IMPLEMENTASI DAN ANALISIS TEXTRANK PADA EKSTRAKSI KATA KUNCI

Hariyati Lubis[1], Warih Maharani[2], Angelina Prima Kurniati[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Banyaknya ketersediaan informasi digital dalam bentuk dokumen teks saat ini tidak berarti pencarian dan identifikasi dokumen yang dibutuhkan mudah dilakukan. Salah satu cara untuk mengidentifikasi dokumen adalah melalui kata kunci. Namun tidak semua dokumen teks dilengkapi kata kunci.

Pengekstraksi kata kunci merupakan sistem yang dapat mengekstrak kata kunci secara otomatis dari teks. Pengekstraksi kata kunci yang dibuat dalam Tugas Akhir ini mengimplementasikan algoritma TextRank yang berbasis graf. TextRank untuk ekstraksi kata kunci mempunyai 3 parameter inputan, yaitu jenis graf, filter kata, dan ukuran window.

Pengujian dilakukan dengan cara melakukan perubahan pengaturan parameter TextRank dan membandingkan kata kunci hasil keluaran sistem dengan kata kunci manual yang telah tersedia di setiap abstrak. Pengujian juga dilakukan terhadap penggunaan beberapa variasi nilai damping factor dan threshold. Parameter uji yang digunakan adalah recall. Dari hasil pengujian, graf undirected dan filter kata benda menghasilkan nilai uji yang lebih baik di hampir semua ukuran window daripada graf dan filter kata lain. Secara umum, nilai recall semakin baik jika nilai damping factor semakin besar meskipun tidak terlalu signifikan dan nilai recall tidak terlalu dipengaruhi nilai threshold.

Kata Kunci : kata kunci, ekstraksi kata kunci, textrank

Abstract

The abundant amount of digital information in the form of text does not mean the searching and the identification of the needed document can be done easily. One way to identify a document is by keywords. However, not all documents are provided with keywords.

Keyword extraction is a system to automatically extract a set of keywords from text. Keyword extraction built for this final work implements TextRank, a graph-based algorithm. TextRank for keyword extraction has 3 input parameters, which are graph type, word filter, and window size. Testing was done by changing the setting of TextRank parameters and comparing system's keywords with human's keyword. The experiment was also done using various damping factor and threshold values. Testing parameter used in this final work are recall. From testing, the using of undirected graph and verb filter results in better testing parameters' values in almost sizes of window than other using of graphs and word filters. In general, recall increases when damping value is bigger although it's not significant and recall is not really influenced by threshold value.

Keywords : keyword, keyword extraction, textrank

1. Pendahuluan

1.1 Latar belakang

Kemudahan digitalisasi teks saat ini semakin meningkatkan jumlah informasi yang tersedia. Namun banyaknya ketersediaan informasi tersebut tidak berarti identifikasi dan pemilihan informasi yang dibutuhkan mudah dilakukan. Identifikasi mengenai informasi apa yang terkandung dari suatu teks dapat diperoleh dengan membaca kata kunci. Namun, banyak dokumen teks yang ada tidak disertai dengan kata kunci.

Pada referensi [3] dinyatakan bahwa kata kunci dapat dianggap sebagai ringkasan yang lebih padat dari suatu teks. Kata kunci tidak hanya berguna untuk mengetahui deskripsi suatu teks, tetapi juga dapat berperan penting untuk proses *indexing* otomatis, peringkasan teks dan meningkatkan performansi kategorisasi teks [3][6].

Mengekstrak kata kunci dari suatu teks tentu saja dapat dilakukan secara manual. Namun, jika teks tersebut merupakan teks yang cukup panjang dan tersedia dalam jumlah yang besar, proses ekstraksi kata kunci secara manual akan memakan waktu yang lama. Oleh karena itu dibutuhkan suatu pengekstrak kata kunci yang dapat mengurangi *resource* yang dibutuhkan jika ekstraksi kata kunci dilakukan secara manual.

Banyak algoritma/metode yang dapat diimplementasikan untuk mengekstrak kata kunci, misalnya algoritma genetika, *naïve bayes*, dan *rule induction*. Dalam Tugas Akhir ini, algoritma yang diimplementasikan adalah *TextRank*. *TextRank* dipilih karena algoritma ini tidak membutuhkan data latih dan bersifat *language independent*, yaitu tidak bergantung pada bahasa tertentu dan dapat digunakan di semua bahasa. *TextRank* termasuk algoritma perangkingan berbasis graf yang menentukan nilai pentingnya sebuah *vertex* berdasarkan informasi global yang didapatkan dari keseluruhan graf. Graf yang dibangun merepresentasikan teks dimana *vertex* mewakili unit teks dan *edge*-nya merupakan relasi antar *vertex*.

Tugas Akhir ini mengimplementasikan *TextRank* pada ekstraksi kata kunci dengan melakukan pengujian pada pengaturan parameter *TextRank*, yaitu jenis graf, filter kata atau pemilihan kata yang direpresentasikan sebagai *vertex*, dan ukuran *window*. Jenis graf yang akan diuji adalah *undirected*, *directed forward*, dan *directed backward*. Filter kata yang digunakan dapat dibagi menjadi 3 kategori, yaitu *all open class words* atau semua kata, kata benda saja, dan kata benda yang digabung dengan kata sifat. Sedangkan ukuran *window* divariasikan antara 2 sampai 10 kata. Pengujian juga dilakukan terhadap variasi nilai *damping factor* dan *threshold* yang digunakan dalam pumus algoritma *TextRank*.

Evaluasi dilakukan dengan cara membandingkan kata kunci hasil ekstraksi sistem dengan kata kunci yang dibuat oleh manusia. Setiap teks masukan diuji terhadap kombinasi parameter *TextRank* yang diimplementasikan dalam sistem ekstraksi yang dibuat. Kata kunci yang dihasilkan kemudian dibandingkan dengan kata kunci asli dengan parameter evaluasi *recall*.

Data uji yang digunakan adalah abstraksi jurnal TA S1 fakultas Informatika ITTelkom dan abstraksi jurnal Seminar Nasional Aplikasi Teknologi Informasi (SNATI). Alasan pemilihan data tersebut adalah karena abstrak pada umumnya

selalu dilengkapi dengan kata kunci yang diberikan oleh penulisnya. Hal ini untuk memudahkan proses evaluasi sistem.

1.2 Perumusan masalah

Perumusan masalah dalam Tugas Akhir ini adalah sebagai berikut:

1. Bagaimana menerapkan algoritma *TextRank* pada sistem ekstraksi kata kunci teks berbahasa Indonesia.
2. Bagaimana pengaruh pengaturan parameter-parameter *TextRank* (jenis graf, jenis filter kata, dan ukuran *window*) terhadap kata kunci hasil ekstraksi sistem bila dibandingkan dengan kata kunci asli dengan parameter uji *recall*.
3. Bagaimana pengaruh penggunaan variasi nilai *damping factor* dan *threshold* terhadap *recall* sistem.

Batasan masalah dalam Tugas Akhir ini adalah :

1. Teks yang digunakan sebagai data dalam Tugas Akhir ini adalah abstrak jurnal-jurnal TA S1 Teknik Informatika dan abstrak jurnal-jurnal Seminar Nasional Aplikasi Teknologi Informasi (SNATI).
2. Format teks dalam bentuk .txt.
3. Tidak melakukan proses *stemming*.
4. Sistem yang dibangun bersifat *offline*.
5. Untuk keperluan evaluasi, kata kunci asli yang telah ditentukan penulis jurnal digunakan sebagai pembanding hasil ekstraksi sistem. Oleh karena itu kata kunci asli dianggap sebagai pilihan kata kunci yang benar.

1.3 Tujuan

Tujuan Tugas Akhir ini adalah:

1. Menerapkan algoritma *TextRank* pada sistem ekstraksi kata kunci teks berbahasa Indonesia.
2. Menganalisis pengaruh pengaturan parameter-parameter *TextRank* (jenis graf, jenis kata, dan ukuran *window*) terhadap kata kunci hasil ekstraksi sistem bila dibandingkan dengan kata kunci asli dengan parameter uji *recall*.
3. Menganalisis pengaruh penggunaan variasi nilai *damping factor* dan *threshold* terhadap *recall* sistem.

1.4 Metodologi penyelesaian masalah

Metodologi yang dilakukan untuk menyelesaikan permasalahan pada Tugas Akhir ini adalah sebagai berikut:

1. Melakukan studi literatur khususnya mengenai ekstraksi kata kunci dan *TextRank*
2. Melakukan pencarian data yang diperlukan untuk mendukung penyelesaian masalah, yaitu jurnal-jurnal TA S1 Teknik Informatika IT Telkom dan jurnal-jurnal SNATI yang berbahasa Indonesia.

3. Melakukan analisis algoritma yang akan diterapkan dalam ekstraksi kata kunci. Algoritma yang dianalisis adalah *TextRank*.
4. Melakukan analisis kebutuhan perangkat lunak dan perancangan perangkat lunak sistem ekstraksi kata kunci yang menerapkan algoritma *TextRank*.
5. Melakukan implementasi sistem sesuai dengan perancangan yang telah dilakukan.
6. Melakukan pengujian sistem dan menganalisis hasil keluaran sistem yang berupa kumpulan kata kunci. Pengujian dilakukan dengan melakukan perubahan pengaturan parameter TextRank dan membandingkan kata kunci hasil sistem dengan kata kunci asli dengan parameter *recall*.
7. Pembuatan laporan Tugas Akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan analisis hasil pengujian, dapat disimpulkan hal-hal sebagai berikut:

1. *TextRank* dapat diimplementasikan pada pengekstraksi kata kunci bahasa Indonesia
2. Secara umum, jenis graf *undirected* dapat memberikan nilai akurasi terbaik jika dibandingkan dengan graf *directed forward* dan *directed backward*. Hal ini disebabkan karena sebenarnya pada teks tidak ada arah yang dibangun dari satu kata ke kata lainnya.
3. Berdasarkan hasil analisis terhadap 3 filter kata yang digunakan, filter kata benda menghasilkan nilai akurasi terbaik. Hal ini disebabkan kata kunci biasanya disusun dari kata benda.
4. Ukuran *window* yang menghasilkan nilai *recall* tertinggi bergantung pada graf dan filter kata yang digunakan. Pada kombinasi graf-filter *undirected*-semua kata dan *undirected*-kata benda, recall tertinggi berada pada *window*=7. Pada kombinasi *undirected*-kata benda&sifat, recall tertinggi berada pada *window*=8. Untuk graf *directed forward*, *window* dengan *recall* tertinggi ada pada ukuran 5 jika menggunakan filter semua kata dan kata benda&sifat, sedangkan untuk filter kata benda *recall* tertingginya ada pada *window*=4. Pada kombinasi graf *directed backward* dan filter semua kata, *recall* tertinggi berada pada *window*=3, sedangkan jika dikombinasikan dengan filter kata benda dan kata benda&sifat, *recall* tertinggi berada pada ukuran *window*=2.
5. Kombinasi parameter yang baik untuk digunakan dalam ekstraksi kata kunci bahasa Indonesia berdasarkan hasil analisis adalah kombinasi graf *undirected*, filter kata benda, dan ukuran *window* 7-8.
6. Pengujian terhadap variasi nilai *damping factor* dan *threshold* menunjukkan bahwa penggunaan *damping factor* yang semakin besar meningkatkan *recall* sistem meskipun tidak terlalu signifikan dan *recall* sistem tidak terlalu dipengaruhi oleh nilai *threshold*.

5.2 Saran

Beberapa saran untuk perbaikan ekstraksi kata kunci dengan *TextRank* adalah sebagai berikut:

1. Menambah kamus jenis kata sehingga penggunaan sistem tidak terbatas pada dokumen yang telah diujikan saja.
2. Melakukan pengujian terhadap jenis dokumen lain yang isi dokumennya lebih banyak.

Referensi

[1] BPPT, “SiDoBi: Sistem Ikhtisar Dokumen untuk Bahasa Indonesia”, Open Source Software oleh Badan Pengkajian dan Penerapan Teknologi. http://www.inn.bppt.go.id/

[2] Giarlo, Michael J., "A Comparative Analysis of Keyword Extraction Techniques". http://lackoftalent.org/michael/papers/596.pdf, di-download pada 26 Juli 2009.

[3] Hulth, Anette dan Be´ata B. Megyesi, 2006, “A Study on Automatically Extracted Keywords in Text Categorization”. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.106.6825, di-download pada 24 Mei 2009.

[4] Learning Center IT Telkom. www.ittelkom.ac.id/library, di-download pada 12 Januari 2010.

[5] Mihalcea, Rada dan Paul Tarau. TextRank: Bringing Order into Texts. http://www.cse.unt.edu/~rada/papers/mihalcea.emnlp04.pdf, di-download pada 24 Mei 2009.

[6] Muresan, Smaranda. Graph-based Algorithms in IR and NLP. http://www.umiacs.umd.edu/~resnik/ling773_sp2007/slides/graph_based_methods.ppt di-download pada 26 Mei 2009.

[7] Prosiding SNATI. www.journal.uii.ac.id/index.php/Snati, di-download pada 11 Januari 2010.

[8] Pusat Bahasa Departemen Pendidikan Nasional, 2002, "Kamus Besar Bahasa Indonesia Edisi Ketiga", Jakarta: Balai Pustaka.

[9] Turney, P. Learning to Extract Keyphrases from Text. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.8505, di-download pada 24 Mei 2009.

[10] Wibisono, Yudi, 2008, “Stop words untuk Bahasa Indonesia”. http://yudiwbs.wordpress.com/2008/07/23/stop-words-untuk-bahasa-indonesia, di-download pada 3 Maret 2010.

[11] ___. Graph (mathematics). http://en.wikipedia.org/wiki/Graph_(graph_theory).htm, di-download pada tanggal 16 Juli 2009

[12] ___. Keyword. http://www.thefreedictionary.com/keyword di-download pada 26 Juli 2009