DETEKSI DAN PENANDAAN NOISE PADA HALAMAN WEB BERITA BERBAHASA INDONESIA MENGGUNAKAN TEKNIK FEATURE WEIGHTING

Rahmat Firdaus[1], Yanuar Firdaus A.w.[2], Zk. Abdurahman Baizal[3]

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

Abstrak

Sebuah website di Internet memiliki banyak konten informasi dalam tiaptiap blok halaman yang ditampilkan. Kemudian tidak seperti kebanyakan data atau teks konvensional lainnya, suatu halaman web selain mengandung konten utama juga mengandung banyak blok informasi yang tidak berhubungan dengan konten utama misalnya, panel navigasi, copyright, user guide, links, sinopsis suatu berita, berbagai macam iklan dan lain-lain. Dalam hal ini blok-blok informasi yang tidak relevan dengan konten utama pada suatu halaman web disebut sebagai web pages noise.
Dalam tugas akhir ini akan digunakan teknik feature weighting untuk meningkatkan performansi hasil klasifikasi dengan mendeteksi noise yang ada pada halaman website. Dengan teknik feature weighting ini suatu halaman web pertama kali akan dimodelkan dengan pohon struktur Dokumen Object Model (DOM) tree dan Compressed structure tree(CST) untuk memperoleh struktur umum dan membandingkan blok-blok informasi dalam suatu website. Informasi yang didapatkan digunakan untuk melakukan pengukuran dan mengevaluasi tingkat kepentingan dari masing-masing node yang terbentuk dari compress struktur tree(CST).
Berdasarkan tree yang terbentuk dan tingkat kepentingan dari nilai node yang didapatkan, metoda ini memberikan bobot pada masing-masing individual word (feature) pada masing-masing blok kontent. Hasil pembobotan (weight) akan digunakan dalam proses web mining.

Kata Kunci : CST, DOM, deteksi noise, eliminasi noise, web mining.

Abstract

A Website on the Internet has shown a lot of information content in each block. Unlike conventional data or text, web pages not only have a main content but also typically contain a large amount of information that is not part of the main content of the pages, e.g., navigation bars, copyright, user guide, links, synopsis and also advertisement. The blocks information that is not the main content or irrelevant information in web pages is called web pages noise.
On this final project, feature weighting technique will be used to improve performance of classification with detection the noisy information in web pages. First, web pages will be modelled with structure tree Documents Object Model (DOM) tree and Compressed Structure Tree (CST) to capture the common structure and compare information block in a website. The Information that is captured will be used to measure and evaluate the importance of each node which is built from Compressed Structure Tree.
Based on the CST and the importance of weighting value, this method will put on a weight to each feature in each content block. The weighting result will be used to web mining process.

Keywords : CST, DOM, Noise Detection, Elimination, web mining

1. Pendahuluan

1.1 Latar belakang

Semakin pesatnya pengunaan Internet telah memacu pertumbuhan dan pertukaran informasi yang sangat pesat dibandingkan era sebelumnya. Sebagai akibatnya, jumlah informasi terus meningkat secara exponensial. Sumber informasi tidak hanya dalam bentuk dokumen atau teks tetapi di Internet jumlah informasi juga di hasilkan, dikoleksi dan dikonsumsi dalam jumlah yang tidak kalah besar di berbagai institusi dan perusahaan.

Sumber informasi di Internet yang dibawa oleh suatu website merupakan sumber informasi yang sangat beraneka ragam. Ada berbagai macam jenis web dengan berbagai macam informasi yang dibawa. Secara umum tidak seperti kebanyakan data atau teks konvensional lainnya suatu halaman web selain mengandung konten utama juga mengandung banyak blok informasi yang tidak berhubungan dengan konten utama misalnya, panel navigasi, *copyright*, *user guide*, kumpulan *link*, sinopsis suatu berita, berbagai macam iklan dan lain-lain yang secara keseluruhan adanya hal tersebut secara tidak langsung bertujuan untuk memudahkan user dalam mengakses web dan atau untuk tujuan bisnis tertentu.

Blok-blok informasi yang tidak relevan dengan konten utama pada suatu halaman web disebut sebagai *web pages noise*[5]. Adanya informasi yang tidak relevan dalam suatu sumber data *(web pages)* dapat menjadi permasalahan yang serius dalam proses data mining misalnya dalam proses klasifikasi atau klasterisasi.

Tugas akhir ini akan menggunakan teknik *feature weighting* dengan membobotkan tiap *feature* pada halaman web untuk kemudian dilakukan proses klasifikasi dengan tujuan meningkatkan performasi pada proses datamining. Dengan metoda ini suatu halaman web akan dimodelkan dengan struktur tree *Dokumen Object Model (DOM)* tree dan *Compressed Structure Tree (CST)* untuk mendapatkan struktur umum dan membandingkan blok-blok informasi dalam suatu website. Informasi yang didapatkan digunakan untuk melakukan pengukuran dan mengevaluasi tingkat kepentingan dari masing-masing node. Proses pengambilan data dari Internet digunakan *webcrawler* yang ada yaitu Teleport.

Hasil akhir dari tugas akhir ini adalah didapatkan suatu sumber data yang baik yang telah dideteksi noisenya dan diujikan dalam proses datamining yaitu proses klasifikasi. C*lassifier* yang digunakan adalah c*lassifier* yang terdapat dalam *tools* Weka. Kemudian akan dilakukan analisis perbandingan antara data yang diberi bobot dengan *feature weighting* dengan data yang diberi bobot dengan TFIDF dengan *clasifier* tersebut, serta akan menganalisis metoda pembobotan antara *feature weighting* dan TFIDF. Parameter yang digunakan untuk mengevaluasi proses deteksi noise adalah dengan menghitung nilai *F-measure*..

1.2 Perumusan masalah

Berdasarkan latar belakang yang dijelaskan diatas, maka permasalahan yang akan diangkat dalam tugas akhir ini adalah :

1. Bagaimana mendeteksi, dan mengurangi noise yang ada dalam suatu halaman web yang akan digunakan dalam proses klasifikasi.
2. Bagaimana menghitung bobot untuk setiap *feature* yang ada pada halaman website.
3. Bagaimana mengevaluasi proses *deteksi dan flagging* yang dilakukan dengan menghitung nilai akurasi dan *F-measure* dari proses klasifikasi.

Batasan masalah dalam tugas akhir ini adalah :

1. Menggunakan teknik *feature weighting* untuk melakukan pembobotan tiap feature ada dalam suatu halaman web.
2. Dataset yang diambil dalam betuk html *pages* secara *offline* yang merupakan berita berbahasa Indonesia.
3. Tiap halaman web berita sudah mempunyai layout standar dalam arti tiap halaman yang satu dengan yang lain memiliki kesesuaian bentuk umum, misalnya dibangun dengan CMS(*Content Managemen System*).
4. Proses pengambilan data menggunakan tools teleport
5. Tiap berita hanya akan dikelompokkan ke dalam satu kelompok berita.

1.3 Tujuan

Berdasarkan rumusan masalah diatas, tujuan yang ingin dicapai dari tugas akhir ini adalah :

1. Mengelola data pada halaman web agar siap dimining dengan teknik *feature weighting*.
2. Membandingkan dan menganalisis data sebelum dan setelah pendeteksian noise dengan mengevaluasi hasil klasifikasi.
3. Membandingkan dan menganalisis metoda pembobotan yang digunakan antara *feature weight dan TFIDF* dengan melihat nilai akurasi dan *f-measure* yang dihasilkan dari proses klasifikasi.

1.4 Metodologi penyelesaian masalah

Metodologi penyelesaian masalah yang digunakan pada Tugas Akhir ini adalah:

- **Studi Literatur**
  a. Pencarian referensi
     Mencari referensi yang layak dan berhubungan dengan topik tugas akhir, memahami dan memperlajari tentang *text mining, web mining, text categorization, web page cleaning,* teknik eliminasi *web pages Noise* dari berbagai jurnal, buku, Internet dan referensi lainnya yang mendukung.

- **Pengumpulan Data**
  Melakukan pengumpulan artikel berita yang berbentuk html yang akan digunakan sebagai data *collection*.

- **Perancangan perangkat lunak yang meliputi :**
  a. Implementasi perangkat lunak
     Mengimplementasikan perancangan menjadi perangkat lunak. Proses implementasi perangkat lunak dilakukan berdasarkan dari proses analisis dan perancangan yang telah dibangun.
  b. Pengujian
     Memeriksa error handling yang ada pada perangkat lunak yang dibangun misalnya kesalahan perhitungan, kesalahan dalam penginputan data, human error dan lain sebagainya.

- **Analisis hasil**
  Melakukan analisis terhadap tiap feature yang berpotensi sebagai noise, kemudian melihat hasil klasifikasi dari dokumen atau artikel web berita dengan mengevaluasi hasil keakuratan pengklasifikasian dokumen. Proses pengukuran keakuratan dari hasil klasifikasi dengan menggunakan konfusion matrik yang ada pada tools WEKA.

- **Pengambilan Kesimpulan dan Pembuatan Laporan**
  Pembuatan laporan Tugas Akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam Tugas Akhir ini.

Telkom University

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Proses deteksi noise pada data yang dilakukan secara manual oleh human analisis lebih baik dari pada yang dilakukan sistem.
2. Teknik *feature weighting* untuk membersihkan blok noise pada halaman website dapat meningkatkan hasil klasifikasi pada datamining berdasarkan hasil ujicoba perhitungan f-measure pada keempat dataset.
3. Metoda pembobotan TFIDF 5,5%-7,5% lebih baik dari pada pembobotan *feature weighting* jika menggunakan dataset yang telah terbebas dari noise berdasarkan hasil uji dari keempat dataset yang dihilangkan noisenya.
4. Blok noise pada halaman web dimana *feature*-nya sesuai dengan konten utama dapat membantu meningkatkan hasil klasifikasi

5.2 Saran

1. Proses pengambilan data dapat dikembangkan secara online, sehingga proses pengambilan data dapat dilakukan secara otomatis.
2. Pembobotan *feature weighting* untuk dokumen website lebih baik jika menggunakan jenis website yang telah terstuktur memiliki bentuk umum secara standart.
3. Implementasi teknik deteksi noise pada halaman website dapat dikembangkan dengan menggunakan metode lain yang diharapkan dapat meningkatkan performansi hasil klasifikasi.

Telkom University

6. Daftar Pustaka

[1]	Bar-Yossef, Z. and Rajagopalan, S. Template Detection via Data Mining and its Applications, WWW 2002, 2002. http://www.cs.uic.edu/%7Eliub/templatedetection.pdf
[2]	Beeferman, D., Berger, A. and Lafferty, J. *Statistical models for text segmentation*. Machine learning, 34(1-3), 1999. Diakses tgl 14 Desember 2007 http://www.w3.org/TR/html4/ml.pdf
[3]	Deepayan Chakrabarti, Ravi Kumar, Kunal Punera. *Page-level Template Detection via Isotonic Smoothing*. Yahoo! Research 701 First Ave Sunnyvale, CA 94089. Diakses tgl 24 Desember 2007 http://citeseer.ist.psu.edu/CIKM630-vieira.pdf
[4]	Karane Vieira, Altigran S. da Silva, Nick Pinto. *A Fast and Robust Method for Web Page Template Detection and Removal.* Universidade Federal do Amazonas Departamento de Ci ˆencia da Computac¸ ˜ao Manaus, AM, Brazil. Diakses Tanggal 1 Januari 2008 http://www.w3.org/TR/xhtml1/CIKM630-vieira.pdf
[5]	Lan Yi, Bing Liu. "*Web Page Cleaning for Web Mining through Feature Weighting*" To appear in *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence* (IJCAI-03), Aug 9-15, 2003, Acapulco, Mexico. Diakses tgl 14 Desember 2007 http://www.cs.uic.edu/%7Eliub/publications/ijcai03-webClean.pdf
[6]	Lan Yi, Bing Liu, Xiaoli Li. "Eliminating Noisy Information in Web Pages for Data Mining." To appear *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (KDD-2003), Washington, DC, USA, August 24 - 27, 2003. Diakses tgl 14 Desember 2007 http://www.cs.uic.edu/%7Eliub/kdd2003-webNoise.pdf
[7]	Ling Ma, Nazli Goharian, Abdur Chowdhury. *Extracting Unstructured Data from Template Generated Web Documents*. Information Retrieval Laboratory Computer Science Department Illinois Institute of Technology Diakses Tgl 1 Januari 2008. http://citeseer.ist.psu.edu/p121-ma.pdf
[8]	Ruihua Song, Haifeng Liu1*, Ji-Rong Wen, Wei-Ying Ma. *Learning Block Importance Models for Web Pages*. Microsoft Research Asia, 49 Zhichun Road, Beijing, 100080, P.R. China, 1Department of Computer Science, University of Toronto, Toronto, ON, Canada http://citeseer.ist.psu.edu/30.pdf
[9]	Shian-Hua Lin and Jan-Ming Ho. *Discovering Informative Content Blocks from Web Documents*, KDD-02, 2002. Diakses Tgl 14 Desember 2007 http://citeseer.ist.psu.edu/lin02discovering.pdf
[10]	Shiren Ye and Tat-Seng Chua. *Learning Object Model from Product Web Pages*. School of Computing, National University of Singapore, Singapore 117543. Diakses Tgl 16 Desember 2007

	http://citeseer.ist.psu.edu/sigir04-semWeb-yeshiren.pdf
[11]	Tsukada, Makoto. Washio, Takashi. and Motoda, Hiroshi. *Automatic Web-Page Classification by Using Machine Learning Methods*. Institute of Scientific and Industrial Research Osaka University. Japan. Diakses Tgl 16 Desember 2007 http://www.ar.sanken.osaka-u.ac.jp/papers/2006-12/wi01_tsukada.pdf
[12]	Yanhong Zhai, Bing Liu. *Web Data Extraction Based on Partial Tree Alignment*. Department of Computer Science University of Illinois at Chicago 851 South Morgan Street, Chicago, IL 60607-7053. http://citeseer.ist.psu.edu/www05-p516.pdf
[13]	Yanuar Firdaus, Vector Similarity Computation with Weights. CS4323/0708-2 YFA
[14]	Yan Puspitaraani, Filter-Based Featurer Selection pada Kategorisasi Artikel Berita Berbahasa Indonesia. 2008. Datamining Center ITTelkom.
[15]	W3school Forum. HTML DOM Tutorial. http://www.w3schools.com/htmldom/default.asp