ISSN: 2089-1121

EKSTRAKSI KEYPHRASE DARI SUATU SITUS DENGAN ALGORITMA KEA

Gunawan*), Arya Tandy Hermawan), Denny Saputra J.P.*), dan Joan Santoso*)**
*) Jurusan Teknik Informatika
Sekolah Tinggi Teknik Surabaya
**) Program Pascasarjana Teknologi Informasi
Sekolah Tinggi Teknik Surabaya
gunawan@stts.edu , arya@stts.edu , - , joan@stts.edu

ABSTRAK

Keyphrase merupakan sarana yang penting dalam meringkas dokumen, mengelompokkan dokumen ataupun melakukan pencarian dokumen dengan topik tertentu. Artikel jurnal umumnya memiliki daftar keyphrase, sedangkan halaman-halaman web umumnya tidak disertai daftar keyphrase. Proses pemberian daftar keyphrase pada dokumen secara manual membutuhkan suatu usaha yang sangat melelahkan. Oleh karena itu dibutuhkan proses yang mampu mengekstraksi keyphrase secara langsung.

Pada penelitian ini diberikan deskripsi mengenai KEA, sebuah algoritma ekstraksi keyphrase secara langsung dari sebuah teks. KEA melakukan identifikasi kandidat frase dengan menggunakan metode leksikal, melakukan kalkulasi nilai feature untuk setiap kandidat frase, dan menggunakan algoritma mesin pembelajaran Naïve Bayes dalam memprediksi kandidat frase mana yang merupakan keyphrase yang baik. Mesin pembelajaran awalnya membentuk sebuah model klasifier dengan menggunakan dokumen pelatihan yang telah disertai daftar keyphrasenya, dan kemudian menggunakan model tersebut dalam mencari keyphrase pada sebuah dokumen yang baru.

Uji coba sistem ekstraksi keyphrase dilakukan pada sebuah corpus "Computer Science Technical Report" untuk menunjukkan seberapa banyak keyphrase yang dipilih oleh penulis, dapat diidentifikasikan oleh KEA. Sistem ekstraksi keyphrase yang ini juga mampu melakukan ekstraksi keyphrase pada sebuah halaman web.

Kata kunci :*Keyphrase extraction*, KEA, *Naive Bayes*

ABSTRACT

Keyphrase is an important means of document summarization, document clustering, and topic search. A journal article usually has author-assigned keyphrases, but web page often does not have author-assigned keyphrases. A process of assigning keyphrases to a document manually needs a very tiring effort. Therefore it needs a process that can extract keyphrases automatically.

This research describes KEA, an algorithm for automatically extracting keyphrases from a text. KEA identifies phrase candidate using lexical method, calculates feature values for each phrase candidate, and uses a machine-learning algorithm, Naïve Bayes to predict which phrase candidates are good keyphrases. The

ISSN: 2089-1121

machine learning first builds a classifier model using training documents with known keyphrases, and then uses the model to find keyphrases in a new document.

The experiment of keyphrase extraction system is using a "Computer Science Technical Report" corpus to show how many author-assigned keyphrases that can be identified by KEA. This keyphrase extraction system is also able to extract keyphrase in a web page.

Keywords :Keyphrase Extraction,KEA, Naive Bayes

1. PENDAHULUAN

Setiap orang yang bergelut dalam dunia web mengetahui bahwa search engine merupakan cara yang paling efisien dan paling murah biayanya untuk mempromosikan sebuah situs web. Jika sebuah situs web merupakan situs yang ditunjuk paling atas peringkatnya oleh sebuah search engine, maka besar kemungkinan akan membuat banyak pengunjung untuk membuka dan menjelajahi situs tersebut. Oleh karenanya dikembangkan suatu metode yang mampu menunjukkan karakteristik dari sebuah halaman web. Pada banyak dokumen, khususnya makalah akademis, biasanya disertai dengan sejumlah keyword yang dipilih oleh penulis untuk memberikan deskripsi tentang dokumen tersebut.

Keyphrase dapat digunakan dalam sistem *information retrieval* sebagai deskripsi dari dokumen yang dikembalikan oleh suatu query atau sebagai salah satu cara untuk menjelajahi sebuah koleksi dokumen dan sebagai teknik pengelompokan dokumen. Sebagai tambahan, *keyphrase* (frase utama) dapat membantu user untuk merasakan isi dari sebuah koleksi, menunjukkan bagaimana query dapat dikembangkan, memfasilitasi pembacaan dokumen secara cepat dengan melakukan visualisasi frase-frase penting yang utama, dan menyediakan sarana yang besar manfaatnya untuk mengukur tingkat kesamaan dokumen.Dalam penelitian ini, yang akan dibahas adalah algoritma KEA, yang merupakan algoritma ekstraksi keyphrase. KEA menggunakan algoritma mesin pembelajaran Naïve Bayes untuk pelatihan dan ekstraksi keyphrase.

2. TINJAUAN PUSTAKA

Keyphrase terdapat dua jenis pendekatan yang berbeda secara dasar, yaitu: *keyphrase assignment* dan *keyphrase extraction*. Keduanya menggunakan metode mesin pembelajaran, dan membutuhkan pelatihan terhadap sejumlah dokumen yang sudah disediakan keyphrasenya.

Keyphrase assignment melakukan pencarian untuk memilih frase-frase dari pembendaharaan kata yang paling baik untuk mendeskripsikan sebuah dokumen. Data pelatihan menghubungkan sejumlah dokumen dengan setiap frase yang terdapat dalam pembendaharaan kata, dan membentuk sebuah pengklasifikasi untuk setiap frase. Sebuah dokumen baru diproses oleh setiap pengklasifikasi, dan mengalokasikan keyphrase dari model apapun yang mengklasifikasikan dokumen tersebut secara benar. Keyphrase yang dipilih adalah semua frase yang muncul dalam data pelatihan.

Sedangkan keyphrase extraction tidak memerlukan pembendaharaan kata, tetapi langsung memilih keyphrase dari teks itu sendiri. Keyphrase extraction menggunakan teknik lexical dan information retrieval untuk mengekstraksi frase dari dokumen teks yang menunjukkan karakteristik dari dokumen tersebut. Dalam pendekatan ini, data pelatihan digunakan untuk menyesuaikan parameter-parameter dari algoritma ekstraksi keyphrase.

ISSN: 2089-1121

3. METODE PENELITIAN

Dalam pelaksanaan penelitian ini dibutuhkan beberapa tahap proses yang harus dilakukan. Tahapan-tahapan yang dilakukan adalah sebagai berikut:

- Tahap pelatihan diberikan inputan berupa dokumen pelatihan dan sebuah corpus, yang merupakan kumpulan dokumen-dokumen pelatihan.Dokumen-dokumen pelatihan adalah dokumen-dokumen yang telah disertai dengan key word atau keyphrase oleh penulis dokumen tersebut. Hasil dari tahap ini adalah model klasifier yang akan digunakan oleh mesin pembelajaran Naive Bayes.
- Tahap esktraksi keyphrase dari sebuah dokumen. Tahap iniakan menghasilkan keyphrase-keyphrasedari dokumen yang akan dicari daftar keyphrasenya.
- Uji coba hasil ekstraksi keyphrase yang dilakukan untuk mendapatkan model klasifier yang mampu memberikan hasil yang sesuai dengan keyphrase yang dipilih oleh penulis/pembuat dokumen.

4. PEMILIHAN KANDIDAT FRASE

Fase pemilihan kandidat frase sangat diperlukan baik pada tahap pelatihan maupun pada tahap ekstraksi karena pada kedua tahap tersebut diperlukan untuk memproses dokumen input menjadi sekumpulan kandidat frase yang memiliki kemungkinan sebagai sekumpulan kandidat keyphrase yang baik. Pada tahap pelatihan, fase ini diperlukan untuk menghasilkan inputan yang digunakan untuk proses pembentukan model klasifier. Sedangkan, pada tahap ekstraksi, fase ini diperlukan untuk menghasilkan sekumpulan kandidat frase yang mungkin akan dipilih sebagai keyphrase yang baik berdasarkan model klasifier yang telah didapatkan pada tahap pelatihan.Fase ini terbagi menjadi tiga tahapan proses, yaitu: preprocessing input, identifikasi kandidat frase, dan case-folding dan stemming.

Fase preprocessing bertujuan untuk mempersiapkan daerah pembentukan kandidat frase dengan menentukan batasan awal dari daerah pembentukan kandidat frase. Dokumen inputan akan dipisah-pisahkan menjadi deretan *token-token* (sederetan huruf, bilangan dan bilangan desimal).Tidak semua token tersebut akan dipilih sebagai bagian dari kandidat frase yang akan dibentuk, sebagai contoh token bilangan akan dihilangkan dari daftar kandidat.Proses preprocessing ini juga melakukan beberapa hal yang lain dalam menentukan token mana yang boleh digunakan sebagai bagian kandidat frase, yaitu:semua tanda baca dan tanda kurung dihilangkan dan digantikan dengan karakter yang digunakan sebagai tanda batasan awal pembentukan kandidat frase,karakter petik tunggal dapat langsung dihilangkan saja, dan kata-kata yang dihubungkan dengan karakter tanda hubung akan dipisahkan menjadi token sendiri-sendiri.Setelah proses tersebut, yang harus dilakukan selanjutnya adalah menghilangkan *proper name* (nama diri).

Fase identifikasi kandidat fraseuntuk mengidentifikasikan dan membentuk kandidat frase yang mungkin pada daerah pembentukan yang telah dipersiapkan oleh tahapan fasepreprocessing.Dalam membentuk sebuah kandidat frase, algoritma KEA menggunakan stopword (sebuah daftar kata yang berisi semua kata sandang, kata depan, kata ganti, kata sifat, dan kata keterangan). Syarat sebuah kandidat frase adalah kandidat tersebut tidak boleh diawali atau diakhiri oleh sebuah kata stopword.Sebagai contoh: misalkan daerah pembentukan kandidat frase memiliki nilai "the programming by demonstration method", akan menghasilkan sejumlah deretan token, antara lain: "programming", "programming by demonstration", "demonstration", "programming by

ISSN: 2089-1121

demonstration method", "demonstration method", dan "method", yang diasumsikan sebagai kandidat-kandidat frase, karena kata "the" dan "by" terdapat pada daftar stopword.

Fase case-folding dan stemming bertujuan untuk penyetaraan bentuk kata dari semua kandidat frase yang telah diidentifikasikan. Proses case-folding dilakukan dengan cara mengubah semua token menjadi token yang hanya terdiri atas huruf bukan capital sedangakn untuk proses stemming dilakukan untuk mencari bentuk kata dasar dari sebuah kata jadian. Sebagai contoh: kata "cut elimination" akan dilakukan proses stemming menjadi "cut elim".

Pada tahap ekstraksi, kandidat frase yang masih belum di-stemming akan disimpan. Untuk mempresentasikan keyphrase yang dihasilkan oleh algoritma KEA, hasil yang diberikan adalah variasi bentuk kata yang paling sering muncul dari kandidat frase tersebut.

5. KALKULASI FEATURE

Berdasarkan algoritma KEA, terdapat dua nilai feature yang dikalkulasikan untuk setiap kandidat frase dan nilai feature ini akan digunakan baik pada tahap pelatihan maupun tahap ekstraksi. Nilai feature yang pertama adalah TFxIDF, yaitu sebuah nilai pengukuran dari frekuensi sebuah kandidat frase pada sebuah dokumen yang dibandingkan dengan tingkat keseringan kandidat frase tersebut digunakan pada keseluruhan dokumen pelatihan (corpus).Nilai feature ini melakukan perbandingan antara jumlah frekuensi dari penggunaan sebuah kandidat frase pada sebuah dokumen tertentu dengan jumlah dokumen yang menggunakan kandidat frase tersebut. Jumlah dokumen dimana sebuah kandidat frase muncul direpresentasikan oleh komponen document frequency. Karena dengan adanya komponen document frequency ini, maka nilai TFxIDF untuk setiap kandidat frase P pada dokumen D dapat dikalkulasikan dengan menggunakan persamaan berikut ini:

$$TFxIDF = \frac{freq(P,D)}{size(D)} \times -\log_2 \frac{df(P)}{N}$$

dimana:

- *freq*(*P,D*) adalah jumlah P yang muncul pada D.
- *size(D*) adalah jumlah kata dalam D.
- *df*(*P*) adalah jumlah dokumen yang mengandung P pada corpus.
- *N* adalah jumlah dokumen pada corpus.

Nilai feature yang kedua adalah pemunculan pertama, dimana nilai yang dihasilkan merupakan nilai posisi pemunculan pertama kandidat frase tersebut dalam sebuah dokumen.Nilai feature ini dikalkulasikan dengan cara menghitung posisi pemunculan pertama dari sebuah kandidat frase pada dokumen kemudian dibagi dengan jumlah kata yang muncul pada dokumen tersebut dimana hasil dari kalkulasi feature pemunculan pertama ini akan selalu bernilai antara 0 dan 1. Nilai feature ini menunjukkan bahwa posisi pemunculan pertama dari kandidat frase mana yang umumnya dipilih sebagai keyphrase yang dipilih oleh penulis dokumen.

6. PEMBENTUKAN MODEL KLASIFIER

Pembentukan model klasifier terdiri atas 3 proses. Proses yang pertama adalah pembentukan document frequency. Untuk mendapatkan nilai feature TFxIDF dari sebuah kandidat frase pada tahap pelatihan maupun pada tahap ekstraksi keyphrase

ISSN: 2089-1121

dilakukan proses pembentukan document frequency, dimana akan diketahui jumlah frekuensi dokumen dari sebuah kandidat frase atau jumlah dokumen dari corpus dimana kandidat frase tersebut muncul.

Proses yang kedua adalah dikritisasi nilai feature. Kedua nilai feature yang didapatkan dari proses kalkulasi merupakan nilai riil, oleh karenanya nilai-nilai tersebut harus dikonversikan menjadi nilai nominal untuk skema mesin pembelajaran atau untuk proses pembentukan model klasifier dan proses ekstraksi keyphrase. Selama proses pelatihan, sebuah tabel diskretisasi untuk masing-masing feature akan dibentuk berdasarkan data pelatihan. Tabel ini memberikan sekumpulan jangkauan nilai untuk masing-masing feature, dan nilai-nilai akan digantikan dengan jangkauan nilai dimana nilai tersebut masuk dalam jangkauan nilai tersebut.Metode diskretisasi yang digunakan antara lain metode equal width, metode equal frequency dan metode Minimum Description Length (MDL). Ketiga metode ini sangat berpengaruh terhadap proses pembentukan model klasifier.

Proses yang ketiga adalah pembentukan model klasifier. Pada tahap ini dimulai dengan dengan melakukan proses pemilihan sekumpulan kandidat frase dari masing-masing dokumen pelatihan, seperti yang dijelaskan sebelumnya. Kandidat frase yang muncul kurang dari minimal pemunculan akan diabaikan/dihilangkan, umumnya minimal pemunculan diisikan dengan 2. Untuk setiap kandidat frase yang masih tersisa, nilai feature TFxIDF dan nilai feature pemunculan pertama dikalkulasikan.Selanjutnya, diperlukan penyimpanan informasi apakah sebuah kandidat frase diidentifikasikan sebagai keyphrase yang baik oleh penulis/pembuat dokumen.Mesin pembelajaran Naïve Bayes melakukan prediksi atribut tujuan dengan cara mengkalkulasikan probabilitas atribut tujuan terhadap suatu data. Oleh karena itu, model klasifier yang akan dibentuk merupakan hasil kalkulasi nilai probabilitas atribut tujuan dari masing-masing interval kelas yang dibentuk pada proses pelatihan terhadap data-data pelatihan. Atribut tujuan yang dikalkulasikan pada model klasifier ini adalah dua macam nilai nominal, yaitu probabilitas sebagai keyphrase dan probabilitas sebagai non-keyphrase untuk masing-masing interval kelas yang terbentuk.

7. EKSTRAKSI KEYPHRASE

Tujuan dari proses ekstraksi keyphrase adalah mengklasifikasikan kandidat frase mana yang dapat dijadikan sebagai keyphrase yang baik berdasarkan model klasifier yang telah dibentuk pada tahap pelatihan.Proses pembentukan feature merupakan proses pengkategorian suatu nilai feature menjadi nilai nominal dari interval kelas yang telah terbentuk pada model klasifier. Proses ini dilakukan untuk masing-masing feature karena setiap feature memiliki interval kelas sendiri-sendiri.

Setelah didapatkan nilai nominal dari masing-masing feature untuk setiap kandiat frase, maka proses selanjutnya adalah mengaplikasikan nilai nominal tersebut pada model klasifier Naïve Bayes. Hal ini dilakukan untuk proses klasifikasi kandidat frase sebagai keyphrase dengan cara mengkalkulasi probabilitas kandidat frase tersebut sebagai keyphrase dengan persamaan sebagai berikut:

$$P[yes] = \frac{Y}{Y+N} P_{TF\times IDF}[t \mid yes] P_{dis\tan ce}[d \mid yes]$$

dimana nilai *Y* merupakan jumlah data pelatihan yang beratribut tujuan sebagai keyphrase dan nilai *N* merupakan jumlah data pelatihan yang beratribut tujuan sebagai non-keyphrase. Notasi t merepresentasikan nilai feature TFxIDF dan notasi d merepresentasikan nilai feature pemunculan pertama. Notasi$P_{TFxIDF}[t|yes]$

ISSN: 2089-1121

merepresentasikan probabilitas suatu kandidat frase dengan nilai feature *TFxIDFt* sebagai keyphrase. Sedangkan notasi $P_{distance}[d \mid yes]$ merepresentasikan probabilitas suatu kandidat frase dengan nilai feature pemunculan pertama *d* sebagai keyphrase.

Selain mengkalkulasikan probabilitas suatu kandidat frase sebagai keyphrase, diperlukan juga kalkulasi probabilitas suatu kandidat frase sebagai non-keyphrase untuk mendapatkan nilai probabilitas secara keseluruhan dari suatu kandidat frase sebagai sebuah keyphrase yang baik. Nilai probabilitas suatu kandidat frase secara keseluruhan dikalkulasikan dengan persamaan sebagai berikut:

$$p = P[yes]/(P[yes] + P[no])$$

dimana nilai *P[yes]* merupakan nilai probabilitas kandidat frase tersebut sebagai keyphrase dan *P[no]* merupakan nilai probabilitas kandidat frase tersebut sebagai non-keyphrase.

Suatu kandidat frase diranking berdasarkan nilai probabilitas ini. Jika terdapat dua kandidat frase yang memiliki nilai probabilitas yang sama, maka nilai feature TFxIDF digunakan sebagai pembanding mana yang lebih baik. Kemudian jika sebuah kandidat frase yang merupakan bagian dari kandidat frase lain yang memiliki nilai probabilitas lebih baik, maka kandidat frase tersebut akan dihilangkan.

Prinsip kerja proses ekstraksi keyphrase dari sebuah halaman HTML adalah sama. Semua tag HTML yang terkandung dalam halaman web tersebut harus dihilangkan terlebih dahulu melalui proses parsing. Namun proses ekstraksi keyphrase dari halaman HTML tidak hanya memperhatikan isi dari web, namun tetap harus memperhatikan isi teks tersebut berada pada tag HTML apa, karena isi teks yang terkandung pada sebuah tag HTML tertentu terkadang memberikan penekanan makna tertentu sehingga proses ekstraksi keyphrase dari halaman HTML memerlukan fasilitas pembobotan khusus pada isi teks yang terkandung pada tag-tag tertentu.

8. UJI COBA MODEL KLASIFIER

Uji coba pembentukan model klasifier dilakukan pada sebuah corpus "Computer Science Technical Report" yang merupakan sekumpulan dokumen atau artikel jurnal mengenai ilmu komputer.Uji coba pembentukan model dilakukan dengan menggunakan metode pengukuran berdasarkan jumlah keyphrase yang berhasil diidentifikasikan.Pada uji coba ini ditunjukkan bagaimana pengaruh jumlah dokumen pelatihan terhadap hasil algoritma ekstraksi keyphrase.

Uji coba dilakukan dengan menggunakan komponen document frequency yang terdiri atas 50 dokumen yang berbeda dengan dokumen pelatihan. Masing-masing model klasifier yang telah dibentuk dari jumlah dokumen pelatihan yang bervariasi itu akan diuji-cobakan pada dokumen percobaan sejumlah 100 dokumen yang dipilih secara acak. Uji coba menggunakan metode diskretisasi MDL dan jumlah keyphrase yang dipilih oleh algoritma dalam uji coba ini adalah sebanyak 15 buah keyphrase dengan nilai probabilitas tertinggi Hasil uji coba pengaruh jumlah dokumen pelatihan ditunjukkan pada tabel 1.

Uji coba kedua dilakukan dengan menggunakan metode diskretisasi yang berbeda.Uji coba dilakukan pada 100 dokumen percobaan.Model klasifier menggunakan document frequency (50 dokumen) dan jumlah keyphrase yang dipilih adalah sebanyak 15 buah keyphrase.Pembentukan model dilakukan dengan 100 dokumen pelatihan.Hasil uji coba ditunjukkan pada tabel 2.

ISSN: 2089-1121

Tabel 1.Hasil Uji Coba Pengaruh Jumlah Dokumen Pelatihan

Jumlah Dokumen Pelatihan	Jumlah Keyphrase yang Sesuai dengan Pilihan Penulis
10	1.03
20	1.25
50	1.42
100	1.44
200	1.46
400	1.48

Tabel 2.Hasil Uji Coba Pengaruh Metode Diskretisasi

Metode Diskretisasi	Jumlah Keyphrase yang Sesuai dengan Pilihan Penulis
Equal Width	1.16
Equal Frequent	1.13
MDL	1.44

9. PENUTUP

Dari hasil penelitian yang dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut Hasil uji coba yang ditunjukkan pada tabel 1 menunjukkan bahwa dengan bertambahnya jumlah dokumen pelatihan akan memberikan hasil yang lebih baik pada hasil ekstraksi algoritma keyphraseHasil uji coba yang ditunjukkan pada tabel 2 menunjukkan bahwa dengan menggunakan jumlah interval kelas yang sama, metode MDL mampu memberikan hasil yang jauh lebih baik daripada dua metode diskretisasi yang lain.

DAFTAR PUSTAKA

Dougherty, James., Kohavi, Ron,Sahami, Mehran. *Supervised and Unsupervised Discretization of Continuous Features,* In Armand Prieditis & Stuart Russell, eds, Machine Learning: Proceedings of the twelfth International Conference, Morgan Kaufmann Publishers, San Fransisco, CA.1995.

Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C. and Nevill-Manning,C.G..*Domain-Specific Keyphrase Extraction.* www.nzdl.org/Kea/Frank-et-al-1999-IJCAI.pdf 1999.

Han, Jiawei.,Camber,Michelin. *Data Mining: Concepts and Techniques,* Morgan Kaufinann Publishers.

Lovins, J.B. *Development of a Stemming Algorithm*. Mechanical Translation and Computational Linguistics, 11, 22-31.1968.

Mitchell, Tom M, *Machine Learning*, The McGraw-Hill Companies, Inc.

Turney, P.D., *Learning to Extract Keyphrase from Text,* National Research Council, Institute for Information Technology, Technical Report ERB-1057.1999.

Witten, Ian H., Paynter,Gordon W., Frank, Eibe, Gutwin, Carl,Nevill-Manning,Craig G..*KEA: Practical Automatic Keyphrase Extraction.* http://www.nzdl.org /Kea/Nevill-et-al-1999-DL99-poster.pdf.