HY463 Αναφορά Project – Α φάση

Υλοποίηση: Αναστάσιος Χατζηαλεξίου - CSD 3902

A. Τι υλοποιήθηκε και τι όχι:

- Υπάρχει γραφική διεπαφή η οποία λειτουργεί πλήρως για τη δημιουργία και το φόρτωμα του ευρετηρίου. Για την αναζήτηση η διεπαφή τροποποιείται κατάλληλα, όμως δεν είναι λειτουργική.
- Η ευρετηρίαση και το φόρτωμα του ευρετηρίου έχουν υλοποιηθεί πλήρως.
- 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 μια και καλή από την κλάση Indexer και παράγεται ένα και μόνο τελικό αρχείο χωρίς να χρειάζεται επιπλέον συνένωση στην περίπτωση πολλαπλών μερικών ευρετηρίων. Οπότε στα μερικά ευρετήρια υπάγονται μόνο αρχεία postings.idx και vocabulary.idx. Η τελική συνένωση των μερικών ευρετηρίων γίνεται με συνένωση όλων των ευρετηρίων ταυτόχρονα χωρίς την ανάγκη να παράγονται επιπλέον μερικά ευρετήρια σαν ενδιάμεσα βήματα. Οσον αφορά τις απαιτήσεις μνήμης, για κάθε μερικό ευρετήριο δε χρησιμοποιήθηκαν πολλές δομές, αλλά μόνο μία δομή που κρατάει ταυτόχρονα πληροφορίες και για το vocabulary file και το postings file. Αυτές οι πληροφορίες είναι μόνο όσες δε μπορούν να υπολογιστούν κατά τη διαδικασία εγγραφής στο δίσκο. Τέλος, στο postings file δεν αποθηκεύεται το docId.

Οσον αφορά τον υπολογισμό των VSM weights αρχικά δοκιμάστηκε πειραματικά να γίνει με ανάγνωση των Postings file, ωστόσο αυτή η διαδικασία ήταν αρκετά χρονοβόρα. Οπότε αποφασίστηκε κατά τη διάρκεια ανάγνωσης του dataset να εγγραφούν στο δίσκο όσα δεδομένα χρειάζονται για τον άμεσο υπολογισμό των weights ο οποίος γίνεται αφού έχουν συνενωθεί όλα τα μερικά ευρετήρια. Αυτός ο τρόπος αυξάνει αρκετά το μέγιστο αποθηκευτικό που χρειαζόμαστε όπως φαίνεται στον πίνακα της παρακάτω ενότητας (υπάρχει τρόπος να μειωθεί αυτός ο χώρος και κατά πάσα πιθανότητα θα υλοποιηθεί στο μέλλον).

Επιπλέον, δεσμεύεται πίνακας μεγέθους όσο και το πλήθος των εγγράφων όπου και αποθηκεύονται οι ενδιάμεσοι υπολογισμοί των βαρών των εγγράφων πριν εγγραφούν στο δίσκο. Αυτό ενδεχομένως να δημιουργήσει πρόβλημα αν το πλήθος τους είναι πολύ μεγάλο.

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 μειώνεται με πιο αργό ρυθμό. Με λίγα λόγια το σύστημα από κάποιο σημείο και πέρα 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) φαίνονται παρακάτω:

	Split at				
	5K	15K	50K	100K	300K
Partial index	195s	207s	212s	217s	255s
# Partial index	200	66	20	10	4
Merge	61s	47s	35s	27s	22s
VSM	45s				
Total	301s	299s	292s	288s	322s

Παρατηρούμε ότι στα 300K ο χρόνος μερικής ευρετηρίασης αυξήθηκε πολύ σε σχέση με τα 100K (+39s), ενώ ο χρόνος συνένωσης μειώθηκε κατά πολύ μικρότερο ποσό (-5s).

Παρακάτω φαίνεται το μέγεθος των παραχθέντων αρχείων στο δίσκο για τη δοσμένη συλλογή μεγέθους 1.5GB με 1M entries.

	Disk usage				
	5K	15K	50K	100K	300K
Partial indexes (w/o documents file)	1068MB	978MB	926MB	901MB	874MB
Documents file	221MB				
VSM files	615MB				
Final index	1093MB				
Max disk usage	2776MB	2686MB	2634MB	2609MB	2582MB

Παρατηρούμε ότι ο αποθηκευτικός χώρος αυξάνεται σημαντικά με τη χρήση των επιπλέον αρχείων για τον γρήγορο υπολογισμό των VSM weights.

Τέλος, όσον αφορά τη μνήμη, ο μέγιστος αριθμός που αναφέρθηκε στον task manager ήταν περίπου 2.3GB για τα 300K. Ωστόσο είναι πολύ πιθανόν αυτός ο αριθμός να περιλαμβάνει και μνήμη που δεν έχει γίνει garbage collected.

D. Ενδεικτικοί χρόνοι αναζήτησης

Με το existential model, η αναζήτηση ενός όρου με εμφανίσεις σε 100 έγγραφα επιστρέφει αποτελέσματα σε 7ms. Η αναζήτηση 7 όρων με εμφανίσεις σε 100 έγγραφα ο κάθε ένας, επιστρέφει αποτελέσματα σε 4ms κατά μέσο όρο.