EKSTRAKSI KATA KUNCI DENGAN MENGGUNAKAN CONDITIONAL RANDOM FIELDS

Minardi Darmawan[1], Imelda Atastina[2], Yanuar Firdaus A.w.[3]

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

Abstrak

Berbagai informasi dalam jumlah besar dapat diperoleh dengan mudah dari Internet. Tetapi, tidak mudah untuk mencari informasi yang berguna diantara sekian banyak sumber informasi yang kita dapatkan. Beberapa search engine dapat melakukan pencarian teks lengkap, tetapi hasilnya kurang memuaskan. Efiesiensi pencarian dari jumlah dokumen yang sangat banyak sangat dipengaruhi oleh kualitas dari kata kunci yang diberikan oleh setiap dokumen. Oleh karena itu, perlu dibutuhkan kata kunci yang berkualitas yang diambil secara otomatis di-ekstrak dari dokumen-dokumen agar mendapatkan nilai efisiensi yang tinggi. Pada Tugas Akhir ini, akan membahas mengenai "Ekstraksi Kata Kunci dengan menggunakan Conditional Random Fields". Conditional Random Fields (CRF) adalah model probabilistik untuk segmentasi dan pelabelan data sekuens. CRF menggunakan teknik yang berbeda dalam melakukan preprocessing dimana kalimat-kalimat yang ada akan disegmentasi dan diberi label, sehingga akan memberikan pengaruh terhadap hasil dari ekstraksi kata kunci tersebut. Pengujian dilakukan dengan menghitung nilai precision, recall dan F-Measure untuk mengetahui performansi sistem yang akan dibuat. Hasil pengujian menunjukkan bahwa ekstraksi kata kunci dengan CRF memiliki akurasi yang lebih baik pada jenis dokumen yang bersifat umum (topik umum) atau dengan kata lain memiliki variasi data yang tinggi. Sedangkan dari sisi fungsi fiturnya, fitur POS (tanpa normalisasi) memberi pengaruh yang cukup signifikan terhadap performansi sistem, sedangkan fitur Len tidak memberikan pengaruh yang signifikan terhadap performansi sistem. Selain itu, berdasarkan jumlah dokumen trainingnya, ekstraksi kata kunci dengan CRF bisa diterapkan secara efektif walaupun jumlah data trainingnya sedikit.

Kata Kunci : Ekstraksi kata kunci, Conditional Random Fields, preprocessing, fitur POS, fitur Len, F-Measure.

Abstract

A variety of large amounts of information can be obtained easily from the Internet. However, it is not easy to find useful information among the many sources of information that we get. Some search engines can perform full text search, but the results are less satisfactory. Efficiency search of the number of documents that are very much very much influenced by the quality of the keywords provided by each document. Therefore, the necessary quality required keywords are retrieved automatically extracted from documents in order to obtain high efficiency values. In this Final Project, will discuss "Keyword Extraction using Conditional Random Fields". Conditional Random Fields (CRF) are probabilistic models for segmenting and labeling sequence data. CRF using different techniques in performing preprocessing in which sentences are to be segmented and labeled, so that will give effect to the results of keyword extraction. Test results showed that the extraction of keywords with CRF have a better accuracy on the document type of a general nature (general topics), or in other words having high data variation. In terms of its functions, POS features (without normalization) gives significant effect on system performance, while Len feature does not have a significant influence on system performance. In addition, training was based on the number of documents, extraction of keywords with CRF can be applied effectively even though the amount of training was a bit of data.

Keywords : Keyword extraction, Conditional Random Fields, preprocessing, POS feature, Len feature, F-Measure.

1. Pendahuluan

1.1 Latar Belakang

Pada saat ini, berbagai informasi dalam jumlah besar dapat diperoleh dengan mudah dari Internet. Tetapi, tidak mudah untuk mencari informasi yang berguna diantara sekian banyak sumber informasi yang didapatkan. Beberapa *search engine* dapat melakukan pencarian teks lengkap, tetapi hasilnya kurang memuaskan. Efisiensi pencarian dari jumlah dokumen yang sangat banyak sangat dipengaruhi oleh kualitas dari kata kunci yang diberikan oleh setiap dokumen. Oleh karena itu, dibutuhkan kata kunci yang berkualitas yang diambil secara otomatis di-ekstrak dari dokumen-dokumen agar mendapatkan nilai efisiensi yang tinggi.

Ekstraksi kata kunci adalah suatu metode untuk meng-identifikasi serangkaian kecil kata-kata, frase kunci, kata kunci, atau segmen kunci dari sebuah dokumen yang dapat menjelaskan makna dari sebuah dokumen. Karena kata kunci adalah unit terkecil yang dapat mengungkapkan makna dokumen, banyak aplikasi *text-mining* dapat mengambil keuntungan dari hal itu seperti *automatic indexing, automatic summarization, automatic classification, clustering, automatic filtering* dan lain sebagainya.

Tugas Akhir ini akan membahas mengenai ”Ekstraksi Kata Kunci dengan menggunakan Conditional Random Fields”. Conditional Random Fields (CRF) adalah model probabilistik untuk segmentasi dan pelabelan data sekuens. CRF menggunakan teknik yang berbeda dalam melakukan *preprocessing* dimana kalimat-kalimat yang ada akan disegmentasi dan diberi label sesuai *part-of-speech (POS)* dari masing-masing kata yang dihasilkan. Fitur lain yang bisa diekstrak dari sebuah dokumen adalah TF-IDF dan *length-of-word* (Len). TF-IDF digunakan sebagai bobot/nilai dari kepentingan sebuah kata, baik dalam sebuah dokumen maupun dalam kumpulan dokumen (*corpus*). Nilai *precision, recall* dan *F-measure* digunakan untuk mengetahui performansi dari hasil penelitian ini. Hipotesis awal dari penelitian ini adalah dengan menggunakan metode Conditional Random Fields akan menghasilkan performansi yang baik.

1.2 Perumusan Masalah

Berdasarkan pada latar belakang diatas, permasalahan yang menjadi fokus pada tugas akhir ini diantaranya yaitu:

a. Bagaimana cara mengimplementasikan ekstraksi kata kunci dengan menggunakan Conditional Random Fields?
b. Bagaimana karakteristik pemodelan Conditional Random Fields dalam melakukan proses ekstraksi kata kunci?
c. Seberapa baikah tingkat performansi yang bisa diperoleh dengan menggunakan Conditional Random Fields?

Sedangkan yang menjadi batasan masalah dalam tugas akhir ini diantaranya yaitu:

a. File yang diinputkan hanya berupa dokumen teks.
b. Dokumen teks yang digunakan dalam tugas akhir ini adalah artikel berbahasa Inggris.
c. Dataset yang digunakan adalah koleksi dokumen dari INSPEC database yang didapat dari https://github.com/snkim/AutomaticKeyphraseExtraction/blob/master/Hulth2003.tar.gz
d. Simulasi yang dibuat berbasis web menggunakan PHP dan basisdata MySQL.
e. Tidak dilakukan stemming terhadap dokumen inputan.
f. Aplikasi hanya melakukan *word* indexing dan tidak melakukan *phrase* indexing.

1.3 Tujuan

Adapun tujuan dari tugas akhir ini adalah :

a. Mengimplementasikan Ekstraksi Kata Kunci dengan menggunakan Conditional Random Fields.
b. Menganalisis Ekstraksi Kata Kunci dengan menggunakan Conditional Random Fields berdasarkan hasil pengukuran precision, recall dan F-measure.

1.4 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang digunakan dalam menyelesaikan penelitian Tugas Akhir ini adalah:

a. Studi literatur
   Pencarian referensi dan sumber-sumber lain yang dapat digunakan sebagai acuan dalam pembuatan tugas akhir ini.
b. Analisis dan Perancangan Perangkat Lunak
   Pada tahap ini dilakukan proses analisis requirement perangkat lunak yang akan dibangun sehingga didapat gambaran mengenai sistem yang akan dibuat.
c. Implementasi Sistem
   Melakukan implementasi sistem sesuai dengan hasil analisis dan perancangan yang telah dilakukan di tahap dua. Tahap-tahap implementasinya antara lain :
   - Preprocessing.
     Merubah dokumen teks tersebut menjadi *text chunking*. *Text chunking* adalah text yang telah disegmentasi dan diberi label *Part-of-speech* (POS). Kemudian melakukan ekstraksi fitur-fitur yang ada di dalam CRF model.
   - CRF model training.
     Melakukan training data untuk menghasilkan file CRF model.
   - CRF labeling and keyword extraction.

Menginputkan data testing kemudian melakukan ekstrasi kata kunci dengan menggunakan file CRF model yang telah dibuat.

d. Analisis Hasil Implementasi
   Menganalisis hasil implementasi aplikasi sehingga didapat data-data mengenai akurasi dari metode yang diimplementasikan.
e. Pembuatan Laporan
   Merupakan tahapan pendokumentasian dari penelitian yang dikerjakan serta mengambil kesimpulan dari penelitian yang dikerjakan

1.5 Sistematika Penulisan

Tugas akhir ini disusun dengan sistematika penulisan sebagai berikut:

BAB I **PENDAHULUAN**

Berisi pemaparan mengenai latar belakang permasalahan, tujuan yang ingin dicapai dengan adanya penelitian ini, perumusan masalah, batasan masalah, metodologi tugas akhir, dan sistematika penulisan.

BAB II **LANDASAN TEORI**

Berisi uraian mengenai landasan teori yang akan digunakan, meliputi teori tentang Conditional Random Fields dan teori-teori lain yang berkaitan dengan penelitian tugas akhir ini

BAB III **ANALISIS DAN PERANCANGAN SISTEM**

Berisi tentang analisa dan perancangan terhadap sistem yang akan dibangun.

BAB IV **IMPLEMENTASI DAN PENGUJIAN**

Berisi implementasi dari hasil analisa dan perancangan sistem yang dibuat, serta pengujian sistem.

BAB V **KESIMPULAN DAN SARAN**

Berisi kesimpulan dan saran-saran untuk pengembangan lebih lanjut terhadap hasil penelitian ini.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari tahapan perancangan hingga pengujian yang dilakukan pada sistem adalah sebagai berikut:

1. Ekstraksi kata kunci dengan CRF memiliki akurasi yang lebih baik pada jenis dokumen yang bersifat umum (topik umum) dengan variasi data yang tinggi.
2. Pada ekstraksi kata kunci dengan CRF, fitur POS (tanpa normalisasi) memberi pengaruh yang cukup signifikan terhadap performansi sistem, sedangkan fitur Len tidak memberikan pengaruh yang signifikan terhadap performansi sistem.
3. Ekstraksi kata kunci dengan CRF bisa diterapkan secara efektif untuk jumlah data training yang sedikit.

5.2 Saran

Berdasarkan hasil analisis dan kesimpulan, terdapat beberapa saran untuk perbaikan pada penelitian ekstraksi kata kunci sebagai berikut:

1. Menggabungkan konsep ekstraksi kata kunci yang diimplementasikan pada tugas akhir ini dengan konsep stemming yang mengembalikan bentuk kata-kata yang berimbuhan menjadi kata dasarnya.
2. Perangkat lunak tidak hanya menangani *word indexing*, tapi juga dapat menangani *phrase indexing*.

Daftar Pustaka

[1] Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, Bo Wang, 2008, Automatic Keyword Extraction from Documents Using Conditional Random Fields.
http://www.sciencenet.cn/upload/blog/file/2008/9/2008931007461407.p
Di download tanggal 20 November 2009

[2] Annete Hulth - Improved Automatic Keyword Extraction Given More Linguistic Knowledge.
http://www.ldc.upenn.edu/acl/W/W03/W03-1028.pdf
Di download tanggal 22 November 2009

[3] Iryna Oelze, 2009, Automatic Keyword Extraction for Database Search.
http://tcc.itc.it/people/pianta/publications/wse2003clustKeywords.pdf
Di download tanggal 2 Desember 2010

[4] Michael J. Giarlo., 2006, A Comparative Analysis of Keyword Extraction Techniques.
http://lackoftalent.org/michael/papers/596.pdf
Di download tanggal 21 November 2009

[5] Yutaka Matsuo, Mitsuru Ishizuka, 2003, Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.5601&rep=rep1&type=pdf
Di download tanggal 21 November 2009

[6] Michael J. Giarlo., 2006, A Comparative Analysis of Keyword Extraction Techniques.
http://lackoftalent.org/michael/papers/596.pdf
Di download tanggal 21 November 2009

[7] Charles Sutton, Andrew McCallum, 2006, An Introduction to Conditional Random Fields for Relational Learning.
http://www.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf
Di download tanggal 24 November 2009

[8] Christopher D. Manning, Prabhakar Raghavan, Hinrich Sch ¨ utze. – Introduction Information Retrieval.
http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
Di download tanggal 22 November 2009

[9] John Kirk, 2005, Word Frequency and Keyword Extraction
http://www.methodsnetwork.ac.uk/redist/pdf/es1abstracts.pdf
Di download tanggal 22 November 2009