CRAWLING WEBSITE E-GOVERNMENT PEMERINTAH DAERAH JAWA TENGAH BERBASIS ONTOLOGY

Bima Jati Wijaya[1], Heru Agus Santoso[2]
[1,2]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Email : bimajatiwijaya@gmail.com[1], herezadi@gmail.com [2]

Abstrak

Agregasi informasi pada situs e-gov Indonesia sangat diperlukan untuk memelihara konten portal web dan memenuhi kebutuhan informasi yang kompleks serta banyaknya jumlah data saat ini sulit untuk mendapatkan informasi yang relevan. Oleh karena itu harus dilakukan crawl pada website pemerintah dengan membuat program crawl. Akan tetapi crawl pada umumnya mendownload seluruh url tanpa melihat content didalamnya sedangkan crawl terfokus memiliki keterbatasan tidak memiliki cakupan yang luas dalam suatu konsep. Oleh karena itu dibutuhkan crawling berbasis ontologi untuk menyelesaikan masalah tersebut dengan disertai pengecekan relevansi menggunakan tfidf dan cosine similarity. Dengan pendekatan ontologi serta pengecekan relevansi tfidf dan cosine similarity diharapkan menghasilkan crawler yang effesien serta mempermudah dalam penggunaan hasil crawling karena sudah terklasifikasi.

Kata Kunci : Web crawl, crawler berdasarkan ontologi, tf-idf

Abstract

Aggregation of information on e-government Indonesian very required to maintain the web portal content and meet complex information needs and currently difficult to obtain relevant information from the large amount of data. Therefore it have to do crawl on the e-government website with create web crawl. However generally crawl download entire URL without seeing the content. While focused crawl has a limitation in the concept coverage. The issue became the basis of research to make web crawler based on ontology to solve the issue which combine with relevant check content use tf-idf and cosine similarity method. The existing approach on web crawler based ontology use tfidf and cosine similarity method is expected to solve the issue and it will becomes efficient web crawler which make easier to reprocess from result of crawling because it was classified.

Keywords : *Web crawl, crawler-based ontology, tf-idf*

1. PENDAHULUAN

Semakin bertambahnya pengguna internet di Indonesia dari berbagai kalangan tidak diikuti perkembangan *e-Government* di Indonesia. Sejauh ini implementasi *e-Government* di Indonesia belum maksimal dapat dilihat berdasarkan laporan *e-Government Development Index (EGDI)* oleh Perserikatan Bangsa Bangsa *EGDI* Indonesia peringkat 97 pada 2012 [11] saat ini semakin menurun menduduki peringkat 106 sedunia [12] yaitu dengan skor EGDI sebesar 0.4487. Hal ini disebabkan oleh beberapa faktor yaitu masalah infrastruktur telekomunikasi dan juga penetrasi ICT diluar Jawa [5]. Sehingga berakibat kurangnya pelayanan pemerintah untuk memberikan informasi dan pelayanan publik kepada masyarakat. *Crawler* adalah program yang digunakan oleh mesin pencari yang mengambil *link* halaman yang terdapat dalam web dari satu *link* ke *link* yang lain [10] secara keseluruhan atau disebut *crawler* tidak

terfokus, tentunya hasil yang didapat tidak seluruhnya relevan dengan yang diharapkan.

C*rawling* terfokus dapat digunakan untuk memenuhi kebutuhan organisasi atau individual dalam membuat topik-topik spesifik, misalnya untuk memelihara konten portal web, untuk mengumpulkan dokumen secara lokal atau memenuhi kebutuhan informasi yang kompleks [7]. Dari kekurangan crawler tersebut peneliti mengajukan pembuatan *crawling* terfokus berdasarkanontologi. Sehingga tujuan khusus yang ingin dicapai pada penelitian ini adalah terbentuknya model agregasi data atau informasi dengan teknik *crawling* terfokus berdasarkanontologi yang dapat diimplementasikan dengan baik dan terukur pada situs-situs *e-Government* untuk pengelompokkan.

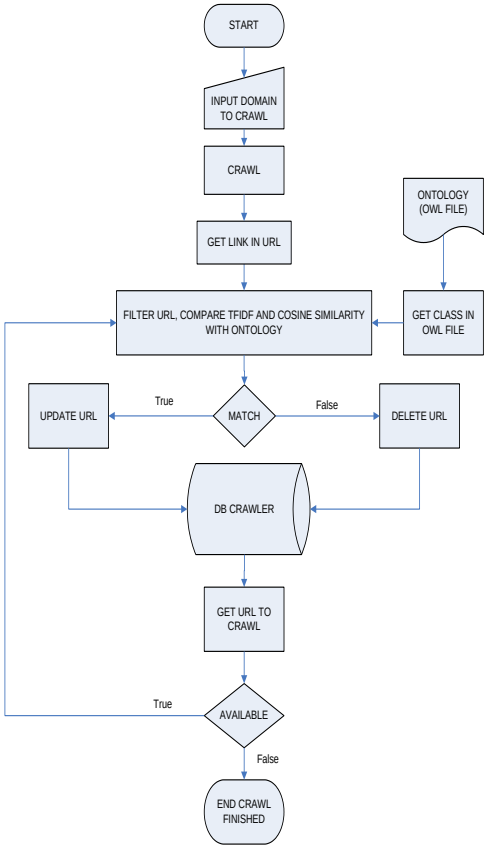
2. METODE YANG DIUSULKAN

Metode yang diajukan peneliti dalam pembuatan *web crawler* berdasarkan ontologi sebagai berikut :

1. Membuat 11 bidang atau domain ontologi berdasarkan peraturan daerah tahun 2008.
2. Metode pembuatan *web crawler* menggunakan pendekatan vertikal *prototyping.*
3. Pengecekan relevansi *web crawler* antara bidang ontologi dan konten dengan menggunakan metode TF-IDF dan *cosine similarity.*
4. Implementasi pada 10 *website* pemerintahan Jawa Tengah.
5. Pengujian metode dan kinerja *web crawler* dengan cara menganalisa ulang relevansi hasil *crawling* setiap halamannya. Sebuah halaman relevan jika memiliki satu kata kunci pada tag *title* atau memiliki minimal 10 kata kunci pada tag *body* [19].
6. Dari hasil analisa ulang relevansi terhadap hasil *web crawling* berdasarkan ontologi. Kemudian hasil tersebut digunakan untuk menghitung rasio *harvest-rate*.

Flowchart web crawler yang peneliti ajukan dapat dilihat pada gambar 1.

Metode *harvest-rate* untuk pengujian kinerja program *web crawlerHarvest-rate* merupakan sebuah metode yang digunakan untuk mengukur kinerja *web crawler* dalam sebuah rasio dengan interval 0 sampai dengan 1. Pada *harvest-rate* terdapat dua variabel r dan p [1]. Variabel r merepresentasikan jumlah halaman web atau url yang relevan dari keseluruhan halaman web yang tercrawling. Sebuah halaman relevan jika memiliki satu kata kunci pada tag *title* atau memiliki minimal 10 kata kunci pada tag *body*[19]. Kemudian variabel p merepresentasikan jumlah keseluruhan halaman web yang ter*download* oleh *web crawler*[1]. *Harvest-rate* maksimal jika nilai r sama dengan nilai p.



Gambar 1 : ***Flowchar web crawler***

3. HASIL DAN PEMBAHASAN

Pembahasan metode TF-IDF dan *cosine similarity* dalam *web crawler* berdasarkan ontologi beserta metode pengujian hasil kinerja *web crwaler* sebagai berikut :

Metode perhitungan relevansi TF-IDF dan *cosine similarity Term frequency* dan *invers document frequency (TF-IDF)* merupakan numerik statistik untuk mempertimbangkan seberapa pentingnya sebuah kata dalam dokumen dan *cosine similarity* dengan nilai *similarity* 0 sampai dengan 1 digunakan untuk membandingkan vektor antara dokumen dengan *keyword*. Tahapan metode ini yaitu :

1. *Tokenization* yaitu merupakan proses merubah deret kata menjadi kata tunggal atau token, menghilangkan tanda baca dan mengubah ke huruf kecil.
2. *Stop words removal* yaitu menghilangkan kata-kata yang dianggap tidak penting seperti kata hubung, imbuhan dan lainnya.
3. *Stemming* yaitu merubah kata yang berimbuhan menjadi kata dasar.
4. Mengurutkan *terms* secara *ascending* , menghitung frekuensi *term*(TF) dan menghitung kemunculan *term* pada dokumen.
   Dokumen1 (baik, jawa,tengah,udinus)
   Dokumen 2 (jawa,tengah,universitas)
   Terms(baik,jawa,tengah,udinus,universitas)
5. Menghitung *invers document frequency(IDF) ,N merupakan jumlah dokumen :*
   $$\mathrm{idf}_t = \log_{10}(N/\mathrm{df}_t)$$
   Dokumen frekuensi(1,2,2,1,1)
   terms
   (baik,jawa,tengah,udinus,universitas)
   IDF (0.301,0,0,0.301,0.301)
6. Memodelkan *vector space model* setiap dokumen dan *keyword*. Contoh *vector space model :*
   dokumen 1 : (1,1,1,1,0)
   dokumen 2 : (0,1,1,0,1)
7. Menghitung bobot(w) dokumen :
   $$\mathrm{w}_{t,d} = \mathrm{tf}_{t,d} \times \log_{10}(N/\mathrm{df}_t)$$
   dokumen 1 : (0.301,0,0,0.301,0)

   dokumen 2 : (0,0,0,0,0.301)
8. Melakukan normalisasi bobot dokumen dan *keyword* yaitu dengan cara setiap elemen *vector* dibagi oleh normalisasi *vector* itu sendiri.
   Normalisasi *vector*:
   $$\|v\|_1 = \sqrt{\sum_i v_i^2}$$
   $$v = \frac{v_0}{\|v\|_2}$$
   Normalisasi setiap elemen *vector*
   Contoh perhitungan :
   dokumen 1 : (0.301,0,0,0.301,0)
   $\|v\|_1 =$ 0.4257
   $v = \frac{0.301}{\|v\|_2}$ =0.7071,elementi+1,..
   v_0 (0.7071,0,0,0.7071,0)

   dokumen 2 : (0,0,0,0,0.301)
   $\|v\|_2 =$ 0.3010
   v_1 (0,0,0,0,1)
9. Tahap terakhir membandingkan *vector* yang setiap dokumen yang telah dinormalisasi dengan *keyword* menggunakan *cosine similarity*. Berikut rumus serta perhitungannya. Asumsi *keyword* adalah udinus jawa tengah diproses hingga normalisasi didapatkan = (0,0,0,1,0).
   $$\cos(q,d) = \frac{\sum_{i=0}^{|V|-1} q_i d_i}{\sqrt{\sum_{i=0}^{|V|-1} q_i^2}\sqrt{\sum_{i=0}^{|V|-1} d_i^2}}$$
   $$\cos(v_k, v_0) =$$
   $$\frac{0x0.7071 + 0x0 + 0x0 + 1x0.7071 + 0x0}{\sqrt{0^2+0^2+0^2+1^2+0^2}\sqrt{0.7071^2+0^2+0^2+0.7071^2+0^2}}$$

$$\cos(v_k, v_1) = 0.7071$$

$$\cos(v_k, v_0) = \frac{0x0 + 0x0 + 0x0 + 1x0 + 0x1}{\sqrt{0^2+0^2+0^2+0^2+1^2}\sqrt{0^2+0^2+0^2+1^2+0^2}}$$

$$\cos(v_k, v_1) = 0$$

Dari hasil perhitungan *TF-IDF* dan *cosine similarity* dapat disimpulkan dokumen ke 1 mempunyai tingkat *similarity* 0.7071 atau 70,7% dengan *keyword* sedangkan dengan dokumen ke 2 adalah 0%.

Hasil penelitian merupakan penyajian data hasil pengelompokan url, evaluasi kinerja web crawling berdasarkan ontologi sesuai peraturan daerah tahun 2008. Implementasi web crawler berdasarkan ontologi dilakukan pada 10 website pemerintahan Jawa Tengah yaitu:

1. www.semarangkota.go.id
2. www.kebumenkab.go.id
3. www.banyumaskab.go.id
4. www.pekalongankab.go.id
5. www.semarangkab.go.id
6. www.magelangkab.go.id
7. www.rembangkab.go.id
8. www.patikab.go.id
9. www.tegalkab.go.id
10. www.cilacapkab.go.id

Pada saat implementasi *crawler* peneliti menentukan minimal relevansi *keyword* terhadap konten suatu URL sebesar 27%.Hasil implementasi didapatkan 3885 halaman *website* yang tercrawling. Klasifikasi dari *web crawler* berdasarkan ontologi pada 11 bidang dapat dilihat pada tabel 1.

Tabel 1: Pengelompokan setiap bidang

No	Class atau Domain	Total
1	Bidang penanaman modal	0
2	Bidang perumahan	177
3	Bidang perencanaan pembangunan	22
4	Bidang lingkungan hidup	24
5	Bidang pertanian dan ketahanan pangan	2317
6	Bidang penataan ruang	43
7	Bidang kepemudaan dan olahraga	50
8	Bidang kesehatan	12
9	Bidang pekerjaan umum	54
10	Bidang pendidikan	219
11	Bidang koperasi dan usaha kecil dan menengah	967
	Total halaman yang tercrawling	3885

Dari total keseluruhan yang didapat oleh *web crawling* peneliti kemudian melakukan analisa relevansi kembali. web crawler dengan metode TF-IDF dan *cosine similarity* mendapatkan 3885 halaman *website* yaitu yang memenuhi 27% relevansi terhadap ontologi yang telah dibuat. Setelah analisa relevansi ulang terhadap hasil *web crawler* didapatkan 2479 halaman memenuhi syarat minimal 1 kata kunci pada tag title atau minimal 10 kata kunci pada tag *body*. Jika dihitung dengan menggunakan metode *harvest-rate* didapatkan *web crawler* berdasarkan ontologi memeliki tingkat akurasi 63,81% dari nilai relevansi minimal 27% pada saat melakukan crawling. Detail data dapat dilihat pada tabel 2.

Tabel 2: Analisa relevansi

No	Domain	Relevan	Hasil Relevan
1	Semarang kota	114	33
2	Kab. Kebumen	1775	1470
3	Kab. Banyumas	666	604
4	Kab.	621	18

	Pekalongan		
5	Kab. Semarang	17	6
6	Kab. Magelang	0	0
7	Kab. Rembang	328	31
8	Kab. Pati	21	1
9	Kab. Tegal	315	314
10	Kab. Cilacap	28	2
	Total relevan	3885	2479
	Harvest-rate		63,81%

4. KESIMPULAN DAN SARAN

Dari hasil penelitian pembuatan *web crawler* berbasis ontologi dengan metode TF-IDF dan *cosine similarity* peneliti dapat menyimpulkan bahwa *crawling* berbasis ontologi dengan pengecekan relevansi TF-IDF dan *cosine similarity* cukup tepat untuk mengkategorikan suatu halaman relevan atau tidak. Hal ini dapat dilihat dari nilai relevansi minimal yang hanya lebih besar atau sama dengan 27% pada saat melakukan *crawling* peneliti mendapatkan rasio *harvest-rate* yang cukup tinggi yaitu 63,81%.

Berikut saran untuk melakkukan penelitian lebih lanjut mengenai *web crawler* yaitu :

1. Diperlukan suatu metode untuk mencari struktur *class* atau id pada tag html yang merupakan konten dari suatu halaman *website* untuk meningkatkan peforma *web crawler*.
2. Penambahan kode program untuk penanganan *file binnary* (pdf atau gambar).
3. Penambahan *class* atau domain untuk membangun ontologi sehingga semakin banyak halaman *website* yang terklasifikasi. Diharapkan dengan semakin banyak *class* atau domain pada penelitian selanjutnya dapat meningkatkan nilai minimal relevansi pada saat melakukan *crawling*.
4. Pada penelitian selanjutnya diharapkan menambahkan metode perhitungan relevansi yang lain guna mencari metode yang lebih baik.
5. Perlunya simulasi pencarian pada web portal dari halaman *website* yang sudah ter*crawling* untuk pengujian lebih lanjut.

5. DAFTAR PUSTAKA

[1] Zuliarso Eri, and Mustofa Khabib, “Crawling Web berdasarkan Ontology” ,*Jurnal Teknologi Informasi DINAMIK*, vol XIV, no. 2,pp. 105-112, July 2009.

[2] Yang Sheng-Yuan, “OntoCrawler: A focused crawler with ontology-supported website models for information agents”, *ELSEVIER*, pp. 5381-5389,2010

[3] Bedi Punam,Thukral Anjali and Banati Hema,”Focused crawling of tagged web resources using ontology”,*ELSEVIER*, pp. 613-628,October 2012.

[4] Pal Anshika,Tomar Deepak Singh and Shivastava S.C, “Effective Focused Crawling Based on Content and Link Structure Analysis”,vol. 2, no. 1, June 2009.

[5] Hermana Budi and Silfianti Widya, “Evaluating E-government Implementation by

Local Government: Digital Divide in Internet Based Public Services in Indonesia", *Internasional Journal of Business and Sosial Science*, vol. 2, no. 3, January 2011.

[6] Kozanidis Lefteris, "An Ontology-Based Focused Crawler", Patras University, Greece,2008.

[7] Batsakis Sotiris, Petrakis Euripides G.M. and Milios Evangelos, "Improving the performance of focused web crawlers", *ELSEVIER*,pp. 1001-1013,April 2009.

[8] More Mangesh R and Govilkar Sharvari, "Profile Based Document Specific Crawling" ,*International Journal of Engineering Research & Technology*, vol.2, issue 2, February 2013.

[9] Nuriana Ayuningtyas, "Implementasi Ontologi web dan Aplikasi Semantik untuk Sistem Sitasi Jurnal Elektronik Indonesia," Universitas Indonesia, Depok, Skripsi Teknik Elektro, Juni 2009.

[10] Kumar Mukesh and Vig Renu, "Learnable Focused Meta Crawling Through Web", *ELSEVIER*, pp. 606-611, 2012.

[11] "United Nations E Government Survey 2012", United Nations, New York,February2012.

[12] "United Nations E Government Survey 2014", United Nations, New York, 2014.

[13] John Hebeler, Matthew Fisher, Ryan Blace and Andrew Perez-Lopez, "Semantic Web Programming", WILLEY, 2009

[14] Liyang Yu, "Introduction to the Semantic Web and Semantic Web Services", Chapman & Hall/CRC, 2007.

[15] Bunafit Nugroho, "Aplikasi Pemrograman Web Dinamis dengan PHP dan MySQL", GAVA MEDIA, Yogyakarta, Oktober, 2004.

[16] Wikipedia,"MySQL"http://id.wikipedia.org/wiki/MySQL (diakses tanggal 6 Januari 2015)

[17] Mesbah Ali,Van Deursen Arie dan Lenselink Stefan," Crawling Ajax-based Web Applications through Dynamic Analysis of User Interface State Changes", ACM Transactions on the Web,2012.

[18] Putra Dwi," Pengembangan model phonetic similarity bahasa indonesia berdasarkan kamus fonetik bahasa indonesia Zahra",Universitas Indonesia,FASILKOM,2009.

[19] Cho, J. and Garcia-Molina, H. and Page, L. (1998) *Efficient Crawling Through URL Ordering.* In:

Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

[20] Gruber,T.,1998,What is a Ontology?, http://www-ksl.stanford.edu/kst/what-isan-ontology.htm (terakhir diakses pada 31 Juli 2015)

[21] https://en.wikipedia.org/wiki/Software_prototyping (terakhir diakses pada 2 Agustus 2015)