3ο Σύνολο Ασκήσεων

Ημερομηνία Παράδοσης: **6 Ιουνίου 2011**, στην εξέταση του μαθήματος (+10% Bonus)
30 Ιουνίου 2011, 13:00 μμ στο γραφείο μου

Ενότητα: Ανάλυση Συνδέσεων (PageRank, HITS) και άλλα
Ποσοστό στον τελικό βαθμό: 30 % ως απαλλακτικές
15 % αν δώσετε τελικό διαγώνισμα

Όλες οι ασκήσεις είναι ατομικές. Οι ασκήσεις είναι απαλλακτικές, με την έννοια ότι μπορεί να αναπληρώσουν το τελικό διαγώνισμα (δείτε και τη σελίδα του μαθήματος).

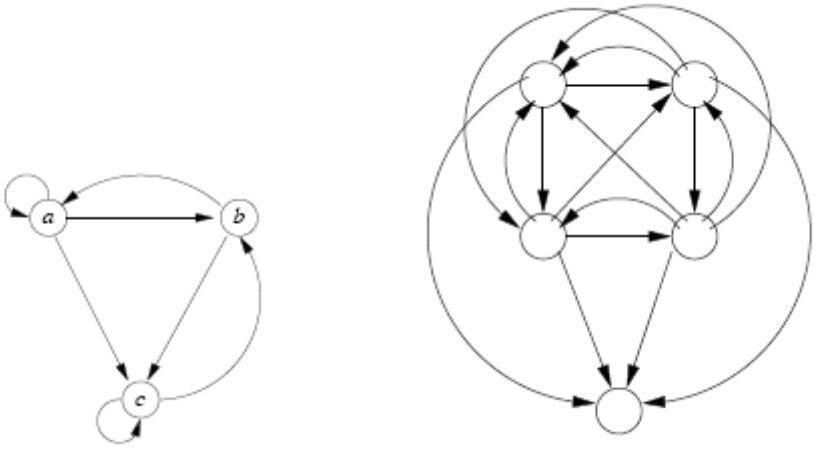
Στο βιβλίο δεν υπάρχει σχετικό κεφάλαιο για Ανάλυση Συνδέσεων. Διαβάστε το Κεφάλαιο 5 (Link Analysis) από το βιβλίο "Mining of Massive Datasets", των Anand Rajaraman και Jeff Ullman.
(online στο: http://infolab.stanford.edu/~ullman/mmds.html)

Άσκηση 1 [30 μονάδες]

Δώστε σύντομες απαντήσεις στα παρακάτω ερωτήματα του αφορούν το PageRank
(α) Μια ιδέα που χρησιμοποιείσαι το google είναι να θεωρήσει τους όρους που εμφανίζονται στα links που δείχνουν σε μια σελίδα ως μέρος της σελίδας. Εξηγείστε γιατί αυτό είναι χρήσιμο.
(β) Τι είναι ένας γράφος μετάβασης (transition graph); Κατασκευάστε τον πίνακα μετάβασης (transition matrix) για το γράφο του Σχήματος 1(αριστερά).
(γ) Υπολογίστε το PageRank για το γράφο του Σχήματος 1(αριστερά).
(δ) Πότε μια σελίδα καλείται αδιέξοδο (dead end); Τροποποιείστε το γράφο του Σχήματος 1(αριστερά), ώστε να έχει 1 αδιέξοδο. Δώστε τον πίνακα μετάβασης για το νέο γράφο. Υπολογίστε το PageRank για το νέο γράφο και εξηγείστε το αποτέλεσμα.
(ε) Δώστε τον ορισμό της «αραχνοπαγίδας» (spider trap) και δώστε ένα παράδειγμα ενός γράφου που να έχει μια τέτοια παγίδα. Πως επηρεάζουν τον υπολογισμό του PageRank;
(στ) Υπολογίστε το PageRank για το γράφο του Σχήματος 1(αριστερά), θεωρώντας παράγοντα β ίσο με 0.9 (όπου 1 - β είναι η πιθανότητα ο τυχαίος περιπατητής να διαλέξει μια τυχαία σελίδα).
(ζ) Περιγράψτε τουλάχιστον έναν τρόπο για να αυξήσει κάποιος το PageRank της σελίδας του.

Άσκηση 2 [10 μονάδες]

Υπολογίστε το PageRank του γράφου του Σχήματος 1(δεξιά) για β = 0.8



Σχήμα 1: (αριστερά) Γράφος για την Άσκηση 1 και 3 (δεξιά) Γράφος για την Άσκηση 2

Άσκηση 3 [15 μονάδες]

Δώστε σύντομες απαντήσεις στα παρακάτω ερωτήματα του αφορούν το HITS.
(α) Δώστε τον πίνακα συνδεσιμότητας (link matrix) για το γράφο του Σχήματος 1(αριστερά).
(β) Υπολογίστε τους συντελεστές κομβικού ρόλου και αυθεντικότητας για το γράφο του Σχήματος 1(αριστερά).
(γ) Περιγράψτε τουλάχιστον ένα τρόπο για να αυξήσει κάποιος το authority score της σελίδας του.

Άσκηση 4 [15 μονάδες]

Άσκηση 7 (σελ 531) του κεφαλαίου 8, προσθέτοντας στον Πίνακα 7.15 την παρακάτω συναλλαγή:

Κωδικός Συναλλαγής	*Αγορασμένα Αντικείμενα*
8	Πατατάκια, Σόδα

Άσκηση 5 [15 μονάδες]

Θεωρείστε το παρακάτω κείμενα:

d1 : Ο κομήτης του Χάλλεϋ μας επισκέπτεται περίπου κάθε εβδομήντα έξι χρόνια.
d2 : Ο κομήτης του Χάλλεϋ πήρε το όνομά του από τον αστρονόμο Έντμοντ Χάλλεϋ.
d3 : Ένας κομήτης διαγράφει ελλειπτική τροχιά.
d4 : Ο πλανήτης Άρης έχει δύο φυσικούς δορυφόρους, το Δείμο και το Φόβο.
d5 : Ο πλανήτης Δίας έχει 63 γνωστούς φυσικούς δορυφόρους.
d6 : Ένας κομήτης έχει μικρότερη διάμετρο από ότι ένας πλανήτης.
d7 : Ο Άρης είναι ένας πλανήτης του ηλιακού μας συστήματος.

και την ερώτηση q = {κομήτης, Χάλεϋ}.

(α) Αναπαραστήστε τα κείμενα και την ερώτηση χρησιμοποιώντας το λογικό (Boolean) (δηλαδή, βάρος 1 αν υπάρχει ο όρος και 0 αλλιώς) και το διανυσματικό μοντέλο (δηλαδή, με τα tfxidf βάρη για τους όρους).
Θεωρείστε ως όρους μόνο τα ουσιαστικά και κύρια ονόματα.
(β) Κατατάξτε τα κείμενα με βάση την απόσταση τους (με βάση το συνημίτονο) από την ερώτηση για κάθε μοντέλο.

Άσκηση 6 [15 μονάδες]

Στο μάθημα είδαμε τρεις διαφορετικές κατηγορίες τεχνικών εξόρυξης δεδομένων: συσταδοποίηση, κατηγοριοποίηση και εύρεση κανόνων αυτοσυσχέτισης.
Για κάθε μία κατηγορία:
(α) Περιγράψτε σύντομα δύο από τους αλγορίθμους που μελετήσαμε και συγκρίνετε τους (δίνοντας τουλάχιστον ένα πλεονέκτημα και ένα μειονέκτημα για τον καθένα)
(β) Δώστε τον ορισμό τουλάχιστον δύο μέτρων για την εκτίμηση της ποιότητάς τους.
(γ) Περιγράψτε μια εφαρμογή τους που θα μπορούσε να χρησιμοποιηθεί στις μηχανές αναζήτησης.