ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
Τμήμα Πληροφορικής

Εξόρυξη Δεδομένων
Ακαδημαϊκό Έτος 2006-2007

2ο Σύνολο Ασκήσεων
Ημερομηνία Παράδοσης: 16 Ιουλίου 2007

Οι ασκήσεις είναι ατομικές (αυστηρά)
50% Τελικού Βαθμού – Το άριστα είναι το 100

Ενότητες: Εξόρυξη Κανόνων Συσχέτισης, Μηχανές Αναζήτησης

Άσκηση 1 [25 μονάδες]
(α) Ποια είναι η εμπιστοσύνη (confidence) των κανόνων $\varnothing \rightarrow A$ και $A \rightarrow \varnothing$
(β) Έστω c1, c2 και c3 η εμπιστοσύνη των κανόνων $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q, r\}$ και $\{p, r\} \rightarrow \{q\}$ αντίστοιχα. Έστω $c1 \neq c2 \neq c3$, ποιες είναι οι πιθανές σχέσεις μεταξύ των c1, c2 και c3; Ποιος κανόνας έχει την μικρότερη εμπιστοσύνη;
(γ) Επαναλάβετε το ερώτημα (β), υποθέτοντας ότι και οι τρεις κανόνες έχουν την ίδια υποστήριξη (support). Ποιος κανόνας έχει τη μεγαλύτερη εμπιστοσύνη;
(δ) Σχολιάστε αν το συμπέρασμα του (γ) χρησιμοποιείται και πως σε κάποιους από τους αλγορίθμους που μελετήσαμε στο μάθημα.
(ε) Μεταβατικότητα: Έστω ότι η εμπιστοσύνη των κανόνων $A \rightarrow B$ και $B \rightarrow C$ είναι μεγαλύτερη κάποιου κατωφλίου minconf. Είναι πιθανόν ο κανόνας $A \rightarrow C$ να έχει εμπιστοσύνη μικρότερη από minconf.

Άσκηση 2 [45 μονάδες]
Έστω οι παρακάτω δοσοληψίες
T1: {M, O, N, K, E, Y}
T2: {D, O, N, K, Y, E}
T3: {M, A, K, E}
T4: {M, U, C, Y, K}
T5: {C, O, K, I, E}
Θεωρείστε ελάχιστη υποστήριξη 60% και ελάχιστη εμπιστοσύνη 80%.
(α) Εφαρμόστε apriori για να βρείτε τα συχνά στοιχειοσύνολα. Δείξτε όλα τα βήματα του αλγορίθμου.
(β) Εφαρμόστε τον FP-Growth για να βρείτε τα συχνά στοιχειοσύνολα. Δείξτε όλα τα βήματα του αλγορίθμου.
(γ) Ποια από τα συχνά στοιχειοσύνολα που βρήκατε είναι κλειστά, και ποια maximal συχνά:
(δ) Ποιο είναι το ποσοστό των συχνών στοιχειοσυνόλων σε σχέση με όλα τα πιθανά στοιχειοσύνολα;
(ε) Δώστε τους κανόνες συσχέτισης που προκύπτουν. Δώστε τα βήματα παραγωγής τους.

Άσκηση 3 [30 μονάδες]
(α) Δώστε ένα παράδειγμα γράφου web που ο PageRank και ο HITS θα έδιναν το ίδιο αποτέλεσμα (συγκεκριμένα τους ίδιους 2 κόμβους ως καλύτερους) για κάποιο θέμα.
(β) Ο PageRank χρησιμοποιεί μια σταθερά απόσβεσης (damping factor) d που πιθανολογείται ότι στο Google είναι ίση με 0.85. Για τα παρακάτω ερωτήματα εξηγείστε πως νομίζετε ότι η τιμή του d επηρεάζει (αν επηρεάζει):

- Τον αριθμό των επαναλήψεων του PageRank
- Την τελική τιμή του PageRank
- Τη διάταξη των σελίδων με βάση το PageRank