ANALISIS DAN IMPLEMENTASI ALGORITMA BLOCK LEVEL HITS

Damangrea Tizar Balamrayoga[1], Yanuar Firdaus A.w.[2], Bayu Munajat[3]

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

Abstrak

Pemeringkatan web berperan sangat penting dalam sebuah sistem pencarian informasi. Proses ini akan menghasilkan rekomendasi web yang dianggap penting, biasanya dapat dihitung dari banyaknya web yang mengacu ke web tersebut. Pemeringkatan ini dilakukan dengan prinsip link analysis, yaitu proses pemeringkatan dokumen berdasarkan informasi yang terkandung di dalam link. Sebagian besar algoritma link analysis yang ada, memberlakukan halaman web sebagai satu node tunggal sehingga pemberian bobot untuk satu page akan selalu sama. Namun dalam berbagai kasus, halaman web berupa semantik. Sehingga tidak mungkin dianggap sebagai satu node tunggal. Sebagai solusinya, muncul metode block level link analysis yang memilah bagian web menjadi satuan block. Dengan demikian, setiap satuan block mendapat bobot yang berbeda.

Pemilahan web menjadi suatu block memiliki berbagai cara. Cara yang dilakukan pada penelitian ini menggunakan algoritma VIPS (Visual-Based Page Segmentation) adalah pembagian berdasarkan garis batas visual yang terlihat oleh user. Bobot yang didapat setelah pemilahan dipergunakan dalam perhitungan nilai pada Block Level HITS yang merupakan turunan dari algoritma block level analysis. Block Level HITS diketahui memiliki dua nilai penentu peringkat, yaitu authority dan hub. Nilai authority adalah jumlah bobot halaman page yang mengacu, sedangkan hub jumlah bobot halaman page yang diacu dalam satu halaman web.

Berdasarkan hasil dari pengujian yang dilakukan, didapatkan bahwa nilai authority menentukan peringkat berdasarkan jumlah dan nilai bobot yang masuk kedalam page. Semakin banyak nilai bobot yang masuk, maka akan semakin tinggi ranking-nya. Nilai hub dapat berubah berdasarkan kombinasi jumlah, bobot, dan link yang diacunya. Semakin tinggi salah satu dari kombinasi tersebut, maka akan semakin tinggi pula nilai hub yang didapat.

Kata Kunci : Kata kunci : authority, block, Block Level HITS, hub, link analysis, ranking

Abstract
Web ranking is crucial to an information retrieval system. This process will result in recommendations that are considered essential web, usually can be calculated from the number of web that refers to the web. The ranking is done by the principle of link analysis, namely the process of ranking the documents based on the information contained in the link. Most of the link analysis algorithms exist, enacting a web page as a single node so that the weight given to one page will always be the same. But in many cases, a semantic web page and therefore can not possibly be considered as a single node. As a solution, block level link analysis methods that sort out the web into a unit block arise. Thus, each unit block have different weights.

Sorting the web into a block having a variety of ways. The way its done in this study using the algorithm of VIPS (Visual-Based Page Segmentation) is a division based on the visual lines are visible to the user. Gained weight after sorting used in calculating the value at Block Level HITS algorithm which is derived from block-level analysis. Block Level HITS are known to have two values determine the rankings, the authority and hub. Authority value is the amount of weight that refers to the page, while the hub weights the number of page referred to in a web page.

According the result from the tests, it was found that the authority determines ranking based on the number and weight values into the page. More weight caused higher ranking. Hub value can be vary based on the combination of the number, weight, and the link to which it refers. The higher either of these combinations, it will be higher the hub value is obtained.

Keywords : authority, block, Block Level HITS, hub, link analysis, ranking

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Saat ini, *web* telah menjadi sarana pencarian data maupun informasi penting yang dibutuhkan oleh masyarakat. Dengan banyaknya data yang tersebar dalam seluruh jaringan internet itu sendiri, perlu dipilah bagian *web* yang mengandung informasi penting atau tautan menuju informasi penting lainnya. Dalam proses pemilahan atau yang sering disebut pemeringkatan *web*, berperan sangat penting dalam sebuah sistem pencarian informasi. Proses ini akan menghasilkan rekomendasi *web* yang dianggap penting, biasanya dapat dihitung dari banyaknya *web* yang mengacu ke *web* tersebut. Diharapkan dengan banyaknya *web* yang mengacu *web* tersebut, menandakan *web* tersebut memiliki informasi yang penting.

Pemeringkatan ini dilakukan dengan prinsip *link analysis*, yaitu proses pemeringkatan dokumen berdasarkan informasi yang terkandung di dalam *link* dari sebuah halaman *web*. *Link analysis* adalah salah satu dari banyak faktor yang dipertimbangkan oleh mesin pencari *web* dalam komputasi skor komposit untuk halaman *web* pada setiap *query*-nya[6]. *Link analysis* telah menunjukkan potensinya dalam peningkatan kinerja pencarian dokumen *web*.

Terdapat dua algoritma pemeringkatan yang populer digunakan, yaitu PageRank dan HITS [2][4][6]. PageRank dan HITS merupakan sebuah algoritma yang berfungsi untuk menentukan situs *web* yang lebih penting atau populer. Dalam penerapannya, algoritma HITS ditekankan pada hubungan yang saling menguatkan pada halaman *web*, sedangkan pada PageRank ditekankan pada normalisasi berat dengan menggunakan model *random walk*[2][6]. Sebagian besar algoritma *link analysis* yang ada, memberlakukan halaman *web* sebagai satu *node* tunggal. Namun dalam berbagai kasus, halaman *web* berupa semantik sehingga tidak mungkin dianggap sebagai satu *node* tunggal. Pada beberapa *web* yang mengandung lebih dari satu semantik dan banyak *link* hanya untuk navigasi atau iklan, akan mengakibatkan kesalahan perhitungan nilai kepentingan oleh PageRank dan pergeseran topik pada HITS[4]. Penempatan *link* pada *web* semantik juga menjadi permasalahan dalam kasus ini.

Pada algoritma *link analysis* berdasarkan pada dua asumsi, yaitu tautan yang disampaikan oleh seseorang dan halaman yang dikutip oleh halaman tertentu dan mungkin memiliki topik yang sama[4]. Pada umumnya dalam perhitungan *link analysis*, *web* dianggap sebagai unit terkecil dalam pembangunan graf *web*. Dengan demikian, setiap link pada setiap *web* akan dianggap sama, tanpa memperhitungkan letak dari *link* tersebut. Akibatnya, *link* yang terletak pada *block* yang kecil kemungkinannya untuk mendapat perhatian dari *user* akan mendapat bobot yang sama pula, dibandingkan dengan *block* yang terletak di tengah halaman *web* yang memiliki kemungkinan untuk mendapat perhatian oleh *user*. Sebagai solusinya, muncul metode *block level link analysis* yang memilah bagian *web* menjadi satuan *block*. Dengan demikian, setiap satuan *block* mendapat bobot yang berbeda, sehingga *link* yang berada pada *block* yang lebih besar dan memungkinkan untuk mendapat perhatian lebih dari *user* akan mendapatkan bobot yang lebih besar pula.

Block level link analysis merupakan sebuah metode pengembangan dari *link analysis* yang menjadikan suatu halaman *web* menjadi bagian-bagian kecil. Dengan menggunakan metode ini, kekurangan *link analysis* dapat tertutupi dengan pemberian bobot yang berbeda pada setiap *block*-nya didasarkan dari besarnya *block* dan kemungkinan *block* tersebut mendapat

perhatian *user*. *Block Level* HITS merupakan penerapan dari metode *block level link analysis* yang juga merupakan pengembangan dari algoritma HITS. *Block level link analysis* adalah metode yang membagi suatu *web* menjadi *block-block* yang merupakan unit terkecil, sehingga perhitungan *link analysis* menjadi lebih dapat dihandalkan. HITS adalah suatu algoritma yang menghitung nilai otoritas dan nilai penghubung untuk menentukan *web* yang lebih penting.

1.2 Perumusan Masalah

1. Bagaimana pengaruh jumlah dan letak *link* terhadap peringkat halaman *web* yang dituju?
2. Bagaimana pengaruh hasil algoritma *Block Level* HITS dalam banyaknya *link* pada *page*?

1.3 Batasan Masalah

1. Penelitian tugas akhir ini hanya fokus dalam pemberian bobot pada algoritma *Block Level* HITS saja.
2. Dataset berupa dokumen *off-line* berupa berkas teks hasil crawling.
3. Dataset yang digunakan diambil dari http://www.cs.toronto.edu/~tsap yang telah melalui tahap *pre-processing*.

1.4 Tujuan Penelitian

1. Menganalisis pengaruh jumlah dan letak *link* terhadap peringkat halaman *web* yang dituju.
2. Menganalisis pengaruh hasil algoritma *Block Level* HITS perbandingan jumlah *link* pada suatu *page*.

1.5 Metodologi Penyelesaian Masalah

1. Studi literatur
   Pada tahap ini akan dilakukan pembelajaran konsep teori-teori tentang *link analysis* dan algoritma HITS yang digunakan, serta informasi lainnya yang menunjang pembuatan tugas akhir ini dari berbagai macam sumber.
2. Pengumpulan data
   Pada tahap ini data yang diambil untuk dataset diolah untuk penentuan peringkat dengan menggunakan algoritma *Block Level* HITS. Data yang diambil berupa data HTML dengan *link* yang menunjuk pada suatu *web* tertentu dalam dokumen *web*.
3. Pemodelan sistem
   Tahap ini meliputi analisis kebutuhan sistem dalam perhitungan pemeringkatan berdasarkan kebutuhan yang telah diidentifikasi.
4. *Testing* dan analisis hasil
   Pada *testing* akan dilakukan percobaan kepada sistem yang selanjutnya akan diketahui sistem tersebut telah berjalan dengan baik sesuai dengan tujuan yang telah ditentukan sebelumnya atau tidak. Hasil yang didapat akan berupa pembobotan *web* yang selanjutnya akan dianalisis mengenai pengaruh letak dan jumlah *link* terhadap peringkat *web* tersebut.

5. KESIMPULAN DAN SARAN

5.1 **Kesimpulan**

Penelitian ini bertujuan untuk menganalisis pengaruh jumlah dan letak *link* terhadap peringkat halaman *web* yang dituju dan pengaruh hasil algoritma *Block Level* HITS terhadap perbandingan jumlah *link* pada suatu *page*. Adapun kesimpulan yang dapat diambil dari penelitian dan analisis yang telah dilakukan, yaitu:

1. Pada algoritma *Block Level* HITS, *ranking* ditentukan dari besarnya nilai *authority* yang merupakan nilai tingkat kepentingan yang didapatkan dari jumlah bobot *block* dari setiap *link* yang mengacu pada *web page* tersebut. Jika suatu *link* yang mengacu suatu *web* terletak pada *block* dengan bobot yang besar, maka akan didapatkan nilai *authority* yang besar pula. Semakin besar jumlah *authority* yang didapatkan maka, *ranking* yang diperoleh juga akan semakin tinggi.
2. Nilai *hub* suatu *block* pada algoritma *Block Level* HITS merupakan kombinasi antara banyaknya *link* yang keluar, nilai *authority* dan bobot *block* itu sendiri. Setiap perubahan komponen akan menentukan besar kecilnya nilai *hub*.
3. Dari alasan penggunaan hasil dari perhitungan algoritma *Block Level* HITS yang telah dijelaskan pada bagian analisis, dapat disimpulkan bahwa algoritma ini lebih dapat digunakan karena kriteria pembobotannya lebih mendekati keadaan sebagian besar *web* yang ada pada saaat ini. Dengan demikian, algoritma *Block Level* HITS dapat digunakan dalam pe-*ranking*-an halaman *web*.

5.2 **Saran**

Saran yang diberikan untuk penelitian selanjutnya, yaitu :

1. Sistem dikembangkan dengan menggabungkan proses pemeringkatan, proses *searching*, serta proses *indexing* sehingga dapat diimplementasikan sebagai sebuah sistem pencarian secara utuh.
2. Algoritma *Block Level* PageRank dengan *Block Level* HITS digabungkan untuk mencari tingkat akurasi terhadap tingkat relevansi dokumen.

DAFTAR PUSTAKA

[1] "*Block Level Analysis : Definition & Examples of Block Level*". Tersedia di: http://seotermglossary.com/block-level-analysis/ [Diakses tanggal 11 Oktober 2011 pukul 16.09 WIB].

[2] Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha, Horst Simon. "PageRank, HITS, *and a Unified Framework for Link Analysis*". ACM. 2002.

[3] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. "VIPS: *a Vision-based Page Segmentation Algorithm*". 2003.

[4] Deng Cai, Xiaofei He, Ji-Rong Wen, Wei-Ying Ma. "*Block-level Link Analysis*". ACM. 2004.

[5] Gautam Pant, Padmini Srinivasan, Filippo Menczer. "*Crawling the Web*". 2004.

[6] "*Link Analysis*". Tersedia di: http://nlp.stanford.edu/IR-book/html/htmledition/link-analysis-1.html [Diakses tanggal 11 Oktober 2011 pukul 16.06 WIB].

[7] Remus, Raluca. "HITS *Algorithm – Hubs and Authorities on the Internet*". 2009.