AUTHORITATIVE DOKUMEN HALAMAN WEB MENGGUNAKAN ALGORITMA HITS

Fretty Herawati Manurung[1], Yanuar Firdaus A.w.[2], Warih Maharani[3]

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

Abstrak

Sulitnya menemukan dokumen yang relevan dengan query user merupakan satu masalah yang ditemukan dalam information retrieval. Penggunaan link based analysis dapat digunakan untuk membantu menyelesaikan masalah tersebut. Struktur link dapat dimanfaatkan karena dalam lingkungan struktur link, suatu dokumen halaman web akan merekomendasikan halaman lain yang memiliki informasi sama dengannya.

Terdapat beberapa algoritma link analysis yang dapat digunakan dalam pencarian dokumen yang relevan dan populer, salah satunya adalah algoritma HITS (Hypertext - Induced Topic Search) yang diimplementasikan dalam Tugas Akhir ini. HITS menggunakan root set sebagai data awal, root set tersebut akan diperluas menjadi base set dan dari base set akan digunakan dalam perhitungan nilai authority dan hub.

Dari pengujian yang dilakukan berdasarkan backlink metric, penggunaan HITS pada information retrieval memberikan pengaruh yang baik terhadap nilai IAP namun tidak demikian halnya pada nilai precision. Penambahan batasan dokumen crawling dan jumlah root set tidak selalu memberikan peningkatan nilai precision, karena dapat dipengaruhi oleh jumlah inlink dan jumlah outlink dokumen tersebut.

Kata Kunci : Link analysis, HITS, authority ,hub, crawler

Abstract

The difficulty of finding documents that relevant to user's queries is a problem in information retrieval. The use of link-based analysis can help to solve the problem. Link structure can be exploited because in the link structure environment, a web page document would recommend other pages that have the same information within.

There are several popular link analysis algorithms that can be used in the search for relevant documents, one of which is the HITS (Hypertext - Induced Topic Search) algorithm, implemented in this thesis. HITS uses the root set as the initial data, the root set will be extended to a base set and from the base set form, it will be used in calculating the authority value and hub.

From conducted test based on backlink metric, the use of HITS on information retrieval provides a better influence towards IAP but not precision. The addition of crawling document and number of root sets do not always provide improved precision value, because it can be influenced by the number of Inlink and outlink document.

Keywords : Link analysis, HITS, authority ,hub, crawler

1. Pendahuluan

1.1 Latar belakang

Dengan berkembangnya teknologi yang ada, maka berkembang pula kebutuhan akan informasi. Oleh sebab itu, banyak dokumen informasi yang tersedia. Banyaknya dokumen yang tersedia menimbulkan permasalahan yakni sulitnya menemukan dokumen informasi yang dibutuhkan secara tepat dan cepat. Dengan banyaknya jumlah dokumen informasi, dibutuhkan waktu yang lama untuk menemukan informasi-informasi yang relevan dengan kebutuhan *user* dari sekian banyak kumpulan dokumen. Oleh karena itu, pada Tugas Akhir ini penulis mencoba mencari dokumen yang relevan berdasarkan prinsip *link-based analysis.* Pada prinsip *link-based analysis*, dokumen yang memiliki topik yang sama dapat merekomendasikan dokumen lain begitu juga sebaliknya. Dalam *link-based analysis* dikenal istilah *authoritative* yang berarti suatu dokumen tidak hanya relevan tetapi juga populer [12]. Dengan *link based analysis*, *user* dapat memperoleh dokumen yang tidak hanya relevan dengan kebutuhannya namun juga populer atau dapat menemukan dokumen yang benar – benar dapat menjadi sumber yang informatif.

Pada Tugas Akhir ini akan dilakukan penelitian mengenai *Authoritative* Dokumen Halaman *Web* dengan menggunakan algoritma HITS (*Hyperlink-Induced Topic Search*). Algoritma HITS merupakan algoritma yang diciptakan untuk membantu menemukan dokumen yang *authoritative* dengan memanfaatkan struktur *link* dalam pencarian dokumen [1,7]. Seorang pencipta halaman *web* akan memberikan *link* lain pada halaman *web*-nya yang dapat memberikan informasi yang sama seperti yang ia miliki. Dalam Algoritma HITS akan dilakukan penghitungan nilai *authority* dan *hub* dalam proses pencariannya. *Authority* ialah dokumen yang memiliki topik serupa dengan topik yang *user* berikan. *Hub* berarti dokumen halaman *web* yang memiki *link* ke *authority* pages [8,11]. Perhitungan nilai *authority* dan *hub* ini akan dilakukan secara *iterative* hingga diperoleh nilai yang *konvergen*. Setelah diperoleh nilai yang *konvergen*, dokumen yang memiliki nilai *authority* yang tinggi merupakan dokumen yang *authoritative* dengan kebutuhan *user* [11].

Oleh sebab itu, dalam Tugas Akhir ini akan diteliti bagaimana pencarian *authoritative* dokumen dengan menggunakan algoritma HITS. Dengan Algoritma HITS, diharapkan dapat memberikan hasil yang berkualitas baik dan mengembalikan dokumen sesuai dengan kebutuhan *user*.

1.2Perumusan masalah

Masalah yang akan diselesaikan dalam Tugas Akhir ini adalah sebagai berikut.

1. Bagaimana efektivitas algoritma HITS dalam menemukan dokumen yang *authoritative* dilihat dari *precision* dan IAP?
2. Bagaimana pengaruh jumlah dokumen ter-*crawling* dan *root set* terhadap dokumen yang dihasilkan dari hasil pengujian yang dilakukan?

1.3 Batasan Masalah

Adapun batasan-batasan yang diberikan dalam penyelesaian masalah Tugas Akhir ini adalah sebagai berikut:

1. Penelitian pada tugas akhir ini hanya fokus pada satu algoritma saja yaitu algoritma HITS (*Hyperlink-Induced Topic Search*).
2. Dalam melakukan *crawling* (penelusuran link) bersifat *online.*
3. Tidak menangani *preprocessing* yaitu proses pencarian dengan *text based* untuk *root set*.
4. Dokumen yang di-*download* merupakan dokumen *web*.

1.4 Tujuan

Tujuan Tugas Akhir ini adalah sebagai berikut.

1. Menganalisis efektivitas algoritma HITS dalam menenemukan dokumen yang *authoritative* dengan mengamati *precision* dan IAP.
2. Menganalisis pengaruh jumlah *root set* dan batasan jumlah dokumen yang di-*crawling* terhadap dokumen hasil pengujian.

1.5 Metodologi penyelesaian masalah

Metode yang digunakan untuk menyelesaikan permasalahan-permasalahan Tugas Akhir ini terdiri dari langkah-langkah sebagai berikut.

1. Studi Literatur
   Mempelajari dan memahami algoritma HITS melalui literatur berupa buku, makalah, atau jurnal dari berbagai media terutama Internet.
2. Perancangan dan Pemodelan Sistem
   Pada tahap ini akan dirancang sebuah *information retrieval system* yang didalamnya dapat melakukan perhitungan nilai *authority* dalam menemukan dokumen yang relevan dengan topik yang diberikan.
3. Implementasi Sistem
   Pada tahap ini dilakukan pembangunan terhadap *system* yang telah dirancang. Pada tahap ini akan dibangun system yang dapat menangani proses *crawling*, proses perhitungan nilai *authority* dan *hub*, dan proses *searching*.
4. Pengujian dan analisis hasil.
   Pengujian dilakukan untuk memperhatikan performansi Algoritma HITS dalam *information retieval*. Pada proses ini akan dilakukan perhitungan nilai *precision* dan IAP.
5. Penyusunan Laporan
   Pada tahap ini, akan dilakukan penyusunan laporan akhir sekaligus dokumentasi dengan mengikuti kaidah penulisan yang benar dan sesuai dengan ketentuan yang ditetapkan oleh Institusi.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan yang dapat diperoleh dari pengujian yang telah dlakukan adalah bahwa :

1. Berdasarkan pengujian yang telah dilakukan dengan menggunakan *backlink metric* Algoritma HITS baik untuk perankingan bila dilihat dari nilai IAP-nya, namun kurang memberikan hasil yang baik bila dilihat dari nilai *precision*-nya. Hal ini disebabkan karena Algoritma HITS tidak mampu menentukan dokumen relevan atau tidak karena algoritma ini fokus pada struktur *link* saja.
2. Penambahan batasan jumlah dokumen yang di *crawling* dan penambahan jumlah *root set* tidak berpengaruh pada peningkatan nilai *precision dan* IAP, hal ini disebakan karena peningkatan nilai *precision* tidak hanya dipengaruhi penambahan batasan dokumen *crawling* dan jumlah *root set* namun dipengaruhi oleh jumlah *inlink* dan jumlah *outlink* yang dimiliki oleh dokumen tersebut.

5.2 Saran

Saran – saran yang dapat dilakukan jika dilakukan pengembangan terhadap TA ini adalah sebagai berikut :

1. Untuk pengujian relevansi dapat menggunakan pengujian lain selain *backlink metric.*
2. Untuk proses *crawling* dilakukan perbandingan *content similarity*.

Daftar Pustaka

[1] Borodin, Allan, Robert, O Gareth, Rosenthal S, Jeffrey, Tsaparas, Panayiotis, Finding Authorities and Hubs From Link Structures on the World Wide Web . http://www10.org/cdrom/papers/pdf/p314.pdf. Diakses pada 11 Maret 2010 pukul 11.00 WIB.

[2] Borodin Allan, Roberts, O Gareth , Rosenthal Jeffrey S, Tsaparas Panayiotis. Link Analysis Ranking : Algorithms, Theory, and Experiments. www.cs.toronto.edu/~tsap/publications/hubs-journal.ps Diakses pada 11 Maret 2010 pukul 11.00 WIB.

[3] Budianto, Arifin Zainal Agus, Lili Suhadi. Perancangan dan Pembuatan Perangkat Lunak Penelusur Web (Web Crawler) Menggunakan Algoritma Pagerank. Teknik Informatika, Institut Teknologi Sepuluh November Surabaya, 2003. http://www.its.ac.id/personal/files/material/1261-agusza-webcrawler.pdf. Diakses pada 10 Agustus 2010 pukul 15.00 WIB.

[4] Farahat, Ayman, Lofaro Thomas, Miller Joel C, Rae Gregory, Ward Lesley A. Authority Rankings From HITS, PAGERANK, AND SALSA : Existence, Uniqueness, and Effect of Initialization. http://www.math.hmc.edu/~ward/paperpdfs/hitsheaderbw6Jan05.pdf. Diakses pada 18 Oktober 2009 pukul 20.13 WIB.

[5] Frederic Stutzman. Social Network Analytic Approaches to the World Wide Web http://fredstutzman.com/pubs/stutzman_wp2.pdf

[6] Firdaus, Yanuar . "Web Search", Slide Kuliah, Institut Teknologi Telkom Bandung,Mei 2008.

[7] Henzinger, Monika . "Link Analysis in Web Information Retrieval" http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/9019.pdf. Diakses pada 6 Oktober 2010 pukul 14.00 WIB

[8] HITS Algorithm. Available at http://en.wikipedia.org/wiki/HITS_algorithm Diakses pada 6 Oktober 2009 pukul 14.32.

[9] Ipek Busra, Isik Selime . Hypertext-Induced Topic Selection.ppt. Diakses tanggal 24 April 2010. http://webcache.googleusercontent.com/search?q=cache:ArxbPVWYy0MJ:www.cmpe.boun.edu.tr/courses/cmpe473/spring2006/presentation/hits.ppt . Diakses pada 6 Oktober 2010 pukul 13.45 WIB.

[10] Irvine . Link Analysis www.**powerpoint**-search.com/**link-ppt**.html Diakses pada 9 Oktober 2010 pukul 12.45 WIB.

[11] Kleinberg M. Jhon . Authoritative Source in a Hyperlinked Environment . 1998. www.cs.cornell.edu/home/kleinber/auth.ps.

Diakses pada 1 Oktober 2009 pukul 15.30 WIB.

[12] Mesin_pencari
http://id.wikipedia.org/wiki/Mesin_pencari
Diakses pada 20 April 2010 pukul 15.30.

[13] Pengertian Search Engine Optimazation. Available at http://belajar-web-ku.blogspot.com/2008/08/pengertian-search-engine-optimization.html.
Diakses tanggal 20 April 2010.

[14] Precision and Recall.
http://en.wikipedia.org/wiki/Precision_%28information_retrieval%29.
Diakses pada 20 Juni 2010 pukul 13.00 WIB.

[15] Signorini Alessio. A survey Ranking Algorithms.
http://www.cs.uiowa.edu/~cremer/courses/cs2/SignoriniRankingAlgSurvey.pdf.
Diakses pada 19 Agustus 2010 pukul 20.22 WIB.