ANALISIS DAN IMPLEMENTASI ALGORITMA LABEL DISCOVERY UNTUK MENCARI STRUKTUR DAN KARAKTERISTIK HALAMAN WEB (STUDI KASUS HALAMAN WEB LOWONGAN KERJA)

Difan Mustafa[1], Ade Romadhony[2], Yanuar Firdaus A.w.[3]

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

Abstrak

Pada World Wide Web atau yang biasa dikenal dengan WWW memiliki berbagai jenis informasi yang terkandung didalamnya. Biasanya user akan menggunakan search engine atau mengikuti link terkait untuk menemukan informasi yang mereka butuhkan. Akan tetapi, pencarian dengan menggunakan search engine terkadang tidak efektif karena menghasilkan banyak data dan link terkait yang membutuhkan waktu tidak sedikit untuk menulusurinya satu persatu, bahkan kadangkala hasil yang keluar sama sekali tidak terkait dengan keyword yang user masukkan. Setelah diteliti ternyata halaman web yang memiliki informasi yang sama memiliki struktur yang sama pula, selain itu diketahui pula bahwa standar dalam menerbitkan halaman web yang tidak ketat menyebabkan si penulis dapat menggunakan kata atau pelabelan yang berbeda untuk halaman web yang memiliki informasi yang sama. Pada tugas akhir ini akan dibangun sebuah sistem yang dapat mengklasifikasi halaman web berdasarkan kelasnya dengan menggunakan algoritma label discovery. Pada awalnya algoritma label discovery (LDA) akan mencari label atau kata yang merepresentasikan kelas halaman web sehingga akan didapatkan struktur dari kelas halaman web yang diinginkan. Setelah struktur ditemukan maka akan digunakan untuk mengklasifikasikan halaman web. Hasil pengujian menunjukkan bahwa LDA dapat menemukan kata atau label yang merepresentasikan suatu kelas dan struktur yang dihasilkan dapat mengklasifikasi halaman web secara akurat.

Kata Kunci : label discovery, struktur, klasifikasi, kelas, halaman web, fungsi kesamaan

Abstract

The World Wide Web (WWW) provides vast resource for information of almost all types. Users commonly use search engine or follow related links to retrieve the information. However, searching information using search engine is not effective because it will provide tedious data and so many related links which are wasting time to read it one by one, even sometimes the result is not related at all to what have user entered. After experiment, it's discovered that web pages that have same information will have same structure too, moreover due to loose standard of web page publishing, different authors can use different wordings (labels) which describe the same information. This thesis builds a system that can classify web pages by class using label discovery algorithm (LDA). First, LDA will find labels or words that represent class of web pages, so that would be obtained the structure of class, finally the structure will be used for classifying web pages. Testing results show that LDA can be used for finding labels or words that represent class of web pages and the structure obtained by this method can classify web pages accurately.

Keywords : label discovery, structure, classification, class, web pages, similarity function

1. Pendahuluan

1.1 Latar Belakang

World Wide Web atau yang biasa dikenal dengan WWW menyediakan berbagai informasi dari segala tipe. *User* pada umumnya mengakses data-data ini dengan menggunakan *search engine*, akan tetapi pencarian dengan menggunakan *search engine* kadang-kadang tidak efektif karena menghasilkan banyak data dan *link* terkait yang membutuhkan waktu yang tidak sedikit untuk menulusurinya satu persatu, bahkan terkadang hasil yang keluar sama sekali tidak terkait dengan *keyword* yang *user* masukkan.

Setelah ditinjau lebih jauh, beberapa halaman web yang dicari dengan informasi yang sama memiliki struktur yang sama pula [8]. Salah satu contohnya adalah halaman web lowongan kerja, namun standard mem-*publish* halaman web yang tidak ketat, sebagai contoh, *publisher* dapat menggunakan kata yang berbeda dalam setiap menampilkan informasi yang sama sebagai contoh, pada situs lowongan kerja ada kata '*skill*' atau '*abilities*' yang sama-sama memberikan informasi bahwa seorang pelamar harus mempunyai kemampuan sesuai dengan kriteria yang diinginkan perusahaan. Hal ini menyulitkan proses ekstraksi informasi secara otomatis, untuk itu diperlukan suatu algoritma yang dapat manggabungkan kata-kata yang memiliki persamaan makna tersebut.

Algoritma yang akan digunakan pada tugas akhir ini adalah algorima *label discovery* (LDA). Algoritma ini dapat menemukan label beserta *feature* data teks dari suatu kelas halaman web yang menjelaskan informasi yang sama dengan cara memanfaatkan struktur hirarki dari kelas halaman web tersebut.

Kelebihan dari LDA adalah mekanisme ektraksi tidak tergantung pada inputan *user* dalam mendeskripsikan struktur HTML dan dapat digunakan untuk membedakan halaman web

1.2 Perumusan Masalah

Perumusan masalah tugas akhir ini adalah:

1. Bagaimana cara menemukan struktur yang khas dari kelas halaman web lowongan kerja.
2. Bagaimana mengimplementasikan LDA untuk mencari label dan *feature* data teks pada kelas halaman web lowongan kerja.
3. Bagaimana membedakan struktur halaman web yang positif kelas halaman web lowongan kerja dan yang negatif.

Adapun batasan masalah tugas akhir ini adalah:

1. Menggunakan studi kasus halaman web lowongan kerja karena pada setiap halaman web lowongan kerja memiliki struktur yang hampir sama walaupun halaman web tersebut berada di situs yang berbeda.
2. Pada halaman web lowongan kerja tersebut hanya berisi satu informasi pekerjaan yang ditawarkan oleh suatu perusahaan bukan berupa *list*

lowongan kerja dan pada halaman web tersebut tidak mengandung informasi lain seperi artikel, berita dan sebagainya.
3. Dataset berupa halaman web lowongan kerja yang berada di wilayah Indonesia.
4. Bahasa yang digunakan Bahasa Indonesia dan Bahasa Inggris.
5. Proses persiapan data berupa ekstraksi halaman web menggunakan *tools* yang sudah ada sehingga tidak dibahas pada tugas akhir ini.

1.3 Tujuan

Tujuan dari tugas akhir ini adalah:
1. Menemukan struktur yang khas dari kelas halaman web lowongan kerja.
2. Mengimplementasikan LDA untuk mencari label dan *feature* data teks pada kelas halaman web lowongan kerja.
3. Membedakan struktur halaman web yang positif kelas halaman web lowongan kerja dan yang negatif.

Hipotesis awal tugas akhir ini:
1. Algoritma *Label Discovery* dapat digunakan untuk mencari label dan *feature* data teks halaman web lowongan kerja yang menggunakan Bahasa Indonesia dan Inggris
2. Algoritma *Label Discovery* dapat digunakan untuk mengklasifikasi halaman web yang positif dan negatif lowongan kerja secara akurat.

1.4 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang digunakan dalam melakukan tugas akhir ini adalah:

1. Studi Literatur

   Tahap ini merupakan proses mendalami materi melalui studi pustaka dan mencari referensi dari berbagai sumber seperti buku, jurnal, dan sumber lainnya tentang LDA , *Tag-Tree*, struktur hirarki, dan HTML.
2. Pengumpulan Data

   Melakukan pengumpulan data berupa halaman web lowongan kerja berupa HTML yang digunakan sebagai *data collection*. Situs lowongan kerja didapat menggunakan *search engine* Google, beberapa *website* yang digunakan antara lain: loker.web.id, jobstreet.co.id, jobloker.co.id, jobmedan.com, id.jobsdb.com, dan karir.com.
3. Analisis dan Perancangan Sistem

   Pada tahap ini dilakukan analisis terhadap system dan algoritma yang digunakan. Sedangkan perancangan system digambarkan dengan menggunakan *flow chart.*
4. Implementasi

   Tahap implementasi perangkat lunak dilakukan dengan menggunakan bahasa pemrograman yang sesuai dengan spesifikasi dan perancangan yang telah ditentukan sebelumnya.
5. Pengujian dan Analisis Hasil

Pada tahap ini dilakukan pengujian terhadap performansi dan akurasi algoritma *label discovery* yang digunakan untuk mendapatkan label-label dari kelas halaman web yang digunakan sebagai *data collection*. Kemudian dilakukan analisis terhadap hasil dari algoritma tersebut dalam mengklasifikasikan halaman web.

6. Dokumentasi

   Tahap akhir dari penelitian adalah penyusunan dokumentasi. Dokumentasi ditulis dalam bentuk buku Tugas Akhir dan berisi dasar teori, tahapan proses penelitian, serta hasil penelitian.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berikut adalah kesimpulan yang dapat diambil dari pengujian dan analisis di atas:

1. Sistem dapat mengklasifikasi halaman web dengan baik saat *minimum support*= 0.1 , *minimum similarity*=0.6, ε_n =5 dan ε_m =0.5.
2. Struktur dan kemiripan kata pada setiap halaman web mempengaruhi akurasi sistem dalam mengklasifikasi halaman web.
3. Perbedaan bahasa tidak mempengaruhi akurasi sistem.
4. Halaman web yang positif dan yang negatif lowongan kerja dapat dibedakan dengan cara mengidentifikasi masing-masing node pada setiap halaman web yang ingin diuji, jika jumlah node yang sama atau mirip dengan label dan *feature* data teks cukup (lebih besar sama dengan *threshold* ε_n) maka halaman web tersebut dikatakan positif halaman web lowongan kerja.
5. Untuk diklasifikasikan sebagai halaman web positif lowongan kerja, halaman tersebut minimal memiliki 3 unsur utama yaitu: *job name*, *job requirement*, dan *job description*.
6. Struktur yang khas dari kelas halaman web lowongan kerja dapat ditentukan dari label dan *feature* data teks yang ditemukan pada proses data *training*. Berikut merupakan label dan *feature* data teks yang ditemukan pada 567 data *training*, dapat dilihat pada tabel 5.1 dan tabel 5.2

Tabel 5.1 Label Kelas Halaman Web Lowongan Kerja

Informasi lowongan
Sumatera utara
Persyaratan
Gaji ditawarkan
nego
karir
Fungsi kerja
Lokasi kerja
Jenjang pendidikan
Pengalaman kerja
Tanggal pemasangan
Tanggal penutupan
Melamar pekerjaan klik tombol lamar
Situs lowongan kerja indonesia
Detil pekerjaan
Deskripsi pekerjaan

Tabel 5.2 *Feature* Data Teks Kelas Halaman Web Lowongan Kerja

Teks	Confidence	Teks	Confidence	Teks	Confidence
persyaratan	0.2178	knowledge	0.0478	skills	0.0690
informasi	0.0925	perusahaan	0.2259	pria	0.0998
wanita	0.1303	S1	0.186	ekonomi	0.0435
memenuhi	0.0180	kualifikasi	0.0294	kirim	0.0456
lamaran	0.1313	cv	0.1379	lambat	0.0812
minggu	0.0036	iklan	0.0849	tayang	0.0016
pt	0.0973	fastrata	0.0016	buana	0.0016
talent	0.0033	development	0.0348	department	0.0096
email	0.2089	recruitment	0.0596	barat	0.0233
fastrabuana	0.0032	id	0.1762	male	0.0690
female	0.0673	max	0.0898	years	0.1505
min	0.2288	degree	0.0450	mechanical	0.0063
electrical	0.0121	industrial	0.0069	engineering	0.0137
proficient	0.0044	english	0.0726	spoken	0.0094
written	0.0499	computer	0.0379	application	0.0409
literacy	0.0016	interpersonal	0.0319	presentation	0.0173
skill	0.0502	analytical	0.0117	thinker	0.0016
good	0.1457	understanding	0.0039	products	0.0401
pumps	0.0016	valves	0.0016	hydraulic	0.0016
equipments	0.0032	palm	0.0016	oil	0.0022
mill	0.0045	electric	0.0016	motors	0.0016
drivers	0.0016	transmission	0.0016	posses	0.0127
communication	0.0643	problem	0.0068	solving	0.0053
follow	0.0051	actions	0.0016	decisions	0.0016
sense	0.0048	urgency	0.0022	punctuality	0.0016
ready	0.0030	work	0.0987	pressure	0.0089
meet	0.2284	deadlines	0.0037	targets	0.0054
sudden	0.0016	environment	0.0076	experience	0.0984
pump	0.0016	automotive	0.0028	industry	0.0081
job	0.2042	order	0.0050	depth	0.0044
production	0.0145	planning	0.0090	material	0.0032
management	0.0302	commitment	0.0048	ability	0.0099
independently	0.0069	driven	0.0034	provide	0.0180
business	0.0476	support	0.0249	partner	0.0030
operations	0.0153	change	0.0030	agent	0.0029
status	0.0059	process	0.0125	improvements	0.0016
located	0.0048	jakarta	0.1111	kalimantan	0.0016
surabaya	0.0257	palembang	0.0024	pekanbaru	0.0016
medan	0.0115	competitive	0.0029	fringe	0.0016
candidate	0.0157	salary	0.0279	transport	0.0117
allowance	0.0064	sales	0.0681	commision	0.0064
health	0.0054	insurance	0.0041	family	0.0016
match	0.0016	requirement	0.0320	send	0.0726
complete	0.0051	previous	0.0024	current	0.0084

***Feature* data teks hingga 3246 kata**

5.2 Saran

Berikut ini merupakan saran-saran yang perlu dipertimbangkan untuk pengembangan lebih lanjut:

1. Dapat menggunakan dataset lowongan kerja yang berupa *list* tidak hanya terpaku pada satu halaman satu informasi lowongan kerja tetapi berupa kumpulan informasi lowongan kerja.
2. Dapat diaplikasikan pada dataset dengan format XML. Data dengan format XML lebih terstruktur dari HTML sehingga jika label dapat ditelusuri secara akurat maka kesalahan atau error dalam mendapatkan label yang mirip dapat berkurang.

6. Referensi

[1] Abiteboul, S., Motwan, R., dan Nestrov, S. 1998. Extracting Schema from Semistructured Data. In Proceedings of the ACM SIGMOD International Conference, Washington, USA. 295-306.

[2] Agrawal, R. dan Srikant, R. 1994. *Fast Algorithms for Mining Association Rules*. In Proceedings of VLDB. 487-499.

[3] Agrawal, R., Imielinski, T., dan Swami, A. 1993. *Mining Association Rules Between Sets of Items in Large Databases*. In Proceedings of the ACM SIGMOD International Conference, Washington,DC. 207-216.

[4] Buneman, P., *et al*. 1996. A Query Language and Optimization Techniques for Unstructured Data. In Proceedings of the ACM SIGMOD International Conference, Montreak, Canada. 505-516.

[5] Chawathe, S., *et al*. 1994. *The TSIMMIS Project: Integration of Heterogeous Information Sources*. In Proceedings of Tenth Anniversary Meeting of the Information Processing Society of Japan, Tokyo, Japan. 7-18.

[6] Embley, D.W., *et al*. 1998. *A Conceptual-modeling Approach to Extracting Data from the Web*. In Proceedings of the 17th International Conference on Conceptual Modeling, Singapore. 78-91.

[7] Embley, D. W., Jiang, Y., dan Ng, Y. K. 1999. *Record-boundry Discovery in Web Documents*. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, Philadelphia, USA. 467-478.

[8] Fu, A.W.C., dan Wong, W.C. 2000. *Finding Structure and Characteristics of Web Documents for Classification*. In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 96-105.

[9] Isahara, H., Sornlertlamvanich, V. Dan Tongchim, S. 2006. Classification of News Web Documents Based on Structural Features. Advances in Natural Language Processing. 153-160.

[10] Liu, Huiqing dan Wang, Ke. 1998. *Discovering Typical Structures of Documents: A Road Map Approach*. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Retrieval, Melbourne, Australia. 146-154.

[11] Quass, D., *et al*. 1995. *Querying Semistructured Heterogeneous Information*. In Deductive and Object-Oriented Databases (DOOD),

Singapore. 319-344.

[12] Wikipedia. 2012. *HyperText Markup Language*. [online]. Tersedia http://en.wikipedia.org/wiki/HTML.