KINERJA COSINE SIMILARITY DAN SEMANTIC SIMILARITY DALAM PENGIDENTIFIKASIAN RELEVANSI NOMOR HALAMAN PADA DAFTAR INDEKS ISTILAH

Sherly Christina
Teknik Informatika Universitas Palangka Raya
Kampus Unpar Tunjung Nyaho Jl. Yos Sudarso, Palangka Raya 73112
E-mail: sherly.christina.upr@gmail.com

ABSTRACT
Index term is a navigation tool for the reader to find the terms on each page of book. Each indexed term on each page should relevant to the meaning of indexed term. The inaccurate indexing process can result in the irrelevant page numbers. This study aim to test the performance of Cosine Similarity and Semantic Similarity framework in detecting the relevancy of page number in the list of index term of electronic books in English. This study will make Cosine Similarity framework and Semantic Similarity framework that will be supported by WordNet database. The result of the testing and evaluation phase shows Kappa values of Semantic Similarity framework is better than Cosine Similarity framework. Kappa value of Semantic Similarity framework shows that the Semantic Similarity method has good performance to detect the relevant page numbers in the list of index terms.
Keywords: cosine similarity, semantic similarity, index, term

ABSTRAK
Indeks istilah adalah alat navigasi bagi pembaca untuk menemukan istilah-istilah pada halaman-halaman buku. Setiap istilah yang diacu oleh daftar indeks pada tiap halaman buku seharusnya saling memiliki relevansi atau keterkaitan makna dengan kata istilah yang diindeks. Tetapi proses pengindeksan yang kurang akurat dapat mengakibatkan nomor halaman yang diacu tidak relevan, sehingga dapat menimbulkan kesulitan bagi pembaca memahami makna suatu istilah. Pada penelitian ini dilakukan penelitian untuk menguji kinerja Cosine Similarity dan Semantic Similarity dalam pendeteksian relevansi nomor halaman pada daftar indeks buku-buku elektronik berbahasa Inggris. Pada penelitian ini dibuat kerangka kerja Cosine Similarity dan kerangka kerja Semantic Similarity dengan dukungan basis data Wordnet. Hasil pengujian dan evaluasi menunjukkan Semantic Similarity menghasilkan nilai Kappa yang lebih baik dibandingkan Cosine Similarity. Nilai kappa yang lebih baik pada kerangka kerja Semantic Similarity menunjukkan bahwa metode Semantic Similarity memiliki kinerja yang baik untuk mendeteksi nomor halaman yang relevan di dalam daftar indeks istilah.
Kata Kunci: cosine similarity, semantic similarity, indeks, istilah

1. PENDAHULUAN

Perkembangan teknologi yang cepat menyediakan model dan kapasitas penyimpanan digital yang semakin luas. Dokumen-dokumen digital adalah sumber daya informasi yang besar oleh karena itu dibutuhkan cara yang efektif untuk memperoleh subyek informasi dari suatu dokumen.

Indeks istilah (*index term*) yang disebut juga indeks subjek, atau deskriptor dalam pencarian informasi, berfungsi menyatakan esensi dari topik dokumen. Daftar indeks istilah pada dokumen seperti pada buku elektronik adalah salah satu cara untuk mempermudah pembaca menemukan informasi yang dibutuhkan.

Pengindeksan istilah-istilah dari suatu dokumen membutuhkan upaya yang besar, karena sifat indeks istilah yang subjektif membutuhkan pengetahuan pakar yang memahami konteks buku serta memerlukan waktu dan energi yang cukup besar.

Masalah keakuratan pada daftar indeks istilah suatu buku menjadi latar belakang penelitian yang dilakukan oleh Christina (2012). Akurasi pada daftar indeks istilah dapat ditingkatkan bila nomor-nomor halaman yang diindeks relevan, dan relevansi ditemukan bila setiap kata istilah yang muncul pada nomor-nomor halaman tersebut memiliki nilai keterkaitan semantik yang tinggi dengan kata istilah yang diindeks. (Christina, 2012).

Pada penelitian ini dilakukan pengujian untuk membandingkan kinerja *Cosine Similarity* dengan kinerja *Semantic Similarity* yang diusulkan oleh Christina(2012). Pengujian dan evaluasi pada penelitian ini dilakukan untuk melihat kemampuan teknik sintatik menggunakan *Cosine Similarity* dan teknik Semantik menggunakan *Semantic Similarity* dengan dukungan basis data Wordnet untuk mendeteksi relevansi antara nomor-nomor halaman yang diindeks.

2. TINJAUAN PUSTAKA

Beberapa penelitian telah dilakukan untuk meningkatkan kinerja dalam pengindeksan buku. Lahtinen menganalisis konten dokumen dengan cara menggabungkan frekuensi kata yang muncul dengan analisis linguistik yang disediakan oleh parser sintaksis (Lahtinen, 2000). Kemudian Duque dan Lobin melakukan penelitian menggunakan Pemrosesan Bahasa Alami dan ontologi untuk mengindeks istilah-istilah dari koleksi teks suatu dokumen (Duque & Lobin, 2004). Serta penelitian yang dilakukan oleh Medelyan dan Witten untuk mengekstraks istilah-istilah dari dokumen dengan menggunakan teknik mesin pembelajaran KEA++ yang dilatih berdasarkan fitur-fitur kandidat indeks istilah yang diperoleh dari skor TFxIDF dan *position of the first occurence* dari kandidat indeks istilah (Medelyan &Witten, 2005).

Penelitian-penelitian tersebut dilakukan untuk meningkatkan kinerja pengindeksan secara otomatis. Sedangkan penelitian yang dilakukan oleh Christina mencoba untuk meningkatkan akurasi dari referensi silang pada daftar indeks istilah, dengan cara mendeteksi ambiguitas yang disebabkan oleh kata istilah yang muncul pada halaman buku yang tidak relevan (Christina,2012).

Fungsi dari daftar indeks istilah sangat penting untuk membantu pembaca menemukan istilah-istilah yang maknanya saling terkait dalam tiap halaman buku. Sehingga nomor-nomor halaman yang diacu dalam daftar indeks seharusnya saling memiliki relevansi atau keterkaitan makna dengan istilah yang diindeks, agar pembaca dapat memahami makna dari istilah tersebut (Diodato, 1991).

Berikut ini adalah contoh nomor halaman yang tidak relevan dalam daftar indeks istilah (Christina, 2012). Misalkan pada sebuah buku berjudul "*Microsoft SQL Server 2000 Database-Design*", terdapat indeks istilah *table* yang mengacu pada dua nomor halaman. Istilah *table* yang muncul di salah satu nomor halaman yang diacu memiliki makna yang terkait dengan konteks buku. Tetapi istilah *table* yang muncul pada nomor halaman lainnya tidak terkait dengan konteks buku tetapi merupakan elemen isi buku seperti gambar atau grafik.

Contohnya pada buku "*Microsoft SQL Server 2000 Database-Design*" (Whallen, 2000), terdapat referensi silang *table 16,23* :

"A table consists of rows and columns; these rows and columns contain the data for the table." (halaman 16)

"If you're like most people, you'll find that you frequently need to use these stored procedures, so you should memorize the stored procedures in this table." (halaman 23)

Kalimat berisi istilah *table* yang muncul pada halaman nomor 16 memiliki makna yang terkait dengan makna istilah *table* yang diindeks pada buku tersebut, bila dicek keterkaitannya dengan konteks buku mengenai *Database Design.* Sedangkan kalimat berisi istilah *table* pada halaman 23, lebih mengacu pada elemen buku. Sehingga dinyatakan halaman nomor 16 relevan sedangkan halaman nomor 23 tidak relevan dengan istilah yang diindeks. Nomor halaman yang tidak relevan dapat menyebabkan kerancuan atau kebingungan dalam pemahaman makna suatu istilah di dalam buku.

Pada penelitian ini digunakan *Cosine Similarity* sebagai teknik Sintatiks untuk mendeteksi relevansi nomor halaman pada daftar indeks. Teknik sintatiks berhubungan dengan jumlah kata istilah pembentuk kalimat atau teks. *Cosine Similarity* mengukur kemiripan antara dua dokumen atau teks. Pada *Cosine Similarity* dokumen atau teks dianggap sebagai vektor (Sighal, 2001). Sehingga nilai *Cosine Similarity* dari Vektor A dan B dapat dihitung seperti persamaan (1) berikut.

$$Cosine\ Similarity = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad \text{.....(1)}$$

Untuk pencocokan text, nilai dari vektor A dan B adalah vektor *term-frequency* dari dokumen. Nilai *cosine similarity* berada pada range 0-1. Pada penelitian ini*, Cosine Similarity* digunakan untuk menghitung jumlah kata istilah yang muncul pada halaman-halaman yang diacu pada daftar indeks. Semakin banyak jumlah kata istilah yang muncul pada suatu halaman semakin tinggi nilai *Cosine Similarity* yang diperoleh.

Teknik Semantik berhubungan dengan makna kata pembentuk kalimat atau teks. Pada penelitian ini digunakan *semantic similarity* untuk menghitung keterkaitan semantik antara kata istilah yang diindeks dengan kata istilah yang muncul pada nomor halaman yang diacu pada daftar indeks. Pada perhitungan nilai *similarity* digunakan basis data leksikal WordNet.

WordNet adalah sebuah basis data leksikal berbahasa Inggris yang diorganisasikan berdasarkan hubungan semantik. Kata benda, kata kerja dan kata sifat diorganisasikan ke dalam *synonym sets* (*synset*). Tiap *synset* mewakili konsep leksikal dasar. Hubungan semantik yang terbentuk antara lain berdasarkan pada sinonim, antonim, hiponim dan meronim (Miller, 1995).

Pada penelitian ini akan dihitung nilai *similarity* antara kata istilah yang diindeks dengan kalimat atau teks berisi kata istilah yang muncul pada halaman-halaman yang diacu pada daftar indeks. Setiap kata yang menyusun suatu teks akan dihitung keterkaitannya maknanya dengan kata istilah yang

diindeks dengan mengacu pada hirarki taksonomi pada WordNet. Gambar 1 menunjukkan contoh hirarki taksonomi hiponim pada WordNet.

Gambar 1. Contoh Hirarki Taksonomi WordNet (Dao et.al., 2006)

Pada Gambar 1 panjang jalur antar simpul dihitung sebagai berikut.

1. *Car* dengan *auto* adalah 1, *LCS* adalah "*car, auto"*.
2. *Car* dengan *truck* adalah 3, *LCS* adalah *"automotive, motor vehicle"*.
3. *Car* dengan *bicycle* adalah 4, *LCS* adalah *"wheeled vehicle"*.
4. *Car* dengan *fork* adalah 12, *LCS* adalah "*artifact"*.

The least common subsumer (LCS) dari dua *synsets* adalah simpul paling dekat dari dua *synsets*. Panjang jalur adalah jalan untuk menghitung nilai relasi antara dua kata yang memiliki *senses* sama.

Langkah-langkah untuk menghitung *Semantic Similarity* adalah sebagai berikut (Dao *et al.*, 2006).

1. Tokenisasi.
   Tahap tokenisasi adalah tahap pemotongan *string* input berdasarkan tiap kata yang menyusunnya. Sebelum melakukan tokenisasi, setiap huruf dalam dokumen harus diubah ke dalam huruf kecil.
2. *Stemming*.
   Tahap *stemming* adalah tahap mencari akar kata dari tiap kata hasil penyaringan. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama.
3. *Speech tagging*.
   Tahap menentukan *part of speech* suatu kata seperti *noun, verb,*atau *adverb* suatu kata dalam kalimat atau teks.
4. *Word sense disambiguation*.
   Tahap *Word Sense Disambiguation* adalah tahap menemukan makna (*sense*) kata-kata yang menyusun suatu kalimat atau teks.
5. Membuat *Semantic Similarity Relative Matrix* R [m,n] untuk tiap pasang *word sense*, di mana R[i,j] adalah kemiripan semantik antara *senses* yang paling cocok dari kata pada posisi *i* dari kata istilah *X* dengan *senses* yang paling cocok dari kata pada posisi *j* dari kalimat *Y*. Jadi, R[i,j] adalah bobot koneksi tepi dari *i* ke *j*.
6. Menghitung *semantic similarity* menggunakan metode heuristik cepat.

Contoh proses perhitungan *semantic similarity* seperti berikut ini. Misalkan terdapat *S* kalimat berisi *T* istilah *table* dari halaman 16 dan 23:

T: Table
S_1 (hal.16) : *"A table consists of rows and columns; these rows and columns contain the data for the table."*
S_2 (hal.23) : *"If you're like most people, you'll find that you frequently need to use these stored procedures, so you should memorize the stored procedures in this table."*

Kemudian pada *T* dan S_1, S_2 dilakukan prapemrosesan teks yaitu proses tokenisasi, penghilangan *stopwords* dan *stemming*. Hasil prapemrosesan teks pada *T* dan S_1, S_2 seperti berikut.

T : "table"
S_1 *: "table", "consist", " row", "column", "row", "column", "contain", "data"*
S_2 *: "like", "people", "find", "frequently", "need", "use", "stored", "procedure", "memorize", "stored", "procedure", "table"*

Kemudian keterkaitan semantik antara *T* dan S_1, S_2 dicek menggunakan WordNet, dan direpresentasikan dalam bentuk matrik seperti ditunjukkan oleh Gambar 2 dan Gambar 3.

S_1 / *T*	table	consist	row	column	row	column	contain	data	table
table	1	0	0,86	0,86	0,86	0,86	0,14	0,62	1

Gambar 2. Matrik Keterkaitan T dengan S_1

S_2 / *T*	like	people	find	frequently	need	used	stored	procedure	memorize	stored	procedure	table
table	0	0,67	0,35	0,1	0	0	0,17	0,4	0,12	0,17	0,4	1

Gambar 3. Matrik Keterkaitan T dengan S_2

Kemudian nilai *semantic similarity T* dengan (S_1,S_2) akan dihitung dengan metode heuristik cepat, seperti berikut:

$$Sim\,(T, S_1) = \frac{1+(1+0{,}86+0{,}86+0{,}86+0{,}86+0{,}14+0{,}62+1)}{10} = 0{,}72$$

$$Sim\ (T, S_2) = \frac{1 + (0 + 0{,}67 + 0{,}35 + 0{,}1 + 0 + 0 + 0{,}17 + 0{,}4 + 0{,}12 + 0{,}17 + 0{,}4 + 1)}{13}$$
$$= 0{,}44$$

3. METODOLOGI

Agar penelitian ini lebih terarah maka disusun langkah-langkah penelitian, seperti pada Gambar 4.

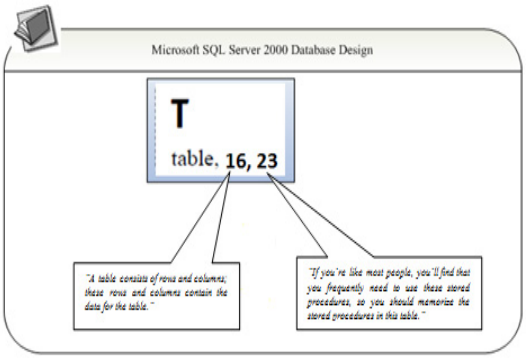
Gambar 4. Langkah-Langkah Penelitian

Langkah Studi Literatur dilakukan untuk memahami konsep-konsep atau dasar-dasar teori yang digunakan dalam tiap tahapan penelitian. Langkah Pengumpulan Data Set adalah tahap mengumpulkan data-data yang diperoleh dari sejumlah buku elektronik berbahasa Inggris. Langkah Perancangan Kerangka Kerja adalah tahap merancang kerangka kerja *Cosine Similarity* dan *Semantic Similarity*. Tahapan Pengujian dan Evaluasi adalah tahap menguji dan mengevaluasi kinerja kerangka kerja *Cosine Similarity* dan *Semantic Similarity*.

4. PEMBAHASAN

Data set yang digunakan pada penelitian ini, diperolah dari beberapa kategori buku elektronik, yaitu, buku-buku politik-sosiologi, ekonomi, bisnis, agama, teknik informatika, studi, umum dan buku terkait binatang. Data set diperoleh dengan mengambil kata istilah dari daftar indeks dan mengambil kalimat atau teks berisi kata istilah dari halaman yang nomornya tercantum pada daftar indeks. Kemudian untuk evaluasi diambil kalimat atau teks kedua dari halaman yang nomornya tidak dicantumkan dalam daftar indeks. Pengambilan kata istilah dan kalimat dilakukan secara acak.

Pada Gambar 5, ditunjukkan ilustrasi pengumpulan data set. Teks berisi kata istilah yang diambil pada halaman 16 adalah halaman yang nomornya telah didefinisikan dalam Daftar Indeks buku. Sedangkan teks berisi kata istilah pada halaman nomor 23 tidak diindeks dalam daftar indeks istilah buku dan akan digunakan untuk evaluasi.

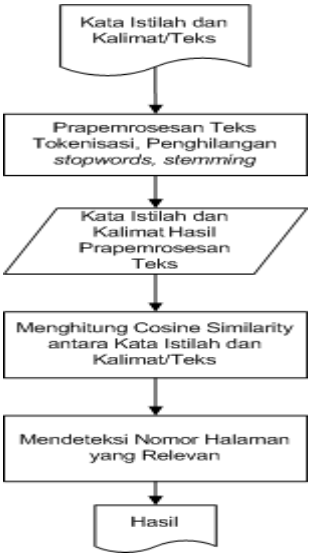


Gambar 5. Ilustrasi Pengumpulan Data Set

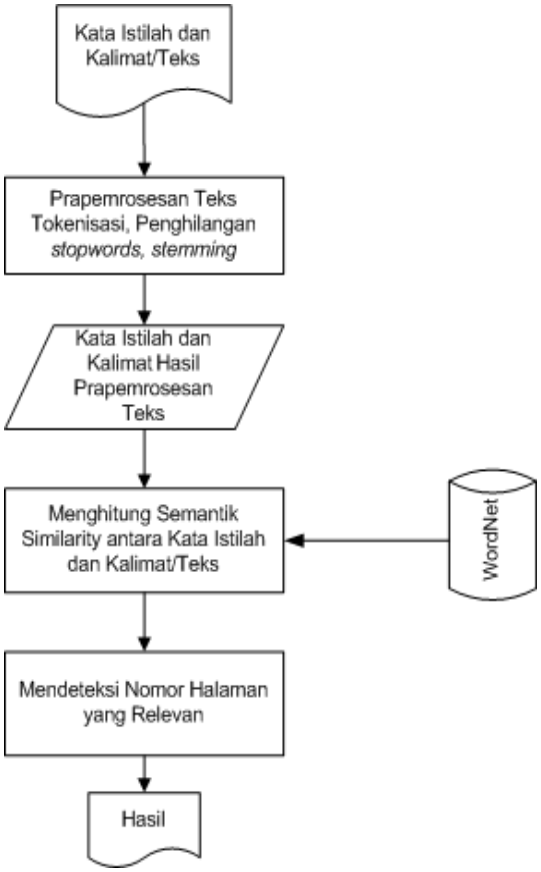
Pada Gambar 6 dan Gambar 7 ditunjukkan kerangka kerja yang digunakan untuk mendeteksi nomor halaman yang relevan dan tidak relevan dengan *Cosine Similarity* dan *Semantic Similarity*. Langkah pertama yang dilakukan pada kerangka kerja *Cosine Similarity* dan *Semantic Similarity* adalah melakukan Prapemrosesan Teks pada kata istilah dan kalimat/teks berisi kata istilah. Prapemrosesan Teks berupa tahap tokenisasi, penghilangan *stopwords* dan *stemming* (Christina, 2012).

Kemudian langkah kedua pada kerangka kerja *Cosine Similarity* adalah menghitung frekuensi kemunculan kata istilah dalam kalimat/teks dan mengukur keterkaitannya dengan kata istilah yang diindeks. Sedangkan langkah kedua pada kerangka kerja *Semantic Similarity* adalah menghitung keterkaitan semantik antara kalimat/teks dan kata istilah dengan mengukur keterkaitan makna antar kata berdasarkan basis data WordNet (Christina, 2012) (Dao et., al, 2006).

Langkah ketiga pada kerangka kerja *Cosine Similarity* dan *Semantic Similarity* adalah membandingkan nilai *similarity* dengan nilai ambang batas. Bila nilai *similarity* lebih dari atau sama dengan ambang batas (θ) maka dinyatakan nomor halaman yang diindeks sudah relevan dengan kata istilah yang diindeks terkait dengan konteks buku.



Gambar 6. Kerangka Kerja *Cosine Similarity*



Gambar 7. Kerangka Kerja *Semantic Similarity*

Pada penelitian ini telah dilakukan pengujian terhadap 727 pasang data set yang diperoleh dari 30 buku dengan kategori yang berbeda. Pengujian dilakukan pada *range* nilai ambang batas 0,45 sampai 0,55. Kemudian dilakukan evaluasi dengan nilai Kappa untuk mengetahui kinerja kerangka kerja dalam mendeteksi nomor halaman yang relevan atau tidak relevan.

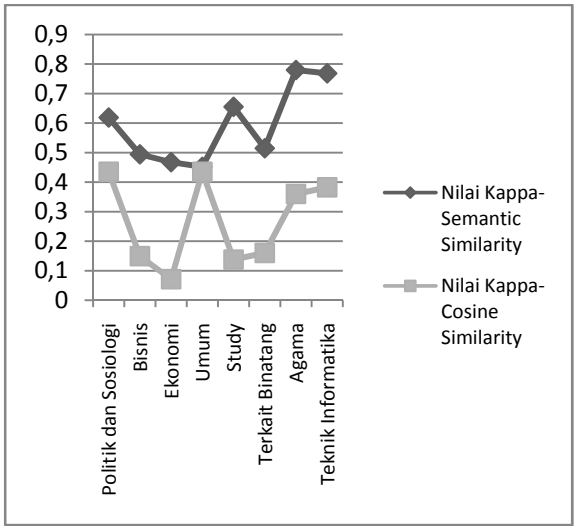
Nilai Kappa akan menunjukkan proporsi kesepakatan antara dua pengamat (Hussain, 2007). Pengamat pada penelitian ini adalah ahli dan kerangka kerja. Peran ahli dalam skenario pengujian pada penelitian ini diwakili oleh daftar indeks yang telah didefinisikan oleh pengindeks buku.

Pengamatan pada hasil pengujian menunjukkan bahwa nilai kappa tertinggi diperoleh pada θ=0,5. Kemudian pada Tabel 1 dan Gambar 8 ditunjukkan hasil evaluasi dengan nilai Kappa terhadap kinerja kerangka kerja *Cosine Similarity* dan *Semantic Similarity* pada θ=0,5 .

Pada Tabel 1 dan Gambar 8 ditunjukkan nilai Kappa pada kerangka kerja *Cosine Similarity* cenderung lebih rendah dibandingkan dengan nilai Kappa pada kerangka kerja *Semantic Similarity*. Nilai kappa yang lebih rendah pada kerangka kerja *Cosine Similarity* menunjukkan kurangnya proporsi kesepakatan antara ahli dengan kerangka kerja dalam menemukan atau mengidentifikasikan nomor halaman yang relevan atau tidak relevan. Sedangkan Nilai kappa yang lebih tinggi pada kerangka kerja *Semantic Similarity* menunjukkan proporsi kesepakatan yang cukup besar antara ahli dan kerangka kerja dalam mengidentifikasikan relevansi nomor halaman dari data set yang tersedia.

Tabel 1. Hasil pengujian dan evaluasi pada kerangka kerja *Cosine Similarity* dan *Semantic Similarity* pada θ=0,5

No	*Data-Set (Buku)*	*Nilai kappa Kerangka Kerja Cosine Similarity*	*Nilai kappa Kerangka Ker Semantic Simi*
1	Politik dan Sosiologi	0,435	0,619
2	Bisnis	0,149	0,494
3	Ekonomi	0,071	0,4678
4	Umum	0,435	0,451
5	Study	0,137	0,655
6	Terkait Binatang	0,160	0,514
7	Agama	0,359	0,779
8	Teknik Informatika	0,382	0,768



Gambar 8. Grafik Hasil Pengujian dan Evaluasi pada Kerangka Kerja Cosine Similarity dan Semantic Similarity

Hasil pengujian dan evaluasi menunjukkan nilai kappa yang diperoleh pada pengujian kerangka kerja *semantic similarity* berkisar antara 0,4-07 yang menunjukkan performa kerangka kerja *semantic similarity* ada pada level *Fair* dan *Moderate* menurut representasi nilai Kappa yang dinyatakan oleh Landis dan Koch (Hussain, 2007). Nilai Kappa yang diperoleh pada kerangka kerja *Semantic Similarity* menunjukkan bahwa proporsi kesepakatan antara ahli dan kerangka kerja *Semantic Similarity* cukup besar dalam mengidentifikasian nomor halaman yang relevan dan nomor halaman yang tidak relevan.

5. KESIMPULAN

Hasil pengujian dan evaluasi pada kerangka kerja *Cosine Similarity* dan *Semantic Similarity* menunjukkan bahwa kerangka kerja *Semantic Similarity* memiliki kinerja yang lebih baik untuk mengidentifikasian nomor halaman yang relevan dan tidak relevan pada daftar indeks istilah. Sehingga kerangka kerja *semantic similarity* dapat berkontribusi untuk menghasilkan daftar indeks yang akurat.

6. SARAN

Pada penelitian yang akan datang disarankan untuk melengkapi basis data Wordnet dengan basis data tesaurus atau Wikipedia untuk mengatasi keterbatasan pada konten Wordnet. Serta menambahkan metode tokenisasi yang dapat mendeteksi frase untuk meningkatkan kinerja kerangka kerja dalam prapemrosesan teks.

Daftar Pustaka

Christina, S. 2012. *Pendeteksian Kerancuan pada Referensi Silang Indeks Istilah*, in Digital Information & Systems Conference. Bandung. hal 56-61

Dao, T. N. & Simpson, T.2006. *Measuring Similarity Between Sentences*

Diodato, V. 1991. Cross-references in back-of-book indexes. *The International Journal of Indexing.* Vol 17. Hal. 178-184

Duque, C. G. & Lobin, Henning. 2004. *Ontology Extraction for Index Generation.* Proceedings of the 8th ICCC. International Conference On Electronic, Uninersidade de Brasilia, hal. 111-120

Hussain, H.M.I. 2007. *Using Text Classsification to Automate Ambiguity Detection in SRS Documents.* Thesis. Concordia University. Montreal. 2007

Lahtinen,T. 2000. *Automatic Indexing: An Approach Using An Index Term Corpus and Combining Linguistic and Statistical Methods.* Thesis. University of Helsinki

Medelyan, O., & Witten, I.H. 2005. *Thesaurus-Based Index Term Extraction for Agricultural Documents.* Proceedings of the 6th Agricultural Ontology Service (AOS). Workshop at EFITA/WCCA. Vila Real. Portugal

Miller, G. 1995. WordNet: A Lexical Database for English. *Communication of The ACM*, volume 38

Singhal, A. 2001. Modern Information Retrieval: A Brief Overview, , *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*

Whalen R. M., *Microsoft SQL Server 2000-Database Design, Instructor Edition*, 2000