SISTEM TEMU KEMBALI INFORMASI

Classic Information Retrieval Model

Extended Boolean

Dosen:

I Made Suwija Putra, ST., MT.

Oleh:

Putu Bagus Candradinatha 1605551105

JURUSAN TEKNOLOGI INFORMASI

FAKULTAS TEKNIK

UNIVERSITAS UDAYANA

2018

BAB I
PENDAHULUAN

1.1. Latar Belakang

Information Retrieval (IR) pada dasarnya adalah keadaan untuk menentukan dokumen mana dalam kumpulan dokumen yang harus diambil untuk memenuhi kebutuhan user akan informasi. Kebutuhan informasi user ditampilkan melalui *query* atau *profile*, dan mengandung satu atau lebih kata pencarian, ditambah mungkin beberapa informasi tambahan seperti bobot dokumen. Karena itu, keputusan pencarian dilakukan melalui membandingkan kata dari *query* dengan *index* kata (kata penting atau ungkapan) yang tampil dalam dokumen itu sendiri. Salah satu cara untuk mendapatkan informasi dari dokumen yang dibutuhkan dalam sistem temu kembali informasi adalah dengan mengimplementasikan teknik extended boolean sesuai P-Norm. Teknik ini akan mengurutkan dokumen yang diinginkan dan nilai P-Norm akan menentukan di urutan ke berapakah dokumen tersebut berada.

1.2. Tujuan

Adapun tujuan dari penulisan ini adalah untuk mempelajari salah satu teknik yang terdapat pada *information retrieval* yaitu model *extended Boolean* dan untuk memenuhi penugasan daripada mata kuliah Sistem Temu Kembali Informasi.

1.3. Manfaat

Penulisan ini akan meningkatkan wawasan dan pengetahuan penulis mengenai *information retrieval system*.

BAB II
TINJAUAN PUSTAKA

2.1. Sistem Temu Kembali Informasi

Sistem temu-kembali informasi pada prinsipnya adalah suatu sistem yang sederhana. Misalkan ada sebuah kumpulan dokumen dan seorang *user* yang memformulasikan sebuah pertanyaan (*request* atau *query*). Jawaban dari pertanyaan tersebut adalah sekumpulan dokumen yang relevan dan membuang dokumen yang tidak relevan. Sistem temu-kembali informasi pada dasarnya dibagi dalam dua komponen utama yaitu sistem pengindeksan (*indexing*) yang menghasilkan basis data sistem dan temu kembali yang merupakan gabungan dari *user interface* dan *look-up-table*. Ada beberapa teknik temu-kembali informasi yang telah dikembangkan salah satunya yaitu teknik *Extended Boolean*. Untuk lebih jelasnya, penjelasan mengenai teknik *Extended Boolean* dapat dilihat pada penjelasan berikut.

2.2. Standard Boolean Model

Dalam model Boolean standar untuk sistem temu kembali informasi, *query* yang digunakan berupa Boolean ekspresi yang mana terdiri dari satu set istilah indeks yang dihubungkan oleh operator Boolean AND, OR, dan NOT.

Dokumen yang diambil berdasarkan *query* yang diberikan adalah yang berisi istilah indeks dalam kombinasi yang ditentukan oleh *query*. Model ini bersifat biner di mana 1 berarti benar atau cocok dan 0 berarti tidak cocok. Untuk mengilustrasikan, pertimbangkan kasus dua kondisi query seperti [q1 ATAU q2] dan [q1 DAN q2]. Gambar 2.1 menunjukkan bahwa tiga kelas dokumen diambil dengan *query*: yang cocok dengan kedua kondisi, yang hanya cocok dengan salah satu kondisi, dan yang tidak cocok dengan semua kondisi.

		Query terms		Similarity with query	
	document	q_1	q_2	[q_1 OR q_2]	[q_1 AND q_2]
Class 1	d_1	1	1	1	1
Class 2	d_2	1	0	1	0
	d_3	0	1	1	0
Class 3	d_4	0	0	0	0

Gambar 2. 1 Boolean Model for Information Retrieval

Query [q1 OR q2] mengasumsikan bahwa dokumen Class 1 dan 2 sama pentingnya dan memiliki kesamaan dokumen-*query* yang sama. Sebaliknya kueri [q1 AND q2] mengasumsikan bahwa hanya Class 1 yang termasuk dokumen penting sementara dokumen Kelas 2 dan 3 tidak berguna.

2.3. Extended Boolean Model

Dalam tesisnya pada tahun 1981, Wu menerapkan konsep *p-norm* ke *Information Retrieval*. Beberapa tahun kemudian, dalam tesisnya pada 1983, Fox menggunakan *p-norm* untuk memperluas model Boolean dan Vector Space, mengembangkan apa yang saat ini dikenal sebagai Extended Boolean Model (Fox, 1983; Salton, Fox, & Wu, 1983a; 1983b; Wu, 1981; Salton, Buckley, & Fox, 1983). Aspek *practical* dari model tersebut adalah subyek dari thesis Smith 1990 (Smith, 1990).

Model Extended Boolean berdasarkan model P-norm merupakan pengembangan lebih lanjut dari model Boolean. Teknik ini memakai operator yang dikomputasi berdasarkan rumus Savoy (1993). Rumus masing-masing operator dapat dilihat pada tabel.

Query	*Retrieval Status Value* (RSV)
A or <p> B	$RSV_{or} = \sqrt[p]{\frac{W_{ia}^{\ p} + W_{ib}^{\ p}}{2}}$
A and <p> B	$RSV_{and} = 1 - \sqrt[p]{\frac{(1-W_{ia})^p + (1-W_{ib})^p}{2}}$
Not A	$RSV_{not} = 1 - W_{ia}$

Gambar 2. 2 Implementasi Teknik Extended Boolean P-Norm Model

Dimana:

- P adalah nilai p-norm yang dimasukkan pada *query*.
- W_{ia} adalah bobot istilah A dalam indeks pada dokumen D_i.
- W_{ib} adalah bobot istilah B dalam indeks pada dokumen D_i.

Untuk menentukan peringkat dapat dilakukan dengan dua cara, yaitu:

- Langsung mengurutkan dokumen (dari besar ke kecil) berdasarkan bobot dokumen yang didapat dengan rumus RSV (*retrieval status value*).
- Memakai rumus *Learning Scheme*.

$$RSV(D_i) = RSV_{init}(D_i) + \sum_{k=1}^{r} \alpha_{ik} \int \text{norm} * RSV_{init}(D_k)$$

Gambar 2. 3 Rumus *Learning Scheme*

Untuk i = 1, 2, ..., n

dimana:

a. $RSV_{init}(D_i)$ merupakan *retrieval status value* dari dokumen Di yang dikomputasi berdasarkan rumus teknik *retrieval* P-norm model.

b. $\propto_{ik}$ merupakan bobot keterhubungan antara dokumen i dan k. Bobot ketergantungan ini diperoleh dari nilai *relevance* link yang merupakan hasil dari proses pembelajaran.

2.4 Analisa Metode

Contoh sederhana berikut dapat digunakan untuk membantu pemahaman mengenai prosedur kerja dari metode Savoy, misalkan terdapat dua buah dokumen yang akan dilakukan pembobotan dan pengindeksan. Isi dari kedua dokumen yang digunakan tersebut dapat dirincikan sebagai berikut.

- **File 1**

  Penelitian di Florida telah menemukan bahwa senyawa alami dalam buah semangka, yaitu citrulline, yang membantu melebarkan dan merilekskan pembuluh darah, sehingga jantung tidak harus bekerja keras untuk memompa darah ke seluruh tubuh. Penelitian lain dari University Kentucky juga menunjukkan bahwa semangka baik dikonsumsi untuk menjaga berat badan dan membantu kesehatan jantung.

- **File 2**

  Melon dan semangka memiliki kandungan air yang cukup banyak sehingga baik untuk mengatasi dehidrasi. Sayangnya, kandungan gula dalam keduanya cukup tinggi (Indeks glikemik mencapai 65-100 persen), sehingga sangat tidak disarankan bagi mereka yang memiliki diabetes atau memiliki riwayat keluarga penderita diabetes untuk mengonsumsi buah ini.

Misalkan kata yang ingin dicari adalah: ‘melon’ dan ‘semangka’, maka proses perhitungannya adalah sebagai berikut:

a. Diketahui:

Frekuensi kata pada File 1 dan File 2

File 1		File 2	
Kata	Frekuensi	Kata	Frekuensi
florida	1	melon	1

temu	1	semangka	1
senyawa	1	milik	2
alami	1	kandung	2
buah	1	air	1
semangka	2	cukup	2
citrulline	1	banyak	1
bantu	1	baik	1
lebar	1	Atas	1
rileks	1	Dehidrasi	1
pembuluh	1	Gula	1
darah	2	Dua	1
jantung	2	Tinggi	1
kerja	1	Indeks	1
keras	1	Glikemik	1
pompa	1	Capai	1
tubuh	1	Persen	1
teliti	1	65	1
university	1	100	1
kentucky	1	Saran	1
tunjuk	1	Mereka	1
baik	1	diabetes	2
konsumsi	1	Riwayat	1
jaga	1	Keluarga	1
berat	1	derita	1
badan	1	konsum	1
sehat	1	buah	1

- Jumlah File (frekuensi dokumen yang mengandung suatu term):
  melon = 1 (df1)
  semangka = 2 (df2)
- File 1:
  melon = 0
  semangka = 2
- File 2:
  melon = 1
  semangka = 1
- Jumlah keseluruhan File (n): 2
  Nilai p = 100 (peng-*input*-an nilai p-norm memiliki batasan maksimal 3 digit bilangan bulat positif)

b. Proses Penyelesaian:

- File 1 (untuk kata melon):
  log(n/dfk) = log (2/1) = 0.3010299956639812

$$\text{nidf}_k = \frac{\log\left[\frac{n}{df_k}\right]}{\log(n)}$$

  nidfk1 = 0.3010299956639812 / log (2) = 1
  keterangan:
  a. n adalah jumlah dokumen dalam kumpulan dokumen.
  b. dfk adalah jumlah dokumen yang mengandung istilah k.
  c. nidfk1 adalah normalisasi jumlah dokumen yang mengandung istilah k.

$$ntf_{ik} = \frac{tf_{ik}}{Max_j tf_{ij}}$$

ntfk1 = 0/2 = 0

keterangan:

a. tf_{ik} adalah frekwensi dari istilah k dalam dokumen i.
b. ntfk1 adalah normalisasi frekuensi dari istilah k dalam dokumen i.
c. $Max_j tf_{ij}$ adalah frekuensi istilah terbesar pada satu dokumen.
d. W_{ik} adalah bobot dari istilah k dalam dokumen i.

$W_{ik} = ntfi_{ik} * nidf_k$

w1 = 0 x 1 = 0

- File 1 (untuk kata semangka)

log (n/dfk) = log (2/2) = 0

dengan rumus yang sama maka didapatkan,

nidfk2 = 0 / log (2) = 0

ntfk1 = 2/2 = 1

w2 = 0.1 x 0 = 0

$RSV = 1 – (((1 – w1)^p + (1 – w2)^p) / 2)^{1/np}$

$RSV = 1 – (((1 – 0)^{100} + (1 – 0)^{100} / 2)^{1/100}$

RSV = 0

Keterangan:

a. $RSV_{init}(Di)$ merupakan retrieval status value dari dokumen Di yang dikomputasi berdasarkan rumus teknik retrieval P-norm model.
b. $\propto_{ik}$ merupakan bobot keterhubungan antara dokumen i dan k. Bobot ketergantungan ini didapat dari nilai relevance link yang merupakan hasil dari proses pembelajaran.

- File 2 (untuk kata melon):

  log(n/dfk) = log (2/1) = 0.3010299956639812

  dengan rumus yang sama maka didapatkan,

  nidfk1 = 0.3010299956639812 / log (2) = 1

  ntfk1 = 2/2 = 1

  w1 = 1 x 1 = 1

- File 2 (untuk kata semangka):

  log(n/dfk) = log (2/2) = 0

  dengan rumus yang sama maka didapatkan,

  nidfk2 = 0 / log (2) = 0

  ntfk2 = 1/2 = 0.5

  w2 = 0.5 x 0 = 0

  $RSV = 1 - (((1 - w1)^{p} + (1 - w2)^{p}) / 2)^{1/np}$

  $RSV = 1 - (((1 - 1)^{100} + (1 - 0)^{100} / 2)^{1/100}$

  RSV = 1

Berdasarkan nilai RSV yang diperoleh maka dokumen ditampilkan pada sistem temu kembali hanyalah dokumen kedua, karena nilai dari RSV pada dokumen pertama bernilai 0.

BAB III
PENUTUP

3.1 Kesimpulan

Setelah menyelesaikan penulisan pada laporan ini, maka dapat ditarik beberapa kesimpulan berikut:

1. Dengan menggunakan *information retrieval system*, maka dapat dilakukan pencarian dokumen yang relevan dengan kata kunci yang dimasukkan.
2. Operator "OR" akan menghasilkan nilai indeks yang jauh lebih besar daripada nilai indeks yang dihasilkan oleh operator "AND". Dengan menggunakan operator "OR" maka hasil yang diperoleh merupakan dokumen yang memiliki salah satu atau kedua kata yang dicari.
3. Operator "AND" akan menghasilkan nilai indeks yang lebih kecil daripada nilai indeks yang dihasilkan oleh operator "OR". Dengan menggunakan operator "AND" maka hasil yang diperoleh merupakan dokumen yang memiliki kedua kata yang dicari.
4. Untuk operator "AND" pada metode Savoy, semakin tinggi nilai p menyebabkan nilai bobot dokumen semakin menurun.

DAFTAR PUSTAKA

[1] The Extended Boolean Model http://www.minerazzi.com/tutorials/term-vector-6.pdf diakses pada 16 September 2018

[2] Perancangan Information Retrieval System dengan Metode Extended Boolean dan Savoy http://download.portalgaruda.org/article.php?article=396937&val=8676&title=Perancangan%20Information%20Retrieval%20System%20Dengan%20Metode%20Extended%20Boolean%20Dan%20Savoy diakses pada 17 September 2018