

## Ekspansi Kueri pada Sistem Temu Kembali Informasi Berbahasa Indonesia Menggunakan Analisis Konteks Lokal

### *Query Expansion on Information Retrieval System for Documents in Indonesian Using Local Context Analysis*

LARAS MUTIARA DIVA, SONY HARTONO WIJAYA\*

#### Abstrak

Pengguna suatu sistem temu kembali sering kali tidak tepat mengungkapkan kebutuhan informasi yang diinginkannya dalam bentuk kueri. Masalah lain ialah adanya perbedaan pilihan kata antara seorang pengguna dalam kuerinya dan penulis dalam dokumennya. Analisis konteks lokal adalah ekspansi kueri otomatis yang mengombinasikan teknik global dan teknik lokal. Analisis konteks lokal mengurutkan konsep berdasarkan pada kemunculannya dengan seluruh *term* kueri pada dokumen peringkat teratas dan menggunakan konsep peringkat teratas untuk ekspansi kueri. Pada dasarnya suatu dokumen mempunyai beberapa topik sehingga pada penelitian ini dokumen peringkat teratas dibagi ke dalam beberapa *passage*. Konsep peringkat teratas diambil dari beberapa *passage* peringkat teratas. Tujuan penelitian ini ialah mengimplementasikan ekspansi kueri menggunakan analisis konteks lokal. Kinerja dari sistem temu kembali informasi menggunakan analisis konteks lokal bagus dengan nilai ketepatan rata-rata sebesar 60%. Hasil penelitian menunjukkan bahwa kinerja sistem dengan analisis konteks lokal secara signifikan meningkat 6.07% dibandingkan dengan sistem tanpa analisis konteks lokal dengan dokumen-dokumen relevan yang ditemukembalikan berada pada posisi teratas penemukembalian. Selain itu, jumlah dokumen dan *passage* peringkat teratas yang terambil secara signifikan tidak mempengaruhi nilai ketepatan rata-rata. Faktor yang lebih mempengaruhi adalah jumlah *term* ekspansi yang ditambahkan. Analisis konteks lokal cukup baik diterapkan pada koleksi dokumen yang memiliki kemiripan cukup tinggi.

Kata kunci: analisis konteks lokal, ekspansi kueri

#### Abstract

*Users express their needs in a query to get information using an information retrieval system. However, many are not able to compose appropriate queries as the average length of their queries were too short. Another problem is word mismatch, which refers to the phenomenon that the users of information retrieval systems often use words to describe the concept in their queries which are different from words that authors use to describe the same concept in their documents. Local context analysis is an automatic query expansion which is a combination of global and local techniques. Local context analysis ranks the concept by their co-occurrences with the query term in the top ranked documents and uses the highest ranked concepts for query expansion. Basically, a document consists of topics, so in this research, the top ranked documents are divided into passages which represent topics in the relevant document. The highest ranked concepts are then taken from top ranked passages. The purpose of this research is to implement query expansion with local context analysis. The performance of information retrieval system with local context analysis gave good result with around 60% average precision. The results showed that the retrieval performance using local context analysis was significantly higher based on statistical analysis using t-test. The average precision increased by 6.07% compared to retrieval without local context analysis, indicating relevant documents occur higher in the retrieval result. The results also showed that the number of top-ranked documents and passages did not significantly affect the average precision. The more influential factor was the number of query expansions added. Local context analysis is quite suitable for collection of relatively similar documents.*

*Keywords: local context analysis, query expansion*

## PENDAHULUAN

Saat ini arus informasi berkembang dengan sangat cepat. Dalam rangka mendapatkan informasi dari Internet dan berbagai tempat penyimpanan data, pengguna mengungkapkan kebutuhan informasi yang diinginkannya dalam bentuk kueri (Yousef *et al.* 2010). Masalah yang dihadapi ialah pengguna mungkin tidak tepat dan tidak mampu merepresentasikan kebutuhan informasi yang diinginkannya dalam bentuk kueri. Masalah lainnya ialah adanya ketidakcocokan kata antara pengguna dan penulis. Persentase tertinggi untuk panjang kata yang digunakan oleh pengguna dalam kuerinya ialah satu sampai tiga kata (EH 2010). Hal ini membuat sistem temu kembali informasi menghasilkan banyak informasi yang tidak terkait dengan kebutuhan pengguna.

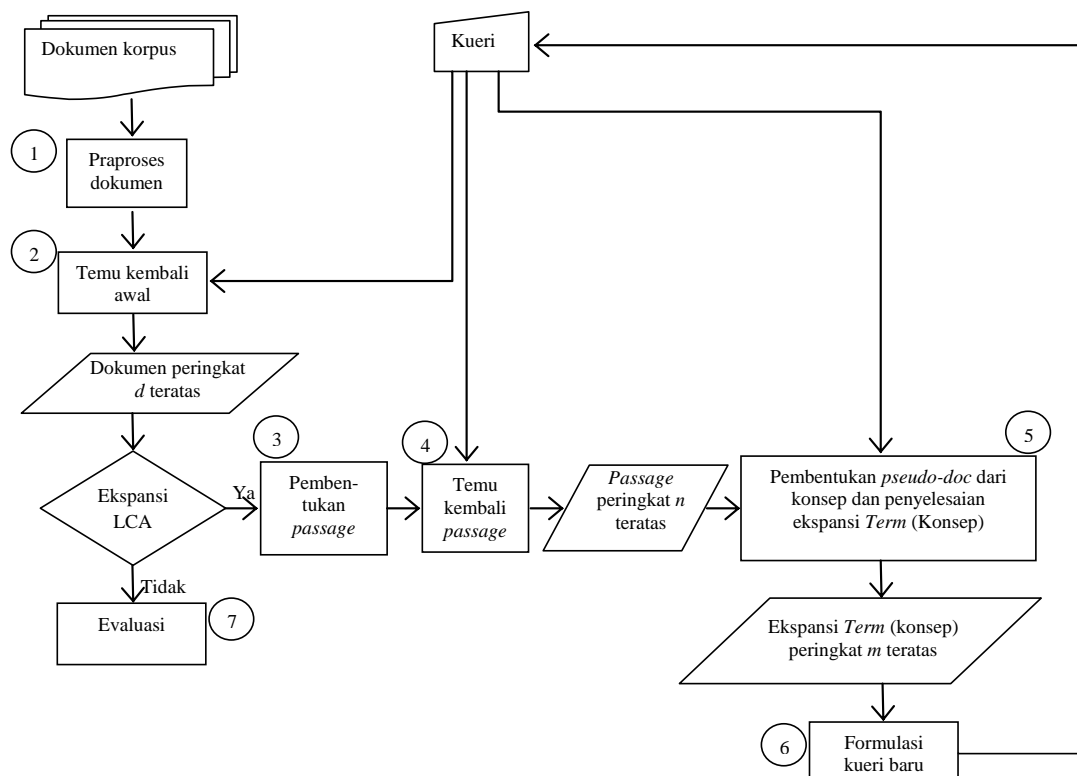
Salah satu cara untuk mengatasi masalah tersebut ialah dengan teknik ekspansi kueri. Pada penelitian ini teknik yang digunakan merupakan perluasan dari teknik global dan teknik lokal yaitu ekspansi kueri menggunakan analisis konteks lokal (Xu dan Croft 2000).

Tujuan penelitian ini ialah mengimplementasikan teknik ekspansi kueri untuk meningkatkan relevansi hasil pencarian pada sistem temu kembali informasi menggunakan analisis konteks lokal.

## METODE

Koleksi dokumen dalam penelitian ini adalah dokumen-dokumen teks berbahasa Indonesia tentang pertanian yang tersedia di Laboratorium Temu Kembali Informasi, Departemen Ilmu Komputer IPB yang berjumlah 1000 dokumen. Kata perluasan (konsep) merupakan seluruh kata tunggal dan unik (tanpa *stopword*) yang ada pada beberapa *passage* peringkat teratas. Evaluasi kinerja sistem temu kembali menggunakan 30 kueri yang masing-masing mempunyai himpunan dokumen yang relevan.

Penelitian ini dilakukan dalam tujuh tahap. Gambaran umum proses temu kembali informasi menggunakan ekspansi kueri dengan LCA dapat dilihat pada Gambar 1.



Gambar 1 Metodologi pelaksanaan penelitian

### **Praproses Dokumen**

Tahapan dimulai dengan mengubah dokumen-dokumen di korpus yang berformat teks menjadi satu *file* dokumen berformat XML. Tiga tahapan praproses yang digunakan pada penelitian ialah tokenisasi, pembuangan *stopword*, dan *stemming*. Tokenisasi adalah proses pembacaan karakter per karakter dan pembersihan separator untuk mengidentifikasi kata di dalam teks. *Stopword* adalah kata yang tidak bisa digunakan sebagai pencari suatu dokumen. *Stemming* adalah proses pembuangan prefiks dan sufiks (secara morfologi) dari suatu kata berimbuhan untuk mendapatkan kata dasar sehingga suatu kata tidak mempunyai banyak variasi bentuk.

### **Temu Kembali Awal**

Pengindeksan dilakukan pada dokumen XML hasil praproses dengan Sphinx. Temu kembali awal menggunakan pembobotan BM25 yang ada pada Sphinx untuk mendapatkan dokumen-dokumen yang sesuai dengan kueri yang diberikan pengguna. Selanjutnya diambil 20 dokumen (nilai optimal pada penelitian ini) peringkat teratas dari hasil pencarian tersebut.

### **Pembentukan Passage**

Dalam proses pembentukan *passage*, teknik *discourse passage* diimplementasikan ke dalam dokumen peringkat teratas. *Discourse passage* merupakan *passage* yang dibentuk berdasarkan pada bagian-bagian dari suatu dokumen, salah satunya menggunakan kalimat.

*Passage* dibentuk dari kumpulan dua kalimat yang saling *overlapping*. Bagian judul pada tiap dokumen dijadikan sebagai kalimat pertama, sedangkan untuk kalimat kedua dan berikutnya diambil dari bagian konten pada dokumen. Kalimat pada konten dipisahkan oleh titik atau suatu baris baru. Setiap dokumen peringkat teratas, kalimat pertama digabung dengan kalimat kedua menjadi satu *passage*, kalimat kedua dan kalimat ketiga menjadi *passage* berikutnya, dan begitu seterusnya untuk setiap dokumen peringkat teratas. Kalimat terakhir pada setiap dokumen dipasangkan lagi dengan judul (kalimat pertama) pada dokumen yang bersangkutan.

### **Temu Kembali Passage**

Langkah awal yang dilakukan ialah tokenisasi untuk mendapatkan semua *term* unik pada semua *passage*. Frekuensi kemunculan *term* unik tersebut di semua *passage* dicari, begitu pula dengan panjang dari masing-masing *passage*. Informasi tersebut digunakan pada saat melakukan pembobotan dengan BM25. Berikutnya dengan menggunakan pembobotan BM25 dan fungsi kesamaannya dicari nilai kemiripan antara *passage* dan kueri pengguna dan didapatkan *passage* peringkat  $n$  teratas yang menjadi hasil dari tahap temu kembali *passage*. Nilai  $n$  optimal pada penelitian ini ialah sebesar 2 *passage*.

### **Pembentukan Pseudo-Doc dari Konsep dan Penyeleksian Term Ekspansi**

Dari *passage* peringkat teratas diambil *term* unik, *Term* unik ini kemudian disebut konsep (calon kueri perluasan). *Term* kueri yang diberikan pengguna dihitung kemunculannya pada setiap *passage* peringkat teratas dengan konsep yang ada menggunakan persamaan:

$$co(c, k_i) = \sum_{j=1}^n pf_{i,j} \cdot pf_{c,j} \quad (1)$$

Informasi *term* kueri dan konsep yang muncul bersamaan disimpan pada *pseudo-doc* beserta nilai kemunculannya. *Pseudo-doc* kemudian digunakan dalam pencarian nilai kemiripan konsep dengan semua *term* yang ada pada kueri menggunakan persamaan:

$$bel(Q, c) = \prod_{k_i \in Q} (\delta + co_{degree}(c, k_i))^{idf_i} \quad (2)$$

Hasil dari tahap ini berupa konsep-konsep yang telah terurut berdasarkan pada kemiripannya dengan keseluruhan kueri.

### Formulasi Kueri Baru

Kata yang muncul dengan seluruh *term* kueri di *passage* peringkat teratas digunakan untuk merumuskan kueri baru yang diformulasikan sebagai berikut:

$$q' = q \cup q_t \quad (3)$$

dengan:

$q'$  = formulasi kueri baru,

$q$  = formulasi kueri awal,

$q_t$  = kata perluasan yang didapat menggunakan teknik LCA.

Pada penelitian ini, banyak kata/konsep yang ditambahkan sehingga mendapatkan hasil yang optimal, yaitu 6 konsep.

### Evaluasi

Evaluasi dilakukan untuk melihat pengaruh jumlah konsep (jumlah kueri perluasan) yang ditambahkan pada kueri awal, pengaruh banyaknya pengambilan dokumen peringkat teratas yang ditemukembali dengan kueri awal, pengaruh banyaknya pengambilan *passage* peringkat teratas yang ditemukembali dengan kueri awal, pengaruh penggunaan LCA, dan pengaruh panjang kueri asli pada nilai AVP sistem temu kembali menggunakan LCA.

## HASIL DAN PEMBAHASAN

### Koleksi Dokumen Pengujian

Deskripsi dokumen penyajian (Tabel 1) ditandai dengan *tag*  $\langle DOC \rangle \langle /DOC \rangle$ , mewakili keseluruhan dokumen dan melingkupi tag-tag lain yang lebih spesifik;  $\langle DOCNO \rangle \langle /DOCNO \rangle$ , menunjukkan *ID* dari dokumen;  $\langle TITLE \rangle \langle /TITLE \rangle$ , menunjukkan judul dari dokumen;  $\langle DATE \rangle \langle /DATE \rangle$ , menunjukkan tanggal dari berita atau penulisan;  $\langle AUTHOR \rangle \langle /AUTHOR \rangle$ , menunjukkan penulis dari berita tersebut;  $\langle TEXT \rangle \langle /TEXT \rangle$ , menunjukkan isi dari dokumen.

### Pengujian Kinerja Sistem

Proses evaluasi menggunakan 30 pasangan kueri uji dan dokumen-dokumen yang relevan dengan kueri uji tersebut. Evaluasi bertujuan mendapatkan nilai *recall* dan *precision* dari sistem berdasarkan pengujian terhadap kueri dan dokumen relevannya. Selanjutnya dilakukan interpolasi maksimum untuk mencari nilai AVP dari sistem.

Kinerja sistem temu kembali tanpa ekspansi yang digunakan sebagai pembanding untuk sistem temu kembali menggunakan ekspansi dengan metode LCA. Nilai AVP yang dihasilkan dari sistem temu kembali tanpa ekspansi ialah sebesar 0.56683, sedangkan nilai *recall* yang dihasilkan sebesar 0.77760. Hal ini menunjukkan bahwa pencarian tanpa ekspansi rata-rata menemukembali 78% dokumen dengan tingkat relevansi sebesar 56.6%.

Tabel 1 Deskripsi koleksi pengujian

Uraian dokumen pertanian	Nilai
Ukuran keseluruhan dokumen	4139332 <i>byte</i>
Ukuran rata-rata dokumen	4139 <i>byte</i>
Ukuran dokumen terbesar	54278 <i>byte</i>
Ukuran dokumen terkecil	451 <i>byte</i>
Rata-rata jumlah kata per dokumen (termasuk <i>stopword</i> )	585.967 kata
Jumlah kueri pengujian	30 kueri
Rata-rata jumlah kata per kueri	2.567 kata
Rata-rata jumlah dokumen relevan per kueri	31.9 dokumen

### Pengaruh Jumlah Konsep (Jumlah Kueri Perluasan) yang Ditambahkan pada Kueri Awal

Nilai AVP untuk setiap jumlah konsep yang ditambahkan dilihat perbandingannya (Tabel 2). Jumlah dokumen dan *passage* dibuat tetap, yaitu dokumen 10 teratas dan *passage* 2 teratas. Semakin banyak konsep yang ditambahkan pada kueri awal, nilai AVP yang diperoleh semakin tinggi. Namun, sampai pada batas tertentu (banyak konsep lebih besar 7) nilai AVP akan kembali turun.

Penambahan lebih dari 8 konsep teratas tidak dapat diketahui hasilnya karena pada kueri pengujian tertentu konsep yang didapatkan pada *passage* peringkat teratas tidak mencapai jumlah lebih dari 8 konsep.

Statistik uji *t* dilakukan untuk membandingkan sistem tanpa ekspansi dan sistem dengan ekspansi LCA dengan menambahkan 7 konsep. Nilai *p-value* yang didapatkan ialah sebesar 0.004 ( $p\text{-value} < 0.05$ , tolak  $H_0$ ) sehingga dapat disimpulkan bahwa kinerja temu kembali menggunakan ekspansi sebanyak 7 konsep berbeda nyata dengan tanpa menggunakan ekspansi.

### Pengaruh Banyaknya Pengambilan Dokumen Peringkat Teratas yang Ditemukembalikan dengan Kueri Awal

LCA mengambil beberapa dokumen peringkat teratas yang diasumsikan sebagai dokumen relevan untuk diproses dalam tahap selanjutnya. Agar pengaruh terlihat pada jumlah dokumen saja, jumlah *passage* peringkat teratas dibuat tetap, yaitu 2 *passage* dan 10 *passage* serta jumlah kata yang diekspansi juga tetap, yaitu 6 kata.

Perbedaan jumlah dokumen peringkat teratas yang diambil secara keseluruhan tidak menimbulkan perbedaan yang nyata pada nilai AVP dari sistem temu kembali menggunakan LCA pada Tabel 3. Seperti yang terlihat, semakin bertambahnya jumlah dokumen peringkat teratas yang diambil tidak selalu mengakibatkan nilai AVP semakin naik ataupun semakin turun, perubahan nilai AVP cenderung tidak berpola dan tidak jauh berbeda (Tabel 3).

Untuk membuktikan tidak terdapat perbedaan yang signifikan terhadap nilai AVP akibat pengambilan jumlah dokumen yang berbeda dilakukan statistik uji *t* yang membandingkan

Tabel 2 Perbandingan nilai AVP pada penambahan sejumlah konsep

Jumlah konsep	Nilai AVP	%
1	0.54576	-3.72
2	0.57024	+0.60
3	0.58739	+3.63
4	0.59559	+5.07
5	0.59879	+5.64
6	0.60009	+5.87
7	0.60035	<b>+5.91</b>
8	0.59999	+5.85

Tabel 3 Perbandingan nilai AVP pada pengambilan sejumlah dokumen peringkat teratas untuk jumlah konsep/kueri perluasan sebanyak 6

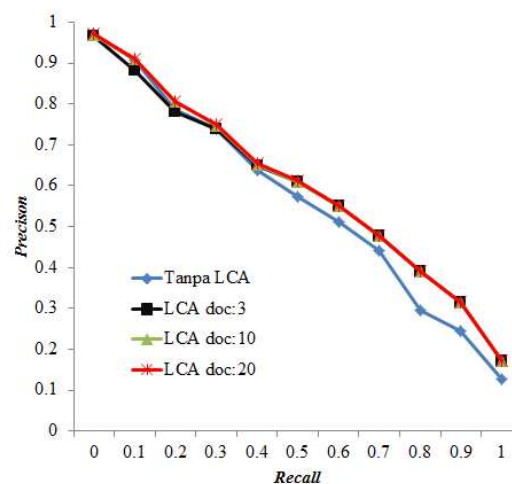
Jumlah dokumen	Jumlah <i>passage</i>	
	2	10
1	0.59678 (+5.28%)	0.59661(+5.25%)
2	0.59677(+5.28%)	0.59765(+5.43%)
3	0.59400(+4.79%)	0.59553 (+5.06%)
5	0.59716 (+5.35%)	0.59687(+5.29%)
10	0.60009 (+5.87%)	0.59658(+5.25%)
20	0.60122 (+ <b>6.07%</b> )	0.59810(+5.52%)

antara jumlah dokumen 3 (nilai AVP terkecil) dan jumlah dokumen 20 (nilai AVP terbesar) dengan jumlah *passage* sebesar 2. Dari hasil uji *t* didapatkan nilai *p-value* sebesar 0.051 ( $p\text{-value} > 0.05$ , terima  $H_0$ ) sehingga dapat disimpulkan bahwa kinerja sistem temu kembali menggunakan jumlah dokumen 3 dan 20 tidak berbeda nyata (Gambar 2). Hal tersebut disebabkan masih terdapatnya dokumen yang tidak relevan di peringkat teratas sehingga semakin banyak dokumen yang diambil mengakibatkan kemungkinan dokumen yang tidak relevan diambil juga semakin banyak.

Nilai *precision* pada tiap titik *recall* untuk pengambilan berbagai jumlah dokumen teratas dengan LCA tidak terdapat perbedaan, kecuali pada pengambilan 3 dokumen peringkat teratas dengan nilai *precision* untuk titik *recall* 0.1 dan 0.2 mengalami sedikit penurunan. Secara keseluruhan dapat disimpulkan bahwa banyaknya jumlah dokumen peringkat teratas yang diambil tidak mempengaruhi kinerja sistem.

### Pengaruh Banyaknya Pengambilan *Passage* Peringkat Teratas yang Ditemukembalikan dengan Kueri Awal

Pada LCA setelah didapatkan dokumen peringkat teratas, berikutnya ialah pengambilan *passage* peringkat teratas. Untuk melihat pengaruh jumlah *passage* saja, jumlah dokumen peringkat teratas dibuat tetap, yaitu 2 dokumen dan 10 dokumen serta jumlah kata yang diekspansi juga tetap, yaitu 6 kata. Perbedaan jumlah *passage* peringkat teratas yang diambil secara keseluruhan tidak menimbulkan perbedaan yang nyata pada nilai AVP dari sistem temu kembali menggunakan LCA (Tabel 4). Terlihat pada Tabel 4, semakin bertambah jumlah *passage* peringkat teratas yang diambil tidak selalu mengakibatkan nilai AVP semakin naik ataupun semakin turun, perubahan nilai AVP cenderung tidak berpola dan tidak jauh berbeda. Perbedaan yang tidak nyata tersebut dibuktikan menggunakan statistik uji *t* dengan melakukan perbandingan kinerja antara jumlah *passage* 2 (nilai AVP terbesar) dan jumlah *passage* 10 (nilai



Gambar 2 Perbandingan grafik R-P tanpa LCA dan dengan LCA pada pengambilan 3, 10, dan 20 dokumen teratas

Tabel 4 Perbandingan nilai AVP pada pengambilan sejumlah *passage* peringkat teratas untuk jumlah konsep/kueri perluasan sebanyak 6

Jumlah <i>passage</i>	Jumlah dokumen	
	2	10
2	0.59677 (+5.28%)	0.60009 (+5.87%)
5	0.59762 (+5.43%)	0.59833 (+5.56%)
10	0.59765 (+5.44%)	0.59658 (+5.25%)
15	0.59765 (+5.44%)	0.59903 (+5.68%)
20	0.59802 (+5.50%)	0.59853 (+5.59%)



AVP terkecil) dengan jumlah dokumen sebesar 10. Hasil uji *t* didapatkan nilai *p-value* sebesar 0.126 (*p-value* > 0.05, terima H0) sehingga dapat disimpulkan bahwa kinerja sistem temu kembali menggunakan jumlah *passage* 2 dan 10 tidak berbeda nyata.

Nilai *precision* di setiap titik *recall* untuk berbagai jumlah pengambilan *passage* teratas pada LCA tidak terdapat perbedaan sehingga dapat disimpulkan bahwa banyaknya jumlah *passage* peringkat teratas yang diambil tidak mempengaruhi kinerja sistem (Gambar 3).

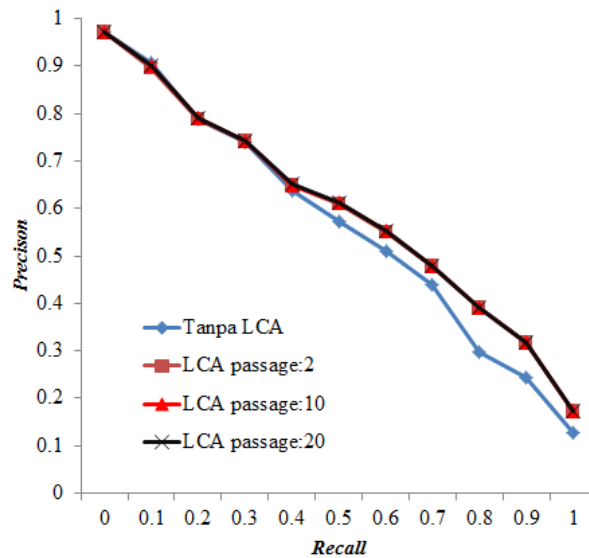
**Pengaruh Penggunaan Metode Analisis Konteks Lokal**

Perbandingan kinerja sistem berdasarkan pada hasil temu kembali dengan ekspansi menggunakan LCA dan temu kembali tanpa ekspansi LCA dapat dilihat pada Tabel 5. Pencarian tanpa LCA didapatkan nilai AVP sebesar 0.56683, sedangkan pada pencarian dengan LCA didapatkan nilai AVP optimal sebesar 0.60122, yakni terjadi peningkatan sebesar 6.07% dari pencarian tanpa LCA (Tabel 5). Hasil penelitian menunjukkan nilai *recall* untuk pencarian dengan LCA sebesar 0.89430 yang menunjukkan bahwa pencarian dengan ekspansi LCA rata-rata menemukembalikan 89% dokumen dengan tingkat relevansi sebesar 60.1% (0.60122). Nilai AVP optimal untuk system dengan LCA yang diperoleh dalam penelitian ini dicapai ketika jumlah dokumen peringkat teratas sebanyak 20, jumlah *passage* peringkat teratas yang diambil sebanyak 2, dan konsep yang ditambahkan sebanyak 6.

**Pengaruh Panjang Kueri Asli**

Pada penelitian ini digunakan tiga kelompok kueri, yaitu kueri dengan panjang 2, 3, dan 4 kata. Ketiga kelompok kueri tersebut selanjutnya diekspansi dengan menambahkan 2 kata ekspansi dan 6 kata ekspansi, seperti yang terlihat pada Tabel 6. Ekspansi kueri untuk setiap kelompok kueri diambil dari 10 dokumen teratas dan 2 *passage* teratas.

Nilai AVP untuk ekspansi kueri dengan penambahan sebanyak 6 kata ekspansi lebih tinggi daripada penambahan 2 kata ekspansi untuk setiap panjang kueri yang digunakan



Gambar 3 Perbandingan grafik R-P tanpa LCA dan dengan LCA pada pengambilan 2, 10, dan 20 *passage* teratas

Tabel 5 Perbandingan nilai AVP tanpa ekspansi LCA dan dengan ekspansi LCA

Perlakuan sistem	AVP
Tanpa LCA	0.56683
LCA (optimal berdasarkan batasan penelitian)	0.60122 (+6.07%)

Tabel 6 Perbandingan Nilai AVP untuk setiap panjang kueri

Panjang kueri	Tanpa ekspansi	Ekspansi 2 kata	Ekspansi 6 kata
2 kata	0.66175	0.66626	0.69071
3 kata	0.54554	0.54417	0.58229
4 kata	0.28213	0.31123	0.33846

(Tabel 6). Selain panjang kueri 3 kata dengan penambahan sebanyak 2 kata ekspansi yang mengalami penurunan dibandingkan dengan tanpa ekspansi (0.54554 ke 0.54417), nilai AVP pada berbagai kelompok panjang kueri yang digunakan meningkat untuk sistem yang menggunakan ekspansi LCA dibandingkan dengan tanpa menggunakan ekspansi LCA. Oleh karena itu, dapat disimpulkan bahwa kueri panjang 2, 3, dan 4 kata sama-sama dapat meningkatkan nilai AVP pada penambahan beberapa kata perluasan.

## SIMPULAN

Penelitian ini menganalisis ekspansi kueri menggunakan metode analisis konteks lokal. Ekspansi kueri menggunakan analisis konteks lokal dapat meningkatkan kinerja sistem temu kembali, yakni dokumen yang relevan berada pada urutan awal pada penemukembalian. Variasi jumlah dokumen dan *passage* peringkat teratas yang diambil tidak memberikan perbedaan yang signifikan satu sama lain pada hasil kinerja sistem temu kembali. Variasi jumlah kueri perluasan yang ditambahkan mempengaruhi kinerja sistem temu kembali. Pada pengambilan 10 dokumen peringkat teratas dan 2 *passage* peringkat teratas semakin banyak konsep yang ditambahkan maka semakin bagus kinerja sistem sampai pada batas tertentu, yaitu 7 konsep. Penambahan beberapa kata perluasan pada kueri dengan panjang 2, 3, dan 4 kata sama-sama dapat meningkatkan kinerja temu kembali.

Saran untuk pengembangan sistem ini selanjutnya ialah penggunaan koleksi yang lebih besar untuk lebih mendekati penggunaan sesungguhnya, penambahan frasa sebagai kata yang akan digunakan untuk ekspansi untuk mengatasi keambiguan kata, dan pembedaan bobot antara kata yang ada pada kueri awal pengguna dan kata ekspansi yang diperoleh.

## DAFTAR PUSTAKA

- [EH] Experian Hitwise. 2010. Google received 70 percent of US searches [Internet]. [diunduh 2011 Des 13]. Tersedia pada: <http://www.hitwise.com/ca/press-center/press-releases/google-searches-mar-10>.
- Xu J, Croft WB. 2000. Improving the effectiveness of informational retrieval with local context analysis. *J ACM Trans Inform Syst.* 18(1):79-112.
- Yousef N, Al-Bidewi I, Fayoumi M. 2010. Evaluation of different query expansion techniques and using different similarity measures in Arabic documents. *Euro J Scic Res.* 43(1):156-166.