NEWS ITEMS EXTRACTION DENGAN MENGGUNAKAN PATTERN BASED STRATEGY

Ratih Retno Indri[1], Yanuar Firdaus A.w.[2], Zk. Abdurahman Baizal[3]

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

Abstrak

Seiring dengan semakin berkembangannya dunia internet, semakin banyak pula web berita online yang tersedia di dunia maya. Pada dasarnya web berita online banyak menyediakan informasi penting. Berbagai teknik dalam text mining dapat diterapkan dengan tujuan untuk memperoleh manfaat yang lebih banyak dari informasi yang disediakan, diantaranya yaitu dengan menggunakan page-based clustering maupun keyword-based search. Namun, page-page pada newspaper biasanya terdiri dari beragam item berita dengan topik yang saling tidak berhubungan satu sama lain, sehingga page-based clustering kurang memberikan hasil yang optimal.

Pada Tugas Akhir ini dilakukan pendekatan dengan melakukan ekstraksi terhadap item-item berita pada web pages secara individual dan melakukan mining secara terpisah dengan menggunakan pattern based strategy. Pendekatan ini menggunakan pattern URL text dan anchor text dalam mengekstrak link berita serta menggunakan crawler untuk penelusuran link berita dalam rangka mencari full story dari masing-masing item berita.

Tahap analisis dan pengujian memberikan hasil bahwa pendekatan pattern based strategy yang dibangun terbukti dapat mengekstrak full story pada halaman Web berita meskipun tidak semua full story dari setiap item berita dapat diekstrak. Hasil ekstraksi item berita akan mencapai nilai optimal jika link-link pada web input bersifat homogen.

Kata Kunci : pattern based strategy , URL text , anchor text, web berita, full story

Abstract

As the increasing of the development of the internet, the number of online news web available in the net is also increasing. Basically, the online news web provides important information. Various techniques in text mining can be applied to gain more advantages from the available information, such as using page-based clustering or even keyword based search. But, pages on newspaper, usually consist of many kinds of news item with unrelated topic each other, so the page-based clustering gives an optimal result deficiently

In this final assignment, an approximation is applied by extracting the news items on web pages individually and by mining it separately using pattern based strategy. This approximation is using the pattern URL text and anchor text in order to extract news link and also using crawler for news link search to seek the full story from each news item.

The analysis and testing phase results that the pattern based strategy approach build, proved can extract the full story of the news web page although not all full stories from each news item is extractable. The extracted news item will reach the optimal value if the links on the input web is homogeny

Keywords : pattern based strategy , URL text , anchor text, newspaper web, full story

BAB I
PENDAHULUAN

1.1. Latar Belakang

Pada dasarnya *web newspaper* menyediakan banyak informasi penting. Berbagai teknik dalam *text mining* dapat diterapkan dengan tujuan untuk memperoleh manfaat yang lebih banyak dari informasi yang disediakan, diantaranya yaitu dengan menggunakan *page-based clustering* maupun *keyword-based search*. Namun, *page-page* pada *newspaper* biasanya terdiri dari beragam *item* berita dengan topik yang saling tidak berhubungan satu sama lain, sehingga *page-based clustering* kurang memberikan hasil yang optimal.

Dalam rangka pengembangan *complete-page mining* maka dapat dilakukan suatu pendekatan dengan melakukan *extracting* terhadap *item-item* berita *web pages* secara individual dan melakukan mining secara terpisah. Pendekatan ini dimungkinkan dapat meningkatkan kualitas dari hasil yang diperoleh. Pendekatan ini juga memiliki keuntungan, dimana *item-item* berita pada *entry page newspaper* menyediakan versi hasil kompresi dari *full story*-nya, sehingga dengan hanya melakukan mining pada main *page website*, dapat mengurangi jumlah data yang harus diambil (mining).

Secara visual, manusia dapat dengan mudah membedakan *item-item* berita pada web *page*, tetapi tidak demikian pada komputer. Pada tugas akhir ini diterapkan suatu strategi berupa pendeteksian *item* berita dengan menggunakan pola-pola yang sering muncul pada *newspaper web pages* (*pattern-based news item extraction*). Pola-pola tersebut diantaranya adalah: *URL-text-URL item, line item, anchor text item, bold header item, and text-based item* [1].

Untuk melihat kualitas dari pendekatan *pattern-based news item extraction*, dapat dilakukan riset dengan membandingkan hasil dari *tools items extraction* terhadap hasil dari inspeksi secara manual dengan menggunakan sejumlah *web pages* atau dengan melakukan perbandingan terhadap subset dari *web pages*.

1.2. Perumusan Masalah

Permasalahan yang dijadikan objek penelitian dalam tugas akhir ini antara lain :

1. Bagaimana menerapkan strategi *pattern based* dalam mengekstraksi *item* berita?
2. Bagaimana mengukur kualitas ektraksi berita dengan menggunakan parameter pengukuran berupa *precision, recall, f measure.*

1.3. Tujuan Pembahasan

Dalam tugas akhir ini, hal-hal yang diharapkan untuk dicapai adalah sebagai berikut :

1. Mengimplementasikan *pattern based strategy* untuk mengekstraksi *item-item* berita pada *web newspaper*.
2. Mengukur kualitas ekstraksi ditinjau dari jumlah *item* berita hasil ekstraksi yang relevan

1.4. Batasan Masalah

Untuk menghindari meluasnya materi pembahasan tugas akhir ini, maka penulis membatasi permasalahan dalam tugas akhir ini hanya mencakup hal-hal berikut :

1. Web *newspaper* yang digunakan sebagai inputan berupa web *offline*.
2. Inputan berupa *web newspaper* berbahasa Indonesia.
3. Menggunakan deskripsi singkat dari berita yang terdapat pada main *page* sebagai basis dari proses *mining*.
4. Output berupa *item-item* berita.
5. Sistem operasi yang digunakan adalah windows Xp .

1.5. Metodologi Penyelesaian Masalah

Metode yang akan digunakan untuk menyelesaikan tugas akhir ini adalah :

1. Studi literatur
   Berupa pencarian sumber-sumber bacaan yang dapat menunjang topik tugas akhir ini. Sumber-sumber bacaan tersebut penulis letakkan pada daftar pustaka. Sumber bacaan berupa *e-book*, jurnal – jurnal yang diperoleh dari internet.
2. Pengumpulan data-data penunjang tugas akhir
   Berupa pengumpulan data penunjang yang dapat membantu perancangan sistem. Data berupa *source code* yang bersifat *open source*, manual pemrograman, maupun data-data lain yang membantu terselesainya tugas akhir ini.
3. Analisis dan perancangan sistem
   Berupa perancangan sistem dari studi pustaka dan data-data penunjang, serta analisis yang dikembangkan.
4. Imlementasi sistem
   Berupa pembangunan perangkat lunak yang mampu mengimplementasikan *pattern based strategy* untuk mengekstraksi *item-item* berita pada *web newspaper*.
5. *Testing* dan analisis hasil
   Berupa pengujian terhadap perangkat lunak yang dibangun sekaligus melakukan analisa terhadap *output* yang dihasilkan oleh perangkat lunak.
6. Penulisan dokumentasi dan laporan
   Berupa proses penulisan dokumentasi dan laporan tugas akhir seperti disyaratkan oleh Departemen Teknik Informatika, Institut Teknologi Telkom.

1.6. Sistematika Penulisan

BAB I PENDAHULUAN
Berisi latar belakang, perumusan masalah, batasan masalah, tujuan pembahasan, metodologi penyelesaian masalah dan sistematika penulisan.

BAB II LANDASAN TEORI
Berisi penjelasan singkat mengenai konsep-konsep yang mendukung dikembangkannya sistem ini.

BAB III DESAIN DAN IMPLEMENTASI

Berisi rincian mengenai desain sistem serta implementasi sistem yang dibuat.

BAB IV PENGUJIAN DAN ANALISA SISTEM

Berisi rincian mengenai pengujian yang dilakukan terhadap sistem yang dikembangkan, disertai analisis terhadap hasil pengujian.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan yang diambil berkaitan dengan sistem yang dikembangkan, serta saran-saran untuk pengembangan lebih lanjut.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan pengujian dan analisa yang telah dilakukan pada bab 4, diperoleh kesimpulan sebagai berikut :

1. Nilai jumlah berita *actual* pada data input tidak mempengaruhi nilai performansi sistem.
2. Performansi pada proses ekstraksi *full story* dipengaruhi oleh nilai batas atas dan batas atas yang digunakan sebagai pengenal *full story* karena nilai batas atas dan batas bawah adalah unik pada masing-masing web uji. Semakin homogen *link-link* berita yang terdapat pada halaman web berita maka akan semakin banyak *full story* yang terekstrak.
3. *Pattern based* yang paling banyak digunakan pada web uji adalah *anchor text* dengan persentase jumlah penggunaan pattern *anchor-text* pada web uji sebesar 90%, sedangkan penggunaan pattern *URL- text* pada *pattern* uji sebesar 10%.

5.2 Saran

1. Sebaiknya dapat diimplementasikan sebuah sistem yang mampu secara otomatis mengekstrak isi berita tanpa melakukan set batas atas dan batas bawah terlebih dahulu.
2. Dapat dilakukan pengujian kualitas hasil ekstraksi untuk dijadikan input aplikasi *text* mining lainnya seperti *tools clustering*, maupun peringkasan dokumen otomatis.
3. Dapat diimplemtasikan pendekatan pattern based strategy dengan menggunakan pattern lainnya seperti *bold header item, text based item,* dan *line item.*

DAFTAR PUSTAKA

[1] Kjetil Nørv°ag_ and Randi Øyri. News Item Extraction for Text Mining in Web Newspapers. *Technical Report IDI-TR-11/2004*. Dikutip: 12 Desember 2008.

[2] D. Cai, S. Yu, J. Wen, and W. Ma. Extracting content structure for web pages based on visual representation. In Y. Z. X. Zhou and M. Orlowska, editors, *Web Technologies and Applications: 5th Asia-Pacific Web Conference, APWeb 2003, Xian, China, April 23-25, 2003*, volume 2642 of *Lecture Notes in Computer Science*, pages 406–417. Springer-Verlag Heidelberg, January 2003. Dikutip: 11 Desember 2008.

[3]. Data mining. Available at : http://one.indoskripsi.com/node/10145.html. Diakses tanggal 30 Juli 2009.

[4] C. Chang, C. Hsu, and S. Lui. Automatic information extraction from semi-structured Web pages by pattern discovery. *Decision Support Systems*, 35(1):129 – 147, April 2003.Dikutip: 11 Desember 2008.

[5] I. S. Dhillon, J. Fan, and Y. Guan. Efficient clustering of very large document collections. In V. K. R. Grossman, C. Kamath and R. Namburu, editors, *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, 2001. Invited Book Chapter. Dikutip: 11 Desember 2008.

[6] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999. Dikutip: 11 Desember 2008.

[7] Ricardo Baeza-Yates, and Berthier Riberio-Neto, *Modern Information Retrieval*, ACM press, 1999

[8] Un Yong Nahm. Text Mining with Information Extraction. Dikutip: 7 Desember 2008.

[9] Lan Yi[1], Bing Liu [2] and Xiaoli Li [3]. Eliminating Noisy Information in Web Pages for Data Mining. 1. School of Computing National University of Singapore 3 Science Drive 2, Singapore. 2. Department of Computer Science University of Illinois at Chicago.3. Singapore-MIT alliance National University of Singapore 3 Science Drive 2 Singapore. Dikutip: 11 Desember 2008.

[10] Bing Liu and Xiaoli Li . OLERA: A Semi-supervised Approach for Web Data Extraction with Visual Support. Dept. of Computer Science and Information Engineering National Central University, Chung-Li 320, Taiwan. Dikutip: 22 Desember 2008.

[11] Lakshmish Ramaswamy[1], Arun Iyengar[2] , Ling Liu[1] and Fred Douglis[2]. Techniques for Efficient Fragment Detection in Web Pages. 1. College of Computing, Georgia Tech _ IBM T.J. Watson Research Center 801 Atlantic Drive, Atlanta. 2. IBM T.J. Watson Research Center. Dikutip: 22 Desember 2008.

[12] Ron Kohavi and Dan Sommerfield. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. *Appears in the First International Conference on Knowledge Discovery and Data Mining (KDD-95)* . Computer Science Department, Stanford University. Dikutip: 22 Desember 2008.