης που χρησιμοποιεί ένας χρήστης προκειμένου να ικανοποιήσει τις πληροφοριακές του ανάγκες. Οι υπηρεσίες αυτές ή αλλιώς Μηχανές Αναζήτησης έχουν ως πρωταρχικό σκοπό να συλλέξουν, να επεξεργαστούν, να ταξινομήσουν και να παρουσιάσουν όσον το δυνατόν πιο σχετικά αποτελέσματα δοθείσης μιας υποβαλλόμενης ερώτησης.

Όμως, λόγω της εξαιρετικά ταχείας ανάπτυξης του Διαδικτύου και κατ' επέκταση και του Παγκόσμιου Ιστού, οι υπάρχουσες μηχανές αναζήτησης δεν μπορούν να εντοπίσουν συγχρόνως και με την ίδια προτεραιότητα όλες τις νέες σελίδες έτσι ώστε να ανανεώσουν τους καταλόγους τους [Lawrence,98a], [Lawrence,99a]. Εκτός αυτού, οι μηχανές αναζήτησης δημιουργούν και συντάσσουν τους καταλόγους και τα ευρετήριά τους με διαφορετικούς αλγορίθμους που έχουν ως αποτέλεσμα διαφορετικό χρόνο απόκρισης στην ενημέρωση των βάσεων δεδομένων τους. Κατά συνέπεια, ο χρήστης που χρησιμοποιεί μια συγκεκριμένη μηχανή αναζήτησης, χάνει σημαντικό ποσό ωφέλιμης πληροφορίας, το οποίο επιστρέφεται από άλλες υπηρεσίες αναζήτησης. Για να αποφευχθεί αυτό, οι χρήστες συνήθως υποβάλουν εκ νέου τις ερωτήσεις τους σε παραπάνω από μια μηχανές αναζήτησης, αυξάνοντας κατ' αυτό τον τρόπο, μια ήδη χρονοβόρα διαδικασία.

Επιπρόσθετα, όλες οι κορυφαίες μηχανές αναζήτησης (Google, MSN, Yahoo) είναι βασισμένες στη λογική διατύπωσης ερωτημάτων όπου ο χρήστης διατυπώνει ένα ερώτημα που αποτελείται από ένα σύνολο λέξεων συχνά συνοδευόμενες με έναν λογικό τελεστή, και το σύστημα επιστρέφει έναν ταξινομημένο κατάλογο εγγράφων. Η αποτελεσματικότητα των μηχανών αναζήτησης Ιστού κρίνεται ικανοποιητική όταν η ανάγκη πληροφόρησης του χρήστη είναι συγκεκριμένη και ακριβώς ορισμένη. Όπως υποστηρίζεται από τον [Marchionini,92], "οι χρήστες θέλουν να επιτύχουν τους στόχους τους με ένα ελάχιστο γνωστικό φορτίο και ένα μέγιστο βαθμό απόλαυσης". Οι χρήστες γενικά διατυπώνουν πολύ σύντομες ερωτήσεις (ένας έως τρεις όροι [CCW95], [Pin94]), μερικές φορές πολύ διφορούμενες, και χωρίς πολλή σκέψη στη διατύπωση ερώτησης. Η ασάφεια της ερώτησης μπορεί εμφανιστεί με διαφορετικούς τρόπους. Ακόμα, είναι συχνά η περίπτωση ότι οι χρήστες οι ίδιοι είναι ασαφείς σχετικά με την ανάγκη πληροφοριών τους.

Με αφορμή τους ανωτέρω περιορισμούς των τεχνικών αναζήτησης που συνοψίζονται κυρίως στον ελλιπή τρόπο διατύπωσης των ερωτημάτων, αλλά και στην αδυναμία παρακολούθησης του βαθμού ανάπτυξης και εντοπισμού νέων πηγών πληροφορίας, η παρούσα διατριβή προτείνει μια νέα μέθοδο αναζήτησης. Περιληπτικά, η διατριβή προτείνει, μια μεθοδολογία αναζήτησης γύρω από μια αρχική πηγή σχετικής πληροφορίας, προσομοιώνοντάς την με την διεργασία αναζήτησης τροφής των μυρμηγκιών γύρω από την αποικία. Η βάση της τεχνικής αναζήτησης, είναι ο αλγόριθμος αποικίας μυρμηγκιών, οποίος αν και σχετικά πρόσφατος έχει ένα ευρύ πεδίο εφαρμογών.

9 Βιβλιογραφικές Αναφορές

[Adar,99] E. Adar, D. Karger, and L. Stein. Haystack: Per-user information environments. In Proceedings of the 1999 Conference on Information and Knowledge Management, CIKM, 1999.

[Agichtein,00] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plaintext collections. In Proceedings of the 5th ACM International Conference on Digital Libraries, 2000.

[Anagnostopoulos,02] "Implementing a customised meta-search interface for user query personalisation", Anagnostopoulos I., Psoroulas I., Loumos V. and Kayafas E., IEEE 24th International Conference on Information Technology Interfaces, ITI 2002 June 24-27, 2002, Cavtat/Dubrovnik, CROATIA.

[Anagnostopoulos,04] "Νέες Μέθοδοι Για Ευφυή Ανάκτηση ,Κατηγοριοποίηση Και Ταξινόμηση Δεδομένων Σε Πληροφοριακά Συστήματα", Διδακτορική διατριβή 2004, ΕΜΠ

[Bressan,97] Bressan S. and Lee T., "Information Brokering on the World Wide Web", WebNet World Conference, 1997.

[Brin,98] S.Brin and L.Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine" Proc. 7 th International World Wide Web Conference, 1998.

[Budzik,00] J. Budzik and K.J. Hammond. User interactions with everyday applications as context for just-in-time information access. In Proceedings of the 2000 International Conference on Intelligent User Interfaces, New Orleans, Louisiana, 2000. ACM Press.

[Buyukkokten,99] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of web pages. In Proceedings of the ACM SIGMOD Workshop on the Web and Databases, WebDB, 1999.

[Chakrabarti,99] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: A new approach to topic-specific web resource discovery. In 8th World Wide Web Conference, Toronto, May 1999.

[Cho,98] Junghoo Cho, Hector Garcia-Molina, Lawrence Page. Efficient Crawling Through URL Ordering. 7th International Web Conference (WWW 98).

[Chu,96] H. Chu and M. Rosenthal, "Search engines for the WWW: A comparative study and evaluation methodology", ASIS 1996 Annual Conference Proceedings, Baltimore, MD, October 19-24, 1996, pp.127-135.

[Gravano,99] L. Gravano, H. Garcia-Molina, and A. Tomasic. GlOSS: Text-source discovery over the Internet. ACM Transactions on Database Systems, 24(2), 1999.

[Croft,95] W. Bruce Croft, Robert Cook, and Dean Wilder. Providing government information on the Internet: Experiences with THOMAS. In Proceedings of DL-95, the 2nd Annual Conference on the Theory and Practice of Digital Libraries, pages 19-24, Austin, Texas, U.S.A., June 1995.

[Lawrence,98a] Steve Lawrence and C. Lee Giles. Searching the World Wide Web. (Vol. 280, pp. 98-100) Science (1998).

[Lawrence,98b] Steve Lawrence and C. Lee Giles. Context and page analysis for improved web search. IEEE Internet Computing, 2(4):38–46, 1998.

[Lawrence,99a] Steve Lawrence and C. Lee Giles. Accecibility of information on the web (Vol. 400, pp. 107-109) Nature (999).

[Lawrence,99b] Steve Lawrence and C. Lee Giles. Searching the web: General and scientific information access. IEEE Communications, 37(1):116–122, 1999.

[Marchionini,92] G. Marchionini. Interfaces for end-user information seeking. Journal of the American Society for Information Science, 43(2):156-163, 1992.

[Pinkerton,94] B Pinkerton. Finding what people want: Experiences with the WebCrawler.In Proceedings of WWW-94, the 2nd International World Wide Web Conference,October 1994.

[Salton,89] Gerald Salton. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison Wesley, 1989.

[Selberg,95] E. Selberg and O. Etzioni. Multi-Engine Search and Comparison using the MetaCrawler. In Proc. of the Fourth Int'l WWW Conference, pages 195-208, Boston, Massachusetts, USA, 1995.

[SER,99] “HotBot Integrates Popularity Into Top Results”, The Search Engine Report, March3, 1999.

[Tomaiuolo,96] N. Tomaiuolo and J. Packer, "An analysis of Internet search engines: assessment of over 200 search queries", Computers in Libraries, vol. 16, no. 6, pp.58 (5).

10 Αναφορές στο Διαδίκτυο

[AJ] Ask Jeeves, http://www.ask.com

[ANAZITISIS] Anazitisis, http://www.anazitisis.gr/

[AOL] America OnLine, http://www.aol.com

[ATW] AllTheWeb, http://www.alltheweb.com

[AV] AltaVista, http://www.altavista.com

[CERN] CERN, History and growth of the Web http://public.web.cern.ch/Public/ACHIEVEMENTS/WEB/history.html

[COP] Copernic, http://www.copernic.com/en/index.html

[DH] DirectHit, http://www.directhit.com

[DMOZ] DMOZ, Open Directory Project, http://www.dmoz.com

[ESEEK] Eseek, http://www.eseek.gr/

[EXCITE] Excite, http://www.excite.com

[GG] GoGreece, http://www.gogreece.com/

[GI] Greek Indexer, http://www.gr-indexer.gr/

[GOOGLE] Google, http://www.google.com

[GWI] Greek Web Index, http://www.webindex.gr/

[HB] HotBot, http://www.hotbot.com

[IN] in.gr, http://www.in.gr/

[INKTOMI] Inktomi, http://www.inktomi.com

[ISC] Internet Software Consortium, http://www.isc.org

[IXQ] Ixquick, http://www.ixquick.com

[LYCOS] Lycos, http://www.lycos.com

[MSN] MSN Search, http://www.msn.com

[NAVNET] Interner Search Engine Guide, http://www.walthowe.com/navnet/faq/search.html

[NL] NorthernLight, http://www.northernlight.com

[PF] Pathfinder, http://www.pathfinder.gr/

[PHANTIS] Phantis, http://www.phantis.com/

[ROBBY] Robby, http://www.robby.gr/

[searchengineshowdown] http://www.searchengineshowdown.com/features/

[searchenginewatch] http://searchenginewatch.com/webmasters/article.php/2167891

[TEOMA] Teoma, http://www.teoma.com

[THEA] Thea, http://www.thea.gr/

[WRP1] http://www.robotstxt.org/wc/exclusion.html#robotstxt, The Web Robots Pages: The Robots Exclusion Protocol.

[WRP2] http://www.robotstxt.org/wc/exclusion.html#meta, The Web Robots Pages: The Robots META tag.

[YAHOO] Yahoo!, http://www.yahoo.com

ΠΑΡΑΡΤΗΜΑ 1ου ΚΕΦΑΛΑΙΟΥ

Α. Εξωτερικά Χαρακτηριστικά Μ.Α.

Χαρακτηριστικά αυτόματης αναζήτησης ιστοσελίδων			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	**Σχόλια**
Βαθιά αναζήτηση	Όλες εκτός …	AltaVista, Teoma, Excite	
Υποστήριξη πλαισίων	Όλες εκτός …	Excite, AllTheWeb	
Χαρτογράφηση εικόνων	Όλες εκτός …	Excite, AllTheWeb, Google, Inktomi	
Αποτροπή αυτόματης ανίχνευσης ιστοχώρου	Όλες		
Αποτροπή αυτόματης ανίχνευσης ιστοσελίδας	Όλες		
Αναφορά από άλλες υπερσυνδέσεις	Όλες		
Ανίχνευση ανανέωσης περιεχομένου	Όλες εκτός …	Excite, AllTheWeb, Google, Northern Light	
Ειδική προβολή με πληρωμή	Όλες εκτός …	Excite, Google	
Έλεγχος διεύθυνσης Ομοιόμορφων Εντοπιστών Πόρων	Όλες		

Χαρακτηριστικά σύνταξης ιστοσελίδων			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	**Σχόλια**
Σύνταξη "ορατού" κειμένου	Όλες		Μερικές κοινές λέξεις ενδέχεται να μην συντάσσονται
Αποβολή κοινών λέξεων	Όλες εκτός …	AllTheWeb	δεν έχει επιβεβαιωθεί για την Teoma
Υποστήριξη πεδίων μετα-ετικετών	Όλες		Οι AltaVista, AllTheWeb και Teoma στηρίζονται κυρίως σε αυτό το χαρακτηριστικό
Δημιουργία παραγώγων	Όλες εκτός …	Northern Light	

Χαρακτηριστικά κατάταξης των αποτελεσμάτων			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	Σχόλια
Στάθμιση των πεδίων μετα-ετικετών	Inktomi	Όλες εκτός Inktomi	
Στάθμιση σε αναφορές από άλλες υπερσυνδέσεις	Όλες		Σημαντικό για Google
Στάθμιση ανάλογα με την επιλογή των αποτελεσμάτων	Hotbot	Όλες εκτός Hotbot	

Χαρακτηριστικά αναγνώρισης και αντιμετώπισης τεχνικών Spam			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	**Σχόλια**
Αντιμετώπιση "αόρατου" κειμένου	Όλες εκτός ...	Excite, AllTheWeb	
Αντιμετώπιση κειμένου ελάχιστου μεγέθους	Όλες εκτός ...	Excite, AllTheWeb, Northern Light	

Β. Εσωτερικά Χαρακτηριστικά Μ.Α.

Μαθηματικές Εντολές αναζήτησης – εντολές Boolean τύπου			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	**Σχόλια**
AND	Όλες		
OR	Όλες		
NOT	Όλες		
ADJ	Όλες		
NEAR	Όλες εκτός...	AllTheWeb, Direct Hit, Google, Inktomi	
FAR	Όλες εκτός...	AllTheWeb, Direct Hit, Google, Inktomi	
Φώλιασμα – Σύνθεση	Όλες εκτός …	AllTheWeb, Inktomi	

Ενισχυμένες εντολές αναζήτησης			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	**Σχόλια**
Αναζήτηση βάσει διεύθυνσης Ιστοχώρου	Όλες εκτός …	Direct Hit, Hotbot, Lycos, Northern Light	
Αναζήτηση βάσει διεύθυνσης Ομοιόμορφου Εντοπιστή Πόρου	Όλες εκτός …	Direct Hit, Hotbot, Lycos	
Αναζήτηση βάσει υπερσυνδέσμου	Όλες εκτός ...	Excite, Direct Hit, Hotbot	
Χρήση χαρακτήρων Μπαλαντέρ	Όλες εκτός …	Hotbot, Lycos	

Χαρακτηριστικά αναζήτησης			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	**Σχόλια**
Σχετικές αναζητήσεις	Όλες		
Συγκέντρωση αποτελεσμάτων	Όλες εκτός ...	Northern Light	
Εύρεση σχετικών ιστοσελίδων	Όλες		
Δημιουργία παραγώγων	Όλες εκτός ...	Northern Light	
Εσωτερική Αναζήτηση	AltaVista, Google, Hotbot, Lycos		
Αναζήτηση κρυφών – αποθηκευμένων ιστοσελίδων	Google		
Αναζήτηση ιστοσελίδων βάσει γλωσσικού περιεχομένου	AltaVista, AllTheWeb, Excite, Google, Hotbot, Lycos		
Μετάφραση ιστοσελίδας	AltaVista, Google, Lycos		
Έλεγχος και φιλτράρισμα “επικίνδυνου” περιεχομένου	AltaVista, AllTheWeb, Lycos		

Χαρακτηριστικά προσαρμογής απεικόνισης και προβολής			
Όνομα	**Υποστηρίζουν**	**Δεν Υποστηρίζουν**	**Σχόλια**
Ταξινόμηση αποτελεσμάτων ανά ημερομηνία	Northern Light		
Ταξινόμηση αποτελεσμάτων σε καθορισμένο εύρος ημερομηνίας	AltaVista, Google, HotBot		
Προβολή ημερομηνίας δημιουργίας ή μορφοποίησης της ιστοσελίδας	AltaVista, HotBot		
Παροχή περιβάλλοντος ενισχυμένης αναζήτησης	Όλες		
Παροχή βοήθειας	Όλες		

ΚΕΦΑΛΑΙΟ

2

ΑΡΧΕΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΚΑΙ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΠΕΡΙΕΧΟΜΕΝΑ 2ου ΚΕΦΑΛΑΙΟΥ

ΑΡΧΕΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΚΑΙ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

1 Εισαγωγή

Στο κεφάλαιο αυτό αναλύεται η έννοια της επεξεργασίας της πληροφορίας, ενώ παράλληλα παρουσιάζονται οι υπάρχουσες παραλλαγές των τεχνικών της επεξεργασίας αυτής. Στη συνέχεια αναλύονται οι μέθοδοι αναπαράστασης κειμένου, οι τεχνικές επιλογής των κατάλληλων ιδιο-χαρακτηριστικών παράλληλα με τις τεχνικές μείωσης της διαστασιολόγησης, με στόχο την αναπαράσταση των εγγράφων ως διανύσματα. Επιπλέον παρουσιάζεται μια αναλυτική περιγραφή των μοντέλων ανάκτησης πληροφορίας και των μεθόδων που χρησιμοποιούν για την αναπαράσταση και παρουσίαση της ανακτημένης πληροφορίας. Τέλος γίνεται μια εκτενής αναφορά στις μεθόδους συσταδοποίησης εγγράφων που χρησιμοποιούνται ευρέως για κατηγοριοποίηση εγγράφων σε δυναμικά περιβάλλοντα.

2 Επεξεργασία της πληροφορίας

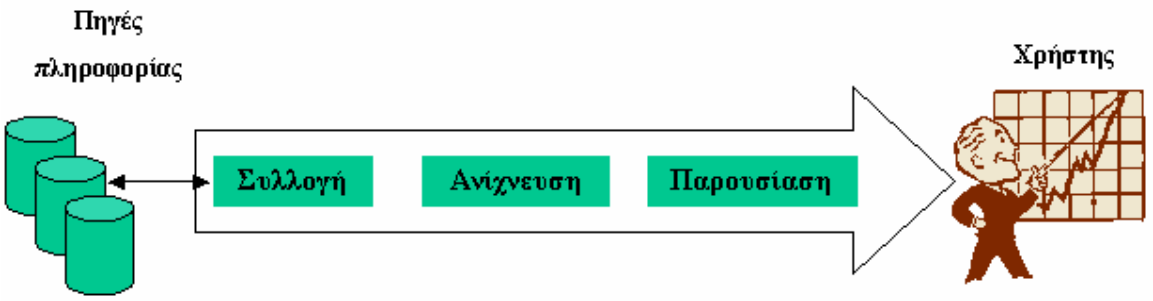
2.1 Ορισμός της επεξεργασίας της πληροφορίας

Ο στόχος των τεχνικών επεξεργασίας της πληροφορίας, είναι να μειωθεί το ποσό της μη σχετικής πληροφορίας που επιστρέφεται στους χρήστες όσον αφορά τους τομείς ενδιαφέροντός τους [Μακρής,02]. Η επεξεργασία της πληροφορίας είναι μια διαδικασία κατά την οποία το μη-σχετικό ποσό πληροφορίας μέσα από ένα σύνολο δεδομένων απορρίπτεται, κατά τέτοιον τρόπο, ώστε να παρουσιάζονται στον χρήστη μόνο οι σχετικές με τις αναζητήσεις του πληροφορίες. Παρακάτω δίνεται ο ορισμός της επεξεργασίας της πληροφορίας όπως αποδίδεται στις αναφορές [Mostafa,97] και [Hull,98], σε ένα γενικότερο πλαίσιο, όπου το προς επεξεργασία μέγεθος ονομάζεται μονάδα πληροφορίας D. Το μέγεθος αυτό αντιπροσωπεύει οντότητες οι οποίες περιέχουν πληροφορίες κειμένου. Στην παρούσα διατριβή, οι μονάδες πληροφορίας αντιστοιχούν σε ιστοσελίδες που διαχέουν και δημοσιεύουν το περιεχόμενό τους στο Διαδίκτυο.

Ας υποτεθεί ένα σύνολο μονάδων πληροφοριών D. Η διαδικασία της Επεξεργασίας Πληροφορίας (ΕΠ) όσον αφορά μια συγκεκριμένη ερώτηση του χρήστη ψ, καθορίζεται ως την χαρτογράφηση $f_\psi : D \mapsto \{0,1\}$, πάνω στο σύνολο των μονάδων πληροφοριών είτε στο μηδέν είτε στο ένα, που αντιστοιχεί στην απόρριψη ή την αποδοχή μιας μονάδας πληροφορίας, αντίστοιχα.

Εναλλακτικά, η διαδικασία αυτή μπορεί να οριστεί ως μια χαρτογράφηση πάνω στο σύνολο των μονάδων πληροφορίας μέσα στο διάστημα $f_\psi^* : D \mapsto \{0,1\}$, όπου το f_ψ^* προσδιορίζει τη σχετικότητα μιας σχετικής μονάδας *d* όσον αφορά την ερώτηση του χρήστη. Πρέπει ωστόσο να σημειωθεί, ότι αυτός ο ορισμός απαιτεί μια περαιτέρω κατωφλίωση των αποτελεσμάτων σχετικότητας, στην περίπτωση κατά την οποία τα έγγραφα πρόκειται να απορριφθούν ρητά ή να γίνουν αποδεκτά.

Η ΕΠ είναι μια διαδικασία που διευθύνεται ενεργά από τους ανθρώπους, με ή χωρίς τη βοήθεια μηχανών, προκειμένου να αντιμετωπιστεί η υπερφόρτωση πληροφοριών [Oard,97]. Παράλληλα, ο στόχος ενός συστήματος ΕΠ είναι η αυτοματοποίηση της διαδικασίας. Η ολοκλήρωση της διαδικασίας αυτής, οδηγεί σε τρεις δευτερεύουσες διαδικασίες, όπως η συλλογή της πληροφορίας, η εύρεση της σχετικής και η απόρριψη της μη-σχετικής πληροφορίας και η παρουσίαση των αποτελεσμάτων στο χρήστη [Oard,97]. Οι λειτουργίες αυτές απεικονίζονται στο Σχήμα 1.



Σχήμα 1. Λειτουργίες κατά την ανίχνευση πληροφορίας

Παρόλα αυτά, αμφισβητείται εάν η συλλογή πληροφοριών πρέπει ή όχι να αποτελεί μέρος της επεξεργασίας. Αυτό εξαρτάται κυρίως από το εάν στα συστήματα αυτά ανατίθεται μια ενεργητική ή παθητική πρωτοβουλία όσον αφορά τη λειτουργία τους. Κατά, τις πρώτες περιγραφές των συστημάτων αυτών, ένα παράδειγμα παθητικής συλλογής πληροφοριών ή εγγράφων, αποτελούσε το εισερχόμενο ηλεκτρονικό ταχυδρομείο [Denning,82]. Εντούτοις, λόγω

ενός αυξανόμενου όγκου ηλεκτρονικά προσιτής πληροφορίας, τα συστήματα που συλλέγουν ενεργά πληροφορία κερδίζουν σε δημοτικότητα. Έτσι παραδείγματος χάριν, ένα σύνολο πρακτόρων θα μπορούσε να χρησιμοποιηθεί για να ανιχνεύσει πολλαπλές πηγές πληροφοριών από το Διαδίκτυο, ερευνώντας για σχετικές πληροφορίες [Balabanovic,97], [Pazzani,97] [Anagnostopoulos,04].

2.2 Ορισμός του Συστήματος Επεξεργασίας της Πληροφορίας

Ένα Σύστημα Επεξεργασίας της Πληροφορίας (ΣΕΠ) αυτοματοποιεί τη διαδικασία επιλογής και φιλτραρίσματος της πληροφορίας με στόχο να μειώσει την υπερφόρτωση πληροφοριών. Το σύστημα είτε παράγει είτε δέχεται ποσά πληροφορίας, υπολογίζει τη σχετικότητα αυτών, όσον αφορά σε σχετικές ερωτήσεις και παρέχει είτε το ανάλογο σχετικό ποσό είτε ένα αποτέλεσμα σχετικότητας για την τελική παρουσίαση στο χρήστη.

Υπάρχουν διάφορα βασικά σημεία τα οποία πρέπει να συγκεκριμενοποιηθούν και στα οποία τα συστήματα πληροφοριών μπορούν να διαφέρουν. Παρακάτω αναλύονται τέσσερις προδιαγραφές: η είσοδος του συστήματος, οι απαιτήσεις για την έξοδο, η κατασκευή του προφίλ του χρήστη και η έννοια της σχετικότητας. Αξίζει να σημειωθεί ότι ένα ΣΕΠ ενδέχεται να είναι εγκατεστημένο στον υπολογιστή του χρήστη, στους ειδικούς ενδιάμεσους κεντρικούς υπολογιστές μεταξύ του προμηθευτή πληροφοριών και των χρηστών ή άμεσα στις πηγές πληροφορίας.

Είσοδος

Ας υποτεθεί ότι δίνεται ένα σύνολο ή ρεύμα πληροφοριών ως εισαγωγή ή είσοδος. Γενικά, η είσοδος μπορεί να είναι οποιοδήποτε είδος πληροφορίας που μπορεί να αντιπροσωπευθεί ψηφιακά. Από την μία πλευρά, η είσοδος θα μπορούσε να είναι υπό μορφή κειμένων ή εγγράφων, όπως το ηλεκτρονικό ταχυδρομείο, άρθρα ομάδων πληροφόρησης, ή άλλες πηγές πληροφορίας (στατικές ή δυναμικές ιστοσελίδες) από το Διαδίκτυο. Από την άλλη, μέσα όπως οι εικόνες, ο ήχος, το βίντεο και άλλα πολυμεσικά έγγραφα μπορούν να θεωρηθούν ως πιθανές είσοδοι. Η παρούσα εργασία σκοπεύει στον περιορισμό του ρεύματος εισόδου σε δυναμικές και στατικές *ιστοσελίδες*. Ως εκ τούτου, από αυτό το σημείο της διατριβής και στο εξής, ο όρος είσοδος ή έγγραφο θα αντιστοιχεί σε ανακτημένες ιστοσελίδες.

Έξοδος

Μόλις ένα φίλτρο υπολογίσει την σχετικότητα των μονάδων πληροφορίας από το εισερχόμενο ρεύμα, τα αποτελέσματα πρέπει να παρουσιαστούν στο χρήστη. Κατά την παρουσίαση των πληροφοριών, τα έγγραφα ταξινομούνται ασύγχρονα, σύμφωνα με μια προσωπική έκθεση των ενδιαφερόντων των χρηστών σε τακτά χρονικά διαστήματα. Αυτή η επεξεργασία κατά δεσμίδες επιτρέπει την ταξινόμηση των μονάδων D σύμφωνα με έναν βαθμό σχετικότητας.

Δημιουργία προφίλ χρήστη

Θεμελιώδες ζήτημα για οποιαδήποτε εργασία ή διαδικασία ανίχνευσης πληροφορίας, συνιστά η γνώση του ενδιαφέροντος του κάθε χρήστη. Στην επεξεργασία πληροφοριών, είναι ωφέλιμη η γνώση των χαρακτηριστικών που συνιστούν στη δημιουργία του προφίλ των ενδιαφερόντων του χρήστη. Στη βιβλιογραφία βάσεων δεδομένων και ανάκτησης πληροφοριών, το παραπάνω είναι γνωστό και ως ερώτηση (query). Εντούτοις, λαμβάνοντας υπ' όψιν την πολλαπλότητα των ενδιαφερόντων του χρήστη, το προφίλ του μπορεί να συνίσταται από ένα σύνολο ερωτήσεων.

Για την κατασκευή και δημιουργία ενός προφίλ χρήστη, παρεμβάλλεται ο ίδιος ο χρήστης ενώ εμπλέκονται παράλληλα και διάφοροι κανόνες τεχνητής αναπαράστασης γνώσης, οι οποίες θα

περιγραφούν παρακάτω. Στην πιο απλή περίπτωση οι χρήστες δημιουργούν μόνοι τους φίλτρα πληροφοριών, με τη διευκρίνιση μερικών απλών κανόνων που ικανοποιούν τις ανάγκες τους. Ειδικευμένοι μηχανικοί πάνω στις διαδικασίες γνώσης μπορούν να παρέχουν συμβουλές πάνω σε συγκεκριμένα σχεδιαγράμματα προφίλ. Ενώ η πρώτη προσέγγιση είναι συχνά ιδιαίτερα χρονοβόρα, η δεύτερη στερεί τη δυνατότητα εξατομίκευσης σύμφωνα με τις ειδικές ανάγκες του χρήστη. Χαρακτηριστικά, τα προφίλ σκιαγραφούνται από ένα σύνολο μονάδων πληροφοριών κατάρτισης για τις οποίες ορίζονται οι βαθμοί σχετικότητας. Η συσχέτιση μιας πληροφορίας με το προφίλ αναζήτησης ενός χρήστη ως δείκτη ομοιότητας, αντιστοιχεί στη χαρτογράφηση της σχετικότητας όπως περιγράφεται στον ορισμό της επεξεργασίας της πληροφορίας.

2.3 Η ιδέα της σχετικότητας

Στη βιβλιογραφία υπάρχουν τρεις κατηγορίες επεξεργασίας πληροφορίας οι οποίες διαφέρουν στα κριτήρια που χρησιμοποιούνται για την επιλογή των σχετικών μονάδων πληροφορίας, όπως περιγράφονται συνοπτικά παρακάτω[Anagnostopoulos,04]:

Επεξεργασία βασιζόμενη στο περιεχόμενο ή γνωστική επεξεργασία

Με την παραπάνω τεχνική λαμβάνεται η απόφαση για την σχετικότητα της πληροφορίας βάσει των χαρακτηριστικών γνωρισμάτων που μπορούν να χρησιμοποιηθούν από το περιεχόμενο του κειμένου της, το οποίο κάθε χρήστης υποτίθεται ότι χρησιμοποιεί ανεξάρτητα. Οι πληροφορίες που πηγάζουν από τα έγγραφα αυτά, έχουν ως κοινά χαρακτηριστικά γνωρίσματα τα n-grams, τις λέξεις, τις ρίζες λέξεων ή φράσεις [Salton,89]. Αυτή είναι η πλέον γνωστή και εφαρμοσμένη τεχνική.

Επεξεργασία βάσει συνεργασίας χρηστών ή κοινωνικό φιλτράρισμα

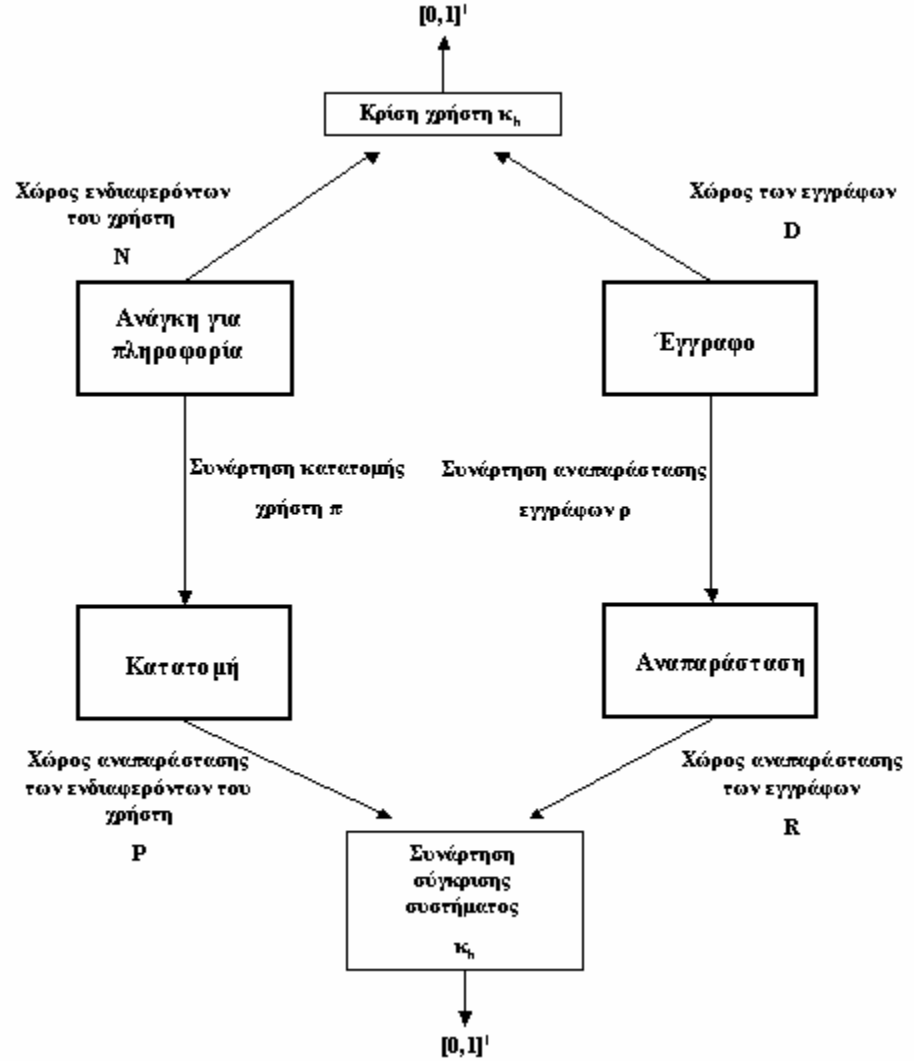
Η τεχνική αυτή, υποθέτει ότι ένας αποτελεσματικός τρόπος για την εύρεση της ενδιαφέρουσας ή σχετικής, προς τον χρήστη, πληροφορίας, αποτελεί η αναζήτηση χρηστών με αντίστοιχα ενδιαφέροντα [Breese,98]. Με άλλα λόγια το φιλτράρισμα της πληροφορίας έγκειται στην επιλογή της σχετικής πληροφορίας από τον χρήστη, η οποία βασίζεται στη χρήση, στις προτιμήσεις, στους σχολιασμούς ή στις απόψεις άλλων χρηστών. Έτσι για παράδειγμα, χαρακτηριστικό κοινωνικού φιλτραρίσματος αποτελεί η συχνότητα ή η αναλογία ανάκτησης μιας συγκεκριμένης πληροφορίας σε ένα σύνολο χρηστών. Με άλλα λόγια μια ιστοσελίδα με μεγαλύτερο ρυθμό καταφόρτωσης (download rate) είναι πιο σημαντική σε σύγκριση με μια άλλη που παρέχει ανάλογο ποσό πληροφορίας.

Επεξεργασία βάσει κόστους

Η τεχνική αυτή βασίζεται σε αξιολογήσεις κόστους-κέρδους θεωρώντας τις πληροφορίες ως προϊόντα σε οικονομική αγορά [Ferguson,96]. Στηριζόμενη σε ρητούς μηχανισμούς κοστολόγησης, η απόφαση για το αν η πληροφορία θεωρηθεί σχετική ή όχι, λαμβάνεται με σκοπό να υπάρχει ισορροπία μεταξύ κόστους και ανάκτησης. Θα πρέπει εδώ να σημειωθεί, ότι ο όρος κόστος δεν αναλογεί σε νομισματικές μονάδες. Παραδείγματος χάριν, η απόφαση να διαβαστεί ή να αγνοηθεί μια ιστοσελίδα που ικανοποιεί τα ενδιαφέροντα ενός χρήστη, μπορεί να βασιστεί απλώς στο μέγεθός της ή στο απαιτούμενο χρόνο καταφόρτωσης.

2.4 Γενικό πλαίσιο Συστήματος Επεξεργασίας Πληροφορίας

Οι μηχανισμοί ΕΠ συσχετίζονται έντονα με μηχανισμούς ανάκτησης πληροφοριών. Στην ενότητα αυτή θα περιγραφεί ένα γενικό μοντέλο ΕΠ που συνεργάζεται με μηχανισμούς εύρεσης και ανάκτησης αυτής, όπως απεικονίζεται στο Σχήμα 2.



Σχήμα 2. Γενικό πλαίσιο ενός συστήματος επεξεργασίας πληροφορίας

Έχοντας υπόψη τον ορισμό της ΕΠ, υποθέτοντας ένα σύνολο μονάδων πληροφορίας D και έναν ανεξάρτητο χρήστη με δεδομένες απαιτήσεις ανάμεσα σε ένα σύνολο χρηστών N, η διαδικασία της επεξεργασίας πληροφοριών αντιστοιχεί σε μια χαρτογράφηση $f_\psi : D \mapsto \{0,1\}$ όπου οι τιμές μηδέν και ένα αντιστοιχούν στην απόρριψη και την αποδοχή μιας μονάδας πληροφορίας αντίστοιχα. Αν κάποιος αναλύσει αυτήν την διαδικασία καθώς διευθύνεται από έναν άνθρωπο χωρίς τη βοήθεια μιας μηχανής, θα ανακαλύψει ότι ο χρήστης προσπαθεί να καλύψει την ανάγκη για πληροφορία μέσα σε ένα σύνολο μονάδων πληροφορίας με στόχο την ανίχνευση του σχετικού ποσού από αυτήν. Βέβαια, το ενδιαφέρον των χρηστών μπορεί να είναι πολύπλευρο και να βασιστεί έτσι σε διαφορετικές πτυχές.

Έτσι, τυποποιείται η κρίση του χρήστη σχετικά με την σχέση μεταξύ των ενδιαφερόντων του και μιας μονάδας πληροφορίας ως λειτουργία σύγκρισης $\kappa_h : N \times D \to [0,1]^I$, όπως φαίνεται στην κορυφή του Σχήματος 2. Βάσει του αποτελέσματος αυτής της σύγκρισης, ο χρήστης μπορεί

έπειτα να αποφασίσει, ακόμα και υποσυνείδητα, είτε να απορρίψει είτε να δεχτεί μια πληροφορία ως σχετική. Η ενέργεια αυτή αντιστοιχεί σε μια χαρτογράφηση $\tau_h : [0,1]^l \to \{0,1\}$. Λαμβάνοντας υπόψη τον ορισμό, για μια δεδομένη αναζήτηση πληροφορίας $\psi \in \mathrm{N}$ και για οποιαδήποτε μονάδα πληροφορίας $d \in D$ ισχύει η παρακάτω Σχέση.

$$f_\psi(d) = (\tau_h \circ \kappa_h)(\psi, d) \qquad \textbf{Σχέση 1}$$

Κάθε προσέγγιση αυτοματισμού αυτής της διαδικασίας έχει τέσσερα βασικά συστατικά:

- Μια τεχνική αντιπροσώπευσης των μονάδων πληροφορίας, που αντιστοιχεί στην συνάρτηση αντιπροσώπευσης $\rho : D \to R$. Αυτή η συνάρτηση χαρτογραφεί μια μονάδα πληροφορίας στην αντιπροσώπευσή της.
- Μια τεχνική για την αναπαράσταση των ερωτήσεων του χρήστη, που αντιστοιχεί στην συνάρτηση προφίλ $\pi : \mathrm{N} \to \mathrm{P}$. Αυτή η συνάρτηση, χαρτογραφεί με την σειρά της τα ενδιαφέροντα των χρηστών επάνω σε ένα σχεδιάγραμμα και δημιουργεί το προφίλ του χρήστη για το διάστημα P.
- Μια τεχνική για την ταυτοποίηση της αναπαράστασης των ερωτήσεων του χρήστη σε σχέση με τις αντιπροσωπεύσεις των μονάδων πληροφοριών, που ορίζεται ως σύγκρισης $\kappa_s : P \times R \to [0,1]^l$. Το διάστημα $[0,1]^l$ απεικονίζει τις διαφορετικές βαθμίδες του ενδιαφέροντος των χρηστών, οι οποίες μετρώνται σε *l* επίπεδα, παρόμοια με την προαναφερθείσα συνάρτηση ανθρώπινης κρίσης κ_h. Παραδείγματος χάριν, στην περίπτωση q διαφορετικών ερωτήσεων που διαμορφώνουν το προφίλ του χρήστη, θα μπορούσαμε να έχουμε *l=q* την τάξη ομοιότητες, μεταξύ μιας μονάδας πληροφορίας και κάθε μιας από τις ερωτήσεις.
- Μια τεχνική για την χρησιμοποίηση των αποτελέσματα της παραπάνω σύγκρισης, που ορίζεται ως συνάρτηση απόφασης του συστήματος $\tau_s : [0,1]^l \to \{0,1\}$. Η συνάρτηση αυτή τις περισσότερες φορές παρουσιάζει παρόμοια χαρακτηριστικά με την συνάρτηση ανθρώπινης απόφασης τ_h. Στην πραγματικότητα βέβαια, ένας άνθρωπος αποφασίζει συνήθως υποσυνείδητα ενώ η αντίστοιχη συνάρτηση για τα συστήματα παρέχει μια εξομοίωση αυτής. Παραδείγματος χάριν, εάν το προφίλ ενδιαφερόντων αποτελείται από l=q αριθμό ερωτήσεων που αντιπροσωπεύουν έναν αριθμό θεμάτων για τα οποία ο χρήστης ενδιαφέρεται ή όχι, αυτή η συνάρτηση θα μπορούσε απλά να επιστρέψει ως τιμή την ‘μονάδα’, εάν η αντίστοιχη μονάδα πληροφορίας είναι η πιο σχετική σε σχέση με τα ενδιαφέροντα του χρήστη. Σε διαφορετική περίπτωση, η συνάρτηση θα επέστρεφε ‘μηδέν’.

Σύμφωνα με αυτό το γενικευμένο μοντέλο, οι τέσσερις συναρτήσεις ρ, π, κ_s και τ_s πρέπει να υλοποιηθούν προκειμένου να θεμελιωθεί ο πυρήνας του ΣΕΠ. Ένας προφανής στόχος για ένα σύστημα επεξεργασίας, είναι η εξομοίωση της συνάρτησης απόφασης τ_s που βασίζεται στην συνάρτηση σύγκρισης κ_s με την συνάρτηση ανθρώπινης απόφαση τ_h που με την σειρά της βασίζεται στην συνάρτηση ανθρώπινης κρίσης κ_h [Oard,97].

$$(\tau_h \circ \kappa_h)(\psi, d) = (\tau_s \circ \kappa_s)(\pi(\psi), \rho(d)) \qquad \forall \psi \in \mathrm{N}, \quad \forall d \in D \qquad \textbf{Σχέση 2}$$

Από την παραπάνω Σχέση 2, φαίνεται ότι η διατύπωση μιας αντικειμενικής λειτουργίας για ένα σύστημα πληροφοριών είναι φαινομενικά απλή. Εντούτοις, η αξιολόγηση της απόδοσης ενός συστήματος που διέπεται από τους παραπάνω ορισμούς, σχέσεις και κανόνες δεν αποτελεί ένα τετριμμένο πρόβλημα, όπως θα αναλυθεί στην επόμενη ενότητα.

2.5 Αξιολόγηση απόδοσης σε Συστήματα Επεξεργασίας Πληροφορίας

Η αξιολόγηση ενός ΣΕΠ χρησιμοποιεί ένα σύνολο μετρήσεων που αφορούν τη δυνατότητα του συστήματος να ικανοποιεί τον χρήστη [vanRijsbergen,79]. Χαρακτηριστικά, περιλαμβάνει μετρήσεις σχετικά με την *αποδοτικότητα* και *την αποτελεσματικότητα* του συστήματος. Η αποτελεσματικότητα είναι ένα μέτρο της δυνατότητας του συστήματος να εντοπίζει την σχετική πληροφορία ενώ συγχρόνως απομονώνει την άσχετη. Με άλλα λόγια όσο πιο αποτελεσματικό είναι ένα σύστημα, τόσο περισσότερο ικανοποιεί το χρήστη [vanRijsbergen,79]. Εντούτοις, ένα αποτελεσματικό σύστημα πρέπει να είναι φιλικό προς τον χρήστη. Εκεί είναι που εισέρχεται η αποδοτικότητα του συστήματος. Η αποδοτικότητα είναι ένα μέτρο της απόδοσης σε σχέση με τους πόρους που καταναλώνονται για να επιτύχουν την επεξεργασμένη έξοδο.

Η απόφαση ταξινόμησης έχει να κάνει με το αν ένα έγγραφο είναι σχετικό ή όχι σε σχέση με μια ερώτηση ενός χρήστη. Έτσι λοιπόν η μέτρηση της αποτελεσματικότητας της επεξεργασίας πληροφορίας μπορεί να θεωρηθεί ως ένα δυαδικό πρόβλημα ταξινόμησης. Στην προηγούμενη ενότητα, περιγράφηκε η οντότητα ενός ΣΕΠ βάσει των συναρτήσεων ρ, π, κ_s και τ_s, με σκοπό να εξομοιωθεί μια ανθρώπινη κρίση κ_h , όπως απεικονίζεται στο Σχήμα 2, σε συνάρτηση με την ακόλουθη απόφαση τ_h. Ο στόχος αυτός βασίζεται στην υπόθεση ότι ένα έγγραφο μπορεί σαφώς να θεωρηθεί σχετικό ή άσχετο σύμφωνα με μια ορισμένη ανάγκη πληροφοριών. Στην πραγματικότητα αυτή η υπόθεση δεν επαληθεύεται πάντα, επειδή η σχετικότητα είναι μια υποκειμενική έννοια. Αυτό συμβαίνει γιατί οι κρίσεις των διαφορετικών χρηστών για τη σχετικότητα συγκεκριμένης πληροφορίας μπορούν να διαφέρουν σημαντικά [vanRijsbergen,79]. Ακόμη και οι κρίσεις του ίδιου χρήστη σε διαφορετικούς χρόνους μπορούν να διαφέρουν. Ως εκ τούτου, η αξιολόγηση της αποτελεσματικότητας ενός ΣΕΠ δεν αποτελεί μια απλή υπόθεση, ιδιαίτερα όταν υπάρχουν πολλές τυποποιημένες κατηγορίες ταξινόμησης, ενώ για κάθε εξεταζόμενο αντικείμενο ανατίθεται από τον ταξινομητή μία μόνο κατηγορία. Σε συστήματα επεξεργασίας, η ανάθεση μιας κατηγορίας εξαρτάται από την ανθρώπινη κρίση. Κατά συνέπεια οι μετρήσεις αποτελεσματικότητας ενός ΣΕΠ δεν είναι πάντα αντικειμενικές. Όμως το πρόβλημα των αντιφατικών ανθρώπινων κρίσεων μπορεί να εκφραστεί σε ένα πιθανολογικό πλαίσιο μέσω "a posteriori" πιθανοτήτων.

2.6 Μεγέθη Αξιολόγησης

Ο όρος Ανάκληση σε προβλήματα επεξεργασίας σχετίζεται με το αν η πληροφορία είναι σχετική ή όχι όσον αφορά τις απαιτήσεις του χρήστη. Έχοντας σκοπό την αξιολόγηση, ας υποτεθεί ένα σύνολο *n* μονάδων πληροφοριών που έχουν ταξινομηθεί από ένα σύστημα, είτε ως δεκτά είτε ως απορριπτέα, ενώ οι πραγματικές κατηγορίες παρέχονται από το χρήστη. Η σχέση μεταξύ των αποφάσεων ταξινόμησης και των πραγματικών κατηγοριών μπορεί να συνοψιστεί σε έναν πίνακα πιθανοτήτων όπως φαίνεται στον Πίνακα 1. Έτσι, η είσοδος *a* είναι ο αριθμός των μονάδων που το σύστημα αποδέχεται ως σχετικές και που είναι, στην πραγματικότητα σχετικές. Αντίστοιχα η είσοδος *d* είναι ο αριθμός των μονάδων που το σύστημα ορθά απορρίπτει ως μη-σχετικές. Από την άλλη, *b* είναι ο αριθμός των μονάδων που το σύστημα λανθασμένα αποδέχεται ως σχετικές, ενώ c είναι οι μονάδες πληροφορίας που λανθασμένα απορρίπτει σε σχέση με την κρίση του χρήστη [Μακρής,02].

2.6.1 Ορθότητα και Ρυθμός Λάθους

Απόφαση Συστήματος / Χρήστη	**Ποσό σχετικών μονάδων**	**Ποσό μη-σχετικών μονάδων**
Ποσό σχετικών	a	b

μονάδων		
Ποσό μη-σχετικών μονάδων	c	d

Πίνακας 1. Παράμετροι για την αξιολόγηση απόδοσης ενός συστήματος επεξεργασίας πληροφοριών

Κοινοί δείκτες μέτρησης για τον υπολογισμό της απόδοσης είναι η ορθότητα και ο ρυθμός λάθους. Ο ρυθμός λάθους υπολογίζει την πιθανότητα της λάθος ταξινόμησης ενώ η ορθότητα υπολογίζει την πιθανότητα σωστής ταξινόμησης μιας μονάδας πληροφορίας, ανεξάρτητα από την κατηγορία που του έχει αποδοθεί. Έτσι, το άθροισμα του λάθους και της ακρίβειας ισούται πάντα με την μονάδα. Βάσει του Πίνακα 1 καθορίζονται από τις παρακάτω σχέσεις:

$$ρυθμός\ λάθους = \frac{b+c}{n}, \qquad ορθότητα = \frac{a+d}{n} = 1 - ρυθμός\ λάθους$$

2.6.2 Ανάκληση και Ακρίβεια

Το πρόβλημα της σχετικής ταξινόμησης μεταξύ κάποιων μονάδων πληροφορίας είναι επίσης πολύ κοινό όσον αφορά την ανάκτηση πληροφοριών. Επιπλέον δείκτες για την μέτρηση της απόδοσης είναι η ανάκληση (*recall*), η ακρίβεια (*precision*) και *η διακοπή* (*fallout*), τα οποία δεν σχετίζονται με τον συνολικό αριθμό μονάδων και μπορούν έτσι να προσδιορίσουν την ακρίβεια και την αποτελεσματικότητα του συστήματος. Με άλλα λόγια, *η ανάκληση* είναι μια εκτίμηση της πιθανότητας που προσδίδει το σύστημα μέσω των σχετικών μονάδων πληροφορίας στο χρήστη, ενώ *η ακρίβεια* είναι μια εκτίμηση της πιθανότητας ότι μια μονάδα που παρουσιάζεται στο χρήστη είναι πράγματι σχετική. Από την άλλη η *διακοπή* είναι μια εκτίμηση της πιθανότητας ότι ένα έγγραφο, παρά του ότι είναι μη-σχετικό, παρουσιάζεται στο χρήστη. Οι δείκτες αυτοί συνοψίζονται από τις παρακάτω σχέσεις αντίστοιχα.

$$recall = \frac{a}{a+c} \qquad precision = \frac{a}{a+b} \qquad fallout = \frac{b}{b+d}$$

Λαμβάνοντας υπόψη την αποκαλούμενη παράμετρο *γενικότητας* g = (a+c)/n, που αποτελεί ένα μέτρο του ποσού των σχετικών μονάδων στο σύνολο όλων των μονάδων πληροφορίας, ορίζεται η ακόλουθη Σχέση 3 που συσχετίζει την παραπάνω παράμετρο, την ανάκληση και το fallout με σκοπό τον υπολογισμό της ακρίβειας.

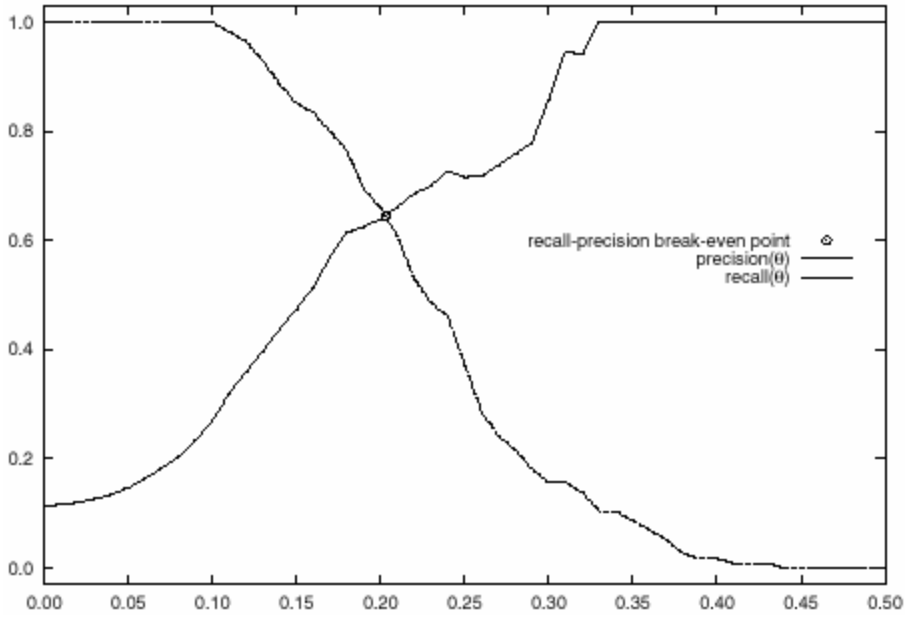
$$precision = \frac{g \cdot recall}{g \cdot recall + (1-g) \cdot fallout} \qquad \textbf{Σχέση 3}$$

Όπως φαίνεται στο Σχήμα 3, η ανάκληση και η ακρίβεια συσχετίζονται με έναν αόριστα διευκρινισμένο τρόπο [vanRijsbergen,79]. Προφανώς, η αποδοχή των περισσότερων μονάδων ως σχετικών παράγει έναν υψηλό βαθμό ανάκλησης σε χαμηλές τιμές ακρίβειας, ενώ αντίθετα απορρίπτοντας τα περισσότερα έγγραφα αποδίδονται στο σύστημα χαμηλές τιμές ανάκλησης σε υψηλές τιμές ακρίβειας. Οι επιλογές μεταξύ αυτών των άκρων συνιστούν κάποια αλληλεπίδραση (trade-off) μεταξύ της ανάκλησης και της ακρίβειας. Έτσι, μετρήσεις που βασίζονται μεμονωμένα από αυτούς τους δείκτες ενδέχεται να είναι παραπλανητικές. Σε προβλήματα ανάκτησης πληροφοριών, είναι αποδεκτό ότι για τη μέτρηση της αποτελεσματικότητας πρέπει να υπολογιστούν οι παράμετροι της ακρίβειας και της ανάκλησης [vanRijsbergen,79].

Έτσι, η ανάκληση και η ακρίβεια πρέπει να εξεταστούν σε συνδυασμό ώστε να εξασφαλίσουν μια μη-τετριμμένη και ουσιαστική αξιολόγηση της αποτελεσματικότητας ενός συστήματος πληροφοριών [Lewis,91]. Όμως, η απαίτηση για την κατοχή ζευγών αριθμών για την μέτρηση

της αποτελεσματικότητας, οδήγησε στον ορισμό σύνθετων μετρήσεων. Ένα κοινό μέτρο που χρησιμοποιείται συχνά για την σύγκριση τέτοιων συστημάτων είναι το σημείο εξισορρόπησης (break-even point) μεταξύ της ανάκλησης και της ακρίβειας [Lewis,92].

Η ιδέα είναι να ρυθμιστούν οι παράμετροι του συστήματος κατά τέτοιο τρόπο ώστε η τιμή της ανάκλησης του συστήματος να είναι ταυτόσημη με την ακρίβειά της, όπως φαίνεται στο παράδειγμα του Σχήματος 3. Όσο μεγιστοποιείται η τιμή αυτή τόσο μεγαλύτερη είναι και η αποτελεσματικότητα του συστήματος. Ένα βασικό μειονέκτημα της μεθόδου αυτής έγκειται στο γεγονός ότι οι τιμές της ανάκλησης και της ακρίβειας δεν μπορούν πάντα είναι ίσες. Επομένως, οι τιμές ανάκλησης και ακρίβειας πρέπει συχνά να παρεμβληθούν κατάλληλα, για να παραγάγουν ένα σημείο εξισορρόπησης που δεν είναι δυνατόν να επιτευχθεί από το σύστημα.



Σχήμα 3. Λειτουργικές χαρακτηριστικές καμπύλες Ανάκλησης και Ακρίβειας

Παρόλα αυτά, ο υπολογισμός του σημείου εξισορρόπησης δεν μπορεί να θεωρηθεί ως πηγή πληροφορίας από την πλευρά του χρήστη, παρά μόνο ένας ρυθμιστικός παράγοντας [Schapire,98]. Ο Van Rijsbergen εισήγαγε ένα σύνολο δεικτών μέτρησης οι οποίοι παραμετροποιούνται βάσει μιας τιμής β που απεικονίζει την σχετική σημασία που ένας χρήστης αποδίδει στην ανάκληση και την ακρίβεια, όπως παρουσιάζεται στην εργασία [vanRijsbergen,79].

$$F_\beta = \frac{(\beta^2 + 1) \cdot recall \cdot precision}{\beta^2 \cdot precision + recall} \qquad \textbf{Σχέση 4}$$

Στην παραπάνω Σχέση 4, ο δείκτης F_0 ταυτίζεται με την τιμή της ακρίβειας όταν $\beta=0$, ενώ αντίθετα ταυτίζεται με την τιμή της ανάκλησης όταν η παράμετρος β τείνει στο άπειρο. Το ίδιο βάρος ανατίθεται στην ανάκληση και την ακρίβεια όταν $\beta=1$, όπου ισχύει η Σχέση 5.

$$F_1 = \frac{2 \cdot recall \cdot precision}{precision + recall} \qquad \textbf{Σχέση 5}$$

Όλοι οι παραπάνω αναφερόμενοι δείκτες απαιτούν "a priori" τη γνώση για τις πραγματικές κατηγορίες που πρόκειται να ταξινομηθούν. Για αυτόν τον λόγο, το διαθέσιμο σύνολο των

εξεταζόμενων μονάδων πληροφορίας χωρίζεται συνήθως στα ανεξάρτητα σύνολα εκπαίδευσης και δοκιμής για την πειραματική αξιολόγηση της αποτελεσματικότητας του συστήματος.

Στο πεδίο της ΕΠ, είναι γνωστό ότι όταν ένα σύστημα αποδίδει καλά κάτω από έναν μεγάλο αριθμό πειραματικών όρων, η πιθανότητα να αποδώσει καλά σε μια λειτουργική κατάσταση όπου δεν είναι γνωστή η σχετικότητα του εγγράφου, είναι αυξημένη [vanRijsbergen,79]. Παράλληλα όμως, εάν το σύνολο των προς επεξεργασία μονάδων αλλάξει σε μια μακροπρόθεσμη εφαρμογή, η αποτελεσματικότητα ενός ΣΕΠ δεν μπορεί να εξασφαλιστεί στην πάροδο του χρόνου. Ιδιαίτερα δε, είναι σχεδόν αδύνατο να διατηρηθεί η αποτελεσματικότητα ενός τέτοιου συστήματος όταν αντλεί τις επεξεργασμένες και προς ταξινόμηση πληροφορίες από δυναμικές και διαρκώς εξελισσόμενες πηγές πληροφορίας, όπως το Διαδίκτυο. Επομένως, είναι απαραίτητο να ελεγχθεί η απόδοση του συστήματος ενώ αυτό είναι σε λειτουργία. Έτσι, πρέπει να αξιολογούνται συνεχώς τα παραπάνω προτεινόμενα μέτρα αποτελεσματικότητας, τα οποία απαιτούν την γνώση των πραγματικών κατηγοριών για ένα τουλάχιστον ποσό από την ταξινομημένη πληροφορία. Ακόμα όμως και αν είναι δυνατό να ληφθεί ανατροφοδοτούμενη πληροφορία από τις σχετικές πηγές, συνήθως το σύστημα αποκλείει την ανάλογη πληροφορία που μπορεί να προέρθει από τις μη-σχετικές πηγές. Ως αποτέλεσμα, αποφεύγεται η λεγόμενη υπερφόρτωση πληροφορίας στο σύστημα, με σημαντική όμως απώλεια στο διακριτικό ποσό αυτής. Αυτό αποτελεί ένα σοβαρό πρόβλημα που πρέπει να αντιμετωπιστεί στα πλαίσια του ποιοτικού ελέγχου για επεξεργασία πληροφορίας.

2.6.3 Αρμονικός Μέσος Όρος

Όπως έχει αναφερθεί προηγουμένως, μία μέθοδος που θα μπορεί να συνδυάζει και ανάκληση και ακρίβεια είναι συχνά επιθυμητή. Τέτοιες μετρήσεις συνδυάζει ο αρμονικός μέσος όρος ανάκλησης και ακρίβειας F που ορίζεται από την Σχέση 6, όπου r(j) είναι η ανάκληση για το j-οστό κείμενο στη διάταξη, P(j) είναι η ακρίβεια για το j-οστό κείμενο στη διάταξη και F(j) είναι ο αρμονικός μέσος όρος των r(j), P(j).

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \qquad \textbf{Σχέση 6}$$

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

2.6.4 Η Μετρική E

Μία άλλη μέθοδος υπολογισμού που συνδυάζει ακρίβεια και ανάκληση είναι η μετρική E. Το βασικό πλεονέκτημα της μεθόδου αυτής είναι ότι επιτρέπει στο χρήστη να προσδιορίσει αν τον ενδιαφέρει περισσότερο η ανάκληση ή η ακρίβεια. Η μετρική E ορίζεται από την Σχέση 7, όπου r(j) είναι η ανάκληση για το j-οστό κείμενο στη διάταξη, P(j) είναι η ακρίβεια για το j-οστό κείμενο στη διάταξη, E(j) είναι η μετρική E των r(j), P(j).

$$E(j) = 1 - \frac{1+b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}} \qquad \textbf{Σχέση 7}$$

Η *b* είναι μία παράμετρος που προσδιορίζεται από τον χρήστη και αντανακλά τη σχετική σημαντικότητα της ακρίβειας και της ανάκλησης. Όταν *b=1* η μετρική είναι το συμπλήρωμα του αρμονικού μέσου όρου, ενώ τιμές του *b* μεγαλύτερες από 1 υποδηλώνουν ότι ο χρήστης ενδιαφέρεται περισσότερο για την ακρίβεια παρά για την ανάκληση. Τέλος, οι τιμές του *b* που είναι μικρότερες από 1 υποδηλώνουν ότι ο χρήστης ενδιαφέρεται περισσότερο για την ανάκληση παρά για την ακρίβεια.

2.7 Σχέση Επεξεργασίας και Ανάκτησης Πληροφορίας

Ένας άλλος ερευνητικός τομέας σχετικός με την ΕΠ είναι αυτός της ανάκτησης πληροφοριών (ΑΠ). Η ΑΠ ενδιαφέρεται για τις διαδικασίες που περιλαμβάνονται στην αντιπροσώπευση, αποθήκευση, έρευνα, εύρεση και παρουσίαση των πληροφοριών που είναι σχετικές με μια απαίτηση για τις πληροφορίες αναζητούνται από έναν χρήστη [Ingwersen,92]. Εξετάζοντας τον μηχανισμό της ανάκτησης πληροφοριών ως τεχνική επιλογής συγκεκριμένων πληροφοριών, το φιλτράρισμα πληροφοριών μπορεί να αντιμετωπισθεί ως ειδική περίπτωση στην οποία το διάστημα των πληροφοριών είναι πολύ εξελισσόμενο και δυναμικό. Μια χρήσιμη περιγραφή της διαφοράς μεταξύ της επεξεργασίας και της ανάκτησης πληροφοριών παρέχεται στην [Belkin,92]. Πιο συγκεκριμένα, περιγράφεται ότι στην ανάκτηση πληροφοριών, η συλλογή των μονάδων πληροφορίας υποτίθεται ότι είναι σχετικά στατική, ενώ ο χρήστης υποβάλλει συχνά ερωτήσεις που αλλάζουν, δηλαδή είναι δυναμικές. Αντίθετα, στην ΕΠ, η ερώτηση είναι σχετικά σταθερή, ενώ η συλλογή των μονάδων υποτίθεται ότι είναι ένα σύνολο μονάδων πληροφορίας που ανανεώνεται συνεχώς και, ως εκ' τούτου, είναι δυναμικό. Συνεπώς, η επεξεργασία πληροφοριών έχει να κάνει περισσότερο με τη *διανομή* της πληροφορίας, ενώ η ανάκτηση πληροφοριών ενδιαφέρεται μάλλον για τη *συλλογή και την οργάνωση* των μονάδων πληροφορίας. Κατά συνέπεια, οι διαδικασίες ΕΠ και ΑΠ μπορούν να θεωρηθούν ως συμπληρωματικές.

3 Ανάκτηση πληροφορίας

3.1 Ορισμός Ανάκτησης Πληροφορίας

Η επιστήμη της Ανάκτησης Πληροφορίας (ΑΠ), ασχολείται με την αναπαράσταση, την αποθήκευση, την οργάνωση και την πρόσβαση σε πληροφοριακές μονάδες. Η αναπαράσταση και η οργάνωση των μονάδων αυτών πρέπει να γίνονται με τρόπο, ώστε να παρέχουν στον ανθρώπινο παράγοντα, εύκολη πρόσβαση στην πληροφορία που τον ενδιαφέρει.

Η ανάκτηση δεδομένων σε ένα περιβάλλον ΑΠ, συνίσταται στην εύρεση όλων των πληροφοριών οι οποίες αντιπροσωπεύονται από κάποιες λέξεις κλειδιά που εμφανίζονται σε ένα ερώτημα προς το ΣΕΠ. Αυτή η προσέγγιση δίνει συχνά κάτι διαφορετικό από αυτό που πραγματικά θέλει ο χρήστης. Στην πράξη, αυτό που περισσότερο ενδιαφέρει τον χρήστη ενός συστήματος ΑΠ, είναι να ανακτήσει πληροφορίες για ένα συγκεκριμένο θέμα, παρά δεδομένα σχετικά με κάποιο ερώτημα. Σε αντίθεση, μια γλώσσα ανάκτησης δεδομένων, στοχεύει στην ανάκτηση όλων των πληροφοριακών μονάδων, που ικανοποιούν ένα σύνολο καλά ορισμένων και διατυπωμένων συνθηκών, με μια κανονική έκφραση ή με σχεσιακή άλγεβρα. Αντίθετα στα συστήματα ΑΠ, τα ανακτώμενα αποτελέσματα μπορεί να είναι ανακριβή και η εμφάνιση κάποιων λαθών στα αποτελέσματα, περνά συχνά απαρατήρητη. Ο λόγος αυτής της διαφοροποίησης είναι ότι το σύστημα ΑΠ, διαχειρίζεται κείμενα γραμμένα σε φυσική γλώσσα, τα οποία δεν είναι πάντα επαρκώς δομημένα ενώ συχνά είναι και αμφίσημα.

Έτσι ενώ η ανάκτηση δεδομένων δίνει λύσεις στο χρήστη ενός συστήματος βάσης δεδομένων, δεν λύνει το πρόβλημα της ανάκτησης πληροφορίας, σχετικής με κάποιο θέμα. Για να μπορέσει ένα σύστημα ΑΠ να ανταποκριθεί στην πληροφοριακή ανάγκη του χρήστη, θα πρέπει να είναι σε θέση, να διερμηνεύσει το σημασιολογικό περιεχόμενο των μονάδων που διαχειρίζεται, και να τις διατάξει σύμφωνα με το βαθμό σχετικότητας του ερωτήματος του χρήστη. Ο κύριος στόχος ενός συστήματος ΑΠ, είναι να μπορεί να επιστρέψει όλες τις σχετικές μονάδες πληροφορίας που ανταποκρίνονται στο ερώτημα ενός χρήστη, επιστρέφοντας όμως παράλληλα και όσο το δυνατόν λιγότερες μη σχετικές. Γι' αυτό το λόγο η έννοια της σχετικότητας, διαδραματίζει κυρίαρχο ρόλο στην ανάκτηση πληροφορίας.

3.2 Ανάκτηση Πληροφορίας στον Παγκόσμιο Ιστό

Η αρχική ανάγκη για ανάπτυξη της ανάκτησης πληροφορίας ήταν η αυτοματοποιημένη δεικτοδότηση πληροφοριακών μονάδων και η ανάπτυξη μεθόδων για την αναζήτηση των σχετικών πηγών σε μια συλλογή. Σήμερα η έρευνα έχει επεκταθεί σε πολλούς παραπάνω τομείς, συμπεριλαμβάνοντας, την μοντελοποίηση, την ταξινόμηση και κατηγοριοποίηση των πληροφοριών, την οπτικοποίηση δεδομένων, την αρχιτεκτονική του συστήματος και τις διεπαφές με τον χρήστη. Η άποψη που επικρατούσε μέχρι στις αρχές τις δεκαετίας του 90, ήταν ότι η ΑΠ απευθυνόταν μόνο σε εφαρμογές βιβλιοθηκονομίας. Όμως η άποψη αυτή άλλαξε δραματικά με τη δημιουργία του Παγκοσμίου Ιστού.

Ο Παγκόσμιος Ιστός γίνεται μια ολοένα και μεγαλύτερη συλλογή ανθρώπινης γνώσης, που επιτρέπει την χωρίς προηγούμενο ανταλλαγή πληροφορίας και ιδεών σε έκταση πολύ μεγαλύτερη από ότι

είχε συλλάβει ο ανθρώπινος νους μέχρι τώρα. Η επιτυχία του Ιστού συνίσταται στην ευκολία που παρέχει στο χρήστη να δημιουργήσει τις δικές του Ιστοσελίδες, όντας έτσι ένα εύκολα προσπελάσιμο και σχετικά φθηνό μέσο προσωπικής έκφρασης. Επιπλέον η ύπαρξη του Ιστού, θέτει νέους τρόπους επικοινωνίας καταργώντας πολλές φορές τις έννοιες της χωρικής και χρονικής απόστασης. Τέλος οι τρέχουσες εξελίξεις στην ολοκλήρωση διαφορετικών υπηρεσιών γύρω από τον Ιστό, έχουν αλλάξει τον τρόπο που ο άνθρωπος βλέπει τον υπολογιστή. Έννοιες όπως Ηλεκτρονικό Εμπόριο και Ηλεκτρονική Διακυβέρνηση είναι δημοφιλείς και δημιουργούν νέες και πολλά υποσχόμενες αγορές.

Παρά την επιτυχημένη διάδοση του Παγκοσμίου Ιστού, η εύρεση χρήσιμης πληροφορίας, γίνεται μια ολοένα και πιο δύσκολη και επίπονη διαδικασία. Μια προσέγγιση θα ήταν η "περιπλάνηση" του χρήστη στον Κυβερνοχώρο, ακολουθώντας συνδέσμους που οδηγούν από σελίδα σε σελίδα, προσπαθώντας έτσι να εντοπίσει την πληροφορία που καλύπτει την πληροφοριακή του ανάγκη. Η παραπάνω διαδικασία περιπλάνησης, είναι συχνά αναποτελεσματική, λόγω του μεγέθους του Παγκοσμίου Ιστού και γιατί τις περισσότερες φορές ο χρήστης δεν γνωρίζει ένα καλό σημείο εκκίνησης. Για τους άπειρους χρήστες, το πρόβλημα της αναζήτησης γίνεται πολύ πιο δύσκολο που συχνά τους οδηγεί σε απογοητευτικά αποτελέσματα. Το κύριο εμπόδιο, είναι η απουσία ενός καλά ορισμένου μοντέλου δεδομένων για τον Παγκόσμιο Ιστό, το οποίο θα επισημαίνει ότι ο ορισμός και η δόμηση της πληροφορίας είναι χαμηλής ποιότητας. Αυτές οι δυσκολίες έστρεψαν το ενδιαφέρον στον τομέα της ΑΠ και οδήγησαν στην υιοθέτηση των τεχνικών που χρησιμοποιούνται στο πεδίο της ΑΠ, ως πολλά υποσχόμενων λύσεων. Πριν αναλυθούν οι βασικοί μηχανισμοί και παράγοντες που εμπλέκονται στις διαδικασίες ΑΠ, πρέπει να σημειωθεί ότι η διατριβή αυτή επικεντρώνεται στις σχετικές διαδικασίες που επιτελούνται στο Διαδίκτυο, όπου οι βασικές μονάδες που παρέχουν πληροφορίες στους χρήστες είναι οι ιστοσελίδες. Συνεπώς, στο υπόλοιπο της διατριβής οι πληροφοριακές μονάδες θα αναφέρονται και ως ιστοσελίδες. Λόγω δε της φυσικής τους απεικόνισης και της συσχέτισης με το αρχικό πρόβλημα διαχείρισης εγγράφων/κειμένων θα αποκαλούνται επιπλέον και ως έγγραφα ή κείμενα.

3.3 Μοντέλα Ανάκτησης Πληροφοριών

Όπως αναφέρθηκε παραπάνω, η πιο συνηθισμένη πρακτική για την δεικτοδότηση και την ανάκτηση κειμένων είναι η χρήση των όρων δεικτοδότησης. Ένας όρος δεικτοδότησης είναι μια λέξη κλειδί ή μια ομάδα εννοιολογικά συσχετιζόμενων λέξεων, η εμφάνιση των οποίων λαμβάνει από μόνη της μια αυτόνομη έννοια. Κατά μια πιο απλοποιημένη εκδοχή, ένας όρος δεικτοδότησης είναι απλά μια λέξη που εμφανίζεται σε ένα κείμενο της συλλογής. Η ανάκτηση που βασίζεται στο ταίριασμα όρων δεικτοδότησης ερωτήματος και κειμένων της συλλογής, είναι πολύ απλή αλλά εισάγει ένα σύνολο προβληματισμών για την αποτελεσματικότητα της. Για παράδειγμα, η βασική υπόθεση που εισάγει η παραπάνω στρατηγική, είναι ότι η σημασιολογία τόσο των κειμένων όσο και της πληροφοριακής ανάγκης του χρήστη, μπορεί να εκφραστεί με φυσικό τρόπο, μέσα από ένα σύνολο λέξεων. Στην πράξη ένα σημαντικό κομμάτι από τη σημασιολογία του κειμένου χάνεται κατά τη μεταφορά στο χώρο του ευρετηρίου.

Ο λόγος γι' αυτήν την απώλεια είναι ότι οι λέξεις αποκτούν την ερμηνεία τους ανάλογα με το πλαίσιο συμφραζομένων στο οποίο εμφανίζονται. Από αυτή την παρατήρηση πηγάζουν τα φαινόμενα της πολυσημίας και της συνωνυμίας. Στην πολυσημία, ο ίδιος όρος λαμβάνει διαφορετικές έννοιες ανάλογα με τα συμφραζόμενα που συνοδεύουν την εμφάνιση του, ενώ στην συνωνυμία, διαφορετικοί όροι μπορούν να περιγράφουν την ίδια έννοια γιατί εμφανίζονται στα ίδια πλαίσια των συμφραζομένων. Η συνωνυμία και η πολυσημία, αποτελούν κλασσικά προβλήματα που συνδέονται με τον τρόπο λογικής αναπαράστασης των κειμένων μέσω ευρετηρίου.

Έχοντας υπόψη τα παραπάνω προβλήματα και με δεδομένο ότι η διαδικασία της αντιστοίχησης του ερωτήματος στη συλλογή των κειμένων, γίνεται στο χώρο του ευρετηρίου, είναι εύκολα κατανοητό

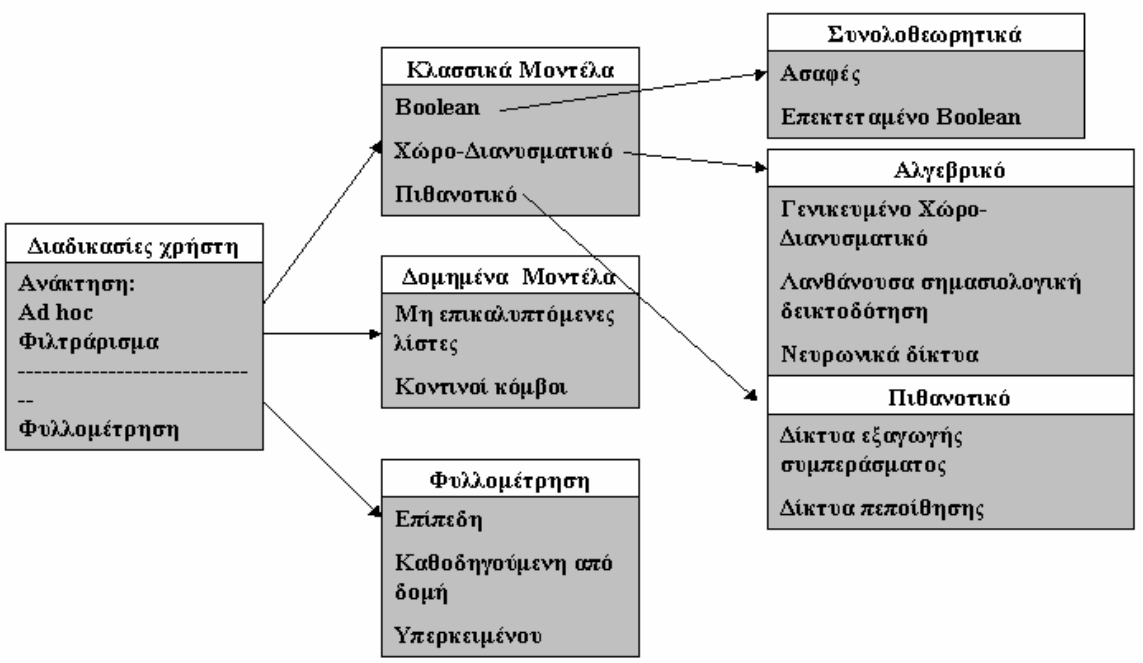
γιατί συχνά τα αποτελέσματα μιας ερώτησης διατυπωμένης με λέξεις-κλειδιά δεν είναι τα αναμενόμενα. Αν μάλιστα ληφθεί υπόψη ότι πολλοί χρήστες δεν είναι σε θέση να επιλέξουν τις κατάλληλες λέξεις-κλειδιά για τον σχηματισμό ερωτήσεων, το πρόβλημα μεγαλώνει. Ένα καλό παράδειγμα του παραπάνω προβλήματος είναι τα απογοητευτικά αποτελέσματα σε πολλά από τα ερωτήματα που υποβάλλονται στις μηχανές αναζήτησης στο Διαδίκτυο. Η πρόκληση για ένα μοντέλο ανάκτησης πληροφορίας, είναι να δημιουργήσει το υπόβαθρο, ώστε να υπάρξει ταίριασμα της πληροφοριακής ανάγκης χρήστη με τα κείμενα της συλλογής, παρά την ανακριβή αναπαράσταση και με όσο το δυνατόν μικρότερες αποκλίσεις.

Στις διαδικασίες ΑΠ, ταίριασμα σημαίνει εκτίμηση από το σύστημα, της σχετικότητας των κειμένων ως προς το δοθέν ερώτημα. Μια τέτοια εκτίμηση επιτυγχάνεται με την χρήση ενός αλγορίθμου κατάταξης, βάσει του οποίου γίνεται μια απλή διάταξη των κειμένων. Τα κείμενα που εμφανίζονται στις πρώτες θέσεις αυτής της διάταξης, θεωρούνται ως το πιο πιθανό να είναι σχετικά με την ερώτηση, με την τιμή της πιθανότητας αυτής να φθίνει, όσο εξετάζουμε τη διάταξη προς τις χαμηλότερες θέσεις. Οι αλγόριθμοι κατάταξης έχουν ζωτική σημασία σε ένα σύστημα ΑΠ. Συνεπώς μια βασική λειτουργία του μοντέλου είναι να παρέχει έναν αλγόριθμο κατάταξης για κάθε ερώτημα που υποβάλλεται.

Ο τρόπος θεώρησης της λογικής αναπαράστασης των κειμένων και η συσχέτιση του με τον αλγόριθμο κατάταξης, είναι το βασικό χαρακτηριστικό που διαφοροποιεί τα μοντέλα ΑΠ. Στην ενότητα αυτή παρατίθεται μια κατηγοριοποίηση των μοντέλων, κάποιοι τυπικοί ορισμοί αυτών και παρουσιάζονται τα κυριότερα μοντέλα ΑΠ.

3.3.1 Ταξινόμηση των Μοντέλων για Ανάκτηση Πληροφορίας

Τα τρία κλασσικά μοντέλα στην Ανάκτηση Πληροφορίας είναι το Boolean, το Χώρο-Διανυσματικό και το Πιθανοτικό. Στο Boolean μοντέλο, τόσο τα κείμενα όσο και τα ερωτήματα αντιμετωπίζονται ως ένα σύνολο από όρους δεικτοδότησης. Στο Χώρο-Διανυσματικό, τα κείμενα και τα ερωτήματα αναπαρίστανται ως διανύσματα σε έναν πολυδιάστατο χώρο. Τέλος το Πιθανοτικό μοντέλο εισάγει έναν τρόπο αναπαράστασης, ο οποίος βασίζεται σε πιθανοθεωρητικές θεωρήσεις. Τα τρία αυτά μοντέλα θεωρούνται ως αντιπρόσωποι των συνολοθεωρητικών, αλγεβρικών και πιθανοτικών μοντέλων.



Σχήμα 4. Ταξινόμηση των μοντέλων Ανάκτησης Πληροφορίας

Με τον καιρό προτάθηκαν διάφορες νέες προσεγγίσεις σε καθεμία από τις κατηγορίες βασικών μοντέλων. Έτσι στο συνολοθεωρητικό πεδίο προτάθηκαν επιπλέον το Boolean ασαφές και το επεκτεταμένο Boolean. Στα αλγεβρικά μοντέλα προτάθηκε αντίστοιχα το γενικευμένο Χώρο-Διανυσματικό μοντέλο, το μοντέλο λανθάνουσας σημασιολογικής δεικτοδότησης (Latent Semantic Indexing, LSI) και το μοντέλο των νευρωνικών δικτύων. Στον πιθανοτικό τομέα εμφανίστηκαν τα δίκτυα εξαγωγής συμπεράσματος και τα δίκτυα πεποίθησης. Το Σχήμα 4 δίνει σχηματικά την κατηγοριοποίηση αυτή.

Εκτός από την χρήση του περιεχομένου των κειμένων, ορισμένα μοντέλα εκμεταλλεύονται και την εσωτερική δομή που φυσιολογικά υπάρχει στο γραπτό λόγο. Σε αυτή την περίπτωση λέμε ότι έχουμε ένα δομημένο μοντέλο. Για τη δομημένη ανάκτηση κειμένου, ορίζονται οι μη-επικαλυπτόμενες λίστες και οι κοντινοί κόμβοι .

Όπως αναφέρθηκε και πριν, οι διαδικασίες του χρήστη ενδέχεται να έχει την μορφή φυλλομέτρησης. Σε αυτή την κατηγορία εντοπίζονται τρία μοντέλα. Την επίπεδη, την καθοδηγούμενη από τη δομή και τη φυλλομέτρηση υπερκειμένου .

Στο κεφάλαιο αυτό αναπτύσσονται μόνο τα συνολοθεωρητικά και αλγεβρικά μοντέλα καθώς και το βασικό πιθανοτικό μοντέλο. Τα υπόλοιπα μοντέλα αναφέρονται για την πληρότητα της διατριβής, ενώ ο αναγνώστης μπορεί να ανατρέξει στην [Baeza-Yates,99] για την εκτενέστερη περιγραφή όλων των μοντέλων.

3.3.2 Ανάκτηση και επεξεργασία ad-hoc

Στα περισσότερα συστήματα ΑΠ, η συλλογή των κειμένων παραμένει σχεδόν στατική, ενώ από την άλλη υποβάλλονται συνέχεια καινούρια ερωτήματα. Αυτός ο τρόπος λειτουργίας έχει ονομαστεί ως ad-hoc ανάκτηση πληροφορίας και είναι η πιο κοινή από τις διαδικασίες του χρήστη. Μια δεύτερη παρόμοια διαδικασία χρήστη είναι το λεγόμενο φιλτράρισμα πληροφορίας. Σε

αντίθεση με την παραπάνω διαδικασία, τα ερωτήματα παραμένουν σχεδόν σταθερά, ενώ η συλλογή των κειμένων μεταβάλλεται με καινούρια κείμενα.

Στο φιλτράρισμα κατασκευάζεται το προφίλ του χρήστη, το οποίο περιγράφει τις προτιμήσεις του. Το προφίλ αυτό συγκρίνεται με κάθε εισερχόμενο κείμενο για να αποφασίσει το σύστημα αν είναι σχετικό ή όχι ανάλογα με τις προτιμήσεις του χρήστη. Με άλλα λόγια το προφίλ είναι μια εναλλακτική μορφή ερωτήματος προς το σύστημα.

Στις διαδικασίες φιλτραρίσματος συνήθως παρέχονται στο χρήστη κείμενα που πιθανόν να τον ενδιαφέρουν χωρίς καμία κατάταξη σχετικότητας. Ο χρήστης καλείται από τον παρεχόμενο αριθμό κειμένων να επιλέξει αυτά που πραγματικά τον αφορούν και να αγνοήσει τα υπόλοιπα. Μερικές φορές παρουσιάζονται στοιχεία κατάταξης στο χρήστη, με τη λογική ότι αυτός θα εξετάσει τα κείμενα με την υψηλότερη κατάταξη, εξετάζοντας έτσι έναν ακόμα μικρότερο αριθμό κειμένων. Η διαδικασία αυτή ονομάζεται δρομολόγηση.

Παρότι ο χρήστης μπορεί να μην έχει εικόνα για την κατάταξη σχετικότητας, μια τέτοια κατάταξη υπολογίζεται εσωτερικά στο σύστημα. Σκοπός του υπολογισμού αυτού είναι να γίνει διαχωρισμός των κειμένων σε σχετικά και μη σχετικά, ως προς το παρεχόμενο προφίλ. Ο διαχωρισμός γίνεται με τον υπολογισμό της σχετικότητας από τον αλγόριθμο που παρέχει το μοντέλο που χρησιμοποιείται και τη σύγκριση με κάποια προκαθορισμένη τιμή κατωφλίου. Τα κείμενα που κατατάσσονται πάνω από αυτό το κατώφλι εκτιμώνται ως σχετικά ενώ τα υπόλοιπα θεωρούνται μη-σχετικά. Για τη διαδικασία κατάταξης μπορεί να χρησιμοποιηθεί οποιοδήποτε μοντέλο ΑΠ, συνήθως όμως για λόγους απλότητας χρησιμοποιείται το Χώρο-Διανυσματικό μοντέλο.

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

Μια πιο καλή στρατηγική είναι η κατασκευή ενός αρχικού προφίλ από τον ίδιο το χρήστη με τη χρήση λέξεων κλειδιών. Στη συνέχεια, ο χρήστης καλείται να αξιολογήσει τη σχετικότητα των κειμένων που του παρουσιάζονται από το σύστημα ως πιθανόν ενδιαφέροντα. Τα κείμενα που αξιολογήθηκαν ως σχετικά αλλά και τα μη σχετικά, χρησιμοποιούνται από το σύστημα για να αναδιατυπώσουν το προφίλ του χρήστη, βάσει ενός κύκλου ανάδρασης. Η παραπάνω διαδικασία θα συγκλίνει σε ένα σχεδόν αμετάβλητο προφίλ, από τη στιγμή και μετά που η πληροφορία που έρχεται από την ανάδραση χρήστη, έχει ήδη χρησιμοποιηθεί σε κάποια προηγούμενη χρονική στιγμή. Έτσι, το φιλτράρισμα μπορεί να θεωρηθεί ως μια διαδικασία ΑΠ, όπου τα ερωτήματα αναπαριστώνται από το προφίλ του χρήστη που παραμένει σταθερά αλλά τα κείμενα αλλάζουν συνεχώς. Το φιλτράρισμα όπως και η αναζήτηση πληροφορίας είναι δύο διαφορετικές μορφές διαδικασίας σε επίπεδο χρήστη. Συνεπώς στο φιλτράρισμα ενδέχεται να χρησιμοποιηθεί οποιοδήποτε μοντέλο χρησιμοποιείται για ανάκτηση πληροφορίας. Η δυσκολία όμως έγκειται συνήθως στον καθορισμό του προφίλ χρήστη. Μια προσέγγιση απαιτεί τον καθορισμό του προφίλ από τον ίδιο το χρήστη με την διατύπωση κάποιων λέξεων κλειδιών, ενώ κάποια άλλη απαιτεί την συλλογή ενός αρχικού συνόλου από το χρήστη και την αξιοποίηση των προτιμήσεων του για τη δυναμική ενημέρωση του προφίλ του. Εξίσου όμως ενθαρρυντικά αποτελέσματα προκύπτουν και από την εφαρμογή τεχνικών όπου το προφίλ του χρήστη προκύπτει τόσο από την δική του συνεισφορά όσο και από αυτόματες διαδικασίες επιλογής όρων βάσει τεχνικών επεξεργασίας της πληροφορίας [Anagnostopoulos,03], [Anagnostopoulos,04].

3.3.3 Ορισμός Μοντέλων Ανάκτησης Πληροφορίας

Πριν εξεταστούν τα επί μέρους μοντέλα, στην υπό-ενότητα αυτή θα δοθεί ένας τυπικός και ακριβής ορισμός για το τι είναι ένα μοντέλο ΑΠ [Baeza-Yates,99].

Ένα μοντέλο ανάκτησης πληροφορίας είναι η τετράδα *[D, Q, F, R(q_i,d_j)]* όπου το *D* είναι ένα σύνολο από λογικές αναπαραστάσεις για τα κείμενα της συλλογής, το *Q* αντιπροσωπεύει ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές ανάγκες (ερωτήσεις) του χρήστη, το *F* αποτελεί το υπόβαθρο για την μοντελοποίηση της αναπαράστασης των κειμένων, των ερωτημάτων και των σχέσεων μεταξύ τους και το *R(q_i,d_j)* είναι μια συνάρτηση κατάταξης, η οποία συνδέει έναν πραγματικό αριθμό με ένα ερώτημα $q_i \in Q$ και μια αναπαράσταση κειμένου $d_j \in D$. Μια τέτοια κατάταξη ορίζει μια διάταξη πάνω στα κείμενα πάντα με βάση το ερώτημα q_i.

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

3.3.4 Κλασσικά Μοντέλα Ανάκτησης Πληροφορίας

Στην ενότητα αυτή θα παρουσιαστούν εν συντομία το Boolean, το Χώρο-Διανυσματικό και το Πιθανοτικό μοντέλο ανάκτησης πληροφοριών.

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

Με δεδομένη την αναπαράσταση των κειμένων ως συλλογές όρων, δεν έχουν όλοι οι όροι την ίδια ισχύ ως προς την περιγραφή ενός κειμένου. Με άλλα λόγια η ερμηνεία ενός όρου συχνά μπορεί να δίνει μια γενικευμένη ή και ασαφή περιγραφή. Τέτοιοι όροι είναι αυτοί που εμφανίζονται με μεγάλη συχνότητα στην πλειονότητα των κειμένων μιας συλλογής. Ας ληφθεί ως παράδειγμα μια συλλογή κειμένων γύρω από την θεματική ενότητα των υπολογιστών. Ο όρος "Υπολογιστής" σε μια τέτοια συλλογή εμφανίζεται με μεγάλη βεβαιότητα σχεδόν σε κάθε κείμενο που αν και περιγράφει κάτι αρκετά συγκεκριμένο, δεν αποτελεί αντιπροσωπευτικό όρου του συγκεκριμένου κειμένου στο οποίο εμφανίζεται. Αντίθετα, αν μια λέξη εμφανίζεται σε μικρό εύρος κειμένων, τότε είναι σχεδόν σίγουρο ότι έχει μεγαλύτερη βαρύτητα στην περιγραφή ενός εξ αυτών. Ο όρος "κληρονομικότητα", εμφανίζεται σίγουρα σε πολύ λιγότερα κείμενα απ' ότι ο όρος "υπολογιστής". Η εμφάνιση του ως όρος δεικτοδότησης για ένα συγκεκριμένο κείμενο, οδηγεί στο συμπέρασμα ότι το συγκεκριμένο κείμενο αναφέρεται στην ιδιότητα της κληρονομικότητας του αντικειμενοστραφούς προγραμματισμού. Για να προσομοιωθεί το γεγονός ότι διαφορετικοί όροι μπορούν να έχουν

διαφορετική βαρύτητα ως προς στην δεικτοδότηση των κειμένων, σε κάθε όρο δεικτοδότησης ανατίθεται ένα αριθμητικό βάρος.

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

3.3.5 Δεικτοδότηση Βάρους Όρου

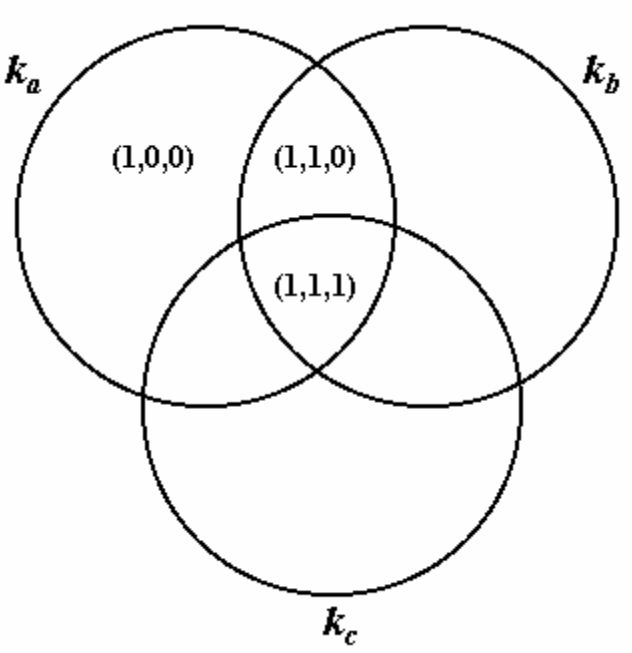
Έστω t είναι ο αριθμός των όρων δεικτοδότησης στο σύστημα και k_i, είναι ένας γενικός όρος δεικτοδότησης. Το σύνολο $K = \{k_1, k_2, \ldots, k_t\}$ είναι το σύνολο όλων των όρων δεικτοδότησης. Ένα βάρος $w_{ij} > 0$ συνδέεται με κάθε όρο k_i, που εμφανίζεται στο κείμενο d_j. Για κάποιον όρο δεικτοδότησης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

Τα παραπάνω βάρη είναι μεταξύ τους ανεξάρτητα, δηλαδή η τιμή του w_{ij} δεν επηρεάζει την τιμή του w_{i+1j}. Αυτή η υπόθεση είναι απλουστευτική δεδομένου ότι συχνά έχουμε συμπλέγματα όρων που εμφανίζονται μαζί. Ένα τέτοιο παράδειγμα είναι οι όροι "δίκτυο" και "υπολογιστής". Σε μια συλλογή με θέμα τα δίκτυα υπολογιστών, αναμένεται αυτοί οι δύο όροι να έχουν παρόμοιες συχνότητες εμφάνισης. Κατά συνέπεια οι δύο αυτοί όροι είναι συσχετισμένοι μεταξύ τους και ο υπολογισμός της ανάθεσης βαρών θα πρέπει να λαμβάνει υπόψη του αυτή τη συσχέτιση. Λαμβάνοντας υπόψη μας τις συσχετίσεις των όρων μεταξύ τους, η πολυπλοκότητα υπολογισμού των βαρών αυξάνει, συμπαρασύροντας και τον υπολογισμό της κατάταξης. Για το λόγο αυτό, οι διακριτοί όροι δεικτοδότησης θα θεωρούνται ότι είναι μεταξύ τους ανεξάρτητοι.

3.3.6 Το Boolean Μοντέλο

Το Boolean μοντέλο, είναι ένα απλό μοντέλο ανάκτησης πληροφορίας που το υπόβαθρό του και τα ερωτήματα που υποβάλει ο χρήστης βασίζονται στη θεωρία συνόλων και στη Boolean άλγεβρα. Συγκεκριμένα στο Boolean μοντέλο, κάθε όρος δεικτοδότησης θεωρείται ότι είτε ανήκει εξ ολοκλήρου σε ένα κείμενο είτε όχι. Κατά συνέπεια τα βάρη θεωρούνται δυαδικά, ως. $w_{ij} \in \{0,1\}$. Το κάθε ερώτημα θεωρείται ότι αποτελείται από όρους δεικτοδότησης οι οποίο συνδέονται με έναν από τους τελεστές and, or, not. Δηλαδή κάθε ερώτημα είναι μια Boolean έκφραση που μπορεί να γραφεί σε Διαζευκτική Κανονική Μορφή (ΔΚΜ). Για παράδειγμα το ερώτημα [$q = k_a \wedge (k_b \vee -k_c)$] μπορεί να γραφεί στην μορφή ΔΚΜ ως [$q_{\Delta KM} = (k_a \wedge k_b) \vee (k_a \vee -k_c)$]. Έστω τώρα ένα διάνυσμα με δυαδικά βάρη που αντιστοιχεί σε ανάθεση αλήθειας σε συζευκτικές εκφράσεις της τριάδας (k_a, k_b, k_c). Για παράδειγμα στην έκφραση $k_a \wedge k_b$ μια ανάθεση αλήθειας είναι η (1,1,0).

Άρα το αρχικό ερώτημα μπορεί να αναλυθεί σε διάζευξη τέτοιων διανυσμάτων ως [$\overline{q}_{\Delta KM} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$]. Τα δυαδικά αυτά διανύσματα εισήχθησαν επειδή υπάρχει απευθείας αντιστοιχία του ερωτήματος $\overline{q}_{\Delta KM}$, όπως φαίνεται και στο Σχήμα 5.



Σχήμα 5. Συζευκτικές συνιστώσες του ερωτήματος $q = k_a \wedge (k_b \vee -k_c)$

3.3.7 Ανάθεση Βαρών Δεικτοδότησης

Στο Boolean μοντέλο, τα βάρη που ανατίθενται στους όρους δεικτοδότησης είναι δυαδικά δηλαδή, $w_{ij} \in \{0,1\}$. Ένα ερώτημα q είναι μια συνήθης Boolean έκφραση. Έστω $\overline{q}_{\Delta KM}$ η διαζευκτική κανονική μορφή του ερωτήματος και $\overline{q}cc$ καθεμία από τις συζευκτικές συνιστώσες του $\overline{q}_{\Delta KM}$. Η ομοιότητα του κειμένου d_j προς το ερώτημα q ορίζεται από την παρακάτω Σχέση 8.

$$sim(d_j, q) = \begin{cases} 1, \text{εάν } \exists \overline{q}_{cc} \text{ ώστε } (\overline{q}_{cc} \in \overline{q}_{\Delta KM}) \wedge (\forall k_i, g_i(d_j) = g_i(\overline{q}_{cc})) \\ 0 \qquad\qquad \text{σε άλλη περίπτωση} \end{cases} \qquad \textbf{Σχέση 8}$$

Αν $sim(d_j, q) = 1$, τότε το Boolean μοντέλο προβλέπει ότι το κείμενο d_j είναι σχετικό με το ερώτημα q, ενώ σε οποιαδήποτε άλλη περίπτωση είναι άσχετο. Με άλλα λόγια στο μοντέλο αυτό δεν υπάρχει η έννοια της μερικής ικανοποίησης των συνθηκών του ερωτήματος. Για παράδειγμα έστω d_j τέτοιο ώστε να είναι d_j = (0,1,0). Το κείμενο αυτό περιέχει τον όρο k_b, αλλά θεωρείται άσχετο ως προς το ερώτημα [$q = k_a \wedge (k_b \vee -k_c)$] . Λόγω αυτής της έλλειψης, το Boolean μοντέλο στην ουσία εκτελεί περισσότερο ανάκτηση δεδομένων παρά πληροφορίας.

Το κύριο πλεονέκτημα του Boolean μοντέλου είναι η απλότητά του. Το κύριο μειονέκτημά του είναι ότι δεν υπάρχει διαβάθμιση σχετικότητας ως προς το ερώτημα κάτι που μπορεί να οδηγήσει σε χαμηλής ποιότητας ανάκτηση πληροφορίας. Ένα δεύτερο μειονέκτημά του είναι ότι συχνά δεν είναι εύκολη η έκφραση της πληροφοριακής ανάγκης του χρήστη με την τυποποίηση που επιβάλλει το μοντέλο αυτό. Λόγω αυτών των χαρακτηριστικών του, το Boolean μοντέλο έχει βρει εφαρμογή σε κυρίως εμπορικά συστήματα βιβλιοθηκών.

3.3.8 Το Χώρο – Διανυσματικό Μοντέλο

Το Χώρο-Διανυσματικό μοντέλο, αντιμετωπίζει την ανεπάρκεια της ανάθεσης δυαδικών βαρών και εισάγει ένα υπόβαθρο, στο οποίο επιτρέπεται το προσεγγιστικό ταίριασμα [Salton,68], [Salton,71]. Τα βάρη που ανατίθενται στους όρους δεικτοδότησης, τόσο για τα κείμενα όσο και για τα ερωτήματα είναι μη δυαδικά και χρησιμοποιούνται για τον υπολογισμό του βαθμού ομοιότητας μεταξύ του ερωτήματος και κάθε αποθηκευμένου κειμένου. Κατόπιν, τα κείμενα διατάσσονται με φθίνουσα σειρά, με κριτήριο τον βαθμό ομοιότητάς τους με το ερώτημα του χρήστη. Έτσι στο μοντέλο αυτό λαμβάνονται υπόψη και κείμενα που ικανοποιούν μερικώς τις συνθήκες του ερωτήματος και το τελικό αποτέλεσμα είναι πολύ πιο ακριβές σε σχέση με την ανάκτηση από το Boolean μοντέλο.

Ανάθεση Βαρών Δεικτοδότησης

Στο Χώρο-Διανυσματικό μοντέλο το βάρος w_{ij} που αντιστοιχεί στο ζεύγος (k_i,d_j) είναι φυσικός αριθμός και όχι δυαδικός. Επιπλέον ανατίθενται βάρη και στους όρους δεικτοδότησης του ερωτήματος που υποβάλλει ο χρήστης. Έστω λοιπόν ότι $w_{i,q}$ το βάρος που αντιστοιχεί στο ζεύγος $[k_i,q]$, όπου $w_{i,q} \geq 0$. Τότε το διάνυσμα του ερωτήματος ορίζεται ως $q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ όπου t είναι ο συνολικός αριθμός των όρων δεικτοδότησης στο σύστημα. Όπως και πριν το διάνυσμα του d_j είναι $d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$.

Μ' αυτόν τον τρόπο το κείμενο d_j και το ερώτημα χρήστη q αναπαρίστανται σα διανύσματα διαστάσεως t. Στο μοντέλο αυτό προτείνεται ο βαθμός της ομοιότητας μεταξύ του κειμένου d_j και του ερωτήματος q να υπολογιστεί ως ο βαθμός συσχέτισης μεταξύ των δύο διανυσμάτων. Μέτρο του βαθμού συσχέτισης αποτελεί το συνημίτονο της γωνίας που περιέχεται μεταξύ των δύο διανυσμάτων, που παρέχεται από την ακόλουθη Σχέση 9, όπου $|d_j|$ και $|q|$ οι νόρμες των διανυσμάτων.

$$sim(d_j, q) = \frac{d_j \cdot q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{t} (w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$ **Σχέση 9**

Εφόσον $w_{ij} \geq 0$ και $w_{i,q} \geq 0$, η τιμή $sim(d_j, q)$ παίρνει τιμές από 0, όπου δεν υπάρχει καθόλου ταύτιση έως 1 και τα διανύσματα ταυτίζονται πλήρως. Έτσι το μοντέλο αυτό, αντί να προσπαθήσει να προσδιορίσει αν ένα κείμενο είναι ή όχι σχετικό, διατάσσει τα κείμενα με κριτήριο τον βαθμό ομοιότητας τους προς το ερώτημα. Με αυτή τη στρατηγική ένα κείμενο μπορεί να ανακτηθεί ακόμα και αν ταιριάζει κατά προσέγγιση με το ερώτημα. Επειδή δεν είναι επιθυμητή η ανάκτηση όλων των κειμένων που έχουν μη μηδενικό βαθμό σχετικότητας με το ερώτημα, αλλά αυτά που ταιριάζουν περισσότερο, ορίζεται ένα κατώφλι ελέγχου για την τιμή που λαμβάνει το μέγεθος $sim(d_j, q)$. Κείμενα με βαθμό ομοιότητας μεγαλύτερο απ' αυτό το κατώφλι επιστρέφονται στο χρήστη ως σχετικά. Πριν όμως ερμηνευτεί ο μηχανισμός κατάταξης των κειμένων πρέπει να εξεταστεί ο τρόπος υπολογισμού των βαρών.

Το πρόβλημα υπολογισμού των βαρών ανάγεται θεωρητικά στο εξής πρόβλημα ομαδοποίησης. Έστω μια συλλογή κειμένων C και ένα σύνολο A από κείμενα της συλλογής. Στο πρόβλημα της ΑΠ, το A είναι το σύνολο εκείνο των κειμένων που απαντούν σε μια πληροφοριακή ανάγκη. Η διατύπωση της πληροφοριακής ανάγκης που καθορίζει το A, μπορεί να είναι σχετικά ασαφής, οπότε τα θέματα που πρέπει να αντιμετωπιστούν είναι δυο ειδών. Πρώτον, πρέπει να καθοριστεί ποια χαρακτηριστικά προσδιορίζουν τα κείμενα του A, ενώ δεύτερον πρέπει να καθοριστεί ποια χαρακτηριστικά

διαχωρίζουν τα κείμενα του συνόλου *A* από τα κείμενα του *C*. Η εξισορρόπηση της επίδρασης αυτών των δύο ομάδων χαρακτηριστικών είναι το αντικείμενο ενός καλού σχήματος ανάθεσης βαρών.

Ένα καλό μέτρο για τον χαρακτηρισμό των στοιχείων εντός του συνόλου *A* είναι η συχνότητα εμφάνισης του όρου k_i, σε κάθε κείμενο d_j. Διαισθητικά όσο πιο συχνά εμφανίζεται ένας όρος k_i, σε ένα κείμενο d_j, τόσο πιο καλή περιγραφή του d_j αποτελεί ο όρος k_i. Η συχνότητα εμφάνισης του όρου, ονομάζεται παράγοντας *tf* (στην Αγγλική term frequency). Επίσης ένα μέτρο για τον διαχωρισμό των συνόλων *A* και *C* αποτελεί η αντίστροφη συχνότητα εμφάνισης του k_i, στα κείμενα της συλλογής. Διαισθητικά αν ο όρος k_i, έχει μεγάλη συχνότητα εμφάνισης στη συλλογή, δεν είναι πολύ χρήσιμος για να χαρακτηρίσει ένα κείμενο και άρα να διαχωρίσει μια ομάδα κειμένων μέσα στην συλλογή. Η αντίστροφη συχνότητα εμφάνισης αναφέρεται συνήθως ως παράγοντας *idf* (*Inverse Document Frequency*). Συνδυάζοντας αυτούς τους δύο παράγοντες προκύπτει το σχήμα υπολογισμού *tf-idf*, όπως ορίζεται παρακάτω.

Ανάθεση Βάρους tf-idf

Έστω *N* ο συνολικός αριθμός των κειμένων και n_i, ο αριθμός των κειμένων στα οποία εμφανίζεται ο όρος k_i. Έστω fr_{ij} η συχνότητα εμφάνισης του όρου k_i στο d_j. Τότε η κανονικοποιημένη συχνότητα f_{ij} του όρου k_i, στο d_j δίνεται από την Σχέση 10, όπου η μέγιστη τιμή *max* υπολογίζεται πάνω σε κάθε όρο που αναφέρεται στο κείμενο d_j. Αν ο όρος k_i δεν εμφανίζεται στο d_j τότε $f_{ij} = 0$. Επιπλέον, έστω idf_i η αντίστροφη συχνότητα εμφάνισης για τον όρο k_i που δίνεται από την Σχέση 11 όπου *N* είναι το σύνολο των εγγράφων της συλλογής και n_i τα έγγραφα που περιέχουν τον όρο k_i. Ο συνδυασμός των δύο αυτών μεγεθών ορίζει το σύστημα ανάθεσης βαρών *tf-idf*, σύμφωνα με την Σχέση 12. Αντίστοιχα, για τα βάρη των όρων στα ερωτήματα ισχύει η Σχέση 13 [Salton,88].

$$f_{i,j} = \frac{fr_{i,j}}{\max_l fr_{l,j}} \qquad \textbf{Σχέση 10}$$

$$idf_i = \log \frac{N}{n_i} \qquad \textbf{Σχέση 11}$$

$$w_{i,j} = f_{i,j} \cdot idf_i = f_{i,j} \times \log \frac{N}{n_i} \qquad \textbf{Σχέση 12}$$

$$w_{i,q} = \left(0.5 + \frac{0.5\, fr_{i,q}}{\max_l fr_{l,q}} \right) \times \log \frac{N}{n_i} \qquad \textbf{Σχέση 13}$$

Στην παραπάνω Σχέση 13, η $fr_{i,q}$ είναι η συχνότητα εμφάνισης του όρου k_i, στο κείμενο που αντιπροσωπεύει την πληροφοριακή ανάγκη *q*. Ο αθροιστικός παράγοντας 0.5, έχει προκύψει πειραματικά και εξισορροπεί το γεγονός ότι το ερώτημα απαρτίζεται συνήθως από πολύ λίγους όρους.

Το Χώρο-Διανυσματικό μοντέλο πλεονεκτεί διότι το σχήμα υπολογισμού των βαρών που χρησιμοποιεί, βελτιώνει την απόδοση της ανάκτησης. Επιπλέον, η στρατηγική προσεγγιστικού ταιριάσματος επιτρέπει την ανάκτηση κειμένων που προσεγγίζουν τις συνθήκες του ερωτήματος που υποβάλλει ο χρήστης. Ακόμα, ο τρόπος του υπολογισμού της κατάταξης με βάση το συνημίτονο επιτρέπει την ταξινόμηση των κειμένων βάσει του βαθμού ομοιότητας τους με την

ερώτηση, ενώ παράλληλα υλοποιείται εύκολα με τις υπάρχουσες δομές δεικτοδότησης. Ένα μειονέκτημα είναι ότι οι όροι δεικτοδότησης θεωρούνται ανεξάρτητοι μεταξύ τους.

Εν τέλει το Χώρο-Διανυσματικό μοντέλο, παρά την απλότητα της σύλληψης και της υλοποίησης του είναι ένα στιβαρό μοντέλο. Η δυνατότητα της εφαρμογής προσεγγιστικού ταιριάσματος, δίνει αποτελέσματα που είναι δύσκολο να βελτιωθούν χωρίς επέκταση του ερωτήματος ή εφαρμογή ανάδρασης του χρήστη. Τα αλγεβρικά μοντέλα που ακολούθησαν το Χώρο-Διανυσματικό μοντέλο αν και έχουν κατά σημεία καλύτερη απόδοση, είναι πιο δύσκολα στην υλοποίηση τους. Πάντως το μοντέλο αυτό δεν αντιμετωπίζει επαρκώς τα προβλήματα της Συνωνυμίας και της Πολυσημίας. Παρόλα αυτά, λόγω της ευκολίας στην υλοποίηση του, παραμένει το πιο δημοφιλές μοντέλο ΑΠ.

3.3.9 Το Πιθανοτικό Μοντέλο

Στην ενότητα αυτή παρουσιάζεται το κλασσικό πιθανοτικό μοντέλο το οποίο είναι γνωστό και ως μοντέλο ανάκτησης δυαδικής ανεξαρτησίας [Robertson,76]. Η αναφορά στο μοντέλο αυτό γίνεται με σκοπό να τονιστούν τα χαρακτηριστικά του.

Το πιθανοτικό μοντέλο επιχειρεί να αντιμετωπίσει το πρόβλημα της ΑΠ παρέχοντας ένα πιθανοτικό υπόβαθρο. Δοθέντος ενός ερωτήματος χρήστη, υπάρχει ένα σύνολο κειμένων που αποτελείται ακριβώς από τα σχετικά κείμενα και μόνο απ' αυτά. Το σύνολο αυτό αναφέρεται με τον όρο ιδανικό σύνολο απάντησης. Δοθείσης της περιγραφής του ιδανικού συνόλου απάντησης, δεν υπάρχει πρόβλημα στην ανάκτηση των κειμένων που το αποτελούν. Συνεπώς μπορεί να θεωρηθεί ότι η διατύπωση ενός ερωτήματος ταυτίζεται με τη διαδικασία καθορισμού των ιδιοτήτων του ιδανικού συνόλου απάντησης. Το πρόβλημα συνίσταται στο ότι δεν είναι γνωστές αυτές οι ιδιότητες, διότι το μόνο που είναι γνωστό είναι μια ομάδα από όρους δεικτοδότησης, η σημασιολογία των οποίων μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει αυτές τις ιδιότητες. Αυτές δεν είναι γνωστές τη στιγμή της διατύπωσης του ερωτήματος, οπότε πρέπει να γίνει μια αρχική προσπάθεια να προσδιοριστούν. Η αρχική αυτή εκτίμηση επιτρέπει τη δημιουργία μιας αρχικής πιθανοτικής περιγραφής του ιδανικού συνόλου απάντησης, η οποία θα χρησιμοποιηθεί για την ανάκτηση ενός πρώτου συνόλου κειμένων. Ακολουθεί αλληλεπίδραση με το χρήστη, με σκοπό τη βελτίωση της περιγραφής του ιδανικού συνόλου απάντησης.

Ο χρήστης εξετάζει το αρχικό σύνολο των επιστρεφόμενων κειμένων και αποφασίζει ποια κείμενα είναι σχετικά και ποια όχι. Κατόπιν το σύστημα αξιοποιεί αυτή την πληροφορία για να βελτιώσει την περιγραφή του συνόλου απάντησης. Επαναλαμβάνοντας αυτή τη διαδικασία αρκετές φορές, αναμένεται ότι η περιγραφή αυτή θα συγκλίνει προς την ιδανική περιγραφή του συνόλου απάντησης. Έτσι πάντα θα πρέπει να λαμβάνεται υπόψη η αρχική περιγραφή του ιδανικού συνόλου απάντησης. Επιπλέον πρέπει να γίνει προσπάθεια να περιγραφεί η παραπάνω διαδικασία πιθανοτικά. Το πιθανοτικό μοντέλο βασίζεται στην ακόλουθη θεμελιώδη υπόθεση.

Δοθέντος ενός ερωτήματος q και ενός κειμένου d_j της συλλογής, το πιθανοτικό μοντέλο προσπαθεί να εκτιμήσει την πιθανότητα ο χρήστης να βρει σχετικό το κείμενο d_j ως προς το ερώτημα q. Υπόθεση του μοντέλου είναι ότι η πιθανότητα της σχετικότητας εξαρτάται μόνο από την αναπαράσταση του ερωτήματος και του κειμένου. Επιπλέον γίνεται η υπόθεση ότι υπάρχει ένα υποσύνολο όλων των κειμένων, το οποίο ο χρήστης προτιμά ως απάντηση στο ερώτημα q. Ένα τέτοιο ιδανικό σύνολο απάντησης, ονομάζεται R και θα πρέπει να μεγιστοποιεί τη συνολική πιθανότητα σχετικότητας προς την πληροφοριακή ανάγκη του χρήστη. Τα κείμενα στο R προβλέπεται ότι είναι σχετικά προς το ερώτημα. Τα κείμενα που δεν ανήκουν σ' αυτό το σύνολο προβλέπεται ότι είναι μη-σχετικά. Μια τέτοια υπόθεση είναι κάπως προβληματική γιατί δεν παρέχει ένα μηχανισμό για τον υπολογισμό των πιθανοτήτων σχετικότητας. Επιπλέον ούτε καν προκύπτει ο δειγματικός χώρος για τον υπολογισμό αυτών των πιθανοτήτων.

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

Ανάθεση Βαρών Δεικτοδότησης

Στο πιθανοτικό μοντέλο όλα τα βάρη των όρων δεικτοδότησης έχουν δυαδική μορφή της μορφής $w_{ij} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$. Ένα ερώτημα *q* είναι ένα υποσύνολο των όρων δεικτοδότησης. Έστω *R* το σύνολο των κειμένων για το οποία υπάρχει η γνώση ή αρχικά η εκτίμηση ότι είναι σχετικά. Έστω $\overline{R}$ το συμπλήρωμα του *R* που αντιπροσωπεύει το σύνολο των μη σχετικών κειμένων. Έστω $P(R \mid d_j)$ η πιθανότητα το κείμενο d_j να είναι σχετικό προς το ερώτημα *q* και $P(\overline{R} \mid d_j)$ η πιθανότητα το κείμενο d_j να μην είναι σχετικό προς το ερώτημα *q*. Η ομοιότητα $sim(d_j, q)$ του κειμένου d_j προς το ερώτημα *q* ορίζεται ως ο λόγος της παρακάτω Σχέσης 14.

$$sim(d_j, q) = \frac{P(R \mid d_j)}{P(\overline{R} \mid d_j)} \qquad \textbf{Σχέση 14}$$

Λαμβάνοντας υπόψη τον κανόνα του Bayes η Σχέση 14 μετασχηματίζεται και για την τιμή ομοιότητας ισχύει $sim(d_j, q) = [P(d_j \setminus R) \times P(R)] \cdot [P(d_j \setminus \overline{R}) \times P(\overline{R})]^{-1}$, όπου $P(d_j \mid R)$ είναι η πιθανότητα το d_j να επιλέχθηκε τυχαία από το σύνολο *R,* δηλαδή να είναι σχετικό. Επιπλέον *P(R)* είναι η πιθανότητα το κείμενο που επιλέξαμε με τυχαίο τρόπο από ολόκληρη τη συλλογή, να είναι τυχαίο. Οι ερμηνείες των ποσοτήτων $P(d_j \mid \overline{R})$ *και* $P(\overline{R})$ είναι δυικές των παραπάνω.

Μια και τα *P(R)* και $P(\overline{R})$ είναι τα ίδια για όλα τα κείμενα της συλλογής, η Σχέση 14 μετασχηματίζεται στην Σχέση 15.

$$sim(d_j, q) \approx \frac{P(d_j \mid R)}{P(d_j \mid \overline{R})} \qquad \textbf{Σχέση 15}$$

Εφόσον υπάρχει στοχαστική ανεξαρτησία οι όροι της παραπάνω σχέσης μπορούν να γραφούν όπως ορίζεται στην Σχέση 16, όπου $P(k_i \mid R)$ είναι η πιθανότητα ο όρος δεικτοδότησης k_i, να εμφανίζεται σε ένα κείμενο το οποίο επιλέχθηκε τυχαία από το σύνολο *R*. Ο όρος $P(\overline{k_i} \mid R)$ δίνει την πιθανότητα ο όρος k_i, να μην εμφανίζεται σε ένα κείμενο το οποίο επιλέχθηκε τυχαία από το σύνολο *R*. Οι πιθανότητες που σχετίζονται με το σύνολο $\overline{R}$ έχουν ανάλογη σημασία.

Λαμβάνοντας υπόψη ότι $P(k_i \mid R) + P(\overline{k_i} \mid R) = 1$ και αγνοώντας τους παράγοντες που είναι σταθεροί για όλα τα κείμενα για ένα συγκεκριμένο ερώτημα, η αρχική Σχέση 14 μετασχηματίζεται στην Σχέση 16 με την οποία τελικά υπολογίζεται η κατάταξη των κειμένων στο πιθανοτικό μοντέλο.

$$sim(d_j, q) \approx \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})} \right) \qquad \textbf{Σχέση 16}$$

Λόγω του ότι το σύνολο *R* δεν είναι εξ' αρχής γνωστό, κρίνεται απαραίτητο να οριστεί μια μέθοδος υπολογισμού για τις πιθανότητες $P(k_i \mid R)$ και $P(\overline{k_i} \mid R)$. Η μέθοδος αυτή θα οριστεί παρακάτω. Αμέσως μετά την διατύπωση του ερωτήματος από το χρήστη, δεν υπάρχουν ακόμα

ανακτημένα κείμενα. Έτσι πρέπει να γίνουν υποθέσεις απλοποίησης σε ότι αφορά τις πιθανότητες. Πρώτον, υποτίθεται ότι η $P(k_i \mid R)$ είναι σταθερή για όλους τους όρους k_i και ίση με 0.5 και δεύτερον υποτίθεται ότι η κατανομή των όρων δεικτοδότησης στα μη-σχετικά κείμενα μπορεί να προσεγγιστεί από την κατανομή των όρων δεικτοδότησης στο σύνολο των κειμένων. Με άλλα λόγια, το μέγεθος του συνόλου των μη σχετικών κειμένων $\overline{R}$, είναι πολύ μεγαλύτερο από το μέγεθος *R*. Ισχύει δηλαδή ότι $P(k_i \mid R) = 0.5$ και $P(k_i \mid \overline{R}) = n_i / N$ όπου, όπως ορίστηκε προηγουμένως n_i, είναι ο αριθμός των κειμένων που περιέχουν τον όρο k_i, και *N* είναι ο συνολικός αριθμός των κειμένων της συλλογής. Έχοντας την αρχική εκτίμηση, μπορεί να ανακτηθεί ένα αρχικό σύνολο κειμένων που περιέχουν όρους που εμφανίζονται στο ερώτημα και να δοθεί μια πιθανοτική κατάταξη γι' αυτά. Κατόπιν ξεκινάει μια διαδικασία βελτίωσης της αρχικής κατάταξης όπως περιγράφεται παρακάτω.

Έστω λοιπόν *V* ένα υποσύνολο των κειμένων που ανακτήθηκαν αρχικά και στα οποία δόθηκε μια κατάταξη από το πιθανοτικό μοντέλο. Το παραπάνω σύνολο θα μπορούσε για παράδειγμα να είναι τα *r* το πλήθος κορυφαία κείμενα, όπου ο δείκτης ομοιότητας τους υπερβαίνει ένα προκαθορισμένο κατώφλι. Έστω επίσης V_i ένα υποσύνολο του *V* το οποίο αποτελείται από τα κείμενα που περιέχουν τον όρο k_i. Για λόγους απλότητας, χρησιμοποιούνται οι όροι *V* και V_i, για να αντιπροσωπεύουν τους πληθυσμούς των αντιστοίχων συνόλων. Για να βελτιωθεί η πιθανοτική κατάταξη, πρέπει να βελτιωθούν οι εκτιμήσεις για τα $P(k_i \mid R)$ και $P(k_i \mid \overline{R})$. Αυτό επιτυγχάνεται αν υποτεθεί ότι μπορεί να προσεγγιστεί η $P(k_i \mid R)$ με την κατανομή του όρου k_i στα κείμενα που ανακτήθηκαν ενώ ταυτόχρονα μπορεί να προσεγγιστεί η τιμή της $P(k_i \mid \overline{R})$ αν θεωρηθούν όλα τα μη ανακτημένα κείμενα ως μη-σχετικά. Έτσι θα ισχύουν οι Σχέσεις 17 και 18.

$$P(k_i \mid R) = \frac{V_i}{V} \qquad \textbf{Σχέση 17}$$

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i}{N - V} \qquad \textbf{Σχέση 18}$$

Αυτή η διαδικασία μπορεί να επαναληφθεί αναδρομικά, υπολογίζοντας κάθε φορά νέα *V* και V_i. Έτσι μπορούν να βελτιωθούν οι εκτιμήσεις για τις τιμές των $P(k_i \mid R)$ και $P(k_i \mid \overline{R})$ χωρίς καμία ανάμιξη του ανθρωπίνου παράγοντα, η οποία ενδέχεται να ζητηθεί στην κατασκευή του συνόλου *V*.

Οι τελευταίοι δυο τύποι για τα $P(k_i \mid R)$ και $P(k_i \mid \overline{R})$ εμφανίζουν προβλήματα για μικρές τιμές των *V* και V_i που εμφανίζονται στην πράξη, όπως για παράδειγμα εάν ισχύει $V = 1$ και $V_i = 0$. Για να αντιμετωπιστούν αυτά τα προβλήματα, χρειάζεται να εισαχθεί ένας προσθετικός παράγοντας οπότε οι παραπάνω Σχέσεις 17 και 18 μετασχηματίζονται στις Σχέσεις 19 και 20.

$$P(k_i \mid R) = \frac{V_i + 0.5}{V + 1} \qquad \textbf{Σχέση 19}$$

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i + 0.5}{N - V + 1} \qquad \textbf{Σχέση 20}$$

Συχνά όμως ένας σταθερός προσθετικός παράγοντας, όπως το 0.5, δεν είναι επαρκής. Μια εναλλακτική λύση είναι να θεωρηθεί ως προσθετικός παράγοντας η ποσότητα n_i/N, οπότε οι παραπάνω σχέσεις μετασχηματίζονται στις Σχέσεις 21 και 22.

$$P(k_i \mid R) = \frac{V_i + \frac{n_i}{N}}{V + 1}$$ **Σχέση 21**

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i + \frac{n_i}{N}}{N - V + 1}$$ **Σχέση 22**

Το κύριο πλεονέκτημα της χρήσης του πιθανοτικού μοντέλου είναι ότι τα κείμενα κατατάσσονται σε φθίνουσα σειρά βάσει της πιθανότητα να είναι σχετικά με το αρχικό ερώτημα του χρήστη. Τα μειονεκτήματα συνίστανται στο ότι χρειάζεται μια αρχική εκτίμηση για τον διαχωρισμό της συλλογής των κειμένων σε σχετικά και μη-σχετικά, ενώ παράλληλα δε λαμβάνεται υπόψη η συχνότητα εμφάνισης του όρου μέσα σε ένα κείμενο. Τέλος η παραδοχή ότι οι όροι είναι μεταξύ τους ανεξάρτητοι δεν μπορεί να περιληφθεί στα θετικά που υποστηρίζει το θεωρητικό υπόβαθρο του μοντέλου αυτού.

4 ΤΕΧΝΙΚΕΣ ΑΝΑΠΑΡΑΣΤΑΣΗΣ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΓΓΡΑΦΩΝ

4.1 Αναπαράσταση Κειμένου

Η εργασία της ταξινόμησης κειμένου αποσκοπεί στην κατηγοριοποίηση των εγγράφων σε μια ή περισσότερες κλάσεις/κατηγορίες. Μια κλάση θεωρείται ως μια σημασιολογική κατηγορία που ομαδοποιεί έγγραφα, τα οποία έχουν κάποιες συγκεκριμένες κοινές ιδιότητες. Γενικά ένα έγγραφο μπορεί να ανήκει σε πολλές, σε μία ή καμία κλάση. Παρόλα αυτά, οι διαδικασίες της επεξεργασίας πληροφορίας στοχεύουν συνήθως στην ταξινόμηση ενός εγγράφου σχετικά με το αν είναι ή όχι όσον αφορά μια κατηγορία πληροφορίας [Μακρής,02], [Anagnostopoulos,04]. Το πρόβλημα τίθεται ως εξής.

Ας υποτεθεί ένα σύνολο εγγράφων *D* και ένα σύνολο από *k* το σύνολο κλάσεις $C = \{c_1, c_2, ..., c_k\}$, το οποίο είναι ένα υποσύνολο του *D*. Η ταξινόμηση κειμένου είναι η χαρτογράφηση $h : D \mapsto C$ από το σύνολο των εγγράφων στο σύνολο των κλάσεων.

Με την σημερινή τεράστια ανάπτυξη των διαθέσιμων πληροφοριών μέσω του Διαδικτύου, το πρόβλημα της αυτόματης ταξινόμησης εγγράφων, βάσει κειμένου σε προκαθορισμένες κλάσεις, είναι σπουδαίας σημασίας για ζητήματα ευφυούς αναζήτησης και οργάνωσης της πληροφορίας.

4.2 Βήματα Μετατροπής σε Διανυσματική Μορφή

Ο σκοπός της διαδικασίας αναπαράστασης κειμένου είναι να μετασχηματίσει ένα έγγραφο κειμένου σε μια κατάλληλη μορφή για περαιτέρω επεξεργασία από αλγόριθμους επεξεργασίας της πληροφορίας [Lewis,92]. Η συνήθης πρακτική είναι η μετατροπή του εγγράφου σε ένα ιδιο-χαρακτηριστικό διάνυσμα (προφίλ) σε ένα υποθετικό χώρο όπου κάθε διάσταση αντιστοιχεί και σε ένα ξεχωριστό συνταγμένο όρο.

Γενικά, το σύνολο των συνταγμένων όρων μπορεί να οριστεί είτε με τη βοήθεια ενός ειδικού συντάκτη είτε αυτόματα από ορισμένες διαδικασίες πάνω σε μια ορισμένη συλλογή εγγράφων.

Ένα δεδομένο διάνυσμα έχει σε κάθε συστατικό του, μια αριθμητική τιμή που προσδιορίζει το πόσο σημαντικό είναι. Αυτή η τιμή συνήθως ορίζεται ως μια αντιστοίχηση της συχνότητας εμφάνισης του συγκεκριμένου όρου σε ένα εξεταζόμενο έγγραφο και πόσο συχνά εμφανίζεται σε όλη την συλλογή των εγγράφων. Η παραλλαγή της συνάρτησης στάθμισης των συχνοτήτων αυτών οδηγεί σε διαφορετικά μοντέλα επεξεργασίας. Το αποτέλεσμα αυτής της αναπαράστασης του κειμένου είναι μια αντιστοίχηση μεταξύ χαρακτηριστικών του εγγράφου και ανατιθέμενων βαρών [Joachims,97].

Η αναπαράσταση στον διανυσματικό χώρο λαμβάνει πληροφορίες από την ακολουθία με την οποία οι συνταγμένοι όροι εμφανίζονται στο εκάστοτε έγγραφο.

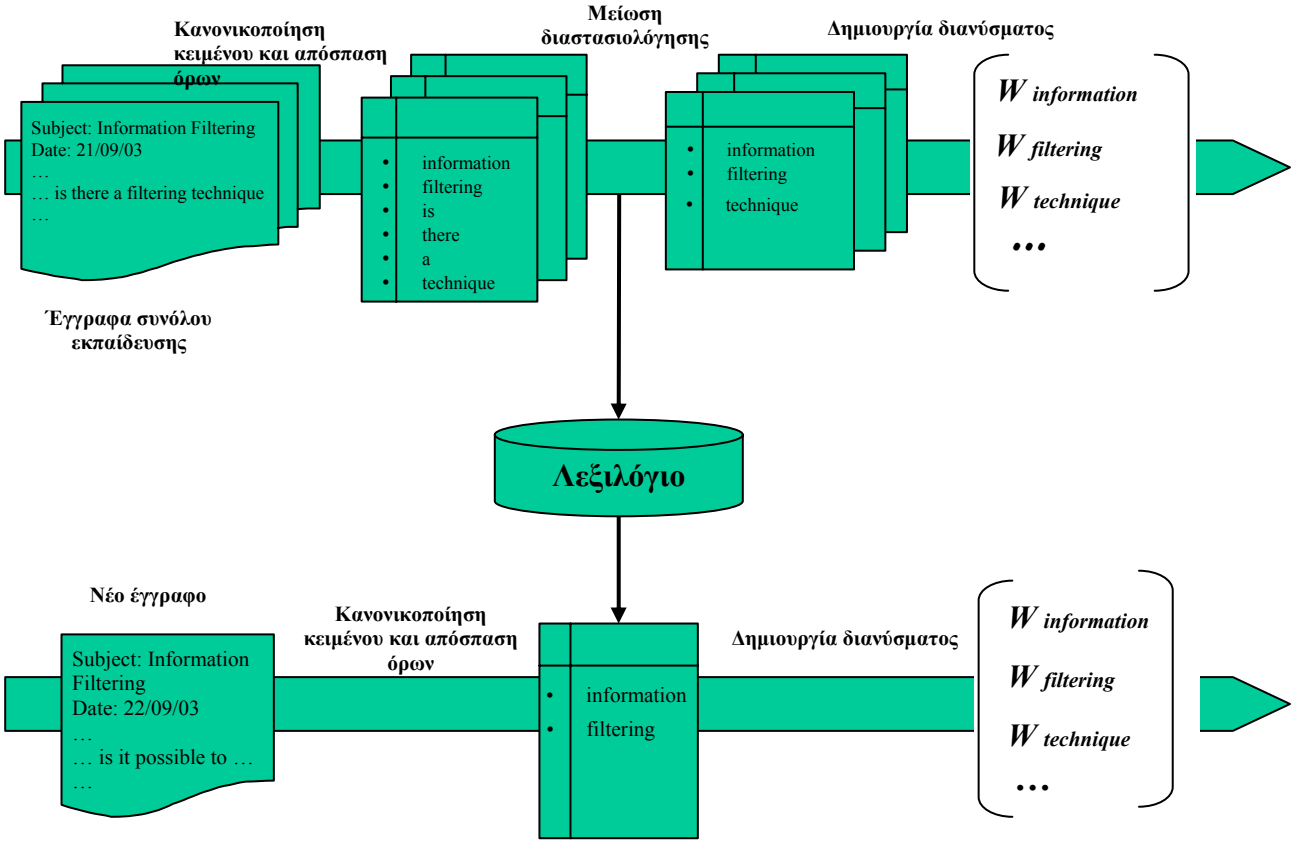
Το Σχήμα 6 παρουσιάζει τα βήματα που λαμβάνουν μέρος στο μετασχηματισμό των εγγράφων του δείγματος εκπαίδευσης και ενός νέου εξεταζόμενου εγγράφου, σε ιδιο-χαρακτηριστικά διανύσματα, όπου όροι ανεξάρτητοι από τη διάταξη της εμφάνισή τους στο κείμενο, χρησιμοποιούνται ως όροι σύνταξης. Για το δείγμα εκπαίδευσης τα βήματα είναι τα εξής:

1. Κανονικοποίηση κειμένου,

2. Απόσπαση όρων,

3. Μείωση διαστασιολόγησης,

4. Δημιουργία διανύσματος περιγραφέα.



Σχήμα 6. Βήματα μετασχηματισμού εγγράφων σε διανύσματα

Το βήμα της κανονικοποίησης του κειμένου μετασχηματίζει οποιοδήποτε τύπο εγγράφου σε μια ακολουθία από λέξεις-πιστοποιητικά. Το κατά πόσο η έξοδος από το βήμα απόσπασης λέξεων χρησιμοποιείται, εξαρτάται από το εάν νέα εξεταζόμενα έγγραφα ή έγγραφα εκπαίδευσης υπόκεινται σε επεξεργασία ή όχι. Σημειώνεται δε, ότι στην περίπτωση αυτή κρίνεται απαραίτητη η παρουσία λεξιλογίου. Έτσι, όλοι οι ξεχωριστοί όροι του δείγματος εκπαίδευσης συλλέγονται κατάλληλα με σκοπό να δημιουργήσουν ένα σύνολο από υποψήφιους όρους, που μπορούν πιθανώς να χρησιμοποιηθούν ως λεξιλόγιο. Αυτή η τεχνική με την οποία αποφεύγεται η χρήση εκτεταμένων λεξιλογίων και εννοιολογικών θησαυρών ονομάζεται δημιουργία ιδιο-χαρακτηριστικών. Συνήθως όμως το σύνολο των συνταγμένων όρων είναι πολύ μεγάλο, με αποτέλεσμα να κρίνεται επιτακτική η ανάγκη τεχνικών μείωσης του ποσού αυτού. Αυτό επιτυγχάνεται με την εφαρμογή της μείωσης της διαστασιολόγησης. Στην περίπτωση βέβαια που χρησιμοποιείται ήδη ένα σταθερό λεξιλόγιο, το βήμα απόσπασης όρων έχει ως αποτέλεσμα την συλλογή όρων που βρίσκονται στο λεξιλόγιο αυτό. Τέλος, το βήμα δημιουργίας του διανύσματος περιγραφέα αφορά την ανάθεση των σχετικών βαρών για όλους τους συνταγμένους όρους, ενός εξεταζόμενου εγγράφου. Αυτά τα τέσσερα βήματα και οι μεταξύ τους συσχετίσεις αναλύονται στις επόμενες παραγράφους.

4.2.1 Κανονικοποίηση Κειμένου

Γενικά, η είσοδος σε ένα σύστημα επεξεργασίας πληροφορίας θεωρείται ένα αρχείο το οποίο αναπαριστά κείμενο. Ο ρόλος της κανονικοποίησης κειμένου είναι ο μετασχηματισμός των

αρχείων εισόδου σε μια ακολουθία γλωσσικών στοιχείων, που ονομάζονται λέξεις-πιστοποιητικά. Στο επόμενο βήμα της απόσπασης όρων, αυτά τα στοιχεία θα χρησιμοποιηθούν με σκοπό την δημιουργία χρήσιμων χαρακτηριστικών που ονομάζονται συνταγμένοι όροι. Υπάρχουν δύο βήματα στην διαδικασία κανονικοποίησης κειμένου.

Αρχικά, κειμενικά χαρακτηριστικά από διαφορετικούς τύπους εγγράφων πρέπει να αναγνωριστούν. Ο προσδιορισμός του τύπου του αρχείου (έγγραφα MS-Word-PDF-postscript, ηλεκτρονικό ταχυδρομείο, Tex-LaTeX, , ιστοσελίδες) πρέπει να είναι γνωστός για να δημιουργηθεί μια ενοποιημένη διαδικασία δημιουργίας εισόδου. Για παράδειγμα, κατά την επεξεργασία εγγράφων του ηλεκτρονικού ταχυδρομείου, συνήθως λαμβάνεται υπόψη το πεδίο του θέματος και το κυρίως κείμενο, ενώ στις ιστοσελίδες η επεξεργασία του κυρίως κειμένου απαιτεί την αφαίρεση των ετικετών της Γλώσσας Υπερκείμενης Σήμανσης (HyperText Markup Language).

Στην συνέχεια, στο δεύτερο βήμα, η ακολουθία των λέξεων-πιστοποιητικών κανονικοποιείται ανάλογα με την εφαρμογή [Fox,92]. Παραδείγματος χάριν, όλοι οι χαρακτήρες μετασχηματίζονται σε πεζούς ενώ αφαιρούνται οι αριθμοί και τα σημεία στίξης.

Σύμφωνα με μια άλλη προσέγγιση, οντότητες όπως ονόματα ανθρώπων, περιοχές, οργανισμοί και προϊόντα μπορούν να αναγνωριστούν και να μετασχηματιστούν σε ανεξάρτητες λέξεις-πιστοποιητικά, μειώνοντας το κόστος επεξεργασίας και αυξάνοντας παράλληλα τον βαθμό συμπίεσης των δεδομένων. Για παράδειγμα, η ακολουθία "C 180" μπορεί να θεωρηθεί και να αναγνωρίζεται ως τύπος προϊόντος μιας συγκεκριμένης αυτοκινητοβιομηχανίας.

4.2.2 Βήμα Απόσπασης Όρων

Ας υποτεθεί ότι μια ακολουθία κανονικοποιημένων λέξεων-πιστοποιητικών παρέχεται από το προηγούμενο βήμα. Ο σκοπός του βήματος απόσπασης όρων είναι να παράγει μια ακολουθία συνταγμένων όρων βάσει αυτών. Ο τρόπος χρήσης αυτής της ακολουθίας εξαρτάται από το εάν τα έγγραφα του συνόλου δοκιμής επεξεργάζονται και είναι απαραίτητη η δημιουργία του λεξιλογίου. Εάν όμως το λεξιλόγιο έχει ήδη δημιουργηθεί, οποιοσδήποτε συνταγμένος όρος που δεν ανήκει στο λεξιλόγιο παραβλέπεται κατά το βήμα της απόσπασης όρων. Στην συγκεκριμένη περίπτωση για κάθε εξεταζόμενο έγγραφο d_j, δημιουργείται το διάνυσμα συχνότητας όρων $d_{tf} = (tf_d(t_1), tf_d(t_2), ..., tf_d(t_m))^T$, όπου ο όρος $tf_d(t)$ προσδιορίζει τον αριθμό εμφάνισης των όρων $t \in V$ στο έγγραφο d, όπου $V = \{t_1, t_2, ...t_m\}$ το σύνολο των ξεχωριστών συνταγμένων όρων. Έχοντας ένα διάνυσμα αυτής της μορφής είναι εύκολη η μετάβαση στα δύο επόμενα βήματα που αποσκοπούν στην μείωση της διαστασιολόγησης και της δημιουργίας του διανύσματος περιγραφέα.

Εξάλλου όταν δεν έχει δημιουργηθεί ένα λεξιλόγιο, οι έξοδοι του βήματος αυτού για όλα τα έγγραφα του συνόλου δοκιμής συνιστούν ένα σύνολο όρων $V^* = \{t_1^*, t_2^*, ...t_m^*\}$, το οποίο ενδέχεται να χρησιμοποιηθεί ως λεξιλόγιο. Ανάλογα με το μέγεθος και το περιεχόμενο του συνόλου εκπαίδευσης, το μέγεθος των συλλεγμένων όρων κυμαίνεται από 10^3 έως 10^4. Συνήθως όμως ένα σύνολο τόσων όρων μεγαλώνει κατά πολύ το χρόνο επεξεργασίας και παρουσίασης των αποτελεσμάτων. Σε αυτό το σημείο παρεμβάλλεται το βήμα της μείωσης της διαστασιολόγησης ή με άλλα λόγια της μείωσης του ποσού των επιλεγμένων και συνταγμένων όρων.

Πριν όμως αναλυθεί το επόμενο βήμα, πρέπει να αναφερθεί ότι στην υπάρχουσα βιβλιογραφία αναφέρονται πέντε διαφορετικές προσεγγίσεις για την γλωσσική ανάλυση και επεξεργασία σε μια συλλογή εγγράφων, όπως αναφέρονται παρακάτω.

1η προσέγγιση: Γραφική ανάλυση: Πραγματοποιείται μια ανάλυση βάσει τμημάτων λέξεων και αφορά συνήθως χαρακτήρες λέξεων.

2η προσέγγιση: Λεξική ανάλυση: Βασίζεται στην ανάλυση ξεχωριστών λέξεων και ριζών-λέξεων.

3η προσέγγιση: Συντακτική ανάλυση: Η προτεινόμενη ανάλυση πραγματοποιείται βάσει της δομής των φράσεων.

4η προσέγγιση: Σημασιολογική ανάλυση: Σχετίζεται με το νόημα και τις αποδιδόμενες έννοιες των λέξεων και των φράσεων.

5η προσέγγιση: Πραγματική ανάλυση: Γλωσσική ανάλυση που συνδέεται με την απόδοση του όρου ανάλογα με την υπάρχουσα εφαρμογή.

Οι πρώτες δύο προσεγγίσεις λειτουργούν βάσει στατιστικών στοιχείων όσον αφορά το κείμενο, όπως συχνότητες συνδυασμού χαρακτήρων ή λέξεων, τα οποία ορίζονται ως συχνότητες όρων. Βέβαια, η αναπαράσταση κειμένου που βασίζεται σε αυτές τις συχνότητες δεν μπορεί να αποτυπώσει και να αποδώσει το νόημα των εγγράφων. Πιο συγκεκριμένα, υπάρχει μόνο μια σχέση εξάρτησης μεταξύ της συχνότητας εμφάνισης των λέξεων και του εξεταζόμενου περιεχομένου [Blair,92]. Σε αντιδιαστολή, άλλες μέθοδοι που αποσκοπούν σε μια ανάλυση υψηλότερου επιπέδου, προσπαθούν να αποτυπώσουν σε μεγαλύτερο βαθμό το σημασιολογικό περιεχόμενο, ερευνώντας ένα ποσό ανταλλαγής πληροφοριών, όπως είναι η δομή των προτάσεων, των παραγράφων ή των εγγράφων. Είναι άλλωστε γενικά παραδεκτό, ότι σε προβλήματα που αφορούν την επεξεργασία φυσικής γλώσσας, η συντακτική και σημασιολογική προσέγγιση παρέχουν μια πιο ολοκληρωμένη αναπαράσταση γνώσης.

Διαισθητικά, φαίνεται λογικό να υποτεθεί ότι οι πιο σύνθετες αντιπροσωπεύσεις κειμένων όπως λαμβάνονται από τα πιο υψηλά επίπεδα ανάλυσης κειμένων θα οδηγούσαν σε αποτελεσματικότερους ταξινομητές κειμένων [Blair,92]. Αυτό μπορεί πραγματικά να ισχύει, έχοντας ένα άπειρο ποσό κειμένου που αποκαλύπτει τις πληροφορίες για τους όρους δεικτών που εξάγονται. Στα πραγματικά προβλήματα, όμως, ο αριθμός των διαθέσιμων εγγράφων για την ανάλυση κειμένων είναι περιορισμένος. Επιπλέον, καθώς πιο σύνθετοι ορισμοί όρων οδηγούν σε πιο σύνθετες αντιπροσωπεύσεις κειμένων, η διάσταση των χαρακτηριστικών γνωρισμάτων στο χώρο αυξάνεται αντίστοιχα. Έχοντας παράλληλα και έναν περιορισμένο αριθμό εγγράφων εκπαίδευσης, η δημιουργία ενός ακριβούς ταξινομητή είναι μια πολύ δύσκολη υπόθεση. Το πρόβλημα αυτό, της ύπαρξης δηλαδή πολλών χαρακτηριστικών γνωρισμάτων σχετικά με τον αριθμό στοιχείων εκπαίδευσης αναφέρεται συνήθως ως η "κατάρα της διαστασιολόγησης".

Εντούτοις, αυτή η εύρεση στηρίζεται στα αποτελέσματα που επιτυγχάνονται κυρίως με έγγραφα στην Αγγλική. Για άλλες γλώσσες, βαθύτερη γλωσσική ανάλυση μπορεί είναι πιο χρήσιμη στην επεξεργασία της πληροφορίας. Στα Γερμανικά, παραδείγματος χάριν, υπάρχουν πολλές σύνθετες λέξεις, γεγονός το οποίο μπορεί να προκαλέσει μια τεράστια αύξηση στον αριθμό χαρακτηριστικών γνωρισμάτων. Έχει ερευνηθεί ότι η ωφέλιμη πληροφορία που μπορεί να ανακτηθεί από τις διαδικασίες της γλωσσικής ανάλυσης εξαρτάται άμεσα από την γλώσσα που υφίσταται την επεξεργασία.

Από τα ανωτέρω φαίνεται ότι η επιλογή ενός κατάλληλου επιπέδου ανάλυσης κειμένων, στο οποίο βασίζεται ο καθορισμός των όρων, αποτελεί μια διαδικασία ανταλλαγής μεταξύ της σημασιολογικής εκφραστικότητας και της αντιπροσωπευτικής πολυπλοκότητας. Αυτή η επιλογή θα ασκήσει μια ευρεία επίδραση πάνω στη δυνατότητα γενίκευσης ενός αλγορίθμου εκμάθησης. Απλοί ορισμοί όρου είναι κυρίαρχοι στην περιοχή της ανάκτησης πληροφοριών [Apte,94]. Χαρακτηριστικά, αυτοί περιλαμβάνουν τις προσεγγίσεις κειμενικής ανάλυσης από τα δύο πρώτα επίπεδα, ενώ παράλληλα κατά περιόδους, κάποιες συντακτικές πληροφορίες μπορούν επίσης να ενσωματωθούν.

Παρακάτω, περιγράφονται τρεις ορισμοί όρου που χρησιμοποιούνται ευρέως, οι οποίοι εκμεταλλεύονται τα σαφή στατιστικά γεγονότα κατά την ανάλυση του κειμένου. Η πρώτη μέθοδος εξαρτάται από τους χαρακτήρες n-grams, που όπως έχει προαναφερθεί αντιστοιχεί στην προσέγγιση γραφικής ανάλυσης. Οι υπόλοιπες δύο μέθοδοι χρησιμοποιούν μεμονωμένες λέξεις και φράσεις και πληροφορίες που πηγάζουν από την συντακτική ανάλυση.

χαρακτήρες n-grams: Επικαλύπτοντας, τις παρακείμενες ακολουθίες n χαρακτήρων των λέξεων, προκύπτει η μέθοδος η οποία ονομάζεται και ως n-grams, όπου το n είναι ένας θετικός ακέραιος αριθμός [de Heer,82], [Cavnar,94]. Σημειωτέον, ότι μερικές φορές ο όρος "*n-grams*" χρησιμοποιείται για να αναφερθεί στις ακολουθίες λέξεων μήκους *n* [[Furnkranz,98]. Παρακάτω, θα γίνει αναφορά σε αυτούς τους όρους πολύ-λέξεων ως φράσεις. Στην ανάλυση αυτή, όταν το *n* ισούται με 3 ή 4 οι εξεταζόμενοι όροι ονομάζονται "*trigrams*" ή "*quadgrams*", αντίστοιχα, αλλά και άλλες τιμές είναι επίσης δυνατές. Παραδείγματος χάριν, η λέξη "monitor" αποτελείται από τα *trigrams* "mon", "oni", "nit", "ito" και "tor". Είναι επίσης δυνατό να περιληφθεί ένα διάστημα στην αρχή και στο τέλος μιας λέξεως κατά την παραγωγή των *n-grams*. Στην περίπτωση της προαναφερθείσας λέξεως, θα λαμβάνονταν επιπρόσθετα τα *trigrams* "_mo" και "or_". Λαμβάνοντας υπόψη τους 26 χαρακτήρες του Αγγλικού αλφάβητου, υπάρχουν 26^3 =17576 ευδιάκριτα *trigrams* και 26^4 =456976 *quadgrams*. Όσο μεγαλύτερη είναι η τιμή του *n*, τόσο πιο ακριβέστερα ορίζονται οι αποστάσεις μεταξύ των *n-gram* διανυσμάτων που αντιστοιχούν με τις σημασιολογικές αποστάσεις από τα εξεταζόμενα έγγραφα [Tauritz,00]. Όμως, υψηλότερες τιμές του *n* οδηγούν σε έναν τεράστιο αριθμό πιθανών όρων, ενώ εάν ισχύει $n<3$, δεν εμφανίζεται ικανοποιητικό ποσό συντακτικής πληροφορίας [Teufel,88]. Το πλεονέκτημα των *n-grams* είναι ότι το σύνολο πιθανών όρων καθορίζεται και είναι γνωστό εκ των προτέρων. Επιπλέον, η μέθοδος των *n-grams* είναι ανεξάρτητη της εξεταζόμενης γλώσσας και αρκετά εύρωστη στις μορφολογικές παραλλαγές και στα λάθη ορθογραφίας. Επιπλέον, τα *n-grams* είναι εύκολο να υπολογιστούν, αλλά η προκύπτουσα αντιπροσώπευση είναι δύσκολο να αναλυθεί από τους ανθρώπους.

Λέξεις: Οι τελευταίες έρευνες στον τομέα ανάκτησης πληροφοριών δείχνουν ότι οι μεμονωμένες λέξεις λειτουργούν αρκετά ικανοποιητικά ως ιδιο-χαρακτηριστικά [Salton,88], [Dumais,98], [Cohen,96]. Έτσι, μια πολύ κοινή προσέγγιση είναι η χρήση κάθε λέξεως, όπως είναι. Προφανώς, αυτή η προσέγγιση στον καθορισμό όρου είναι ανεξάρτητη της γλώσσας και υπολογιστικά πιο αποδοτική. Εντούτοις, ένα μειονέκτημα είναι ότι κάθε παραλλαγή μιας λέξης είναι ένα πιθανό χαρακτηριστικό γνώρισμα ενώ παράλληλα ο αριθμός πιθανών χαρακτηριστικών γνωρισμάτων μπορεί να μην είναι απαραίτητα μεγάλος. Επιπλέον, οι μορφολογικές παραλλαγές μιας λέξης δεν αναγνωρίζονται πάντα ως παρόμοιες. Μια αντιμετώπιση του περιορισμού αυτού είναι να μειωθεί κάθε λέξη στην αντίστοιχη ρίζα της προκειμένου να συγχωνευθούν οι μορφολογικές παραλλαγές της.

Φράσεις: Οι συνδυασμοί των λέξεων-πιστοποιητικών αναφέρονται ως φράσεις. Η χρησιμοποίηση των φράσεων δικαιολογείται από την παρατήρηση ότι, ειδικά στα αγγλικά, πολλές εκφράσεις είναι στην πραγματικότητα πολύ-λεξικοί όροι όπως για παράδειγμα οι "information retrieval" ή "neural networks". Χαρακτηριστικά, μόνο οι εξαρτώμενες φράσεις από το γνωστικό πεδίο εξετάζονται, επειδή, διαφορετικά, ο αριθμός πιθανών όρων θα αυξανόταν δραστικά. Μια εύκολη προσέγγιση για τον προσδιορισμό φράσεων είναι να πληκτρολογηθεί ένα σύνολο φράσεων για μια ιδιαίτερη περιοχή πληροφορίας [Spertus,97].

4.2.3 Μείωση Διαστασιολόγησης

Το βήμα της μείωσης της διαστασιολόγησης εφαρμόζεται στην περίπτωση κατασκευής του λεξιλογίου, όταν δηλαδή υποβάλλονται σε επεξεργασία τα έγγραφα εκπαίδευσης. Υπενθυμίζεται

ότι το σύνολο των πιθανών συντασσόμενων όρων $V = \{t_1, t_2, ...t_m\}$ το οποίο προέρχεται από το αρχικό βήμα εξαγωγής των όρων για τα έγγραφα εκπαίδευσης, είναι συνήθως πολύ μεγάλο. Ο στόχος του βήματος αυτού είναι να μειωθεί ο αριθμός των ιδιο-χαρακτηριστικών γνωρισμάτων που χρησιμοποιούνται τελικά για να αντιπροσωπεύσουν τα έγγραφα, κάτω από την προϋπόθεση βέβαια, ότι το νέο προκύπτον σύνολο χαρακτηριστικών ιδιο-γνωρισμάτων πρέπει ακόμα να έχει την δυνατότητα να διακρίνει τις διαφορετικές κατηγορίες. Κατά συνέπεια, λαμβάνεται ένα μικρότερο σύνολο συνταγμένων όρων $\dot{V} = \{t_1, t_2, ...t_{\dot{m}}\}$, όπου $\dot{V}$ το λεξιλόγιο και ο δείκτης $\dot{m} < m$ δείχνει το ποσό των όρων που παραμένουν μετά την διαδικασία μείωσης.

Ο έλεγχος της διαστασιολόγησης του διανυσματικού χώρου είναι ουσιαστικός για δύο κυρίως λόγους. Η πολυπλοκότητα πολλών αλγορίθμων εκμάθησης εξαρτάται σε μεγάλο βαθμό όχι μόνο από τον αριθμό των δειγμάτων εκπαίδευσης αλλά και από τον αριθμό των ιδιο-χαρακτηριστικών γνωρισμάτων. Επίσης, αν υποτεθεί ότι περισσότερα ιδιο-χαρακτηριστικά υποτίθεται ότι μπορούν να "μεταφέρουν" περισσότερες πληροφορίες, πρέπει επομένως ένας ταξινομητής με μεγάλη ακρίβεια να αποτελείται από όσο το δυνατόν μεγαλύτερο αριθμό χαρακτηριστικών ιδιο-γνωρισμάτων. Παρόλα αυτά όμως, αν εξαιρέσει κανείς την σοβαρή επίπτωση στο χρόνο απόκρισης του ταξινομητή, η ύπαρξη ενδεχομένως πολλών άσχετων όρων, μπορεί στην ουσία να εμποδίσει έναν αλγόριθμο εκμάθησης που κατασκευάζει έναν ταξινομητή [Lewis,92]. Λαμβάνοντας υπόψη έναν σταθερό αριθμό συνόλου εκπαίδευσης, τα θεωρητικά και εμπειρικά αποτελέσματα στην μηχανική εκμάθηση έχουν πιστοποιήσει ότι υπάρχει συχνά ένας μέγιστος αριθμός ιδιο-χαρακτηριστικών γνωρισμάτων πέρα από τον οποίο η αποτελεσματικότητα ενός ταξινομητή θα αρχίσει να μειώνεται [Lewis,92]. Ως εκ τούτου, η αφαίρεση των λιγότερων πληροφοριακών αυτών ιδιο-χαρακτηριστικών γνωρισμάτων μπορεί πραγματικά να αυξήσει την απόδοση ταξινόμησης.

Το βήμα της μείωσης της διαστασιολόγησης των όρων εντάσσει οποιεσδήποτε τεχνικές που στοχεύουν στον έλεγχο της διάστασης του διανυσματικού χώρου. Αυτό περιλαμβάνει τις τεχνικές επιλογής χαρακτηριστικών γνωρισμάτων που προσπαθούν να εντοπίσουν ένα κατάλληλο υποσύνολο του δεδομένου συνόλου ιδιο-χαρακτηριστικών με σκοπό να συμπεριλάβουν κάποιους ιδιαίτερους συνταγμένους όρους στο λεξιλόγιο.

4.3 Επιλογή των Ιδιο-Χαρακτηριστικών

Οι τεχνικές επιλογής για τη μείωση της διαστασιολόγησης δέχονται ως εισαγωγή ένα σύνολο ιδιο-χαρακτηριστικών γνωρισμάτων και παράγουν ένα υποσύνολο αυτών των χαρακτηριστικών γνωρισμάτων, τα οποία είναι υπεύθυνα για τη διάκριση μεταξύ των κατηγοριών [Dash,97]. Γενικά, είναι επιθυμητό το υποσύνολο αυτό να προκαλεί την καλύτερη δυνατή υπόθεση, όσον αφορά ένα δεδομένο μέτρο αποτελεσματικότητας [Kohavi,97]. Έτσι, η επιλογή ιδιο-χαρακτηριστικών γνωρισμάτων μπορεί να θεωρηθεί ως πρόβλημα βελτιστοποίησης όπου το ζητούμενο διάστημα αντιστοιχεί στη δύναμη του συνόλου όλων των ιδιο-χαρακτηριστικών [Blum,97]. Προφανώς, αυτό απαιτεί μια εξαντλητική αναζήτηση δεδομένου ότι ο αριθμός ιδιο-χαρακτηριστικών γνωρισμάτων είναι συνήθως πολύ μεγάλος στις περιοχές κειμένων. Επομένως, η επιλογή πρέπει να καθοδηγηθεί από ευρετικές μεθόδους (heuristics) [Schurmann,96]. Στην [Langley,94] προσδιορίζονται τέσσερις διαστάσεις, βάσει των οποίων οι ευρετικές μέθοδοι αναζήτησης μπορούν να ποικίλουν. Αυτές είναι το σημείο εκκίνησης, η οργάνωση της αναζήτησης, η στρατηγική αξιολόγησης, και το κριτήριο τερματισμού.

Παρακάτω, θα αναλυθεί μια ευρέως εφαρμοσμένη γλωσσική προσέγγιση γνωστή ως αποβολή κοινών λέξεων. Κατόπιν, θα αναφερθούν κάποιοι συχνοί αριθμητικοί δείκτες που προσδιορίζουν την ποιότητα του εξεταζόμενου όρου όσον αφορά τη δυνατότητά του να κάνει διακρίσεις μεταξύ των κατηγοριών που αναφέρονται παραπάνω. Χαρακτηριστικά, οι δείκτες αυτοί βασίζονται στην

συχνότητα με την οποία οι όροι εμφανίζονται στα έγγραφα, τις κατηγορίες ή την συλλογή των εγγράφων. Να σημειωθεί, ότι οι συχνότητες όρου (*tf*) απεικονίζουν τον πραγματικό αριθμό εμφανίσεων των όρων σε συγκεκριμένα έγγραφα, ενώ άλλοι αριθμοί συχνότητας εγγράφων εξαρτώνται από όρους που βασίζονται σε δυαδικούς δείκτες ανάλογα με την παρουσίας τους ή όχι στα αντίστοιχα έγγραφα. Οι ακόλουθες Σχέσεις 23 και 24 ισχύουν για μια συλλογή εγγράφων που χωρίζεται σε *k* κλάσεις. Το συμπλήρωμα των μεγεθών $n(t)$ και $n_{c_i}(t)$ δίδεται από τις Σχέσεις 25 και 26, ενώ αθροιστικά για τις συναρτήσεις *tf* και *tf(t)*, ισχύουν οι Σχέσεις 27 και 28. Ο συντελεστής *m* αναφέρεται στον αριθμό των ιδιο-χαρακτηριστικών στο λεξιλόγιο *V*, πριν από την πιθανή εφαρμογή του βήματος μείωσης της διαστασιολόγησης.

$$n = \sum_{i=1}^{k} n_{c_i} \qquad \textbf{Σχέση 23}$$

$$n(t) = \sum_{i=1}^{k} n_{c_i}(t) \qquad \textbf{Σχέση 24}$$

$$\overline{n}(t) = n - n(t) \qquad \textbf{Σχέση 25}$$

$$\overline{n}_{c_i}(t) = n_{c_i} - n_{c_i}(t) \qquad \textbf{Σχέση 26}$$

$$tf = \sum_{i=1}^{m} tf(t_i) \qquad \textbf{Σχέση 27}$$

$$tf(t) = \sum_{j=1}^{n} tf_{d_j}(t) \qquad \textbf{Σχέση 28}$$

όπου:

n Ο αριθμός των εγγράφων στο σύνολο εκπαίδευσης *D*,

n_{c_i} Ο αριθμός των εγγράφων της κλάσης c_i,

$n(t)$ Ο αριθμός των εγγράφων στα οποία ο όρος *t* εμφανίζεται τουλάχιστον μια φορά,

$\overline{n}(t)$ Ο αριθμός των εγγράφων στα οποία ο όρος *t* δεν εμφανίζεται,

$n_{c_i}(t)$ Ο αριθμός των εγγράφων της κλάσης c_i, στα οποία ο όρος *t* εμφανίζεται τουλάχιστον μια φορά,

$\overline{n}_{c_i}(t)$ Ο αριθμός των εγγράφων της κλάσης c_i, στα οποία ο όρος *t* δεν εμφανίζεται.

4.4 Αποβολή Κοινών Λέξεων

Κατά την κειμενική ανάλυση μιας συλλογής εγγράφων, υπάρχουν πολλές λέξεις που εμφανίζονται σε όλα τα έγγραφα και επομένως προσδίδουν ελάχιστη ή σχεδόν καμία διακριτική πληροφορία. Αυτές οι λέξεις παρουσιάζουν υψηλή συχνότητα και αναφέρονται στην βιβλιογραφία ως κοινές λέξεις [Salton,83], [vanRijsbergen,79]. Πιο συγκεκριμένα, αυτές οι λέξεις είναι άρθρα, προθέσεις, ή αντωνυμίες και παρέχουν τη δομή στη γλώσσα παρά το περιεχόμενο [Sahami,98a], [Sahami,98b]. Δεδομένου ότι οι κοινές λέξεις στερούνται τη χαρακτηριστική διακριτική δύναμη, είναι λογικό να αποβάλλονται από το σύνολο των πιθανών συνταγμένων όρων του συνόλου *V*.

Υπάρχουν δύο τρόποι δημιουργίας ενός καταλόγου με κοινές λέξεις. Μια κοινή προσέγγιση είναι να δημιουργηθεί υποκειμενικά, ενώ πολλοί τέτοιοι κατάλογοι μπορούν να βρεθούν στο Διαδίκτυο και την βιβλιογραφία [Fox,92]. Ο Πίνακας 2 παρουσιάζει ένα απόσπασμα από έναν κατάλογο συχνά χρησιμοποιημένων αγγλικών λέξεων. Σε μερικές εφαρμογές, ενδέχεται να είναι πιο χρήσιμο να παρασχεθούν κατάλογοι κοινών λέξεων που εξαρτώνται από την περιοχή πληροφορίας. Η αφαίρεση των κοινών λέξεων μπορεί εύκολα προσεγγιστεί ως μια συνάρτηση που αποκρίνεται σε κάθε όρο *t* με μια ορισμένη αξία, δείχνοντας εάν ο όρος *t* θεωρείται ή όχι κοινή λέξη. Κατόπιν, οι κοινές λέξεις θα αποβάλλονταν σύμφωνα με αυτήν την αξία.

a	be	Each	if	last	Near	That
about	but	Else	in	late	no	The
all	by		is	like		They
an		For	it		of	To
and	did	From	into	many	often	
are	do	Further	itself	much	on	With
as	down			more	once	Which
at	during	Get	just	must	or	wheather

Πίνακας 2. Ένα απόσπασμα μιας λίστας κοινών λέξεων

Μια δεύτερη προσέγγιση στο πρόβλημα αυτό είναι να κατασκευαστεί ο εν λόγω κατάλογος βασισμένος αυτόματα σε μια συλλογή εγγράφων υπό εξέταση. Έτσι, μπορεί για παράδειγμα να θεωρηθούν ως κοινές λέξεις αυτές που στην συγκεκριμένη συλλογή εμφανίζουν συχνότητες εμφάνισης επάνω από ένα δεδομένο κατώφλι [Lang,95]. Το κατώφλι αυτό μπορεί να καθοριστεί είτε υποκειμενικά εκ των προτέρων είτε να οριστεί πάνω στην πληροφορία που παρέχει το ιστόγραμμα συχνότητας της κατανομής του όρου. Η προσέγγιση αυτή σε αντίθεση με την προηγούμενη, αποβάλλει τις κοινές λέξεις της συγκεκριμένης συλλογής εγγράφων υπό εξέταση και επομένως είναι εξαρτώμενη από το πεδίο πληροφορίας. Έτσι, όχι μόνο οι συνήθως αποδεκτές κοινές λέξεις ενδέχεται να αποβληθούν, αλλά και άλλες λέξεις όπως ουσιαστικά, ρήματα, ή επίθετα.

5 Τεχνικές Συσταδοποποίησης Εγγράφων

5.1 Εισαγωγή

Από τη στιγμή που η καταχώριση πληροφορίας στον παγκόσμιο ιστό έγινε εύκολη και δημοφιλής, ο εντοπισμός των πηγών, με υψηλή ποιότητα και σχετικού περιεχομένου προς μια δεδομένη ανάγκη πληροφόρησης, γίνεται όλο και πιο δύσκολη. Οι τεχνικές ανάκτησης πληροφορίας που αναπτύχθηκαν στο πέρασμα των χρόνων, έδωσαν μια ικανοποιητική απάντηση. Το γνωστό πρόβλημα ανάκτησης πληροφοριών, "Δεδομένου ενός συνόλου εγγράφων και μιας ερώτησης, καθόρισε το υποσύνολο των εγγράφων σχετικών με την ερώτηση ", στο πρόβλημα αναζήτησης στον παγκόσμιο ιστό μπορεί να μετατραπεί ως "Πως ένας χρήστης μπορεί να βρει το σύνολο σχετικών με ένα θέμα εγγράφων που βρίσκονται στον παγκόσμιο ιστό".

Οι υπηρεσίες αναζήτησης πληροφορίας στον παγκόσμιο ιστό, οι οποίες ουσιαστικά αποτελούν συστήματα ανάκτησης πληροφορίας, κατέχουν μια εξέχουσα θέση μεταξύ όλων των υπηρεσιών του διαδικτύου και λαμβάνουν και επεξεργάζονται εκατομμύρια ερωτήσεων καθημερινά. Εντούτοις, όλες οι κορυφαίες μηχανές αναζήτησης (Google, Yahoo, MSN κλπ) είναι βασισμένες στη λογική διατύπωσης ερωτημάτων όπου ο χρήστης διατυπώνει ένα ερώτημα που αποτελείται από ένα σύνολο λέξεων συχνά συνοδευόμενες με έναν λογικό τελεστή, και το σύστημα επιστρέφει έναν ταξινομημένο κατάλογο εγγράφων.

Αυτό που ποικίλλει συχνά μεταξύ των υπηρεσιών αναζήτησης, είναι οι αλγόριθμοι ταξινόμησης, που κυμαίνονται από μια απλή αναζήτηση των όρων της ερώτησης στο έγγραφο, ως και μια προηγμένη ανάλυση γραφικής αναπαράστασης ή και χρήση ανατροφοδότησης χρηστών. Η αποτελεσματικότητα των μηχανών αναζήτησης στον παγκόσμιο ιστό, κρίνεται ικανοποιητική όταν η ανάγκη πληροφόρησης του χρήστη είναι συγκεκριμένη και ακριβώς ορισμένη. Ωστόσο, οι υπάρχουσες μηχανές αναζήτησης δεν βρίσκονται ακόμη σε θέση να αντιμετωπίσουν πιο σύνθετες αναζητήσεις όπως για παράδειγμα αναζήτηση εννοιών.

Όπως υποστηρίζεται από τον [Marchionini,92], "οι χρήστες θέλουν να επιτύχουν τους στόχους τους με ένα ελάχιστο γνωστικό φορτίο και ένα μέγιστο βαθμό ευχρηστίας". Οι χρήστες γενικά διατυπώνουν πολύ σύντομες ερωτήσεις (ένας έως τρεις όροι [Croft,95], [Pinkerton,94]), μερικές φορές πολύ διφορούμενες, και χωρίς πολλή σκέψη στη διατύπωση ερώτησης. Ένα κλασικό παράδειγμα αυτού του προβλήματος είναι η έννοια της “Java” ως ερώτηση: όπου μπορεί να σημαίνει ότι υπάρχει ενδιαφέρον για τον καφέ, για τη γλώσσα προγραμματισμού, ή το νησί της Ινδονησίας. Η ασαφής διατύπωση μιας ερώτησης μπορεί εμφανιστεί με διαφορετικούς τρόπους και σε μεγάλη συχνότητα. Ακόμα, συνήθης είναι η περίπτωση, που οι ίδιοι χρήστες είναι ασαφείς σχετικά με την ανάγκη πληροφοριών τους [Efthimiadis,93]. Το αποτέλεσμα τέτοιων ερωτήσεων είναι η δημιουργία μια τεράστιας λίστας αποτελεσμάτων εκ των οποίων ένας ελάχιστος αριθμός είναι σχετικός με τις ανάγκες του χρήστη.

Η συσταδοποίηση μεγάλων συλλογών κειμένων είναι μια σημαντική διαδικασία για την παροχή ενός υψηλότερου επιπέδου γνώσης σχετικά με την θεμελιώδη ταξινόμηση των εγγράφων. Η ανάλυση των εγγράφων του Διαδικτύου, ειδικότερα, αποκτά μεγαλύτερο ενδιαφέρον από τη στιγμή που η πρόσβαση, η διαχείριση, η αναζήτηση και η περιήγηση στην αχανή βάση δεδομένων πληροφορίας του Διαδικτύου απαιτεί μεγάλο επίπεδο οργάνωσης.

Σύμφωνα με την υπόθεση συσταδοποίησης εγγράφων [vanRijsbergen,79], τα σχετικά έγγραφα είναι πιθανότερο να είναι παρόμοια μεταξύ τους παρά με τα άσχετα. Για τη συσχέτιση των εγγράφων χρησιμοποιείται η έννοια της ομοιότητας εγγράφων. Ειδικά σε περιπτώσεις που υπάρχει έγγραφο αναφοράς η ομοιότητα ενός εγγράφου υπολογίζεται με βάση το έγγραφο αναφοράς.

Αυτό αποτελεί μια μάλλον διαφορετική προσέγγιση στην αναζήτηση στον παγκόσμιο ιστό, δεδομένου, ότι η είσοδος σε αυτήν την διαδικασία ανάκτησης δεν είναι ένα σύνολο όρων, αλλά ένα έγγραφο αναφοράς. Αυτό είναι ιδιαίτερα χρήσιμο όταν οι χρήστες δυσκολεύονται να εκφράσουν την ανάγκη για συγκεκριμένη πληροφορία με ένα σύνολο λέξεων κλειδιών, παρόλο που διαθέτουν ένα παράδειγμα αυτού που ψάχνουν. Με τη χρήση ενός μέτρου της ομοιότητας εγγράφων, δίνεται η δυνατότητα συσταδοποίησης των παρόμοιων σχετικών έγγραφων ενώ ταυτόχρονα, η απομάκρυνση των μη σχετικών εγγράφων.

5.2 Μέθοδοι ιεραρχικής συσταδοποίησης

Οι μέθοδοι ιεραρχικής συσταδοποίησης βασίζονται στην ιδέα της ομαδοποίησης των μονάδων πληροφορίας (έγγραφα) σε μια δενδρική μορφή. Η ενδιάμεση δομή δεδομένων που κατασκευάζεται κατά τη διάρκεια της επεξεργασίας καλείται *dendogram*. Ανάλογα με τον τρόπο που δημιουργείται η δομή, οι ιεραρχικοί αλγόριθμοι διακρίνονται σε συσωρευτικούς και διαχωριστικούς και αναφέρονται παρακάτω.

5.2.1 Η συσσωρευτική προσέγγιση

Η συσσωρευτική προσέγγιση (Hierarchical Agglomerative Clustering, HAC) δημιουργεί την ιεραρχία από κάτω προς τα επάνω. Αρχικά ορίζει την κάθε μονάδα πληροφορίας ως μεμονωμένη συστάδα. Επαναληπτικά συγκρίνει αυτές ανά ζεύγη και συγχωνεύει το ζεύγος συστάδων με τη μεγαλύτερη ομοιότητα μέχρι τελικά να ενωθούν όλες οι συστάδες. σε μια που αποτελεί και το ανώτερο επίπεδο ιεραρχίας. Οι αλγόριθμοι αυτής της οικογένειας συνήθως ακολουθούν τα παρακάτω βήματα:

1. Δημιουργία μιας μήτρας (i,j) η οποία περιέχει την ομοιότητα της i συστάδας με την j.
2. Το ζευγάρι με τον μεγαλύτερο βαθμό ομοιότητας επιλέγεται και συγχωνεύεται σε μια νέα συστάδα, ενώ ταυτόχρονα ο ολικός αριθμός συστάδων μειώνεται κατά ένα.
3. Υπολογισμός εκ νέου του βαθμού ομοιότητας της δημιουργηθείσας συστάδας με τις υπόλοιπες.
4. Επανάληψη των βημάτων 2 και 3 μέχρι να ενωθούν όλες οι συστάδες σε μια ενιαία.

Η κύρια διαφοροποίηση των αλγορίθμων έγκειται, συνήθως, στο διαφορετικό τρόπο υπολογισμού του μέτρου της ομοιότητας μεταξύ των συστάδων. Οι επικρατέστεροι είναι τρεις, από τους ο οποίους αυτός που κάνει τη χρήση της μέσης τιμής της ομοιότητας, οδηγεί σε πιο ακριβή αποτελέσματα.

1. Συσταδοποίηση απλής διασύνδεσης. Η ομοιότητα μεταξύ δυο συστάδων είναι ίση με την μέγιστη ομοιότητα που παρουσιάζεται μεταξύ ενός μέλους της μιας συστάδας με ένα άλλο της άλλης.
2. Συσταδοποίηση πλήρους διασύνδεσης. Η ομοιότητα μεταξύ δυο συστάδων είναι ίση με την ελάχιστη ομοιότητα που παρουσιάζεται μεταξύ ενός μέλους της μιας συστάδας με ένα άλλο της άλλης.
3. Συσταδοποίηση μέσης τιμής διασύνδεσης. Η ομοιότητα μεταξύ δυο συστάδων είναι ίση με την μέση τιμή ομοιότητας που παρουσιάζεται μεταξύ όλων των μελών της μιας συστάδας με όλα της άλλης.

5.2.2 Η διαχωριστική προσέγγιση

Αντίθετα με τη συσσωρευτική προσέγγιση, η διαχωριστική (Hierarchical Divisive Clustering, HDC) δημιουργεί την ιεραρχία από πάνω προς τα κάτω. Δηλαδή αρχικά όλες οι μονάδες πληροφορίας ορίζουν μια ενιαία συστάδα και επαναληπτικά διαιρείται σε μικρότερες μέχρι κάθε συστάδα να αποτελείται από μόνο μια μονάδα πληροφορίας . Οι πιο κοινά χρησιμοποιούμενες τεχνικές είναι οι συσσωρευτικές γιατί εκτός από την καλύτερη απόδοση απαιτούν και μικρότερη υπολογιστική ισχύ. Ωστόσο οι υπολογιστική ισχύ που απαιτούνται και οι δυο μέθοδοι καθιστούν εξαιρετικά δύσκολη την εφαρμογή τους στον παγκόσμιο ιστό.

5.3 Μέθοδοι διαιρετικής συσταδοποίησης

Μια μέθοδος χωρισμού δημιουργεί συσταδοποίηση ενός επιπέδου αλλά η επαναληπτική εφαρμογή της μπορεί να παρέχει και ιεραρχική δομή. Η λογική της μεθόδου είναι η εξαρχής τυχαία δημιουργία ενός αριθμού *k* συστάδων τις οποίες βελτιώνει, μεταφέροντας επαναληπτικά μονάδες πληροφορίας από τη μια συστάδα στην άλλη. Ο πιο γνωστός αλγόριθμος είναι ο K-means και οι παραλλαγές του. Ένα σημαντικό μειονέκτημα της μεθόδου συσταδοποίησης K-means είναι ότι απαιτεί το χρήστη για να διευκρινίσει τον αριθμό Κ συστάδων εκ των προτέρων, ο οποίος μπορεί σε μερικές περιπτώσεις να είναι αδύνατος να υπολογιστεί με αποτέλεσμα να οδηγεί σε χαμηλής ποιότητας συσταδοποίηση. Τα βήματα του αλγορίθμου είναι:

1. *K* μονάδες πληροφορίας επιλέγονται τυχαία ως αρχικοί πυρήνες των *k* συστάδων.
2. Η κάθε μονάδα πληροφορίας εντάσσεται στην συστάδα με την μεγαλύτερη συσχέτιση (Γίνεται υπολογισμός της ομοιότητας μεταξύ του εγγράφου και των πυρήνων κάθε συστάδας).
3. Εκ νέου υπολογισμός του πυρήνα κάθε συστάδας.
4. Επανάληψη των βημάτων 2 και 3 μέχρι οι πυρήνες να συγκλίνουν σε μια σταθερή κατάσταση.

5.4 Συσταδοποίση με τη χρήση δένδρων προσφυμάτων

Η ιδέα της χρήσης δένδρων προσφυμάτων στη συσταδοποίηση εγγράφων, αρχικά προτάθηκε στο [Zamir,97], όπου ο προτεινόμενος αλγόριθμος (Suffix Tree Clustering STC) αρχίζει με την κατασκευή του δένδρου προσφυμάτων για όλες τις προτάσεις των εγγράφων που ανήκουν στη συλλογή. Στην προσέγγιση του αλγορίθμου το κάθε έγγραφο ορίζεται ως μια ακολουθία χαρακτήρων. Ένα δένδρο προσφυμάτων για μια σειρά είναι ένα συμπαγές δένδρο που περιέχει προσφύματα [Grossi,93] [Andersson,95], [Andersson,99], Cox,99].Ο συγκεκριμένος αλγόριθμος συσταδοποίησης μεταχειρίζεται τα έγγραφα ως αλληλουχία λέξεων και όχι χαρακτήρων, έτσι κάθε πρόσφυμα περιέχει μια ή περισσότερες ολόκληρες λέξεις. Με άλλα λόγια, ισχύουν τα παρακάτω:

- Ένα δένδρο προσφυμάτων μιας ακολουθίας *S* είναι ένα κατευθυνόμενο δέντρο.
- Κάθε εσωτερικός κόμβος του δένδρου έχει τουλάχιστον 2 παιδιά.
- Σε κάθε ακμή ορίζεται μια ετικέτα η οποία περιέχει μια υπο-ακολουθία του *S*. Η ετικέτα ενός κόμβου καθορίζεται για να είναι η αλληλουχία των ετικετών των ακμών στην πορεία από τη ρίζα μέχρι τον συγκεκριμένο κόμβο.
- Δυο ακμές του ίδιου κόμβου δεν μπορούν να έχουν ετικέτες που αρχίζουν με την ίδια λέξη.

- Για κάθε πρόσφυμα *s* του *S*, υπάρχει ένα επίθημα - κόμβος του οποίου η ετικέτα είναι ίση με *το s*.

Το δέντρο προσφυμάτων μιας συλλογής ακολουθιών, είναι ένα συμπαγές δένδρο που περιέχει όλα τα προσφύματα όλων των σειρών στη συλλογή. Κάθε επίθημα - κόμβος είναι σημειωμένος να προσδιορίζει τη συμβολοσειρά από την οποία προέρχεται.

Το σχήμα 3.6 είναι ένα παράδειγμα δέντρου επιθημάτων των ακολουθιών χαρακτήρων "Η γάτα έφαγε το τυρί", "Το ποντίκι έφαγε το τυρί επίσης" και "Η γάτα έφαγε το ποντίκι επίσης" που οι κόμβοι του δέντρου επιθήματος σχεδιάζονται ως κύκλοι. Ο κάθε κόμβος - επίθημα, συνδέεται με μια η ή περισσότερες ετικέτες, που υποδεικνύουν την ακολουθία (*S*) από την οποία προήλθε. Ο πρώτος αριθμός σε κάθε ετικέτα, υποδεικνύει την ακολουθία προέλευσης και ο δεύτερος αριθμός υποδεικνύει ποιο επίθημα της ακολουθίας ονομάζει τον κόμβο - επίθημα.

Κάθε κόμβος του δέντρου επιθημάτων, αντιπροσωπεύει ένα σύνολο εγγράφων και μιας φράσης που είναι κοινή για όλα τα έγγραφα του συνόλου αυτού. Η ετικέτα του κόμβου, αντιπροσωπεύει μια φράση η οποία είναι κοινή σε ένα σύνολο εγγράφων. Επομένως, κάθε κόμβος αντιπροσωπεύει μια βασική συστάδα εγγράφων, και όλες οι πιθανές βασικές συστάδες (περιέχοντας 2 ή περισσότερα έγγραφα) εμφανίζονται ως κόμβοι στο δέντρο επιθημάτων.

Σε κάθε βασική συστάδα ορίζεται μια τιμή $s(B)$ που είναι συνάρτηση του αριθμού εγγράφων που περιέχει, και των λέξεων που αποτελούν τη φράση της. Η τιμή s (B) της βασικής συστάδας β με τη φράση P δίνεται από τη σχέση $s(B)=|B|* f(|P|)$, όπου $|B|$ είναι ο αριθμός εγγράφων στη βασική συστάδα β, και $|P|$ είναι ο αριθμός λέξεων *στο* P που είναι διάφορος του μηδενός.

Τα έγγραφα μπορούν να περιέχουν περισσότερες από μια κοινές φράσεις. Κατά συνέπεια, τα σύνολα των εγγράφων που περιέχονται στις βασικές συστάδες μπορεί να έχουν έναν μεγάλο αριθμό επικάλυψης ή ακόμη να είναι και ίδια. Για την αποφυγή παραγωγής σχεδόν όμοιων συστάδων, ο αλγόριθμος συγχωνεύει τις βασικές συστάδες με μεγάλο ποσοστό επικάλυψης στο σύνολο των εγγράφων τους . Με αυτόν τον τρόπο ορίζεται ένα δυαδικό μέτρο ομοιότητας μεταξύ των βασικών συστάδων βασιζόμενο στην επικάλυψη των συνόλων των εγγράφων τους. Δοθέντων δυο βασικών συστάδων B_m και B_n, με μέγεθος $|B_m|$ και $|B_n|$ αντίστοιχα, και $\left|B_m \cap B_n\right|$ ο κοινός αριθμός εγγράφων στις δυο συστάδες, η ομοιότητα των δυο συστάδων παίρνει την τιμή 1 όταν ισχύει:

$$\frac{\left|B_m \cap B_n\right|}{B_m} > 0.5 \text{ και } \frac{\left|B_m \cap B_n\right|}{B_n} > 0.5$$

Σε κάθε άλλη περίπτωση η ομοιότητα είναι ίση με 0.

Στη συνέχεια, ο αλγόριθμος εξετάζει τη "γραφική παράσταση συστάδων βάσεων", όπου οι κόμβοι είναι συστάδες βάσεων. Δύο κόμβοι συνδέονται εάν οι δύο συστάδες βάσεων έχουν ομοιότητα 1. Μια συστάδα ορίζεται ένα συνδεδεμένο συστατικό στη γραφική παράσταση των συστάδων βάσεων. Κάθε συστάδα περιέχει την ένωση των εγγράφων όλων των συστάδων βάσεών της. Αυτός ο αλγόριθμος συγκέντρωσης είναι επαυξητικός και ανεξάρτητος διαταγής.

Τελικά, οι συστάδες που προκύπτουν, βαθμολογούνται και ταξινομούνται, με βάση τις τιμές των βασικών συστάδων τους και της επικάλυψή τους. Καθώς ο τελικός αριθμός συστάδων μπορεί να ποικίλει, ο αλγόριθμος παρουσιάζει μόνο ένα μικρό αριθμό συστάδων με την καλύτερη βαθμολογία (ένας τυπικός αριθμός είναι το 10). Για κάθε συστάδα αναφέρεται ο αριθμός των εγγράφων που περιέχει καθώς και τις φράσεις των βασικών συστάδων της.

5.5 Επαυξητική συσταδοποίηση

Οι αλγόριθμοι επαυξητικής συσταδοποίησης συνήθως ανάγονται συνήθως στις παραδοσιακές τεχνικές συσταδοποίησης, που μπορούν να εφαρμοστούν σε ένα δυναμικό περιβάλλον όπως ο παγκόσμιος Ιστός.

Οι Μέθοδοι επαυξητικής συσταδοποίησης παρουσιάζουν ιδιαίτερο ενδιαφέρον κατά την εφαρμογή τους στο περιεχόμενο του παγκόσμιου ιστού λόγω της δυνατότητάς τους να αντιμετωπίσουν τέτοιας τάξης μεγέθους μεταβαλλόμενο περιεχόμενο. Οι τεχνικές συσταδοποίησης εγγράφων συνήθως εφαρμόζονται για την ομαδοποίηση των ανακτημένων εγγράφων και την παρουσίαση οργανωμένων και κατανοητών αποτελεσμάτων στο χρήστη, την ομαδοποίηση εγγράφων σε μια συλλογή όπως για παράδειγμα ψηφιακές βιβλιοθήκες, την αυτόματη ή ημιαυτόματη δημιουργία κατηγοριών ταξινομιών εγγράφων (π.χ. Yahoo και ανοικτές μορφές καταλόγου) και την αποδοτικότερη ανάκτηση πληροφορίας με ιδιαίτερη βαρύτητα σε σχετικά υποσύνολα (συστάδες) παρά σε ολόκληρη τη συλλογή. Η διαφορά μεταξύ των παραδοσιακών μεθόδων συσταδοποίησης με την επαυξητική συσταδοποίηση είναι η δυνατότητα επεξεργασίας των νέων στοιχείων που προστίθενται στη συλλογή δεδομένων. Αυτό επιτρέπει τη δυναμική παρακολούθηση του καθημερινά αυξανόμενου ρυθμού τοποθέτησης πληροφορίας στον παγκόσμιο ιστό χωρίς να απαιτείται επανασυσταδοποίηση. Ο αλγόριθμος επαυξητικής συσταδοποίησης βασίζεται στη διατήρηση της υψηλής συνεκτικότητας των συστάδων, που αναφέρεται και ως ιστόγραμμα ομοιότητας συστάδων, το οποίο αποτελεί και μια συνοπτική στατιστική αναπαράσταση της κατανομής της ομοιότητας που παρουσιάζουν ανά ζεύγη τα έγγραφα μέσα σε κάθε συστάδα. Όσο προστίθενται νέα έγγραφα, οι συστάδες απαιτείται να διατηρούν ένα υψηλό επίπεδο συνεκτικότητας, που εκφράζεται με τους όρους του ιστογράμματος, Η διαδικασία επιτρέπει την επανατοποθέτηση των εγγράφων σε συστάδες που δημιουργήθηκαν μετά την εισαγωγή του έγγραφου. Για τη δημιουργία συνεκτικών συστάδων συνοχής ακολουθείται η τακτική διατήρησης σε υψηλά επίπεδα της ομοιότητας μέσα στις συστάδες.

Η μεθοδολογία επαυξητικής συσταδοποίησης που αναλύεται στην παράγραφο αυτή, προτάθηκε από τον [Hammouda,03] και βασίζεται σε ιστόγραμμα ομοιοτήτων (Similarity Histogram Clustering SHC) που αντιπροσωπεύει τον βαθμό συνοχής των συστάδων. Με άλλα λόγια αποτελεί μια περιεκτική στατιστική αναπαράσταση της κατανομής των ομοιοτήτων των ανά ζεύγη εγγράφων μέσα στη συστάδα. Το ιστόγραμμα διαιρείται σε ένα σύνολο περιοχών που αντιστοιχούν σε σταθερά διαστήματα τιμών ομοιότητας. Κάθε περιοχή περιλαμβάνει ένα σύνολο τιμών ομοιότητας για το αντίστοιχο διάστημα.

Ο κύριος στόχος της μεθοδολογίας επικεντρώνεται στην επίτευξη συνεκτικών συστάδων ώστε να διατηρηθεί η υψηλός-ομοιότητα μέσα στις συστάδες. Στο ιστόγραμμα ομοιότητας, αυτό σημαίνει ότι η κατανομή των τιμών των ομοιοτήτων επεκτείνεται προς τα δεξιά. Κάθε νέο έγγραφο που πρόκειται να τοποθετηθεί σε μια συστάδα, συγκρίνεται με το αντίστοιχο ιστόγραμμα κάθε συστάδας. Αν η προσθήκη του νέου εγγράφου χειροτερεύει την κατανομή του ιστογράμματος τότε το έγγραφο απορρίπτεται, αλλιώς προστίθεται. Η ιδανικότερη περίπτωση θα ήταν η ενίσχυση της κατανομής της ομοιότητας, αλλά αυτό πρακτικά οδηγεί σε "τέλειες" συστάδες και θα απέρριπταν σχεδόν κάθε έγγραφο.

Η τιμή της ποιότητας ενός ιστογράμματος ομοιότητας (και κατ' επέκταση της συνεκτικότητας της συστάδας), υπολογίζεται με βάση το λόγο του αριθμού των ομοιοτήτων μέσα στη συστάδα, που έχουν τιμή πάνω από ένα κατώφλι S_T, προς το συνολικό αριθμό ομοιοτήτων και δίνεται από τη Σχέση 29. Όσο υψηλότερη αυτή η τιμή, τόσο πιο συνεκτική είναι η συστάδα.

$$HR_c = \frac{\sum_{i=T}^{B} h_i}{\sum_{j=1}^{B} h_j} \text{ με } T = [S_T \cdot B] \qquad \textbf{Σχέση 29}$$

Όπου S_T η τιμή κατωφλίου και T ο αριθμός της περιοχής του ιστογράμματος που ανταποκρίνεται στο κατώφλι. Θεωρώντας ότι ο αριθμός των εγγράφων σε μια συστάδα είναι n_c, Ο αριθμός των ανά ζεύγη ομοιοτήτων είναι $m_c = n_c(n_c+1)/2$. Αν $S = \{s_i : i = 1,...m_c\}$ είναι το σύνολο των ομοιοτήτων στη συστάδα τότε το ιστόγραμμα ομοιοτήτων της συστάδας δίνεται από τις σχέσεις:

$$H = \{h_i : i = 1,...B\} \qquad \textbf{Σχέση 30α}$$

$$h_i = count(s_k) \text{ με } s_{li} \le s_k \le s_{ui} \qquad \textbf{Σχέση 30β}$$

Όπου B ο αριθμός των περιοχών του ιστογράμματος, h_i ο αριθμός των ομοιοτήτων στην περιοχή i, s_{ji} το ελάχιστο όριο της τιμής της ομοιότητας στην περιοχή i και s_{ui} το υψηλότερο όριο της τιμής της ομοιότητας στην περιοχή i. Αν και η βασική ιδέα της συγκεκριμένης μεθόδου είναι η διατήρηση υψηλής συνοχής στις συστάδες, εντούτοις η προσθήκη εγγράφων που υποβιβάζουν την ποιότητα θα μπορούσε σταδιακά να καταστρέψει τη συνοχή της συστάδας. Αυτό αποφεύγεται με τη χρήση μιας ελάχιστης τιμής συνοχής HR_{min}.

Στο σχήμα 7 παρουσιάζεται με τη μορφή ψευδοκώδικα ο αλγόριθμος επαυξητικής συσταδοποίησης. Για κάθε νέο έγγραφο που καταφθάνει προς ταξινόμηση, υπολογίζεται το ιστόγραμμα ομοιότητας της κάθε συστάδας. Για κάθε συστάδα γίνεται προσομοίωση προσθήκης του εγγράφου στη συστάδα και υπολογίζεται η ποιότητα συνεκτικότητας αυτής. Όταν η νέα τιμή της ποιότητας είναι μεγαλύτερη από ένα κάτω όριο HR_{min} και είναι μικρότερη κατά ε από την τιμή πριν την προσθήκη ή και μεγαλύτερη τότε το έγγραφο προστίθεται στη συστάδα, αλλιώς όχι. Αν τελικά δεν προστεθεί σε καμία συστάδα, τότε δημιουργείται μια νέα συστάδα και το έγγραφο προστίθεται σ' αυτή.

```
L ← Empty List {Cluster List}
 for each document D do
          for each cluster C in L do
                    HRold = HRC
                    Simulate adding D to C
                    HRnew = HRC
                     if (HRnew ≥ HRold) OR ((HRnew > HRmin) AND (HRold − HRnew < ε)) then
                              Add D to C
                    end if
          end for
          if D was not added to any cluster then
                    Create a new cluster C
                    ADD D to C; ADD C to L
          end if
end for
```

Σχήμα 7. Ψευδοκώδικας του αλγορίθμου επαυξητικής συσταδοποίησης

6 Αναφορές

[Anagnostopoulos,03] I. Anagnostopoulos, C. Anagnostopoulos, Vassili Loumos, Eleftherios Kayafas, "Taxonomy of E-commerce Web Pages employing a Probabilistic Neural Network Classifier", submitted to IEE Proceedings - Software (under review process).

[Anagnostopoulos,04] "Νέες Μέθοδοι Για Ευφυή Ανάκτηση ,Κατηγοριοποίηση Και Ταξινόμηση Δεδομένων Σε Πληροφοριακά Συστήματα", Διδακτορική διατριβή 2004, ΕΜΠ

[Andersson,95] A. Andersson and Stefan Nilsson. Efficient implementation of suffix trees. Software - Practice and Experience, 25(2):129-141, 1995.

[Andersson,99] A. Andersson, N. Jesper Larsson, and Kurt Swanson. Suffix trees on words. Algorithmica, 23(3):246-260, 1999.

[Apte,94] Apte, C., Damerau, F., and Weiss, S. M. (1994). Automated learning of decision rules for text categorization. ACM Transactions on Information Systems 12(3), 233–251.

[Baeza-Yates,99] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley Longman Inc., 1999.

[Balabanovic,97] Balabanovic, M. (1997). An adaptive web page recommendation service. In Proceedings of the First International Conference on Autonomous Agents, Marina del Rey, CA, pp. 378–385.

[Belkin,92] Belkin, N. J. and Croft, W. B. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM 35(12), 29–38.

[Blair,92] Blair, D. (1992). Information retrieval and the philosophy of language. The Computer Journal 35(3), 200–207.

[Breese,98] Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, pp. 43–52. Morgan Kaufmann Publisher.

[Cohen,96] Cohen, W. W. and Singer, Y. (1996). Context-sensitive learning methods for text categorization. In Proceedings of the Nineteenth Annual International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 307–315.

[Cox,99] Sarah Cox. "Suffix trees in java". Master's thesis, Department of Computing Scienceof the University of Glasgow, 1999.

[Croft,95] W. Bruce Croft, Robert Cook, and Dean Wilder. Providing government information on the Internet: Experiences with THOMAS. In Proceedings of DL-95, the 2nd Annual Conference on the Theory and Practice of Digital Libraries, pages 19-24, Austin, Texas, U.S.A., June 1995.

[Dash,97] Dash, M. and Liu, H. (1997). Feature selection for classification. Intelligent Data Analysis 1(3). An online version is avaible at http://www-east.

[Denning,82] Denning, P. J. (1982). Electronic Junk. Communications of the ACM 25(3), 163–165.

[Dumais,98] Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representation for text categorization. In Proceedings of the Seventh International Conference on Information and Knowledge Management, pp. 148–155.

[Ferguson,96] Ferguson, I. A. and Karakoulas, G. J. (1996). Multiagent learning and adaptation in an information filtering market. In Proceedings AAAI Spring Symposium on Adaptation,Co-evolution and Learning in Multiagent Systems, pp. 28–32.

[Fox,92] Fox, C. (1992). Lexical analysis and stoplists. See Frakes and Baeza-Yates (1992), pp. 102–130.

[Furnkranz,98] Furnkranz, J., Mitchell, T., and Riloff, E. (1998). A case study in using linguistic phrases for text categorization on the www. In M. Sahami (Ed.), Proceedings of the AAAI/ICML'98 Workshop on Learning for Text Categorization, Madison, WI, pp. 5–12.

[Grossi,93] R. Grossi and Giuseppe F. Italiano. Suffix trees and their applications in string algorithms. In Proceedings of WSP-93, the 1st South American Workshop on String Processing, pages 57-76, September 1993.

[de Heer,82] de Heer, T. (1982). The application of the concept of homeosemy to natural language information retrieval. Information Processing Management 18(5), 229–236.

[Efthimiadis,93] E. N. Efthimiadis. A user-centred evaluation of ranking algorithms for interactive query expansion. In Proceedings of SIGIR-93,the 16th ACM International Conference on Research and Development in Information Retrieval, pages 146-159. ACM Press, 1993.

[Hammouda,03] K.M. Hammouda, M. S. Kamel, "Incremental Document Clustering Using Cluster Similarity Histograms", WIC International Conference on Web Intelligence, (WI 2003), 13-17 October 2003, Halifax, Canada. IEEE Computer Society 2003, ISBN 0-7695-1932-6, pp. 597-601.

[Hull,98] Hull, D. A. (1998). The TREC-6 Filtering Track: Description and Analysis. In Proceedings of the Sixth Text REtrieval Conference (TREC-6), Gaithersburg, MD, pp. 45–67. National Institute of Standards and Technology.

[Ingwersen,92] Ingwersen, P. (1992). Information Retrieval Interaction. Taylor Graham.

[Joachims,97] Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In International Conference on Machine Learning, pp. 143–151.

[Kohavi,97] Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence 97(1–2), 273–324.

[Lang,95] Lang, K. (1995). Newsweeder: Learning to filter netnews. In Proceedings of the Twelfth International Conference on Machine Learning, pp. 331–339.

[Lewis,91] Lewis, D. D. (1991). Evaluating text categorization. In Proceedings of the Speech and Natural Language Workshop, Asilomar, pp. 312–318. Morgan Kaufmann Publishers.

[Lewis,92] Lewis, D. D. (1992a). An evaluation of phrasal and clustered representations on a text categorization task. In Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, pp. 37–50.

[Marchionini,92] G. Marchionini. Interfaces for end-user information seeking. Journal of the American Society for Information Science, 43(2):156-163, 1992.

[Mostafa ,97] Mostafa, J., Mukhopadhyay, S., Lam, W., and Palakal, M. (1997). A multilevel approach to intelligent information filtering: Model, system, and evaluation. ACM Transactions on Information Systems 15(4), 368–399.

[Oard,97] Oard, D. W. (1997). The state of the art in text filtering. User Modeling and User-Adapted Interaction 7(3), 141–178.

[Pazzani,97] Pazzani, M. and Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. Machine Learning 27, 313–331.

[Pinkerton,94] B. Pinkerton. "Finding what people want: Experiences with the WebCrawler".In Proceedings of WWW-94, the 2nd International World Wide Web Conference,October 1994.

[Robertson,76] S. E. Robertson, K. Sparck Jones, Relevance weighting of search terms. Journal of the American Society for Information Sciences, 27(3): 129-146, 1976.

[Sahami,98a] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In M. Sahami (Ed.), Proceedings of the AAAI/ICML'98 Workshop on Learning for Text Categorization, Madison, WI.

[Sahami,98b] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In M. Sahami (Ed.), Proceedings of the AAAI/ICML'98 Workshop on Learning for Text Categorization, Madison, WI.

[Salton,68] G. Salton, M. E. Lesk. Computer evaluation of indexing and text processing, Journal of the ACM, 15(1): 8-36, January 1968.

[Salton,71] G. Salton. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice Hall Inc., 1971.

[Salton,83] G. Salton, E.A. Fox, H. Wu. Extended Boolean information retrieval. Communications of the ACM, 26(11): 1022-1036, November 1983.

[Salton,88] G. Salton, C. Buckley. Term-weighting approaches in automatic retrieval, Information Processing & Management, 24(5): 513-523, 1988

[Salton,89] Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley.

[Schapire,98] Schapire, R. E., Singer, Y., and Singhal, A. (1998). Boosting and Rocchio applied to text filtering. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pp. 215–223.

[Schurmann,96] Schurmann, J. (1996). Pattern Classification: A unified view of statistical and neural approaches. New York: Wiley-Interscience.

[Spertus,97] Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In Proceedings of Innovative Applications of Artificial Intelligence (IAAI), pp. 1058–1065.

[Tauritz,00] Tauritz, D. R., Kok, J. N., and Sprinkhuizen-Kuyper, I. G. (2000). Adaptive information filtering using evolutionary computation. Information Sciences 12-2, 121–140.

[Teufel,88] Teufel, B. and Schmidt, S. (1988). Full text retrieval based on syntactic similarities.Information Systems 13(1), 65–70.

[vanRijsbergen,79] van Rijsbergen, C. J. (1979). Information Retrieval (2 ed.). London: Butterworths. Online version: http://www.dcs.gla.ac.uk/Keith.

[Zamir,97] Oren Zamir, Oren Etzioni, Omid Madani, and Richard M. Karp. Fast and intuitive clustering of Web documents. In Proceedings of KDD-97, the 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 287-290, 1997.

[Μακρής,02] Μακρής Χ., Σημειώσεις του μαθήματος "Ανάκτηση Πληροφορίας", Πανεπιστήμιο Πατρών, 2002 URL: http://thalis.cs.unipi.gr/~ir/.

ΚΕΦΑΛΑΙΟ

3

ΑΛΓΟΡΙΘΜΟΣ ΑΠΟΙΚΙΑΣ ΜΥΡΜΗΓΚΙΩΝ ΚΑΙ ΧΡΗΣΗ ΣΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ ΠΛΗΡΟΦΟΡΙΑΣ

ΠΕΡΙΕΧΟΜΕΝΑ 3ου ΚΕΦΑΛΑΙΟΥ

ΑΛΓΟΡΙΘΜΟΣ ΑΠΟΙΚΙΑΣ ΜΥΡΜΗΓΚΙΩΝ ΚΑΙ ΧΡΗΣΗ ΣΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ ΠΛΗΡΟΦΟΡΙΑΣ

1 Εισαγωγή

Στο κεφάλαιο αυτό γίνεται μια εισαγωγή στους αλγορίθμους νοημοσύνης σμηνών και στα ιδιαίτερα χαρακτηριστικά τους που κατά κύριο λόγο εμπνέονται από τις συλλογικές συμπεριφορές σμηνών που συναντώνται στη φύση. Στη συνέχεια γίνεται μια περιληπτική αναφορά στους κύριους αντιπροσώπους των αλγορίθμων νοημοσύνης σμηνών άλλα και των γενετικών αλγορίθμων δίνοντας ιδιαίτερη βαρύτητα στην περιγραφή του αλγορίθμου αποικίας μυρμηγκιών. Κλείνοντας περιγράφεται η εφαρμογή του αλγορίθμου αποικίας μυρμηγκιών στο πεδίο της επεξεργασίας της πληροφορίας και δη στην κατηγοριοποίηση κειμένων.

2 Νοημοσύνη Σμηνών

Η έρευνα στην κοινωνική συμπεριφορά εντόμων έχει προσφέρει στους επιστήμονες που ασχολούνται με υπολογιστικά συστήματα ισχυρές μεθόδους για τη σχεδίαση αλγορίθμων κατανεμημένου ελέγχου και βελτιστοποίησης. Αυτές οι τεχνικές εφαρμόζονται επιτυχώς σε ποικίλα επιστημονικά προβλήματα καθώς και σε προβλήματα μηχανικής. Εκτός από την επίτευξη καλής απόδοσης σε ένα ευρύ φάσμα στατικών 'προβλημάτων', τέτοιες τεχνικές τείνουν να παρουσιάζουν υψηλό βαθμό ευελιξίας και ευρωστίας σε ένα δυναμικό περιβάλλον.

Οι επιστήμονες που μελετούν την εξέλιξη των ζώων (*ethologists*) χρησιμοποιούν την μοντελοποίηση για να καταλάβουν τη ζωική συμπεριφορά. Έρευνες στην κοινωνική συμπεριφορά εντόμων δείχνουν ότι πρότυπα βασισμένα στην αυτό-οργάνωση μπορούν να βοηθήσουν στην εξήγηση του πως προκύπτει η περίπλοκη συμπεριφορά των αποικιών από τις αλληλεπιδράσεις των μεμονωμένων μελών αυτής. Αν και ο στόχος του επιστήμονα (*modeler*) είναι γενικά να καταλάβει τη διαβίωση, ένα μοντέλο μπορεί σε γενικές γραμμές να εξερευνηθεί πέρα από την βιολογικά εύλογη πλευρά του. Αν και η βιολογία δεν ωφελείται απαραιτήτως από μια τέτοια εξερεύνηση, οι επιστήμονες και οι μηχανικοί υπολογιστών είναι σε θέση να μετασχηματίσουν τα πρότυπα της κοινωνικής συλλογικής συμπεριφοράς εντόμων σε χρήσιμους αλγόριθμους βελτιστοποίησης και ελέγχου. Αυτή η νέα γραμμή έρευνας αφορά το μετασχηματισμό, της γνώσης σχετικά με το πώς είδη με έντονη κοινωνική δραστηριότητα συλλογικά λύνουν προβλήματα, σε τεχνητές μεθόδους επίλυσης προβλημάτων [Darken,01]– παράγοντας μια μορφή Τεχνητής Νοημοσύνης, ή Νοημοσύνη σμηνών (*Swarm Intelligence*), στην οποία το βαθύτερο μοντέλο νοημοσύνης είναι η συλλογική νοημοσύνη μιας κοινωνικής αποικίας εντόμων [Bonabeau,99].

2.1 Συμπεριφορά των Σμηνών

Η συμπεριφορά των σμηνών παρουσιάζει μεγάλη ποικιλομορφία εντούτοις κάποιες βασικές συμπεριφορές που αφορούν την κατανομή εργασίας και την επικοινωνία μεταξύ των μελών παρουσιάζουν μεγάλη ομοιότητα [Kennedy,01], [Lai,01].

Ο έλεγχος μιας εργασίας όπως για παράδειγμα η αναζήτηση τροφής δεν είναι ιδιότητα ενός κεντρικού σημείου η ατόμου αλλά κατανέμεται σε όλα τα μέλη που σχετίζονται ή συμμετέχουν στην αποπεράτωση αυτής της εργασίας.

Η επικοινωνία μεταξύ των μελών δεν περιορίζεται σε συγκεκριμένο σημείο ή σημεία (για παράδειγμα στην αποικία) αλλά λαμβάνει χώρα σε οποιοδήποτε σημείο ή χώρο που χρησιμοποιείται για την αποπεράτωση της εργασίας και οποιαδήποτε στιγμή μεταξύ όλων των μελών.

Η συμπεριφορά των μελών μπορεί να διαφέρει μεταξύ τους και μάλιστα η συμπεριφορά μερικών εξ' αυτών μπορεί να παρεκκλίνει σημαντικά από την επιθυμητή. Εντούτοις η συμπεριφορά της κοινωνίας υπερκαλύπτει τις επιμέρους ιδιαιτερότητες με αποτέλεσμα να επιτυγχάνεται ισορροπία που οδηγεί στο επιθυμητό αποτέλεσμα. Η όλη διαδικασία αποπεράτωσης εργασίας μπορεί να εξαρτάται από εξωγενείς παράγοντες αλλά η δομή της κοινωνίας και οι κανόνες επικοινωνίας μεταξύ των μελών παρέχουν τους απαραίτητους μηχανισμούς προσαρμογής της διαδικασίας στις νέες συνθήκες.

2.1.1 Ενδιαφέρουσες Συλλογικές Διεργασίες

Με τη χρήση των κανόνων συμπεριφοράς οι αποικίες σμηνών αποπερατώνουν διάφορες βασικές αλλά και σύνθετες διεργασίες όπως είναι:

- Η δημιουργία αποικιών και η συντήρηση αυτών
- Κατάτμηση εργασίας και προσαρμοστική κατανομή πόρων στις επιμέρους διεργασίες Τμήμα της εργασίας και προσαρμοστική κατανομή στόχου
- Ανακάλυψη τροφής και εύρεση βέλτιστης διαδρομής μεταξύ της αποικίας και της τροφής
- Συγκέντρωση και ταξινόμηση (π.χ., νεκροί οργανισμοί, αυγά)
- Σχηματισμός δομών
- Στρατολόγηση για την προμήθεια τροφής
- Ομαδική Μεταφορά (π.χ., τρόφιμα)

2.2 Αυτό – Οργάνωση

Η απάντηση στο βασικό ερώτημα για το πώς οι συμπεριφορά των κοινωνιών σμηνών οδηγεί στην επιτυχή αποπεράτωση βασικών αλλά και πολύπλοκων εργασιών, δίνεται με την εισαγωγή του όρου της «αυτό – οργάνωσης».

Η αυτό – οργάνωση ορίζεται ως ένα σύνολο δυναμικών μηχανισμών μέσω των οποίων η δομή και η συμπεριφορά του συνόλου εμφανίζεται ως το αποτέλεσμα των αλληλεπιδράσεων μεταξύ των μελών αυτού του συνόλου. Οι κανόνες που διευκρινίζουν τις αλληλεπιδράσεις μεταξύ των μελών του συνόλου, εξαρτώνται αποκλειστικά και μόνο από τις τοπικές πληροφορίες που συναλλάσσουν μεταξύ τους τα μέλη του συνόλου. Το σύνολο αυτών των κανόνων είναι μια γηγενή ιδιότητα του συνόλου και δεν επιβάλλεται από εξωτερική πηγή απόφασης [Bonabeau,97].

2.2.1 Χαρακτηριστικά της Αυτό – Οργάνωσης

Σε ένα σύστημα που βασίζεται στην αυτό – οργάνωση, εμφανίζεται ένα σύνολο διαδικασιών οι οποίες είναι απαραίτητες στην κατανόηση της λειτουργίας του συστήματος.

- Πολλαπλές αλληλεπιδράσεις και ανταλλαγές πληροφορίας μεταξύ των μελών του συστήματος
- ενίσχυση των διακυμάνσεων και του τυχαίου
- Θετική ανατροφοδότηση
- Αρνητική ανατροφοδότηση

Η ροή πληροφοριών μεταξύ των μελών του συστήματος, είναι συνεχής και γίνεται σε όλο το εύρος που καλύπτει το σύστημα [Franks,92]. Με αυτόν τον τρόπο, όλα τα μέλη έχουν την απαραίτητη γνώση για το πως θα συμμετέχουν με τη μέγιστη δυνατή απόδοση στην αποπεράτωση της εργασίας. Με την αδιάκοπη διακίνηση της πληροφορίας σε όλα τα μέλη του συστήματος, παρέχεται η δυνατότητα της άμεση ανταπόκρισης αυτού σε ανάγκες που προκύπτουν καθ' όλη τη διάρκεια της εργασίας, για την αποκατάσταση ισορροπίας. Η κύρια παρέμβαση του συστήματος για την αποκατάσταση της ισορροπίας, γίνεται κατά κύριο λόγο με τη χρήση ανατροφοδότησης. Η ανατροφοδότηση ανάλογα, μπορεί να είναι αρνητική αλλά και θετική.

2.2.2 Στατιστική Απεικόνιση

Γενικά μπορούμε να πούμε ότι όταν ένα σύστημα βασίζεται στην αυτό – οργάνωση η ροή πληροφοριών μεταξύ των μελών του συστήματος αυξάνει έτσι ώστε να επιτευχθεί με τον καλύτερο και αποδοτικότερο τρόπο [Sole,99]. Ουσιαστικά για την ορθή λειτουργία ενός συστήματος αυτό – οργάνωσης απαιτείται η αποθήκευση μεγαλύτερου αριθμού πληροφορίας. Από τη σκοπιά της στατιστικής, η πολυπλοκότητα ενός συστήματος είναι άμεσα εξαρτώμενη από το μέγεθος της πληροφορίας που διαχειρίζεται. Πιο συγκεκριμένα, θεωρώντας ένα σύστημα με ένα σύνολο αιτιωδών καταστάσεων S και ένα σύνολο μ εισόδων, ισχύουν οι παρακάτω προτάσεις:

- Η στατιστική πολυπλοκότητα $C_\mu(S)$ των αιτιωδών καταστάσεων S, ορίζεται ως η τιμή της εντροπίας H[S] πάνω στην κατανομή πληροφοριών μ στις εισόδους των καταστάσεων.
- Η στατιστική πολυπλοκότητα $C_\mu(S)$ είναι η μέση τιμή του μεγέθους της πληροφορίας που διατηρείται ως είσοδο σε κάθε κατάσταση.
- Για μια διαδικασία, η στατιστική πολυπλοκότητα αυξάνει όταν εφαρμόζεται η αυτό - οργάνωση $C_\mu(S_t) < C_\mu(S_t+T)$.

2.2.3 Είδη της Αυτό – Οργάνωσης

Όπως προαναφέρθηκε, σε ένα σύστημα αυτό – οργάνωσης απαραίτητο συστατικό είναι η επικοινωνία μεταξύ των μελών. Η επικοινωνία γίνεται με ποικίλους τρόπους και εξαρτάται συνήθως από τη δομή του συστήματος και επηρεάζει το είδος της αυτό – οργάνωσης. Δυο είναι οι σημαντικότεροι τρόποι επικοινωνίας οι οποίοι προσδιορίζουν και τα αντίστοιχα είδη αυτό – οργάνωσης, η άμεση και η έμμεση επικοινωνία.

Άμεση επικοινωνία

Κατά την άμεση επικοινωνία τα μέλη της κοινωνίας ανταλλάσσουν πληροφορίες με άμεσο τρόπο και τη χρήση κάποιας ή κάποιων αισθήσεων (αφή, γεύση), όπως για παράδειγμα με ανταλλαγή χημικών ουσιών ή με άγγιγμα κεραιών. Με άλλα λόγια το ένα μέλος μεταφέρει την γνώση και την εμπειρία του στο άλλο μέλος και αντίστροφα, τη στιγμή που γίνεται η συνάντηση. Η άμεση επικοινωνία καλείται και επικοινωνία τύπου εκπομπής (Broadcast-like).

Έμμεση επικοινωνία

Κατά την έμμεση επικοινωνία δυο μέλη του συνόλου δεν αλληλεπιδρούν ταυτόχρονα, αλλά το ένα τροποποιεί το περιβάλλον αφήνοντας κάποιο ίχνος συνήθως κάποια χημική ουσία, ενώ το δεύτερο μέλος σε κάποια άλλη χρονική στιγμή ανταποκρίνεται σ' αυτό το ίχνος. Ουσιαστικά το χημικό ίχνος αναπαριστά την εμπειρία του μέλους που το τοποθετεί, το οποίο είναι διαθέσιμο προς «ανάγνωση» και χρήση από τα υπόλοιπα μέλη που θα βρεθούν κάποια χρονική στιγμή στο σημείο. Ο κοινός όρος που χρησιμοποιείται για την έμμεση επικοινωνία καλείται επικοινωνία στίγματος (stigmergy).

3 Αλγόριθμοι Νοημοσύνης Σμηνών

Ανάλογα με το είδος επικοινωνίας που χρησιμοποιούν οι αλγόριθμοι νοημοσύνης σμηνών, χωρίζονται σε δύο κατηγορίες. Στην πρώτη κατηγορία ανήκουν οι αλγόριθμοι που βασίζονται στην άμεση επικοινωνία και ο κύριος αντιπρόσωπος είναι ο αλγόριθμος «Particle Swarm Optimization». Η δεύτερη κατηγορία χαρακτηρίζεται από τη χρήση της έμμεσης επικοινωνίας που δίνεται και με τον όρο «Stigmergy». Ο πιο σημαντικός είναι οι αλγόριθμοι αποικίας μυρμηγκιών. Παρακάτω αναφέρονται οι κύριοι αντιπρόσωποι των παραπάνω κατηγοριών αλλά και μια άλλη οικογένεια αλγορίθμων, οι γενετικοί αλγόριθμοι που βρίσκουν ένα ευρύ πεδίο εφαρμογών.

3.1 Αλγόριθμοι Αποικίας Μυρμηγκιών

Η οικογένεια αλγορίθμων αποικίας μυρμηγκιών αποτελεί ουσιαστικά τον κύριο αντιπρόσωπο των αλγορίθμων που στηρίζονται στην έμμεση επικοινωνία, και εμπνεύστηκε από τη βιολογία και πιο συγκεκριμένα από τη μελέτη της συμπεριφοράς των μυρμηγκιών σε μια οργανωμένη κοινωνία όπως είναι η αποικία. Η διατήρηση της ζωής σε μια αποικία μυρμηγκιών καθώς και η εξέλιξή της βασίζεται σε μια πολύ καλή οργάνωση και κατανομή εργασιών σε όλα τα μέλη της αποικίας. Για παράδειγμα κάποια μέλη αναλαμβάνουν την εύρεση και μεταφορά του φαγητού κάποια άλλα μέλη την προάσπιση της αποικίας άλλα την επέκταση αυτής κλπ. Παρόλο που δεν υπάρχει ένας ενιαίος συντονισμός για την εκτέλεση των εργασιών, αυτές αποπερατώνονται με μια αξιοθαύμαστη λεπτομέρεια. Το γεγονός αυτό οφείλεται στο εξελιγμένο σύστημα που χρησιμοποιούν για την μεταξύ τους επικοινωνία το οποίο στηρίζεται στην ανταλλαγή χημικών ερεθισμάτων «σημάτων». Καθώς βαδίζουν τα μυρμήγκια, εναποθέτουν στο έδαφος μια χημική ουσία η οποία καλείται φερομόνη. Η εναπόθεση της φερομόνης οδηγεί σε δημιουργία μονοπατιών. Τα υπόλοιπα μυρμήγκια αν και σχεδόν τυφλά ακολουθούν τα μονοπάτια φερομόνης. Τα μονοπάτια που περιέχουν μεγαλύτερες ποσότητες φερομόνης έχουν μεγαλύτερη πιθανότητα να ακολουθηθούν από τα μυρμήγκια. Επιπρόσθετα η φερομόνη εξατμίζεται με τον χρόνο. Αυτό έχει ως αποτέλεσμα τα μονοπάτια τα οποία δεν προτιμώνται, να εξαφανίζονται με τον χρόνο. Αυτό τους δίνει τη δυνατότητα να βρίσκουν τον δρόμο προς την τροφή ή πίσω προς την αποικία. Εκτενής αναφορά για τον τρόπο λειτουργίας των αλγορίθμων αποικίας μυρμηγκιών θα γίνει στις επόμενες παραγράφους αυτού του κεφαλαίου και αποτελεί τη θεμελιώδη βάση της παρούσας διατριβής.

3.2 Αλγόριθμος Βελτιστοποίησης Σμήνους Μορίων

Οι αλγόριθμοι νοημοσύνης σμηνών εμπνεύστηκαν από τη μελέτη και μοντελοποίηση της συλλογικής συμπεριφοράς των αποικιών των εντόμων, των σμηνών, και άλλων ζωικών κοινωνιών, και χρησιμοποιούν την αυτό-οργάνωση και την κατανομή εργασίας για την επίλυση προβλημάτων. Η μέθοδος βελτιστοποίησης σμήνους μορίων (PSO) αποτελεί μέλος του ευρύτερου τομέα νοημοσύνης σμηνών. Οι πρώτοι που εισήγαγαν αυτή τη μέθοδο ήταν οι James Kennedy και Russell Eberhart [Kennedy,95]. Όπως και οι γενετικοί αλγόριθμοι, η προσέγγιση λύσης βασίζεται στον πληθυσμό ατόμων μελών. Η εύρεση λύσης στηρίζεται στην τροποποίηση της κυκλοφορίας και της θέσης των ατόμων του συνόλου, που αποκαλούνται μόρια, χωρίς την παραγωγή εκ' νέου του πληθυσμού. Η ταχύτητα και η τροχιά κάθε μορίου, τροποποιούνται με βάση την καλύτερη θέση που κατείχαν σε περασμένο χρόνο.

3.2.1 Θεωρητικό Μοντέλο

Θεωρώντας έναν χώρο *D* διαστάσεων και ένα σμήνος *N* μορίων, η θέση ενός μορίου *i* με $i = 1,2,...D$ στο χώρο μπορεί να οριστεί ως ένα διάνυσμα $X_i = (x_{i1}, x_{i2}..., x_{iD})$, με τις επιμέρους θέσεις σε κάθε διάσταση $d = 1,2,..., D$. Αντίστοιχα, η ταχύτητα του *i* μορίου μπορεί να οριστεί ως ένα αντίστοιχο διάνυσμα $U_i = (u_{i1}, u_{i2}..., u_{iD})$. Ως $p_i = (p_{i1}, p_{i2}..., p_{iD})$ ορίζεται η καλύτερη θέση του *i* μορίου στο παρελθόν. Οι εξισώσεις που ορίζουν τη θέση και την ταχύτητα κάθε μορίου του συνόλου στο χώρο, δίνονται από τις παρακάτω σχέσεις:

$$x_{id} = x_{id} + u_{id} \qquad \textbf{Σχέση 1}$$

$$u_{id} = \chi(w_k u_{id} + \phi_1 r_1 (p_{id} - x_{id}) + \phi_2 r_2 (p_{gd} - x_{id})) \qquad \textbf{Σχέση 2}$$

Όπου w η τιμή αδράνειας, g είναι το μόριο με την καλύτερη θέση, φ_1 και φ_2 δυο θετικές σταθερές που συχνά αποκαλούνται ως *γνωστική* και *κοινωνική* παράμετρος αντίστοιχα, ενώ r_1 και r_2 είναι δυο τυχαίοι αριθμοί μεταξύ 0 και 1. Ο όρος χ είναι μια τιμή συστολής η οποία δίνεται από τη σχέση:

$$\chi = \frac{2}{\left|2 - \phi - \sqrt{\phi^2 - 4\phi}\right|} \qquad \textbf{Σχέση 3}$$

Με $\varphi = \varphi_1 + \varphi_2$ και $\varphi > 4$.

Η εξίσωση της ταχύτητας (2), υπολογίζει τη νέα ταχύτητα του κάθε μορίου σε συνάρτηση με την τρέχουσα ταχύτητα του μορίου. Ο δεύτερος όρος είναι συνάρτηση της διαφοράς της καλύτερης θέσης του μορίου στο παρελθόν με την τρέχουσα θέση του. Αντίστοιχα ο τρίτος όρος είναι μια συνάρτηση της διαφοράς της καλύτερης θέσης όλων των μορίων στο παρελθόν με την τρέχουσα θέση του. Οι τρεις όροι της εξίσωσης 2 πολλαπλασιάζονται με μια τιμή συστολής, η οποία δίνει τη δυνατότητα στον αλγόριθμο να περιορίσει και να ελέγξει τις ταχύτητες.

Ο τιμή της αδράνειας *w* στον πρώτο όρο, είναι μια παράμετρος που χρησιμοποιείται για τον έλεγχο της συμμετοχής της προηγούμενης ταχύτητας στην τρέχουσα ταχύτητα του μορίου. Οι άλλοι δυο όροι ουσιαστικά προσδιορίζουν το ποσοστό της συνολικής (εξερεύνηση) και τοπικής (εκμετάλλευση) ικανότητας του μορίου. Συνήθως, στην αρχή της λειτουργίας του αλγορίθμου, ή τιμή της αδράνειας είναι μεγάλη και στη συνέχεια φθίνει. Αυτή η μέθοδος ενισχύει τις δυνατότητες εξερεύνησης του αλγορίθμου νωρίς στη διαδικασία και βαθμιαία τις μεταβάσεις σε έναν τρόπο εκμετάλλευσης τοπικών περιοχών ενδιαφέροντος.

3.2.2 Δομή του Αλγορίθμου

Αν και έχουν προταθεί διάφορες προσεγγίσεις στον τρόπο λειτουργίας των αλγορίθμων [Eberhart,01], [Løvbjerg,01], [Shi,98] βελτιστοποίησης σμήνους μορίων, ωστόσο όλοι στηρίζονται πάνω σε ένα γενικό πλαίσιο. Οι διαφορές στους έγκειται κυρίως στον τρόπο αρχικοποίησης του αλγορίθμου. Πάντως τα βασικά βήματα όλων των προσεγγίσεων είναι:

1. Αρχικοποίηση του αλγορίθμου. Ορίζεται ως αρχική χρονική στιγμή 0. Δημιουργία αριθμού *N* ατόμων και αρχικοποίηση της θέσης και της ταχύτητάς τους μέσα στο χώρο *D* διαστάσεων.
2. Δημιουργία αντικειμενικής συνάρτησης για την αξιολόγηση των μορίων.
3. Αρχικοποίηση των βέλτιστων κάθε μορίου. Αυτό ορίζεται ως η τρέχουσα θέση του.

4. Αρχικοποίηση του συνολικού βέλτιστου. Αυτό ορίζεται ως η τρέχουσα θέση του μορίου με την καλύτερη θέση.
5. Αύξηση της τιμής του χρόνου κατά ένα, και ενημέρωση της τιμής της αδράνειας
6. Ανανέωση της ταχύτητας των μορίων με βάση τη σχέση 2.
7. Ανανέωση της θέσης των μορίων με βάση τη σχέση 1 και χρήση της ανανεωμένης σχέσης 2 από το βήμα 6.
8. Υπολογισμός αντικειμενικής συνάρτησης και αξιολόγηση κάθε μορίου.
9. Ανανέωση των βέλτιστων κάθε μορίου με βάση την υπάρχουσα θέση του και την καλύτερη θέση στο παρελθόν.
10. Ανανέωση του συνολικού βέλτιστου με βάση το προηγούμενο ολικό βέλτιστο και το βέλτιστο των τρεχόντων βέλτιστων των μορίων.
11. Έλεγχος τερματισμού. Αν ικανοποιείται το κριτήριο η τα κριτήρια τερματισμού, ο αλγόριθμος υπολογίζει την τελική λύση και τερματίζει τη λειτουργία του, αλλιώς επαναλαμβάνει τα βήματα από το 5 και μετά.

Αρχικοποίηση Αλγορίθμου

Ένα σημαντικό τμήμα του αλγορίθμου για το οποίο έχουν προταθεί πολλές προσεγγίσεις είναι η αρχικοποίηση του πληθυσμού των μορίων του σμήνους. Ωστόσο δυο είναι οι σημαντικότερες μεθοδολογίες που αναλύονται παρακάτω:

Μέθοδοι τυχαίας αρχικοποίησης

Ο συνήθης τρόπος αρχικοποίησης του πληθυσμού των σμηνών είναι με τη χρήση μιας γεννήτριας ψευδό-τυχαίων αριθμών. Αυτό συνήθως οδηγεί σε μη ομοιόμορφη κατανομή των μορίων του σμήνους στο χώρο των D διαστάσεων. Αν και έχουν προταθεί τρόποι αρχικοποίησης, με διάφορες κατανομές, οι οποίοι οδηγούν σε πιο ομοιόμορφη κατανομή των μορίων στο χώρο, αλλά και σε καλύτερη ποιότητα αποτελεσμάτων, ωστόσο έχουν το μειονέκτημα ότι υστερούν σε προγραμματιστική υλοποίηση. Παρόμοια με την αρχικοποίηση της θέσης γίνεται και η αρχικοποίηση της ταχύτητας των μορίων με τη διαφορά ότι κατανέμονται ομοιόμορφα μεταξύ μιας ελάχιστης και μιας μέγιστης τιμής.

Μέθοδοι ευρετικής αρχικοποίησης

Μια άλλη διαφορετική προσέγγιση για την αρχικοποίηση του πληθυσμού αποτελεί η χρήση ευρετικής συνάρτησης. Συνήθως η ευρετική συνάρτηση που χρησιμοποιείται πηγάζει από της ανάγκες του εκάστοτε προβλήματος και έχει το πλεονέκτημα να συγκεκριμενοποιεί τον αλγόριθμο στο υπάρχον πρόβλημα.

Κριτήρια Τερματισμού

Συνήθως τα κριτήρια τερματισμού της λειτουργίας του αλγορίθμου εξαρτώνται από τις ανάγκες του προβλήματος αλλά και του χρήστη. Παρόλα αυτά τα κυριότερα και τα πιο χρησιμοποιήσιμα είναι:

- Ο αριθμός των επαναλήψεων δεν μπορεί να υπερβαίνει έναν μέγιστο αριθμό C_{max}
- Ορίζεται ένας μέγιστος πεπερασμένος αριθμός χρονικών περασμάτων
- Ορίζεται μια μέγιστη χρονική υπολογιστική ισχύς που θα καταναλωθεί από τον επεξεργαστή.

- Ορίζεται ένας μέγιστος αριθμός επαναλήψεων μέσα στον οποίο η καλύτερη λύση δε θα αλλάζει από επανάληψη σε επανάληψη.

3.3 Γενετικοί Αλγόριθμοι

Ο γενετικός αλγόριθμος (Genetic Algorithm GA), ο οποίος εφευρέθηκε από τον [Holland,75], είναι μια κατευθυνόμενη στοχαστική τεχνική αναζήτησης που μπορεί να βρει την συνολική βέλτιστη λύση σε πολυσύνθετους χώρους αναζήτησης με πολλές διαστάσεις. Ένας γενετικός αλγόριθμος (GA) μοντελοποιείται πάνω στη φυσική εξέλιξη δεδομένου ότι οι τελεστές που χρησιμοποιεί επηρεάζονται από την διαδικασία εξέλιξης. Αυτοί οι τελεστές, γνωστοί σαν γενετικοί τελεστές, εφαρμόζονται σε άτομα ενός πληθυσμού μέσω μερικών γενεών για να βελτιώσουν σταδιακά την καταλληλότητά τους. Τα άτομα σε έναν πληθυσμό παρομοιάζονται με τα χρωμοσώματα και συνήθως αναπαριστάνονται σαν ακολουθίες από δυαδικούς αριθμούς.

Η εξέλιξη ενός πληθυσμού ατόμων κατευθύνεται από την «schema theorem» [Holland,75]. Ένα «σχήμα» αναπαριστά ένα σύνολο από άτομα, π.χ. ένα υποσύνολο του πληθυσμού, από την άποψη της ομοιότητας των bits σε ορισμένες θέσεις αυτών των ατόμων. Για παράδειγμα, το σχήμα 1*0* περιγράφει το σύνολο των ατόμων των οποίων το πρώτο και το τρίτο bit είναι 1 και 0,αντίστοιχα.Εδώ,το σύμβολο * σημαίνει ότι οποιαδήποτε τιμή θα ήταν αποδεκτή. Με άλλα λόγια, οι τιμές των bits στις θέσεις που περιέχουν * θα μπορούσαν να είναι είτε 0 ή 1 σε μια δυαδική ακολουθία. Ένα σχήμα χαρακτηρίζεται από δύο παραμέτρους που είναι το καθοριζόμενο μήκος και η τάξη. Το καθοριζόμενο μήκος είναι το μήκος μεταξύ του πρώτου και του τελευταίου bit που έχουν αμετάβλητες τιμές. Η τάξη του σχήματος είναι ο αριθμός των bits με καθορισμένες τιμές. Σύμφωνα με τη θεωρία του σχήματος, η κατανομή του σχήματος μέσω του πληθυσμού από μία γενιά σε μία άλλη εξαρτάται από την τάξη του, που καθορίζει το μήκος και την καταλληλότητα.

Οι γενετικοί αλγόριθμοι δεν χρησιμοποιούν πολλές γνώσεις από το πρόβλημα που βελτιστοποιείται και δεν χειρίζονται άμεσα τις παραμέτρους του προβλήματος. Δουλεύουν με κώδικες που αναπαριστούν τις παραμέτρους. Επομένως, το πρώτο θέμα σε μία εφαρμογή GA είναι πως κωδικοποιούμε το υπό μελέτη πρόβλημα, π.χ. πως αναπαριστούμε τις παραμέτρους ενός προβλήματος. Οι GA δουλεύουν με ένα πληθυσμό πιθανών λύσεων, όχι μόνο μια πιθανή λύση, και το δεύτερο θέμα είναι επομένως πως δημιουργείται ο αρχικός πληθυσμός των πιθανών λύσεων. Το τρίτο θέμα σε μία εφαρμογή GA είναι πως επιλέγεται ή εφευρίσκεται ένα κατάλληλο σύνολο γενετικών τελεστών,. Τέλος, όπως συμβαίνει και σε άλλους αλγορίθμους αναζήτησης, οι GA πρέπει να γνωρίζουν τα χαρακτηριστικά των λύσεων που έχουν ήδη βρεθεί για να τις βελτιώσουν περαιτέρω. Άρα, υπάρχει η ανάγκη για μια διασύνδεση μεταξύ του περιβάλλοντος του προβλήματος και του ίδιου του GA για να έχει τη δυνατότητα ο αλγόριθμος να έχει αυτή τη γνώση. Ο σχεδιασμός μιας τέτοιας διασύνδεσης μπορεί να θεωρηθεί σαν το τέταρτο σημαντικό ζήτημα.

3.3.1 Αναπαράσταση Πληθυσμού

Οι GA λειτουργούν πάνω σε έναν αριθμό πιθανών λύσεων, όπως έχει αναφερθεί στην παραπάνω παράγραφο, που ονομάζεται πληθυσμός, και αποτελείται από κάποια κωδικοποίηση του συνόλου των παραμέτρων. Οι παράμετροι προς βελτιστοποίηση συνήθως αναπαριστάνονται με έναν τύπο ακολουθίας (string) αφού οι γενετικοί τελεστές είναι κατάλληλοι γι' αυτόν τον τύπο αναπαράστασης. Υπάρχουν δύο γενικές μέθοδοι αναπαράστασης για αριθμητικά προβλήματα βελτιστοποίησης [Michalewicz,92], [Davis - 1991].

Μέθοδος αναπαράστασης με μια δυαδική ακολουθία

Σε αυτή την μέθοδο αναπαράστασης, κάθε μεταβλητή απόφασης στο σύνολο παραμέτρων κωδικοποιείται ως μια δυαδική σειρά και αυτές συνδέονται για να σχηματίσουν ένα χρωμόσωμα. Διάφορα σχήματα δυαδικής κωδικοποίησης μπορούν να βρεθούν στη βιβλιογραφία, όπως είναι η ομοιόμορφη κωδικοποίηση και η Gray scale κωδικοποίηση. Η χρήση της κωδικοποίησης Gray έχει υποστηριχτεί σαν μια μέθοδο για να αντιμετωπιστεί η κρυμμένη εύνοια (representational bias) στη συμβατική δυαδική αναπαράσταση αφού η απόσταση Hamming μεταξύ διπλανών τιμών είναι σταθερή. Εμπειρικά στοιχεία του Caruana και του Schaffer δείχνουν ότι μεγάλες αποστάσεις Hamming στον χάρτη αναπαράστασης μεταξύ διπλανών τιμών, όπως συμβαίνει στην κανονική δυαδική αναπαράσταση, μπορούν να οδηγήσουν τη διαδικασία αναζήτησης σε μη αξιόπιστη ή μπορεί να θεωρηθεί ανίκανη για να εντοπίσει αποτελεσματικά το σφαιρικό ελάχιστο. Μια περαιτέρω προσέγγιση του [Schmitendorgf,92], είναι η χρήση λογαριθμικής κλίμακας στην μετατροπή των δυαδικά κωδικοποιημένων χρωμοσωμάτων στις πραγματικές φαινοτυπικές τιμές τους. Αν και η ακρίβεια των τιμών των παραμέτρων είναι ενδεχομένως λιγότερο σταθερή όσον αφορά το επιθυμητό εύρος, σε προβλήματα όπου το εύρος των εφικτών παραμέτρων είναι άγνωστο, ένα μεγαλύτερο διάστημα αναζήτησης μπορεί να καλυφθεί με τον ίδιο αριθμό bits σε σχέση με ένα γραμμικό σχήμα απεικόνισης επιτρέποντας την μείωση του υπολογιστικού κόστους για την αναζήτηση αγνώστων περιοχών αναζήτησης.

Χρήση διανύσματος αριθμών

Η δεύτερη μέθοδος κωδικοποίησης είναι η χρησιμοποίηση ενός διανύσματος ακεραίων ή πραγματικών αριθμών, με τον κάθε ακέραιο ή πραγματικό αριθμό να αναπαριστά μια μοναδική παράμετρο. Για μερικές περιοχές ενός προβλήματος, υποστηρίζεται ότι η δυαδική αναπαράσταση είναι στην πραγματικότητα παραπλανητική γιατί κρύβει τη φύση της αναζήτησης . Στο πρόβλημα επιλογής υποσυνόλων, παραδείγματος χάρη, η χρήση μιας αναπαράστασης ακέραιων αριθμών και look-up πινάκων παρέχει έναν κατάλληλο και φυσικό τρόπο να εκφράσουμε την απεικόνιση (mapping) από την αναπαράσταση στην περιοχή του προβλήματος.

3.3.2 Δημιουργία Αρχικού Πληθυσμού

Υπάρχουν δύο τρόποι σχηματισμού του αρχικού πληθυσμού. Ο πρώτος προκύπτει από την τυχαία παραγωγή λύσεων που δημιουργούνται από μια γεννήτρια τυχαίων αριθμών. Αυτή η μέθοδος προτιμάται για προβλήματα για τα οποία δεν υπάρχει γνώση από πριν ή για την αξιολόγηση της επίδοσης ενός αλγορίθμου.

Η δεύτερη μέθοδος χρησιμοποιεί γνώση από πριν σχετικά με το δοθέν πρόβλημα βελτιστοποίησης. Χρησιμοποιώντας αυτή τη γνώση, ένα σύνολο προδιαγραφών λαμβάνεται υπόψη και λύσεις οι οποίες ικανοποιούν αυτές τις προδιαγραφές συλλέγονται για να σχηματίσουν έναν αρχικό πληθυσμό. Σε αυτή την περίπτωση, ο GA αρχίζει με ένα σύνολο κατά προσέγγιση γνωστών λύσεων και άρα συγκλίνει σε μια βέλτιστη λύση σε λιγότερο χρόνο σε σχέση με την προηγούμενη μέθοδο.

3.3.3 Γενετικοί Τελεστές

Υπάρχουν τρεις γενικοί γενετικοί τελεστές: η επιλογή, η διασταύρωση και η μετάλλαξη (mutation). Ένας πρόσθετος τελεστής αναπαραγωγής, η αντιστροφή, εφαρμόζεται μερικές φορές. Μερικοί από αυτούς τους τελεστές επηρεάστηκαν από τη φύση. Δεν είναι απαραίτητο να χρησιμοποιηθούν όλοι αυτοί οι τελεστές σε έναν GA. Η επιλογή ή ο σχεδιασμός των τελεστών εξαρτάται από το πρόβλημα και το σχήμα αναπαράστασης που χρησιμοποιείται. Λόγου χάρη, οι τελεστές που προορίζονται για δυαδικές ακολουθίες δεν μπορούν άμεσα να κωδικοποιηθούν με ακέραιους ή πραγματικούς αριθμούς.

Επιλογή

Ο σκοπός της διαδικασίας επιλογής είναι η αναπαραγωγή περισσότερων αντιγράφων ατόμων των οποίων οι τιμές καταλληλότητας (fitness) είναι υψηλότερες από τα αντίγραφα των οποίων οι τιμές καταλληλότητας είναι χαμηλές. Η διαδικασία επιλογής έχει σημαντική επίδραση στην οδήγηση της αναζήτησης προς μια περιοχή με καλές προοπτικές και στην ανεύρεση καλών λύσεων σε σύντομο χρονικό διάστημα. Όμως, η ποικιλομορφία του πληθυσμού πρέπει να διατηρηθεί για να αποφευχθεί η πρόωρη σύγκλιση και να προσεγγιστεί η συνολική βέλτιστη λύση. Στους GA υπάρχουν δύο κυρίως διαδικασίες επιλογής: η αναλογική επιλογή και η επιλογή που βασίζεται σε στοίχιση.

Η επιλογή ανάλογα με την ικανότητα χρησιμοποιείται στην αρχική παρουσίαση του Holland. Αναφέρει ότι οι φορές που ένα άτομο αναμένεται να αναπαραγάγει υπολογίζεται από το πηλίκο της ικανότητας του ατόμου με το μέσο όρο της ικανότητας του πληθυσμού. Το ίδιο σχήμα συναντάται στη φύση και οι βιολόγοι αναφέρονται σε αυτό ως επιλογή βιωσιμότητας (Viability Selection).

Υπάρχουν δύο δειγματοληπτικές μέθοδοι για αυτήν την επιλογή γνωστές ως δειγματοληψία τροχού ρουλέτας (Roulette Wheel Sampling) και στοχαστική καθολική δειγματοληψία (Stochastic Universal Sampling). Η πρώτη αναθέτει τα μέρη - φέτες ενός τροχού σε κάθε άτομο. Το μέγεθος κάθε φέτας είναι ανάλογο με την ικανότητα του ατόμου. Μετά από μια τυχαία περιστροφή του τροχού για την επιλογή ενός ατόμου για την επόμενη γενιά, τα άτομα που βρίσκονται σε slots με μεγάλα εύρη τα οποία συμβολίζουν υψηλές τιμές καταλληλότητας θα έχουν μεγαλύτερη πιθανότητα να επιλεγούν. Με την περιστροφή του τροχού Ν φορές θα επιλεγούν Ν άτομα (με επαναλήψεις) που θα αποτελέσουν την ομάδα των πιθανών γονέων.

Ωστόσο, λόγω της στοχαστικής φύσης της δειγματοληψίας, είναι δυνατόν, ο τροχός να σταματήσει στην μικρή φέτα του χειρότερου ατόμου (στη χειρότερη περίπτωση, αυτό μπορεί να συμβεί Ν φορές). Επίσης, το μεγάλο μέρος του καταλληλότερου ατόμου είναι πιθανό να χάνεται κάθε φορά.

Τα παραπάνω πιθανά προβλήματα οδήγησαν στην πρόταση της καθολικής στοχαστικής δειγματοληπτικής μεθόδου. Εδώ, ο τροχός της ρουλέτας χωρίζεται όπως πριν, αλλά αντί της περιστροφής του τροχού Ν φορές με έναν δείκτη, περιστρέφεται μια φορά μόνο αλλά με Ν δείκτες που καταλαμβάνουν τον ίδιο χώρο. Αυτό εξασφαλίζει ότι τα καλά άτομα δεν μπορούν να λείψουν στην ομάδα από ένα λάθος της φύσης.

Εάν η ικανότητα ενός ατόμου που διαιρείται με τη μέση ικανότητα του πληθυσμού συμβολίζεται με x και xi είναι το ακέραιο μέρος του x, τότε η στοχαστική καθολική δειγματοληψία εγγυάται να επιλέξει αυτό το άτομο τουλάχιστον xi και το πολύ-πολύ xi +1 φορές.

Διασταύρωση (Crossover) ή Επανασύνδεση (Recombination)

Χρησιμοποιείται για τη δημιουργία δύο νέων ατόμων (παιδιά) από δύο άτομα που υπάρχουν (γονείς) τα οποία λαμβάνονται από τον τρέχοντα πληθυσμό με την λειτουργία της επιλογής. Υπάρχουν πολλοί τρόποι για να γίνει αυτό. Μερικές γνωστές λειτουργίες διασταύρωσης είναι η διασταύρωση ενός σημείου, η διασταύρωση δύο σημείων, η κυκλική διασταύρωση και η ομοιόμορφη διασταύρωση.

Μία διασταύρωση ενός σημείου είναι η πιο απλή λειτουργία διασταύρωσης. Δύο άτομα επιλέγονται τυχαία ως γονείς από την ομάδα των ατόμων που έχει σχηματιστεί από την διαδικασία επιλογής και κόβονται σε ένα τυχαία επιλεγόμενο σημείο. Οι ουρές, που είναι τα μέρη

μετά το σημείο αποκοπής, αντιμετατίθενται και δύο νέα άτομα (παιδιά) παράγονται. Πρέπει να σημειώσουμε ότι αυτή η λειτουργία δεν αλλάζει τις τιμές των bits.

Στην ομοιόμορφη διασταύρωση μια μάσκα διασταύρωσης που έχει το ίδιο μήκος με τις δομές χρωμοσωμάτων δημιουργείται τυχαία, και η ισοτιμία (parity) των bits στην μάσκα δείχνει ποιος από τους δύο γονείς θα δώσει bits στους απόγονους.

Η ομοιόμορφη διασταύρωση, όπως και η διασταύρωση σε πολλά σημεία, έχει θεωρηθεί ότι μειώνει την εύνοια (bias) που συνδέεται με το μήκος της δυαδικής αναπαράστασης που χρησιμοποιείται και την συγκεκριμένη κωδικοποίηση για ένα δεδομένο σύνολο παραμέτρων. Αυτό βοηθά στην αντιμετώπιση της εύνοιας προς τα κοντά (short) substrings που υπάρχει στην διασταύρωση ενός σημείου χωρίς να είναι αναγκαία η ακριβής κατανόηση της σημασίας των bits των ατόμων στην αναπαράσταση χρωμοσωμάτων. Οι Spears και De Jong στην εργασία τους [DeJong,93] έδειξαν πως η ομοιόμορφη διασταύρωση μπορεί να παραμετροποιηθεί με την εφαρμογή μιας πιθανότητας στην ανταλλαγή των bits. Αυτή η πρόσθετη παράμετρος μπορεί να χρησιμοποιηθεί για να ελέγξει το ποσό της αποσύνθεσης (disruption) κατά τη διάρκεια της επανασύνδεσης χωρίς να εισάγει μια εύνοια (bias) προς το μήκος της αναπαράστασης που χρησιμοποιείται. Όταν χρησιμοποιείται η ομοιόμορφη διασταύρωση με αλληλόμορφα γονίδια (alleles) πραγματικών τιμών, αναφέρεται συνήθως ως διακριτή επανασύνδεση.

Μετάλλαξη (Mutation)

Σε αυτή την διαδικασία, όλα τα άτομα ελέγχονται από bit σε bit και οι τιμές των bit αντιστρέφονται τυχαία σύμφωνα με έναν καθορισμένο ρυθμό. Σε αντίθεση με τη διασταύρωση, αυτή είναι μια μοναδιαία λειτουργία. Δηλαδή, μια ακολουθία παιδιού παράγεται από μία μόνο ακολουθία γονέα. Ο τελεστής μετάλλαξης αναγκάζει τον αλγόριθμο να ψάξει νέες περιοχές. Τελικά, βοηθάει τον GA να αποφύγει την πρόωρη σύγκλιση και να βρει την συνολική βέλτιστη λύση.

Όπως και σε άλλους τελεστές, υπάρχουν πολλές παραλλαγές αυτού του τελεστή στη βιβλιογραφία. Στους GA, η μετάλλαξη εφαρμόζεται τυχαία με χαμηλή πιθανότητα, συνήθως κυμαίνεται μεταξύ των τιμών 0,001 και 0,01, και τροποποιεί στοιχεία στα χρωμοσώματα. Συνήθως θεωρείται ως τελεστής υποβάθρου (backround operator), καθώς η μετάλλαξη θεωρείται συχνά ότι παρέχει μια εγγύηση ότι η πιθανότητα αναζήτησης οποιασδήποτε δεδομένης σειράς δεν θα είναι ποτέ μηδέν και ενεργεί ως δίχτυ ασφάλειας για να ανακτήσει το καλό γενετικό υλικό που μπορεί να χαθεί κατά τις διαδικασίες της επιλογής και της διασταύρωσης.

Αντιστροφή

Αυτός ο πρόσθετος τελεστής χρησιμοποιείται για έναν αριθμό προβλημάτων όπως είναι το πρόβλημα τοποθέτησης κελιού, προβλήματα διάταξης και το πρόβλημα του περιπλανώμενου πωλητή. Ενεργεί επίσης σε ένα άτομο την φορά. Δύο σημεία επιλέγονται τυχαία από ένα άτομο και το μέρος της ακολουθίας μεταξύ αυτών των δύο σημείων αντιστρέφεται.

4 Οικογένεια Αλγορίθμων Αποικίας μυρμηγκιών (ACO)

Στην ενότητα αυτή γίνεται μια εκτενής αναφορά στη δομή και τα χαρακτηριστικά των αλγορίθμων αποικίας μυρμηγκιών. Πιο συγκεκριμένα γίνεται ένας παραλληλισμός με τα φυσικά μυρμήγκια και στην ιδέα χρήσης των πρακτόρων ως τεχνητά μυρμήγκια. Στη συνέχεια αναλύεται το θεωρητικό μοντέλο λειτουργίας που είναι η βάση για όλες τις παραλές και τροποποιήσεις των αλγορίθμων αποικίας μυρμηγκιών.

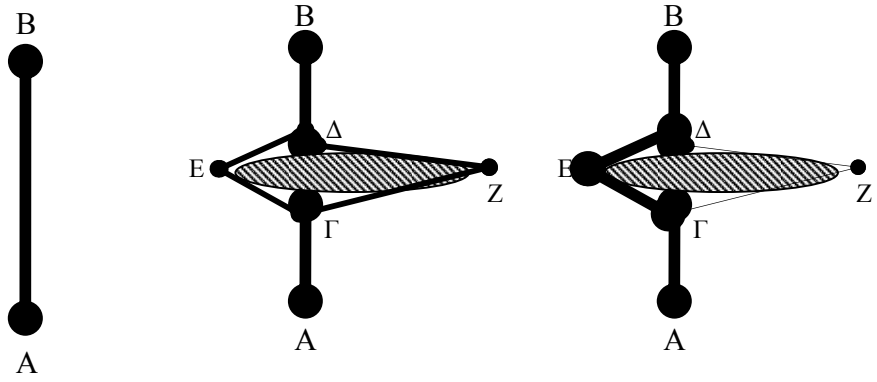
4.1 Γενικά Χαρακτηριστικά των Αλγορίθμων Αποικίας Μυρμηγκιών

4.1.1 Πραγματικά Μυρμήγκια

Η οικογένεια αλγορίθμων αποικίας μυρμηγκιών εμπνεύστηκε από τη βιολογία και πιο συγκεκριμένα από τη μελέτη της συμπεριφοράς των μυρμηγκιών σε μια οργανωμένη κοινωνία όπως είναι η αποικία [Bonabeau,99]. Η διατήρηση της ζωής σε μια αποικία μυρμηγκιών καθώς και η εξέλιξή της βασίζεται σε μια πολύ καλή οργάνωση και κατανομή εργασιών σε όλα τα μέλη της αποικίας [Dorigo,99]. Για παράδειγμα κάποια μέλη αναλαμβάνουν την εύρεση και μεταφορά του φαγητού κάποια άλλα μέλη την προάσπιση της αποικίας άλλα την επέκταση αυτής κλπ. Παρόλο που δεν υπάρχει ένας ενιαίος συντονισμός για την εκτέλεση των εργασιών, αυτές αποπερατώνονται με μια αξιοθαύμαστη λεπτομέρεια. Τι είναι όμως αυτό που κάνει τα μέλη μιας αποικίας μυρμηγκιών να συνεργάζονται τόσο στενά; Πως τα μέλη μιας κοινωνίας χωρίς κάποια ιδιαίτερη νοημοσύνη καταφέρνουν να πετύχουν πολύπλοκες εργασίες χωρίς ενιαίο κέντρο συντονισμού; Πως μπορούν να βρουν το δρόμο προς την τροφή όντας σχεδόν τυφλά; Η απάντηση σε ερωτήματα όπως τα παραπάνω, δόθηκε από τους βιολόγους οι οποίοι παρατήρησαν ότι τα μυρμήγκια συμπεριφέρονται σαν απλές αυτόνομες μονάδες με μια στοιχειώδη νοημοσύνη, αλλά, ταυτόχρονα ανέπτυξαν ένα εξελιγμένο σύστημα για την μεταξύ τους επικοινωνία το οποίο στηρίζεται στην ανταλλαγή χημικών ερεθισμάτων «σημάτων». Ένα καλό παράδειγμα για την κατανόηση της συμπεριφοράς των μυρμηγκιών είναι η μελέτη του τρόπου με τον οποίο βρίσκουν την τροφή τους και τη μεταφέρουν πίσω στην αποικία. Καθώς βαδίζουν τα μυρμήγκια, εναποθέτουν στο έδαφος μια χημική ουσία η οποία καλείται φερομόνη. Η εναπόθεση της φερομόνης οδηγεί σε δημιουργία μονοπατιών. Τα υπόλοιπα μυρμήγκια αν και σχεδόν τυφλά ακολουθούν τα μονοπάτια φερομόνης. Τα μονοπάτια που περιέχουν μεγαλύτερες ποσότητες φερομόνης έχουν μεγαλύτερη πιθανότητα να ακολουθηθούν από τα μυρμήγκια. Επιπρόσθετα η φερομόνη εξατμίζεται με τον χρόνο. Αυτό έχει ως αποτέλεσμα τα μονοπάτια τα οποία δεν προτιμώνται να εξαφανίζονται με τον χρόνο. Αυτό τους δίνει τη δυνατότητα να βρίσκουν τον δρόμο προς την τροφή ή πίσω προς την αποικία. Πειράματα που πραγματοποιήθηκαν, απέδειξαν ότι η χρήση της φερομόνης από τα μυρμήγκια οδηγεί στην εύρεση και τη δημιουργία συντομότερων δρόμων από την αποικία στην τροφή και αντίστροφα [Dorigo,96].

Στο σχήμα 1 που ακολουθεί αποτυπώνεται ένα πείραμα με πραγματικ μυρμήγκια που επιλέγουν την συντομότερη διαδρομή [Dorigo,96]. Αρχικά υποθέτουμε ότι υπάρχει μια διαδρομή που συνδέει μια αποικία μυρμηγκιών (σημείο Α) με την τροφή (σημείο Β) και ότι όλα τα μυρμήγκια την ακολουθούν. Επίσης θεωρούμε ότι ταχύτητα με την οποία βαδίζουν τα μυρμήγκια είναι ίδια καθώς και η ποσότητα φερομόνης που εναποθέτουν. Κάποια στιγμή, τοποθετείται ένα εμπόδιο με τέτοιο τρόπο ώστε η διαδρομή να κόβεται στα δυο (ΑΓ και ΔΒ) και το ένα άκρο του εμποδίου (σημείο Ε) να είναι πιο κοντά στη διαδρομή από το δεύτερο (σημείο Ζ). Τα πρώτα μυρμήγκια που θα φτάσουν στο σημείο Γ θα πρέπει να αποφασίσουν ποια διαδρομή θα ακολουθήσουν μεταξύ των ΓΕ και ΓΖ. Επειδή σε και τις δύο διαδρομές τα επίπεδα φερομόνης είναι τα ίδια (=0) οι πιθανότητα επιλογής είναι η ίδια για και τις δυο διαδρομές. Οπότε αρχικά ίδιος (ίσος) αριθμός μυρμηγκιών θα ακολουθήσει τα δύο μονοπάτια ΓΕΔ και ΓΖΔ και συνεπώς ίση ποσότητα φερομόνης θα εναποθετηθεί σ' αυτά. Επειδή όμως η διαδρομή ΓΖΔ είναι μεγαλύτερη από τη

διαδρομή ΓΕΔ, τα μυρμήγκια θα χρειαστούν να περισσότερο χρόνο να διανύσουν τη διαδρομή ΓΖΔ με αποτέλεσμα μεγαλύτερο ποσοστό φερομόνης θα εξατμιστεί από την μακρινή διαδρομή.



Σχήμα 1. Επιλογή ελάχιστης διαδρομής από τα μυρμήγκια

Στη συνέχεια όταν τα επόμενα μυρμήγκια θα χρειαστεί να επιλέξουν μεταξύ των δυο διαδρομών τα περισσότερα θα ακολουθήσουν το μονοπάτι με τη μεγαλύτερη ποσότητα φερομόνης το οποίο είναι και το μικρότερο σε απόσταση εκ των δύο (ΓΕΔ). Όσο περνά ο χρόνος τα όλο και περισσότερα μυρμήγκια θα ακολουθούν τη μικρότερη σε απόσταση διαδρομή (ΓΕΔ) με αποτέλεσμα η ποσότητα φερομόνης να αυξάνει συνεχώς. Από την άλλη πλευρά καθώς η μεγαλύτερη διαδρομή θα επιλέγεται από λιγότερα μυρμήγκια, η ποσότητα φερομόνης θα μειώνεται συνεχώς. Τελικά μετά από κάποιο χρονικό διάστημα όλα τα μυρμήγκια θα ακολουθούν τη συντομότερη διαδρομή ενώ η άλλη διαδρομή θα εξαφανιστεί.

Ο τρόπος με τον οποίο χρησιμοποιούν τα μυρμήγκια την επονομαζόμενη χημική ουσία της φερομόνης τα οδηγεί στην εύρεση των συντομότερων διαδρομών προς την τροφή και θυμίζει έντονα το φαινόμενο της θετικής ανάδρασης.

4.1.2 Συμπεριφορά και Μοντελοποίηση

Το τελικό αποτέλεσμα είναι σε σύντομο διάστημα όλα τα μυρμήγκια θα επιλέξουν την βέλτιστη διαδρομή. Εντούτοις, η απόφαση για το εάν μια πορεία ακολουθηθεί ή όχι, δεν είναι ποτέ αιτιοκρατική αλλά πιθανοτική επιτρέποντας κατά συνέπεια μια συνεχόμενη εξερεύνηση εναλλακτικών διαδρομών. [Dorigo,04]

Τα υπολογιστικά πρότυπα που έχουν αναπτυχθεί για να μιμηθούν τη διαδικασία εύρεσης τροφής των φυσικών παρουσιάζουν ικανοποιητικά αποτελέσματα, δείχνοντας ότι ένα απλό πιθανολογικό πρότυπο βασισμένο στη νοημοσύνη σμηνών, είναι αρκετό να προσομοιώσει σύνθετες και συλλογικές διαδικασίες. Αυτό είναι ένα σημαντικό αποτέλεσμα, όπου ένα ελάχιστο επίπεδο μεμονωμένης πολυπλοκότητας μπορεί να εξηγήσει μια σύνθετη συλλογική συμπεριφορά. Μια αύξηση στην υπολογιστική πολυπλοκότητα κάθε ατόμου του σμήνους, μόλις καθιερωθεί το χαμηλότερο όριο που απαιτείται για να αποτελέσει τις επιθυμητές συμπεριφορές, μπορεί να βοηθήσει στη διαφυγή από τα τοπικά βέλτιστα και να αντιμετωπίσει τις αλλαγές του περιβάλλοντος.

Η τοποθέτηση ιχνών που ρυθμίζεται από το σύστημα ανατροφοδότησης πληροφοριών, διευκολύνει τα τεχνητά μυρμήγκια «πράκτορες» να ακολουθήσουν πορείες που ακολουθήθηκαν από προηγούμενα μυρμήγκια «πράκτορες». Αυτό πρακτικά, μετά από ένα χρονικό διάστημα οδηγεί τα τεχνητά μυρμήγκια να συγκλίνουν στην εύρεση μιας βέλτιστης κοινής λύσης.

4.1.3 Τεχνητά Μυρμήγκια «Πράκτορες»

Η ικανότητα των μυρμηγκιών στην εύρεση της συντομότερης διαδρομής μέσω μηχανισμών συνεργασίας οδήγησε τους επιστήμονες στη μελέτη της συμπεριφοράς τους και τη μεταφορά των φυσικών διεργασιών, που ακολουθούν τα μυρμήγκια, στον τεχνητό κόσμο με σκοπό τη δημιουργία μοντέλων που προσομοιώνουν αυτές τις φυσικές διεργασίες και την ανάπτυξη υπολογιστικών συστημάτων ικανά να παρέχουν λύσεις σε σύνθετα προβλήματα. Η αρχική μεταφορά της νοημοσύνης των μερμηγκιών στον κόσμο των υπολογιστών προτάθηκε από τον ιταλό Dorigo με τη δημιουργία του Συστήματος Αποικίας Μυρμηγκιών (Ant Colony System (ACS)) [Dorigo,96]. Το ανωτέρω σύστημα χρησιμοποιεί τον προτεινόμενο από τον Dorigo, αλγόριθμο βελτιστοποίησης αποικίας μυρμηγκιών ο οποίος προσομοιώνει τη φυσική συμπεριφορά των μυρμηγκιών για την επίλυση προβλημάτων βελτιστοποίησης. Πιο συγκεκριμένα ο αλγόριθμος αρχικά σχεδιάστηκε για την επίλυση προβλημάτων τύπου περιπλανώμενου πωλητή (TSP). Δυο σημαντικά χαρακτηριστικά του αλγορίθμου είναι η επεκτασιμότητα και η ευρωστία του. Έχει δε τα ακόλουθα επιθυμητά χαρακτηριστικά:

- Είναι άμεσα επεκτάσιμος, καθώς έχει τη δυνατότητα να εφαρμοστεί σε παρόμοιες εκδοχές του ίδιου προβλήματος. Για παράδειγμα, μπορεί να επεκταθεί απλά από το πρόβλημα του περιπλανώμενου πωλητή TSP στο αντίστοιχο ασύμμετρο ATSP.
- Είναι εύρωστος. Μπορεί να εφαρμοστεί με ελάχιστες αλλαγές σε άλλα συνδυαστικά προβλήματα βελτιστοποίησης, όπως είναι το τετραγωνικό πρόβλημα ανάθεσης QAP (Quadratic Assignment Problem) και η δρομολόγηση εργασιών JSP (Job – Shop scheduling).

Επιπλέον ο αλγόριθμος ακολουθώντας την λογική όλων των αλγορίθμων νοημοσύνης σμήνους, βασίζεται στη συνεργασία ενός συνόλου τεχνιτών μυρμηγκιών με "κρίση", και δυνατότητα μερικής μνήμης τα οποία με τη χρήση ενός συνόλου κανόνων επικοινωνίας, έχουν τη δυνατότητα αποπεράτωσης μιας διαδικασίας βελτιστοποίησης. Τα τεχνητά μυρμήγκια χαρακτηρίζονται ως πράκτορες που μιμούνται τη συμπεριφορά των πραγματικών μυρμηγκιών.

4.1.4 Χαρακτηριστικά των Πρακτόρων

Σε ένα τεχνητό σύστημα αυτό – οργάνωσης τα μέλη του συστήματος έχουν ένα σύνολο κανόνων και χαρακτηριστικών για να επιτυγχάνεται η μεταξύ τους συνεργασία. Για τον αλγόριθμο αποικίας μυρμηγκιών ένα μέρος των χαρακτηριστικών των πρακτόρων ορίζονται από τη συμπεριφορά των αντίστοιχων φυσικών συγγενών τους:

- Τα τεχνητά μυρμήγκια «πράκτορες» επιλέγουν την επόμενης κατάστασή τους με βάση το σύνολο των ερεθισμάτων που δέχθηκαν στην προηγούμενη κατάσταση.
- Το σύνολο των ερεθισμάτων της προηγούμενης κατάστασης είναι η πληροφορία που εναπόθεσαν στη συγκεκριμένη κατάσταση τα υπόλοιπα μυρμήγκια πράκτορες σε προηγούμενη χρονική στιγμή και προσομοιώνει την φερομόνη των πραγματικών μυρμηγκιών.
- Οι αποδοτικότερες μεταπτώσεις μεταξύ δύο καταστάσεων συγκεντρώνουν μεγαλύτερο σύνολο ερεθισμάτων με αποτέλεσμα τη συχνότερη χρήση από τα μέλη του συστήματος κατά αναλογία με τον πραγματικό κόσμο όπου τα μικρότερα μονοπάτια τείνουν να αυξάνουν το επίπεδο της φερομόνης με μεγαλύτερο ρυθμό

Έκτός των βασικών «φυσικών» χαρακτηριστικών τους, τα τεχνητά μυρμήγκια «πράκτορες» χρησιμοποιούν και επιπρόσθετα χαρακτηριστικά τα οποία τους επιτρέπουν να δημιουργήσουν αποδεκτές λύσεις. Τα πιο σημαντικά, είναι η χρήση μνήμης και η διακριτοποίηση του χρόνου.

Με την προσθήκη μνήμης δίνεται η δυνατότητα διαχείρισης των ιχνών φερομόνης σε κάθε χρονική στιγμή, ενώ η διακριτοποίηση χρόνου μειώνει κατά πολύ την υπολογιστική τους πολυπλοκότητα.

4.1.5 Εύρεση Λύσης

Η βασική ιδέα λειτουργίας των αλγορίθμων αποικίας μυρμηγκιών είναι ότι, όταν ένα δεδομένο τεχνητό μυρμήγκι πρέπει να επιλέξει μεταξύ δύο ή περισσότερων πορειών, η πορεία που επιλέχτηκε συχνότερα από άλλα μυρμήγκια στο παρελθόν θα έχει μια μεγαλύτερη πιθανότητα επιλογής από το μυρμήγκι. Επομένως, τα ίχνη με το μεγαλύτερο ποσό φερομόνης είναι συνώνυμα των βέλτιστων καταστάσεων. Στην ουσία, ένας αλγόριθμος αποικίας μυρμηγκιών εκτελεί επαναληπτικά έναν βρόχο που περιέχει δύο βασικές διαδικασίες, δηλαδή:

1. Μια διαδικασία που καθορίζει τον τρόπο με τον οποίο τα τεχνητά μυρμήγκια κατασκευάζουν ή τροποποιούν τις αποδεκτές λύσεις στο δεδομένο προς επίλυση πρόβλημα
2. Μια διαδικασία η οποία θα ανανεώνει τα επίπεδα φερομόνης στα επιμέρους τμήματα που θα απαρτίσουν την τελική λύση.

Η κατασκευή ή τροποποίηση μιας λύσης εκτελείται με έναν πιθανοτικό τρόπο. Η πιθανότητα της προσθήκης ενός νέου στοιχείου σε μια τρέχουσα μερική λύση δίνεται από μια συνάρτηση που εξαρτάται από μια ευρετική συνάρτηση (η) άμεσα εξαρτώμενη από το εκάστοτε πρόβλημα και αναπαριστά την σημαντικότητα του εκάστοτε στοιχείου στο συγκεκριμένο σημείο της μερικής λύσης και από το ποσό φερομόνης (τ) που εναποθέτεται από τα μυρμήγκια «πράκτορες» στο αντίστοιχο στοιχείο στο παρελθόν. Η διαδικασία ανανέωσης των ιχνών φερομόνης εξαρτάται από το ποσοστό εξάτμισης των μονοπατιών φερομόνης και από την ποιότητα της παραχθείσας λύσης. Για την επίλυση ενός προβλήματος με τη χρήση αλγορίθμων αποικίας μυρμηγκιών είναι απαραίτητος ο καθορισμός ορισμένων χαρακτηριστικών:

- Μια κατάλληλη αντιπροσώπευση του προβλήματος, που επιτρέπει τα μυρμήγκια να κατασκευάσουν η να τροποποιήσουν κλιμακωτά τις λύσεις μέσω της χρήσης ενός πιθανοτικού μοντέλου.
- Κανόνες μετακίνησης από ένα σημείο σε ένα άλλο βασισμένοι στην τιμή του ίχνους φερομόνης.
- Μια ευρετική συνάρτηση (η) που μετρά την ποιότητα των στοιχείων που μπορεί να προστεθούν στην τρέχουσα μερική λύση.
- Μια μέθοδος ελέγχου ποιότητας και εγκυρότητας παραχθέντων λύσεων, δηλαδή οι λύσεις θα πρέπει να ικανοποιούν τους περιορισμούς του προβλήματος.
- Ένας κανόνας για την ενημέρωση των ιχνών φερομόνης.
- Ένας πιθανοτικός κανόνας της μετάβασης βασισμένος στην τιμή της ευρετικής συνάρτησης (η) και στο περιεχόμενο των ιχνών φερομόνης (τ).

4.2 Θεωρητικό Μοντέλο Αλγορίθμων Αποικίας Μυρμηγκιών

Στην παράγραφο αυτή παρέχεται το γενικό μοντέλο λειτουργίας των αλγορίθμων αποικίας μυρμηγκιών όπως περιγράφεται στο [Dorigo,04]. Η κατανόηση των χαρακτηριστικών του αλγορίθμου καθώς και ο τρόπος σύγκλισης σε αποδεκτή λύση γίνεται με την εισαγωγή ενός θεωρητικού προβλήματος βελτιστοποίησης.

4.2.1 Ορισμός Προβλήματος

Για το πως λειτουργούν οι αλγόριθμοι αποικίας μυρμηγκιών θεωρούμε ένα πρόβλημα ελαχιστοποίησης *(S,f,Ω)*, όπου *S* είναι ένα σύνολο υποψήφιων λύσεων, *f* είναι η αντικειμενική συνάρτηση η οποία προσδιορίζει μια συνάρτηση κόστους για κάθε υποψήφια λύση $s \in S$, και Ω είναι ένα σύνολο μεταβλητών και παραμέτρων, οι οποίες ορίζουν το σύνολο των εφικτών λύσεων. Ο στόχος είναι να βρεθεί μια βέλτιστη λύση *s**, η οποία είναι μια εφικτή υποψήφια λύση ελάχιστου κόστους. Επιπρόσθετα το θεωρητικό πρόβλημα ελαχιστοποίησης εμπεριέχει ένα σύνολο ιδιοτήτων – χαρακτηριστικών:

- Ένα πεπερασμένο σύνολο δομικών συστατικών $C = \{c_1, c_2, c_3, ..., c_{Nc}\}$
- Ένα πεπερασμένο σύνολο X καταστάσεων του προβλήματος, το οποίο ορίζεται ως το σύνολο $\chi = \{c_i, c_j, ..., c_h, ...\}$ όλων των πιθανών αλληλουχιών των δομικών συστατικών του προβλήματος. Το μήκος μιας αλληλουχίας χ, το οποίο είναι και το σύνολο των στοιχείων στην αλληλουχία, εκφράζεται με το $|x|$. Το μέγιστο δυνατό μήκος μιας αλληλουχίας περιορίζεται από μια θετική σταθερά $n < +\infty$.
- Το σύνολο των υποψήφιων λύσεων αποτελεί ένα υποσύνολο του X.
- Ορίζεται ένα σύνολο εφικτών καταστάσεων $\widetilde{X}$ με $\widetilde{X} \subseteq X$,το οποίο αποδεικνύει ότι είναι εφικτή η ολοκλήρωση μιας αλληλουχίας $\chi \in \widetilde{X}$ που αντιστοιχεί σε μια υποψήφια λύση που ικανοποιεί τις παραμέτρους Ω.
- Ένα μη κενό σύνολο S^* βέλτιστων λύσεων με $S^* \subseteq \widetilde{X}$ και $S^* \subseteq S$.
- Επιπρόσθετα ένα κόστος g(s) ανατίθεται και κάθε υποψήφια λύση $s \in S$ για το οποίο ισχύει ότι $g(s) \equiv f(s) \forall s \in S$.

Τα τεχνητά μυρμήγκια με βάση τους παραπάνω ορισμούς χτίζουν υποψήφιες λύσεις, εκτελώντας τυχαίες μετακινήσεις πάνων σε έναν πλήρως συνδεδεμένο γράφο $G_C = (C, L)$ με κόμβους τα δομικά συστατικά *C*, και *L* το σύνολο των συνδέσεων μεταξύ των κόμβων. Η τυχαία περιήγηση των τεχνητών μυρμηγκιών εξαρτάται από τα ίχνη φερομόνης *τ* που συσσωρεύονται σε μια μήτρα *T*. Τα ίχνη φερομόνης αντιστοιχούν με τις συνδέσεις των κόμβων του γράφου, έτσι το τ_{ij} ορίζεται ως η τιμή της φερομόνης που αντιστοιχεί στην σύνδεση των δομικών συστατικών *i* και *j*.

4.2.2 Κατασκευή Λύσης

Ο αλγόριθμος αρχικοποιείται θέτοντας μια αρχική τιμή $\tau_0>0$ στα ίχνη φερομόνης όλων των συνδέσεων. Σε κάθε επανάληψη του αλγορίθμου, τα τεχνητά μυρμήγκια τοποθετούνται σε κόμβους που επιλέγονται με βάση τα κριτήρια του εκάστοτε προβλήματος. Κατά τη διάρκεια μετακίνησης του τεχνητού μυρμηγκιού μεταξύ των κόμβων του γράφου G_C, το σύνολο των μεταβλητών Ω του προβλήματος χρησιμοποιούνται για να αποτρέψουν τη δημιουργία μη αποδεκτών λύσεων. Η γενική διαδικασία κατασκευής λύσης καλείται *AntSolutionConstruction* και περιγράφεται στο σχήμα 2.

Procedure AntSolutionConstruction
Επέλεξε έναν αρχικό κόμβο που ικανοποιεί τα κριτήρια του προβλήματος
$h \leftarrow 1$
$\chi_h \leftarrow \langle c_1 \rangle$
Επανέλαβε όσο ($x_h \notin S$ *and* $\mathrm{N}_\mathrm{i}^\mathrm{k} \neq \emptyset$)
$j \leftarrow selecNextNode(x_h, T)$
$\chi_h \leftarrow \chi_h \otimes j$
Τέλος Επανάληψης
Αν $x_h \in S$ Τότε
Επέστρεψε x_h
Αλλιώς
Ακύρωσε διαδικασία
Τέλος Αν
Τέλος διαδικασίας

Σχήμα 2. Λογική αναπαράσταση της διαδικασίας κατασκευής λύσης

Η συνάρτηση $selecNextNode(x_h, T)$ επιστρέφει τον επόμενο προς επίσκεψη κόμβο σύμφωνα με την συνάρτηση πιθανότητας που δίνεται από τη σχέση 4.

$$P_T(c_{h+1} = j \mid x_h) = \begin{cases} \dfrac{F_{ij}(\tau_{ij})}{\sum_{i,l \in N_i^k} F_{il}(\tau_{il})}, & \text{if } (i,j) \in N_i^k; \\ 0, & \text{αλλιώς;} \end{cases} \qquad \textbf{Σχέση 4}$$

Στην παραπάνω σχέση το j είναι το δομικό στοιχείο που προστίθεται στη υπάρχουσα κατάσταση x_h και επιλέγεται μόνο αν η νέα κατάσταση $\chi = \{c_i, c_j, ..., c_h, j\}$ ικανοποιεί τους περιορισμούς Ω του προβλήματος, και $F_{ij}(z)$ μια συνάρτηση που εξαρτάται από την τιμή της φερομόνης σε κάθε σύνδεση μεταξύ των στοιχείων i και j, αλλά και από την τιμή μιας ευρετικής συνάρτησης η_{ij} που υποδηλώνει την "προσβασιμότητα" του στοιχείου j από το τεχνητό μυρμήγκι από την θέση i. Στις περισσότερες εφαρμογές του αλγορίθμου ισχύει ότι $F_{ij}(z) = z^a \eta^b$ με $a, b > 0$.

Αν κατά τη διάρκεια κατασκευής λύσης, η δημιουργηθείσα κατάσταση δεν ανήκει στο σύνολο των υποψήφιων λύσεων ή δεν υπάρχει διαθέσιμο στοιχείο για να προστεθεί στη ήδη υπάρχουσα κατάσταση, δηλαδή ισχύει ότι $x_h \notin S$ και $N_i^k = \emptyset$, τότε η διαδικασία κατασκευής λύσης διακόπτεται, απαγορεύοντας τη δημιουργία μη εφικτών λύσεων.

4.2.3 Συνάρτηση Φερομόνης

Όταν όλα τα τεχνητά μυρμήγκια εκτελέσουν τη διαδικασία κατασκευής λύσης (*AntSolutionConstruction*), ακολουθεί η διαδικασία ενημέρωσης των ιχνών φερομόνης, που ουσιαστικά περιλαμβάνει δυο φάσεις. Κατά την πρώτη φάση μειώνεται η τιμή των ιχνών της φερομόνης σε όλες τις συνδέσεις L μεταξύ των κόμβων του γράφου G_C, με βάση μια σταθερά ρ, η οποία προσομοιώνει την εξάτμιση της φερομόνης και καλείται ρυθμός εξάτμισης. Κατά τη δεύτερη φάση μόνο για ορισμένα ίχνη φερομόνης αυξάνεται η τιμή τους, και ο τρόπος ορίζεται από δύο βασικά μοντέλα. Στο πρώτο μοντέλο, η ενημέρωση της τιμής της φερομόνης γίνεται

μόνο στα ίχνη που αντιστοιχούν σε διαδρομές του γράφου που χρησιμοποιήθηκαν από ένα τουλάχιστο τεχνητό μυρμήγκι.

Procedure GenericPheromoneUpdate
Για κάθε $(i, j) \in L$ επανέλαβε
$\tau_{ij} \leftarrow (1-\rho)\tau_{ij}$
Τέλος επανάληψης
Για κάθε $s \in \hat{S}_\theta$ επανέλαβε
Για κάθε $(i, j) \in s$ επανέλαβε
$\tau_{ij} \leftarrow \tau_{ij} + q_f(s \mid S_1,...S_\theta)$
Τέλος επανάληψης
Τέλος επανάληψης
Τέλος διαδικασίας

Σχήμα 3. Λογική αναπαράσταση της διαδικασίας ανανέωσης φερομόνης

Στο δεύτερο μοντέλο, υποθέτοντας ότι η s^{bs} είναι η καλύτερη μέχρι τώρα αποδεκτή λύση, με $f(s^{bs})$ την αντίστοιχη συνάρτηση κόστους, τότε η διαδικασία ενημέρωσης της φερομόνης αυξάνει την τιμή των ιχνών της φερομόνης στις συνδέσεις που ανήκουν στη βέλτιστη μέχρι τώρα αποδεκτή λύση s^{bs}.

Η διαδικασία ανανέωσης της τιμής των ιχνών της φερομόνης μπορεί να διαφέρει από μοντέλο σε μοντέλο αλλά το σχήμα 3 περιγράφει τη γενική διαδικασία ενημέρωσης της φερομόνης.

4.2.4 Σύγκλιση Αλγορίθμου

Η ενότητα αυτή, διαπραγματεύεται με τη σύγκλιση του αλγορίθμου.

Αρχικά ορίζεται ο αλγόριθμος $ACO_{bs,\tau_{\min}}$ για τον οποίο αποδεικνύεται η σύγκλιση σε τιμή. Στη συνέχεια με τον ορισμό του αλγορίθμου $ACO_{bs,\tau_{\min}(\theta)}$ μελετάται η σύγκλιση το αλγορίθμου σε λύση.

Ο αλγόριθμος $ACO_{bs,\tau_{\min}}$ ορίζεται όπως και στην προηγούμενη ενότητα, με τη μόνη διαφορά, ότι η επιλογή του επόμενου τμήματος της λύσης από το μυρμήγκι, εξαρτάται αποκλειστικά και μόνο από τα ίχνη φερομόνης των συνδέσεων του γράφου G_C. Με άλλα λόγια ισχύει ότι $F_{ij}(\tau_{ij}) \equiv F(\tau_{ij})$. Επιπρόσθετα για τον ορισμό της συνάρτησης F χρησιμοποιείται ο τύπος που συναντάται στους περισσότερους αλγορίθμους της οικογένειας αποικίας μυρμηγκιών και είναι της μορφής $F(\tau_{ij}) \equiv \tau_{ij}^a$,όπου α είναι μια παράμετρος για την οποία ισχύει $0 < a < +\infty$, οπότε η πιθανοτική συνάρτηση κατασκευής λύσεων γίνεται:

$$P_T(c_{h+1} = j \mid x_h) = \begin{cases} \dfrac{\tau_{ij}^a}{\sum_{i,l \in N_i^k} \tau_{ij}^a}, \text{if } (i, j) \in N_i^k; \\ 0 \qquad\qquad , \text{αλλιώς}; \end{cases} \qquad \textbf{Σχέση 5}$$

Στη συνέχεια ως μοντέλο ανανέωσης της φερομόνης επιλέγεται το μοντέλο ανανέωσης ολικού μέγιστου $\hat{S}_\theta = s^{bs}$, προσθέτοντας και ένα ελάχιστο όριο $\tau_{\min} > 0$ στην τιμή του ίχνους φερομόνης για το οποίο θα ισχύει ότι $\tau_{\max} < q_f(s^*)$.

Σύγκλιση σε τιμή

Η σύγκλιση σε τιμή ουσιαστικά είναι η σύγκλιση σε μια λύση με μια πιθανότητα κοντά στο 1, με δεδομένο ότι ο χρόνος σύγκλισης είναι αρκετά μεγάλος. Πρακτικά γίνεται η θεώρηση ότι ο αριθμός των επαναλήψεων τείνει στο άπειρο $\theta \to \infty$. Η σύγκλιση στηρίζεται στο γεγονός ότι η μέγιστη τιμή της φερομόνης που μπορεί να έχει μια σύνδεση του γράφου G_C, έχει ένα άνω όριο σύγκλισης το οποίο δίνεται από τη σχέση 6.

$$\lim_{\theta\to\infty} \tau_{ij} \le \tau_{\max} = \frac{q_f(s^*)}{\rho} \qquad \textbf{Σχέση 6}$$

Το παραπάνω άνω ασυμπτωματικό όριο της τιμή της φερομόνης επιτυγχάνεται ουσιαστικά όταν βρεθεί η βέλτιστη λύση. Αν υποτεθεί ότι ο αλγόριθμος ανακαλύπτει μία βέλτιστη λύση τουλάχιστο μια φόρα στις θ πρώτες επαναλήψεις, με μια πιθανότητα $P^*(\theta)$, τότε για μια μικρή διακύμανση $\varepsilon > 0$ και για έναν μεγάλο αριθμό επαναλήψεων τότε η πιθανότητα εύρεσης της λύσης αυτής είναι $P^*(\theta) > 1-\varepsilon$, ενώ το όριο είναι $\lim_{\theta\to\infty} P^*(\theta) = 1$.

Σύγκλιση σε λύση

Κατά τη σύγκλιση στη λύση αποδεικνύεται ότι κάθε τυχαίο μυρμήγκι θα συγκλίνει σε μια βέλτιστη λύση με πιθανότητα 1. Για τη σύγκλιση στη λύση χρησιμοποιείται μια άλλη θεωρητική παραλλαγή του αλγορίθμου, ο αλγόριθμος $ACO_{bs,\tau_{\min}(\theta)}$, με βάση τον οποίο η ελάχιστη τιμή του ίχνους φερομόνης μπορεί να μεταβάλλεται.

Η σύγκλιση ουσιαστικά στηρίζεται στο γεγονός ότι, οι τιμές των ιχνών φερομόνης, για τις συνδέσεις του γράφου G_C, που δεν χρησιμοποιούνται για την κατασκευή μιας λύσης, μονίμως θα μειώνονται και μετά από πολλές επαναλήψεις ασυμπτωματικά θα τείνουν στο 0. Συγκεκριμένα, η ελάχιστη τιμή της φερομόνης δίνεται από τη σχέση 7.

$$\tau_{\min} = \frac{d}{\ln(\theta+1)}, \quad \forall \theta \ge 1 \qquad \textbf{Σχέση 7}$$

Ενώ αν $P^*(\theta)$ η πιθανότητα να συγκλίνει ο αλγόριθμος σε μια βέλτιστη λύση, τότε αυτή τείνει στο 1, δηλ:

$$\lim_{\theta\to\infty} P^*(\theta) = 1 \qquad \textbf{Σχέση 8}$$

Το εξαγόμενο συμπέρασμα από τα παραπάνω είναι το γεγονός ότι ουσιαστικά ο αλγόριθμος όταν μετά από έναν αριθμό επαναλήψεων θ^* κατασκευάσει μια βέλτιστη λύση, τότε η πιθανότητα $P^*(s^*,\theta,k)$ να συγκλίνει στη λύση αυτή το κάθε μυρμήγκι k και για κάθε επιπλέον επανάληψη θ, είναι ίση με 1, δηλαδή:

$$\lim_{\theta\to\infty} P^*(s^*,\theta,k) = 1 \qquad \textbf{Σχέση 9}$$

Συμπεράσματα της σύγκλισης

Με τη μελέτη της σύγκλισης του αλγορίθμου, πρακτικά αποδεικνύεται ότι ο αλγόριθμος μπορεί να κατασκευάσει βέλτιστες λύσεις και να συγκλίνει σ' αυτές. Συνοπτικά τα συμπεράσματα της μελέτης της σύγκλισης είναι:

- Όταν γίνεται η χρήση ενός ελάχιστου κάτω ορίου για την τιμή της φερομόνης, ο αλγόριθμος συγκλίνει στην βέλτιστη δυνατή λύση.
- Χωρίς τη χρήση του ελάχιστου κάτω ορίου για την τιμή της φερομόνης, ο αλγόριθμος εξακολουθεί να συγκλίνει σε μια βέλτιστη λύση.
- Κατά τη σύγκλιση του αλγορίθμου σε μια βέλτιστη λύση, η τιμή των ιχνών φερομόνης των στοιχείων που απαρτίζουν τη λύση, συγκλίνει σε ένα μέγιστο άνω όριο.
- Μετά την εύρεση μιας βέλτιστης λύσης ο αλγόριθμος τελικά επέρχεται σε μια κατάσταση ισορροπίας κατά την οποία όλα τα μυρμήγκια κατασκευάζουν συνεχώς αυτή τη λύση.

4.3 Πεδία εφαρμογής του αλγορίθμου

Οι αλγόριθμοι αποικίας μυρμηγκιών αν και σχετικά πρόσφατοι, έχουν ένα μεγάλο πεδίο εφαρμογών σε κλασσικά αλλά και σε σύγχρονα προβλήματα. Ωστόσο οι σημαντικότερες εφαρμογές είναι στις εξής κατηγορίες προβλημάτων:

- Προβλήματα δρομολόγησης
- Προβλήματα ανάθεσης
- Προβλήματα οργανογραμμάτων
- Μηχανική μάθηση

Στα προβλήματα δρομολόγησης από τις κυριότερες εφαρμογές είναι ο Ant Colony System (ACS), που χρησιμοποιείται για την επίλυση προβλημάτων TSP (Traveling Salesman Problem). Παραλλαγές του ACS όπως οι Hybrid AS-SOP και AS_{rank}-CVRP χρησιμοποιούνται για την επίλυση προβλημάτων SO (Sequential Ordering) και CVRP (Capacitated Vechicle Route Problem) αντίστοιχα. Επίσης στην κατηγορία αυτή μπορεί να προστεθεί και ο Ant-Net ο οποίος χρησιμοποιείται για δρομολόγηση πακέτων σε τηλεπικοινωνιακά δίκτυα.

Στην κατηγορία των προβλημάτων ανάθεσης οι κύριοι εκφραστές είναι οι AS-QAP, *MM*AS-QAP και ANTS-QAP. Το πεδίο εφαρμογών περιλαμβάνει τετραγωνική ανάθεση (QA), ανάθεση συχνοτήτων (FA), αλλά και άλλα όπως το "Graph Coloring Problem (GCP)".

Στα προβλήματα οργανογραμμάτων ιδιαίτερη αξία έχει η εφαρμογή του αλγορίθμου στις υποκατηγορίες των ,Job Shop Problem (JSP), Open Shop Problem (OSP) και Group Shop Scheduling Problem (GSP).

Τελειώνοντας με τα προβλήματα μηχανικής εκμάθησης αναφέρεται ο Ant-Miner ο οποίος σχεδιάστηκε για τη δημιουργία κανόνων ταξινόμησης καθώς και ο ACS-BN ο οποίος χρησιμοποιείται για την εκμάθηση δικτύων Bayes . Η λειτουργία του Ant-Miner παρουσιάζεται εκτενώς στην επόμενη ενότητα του κεφαλαίου.

5 Χρήση ACO στην ΕΠ

Αν και οι αλγόριθμοι αποικίας μυρμηγκιών είναι σχετικά πρόσφατοι σε διάρκεια ζωής [Dorigo,96], εντούτοις η χρήση τους έχει επεκταθεί σε πολλά επιστημονικά πεδία, ένα εκ' των οποίων είναι και το πεδίο της επεξεργασίας της πληροφορίας. Πιο συγκεκριμένα αξιόλογη είναι η εφαρμογή στη δημιουργία κανόνων ταξινόμησης μονάδων πληροφορίας σε διάφορες θεματικές ενότητες ή κλάσεις. Στη παρακάτω παράγραφο αναφέρεται ο τρόπος με τον οποίο ο αλγόριθμος αποικίας μυρμηγκιών εφαρμόζεται για τη δημιουργία κανόνων ταξινόμησης

5.1 Ant Miner

Η ανακάλυψη κανόνων ταξινόμησης είναι ένα σημαντικό κομμάτι της ανάκτησης πληροφορίας, διότι με τη χρήση ενός συνόλου συμβολικών κανόνων, μπορεί να ταξινομηθεί ένα σύνολο εγγράφων ή μονάδων πληροφορίας D, σε ένα σύνολο κλάσεων ή κατηγοριών k, με έναν φυσικό τρόπο. Το ανθρώπινο μυαλό είναι σε θέση να καταλάβει τους κανόνες καλύτερα από οποιοδήποτε άλλο πρότυπο ανάκτησης πληροφορίας. Εντούτοις, αυτοί οι κανόνες πρέπει να είναι απλοί και περιεκτικοί διαφορετικά, ένας άνθρωπος δεν θα είναι σε θέση να τους κατανοήσει. Πολλοί αλγόριθμοι έχουν χρησιμοποιηθεί ευρέως για τη δημιουργία κανόνων ταξινόμησης.

Στην ουσία, η διαδικασία της ταξινόμησης αποτελεί μια αντιστοίχηση της κάθε μονάδας πληροφορίας (αντικείμενο, ή εγγραφή) σε μια κατηγορία, από ένα σύνολο προκαθορισμένων κατηγοριών. Αυτή η αντιστοίχηση βασίζεται στις τιμές μερικών ιδιοτήτων (οι οποίες αποκαλούνται ιδιότητες πρόβλεψης) για την κάθε περίπτωση. Πολλές φορές, οι κανόνες που προκύπτουν είναι της μορφής:

ΑΝ <*υπόθεση*> ΤΟΤΕ <*κλάση X*>

Το τμήμα του υπόθεσης του κανόνα στην ουσία αποτελεί έναν σύνολο όρων οι οποίοι συνδέονται μεταξύ τους με τους βασικούς κανόνες Boole και έχει τη μορφή:

term1 AND *term2* AND ... AND *termN*

Ο κάθε όρος είναι μια τριπλέτα της μορφής <*Χαρακτηριστικό, τελεστής, τιμή*>. Το κάθε χαρακτηριστικό αποτελεί μια ιδιότητα η οποία έχει ένα σύνολο προκαθορισμένων τιμών, όπως για παράδειγμα στην ιδιότητα «*φύλο*» μπορούν να ανατεθούν δύο τιμές «*Άνδρας*» ή «*Γυναίκα*». Ο τελεστής ουσιαστικά αναθέτει μια από τις προκαθορισμένες τιμές στο αντίστοιχο χαρακτηριστικό, όπως για παράδειγμα <*Φύλο = Άνδρας*>. Το τμήμα <*κλάση X*> υποδηλώνει την κατηγορία στην οποία πρόκειται να αντιστοιχηθεί η μονάδας πληροφορίας.

Στη παρακάτω παράγραφο αναφέρεται ο τρόπος με τον οποίο ο αλγόριθμος αποικίας μυρμηγκιών εφαρμόζεται για τη δημιουργία κανόνων ταξινόμησης [Parpinelli,02a], [Parpinelli,02β].

5.1.1 Δημιουργία Κανόνων Ταξινόμησης

Το κάθε μυρμήγκι χρησιμοποιείται ως ένας πράκτορας οποίος αυξητικά δημιουργεί ή τροποποιεί έναν κανόνα ταξινόμησης.

Κάθε μυρμήγκι αρχίζει με έναν κενό κανόνα και κάθε φορά προσθέτει έναν όρο. Ο τρέχων μερικός κανόνας που κατασκευάζεται από ένα μυρμήγκι αντιστοιχεί στην πορεία που ακολουθείται από το συγκεκριμένο μυρμήγκι. Ομοίως, η επιλογή ενός όρου που προστίθεται στον τρέχοντα κανόνα αντιστοιχεί στην επιλογή της κατεύθυνσης στην οποία η τρέχουσα πορεία θα επεκταθεί, μεταξύ ενός συνόλου πιθανών κατευθύνσεων (όλοι οι όροι που θα μπορούσαν να προστεθούν στον τρέχοντα μερικό κανόνα).

Η επιλογή του όρου που προστίθεται στον τρέχοντα μερικό κανόνα, εξαρτάται από μια ευρετική συνάρτηση και από την ποσότητα φερομόνης που αντιστοιχεί στον συγκεκριμένο όρο. Η ευρετική συνάρτηση καθορίζει την διακριτική ικανότητα του εκάστοτε όρου ενώ η συνάρτηση φερομόνης καθορίζει τη συχνότητα με την οποία επιλέγεται ο κάθε όρος από το σύνολο των μυρμηγκιών.

Το κάθε μυρμήγκι συνεχίζει να προσθέτει έναν όρο τη φορά στον τρέχοντα μερικό κανόνα μέχρι να ικανοποιήσει ένα από τα δυο κριτήρια ολοκλήρωσης του κανόνα. Το πρώτο κριτήριο είναι η χρήση από το μυρμήγκι όλων των διαθέσιμων ιδιοτήτων στην κατασκευή του κανόνα. Θα πρέπει να σημειωθεί ότι κατά την κατασκευή ενός κανόνα ένα χαρακτηριστικό δεν μπορεί να χρησιμοποιηθεί περισσότερες από μια φορά. Το δεύτερο κριτήριο είναι ότι ο κανόνας που θα δημιουργηθεί θα πρέπει να επαληθεύει έναν ελάχιστο αριθμό περιπτώσεων (Min_cases_per_rule).

Όταν ολοκληρωθεί η προσθήκη όρων, ακολουθεί η διαδικασία συρρίκνωσης του κανόνα. Πολλοί από τους όρους που προστέθηκαν δεν προσδίδουν ουσιαστική διακριτική ικανότητα στον κανόνα οπότε αφαιρούνται με αποτέλεσμα να παραμένουν μόνο οι όροι με μεγάλη διακριτική ικανότητα. Η ενσωμάτωση άσχετων όρων στον αρχικό κανόνα δικαιολογείται από το γεγονός ότι η επιλογή τους γίνεται με έναν πιθανοτικό τρόπο και ότι κατά την επιλογή τους δεν ελέγχεται η συνολική διακριτική ικανότητα του κανόνα.

```
TrainingSet = {all training cases};
DiscoveredRuleList = [ ]; /* Αρχικά η λίστα κανόνων είναι κενή */
        WHILE (TrainingSet > Max_uncovered_cases)
t = 1;
j = 1;
Αρχικοποίηση των ιχνών φερομόνης με την ίδια τιμή;
REPEAT
        ConstructRule(Ant_t);
        PruneRule (Ant_t, R_t,Q_Rt);
        UpdatePheromone(Ant_t, R_t, Q_Rt);
        IF (R_t is equal to R_t-1) {
                j = j + 1;
        }
        ELSE {
                j = 1;
        }
        t = t + 1;
UNTIL (i > No_of_ants) OR (j > No_rules_converg)
R_best = argmax(R_t);
AddRule(R_best, DiscoveredRuleList);
TrainingSet = TrainingSet - {set of cases correctly covered by R_best};
END WHILE
```

Σχήμα 4. Ο αλγόριθμος Ant-miner

Όταν ένα μυρμήγκι ολοκληρώνει τον κανόνα του, τα επίπεδα φερομόνης για τους όρους που χρησιμοποιήθηκαν αυξάνουν. Η αύξηση της τιμής της φερομόνης για έναν όρο ουσιαστικά υποδηλώνει ότι ο όρος αυτός χρησιμοποιήθηκε στην κατασκευή ενός κανόνα άρα είναι περισσότερο χρήσιμος για τη χρήση στο μελλοντικό χτίσιμο ενός άλλου κανόνα.

Το επόμενο μυρμήγκι που θα ξεκινήσει την κατασκευή ενός νέου κανόνα θα ακολουθήσει την ίδια λογική χρησιμοποιώντας όμως τα νέα επίπεδα φερομόνης. Η διαδικασία επαναλαμβάνεται για έναν συγκεκριμένο αριθμό μυρμηγκιών No_of_ants. Εντούτοις, αυτή η επαναληπτική διαδικασία μπορεί να διακοπεί νωρίτερα, στην περίπτωση που ο κάθε κανόνας που δημιουργείται είναι ίδιος με τον προηγούμενο. Ουσιαστικά το δεύτερο κριτήριο ανιχνεύει ότι τα μυρμήγκια έχουν συγκλίνει ήδη στον ίδιο κατασκευασμένο κανόνα, το οποίος είναι ομότιμο με τη σύγκλιση

στην ίδια πορεία των πραγματικών συστημάτων αποικίας μυρμηγκιών. Από το σύνολο των δημιουργηθέντων κανόνων επιλέγεται ο καλύτερος και οι περιπτώσεις που ικανοποιούν τον κανόνα αφαιρούνται από το σετ εκπαίδευσης. Η όλη διαδικασία αποκαλείται μια επανάληψη και επαναλαμβάνεται μέχρι το σύνολο των προς εκπαίδευση περιπτώσεων εξαντληθεί ή περιοριστεί κάτω από κατώφλι που ονομάζεται Max_uncovered_cases. Στο παρακάτω σχήμα δείχνει τον αλγόριθμο

5.1.2 Ευρετική Συνάρτηση

Για κάθε όρο $term_{ij}$ που μπορεί να προστεθεί στον τρέχοντα κανόνα, το μυρμήγκι πράκτορας υπολογίζει την αξία η_{ij} μιας ευρετικής συνάρτησης που είναι μια εκτίμηση της ποιότητας αυτού του όρου, όσον αφορά τη δυνατότητά της να βελτιώσει τη διακριτική ικανότητα του κανόνα για μια συγκεκριμένη κλάση. Αυτή η ευρετική λειτουργία είναι βασισμένη στη θεωρία πληροφοριών. Πιο αναλυτικά, η αξία της ευρετικής συνάρτησης η_{ij} για τον όρο $term_{ij}$ περιλαμβάνει ένα μέτρο της εντροπίας (ή του ποσού πληροφοριών) που συνδέεται με τον όρο αυτό. Κάθε όρος $term_{ij}$ είναι της μορφής $Attr_i = Value_{ij}$, με $Attr_i$ να είναι η i ιδιότητα και $Value_{ij}$ να είναι η j τιμή που μπορεί να ανατεθεί στην ιδιότητα $Attr_i$. Η τιμή της εντροπίας δίνεται από τη σχέση:

$$\text{infoT}_{ij} = -\sum_{w=1}^{k}\left[\frac{freqT_{ij}^{w}}{|T_{ij}|}\right] * \log 2\left[\frac{freqT_{ij}^{w}}{|T_{ij}|}\right] \qquad \textbf{Σχέση 10}$$

Όπου:

- k είναι ο συνολικός αριθμός των κλάσεων
- $|T_{ij}|$ είναι ο συνολικός αριθμός των περιπτώσεων όπου ο όρος $Attr_i$ έχει τιμή $Value_{ij}$.
- $freqT_{ij}^{w}$ είναι ο συνολικός αριθμός των περιπτώσεων όπου ο όρος $Attr_i$ έχει τιμή $Value_{ij}$ και ανήκουν στην κλάση w

Όσο υψηλότερη η τιμή της συνάρτησης infoT_{ij}, τόσο πιο ομοιόμορφα διανεμημένη στις κατηγορίες είναι, και έτσι τόσο χαμηλότερη είναι η διακριτική ικανότητα όρου $term_{ij}$. Ο στόχος προφανώς, είναι η προσθήκη στον τρέχοντα μερικό κανόνα, όρων με υψηλή διακριτική ικανότητα. Επομένως, όσο υψηλότερη είναι η αξία της συνάρτησης infoT_{ij}, τόσο μικρότερη η πιθανότητα ενός μυρμηγκιού πράκτορα να επιλέξει τον όρο $term_{ij}$. Όπως γίνεται φανερό η τιμή της συνάρτησης infoT_{ij} κυμαίνεται στα όρια $0 \le \text{infoT}_{ij} \le \log_2(k)$.

$$\eta_{ij} = \frac{\log_2(k) - \text{info}_{ij}}{\sum_{i}^{\alpha}\sum_{j}^{b_i}\log_2(k) - \text{info}_{ij}}, \forall \text{i} \in \text{I} \qquad \textbf{Σχέση 11}$$

Όπου:

- α ο συνολικός αριθμός των χαρακτηριστικών
- b_i ο αριθμός των εναλλακτικών τιμών για το χαρακτηριστικό i.

Το μέτρο εντροπίας που χρησιμοποιείται από τον Ant-Miner ως ευρετική συνάρτηση, είναι παρεμφερής με την αντίστοιχη ευρετική συνάρτηση που χρησιμοποιείται από τους αλγορίθμους δέντρων αποφάσεων όπως ο C4.5 [19]. Η κύρια διαφορά μεταξύ των δέντρων απόφασης και του Ant – miner , όσον αφορά την ευρετική συνάρτηση, είναι ότι στα δέντρα απόφασης η εντροπία χρησιμοποιείται ως μια ιδιότητα, ενώ στον Ant – miner η εντροπία χρησιμοποιείται ως ζευγάρι ιδιοτήτων. Επιπλέον, θα πρέπει να τονιστεί ότι στους συμβατικούς αλγορίθμους δέντρων απόφασης το μέτρο εντροπίας είναι η μόνη ευρετική λειτουργία που χρησιμοποιείται κατά τη διάρκεια δημιουργίας των δέντρων, ενώ στον Ant – miner το μέτρο εντροπίας χρησιμοποιείται μαζί με τη συνάρτηση φερομόνης. Αυτό καθιστά τη διαδικασία κατασκευής κανόνων του πιο δυνατή και λιγότερο επιρρεπή σε καταστάσεις εγκλωβισμού σε τοπικά βέλτιστα, δεδομένου ότι η ανατροφοδότηση που παρέχεται με την ενημέρωση των ιχνών φερομόνης βοηθά να διορθώσει μερικά λάθη που εξάγονται από μια τέτοια συμπεριφορά. Θα πρέπει να σημειωθεί ότι το μέτρο εντροπίας είναι ένα τοπικό ευρετικό μέτρο, το οποίο εξετάζει μόνο μια ιδιότητα τη φορά, και είναι έτσι ευαίσθητο στα προβλήματα αλληλεπίδρασης ιδιοτήτων. Αντίθετα, η ενημέρωση των ιχνών φερομόνης τείνει να αντιμετωπίσει καλύτερα τις αλληλεπιδράσεις ιδιοτήτων, δεδομένου ότι είναι άμεσα βασισμένη στη συνολική απόδοση του κανόνα. Αρχικά, όσο οι τιμές των ιχνών φερομόνης στις επιμέρους ιδιότητες δε διαφέρουν, η διαδικασία της κατασκευής κανόνα που χρησιμοποιείται από τον Ant-Miner, αρχικά οδηγεί σε πολύ κακής ποιότητας κανόνες. Στη συνέχεια όμως με τη διαφοροποίηση των τιμών σε κάθε ίχνος, συγκλίνει σε καλύτερης ποιότητας κανόνες.

5.1.3 Κατασκευή Κανόνα

Όπως σε κάθε αλγόριθμο αποικίας μυρμηγκιών, έτσι και στον αλγόριθμο Ant – miner, η επιλογή και η προσθήκη ενός όρου στον τρέχοντα κανόνα, που κατασκευάζεται από ένα μυρμήγκι πράκτορα, γίνεται με μια πιθανοτική συνάρτηση. Η συνάρτηση αυτή εξαρτάται από την ευρετική συνάρτηση (η) που αναλύθηκε στην προηγούμενη παράγραφο και αναπαριστά τη διακριτική ικανότητα του προς προσθήκη όρου, και από την τιμή του ίχνους φερομόνης (τ) που εναποθέτεται από τα μυρμήγκια πράκτορες σε αυτόν τον όρο στο παρελθόν. Η πιθανοτική συνάρτηση δίνεται από τη σχέση 12.

$$P_{ij}(t) = \frac{\tau_{ij}(t) \cdot \eta_{ij}}{\sum_{i}^{\alpha} \sum_{j}^{b_i} \tau_{ij}(t) \cdot \eta_{ij}}, \forall i \in I \qquad \textbf{Σχέση 12}$$

Όπου:

- η_{ij} η τιμή της ευρετικής συνάρτησης για τον όρο term_{ij}.
- $\tau_{ij}(t)$ η ποσότητα της φερομόνης που αντιστοιχεί στον όρο $term_{ij}$ τη χρονική στιγμή t.
- α ο συνολικός αριθμός των χαρακτηριστικών.
- b_i ο αριθμός των εναλλακτικών τιμών για το χαρακτηριστικό i.
- I είναι τα χαρακτηριστικά i που δεν έχουν χρησιμοποιηθεί.

Η ευρετική συνάρτηση η_{ij} που είναι άμεσα εξαρτώμενη από τη φύση του προβλήματος, είναι ένα μέτρο της διακριτικής ικανότητας του όρου $term_{ij}$. Όσο υψηλότερη η τιμή του της ευρετικής

συνάρτησης η_{ij} τόσο πιο σχετικός είναι για την ταξινόμηση ο όρος $term_{ij}$, και έτσι τόσο υψηλότερη η πιθανότητα επιλογής του.

Η τιμή των ιχνών φερομόνης $\tau_{ij}(t)$ είναι επίσης ανεξάρτητο από τους όρους που εμφανίζονται στον τρέχοντα μερικό κανόνα, αλλά εξαρτάται εξ ολοκλήρου από τις πορείες που ακολουθούνται από τα προηγούμενα μυρμήγκια. Ως εκ τούτου, τα ίχνη φερομόνης $\tau_{ij}(t)$ ενσωματώνουν μια έμμεση μορφή επικοινωνίας μεταξύ των μυρμηγκιών, όπου τα μυρμήγκια αφήνουν μια "ένδειξη" (φερομόνη) υποδεικνύοντας την καλύτερη πορεία στα μελλοντικά μυρμήγκια. Όταν το πρώτο μυρμήγκι αρχίζει να χτίζει τον κανόνα του, όλες οι θέσεις ιχνών *i,j* έχουν το ίδιο ποσό φερομόνης.

Στη συνέχεια, μόλις τελειώνει ένα μυρμήγκι την πορεία του τα ποσά φερομόνης σε κάθε θέση *i,j* που επισκέπτεται από το μυρμήγκι ενημερώνονται, όπως θα εξηγηθεί παρακάτω. Με άλλα λόγια η βασική ιδέα είναι ότι όσο πιο μεγάλη είναι η ποιότητα του κανόνα που κατασκευάζεται από το μυρμήγκι, τόσο υψηλότερο είναι το ποσό φερομόνης προστίθεται στα ίχνη που επισκέφτηκε το μυρμήγκι. Ως εκ τούτου, με το χρόνο τα ίχνη των καλύτερων όρων ($term_{ij}$) που προστίθενται σε έναν κανόνα, θα έχουν ολοένα και μεγαλύτερα ποσά φερομόνης, πράγμα που αυξάνει την πιθανότητα επιλογής τους στις μελλοντικές επαναλήψεις.

Ο όρος $term_{ij}$ με την υψηλότερη τιμή στην Σχέση (12) επιλέγεται για να προστεθεί στον τρέχοντα μερικό κανόνα υπό το πρίσμα δύο περιορισμών. Ο πρώτος περιορισμός είναι ότι το χαρακτηριστικό i δεν μπορεί να προστεθεί αν ήδη υπάρχει. Για να ικανοποιηθεί ο περιορισμός αυτός τα μυρμήγκια εμπεριέχουν μια στοιχειώδη μνήμη ώστε να "θυμούνται" τα ζεύγη (χαρακτηριστικό, τιμή) που περιλαμβάνονται στον τρέχοντα μερικό κανόνα.

Ο δεύτερος περιορισμός είναι ότι ένας όρος $term_{ij}$ δεν μπορεί να προστεθεί στον τρέχοντα μερικό κανόνα εάν οι ο κανόνας ικανοποιεί μικρότερο από έναν ελάχιστο αριθμό περιπτώσεων, αποκαλούμενο *Min_cases_per_rule*. Ωστόσο η απόφαση λαμβάνεται μόνο εφόσον ολοκληρωθεί η κατασκευή του κανόνα.

5.1.4 Κατάτμηση Κανόνα

Όπως αναφέρεται ανωτέρω, ο κύριος στόχος της κατάτμησης κανόνα είναι να αφαιρεθούν οι άσχετοι όροι που συμπεριλήφθηκαν αδικαιολόγητα στον κανόνα. Η κατάτμηση κανόνα αυξάνει ενδεχομένως την ποιότητα του κανόνα. Ένα άλλο κίνητρο για την κατάτμηση του κανόνα είναι ότι βελτιώνει την απλότητα αυτού, δεδομένου ότι ένας πιο μικρός κανόνας είναι ευκολότερα ερμηνεύσιμος από το χρήστη σε σχέση με έναν μεγαλύτερο κανόνα.

Η διαδικασία κατάτμησης κανόνα εκτελείται για κάθε μυρμήγκι μόλις αυτό ολοκληρώνει την κατασκευή του κανόνα. Η βασική ιδέα είναι ότι επαναληπτικά αφαιρείται ένας όρος από τον κανόνα και παράλληλα ελέγχεται η ποιότητα του νέου κανόνα σε σχέση με τον προηγούμενο. Πιο συγκεκριμένα, στην πρώτη επανάληψη το κάθε μυρμήγκι αρχίζει με τον πλήρη κανόνα. Κατόπιν δοκιμαστικά αφαιρεί έναν προς έναν τους όρους του κανόνα και υπολογίζει την ποιότητα του νέου κανόνα, με βάση τη σχέση 13. Ο όρος ο οποίος με την αφαίρεσή του βελτιώνει περισσότερο την ποιότητα του κανόνα, αφαιρείται, ολοκληρώνοντας την πρώτη επανάληψη. Αυτή η διαδικασία επαναλαμβάνεται έως ότου στον κανόνα παραμείνει μόνο ένας όρος ή έως ότου δεν υπάρχει κανένας άλλος όρος, του οποίου η αφαίρεση θα βελτιώσει την ποιότητα του κανόνα.

$$Q = \left(\frac{TruePos}{TruePos + FalseNeg} \right) * \left(\frac{TrueNeg}{FalsePos + TrueNeg} \right) \qquad \textbf{Σχέση 13}$$

Όπου:

- *TruePos* είναι ο αριθμός των περιπτώσεων που κάλυψε ο κανόνας και κατέταξε στη σωστή κατηγορία.
- *FalsePos* είναι ο αριθμός των περιπτώσεων που κάλυψε ο κανόνας αλλά κατέταξε σε λάθος κατηγορία.
- *FalseNeg* είναι ο αριθμός περιπτώσεων που δεν καλύπτονται από τον κανόνα αλλά ανήκουν στην κατηγορία που προβλέπει ο κανόνας
- *TrueNeg* είναι ο αριθμός περιπτώσεων που δεν καλύπτονται από τον κανόνα και δεν ανήκουν στην κατηγορία που προβλέπει ο κανόνας

5.1.5 Ενημέρωση Φερομόνης

Κατά την αρχικοποίηση του αλγορίθμου, σε όλους τους όρους $term_{ij}$ ανατίθεται η ίδια ποσότητα φερομόνης, έτσι ώστε όταν αρχίζει το πρώτο μυρμήγκι την αναζήτησή του, όλες οι πορείες έχουν το ίδιο ποσό φερομόνης. Το αρχικό ποσό φερομόνης που ανατίθεται σε κάθε όρο είναι αντιστρόφως ανάλογο προς το άθροισμα των πιθανών τιμών όλων των ιδιοτήτων, και καθορίζεται από τη σχέση 14.

$$\tau_{ij}(t=0) = \frac{1}{\sum_{i=1}^{a} b_i} \qquad \textbf{Σχέση 14}$$

Όπου a είναι ο συνολικός αριθμός των ιδιοτήτων και b_i ο αριθμός των πιθανών τιμών της ιδιότητας A_i.

Η τιμή που επιστρέφεται από την παραπάνω εξίσωση είναι κανονικοποιημένη για τη διευκόλυνση της χρήσης της, στη σχέση 12, η οποία συνδυάζει αυτήν την τιμή και την τιμή της ευρετικής συνάρτησης. Όποτε ένα μυρμήγκι ολοκληρώνει την κατασκευή και την κατάτμηση ενός κανόνα, η τιμή των ιχνών φερομόνης ενημερώνεται. Η ενημέρωση αυτή έχει δύο βασικά χαρακτηριστικά:

- Η τιμή του ίχνους φερομόνης που συνδέεται με κάθε όρο *termij* που εμφανίζεται στον κανόνα που δημιουργήθηκε από το μυρμήγκι αυξάνεται ανάλογα προς την ποιότητα του κανόνα.
- Η τιμή του ίχνους φερομόνης που συνδέεται με κάθε *termij* που δεν εμφανίζεται στον κανόνα μειώνεται, προσομοιώνοντας τη φυσική εξάτμιση των φερομονών στις πραγματικές αποικίες μυρμηγκιών.

Αύξηση φερομόνης

Η αύξηση του ποσού φερομόνης που συνδέεται με κάθε όρο $term_{ij}$ που εμφανίζεται στον κανόνα αντιστοιχεί στην αύξηση του ποσού φερομόνης κατά μήκος της πορείας που ολοκληρώνεται από το μυρμήγκι. Αυτό αντιστοιχεί στην αύξηση της πιθανότητας ο όρος $term_{ij}$ να επιλεγεί μελλοντικά από άλλα μυρμήγκια και η οποία αύξηση είναι ανάλογη προς την ποιότητα του κανόνα που δίνεται από τη σχέση 13. Αντίστοιχα η αύξηση της τιμής της φερομόνης δίνεται από την σχέση 15.

$$\tau_{ij}(t+1) = \tau_{ij}(t) + \tau_{ij}(t) \cdot Q, \forall i, j \in στον\ \ κανόνα \qquad \textbf{Σχέση 15}$$

Ελάττωση φερομόνης

Όπως προαναφέρθηκε, η τιμή φερομόνης, που συνδέεται με κάθε όρο $term_{ij}$ που δεν εμφανίζεται στον κανόνα που δημιουργήθηκε από το τρέχον μυρμήγκι, πρέπει να μειωθεί προκειμένου προσομοιώσει τη φυσική εξάτμιση των φερομονών στις πραγματικές αποικίες μυρμηγκιών. Το μοντέλο εξάτμισης που χρησιμοποιείται μειώνει την τιμή με έμμεσο τρόπο. Ακριβέστερα, η επίδραση της εξάτμισης φερομονών για τους αχρησιμοποίητους όρους επιτυγχάνεται με την κανονικοποίηση της τιμής κάθε ίχνους τ_{ij}. Η κανονικοποιημένη τιμή ουσιαστικά εξάγεται με τη διαίρεση της τιμής του κάθε ίχνους τ_{ij} με το άθροισμα όλων των τιμών των ιχνών. Για την κατανόηση του τρόπου που επιδρά η κανονικοποίηση ως τεχνητή εξάτμιση, αναφέρεται ότι, μόνο οι όροι που χρησιμοποιούνται από έναν κανόνα αυξάνουν την τιμή της φερομόνης τους από την σχέση 15. Επομένως, κατά την κανονικοποίηση, η τιμή της φερομόνης ενός αχρησιμοποίητου όρου θα υπολογιστεί με τη διαίρεση της τρέχουσας τιμής του η οποία δεν τροποποιείται από την εξίσωση (15) με το συνολικό άθροισμα των τιμών της φερομόνης για όλους τους όρους (που αυξήθηκε ως αποτέλεσμα της εφαρμογής της εξίσωσης (14) σε όλους τους χρησιμοποιημένους όρους). Η τελική επίδραση θα είναι η μείωση της κανονικοποιημένης τιμής της φερομόνης για κάθε αχρησιμοποίητο όρο.

6 Αναφορές

[Bonabeau,97] Bonabeau, E. (1997), ``From classical models of morphogenesis to agent-based models of pattern iformation", Artificial Life, Vol. 3, pp. 191-211.

[Bonabeau,99] Bonabeau, E., Dorigo, M., and Theraulaz, G. “Swarm Intelligence: From Natural to Artificial Systems”. Oxford University Press, New York, 1999.

[Darken,01] Darken, R. P. and Sibert, J. L. Wayfinding Strategies and Behaviors in Large Virtual Worlds, In Proceedings of CHI’96, ACM Press, New York, 2001, 142-149.

[Davis,91] Davis, T. and Principe, J.C. (1991) A simulated annealing like convergence theory for the simple genetic algorithm. In Belew, R. and Bookers, L., editors, Proc. of the Fourth International Conference on genetic Algorithms, pages 174-181, San Mateo, CA. Morgan Kaufmann. W. E.

[DeJong,93] K. DeJong, M. Spears, and D. Gordon. Using genetic algorithms for concept learning. Machine Learning, 13:161–188, 1993.

[Dorigo,96] Dorigo, M., Maniezzo, V., and Colorni, A. The ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics-Part B, 26, 1, 29–41, 1996.

[Dorigo,99] Dorigo M. and Caro G.D., 1999, “The Ant Colony Optimization Meta-heuristic,” in New Ideas in Optimization, D. Corne, M. Dorigo, and F. Glover, Eds. London: McGraw-Hill, pp. 11-32.

[Dorigo,04] M. Dorigo and T. St¨utzle. Ant Colony Optimization. The MIT Press, 2004.

[Eberhart,01] R. Eberhart and Y. Shi, “Particle swarm optimization: applications and resources,” Proceedings of the 2001 Congress on Evolutionary Computation, vol. 1 , 27-30 May 200 1, pp. 8 1-86.

[Franks,92] Franks, N.R., Wilby, A., Silverman, B.W. and Tofts, C. (1992), “Self-organizing nest construction in ants: sophisticated building by blind bulldozing”", Animal Behaviour, Vol. 44, pp. 357-75.

[Holland,75] J. Holland, Adaptation in natural and artificial systems (AnnArbor, MI: University of Michigan Press, 1975).

[Kennedy,95] Kennedy, J. & Eberhart, R. C., 1995, “Particle Swarm Optimization.” In Proceeding of the IEEE International Conference on Neural Networks, Perth, Australia, IEEE Service Center, 12-13.

[Kennedy,01] Kennedy, J., & Eberhart, R. C. (2001). Swarm Intelligence (Denise E.M. Penrose). San Francisco: Morgan Kaufmann Publishers.

[Lai,01] Lai W. K., Lim Keh Long and Aw Yit Mei, "Understanding Swarm Intelligence in Solving Combinatorial Optimization Problems", Proceedings of the Third MIMOS R&D Symposium on ICT and Microelectronics, 8th November 2001.

[Løvbjerg,01] M. Løvbjerg, T. Rasmussen, and T. Krink. Hybrid particle swarm optimiser with breeding and subpopulations. In Proceedings of the third Genetic and Evolutionary Computation Conference 2001.

[Michalewcz,92] Z. Michalewcz (1992) Genetic Algorithms + Data Structures = Evolution Programs. Extended Edition, Springer-Verlag.

[Parpinelli,02a] R.S. Parpinelli, H.S. Lopes and A.A. Freitas. An Ant Colony Algorithm for Classification Rule Discovery. In: H.A. Abbass, R.A. Sarker, C.S. Newton. (Eds.) Data Mining: a Heuristic Approach, pp. 191-208. London: Idea Group Publishing, 2002.

[Parpinelli,02b] R.S. Parpinelli, H.S. Lopes, and A.A. Freitas. Data mining with an ant colony optimization algorithm. IEEE Transactions onEvolutionary Computing 6(4), 2002, pp. 321–332.

[Schmitendorgf,92] Schmitendorgf, O. Shaw, R. Benson and S. Forrest, "Using Genetic Algorithms for Controller Design: Simultaneous Stabilization and Eigenvalue Placement in a Region", Technical Report No. CS92-9, Dept. Computer Science, College of Engineering, University of New Mexico, 1992.

[Shi,98] Y. Shi and R.C. Eberhart. A modified particle swarm optimizer. In Proceedings of the IEEE International Conference on Evolutionary computation 1998; 69–73.

[Sole,99] Sole, Ricard V. and Bartolo Luque (1999). Statistical Measures of Complexity for Strongly Interacting Systems." E-print, arxiv.org, adap-org/9909002.

ΚΕΦΑΛΑΙΟ

4

ΜΕΘΟΔΟΣ ΔΙΑΔΙΚΤΥΑΚΗΣ ΑΝΑΖΗΤΗΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΜΕ ΧΡΗΣΗ
ΑΛΓΟΡΙΘΜΟΥ ΑΠΟΙΚΙΑΣ ΜΥΡΜΗΓΚΙΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ 4ου ΚΕΦΑΛΑΙΟΥ

ΜΕΘΟΔΟΣ ΔΙΑΔΙΚΤΥΑΚΗΣ ΑΝΑΖΗΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΟΥ ΑΠΟΙΚΙΑΣ ΜΥΡΜΗΓΚΙΩΝ

1 Εισαγωγή

Στο κεφάλαιο αυτό προτείνεται μια νέα μεθοδολογία διαδικτυακής αναζήτησης πληροφορίας στηριζόμενη στον αλγόριθμο αποικίας μυρμηγκιών. Πιο συγκεκριμένα, προτείνεται μια προσέγγιση του αλγορίθμου αποικίας μυρμηγκιών (Ant-Seeker) με δυνατότητα δρομολόγησης της αναζήτησης πληροφορίας σε δυναμικά περιβάλλοντα, όπως είναι ο παγκόσμιος ιστός. Παράλληλα με τη δρομολόγηση της αναζήτησης, γίνεται χρήση τεχνικών ανάκτησης πληροφοριών για τον εντοπισμό και την αξιολόγηση της πληροφορίας που βασίζονται στην ομοιότητα εγγράφων και επιπρόσθετα γίνεται χρήση μοντέλων συσταδοποίησης εγγράφων. Στη συνέχεια παρουσιάζονται πειραματικές μετρήσεις για την προτεινόμενη μεθοδολογία καθώς και ποιοτικά συμπεράσματα.

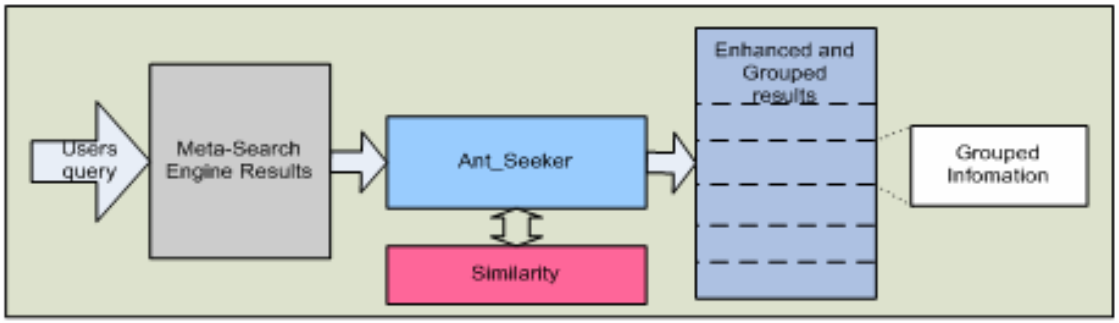
2 Μεθοδολογία

Κεντρικός στόχος της παρούσας διατριβής είναι η εισήγηση μιας μεθοδολογίας η οποία είναι σε θέση να εντοπίζει διαδρομές στον ιστοχώρο που συνδέουν δύο ή και περισσότερες πληροφοριακές μονάδες με σχετικό περιεχόμενο και με το ελάχιστο δυνατό κόστος.

Αρχικά θεωρούμε μια δικτυακή μονάδα πληροφορίας (ιστοσελίδα) σχετική με τις ανάγκες του χρήστη. Αυτή η ιστοσελίδα θα αποτελέσει την αφετηρία της αναζήτησης. Η αναζήτηση στηρίζεται στη φιλοσοφία ότι όταν σε ένα σημείο-κόμβο του παγκόσμιου ιστού υπάρχει μια πληροφορία ανάλογη με τις επιταγές του χρήστη, τότε ένα άλλο σημείο-κόμβος σε "κοντινή απόσταση", θα περιέχει πληροφορία σχετική με αυτό[Kouzas,06]. Ως μονάδα απόστασης στον παγκόσμιο ιστό ορίζεται ο υπερσύνδεσμος. Με την έννοια "απόσταση" μεταξύ δύο κόμβων, μπορούμε να ορίσουμε των αριθμό των υπερσυνδέσμων που απαιτείται να ακολουθήσουμε για την μεταφορά από τον έναν κόμβο στον άλλο. Εδώ θα πρέπει να αναλυθεί η χρησιμότητα των υπερσυνδέσμων στον παγκόσμιο ιστό. Η δημιουργία των υπερσυνδέσμων αποσκοπούσε στην εύκολη μετακίνηση μεταξύ πληροφοριακών μονάδων του παγκόσμιου ιστού. Επίσης η χρήση του επεκτείνεται στη δημιουργία αναφορών περιεχομένων αλλά και στην σύνδεση σχετικών δικτυακών τόπων. Με την αλματώδη εξάπλωση του διαδικτύου τα τελευταία χρόνια χρησιμοποιείται εκτεταμένα και για την προώθηση διαφημίσεων.

2.1 Αρχιτεκτονική Συστήματος

Η μεθοδολογία περιλαμβάνει 3 στάδια. Στο πρώτο στάδιο γίνεται ο ορισμός των μονάδων αναζήτησης που θα αποτελέσουν τα έγγραφα αναφοράς και την αφετηρία της αναζήτησης. Ανάλογα με τις ανάγκες του χρήστη ως σημεία εκκίνησης ορίζονται πληροφοριακές μονάδες που ενδιαφέρουν το χρήστη. Αυτές μπορεί να είναι είτε ιστοσελίδες που ο χρήστης έχει αξιολογήσει στο παρελθόν είτε αποτελέσματα μηχανών αναζήτησης που πληρούν τις ανάγκες του χρήστη. Στο δεύτερο στάδιο λαμβάνει χώρα η αναζήτηση με τη χρήση του προτεινόμενου τροποποιημένου αλγορίθμου αποικίας μυρμηγκιών. Το μέτρο αξιολόγησης, που χρησιμοποιεί ο προτεινόμενος αλγόριθμος για την εύρεση της σχετικής πληροφορίας, βασίζεται ουσιαστικά σε τεχνικές εύρεσης ομοιότητας εγγράφων. Η διαδικασία του αλγορίθμου δρα επαναληπτικά, οπότε κάθε φορά που συγκλίνει σε μια επιθυμητή πληροφοριακή μονάδα, αυτή ορίζει ένα νέο σημείο αφετηρίας. Το τρίτο και τελευταίο στάδιο αποτελεί την ομαδοποίηση των αποτελεσμάτων με βάση το βαθμό συσχέτισης των περιεχομένων τους. Στις επόμενες παραγράφους αναλύονται οι τεχνικές ομοιότητας και συσταδοποίησης των οποίων το θεωρητικό υπόβαθρο αναλύθηκε στο κεφάλαιο 2. Η δομή και λειτουργία του προτεινόμενου αλγορίθμου αναζήτησης αναλύεται στην επόμενη ενότητα.



Σχήμα 1. Αρχιτεκτονική Συστήματος

2.2 Επεξεργασία εγγράφου

Η πληροφορία στον παγκόσμιο ιστό παρουσιάζεται με την μορφή εγγράφων κειμένων υπερκειμένου (HTML). Το μεγαλύτερο μέρος της πληροφορίας ενός τέτοιου εγγράφου χρησιμοποιείται για την παρουσίαση της βασικής πληροφορίας στο χρήστη με ένα πιο φιλικό τρόπο όπως για παράδειγμα πίνακες υπογραμμίσεις χρωματισμός και λοιπά. Ωστόσο, βάσει της δομής του εγγράφου, είναι δυνατή η αναγνώριση σημαντικών σημείων αυτού. Αυτό προκύπτει από το γεγονός ότι κάποια μέρη του εγγράφου δύναται να περιέχουν περισσότερο όγκο πληροφοριών από άλλα, με αποτέλεσμα να καθίστανται πιο σημαντικά μέσα στο κείμενο. Για παράδειγμα η επικεφαλίδα και το κυρίως σώμα του κειμένου δεν γίνεται να έχουν την ίδια αντιμετώπιση.

Κατά την προεπεξεργασία των εγγράφων υπερκειμένων, το πρώτο βήμα είναι η αποκοπή της κειμενικής πληροφορίας από το σώμα των HTML ετικετών και αποθήκευση. Στη συνέχεια γίνεται διαχωρισμός των υπερσυνδέσμων προς άλλα υπερκείμενα, για τον σχηματισμό του γράφου στον οποίο λαμβάνει χώρα η αναζήτηση. Το τελικό βήμα της προεπεξεργασίας είναι ο υπολογισμός των συχνοτήτων των όρων για κάθε έγγραφο και ο υπολογισμός της ομοιότητας των εγγράφων. Στην παρούσα μεθοδολογία για την εύρεση της ομοιότητας εγγράφων χρησιμοποιείται η τεχνική του [Hammouda,03] στην οποία λαμβάνεται υπόψη και η δομή του εγγράφου και αποτελείται από το στάδιο προεπεξεργασίας εγγράφου, και το στάδιο υπολογισμού της ομοιότητας.

2.2.1 Προεπεξεργασία

Ανάλογα με τις ετικέτες που περικλείουν το κείμενο, ορίζονται τρία επίπεδα σημασίας : το ΥΨΗΛΟ, το ΜΕΣΑΙΟ και το ΧΑΜΗΛΟ. Παραδείγματα μερών υψηλής σημασίας αποτελούν ο τίτλος, οι επικεφαλίδες των παραγράφων και οι λέξεις-κλειδιά των μετά-ετικετών. Παραδείγματα μερών μεσαίας σημασίας είναι τα τμήματα που εμφανίζονται έγχρωμα, υπογραμμισμένα, πλάγια ή έντονα, ενώ χαμηλής σημασίας παράδειγμα αποτελεί το υπόλοιπο κυρίως σώμα του κειμένου. Ο πίνακας 1 δείχνει ορισμένες ετικέτες και το επίπεδο στο οποίο κατατάσσονται. Η ανωτέρω προσέγγιση διευκολύνει τη σύγκριση των ομοιοτήτων δυο εγγράφων. Για παράδειγμα θεωρείται ότι δυο κείμενα διαπραγματεύονται κοινό θέμα, όταν κατά τη σύγκρισή τους εμφανίζονται υψηλής σημασίας μέρη και όχι χαμηλής, δηλαδή όταν συμφωνούν οι επικεφαλίδες τους και όχι κάποια πρόταση που εμφανίζεται κοινή στο κυρίως σώμα τους.

Ετικέτα HTML	Επίπεδο σημαντικότητας
<TITLE> <META NAME=”description”> <META NAME=”keyword”> <HEAD>	3
<H1>,<H2> ,<st> <B>,<I>,<U> <TABLE>	2
<P>	1

Πίνακας 2. Επίπεδο σημαντικότητας HTML Ετικετών

2.3 Προσδιορισμός Ομοιότητας Εγγράφων

Όπως αναφέρθηκε και παραπάνω, οι προτάσεις μεταφέρουν πληροφορίες σχετικά με το περιεχόμενο του κειμένου που είναι πολύ σημαντικές για τον καθορισμό της ομοιότητας των εγγράφων. Με βάση αυτό χρησιμοποιήθηκε ένα μέτρο σύγκρισης που βασίζεται όχι μόνο στην ομοιότητα των μεμονωμένων όρων αλλά και στην ομοιότητα των προτάσεων σύμφωνα με την αναφορά [Isaacs,99]. Το μέτρο αυτό, εκμεταλλεύεται τις πληροφορίες που ανακτώνται από τον προηγούμενο αλγόριθμο για να επανεξετάσει τις ομοιότητες μεταξύ των κειμένων.

Η ομοιότητα των προτάσεων μεταξύ δυο κειμένων υπολογίζεται βάσει της λίστας των κοινών φράσεων τους. Αυτό το μέτρο σύγκρισης εξαρτάται από τέσσερις παράγοντες :

- Τον αριθμό των κοινών φράσεων P
- Το μήκος των κοινών φράσεων (l_i : $i=1,2,\ldots\ldots P$)
- Τη συχνότητα των κοινών φράσεων στα δυο κείμενα (f_{i1} και f_{i2}: i=1,2,…,P)
- Τα επίπεδα σημασίας (βάρος) των κοινών φράσεων στα δυο κείμενα (w_{i1} και w_{i2} i=1,2,……..P)

Η συχνότητα των φράσεων αποτελεί πολύ σημαντικό παράγοντα για το μέτρο σύγκρισης. Όσο πιο συχνά εμφανίζεται μια φράση μέσα στα έγγραφα τόσο πιο σχετικά τείνουν να είναι. Αντίστοιχα πρέπει να λαμβάνεται υπόψη και το επίπεδο σημασίας των όμοιων φράσεων. Η ομοιότητα μεταξύ δύο εγγράφων, d_1 και d_2 υπολογίζεται βάσει της Σχέσης 1.

$$S_p(d_1,d_2)=\frac{\sqrt{\sum_{i=1}^{P}[g(l_i)\cdot(f_{i1}w_{i1}+f_{i12}w_{i2})]^2}}{\sum_j|s_{j1}|\cdot w_{j1}+\sum_j|s_{k2}|\cdot w_{k2}} \qquad \textbf{Σχέση 1}$$

$$g(l_i)=\left(\frac{|ms_i|}{|s_i|}\right)^{\gamma} \qquad \textbf{Σχέση 2}$$

όπου $g(l_i)$ μια συνάρτηση που βαθμολογεί το μήκος της κοινής φράσης και δίνει υψηλότερη βαθμολογία όσο το μήκος της κοινής φράσης προσεγγίζει το μήκος της αρχικής ενώ το $|s_{j1}|$ και $|s_{k2}|$ αναπαριστούν το αρχικό μήκος των προτάσεων των εγγράφων d_1 και d_2 αντίστοιχα. Η συνάρτηση $g(l_i)$ είναι ανάλογη του λόγου του μήκους του κοινού τμήματος της πρότασης προς το συνολικό μήκος της πρότασης και δίνεται από τη σχέση 2. Το γ είναι ένας δείκτης τεμαχισμού της πρότασης με τιμή μεγαλύτερη ή ίση του 1. Αν το γ είναι ίσο με 1, τα δυο μισά μιας πρότασης μπορούν να βρουν αντιστοιχία, ανεξάρτητα το ένα από το άλλο, Αυξάνοντας το γ σε τιμές >1 τότε οι ολόκληρες προτάσεις βαθμολογούνται υψηλότερα από τα επιμέρους τμήματά τους. Μια τιμή της τάξεως του 1,2 για το γ επιφέρει τα βέλτιστα αποτελέσματα.

Παράλληλα με την ομοιότητα μεταξύ των φράσεων των δύο κειμένων γίνεται και ο υπολογισμός της ομοιότητας με βάση της συχνότητας των κοινών όρων μεταξύ των δυο εγγράφων. Ο υπολογισμός με βάση τη συχνότητα των όρων έγινε με τη χρήση του χωροδιανυσματικού μοντέλου [Salton,68], [Salton,71]όπως περιγράφεται στο Κεφάλαιο 2 παρ. 3.3.8 και δίνεται από τη Σχέση 3. Ωστόσο στην ανάθεση βαρών δεν λαμβάνεται υπόψη η αντίστροφη συχνότητα εμφάνισης του κάθε όρου γιατί η ομοιότητα γίνεται αποκλειστικά μεταξύ δυο εγγράφων και όχι μιας συλλογής. Η αντίστροφη συχνότητα εμφάνισης στην αναζήτηση δεν έχει νόημα διότι δεν

υπάρχει προκαθορισμένος αριθμός εγγράφων όπως απαιτεί το χωροδιανυσματικό μοντέλο [Salton,88].

$$sim(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i| \times |d_j|} = \frac{\sum_{k=1}^{t} (w_{k,i} \times w_{k,j})}{\sqrt{\sum_{k=1}^{t} w_{k,i}^2} \times \sqrt{\sum_{k=1}^{t} w_{k,j}^2}}$$ **Σχέση 3**

Οπότε η τελική τιμή ομοιότητας δυο εγγράφων δίνεται από τη Σχέση 4.

$$S(d_1, d_2) = 0.5 S_p(d_1, d_2) + 0.5 S_t(d_1, d_2)$$ **Σχέση 4**

2.4 Ομαδοποίηση Αποτελεσμάτων

Η ομαδοποίηση εγγράφων χρησιμοποιείται για την καλύτερη παρουσίαση των αποτελεσμάτων της αναζήτησης. Οι ανακατωμένες πληροφοριακές μονάδες αναλύονται και παρουσιάζονται σε συστάδες. Η κάθε συστάδα περιέχει ένα σύνολο πληροφοριακών μονάδων υψηλής σχετικότητας. Η δημιουργία των συστάδων βασίζεται στη μεθοδολογία που προτείνεται στο [Hammouda,03] και βασίζεται στην ανάλυση ιστογράμματος ομοιοτήτων. Για τον προσδιορισμό ομοιότητας χρησιμοποιείται το κλασικό χωροδιανυσματικό μοντέλο [Salton,68], [Salton,71].

3 Ant - Seeker

3.1 Εισαγωγή

Στη ενότητα αυτή παρουσιάζεται ο προτεινόμενος αλγόριθμος αναζήτησης, ο οποίος ουσιαστικά αποτελεί μια τροποποίηση του θεωρητικού μοντέλου λειτουργίας του αλγορίθμου αποικίας μυρμηγκιών που αναφέρθηκε στο κεφάλαιο 3 [Dorigo,04]. Η αρχή λειτουργίας του στηρίζεται στη θεωρητική αρχή σύγκλισης και ουσιαστικά αποτελεί μια νέα προσέγγιση στον τρόπο αναζήτησης μέσα στον παγκόσμιο ιστό. Αν και διατηρεί τις περισσότερες ιδιότητες που χαρακτηρίζουν τους αλγορίθμους αποικίας μυρμηγκιών, η ιδιαιτερότητά του έγκειται στο γεγονός ότι η εφαρμογή του λαμβάνει χώρα σε ένα δυναμικό περιβάλλον του οποίου η δομή δεν είναι εξ' αρχής ορισμένη. Επιπρόσθετα ως ευρετική συνάρτηση χρησιμοποιεί τεχνικές επεξεργασίας πληροφορίας κατ' αναλογία με τον αλγόριθμο Ant-miner [Parepinelli,02] και πιο συγκεκριμένα της ομοιότητας εγγράφων που ορίζεται στην προηγούμενη ενότητα από τη Σχέση 4.

3.2 Ορισμός Προβλήματος

Ο προτεινόμενος αλγόριθμος πραγματεύεται το πρόβλημα αναζήτησης πληροφορίας στον παγκόσμιο ιστό. Η δομή του παγκόσμιου ιστού αποτελείται ουσιαστικά από ένα σύνολο πληροφοριακών μονάδων (ιστοσελίδες) και ένα σύνολο συνδέσεων (υπερσύνδεσμοι) μεταξύ αυτών. Μεταφέροντας το πρόβλημα της αναζήτησης σε έναν γράφο $G_p = (P, L)$ με κόμβους τα δομικά συστατικά P, και L το σύνολο των συνδέσεων μεταξύ των κόμβων, μπορούμε να θεωρήσουμε τον παγκόσμιο ιστό ως έναν γράφο G με άπειρες διαστάσεις. Οι κόμβοι του γράφου πλέον αντιστοιχούν στις ιστοσελίδες ενώ οι υπερσύνδεσμοι καθορίζουν τις συνδέσεις μεταξύ των κόμβων. Στην προκείμενη περίπτωση ο στόχος δεν είναι η σάρωση του γράφου και η εύρεση βέλτιστης διαδρομής αλλά η εύρεση κόμβων με σχετικό περιεχόμενο και των διαδρομών που συνδέουν αυτούς. Ο καθορισμός σχετικότητας μεταξύ των δύο κόμβων γίνεται με την Σχέση 4 που αναφέρθηκε στη προηγούμενη ενότητα.

3.3 Προτεινόμενος Αλγόριθμος (Ant-Seeker)

Ο προτεινόμενος αλγόριθμος ουσιαστικά αποτελεί μια τροποποίηση του αλγορίθμου αποικίας μυρμηγκιών [Kouzas, 06a] και ως εκ τούτου υιοθετεί τα περισσότερα βασικά χαρακτηριστικά της οικογένειας αποικίας αλγορίθμων [Dorigo,96]. Πιο συγκεκριμένα:

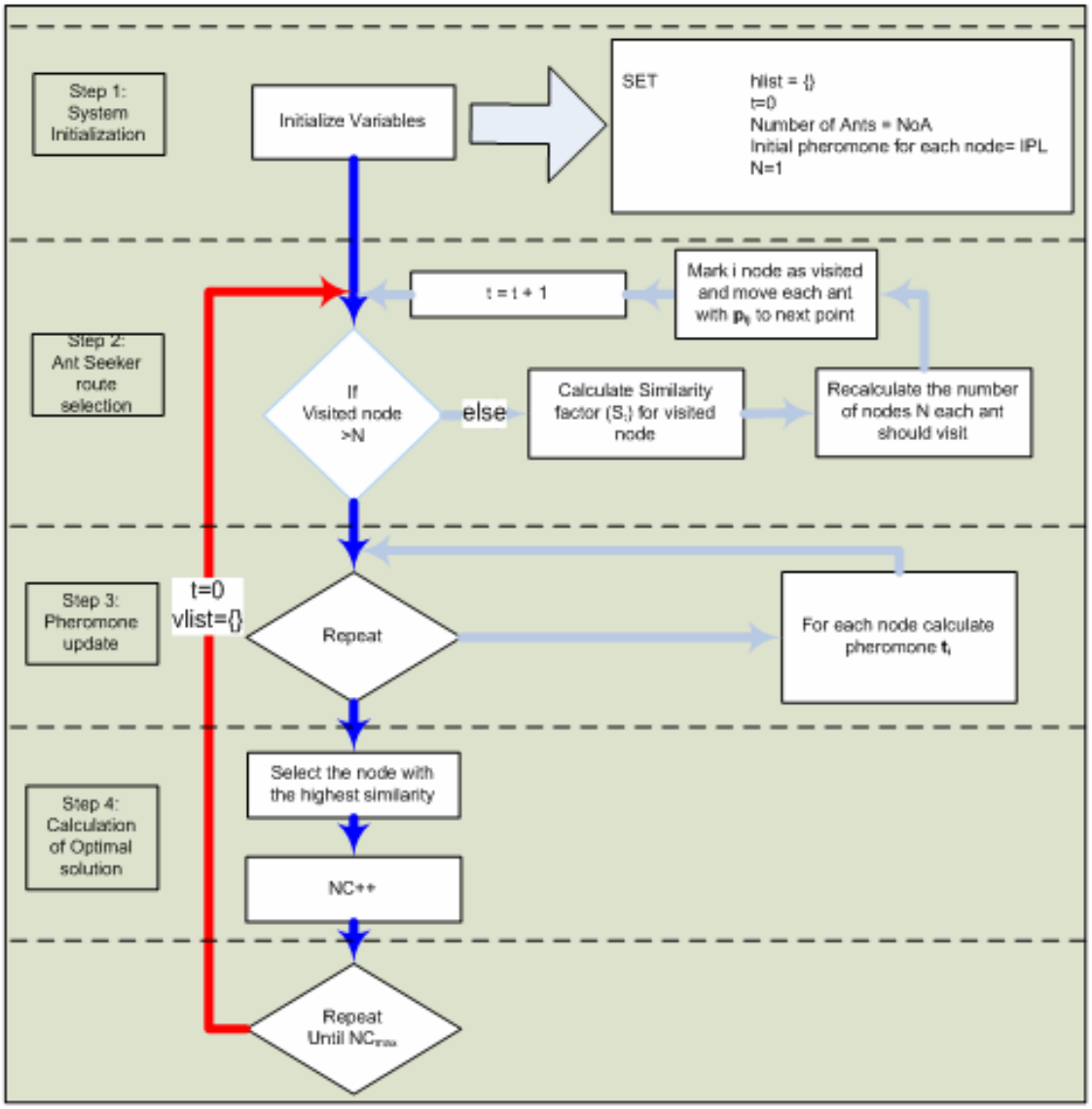
- Τα τεχνητά μυρμήγκια «πράκτορες» επιλέγουν τον επόμενο κόμβο με βάση το σύνολο των ερεθισμάτων που δέχθηκαν στον προηγούμενο κόμβο.
- Το σύνολο των ερεθισμάτων του προηγούμενου κόμβου είναι η πληροφορία που εναπόθεσαν στον συγκεκριμένο κόμβο τα υπόλοιπα μυρμήγκια πράκτορες σε προηγούμενη χρονική στιγμή και προσομοιώνει την φερομόνη των πραγματικών μυρμηγκιών.
- Οι αποδοτικότερες μετακινήσεις μεταξύ δύο κόμβων συγκεντρώνουν μεγαλύτερο σύνολο ερεθισμάτων με αποτέλεσμα τη συχνότερη χρήση από τα μέλη του συστήματος κατά αναλογία με τον πραγματικό κόσμο όπου τα μικρότερα μονοπάτια τείνουν να αυξάνουν το επίπεδο της φερομόνης με μεγαλύτερο ρυθμό.

Εκτός των βασικών χαρακτηριστικών τα οποία ουσιαστικά δίνουν τη δυνατότητα εξερεύνησης και αναζήτησης λύσεων, η εφαρμογή του στο συγκεκριμένο πρόβλημα απαιτεί την προσθήκη επιπρόσθετων ιδιοτήτων:

- Το κάθε τεχνητό μυρμήγκι μπορεί να επισκεφτεί έναν μέγιστο αριθμό κόμβων. Ο περιορισμός αυτός είναι απαραίτητος για την ορθή λειτουργία του αλγορίθμου. Ο παγκόσμιος ιστός δεν αποτελεί γράφο πεπερασμένων διαστάσεων και ως εκ' τούτου η ατέρμονα αναζήτηση ενός τεχνητού μυρμηγκιού «πράκτορα» θα οδηγούσε στην καλύτερη περίπτωση σε σάρωση ολόκληρου του παγκόσμιου ιστού (crawler) ενώ ο χρόνος εκτέλεσης θα καθιστούσε απαγορευτική την εφαρμογή του αλγορίθμου.
- Όλα τα τεχνητά μυρμήγκια ξεκινούν από τον κόμβο αφετηρία. Ουσιαστικά κατά την εκκίνηση της αναζήτησης δεν υπάρχει κανένα στοιχείο για τη δομή του τμήματος του γράφου στον οποίο θα εφαρμοστεί η αναζήτηση, παρά μόνον ένα σημείο εκκίνησης. Με αυτόν τον τρόπο προσομοιώνεται μια αποικία μυρμηγκιών και η εκκίνηση της διαδικασίας εξερεύνησης της γύρω περιοχής για την εύρεση τροφής.
- Η διαδικασία αναγνώρισης κόμβων με σχετικό προς τον αρχικό κόμβο περιεχόμενο βασίζεται στον παράγοντα συσχέτισης εγγράφων όπως αναλύθηκε στην προηγούμενη ενότητα.

Η εκκίνηση της αναζήτησης γίνεται από τον αρχικό κόμβο αφετηρία ο οποίος δίνεται από το χρήστη και μπορεί να είναι είτε μια σελίδα ενδιαφέροντος, είτε αποτέλεσμα μιας μηχανής αναζήτησης. Σε κάθε βήμα του αλγορίθμου κάθε μυρμήγκι πράκτορας κινείται από έναν κόμβο *i* προς έναν κόμβο *j*. Η δυνατότητα μετακίνησης από έναν κόμβο *i* προς έναν κόμβο *j* καλείται προσβασιμότητα και είναι εφικτή μόνον εφόσον οι δυο κόμβοι είναι άμεσα συνδεδεμένοι. Αν υποτεθεί ότι στον κόμβο *j* η τιμή της ποσότητας της φερομόνης τη στιγμή *t* είναι $\tau_j(t)$, τότε η επιλογή του κόμβου *j* από τον κόμβο *i* είναι συνάρτηση της φερομόνης. Γενικά όσο πιο μεγάλη η τιμή της φερομόνης τόσο πιο μεγάλη η πιθανότητα επιλογής του κόμβου από τα μυρμήγκια πράκτορες. Η διαδικασία επιλογής κόμβων επαναλαμβάνεται μέχρι το κάθε μυρμήγκι να επισκεφτεί έναν μέγιστο αριθμό κόμβων. Μετά την ολοκλήρωση της δημιουργίας των διαδρομών από τα μυρμήγκια πράκτορες, υπολογίζεται η καλύτερη διαδρομή και ανανεώνονται οι τιμές φερομόνης των κόμβων. Η άνω διαδικασία επαναλαμβάνεται μέχρι να υπάρξει σύγκλιση σε κάποια συγκεκριμένη διαδρομή. Ο κόμβος της διαδρομής αυτής με την μέγιστη τιμή ομοιότητας προς τον κόμβο αφετηρία επιλέγεται και στη συνέχεια ορίζεται ως αφετηρία για νέα αναζήτηση. Το σχήμα 2 αναπαριστά την αρχή λειτουργίας του αλγορίθμου. Κατά την αρχικοποίηση του αλγορίθμου, ορίζονται και οι παρακάτω μεταβλητές.

- Ο συνολικός αριθμός μυρμηγκιών πρακτόρων *NoA*, που λαμβάνουν μέρος στην αναζήτηση. Όσο μεγαλύτερος είναι ο αριθμός των μυρμηγκιών τόσο καλύτερα είναι τα αποτελέσματα της αναζήτησης. Ωστόσο μεγάλος αριθμός μυρμηγκιών οδηγεί σε μεγάλους χρόνους εκτέλεσης ενώ η απόδοση δεν αυξάνει γραμμικά με τον αριθμό μυρμηγκιών που χρησιμοποιείται.
- Σε κάθε νέο κόμβο που προστίθεται στον γράφο, ορίζεται μια αρχική τιμή φερομόνης *IPV*. Η αρχική τιμή δεν πρέπει να είναι πολύ μεγάλη για να μην αναγκάζει την αναζήτηση να κατευθυνθεί αποκλειστικά προς ανεξερεύνητους κόμβους, αλλά ούτε πολύ μικρή που αποτρέπει την προσθήκη νέων κόμβων.
- Ο αριθμός N_{max} είναι ο μέγιστος αριθμός κόμβων που μπορεί να επισκεφτεί ένα μυρμήγκι μέχρι την ολοκλήρωση της διαδρομής του. Η τιμή αυτή ουσιαστικά προσδιορίζει και το βάθος της αναζήτησης.



Σχήμα 2. Διάγραμμα ροής του προτεινόμενου Αλγορίθμου

3.3.1 Ευρετική Συνάρτηση

Κατά τη δημιουργία μιας διαδρομής από ένα μυρμήγκι, γίνεται έλεγχος της κειμενικής ομοιότητας του κάθε κόμβου που προστίθεται στη διαδρομή, με το αρχικό έγγραφο αναφοράς. Οπότε κάθε κόμβος χαρακτηρίζεται από μια τιμή ομοιότητας, η οποία υποδηλώνει και την ποιότητα ενός κόμβου και δίνεται από την σχέση 4 όπως αναλύθηκε στην προηγούμενη ενότητα.

$$S_i = S(d_1, d_2) = 0.5S_p(d_1, d_2) + 0.5S_t(d_1, d_2) \qquad \textbf{Σχέση 5}$$

Για να αποτυπωθεί πλήρως η ποιότητα χρήσης ενός κόμβου, εκτός από τη σχέση ομοιότητας με το αρχικό έγγραφο, θα πρέπει να τον χαρακτηρίζει και μια δεύτερη τιμή η οποία υποδηλώνει ότι πιθανόν οδηγεί σε κόμβο υψηλής ποιότητας. Αν θεωρηθεί ότι ο παγκόσμιος ιστός είναι μια

κατανομή ομοιοτήτων εγγράφων στο χώρο, τότε η κατανομή αυτή παρουσιάζει πυκνώματα και αραιώματα (υψηλές και χαμηλές τιμές). Κατά τη διαδικασία αναζήτησης περιοχών υψηλής πυκνότητας, η διαδρομή ενδέχεται να διασχίζει περιοχές χαμηλής πυκνότητας. Συνεπώς, τα σημεία της περιοχής με χαμηλή τιμή ομοιότητας αυξάνουν την σπουδαιότητά τους όταν οδηγούν σε περιοχές υψηλής ποιότητας. Με άλλα λόγια, εάν ένας κόμβος με χαμηλή τιμή ομοιότητας προς το αρχικό έγγραφο, οδηγεί σε κόμβο υψηλής τιμής ομοιότητας, τότε η ποιότητα αυτού του κόμβου κρίνεται ως υψηλή, οπότε έχει και μεγαλύτερη πιθανότητα να επιλεγεί. Ο υπολογισμός της ποιότητας ενός κόμβου αποτελεί ουσιαστικά την ευρετική συνάρτηση και δίνεται από την Σχέση 6:

$$h_i(t+1) = \max_{i<j<N_{\max}}(S_j^d , S_i, h_i(t)) \qquad \textbf{Σχέση 6}$$

Όπου d είναι η διαδρομή ενός μυρμηγκιού στη δομή της οποίας συμμετέχει ο κόμβος i ($0 < d < NoA$), S_i η συνάρτηση ομοιότητας του κόμβου i που δίνεται από τη σχέση 5, S_j^d η συνάρτηση ομοιότητας του κόμβου j που αποτελεί μέρος της διαδρομής d μετά όμως από τον κόμβο i, δηλαδή ισχύει $i < j < N_{\max}$.

3.3.2 Μοντέλο Φερομόνης

Αρχικά ο κάθε κόμβος που εισάγεται στο γράφο, έχει μια αρχική τιμή φερομόνης *IPV*. Για τη διαδικασία ανανέωσης φερομόνης, επιλέγεται το δεύτερο μοντέλο που αναγράφεται στο 3ο κεφάλαιο (5.2.3), δηλαδή η ανανέωση λαμβάνει χώρα στους κόμβους οι οποίοι χρησιμοποιήθηκαν ως ενδιάμεσοι ή τελικοί σταθμοί στις διαδρομές των μυρμηγκιών και δίνεται από τη Σχέση 7 και 8.

$$\Delta\tau_i = kh_i \qquad \textbf{Σχέση 7}$$

$$\tau_i(t+1)' = \tau_i(t) + \Delta\tau_i \qquad \textbf{Σχέση 8}$$

Όπου h_i η ευρετική συνάρτηση που δίνεται από τη Σχέση 6 ενώ k είναι ο αριθμός των μυρμηγκιών που χρησιμοποίησαν τον κόμβο i για τη δημιουργία της διαδρομής τους. Σύμφωνα με τη σχέση 8, οι κόμβοι που οδηγούν σε υψηλής ποιότητας διαδρομές, αυξάνουν σημαντικά τη φερομόνη τους με το πέρασμα των επαναλήψεων του αλγορίθμου. Για την αποφυγή απεριόριστων εναποθέσεων φερομονών σε ορισμένους κόμβους, η τιμή της φερομόνης κανονικοποιείται και λαμβάνει τιμές στο διάστημα [0,1] σύμφωνα με τη σχέση 9.

$$\tau_i(t+1) = \frac{\tau_i(t+1)'}{\tau_{\max}(t+1)} \qquad \textbf{Σχέση 9}$$

Όπου $\tau_{\max}(t+1)$, η μέγιστη τιμή φερομόνης που εμφανίστηκε στην τρέχουσα επανάληψη του αλγορίθμου. Επίσης για κάθε κόμβο ο οποίος δεν έχει εξερευνηθεί, ορίζεται μια αρχική τιμή φερομόνης *IPV*. Η αρχική τιμή της φερομόνης ορίζεται επίσης στο διάστημα [0,1] και καθορίζει την κατεύθυνση της αναζήτησης προς ανεξερεύνητους ή όχι κόμβους. Η χρήση της κανονικοποίησης βελτιώνει την ικανότητα αναζήτησης του αλγορίθμου διότι κόμβοι χαμηλής ποιότητας αφαιρούνται από τον γράφο αναζήτησης ενώ τη θέση τους καταλαμβάνουν ανεξερεύνητοι κόμβοι.

3.3.3 Επιλογή Κόμβων

Κάθε στιγμή που ένα τεχνητό μυρμήγκι βρίσκεται σε έναν κόμβο *i*, καλείται να επιλέξει τον επόμενο κόμβο *j* εφόσον η διαδρομή του δεν έχει ολοκληρωθεί. Κάθε φορά, οι εναλλακτικοί προς επίσκεψη κόμβοι είναι οι κόμβοι που συνδέονται άμεσα με τον τρέχοντα κόμβο, δηλαδή οι σελίδες στις οποίες κάνουν αναφορά οι υπερσύνδεσμοι τις τρέχουσας σελίδας, και ορίζεται ως προσβασιμότητα. Η προσβασιμότητα δίνεται από τη Σχέση 10. Για την αποφυγή κυκλικών διαδρομών και κατ' επέκταση δημιουργίας λιμνάζουσας κατάστασης, η προσβασιμότητα αποκλείει κόμβους οι οποίοι ήδη αποτελούν μέρος της διαδρομής. Με αυτόν τον τρόπο επιτυγχάνεται μια συνέχεια της κίνησης στον γράφο καθώς και η δημιουργία φυσικών διαδρομών στον παγκόσμιο ιστό με αρχή και τέλος. Ο αλγόριθμος χρησιμοποιεί το κλασικό πιθανοτικό μοντέλο επιλογής των αλγορίθμων αποικίας μυρμηγκιών και δίνεται από τη Σχέση 11.

$$\eta_{ij} = \begin{cases} 1 & \text{if node j is directly linked from node i} \\ 0 & \textit{otherwise} \end{cases}$$ **Σχέση 10**

$$\mathrm{P}_{ij} = \frac{\tau_j \cdot \eta_{ij}}{\sum_{k \in allowed_k} \tau_k \cdot \eta_{ik}}$$ **Σχέση 11**

3.3.4 Εύρεση Λύσης

Σε κάθε επανάληψη του αλγορίθμου τα μυρμήγκια δημιουργούν μια διαδρομή το καθένα, με βάση την τιμή φερομόνης και την ποιότητα των κόμβων που βρίσκονται στο γράφο αναζήτησης. Όπως αναφέρθηκε η επιλογή του εκάστοτε κόμβου γίνεται με τη σχέση πιθανότητας που περιγράφεται στην προηγούμενη παράγραφο. Όσο προστίθενται επαναλήψεις κατά την εφαρμογή του αλγορίθμου, οι κόμβοι που παρουσιάζουν υψηλότερη ποιότητα αυξάνουν την τιμή φερομόνης τους, οπότε και έχουν μεγαλύτερη πιθανότητα να επιλεγούν. Ουσιαστικά το μοντέλο φερομόνης που ακολουθείται τείνει να δημιουργεί διαδρομές υψηλής και χαμηλής ποιότητας, εφόσον η ποιότητα του κόμβου εξαρτάται και από την ποιότητα της διαδρομής στην οποία συμμετέχει. Οπότε όταν αναφερόμαστε σε λύση αναφερόμαστε σε επιλογή διαδρομής και όχι μεμονωμένου κόμβου. Τελικά, μετά από έναν αριθμό επαναλήψεων το σύνολο των μυρμηγκιών τείνει να επιλέξει τη διαδρομή με τη μέγιστη τιμή της φερορμόνης. Ως λύση ορίζεται ο κόμβος της τελικής διαδρομής που έχει την μεγαλύτερη τιμή ομοιότητας. Ο κόμβος αυτός προστίθεται στη λίστα των λύσεων του αλγορίθμου. Για το πρόβλημα της αναζήτησης στον παγκόσμιο ιστό δεν απαιτείται η εύρεση μιας βέλτιστης λύσης αλλά η επιστροφή ενός συνόλου αποτελεσμάτων σχετικών με το αρχικό έγγραφο ερώτημα. Οπότε ο αλγόριθμος δρα επαναληπτικά μέχρι να σχηματιστεί ένα σύνολο αποτελεσμάτων, ικανοποιητικό προς τις ανάγκες του χρήστη.

Σημαντικό ρόλο στην αναζήτηση του αλγορίθμου, κατέχουν οι μεταβλητές: του αριθμού των μυρμηγκιών *NoA*, του αριθμού επαναλήψεων ανά εφαρμογή του αλγορίθμου *NC*, της αρχικής τιμής της φερορμόνης *IPV* σε κάθε νέο κόμβο που προστίθεται στον γράφο αναζήτησης και του μέγιστου αριθμού κόμβων N_{max}, που μπορεί να επισκεφτεί το κάθε μυρμήγκι πριν σχηματίσει την διαδρομή του. Ο τρόπος που επηρεάζουν την αναζήτηση οι παραπάνω μεταβλητές, αναλύεται στην επόμενη ενότητα.

4 Αποτελέσματα

Η ενότητα αυτή παρουσιάζει τα αποτελέσματα της προτεινόμενης μεθοδολογίας. Ουσιαστικά η αξιολόγηση έγινε σε δυο στάδια πειραμάτων. Στο πρώτο στάδιο πειραμάτων αξιολογείται η απόδοση του προτεινόμενου αλγορίθμου Ant-Seeker [Kouzas, 06a], [Kouzas, 06b] εφαρμόζοντάς τον σε τρία διαφορετικά ερωτήματα. Στο δεύτερο αξιολογείται και η προσθήκη στην μεθοδολογία των τεχνικών συσταδοποίησης για την ομαδοποίηση των επιστρεφομένων αποτελεσμάτων. Για την πραγματοποίηση και ανάλυση των αποτελεσμάτων απαιτήθηκε η ολοκληρωμένη σάρωση των περιοχών αναζήτησης πράγμα που αποτέλεσε μια επίπονη διαδικασία.

4.1 Αποτελέσματα Αναζήτησης

Η διαδικασία της αναζήτησης εξετάστηκε σε τρία διαφορετικά ερωτήματα σε αντίστοιχα τμήματα του παγκόσμιου ιστού. Στην όλη διαδικασία χρησιμοποιήθηκαν μόνο ιστοσελίδες των οποίων το περιεχόμενο ήταν στην αγγλική γλώσσα. Για την ελαχιστοποίηση του χρόνου εκτέλεσης και την καλύτερη αξιολόγηση των αποτελεσμάτων του αλγορίθμου έγινε η καταφόρτωση των περιοχών του παγκόσμιου ιστού γύρω από το έγγραφο αναφοράς-ερώτημα όπως δείχνει ο πίνακας 2.

Δείγματα	Αριθμός Ιστοσελίδων	Αριθμός Ιστοσελίδων προς Αναζήτηση
1	140594	43
2	192155	34
3	175977	72

Πίνακας 2. Δείγματα αξιολόγησης του αλγορίθμου. Αριθμός συνολικών ιστοσελίδων και αριθμός σχετικών σελίδων προς το ερώτημα

Εδώ θα πρέπει να αναφερθεί ότι η διαδικασία καταφόρτωσης ενός μεγάλου αριθμού σελίδων από τον παγκόσμιο ιστό υπήρξε μια επίπονη διαδικασία. Λόγω της α^x εξάπλωσης του παγκόσμιου ιστού, όπου α ο μέσος αριθμός υπερσυνδέσμων ανά ιστοσελίδα, ο συνολικός αριθμός ιστοσελίδων που καταφορτώνεται για μια αναζήτηση σε βάθος 10 επιπέδων είναι πρακτικά τεράστιος. Ενδεικτικά για έναν μέσο όρο 10 υπερσυνδέσμων ανά ιστοσελίδα, που συνήθως είναι ένας λογικός αριθμός, ο συνολικός αριθμός ιστοσελίδων που καταφορτώνονται για τη σάρωση βάθους 10 επιπέδων είναι της τάξης των 10^{10}!

Επίπεδο καταφόρτωσης	**Συνολικές σελίδες καταφόρτωσης**
1	61
2	579
3	4057
4	16307

Πίνακας 3. Αριθμός σελίδων ανά επίπεδο καταφόρτωσης του Δικτυακού τόπου του ΕΜΠ

Ο πίνακας 3 περιγράφει των αριθμό ιστοσελίδων που καταγράφτηκαν κατά την καταφόρτωση 4 μόνο επιπέδων με αφετηρία τη αρχική σελίδα του ιστοχώρου του ΕΜΠ. Στα δείγματα που

επιλέχθηκαν ο αριθμός των σελίδων περιορίστηκε σε έναν αριθμό κάτω των 200000 με αποτέλεσμα το βάθος καταφόρτωσης να περιοριστεί σε έναν αριθμό από 5 έως 7.

Για την εφαρμογή και αξιολόγηση του αλγορίθμου, ακολουθήθηκαν τρία βήματα. Στο πρώτο βήμα έλαβε χώρα η προεπεξεργασία των ιστοσελίδων, στο δεύτερο βήμα η ισοστάθμιση των μεταβλητών NoA, Nmax και IPV, ενώ στο τρίτο βήμα η εκτέλεση του αλγορίθμου.

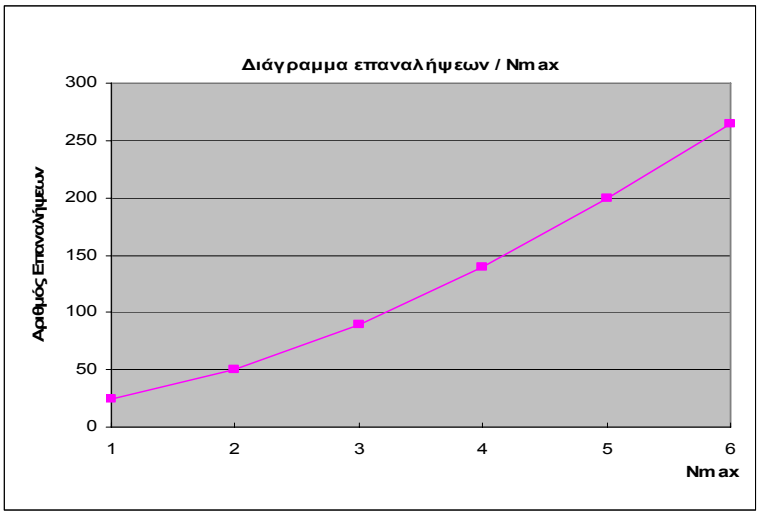
4.1.1 Προ επεξεργασία

Κατά την καταφόρτωση των δειγμάτων ακολουθήθηκαν δυο βήματα προ-επεξεργασίας των εγγράφων - ιστοσελίδων. Το πρώτο στάδιο αποτέλεσε η αποθήκευση σε βάση δεδομένων καθώς και η εξαγωγή του καθαρού περιεχομένου από τα υπερκείμενα, με την αποκοπή των HTML ετικετών. Εκτός του καθαρού κειμένου στη βάση δεδομένων αποθηκεύτηκε και η δομή του ιστού (σύνδεση μεταξύ κόμβων του γράφου). Με αυτόν τον τρόπο ανακτήθηκε η πληροφορία των υπερσυνδέσμων (σημείο εκκίνησης και σημείο προορισμού), οπότε και η δομή του γράφου (κόμβοι και συνδέσεις) στον οποίο εφαρμόζεται ο αλγόριθμος είναι πλέον γνωστός.

Στο δεύτερο στάδιο υπολογίστηκε η τιμή ομοιότητας των κειμενικών περιεχομένων των ιστοσελίδων με τον αρχικό έγγραφο αναφοράς καθώς και το σύνολο των εγγράφων που ικανοποιούν την ερώτηση. Ο βέλτιστος τρόπος για τον ορισμό των σχετικών προς ένα ερώτημα σελίδων, είναι η ανθρώπινη σκέψη και συγκεκριμένα του χρήστη που θέτει το ερώτημα. Ωστόσο η καταμέτρηση και αξιολόγηση ενός τέτοιου αριθμού εγγράφων καθιστά πρακτικώς αδύνατη τη συμμετοχή του ανθρώπινου παράγοντα. Λαμβάνοντας υπόψη τον παραπάνω περιορισμό, για τον προσδιορισμό των σελίδων που ικανοποιούν τον χρήστη, χρησιμοποιήθηκε η τιμή της ομοιότητας που παρουσίασε το κειμενικό περιεχόμενο των σελίδων με την αρχική σελίδα που ουσιαστικά αποτελεί το ερώτημα του χρήστη. Οι σελίδες με τον μεγαλύτερο βαθμό ομοιότητας ορίστηκαν ως σελίδες ενδιαφέροντος.

4.1.2 Ορισμός Μεταβλητών

Η παράμετρος N_{max}

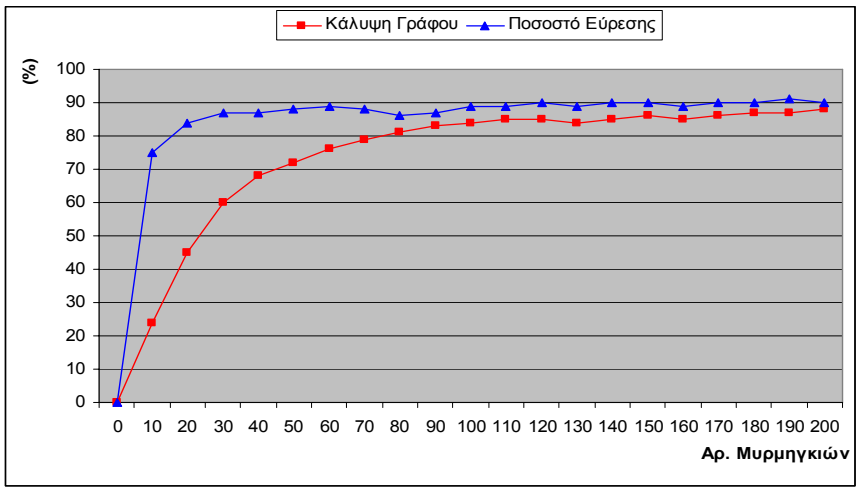


Σχήμα 3. Μέσος Αριθμός επαναλήψεων ανά μήκος διαδρομής

Ο περιορισμός του μεγάλου μεγέθους καταφόρτωσης δεν έδωσε τη δυνατότητα εκτεταμένης μελέτης της επίδρασης την μεταβλητής N_{max}, ωστόσο φαίνεται ότι επηρεάζει κυρίως τον χρόνο εκτέλεσης. Όσο μεγαλύτερος ο αριθμός των κόμβων που επισκέπτεται ένα μυρμήγκι κατά το σχηματισμό μιας διαδρομής τόσο περισσότερος χρόνος απαιτείται για την σύγκλιση του αλγορίθμου. Όπως φαίνεται και στο Σχήμα 3 η αύξηση του αριθμού των επαναλήψεων είναι περίπου ανάλογη με την αύξηση της τιμής N_{max}. Ωστόσο η επιλογή μιας μικρής τιμής για *το* N_{max} ουσιαστικά οδηγεί σε περισσότερες εφαρμογές του αλγορίθμου για την εύρεση των σελίδων ενδιαφέροντος. Ειδικά όταν *το* N_{max}=1 θα απαιτηθεί σχεδόν η πλήρης κάλυψη των σελίδων για την εύρεση των σημείων ενδιαφέροντος. Στις μετρήσεις αξιολόγησης του αλγορίθμου χρησιμοποιήθηκε η τιμή N_{max}=3.

Η παράμετρος *NoA*

Ο αριθμός των μυρμηγκιών δεν είναι ιδιαίτερα σημαντικός στην ποιότητα αναζήτησης. Ειδικά λαμβάνοντας υπόψη το Σχήμα 4, κρατώντας τον αριθμό των μυρμηγκιών σχετικά μικρό (τιμές από 5 με 20) επιτυγχάνεται μια αρκετά ικανοποιητική απόδοση του αλγορίθμου (70-80%), ενώ το κόστος εξερεύνησης που αποτελεί το ποσοστό κάλυψης του γράφου, παραμένει σχετικά χαμηλό (20-60%). Το επιπλέον κέρδος για τον χαμηλό αριθμό μυρμηγκιών αποτελεί και η ελάττωση του χρόνου εκτέλεσης του αλγορίθμου.

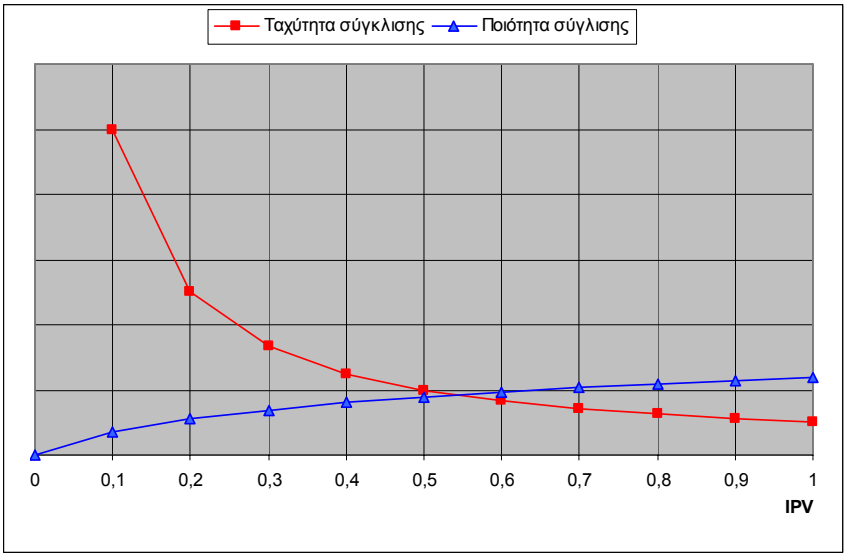


Σχήμα 4. Διάγραμμα κόστους (ποσοστό κάλυψης) και εύρεσης λύσεων ανά Αριθμού μυρμηγκιών

Η παράμετρος *IPV*

Η αρχική τιμή φερομόνης *IPV* που εναποτίθεται στους κόμβους που εισέρχονται στο χώρο αναζήτησης σύμφωνα με το Σχήμα 5, επηρεάζει την ταχύτητα σύγκλισης καθώς και την ποιότητα των αποτελεσμάτων. Πιο συγκεκριμένα, όσο πιο μικρή είναι η τιμή *IPV* τόσο πιο γρήγορα ο αλγόριθμος συγκλίνει σε μια λύση. Ωστόσο η λύση αυτή είναι τυχαία και εξαρτάται από τις αρχικές διαδρομές των μυρμηγκιών. Οπότε η χρήση μικρής τιμής στη μεταβλητή *IPV*, οδηγεί σε κακής ποιότητας λύσεις. Από την άλλη πλευρά μια μεγάλη τιμή δεν βελτιώνει θεαματικά την ποιότητα της αναζήτησης ενώ ταυτόχρονα μειώνει την ταχύτητα σύγκλισης (πολύ μεγάλος χρόνος απόκρισης). Ουσιαστικά η μεταβλητή *IPV* υποδηλώνει τον τρόπο που συμπεριφέρεται το μυρμήγκι στους κόμβους οι οποίοι εισέρχονται στο χώρο αναζήτησης. Δεδομένου ότι η ποιότητα

του κόμβου ορίζεται από την εξίσωση ομοιότητας που δίνεται από τη σχέση 5 και κυμαίνεται στο διάστημα [0,1], η ιδανική τιμή για την παράμετρο *IPV* είναι αυτή που προσδίδει ουδέτερη συμπεριφορά σε κάθε νέο κόμβο. Μια καλή προσέγγιση είναι η μέση τιμή ομοιοτήτων ή ο μέσος όρος ομοιοτήτων οι οποίες κατά τη διάρκεια αναζήτησης δεν είναι γνωστές. Για τα επόμενα πειράματα επιλέχθηκε η τιμή *IPV* =0.4.



Σχήμα 5. Διάγραμμα Ταχύτητας σύγκλισης και ποιότητα σύγκλισης σε διάφορες τιμές της μεταβλητής *IPV*

4.1.3 Παρουσίαση Αποτελεσμάτων Εφαρμογής του Αλγορίθμου

Στους πίνακες 4, 5 και 6 παρουσιάζονται τα πειραματικά αποτελέσματα αναζήτησης του αλγορίθμου. Για όλα τα πειράματα η τιμές των μεταβλητών που επιλέχθηκαν είναι *NoA*=10, N_{max}=3, *NC*=100 και *IPV*=0,4. Για κάθε σύνολο σελίδων υπολογίζεται ο αριθμών των επιστρεφόμενων αποτελεσμάτων – λύσεων, το κόστος εξερεύνησης, ο αριθμός των σχετικών προς το ερώτημα αποτελεσμάτων και ένας δείκτης ποιότητας.

Όπως προαναφέρθηκε στην προηγούμενη ενότητα σε κάθε εφαρμογή του αλγορίθμου προστίθεται και ένα αποτέλεσμα είτε είναι σχετικό είτε άσχετο με το αρχικό σύνολο των προς αναζήτηση σελίδων. Με άλλα λόγια, για κάθε αναζήτηση το σύνολο επιστρεφομένων αποτελεσμάτων είναι ίσο με τον αριθμό των επαναληπτικών εφαρμογών του αλγορίθμου. Αυτό ουσιαστικά μειώνει την ποιότητα των αποτελεσμάτων, μόνο το 40% των αποτελεσμάτων είναι σχετικό προς το αρχικό ερώτημα, αλλά επιτρέπει την περαιτέρω εξερεύνηση του γράφου όταν δε βρεθεί σχετική πληροφορία. Αυτό είναι σημαντικό ιδιαίτερα σε περιπτώσεις που ο αλγόριθμος δεν καταφέρει να συγκλίνει σε σχετική προς το ερώτημα λύση στα αρχικά στάδια της αναζήτησης. Με άλλα λόγια με κόστος την ελάττωση της ποιότητας δίνεται η δυνατότητα στον αλγόριθμο να αναζητήσει περαιτέρω λύσεις.

Δείγμα	Αρ. Σελίδων	NoA	N_{max}	NC	Σχετικές Σελίδες
1	140594	10	3	100	43
Αρ. Εφαρμογών του Αλγορίθμου	Κόστος Αναζήτησης	Ποσοστό Κόστους (%)	Αριθμών Σχετικών Σελίδων	Ποσοστό Ανάκτησης (%)	Ποσοστό ποιότητας (%)
0	0	0,00	0	0,00	0,00
5	4563	3,25	1	2,33	20,00
10	7659	5,45	3	6,98	30,00
15	11243	8,00	4	9,30	26,67
20	12732	9,06	7	16,28	35,00
25	16023	11,40	9	20,93	36,00
30	19483	13,86	10	23,26	33,33
35	21761	15,48	14	32,56	40,00
40	24367	17,33	17	39,53	42,50
45	27337	19,44	18	41,86	40,00
50	30554	21,73	23	53,49	46,00
55	33772	24,02	25	58,14	45,45
60	36990	26,31	28	65,12	46,67
65	37813	26,90	32	74,42	49,23
70	39208	27,89	34	79,07	48,57
75	41716	29,67	36	83,72	48,00
80	43426	30,89	37	86,05	46,25
85	46238	32,89	37	86,05	43,53
90	48644	34,60	37	86,05	41,11
95	51634	36,73	38	88,37	40,00
100	54367	38,67	38	88,37	38,00

Πίνακας 4. Αποτελέσματα εφαρμογής του αλγορίθμου στο πρώτο δείγμα εγγράφων

Δείγμα	Αρ. Σελίδων	NoA	N_{max}	NC	Σχετικές Σελίδες
2	192155	10	3	100	34
Αρ. Εφαρμογών του Αλγορίθμου	Κόστος Εξερεύνησης	Ποσοστό Κόστους (%)	Αριθμών Σχετικών Σελίδων	Ποσοστό Ανάκτησης (%)	Ποσοστό ποιότητας (%)
0	0	0,00	0	0,00	0,00
5	6535	3,40	2	5,88	40,00
10	11436	5,95	3	8,82	30,00
15	15563	8,10	5	14,71	33,33
20	19689	10,25	7	20,59	35,00
25	23311	12,13	8	23,53	32,00
30	28687	14,93	10	29,41	33,33
35	32144	16,73	12	35,29	34,29
40	37268	19,39	13	38,24	32,50
45	41623	21,66	15	44,12	33,33
50	45835	23,85	19	55,88	38,00
55	50130	26,09	20	58,82	36,36
60	56210	29,25	22	64,71	36,67
65	59644	31,04	23	67,65	35,38
70	63192	32,89	24	70,59	34,29
75	67895	35,33	26	76,47	34,67
80	69341	36,09	28	82,35	35,00
85	74868	38,96	29	85,29	34,12
90	78156	40,67	29	85,29	32,22
95	79929	41,60	30	88,24	31,58
100	84367	43,91	30	88,24	30,00

Πίνακας 5. Αποτελέσματα εφαρμογής του αλγορίθμου στο δεύτερο δείγμα εγγράφων

Παρατηρώντας του πίνακες 3, 4 και 5 με κόστος αναζήτησης 30% με 50 % επιτυγχάνεται η ανάκτηση του 85 με 90% των εγγράφων.

Δείγμα	Αρ. Σελίδων	NoA	N_{max}	NC	Σχετικές Σελίδες
3	175997	10	3	100	72
Αρ. Εφαρμογών του Αλγορίθμου	Κόστος Εξερεύνησης	Ποσοστό Κόστους (%)	Αριθμών Σχετικών Σελίδων	Ποσοστό Ανάκτησης (%)	Ποσοστό ποιότητας (%)
0	0	0,00	0	0,00	0,00
5	5134	2,92	3	4,17	60,00
10	9346	5,31	5	6,94	50,00
15	13689	7,78	9	12,50	60,00
20	17235	9,79	11	15,28	55,00
25	20463	11,63	12	16,67	48,00
30	22674	12,88	16	22,22	53,33
35	26107	14,83	17	23,61	48,57
40	28438	16,16	19	26,39	47,50
45	30722	17,46	20	27,78	44,44
50	32283	18,34	24	33,33	48,00
55	36641	20,82	26	36,11	47,27
60	40762	23,16	29	40,28	48,33
65	42577	24,19	30	41,67	46,15
70	44961	25,55	34	47,22	48,57
75	49388	28,06	39	54,17	52,00
80	53676	30,50	43	59,72	53,75
85	57169	32,48	46	63,89	54,12
90	61250	34,80	49	68,06	54,44
95	64934	36,89	52	72,22	54,74
100	66599	37,84	54	75,00	54,00
105	67124	38,14	54	75,00	51,43
110	71604	40,68	55	76,39	50,00
115	75823	43,08	55	76,39	47,83
120	79368	45,10	59	81,94	49,17
125	83492	47,44	61	84,72	48,80
130	86082	48,91	64	88,89	49,23
135	89562	50,89	65	90,28	48,15
140	91534	52,01	66	91,67	47,14
145	93627	53,20	66	91,67	45,52
150	94311	53,59	66	91,67	44,00

Πίνακας 6. Αποτελέσματα εφαρμογής του αλγορίθμου

στο τρίτο σύνολο εγγράφων

4.2 Εφαρμογή με τη χρήση τεχνικών ομαδοποίησης

Η χρήση της συσταδοποίησης στην προτεινόμενη μεθοδολογία έγινε για την βελτίωση της ποιότητας αποτελεσμάτων της αναζήτησης. Ουσιαστικά έχοντας ένα σύνολο εγγράφων η ομαδοποίηση αυτών σε κατηγορίες με υψηλό βαθμό συνοχής αποκόπτει τα αποτελέσματα των οποίων η σχετικότητα με το αρχικό ερώτημα είναι χαμηλή.

Πείραμα	**Σχετικές Σελίδες**	**Αλγόριθμος**		**Συσταδοποίηση**		**Ανάκτηση (%)**	**Ακρίβεια (%)**
		Συνολικός Αριθμός	Βρέθηκαν	Ορισμένες από το Σύστημα ως Σχετικές	Σωστά τοποθετημένες		
1	43	80	37	42	35	81,40	83,33
2	34	85	29	32	26	76,47	81,25
3	72	135	65	70	65	90,28	92,86

Πίνακας 7. Απόδοση του συστήματος με τη χρήση τεχνικών συσταδοποίησης

Στον πίνακα 7 παρουσιάζονται τα αποτελέσματα της εφαρμογής της συσταδοποίησης στα 3 σύνολα αξιολόγησης του αλγορίθμου. Ουσιαστικά με την εφαρμογή μεθόδων συσταδοποίησης στα επιστρεφόμενα αποτελέσματα, ελαττώνεται το ποσοστό ανάκτησης (κατα 2% με 5%) των σχετικών σελίδων αλλά ταυτόχρονα αυξάνεται η ποιότητα αυτών (από 30-50% στο 80-90%). Αυτό είναι λογικό διότι με την συσταδοποίηση οι σελίδες-κόμβοι, που χρησιμοποιήθηκαν μόνο για τη συνέχιση της αναζήτησης, αποκόπτονται λόγω της χαμηλής τιμής ομοιότητας με το αρχικό έγγραφο. Ωστόσο θα πρέπει να σημειωθεί ότι οι ένα μικρό μέρος των σωστών αποτελεσμάτων που όμως δεν κατατάσσονται σωστά κατά την εφαρμογή της συσταδοποίησης, οφείλεται εν μέρει στον διαφορετικό τρόπο υπολογισμού της ομοιότητας. Ενώ κατά την εφαρμογή του αλγορίθμου χρησιμοποιείται η Σχέση 4 που προσδιορίζει την ομοιότητα μεταξύ δυο εγγράφων, κατά τη συσταδοποίηση χρησιμοποιείται το χωροδιανυσματικό μοντέλο [Salton,71]. Η χρήση διαφορετικής συνάρτησης για τον υπολογισμό της ομοιότητας οφείλεται στο γεγονός ότι κατά την αναζήτηση δεν είναι γνωστό το σύνολο της συλλογής των εγγράφων οπότε χρησιμοποιείται η συνάρτηση της Σχέσης 4 που ορίζει την ομοιότητα ανά ζεύγος εγγράφων. Αντίθετα κατά την εφαρμογή της συσταδοποίησης το σύνολο των επιστρεφομένων αποτελεσμάτων ουσιαστικά ορίζει τη συλλογή. Στον πίνακα 8 αποτυπώνονται τα αποτελέσματα εφαρμογής της μεθοδολογίας για 6 τυχαία ερωτήματα στον παγκόσμιο ιστό.

Πείραμα	Σχετικές Σελίδες	Ορισμένες από το Σύστημα ως Σχετικές	Σωστά τοποθετημένες	Ανάκτηση (%)	Ακρίβεια (%)
1	32	37	28	87,50	75,68
2	40	56	32	80,00	57,14
3	21	25	19	90,48	76,00
4	10	8	8	80,00	100,00
5	17	17	15	88,24	88,24
6	36	42	33	91,67	78,57

Πίνακας 8. Απόδοση του συστήματος σε τυχαία ερωτήματα στο διαδίκτυο

Εδώ θα πρέπει να τονιστεί ότι οι όροι "*Ανάκτηση*" και "*Ακρίβεια*" χρησιμοποιούνται στο κεφάλαιο αυτό με την ευρύτερη έννοια ως ένα γενικό πλαίσιο αξιολόγησης και όχι με την έννοια "*recall*" και "*precision*" που χρησιμοποιούνται ευρέως για την αξιολόγηση συστημάτων επεξεργασίας πληροφορίας. Αυτό γίνεται διότι το μέγεθος του συνόλου των εγγράφων καθιστά απαγορευτική την πλήρη αξιολόγησή του με αποτέλεσμα να μην υπάρχει ένα σταθερό μέτρο σύγκρισης.

4.3 Αξιολόγηση

4.3.1 Περιορισμοί

Στα πειραματικά αποτελέσματα αποτυπώνεται η δυνατότητα του προτεινόμενου αλγορίθμου να αναζητεί και να εξάγει πληροφορίες από τον παγκόσμιο ιστό σχετικές με ένα ερώτημα. Ωστόσο θα πρέπει να σημειωθεί ότι κατά τη διάρκεια των πειραμάτων δημιουργήθηκε ένα σύνολο περιορισμών. Ο πιο σημαντικός από αυτούς είναι η έκταση του παγκόσμιου ιστού η οποία δεν επέτρεψε την πιλοτική εφαρμογή του αλγορίθμου σε βαθύτερης κλίμακας αναζητήσεις. Όπως προαναφέρθηκε, η ραγδαία εξάπλωση περιόρισε το δείγμα δοκιμών σε ένα βάθος 5-7 υπερσυνδέσμων από τον κόμβο αφετηρίας της αναζήτησης.

Το μέγεθος των δειγμάτων ήταν της τάξης των 200.000. Η ορθή αξιολόγηση του αλγορίθμου προαπαιτεί τον πλήρη καθορισμό των μελών δείγματος, όσον αφορά την πληροφοριακή σχετικότητά τους με τον κόμβο αναφοράς. Η κατάταξη του συνόλου του δείγματος με βάση την ομοιότητα δίνει μια εκτίμηση της σχετικότητας των εγγράφων και κατά συνέπεια ένα μέτρο ταξινόμησης αλλά παραμένει να είναι μια μηχανική μέθοδος κατάταξης και δεν μπορεί να αντικαταστήσει τον παράγοντα άνθρωπο. Το μέγεθος των δειγμάτων είναι αρκετά μεγάλο για να κατηγοριοποιηθεί από ανθρώπινο χέρι. Για την αξιολόγηση θα μπορούσαν να χρησιμοποιηθούν τεχνικές μηχανικής εκμάθησης όπως είναι τα νευρωνικά δίκτυα [Anagnostopoulos,03a], [Anagnostopoulos,03b], αλλά απαιτείται ένα σύνολο έτοιμων και ταξινομημένων εγγράφων για την εξόρυξη των σχετικών εγγράφων.

Ο δεύτερος περιορισμός που αποτελεί ουσιαστικά αποτέλεσμα του προηγούμενου περιορισμού, είναι η επικάλυψη μεταξύ των αναζητήσεων. Ο αλγόριθμος περιλαμβάνει μηχανισμό αποτροπής επικαλυπτόμενων αναζητήσεων για την αποφυγή δημιουργίας κυκλικών διαδρομών αναζήτησης. Ωστόσο σε ένα περιορισμένο τμήμα του ιστού η απαγόρευση κίνησης προς τα πίσω θα προκαλούσε τερματισμό της αναζήτησης σε ελάχιστα μόλις βήματα (συνολικά 2 με 5 αναζητήσεις, ανά δείγμα). Για αυτό το λόγο ο μόνος περιορισμός που ορίστηκε για τη δημιουργία διαδρομών, ήταν η εμπόδιση προσθήκης κόμβου, που ανήκει στο τρέχον σύνολο των λύσεων του αλγορίθμου.

Ιδιαίτερο ενδιαφέρον παρουσιάζει και η συμπεριφορά του αλγορίθμου στα οριακά σημεία του γράφου αναζήτησης. Ως οριακά σημεία ορίζονται οι κόμβοι που ανήκουν στο τελευταίο επίπεδο καταφόρτωσης. Η συμμετοχή αυτών των κόμβων στο γράφο κρίνεται ελλιπής διότι το σύνολο των υπερσυνδέσμων οδηγούν σε ιστοχώρους έξω από το χώρο αναζήτησης. Όπότε κάθε φορά που ο αλγόριθμος δημιουργούσε διαδρομές πολλές από αυτές τερματιζόντουσαν σε κόμβους των ορίων. Αυτό οδηγούσε τον αλγόριθμο καταστάσεις αδράνειας (stagnant).

4.3.2 Θετικά

Αξιολογώντας τη λειτουργία και τη συμπεριφορά του αλγορίθμου παρατηρούμε ότι παρέχει τη δυνατότητα αναζήτησης σε πραγματικό χρόνο. Αξιοσημείωτη είναι η δυνατότητα να επιλέγει αυτόνομα την κατεύθυνση της αναζήτησης επιτρέποντας ταυτόχρονα και την εξερεύνηση "άγνωστων" περιοχών.

Ένα σημαντικό πλεονέκτημα σε σχέση με άλλες κλασσικές μεθόδους αναζήτησης και κατηγοριοποίησης είναι το γεγονός ότι δεν απαιτεί την κάλυψη του συνόλου της περιοχής αναζήτησης. Στα πειράματα που έγιναν είχε ένα ποσοστό κάλυψης περίπου 40%, λαμβάνοντας υπόψη όλους τους ανωτέρω περιορισμούς. Ωστόσο αυτό έχει και το τίμημά του. Το ποσοστό των ανακτηθέντων εγγράφων περιορίζεται σε ένα ποσοστό κοντά στο 80%.

Η δομή του γράφου στον οποίο λαμβάνει χώρα η αναζήτηση δεν φαίνεται να επηρεάζει ιδιαίτερα την αναζήτηση. Μάλιστα όσο η δομή τείνει σ' αυτή ενός πλήρους συνδεδεμένου γράφου η απόδοση της αναζήτησης αυξάνει.

Σε αντίθεση με τις περισσότερες τεχνικές αναζήτησης, η δυνατότητα χρήσης κειμενικού ερωτήματος επιτρέπει ποιοτικότερες αναζητήσεις. Η αναζήτηση με ερωτήματα της μορφής ενός συνόλου όρων έχει το πλεονέκτημα ότι είναι σύντομη και περιεκτική όσο η διατύπωση της ερώτησης είναι ακριβής. Όταν όμως το ερώτημα είναι ασαφές τότε στα αποτελέσματα των αναζητήσεων επικρατεί σύγχυση. Η χρήση ενός εγγράφου αναφοράς για την αναζήτηση βελτιώνει την ποιότητα των επιστρεφόμενων αποτελεσμάτων. Βέβαια, έγγραφα σχετικά με την πληροφοριακή ανάγκη, αλλά με χαμηλή τιμή ομοιότητας με το έγγραφο αναφοράς, δεν ανακτούνται.

4.3.3 Αρνητικά

Σύμφωνα με τα αποτελέσματα των πειραμάτων το ποσοστό κάλυψης είναι περίπου στο 40% με έναν μέσο όρο ανάκτησης στο 90%. Παρόλο, που το ποσοστό ανάκτησης για το συγκεκριμένο ποσοστό κάλυψης, είναι ικανοποιητικό, καθιστά όμως την εφαρμογή της μεθοδολογίας αναζητήσεις μεγάλης κλίμακας απαγορευτική. Η εφαρμογή του αλγορίθμου στον παγκόσμιο ιστό για το δεύτερο σύνολο πειραμάτων ήταν αρκετά χρονοβόρα. Για την επιστροφή 100 αποτελεσμάτων απαιτείται περισσότερο από 1 ώρα.

Επιπρόσθετα η προσθήκη λύσεων μη σχετικών με το ερώτημα προκαλεί πτώση της ποιότητας των αποτελεσμάτων. Η προσθήκη τεχνικών συσταδοποίησης στα ανακτηθέντα έγγραφα βελτιώνει την ποιότητα των αποτελεσμάτων, αλλά για κάθε επιστροφή άσχετου με το ερώτημα εγγράφου προστίθεται και το ανάλογο κόστος εξερεύνησης στον ιστό. Ως εναλλακτική μέθοδος για την αποκοπή των κακών λύσεων μπορεί να οριστεί μια επιπρόσθετη μεταβλητή στον κανόνα δημιουργίας λύσεων του αλγορίθμου. Αυτή η τιμή θα ορίζει ένα ελάχιστο όριο στο βαθμό ομοιότητας της υποψήφιας λύσης με το έγγραφο αναφοράς.

5 Αναφορές

[Anagnostopoulos,03a] I. Anagnostopoulos, C. Anagnostopoulos, D. Vergados, Vassili Loumos and Eleftherios Kayafas "Classification of a large web page collection applying a GRNN architecture", ISCIS 03, pp. 34-41, LNCS Springer-Verlag.

[Anagnostopoulos,03b] I. Anagnostopoulos, C. Anagnostopoulos, Vassili Loumos, Eleftherios Kayafas, "Taxonomy of E-commerce Web Pages employing a Probabilistic Neural Network Classifier", submitted to IEE Proceedings - Software (under review process).

[Broder, 97] A. Broder, S. Glassman, M. Manasse, Zweig G., "Syntactic clustering of the Web". Proceedings of the 6th International World Wide Web Conference, April 1997; 391–404.

[Dorigo,04] M. Dorigo and T. St¨utzle. Ant Colony Optimization. The MIT Press, 2004.

[Dorigo,96] Dorigo, M., Maniezzo, V., and Colorni, A. The ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics-Part B, 26, 1, 29–41, 1996.

[Fetterly,04] Dennis Fetterly, et al. "A large-scale study of the evolution of Web pages" SOFTWARE—PRACTICE AND EXPERIENCE 2004; 34:213–237.

[Hammouda,02] K.M. Hammouda, M. S. Kamel,"Phrase-based Document Similarity Based on an Index Graph Model|", Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan. IEEE Computer Society 2002, ISBN 0-7695-1754-4:pp. 203-210.

[Hammouda,03] K.M. Hammouda, M. S. Kamel, "Incremental Document Clustering Using Cluster Similarity Histograms", WIC International Conference on Web Intelligence, (WI 2003), 13-17 October 2003, Halifax, Canada. IEEE Computer Society 2003, ISBN 0-7695-1932-6, pp. 597-601.

[Isaacs,99] J. D. Isaacs and J. A. Aslam. "Investigating measures for pairwise document similarity. Technical Report PCS-TR99- 357, Dartmouth College, Computer Science, Hanover, NH, June 1999.

[Kouzas,06a] G. Kouzas, E. Kayafas, V. Loumos: "Ant Seeker: An algorithm for enhanced web search", 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006, June 7-9, 2006, Athens, Greece. IFIP 204 Springer 2006, ISBN 0-387-34223-0 pp 649-656.

[Kouzas,06b] Kouzas G., Anagnostopoulos I., Maglogiannis I. and Anagnostopoulos C. "Bridging the syntactic and the semantic web search"-ICANN 2006, 16th International Conference, Athens, Greece, September 10-14, 2006. Proceedings, Part II. Lecture Notes in Computer Science 4132 Springer 2006, ISBN 3-540-38871-0, pp. 104-112.

[Kouzas,06c] Kouzas G., E. Kayafas, V. Loumos "Web Similarity Measurements using Ant – Based Search Algorithm" XVIII IMEKO WORLD CONGRESS Metrology for a Sustainable Development September, 17 – 22, 2006, Rio de Janeiro, Brazil.

[Parpinelli,02] R.S. Parpinelli, H.S. Lopes, and A.A. Freitas. Data mining with an ant colony optimization algorithm. IEEE Transactions onEvolutionary Computing 6(4), 2002, pp. 321–332.

[Salton,68] G. Salton, M. E. Lesk. Computer evaluation of indexing and text processing, Journal of the ACM, 15(1): 8-36, January 1968.

[Salton,71] G. Salton. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice Hall Inc., 1971.

[Salton,88] G. Salton, C. Buckley. Term-weighting approaches in automatic retrieval, Information Processing & Management, 24(5): 513-523, 1988

ΚΕΦΑΛΑΙΟ

5

ΣΥΜΠΕΡΑΣΜΑΤΑ - ΣΥΖΗΤΗΣΗ

ΠΕΡΙΕΧΟΜΕΝΑ 5ου ΚΕΦΑΛΑΙΟΥ
ΣΥΜΠΕΡΑΣΜΑΤΑ - ΣΥΖΗΤΗΣΗ

1 ΑΝΑΚΕΦΑΛΑΙΩΣΗ της ΔΙΑΤΡΙΒΗΣ

Στην παρούσα διατριβή προτείνεται μια μεθοδολογία οδήγησης και εντοπισμού πληροφορίας στον Παγκόσμιο Ιστό. Αρχικά πραγματοποιείται μια αναφορά στην εξέλιξη του Παγκοσμίου Ιστού και στα προβλήματα που κυριαρχούν σ' αυτόν. Τα δυο σημαντικότερα προβλήματα είναι η τεράστια εξάπλωση και η άναρχη δομή του. Αποτέλεσμα αυτών είναι η αδυναμία αναζήτησης πληροφορίας καθώς και η αδυναμία εκτίμησης του πληθυσμού.

Στη συνέχεια γίνεται μια αναφορά στης τεχνικές αναζήτησης καθώς και οι αδυναμίες τους. Αναλύονται διεξοδικά τεχνικές επεξεργασίας πληροφορίας και μοντέλα ταξινόμησης εγγράφων, ενώ παρουσιάζονται και αλγόριθμοι βελτιστοποίησης με ιδιαίτερη έμφαση στον αλγόριθμο αποικίας μυρμηγκιών.

Κατόπιν εισάγεται η μεθοδολογία αναζήτησης της οποίας τα βασικά στάδια είναι:

- Ο αλγόριθμος αναζήτησης
- Ο έλεγχος ομοιότητας
- Η ομαδοποίηση των αποτελεσμάτων.

Ο αλγόριθμος αναζήτησης στηρίζεται στον αλγόριθμο αποικίας μυρμηγκιών και προσομοιώνει την αναζήτηση πληροφορίας με τη διαδικασία εύρεσης τροφής των φυσικών μυρμηγκιών. Ξεκινώντας από έναν αρχικό κόμβο – ιστοσελίδα, αναζητεί διαδρομές που οδηγούν σε κόμβους με σχετικό, προς τον κόμβο αναφοράς, περιεχόμενο.

Το στάδιο ελέγχου ορίζει το ποσοστό κειμενικής ομοιότητας του εκάστοτε κόμβου με τον κόμβο αναφοράς. Το μέτρο της ομοιότητας αποτελεί το κριτήριο επιλογής του κάθε κόμβου ως αποδεκτή λύση της αναζήτησης. Η επιλογή βασίζεται στις κλασικές τεχνικές ομοιότητας εγγράφων λαμβάνοντας υπόψη και τη δομή του κειμένου του κόμβου.

Το στάδιο ομαδοποίησης περιλαμβάνει την αξιολόγηση και κατηγοριοποίηση των αποτελεσμάτων. Για την ομαδοποίηση επιλέχθηκαν τεχνικές συταδοποίησης εγγράφων και συγκεκριμένα ο αλγόριθμος επαυξητικής συσταδοποίησης με τη χρήση ιστογράμματος ομοιοτήτων.

1.1 Περαιτέρω Έρευνα

Η παρούσα διατριβή διαπραγματεύτηκε μια νέα μεθοδολογία αναζήτησης στον Παγκόσμιο Ιστό με τη χρήση αλγορίθμων αποικίας μυρμηγκιών. Ο προτεινόμενος αλγόριθμος επιτυγχάνει ικανοποιητικά ποσοστά αναζήτησης ενώ παράλληλα διατηρεί σχετικά μικρό το κόστος εξερεύνησης.

Ωστόσο περαιτέρω έρευνα θα μπορούσε να βελτιώσει σημαντικά την απόδοση του αλγορίθμου. Για την εφαρμογή του αλγορίθμου σε μεγαλύτερης κλίμακας αναζητήσεις απαιτείται η ελαχιστοποίηση του κόστους εξερεύνησης. Μια πιθανή τροποποίηση του αλγορίθμου είναι η αναζήτηση περισσοτέρων λύσεων ανά εφαρμογή. Αυτό θα ελάττωνε κατά πολύ το χρόνο εκτέλεσης καθώς και το κόστος εξερεύνησης του Ιστού. Βέβαια αυτό απαιτεί και την προσθήκη μιας επιπλέον μεταβλητής στην αρχικοποίηση του αλγορίθμου, η οποία θα αποτελεί ένα ελάχιστο όριο στην τιμή της ομοιότητας για την ορισμό των κόμβων ως αποδεκτές λύσεις. Μια ακόμη πιο εκτεταμένη προσέγγιση είναι η παράλληλη αναζήτηση.

Η δομή του αλγορίθμου είναι σχεδιασμένη ώστε να δρα ανεξάρτητη από την συνάρτηση ομοιότητας οπότε είναι δυνατή η εφαρμογή του και με άλλες τεχνικές επεξεργασίας κειμένων. Στην παρούσα προσέγγιση χρησιμοποιήθηκε μια μεθοδολογία υπολογισμού ομοιότητας που βασίζεται στη δομή του εγγράφου (η δομή των ετικετών HTML, και η συντακτική δομή του καθαρού κειμένου) αλλά και στη συσχέτιση όσον αφορά τη συχνότητα εμφάνισης όρων. Γενικά το πλαίσιο που ακολουθείται είναι η επιλογή ενός δείκτη συσχέτισης εγγράφων ανά ζεύγη, διότι κατά τη διάρκεια της αναζήτησης δεν υπάρχει συγκεντρωμένο

σύνολο εγγράφων ώστε να εφαρμοστεί το κλασσικό χωροδιανυσματικό μοντέλο (η ανάστροφη συχνότητα χάνει το νόημα της ύπαρξής της).

1.2 Προτεινόμενη Μέθοδος Εκτίμησης του Μεγέθους του Παγκόσμιου Ιστού

Μια πολύ καλή προσέγγιση στον υπολογισμό του μεγέθους του διαδικτύου παρέχεται από τη μελέτη της φύσης. Η άγρια φύση προσελκύει το ενδιαφέρον του σύγχρονου κόσμου. Ιδιαίτερα οι μετρήσεις του πληθυσμού των άγριων ζώων ήταν πάντοτε μια πρόκληση για τον άνθρωπο. Παλιότερα η γνώση των πληθυσμών ήταν πιο πολύ θέμα επιβίωσης ενώ τώρα λαμβάνει περισσότερο εγκυκλοπαιδικό χαρακτήρα. Οι βιολόγοι με τη βοήθεια των στατιστικών κατάφεραν να αναπτύξουν μεθοδολογίες για την καταμέτρηση των πληθυσμών διαφόρων αγρίων ζώων. Μια από τις πιο γνωστές είναι η μεθοδολογία *Capture – Recapture* [Kendall,95]. Διάφορες προσεγγίσεις της μεθοδολογίας χρησιμοποιήθηκαν και σε άλλα προβλήματα όπως η μέτρηση δημογραφικών παραμέτρων στην ανθρώπινη κοινωνία. Με την αλματώδη ανάπτυξη του διαδικτύου, δημιουργήθηκε έντονο ενδιαφέρον για εφαρμογή της μεθοδολογίας στην εκτίμηση οντοτήτων στον παγκόσμιο ιστό[Anagnostopoulos,06], [Kouzas,03]. Με την έννοια οντότητα στον παγκόσμιο ιστό μπορεί να οριστεί είτε μια απλή ιστοσελίδα είτε μια σελίδα το περιεχόμενο της οποίας ανήκει σε μια συγκεκριμένη κατηγορία, είτε οποιοδήποτε άλλο μετρήσιμο μέγεθος.

Σύμφωνα με την προσέγγιση οι οντότητες του διαδικτύου προσομοιώνονται ως άγρια ζώα ελεύθερα στη φύση, που γεννιούνται, ζουν και πεθαίνουν. Κάθε φορά που μια οντότητα εισέρχεται στο διαδίκτυο, όπως για παράδειγμα η δημιουργία μιας ιστοσελίδας, λαμβάνει χώρα μια γέννηση. Η βασική ιδέα της μεθοδολογίας είναι εκτίμηση του συνολικού πληθυσμού από την μελέτη ενός τυχαίου δείγματος αυτού. Η συνήθης εφαρμογή της μεθοδολογίας γινόταν με τη χρήση παγίδων στην ευρύτερη περιοχή που κατοικεί ο προς παρακολούθηση πληθυσμός. Κάθε φορά που ένας αριθμός άγριων ζώων εγκλωβίζεται στις παγίδες, γίνεται η καταμέτρηση των χαρακτηριστικών του και απελευθερώνεται. Η διαδικασία εγκλωβισμού μέρους του πληθυσμού επαναλαμβάνεται σε τακτά χρονικά διαστήματα. Με τη στατιστική μελέτη των δειγμάτων, όπως για παράδειγμα πόσες φορές ένα μέλος εγκλωβίστηκε στις παγίδες κατά το παρελθόν, γίνεται μια εκτίμηση του πληθυσμού καθώς επίσης και του ρυθμού γεννήσεων και θανάτου.

Μεθοδολογία

Παρακάτω αναλύεται η μεθοδολογία capture – recapture σύμφωνα με το μοντέλο του [Kendall,95] για την εκτίμηση πληθυσμού ιστοσελίδων. Ανά τακτά χρονικά διαστήματα λαμβάνει χώρα μια περίοδος κύριας δειγματοληψίας η οποία χωρίζεται σε δύο δευτερεύουσες δειγματοληψίες χρονικά πολύ κοντινές μεταξύ τους.. Το χρονικό διάστημα μεταξύ των δυο δευτερευούσων δειγματοληψιών πρέπει να είναι όσο το δυνατόν μικρότερο, ώστε ο πληθυσμός να θεωρείται σταθερός. Σε κάθε μια από τις δευτερεύουσες δειγματοληψίες, παγιδεύεται τυχαία ένα σύνολο ιστοσελίδων το οποίο μαρκάρεται. Όταν ολοκληρωθούν οι δυο δευτερεύουσες δειγματοληψίες ολοκληρώνεται και η κύρια. Η επόμενη κύρια δειγματοληψία λαμβάνει χώρα μετά από ένα χρονικό διάστημα στο οποίο θεωρείται ότι ο πληθυσμός έχει αλλάξει. Η στατιστική επεξεργασία των μετρήσεων που εξάγονται σε κάθε κύρια δειγματοληψία (στις δύο δευτερεύουσες) προσδίδει το μέγεθος του πληθυσμού, ενώ η επεξεργασία των μετρήσεων μεταξύ δύο κύριων δειγματοληψιών εξάγει το ρυθμό γεννήσεων και θανάτου.

Έστω ότι σε κάθε κύρια περίοδο δειγματοληψίας i υπάρχουν N_i σελίδες και ότι επιλέχτηκαν κάποιες από αυτές τις σελίδες U_i. Οι υπόλοιπες πού δεν επιλέχτηκαν δίνονται από τη Σχέση 1. Φυσικά, μεταξύ δύο κύριων περιόδων δειγματοληψίας, κάποιες σελίδες θα γίνουν ενεργές. Εάν η πιθανότητα να γίνει ενεργή μια νέα σελίδα είναι φ_k, τότε: ο πληθυσμός των σελίδων σε συνάρτηση με την στιγμή της προηγούμενης δειγματοληψίας δίνεται από τη σχέση 2.

$$M_i = N_i - U_i \qquad \textbf{Σχέση 1}$$

$$M_{i+1} = \phi_i * N_i + \mathrm{B}_i \qquad \textbf{Σχέση 2}$$

Σε κάθε κύρια περίοδο δειγματοληψίας, επιλέγεται ένας αριθμός σελίδων. Κάποιες από αυτές επιλέγονται για πρώτη φορά (u_i) ενώ οι υπόλοιπες επιλέχτηκαν και στην κύρια περίοδο h (m_{hi}). Ο συνολικός αριθμός ιστοσελίδων, που επιλέχτηκαν για τουλάχιστον μία φορά μέχρι την κύρια περίοδο k, δίνονται από τη σχέσεις 3 και XX4XX:

$$n_i = u_i + m_i \qquad \textbf{Σχέση 3}$$

$$\text{όπου } m_i = \sum_{h=1}^{i-1} m_{hi} \qquad \textbf{Σχέση 4}$$

Όπως αναφέρθηκε παραπάνω, υπάρχουν δύο δευτερεύουσες περίοδοι δειγματοληψίας για κάθε κύρια περίοδο. Κάθε σελίδα μπορεί να επιλεχτεί ή σε μία δευτερεύουσα περίοδο ή και στις δύο. Έστω ω, η μεταβλητή που ορίζει τη δευτερεύουσα δειγματοληψία. Για την στατιστική μελέτη των δειγμάτων ορίζονται οι παρακάτω μεταβλητές:

χ_{0i}^{ω} είναι ο αριθμός των ιστοσελίδων μεταξύ των u_i, οι οποίες έχουν ιστορία επιλογής ω στην κύρια περίοδο i.

χ_{hi}^{ω} είναι ο αριθμός των ιστοσελίδων μεταξύ των m_{hi} οι οποίες έχουν ιστορία επιλογής ω στην κύρια περίοδο i.

$\chi_{.i}^{\omega}$ είναι το άθροισμα των χ_{0i}^{ω} και χ_{hi}^{ω} για κάθε $h < i$.

Σε κάθε δευτερεύουσα δειγματοληψία της κύριας περιόδου i, οποιαδήποτε ιστοσελίδα, έχει πιθανότητα p_i να επιλεχθεί. Προφανώς, η πιθανότητα που έχει οποιαδήποτε ιστοσελίδα, να επιλεχθεί, τουλάχιστον μία φορά κατά τη διάρκεια της κύρια περιόδου i, δίνεται από τη σχέση χχχχ. Τέλος, ως χ_i ορίζεται ως η πιθανότητα μιας ιστοσελίδας να επιλεχθεί στην κύρια περίοδο i και να μην ξαναεπιλεχθεί ποτέ.

$$p_i^* = 1 - (1 - p_i)^2 \qquad \textbf{Σχέση 5}$$

Η στατιστική απεικόνιση περιλαμβάνει δυο σύνολα δεικτών με βάση τα οποία εξάγονται τα συμπεράσματα εκτίμησης του πληθυσμού. Η πρώτη κατηγορία (κατηγορία κύριων δεικτών) περιλαμβάνει τους δείκτες, οι οποίοι είναι σχετικοί με την τρέχουσα κύρια περίοδο δειγματοληψίας. Η δεύτερη κατηγορία (κατηγορία σχετικών δεικτών) περιλαμβάνει τους δείκτες, οι οποίοι αναφέρονται στη σχέση μεταξύ της τρέχουσας κύριας περιόδου δειγματοληψίας και όλων των προηγούμενων.

<u>Κατηγορία κύριων δεικτών</u>

NT: Ο αριθμός των σελίδων, που εμφανίστηκε στο δείγμα της τελευταίας περιόδου δειγματοληψίας. Αυτές οι σελίδες δεν πρέπει να έχουν εμφανιστεί σε καμία από τις προηγούμενες αρχικές περιόδους δειγματοληψίας.

NB: Ο αριθμός των νέων σελίδων, που εμφανίστηκαν στο δείγμα και της πρώτης και της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας.

NF: Ο αριθμός των νέων σελίδων που εμφανίστηκαν στο δείγμα της πρώτης δευτερεύουσας περιόδου δειγματοληψίας αλλά δεν εμφανίστηκαν στο δείγμα της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας.

NS: Ο αριθμός των νέων σελίδων που εμφανίστηκαν στο δείγμα της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας αλλά δεν εμφανίστηκαν στο δείγμα της πρώτης δευτερεύουσας περιόδου δειγματοληψίας.

PT: Ο αριθμός των σελίδων που εμφανίστηκαν στην τελευταία κύρια περίοδο δειγματοληψίας και επίσης εμφανίστηκαν τουλάχιστον σε μια προηγούμενη κύρια περίοδο δειγματοληψίας. Πρόκειται δηλαδή για παλιές σελίδες, οι οποίες εμφανίστηκαν στην τελευταία κύρια περίοδο δειγματοληψίας.

TT: Ο συνολικός αριθμός των σελίδων που εμφανίστηκαν στο δείγμα της τελευταίας κύριας περιόδου δειγματοληψίας.

TB: Ο συνολικός αριθμός των σελίδων που εμφανίστηκαν στο δείγμα και της πρώτης και της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας της τελευταίας κύριας περιόδου δειγματοληψίας.

TF: Ο συνολικός αριθμός των σελίδων που εμφανίστηκαν στο δείγμα της πρώτης δευτερεύουσας περιόδου δειγματοληψίας αλλά δεν εμφανίστηκαν στο δείγμα της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας.

TS: Ο συνολικός αριθμός των σελίδων που εμφανίστηκαν στο δείγμα της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας αλλά δεν εμφανίστηκαν στο δείγμα της πρώτης δευτερεύουσας περιόδου δειγματοληψίας.

Κατηγορία σχετικών δεικτών

Σε κάθε κύρια περίοδο δειγματοληψίας δημιουργείται ένα σύνολο σειρών με αυτά τα στατιστικά αποτελέσματα. Εάν η τρέχουσα κύρια περίοδος δειγματοληψίας είναι «n» (όπου n είναι φυσικός αριθμός), δημιουργούνται n-1 σειρές με αυτά τα στατιστικά αποτελέσματα. Κάθε σειρά περιέχει τα ακόλουθα πεδία:

- H_T: Ο αριθμός των σελίδων που εμφανίστηκαν στο δείγμα της τελευταίας κύριας περιόδου δειγματοληψίας και τελευταία εμφανίστηκαν στην κύρια περίοδο δειγματοληψίας *h*.
- H_B: Ο αριθμός HT των σελίδων που εμφανίστηκαν στο δείγμα και της πρώτης και της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας της τελευταίας κύριας περιόδου δειγματοληψίας.
- H_F: Ο αριθμός HT των σελίδων που εμφανίστηκαν στο δείγμα της πρώτης δευτερεύουσας περιόδου δειγματοληψίας αλλά δεν εμφανίστηκαν στο δείγμα της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας της τελευταίας κύριας περιόδου δειγματοληψίας.

H_S: Ο αριθμός HT των σελίδων που εμφανίστηκαν στο δείγμα της δεύτερης δευτερεύουσας περιόδου δειγματοληψίας αλλά δεν εμφανίστηκαν στο δείγμα της πρώτης δευτερεύουσας περιόδου δειγματοληψίας της τελευταίας κύριας περιόδου δειγματοληψίας.

2 ΤΟ ΜΕΛΛΟΝ ΤΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ

Ξεκινώντας από μια ιδέα διασύνδεσης υπολογιστών το 1969 για στρατιωτικούς και ερευνητικούς σκοπούς ουσιαστικά δημιουργήθηκε ο σκελετός και η βάση του διαδικτύου. Στην πορεία η προσθήκη και άλλων ιδεών με σημαντικότερη αυτή του Tim Berners-Lee το 1989 από το ερευνητικό ινστιτούτο CERN της Ελβετίας, δημιούργησε τον παγκόσμιο ιστό. Έκτοτε ο Παγκόσμιος Ιστός παρουσιάζει ραγδαία ανάπτυξη μετρώντας 10.000 εξυπηρετητές και 10 εκατομμύρια χρήστες το 1994, 5 μόλις χρόνια μετά τη δημιουργία του. Η εκθετική εξάπλωση αποτυπώνεται στις μέρες μας όπου το αριθμός των συνδεδεμένων υπολογιστών ξεπερνά τα 300.000.000! Η μεγάλη επιτυχία του Ιστού έγκειται στην απλότητα των κανόνων που το διέπουν που συνοψίζονται στον εξής έναν "Δεν υπάρχει κανένας περιορισμός". Αυτό γίνεται εύκολα αντιληπτό αν παρατηρήσουμε τη δομή του. Επικρατεί πλήρης αναρχία.

Στις μέρες μας το Διαδίκτυο καλύπτει πλέον ένα μεγάλο μέρος της ανάγκης μας για πληροφορία (ενημέρωση, έρευνα, ψυχαγωγία, επικοινωνία). Ωστόσο η άναρχη δομή του, και ο καταιγισμός πληροφορίας πολλές φορές προκαλούν το αντίθετο αποτέλεσμα. Οι χρήστες δεν μπορούν αρχικά να αναζητήσουν και στη συνέχεια να επεξεργαστούν τον μεγάλο όγκο πληροφορίας με τον οποίο έρχονται "αντιμέτωποι". Για να καλυφθούν οι ανάγκες των χρηστών, αναπτύχθηκαν διάφορες τεχνικές και εργαλεία αναζήτησης και επεξεργασίας πληροφορίας. Μια άλλη προσέγγιση για τη επίλυση των προβλημάτων είναι η προσπάθεια εφαρμογής κανόνων και δομής στον Παγκόσμιο Ιστό. Κάποιες από αυτές κρίνονται ικανοποιητικές ενώ άλλες όχι.

Τα βασικά πλέον ερωτήματα που κυριαρχούν στον Παγκόσμιο Ιστό είναι το μέγεθός του καθώς και η διαχείριση της πληροφορίας που περιέχει. Από τη μόνη μελέτη για το ποσοστό κάλυψης του παγκόσμιου ιστού το 1999 [Lawrence,98a], [Lawrence,98b] το διαδίκτυο άλλαξε δραματικά. Νέες τεχνολογίες προστέθηκαν άλλες αφαιρέθηκαν ενώ η ανάπτυξή του παραμένει ραγδαία. Τελικά μήπως είναι το μόνο ανθρώπινο κατασκεύασμα του ανθρώπου που δεν έχει ούτε αρχή ούτε τέλος;

Τα τελευταία χρόνια γίνεται μια προσπάθεια δημιουργίας δομών στον Παγκόσμιο Ιστό με στόχο την καλύτερη οργάνωση, διάδοση και επεξεργασία της πληροφορίας. Επιχειρείται δηλαδή μια πιο ορθολογιστική προσέγγιση στην διαχείριση της πληροφορίας. Μια από τις προσεγγίσεις προτείνει τη μετατροπή της πληροφορίας από ημιδομημένη σε δομημένη μορφή με στόχο την αυτοματοποιημένη ανταλλαγή πληροφορίας. Η επεκτάσιμη γλώσσα σήμανσης γνωστή και ως XML (eXtensible Markup Language). Η χρήση της είναι πλέον δεδομένη σε όλο το εύρος των εφαρμογών του διαδικτύου και όχι μόνο.

Μια πιο ολοκληρωμένη πρόταση στην προσπάθεια αναδόμησης του Ιστού προέρχεται από τον ίδιο του τον δημιουργό, τον Berners-Lee, και ονομάζεται Σημασιολογικός Ιστός. Ο Σημασιολογικός Ιστός, αποτελεί ένα σχετικά καινούριο όραμα για την κοινωνία της πληροφορίας του μέλλοντος. Ο ίδιος υποστήριξε πως η κατανόηση του περιεχομένου που είναι διαθέσιμο στον Παγκόσμιο Ιστό από τους ηλεκτρονικούς υπολογιστές μπορεί να προσφέρει μία επανάσταση από νέες δυνατότητες και πληροφορικές εφαρμογές (Berners-Lee 2001).

Ο Σημασιολογικός Ιστός αποτελεί ουσιαστικά μια επέκταση του παρόντος Ιστού. Η διαφορά έγκειται στο γεγονός ότι αποδίδεται πολύ καλά το νόημα της διαθέσιμης πληροφορίας σε μορφή κατανοητή τόσο από ανθρώπους, όσο και από ηλεκτρονικούς υπολογιστές, διευκολύνοντας τη μεταξύ τους συνεργασία. Η σημασιολογική περιγραφή των δεδομένων και του περιεχομένου πραγματοποιείται σε σταθερά λεξιλόγια, τα οποία είναι διαθέσιμα στον Ιστό και λειτουργούν ως σημεία αναφοράς. Για την περιγραφή των δεδομένων και τη σύνδεσή τους με σταθερά λεξιλόγια απαιτείται η χρήση ενός απλού μοντέλου μετα-δεδομένων και μιας σύνταξης για αυτά. Το επικρατέστερο μοντέλο είναι το Πλαίσιο Περιγραφής Πόρων (Resource Description Framework ή RDF), το οποίο χρησιμοποιώντας τριαδικούς συνδυασμούς πόρων Ιστού καταφέρνει να προβάλει μία σαφή μέθοδο έκφρασης σημασιολογίας, αναγνώσιμη από μηχανικά πληροφοριακά συστήματα.

Το επόμενο βήμα για την οριστική εφαρμογή του Σημασιολογικού Ιστού είναι η δημιουργία μηχανισμών και συστημάτων που θα αξιοποιήσουν της δυνατότητες που προσφέρει το μέχρι τώρα θεωρητικό υπόβαθρο ώστε να ικανοποιήσουν της υπάρχουσες ανάγκες. Ακόμη μεγαλύτερη πρόκληση είναι η δημιουργία στιβαρών δομών οι οποίες θα είναι σε θέση να βάλουν σε τάξη την άναρχη δομή του υπάρχοντος Παγκόσμιου Ιστού.

Ωστόσο η μετατροπή ή η αντικατάσταση του υπάρχοντος Παγκόσμιου Ιστού σε Σημασιολογικό Ιστό ακόμη φαντάζει δύσκολη. Αν και τις περισσότερες φορές για την καθιέρωση ενός προτύπου δεν αποτελεί κριτήριο η απόφαση μιας κοινότητας ειδικών αλλά ο τελικός χρήστης. Η εξέλιξη των επόμενων ετών θα δείξει αν ο Σημασιολογικός Ιστός κυριαρχήσει ή πεθάνει ή ακόμη περισσότερο αν οδηγηθούμε σε έναν Παγκόσμιο Ιστό δυο ταχυτήτων...

3 Βιβλιογραφικές αναφορές

[Kendall,95] Kendall, W. L., Pollock, K. H. and Brownie, C. (1995). A likelihood-based approach to capture-recapture estimation of demographic parameters under the robust design. Biometrics, 51, 293-308.

[ERMIS,02] Statistical Methodology for measurement, WP1, Deliverable D1.3, ERMIS Consortium, http://www.ermisproject.gr, May 2002.

[Anagnostopoulos,06] "Estimating the evolution of categorized web page populations", I. Anagnostopoulos, P. Stavropoulos, G. Kouzas, C. Anagnostopoulos, D.D.Vergados, proceeding of the 1st International Workshop on Adaptation and Evolution in Web Systems Engineering (AEWSE 06), 6th International Conference of Web Engineering (ICWE 06), Palo Alto, CA, July 10-14, 2006, ISBN: 1-59593-435-9.

[Kouzas,03] "Measuring the population of web pages in the wild web", Kouzas G.S., Stavropoulos P., Anagnostopoulos I., Anagnostopoulos C., Loumos V. and Kayafas E., XVII IMEKO World Congress, pp. 720-725, June 22-27, 2003, Dubrovnik, CROATIA

[Lawrence,98a] Steve Lawrence and C. Lee Giles. Searching the World Wide Web. (Vol. 280, pp. 98-100) Science (1998).

[Lawrence,98b] Steve Lawrence and C. Lee Giles. Context and page analysis for improved web search. IEEE Internet Computing, 2(4):38–46, 1998.

ΚΕΦΑΛΑΙΟ

6

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

ΠΕΡΙΕΧΟΜΕΝΑ 6ου ΚΕΦΑΛΑΙΟΥ

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

Βιογραφικό Σημείωμα

Ο Γεώργιος Σ. Κούζας γεννήθηκε στην Ελβετία, το 1976. Το 1999 έγινε κάτοχος του διπλώματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, του Εθνικού Μετσοβίου Πολυτεχνείου. Τον Ιανουάριο 2001 έγινε δεκτός στον Τομέα «Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής» του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Ε.Μ.Π, για μεταπτυχιακές σπουδές, με σκοπό την απόκτηση Διδακτορικού Διπλώματος.

Από το 2001 έως σήμερα έχει εργαστεί στα ερευνητικά προγράμματα ERMIS (Electronic commeRce Measurements through Intelligent agentS – IST 1999 – 21051) (2001-2003) και CHIME (2003-…). Το ερευνητικό του ενδιαφέρον εστιάζεται στην εξόρυξη και διαχείριση πληροφοριών του διαδικτύου. Σε αυτό το πλαίσιο, τα ενδιαφέροντα του εστιάζονται κυρίως στην εφαρμογή τεχνικών ανάκτησης πληροφορίας, στον τομέα της διαχείρισης της διαδικτυακής πληροφορίας, καθώς και στη χρήση αλγορίθμων βελτιστοποίησης για έξυπνα συστήματα δυναμικής αναζήτησης.

Από το 2000 ως το 2001 εργάστηκε στην ΚΑΠΑ-ΤΕΛ ως υπεύθυνος ανάπτυξης εφαρμογών κινητής τηλεφωνίας, με τη χρήση του πρωτοκόλλου WAP. Από το 2001 είναι Επιστημονικός Συνεργάτης και βοηθός των Καθ. Ε. Καγιάφα, Καθ. Β. Λούμου του Εργαστηρίου Τεχνολογίας Πολυμέσων του Ε.Μ.Π. Έχει συμμετάσχει στην οργάνωση και διδασκαλία των εργαστηριακών μαθημάτων και ασκήσεων “Ηλεκτρονική Ι” και “εργαστήριο Αναλογικών Κυκλωμάτων”, μαθήματα που περιλαμβάνουν την διδασκαλία δομικών στοιχείων σύγχρονων υπολογιστικών λειτουργικών συστημάτων και συστημάτων επικοινωνιών, καθώς και του μαθήματος “Γραφικά με Υπολογιστές” που αφορά τη διδασκαλία των σύγχρονων γλωσσών προγραμματισμού (Java), τα οποία προσφέρονται από το Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ.

Δημοσιεύσεις

Περιοδικά

Π1 "Urban solid waste collection and routing: the ant colony strategic approach", N. V. Karadimas, **G. Kouzas**, I. Anagnostopoulos, V. Loumos, International Journal of Simulation Systems, Science & Technology, Special Issue on: Modelling & Simulation in Industry, Enterprises & Organisations, Vol. 6, No. 12-13, pp. 45-53, November 2005.

Π2 "A Generalised Regression algorithm for web page categorisation", I. Anagnostopoulos, C. Anagnostopoulos, **G. Kouzas** and D. Vergados, Neural Computing & Applications journal, Springer-Verlag, Vol. 13, no. 3, pp. 229 – 236, 2004.

Π3 "Classification of e-commerce web pages using statistical descriptor vectors", Anagnostopoulos I., **Kouzas G**., Anagnostopoulos C., Kotsilieris T., Kalogeropoulos S., Loumos V. and Kayafas E, Rivista di Statistica Applicata, special issue on new methodologies and applications in the E-Commerce field, vol. 1, pp. 91-107, 2003.

Π4 "An Intelligent Agent Based QoS Provisioning and Network Management System" A. Michalas, M. Louta, **G. Kouzas**, WSEAS Transactions on Computer, Issue 11, Volume 5, November 2006, ISSN 1109-2750, pp. 2710-2717, 2006

Κεφάλαια σε Βιβλία

Β1 "Bridging the Syntactic and the Semantic Web Search", **G. Kouzas**, I. Anagnostopoulos, I. Maglogiannis, and C. Anagnostopoulos, in S. Kollias et al. (Eds.): ICANN 2006, Part II, LNCS 4132, pp. 104–112, 2006.

Β2 "Ant Seeker: An algorithm for enhanced web search", **G. Kouzas**, E. Kayafas and V. Loumos , 2006 in IFIP International Federation for Information Processing, Volume 204, Artificial Inteligence Applications and Innovation, eds. pp 649-656

Β3 "An intelligent system for detecting web commerce transactions", I. Anagnostopoulos, **G. Kouzas**, C. Anagnostopoulos, E. Kayafas and V. Loumos, P.S. Szczepaniak et al. (Eds.): AWIC 2005, LNAI 3528, pp. 38 – 43, Springer-Verlag Berlin Heidelberg 2005.

Β3 "Precise photo retrieval on the web with a fuzzy logic\neural network-based meta-search engine", I. Anagnostopoulos, C. Anagnostopoulos, **G. Kouzas** and D. Vergados, Lecture Notes in Computer Science – LNCS/LNAI, Vol. 3025, pp. 43-53, May 2004.

Πρακτικά Συνεδρείων

Σ1. "WEB Similarity Measurements using Ant-Based Search Algorithm", **G. Kouzas**, E. Kayafas, V. Loumos, XVIII IMEKO World Congress, Metrology for a Sustainable Development , September, 17-22, 2006, Rio de Janeiro, Brazil, 2006

Σ2 "Estimating the evolution of categorized web page populations", I. Anagnostopoulos, P. Stavropoulos, **G. Kouzas**, C. Anagnostopoulos, D.D.Vergados, proceeding of the 1st International Workshop on Adaptation and Evolution in Web Systems Engineering (AEWSE 06), 6th International Conference of Web Engineering (ICWE 06), Palo Alto, CA, July 10-14, 2006, ISBN: 1-59593-435-9.

Σ3 "Ant Colony Self-Healing Schemes for Survivable Optical Networks", **G. Kouzas**, I. Anagnostopoulos, A. Rouskas, N.V. Karadimas and V.G. Loumos, IEEE EUROCON 2005 – The International Conference on 'Computer as a Tool', pp. 1345-1348, Belgrade, Serbia and Montenegro, November 21-24, 2005.

Σ4 "Using Sliding Concentric Windows for License Plate Segmentation and Processing", C. Anagnostopoulos , I. Anagnostopoulos, G. Tsekouras, **G. Kouzas**, V. Loumos, E. Kayafas, IEEE 2005 Workshop on Signal Processing Systems (SIPS'05), pp. 337-342, Athens, Greece, November 2-4, 2005.

Σ5 "Ant Colony Route Optimization for Municipal Services", Nikolaos V. Karadimas, **G. Kouzas**, I. Anagnostopoulos, Vassili Loumos and Elefterios Kayafas, paper submitted to the 19th European Simulation Multiconference (SCS-ESM 2005), pp. 381-386, June 1-4, Riga, Latvia.

Σ6 "Measuring the population of web pages in the wild web", **Kouzas G.S**., Stavropoulos P., Anagnostopoulos I., Anagnostopoulos C., Loumos V. and Kayafas E., XVII IMEKO World Congress, pp. 720-725, June 22-27, 2003, Dubrovnik, CROATIA.

Σ7 "A real-time human activity monitoring system for measuring customer behaviour in retail stores", Anagnostopoulos I., Anagnostopoulos C., Pavlaki V., **Kouzas G**., Loumos V. and Kayafas E., International Conference Automatics and Informatics '03, Sofia, Bulgaria, October 6-8 2003, Vol 1, p.p. 21-24.

Σ8 "An Artificial Neural Network Approach for Classifying E-Commerce Web Pages", Anagnostopoulos I., **Kouzas G**., Anagnostopoulos C., Psoroulas I., Vergados D., Loumos V. and Kayafas E., pp. 237-242, IASTED AI 2003, February 10-13, 2003, Innsbruck, Austria.

Σ9 "Automatic Site Classification in a Large Repository under information Filtering and Retrieval Techniques", Anagnostopoulos I., **Kouzas G**., Anagnostopoulos C., Vergados D., Papaleonidopoulos I., Loumos V. and Kayafas E., 11th Mediterranean Electrotechnical Conference, MELECON 2002, pp. 279-283, 7-9 May 2002, Cairo, Egypt.

Σ10 "Neural Network-based identification and diagnostic systems for Industrial applications", Anagnostopoulos C., Anagnostopoulos I., **Kouzas G**., Papadoyiannis K., Kayafas E. and Loumos V., Automatics – Informatics 2001, 30 May – 2 June 2001, Sofia, Bulgaria, pp I-57 – I60.

Σ11 "Image processing and database architecture for Intelligent Transportation Systems", I. Anagnostopoulos, C. Anagnostopoulos, **G. Kouzas**, S. Kotsakis, E. Kayafas and V. Loumos, KTISIVIOS Conference, 27-29 June 2001,pp. 223-231, Santorini, Greece.

Σ12 "High Performance Computing Algorithms for Textile Quality Control", C. Anagnostopoulos, I. Anagnostopoulos, D. Vergados, **G. Kouzas**, E. Kayafas, V. Loumos and G. Stassinopoulos, KTISIVIOS Conference, pp. 97-106, 28-30 June 2001, Santorini, Greece.