Perancangan dan Pembangunan Perangkat Lunak *Indexing* Buku Berbahasa Indonesia

Hari Prasetyo, Daniel O. Siahaan, dan Ahmad Saikhu
Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: daniel@if.its.ac.id

Abstrak—**Pengindeksan buku merupakan sebuah proses yang membutuhkan waktu dan keahlian. Ketika kebutuhan indeks buku yang diperlukan cukup banyak, sumber daya yang dibutuhkan untuk dapat menghasilkan indeks tersebut juga akan meningkat. Sumber daya yang diperlukan meliputi waktu, tenaga, dan pikiran. Jika proses pengindeksan dapat dikerjakan secara otomatis oleh komputer, maka keperluan sumber daya tadi dapat dialihkan untuk keperluan lain.**

Artikel ini menjelaskan tentang pembuatan perangkat lunak untuk mengindeks sebuah buku. Buku yang dapat diindeks oleh perangkat lunak adalah buku yang berbahasa Indonesia. Dalam artikel ini terdapat implementasi dari algoritma ekstraksi kata kunci pada dokumen tunggal yang digunakan untuk mencari kata kandidat indeks. Setelah mendapatkan kata kunci yang dianggap sebagai kandidat indeks, akan dilakukan perhitungan keterkaitan semantik dengan menggunakan WordNet. Hasil dari perhitungan keterkaitan semantik dapat dikatakan sebagai indeks dari sebuah buku.

Ekstraksi kata kandidat indeks menggunakan informasi statistik dari kemunculan seluruh kata-kata yang ada. Hasil dari uji coba fungsional menunjukan bahwa perangkat lunak ini berjalan baik dari berbagai skenario yang telah dibuat. Hasil dari uji coba presisi dan sensitivitas masih dipengaruhi oleh banyak faktor. Hasil akhir yang diperoleh dari perangkat lunak ini adalah sebuah dokumen teks berisi daftar indeks beserta halaman tempat kata indeks tersebut dapat ditemukan.

Kata Kunci—**Ekstraksi Kata Kunci, Indeks, Token, WordNet.**

I. PENDAHULUAN

INDEKS adalah daftar kata atau istilah penting yang terdapat dalam buku teks. Indeks tersusun menurut abjad dan memberikan informasi mengenai halaman tempat kata atau istilah itu ditemukan dalam buku. Indeks sangat berguna untuk mempermudah pencarian keterangan di dalam buku.

Pada masa sekarang untuk dapat membuat indeks dapat digunakan *tools* Index dari Microsoft Word, tetapi kata–kata yang akan dimasukkan ke dalam indeks masih dilakukan dengan cara memasukkan kata satu persatu. Perangkat lunak lainnya dapat ditemui di internet antara lain Adobe FrameMaker, dan SKY Index. Perangkat lunak tersebut memiliki cara kerja yang hampir sama dengan Microsoft Word, yaitu dapat mengubah dan mengorganisasi indeks yang dibuat.

Mengindeks buku merupakan sebuah proses yang cukup lama. Ketika kebutuhan indeks buku yang diperlukan cukup banyak, sumber daya yang dibutuhkan untuk dapat menghasilkan indeks tersebut juga akan meningkat. Sumber daya yang diperlukan meliputi waktu, tenaga, dan pikiran. Jika proses pengindeksan dapat dikerjakan oleh komputer, maka keperluan sumber daya tersebut dapat dialihkan untuk keperluan lain. Pengindeksan dengan komputer tidak sesederhana jika dilakukan oleh manusia, oleh karena itu pada artikel ini dibuat sebuah perangkat lunak yang dapat mengatasi masalah-masalah di atas.

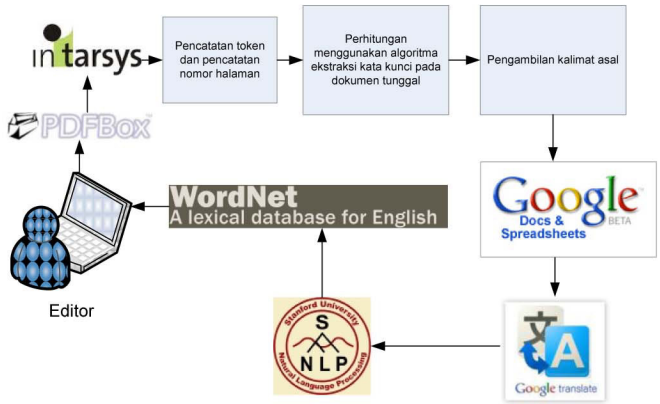
II. TINJAUAN PUSTAKA

A. Indeks

Indeks adalah daftar kata atau istilah penting (*headings*) dan pointer yang terkait (*locators*) yang terdapat dalam buku cetak. Indeks tersusun menurut abjad dan memberikan informasi mengenai halaman tempat kata atau istilah itu ditemukan dalam buku. Indeks istilah membuat sebuah kontrol kosa kata untuk digunakan dalam catatan bibliografi. Indeks istilah yang baik tidak memuat istilah yang tidak terdapat pada buku. Indeks didesain untuk pencarian informasi secara cepat dan mudah. Sebuah indeks yang lengkap dan benar-benar berguna bukan hanya sebuah urutan kata-kata dan frase yang digunakan dalam publikasi, tetapi pemetaan konten yang terorganisir, termasuk referensi silang, pengelompokan konsep yang serupa, dan analisis intelektual yang berguna lainnya [1].

B. Algoritma Ekstraksi Kata Kunci pada Dokumen Tunggal

Algoritma ekstraksi kata kunci pada dokumen tunggal dapat mengekstrak kata kunci dari sebuah dokumen tanpa menggunakan kumpulan dokumen. Dalam kasus artikel, ekstraksi kata kunci dalam satu dokumen sangat berguna. Pertama-tama, kata istilah yang sering muncul diekstraksi terlebih dahulu. Kemunculan dari sebuah kata istilah dengan kata istilah yang sering muncul juga dihitung. Jika sebuah kata istilah sering muncul bersama bagian tertentu dari kata istilah, maka kata istilah yang sering muncul tadi akan terlihat memiliki makna penting. Derajat simpangan dari distribusi kemunculan dihitung dengan pengukuran χ^2. Simpangan ini diturunkan baik dari semantik, leksikal, atau relasi lainnya diantara dua kata istilah. Oleh karena itu, derajat simpangan dari kemunculan kata istilah bisa digunakan sebagai indikator dari kepentingan suatu kata



Gambar. 1. Arsitektur perangkat lunak

istilah [2].

C. *WordNet*

WordNet adalah sebuah basis data leksikal bahasa Inggris. Kata benda, kata kerja, dan kata sifat diorganisasikan ke dalam *synonym sets* (*synset*). Tiap *synset* mewakili konsep leksikal dasar. WordNet diorganisasikan berdasarkan hubungan semantik. Hubungan semantik adalah hubungan antara makna yang diwakili dengan *synset* maka dapat dikatakan bahwa antar *synset* terjalin hubungan semantik. Hubungan yang terbentuk antara lain sinonim, antonim, hiponim, dan meronim. WordNet membagi kata menjadi kata benda, kata kerja, kata sifat, dan kata keterangan. Jika konsep direpresentasikan sebagai *synset* dan jika sinonim harus dapat dipertukarkan, maka kata dalam kategori yang berbeda dianggap tidak bersinonim. Kata benda menyatakan konsep nominal, kata kerja menyatakan konsep verbal, dan kata sifat adalah cara untuk membatasi konsep [3].

III. METODOLOGI

A. *Analisis*

1) *Deskripsi Umum Sistem*

Artikel ini akan membuat suatu perangkat lunak yang dapat menghasilkan indeks dari sebuah buku elektronik dengan format PDF. Perangkat lunak pengindeks buku berbasis *desktop* dengan platform Java. Perangkat lunak pengindeks buku terdiri dari beberapa langkah dalam menjalankan fungsi utamanya.

Setelah pengambilan konten dari buku elektronik berhasil, maka langkah selanjutnya adalah pemrosesan yang melibatkan sistem temu kembali informasi. Konten yang berhasil diambil dari buku elektronik adalah berupa data teks. Setelah mendapatkan data teks, akan dilakukan prapemrosesan data teks tadi. Prapemrosesan data teks meliputi tokenisasi dan penghilangan kata umum.

Langkah berikutnya adalah mencatat semua token hasil dari prapemrosesan data teks. Pencatatan token meliputi isi token dan halaman di mana token tersebut ditemukan. Pencatatan ini berguna pada saat token tersebut terpilih menjadi kata indeks maka perangkat lunak tidak perlu mencari lagi pada halaman berapa token tersebut berada.

Tabel 1.
Uji proses perangkat lunak pengindeks buku

Skenario	Hasil Uji	Status
Direktori dokumen tidak lengkap	Perangkat lunak menampilkan peringatan.	Berhasil
Tidak ada koneksi internet	Perangkat lunak menampilkan peringatan,	Berhasil
Ada koneksi internet	Sebuah berkas teks yang berisi indeks dari buku	Berhasil

Pencatatan halaman juga berguna untuk penghitungan jumlah suatu token pada dokumen tersebut. Jumlah tersebut akan digunakan dalam algoritma ekstraksi kata kunci pada dokumen tunggal.

Kandidat indeks yang telah didapatkan tidak langsung menjadi indeks. Suatu token yang muncul pada setiap kalimat, belum tentu memiliki keterkaitan semantik dengan kalimat tersebut. Token yang sudah menjadi kandidat indeks akan diperiksa kembali apakah kemunculannya pada kalimat tersebut memiliki keterkaitan semantik. Berdasarkan penelitian sebelumnya, jika suatu token memiliki keterkaitan semantik di atas ambang batas maka token tersebut layak untuk dijadikan indeks.

2) *Arsitektur Sistem*

Pada Gambar 1 diilustrasikan bagaimana arsitektur perangkat lunak dari artikel ini. Hanya terdapat satu aktor yaitu *editor*. *Editor* terlebih dahulu menginstal perangkat lunak pengindeks buku pada komputer mereka. Selanjutnya perangkat lunak pengindeks buku akan menampilkan dialog untuk memilih dokumen PDF yang akan dibuatkan indeks.

Apabila *editor* sudah memilih dokumen PDF yang akan diindeks maka perangkat lunak pengindeks buku akan membaca dokumen PDF tersebut. Setelah isi konten terambil, akan dilakukan pencatatan token dan pencatatan nomer halaman. Hasil dari pencatatan pada proses sebelumnya digunakan untuk perhitungan menggunakan algoritma ekstraksi kata kunci pada dokumen tunggal.

Hasil penghitungan pada proses sebelumnya akan menghasilkan kandidat indeks. Kalimat asal dari kandidat indeks yang sudah ada akan diambil untuk kemudian dicari keterkaitan semantik antara token dengan kalimat asal tersebut. Cara penghitungan keterkaitan semantik kalimat yang digunakan dalam perangkat lunak pengindeks buku adalah dengan menghitung token kandidat indeks dengan setiap kata dalam kalimat asal tersebut. Hasil dari semua penghitungan tersebut akan dijumlah lalu dirata-rata. Jika hasil dari rata-rata nilai keterkaitan semantik melewatai ambang batas atas maka token kalimat tersebut akan dimasukkan ke dalam indeks.

B. *Perancangan*

Perangkat lunak pengindeks buku terdiri dari beberapa proses. Proses-proses yang ada berjalan secara berurutan sehingga tiap proses bergantung pada proses sebelumnya.

1) *Proses Mengekstrak Teks*

Proses mengekstrak teks merupakan proses awal dalam

perangkat lunak pengindeks buku. Proses ini akan menghasilkan keluaran berupa kumpulan teks dalam suatu struktur data.

2) Proses Prapemrosesan Teks

Proses prapemrosesan teks terdiri dari dua bagian yaitu tokenisasi dan penghapusan kata umum. Proses ini menjadikan teks halaman-halam dokumen tadi menjadi token-token dan sudah tidak ada kata umum yang terkandung di dalam daftar token.

Tabel 2.
Uji coba presisi dan sensitivitas perangkat lunak pengindeks buku

Buku ke-i	Presisi	Sensitivitas
1	35,75%	8,16%
2	38,94%	12,62%
3	9,23%	12,45%
4	23,84%	14,89%
5	15,51%	18,30%

3) Proses Pencatatan Nomer Halaman

Proses pencatatan nomor halaman berjalan setelah proses penghapusan kata umum. Token yang dihasilkan dari proses tokenisasi dan penghapusan kata umum sudah dibagi menjadi perhalaman.

4) Proses Ekstraksi Kata Kunci

Proses ekstraksi kata kunci adalah proses pencarian kata-kata yang dianggap penting. Pertama-tama akan diambil sebanyak 30% dari jumlah token yang memiliki frekuensi terbanyak. Setelah penghitungan selesai, akan diambil sebanyak 70% kata yang memiliki nilai terbesar berdasarkan penghitungan tersebut. Kata-kata yang dianggap penting tersebut nantinya akan menjadi kandidat indeks.

5) Proses Pengambilan Kalimat Asal

Proses pengambilan kalimat asal merupakan proses lanjutan setelah perangkat lunak pengindeks buku berhasil mendapatkan kandidat indeks. Pada proses ini, perangkat lunak pengindeks buku akan membuat daftar yang berisi pasangan antara kandidat indeks dengan kalimat asal tempat kandidat indeks tersebut ditemukan.

6) Proses Penerjemahan

Proses penerjemahan merupakan salah satu syarat untuk proses menghitung keterkaitan semantik kandidat indeks dengan kalimat asal. Proses penerjemahan ini melibatkan Google Spreadsheets dan Google Translate.

7) Proses Penghitungan Keterkaitan Semantik

Proses penghitungan keterkaitan semantik menghitung keterkaitan semantik kandidat indeks dengan kalimat asal tempat kandidat indeks ditemukan. Jika kata indeks dan kalimat asal memiliki nilai rata-rata keterkaitan semantik lebih dari 0.56 maka, halaman tempat kata indeks tersebut ditemukan akan dimasukkan kedalam indeks.

IV. UJI COBA DAN EVALUASI

A. Uji Coba Fungsionalitas

Tabel 1 merupakan hasil uji coba perangkat lunak pengindeks buku pada proses utama yaitu mengindeks buku. Setiap poin-poin telah dilakukan pengujian berdasarkan skenario tertentu. Semua uji coba yang dilakukan menghasilkan status berhasil.

B. Uji Coba Non Fungsionalitas

Tabel 2 merupakan hasil uji coba perangkat lunak pengindeks buku untuk mencari nilai presisi dan sensitivitas. Uji coba tersebut menggunakan 5 buku yang berbeda. Berdasarkan uji coba tersebut diperoleh rata-rata nilai presisi sebesar 24,66% dan rata-rata nilai sensitivitas sebesar 13,29%. Rincian uji coba non fungsionalitas dapat dilihat pada Tabel 3.

Tabel 3.
Rincian uji coba non fungsionalitas

No.	Nama Buku	Jumlah indeks yang terdapat pada buku	Jumlah indeks yang dihasilkan oleh perangkat lunak	Presisi	Sensitivitas
1	Algorithm - Design Techniques and Analysis	2107	481	35,75%	8,16%
2	Graph Theory	1465	475	38,94%	12,62%
3	Machine Learning, Neural and Statistical Classification	562	758	9,23%	12,45%
4	Mathematics For Game Developers	1349	843	23,84%	14,89%
5	Simulation Modeling and Analysis with ARENA	612	722	15,51%	18,30%

V. KESIMPULAN/RINGKASAN

Perangkat lunak pengindeks buku dapat menghasilkan keluaran berupa indeks dari sebuah buku. Buku yang digunakan menggunakan format PDF. Berdasarkan uji coba non fungsionalitas, didapatkan rata-rata nilai presisi sebesar 24,66% dan rata-rata nilai sensitivitas sebesar 13,29%. Nilai tersebut dipengaruhi oleh dua faktor. Faktor pertama adalah keterbatasan pustaka yang digunakan. Faktor kedua adalah tidak digunakannya pengambilan kata dasar (*stemming*).

UCAPAN TERIMA KASIH

Penulis H.P. mengucapkan terima kasih kepada orang tua dan keluarga penulis, dosen pembimbing, dosen dan kepala jurusan Teknik Informatika, kerabat-kerabat dekat, serta berbagai pihak yang telah membantuk penulis dalam menyelesaikan artikel ini.

DAFTAR PUSTAKA

[1] Knight, G. Norman. 1979. Indexing, the Art of: A Guide to the Indexing of Books and Periodicals, HarperCollins.
[2] Matsuo, Y. Ishizuka, M., 2003. Keyword Extraction from a Single Document using Word Co-ocurrence Statistical Information. International Journal on Artificial Intelligence Tools, Vol. 13.
[3] Miller, George A. 1995. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38.