ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εξαγωγή Γεωγραφικής Πληροφορίας από Ημιδομημένο Κείμενο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΛΒΕΡΤΟΥ-ΔΑΥΪΔ Α. ΑΝΤΖΕΛ

Επιβλέπων : Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2006

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΈΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εξαγωγή Γεωγραφικής Πληροφορίας από Ημιδομημένο Κείμενο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΑΛΒΕΡΤΟΥ-ΔΑΥΪΔ Α. ΑΝΤΖΕΛ

Επιβλέπων : Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Ιουλίου 2006.

.....	.....	.....
Τιμολέων Σελλής	Ιωάννης Βασιλείου	Κων/νος Σαγώνας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπλ. Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2006

.....

ΑΛΒΕΡΤΟΣ-ΔΑΥΪΔ Α. ΑΝΤΖΕΛ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η αναζήτηση, πλοήγηση, ευρετηριοποίηση, οργάνωση του παγκόσμιου ιστού μπορεί να γίνει πάνω σε διάφορους άξονες χαρακτηριστικών (π.χ. με λέξεις κλειδιά, τοπολογικά-μέσω υπερσυνδέσμων, θεματικά, χρονολογικά, γεωγραφικά). Στην διπλωματική εργασία αυτή, μελετώνται διάφορες προσεγγίσεις ανακάλυψης γεωγραφικής-χωρικής πληροφορίας σε ιστοσελίδες (geoparsing), και απόδοσης ακριβών συντεταγμένων στην πληροφορία αυτή (γεωκωδικοποίηση).

Για την πραγματοποίηση των στόχων αυτών, απαιτούνται αποδοτικοί αλγόριθμοι προσεγγιστικού και φωνητικού ταιριάγματος συμβολοσειρών (για παράδειγμα, για το ταίριαγμα πιθανών τοπωνυμίων από μια ιστοσελίδα με μια μεγάλη βάση τοπωνυμίων, ή για τον καθαρισμό των δεδομένων στης βάσης αυτής). Γι'αυτό, εξετάζονται οι υπάρχοντες αλγόριθμοι ταιριάγματος, και προτείνονται νέοι, καθώς και παραλλαγές τους για την Ελληνική γλώσσα. Οι ιδιοτροπίες που προκύπτουν από την χρήση της τελευταίας (π.χ. ύπαρξη μεγάλου όγκου πληροφορίας σε greeklish), επισημαίνονται, παράλληλα με τους τρόπους αντιμετώπισής τους.

Επίσης, για έναν τέτοιο στόχο, απαιτούνται υψηλής ποιότητας γεωγραφικά δεδομένα. Προτείνονται διάφορες μέθοδοι απόκτησης και καθαρισμού αυτών, προσαρμοσμένες στην Ελληνική πραγματικότητα.

Τέλος, αναπτύσσουμε ένα πρωτότυπο εργαλείο για την γεωγραφική ευρετηριοποίηση του Ελληνικού ιστοχώρου, που υλοποιεί τις παραπάνω ιδέες, και προσφέρεται για ποικίλες εφαρμογές (π.χ. γεωγραφική αναζήτηση, εύρεση σημείων ενδιαφέροντος στην εγγύτητα κ.λ.π.)

Λέξεις-Κλειδιά: εξαγωγή γεωγραφικής πληροφορίας, γεωκωδικοποίηση, αλγόριθμοι προσεγγιστικού και φωνητικού ταιριάγματος, καθαρισμός δεδομένων, γεωγραφική αναζήτηση

Abstract

Web pages may be organized, indexed, searched, and navigated along several different feature dimensions (e.g. keywords, theme, geography, time). In this thesis we investigate different approaches of discovering spatial context for web pages (geoparsing), as well as for providing accurate coordinates for said spatial context (geocoding).

For both of these goals to be realised, efficient algorithms for imprecise and phonetic string matching are needed (e.g. for matching potential feature names encountered in a web page with a large feature name database, or for cleaning geographic data in said database). Thus, existing matching algorithms are examined. In addition, several new variants are proposed, and are customised for use in a Greek context; the peculiarities presented by the latter are explored as well.

Furthermore, high-quality geographic datasets are required for such a task. Alternative methods of obtaining and cleaning these are presented.

Finally, a prototype tool for the geographic indexing of the Greek web is developed, implementing the aforementioned concepts and allowing for a multitude of applications (e.g. searching, or ranking search results by geographic relevance, finding points of interest in the vicinity e.t.c.)

Keywords: geoparsing, geocoding, imprecise and phonetic string matching, data cleaning, geographic information extraction

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους τους ανθρώπους που με διάφορους τρόπους με βοήθησαν στην εκπόνηση αυτής της εργασίας.

Κατ'αρχήν, τον επιβλέποντα καθηγητή, κ. Τ. Σελλή, του οποίου η διδακτική, και η γενικότερη υποδειγματική στάση ως καθηγητή, με ώθησαν να ασχοληθώ με τον κλάδο των Βάσεων Δεδομένων - η βοήθειά του τόσο σε επίπεδο διπλωματικής εργασίας, όσο και σε ζητήματα διαμόρφωσης επιλογών και σχεδίων για την μετέπειτα ακαδημαϊκή μου πορεία, υπήρξε καθοριστική.

Επίσης, τον συνεπιβλέποντα της εργασίας, κ. D.Pfoser, PhD., για την εποικοδομητική συνεργασία που είχαμε, στα πλαίσια της διπλωματικής. Ο κ. Pfoser αφιέρωσε με ενθουσιασμό, προσωπικό του χρόνο, για την καθοδήγηση, διαμόρφωση, και γενικότερη υποστήριξη της εργασίας αυτής. Έτσι, αναπτύχθηκε μεταξύ μας μια άριστη διαλεκτική και επικοινωνία, στην οποία οφείλεται, πιστεύω, μέγαλο μέρος της -όποιας- επιτυχίας του εγχειρήματος. Πέραν των ουσιαστικών συζητήσεων επί της εργασίας που είχαμε, ο κ.Pfoser συντέλεσε καταλυτικά στην διαμόρφωση του προσανατολισμού και των ενδιαφερόντων μου, στο πεδίο των Βάσεων Δεδομένων.

Τέλος, είναι γεγονός πως αυτή η εργασία δεν θα είχε φθάσει σε αίσιο τέλος, δίχως την απεριόριστη και ποικιλότροπη στήριξη που έλαβα από την συνάδελφο και φίλη Α.Κούρτη. Η Αμαλία συνέβαλε με χρήσιμες ιδέες και προτάσεις κατά την εκπόνηση της εργασίας -άλλωστε μέρος της βασίζεται σε παλαιότερη εργασία που είχαμε εκπονήσει από κοινού- , βοήθησε αποφασιστικά στην διαμόρφωση και διόρθωση του τόμου αυτού, ενώ πολύτιμη μου στάθηκε η συνεχής υποστήριξή της όλον αυτόν τον καιρό.

Πίνακας περιεχομένων

1 *Εισαγωγή*

Η οικογένεια Σπανοδημήτρη θέλει να πάει για σκι στον Παρνασσό. Αναλαμβάνει, λοιπόν η μεγαλύτερη κόρη, η Αφροξυλάνθη (φοιτήτρια του Ε.Μ.Π.) να ψάξει για σχετικές πληροφορίες στο διαδίκτυο: Καταλύματα, χιονοδρομικά κέντρα, αλλά και γενικότερες πληροφορίες για την περιοχή. Γράφει λοιπόν στο Google™ μια ερώτηση της μορφής: +σκι +παρνασσός, ή +εστιατόρια +παρνασσός. Προς μεγάλη έκπληξή της, βρίσκει δεκάδες μάλλον άσχετα αποτελέσματα, όπως το κατάστημα ειδών σκι "Ο Παρνασσός" που εδρεύει στην Ομόνοια, το ψητοπωλείο "Ο Παρνασσός" στα Εξάρχεια ή και τον διάσημο τενόρο Άλαν Σκί που τραγουδά για μια φιλανθρωπική συναυλία στην αίθουσα "Παρνασσός".

Παραπλήσια προβλήματα αντιμετωπίζει και ο θεατρόφιλος Χανς που ψάχνει ένα θέατρο στην γειτονιά του, τον Ζωγράφο, που να παίζει τον "Οθέλλο".

Ο γνωστός προγραμματιστής Douda G. αναπτύσσει μια εφαρμογή που παρέχει υπηρεσίες βασισμένες στη θέση (Location Based Services). Αποφασίζει πως χρειάζεται ένα ψηφιακό χάρτη με σημεία ενδιαφέροντος στην Χαλκίδα (κέντρα διασκέδασης, πάρκα, εστιατόρια, φαρμακεία κ.ο.κ.), ο οποίος δυστυχώς δεν διατίθεται στην αγορά.... Επειδή η συλλογή αυτών των δεδομένων μπορεί να είναι αρκετά δύσκολη ή δαπανηρή, θα ήθελε έναν εύκολο, αυτόματο τρόπο να τα συλλέξει, ει δυνατόν από δημόσια, δωρεάν διαθέσιμες πληροφορίες (π.χ. το διαδίκτυο).

Η Ζουμπουλία, ο Φώτης, ο Σπύρος, η Ντάλια, και η Αγγέλα ερευνώντας έναν φόνο, ψάχνουν στο διαδίκτυο για να βρουν πληροφορίες για παράνομες πράξεις που διαπράχθηκαν στην περιοχή της Θεσσαλονίκης πριν από χρόνια. Επειδή οι ειδησεογραφικές ιστοσελίδες δεν ταξινομούν τις παλιές ειδήσεις με βάση τον χώρο, ψάχνοντας για: Θεσσαλονίκη φόνος, οι πέντε δεν θα βρουν το δημοσίευμα που ψάχνουν (που αναφέρεται σε έναν φόνο στα Λαδάδικα)....

1.1 Ανάγκες που οδηγούν την εργασία (Motivation)

1.1.1 Εναλλακτική αλληλεπίδραση με τον παγκόσμιο ιστό

Σε τέτοια προβλήματα, αλλά και άλλα πιο γενικά, παρατηρούμε την αδυναμία των υπαρχουσών μεθόδων αλληλεπίδρασης με τον παγκόσμιο ιστό. Πράγματι, ο παγκόσμιος ιστός αποτελεί μια τεράστια πηγή πληροφοριών, με συγκεκριμένους και περιορισμένους, όμως, τρόπους αλληλεπίδρασης. Ο κυριότερος, που είναι η αναζήτηση με λέξεις κλειδιά, έχει εμφανή μειονεκτήματα - ορισμένα διαφαίνονται στα ανωτέρω παραδείγματα. Θα μπορούσε πει κανείς ότι τα κενά αυτά καλύπτονται από τα θεματικά ευρετήρια (Web directories). Τα ευρετήρια, δυστυχώς, έχοντας δημιουργηθεί από ανθρώπους, κατηγοριοποιούν μονάχα μικρό όγκο πληροφορίας, ανανεώνονται δε με βραδύ ρυθμό, που αδυνατεί να καλύψει τις ταχείες αλλαγές στον παγκόσμιο ιστό.

Συνεπώς, για να μπει μια τάξη στο "χάος", θα ήταν εποικοδομητική η χρήση μιας άλλης προσέγγισης στην αλληλεπίδραση με τον παγκόσμιο ιστό, πέραν των υπαρχουσών. Ο οποίος ιστός μπορεί, και έχει νόημα, να οργανωθεί, να ευρετηριοποιηθεί, να αναζητηθεί και να πλοηγηθεί πάνω σε αρκετούς άξονες χαρακτηριστικών. Τέτοιοι είναι, για παράδειγμα, ο χρόνος, η γεωγραφική θέση, το θέμα κ.ο.κ. Έχουν, όντως, προταθεί τελευταία λύσεις για την αυτόματη θεματική κατηγοριοποίηση του παγκόσμιου ιστού (κάποιες στα: [HNV+03], [SCY+04], [Cir04], [KSS+03]), όπως και για την γεωγραφική κατηγοριοποίησή του (μερικές ενδιαφέρουσες στα: [Mcc01], [WA03], [Clo05], [JPR+02], [DGS00], [BLB+03], [BCG+99]).

1.1.2 Χτίζοντας σημασιολογικές υπηρεσίες πάνω στον (ημιδομημένο) ιστό

Βλέποντας λίγο πιο μακριά, ας θεωρήσουμε λίγο τις σημερινές τάσεις σχετικά με τον παγκόσμιο ιστό. Από την μία, ο σημασιολογικός ιστός (Semantic Web [LHL01]) οδηγεί σε μια νεα φιλοσοφία δημιουργίας περιεχομένου για τον ιστό. Από την άλλη, καταννοώντας ότι δεν είναι δυνατή η αναδόμηση όλου του ιστού εκ του μηδενός, ώστε να επέλθει ο σημασιολογικός ιστός, οδηγούμαστε στον ορισμό δομής και υπηρεσιών πάνω στον παγκόσμιο ιστό. Κατά τα λεγόμενα του AnHai Doan: "Μακροπρόθεσμα, πιστεύουμε πως τα συστήματα διαχείρησης πληροφοριών της επόμενης γενιάς, πρέπει να χειρίζονται εξίσου δομημένα και κειμενικά δεδομένα, να παρέχουν πολλαπλές υπηρεσίες, από αναζήτηση με λέξεις-κλειδιά εως ερωτήματα SQL, και να αλληλεπιδρούν ευφυώς με τους χρήστες. Ο στόχος μας είναι να αναπτύξουμε τέτοια συστήματα, παίρνοντας τεχνικές από τις Βάσεις Δεδομένων, την Ανάκτηση Πληροφορίας, τον Παγκόσμιο Ιστό, την Εξόρυξη Δεδομένων, και τον χώρο της Τεχνητής Νοημοσύνης"[1]. Η εξόρυξη γεωγραφικής πληροφορίας από ημιδομημένο κείμενο

[1] "In the long term, we believe that next-generation information management systems must handle both structured and textual data, provide multiple services, ranging from keyword search to SQL querying, and

(όπως είναι οι ιστοσελίδες) εντάσσεται, πιστεύουμε, σε αυτήν την προσπάθεια, επιχειρώντας να προσδώσει δομή και τοπολογία στον παγκόσμιο ιστό, με τρόπο αυτόματο.

1.2 Αντικείμενο της διπλωματικής

Σε αυτήν την διπλωματική εργασία, προτείνουμε μια λύση σε ανάλογα προβλήματα, εκμεταλλευόμενοι την γεωγραφική πληροφορία που μπορεί να έχει μια ιστοσελίδα. Η επιλογή της γεωγραφικής διάστασης (και όχι, για παράδειγμα, της θεματικής) δεν είναι τυχαία, καθώς πολλές σελίδες περιέχουν πλήθος γεωγραφικών πληροφοριών, ή έχουν έντονη γεωγραφική χροιά. (Σύμφωνα με τον [Mcc01], σχεδόν 10% των σελίδων περιέχουν προφανή γεωγραφικά στοιχεία).

Πιο συγκεκριμένα, ξεκινώντας από μια ιστοσελίδα, επιχειρούμε να της αποδώσουμε μία ή περισσότερες γεωγραφικές συντεταγμένες, με βάση το περιεχόμενό της. Προφανώς, η πληροφορία εξάγεται με τρόπο αυτόματο, ώστε να είναι εφικτή η λύση μας. Το αποτέλεσμα της διαδικασίας αυτής είναι ένα γεωγραφικό ευρετήριο για τις ιστοσελίδες που εξετάστηκαν. Τέλος, επικεντωθήκαμε στον ελληνικό κυβερνοχώρο (ο οποίος άλλωστε πάσχει από ένδεια τέτοιων υπηρεσιών) γεγονός που, όπως θα δούμε, δημιουργεί επιπλέον προβλήματα, με ιδιαίτερο ενδιαφέρον.

1.2.1 Έννοιες

Κεντρικές έννοιες στην εργασία αυτή είναι το geoparsing, και η γεωκωδικοποίηση (geocoding). Ακολουθεί σύντομη περιγραφή τους - λεπτομερής εξέτασή τους, σε πολλαπλά επίπεδα, γίνεται στα προσεχή κεφάλαια.

1.2.1.1 Geoparsing

Το geoparsing[2] ([Mcc01]) σημαίνει την λεκτική και γραμματική ανάλυση μιας ιστοσελίδας, πιθανώς υποβοηθούμενη από κάποιο gazetteer (ευρετήριο τοπωνυμίων), με στόχο την ανακάλυψη πιθανής γεωγραφικής πληροφορίας. Η πληροφορία αυτή μπορεί να εμφανίζεται με διάφορες μορφές (από το επίπεδο του δικτύου, όπως διεύθυνση IP, από επίπεδο κειμένου, όπως τηλέφωνα, ταχυδρομικοί κώδικες, από επίπεδο σημασιολογίας, όπως διευθύνσεις, τοπωνύμια, κατευθύνσεις, από επίπεδο τοπολογίας κ.α.), τις οποίες θα εξετάσουμε ενδελεχώς σε επόμενο κεφάλαιο.

---

interact intelligently with users. Our goal is to develop such systems, leveraging techniques from database, IR, Web, data mining, and AI communities." AnHai Doan, σε προσωπική του αλληλογραφία

[2]Μια μάλλον ατυχής απόδοση του όρου geoparsing στα Ελληνικά είναι ο όρος "γεωτεχνολόγηση", ο οποίος εφεξής θα αποφεύγεται. Περιφραστικά αποδίδεται ως "εξαγωγή γεωγραφικής πληροφορίας μέσω parsing (τεχνολόγησης)"

1.2.1.2 Γεωκωδικοποίηση

Η γεωκωδικοποίηση, από την άλλη, είναι η μετατροπή αυτής της πληροφορίας σε συντεταγμένες, σε κάποιο γεωγραφικό σύστημα αναφοράς. Αν και αυθύπαρκτη διαδικασία, στα πλαίσια της εργασίας αυτής νοείται ως φυσική συνέχεια του geoparsing.

1.3 Οργάνωση του τόμου

Στο κεφάλαιο αυτό είδαμε μέσω μιας (μάλλον χαρωπής) εισαγωγής, γιατί αξίζει να ασχοληθεί κανείς με την εξαγωγή γεωγραφικής πληροφορίας από τον παγκόσμιο ιστό.

Στο κεφάλαιο **2** "Περιγραφή" οριοθετείται πιο αυστηρά ο στόχος της παρούσας εργασίας. Ακόμη, εξετάζουμε σε αδρές γραμμές την αρχιτεκτονική του δημιουργηθέντος συστήματος, και τις αιτίες που οδήγησαν σε αυτήν. Έχοντας παρουσιάσει τις βασικές αρχές του συστήματος, μπορούμε πλέον να εξετάσουμε θεωρητικά και σε επίπεδο υλοποίησης τους διάφορους επιμέρους τομείς του.

Στο **3**ο κεφάλαιο μιλάμε για τις βασικές έννοιες της εργασίας, δηλαδή το geoparsing και την γεωκωδικοποίηση. Ξεκινώντας από το geoparsing, εξετάζουμε συγκριτικά τις διάφορες προσεγγίσεις στο θέμα αυτό, και παρουσιάζουμε την δική μας, μαζί με τους λόγους που οδήγησαν σε αυτήν. Αξιολογούμε, ακόμη, άλλες πηγές γεωγραφικής πληροφορίας που μπορούμε να χρησιμοποιήσουμε. Βλέπουμε, έτσι, τις σχεδιαστικές και αλγοριθμικές επιλογές στο τμήμα αυτό, και ορισμένα στοιχεία επίδοσης.

Συνεχίζουμε με την φυσική συνέχεια του geoparsing, την γεωκωδικοποίηση: Αναφέρουμε την προηγούμενη δουλειά στον τομέα και την δική μας προσέγγιση. Αφού ευρεθούν συντεταγμένες για μια ιστοσελίδα, έχει νόημα η ολοκλήρωση και συνάθροιση των αποτελεσμάτων, ώστε να δοθεί κάποια γενικότερη πληροφορία για την σελίδα. Επομένως, πραγματευόμαστε το ζήτημα της ολοκλήρωσης και παρουσίασης των δεδομένων.

Τέλος, η διαδικασία της γεωκωδικοποίησης απαιτεί γεωγραφικά δεδομένα υψηλής ποιότητας, και ταχείς αλγόριθμους ταιριάγματος. Κάνουμε, λοιπόν, μια νύξη πάνω στο θέμα του ταιριάγματος (matching) που θα μας απασχολήσει σε επόμενο κεφάλαιο. Επίσης, αναφερόμαστε στα "γεωδεδομένα", από την απόκτησή τους μέχρι τον καθαρισμό τους. Με την ευκαιρία αυτή, μιλάμε για τον καθαρισμό δεδομένων, τόσο γενικότερα, όσο και στην εφαρμογή του στην περίπτωσή μας. Το κεφάλαιο κλείνει με την περιγραφή της βάσης γεωδεδομένων που χτίστηκε για τις ανάγκες της εργασίας.

Όπως έχει αναφερθεί στο προηγούμενο κεφάλαιο, για την γεωκωδικοποίηση απαιτούνται ποιοτικοί αλγόριθμοι ταιριάγματος. Το **4**ο κεφάλαιο, λοιπόν, παρουσιάζει συγκριτικά διάφορους αλγόριθμους ταιριάγματος που έχουν προταθεί, μερικές δικές μας προτάσεις και παραλλαγές επί αυτών. Μιλάμε επίσης για τις προκλήσεις που δημιουργεί η Ελληνική γλώσσα σε μια τέτοια

εργασία, όπως το ακανθώδες ζήτημα των Λατινοελληνικών ή Φραγκολεβαντίνικων (greeklish)[3], και τις μεθοδολογίες που έχουν προταθεί για την αντιμετώπιση αυτού.

Στο **5**ο κεφάλαιο βλέπουμε στοιχεία της υλοποίησης του συστήματος, δηλαδή το πώς μεταφράζονται όσα έχουμε πει σε κλάσεις και μεθόδους.

Έπειτα, στο **6**ο κεφάλαιο, αναλύουμε τον έλεγχο του δημιουργηθέντος συστήματος, και βλέπουμε κάποια αποτελέσματα εκτέλεσής του.

Το **7**ο κεφάλαιο αποτελεί τον επίλογο της εργασίας, συνθέτοντας τα συμπεράσματα αυτής, και δίνοντας μερικές κατευθύνσεις για μελλοντικές επεκτάσεις.

Κλείνοντας, στο **8**ο κεφάλαιο παρουσιάζεται ενδεικτική βιβλιογραφία για όλα όσα εξετάζονται στον τόμο αυτό.

Καλή ανάγνωση!

[3]Μάλλον ατυχής απόδοση του όρου greeklish από την δημοσιογραφική κοινότητα.

2 Περιγραφή

Στο κεφάλαιο αυτό θα κάνουμε μια ανασκόπηση ορισμένων εφαρμογών του χώρου, θα ορίσουμε τυπικά τον στόχο της εργασίας μας, θα παρουσιάσουμε την αρχιτεκτονική του συστήματός μας, και θα δούμε μερικές λεπτομέρειες σχετικές με την υλοποίησή του.

2.1 Προηγούμενη εργασία

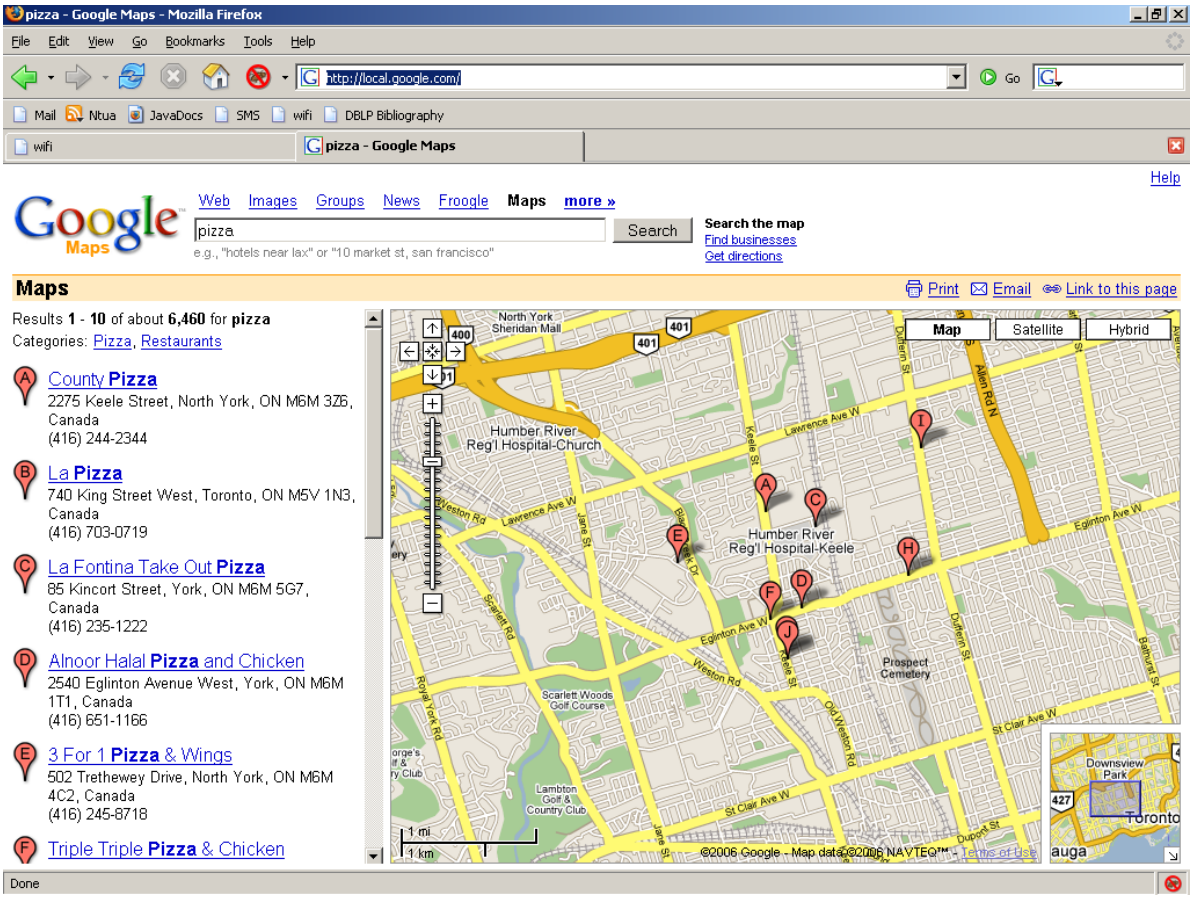
Προκειμένου να οριοθετήσουμε τον χώρο που πραγματεύεται η εργασία αυτή, χρήσιμη θα ήταν μια σύντομη ανασκόπηση ορισμένων αποτελεσμάτων που παρουσιάζουν ενδιαφέρον. Θα δούμε, συνοπτικά, κάποιες ολοκληρωμένες προτάσεις, ερευνητικές και εμπορικές. Αργότερα, σε επόμενα κεφάλαια, θα εξετάσουμε επιμέρους ιδέες που έχουν προταθεί.

Οποιαδήποτε επισκόπηση εφαρμογών "εξαγωγής πληροφορίας" θα είναι ελλιπής χωρίς την αναφορά του κολοσσού που ακούει στο όνομα Google™[4]. Τον τελευταίο χρόνο, το Google έχτισε, και έκτοτε αναβαθμίζει διαρκώς, την υπηρεσία Google Local[5] (περιλαμβάνει και το Google Maps και Google Earth). Με την υπηρεσία αυτή ο χρήστης μπορεί, μεταξύ άλλων, να αναζητήσει με βάση γεωγραφικά κριτήρια σε συνδυασμό με λέξεις-κλειδιά ("βρες πιτσαρίες σε αυτήν την περιοχή του χάρτη"). Η δυνατότητα αυτή, βέβαια, λειτουργεί ικανοποιητικά, προς το παρόν, μόνο για τις

[4]Στο [GH05] αναφέρεται, σχεδόν ανεκδοτολογικά, ότι "Μια από τις καλύτερες ερωτήσεις που μπορείτε να ρωτήσετε (τον εαυτό σας), όσο σκέφτεστε τη διαδικασία της ανάλυσης, είναι "Τι θα έκανε το Google;". Αν και οι αλγόριθμοι του Google είναι κλειστοί και κρατιούνται αρκετά μυστικοί, τα αποτελέσματα των αναζητήσεων βοηθούν στην κατανόησή του."

[5]http://local.google.com , ίσχυε την 28/6/2006

Ηνωμένες Πολιτείες. Εξετάζοντας την λειτουργία της υπηρεσίας, παρατηρούμε ότι λειτουργεί ψάχνοντας στις σελίδες που, ούτως ή άλλως η Google έχει στα ευρετήριά της, ταχυδρομικούς κώδικες και τοπωνύμια, ενώ κάνει και χρήση κατηγοριοποίησεων των ιστοσελίδων από ανθρώπους[6].



Σχήμα 1: Ψάχνοντας για πίτσα στο Toronto με το Google Local

Αντίστοιχα, ίσως με λιγότερο εντυπωσιακό παρουσιαστικό, φαίνεται να λειτουργεί ο χρυσός οδηγός του Yahoo™, αν και εδώ οι πληροφορίες προέρχονται μονάχα από την ανθρώπινη κατηγοριοποίηση ιστοσελίδων.

Ένα άλλο εμπορικό προϊόν που εντοπίζει έναν εξυπηρετητή ιστού στο χάρτη (με βάση τη διεύθυνση IP του, την καταχώρησή του στην υπηρεσία WhoIs, κ.α.), είναι το NeoTrace[7].

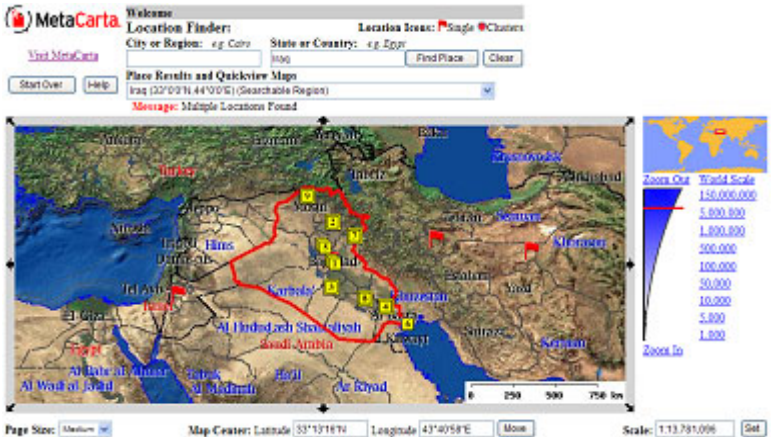
Περισσότερο ενδιαφέρον, για εμάς, παρουσιάζουν οι, εμπορικές, υπηρεσίες της MetaCarta[8]. Σε αυτές βρίσκουμε μια ενοποιημένη αντιμετώπιση geoparsing/γεωκωδικοποίησης, που χρησιμοποιεί

[6]Δηλαδή, κάποιος υπάλληλος της εταιρίας πρέπει να εξετάσει τη σελίδα, και να την κατατάξει στην κατάλληλη κατηγορία.

[7]http://www.networkingfiles.com/PingFinger/Neotraceexpress.htm ,ίσχυε την 28/6/2006

[8]http://www.metacarta.com ,ίσχυε την 28/6/2006

στοιχεία επεξεργασίας φυσικής γλώσσας, στατιστικής, και γεω-δεδομένα υψηλής λεπτομέρειας[9]. Τα αποτελέσματα μπορούν να χρησιμοποιηθούν για γεωγραφικές αναζητήσεις στο διαδίκτυο, σε εφαρμογές GIS, για γεωγραφική ομαδοποίηση/συσταδοποίηση εγγράφων κ.ο.κ.



Σχήμα 2: Γεωγραφική αναζήτηση με την υπηρεσία της Metacarta

Τέλος, μια πολύ ενδιαφέρουσα και πρόσφατη υπηρεσία ανοιχτού κώδικα, είναι ο μετατροπέας RSS σε GeoRSS [Exp06]. Με χρήση επεξεργασίας φυσικής γλώσσας (για την εύρεση της πληροφορίας) και τεχνικών μηχανικής μάθησης (για να βελτιώνεται διαρκώς, αλληλεπιδρώντας με τους χρήστες), και βρίσκει γεωγραφική πληροφορία μέσα σε RSS feeds. Με εφαρμογή κυρίως σε ειδησεογραφικές ιστοσελίδες, παρουσιάζει τελικά έναν παγκόσμιο χάρτη, που δείχνει τα κατά τόπους γεγονότα. Έτσι η αναζήτηση και ανάγνωση των ειδήσεων μπορεί να γίνει με βάση γεωγραφικά κριτήρια.

Έχοντας κάνει μια μικρή επισκόπηση του εμπορικού τοπίου, ας δούμε και μερικές πρόσφατες ερευνητικές προτάσεις.

Στο [GHL03] γίνεται εξαγωγή γεωγραφικής πληροφορίας από ιστοσελίδες, προκειμένου να ταξινομηθούν ως προς την γεωγραφική θέση τα αποτελέσματα μιας αναζήτησης. Δίνεται βαρύτητα και ανάλογα με την τοπικότητα μιας σελίδας (π.χ. μια σελίδα που αναφέρεται στα Θεωδωριανά Άρτης[10] αναμένεται να ενδιαφέρει λιγότερο κόσμο από μια που αναφέρεται στα Βαλκάνια γενικά).

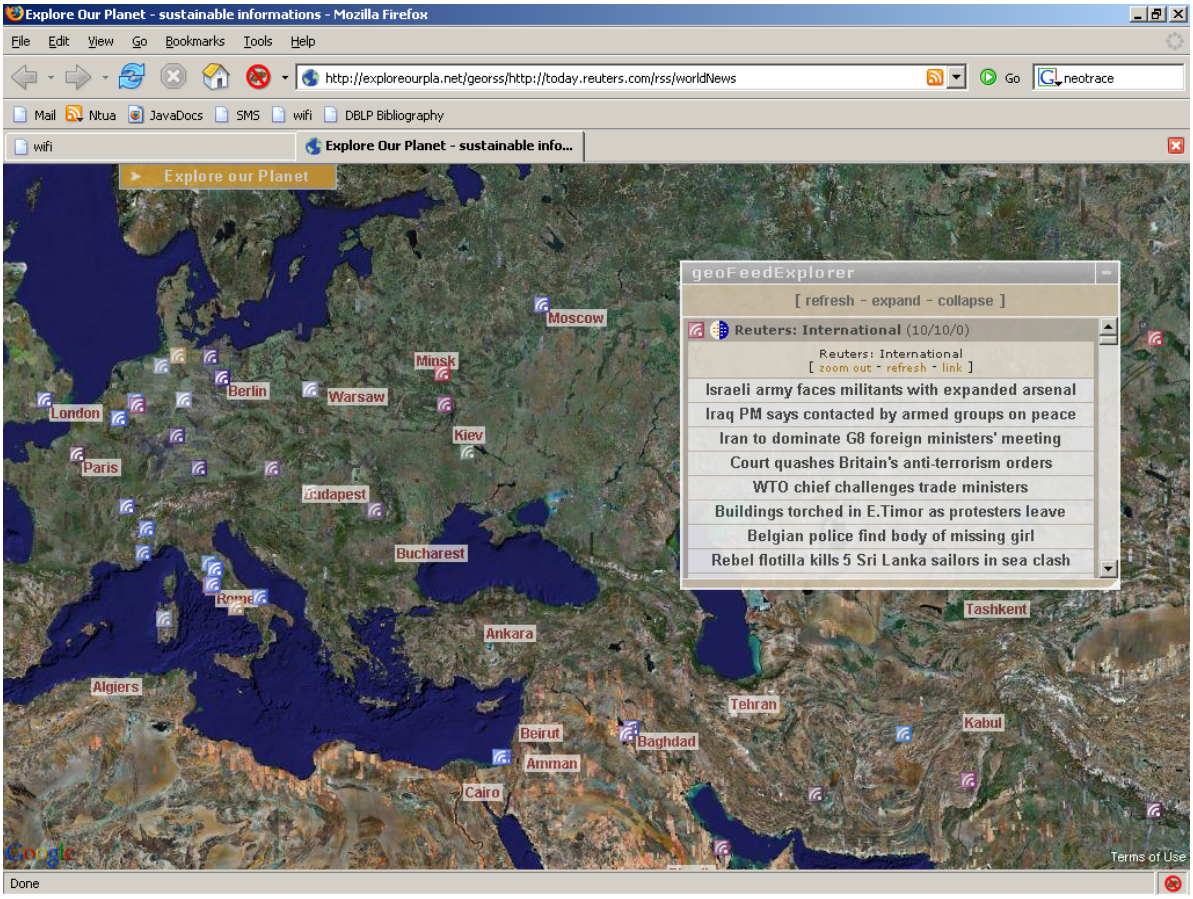
Στο κλασσικό [Mcc01] περιγράφεται μια μηχανή γεωγραφικής πλοήγησης στο διαδίκτυο, για την οποία έχει προηγουμένως εξαχθεί γεωγραφική πληροφορία από αυτό. Προκύπτει έτσι ένας εναλλακτικός τρόπος πλοήγησης στο διαδίκτυο (βλ. σχ.4).

[9]Για κάθε τοπωνύμιο περιέχονται, πέραν των συντεταγμένων του, διάφορα δεδομένα που βοηθούν στην ταυτοποίησή του. Για παράδειγμα, χρησιμοποιείται ο πληθυσμός ενός τόπου ως (ένας εκ των) συντελεστών βεβαιότητας για τον εντοπισμό του - επομένως ελλείψει άλλων στοιχείων το "Ρίο" θα μας οδηγήσει στο Rio de Janeiro και όχι στην συμπαθή πόλη της Πελοποννήσου.

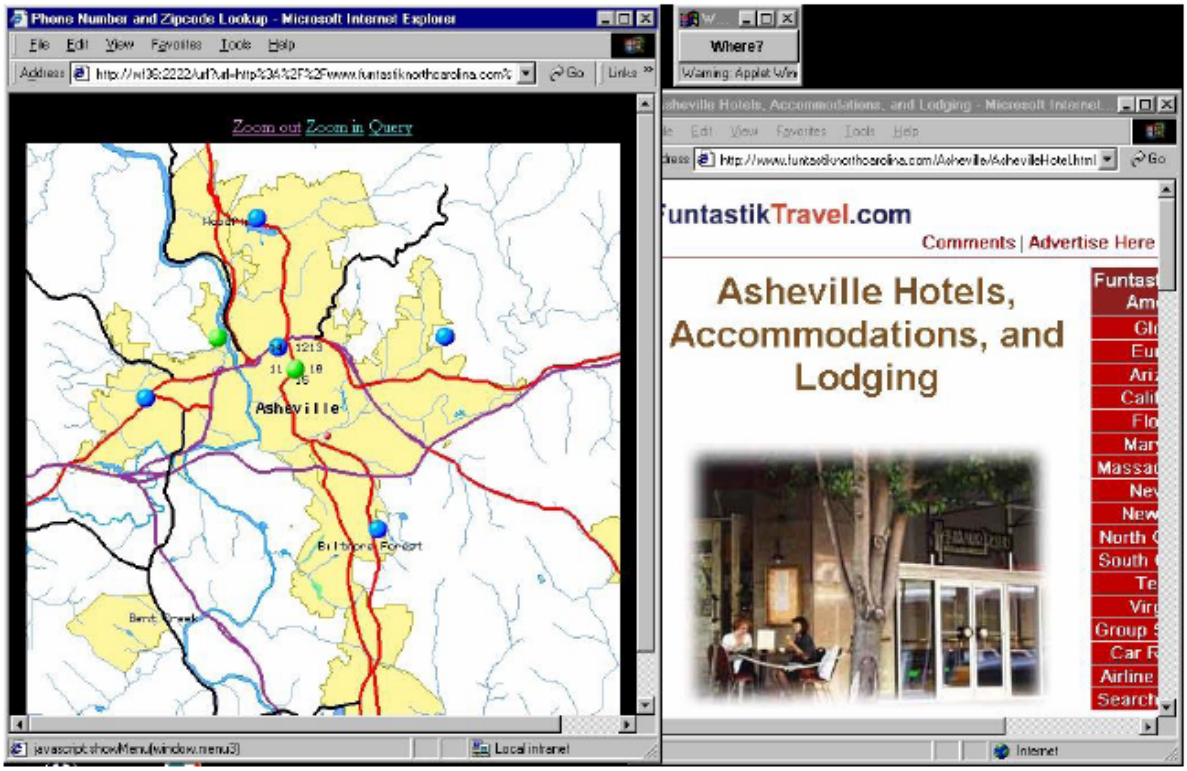
[10]http://www.epirus.com/theodoriana , ίσχυε την 28/6/2006

Στο [BLB+03] προτείνεται η χρήση της γεωγραφικής πληροφορίας που διατίθεται στον παγκόσμιο ιστό για την πλήρωση με δεδομένα μιας χωρικής Β.Δ. (για παράδειγμα, στο περιβάλλον μιας υπηρεσίας βασισμένης στη θέση - Location Based Service). Εναλλακτικά, η πληροφορία μπορεί να χρησιμοποιηθεί για την αρχικοποίηση (bootstrapping) της βάσης.

Τέλος, στο [DGS00] περιγράφεται μέθοδος υπολογισμού του γεωγραφικού εύρους μιας ιστοσελίδας, ενώ υλοποιείται και μια μηχανή γεωγραφικής αναζήτησης. Η μηχανή αυτή αξιοποιείται και στον τομέα της εξατομικευμένης αναζήτησης (personalised search), όπου η γεωγραφική θέση του χρήστη θεωρείται ως στοιχείο του προφίλ του, και άρα η εγγύτητα της γεωγραφικής θέσης μιας σελίδας σε αυτήν του χρήστη αποτελεί κριτήριο για την ιεράρχηση της σελίδας στην λίστα αποτελεσμάτων.



Σχήμα 3: Διαβάζοντας τα νέα της μέρας με το RSS to GeoRSS converter



Σχήμα 4: Γεωγραφική πλοήγηση στο διαδίκτυο. Από το [Mcc01]

2.2 Στόχος της διπλωματικής

Έχοντας δει ορισμένες προτάσεις που εκμεταλλεύονται την γεωγραφική πληροφορία ιστοσελίδων, είναι πλέον καιρός να ορίσουμε τον στόχο της διπλωματικής αυτής. Δηλαδή, να εξετάσουμε σε τι αποσκοπεί η μελέτη των διαφόρων αποτελεσμάτων και η πρόταση νέων λύσεων.

Θέλουμε, λοιπόν, να δημιουργήσουμε ένα γεωγραφικό ευρετήριο του ελληνικού ιστοχώρου. Δηλαδή ξεκινώντας από μια ιστοσελίδα του, να της αποδώσουμε γεωγραφικές συντεταγμένες με βάση το τι περιγράφει (το περιεχόμενό της). Ο μετασχηματισμός αυτός πρέπει να γίνεται με μεγάλη ταχύτητα, εάν λάβουμε υπ'όψη το πλήθος των ιστοσελίδων που θα πρέπει να γεωκωδικοποιηθούν για να δημιουργηθεί το ευρετήριο.

2.2.1 Έμφαση της εργασίας

Έχουμε ήδη περιγράψει κάποια κίνητρα που θα μπορούσαν να μας ωθήσουν σε αυτήν την προσπάθεια. Έχοντας αυτά υπ'όψη, και εξετάζοντας τον ελληνικό ιστοχώρο, παρατηρήσαμε ότι δύο κατηγορίες ιστοσελίδων παρουσιάζουν ιδιαίτερο ενδιαφέρον: Πρώτον, οι εμπορικές ιστοσελίδες, που περιέχουν σχεδόν πάντα γεωγραφική πληροφορία εν είδη τηλεφώνων ή διευθύνσεων (και γενικά, οι ιστοσελίδες που περιέχουν λίγες, αλλά βέβαιες γεωγραφικές πληροφορίες τέτοιου τύπου). Δεύτερον, μας ενδιαφέρουν οι ειδησεογραφικές και τουριστικές ιστοσελίδες. Αυτές περιέχουν πολλαπλάσιες γεωγραφικές πληροφορίες, υπό την μορφή τοπωνυμίων. Δυστυχώς, η πληροφορία

που παρέχει ένα τοπωνύμιο που βρίσκουμε σε μια σελίδα έχει μικρή βεβαιότητα, αφού μπορεί η λέξη να μην χρησιμοποιείται ως τοπωνύμιο, ή ένα τοπωνύμιο να αναφέρεται σε πάνω από μια τοποθεσίες. Στην περίπτωση αυτή πρέπει με κατάλληλο τρόπο να σταθμίσουμε την πληροφορία που βρίσκουμε, ώστε να βεβαιωθούμε για το αληθές της.

Ως απόρροια των παρατηρήσεων αυτών, στην εργασία δίνουμε έμφαση στην εξαγωγή πληροφορίας από τέτοιου είδους ιστοθέσεις. Οι πληροφορίες που αναζητούμε είναι τηλέφωνα, ταχυδρομικοί κώδικες, διευθύνσεις και τοπωνύμια. Αξιοποιούνται επίσης άλλες μορφές πληροφορίας, όπως η διεύθυνση IP του εξυπηρετητή.

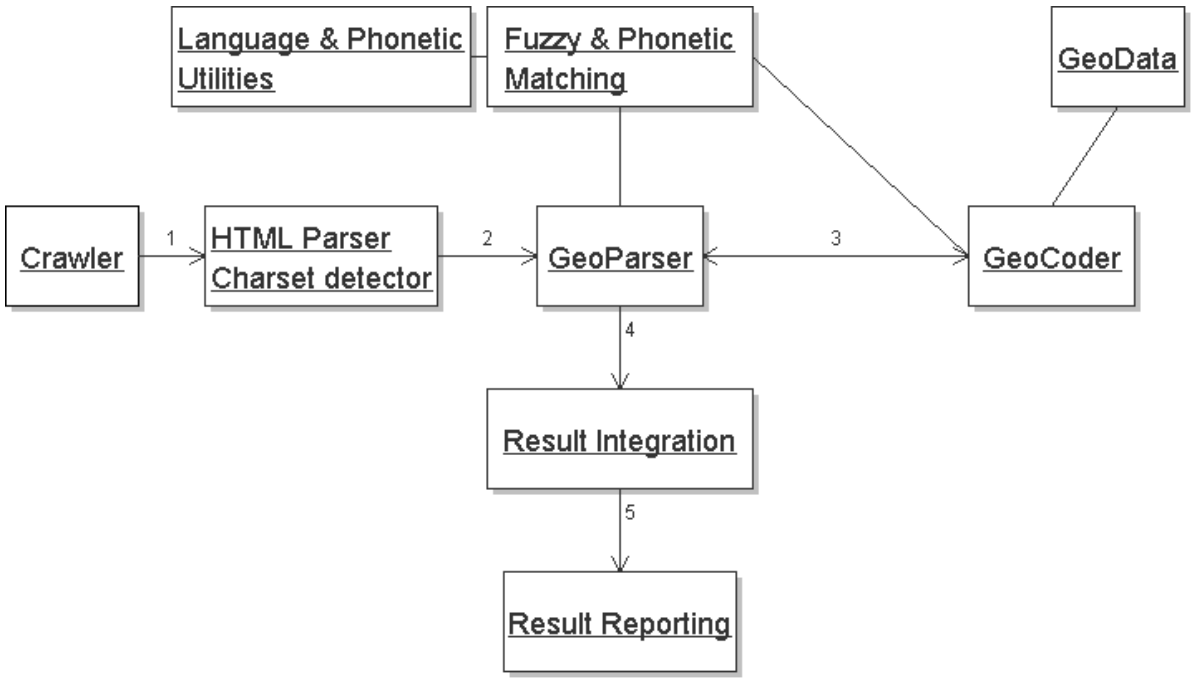
2.3 Αρχιτεκτονική συστήματος

Κατά τον σχεδιασμό της αρχιτεκτονικής του συστήματος, ακολουθήσαμε την κοινώς αποδεκτή μέθοδο της εξαγωγής γεωγραφικής πληροφορίας οδηγούμενης από το geoparsing. Δηλαδή, οι ιστοσελίδες προς επεξεργασία συλλέγονται ξεχωριστά, και δίνονται στον geoparser για επεξεργασία. Αυτός φροντίζει για την εξαγωγή γεωγραφικής πληροφορίας από το κείμενο, την οποία και παραδίδει στον γεωκωδικοποιητή για γεωκωδικοποίηση. Τέλος, μια μονάδα του φροντίζει για την ολοκλήρωση της πληροφορίας, και της αναφοράς/παρουσίασης των αποτελεσμάτων. Σε ξεχωριστές μονάδες έχουν τοποθετηθεί, επίσης, οι μέθοδοι που αφορούν την προσεγγιστική αναζήτηση και το προσεγγιστικό ταίριαγμα (βλ. κεφ. 5). Το σύστημα έχει δομηθεί, ουσιαστικά, πάνω στο αρχιτεκτονικό μόρφημα της αλυσίδας επεξεργασίας. Στο σχ. 5 εικονίζεται ένα σκαρίφημα της αρχιτεκτονικής του[11], καθώς και η τυπική σειρά ακολουθιών για την επεξεργασία μιας ιστοσελίδας.

Στο σημείο αυτό, αξίζει να αναφέρουμε και μια άλλη προσέγγιση, την εξαγωγή γεωγραφικής πληροφορίας οδηγούμενης από την γεωκωδικοποίηση. Αυτή ακολουθείται από υπηρεσία γεωκωδικοποίησης της Metacarta (βλ. κεφ 2., υποσημειώσεις 8,9).

Στην υπόλοιπη εργασία, πραγματευόμαστε τα επιμέρους τμήματα του συστήματος, τόσο από θεωρητική/ερευνητική σκοπιά, όσο και από την πλευρά της υλοποίησης.

[11]Η βαθμίδα crawler υποννοεί οποιονδήποτε τρόπο απόκτησης της ιστοσελίδας, είτε έναν κανονικό crawler, είτε μια προσομοίωσή του, πάνω σε σελίδες που έχουν ήδη ανακληθεί τοπικά.



Σχήμα 5: Αρχιτεκτονική του συστήματος, και τυπική ροή επεξεργασίας

2.3.1 Περί ιστοσελίδων

Προτού επεκταθούμε στην περιγραφή του συστήματος, ας μιλήσουμε λίγο για την πρώτη ύλη που αυτό καλείται να διαχειριστεί, τις ιστοσελίδες.

2.3.1.1 Web crawlers

Για την απόκτηση ικανού πλήθους ιστοσελίδων για τέτοιου τύπου εφαρμογές, συνήθως χρησιμοποιούνται εφαρμογές που καλούνται web crawlers ή spiders. Ένας web crawler, με αφετηρία κάποιες αρχικές σελίδες, και ακολουθώντας τους υπερσυνδέσμους κάθε σελίδας, θα ανασύρει, σταδιακά, όλες τις ιστοσελίδες του ελληνικού διαδικτύου. Μετά την μετατροπή τους σε ενδιάμεση μορφή, κατάλληλη για αρχειοθέτηση και επεξεργασία, θα τις αποθηκεύσει τοπικά. Θα φροντίσει, επίσης, για την τακτική ανανέωσή τους.

Ορισμένα προβλήματα που καλείται να αντιμετωπίσει είναι οι δομές δεδομένων που θα χρησιμοποιηθούν για να επιτευχθεί ικανοποιητική παραλληλία και απόδοση, όπως και κάποια ζητήματα ηθικής (π.χ. πώς να επιτελέσει το έργο του αποδοτικά, χωρίς να δημιουργεί πρόβλημα φόρτου στον εξυπηρετητή ιστού). Τέτοια προβλήματα έχουν λυθεί σε ερευνητικό επίπεδο (π.χ. [KSS+03]), όπως και στους διαθέσιμους web crawlers ανοιχτού κώδικα[12] (Για εκτενή παρουσίαση τέτοιων ζητημάτων, βλ. [CH03],[Hea02])

[12]http://java-source.net/open-source/crawlers , ίσχυε την 5/7/2006

Μερικά άλλα προβλήματα, χρειάζονται εξειδικευμένες λύσεις. Ένα βασικό, είναι ο ορισμός του "Ελληνικού διαδικτύου". Υπάρχει αρκετή ερευνητική εργασία πάνω σε αυτό το ζήτημα, με απαντήσεις που κυμαίνονται από την σχετικά απλή "Ελληνική ιστοσελίδα είναι όποια προέρχεται από το .gr domain, ή είναι γραμμένη στα Ελληνικά" (λύση την οποία χρησιμοποιήσαμε κι εμείς), εως την πιο πολύπλοκη αντιμετώπιση του [LEJ+04], που κάνει χρήση τεχνικών Ανάκτησης Πληροφορίας και Εξόρυξης Δεδομένων για την ταυτοποίηση της "ελληνικότητας" μιας ιστοσελίδας.

2.3.1.2 HTML Parsing

Ένα ακόμη πρόβλημα είναι πως πολλές σελίδες είναι γραμμένες σε συντακτικά λάθος HTML, με αποτέλεσμα να χρειάζονται ιδιαίτερα ευφυείς parsers για να την αναγνωρίσουν και να την απεικονίσουν. (Αυτός που αναπτύξαμε εμείς, βασίζεται στον HTML parser ανοιχτού κώδικα Tagsoup[13]).

2.3.1.3 Ανίχνευση κωδικοσελίδας

Επίσης, πρόβλημα του ελληνικού ιστού, είναι και η επιλογή της σωστής κωδικοσελίδας[14] για την χρήση μιας ιστοσελίδας. Για λόγους ιστορικούς, ο Ελληνικός ιστός είναι γραμμένος σε τρεις, κυρίως, κωδικοσελίδες: την ISO-8859-7, την Cp-1253[15], καθώς και την νεότερη (unicode) UTF-8. Δυστυχώς, η μέση ιστοσελίδα δεν προσδιορίζει σε ποιά κωδικοποίηση έχει γραφεί. Διάφοροι τρόποι έχουν προταθεί για την ανίχνευση της κωδικοσελίδας που χρησιμοποιεί μια ιστοσελίδα, όπως ανίχνευση συγκεκριμένων χαρακτήρων, χρήση συχνοτήτων εμφάνισης ν-γραμμάτων σε κάθε γλώσσα[16], ή και χρήση των πληροφοριών που μπορεί η ίδια η σελίδα να δηλώνει, μέσω των ετικετών μεταχαρακτηριστικών κεφαλίδας (header meta tags). Ένα συνδυασμό των ανωτέρω, προσαρμοσμένο στην Ελληνική πραγματικότητα, ακολουθούμε κι εμείς.

2.3.1.4 Διαχείριση ιστοσελίδων στο σύστημά μας

Τέλος, για τις ανάγκες της εργασίας, θεωρήσαμε υπερβολική την χρήση ενός crawler. Αντ'αυτού, προτιμήσαμε να ανασύρουμε από τον ελληνικό ιστό περίπου 5 GB σελίδων γενικού περιεχομένου, και περίπου 100 σελίδων στοχευμένου περιεχομένου, από ιστοτόπους ειδησεογραφικούς, τουριστικούς και εμπορικούς. Πάνω σε αυτές τις ιστοσελίδες εκτελέσαμε τα πειράματά μας, και από αυτές βγάλαμε τα συμπεράσματά μας.

---

[13]http://mercury.ccil.org/~cowan/XML/tagsoup/ , ίσχυε την 5/7/2006

[14]Κωδικοποίηση χαρακτήρων: αντιστοίχιση ακολουθίας bit (συνηθέστερα 8 ή 16 bits) σε χαρακτήρες

[15]Πρόκειται για τροποποίηση της πρότυπης ISO-8859-7 από την Microsoft...

[16]Οι μέθοδοι αυτές χρησιμοποιούνται από τον φυλλομετρητή ανοιχτού κώδικα Mozilla, και είναι διαθέσιμες στη γλώσσα Java στο: http://jchardet.sourceforge.net/ , ίσχυε την 5/7/2006

2.4 Σύνοψη

Στο κεφάλαιο αυτό, προδιαγράψαμε την έμφαση και τις λειτουργίες του συστήματος. Περιγράψαμε την αλυσωτή αρχιτεκτονική του, και τους λόγους που οδήγησαν σε αυτήν. Το σύστημα λειτουργεί με πρώτη ύλη τις ιστοσελίδες - γι'αυτό αναλύσαμε την διαχείριση ιστοσελίδων, τόσο σε επίπεδο προηγούμενης δουλειάς, όσο και ως προς τη δική μας προσέγγιση.

3 *Geoparsing & Geocoding*

Στο κεφάλαιο αυτό πραγματευόμαστε τις βασικές έννοιες της εργασίας: το geoparsing και την γεωκωδικοποίηση. Ξεκινάμε με την ανάλυση του geoparsing, παρουσιάζοντας την προηγούμενη εργασία, και την δική μας προσέγγιση. Μετά την παρουσίαση ορισμένων στοιχείων επίδοσης, στρεφόμαστε στην γεωκωδικοποίηση.

3.1 Geoparsing - Προσεγγίσεις

Πάνω στο θέμα της εξαγωγής γεωγραφικής πληροφορίας από ιστοσελίδες, έχει γίνει τα τελευταία χρόνια αρκετή εργασία, τόσο σε ερευνητικό όσο και σε εμπορικό επίπεδο. Πολλοί ερευνητές στο παρελθόν έχουν επισημάνει τον όγκο της γεωγραφικής πληροφορίας που μπορεί να φέρει μια σελίδα, όπως και την χρησιμότητα αυτής (ενδεικτικά [Mcc01], [WA03], [DGS00], [MAH+03]). Σε γενικές γραμμές προτείνονται τέσσερα επίπεδα από τα οποία μπορεί να εξαχθεί η πληροφορία αυτή: Το επίπεδο δικτύου, το συντακτικό επίπεδο, το σημασιολογικό επίπεδο και το επίπεδο της τοπολογίας.

Επίπεδο	*Ενδεικτικές πηγές πληροφορίας*	*Παράδειγμα πληροφορίας*
Δικτύου	Διεύθυνση IP, Υπηρεσία whois	Διεύθυνση εξυπηρετητή 147.102.1.1
Σύνταξης	Τηλέφωνα, Τ.Κ., Διευθύνσεις	Αναφορά: "Ηρώων Πολυτεχνείου 12, Ζωγράφου"
Σημασιολογίας	Προτάσεις σε φυσική γλώσσα	Αναφορά: "... κοντά στη βιβλιοθήκη, με θέα τον Υμηττό,..."
Τοπολογίας	Υπερσύνδεσμοι	Η κεντρική σελίδα του Πολυτεχνείου δείχνει στη σελίδα

Πίνακας 1: Επίπεδα εξαγωγής γεωγραφικής πληροφορίας

3.1.1 Επίπεδο δικτύου

Οι πρώτες προσπάθειες στο χώρο επιχείρησαν την εξαγωγή της από τη διεύθυνση IP του εξυπηρετητή (π.χ. [Wis01], MaxMind GeoIP[17]). Τέτοιες προσπάθειες συνεχίστηκαν με την εκμετάλλευση της υπηρεσίας whois ([Mcc01], Neotrace[18]), που δίνει ποικίλες πληροφορίες (μεταξύ άλλων, ενίοτε και μια διεύθυνση ή τηλέφωνο του υπεύθυνου για τον εξυπηρετητή). Δυστυχώς, αυτές οι μέθοδοι έχουν μειωμένη χρησιμότητα, ή/και χαμηλή βεβαιότητα, επειδή συνήθως η γεωγραφική θέση του εξυπηρετητή δεν συμπίπτει με την θέση στην οποία αναφέρεται μια ιστοσελίδα.

[17]http://www.maxmind.com/ , ίσχυε την 3/7/2006

[18]http://www.networkingfiles.com/PingFinger/Neotraceexpress.htm ,ίσχυε την 28/6/2006

3.1.2 Επίπεδο κειμένου-σύνταξης

3.1.2.1 Αναζητούμενες πληροφορίες

Μια διαφορετική οπτική γωνία είναι η εκμαίευση πληροφορίας από το ίδιο το κείμενο. Με την χρήση απλών γραμματικών μπορούν να εντοπιστούν σε αυτό τηλέφωνα, ταχυδρομικοί κώδικες, ή και τοπωνύμια. Προβλήματα που έχουν επισημανθεί σε αυτήν την φάση αφορούν στην επαλήθευση των αποτελεσμάτων (πρόκειται πράγματι για ταχυδρομικό κώδικα, ή είναι ένας άσχετος αριθμός;), στην δυσκολία εύρεσης μιας ενιαίας γραμματικής που να καλύπτει όλες τις μορφές αναπαράστασης τηλεφώνων, ταχυδρομικών κωδίκων κ.λ.π. ανά τον κόσμο (στην πράξη συνήθως γράφεται μια γραμματική ανά χώρα που ενδιαφέρει), όπως και την συνάθροιση των αποτελεσμάτων που βρίσκονται σε μια σελίδα. Με χρήση πιο πολύπλοκων γραμματικών, γεωγραφική πληροφορία μπορεί να αναζητηθεί και υπό την μορφή διευθύνσεων. Οι γραμματικές αυτές εξαρτώνται από τον τρόπο με τον οποίο γράφονται τυπικά οι διευθύνσεις στη χώρα στην οποία θα χρησιμοποιηθούν. Επίσης, συνεπικουρούνται από επιπλέον εργαλεία και πληροφορίες, όπως αναζήτηση στο κείμενο λέξεων-κλειδιών, ή τοπωνυμίων (βλ. κεφ.5).

3.1.2.2 Προβλήματα - Προσεγγίσεις

Ενδεικτικά και μόνο, αναφέρουμε: Το [BCG+99], όπου προτείνεται η απλή αναζήτηση τοπωνυμίων και ταχυδρομικών κωδίκων, ιδέα που επεκτείνεται στο [Mcc01] με την χρήση τηλεφώνων. Στα [BLB+03],[MAH+03] εντοπίζονται διευθύνσεις σε ιστοσελίδες, ενώ τα τελικά αποτελέσματα αξιοποιούνται με ποικίλους τρόπους. Στο [WA03] προτείνεται η εξέταση των λέξεων που απαρτίζουν ένα URL για γεωγραφική πληροφορία (π.χ. www.somesite.**gr**/text/**macedonia**.html). Στο SPIRIT (βλ. [Clo05]) παρουσιάζονται τρόποι για την άρση της αμφισημίας ενός τοπωνυμίου (π.χ. δύο πόλεις που έχουν το ίδιο όνομα).

Τέλος, αξίζει να αφιερώσουμε μια παράγραφο στην δουλειά της S.Sarawagi στον τομέα αυτό ([BDS01],[Sar02]) η οποία διαφέρει σημαντικά από τις υπόλοιπες προτάσεις. Αν και εκμεταλλεύεται το κείμενο σε επίπεδο σύνταξης, εντούτοις προτείνει τη χρήση κρυμμένων μοντέλων Markov και τεχνικών μηχανικής μάθησης για την εξαγωγή δομής και πληροφορίας από αδόμητο κείμενο. Οι τεχνικές που προτείνονται έχουν βρει εφαρμογή στο γενικής χρήσης εργαλείο καθαρισμού DATAMOLD, όπως και σε συγκεκριμένες εφαρμογές σε τομείς geoparsing.

3.1.3 Επίπεδο κειμένου-σημασιολογίας

Προχωρώντας ένα επίπεδο παραπέρα, έχουν αναφερθεί μέθοδοι εξαγωγής γεωγραφικής πληροφορίας από τη σημασιολογία ενός κειμένου. Ίσως η πρώτη σχετική δουλειά ειναι στο [WP94], όπου κάθε έγγραφο εξετάζεται αφ'ενός για λεξιλόγιο με γεωγραφική χροιά (π.χ. ένα κείμενο μιλά

για λιμνοθάλασσες και καταρράκτες), αφ'ετέρου για τοπωνύμια που αναζητούνται σε κατάλογο τοπωνυμίων (π.χ. αναφέρεται η "Μούτελη"). Λαμβάνοντας αυτά υπ'όψη, μπορεί να υπολογιστεί κατά προσέγγιση η θέση της ιστοσελίδας (εδώ: αναφέρεται στη Λευκάδα).

Στο ίδιο επίπεδο, στο [Mcc01] αναφέρεται η χρήση γλωσσικής ανάλυσης για τον εντοπισμό ονομάτων φυσικών προσώπων, εταιριών κ.ο.κ., τα οποία συνοδεύονται από γεωγραφική πληροφορία. Αυτή η ιδέα συναντάται και στην κοινότητα του GATE ([CMB+02], [CMB+06]).

Το μειονέκτημα αυτών των μεθόδων δεν είναι άλλο από την αυξημένη πολυπλοκότητα, και εξάρτηση από την γλώσσα, που παρουσιάζουν, γι'αυτό και οι περισσότερες μέθοδοι στράφηκαν σε απλούστερες και ταχύτερες ιδέες.

Μια πολύ πρόσφατη εφαρμογή, που δείχνει να ξεπερνά αυτό το μειονέκτημα μειώνοντας τον όγκο των δεδομένων που επεξεργάζεται, είναι η υπηρεσία RSS to GeoRSS του Exploreourpla.net [Exp06]. Σε αυτήν εντοπίζονται τοπωνύμια μέσα στους τίτλους ενός RSS feed, τα οποία αναζητούνται σε μια βάση τοπωνυμίων. Για να μην εκλαμβάνεται η κάθε λέξη ως πιθανό τοπωνύμιο, όμως (κάτι που θα είχε άσχημα αποτελέσματα), χρησιμοποιούνται και γραμματικοί κανόνες, προσαρμοσμένοι στην γλώσσα του κειμένου, καθώς και τεχνικές μηχανικής μάθησης. Τα αποτελέσματα είναι αρκετά ελπιδοφόρα, όπως μπορεί να διαπιστώσει ο επισκέπτης στον δικτυακό τόπο της υπηρεσίας.

3.1.4 Επίπεδο τοπολογίας

Πληροφορία μπορούμε επίσης να αναζητήσουμε στην τοπολογία ενός συνόλου ιστοσελίδων, όπως αυτή ορίζεται από τους υπερσυνδέσμους που τις συνδέουν. Κατά τον Μ. Βαζιργιάννη ([Vaz05]), οι υπερσύνδεσμοι από μια σελίδα Πηγή σε μια σελίδα Στόχο μπορεί να δείχνουν ότι η Πηγή σχετίζεται με το Στόχο, ή και ότι η Πηγή συνιστά, αναφέρει ή επικυρώνει τον Στόχο.

Μεταφέροντας την επιτυχημένη ιδέα της αξιοποίησης των υπερσυνδέσμων από άλλους τομείς, όπως η εκτίμηση της σημαντικότητας μιας σελίδας ([BP98], [Vaz05]), η θεματική κατηγοριοποίηση (π.χ. THESUS [HNV+03]) ή το στοχευμένο web crawling ([KSS+03]), στην αναζήτηση γεωγραφικής πληροφορίας, θα μπορούσαμε να βελτιώσουμε την τελευταία, τόσο ποσοτικά όσο και ποιοτικά. Πέρα από γενικές προτάσεις και μικρά proofs-of-concept, η ιδέα αυτή υλοποιείται στο [DGS00].

3.2 Geoparsing - Η δική μας προσέγγιση

Έχοντας κατά νου τις ανωτέρω προσεγγίσεις, καθώς και την έμφαση που αποφασίσαμε να δώσουμε σε εμπορικές, ειδησεογραφικές και τουριστικές ιστοσελίδες (βλ. και κεφ.2), ας μιλήσουμε για την προσέγγιση που προτείνουμε. Παρουσιάζουμε, αρχικά, μια πρώτη υλοποίηση που κάναμε στο περιβάλλον επεξεργασίας φυσικής γλώσσας GATE, και έπειτα την τρέχουσα υλοποίηση του συστήματος, με χρήση διαδοχικών κανονικών γραμματικών.

3.2.1 Προσέγγιση στο GATE

3.2.1.1 Το περιβάλλον GATE

Το περιβάλλον GATE (General Architecture for Text Engineering, [CMB+02]), είναι ένα πλαίσιο (framework) που μοντελοποιεί υποσημειώσεις (annotations) επί ενός κειμενικού εγγράφου ως ακμές με άκρα σημεία του κειμένου. Μια υποσημείωση έχει, εκτός των άκρων της, και άλλα χαρακτηριστικά, οριζόμενα από τον προγραμματιστή. (π.χ. μια υποσημείωση τύπου "Λέξη" ξεκινά από τον 242ο χαρακτήρα του κειμένου, τελειώνει στον 247ο, και έχει ιδιότητες "κείμενο"="υγεία", "ορθογραφία"="κεφαλαία"). Σημαντικό στοιχείο του μοντέλου του GATE είναι πως δυο υποσημειώσεις δεν είναι αμοιβαία αποκλειόμενες, αλλά μπορούν να υφίστανται επί κοινού κειμένου (π.χ. η ανωτέρω υποσημείωση μπορεί να συνυπάρχει με μια υποσημείωση τύπου "Φράση" από τον 240ο εως τον 251ο χαρακτήρα με "κείμενο"="η υγεία σου"). Υποσημειώσεις μπορούν να δημιουργηθούν με διάφορους τρόπους. Οι τρεις κυριότεροι είναι:

1. Κατά την δημιουργία του εγγράφου, για τις βασικές μονάδες του (π.χ. λέξεις, κενά, αριθμούς)
2. Μέσω αναζήτησης φράσεων σε ευρετήριο (lookup)
3. Μέσω κανονικών γραμματικών που ενεργούν επί των υποσημειώσεων. Οι γραμματικές αυτές πρέπει να είναι γραμμένες σε JAPE (λίγο στρυφνή γλώσσα, που μοιάζει με την CPSL).

3.2.1.2 Υλοποίηση σε GATE

Στο περιβάλλον αυτό υλοποιήθηκε ένας πλήρης geoparser, μέσω γραμματικών JAPE, καθώς και διαφόρων μονάδων (για υποστήριξη ελληνικών, προσεγγιστικής αναζήτησης, κ.ο.κ.) που γράφτηκαν εξ'αρχής. Η εφαρμογή δοκιμάστηκε επί αρκετών ιστοσελίδων, αλλά παρά τις προσπάθειες ρύθμισης και βελτίωσής της, τα αποτελέσματα παρέμεναν μη ικανοποιητικά.

3.2.1.3 Συμπεράσματα

Μετά τα ανωτέρω, φάνηκε αφ'ενός ότι το μοντέλο του GATE δεν είναι κατάλληλο για μια τέτοια εφαρμογή (κυρίως λόγω της αμφισημίας που εισάγει η δυνατότητα πολλαπλών υποσημειώσεων επί του ίδιου κειμένου), αφ'ετέρου ότι το ίδιο το GATE, παρά τα 10 χρόνια ύπαρξής του, δεν είναι ακόμη αρκετά ώριμο (από πλευράς ορθότητας και σταθερότητας), για να στηρίξει μια εφαρμογή ελαφρώς διαφορετική από αυτές για τις οποίες έχει σχεδιαστεί. Συνεπώς, εγκαταλείφθηκε ως πλατφόρμα υλοποίησης.

3.2.2 Προσέγγιση με χρήση διαδοχικών κανονικών γραμματικών

Χρησιμοποιώντας ιδέες που αποκομίσθηκαν από την χρήση του GATE, το σύστημα ξαναγράφτηκε εξ'αρχής. Στην νέα του μορφή κάνει χρήση διαδοχικών κανονικών γραμματικών (cascaded regular grammar transducers), συνεπικουρούμενων από προσεγγιστική αναζήτηση λέξεων σε ευρετήριο (approximate string lookup - βλ. κεφ.5). Κάθε επίπεδο κανονικής γραμματικής χρησιμοποιεί τα αποτελέσματα του προηγούμενου, και εντοπίζει πληροφορίες διαφορετικής υφής. Για παράδειγμα, το πρώτο επίπεδο παίρνει ως είσοδο χαρακτήρες, και βγάζει λέξεις, αριθμούς, κενά, πιθανά τοπωνύμια κ.λ.π., ενώ το τελευταίο επίπεδο, δεχόμενο ως είσοδο ταχυδρομικούς κώδικες, τοπωνύμια, τηλέφωνα κ.ο.κ. βγάζει ως αποτελέσματα ολοκληρωμένες διευθύνσεις. Η προσεγγιστική αναζήτηση που αναφέρθηκε, χρησιμοποιείται για τον εντοπισμό πιθανών τοπωνυμίων στο κείμενο.

Παράλληλα, η πληροφορία που ανιχνεύεται, κανονικοποιείται και τεχνολογείται (standardised & parsed) (π.χ. η "οδ. Αγ.Σώστη 3, Χαλάνδρι 231-35" γίνεται {τύπος: "οδός", οδός: "Αγίου Σώστη", αρ: 3, περιοχή:"Χαλάνδρι", ΤΚ:"23135"}). Η κανονικοποίηση και ο διαχωρισμός της πληροφορίας σε πεδία, είναι απαραίτητα για την μετέπειτα γεωκωδικοποίησή της.

3.2.2.1 Ταίριαγμα τύπου Brill για πιθανά τοπωνύμια

Τέλος, όσον αφορά τα πιθανά τοπωνύμια που ευρίσκονται, γίνεται χρήση μιας ιδέας του Brill ([Bri92], μια εφαρμογή στα Ελληνικά στο [PPK+99]) για να βεβαιωθούμε ότι, θεωρώντας μια αλληλουχία λέξεων ως πιθανό τοπωνύμιο, δεν αποκλείουμε κάποια άλλη αλληλουχία, υποσυμβολοσειρά της πρώτης, από το να επιλεγεί.

Για παράδειγμα, από την φράση "η ομάδα του Βόλου Μαγνησίας 'Ατρόμητος'", θέλουμε να αναζητηθούν ως ξεχωριστά πιθανά τοπωνύμια οι συμβολοσειρές: "Βόλος", "Μαγνησίας", και "Βόλος Μαγνησίας", έστω και χρειαστεί να αναζητήσουμε - δίχως επιτυχία - και τις "Ατρόμητος" και "Βόλος Μαγνησίας Ατρόμητος".

Σημειωτέον ότι η λειτουργία αυτή δεν επιβαρύνει αισθητά τον χρόνο επεξεργασίας μιας σελίδας, όπως ίσως θα περίμενε κανείς, επειδή η μέση σελίδα περιέχει συνήθως λίγα, και διασκορπισμένα, πιθανά τοπωνύμια.

3.2.3 Εκμετάλλευση υπερσυνδέσμων

Σχετικά με την εκμετάλλευση υπερσυνδέσμων για την εξαγωγή γεωγραφικής πληροφορίας, είναι λογικό να υποθέσουμε πως ένα link που βρίσκεται οπτικά κοντά σε γεωγραφική πληροφορία, οδηγεί σε σελίδα που πιθανώς έχει κι αυτή παρεμφερές γεωγραφικό περιεχόμενο. Δηλαδή, υπό προϋποθέσεις, η γεωγραφική θέση είναι ιδιότητα μεταβατική (μέσω υπερσυνδέσμων). Οδηγούμαστε έτσι στο να υπολογίσουμε ένα "μεταβατικό κλείσιμο" της κάθε γεωγραφικής πληροφορίας. Έτσι, έχουμε μια άλλη πηγή γεωγραφικής πληροφορίας.

Δύο σημεία χρήζουν προσοχής: Κατ'αρχήν, η βεβαιότητα της πληροφορίας πρέπει να εξασθενεί με κάθε υπερσύνδεσμο που ακολουθούμε, αλλιώς θα καταλήξουμε σε εσφαλμένα συμπεράσματα.

Ακόμη, έχει σημασία ο υπερσύνδεσμος και η αρχική γεωγραφική πληροφορία να βρίσκονται κοντά. Η εγγύτητα αυτή ιδανικά θα ήταν το πόσο κοντά θα φαίνονταν σε έναν φυλλομετρητή. Για τον προσδιορισμό της σε εύλογο χρόνο θα μπορούσαμε να κάνουμε μερικώς render τη σελίδα, και να μετρήσουμε την απόσταση των δύο στοιχείων (όπως προτείνεται στο [RM01]). Μια ευκολότερη, εναλλακτική μετρική, είναι η κειμενική απόσταση υπερσυνδέσμου και γεωγραφικής πληροφορίας. Στο [HNV+03], για παράδειγμα, επισημαίνεται ότι το κείμενο από 5 λέξεις πριν έως και 5 λέξεις μετά από έναν υπερσύνδεσμο είναι συνήθως σχετικό με την σελίδα-στόχο.

3.2.3.1 Συμπεράσματα από την εκμετάλλευση υπερσυνδέσμων

Έπειτα από μια ταχεία προτυποποίηση της ιδέας αυτής (εκμετάλλευση υπερσυνδέσμων), όμως, παρατηρήθηκε ότι η αναμενόμενη ποσότητα γεωγραφικής πληροφορίας που θα κερδίζαμε ήταν αρκετά μικρή για να δικαιολογήσει τον κόπο απόκτησής της. Έτσι η ιδέα εγκαταλείφθηκε, χάριν βελτίωσης του υπόλοιπου συστήματος.

3.3 Geoparsing - Επίδοση

Κλείνοντας τα περί geoparsing, αναφέρουμε μερικά ποιοτικά στοιχεία σχετικά με την επίδοση του συστήματος, όσον αφορά τη διαδικασία αυτή. Η χρήση του (γρήγορου) JFlex[19] για την υλοποίηση των διαδοχικών κανονικών γραμματικών, σε συνδυασμό με τον βελτιστοποιημένο κώδικα, οδήγησε σε χρόνο ανάλυσης μιας σελίδας ακόμη καλύτερο από τον ήδη μικρό χρόνο που επιτυγχάνετο με το GATE. Ενδεικτικά, με την πρώτη υλοποίηση, στο περιβάλλον GATE, ο μέσος χρόνος ανάλυσης σελίδας ήταν περίπου 2"-2.5" (σε υπολογιστή παλαιάς τεχνολογίας, συνυπολογίζοντας τον χρόνο φόρτωσης της σελίδας), ενώ με την επανϋλοποίηση του συστήματος, ο χρόνος αυτός κατήλθε στα 1" - 1.5" (στον ίδιο υπολογιστή).

3.4 Geoparsing - Ανακεφαλαίωση

Μέχρι στιγμής εξετάσαμε την διαδικασία του geoparsing. Είδαμε τα 4 επίπεδα από τα οποία μπορεί να εξαχθεί γεωγραφική πληροφορία από μια ιστοσελίδα - δικτύου, σύνταξης, σημασιολογίας και τοπολογίας. Εξετάσαμε υπάρχουσες προσεγγίσεις, και παρουσιάσαμε την δική μας, που συνίσταται σε διαδοχικούς μετατροπείς κανονικών γραμματικών, συνεπικουρούμενους από προσεγγιστική αναζήτηση (για τα πιθανά τοπωνύμια).

3.5 Γεωκωδικοποίηση - Εισαγωγή

Θα στραφούμε τώρα στη διαδικασία της γεωκωδικοποίησης. Θα περιγράψουμε τις υπάρχουσες πρακτικές στον χώρο, όπως και για την δική μας προσέγγιση.

Θα συνεχίσουμε με το άμεσα συνδεόμενο θέμα των γεωγραφικών δεδομένων. Αφού αναφέρουμε λίγα απαραίτητα στοιχεία γεωδαισίας, θα συνεχίσουμε με τα προβλήματα των γεωδεδομένων στην Ελλάδα. Ορμώμενοι από αυτό, θα αναφέρουμε μερικά στοιχεία για τον καθαρισμό δεδομένων, ζήτημα που θα μας απασχολήσει και σε επόμενο κεφάλαιο. Τέλος, θα περιγράψουμε τον τρόπο δημιουργίας και οργάνωσης της βάσης γεωγραφικών δεδομένων του συστήματός μας.

[19] http://jflex.de/ ,ίσχυε την 5/7/2006

3.6 Γεωκωδικοποίηση - Προσεγγίσεις

Άμα τη ευρέσει της γεωγραφικής πληροφορίας, πρέπει να ακολουθήσει κάποιος μετασχηματισμός της, ώστε να μετατραπεί σε ένα ζεύγος ή μια τριάδα συντεταγμένων. Εάν η πληροφορία ήταν κατά κάποιο τρόπο μονοσήμαντη, π.χ. ένα τηλέφωνο, ή ένας ταχυδρομικός κώδικας, ο μετασχηματισμός είναι σχετικά προφανής: το στοιχείο που βρέθηκε θα αναζητηθεί σε έναν πίνακα, όπου θα αντιστοιχίζεται μονοσήμαντα σε μια θέση στο χώρο (Τι γίνεται αν δεν είναι διαθέσιμος αυτός ο πίνακας; Θα εξετάσουμε αυτήν την περίπτωση σύντομα...).

3.6.1 Ανάγκη για προσεγγιστική αναζήτηση συμβολοσειρών

Εάν, όμως, η πληροφορία είναι αμφίσημη (π.χ. διεύθυνση, τοπωνύμιο), η πολυπλοκότητα του εγχειρήματος αυξάνεται... Πλέον, απαιτούνται τεχνικές από τον καθαρισμό δεδομένων, ώστε να αποφασίσουμε ποια γεωγραφική οντότητα αντιπροσωπεύει η πληροφορία που βρήκαμε. Η τυπική μεθοδολογία που ακολουθείται είναι αυτή που παρατηρείται στο διαδεδομένο πακέτο γεωκωδικοποίησης της ESRI. Η πληροφορία αρχικά κανονικοποιείται, δηλαδή τα πεδία της γράφονται με μια τυπική σειρά (π.χ. για διεύθυνση: Οδός, Αριθμός, Περιοχή, Τ.Κ.), ενώ λέξεις με πολλαπλές γραφές, μεταγράφονται στην συνηθέστερη (π.χ. οι "οδ.", "ΟΔΟΣ", "οδού" γίνονται όλες "οδός"). Έπειτα, αναζητούμε εγγραφές, από έναν κατάλογο, που να είναι παρόμοιες με την ευρεθείσα. Το πρόβλημα που καλούνται να λύσουν οι διάφορες λύσεις και παραλλαγές τους είναι η επίδοση, καθώς η η αναζήτηση "ομοιότητας" είναι διαδικασία χρονοβόρα και δαπανηρή σε πόρους. Περισσότερα για αυτό το σημαντικό ζήτημα, θα αναφέρουμε στο κεφάλαιο 5.

3.6.2 Ολοκλήρωση αποτελεσμάτων

Αφού βρεθούν και γεωκωδικοποίηθούν οι διάφορες πληροφορίες, πρέπει να συνεκτιμηθούν και να ολοκληρωθούν, ώστε να προκύψει μια γενικότερη γνώση για την ιστοσελίδα που τις περιέχει. Βασικές παράμετροι σε αυτήν την ολοκλήρωση είναι η γεωγραφική κατανομή των αποτελεσμάτων, όπως και η κατανομή τους στην ιστοσελίδα.

Στο [MAH+03] προτείνεται η συσταδοποίηση των αποτελεσμάτων, με τεχνικές από το αντίστοιχο γνωστικό πεδίο. Οι συστάδες πληροφορίας χρησιμοποιούνται, έπειτα, για την εξαγωγή γνώσης από τη σελίδα. Για την συσταδοποίηση αυτή, βέβαια, όπως και για οποιαδήποτε εκμετάλλευση της κατανομής των αποτελεσμάτων στη σελίδα, χρειαζόμαστε μια μετρική (οπτικής) απόστασης δύο στοιχείων σε μια ιστοσελίδα (δηλαδή, πόσο κοντά θα φαίνονταν σε ένα χρήστη). Για λόγους απόδοσης, δεν μπορούμε να απεικονίσουμε τη σελίδα όπως θα έκανε ένας φυλλομετρητής, και μετά να μετρήσουμε τη ζητούμενη απόσταση. Στο HiWE (Hidden Web Explorer, [RM01]) προτείνεται το μερικό rendering της ιστοσελίδας (με παράλληλο "κλάδεμα" του συντακτικού δέντρου HTML στους κόμβους που δεν μας ενδιαφέρουν), και έπειτα η προσεγγιστική

μέτρηση της ζητούμενης απόστασης. Μια ευκολότερη, εναλλακτική μετρική, είναι η κειμενική απόσταση υπερσυνδέσμου και γεωγραφικής πληροφορίας. Όπως έχουμε ήδη αναφέρει, στο [HNV+03], για παράδειγμα, επισημαίνεται ότι το κείμενο από 5 λέξεις πριν, έως και 5 λέξεις μετά από έναν υπερσύνδεσμο είναι σχετικό με την σελίδα-στόχο.

Στον αντίποδα, δηλαδή με ζητήματα γεωγραφικής κατανομής των αποτελεσμάτων σε μια ιστοσελίδα, ασχολείται το [LST+03]. Σε αυτό, επιχειρείται να προσδιοριστεί η βέλτιστη γεωγραφική περιοχή στην οποία αναφέρεται μια ιστοσελίδα, με την χρήση ελαχίστων περικλειόντων ορθογωνίων, και χωρικών λειτουργιών (χωρική ένωση, τομή, κ.ο.κ.).

3.6.3 Επεξεργασία οδηγούμενη από την γεωκωδικοποίηση

Τέλος, μια ολότελα διαφορετική προσέγγιση, όπως έχουμε ήδη αναφέρει, παρουσιάζεται στην μηχανή γεωγραφικής αναζήτησης της Metacarta (βλ. κεφ. 2, Αρχιτεκτονική, και κεφ. 2, υποσημείωση 8). Εκεί, η επεξεργασία μιας ιστοσελίδας οδηγείται από την γεωκωδικοποίηση (αντί το geoparsing), δημιουργώντας διαφορετικές προκλήσεις και απαιτήσεις από την μηχανή.

3.7 Γεωκωδικοποίηση - Η δική μας προσέγγιση

Έχοντας κατά νου τα προηγούμενα, αναπτύξαμε το κομμάτι του συστήματος που ασχολείται με την γεωκωδικοποίηση. Αρχικά συγκεντρώσαμε επαρκή γεωγραφικά δεδομένα για τον ελλαδικό χώρο (βλ. παρακάτω). Έπειτα, αναπτύξαμε αλγορίθμους για την προσεγγιστική αναζήτηση (βλ. κεφ. 5) όσων συμβολοσειρών χρειάζεται να αναζητηθούν (τοπωνύμια, διευθύνσεις κ.λ.π.). Ακολούθως, μια απλή λογική, με βάση την οποία γεωκωδικοποιείται η γεωγραφική πληροφορία, και της αντιστοιχίζεται συντελεστής βεβαιότητας, και ελάχιστο περιβάλλον ορθογώνιο. Ο συντελεστής βεβαιότητας συνοψίζει, αριθμητικά, την βεβαιότητα, αφ'ενός, ότι πρόκειται περί γεωγραφικής πληροφορίας, αφ'ετέρου, πως η γεωγραφική θέση αυτής είναι η ευρεθείσα. Το δε ελάχιστο περιβάλλον ορθογώνιο περιγράφει την έκταση που έχει η θέση της πληροφορίας (δηλ. το γεωγραφικό της εύρος).

3.7.1 Παράδειγμα γεωκωδικοποίησης

Για παράδειγμα, ας θεωρήσουμε μια διεύθυνση: Εφόσον συνοδεύεται από όνομα περιοχής, ή Τ.Κ., θα αναζητήσουμε τα στοιχεία αυτά στους αντίστοιχους καταλόγους. Εάν η διεύθυνση βρίσκεται εντός Αττικής (όπου τα γεωδεδομένα μας έχουν πυκνότερη κάλυψη), θα αναζητήσουμε και το όνομα του δρόμου στον αντίστοιχο κατάλογο. Τέλος, θα ελέγξουμε ότι οι γεωγραφικές θέσεις που λαμβάνουμε από τα επιμέρους στοιχεία συμφωνούν μεταξύ τους (δηλ. έχουν μη κενή τομή). Σε αντίθετη περίπτωση, μειώνουμε τον συντελεστή βεβαιότητας κατά ποσοστό ανάλογο της εντοπιζόμενης ασυμφωνίας.

3.7.2 Ολοκλήρωση αποτελέσμάτων

Ακολούθως, τα αποτελέσματα συναθροίζονται και συνεκτιμώνται, με βάση τον συσχετισμό της γεωγραφικής και κειμενικής τοπολογίας της σελίδας. Δηλαδή, εξετάζουμε παράλληλα την θέση της κάθε πληροφορίας στο χάρτη, όπως και μέσα στο κείμενο.

3.7.2.1 Κειμενική τοπολογία

Για την εκτίμηση της κειμενικής τοπολογίας, λαμβάνουμε υπ'όψη την ενότητα του κειμένου (όπως αυτή ορίζεται από τις σχετικές ετικέτες HTML, όπως br, p, κ.λ.π.) όπου ανήκει η κάθε γεωγραφική πληροφορία, όπως και τις (κειμενικές) αποστάσεις της από άλλες γεωγραφικές πληροφορίες (δηλ. τον αριθμό χαρακτήρων που μεσολαβούν ανάμεσα στις δύο πληροφορίες). Η μετρική αποστάσεων αυτή επιλέχθηκε, έναντι άλλων (βλ. εισαγωγή), αφ'ενός χάριν απλότητος, αφ'ετέρου επειδή η περαιτέρω αύξηση της πολυπλοκότητάς της δεν θα επέφερε σημαντική βελτίωση των αποτελεσμάτων, σύμφωνα με διερευνητικές δοκιμές που εκτελέσαμε.

3.7.2.2 Γεωγραφική τοπολογία

Για την εκτίμηση της γεωγραφικής τοπολογίας, χρησιμοποιούμε μια μετρική απόστασης που συνυπολογίζει:

- Τη γεωγραφική απόσταση των κεντροειδών των επιμέρους γεωγραφικών πληροφοριών (δηλαδή, το πόσο απέχουν)
- Την απόσταση μεταξύ των ελαχίστων περιβαλλόντων ορθογωνίων αυτών (για την συνεκτίμηση του εύρους των θέσεών τους), και
- Τον συντελεστή βεβαιότητας της κάθε πληροφορίας

3.7.2.3 Συνεκτίμηση τοπολογιών

Για την συνεκτίμηση των δύο τοπολογιών, αρχικά χρησιμοποιούμε τεχνικές συσταδοποίησης, με μετρική απόστασης έναν σταθμισμένο μέσο γεωγραφικής και κειμενικής απόστασης.

Το αποτέλεσμα αυτής της συσταδοποίησης για κάθε πληροφορία είναι η λήψη μιας από τις ακόλουθες αποφάσεις:

- Αποδοχή της πληροφορίας ως έχει
- Συμψηφισμός (χωρική ένωση ή τομή) της πληροφορίας με γειτονικές της πληροφορίες, ή
- Απόρριψη της πληροφορίας (π.χ. αν αναφέρεται σε θέση παντελώς άσχετη με το υπόλοιπο κείμενο)

Έπειτα, δεδομένων των κειμενικών αποστάσεων των συστάδων που υπολογίστηκαν, υπολογίζουμε το κειμενικό εύρος της κάθε συστάδας γεωγραφικής πληροφορίας.

3.7.2.4 Παράδειγμα ολοκλήρωσης

Παραδείγματος χάριν, ας υποθέσουμε ότι κατόπιν γεωκωδικοποίησης έχουμε το ακόλουθο σενάριο (εντός των {} αναφέρεται περιγραφικά η γεωγραφική θέση της κάθε πληροφορίας):

Πληροφορία_1{Αθήνα} .... Πληροφορία_2{Χαλάνδρι}

.....

Αλλαγή ενότητας

Πληροφορία_3{Πειραιάς} .... Πληροφορία_4{Θεσ/νίκη}

Είναι λογικό να θεωρήσουμε πως οι πληροφορίες 1 και 2 αναφέρονται από κοινού στην περιοχή του Χαλανδρίου (χωρική τομή), και η πληροφορία 3 αναφέρεται στην περιοχή του Πειραιά, άσχετα από τις 1 και 2. Τέλος, η πληροφορία 4 είναι πιθανώς εσφαλμένη, και, εκτός εαν υποστηρίζεται από υψηλό συντελεστή βεβαιότητας, καλό θα είναι να απορριφθεί.

3.8 Γεωκωδικοποίηση - Επίδοση

Κλείνοντας τα περί γεωκωδικοποίησης, αναφέρουμε μερικά ποιοτικά στοιχεία σχετικά με την επίδοση του συστήματος συνολικά, δηλαδή από την φόρτωση της σελίδας από τον σκληρό εώς το geoparsing και τη γεωκωδικοποίηση, και συνάθροιση των αποτελεσμάτων. Βελτιστοποιώντας κρίσιμα τμήματα του κώδικα της γεωκωδικοποίησης, ο χρόνος επεξεργασίας διαμορφώθηκε περίπου σε 4.5" ανά σελίδα (σε υπολογιστή παλαιάς τεχνολογίας).

3.9 Γεωκωδικοποίηση - Ανακεφαλαίωση

Αναλύσαμε τα σχετικά με την διαδικασία της γεωκωδικοποίησης. Είδαμε τις υπάρχουσες προσεγγίσεις, και καταδείξαμε την ανάγκη ύπαρξης μεθόδων προσεγγιστικής αναζήτησης, προετοιμάζοντας το έδαφος για το επόμενο κεφάλαιο.

Περιγράψαμε την δική μας προσέγγιση στη γεωκωδικοποίηση, δίνοντας έμφαση στην ολοκλήρωση και συνάθροιση των αποτελεσμάτων της, με χρήση τεχνικών συσταδοποίησης.

3.10 Γεωδεδομένα

Ακολούθως, θα μιλήσουμε για τα γεωγραφικά δεδομένα, τα προβλήματα δημιουργίας μιας βάσης Ελληνικών γεωδεδομένων, και τους τρόπους με τους οποίους τα ξεπεράσαμε. Όπως αναφέρθηκε, για την διαδικασία της γεωκωδικοποίησης απαιτούνται δεδομένα υψηλής ακρίβειας και ποιότητας. Τα προβλήματα που ανακύπτουν σε αυτήν την διαδικασία είναι αρκετά, όχι ιδιαίτερα συναρπαστικά per se, ενώ χρειάζεται αρκετό χρόνο η επίλυσή τους.

Κατ' αρχήν, εύδηλο είναι ότι τα δεδομένα αυτά είναι, στην απλούστερη των περιπτώσεων[20], κατάλογοι αντιστοίχισης τοπωνυμίων σε γεωγραφικές συντεταγμένες. Τι αναπαριστούν, όμως, αυτές οι συντεταγμένες; Για να απαντήσουμε σε αυτό το ερώτημα, χρειαζόμαστε -λίγα- στοιχεία γεωδαισίας.

Έπειτα, επειδή αναφερόμαστε στην ελληνική πραγματικότητα, αναγκαζόμαστε να απαντήσουμε σε ορισμένα ιδιαίτερα προβλήματα, όπως την έλλειψη γεωδεδομένων, τις ιδιαιτερότητες που έχουν να κάνουν με γεωγραφικά συστήματα συντεταγμένων, καθώς και την ποιότητα των δεδομένων. Ορμώμενοι από το ζήτημα της ποιότητας, κάνουμε μια μικρή εισαγωγή στον καθαρισμό δεδομένων (που στόχο έχει την βελτίωση της ποιότητάς τους). Σε ζητήματα που άπτονται του καθαρισμού δεδομένων αναφερόμαστε και στο επόμενο κεφάλαιο.

Ολοκληρώνοντας την ενότητα, περιγράφουμε την οργάνωση της βάσης γεωδεδομένων που χτίσαμε στα πλαίσια της εργασίας, όπως και την -τυπική- περίπτωση απόκτησης και επεξεργασίας ενός συνόλου γεωδεδομένων.

3.10.1 Στοιχεία γεωδαισίας[21]

Δίχως να μπούμε σε υπερβολικές λεπτομέρειες, αναφέρουμε εν τάχει ορισμένες βασικές έννοιες γεωδαισίας, τις ελάχιστες δυνατές για την υλοποίηση αυτής της εργασίας, και πάντα από την οπτική γωνία ενός μηχανικού υπολογιστών.

Το σχήμα της γης (γεωειδές) είναι αρκετά ακανόνιστο, και μπορεί να προσεγγιστεί με διάφορους τρόπους. Συνηθέστερα, επιλέγεται ένα ελλειψοειδές. Ανάλογα με την εφαρμογή, επιλέγεται είτε ένα ελλειψοειδές που να μειώνει το μέσο σφάλμα προσέγγισης του γεωειδούς σε παγκόσμια κλίμακα (π.χ. το ελλειψοειδές του παγκόσμιου γεωγραφικού συστήματος αναφοράς WGS84), είτε σε επίπεδο τοπικό (π.χ. για τον ελλαδικό χώρο το ελλειψοειδές του συστήματος ΕΓΣΑ87, [Vei88]), (βλ. σχ. 6). Το ελλειψοειδές αυτό, σε συνδυασμό με άλλες παραμέτρους, καλείται γεωγραφικό δεδομένο (datum). Έχοντας ένα γεωγραφικό δεδομένο, μπορούμε πλέον να ορίσουμε γεωγραφικές συντεταγμένες με βάση αυτό (η ακρίβειά τους εξαρτάται από το γεωγραφικό δεδομένο που επιλέξαμε). Οι συντεταγμένες, εφόσον πρόκειται για διδιάστατες, συνηθέστερα δίδονται σε γεωδαιτική μορφή (δηλαδή δύο γωνίες). Υπάρχουν τρόποι μετασχηματισμού συντεταγμένων μεταξύ δύο διαφορετικών γεωγραφικών συστημάτων αναφοράς, αλλά είναι πάντα προσεγγιστικοί, ενώ αν απαιτείται από αυτούς μεγάλη ακρίβεια, αυτοί γίνονται ιδιαίτερα πολύπλοκοι. Συνήθως

[20]Για λιγότερο απλές περιπτώσεις, βλ. κεφ. 2, υποσημείωση 9, περί των γεωδεδομένων που χρησιμοποιεί η υπηρεσία γεωκωδικοποίησης της Metacarta.

[21]Η παράγραφος αυτή βασίζεται στο [Syn04], σε σημειώσεις διαλέξεων του καθ. Ε. Στεφανάκη: http://www.dbnet.ntua.gr/~stefanak/unipi-gis/lect-3.pdf , ίσχυε την 4/7/2006 , καθώς και σε σημειώσεις διαλέξεων του Γ.Σ.Βέργου http://www.teiser.gr/geoplir/mathima405.htm , ίσχυε την 9/6/2006

υποστηρίζονται μόνο από εξειδικευμένα πακέτα λογισμικού (εμπορικά πακέτα, ή, για τον ελλαδικό χώρο, το δωρεάν COORD_GR [Syn04]).

Σχήμα 6: Τοπικό και Παγκόσμιο Σύστημα Αναφοράς

3.10.2 Προβλήματα... Ελληνικά

Έχοντας δει τις απολύτως απαραίτητες έννοιες από τον χώρο της γεωδαισίας, ας επικεντρώσουμε την προσοχή μας στην ελληνική πραγματικότητα και τα προβλήματά της.

3.10.2.1 Έλλειψη δεδομένων

Κατ'αρχήν, επισημαίνουμε την έλλειψη γεωγραφικών δεδομένων. Ενώ σε πολλές χώρες του εξωτερικού, διατίθενται δωρεάν και δημόσια γεωγραφικά δεδομένα υψηλής ποιότητας (λ.χ. στις Η.Π.Α. διατίθεται, μεταξύ άλλων[22], γεωκωδικοποιημένος κατάλογος οδών, πόλεων, ταχυδρομικών κωδίκων, κ.ο.κ.[23], στις περισσότερες χώρες διατίθεται τηλεφωνικός κατάλογος σε ηλεκτρονική μορφή, κ.λ.π.), στην Ελλάδα δεν υπάρχει κάποια κεντρική υπηρεσία διάθεσής τους. Ορισμένες υπηρεσίες (π.χ. ΕΛΤΑ, ΟΤΕ) προσφέρουν τα δεδομένα τους, αλλά μόνο για μεμονωμένα ερωτήματα στον εξυπηρετητή τους, κάνοντας την χρήση τους από προγράμματα δυσχερή. Συνεπώς, η ανάγκη για γεωδεδομένα καλύπτεται είτε καταφεύγοντας στην δαπανηρή λύση της αγοράς τους από εταιρείες του χώρου, είτε ψάχνοντας εναλλακτικούς τρόπους απόκτησής τους.

[22]http://www.geodata.gov , ίσχυε την 4/7/2006

[23]http://factfinder.census.gov , ίσχυε την 4/7/2006

3.10.2.2 Ποιότητα δεδομένων

Αγνοώντας προς το παρόν την ένδειά τους, τα ελληνικά γεωδεδομένα υποφέρουν και από προβλήματα πληρότητας, ποιότητας και ακρίβειας. Σχετικά με την πληρότητα, επειδή ο πληθυσμός της χώρας μας είναι ανισομερώς κατανεμημένος, είναι πολύ ευκολότερο ή/και φθηνότερο να βρεθούν γεωδεδομένα για μια πυκνοκατοικημένη περιοχή, π.χ. για τον νομό Αττικής (όπου, άλλωστε, το κόστος απόσβεσης της απόκτησης των σχετικών γεωδεδομένων καλύπτεται γρήγορα), απ'ότι για μια επαρχιακή. Η υποβέλτιστη ποιότητα των δεδομένων προκύπτει από την άντλησή τους από ετερόκλητες πηγές, πολλές από τις οποίες δεν διασφαλίζουν την ποιότητά των δεδομένων τους. Ακόμη, επειδή τα περισσότερα ελληνικά γεωδεδομένα διατίθενται από ξένες υπηρεσίες (π.χ. την γεωγραφική υπηρεσία των ΗΠΑ), και άρα με λατινικούς χαρακτήρες, τίθεται ζήτημα μεταγραφής τους στα ελληνικά. Τέλος, η χρήση κατά το παρελθόν ποικίλων συστημάτων συντεταγμένων (Bessel, HATT κ.λ.π.), τα οποία τροποποιούνταν συνεχώς, αλλά και η έλλειψη εύχρηστων εφαρμογών μετατροπής συντεταγμένων μεταξύ του παγκόσμιου συστήματος αναφοράς WGS84, και του τοπικά χρησιμοποιούμενου ΕΓΣΑ87, έχει οδηγήσει αρκετά datasets γεωδεδομένων στο να περιέχουν σημαντικά σφάλματα ή μειωμένη ακρίβεια στις συντεταγμένες τους.

3.10.3 Καθαρισμός δεδομένων

Ο καθαρισμός δεδομένων, που αναλύεται περαιτέρω στο επόμενο κεφάλαιο, είναι μια διαδικασία που αποσκοπεί στην βελτίωση της ποιότητας (και, άρα, της αξίας) ενός συνόλου δεδομένων (dataset). Αυτό επιτυγχάνεται με διάφορες τεχνικές, που συνήθως αποτελούν μείξη τυπικών μεθόδων και εξειδικευμένων, για την συγκεκριμένη περίπτωση, λύσεων. Ο κύριος στόχος της διαδικασίας είναι η εξασφάλιση της ορθότητας του κάθε δεδομένου (π.χ. μια τοποθεσία στην Ελλάδα δεν μπορεί να έχει συντεταγμένες που αναφέρονται στη Σιβηρία), της συνέπειας των δεδομένων ενός συνόλου (π.χ. μπορεί να μην είναι επιθυμητό δύο εγγραφές να αναφέρονται στην ίδια τοποθεσία), καθώς και της συνέπειας μεταξύ των διαφόρων συνόλων δεδομένων (π.χ. σε ένα σύνολο δεδομένων ο Τ.Κ. 35323 αντιστοιχίζεται σε μια περιοχή στη Μακεδονία, ενώ σε ένα άλλο σύνολο, μια οδός στη Λαμία έχει τον ίδιο Τ.Κ.). Ακόμη, η διαδικασία πρέπει να είναι οργανωμένη σε μορφή ροής, στην είσοδο της οποίας δίνονται τα ακατέργαστα δεδομένα, και από έξοδο της οποίας λαμβάνονται καθαρά. Ο λόγος για αυτή την αρχιτεκτονική είναι, αφ'ενός η αυτοματοποίηση της διαδικασίας, αφ'ετέρου η δυνατότητα αυξητικής (incremental) εισαγωγής νέων δεδομένων στο σύστημα. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο [RH00] για περισσότερα σχετικά με το γνωστικό αντικείμενο αυτό. Στην μεθεπόμενη παράγραφο, θα δούμε μια τυπική εφαρμογή καθαρισμού δεδομένων, στα πλαίσια ενός συνόλου γεω-δεδομένων.

3.10.4 Οργάνωση της βάσης γεω-δεδομένων του συστήματος

Ξεπερνώντας, ενίοτε με τρόπους ευρηματικούς, την έλλειψη δωρεάν γεωδεδομένων, και καθαρίζοντάς τα σχολαστικά, καταλήξαμε στην δημιουργία μιας καλής βάσης ελληνικών γεωδεδομένων. Εκτός των εκάστοτε πληροφοριών του κάθε πίνακα, αυτός περιέχει πάντα το τοπωνύμιο γραμμένο με εναλλακτικές μορφές (π.χ. "οδ. Ελευθερίου Βενιζέλου"→ "οδός Ελευθερίου Βενιζέλου", "Ελευθερίου Βενιζέλου"), και πάντα σε φωνητικό αλφάβητο, καθώς και τις συντεταγμένες της τοποθεσίας στο datum WGS84. Ακολουθεί μια σύντομη περιγραφή των κυριότερων πινάκων της που χρησιμοποιούνται τελικά στην γεωκωδικοποίηση, όπως έχουν προκύψει από τη σχετική επεξεργασία:

- Πίνακες attiki: Το σύνολο δεδομένων αυτό περιέχει τις περισσότερες οδούς (~6.100) και πλατείες (~500) της Αττικής, γεωκωδικοποιημένες με ακρίβεια 300 μέτρων.
- Πίνακας gns: Το σύνολο δεδομένων αυτό, περιέχει πολλά (~44.000) ελληνικά τοπωνύμια, με αρκετούς εναλλακτικούς τρόπους γραφής, σε συνδυασμό με την προσεγγιστική τους θέση στο χάρτη. Προέρχεται από την στρατιωτική χαρτογραφική υπηρεσία των ΗΠΑ[24]. Έχει υποστεί αρκετή επεξεργασία, κυρίως στο ζήτημα της μετατροπής από greeklish σε Ελληνικά, καθώς και στην δημιουργία των εναλλακτικών τρόπων γραφής του κάθε τοπωνυμίου (π.χ. "Porto di Accandia"→ "Λιμένας Ακανδίας", "Ακανδία", "Πόρτο ντι Ακάντια"). Μεταξύ άλλων περιέχει πεδία: μοναδικής ταυτοποίησης τοποθεσίας (UFI - Unique Feature Identifier), τοπωνυμίου (UNI - Unique Name Identifier), τύπου τοποθεσίας (π.χ. κατοικημένη περιοχή, πρωτεύουσα νομού κ.ο.κ. - δυστυχώς το πεδίο αυτό είναι αρκετά ανακριβές).
- Πίνακες mapdekode: Το σύνολο δεδομένων αυτό, περιέχει ελληνικά τοπωνύμια (6000 το πλήθος), και τις κύριες οδούς της Ελλάδας (36.000 οδούς). Έχει δημιουργηθεί ερασιτεχνικά από τον φιλέλληνα Γερμανό Peter, και διατίθεται δωρεάν στο διαδίκτυο[25]. Δυστυχώς, στην παρούσα έκδοση, οι γεωγραφικές συντεταγμένες που περιέχει δεν είναι απολύτως ορθές (όπως μπορεί να πειστεί ο καθένας με χρήση GPS), αλλά το dataset αυτό αποτελεί ένα καλό ξεκίνημα για το χτίσιμο της βάσης. Η επεξεργασία του εστιάστηκε, κι εδώ, κυρίως στην γλωσσική επεξεργασία των τοπωνυμίων.
- Πίνακας mjroads: Το σύνολο δεδομένων αυτό δεν είναι ελεύθερα διαθέσιμο, περιέχει, δε, τις κύριες οδικές αρτηρίες της Ελλάδας (2.500 το πλήθος), γραμμένες, ευτυχώς, στα Ελληνικά. Η επεξεργασία του αφορούσε κυρίως τις συντεταγμένες του, που είναι στο τοπικό datum ΕΓΣΑ87.

[24]http://gnswww.nga.mil/geonames/GNS/index.jsp , ίσχυε την 10/4/2006

[25]http://two.fsphost.com/elsinga/gps/maps/ ή http://www.elsinga.org , ίσχυε την 5/7/2006

- Πίνακες telephone: Περιέχουν όλα τα προθέματα τηλεφώνων της Ελλάδας (~240), καθώς και πολλά (~1.800.000) τηλέφωνα στην Αττική, αμφότερα γεωκωδικοποιημένα. Χρειάστηκε εκτεταμένη επεξεργασία, τόσο για τον καθαρισμό των αρχικών δεδομένων, όσο και για την γεωκωδικοποίησή τους, με χρήση των υπόλοιπων datasets.
- Πίνακες postal: Περιέχουν όλους τους ταχυδρομικούς κώδικες της Ελλάδας (~900), γεωκωδικοποιημένους. Τα δεδομένα προέρχονται από το διαδίκτυο[26], ενώ η γεωκωδικοποίησή τους έγινε με χρήση των υπόλοιπων datasets.

3.10.5 Δημιουργία των συνόλων δεδομένων

Θα περιγράψουμε, τώρα, σε αδρές γραμμές, τον τρόπο απόκτησης και καθαρισμού δυο εκ των ανωτέρω συνόλων δεδομένων.

3.10.5.1 Πίνακες postal

Έχοντας παρατηρήσει πως οι ταχυδρομικοί κώδικες της Ελλάδας είναι δημόσια διαθέσιμοι [26], δημιουργήθηκε μια μικρή εφαρμογή σε C για την απόκτηση όλων των δεδομένων από τον εξυπηρετητή. Τα δεδομένα που αποκτήθηκαν ήταν της μορφής: {Ταχυδρομικός Κώδικας, Τοποθεσία, Δήμος, Νομός}. Αφού αποκτήθηκαν, μορφοποιήθηκαν και φορτώθηκαν στην βάση δεδομένων, τα τοπωνύμια μεταγράφηκαν σε φωνητικό αλφάβητο και δημιουργήθηκαν εναλλακτικοί τρόποι γραφής του κάθε τοπωνυμίου. Τέλος, κάθε τοπωνύμιο γεωκωδικοποιήθηκε με βάση τα υπόλοιπα datasets, και η ένωση των επιμέρους θέσεών του κάθε ταχυδρομικού κώδικα καθόρισε το γεωγραφικό εύρος του.

3.10.5.2 Πίνακες mapdekode

Χρησιμοποιώντας την εφαρμογή mapdekode[27], και γράφοντας αρκετό κώδικα μετατροπής δεδομένων (αλλαγή μορφοποίησης, μετατροπή και διόρθωση συντεταγμένων, κ.ο.κ.) σε Java, φορτώθηκαν τα δεδομένα στην βάση. Έπειτα, έγινε μια μελέτη της μορφής των greeklish που χρησιμοποιήθηκαν στο dataset αυτό, όπως και της μορφής των εγγραφών (π.χ. ποιες εγγραφές αντιστοιχούν σε πόλεις, και άρα πρέπει να διατηρηθούν, και ποιες σε τράπεζες, φαρμακεία κ.λ.π., και άρα δεν μας αφορούν). Κατόπιν, γράφτηκε εξειδικευμένος κώδικας καθαρισμού του συνόλου δεδομένων, ο οποίος έκανε και χρήση των ήδη αποκτηθέντων δεδομένων (π.χ. για να σιγουρευτούμε ότι η μετατροπή ενός ονόματος οδού έγινε σωστά, κάνουμε προσεγγιστική αναζήτησή του στα ήδη υπάρχοντα ονόματα οδών).

[26]http://www.postal.gr , ίσχυε την 5/7/2006

[27]http://paginas.terra.com.br/informatica/download1/dekode_download.htm , ίσχυε την 5/7/2006

Σημειωτέον ότι συγγραφέας της εφαρμογής είναι ο ίδιος ο Peter που έφτιαξε το dataset από το οποίο προήλθαν οι πίνακες mapdekode της βάσης.

3.11 Σύνοψη κεφαλαίου

Στο κεφάλαιο αυτό, μιλήσαμε για την διαδικασία του geoparsing και της γεωκωδικοποίησης.

Σχετικά με το geoparsing, είδαμε τα 4 επίπεδα από τα οποία μπορεί να εξαχθεί γεωγραφική πληροφορία από μια ιστοσελίδα - δικτύου, σύνταξης, σημασιολογίας και τοπολογίας, εξετάσαμε υπάρχουσες προσεγγίσεις, και παρουσιάσαμε την δική μας, που συνίσταται σε διαδοχικούς μετατροπείς κανονικών γραμματικών, συνεπικουρούμενους από προσεγγιστική αναζήτηση (για τα πιθανά τοπωνύμια).

Περιγράψαμε και την υπάρχουσα πρακτική όσον αφορά τη γεωκωδικοποίηση, την ανάγκη μεθόδων προσεγγιστικής αναζήτησης, και τη δική μας προσέγγιση στη γεωκωδικοποίηση, με έμφαση στην ολοκλήρωση και συνάθροιση των αποτελεσμάτων της.

Τέλος, αναφερθήκαμε στα γεωδεδομένα, τα προβλήματα δημιουργίας μιας βάσης Ελληνικών γεωδεδομένων, και τους τρόπους με τους οποίους τα ξεπεράσαμε.

4 *Προσεγγιστικό ταίριαγμα και αναζήτηση*

Όπως έχει ήδη αναφερθεί, κεντρικό ρόλο στην εργασία αυτή παίζουν οι αλγόριθμοι ταιριάγματος και αναζήτησης. Είτε στην φάση του καθαρισμού δεδομένων, όπου καλούμαστε να αποφανθούμε για την εννοιολογική ταύτιση δύο συμβολοσειρών (δηλ. η οδός "Μ. Μπότσαρη" είναι ίδια με την "Μπότσαρι Μάρκου" ;), είτε κατά την προσεγγιστική αναζήτηση μιας συμβολοσειράς σε ευρετήριο (π.χ. βρες την οδό "Μπότσαρι Μάρκου" στο ευρετήριο).

Γιατί όμως να έχουμε προσεγγιστικό ταίριαγμα; Η κλασσική αναζήτηση ισότητας (δηλ. δύο συμβολοσειρές αντιπροσωπεύουν το ίδιο αντικείμενο ανν ταυτίζονται), προφανώς δεν λειτουργεί στην περίπτωσή μας. Επομένως, χρειαζόμαστε μετρικές ομοιότητας, προκειμένου να αποφανθούμε για την ομοιότητα ή όχι δυο συμβολοσειρών.

Επιπλέον, γιατί να απαιτούνται εξειδικευμένοι αλγόριθμοι για προσεγγιστική αναζήτηση; Δυστυχώς, όλες οι μετρικές ομοιότητας έχουν μεγάλο κόστος εκτέλεσης (μετρούμενο σε χρονικούς και χωρικούς πόρους). Έτσι, χρειαζόμαστε "έξυπνες" ιδέες για την μείωση του αριθμού των συγκρίσεων που γίνονται, ώστε να καταστεί εφικτός ένας αλγόριθμος προσεγγιστικής αναζήτησης.

4.1 Οργάνωση του κεφαλαίου

Η επισκόπηση του χώρου ξεκινά με ένα γνωστό πρόβλημα από το πεδίο του καθαρισμού δεδομένων, το πρόβλημα ταυτότητας αντικειμένου, δηλαδή, δοθεισών δύο πλειάδων, να αποφασιστεί αν αυτές αντιπροσωπεύουν το ίδιο πραγματικό αντικείμενο. Από την δουλειά που έχει γίνει πάνω σε αυτό το πρόβλημα, αντλούμε βασικές έννοιες και μεθοδολογίες για την αντιμετώπιση του προσεγγιστικού ταιριάγματος και αναζήτησης. Πράγματι, στα πλαίσια αυτού, παρουσιάζουμε

μερικές από τις πιο διαδεδομένες μετρικές ομοιότητας συμβολοσειρών, όπως και αρκετές γνωστές λύσεις στο πρόβλημα της προσεγγιστικής αναζήτησης.

Έπειτα, αναλύουμε την προσέγγισή μας, όσον αφορά το προσεγγιστικό και φωνητικό ταίριαγμα συμβολοσειρών, και την προσαρμογή της στην ελληνική πραγματικότητα. Περιγράφουμε το φαινόμενο των greeklish, τα προβλήματα που μας προξενεί, και τους διάφορους τρόπους αντιμετώπισής τους, μαζί με την δική μας προσέγγιση στο ζήτημα.

Το κεφάλαιο κλείνει με την παρουσίαση των αλγορίθμων που αναπτύξαμε για την επίλυση του προβλήματος της προσεγγιστικής αναζήτησης.

4.2 Το πρόβλημα ταυτότητας αντικειμένου[28]

Στην βιβλιογραφία υπάρχει μια σχετική σύγχυση για την ονοματολογία του προβλήματος αυτού. Εμείς επιλέξαμε την πρόταση των [GFS+01], που το ονομάζουν Object Identity Problem. Πάντως, αναφέρεται και ως Merge/Purge Problem [HS98], και Data De-duplication [Mon00], ή Data Cleansing.

4.2.1 Ορισμός

Γενικά, μιλάμε για πρόβλημα ταυτότητας αντικειμένου όταν υπάρχει δυσκολία να "καταλάβει" ένα σύστημα πως τουλάχιστον δύο πλειάδες σε μία ή περισσότερες σχέσεις αναπαριστούν την ίδια οντότητα.

4.2.2 Σύνδεση με την εργασία

Αν και, εκ πρώτης όψεως, μπορεί το πρόβλημα αυτό να φαντάζει απομακρυσμένο από το κυρίως αντικείμενο της εργασίας αυτής, εντούτοις οι λύσεις του σχετίζεται στενά με αυτήν, με δύο τρόπους. Αφ'ενός, άμεσα, στο χρονοβόρο στάδιο του καθαρισμού των γεωδεδομένων (βλ. σχετική παράγραφο στο κεφ. 4), όπου χρησιμοποιήθηκαν, αυτούσιοι ή τροποποιημένοι, αλγόριθμοι στους οποίους αναφερόμαστε παρακάτω. Αφ'ετέρου, έμμεσα, ως βάση για την πρόταση λύσεων στο συναφές πρόβλημα του προσεγγιστικού ταιριάγματος και της προσεγγιστικής αναζήτησης.

4.2.3 Επιμέρους προβλήματα

Για τον αποδοτικό εντοπισμό "διπλότυπων" πλειάδων, λοιπόν, έχουν προταθεί αρκετοί αλγόριθμοι, τους οποίους θα παρουσιάσουμε συγκριτικά ακολούθως. Ακόμη, όταν εντοπιστούν οι πλειάδες αυτές, πρέπει να γίνει κάποια επιλογή: Είτε θα πρέπει να επιλεγεί μια εξ αυτών (π.χ. αυτή

[28]Η παράγραφος αυτή βασίζεται, εν μέρει, σε εργασία του γράφοντος και της Α.Κούρτη στα πλαίσια του μαθήματος "Προχωρημένα Θέματα Βάσεων Δεδομένων", 2006, με τίτλο "Καθαρισμός δεδομένων και το πρόβλημα ταυτότητας αντικειμένου".

με τα περισσότερα μη κενά πεδία) ως η πλέον ορθή, και οι υπόλοιπες να απορριφθούν, ή με κάποιο τρόπο η πληροφορία που περιέχεται σε αυτές, να ενσωματωθεί σε μία ακριβώς πλειάδα, η οποία θα περιέχει ορθότερη πληροφορία, από καθεμία από τις αρχικές. Και οι δύο επιλογές παρουσιάζουν δυσκολίες (π.χ. πώς ξέρουμε ποιά είναι η ορθότερη εγγραφή; Πώς ενσωματώνουμε πληροφορία από πολλαπλές πλειάδες σε μία;). Σε αυτό το σημείο καλείται η γνώση επί του πεδίου των δεδομένων (domain knowledge) να λύσει το πρόβλημα – μάλιστα δεν γνωρίζουμε κάποια γενική μέθοδο αντιμετώπισης διπλοτύπων.

4.2.4 Αιτίες του προβλήματος ταυτότητας αντικειμένου

Ας δούμε, συνοπτικά, μερικές από τις αιτίες του προβλήματος ταυτότητας αντικειμένου:

- Λάθος τιμές πεδίων, ή πεδία που λείπουν, εξ'αιτίας σφαλμάτων κατά την εισαγωγή δεδομένων. Π.χ.

Γεράσιμος	Σπανοδημήτρης	2102252252	Αθήνα	Άρρεν
Άσιμος	Σπανοδημήτρης	2102222252	**null**	Άρρεν

- Διαφορά στην ονοματολογία (π.χ. σε μία βάση έχουμε την κωδικοποίηση true= =Φίλαθλος false= =Μη φίλαθλος, ενώ στην άλλη έχουμε Φ= =Φίλαθλος και ΜΦ= =Μη φίλαθλος). Αυτό το πρόβλημα μπορεί να λυθεί εύκολα με μεθόδους μετάφρασης σχημάτων, και δεν χρειάζεται να μας απασχολήσει ιδιαίτερα εδώ.
- Μη πλήρης πληροφορία, γιατί τα απαραίτητα δεδομένα δεν καταχωρήθηκαν ή δεν είναι διαθέσιμα (π.χ. μια δημοσκόπηση καταγράφει μόνο τα αρχικά των ονομάτων των συμμετεχόντων)
- Αλλαγή σε κάποιο πεδίο ενδεχομένως δεν μεταβάλλει την ταυτότητα του αντικειμένου (π.χ. ο Γεράσιμος του παραδείγματος μετακομίζει στην Λάρισα)
- (Υστερόβουλη) Ψευδής δήλωση στοιχείων: Με τις αποθήκες δεδομένων να χρησιμοποιούνται για ανίχνευση απατών, ξεπλύματος χρήματος, αλλά και παρακολούθησης πολιτών, είναι αναμενόμενο κάποιος να εισαγάγει ψευδή στοιχεία, ή επιτηδευμένα ανορθόγραφα στοιχεία.

4.2.5 Λύσεις στο πρόβλημα ταυτότητας αντικειμένου

Ας περάσουμε τώρα σε προτάσεις για την λύση του προβλήματος αυτού. Θα εξετάσουμε πρώτα τις λύσεις στο υποπρόβλημα του ταιριάγματος πεδίων (προσεγγιστικό ταίριαγμα), κι έπειτα στο υποπρόβλημα της απαλοιφής διπλοτύπων.

4.2.6 Το υποπρόβλημα ταιριάγματος πεδίων

4.2.6.1 Ορισμός

Το πρώτο ερώτημα που θα μας απασχολήσει είναι το "πώς θα καταλάβουμε ότι δυο πεδία είναι ισοδύναμα" (αναφέρεται ως field matching problem [ME96]). Ακριβέστερα, επειδή ζούμε σ'έναν κόσμο ατελή, ψάχνουμε να υπολογίσουμε τον βαθμό ομοιότητας δύο πεδίων. Στη συνέχεια, εξετάζοντας τους βαθμούς ομοιότητας μεταξύ όλων των πεδίων δύο εγγραφών (κατάλληλα σταθμισμένων), θα μπορέσουμε να αποφανθούμε για το αν αναπαριστούν την ίδια οντότητα.

4.2.6.2 Προεπεξεργασία πεδίων

Σε πολλές περιπτώσεις είναι ωφέλιμη η προεπεξεργασία των πεδίων. Με την χρήση εξωτερικών λεξικών, πολλές φορές σχετικών με τον τομέα του πεδίου εφαρμογής, μπορούμε να διορθώσουμε τυχόν τυπογραφικά λάθη και να κανονικοποιήσουμε τυχόν ακρωνύμια. (π.χ. για τον καθαρισμό ενός ευρετηρίου οδών, θα χρησιμοποιούσαμε έναν κατάλογο τοπωνυμίων της Ελλάδας, καθώς και έναν μικρό κατάλογο ακρωνυμίων της μορφής Α., Αγ. -> Άγιος|Αγία, Λ.->Λεωφόρος κ.ο.κ.) Έτσι ελπίζουμε να έχουμε καλύτερα αποτελέσματα στις επόμενες φάσεις καθαρισμού. Τα πειραματικά αποτελέσματα υποστηρίζουν αυτήν την υπόθεση ([LLL+99],[RH00],[LLL00]).

4.2.6.3 Μετρικές ομοιότητας συμβολοσειρών

Έπειτα, χρειαζόμαστε συναρτήσεις που να μας δείχνουν την ομοιότητα δύο πεδίων. Για να υπολογίσουμε αυτές, χρησιμοποιούμε συναρτήσεις που ποσοτικοποιούν τις διαφορές τους. Στην βιβλιογραφία ονομάζονται **edit distance** συναρτήσεις, έχουν δε πλείστες παραλλαγές. Γενικά υλοποιούνται με μεθόδους δυναμικού προγραμματισμού, με κόστος O(mn), όπου m,n τα μήκη των πεδίων που συγκρίνονται. Στην βασική τους μορφή, όπως προτάθηκαν από τους Damerau και **Levenshtein**, η απόσταση δύο συμβολοσειρών ορίζεται ως ο αριθμός χαρακτήρων που πρέπει να εισαγάγουμε, να διαγράψουμε ή να τροποποιήσουμε προκειμένου να μεταβούμε από την μία συμβολοσειρά στην άλλη. Προφανώς, όσο μικρότερη είναι η απόσταση τόσο πιο όμοια είναι τα πεδία.

Στην παραλλαγή των **Smith και Waterman**, για την ανίχνευση κοινών μοριακών υπακολουθιών, εισάγεται μια νέα παράμετρος $\beta<1$. Μπορούμε ακόμη να εισαγάγουμε, να διαγράψουμε ή να τροποποιήσουμε έναν χαρακτήρα με μοναδιαίο "κόστος", αλλά η διαγραφή ενός block από $\kappa+1$ χαρακτήρες έχει κόστος $\beta\kappa+1$. Με αυτήν την μέθοδο, που χρησιμοποιείται στο [ME96], επιτυγχάνουμε ανίχνευση ακρωνυμίων, π.χ. το "Εργ. Συστ. Β.Δ." θα ταιριάξει με το "Εργαστήριο Συστημάτων Βάσεων Δεδομένων", σε αντίθεση με την κλασσική απόσταση Levenshtein, όπου αυτά θα χαρακτηριστούν ως ανόμοια.

Στην παραλλαγή της **φωνητικής απόστασης**, η τροποποίηση ενός χαρακτήρα δεν έχει μοναδιαίο κόστος, αλλά το κόστος αυτό εξαρτάται από την φωνητική ομοιότητα του παλαιού με τον νέο χαρακτήρα. Π.χ., στα Ελληνικά, η αντικατάσταση του 'π' από το 'φ' δεν είναι τόσο "δαπανηρή" σε απόσταση όσο του 'γ' από το 'α', καθότι τα πρώτα είναι αμφότερα χειλικά σύμφωνα, ενώ τα δεύτερα παντελώς άσχετα μεταξύ των. Σε αυτήν την παραλλαγή, υποθέτουμε πως σημαντική πηγή των λαθών είναι τα ορθογραφικά λάθη μεταξύ λέξεων ηχητικά όμοιων, π.χ. να εισαχθεί η λέξη "Ανασίας" αντί του "Αμασείας".

Στην παραλλαγή της **απόστασης γραφομηχανής** το κόστος τροποποίησης ενός χαρακτήρα εξαρτάται από την απόσταση πάνω στο πληκτρολόγιο του παλαιού με τον νέο χαρακτήρα. Π.χ. για QWERTY πληκτρολόγια, η απόσταση των A,S θεωρείται μικρότερη από αυτήν των Q,P Προφανώς, η παραλλαγή αυτή υποθέτει ότι σημαντική πηγή λαθών είναι τα ορθογραφικά λάθη λόγω λανθασμένης πληκτρολόγησης.

4.2.6.4 Εναλλακτικές λύσεις : Tokenize and sort

Στη βιβλιογραφία ([Mon00],[ME96],[HS98]) έχει προταθεί και η εξής ιδέα (**tokenize and sort**): Δεδομένου ενός ατομικού πεδίου[29], π.χ. Ονόματα Συγγραφέων, μπορούμε να χωρίσουμε το πεδίο σε ατομικές υποσυμβολοσειρές (δηλ. συμβολοσειρές που δεν περιέχουν κενά), και να ταξινομήσουμε αυτές. Έπειτα, υπολογίζουμε κανονικά την εκάστοτε συνάρτηση απόστασης. Η ιδέα αυτή έχει το πλεονέκτημα ότι μια τυχαία μετάθεση, άνευ σημασίας για τον βαθμό ισοδυναμίας δύο πεδίων, δεν επιδρά καθόλου στον υπολογισμό του. Εντούτοις, η μέθοδος είναι δαπανηρότερη, και δεν αποδίδει καλά παρουσία ορθογραφικών σφαλμάτων.

4.2.6.5 Εναλλακτικές λύσεις : Αναδρομικό ταίριαγμα πεδίων

Τέλος, ένας άλλος, **αναδρομικός αλγόριθμος**, που προτείνεται από τους Monge και Elkan στο [ME96] για ταυτοποίηση πεδίου είναι ο εξής: Δυο συμβολοσειρές A και B ταιριάζουν με βαθμό 1 αν ταυτίζονται, ή αν η μία είναι σύντμηση της άλλης[30], αλλιώς ταιριάζουν με βαθμό 0. Κάθε υποπεδίο της A θεωρούμε ότι ταιριάζεται με το υποπεδίο της B με το οποίο έχει μεγαλύτερο βαθμό ομοιότητας. Τότε ο βαθμός ομοιότητας του A με το B (από 0 εως 1) είναι

[29]Ατομικό όχι με την έννοια της πρώτης κανονικής μορφής, αλλά με την έννοια ότι το πεδίο δεν μπορεί να τεχνολογηθεί (τεχνολόγηση = parsing) περαιτέρω σε άλλα αυτοτελή πεδία, όπως για παράδειγμα η Διεύθυνση που μπορεί μέσω τεχνολόγησης να διαιρεθεί σε Όνομα Οδού, Τύπος Οδού, Αριθμός, Περιοχή κ.ο.κ.

[30]Στα πλαίσια αυτού του αλγορίθμου, μια συμβολοσειρά A είναι σύντμηση της B αν

1. A είναι πρόθεμα της B
2. A αποτελείται από ένα πρόθεμα και ένα επίθεμα της B (π.χ. **Dept**. - **Dep**artmen**t**)
3. A αποτελείται από παράθεση προθεμάτων της B (π.χ. ΕΜΠ – Εθνικό Μετσόβιο Πολυτεχνείο, CalTech – California Institute of Technology)

$$match(A, B) = \frac{1}{|A|} \sum_{i=1}^{|A|} \max_{j=1}^{|B|} match(A_i, B_j).$$

Ας σημειωθεί ότι οι ανωτέρω παραλλαγές δίνουν στην πράξη παραπλήσια αποτελέσματα τόσο από πλευράς ποιότητας[31], όσο και πολυπλοκότητας. Στην πράξη συνήθως χρησιμοποιείται κάποια παραλλαγή της απόστασης Levenshtein, ανάλογα και με το πεδίο εφαρμογής.

Έχοντας μια συνάρτηση απόστασης, μπορούμε να υπολογίσουμε την ομοιότητα δύο εγγραφών σαν το σταθμισμένο άθροισμα των ομοιοτήτων των επιμέρους πεδίων τους. Η κατάλληλη **στάθμιση των πεδίων** έχει αποδειχθεί πειραματικά ότι αποδίδει μεγάλη βελτίωση στην ποιότητα των αποτελεσμάτων- εντούτοις μπορεί να προκύψει μόνο από γνώση του πεδίου των δεδομένων, ή από τεχνικές μηχανικής μάθησης.

4.2.7 Το πρόβλημα εντοπισμού (προσεγγιστικών) διπλότυπων

4.2.7.1 Ορισμός, ανάγκες

Με όπλο την μετρική ομοιότητας μεταξύ εγγραφών, πλεον πρέπει να δούμε πώς θα **εντοπίσουμε** όλα τα (**προσεγγιστικά**) **διπλότυπα**. Εάν επρόκειτο για ακριβή διπλότυπα, θα μπορούσαμε να εφαρμόσουμε την κλασσική ιδέα από τις βάσεις δεδομένων : Ταξινόμησε όλον τον πίνακα, και εξέτασε αν τα γειτονικά στοιχεία ταυτίζονται. Όμως εδώ, λόγω των λαθών που προαναφέρθηκαν, η ταξινόμηση μπορεί να μην φέρει σε γειτονικές θέσεις εγγραφές που μοιάζουν. (Π.χ. ας σκεφτούμε το παράδειγμά μας, με ταξινόμηση επί του ονοματεπωνύμου.)

Θα μπορούσε επίσης να σκεφτεί κανείς να συγκρίνει κάθε στοιχείο του πίνακα με κάθε άλλο (**brute force**). Όμως, η αυξημένη πολυπλοκότητα των αλγορίθμων σύγκρισης (είναι τετραγωνική ως προς το μήκος των εγγραφών) θα οδηγούσε σε πολυπλοκότητα $O(m^2n^2)$ (όπου m το μέσο μήκος των εγγραφών, και n το πλήθος τους), που είναι ανεπίτρεπτη στην πράξη (στο [GIJ+01]ανεφέρεται ανεκδοτολογικά πως μια τέτοια ερώτηση επί δεδομένων αποκτηθέντων από πρακτικό πρόβλημα είχε αναμενόμενο χρόνο εκτέλεσης 3 ημερών!).

4.2.7.2 Υλοποίηση εντός του ΣΔΒΔ

Έτσι, πρέπει να χρησιμοποιηθεί κάποια καλύτερη ιδέα. Στο [GIJ+01] προτείνεται η υλοποίηση των συναρτήσεων προσεγγιστικού ταιριάσματος μέσα στα **ενδότερα της βάσης** δεδομένων, χρησιμοποιώντας έτσι ήδη έτοιμη τεχνολογία για την βελτιστοποίηση του χρόνου εκτέλεσης. Αν και

[31]Η ποιότητα σε αυτές τις περιπτώσεις μετράται με βάση δύο μετρικές δανεισμένες από την Εξόρυξη Πληροφορίας: την ανάκληση (recall), που ορίζεται ως το ποσοστό εκ των ταιριαγμάτων που έπρεπε να γίνουν που τελικά έγιναν, και την ακρίβεια (precision), που είναι το ποσοστό εκ των ταιριαγμάτων που έγιναν που πραγματικά έπρεπε να γίνουν.

η ιδέα αυτή δεν μπορεί να βρει εφαρμογή σε όλες τις περιπτώσεις, αξίζει εντούτοις να την αναφέρουμε σύντομα, προτού προχωρήσουμε στην συγκριτική παρουσίαση των πιο επικρατουσών.

Σε γενικές γραμμές, όταν θέλουμε να βρούμε προσεγγιστικά διπλότυπα ως προς ένα πεδίο, δημιουργούμε και αποθηκεύουμε σε ξεχωριστή σχέση στη βάση όλες τις υποσυμβολοσειρές του μήκους q. Έπειτα, εκμεταλλευόμενοι ιδιότητες της συνάρτησης Levenshtein, μπορούμε για κάθε πεδίο να φιλτράρουμε μέσω ερωτημάτων σε απλή SQL τα περισσότερα από τα υπόλοιπα πεδία που δεν ταιριάζουν με αυτό. Για τα λίγα εναπομείναντα πεδία, υπολογίζουμε τελικά τη συνάρτηση Levenshtein. Τελικά, όλη η δουλειά γίνεται μέσω ενός (πολύπλοκου) ερωτήματος SQL, το οποίο μπορεί να βελτιστοποιηθεί από τον βελτιστοποιητή ερωτημάτων του ΣΔΒΔ, με λίγες μόνο (δαπανηρές) κλήσεις της συνάρτησης υπολογισμού απόστασης Levenshtein. Τα πειραματικά αποτελέσματα συνηγορούν στην αποτελεσματικότητα της μεθόδου, η οποία ίσως, μελλοντικά, να υποστηρίζεται εγγενώς από τα ΣΔΒΔ.

4.2.7.3 Μέθοδος Ταξινομημένης Γειτονιάς

Αλλά ας επιστρέψουμε σε λιγότερο εξωτικές λύσεις. Αντί να συγκρίνουμε κάθε στοιχείο του πίνακα με κάθε άλλο, μπορούμε κατ'αρχήν να ταξινομήσουμε τον πίνακα επί ενός πεδίου, και έπειτα να συγκρίνουμε κάθε στοιχείο με τα προηγούμενα w του στοιχεία, όπου το w καλείται μέγεθος παραθύρου (γιατί είναι σαν να κυλάμε ένα παράθυρο σταθερού μεγέθους w επί των δεδομένων) ή γειτονιάς (γιατί συγκρίνουμε το κάθε στοιχείο με τα γειτονικά του). Το κόστος της μεθόδου τώρα γίνεται γραμμικό (O(wn)) ως προς το πλήθος των στοιχείων n, αφού το w είναι τυπικά ένας μικρός, σταθερός αριθμός (για λόγους οικονομίας δεν θα αναφέρουμε εφεξής τον πολλαπλασιαστικό παράγοντα O((μέσο μήκος πεδίου)^2), που επιφέρει η χρήση συναρτήσεων ομοιότητας). Η ιδέα πίσω από αυτήν την πρόταση (που, παρεμπιπτόντως, αναφέρεται ως **Μέθοδος Ταξινομημένης Γειτονιάς**) είναι ότι ελπίζουμε ότι η ταξινόμηση επί ενός πεδίου θα φέρει σχετικά κοντά δύο εγγραφές που αναπαριστούν την ίδια οντότητα.

Μια απλή βελτίωση της μεθόδου αυτής ελέγχει πρώτα για ακριβή αντίγραφα, τα οποία και απορρίπτει, και εκτελεί έπειτα τον ανωτέρω αλγόριθμο, επιφέροντας μικρή βελτίωση στην απόδοση.

4.2.7.4 Tokenize and sort, κλειδιά ταξινόμησης

Ουσιωδέστερες είναι άλλες τροποποιήσεις της μεθόδου, που προσπαθούν να επιτύχουν ποιοτικότερα αποτελέσματα. Για παράδειγμα, αντί να ταξινομούμε επί ενός πεδίου, μπορούμε πρώτα να εφαρμόσουμε μια προαναφερθείσα ιδέα : χωρίζουμε το πεδίο σε ατομικές υποσυμβολοσειρές (**tokenizing**), και τις **ταξινομούμε**. Έτσι είναι πιθανότερο δυο όμοιες εγγραφές να πλησιάσουν με την ταξινόμηση.

Επεκτείνοντας την ιδέα αυτή, μπορούμε να δημιουργήσουμε κάποιο **κλειδί** κάθε εγγραφής, με την προσδοκία δύο τέτοια κλειδιά να είναι λεξικογραφικά κοντά σε εγγραφές που αναπαριστούν την

ίδια οντότητα. Παράδειγμα τέτοιου κλειδιού θα μπορούσε να είναι (στην περίπτωσή π.χ. του πελατολογίου που αναφέρθηκε) τα τρία πρώτα γράμματα του ονοματεπώνυμου (κατόπιν tokenization), ακολουθούμενα από τα ψηφία 4,5 και 6 του τηλεφώνου, ακολουθούμενα από τα πρώτα τρια γράμματα της διεύθυνσης. Έτσι, τα κλειδιά των δύο εγγραφών θα γίνονταν : ΓΑΣ225ΑΘΗ και ΓΕΡ222ΑΘΗ, που είναι όντως λεξικογραφικά κοντά. Μια κλασσική περίπτωση τέτοιου κλειδιού είναι και ο κώδικας Soundex, καθώς και ο διάδοχός του Metaphone, για την περίπτωση κυρίων ονομάτων, διευθύνσεων κ.ο.κ.

4.2.7.5 Φωνητικοί κώδικες

Ο κώδικας Soundex συνίσταται στην επιλογή του πρώτου συμφώνου μιας αγγλικής λέξης, ακολουθούμενου από τρεις αριθμούς που χαρακτηρίζουν τις οικογένειες στις οποίες ανήκουν τα επόμενα τρία σύμφωνα της λέξης. (π.χ. Hampshire → H512). Αν και δεν λειτουργεί καλά ως κώδικας φωνητικής ομοιότητας, εντούτοις είναι ικανοποιητικός για κλειδί για αυτήν την μέθοδο. Ο κώδικας Soundex χρησιμοποιείται κατά κόρον, μεταξύ άλλων, στα προγράμματα της ESRI[32] για την προσεγγιστική φωνητική αναζήτηση συμβολοσειρών κατά την διαδικασία της γεωκωδικοποίησης. Ο κώδικας Metaphone επιχειρεί να λύσει ορισμένα από τα προβλήματα του Soundex, δημιουργώντας κώδικες μεγαλύτερου μήκους, που εκφράζουν περισσότερο τα χαρακτηριστικά της λέξης. Επίσης, χρησιμοποιεί κανόνες αγγλικής φωνητικής, προκειμένου να αντιμετωπίσει συμπλέγματα χαρακτήρων (π.χ. η λέξη though να γίνει αντιληπτή ως "δόου" και όχι π.χ. ως "τχόουγκχ"). (Για μια παρουσίαση των κωδίκων αυτών, καθώς και άλλων παραμφερών, βλ. [LR93])

4.2.7.6 Χρήση διαδοχικών περασμάτων

Έχει παρατηρηθεί πειραματικά, πως ενώ η αύξηση του μεγέθους του παραθύρου επιφέρει κάποιες βελτιώσεις στην ποιότητα των αποτελεσμάτων, η ποιότητα βελτιώνεται πολύ περισσότερο (για το ίδιο χρονικό κόστος), αν εκτελεστούν **διαδοχικά περάσματα** ενός μικρού παραθύρου, χρησιμοποιώντας κάθε φορά διαφορετικό κλειδί.

4.2.7.7 Μεταβατικότητα της ταύτισης

Σε κάθε περίπτωση (ένα ή πολλαπλά περάσματα), τα αποτελέσματα μπορούν να βελτιωθούν χρησιμοποιώντας το **μεταβατικό κλείσιμο** της ιδιότητας "ταύτιση οντοτήτων". Δηλαδή, αν θεωρούμε πως οι εγγραφές Α και Β αντιπροσωπεύουν την ίδια οντότητα, και οι Β και Γ επίσης, τότε μπορούμε να θεωρήσουμε πως και οι τρεις εγγραφές αναπαριστούν την ίδια οντότητα, και άρα πρέπει να συγχωνευτούν. Ο υπολογισμός του μεταβατικού κλεισίματος μιας σχέσης μπορεί να υλοποιηθεί αποδοτικά με μεθόδους γραφοθεωρητικές, και με χρήση της δομής Union-Find.

[32]http://www.esri.com ,ίσχυε την 29/6/2006

Βέβαια, η ιδέα αυτή είναι λίγο αστήρικτη, καθότι το αν δυο εγγραφές αναπαριστούν με μεγάλη πιθανότητα την ίδια οντότητα δεν είναι σχέση κατ'ανάγκην μεταβατική. Συνεπώς, με το να απαιτούμε μεταβατικότητα από την σχέση αυτή, βοηθάμε μια λάθος ταύτιση να διαδοθεί (π.χ. αν η "Αναΐς" ταυτίστηκε ορθά με την "Αναής" και η "Αναής" λανθασμένα με τον "Παναής", θεωρώντας το μεταβατικό κλείσιμο, θα έχουμε (λανθασμένα) ταύτιση της "Αναΐς" με τον "Παναής"[33]). Εντούτοις, σύμφωνα με τη βιβλιογραφία αυτό σπάνια συμβαίνει, και πρακτικά, η τεχνική αυτή βοηθάει στην αύξηση της ανάκλησης, χωρίς μεγάλη απώλεια στην ακρίβεια.

4.2.7.8 Προσεγγιστική μεταβατικότητα της ταύτισης

Εάν, δε, μας απασχολήσει αυτή η αύξηση ψευδώς θετικών απαντήσεων, μπορούμε να χρησιμοποιήσουμε την εξής ιδέα από τα **ασαφή σύνολα**: Κάθε ομάδα εγγραφών που θεωρούμε πως αναπαριστούν την ίδια οντότητα θα την συσχετίζουμε με ένα συντελεστή βεβαιότητας $\in [0,1]$. Κατά την εύρεση του μεταβατικού κλεισίματος, θα υπολογίζουμε πρώτα τον συντελεστή βεβαιότητας της ομάδας που θα προκύψει, συναρτήσει των συντελεστών βεβαιότητας των δύο ομάδων που την γεννούν (π.χ. με την πράξη του γινομένου). Εάν η βεβαιότητα αυτή είναι κάτω από ένα ορισμένο κατώφλι, δεν εκτελούμε την μεταβατική αυτή επέκταση, αποφεύγοντας προβλήματα σαν τα προαναφερθέντα.

4.2.7.9 Προσαρμοστικό μέγεθος παραθύρου

Μια αδυναμία όλων των ανωτέρω μεθόδων είναι το σταθερό μέγεθος παραθύρου με το οποίο σαρώνεται η βάση. Και αυτό γιατί αν μια συστάδα όμοιων εγγραφών έχει περισσότερες εγγραφές από το μέγεθος του παραθύρου, ενδεχομένως να μην ανιχνευθούν όλες ως διπλότυπα, γιατί δεν θα γίνουν αρκετές συγκρίσεις. Ακόμη, αν μια συστάδα έχει μια ή λίγες εγγραφές, τότε γίνονται υπερβολικά πολλοί έλεγχοι. Μια έξυπνη ιδέα για την αντιμετώπιση αυτού βρίσκουμε στο [Mon00]. Εδώ, για να προσδοθεί **προσαρμοστικότητα** ως προς το **μέγεθος** και την **ομοιογένεια** των **συστάδων**, το σταθερού μεγέθους παράθυρο αντικαθίσταται από μια ουρά προτεραιότητας συστάδων διπλοτύπων, σταθερού και μικρού μεγέθους. Στην ουρά αυτή τοποθετούνται σύνολα αντιπροσωπευτικών εγγραφών από τις τελευταίες ανιχνευθείσες συστάδες όμοιων εγγραφών. Κατ'αυτόν τον τρόπο, κάθε νέα εγγραφή που συναντάται κατηγοριοποιείται με βάση γενικευμένους (domain independent) και υπολογιστικά φθηνούς[34] κανόνες σε κάποια συστάδα της ουράς, ή τοποθετείται σε νέα συστάδα εάν κριθεί ότι δεν ανήκει με μεγάλη βεβαιότητα σε κάποια προϋπάρχουσα. Αν και η ασυμπτωτική πολυπλοκότητα του αλγορίθμου δεν είναι καλύτερη από

[33]Κατά το γνωστό ανέκδοτο.

[34]Λέγοντας υπολογιστικά φθηνούς, εννοούμε ότι στην πράξη δεν θα χρειαστεί μια εγγραφή να συγκριθεί με όλες τις εγγραφές στην ουρά, αλλά με τη χρήση κατάλληλων γενικευμένων ευριστικών που προτείνονται, οι συγκρίσεις αυτές μπορούν να περιοριστούν στο ελάχιστο.

αυτήν των άλλων αλγορίθμων, πειραματικά δεδομένα επί ευρείας κατηγορίας δεδομένων καταδεικνύουν την αποδοτική συμπεριφορά του.

4.2.7.10 Εντοπισμός διπλοτύπων βασισμένος στη γνώση

Μια ακόμη διαφορετική προσέγγιση στο ζήτημα, αν και ξεπερασμένη πλέον, είναι η προσέγγιση του IntelliClean, **βασισμένη στη γνώση** του πεδίου των δεδομένων, που περιγράφεται στο [LLL00]. Γνωρίζοντας καλά τις ιδιότητες των δεδομένων (έπειτα από εξόρυξη γνώσης, συζήτηση με ειδικούς κ.α.), ο μηχανικός γνώσης δημιουργεί κανόνες παραγωγής που προδιαγράφουν πότε δύο εγγραφές αναπαριστούν την ίδια οντότητα, καθώς και τον τρόπο με τον οποίο θα ενσωματωθούν σε μία. Με την χρήση του αλγορίθμου του Rete για την αποδοτικότερη υλοποίηση του έμπειρου συστήματος, επιτυγχάνεται ασυμπτωτικά καλός χρόνος απόκρισης, εις βάρος αυξημένης χρήσης μνήμης. Στα αρνητικά αυτής της προσέγγισης συγκαταλέγουμε την ισχυρή της εξάρτηση από το πεδίο εφαρμογής, που την καθιστά δύσκολα μεταφέρσιμη σε παρεμφερή προβλήματα, και απρόσιτη για τον μέσο χρήστη.

4.2.7.11 Αυξητικός εντοπισμός διπλοτύπων

Τέλος, μια ακόμη ιδέα, που σκιαγραφείται στο [HS98] είναι ένας **αυξητικός αλγόριθμος**. Πράγματι, σε όλες τις ανωτέρω περιπτώσεις υποθέταμε ότι ο καθαρισμός δεδομένων από διπλότυπα εκτελείται μια φορά. Στην πράξη, όμως, νέα δεδομένα φθάνουν περιοδικά στην αποθήκη. Η απαλοιφή διπλοτύπων μέσω της εκ νέου εφαρμογής της όποιας μεθόδου επί όλων των δεδομένων είναι σαφώς αντιπαραγωγική. Ο Incremental Merge/Purge αλγόριθμος, λοιπόν, κάνει χρήση "αντιπροσωπευτικών εγγραφών" από κάθε ομάδα εγγραφών που βρέθηκε να αναπαριστούν την ίδια οντότητα. Έτσι, οι νέες εγγραφές που προστίθενται συγκρίνονται μόνο με τις αντιπροσωπευτικές αυτές εγγραφές, γλιτώνοντας το σύστημα από επιπλέον περιττούς υπολογισμούς. Η κυριότερη δυσκολία σε αυτήν την περίπτωση είναι η επιλογή των "αντιπροσωπευτικών πλειάδων", αν και έχουν προταθεί γενικές μέθοδοι γι'αυτό το σκοπό.

4.2.8 Σύνοψη

Σε αυτήν την ενότητα, παρουσιάσαμε το πρόβλημα ταυτότητας αντικειμένου. Περιγράψαμε διάφορες μετρικές ομοιότητας συμβολοσειρών, και αλγορίθμους προσεγγιστικού ταιριάγματος συμβολοσειρών. Είδαμε την ανάγκη για προηγμένους αλγορίθμους προσεγγιστικής αναζήτησης συμβολοσειρών / απαλοιφής διπλοτύπων συμβολοσειρών, και περιγράψαμε τις βασικές προταθείσες ιδέες. Κοινό χαρακτηριστικό των περισσότερων είναι η χρήση κωδίκων ομαδοποίησης πιθανά όμοιων συμβολοσειρών, για την μείωση των απαιτούμενων προσεγγιστικών συγκρίσεων μεταξύ συμβολοσειρών.

4.3 Προσεγγιστικό ταίριαγμα συμβολοσειρών - Η προσέγγισή μας

Εξετάζουμε τώρα την εφαρμογή των ανωτέρω στην περίπτωσή μας.

Το πρώτο πρόβλημα που συναντάμε είναι η προσεγγιστική σύγκριση δύο συμβολοσειρών. Δηλαδή, δοθείσων δύο συμβολοσειρών, να αποφασιστεί αν είναι όμοιες. Βασισμένοι και στις υπάρχουσες προτάσεις, δώσαμε την ακόλουθη λύση:

4.3.1.1 Κανονικοποίηση και τεχνολόγηση

Οι δύο συμβολοσειρές κανονικοποιούνται και τεχνολογούνται (standardised & parsed)[35]. Δηλαδή, (κανονικοποίηση) λέξεις-κλειδιά με υψηλή συχνότητα εμφάνισης ταυτοποιούνται, και μετατρέπονται σε μια κανονική μορφή (π.χ. οι "οδ.", "οδού" κ.λ.π. θα γίνουν όλες "οδός"). Επίσης, δημιουργείται μια εναλλακτική μορφή του ονόματος προς αναζήτηση, η οποία δεν περιέχει τις λέξεις αυτές (π.χ. η "οδός ανθυπολοχαγού Μ. Μπότσαρη" θα αναζητηθεί και ως "Μ.Μπότσαρη", αν και με ελαφρώς μειωμένο συντελεσή βεβαιότητας αποτελεσμάτων). Αυτό το βήμα βελτιώνει κατά πολύ την ποιότητα των αποτελεσμάτων μας, αν αναλογιστούμε ότι η παρουσία ή μη των λέξεων αυτών, δεν προσφέρει σημαντική πληροφορία. Τα σχετικά με την τεχνολόγηση έχουν ήδη αναφερθεί στο κεφ. 3.

4.3.1.2 Μετατροπή σε φωνητικό αλφάβητο

Οι συμβολοσειρές μετατρέπονται σε ένα ενδιάμεσο φωνητικό αλφάβητο που αναπτύξαμε για την ελληνική γλώσσα. Αυτό αποφασίστηκε έπειτα από την παρατήρησή μας ότι τα περισσότερα ορθογραφικά λάθη στα Ελληνικά διατηρούν την προφορά (π.χ. "Πελοπόνησος", "Πελοπόννισος"), ενώ ελάχιστες είναι οι λέξεις με ίδια προφορά και διαφορετική έννοια (ακόμη λιγότερες από αυτές είναι τοπωνύμια). Η μετατροπή στο φωνητικό αλφάβητο βασίζεται στους κανόνες προφοράς των Ελληνικών, όπως και σε μερικές ευριστικές παρατηρήσεις.

Το φωνητικό αλφάβητό μας βασίζεται σε ανάλυση των φωνητικών ιδιοτήτων των φθόγγων της Ελληνικής (π.χ. [Fou03],[SFK98],[Arv99]), και περιέχει 23 φωνητικά σύμβολα. Μεταξύ των φθόγγων που αντιπροσωπεύουν, ορίζουμε αποστάσεις, ανάλογα π.χ. με το αν ανήκουν στην ίδια ή σε όμοια "ομάδα" εκφοράς (π.χ. χειλικά, οδοντικά), ή αν έχουν το ίδιο "μήκος" (μακρά, βραχέα). Για παράδειγμα, όπως είναι και εμφανές ακουστικά, ο φθόγγος "μπ" είναι πιο κοντά στον φθόγγο "γκ" από τον "β".

Στο φωνητικό αλφάβητο συμπεριλαμβάνουμε και τον χαρακτήρα "/", ο οποίος χρησιμοποιείται ευρέως και υποδηλώνει την ύπαρξη τουλάχιστον ενός χαρακτήρα ακόμη (π.χ. "Αγ. Ευστράτιος" → "αγ/ efstratios", "Αιτ/νία"→ "et/nia"). Όπως θα δούμε, η ύπαρξη του χαρακτήρα αυτού αφ'ενός

[35]Αυτη η λειτουργία λαμβάνει χώρα στα πλαίσια του geoparsing.

προσδίδει δύναμη στην προσέγγισή μας (αφού επιτρέπει την ύπαρξη συντμήσεων), αφ'ετέρου δημιουργεί επιπλέον προκλήσεις για τις επόμενες φάσεις.

Σε αυτήν την φάση ανακύπτει και το πρόβλημα των greeklish (αν κάποια από τις δύο συμβολοσειρές είναι γραμμένη σε greeklish) - σχετική συζήτηση γίνεται σε ξεχωριστή ενότητα αργότερα.

Μέχρι τώρα αναφέραμε προεπεξεργαστικά βήματα στο πρόβλημά μας. Ας περάσουμε, τώρα, στο κυρίως κομμάτι του προτεινόμενου αλγορίθμου.

4.3.1.3 Τροποποιημένη απόσταση Levenshtein

Οι δύο μετετραμμένες συμβολοσειρές συγκρίνονται με βάση την απόσταση Levenshtein, που ήδη αναφέρθηκε. Τροποποιήσαμε τον κλασσικό αλγόριθμο υπολογισμού αυτής (βλ. [Gil00], την παραλλαγή που αποδίδεται στον Chas Emerick), ώστε να λαμβάνει υπόψη αφ'ενός τις φωνητικές ομοιότητες δύο φθόγγων (άρα π.χ. η "Μάνου" μοιάζει περισσότερο στην "Νάνου" απ'ότι στην "Πάνου") , αφ'ετέρου τον χαρακτήρα "/" που ήδη αναφέραμε (επομένως η "Αιτ/νία" μοιάζει πολύ με την "Αιτωλοακαρνανία"). Σημειωτέον ότι οι τροποποιήσεις που έγιναν δεν μεταβάλλουν την πολυπλοκότητα του αλγορίθμου, που παραμένει τετραγωνική. Για την ακριβή περιγραφή του αλγορίθμου, ο αναγνώστης παραπέμπεται στον αντίστοιχο πηγαίο κώδικα.

4.3.1.4 Σύγκριση με απόσταση κατωφλίου

Η υπολογισθείσα απόσταση των δύο λέξεων συγκρίνεται με μια απόσταση κατωφλίου, εξαρτώμενη από το μήκος των λέξεων. Η συνάρτηση κατωφλίου υπολογίστηκε με μεθόδους ευριστικές (ανάλυση μεγάλου πλήθους ομόηχων λέξεων, και καθορισμός βέλτιστης απόστασης κατωφλίου, με κριτήριο το F-measure[36]). Εφόσον η απόσταση είναι μικρότερη του κατωφλίου, οι λέξεις θεωρούνται προσεγγιστικά ίσες, και μάλιστα με συντελεστή βεβαιότητας (ποιοτικά) αντιστρόφως ανάλογο της απόστασης μεταξύ τους.

4.3.2 Το πρόβλημα των Greeklish

Έχουμε αφήσει εκκρεμές το ζήτημα των greeklish. Αυτή η μορφή επικοινωνίας, γνωστή στους περισσότερους χρήστες του ελληνικού κυβερνοχώρου, συνίσταται στην γραφή ελληνικών λέξεων, μεταγραμμένων σε λατινικούς χαρακτήρες με αυθαίρετο τρόπο.

[36]Το F-measure είναι ένας σταθμισμένος μέσος precision και recall, βλ. και υποσημείωση 31). Στην περίπτωσή αυτή ως false positive θεωρούμε δύο λέξεις που ταυτοποίηθηκαν ενώ δεν αναπαριστούσαν την ίδια οντότητα, και ως false negative δύο λέξεις που δεν ταυτοποιήθηκαν, ενώ αναπαριστούσαν την ίδια οντότητα.

4.3.2.1 Λόγοι εξάπλωσης

Τα greeklish εξαπλώθηκαν όταν η πληροφορική βρισκόταν ακόμη στα σπάργανά της στην Ελλάδα, και η υποστήριξη των Ελληνικών σε υπολογιστικά συστήματα ήταν περιορισμένη, ή προβληματική. Πέραν από τον μεγάλο όγκο πληροφορίας που υφίσταται ήδη σε greeklish, η χρήση τους συνεχίζεται αθρόα μέχρι τις μέρες μας, λόγω της υψηλής πιθανότητας οι διάφοροι συμμετέχοντες σε μια πράξη επικοινωνίας (είτε μέσω ηλεκτρονικού ταχυδρομείου, είτε μέσω του παγκόσμιου ιστού, ή άλλου ηλεκτρονικού μέσου), να μην έχουν μεταξύ τους συμβατές γλωσσικές ρυθμίσεις (π.χ. κωδικοσελίδας). Αν και έχουν κατά καιρούς προταθεί πρότυπα για την κανονικοποίηση των greeklish (π.χ. από τον ΕΛΟΤ), ουδέποτε έχουν γνωρίσει ευρεία αποδοχή.

4.3.2.2 Τύποι Greeklish

Ο κάθε χρήστης γράφει με τρόπο προσωπικό, χρησιμοποιώντας, ανάμεικτα, φωνητική (π.χ. "αθηναϊκός"→"athinaikos") μεταγραφή, με γνώμονα την αναπαράσταση φθόγγων και συνεπαγόμενη απλοποίηση της ιστορικής ορθογραφίας και ορθογραφική (π.χ. "Εύηνος"→"Euhnos") μεταγραφή, με γνώμονα την αναπαράσταση των ελληνικών φθογγοσήμων, όσο αυτό είναι δυνατό με λατινικούς χαρακτήρες ([And99]). Όπως είναι φυσικό, η κατάσταση αυτή δημιουργεί προβλήματα σε μια εφαρμογή που χειρίζεται κείμενα από τον ιστό, που είναι συχνά γραμμένα σε greeklish.

4.3.2.3 Λύσεις

Για την αντιμετώπιση του προβλήματος έχουν προταθεί διάφορες λύσεις. Στις πιο απλές, υποτίθεται ένας και μοναδικός τρόπος μεταγραφής, με αποτέλεσμα ένα συντριπτικό ποσοστό των λέξεων να μην μεταφέρεται σωστά στην Ελληνική. Με μικρή αύξηση της πολυπλοκότητας, βρίσκουμε άλλες λύσεις που προσπαθούν να "μαντέψουν" με ποιον τρόπο έχει γίνει η μεταγραφή σε greeklish, και να μεταφέρουν έπειτα το κείμενο στα Ελληνικά. Η εύρεση του τρόπου μεταγραφής γίνεται με την ανίχνευση ορισμένων χαρακτηριστικών ν-γραμμάτων, και ανάλυση συχνοτήτων εμφάνισής τους (π.χ. αν εμφανίζεται πολλές φορές το δίγραμμα "th", μπορούμε να υποθέσουμε ότι θα αντιπροσωπεύει το γράμμα "θ"). Μια λύση ανάμεσα στις δύο αυτές, υιοθετήσαμε σε μερικές περιπτώσεις στην εργασία αυτή (βλ. παρακάτω).

4.3.2.4 Η προσέγγιση του Ινστιτούτου Επεξεργασίας του Λόγου

Μια μέθοδος που αξίζει της προσοχής μας, τόσο λόγω της κομψότητάς της (αλλά και αυξημένης πολυπλοκότητάς της), όσο και λόγω της ποιότητας των αποτελεσμάτων της, είναι αυτή που

προτείνεται στο [CTR+04], και έχει υλοποιηθεί στο προϊόν All Greek To Me![37] Του Ινστιτούτου Επεξεργασίας του Λόγου. Η προσέγγιση αυτή συνίσταται στα εξής:

1. Κάθε λέξη στα greeklish, μεταγράφεται σε όλες τις δυνατές φωνητικές της αναπαραστάσεις, λαμβάνοντας υπόψη όλους τους διαφορετικούς τύπους greeklish, καθώς και όλους τους πιθανούς συνδυασμούς τους.
2. Εξ'αυτών, απορρίπτεται ένα μεγάλο ποσοστό, με βάση την συμφωνία κάθε λέξης ή όχι με ένα ακουστικό μοντέλο, σχεδιασμένο για την ελληνική γλώσσα.
3. Εξ'αυτών, αναζητούνται οι πιθανότερες λύσεις σε ένα λεξικό με ειδικές δομές δεδομένων, που έχει προκύψει από την αυτόματη εξέταση πληθώρας Ελληνικών κειμένων.
4. Εξ'αυτών, επιλέγεται η καλύτερη λύση, με βάση υπό συνθήκη πιθανότητες, καθώς και κανόνες εξαρτώμενους από το περιέχον κείμενο.
5. Παράλληλα, εκτελείται ένας αλγόριθμος ανίχνευσης γλώσσας, ώστε να μην επιχειρηθεί η μεταγραφή μιας Αγγλικής λέξης στα Ελληνικά. (δηλ. να διαχωριστούν οι λέξεις σε greeklish από αυτές που είναι όντως Αγγλικές). Ο αλγόριθμος βασίζεται σε πιθανοτικούς κανόνες και στατιστικά μοντέλα σχεδιασμένα για την Ελληνική γλώσσα.

4.3.2.5 Η δική μας προσέγγιση

Εξ'αιτίας της ποιότητας των αποτελεσμάτων αυτής της μεθόδου, στην εργασία χρησιμοποιήσαμε, όπου ήταν δυνατόν, το εργαλείο All Greek To Me!. Για τις περιπτώσεις όπου απαιτείτο μεγάλη ταχύτητα, ή που το εργαλείο αποτύγχανε, δημιουργήσαμε μια εναλλακτική συνάρτηση ταχείας μετατροπής από greeklish απευθείας στο φωνητικό μας αλφάβητο, βασισμένη στις ιδέες που ήδη αναφέραμε (βλ. παραπάνω).

4.4 Προσεγγιστική αναζήτηση συμβολοσειρών -Η προσέγγισή μας

4.4.1 Ορισμός του προβλήματος

Πλέον, έχοντας ορίσει τον αλγόριθμο για τη σύγκριση δυο λέξεων, μπορούμε να εξετάσουμε τον κυρίως στόχο μας. Ο οποίος δεν είναι άλλος από την προσεγγιστική αναζήτηση μιας συμβολοσειράς (π.χ. ενός τοπωνυμίου) από έναν κατάλογο, για την ανάκτηση κάποιων στοιχείων του (π.χ. της θέσης του, αλλά, γενικότερα, ενός κωδικού αριθμού (ταυτότητα αντικειμένου- object id) που να το χαρακτηρίζει μοναδικά (με τον κωδικό αυτό μπορούμε μετά να αναζητήσουμε εύκολα και αποδοτικά περαιτέρω πληροφορίες).

[37]http://www.ilsp.gr/greeklish.html , ίσχυε την 29/6/2006

4.4.2 Γιατί προσεγγιστική αναζήτηση

Ας υπενθυμίσουμε ότι η απευθείας σύγκριση της αναζητούμενης συμβολοσειράς με όλες τις υπόλοιπες είναι υπολογιστικά ανέφικτη, και γι'αυτό απαιτείται κάποιος γρήγορος τρόπος "φιλτραρίσματος" των υποψηφίων ταιριαγμάτων έτσι ώστε αφ'ενός να είναι απίθανο το να απορριφθεί εσφαλμένα κάποια συμβολοσειρά, αφ'ετέρου να απορριφθούν (ορθά) όσο το δυνατόν περισσότερες συμβολοσειρές. Αν εξαιρέσουμε κάποιες "εξωτικές" λύσεις ([GIJ+01]), η συνήθης πρακτική είναι η ακριβής αναζήτηση των συμβολοσειρών που έχουν τον ίδιο κώδικα-κλειδί με την αναζητούμενη, και έπειτα η προσεγγιστική της σύγκριση με αυτές.

4.4.3 Ελληνικός φωνητικός κώδικας

Εμείς ορίσαμε έναν κώδικα-εξέλιξη του Metaphone, με βάση το φωνητικό αλφάβητο που φτιάξαμε. Κάθε ομάδα όμοιων φθόγγων αντιστοιχίζεται στο ίδιο σύμβολο, τα τελικά 'ς' απορίπτονται, όπως και τα φωνήεντα που δεν είναι στην αρχή της λέξης. Ακόμη, ο χαρακτήρας "/" διατηρείται: έτσι είναι δυνατόν μια λέξη να ταιριάξει με την σύντμησή της, αλλά έχουμε ένα επιπρόσθετο κόστος κατά την σύγκριση κωδίκων: δεν μπορούμε πλέον να ελέγχουμε για ισότητα κωδίκων, αλλά για συμβατότητα (π.χ. οι κώδικες 'edlgm' (της λέξης "αιτωλοακαρνανίας") και 'ed/m' (της λέξης "αιτ/νίας") δεν είναι ίσοι, αλλά είναι συμβατοί μεταξύ τους).

4.4.4 Αλγόριθμοι αναζήτησης

Έπειτα, κατασκευάσαμε έναν αλγόριθμο αναζήτησης σε τρεις παραλλαγές, ανάλογα με το μέγεθος του καταλόγου αναζήτησης, του τύπου των δεδομένων που αναζητούνται, και της ταχύτητας που επιδιώκεται. Το **ευρετήριο κατακερματισμού μεμονωμένων λέξεων** βρίσκει (προσεγγιστικά) μεμονωμένες λέξεις σε έναν κατάλογο. Το **ευρετήριο κατακερματισμού πολλαπλών λέξεων** βρίσκει φράσεις, ή προθέματα φράσεων, σε έναν κατάλογο. Οι δύο δομές αυτές χρησιμοποιούν μόνο την κύρια μνήμη του υπολογιστή, σε αντίθεση με το **φωνητικό ευρετήριο**, που προσφέρει την δυνατότητα προσεγγιστικής αναζήτησης σε κατάλληλα διαμορφωμένο πίνακα μιας βάσης δεδομένων.

Ακολούθως, περιγράφουμε τις τρεις αυτές παραλλαγές.

4.4.5 Ευρετήριο κατακερματισμού μεμονωμένων λέξεων

Η παραλλαγή αυτή χρησιμοποιείται για την αναζήτηση εντός μεσαίου μεγέθους συνόλου δεδομένων (διότι το τελευταίο πρέπει να χωράει ολόκληρο στην μνήμη).

4.4.5.1 Δομή

Βασίζεται σε έναν πίνακα κατακερματισμού, που αντιστοιχίζει κώδικες σε σύνολα λέξεων σε φωνητικό αλφάβητο (π.χ. 'edlgm' → {"etoloakarnanias", "edrikania"}). Περιέχει, επίσης, δυο πίνακες κατακερματισμού που αντιστοιχίζουν κώδικες που περιέχουν τον χαρακτήρα "/" σε συμβατούς κώδικες που δεν τον περιέχουν, και αντιστρόφως. Ακόμη, έναν τέταρτο πίνακα κατακερματισμού, ο οποίος αντιστοιχίζει λέξεις σε μοναδικούς κωδικούς - ταυτότητες αντικειμένου.

4.4.5.2 Αναζήτηση λέξης

Κατά την αναζήτηση μιας λέξης, ανασύρουμε από τους δεύτερους πίνακες κατακερματισμού τους κώδικες που είναι συμβατοί με τον κώδικα της αναζητούμενης λέξης, κι έπειτα, από τον πρώτο πίνακα, τις λέξεις που αντιστοιχούν σε αυτούς τους κώδικες. Κατόπιν, συγκρίνονται προσεγγιστικά οι λέξεις αυτές με την αναζητούμενη λέξη. Τα αποτελέσματα της σύγκρισης χρησιμοποιούνται για την ανάκτηση από τον τέταρτο πίνακα κατακερματισμού των ζητούμενων ταυτοτήτων.

4.4.5.3 Εισαγωγή λέξης

Κατά την εισαγωγή μιας λέξης στο ευρετήριο, καταχωρείται η αντιστοίχιση του κώδικά της στην ίδια στον πρώτο πίνακα, καθώς και η αντιστοίχιση του κώδικα αυτής με όλους τους κώδικες που είναι συμβατοί με αυτόν, εάν υπάρχουν ήδη στον δεύτερο ή τρίτο πίνακα. Ακόμη, εισάγεται στον τέταρτο πίνακα η αντιστοίχιση της λέξης με την αντίστοιχη πληροφορία (π.χ. ταυτότητα αντικειμένου).

4.4.5.4 Αξιολόγηση

Εάν η αναζητούμενη συμβολοσειρά αποτελείται από πολλές λέξεις, πρέπει να αναζητήσουμε την κάθε λέξη χωριστά, και έπειτα να συνυπολογίσουμε τα επιμέρους αποτελέσματα. Επειδή η διαδικασία αυτή είναι δαπανηρή, η παραλλαγή του ευρετηρίου κατακερματισμού μεμονωμένων λέξεων χρησιμοποιείται σε περιπτώσεις που δεν ενδιαφέρει η ακριβής σειρά των λέξεων, ούτε η παρουσία όλων των λέξεων μιας φράσης. Ως αντιστάθμιση, η μέθοδος είναι πολύ γρήγορη: Κάνοντας παραδοχές που ισχύουν στις πρακτικές περιπτώσεις (π.χ. περιορισμένο μέγιστο μήκος λέξης, και άρα μικρό και περιορισμένο μήκος κώδικα, ομοιόμορφη κατανομή των κωδίκων των λέξεων, περιορισμένο σχετικό ποσοστό λέξεων με τον χαρακτήρα "/", κ.ο.κ.), η αναζήτηση μιας λέξης έχει σταθερό χρονικό κόστος, ανεξάρτητο του μεγέθους του ευρετηρίου. Η μέθοδος αυτή βρίσκει ιδανική εφαρμογή κατά την διαδικασία του geoparsing, στην εύρεση **πιθανών** τοπωνυμίων.

4.4.6 Ευρετήριο κατακερματισμού πολλαπλών λέξεων

Μια αδυναμία της προηγούμενης μεθόδου, είναι το ότι δεν υποστηρίζει άμεσα την αναζήτηση συμβολοσειρών που περιέχουν περισσότερες από μια λέξεις. Γι'αυτό το λόγο, αναπτύξαμε την παραλλαγή της δομής του ευρετηρίου πολλαπλών λέξεων, που επεκτείνει το ευρετήριο μεμονωμένων λέξεων.

4.4.6.1 Περιγραφή

Ουσιαστικά, πρόκειται για ένα ευρετήριο κατακερματισμού μεμονωμένων λέξεων πολλαπλών βαθμίδων. Δηλαδή, η πρώτη αναζήτηση στο ευρετήριο μας επιστρέφει ένα σύνολο που περιέχει:

- Τις πληροφορίες (ταυτότητες αντικειμένου) που αντιστοιχούν στις συμβολοσειρές που αποτελούνται από ακριβώς μία λέξη, και μοιάζουν με την αναζητούμενη συμβολοσειρά, αν αυτή αποτελείται από μια και μόνο λέξη, ή,
- Τα ευρετήρια της επόμενης βαθμίδας που περιέχουν συμβολοσειρές που αποτελούνται από παραπάνω από μια λέξη, και η πρώτη λέξη τους μοιάζει με την πρώτη λέξη της αναζητούμενης συμβολοσειράς, αν αυτή αποτελείται από παραπάνω από μια λέξη.

Η διαδικασία αυτή συνεχίζεται αναδρομικά, μέχρις ότου εξαντληθούν οι λέξεις της αναζητούμενης συμβολοσειράς, ή οι φράσεις του ευρετηρίου. Σε κάθε περίπτωση, επειδή τα περισσότερα τοπωνύμια περιέχουν λίγες λέξεις, ενώ ελάχιστα έχουν πάνω από 4, οι απαιτούμενοι πόροι σε χρόνο και χώρο είναι λογικοί και περιορισμένοι, ακόμη και για μεγάλο όγκο ευρετηριοποιημένης πληροφορίας.

Για τον διασαφηνισμό της όλης διαδικασίας, ας δούμε το ακόλουθο παράδειγμα:

4.4.6.2 Παράδειγμα: Εισαγωγή στοιχείου

Έστω ότι θέλουμε να εισαγάγουμε στο ευρετήριο την φράση "leof/ panaγi tsalδari", αντιστοιχίζοντάς την στον μοναδικό κωδικό 1873. Εάν η λέξη "leof/" είναι ήδη καταχωρημένη στο πρώτο ευρετήριο, ως δεύτερο ευρετήριο θεωρούμε το ευρετήριο που της έχει αντιστοιχηθεί, εάν όχι δημιουργούμε ένα νέο ευρετήριο, και καταχωρούμε στο πρώτο ευρετήριο, κατά τα γνωστά, την λέξη "leof/", με αντίστοιχη πληροφορία το νέο ευρετήριο που δημιουργήσαμε. Συνεχίζουμε όμοια: Εάν στο δεύτερο ευρετήριο είναι καταχωρημένη η λέξη "panaγi", ως τρίτο ευρετήριο θεωρούμε αυτό που της αντιστοιχεί, ει δε μη το δημιουργούμε και το καταχωρούμε. Όμοια πράττουμε για την λέξη "tsalδari", μόνο που εδώ ως πληροφορία αντιστοιχίζουμε τον κωδικό 1873.

4.4.6.3 Παράδειγμα: Αναζήτηση στοιχείου

Έστω ότι θέλουμε να αναζητήσουμε στο ευρετήριο την φράση "leoforos pan/ tsalδali". Το πρώτο ευρετήριο μας επιστρέφει ένα δεύτερο ευρετήριο, που αντιστοιχεί σε φράσεις που ξεκινούν λέξεις που μοιάζουν με την "leoforos". Όμοια, το δεύτερο ευρετήριο μας επιστρέφει ένα τρίτο, για το πρόθεμα "leoforos pan/". Τέλος, η αναζήτηση της λέξης "tsalδali" στο τρίτο ευρετήριο θα μας επιστρέψει τον κωδικό 1873 που είχαμε εισαγάγει νωρίτερα. Το ταίριαγμα είναι αποδεκτό, διότι η διαφορά των δύο συμβολοσειρών, της αναζητούμενης "leoforos pan/ tsalδali" και της καταχωρημένης "leof/ panaγi tsalδari", πέραν των συντμήσεων, είναι η τροποποίηση ενός χαρακτήρα, με έναν αρκετά όμοιό του.

4.4.6.4 Αναζήτηση προθέματος

Προσφέρεται, επίσης, η δυνατότητα αναζήτησης φράσεων-συμβολοσειρών που μοιάζουν με ένα πρόθεμα της αναζητούμενης φράσης (αποτελούμενο από ολόκληρες λέξεις μόνο, π.χ. για την "Λεωφ. Παναγιώτη Τσαλδάρη" θα βρεθεί η "Λεωφ. Παναγιώτη", αλλά όχι η "Λεωφ. Πάνα"). Στην περίπτωση αυτή, επιστρέφεται το αποτέλεσμα που αντιστοιχεί στην φράση με το καλύτερο ταίριαγμα, καθώς και ο αριθμός των λέξεων που χρησιμοποιήθηκαν από την αναζητούμενη φράση.

Η δυνατότητα αυτή καθιστά το ευρετήριο αυτό χρήσιμο στην διαδικασία του geoparsing (αντί του ευρετηρίου μεμονωμένων λέξεων), αφού εκεί δεν είναι γνωστό εκ των προτέρων ποιες λέξεις αποτελούν την αναζητούμενη συμβολοσειρά (π.χ. από το κείμενο "η ομάδα του Βόλου Μαγνησίας 'Ατρόμητος'", θα πρέπει να αναζητηθεί μονάχα η φράση "Βόλου Μαγνησίας", δίχως τον "Ατρόμητο").

Ας σημειωθεί πάντως, ότι στην πράξη, στο geoparsing, χρησιμοποιείται το προηγούμενο ευρετήριο. Κι αυτό, διότι, σύμφωνα με εμπειρικές μας διαπιστώσεις, η σειρά των λέξεων στα τοπωνύμια δεν έχει τόσο μεγάλη σημασία (π.χ. είναι προφανές ότι η "Λεωφόρος Παναγή Τσαλδάρη" είναι η ίδια με την "Τσαλδάρη Παν., Λεωφ."). Την παρατήρηση αυτή ενσωματώσαμε ως δυνατότητα στην δομή του φωνητικού ευρετηρίου, της οποίας περιγραφή ακολουθεί.

4.4.7 Φωνητικό ευρετήριο

Οι προηγούμενες δύο προσεγγίσεις απαιτούν όλα τα δεδομένα που ευρετηριοποιούνται να βρίσκονται στην κύρια μνήμη. Συνεπώς, είναι ακατάλληλες για μεγάλου μεγέθους σύνολα δεδομένων. Η προσέγγιση του φωνητικού ευρετηρίου αναπτύχθηκε για την μεταφορά των σχετικών ιδεών σε μια ταχεία υλοποίηση μέσα στα πλαίσια ενός Συστήματος Διαχείρησης Βάσεων Δεδομένων, ορμώμενοι και από τις ιδέες των [GIJ+01] και [FFM05].

4.4.7.1 Προϋποθέσεις χρήσης

Στην προσέγγιση αυτή, υποθέτουμε ότι έχουμε έναν πίνακα σε μια βάση δεδομένων, τον οποίο θέλουμε να ευρετηριοποιήσουμε (index), πάνω σε μια στήλη που περιέχει μια περιγραφική συμβολοσειρά (π.χ. τοπωνύμιο). Κάθε γραμμή του πίνακα έχει ένα ατομικό πρωτεύον κλειδί, που αποτελεί την πληροφορία που θέλουμε να επιστρέψει η αναζήτηση (π.χ. κωδικός ταυτότητας αντικειμένου).

4.4.7.2 Δημιουργία ευρετηρίου

Κατά την δημιουργία του ευρετηρίου (που χρειάζεται να γίνει μια φορά για κάθε πίνακα που ευρετηριοποιούμε), δημιουργείται ένας ξεχωριστός πίνακας-ευρετήριο στη βάση. Αυτός περιέχει τις εξής στήλες: Ταυτότητα αντικειμένου, Αριθμός λέξης, Φωνητικός κώδικας, Λέξη, Μήκος λέξης, Μήκος φράσης, Είναι σύντμηση. (Ο σκοπός της κάθε στήλης θα φανεί κατά την περιγραφή των λειτουργιών του φωνητικού ευρετηρίου.) Ο πίνακας-ευρετήριο αρχικοποιείται με πληροφορία από τον αρχικό πίνακα ως εξής: Για κάθε πλειάδα του αρχικού πίνακα, η συμβολοσειρά που θα αναζητηθεί χωρίζεται σε λέξεις. Έπειτα, δημιουργούνται τόσες εγγραφές στον πίνακα-ευρετήριο, όσες οι λέξεις που προέκυψαν, ως εξής:

- Ταυτότητα αντικειμένου: ο κωδικός ταυτότητας αντικειμένου της πλειάδας του αρχικού πίνακα
- Αριθμός λέξης: Με ποια σειρά εμφανίζεται η λέξη μέσα στην αρχική συμβολοσειρά, π.χ. 1,2,3...
- Λέξη , Φωνητικός κώδικας: Η ίδια η λέξη (προφανώς σε φωνητικό αλφάβητο) και ο φωνητικός της κώδικας.
- Μήκος λέξης, Μήκος φράσης: Το μήκος της λέξης, και ολόκληρης της φράσης-συμβολοσειράς, σε χαρακτήρες, αντίστοιχα
- Είναι σύντμηση: Η λέξη περιέχει τον χαρακτήρα "/" ;

Ο πίνακας-ευρετήριο έχει ο ίδιος ευρετήριο Β-δένδρου πάνω στη συμπαράθεση των πεδίων Είναι σύντμηση, Φωνητικός κώδικας, Ταυτότητα αντικειμένου, και Αριθμός λέξης.

4.4.7.3 Αναζήτηση συμβολοσειράς

Κατά την αναζήτηση μιας συμβολοσειράς στο ευρετήριο, ακολουθείται, σε αδρές γραμμές, η εξής διαδικασία:

- Δημιουργείται ένας προσωρινός πίνακας στην βάση, στον οποίο το ΣΔΒΔ εισαγάγει, κατόπιν κατάλληλου ερωτήματος, τις εγγραφές του πίνακα-ευρετηρίου που έχουν φωνητικό κώδικα συμβατό με τον φωνητικό κώδικα κάποιας λέξης της αναζητούμενης φράσης. Η

συμβατότητα των φωνητικών κωδίκων ελέγχεται μέσα στο ΣΔΒΔ, με τρόπο αποδοτικό, που εκμεταλλεύεται στο έπακρο το ευρετήριο B-δένδρου του πίνακα-ευρετηρίου.

- Από τον πίνακα αυτό, απορρίπτονται όσες ομάδες εγγραφών (που αντιστοιχούν στο ίδιο αντικείμενο-φράση) διαφέρουν σημαντικά από την αναζητούμενη φράση ως προς τον αριθμό και το μήκος των λέξεων που πρέπει να προστεθούν ή να αφαιρεθούν προκειμένου να ταιριάξουν. Για τον σκοπό αυτό χρησιμοποιούνται τα πεδία Μήκος λέξης, Μήκος φράσης του πίνακα-ευρετηρίου. (π.χ. αναζητώντας την φράση "Λεωφόρος Ελευθερίου Βενιζέλου", στο πρώτο βήμα θα ταιριάξει και η "Πλατεία Βενιζέλου Σοφοκλή", επειδή μοιράζονται την λέξη "Βενιζέλου". Όμως, για να ταιριάξουν οι δύο λέξεις πρέπει στην δεύτερη να αφαιρεθούν 2 λέξεις ("Πλατεία","Σοφοκλή"), και να προστεθούν 2 λέξεις ("Λεωφόρος","Βενιζέλου"), οπότε στην φάση αυτή θα απορριφθεί η "Πλατεία Βενιζέλου Σοφοκλή".)
- Από το ΣΔΒΔ επιστρέφονται στο πρόγραμμα λίγες εγγραφές, που αντιστοιχούν σε φράσεις-συμβολοσειρές που μοιάζουν με την αναζητούμενη σε επίπεδο φωνητικών κωδίκων. Το φιλτράρισμα αυτό έχει γίνει με μεγάλη ταχύτητα, καθότι το έχει εκτελέσει το ίδιο το ΣΔΒΔ, βελτιστοποιώντας το, και χρησιμοποιώντας ευρετήρια. Σημειωτέον ότι στο μέχρι τώρα ταίριαγμα, δεν λαμβάνεται υπόψη η σειρά των λέξεων στη φράση.
- Οι εγγραφές αυτές συγκρίνονται (εκτός ΣΔΒΔ), λέξη προς λέξη, σύμφωνα με την τροποποιημένη απόσταση Levenshtein που φτιάξαμε. Εάν κατά την εκτέλεση της σύγκρισης διαπιστωθεί ότι κάποια φράση διαφέρει υπερβολικά από την αναζητούμενη, αυτή απορρίπτεται. Αφού συγκριθούν όλες οι λέξεις μιας φράσης, λαμβάνουμε υπόψη (κατάλληλα σταθμισμένο) και τον παράγοντα της σειράς των λέξεων στη φράση.
- Οι κωδικοί ταυτότητας αντικειμένου των φράσεων με αποδεκτό τελικό κόστος-διαφορά από την αναζητούμενη φράση επιστρέφονται στον χρήστη, ταξινομημένη κατ'αύξουσα σειρά κόστους.

4.4.7.4 Αξιολόγηση

Η δομή του φωνητικού ευρετηρίου επιτυγχάνει, πρακτικά, ταχύτατη προσεγγιστική φωνητική αναζήτηση. Αν και η πολυπλοκότητα της αναζήτησης την κάνει υπολογιστικά δυσχερή στην χειρότερη περίπτωση, θεωρώντας πάλι τους περιορισμούς που ισχύουν σε πλείστες πρακτικές περιπτώσεις (της υπό μελέτη συμπεριλαμβανομένης), βλέπουμε ότι η αναζήτηση απαιτεί χρόνο σταθερό και μικρό.

Τα χαρακτηριστικά του φωνητικού ευρετηρίου προσιδιάζουν σε εφαρμογές καθαρισμού δεδομένων και γεωκωδικοποίησης, στις οποίες άλλωστε χρησιμοποιήθηκε στα πλαίσια της εργασίας. Είναι όμως ικανοποιητικό και για την εκτέλεση γενικότερων λειτουργιών που απαιτούν

προσεγγιστικό φωνητικό ταίριαγμα μεταξύ συμβολοσειρών (π.χ. προσεγγιστικός σύνδεσμος (join) συμβολοσειρών).

Ενδεικτικά, ας αναφέρουμε τα αποτελέσματα ενός σχετικού πειράματος[38]. Έχοντας ως αφετηρία έναν κατάλογο 1800 διευθύνσεων στην Αττική, με την μορφή που είχαν δηλωθεί από τους πελάτες ενός καταστήματος (δηλαδή με ποικίλα λάθη: ορθογραφικά, ιδιωματισμούς, διαφορετικούς τρόπους περιγραφής του ίδιου τοπωνυμίου κ.λ.π.), επιχειρήσαμε την αναζήτηση των ονομάτων των οδών και περιοχών σε έναν πίνακα με όλες τις οδούς και περιοχές της Αττικής. Η αναζήτηση κράτησε λιγότερο από 2 λεπτά, εκτελούμενη σε υπολογιστή παλαιάς τεχνολογίας και δίχως βελτιστοποιήσεις στις παραμέτρους του ΣΔΒΔ. Συγκριτικά, ας αναφέρουμε ότι η ίδια αναζήτηση, εκτελούμενη από το διαδεδομένο πακέτο γεωκωδικοποίησης της ESRI (το οποίο, βέβαια, δεν είναι προσαρμοσμένο για τα Ελληνικά), διήρκεσε περίπου μια ώρα (σε πιο σύγχρονο υπολογιστή).

4.4.8 Σύνοψη

Έχοντας περιγράψει τις τρεις παραλλαγές του αλγορίθμου αναζήτησης, ας συνοψίσουμε αναφέροντας τα κύρια χαρακτηριστικά του καθενός υπό μορφή πίνακα.

Χαρακτηριστικό	***Ευρετήριο μεμονωμένων λέξεων***	***Ευρετήριο πολλαπλών λέξεων***	***Φωνητικό ευρετήριο***
Πλήθος δεδομένων προς ευρετηριοποίηση	Μεσαίο (περιορισμένο από τη μνήμη)	Μεσαίο (περιορισμένο από τη μνήμη)	Απεριόριστο
Ταχύτητα αναζήτησης	Εξαιρετικά γρήγορη	Πολύ γρήγορη	Γρήγορη
Εξάρτηση από σειρά λέξεων στη φράση	Καμμία	Απόλυτη	Ρυθμιζόμενη
Απαιτήσεις σε μνήμη	Μεγάλες (= πλήθος δεδομένων προς ευρετηριοποίηση)	Μεγάλες (= πλήθος δεδομένων προς ευρετηριοποίηση)	Μικρές
Τυπικές εφαρμογές	Εντοπισμός πιθανών ταιριαγμάτων σε κείμενο	Εντοπισμός ταιριαγμάτων σε κείμενο	Αναζήτηση εγγραφών

Πίνακας 2: Αλγόριθμοι προσεγγιστικής αναζήτησης

[38]Το πείραμα ήταν ποιοτικό, με την έννοια ότι δεν επιδιώξαμε ακριβείς μετρήσεις ή συγκρίσεις με άλλα διαδεδομένα πακέτα, αλλά θέλαμε να δώσουμε μια τάξη μεγέθους για τις επιδόσεις της αναζήτησης σε φωνητικό ευρετήριο.

4.5 Σύνοψη κεφαλαίου

Στο κεφάλαιο αυτό περιγράψαμε τρία - στενά σχετιζόμενα - προβλήματα.

Το πρόβλημα ταυτότητας αντικειμένου, μεταξύ άλλων, μας εισήγαγε στις μετρικές ομοιότητας συμβολοσειρών, και στην χρήση φωνητικών κωδίκων για την ταχεία προσεγγιστική αναζήτηση.

Στο προσεγγιστικό ταίριαγμα συμβολοσειρών παρουσιάσαμε την δική μας προσέγγιση, που περιλαμβάνει κανονικοποίηση και τεχνολόγηση των συμβολοσειρών, μεταγραφή τους στο φωνητικό αλφάβητο, και σύγκριση με την τροποποιημένη απόσταση Levenshtein. Με την ευκαιρία, εξετάσαμε το πρόβλημα των greeklish, και είδαμε ορισμένες λύσεις του.

Τέλος, ως προς την προσεγγιστική αναζήτηση συμβολοσειρών, παρουσιάσαμε και συγκρίναμε τις τρεις προσεγγίσεις μας : Ευρετήριο μεμονωμένων λέξεων, Ευρετήριο πολλαπλών λέξεων, και Φωνητικό ευρετήριο.

5 Υλοποίηση

5.1 Πλατφόρμες και προγραμματιστικά εργαλεία

Ας δούμε, τώρα, μερικές λεπτομέρειες σχετικά με το υλισμικό και λογισμικό που χρησιμοποιήθηκε.

5.1.1 Λογισμικό

Η εργασία αυτή αναπτύχθηκε στην γλώσσα προγραμματισμού Java[39], έκδοση 1.5 -για την εκτέλεση του συστήματος απαιτείται η ύπαρξη εγκατεστημένης Java έκδοσης 1.5 ή νεώτερης - στο περιβάλλον Eclipse, έκδοση 3.1[40]. Για την υλοποίησή της έγινε χρήση αρκετά μεγάλου αριθμού πακέτων. Ενδεικτικά, αναφέρουμε τα ακόλουθα: Σε μια πρώτη πρωτοτυποποίηση χρησιμοποιήθηκε το περιβάλλον επεξεργασίας φυσικής γλώσσας GATE ([CMB+02], [CMB+06],[41]), το οποίο αργότερα αντικαταστάθηκε από κώδικα που έκανε χρήση του γεννήτορα τεχνολογητών κανονικών γραμματικών JFlex[42]. Τα γεωδεδομένα αποθηκεύθηκαν σε βάση δεδομένων που διαχειρίζεται το ΣΔΒΔ MySQL[43], έκδοση 5 (απαιτείται η εγκατάστασή του, πριν την χρήση του συστήματος). Για

[39]http://java.sun.com , ίσχυε την 5/7/2006

[40]http://www.eclipse.org , ίσχυε την 5/7/2006

[41]http://www.gate.ac.uk , ίσχυε την 5/7/2006

[42]http://jflex.de , ίσχυε την 5/7/2006

[43]http://www.mysql.com , ίσχυε την 5/7/2006

τις μετατροπές από greeklish σε ελληνικά, χρησιμοποιήθηκε και η εφαρμογή All Greek To Me![44] του Ινστιτούτου Επεξεργασίας του Λόγου. Για την σύνδεση αυτής με το υπόλοιπο σύστημα, μιας και δεν προσφέρεται κάποιο δημόσιο API, έγινε χρήση της εφαρμογής AutoHotKey[45]. Όπως αναφέρθηκε ήδη, στα πλαίσια της εργασίας, επεκτάθηκαν τα πακέτα JChardet [16] και Tagsoup [13]. Χρησιμοποιήθηκαν ελάχιστα τμήματα κώδικα από τη συλλογή Jakarta Commons[46]. Οι ιστοσελίδες για τα πειράματα αποκτήθηκαν μέσω του εργαλείου ανοιχτού κώδικα Wget[47]. Πολλά αρχεία γεωδεδομένων προσπελάστηκαν με τις βιβλιοθήκες JavaDBF[48]. Για ανάγνωση γεωδεδομένων έγινε επίσης χρήση της εφαρμογής Mapdekode [49], και των βιβλιοθηκών OpenMap[50]. Για την μετατροπή συντεταγμένων μεταξύ datum, έγινε χρήση του προγράμματος CoordGR[51]. Τέλος, για την γεωκωδικοποίηση διευθύνσεων IP χρησιμοποιήθηκε η βιβλιοθήκη GeoIP[52].

5.1.2 Υλικό

Ως προς τις απαιτήσεις σε υλικό του συστήματος, αυτές είναι αρκετά μικρές. Σε έναν φορητό υπολογιστή παλαιάς τεχνολογίας (Celeron M 1.5 Ghz, 256 MB RAM), στον οποίο έγιναν και οι περισσότερες δοκιμές, η εφαρμογή μας έτρεχε με ικανοποιητική ταχύτητα - ο μέσος χρόνος επεξεργασίας μιας σελίδας κυμαινόταν περί τα 4.5" (βλ. σχετικά κεφάλαια για τις επιδόσεις των επιμέρους τμημάτων).

5.2 Λεπτομέρειες υλοποίησης

Ακολούθως, περιγράφουμε τα βασικά πακέτα και κλάσεις που απαρτίζουν το σύστημα. Η περιγραφή αυτή έχει στόχο να δείξει την αντιστοίχιση μεταξύ των αλγορίθμων που περιγράφηκαν, της αρχιτεκτονικής του συστήματος, κ.ο.κ, με τις δημιουργηθείσες μονάδες κώδικα. Ως εκ τούτου, οι περιγραφές είναι, κατ'ανάγκη, περιληπτικές. Ο ενδιαφερόμενος αναγνώστης καλείται να εξετάσει τον ίδιο τον κώδικα, σε περίπτωση που χρειάζεται περισσότερες πληροφορίες - άλλωστε η καλύτερη τεκμηρίωση ενός τμήματος κώδικα είναι ο ίδιος ο σχολιασμένος κώδικας.

[44]http://www.ilsp.gr/greeklish.html , ίσχυε την 5/7/2006

[45]http://www.autohotkey.com , ίσχυε την 5/7/2006

[46]http://jakarta.apache.org/commons/ , ίσχυε την 5/7/2006

[47]http://www.gnu.org/software/wget/ , ίσχυε την 5/7/2006

[48]http://sarovar.org/projects/javadbf , ίσχυε την 5/7/2006

[49]http://paginas.terra.com.br/informatica/download1/dekode_download.htm , ίσχυε την 5/7/2006

[50]http://openmap.bbn.com/ , ίσχυε την 5/7/2006

[51]http://www.env.gr/myenv/meletes_iliko/yliko/coords_gr_1.6.0.zip , ίσχυε την 5/7/2006

[52]http://www.maxmind.com/app/city , ίσχυε την 5/7/2006

5.3 Λεπτομέρειες υλοποίησης: Περιγραφή πακέτων

Πακέτο	*Περιέχει κλάσεις που αφορούν...*
geo	Βασικές κλάσεις του συστήματος
geo.coder	Την γεωκωδικοποίηση
geo.dcle	Τον καθαρισμό των γεωδεδομένων
geo.parser	Το geoparsing
geo.parser.lookup	Την προσεγγιστική αναζήτηση σε κείμενο
geo.lang	Το φωνητικό αλφάβητο και το προσεγγιστικό ταίριαγμα
geo.systemdependent	Την ομαλή ενσωμάτωση εφαρμογών που τρέχουν μόνο σε Windows, στο υπόλοιπο σύστημα
geo.tests	Τον αυτόματο και ημιαυτόματο έλεγχο του συστήματος

Πίνακας 3: Πακέτα του συστήματος

5.4 Λεπτομέρειες υλοποίησης: Περιγραφή κλάσεων

5.4.1 Πακέτο geo

5.4.1.1 geo.AbstractFileCrawler

- Παρέχει βασικές λειτουργίες εξομοίωσης ενός web crawler, πάνω σε σελίδες που έχουν ήδη ανακληθεί τοπικά στον δίσκο. Κληρονομώντας από αυτήν την κλάση, μπορούν να οριστούν αφ'ενός ο τύπος αρχείων που θα εξεταστούν (protected boolean accept(String filename)), αφ'ετέρου η επεξεργασία που θα υποστούν (protected void process(File file)).
- Χρήσιμες μέθοδοι:
  - public AbstractFileCrawler(String crawlRoot): Δημιουργείται ένας νέος AbstractFileCrawler, με τον κατάλογο crawlRoot σαν βάση. Εξετάζονται όλα τα αρχεία του καταλόγου.

5.4.1.2 geo.FileCrawler

- Υλοποιεί τον ανωτέρω geo.AbstractFileCrawler, προκειμένου για την επεξεργασία html αρχείων, και την εξαγωγή γεωγραφικής πληροφορίας από αυτά. Είναι η βασική κλάση που καλείται για την εκτέλεση του συστήματος, με ορίσματα τον κατάλογο που περιέχει τα αρχεία προς επεξεργασία, και, προαιρετικά, παραμέτρους βελτιστοποίησης (π.χ. απενεργοποίηση διαδικασίας Logging).

5.4.1.3 geo.CharsetDetector

- Ανιχνεύει, με διάφορους τρόπους, την πιθανότερη κωδικοσελίδα με την οποία έχει γραφεί μια ιστοσελίδα. Υλοποιεί τα όσα αναφέρονται στο 2.3.1.2.
- Χρήσιμες μέθοδοι:
  - static String getCharset(byte[] input) : Επιστρέφει την πιθανότερη κωδικοσελίδα με την οποία έχει γραφεί μια ιστοσελίδα, της οποίας το περιεχόμενο έχει φορτωθεί στην μνήμη, στον πίνακα input.
  - static String getReader(byte[] input) : Το ίδιο, αλλά επιστρέφει έναν Reader για την ιστοσελίδα.

5.4.1.4 geo.ParsingDetector

- Χρησιμοποιείται από τον ανωτέρω geo.CharsetDetector, για την ανίχνευση της κωδικοσελίδας που πιθανώς δηλώνεται στην ίδια ιστοσελίδα.

5.4.1.5 geo.DBInterfacer

- Χρησιμοποιείται από πολλές άλλες κλάσεις, κυρίως στην φάση καθαρισμού δεδομένων, για επικοινωνία με την βάση δεδομένων.
- Χρήσιμες μέθοδοι:
  - public void connect(), και οι παραλλαγές της: Γίνεται σύνδεση με την βάση, και εκτελείται η (αφηρημένη) μέθοδος protected void doStuffWith(Statement stmt). (Προφανώς έχει υλοποιηθεί η μέθοδος στην υλοποιούσα κλάση)

5.4.1.6 geo.LoggingUtils

- Περιέχει συναρτήσεις για την υποβοήθηση της διαδικασίας του Logging.
- Χρήσιμες μέθοδοι:
  - public static Logger initLogger(String className) : Αρχικοποιεί τον logger για μια κλάση.

5.4.1.7 geo.ResultReporter

- Δημιουργεί μια αναφορά σχετικά με τις γεωγραφικές πληροφορίες που βρέθηκαν σε μια σελίδα.
- Χρήσιμες μέθοδοι:
  - public static String report(String content, TokenBuffer annotations)

5.4.1.8 geo.Settings

- Περιέχει συγκεντρωμένες τις βασικές ρυθμίσεις του συστήματος.

5.4.1.9 geo.Timer

- Μια υποβοηθητική κλάση για την μέτρηση της επίδοσης του συστήματος

5.4.1.10 geo.WgetUtils

- Μια υποβοηθητική κλάση για την μετατροπή των ονομάτων των αρχείων HTML που βρίσκονται στο σκληρό δίσκο, και έχουν ανακτηθεί με το εργαλείο Wget, σε έγκυρες διευθύνσεις URL.

5.4.2 Πακέτο geo.coder

5.4.2.1 geo.coder.Geocoder

- Η βασική κλάση γεωκωδικοποίησης. Υλοποιεί τα όσα αναφέρονται στο 3.7.
- Χρήσιμες μέθοδοι:
  - GeocodingResult geocode(Map datum) : Γεωκωδικοποιεί την γεωγραφική πληροφορία datum (διεύθυνση, Τ.Κ., τηλέφωνο, κ.λ.π.).
  - void reIndex(): Ξαναδημιουργεί τα φωνητικά ευρετήρια στην βάση δεδομένων.

*5.4.2.2 geo.coder.*Geocoder*

geo.coder.AttikiGeocoder

geo.coder.GnsGeocoder

geo.coder.MapdekodeGeocoder

geo.coder.OldAttikiGeocoder

geo.coder.OteGeocoder

geo.coder.PostalGeocoder

geo.coder.ScannedGeocoder

- Κλάσεις για την γεωκωδικοποίηση δεδομένων από τα ομώνυμα datasets. Χρησιμοποιούνται από τον ανωτέρω geo.coder.Geocoder.
- Κοινή μέθοδος η void reIndex(), όπως και στον geo.coder.Geocoder.

5.4.2.3 geo.coder.DistrictInclusions

- Κλάση υποβοηθητική του geo.coder.Geocoder.

5.4.2.4 geo.coder.ScannedMapGeocoder

- Κλάση υποβοηθητική του geo.coder.Geocoder.

*5.4.2.5 geo.coder.*GeocodingResult*

geo.coder.GeocodingResult

geo.coder.IntermediateGeocodingResult

geo.coder.LocalisedGeocodingResult

- Αντιπροσωπεύουν αποτελέσματα γεωκωδικοποίησης, στην βασική τους μορφή (GeocodingResult), στην ενδιάμεση μορφή στην οποία περιέρχονται κατά την γεωκωδικοποίηση (IntermediateGeocodingResult), και στην τελική τους μορφή, όπου ρόλο παίζει και η θέση τους στο κείμενο (LocalisedGeocodingResult)
- Χρήσιμες μέθοδοι:
  - int getConfidencePercent()

    void setConfidencePercent(int confidencePercent)
  - double getLatitude()
  - double getLongitude()
  - double getMaxLatitude()
  - double getMaxLongitude()
  - double getMinLatitude()
  - double getMinLongitude()

    Getters και setters για τις ιδιότητες του αποτελέσματος.
  - boolean intersect(GeocodingResult gc)

    Το αποτέλεσμα τέμνεται με το gc, και επιστρέφεται true εαν τέμνονταν προσεγγιστικά, false ειδεμή
  - double intersectionDistance(GeocodingResult gc)

    Η ελάχιστη παράλληλη μετατόπιση του τρέχοντος αποτελέσματος, για να τμήσει το gc

5.4.2.6 geo.coder.PhoneticIndex

- Η βασική κλάση υλοποίησης του Φωνητικού Ευρετηρίου (βλ. 4.4.7)

- Χρήσιμες μέθοδοι:
  - void createIndex() : Δημιουργία του πίνακα-ευρετηρίου στη βάση
  - void populateIndex() : Γέμισμα του πίνακα-ευρετηρίου στη βάση, από τις εγγραφές του πίνακα
  - int[][] getCandidatePhrases(String phrase) : Εύρεση υποψηφίων εγγραφών στη βάση, που μοιάζουν με την phrase. Επιστρέφει πίνακα της μορφής: {{Κωδικός εγγραφής1, Κόστος1},{Κωδικός εγγραφής2, Κόστος2},...}
  - String getKeyColumn()

    String getPhoneticTableName()

    String getTableName()

    String getWordColumn()

    void setKeyColumn(String keyColumn)

    void setPhoneticTableName(String phoneticTableName)

    void setTableName(String tableName)

    void setWordColumn(String wordColumn)

    Getters και setters για τις βασικές ιδιότητες του ευρετηρίου
  - static int missingCharThreshold(int phraseLength) : Ένα ευριστικό κατώφλι για την σύγκριση συμβολοσειρών, με βάση το μήκος τους

5.4.2.7 geo.coder.ResultIntegrator

- Ολοκλήρωση/ συνεκτίμηση αποτελεσμάτων γεωκωδικοποίησης. Υλοποιεί τα όσα αναφέρονται στο 3.7.2.
- Χρήσιμες μέθοδοι:
  - static GeocodingResult geocodeIp(URL pageUrl) : Εύρεση πιθανής γεωγραφικής θέσης σελίδας με βάση τη διεύθυνση IP της
  - static List<LocalisedGeocodingResult> integrateResults (TokenBuffer in, URL pageUrl) : Ολοκλήρωση/ συνεκτίμηση αποτελεσμάτων γεωκωδικοποίησης.

5.4.3 Πακέτο geo.lang

5.4.3.1 geo.lang.FuzzyComparator

- Περιέχει συναρτήσεις που χρησιμοποιούνται στο προσεγγιστικό ταίριαγμα, και περιγράφονται στο κεφ 4.3.1.3.

- Χρήσιμες μέθοδοι:
  - static boolean areSimilar(String word1, String word2, boolean phonetic) : Είναι οι συμβολοσειρές word1, word2 όμοιες; (αν phonetic = true, τότε ο έλεγχος γίνεται με βάση την φωνητική ομοιότητα, ειδεμή οι φωνητικές ομοιότητες δεν λαμβάνονται υπόψη).
  - static int getDistance(String s, String t, boolean phonetic): Υπολογίζει την τροποποιημένη απόσταση Levenshtein των s,t. (αν phonetic = true, τότε ο έλεγχος γίνεται με βάση την φωνητική ομοιότητα, ειδεμή οι φωνητικές ομοιότητες δεν λαμβάνονται υπόψη). Καλεί μια από τις ακόλουθες, ανάλογα με τις λέξεις:
  - static int getEditDistance(String s, String t)

    static int getEditDistanceWithEllipsis(String s, String t)

    static int getPhoneticEditDistance(String s, String t)

    static int getPhoneticEditDistanceArray(String s, String t)

    static int getPhoneticEditDistanceArrayWithEllipsis(String s, String t)

    static int getPhoneticEditDistanceWithEllipsis(String s, String t)
  - static int threshold(int wordLength): Ένα ευριστικό κατώφλι για την σύγκριση συμβολοσειρών, με βάση το μήκος τους. Χρησιμοποιείται και στην areSimilar(String,String).

5.4.3.2 geo.lang.GreekUtils

- Περιέχει μερικές συναρτήσεις για μετατροπή προς το φωνητικό αλφάβητο που χρησιμοποιείται (βλ. 4.3.1.2).
- Χρήσιμες μέθοδοι:
  - static String getGreekSoundex(String phonetic): Επιστρέφει τον φωνητικό κώδικα μιας λέξης γραμμένης σε φωνητικό αλφάβητο.
  - static String getGreekSoundexAnyLength(String phonetic) : Το ίδιο με την προηγούμενη getGreekSoundex, αλλά ο φωνητικός κώδικας δεν έχει περιορισμένο μήκος.
  - static String getGreekSoundexMultiWord(String phonetic): Το ίδιο με την προηγούμενη getGreekSoundex, αλλά η συμβολοσειρά phonetic μπορεί να είναι και φράση (δηλ. να αποτελείται από πολλές λέξεις).
  - static String greek2Phonetic(String greek) :Μετατροπή μιας συμβολοσειράς σε πεζούς ελληνικούς χαρακτήρες, στο φωνητικό αλφάβητο.

- static String greek2PhoneticBestEffort(String greek) : Το ίδιο με πριν, αλλά σε περίπτωση μη αποδεκτών χαρακτήρων, το σύστημα δεν διαμαρτύρεται, και απλά τους αγνοεί.
- static String greeklish2Phonetic(String greeklish) :Μετατροπή μιας συμβολοσειράς σε πεζούς λατινικούς χαρακτήρες, στο φωνητικό αλφάβητο.
- static String greeklish2PhoneticBestEffort(String greeklish) : Το ίδιο με πριν, αλλά σε περίπτωση μη αποδεκτών χαρακτήρων, το σύστημα δεν διαμαρτύρεται, και απλά τους αγνοεί.
- static String greekStripDiacritics(String greek): Αφαίρεση τόνων, σημείων στίξεως, κ.λ.π., από μια συμβολοσειρά σε πεζούς ελληνικούς χαρακτήρες.
- static boolean isGreekLower(char ch) : Έλεγχος για ελληνικό πεζό χαρακτήρα
- static boolean longSoundexesEqual(String longSoundex1, String longSoundex2) : Έλεγχος συμβατότητας (όχι ισότητας!) δύο φωνητικών κωδικών μη περιορισμένου μήκους.
- static String mixed2Greek(String greek) : Μετατροπή μιας συμβολοσειράς σε πεζούς ελληνικούς ή λατινικούς χαρακτήρες, σε ελληνικούς χαρακτήρες.
- static String mixed2GreekBestEffort(String greek) : Το ίδιο με πριν, αλλά σε περίπτωση μη αποδεκτών χαρακτήρων, το σύστημα δεν διαμαρτύρεται, αλλά τους συμπεριλαμβάνει στο τελικό αποτέλεσμα.

5.4.3.3 geo.lang.Phoneme

- Αναπαριστά τους φθόγγους του φωνητικού αλφαβήτου.

5.4.3.4 geo.lang.PronunciationGroup

- Αναπαριστά τις ομαδοποιήσεις των φθόγγων του φωνητικού αλφαβήτου

5.4.4 Πακέτο geo.parser

5.4.4.1 geo.parser.GeoLexer

- Βασική κλάση για όλους τους τεχνολογητές κανονικών γραμματικών (βλ. 3.2.2).
- Χρήσιμες μέθοδοι:
  - void close() : Κλείνει την ροή εισόδου
  - TokenBuffer parse() : Τεχνολογεί την είσοδο, επιστρέφει μια ροή εξόδου λεκτικών μονάδων, ή μονάδων γεωγραφικής πληροφορίας.

- void setInput(Reader r)

  void setInputTokenBuffer(TokenBuffer in) : Θέτουν την ροή εισόδου του λεκτικού αναλυτή
- abstract void yyreset(Reader r) : Επαναρχικοποίηση του αναλυτή, με νέα είσοδο

*5.4.4.2 geo.parser.GeoLexer**

geo.parser.GeoLexerOne

geo.parser.GeoLexerTwo

geo.parser.GeoLexerTwoFeatureNames

geo.parser.GeoLexerThree

- Τα τρία επίπεδα τεχνολογητών που χρησιμοποιούνται (βλ. 3.2.2). Κληρονομούν την GeoLexer

geo.parser.GeoLexerThreePatterns

- Υποβοηθητική κλάση για τον GeoLexerThree, αναλαμβάνει την τεχνολόγηση επιμέρους μονάδων (Τοπωνύμια, διευθύνσεις κ.λ.π.), που ανιχνεύει ο τελευταίος. Χρησιμοποιείται και κατά την διαδικασία του Brill-style matching (βλ. 3.2.2.1).

5.4.4.3 geo.parser.GeoParser

- Συντονίζει το έργο του geoparsing και της γεωκωδικοποίησης και ολοκλήρωσης αποτελεσμάτων.
- Χρήσιμες μέθοδοι:
  - TokenBuffer parse(Reader in) : Κάνει geoparsing και γεωκωδικοποίηση επί των ευρεθεισών πληροφοριών, στο κείμενο που διαβάζει από την είσοδο in.

5.4.4.4 geo.parser.HTMLParser

- Μετατρέπει ένα κείμενο σε (πιθανώς εσφαλμένη) HTML, σε απλό κείμενο, λαμβάνοντας υπόψη τη δομή του πρώτου, όπως περιγράφεται στο 2.3.1.2.
- Χρήσιμες μέθοδοι:
  - static String getString(URL url, Reader in) : Κάνει την μετατροπή του εγγράφου HTML που διαβάζει από την είσοδο in, σε απλό κείμενο.

5.4.4.5 geo.parser.TokenBuffer

- Περιέχει αναγνωρισθείσες λεκτικές μονάδες, και ενδιάμεσες γεωγραφικές πληροφορίες, σε μορφή αναγνωρίσιμη τόσο από τους λεκτικούς αναλυτές, όσο και από άλλες μονάδες του προγράμματος.
- geo.parser.TokenBuffer.TokenBufferException : Η εξαίρεση που πετάει ο TokenBuffer σε περίπτωση σφάλματος.
- Χρήσιμες μέθοδοι:
  - void add(Map gt) : Προσθήκη δεδομένου
  - Map getToken(int i) : Ανάκτηση δεδομένου
  - int length() : Πλήθος δεδομένων

5.4.5 Πακέτο geo.parser.lookup

5.4.5.1 geo.parser.lookup.FeatureNameLookup

- Αναζήτηση πιθανών τοπωνυμίων. Χρησιμοποιεί το ευρετήριο κατακερματισμού μεμονωμένων λέξεων (βλ. 4.4.5)
- Χρήσιμες μέθοδοι:
  - String get(String word) : Αναζήτηση μιας λέξης που πιθανώς ανήκει σε τοπωνύμιο. Επιστρέφει τον πιθανότερο τύπο της, ή null αν δεν αποτελεί μέρος τοπωνυμίου. Κάνει χρήση του PhoneticWordMap.
  - static FeatureNameLookup getInstance() : Επιστροφή του μοναδικού (singleton) αντικειμένου FeatureNameLookup.

5.4.5.2 geo.parser.lookup.TokenLookup

- Αναζήτηση λέξεων-κλειδιών (π.χ. προσδιοριστικά οδών, τηλεφώνων κ.λ.π.). Χρησιμοποιεί απλό ευρετήριο κατακερματισμού.
- Χρήσιμες μέθοδοι :
  - String[] get(String word) : Αναζήτηση μιας λέξης που πιθανώς είναι λέξη-κλειδί. Επιστρέφει τον πιθανότερο τύπο της, και την κανονικοποιημένη μορφή της ή null αν δεν πρόκειται για λέξη-κλειδί. Κάνει χρήση του PhoneticWordMap.

5.4.5.3 geo.parser.lookup.TokenLookupWrapper

- Singleton wrapper για την TokenLookup

- Χρήσιμες μέθοδοι:
  - static TokenLookup getInstance() : Επιστροφή του μοναδικού (singleton) αντικειμένου TokenLookup

5.4.5.4 geo.parser.lookup.PhoneticMultiWordMap<T>

- Υλοποίηση του ευρετηρίου κατακερματισμού πολλαπλών λέξεων (βλ. 4.4.6)
- Χρήσιμες μέθοδοι:
  - Set<T> get(String phrase) : Αναζήτηση συμβολοσειράς. Επιστρέφει ένα σύνολο με τους πιθανούς τύπους/υποσημειώσεις της.
  - Set<T>[] getAllCombos(String phrase) : Το ίδιο με πριν, αλλά είναι δυνατό το ταίριαγμα προθέματος της φράσης. Επιστρέφονται σύνολα για το ταίριαγμα της πρώτης λέξης, των πρώτων δύο λέξεων κ.ο.κ.
  - void put(String phrase, T value) : Εισαγωγή φράσης με τον τύπο/υποσημείωσή της στο ευρετήριο.

5.4.5.5 geo.parser.lookup.PhoneticWordMap<T>

- Υλοποίηση του ευρετηρίου κατακερματισμού μεμονωμένων λέξεων (βλ. 4.4.5)
- Χρήσιμες μέθοδοι:
  - T get(String word) : Αναζήτηση λέξης. Επιστρέφει τον τύπο/υποσημείωσή της, ή null αν δεν βρεθεί.
  - void put(String word, T value) : Εισαγωγή λέξης με τον τύπο/υποσημείωσή της στο ευρετήριο.

5.4.5.6 geo.parser.lookup.ZipCodeLookup

- Αναζήτηση ταχυδρομικού κώδικα. Αποφαίνεται για την εγκυρότητα ή μη ενός ταχυδρομικού κώδικα.
- Χρήσιμες μέθοδοι:
  - boolean contains(String zip) : Είναι έγκυρος ο Τ.Κ. zip?
  - static ZipCodeLookup getInstance() : Επιστροφή του μοναδικού (singleton) αντικειμένου ZipCodeLookup.

5.4.6 Πακέτο geo.systemdependent

5.4.6.1 geo.systemdependent.AllGreekInterfacer

- Αναλαμβάνει την διασύνδεση με το πρόγραμμα All Greek To Me!
- Χρήσιμες μέθοδοι:
  - static String greeklish2Greek(String greeklish) : Μετατροπή συμβολοσειράς από greeklish σε ελληνικά, μέσω του All Greek To Me!. Λειτουργεί μόνο σε σύστημα Windows.

5.4.6.2 geo.systemdependent.CoordGRInterfacer

- Αναλαμβάνει την διασύνδεση με το πρόγραμμα CoordGR.
- Χρήσιμες μέθοδοι:
  - static double[] egsa2wgs(double easting, double northing) : Μετατροπή συντεταγμένων από ΕΓΣΑ87 σε WGS84, μέσω του CoordGR. Λειτουργεί μόνο σε σύστημα Windows.

5.4.7 Πακέτο geo.tests

5.4.7.1 geo.tests.AllGreekInterfacerTest

- Ελέγχει την ορθή λειτουργία του geo.systemdependent.AllGreekInterfacer

5.4.7.2 geo.tests.GeocodersTest

- Ελέγχει την ορθή λειτουργία των επιμέρους γεωκωδικοποιητών

5.4.7.3 geo.tests.GeocodingResultMBRTest

- Ελέγχει την ορθή λειτουργία της γεωγραφικής/χωρικής ένωσης.

5.4.7.4 geo.tests.GreekUtilsTest

- Ελέγχει την ορθή λειτουργία των μεθόδων που σχετίζονται με το φωνητικό αλφάβητο, και με το προσεγγιστικό ταίριαγμα.

5.4.7.5 geo.tests.IPLookupTest

- Ελέγχει την ορθή λειτουργία της εξαγωγής γεωγραφικής πληροφορίας από διευθύνσεις IP.

5.4.7.6 geo.tests.PhoneticIndexTest

- Ελέγχει την ορθή λειτουργία του Φωνητικού Ευρετηρίου.

5.4.7.7 geo.tests.PhoneticParamSet

- Ελέγχει την βελτιστότητα των παραμέτρων του Φωνητικού Ευρετηρίου, ως προς ένα σύνολο δεδομένων.

5.4.7.8 geo.tests.TestFuzzyComparator

- Ελέγχει την ορθή λειτουργία των επιμέρους συναρτήσεων προσεγγιστικού ταιριάγματος.

5.4.8 Πακέτο geo.dcle

*5.4.8.1 geo.dcle.**

- Επιμέρους κλάσεις καθαρισμού δεδομένων. Σχετίζονται άμεσα με τα σύνολα δεδομένων που καθαρίζουν, ως εκ τούτου δεν έχει νόημα η λεπτομερής παρουσίασή τους. Ο ενδιαφερόμενος αναγνώστης παραπέμπεται απ'ευθείας στον κώδικα...

6 Έλεγχος και Αποτελέσματα

6.1 Μεθοδολογία Ελέγχου

Για την εξασφάλιση της ορθότητας και ποιότητας του κώδικα, χρησιμοποιήθηκε η μέθοδος αυτοματοποιημένου και ημιαυτοματοποιημένου ελέγχου με τη μέθοδο του "μαύρου κουτιού", καθώς και η μέθοδος ελέγχου παλινδρόμησης (regression testing). Συγκεκριμένα, για όσες μονάδες του κώδικα κρίθηκε αναγκαίο, γράφτηκαν περιπτώσεις δοκιμής, ή διαλογικά προγράμματα δοκιμής των μονάδων αυτων, τα οποία εκτελούνταν σε κάθε σημαντική αλλαγή του κώδικα, για την πιστοποίηση της ορθότητάς του. Σημαντική συνεισφορά στον έλεγχο είχε ο καθαρισμός των γεωδεδομένων, αφού αυτός έγινε με μεθόδους του ίδιου του συστήματος. Έτσι, το σύστημα έτρεξε για αρκετές συνεχόμενες ώρες, αποδεικνύοντας τόσο την ορθότητά του, όσο και την σταθερότητά του και την ανοχή του σε σφάλματα και υψηλό φόρτο. Τέλος, το σύστημα χρησιμοποιήθηκε για την γεωκωδικοποίηση μεγάλου αριθμού ιστοσελίδων, διαφόρων κατηγοριών, και τα αποτελέσματα ελέγχθηκαν για επαλήθευση της ποιότητάς τους.

6.2 Αποτελέσματα[53]

Ακολουθούν ενδεικτικά αποτελέσματα εκτέλεσης του προγράμματος, επί διαφόρων σελίδων, ποικίλων κατηγοριών, για την έμπρακτη απόδειξη των προαναφερθέντων. Τα αποτελέσματα κρίνονται πολύ ικανοποιητικά, αφού συγκρίθηκαν με τις αντίστοιχες πληροφορίες που θα επεσήμαινε ένας ανθρώπινος αναγνώστης της σελίδας, και δεν βρέθηκαν να διαφέρουν σημαντικά.

[53]Οι χάρτες που εικονίζονται έχουν ληφθεί από το Google Earth, http://maps.google.com , ίσχυε την 6/7/2006

6.2.1 Τουριστικές ιστοσελίδες[54]

Σχήμα 7: Τουριστική ιστοσελίδα για το Πανταβρέχει

- Το σύστημα εντόπισε τις ακόλουθες γεωγραφικές πληροφορίες στη σελίδα:

Παρ' ότι το Πανταβρέχει δεν είναι κλασικό φαράγγι με την έννοια του ορισμού αλλά το στενότερο σημείο του Κρικελοπόταμου, που ρέει ανάμεσα στα βουνά Καλιακούδα (2.101 μ.) και Πλατανάκι (1.777 μ.), το εντάξαμε στις διαδρομές μας γιατί το θέαμα στη συγκεκριμένη θέση είναι τόσο εντυπωσιακό που όποιος δεν το έχει αντικρίσει είναι σα να μην έχει δει ποτέ στη ζωή του φαράγγι. Στη συγκεκριμένη θέση το ποτάμι στενεύει πολύ σχηματίζοντας στενό φαράγγι μήκους 100 περίπου μέτρων, με τοπίο σπάνιας ομορφιάς. Τα νερά της απόκρημνης Καλιακούδας στη νότια πλευρά του βουνού βρίσκουν διέξοδο και σχηματίζουν πηγές που χύνονται από μεγάλο υψόμετρο στον Κρικελοπόταμο. Το σημείο αυτό ονομάζεται Πανταβρέχει

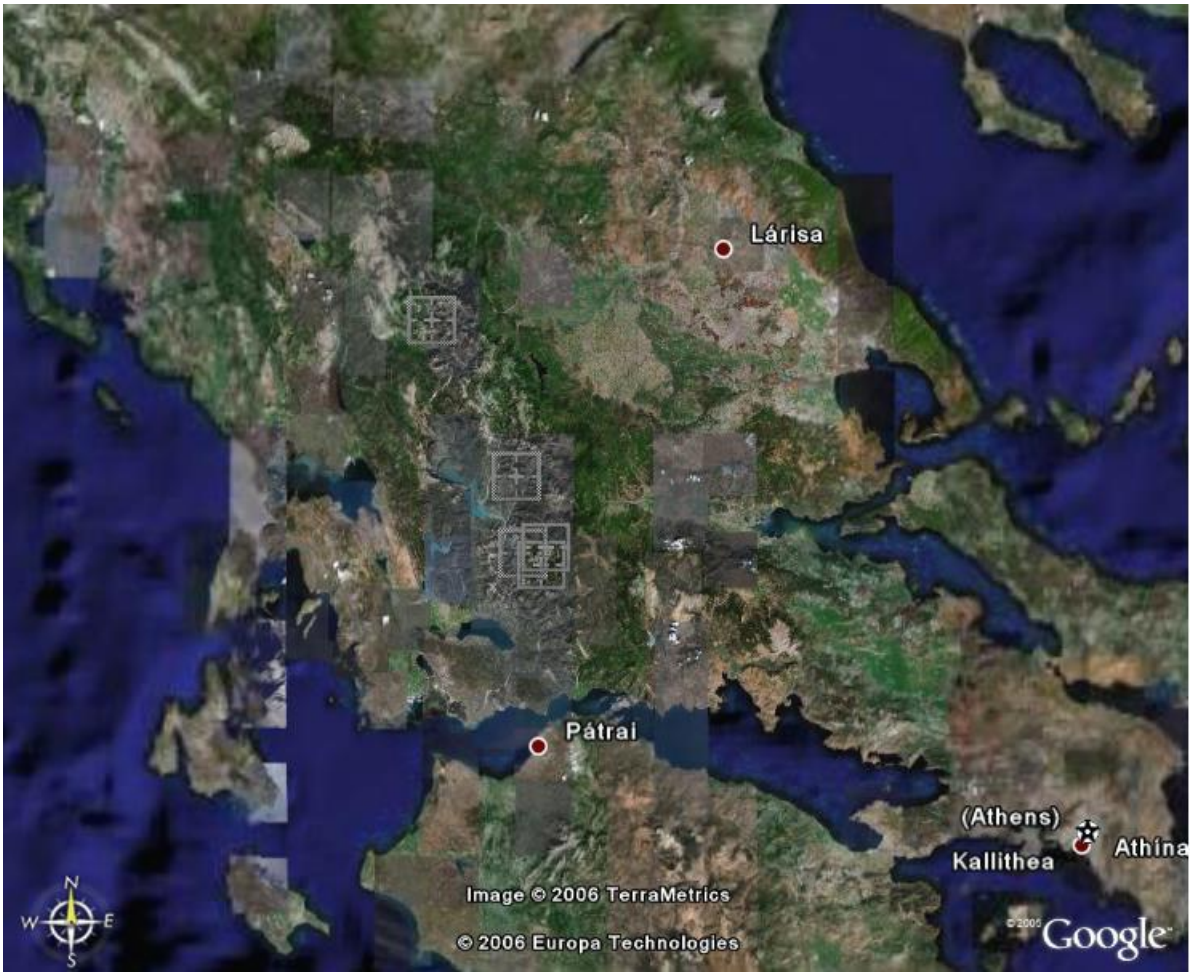
Τα νερά εδώ έχουν διαβρώσει το έδαφος που είναι σκεπασμένα με μακριά βρύα, ρέουν σε αρνητική κλίση και έτσι μετατρέπονται σε υδάτινα παραπετάσματα. Το νερό δημιουργεί λίμνες που για να τις προσπελάσει κανείς χρειάζεται να βραχεί ως την μέση. Η φύση εδώ είναι οργιώδης σε μία από τις αγριότερες και συνάμα καθαρότερες περιοχές της Ευρώπης. (Υπάρχουν σπάνια είδη χλωρίδας όπως τα κρίνα Lilium Heildrechii και πολυάριθμες φυτοκοινωνίες των σαρκοβόρων Pinguicola hirtiflora). Στον Κρικελοπόταμο ζουν πέστροφες που με τη σειρά τους συντηρούν της βίδρες του ποταμού.

[54]http://www.e-city.gr/evrytania/home/view/1312.php και http://www.e-city.gr/evrytania/home/view/1200.php , ίσχυαν την 6/7/2006

Η προσέγγιση στο Πανταβρέχει είναι όμως από δύσκολη έως προβληματική και γι' αυτό δεν είναι ευρέως γνωστό. Οι ανεκτικότερες διαδρομές είναι από Δομνίστα Ευρυτανίας και Κόνισκα Αιτωλοακαρνανίας. Πλησιέστερος οικισμός από πλευράς Ευρυτανίας το άλλοτε κεφαλοχώρι της περιοχής, η γεωργοκτηνοτροφική Ροσκά που κάποτε συντηρούσε εκατοντάδες πολυμελείς οικογένειες. Το χωριό είναι χτισμένο σε εκπληκτική τοποθεσία σε υψόμετρο 1000 μ., σε λάκα του βουνού Πλατανάκι.

(Οι περισσότεροι κάτοικοι της Ροσκάς σήμερα ζουν στην Αιτωλοακαρνανία και επισκέπτονται το χωριό τους καλοκαιρινούς μήνες. Σήμερα οι λιγοστοί κάτοικοι που απόμειναν συνάζονται στο μοναδικό καφενείο του χωριού).

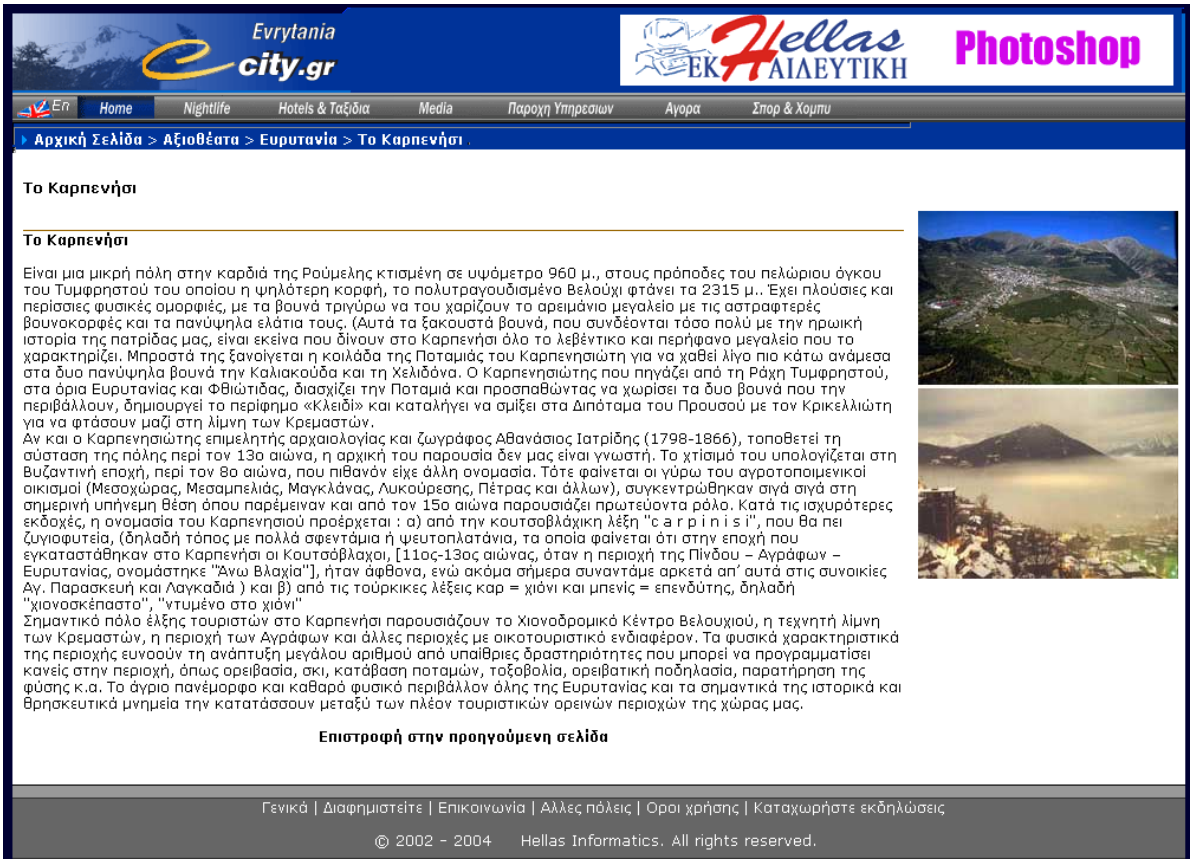
- Και τις τοποθέτησε ως εξής στον χάρτη:



Σχήμα 8: Οι επιμέρους τοποθεσίες της σελίδας

- Έπειτα από συνεκτίμηση των αποτελεσμάτων, προέκυψε η ακόλουθη θέση για την ιστοσελίδα:

Σχήμα 9: Το γεωγραφικό εύρος της σελίδας



Σχήμα 10: Τουριστική Ιστοσελίδα για το Καρπενήσι

- Το σύστημα εντόπισε τις ακόλουθες γεωγραφικές πληροφορίες στη σελίδα:

Το Καρπενήσι

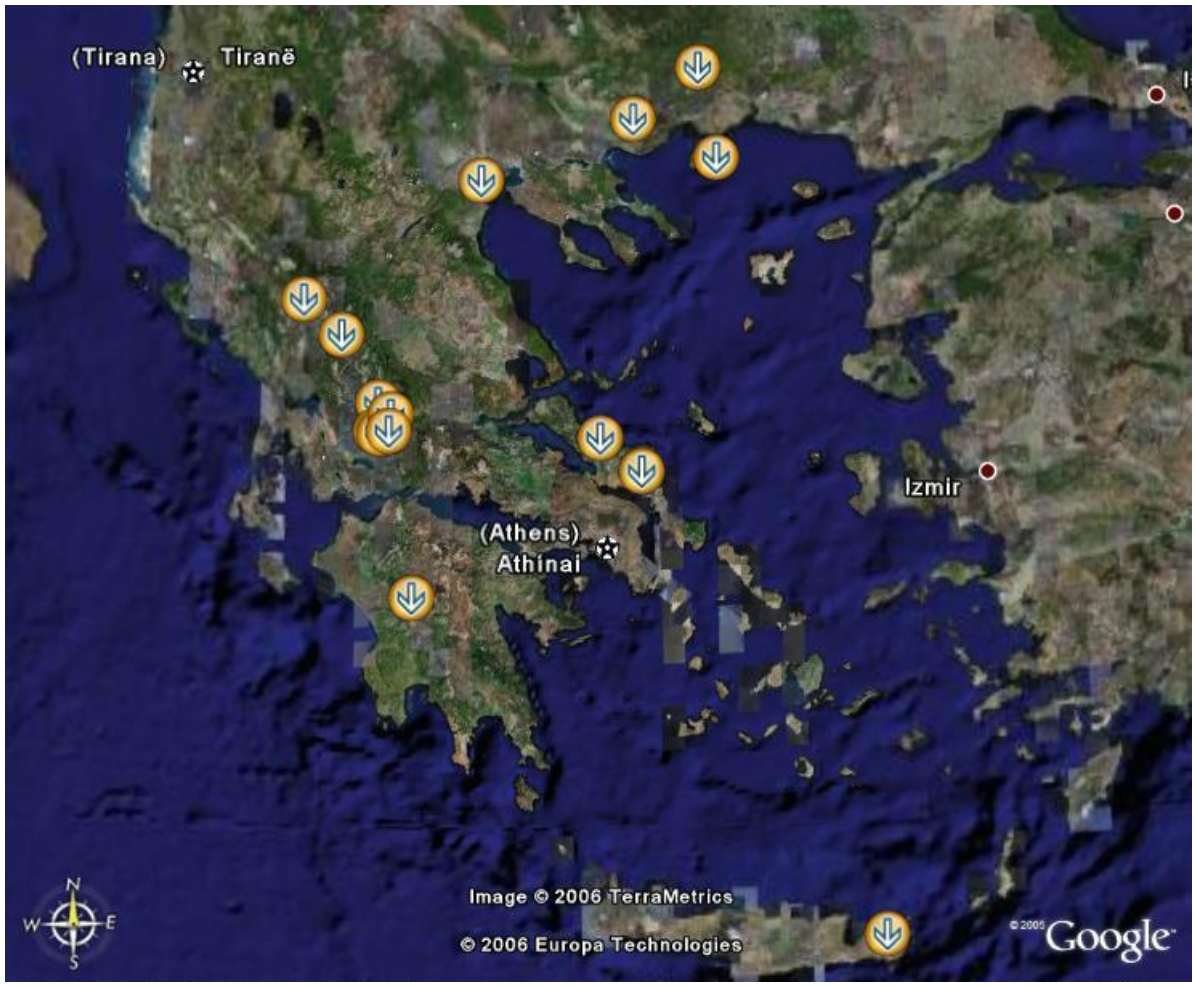
Είναι μια μικρή πόλη στην καρδιά της Ρούμελης κτισμένη σε υψόμετρο 960 μ., στους πρόποδες του πελώριου όγκου του Τυμφρηστού του οποίου η ψηλότερη κορφή, το πολυτραγουδισμένο Βελούχι φτάνει τα 2315 μ.. Έχει πλούσιες και περίσσιες φυσικές ομορφιές, με τα βουνά τριγύρω να του χαρίζουν το αρειμάνιο μεγαλείο με τις αστραφτερές βουνοκορφές και τα πανύψηλα ελάτια τους. (Αυτά τα ξακουστά βουνά, που συνδέονται τόσο πολύ με την ηρωική ιστορία της πατρίδας μας, είναι εκείνα που δίνουν στο Καρπενήσι όλο το λεβέντικο και περήφανο μεγαλείο που το χαρακτηρίζει. Μπροστά της ξανοίγεται η κοιλάδα της Ποταμιάς του Καρπενησιώτη για να χαθεί λίγο πιο κάτω ανάμεσα στα δυο πανύψηλα βουνά την Καλιακούδα και τη Χελιδόνα. Ο Καρπενησιώτης που πηγάζει από τη Ράχη Τυμφρηστού, στα όρια Ευρυτανίας και Φθιώτιδας, διασχίζει την Ποταμιά και προσπαθώντας να χωρίσει τα δυο βουνά που την περιβάλλουν, δημιουργεί το περίφημο Κλειδί και καταλήγει να σμίξει στα Διπόταμα του Προυσού με τον Κρικελλιώτη για να φτάσουν μαζί στη λίμνη των Κρεμαστών.

Αν και ο Καρπενησιώτης επιμελητής αρχαιολογίας και ζωγράφος Αθανάσιος Ιατρίδης (1798-1866), τοποθετεί τη σύσταση της πόλης περί τον 13ο αιώνα, η αρχική του παρουσία δεν μας είναι γνωστή. Το χτίσιμό του υπολογίζεται στη Βυζαντινή εποχή, περί τον 8ο αιώνα, που πιθανόν είχε άλλη ονομασία. Τότε φαίνεται οι γύρω του αγροτοποιμενικοί οικισμοί (Μεσοχώρας, Μεσαμπελιάς,

Μαγκλάνας, Λυκούρεσης, Πέτρας και άλλων), συγκεντρώθηκαν σιγά σιγά στη σημερινή υπήνεμη θέση όπου παρέμειναν και από τον 15ο αιώνα παρουσιάζει πρωτεύοντα ρόλο. Κατά τις ισχυρότερες εκδοχές, η ονομασία του Καρπενησιού προέρχεται : α) από την κουτσοβλάχικη λέξη c a r p i n i s i, που θα πει ζυγιοφυτεία, (δηλαδή τόπος με πολλά σφεντάμια ή ψευτοπλατάνια, τα οποία φαίνεται ότι στην εποχή που εγκαταστάθηκαν στο Καρπενήσι οι Κουτσόβλαχοι, [11ος-13ος αιώνας, όταν η περιοχή της Πίνδου Αγράφων Ευρυτανίας, ονομάστηκε Άνω Βλαχία], ήταν άφθονα, ενώ ακόμα σήμερα συναντάμε αρκετά απ' αυτά στις συνοικίες Αγ. Παρασκευή και Λαγκαδιά) και β) από τις τούρκικες λέξεις καρ-χιόνι και μπενίς-επενδύτης, δηλαδή χιονοσκέπαστο, ντυμένο στο χιόνι.

Σημαντικό πόλο έλξης τουριστών στο Καρπενήσι παρουσιάζουν το Χιονοδρομικό Κέντρο Βελουχιού, η τεχνητή λίμνη των Κρεμαστών, η περιοχή των Αγράφων και άλλες περιοχές με οικοτουριστικό ενδιαφέρον. Τα φυσικά χαρακτηριστικά της περιοχής ευνοούν τη ανάπτυξη μεγάλου αριθμού από υπαίθριες δραστηριότητες που μπορεί να προγραμματίσει κανείς στην περιοχή, όπως ορειβασία, σκι, κατάβαση ποταμών, τοξοβολία, ορειβατική ποδηλασία, παρατήρηση της φύσης κ.α. Το άγριο πανέμορφο και καθαρό φυσικό περιβάλλον όλης της Ευρυτανίας και τα σημαντικά της ιστορικά και θρησκευτικά μνημεία την κατατάσσουν μεταξύ των πλέον τουριστικών ορεινών περιοχών της χώρας μας.

- Και τις τοποθέτησε ως εξής στον χάρτη:



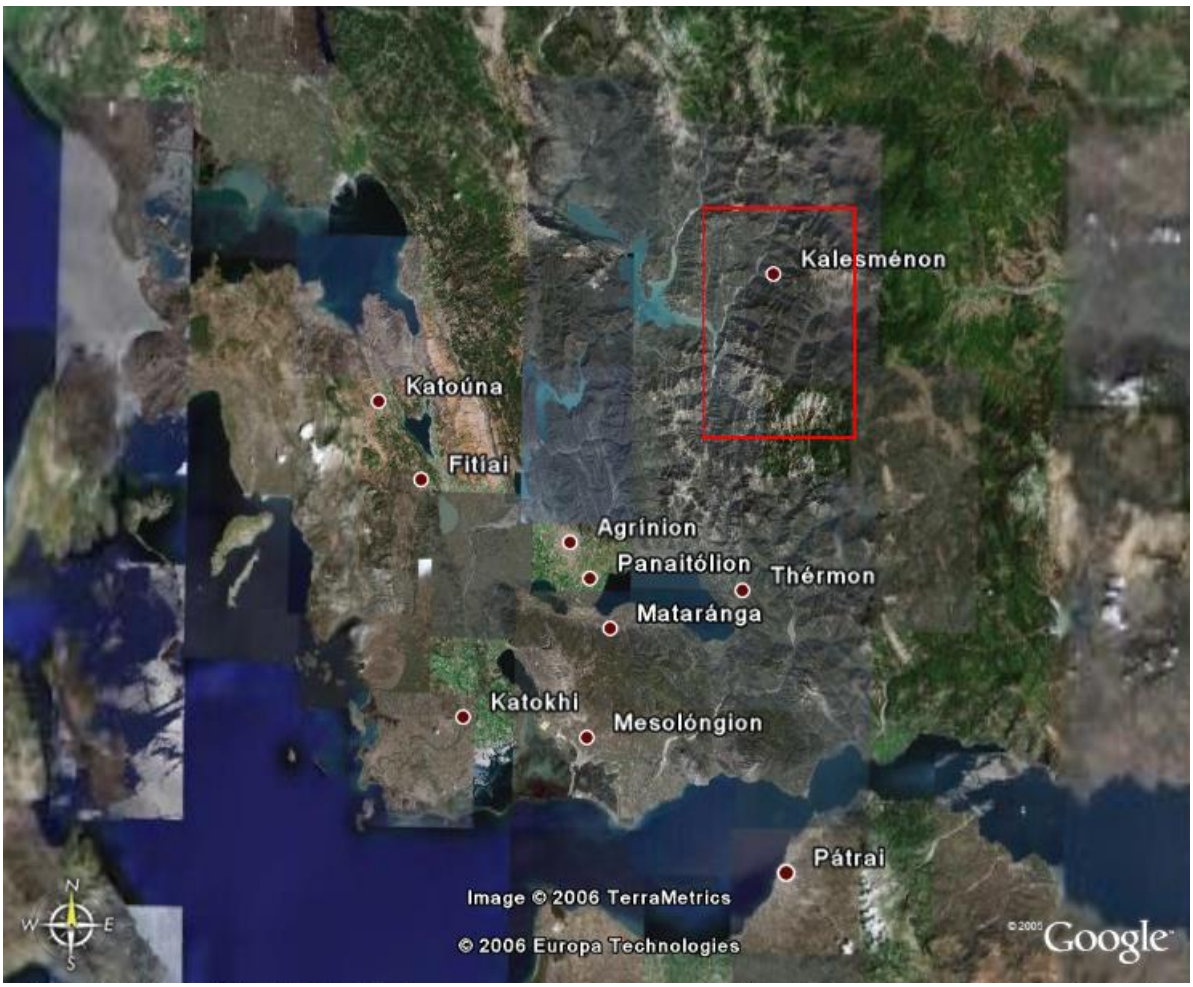
Σχήμα 11: Οι επιμέρους τοποθεσίες της σελίδας

Παρατηρούμε ότι οι τοποθεσίες είναι διάσπαρτες στον χάρτη. Τονίζουμε ότι σε όλες τις τοποθεσίες έχουν αντιστοιχηθεί χαμηλές βεβαιότητες (επειδή πρόκειται για σκέτα τοπωνύμια).

Επίσης, αν παρατηρήσουμε την (ορθή) περιοχή της Ευρυτανίας, παρατηρούμε αυξημένη συγκέντρωση τοποθεσιών (βλ. σχ. 12). Αναμένουμε (όπως και γίνεται), το εύρος της σελίδας που θα υπολογιστεί να βρίσκεται όντως στην Ευρυτανία.

Σχήμα 12: Οι επιμέρους τοποθεσίες στην περιοχή της Ευρυτανίας

- Έπειτα από συνεκτίμηση των αποτελεσμάτων, προέκυψε, όντως, η ακόλουθη θέση για την ιστοσελίδα:



Σχήμα 13: Το γεωγραφικό εύρος της ιστοσελίδας

6.2.2 Εμπορικές ιστοσελίδες[55]

Σχήμα 14: Εμπορική Ιστοσελίδα 1

- Το σύστημα εντόπισε τις ακόλουθες γεωγραφικές πληροφορίες στη σελίδα:

Catering Services Athens: Aidinis - Restaurants Catering Athens - Greece

Ηρώων Πολυτεχνείου 106 - Κάτω Χαλάνδρι Τ.Κ 152 31

Tel: 010 6713883, 010 6713179, 010 6813865

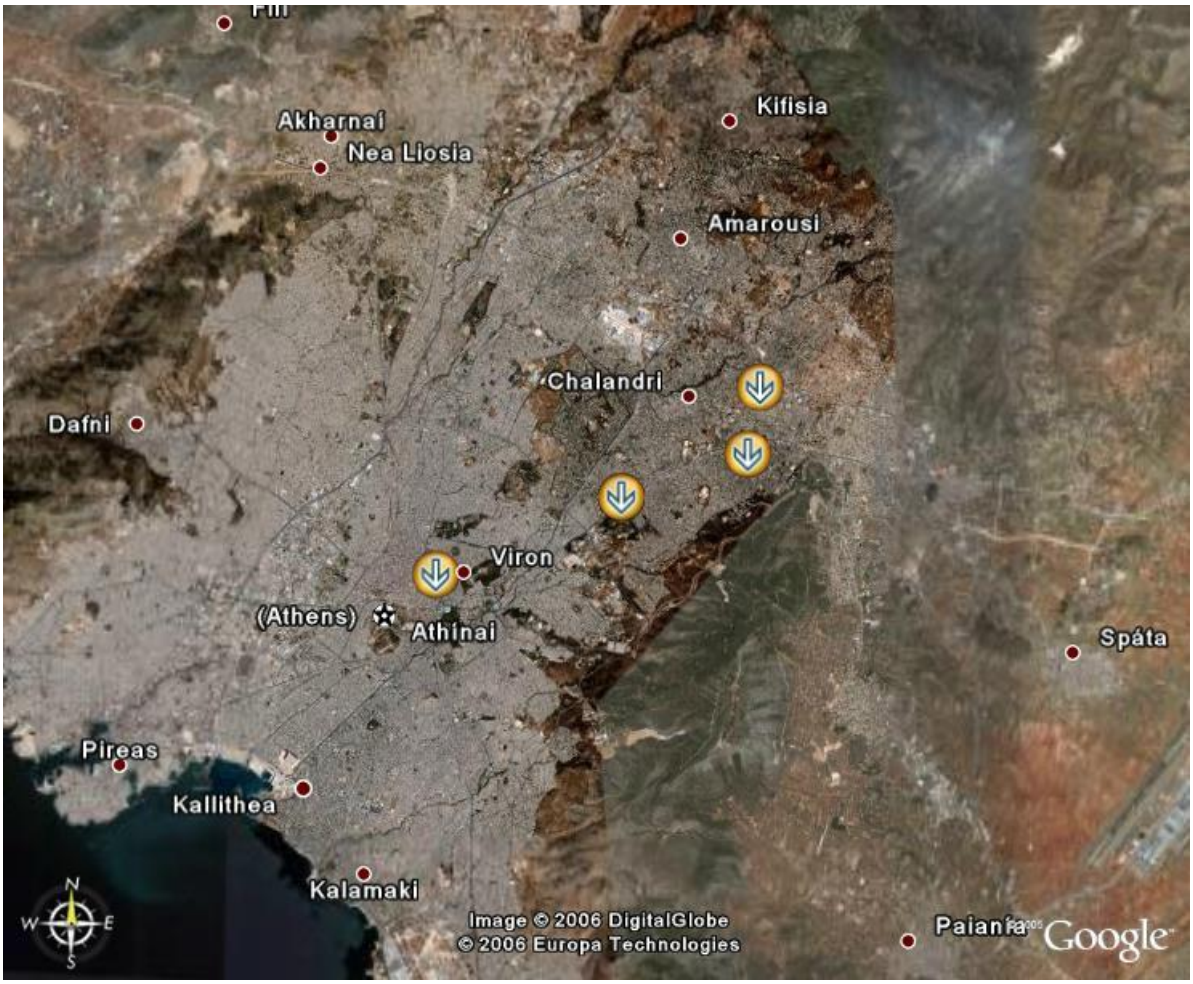
Fax: 010 6713748

e-mail: info@aidinis-since1931.gr

Powered by MARINET

[55]http://www.aidinis-since1931.gr/ , http://www.astrafoods.gr/ , http://www.alfacatering.gr/epikoinonia.htm , ίσχυαν την 6/7/2006

- Και τις τοποθέτησε ως εξής στον χάρτη:



Σχήμα 15: Οι επιμέρους τοποθεσίες της σελίδας

- Έπειτα από συνεκτίμηση των αποτελεσμάτων, προέκυψε η ακόλουθη θέση για την ιστοσελίδα:

Σχήμα 16: Το γεωγραφικό εύρος της ιστοσελίδας

Σχήμα 17: Εμπορική Ιστοσελίδα 2

- Το σύστημα εντόπισε τις ακόλουθες γεωγραφικές πληροφορίες στη σελίδα:

Astrafoods Catering

Chief Executive Officers

H. Papadopoulos, M. Papadopoulos

Headquarters

I str. Industrial Area

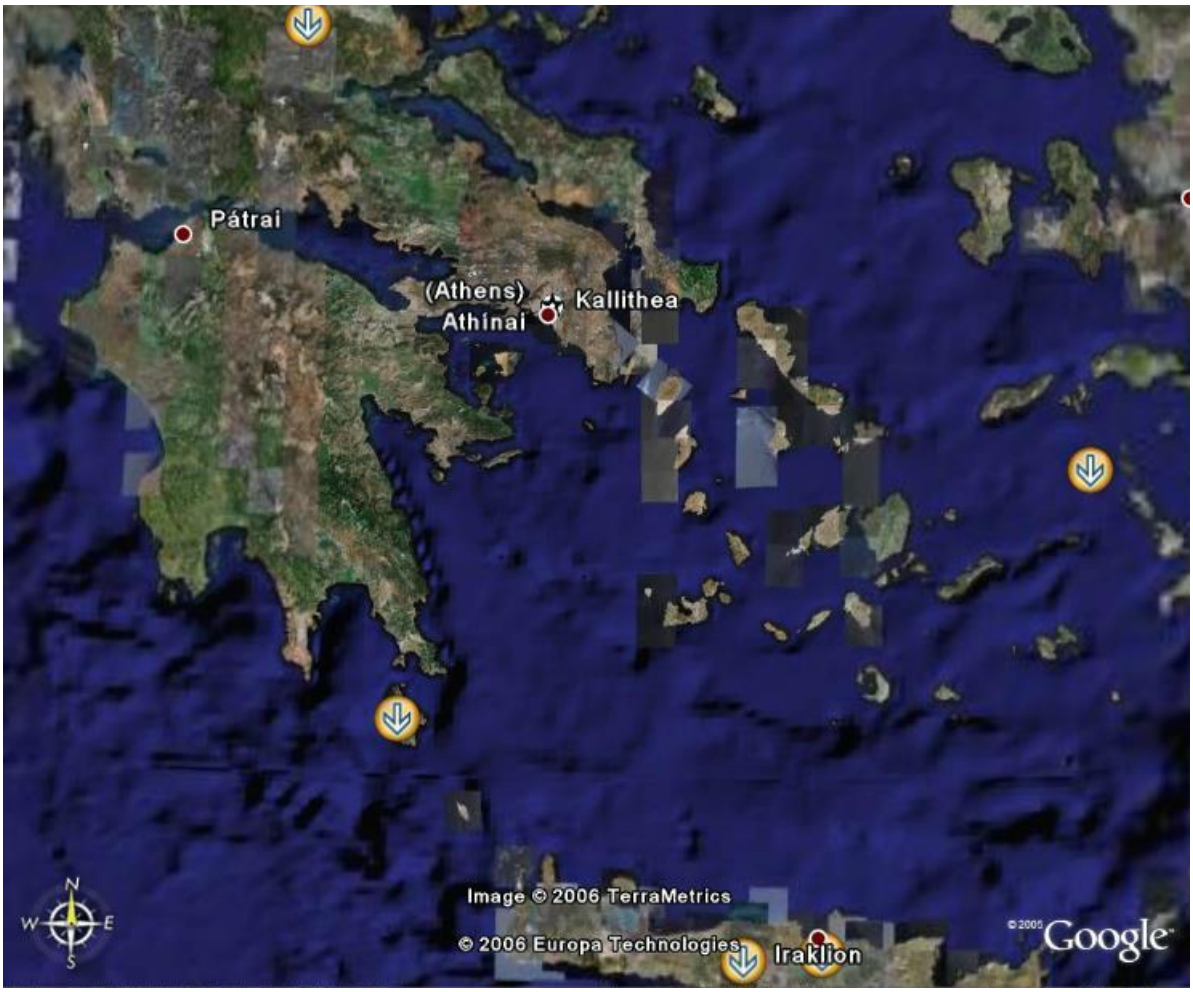
Heraklion, Crete, Greece

Tel.: +30 81 381038

Fax.: +30 81 381138

SITA: HERAFXH

- Και τις τοποθέτησε ως εξής στον χάρτη:



Σχήμα 18: Οι επιμέρους τοποθεσίες της ιστοσελίδας

Ας σημειωθεί ότι, παρότι εμφανίζονται αρκετά διάσπαρτα αποτελέσματα, εντούτοις τα αποτελέσματα που αναφέρονται στην περιοχή του Ηρακλείου έχουν σημαντικά μεγαλύτερο συντελεστή βεβαιότητας, και άρα αναμένουμε να αγνοηθούν τα άσχετα αποτελέσματα.

- Πράγματι, έπειτα από συνεκτίμηση των αποτελεσμάτων, προέκυψε η ακόλουθη θέση για την ιστοσελίδα:

Σχήμα 19: Το γεωγραφικό εύρος της ιστοσελίδας

Σχήμα 20: Εμπορική Ιστοσελίδα 3

- Το σύστημα εντόπισε τις ακόλουθες γεωγραφικές πληροφορίες στη σελίδα:

Επικοινωνία

Έδρα : Κεντρική Λαχαναγορά Αθηνών - Υπόστεγα 58-59

Τηλέφωνα : 210-4833996, 210-4838348

Μονάδα επεξεργασίας : Ομήρου 140, Μοσχάτο Τηλ.210-9417977

Τηλέφωνο - FAX παραγγελιών : 210-9402241

Κινητό : 6944-882366

E-Mail : alexandrosfruits@hotmail.com

Copyright 2004 Alfa Catering LTD. All rights reserved.

- Και τις τοποθέτησε ως εξής στον χάρτη:

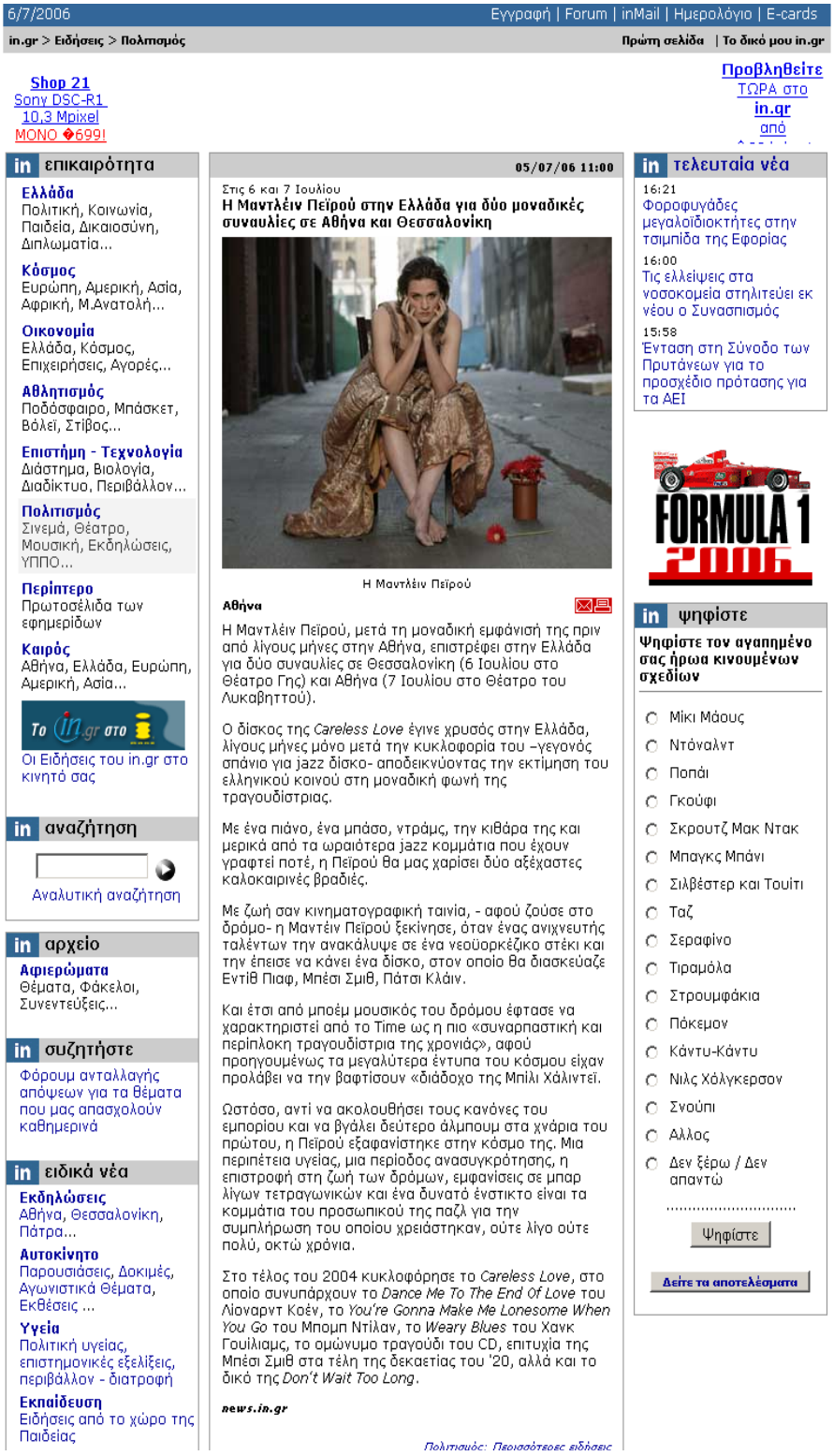
Σχήμα 21: Οι επιμέρους τοποθεσίες της ιστοσελίδας

(Σημείωση: Βρέθηκε και μια τοποθεσία στην Κρήτη (το τοπωνύμιο "Άλφα"), αλλά με ιδιαίτερα χαμηλή βεβαιότητα. Σε συνδυασμό με την μεγάλη της απόσταση από τις υπόλοιπες τοποθεσίες, αναμένουμε να απορριφθεί στην τελική συνεκτίμηση).

- Έπειτα από συνεκτίμηση των αποτελεσμάτων, προέκυψε η ακόλουθη θέση για την ιστοσελίδα:

Σχήμα 22: Το γεωγραφικό εύρος της ιστοσελίδας

6.2.3 *Ειδησεογραφικές ιστοσελίδες* [56]



Σχήμα 23: Ειδησεογραφική Ιστοσελίδα από το in.gr

[56]http://www.in.gr/news/article.asp?lngEntityID=719594&lngDtrID=253 , ίσχυε την 6/7/2006

- Το σύστημα εντόπισε τις ακόλουθες γεωγραφικές πληροφορίες στη σελίδα:

Στις 6 και 7 Ιουλίου

Η Μαντλέιν Πεϊρού στην Ελλάδα για δύο μοναδικές συναυλίες σε Αθήνα και Θεσσαλονίκη

Η Μαντλέιν Πεϊρού

Αθήνα

Η Μαντλέιν Πεϊρού, μετά τη μοναδική εμφάνισή της πριν από λίγους μήνες στην Αθήνα, επιστρέφει στην Ελλάδα για δύο συναυλίες σε Θεσσαλονίκη (6 Ιουλίου στο Θέατρο Γης) και Αθήνα (7 Ιουλίου στο Θέατρο του Λυκαβηττού).

Ο δίσκος της Careless Love έγινε χρυσός στην Ελλάδα, λίγους μήνες μόνο μετά την κυκλοφορία του, γεγονός σπάνιο για jazz δίσκο- αποδεικνύοντας την εκτίμηση του ελληνικού κοινού στη μοναδική φωνή της τραγουδίστριας..

Με ένα πιάνο, ένα μπάσο, ντράμς, την κιθάρα της και μερικά από τα ωραιότερα jazz κομμάτια που έχουν γραφτεί ποτέ, η Πεϊρού θα μας χαρίσει δύο αξέχαστες καλοκαιρινές βραδιές.

Με ζωή σαν κινηματογραφική ταινία, - αφού ζούσε στο δρόμο- η Μαντέιν Πεϊρού ξεκίνησε, όταν ένας ανιχνευτής ταλέντων την ανακάλυψε σε ένα νεοϋορκέζικο στέκι και την έπεισε να κάνει ένα δίσκο, στον οποίο θα διασκεύαζε Εντίθ Πιαφ, Μπέσι Σμιθ, Πάτσι Κλάιν.

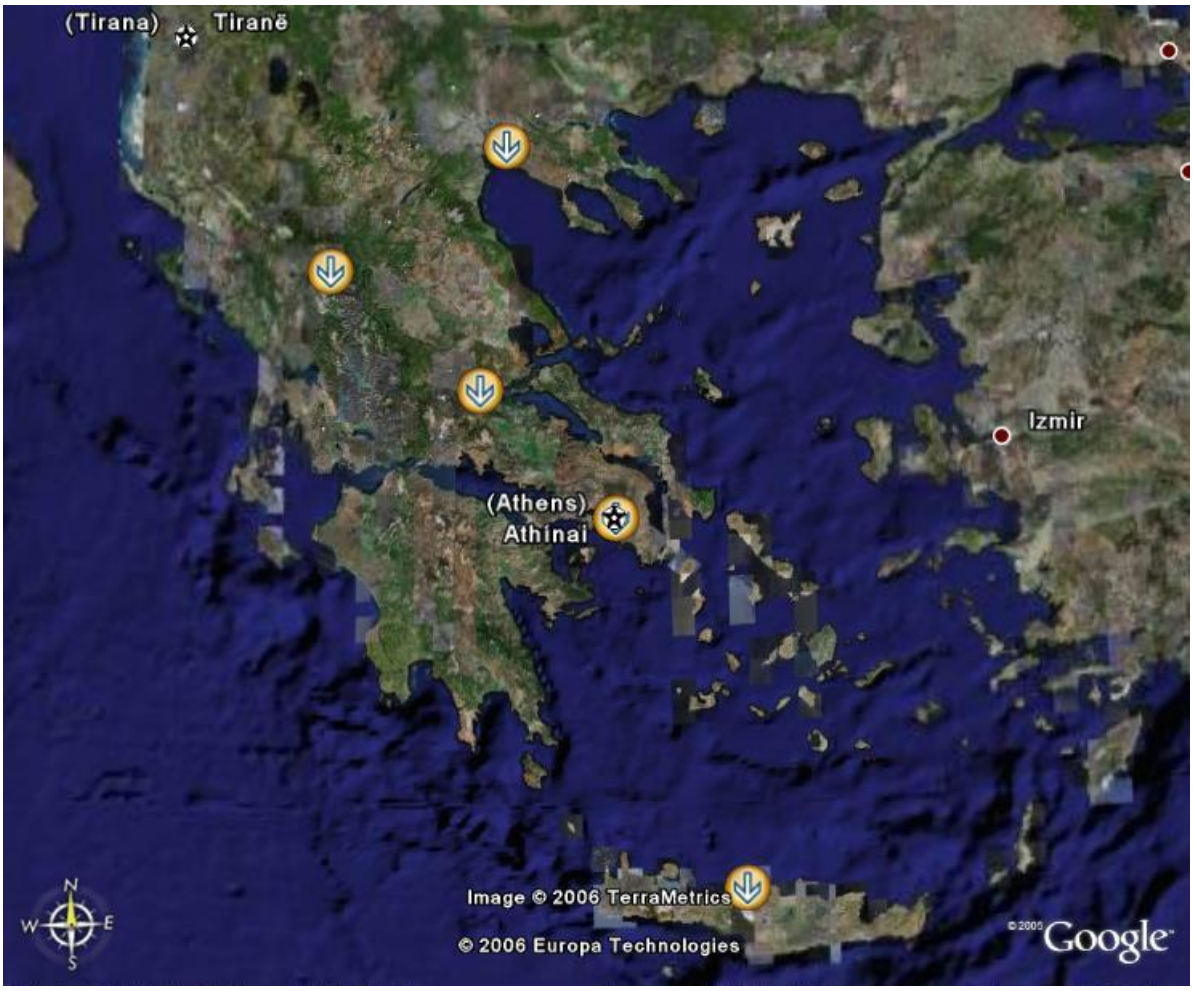
Και έτσι από μποέμ μουσικός του δρόμου έφτασε να χαρακτηριστεί από το Time ως η πιο συναρπαστική και περίπλοκη τραγουδίστρια της χρονιάς, αφού προηγουμένως τα μεγαλύτερα έντυπα του κόσμου είχαν προλάβει να την βαφτίσουν διάδοχο της Μπίλι Χάλιντεϊ.

Ωστόσο, αντί να ακολουθήσει τους κανόνες του εμπορίου και να βγάλει δεύτερο άλμπουμ στα χνάρια του πρώτου, η Πεϊρού εξαφανίστηκε στην κόσμο της. Μια περιπέτεια υγείας, μια περίοδος ανασυγκρότησης, η επιστροφή στη ζωή των δρόμων, εμφανίσεις σε μπαρ λίγων τετραγωνικών και ένα δυνατό ένστικτο είναι τα κομμάτια του προσωπικού της παζλ για την συμπλήρωση του οποίου χρειάστηκαν, ούτε λίγο ούτε πολύ, οκτώ χρόνια.

Στο τέλος του 2004 κυκλοφόρησε το Careless Love , στο οποίο συνυπάρχουν το Dance Me To The End Of Love του Λίοναρντ Κοέν, το You ' re Gonna Make Me Lonesome When You Go του Μπομπ Ντίλαν, το Weary Blues του Χανκ Γουίλιαμς, το ομώνυμο τραγούδι του CD, επιτυχία της Μπέσι Σμιθ στα τέλη της δεκαετίας του '20, αλλά και το δικό της Don ' t Wait Too Long .

news.in.gr

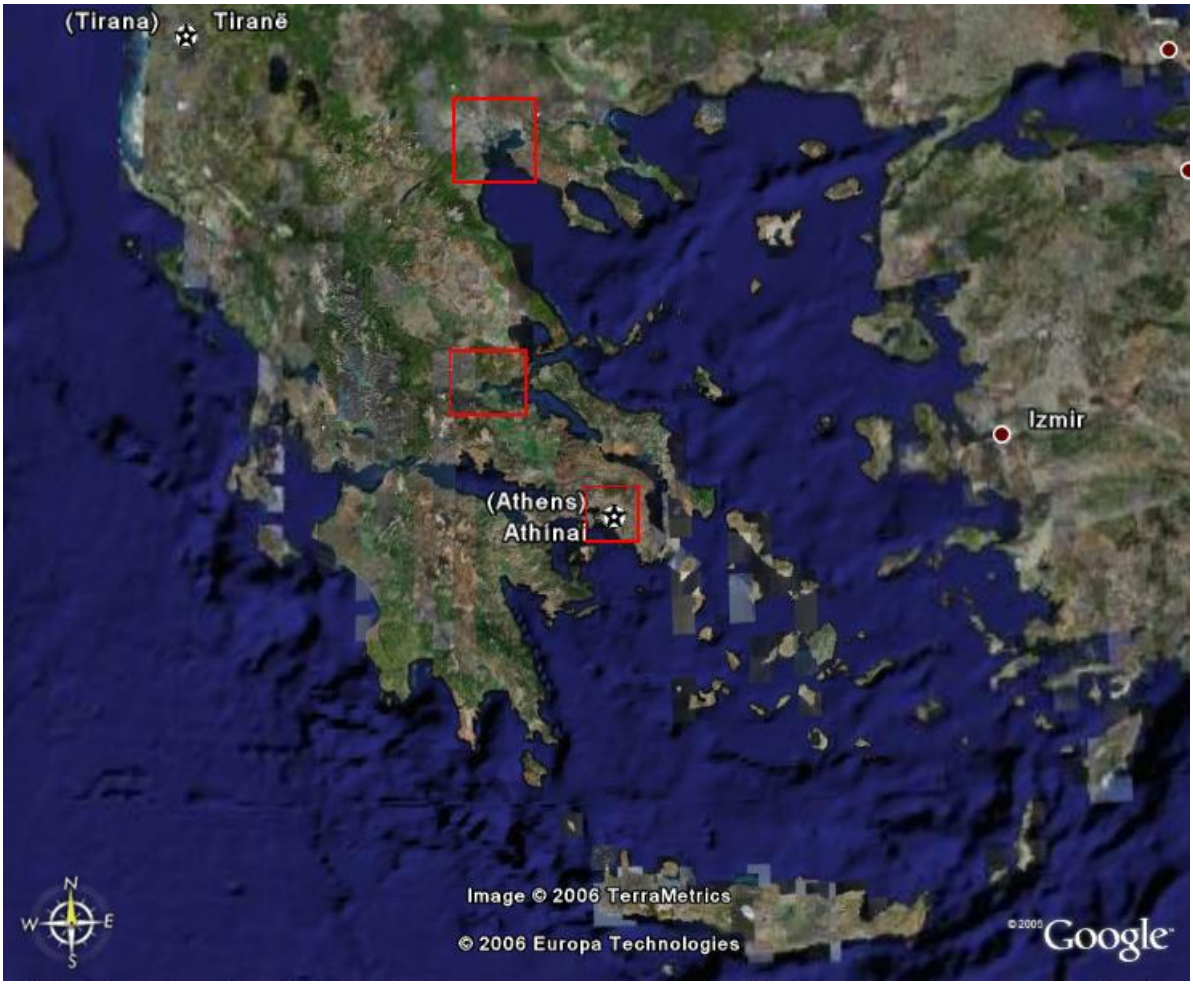
- Και τις τοποθέτησε ως εξής στον χάρτη:



Σχήμα 24: Οι επιμέρους τοποθεσίες της ιστοσελίδας

Παρατηρούμε πάλι ορισμένα outliers, τα οποία όμως απορρίπτονται στην επόμενη φάση.

- Έπειτα από συνεκτίμηση των αποτελεσμάτων, προέκυψαν οι ακόλουθες θέσεις για την ιστοσελίδα:



Σχήμα 25: Οι θέσεις που περιγράφονται στην ιστοσελίδα

7 Επίλογος...

7.1 Σύνοψη και συμπεράσματα

Για την παρουσίαση μιας ολοκληρωμένης πρότασης γεωκωδικοποίησης, δημιουργήσαμε ένα σύστημα που αποδίδει σε μια ιστοσελίδα γεωγραφικές συντεταγμένες σύμφωνα με το περιεχόμενό της (π.χ. αν αναφέρεται στην ακρόπολη, να αποκτήσει τις συντεταγμένες του μνημείου). Εκμεταλλευτήκαμε διάφορα στοιχεία από τη σελίδα, όπως ταχυδρομικούς κώδικες, τηλέφωνα, διευθύνσεις, τοπωνύμια, τη διεύθυνση IP του εξυπηρετητή. Επικεντρωθήκαμε στον ελληνικό ιστοχώρο, ο οποίος αφ'ενός έχει έλλειψη τέτοιων υπηρεσιών, αφ'ετέρου παρουσιάζει επιπλέον ενδιαφέροντα προβλήματα .

Για την υποστήριξη της προσπάθειας αυτής, προτείναμε νέες παραλλαγές αλγορίθμων προσεγγιστικής και φωνητικής αναζήτησης και ταιριάγματος, τις οποίες υλοποιήσαμε προσαρμοσμένες στα ελληνικά δεδομένα. Οι αλγόριθμοι αυτοί παρέχουν ποιοτικά αποτελέσματα σε μικρό χρόνο.

Από τα αποτελέσματα που προέκυψαν, φαίνεται πως είναι εφικτή η γεωγραφική ευρετηριοποίηση του ελληνικού ιστού, και άρα η δημιουργία εφαρμογών που βασίζονται στην γεωγραφική του τοπολογία (μηχανές γεωγραφικής αναζήτησης, εναλλακτικοί τρόποι πλοήγησης, παροχή νέων υπηρεσιών, κ.ο.κ.).

7.2 Μελλοντικές επεκτάσεις

Έχοντας δει, πλέον, το εύρος και τα αποτελέσματα της διπλωματικής αυτής εργασίας, ας αναφερθούμε και σε δυνατές μελλοντικές της επεκτάσεις.

7.2.1 Τεχνικές Επεκτάσεις

Κατ'αρχήν, υπάρχουν ορισμένες απλές, και καθαρά τεχνικές επεκτάσεις, όπως η δημιουργία μιας γεωγραφικής μηχανής αναζήτησης, με οπτικοποίηση των αποτελεσμάτων. Μια τέτοια επέκταση θα μπορούσε να χρησιμοποιήσει την (έτοιμη) μηχανή αρχειοθέτησης ανοιχτού κώδικα Apache Lucene[57], ή την μηχανή αρχειοθέτησης του Google Desktop (που έχει μια υποτυπώδη διαπροσωπεία σε Java, αλλά θα υποστηρίζει στο άμεσο μέλλον την Ελληνική γλώσσα), για αναζητήσεις με λέξεις-κλειδιά, και το δημοφιλές API του Google Earth/ Google Maps για την οπτικοποίηση. Μια ιδέα σε αυτήν την περίπτωση θα ήταν οι λέξεις κλειδιά να ελέγχονται για γεωγραφική πληροφορία (π.χ. αν είναι τοπωνύμια), και σε αυτήν την περίπτωση τα αποτελέσματα της αναζήτησης να ιεραρχούνται και με βάση την γεωγραφική εγγύτητα στα τοπωνύμια αυτά.

Ακόμη, το προϊόν της εργασίας αυτής θα μπορούσε να συζευχθεί με κάποιον έτοιμο web crawler (βλ. κεφάλαιο 2) για την συνεχή ανανέωση του ευρετηρίου του.

Επίσης, τμήματα αυτής της εργασίας μπορούν να χρησιμοποιηθούν και σε εμπορικές εφαρμογές, όπως για παράδειγμα σε μεσιτικά γραφεία (για την εύρεση, σε κατάλληλες ιστοσελίδες, αγγελιών αγοράς/πώλησης/ενοικίασης σπιτιών, την γεωκωδικοποίηση των οικείων διευθύνσεων, και την παρουσίαση των αποτελεσμάτων στον μεσίτη).

Οι επεκτάσεις αυτές είναι, όπως προαναφέρθηκε, καθαρά τεχνικής φύσης. Εντούτοις, η παρούσα εργασία θα μπορούσε να επεκταθεί και με άλλους, ουσιωδέστερους, τρόπους:

7.2.2 Εκμετάλλευση γεωγραφικής και χρονικής πληροφορίας

Για την εκμετάλλευση της χρονικής πληροφορίας που συχνά συνοδεύει την γεωγραφική (για παράδειγμα σε ειδησεογραφικές ιστοσελίδες), θα μπορούσε να γίνει εξαγωγή χρονικής πληροφορίας παράλληλα με την γεωγραφική, και η δημιουργία μιας γεω-χρονικής μηχανής αναζήτησης. Άλλωστε, η εξαγωγή χρονικής πληροφορίας αφ'εαυτής δεν ενέχει κάποια ιδιαίτερη δυσκολία, αφού μπορεί να υλοποιηθεί μέσω απλών γραμματικών (π.χ. "13 Ιουλίου 2006" ή "χθες"[58]).

[57]http://lucene.apache.org/

[58]Στην περίπτωση αυτή απαιτείται συσχετισμός με την ημερομηνία τελευταίας τροποποίησης του εγγράφου που δίνεται από τον εξυπηρετητή, δηλ. αν χθες & (lastModified = 13/6/2006)→12/6/2006

7.2.3 Γεωγραφική και θεματική κατηγοριοποίηση

Μια άλλη διάσταση του παγκόσμιου ιστού που έχει μεν εξερευνηθεί, αλλά θα είχε ενδιαφέρον η εξέτασή σε παράλληλία με την γεωγραφική είναι η θεματική. Συνδυάζοντας το παρόν σύστημα με κάποιο σύστημα θεματικής κατηγοριοποίησης του ιστού, (π.χ. το σύστημα THESUS ([HNV+03]) , ή βλ. [MNR+99]), θα μπορούσε να δημιουργηθεί μια μηχανή γεω-θεματικής αναζήτησης. Έτσι, θα ήταν δυνατές ερωτήσεις του τύπου "ψάξε για σελίδες με τις λέξεις "Οθέλλος", στην κατηγορία "Τέχνη, Θέατρο", στην ευρύτερη περιοχή του Ζωγράφου", φέρνοντας την αναζήτηση και πλοήγηση στον παγκόσμιο ιστό ακόμη πιο κοντά στο χρήστη.

7.2.4 Εξαγωγή γεωγραφικής πληροφορίας από το "βαθύ ιστό" (deep web geocrawling)

Ενδιαφέρον θα είχε η εξαγωγή γεωγραφικής πληροφορίας από το λεγόμενο "βαθύ ιστό" (deep web), δηλαδή τις ιστοσελίδες που δεν είναι αποθηκευμένες κάπου, αλλά δημιουργούνται σαν απόκριση σε κάποιο ερώτημα από έναν εξυπηρετητή. Ο βαθύς ιστός έχει μελετηθεί ενδελεχώς (π.χ. [RM01], [Ber01]), και σύμφωνα με το τελευταίο το μέγεθός του είναι μεγαλύτερο από αυτό του "επιφανειακού" ιστού κατά ένα παράγοντα μεγαλύτερο του 400. Αυτό, σε συνδυασμό με το ότι είναι σαφώς πιο δομημένος από τον "επιφανειακό" ιστό[59], και το ότι περιέχει, εν δυνάμει, πληθώρα γεωγραφικής πληροφορίας, δημιουργεί κατάλληλες συνθήκες για την αναζήτηση και εξαγωγή γεωγραφικής πληροφορίας από το βαθύ ιστό.

7.2.5 Εξαγωγή γεωγραφικής πληροφορίας από ειδικές κατηγορίες ιστοσελίδων

Τα weblogs, ή blogs, είναι μια σχετικά νέα τάση στο διαδίκτυο, όπου ο κάθε δικτυακός πολίτης μπορεί να μετατραπεί σε επίδοξο δημοσιογράφο, εκφέροντας την προσωπική του άποψη επί παντός επιστητού. Γεγονός είναι ότι συχνά περιέχουν γεωγραφικές αναφορές (π.χ. κατά τον σχολιασμό γεγονότων που έγιναν "κάπου"), κάνοντάς τα πρόσφορο έδαφος για αναζήτηση γεωγραφικής πληροφορίας. Σαφώς και υπάρχουν άλλες τέτοιες ειδικές κατηγορίες ιστοσελίδων (για παράδειγμα, οι ειδησεογραφικές, όπως έχουμε ήδη αναφέρει. βλ. και [Exp06]).

[59]Αρκεί να σκεφτούμε ότι μια τυπική σελίδα του "βαθέως ιστού" αποτελείται από μια λίστα αποτελεσμάτων κάποιου ερωτήματος, επομένως παρουσιάζει μεγάλη δομική και σημασιολογική κανονικότητα. Αυτήν την κανονικότητα μπορούμε να εκμεταλλευτούμε με διάφορες τεχνικές (π.χ. μηχανική μάθηση ([CHL03],[Exp06]), μοντέλα Markov ([BDS01]), ή και κάποια μετρική αυτοσυσχέτισης), ώστε να εξαγάγουμε αποτελεσματικά πληροφορία από την σελίδα.

8 *Βιβλιογραφία*

And99, Γιάννης Κ. Ανδρουτσόπουλος, Λατινο-ελληνική ορθογραφία στο ηλεκτρονικό ταχυδρομείο:χρήση και στάσεις, 20η Συνάντηση Εργασίας του Τομέα Γλωσσολογίας,Αριστοτέλειο Παν/μιο Θεσσαλονίκης, 1999

Arv99, Arvaniti A., Illustrations of the IPA: Modern Greek, Journal of the International Phonetic Association, vol 19, pages 167-172, 1999

BCG+99, Orkut Buyukkokten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, Narayan, Exploiting Geographical Location Information of Web Pages., WebDB (Informal Proceedings), pages 91-96, 1999

BDS01, Vinayak Borkar, Kaustubh Deshmukh, Sunita Sarawagi, Automatic segmentation of text into structured records, SIGMOD, 2001

Ber01, Bergman M., The deep Web: Surfacing Hidden Value, The Journal of Electronic Publishing from the University of Michigan, 2001

BLB+03, Karla A. V. Borges,Alberto H. F. Laender,Claudia Bauzer Medeiros ,Altigran, The Web as a Data Source for Spatial Databases., GeoInfo, 2003

BP98, Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks vol 30(1-7), pages 107-117, 1998

Bri92, E. Brill, A Simple Rule-based Part of Speech Tagger, Third Conference on Applied Natural Language Processing, 1992

CH03, Tara Calishain, Kevin Hemenway, Spidering Hacks, O'Reilly, ISBN 0-596-00577-6, 2003

CHL03, Chia-Hui Chang, Chun-Nan Hsu, Shao-Chen Lui, Automatic information extraction from semi-structured Web pages by pattern discovery., Decision Support Systems, vol 35(1), pages 129-147, 2003

Cir04, Fabio Ciravegna, Armadillo: harvesting information for the semantic web., SIGIR, page 598, 2004

Clo05, Paul Clough, Extracting metadata for spatially-aware information retrieval, GIR, pages 25-30, 2005

CMB+02, H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, ACL, 2002

CMB+06, Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Cristian Ursu, Marin Dimitrov, Mike Dowman, Niraj Aswani, Ian Roberts, Developing Language Processing Components with GATE, http://gate.ac.uk/sale/tao/, ίσχυε την 3/7/2006 2006

CTR+04, A. Chalamandaris, P. Tsiakoulis, S. Raptis, G. Giannopoulos and G. Carayannis, Bypassing Greeklish!, LREC (Fourth International Conference on Language Resources and Evaluation), pages 275-278, 2004

DGS00, Junyan Ding,Luis Gravano,Narayanan Shivakumar, Computing Geographical Scopes of Web Resources., VLDB, pages 545-556, 2000

Exp06, Explore Our Pla.Net, RSS to GeoRSS converter, found on web 15/6/2006 , http://exploreourpla.net/2006-06-08/georss-feed-reader-shows-podcasts.html , 2006

FFM05, Ariel Fuxman, Elham Fazli, Renée J. Miller, ConQuer: Efficient Management of Inconsistent Databases., SIGMOD, pages 155-166, 2005

Fou03, Harry Foundalis, The Details of Modern Greek Phonetics and Phonology, http://www.cogsci.indiana.edu/farg/harry/lan/grphdetl.htm, ίσχυε την 29/6/2006, 2003

GFS+01, Helena Galhardas, Daniela Florescu,Dennis Shasha, Eric Simon, Cristian-Augustin Saita, Declarative Data Cleaning: Language, Model and Algorithms, VLDB, 2001

GH05, Otis Gospodnetic,Erik Hatcher, Lucene in Action, Manning, page 138, 2005

GHL03, Luis Gravano, Vasileios Hatzivassiloglou, Richard Lichtenstein, Categorizing web queries according to geographical locality., CIKM, pages 325-333, 2003

GIJ+01, Luis Gravano, Panagiotis G. Ipeirotis, H. V. Jagadish, Nick Koudas, S. Muthukrishnan, Divesh Srivastava, Approximate String Joins in a Database (Almost) for Free, VLDB, 2001

Gil00, Michael Gilleland, Levenshtein Distance, in Three Flavors, http://www.merriampark.com/ld.htm, ίσχυε την 29/6/2006, 2000

Hea02, Jeff Heaton, Programming Spiders, Bots, and Aggregators in Java, Sybex, ISBN 0782140408, 2002

HNV+03, Maria Halkidi, Benjamin Nguyen, Iraklis Varlamis,Michalis Vazirgiannis, THESUS: Organizing Web document collections based on link semantics., VLDB J., pages 320-332, 2003

HS98, M.A. Hernández, S.J. Stolfo, Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem, Data Mining and Knowledge Discovery, 1998

JPR+02, Christopher B. Jones,Ross Purves,Anne Ruas,Mark Sanderson,Monika Sester,Mar, Spatial information retrieval and geographical ontologies an overview of th, SIGIR, pages 387-388, 2002

KSS+03, Vangelis Karkaletsis, Constantine D. Spyropoulos,Dimitris Souflis,Claire Gr, Demonstration of the CROSSMARC System., HLT-NAACL, 2003

LEJ+04, Charalampos Lampos, Magdalini Eirinaki, Darija Jevtuchova,Michalis Vazirgiannis, Archiving the Greek Web, IWAW, 2004

LHL01, Tim Berners-Lee, James Hendler,Ora Lassila, The Semantic Web, Scientific American, 2001

LLL+99, Mong Li Lee, Hongjun Lu, Tok Wang Ling, Yee Teng Ko, Cleansing Data for Mining and Warehousing, Lecture Notes In Computer Science, 1999

LLL00, M.L. Lee, T.W. Ling, W.L. Low, IntelliClean: A Knowledge-Based Intelligent Data Cleaner, KDD, 2000

LR93, A.J. Lait, B.Randell, An Assessment of Name Matching Algorithms, Technical Report, Dept. of Comp.Sci., University of Newcastle upon Tyne , 1993

LST+03, R. Lee, H. Shiina, H. Takakura, Y.J. Kwon, Y. Kambayashi, Optimization of Geographic Area to a Web Page for Two-Dimensional Range Query Processing, Web Information Systems Engineering Workshops, 2003

MAH+03, Yasuhiko Morimoto,Masaki Aono, Michael E. Houle,Kevin S. McCurley, Extracting Spatial Knowledge from the Web., SAINT, pages 326-333, 2003

Mcc01, Kevin S. McCurley, Geospatial mapping and navigation of the web., WWW, pages 221-229, 2001

ME96, A. E. Monge, C. Elkan, The field matching problem: Algorithms and applications, KDD, pages 267–270, 1996

MNR+99, Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore, A Machine Learning Approach to Building Domain-Specific Search Engines., IJCAI, pages 662-667, 1999

Mon00, A. E. Monge, Matching algorithms withing a Duplicate Detection System, IEEE Data Engineering Bulletin, 2000

PPK+99, Georgios Petasis, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos, Ion Androutsopoulos, Resolving Part-Of-Speech Ambiguity in the Greek Language Using Learning Techniques, CoRR, 1999

RH00, Rahm E., Do H.H, Data Cleaning: Problems and Current Approaches, IEEE Bulletin on Data Engineering, vol 23(4), 2000

RM01, Sriram Raghavan, Hector Garcia-Molina, Crawling the Hidden Web, VLDB, pages 129-138, 2001

Sar02, Sunita Sarawagi, Automation in Information Extraction and Integration, VLDB, 2002

SCY+04, Dou Shen, Zheng Chen, Qiang Yang,Hua-Jun Zeng,Benyu Zhang,Yuchang Lu,Wei-Yi, Web-page Classification through Summarization, SIGIR, pages 242-249, 2004

SFK98, K. Sgarbas, N.Fakotakis, G.Kokkinakis, A PC-KIMMO-Based Bi-directional Graphemic/Phonetic Converter for Modern Greek, Literary & Linguistic Computing, Oxford University Press, vol 13(2), pages 65-75, 1998

Syn04, Ιωάννης Συγγρός, Μετασχηματισμοί Συντεταγμένων των Γεωγραφικών Δεδομένων στον Ελληνικό Χώρο, HellasGI, 2004

Vaz05, Michalis Vazirgiannis, Introduction to link analysis &Temporal/Trend extensions of Pagerank, http://pages.cs.aueb.gr/sdep/slides/slides_vazirgiannis_11_01_2005.pdf, ίσχυε την 3/7/2006, 2005

Vei88, Βεης, Γ., Το χρησιμοποιούμενο πλέον σήμερα Ελληνικό Datum (ΕΓΣΑ87) , Δελτίο ΠΑ.Σ.Δ.Α.Τ.Μ., τεύχος 80, 1988

WA03, Carolyn R. Watters, Ghada Amoudi, Geosearcher: Location-based Ranking of Search Engine Results., JASIST, vol 54(2), pages 140-151, 2003

Wis01, Simon Wistow, IP2LL: Ways of finding Location from IP Address , http://www.thegestalt.org/simon/ip2ll/, ίσχυε την 3/7/2006, 2001

WP94, Allison Woodruff, Christian Plaunt, GIPSY: Automated Geographic Indexing of Text Documents., JASIS, pages 645-655, 1994