
RANCANG BANGUN APLIKASI SERVER CRAWLING BERITA ONLINE SEBAGAI PENYEDIA BERITA UP TO DATE PADA HANDPHONE YANG MENDUKUNG WAP

Moh Badrullami, Idris Winarno, S.ST, M.Kom, Entin Martiana K., S.Kom, M.Kom

Program D IV Jurusan Teknik Informatika
Politeknik Elektronika Negeri Surabaya-Institut Teknologi Sepuluh Nopember
Kampus ITS Keputih Sukolilo Surabaya 60111
Tel: (+62)31-5910040 Fax: (+62)31-5910040
E-mail: cyberbreaks88@yahoo.co.id

ABSTRAK

Berita merupakan sesuatu hal yang memiliki peranan penting dalam kehidupan manusia. Karena manusia hidup pasti membutuhkan sebuah informasi, Seseorang bisa mengakses berita melalui berbagai media, antara lain : media cetak, media elektronik, dan media internet. Kebanyakan orang menginginkan sebuah berita dapat diakses di mana saja dan kapan saja Oleh karena itu timbul sebuah pemikiran bagaimana menjadikan berita itu menjadi portable. Pada proyek akhir ini, akan dibangun sebuah aplikasi server yang bertugas untuk crawling berita dan memanager nya, berita tersebut ditujukan kepada mobile user seperti handphone. Tujuan dari proyek akhir ini yakni memberikan kemudahan bagi pengguna handphone untuk mendapatkan berita-berita terbaru dari situs penyedia berita di internet. Dengan memanfaatkan fitur RSS dari situs, sistem akan mendapatkan berita yang terbaru dari situs tersebut. Kemudian mencari berita secara lengkap dengan mengambil konten halaman di dalam link hasil dokumen RSS. Berita yang masih melekat pada dokumen HTML akan dipisahkan dengan menggunakan metode RegEx atau *Regular Expression*. Kemudian berita dikategorikan dengan menggunakan proses *text mining* dan analisis korelasi. Pada perhitungan korelasi, yang menjadi *variable* adalah dokumen berita dan dokumen kategori yang telah disediakan oleh sistem. Kemudian disimpan ke dalam database yang kemudian akan diakses oleh server WML. Hasil yang diharapkan adalah akses berita ke server WML dengan load data yang lebih cepat dibandingkan dengan akses langsung ke situs aslinya. Dan pengkategorian berita dengan tingkat error 20 sampai dengan 30%.

Kata kunci : berita, rss, regex, *text mining*, analisis korelasi, wml.

I. PENDAHULUAN

1.1 Latar Belakang

Seiring dengan perkembangan jaman, teknologi di dunia ini mengalami kemajuan yang sangat pesat. Salah satunya perkembangan di dunia *mobile phone* atau *handphone*. Yang sekarang ini sedang populer adalah fasilitas WAP, yaitu sebuah protocol untuk menghubungkan ke dalam dunia internet yang cukup melalui sebuah handphone, tentunya yang mendukung WAP.

Berita merupakan sesuatu hal yang dapat dikatakan sebagai kebutuhan pokok dalam diri seseorang. Karena manusia hidup pasti membutuhkan sebuah informasi entah itu tentang pekerjaan, pendidikan, dsb. Dengan berita seseorang tahu akan sesuatu yang baru. Oleh karena itu tanpa berita seseorang bisa dikatakan sebagai katak dalam

tempurung. Seseorang bisa mengakses sebuah berita melalui berbagai media, antara lain : media cetak, media elektronik, dan media internet. Kebanyakan orang menginginkan sebuah berita dapat diakses di mana saja dan kapan saja berada. Oleh karena itu timbul sebuah pemikiran bagaimana menjadikan berita itu menjadi portable.

Pada proyek akhir ini, akan dibangun sebuah aplikasi server yang bertugas untuk crawling berita dan memanager nya. Sasaran utama pengguna layanan berita ini adalah mobile user seperti handphone.

1.2 Tujuan

Tujuan proyek akhir ini adalah memberikan kemudahan bagi pengguna handphone untuk mendapatkan berita-berita terbaru dari situs penyedia berita di internet. Dengan load data lebih cepat dibandingkan akses berita ke situs aslinya.

1.3 Batasan Masalah

Berdasarkan uraian di atas, maka permasalahan yang timbul dalam pengerjaan proyek akhir ini antara lain adalah:

1. Membangun sebuah perangkat lunak yang mampu crawling berita kedalam sebuah URL yang telah terdaftar.
2. Menjadikan sebuah sistem dapat mengelompokkan berita berdasarkan kategori yang telah disediakan.
3. Membangun sebuah perangkat lunak yang mampu menampilkan berita yang telah dikelompokkan sehingga dapat diakses oleh handphone.

Adapun permasalahan yang perlu dibatasi dalam penyusunan proyek akhir ini adalah :

1. URL didaftarkan dengan manual yakni diinputkan oleh user.
2. Dibatasi hanya 3 URL yang didaftarkan, yakni rss detik, rss okezone, dan rss antaranews.
3. URL yang didaftarkan harus mempunyai fasilitas RSS.
4. Berita yang dicari hanya berita yang berbahasa Indonesia.
5. Kategori yang disediakan antara lain : hukum dan kriminal, ekonomi dan bisnis, olahraga, politik, teknologi, dan bencana.
6. Berita yang disajikan khusus bagi pengguna handphone yang mendukung WAP.

II. TINJAUAN PUSTAKA

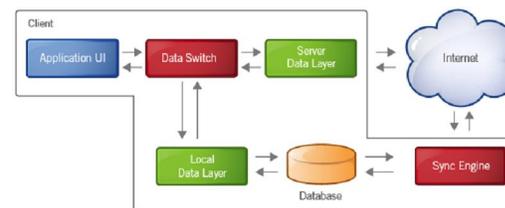
2.1 Algoritma Crawling

Beberapa mesin pencari atau *search engine* yang tersedia di internet memanfaatkan *web crawler* untuk mendapatkan informasi yang dibutuhkan oleh user. *Web crawler* bisa dikatakan sebagai sebuah aplikasi yang berjalan menjelajahi halaman – halaman situs di internet dan kemudian mengambil informasi yang tersedia pada masing – masing halaman tersebut.

Pada referensi^[2] akan dibuat sebuah aplikasi *web crawler* yang akan mencari informasi tentang

komputer rakitan yang tersedia pada beberapa situs di internet. Situs – situs yang akan dikunjungi telah ditentukan sebelumnya. Data – data yang diperoleh dari proses crawling akan disimpan di dalam database yang kemudian dapat digunakan oleh *user* untuk melakukan kombinasi spesifikasi komputer sesuai dengan yang di inginkan.

Dalam kasus ini dikenalakan istilah *Google Gears*, aplikasi yang akan dibuat memanfaatkan kinerja dari *Google Gears*. Pemanfaatan *Google* dalam aplikasi yang akan dibuat, yakni digunakan agar web dapat berjalan secara *offline* tanpa adanya koneksi dengan jaringan internet. *Google Gears* memungkinkan aplikasi untuk berjalan di komputer lokal sehingga proses pencarian semakin mudah dan cepat^[2].



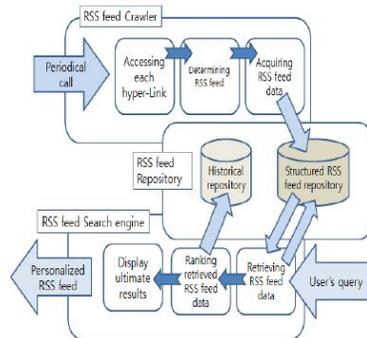
Gambar 2.1 Arsitektur *Google Gears*

2.2 RSS

Meskipun banyak sekali RSS feed yang tersedia, namun bukan suatu hal yang mudah untuk mencari informasi yang relevan dengan pengguna. Layanan RSS feed yang diambil ke dalam *account* RSS sebelumnya tidak menggunakan RSS karakteristik informasi dan *user context*. Sehingga hal ini sangat membatasi layanan informasi kepada user. Kemudian diusulkan sebuah personalisasi baru dari layanan RSS dengan menggunakan struktur RSS *feed* dan *user context*, Tujuan dari usulan ini adalah mengarahkan informasi yang sangat relevan kepada pengguna sesuai dengan yang pengguna inginkan yakni mengumpulkan data dari layanan RSS situs dengan mengkategorikan struktur RSS *feed* dan merangking RSS *channel* menggunakan tag RSS karakteristik dan *user context*. Salah satu contoh penyedia layanan personalisasi pada web, yakni Pandora musik genome menawarkan layanan kepada pengguna berupa *streaming* musik berdasarkan preferensi dari pengguna. Dengan layanan tersebut, pengguna dapat mendengarkan musik sesuai dengan preferensi pengguna.

Untuk mencapai tujuan tersebut diperlukan sebuah desain efisiensi web crawler yang spesifik

untuk mendapatkan RSS *feed* yang bukan tipe dari dokumen web dan repositori RSS *feed* yang berdasarkan RSS spesifikasi yang digunakan untuk meningkatkan *performance* dari RSS *feed retrieval*. Selain itu juga dikenalkan sebuah algoritma peringkat baru menggunakan berbagai faktor karakteristik yang relevan dari RSS dan konteks dari setiap pengguna^[3].



Gambar 2.2 Arsitektur sistem pencarian

2.3 Formula RegExp

Seiring dengan perkembangan jaman yang banyak membawa dampak positif juga disisi lain terdapat dampak negatif, yakni dengan munculnya spam robot. Dimana spam robot akan mencari setiap alamat email yang ada di seluruh situs dan kemudian dikirim iklan yang kebanyakan tidak bermanfaat. Ada beberapa cara untuk melindungi alamat email pada suatu situs. Salah satunya yakni dengan mengganti alamat email yang berbasis teks tersebut menjadi gambar/image. Sehingga mencegah spam robot melakukan parsing terhadap alamat email.

Pada referensi^[4] dilakukan sebuah penelitian dengan mengambil studi kasus pada situs yang menggunakan *Content Management System (CMS)* wordpress yang bertujuan untuk melakukan pengembangan pada konten situs wordpress sehingga setiap alamat email yang ditulis secara standard nantinya secara otomatis akan diubah menjadi gambar.

Terdapat beberapa cara untuk mendapatkan alamat email diantaranya adalah dengan menggunakan *engine/program* yang biasa disebut sebagai *spam robot* yang mencari alamat email di seluruh konten situs web di dunia. Saat alamat email ditemukan, maka sejak saat itulah iklan akan mulai dikirimkan^[4].

2.4 Text Mining

Dengan berkembangnya internet, semakin banyak pula bermunculan situs atau blog yang menyediakan berbagai macam artikel berita secara online. Sebuah artikel, sebelum dapat diterbitkan, awalnya dikirim oleh wartawan ke editor untuk dipilah. Pemilahan jenis berita relatif mudah dilakukan oleh manusia, tetapi jika kasus ini dibawa ke level pemilahan secara otomatis dengan komputer akan membawa permasalahan tersendiri, walaupun untuk berita yang lebih pendek. Text mining adalah salah satu cara yang diharapkan dapat mengatasi permasalahan di atas. Dengan text mining, dapat dicari kata-kata yang dapat mewakili isi dari artikel berita, lalu ditentukan kategorinya berdasarkan frekuensi kata-kata yang terdapat di dalamnya.

Pada referensi^[1] dibuat sebuah sistem otomatis pemilahan berita dengan menggunakan text mining sehingga diharapkan dapat membantu sistem redaksi elektronik untuk dapat memilah atau mengetahui kategori dari sebuah artikel berita tanpa memerlukan seorang editor sehingga menghemat waktu dan biaya dalam menjalankan bisnis pada model kantor berita elektronik on-line berbasis internet.

Proses pemilahan berita yang dilakukan terdiri dari beberapa proses, antara lain : pembuatan database, dan pemilihan berita^[1].

2.5 Analisis Korelasi

Bencana alam merupakan sesuatu hal yang terjadi secara alami dan tidak dapat dihindari. Sementara saat ini teknologi yang kita miliki hanya bisa memprediksi kapan dan hanya membaca tanda – tanda terjadinya bencana. Pada dasarnya manusia telah memiliki data – data kongkret dari bencana yang telah terjadi sebelumnya. Dari data – data tersebut sebenarnya sangat bisa digunakan untuk memprediksi dan menganalisa dampak dari bencana yang akan terjadi selanjutnya. Dan selain itu bencana alam adalah sebuah rangkaian sistem yang memiliki ciri yang tidak jauh berbeda, serta memiliki waktu periodik tertentu.

Pada referensi^[5] dibuat sebuah sistem pengetahuan global yang dapat membantu dalam menganalisa dampak terjadinya bencana. Sistem pengetahuan tersebut tersebut berjalan melalui informasi yang tersimpan dalam Wikipedia. Wikipedia disini berfungsi sebagai sumber informasi. Setelah itu informasi akan dipresentasikan ke dalam bentuk matriks yang menyatakan hubungan dari tiap kata kunci. Matriks tersebut melalui metode Mathematical Model Of Meaning akan menghasilkan data vektor. Dengan pemodelan matematik, sistem akan mampu

mengenal nilai keterkaitan antara kata kunci. Nilai tersebut di dapat dari rumus korelasi serta inner product. Dengan hubungan Keterkaitan tersebut, maka dapat diketahui nilai dampak yang ditimbulkan dari suatu gejala atau bencana. Tujuan dari sistem sendiri adalah menciptakan pusat pengetahuan global tentang bencana alam sehingga dapat menjadi sumber informasi yang dapat membantu menganalisa dampak dari bencana alam secara akurat^[5].

Berikut merupakan rumus dari perhitungan korelasi :

$$correl(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{(\sum (x - \bar{x})^2)(\sum (y - \bar{y})^2)}} \dots\dots\dots(1)$$

Rumus di atas digunakan untuk menentukan nilai keterkaitan antar kata kunci dalam suatu matriks. Dimana :

- x adalah kata kunci dari dokumen artikel Wikipedia
- y adalah kata kunci dari dokumen database.

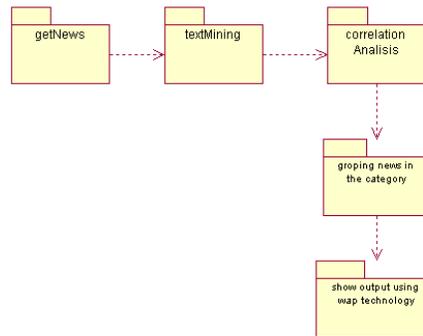
III. PERANCANGAN DAN IMPLEMENTASI

3.1 Desain Input

Input dari sistem ini adalah berupa RSS feed dari sebuah situs. Dan kemudian situs tersebut akan diolah oleh sistem untuk mencari elemen – elemen yang ada di dalamnya.

3.2 Desain Proses

UML (*Unified Modelling Language*) adalah salah satu alat bantu yang sangat handal di dunia pengembangan sistem yang berorientasi obyek. Hal ini disebabkan karena UML menyediakan bahasa pemodelan visual yang memungkinkan bagi pengembang sistem untuk membuat cetak biru (*blueprint*) atas visi mereka dalam bentuk yang baku, mudah dimengerti, serta dilengkapi dengan mekanisme yang efektif untuk berbagi (*sharing*) dan mengkomunikasikan rancangan mereka dengan yang lain. UML memiliki banyak model diagram dan yang digunakan disini adalah Use Case Diagram dan Activity Diagram.



Gambar 3.1 Use Case utama

3.2 Desain Output

Output yang dihasilkan berupa berita yang dapat diakses melalui handphone. Agar menjadi dinamis WML akan diintegrasikan dengan php. Sebelumnya php akan menampilkan potongan dari dokumen berita, dilengkapi dengan link berita selengkapnya. Apabila link berita selengkapnya di klik maka window akan secara otomatis menampilkan content berita secara lengkap. Berikut merupakan gambaran dari pernyataan diatas :



Gambar 3.2 Hasil output system

IV. UJI COBA DAN ANALISA

4.1 Analisa Pemisahan Tag

Data berita pada awalnya di dapatkan dari halaman HTML, sehingga masih tercampur dengan tag – tag HTML. Jadi perlu adanya pemisahan tag – tag tersebut dari dokumen berita. Agar di dapatkan sebuah berita yang bersih dari tag dan mudah dibaca, langkah – langkahnya antara lain :

1. Menghilangkan tag `<a href>`
2. Menghilangkan main tag `<. *?>`
3. Menghilangkan tanda `&. *? ;`
4. Menghilangkan tanda `"\t\n| ->"`

4.2 Analisa Pengkategorian Berita

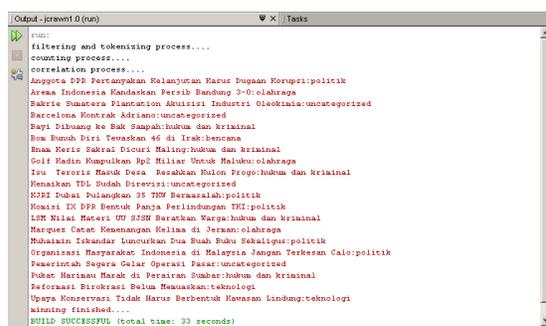
Pengkategorian berita disini dilakukan dengan menggunakan text mining dan analisis korelasi. Nilai analisis korelasi disini merupakan penentu pengkategorian berita, misalnya : berita dapat dikatakan dapat masuk kedalam kategori x apabila nilai korelasi dokumen berita dengan x paling besar dibanding yang lain. Maka dokumen berita tersebut masuk kedalam kategori x. Seperti yang ada pada contoh, nilai korelasi di dapatkan dari dokumen berita dan dokumen kamus kategori. Kemudian hasil dari nilai korelasi di shorting dengan nilai minimum berupa nilai threshold. Hasil dari percobaan yaitu kategori dari berita itu sendiri. Kategori yang disediakan antara lain : bencana, ekonomi dan bisnis, hokum dan kriminal, olahraga, politik, dan teknologi. Yang mana pada setiap kategori memiliki kamus sebagai nilai bobot yang akan digunakan pada perhitungan korelasi.

4.2.1 Percobaan 1

Pada percobaan ini akan dilakukan ujicoba untuk beberapa data campuran dengan data kamus sesuai dengan data yang telah di update terakhir kali, threshold -0.55 dan jumlah pembatasan kata minimal > 0. Dan data campuran tersebut diambil ng dari situs di bawah ini pada tanggal 18 Juli 2010 jam 20.41

<http://www.antaraneews.com/rss/news.xml>

Berikut merupakan tampilan dari percobaan 1



```

C:\Users\jcravnt1> java -jar news1_0.jar
filtering and tokenizing process...
counting process...
correlation process...
Anggota DPR Pertanyakan Kelanjutan Kasus Dugaan Korupsi:politik
Arem Indonesia Kandaskan Persib Bandung 3-0:olahraga
Bakrie Sumatera Plantation Akuisisi Industri Oleokimia:uncategorized
Barcelona Kontrak Adriano:uncategorized
Bayi Dibuang ke Bak Sampah:hukum dan kriminal
Bom Bunuh Diri Tewaskan 46 di Irak:bencana
Enam Keris Sakral Dicuri Maling:hukum dan kriminal
Golf Kadin Kumpulkan Rp2 Miliar Untuk Maluku:olahraga
Isu Teroris Masuk Desa Resahkan Kulon Progo:hukum dan kriminal
Kenaikan TDL Sudah Direvisi:uncategorized
KJRI Dubai Pulangkan 35 TKW Bermasalah:politik
Komisi IX DPR Bentuk Panja Perlindungan TKI:politik
LSM Nilai Materi UU SJSN Beratkan Warga:hukum dan kriminal
Marquez Catat Kemenangan Kelima di Jerman:olahraga
Muhaimin Iskandar Luncurkan Dua Buah Buku Sekaligus:politik
Organisasi Masyarakat Indonesia di Malaysia Jangan Terkesan Calo:politik
Pemerintah Segera Gelar Operasi Pasar:uncategorized
Pukat Harimau Marak di Perairan Sumbar:hukum dan kriminal
Reformasi Birokrasi Belum Memuaskan:teknologi
Upaya Konservasi Tidak Harus Berbentuk Kawasan Lindung:teknologi
warning finished...
BUILD SUCCESSFUL (total time: 33 seconds)
  
```

Gambar 4.1 Hasil percobaan 1

Agar lebih jelas maka hasil keluaran diletakkan pada sebuah table

Tabel 4.1 Hasil percobaan 1

Judul Berita	Kategori
Anggota DPR Pertanyakan Kelanjutan Kasus Dugaan Korupsi	politik
Arem Indonesia Kandaskan Persib Bandung 3-0	olahraga
Bakrie Sumatera Plantation Akuisisi Industri Oleokimia	uncategorized
Barcelona Kontrak Adriano	uncategorized
Bayi Dibuang ke Bak Sampah	hukum dan kriminal
Bom Bunuh Diri Tewaskan 46 di Irak	bencana
Enam Keris Sakral Dicuri Maling	hukum dan kriminal
Golf Kadin Kumpulkan Rp2 Miliar Untuk Maluku	olahraga
Isu Teroris Masuk Desa Resahkan Kulon Progo	hukum dan kriminal
Kenaikan TDL Sudah Direvisi	uncategorized
KJRI Dubai Pulangkan 35 TKW Bermasalah	politik
Komisi IX DPR Bentuk Panja Perlindungan TKI	politik
LSM Nilai Materi UU SJSN Beratkan Warga	hukum dan kriminal
Marquez Catat Kemenangan Kelima di Jerman	olahraga
Muhaimin Iskandar Luncurkan Dua Buah Buku Sekaligus	politik
Organisasi Masyarakat Indonesia di Malaysia Jangan Terkesan Calo	politik
Pemerintah Segera Gelar Operasi Pasar	uncategorized
Pukat Harimau Marak di Perairan Sumbar	hukum dan kriminal
Reformasi Birokrasi Belum Memuaskan	teknologi
Upaya Konservasi Tidak Harus Berbentuk Kawasan Lindung	teknologi

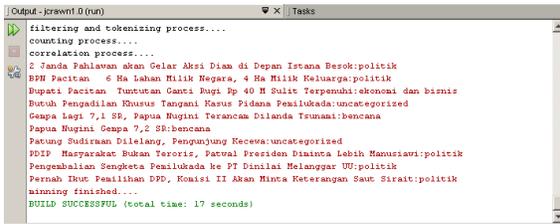
Dari hasil percobaan di atas, ada 4 berita yang tidak memiliki dan 2 berita yang salah dalam pengkategorian. Sehingga apabila dipersenkan error yang muncul pada percobaan ini adalah 30%. Kalau dicermati dari kasus – kasus sebelumnya, memang kamus untuk teknologi dan politik agak kurang. Sehingga ini menyebabkan nilai korelasi yang tidak pas.

4.2.2 Percobaan 2

Pada percobaan ini akan dilakukan ujicoba untuk beberapa data campuran dengan data kamus sesuai dengan data yang telah di update terakhir kali, threshold -0.55 dan jumlah pembatasan jumlah minimal > 0. Dan data campuran tersebut diambil ng dari situs di bawah ini pada tanggal 18 Juli 2010 jam 21.50

<http://rss.detik.com/index.php>

Berikut merupakan output dari percobaan 2 :



Gambar 4.2 Hasil percobaan 2

Agar lebih jelas maka hasil keluaran diletakkan pada sebuah table.

Tabel 4.2 Hasil percobaan 1

Judul Berita	Kategori
2 Janda Pahlawan akan Gelar Aksi Diam di Depan Istana Besok	politik
BPN Pacitan 6 Ha Lahan Milik Negara, 4 Ha Milik Keluarga	politik
Bupati Pacitan Tuntutan Ganti Rugi Rp 40 M Sulit Terpenuhi	ekonomi dan bisnis
Butuh Pengadilan Khusus Tangani Kasus Pidana Pemilukada	bencana
Gempa Lagi 7,1 SR, Papua Nugini Terancam Dilanda Tsunami	bencana
Papua Nugini Gempa 7,2 SR	bencana
Patung Sudirman Dilelang, Pengunjung Kecewa	politik
PDIP Masyarakat Bukan Teroris, Patwal Presiden Diminta Lebih Manusiawi	politik
Pengembalian Sengketa Pemilukada ke PT Dinilai Melanggar UU	politik
Pernah Ikut Pemilihan DPD, Komisi II Akan Minta Keterangan Saut Sirait	politik

Dari hasil percobaan di atas, ada 2 berita yang masuk pada kategori yang tidak benar. Sehingga apabila dipersenkan error yang muncul pada percobaan ini adalah 20%.

V. KESIMPULAN

Dari hasil percobaan dan analisa yang dilakukan, maka dapat diambil kesimpulan :

1. Pengaksesan berita pada server WML akan menjadi lebih cepat apabila dibandingkan

langsung ke situs aslinya karena sistem meminimalisir image yang ditampilkan pada server, sehingga *load* data dari server menuju klien menjadi lebih lancar.

2. Analisis korelasi yang digunakan mampu mengkategorikan berita yang didapat dengan nilai persen error 20 s.d 30.

VI. DAFTAR PUSTAKA

- [1]. Adrifina Arrummaisha, Utami P. Juwita, 2008, Pemilihan Artikel Berita dengan Text Mining, Seminar Ilmiah Nasional Komputer dan Sistem Intelijen, Auditorium Universitas Gunadarma, Depok.
- [2]. D.U Pawestri, Muslim Royyana, T.C Henning, 2009, Perancangan Dan Pembuatan Web Crawler Aplikasi Panduan Pembelian Spesifikasi Komputer Rakitan Online Dengan Memanfaatkan Google Gears, Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember, Surabaya.
- [3]. Lee Haesung, Kwon Joonhee, 2008, *Personalized RSS Search Service Using RSS Characteristics and User Context*, Proceedings of the International MultiConference of Engineers and Computer Scientists. Hong Kong.
- [4]. Muliantara Agus, Penerapan *Reguler Expression* dalam Melindungi Alamat Email dari Spam Robot pada Konten Wordpress, Program Studi Teknik Informatika, Jurusan Ilmu Komputer, Universitas Udayana, Bali.
- [5]. Nalendra Adiwana, 2008, Pembuatan Global Metadata Berdasarkan Wikipedia Menggunakan *Mathematical Model of Meaning*, Jurusan Teknologi Informasi, Politeknik Elektronika Negeri Surabaya.
- [6]. Triyo Machfudzain, 2009, Sistem Manajemen Infaq Berbasis WAP (studi kasus di ydsf), Jurusan Teknologi Informasi, Politeknik Elektronika Negeri Surabaya.