Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics
JLK
Vol. 1, No. 1
Maret 2018

Pembobotan Kata pada Query Expansion dengan Tesaurus dalam Pencarian Dokumen Bahasa Indonesia

Fatra Nonggala Putra[#1], Ari Effendi[#2], Agus Zainal Arifin[#3]

[#]*Jurusan Teknik Informatika Institut Teknologi Sepuluh Nopember*
Kampus ITS Keputih, Sukolilo, Surabaya, 60111, Jawa Timur, Indonesia

[1]putra.fatra08@gmail.com

[2]arieffendi@gmail.com

[3]agusza@cs.its.ac.id

Abstract— ***In this paper we aim to expand the search results in the retrieval system by expanding words in queries using synonyms of words in a synonymous thesaurus, which then queries are weighted based on word combinations in queries based on the main query and the number of document frequency of expansion terms. Our contribution in this paper is: first, query expansion using thesaurus word synonym Indonesian. Second, new term weighting of query for ranking the retrieval documents. From the experiment result, we can be concluded, firstly, result of comparison of precision and recall between using query expansion and without query expansion each up by 2.05% and 54.24%. The value of precision and recall without query expansion 80% and 43.54%. While the query expansion has the value of precision and recall 82.05% and 97.78%. Second, new term weighting of query our proposed on query expansion is able to place the retrieval result from the top query in the top ranking compared to query expansion.***

Keywords- term weighting, query expansion, thesaurus, tf-idf

Abstrak— **Dalam paper ini kami bertujuan untuk memperluas hasil pencarian dalam sistem temu kembali dengan melakukan expansi kata dalam *query* menggunakan sinonim kata dalam tesaurus sinonim, yang kemudian query diberi bobot berdasarkan kombinasi kata dalam query berdasarkan query utama dan banyaknya df dari term hasil expansi. Kontribusi kami dalam paper ini adalah: pertama, query expansion menggunakan tesaurus sinonim kata bahasa indonesia. Kedua, pembobotan kata dalam query untuk perankingan dokumen hasil temu kembali. Dari hasil uji coba dapat disimpulkan, pertama, mendapatkan hasil perbandingan precison dan recall antara menggunakan query expand dan tanpa query expand masing-masing naik sebesar 2.05% dan 54.24%. Nilai *precision* dan *recall* tanpa query expansion 80% dan 43.54%. Sedangkan dengan *query expansion* emiliki nilai *precision* dan *recall* 82.05% dan 97.78%. Kedua, pembobotan kata dalam *query expansion* yang kami usulkan mampu menempatkan hasil temu kembali dari *query* utama pada peringkat teratas dibandingkan hasil temu kembali *query expansion*.**

Kata kunci **- pembobotan kata, query expansion, tesaurus, tf-idf**

I. PENDAHULUAN

Dengan terus bertambahnya jumlah dokumen yang tersedia secara daring. Layanan mesin pencari satu persatu bermunculan guna membantu pengguna menemukan dokumen yang dibutuhkan. Secara umum mesin pencari melakukan pencarian dokumen bedasarkan masukan kata kunci yang diberikan. Pada kondisi tertentu, mesin pencari tidak mengembalikan dokumen apapun dikarenakan ketidaksesuaian kata kunci dengan dokumen yang tersedia[4].

Tesaurus merupakan sebuah rujukan yang berisi kata-kata yang memiliki pertalian makna berupa kesamaan, perlawanan, hubungan superordinat dan subordinat, atau hubungan antarbagian. Kesamaan makna (sinonim) dapat digunakan untuk memperkaya query pada proses Temu Kembali Informasi. Diharapkan dengan pemanfaatan sinonim untuk memperkaya query dapat menemukan dokumen yang lebih banyak dibanding sebelum menggunakan pengayaan query.

Penelitian dengan memanfatkan tesaurus persamaan kata (sinonim) juga pernah dilakukan oleh Hazra Imrandan[1] dan Percy Nohama et.al[2]. Berbeda dengan penelitian[3] yang menggunakan selektif *query expansion* (QE) untuk melakukan prediksi terhadap kata yang akan di-*expan*, metode yang kami gunakan adalah dengan mencari sinonim *term* dari tesaurus dan melakukan perankingan terhadap kandidat *term* k QE, yaitu hanya mengambil lima sinonim dari setiap kata yang memiliki nilai *documen frequency*(df) tertinggi.

Dalam penelitian ini kami mengusulkan, pertama, QE menggunakan tesaurus sinonim kata bahasa indonesia. Kedua, pembobotan kata dalam *query* untuk perangkingan dokumen hasil temu kembali.

Submitted 23-12-2017; Revised 9-03-2018; Accepted 11-03-2018

II. METODOLOGI

Dalam Sistem temu kembali informasi pengguna menggunakan keyword untuk melakukan pencarian di dalam sistem. Keyword yang diinputkan oleh pengguna dikenali sebagai query oleh sistem dan nantinya akan di-expand oleh sistem berdasarkan tesaurus sinonim yang ada.

Pada penelitian ini, tesaurus sinonim yang kami buat berupa kumpulan sinonim yang merupakan sinonim dari kata hasil ekstrasi dari dokumen berbahasa indonesia. Tesaurus sinonim ini yang dijadikan sumber expansi kata dari keyword yang dimasukan oleh user.

A. Pengumpulan Data

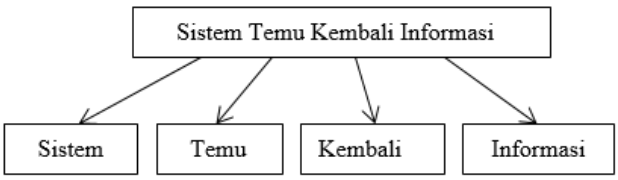
Data yang digunakan sebagai corpus dalam penelitian ini berupa dokumen berita dari berbagai media daring nasional berbahasa indonesia dengan topik politik pada tahun 2016. Jumlah dokumen berita yang dikumpulkan sebanyak 400 dokumen. Dokumen berita memiliki metadata yang terdiri dari: <id>, <judul>, <tanggal>, <kata_kunci>, <isi>, dan <link>. Kemudian dokumen disimpan di database sistem.

B. Preprocessing

Sebelum data diolah lebih lanjut perlu dilakukan tahap preprocessing agar data siap untuk diproses. Tahapan preprocessing ini terdiri dari 4 (empat) tahap yaitu casefolding, tokenizer, stopwords removal, dan stemming.

1) *Casefolding:* Casefolding adalah proses penyeragaman huruf menjadi huruf kecil (lowercase) agar tidak terjadi kesalahan sistem dalam mengenali dua kata atau lebih yang sama namun berbeda dalam penulisannya, misalnya kata 'KPK' dengan 'kpk'.

2) *Tokenizing:* Tokenizer adalah proses pemecahan kumpulan kata dalam dokumen atau kalimat yang dipisahkan oleh tanda baca spasi menjadi kumpulan token yang biasanya berupa kata [5]. Gambar.1 merupakan contoh tokenizing pada sebuah kalimat.



Gambar. 1 Contoh Tokenizer pada sebuah kalimat

3) *Stopwords Removal:* tahap ini adalah proses penghapusan kata yang dianggap tidak penting (*stopword*) hasil proses *tokenizing* apakah termasuk dalam daftar kata tidak penting (*stop list*) [5]. Contoh *stopword* tertera pada Tabel I.

TABEL I
CONTOH STOPWORD

ada	pada
adalah	di
akan	yang
mungkin	oleh
jika	dll

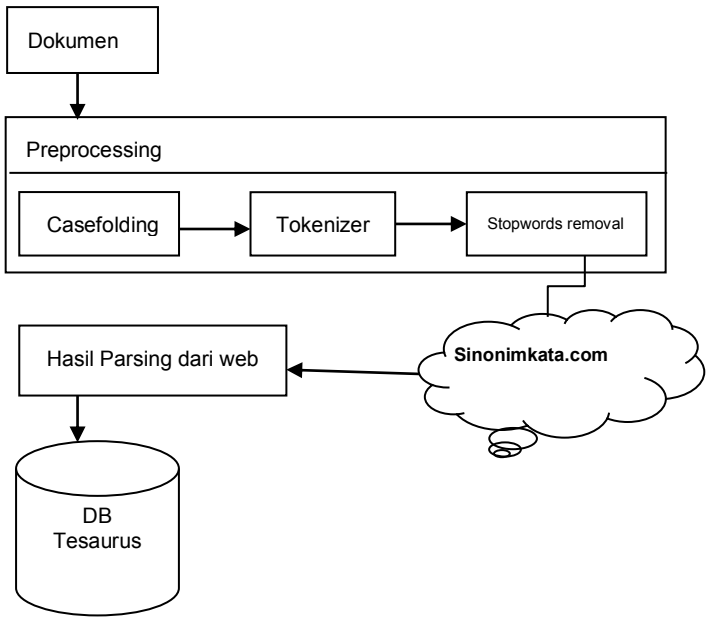
4) *Stemming:* Stemming adalah proses pengubahan kata berimbuhan menjadi kata dasar sesuai dengan kamus kata yang digunakan. Dalam penelitian ini kami menggunakan algoritma dan kamus kata *stemming* milik sastrawi. Contoh dari proses stemming ditunjukan pada Tabel II.

TABEL II
CONTOH STEMMING KATA

Sebelum Stemming	**Setelah Stemming**
mencari	cari
melihat	lihat
pembobotan	bobot
pencalonan	calon

C. Pembentukan Tesaurus Sinonim kata

Pada tahap ini dilakukan pembentukan tesaurus sinonim yang memuat semua sinonim kata dari kata yang terkandung di dalam korpus. Semua kata yang dihasilkan dari tahapan *preprocessin*g digunakan untuk melakukan proses scraping dari laman sinonimkata.com untuk mendapatkan sinonim kata dari semua kata/token yang dihasilkan pada tahapan preprocessing dan hasilnya berupa kumpulan sinonim kata disimpan di database sistem. Alur tahap ini ditunjukan pada Gambar. 2

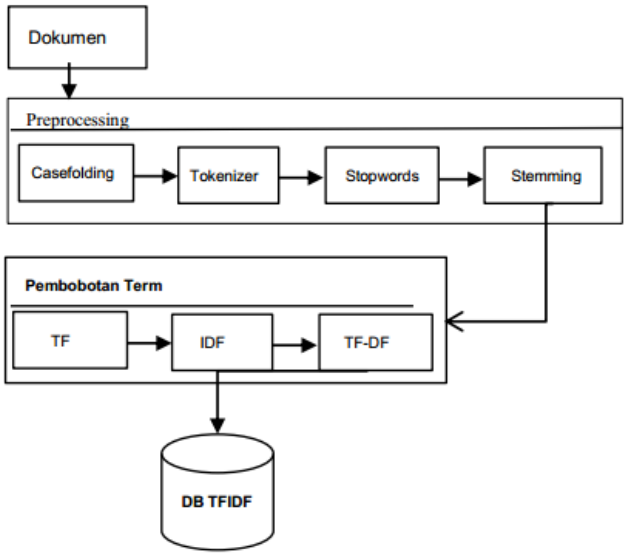


Gambar. 2 Proses pembentukan tesaurus

D. Perhitungan pembobotan TF-IDF

Pada tahap ini setiap kata (*term*) hasil *pre-processing* dari seluruh dokumen yang akan digunakan untuk perhitungan *cosine similarity* untuk dilakukan perhitungan bobot setiap dalam tiap dokumen. Perhitungan bobot yang digunakan adalah *term frequency* (tf) , *inverse document frequency* (idf), dan tf-idf.

Gambar.3 adalah alur proses pemberian bobot kata dengan metode tf-idf. Metode pembobotan tf-idf ini digunakan untuk menilai bobot relevansi term dari sebuah dokumen terhadap seluruh dokumen dalam korpus[6].



Gambar. 3 Proses perhitungan bobot kata dengan metode TFIDF

E. Query Expansion

Tesaurus sinonim kata berbahasa indonesia yang telah dibuat digunakan untuk memperluas pencarian dengan memperbanyak query asli menjadi beberapa QE.

TABEL III
CONTOH SINONIM KATA BERDASARKAN TESAURUS YANG DIBANGUN

Query	Kata	Sinonim
Kasus korupsi	Kasus	kejadian, masalah, peristiwa, perkara, persoalan, skandal, urusan
	Korupsi	kecurangan, manipulasi, penggelapan, penyelewengan,

F. Perhitungan Cosine Similarity

Dalam proses perhitungan *cosine similarity, input* yang digunakan adalah bobot dari term setiap dokumen, bobot *term* yang digunakan adalah bobot tf-idf yang sudah dihitung pada proses sebelumnya. Persamaan *cosine similarity* yang digunakan sebagai berikut:

$$CosSim(d_i, q_i) = \frac{q_i . d_i}{|q_i||d_i|} = \frac{\sum_{j=1}^{t}(q_{ij} . d_{ij})}{\sqrt{\sum_{j=1}^{t}(q_{ij})^2 . \sum_{j=1}^{t}(d_{ij})^2}} \quad (1)$$

Keterangan:
q_{ij} = bobot istilah j pada dokumen i = tf_{ij} . idf_j
d_{ij} = bobot istilah j pada dokumen i = tf_{ij} . idf_j

G. Pembobotan Kata pada Query

Pembobotan kata pada query yang kami usulkan ini untuk memberikan bobot pada kata(t*erm*) yang digunakan untuk pencarian dokumen. Dalam pembobotan yang diusulkan, bobot kata yang merupakan bagian dari query asal diberikan bobot lebih dibandingkan kata hasil expansi dari tesaurus sinonim yaitu 1 untuk term dari query asli dan 1-1/log(df_{term}), sehingga hasil pencarian temu kembali oleh sistem akan menempatkat dokumen hasil dari *query* asli pada posisi teratas dibandingkan dokumen hasil temu kembali oleh *QE*.

Persaman yang diusulkan untuk pembobotan terbaru kata pada *query* tercantum pada Persamaan. 2 dan contoh pemberian bobot pada Tabel IV. Sedangkan untuk alur temu kembali informasi ditunjukan pada Gambar. 5.

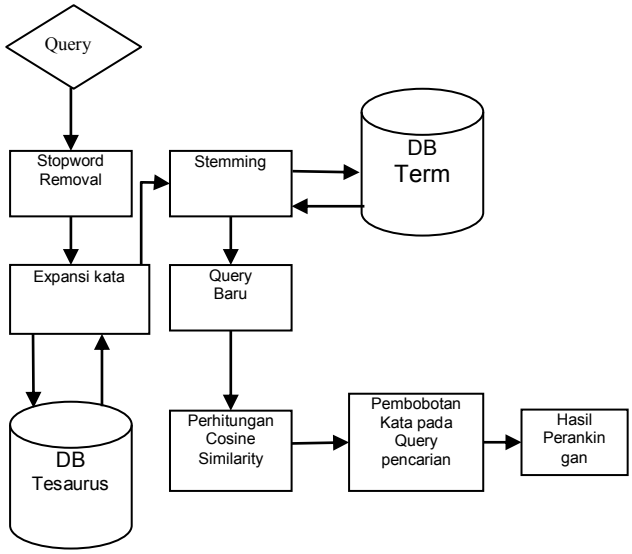
$$Bobot\ Baru = \sum_{k=0}^{n}(WTqa) + \sum_{k=0}^{n}(WTqe) + CosSim \quad (2)$$

Keterangan:
WT_{qa} = bobot kata jika kata termasuk anggota kata dari query asli = 1
WT_{qe} = bobot kata untuk kata hasil expansi (sinonim) = 1 - 1/log(df_{term})

*jika *document frequency* (df) kata expansi (sinonim) < = 10 maka WTqe = 0

TABEL IV
CONTOH PERHITUNGAN BOBOT KATA DALAM QUERY

No	Query	Bobot	Total Bobot Baru
1	Kasus Uang Suap	1 + 1 + 1 + Cosim	3 + Cosim
2	**Skandal** Uang Suap	WTqe + 1 + 1 + Cosim	TWqe + 2 + Cosim
3	**Skandal Duit** Suap	WTqe + WTqe + 1 + Cosim	2WTqe + 1 + Cosim
4	**Skandal Duit Korupsi**	WTqe + WTqe + WTqe + Cosim	3WTqe + Cosim



Gambar.5 Metodologi Temu Kembali Informasi

III. HASIL UJI COBA

Uji coba dilakukan menggunakan 400 dokumen berita dengan topik politik digunakan sebagai dataset. Dokumen berita memiliki keunikan dalam penyajian penulisan berita antar tiap penulis atau antar media. Keunikan yang paling mencolok adalah perbedaan pemilihan kata yang digunakan untuk menyampaikan topik berita yang sama. Hal ini menyebabkan pencarian dengan kata kunci tertentu akan hanya mengembalikan dokumen berita yang memiliki kata kunci sama dengan kata kunci yang digunakaan oleh pengguna sistem dan tidak dapat menemukan dokumen dengan kata kunci berbeda meskipun dengan topik dan isi berita yang sama. Gambar.7 menunjukan hasil pencarian dengan kata kunci utama yaitu "demokrat pecat ruhut", sedangkan Gambar.8 menunjukan hasil pencarian dari kata kunci utama yang telah diekspan menjadi "demokrat henti ruhut" dan "demokrat lepas ruhut".

Untuk menguji performa sistem dari metode yang diusulkan, perhitungan *recall* dan *precision* digunakan untuk mengukur kemampuan sistem pencarian dalam menemukan dokumen yang relevan (*recall*) dan mengukur kemampuan sistem untuk tidak memberikan peringkat pada dokumen yang tidak relevan (*precision*)[7]

Berdasarkan hasil ujicoba dengan metodologi pencarian pada gambar.5 menggunakan lima *query*, kami membandingkan antara pencarian biasa (tanpa QE) dengan pencarian menggunakan QE. Dari tabel V dan VI, terlihat perbandingan hasil temu kembali dari pencarian dengan *query* asli dan hasil pencarian *QE*. Pada kata kunci 'demokrat pecat ruhut', hasil temu kembali dari *query* asli hanya mampu mengembalikan 17 dokumen berita dengan nilai precision 100 % dan recall 65,38%. Dengan kata lain hasil pencarian dengan kata kunci asli tidak mampu mengembalikan 34,62% dokumen berita yang mempunya topik sama. Sedangkan hasil temu kembali oleh *QE* mampu mengembalikan 30 dokumen dengan nilai precision 86,67% dan recall 100%.

Penurunan *precision* atau dengan kata lain ada beberapa dokumen yang tdak relevan dikembalikan oleh sistem pada hasil temu kembali oleh *QE* dapat diatasi dengan pembobotan baru kata dalam *query* yang kami usulkan. Dari gambar.7 dan gambar.8, terlihat bahwa hasil temu kembali dengan kata kunci asli yaitu 'demokrat pecat ruhut' mepunyai bobot 3,6 dan 3,5 sedangkan dokumen hasil QE dengan kata kunci 'demokrat lepas ruhut' dan 'demokrat henti ruhut' masing-masing 2,6 dan 2,5. Pembobotan baru kata dalam *query* yang kami usulkan memberikan bobot pada kata dalam *query* asli lebih besar daripada kata hasil expansi. Bobot query asli 'demokrat pecat ruhut' mempunyai bobot 3,5 hasil dari penjumlahani (3 + 0.5), 3 merupakan bobot tambahan karena terdapat tiga kata *query* yang masing-maisng kata bernilai 1 dan 0,5 merupakan hasil perhitungan *cosine similarity*. Sedangkan query hasil expansi 'demokrat lepas ruhut' mempunyai bobot 2,6 hasil dari penjumlahan (2 + 0.6), 2 merupakan bobot tambahan karena terdapat 2 unsur kata query asli dan 0.6 merupakan hasil perhitungan *cosine similarity*.

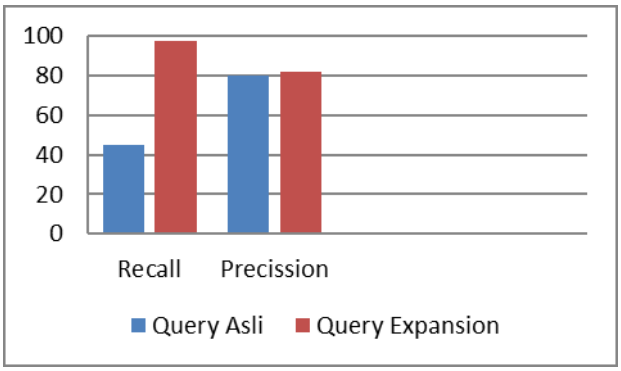
Pembobotan kata dalam query yang kami usulkan memeberikan bobot lebih besar kepada term yang bersal dari query asli daripada term dari hasil expansi. Sehingga dokumen hasil temu kembali dari query asli akan mempunya peringkat yang lebih tinggi daripada dokumen hasil *QE* yang membuat dokumen temu kembali dari query asli akan selalu berada pada peringkat teratas.

TABEL V
HASIL PERCOBAAN TEMU KEMBALI SISTEM

No	Keyword	Hasil Temu Kembali			
		Query Asli	Relevan	Query Ekspansi	Relev-an
1	Demokrat pecat ruhut	17	17	30	26
2	Penistaan agama oleh ahok	10	8	10	8
3	Agus maju gubernur dki	4	4	8	8
4	Kasus suap partai golkar	1	1	17	12
5	Kasus suap E-ktp	0	0	4	3

TABEL VI
NLAI PRECISION DAN RECALL HASIL UJICOBA SISTEM

No	Keyword	Tanpa QE		QE	
		Precission (%)	Recall (%)	Precision (%)	Recall (%)
1	Demokrat pecat ruhut	100	65,38	86.67	100
2	Penistaan agama oleh ahok	100	100	80	100
3	Agus maju gubernur dki	100	44.45	100	88.9
4	Kasus suap partai golkar	100	8.33	70.5	100
5	Kasus suap E-ktp	0	0	75	100
Rata-rata		**80**	**43.54**	**82.05**	**97.78**



Gambar.6 Grafik nilai precision dan recall antara query asli dan QE

Keputusan Dewan Kehormatan Demokrat: Ruhut Sitompul Dipecat dari Partai morelikethis

Score: 3.6460315970831827
Query: demokrat pecat ruhut

Pantaskah Demokrat Pecat Ruhut dan Hayono? morelikethis

Score: 3.5705754090155075
Query: demokrat pecat ruhut

Gambar. 7 Hasil temu kembali dengan query asli "demokrat pecat ruhut"

Komisi Pengawas Demokrat Segera Beri Sanksi pada Ruhut morelikethis

Score: 2.636941393844879
Query: demokrat lepas ruhut

Demokrat Tutup Rapat Alasan Pemberhentian Ruhut dari Jubir morelikethis

Score: 2.5918871265065198
Query: demokrat henti ruhut

Gambar. 8 Hasil temu kembali dengan *QE*

IV. ANALISA DAN PEMBAHASAN

Pada proses expansi kata dalam query tidak dilakukan tahapan *stemming* sebelum kata itu telah di-expansi atau dicari persamaan kayanya, namun *stemming* dilakukan setelah proses *QE*. Hal ini dilakukan karena berdasarkan pembacaan terhadap sinonim kata, kata berimbuhan dan kata dasar memiliki beberapa sinonim yang berbeda yang dapat menyebabkan tidak relevannya antara kata kunci pencarian dengan dokumen hasil temu kembali.

Pada hasil ujicoba dengan menggunakan QE terjadi kenaikan recall yang signifikan yaitu 54.24% dibanding tanpa menggunakan QE. Sedangkan nilai precision cenderung stabil meskipun ada kenaikan sebesar 2.05%. Kenaikan recall yang cukup tinggi dikarenakan data yang kami pakai adalah dokumen berita yang kebanyakan dari redaksi media banyak menggunakan kata ganti sinonim ataupun majas untuk menggantikan topik pada umumnya. Contoh pada pencarian dengan kata kunci 'suap e-ktp' dokumen yang dikembalikan sistem sebanyak nol atau dengan kata lain sistem tidak dapat menemukan dokumen dengan topik 'suap e-ktp'. Namun, dengan diterapkannya QE dokumen yang dikembalikan oleh sistem sebanyak empat dokumen dikarenakan kata 'suap' telah dicari persamaan katanya (sinonim) menjadi 'korupsi', sehingga kata kunci pencarian hasil *QE* menjadi 'korupsi e-ktp'.

V. KESIMPULAN

Dalam penelitian ini telah diusulkan metode baru untuk melakukan pembobotan query. Serangkaian percobaan telah dilakukan untuk menguji kinerja dari metode yang diusulkan. Dari hasil pengujian didapatkan perbandingan *precision* dan *recall* antara penggunaan *query expand* dan tanpa *query expand* masing-masing naik sebesar 2.05% dan 54.24%. Nilai *precision* dan *recall* tanpa *QE* 80% dan 43.54%. Sedangkan dengan *QE* meiliki nilai *precision* dan *recall* 82.05% dan 97.78%. Pembobotan kata dalam penelitian yang diusulkan mampu menempatkan hasil temu kembali dari *query* utama pada peringkat teratas dibandingkan hasil temu kembali *QE*.

Penelitian lebih lanjut dapat dilakukan untuk menyelidiki syarat perluasan query agar tidak merubah makna dari query asal. Sehingga perluasan relevansi dokumen yang ditampilkan dari hasil pencarian tetap terjaga.

REFERENSI

[1] H. Imran, & A. Sharan. "Thesaurus and query expansion," *International journal of computer science & information Technology (IJCSIT)*, *1*(2), 89-97, 2009

[2] P. Nohama., Pacheco, E.J., Andrade, R.L de,Bitencourt, J.L., Markó, K., Schulz, S., "Quality issues in thesaurus building: a case study from the medical Domain,". Brazilian Journal of Biomedical Engineering Volume 28, Número 1, p. 11-22, 2012.

[3] Amati G., Carpineto C., Romano G. "Query Difficulty, Robustness, and Selective Application of Query Expansion", McDonald S., Tait J. (eds) Advances in Information Retrieval. ECIR 2004. Lecture Notes in Computer Science, vol 2997, 2004.

[4] Hang Cui, Ji-Rong Wen, Jian-Yun Nie and Wei-Ying Ma, "Query expansion by mining user logs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 829-839, July-Aug. 2003.

[5] Arifin AZ , SA Novan. "Klasifikasi dokumen berita kejadian berbahasa Indonesia dengan algoritma single pass clustering", Proceeding of seminar on intelegent Technology and Its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh, 2002.

[6] Munjiah NS, Atmagi RW, Rahayu DS, dan Arifin AZ, "Sistem Temu Kembali Dokumen Teks Dengan Pembobotan Tf-Idf Dan Lcs", JUTI, vol. 11, pp. 17-20, 2013.

[7] Carterette B. Precision and Recall. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, 2009.