

TAKE HOME EXAM

DUE DATE: **Monday, January 31 2015 at 20:00 pm (max time limit)**

SEND TO: seminarkudangboro@gmail.com

1. Ulas **perbedaan mendasar** (bukan definisi) antara (a) sistem komputasi terpusat, (b) sistem komputasi tersebar, (c) sistem basis data terpusat, dan (d) sistem basis data tersebar!
2. Identifikasi minimal tiga kompleksitas yang terjadi ketika berhadapan dengan sistem komputasi tersebar dan sistem basis data tersebar dibandingkan yang terpusat!
3. Pada sistem basis data tersebar di mana data tersimpan pada berbagai lokasi fisik yang berbeda, pencarian dan pengambilan data perlu dioptimasi secara signifikan. Identifikasi apa saja dimensi optimasi yang bias dilakukan dan jelaskan argumentasinya!
4. Identifikasi minimal dua parameter yang menentukan kualitas atau keberhasilan dari suatu scenario IR (*information Retrieval*)!
5. Jelaskan keterkaitan yang sangat erat antara Database Management dengan IR?
6. Ulas peran *user feedback* pada peningkatan kinerja IR? (Tip: Merujuk pada *IR System Architecture*)
7. Perhatikan ukuran kesesuaian antara dua *strings* dengan menggunakan *Kendall Tau (T)* dimana
$$T = \frac{P-Q}{\frac{1}{2}n(n-1)}$$
 dimana P: adalah pasangan string yang bersesuaian dan Q adalah pasangan string yang tak bersesuaian, dan *n* adalah jumlah characters pada kedua string. Misal ada dua string "ABC" dan "CAB", maka P= 1 (karena hanya pasangan substring "AB" yang bersesuaian) dan Q = 2, krn ada dua pasang sub-string yang tak bersesuaian yaitu "BC" dan "CA", maka
$$T = \frac{P-Q}{\frac{1}{2}n(n-1)} = \frac{1-2}{\frac{1}{2}(3)(3-1)} = -1/3.$$
 - (a) Hitunglah nilai *Kendall Tau* untuk dua string "Teknologi Informasi" dengan "Informasi Teknologi"?
 - (b) Buktikan untuk sebarang *n* maka kalau dua string itu sepenuhnya bersesuaian maka nilai $T= 1$
 - (c) Buktikan untuk sebarang *n* maka kalau dua string itu sepenuhnya tak bersesuaian maka nilai $T= -1$
 - (d) Identifikasi minimal 2 kelemahan dari penggunaan *Kendall Tau* sebagai alat ukur kesesuaian dua string dalam konteks IR!
8. Ulas kelebihan dan kekurangan dari tiga model IR (a) *Boolean models* (b) *Vector space models*, dan (c) *Probabilistic models*! Jawaban bias dibuat dalam bentuk table dua dimensi dimana barisnya merepresentasikan tiga jenis model tsb, dan kolomnya ada dua yaitu kelebihan dan kelemahan dari masing-masing model. Dengan tabel seperti itu akan lebih jelas ulasan saudara.
9. Suatu dokumen terdiri dari 4 *terms* A B C dan D dengan frekuensi masing masing *terms* A(20), B(15), C(10), D(5) dan koleksi dokumen yang ada terdiri 10,000 documents dan frekuensi (jumlah) dokumen yang mengandung *terms* A, B, C dan D masing masing adalah 50, 1200, 200, 450 yang dinotasikan sebagai A(50), B(1200), C(200), D(450).
 - (a) Hitunglah *tf-idf* untuk masing-masing *terms* A, B, C, D!
 - (b) Apa kesimpulan yang bisa saudara ambil dari hasil perhitungan *tf-idf* tsb.
10. Diketahui pilihan menu A, B, dan C di suatu restoran dimana masing masing menu diekspresikan dengan unsur penyusunnya yang terdiri dari maksimum 10 komposisi penyusun: (1) *nasi*, (2) *kentang*, (3) *daging*, (4) *ikan*, (5) *udang*, (6) *telur*, (7) *garlic*, (8) *slada*, (9) *green pepper*, (10) *tomat*. Komposisi penyusun itu pada menu A, B dan C direpresentasikan sebagai vektor biner yang

menunjukkan ada tidaknya unsur penyusun tsb pada tiap menu dari unsur komposisi sebagai berikut:

$$A = (1, 0, 1, 0, 1, 0, 0, 1, 0, 1)$$

$$B = (0, 1, 0, 1, 0, 1, 0, 0, 1, 1)$$

$$C = (0, 1, 0, 1, 1, 1, 1, 0, 0, 1)$$

Diketahui bahwa masing-masing unsur ada bobot preferensinya seperti berikut:

$$W_{i,j} = \begin{vmatrix} 3 & 0 & 2 & 0 & 4 & 0 & 0 & 3 & 0 & 4 \\ 0 & 4 & 0 & 3 & 0 & 2 & 0 & 0 & 3 & 4 \\ 0 & 3 & 0 & 3 & 2 & 4 & 3 & 0 & 0 & 3 \end{vmatrix}$$

dimana $i = A \dots C$ dan $j = 1 \dots 10$

- (a) Hitung sim (A,C) dan sim (B,C) yg terbobot menggunakan *inner product measure*!
- (b) Hitung sim (A,C) dan sim (B,C) yg terbobot menggunakan *cosin similarity measure*!
- (c) Apa saja kesimpulan yang dapat saudara ambil dari hail komputasi (a) dan (b)?

SELESAI