Perangkat Lunak Penganalisis Kemiripan *Webpage* Berdasarkan Konten Presentasional

Hadipurnawan Satria
Universitas Sriwijaya
hadipurnawan.satria@gmail.com

Anggina Primanita
Universitas Sriwijaya
anggina.primanita@gmail.com

Abstract – Sebuah webpage selain berisi sekumpulan informasi utama (konten) juga mengandung konten presentasional yang digunakan untuk menampilkan isi informasi utama. Pada sebuah website, konten presentasional sebuah webpage cenderung mirip dengan konten presentasional dalam webpage lainnya di website tersebut. Meskipun mirip ataupun identik, setiap kali sebuah webpage dimuat dalam browser konten presentasional ini tetap mengalami proses pemuatan ulang. Jika kemiripan konten presentasional cukup besar, maka akan terjadi banyak pemborosan konten yang dimuat dari server. Penelitian ini bertujuan untuk mengembangkan perangkat lunak yang dapat menganalisis kemiripan sekelompok webpage dalam sebuah website. Data yang digunakan adalah kumpulan webpage dari sebuah website yang diunduh menggunakan web crawler. Berdasarkan hasil analisis pada website www.pusbangdik.unsri.ac.id , didapatkan bahwa konten presesentasional dari masing-masing webpage cukup mirip, dengan rata-rata kemiripan 67% untuk semua webpage dan 58% untuk webpage yang terhubung saja.

Index Terms – webpage similarity, web processing

PENDAHULUAN

Pada era internet sekarang ini, *website* adalah sumber informasi yang sering dipakai. Sebuah *website* pada umumnya berisi sekumpulan informasi (selanjutnya disebut konten) yang berhubungan satu sama lain. Konten dalam sebuah *website* biasanya tidak ditampilkan secara keseluruhan dalam satu dokumen tapi disusun menjadi beberapa dokumen (disebut *webpage*). Konten pada sebuah *webpage* dapat berupa teks dan gambar yang disusun dan ditampilkan dalam format HTML.

Selain konten utama, sebuah dokumen HTML seperti *webpage* berisi konten tambahan yang menentukan bagaimana konten-konten utama pada dokumen tersebut disusun dan ditampilkan bersamaan. Disamping itu, sebuah dokumen HTML juga dapat mengandung informasi tambahan yang mengatur bagaimana cara berinteraksi dengan pengguna. Konten-konten ini selanjutnya disebut konten presentasional karenanya sifatnya yang berhubungan dengan cara penampilan konten utama. Sering kali, dalam sebuah dokumen HTML beragam konten-konten tambahan ini lebih banyak jumlahnya daripada konten utamanya sendiri.

Webpage yang dihasilkan dari *server-side script* yang sama biasanya memiliki tampilan mirip satu sama lain. Banyak *website* yang menggunakan *themes* untuk mendapatkan tampilan yang konsisten pada setiap *webpage*-nya. Banyak pula *website* yang menggunakan Content Management System (CMS) untuk memudahkan pengorganisasian dan manajemen informasi dalam *website* tersebut. Hal-hal di atas menyebabkan *webpage* terlihat mirip satu sama lain, yang berarti bahwa *webpage* tersebut memiliki konten presentasional yang mirip pula.

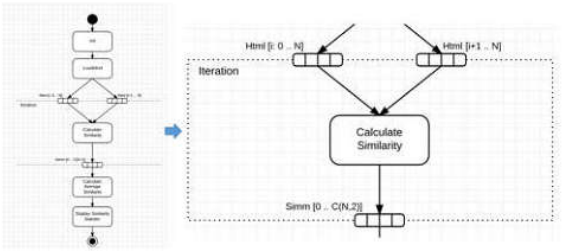
Pada kondisi jaringan yang tidak baik, besarnya konten presentasional suatu *webpage* dalam meningkatkan proses loading secara signifikan. Bahkan pada kondisi jaringan yang lebih baik, pemuatan konten presentasional secara berulang dapat membebani *transfer bandwidth* pengguna. Semakin besar kemiripan konten presentasional sebuah *website*, maka akan semakin besar pula pemborosan *transfer bandwidth*.

Penelitian yang membahas tentang kemiripan *webpage* telah banyak dilakukan sebelumnya, namun belum ada yang secara teknis mengkhususkan pembahasan kemiripan konten presentasional dan konsekuensinya. Penelitian seperti [1], [2], [3] dan [4] menggunakan nilai kemiripan antar *web page* hanya untuk menentukan letaknya pada kluster yang sudah ditentukan baik secara manual maupun otomatis. Sedangkan penelitian lainnya seperti [5], dan [6] berfokus pada konten informasi pada suatu *web page,* dengan mengindahkan konten presentational didalamnya.

Pada penelitian ini penulis melakukan pengembangan perangkat lunak penganalisis kemiripan *webpage* dalam sebuah *website.* Kemudian, dengan menggunakan perangkat lunak yang telah dikembangkan, penguji melakukan studi dan analisis kemiripan *webpage* pada sebuah *website* yang telah ditentukan.

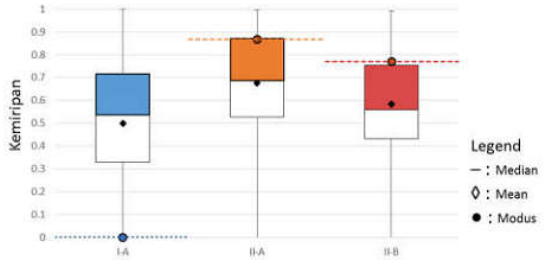
HASIL DAN ANALISIS HASIL

Untuk mencari kemiripan konten presentasional setiap *webpage* dalam sebuah *website* digunakan metode sebagai berikut (gambar 1). Pertama-tama, perangkat lunak melakukan inisialiasi (*init*) lalu kemudian dokumen-dokumen yang telah dikumpulkan oleh *web crawler* dan disimpan dalam *database* di-*load* ke *memory* agar dapat dimanipulasi. Selanjutnya, setiap dokumen HTML (dilambangkan dengan *i*) dicari kemiripannya dengan setiap dokumen HTML yang berada setelahnya (*i+1*) hingga dokumen terakhir. Total kemiripan yang dihitung berjumlah ${}_{N}C_{2}$. Dari semua kemiripan yang didapat, selanjutnya dihitung nilai-nilai statistik untuk dianalisis.



GAMBAR 1. ALUR PROSES PERHITUNGAN KEMIRIPAN N-*WEBPAGE*

Pada penelitian ini dilakukan studi analisis kemiripan *webpage* yang datanya diambil dari *website* pusbangdik UNSRI (pusbangdik.unsri.ac.id), yang dibagi menjadi tiga data set I-A, II-A dan II-B. Pada pengumpulan dataset I, *web crawler* mendapatkan 140 *webpage* sedangkan pada data set II 286 *webpage*. Perbedaan antara dataset II-A dan II-B adalah dataset II-B merupakan subset dari dataset II-A dimana hanya *webpage* yang terhubung saja yang diambil. Hasil analisis dapat dilihat di gambar 2.



GAMBAR 2. PLOT BOX DAN WHISKER UNTUK DATA SET I-A, II-A & II-B

Pada data set I-A, terlihat nilai median dan mean berdekatan dan bernilai sekitar 0.5, akan tetapi nilai modus sangat timpang yaitu 0.0. Nilai modus ini diimbangi oleh distribusi tertinggi selanjutnya yaitu 0.9 dan 0.8 yang menyebabkan nilai mean dan median tetap ditengah. Timpangnya nilai modus disebabkan oleh adanya dua tipe presentasi konten di *website* pusbangdik, yaitu tipe biasa dan tipe cetak.

Pada studi selanjutnya (data set II), *webpage* tipe cetak dikeluarkan dari dataset, dan total jumlah *webpage* yang dikumpulkan *web crawler* juga ditambah. Dalam dataset II-A terlihat efek dari pengeluaran tipe cetak (gambar 2). Sebaran nilai lebih condong ke atas dan titik kuartil bertama berada di atas 0.5, yang berarti 75 persen kemiripan berada di atas 0.5. Kemudian, nilai mean dan median berdekatan di sekitar 0.67 dan nilai modus berada di sekitar 0.88.

Dataset II-A melihat kemiripan semua kemungkinan *webpage* yang satu dengan yang lain. Namun pada kenyataannya, pengguna lebih mungkin berpindah dari satu *webpage* ke *webpage* lain bila terdapat *link* dari sebuah *webpage* ke *webpage* lain. Karena itulah studi selanjut dilakukan menggunakan dataset II-B dimana isi dataset ini hanya *webpage* yang terhubung saja. Pada data set II-B terjadi penurunan nilai kemiripan tetapi mean, median dan modus masih tetap di atas nilai 0.5 .

KESIMPULAN DAN SARAN

Berdasarkan data hasil dan analisis terhadap data hasil tersebut, dapat disimpulkan bahwa perangkat lunak yang dibangun berhasil menghitung kemiripan laman-laman web dalam sebuah *website*, lalu kemudian menampilkan statistik kemiripan tersebut, baik untuk semua laman tanpa terkecuali, maupun hanya untuk laman-laman yang terhubung (*linked*). Hasil data kemiripan dan statistiknya telah dapat dipergunakan untuk proses analisis lebih lanjut.

Untuk *website* yang diuji, yaitu *website* Pusbangdik UNSRI, berdasarkan hasil perhitungan dan analisa hasil tersebut, didapatkan bahwa laman-laman dalam *website* Pusbangdik cukup mirip satu sama lain, dengan rata-rata kemiripan sekitar 67% untuk semua laman dan 58% untuk laman-laman yang terhubung saja.

Penelitian ini masih jauh dari sempurna. Perlu dilakukan lebih banyak lagi pengujian ke *website-website* yang lebih besar dan jumlah lamannya lebih banyak, kemudian dianalisis dan dibandingkan hasil dari satu *website* dengan *website* yang lain. Jenis-jenis *website*-nya pun bisa dibuat beragam.

DAFTAR PUSTAKA

[1] R. Akiyama, K. Kanamori dan H. Ohwada, "Clustering Web Pages Considering the Position of Each Word and the Search Term," *International Journal of e-Education, e-Business, e-Management and e-Learning,* vol. 3, no. 6, p. n/a, 2013.

[2] Z. Yao dan B. Choi, "Clustering Web Pages into Hierarchical Categories," *International Journal of Intelligent Information Technologies,* vol. 3, no. 2, pp. 17-28,30-35., 2007.

[3] J. Wu, L. Chen, Z. Zheng, M. R. Lyu dan Z. Wu, "Clustering Web services to facilitate service discovery,"

Knowledge and Information Systems, vol. 38, no. 1, pp. 207-229, 2014.

[4] Q. Tan dan P. Mitra, "Clustering-based incremental web crawling," *ACM Transactions on Information Systems ,* vol. 28, no. 4, 2010.

[5] D. Sánchez, M. Batet, A. Valls dan K. Gibert, "Ontology-driven web-based semantic similarity," *Journal of Intelligent Information Systems,* vol. 35, no. 3, pp. 383-413, 2010.

[6] K. C. Srikantaiah, S. M., K. R. Venugopal dan L. M. Patnaik, "Similarity based Dynamic Web Data Extraction and Integration System from Search Engine Result Pages for Web Content Mining," *ACEEE International Journal on Information Technology,* vol. 3, no. 1, p. n/a, 2013.

ISBN: 979-587-573-6