ANALISIS DAN IMPLEMENTASI ASSOCIATIVE NETWORK DALAM PEMBENTUKAN METADATA SECARA OTOMATIS PADA DOKUMEN WEB
ANALYSIS AND IMPLEMENTATION OF ASSOCIATIVE NETWORKS IN THE GENERATION OF AUTOMATIC METADATA IN WEB DOCUMENTS

Darmawan Fatriananda[1], Angelina Prima Kurniati[2], Bayu Erfianto[3]

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

Abstrak

Metadata merupakan istilah dari proses identifikasi suatu atribut dan struktur dari sebuah data atau informasi. Metadata ini sebagai data yang menjelaskan sebuah data itu sendiri. Dalam dokumen web, metadata ini membantu dalam menentukan dokumen web yang relevan. Pembentukan metadata secara otomatis dapat membantu penemuan informasi yang relevan. Dalam Tugas Akhir ini dibangun aplikasi untuk membentuk metadata secara otomatis dengan menggunakan metode associative network. Metode associative network digunakan untuk menentukan keterhubungan antar dokumen. Terdapat dua algoritma yang digunakan untuk membuat associative network, yaitu occurrence associative network dan coocurrence associative network. Setelah ditentukan dokumen yang memiliki keterhubungan, kemudian dilakukan propagasi metadata dengan algoritma particle swarm untuk proses shared metadata dari dokumen web yang memiliki metadat a lengkap ke dokumen web yang metadatanya kurang atau tidak lengkap. Hasil dari rekomendasi metadata dapat dilakukan proses filter energi untuk menambah keakuratan dari hasil rekomendasi. Dengan metode associative network, tidak semua dokumen berhasil mend apat rekomendasi metadata. Hasil pembentukan metadata secara otomatis bisa dipilih berdasar ketepatan atau banyaknya properti metadata yang dihasilkan dengan cara penyaringan energi terhadap metadata yang direkomendasikan. Sedangkan untuk membentuk metadat a secara otomatis dengan rekomendasi properti metadata keyword hasilnya lebih akurat jika menggunakan associative network berdasar cooccurrence.

Kata Kunci : metadata, associative network, particle swarm, dan pembentukan metadata otomatis

Abstract

Metadata is the term of the process of identifying an attribute and structure of a data or information. Metadata as data that describes a data itself . In a web document, the metadata is helpful in determining the relevant web documents. Metadata generation can help finding relevant information. In this Final Project built an application to create metadata automatically by using the method of associative network. The associative network method used to determine connectivity between documents. There are two algor ithms used to create associative network, the occurrence and coocurrence associative network. Having determined that the documents have connectivity, then performed the metadata propagation with particle swarm algorithm for metadata shared processes from web documents that have a complete metadata to web documents whose metadata lacking or incomplete. The results of the metadata recommendation process can be done to increase the accuracy of the energy filter of the metadata recommendations. With the method of associative networks, not all document managed to get a metadata recommendation. The result of the establishment of metadata can be automatically chosen based on accuracy or the number of metadata properties generated by the application of energy filter ing of the metadata recommended. Meanwhile, to create the metadata automatically with metadata keywords property recommendations more accurate results when using the associative network based on cooccurrence.

Keywords : metadata, associative network, part icle swarm, and metadata generation

1. Pendahuluan

1.1 Latar Belakang

Seiring dengan berkembangnya Teknologi Informasi dan ilmu pengetahuan, maka informasi yang terdapat di Internet pun semakin banyak dan luas. Dengan banyaknya informasi yang tersedia di Internet, informasi apapun yang dibutuhkan bisa dicari di Internet. Dengan bantuan *search engine*, informasi yang dibutuhkan akan didapat, hanya dengan cara mengetikan suatu *keyword*. Namun seringkali informasi yang didapat tidak relevan dengan apa yang dicari.

Informasi yang didapat dari Internet biasanya berupa dokumen web yang memiliki elemen-elemen teks yang dapat diproses lebih lanjut untuk membantu penemuan dokumen yang relevan. Suatu dokumen web memiliki metadata berupa teks yang terstruktur, sedangkan di media lain metadata bisa berupa suara atau gambar. Dalam dokumen web, metadata ini membantu dalam menentukan dokumen web yang relevan. Meskipun memiliki nilai yang sangat penting, metadata pada umumnya diabaikan dan kalaupun ada seringkali isinya tidak lengkap [4], karena menemukan, membuat atau memelihara metadata pada suatu halaman web secara manual ini terbilang sulit [3]. Hal ini tentu saja menghambat efektivitas pelayanan atau pemberian informasi yang relevan.

Pembentukan metadata secara otomatis dapat membantu penemuan informasi yang relevan. Selain itu bagi pembuat dokumen web, pembentukan metadata secara otomatis dalam suatu dokumen web ini menjadi lebih efisien dan lebih konsisten daripada proses manual (yang berorientasi pada manusia) [5]. Penelitian menunjukkan bahwa pembentukan metadata secara otomatis dapat menghasilkan metadata yang dapat diterima. Namun, pembentukan metadata ini tetap memerlukan pertimbangan intelektual manusia. Para peneliti telah menyimpulkan bahwa pembentukan metadata yang paling efektif adalah mempertimbangkan penilaian manusia maupun metode otomatis [6].

Teknik yang dilakukan dalam pembentukan metadata secara otomatis ini adalah dengan melakukan ekstraksi metadata dan isi dari dokumen. Teknik ekstraksi yang berdasarkan isi*(content)* ini kurang efisien untuk jenis dokumen multimedia (audio, video dan gambar) [16]. Selain itu, proses ekstraksi yang berdasar pada isi*(content)* ini ternyata menghabiskan biaya komputasi yang sangat besar [11].

Dalam suatu situs di Internet tentu ada dokumen yang saling berhubungan, contohnya dokumen-dokumen yang terhubung melalui *hyperlink* (sintax *href* dalam HTML) pada suatu halaman web. Contoh lain, dalam situs detik.com keterhubungan suatu artikel dapat dilihat dari data tentang penulis artikel tersebut. Dokumen-dokumen web yang saling berhubungan ini dapat berbagi(*shared)* metadata dari dokumen web yang memiliki metadata lengkap dapat berbagi ke dokumen web yang memiliki metadata kurang atau tidak lengkap. Teknik berbagi(*shared)* metadata ini merupakan teknik ekstraksi dengan pendekatan ekstraksi berdasar pada konten yang dapat membantu mengurai biaya komputasi [13].

Dalam Tugas Akhir ini dibangun aplikasi untuk membentuk metadata secara otomatis dengan menggunakan metode *associative network*. Metode *associative network* digunakan untuk menentukan keterhubungan antar dokumen. Terdapat dua algoritma yang digunakan untuk membuat *associative network*, yaitu *occurrence associative network* dan *coocurrence associative network*. Algoritma *occurrence associative network* melakukan pendekatan pada kutipan (biasanya berupa hyperlink) yang ada pada dokumen web, sedangkan *coocurance associative network* melakukan pendekatan pada nilai properti-properti pada metadata seperti *author, keyword, identifier, description, publisher, format* dan tanggal *publish*. Setelah ditentukan dokumen yang memiliki keterhubungan, kemudian dilakukan propagasi metadata dengan algoritma *particle swarm* untuk proses *shared* (berbagi) metadata dari dokumen web yang memiliki metadata lengkap ke dokumen web yang metadatanya kurang atau tidak

lengkap[14]. Teknik propagasi metadata menggunakan algoritma *particle swarm* dipilih karena dalam proses komputasi algoritma *particle swarm* akan lebih unggul daripada algoritma sejenis seperti algoritma genetik ,karena pada algoritma *particle swarm* tidak ada proses evolusi (mutasi dan *crossover*) [15]. Proses selanjutnya adalah mekanisme *filter* (penyaringan) terhadap energi dari metadata yang direkomendasikan, proses ini bersifat *optional,* karena untuk kebutuhan tertentu proses *filter* bisa tidak dilakukan.

1.2 Perumusan Masalah

Perumusan masalah yang diambil dalam penelitian Tugas Akhir ini adalah bagaimana mengatasi kesulitan untuk membuat, menemukan atau melengkapi metadata secara manual. Pembentukan metadata secara otomatis dengan metode *associative network* menggunakan algoritma *occurrence* dan *cooccurrence* dipilih untuk membantu permasalahan yang ada. Selain proses ekstraksi dilakukan juga proses proses propagasi menggunakan algoritma *particle swarm*. Hasil dari pembentukan metadata secara otomatis ini harus menginformasikan isi dari dokumen. Keakuratan dari hasil pembentukan metadata ini yang kemudian dievaluasi.

1.3 Batasan Masalah

Implementasi Tugas Akhir ini dibatasi oleh beberapa hal, sebagai berikut :

1. Data yang digunakan untuk membentuk metadata bentuk fisiknya adalah halaman web.
2. *Resource* metadata menggunakan data yang berasal dari lingkungan *Knowledge Discover Labotary* berupa file xml yang didapat dari http://kdl.cs.umass.edu/proximity/proximity.html , kemudian di-*export* ke mysql sebagai database. Jumlah data tersebut sebanyak 4135 data. Untuk file fisik dokumen html dapat di unduh di http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/.
3. Properti-properti metadata yang digunakan dalam penerapan algoritma *coocurence associative network* adalah *keyword*. *Keyword* dipilih karena merupakan properti yang sering digunakan. Properti pada dataset yang digunakan merupakan properti yang didefinisikan oleh suatu institusi yang bernama *Dublin Core.*

1.4 Tujuan

Tujuan dari Tugas Akhir ini adalah

1. Membangun suatu aplikasi untuk membentuk metadata secara otomatis dengan metode *associative network* dengan algoritma *occurrence associative network* dan *cooccurrence associative network* untuk teknik ektraksi. Kemudian mengimplementasikan teknik propagasi metadata dengan menerapkan algoritma *particle swarm* yang berfungsi menyebarkan properti metadata .
2. Evaluasi performansi algoritma *occurrence associative network* dan *cooccurrence associative network* yang dijadikan sumber pada teknik propagasi metadata (menyebarkan properti metadata) dengan algoritma *particle swarm* menggunakan nilai *precision, recall* dan *F-score.*

1.5 Metode Penyelesaian Masalah

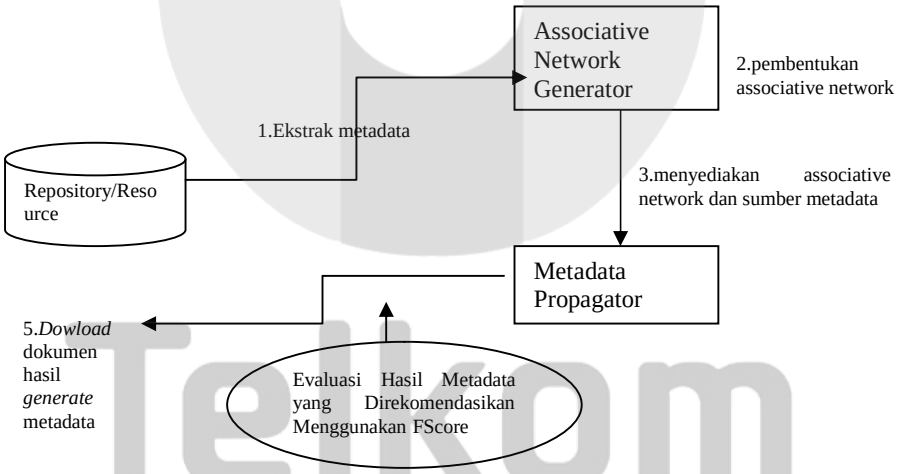
a. Studi literatur

Studi Literatur dengan mempelajari literatur-literatur yang relevan dengan permasalahan yang meliputi : melakukan studi pustaka dan referensi mengenai *automatic metadata generation, associative network, occurace associative network, coocurence associative network,* algoritma particle swarm, library Java untuk algoritma particle swarm (JSwarm), *Dublin Core metadata element set*.

b. Analisis Perancangan
Pada tahap ini dilakukan analisis, pengkajian dan pemahaman terhadap tahapan dan cara kerja *occurrence associative network, coocurrence associative network* serta algoritma *particle swarm* untuk *metadata generation*. Dari dua algoritma ekstraksi (*occurrence* dan *cooccurrence*) dianalisis mana yang menghasilkan rekomendasi metadata yang lebih ba ik.

c. Implementasi
Melakukan pembangunan aplikasi untuk implementasi metode yang digunakan pada sistem pembentukan metadata secara otomatis. Aplikasi yang dibuat bersifat web-based. Aplikasi dapat meng-*insert* metadata yang direkomendasikan ke dokumen web yang tidak memiliki atau tidak lengkap metadata.

d. Testing dan Analisa Hasil
Pada tahap testing dan analisis hasil ini Implementasi algoritma akan dilakukan berdasarkan hasil analisis dan perancangan algoritma pada tahap sebelumnya. Pengujian algoritma akan dilakukan dengan menggunakan input berupa koleksi dokumen yang akan di lakukan proses pembuatan metadata asecara otomatis kemudian dilakukan analisis hasil berdasar parameter yang telah ditentukan. Pengujian akan didasarkan pada 2 parameter, yaitu :
   a. Hasil dari pembentukan metadata secara otomatis ini dibandingkan dengan metadata asli yang ada di *resource*.
   b. Membentuk beberapa *associative network* menggunakan algoritma *occurrence* dan *cooccurrence* (properti metadata yang akan digunakan *author,date publish* dan *keyword*). Kemudian dari *associative network* yang tersebut dilakukan propagasi metadata dengan menyebarkan (*propagating*) properti metadata. Evaluasi dilakukan dengan menghitung *F-score*.



Gambar 1.1 Deskripsi Sistem

5 Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang dapat diperoleh dari Tugas Akhir ini adalah sebagai berikut :

1 Keberhasilan dalam membentuk metadata otomatis lebih banyak ketika menerapkan *Associative Network* berdasar *Cooccurrence* tanpa menerapkan penyaringan energi terhadap metadata yang direkomendasikan.

2 Hasil pembentukan metadata secara otomatis bisa dipilih dalam dua tipe, yaitu berdasar ketepatan dan banyaknya properti metadata yang dihasilkan. Caranya adalah dengan penerapan penyaringan energi terhadap metadata yang direkomendasikan. Penyaringan energi terhadap metadata yang direkomendasikan baik diterapkan jika ingin menghasilkan rekomendasi metadata yang lebih sedikit tapi lebih tepat, tetapi bila ingin menghasilkan rekomendasi metadata yang banyak maka jangan menerapkan penyaringan energi.

3 *Associative Network* berdasar *Cooccurrence* membentuk metatada secara otomatis lebih akurat dibanding berdasar *Occurrence*.

4 Sejumlah 50% dokumen yang di-*generate* dari *resource* tidak berhasil mendapat rekomendasi metadata.

5.2 Saran

Saran yang dapat diuraikan untuk keperluan analisis selanjutnya adalah sebagai berikut

1 Algoritma propagasi yang proses komputasi nya lebih cepat dibanding algorima *particle swarm*, karena untuk performansi proses penyebaran metadata menggunakan algoritma *particle swarm* terhitung lama.

2 Penggunaan dataset yang memiliki keragaman properti metadata .

Daftar Pustaka

[1] Devlin, B. (1997) Data Warehouse: from architecture to implementation, Addison - Wesley, Reading.
[2] Dublin Core Automatic Metadata Generation, http://www.ukoln.ac.uk/cgi-bin/dcdot.pl
[3] Duval, E., Hodgins, W., Sutton, S., and Weibel, S. L.2002.Metadata principles and practices. D-Lib Mag.
[4] Greenberg, J. 2004. Metadata extraction and harvesting: A comparison of two automatic meta- data generation applications. J. Intern. Catalog.
[5] Greenberg, J.,Spurgin, K., Crystal Abe 2005.AMeGA(Automat ic Metadata Generation Application) Project. UNC School Of Information And Library Science .
[6] Han, H. C., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E.A. (2003). Automatic document metadata extraction using support vector machines. In Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 37 – 48). New York: ACM Press.
[7] Haynes, D. (2004) Metadata for Information management an d retrieval, Facet, London.
[8] Hilmann, Diane, Dublin Core Element. http://dublincore.org/documents/usageguide/elements.shtml, diakses pada 14 juni 2010
[9] Jensen, David . 2004-2007. Proximity 4.3 Tutorial. Knowledge Discovery Laboratory.
[10] Jensen. D and Neville. J. 2002. Schemas and Models. *Proceedings of the Multi-Relational Data Mining Workshop, 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[11] Kuwano H., Matsuo, Y., and Kawazoe, K. 2004. Reducing the cost of metadata generation by using video/audio indexing and natural language processing tec hniques. NTT Technical Review.
[12] P. A. Boncz and M. L. Kersten. 1995. Monet: An Impressionist Sketch of an Advanced Database System. *Proceedings Basque International Workshop on Information Technology* .
[13] Rodriguez, M. A., Bollen, J., and Van De Sompel, H. 2009. Automatic metadata generation using associative networks. AC M Trans. Inform. Syst. 27,
[14] Rodriguez, M. A. 2007. Social decision making with multi -relational networks and grammar- based particle swarms. In Proceedings of the 40th Annual Hawaii International Conference on Systems Science (HICSS'07).
[15] Xihaou, H, Particle Swarm.http://www.swarmintelligence.org/tutorials.php, diakses pada 3 januari 2010.
[16] Yang, H.-C. and Lee, C.-H.2005.Automatic metadata generation for Web pages using a text mining approach. In Proce edings of the Internationalv Workshop on Challenges in Web Information Retrie val and Integration. IEEE,