Klasterisasi Judul Buku dengan Menggunakan Metode K-Means

Deka Dwinavinta Candra Nugraha
Teknik Informatika, Fakultas Teknologi Industri,
Universitas Islam Indonesia
dekadwinavinta_cn@ymail.com

Zumrotun Naimah
Teknik Informatika, Fakultas Teknologi Industri,
Universitas Islam Indonesia
zumrotunnaimah@gmail.com

Makhfuzi Fahmi
Teknik Informatika, Fakultas Teknologi Industri,
Universitas Islam Indonesia
makhfuzi.fahmi@gmail.com

Novi Setiani
Teknik Informatika, Fakultas Teknologi Industri,
Universitas Islam Indonesia
115230101@uii.ac.id

Abstrak*—Secara umum perpustakaan merupakan tempat yang menyediakan berbagai bahan pustaka yang digunakan untuk memenuhi kebutuhan semua orang. Koleksi buku yang semakin banyak di Perpustakaan akan memudahkan banyak orang untuk mencari studi pustaka yang diinginkan. Namun banyaknya koleksi buku juga akan menyulitkan dalam pengelolaan letak buku di perpustakaan sehingga ada beberapa buku yang tidak terbaca dan tidak terpinjam. Kemudian untuk penyelesaian masalah tersebut maka akan diterapkan teknik clustering dengan metode K-Means pada pengelolaan buku di perpustakaan. Teknik clustering merupakan sebuah teknik pengelompokan sejumlah data/obyek ke dalam cluster (group) sehingga dalam setiap cluster akan berisi data yang semirip mungki.n Pada penelitian ini, teknik clustering tersebut akan diterapkan pada Perpustakaan Pusat Universitas Islam Indonesia. Teknik clustering akan mengelompokan judul buku sesuai dengan kategorinya. Buku-buku yang memiliki cluster yang sama akan digunakan sebagai bahan untuk analisis dalam pengambilan keputusan yang bertujuan untuk mempermudah pustakawan(petugas perpustakaan) dalam pengelolaan peletakan buku yang diminati dan merancang strategi dalam meningkatkan minat baca mahasiswa.

Kata kunci— perpustakaan, clustering, K-Means

I. PENDAHULUAN

Seiring dengan meningkatnya pertumbuhan teknologi, Seiring dengan pesatnya perkembangan ilmu pengetahuan, jumlah koleksi buku yang ada di perpustakaan juga mengalami peningkatan. Dengan koleksi buku yang sangat banyak tersebut, salah satu masalah yang ditimbulkan adalah sulitnya proses pengelolaan letak buku. Yang tentu saja akan menyulitkan para anggota perpustakaan dalam mencari buku yang diinginkan. Sehingga diperlukan suatu solusi yang mempermudah proses pengelolaan letak buku tersebut.

Inti dari permasalahan tersebut ialah bagaimana data-data koleksi tersebut diolah untuk dikelompokkan. Untuk itu data yang diperlukan dalam penelitian ini ialah data koleksi buku yang merupakan data digital koleksi perpustakaan. Dari banyaknya data tersebut kita mengambil sampel berjumlah 500. Data tersebut berbentuk tabel yang terdiri dari kolom judul, tahun terbit, jenis, bahasa yang digunakan, nomor DDC, dan eksemplar.

Tujuan utama dari penelitian ini adalah penggunaan metode *data mining* dalam membantu proses pengelolaan letak buku di perpustakaan. *Data mining* adalah sekumpulan metode yang digunakan untuk mendapatkan informasi dari data dengan cara mempelajari pola dari data tersebut. Selain itu, *data mining* sering juga disebut *Knowledge Discovery in Database* (KDD), yaitu kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [1].

Sehingga dalam penelitian ini fokus pada metode *data mining* dengan kasus pengelompokan data (*clustering*). Adanya data dalam skala besar memungkinkan metode data mining dengan teknik *clustering* yang dapat mengelompokkan data ke dalam beberapa kelompok yang diinginkan. Teknik *clustering* yang digunakan yaitu *K-Means.*

K-Means adalah suatu teknik pengelompokan data yang mana keberadaan tiap-tiap titik data dalam suatu cluster ditentukan oleh derajat keanggotaan. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981[2]. Data yang akan digunakan dalam teknik *K-Means* ini ialah sampel data yang hanya diambil kolom judul buku. Yang kemudian akan dikelompokkan berdasarkan berdasarkan kemiripan derajat kanggotaan data tersebut di dalam set data.

II. TEORI PENDUKUNG

Berikut adalah beberapa teori yang mendukung penelitian ini.

A. Konsep Data Mining

Data mining adalah suatu istilah yang digunakan untuk untuk menguraikan penemuan pengetahuan di dalam database.*Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai database besar[3].

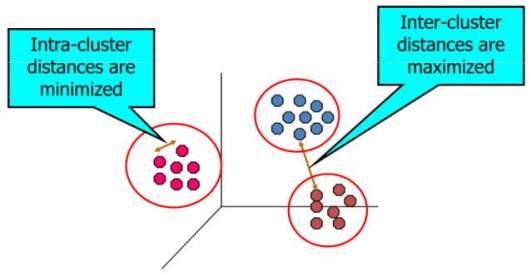
Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual [4].

Dimana hasil dari proses penggalian tersebut akan membentuk pola-pola dari kumpulan data, yang sering disebut dengan pengenalan pola (*pattern recognition*). Pengenalan pola merupakan bagian dari *data mining*. *Data mining* sering juga disebut *Knowledge Discovery in Database* (KDD), yaitu kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [1].

B. *Clustering*

Clustering merupakan salah satu teknik data mining yang digunakan untuk mendapatkan kelompok-kelompok dari obyek-obyek yang mempunyai karakteristik yang umum di data yang cukup besar. Tujuan utama dari metode clustering adalah pengelompokan sejumlah data/obyek ke dalam cluster (*group*) sehingga dalam setiap cluster akan berisi data yang semirip mungkin [1].

Clustering melakukan pengelompokan data yang didasarkan pada kesamaan antar objek, oleh karena itu klasterisasi digolongkan sebagai metode *unsupervised learning*.



Gambar 1. Konsep Clustering

III. HASIL DAN PEMBAHASAN

A. *Preprocessing Data*

Sebelum proses klasifikasi dilakukan, diperlukan *preprocessing* data terlebih dahulu. Tahap *preprocessing* yang dilakukan:

1. *Tokenizing*

Tokenizing adalah proses untuk memecah kalimat pada kolom Title menjadi kata-kata yang terpisah. Hal ini dilakukan untuk membandingkan nilai setiap kata yang ada di setiap judul.

Kata-kata yang ada pada setiap kalimat judul dipisahkan dengan mendeteksi karakter spasi di dalamnya, spasi digunakan sebagai pemisah antara satu kata dengan kata yang lain, sehingga setiap kata menjadi suatu kesatuan tersendiri seperti yang terlihat pada gambar 2 berikut.

fundamental	of	metal	forming					
introduction	to	asean	librarianship	academic	libraries			
basic	principles	of	chemistry					
defining	and	measuring	democracy					
proceedings	of	the	regional	seminar	on	conservation	of	biodiversity
time	series	analysis	forecasting	and	control			
research	a	practical	guide	to	finding	information		
pascal								
design	of	prestressed	concrete	structures				
plastic	design							
an	introduction	to	visicalc	matrixing	for	apple	and	ibm
power	pack	for	the	ibm	pc			
essentials	of	information	processing					
increasing	your	productivity	lotus	release				

Gambar 2. Hasil Tokenizing

2. *Remove Duplicate*

Antara judul satu dengan yang lain memiliki kemungkinan besar terdapat kata-kata yang sama, sehingga terdapat beberapa kata yang jumlahnya lebih dari satu seperti yang terlihat di gambar 3. Dengan melakukan langkah ini maka kita dapat mendapatkan kata yang *unique*. Kata-kata ini digunakan untuk melakukan langkah berikutnya yaitu pembobotan.

fundamental
introduction
basic
defining
proceedings
time
research
pascal
design
plastic
an
power
essentials
increasing
of
to
principles
and
series
a
pack
your
metal
asean
measuring

Gambar 3. Hasil Remove Duplicate

3. Menghapus kata-kata yang memiliki unsur angka dan kata-kata yang tidak dibutuhkan

Langkah ini adalah langkah yang digunakan untuk menghapus kata-kata yang tidak diperlukan dan kata-kata yang mengandung angka seperti yang terlihat pada gambar 4.

fundamental
introduction
basic
defining
proceedings
time
research
pascal
design
plastic
power
essentials
increasing
principles
pack
metal
asean
measuring
analysis
practical
prestressed
information
productivity
forming

Gambar 4. Hasil Menghapus kata-kata yang memiliki unsur angka dan kata-kata yang tidak dibutuhkan

4. Menghapus tanda baca dalam kata

Tanda baca dalam kata tentu bukan hal penting dalam *clustering*, sehingga kita dapat menghapusnya. Pada langkah ini kata yang memiliki tanda baca tidak dihapus semua, namun penghapusan hanya dilakukan oleh tanda bacanya saja seperti yang terlihat pada gambar 5.

fundamental
introduction
basic
define
proceed
time
research
pascal
design
plastic
power
essential
increase
principle
pack
metal
asean
measure
analysis
practical
prestressed
information
productivity
form
librarianship

Gambar 5. Hasil Menghapus tanda baca dalam kata

B. Klasterisasi *K-Means*

K-Means adalah suatu teknik pengelompokan data yang mana keberadaan tiap-tiap titik data dalam suatu cluster ditentukan oleh derajat keanggotaan. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981[2].

K-Means merupakan algoritma *clustering* yang berulang-ulang. Algoritma *K-Means* dimulai dengan pemilihan secara acak K, K disini merupakan banyaknya *cluster* yang ingin dibentuk. Kemudian tetapkan nilai K secara random, untuk sementara nilai tersebut menjadi pusat dari cluster atau bisa disebut dengan centeroid menggunakan rumus hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Klasifikasikan setiap data berdasarkan kedeketannya dengan centroid. Lakukan langkah tersebut hingga nilai *centroid* tidak berubah (stabil) [5].

Dasar algoritma *K-Means*:

1. Pilih K sebagai centorid awal
2. Ulangi
3. Bentuk K cluster dengan menetapkan semua poin ke centroid terdekat.
4. Menghitung berubah centroid setiap cluster
5. Sampai centroid tidak

Proses klasterisasi *K-Means* ini menggunakan aplikasi Weka dengan pembentukan 6 klaster dan 1000 iterasi. Data yang digunakan untuk proses klasterisasi ini adalah data judul buku dengan judul bahasa inggris sebanyak 500 judul. Berikut adalah hasil dare proses klasterisai yang dapat dilihat pada gambar 6 dan gambar 7.

```
=== Model and evaluation on training set ===

Clustered Instances

0        1 (  0%)
1        2 (  0%)
2      428 ( 86%)
3       31 (  6%)
4        4 (  1%)
5       34 (  7%)
```

Gambar 6. Hasil Klasterisasi K-Means

cluster 1	cluster 2	cluster 3		
intermediate	power	forming	analysis	ninth
problems	engineering	fundamental	control	beginners
exercises	statistics	metal	forecasting	boundary
manual	engineers	academic	series	complete
questions	civil	asean	time	element
case	textbook	introduction	finding	advanced
	hydro	librarianship	guide	turbo
	environmental	libraries	information	framework
	reliability	basic	practical	managing
	probability	chemistry	research	improvement
		principles	pascal	architektur
		defining	concrete	china
		democracy	design	impelmentation
		measuring	prestressed	intelligent
		biodiversity	structures	manufacturing
		conservation	plastic	behaviour
		proceedings	apple	charecteristics
		regional	ibm	deformation
		seminar	matrixing	pavement
		transport	visicalc	sections
		package	pack	strength
		second	pc	contruction

Gambar 7. Daftar Kata Hasil Klasterisasi K-Mean

cluster 4	cluster 5	cluster 6	
engineering	combined	principles	correlation
journal	parts	control	tqm
geoenvironmental	physics	information	approachess
geotechnical	fundamentals	research	responsiveness
dobel	modern	productivity	production
	university	applications	pavement
	automatic	business	contruction
	nuclear	computer	apllications
	extended	programs	xmp
		construction	protocols
		engineering	api
		theory	strategies
		management	global
		project	property
		total	professionals
		based	professional
		statistics	practic
		methods	ethical
		statistical	edited
		systems	aspect
		human	sales
		marketing	service

Gambar 8. Daftar Kata Hasil Klasterisasi K-Mean

Untuk mengetahui validitas dari hasil klasterisasi, maka dilakukan ujicoba dengan cara mencocokkan hasil klasterisasi dengan judul buku yang sebenarnya. Sebagai contoh yakni beberapa kata yang ada pada *cluster*2 yaitu *power*, *engineering, statistics, engineer, civil, environmental*.

TABEL 1. VALIDASI HASIL KLASTERISASI K-MEAN

Kata	Judul Buku
power	power transmission elements american locomotives a pictorial record of steam power water power engineering
engineer	control system engineering
ing	corrosion engineering applied mechanics for engineering technology
engineer	vector mechanics for engineers dynamics
civil	reliability based design in civil engineering civil engineering materials
environmental	aplication of biotechnology environmental and policy issues environmental conservation

Setelah dicocokkan dengan judul buku yang mengandung kata tersebut, ternyata kata-kata yang ada pada *cluster* 2menunjukkan buku-buku dengan bidang ilmu teknik.

IV. KESIMPULAN

Dari hasil klasterisasi menggunakan K-Means, terbentuklah beberapa *cluster* yang di dalamnya berisi kata-kata yang memiliki jarak yang berdekatan. Setelah melalui proses validasi, ternyata buku-buku yang berada dalam satu *cluster* memiliki kategori yang serupa. Dengan menganalisis hasil kalsterisasi menggunakan K-Mean maka buku-buku yang memiliki *cluster*serupabisa digunakan sebagai bahan untuk analisis dalam pengambilan keputusan yang bertujuan untuk mempermudah pustakawan (petugas perpustakaan) dalam pengelolaan peletakan buku yang diminati dan merancang strategi dalam meningkatkan minat baca mahasiswa.

DAFTAR PUSTAKA

[1] Santosa, Budi. 2007. Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu.

[2] Kusumadewi dan Purnomo. 2010. Aplikasi Logika Fuzzy untuk mendukung keputusan. Yogyakarta: Graha Ilmu.

[3] Kusrini dan Lutfi, E.T. 2009. Alogoritma Data Mining. Yogyakarta: Andi Offset.

[4] Pramudiono, 2006, Apa Itu Data Mining?, [online], (http://gunawan-ndra.blogspot.com/2013/03/pengertian-data-mining-menurut-para.html. diakses tanggal 26 September 2013)

[5] Rismawan, T. (2008). Aplikasi K-Means Untuk Pengelompokkan Mahasiswa, 2008(Snati).