ANALISIS PENGARUH METODE UNTUK PENGINDEKSAN BUKU OTOMATIS
(INFLUENCE ANALYSIS METHODS FOR AUTOMATIC BOOK INDEXING)

I Gusti Lanang Putra Eka Prismana
Teknik Informatika, FTIf, Institut Teknologi Sepuluh Nopember Surabaya, glan.putra@gmail.com

Daniel Oranova Siahaan
Teknik Informatika, FTIf, Institut Teknologi Sepuluh Nopember Surabaya, daniel@its-sby.edu

Ahmad Saikhu
Teknik Informatika, FTIf, Institut Teknologi Sepuluh Nopember Surabaya, saikhu@its-sby.edu

Abstrak

Salah satu sumber dalam mencari pengetahuan atau informasi adalah Buku. Buku merupakan salah satu bentuk komunikasi tertua yang berupa tulisan. Saat ini, buku tidak hanya terdapat di dalam media kertas saja, tetapi juga sudah ada di media elektronik (file). Tidak semua buku terdapat indeks buku, Untuk membuat indeks dibutuhkan waktu yang relatif lama dan tenaga ahli / profesional dalam menyusunnya. Bagian terbesar dari indeks buku menjadi konsep penting yang berkaitan dengan buku tersebut dan memberikan informasi halaman terkait dengan konsep tersebut. Banyak metode yang dapat digunakan untuk melakukan pengindeksan buku secara otomatis.

Metode untuk melakukan penghitungan pengaruh pada masing-masing metode dengan menggunakan infogain. Kelima metode yang akan diuji adalah informativeness, phraseness, syntactic, first position of occurence, dan tesaurus. Kelima metode ini yang merupakan metode unggulan dari penelitian yang dilakukan oleh Csomai (Csomai dan Mihalcea, 2008) dan Medelyan (Medelyan dan Witten, 2009).

Metode informativeness mempunyai dampak yang cukup signifikan pada penelitian ini. Berikutnya metode syntactic, dan tesaurus berada di peringkat selanjutnya. Penggunaan ketiga metode tersebut yang memberikan sumbangsih terbesar pada proses pengindeksan otomatis.

Metode phraseness dan FPOC ternyata tidak memberikan dampak yang signifikan terhadap proses pengindeksan otomatis. Detail implementasi diyakini merupakan faktor kunci dalam peningkatan pengaruh masing-masing metode.

Kata kunci: Pengindeksan buku otomatis, *informativeness*, *phraseness, syntactic, first position of occurence,* tesaurus

Abstract

One source in the search for knowledge or information is the book. The book is one of the oldest forms of communication in the form of writing. Currently, the book is not only found in paper media, but also exist in electronic media (files). Not all books are books index, To create the index takes a relatively long time and experts / professionals in arranging. The largest part of the index of the book to be an important concept related to the book and page information associated with the concept. Many methods can be used to perform automatic indexing books.

The method for calculating the effect of each method using Infogain. The fifth method to be tested is informativeness, phraseness, syntactic, first position of occurence, and a thesaurus. The fifth method is a superior method of research conducted by Csomai (Csomai and Mihalcea, 2008) and Medelyan (Medelyan and Witten, 2009).

Informativeness method has a significant impact on this study. Next syntactic method, and the thesaurus is ranked next. The use of these three methods that provide the largest contribution to the process of automatic indexing.

FPOC phraseness method and did not have a significant impact on the process of automatic indexing. Implementation details is believed to be a key factor in the increase of the influence of each method.

Keyword: automatic book indexing, informativeness, phraseness, syntactic, first position of occurence,thesaurus

PENDAHULUAN

Salah satu sumber dalam mencari pengetahuan atau informasi adalah Buku. Buku merupakan salah satu bentuk komunikasi tertua yang berupa tulisan. Saat ini, buku tidak hanya terdapatdi dalam media kertas saja, tetapi juga sudah ada di media elektronik (*file*).

Mudahnya pencarian buku di internet justru berbanding terbalik dengan pencarian informasi di dalam buku. Letak rangkuman informasi yang terdapat di dalam sebuah buku biasanya terletak pada daftar isi, *glossary*, dan indeks. Di dalam daftar isi sebuah buku pada umumnya berisi informasi tentang bab, sub bab, dan sub-sub bab yang terdapat di dalam buku. Informasi yang terdapat di dalam daftar isi tentunya kurang dapat memberikan gambaran tentang isi buku. Bagian lain dari buku yang dapat memberikan gambaran yang cukup dalam adalah bagian indeks.

Tidak semua buku terdapat indeks buku, Untuk membuat indeks dibutuhkan waktu yang relatif lama dan tenaga ahli / profesional dalam menyusunnya. Bagian terbesar dari indeks buku menjadi konsep penting yang berkaitan dengan buku tersebut dan memberikan informasi halaman terkait dengan konsep tersebut.

The National Information Standards Organization mendefinisikan bahwa indeks buku merupakan "Panduan sistematis yang dirancang untuk menunjukkan topik atau metode dokumen dalam rangka memfasilitasi pengambilan dokumen atau bagian dari dokumen" (Anderson dan NISO, 1997). British Standard of 1976 sebagai "Panduan sistematis untuk lokasi kata-kata, konsep atau hal lain di buku, majalah, atau publikasi lainnya. Indeks buku terdiri dari serangkaian daftar, tidak dalam urutan kemunculan dalam publikasi, tetapi dalam beberapa urutan lainnya (misalnya abjad) yang dipilih untuk memungkinkan pengguna untuk menemukannya dengan cepat, bersama-sama dengan referensi yang menunjukkan dimana hal tersebut berada". Dari kedua definisi tersebut dapat disimpulkan bahwa indeks buku merupakan sebuah daftar terstruktur yang berisi informasi penting mengenai isi dokumen sehingga memudahkan pembaca untuk menemukan penjelasan informasi tersebut di dalam dokumen.

Makalah*Linguistically Motivated Features for Enhanced Back-of-the-Book Indexing*(LMFEBI) (Csomai dan Mihalcea, 2008) menegaskan bahwa dalam membuat indeks buku masih diperlukan tenaga manusia atau ahli. Telah ada beberapa bagian yang dapat dibantu oleh komputer dengan perkakas yang dapat mengorganisasikan dan mengubah indeks buku, tetapi tidak ada satu metode yang dapat menggantikannya secara otomatis keseluruhan atau hampir seluruhnya proses pengindeksan.Penelitian lain (Medelyan dan Witten, 2006) menyebutkan bahwa terjadi peningkatan sebesar 110% pada F-Measure dengan menggunakan tesaurus domain terkait.

Dari sekian banyak metode, penelitian ini akan merumuskan pengaruh dari masing-masing metode terhadap proses pengindeksan secara otomatis.

PENGINDEKSAN BUKU OTOMATIS

Pada saat ini dikenal ada dua pendekatan dalam melakukan pengindeksan buku secara otomatis. Pendekatan pertama adalah dengan menggunakan model *supervised* dan pendekatan kedua adalah dengan menggunakan model *unsupervised*. Kedua model tersebut mempunyai kelebihan dan kekurangan yang apabila dipadukan dapat memberikan hasil yang memuaskan.

Model *supervised* menggunakan daftar frasa kunci yang sebelumnya telah dikumpulkan terlebih dahulu. Kelebihan dari model ini adalah pemilihan frasa kunci untuk dijadikan indeks lebih terarah. Dengan menggunakan daftar indeks dari *corpus* yang telah diproses maka pemilihan frasa akan memberikan nilai *precision* dan *recall* lebih tinggi. Kelemahan dari model ini tidak bisa digunakan untuk buku atau dokumen diluar domain ilmu. Kelemahan berikutnya adalah kurang bisa mencari frasa kunci baru (selain yang terdapat di daftar indeks yang telah dibuat) untuk dijadikan kandidat indeks. Implementasi dari model ini antara lain dengan menggunakan tesaurus atau dengan menggunakan probabilitas kemungkinan frasa kunci dipilih menjadi frasa kunci indeks buku.

Model *unsupervised* berkebalikan dari model *supervised*. Model *unsupervised* menggunakan *corpus* sebagai pembanding apakah sebuah frasa kunci layak untuk dijadikan frasa kunci indeks buku. Kelebihan dari model ini adalah kemampuan untuk menemukan frasa kunci baru. Kelemahan dari model ini membutuhkan *corpus* yang baik serta kemungkinan nilai *precision* dan *recall* menjadi kurang memuaskan bisa terjadi.

Pengembangan kedua model berjalan beriringan. Dari kedua model tersebut saat ini masih dilakukan banyak penelitian.

DATA

Dataset pada penelitian ini dibagi menjadi lima domain ilmu. Pembatasan dataset sesuai dengan domain ilmu karena ada beberapa metode yang merupakan *supervised* sehingga diperlukan pembagian dataset ke dalam domain. Ada lima domain yang digunakan yaitu rekayasa perangkat lunak, agrikultur, psikologi, antropologi, dan sosiologi.

Buku diambil dari berbagai sumber dengan daftar buku berdasar dari referensi pengelompokkan yang sudah diakui. Untuk rekayasa perangkat lunak digunakan referensi *Software Engineering Body of Knowledge*. Untuk keempat domain lainnya digunakan daftar buku dari *Library of Congress*.

Masing-masing domain diambil sampel 20 buku. Buku-buku tersebut dibagi menjadi dataset untuk data latih dan dataset untuk data uji.

EKSTRAKSI FRASA KUNCI

Untuk mendapatkan informasi dari teks atau dokumen (*information retrieval*) pada umumnya menggunakan kerangka kerja yang sama, diawali dengan persiapan untuk melakukan ekstraksi, melakukan penyaringan, dan memberikan bobot pada frasa kunci yang diambil.Proses ekstrasi pada penelitian ini menggunakan metode n-gram. Metode n-gram yang digunakan adalah *word n-grams*. Metode ini membagi sebuah kalimat menjadi beberapa kata dengan pembatas spasi (*whitespace*). Metode ini dapat diterapkan untuk mengambil satu atau beberapa kata pada teks. *Unigram* untuk mengekstrak satu kata, *bigram* untuk mengekstrak dua kata, *trigram* untuk mengekstrak tiga kata, dan seterusnya disebut n-gram.

Contoh ekstraksi n-gram kata dapat digambarkan sebagai berikut. Sebuah kalimat "The construction integration model is not directly applicable". Bila dilakukan ekstraksi unigram maka hasilnya adalah "The", "construction", "integration", "model", "is", "not", "directly", "applicable". Hasil untuk ekstraksi bigram adalah "The construction", "constraction integration", "integration model", "model is", "is not", "not directly", "directly applicable". Hasil untuk trigram adalah "The construction integration", "construction integration model", "integration model is", "model is not", "is not directly", "not directly applicable". Dan hasil untuk 4-gram adalah "The construction integration model", "construction integration model is", "integration model is not", "model is not directly", "is not directly applicable".

Untuk meminimalkan hasil ekstraksi frasa maka diperlukan penyaringan frasa yang tidak mungkin terdapat pada indeks buku. Aturan yang digunakan adalah :

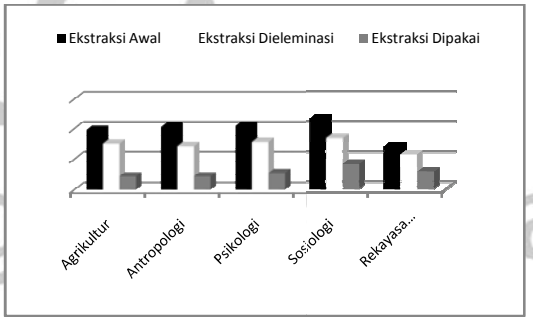
a. Frasa hasil ekstraksi tidak boleh terdapat angka.
b. Frasa hasil ekstraksi tidak boleh terdapat tanda baca seperti koma, titik, tanda kurung, tanda tanya, tanda seru, tanda dan, tanda atau, tanda bintang, tanda mata uang, tanda pagar, tanda tambah, tanda sama dengan.
c. Frasa hasil ekstraksi tidak boleh diawali dan diakhiri dengan *common word*.

Dari proses ekstraksi frasa kunci per domain didapat hasil ekstrasi seluruh n-gram frasa kunci antara 1.800.000an sampai 5.600.000an frasa kunci. Tabel 1 menggambarkan jumlah total hasil ekstraksi frasa kunci per domain.

Tabel 1. Jumlah Total Ekstraksi N-Gram Per Domain

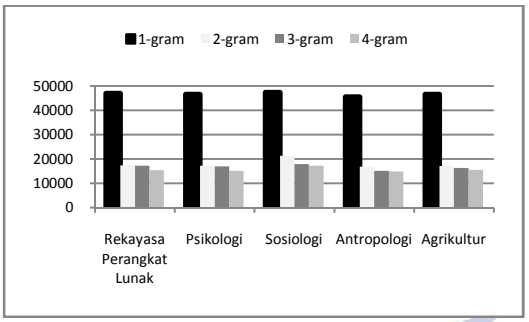
Domain	Total N-Gram
Rekayasa Perangkat Lunak	2.820.242
Psikologi	2.302.170
Sosiologi	5.646.962
Antropologi	1.344.766
Agrikultur	1.987.072

Tabel 1 adalah hasil ekstraksi frasa yang sudah dilakukan penyarigan. Proses penyaringan ini menghilangkan sekitar 70% frasa kunci. Gambar 1 menggambarkan hasil ekstraksi frasa yang belum dilakukan proses penyaringan, frasa kunci yang disaring, dan hasil penyaringan.



Gambar 1. Perbandingan Hasil Ekstraksi Frasa Kunci

Bila dilihat lebih detail, maka per buku akan dilakukan pembobotan dengan menggunakan kelima metode terhadap sekitar 100.000an frasa kunci. Hal ini tergambar pada Gambar 2.



Gambar 2. Rata-Rata Hasil Ekstraksi Per N-Gram Per Buku

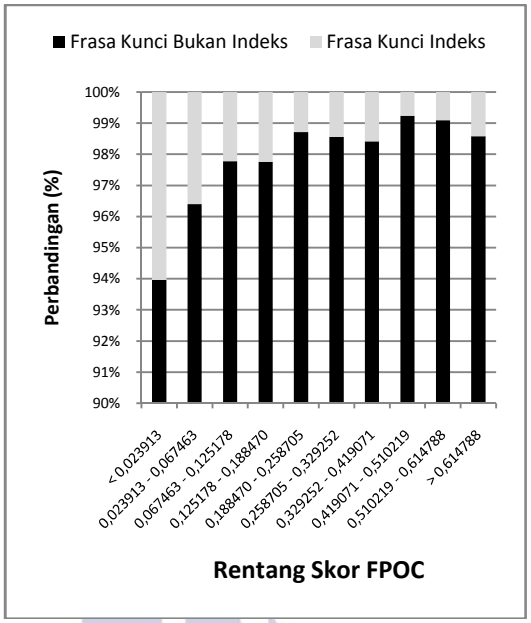
METODE *FIRST POSITION OF OCCURENCE*

Buku mempunyai beberapa bagian, dimana ada beberapa bagian dari buku merupakan bagian yang lebih penting daripada bagian yang lain. Posisi kemunculan pertama adalah metode dengan menghitung posisi pertama kali ditemukan frasa kunci dari awal dokumen atau akhir dokumen. Frasa kunci yang mempunyai nilai paling kecil atau paling besar secara statistik mempunyai kecenderungan untuk dijadikan kandidat utama di dalam indeks buku. Metode ini dibahas pada *Thesaurus-Based Index Term Extraction for Agricultural Documents* (Medelyan dan Witten, 2009).

Metode ini menghitung posisi kandidat frasa kunci pertama kali ditemukan di dalam buku. Posisi kandidat dilakukan normalisasi dengan keseluruhan isi buku. Nilai proporsional inilah yang akan digunakan sebagai penentu keutamaan kandidat sebagai frasa kunci indeks buku. Kandidat frasa kunci dengan nilai tinggi mempunyai keutamaan lebih dari kandidat frasa kunci yang lain.

Nilai yang dihasilkan pada metode ini mempunyai rentang antara 0 hingga 1. Nilai ini didapatkan dari normalisasi posisi kandidat frasa kunci terhadap dokumen. Dengan normalisasi ini, nilai yang paling kecil (mendekati nol) dan paling besar (mendekati satu) dapat dijadikan kandidat utama pada indeks buku.

$$Position = \frac{Sentencepositionindex}{\sum Sentencesindocument}$$



Gambar 3. Distribusi Frasa Kunci Pada Metode First Position of Occurence (FPOC)

Tujuan dari metode ini adalah menghitung jarak ditemukannya frasa kunci di dalam buku karena diyakini bahwa penulisan subjek frasa kunci yang merupakan frasa kunci penting berada di awal atau di akhir. Semakin kecil skor yang didapat maka semakin dekat lokasi frasa kunci dengan awal buku. Semakin besar skor yang didapat maka semakin dekat lokasi frasa kunci dengan akhir buku. Asumsi skor yang berpengaruh pada pemilihan frasa kunci yang terdapat di dalam indeks buku adalah skor yang paling kecil dan yang paling besar. Gambar 3 menunjukkan distribusi frasa kunci.

METODE TESAURUS

Metode ini dikembangkan oleh Medelyan dengan basis domain ilmu agrikultur. Dalam menggunakan metode ini dibutuhkan sebuah tesaurus yang telah ada sebelumnya. Tesaurus berisi daftar frasa kunci yang terdapat pada indeks buku. Pada penelitiannya, Medelyan menggunakan agrovoc yang berisi tesaurus tentang domain agrikultur.Pembuatan tesaurus membutuhkan koleksi daftar indeks buku yang lengkap.Metode ini dibahas pada *Thesaurus-Based Index Term Extraction for Agricultural Documents* (Medelyan dan Witten, 2009).

Tesaurus dapat dibentuk dari kumpulan indeks buku. Pembentukan tesaurus ini tentunya mengikuti kaidah penulisan dalam indeks. Penulisan indeks dibagi menjadi dua bagian yaitu *heading* dan *references*. *Heading* berisi

frasa kunci yang biasanya merupakan subjek yang terdapat pada isi buku. *References* adalah informasi tentang letak frasa tersebut di dalam buku. Dibagian *heading* sendiri terdapat dua kelompok, yaitu kelompok yang berisi frasa kunci dan kelompok yang merupakan rujukan tambahan dari frasa kunci yang dimaksud. Rujukan tambahan ini biasanya ditulis dengan "*see also ...*" atau "*see...*".

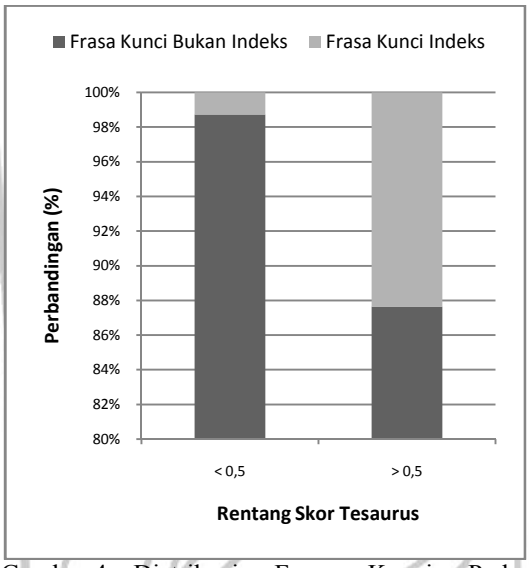
Penulisan frasa kunci juga mempunyai tingkatan. Secara umum ada tiga tingkatan yang biasa digunakan pada penulisan frasa kunci. Tingkatan teratas dapat disebut sebagai induk frasa. Induk frasa ini ditulis paling kiri. Tingkatan dibawahnya adalah anak frasa. Anak frasa ini ditulis lebih kekanan daripada induk frasa. Dan tingkatan berikutnya disebut cucu frasa. Cucu frasa ini ditulis lebih kekanan daripada anak frasa. Penulisan berjenjang dimaksudkan untuk memudahkan pembaca dalam mencari frasa kunci yang diinginkan. Anak frasa merupakan penjelasan tambahan terkait dengan induk frasa demikian juga cucu frasa merupakan penjelasan tambahan dari anak frasa.

Dalam merangkaikan frasa indeks, dilakukan penghapusan frasa rujukan (frasa setelah *see...*). Hal ini dilakukan karena frasa rujukan telah ada di dalam daftar indeks. Frasa rujukan tidak mungkin merujuk ke frasa yang tidak terdapat di dalam indeks.

Setelah rangkaian frasa disatukan menjadi frasa utuh dengan kaidah yang telah ditentukan, kemudian dilakukan pengecekan. Pengecekan ini berfungsi untuk menyakinkan bahwa frasa yang dibentuk terdapat di dalam isi buku. Ekstraksi frasa kunci untuk indeks buku tidak mungkin menemukan frasa kunci yang tidak terdapat di dalam buku. Dari pengujian yang dilakukan hanya 40% dari total rangkaian frasa kunci indeks yang ditemukan padanannya di dalam buku.

Tidak semua frasa kunci indeks buku ditulis seperti yang terdapat di dalam isi buku. Hal ini terkait dengan sisi manusia dari pengindeks. Setiap pengindeks mempunyai kebiasaan, wawasan dan pengetahuan yang berbeda. Inilah yang menjadi salah satu kesulitan untuk melakukan pengindeksan otomatis dengan hasil yang menyamai hasil pengindeksan yang dilakukan oleh profesional.

Metode ini cukup berperan dalam menentukan frasa kunci mana yang ada di dalam indeks. Daftar tesaurus diambil dari data latih. Keterbatasan data latih dapat menyebabkan kondisi pembagian hampir sama rata. Dalam distribusi frasa kuncinya hampir sama terbagi menjadi dua kelompok. Pada Gambar 4 dapat dilihat distribusi pengelompokkannya. Pembagian menjadi dua kelompok ini dapat memberikan dampak yang cukup besar.



Gambar 4. Distribusi Frasa Kunci Pada Metode Tesaurus.

METODE *INFORMATIVENESS*

Metode ini dikembangkan oleh Csomai (Csomai dan Mihalcea, 2008). Basis penghitungan pada metode ini menggunakan *chi-square*. Pengembangan dilakukan pada isi setiap sel dari tabel contigency yang digunakan. Tabel 1 menggambarkan rumus yang digunakan untuk mengisi sel pada tabel contigency.

Tabel 1. Tabel Contigency Informativeness

Jumlah (Frasa di dokumen)	Jumlah (Frasa lain di dokumen)
Jumlah (Frasa di *corpus*)	Jumlah (Frasa lain di *corpus*)

Sumber: Diedit dari LMFEBI (Csomai dan Mihalcea, 2008).

Untuk menghitung skor akhir menggunakan rumus

$$x^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O (*Observed*) adalah jumlah sel pada tabel dan E (*Expected*) adalah hasil perhitungan probabilitas jarak yang dikonversikan ke dalam proporsi dengan membaginya terhadap jumlah total kejadian O.

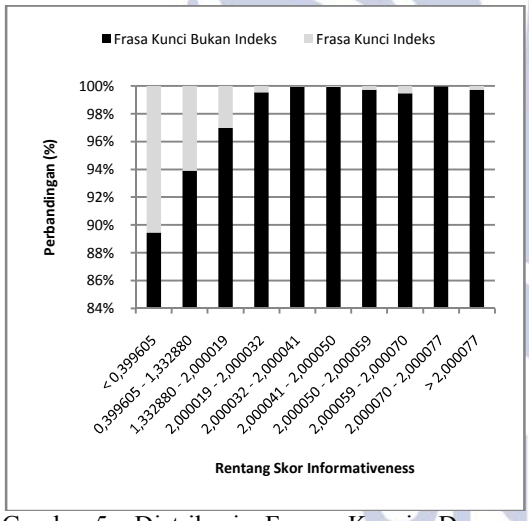
Untuk mengisi nilai pada expected menggunakan rumus

$$E_{11} = \frac{O_{11} + O_{12}}{N} \times \frac{O_{11} + O_{21}}{N} \times N$$

Dan nilai N didapat dari rumus

$$N = O_{11} + O_{12} + O_{21} + O_{22}$$

Pada metode ini, kandidat frasa kunciyang terdapat pada indeks berhasil dikelompokkan menjadi beberapa rentang skor. Di bagian kiri diagram batang dengan rentang skor kurang dari 0,3996 memiliki lebih banyak kandidat frasa kunci yang ada di daftar indeks (11%). Berikutnya adalah rentang antara 0,3996 – 1,3328 dan 1,3328 – 2,0. Dari Gambar 5 dapat dilakukan *tuning* parameter pemilihan *informativeness* dilakukan pada rentang kurang dari 2,000050. Frasa kunci yang dapat dikelompokkan mencapai hampir 70% berada di rentang 0 hingga 2,000050.



Gambar 5. Distribusi Frasa Kunci Dengan Metode *Informativeness*

METODE *PHRASENESS*

Metode ini menggunakan *chi-squareindependence test*. Metode ini digunakan untuk mengidentifikasi apakah frasa yang ada terbentuk bukan dari sebuah kebetulan. Dengan skor yang melewati *threshold* yang telah ditetapkan, maka diyakini bahwa frasa tersebut muncul bukan dari sebuah kebetulan.

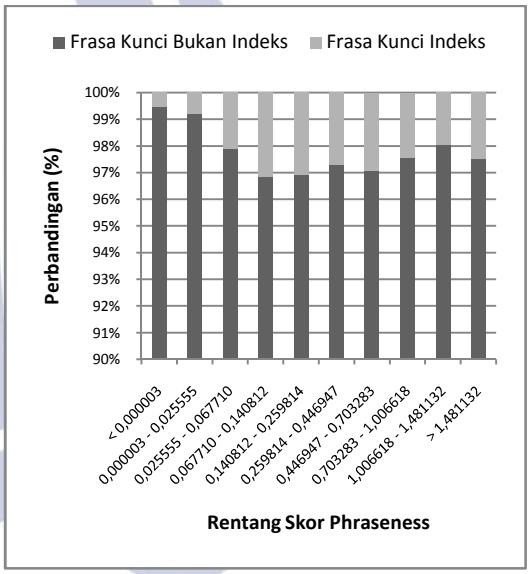
Metode ini merupakan salah satu dari model dengan peringkat terbaik pada *coolocation discovery*(Pecina dan Schlesinger, 2006). Pada metode ini tetap digunakan tabel contigency seperti pada metode *informativeness* tetapi menggunakan rumus yang berbeda untuk isian di tiap sel. Rumus yang digunakan terdapat pada Tabel 2.

Tabel 2. Tabel Contigency Phraseness

$f(x,y)$	$f(x,\bar{y})$	$f(x,*)$
$f(\bar{x},y)$	$f(\bar{x},\bar{y})$	$f(\bar{x},*)$
$f(*,y)$	$f(*,\bar{y})$	N

Sumber: *Combining Association Measure for Collocation Extraction* (Pecina dan Schlesinger, 2006)

Karena metode ini digunakan untuk menguji sebuah frasa, maka frasa harus terbentuk dari 2 kata. Apabila frasa yang diuji lebih dari 2 kata maka harus dilakukan konversi ke dalam bentuk frasa dengan 2 kata.



Gambar 6. Distribusi Frasa Kunci Pada Metode *Phraseness*

Penyebaran yang hampir merata disemua rentang skor membuat penentuan *threshold* skor frasa kunci yang dipilih menjadi sulit. Hal ini dimungkinkan terjadi karena frasa kunci pada buku tidak muncul sesering yang diperkirakan. Sebuah buku membahas subjek yang cukup luas sehingga jumlah kemunculan frasa kunci yang dipilih tidak signifikan. Gambar 6 menggambarkan distribusi frasa kunci pada metode ini.

METODE *SYNTACTIC*

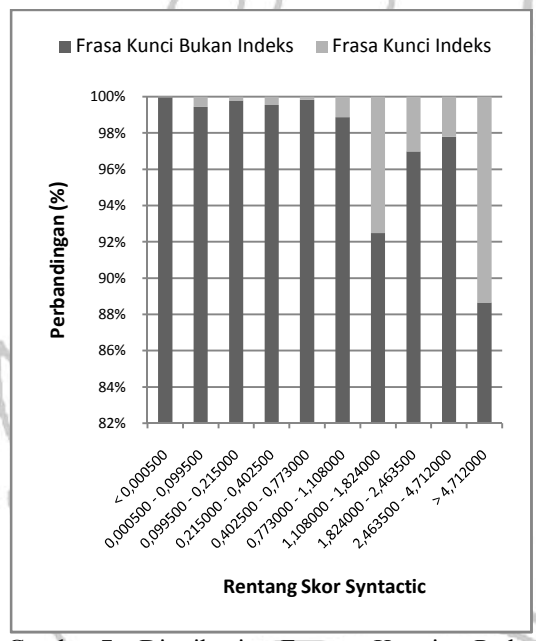
Metode ini dikembangkan oleh Csomai (Csomai dan Mihalcea, 2008). Metode *syntactic* ini pada dasarnya berupa pengembangan dari *part-of-speech (POS) tagging*. Pada metode ini

tidak digunakan pola-pola yang terbentuk secara langsung, tetapi hasil perhitungan probabilitas yang terjadi di data latih yang digunakan sebagai pembobot pada data *testing*.

Rumus yang digunakan adalah

$$P_{(pattern)} = \frac{C_{(patternpositive\ )}}{C_{(pattern)}}$$

Dimana $C_{(pattern,positive)}$ adalah jumlah frasa unik yang mempunyai pola POS tertentu dan terdapat pada indeks buku dan $C_{(pattern)}$ adalah jumlah frasa unik dengan pola POS. Pola POS menggunakan Penn Treebank.



Gambar 7. Distribusi Frasa Kunci Pada Metode Syntactic

Dari Gambar 7 menunjukkan hampir 80% frasa kunci yang ada di indeks buku berhasil dikelompokkan pada rentang skor 1,108000 keatas. Dengan keberhasilan ini, penentuan *threshold* untuk mempersempit rentang skor dapat dilakukan. Pada penelitian ini, metode syntactic memiliki peran dominan. Kemampuan metode ini mengelompokkan frasa kunci pada rentang tertentu telah terbukti pada penelitian ini.

HASIL ANALISA

Untuk membandingkan pengaruh dari masing-masing metode terhadap pengindeksan otomatis digunakan infogain. Dari hasil analisis infogain didapat hasil yang terangkum di Tabel 3.

Tabel 3. Hasil Analisa Infogain

Skor	Keterangan
0.03092	Informativeness
0.03068	Syntactic
0.01666	Tesaurus
0.00648	First Position of Occurence
0.00343	Phraseness

Dari Tabel3 dapat dilihat bahwa metode *informativeness* mempunyai pengaruh yang yang cukup tinggi dibanding dengan metode yang lainnya. Dengan skor 0,03092 menggunakan *attribute ranking* dari *information gain* menempatkan metode ini di posisi pertama diantara kelima metode yang lain. Skor yang didapat hampir sama dengan skor yang diperoleh metode *syntactic*.

Metode *Phraseness* kurang berhasil mengelompokkan kandidat frasa kunci yang ada di dalam indeks. Skor yang didapat pada Tabel 3 adalah 0,00343. Metode ini mendapatkan peringkat terakhir pada tabel tersebut. Bila dilihat pada Gambar 6 maka akan terlihat jelas distribusi kandidat frasa kunci yang terdapat di indeks buku hampir merata di semua kelompok.

Distribusi frasa kunci yang ada di dalam indeks buku pada rentang skor 0,025555 ke atas mempunyai prosentase rata-rata 12% disetiap rentang skornya. Faktor inilah yang menyebabkan tingkat pengaruh yang rendah untuk metode ini.

Metode*syntactic* mendapatkan skor yang hampir sama dengan metode *informativeness*. Skor 0,03068 dari Tabel 3 diperingkat kedua setelah *informativeness*. Metode ini berhasil mengelompokkan kandidat frasa kunci ke bagian kanan dari kelompok yang ada. Pada Gambar 7 dapat dilihat kelompok sebelah kanan lebih banyak memiliki frasa kunci yang ada di dalam indeks. Rentang skor lebih dari 4,712 dan skor antara 1,108 – 1,824 memiliki lebih banyak frasa kunci yang ada di dalam indeks dari pada rentang kelompok lainnya.

Dengan penggunaan domain tesaurus diyakini dapat meningkatkan tingkat akurasi pembobotan. Domain tesaurus pada penelitian ini belum dikembangkan dengan maksimal. Penambahan frasa kunci yang mempunyai tingkat similarity tertentu diyakini dapat meningkat kemampuan dari domain tesaurus tersebut.

Pada Tabel 3 skor 0,00648 didapat untuk metode FPOC. Dari Gambar 3 dapat dilihat bahwa sebagian besar frasa kunci yang ada di dalam indeks berada di sebelah kiri. Pada setiap kelompok rentang skor masih terdapat frasa kunci yang ada di dalam daftar indeks. Hal ini

yang menyebabkan skor peringkat metode ini tidak terlalu bagus.

Pada penelitian ini, kondisi yang terjadi identifikasi frasa kunci berada merata disemua rentang skor. Kondisi ini membuat hasil akhir yang tidak optimal. Penggunaan metode ini sebaiknya dilakukan dengan dilakukan pemisahan setiap bab yang terdapat di buku. Dengan pemisahan buku menjadi bagian-bagian yang lebih kecil maka efektifitas penggunaan metode ini dapat ditingkatkan. Didalam penulisan sebuah buku, disetiap bab memiliki pokok pikiran yang berbeda-beda walaupun masih dalam satu kajian. Metode ini sangat cocok untuk mengidentifikasi pokok pikiran tersebut. Penulisan pokok pikiran dalam sebuah bab biasanya lebih banyak berada di awal atau akhir bab. Indeks buku sebagian besar berisi pokok-pokok pikiran tersebut.

Pada metode tesaurus memberikan dampak pengaruh sebesar 0,01666 seperti yang terdapat pada Tabel 3. Metode ini cukup berperan dalam menentukan frasa kunci mana yang ada di dalam indeks. Daftar tesaurus diambil dari data latih. Keterbatasan data latih dapat menyebabkan kondisi pembagian hampir sama rata. Dalam distribusi frasa kuncinya hampir sama terbagi menjadi dua kelompok. Pada Gambar 4 dapat dilihat distribusi pengelompokkannya. Pembagian menjadi dua kelompok ini dapat memberikan dampak yang cukup besar.

Faktor gaya penulisan dari masing-masing pengindeks menjadi faktor dominan yang menyebabkan kecilnya hasil pencarian. Gaya penulisan ini juga menjadi kendala diantara pengindeks dalam melakukan pembuatan indeks buku (Mulvany, 2005). Faktor kesepahaman dan wawasan menjadi faktor subjektif yang tidak bisa diperdebatkan diantara pengindeks. Adakalanya pengindeks memberikan frasa kunci yang berupa padanan dari frasa kunci yang ada di isi buku. Hal ini dilakukan untuk menghasilkan indeks buku yang lebih terstruktur dan lebih mudah digunakan.

KESIMPULAN DAN SARAN

Metode *informativeness* mempunyai dampak yang cukup signifikan pada penelitian ini. Berikutnya metode *syntactic*, dan tesaurus berada di peringkat selanjutnya. Penggunaan ketiga metode tersebut yang memberikan sumbangsih terbesar pada proses pengindeksan otomatis.

Metode *phraseness* dan FPOC ternyata tidak memberikan dampak yang signifikan terhadap proses pengindeksan otomatis. Detail implementasi diyakini merupakan faktor kunci dalam peningkatan pengaruh masing-masing metode.

Pengembangan isi dari tesaurus dengan menyertakan frasa kunci – frasa kunci yang memiliki tingkat *similarity* tinggi dapat dilakukan. Dengan pengembangan ini diharapkan ruang lingkup tesaurus dapat lebih luas dan sekaligus akan mampu menghasilkan performa yang lebih tinggi. Pada penelitian ini tidak dilakukan pengembangan frasa kunci tesaurus sehingga hasil yang didapat tidak mengalami peningkatan yang signifikan.

DAFTAR PUSTAKA

Anderson, J. D. dan NISO, (1997), *Guidelines for Indexes and Related Information Retrieval Devices*, Niso Press, Maryland.

Csomai, A. dan Mihalcea, R. (2008), "Lingustically Motivated Features for Enhanced Back-of-the-Book Indexing.", *Computer Linguistic*, Vol.16,No.2.

Medelyan, O. dan Witten, I. H. (2006), "Thesaurus Based Automatic Keyphrase Indexing.", *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, hal.296-297.

Mulvany, N. C., (2005), *Indexing Books*, 2nd ed. The University of Chicago Press, Chicago.

Pecina, P. dan Schlesinger, P. (2006), "Combining Association Measures for Collocation Extraction.", *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, hal.651-658.