Tugas Kelompok Rekayasa Data Terapan

Nama Kelompok : Nougat

Nama Anggota :

1. Suryadi – 1706071245
2. Kukuh Nur Aji – 2006494464
3. Jelita Permatasari – 2006547123

Untuk penerapan IR, kelompok kami mengambil referensi dari dua paper dibawah ini:

1. A Boolean Model in Information Retrieval For Search Engines - https://ieeexplore.ieee.org/document/5077062

   Proses pengambilan informasi (IR) dimulai saat pengguna memasukkan kueri ke dalam sistem. Query adalah pernyataan formal dari kebutuhan informasi, misalnya string pencarian di mesin pencari web. Dalam IR, kueri tidak secara unik mengidentifikasi satu objek dalam koleksi. Sebaliknya, beberapa objek mungkin cocok dengan kueri, mungkin dengan tingkat relevansi yang berbeda.

   Objek adalah entitas yang menyimpan atau menyimpan informasi dalam database. Kueri pengguna dicocokkan dengan objek yang disimpan dalam database. Tergantung pada aplikasinya, objek data dapat berupa, misalnya, dokumen teks, gambar atau video. Dokumen itu sendiri tidak disimpan atau disimpan secara langsung di sistem IR, tetapi diwakili dalam sistem oleh pengganti dokumen.

   Sebagian besar sistem IR menghitung skor numerik tentang seberapa cocok setiap objek dalam database dengan kueri, dan memberi peringkat objek sesuai dengan nilai ini. Objek peringkat teratas kemudian ditampilkan kepada pengguna. Prosesnya kemudian dapat diulang jika pengguna ingin memperbaiki kueri.

   Dalam tulisan ini kami mencoba menjelaskan metode IR dan menilai mereka dari dua sudut pandang dan akhirnya mengusulkan metode sederhana untuk peringkat istilah dan dokumen di IR dan menerapkan metode tersebut dan memeriksa hasilnya.

   That each query terms specifies a set of documents containing the term:

   And (∧): The intersection of two sets.

   OR (∨): The union of two sets

   Not (¬): Set inverse or really set difference

   For example:

   If we have 4 documents as:

   Doc1: Information Retrieval has 2 models and Information.

   Doc2: Boolean is a basic Information Retrieval classic model.

   Doc3: Information is a data that processed, Information.

Doc4: When a Data Processed the result is Information, Data.

And D is: Information, Data, and Retrieval

R	Doc1	Doc2	Doc3	Doc4
Data	0	0	1	1
Retrieval	1	1	0	0
Information	1	1	1	1

If Q is: (Data ∧ Information) ∨ (¬ Retrieval)

You have:
Data: Doc3, Doc4
Retrieval: Doc1, Doc2
Information: Doc1, Doc2, Doc3, Doc4

Then:
Data: Doc3, Doc4
¬ Retrieval: Doc3, Doc4
Information: Doc1, Doc2, Doc3, Doc4

Then:
Data ∧ Information: Doc3, Doc4
¬ Retrieval: Doc3, Doc4

Then Result is:
(Data ∧ Information) ∨ (¬Retrieval): Doc3, Doc4

2. An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia - https://ieeexplore.ieee.org/document/5076348

Makalah ini menyelidiki kinerja pencarian semantik mesin pencari. Awalnya, tiga mesin pencari berbasis kata kunci (Google, Yahoo dan Msn) dan mesin pencari semantik (Hakia) dipilih.

Kemudian, sepuluh kueri, dari berbagai topik, dan empat frasa, yang memiliki sintaks berbeda tetapi artinya serupa, ditentukan. Setelah setiap permintaan dijalankan di setiap mesin pencari; dan setiap frasa yang berisi kueri dijalankan di mesin telusur semantik, dua puluh dokumen pertama pada setiap keluaran pengambilan diklasifikasikan sebagai "relevan" atau "tidak relevan". Setelah itu, presisi dan rasio penarikan yang dinormalisasi dihitung pada berbagai titik potong untuk mengevaluasi mesin pencari berbasis kata kunci dan mesin pencari semantik. Secara keseluruhan, Yahoo menunjukkan kinerja terbaik dalam hal rasio presisi, sedangkan Google ternyata menjadi mesin pencari terbaik dalam hal rasio penarikan yang dinormalisasi. Namun, ditemukan bahwa kinerja pencarian semantik mesin pencari rendah untuk mesin pencari berbasis kata kunci dan mesin pencari semantik.

Rnorm dihitung pada empat titik cut-off (cut-off 5, cut-off 10, cut-off 15 dan cut-off 20) untuk mendapatkan nilai yang sejajar dengan yang untuk presisi. Formula 1, yang dikemukakan oleh Bollmann et al. [17], digunakan untuk menghitung nilai penarikan yang dinormalisasi di berbagai titik potong.

$$R_{norm}(\Delta) = \frac{1}{2}\left(1 + \frac{R^{+} - R^{-}}{R^{+}_{\max}}\right)$$

Rumus normalized recall, dimana:

R+ adalah jumlah pasangan dokumen yang memiliki dokumen relevan dengan peringkat lebih tinggi dari dokumen tidak relevan

R− adalah jumlah pasangan dokumen yang memiliki dokumen tidak relevan berperingkat lebih tinggi dari yang relevan

Rmax+ memberikan jumlah maksimum R+