Αναζήτηση και Ανάκτηση Πληροφοριών από το Διαδίκτυο - Εφαρμογή στην Εύρεση Αγγελιών Θέσεων Εργασίας

Τσιρογιάννης Δημήτρης, Α.Μ. 1169
E-mail: stud1169@di.uoa.gr

Τμήμα Πληροφορικής & Τηλ/ών

Περίληψη Το πρόβλημα της αναζήτησης πληροφοριών στο διαδίκτυο απασχολεί εδώ και αρκετό καιρό την επιστημονική κοινότητα. Η ανομοιογένεια που παρουσιάζει το διαδικτυακό περιβάλλον μαζί με τον μεγάλο όγκο δεδομένων, ανάγει την εύρεση συγκεκριμένης πληροφορίας σε μία ιδιαίτερα πολύπλοκη και δύσκολη διαδικασία. Στην παρούσα εργασία, αρχικά μελετήθηκαν υπάρχουσες τεχνικές για αναζήτηση πληροφοριών από το διαδίκτυο, καθώς και για την εξαγωγή τους από ιστοσελίδες. Το αποτέλεσμα ήταν ο σχεδιασμός ενός συστήματος ανάκτησης πληροφοριών, με χρήση τεχνικών μηχανικής μάθησης και ευριστικών για την εξαγωγή πληροφοριών και το φιλτράρισμα των αποτελεσμάτων. Σαν υλοποίηση, αναπτύχθηκε ένα σύστημα εύρεσης αγγελιών για θέσεις εργασίας, μέσα από ιστοσελίδες.[1]

1 Εισαγωγή

Η πλειοψηφία των ανθρώπων που χρησιμοποιεί το διαδίκτυο, έχουν βρεθεί αντιμέτωποι με το πρόβλημα της αναζήτησης συγκεκριμένου τύπου πληροφοριών. Το σύνηθες εργαλείο είναι οι μηχανές αναζήτησης, οι οποίες μέσω των κριτηρίων που εισάγει ο εκάστοτε χρήστης επιστρέφουν στη πλειοψηφία των περιπτώσεων, ένα αρκετά μεγάλο πλήθος αποτελεσμάτων, πολλά από τα οποία δεν αναφέρονται σε αυτό το οποίο είχε αρχικά υπόψιν. Ακόμα και αν τα αποτελέσματα είναι σχετικά με την υπό αναζήτηση θεματική ενότητα, το αποτέλεσμα της αναζήτησης είναι ιστοσελίδες, τις οποίες πρέπει ο χρήστης να μελετήσει, ώστε να εντοπίσει την πληροφορία που τον ενδιαφέρει. Είναι εμφανής η ύπαρξη τριών αναγκών. Η πρώτη χαρακτηρίζει τον τρόπο με τον οποίο ο χρήστης εκφράζει την πληροφορία που τον ενδιαφέρει. Η συνήθης πρακτική της εισαγωγής λέξεων κλειδιών δεν είναι αποδεκτή, καθώς δεν χαρακτηρίζεται από σαφήνεια. Μια αρκετά καλή προσέγγιση θα ήταν μέσω περιγραφής σε ελεύθερο ή ημι-δομημένο κείμενο. Η ύπαρξη ενός αποδοτικού συστήματος ανάκτησης πληροφοριών, προσανατολισμένο στον εντοπισμό πληροφοριών σε συγκεκριμένες θεματικές περιοχές, αποτελεί την δεύτερη ανάγκη. Τέλος, η αυτόματη επεξεργασία των αποτελεσμάτων και η εξαγωγή μόνο των χρήσιμων πληροφοριών που περιέχουν, αποτελεί μια εξίσου σημαντική απαίτηση.

[1] Επιβλέπουσα Καθηγήτρια: Ι. Καράλη

Η συγκεκριμένη εργασία εστίασε στη διαδικασία εντοπισμού ιστοσελίδων που ανήκουν σε συγκεκριμένη θεματική περιοχή. Παράλληλα, αναπτύχθηκαν ευριστικές μέθοδοι, για το φιλτράρισμα των αποτελεσμάτων και την εξαγωγή πληροφοριών από τις ιστοσελίδες.

Η κατασκευή ενός συστήματος για την αναζήτηση συγκεκριμένων πληροφοριών από το διαδίκτυο, ανάγεται στην αντιμετώπιση δύο ουσιαστικών προβλημάτων. Το πρώτο πρόβλημα αναφέρεται στον τρόπο με τον οποίο θα γίνει η αναζήτηση μέσα στο διαδίκτυο, κυρίως λόγω του όγκου και της μεταβλητότητάς του [2]. Λαμβάνοντας υπόψιν τις δύο αυτές παραμέτρους, θα πρέπει το σύστημα να είναι ιδιαίτερα αποτελεσματικό, όσον αφορά το πλήθος των χρήσιμων ιστοσελίδων που εντοπίζει στη μονάδα του χρόνου. Υπάρχουν διάφορες προσεγγίσεις για την αντιμετώπιση του συγκεκριμένου προβλήματος, κυρίως από την τεχνητή νοημοσύνη, όπως αναζήτηση-κατά-πλάτος, αναζήτηση-κατά-βάθος και ευριστική αναζήτηση. Στη συγκεκριμένη εργασία χρησιμοποιήθηκε ένας συνδυασμός αναζήτησης-κατά-πλάτος και ευριστικής αναζήτησης. Οι ευριστικές μέθοδοι αναπτύχθηκαν εκμεταλλευόμενοι χαρακτηριστικά γνωρίσματα από τη περιοχή εφαρμογής, η οποία είναι η αγορά εργασίας. Το μειονέκτημα των μεθόδων ευριστικής αναζήτησης, είναι η έλλειψη γενικότητας και επεκτασιμότητας.

Η μέθοδος αναζήτησης που χρησιμοποιείται, προσδιορίζει το δρόμο που θα πρέπει να ακολουθήσει το σύστημα κατά την πλοήγηση στο διαδίκτυο, ώστε να εντοπίσει τη ζητούμενη πληροφορία. Αδυνατεί όμως να αναγνωρίσει στο προτεινόμενο μονοπάτι την ιστοσελίδα "στόχο", το οποίο και αποτελεί το δεύτερο πρόβλημα. Για την αντιμετώπιση αυτού του προβλήματος χρησιμοποιήθηκε μηχανική μάθηση. Συγκεκριμένα, με βάση τη *θεωρία εκμάθησης* του Bayes (Bayesian Learning) [1], κατασκευάστηκε ένας ταξινομητής, ο οποίος δύναται να αναγνωρίσει, αποκλειστικά με βάση το περιεχόμενο, αν μία ιστοσελίδα περιέχει χρήσιμες πληροφορίες ή όχι. Η απόδοση ενός συστήματος ταξινόμησης εξαρτάται από διάφορες επιλογές κατά τη φάση της εκπαίδευσης. Συγκεκριμένα, ουσιαστικό παράγοντα αποτελεί το πλήθος και η ποιότητα των ιστοσελίδων που θα χρησιμοποιηθούν για την εκπαίδευση, καθώς και των χαρακτηριστικών (features) που θα επιλεγούν.

Το σύστημα περιλαμβάνει ένα ευριστικό συστήμα εξαγωγής πληροφορίας. Από το καθαρό κείμενο των ιστοσελίδων που έχουν εντοπισθεί, εξάγεται η πληροφορία που ενδιαφέρει το χρήστη (τίτλος εργασίας, απαιτούμενα προσόντα κ.α.) και παρουσιάζεται σε ημιδομημένη μορφή, μέσω XML αρχείων. Παράλληλα, χρησιμοποιήθηκαν ευριστικές για το ταίριασμα των αποτελεσμάτων με τα χαρακτηριστικά του χρήστη, αξιοποιώντας την μεθοδολογία αναπαράστασης της XML.

Αρχικές προσπάθειες χρησιμοποίησαν μεθόδους αξιολόγησης ιστοσελίδων ανεξάρτητες από θεματικές ενότητες, ώστε να μειώσουν το πλήθος των ιστοσελίδων που προσπελαύνουν οι αναζητητές [7]. Προσπάθειες κατασκευής προσανατολισμένων αναζητητών έγιναν από τους M.Diligenti [5] και J.Rennie [4]. Στο [5] η προσανατολισμένη αναζήτηση σε μία θεματική περιοχή, μοντελοποιείται μέσω γράφων συμφραζομένων και της κατασκευής ενός συνόλου ταξινομητών. Στο [4] χρησιμοποιείται ενισχυτική εκμάθηση (reinforcement learning) για την κατασκευή ενός προσανατολισμένου ταξινομητή, αντιμετωπίζοντας το πρόβλημα σαν μία απεικόνιση ενεργειών σε καταστάσεις, μέσω μιας συνάρτησης αμοιβών.

Η προσέγγιση που ακολουθήσαμε εντάσσεται στη περιοχή της προσανατολισμένης αναζήτησης, η οποία και περιγράφεται στο επόμενο κεφάλαιο. Ακολουθεί η περιγραφή της προσέγγισης που υλοποιήθηκε. Τέλος, παρουσιάζονται συνοπτικά τα συμπεράσματα από την εργασία.

2 Προσανατολισμένη Αναζήτηση (Focused Crawling)

Η ανάκτηση πληροφοριών από το διαδίκτυο (Web Information Retrieval) στηρίζεται κυρίως στις μηχανές αναζήτησης. Ένα από τα πλέον σημαντικά συστατικά μιας μηχανής αναζήτησης είναι οι αναζητητές (crawlers). Το έργο των αναζητητών είναι η προσπέλαση ιστοσελίδων από διακομιστές διαδικτύου (web servers) και η αποθήκευσή τους τοπικά για επεξεργασία. Σημαντικό χαρακτηριστικό των αναζητητών είναι η πολιτική προσπέλασης των ιστοσελίδων. Συγκεκριμένα, υπάρχουν δύο εναλλακτικές προσεγγίσεις, η πρώτα - κατά - πλάτος και η πρώτα - κατά - βάθος. Οι δύο αυτές προσεγγίσεις θεωρήθηκαν ικανοποιητικές στη πρώτη γενιά μηχανών αναζήτησης, όταν το κυριότερο μέλημά τους ήταν η όσο το δυνατόν μεγαλύτερη κάλυψη του διαδικτύου. Το πολύ μεγάλο πλήθος των ιστοσελίδων στη τρέχουσα μορφή του διαδικτύου, καθιστά απαγορευτική τη χρήση των συγκεκριμένων μεθόδων.

Μια άλλη προσέγγιση, είναι η κατασκευή συστημάτων ανάκτησης πληροφοριών προσανατολισμένων σε μία συγκεκριμένη θεματική περιοχή. Οι αναζητητές που υποστηρίζουν τη λειτουργία τέτοιων συστημάτων πρέπει να λειτουργούν με τελείως διαφορετική φιλοσοφία. Συγκεκριμένα, ο στόχος δεν είναι η κάλυψη όσο το δυνατόν μεγαλύτερου πλήθους σελίδων, αλλά ο εντοπισμός του μέγιστου αριθμού σελίδων με σχετικό περιεχόμενο, ενώ ταυτόχρονα γίνεται προσπέλαση του ελάχιστου αριθμού άσχετων σελίδων. Οι αναζητητές εντοπίζουν σελίδες που ανήκουν σε μία συγκεκριμένη θεματική ενότητα και καθοδηγούν την αναζήτηση βασιζόμενοι στο περιεχόμενο των σελίδων και τη δομή των συνδέσμων στο διαδίκτυο. Συγκεκριμένα, υλοποιούν μια στρατηγική, η οποία συσχετίζει με ένα βαθμό τους συνδέσμους κάθε σελίδας που προσπελαύνει, ανάλογα με τη δυνατότητα υπόδειξης σχετικών ιστοσελίδων. Οι σύνδεσμοι ταξινομούνται με βάση το βαθμό τους και εισάγονται σε μια ουρά. Στη συνέχεια, υλοποιείται μια μέθοδος (πρώτα - ο - καλύτερος) παίρνοντας τον καλύτερο σύνδεσμο από την ουρά. Αυτή η στρατηγική εξασφαλίζει σε μεγάλο βαθμό ότι ο αναζητητής προτιμάει τα πλέον υποσχόμενα μονοπάτια για τον εντοπισμό των ζητούμενων σελίδων. Το πρόβλημα έγκειται στην επιλογή του κατάλληλου συστήματος βαθμολόγησης των συνδέσμων.

Το κύριο πρόβλημα που αντιμετωπίζουν οι προσανατολισμένοι αναζητητές, είναι η δυσκολία τους να μάθουν πως από ένα σύνολο από, άσχετες με το περιεχόμενο, σελίδες μπορούν να οδηγηθούν σε κείμενα με υψηλό βαθμό σχετικότητας. Το πρόβλημα έγκειται στο γεγονός ότι δεν υπάρχει πληροφορία η οποία να καθορίζει πότε μια ιστοσελίδα προσπελαύνεται επειδή έχει σχετικό περιεχόμενο ή λόγω της θέσης της σε ένα δικτυακό τόπο. Το γεγονός ότι οι σύνδεσμοι είναι μονής κατεύθυνσης, αυξάνει τη δυσκολία κατασκευής ευέλικτων προσανατολισμένων αναζητητών.

Στη συνέχεια, περιγράφονται δύο από τις πλέον ενδιαφέρουσες προσεγγίσεις στη κατασκευή προσανατολισμένων αναζητητών.

Προσανατολισμένη Αναζήτηση με χρήση Γράφων Συμφραζομένων (Focused Crawling Using Context Graphs).

Η συγκεκριμένη προσέγγιση ενισχύει το σύστημα βαθμολόγησης των συνδέσμων εφοδιάζοντας τον αναζητητή με τη δυνατότητα να μοντελοποιεί τα συμφραζόμενα μέσα από τα οποία εντοπίζονται θεματικές ενότητες στο διαδίκτυο [5]. Τέτοιο μοντέλο συμφραζομένων προσπαθεί να καθορίσει ιεραρχίες συνδέσμων, μέσα από τις οποίες εντοπίζονται οι ενδιαφέρουσες σελίδες καθώς και να περιγράψει άσχετο θεματικά περιεχόμενο το οποίο όμως σχετίζεται σε μεγάλο βαθμό με τις σελίδες αυτές. Ο προσανατολισμένος αναζητητής, ο οποίος καλείται και CFC (Context Focused Crawler) χρησιμοποιεί τις δυνατότητες κάποιων μηχανών αναζήτησης όπως η Google [2], παρέχοντας τη δυνατότητα εντοπισμού ιστοσελίδων οι οποίες δείχνουν σε μία συγκεκριμένη. Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί για τη κατασκευή μιας σημασιολογικής αναπαράστασης των σελίδων που βρίσκονται συγκεκριμένο αριθμό βημάτων μακριά από μία ιστοσελίδα στόχο. Η αναπαράσταση χρησιμοποιείται για την εκπαίδευση ενός συνόλου από ταξινομητές, οι οποίοι βελτιστοποιούνται στον εντοπισμό και την κατάταξη κειμένων σε κατηγορίες, οι οποίες καθορίζονται με βάση την απόσταση που έχουν από ιστοσελίδες στόχους. Δηλαδή, κατασκευάζεται ένας ταξινομητής για σελίδες που απέχουν ένα βήμα από μια σελίδα στόχο, ένας για σελίδες με απόσταση δύο βημάτων κ.ο.κ. Το πλήθος των κατηγοριών και κατά συνέπεια των ταξινομητών εξαρτάται από την εφαρμογή.

Κατά τη διάρκεια της αναζήτησης οι ταξινομητές χρησιμοποιούνται για να προβλέψουν την απόσταση ενός κειμένου στόχου από το συγκεκριμένο κείμενο το οποίο εξετάζουν τη δεδομένη χρονική στιγμή, ανάλογα με την κατηγορία στην οποία κατατάσσεται. Για παράδειγμα, αν οι ταξινομητές κατατάξουν τη σελίδα στη πρώτη κατηγορία, τότε θα μπορούμε να πούμε με βεβαιότητα ότι σε απόσταση ενός βήματος θα εντοπίσουμε ιστοσελίδα με ζητούμενο περιεχόμενο. Με τον τρόπο αυτό βελτιστοποιείται η αναζήτηση, εξετάζοντας πρώτα μονοπάτια που ενδέχεται να οδηγήσουν άμεσα σε ενδιαφέρουσες ιστοσελίδες [5].

Ένα σοβαρό μειονέκτημα του αλγορίθμου είναι η απαίτηση για ύπαρξη αντίστροφων συνδέσμων σε μια μηχανή αναζήτησης, για την κατασκευή των γράφων συμφραζομένων και των ταξινομητών.

Προσανατολισμένη Αναζήτηση με Ενισχυτική Εκμάθηση (Focused Crawling with Reinforcement Learning).

Η ενισχυτική εκμάθηση (Reinforcement Learning) αναφέρεται στη βιβλιογραφία, από τη περιοχή της μηχανικής μάθησης [1], σαν την εκμάθηση της διαδικασίας βέλτιστης λήψης αποφάσεων μέσω ποινών ή αμοιβών.

Στη περίπτωση της προσανατολισμένης αναζήτησης, οι ιστοσελίδες που έχουν το θέμα που μας ενδιαφέρει είναι οι άμεσες αμοιβές. Οι ενέργειες είναι η επιλογή των συνδέσμων. Το πλήθος των καταστάσεων είναι πολύ μεγάλο και συγκεκριμένα 2^N, όπου N το πλήθος των ιστοσελίδων με σχετικό περιεχόμενο. Το πλήθος των ενεργειών είναι εξίσου μεγάλο και ισοδυναμεί με το πλήθος των συνδέσμων (URLs). Στόχος της μεθόδου είναι η εκμάθηση μιας συνάρτησης

[2] http://www.google.com

αμοιβών, ώστε σε δεδομένη κατάσταση (ιστοσελίδα) να επιλέγεται η ενέργεια (σύνδεσμος) με τη μεγαλύτερη αμοιβή. Η μέθοδος παρέχει τη δυνατότητα μοντελοποίησης της διαδικασίας εύρεσης σχετικών σελίδων, μέσα από φαινομενικά άσχετα μονοπάτια.
Το μειονέκτημα της χρήσης ενισχυτικής μάθησης για προσανατολισμένη αναζήτηση, είναι ότι απαιτεί από το χρήστη να ορίσει αντιπροσωπευτικές ιστοσελίδες, τις οποίες πρέπει να χρησιμοποιήσει ο αναζητητής για την εκμάθηση της συνάρτησης αμοιβών [4].

3 Σύστημα Εύρεσης Αγγελιών για Θέσεις Εργασίας (JobFinder)

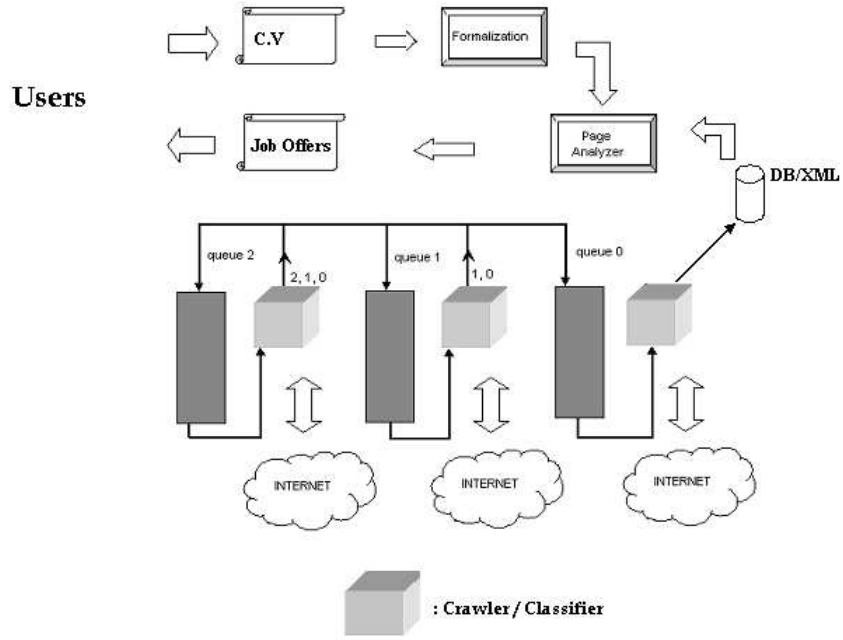
Στα πλαίσια της εργασίας αναπτύχθηκε ένα σύστημα ανάκτησης πληροφοριών, προσανατολισμένο στην εύρεση αγγελιών για θέσεις εργασίας. Παράλληλα, αναπτύχθηκαν επιμέρους υποσυστήματα για την εξαγωγή πληροφοριών από ιστοσελίδες, καθώς και για το φιλτράρισμα των αποτελεσμάτων, με βάση τα κριτήρια των χρηστών. Παρέχεται η δυνατότητα στους χρήστες που το χρησιμοποιούν, να εισάγουν τα κριτήρια αναζήτησης, συμπληρώνοντας μία φόρμα με στοιχεία από το βιογραφικό τους. Το σύστημα αρχικά αναλαμβάνει να εντοπίσει αγγελίες από το διαδίκτυο, εξάγει τις πληροφορίες που περιέχουν χρησιμοποιώντας το υποσύστημα *Αναλυτής Σελίδων* και τις αποθηκεύει τοπικά με τη μορφή XML αρχείων. Οι χρήστες εισάγουν τα κριτήρια αναζήτησης (στοιχεία βιογραφικού) και το υποσύστημα *Αρχικοποίησης* επιστρέφει τις αγγελίες που ταιριάζουν με τα κριτήριά τους.

Στη συνέχεια περιγράφεται η αρχιτεκτονική του συστήματος, παρουσιάζοντας τα τμήματα από τα οποία αποτελείται και τις συσχετίσεις μεταξύ τους, όπως φαίνεται και στο σχήμα 1.

Το σύστημα αποτελείται από τα εξής συστατικά:

- *Αναζητητής Διαδικτύου (Web Crawler).* Εκτελεί την αναζήτηση και την προσωρινή αποθήκευση για επεξεργασία, των ιστοσελίδων στο δίκτυο.
- *Ταξινομητής Κειμένου (Text Classifier).* Χρησιμοποιείται για τον εντοπισμό των σελίδων, οι οποίες περιέχουν αγγελίες με θέσεις εργασίας.
- *Σύστημα Αρχικοποίησης (Formalization System).* Αποτελεί το περιβάλλον, το οποίο δέχεται τις παραμέτρους αναζήτησης από κάθε χρήστη και τις επεξεργάζεται για το αποτελεσματικό φιλτράρισμα των αποτελεσμάτων.
- *Αναλυτής Σελίδων (Page Analyzer).* Είναι το σύστημα για την εξαγωγή χρήσιμων πληροφοριών από τις ιστοσελίδες.

Τα πλέον σημαντικά συστατικά του συστήματος είναι ο αναζητητής διαδικτύου και ο ταξινομητής, οι οποίοι αναλαμβάνουν τον κύριο όγκο επεξεργασίας. Υπάρχουν τρεις αναζητητές διαδικτύου, οι οποίοι λειτουργούν παράλληλα, κατεβάζοντας ιστοσελίδες από το διαδίκτυο, αυξάνοντας με τον τρόπο αυτό την απόδοση του συστήματος. Ο κάθε αναζητητής χρησιμοποιεί μία συγκεκριμένη ουρά, από την οποία παίρνει σαν είσοδο τους συνδέσμους που περιέχει. Κάθε αναζητητής συνδέεται με ένα διαφορετικό στιγμιότυπο ταξινομητή. Η λογική χρήσης των τριών αναζητητών



Σχήμα 1. Η αρχιτεκτονική του συστήματος ανάκτησης πληροφοριών.

με τις αντίστοιχες ουρές έγκειται στον αλγόριθμο προσανατολισμένης αναζήτησης που χρησιμοποιήθηκε. Συγκεκριμένα, αναπτύχθηκε ένας αλγόριθμος, ο οποίος αποτελεί ένα συνδυασμό της πρώτα - κατά - πλάτος προσέγγισης και μίας ευριστικής μεθόδου, κατάλληλης για τη θεματική περιοχή της αγοράς εργασίας. Συγκεκριμένα, εξετάζοντας τις σελίδες στο διαδίκτυο, όταν η ευριστική μέθοδος θεωρήσει, με βάση τα κριτήρια που περιγράφονται σε επόμενη παράγραφο, ότι οι αναζητητές βρίσκονται κοντά σε ιστοσελίδα στόχο, εισάγει τους συνδέσμους σε διαφορετική ουρά. Οι ουρές δημιουργούν μια ιεραρχία σελίδων με βάση την απόστασή τους από ιστοσελίδες που ανήκουν στη συγκεκριμένη θεματική περιοχή. Κάθε αναζητητής αντιμετωπίζει διαφορετικά του συνδέσμους από κάθε ουρά, μετατρέποντας δυναμικά την διαδικασία αναζήτησης, από πλοήγηση μεταξύ δικτυακών τόπων (crawling) σε πλοήγηση εντός των δικτυακών τόπων (spidering) ακολουθώντας συγκεκριμένα μονοπάτια και μέχρι ενός συγκεκριμένου βάθους.

Κάθε σελίδα την οποία προσπελαύνει ένας αναζητητής, δίνεται σαν είσοδος σε ένα στιγμιότυπο ταξινομητή. Ο ρόλος του ταξινομητή είναι να κατηγοριοποιήσει τις σελίδες, ανάλογα με το περιεχόμενό τους, σε δύο κλάσεις. Στη πρώτη κλάση ανήκουν σελίδες που περιέχουν αγγελίες θέσεων εργασίας, ενώ στη δεύτερη όλες οι υπόλοιπες. Η κατασκευή του ταξινομητή αποτελεί μία χρονοβόρα διαδικασία, καθώς θα πρέπει να καθορισθεί ο αλγόριθμος που θα χρησιμοποιηθεί και στη συνέχεια

να συλλεχθεί ένα κατάλληλο σύνολο ιστοσελίδων, το οποίο θα χρησιμοποιηθεί για την εκπαίδευσή του.

Ιστοσελίδες οι οποίες περιέχουν αγγελίες θέσεων εργασίας υφίστανται επεξεργασία για την εξαγωγή της πληροφορίας που περιέχουν και η οποία ενδιαφέρει τον χρήστη. Συγκεκριμένα, ο *Αναλυτής Σελίδων* χρησιμοποιεί μια ευριστική μέθοδο εξαγωγής πληροφορίας. Χρησιμοποιεί λέξεις κλειδιά που περιέχονται σε τέτοιου είδους σελίδες για τον προσδιορισμό των ορίων κάθε πληροφορίας (τίτλος εργασίας, απαιτούμενες γνώσεις), ενώ παράλληλα εισάγει την εξαχθήσα πληροφορία από κάθε σελίδα σε XML αρχεία.

Το φιλτράρισμα των αποτελεσμάτων με βάση τα κριτήρια του χρήστη γίνεται από το *Σύστημα Αρχικοποίησης*. Συγκεκριμένα, το βιογραφικό που εισάγει ο χρήστης σε κατάλληλα διαμορφωμένη φόρμα, μετατρέπεται σε αρχείο XML. Στη συνέχεια, γίνεται σύγκριση με τα αρχεία XML που περιέχουν αγγελίες για θέσεις εργασίας, ώστε να εντοπισθούν αυτές που ταιριάζουν με τα χαρακτηριστικά του εκάστοτε χρήστη.

Αναζητητής Διαδικτύου (Web Crawler). Η αλγόριθμος αναζήτησης εκτελείται είτε μέχρι να αδειάσει το περιεχόμενο των ουρών είτε ένα συγκεκριμένο χρονικό διάστημα το οποίο ορίζεται από τον διαχειριστή του συστήματος, ικανό για την ανάκτηση ενός ικανοποιητικού αριθμού αγγελιών. Για κάθε σύνδεσμο που παίρνει ο αλγόριθμος μέσα από μία ουρά, ελέγχει αν έχει ήδη εξετασθεί κατά το παρελθόν, ώστε να μην διαγράφει κύκλους ο αλγόριθμος αναζήτησης. Για την αντιμετώπιση του προβλήματος αποθήκευσης των συνδέσμων, χρησιμοποιείται μια πεπερασμένη δομή, στην οποία αποθηκεύεται ο αρχικός σύνδεσμος από κάθε δικτυακό τόπο. Οι αναζητητές είναι υλοποιημένοι με τέτοιο τρόπο, ώστε να γνωρίζουν πότε διενεργούν προσπέλαση σελίδων εντός του ίδιου δικτυακού τόπου ώστε να μην αποθηκεύουν όλους τους συνδέσμους που εξετάζουν.

Οι αναζητητές προσπελαύνουν τις ιστοσελίδες από διακομιστές διαδικτύου. Η λειτουργία των διακομιστών δεν πρέπει να επηρεάζεται από τα συγκεκριμένα προγράμματα. Για το λόγο αυτό, για κάθε ιστοσελίδα που προσπελαύνεται από τον ίδιο διακομιστή, εισάγεται μια μικρή τεχνητή καθυστέρηση, ώστε να μην προκαλείται υπερφόρτωση.

Για τον εντοπισμό των συνδέσμων σε μία ιστοσελίδα αξιοποιούμε τις ιδιότητες τις γλώσσας HTML, στην οποία υπάρχουν ζευγάρια πεδίων (tags) για τη περιγραφή κάθε συστατικού της σελίδας (σύνδεσμοι, εικόνες, κώδικας). Η εξαγωγή των συνδέσμων ανάγεται στον εντοπισμό των κατάλληλων πεδίων εντός της ιστοσελίδας. Στη διαδικασία εξαγωγής των συνδέσμων, εξετάζοντας τα πεδία της γλώσσας HTML, δεν παίρνουμε μόνο το σύνδεσμο, αλλά και το κείμενο που τον περιγράφει (anchor text). Η συγκεκριμένη πληροφορία είναι ιδιαίτερα χρήσιμη, καθώς με βάση αυτή λειτουργεί η ευριστική μέθοδος προσανατολισμένης αναζήτησης. Συγκεκριμένα, ύστερα από μελέτη ενός συνόλου δικτυακών τόπων, διαπιστώθηκε ότι οι σύνδεσμοι που οδηγούν σε αγγελίες με θέσεις εργασίας, περιέχουν ένα συγκεκριμένο σύνολο λέξεων στο συνοδευτικό τους κείμενο. Επίσης, διαπιστώθηκε ότι η απόσταση μεταξύ της σελίδας που περιέχει τον σύνδεσμο και της σελίδας που περιέχει αγγελίες είναι το πολύ *πέντε* βήματα, εντός του ίδιου δικτυακού τόπου.

Εκμεταλλευόμενοι τη συγκεκριμένη πληροφορία, περιορίζουμε την αναζήτηση εντός του δικτυακού τόπου, μόνο σε σελίδες που προέρχονται από τον συνδέσμο με την προανεφερθείσα ιδιότητα. Οι ιστοσελίδες με αγγελίες για θέσεις εργασίας, κυρίως περιέχονται σε δικτυακούς τόπους εταιρειών. Κατά την πλοήγηση στο διαδίκτυο μεταξύ δικτυακών τόπων δεν έχουμε άλλου είδους πληροφορία, η οποία να υποδεικνύει την ύπαρξη σε κοντινή απόσταση, σελίδων με σχετικό περιεχόμενο. Ενδεχομένως, για κάποια άλλη θεματική περιοχή, να υπήρχε διαθέσιμη ευρύτερη πληροφορία. Σε μία τέτοια περίπτωση η ευριστική συνάρτηση θα ήταν σίγουρα διαφορετική ή δεν θα υπήρχε και καθόλου και η καθοδήγηση θα γίνονταν με άλλες μεθόδους όπως αυτές που περιγράφονται στο [4] και στο [5]. Για την συγκεκριμένη όμως περιοχή, οι μέθοδοι αυτές θα εισήγαγαν περιττή πολυπλοκότητα και καθυστέρηση, χωρίς να εξασφαλίζουν καλύτερα αποτελέσματα από τη μέθοδο που ακολουθήθηκε.

Ταξινόμηση Κειμένου (Text Classification). Σημαντικό τμήμα του συστήματος είναι η διαδικασία αναγνώρισης των σελίδων που περιέχουν σχετικό περιεχόμενο. Μια πρώτη προσέγγιση είναι η μελέτη συνδέσμων που δείχνουν σε τέτοιου είδους σελίδες και του συνοδευτικού τους κειμένου, προκειμένου να εντοπισθεί πληροφορία, η οποία θα καταδείκνυε την ύπαρξη του ζητούμενου περιεχομένου. Η συγκεκριμένη προσέγγιση όμως δεν είχε τα επιθυμητά αποτελέσματα.

Τελικά, χρησιμοποιήθηκε μία τεχνική από την μηχανική μάθηση ονομαζόμενη **κατηγοριοποίηση κειμένου (text classification)**, για την αναγνώριση των ζητούμενων ιστοσελίδων με βάση το περιεχόμενό τους. Υπάρχουν διάφοροι αλγόριθμοι στη περιοχή, αλλά αυτός ο οποίος χρησιμοποιείται λόγο της απόδοσής του είναι ο αλγόριθμος του Bays, Naive Bayes Text Classification [1]. Για την κατασκευή του ταξινομητή πρέπει να ορισθούν οι κατηγορίες (κλάσεις), να συλλεγούν κείμενα για την εκπαίδευση και να επιλεγούν τα χαρακτηριστικά (attributes), με βάση τα οποία θα κατασκευασθεί το μοντέλο του ταξινομητή για κάθε κλάση.

Για τη θεματική ενότητητα της αγοράς εργασίας θεωρήθηκαν δύο κλάσεις, η μία αναφέρεται στις σελίδες που περιέχουν σχετικό περιεχόμενο, ενώ η δεύτερη σε όλες τις υπόλοιπες. Σημαντική διαδικασία από την οποία εξαρτάται η απόδοση του ταξινομητή, είναι η επιλογή των ιστοσελίδων οι οποίες θα χρησιμοποιηθούν για εκπαίδευση. Οι ιστοσελίδες χωρίζονται σε δύο κατηγορίες, τα θετικά παραδείγματα, από τα οποία ο ταξινομητής θα μάθει να ξεχωρίζει ποιες ιστοσελίδες ανήκουν στη κλάση των σχετικών σελίδων και τα αρνητικά. Τα αρνητικά παραδείγματα δεν πρέπει να είναι τυχαίες σελίδες ασχέτου περιεχομένου, αλλά σελίδες που αναφέρονται στην ίδια θεματική ενότητα, χωρίς όμως να περιέχουν πληροφορίες που ενδιαφέρουν τους χρήστες. Όσο καλύτερη είναι η επιλογή των αρνητικών παραδειγμάτων, τόσο αυξάνει η διακριτότητα του ταξινομητή, ώστε να μπορεί με βεβαιότητα να ξεχωρίζει τις ζητούμενες σελίδες.

Έχοντας προσδιορίσει τα θετικά και αρνητικά παραδείγματα για τον ταξινομητή, θα πρέπει να επιλεγούν τα χαρακτηριστικά εκείνα με βάση το οποίο θα κατασκευασθεί το μοντέλο του ταξινομητή. Για την ταξινόμηση κειμένου, τα χαρακτηριστικά είναι λέξεις που υπάρχουν στις εν λόγω ιστοσελίδες. Για την επιλογή των χαρακτηριστικών χρησιμοποιήθηκε ο αλγόριθμος TF-IDF [6], ο οποίος με βάσει τις

συχνότητες εμφάνισης των λέξεων στις ιστοσελίδες εκπαίδευσης, επιλέγει αυτές που θεωρεί τις πλέον σημαντικές.

Έχοντας επιλέξει τα χαρακτηριστικά, δημιουργείται ένα ειδικά διαμορφωμένο αρχείο, το οποίο θα δωθεί σαν είσοδος για τη κατασκευή του μοντέλου του ταξινομητή. Το αρχείο αυτό αποτελείται από ένα σύνολο διανυσμάτων, ένα για κάθε ιστοσελίδα εκπαίδευσης. Τα διανύσματα είναι n θέσεων, όπου n το πλήθος των χαρακτηριστικών, ενώ σε κάθε θέση αναφέρεται η συχνότητα εμφάνισης της συγκεκριμένης λέξης στην ιστοσελίδα. Για τη περιοχή της αγοράς εργασίας χρησιμοποιήθηκαν 19 χαρακτηριστικά.

Για την ταξινόμηση κάθε σελίδας που προσπελαύνουν οι αναζητητές, πρέπει να γίνει μια προεπεξεργασία. Συγκεκριμένα, αφαιρούνται τα στοιχεία της γλώσσας HTML, ώστε να μείνει το καθαρό κείμενο, όπως επίσης και τα σημεία στίξης, καθώς και ένα σύνολο από κοινά εμφανιζόμενες λέξεις όπως άρθρα και γενικά λέξεις που δεν παρέχουν πληροφορία. Στη συνέχεια, από το κείμενο κατασκευάζεται ένα διάνυσμα παρόμοιο με αυτά τα οποία χρησιμοποιήθηκαν στην εκπαίδευση του ταξινομητή. Το διάνυσμα δίνεται σαν είσοδος στον ταξινομητή, ο οποίος το κατηγοριοποιεί σε μία από τις δύο κλάσεις. Η απόδοση του ταξινομητή είναι αρκετά υψηλή, προσεγγίζοντας το 95%. Περιθώρια περαιτέρω βελτίωσης υπάρχουν επιλέγοντας μεγαλύτερο πλήθος από κείμενα εκπαίδευσης.

Σύστημα Αρχικοποίησης (Formalization System). Οι πιο σημαντικές λειτουργίες που παρέχονται στο χρήστη από το *Σύστημα Αρχικοποίησης* είναι:

- Εισαγωγή βιογραφικού. Ο χρήστης εισάγει σε μία ειδικά διαμορφωμένη φόρμα στοιχεία από το βιογραφικό του σημείωμα σε ελεύθερο κείμενο. Από τα στοιχεία του χρήστη δημιουργείται ένα αρχείο XML με συγκεκριμένο σχήμα (DTD).
- Αναζήτηση Αγγελιών. Το σύστημα ψάχνει να βρει αγγελίες που ταιριάζουν με το βιογραφικό του χρήστη. Η αναζήτηση γίνεται ανάμεσα στις αγγελίες που έχει βρει ο αναζητητής και οι οποίες έχουν μετατραπεί σε XML αρχεία. Για τη διαδικασία ταιριάσματος έχει δημιουργηθεί μια αντιστοίχιση ανάμεσα στα πεδία του σχήματος που περιγράφει τις αγγελίες και τα πεδία του σχήματος που περιγράφει τα βιογραφικά. Το ταίριασμα γίνεται, ελέγχοντας την ύπαρξη κοινών λέξεων ανάμεσα στα συσχετιζόμενα πεδία των δύο αρχείων. Εισάγοντας βάρη σπουδαιότητας σε κάθε διαφορετικό πεδίο, παρέχεται ένα ποιοτικό φιλτράρισμα των αποτελεσμάτων, ώστε να εξασφαλίζεται η όσο το δυνατόν μεγαλύτερη συνέχεια μεταξύ των αγγελιών που επιστρέφονται στο χρήστη και των χαρακτηριστικών του τελευταίου.

Αναλυτής Σελίδων (Page Analyzer). Για να είναι όσο το δυνατόν πιο αποτελεσματικό το φιλτράρισμα των αποτελεσμάτων σε σχέση με τα χαρακτηριστικά του χρήστη, αλλά και για μεγαλύτερη ευελιξία, από τις ιστοσελίδες που έχει εντοπίσει ο αναζητητής, εξάγεται η πληροφορία που περιέχεται και εισάγεται σε ένα αρχείο XML. Για τη διαδικασία αυτή, έχουμε ορίσει ένα προκαθορισμένο σχήμα (DTD) που θα έχουν τα αρχεία XML και για κάθε πεδίο του σχήματος έχουν προσδιορισθεί λέξεις κλειδιά, οι οποίες μέσα στο κείμενο προσδιορίζουν την αρχή και το τέλος

της πληροφορίας. Η συγκεκριμένη ευριστική προσέγγιση είναι αποδοτική μόνο στη περίπτωση που οι ιστοσελίδες που εξετάζουμε είναι ημιδομημένες, όπως ισχύει στη περίπτωση των ιστοσελίδων που περιέχουν αγγελίες για θέσεις εργασίας.

4 Συμπεράσματα

Δεδομένου του όγκου των πληροφοριών στο διαδίκτυο, το πρόβλημα της ανάκτησης πληροφοριών είναι ιδιαίτερο σημαντικό. Στη παρούσα εργασία μελετήθηκαν διάφορες προσεγγίσεις, εστιάζοντας στο πρόβλημα της προσανατολισμένης αναζήτησης. Αναπτύχθηκε μία ευριστική μέθοδος προσανατολισμένης αναζήτησης και ενισχύθηκε με τεχνικές από τη περιοχή της μηχανικής μάθησης, με σκοπό την αποτελεσματική ανάκτηση πληροφοριών, για τη περιοχή της αγοράς εργασίας. Παράλληλα, αναπτύχθηκαν ευριστικές προσεγγίσεις για την εξαγωγή πληροφοριών από τις ιστοσελίδες, καθώς και του φιλτραρίσματος των αποτελεσμάτων με βάση τα χαρακτηριστικά του χρήστη. Σε αρκετά σημεία χρησιμοποιήθηκε η μεθοδολογία αναπαράστασης της XML, αξιοποιώντας τη δόμηση που παρέχει στις πληροφορίες. Υπάρχουν διάφορα σημεία τα οποία μπορούν να βελτιωθούν, όπως η χρήση μεθόδων για προσανατολισμένη αναζήτηση, οι οποίες να επιτρέπουν την εύκολη μετάβαση μεταξύ θεματικών περιοχών, ενώ για τη διαδικασία εξαγωγής πληροφορίας μπορούν να χρησιμοποιηθούν διάφορες τεχνικές, όπως αυτόματη εκμάθηση περιτυλιγμάτων (wrapper induction). Περιθώρια περαιτέρω βελτίωσης του ταξινομητή υπάρχουν, επιλέγοντας μεγαλύτερο πλήθος από κείμενα εκπαίδευσης.

References

1. Tom M. Mitchell McGraw-Hill 1997. Machine Learning.
2. Marc Najork, Janet L. Wiener. Breadth - first Search Crawling Yields High Quality Pages, Proceedings of the www10 2001.
3. Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore. Automating the Construction of Internet Portals with Machine Learning, Information Retrieval vol. 3, 2000.
4. Jason Rennie, Andrew McCallum. Efficient web Spidering with Reinforcement Learningm Information Retrieval vol.3 2000.
5. M.Diligenti, F.M.Coetzee, S. Lawrence, C.L.Giles and M.Gori. Focused Crawling using Context Graphs, 26th International Conference on Very Large Databases, VLDB2000.
6. Fabrizio Sebastiani. Machine Learning in Automated Text Categorization, ACM Computing Surveys(34)1, 2002.
7. Monika Henzinger. Link Analysin in Web Information Retrieval, IEEE Data Engineering Bulletin 23(3) 2000.
8. S. Brin and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems, vol. 30, 1998.