Optimasi Pembobotan pada *Query Expansion* dengan *Term Relatedness to Query-Entropy based* (TRQE)

Resti Ludviani[1], Khadijah F. Hayati[2], Agus Zainal Arifin[3], Diana Purwitasari[4]
Program Studi Teknik Informatika, Fakultas Teknik Informatika, Institut Teknologi Sepuluh November
Jl. Sukolilo, Surabaya 60111, Jawa Timur
E-mail: [1]restiludvi@gmail.com, [2]dee.jafa@gmail.com, [3]agusza@cs.its.ac.id, [4]diana@if.its.ac.id

Masuk: 31 Desember 2014; Direvisi: 20 Januari 2015; Diterima: 21 Januari 2015

Abstract. *An appropriate selection term for expanding a query is very important in query expansion. Therefore, term selection optimization is added to improve query expansion performance on document retrieval system. This study proposes a new approach named Term Relatedness to Query-Entropy based (TRQE) to optimize weight in query expansion by considering semantic and statistic aspects from relevance evaluation of pseudo feedback to improve document retrieval performance. The proposed method has 3 main modules, they are relevace feedback, pseudo feedback, and document retrieval. TRQE is implemented in pseudo feedback module to optimize weighting term in query expansion. The evaluation result shows that TRQE can retrieve document with the highest result at precission of 100% and recall of 22,22%. TRQE for weighting optimization of query expansion is proven to improve retrieval document.*
Keywords: *TRQE, query expansion, term weighting, term relatedness to query, relevance feedback*

Abstrak..*Pemilihan term yang tepat untuk memperluas queri merupakan hal yang penting pada query expansion. Oleh karena itu, perlu dilakukan optimasi penentuan term yang sesuai sehingga mampu meningkatkan performa query expansion pada system temu kembali dokumen. Penelitian ini mengajukan metode Term Relatedness to Query-Entropy based (TRQE), sebuah metode untuk mengoptimasi pembobotan pada query expansion dengan memperhatikan aspek semantic dan statistic dari penilaian relevansi suatu pseudo feedback sehingga mampu meningkatkan performa temukembali dokumen. Metode yang diusulkan memiliki 3 modul utama yaitu relevan feedback, pseudo feedback, dan document retrieval. TRQE diimplementasikan pada modul pseudo feedback untuk optimasi pembobotan term pada ekspansi query. Evaluasi hasil uji coba menunjukkan bahwa metode TRQE dapat melakukan temukembali dokumen dengan hasil terbaik pada precision 100% dan recall sebesar 22,22%.Metode TRQE untuk optimasi pembobotan pada query expansion terbukti memberikan pengaruh untuk meningkatkan relevansi pencarian dokumen.*
Kata Kunci: *TRQE, ekspansi query, pembobotan term, term relatedness to query, relevance feedback*

1. Pendahuluan

Jumlah informasi yang tersedia secara elektronik meningkat secara dramatis. Dalam melakukan pencarian informasi membutuhkan metode untuk mengidentifikasi dokumen yang relevan terhadap *query user* (Boston, 2014). Hanya saja, *query user* untuk me-*retrieve* dokumen sering kali terlalu pendek atau ambigu (Araujo, 2010). Aroujo menggunakan teknik *query reformulation* untuk meningkatkan hasil temu kembali (*retrieval*).

Ekspansi *query* (*query expansion*) dikenal juga dengan pengayaan *query* (*query enrichment*) digunakan untuk meningkatkan kinerja *retrieval* (Saneifar, 2014). Cara yang digunakan dengan merumuskan dan menambahkan *term* ke *query* awal (*initial query*) untuk memperjelas *query*. Sehingga diharapkan mampu menangani masalah ketidakjelasan *query* (*disambiguate query*). Diantara teknik ekspansi *query* yang mulai dikenalkan pada pertengahan 1960-an adalah *relevance feedback* (umpan balik relevansi). *Relevance feedback* adalah teknik

dimana pengguna dapat memberikan informasi relevansi pada dokumen tertentu atau *term query* agar sistem menemukan tambahan dokumen yang relevan.

Terdapat beberapa jenis teknik *relevance feedback*, diantaranya adalah jenis eksplisit. Jenis eksplisit *feedback* dapat digunakan ketika penilai secara langsung mengetahui bahwa *feedback* yang diberikan relevan (Carpineto, 2001). Hal ini mengharuskan penilai memiliki informasi yang cukup. Hasil *retrieval* kurang sesuai jika penilai tidak dapat menilai *feedback* yang relevan.

Saneifar, dkk (2014) mengajukan pendekatan *query expansion* berbasis eksplisit dan *pseudo relevance feedback* yang baru dengan pembobotan yang disebut dengan TRQ (*term relatedness to query*). Pembobotan tersebut memberikan nilai pada *term* berdasarkan hubungannya dengan *query*. Metode TRQ memperhatikan aspek semantik dan statistik dengan menggabungkan *lexical word frequency* (lwf) dengan *invers document frequency* (idf). Saneifar juga mengembangkan metode tersebut menjadi TRQ_{ext}. Hasil implementasi metode TRQ pada *log file* menunjukkan bahwa metode yang diajukan mampu meningkatkan performa *passage retrieval*.

Wu, dkk. (2013) menyatakan bahwa entropi dapat digunakan sebagai pengukuran seleksi fitur. Pada penelitiannya, mereka mengajukan pembobotan berbasis entropi dan memberikan hasil yang lebih baik dibandingkan dengan pembobotan TF.IDF. Pendekatan *query expansion* berbasis entropi mempertimbangkan distribusi *term* yang muncul dalam dokumen yang relevan dan tidak relevan hingga sampai pada nilai diskriminasi *term.* Oleh karena itu, dibutuhkan optimasi *pseudo relevance feedback* pada *query expansion* menggunakan konsep entropi untuk meningkatkan performa dokumen *retrieval*. Pada penelitian ini diajukan sebuah metode optimasi pembobotan yang memperhatikan aspek semantik dan statistik dari penilaian relevansi *pseudo feedback* sehingga mampu meningkatkan performa temu kembali dokumen. Metode yang diajukan ini disebut dengan TRQE (*Term Relatedness to Query – Entropy based*).

2. Tinjauan Pustaka

Manning (2008) menjelaskan dua tipe metode *Query Expansion*, yaitu: (1) global dan (2) teknik lokal. Metode global membutuhkan adanya *lexical-semantic* seperti ontologi. Metode global banyak digunakan dalam *question answering* (QA) dimana QA menggunakan basis pengetahuan untuk mengidentifikasi variasi leksikal dari *term* pertanyaan. Variasi tersebut digunakan untuk menganalisis dan mengambil jawaban yang relevan. Pada tahap ini, sistem biasanya melibatkan pengetahuan morfologi dan semantik dari kamus elektronik yang ada dan sumber informasi leksikal seperti WordNet. Pasca (2001) mengajukan metode *query expansion* dimana *query* diperluas menggunakan morfologi, derivasi leksikal, dan kesamaan semantik semacam sinonim yang sangat bergantung pada sumber informasi (*resource*). Selain itu, Agichtein (2001) mengusulkan *query expansion* berbasis *web resources*. Mereka mengambil sejumlah *keyword* dari pertanyaan awal untuk membentuk *query* yang diperluas dengan frasa yang cenderung terjadi dalam kalimat deklaratif yang mengandung jawaban.

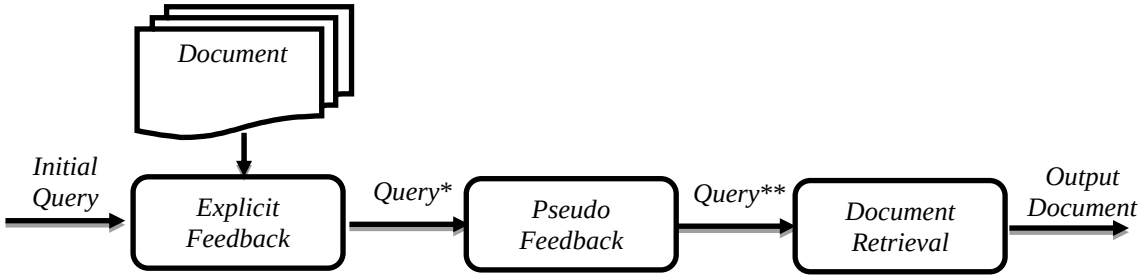
Metode lokal, dikenal dengan *relevance feedback*, merujuk pada suatu proses interaktif yang membantu untuk meningkatkan performa temu kembali (Saneifar, 2014). Pada metode ini, ketika pengguna memasukkan *query*, sistem temu kembali informasi mengembalikan set awal dokumen hasil pencarian kemudian meminta pengguna untuk menilai apakah beberapa dokumen relevan atau tidak. Setelah itu, sistem merumuskan *query* berdasarkan penilaian pengguna, dan mengembalikan satu set hasil baru. Ada tiga jenis *relevance feedback* (Saneifar, 2014), yaitu: eksplisit, *pseudo*/semu (*blind*), dan implisit. Jenis eksplisit feedback dapat digunakan ketika penilai secara langsung mengetahui bahwa *feedback* yang diberikan relevan (Carpineto, 2001). Ketika tidak ada penilaian relevansi yang tersedia, sebagai alternatif, *pseudo relevance feedback* atau *blind relevance feedback* dapat dilakukan dengan mengasumsikan bahwa sejumlah kecil dokumen peringkat teratas (*top-rank*) dalam hasil pencarian awal yang relevan dan kemudian menerapkan umpan balik relevansi berdasarkan asumsi tersebut. Metode ini mampu melakukan analisis local secara otomatis. Selain *relevance feedback* dan *psudo relevance feedback*, terdapat *implicit feedback*, dimana tidak ada informasi langsung yang

menunjukkan relevansi dokumen. Pada implisit *feedback* tindakan atau perilaku pengguna dalam berinteraksi dengan sistem digunakan untuk menyimpulkan informasi kebutuhan pengguna. Perilaku pengguna tersebut misalkan lamanya waktu yang digunakan pengguna untuk suatu dokumen. (Saneifar, 2014)

Xu dan Croft (2000) memperkenalkan metode *query expansion* dengan gabungan metode lokal (*relevance feedback*) dan metode global, yang disebut *local context analysis*. Pada metode yang mereka ajukan, pemilihan *term* ekspansi ditingkatkan dengan mempertimbangkan konsep dalam dokumen peringkat atas (*top-rank*) yang sering memuat banyak *term query* di seluruh koleksi. Dibandingkan dengan *relevance feedback* klasik, kandidat *term* ekspansi lebih relevan dengan *query*, karena *term* tersebut telah diteliti sering muncul pada dokumen.

3. Metode yang diusulkan

Secara garis besar, tahapan *query expansion* dapat dilihat pada Gambar 1. *Query* awal yang dimasukkan oleh pengguna diperluas menggunakan langkah *explicit feedback.* Pada tahap ini pengguna dapat menentukan perluasan *term* yang sesuai untuk menjadi *query* baru (*query**). *Query** diperluas lagi menggunakan langkah *pseudo feedback*. Secara otomatis, sistem menentukan *term* yang sesuai untuk perluasan *query* sehingga menjadi *query***. Pada tahap akhir, pencarian dokumen dilakukan berdasarkan *query* yang telah diekspasi yaitu *query***. Hasil akhir yang diperoleh adalah output dokumen yang relevan.



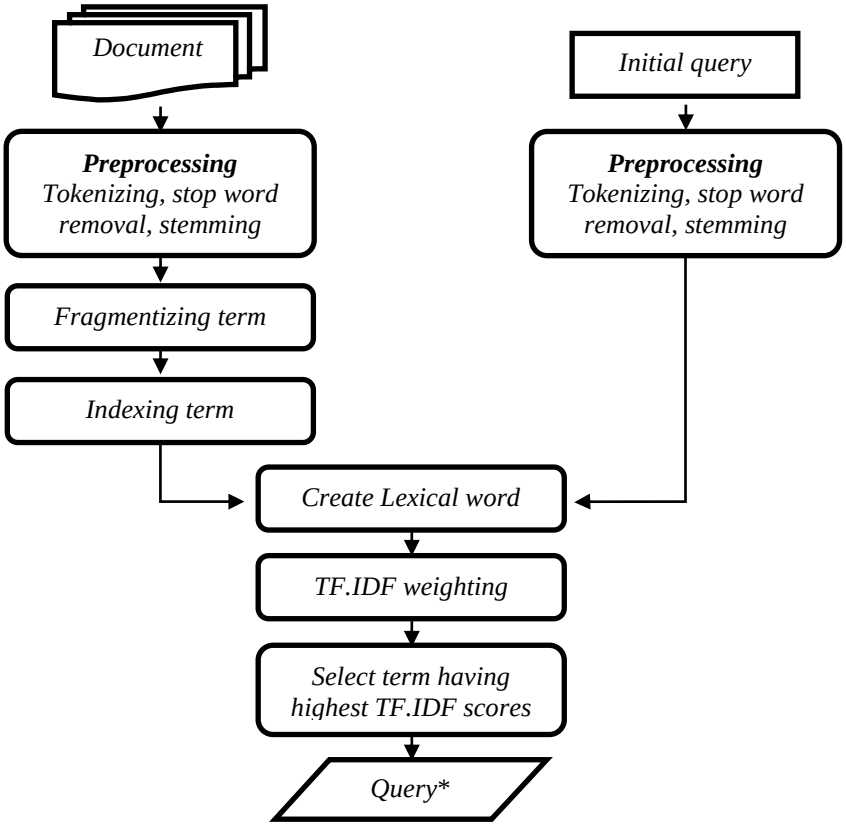
Gambar 1. Diagram Tahapan *Query Expansion*

3.1. *Explicit Feedback*

Langkah dalam *explicit feedback* ditunjukkan pada Gambar 2, yaitu dimulai dengan *preprocessing, fragmentizing, indexing*, pembentukan *lexical word*, pembobotan, serta pemilihan *term* untuk ekspansi *query*. Normalisasi dokumen dan *query* dilakukan dengan tahapan *preprocessing text*, yaitu dengan melakukan pemotongan teks dokumen menjadi kumpulan kata (*tokenizing*), menghilangkan kata yang tidak representatif (*stopword removal*), dan menjadikan kata berimbuhan menjadi akar kata (*stemming*).

Fragmentizing serta *indexing term* dokumen dan *query* diperlukan untuk membentuk *lexical word*. *Fragmentizing* merupakan pemenggalan dokumen menjadi beberapa penggalan kecil. Tiap *fragment* (penggalan) terdiri dari beberapa *term*. Pada penelitian ini, tiap *fragment* terdiri dari paling banyak sepuluh *term*. Sedangkan *lexical word* merupakan penggalan kecil (*small fragment*) dari dokumen yang memuat *term query*. Tujuan pembentukan *lexical word* adalah untuk menentukan nilai semantik dari suatu *term*. Beberapa *term* yang berada dalam suatu *fragment* secara umum memiliki hubungan kontekstual atau semantik yang kuat. Sehingga dengan *lexical word* dapat diidentifikasi kumpulan *term* yang muncul di sekitar suatu *term* dalam dokumen.

Pada tahap pembobotan, dilakukan perhitungan bobot *term* menggunakan pembobotan TF-IDF (*Term frequency – invers document frequency*) untuk tiap *term* dalam kumpulan *lexical word.* Pada tahap ini *lexical word* yang telah dibentuk dianggap sebagai dokumen. Bobot tiap *term* diurutkan dan dipilih sejumlah *term* dengan nilai bobot tertinggi. Berdasarkan *term* dengan nilai bobot tertinggi itulah pengguna memilih *term* yang dianggap relevan. *Term* yang terpilih tersebut menjadi perluasan *term query* sehingga menghasilkan *query* baru (*query**).



Gambar 2. Tahapan *Explicit Feedback*

3.2. *Pseudo Feedback*

*Query** sebagai hasil dari tahap sebelumnya diperluas lagi dengan *pseudo feedback*. Metode pembobotan *term* yang digunakan pada tahap ini adalah pembobotan TRQE (*Term Relatedness to Query – Entropy based*). Ide dari metode yang diajukan berdasarkan penelitian sebelumnya yang menggunakan pembobotan TRQ. Pembobotan TRQ memberikan nilai lebih pada *term* yang berhubungan dengan *query*. Rumusan dari TRQ memadukan antara perhitungan IDF (*invers document frequency*) dengan LWF (*lexical word frequency*), dimana formula lwf menggunakan Persamaan 1 dan TRQ dihitung menggunakan Persamaan 2. Notasi K pada Persamaan 1 adalah jumlah *query keyword* dan K_i adalah jumlah *query keyword* yang berhubungan dengan *lexical word i*. Konstanta α digunakan untuk mengatur keberimbangan bobot LWF dan IDF. Berdasarkan ekperimen yang dilakukan oleh Saneifar (2014) hasil optimal ketika $\alpha = 0{,}25$.

$$lwf_{ti} = \frac{1}{1 + \log(\frac{K}{K_i})} \quad (1)$$

$$TRQ_{ti} = \alpha \times lwf_{ti} + (1 - \alpha) \times idf_t \quad (2)$$
$$\alpha \in [0,1]$$

Tujuan utama entropi adalah menghitung jumlah rata-rata informasi yang diperlukan untuk mengidentifikasi label kelas di *data training*. Pendekatan *query expansion* berbasis entropi mempertimbangkan distribusi *term* yang muncul dalam dokumen yang relevan dan tidak relevan hingga sampai pada nilai diskriminasi *term*. Proporsi kemunculan *term* dalam dokumen

yang relevan dan tidak relevan ditunjukkan pada Persamaan 3 sedangkan perhitungan nilai entropi ditunjukkan pada Persamaan 4. G_R menunjukkan jumlah informasi dalam dokumen yang relevan, sedangkan G_{RN} menunjukkan jumlah informasi dalam dokumen yang tidak relevan. R adalah jumlah dokumen yang relevan, dan R_N adalah jumlah dokumen yang tidak relevan. R_o menunjukkan jumlah dokumen yang relevan yang memuat *term* tertentu, dan R_q menunjukkan jumlah dokumen relevan yang tidak memuat *term* tertentu. Demikian pula, R_{NO} menunjukkan jumlah dokumen tidak relevan yang memuat *term* tertentu, dan R_{Nq} menunjukkan jumlah dokumen tidak relevan yang tidak memuat *term* tertentu.

$$G_R = \begin{cases} -\frac{R_0}{R} log_2 \frac{R_0}{R} - \frac{R_q}{R} log_2 \frac{R_q}{R}, & \frac{R_0}{R} > \frac{1}{2} \\ 1, & \frac{R_0}{R} \leq \frac{1}{2} \end{cases} \quad (3)$$

$$G_{RN} = \begin{cases} -\frac{R_{N0}}{R_N} log_2 \frac{R_{N0}}{R_N} - \frac{R_{Nq}}{R_N} log_2 \frac{R_{Nq}}{R_N}, & \frac{R_{N0}}{R_N} < \frac{1}{2} \\ 1, & \frac{R_{N0}}{R_N} \geq \frac{1}{2} \end{cases} \quad (4)$$

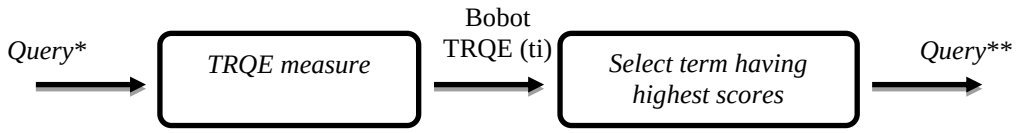
Karakteristik utama dari entropi adalah entropi dapat memperkirakan distribusi dari sample. Dengan demikian, jika *term* muncul dalam dokumen yang relevan 25 kali, nilai entropi akan sama seperti jika *term* itu muncul dalam dokumen yang relevan 75 kali. Untuk mengatasi masalah ini, maka diterapkan pengkondisian pada Persamaan 3 dan Persamaan 4. Pada Persamaan 3, $\frac{R_0}{R} > \frac{1}{2}$ untuk memastikan pemilihan *term* yang muncul di sebagian besar dokumen dari set dokumen yang relevan. Selain itu, pada Persamaan 4, $\frac{R_{N0}}{R_N} < \frac{1}{2}$ untuk memastikan bahwa *term* tersebut tidak muncul dalam sebagian besar dokumen dari set dokumen yang tidak relevan. Nilai *term* diskriminan dirumuskan dalam Persamaan 5.

$$TG_t = 1 - \left(\frac{G_R + G_{RN}}{2}\right) \quad (5)$$

TG_t berarti nilai diskriminasi untuk *term* t dengan nilai antara nol hingga satu. Persamaan 5 adalah normalisasi untuk Persamaan 3 dan Persamaan 4. Sebagai contoh, jika *term* yang relevan, t, muncul kurang dari atau sama dengan setengah dari dokumen yang relevan, nilai G_R menjadi satu. Demikian pula, jika *term* yang relevan t, juga muncul dalam lebih dari atau sama dengan setengah dari dokumen yang tidak relevan, nilai G_{RN} juga menjadi satu. Berdasarkan Persamaan 5, nilai TG_t akan menjadi nol, yang berarti *term* t, tidak akan dipilih karena memiliki nilai diskriminasi yang rendah.

Rumusan dari metode yang diajukan pada penelitian ini, yaitu TRQE dapat dilihat pada Persamaan 6. TG_t yang merupakan nilai diskiminan dari *term* t dari kumpulan dokumen relevan dan tiak relevan adalah bobot global yang menggantikan bobot IDF pada Persamaan 2. Pada TRQE disertakan pula faktor bobot lokal dari *term* yaitu f_t (frekuensi *term*) dengan cara mengalikannya dengan bobot global, lwf dan TG_t. Hasil akhir bobot tiap *term* diurutkan berdasarkan nilai bobot tertinggi. *Term* dengan bobot tertinggi menjadi perluasan *term query* sehingga menghasilkan *query* baru (*query***) sebagai *expanded query/enriched query*. Langkah *pseudo feedback* dapat dilihat pada Gambar 3.

$$TRQE_{ti} = f_t \times (\alpha \times lwf_{ti} + (1 - \alpha) \times TG_t)$$
$$\alpha \in [0, 1] \quad (6)$$

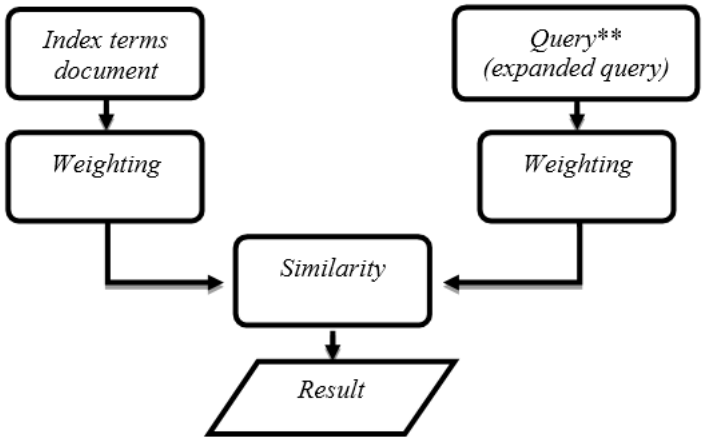


Gambar 3. Tahapan *Pseudo Feedback*

3.3. *Document retrieval*

Untuk mengembalikan dokumen hasil pencarian berdasarkan *query* yang telah dikembangkan disini menggunakan pertihungan ukuran kemiripan (*similarity*). Sebagaimana tahapan yang ditunjukkan pada Gambar 4, pembobotan dilakukan terhadap *term document* dan *query***, dari vektor bobot tersebut dilakukan perhitungan *similarity* menggunakan metode yang telah banyak digunakan yaitu *cosine similarity* (Tata, 2007). Dari ukuran kemiripan tersebut maka diperoleh hasil dokumen yang relevan dengan *query*. Nilai hasil perhitungan *cosine similarity* berkisar pada nol sampai dengan satu, dimana nol menandakan bahwa kedua dokumen tidak mirip sama sekali, dan satu menandakan bahwa antara *query* dan dokumen benar-benar identik. Rumusan *cosine* (Garcia, 2006) dinyatakan sebagaimana Persamaan 7 dimana $cos(q,d_j)$ merupakan nilai kosinus antara *query* dan dokumen *j*, sedangkan $TFIDF(t_k,q)$ dan $TFIDF(t_k,d_j)$ adalah pembobotan *TFIDF* kata t_k pada *query* dan dokumen *j*. $|TFIDFq|$ dan $|TFIDFd_j|$ adalah panjang dari vektor *query* q dan dokumen. Sebagai contoh $||d_i||^2 = (TFIDFt_1^2 + TFIDFt_2^2 + TFIDFt_3^2 + ... + TFIDFt_k^2)^{1/2}$, dimana $TFIDFt_k$ adalah bobot kata ke-t_k pada vektor dokumen d_i.

$$\cos(q,d_j) = \frac{\sum_{t_k}\left[TFIDF(t_k,q)\right]\bullet\left[TFIDF(t_k,d_j)\right]}{\sqrt{\sum\left|TFIDFq\right|^2}\bullet\sqrt{\sum\left|TFIDFd_j\right|^2}}, \quad (7)$$



Gambar 4. Tahapan *Document Retrieval*

Pada penelitian ini, *document retrieval* dilakukan dengan bantuan *library java* yang sudah ada yaitu *lucene*. *Library* ini bersifat *open source* dan dapat diunduh di http://lucene.apache.org/.

4. Skenario dan Hasil Pengujian

4.1. Skenario Pengujian

Data pengujian berupa dokumen berita berbahasa Inggris yang diambil dari situs New York Times (http://www.nytimes.com/). *Dataset* yang digunakan berjumlah 20 dokumen, terdiri dari berbagai topik meliputi kesehatan, politik, dan teknologi. Sedangkan kata kunci (*query*) yang digunakan berjumlah tujuh, digunakan secara konsisten pada pengujian metode ekspansi

query usulan dengan metode pembanding (TRQE dan TRQ). *Query* yang digunakan sebagai uji coba pada penelitian ini terdiri dari jumlah yang beragam, mulai dari satu kata hingga enam kata. *Query* tersebut terbatas pada kata kunci yang relevan dengan topik kesehatan dan politik. Kedua topik tersebut merupakan topik yang berkaitan dengan *dataset* yang digunakan pada penelitian ini. Sama halnya dengan data pengujian, *query* yang digunakan juga berbahasa Inggris. Pendapat seorang ahli (*expert*) digunakan untuk menentukan apakah dokumen hasil pencarian sistem relevan terhadap *query* yang dimasukkan.

Hasil uji coba pencarian dokumen yang menggunakan metode ekspansi *query* usulan dengan metode pembanding (TRQE dan TRQ) dievaluasi menggunakan perhitungan *recall* dan *precision* (Manning, 2008), seperti pada Persamaan 8 dan Persamaan 9 berikut ini.

$$precision = \frac{TP}{TP + FP} \tag{8}$$

$$recall = \frac{TP}{TP + TN} \tag{9}$$

Sebagaimana yang ditunjukkan pada Tabel 1, TP (*true positive*) merupakan jumlah dokumen yang ter-*retrieve* secara tepat oleh sistem. FP (*false positive*) merupakan jumlah dokumen yang tidak tepat ter-*retrieve*. TN (*true negative*) merupakan jumlah dokumen relevan yang tidak ter-*retrieve*. Sedangkan *false negative* merupakan jumlah dokumen tidak relevan yang tidak ditemukan. Berdasarkan evaluasi menggunakan *precission*, maka efektivitas dari metode yang digunakan dapat diketahui. Evaluasi dengan *recall* menunjukkan kemampuan metode dalam mengembalikan dokumen dengan tepat. Hasil evaluasi *recall* dan precission metode usulan TRQE kemudian dibandingkan dengan metode TRQ untuk dianalisis kemampuan TRQE dalam memperbaiki metode TRQ.

Tabel 1. *Recall* dan *Precision*

		***Document retrieval* secara Manual**	
		Dokumen yang Relevan	**Dokumen yang Tidak Relevan**
***Document retrieval* menggunakan TRQE**	**Dokumen yang Ditemukan**	*True Positive*	*False Positive*
	Dokumen yang Tidak Ditemukan	*True Negative*	*False Negative*

4.2 Hasil Pengujian

Tabel 2 dan Tabel 3 masing-masing menunjukkan hasil ekspansi *query* menggunakan metode TRQ dan TRQE. Sedangkan hasil pengujian pencarian dokumen ditunjukkan pada Tabel 4.

Tabel 2. Hasil Ekspansi *Query* Menggunakan Metode TRQ

Query	**Relevan *Feedback***		***Pseudo Feedback***
	Suggestions	***Selected query (query *)***	
Ebola Virus	*- ebola virus parasite* *- ebola virus bat* *- ebola virus sicken*	*ebola virus parasite*	*ebola virus parasite*
Ebola patient	*- ebola patient parasite* *- ebola patient bat* *- ebola patient sicken*	*Ebola patient*	*Ebola patient*
Respiratory virus	*- respiratory virus parasite* *- respiratory virus bat* *- respiratory virus sicken*	*respiratory virus parasite*	*respiratory virus parasite headline*
Democrate and Republic party in election	*- democrate republic party election conservative* *- democrate republic party election peace* *- democrate republic party election arm*	*democrate republic party election conservative*	*democrate republic party election conservative edition*
Virus	*- virus parasite* *- virus bat* *- virus sicken*	*virus sicken*	*virus sicken*
Cause of Children desease	*- child desease secretion* *- child desease unconventional* *- child desease spread*	*child desease spread*	*child desease spread*

Pada Tabel 2, kolom *query* berisi kata kunci pencarian yang dimasukkan ke sistem. Kolom *suggestion* berisi saran pencarian (*relevan feedback*) yang diusulkan oleh sistem, sedangkan kolom *selected query* (*query*)* berisi kata kunci pencarian yang dipilih dari relevan *feedback*. Kolom *pseudo feedback* berisi perbaikan *query* di dalam sistem yang tidak ditunjukkan ke pengguna. Sebagai contoh, *query* yang dimasukkan ke sistem adalah *ebola virus*. Setelah *query* tersebut diproses, maka sistem akan mengeluarkan saran pencarian berupa *ebola virus parasite*, *ebola virus bat*, dan *ebola virus sicken*. Dari ketiga saran pencarian tersebut, *query* yang dipilih adalah *ebola virus parasite*. *Query* tersebut kemudian diproses kembali ke sistem sehingga menghasilkan *pseudo feedback* berupa *ebola virus parasite*. *Query* akhir inilah yang digunakan sistem untuk melakukan pencarian dokumen.

Cara membaca Tabel 3 sama dengan Tabel 2. Pembeda kedua tabel tersebut terletak pada metode yang digunakan, Tabel 2 adalah hasil ekspansi *query* dengan metode TRQ, sedangkan Tabel 3 adalah hasil ekspansi *query* dengan metode TRQE. *Query* dan relevan *feedback* yang diujicobakan sama, tetapi dapat menghasilkan *pseudo feedback* yang berbeda sesuai dengan algoritma metode masing-masing.

Tabel 3. Hasil Ekspansi *Query* Menggunakan Metode TRQE

Query	**Relevan *Feedback***		***Pseudo Feedback***
	Suggestions	***Selected (query *)***	
Ebola Virus	*- ebola virus parasite* *- ebola virus bat* *- ebola virus sicken*	*ebola virus parasite*	*ebola virus parasite*
Ebola patient	*- ebola patient parasite* *- ebola patient bat* *- ebola patient sicken*	*Ebola patient*	*Ebola patient prodigiously*
Respiratory virus	*- respiratory virus parasite* *- respiratory virus bat* *- respiratory virus sicken*	*respiratory virus parasite*	*respiratory virus parasite*
Democrate and Republic party in election	*- democrate republic party election conservative* *- democrate republic party election peace* *- democrate republic party election arm*	*democrate republic party election conservative*	*democrate republic party election conservative*
Virus	*- virus parasite* *- virus bat* *- virus sicken*	*virus sicken*	*virus sicken*
Cause of Children desease	*- child desease secretion* *- child desease unconventional* *- child desease spread*	*child desease spread*	*child desease spread*

Tabel 4 berikut ini menunjukkan hasil evaluasi *recall* dan *precission* metode TRQ dan TRQE. Berdasarkan hasil tersebut, TRQE kemudian dibandingkan dengan TRQ untuk mengetahui apakah metode tersebut dapat memperbaiki metode pembandingnya atau tidak.

Tabel 4. Hasil Evaluasi *Recall* dan *Precission*

***Query* ke-**	**TRQ**		**TRQE**	
	recall%	***precision%***	***recall%***	***precision%***
1	22.22	100.00	22.22	100.00
2	22.22	100.00	22.22	100.00
3	5.26	50.00	5.26	50.00
4	**25.00**	**50.00**	**23.53**	**57.00**
5	10.00	100.00	10.00	100.00
6	5.26	50.00	5.26	50.00
7	22,22	100.00	22.22	100.00

5. Diskusi

Berdasarkan hasil pengujian pada Tabel 2 dan Tabel 3, dapat dilihat beberapa ekspansi *query* yang dihasilkan menghasilkan *query** yang. Seperti pada uji coba ke-4, yaitu ketika *query* yang dimasukkan adalah *democrate and republic party in election*. Pada ekspansi *query* TRQ (Tabel 2), *pseudo feedback* yang dihasilkan adalah *democrate republic party election conservative edition*. Sedangkan pada ekspansi *query* TRQE (Tabel 3), *pseudo feedback* yang dihasilkan adalah *democrate republic party election conservative*. Adapun beberapa uji coba

menghasilkan ekspansi *query* yang sama yaitu pada uji coba 1, uji coba 5, uji coba 6, dan uji coba 7.

Secara umum, hasil evaluasi *recall* dan *precission* pada TRQ dan TRQE hampir sama untuk setiap uji coba. Hanya pada uji coba ke-4 yang menghasilkan nilai *recall* dan *precission* yang berbeda. Ekspansi *query* dengan metode TRQE memiliki kemampuan meningkatkan *recall* dan *precision* dari metode TRQ. Peningkatan tersebut tampak pada *query* kedua, dimana TRQ memiliki kemampuan mengembalikan data dengan *precision* 50% sedangkan TRQE mampu mengembalikan data dengan nilai *precision* 57%. Keenam *query* lainnya yang diujicobakan tergolong lebih pendek daripada *query* kedua (*query* ke-2: *democrate and republic party in election*). Berdasarkan hasil evaluasi, metode TRQE yang diusulkan terlihat memberikan hasil yang lebih baik untuk *query* yang relatif lebih panjang.

Entropi mampu mempertimbangkan distribusi *term* yang muncul dalam dokumen yang relevan dan tidak relevan hingga sampai pada nilai diskriminasi *term.* Dengan menerapkan konsep entropi pada pembobotan *term* untuk ekspansi *query* maka dapat diperoleh *term* yang tepat sebagai tambahan *query*. Sehingga ekspansi *query* lebih optimal dan dapat meningkatkan kemampuan pencarian dokumen. Hal ini tampak pada *query* kedua, kata "*edition*" dianggap perlu ditampahkan pada *query* ketika diukur dengan metode TRQ. Berbeda ketika menggunakan metode TRQE, kata "*edition*" dinilai tidak tepat untuk ditambahkan pada *query*. Entropi pada TRQE memperhatikan dokumen relevan dan tidak relevan. Sehingga, kata "*edition*" yang lebih banyak dimuat pada dokumen tidak relevan tidak dianggap sebagai *term* yang tepat untuk ditambahkan.

Pada tahap perhitungan entropi diperlukan penentuan kelompok dokumen relevan dan tidak relevan. Penentuan kelompok tersebut didasarkan pada suatu nilai ambang batas yang didapatkan dari rata-rata selisih nilai tertinggi dan terendah TFIDF *lexical word* terhadap *query*. Penelitian ini belum menggunakan batas *threshold* yang optimal dalam menentukan dokumen relevan dan tidak relevan. Beberapa dokumen yang seharusnya relevan dianggap tidak relevan oleh sistem. Hal ini menjadi salah satu faktor yang mempengaruhi hasil metode yang diajukan. Selain itu, keterbatasan data juga mempengaruhi hasil evaluasi, dimana rentang nilai *precision* untuk setiap uji coba terpaut jauh.

6. Kesimpulan

Pada penelitian ini telah diajukan satu metode *query expansion* yang baru berbasis eksplisit dan *pseudo relevance feedback*. Metode pembobotan dalam *pseudo relevance feedback* dalam penelitian ini disebut dengan *Term Relatedness to Query – Entropy based* (TRQE).*Query expansion* ini diterapkan dalam *document retrieval* dengan menggunakan *dataset* dokumen berita dari New York Times untuk uji coba.

Penambahan entropi untuk pembobotan pada *query expansion* memberikan pengaruh untuk meningkatkan relevansi pencarian dokumen dengan nilai *recall* 25,53% dan presisi 57%. Berdasarkan beberapa *query* yang diuji, tampak bahwa pembobotan dengan entropi kurang berpengaruh untuk *query* yang pendek.

Dataset yang diujikan dalam penelitian ini dilakukan pada jumlah dokumen yang terbatas, oleh karena itu perlu adanya ujicoba terhadap jumlah data yang lebih besar dan beragam sehingga pengaruh optimasi ekspansi *query* terhadap pencarian lebih terlihat.

Referensi

Agichtein, E., Lawrence, S., & Gravano, L. 2001. Learning search engine specific *query* transformations for question answering. In *Proceedings of the 10th international conference on World Wide Web* (pp. 169-178). ACM.

Araujo, L., Zaragoza, H., Pérez-Agüera, J. R., & Pérez-Iglesias, J. 2010. Structure of morphologically expanded queries: A genetic algorithm approach. *Data & Knowledge Engineering*, *69*(3), 279-289.

Boston, C., Fang, H., Carberry, S., Wu, H., & Liu, X. 2014. Wikimantic: Toward effective disambiguation and expansion of queries. *Data & Knowledge Engineering*, *90*, 22-37.

Carpineto, C., De Mori, R., Romano, G., & Bigi, B. 2001. An information-theoretic approach to automatic *query* expansion. *ACM Transactions on Information Systems (TOIS)*, *19*(1), 1-27.

Garcia, DR. E., 2006, *The Classic Vector Space Model*, (Online), (http://www.miislita.com/term-vector/term-vector-3.html, diakses 21 Oktober 2011)

Manning, C. D., Raghavan, P., & Schütze, H. 2008. *Introduction to information retrieval* (Vol. 1, p. 6). Cambridge: Cambridge university press.

Pasca, M. A., & Harabagiu, S. M. 2001. High performance question/answering. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 366-374). ACM.

Saneifar, H., Bonniol, S., Poncelet, P., & Roche, M. 2014. Enhancing passage retrieval in log files by *query* expansion based on explicit and pseudo relevance feedback. *Computers in Industry*, 65(6), 937-951.

Tata, Sandeep, Patel M, Jignesh. 2007. Estimating he Selectivity of tf-idf based Cosine similarity Predicates, *SIGMOD Record* December 2007 Vol 36 No. 2

Wu, I., Chen, G. W., Hsu, J. L., & Lin, C. Y. (2013). An entropy-based *query* expansion approach for learning researchers' dynamic information needs. *Knowledge-Based Systems*, *52*, 133-146.

Xu, J., & Croft, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, *18*(1), 79-112.