SISTEM PEMEROLEHAN INFORMASI KARYA ILMIAH BERBASIS CLUSTER DENGAN G-MEANS CLUSTERING

Agustinus Agri Ardyan[1)], J.B. Budi Darmawan[2)]
[1, 2)]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi,
Universitas Sanata Dharma
Mrican, Tromol Pos 29, Yogyakarta 55002
e-mail: agri.ardyan@gmail.com[1)], jbbudi@gmail.com[2)]

ABSTRAK

Dalam kurun waktu terakhir,jumlah publikasi karya ilmiah berbahasa Indonesia berkembang sangat pesat. Tanpa adanya pengembangan dalam sistem pemerolehan informasi, penambahan volume data ini dapat berdampak pada performa sistem, terutama di bidang waktu retrieval.

Metode yang diusulkan penulis untuk menurunkan waktu retrieval adalah pengelompokan koleksi. G-Means dipilih sebagai algoritma pemodelan cluster. Keuntungan penggunaan G-Means adalah kemampuan algoritma tersebut untuk memilih jumlah cluster yang optimal. Hasil pengelompokan dokumen kemudian diuji dalam sistem pencarian dokumen untuk melihat seberapa besar metode pengelompokan dokumen dalam menurunkan waktu retrieval dan dampaknya terhadap precision.

Data yang digunakan sebagai koleksi dalam percobaan ini adalah 100 karya ilmiah berbahasa Indonesia. Pengelompokan dokumen menghasilkan jumlah cluster sebanyak 15 cluster dengan nilai purity sebesar 75%. Berdasarkan hasil pengujian, waktu retrieval turun hingga 16.14% dibandingkan tanpa pengelompokan dokumen, dengan rerata waktu retrieval 12,88 detik. Rerata precision yang didapatkan sebesar 48%.

***Kata Kunci**:pemerolehan informasi, clustering, g-means*

ABSTRACT

In recent years, Indonesian-written scientific papers grow significantly in term of number. Without any improvement in information retrieval systems, increasing data volume could lead to poor system performance, especially in its retrieval time.

One proposed method to improve retrieval time is collection clustering. G-Means was chosen for cluster modeling algorithm, as it can determine the number of generated clusters automatically. Clustering collection results are tested in information retrieval system to find how significant clustering can reduce retrieval time, and whether it has impact to system's average precision.

In this experiment, we use 100 Indonesian scientific papers as collection. Clusters's purity are 75%. Based from the retrieval results, retrieval time gain 2.7% faster, with average retrieval time is about 12.88 seconds and average precision is about 48%.

***Keyword**:information retrieval, clustering, g-means*

I. PENDAHULUAN

Jumlah pertambahan publikasi karya ilmiah di Indonesia tercatat cukup tinggi. Pada tahun 2013, terdapat 4.881 publikasi internasional dan pada tahun 2014, terdapat 5.499 publikasi internasional [1]. Sementara itu, pada tahun 2015 terdapat tambahan 5.421 publikasi internasional yang baru [2]. Jumlah karya ilmiah yang tidak masuk dalam publikasi internasional tersebut tentunya jauh lebih besar lagi.

Dengan volume data yang semakin besar, waktu retrieval menjadi lebih lama [3]. Untuk itu, diperlukan beberapa perbaikan dalam sistem pemerolehan informasi. Salah satu perbaikan yang dapat dilakukan antara lain dengan menerapkan clustering pada koleksi dokumen yang ada.

Penelitian ini mencakup tiga hal, yaitu implementasi G-Means sebagai pemodelan *cluster* yang memiliki kemampuan menentukan jumlah *cluster* optimum, penghitungan nilai purity sebagai tolok ukur kualitas *cluster*, serta penghitungan average precision dari sistem pemerolehan informasi berbasis *cluster*.

Sistem yang akan dikembangkan dalam penelitian ini adalah sebuah sistem pengelompokan koleksi dan pencarian dokumen berdasarkan input kueri pengguna. Sistem ini terdiri dari dua sub sistem, yaitu sub sistem pengelompokan dokumen dan sub sistem pencarian dokumen.

Tujuan dari penelitian ini adalah melihat seberapa baik sistem pemerolehan informasi berbasis *cluster* dalam menurunkan waktu retrieval, dan seberapa besar pengaruhnya terhadap *precision*.

Pemerolehan informasi (*Information Retrieval*) adalah aktivitas menemukan materi dalam koleksi yang tidak terstruktur yang memenuhi kebutuhan informasi, pada suatu koleksi data yang besar [4]. Pemrosesan teks dilakukan pada tahap awal yang meliputi beberapa proses seperti tokenisasi penghilangan stopword, stemming.

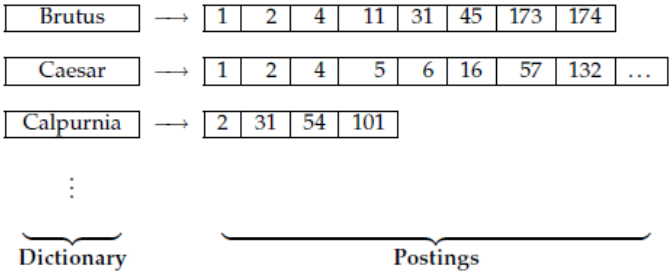
Tokenisasi adalah proses pemisahan kata menjadi bagian-bagian kecil, yang disebut dengan token. Token sering diterjemahkan secara bebas sebagai suku kata, meskipun penting adanya suatu pembedaan dalam terhadap istilah token dan type [5]. Contoh dari input dan output dari tokenisasi adalah sebagai berikut :

Input : Suatu deret angka genap → Output: suatu, deret, angka, genap

Stopword adalah suatu kata yang sangat sering muncul dalam berbagai dokumen adalah diskriminator yang buruk dan tidak berguna dalam temu kembali informasi. Contoh stopword dalam bahasa Indonesia, yaitu kata ganti orang ("aku", "kamu", "kita", dsb.), konjungsi ("dan", "atau", dsb.), dan beberapa kata lainnya.

Stemming adalah proses pengenalan suatu kata. Stemming sering melibatkan pemisahan kata dari imbuhan dan tanda baca [6]. Menurut Agusta [7], pola suatu kata dalam bahasa Indonesia adalah sebagai berikut : Prefiks I + Prefiks II + kata dasar + Sufiks III + Sufiks II + Sufiks I

Inverted index adalah salah satu bentuk struktur data pokok yang terdapat di sistem pemerolehan informasi [8]. Visualisasi inverted index terdapat pada gambar berikut ini :



Gambar 1. Visualisasi inverted index [9]

Konsep dasar inverted index diperlihatkan di gambar 1. diatas. Kumpulan dari berbagai *term* disebut dengan *dictionary*, atau yang disebut juga dengan *vocabulary* atau *lexicon*. Sementara itu, informasi tentang id dokumen tempat keberadaaan *term* terkait (*posting*) akan disimpan dalam suatu list yang disebut dengan *posting list*. Gambar diatas menunjukkan bahwa *term* "Brutus" berada pada dokumen dengan id 1, 2, 4, 11, 31, 45, 173, dan 174. Begitu pula dengan *term* "Caesar" dan "Calipurnia".

***Terms Frequency – Inverse Documents Frequency* (TF-IDF)** adalah skema pembobotan *term* yang paling populer dalam ranah pemerolehan informasi [10].

Formula pembobotan TF-IDF adalah sebagai berikut :

$$w_{ij} = ntf_{ij} * nidf_{ij} \quad (1)$$

Dimana,

$$ntf_{ij} = \frac{tf_{ij}}{\max tf_i} \text{ dan } nidf_{ij} = \frac{\log\left(\frac{m}{df_j}\right)}{\log(m)}$$

Keterangan

- w = bobot *term* (T_j) pada dokumen D_i
- tf_{ij} = frekuensi kemunculan *term* (T_j) pada dokumen D_i
- m = jumlah dokumen D_i pada kumpulan dokumen
- df_j = jumlah dokumen yang mengandung *term* (T_j)
- idf_j = invers frekuensi dokumen (*inverse document frequency*)
- max tf_i = frekuensi *term* terbesar dalam suatu dokumen

Gaussian-Means (G-Means) adalah salah satu jenis pemodelan *cluster* yang dapat menentukan jumlah *cluster* secara otomatis [11]. Algoritma ini terbukti memiliki hasil yang lebih baik dibandingkan X-means dan lainnya. G-Means dimulai dengan jumlah *cluster* yang kecil. Uji statistik dilakukan untuk melihat apakah anggota suatu *cluster* sudah terdistribusi secara normal atau belum. Apabila belum, maka *cluster* tersebut akan dipecah menjadi dua *cluster*.

Algoritma G-Means secara detail adalah sebagai berikut [11] :

```
1. Pilih C sebagai sekumpulan pusat cluster (centroid) awal
2. Lakukan K-Means pada dataset X dengan C sebagai pusat-pusat clusternya.
```

3. x_i adalah sekumpulan datapoint yang menjadi member centroid c_j, dimana { x_i | class(x_i) = j }
4. Gunakan uji statistik untuk melihat apakah tiap { x_i | class(x_i) = j } mengikuti distribusi normal (pada suatu confidence level α).
5. Jika data terlihat terdistribusi normal, maka c_j tidak berubah. Namun jika sebaliknya, maka c_j diganti menjadi dua pusat cluster
6. Ulangi langkah no. 2 hingga tidak ada lagi pusat cluster yang ditambahkan.

Terdapat dua hipotesis dalam uji statistik pada no. 4, yaitu sebagai berikut [11]:

- H_0 : data disekitar pusat *cluster* terdistribusi normal
- H_1 : data disekitar pusat *cluster* tidak terdistribusi normal

Jika H_0 diterima, maka pusat *cluster* tidak perlu dipisah lagi menjadi dua. Sementara itu, jika H_1 diterima, maka pusat *cluster* harus dipecah menjadi dua.

Uji statistik yang digunakan adalah adalah uji Anderson-Darling, dengan formula sebagai berikut [11]:

$$A_*^2(Z) = A^2(Z)(1 + \frac{4}{n} - \frac{25}{n^2}) \quad (2)$$

dengan :

$$A^2(Z) = -\frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log(z_i) + \log(z_{n+1-i})] - n \quad (3)$$

X adalah subset dengan pusat *cluster* C. Tiap instance dari X diwakili dengan x_i, x_{i+1}, ..., x_{n-1}, x_n. Sementara itu, z_i adalah hasil dari fungsi distribusi kumulatif untuk distribusi normal baku terhadap nilai x_i.

Untuk melakukan uji statistik diatas, dilakukan langkah seperti berikut ini :

1. *Ambil suatu subset X*
2. *Pilih level signifikan α untuk uji.*
3. *Dari pusat cluster tersebut, ambil dua buah "anak" pusat cluster, dinotasikan dengan c1 dan c2. Caranya dengan menggunakan rumus c±m, dimana m adalah random*
4. *Hitung nilai vektor v dengan v = c1 – c2.*
5. *Proyeksikan X ke v, menjadi X', dengan rumus sebagai berikut :*
$$x_i' = \frac{(x_i, v)}{\|v\|^2}$$
6. *Normalisasi X` sehingga memiliki rerata 0 dan varian 1.*
7. *Hitung z_i dengan rumus $z_i = F(x_i)$.*
8. *Hitung $A_*^2(Z)$. Apabila $A_*^2(Z)$ berada pada daerah non-kritis, maka H_0 diterima. Sebaliknya apabila $A_*^2(Z)$ berada di dalam daerah kritis, maka H_1 diterima dan pusat cluster yang baru adalah c1 dan c2.*

Recall digunakan untuk mengukur seberapa baik suatu sistem melakukan pencarian terhadap dokumen yang relevan terhadap suatu query pengguna. **Precision** digunakan untuk melihat seberapa baik sistem pemerolehan informasi mengeliminasi dokumen yang tidak relevan [12].

Formula recall dan precision adalah sebagai berikut [13] :

$$Recall = \frac{\sum dokumen\ relevan\ yang\ diperoleh}{\sum seluruh\ dokumen\ relevan} \quad (4)$$

$$Precision = \frac{\sum dokumen\ relevan\ yang\ diperoleh}{\sum dokumen\ yang\ diperoleh} \quad (5)$$

Purity adalah salah satu pengukuran dalam evaluasi *cluster*. Untuk menghitung purity, tiap *cluster* diberikan label kelas berdasarkan label yang paling sering muncul dalam *cluster* tersebut, dan kemudian akurasi *cluster* dihitung dengan jumlah data yang benar dibagi dengan banyak data [14]. Rentang purity dari 0 hingga 1. Semakin besar nilai purity, semakin baik *cluster* tersebut.

Formula purity adalah sebagai berikut [14] :

$$purity(\Omega, \Gamma) = \frac{1}{N}\sum_k max|\omega_k \cap c_j| \quad (6)$$

II. METODE PENELITIAN

Penelitian ini dilakukan dengan terlebih dahulu membangun sistem untuk pengelompokan dan pencarian dokumen. Sistem yang telah selesai dibangun kemudian diuji oleh 10 responden untuk mendapatkan nilai recall dan precision yang digunakan sebagai bahan analisa.

A. Sistem yang Dikembangkan

Sub sistem pengelompokan dokumen bertindak sebagai modul *clustering* dokumen. Nantinya koleksi dokumen yang diunggah oleh Administrator kedalam sistem mula-mula diproses oleh

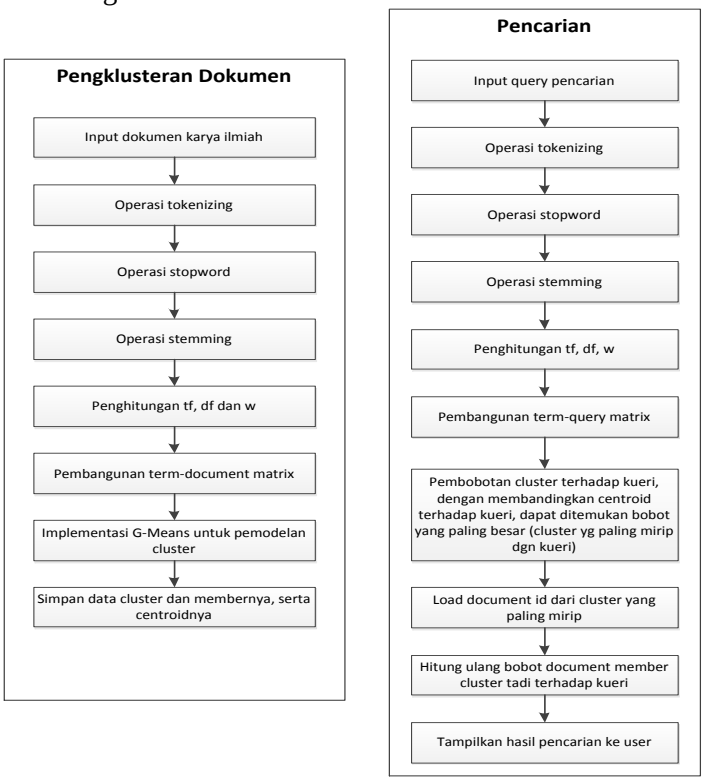
subsistem ini. Proses yang terjadi adalah tokenisasi, eliminasi stopword, stemming, lalu dilanjutkan dengan pembangunan term-document matrix. Dari term-document matrix inilah akan dilakukan pengelompokan koleksi.

Jumlah *cluster* optimum akan dicari secara otomatis oleh sistem menggunakan algoritma G-Means, yaitu pemodelan *cluster* dengan memperhitungkan kenormalan distribusi dari tiap anggota *cluster* terhadap pusatnya masing-masing. Dari situ, dapat diketahui berapa jumlah *cluster* yang optimum.

Sub sistem pencarian dokumen berfungsi untuk mencari dokumen yang memiliki kemiripan atau relevan dengan kueri yang diberikan oleh pengguna sistem. Kueri hanya akan dicocokkan dengan *centroid* tiap *cluster* dengan menggunakan operator boolean AND dan fungsi jarak *Euclidean distance*. *Cluster* yang memiliki *centroid* dengan kemiripan yang tertinggi terhadap kueri user akan dicatat oleh sistem. Apabila tidak ada kecocokan dengan semua centroid, maka dicoba pencocokan dengan menggunakan operator OR.

Apabila sudah ditemukan *cluster* yang sesuai, dokumen yang berada dalam *cluster* tersebut akan dibobot ulang oleh sistem menggunakan TF-IDF untuk kemudian ditampilkan urut ke pengguna berdasarkan bobot terhadap kueri yang diberikan oleh pengguna. Jumlah dokumen untuk penghitungan IDF didasarkan pada jumlah dokumen yang berada pada *cluster* terpilih.

Alur sistem ini adalah sebagai berikut :



Gambar 2. Alur sistem

B. DATA

Data yang digunakan dalam penelitian ini adalah 100 karya ilmiah berbahasa Indonesia yang diambil dari prosiding berbagai seminar yang memiliki ranah teknologi informasi. Pemilihan dokumen tersebut dilakukan secara acak.

C. EVALUASI HASIL

Evaluasi hasil pengelompokan dokumen dilakukan dengan melibatkan penilaian manusia. Penilaian ini disebut juga dengan gold standard atau ground truth [14]. Dari penilaian tersebut, dilakukan evaluasi secara matematis dengan penghitungan purity. Penghitungan purity untuk menghitung kualitas kluster salah satunya pernah dilakukan pada penelitian oleh Rendy *et al* [15].

Pengukuran dan evaluasi hasil retrieval dapat dilakukan dengan penghitungan *recall* dan precision, serta average precision. Data relevansi ini didapatkan dengan melibatkan 10 responden untuk memberikan kueri.

Dengan responden yang sama, hasil retrieval Sistem Pemerolehan Informasi Berbasis Cluster ini juga dibandingkan dengan sistem pemerolehan informasi tanpa *cluster*. Dalam paper ini, sistem pemerolehan informasi tanpa *cluster* disebut dengan Sistem Pemerolehan Informasi Konvensional.

III. HASIL

A. *Hasil Pengelompokan Dokumen*

Pengelompokan dokumen oleh sistem menghasilkan 15 *cluster*. Dengan term-document matrix berdimensi 100 x 4067, waktu eksekusi sub sistem pengelompokan dokumen sekitar 21,5 menit.

Dengan melibatkan responden, evaluasi hasil *clustering* menghasilkan nilai purity sebesar 0.75. Hasil tersebut terlihat dalam tabel berikut ini :

Tabel I.
Hasil penghitungan purity *cluster*

CLUSTER	MATCH	TOPIK
1	16	diagnosis penyakit
2	1	tindak pidana
3	2	metode certainty factor
4	1	metode single moving average
5	8	data mining
6	11	klasifikasi
7	15	sistem pendukung keputusan
8	1	manajemen qos
9	2	protokol routing ad-hoc
10	8	jaringan
11	1	implementasi ospf
12	1	SIG wisata
13	1	SIG prediksi bencana
14	6	SIG
15	1	SIG penduduk
TOTAL	75	

Kolom MATCH berisi jumlah dokumen yang paling banyak memiliki kesamaan dalam satu *cluster*. Jumlah nilai pada kolom MATCH tersebut adalah 75. Dengan total data sebanyak 100 dokumen, dapat dihitung nilai purity dari pemodelan *cluster* tersebut yaitu 0.75.

B. *Hasil Pencarian Berdasar Kueri Pengguna*

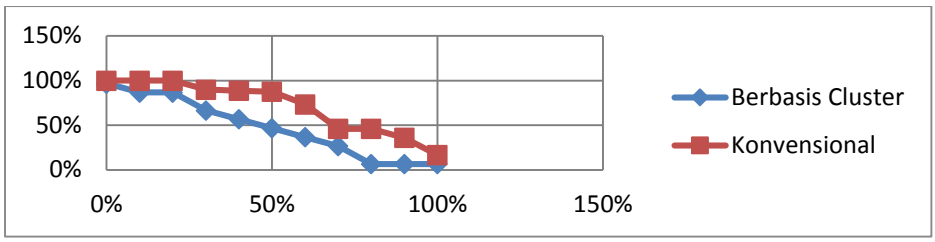
Dari hasil pengujian sub sistem pencarian dokumen, pengguna menghitung berapa dan apa saja dokumen relevan yang ditampilkan oleh sistem. Setelah jumlah dokumen relevan ditemukan, dihitung nilai rata-rata precision dari kedua sistem. Hasilnya adalah Sistem Pemerolehan Informasi Berbasis Cluster menghasilkan rata-rata precision sebesar 48%, sementara Sistem Pemerolehan Informasi Konvensional menghasilkan rata-rata precision sebesar 71%.

Penyajian hasil perhitungan interpolasi 11 titik recall precision untuk kedua sistem disajikan dalam tabel berikut ini :

TABEL II.
RATA-RATA INTERPOLASI 11 TITIK RECALL PRECISION DARI PENGUJIAN DENGAN 10 KUERI PENCARIAN DARI KEDUA SISTEM

RECALL	BERBASIS CLUSTER	KONVENSIONAL
0%	97%	100%
10%	87%	100%
20%	87%	100%
30%	67%	90%
40%	57%	89%
50%	47%	88%
60%	37%	74%
70%	27%	46%
80%	7%	46%
90%	7%	36%
100%	7%	17%
AVE	48%	71%

Visualisasi hasil perhitungan interpolasi 11 titik recall precision untuk kedua sistem dalam bentuk grafik dapat dilihat pada gambar 1 di bawah ini.



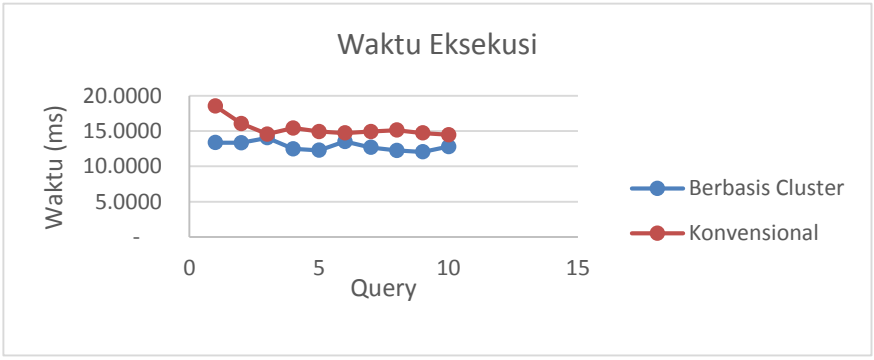
Gambar 3. Grafik Recall - Precision dari Sistem Pemerolehan Informasi Berbasis Cluster dan Konvensional

Dari grafik tersebut, terlihat luas bidang dibawah grafik yang mewakili Sistem Pemerolehan Informasi Konvensional lebih luas dibandingkan luas bidang dibawah grafik yang mewakili Sistem Pemerolehan Informasi Berbasis Cluster. Sehingga dapat disimpulkan bahwa Sistem Pemerolehan Informasi Konvensional memiliki precision yang lebih baik dibandingkan dengan Sistem Pemerolehan Informasi Berbasis Cluster.
Pencatatan waktu retrieval disajikan dalam TABEL III di bawah ini.

TABEL III
WAKTU RETRIEVAL SISTEM DI TIAP KUERI DARI PENGGUNA

QUERY	BERBASIS CLUSTER (s)	KONVENSIONAL (s)
1	13.3721	18.5425
2	13.3514	16.0667
3	14.0447	14.5661
4	12.4743	15.4405
5	12.2814	14.9437
6	13.5197	14.7423
7	12.681	14.9348
8	12.2554	15.1375
9	12.0591	14.7241
10	12.7906	14.4841
AVE	**12.88297**	**15.35823**

Sementara itu, visualisasi waktu retrieval kedua sistem disajikan dalam bentuk grafik dibawah ini :



Gambar 4 Grafik waktu retrieval dari Sistem Pemerolehan Informasi Berbasis Cluster

Dari tabel dan grafik tersebut, terlihat bahwa Sistem Pemerolehan Informasi Berbasis Cluster memiliki waktu eksekusi yang lebih baik dibandingkan dengan Sistem Pemerolehan Informasi Konvensional.

IV. PEMBAHASAN

A. EVALUASI CLUSTER

Terdapat 7 *cluster* yang masing-masing berisi satu dokumen saja. Hal ini dimungkinkan karena dokumen tersebut memiliki distribusi yang berbeda dengan dokumen lainnya, sehingga ketika terjadi penghitungan nilai kritis (*lih.* **Error! Reference source not found.** point no.4), *cluster* dianggap tidak terdistribusi normal dan dokumen tersebut disendirikan dalam *cluster* tersendiri oleh sistem.

B. WAKTU EKSEKUSI DAN AVERAGE PRECISION

Sistem pemerolehan informasi berbasis *cluster* dalam pengujian selalu unggul dalam waktu retrieval yang lebih singkat 16.14 % dibandingkan sistem pemerolehan informasi konvensional,

dengan rerata waktu retrieval sebesar 12,88 detik. Hal ini disebabkan karena jumlah dokumen yang harus diretrieve menjadi lebih sedikit karena sudah dikelompokkan oleh sistem.

Sebagai tradeoff, nilai average precision menurun dari 71 % pada sistem pemerolehan informasi konvensional menjadi 48 % pada sistem pemerolehan informasi berbasis *cluster* . Hal ini disebabkan karena beberapa dokumen relevan berada di *cluster* yang berbeda dengan *cluster* yang dipilih sistem untuk di*retrieve*.

Untuk mengatasi penurunan average precision, pemodelan *cluster* akan lebih tepat apabila menggunakan soft assignment atau soft *clustering*. Soft *clustering* dinilai lebih tepat untuk ranah pemerolehan informasi dan natural language processing (NLP) [16]. Pemodelan *cluster* soft assignment memungkinkan satu dokumen berada di beberapa *cluster*, sehingga jumlah miss (dokumen relevan yang tidak di*retrieve*) bisa dikurangi dan recall dapat meningkat.

V. SIMPULAN DAN SARAN

A. SIMPULAN

Dalam penelitian ini, diketahui bahwa sistem pemerolehan informasi berbasis *cluster* menghasilkan waktu retrieval yang lebih singkat. Dalam pengujian, diketahui rata-rata waktu retrieval sekitar 12.88 detik. Lebih singkat 16.14% dibandingkan Sistem Pemerolehan Informasi Konvensional. Sebagai tradeoff, nilai rata-rata precision cenderung menurun. Dalam pengujian, didapatkan nilai rata-rata precision sebesar 48%. Hal ini terjadi karena pemodelan *cluster* yang menggunakan pemodelan *hard clustering*, dimana satu dokumen hanya bisa menjadi anggota satu *cluster* saja. Selain itu, retrieval sistem ini dibatasi dengan hanya mengambil satu *cluster* dokumen saja. Sehingga dokumen relevan yang berada di *cluster* lain tidak ikut terambil.

B. SARAN

Penggunaan pemodelan *cluster* dengan jenis *soft clustering* dirasa lebih tepat untuk kasus pengelompokan dokumen. Misalnya pengubahan algoritma pemodelan cluster dari G-Means ke Fuzzy c-Means Clustering (FCM).

Untuk inisialisasi centroid awal, dapat ditambahkan algoritma inisialisasi centroid seperti k-Means++ agar menghasilkan *cluster* yang lebih baik.

Reduksi dimensi dimungkinkan dapat meningkatkan purity *cluster* dan meningkatkan precision sistem. Reduksi dimensi dapat dilakukan dengan feature selection. Dalam konteks pemerolehan informasi, salah satu metode feature selection yang efisien adalah metode DF [17].

REFERENSI

[1] Scimagojr.com,. 2016. *SJR - International Science Ranking*. Diakses pada 7 Januari 2016, dari http://scimagojr.com/countryrank.php

[2] KOMPAS.com. 2016. *Kemenristekdikti Nyatakan Indonesia Lampaui Target Publikasi Internasional - Kompas.com*. Diakses pada 7 Januari 2016, dari http://sains.kompas.com/read/2015/10/30/16544281/Kemenristekdikti.Nyatakan.Indonesia.Lampaui.Target.Publikasi.Internasional

[3] Chen, Berlin. *Clustering Techniques for Information Retrieval*. Department of Computer Science & Information Engineering. National Taiwan Normal University.

[4] Manning, C., Raghavan, P., Schutze, H. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. 2009: 1

[5] Manning, C., Raghavan, P., Schutze, H. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. 2009: 22

[6] Göker, Ayşe., Davies, John. *Information Retrieval Searching in 21st Century*. West Sussex: John Wiley & Sons. 2009.

[7] Agusta, Ledy. *Perbandingan Algoritma Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia*. Jurnal Konferensi Nasional Sistem dan Informatika 2009. 2009.

[8] Büttcher, Stefan., Clarke, L.A. Charles., Cormack, V. Gordon. *Information Retrieval Implementing and Evaluating Search Engine*. Massachusetts: MIT Press. 2010.

[9] Göker, Ayşe., Davies, John. *Information Retrieval Searching in 21st Century*. West Sussex: John Wiley & Sons. 2009.

[10] Baeza-Yates, R., Ribeiro-Neto, B. *Modern Information Retrieval the Concept and Technology Behind Search*. England: ACM Press. 1999.

[11] Hamerly, Greg., Elkan, Charles. *Learning the k in k-means*. Electronic Proceeding of Advances in Neural Information Processing Systems 16 (NIPS). 2004.

[12] Croft, Bruce W., Meltzer, Donald., Strohman, Trevor. *Search Engines Information Retrieval in Practice*. Massachusetts: Amherst. Pearson Education USA. 2010.

[13] Manning, C., Raghavan, P., Schutze, H. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. 2009: 155

[14] Manning, C., Raghavan, P., Schutze, H. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. 2009: 356

[15] Handoyo, Rendy. Rumani, R.M., Nasution, S.M. *Perbandingan Metode Clustering Menggunakan Metode Single Linkage dan K-Means pada Pengelompokan Dokumen*. Jurnal SIFO Mikroskil. 2014; 15(02):73.

[16] Chen, Berlin. *Clustering Techniques for Information Retrieval*. Department of Computer Science & Information Engineering. National Taiwan Normal University.

[17] Yang, Yiming. Pedersen, Jan O. *A Comparative Study on Feature Selection in Text Categorization*. Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco. 1997. 412-420