ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΔΙΔΑΚΤΙΚΗΣ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ & ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ:

«Διδακτική της Τεχνολογίας και Ψηφιακών Συστημάτων» -

«ΔΙΚΤΥΟΚΕΝΤΡΙΚΑ ΣΥΣΤΗΜΑΤΑ»

Εξόρυξη γνώσης από ειδησεογραφικούς ιστοχώρους και σύνδεσή τους με πραγματικά γεγονότα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Γεώργιου Τζούμη

Επιβλέπων: Βασιλακόπουλος Γεώργιος

Καθηγητής Παν. Πειραιώς

Αθήνα, Αύγουστος 2011

Περιεχόμενα

Περιεχόμενα 2
1 Εισαγωγή 5
1.1 Το Κίνητρο μελέτης 5
1.2 Αντικείμενο 7
1.3 Διάρθρωση της εργασίας 8
2 Η χρήση των Υπηρεσιών Ιστού (Web Services) στον Παγκόσμιο ιστό 10
2.1 Ανακάλυψη Υπηρεσιών Ιστού 12
2.2 XML 14
3 Μέθοδοι εξόρυξης κειμένων 18
3.1 Επεξεργασία φυσικής γλώσσας 19
3.1.1 Σημειολογική επεξεργασία 20
3.1.2 Προτασιακή επεξεργασία 20
3.1.3 Μορφολογική ανάλυση 21
3.2 Κατηγοριοποίηση 21
3.2.1 Διανυσματικά μοντέλα 22
4 Αποθήκες δεδομένων 26
4.1 Αρχιτεκτονική 27
4.2 Μεταφορά Δεδομένων από τις Πηγές 30
4.3 Αποθήκη Δεδομένων 30
4.4 Μοντέλα Ανάλυσης Δεδομένων 32
4.4.1 Πολυδιάστατα Μοντέλα Δεδομένων 33
4.4.2 Πράξεις στους υπερκύβους 34
5 Περιγραφή συστήματος 36
5.1 Υποσυστήματα 36
5.1.1 Υπηρεσίες Ιστού 37
5.1.2 Υποσύστημα εύρεσης Blogs(crawler) 39
5.1.3 Υποσύστημα εξόρυξης κειμένων 42
5.1.4 Υποσύστημα Αποθήκης & Ανάλυσης Δεδομένων 45
5.2 Βάση Δεδομένων 47
5.3 Αποθήκη δεδομένων ιστολόγιων 49

6 Παραδείγματα 52
6.1 Γραφήματα 52
6.1.1 Δημοσιεύσεις ανά κατηγορία 52
6.1.2 Σχόλια-Δημοσιεύσεις ανά ημέρα 53
6.1.3 Σχόλια ανά ώρα σε περίοδο μιας μέρας 54
6.1.4 Σχόλια-Δημοσιεύσεις ανά ημέρα 54
6.1.5 Σχόλια-Δημοσιεύσεις ανά ημέρα 55
6.1.6 Πλήθος σχολίων από την ώρα δημοσίευσης 56
6.1.7 Πλήθος σχόλιων και δημοσιεύσεων ανά κατηγορία 56
6.1.8 Συσχέτιση δημοσιεύσεων με πραγματικά γεγονότα 57
6.1.9 Συσχέτιση σχολίων με πραγματικά γεγονότα 57
7 Σύνοψη και Συμπεράσματα 59
8 Βιβλιογραφία 61

Περίληψη

Στις μέρες μας, το διαδίκτυο είναι η μεγαλύτερη πηγή πληροφορίας και οι περισσότεροι από εμάς το χρησιμοποιούμε για την εύρεση στοιχείων που μας ενδιαφέρουν. Εκτός από τα ειδησεογραφικά portals που κατέχουν την περισσότερη επισκεψιμότητα, υπάρχουν και τα προσωπικά ιστολόγια (blogs) που ανάλογα με το θέμα που ασχολούνται αρχίζουν να αναπτύσσονται. Η πληθώρα ιστολόγιων που υπάρχει, καταστεί αδύνατη την παρακολούθηση των θεμάτων που απασχολούν τους χρήστες. Στόχος αυτής της διπλωματικής εργασίας είναι η υλοποίηση ενός συστήματος που βασίζεται σε Web Services και παρακολουθεί την ελληνική κοινότητα ιστολόγιων και παρουσιάζει τα σημαντικότερα θέματα κατηγοριοποιημένα ανά θεματική ενότητα και ταξινομημένα με βάση το πλήθος των σχολίων που ακολουθούν τις δημοσιεύσεις. Αυτή η διαδικασία εξόρυξης γνώσης είναι ιδιαίτερα χρήσιμη για τους χρήστες που θέλουν να παρακολουθήσουν ένα θέμα καθώς και για την καταγραφή των κοινωνικών δικτύων που αναπτύσσονται μέσα από τα ιστολόγια.

Λέξεις Κλειδιά

ενημέρωση, blogs, ιστολόγια, ειδήσεις, bloggers, εξόρυξη δεδομένων, κατηγοριοποίηση

1 Εισαγωγή

1.1 Το Κίνητρο μελέτης

Το πρόβλημα που καλείται να λύσει η συγκεκριμένη πτυχιακή είναι μια ανάγκη για άμεση ενημέρωση που προέκυψε στη σημερινή πραγματικότητα που ο ψηφιακός κόσμος κατακλύζεται από πληροφορίες που προέρχονται όχι μόνο από δημοσιογραφικούς ιστοχώρους αλλά και από προσωπικά διαδικτυακά ημερολόγια (blogs) που παρουσιάζουν τρομερή άνθηση στις μέρες μας. Βασικός συντελεστής της ηλεκτρονικής δημοσίευσης αποτελεί το διαδίκτυο. Η διασύνδεση όλων των υπολογιστικών συστημάτων ανά τον κόσμο έδωσε την δυνατότητα για άμεση διάδοση της πληροφορίας χωρίς χρονικούς ή γεωγραφικούς περιορισμούς. Το ηλεκτρονικό περιοδικό, το ηλεκτρονικό βιβλίο ή ακόμη και η απευθείας δημοσίευση στο διαδίκτυο αναπτύσσονται όλο και περισσότερο, απαιτώντας παράλληλα νέες τεχνολογίες και εργαλεία για την διαχείριση αυτής της μορφής την πληροφορία. Η πληροφορία που περιέχουν τα blogs έχει ενδιαφέρον διότι όπως και οι ειδήσεις έτσι και οι δημοσιεύσεις (posts) ενός συγγραφέα (blogger) διαμορφώνουν την κοινή γνώμη μέσα από ένα διαφορετικό πρίσμα.

Έτσι, παρακολουθώντας τις δημοσιεύσεις υπάρχει η δυνατότητα να παρουσιαστούν μέσω μιας διαδραστικής διεπαφής, λέξεις, καθώς και έννοιες-κλειδιά, σχετικά με την πρόσφατη ειδησεογραφία.

Οι ιστοσελίδες με γενικό περιεχόμενο έχουν αντικατασταθεί με ειδησεογραφικά Portal που διαθέτουν κατηγορίες και άρθρα για οποιοδήποτε θέμα. Τα blogs αντίστοιχα έχουν μια κατεύθυνση ως προς το περιεχόμενο αλλά δεν έχουν απολύτως υιοθετήσει την

δεοντολογία των δημοσιογράφων με αποτέλεσμα να δημοσιεύονται ειδήσεις που δημιουργούν πληθώρα σχολίων.

Η δημιουργία μιας τέτοιας εφαρμογής θα δώσει τη δυνατότητα σε επιστήμονες του χώρου της πληροφορικής αλλά και της διαφήμισης να αποκτήσουν εικόνα σχετικά με τα θέματα που συζητούνται στα blogs και που θα ενδιέφεραν να γίνουν αντικείμενο διαφήμισης ή παρακολούθησης. Οι πληροφορίες θα είναι προσβάσιμες από το κοινό έτσι ώστε όταν κάποιος θελήσει να δει σχετικά με ποια θέματα ασχολείται ένα blog να έχει αυτή τη δυνατότητα.

Οι συνομιλίες που καταγράφονται στα blogs είναι αντικείμενο προς μελέτη και επεξεργασία. Μέσα από την εφαρμογή, μια εταιρεία θα μπορεί να αναζητήσει πληροφορίες σχετικά με το πότε αναφέρθηκε κάποιος σε αυτή ή ακόμα και να παρακολουθήσει τον τρόπο που εξελίσσεται μια συζήτηση γύρω από ένα θέμα που την ενδιαφέρει.

Πολλές υποθέσεις σχετικά με σκάνδαλα αλλά και ζητήματα άμεσης επέμβασης έχουν προκύψει μέσα από blogs. Η δυνατότητα να παρέχεται ένα σύστημα ειδοποίησης με ηλεκτρονικό ταχυδρομείο λέξεων-κλειδιών σχετικά με ευαίσθητα θέματα θα είχε μεγάλο ενδιαφέρον από επιχειρήσεις. Επίσης, πολλοί καταναλωτές έχουν την ευκαιρία να σχολιάσουν αρνητικά ή θετικά προϊόντα και να διαμορφώσουν άποψη σε μια μερίδα εν δυνάμει πελατών που κάνουν έρευνα. Μέσα από την εφαρμογή θα μπορεί κάποιος να ενημερωθεί και να αξιοποιήσει τις πληροφορίες προς όφελός του.

Οι δουλειές που έχουν πραγματοποιηθεί αντίστοιχα στον χώρο είναι σχετικά λίγες και ασχολούνται, κυρίως, με την παρακολούθηση των ειδήσεων μέσα από

ειδησεογραφικά portals και όχι μέσα από blogs. Η ιδιαιτερότητα της εργασίας έχει να κάνει με το γεγονός ότι τα blogs είναι σελίδες με περιεχόμενο που αφ' ενός, δεν ασχολείται αποκλειστικά με ένα θέμα και αφ' ετέρου, η δημοσίευση (post) μπορεί να σχολιαστεί από άλλους. Χρειάζεται, λοιπόν, συνεχής παρακολούθηση και ενημέρωση.

Οι δυσκολίες που υπάρχουν σε ένα ολοκληρωμένο σύστημα καταγραφής των στοιχείων και παρακολούθησης των blogs είναι ποικίλες και διαφορετικής φύσης. Η εύρεση των κειμένων είναι δύσκολη διότι κάθε πάροχος blog (blogspot, wordpress) διαθέτει διαφορετικά πρότυπα εμφάνισης των δημοσιεύσεων και ο τρόπος συλλογής τους πρέπει να είναι όσο το δυνατόν ανεξάρτητος του προτύπου. Επίσης, η κατηγοριοποίηση σε κείμενα που περιέχουν περιορισμένο αριθμό λέξεων είναι πολύ δύσκολη και απαιτεί μεγάλο όγκο κειμένων για την εκπαίδευση του κατηγοριοποιητή. Επιπλέον, η διαχείριση κειμένων που δημοσιεύονται σε καθημερινή βάση από χιλιάδες bloggers απαιτεί σωστή οργάνωση και σχεδίαση της βάσης δεδομένων που θα φυλάσσονται.

1.2 Αντικείμενο

Η εργασία έχει σκοπό να ακολουθήσει όλα τα βήματα της διαδικασίας εξόρυξης γνώσης από τα ελληνικά διαδικτυακά ιστολόγια, έτσι ώστε να καταγραφεί η κίνηση στους ιστοτόπους και να παρουσιαστούν στατιστικά στοιχεία σχετικά με την ελληνική πραγματικότητα. Επίσης, θα δημιουργηθεί ένα σύστημα από προγράμματα το οποίο μπορεί να παρέχει υπηρεσίες πληροφόρησης για ιστολόγια και τις κατηγορίες που ανήκουν, ώστε να βασιστούν αναλύσεις περιεχομένου και μέθοδοι ανάκτησης δεδομένων σε μελλοντικές εργασίες.

1.3 Διάρθρωση της εργασίας

Η δομή της διπλωματικής εργασίας είναι η ακόλουθη:

Στο πρώτο κεφάλαιο αναφέρονται σε εισαγωγικό επίπεδο οι βασικές έννοιες που θα απασχολήσουν την όλη εργασία και οι ανάγκες που αυτή εξυπηρετεί.

Στο δεύτερο κεφάλαιο αναλύεται σε βάθος η έννοια της υπηρεσιοστρεφούς αρχιτεκτονικής καθώς και η δομή της. Ακόμα, εξηγείται η σημαντική ευελιξία που παρέχουν οι υπηρεσίες ιστού(Web Services) στην διαμόρφωση του Διαδικτύου στις μέρες μας με αναφορές στις τεχνολογίες που τις απαρτίζουν.

Στο τρίτο κεφάλαιο παρουσιάζεται η έννοια της εξόρυξης των δεδομένων και γνώσης καθώς και όλα τα στάδια επεξεργασίας φυσικής γλώσσας που λαμβάνουν μέρος κατά τη διαδικασία κανονικοποίησης των δεδομένων.

Στο τέταρτο κεφάλαιο, περιγράφονται αναλυτικά οι έννοιες της αποθήκης δεδομένων και οι μέθοδοι που χρησιμοποιούνται σε περιπτώσεις ανάλυσης και επεξεργασίας των δεδομένων.

Στο πέμπτο κεφάλαιο αναλύεται η λειτουργική αρχιτεκτονική του συστήματος που αναπτύχθηκε παρουσιάζοντας όλα τα διαγράμματα περιπτώσεων χρήσης που αφορούν το σύστημα. Επιπλέον, παρουσιάζεται ένα σενάριο χρήσης με μερικά ενδεικτικά στιγμιότυπα από το σύστημα.

Στο έκτο κεφάλαιο αναλύεται η τεχνική αρχιτεκτονική του συστήματος που αναπτύχθηκε και παρουσιάζονται βασικές τεχνολογίες-κλειδιά που χρησιμοποιήθηκαν.

Τέλος, στο έβδομο κεφάλαιο καταγράφονται τα βασικά συμπεράσματα που προέκυψαν από την εκπόνηση της παρούσας διπλωματικής εργασίας.

2 Η χρήση των Υπηρεσιών Ιστού (Web Services) στον Παγκόσμιο ιστό

Ο σχεδιασμός και η ανάπτυξη ολοκληρωμένων συστημάτων εφαρμογών καθώς επίσης και οι αυξανόμενες προσδοκίες των εκάστοτε χρηστών έχουν αλλάξει δραματικά τα τελευταία χρόνια, κυρίως όσον αφορά την τεχνολογική εξέλιξη στην ανάπτυξη προσβάσιμων επιχειρησιακών συστημάτων.

Έτσι η επιτυχία που σημείωσε η εμφάνιση του διαδικτύου, οι διαδικτυακές υπηρεσίες και γενικότερα η εμφάνιση τεχνολογιών σημασιολογικού ιστού, ήταν αρκετή για να πείσει σχεδόν τους περισσότερους ότι η χρήση των τεχνολογιών αυτών είναι καθοριστικής σημασίας για τον σχεδιασμό, την ανάπτυξη αλλά και την διασύνδεση συστημάτων εφαρμογών. Κατά συνέπεια, αυτό είχε σαν αποτέλεσμα να κάνει την εμφάνισή της δειλά– δειλά η υπηρεσιοστρεφής αρχιτεκτονική (service oriented architecture) η οποία συνθέτει όλες τις παραπάνω τεχνολογίες και καθορίζει ένα σύνολο κατάλληλων συνιστωσών με στόχο αφενός την διασύνδεση ετερογενών συστημάτων και αφετέρου το σχεδιασμό και την ανάπτυξη μιας νέας μορφής προσβάσιμων επιχειρησιακών συστημάτων.

Η Υπηρεσιοστρεφής Αρχιτεκτονική (ΥΑ) αποτελεί μια πολύ ενδιαφέρουσα αρχιτεκτονική ιδέα. Βασίζεται στη συνεργασία σύνθετων στοιχείων δυναμικά μέσω ανακάλυψης (discovery), σύνθεσης (composition) και διαλειτουργικότητας (interoperability). Οι Υπηρεσίες Ιστού είναι μία τεχνολογία που χρησιμοποιείται για να υλοποιηθεί η ΥΑ και έχει σχεδόν καταφέρει να αποτελεί την καλύτερη επιλογή.

Οι Υπηρεσίες Ιστού είναι λογισμικό που διαθέτει κάποια χαρακτηριστικά που του επιτρέπουν να θεωρείται τμήμα υπηρεσιοστρεφούς αρχιτεκτονικής, όπως επικοινωνία μέσω μηνυμάτων, ανακάλυψη, πύλες, ρόλοι και εναρμόνιση.

Ο πιο ευρέως διαδεδομένος και επιτυχημένος τύπος υπηρεσιών είναι οι διαδικτυακές υπηρεσίες XML Services γνωστές ως Web Services. Αυτός ο τύπος υπηρεσίας έχει δύο βασικές προαπαιτήσεις:

- επικοινωνεί μέσω πρωτοκόλλου Internet (κυρίως HTTP)
- στέλνει και δέχεται δεδομένα μέσα από XML αρχεία

Η ευρεία αποδοχή του μοντέλου των Web Services είχε ως αποτέλεσμα την ανάγκη πρόσθετων τεχνολογιών που βασίζονται σε αυτές και τη δημιουργία καινούργιων προτύπων. Έτσι η "παραγωγή" τέτοιων υπηρεσιών απαιτεί:

- τη περιγραφή της υπηρεσίας αναλυτικά, τουλάχιστον με ένα WSDL έγγραφο
- τη δυνατότητα μεταφοράς ενός XML εγγράφου χρησιμοποιώντας SOAP μέσω HTTP.

Επιπρόσθετα είναι συνηθισμένο μια υπηρεσία να λειτουργεί και ως πελάτης (client/requestor) και ως πάροχος (provider) υπηρεσίας. Αναλόγως λοιπόν με τη δραστηριότητα της υπηρεσίας κάθε στιγμή, μετατρέπεται από το ένα στο άλλο.

2.1 Ανακάλυψη Υπηρεσιών Ιστού

Η μεταφορά μηνυμάτων μεταξύ των φορέων γίνεται μέσω του πρωτοκόλλου Simple Object Access Protocol (SOAP) για να οριστεί η μορφή των μηνυμάτων δημιουργήθηκε η γλώσσα Web Services Description Language (WSDL). Η WSDL είναι μία γλώσσα μορφής XML η οποία περιγράφει δικτυακές υπηρεσίες.

Μία WSDL περιγραφή αποτελείται από δύο τμήματα.

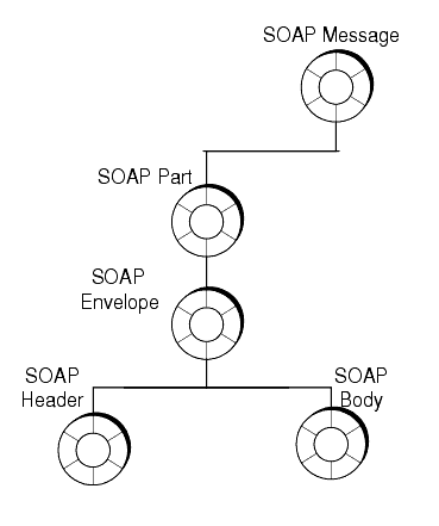
- Abstract
    - Περιγραφή των τύπων δεδομένων που αποστέλλονται σε μια αίτηση προς την υπηρεσία.
    - Η δομή των μηνυμάτων που αποστέλλονται προς την υπηρεσία.
    - Οι επιμέρους λειτουργίες που παρέχονται από την υπηρεσία.
    - Η οργάνωση σχετικών λειτουργιών σε επιμέρους διεπαφές που προσφέρονται από την υπηρεσία.
- Concrete: περιγραφή μιας ή περισσότερων υπηρεσιών που προσφέρουν στιγμιότυπα των διαπροσωπειών που ορίζονται στο abstract κομμάτι.
    - Η συσχέτιση των διεπαφών με κάποια συγκεκριμένα πρωτόκολλα επικοινωνίας (SOAP, …).
    - Τα στιγμιότυπα της κάθε διεπαφής, τα οποία προσφέρονται σε συγκεκριμένες διευθύνσεις (URIs).

- Τα στιγμιότυπα των υπηρεσιών, ορισμένα σαν συλλογές από στιγμιότυπα διαπροσωπειών.

Κάθε υπηρεσία ιστού δημιουργεί μία WSDL διεπαφή η οποία παρέχει πληροφορίες σε κάθε ενδιαφερόμενο που θέλει να χρησιμοποιήσει την υπηρεσία. Οι υπηρεσίες ιστού μπορούν να επικοινωνήσουν με νέες υπηρεσίες χωρίς την παραμικρή αλλαγή. Η ανακάλυψη των νέων υπηρεσιών ιστού γίνεται μέσω μιας υπηρεσίας που λέγεται Universal Description, Discovery and Integration (UDDI). Ο στόχος του UDDI είναι να παρέχει την αναγκαία υποδομή για την περιγραφή και αναζήτηση των υπηρεσιών ιστού, να ορίζει τον τρόπο καταχώρησής τους σε μητρώο και να παρέχεται εύκολα μέσω XML προτύπου.

Αν και αρχικά είχε θεωρηθεί ως η τεχνολογία που θα γεφυρώσει το κενό μεταξύ ανόμοιων πλατφόρμων βασισμένων σε RPC επικοινωνία, το SOAP έχει εξελιχθεί στο πλέον ευρέως χρησιμοποιούμενο πρότυπο επικοινωνίας για τη χρήση XML Υπηρεσιών Διαδικτύου (Web Services). Μετά από αυτή την κατάσταση γίνεται πολλές φορές η παράφραση του ακρωνύμιου από Simple Object Access Protocol σε Service Oriented Architecture (or Application) Protocol.

Το πρωτόκολλο SOAP διαμορφώνει ένα πρότυπο μήνυμα που αποτελείται από ένα XML έγγραφο ικανό να περιγράψει δεδομένα όπως RPC κλήσεις κ.α. Το μήνυμα αυτό μεταφέρεται μεταξύ των υπηρεσιών και των εφαρμογών, χρησιμοποιώντας κυρίως το HTTP πρωτόκολλο δικτύου. Με τον τρόπο αυτό ολοκληρώνεται το πλαίσιο λειτουργίας και επικοινωνίας στην SOA δομή, αφού με την βοήθεια της περιγραφής WSDL είναι εφικτή η επικοινωνία και συνεργασία οποιωνδήποτε υπηρεσιών στο δίκτυο.



Εικόνα 1

2.2 XML

Σε ένα κόσμο όπου οι πληροφορίες παρέχονται μέσω του παγκόσμιου διαδικτύου, τα έγγραφα πρέπει να είναι εύκολα προσβάσιμα, μεταφέρσιμα και ευέλικτα. Πρέπει επίσης να είναι ανεξάρτητα οποιουδήποτε συστήματος και περιεχομένου. Οι γενικευμένες γλώσσες έχουν τέτοια χαρακτηριστικά , παρέχοντας στα έγγραφα αυτά μια δυνατότητα η οποία δεν υπάρχει σε άλλες γλώσσες περιγραφής εγγράφων. Η HTML είναι προβληματική και περιοριστική γλώσσα. Η XML έλυσε πολλά από τα προβλήματα που αντιμετώπισαν οι σχεδιαστές του web και είναι υπεύθυνη για την XHTML, μια ανασχεδιασμένη HTML. Θα χρησιμοποιείται για πολλά χρόνια επειδή προσφέρει αποτελεσματικές και δυναμικές πολυμεσικές λύσεις.

Η XML αποτελεί σήμερα το πρότυπο για την αποθήκευση δεδομένων που ανταλλάσσονται μεταξύ των εφαρμογών χάρη στα ακόλουθα χαρακτηριστικά που παρουσιάζει :

- Υποστηρίζει ανεξαρτησία από τα δεδομένα και διαχωρίζει τα περιεχόμενα από τον τρόπο εμφάνισής τους και τον χειρισμό τους, οπότε διευκολύνεται η λεκτική ανάλυσή τους (parsing).

- Διατίθενται έτοιμοι τρόποι σύνδεσης των κειμένων XML με τα πλέον σύγχρονα προγραμματιστικά περιβάλλοντα, όπως το Document Object Model (DOM) και το Simple API for XML (SAX).

- Είναι επεκτάσιμη και ανεξάρτητη από πλατφόρμες, γεγονός που την καθιστά απρόσβλητη σε τεχνολογικές αλλαγές.

- Τα έγγραφα XML είναι αναγνώσιμα από ανθρώπους και μηχανές και παρότι δεν προορίζονται για ανάγνωση προσφέρουν αυτή τη δυνατότητα στο χρήστη εάν κριθεί αναγκαίο.

- Είναι πλήρως συμβατή με Unicode, οπότε μπορεί να χειριστεί την πληροφορία που έχει γραφεί σε οποιαδήποτε ανθρώπινη γλώσσα. Παράλληλα, υποστηρίζει διεθνείς και τοπικές προσαρμογές.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ARTICLES SYSTEM
"D:\Projects\Clients\XML\Contents\Temp\articlelist.dtd">
<?xml-stylesheet type="text/xsl"
href="D:\xmltohtml.xslt"?>
<ARTICLES>
    <ARTICLE>
        <ARTICLEDATA>
            <TITLE>XML Demystified</TITLE>
            <AUTHOR>Jaidev</AUTHOR>
        </ARTICLEDATA>
    </ARTICLE>
    <ARTICLE>
        <ARTICLEDATA>
            <TITLE>XSLT Demystified</TITLE>
            <AUTHOR>X S Cel Tea </AUTHOR>
        </ARTICLEDATA>
    </ARTICLE>
    <ARTICLE>
        <ARTICLEDATA>
            <TITLE>C# Demystified</TITLE>
            <AUTHOR>Aleksey N</AUTHOR>
        </ARTICLEDATA>
    </ARTICLE>
</ARTICLES>
```

Εικόνα 2

Η XML σχεδιάστηκε να ικανοποιήσει πολλές ανάγκες δίνοντας στα έγγραφα ένα μεγαλύτερο επίπεδο προσαρμοστικότητας στο στυλ και τη δομή από αυτό που υπήρχε παλαιότερα στην HTML. Η XML προσφέρει στους σχεδιαστές της HTML τη δυνατότητα να προσθέτουν περισσότερα στοιχεία στη γλώσσα. Δεν αναφέρεται μονάχα στους σχεδιαστές του web αλλά σε οποιονδήποτε ασχολείται με εκδόσεις.

Στην πραγματικότητα, η XML είναι markup γλώσσα για έγγραφα που περιέχουν δομημένες πληροφορίες. Markup γλώσσα είναι ένας μηχανισμός που καθορίζει δομές σε ένα έγγραφο. Οι δομημένες πληροφορίες περιλαμβάνουν περιεχόμενο και κάποιες διευκρινίσεις για το ρόλο που παίζει το περιεχόμενο σχεδόν όλα τα έγγραφα έχουν την ίδια δομή.

Η XML είναι κάτι περισσότερο από markup language είναι metalanguage, δηλαδή μια γλώσσα που χρησιμοποιείται για να καθορίσει νέες markup γλώσσες. Η XML συμπληρώνει και δεν αντικαθιστά την HTML, ενώ η HTML χρησιμοποιείται στη

διατύπωση και την εμφάνιση των δεδομένων η XML αναπαριστά τη συναφή έννοια των δεδομένων. Στην HTML τα tags είναι προκαθορισμένα ενώ η XML παρέχει τη δυνατότητα να καθορίζουν οι χρήστες τα tags και τις δομημένες μεταξύ τους σχέσεις.

Τα XML έγγραφα δεν είναι πολύπλοκα αλλά απλά και πολύ αποτελεσματικά. Το διδακτικό υλικό της well-formed XML αναλύει τη δημιουργία των XML εγγράφων, η οποία είναι κατά κάποιο τρόπο ίδια με την HTML καθώς επιτρέπει τη μη δομημένη δημιουργία εγγράφου. Η valid XML είναι πιο σύνθετη. Απαιτεί την ύπαρξη ενός Document Type Definition πριν να γραφεί το έγγραφο αλλά παρέχει μια γενική δομή με βάση την οποία τη δημιουργούμε.

Η γλώσσα προγραμματισμού XML περιγράφει μια κατηγορία πληροφοριών (data objects) που καλούνται XML έγγραφα (documents) καθώς επίσης περιγράφει τμηματικά τη συμπεριφορά των προγραμμάτων που τα επεξεργάζονται.

Τα XML έγγραφα αποτελούνται από μονάδες αποθήκευσης που καλούνται entities (οντότητες), οι οποίες περιέχουν πληροφορίες αναλυμένες ή μη. Οι αναλυμένες πληροφορίες αποτελούνται από χαρακτήρες (characters) οι οποίοι συνθέτουν character data και άλλοι οι οποίοι συνθέτουν markup. Η μορφή markup κωδικοποιεί την περιγραφή της τελικής αποθήκευσης του εγγράφου καθώς και τη λογική δομή.

3 Μέθοδοι εξόρυξης κειμένων

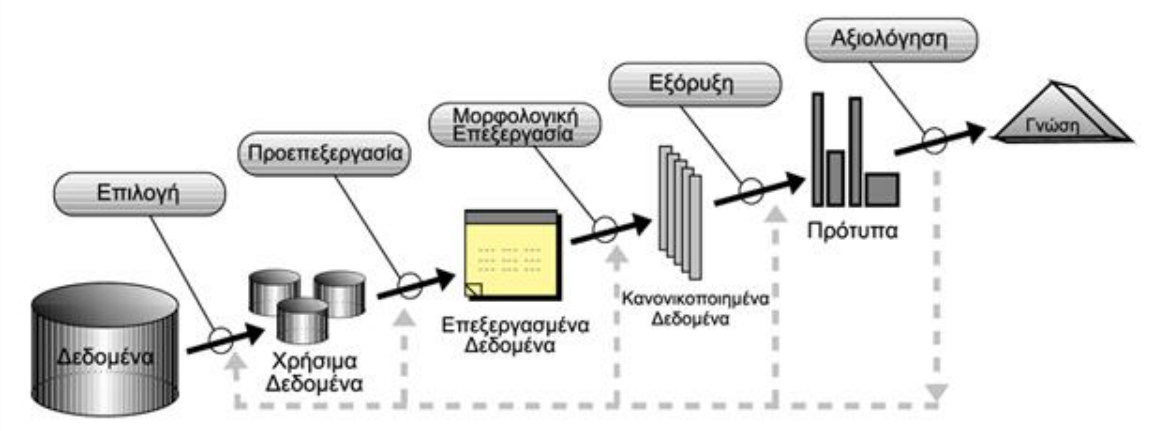
Εξόρυξη κειμένων είναι η ανακάλυψη, με αυτόματο τρόπο, πληροφορίας που βρίσκεται σε γραπτές πηγές. Το πιο σημαντικό είναι να συνδυαστούν τα αποτελέσματα της εξαγομένης πληροφορίας έτσι ώστε να δημιουργηθούν δεδομένα που θα επεξεργαστούν για επιπλέον ανάπτυξη.

Η εξόρυξη κειμένων είναι διαφορετική από την αναζήτηση λέξεων στο διαδίκτυο, διότι ο χρήστης εκεί, συνήθως, ψάχνει για κάτι ήδη γραμμένο από κάποιον άλλο. Το πρόβλημα στην διαδικτυακή αναζήτηση είναι να απομονώσεις τα στοιχεία που σε ενδιαφέρουν έτσι ώστε να βρεις τις σχετικές πληροφορίες. Αντιθέτως, στην εξόρυξη κειμένων ο στόχος είναι να ανακαλύψεις άγνωστες μέχρι τώρα πληροφορίες, κάτι που κανένας δεν έχει γράψει.

Μεγάλες βάσεις δεδομένων χρησιμοποιούνται στην εξόρυξη δεδομένων, για να ανακαλυφθούν ενδιαφέροντα μοτίβα πληροφορίας, αντίστοιχα ένα πεδίο της εξόρυξης δεδομένων είναι και η εξόρυξη κειμένων. Ένα τυπικό παράδειγμα για την εξόρυξη δεδομένων είναι η χρήση στοιχείων των πωλήσεων ενός καταστήματος σχετικά με συνδυασμούς προϊόντων που πωλούνται παρέα έτσι ώστε να είναι στο ίδιο ράφι ή να δοθούν κουπόνια εκπτώσεων. Οι αναλυτές επίσης μπορούν μέσα από τεράστιες βάσεις δεδομένων με στοιχεία πιστωτικών καρτών να ανακαλύψουν απάτες σε συναλλαγές.

Η διαφορά μεταξύ εξόρυξης δεδομένων και εξόρυξης κειμένων είναι ότι στα κείμενα τα μοτίβα προέρχονται κυρίως από κείμενο φυσικής γλώσσας και όχι από δομημένες βάσεις δεδομένων από αριθμούς και στοιχεία. Οι βάσεις δεδομένων είναι σχεδιασμένες για να επικοινωνούν με προγράμματα, αντιθέτως το κείμενο είναι για να το διαβάζουν

άνθρωποι. Αυτή τη στιγμή δεν υπάρχουν προγράμματα που να «διαβάζουν» κείμενο και δεν θα υπάρχουν σίγουρα για το προσεχές μέλλον. Οι περισσότεροι ερευνητές πιστεύουν ότι πρέπει να προσομοιωθεί πλήρως η διαδικασία πως δουλεύει το μυαλό και μετά να υλοποιηθούν προγράμματα που διαβάζουν όπως οι άνθρωποι.



Εικόνα 3

Ωστόσο, η επεξεργασία φυσικής γλώσσας είναι ένας κλάδος που αρχίζει να αναπτύσσεται, δείχνει να έχει καταφέρει να δημιουργήσει μικρά κομμάτια κειμένου, που κυρίως χρησιμεύουν ως αυτόματη σύνοψη σε μεγάλα κείμενα. Επίσης, υπάρχουν προγράμματα που με σχετική ακρίβεια αντλούν πληροφορίες από κείμενα με κάποια δομή. Για παράδειγμα, προγράμματα διαβάζουν συνόψεις και βρίσκουν κύρια ονόματα, διευθύνσεις κ.α. σε ποσοστό 80 τοις εκατό.

3.1 Επεξεργασία φυσικής γλώσσας

Η επεξεργασία φυσικής γλώσσας έδωσε τη δυνατότητα να κατανοηθεί η δομή της κάθε γλώσσας και οι κανόνες που την διέπουν με έναν πιο ξεκάθαρο τρόπο, βασισμένο πάνω σε υπολογιστικές αρχές και εργαλεία.

Σημαντικά πλεονεκτήματα που μας δίνει η επεξεργασία φυσικής γλώσσας, είναι η μελέτη πραγματικών κειμένων που έχουν γραφτεί με φυσική ροή και όχι δομημένα όπως είναι παράδειγμα σε ένα βιβλίο. Επίσης, βοήθησε στην αποδόμηση των γλωσσών σε κανόνες και με την χρήση δυνατών, ανεκτών σε λάθη αλγορίθμων να εξαχθεί πληροφορία. Ακόμη, η βασική προσέγγιση με βάσει κανόνες εξελίχθηκε με αποτέλεσμα να χρησιμοποιηθούν στατιστικές μέθοδοι και μηχανική μάθηση με πολύ καλύτερα αποτελέσματα. Τέλος, έδωσε την ώθηση για εξέλιξη του σημασιολογικού ιστού βάζοντας τα θεμέλια για την εξέλιξη του διαδικτύου και την έλευση του Web 2.0.

3.1.1 Σημειολογική επεξεργασία

Η πρώτη φάση της επεξεργασίας των κειμένων προς ανάλυση είναι αυτή της αναγνώρισης των στοιχείων μιας πρότασης και των λέξεων που την απαρτίζουν. Πριν ξεκινήσει η αναγνώριση των λέξεων πρέπει να απαλλαχθεί το κείμενο από ανεπιθύμητες λέξεις όπως διαφημίσεις ή μενού και στοιχεία πλοήγησης των ιστοσελίδων.

Στη συνέχεια, γίνεται η τμηματοποίηση της πρότασης σε λέξεις, αριθμούς, κενά, σημεία στίξης και χαρακτήρες που δεν ανήκουν σε κάποια κατηγορία όπως σύμβολα νομίσματος, εισαγωγικά και άλλα. Η προσπάθεια εντοπισμού των σωστών στοιχείων στην ελληνική γλώσσα είναι ένα δύσκολο κομμάτι λόγω της ιδιομορφίας της και των ιδιωματισμών.

3.1.2 Προτασιακή επεξεργασία

Ο εντοπισμός των προτάσεων είναι επίσης πολύ σημαντικός για κάθε επεξεργασία που θα ακολουθήσει στη συνέχεια. Η κάθε πρόταση εντοπίζεται συνήθως με τα σημεία στίξεως αλλά και πάλι υπάρχει πιθανότητα να μην είναι ακριβές κάτι τέτοιο.

Τα συστήματα που καθορίζουν τα μέρη του λόγου σε μια πρόταση καθώς και συστήματα αυτόματης περίληψης κειμένου βασίζονται πάνω σε πολύ καλή ανάλυση αυτής της φάσης και σωστού εντοπισμού των προτάσεων.

3.1.3 Μορφολογική ανάλυση

Οι λέξεις σε κάθε κείμενο μπορούν να εμφανιστούν σε διαφορετικό γένος, πλήθος και πτώση με αποτέλεσμα ενώ είναι η ίδια λέξη να μην εντοπίζεται κυρίως λόγω των καταλήξεων. Σκοπός αυτής της φάσης είναι η κανονικοποίηση των λέξεων έτσι ώστε ανεξάρτητα από την κατάληξη να μπορούν να καταχωρηθούν, κατά την ανάλυση του κειμένου, ως ίδιες.

Η διαδικασία καθαρισμού των λέξεων λέγεται Stemming και βασίζεται πάνω σε κανόνες αναγνώρισης των καταλήξεων, έτσι, για παράδειγμα, οι λέξεις «καθηγητής» και «καθηγήτρια» ανήκουν στην ίδια ρίζα «καθηγητ».

3.2 Κατηγοριοποίηση

Μετά την ολοκλήρωση της κανονικοποίησης των κειμένων και της εφαρμογής των φίλτρων, είναι το στάδιο της κατηγοριοποίησης.

Η κατηγοριοποίηση βοηθάει στην ανάθεση των κειμένων σε κατηγορίες ανάλογα με το περιεχόμενό τους. Οι μέθοδοι που επιτυγχάνεται η κατηγοριοποίηση είναι δύο:

1. **Επιβλεπόμενη κατηγοριοποίηση**, όπου δίνονται στο σύστημα παραδείγματα κειμένου με σημειωμένη την κατηγορία που ανήκουν και αυτό πρέπει να

εντοπίσει τα κοινά χαρακτηριστικά τους. Από τη διαδικασία αυτή εξάγονται κανόνες για το πώς θα κατηγοριοποιηθούν όλα τα άγνωστα κείμενα.

2. **Μη επιβλεπόμενη κατηγοριοποίηση**, όπου το σύστημα προσπαθεί να εντοπίσει κάποια κοινά χαρακτηριστικά και συσχετίσεις σε μια ομάδα κειμένων ώστε είτε να τα τοποθετήσει σε κατηγορίες ή πόσες είναι σε αριθμό, είτε να βγάλει κάποια συμπεράσματα υπο μορφή προτύπων.

Οι βασικές μέθοδοι που χρησιμοποιούνται για την κατηγοριοποίηση κειμένων είναι: Μέθοδος Bayes, Δέντρα Αποφάσεων, Νευρωνικά Δίκτυα, k-Nearest Neighbour και Διανυσματικά μοντέλα(Vector Space Model).

3.2.1 Διανυσματικά μοντέλα

Η συγκεκριμένη εργασία βασίστηκε σε Διανυσματικά μοντέλα και πιο συγκεκριμένα σε μια υλοποίηση της βιβλιοθήκης libsvm σε C#. Τα διανυσματικά μοντέλα (SVM, Vapnik, 1995; Vapnik, 1998). Τα διανυσματικά μοντέλα ανήκουν στην επιβλεπόμενη κατηγοριοποίηση όπου απαιτείται ένα μέρος δεδομένων για την εκπαίδευση του μοντέλου για να ξεκινήσει η αυτόματη κατηγοριοποίηση έπειτα.

Τα διανυσματικά μοντέλα είναι αλγεβρικά μοντέλα που χρησιμοποιούνται για Φιλτράρισμα Πληροφορίας, Εξόρυξη Πληροφορίας, Ευρετηρίαση, κατηγοριοποίηση σχετικότητας. Αναπαριστούν κείμενα φυσικής γλώσσας με οργανωμένο τρόπο με τη χρήση διανυσμάτων σε έναν πολυδιάστατο χώρο με θετικές μόνο τιμές. Η τυπική διαδικασία κατηγοριοποίησης περιλαμβάνει τρία στάδια. Το πρώτο στάδιο είναι η ευρετηρίαση των λέξεων των κειμένων, όπου ξεκαθαρίζουν από τις άχρηστες λέξεις. Το δεύτερο βήμα περιλαμβάνει την σύνδεση της κάθε λέξεως με κάποιο βάρος επίδρασης

στο αν ανήκει ή όχι σε αυτή την κατηγορία το κείμενο. Το τρίτο βήμα αναλαμβάνει να υπολογίσει τις ομοιότητες των κειμένων που έρχονται προς κατηγοριοποίηση και του μοντέλου που έχει δημιουργηθεί από την εκπαίδευση που έγινε.

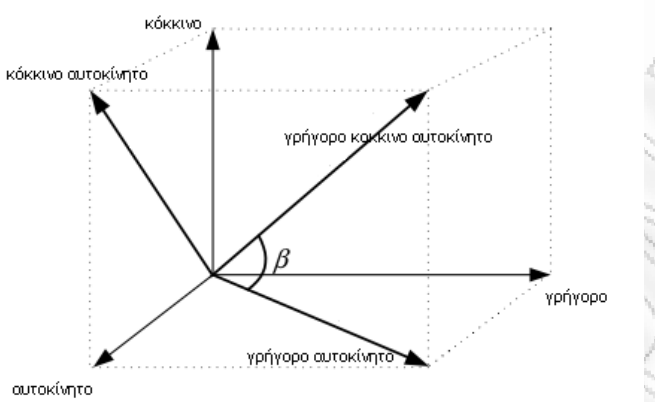
Η ευρετηρίαση των κειμένων ίσως τις περισσότερες φορές προϋποθέτει και την προεπεξεργασία των κειμένων αποκλείοντας τους συνδέσμους, τα σημεία στίξης και καθαρισμού των λέξεων από τις καταλήξεις τους (Stemming).

Η σύνδεση κάθε λέξης με βάρος επίδρασης είναι υπολογίσιμο με βάση τα στατιστικά. Υπάρχουν τρεις βασικοί παράγοντες που επηρεάζουν αυτό το δείκτη:

1. Συχνότητα λέξης στο κείμενο
2. Συχνότητα λέξης στο εκπαιδευτικό σύνολο κειμένων
3. Μήκος κανονικοποιημένου διανύσματος κειμένου.

Το συνολικό βάρος μπορεί να προκύψει από κάποιον ή όλους τους παράγοντες, αν και δοκιμές έδειξαν ότι η χρήση και των τριών παραγόντων δίνει το καλύτερο αποτέλεσμα.

Η ομοιότητα των κειμένων καθορίζεται από τη χρήση συνειρμικών συντελεστών βασισμένων στο διάνυσμα που προέκυψε από τα κείμενα εκπαίδευσης και ενός κειμένου προς κατηγοριοποίηση όπου η επικάλυψη λέξεων δείχνει την ομοιότητα. Στις περισσότερες περιπτώσεις ο συντελεστής συνημίτονου, που μετρά τη γωνία μεταξύ των διανυσμάτων κειμένων χρησιμοποιείται ως μέτρο ομοιότητας.



Εικόνα 4

Το παραπάνω σχήμα δίνει ένα απλό παράδειγμα αναπαράστασης των διανυσμάτων που δημιουργούνται. Οι τρεις διαστάσεις αναπαριστούν τις λέξεις κόκκινο, αυτοκίνητο και γρήγορο. Τα διανύσματα «κόκκινο αυτοκίνητο», «γρήγορο αυτοκίνητο» και «γρήγορο κόκκινο αυτοκίνητο» είναι τρία διανύσματα που επίσης αναπαρίστανται, από την εκπαίδευση του συστήματος. Η ομοιότητα των κειμένων εκφράζεται ως η γωνία μεταξύ των διανυσμάτων, δηλαδή, η γωνία β εκφράζει την ομοιότητα μεταξύ των κειμένων «γρήγορο αυτοκίνητο» και «γρήγορο κόκκινο αυτοκίνητο» και η φόρμα αναπαράστασης του σχήματος είναι η εξής:

$$
\begin{aligned}
\vec{d_i} &= \left(\omega_{d_i,t_1}, \omega_{d_i,t_2}, \ldots, \omega_{d_i,t_{\#T}}\right) \\
\vec{d_j} &= \left(\omega_{d_j,t_1}, \omega_{d_j,t_2}, \ldots, \omega_{d_j,t_{\#T}}\right) \\
sim(d_i, d_j) &= \frac{\vec{d_i}\vec{d_j}}{|\vec{d_i}||\vec{d_j}|} \\
&= \frac{\sum_{t \in T} \omega_{d_i,t}\omega_{d_j,t}}{\sqrt{\sum_{t \in T} \omega_{d_i,t}^2}\sqrt{\sum_{t \in T} \omega_{d_j,t}^2}}
\end{aligned}
$$

Εικόνα 5

και η εξίσωση εισαγωγής της μεταβλητής επηρεασμού απο το σύνολο των κειμένων, έτσι όπως την πρότειναν οι Salton, Wong και Yang είναι:

$$\omega_{d,t_i} = \frac{\alpha_{d,t_i}}{\max_{t \in T} \alpha_{d,t}} \log \frac{\#D}{\#\{e \in D : \alpha_{e,t_i} > 0\}}$$

Εικόνα 6

γνωστή ως tf-idf και επιτρέπει οι λέξεις να έχουν διαφορετικό βάρος για κάθε κείμενο της συλλογής.

4 Αποθήκες δεδομένων

Με τον όρο Αποθήκες Δεδομένων (Data Warehouses) χαρακτηρίζουμε ένα σύνολο τεχνολογιών που επιτρέπει στους αναλυτές ενός οργανισμού στη σχεδίαση της πολιτικής του έχοντας αποδοτική πρόσβαση στα δεδομένα του οργανισμού. Μία Αποθήκη Δεδομένων διατηρεί δεδομένα που αντλεί από τις βάσεις δεδομένων των πληροφοριακών συστημάτων του οργανισμού αλλά και άλλες πηγές δεδομένων, όπως αρχεία του οργανισμού ή δεδομένα που προέρχονται από εξωτερικές πηγές. Αυτά τα δεδομένα οργανώνονται στην Αποθήκη δεδομένων σε δομές κατάλληλες να απαντήσουν τις απαιτήσεις των αναλυτών - χρηστών των συστημάτων στήριξης αποφάσεων. Τα συστήματα στήριξης αποφάσεων αποκτούν πρόσβαση στα δεδομένα λειτουργίας του οργανισμού χωρίς την παρουσία των προαναφερθέντων προβλημάτων. Οι Αποθήκες δεδομένων παρέχουν τη δυνατότητα για Συνεχή Αναλυτική Επεξεργασία (On-Line Analytical Processing- OLAP) των δεδομένων περιέχοντας συνήθως ιστορικά και συγκεντρωτικά δεδομένα που συνήθως αποδεικνύονται χρήσιμα για υποστήριξη αποφάσεων. Επίσης, παρέχουν μία ολοκληρωμένη εικόνα του σχήματος των δεδομένων του οργανισμού.

Η σχεδίαση των Αποθηκών Δεδομένων έχει σαν στόχο την αποδοτική απάντηση των πολύπλοκων ερωτήσεων που θέτονται κατά την αναλυτική επεξεργασία δεδομένων από τις εφαρμογές στρατηγικού σχεδιασμού. Η δημιουργία και η συντήρηση μιας Αποθήκης Δεδομένων είναι μία πολύπλοκη διαδικασία καθώς πολλές διαφορετικές προσεγγίσεις είναι εφικτές. Αρκετοί οργανισμοί επιδιώκουν να δημιουργήσουν μία Αποθήκη Δεδομένων που θα περιέχει αναλυτικά δεδομένα από όλες τις δραστηριότητες του

οργανισμού. Πρόκειται για ένα πολύπλοκο εγχείρημα που απαιτεί μεγάλο κόστος για να επιτύχει. Μία άλλη λύση είναι η δημιουργία Επιμέρους Συλλογών Δεδομένων (data marts) με κριτήριο το αντικείμενο των εφαρμογών από τις οποίες προέρχονται ή το τμήμα του οργανισμού που τις χρησιμοποιεί. Πρόκειται για πιο ευέλικτα συστήματα στη δημιουργία τους, τα οποία όμως δεν παρέχουν ενιαία λύση, δημιουργώντας προβλήματα σε περίπτωση μακρόχρονης χρήσης τους.

Τα τελευταία χρόνια η ανάπτυξη και λειτουργία Αποθηκών δεδομένων κρίνεται κρίσιμη για την λειτουργία των οργανισμών. Τεράστια ποσά επενδύονται σε αυτή τη δραστηριότητα ενώ τα οφέλη από τη λειτουργία τέτοιων συστημάτων κρίνονται ήδη ως ιδιαίτερα σημαντικά. Όπως είναι φυσικό, όλες οι μεγάλες εταιρείες του χώρου των Βάσεων δεδομένων και των πληροφοριακών συστημάτων αναπτύσσουν και προτείνουν προϊόντα στο χώρο των Αποθηκών δεδομένων. Τα επόμενα χρόνια αναμένονται ακόμα μεγαλύτερες επενδύσεις σε τεχνολογία αιχμής του χώρου. Για την παρουσίαση του κεφαλαίου αυτού, στηριχτήκαμε κυρίως στα [BS96], [CD 96], [CD97], [Coll96], [Inm96], [Kena95], [RedB97], [Wido95].

4.1 Αρχιτεκτονική

Η επιλογή της αρχιτεκτονικής μιας αποθήκης δεδομένων πρέπει να ικανοποιεί τις συγκεκριμένες ανάγκες του οργανισμού για τις οποίες δημιουργήθηκε και να εξασφαλίζει τη διαθεσιμότητα και την αποδοτικότητα του συστήματος. Υπάρχουν τρεις Αρχιτεκτονικές με βάση την πολυπλοκότητα του οργανισμού που θα εξυπηρετήσουν:

1. Βασική Αρχιτεκτονική

Η βασική αρχιτεκτονική περιέχει μόνο τις απολύτως απαραίτητες μονάδες για μια Αποθήκη Δεδομένων όπου οι τελικοί χρήστες έχουν άμεση πρόσβαση στα δεδομένα.

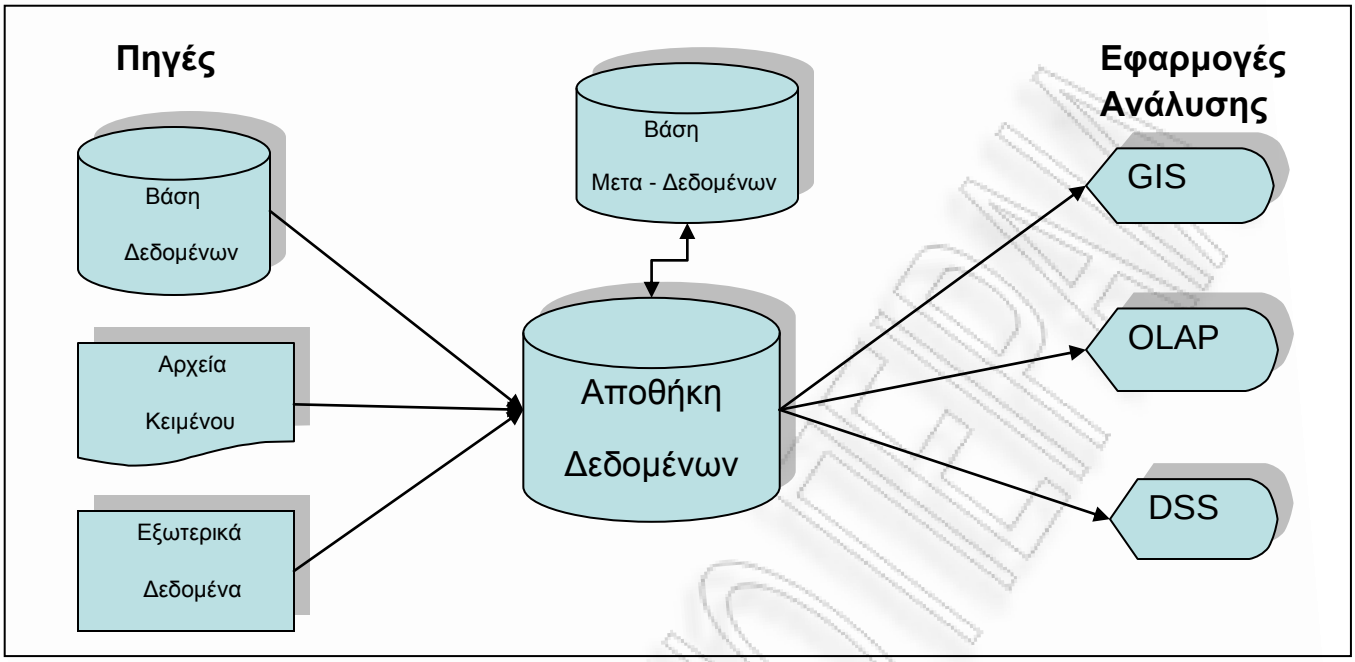
2. Αρχιτεκτονική με Προσωρινή Περιοχή

Η προπαρασκευή των δεδομένων πριν μεταφερθούν στην Αποθήκη είναι συνήθως απαραίτητη. Η διαδικασία αυτή στις περισσότερες Αποθήκες Δεδομένων επιτυγχάνεται και μέσω ενός επιπλέον βήματος προπαρασκευαστικό στο οποίο απλοποιείται η δημιουργία συνόψεων και γενικά διαχείρισης της Αποθήκης.

3. Αρχιτεκτονική με Προσωρινή Περιοχή και Επιμέρους Συλλογές Δεδομένων

Η αρχιτεκτονική που συνδυάζει και επιμέρους Συλλογές Δεδομένων εξυπηρετεί τον οργανισμό ώστε να έχει διαφορετικές συλλογές για τα τμήματά του.

Το παρακάτω σχήμα παρουσιάζει μια γενική αρχιτεκτονική ενός συστήματος Αποθήκης Δεδομένων. Στο σχήμα σημειώνονται τα βασικά δομικά στοιχεία μίας Αποθήκης Δεδομένων, η διασύνδεση των στοιχείων τους, καθώς και η ροή των δεδομένων.



Εικόνα 7

Τα βασικά μέρη της αρχιτεκτονικής ενός συστήματος Αποθήκης Δεδομένων είναι τα ακόλουθα:

- **Πηγές:** Κάθε πηγή από την οποία η Αποθήκη Δεδομένων αντλεί δεδομένα.
- **Αποθήκη Δεδομένων:** Τα συστήματα που αποθηκεύονται δεδομένα που παρέχονται προς τους χρήστες.
- **Βάση Μεταδεδομένων:** Σύστημα αποθήκευσης πληροφορίας σχετικά με τη δομή και τη λειτουργία του συστήματος.
- **Εφαρμογές Ανάλυσης:** Εφαρμογές που έχουν πρόσβαση στην Αποθήκη Δεδομένων.

4.2 Μεταφορά Δεδομένων από τις Πηγές

Βασικός παράγοντας για την επιτυχία των Αποθηκών Δεδομένων είναι η ορθή τροφοδοσία της Αποθήκης Δεδομένων από τις πηγές. Η διαδικασία μεταφοράς δεδομένων από τις πηγές στην Αποθήκη δεδομένων είναι αρκετά πολύπλοκη καθώς πολλά προβλήματα πρέπει να αντιμετωπισθούν:

1. Εξαγωγή Δεδομένων απο τις πηγές.
2. Καθαρισμός των δεδομένων με την διάγνωση πιθανών ασυνεπειών και τη μεταφορά μόνο των πραγματικά χρήσιμων δεδομένων.
3. Μετατροπή των δεδομένων μεταξύ διαφορετικών μοντέλων και προτύπων.
4. Διάγνωση αλλαγών στα δεδομένα των πηγών και μεταφορά των νέων δεδομένων.
5. Ανάλυση των μεταφερόμενων δεδομένων για τη διάγνωση μη ορθής πληροφορίας.
6. Έλεγχος πληρότητας.

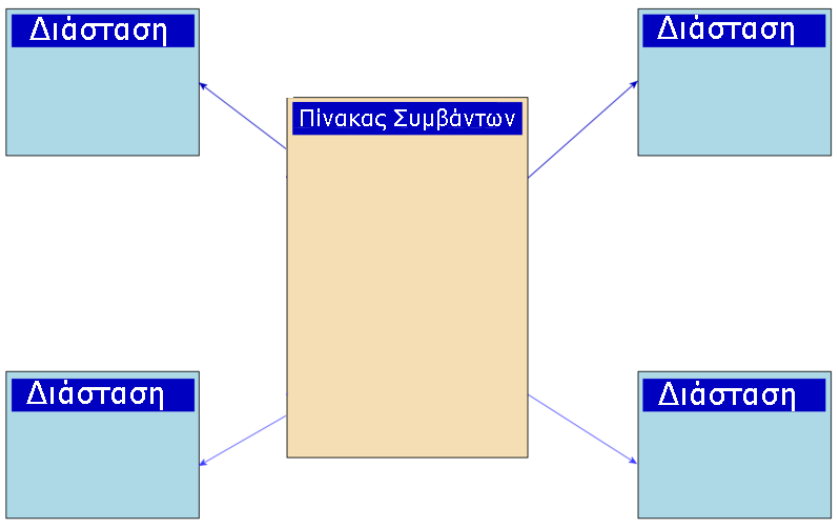
4.3 Αποθήκη Δεδομένων

Οι αποθήκες δεδομένων είναι ένας χώρος αποθήκευσης πληροφορίας που συλλέγεται από πολλές πηγές με ένα ορισμένο τρόπο. Επίσης είναι σχεδιασμένες με σκοπό την ανάλυση των δεδομένων, τα οποία είναι πολυδιάστατα με μετρήσιμες ιδιότητες και διαστάσεις. Οι μετρήσιμες ιδιότητες είναι αυτές που περιέχουν κάποιου είδους τιμή και μπορούν να υπολογιστούν. Για παράδειγμα, σε ένα σύστημα πωλήσεων, ο αριθμός των προϊόντων που πωλήθηκαν είναι μία μετρήσιμη ιδιότητα. Οι διαστάσεις

είναι ιδιότητες πάνω στις οποίες μία μετρήσιμη ιδιότητα μπορεί να προβληθεί, για παράδειγμα, το όνομα ενός προϊόντος ή η ημερομηνία πώλησης είναι διαστάσεις.

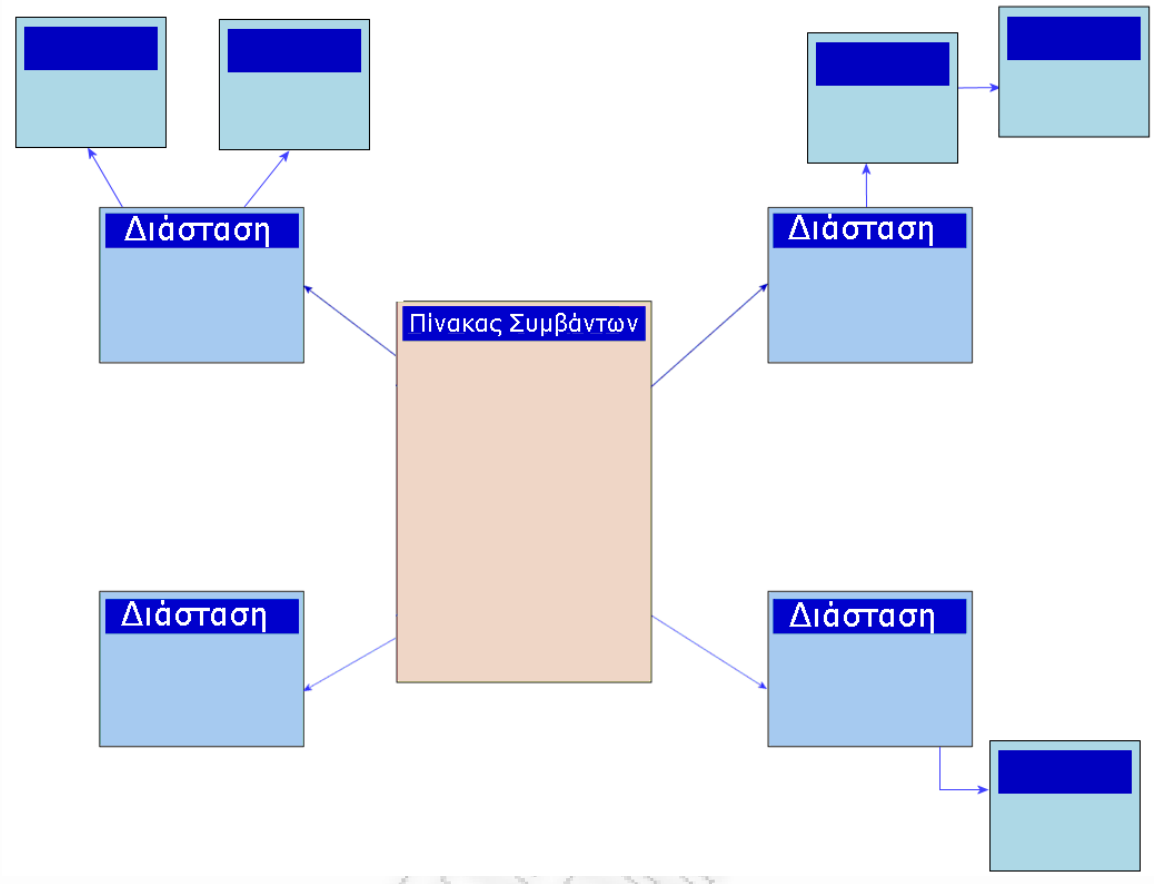
Στις αποθήκες δεδομένων, υπάρχει ο πίνακας συμβάντων (fact table), που περιέχει τα πολυδιάστατα δεδομένα και συνήθως είναι πολύ μεγάλος. Για να μειωθούν οι απαιτήσεις, οι διαστάσεις έχουν μικρά πεδία που είναι ξένα κλειδιά (foreign keys) σε άλλους πίνακες που λέγονται πίνακες διαστάσεων(dimensional tables). Ο πίνακας συμβάντων διατηρείται κανονικοποιημένος όσο το δυνατόν περισσότερο γι α να μειωθούν οι απαιτήσεις σε χώρο ενώ, οι πίνακες διαστάσεων έχουν πολλά επιπλέον πεδία έτσι ώστε να είναι εύκολη η εμφάνιση των στοιχείων που μας ενδιαφέρουν.

Τα πιο δημοφιλή σχήματα αποθηκών δεδομένων είναι το αστεροειδές, στο οποίο ο πίνακας συμβάντων περιέχει ξένα κλειδιά σε πίνακες διαστάσεων



Εικόνα 8

ενώ, αν οι πίνακες διαστάσεων περιέχουν και εκείνοι με τη σειρά τους ξένα κλειδιά σε περισσότερους πίνακες διαστάσεων τότε το σχήμα ονομάζεται χιονονιφάδα.



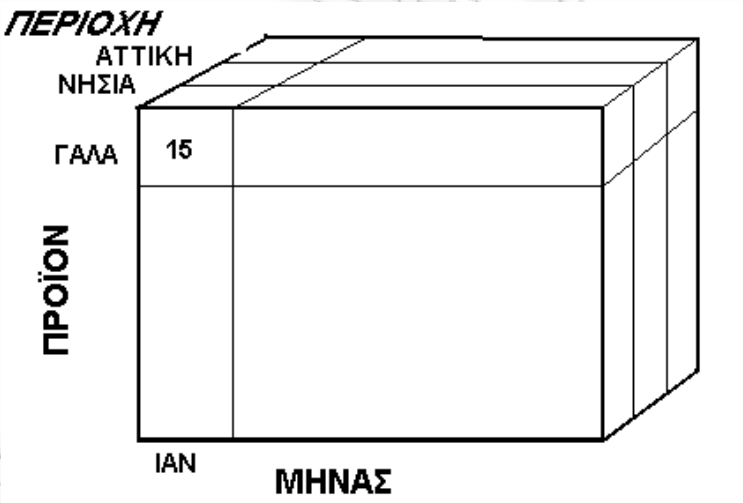
Εικόνα 9

4.4 Μοντέλα Ανάλυσης Δεδομένων

Στο πρόγραμμα που αναπτύχθηκε χρειάστηκε να γίνει επεξεργασία των δεδομένων για την εξαγωγή συμπερασμάτων σχετικά με τα σχόλια των ιστολόγιων σε συνάρτηση με άλλες διαστάσεις όπως ο χρόνος ή η κατηγορία του κειμένου. Η αναλυτική επεξεργασία δεδομένων είναι τμήμα των εφαρμογών στήριξης αποφάσεων και των στρατηγικών πληροφοριακών συστημάτων. Η παρουσίαση των αποτελεσμάτων απαιτεί την ενδελεχή ανάλυση των δεδομένων και την λεπτομερή ανάλυση σε επιμέρους αποτελέσματα βάσει των οποίων θα ληφθούν αποφάσεις.

4.4.1 Πολυδιάστατα Μοντέλα Δεδομένων

Οι πίνακες των σχεσιακών βάσεων δεδομένων περιέχουν εγγραφές οι οποίες αποτελούνται από πεδία. Σε φυσιολογικές σχεσιακές βάσεις δεδομένων, ένα υποσύνολο των πεδίων ενός πίνακα συνθέτουν το κλειδί του. Αντίθετα τα πολυδιάστατα μοντέλα δεδομένων περιέχουν ν-διάστατους πίνακες που συχνά αποκαλούνται υπερκύβοι (cubes ή hyper cubes). Κάθε διάσταση έχει μία ιεραρχία επιπέδων. Για παράδειγμα, η διάσταση "Γεωγραφική τοποθεσία" έχει τα επίπεδα πόλη, νομός, χώρα. Οι τιμές (μετρικές) που περιέχουν οι υπερκύβοι αντιστοιχούν στις στήλες των σχεσιακών πινάκων.



Εικόνα 10

Στο σχήμα εμφανίζεται ένα μοντέλο των δεδομένων ενός υπερκύβου που παρέχει δεδομένα για τις πωλήσεις προϊόντων. Σύμφωνα με το παράδειγμα οι πωλήσεις του προϊόντος «ΓΑΛΑ» το μήνα Ιανουάριο στα νησιά ήταν 15. Οι τιμές «ΓΑΛΑ», «ΙΑΝ» και «ΝΗΣΙΑ» στο συγκεκριμένο παράδειγμα είναι οι τιμές των διαστάσεων «ΠΡΟΪΟΝ», «ΜΗΝΑΣ» και «ΠΕΡΙΟΧΗ» αντίστοιχα. Ανάλογα με τις ερωτήσεις που κάνουμε και τις

διαστάσεις που θα συμπεριλάβουμε στις ερωτήσεις μας έχουμε τη δυνατότητα να παίρνουμε συγκεντρωτικές πληροφορίες για τις τιμές των δεδομένων που ζητήσαμε. Αν στο παράδειγμα του σχήματος εκτελέσουμε μια αναζήτηση στα δεδομένα του κύβου με βάση μόνο συγκεκριμένες τιμές για δύο διαστάσεις (ΠΡΟΪΟΝ = «ΓΑΛΑ» και ΜΗΝΑΣ = «ΙΑΝ») τότε θα δούμε αποτελέσματα συγκεντρωτικά για το «ΓΑΛΑ» στον Ιανουάριο για όλες τις περιοχές. Επίσης υπάρχει η δυνατότητα να κάνουμε ερωτήσεις με εύρος τιμών όπως, για παράδειγμα, να πάρουμε τις πωλήσεις για το «ΓΑΛΑ» σε όλες τις περιοχές το τελευταίο έτος.

4.4.2 Πράξεις στους υπερκύβους

Οι υπερκύβοι μας δίνουν τη δυνατότητα πλοήγησης στις ιεραρχίες των διαστάσεών τους. Η πλοήγηση είναι δυνατή από τις πράξεις οι οποίες μας παρέχονται. Οι πράξεις που συνήθως γίνονται στους υπερκύβους είναι οι παρακάτω:

4.4.2.1 ROLL-UP

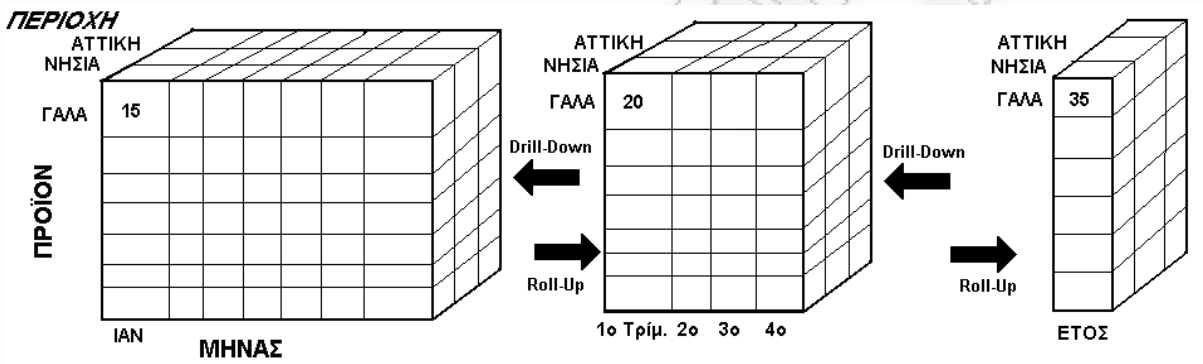
Οι διαστάσεις ενός υπερκύβου πολλές φορές έχουν επίπεδα όπως, για παράδειγμα, στη διάσταση ΜΗΝΑΣ θα μπορούσε να υπάρχουν τα ανώτερα επίπεδα:

- ΤΡΙΜΗΝΟ
- ΕΤΟΣ

Η πράξη, σε έναν υπερκύβο, που εκτελούμε για να ανέβουμε στην ιεραρχία μιας διάστασης είναι Roll-up. Τα αποτελέσματα μετά από μια πράξη Roll-up στο παράδειγμα θα έδινε έναν νέο κύβο που θα περιείχε αθροιστικές πωλήσεις για το «ΓΑΛΑ» στα «ΝΗΣΙΑ» για ένα τρίμηνο ή έτος, δηλαδή τα δεδομένα θα εμφανίζονταν ομαδοποιημένα με βάση τη διάσταση στην οποία έγινε και στο επίπεδο που επιλέξαμε.

4.4.2.2 DRILL-DOWN

Η πράξη drill-down προσφέρει τη δυνατότητα να αναλύσουμε τα δεδομένα ανάλογα με το επίπεδο που έχει γίνει η ομαδοποίηση σε μια διάσταση. Αντίστροφα με το roll-up, εδώ πάμε σε ένα χαμηλότερο επίπεδο ιεραρχίας μιας διάστασης. Για παράδειγμα, η πράξη drill-down από το επίπεδο τριμήνου ή έτους σε επίπεδο μήνα θα μας έδινε τον αρχικό κύβο.



Εικόνα 11

4.4.2.3 SLICING

Αν θέλουμε να δημιουργήσουμε έναν κύβο που να είναι μόνο από το χαμηλότερο επίπεδο των διαστάσεων του υπερκύβου, τότε επιλέγουμε την πράξη slicing. Στο παράδειγμα, θα πρέπει να επιλέξουμε μόνο το «ΓΑΛΑ» από το ΠΡΟΪΟΝ, τα «ΝΗΣΙΑ» και την «ΑΤΤΙΚΗ» από την ΠΕΡΙΟΧΗ και τον «ΙΑΝ» από τους ΜΗΝΕΣ.

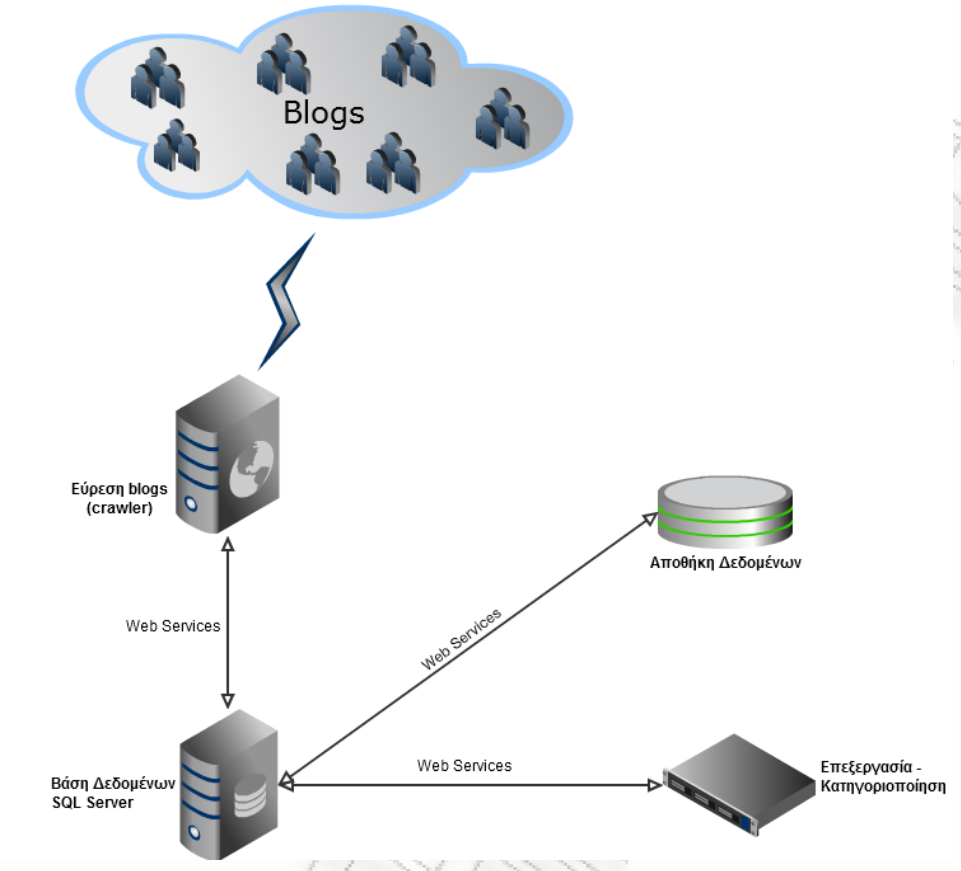
5 Περιγραφή συστήματος

5.1 Υποσυστήματα

Η εργασία έχει σκοπό να ακολουθήσει όλα τα βήματα της διαδικασίας εξόρυξης γνώσης από τα ελληνικά διαδικτυακά ιστολόγια, έτσι ώστε να καταγραφεί η κίνηση στους ιστοτόπους και να παρουσιαστούν στατιστικά σχετικά με την ελληνική πραγματικότητα.

Πιο συγκεκριμένα, τα υποσυστήματα που θα αποτελείται η εργασία είναι:

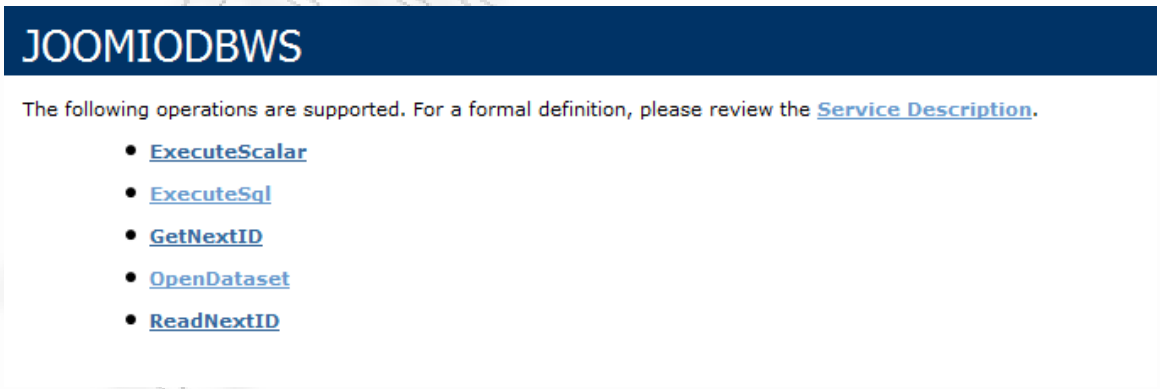
1. Υποσύστημα εύρεσης Blogs.
2. Υποσύστημα εύρεσης κειμένων στα Blogs.
3. Υποσύστημα εξόρυξης κειμένων.
4. Υποσύστημα Αποθήκης & Ανάλυσης Δεδομένων με κείμενα.



Εικόνα 12

5.1.1 Υπηρεσίες Ιστού

Οι υπηρεσίες ιστού που υλοποιήθηκαν ήταν οι υπηρεσίες σχετικά με τη Βάση Δεδομένων :



Εικόνα 13

1. ExecuteScalar: η υπηρεσία παίρνει ως παραμέτρους μία ή περισσότερες sql εντολές και τις εκτελεί στη βάση δεδομένων που είναι ορισμένη από τις ρυθμίσεις του εξυπηρετητή.

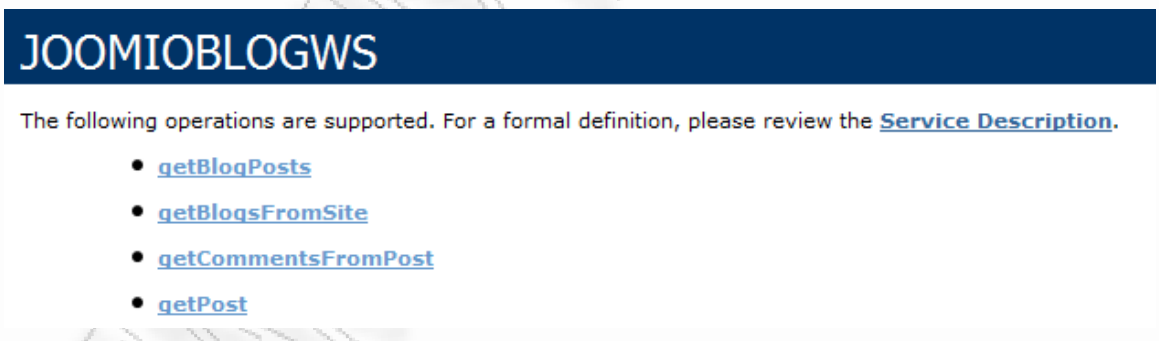
2. ExecuteSql: η υπηρεσία παίρνει ως παράμετρο μία εντολή sql και την εκτελεί στη βάση δεδομένων, σε περίπτωση που εκτελεστεί σωστά επιστρέφει τον αριθμό των εγγραφών που επηρεάστηκαν, ενώ αν εκτελεστεί λάθος επιστρέφει -1.

3. GetNextID: η υπηρεσία παίρνει ως παράμετρο ένα όνομα πίνακα και επιστρέφει το επόμενο id που θα αποθηκευθεί στη βάση δεδομένων.

4. OpenDataset: η υπηρεσία παίρνει ως παράμετρο ένα sql query και επιστρέφει ένα DataSet με όλες τις τιμές και τους πίνακες που ζητούσε το query.

5. ReadNextID: η υπηρεσία παίρνει ως παράμετρο ένα όνομα πίνακα και επιστρέφει το επόμενο id χωρίς να αυξάνει τον αριθμό.

Επίσης, δημιουργήθηκαν υπηρεσίες ιστού για την προσπέλαση των blogs, posts και comments:



Εικόνα 14

1. getBlogPosts: η υπηρεσία παίρνει ως παράμετρο ένα blog url και επιστρέφει ένα Dataset με όλα τα post url που βρέθηκαν.

2. getBlogFromSite: η υπηρεσία παίρνει ως παράμετρο ένα blog url και επιστρέφει ένα Dataset με όλα τα φιλικά blog που βρήκε, έτσι ώστε να προσπελαστούν στο μέλλον.

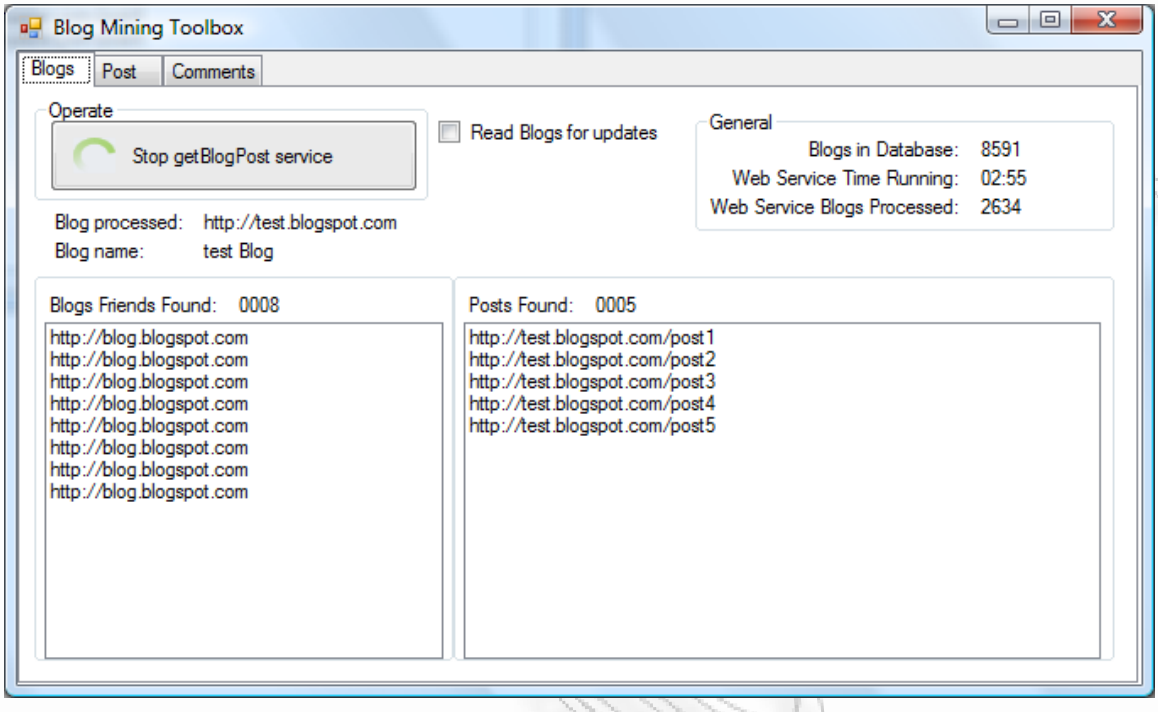
3. getCommentsFromPost: η υπηρεσία παίρνει ως παράμετρο ένα post url και επιστρέφει ένα Dataset με όλα τα comments καθώς και πληροφορίες, αν υπάρχουν, σχετικά με τον συγγραφέα και την ώρα καταχώρησης.

4. getPost: η υπηρεσία παίρνει ως παράμετρο ένα post url και επιστρέφει ένα Dataset με όλα τα στοιχεία σχετικά με αυτό, όπως κείμενο, ημερομηνία, κατηγορίες.

5.1.2 Υποσύστημα εύρεσης Blogs**(crawler)**

Η πρώτη εργασία που λαμβάνει χώρα είναι η διαδικασία εύρεσης των blogs. Η βάση δεδομένων περιέχει κάποια blogs και με αφετηρία αυτά, το σύστημα αναζητά φιλικά blogs που εμφανίζονται συνήθως στις στήλες αριστερά ή δεξιά των blogs.

Το υποσύστημα αυτό λειτουργεί στον εξυπηρετητή που είναι εγκατεστημένο ως υπηρεσία χωρίς την παρεμβολή των χρηστών και αποθηκεύει στη Βάση Δεδομένων που έχει οριστεί, τα στοιχεία που είναι απαραίτητα για την αναζήτηση των κειμένων. Για τη λειτουργία του είναι απαραίτητη η σύνδεση στο Διαδίκτυο, ώστε από το δίκτυο των blogs που αποθηκεύει να είναι σε θέση να βρει νέα.

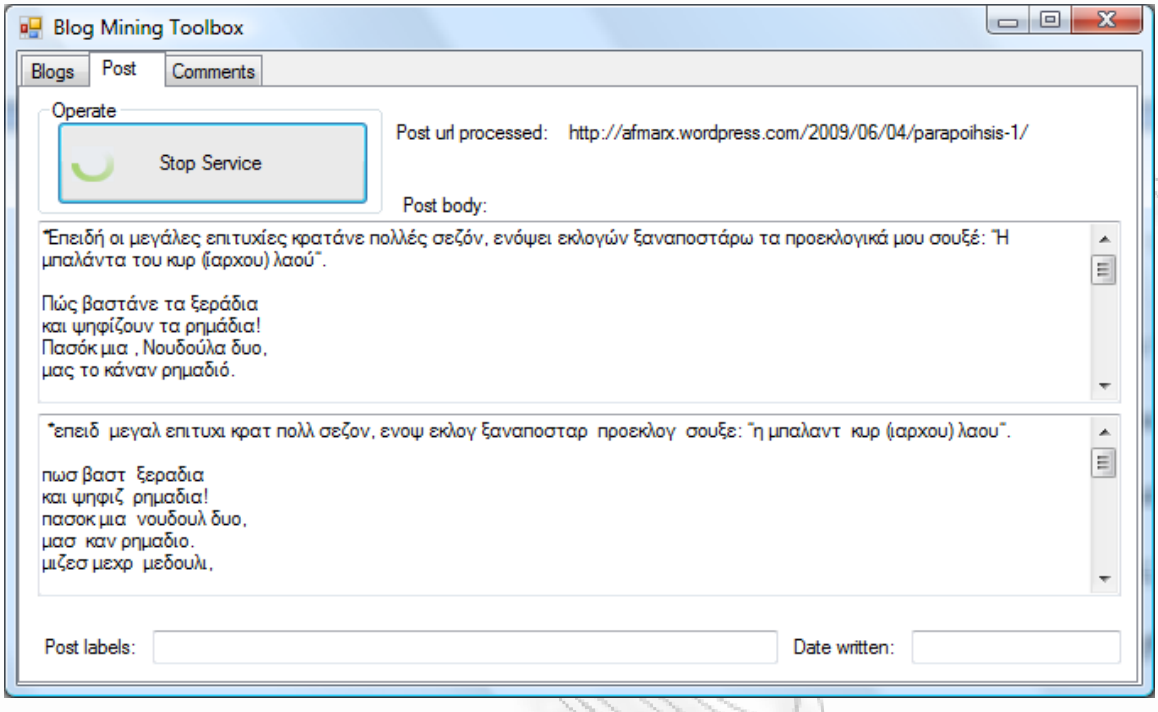


Εικόνα 15

Η νοοτροπία που βασίζεται το πρόγραμμα είναι η αλληλο-διαφήμιση που λαμβάνει μέρος σε όλα σχεδόν τα blogs. Όταν, για παράδειγμα, εμφανιστεί η πρώτη σελίδα ενός blog, συνήθως, στις κολώνες αριστερά ή δεξιά υπάρχει η λίστα με τα αγαπημένα blogs του συγγραφέα, έτσι σιγά-σιγά αναπτύσσεται ένα δίκτυο με «φιλικά» blogs.

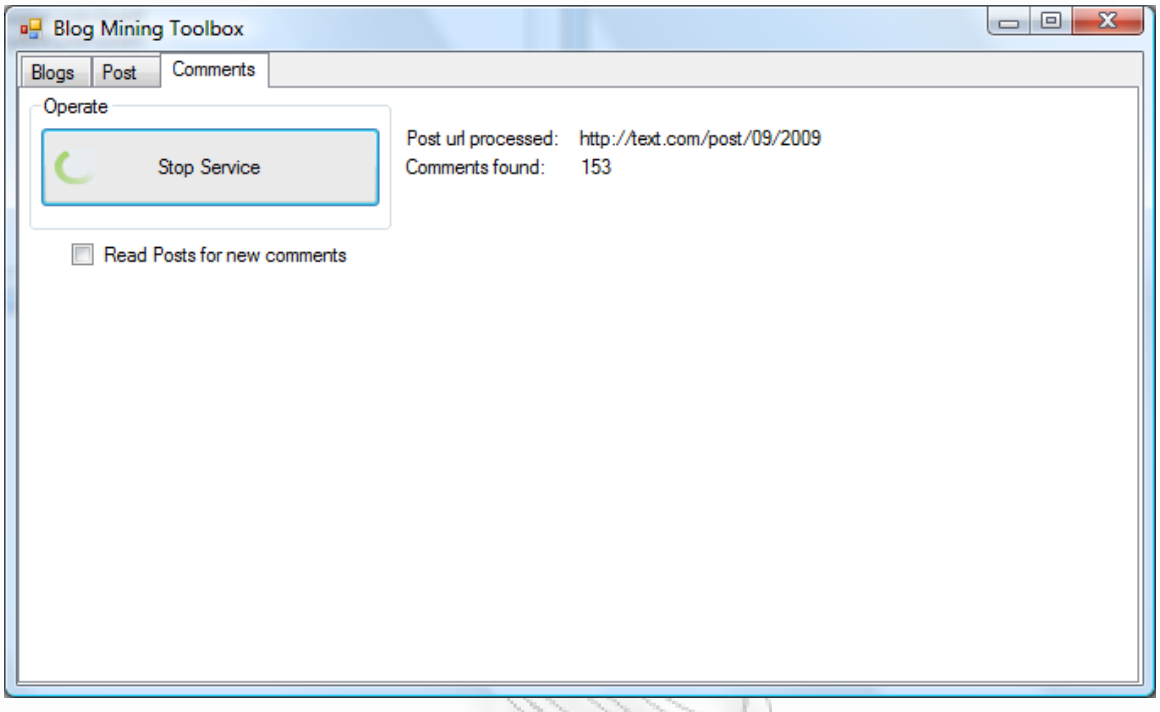
Για παράδειγμα, ζητείται το πρώτο blog που είναι στη Βάση Δεδομένων και δεν έχει προσπελαστεί ακόμα ή έχει προσπελαστεί σε περισσότερο από 1 ημέρα,

Στη συνέχεια, χρειάζεται να βρεθούν τα κείμενα σε κάθε blog και να καταχωρηθούν στη Βάση Δεδομένων. Με τεχνολογία υπηρεσιο-στραφούς αρχιτεκτονικής και web services για ελευθερία χρήσης των υπηρεσιών από οποιοδήποτε σημείο, υλοποιήθηκαν ρουτίνες εύρεσης κειμένων σε δημοφιλείς πλατφόρμες δημιουργίας blogs (blogspot, wordpress, pblogs.gr). Η διαδικασία εύρεσης των κειμένων αναπτύχθηκε με σκοπό να παρακάμπτει τα πρότυπα που χρησιμοποιούν οι πάροχοι.



Εικόνα 16

Το υποσύστημα καταχωρεί όλη την πληροφορία που υπάρχει σχετικά με τα blogs όπως κείμενα, σχόλια, «φιλικά» blogs. Η εύρεση επιπλέον σχόλιων υπάρχει ως επιλογή στην εφαρμογή έτσι ώστε να ενημερωθεί η βάση δεδομένων μετά από λίγες ώρες με όλα τα νέα σχόλια που έχουν καταχωρηθεί απο την τελευταία φορά που έγινε η αναζήτηση.



Εικόνα 17

Η εύρεση των κειμένων σε κάθε blog επιτυγχάνεται μέσω των υπερσυνδέσμων που υπάρχουν στους τίτλους των δημοσιεύσεων. Κάθε δημοσίευση –στην πλειοψηφία των blogs- έχει ξεχωριστή σελίδα με το κείμενο σε πλήρη ανάπτυξη και τα σχόλιά της.

Το συγκεκριμένο υποσύστημα καταχωρεί όλα τα στοιχεία των blogs στη βάση δεδομένων όπως «φιλικά» blogs, δημοσιεύσεις κάθε blog, σχόλια σε κάθε δημοσίευση καθώς και στοιχεία σχολιαστή αν είναι ενυπόγραφα.

5.1.3 Υποσύστημα εξόρυξης κειμένων

Το υποσύστημα εξόρυξης κειμένων, αναλαμβάνει να μετατρέψει τα κείμενα των δημοσιευμάτων σε κατάλληλη μορφή για την κατηγοριοποίησή τους. Η μετατροπή περιλαμβάνει μια σειρά από διαδικασίες.

Αρχικά, τα κείμενα υφίστανται μορφολογική επεξεργασία (Stemming) όπου το πρόγραμμα πρέπει:

1. Να ξεχωρίσει τις λέξεις, όπως στο «σ' όλα» σε «σ» και «όλα».
2. Να αναγνωρίσει τις καταλήξεις, όπως την κατάληξη «-μένος» στη λέξη «κουρασμένος».

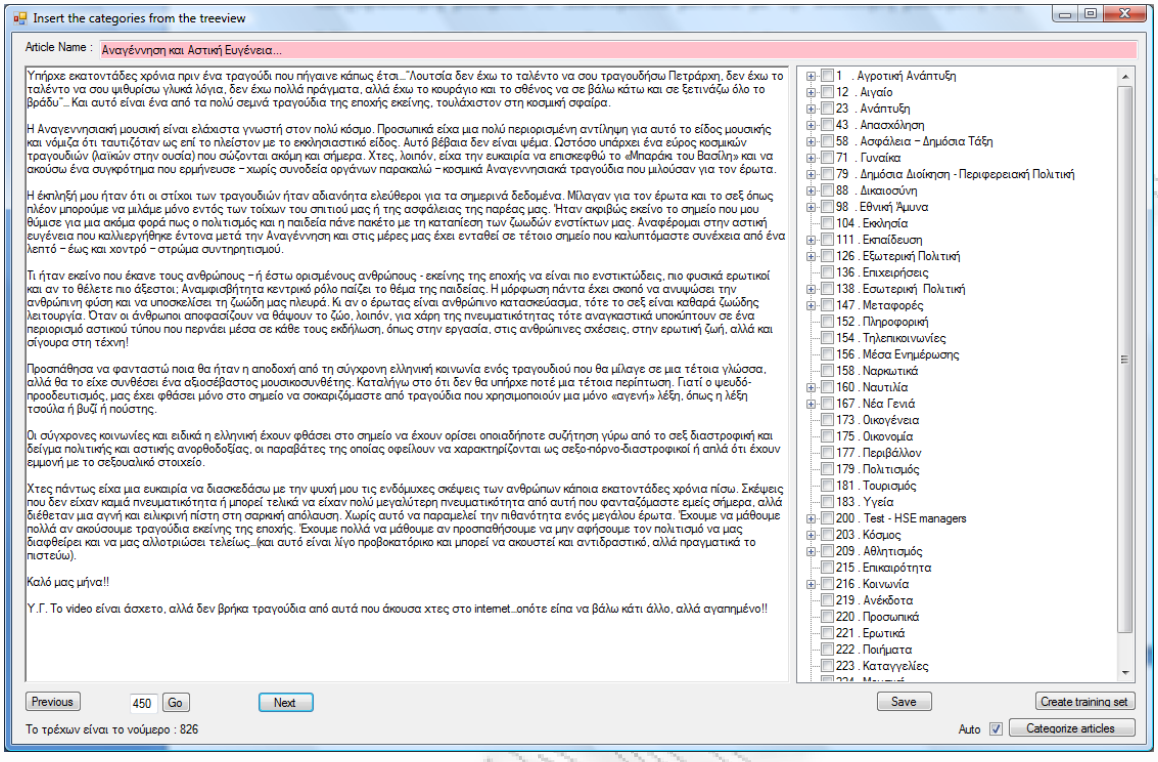
Τα κείμενα καταλήγουν να περιέχουν μόνο τη ρίζα κάθε λέξης οπότε είναι πιο εύκολο στη συνέχεια το πρόγραμμα να ξεκαθαρίσει τα κείμενα επιλέγοντας τις λέξεις που έχουν κάτω από τρεις χαρακτήρες και όλες τις λέξεις που έχουν ρυθμό εμφάνισης σε κάθε κείμενο περισσότερο από 5.

Έπειτα, δημιουργείται ένας πίνακας στη βάση δεδομένων που περιέχει όλες τις ρίζες των λέξεων των κειμένων που θα αποτελέσουν το δείγμα εκπαίδευσης για τον κατηγοριοποιητή. Η κατηγοριοποίηση βασίζεται σε Διανυσματικά μοντέλα με την υλοποίηση βασισμένη στη βιβλιοθήκη «LIBSVM -- A Library for Support Vector Machines».

Στα διανυσματικά μοντέλα μπορούμε να αναπαραστήσουμε ένα έγγραφο dj σαν ένα διάνυσμα (*w1j, w2j, …, wt,j*), όπου *t* το πλήθος όρων και ένα ερώτημα q σαν (*w1q, w2q, …, wtq*). Για τον καθορισμό του βάρους μιας λέξης καθοριστικό ρόλο παίζουν:

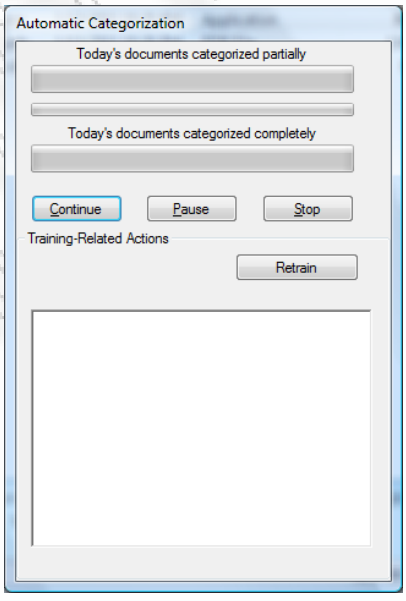
- η συχνότητα του όρου στο κείμενο του εγγράφου
- Ο αριθμός του εγγράφων στα οποία συμμετέχει ο όρος

Στην αρχή δημιουργείται ένα σύνολο κειμένων που δεν κατηγοριοποιήθηκαν αυτόματα και αποτελεί το εκπαιδευτικό για να καταφέρει στη συνέχεια το πρόγραμμα να κατηγοριοποιήσει αυτόματα.



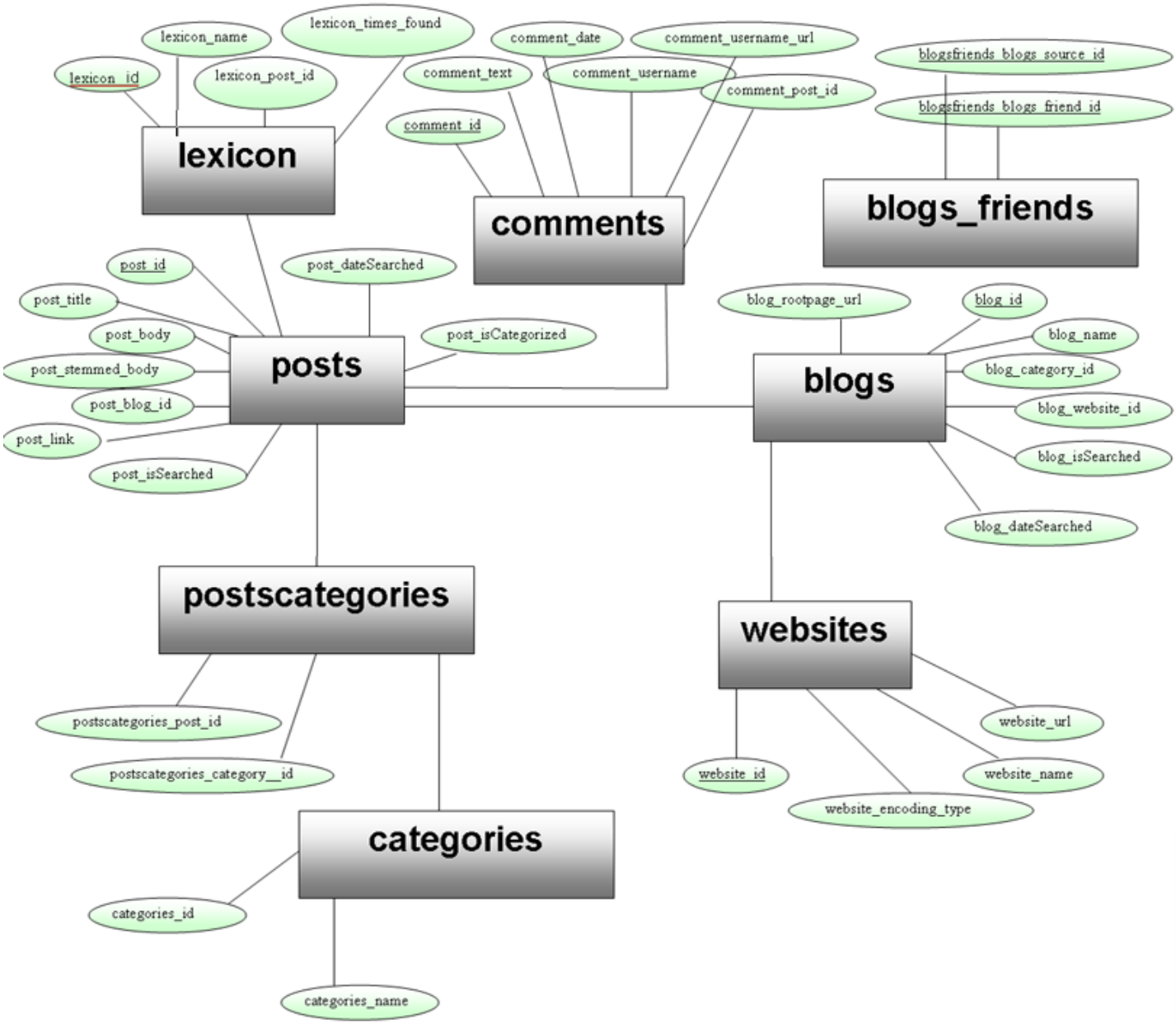
Εικόνα 18

Το σύστημα αφού γίνει η εκπαίδευσή του αρχίζει να κατηγοριοποιεί ανεξάρτητα για κάθε κατηγορία τα κείμενα καθώς αυτά εισέρχονται στη βάση δεδομένων.



Εικόνα 19

5.1.4 Υποσύστημα Αποθήκης & Ανάλυσης Δεδομένων



Εικόνα 20

Το παραπάνω σχήμα εμφανίζει το σχήμα της Βάσης Δεδομένων από το οποίο προέκυψε η Αποθήκη Δεδομένων. Ο πίνακας «posts» περιέχει την κάθε δημοσίευση η οποία δημοσιεύεται σε ένα «blog» το οποίο ανήκει σε ένα «website», το οποίο είναι ο πάροχος της υπηρεσίας όπως

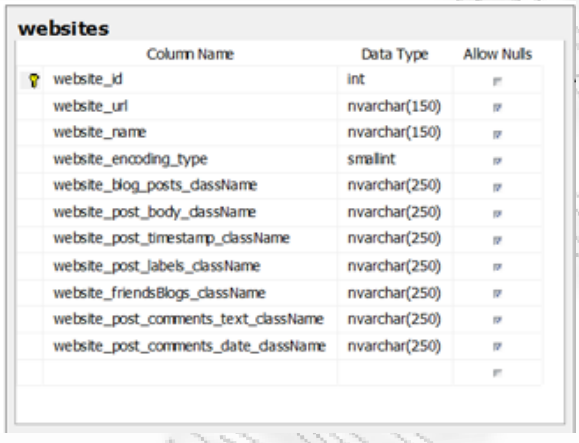
“blogspot.com”, “wordpress.com”, “pblogs.gr”. Τα σχόλια για κάθε δημοσίευση περιέχονται στον πίνακα «comments». Στον πίνακα «lexicon» αποθηκεύονται στοιχεία σχετικά με τις λέξεις που περιέχουν τα κείμενα και χρησιμεύει στην κατηγοριοποίηση. Τέλος, οι κατηγορίες έχουν δημιουργηθεί έτσι ώστε να μπορεί ένα κείμενο να ανήκει σε πολλές κατηγορίες με σκοπό να χρησιμεύσει σε κατηγορίες που εννοιολογικά υπερκαλύπτονται π.χ. «Πολιτική» > «Εξωτερική Πολιτική» > «Τουρκία».

Το υποσύστημα Αποθήκης Δεδομένων υλοποιήθηκε με SQL SERVER 2005. Τα στοιχεία των blogs που υπάρχουν ήδη στη βάση Δεδομένων δεν έχουν καλή ποιότητα και παρουσιάζουν κενά. Επίσης, η ανακάλυψη συσχετίσεων γίνεται μέσα από περίπλοκα ερωτήματα που δεν μπορεί να διαχειριστεί η βάση δεδομένων. Το υποσύστημα Αποθήκης Δεδομένων έγινε με σκοπό να μοντελοποιηθούν τα στοιχεία που υπάρχουν ώστε να εξαχθούν συμπεράσματα σχετικά με τη συχνότητα δημοσίευσης blogs καθώς και τη συχνότητα απαντήσεων-σχολίων σε αυτές. Η αποθήκη δεδομένων επιτρέπει την παρακολούθηση των πληροφοριών αυτών σε βάθος χρόνου με την επιλογή προσθήκης της διάστασης «Χρόνος» στον κύβο που δημιουργήθηκε.

Ο κύβος παρέχει τις ευκολίες ανάλυσης των στοιχείων που έχουν εισαχθεί από την Βάση Δεδομένων. Η προπαρασκευή των δεδομένων απαιτούσε τον μετασχηματισμό των σχολίων σε αριθμό, των κατηγοριών σε αριθμό ώστε να πραγματοποιηθεί η διαστασιοποίηση. Ο κύβος υλοποιήθηκε σε σχήμα αστέρα με διαστάσεις τον χρόνο, την κατηγορία, τον αριθμό σχολίων, το blog και τον αριθμό των φιλικών blogs.

5.2 Βάση Δεδομένων

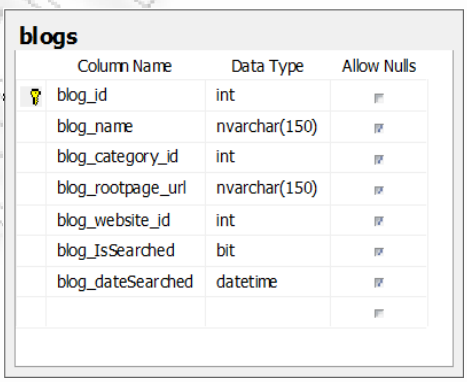
Η Βάση Δεδομένων δημιουργήθηκε με βάση τις ανάγκες του συστήματος, με βασικό πίνακα τον πίνακα «Posts», η αναζήτηση ενός post-δημοσίευσης ξεκινά από την αναγνώριση του βασικού παρόχου που ανήκει δηλαδή «blogspot», «wordpress», «pblogs».



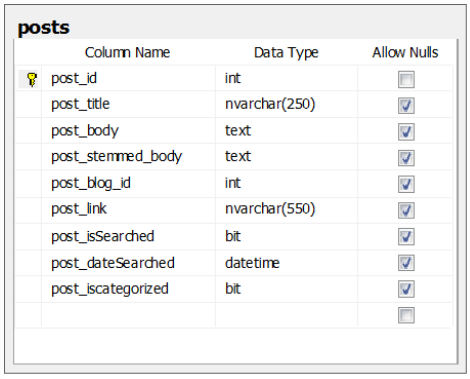
Εικόνα 21

Ο πίνακας websites προς το παρόν περιέχει τρείς βασικούς παρόχους και περιέχει τις απαραίτητες πληροφορίες για να γίνει σωστά η ανάγνωση και εύρεση των posts σε κάθε blog.

Στη συνέχεια, αναγνωρίζονται τα posts μέσα από τα blogs που έχουν καταχωρηθεί στη βάση αρχικά.



Εικόνα 22

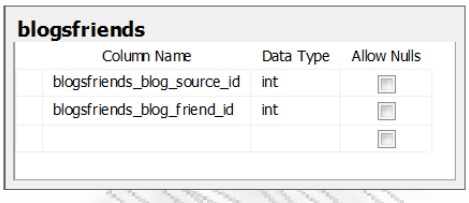


posts

Column Name	Data Type	Allow Nulls
post_id	int	☐
post_title	nvarchar(250)	☑
post_body	text	☑
post_stemmed_body	text	☑
post_blog_id	int	☑
post_link	nvarchar(550)	☑
post_isSearched	bit	☑
post_dateSearched	datetime	☑
post_iscategorized	bit	☑
		☐

Εικόνα 23

Μαζί με τα posts αναγνωρίζονται και όλα τα φιλικά blogs για να ερευνηθούν αργότερα.

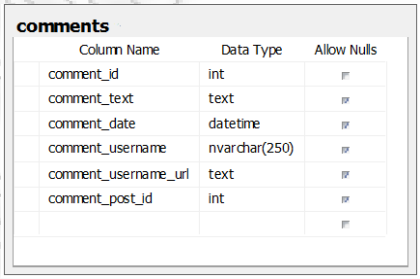


blogsfriends

Column Name	Data Type	Allow Nulls
blogsfriends_blog_source_id	int	☐
blogsfriends_blog_friend_id	int	☐
		☐

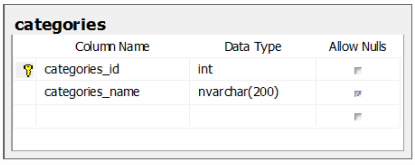
Εικόνα 24

Προχωρώντας στην ανίχνευση του Blog αρχίζει η εύρεση όλων των post και για κάθε post αποθηκεύονται τα σχόλια που έγιναν καθώς και οι κατηγορίες που ανήκει σύμφωνα με τον συγγραφέα.



comments

Column Name	Data Type	Allow Nulls
comment_id	int	☐
comment_text	text	☐
comment_date	datetime	☐
comment_username	nvarchar(250)	☐
comment_username_url	text	☐
comment_post_id	int	☐
		☐

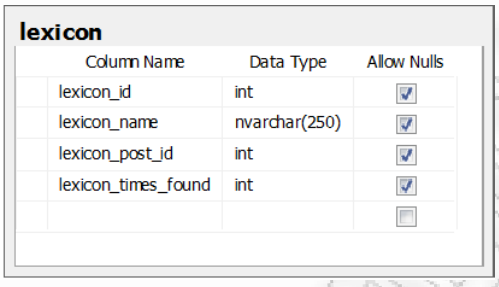


categories

Column Name	Data Type	Allow Nulls
categories_id	int	☐
categories_name	nvarchar(200)	☐
		☐

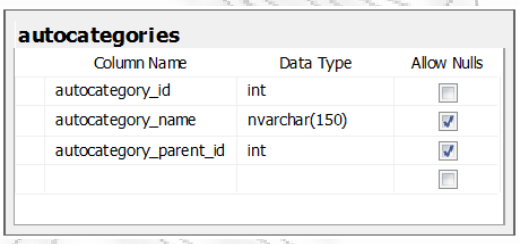
Εικόνα 25

Στη συνέχεια, το κάθε post υφίσταται επεξεργασία φυσικής γλώσσας όπου συλλέγονται οι λέξεις κλειδιά και καταχωρούνται στον πίνακα lexicon.



Εικόνα 26

Τέλος, ο αλγόριθμος αυτόματης κατηγοριοποίησης τοποθετεί το post στην ανάλογη κατηγορία με βάση τις λέξεις που περιέχει και που ακολουθούν δενδρική μορφή.



Εικόνα 27

5.3 Αποθήκη δεδομένων ιστολόγιων

Η Αποθήκη Δεδομένων είναι μία βάση δεδομένων με προσανατολισμό στην από-κανονικοποίηση της ήδη υπάρχουσας βάσης και την υλοποίησή της έτσι ώστε να συγκεντρώνει τα απαραίτητα για ανάλυση δεδομένα σε ένα πίνακα.

Ο πίνακας γεγονότων αποφασίστηκε να είναι μία προβολή με βάση τον πίνακα “posts” διότι περιέχει τις απαραίτητες πληροφορίες που μπορούν να εμπλουτιστούν και στη συνέχεια να αναλυθούν. Η επιλογή έγινε για να συμπτυχθούν τα στοιχεία των υπόλοιπων πινάκων στα απαραίτητα και να δημιουργήσουμε τις διαστάσεις του κύβου. Η αποθήκη δεδομένων αναπτύχθηκε με το εργαλείο Microsoft Analysis Services σε Visual Studio 2008.

Fact Table: Η προβολή «vFact» περιέχει τα κλειδιά από όλους του πίνακες που θα δημιουργηθούν ως διαστάσεις δηλαδή το “post_id” από τον πίνακα “posts”, το “blog_id” από τον πίνακα “blogs”, το “comment_id” από τον πίνακα “comments”, το “PK_Date” είναι ένα επιπλέον κλειδί σε έναν πίνακα που δημιουργήθηκε ειδικά για τις ανάγκες τις ανάλυσης στην αποθήκη δεδομένων σχετικά με τον χρόνο και το “categories_id” που συνδέει το κάθε post με τις κατηγορίες που έχει.

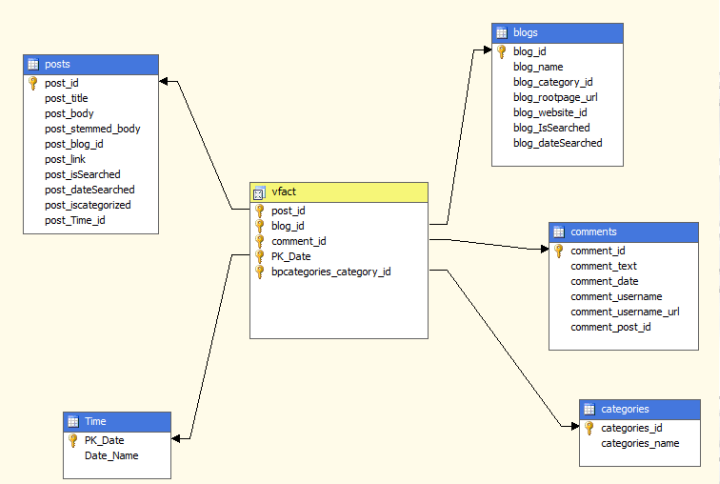
Dimension “Posts”: Η διάσταση posts χρησιμεύει για την εμφάνιση των posts.

Dimension “Blogs”: Η διάσταση blogs χρησιμεύει για την εμφάνιση των ιστολόγιων που ανήκουν τα posts.

Dimension “Comments”: Η διάσταση comments χρησιμεύει για την εμφάνιση των σχολίων που έχουν γίνει σε κάθε post.

Dimension “Time”: Η διάσταση time χρησιμεύει για την εμφάνιση των posts σε περίπτωση που θέλουμε να τα συσχετίσουμε με τον χρόνο.

Dimension “Categories”: Η διάσταση categories χρησιμεύει για την εμφάνιση των κατηγοριών που τα posts ανήκουν.


posts
post_id
post_title
post_body
post_stemmed_body
post_blog_id
post_link
post_isSearched
post_dateSearched
post_iscategorized
post_Time_id
blogs
blog_id
blog_name
blog_category_id
blog_rootpage_url
blog_website_id
blog_IsSearched
blog_dateSearched
vfact
post_id
blog_id
comment_id
PK_Date
bpcategories_category_id
comments
comment_id
comment_text
comment_date
comment_username
comment_username_url
comment_post_id
Time
PK_Date
Date_Name
categories
categories_id
categories_name

Εικόνα 28

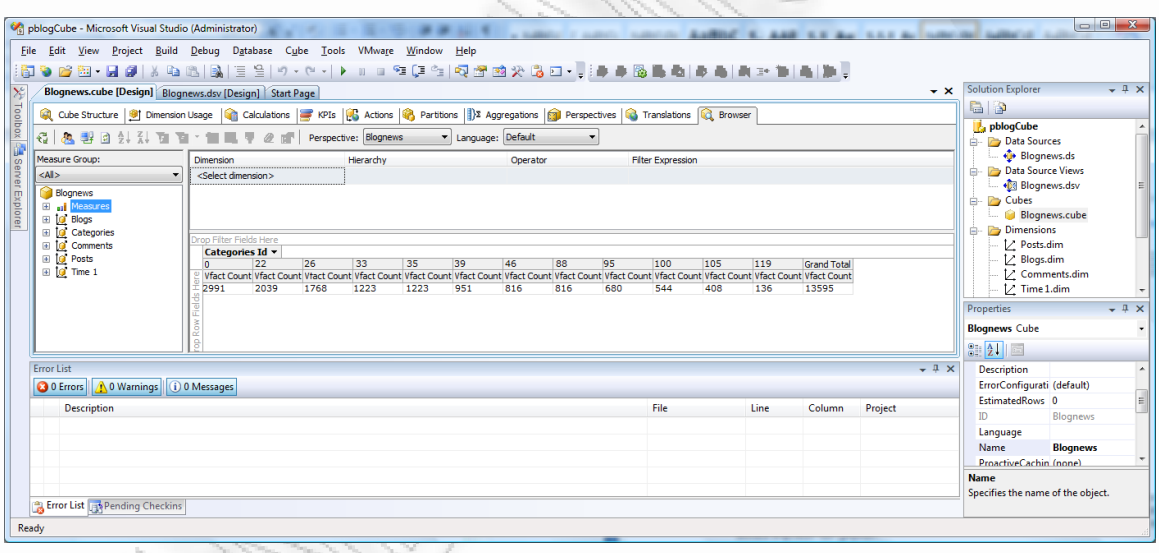
6 Παραδείγματα

6.1 Γραφήματα

Στα παρακάτω γραφήματα εμφανίζονται κάποια ενδεικτικά αποτελέσματα από την Αποθήκη Δεδομένων που έχει δημιουργηθεί. Η εικόνα που εμφανίζεται είναι για κάποια ποσοστά που προέκυψαν βάσει των αρχικών υποθέσεων που είχαν τεθεί για το σκοπό της παρούσης διπλωματικής.

6.1.1 Δημοσιεύσεις ανά κατηγορία

Κάθε δημοσίευση ανήκει σε μία κατηγορία που έχει συνδεθεί αυτόματα από το σύστημα κατηγοριοποίησης. Από τις δημοσιεύσεις που έχουμε αυτή τη στιγμή στην αποθήκη δεδομένων τα σύνολα που προέκυψαν από τις 12 βασικότερες κατηγορίες είναι τα παρακάτω:



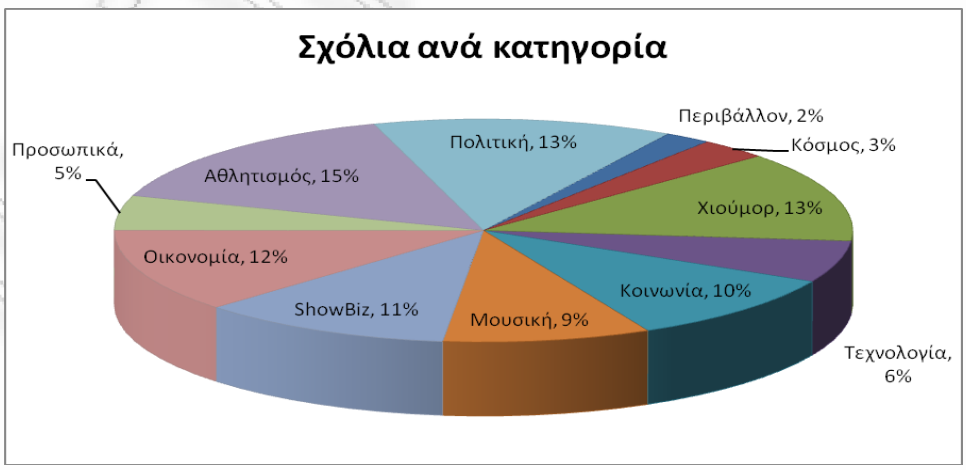
Εικόνα 29

Με βάση τα παραπάνω σύνολα προκύπτει η παρακάτω στατιστική πίτα:

Εικόνα 30

6.1.2 Σχόλια-Δημοσιεύσεις ανά ημέρα

Τα σχόλια που γίνονται σε κάθε δημοσίευση έχει ενδιαφέρον να εμφανιστούν για κάθε κατηγορία, τα αποτελέσματα συνοψίζονται στο παρακάτω γράφημα. Παρατηρούμε ότι τα περισσότερα σχόλια γίνονται στις δημοσιεύσεις που ανήκουν στις κατηγορίες «Πολιτική», «Αθλητισμός» και «Χιούμορ».



Εικόνα 31

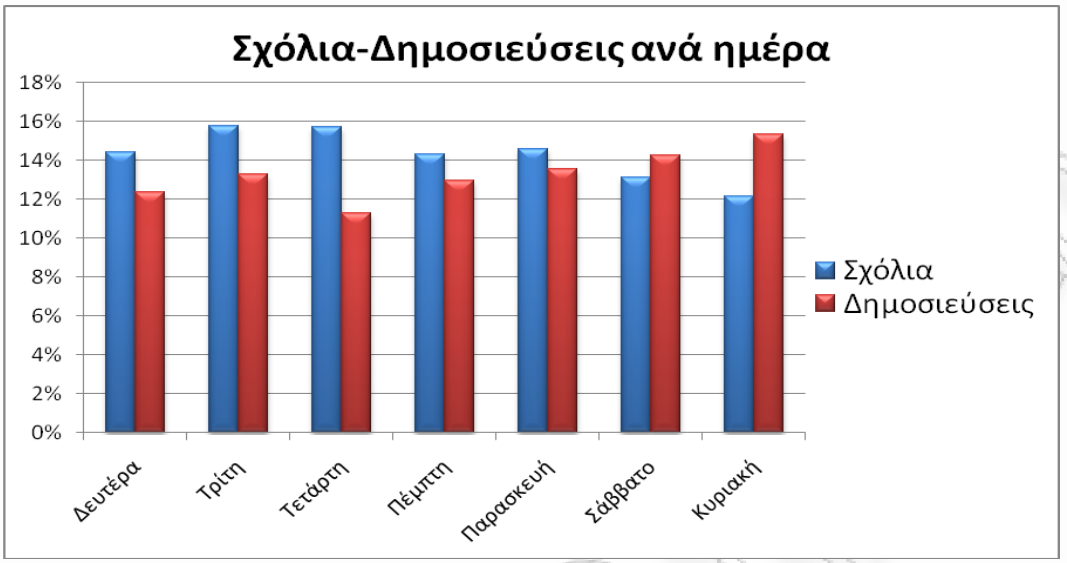
6.1.3 Σχόλια ανά ώρα σε περίοδο μιας μέρας

Σύμφωνα με τα στοιχεία της Αποθήκης Δεδομένων προκύπτουν τα εξής στοιχεία σχετικά με το ποιά ώρα της ημέρας γίνονται τα περισσότερα σχόλια.

Εικόνα 32

6.1.4 Σχόλια-Δημοσιεύσεις ανά ημέρα

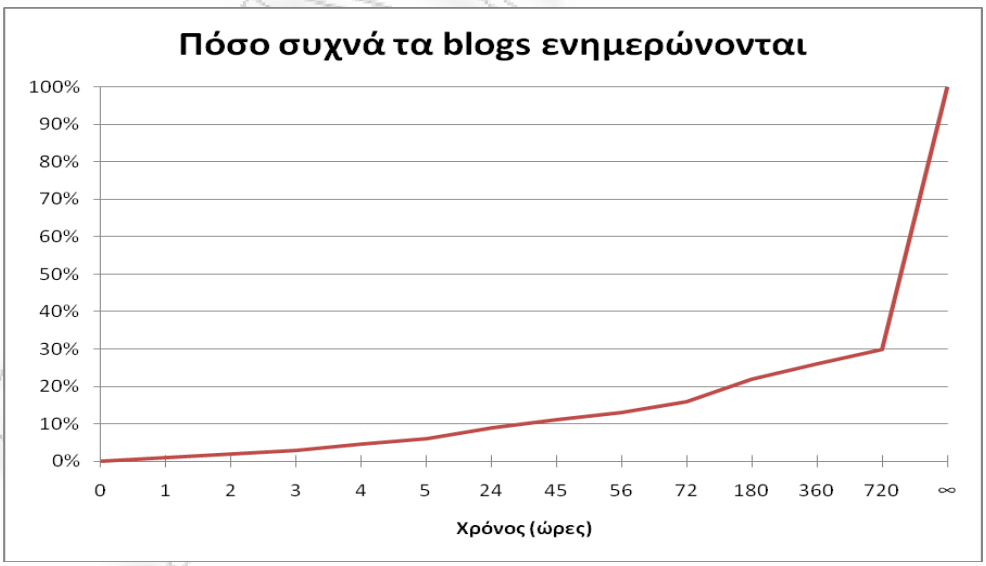
Το παρακάτω γράφημα απεικονίζει το ποσοστό των σχολίων και δημοσιεύσεων που γίνονται σε περίοδο μιας εβδομάδας και εμφανίζεται η χαρακτηριστική αύξηση των σχολίων στην αρχή της εβδομάδας και η αύξηση των δημοσιεύσεων προς το τέλος της.



Εικόνα 33

6.1.5 Σχόλια-Δημοσιεύσεις ανά ημέρα

Το παρακάτω γράφημα εμφανίζει τα στοιχεία σχετικά με την ενημέρωση των ιστολόγιων και την περίοδο που ανανεώνονται. Παρατηρείται, ότι το μεγαλύτερο ποσοστό των ιστολόγιων ενημερώνεται ανά εβδομάδα ή ανά μήνα.



Εικόνα 34

6.1.6 Πλήθος σχολίων από την ώρα δημοσίευσης

Επίσης, ένα πολύ σημαντικό στατιστικό στοιχείο είναι η περίοδος που μεσολαβεί και εμφανίζονται τα περισσότερα σχόλια σε κάθε δημοσίευση. Παρατηρείται, ότι σε διάστημα από 6 μέχρι 12 ώρες είναι τα περισσότερα σχόλια.

Εικόνα 35

6.1.7 Πλήθος σχόλιων και δημοσιεύσεων ανά κατηγορία

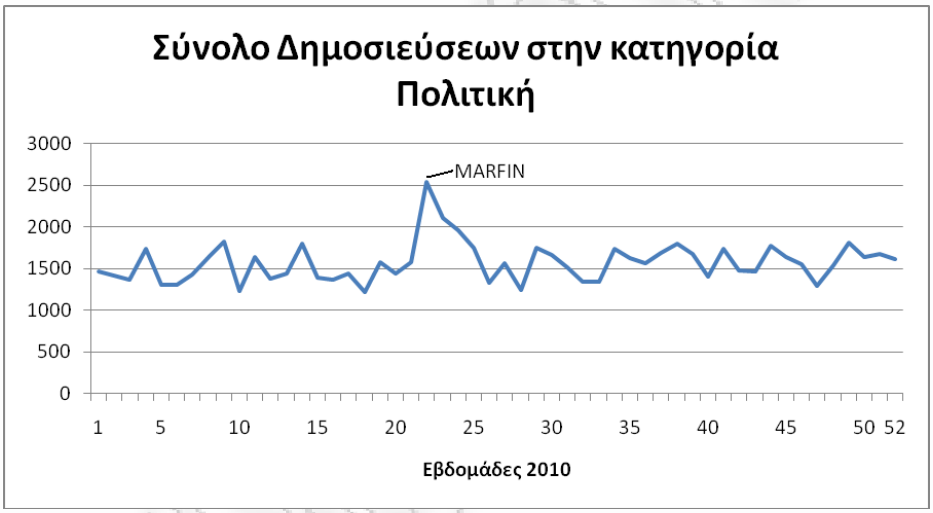
Στο παρακάτω γράφημα εμφανίζεται το σύνολο των σχολίων και δημοσιεύσεων ανά κατηγορία.

Εικόνα 36

6.1.8 Συσχέτιση δημοσιεύσεων με πραγματικά γεγονότα

Πιο συγκεκριμένα, για ένα γεγονός που πραγματοποιήθηκε τον Μάιο του 2010, που είναι ο θάνατος τριών εργαζομένων της τράπεζας Marfin Bank κατά τη διάρκεια πορείας διαμαρτυρίας στο κέντρο της Αθήνας, το παρακάτω γράφημα εμφανίζει το πλήθος των δημοσιεύσεων για την κατηγορία «Πολιτική» για όλο το έτος 2010, περίοδο 52 εβδομάδων.

Παρατηρούμε ότι την περίοδο που έγινε το γεγονός εμφανίζεται άνοδος των δημοσιεύσεων και διατηρείται για κάποιο διάστημα μερικών εβδομάδων.

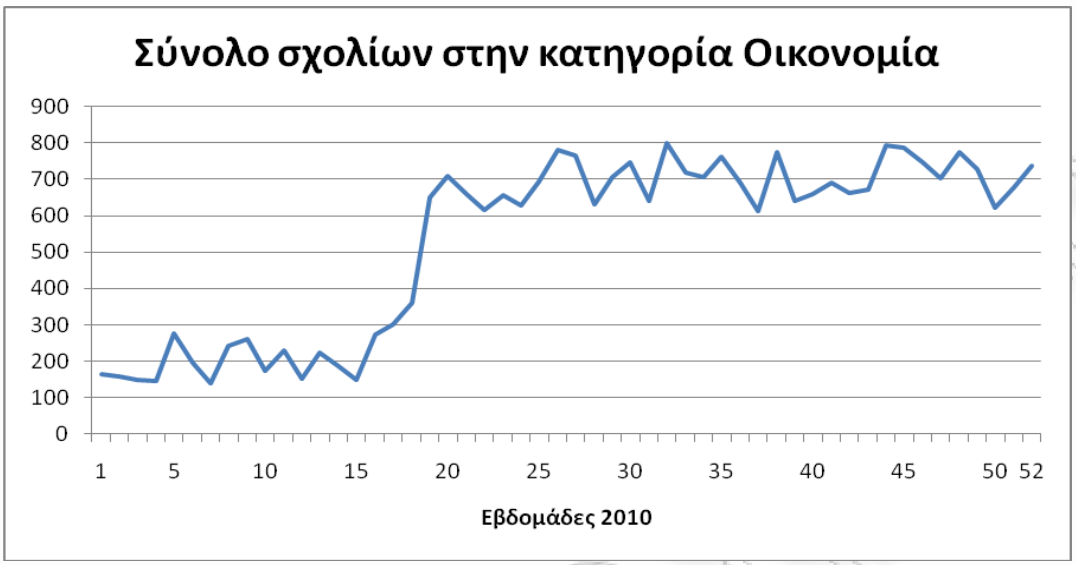


Εικόνα 37

6.1.9 Συσχέτιση σχολίων με πραγματικά γεγονότα

Αντίστοιχα το σύνολο των σχολίων για την κατηγορία «Οικονομία» φαίνεται να αυξάνεται όταν αρχίζει να εμφανίζεται η προοπτική του μνημονίου για την οικονομική Πολιτική της Ελλάδας, περίπου την περίοδο Απρίλιο-Μάιο.

Παρατηρούμε ότι την περίοδο που έγινε το γεγονός εμφανίζεται άνοδος των σχολίων και διατηρείται για όλη την υπόλοιπη περίοδο του έτους.


Σύνολο σχολίων στην κατηγορία Οικονομία
900
800
700
600
500
400
300
200
100
0
1
5
10
15
20
25
30
35
40
45
50
52
Εβδομάδες 2010

Εικόνα 38

7 Σύνοψη και Συμπεράσματα

Σύμφωνα με τις τελευταίες εξελίξεις στον τομέα της ειδησεογραφίας και της ενημέρωσης, παρατηρείται η ανάπτυξη των ηλεκτρονικών μέσων πληροφόρησης και ειδικότερα των ιστολόγιων. Η διαδραστικότητα που προσφέρουν τα ιστολόγια με τη δυνατότητα σχολιασμού συντελεί στην ανάπτυξή τους. Μέσω της εφαρμογής καταφέρνουμε να αποτυπώσουμε ένα κομμάτι της μπλογκόσφαιρας και να καταγράψουμε τις τάσεις και τα θέματα που αναπτύσσονται.

Οι τεχνολογίες υπηρεσίες ιστού (Web Services) καθώς και η ανάπτυξη των αποθηκών δεδομένων βοήθησε στον να ολοκληρωθεί η συγκεκριμένη εργασία. Μέσα από τις υπηρεσίες που αναπτύχθηκαν ο χρήστης μπορεί να αναζητήσει δημοσιεύσεις και σχόλια για κάθε ιστολόγιο καθώς και να αντλήσει πληροφορίες που έχουν προκύψει ύστερα από ανάλυση και επεξεργασία στην Αποθήκη Δεδομένων.

Οι υπομονάδες που αποτελούν την πλατφόρμα που βασίστηκε η εργασία είναι:

1. ένας εξυπηρετητής ιστού ο οποίος είναι υπεύθυνος για την συγκέντρωση των δημοσιεύσεων και σχολίων τους από όλα τα ιστολόγια,
2. ένας εξυπηρετητής για την επεξεργασία των κειμένων που έχουν συλλεχθεί όπως επεξεργασία φυσικής γλώσσας και κατηγοριοποίηση των δημοσιεύσεων,
3. μία βάση δεδομένων SQL Server για την αποθήκευση των δεδομένων καθώς συλλέγονται και αφού επεξεργαστούν από τον κατηγοριοποιητή και
4. μία Αποθήκη Δεδομένων για την εξόρυξη γνώσης από τα δεδομένα.

Σύμφωνα με τα παραδείγματα που παρατέθηκαν παραπάνω εμφανίζονται αξιόλογα στατιστικά σχετικά με τις δημοσιεύσεις και τα σχόλια συναρτήσει των κατηγοριών. Παρατηρείται ότι σε γεγονότα με υψηλό ενδιαφέρον υπάρχει εξαιρετική αύξηση της δραστηριότητας στην μπλογκόσφαιρα, τόσο στις δημοσιεύσεις όσο και στα σχόλια.

Η εργασία που αναπτύχθηκε και υλοποιήθηκε είχε ως σκοπό την δημιουργία μιας ενιαίας πλατφόρμας συλλογής, επεξεργασίας και αποθήκευσης στοιχείων από την ελληνική κοινότητα ιστολογίων. Τα παραδείγματα που δόθηκαν ήταν ενδεικτικά της περαιτέρω ανάλυσης που μπορεί να υποβληθεί στην Αποθήκη Δεδομένων και στα στοιχεία που περιέχει. Οι μελλοντικές εργασίες θα μπορούσαν να ασχοληθούν ειδικότερα με γεγονότα και την αντίστοιχη δραστηριότητα που παρουσιάζεται καθώς και με επιπλέον επεξεργασία των δεδομένων που έχουν αποθηκευθεί.

8 Βιβλιογραφία

Κατά τη συγγραφή του παρόντος τόμου και του συστήματος της εφαρμογής χρησιμοποιήθηκε η βιβλιογραφία που ακολουθεί :

[AM05] Blog mining in a corporate environment, Andreas Aschenbrenner, Silvia Miksch, September 2005.

[BM06] Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering , Christopher H. Brooks and Nancy Montanez, WWE 2006.

[OKN06] Browsing System for Weblog Articles based on Automated Folksonomy ,Tsutomu Ohkura, Yoji Kiyota, Hiroshi Nakagawa, WWE 2006.

[AZA04] Implicit structure and the dynamics of blogspace, Eytan Adar, Li Zhang, Lada A. Adamic, Rajan M. Lukose, Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, May 2004.

[EF04] Learning webs: Learning in weblog networks , L. Efimova, S. Fiedler , Proceedings of the IADIS International Conference Web Based Communities 2004, Lisbon, March 2004 , IADIS Press.

[Rol04] Business Weblogs - A pragmatic Approach to introducing Weblogs in medium and large Enterprises , Martin Röll, BlogTalk, Vienna Austria, 2004.

[MG06] Leave a Reply: An Analysis of Weblog Comments , Gilad Mishne and Natalie Glance, WWE 2006.

[IKT05] Text Classification Using Machine Learning Techniques, M. Ikonomakis, S. Kotsiantis, V. Tampakas, WSEAS TRANSACTIONS on COMPUTERS, August 2005.

[ACK04] Web Services Concepts, Gustavo Alonso, Fabio Casati, Harumi Kuno and Vijay Machiraju, Architectures and Applications, Springer, 2004.

[FMU04] Programming Microsoft .NET XML Web Services, Damien Foggon, Daniel Maharry, Chris Ullman and Karli Watson, Microsoft Press, 2004.