EKSTRAKSI KATA KUNCI PADA DOKUMEN TEKS MENGGUNAKAN METODE NAIVE BAYES

Lisa Maharsi[1], Yanuar Firdaus A.w.[2], Arie Ardiyanti Suryani[3]

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

Abstrak

Perkembangan teknologi menyebabkan terjadinya penumpukan data berupa dokumen teks baik secara online maupun offline. Dokumen teks yang menumpuk menyebabkan sulitnya mencari dokumen yang sesuai dengan kebutuhan. Untuk kemudahan pencarian dokumen yang sesuai, maka dibutuhkan kata kunci yang mengiringi dokumen. Kata kunci mewakili isi dokumen secara keseluruhan, sehingga pembaca dapat dengan mudah mencari dokumen sesuai dengan yang dibutuhkan. Pada tugas akhir ini, diimplementasikan metode Naive Bayes untuk mengekstraksi kata kunci dokumen. Proses ini membutuhkan dokumen training yang telah disertai kata kunci, sehingga dengan menggunakan fitur-fitur tertentu dapat memberikan learning kepada sistem bagaimana probabilitas atau peluang kata pada dokumen masukan menjadi kata kunci. Pengujian dilakukan untuk mengetahui keakurasian kata kunci yang dihasilkan oleh sistem dengan menggunakan metode Naive Bayes berdasarkan parameter precision, recall, dan f-measure. Penambahan jumlah dokumen training, menyebabkan meningkatnya keakurasian hasil ekstraksi kata kunci. Namun, penggunaan jumlah dokumen training yang terlalu besar menyebabkan penurunan nilai keakurasian. Penggunaan 4 fitur pada proses ekstraksi memberikan hasil keakurasian yang lebih baik dibandingkan dengan penggunaan 2 fitur. Kemampuan Naive Bayes dalam mengekstraksi kata kunci dengan benar dapat dilihat dari nilai recall. Dalam rentang 20 jumlah kata kunci yang dihasilkan sistem, Naive Bayes mampu memberikan nilai recall sebesar 0,75 (sekitar 75% sistem mampu mengekstraksi benar kata kunci yang sesuai dengan dokumen masukan). Ekstraksi kata kunci dengan menambahkan eliminasi stopwords memberikan hasil keakurasian yang lebih baik dibanding dengan tanpa eliminasi stopwords.

Kata Kunci : Ekstraksi kata kunci, kata kunci, naive bayes

Abstract

Technological developments lead to the accumulation of data in the form of text documents either online or offline. Text documents that accumulate causing difficulty to find documents as needed. For ease of searching the appropriate documents, it is required to accompany the keywords document. Keywords representing the contents of the document as a whole, so that readers can easily search for documents in accordance with the required. In this final project, implemented Naive Bayes method to extract the document keywords. This process requires training documents that have been accompanied by keywords, so by using certain features can provide the system of learning how the probability or the chance to document the inputs into the keyword. Tests performed to determine the accuracy of the keywords generated by the system using Naive Bayes method based on the parameters of precision, recall, and f-measure. The addition of training documents, increase in the accuracy of the results of keyword extraction. However, the use of the number of training documents is too large causes a decrease in the value of accuracy. Use of the 4 feature extraction processes yield better accuracy than the use of feature 2. Naive Bayes capability in extracting the correct keywords can be seen from the value of recall. In the span of 20 the number of keywords that generated the system, Naive Bayes can provide a recall value of 0.75 (about 75% of the system is capable of extracting keywords according to the input document). Extraction of keywords by adding stopwords elimination yield better accuracy than without stopwords elimination.

Keywords : Keyword extraction, keyword, naive bayes

1. Pendahuluan

1.1 Latar Belakang

Seiring dengan perkembangan jaman, maka kebutuhan manusia akan teknologi dan informasi semakin meningkat. Bahkan dengan adanya Internet, segala kebutuhan akan informasi menjadi lebih mudah didapatkan. Namun perkembangan Internet yang sangat pesat, menyebabkan penumpukan data dan informasi yang sangat banyak di Internet sehingga seringkali orang bingung untuk memilih informasi mana yang sesuai dengan yang diinginkan karena hanya dengan judul belum cukup untuk menggambarkan keseluruhan isi dari sebuah dokumen. Maka dari itu orang membutuhkan sesuatu yang mudah dan praktis untuk mengetahui apakah suatu dokumen itu sesuai dengan informasi yang dibutuhkan.

Kata kunci adalah kata atau kumpulan kata pada dokumen yang mampu memberikan deskripsi dari keseluruhan isi dokumen. Maka dengan adanya kata kunci dalam dokumen, pembaca akan lebih mudah mendapatkan dokumen yang isinya sesuai dengan yang dibutuhkan. Namun, kebanyakan dokumen justru tidak disertai dengan kata kunci. Pengekstraksian kata kunci secara manual untuk dokumen dalam jumlah besar membutuhkan waktu dan tenaga yang besar. Oleh karena itu, dibutuhkan suatu automasi pengekstraksian kata kunci. Selain itu dengan adanya kata kunci pada dokumen akan memudahkan proses information retrieval yang lain, seperti pengindeksian, klasterisasi dokumen, klasifikasi dokumen, text summarization, dan lain-lain.

Ekstraksi kata kunci merupakan suatu permasalahan klasifikasi, yaitu bagaimana mengklasifikasikan kata pada dokumen masukan menjadi kata kunci atau kata biasa (bukan kata kunci). Pada tugas akhir ini, akan diimplementasikan metode Naïve Bayes untuk mengekstraksi kata kunci pada dokumen teks. Naive Bayes merupakan sebuah pendekatan statistik yang mampu memutuskan sebuah pilihan berdasarkan probabilitas dan minimum error yang ditimbulkan atas pilihan tersebut berdasarkan training set yang diberikan. Metode Naïve Bayes ini sederhana dan tidak membutuhkan aturan semantik untuk proses ekstraksi. Naïve Bayes yang merupakan suatu metode learning, diharapkan mampu mengekstraksi kata kunci pada dokumen dan memberikan hasil yang baik dengan beberapa fitur yang digunakan pada training set, yaitu TFxIDF dan posisi kata-kata kunci pada dokumen.

1.2 Perumusan Masalah

Tugas Akhir ini mempunyai perumusan masalah sebagai berikut :

1. Bagaimana metode Naive Bayes dapat menemukan kata-kata yang merupakan kata kunci pada dokumen.
2. Bagaimana menentukan fitur-fitur yang digunakan untuk pengekstraksian kata kunci.
3. Bagaimana pengaruh penggunaan variasi fitur yang digunakan terhadap hasil ekstraksi.
4. Bagaimana pengaruh eliminasi *stopwords* sebagai tambahan pengetahuan kebahasaan.

5. Apakah metode yang diimplementasikan dapat menghasilkan kata-kata kunci yang sesuai dengan dokumen input.

Adapun batasan masalah dari tugas akhir ini adalah :

1. Dokumen teks yang digunakan dalam tugas akhir ini adalah dokumen abstraksi tugas akhir mahasiswa IT Telkom yang diambil dari *website* perpustakaan IT Telkom. Dokumen abstraksi tugas akhir yang digunakan sudah disertai dengan kata kunci, yang akan digunakan dalam proses *training* dan pengujian untuk mengetahui nilai keakurasian dari kata kunci yang dihasilkan sistem.
2. Term-term yang digunakan dalam dokumen latih dan dokumen uji merupakan kata. Frase penting dianggap sebagai term terkecil (kata).
3. Tidak melakukan penanganan terhadap frase.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Mengimplementasikan metode Naïve Bayes untuk mengekstraksi kata kunci pada dokumen teks.
2. Menganalisis keakurasian metode Naïve Bayes dalam mengekstrak kata kunci pada dokumen teks berdasarkan kesesuaian antara kata kunci dan dokumen teks dengan parameter *precision*, *recall,* dan *f-measure*.
3. Menganalisis jumlah fitur yang digunakan dalam mengekstraksi kata kunci.
4. Menganalisis pengaruh eliminasi *stopwords* terhadap hasil ekstraksi kata kunci.

2.1 Metodologi Penyelesaian Masalah

Adapun metodologi yang digunakan dalam penulisan tugas akhir ini adalah:

1. Melakukan studi literatur mengenai ekstraksi kata kunci dan metode ekstraksi khususnya metode Naive Bayes.
2. Melakukan analisis kebutuhan perangkat lunak dan perancangan perangkat lunak ekstraksi kata kunci otomatis yang menerapkan metode Naive Bayes.
3. Melakukan implementasi perangkat lunak sesuai dengan perancangan yang telah dilakukan.
4. Melakukan pencarian data yang diperlukan untuk penelitian berupa dokumen abstraksi tugas akhir mahasiswa IT Telkom dari *website* perpustakaan IT Telkom.
5. Melakukan pengujian pada implementasi metode ekstraksi kata kunci terhadap dokumen abstraksi ta.
6. Melakukan analisis terhadap hasil pengujian.
7. Pembuatan laporan tugas akhir.

5. Penutup

5.1 Kesimpulan

Berdasarkan hasil analisis terhadap hasil pengujian dapat disimpulkan sebagai berikut:

1. Metode Naïve Bayes dapat diimplementasikan pada pengekstraksian kata kunci dan menghasilkan kata kunci yang ekstraktif.
2. Pengunaan 4 fitur dalam pengekstraksian kata kunci memberikan hasil yang lebih baik dibandingkan dengan penggunaan 2 fitur.
3. Semakin besar dokumen training yang digunakan, memberikan hasil yang semakin baik pada keakurasian sistem dalam mengekstraksi kata kunci. Namun jumlah dokumen training yang terlalu besar tidak selalu memberikan hasil yang optimum. Jumlah dokumen training pada penelitian ini yang memberikan hasil optimum adalah 30 dokumen training.
4. Ekstraksi kata kunci dengan Naïve Bayes mampu mengekstraksi kata kunci dengan keakurasian *recall* 0,725 dengan pengujian menggunakan 30 dokumen training dengan 20 jumlah kata kunci yang diekstraksi.
5. Eliminasi stopwords mempengaruhi hasil pengujian. Ekstraksi kata kunci dengan mengeliminasi stopwords memberikan hasil yang lebih baik dibandingkan dengan melibatkan stopwords dalam proses ekstraksi.

5.2 Saran

Berdasarkan hasil saran dan kesimpulkan, terdapat beberapa saran untuk perbaikan penelitian terhadap ekstraksi kata kunci:

1. Dilakukan pelitian terhadap pengaruh fitur lain yang dapat digunakan dalam proses ekstraksi.
2. Melakukan penelitian pada data uji yang tidak domain.

Referensi

[1] Uzun, Yasin., *Keyword Extraction Using Naïve Bayes*, Bilkent University, Department of Computer Science, Turkey. http://www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Workshop/Yasin_Uzun.pdf

[2] Ramesh Nallapati, James Allan, Sridhar Mahadevan, *Extraction of Key Words from News Stories*, Department of Computer Science University of Massachuset. http://ai.stanford.edu/~nmramesh/synthesis_report.pdf

[3] Huaizhong KOU and Georges Gardarin. Keywords Extraction, *Document Similarity and Categorization*. PRiSM Lab, University of Versailles, Franch. http://www.prism.uvsq.fr/rapports/2002/document_2002_22.pdf

[4] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. *Introduction to Data Mining.*Addison Wisley.USA.

[5] Jiawei Han, Micheline Kamber. *Data Mining Concept and Technique.*Morgan Kaufman.USA.

[6] Matsuo, Ishizuka. *Keyword Extraction from a Single Document using Word Co-occurance Statistical Information.*National Institute of Advanced Industrial Statistical Information. Japan. http://ymatsuo.com/papers/ijait04.pdf

[7] Renz, Ingrid., Ficzay, Andrea., Hitzler, Holger. *Keyword Extraction for Text Characterization.* DaimlerChrysler AG, Research and Technology. Germany. http://subs.emis.de/LNI/Proceedings/Proceedings29/GI-Proceedings.29-19.pdf

[8] Tonella, Paolo.,Ricca, Filippo., Pianta, Emanuale., Girardi, Christian., *Using Keyword Extraction for Web Site Clustering*, Italy. http://tcc.itc.it/people/pianta/publications/wse2003clustKeywords.pdf

[9] Hulth, Anette. *Improved Automatic Keyword Extraction Given More Linguistic Knowledge.* Department of Computer and Systems Sciences Stockholm University. Sweden. http://acl.ldc.upenn.edu/acl2003/emnlp/pdfs/Hulth.pdf

10] Kirk, John M. *Word Frequency and Keyword Extraction.* AHRC ICT Methods Network Expert Seminar on Linguistics. http://www.methodsnetwork.ac.uk/redist/pdf/es1abstracts.pdf