HY463 Αναφορά Project – Α φάση

Υλοποίηση: Αναστάσιος Χατζηαλεξίου - CSD 3902

A. Τι υλοποιήθηκε και τι όχι:

- Υπάρχει γραφική διεπαφή η οποία λειτουργεί για τη δημιουργία, το φόρτωμα του ευρετηρίου και για την αναζήτηση. Η ευρετηρίαση και το φόρτωμα του ευρετηρίου έχουν υλοποιηθεί πλήρως.
- Δεν υπάρχει για την ώρα σύνδεση της αναζήτησης με το GUI.
- Εχουν επίσης υλοποιηθεί οι συναρτήσεις που επιστρέφουν τις plain, essential, full αναπαραστάσεις.
- Από τα μοντέλα αναζήτησης έχει υλοποιηθεί μόνο το existential.

B. Επισκόπηση υλοποίησης για την ευρετηρίαση.

Σε γενικές γραμμές η μέθοδος ευρετηρίασης που εφαρμόστηκε διαφέρει από τη μέθοδο που περιγράφηκε στις διαλέξεις. Η εγγραφή στο documents file γίνεται από την κλάση Indexer και παράγεται ένα και μόνο τελικό document αρχείο χωρίς να χρειάζεται επιπλέον συνένωση στην περίπτωση πολλαπλών μερικών ευρετηρίων. Οπότε στα μερικά ευρετήρια υπάγονται μόνο αρχεία postings και vocabulary. Η τελική συνένωση των μερικών ευρετηρίων γίνεται ως: 1) συνένωση όλων των vocabulary ταυτόχρονα 2) υπολογισμός των VSM weights 3) συνένωση όλων των postings ταυτόχρονα. Ενδιάμεσα indexes δεν παράγονται. Η συνένωση των vocabulary παράγει επιπλέον ένα προσωρινό αρχείο που χρησιμοποιείται για τη συνένωση των postings αργότερα. Οσον αφορά τις απαιτήσεις μνήμης, για κάθε μερικό ευρετήριο δεν χρησιμοποιήθηκαν πολλές δομές, αλλά μόνο μία δομή που κρατάει ταυτόχρονα πληροφορίες και για το vocabulary και τα postings. Αυτές οι πληροφορίες είναι μόνο όσες δε μπορούν να υπολογιστούν κατά τη διαδικασία εγγραφής των τελικών αρχείων στο δίσκο. Τέλος, στο postings file δεν αποθηκεύεται το docId.

Για τον υπολογισμό των VSM weights αποφασίστηκε κατά τη διάρκεια ανάγνωσης του dataset να εγγραφούν στο δίσκο όσα δεδομένα χρειάζονται ώστε να μην χρειαστεί να αναγνωστεί το τελικό postings file. Συγκεκριμένα, κατά τη διάρκεια της παραγωγής των μερικών ευρετηρίων, παράγονται 2 επιπλέον προσωρινά αρχεία τα οποία χρησιμοποιούνται για γρήγορο και άμεσο υπολογισμό των weights.

Τέλος, έχει δοθεί η απαραίτητη προσοχή ώστε να σβήνονται άμεσα όσα αρχεία δε χρειάζονται πια κατά τη διάρκεια της ευρετηρίασης ώστε να μειωθεί ο απαιτούμενος αποθηκευτικός χώρος.

C. Ενδεικτικοί χρόνοι & αποθηκευτικός χώρος για την ευρετηρίαση.

Η διαδικασία της συνένωσης των μερικών ευρετηρίων είναι αρκετά γρήγορη με τον τρόπο που γίνεται. Αυτό έχει ως συνέπεια να μην αυξάνει σημαντικά ο χρόνος που χρειάζεται στην

περίπτωση που το πλήθος των μερικών ευρετηρίων είναι μεγάλο. Μάλιστα μετά από αρκετές δοκιμές βρέθηκε ότι όσο μειώνεται ο αριθμός των μερικών ευρετηρίων για δεδομένη συλλογή, ο χρόνος παραγωγής τους αυξάνεται πολύ γρήγορα ενώ ο χρόνος συνένωσης τους μειώνεται με πιο αργό ρυθμό. Το σύστημα φαίνεται ότι αποδίδει χειρότερα όσο περισσότερη μνήμη του δίνουμε.

Σύστημα δοκιμής:
OS: Windows 7
RAM: 8GB DDR2
DISK: 500GB SSD
CPU: Dual core @ 3GHz, 6MB L2 cache
VM options: -Xmx3000m

Ενδεικτικοί χρόνοι χωρίς stemming και stopwords για τη συλλογή μεγέθους 1.5GB με 1M έγγραφα:

	Split at article				
	5K	15K	50K	100K	300K
# Partial index	200	67	20	10	4
Partial vocabularies, partial postings + documents	192s	205s	202s	235s	254s
Merge vocabularies	14s	10s	7s	6s	5s
Merge postings	8s	6s	5s	4s	4s
VSM weights	45s				
Total	259s	266s	259s	290s	308s

Παρακάτω φαίνεται το μέγεθος των παραχθέντων αρχείων:

	Disk usage				
	5K	15K	50K	100K	300K
Partial indexes (w/o documents file)	1068MB	978MB	926MB	901MB	874MB
Documents file	225MB				
Vocabulary file	68MB				
Postings file	778MB				
term_tf temp file	77MB				
doc_tf temp file	615MB				
doc_size temp file	4MB				
Final index	1098MB				
Max disk usage (w/ final index	1989MB < (2x final index size)				

D. Ενδεικτικοί χρόνοι αναζήτησης

Existential model χωρίς stemming και stopwords για 1,3,8 διαφορετικούς όρους ταυτόχρονα με εμφανίσεις σε 100 documents κάθε ένας. Οι χρόνοι είναι για αναζήτηση των plain, essential, full αναπαραστάσεων για όλους τους όρους.

	1 term	3 terms	8 terms
Plain	5ms	13ms	29ms
Essential	17ms	42ms	104ms
Full	35ms	68ms	142ms