ISSN 2087-0256

smatika Jurnal

STIKI Informatika Jurnal

Volume 05, Nomor 02 Tahun 2015

ISSN 2087-0256

smatika Jurnal

STIKI Informatika Jurnal

Volume 05, Nomor 02 Tahun 2015

Lembaga Penelitian & Pengabdian Masyarakat
SEKOLAH TINGGI INFORMATIKA & KOMPUTER INDONESIA

PENGANTAR REDAKSI

STIKI Informatika Jurnal (SMATIKA Jurnal) merupakan jurnal yang diterbitkan oleh Lembaga Penelitian & Pengabdian kepada Masyarakat (LPPM), Sekolah Tinggi Informatika & Komputer Indonesia (STIKI) Malang.

Pada edisi ini, SMATIKA Jurnal menyajikan 14 (*empat belas*) naskah dalam bidang sistem informasi, jaringan, pemrograman web, perangkat bergerak dan sebagainya. Redaksi mengucapkan terima kasih dan selamat kepada Pemakalah yang diterima dan diterbitkan dalam edisi ini, karena telah memberikan kontribusi penting pada pengembangan ilmu dan teknologi.

Pada kesempatan ini, redaksi kembali mengundang dan memberi kesempatan kepada para Peneliti di bidang Teknologi Informasi untuk mempublikasikan hasil-hasil penelitiannya melalui jurnal ini. Bagi para pembaca yang berminat, Redaksi memberi kesempatan untuk berlangganan.

Akhirnya Redaksi berharap semoga artikel-artikel dalam jurnal ini bermanfaat bagi para pembaca khususnya dan bagi perkembangan ilmu dan teknologi di bidang Teknologi Informasi pada umumnya.

REDAKSI

ISSN 2087-0256

smatika Jurnal

STIKI Informatika Jurnal

Volume 05, Nomor 02 Tahun 2015

SEKRETARIAT
Lembaga Penelitian & Pengabdian kepada Masyarakat
Sekolah Tinggi Informatika & Komputer Indonesia (STIKI)
Malang
smatika Jurnal
Jl. Raya Tidar 100 Malang 65146
Tel. +62-341 560823
Fax. +62-341 562525
Website: jurnal.stiki.ac.id
E-mail: lppm@stiki.ac.id

ISSN 2087-0256
Volume 05, Nomor 02 Tahun 2015

DAFTAR ISI

Undangan Makalah

smatika Jurnal Volume 06, Nomor 01 Tahun 2016

Temu Kembali Informasi *Big Data* Menggunakan *K-Means Clustering*

Imam Marzuki
Program Studi Teknik Elektro, Universitas Panca Marga
Jln. Yos Sudarso 107 Pabean Dringu Probolinggo 67271, Indonesia
Email: imammarzuki32@gmail.com

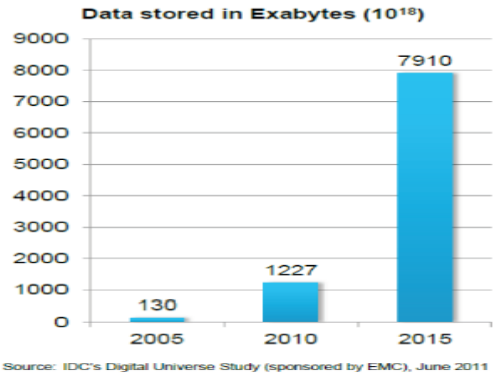
ABSTRAK

Saat ini manusia hidup selalu berhubungan dengan data, dimana data dibuat dan dikirimkan tiap detiknya di seluruh dunia. Hal ini menyebabkan data di jaringan bertambah secara massive (besar-besaran). Oleh karena itu kebutuhan akan pengelolaan data tersebut semakin meningkat. Salah satu bagian yang penting dalam pengelolaan data adalah proses pencarian informasi yang diinginkan oleh pengguna atau biasa disebut dengan temu kembali informasi (information retrieval). Tujuan utama dari temu kembali informasi adalah menemukan kembali dokumen yang berisi informasi yang relevan dengan query yang di inputkan oleh pengguna. *Sudah banyak metode yang diusulkan untuk temu kembali informasi. Namun dari sekian teknik masih menyisakan permasalahan terkait kecepatan dan akurasi pencarian. Pada tesis ini, penulis mengusulkan metode terbaik temu kembali informasi pada pencarian big data. Permasalahan muncul ketika melakukan proses pencarian informasi. Hal ini disebabkan big data didominasi oleh data tidak terstruktur. Data tidak terstruktur memiliki sifat sulit diorganisir. Oleh karena itu, diperlukan suatu teknik khusus untuk mengatasinya. Salah satu solusi untuk mengatasi permasalahan tersebut adalah dengan menambahkan klasterisasi dalam indexing informasi. Dalam penelitian ini digunakan metode klasterisasi menggunakan k-means clustering. Berdasarkan hasil percobaan menggunakan k-means clustering didapatkan nilai rata-rata precision 0.8 nilai rata-rata recall 0.741, dan nilai rata-rata waktu komputasi 0.579 detik.*

Kata kunci: *big data, temu kembali informasi, k-means clustering*

1. PENDAHULUAN

Perkembangan teknologi informasi yang semakin pesat telah menyebabkan aktivitas manusia tidak terlepas dari data. Setiap individu dapat membuat data dan mengirimkannya setiap detiknya. Hal ini menyebabkan data di jaringan bertambah secara *massive* (besar-besaran). Pertambahan data yang *massive* tersebut didominasi oleh data tidak terstruktur seperti teks, citra, audio, video, email, slide persentasi, animasi, dll. Data ini berasal dari berbagai macam sumber, misalnya dari jejaring sosial dan situs-situs portal berita. Bahkan penggunaan perangkat genggam untuk berkomunikasi sehari-hari membuat trafik data semakin membengkak. Tidak dapat dipungkiri lagi bahwa penyimpanan sebesar *petabyte* bahkan *exabyte* kerapkali dijumpai. Penggunaan database relasional sudah terbukti tidak mampu menangani *big data* (hariadi, 2013). Hal ini dikarenakan *big data* didominasi oleh data tidak terstruktur. Data tidak terstruktur memiliki karakteristik sulit diorganisir dan tidak memiliki struktur hirarki relasional. Pada gambar 1 diperlihatkan perkembangan data digital dalam rentang waktu 10 tahun terakhir.



Gambar 1. Perkembangan data digital dalam rentang waktu 10 tahun terakhir.

Perkembangan trafik data tersebut telah membawa manusia menuju era *big data*. Dalam pengertian teknis, *big data* didefinisikan sebagai sebuah problem domain dimana teknologi tradisional seperti database relasional tidak mampu lagi untuk melayani. *Big data* mempunyai tiga karakteristik yaitu volume, velositas, dan variasi datanya.

Peningkatan volume, velositas, dan variasi data banyak diakibatkan oleh adopsi internet dimana setiap individu memproduksi konten atau paling tidak meninggalkan sidik jari digital yang berpotensi digunakan untuk hal-hal baru.

Dalam Peningkatan data yang pesat ini telah membuat kebutuhan akan pengelolaan data semakin meningkat. Salah satu bagian yang penting dalam pengelolaan data adalah proses pencarian informasi yang diinginkan oleh pengguna atau biasa disebut dengan temu kembali informasi (*information retrieval*). Tujuan utama dari temu kembali informasi adalah menemukan kembali dokumen yang berisi informasi yang relevan dengan *query* yang di inputkan oleh pengguna.

Salah satu bagian penting dalam optimasi mesin pencari (*search engine optimization*) adalah mengoptimalkan proses temu kembali informasi yang diinginkan oleh pengguna. Sudah banyak teknik yang diusulkan untuk temu kembali informasi. Namun dari sekian teknik masih menyisakan persoalan terkait penanganan *big data*. Hal ini dikarenakan *big data* memiliki data tidak terstruktur dengan volume yang besar dan *outlier*.

Salah satu solusi untuk menangani permasalahan tersebut adalah dengan menambahkan klasterisasi dalam *indexing* informasi. Penambahan klasterisasi ini sebagian besar mengunakan metode hirarki dalam pengelompokannya seperti pada [2] dan [3]. Pada [2] dan [3] melakukan pencarian informasi dengan inputan berupa kata kunci dan penambahan klasterisasi menggunakan *centroid linkage hierarchical method*. Kekurangan muncul pada kompleksitas waktu komputasi dari metode hierarki ini meningkat ketika jumlah data semakin besar.

Baru-baru ini, [3] memperkenalkan sebuah teknik pencarian yang dapat meningkatkan peringkat hasil pencarian informasi dengan waktu yang cepat menggunakan *k-means clustering* yang merupakan klasterisasi dengan metode partisi. Namun belum dilakukan penelitian mendalam mengenai teknik ini.

Dalam penelitian ini sekumpulan *big data* berupa dokumen tidak terstruktur yang bersumber dari internet yang bervariasi jenisnya akan disesuaikan dengan kata kunci yang diinputkan melalui proses *text mining*. Selanjutnya hasil proses ini akan diklasterisasi menggunakan *k-means clustering.* Hasil akhirnya adalah sekumpulan dokumen tersebut akan ditampilkan lagi ke web yang diurutkan sesuai dengan tingkat relevansinya terhadap kata kunci.

2. TINJAUAN PUSTAKA

Dalam bagian ini akan dibahas mengenai teori-teori penting yang dapat menunjang dan menjadi acuan dalam perancangan penelitian. Bagian tersebut meliputi teori dasar mengenai *big data*, data tidak terstruktur, optimasi mesin pencari, temu kembali informasi (*information retrieval*), *text mining*, dan klasterisasi.

A. *Big Data*

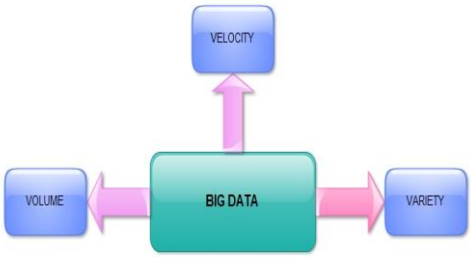
Dalam pengertian teknis, big data didefinisikan sebagai sebuah problem yang terjadi ketika teknologi tradisional seperti relasional database tidak mampu untuk memberikan layanan. Big data mempunyai tiga karakteristik yaitu volume, velositas, dan variasi datanya. Peningkatan volume, velositas, dan variasi data diakibatkan oleh adopsi internet. Selain itu, peningkatan tersebut juga diakibatkan oleh penggunaan perangkat genggam untuk berkomunikasi sehari-hari. Setiap individu memproduksi konten atau paling tidak meninggalkan sidik jari digital yang berpotensi digunakan untuk hal-hal baru. Beberapa prinsip dari big data adalah tidak membuang data apapun karena residu tersebut mungkin akan menjadi penting sejalannya waktu. Sedangkan untuk menghadapi variasi data yang tinggi, big data menciptakan struktur melalui ekstraksi, transformasi, tanpa harus membuang data mentah yang dimiliki.

Sejumlah data atau informasi dikatakan *big data* apabila memenuhi tiga karakteristik, antara lain :

1. *Volume*
   Ciri ini menandakan bahwa ukuran dan kapasitas data tersebut besar dan memungkinkan selalu bertambah seiring dengan pertambahan waktu. Dengan data yang semakin besar merupakan tantangan bagi media penyimpanan.
2. *Velocity*
   Pengolahan data maupun pemrosesan informasi memerlukan kecepatan (*real time processing*).
3. *Variety*
   Data atau informasi bervariasi jenisnya baik data terstruktur maupun tidak terstruktur. Data terstruktur adalah data yang mudah dianalisa menggunakan database relasional. Sedang data tidak terstruktur tidak

bisa diolah menggunakan database relasional. Big data didominasi oleh data tidak terstruktur.

Gambar 2 berikut ini mengilustrasikan mengenai karakteristik dari *big data.*



Gambar 2. Karakteristik *big data*

B. Data Tidak Terstruktur

Salah satu tantangan dalam pengolahan *big data* adalah data tidak terstruktur dimana tidak memiliki hirarki relasional dan tidak cocok dengan database tradisional seperti *Relational Database Management System* (RDBMS).

Beberapa karakteristik dari data tidak terstruktur, antara lain sebagai berikut :

1. Berisikan obyek atau dokumen baik ukuran maupun tipe datanya bebas.
2. Tidak terorganisir.
3. Organisasi dan informasi tidak konsisten.
4. Berisikan teks, image, audio, video, email dan persentasi powerpoint.
5. Data yang ditampilkan pada halaman web.

C. Optimasi Mesin Pencari (*Search Engine Optimization*)

Sejalan dengan adanya fenomena big data yang didominasi data tidak terstruktur, maka diperlukan sebuah teknik pencarian informasi (information retrieval) yang efisien. Teknik tersebut dinamakan dengan optimasi mesin pencari (search engine optimization). Optimasi mesin pencari mengacu pada peningkatan kinerja mesin pencari dalam hal akurasi dan kecepatan. Tujuan dari optimasi mesin pencari adalah menempatkan sebuah situs web pada posisi teratas, atau setidaknya halaman pertama hasil pencarian berdasarkan kata kunci tertentu yang ditargetkan. Secara logis, situs web yang menempati posisi teratas pada hasil pencarian memiliki peluang lebih besar untuk mendapatkan pengunjung.

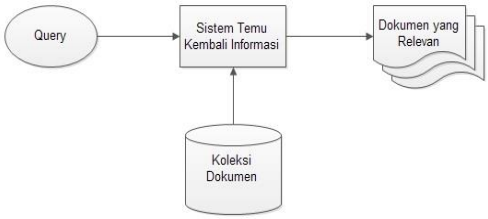
Mesin pencari telah menjadi bagian yang paling dominan dalam hidup pengguna internet. Dengan makin berkembangnya pemanfaatan jaringan internet sebagai media bisnis, kebutuhan akan optimasi mesin pencari (*search engine optimization*) juga semakin meningkat. Berada pada posisi teratas hasil pencarian akan meningkatkan peluang sebuah perusahaan pemasaran berbasis web untuk mendapatkan pelanggan baru. Peluang ini dimanfaatkan sejumlah pihak untuk menawarkan jasa optimasi mesin pencari bagi perusahaan-perusahaan yang memiliki basis usaha di internet.

Mesin pencari telah menjadi bagian yang paling dominan dalam hidup pengguna internet. Dengan makin berkembangnya pemanfaatan jaringan internet sebagai media bisnis, kebutuhan akan optimasi mesin pencari (*search engine optimization*) juga semakin meningkat. Berada pada posisi teratas hasil pencarian akan meningkatkan peluang sebuah perusahaan pemasaran berbasis web untuk mendapatkan pelanggan baru. Peluang ini dimanfaatkan sejumlah pihak untuk menawarkan jasa optimasi mesin pencari bagi perusahaan-perusahaan yang memiliki basis usaha di internet.

D. Temu Kembali Informasi *(Information Retrieval)*

Temu kembali informasi merupakan bagian dari computer science yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Berdasarkan referensi dijelaskan bahwa temu kembali informasi (*information retrieval*) merupakan suatu pencarian informasi yang didasarkan pada suatu kata kunci yang diharapkan dapat memenuhi keinginan user dari kumpulan dokumen yang ada. Selain itu referensi lain menyebutkan bahwa temu kembali informasi merupakan studi tentang sistem pengindeksan, pencarian, dan mengingat data, khususnya teks atau bentuk tidak terstruktur lainnya.

Informasi atau data yang dicari dapat berupa berupa teks, image, audio, video dan lain-lain. Koleksi data teks yang dapat dijadikan sumber pencarian juga dapat berupa pesan teks, seperti e-mail, fax, dan dokumen berita, bahkan dokumen yang beredar di internet. Dengan jumlah dokumen koleksi yang besar sebagai sumber pencarian, maka dibutuhkan suatu sistem yang dapat membantu user menemukan dokumen yang relevan dalam waktu yang singkat dan tepat.



Gambar 3. Ilustrasi proses temu kembali informasi

E. Text Mining

Definisi dari *text mining* sudah sering diberikan oleh banyak ahli riset dan praktisi. Seperti halnya *data mining*, *text mining* adalah proses penemuan akan informasi yang sebelumnya tidak terungkap dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan data teks tidak terstruktur, *text mining* mencoba untuk mengasosiasikan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu. Hasil yang diharapkan adalah informasi baru yang tidak terungkap jelas sebelumnya.

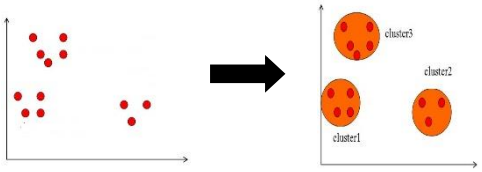
Seperti halnya *data mining*, *text mining* juga menghadapi masalah yang sama, termasuk jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah. Berbeda dengan *data mining* yang utamanya memproses data terstruktur, data yang digunakan *text mining* pada umumnya dalam bentuk tidak terstruktur. Akibatnya, *text mining* mempunyai tantangan tambahan yang tidak ditemui di *data mining*, seperti struktur teks yang kompleks dan tidak lengkap, arti yang tidak jelas dan tidak standard, dan bahasa yang berbeda ditambah *translasi* yang tidak akurat. Dikarenakan struktur data ditujukan agar mudah di proses komputer secara otomatis, *pre-processing* data di *data mining* jauh lebih mudah dilakukan daripada di *text mining*.

Teks diciptakan bukan untuk digunakan oleh mesin, tapi untuk dikonsumsi manusia langsung. Text mining telah mengadopsi teknik yang digunakan di bidang natural language processing dan computational linguistics. Walaupun teknik di computational linguistics bisa dibilang maju dan cukup akurat untuk mengekstrak informasi, tujuan text mining bukan hanya mengekstrak informasi. Melainkan untuk menemukan pola dan informasi baru yang belum terungkap.

F. Klasterisasi

Klaster adalah suatu kumpulan objek atau data yg mimiliki kesamaan diantara mereka dan data yg tidak mmiliki kesamaan dimasukkan kedalam klaster lain. Sedangkan klasterisasi proses pengelompokan objek atau data kedalam grup yang anggotanya memiliki kesamaan tertentu. Klasterisasi merupakan metode penganalisaan data, yang sering dimasukkan sebagai salah satu metode *data mining*, yang tujuannya adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu wilayah yang sama dan data dengan karakteristik yang berbeda ke wilayah yang lain.

Gambar 3 adalah contoh klasterisasi:



Gambar 4. Klasterisasi berdasarkan kedekatan jarak

Gambar 4 diatas adalah gambar klasterisasi data dengan menggunakan kedekatan jarak sebagai parameternya. Data-data yang jaraknya saling berdekatan akan bergabung sebagai anggota dari klaster.

3. METODOLOGI PENELITIAN

Pada bagian ini akan dibahas mengenai perancangan sistem dan implementasi. Yang di dalamnya terdapat blok diagram alur sistem secara keseluruhan dan penjelasan proses-proses secara detail yang diilustrasikan dengan *flowchart*.

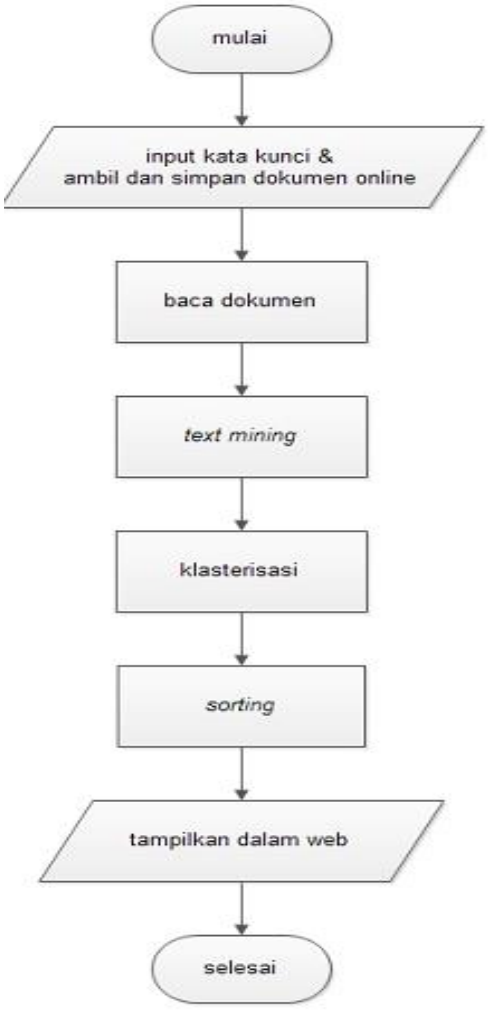
A. Gambaran Umum

Penggunaan *k-means clustering* dalam penelitian ini untuk mengelompokkan *big data* berupa dokumen yang berisi informasi tertentu dari kumpulan dokumen yang ada pada beberapa alamat web. Kemudian akan dilakukan pencocokan jumlah kata kunci hasil *text mining* dalam tiap-tiap dokumen. Dari masing-masing jumlah tersebut kemudian dilakukan pengklasteran dengan *k-means clustering* terhadap koordinat titik yang menunjukkan jumlah kata kunci dari masing-masing dokumen. Kemudian dihitung nilai setiap klaster. Setiap klaster dihitung nilainya menggunakan perkalian matrik. Selanjutnya dibandingkan nilai antar setiap klaster. Klaster dokumen yang mempunyai jumlah nilai paling tinggi merupakan klaster dokumen yang terpilih sebagai hasil pencarian. Untuk

menampilkan hasilnya dalam halaman web yaitu dengan mencari nilai setiap dokumen dengan cara mengalikan setiap matrik jumlah kata kunci dari setiap dokumen dengan matrik transposnya. Kemudian menampilkan hasilnya dengan urutan dokumen yang mempunyai nilai yang paling besar sampai paling kecil.

B. Desain Sistem

Dalam sub bagian ini akan digambarkan desain sistem dari penelitian ini. Input dari penelitian ini merupakan kata kunci yang dimasukkan oleh pengguna dan data yang diambil dari internet pada saat itu juga. Sedangkan output dari penelitian ini adalah beberapa dokumen yang diperingkat menurut relevansinya. Gambar 5 berikut menjelaskan alur proses dari sistem.



Gambar 5. Diagram alur desain sistem

C. Evaluasi Hasil

Untuk mengevaluasi hasil pencarian, diperlukan suatu pengukuran. Pada Sub bab ini menjelaskan cara mengukur hasil pencarian. Pengukuran mengacu pada optimasi mesin pencari yaitu peningkatan kinerja mesin pencari dalam hal akurasi dan kecepatan. Untuk pengukuran akurasi ditinjau dari nilai precision dan recall dari sistem. Semakin tinggi nilai precision dan recall, maka semakin tinggi pula tingkat akurasi. Sebaliknya semakin rendah nilai precision dan recall, maka semakin rendah pula tingkat akurasi. Atau dengan kata lain akurasi berbanding lurus dengan precision dan recall. Sedangkan untuk kecepatan ditinjau dari nilai waktu komputasi. Semakin tinggi nilai waktu komputasi, maka semakin rendah tingkat kecepatan. Sebaliknya semakin rendah nilai waktu komputasi maka semakin tinggi kecepatan. Atau dengan kata lain kecepatan berbanding terbalik dengan waktu komputasi.

Dengan demikian, pengukuran yang digunakan dalam penelitian ini yaitu precision recall, dan waktu komputasi.

1. *Precision*
   Untuk *Precision* adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. *Precision* merupakan perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen yang terambil oleh sistem baik relevan maupun tidak relevan. Menurut pengertian ini dapat dinyatakan dalam persamaan (1).

$$P = \frac{TP}{TP + FP} \quad \text{.....} \quad (1)$$

   Dimana :
   P (Precision) = Tingkat *precision* pencarian
   TP (True Positive) = Dokumen relevan yang ditemukan
   FP (False Positive) = Dokumen tidak relevan yang ditemukan
   TN (True Negative) = Dokumen relevan yang ditemukan
   FN (False Negative) = Dokumen tidak relevan yang tidak ditemukan

2. *Recall*
   Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. *Recall* merupakan perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah

seluruh dokumen relevan yang ada dalam koleksi dokumen (terambil ataupun tak terambil sistem). Menurut pengertian ini dapat dinyatakan dalam persamaan (2).

$$R = \frac{TP}{TP + FN} \quad \text{.....} \quad (2)$$

Dimana :
R (Recall) = Tingkat *recall*

3. Waktu Komputasi
Pengukuran ini dilakukan dengan cara mengetahui waktu tempuh yang terjadi ketika proses *text mining* berjalan hingga menampilkan hasil klasterisasi.

4. HASIL DAN PEMBAHASAN

Pada bagian ini akan dipaparkan hasil dari uji coba sistem dan analisanya. Pada pengujian akan diterapkan penambahan klasterisasi menggunakan *k-means clustering* yang telah dibahas terhadap 80% dokumen yang relevan dengan kata kunci "gol ronaldo" dan 20% tidak relevan dengan sejumlah dokumen yang berbeda.

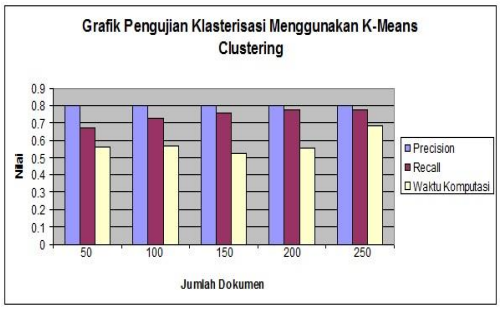
Pada pengujian akan dilakukan evaluasi keberhasilan sistem dengan penambahan klasterisasi menggunakan *k-means clustering*. Berdasarkan persamaan (1) dan persamaan (2) didapatkan nilai precision dan nilai recall. Sedangkan waktu komputasi dihitung berdasarkan waktu proses yang diperlukan komputer antara mengetikkan kata kunci yang diminta sampai menampilkan hasil pencarian. Pengujian dilakukan pada sejumlah dokumen yang berbeda. Dari hasil pengujian tersebut dapat dituliskan dalam tabel berikut.

Dari hasil pengujian tersebut dapat dituliskan dalam tabel 1.

Tabel 1. Hasil pengujian menggunakan klasterisasi *k-means Clustering*

Jumlah Dokumen	*Precision*	*Recall*	Waktu Komputasi
50	0.8	0.673	0.563 detik
100	0.8	0.727	0.565 detik
200	0.8	0.754	0.527 detik
400	0.8	0.777	0.553 detik
800	0.8	0.776	0.685 detik
Rata-rata	0.8	0.741	0.579 detik

Tabel 1 didapatkan melalui perhitungan *precision*, *recall* dan waktu komputasi pada sejumlah dokumen yang berbeda. Dari tabel tersebut didapatkan nilai rata-rata *precision* 0.8, nilai rata-rata *recall* 0.857, dan nilai rata-rata waktu komputasi 0.958 detik. Berdasarkan perhitungan tersebut juga dapat digambarkan dalam grafik gambar 6.



Gambar 6. Grafik pengujian menggunakan *k-means Clustering*

Gambar 6 adalah grafik hasil perhitungan *precision*, *recall*, dan waktu komputasi penambahan klasterisasi dalam *indexing* informasi menggunakan *k-means clustering*. Dari sejumlah dokumen yang berbeda terlihat bahwa nilai *precision* menunjukkan angka tetap. Namun nilai *recall* dan waktu komputasi bervariasi seiring dengan bertambahnya jumlah dokumen. Variasi nilai *recall* dan waktu komputasi ini disebabkan oleh penentuan centroid awal secara random.

5. KESIMPULAN

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

- Salah satu solusi untuk menangani *big data* yang terdiri dari data tidak terstruktur yang besar dan *outlier* adalah dengan menambahkan klasterisasi dalam *indexing* informasi.
- Berdasarkan hasil percobaan menggunakan *k-means clustering* didapatkan nilai rata-rata precision 0.8

nilai rata-rata recall 0.741, dan nilai rata-rata waktu komputasi 0.579 detik.

6. REFERENSI

[1] A.R. Barakbah, K. Arai, "A New Algorithm For Optimization Of K-Means Clustering With Determining Maximum Distance Between Centroids", *In. IES 2006, Politeknik Elektronika Negeri Surabaya, ITS.*

[2] A.S.N.Chakravarthy, Deepthi.S, K.Satyatej, Sk.Nizmi, S.Sindhura, "Document Clustering in Web Search Engine", *International Journal of Computer Trends and Technology- volume 3 Issue 2- 2012.*

[3] Damayanti, Nadia, "Temu Kembali Informasi Teks Berdasarkan Lokasi Pada Dokumen Yang Dikelompokkan Menggunakan Metode Centroid Linkage Hierarchical Method", *Tugas Akhir D4, Politeknik Elektronika Negeri Surabaya, 2011.*

[4] E. Martiana, N. R. Mubtada, U. Agusetia, "Mesin Pencari Dokumen Dengan Pengklasteran Secara Otomatis", *Politeknik Elektronika Negeri Surabaya, Telkomnika Vol. 8, No. 1, April 2010 : 41 - 48.*

[5] Hariadi, Mochammad, "Big Data From Infrastructured To Analytics", *Telematics Lab, Multimedia Network Dept. ITS,2013.*

[6] K. Supreet, K. Usvir, "An Optimizing Technique for Weighted Page Rank With K-Means Clustering", *International Journal of Advanced Research in Computer Science and Software Engineering, July 2013.*

[7] K.K. Kattamuri, R. Chiramdasu, "Search Engine With Parallel Processing And Incremental K-Means For Fast Search And Retrieval", *International Journal of Advances in Engineering & Technology, Jan. 2013*

[8] R. Vadivel , K. Baskaran, " Enrich the E-publishing Community Website with Search Engine Optimization Technique", *IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011.*

[9] Y.SureshBabu, K.Venkat Mutyalu, Y.A.Siva Prasad, "A Relevant Document Information Clustering Algorithm for Web Search Engine", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 8, October 2012.*

Biografi Singkat Penulis

Imam Marzuki, S.ST, M.T lahir di Probolinggo Tahun 1987. Tahun 2005 lulus dari SMAN 1 Gending. Tahun 2012 menyelesaikan studi D4 dari Politeknik Elektronika Negeri Surabaya dan tahun 2014 menyelesaikan studi S2 dari Teknik Elektro Institut Teknologi Sepuluh Nopember dalam Bidang Telematika. Setelah memperoleh gelar Magister aktif sebagai staf pengajar di Universitas Panca Marga Program Studi Teknik Elektro Konsentrasi Informatika dan Komputer. Selain itu juga aktif sebagai staf pengajar di Akademi Manajemen Informatika dan Komputer Taruna Program Studi Teknik Komputer.