Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463, Συστήματα Ανάκτησης Πληροφοριών
2005-2006 Εαρινό Εξάμηνο

Εργασία Μαθήματος: "GRoogle2006"

Σχεδιασμός και Υλοποίηση μιας Μηχανής Αναζήτησης για τον Παγκόσμιο Ιστό

Ανάθεση: 24/3/2006
Τελική Παράδοση: 15/6/2006

Γενική Περιγραφή

Στόχος

Στόχος της εργασίας αυτής είναι ο σχεδιασμός και η υλοποίηση μιας μηχανής αναζήτησης για τον Παγκόσμιο Ιστό. Σε γενικές γραμμές, η μηχανή αυτή θα πρέπει να λειτουργεί όπως το Google, και γι αυτό και θα μπορούσαμε να την ονομάσουμε GRoogle2006 (όπου GR από το Greece). Χονδρικά, μια τέτοια μηχανή συγκροτείται από τα εξής υποσυστήματα:

- Ερπυστής (Crawler): Με σκοπό τη διάσχιση του ιστού, τη συλλογή και αποθήκευση σε τοπικό χώρο των ιστοσελίδων.
- Ευρετηριαστής (Indexer): Με σκοπό την συντακτική ανάλυση των σελίδων και κατασκευή των ευρετηρίων που απαιτούνται για γρήγορη αποτίμηση επερωτήσεων.
- Διαβαθμιστής (Ranker): Με σκοπό τη διαβάθμιση των σελίδων με τεχνικές ανάλυσης συνδέσμων (τύπου PageRank).
- Αποτιμητής Επερωτήσεων (Query Evaluator): Με σκοπό την αποτίμηση των επερωτήσεων του χρήστη.
- Ομαδοποίηση αποτελεσμάτων και Τοπική Ανάλυση (Document Clustering): Με σκοπό τη δυνατότητα ομαδοποίησης των αποτελεσμάτων μιας επερώτησης, καθώς και κατ' επιλογής επέκτασης της αρχικής επερώτησης με τις πιο συχνά εμφανιζόμενες λέξεις των κορυφαίων εγγράφων.
- Επαφή Χρήσης (User Interface): Με σκοπό την υποστήριξη του διαλόγου μεταξύ χρήστη και μηχανής.

Οργάνωση Εργασιών

Την προηγούμενη χρονιά οι συνάδελφοι σας έπρεπε να αναπτύξουν ένα ολόκληρο σύστημα ανάκτησης πληροφοριών για συλλογές εγγράφων κειμένου. Εφέτος ο σκοπός είναι να αναπτυχθεί ένα μόνο σύστημα, και για το λόγο αυτό το σύστημα θα αναλυθεί σε υποσυστήματα και κάθε ομάδα (2 ή 3 φοιτητών) θα είναι υπεύθυνη για ένα ή δύο μόνο υποσυστήματα. Κάθε ομάδα πρέπει να δώσει έμφαση στον καλό σχεδιασμό, στις επιδόσεις (αφού θα έχουμε να κάνουμε με μεγάλους όγκους πληροφοριών) και στην καινοτομία (οι πρωτοβουλίες και οι νέες ιδέες είναι παραπάνω από ευπρόσδεκτες). Επίσης θα πρέπει να δοθεί έμφαση και στην συνεργασία μεταξύ των ομάδων. Για το λόγο αυτό μια ομάδα δεν θα αναλάβει προγραμματιστικό έργο αλλά τον συντονισμό των υπολοίπων ομάδων (οργάνωση συναντήσεων) και την ενοποίηση των υποσυστημάτων που θα προκύπτουν. Η ομάδα αυτή θα πρέπει να είναι σε θέση να παρουσιάζει την πρόοδο του έργου καθώς και το τελικό έργο. Τα φροντιστήρια του μαθήματος μπορεί να αποτελέσουν τον τόπο συνάντησης για τον συντονισμό των ομάδων και για συζήτηση με τους βοηθούς.

Διαθέσιμος κώδικας Εργαλεία

Για την υλοποίηση του συστήματος μπορείτε να χρησιμοποιήσετε Java. Για να μπείτε γρήγορα στο «παιχνίδι» και να διευκολυνθεί ο συντονισμός των ομάδων σας έχει δοθεί ήδη ο κώδικας δύο συστημάτων που έχουν αναπτύξει οι βοηθοί του μαθήματος την προηγούμενη χρονιά. Επίσης θα σας προταθεί ένα σχέδιο (design) για το σύστημα που πρέπει να αναπτύξετε. Μπορείτε να χρησιμοποιήσετε τα παραπάνω ως αφετηρία αλλά μπορείτε να αναδομήσετε ό,τι επιθυμείτε κατά βούληση. Περισσότερες πληροφορίες μπορείτε να βρείτε στην διεύθυνση: http://www.csd.uoc.gr/~hy463/2006/el/assignments.html

Χρονοδιάγραμμα

Φάση	**Ημερομηνίες**
Φ1: GRoogle Ver 0.1	24 Μαρτίου – 2 Μαΐου
Φ2: GRoogle Ver 1.0 (τελική έκδοση)	3 Μαΐου – 15 Μαΐου
Φ3: Δοκιμές και βελτιώσεις	15 Μαΐου – 15 Ιουνίου

Παραδοτέα

Σε όλες τις φάσεις πρέπει να φορτώσετε (submit) τον κώδικα του συστήματος σας με τρόπο που θα ανακοινωθεί. Κάθε ομάδα θα υποβάλει ξεχωριστά το αποτέλεσμα της εργασίας της, ενώ η ομάδα συντονισμού θα υποβάλει το ενοποιημένο σύστημα. Επίσης με την παράδοση της τελικής φάσης θα πρέπει να παραδώσετε τυπωμένη αναφορά η οποία να περιέχει:
(α) περιγραφή της λειτουργικότητας και των βασικών ατού του συστήματος σας,
(β) περιγραφή και ανάλυση των πειραμάτων που κάνατε και του τι μάθατε από αυτά.
Επίσης θα χρειαστεί να κάνετε μια επίδειξη της λειτουργίας του συστήματος σας στο τέλος του εξαμήνου.

Αξιολόγηση Συστήματος

Το τελικό αποτέλεσμα θα αξιολογηθεί (εμπειρικά) σε πραγματικά δεδομένα. Η «ορθότητα» της λειτουργίας του θα δοκιμαστεί σε μικρές συλλογές σελίδων ενώ για την αξιολόγηση των επιδόσεων θα χρησιμοποιηθούν μεγαλύτερες συλλογές, για παράδειγμα όλες οι σελίδες του csd.uoc.gr και του www.csi.forth.gr. Όσον αφορά τον σχεδιασμό του συστήματος κάντε την υπόθεση ότι το σύστημα θα πρέπει να δουλεύει ικανοποιητικά για τουλάχιστον 1.000.000 έγγραφα.

Βαθμολόγηση

Θα υπάρξει ξεχωριστή βαθμολογία για κάθε ομάδα και για κάθε φάση. Το τελικό αποτέλεσμα (ήτοι το GRoogle2006 ver 1.0) θα λειτουργήσει ως καταλύτης (θετικός ή αρνητικός) στη βαθμολογία όλων των ομάδων. Αυτό για να ενθαρρύνουμε την συνεργασία και τον συντονισμό μεταξύ των ομάδων και για να βελτιωθεί το τελικό αποτέλεσμα.

Αναλυτικότερη Περιγραφή Υποσυστημάτων

Περισσότερες πληροφορίες για το κάθε υποσύστημα θα δοθούν στο μάθημα και στα φροντιστήρια. Κάποια επιπλέον στοιχεία ακολουθούν.

Ερπυστής (Crawler):
Σκοπός του «ερπυστή» είναι η διάσχιση του ιστού και η συλλογή και αποθήκευση των σελίδων σε τοπικό χώρο. Τα σημεία εκκίνησης θα καθορίζονται από εξωτερικό αρχείο. Ο ερπυστής θα πρέπει επίσης να λαμβάνει παραμέτρους που θα καθορίζουν την στρατηγική διάσχισης (dfs, bfs, important first, depth within a site, κλπ). Το αποτέλεσμα της λειτουργία του θα είναι μια τοπική αποθήκη σελίδων η οποία θα συγκροτείται από ένα ευρετήριο, π.χ. ένα αρχείο με εγγραφές της μορφής "docId, URI, type, lastUpdate, lastFetched, ChangeFrequency" (και ότι άλλο κρίνετε αναγκαίο), καθώς και τοπικά αντίγραφα των σελίδων με τοπικό όνομα docId. Ο ερπυστής πρέπει επίσης να συντηρεί το ημερολόγιο διάσχισης (ανανέωση σελίδων, προσθήκη, διαγραφή, διάσχιση, κλπ). Για να επιταχυνθεί η διαδικασία της διάσχισης ο ερπυστής μπορεί να δημιουργεί πολλά νήματα για το κατέβασμα (downloading) των σελίδων και πρέπει να αποφεύγει τις συνεχείς κλήσεις στο ίδιο ιστόχωρο.

Λεξιλογικός Αναλυτής.
Για αρχεία κειμένου, html (και ότι άλλο θέλετε, π.χ. pdf).

Θα μπορεί να λαμβάνει και αρχείο με λέξεις αποκλεισμού (stoplist). Μια stoplist για την αγγλική γλώσσα είναι διαθέσιμη στη δνση http://www.csd.uoc.gr/~hy463/project/.

Στελεχωτής για την Ελληνική Γλώσσα
Για τη στελέχωση των κειμένων (και των επερωτήσεων) της αγγλικής γλώσσας μπορείτε να χρησιμοποιήσετε τον αλγόριθμο του Porter τον οποίο μπορείτε να βρείτε υλοποιημένο σε Java στην διεύθυνση: http://www.tartarus.org/~martin/PorterStemmer/.
Για την ελληνική γλώσσα θα πρέπει να φτιάξετε τον δικό σας, βασισμένο σε κανόνες όπως αυτούς του Porter, ή σε κάποια άλλη τεχνική από αυτές που θα δούμε στο μάθημα (π.χ. successor variety).

Ευρετηριαστής
Υπεύθυνος για την κατασκευή και συντήρηση του ευρετηρίου της μηχανής. Το ευρετήριο θα αποτελείται από (α) το αρχείο λεξιλογίου, (β) το αρχείο με τις ανεστραμμένες λίστες, και (γ) το ευρετήριο συνδέσμων. Το (α) θα περιέχει δείκτες προς το (β). Για μείωση του χώρου του (β), μπορεί να χρησιμοποιηθεί block addressing στις ανεστραμμένες λίστες. Το (γ) θα καταχωρεί τους συνδέσμους. Για την βάρυνση των όρων μπορεί να λαμβάνετε υπόψη τη δομή του εγγράφου (τίτλος, h1, h2, κλπ). Επίσης anchor indexing.

Διαβαθμιστής (Global Ranker)
Θα διαβάζει το ευρετήριο συνδέσμων και θα υπολογίζει τα σκορ αλα PageRank. Πρόβλεψη για γρήγορο υπολογισμό και biased Page Rank. Επίσης μπορεί υποστηρίζονται τεχνικές για αποκλεισμό των spam σελίδων, για παράδειγμα Inverse PageRank (όπου τότε θα πρέπει να λαμβάνει υπόψη του και αρχείο με διευθύνσεις spam sites/pages).

Αποτιμητής επερωτήσεων
Αξιοποιώντας τα ευρετήρια θα υπολογίζει την απάντηση της επερώτησης. Υποστήριξη επερωτήσεων ελεύθερου κειμένου, Boolean εκφράσεων, τελεστών εγγύτητας (proximity operators). Ο υπολογισμός του βαθμού συνάφειας ενός εγγράφου θα εξαρτάται και από το βαθμό PageRank της σελίδας και από το βαθμό ομοιότητας βάσει του μοντέλου ανάκτησης (το οποίο μπορεί να είναι το Διανυσματικό, το Boolean, το Simple Fuzzy, το Extended Boolean, κλπ).

Επαφή Χρήσεως
Θα παραγάγει και θα παρουσιάζει την λίστα των αποτελεσμάτων της επερώτησης. Η απάντηση μιας επερώτησης θα δίνεται στο χρήστη ως ακολουθία σελίδων με ελεγχόμενο (από το χρήστη) αριθμό στοιχείων ανά σελίδα. Για κάθε στοιχείο της απάντησης πρέπει να παρουσιάζεται ένα ενδεικτικό αποσπάσματα του αντίστοιχου εγγράφου (το οποίο θα εξάγεται από το αρχείο των σελίδων στον τοπικό χώρο). Αν το ανεστραμμένο ευρετήριο χρησιμοποιεί block addressing, τότε θα πρέπει να γίνεται αναζήτηση κειμένου (text searching) στο σχετικό block. Επίσης κάθε στοιχείο της απάντησης θα συνοδεύεται από σύνδεσμο προς την αυθεντική σελίδα, καθώς και από σύνδεσμο προς το τοπικό αντίγραφο της σελίδας (αν αυτό υπάρχει). Δίπλα σε κάθε σημείο μπορεί να υπάρχει και μια επιλογή MarkThisAsSpam (και η επιλογή του θα πρέπει να ενημερώνει κατάλληλα των αρχείων σελίδων αποκλεισμού). Επίσης η παρουσίαση των αποτελεσμάτων μπορεί να συνοδεύεται από το αποτέλεσμα εφαρμογής ιεραρχικής ομαδοποίησής σε αυτήν (όπως γίνεται στο www.vivisimo.com). Η επαφή χρήσεως πρέπει να έχει δυο εκδόσεις: μια με Web interface υποθέτοντας tomcat και μία text-based.

Προαιρετικές Λειτουργίες

- Επέκταση της ερωτηματικής γλώσσας ώστε να περιλαμβάνει τελεστές Edit Distance, κλπ.
- Υποστήριξη Προφίλ Χρηστών
- Μηχανισμό ψευδο-ανάδρασης συνάφειας (pseudo relevance feedback).
- Και ό,τι άλλο θέλετε ή φανταστείτε.

Καλή δουλειά & διασκέδαση