Jurnal SimanteC
Vol. 4, No. 3 Juni 2015 ISSN 2088-213

Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma K-Means

Husni, Yudha Dwi Putra Negara, M. Syarief

Jurusan teknik Informatika, Fakutas Teknik, Universitas Trunojoyo
Jl. Raya Telang PO. BOX 2 Kamal, Bangkalan, Madura, 691962
Email: husni@if.trunojoyo.ac.id , yudhabary@yahoo.com, ayibnya@gmail.com

Abstrak

Informasi yang tersedia pada halaman-halaman web trunojoyo.ac.id semakin besar, belum tertata dengan baik, belum terstruktur atau terkategori mengikuti kaidah tertentu dan tersebar pada banyak sub-domain. Sejauh ini, tidak ada gerbangatau portal web yang menyediakan akses ke berbagai situs webyang dihosting oleh data center PTIK Universitas Trunojoyo. Salah satu masalah yang telah diselesaikan adalah pengelompokan informasi atau berita web tersebut secara otomatis menggunakan algoritma clustering K-Means. Search engine RISE yang telah berjalan menghimpun semua halaman web yang ditulis dalam bahasa Indonesia di bawah domain trunojoyo.ac.id menggunakan teknik crawling. Halaman-halaman tersebut kemudian di*pre-processing* menggunakan teknik standar dalam text minig (informationm retrieval). Proses utamanya adalah penerapan teknik k-menas sehingga terbentuk kelompok-kelompok berita otonom. Pengujian yang telah dilakukan menunjukkan bahwa teknik *clustering* yang diterapkan mampu bekerja dengan baik dan memberikan akurasi yang memuaskan. Ada sekitar 300 halaman web yang dilibatkan dalam proses clustering dimana diperoleh ukuran rata-rata *F-Measure* sebesar 0.6129192 dan *Purity* bernilai 0.67294195. Faktor yang cukup berpengaruh dalam clustering dan klasifikasi teks bahasa Indonesia adalah fase pre-processing, terutama pada pendekatan stemming. Perbaikan terhadap teknik stemming diyakini akan meningkatkan akurasi pengelompokan dokumen.

Kata Kunci : *Clustering, K-Means, F-Measure, Purity*

Abstract

The information available on the web pages trunojoyo.ac.id getting bigger, not well ordered, yet structured or terkategori follow certain rules and scattered in many sub-domains. So far, no gerbangatau web portal that provides access to a variety of sites hosted by the data center webyang PTIK Trunojoyo University. One problem that has been solved is the grouping of information or news Web site automatically using the K-Means clustering algorithm. RISE search engines that have been running together all the web pages are written in Indonesian under trunojoyo.ac.id domain using crawling techniques. The pages are then dipre-processing using standard techniques in text Minig (informationm retrieval). The main process is the application of K-menas technique to form groups of autonomous news. Tests have shown that the clustering technique applied is able to work well and give satisfactory accuracy. There are about 300 web pages that are involved in the process of clustering which gained an average size of F-Measure 0.67294195 and Purity 0.6129192. Factors influential in clustering and classification Indonesian text is pre-processing phase, especially on the stemming approach. Repairs to stemming technique is believed to improve the accuracy of the document grouping.

Keywords : *Clustering, K-Means, F-Measure, Purity*

1. Pendahuluan

Hampir semua universitas atau perguruan tinggi di Indonesia mempunyai situs, termasuk Universitas Trunojoyo Madura. Jumlah halaman web yang tersebar pada banyak situs web di bawah domain trunojoyo.ac.id semakin besar, tidak tertata atau terkelompok mengikuti kaidah tertentu dan tidak ada portal atau gerbang masuk untuk semua situs web tersebut. Hal ini tentu mempersulit pengguna menemukan informasi tertentu yang diharapkan. Sangat tidak mungkin jika setiap pengguna harus mengetahui dan menghapal URL setiap situs web. Search engine khusus bahasa Indonesia RISE hadir untuk menghimpun semua informasi yang tersebar pada banyak halaman web tersebut. Aplikasi berupa portal ini dilengkapi program crawler yang berjalan mengililingi halaman web dan menghimpun semua halaman web berbahasa Indonesia yang di*hosting* pada data center Universitas Trunojoyo Madura.

Masalah lain yang perlu penyelesaian adalah beragamnya bahasa atau topik dari halaman web. Crawler telah menghimpun banyak halaman web tetapi bagaimana pengguna mendapatkan informasi khusus, misalnya tentang beasiswa atau penerimaan mahasiswa baru? Halaman-halaman tersebut harus dikelompokkan. Ada 2 pendekatan pengelompokan data, yaitu klasifikasi dan clustering [1]. Klasifikasi adalah mengelompokkan data ke dalam kelas-kelas yang telah ditentukan sebelumnya, artinya nama dan jumlah kelas sudah terdefinisi sebelum proses klasifikasi dilakukan. Pendekatan berbeda dilakukan pada clustering. Tidak ada nama kelas atau batasan jumlah yang ditentukan. Halaman-halaman web seolah diperintahkan untuk mencari teman terdekatnya dan membuat kelompok sendiri. Penyederhanaan dapat dilakukan setelah proses clustering selesai, misalnya dengan melakukan clustering lanjutan, menggabungkan beberapa kelas yang mirip, atau menghapus kelas yang hanya beranggotakan satu halaman web.

Tulisan ini mencoba melaporkan hasil penelitian yang kami lakukan, yaitu clusterisasi terhadap halaman-halaman web yang telah dihimpun dalam Search Engine. Algoritma yang digunakan adalah k-means karena beberapa hasil penelitian sebelumnya memperlihatkan bahwa k-means mampu memberikan akurasi terbaik dalam pengelompokan teks [2]. Bagian berikutnya dari tulisan ini akan menjelaskan konsep information retrieval, clustering k-means, implementasi dan pengujian yang telah dilakukan dan terakhir kesimpulan yang diperoleh.

2. Temu-Kembali Informasi

Information Retrieval (IR) atau temu-kembali informasi adalah ilmu yang mempelajari tentang cara mendapatkan kembali informasi yang pernah tersedia. IR berisi tindakan, metode dan prosedur untuk menemukan kembali data yang tersimpan untuk menyediakan informasi mengenai subyek yang dibutuhkan. Tindakan tersebut mencakup *text indexing, inquiry analysis*, dan *relevance analysis* [11].

Menurut [4] terdapat 5 langkah pembangunan *inverted index*, yaitu:

1. Penghapusan format dan *markup* dari dalam dokumen (halaman web)
2. Pemisahan rangkaian term (*tokenization*). Term biasanya berupa kata atau frasa di di dalam dokumen. Namun, kata-kata yang tidak memberikan perbedaan seperti ini, itu, saya, kamu, serta tanda-tanda baca dihilangkan (tidak dianggap sebagai term).
3. Pengembalian term ke bentuk akar kata (*stemming*) atau bentuk umum yang disepakati.
4. Pemberian bobot terhadap term (*weighting*) dengan memberlakukan

kombinasi perkalian bobot lokal *term frequency* dan bobot global *inverse document frequency*, ditulis *tf .idf*.

5. Menyimpan term yang diperoleh disertai oleh nomor dokumen dimana term tersebut muncul dan jumlah kemunculannya. Daftar term ini dinamakan index atau inverted Index.

Index tersebut selanjutnya digunakan oleh berbagai metode information retrieval, seperti asosiasi, klasifikasi dan clustering untuk menemukan suatu kesimpulan mengenai suatu himpunan dokumen.

3. Algoritman *Clustering* K-MEANS

Algoritma yang paling umum digunakan dalam *clustering* yaitu algoritma K-Means. Algoritma ini populer karena mudah diimplementasikan dan kompleksitas waktunya linear. Kelemahannya adalah algoritma ini sensitif terhadap inisialisasi *cluster* .

Dasar algoritmanya adalah sebagai berikut:

1) Inisialisasi *cluster*
2) Masukkan setiap dokumen ke *cluster* yang paling cocok berdasarkan ukuran kedekatan dengan centroid. Centroid adalah vektor term yang dianggap sebagai titik tengah *cluster*. Persamaan kedekatannya sebagai berikut

$$d(P,Q)=\sqrt{\sum_{j=1}^{p}(x_j(P)-x_j(Q))^2} \quad (1)$$

3) Setelah semua dokumen masuk ke *cluster*. Hitung ulang *centroid cluster* berdasarkan dokumen yang berada di dalam *cluster* tersebut dengan mencari rata-rata centroid tiap *cluster*.
4) Jika centroid tidak berubah (dengan treshold tertentu) maka stop. Jika tidak, kembali ke langkah 2.

4. Rancangan Sistem

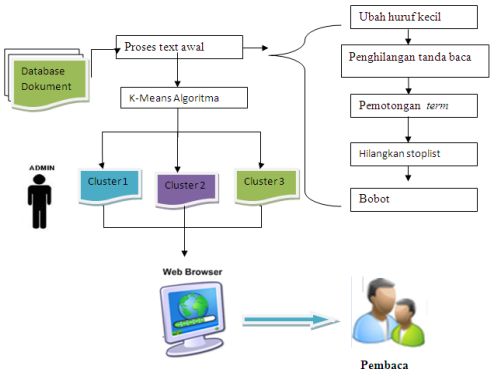
Metode yang digunakan pada sistem ini adalah bagian dari *text mining.* Tahapan dalam pengelompokan berita terdari dari 2 tahapan utama. Tahap pertama yaitu *pre-processing* atau *text-processing* terhadap himpunan dokumen berita, sedangkan tahap kedua yaitu proses pengkelompokan berita berdasarkan bobot yang telah diketahui dengan menggunakan metode *K-means Clustering*. Akan ada pekerjaan lanjutan yaitu analisis terhadap data yang sudah dikelompokkan sehingga berita atau dokumen web dapat dengan mudah diakses oleh *user.*

Rancangan sistem *Clustering* dokumen berita ini dalam penelitian ini dapat terlihat pada Gambar 1. dapat dijabarkan proses-proses yang terjadi di dalam sistem adalah sebagai berikut

1) Pembaca, adalah pembaca yang akan melihat informasi.
2) *Admin*, yang berhak melakukan *update* dan kontrol pada *database*.
3) *Database dokumen*, sebagai tempat penyimpanan *URL* web portal berita dan data penting lainnya, seperti konfigurasi web portal, dan eleman dari berita yang ada.
4) *Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.
5) *Tokenizing*, proses pemenggalan kata pada dokumen berdasarkan *spasi* dan tanda – (penghubung).
6) *Filtering*, proses penghilangan kata-kata (yang dianggap) sebagai kata yang jarang dicari atau jarang digunakan sebagai *keywords* pada

proses pencarian.

7) *Remove stoplist*, proses penghapusan kata yang dianggap tidak penting seperti dan, tidak, yang, dan lain-lain.
8) *K-Means* adalah sebuah algoritma *Clustering* dokumen text yaitu dengan mengelompokkan n buah objek ke dalam k kelas berdasarkan jaraknya dengan pusat kelas.
9) *Weighting*, proses pembobotan yang nanti akan di inputkan pada metode *K-Means*.



Gambar 1. Proses terhadap dokumen yang telah dihimpun pada Search Engine, salah satunya clustering

Pada tahap *pre-processing*, proses-proses yang dilakukan adalah:

1) Sistem membaca dokumen berita. Dokumen berita disimpan kedalam RDBMS yang terdiri dari satu tabel yaitu, misalnya tabel tb_raw. Strukturnya diperlihatkan pada Gambar 2.

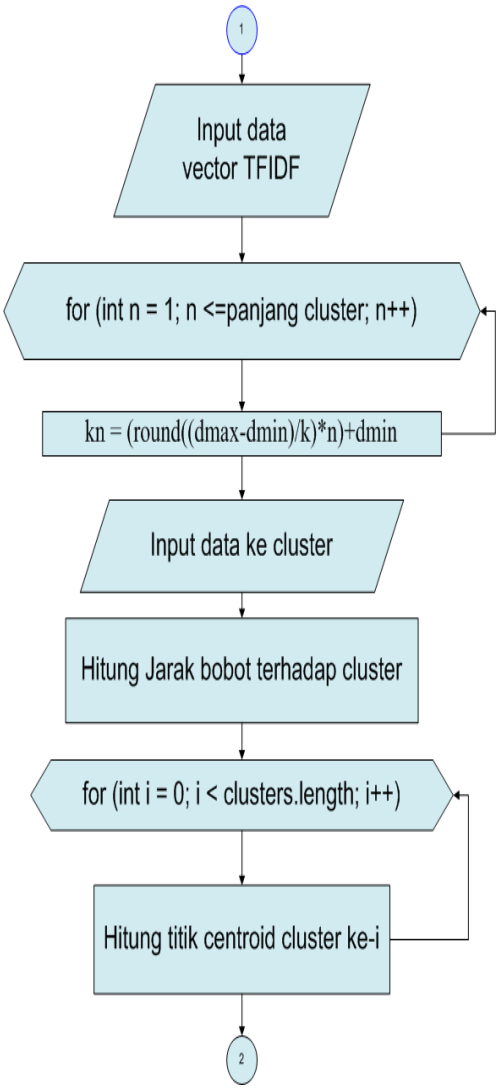
Field	Type	Comment
id_raw	int(50)	
raw	varchar(500)	
vecidf	double	
dokumen	varchar(15)	
cluster	int(50)	
tanggal	date	
link	varchar(100)	
counter	int(10)	
title	varchar(50)	

Gambar 2. Contoh struktur data untuk menyimpan dokumen web yang dikelompokkan

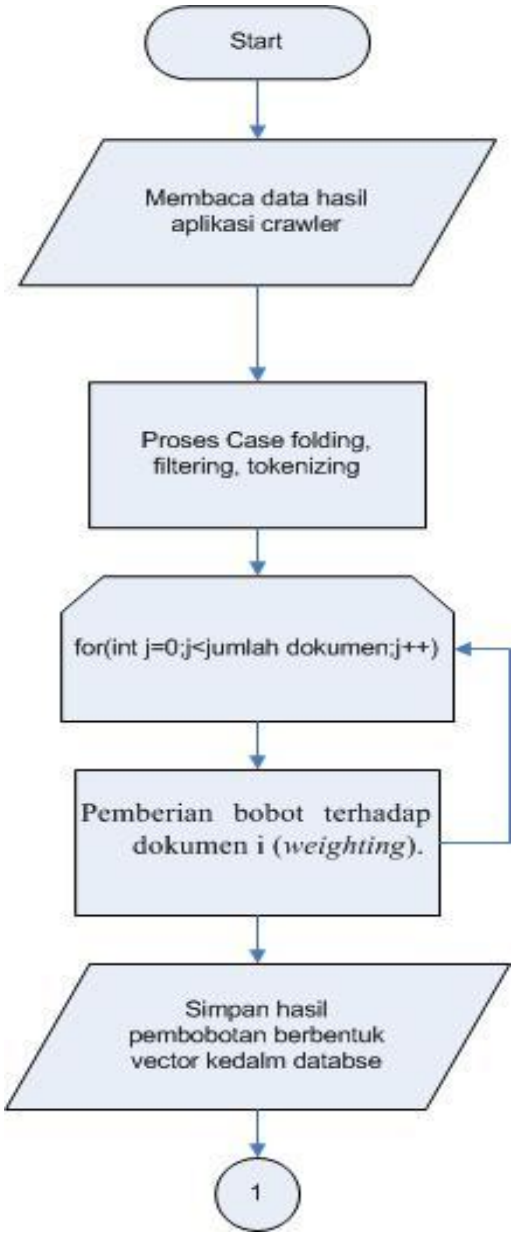
2) Pembacaan dokumen menggunakan aplikasi crawler karena dengan aplikasi crawler dapat mempercepat pengambilan data dengan jumlah yang banyak dalam waktu yang singkat.
3) Kemudian sistem, teks berita tersebut dilakukan proses *filtering* yaitu hanya mengambil kata yang bermakna atau menghilangkan tanda baca seperti titik, koma dan lain-lain karena tanda tersebut tidak digunakan dalam proses pembobotan.
4) Proses selanjutnya adala*h case folding* yaitu mengubah semua dokumen berita menjadi huruf kecil sehingga sistem membaca sama pada semua kata.
5) Proses berikutnya adalah *tokenizing* yaitu pemotongan dokumen sehingga menjadi *term.* Fungsi dari pemotongan kata supaya kata dapat dihitung frekuensinya sehingga dapat diketahui bobot dokumen berita berdasarkan kesamaan kata. Dalam penelitian pembobotan mengunaka metode TF-IDF (*term frequency–inverse document frequency*) karena pembobotan paling baik adalah berdasarkan frekuensi kata yang sama yang diterapkan pada metode TF-IDF. Penggunaan *TF-IDF* tanpa *stemming* menghasilkan kualitas *cluster* terbaik pada *K-Means Clustering*, Sehingga pada penelitian ini tidak menggunakan *stemming* kata dasar.
6) *Term* yang dihasilkan dihitung frekuensi supaya diketahui tingkat kesamaan berita dan dibobotkan dengan menggunakan metode *TF-IDF* terhadap semua dokumen.
7) Kemudian bobot hasil TF-IDF akan simpan kedalan RDBMS.

8) Tahap selanjutnya adalah clustering dokumen berita dengan menggunakan *K-Means*. Hasil dari pembobotan dokumen di inputkan kedalam algoritma K-means Sehingga dapat ter*cluster*.

Diagram alir dari sistem clustering k-means ini secara lengkap diperlihatkan pada Gambar 3.



Gambar 3a. Diagram alir dari sistem clustering k-means (bagian 1)

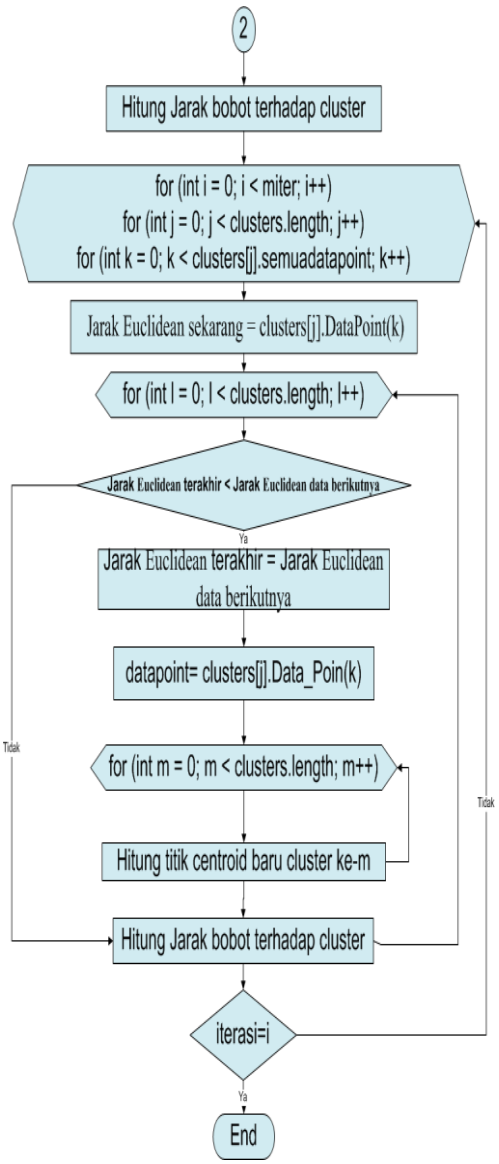


Gambar 3b. Diagram alir dari sistem clustering k-means (bagian 2)

5. Uji Coba

Gambar 4 memperlihatkan contoh hasil clustering untuk membentuk kelompok-kelompok halaman web. Hasil dari *clustering* dokumen akan disimpan kedalam database, kemudian ditampilkan dalam sebuah *web* disertai dengan berbagai macam fasilitas untuk memudahkan mencari berita seperti berita terbaru dan berita terpopuler.

Pada pengujian ini, digunakan acuan data berupa sekumpulan dokumen jumlahnya 300 dokumen berita yang telah diklasifikasi atau dikelompokan secara manual bedasarkan kemiripan berita. Kemudian hasil *cluster* dengan *K-Means* dihitung nilai *F-Measure* dan *Purity* berdasarkan kelas-kelas dokumen tersebut berdasarkan pengelompokan secara manual oleh 3 orang berbeda.

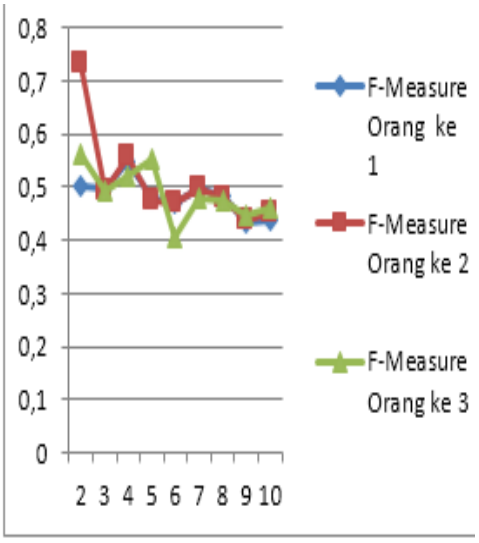


Gambar 3c. Diagram alir dari sistem clustering k-means (bagian 3)

Berdasarkan perhitungan analisa menggunakan *F-Measure* untuk menghitung tingkat akurasi dan *Purity* untuk menghitung tingkat kemurnian *cluster* maka diperoleh nilai *F-Measure* dan nilai *Purity* untuk jumlah *cluster* = 2 sampai *cluster* = 10 . Sehingga diperoleh nilai *F-Measure* pada tabel 1 dan nilai *Purity* hasil perhitungan analisis uji coba pada tabel 2

Tabel 1. Nilai *F-Measure* dari 3 Orang Berbeda

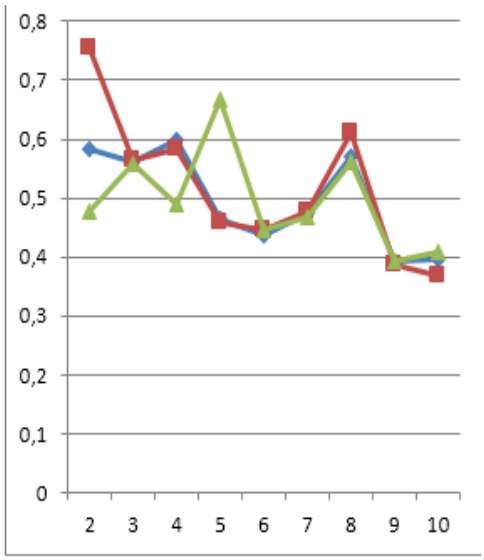
Jumlah Cluster	F-Measure		
	Orang ke 1	Orang ke 2	Orang ke 3
2	0.50367212	0.730666369	0.560351689
3	0.49637684	0.495027054	0.493931541
4	0.54773972	0.556924545	0.521142812
5	0.47948964	0.475916868	0.551235082
6	0.46997678	0.467789437	0.405658036
7	0.49715791	0.497338831	0.479508719
8	0.48304734	0.480924474	0.475128531
9	0.43518951	0.43825144	0.445679431
10	0.43859815	0.450356173	0.462563857



Gambar 4. Grafik *F-Measure* dari 3 Orang Berbeda

Tabel 2. Nilai Purity dari 3 Orang Berbeda

Jumlah Cluster	Putitas		
	Orang ke 1	Orang ke 2	Orang ke 3
2	0.583333	0.7547582	0.47758795
3	0.561528	0.5647739	0.55882546
4	0.597513	0.5837604	0.49031643
5	0.465909	0.459294	0.66655459
6	0.437848	0.446533	0.44653297
7	0.474943	0.4765357	0.46715661
8	0.571386	0.6102297	0.56201886
9	0.393025	0.3857542	0.39229242
10	0.394941	0.3678502	0.40872241



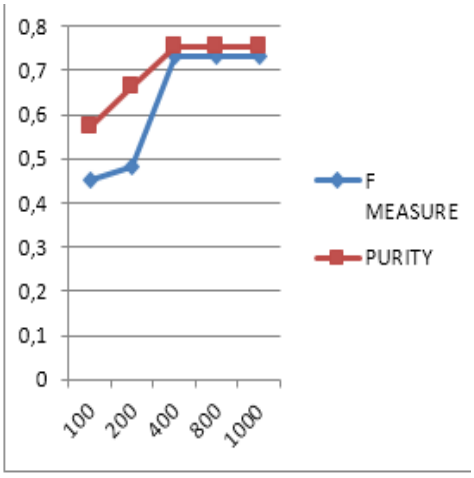
Gambar 5. Grafik Puritas

Berdasarkan perhitungan dari beberapa percobaan maka diperoleh perbandingan nilai *F-measure* yang paling besar adalah percobaan dengan jumlah *cluster* = 2 dengan nilai *F-Measure* untuk *cluster* = 2 sebesar 0.730666369 Dengan nilai rata-rata 0.612919. Dan nilai *Purity* paling besar pada jumlah *cluster* = 2 adalah 0.729723 dengan rata-rata 0.672942.

Pada pengujian Selanjutnya dilakukan uji coba jumlah iterasi yang berbeda-beda seperti pada tabel 3.

Tabel 3. Uji Coba Jumlah Iterasi Berbeda

Iterasi	*F Measure*	*Purity*
100	0.45297778	0.574467
200	0.48337963	0.664161
400	0.73066637	0.754758
800	0.73066637	0.754758
1000	0.73066637	0.754758



Gambar 6. Grafik Perubahan Iterasi

6. Kesimpulan

Berdasarkan analisis terhadap hasil ujicoba, penelitian clusterisasi dokumen web menggunakan k-means ini menghasilkan 2 kesimpulan penting, yaitu: (1) Dokumen berita berhasil dikelompokan secara otomatis sesuai dengan derajat kesamaan berita sehingga menjadi kelompok dokumen berita yang terstruktur dengan diperoleh nilai rata-rata *F-Measure* 0.6129. (2) Jumlah *cluster* dengan nilai puritas terbaik 0.75475adalah 2 *cluster*. Akurasi yang masih belum sempurna dapat diperbaiki dengan memperbaiki teknik *pre-processing* dan melakukan analisis mengenai keakuratan dari metode lain, serta perlunya riset lanjutan untuk menentukan nilai k dengan lebih banyak proses percobaan pengukuran tingkat akurasi dan kemurnian *cluster*.

Referensi

[1] Husni.2011.Web Portal + Search Engine trunojoyo.ac.id.<URL: http://komputasi.wordpress.com/2011/01/03/true-se-web-portal-search-engine-trunojoyo-ac-id/> . Diakses 3 Januari 2011

[2] Wibisono,Y. 2005. "***Clustering* Berita Berbahasa Indonesia**".KK Informatika Sekolah Teknik Elektro dan Informatika I T B. <URL: http://fpmipa.upi.edu/staff/yudi/KNSI_*Clustering*_yudi_masayu.pdf>. Diakses 01 Januari 2011.

[3] Jain, M. N. Murty, and P. J. Flynn. 1999. **Data *clustering***: a review.*ACM Computing Surveys*, 31(3):264–323. URL http://www.csc.kth.se/~rosell/undervisning/sprakt/irintro090824.pdf.> Diakses 31 Januari 2011

[4] Manning, P. Raghavan, and H. Schütze. 2008. **Introduction to Information Retrieval**. Cambridge University Press. ISBN 978-0521865715.<URL http://www.csc.kth.se/~rosell/undervisning/sprakt /irintro090824.pdf.>Diakses 31 Januari 2011

[5] Hand, H. Mannila, and P. Smyth. 2001. **Principles of data mining**. MIT Press, Cambridge, MA, USA. ISBN 0-262-08290-X.

[6] EVERITT, B.S. 1993. *Cluster Analysis* (3ed). London: Edward Arnold

[7] Zhao and G. Karypis. 2004. **Empirical and theoretical comparisons of selected criterion functions for document *clustering***. *Mach. Learn.*, 55(3):311–331. ISSN 0885-6125.

[8] Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. ***Journal of the Royal Statistical Society***, *Series B*, 39(1):1–38.

[9] Feldman, R., dan Sanger, J. 2007**. The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data.** Cambridge University Press.

[10] Rismawan, T., dan Kusumadewi, S., 1991. "**Aplikasi k-means untuk pengelompokkan mahasiswa Berdasarkan nilai body mass index (bmi) & ukuran kerangka**". UII. 1907-5022

[11] Cios, Krzysztof J. Etc. 2007. **Data Mining A Knowledge Discovery Approach**. Springer.

[12] Murad, Azmi MA., Martin, Trevor. .2007. "Word Similarity for Document Gouping using Soft Computing". **IJCSNS International Journal of Computer Science and Network Security**, Vol.7 No.8, August 2007, pp. 20- 27

[13] Roy. 2007. **Berita**. < URL: http://www.beritanet.com/Education/Berita-Jurnalistik/berita.html>. Diakses 15 Maret 2010.

[14] Wikipedia. 2010. **KNN <** URL-http://id.wikipedia.org/wiki/KNN>. Diakses 27 juli 2011.

[15] Rachli. 2007. ***Email Filtering* Menggunakan *Naive Bayesian*.** < URL-www.cert.or.id/~budi/courses/security/2006.../Report-Muhamad-Rachli.doc>Diakses 27 Juli 2011