Οικονομικό Πανεπιστήμιο Αθηνών

Τμήμα Πληροφορικής

Πτυχιακή Εργασία

Αυτόματη κατασκευή παραδειγμάτων εκπαίδευσης για το χειρισμό ερωτήσεων ορισμού σε συστήματα ερωταποκρίσεων που χρησιμοποιούν μηχανική μάθηση

Γαλάνης Δημήτριος

Α.Μ. 3990036

Επιβλέπων Καθηγητής : Ίων Ανδρουτσόπουλος

Αθήνα 2004

Περιεχόμενα

Περιεχόμενα 1

1. Εισαγωγή 2

1.1Αντικείμενο εργασίας. 2

2. Συστήματα ερωταποκρίσεων 4

2.1 Κατηγορίες ερωτήσεων 4

2.2 Αρχιτεκτονική Συστημάτων Ερωταποκρίσεων 4

3. Μηχανική Μάθηση 8

3.1 Βασικές Έννοιες Μηχανικής Μάθησης 8

3.2 Αλγόριθμός Μηχανών Διανυσμάτων Υποστήριξης (SVM) 9

4. Χειρισμός ερωτήσεων ορισμού με Μηχανική Μάθηση 13

4.1 Συλλογές δεδομένων διαγωνισμών TREC 13

4.2 Νέο σύστημα ερωταποκρίσεων 14

4.3 Νέα μέθοδος ταξινόμησης 15

4.3.1 Βασική ιδέα 15

4.3.2 Πρώτη μορφή αλγόριθμου ομοιότητας 15

4.3.3 Βελτιώσεις αλγορίθμου ομοιότητας 17

4.3.4 Κατώφλια 22

4.3.5 Επιλογή κατωφλίων 26

4.4 Ιδιότητες Μηχανικής Μάθησης 27

4.4.1 Χειρωνακτικά επιλεγμένες ιδιότητες 27

4.4.2 Επιλογή ιδιοτήτων με την χρήση της ακρίβειας 29

5. Πειράματα – Αξιολόγηση συστήματος 33

5.1 Μέτρα αξιολόγησης 33

5.2 Διασταυρωμένη επικύρωση 33

5.3 Αξιολόγηση νέου συστήματος ερωταποκρίσεων-Πειράματα 34

6. Συμπεράσματα-Μελλοντικές προσεγγίσεις 38

1. Εισαγωγή

1.1 Αντικείμενο της εργασίας

Τα τελευταία χρόνια, η εξάπλωση του διαδικτύου είναι ραγδαία και ο όγκος των πληροφοριών στις οποίες ο χρήστης έχει πρόσβαση είναι τεράστιος. Η χρησιμοποίηση αυτού του τεράστιου όγκου πληροφοριών, ο οποίος αφορά μία μεγάλη ποικιλία θεμάτων είναι πολύ σημαντική υπόθεση για ένα μεγάλο αριθμό ανθρώπων. Επιστήμονες, «ερευνητές» ,μαθητές, φοιτητές και πολλοί εργαζόμενοι χρησιμοποιούν το διαδίκτυο για την άντληση πληροφοριών που είναι χρήσιμες για την εργασία τους. Όμως, ο εντοπισμός των απαιτούμενων πληροφοριών δεν είναι μια εύκολη υπόθεση, αφού απαιτείται η γνώση των συγκεκριμένων ιστοσελίδων οι οποίες τις φιλοξενούν. Για να είναι δυνατή η αναζήτηση και εύρεση των κατάλληλων ιστοσελίδων έχουν δημιουργηθεί οι μηχανές αναζήτησης, οι οποίες επιτρέπουν στο χρήστη να αναζητήσει ιστοσελίδες που περιέχουν συγκεκριμένους όρους. Για να είναι ακόμα πιο εύκολος ο εντοπισμός των επιθυμητών ιστοσελίδων οι μηχανές αναζήτησης δεν παρουσιάζουν με τυχαία σειρά τις σελίδες που εντοπίζουν, αλλά τις κατατάσσουν με κάποια κριτήρια προσπαθώντας να παρουσιάσουν στις πρώτες θέσεις αυτές οι οποίες είναι οι πιο χρήσιμες στο χρήστη.

Παρά τις πολλές υπηρεσίες που προσφέρουν οι σημερινές μηχανές αναζήτησης είναι επιθυμητό να εξελιχθούν κατάλληλα έτσι ώστε να γίνουν πιο αποδοτικές. Συγκεκριμένα, σε μια επόμενη γενιά μηχανών αναζήτησης είναι επιθυμητό να γίνονται ερωτήσεις σε φυσική γλώσσα και να μην επιστρέφεται στο χρήστη μία λίστα ιστοσελίδων, αλλά μία σύντομη απάντηση και συγχρόνως ένας σύνδεσμος προς την ιστοσελίδα στην οποία περιέχεται η απάντηση . Ένα σύστημα το οποίο θα επιστρέφει απαντήσεις στο χρήστη και όχι μια λίστα με έγγραφα ονομάζεται σύστημα ερωταποκρίσεων (Question Answering System).Πολλά ελπιδοφόρα αποτελέσματα έχουν αναφερθεί για τα συστήματα ερωταποκρίσεων τα τελευταία χρόνια και κυρίως μετά το 1999 στα πλαίσια του TREC (Text Retrieval Conference). Τα καλύτερα συστήματα μπορούν και απαντούν σωστά πάνω από τα 2/3 των ερωτήσεων. Τα αποτελέσματα αυτά, καθώς και η απαίτηση των χρηστών για την δημιουργία τέτοιων συστημάτων με καλές επιδόσεις, έχουν προκαλέσει μεγάλο ενδιαφέρον και πολλή ερευνητική δραστηριότητα σε αυτόν το τομέα.

Η συγκεκριμένη εργασία αποτελεί μια προσπάθεια διερεύνησης πάνω σε θέματα που αφορούν τον τομέα των συστημάτων ερωταποκρίσεων. Συγκεκριμένα, στοχεύει στη διερεύνηση θεμάτων που αφορούν στην κατασκευή ενός συστήματος, το οποίο χρησιμοποιεί μηχανική μάθηση και εξάγει κατάλληλες απαντήσεις για τις αντίστοιχες ερωτήσεις από τις ιστόσελίδες που επιστρέφει μία μηχανή αναζήτησης. Το σύστημα που δημιουργήθηκε

προσπαθεί να αντιμετωπίσει το πρόβλημα κατασκευής δεδομένων εκπαίδευσης με πρωτότυπο και αυτόματο τρόπο. Για την ολοκλήρωση όλων των σταδίων που είναι απαραίτητα σε ένα τέτοιο σύστημα χρησιμοποιήθηκαν συμπεράσματα και από προηγούμενες ερευνητικές προσπάθειες. Η όλη προσπάθεια εστιάζεται σε μια συγκεκριμένη κατηγορία ερωτήσεων τις ερωτήσεις ορισμού (π.χ. «Τι είναι η θαλασσαιμία;»). Πιο συγκεκριμένα, βασίζεται σε μια μέθοδο εντοπισμού σύντομων απαντήσεων σε ερωτήσεις ορισμού, που προτάθηκε σε προηγούμενη εργασία (Μηλιαράκη 2003) και χρησιμοποιεί έναν αλγόριθμο επιβλεπόμενης μηχανικής μάθησης. Η παρούσα εργασία διερευνά τρόπους αυτόματης παραγωγής παραδειγμάτων εκπαίδευσης για την προηγούμενη μέθοδο, μετατρέποντάς την ουσιαστικά σε μη επιβλεπόμενη.

2. Συστήματα ερωταποκρίσεων

2.1 Κατηγορίες ερωτήσεων

Το σύνολο των ερωτήσεων μπορεί να διακριθεί σε τρεις διαφορετικές κατηγορίες χρησιμοποιώντας ως κριτήριο τον τύπο της απάντησης .

- Ερωτήσεις με καθορισμένη απάντηση (factual questions), οι οποίες διαιρούνται σε υποκατηγορίες όπως:
  - Ερωτήσεις προσώπου π.χ. «Ποιος είναι ο πρωθυπουργός της Ελλάδας;».
  - Ερωτήσεις τοποθεσίας π.χ. «Πού βρίσκεται το Στάδιο Ειρήνης και Φιλίας;».
  - Ερωτήσεις ορισμού π.χ. «Τι είναι η καφεΐνη;».
  - Ερωτήσεις ποσότητας π.χ. «Πόσες μέρες διάρκεσαν οι Ολυμπιακοί Αγώνες του 2004;».
  - Ερωτήσεις χρόνου π.χ. «Πότε έγινε η Άλωση της Κωνσταντινούπολης;»
- Ερωτήσεις γνώμης (opinion questions) π.χ. «Ποιες είναι οι επιπτώσεις της αύξησης της τιμής των καυσίμων στην οικονομία της Ελλάδας;».
- Ερωτήσεις περίληψης (summary questions) π.χ. «Ποια είναι η βασική ιστορία η οποία ξετυλίγεται στις σελίδες του βιβλίου Κώδικας Da Vinci; ».

Η εργασία εστιάζεται σε μια συγκεκριμένη κατηγορία ερωτήσεων στις ερωτήσεις ορισμού διατυπωμένες στα Αγγλικά. Η γενική μορφή των ερωτήσεων ορισμού είναι:

«What/Who is/are/was <*ονομαστική φράση*> ?»

π.χ.

Who was Galileo?

What is poliomyelitis?

What are sunspots?

What is bipolar disorder?

What is Teflon?

Η ονομαστική φράση ονομάζεται όρος της ερώτησης και είναι η φράση η οποία ζητείται να οριστεί. Ο όρος όπως είναι φανερό μπορεί να αποτελείται από 1 ή παραπάνω λέξεις.

2.2 Αρχιτεκτονική συστημάτων ερωταποκρίσεων

Τα συστήματα ερωταποκρίσεων σχεδιάζονται συνήθως ακολουθώντας μια γενική αρχιτεκτονική της οποίας μια αναπαράσταση φαίνεται στο σχήμα 1. Η δημιουργία ενός

συστήματος ερωταποκρίσεων μπορεί να θεωρηθεί ως στιγμιότυπο αυτής της γενικής αρχιτεκτονικής, όμως είναι φανερό ότι δεν είναι απαραίτητο όλα τα συστήματα να υλοποιούν όλες τις δομικές μονάδες του προτεινόμενου μοντέλου.

Για να γίνει αντιληπτό ποιες είναι οι γενικές αρχές λειτουργίας ενός συστήματος ερωταποκρίσεων περιγράφονται συνοπτικά οι μονάδες του προτεινόμενου μοντέλου.

1.Ανάλυση ερωτήσεων

Η ερώτηση αναλύεται, π.χ. μορφολογικά ή συντακτικά, και από την ανάλυση προκύπτουν πληροφορίες όπως η κατηγορία της ερώτησης και οι όροι της.

2.Προεπεξεργασία της συλλογής εγγράφων

Υποθέτοντας ότι το σύστημα έχει πρόσβαση σε μια μεγάλη συλλογή εγγράφων που χρησιμοποιείται σαν βάση γνώσης για την απάντηση ερωτήσεων, ίσως αυτή η συλλογή να απαιτείται να επεξεργαστεί έτσι ώστε να μετασχηματιστεί τελικά σε μια μορφή κατάλληλη για ένα σύστημα πραγματικού χρόνου (π.χ. η εξαγωγή των HTML tags στην περίπτωση που η συλλογή κειμένων έχει προκύψει από μία μηχανή αναζήτησης).

3.Επιλογή υποψηφίων εγγράφων

Σε αυτό το στάδιο επιλέγεται ένα υποσύνολο της συλλογής εγγράφων που αποτελείται από έγγραφα με μεγάλη πιθανότητα να περιέχουν απαντήσεις. Συνήθως χρησιμοποιείται μια μηχανή αναζήτησης και επιλέγονται τα έγγραφα που επιστρέφονται στις υψηλότερες θέσεις.

4.Εξαγωγή υποψήφιων απαντήσεων

Για κάθε ερώτηση εξάγονται από τα αντίστοιχα επιλεγμένα έγγραφα τμήματα κειμένου ως υποψήφιες απαντήσεις. Το κάθε τμήμα κειμένου που εξάγεται είτε αποτελεί μια ολόκληρη πρόταση είτε αποτελεί μια συμβολοσειρά συγκεκριμένου μήκους. Οι υποψήφιες απαντήσεις ταξινομούνται με κριτήρια που υπολογίζουν τη πιθανότητα να είναι σωστές απαντήσεις. Στην περίπτωση του συστήματος της εργασίας, η ταξινόμηση αυτή γίνεται με έναν αλγόριθμο μηχανικής μάθησης. Οι υποψήφιες απαντήσεις είναι όλες οι συμβολοσειρές μήκους 250 χαρακτήρων των επιλεγμένων εγγράφων οι οποίες περιέχουν τον όρο που πρέπει να οριστεί στο κέντρο τους. Στις επόμενες ενότητες ονομάζουμε τις υποψήφιες αυτές απαντήσεις «παράθυρα». Ακολουθούν τρία παραδείγματα παραθύρων (υποψηφίων απαντήσεων) για την ερώτηση «Who was Archimedes?». Από αυτά μόνο το δεύτερο είναι αποδεκτό ως ορισμός του όρου της ερώτησης.

Π.χ
Ερώτηση: Who was **Archimedes**?

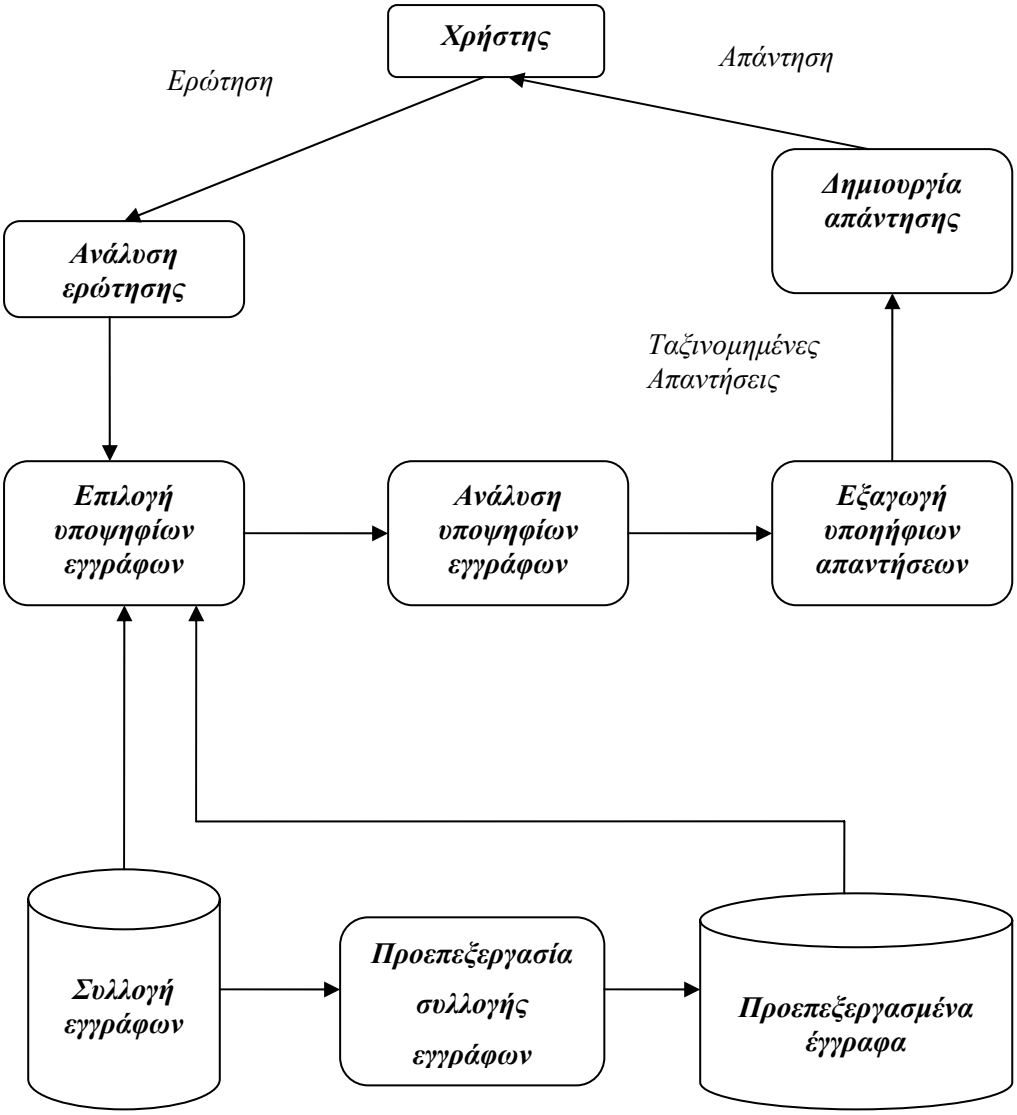
estimating the value of pi. for grades 6-8, 9-12. find it at: www.pbs.org/nova/teachers/activities/3010_archimed.html **archimedes** and the palimpsest learn about archimedes, his hidden manuscript, and the nova program that features him.

nova | infinite secrets | library resource kit | who was archimedes? | pbs who was archimedes? by [author] infinite secrets homepage **archimedes** of syracuse was one of the greatest mathematicians in history.

comes from his writings and those of his contemporaries. born in syracuse, sicily (then part of greece), in about 287 b.c., **archimedes** traveled to egypt at the age of 18 to study at the great library of alexandria. upon completing his studies, he

4.Δημιουργία απάντησης

Σε αυτό το στάδιο μια ή περισσότερες από τις υποψήφιες απαντήσεις επιστρέφονται στο χρήστη. Στην περίπτωση των συστημάτων ερωταποκρίσεων του TREC επιστρέφεται, ανάλογα με το έτος του διαγωνισμού και την κατηγορία των ερωτήσεων, η υποψήφια απάντηση που το σύστημα θεωρεί ότι έχει τη μεγαλύτερη πιθανότητα να είναι σωστή ή μία ταξινομημένη λίστα με τις πέντε «καλύτερες» υποψήφιες απαντήσεις.» Στην περίπτωση αυτής της εργασίας, επιστρέφονται οι πέντε καλύτερες υποψήφιες απαντήσεις, δηλαδή πέντε παράθυρα μήκους 250 χαρακτήρων το καθένα, κάτι που είναι συμβατό με τους κανονισμούς των διαγωνισμών TREC 2000 και TREC 2001 για τις ερωτήσεις ορισμού. Επιπλέον, επιστρέφονται σύνδεσμοι προς τα κείμενα από τα οποία προέρχονται οι απαντήσεις.


Χρήστης
Ερώτηση
Απάντηση
Ανάλυση ερώτησης
Δημιουργία απάντησης
Ταξινομημένες Απαντήσεις
Επιλογή υποψηφίων εγγράφων
Ανάλυση υποψηφίων εγγράφων
Εξαγωγή υποηήφιων απαντήσεων
Συλλογή εγγράφων
Προεπεξεργασία συλλογής εγγράφων
Προεπεξεργασμένα έγγραφα

Σχήμα 1

3. Μηχανική Μάθηση

3.1 Βασικές έννοιες της μηχανικής μάθησης

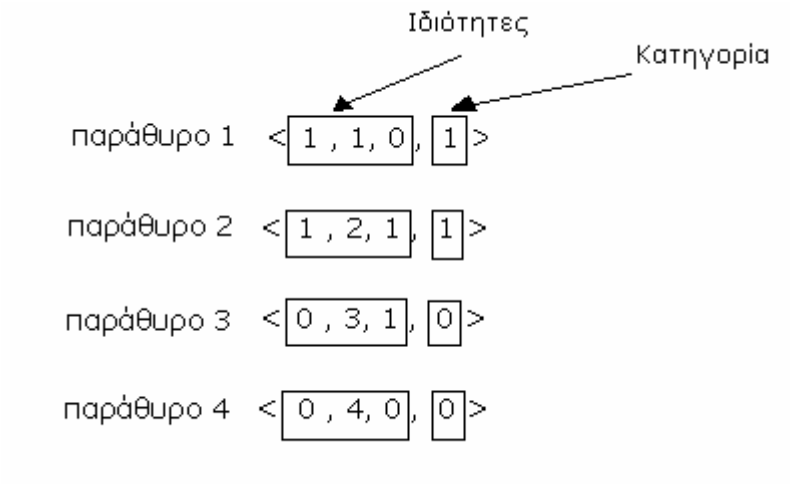
Η μηχανική μάθηση (Mitchell 1997) είναι ένας από τους παλιότερους ερευνητικούς τομείς της ΤΝ(Τεχνητής Νοημοσύνης). Αντικείμενο της μηχανικής μάθησης είναι η κατασκευή προγραμμάτων-συστημάτων ικανών να βελτιώνουν αυτόματα τις επιδόσεις που επιτυγχάνουν κατά την εκτέλεση συγκεκριμένων εργασιών, εκμεταλλευόμενα προηγούμενα εμπειρικά δεδομένα από την εκτέλεση των εργασιών αυτών.

Στην συγκεκριμένη εργασία, η μηχανική μάθηση χρησιμοποιείται για την κατάταξη των υποψηφίων απαντήσεων (των «παραθύρων» της ενότητας 2.2) σε μία από τις εξής δύο κατηγορίες: ορισμός (παράθυρα που είναι πράγματι ορισμοί) και μη-ορισμός. Για την ακρίβεια, υπολογίζονται βαθμοί βεβαιότητας, που δείχνουν πόσο σίγουρο είναι το σύστημα ότι το κάθε παράθυρο ανήκει στη μία ή την άλλη κατηγορία. Οι βαθμοί βεβαιότητας χρησιμοποιούνται κατόπιν κατά την επιλογή των υποψηφίων απαντήσεων (παραθύρων) που θα επιστραφούν στο χρήστη. Όπως αναφέρθηκε ήδη στην ενότητα 2.2, η παρούσα εργασία βασίζεται σε μια μέθοδο προηγούμενης εργασίας (Μηλιαράκη 2003), στην οποία η κατάταξη των παραθύρων γίνεται με τη χρήση ενός αλγορίθμου επιβλεπόμενης μηχανικής μάθησης. Ο αλγόριθμος αυτός απαιτεί παραδείγματα εκπαίδευσης, δηλαδή, στην περίπτωσή μας, παραδείγματα παραθύρων που έχουν καταταγεί χειρωνακτικά στις σωστές κατηγορίες. Η παρούσα εργασία διερευνά πώς τα παραδείγματα εκπαίδευσης μπορούν να παραχθούν αυτόματα αξιοποιώντας ηλεκτρονικές εγκυκλοπαίδειες, ώστε να παρακαμφθεί η χειρωνακτική παραγωγή παραδειγμάτων εκπαίδευσης, η οποία είναι χρονοβόρα και επίπονη

Το σύνολο των παραθύρων κειμένου, που χρησιμοποιούνται για την εκπαίδευση του συστήματος ονομάζεται σώμα εκπαίδευσης, ενώ το σύνολο των παραθύρων που χρησιμοποιούνται για την αξιολόγηση της επίδοσης του εκπαιδευμένου συστήματος , ονομάζεται σώμα αξιολόγησης. Και τα δύο παραπάνω σώματα παριστάνονται σε διανυσματική μορφή, δηλαδή κάθε παράθυρο κειμένου αντιστοιχεί σε ένα διάνυσμα που δίνει τιμές σε ένα συγκεκριμένο σύνολο ιδιοτήτων. Παρακάτω παρουσιάζεται ένα παράδειγμα που δείχνει τις τιμές τριών ιδιοτήτων μερικών παραθύρων και τις κατηγορίες στις οποίες ανήκουν τα παράθυρα

	Ιδιότητα 1: Εμφάνιση της λέξης is μετά τον όρο	*Ιδιότητα 1:* Θέση του παραθύρου στο κείμενο	*Ιδιότητα 3:* Εμφάνιση κόμματος μετά τον όρο	Κατηγορία
Παράθυρο 1	Ναι(1)	1	0	Ορισμός(1)
Παράθυρο 2	Ναι(1)	2	1	Ορισμός(1)
Παράθυρο 3	Όχι(0)	3	1	μη-ορισμός(0)
Παράθυρο 4	Όχι(0)	4	0	μη-ορισμός(0)

Τα παραπάνω παράθυρα παριστάνονται διανυσματικά ως εξής:



3.2 Αλγόριθμος Μηχανών Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVM) είναι μία οικογένεια αλγορίθμων επιβλεπόμενης μάθησης που αναπτύχθηκαν από τον Vapnik και χρησιμοποιούνται ευρύτατα σε προβλήματα κατάταξης [1].

Σε γενικές γραμμές, οι SVM αναπαριστούν τα δεδομένα εκπαίδευσης, στην περίπτωσή μας το σώμα εκπαίδευσης, ως διανύσματα ενός διανυσματικού χώρου μεγάλου πλήθους διαστάσεων και προσπαθούν να εντοπίσουν σε αυτό το χώρο ένα υπερεπίπεδο (επίπεδο σε ένα χώρο με περισσότερες των 3 διαστάσεων) , το οποίο διαχωρίζει τα θετικά από τα αρνητικά παραδείγματα εκπαίδευσης. Θεωρούμε ότι υπάρχουν μόνο δύο κατηγορίες, στην περίπτωση μας «ορισμός» και «μη-ορισμός», μία εκ των οποίων θεωρείται θετική και η άλλη αρνητική. Έχοντας βρει ένα τέτοιο υπερεπίπεδο οι SVM μπορούν να κατατάξουν σε μία από τις δύο κατηγορίες ένα

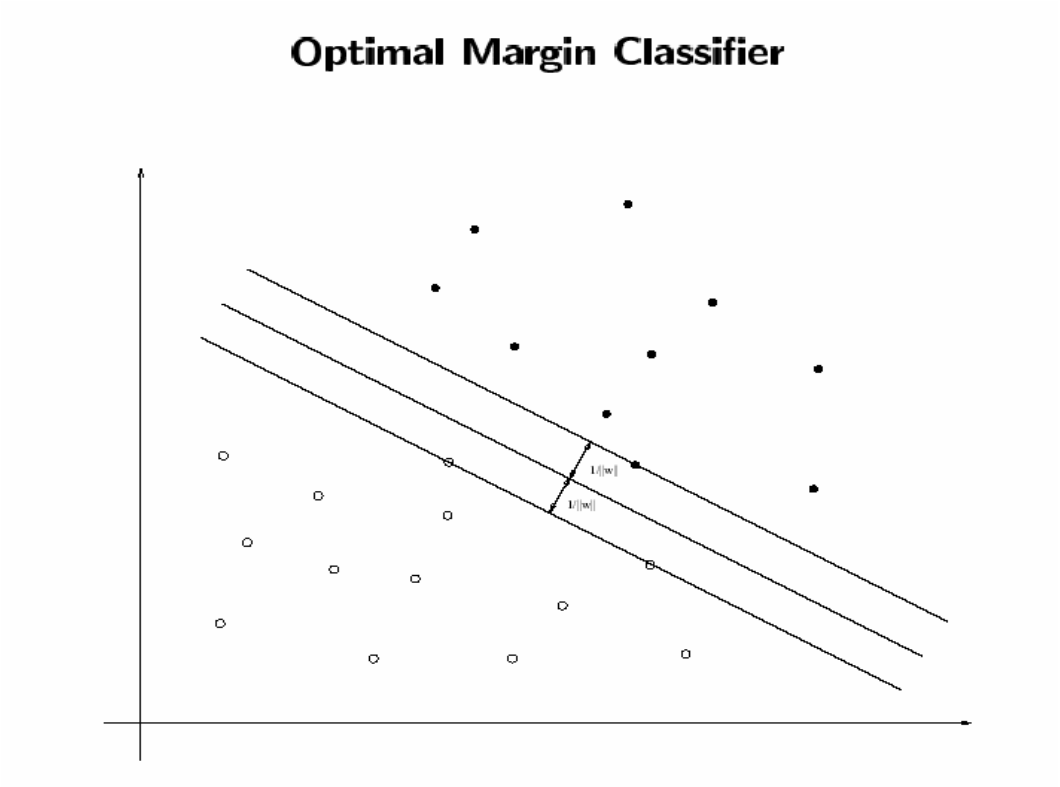
παράδειγμα του οποίου η κατηγορία είναι άγνωστη, αναπαριστώντας το στο διανυσματικό χώρο και ερευνώντας σε ποια πλευρά του υπερεπιπέδου διαχωρισμού βρίσκεται. Μεγάλο πλεονέκτημα των SVM είναι ο τρόπος με τον οποίο επιλέγουν το υπερεπίπεδο διαχωρισμού όταν υπάρχουν πολλά υποψήφια υπερεπίπεδα. Οι SVM επιλέγουν το επίπεδο που διατηρεί το μεγαλύτερο περιθώριο διαχωρισμού από κάθε σημείο του σώματος δεδομένων (Σχήμα 2 [Dobra 2000]), γεγονός πολύ σημαντικό γιατί σύμφωνα με την στατιστική θεωρία μάθησης αυτό μπορεί να οδηγήσει στο μέγιστο βαθμό γενικότητας κατά την κατάταξη αταξινόμητων παραδειγμάτων. Άλλο πλεονέκτημα των SVM είναι η ανοχή που παρουσιάζουν εν γένει στην ύπαρξη θορύβου στα παραδείγματα εκπαίδευσης (π.χ. λόγω λαθών κατά την χειρωνακτική κατάταξη των παραδειγμάτων).

Αν υποθέσουμε ότι έχουμε στη διάθεση μας μία συλλογή δεδομένων εκπαίδευσης η οποία περιέχει n παραδείγματα, που το κάθε ένα είναι ένα διάνυσμα m αριθμών τότε αυτά μπορούν θεωρηθούν ως σημεία σε ένα m-διάστατο χώρο. Ένας απλός τρόπος να κατασκευαστεί ένας δυαδικός ταξινομητής (binary classifier) είναι η δημιουργία ενός υπερεπιπέδου, το οποίο θα διαχωρίζει τα μέλη της μίας κατηγορίας από εκείνα της άλλης. Δυστυχώς, συχνά ο πληθυσμός από τον οποίο προέρχονται τα δεδομένα εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμος στο διανυσματικό χώρο που προκύπτει με αυτόν τον τρόπο, δηλαδή δεν είναι δυνατόν να βρεθεί υπερεπίπεδο που να διαχωρίζει πλήρως τα θετικά από τα αρνητικά παραδείγματα. Μια λύση είναι η απεικόνιση των δεδομένων εκπαίδευσης σε έναν χώρο με μεγαλύτερο πλήθος διαστάσεων, όπου το πρόβλημα είναι γραμμικά διαχωρίσιμο. Με την επιλογή ενός κατάλληλου νέου χώρου με επαρκώς μεγάλο πλήθος διαστάσεων οποιοδήποτε συνεπές σώμα εκπαίδευσης μπορεί να γίνει τελικά γραμμικά διαχωρίσιμο. Ωστόσο, η μετατροπή των δεδομένων σε διανύσματα ενός χώρου με πολύ περισσότερες διαστάσεις επιφέρει υπολογιστικό κόστος. Επιπλέον μεταβαίνοντας σε χώρο πολύ περισσότερων διαστάσεων διατρέχουμε τον κίνδυνο της εύρεσης τετριμμένων λύσεων (trivial solutions) που υπερεφαρμόζουν (over-fitting) τα δεδομένα εκπαίδευσης.

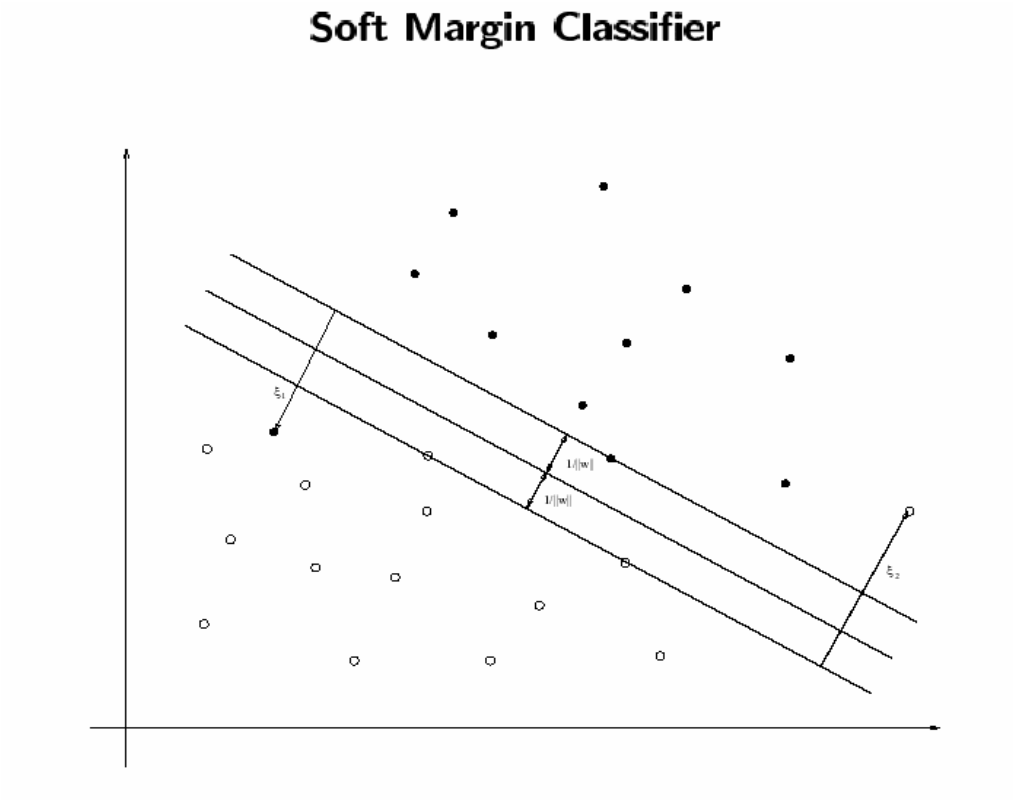
Οι SVM παρακάμπτουν αυτά τα προβλήματα. Το πρόβλημα της υπερ-εφαρμογής αποφεύγεται με την επιλογή του υπερεπιπέδου διαχωρισμού με το μέγιστο περιθώριο. Επειδή ο εντοπισμός του υπερεπιπέδου διαχωρισμού στο νέο χώρο μπορεί γίνει υπολογίζοντας μόνο με εσωτερικά γινόμενα των διανυσμάτων του νέου χώρου, οι SVM μπορούν να εντοπίσούν το υπερεπίπεδο χωρίς ποτέ να αναπαραστήσουν το νέο χώρο ρητά, αλλά ορίζοντας μια συνάρτηση που αποκαλείται συνάρτηση πυρήνα (kernel function), που παίζει το ρόλο του εσωτερικού γινομένου στο νέο χώρο. Αυτή η τεχνική αποφεύγει το υπολογιστικό φορτίο της ρητής αναπαράστασης των διανυσμάτων στο νέο χώρο. Στην περίπτωσή μας, τα αποτελέσματα της προηγούμενης εργασίας (Μηλιαράκη

2003) δείχνουν πως δεν είναι απαραίτητη η μετάβαση σε νέο διανυσματικό χώρο, οπότε ως συνάρτηση πυρήνα χρησιμοποιείται το εσωτερικό γινόμενο στον αρχικό χώρο.

Παρά τη χρήση του νέου διανυσματικού χώρου, οι SVM ενδέχεται να μην μπορούν να εντοπίσουν υπερεπίπεδο που να διαχωρίζει πλήρως τα δεδομένα, είτε γιατί η συνάρτηση πυρήνα είναι ακατάλληλη για τα δεδομένα εκπαίδευσης, είτε γιατί τα δεδομένα περιέχουν λάθος κατηγοριοποιημένα παραδείγματα. Το δεύτερο πρόβλημα μπορεί να λυθεί με την χρήση εύκαμπτου περιθωρίου (Σχήμα 3 [Dobra 2000]) το οποίο επιτρέπει σε κάποια παραδείγματα εκπαίδευσης να βρίσκονται σε λάθος πλευρά του υπερεπιπέδου αντί κάποιας ποινής. Αν θέλουμε να ορίσουμε μια μηχανή SVM πλήρως πρέπει, μεταξύ άλλων να γίνουν συγκεκριμένες δύο παράμετροι: η συνάρτηση πυρήνα και η ποινή από την παραβίαση του εύκαμπτου περιθωρίου.



Σχήμα 2



Σχήμα 3

4. Χειρισμός ερωτήσεων ορισμού με Μηχανική Μάθηση

4.1 Συλλογές δεδομένων διαγωνισμών TREC

Οι διοργανωτές των διαγωνισμών TREC παρέχουν μια συλλογή ερωτήσεων που συμπεριλαμβάνει ερωτήσεις ορισμού. Για κάθε ερώτηση παρέχονται επίσης 50 κείμενα που προέρχονται από άρθρα εφημερίδων και έχουν επιστραφεί από μια μηχανή αναζήτησης χρησιμοποιώντας τις λέξεις της αντίστοιχης ερώτησης ως όρους αναζήτησης. Κάθε κείμενο συνοδεύεται από έναν αριθμό (1-50) ο οποίος δηλώνει την σειρά (κατάταξη) με την οποία επιστράφηκε το κείμενο από την μηχανή αναζήτησης. Οι διαγωνισμοί TREC εκτός από τα κείμενα, παρέχουν και μία λίστα από πρότυπα απαντήσεων για κάθε ερώτηση, με τα οποία γίνεται η κατηγοριοποίηση κάθε παραθύρου κειμένου. Για κάθε ερώτηση, παρέχεται επίσης μια λίστα προτύπων (patterns) γραμμένων στη γλώσσα Perl, που μπορούν να χρησιμοποιηθούν για την αξιολόγηση των απαντήσεων που επιστρέφει ένα σύστημα ως ορθών ή λανθασμένων. Τα πρότυπα έχουν κατασκευαστεί από τους διοργανωτές των διαγωνισμών TREC λαμβάνοντας υπόψη τους όλες τις σωστές απαντήσεις που επέστρεψαν τα συστήματα των διαγωνισμών για τις αντίστοιχες ερωτήσεις Στην προσέγγιση της εργασίας (Μηλιαράκη 2003) στην οποία βασίζεται η παρούσα εργασία, τα πρότυπα αυτά χρησιμοποιούνταν για τη δημιουργία των δεδομένων εκπαίδευσης. Αντί να κατατάσσονται χειρωνακτικά τα παραδείγματα εκπαίδευσης στις δύο κατηγορίες, αν ένα παράθυρο εκπαίδευσης ικανοποιούσε τουλάχιστον ένα πρότυπο της αντίστοιχης ερώτησης εθεωρείτο ότι ανήκει στην κατηγορία «ορισμός» και διαφορετικά στην κατηγορία «μη ορισμός.

Τα πρότυπα απαντήσεων μπορούν επίσης να χρησιμοποιηθούν κατά την αξιολόγηση του συστήματος. Για μια ερώτηση η οποία δεν χρησιμοποιήθηκε για την εκπαίδευση του συστήματος μπορούν να δημιουργηθούν παράθυρα υποψηφίων απαντήσεων, όπως έχει ήδη περιγραφεί και αυτά να δοθούν σαν είσοδος στον ταξινομητή (classifier) που παράγει ο αλγόριθμος μηχανικής μάθησης. Λαμβάνοντας υπόψη τις αποκρίσεις του ταξινομητή, θα επιστραφούν 5 παράθυρα τα οποία έχουν την μεγαλύτερη πιθανότητα να αποτελούν ορθές απαντήσεις, δηλαδή να ανήκουν στην κατηγορία «ορισμός». Με τη χρήση προτύπων απαντήσεων μπορεί να αξιολογηθεί πόσο καλό είναι το σύστημα ερωταποκρίσεων που δημιουργήθηκε, εξετάζοντας αν στις 5 απαντήσεις που επιστράφηκαν υπάρχει κάποια σωστή, δηλαδή κάποια που ταιριάζει με τουλάχιστον ένα πρότυπο της αντίστοιχης ερώτησης.

Παρακάτω δίνονται κάποια πρότυπα για τις ερωτήσεις «Who was Galileo?» και «What is cholesterol?»

Who was Galileo ?

astronomer
the Italian sunspots expert

What is cholesterol ?
steroidlike compound
(fatty|waxy) substance
fat(ty|s)?

Η δημιουργία των προτύπων απαντήσεων γίνεται χειρωνακτικά και δεν είναι μια εύκολη διαδικασία. Σκοπός της παρούσας εργασίας είναι η εξεύρεση μιας εναλλακτικής μεθόδου δημιουργίας παραδειγμάτων εκπαίδευσης, που δεν θα απαιτεί χειρωνακτική κατάταξη των παραδειγμάτων εκπαίδευσης σε κατηγορίες ούτε χειρωνακτική δημιουργία προτύπων απαντήσεων.

4.2 Υπάρχουσες μέθοδοι δημιουργίας παραδειγμάτων εκπαίδευσης

Όπως αναφέρθηκε στις προηγούμενες ενότητες, η μέθοδος χειρισμού ερωτήσεων ορισμού στην οποία βασίζεται η παρούσα εργασία προϋποθέτει μια συλλογή παραδειγμάτων (παραθύρων) εκπαίδευσης, στην οποία πρέπει να φαίνεται η κατηγορία (ορισμός ή μη-ορισμός) στην οποία ανήκει κάθε παράδειγμα. Για την κατάταξη των παραθύρων εκπαίδευσης υπάρχουν δύο μέθοδοι:

a) Η χρήση προτύπων απαντήσεων όπως και στα δεδομένα των διαγωνισμών TREC (ενότητα 4.1).
b) Η χειρωνακτική κατάταξη κάθε παραθύρου εκπαίδευσης ξεχωριστά.

Οι δύο προαναφερθείσες λύσεις παρουσιάζουν η κάθε μία πλεονεκτήματα και μειονεκτήματα. Για την a) τα πλεονεκτήματα είναι:

- Δεν απαιτείται η χειρωνακτική κατάταξη κάθε παραθύρου αλλά η δημιουργία ενός αριθμού προτύπων απαντήσεων για κάθε ερώτηση.
- Τα πρότυπα απαντήσεων χρησιμοποιούνται ήδη από τους διαγωνισμούς TREC με αποτελεσματικό τρόπο.

Το μειονεκτήματα της a) είναι:

- Στην περίπτωση που απαιτείται η προσθήκη ή η ανανέωση του σώματος εκπαίδευσης του συστήματος, πρέπει να δημιουργηθούν και πάλι πρότυπα απαντήσεων για τις νέες ερωτήσεις. Η δημιουργία προτύπων απαντήσεων δεν είναι μια εύκολη διαδικασία, αφού τα πρότυπα απαντήσεων πρέπει να

ταιριάζουν με τον μεγαλύτερο αριθμό δυνατών απαντήσεων που μπορεί να έχει μια ερώτηση.

Για την μέθοδο b) το πλεονέκτημα είναι ότι εξαιτίας του ότι η κατάταξη γίνεται χειρωνακτικά ο θόρυβος στα δεδομένα εκπαίδευσης είναι μικρότερος, από ό,τι στη μέθοδο (a). Όμως, όπως και στα πρότυπα απαντήσεων η δημιουργία ενός νέου σώματος εκπαίδευσης ή η επέκταση ενός παλιού προϋποθέτει τη χειρωνακτική κατάταξη ενός νέου μεγάλου όγκου παραθύρων κειμένου που είναι μια διαδικασία επίπονη και χρονοβόρα.

Όλα τα παραπάνω κάνουν επιθυμητή τη δημιουργία μιας νέας μεθόδου κατάταξης παραθύρων εκπαίδευσης, η οποία θα είναι σε μεγάλο βαθμό αυτόματη.

4.3 Νέα μέθοδος δημιουργίας παραθύρων εκπαίδευσης

4.3.1 Βασική ιδέα

Έστω ότι έχουμε στην διάθεση μας μια ερώτηση ορισμού από την οποία έχει εξαχθεί ο όρος που πρέπει να οριστεί. Έστω επίσης ότι έχουμε και ένα παράθυρο κειμένου που περιέχει τον όρο, το οποίο παράθυρο έχει εξαχθεί από μια ιστοσελίδα με τον τρόπο που έχει περιγραφεί παραπάνω (μήκους 250 χαρακτήρων). Αν είχαμε στην διάθεση μας ένα τουλάχιστον ορισμό του όρου, τότε θα μπορούσαμε δημιουργώντας ένα μέτρο ομοιότητας να συγκρίνουμε τα 2 κομμάτια κειμένου και να αποφασίσουμε αν το παράθυρο κειμένου είναι και αυτό ορισμός του όρου, βάσει της ομοιότητας του με τον ορισμό που διαθέτουμε. Για να υλοποιηθεί αυτή η ιδέα πρέπει να κατασκευαστεί ένας αλγόριθμος που θα συγκρίνει τα δύο προαναφερθέντα κομμάτια κειμένου και να οριστεί πότε αυτά θα θεωρούνται ότι ορίζουν τον ίδιο όρο.

4.3.2 Πρώτη μορφή του αλγόριθμου ομοιότητας

Μια αρχική προσέγγιση της κατασκευής ενός τέτοιου αλγορίθμου είναι να μετρηθεί ο αριθμός των κοινών λέξεων που έχουν τα δύο κείμενα (το κείμενο του ορισμού που διαθέτουμε και το παράθυρο εκπαίδευσης του οποίου δεν γνωρίζουμε την κατηγορία). Στο παρακάτω παράδειγμα φαίνεται ο τρόπος λειτουργίας αυτής της μεθόδου:

Ερώτηση: Who was Archimedes?

Παράθυρο κειμένου

nova | infinite secrets | library resource kit | who was archimedes? | pbs who was **archimedes**? by [author] infinite secrets homepage **archimedes** of syracuse was one of the greatest mathematicians in history.

Ορισμός

A Greek mathematician living from approximately 287 BC to 212 BC in Syracuse. He invented much plane geometry, studying the circle, parabola and three-dimensional geometry of the sphere as well as studying physics. See also Archimedean solid.

Κοινές λέξεις

of, the, in, Syracuse

Παρότι τα δύο παράθυρα είναι και τα δύο ορισμοί του όρου *Archimedes* οι κοινές λέξεις που βρίσκει ο αλγόριθμος δεν φανερώνουν την ομοιότητα τους, καθώς είναι λέξεις που μπορούν να είναι κοινές σε οποιαδήποτε 2 κείμενα που αναφέρονται στον Αρχιμήδη. Επίσης, είναι φανερό ότι ο αλγόριθμος αποτυγχάνει να εντοπίσει κοινές λέξεις που φανερώνουν ότι τα δύο κείμενα είναι και τα δύο ορισμοί του Αρχιμήδη λόγω διαφορετικών καταλήξεων των λέξεων (π.χ. mathematicians - mathematician). Ένα επιπλέον πρόβλημα είναι πως για τον ορισμό ενός όρου είναι δυνατό να χρησιμοποιηθεί μια μεγάλη ποικιλία διαφορετικών εκφράσεων και λέξεων. Επομένως ο αλγόριθμος θα εντοπίσει μόνο τις λέξεις του παραθύρου που χρησιμοποιούνται και στον συγκεκριμένο ορισμό που έχουμε στη διάθεσή μας». Το πρόβλημα αυτό γίνεται φανερό αν δοθεί ένα παράδειγμα ενός όρου ο οποίος μπορεί να οριστεί με διαφορετικούς τρόπους.

What is a **galaxy**?

- A large aggregation of stars, bound together by gravity. There are three major classifications of galaxies-spiral, elliptical, and irregular.
- a very large cluster of stars (tens of millions to trillions of stars) gravitationally bound together.
- an organized system of many hundreds of millions of stars, often mixed with gas and dust. The universe contains billions of galaxies.
- a component of our Universe made up of gas and a large number (usually more than a million) of stars held together by gravity.
- A large grouping of stars. Galaxies are found in a variety of sizes and shapes. Our own Milky Way galaxy is spiral in shape and contains several billion stars. Some galaxies are so distant the their light takes millions of years to reach the Earth.

Για όλους τους παραπάνω λόγους ο απλοϊκός αυτός αλγόριθμος ομοιότητας αποτυγχάνει στην συντριπτική πλειοψηφία των περιπτώσεων.

4.3.3 Βελτιώσεις αλγορίθμου ομοιότητας

Με σκοπό την εξάλειψη των λόγων που κάνουν τον προηγούμενο αλγόριθμο ομοιότητας να αποτυγχάνει υιοθετήθηκαν κάποιες βελτιώσεις οι οποίες είναι:

- Αφαίρεση από κάθε παράθυρο κειμένου των 100 συχνότερων λέξεων που εμφανίζονται σε αγγλικά κείμενα π,χ ("the", "be", "of", "and", "a", "in", "to", "have", "it", "to", "for", "i", "that", "you", "he", "on", "with", "do", "at", "by", "not", "this"). Οι 100 συχνότερες λέξεις έχουν προκύψει από το British National Corpus (http://www.itri.bton.ac.uk/~Adam.Kilgarriff/bnc-readme.html). Η αφαίρεση γίνεται διότι κατά την σύγκριση δύο κειμένων η εύρεση κοινών λέξεων που είναι πολύ συχνές δεν φανερώνει ομοιότητα.
- Εφαρμογή ενός αλγορίθμου που αποκόπτει την κατάληξη κάθε λέξης αφήνοντας προς σύγκριση μόνο την ρίζα της (stemmer π.χ. το approximately γίνεται approxim, το invented γίνεται invent). Ο αλγόριθμος που χρησιμοποιήθηκε είναι ο Porter Stemming Algorithm (http://www.tartarus.org/~martin/PorterStemmer)
- Διαγραφή από κάθε παράθυρο των ειδικών συμβόλων (!@&^%$#., κ.λ.π.).
- Σύγκριση κάθε παραθύρου εκπαίδευσης με περισσότερους από έναν ορισμούς του αντιστοίχου όρου, διότι ο τρόπος με τον οποίο μπορεί να οριστεί κάποιος όρος ποικίλει και μπορεί να γίνει με την χρησιμοποίηση αρκετών διαφορετικών λέξεων. Για να επιτευχθεί αυτό χρησιμοποιήθηκε η μηχανή αναζήτησης Google η οποία παρέχει μια υπηρεσία αυτόματης παρουσίασης διαφορετικών ορισμών που προέρχονται από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια εάν χρησιμοποιηθεί μια έκφραση αναζήτησης της μορφής define: <όρος>.

Συγκεκριμένα ο βελτιωμένος αλγόριθμος υπολογίζει έναν αριθμό (score), ο οποίος μετρά την ομοιότητα ενός παραθύρου εκπαίδευσης με όλους τους ορισμούς εγκυκλοπαιδειών και γλωσσαρίων του ιδίου όρου που έχουμε στη διάθεσή μας. Στόχος είναι κάθε παράθυρο εκπαίδευσης το οποίο έχει κοινό περιεχόμενο με τους ορισμούς του όρου που διαθέτουμε, να έχει υψηλό score, ενώ αντίθετα κάθε παράθυρο που δεν έχει κοινό περιεχόμενο, να έχει χαμηλό score. Με αυτό τον τρόπο θα καταστεί δυνατό να γίνει ο διαχωρισμός των παραθύρων εκπαίδευσης που ανήκουν στην κατηγορία «ορισμός» από τα παράθυρα εκπαίδευσης που ανήκουν στην κατηγορία «μη ορισμός». Για να εφικτή η παραγωγή αυτού

του score δημιουργήθηκε ένα μαθηματικό μοντέλο με το οποίο υπολογίζεται το score. Ο επιστρεφόμενος αριθμός (score) κάθε παραθύρου είναι το άθροισμα των βαρών των λέξεων του παραθύρου δια τον αριθμό των λέξεων του παραθύρου.

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

n ο αριθμός των λέξεων του παραθύρου. Λέξεις που εμφανίζονται στο παράθυρο πολλές φορές υπολογίζονται μόνο μία φόρά

w_i το βάρος της λέξης i.

Το βάρος κάθε λέξης υπολογίζεται ως εξής:

$$w = fdef * idf$$

w το βάρος της λέξης

fdef το ποσοστό των ορισμών που περιέχουν την λέξη

$fdef = \frac{cdefs}{defs}$ cdefs = ορισμοί που διαθέτουμε για τον όρο που πρέπει να οριστεί και που περιέχουν τη λέξη , defs = ολικός αριθμός ορισμών που διαθέτουμε για τον όρο που πρέπει να οριστεί.

idf (inverse document frequency) η αντίστροφη συχνότητα της λέξης.

Το idf ορίζεται ως εξής:

$$idf = 1 + \log \frac{N}{df}$$

N ο ολικός αριθμός των εγγράφων του British National Corpus (BNC)

df ο αριθμός των εγγράφων του British National Corpus που περιέχουν την λέξη

Όπως είναι φανερό κάθε λέξη που υπάρχει στο παράθυρο αλλά και στους ορισμούς δεν συμβάλει ισόποσα στην διαμόρφωση του score που φανερώνει την ομοιότητα του παραθύρου με τους ορισμούς. Οι λέξεις που εμφανίζονται σε περισσότερους ορισμούς, δηλαδή αυτές που πιθανότατα χρησιμοποιούνται συχνότερα για την περιγραφή του όρου που πρέπει να οριστεί έχουν μεγαλύτερη βαρύτητα όταν βρεθούν σε κάποιο παράθυρο

εκπαίδευσης (fdef). Επίσης, όσο πιο σπάνια είναι μια λέξη «στην αγγλική γλώσσα, τόσο μεγαλύτερη είναι η βαρύτητά της, αφού είναι λιγότερο πιθανό η εμφάνισή της στο παράθυρο και τους ορισμούς να οφείλεται στο ότι είναι πολύ συνηθισμένη λέξη (για αυτό το λόγο έχουν ήδη αφαιρεθεί από όλα τα παράθυρα οι 100 πιο συχνές λέξεις). Οι αριθμητικές τιμές των συχνοτήτων εμφάνισης των λέξεων που χρησιμοποιήθηκαν, έχουν προκύψει και αυτές από το BNC. Για να γίνει αντιληπτό πως λειτουργεί ο αλγόριθμος που παρουσιάστηκε ακολουθεί ένα παράδειγμα

Ερώτηση: What is a **palindrome?**

Οι ορισμοί που επέστρεψε το Google δίνοντας σαν όρο το palindrome (define: palindrome) είναι:

Words, numbers and phrases that can be read the same backwords as forwards. Some examples include: "mom", "racecar", "34543", or the phrase "never odd or even".

A word or phrase that reads the same forwards and backwards. Examples in English are repaper and Able was I ere I saw Elba (facetiously attributed to Napoleon).

a word, phrase, clause, or sentence that reads the same regularly as it does when its letters are reversed; a type of palingram. "A man, a plan, a canal, Panama." See also: palingram.

In molecular biology, a nucleotide sequence in which the 5'to 3' sequence of 1 strand of a segment of DNA is the same as that of its complementary strand. The sites of many restriction enzymes are palindromes.

A positive integer whose digits read the same forward and backwards.

A palindrome is a number that reads the same from left to right and from right to left. 101 is the smallest 3-digit palindrome. 123454321 is a palindrome.

(3 syl.). A word or line which reads backwards and forwards alike, as Madam, also Roma tibi subito motibus ibit amor. (Greek, palin dromo, to run back again.) (See Sotadic.) The following Greek palindrome is very celebrated:- NIsiONANOMHMATAMHMONANOsiIN (Wash my transgressions, not only my face). The legend round the font at St. Mary's, Nottingham. Also on the font in the basilica of St. Sophia, Constantinople; also on the font of St. Stephen d'Egres, Paris; at St. Menin's Abbey, Orléans; at Dulwich College; and at the following churches: Worlingsworth (Suffolk), Harlow (Essex), Knapton (Norfolk), Melton Mowbray (it has been removed to a neighbouring hamlet), St. Martin's, Ludgate (London), and Hadleigh (Suffolk). (See Ingram: Churches of London, vol. ii.; Malcolm: Londinum Redivivum, vol. iv. p. 356; Allen: London, vol. iii. p. 530.) It is said that when Napoleon was asked whether he could have invaded England, he answered “Able was I ere I saw Elba.”

(PAL-uhn-drohm), n.: a word or phrase that reads the same backward as forward

a word or phrase that reads the same backward as forward

Παίρνοντας τον πρώτο μόνο ορισμό και ακολουθώντας τα βήματα του αλγορίθμου έχουμε:

Διαγραφή συμβόλων !@#$%^&*() κ.λ.π.

words numbers and phrases that can be read the same backwords as forwards some examples include mom racecar 34543 or the phrase never odd or even

Διαγραφή λέξεων που εμφανίζονται πάνω από μια φορά έτσι ώστε να περιέχονται στο παράθυρο μόνο μια φορά. Οι λέξεις που διαγράφονται σημειώνονται με έντονη γραφή.

words numbers and phrases that can be read the same backwords as forwards some examples include mom racecar 34543 or **the** phrase never odd **or** even

Διαγραφή των λέξεων που συμπεριλαμβάνονται στις 100 πιο συχνές του BNC. Οι λέξεις που διαγράφονται σημειώνονται με έντονη γραφή.

words numbers **and** phrases **that can be** read **the** same backwords **as** forwards **some** examples include mom racecar 34543 **or** phrase never odd even

Αποκοπή των καταλήξεων των λέξεων

word number phrase read same backword forward exampl includ mom racecar 34543 phrase never odd even

Αν εφαρμόσουμε τον ίδιο αλγόριθμο σε όλους τους ορισμούς τα κομμάτια κειμένου που θα προκύψουν θα είναι.

a) word number phrase read same backword forward exampl includ mom racecar 34543 phrase never odd even
b) word phrase read same forward backward exampl english ar repap abl wa er saw elba faceti attribut napoleon
c) word phrase claus sentenc read same regularli doe letter ar revers type palingram man plan canal panama
d) molecular biologi nucleotid sequenc 5 3 1 strand segment dna is same complementari site mani restrict enzym ar palindrome
e) posit integ whose digit read same forward backward
f) palindrom is number read same left right 101 smallest 3 digit 123454321

g) 3 syl word line read backward forward alik madam roma tibi subito motibu ibit amor greek palin dromo run back again sotad follow palindrom is celebr ni si onanomhmatamhmonano wah transgress face legend round font st mari s nottingham basilica sophia constantinopl stephen d egr pari menin abbei orlıan dulwich colleg church worlingsworth suffolk harlow essex knapton norfolk melton mowbrai ha been remov neighbour hamlet martin ludgat london hadleigh ingram vol ii malcolm londinum redivivum iv p 356 allen iii 530 said napoleon wa ask whether invad england answer 147 abl er saw elba 148

h) pal uhn drohm n word phrase read same backward forward

i) word phrase read same backward forward

Έστω το παρακάτω παράθυρο εκπαίδευσης κειμένου στο οποίο εφαρμόζονται τα ίδια βήματα που εφαρμόστηκαν σε κάθε ένα από τα παράθυρα ορισμού.

welcome! bradford elementary school palindromes what is a palindrome?by courtney what is a palindrome? a palindrome is a word or a sentence or number that is the same turned around,

Τότε θα είχαμε σαν αποτέλεσμα το ακόλουθο κομμάτι κειμένου.

welcom bradford elementari school palindrom is palindrom courtnei word sentenc number same turn around

Για κάθε λέξη του τελικού κειμένου υπολογίζουμε τον βάρος της σύμφωνα με τον τύπο $w = fdef * idf$.

Π.χ

$$w_{(welcom)} = fdef_{(welcom)} * idf_{(welcom)} = \frac{cdefs}{defs} * \left(1 + \log\frac{N}{df}\right) =$$

$$= \frac{0}{9}\left(1 + \log\frac{4124}{1039}\right) = 0$$

$$w_{(\text{word})} = fdef_{(\text{word})} * idf_{(\text{word})} = \frac{6}{9}\left(1 + \log\frac{4124}{2683}\right) = 0.95$$

$$w_{(\text{same})} = fdef_{(\text{same})} * idf_{(\text{same})} = \frac{8}{9}\left(1 + \log\frac{4124}{3723}\right) = 0.97$$

Έχοντας υπολογίσει το βάρος για κάθε λέξη μπορούμε να υπολογίσουμε το score του παραθύρου αθροίζοντας και διαιρώντας με τον αριθμό των λέξεων που περιέχει το συγκεκριμένο παράθυρο.

$$score = \frac{\sum_{i=1}^{n} w_i}{n} = \frac{w_{(welcom)} + w_{(bradford)} + \ldots + w_{(around)}}{14} = 0.59$$

4.3.4 Κατώφλια

Με τον αλγόριθμο που περιγράφτηκε παραπάνω για κάθε παράθυρο εκπαίδευσης που δεν γνωρίζουμε αν ανήκει στην κατηγορία ορισμός ή στην κατηγορία μη-ορισμός παράγουμε ένα αριθμό, ο οποίος δηλώνει την ομοιότητα του με τους ορισμούς του όρου της ερώτησης που διαθέτουμε από εγκυκλοπαίδειες και γλωσσάρια. Για την ταξινόμηση των παραθύρων εκπαίδευσης στις κατηγορίες ορισμός και μη ορισμός θα πρέπει να βρεθούν 2 αριθμητικά κατώφλια κοινά για όλες τις ερωτήσεις για τα οποία θα ισχύει:

- Το κατώφλι που είναι μεγαλύτερο θα ονομάζεται άνω κατώφλι.
- Το κατώφλι που είναι μικρότερο θα ονομάζεται κάτω κατώφλι.
- Τα παράθυρα που θα έχουν score μεγαλύτερο του άνω κατωφλιού θα είναι παράθυρα ορισμού με μεγάλη πιθανότητα.
- Τα παράθυρα που θα έχουν score μικρότερο του κάτω κατωφλιού θα είναι παράθυρα μη-ορισμού με μεγάλη πιθανότητα.

Η εύρεση του άνω και κάτω κατωφλιού θα διαχωρίσει τα παράθυρα εκπαίδευσης σε 3 είδη, α) παράθυρα για τα οποία είμαστε σχεδόν σίγουροι ότι είναι ορισμοί, β) παράθυρα για τα οποία είμαστε σχεδόν σίγουροι ότι δεν είναι ορισμοί και γ) σε παράθυρα για τα οποία δεν μπορεί να αποφασιστεί με την απαιτούμενη βεβαιότητα σε ποια κατηγορία ανήκουν. Για την εκπαίδευση του ταξινομητή της προηγούμενης εργασίας θα χρησιμοποιήσουμε μόνο παράθυρα των πρώτων δύο ειδών. Αν τα περισσότερα παράθυρα ανήκουν στο τρίτο είδος, τότε τα παράθυρα που απομένουν για την εκπαίδευση του ταξινομητή θα είναι πολύ λίγα και κινδυνεύουμε ο ταξινομητής να εκπαιδευθεί σε ένα δείγμα παραθύρων που δεν θα είναι αντιπροσωπευτικό του συνολικού πληθυσμού παραθύρων που θα χρειαστεί να κατατάξει μετά την εκπαίδευσή του. Επίσης, όσο αυξάνουν τα παράθυρα του τρίτου είδους, τόσο μεγαλώνει ο αριθμός των ερωτήσεων, ορισμών από εγκυκλοπαίδειες και παραθύρων ιστοσελίδων που πρέπει να συγκεντρωθούν αρχικά, ώστε να απομείνει ικανός αριθμός παραθύρων εκπαίδευσης. Επίσης αν τα παράθυρα του πρώτου είδους είναι πολύ λιγότερα από εκείνα του δεύτερου, ή αντίστροφα, τότε και πάλι θα δημιουργείται πρόβλημα κατά την εκπαίδευση του ταξινομητή, επειδή ο ταξινομητής είναι πιθανόν να μάθει να κατατάσσει όλα τα παράθυρα είτε ως ορισμούς είτε ως μη-ορισμούς.

Για να διερευνηθεί αν ο προτεινόμενος αλγόριθμος με τα κατώφλια μπορεί να χρησιμοποιηθεί για τη δημιουργία ενός ικανοποιητικού σε μέγεθος και σε ορθότητα σώματος εκπαίδευσης, διενεργήθηκε το εξής πείραμα. Χρησιμοποιώντας 130 ερωτήσεις ορισμού για

τις οποίες διαθέταμε ορισμούς από ηλεκτρονικές εγκυκλοπαίδειες και γλωσσάρια δημιουργήθηκαν παράθυρα κειμένου από τις ιστοσελίδες που επέστρεψε για την κάθε μια η μηχανή αναζήτησης Altavista. Για κάθε ερώτηση κρατήθηκαν οι 10 πρώτες ιστοσελίδες και τα 5 πρώτα παράθυρα κειμένου των ιστοσελίδων, διότι σε αυτά είναι πιθανότερο να περιέχονται ορισμοί. Με τυχαίο τρόπο επιλέχθηκαν 400 παράθυρα από το σύνολο αυτών και κατατάχθηκαν χειρωνακτικά στις δύο κατηγορίες (ορισμοί και μη-ορισμοί. Ύστερα, υπολογίσθηκαν για διάφορες τιμές κατωφλιών 4 μεγέθη. Η **ακρίβεια** (Precision) και η **ανάκληση** (Recall) για τα παράθυρα ορισμού και για τα παράθυρα μη-ορισμού. Ο ορισμός των μεγεθών δίνεται παρακάτω.

$$precision_{\mu\eta-ορισμών} = \frac{x = 0 \rightarrow y = 0}{x = 0 \rightarrow y = 0 + x = 1 \rightarrow y = 0}$$

$$recall_{\mu\eta-ορισμών} = \frac{x = 0 \rightarrow y = 0}{x = 0 \rightarrow y = 0 + x = 0 \rightarrow y = 1}$$

$$precision_{ορισμών} = \frac{x = 1 \rightarrow y = 1}{x = 1 \rightarrow y = 1 + x = 0 \rightarrow y = 1}$$

$$recall_{ορισμών} = \frac{x = 1 \rightarrow y = 1}{x = 1 \rightarrow y = 1 + x = 1 \rightarrow y = 0}$$

Το 0 παριστάνει την κατηγορία των μη-ορισμών και το 1 την κατηγορία των ορισμών. Με x συμβολίζεται η κατηγορία σύμφωνα με την χειρωνακτική μέθοδο ταξινόμησης και με y η κατηγορία σύμφωνα με τη μέθοδο που χρησιμοποιεί τον αλγόριθμο ομοιότητας και τα κατώφλια. Η έκφραση $x = a \rightarrow y = b$, όπου $a, b \in \{0,1\}$ είναι ο αριθμός των παραθύρων για τα οποία ισχύει ταυτόχρονα $x = a$ και $x = b$. Για παράδειγμα η έκφραση $x = 0 \rightarrow y = 0$ είναι ο αριθμός των παραθύρων για τα οποία και η χειρωνακτική και η μέθοδος με τα κατώφλια έχουν συμφωνήσει ότι είναι παράθυρα μη ορισμού. Τα 4 παραπάνω μεγέθη υπολογίστηκαν θέτοντας το άνω κατώφλι ίσο με το κάτω και μεταβάλλοντας την τιμή του κοινού κατωφλιού. Τα αποτελέσματα αποτυπώνονται στον Πίνακα 1. Αν υποθέσουμε ότι η χειρωνακτική ταξινόμηση των παραθύρων έγινε με αντικειμενικό τρόπο και χωρίς λάθη τότε:

- το $precision_{ορισμών/\mu\eta-ορισμών}$ εκφράζει το ποσοστό των παραθύρων που ο αλγόριθμος ομοιότητας ταξινομεί ως παράθυρα ορισμού/μη-ορισμού που πράγματι είναι παράθυρα ορισμού/μη-ορισμού.
- το $recall_{ορισμών/\mu\eta-ορισμών}$ εκφράζει το ποσοστό των συνολικών παραθύρων ορισμού/μη-ορισμού (χειρωνακτικά κατηγοριοποιημένα) που

κατηγοριοποιήθηκαν από τον αλγόριθμο ομοιότητας ως παράθυρα ορισμού/μη-ορισμού.

Threshold	Precision-0	Recall-0	Precision-1	Recall-1
0,02	1	0,089655172	0,294118	1
0,04	1	0,120689655	0,30137	1
0,06	0,980392	0,172413793	0,312321	0,990909091
0,08	0,984127	0,213793103	0,323442	0,990909091
0,1	0,987179	0,265517241	0,338509	0,990909091
0,12	0,989011	0,310344828	0,352751	0,990909091
0,14	0,980769	0,351724138	0,364865	0,981818182
0,16	0,974576	0,396551724	0,379433	0,972727273
0,18	0,970588	0,455172414	0,401515	0,963636364
0,2	0,97351	0,506896552	0,425703	0,963636364
0,22	0,976471	0,572413793	0,46087	0,963636364
0,24	0,962963	0,627586207	0,488152	0,936363636
0,26	0,965	0,665517241	0,515	0,936363636
0,28	0,951923	0,682758621	0,520833	0,909090909
0,3	0,933921	0,731034483	0,549133	0,863636364
0,32	0,915966	0,751724138	0,555556	0,818181818
0,34	0,905738	0,762068966	0,557692	0,790909091
0,36	0,899614	0,803448276	0,595745	0,763636364
0,38	0,874552	0,84137931	0,619835	0,681818182
0,4	0,862543	0,865517241	0,642202	0,636363636
0,42	0,847682	0,882758621	0,653061	0,581818182
0,44	0,846906	0,896551724	0,677419	0,572727273
0,46	0,834921	0,906896552	0,682353	0,527272727
0,48	0,829721	0,924137931	0,714286	0,5
0,5	0,827692	0,927586207	0,72	0,490909091
0,52	0,819277	0,937931034	0,735294	0,454545455
0,54	0,810089	0,94137931	0,730159	0,418181818
0,56	0,805882	0,944827586	0,733333	0,4
0,58	0,797101	0,948275862	0,727273	0,363636364
0,6	0,777159	0,962068966	0,731707	0,272727273
0,62	0,773481	0,965517241	0,736842	0,254545455
0,64	0,769231	0,965517241	0,722222	0,236363636
0,66	0,764228	0,972413793	0,741935	0,209090909
0,68	0,762803	0,975862069	0,758621	0,2
0,7	0,763441	0,979310345	0,785714	0,2
0,72	0,757333	0,979310345	0,76	0,172727273
0,74	0,755319	0,979310345	0,75	0,163636364
0,76	0,755319	0,979310345	0,75	0,163636364
0,78	0,751323	0,979310345	0,727273	0,145454545
0,8	0,747368	0,979310345	0,7	0,127272727
0,82	0,744125	0,982758621	0,705882	0,109090909
0,84	0,742188	0,982758621	0,6875	0,1
0,86	0,740933	0,986206897	0,714286	0,090909091
0,88	0,741602	0,989655172	0,769231	0,090909091
0,9	0,741602	0,989655172	0,769231	0,090909091
0,92	0,742268	0,993103448	0,833333	0,090909091
0,94	0,742268	0,993103448	0,833333	0,090909091
0,96	0,74359	1	1	0,090909091

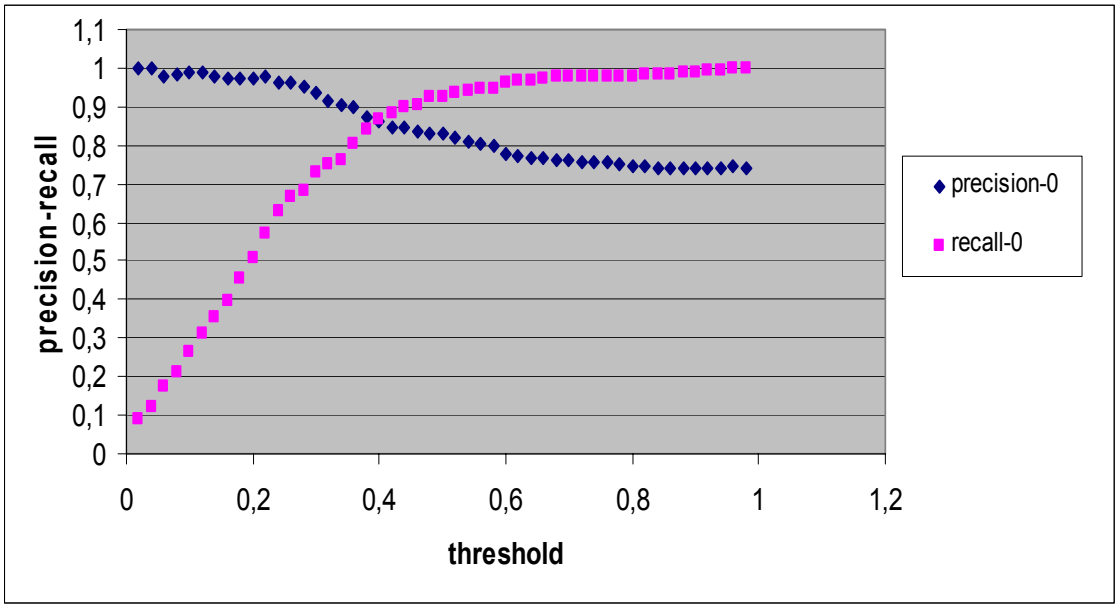
0,98	0,741688	1	1	0,081818182
1	0,736041	1	1	0,054545455

Πίνακας 1

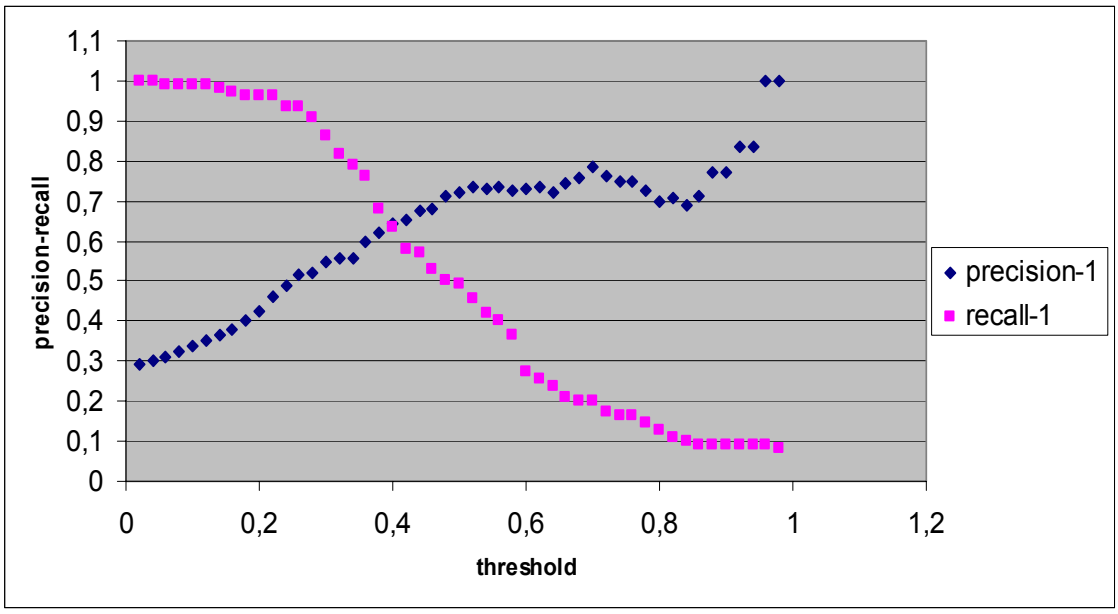
Από την μελέτη του παραπάνω πίνακα γίνεται αντιληπτό ότι:

a) Μείωση του κατωφλιού έχει ως συνέπεια την αύξηση του $precision_{μη-ορισμών}$ και μείωση του $recall_{μη-ορισμών}$ (Σχήμα 1).

b) Αύξηση του κατωφλιού έχει ως συνέπεια την αύξηση του $precision_{ορισμών}$ και μείωση του $recall_{ορισμών}$ (Σχήμα 2).



Σχήμα 1



Σχήμα 2

4.3.5 Επιλογή κατωφλίων

Από την Σχήματα 1 και 2 γίνεται φανερό πως ο βελτιωμένος αλγόριθμος ομοιότητας επιτυγχάνει να διαχωρίσει τα παράθυρα κειμένου με ικανοποιητικό τρόπο, έτσι ώστε όσο μεγαλύτερο score έχουν, τόσο μεγαλύτερη πιθανότητα έχουν να είναι παράθυρα ορισμού. Το ίδιο συμβαίνει και με τα παράθυρα μη-ορισμού αφού όσο μικρότερο score έχει ένα παράθυρο, τόσο μεγαλύτερη πιθανότητα έχει να είναι παράθυρο ορισμού. Το πρόβλημα όμως που δημιουργείται είναι ότι όταν μεγαλώνει η ακρίβεια (precision) για οποιαδήποτε από τις δύο κατηγορίες η αντίστοιχη ανάκληση (recall) μικραίνει. Αυτό σημαίνει ότι κρατώντας στο σώμα εκπαίδευσης παράθυρα με υψηλό βαθμό ομοιότητας μπορούμε να είμαστε σχεδόν βέβαιοι ότι πρόκειται για παράθυρα ορισμού αλλά δεν συμπεριλαμβάνουμε πολλά άλλα παράθυρα ορισμού. Αντίστοιχα, κρατώντας παράθυρα με χαμηλό βαθμό ομοιότητας μπορούμε να είμαστε σχεδόν βέβαιοι ότι πρόκειται για παράθυρα μη-ορισμού αλλά δεν συμπεριλαμβάνουμε πολλά άλλα παράθυρα της κατηγορίας αυτής. Επομένως, το άνω και κάτω κατώφλι πρέπει να επιλεγούν έτσι, ώστε να είμαστε σχεδόν σίγουροι για τις κατηγορίες των παραδειγμάτων εκπαίδευσης (υψηλή ακρίβεια) συμπεριλαμβάνοντας στο σώμα εκπαίδευσης όσο το δυνατόν περισσότερα παράθυρα (υψηλή ανάκληση).

Αν ληφθούν υπόψιν όλα τα παραπάνω το βέλτιστο άνω κατώφλι «(για την επιλογή παραδειγμάτων ορισμών) φαίνεται να είναι το 0.5, γιατί έχει υψηλό precision και υψηλό recall. Επίσης, όπως φαίνεται στο Σχήμα 2 αν χρησιμοποιηθεί κάποιο μεγαλύτερο άνω κατώφλι δεν αυξάνεται σημαντικά η ακρίβεια (precision), ενώ η ανάκληση (recall) μειώνεται δραστικά. Όσο για το κάτω κατώφλι οποιαδήποτε τιμή από 0.12 έως 0.34 είναι

ικανοποιητική (Σχήμα 1), καθώς σε αυτό το εύρος συνδυάζεται εξαιρετικά υψηλή ακρίβεια και ικανοποιητική ανάκληση.

Ακόμη για την επιλογή των τελικών κατωφλίων θα πρέπει να ληφθεί υπόψιν και μια πρόσθετη παράμετρος. Στην περίπτωση της συλλογής παραθύρων του σώματος εκπαίδευσης που χρησιμοποιήσαμε, αν επιλέξουμε για κατώφλια το 0.15 και το 0.5 η αναλογία παραθύρων μη-ορισμών/ορισμών είναι 60%-40%. Όμως, όπως έδειξε η χειρωνακτική ταξινόμηση η πραγματική αναλογία είναι 73%-27%.Προκειμένου να μην κλίνουν (bias) οι αποφάσεις του ταξινομητή που θα προκύψει από την εκπαίδευση υπέρ της μίας ή της άλλης κατηγορίας, θα πρέπει το νέο σώμα εκπαίδευσης που προκύπτει με τη χρήση των δύο κατωφλίων να διατηρεί την αναλογία της αρχικής συλλογής παραθύρων. Άρα, για την δημιουργία της σωστής αναλογίας θα πρέπει είτε να αυξηθεί το κάτω κατώφλι είτε το πάνω. Όμως, όπως έχει ήδη αναφερθεί η αύξηση του πάνω κατωφλίου έχει ως συνέπεια την σημαντική μείωση της ανάκλησης για την κατηγορία ορισμών χωρίς την αύξηση της ακρίβειας. Για αυτό επιλέγεται η αύξηση του κάτω κατωφλιού γεγονός που δεν μειώνει το precision της κατηγορίας μη-ορισμών σε μεγάλο βαθμό. Επιλέγοντας για κάτω κατώφλι την τιμή 0.32 η επιθυμητή αναλογία αποκαθίσταται.

Η χρήση του αλγορίθμου ομοιότητας και των παραπάνω κατωφλίων οδηγεί σε ένα σώμα εκπαίδευσης με ένα ποσοστό λαθών, δηλαδή παράθυρα που είναι σημειωμένα πως ανήκουν σε λάθος κατηγορία, αφού για καμία από τις δύο κατηγορίες η ακρίβεια δεν είναι 100%. Τα λάθη αυτά εισάγουν θόρυβο στα δεδομένα εκπαίδευσης αλλά οι αλγόριθμοι μηχανικής μάθησης καταφέρνουν συχνά να μην επηρεάζονται από θόρυβο αυτής της μορφής.

4.4 Ιδιότητες μηχανικής μάθησης

Ο ταξινομητής που κατατάσσει τα παράθυρα σε ορισμούς και μη-ορισμούς χρησιμοποιεί τον αλγόριθμο μάθησης της ενότητας 3.2 και 222 ιδιότητες από τις οποίες οι 22 επιλέχθηκαν χειρωνακτικά ενώ οι υπόλοιπες παράγονται με αυτόματο τρόπο από σώμα εκπαίδευσης. Οι συγκεκριμένες ιδιότητες επιλέχθηκαν, διότι πειράματα σε δεδομένα των διαγωνισμών TREC (Μηλιαράκη 2003) έδειξαν ότι συμβάλλουν στην καλύτερη δυνατή εκπαίδευση του συστήματος.

4.4.1 Χειρωνακτικά επιλεγμένες ιδιότητες

1. Η **κατάταξη (ranking) του κειμένου** που από το οποίο προέρχεται το παράθυρο, η οποία στο σύστημα που δημιουργήθηκε ταυτίζεται με τη σειρά που επέστρεψε η μηχανή

αναζήτησης το έγγραφο Χρησιμοποιείται ως ιδιότητα διότι συνήθως οι ζητούμενοι ορισμοί βρίσκονται στα κορυφαία κείμενα που επιστρέφονται και όχι στα τελευταία.

2. Η **θέση του παραθύρου** μέσα στο έγγραφο. Η ιδιότητα δείχνει αν πρόκειται για το πρώτο, δεύτερο, κ.ο.κ. παράθυρο του εγγράφου. Χρησιμοποιείται διότι συνήθως τα παράθυρα ορισμού εντοπίζονται στην αρχή των κειμένων.
3. Το **πλήθος των κοινών λέξεων του παραθύρου.** Η ιδιότητα αυτή προκύπτει λόγω της παρατήρησης ότι τα παράθυρα ορισμού ενός όρου συνήθως έχουν κοινές λέξεις. Επομένως, τα παράθυρα που έχουν υψηλό αριθμό λέξεων από τις κοινές έχουν μεγαλύτερη πιθανότητα να είναι ορισμοί. Για να έχει νόημα αυτή η ιδιότητα θα πρέπει να αφαιρεθούν οι πολύ συχνές λέξεις της αγγλικής. Για το σκοπό αυτό χρησιμοποιείται πάλι η λίστα συχνών λέξεων (stop-words) του BNC της ενότητας 4.3.3 Η λίστα των κοινών λέξεων που δημιουργήθηκε για κάθε όρο είχε μέγεθος 20.
4. Η φράση **"such <…> as *όρος*"**

   Παράδειγμα : *"such antibiotics as amoxicillin"*
5. Η φράση **"*όρος* and other <…>"**

   Παράδειγμα : *"broken bones and other injuries"*
6. Η φράση **"*όρος* or other <…>"**

   Παράδειγμα : *"cats or other animals"*
7. Η φράση **"especially *όρος*"**

   Παράδειγμα : *"some plastics especially Teflon"*
8. Η φράση **"including *όρος*"**

   Παράδειγμα : *"some amphibians including frog"*
9. **Παρενθέσεις μετά** τον όρο

   Παράδειγμα : *"sodium chloride (salt)"*
10. **Παρενθέσεις πριν** τον όρο

    Παράδειγμα : *"(Vitamin B1) thiamine"*
11. Η φράση **"*όρος* is a"**

    Ακριβέστερα αναζητείται η πληρέστερη φράση της μορφής "*όρος* is/are/was/were a/an/the <…>"

    Παράδειγμα : *"Galileo was a great astronomer"*
12. **Κόμμα μετά** τον όρο

    Παράδειγμα : *"amoxicillin, an antibiotic"*
13. Η φράση **"*όρος* which is/was/are/were <…>"**

    Παράδειγμα : *"tsunami which is a giant wave"*
14. Η φράση **"*όρος* like <…>"**

    Παράδειγμα : *"antibiotics like amoxicillin"*
15. Η φράση **"*όρος* , <…> , is/was/are/were"**

Παράδειγμα : *"amphibians, like frogs, are animals that can live both on land and in water"*

16. Η φράση **"*όρος* or <…>"**

    Παράδειγμα : *"autism or some other type of disorder"*

17. Ένα από τα ρήματα **"can", "refer", "have"** μετά τον όρο **(3 ιδιότητες)**

    Παράδειγμα : *"Amphibians can live both on land and in water"*

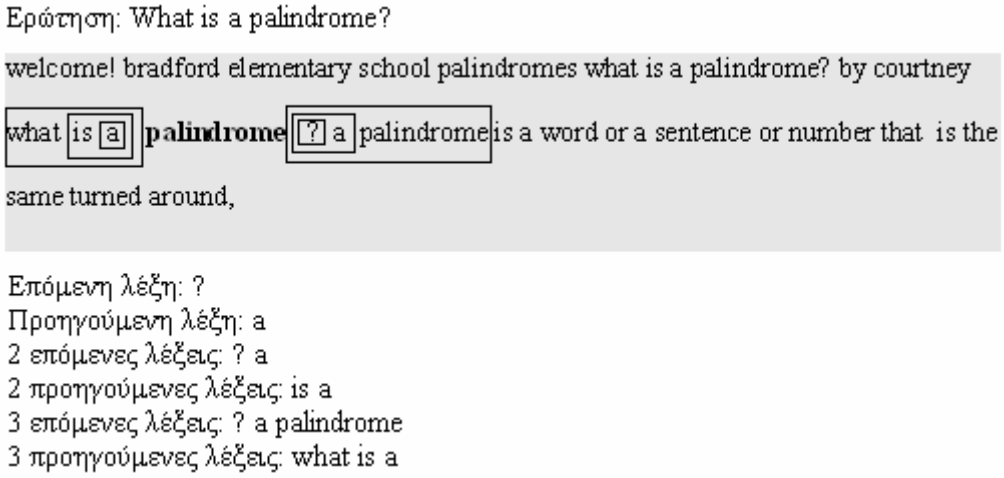
18. Ένα από τα ρήματα **"called", "known","defined"** πριν τον όρο **(3 ιδιότητες)**

    Παράδειγμα : *" The giant wave known as tsunami"*

4.4.2 Επιλογή ιδιοτήτων με την χρήση της ακρίβειας

Οι υπόλοιπες 200 ιδιότητες επιλέγονται αυτόματα από το σώμα εκπαίδευσης που κατασκευάζεται με τη χρήση του αλγόριθμου ομοιότητας και των κατωφλιών, χρησιμοποιώντας σαν μέτρο επιλογής την ακρίβεια της κατηγορίας των ορισμών (precision1). Η διαδικασία επιλογής αυτών των ιδιοτήτων είναι η εξής:

- Δημιουργείται μια κενή λίστα φράσεων.
- Για κάθε παράθυρο του σώματος εκπαίδευσης εξάγονται οι φράσεις 1) επόμενη λέξη του όρου (όρου που πρέπει να οριστεί) 2) 2 επόμενες λέξεις του όρου 3) 3 επόμενες λέξεις του όρου 4) προηγούμενη λέξη του όρου 5) 2 προηγούμενες λέξεις του όρου 6)3 προηγούμενες λέξεις του όρου (Σχήμα 3).
- Για κάθε φράση που εξάγεται ελέγχεται αν ήδη υπάρχει στην λίστα. Αν υπάρχει τότε απλά αυξάνεται ένας μετρητής εμφανίσεων της στο σώμα εκπαίδευσης κατά 1, αλλιώς εισάγεται στην λίστα και ο μετρητής της αρχικοποιείται στην τιμή 1.



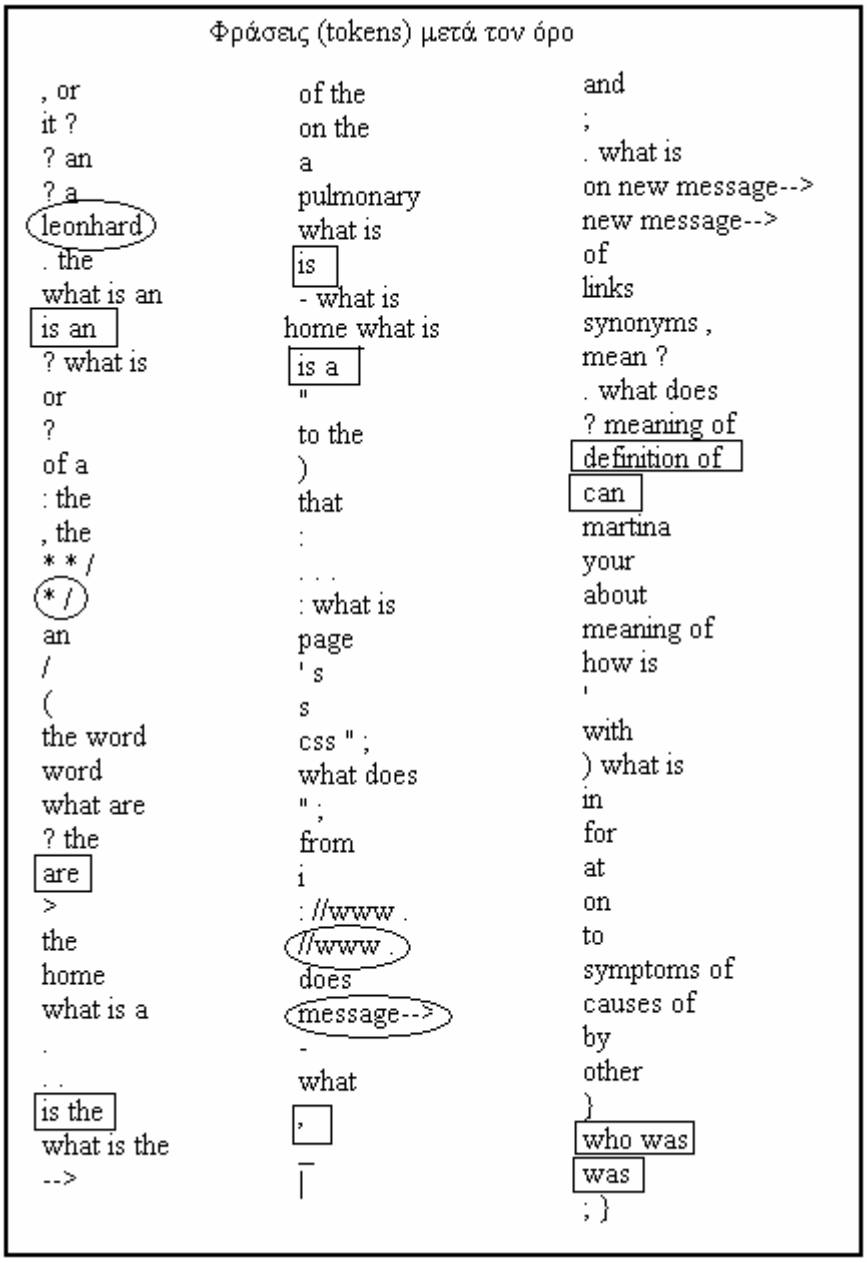
Σχήμα 3

Όταν ολοκληρωθεί η διαδικασία κάθε φράση μήκους 1 2 ή 3 λέξεων που εμφανίζεται στο σώμα εκπαίδευσης ακριβώς πριν ή μετά από κάποιο όρο θα βρίσκεται στη λίστα συνοδευόμενη από τον μετρητή εμφανίσεων της στο σώμα εκπαίδευσης. Με αυτό τον τρόπο όμως δημιουργείται μια τεράστια σε μέγεθος λίστα ακόμη και για σχετικά μικρά σώματα εκπαίδευσης. Για αυτό το λόγο διαγράφονται από την λίστα οι φράσεις που έχουν αριθμό εμφανίσεων μικρότερο από μια τιμή. Ύστερα υπολογίζεται η ακρίβεια (precision) για κάθε φράση και επιλέγονται ως ιδιότητες αυτές που θα έχουν υψηλότερες τιμές ακρίβειας.

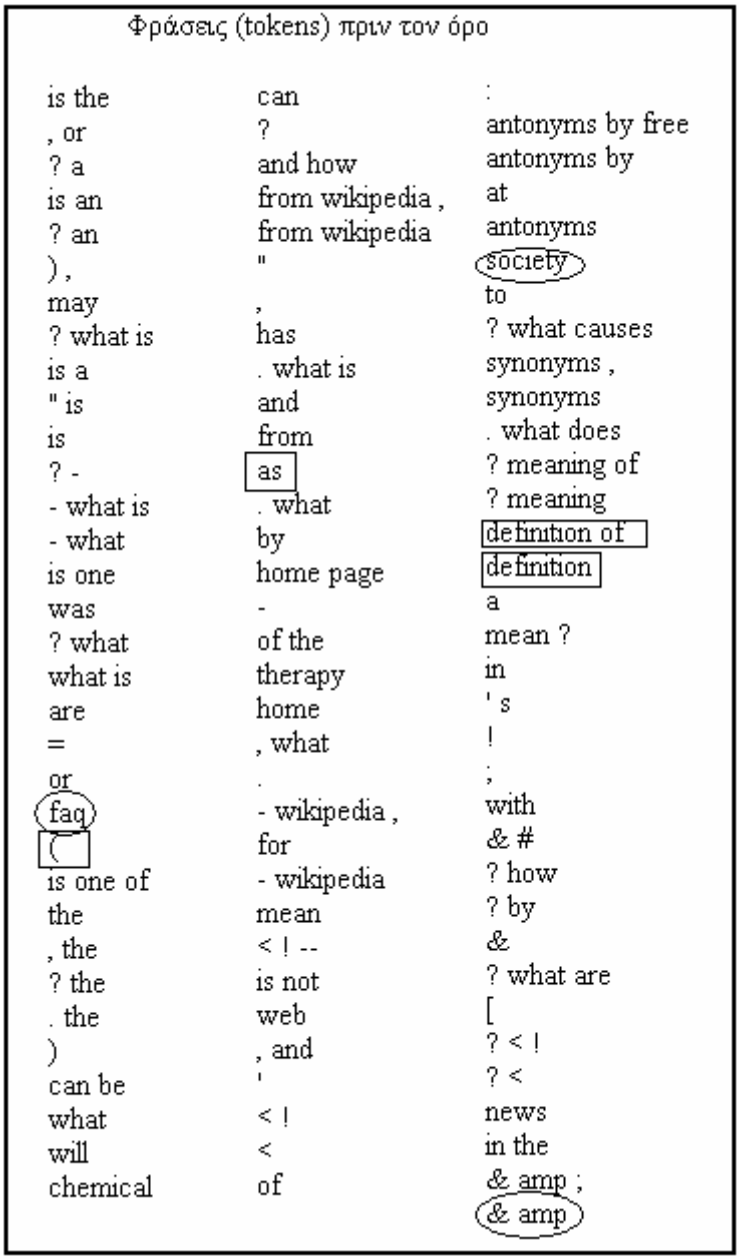
Η ακρίβεια μιας φράσης υπολογίζεται ως ο λόγος των παραθύρων ορισμού που περιέχουν τη φράση εμφανιζόμενη ακριβώς πριν η μετά από τον όρο δια τα συνολικά παράθυρα (ορισμού ή μη-ορισμού) που την περιέχουν εμφανιζόμενη ακριβώς πριν η μετά από τον όρο. Η ακρίβεια ενός token μας δίνει ένα μέτρο βεβαιότητας για το αν ένα παράθυρο που το περιέχει είναι και ορισμός.

Χρησιμοποιώντας 480 ερωτήσεις οι οποίες συλλέχθηκαν από μία online εγκυκλοπαίδεια (http://www.encyclopedia.com) και εξάγοντας από τα κείμενα που επέστρεψε μια μηχανή αναζήτησης για αυτές τα αντίστοιχα παράθυρα δημιουργούνται 7200 παράθυρα κειμένου, τα οποία κατατάσσονται με την μέθοδο με τα κατώφλια. Από αυτό το σώμα κειμένων προέκυψαν οι 200 ιδιότητες με την υψηλότερη ακρίβεια. Οι ιδιότητες αυτές παρουσιάζονται στο σχήμα 4 και 5. Με τετράγωνο πλαίσιο έχουν σημειωθεί μερικές φράσεις (tokens) που φαίνεται να είναι καλές ιδιότητες, δηλαδή αποτελούν ισχυρή ένδειξη ότι ένα παράθυρο κειμένου είναι και παράθυρο ορισμού και με στρογγυλό πλαίσιο μερικές που φαίνεται να έχουν προκύψει τυχαία. Μερικές από τις καλές ιδιότητες που σημειώθηκαν έχουν ήδη εντοπιστεί αφού υπάρχουν στην λίστα των χειρωνακτικά επιλεγμένων ιδιοτήτων. Τέτοιες είναι οι <is>, <is an>, <is a>, <are>, <can>, <,>, <is the> μετά τον όρο και οι <definition of>, <as>, <(> πριν τον όρο.

Στην συγκεκριμένη περίπτωση, ο αριθμός των ιδιοτήτων που προκύπτουν είναι ίσες για τις δύο κατηγορίες (φράσεις πριν και μετά τον όρο) χωρίς αυτό να σημαίνει ότι κάθε φορά επιλέγονται οι 100 ιδιότητες από κάθε κατηγορία με το υψηλότερο precision. Επιλέγονται συνολικά 200 ιδιότητες, που ήταν ο βέλτιστος αριθμός ιδιοτήτων στα πειράματα της προηγούμενης εργασίας (Μηλιαράκη 2003)


Φράσεις (tokens) μετά τον όρο
, or
it ?
? an
? a
leonhard
. the
what is an
is an
? what is
or
?
of a
: the
, the
* * /
* /
an
/
(
the word
word
what are
? the
are
>
the
home
what is a
.
. .
is the
what is the
-->
of the
on the
a
pulmonary
what is
is
- what is
home what is
is a
"
to the
)
that
:
. . .
: what is
page
' s
s
css " ;
what does
" ;
from
i
: //www .
//www .
does
message-->
-
what
,
_
|
and
;
. what is
on new message-->
new message-->
of
links
synonyms ,
mean ?
. what does
? meaning of
definition of
can
martina
your
about
meaning of
how is
'
with
) what is
in
for
at
on
to
symptoms of
causes of
by
other
}
who was
was
; }

Σχήμα 4


Φράσεις (tokens) πριν τον όρο
is the
, or
? a
is an
? an
) ,
may
? what is
is a
" is
is
? -
- what is
- what
is one
was
? what
what is
are
=
or
faq
(
is one of
the
, the
? the
. the
)
can be
what
will
chemical
can
?
and how
from wikipedia ,
from wikipedia
"
,
has
. what is
and
from
as
. what
by
home page
-
of the
therapy
home
, what
.
- wikipedia ,
for
- wikipedia
mean
< ! --
is not
web
, and
'
< !
<
of
:
antonyms by free
antonyms by
at
antonyms
society
to
? what causes
synonyms ,
synonyms
. what does
? meaning of
? meaning
definition of
definition
a
mean ?
in
' s
!
;
with
& #
? how
? by
&
? what are
[
? < !
? <
news
in the
& amp ;
& amp

Σχήμα 5

5. Πειράματα-Αξιολόγηση συστήματος

5.1 Μέτρα αξιολόγησης

Για την αξιολόγηση ενός συστήματος ερωταποκρίσεων συνήθως χρησιμοποιούνται 2 δείκτες, ο αριθμός των ερωτήσεων για τις οποίες το σύστημα επέστρεψε σωστή απάντηση και η μέση αντίστροφη κατάταξη (**M**ean **R**eciprocal **R**ank). Τα μέτρα αυτά χρησιμοποιούνται και για την αξιολόγηση συστημάτων που συμμετέχουν στους διαγωνισμούς TREC.

Ένα σύστημα ερωταποκρίσεων όπως ήδη έχει αναφερθεί αξιολογεί όλες τις υποψήφιες απαντήσεις (στην περίπτωσή μας παράθυρα κειμένου) μιας ερώτησης. Ακολουθώντας τους κανονισμούς που ίσχυαν για τις ερωτήσεις ορισμού στους διαγωνισμούς TREC 2000 και 2001, το σύστημα αυτής της εργασίας επιστρέφει στο χρήστη μια ταξινομημένη λίστα με τα 5 παράθυρα (υποψήφιες απαντήσεις) που έχουν την μεγαλύτερη πιθανότητα να είναι πράγματι ορισμοί του όρου της ερώτησης. Αν έστω μια από αυτές που επεστράφησαν αποτελεί αποδεκτό ορισμό τότε θεωρείται ότι σύστημα επέστρεψε σωστή απάντηση.

Όμως, είναι επιθυμητό για ένα σύστημα όχι μόνο να απαντάει με σωστό τρόπο μεγάλο αριθμό ερωτήσεων, αλλά να κατατάσσει τις καλύτερες απαντήσεις στις κορυφαίες θέσεις της λίστας των υποψήφιων απαντήσεων που επιστρέφει. Χρησιμοποιώντας ένα σύστημα με αυτό το χαρακτηριστικό είναι πολύ πιο εύκολο να εντοπιστούν οι επιθυμητές απαντήσεις καθώς απαιτείται λιγότερος χρόνος και λιγότερο κόπος. Η μέση αντίστροφη κατάταξη (MRR) είναι ένας δείκτης που μετρά το κατά πόσο ένα σύστημα επιστρέφει σωστές απαντήσεις στις υψηλές θέσεις της λίστας των υποψήφιων απαντήσεων. Για τον υπολογισμό της μέσης αντίστροφης κατάταξης κάθε ερώτηση της συλλογής αξιολόγησης λαμβάνει ένα βάρος, το οποίο είναι ίσο με 1 δια την σειρά της πρώτης σωστής απάντησης (1-5). Αν στις επιστρεφόμενες απαντήσεις δεν υπάρχει σωστή, το βάρος της ερώτησης λαμβάνει την τιμή 0. Τα βάρη όλων των ερωτήσεων αθροίζονται και το τελικό αποτέλεσμα διαιρείται με το συνολικό αριθμό των απαντήσεων οπότε προκύπτει η αριθμητική τιμή του δείκτη MRR.

5.2 Διασταυρωμένη επικύρωση

Για την αξιολόγηση ενός συστήματος μηχανικής μάθησης είναι απαραίτητος ο διαχωρισμός των διαθέσιμων δεδομένων σε δεδομένα εκπαίδευσης και δεδομένα αξιολόγησης. Σε περιπτώσεις που τα διαθέσιμα δεδομένα είναι σχετικά λίγα, χρησιμοποιείται μια διαδικασία που ονομάζεται δεκαπλή διασταυρωμένη επικύρωση (10-fold cross

validation). Τα διαθέσιμα δεδομένα χωρίζονται σε 10 κομμάτια το καθένα με τον ίδιο αριθμό παραδειγμάτων. Ύστερα γίνονται 10 επαναλήψεις, όπου κάθε φορά χρησιμοποιείται για αξιολόγηση ένα διαφορετικό από τα 10 κομμάτια και για εκπαίδευση τα υπόλοιπα 9 και υπολογίζεται ο μέσος όρος των μέτρων αξιολόγησης. Στην περίπτωση μας, το κάθε κομμάτι αποτελείται από το 1/10 των διαθέσιμων ερωτήσεων και τα αντίστοιχα παράθυρα των ιστοσελίδων που επέστρεψε η μηχανή αναζήτησης. Σε κάθε επανάληψη ο ταξινομητής εκπαιδεύεται στα παράθυρα των 9 κομματιών και αξιολογείται στο δέκατο κομμάτι. Σε κάθε επανάληψη υπολογίζονται, ο αριθμός των ερωτήσεων για τις οποίες το σύστημα επιστρέφει σωστή απάντηση, αλλά και η μέση αντίστροφη κατάταξη και υπολογίζεται τελικά ο μέσος όρος αυτών των μέτρων στις 10 επαναλήψεις. Η τεχνική αυτή έχει το πλεονέκτημα ότι χρησιμοποιεί όλο το σύνολο δεδομένων για εκπαίδευση και για αξιολόγηση, αλλά και ότι τα πειράματα επαναλαμβάνονται πολλές φορές με διαφορετικά κάθε φορά δεδομένα εκπαίδευσης και αξιολόγησης, γεγονός που συμβάλλει στην αξιολόγηση του συστήματος με αντικειμενικότερο τρόπο.

5.3 Αξιολόγηση νέου συστήματος ερωταποκρίσεων-Πειράματα

Για να αξιολογηθεί το νέο σύστημα ερωταποκρίσεων που δημιουργήθηκε επιχειρήθηκε η σύγκριση του με ένα ήδη υπάρχον. Πιο συγκεκριμένα, δημιουργήθηκαν 2 συστήματα, ένα εκπαιδευμένο με δεδομένα TREC και πρότυπα απαντήσεων, σε γενικές γραμμές όπως στην προηγούμενη εργασία (Μηλιαράκη 2003) και ένα άλλο εκπαιδευμένο με παράθυρα κειμένου που προέκυψαν από ιστοσελίδες με την μέθοδο των κατωφλίων. Ονομάζουμε σύστημα TREC και σύστημα WEB τα δύο συστήματα αντίστοιχα. Και τα δύο συστήματα χρησιμοποίησαν την ίδια μηχανή SVM, τις ίδιες χειρωνακτικά επιλεγμένες ιδιότητες και τον ίδιο αυτόματο τρόπο επιλογής 200 επιπλέον ιδιοτήτων που περιγράφηκε στην ενότητα 4. Και στα δύο συστήματα χρησιμοποιήθηκαν για κάθε ερώτηση τα 10 πρώτα έγγραφα που επιστρέφονται από τη μηχανή αναζήτησης και τα 5 πρώτα παράθυρα κειμένου κάθε εγγράφου. Αυτό είναι απαραίτητο, γιατί με ανάλογο τρόπο έχουν προκύψει τα παράθυρα με τα οποία διερευνήθηκαν οι τιμές των κατωφλίων. Τα δεδομένα κάθε συστήματος (ερωτήσεις και δεδομένα TREC, ερωτήσεις και δεδομένα ιστοσελίδων) χωρίστηκαν σε 10 μέρη, όπως στην περίπτωση της 10-πλής διασταυρωμένης επικύρωσης. Σε κάθε επανάληψη, κάθε σύστημα εκπαιδευόταν στα 9/10 των δεδομένων του, ενώ η αξιολόγηση γινόταν στο 1/10 των δεδομένων TREC (κοινά δεδομένα αξιολόγησης και για τα δύο συστήματα). Φροντίσαμε σε κάθε μία από τις 10 επαναλήψεις η αναλογία παραθύρων ορισμού/μη-ορισμού στα δεδομένα εκπαίδευσης του συστήματος WEB να είναι η ίδια με την αντίστοιχη αναλογία που υπήρχε στα δεδομένα εκπαίδευσης του συστήματος TREC Αυτό

επιτυγχάνεται με την προσαρμογή των κατωφλίων. Το άνω κατώφλι παραμένει σταθερό για λόγους που έχουν ήδη αναφερθεί (ενότητα 4.3.5) και προσαρμόζεται το κάτω κατώφλι έτσι ώστε να επιτευχθεί η επιθυμητή αναλογία. Επίσης έχει γίνει η απαραίτητη επιλογή του αριθμού ερωτήσεων των ερωτήσεων στα δεδομένα εκπαίδευσης του συστήματος WEB έτσι ώστε κάθε φορά τα δεδομένα εκπαίδευσης του ενός και του άλλου συστήματος να έχουν τον ίδιο αριθμό διανυσμάτων εκπαίδευσης.

Αφού έγινε η σύγκριση των δύο συστημάτων με ίσο αριθμό διανυσμάτων εκπαίδευσης επιχειρήθηκε η αύξηση των δεδομένων εκπαίδευσης του συστήματος WEB(αύξηση ερωτήσεων που συλλέχθηκαν από την online εκυκλοπαίδεια). Αυτό είναι εύκολο, αφού τα δεδομένα εκπαίδευσης του συστήματος WEB παράγονται αυτόματα. Τα αποτελέσματα που προέκυψαν παρουσιάζονται στον Πίνακα 2.

Σύστημα Ερωταποκρίσεων	**Αριθμός διανυσμάτων εκπαίδευσης**	**Ποσοστό ερωτήσεων με σωστή απάντηση**	**MRR**
TREC	3369	62%	0,51
WEB	3397	50%	0,30
WEB	5392	47%	0,35
WEB	7999	48%	0,35
Baseline	-	50%	-

Πίνακας 2

Από την παρατήρηση των αποτελεσμάτων προκύπτει ότι:

1. Για τον ίδιο αριθμό διανυσμάτων εκπαίδευσης το σύστημα TREC απαντάει μεγαλύτερο ποσοστό ερωτήσεων και ο δείκτης MRR του είναι σημαντικά υψηλότερος από τον αντίστοιχο δείκτη για το WEB σύστημα.
2. Κατά την αύξηση των δεδομένων εκπαίδευσης του συστήματος WEB ο αριθμός των ερωτήσεων με σωστή απάντηση δεν αυξάνεται αλλά μειώνεται (2-3%) παράλληλα όμως ο δείκτης MRR αυξάνεται κατά 0.05.

Για την καλύτερη αξιολόγηση των δύο συστημάτων, οι επιδόσεις τους συγκρίθηκαν και με τις επιδόσεις ενός αφελούς συστήματος (baseline) το οποίο απλά επιλέγει για κάθε ερώτηση με τυχαίο τρόπο 5 υποψήφιες απαντήσεις, χωρίς να χρησιμοποιεί μηχανική μάθηση. Είναι προφανές ότι ένα τέτοιο σύστημα δεν θα επιστρέφει κάθε φορά τις ίδιες απαντήσεις, διότι αυτές επιλέγονται τυχαία. Επομένως, κάθε φορά που θα επιχειρείται να υπολογιστεί ο δείκτης MRR και το ποσοστό ερωτήσεων που απαντά το σύστημα θα προκύπτουν διαφορετικές τιμές. Για να αποφευχθεί αυτό και να εξαχθούν αντικειμενικότερες τιμές υπολογίστηκαν οι αναμενόμενες τιμές αυτών των μεγεθών. Χρησιμοποιήσαμε τα ίδια δεδομένα αξιολόγησης με εκείνα που είχαν χρησιμοποιηθεί στην αξιολόγηση των

συστημάτων TREC και WEB. Για κάθε ερώτηση αξιολόγησης υπολογίστηκε η αναλογία των παραθύρων ορισμού/μη-ορισμού και ύστερα η πιθανότητα το αφελές σύστημα να απαντήσει τη συγκεκριμένη ερώτηση. Η πιθανότητα αυτή προκύπτει ως εξής:

P(το σύστημα να απαντάει την ερώτηση) =

= P(το σύστημα να επιστρέφει τουλάχιστον 1 παράθυρο ορισμού) =

= 1- P(το σύστημα να επιστρέφει μόνο παράθυρα μη ορισμού) =

= 1- P(να επιλεχθεί 1 παράθυρο μη-ορισμού)^5 =

= 1 – (αριθμός παραθύρων μη-ορισμού/συνολικός αριθμός παραθύρων)^5.

Αθροίζοντας τις τιμές των πιθανοτήτων που προκύπτουν και διαιρώντας με τον αριθμό των ερωτήσεων προκύπτει το αναμενόμενο ποσοστό των ερωτήσεων που το σύστημα απαντά. Στην συγκεκριμένη περίπτωση το ποσοστό που προκύπτεί είναι 50% δηλαδή περίπου ίσο με τα αντίστοιχα ποσοστά που επιτυγχάνει το σύστημα WEB web σύστημα για τα 3 διαφορετικά μεγέθη δεδομένων εκπαίδευσης. Το γεγονός αυτό δείχνει ότι το WEB σύστημα δεν έχει ικανοποιητικές επιδόσεις, αφού επιτυγχάνει τα ίδια ποσοστά επιτυχίας με αυτά που θα είχε ένα σύστημα το οποίο επιστρέφει με τυχαίο τρόπο απαντήσεις. Αντίθετα το TREC σύστημα επιτυγχάνει ικανοποιητικά αποτελέσματα τα οποία ξεπερνούν σε σημαντικό βαθμό τα αποτελέσματα που θα είχε ένα σύστημα που επιστρέφει με τυχαίο τρόπο απαντήσεις στο χρήστη.

Παρότι το σύστημα WEB φαίνεται να αποτυγχάνει υπάρχουν κάποια σημεία τα οποία δείχνουν ότι η αξιολόγηση ενδεχομένως το αδικεί. Όπως ήδη έχει αναφερθεί οι ιδιότητες μηχανικής μάθησης οι οποίες χρησιμοποιούνται προκύπτουν σε μεγάλο ποσοστό αυτόματα από το σώμα εκπαίδευσης. Στο πείραμα που διενεργήθηκε για την σύγκριση των 2 συστημάτων, σε κάθε επανάληψη της διασταυρωμένης επικύρωσης εξάγονται από το σώμα εκπαίδευσης οι 200 από τις 220 ιδιότητες οι οποίες θα χρησιμοποιηθούν για την δημιουργία των διανυσμάτων εκπαίδευσης και αξιολόγησης. Οι ιδιότητες που επιλέγουν με αυτόν τον τρόπο τα δύο συστήματα φαίνεται να διαφέρουν σε μεγάλο βαθμό. Αυτό πιθανότατα να αδικεί το σύστημα WEB, διότι αξιολογείται σε δεδομένα των διαγωνισμών TREC με ιδιότητες οι οποίες έχουν προκύψει από ιστοσελίδες, που φαίνεται να διαφέρουν αρκετά από τα κείμενα των διαγωνισμών TREC. Όπως έχουν δείξει πειράματα σε δεδομένα TREC (Μηλιαράκη 2003)οι ιδιότητες που προκύπτουν από το σώμα εκπαίδευσης βοηθούν ένα σύστημα να ανεβάσει αρκετά τα ποσοστά επιτυχίας του σε σχέση με ένα σύστημα το οποίο χρησιμοποιεί μόνο τις 22 χειρωνακτικά επιλεγμένες ιδιότητες. Επομένως είναι απαραίτητο όχι μόνο να χρησιμοποιούνται αλλά και να είναι αντιπροσωπευτικές των δεδομένων αξιολόγησης. Δεδομένου ότι τα κείμενα των TREC διαφέρουν από τα κείμενα των ιστοσελίδων, είναι πιθανό οι ιδιότητες που ανακαλύπτει το σύστημα TREC στις ιστοσελίδες εκπαίδευσης να μην το βοηθούν κατά την αξιολόγησή του σε κείμενα TREC

Για όλους τους παραπάνω λόγους διενεργήθηκε ένα ακόμη πείραμα στο οποίο το σύστημα WEB αξιολογήθηκε αυτή τη φορά σε δεδομένα ιστοσελίδων. Πιο συγκεκριμένα δημιουργήθηκε μια συλλογή εκπαίδευσης μεγέθους 7200 διανυσμάτων και χρησιμοποιήθηκαν 81 ερωτήσεις αξιολόγησης (οι οποίες προέρχονται και αυτές από την online εγκυκλοπαίδεια που έχει αναφερθεί). Μετά την εκπαίδευσή του, το σύστημα επέστρεψε για κάθε μία ερώτηση το παράθυρο με την μεγαλύτερη πιθανότητα να είναι ορισμός. Το κάθε επιστρεφόμενο παράθυρο αξιολογήθηκε χειρωνακτικά χωρίς την χρήση προτύπων απαντήσεων. Επιτρέψαμε στο σύστημα να επιστρέφει μόνο ένα παράθυρο ανά ερώτηση, γιατί διαφορετικά η χειρωνακτική αξιολόγηση θα ήταν εξαιρετικά χρονοβόρα. Δεν θα ήταν δυνατόν η αξιολόγηση να γίνει με διασταυρωμένη επικύρωση, γιατί οι ερωτήσεις της συλλογής των ιστοσελίδων δεν συνοδεύονται από πρότυπα απαντήσεων, αντίθετα από τις ερωτήσεις των διαγωνισμών TREC. Επίσης, δεν θα μπορούσε να χρησιμοποιηθεί κατά την αξιολόγηση ο αλγόριθμος ομοιότητας/κατωφλίων αντί για πρότυπα απαντήσεων, γιατί δεν είναι αρκετά ακριβής ώστε να χρησιμοποιηθεί κατά την αξιολόγηση ούτε δίνει απαντήσεις για παράθυρα των οποίων οι βαθμοί ομοιότητας βρίσκονται μεταξύ των δύο κατωφλίων. Τα αποτελέσματα έδειξαν ότι το 56% (46/81) των ερωτήσεων απαντήθηκαν σωστά. Επίσης για κάθε ερώτηση επιλέχθηκε με τυχαίο τρόπο (baseline) ένα παράθυρο κειμένου και αξιολογήθηκε χειρωνακτικά. Το ποσοστό των ερωτήσεων που απαντήθηκαν ήταν 17% (14/81). Το ποσοστό που πέτυχε το σύστημα WEB σε σχέση με αυτό πού πέτυχε το τυχαίο σύστημα είναι κατά πολύ μεγαλύτερο, γεγονός που δείχνει ότι πιθανότατα η αξιολόγηση στα TREC δεδομένα αδίκησε το WEB σύστημα.

Επίσης, διενεργήθηκε ένα ακόμη πείραμα. Δημιουργήθηκε μια συλλογή εκπαίδευσης, η οποία χρησιμοποιεί όλα τα δεδομένα TREC (μεγέθους 3800 διανυσμάτων) και χρησιμοποιήθηκαν για αξιολόγηση οι 81 ερωτήσεις του παραπάνω πειράματος. Όπως και στο προηγούμενο πείραμα επιστράφηκε μόνο ένα παράθυρο ανά ερώτηση, το οποίο αξιολογήθηκε χειρωνακτικά. Το σύστημα απάντησε το 20% (17/81) των ερωτήσεων ποσοστό σημαντικά χαμηλότερο από το ποσοστό που είχε επιτευχθεί από το σύστημα WEB στις ίδιες ερωτήσεις. Επιπλέον το ποσοστό του TREC συστήματος προσεγγίζει το ποσοστό του τυχαίου συστήματος (baseline) στα ίδια δεδομένα αξιολόγησης. Πιο συγκεκριμένα το TREC σύστημα απλά καταφέρνει να απαντήσει ελάχιστες ερωτήσεις (3%) παραπάνω από το τυχαίο σύστημα (baseline). Αυτό δείχνει ότι και το TREC σύστημα αδικείται όταν αξιολογείται σε WEB δεδομένα, όπως αδικείται το WEB σύστημα όταν αξιολογείται σε TREC δεδομένα.

6. Συμπεράσματα και μελλοντικές προσεγγίσεις

Τα πειράματα που έγιναν στα πλαίσια της εργασίας δείχνουν ότι:

- Η νέα μέθοδος κατασκευής παραθύρων εκπαίδευσης μπορεί να δημιουργήσει ένα μεγάλο σε αριθμό διανυσμάτων σώμα εκπαίδευσης με αυτόματο τρόπο. Το μόνο που απαιτείται είναι η συλλογή μεγάλου αριθμού ερωτήσεων ορισμού.
- Η νέα μέθοδος κατασκευής παραθύρων εκπαίδευσης μπορεί να συμβάλλει στη δημιουργία ενός συστήματος ερωταποκρίσεων (για ερωτήσεις ορισμού) με αρκετά καλή επίδοση.
- Τα συστήματα ερωταποκρίσεων που εξετάστηκαν σε αυτή την εργασία (TREC σύστημα, WEB σύστημα) έχουν ικανοποιητικές επιδόσεις όταν αξιολογούνται σε δεδομένα που έχουν την ίδια προέλευση με αυτά που εκπαιδεύτηκαν. Σε αντίθετη περίπτωση τα αποτελέσματα που επιτυγχάνουν δεν είναι ικανοποιητικά, καθώς ακόμη και ένα τυχαίο σύστημα (baseline) έχει σχεδόν την ίδια επίδοση.

Ακόμη, είναι φανερό πως μπορεί για το WEB σύστημα να διερευνηθούν ζητήματα που αφορούν τον τρόπο επιλογής των ιδιοτήτων και τον αριθμό αυτών. Στην εργασία αυτή σε όλα τα πειράματα που έγιναν χρησιμοποιήθηκαν 200 ιδιότητες επιλεγμένες με κριτήριο την ακρίβεια της κατηγορίας των ορισμών. Αυτό έγινε διότι πειράματα που έγιναν σε προηγούμενη εργασία (Μηλιαράκη 2003) έδειξαν ότι με το συγκεκριμένο αριθμό ιδιοτήτων και με το συγκεκριμένο τρόπο επιλογής αυτών το σύστημα TREC επιτυγχάνει τα καλύτερα αποτελέσματα. Όμως αυτό δεν είναι απαραίτητο να συμβαίνει και για το WEB σύστημα, επομένως μπορεί να διερευνηθεί για ποιό αριθμό ιδιοτήτων (π.χ 100, 200, 300) και με ποιό τρόπο επιλογής αυτών (ακρίβεια, ανάκληση, πληροφοριακό κέρδος) επιτυγχάνει το καλύτερο αποτέλεσμα στα ίδια δεδομένα αξιολόγησης.

Επίσης για τις παραμέτρους που θα προκύψουν από τα προτεινόμενα πειράματα μπορεί να υπολογισθεί το ποσοστό των σωστών απαντήσεων και το MRR για διαφορετικά μεγέθη σωμάτων εκπαίδευσης για τις ίδιες ερωτήσεις αξιολόγησης. Η αξιολόγηση θα γίνεται βέβαια χειρωνακτικά, διότι όπως έδειξαν τα πειράματα η χρησιμοποίηση των TREC δεδομένων και των προτύπων απαντήσεων αδικούν το σύστημα WEB. Είναι σημαντικό να διερευνηθεί η απόδοση του συστήματος όταν αυξάνεται διαδοχικά το μέγεθος των δεδομένων εκπαίδευσης.

Αναφορές

Voorhees Ellen M., *"Overview of the TREC2001 Question Answering Track"*,National Institute of Standards and Technology, 2001

Voorhees Ellen M., *"Overview of the TREC-9 Question Answering Track"*, National Institute of Standards and Technology, 2000

Voorhees Ellen M., *"The TREC-8 Question Answering Track Report"*, National Institute of Standards and Technology, 1999

Schölkopf Bernhard, Alex Smola, "*Learning with Kernels*", MIT Press, Cambridge, MA, 2002

Simmons R. F., *"Answering English questions by computer : A survey"*, Communications Association for Computing Machinery (ACM), 8(1): 53-70, 1965

Mitchell, T.M. "*Machine Learning",* McGraw-Hill International Editions, 1997

Prager John, Brown Eric, Radev Dragomir R., Krzysztof Czuba, *"One Search Engine or Two for Question-Answering"*, TREC9 QA-Track Notebook Paper, NIST, 2000

Prager John, Dragomir Radev, Brown Eric, Coden Anni, Samn Valerie, *"The use of Predictive Annotation for Question-Answering in TREC8"*, in Proceedings of TREC8, 1999

Prager John, Dragomir Radev, Krzysztof Czuba, *"Answering What-Is Questions by Virtual Annotation"*, 2001

Prager John, Jennifer Chu-Carroll, Krzysztof Czuba, *"Use of Wordnet Hypernyms for answering What-Is Questions"*, 2002

Hirschamn L., Gaizauskas R. *"Natural language question answering: the view from here",* Cambridge University Press, 2001

Alin Dobra, Support Vector Machine Learning CS478 Machine Learning May 2, 2000

S. Miliaraki and I. Androutsopoulos, "Learning to Identify Single-Snippet Answers to Definition Questions". Proceedings of the *20th International Conference on Computational Linguistics* (COLING 2004), Geneva, Switzerland, pp. 1360-1366, 2004.

Reiter E., Dale R., "*Building Natural Language Generation Systems",* Cambridge University Press, 2000

Radev Dragomir R., Prager John, Samn Valerie, *"Ranking suspected answers to natural language questions using predictive annotation",*1999

Ιστοσελίδες

[1] http://svm.sdsc.edu/svm-overview.html.
Overview of the SVM algorithm