OTOMATISASI PENGELOMPOKKAN KOLEKSI PERPUSTAKAAN DENGAN PENGUKURAN *COSINE SIMILARITY* DAN *EUCLIDEAN DISTANCE*

Heri Kurniawan, Rizal Fathoni Aji
Fakultas Ilmu Komputer, Universitas Indonesia Jakarta
E-mail: {hek50, rfa51}@ui.edu

ABSTRAKSI

Sebuah perpustakaan mempunyai bilangan katalog yang sangat besar. Katalog tersebut bertambah setiap waktu seiring dengan pertambahan cirinya yang sangat beragam. Otomatisasi pengklasifikasian katalog tanpa mendefinisikan kategori sebelumnya akan mempercepat dan mempermudah pengelompokkan kategori secara lebih objektif. Otomatisasi terdiri dari dua langkah yaitu pengelompokkan dan pemberian label. Sebagai langkah awal implementasi otomatisasi yang efektif, paper ini mencoba melihat kinerja pengelompokkan dokumen melalui penerapan Cosine Similarity dan Euclidean Distace sebagai distance measure dalam algoritma Kmeans.

Kata kunci: *Clustering, Kmeans, Euclidean, Cosine Vector, Frekuensi term, Unsupervised*

1. PENDAHULUAN

Jumlah koleksi dalam sebuah perpustakaan bersifat dinamis, artinya selalu terjadi pertambahan koleksi dari waktu ke waktu. Kategorisasi koleksi diperlukan untuk mempermudah pemakai dalam mencari disiplin ilmu yang diinginkan. Kategorisasi secara manual akan menyulitkan dan membutuhkan waktu yang lama. Perlu ada mekanisme yang cepat dan objektif untuk kategorisasi koleksi. Dalam paper ini, kami menggunakan teknik *unsupervised clustering* untuk pengelompokkan dokumen[6]. Teknik ini telah banyak digunakan pada area *text mining* dan *information retrieval*.

Teknik pengelompokkan (*clustering*) dokumen dikembangkan untuk meningkatkan *precission* atau *recall* dalam *information retrieval*[2]. Cara ini dilakukan untuk meningkatkan efisiensi dalam menemukan dokumen yang saling berkaitan[3]. Saat ini terdapat dua teknik pengelompokkan dokumen yaitu secara hirarkis dan non hirarkis[1]. Teknik yang dipakai pada paper ini adalah non hirarkis (*partition*) dengan menggunakan algoritma *Kmeans*[7].

Penelitian ini bertujuan untuk melihat kinerja pengelompokkan dokumen melalui penerapan *Cosine Similarity* dan *Euclidean Distace* sebagai *distance measure* dalam algoritma kmeans. Ujicoba dilakukan dalam dua tahap, yaitu mengelompokkan koleksi dokumen dan memberikan label sebagai represetasi isi kelompok dokumen.

2. LANDASAN TEORI

2.1 Pengelompokkan

Pengelompokkan dengan teknik non hirarkis akan menghasilkan sejumlah K *cluster* yang saling terpisah. Bilangan K ini ditentukan sebelumnya secara manual. Implementasi algoritma KMeans diawali dengan inisialisasi titik *centroid* (pusat *cluster*). Banyaknya titik *centroid* tergantung dari bilangan K *cluster* yang dipilih. Bila K=4, maka akan terdapat 4 titik *centroid*. Nilai titik awal *centroid* dipilih secara manual [6]. Pada umumnya titik dokumen ($D_{i..j}$) menjadi titik awal K *centroid* (C_1=D_1,C_2=D_2). Selanjutnya antara titik *centroid* dengan dokumen-dokumen dihitung jaraknya dengan menggunakan standar pengukuran. Pengukuran yang digunakan pada paper ini adalah *Euclidean*[8] dan *cosine similarity*[4].

$$D_{ij} = \sqrt{\sum_{k-1}^{n} (d_{jk} - c_{jk})^2}$$

Euclidean

$$C(d_i, c_j) = 1 - \frac{\vec{di} \bullet \vec{cj}}{|\vec{di}| \cdot |\vec{cj}|}$$

Cosine similarity

Jarak yang terpendek antara *centroid* dengan dokumen menentukan posisi *cluster* suatu dokumen. Misalnya dokumen A mempunyai jarak yang paling pendek ke *centroid* 1 dibanding ke yang lain, maka dokumen A masuk ke group 1. Hitung kembali posisi *centroid* baru untuk tiap-tiap *centroid* ($C_{i..j}$)dengan mengambil rata-rata dokumen yang masuk pada *cluster* awal ($G_{i..j}$). Iterasi dilakukan terus hingga posisi group tidak berubah.

$$\mathrm{C(i)} = \frac{1}{|G_i|} \sum_{\vec{x} \in c} \vec{dx}$$

Penentuan *centroid*

Berikut adalah contoh implementasi *Kmeans* dengan menggunakan pengukuran *Euclidean* (K=2).

Tabel 1. Sample data

Dok	***Term sample***
D_1	Kesejahteraan sosial Indonesia
D_2	Sosial masyarakat surabaya
D_3	Peta kemiskinan Indonesia
D_4	Kesejahteraan pegawai Indonesia

Kesejahteraan=T_1, sosial =T_2, Indonesia=T_3, Masyarakat=T_4, Surabaya=T_5, Peta=T_6, Kemiskinan=T_7, Pegawai=T_8

Tabel 2. Jumlah term pada masing-masing dokumen

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
D_1	4	5	2					
D_2		10		4	1			
D_3		1				5	2	
D_4	6		4					2

Inisialisasi *centroid*, dengan C_1=D_1, C_2=D_2, C_1(4,5,2,0,0,0,0,0) C_2(0,10,0,4,1,0,0,0).

Jarak antara D_3(0,1,0,0,0,5,2,0) dengan C_1: $\sqrt{(0-4)^2+(1-5)^2+(0-2)^2+(0-0)^2+(0-0)^2+(5-0)^2+(2-0)^2+(0-0)} = 8.06$
Jarak antara D_3 dengan C_2 adalah 11.27.

Tabel 3. Jarak centroid dengan dokumen

Jarak	D_1	D_2	D_3	D_4
C_1	0	7.87	8.06	6.08
C_2	7.87	0	11.27	13.11

Pada tabel 3, D_1 berjarak 0 dari C_1 dan berjarak 2 dari C_2. Karena jarak D_1 ke C_1 lebih kecil dibanding jarak D_1 ke C_2, maka D_1 masuk kedalam *cluster* G_1. Begitu juga dengan dokumen D_2, D_3, D_4. Hasil pengelompokkan dokumen bisa dilihat pada tabel 4. Selanjutnya dilakukan reformulasi *centroid* hingga posisi group tidak berubah.

Tabel 4. Posisi grup pada interasi awal

GC_1	D_1	D_2	D_3	D_4
G_1	1	0	1	1
G_2	0	1	0	0

Nilai *centroid* baru pada C_1 (T_1,T_2,T_3,T_4,T_5,T_6,T_7,T_8), misalnya pada titik T_1, T_1 =($D_{1..n}T_{1..n}$)/(jumlah dokumen pada *cluster* G_1), T_1= (D_1T_1+D_3T_1+D_4T_1)/3 =(4+0+6)/3= 3.33. Selanjutnya nilai *centroid* baru adalah C_1(3.33, 2, 2, 0, 0, 2.5, 1, 1) ; C_2(0,10,0,4,1,0,0,0).

Berikut adalah tabel jarak dan group dengan mengunakan nilai *centroid* yang baru.

Tabel 5. Jarak cluster dengan dokumen

Jarak	D_1	D_2	D_3	D_4
C_1	3.51	9.76	4.93	4.86
C_2	7.87	0	11.27	13.11

Tabel 6. Posisi grup pada interasi kedua

GC_2	*D1*	*D2*	*D3*	*D4*
G1	1	0	1	1
G2	0	1	0	0

Pada tabel 6 terlihat bahwa posisi dokumen pada tabel group iterasi kedua sama dengan posisi dokumen pada tabel grup iterasi pertama. Ini menandakan komputasi dengan *Kmeans* clustering telah mencapai kestabilan. Selanjutnya iterasi untuk reformulasi titik centroid tidak diperlukan lagi. Proses diatas mengelompokkan dokumen D_1, D_3 dan D_4 pada cluster G_1 sedangkan dokumen D_2 pada cluster G_2.

2.2 Pemberian Label

Yang dimaksud *labeling* disini adalah pemberian nama untuk masing-masing *cluster*. Pencarian label untuk *cluster* dilakukan dengan tahap-tahap sebagai berikut [4]:

1. Ambil pasangan *term* dan *frequency* (frekwensi bisa berupa tf maupun df) dari setiap dokumen dalam *cluster* tersebut.
2. Urutkan pasangan <term,freq> berdasarkan freq.
3. Ambil beberapa term teratas sebagai hasil label

Contoh: Misalkan pasangan `<term, freq>` yang sudah terurut adalah sebagai berikut:

```
informasi -> 637
teknologi -> 610
perusahaan -> 569
bisnis -> 552
sistem -> 509
proses -> 484
```

label cluster tersebut adalah ‘informasi teknologi’.

Sebelum digabungkan dengan *Kmeans* clustering, algoritma ini di uji coba dengan menggunakan data yang sudah di kelompokkan secara manual. Hasil uji coba labeling dapat dilihat dalam tabel 7.

Tabel 7. Hasil label

Kelompok Dokumen	*Hasil label dengan freq tf*	*Hasil label dengan freq. df*
Tesis MTI Fasilkom	Sistem Informasi	Studi Informasi
Tesis MIK Fasilkom	Informasi Sistem	Sistem Informasi
Buku dengan subjek Computer Graphics	Graphics Computer	Graphics Computer
Buku dengan subjek Information System Management	Information Management	Information Management

3. EKSPERIMEN

3.1 Dataset

Eksperimen dilakukan dengan menggunakan data judul dari Buku perpustakaan Fakultas Ilmu Komputer (Fasilkom) 281 judul dan buku perpustakaan pusat UI 778 judul.

3.2 Metode

Jumlah *cluster* yang dibentuk (k) adalah 10 *cluster*. Algoritma penghitungan jarak antara *centroid* dengan dokumen menggunakan *euclidean*

distance dan *cosine similarity*, sedangkan dalam proses *labeling* digunakan tf untuk pengurutan *term*. Hasil dari cluster dan labeling dapat dilihat pada daftar dibawah ini. Dalam satu cluster, hanya ditampilkan lima dokumen saja.

3.3 Hasil eksperimen

Berikut adalah waktu yang dibutuhkan dalam proses clustering.

Tabel 8. Waktu eksekusi

Time	Fasilkom	UI
Euclidean Distance	5707 ms	26234 ms
Cosine Similarity	6348 ms	34750 ms

Pada Tabel 8 terlihat, proses *clustering* dengan *euclidean distance* berlangsung lebih cepat bila dibandingkan dengan menggunakan *cosine similarity*. Ini berlaku untuk sampel data Fasilkom maupun data UI. Berikut ini adalah hasil proses cluster dan label dibandingkan dengan subjek buku yang dibuat oleh pustakawan.

Tabel 9. *Clustering* data Fasilkom dengan *Cosine similarity*

Hasil Cluster	***Subjek yang dibuat pustakawan***
Artificial Intelligence Java Gem Programming (69 dokumen)	Artificial intelligence, Calculus, Computer architecture, Computer graphics, Database management, Java (Computer program language), Management Information Systems, Software engineering
Management Systems Information Database Mcleod (48 dokumen)	Computer architecture, Database management Management information systems, Software engineering
Computer Architecture Symposium Approach Edition (25 dokumen)	Computer architecture, Computer networks, Java (Computer program language)
Software Engineering Pressman Int Roger (44 dokumen)	Software engineering
Ibm Network Graphics Waite Study (3 dokumen)	Computer network, Computer graphics
Information Steven Technology Managing System (8 dokumen)	Management information systems, Artificial intelligence, Computer networks
Calculus Edwards Henry Thomas Analytic (13 dokumen)	Artificial intelligence, Calculus
Networking Guide Essentials Implementation Computer (9 dokumen)	Computer networks, Software engineering
Laudon Kenneth Management Information Systems (11 dokumen)	Management Information systems
Computer Networks Graphics Approach 3rd (51 dokumen)	Computer networks, Computer architecture, Management information systems, Computer graphics, Software engineering, Calculus

Beberapa dokumen terkumpul dengan baik dalam satu *cluster* dan memiliki label yang dapat merepresentasikan isi cluster. Selain itu, ada beberapa *cluster* yang memipengelompokannya tidak jelas. Pengelompokan *cluster* yang tidak jelas, lebih banyak terjadi pada data perpustakaan pusat UI. Perpustakaan UI memiliki data yang lebih beragam bila dibandingkan dengan data Fasilkom yang hanya berisi buku-buku sekitar komputer.

Tabel 10. *Clustering* data Fasilkom dengan *Euclidean Distance*

Hasil Cluster	***Subjek yang dibuat pustakawan***
Computer Graphics Artificial Intelligence Java (163 dokumen)	Artificial intelligence, Calculus, Computer architecture, Computer graphics, Computer network, Database management, Java (Computer program language), Management Information Systems, Software engineering
Information Management Systems Mcleod Database (24 dokumen)	Database management, Management information systems
Approach Quantitative Architecture Computer Patterson (5 dokumen)	Computer architecture, Computer networks
Software Engineering Int Pressman Conference (36 dokumen)	Software engineering
Joint Study Ibm Siements Bell (1 dokumen)	Computer network
Information Steven System Alter Martin (6 dokumen)	Management information systems
Thomas Calculus 10 Applied Wonnacott (2 dokumen)	Calculus
Networking Computer Internet Essentials Guide (9 dokumen)	Computer networks
Laudon Kenneth Management Information Systems (10 dokumen(	Management Information systems
Computer Networks Graphics Approach 3rd (25 dokumen)	Computer networks, Computer architecture, Management information systems

Berdasarkan hasil eksperimen, *cosine similarity* menghasilkan *cluster* yang lebih merata bila dibandingkan dengan *euclidean distance. Euclidean distance* menghasilkan beberapa *cluster* yang hanya terdiri dari satu atau dua dokumen, sementara terdapat *cluster* lain yang terdiri dari banyak dokumen. Hasil pemberian label sangat tergantung pada kualitas *cluster* yang terbentuk. *Cluster* dengan subjek yang bersesuaian akan menghasilkan label yang representatif.

Pada percobaan lain digunakan data Fasilkom sebanyak 232 dokumen dengan subjek yang sudah diketahui. Percobaan tersebut bertujuan untuk membandingkan hasil *cluster* dengan *cosine similarity* dan *euclidean distance*. Subjek digunakan sebagai acuan untuk melihat kesesuaian ciri antara hasil cluster dengan data sebenarnya.

Tabel 14 menunjukkan hasil pemberian label dengan menggunakan cosine similarity lebih mendekati bila dibandingkan dengan euclidean distance.

Tabel 11. *Clustering* data Perpustakaan Pusat UI dengan *Cosine similarity*

Hasil Cluster	***Subjek yang dibuat pustakawan***
Novel Systems Cinta Computers Programming (408 dokumen)	Bank dan perbankan, Fiction, Fiksi Amerika, Fiksi Indonesia, Komputer, Komputer Penyusunan program, Manajemen, Manajemen Eksekutif , Manajemen keuangan, Manajemen pegawai
Computer Programming Introduction Technology How (55 dokumen)	Komputer, Komputer Penyusunan program, Manajemen Eksekutif, Manajemen keuangan
Bahasa Alih Anak Editor Orang (39 dokumen)	Bank dan perbankan, Fiction, Fiksi Amerika, Fiksi Indonesia, Manajemen, Manajemen eksekutif, Manajemen keuangan, Manajemen pegawai
Bank Pt Studi Kasus Indonesia (88 dokumen)	Bank dan Perbankan, Fiksi Amerika, Fiksi Indonesia, Manajemen, Manajemen eksekutif, Manajemen keuangan, Manajemen pegawai
Kisah Jiwa Mata Gadis Mencari (17 dokumen)	Bank dan Perbankan, Fiksi Amerika, Fiksi Indonesia, Manajemen keuangan
Atmowiloto Arswendo Pengkhianatan G30s/pki (1 dokumen)	Fiksi Indonesia
Management Financial Human James Concepts (68 dokumen)	Komputer, Komputer Penyusunan program, Manajemen, Manajemen eksekutif, Manajemen keuangan, Manajemen pegawai
Business Harvard Review Essentials Leadership (27 dokumen)	Fiction, Fiksi Indonesia, Manajemen, Manajemen eksekutif, Manajemen keuangan, Manajemen pegawai
Stephen Dark King Dean Hari (26 dokumen)	Fiction, Fiksi Amerika, Fiksi Indonesia, Manajemen, Manajemen eksekutif, Manajemen pegawai
Manajemen Manusia Daya Sumber Editor (49 dokumen)	Fiction, Manajemen, Manajemen Eksekutif , Manajemen keuangan, Manajemen pegawai

Tabel 12. *Clustering* data Perpustakaan Pusat UI dengan *Euclidean Distance*

Hasil Cluster	***Subjek yang dibuat pustakawan***
Eric Decision Concepts Control Noreen (1 dokumen)	Manajemen keuangan
Computer Programming Introduction Technology Design (51 dokumen)	Komputer, Komputer Penyusunan program, Manajemen keuangan
Anak Yun Ching Bahasa Bezine (2 dokumen)	Fiksi Amerika
Kasus Studi Bank Pt Indonesia (52 dokumen)	Bank dan perbankan, Manajemen, Manajemen keuangan, Manajemen pegawai
Risk Value Perhitungan Singapore Var (3 dokumen)	Bank dan perbankan, Manajemen keuangan
Manajemen Bahasa Alih Editor Novel (579 dokumen)	Bank dan perbankan, Fiction, Fiksi Amerika, Fiksi Indonesia, Komputer, Komputer Penyusunan program, Manajemen, Manajemen Eksekutif , Manajemen keuangan, Manajemen pegawai
Management Financial Human James Concepts (64 dokumen)	Komputer, Manajemen, Manajemen eksekutif, Manajemen keuangan, Manajemen pegawai
Business Harvard Review Essentials Leadership (24 dokumen)	Manajemen, Manajemen Eksekutif Manajemen keuangan, Manajemen pegawai
Dark Frank Australia Folk Story (1 dokumen)	Fiction
Daerah Kepala Konsep Konteks Era (1 dokumen)	Manajemen

Tabel 13. Daftar subjek

Class	***Jumlah***
Computer Architecture	48
Calculus	32
Computer Networks	64
Computer Graphics	54
Multimedia Systems	20
Project Management	14
Total	232 dokumen

Tabel 14. Hasil *Cluster*

Cosine Similarity (7640 ms)	***Jumlah***	***Euclidean Distance (7109 ms)***	***Jumlah***
Calculus Edwards	66	Calculus Multimedia	115
Computer Networks	67	Computer Approach	51
Computer Graphics	47	Computer Graphics	46
Design Computer	15	Computer Design	10
Multimedia Computing	26	Technology Computer	1
Graphics Computer	11	Graphics Computer	9

4. KESIMPULAN

Dari sisi ketepatan hasil, *cluster Kmeans* dengan menggunakan *cosine similarity* memberikan hasil yang lebih baik bila dibandingkan dengan pengukuran *euclidean distance*. Namun waktu proses yang dibutuhkan *euclidean distance* lebih cepat bila dibandingkan dengan *cosine similarity*. Kualitas label sangat tergantung pada kualitas hasil cluster. Koleksi data yang sangat beragam dapat menurunkan kualitas pengelompokkan.

DAFTAR PUSTAKA

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, Data Clustering: A Review, *ACM Computing Surveys*, vol.31, no.3, pp. 264–323. (1999).
[2] C. J. van Rijsbergen, (1989), *Information Retrieval*, Buttersworth, London, second edition.
[3] Chris Buckley and Alan F. Lewit, Optimizations of inverted vector searches, *SIGIR '85*, Pages 97-110, 1985.
[4] Filippo Ricca, Emanuele Pianta and Christian Girardi,Using Keyword Extraction for Web Site Clustering Paolo Tonella, *ITC-irst Centro per la Ricerca Scientifica e Tecnologica*, Italy, 2003.
[5] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison-Wesley, Reading, MA,1989).
[6] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. *In Proc. 6 th Europ. Conf. Comput. Vision*, Dublin, Ireland, June 2000.
[7] Richard C. Dubes and Anil K. Jain, *Algorithms for Clustering Data*, Prentice Hall, 1988.
[8] Tarazaga P. Hayden, *Linear Algebra and its Applications*, 1996.