Pengelompokan Dokumen Menggunakan Dokumen Berlabel dan Tidak Berlabel Dengan Pendekatan *Modified Heuristic Fuzzy Co-Clustering*

Khadijah F. Hayati[1], Resti Ludviani[2], Arini R. Rosyadi[3]
Program Studi Teknik Informatika, Institut Teknologi Sepuluh November
Jl. Sukolilo, Surabaya 60111, Jawa Timur
E-mail: [1]dee.jafa@gmail.com, [2]restiludvi@gmail.com, [3]arini.rosyadi@gmail.com

Masuk: 12 Juni 2014; Direvisi: 30 Juni 2014; Diterima: 10 Juli 2014

Abstract. *Document clustering is a method used to cluster the data of text documents in accordance with the categories of information held. However, a varyed number of text documents cause some problems in the process of document clustering. One of them is the result of hard clustering caused by unsupervised method. Thus proposed is a clustering using Modified Heuristic Fuzzy Co-clustering (HFCR) Algorithm with semi-supervised method that uses a document label as its learning process. The result of the experiments show that the proposed method achieves the best result when Tu=Tv=0.01, and it has precission value of 0.19 and recall value of 0.16. Modified Heuristic Fuzzy Co-clustering algorithm proposed gives more stable results compared to the results of standart HFCR method.*
Keywords: *Documents clustering, Fuzzy Co-clustering, Semi-supervised learning.*

Abstrak. *Pengelompokan dokumen merupakan suatu metode yang digunakan untuk dapat mengelompokkan suatu data berupa dokumen teks sesuai dengan kategori dari informasi yang dimiliki. Akan tetapi dengan banyaknya dokumen teks yang bervariasi menyebabkan beberapa masalah timbul dari proses pengelompokan dokumen. Salah satu diantaranya adalah hasil pengelompokan yang bersifat hard clustering. Hal ini disebabkan karena proses pengelompokan yang diterapkan merupakan metode unsupervised. Berdasarkan hal tersebut maka diajukan suatu metode pengelompokan yang mengunakan Algoritma Heuristic Fuzzy Co-clustering dengan menerapkan metode semi-supervised yang menggunakan dokumen berlabel sebagai proses pembelajarannya. Hasil uji coba terhadap metode yang diusulkan menunjukkan Algoritma Heuristic Fuzzy Co-clustering usulan terbaik dicapai pada kondisi Tu=Tv=dan 0,01 dengan nilai precission 0,19 dan recall 0,16. Algoritma Modified Heuristic Fuzzy Co-clustering yang diusulkan memberikan hasil lebih stabil dibandingkan dengan hasil pengelompokan Algoritma Heuristic Fuzzy Co-clustering.*
Kata Kunci: *Pengelompokan dokumen, Fuzzy Co-clustering, Semi-supervised learning.*

1. Pendahuluan

Perkembangan data meningkat dengan sangat signifikan dewasa ini. Jumlahnya meningkat dalam waktu yang singkat. Diantara data-data tersebut, data teks merupakan data yang banyak tersimpan (Cai, 2012). Akan tetapi, banyaknya data teks yang tersimpan mempersulit pengguna untuk dapat mendapatkan informasi yang sesuai dengan topik yang diinginkan. Oleh karena itu, pengguna membutuhkan suatu proses *filtering* untuk dapat mendapatkan informasi yang tepat.

Metode pengelompokan dokumen merupakan metode yang tepat untuk dapat melakukan pengelompokan suatu data dokumen teks yang besar, sehingga menjadi kelompok-kelompok kecil yang sesuai dengan kategori dari informasi yang terdapat dalam data teks (Yan, 2013). Pada prosesnya, pengelompokan dokumen merupakan teknik *unsupervised* yang tidak menggunakan data latih, yaitu berupa dokumen berlabel, untuk mendapatkan *learning model* (Yan, 2013) dan (Liu, 2013). Hal tersebut menyebabkan hasil yang didapatkan dari proses pengelompokan menjadi tidak optimal. Selain itu, kendala dalam dokumen teks adalah dapat ditemukannya lebih dari satu topik dalam satu dokumen, hal ini menyebabkan terjadinya *hard clustering*.

Penerapan metode *semi-supervised* pada proses pengelompokan dokumen dapat mengoptimalkan hasil pengelompokan menjadi lebih efektif. Karena metode *semi-supervised* melibatkan beberapa dokumen yang telah berlabel untuk digunakan dalam proses pengelompokan (Yan, 2013) dan (Liu, 2013).

Penelitian ini menerapkan Algoritma *Heuristic Fuzzy Co-clustering*, selain menggunakan metode *semi-supervised*. Algoritma *Heuristic Fuzzy Co-clustering* merupakan suatu algoritma yang cukup efektif untuk melakukan proses pengelompokan terhadap dokumen. Algoritma *Heuristic Fuzzy Co-clustering* digunakan pada banyak data yang bi-partisi, misalnya data bi-partisi berdasarkan gambar spektral dan juga dokumen teks yaitu frekuensi kata yang terdapat dalam suatu dokumen (Yan, 2013). Disamping itu, Algoritma *Heuristic Fuzzy Co-clustering* memiliki kelebihan untuk dapat menghasilkan *group* yang tidak bersifat *hard cluster*, yaitu anggota *group* yang satu dapat menjadi bagian dari *group* yang lain (Tjhi, 2008). Adapun tujuan dari penelitian ini adalah untuk melakukan perbandingan antara algoritma yang telah ada HFCR dengan modifikasi algoritma yang diajukan, HFCR dengan menerapkan metode *semi-supervised* pada proses pengelompokan dokumen.

Dalam penulisan hasil penelitian ini, dijabarkan menjadi beberapa bagian. Pada sub-bab kedua, diberikan beberapa penelitian terkait. Selanjutnya pada sub-bab ketiga diberikan penjelasan terkait dengan metode yang akan diusulkan dalam penelitian ini. Dan pada sub-bab keempat dan kelima diberikan skenario dan hasil pengujian dari sistem yang dibuat dan juga analisa dari hasil pengujian. Dan terakhir pada sub-bab keenam dipaparkan kesimpulan dari apa yang telah dilakukan dalam penelitian.

2. Tinjauan Pustaka

Dalam penelitian yang telah dilakukan sebelumnya (Tjhi, 2008), penelitian dilakukan untuk dapat memberikan jalan keluar terhadap permasalahan yang timbul dalam penerapan Algoritma *Fuzzy Clustering for Categorical Multivariate Data* (*FCCM*) dan *Fuzzy Co-Clustering of Document and Keyword* (*Fuzzy Codok*), yang menerapkan pendekatan *partitioning-ranking*. Permasalahan pertama adalah akurasi dalam proses pengelompokan. Hal ini disebabkan adanya tumpang tindih antara *feature* yang dilakukan perangkingan. Permasalahan kedua adalah pemrosesan antara *object* dan *feature* yang dilakukan secara berbeda.

Berdasarkan permasalahan tersebut, penelitian tersebut mengajukan suatu algoritma yang disebut dengan *Heuristic Fuzzy Co-clustering with the Ruspini's condition* (HFCR) yang menerapkan pendekatan *dual-partitioning*. Algoritma ini diajukan untuk dapat memberikan solusi pada kedua permasalahan, yaitu dengan menerapkan pendekatan partisi untuk digunakan pada *object* dan juga *membership feature*. Dengan demikian maka akurasi dari hasil pengelompokan menjadi lebih alami. Selain itu, algoritma HFCR diketahui merupakan algoritma yang mampu menangani *noise* yang terdapat dalam suatu dokumen.

Penelitian (Yan, 2013) merupakan penelitian lanjutan dari (Tjhi, 2008) yang memaparkan beberapa permasalahan yang terjadi pada pengelompokan dokumen antara lain ukuran dataset teks yang besar, *noise*, *overlapping*, dan ukuran dimensi yang tinggi. Berdasarkan hal tersebut, beberapa teknik dikembangkan seperti (a) *Co-clustering* yang efektif dalam menangani data dengan dimensi tinggi. (b) *Fuzzy clustering* yang merepresentasikan overlapping cluster. (c) *Model-based clustering* yang mampu menangani data *outlier*, tetapi memiliki kompleksitas tinggi. Di lain pihak, pengelompokan merupakan metode *unsupervised learning* yang masih sulit untuk menyelesaikan *dataset* yang kompleks. Untuk itu, diperlukan adanya penerapan pendekatan *semi-supervised*.

Penelitian ini mengusulkan metode dokumen pengelompokan yang mengkombinasikan kekuatan dari Teknik *Fuzzy Co-clustering* (Tjhi, 2008), dan metode *semi-supervised* yang disebut dengan *Heuristic Semi-Supervised Fuzzy Co-Clustering Algorithm* (SS-HFCR). Pengembangan metode ini bertujuan untuk meningkatkan akurasi pengelompokan, mengurangi sensitivitas terhadap parameter fuzzy dengan pengetahuan terbatas, serta menjaga kompleksitas algoritma relatif rendah.

Prinsip dasar *semi-supervised* dalam penelitian ini adalah memanfaatkan pengetahuan yang terbatas dalam bentuk kendala berpasangan (*pair-wise constraint*) pada dokumen. Batasan ditentukan dari dua dokumen dengan mempertimbangkan similaritas satu sama lain.

Pengujian dilakukan pada kasus *toy problem* dan beberapa dataset besar dengan kriteria evaluasi berdasarkan akurasi, stabilitas, dan efisiensi. Pada pengujian akurasi, SS-HFCR dibandingkan dengan PMFCC, SS-HFCR menghasilkan akurasi yang lebih baik untuk semua dataset. Pada pengujian stabilitas, SS-HFCR menunjukkan hasil yang lebih baik dari PMFCC pada jumlah *constraint* yang terus meningkat. Pada pengujian efisiensi waktu, SS-HFCR memiliki iterasi proses paling sedikit dibandingkan metode lainnya (SS-WNMF, Sd-Kmeans, SFCM) dengan kompleksitas algoritma yang setara dengan Sd-*Kmeans* & SFCM.

Di lain pihak, penelitian (Liu, 2013) mengemukakan bahwa metode pengelompokan yang banyak digunakan adalah metode yang bertipe diskriminatif atau generatif. Dalam proses pengelompokan ini, sering kali menghasilkan *group* yang bersifat *hard cluster*.

Pada penelitian ini, penulis menerapkan metode *semi-unsupervised* untuk melakukan proses pengelompokan dokumen, yaitu dengan menggunakan dokumen berlabel dan juga tidak berlabel, sehingga hasil pengelompokan yang didapatkan bersifat *soft cluster*.

Metode yang diajukan dalam penelitian ini adalah mengadopsi teknik dari metode pengelompokan dalam tipe diskriminatif dan juga generatif. Pada jenis diskriminatif, metode ini menggunakan teknik *K-Means* yang bertujuan untuk meminimalisir nilai jarak rata-rata antar *object* dan juga *centroid* dari tiap-tiap *group*. Dalam mengadaptasi jenis generatif, metode ini bertujuan untuk dapat menggabungkan fungsi *membership*, sehingga pada tiap satu *object* dapat menjadi bagian dari beberapa kelompok.

Pada penelitian ini dihasilkan bahwa dengan menggunakan Algoritma *Fuzzy Semi K-Means* ini, dapat menghasilkan *group* yang optimal. Walaupun terdapat beberapa batasan-batasan yang kurang jelas.

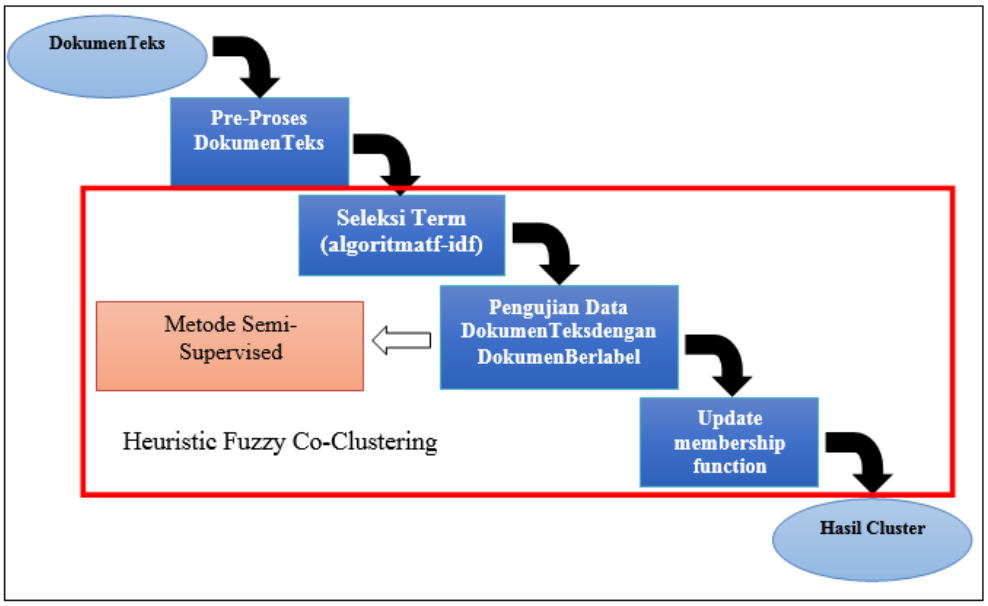
3. Metode yang diusulkan

Dari penelitian yang telah dianalisa pada sub-bab sebelumnya, maka dalam penelitian ini akan mengajukan suatu proses pengelompokan dokumen dengan menggunakan Algoritma *Heuristic Fuzzy Co-clustering* dengan menerapkan metode *semi-supervised*.

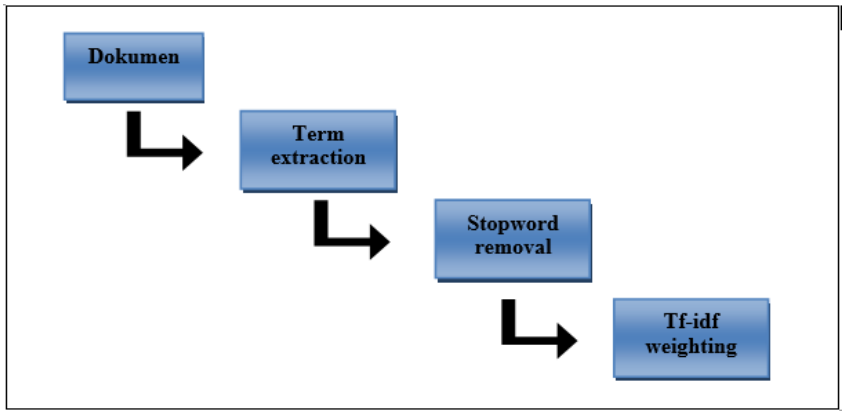
Dalam penelitian ini dilakukan kombinasi dari dua penelitian yang telah dilakukan. Penggunaan Algoritma *Heuristic Fuzzy Co-clustering* didasarkan pada penelitian (Tjhi, 2008) dan penerapan metode *semi-supervised* dilandaskan pada penelitian(Liu, 2013). Berbeda dengan pendekatan *semi-supervised* SS-HFCR yang menggunakan *pair-wise constraint* (Yan, 2013), metode yang diusulkan menerapkan *semi-supervised* dengan menggunakan dokumen berlabel dan tidak berlabel (Liu, 2013).

Usulan dari penelitian ini dijabarkan pada Gambar 1. Pada Alur metode digambarkan, diberikan suatu masukan yaitu suatu data berupa dokumen teks. Dokumen teks digunakan sebagai masukanan merupakan data yang didapatkan dari 20NewsGroup. Selanjutnya dokumen masukkan dilakukan proses pre-proses untuk mendapatkan term-term yang akan digunakan dalam proses pengelompokan dokumen(Chen, 2010).

Pada Gambar 1, proses pengelompokan digambarkan dalam suatu kotak yang berisikan beberapa proses. Termasuk didalamnya merupakan proses dari penerapan metode *semi-supervised*. Proses pengelompokan akan diawali dengan melakukan pre-proses terhadap dokumen masukan dengan melewati beberapa tahapan pre-proses pada dokumen teks (Gambar 2). Pada tahap pre-proses, dilakukan ekstraksi term dari masukan dokumen. Term yang tidak memberikan makna terhadap dokumen, yaitu term yang termasuk stopword dihapus. Proses selanjutnya adalah menyeleksi term-term yang telah didapatkan pada pre-proses. Penyeleksian term-term, sebelumnya dilakukan pembobotan term-term tersebut dengan menggunakan Algoritma tf-idf. Pada proses ini term-term yang diseleksi tidak lagi berupa kata akan tetapi berupa suatu nilai bobot dari kata tersebut (Chen, 2010) dalam bentuk *vector* bobot dokumen.



Gambar 1. Alur metode yang diajukan



Gambar 2. Pre-proses pada dokumen teks

Pada tahap selanjutnya merupakan metode *semi-supervised*. Tahapan ini dilakukan dengan melakukan pengujian terhadap dokumen masukan dengan dokumen berlabel yang telah tersimpan. Pengujian dilakukan untuk mencocokkan kedua dokumen tersebut. Dokumen masukan yang dianggap sama dengan dokumen berlabel akan dikelompokkan pada kluster yang sesuai dengan dokumen berlabel. Proses pencocokan ini merupakan suatu langkah untuk menciptakan *learning model*.

Setelah tahapan pengujian dokumen masukan dengan dokumen berlabel, proses dilanjutkan dengan melakukan perubahan terhadap nilai membership pada formula (1) dan juga nilai objek pada formula (2). Proses perubahan terhadap nilai tersebut, dilakukan berdasarkan dari hasil proses sebelumnya. Perubahan akan dilakukan jika dokumen masukan tidak termasuk dalam dokumen berlabel.

Berdasarkan penelitian Tjhi, rumus perubahan derajat keanggotaan objek dan fitur ditunjukkan pada persamaan (1) dan (2).

$$u_{ci} = \frac{\left\{\frac{\sum_{j=1}^{K} v_{cj} d_{ij}}{T_u \sum_{j=1}^{K} v_{cj}}\right\}}{\sum_{f=1}^{C} \left\{\frac{\sum_{j=1}^{K} v_{fj} d_{ij}}{T_u \sum_{j=1}^{K} v_{fj}}\right\}} \quad (1)$$

$$v_{cj} = \frac{\left\{\frac{\sum_{i=1}^{N} u_{ci} d_{ij}}{T_v \sum_{i=1}^{N} u_{ci}}\right\}}{\sum_{f=1}^{C} \left\{\frac{\sum_{i=1}^{N} u_{fi} d_{ij}}{T_v \sum_{i=1}^{N} u_{fi}}\right\}} \quad (2)$$

Keterangan:

K = jumlah fitur (*key term*)
N = jumlah objek (dokumen)
C = jumlah *co-cluster*
d_{ij} = nilai ukuran antara objek dengan fitur, dalam kasus ini $d_{ij} \geq 0$
T_u, T_v = nilai keanggotaan *co-cluster*
u_{ci}, v_{cj} = nilai keanggotaan dari objek dan fitur

Proses diatas berulang sesuai dengan jumlah dokumen masukan atau akan terhenti jika telah mendekati nilai *error rate* yang telah ditentukan pada awal proses berjalan. *Pseudo-code* HFCR (Tjhi, 2008) ditunjukkan Kode 1, sedangkan *pseudo-code* dari metode yang diusulkan diberikan pada Kode 2.

Kode 1. Algoritma *Heuristic Fuzzy Co-clustering* (HFCR)

```
1.  Inisialisasi Tu , Tv , C, tmax, ε
2.  Masukan: matrix term selection, SC
3.  Begin
4.      set parameters Tu , Tv , C, tmax, ε, N
5.      set t=0
6.      for i=1 to N do
7.          for c=1 to C do
8.                Uci = random()
9.                Repeat
10.                  Update Vcj
11.                  Update Uci
12.                  Update t = t + 1
13.               until max(c,i)|Uci(t) – Uci(t-1)|<= ε or t=tmax
14. End
```

Kode 2. Modifikasi Algoritma *Heuristic Fuzzy Co-clustering* (HFCR) yang diusulkan

```
1.  Inisialisasi Tu , Tv , C, tmax, ε
2.  Masukan: matrix term selection, SC
3.  Begin
4.      set parameters Tu , Tv , C, tmax, ε, N
5.      set t=0
6.      for i=1 to N do
7.          for c=1 to C do
8.              if  di is document of SC then
9.                                  Uci  ← 1
10.                                 Uci' ← 0
11.                       Break
12.                     Else
13.                      Uci = random()
14.                             Repeat
15.                                    Update Vcj
16.                                    Update Uci
17.                     Update t = t + 1
18.                             until max(c,i)|Uci(t) – Uci(t-1)|<= ε or t=tmax
19. End
```

4. Skenario dan Hasil Pengujian

4.1 Skenario Pengujian

Pada tahap ini, pengujian terhadap metode yang diajukan dilakukan dengan menggunakan dataset sebanyak 600 dokumen teks yang didapatkan dari situs penyedia dataset dokumen teks, yaitu 20NewsGroup. Dokumen-dokumen tersebut merupakan dokumen berita yang tidak berlabel yang termasuk dalam beberapa kategori yang berbeda. Pada pengujian ini dokumen berlabel digunakan dalam penggujian metode *semi-supservised* menggunakan data training sebanyak 200 dokumen berlabel yang didapatkan dari 20NewsGroup. Dokumen berlabel tersebut diambil secara acak dari 6 kategori berita dalam 20NewsGroup. Dokumen berlabel berperan sebagai pembanding terhadap dokumen dataset yang dikelompokkan.

Pengujian dilakukan untuk membandingkan hasil dari pengelompokan dokumen dari 2 metode, yaitu metode HFCR dan HFCR yang diusulkan. Dokumen masukan sejumlah 600 dokumen dikelompokkan kedalam enam kluster menggunakan metode HFCR dan HFCR yang diusulkan. Pengujian terhadap masing-masing metode dilakukan sebanyak empat kali pengujian dengan variasi inisialisasi nilai variabel berturut-turut adalah Tv=Tu = 0,001, Tv=Tu=0,01, Tv=Tu=0,1, dan Tv=Tu=1. Hasil uji coba tersebut dievaluasi menggunakan perhitungan *recall* dan *precision* pada persamaan 3 dan 4.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+TN} \quad (4)$$

Sebagaimana yang ditunjukkan pada tabel 1, TP (*true positif*) merupakan jumlah dokumen yang dikelompokkan secara tepat oleh system. FP (*false positif*) merupakan jumlah dokumen yang berada pada kelompok dokumen yang tidak tepat. TN (*true negatif*) merupakan jumlah dokumen yang tidak ter-*retrieve*. Berdasarkan evaluasi menggunakan *precission*, maka efektivitas dari metode yang digunakan dapat diketahui. Evaluasi dengan *recall* akan menunjukkan kemampuan metode dalam mengelompokkan dokumen dengan tepat.

Tabel 1. *Recall* dan Precision

		Clustering secara Manual	
		Dokumen yang Relevan	**Dokumen yang Tidak Relevan**
Clustering menggunakan modifikasi algoritma	**Dokumen yang Ditemukan**	*True Positive*	*False Posotive*
	Dokumen yang Tidak Ditemukan	*True Negative*	*False Negative*

4.2 Hasil Pengujian

Hasil pengujian dan perbandingan terhadap Algoritma *Heuristic Fuzzy Co-clustering* dengan Algoritma *Heuristic Fuzzy Co-clustering* usulan ditunjukkan pada Tabel 2 sampai Tabel 5 berikut ini. Pengujian dilakukan dengan variasi inisialisasi nilai variabel berturut-turut (a) Tv=Tu = 0,001. (b) Tv=Tu=0,01. (c) Tv=Tu=0,1. (d) Tv=Tu=1 terhadap masing-masing metode. Variasi inisialisasi variabel Tv dan Tu tersebut menjadi kriteria pengukuran *recall* dan precision pada penelitian ini. Variabel Tv dan Tu yang digunakan pada penelitian ini merupakan nilai keanggotaan co-clustering yang digunakan pada perhitungan nilai membership (Uci) dengan formula 1 dan nilai objek (Vcj) dengan formula 2. Dengan demikian, perbedaan nilai Tv dan Tu diasumsikan akan mempengaruhi hasil dari pengelompokan dokumen. Hal ini dibuktikan dari hasil uji coba yang telah dilakukan.

Pada Tabel 2 diberikan hasil perhitungan Precision pada masing-masing hasil pengelompokan beserta dengan nilai rata-rata dari Algoritma *Heuristic Fuzzy Co-clustering*. Dan Tabel 3 adalah nilai dari precision dari Algoritma *Heuristic Fuzzy Co-clustering* yang

diusulkan. Tabel 4 dan Tabel 5 merupakan hasil perhitungan *recall* dari Algoritma *Heuristic Fuzzy Co-clustering* dan Algoritma *Heuristic Fuzzy Co-clustering* yang diusulkan.

Tabel 2. Nilai *Precision* Algoritma *Heuristic Fuzzy Co-clustering*

Cluster	HFCR			
	Nilai Tv Tu			
	0,001	**0,01**	**0,1**	**1**
1	0,26	0,65	0,39	0,31
2	0,17	0,17	0,24	0
3	0,16	0,08	0,17	0,12
4	0	0	0	0,01
5	0	0,42	0,03	0,06
6	0,06	0,04	0,16	0
Rata-rata	**0,11**	**0,23**	**0,17**	**0,08**

Tabel 3. Nilai *Precision* Algoritma *Heuristic Fuzzy Co-clustering* Usulan

Cluster	HFCR USULAN			
	Nilai Tv Tu			
	0,001	**0,01**	**0,1**	**1**
1	0,28	0,54	0,48	0,5
2	0,38	0,08	0,04	0
3	0,12	0,21	0,05	0,22
4	0	0	0	0
5	0,13	0,25	0,36	0,2
6	0,17	0,05	0	0,04
rata-rata	**0,17**	**0,19**	**0,15**	**0,16**

Tabel 4. Nilai *Recall* Algoritma *Heuristic Fuzzy Co-clustering*

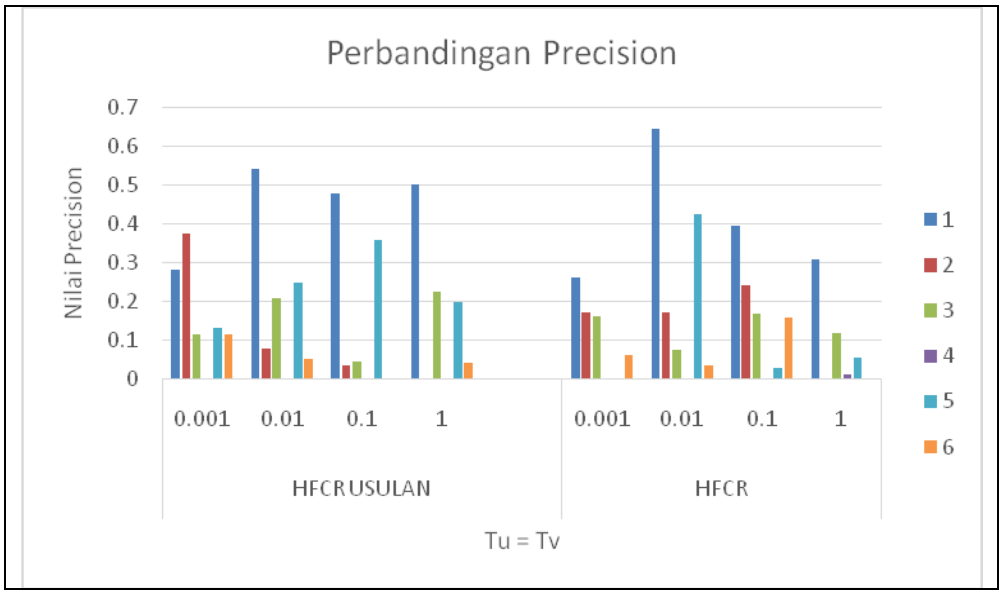
Cluster	HFCR			
	Nilau Tv Tu			
	0,001	**0,01**	**0,1**	**1**
1	0,10	0,07	0,36	0,06
2	0,22	0,06	0,09	0
3	0,06	0,01	0,10	0,08
4	0	0	0	1
5	0	0,43	0,05	0,26
6	0,30	0,47	0,45	0
Rata-rata	**0,11**	**0,17**	**0,17**	**0,23**

Tabel 5. Nilai *Recall* Algoritma *Heuristic Fuzzy Co-clustering* Usulan

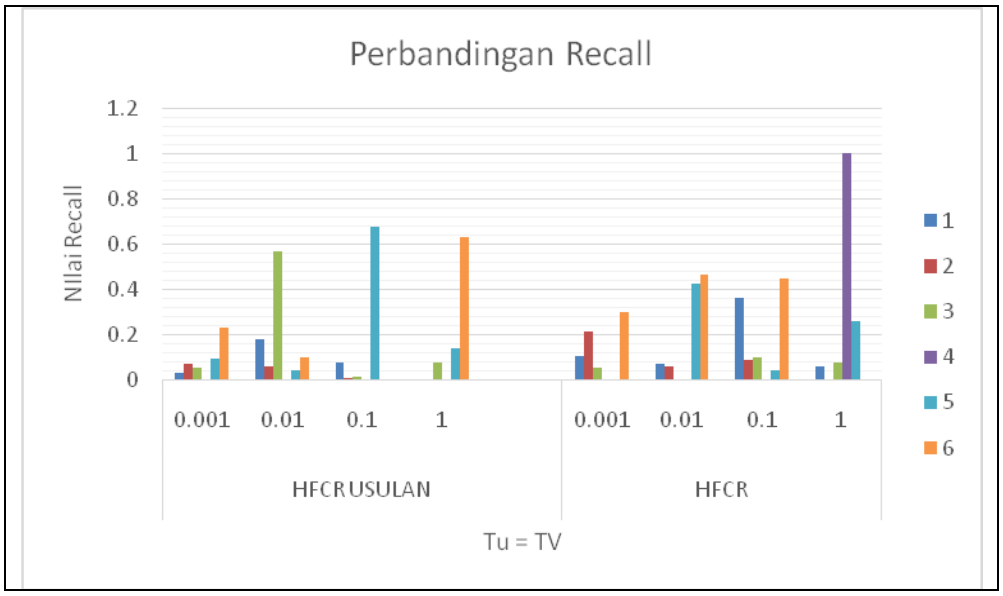
Cluster	HFCR USULAN			
	Nilai Tv Tu			
	0,001	**0,01**	**0,1**	**1**
1	0,03	0,18	0,08	0,01
2	0,08	0,06	0,01	0
3	0,06	0,57	0,01	0,08
4	0	0	0	0
5	0,10	0,05	0,68	0,14
6	0,23	0,1	0	0,63
rata-rata	**0,08**	**0,16**	**0,13**	**0,14**

5. Diskusi

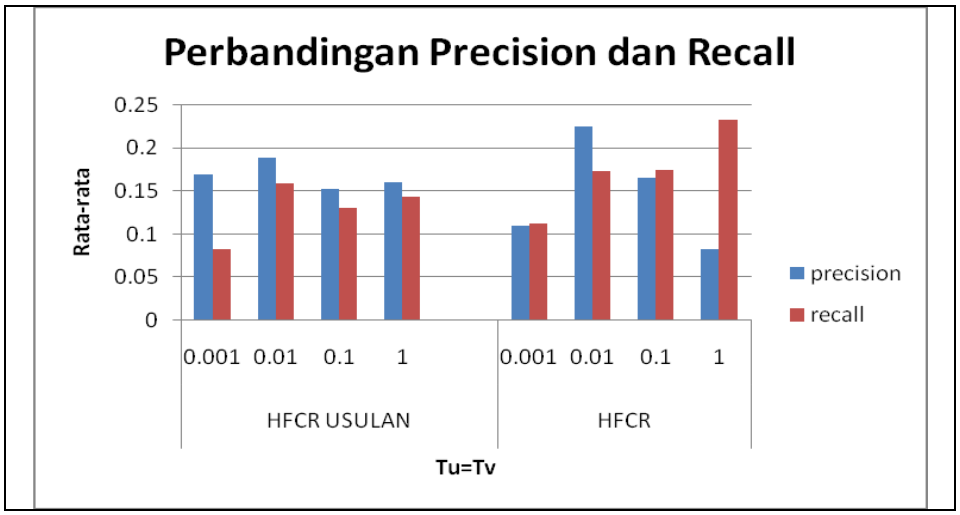
Pada Gambar 3, nilai *precission* terbaik pada Algoritma *Heuristic Fuzzy Co-clustering* usulan dicapai pada kondisi Tu=Tv=0,01 dengan nilai 0,19. Demikian pula pada Algoritma *Heuristic Fuzzy Co-clustering* R, nilai *precission* terbaik terjadi pada kondisi Tu=Tv=0,01 dengan nilai 0,26. Dari perbandingan rata-rata *precission* antara kedua metode tersebut (gambar 5), tampak bahwa HFCR usulan memiliki nilai *precission* yang lebih tinggi dibandingkan metode HFCR pada kondisi Tu=Tv=0,001 dan 1. Sedangkan pada kondisi Tu=Tv=0,01 dan 0,1, metode HFCR memiliki nilai *precission* lebih tinggi dibandingkan metode usulan. *Precission* pada tiap enam kluster hasil pengujian menunjukkan bahwa data kluster pertama memiliki nilai *precission* yang lebih tinggi dibandingkan kelima kluster lainnya baik pada metode HFCR usulan maupun HFCR, kecuali pada kodisi Tu=Tv=0,001 dengan metode HFCR usulan.



Gambar 3. Perbandingan Nilai *Precision* Algoritma *Heuristic Fuzzy Co-clustering* dengan Algoritma *Heuristic Fuzzy Co-clustering* usulan



Gambar 4. Perbandingan Nilai *Recall* Algoritma *Heuristic Fuzzy Co-clustering* dengan Algoritma *Heuristic Fuzzy Co-clustering* usulan



Gambar 5. Perbandingan Nilai rata-rata *Precision* dan *Recall* Algoritma *Heuristic Fuzzy Co-clustering* dengan Algoritma *Heuristic Fuzzy Co-clustering* usulan

Pada Gambar 4, nilai *recall* terbaik pada Algoritma *Heuristic Fuzzy Co-clustering* usulan dicapai pada kondisi Tu=Tv=0,01 dengan nilai sebesar 0,16. Sedangkan pada Algoritma *Heuristic Fuzzy Co-clustering*, nilai *recall* terbaik terjadi pada kondisi Tu=Tv=1 dengan nilai 0,23. Perbandingan rata-rata *recall* antara kedua metode tersebut (gambar 5) menunjukkan bahwa nilai *recall* pada metode HFCR usulan pada masing-masing kondisi Tu=Tv lebih rendah dibandingkan dengan nilai *recall* metode HFCR.

Secara keseluruhan, hasil pengujian menunjukkan bahwa metode HFCR usulan lebih stabil terhadap kondisi Tu=Tv dibandingkan metode HFCR. Hal ini dapat dilihat dari nilai *precission* pada tiap kondisi Tu=Tv yang cenderung rata dengan selisih nilai yang kecil. Keberagaman nilai *pecission* dan *recall* terhadap enam kluster pada tiap metode dipengaruhi oleh pengambilan 600 data cluster dan 200 data berlabel secara acak dengan komposisi jumlah yang tidak merata dari tiap kategori dokumen berita pada 20NewsGroup.

Uji coba dengan parameter kondisi Tu=Tv yang berbeda ini menunjukkan bahwa nilai Tv dan Tu memberikan pengaruh dalam perhitungan *fuzzy* dalam menentukan derajat keanggotaan yang digunakan dalam penentuan kelompok kluster.

6. Kesimpulan

Berdasarkan hasil pengujian, hasil pengelompokan Algoritma *Heuristic Fuzzy Co-clustering* usulan terbaik dicapai pada kondisi Tu=Tv=dan 0,01 dengan nilai *precission* 0,19 dan 0,16. Algoritma *Heuristic Fuzzy Co-clustering* usulan memberikan hasil lebih stabil dibandingkan dengan hasil pengelompokan Algoritma *Heuristic Fuzzy Co-clustering*.

Penelitian selanjutnya dapat difokuskan pada tahap seleksi term untuk menangani masalah dimensi tinggi. Hal ini memberikan pengaruh pada waktu komputasi yang dibutuhkan pada tahap *clustering* karena semakin banyak jumlah *term* maka dimensi dari vektor dokumen, sebagai masukan dari metode HFCR yang diusulkan, akan semakin besar.

Referensi

Cai, Xiaoyan., Li, Wenjie., 2012., Mutually Reinforced Manifold-Ranking Based Relevance Propagation Model for Query-Focused Multi-Document Summarization., Ieee Transactions on Audio, Speech, and Language Processing., VOL. 20,

Yan, Yang., Chen, Lihui., Tjhi, W-C., 2013., Fuzzy semi-supervised co-clustering for Text Document. Fuzzy Sets and Systems., Vol. 215., Pg. 74 – 89.

Liu, Chien-Liang., Chang, Tao-Hsing., Li, Hsuan-Hsun., 2013., Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans., Fuzzy Sets and Systems., Vol. 221., Pg. 48 – 64.

Tjhi, William C., Chen, Lihui., 2008., A heuristic-based fuzzy co-clustering algorithm for categorization of high-dimensional data., Fuzzy Sets and Systems., Vol. 159., Pg. 371 – 389.

Chen, Chun-Ling., Tseng, Frank S.C., Liang, Tyne., 2010,, Mining fuzzy frequent itemsets for hierarchical document clustering., Information Processing and Management., Vol. 46., Pg. 193–211.