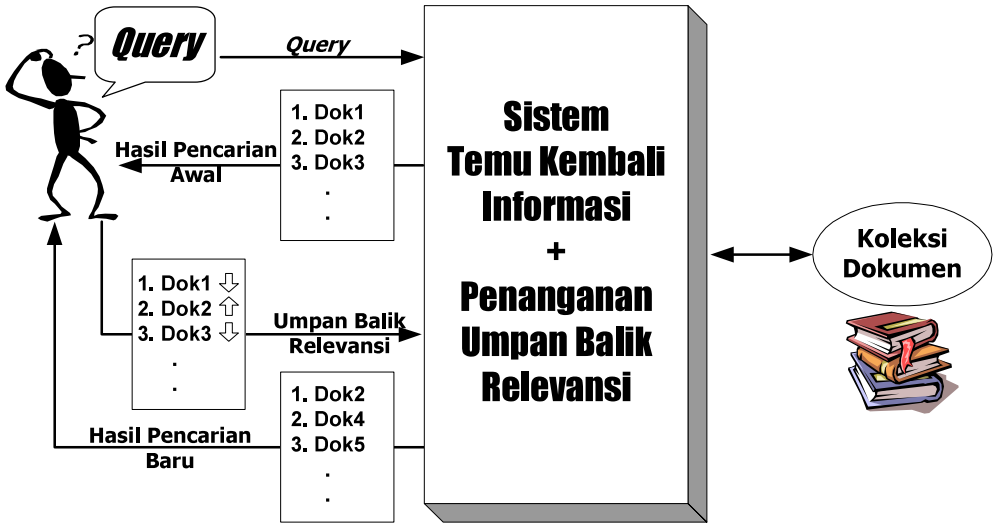
UMPAN BALIK RELEVANSI

1.1 Pengantar Umpan Balik Relevansi

Mendapatkan hasil pencarian yang sesuai dengan kebutuhan dalam suatu koleksi dokumen yang besar merupakan hal sulit. Usaha pengguna secara manual untuk memilah-milah dokumen yang sesuai dengan kebutuhannya ternyata sangat besar. Hasil pencarian merupakan sejumlah dokumen yang relevan menurut sistem, namun relevansi merupakan hal yang subjektif.

Query yang baik adalah *query* yang mampu merangkum kebutuhan informasi pengguna. Kunci pencarian yang tepat adalah formulasi *query* yang baik dan sesuai. Namun bagi kebanyakan pengguna, memformulasikan *query* yang baik tidak mudah. Karena sangat bergantung berbagai faktor seperti latar belakang pengetahuan pengguna terhadap koleksi dokumen, lingkungan sistem temu kembali informasi, maupun pengetahuan pengguna mengenai koleksi dokumen maupun topik kebutuhan yang dicari..

Penanganan umpan balik relevansi merupakan proses formulasi ulang *query* awal berdasarkan informasi umpan balik relevansi dari pengguna terhadap dokumen-dokumen hasil pencarian awal. Berdasarkan umpan balik, sistem secara otomatis akan menentukan *query* baru dan melakukan pencarian berdasarkan *query* baru tersebut. Proses umpan balik diulang pengguna menilai bahwa kebutuhannya sudah terpenuhi.

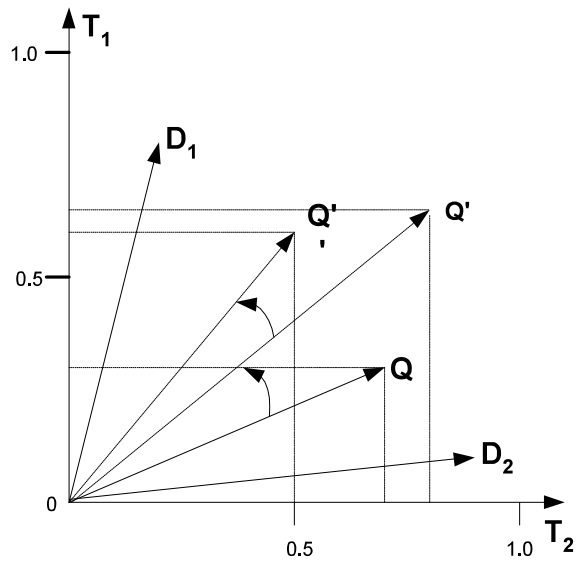


Gambar 3.1 Ilustrasi umpan balik relevansi

Metode umpan balik relevansi yang ada sangat beragam, berbagai aspek ditekankan oleh masing-masing metode. Secara umum metode umpan balik pada model ruang vektor mengandung dua unsur yaitu pembobotan ulang kata dalam *query* (*query reweighting*) dan perluasan *query* (*query expansion*). Pada pembobotan ulang *query*, sistem melakukan perhitungan ulang bobot kata-kata pada *query*. Pada perluasan *query*, *query* awal ditambahkan sejumlah kata yang berasal dari dokumen-dokumen yang relevan.

1.2 Umpan Balik Relevansi dalam Model Ruang Vektor

Secara umum umpan balik relevansi pada model ruang vektor dapat dijelaskan sebagai penggeseran vektor *query* mendekati vektor dokumen relevan dan menjauhi vektor dokumen tidak relevan.



Gambar 3.2 Ilustrasi Umpan Balik Relevansi pada Model Ruang Vektor

Sebagai ilustrasi, sebuah *query* Q dan dua buah dokumen D_1 dan D_2, masing-masing terdiri dari dua kata yaitu T_1 dan T_2, D_1 adalah dokumen relevan sedangkan D_2 adalah dokumen tidak relevan. Pengaruh dokumen relevan D_1 akan menggeser Q menjadi Q' dengan penambahan bobot dari pada setiap kata. Pengaruh dokumen tidak relevan terhadap Q' (vektor *query* hasil pengaruh dokumen relevan D_1) akan menggeser Q' menjadi Q'' yaitu dengan mengurangi bobot.

1.2.1 *Metode Rocchio*

Diasumsikan terdapat sekumpulan dokumen D_R yang merupakan bagian dari koleksi dokumen D. Kumpulan dokumen D_R ini merupakan kumpulan dokumen relevan terhadap *query* Q. Dengan telah terdefinisinya kumpulan dokumen relevan $D_{R,}$, maka suatu *query* Q_{opt} optimal dapat ditentukan. *Query* Q_{opt}

akan menyebabkan kumpulan dokumen D_R memiliki rangking lebih besar atau memiliki kesesuaian yang lebih besar daripada dokumen-dokumen lain dalam koleksi *D*.

Menurut *Rocchio* [4], *query* optimal Q_{opt} adalah *query* yang memaksimalkan perbedaan antara rata-rata kesesuaian dokumen-dokumen relevan (anggota dari D_R) dan rata-rata kesesuaian dokumen-dokumen tidak relevan (anggota *D* yang bukan anggota D_R). Secara matematis vektor *query* optimal *Q* menurut $D_R \subset D$ adalah :

$$C = \frac{1}{n_0} \sum_{Di \subset DR} sim(Q, Di) - \frac{1}{n - n_0} \sum_{Di \not\subset DR} sim(Q, Di) \quad \text{.....(3.1)}$$

dimana n_0 adalah jumlah dokumen dalam kumpulan dokumen D_R dan $n = n(D)$ yaitu jumlah total dokumen dalam koleksi.

Bila kesesuaian persamaan *sim(Q,D)*, disubstitusikan pada persamaan diatas maka persamaan menjadi :

$$C = \frac{1}{n_0}\left(\frac{Q}{|Q|}\right) \bullet \sum_{Di \subset DR} \frac{Di}{|Di|} - \frac{1}{n - n_0}\left(\frac{Q}{|Q|}\right) \bullet \sum_{Di \not\subset DR} \frac{Di}{|Di|} \quad \text{.....(3.2)}$$

atau $$C = \left(\frac{Q}{|Q|}\right) \bullet \left[\frac{1}{n_0} \sum_{Di \subset DR} \frac{Di}{|Di|} - \frac{1}{n - n_0} \sum_{Di \not\subset DR} \frac{Di}{|Di|}\right] \quad \text{.....(3.3)}$$

Persamaan terakhir ini ekuivalen dengan persamaan $C = Q^* \bullet A$, dengan Q^* adalah suatu vektor unit dan $A = \left[\frac{1}{n_0} \sum_{Di \subset DR} \frac{Di}{|Di|} - \frac{1}{n - n_0} \sum_{Di \not\subset DR} \frac{Di}{|Di|}\right]$ .....(3.4)

Dapat disimpulkan bahwa $Q_{opt} = kA$ (*k* adalah nilai sembarang), atau

$$Qopt = \frac{1}{n_0} \sum_{Di \subset DR} \frac{Di}{|Di|} - \frac{1}{n - n_0} \sum_{Di \not\subset DR} \frac{Di}{|Di|} \quad \text{.....(3.5)}$$

Persamaan ini tidak dapat digunakan karena pada saat pencarian awal, kumpulan dokumen relevan tidak diketahui. Maka umpan balik relevansi digunakan untuk mendekatkan vektor *query* awal ke vektor *query* optimal. Metode umpan balik relevansi yang diajukan oleh *Rocchio* adalah

$$Q1 = Q0 + \beta \sum_{k=1}^{n1} \frac{Rk}{n1} - \gamma \sum_{k=1}^{n2} \frac{Sk}{n2} \quad \text{.....(3.6)}$$

dimana Q_1 adalah vektor *query* baru

Q_0 adalah vektor *query* awal

R_k adalah vektor dokumen yang relevan ke k

S_k adalah vektor dokumen yang tidak relevan ke k

n_1 adalah jumlah dari dokumen yang relevan

n_2 adalah jumlah dari dokumen yang tidak relevan

Parameter β dan γ yang menentukan kontribusi dokumen-dokumen relevan dan dokumen-dokumen tidak relevan. Perluasan *query* dapat dilakukan dengan memberi nilai 0 untuk vektor *query* awal pada persamaan.

1.2.2 *Metode Ide*

Setelah *Rocchio* mengusulkan metode umpan baliknya, kemudian *E. Ide* [6] melakukan eksperimen dengan merubah persamaan pada metode *Rocchio*. Kemudian *Ide* melakukan eksperimen terhadap berbagai koleksi dokumen dan membandingkan berbagai metode dan strategi umpan balik relevansi. Metode yang digunakan oleh *Ide* adalah

Ide Regular : $$Q1 = Q0 + \sum_{k=1}^{n1} Rk - \sum_{k=1}^{n2} Sk \quad \text{.....(3.7)}$$

Ide dec-hi : $$Q1 = Q0 + \sum_{k=1}^{n1} Rk - Sk \quad \text{.....(3.8)}$$

dimana Q_1 adalah vektor *query* baru

Q_0 adalah vektor *query* awal

R_k adalah vektor dokumen yang relevan ke k

S_k adalah vektor dokumen yang tidak relevan ke k

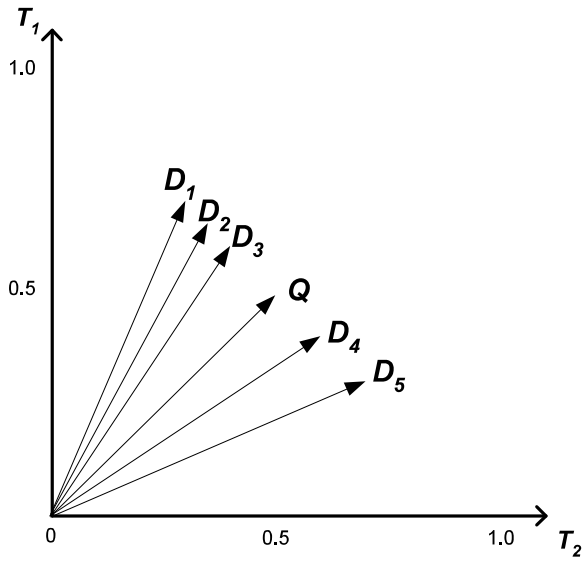
Kedua metode ini menghilangkan faktor normalisasi terhadap jumlah umpan balik pada metode *Rocchio*. Metode *Ide dec-hi* hanya menggunakan sebuah dokumen tidak relevan pada rangking teratas sebagai umpan balik. Hal ini dimaksudkan untuk mengurangi kontribusi dokumen tidak relevan terhadap *query* yang dihasilkan. Sedangkan metode *Ide Regular* menggunakan seluruh dokumen tidak relevan sebagai umpan balik. Sama seperti metode *Rocchio*, perluasan *query* dilakukan dengan menganggap bahwa bobot kata baru tersebut pada *query* awal bernilai 0.

Berikut adalah contoh dari formulasi ulang *query* dengan menggunakan metode *Rocchio*, *Ide Regular* dan *Ide dec Hi*. Misalkan pengguna memberikan umpan balik terhadap hasil pencarian berdasarkan *query Q* berupa 3 buah dokumen relevan D_R dan 2 buah dokumen tidak relevan D_{NR}. D_R terdiri dari D_1, D_2 dan D_3 sedangkan D_{NR} terdiri dari D_4 dan D_5.

Query Q terdiri dari dua kata yaitu T_1 dan T_2. Bobot T_1 dan T_2 pada masing-masing vektor adalah tertera pada tabel 3.1 dan gambaran vektor-vektor pada ruang vektor pada gambar 3.3.

Tabel 3.1 Bobot vektor untuk contoh pengananan umpan balik relevansi

Vektor	Bobot T_1	Bobot T_2
Q	0.5	0.5
D_1	0.3	0.7
D_2	0.35	0.65
D_3	0.4	0.6
D_4	0.6	0.4
D_5	0.7	0.3

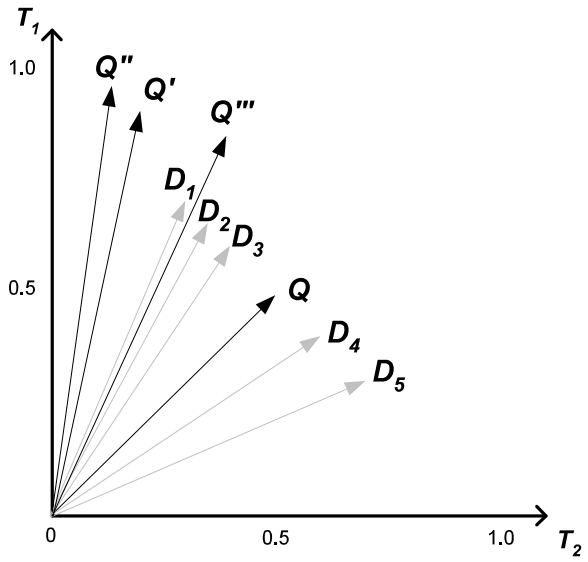


Gambar 3.3 Vektor dokumen dan *query* pada contoh

Tabel 3.2 Contoh perhitungan formulasi ulang *query*

Rocchio (Q`)	T_1	= 0.5 + (0.3+0.35+0.4)/3 – (0.6+0.7)/2 = 0.2
	T_2	= 0.5 + (0.7+0.65+0.6)/3 - (0.4+0.3)/2 = 0.8
Ide Regular (Q``)	T_1	= 0.5 + (0.3+0.35+0.4) – (0.6+0.7) = 0.25
	T_2	= 0.5 + (0.7+0.65+0.6) - (0.4+0.3) = 1.75
Ide dec Hi (Q```)	T_1	= 0.5 + (0.3+0.35+0.4) – 0.6 = 0.95
	T_2	= 0.5 + (0.7+0.65+0.6) - 0.4 = 2.05

Formulasi ulang *query* baru berdasarkan umpan balik dapat dilihat pada tabel 3.2. Gambaran *query* hasil formulasi ulang tersebut dalam ruang vektor adalah seperti pada gambar 3.4. Dari gambar 3.4 terlihat bahwa penggeseran membuat vektor *query* mendekati kumpulan dokumen relevan dan menjauhi dokumen tidak relevan.



Gambar 3.4 Gambaran penggeseran vektor *query* dalam ruang vektor