bidang
EKONOMI

PENENTUAN *MULTIPLE MEMBERSHIP* DOKUMEN

STEPHANIE BETHA R.H
Program Studi Akuntansi, Fakultas Ekonomi dan Bisnis
AMIK Purnama Niaga Indramayu

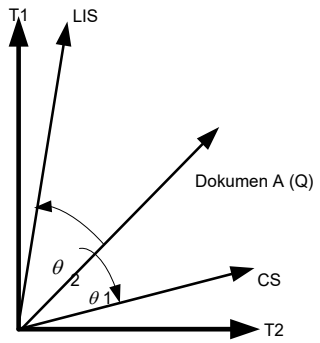
Multiple membership merupakan keanggotaan yang dimiliki oleh seseorang pada beberapa komunitas. *Multiple membership* pada dokumen artinya suatu dokumen dapat mengandung konten dari beberapa jenis kategori. Jenis kategori pada dokumen dapat ditentukan dengan mengukur kemiripan dokumen tersebut dengan kategori yang ada. *Vector Space Model* adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dan suatu *query* dengan mewakili setiap dokumen dalam sebuah koleksi sebagai sebuah titik dalam ruang vektor. Hasil dari pengukuran kemiripan tersebut merupakan nilai *cosine similarity* antara vektor *query* dari dokumen terhadap vektor kategori. Permasalahan yang terjadi adalah suatu pengukuran kemiripan vektor *query* dokumen, dapat menghasilkan nilai *cosine similarity* dengan selisih yang kecil antara vektor kategori satu dengan vektor kategori lain. Hal ini menyebabkan kedua vektor kategori tersebut menjadi saling dominan satu sama lain pada dokumen. Oleh karena itu, dibutuhkan suatu nilai batas untuk menentukan kondisi kapan suatu vektor kategori dapat dinyatakan sebagai vektor kategori yang saling dominan. Penetapan nilai batas ini menggunakan *K-Means Clustering*. Nilai batas ini ditetapkan berdasarkan pengelompokkan nilai jarak antar presentase *cosine similarity* pada suatu dokumen. Penentuan *multiple membership* dokumen ini akan dilakukan pada atribut judul dan kata kunci pada dokumen publikasi ilmiah.

Keywords **:** *Dokumen, Multiple Membership, Nilai Batas, K-Means*

PENDAHULUAN

Multiple membership merupakan keanggotaan yang dimiliki oleh seseorang pada beberapa komunitas (Nevedov, N, 2011). *Multiple membership* kategori pada dokumen artinya suatu dokumen dapat mengandung konten dari beberapa jenis kategori. Suatu publikasi ilmiah dapat memiliki beberapa jenis bidang atau topik penelitian (Meng, Qinxue dan Kennedy, Paul, 2012) Penentuan jenis kategori pada dokumen ini dapat dilakukan melalui pengukuran kemiripan *query* dokumen terhadap vektor kategori tertentu. Kesamaan antar jenis kategori dan *query* tersebut dinyatakan dengan *cosinus* sudut antar keduannya (Ning Liu, dkk, 2004). Nilai *cosinus* yang besar mengindikasikan bahwa semakin dekat dokumen tersebut terhadap suatu *query*. Nilai *cosinus* sama dengan satu mengindikasikan bahwa dokumen sama dengan *query* (Mandala, R dan Setiawan, H, 2002).

Vector Space Model adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dan suatu *query* dengan mewakili setiap dokumen dalam sebuah koleksi sebagai sebuah titik dalam ruang vektor (Turney, Peter dan Pantel, Patrick, 2010). Suatu *query* Dokumen A memiliki kesamaan dengan beberapa jenis kategori yaitu nilai *cosine similarity* terhadap vektor kategori *Computer Software* (CS) sebesar 0,228 dan nilai *cosine similarity* terhadap vektor kategori *Library Information Science* (LIS) sebesar 0,011. Pengukuran kemiripan ini akan diilustrasikan pada Gambar 1.



Gambar 1. Pengukuran kemiripan *query* dokumen A terhadap vektor kategori

Gambar 1. menunjukkan Dokumen A lebih memiliki kemiripan terhadap vektor kategori CS dibandingkan vektor kategori LIS. Hal ini dibuktikan dengan sudut cosinus yang terbentuk antara *query* dengan kategori CS lebih kecil, dibandingkan sudut cosinus yang terbentuk antara *query* dengan vektor kategori LIS.

Permasalahan yang terjadi adalah suatu dokumen dapat memiliki kemiripan dengan selisih yang kecil pada kedua nilai *cosine similarity* vektor kategori. Hal ini mengakibatkan vektor kategori satu dengan vektor kategori lain saling dominan. Oleh karena itu, dibutuhkan suatu nilai batas untuk menentukan kondisi kapan suatu dokumen memiliki kecenderungan *multiple membership* atau tidak. Penetapan nilai batas dilakukan menggunakan *K-Means Clustering*. *K-Means* adalah algoritma sederhana untuk menganalisis data, bersifat simple*, robust* dan efisien (Wesan, Barbakh And Colin Fyfe, 2008). Atribut dokumen yang digunakan dalam penentuan *multiple membership* dokumen ini adalah judul dan kata kunci.

Paper ini terdiri atas beberapa bagian. Beberapa penelitian terkait dengan *multiple membership*, VSM dan *K-Means* akan dijelaskan pada bagian II. Proses penentuan batas nilai *multiple membership* akan dijelaskan pada bagian III. Bagian IV menjelaskan tentang implementasi. Paper ini diakhiri dengan kesimpulan pada bagian V.

PENELITIAN TERKAIT

Penelitian (Nevedov, N, 2011) mengusulkan suatu metode untuk mendeteksi *multiple membership* di suatu komunitas serta memprediksi rekomendasi *link* pada *multi layer graph* berdasarkan *network topology*. Metode ini memiliki proses awal yaitu deteksi komunitas menggunakan *modularity maximation. Modularity maximation* adalah salah satu ukuran yang digunakan untuk mendeteksi komunitas pada suatu *network*.

VSM digunakan untuk merepresentasikan dokumen menjadi informasi *multiple membership* dokumen. Beberapa jenis *Information Retrieval Model* dapat digunakan untuk merepresentasikan dokumen. Namun, VSM adalah metode yang paling layak untuk merepresentasikan dokumen Guo, Qinglin, 2008). Hal itu disebabkan oleh VSM disertai model aljabar yang mampu merepresentasikan dokumen dengan baik dan disertai dengan tingkat fleksibilitas yang lebih tinggi dibandingkan dengan model Probabilitas ataupun Boolean (Pannu Mandeep, et al, 2014).

K-Means membagi data ke dalam k

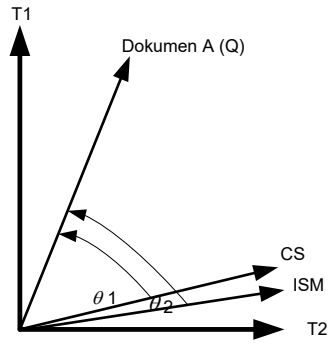
kelompok yang ditentukan secara manual terlebih dahulu. *K-Means clustering* digunakan sebagai salah satu dasar prediksi kinerja akademik siswa (M. Durairaj, dkk, 2014). *K-Means* digunakan untuk memprediksi kelulusan dan kegagalan mahasiswa. Hasil akurasi menunjukkan bahwa algoritma ini dapat memprediksi kinerja akademik siswa secara akurat. Selain itu, algoritma *K-Means* juga mudah diimplementasikan dan memiliki kompleksitas waktu sebesar $O(n)$ dengan n jumlah pola atau kelompok (Jain,A.K, dkk, 1999).

USULAN METODE

Multiple membership pada dokumen yang dimaksud adalah suatu dokumen dapat mengandung konten dengan jenis kategori yang saling mendominasi satu sama lain. *Multiple membership* dokumen diperoleh dari jenis kategori yang terkandung pada dokumen tersebut. Penentuan kategori dokumen dilakukan dengan mengukur kemiripan (*similarity document*) antara vektor *query* dokumen terhadap vektor jenis kategori yang ada. Nilai *cosine similarity* yang dihasilkan dari proses pengukuran kemiripan tersebut, tidak dapat digunakan untuk menentukan *multiple membership* dokumen secara langsung. Hal ini disebabkan oleh nilai *cosine similarity* belum dapat mewakili kecenderungan *multiple membership* suatu dokumen. Permasalahan dalam penentuan jenis kategori dokumen adalah suatu konten dokumen dapat mengandung beberapa jenis kategori, sehingga dibutuhkan suatu cara untuk menentukan kondisi, kapan suatu dokumen memiliki kategori yang saling dominan satu sama lain atau tidak (*multiple membership*). Ilustrasi permasalahannya sebagai berikut :

Dokumen A mengandung presentase jenis kategori *Computer Software* (CS) sebesar 57% dan presentase jenis kategori *Information System* (ISM) sebesar 43%. Nilai presentase setiap kategori ini diperoleh dari presentase nilai *cosine similarity* per kategori dibagi dengan total *nilai cosine similarity* pada suatu dokumen. Kategori *Computer Software* (CS) memiliki presentase lebih besar presentase jenis kategori *Computer Software* (CS). Namun, belum dapat dipastikan bahwa kategori *Computer Software* (CS) lebih dominan dari kategori lain pada dokumen tersebut. Nilai presentase tersebut diperoleh dari pengukuran kemiripan *query* pada Dokumen A terhadap vektor kategori.

Pengukuran kemiripan *query* pada Dokumen A menghasilkan nilai *cosine similarity* terhadap vektor kategori *Computer Software* (CS) sebesar 0,08 dan nilai cosine similarity terhadap vektor kategori *Information System* (ISM) sebesar 0,06. Hasil pengukuran kemiripan Dokumen A tersebut dapat digambarkan pada model ruang vektor sebagai berikut :



Gambar 2. Dokumen A pada ruang vektor

Gambar 2. menunjukkan bahwa *query* dari Dokumen A lebih memiliki kemiripan terhadap vektor kategori CS. Hal ini disebabkan oleh sudut yang dihasilkan antara *query* dokumen dengan vektor kategori CS lebih kecil dibandingkan sudut antara *query* Dokumen A dengan vektor kategori ISM. Namun, kondisi tersebut belum dapat digunakan sebagai justifikasi *multiple membership* suatu dokumen.

Oleh karena itu, dibutuhkan suatu nilai batas untuk menentukan kapan suatu dokumen dapat disebut memiliki *multiple membership* atau tidak.

1. Pengolahan Data Latih

Penentuan nilai batas diawali dengan mengolah data latih. Data latih ini yang akan digunakan sebagai input pada algoritma *K-Means Clustering*. Data latih diperoleh dari kumpulan judul dan kata kunci dokumen publikasi ilmiah. Setiap judul dan kata kunci pada data latih diukur kemiripannya berdasarkan jenis kategori yang ada. Pengukuran kemiripan ini bertujuan untuk memperoleh nilai *cosine similarity* (kemiripan) *query* judul atau kata kunci terhadap vektor kategori yang ada.

Data latih mengalami pra proses dokumen yang terdiri atas *case folding*, tokenisasi, penghapusan *stop words* dan *stemming* pada judul dan kata kunci. Kemudian, proses pembelajaran dilakukan pada data latih. Proses pembelajaran ini berfungsi untuk menghasilkan vektor yang dapat merepresentasikan setiap kategori. Input dari proses ini adalah daftar kata judul dan kata kunci dari hasil pra proses dokumen. Pembobotan TFIDF digunakan untuk membobotkan kata pada kata judul dan kata kunci. Pembobotan TFIDF dilakukan dengan perhitungan sebagai berikut :

$$Term\ Weight : w_i = tf_i * \log\left(\frac{D}{df_i}\right) \quad \text{(III.1)}$$

tf_i adalah banyaknya *term* i yang muncul pada sebuah dokumen. df_i merupakan banyaknya dokumen yang mengandung *term* i. D adalah jumlah seluruh dokumen.

Panjang vektor kategori dapat dihitung dengan perhitungan sebagai berikut :

$$|D_i| : \sqrt{\sum wij}^2 \quad \text{(III.2)}$$

Perhitungan panjang vektor kategori dilakukan dengan menghitung total bobot kata kuadrat (*term weight*) pada kategori.

Panjang vektor *query* dapat dihitung dengan perhitungan sebagai berikut :

$$|Q| : \sqrt{\sum wqj}^2 \quad \text{(III.3)}$$

Setelah itu, *query* dari judul dan kata kunci diukur kemiripannya terhadap vektor kategori yang telah dibangun. Pengukuran kemiripan ini akan menghasilkan nilai *cosine similarity* pada judul dan kata kunci terhadap vektor kategori. Sudut kemiripan antara vektor *query* dan vektor kategori dihitung dengan perhitungan *cosine similarity* sebagai berikut :

$$Cos(Q,D) : \frac{Q \cdot D}{|Q| * |D|} \quad \text{(III.4)}$$

Q adalah *query* dan D adalah kategori. $|Q|$ merupakan panjang vektor query. $|D|$ adalah panjang vektor kategori. Pengukuran kemiripan *query* dokumen terhadap vektor kategori menghasilkan dua jenis nilai *cosine similarity* terhadap jenis vektor kategori, yaitu, *cosine similarity* pada peringkat pertama dan peringkat kedua.

Semua nilai *cosine similarity* pada peringkat pertama dan kedua dalam data latih diubah ke bentuk presentase. Tujuannya adalah untuk mengetahui besarnya presentase kategori terhadap suatu dokumen. Perhitungan presentase nilai *cosine*

similarity terhadap vektor kategori dihitung dengan rumus sebagai berikut :

$$\frac{nilai\ cosine\ similarity\ terhadap\ suatu\ vektor\ kategori}{Total\ nilai\ \ cosine\ similarity\ pada\ satu\ dokumen} x\ 100\%$$

Kedua nilai presentase antar kategori ini akan dihitung jaraknya. Perhitungan jarak dilakukan menggunakan rumus jarak *euclidean.* Apabila jarak antara nilai presentase *cosine similarity* pada suatu dokumen dinotasikan dengan $d\ (x, y)$ maka jarak antar nilai presentase *cosine similarity* dihitung dengan persamaan berikut ini (M. Durairaj, dkk, 2014) :

$$d(x, y) = \sqrt{(a_{rx} - a_{ry})^2} \quad \text{(III.4)}$$

$d(x, y)$ merupakan jarak antara nilai presentase *cosine similarity*, a_{rx} adalah nilai presentase *cosine similarity* dokumen r pada kategori x dan a_{ry} adalah nilai presentase *cosine similarity* dokumen r pada kategori y . *Pseudo code* perhitungan jarak akan ditampilkan pada tabel di bawah ini.

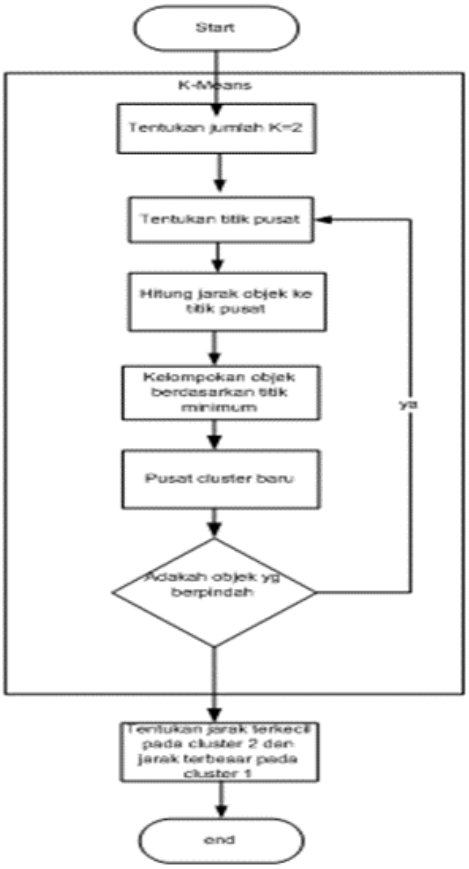
```
/*3. calculate distance */
for (i<-1 to nPaper) do
  sumCosineCategory <- 0
    for (j<-1 to j<nCategory-1) do
    x <- DataPaper[i].DataCategory[j].presentase
   y <- DataPaper[i].DataCategory[j+1].presentase
   D <- sqrt((x-y)^2)


 DataPaper[i].distanceKategory <- D
    end for
end for
```

2. Penentuan Nilai Batas

Penentuan nilai batas diawali dengan mengukur jarak antar presentase nilai *cosine similarity*. Daftar nilai jarak inilah yang akan diolah dengan *K-Means Clustering* pada *tools* RapidMiner 5.3. Algoritma ini diawali dengan proses penentuan jumlah *cluster*. Jumlah *cluster* yang digunakan pada penelitian ini sejumlah 2 cluster. Nilai k=2 ini ditetapkan berdasarkan kondisi suatu dokumen bersifat *multiple membership* atau tidak. Kemudian, titik pusat ditentukan pada langkah kedua. Setelah itu, jarak dihitung dari setiap objek ke titik pusat. Selanjutnya, objek dikelompokkan berdasarkan titik minimum. Semua proses akan berhenti ketika objek tidak mengalami perpindahan lagi. Gambar 3. menunjukkan alur proses penentuan nilai batas menggunakan *K-Means Clustering*.

Tabel 1. Daftar *centroid* pada setiap iterasi

Iterasi	*Centorid Cluster* 1	*Centorid Cluster* 2
1	79,844	19,315
2	80,214	19,674
3	80,242	19,702
4	19,702	80,242
5	19,702	80,242



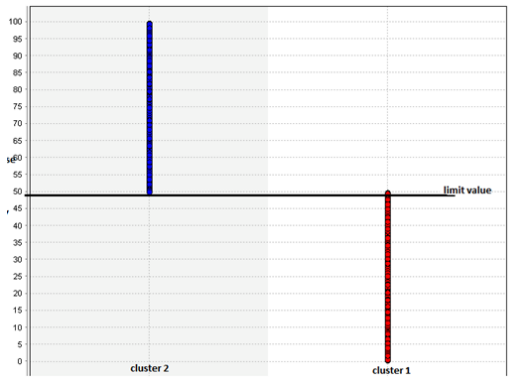
Gambar 3. Alur algoritma *K-Means* untuk menentukan nilai batas

Nilai batas kecenderungan *multiple membership* ini ditentukan sesuai dengan nilai jarak terbesar pada *cluster* 1 dan nilai jarak terkecil pada *cluster* 2. Jumlah iterasi yang dilakukan adalah sebanyak 4 iterasi hingga mencapai *cluster* yang konvergen.

Tabel 1 menunjukkan daftar *centroid* atau titik pusat pada setiap iterasi. Iterasi keempat (kolom yang berwarna abu-abu) pada tabel 1. menunjukkan bahwa nilai titik pusat tidak mengalami perpindahan lagi.

Tabel 1 menunjukkan bahwa *cluster* bersifat konvergen pada iterasi ke-4. Pemodelan *K-Means* ini menghasilkan model *cluster* yang terdiri atas 2 *cluster*. *Cluster* 1 memiliki 1078 *items* dan cluster 2 memiliki 1080 *items*. *Centroid* akhir *cluster* 1 yaitu 19,072. *Centroid* akhir *cluster* 2 yaitu 80,242. Gambar 3. menunjukkan grafik hasil algoritma *K-Means.*

Gambar 4. menunjukkan bahwa perpotongan garis antara *cluster* 1 dengan *cluster* 2 berada pada nilai mendekati 50. Hal ini dibuktikan dengan jarak terkecil pada cluster 2 sebesar 49,982 dan jarak terbesar pada cluster 1 sebesar 49,701. Oleh karena itu, jarak antara 49,701 – 49,982 akan digunakan sebagai nilai batas untuk menentukan kecenderungan *multiple membership* dokumen pada penelitian ini.



Gambar 4. Grafik hasil algoritma *K-Means*

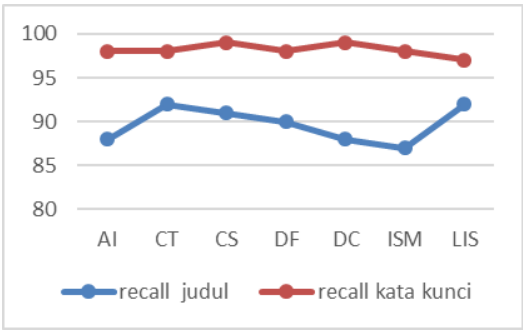
Apabila suatu dokumen memiliki jarak antar presentase nilai *cosine similarity* kategori di bawah nilai batas, maka dokumen tersebut dapat dikatakan memiliki kecenderungan *multiple membership.* Sebaliknya, apabila suatu dokumen memiliki jarak antar nilai *cosine similarity* kategori di atas nilai batas, maka dokumen tersebut tidak memiliki kecenderungan *multiple membership.*

Pseudo code dalam penentuan *multiple membership* dokumen akan dijelaskan secara garis besar sebagai berikut :

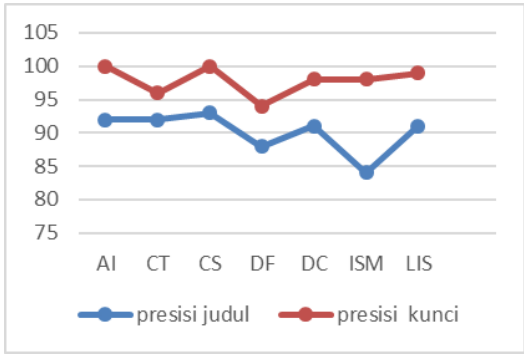
```
/* determinant multimember of paper*/
for (i<-1 to nPaper) do
 DataPaper[i].statusMultimember <- false
  for (j<-1 to j<nCategory) do
   if (DataPaper[i].distanceKategory
     < paramBoundary) then
        DataPaper[i].statusMultimember <- true
    end if
   end for
end for

/*5. Output*/
write ('Multimember Paper : ')
for (i<-1 to nPaper) do
  if (DataPaper[i].statusMultimember = true)
then
    write ('Paper ke-', i, ', ')
  end if
end for
```

PENGUJIAN DAN IMPLEMENTASI

Pengujian bertujuan untuk menguji metode VSM menggunakan atribut kata kunci dan judul dari dokumen. Data pengujian ini berasal dari jurnal yang diambil dari DBLP dan terdapat kategorinya pada divisi *Information and Computing Science* pada standar klasifikasi *Field of Research* (FoR). Kategori yang digunakan sejumlah 7 bidang penelitian. Total data latih sejumlah 2300 dan total data uji sejumlah 575 dokumen.



Gambar 5. Perbandingan *recall* judul dengan kata kunci



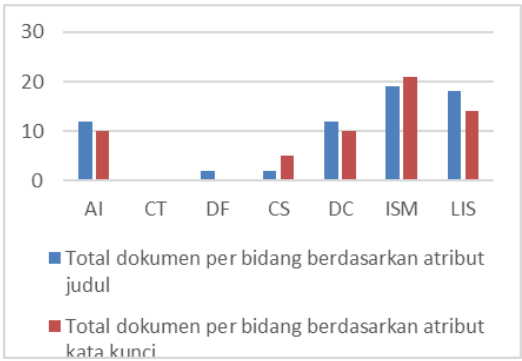
Gambar 6. Perbandingan presisi judul dengan kata kunci

Gambar 4. Dan Gambar 5. menunjukkan bahwa penggunaan kata kunci dapat menghasilkan nilai *recall* dan presisi yang lebih tinggi dibandingkan penggunaan judul. Hal ini disebabkan oleh *term* pada kumpulan kata kunci memiliki jenis kata yang unik (setiap dokumen dapat memiliki jenis kata kunci yang berbeda).

Perbedaan nilai *recall* dan presisi menggunakan atribut judul dan kata kunci akan menghasilkan informasi *multiple membership* dokumen yang berbeda juga. Semakin tinggi nilai *recall* dan presisi pada suatu atribut, belum tentu menghasilkan lebih banyak variasi kecenderungan multiple membership dokumen. Hal tersebut akan dibuktikan dengan mengimplementasikan penggunaan kedua atribut pada dokumen studi kasus.

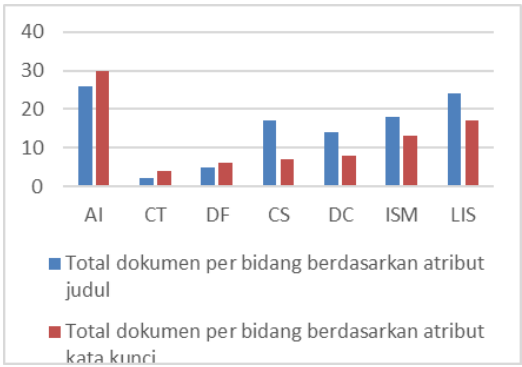
Implementasi dilakukan pada data studi kasus yang akan ditentukan *multiple membership* dokumennya. Data ini berasal dari dokumen publikasi ilmiah pada Kelompok Keilmuan (KK) RPL dan Informatika STEI ITB. Atribut dokumen publikasi ilmiah yang digunakan adalah judul dan kata kunci. Total publikasi ilmiah

yang digunakan pada kedua KK tersebut sejumlah 112 publikasi.



Gambar 7. Hasil *multiple membership* dokumen per kategori pada KK RPL

Gambar 7. menunjukkan bahwa total dokumen per bidang berdasarkan judul memiliki nilai yang lebih tinggi pada bidang AI, DC, LIS. Kemudian, Gambar 7. hasil *multiple membership* dokumen per kategori berdasarkan judul dan kata kunci pada KK Informatika.



Gambar 8. Hasil *multiple membership* dokumen per kategori pada KK Informatika

Gambar 8. menunjukkan bahwa total dokumen per bidang berdasarkan judul memiliki nilai yang lebih tinggi pada bidang CS, DC, ISM dan LIS.

Hasil grafik di atas menunjukkan bahwa penentuan *multiple membership* pada dokumen dengan atribut judul lebih banyak menghasilkan dokumen bersifat *multiple membership*. Dokumen ini bersifat *multiple membership* karena memiliki jarak antar presentase nilai *cosine similarity* antar vektor kategori berada di bawah nilai batas yang telah ditetapkan. Apabila jarak antara vektor kategori semakin kecil, maka dokumen tersebut semakin memiliki kemiripan pada kedua bidang tersebut.

KESIMPULAN

K-Means Clustering dapat digunakan untuk menentukan nilai batas *multiple membership* dokumen. Penggunaan judul dan kata kunci sama-sama dapat menghasilkan informasi *multiple membership* dokumen. Penentuan *multiple membership* dokumen dengan atribut kata kunci lebih jarang menghasilkan dokumen bersifat *multiple membership*. Kondisi ini terjadi karena jarak antar nilai presentase *cosine similarity* antar vektor kategori terlalu jauh dan berada di atas nilai batas yang telah ditetapkan.

DAFTAR PUSTAKA

Nevedov, N. (2011) : Community Detection and Its Applications for Mobile Networks “, Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Article No. 64,

Meng, Qinxue dan Kennedy, Paul (2012) : Using Field of Research to Discover Research Group from Co-Authorship. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*

Ning Liu, dkk (2004) : *Learning Similarity Measures in Non-orthogonal Space. Washington D.C. CIKM 04*.

Mandala, R dan Setiawan, H. (2002) : Peningkatan Performansi Sistem Temu Kembali Informasi dengan Perluasan Query Secara Otomatis.

Institut Teknologi Bandung.

Turney, Peter dan Pantel, Patrick. (2010) : From Frequency to Meaning : Vector Space Models of Semantics, *Journal of Artificial Intelligence Research, 37:141-188.*

Wesan, Barbakh And Colin Fyfe (2008) : *Local vs global interactions in clustering algorithms: Advances over K-means*. International Journal of knowledge-based and Intelilligent Engineering Systems 12.83 – 99.

Guo, Qinglin (2008) : The similarity Computing of Document based on VSM. Annual IEEE International Computer Software and Applications Conference

Pannu Mandeep, et al. (2014) : A Comparision of Information Retrieval Model, ACM, 978-1-4503-2899-9/14/05.

M. Durairaj, dkk. (2014) : Educational Data Mining for Prediction of Student Performance using Clustering Algorithms. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5987-5991

Jain,A.K, dkk. (1999) : Data Clustering: A Review. ACM Comput. Surv., sept, Volume 31, pp. 264—323Guo, Qinglin (2008) : The similarity Computing of Document based on VSM. Annual IEEE International Computer Software and Applications Conference

Trstenjak, B, dkk. (2013) : KNN with TF-IDF Based Framework for Text Categorization. *24th DAAAM International Symposium on Intelligent Manufacturing and Automation*.