EKSTRAKSI KATA KUNCI BERDASARKAN HIPERNIM DENGAN INISIALISASI KLASTER MENGGUNAKAN *FUZZY ASSOCIATION RULE MINING* PADA PENGELOMPOKAN DOKUMEN

Fahrur Rozi[1)], Chastine Fatichah[2)], dan Diana Purwitasari[3)]
[1, 2, 3)]Institut Teknologi Sepuluh Nopember
Jurusan Teknik Informatika, Fakultas Teknologi Informasi, ITS Surabaya 60111
e-mail: rozi.fahrur04@gmail.com[1)],
chastine.fatichah@gmail.com[2)], diana.purwitasari@gmail.com[3)]

ABSTRAK

Pertumbuhan dunia digital dalam dokumen tekstual terutama di World Wide Web mengalami pertumbuhan pesat. Peningkatan dokumen tekstual ini menyebabkan terjadinya penumpukan informasi, sehingga diperlukan sebuah pengorganisasian yang efisien untuk pengelolaan dokumen tekstual. Salah satu metode yang dapat mengelompokkan dokumen dengan tepat adalah menggunakan fuzzy association rule. Tahap ekstraksi kata kunci serta tipe fuzzy yang digunakan berpengaruh terhadap kualitas pengelompokan dokumen. Penggunaan hipernim dalam ekstraksi kata kunci untuk mendapatkan suatu klaster label dapat memperluas makna dari klaster label, sehingga dapat diperoleh suatu meaningful klaster label, selain itu ambiguitas dan uncertainties yang terjadi di dalam aturan fuzzy logic systems (FLS) tipe-1 dapat diatasi dengan fuzzy set tipe-2. Penelitian ini mengusulkan sebuah metode yaitu ekstraksi kata kunci berdasarkan hipernim dengan inisialisasi klaster menggunakan fuzzy association rule mining pada pengelompokan dokumen. Metode ini terdiri dari empat tahap, yaitu : preprocessing dokumen, ekstraksi key terms dari hipernim, ekstraksi kandidat klaster, dan konstruksi klaster tree. Pengujian terhadap metode ini dilakukan dengan tiga jenis data berbeda, yaitu Classic, Reuters, dan 20 Newsgroup. Pengujian dilakukan dengan membandingkan nilai overall f-measure dari metode tanpa hipernim (level 0), hipernim level 1, dan hipernim level 2. Berdasarkan pengujian didapatkan bahwa penggunaan hipernim dalam ektraksi kata kunci mampu menghasilkan rata-rata overall f-measure sebesar 0.5783 untuk data classic, 0.4001 untuk data reuters, dan 0.5269 untuk data 20 newsgroup.

***Kata Kunci**: Fuzzy set tipe-2, hipernim, association rule, clustering dokumen.*

ABSTRACT

The growth of the digital world in textual documents, especially on the World Wide Web is incredibly fast. Increase of textual document causes the accumulation of information, so we need an efficient organization to manage textual documents. One of method that can accurately classify documents is using fuzzy association rule. Phase extraction of key terms and type of fuzzy that used for clustering affected on the quality of the document clustering. Hypernym that use in the extraction of key terms to obtain a cluster label can expand the meaning of cluster labels and obtain a meaningful cluster labels, in addition to the ambiguities and uncertainties that occur in the rules of fuzzy logic systems (FLS) type-1 can be overcome with fuzzy sets type-2. This study propose a method of key terms extraction based on hypernym with initialization cluster using fuzzy association rule mining in document clustering. This method consists of four stages, that is: preprocessing documents, extracting key terms with hypernym, extraction of candidate clusters, and cluster tree construction. Testing of this method is done by three different types of data, that is : Classic, Reuters, and 20 Newsgroup. Testing is done by comparing overall f-measure of method without hypernym (level 0), hypernym level 1, and hypernym level 2. Based on testing, method with hypernym in the extraction of keyword can produce overall f-measure 0.5783 for classic data, 0.4001 for reuters data, and 0.5269 for 20 newsgroup data.

***Keywords**: Fuzzy set type-2, hypernym, association rule, document clustering.*

I. PENDAHULUAN

CLUSTERING dokumen (pengelompokan teks) merupakan salah satu metode *text mining* yang dikembangkan untuk mengefisienkan pengelolaan teks serta peringkasan teks [1]. Beberapa hal yang dapat meningkatkan kualitas *clustering* dokumen antara lain : mengatasi dimensi tinggi yang diakibatkan besarnya jumlah dokumen dan jumlah kata dalam dokumen, meningkatkan skalabilitas agar mampu bekerja dengan jumlah dokumen dalam skala kecil ataupun besar (*scalable*), meningkatkan akurasi, memberikan *label cluster* yang bermakna, mampu mengatasi *overlapping*, serta memperhitungkan kesamaan konseptual istilah dari kata [2] .

Beberapa metode telah dikembangkan untuk mendapatkan *clustering* dokumen dengan kualitas yang baik. Penggunaan *fuzzy* untuk *clustering* dokumen [3] dengan cara menerapkan *α-threshold Fuzzy Similiarity Classification Method* (α-FSCM) dan *Multiple Categories Vector Method* (MCVM). Penggunaan metode *fuzzy tipe-1* ini mampu menghasilkan *cluster* yang *overlapping*. *High dimensionality* merupakan salah satu permasalahan dari *clustering* dokumen, untuk mengatasi permasalahan ini Beil dkk [4] mengembangkan

algoritma *frequent itemset* yaitu *Hierarchical Frequent Term-based Clustering* (HFTC). Namun, berdasarkan penelitian Fung dkk [5] bahwa HFTC tidak *scalable*. Sehingga untuk menghasilkan metode yang *scalable*, Fung dkk mengembangkan metode *Frequent Itemset Hierarchical Clustering* (FIHC) yang merupakan algoritma hasil pengembangan *frequent-itemset* yang berasal dari *association rule mining* untuk membangun *hierarchical tree* untuk topik *cluster*. Penggabungan antara *fuzzy* dan *association rule mining* [6] yaitu *Fuzzy Frequent Itemset-Based Hierarchical Clustering* (F^2IHC) mampu meningkatkan tingkat akurasi serta menghasilkan *cluster* yang *overlapping* dalam *clustering* dokumen.

Beberapa penelitian *clustering* dokumen HFTC [4], FIHC [5], dan F^2HIC dengan *fuzzy* set tipe-2 [7] masih menggunakan *term* yang berada dalam dokumen teks sebagai *label cluster*. Meskipun hal tersebut dibenarkan, namun pelabelan *cluster* yang lebih umum akan memudahkan melakukan analisis terutama dalam domain pengetahuan [8]. Penggunaan hipernim berdasarkan *Wordnet* dapat memperluas dalam pencarian *hidden similiarities* untuk mengidentifikasi topik dalam dokumen [2]. Sebagai contoh, jika beberapa dokumen memiliki topik "kursi","meja", dan "almari", maka label *cluster* yang terbaik adalah hipernimnya yaitu "*furniture*". Pada penelitiannya yang lain Chen dkk [2] mengembangkan sebuah metode untuk *Fuzzy based Multi-label Document Clustering* (FMDC) dengan menggunakan *fuzzy assocation mining* yang terintegrasi dengan *Wordnet*. Metode yang dikembangkan Chen dkk, terdapat suatu tahap untuk menemukan *key term set* dengan menggunakan hipernim berdasarkan *Wordnet*. Penambahan *key term* dari hasil hipernim akan memperluas makna dari label *cluster* sehingga mendapatkan suatu *meaningful* label *cluster*. Sementara, Tseng [8] dalam penelitiannya menggunakan algoritma *hypernym search* untuk mendapatkan *label clustering* yang *generic*. Tseng, mengekstraksi *term* yang memiliki kategori secara spesifik untuk digunakan sebagai calon label *cluster*. Setiap calon label *cluster* akan diperluas menjadi *term* yang lebih umum berdasarkan *hypernym search*

Ambiguitas dan *uncertainties* yang terjadi di dalam aturan *fuzzy logic systems* (FLS) tipe-1 [9] dapat mengurangi tingkat akurasi dalam *clustering* dokumen. Terdapat beberapa ambiguitas dan *uncertainties* pada *fuzzy* set tipe-1, yaitu : kata yang digunakan dalam *antecendent* dan *consequent* memiliki arti berbeda bagi setiap orang dan data masukan yang digunakan dalam *fuzzy* tipe-1 dapat dimungkinkan merupakan *noise* akibat dari penentuan *range* untuk variabel linguistik yang berbeda bagi setiap orang. *Fuzzy set* tipe-2 mampu menutupi kelemahan yang terdapat dalam *fuzzy* tipe-1 [9]. Penelitian [7] [9] [10] [11] [12] dengan menggunakan *fuzzy* set tipe-2, memberikan hasil bahwa *fuzzy* tipe-2 mampu mengatasi kelemahan yang terjadi pada *fuzzy* tipe-1 serta penggunaan *fuzzy* tipe-2 lebih baik dibanding dengan menggunakan *fuzzy* tipe-1. Selain itu penggunaan hipernim untuk mendapatkan suatu *cluster* label dapat memperluas makna dari *cluster* label, sehingga dapat diperoleh suatu *meaningful cluster* label. Oleh karena itu, penelitian ini bertujuan membangun metode ekstraksi kata kunci berdasarkan hipernim dengan inisialisasi klaster menggunakan *fuzzy association rule mining* pada pengelompokan dokumen.

II. METODE

Metode penelitian ini terdiri atas empat bagian utama yaitu : *preprocessing* dokumen, ekstraksi *key term* dari hipernim, ekstraksi *candidate cluster*, dan konstruksi *cluster tree*.

A. Preprocessing Dokumen

Terdapat beberapa tahap yang dilakukan dalam *preprocessing* dokumen, yaitu : ektraksi *term*, penghilangan stopwords, stemming, dan seleksi *term*. Pada tahap awal, hasil dari ekstraksi dokumen dikumpulkan dalam suatu koleksi *single word* $T_D = \{t_1, t_2, \ldots, t_n\}$. T_D menyatakan koleksi *term* (t) dalam dokumen (D), n menyatakan jumlah *term* dalam T_D. Hasil yang didapatkan dari ekstraksi *term* T_D digunakan sebagai *input* untuk dilanjutkan dengan penghilangan *stopwords* dan proses *stemming*. Algoritma *stemming* yang digunakan dalam penelitian ini adalah Porter stemmer yang ditemukan oleh Martin Porter pada tahun 1980. Langkah terakhir yang dilakukan dalam *preprocessing* dokumen adalah seleksi *term* dengan menghitung bobot $tfidf$ (1) setiap *term* dalam T_D.

$$tf.idf_{ij} = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}} \times \log\left(\frac{|D|}{|\{d_i | t_j \in d_i, d_i \in D|\}|}\right), \quad (1)$$

dimana $tf.idf_{ij}$ adalah bobot *term* t_j dalam dokumen d_i. Untuk mencegah bias dokumen yang panjang, bobot frekuensi *term* f_{ij} dinormalisasi dengan total frekuensi semua *term* dalam dokumen d_i. Variabel $|D|$ adalah jumlah seluruh dokumen dan $\left|\{d_i | t_j \in d_i, d_i \in D\}\right|$ adalah jumlah dokumen yang memiliki *term* t_j.

B. Ekstraksi Key terms dari Hipernim

Hipernim dari suatu term dilakukan pencarian berdasarkan dari *Wordnet*. Contoh urutan peringkat dari hipernim h adalah $h_1 \leq h_2$, yaitu peringkat hipernim h_1 lebih kecil dari h_2 jika h_2 adalah hipernim dari h_1. Perhitungan frekuensi dari hipernim dilakukan dengan menggunakan persamaan (2).

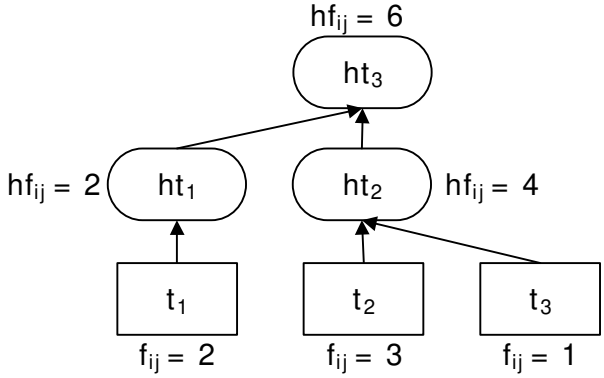
$$hf_{ij} = hf_{ij} + f_{ij}, \quad (2)$$

dimana f_{ij} adalah frekuensi *term* j dalam dokumen i, dan hf_{ij} adalah frekuensi hipernim dari *term* t_j dalam dokumen d_i. Contoh perhitungan dari frekuensi hipernim terdapat dalam Gambar 1. Berdasarkan Gambar 1, diketahui bahwa *term* t_1 dengan frekuensi $f_{ij} = 2$ memiliki hipernim ht_1, *term* t_2 dengan frekuensi $f_{ij} = 3$ memiliki hipernim ht_2 dan ht_3, sementara *term* t_3 dengan frekuensi $f_{ij} = 1$ memiliki hipernim ht_2. Frekuensi dari *term* ht_1 adalah $hf_{ij} = 2$, karena ht_1 merupakan hipernim dari t_1. Untuk frekuensi ht_2 adalah $hf_{ij} = 4$, karena ht_2 merupakan hipernim dari t_2 dan dan t_3 yang menjumlahkan frekuensi masing – masing yaitu : 3 dan 1. Sementara *term* ht_3 memiliki frekuensi $hf_{ij} = 6$, karena menjumlahkan frekuensi dari ht_1 dan ht_2 yaitu : 2 dan 4.

C. Ekstraksi Kandidat Cluster

Terdapat empat proses yang harus dilalui untuk mendapatkan kandidat *cluster*, diantaranya : menghitung nilai membership function dengan *fuzzy* set tipe-2, menemukan *candidate*-1 *itemset*, menemukan *candidate*-2 *itemset*, dan seleksi kandidat *cluster*. *Fuzzy* set tipe-2 dalam penelitian ini menggunakan dua jenis tipe fungsi keanggotaan, yaitu : fungsi kenggotaan jenis *triangular* sebagai LMF (*Lower Membership Function*) dan fungsi keanggotaan jenis *trapezoidal* sebagai UMF (*Upper Membership Function*). Setiap *term* j dalam dokumen i dengan frekuensi f_{ij} memiliki bobot $w_{ij}^{r,z}$ yang menyatakan bobot atau fungsi keanggotaan *term* j dalam dokumen i yang terdapat dalam wilayah fungsi keanggotaan *fuzzy* set tipe-2. Variabel r dalam $w_{ij}^{r,z}$ merupakan variabel linguistik, yaitu : *Low*, *Medium*, dan *High*. Sementara z merepresentasikan LMF dan UMF.

Hasil bobot *fuzzy* tipe-2 dari setiap *term* selanjutnya akan digunakan untuk menentukan *candidate* 1-frequent *itemset*. Untuk menemukan *term* yang digunakan sebagai *candidate* 1-*itemset*, setiap *term* dilakukan perhitungan nilai *support*. Perhitungan nilai *support* didapatkan dari hasil perbandingan antara nilai bobot *fuzzy* dengan jumlah dokumen. Hasil *term* j yang diperoleh dari *candidate* 1-*itemset* akan diasosiasikan terhadap *term* yang lain untuk mendapatkan *candidate* 2-*itemset*. Setiap pasang term yang memiliki nilai *support* dan *confidence* lebih dari minimum support dan minimum confidence akan dijadikan sebagai *candidate* 2-*itemset*. Hasil dari *candidat* 1-*itemset* dan *candidate* 2-*itemset* dijadikan sebagai candidat *cluster* set $\tilde{C}_D = \{\tilde{c}_1^1, \ldots, \tilde{c}_{l-1}^2, \tilde{c}_l^q, \ldots, \tilde{c}_k^q\}$, dimana D merupakan koleksi dokumen, q merupakan jumlah q-*itemset*, dan k adalah jumlah semua kandidat *cluster* c yang didapatkan dari *candidat* 1-*itemset* dan *candidate* 2-*itemset*.



Gambar 1. Contoh Perhitungan Frekuensi Hipernim

D. Konstruksi Tree

Untuk membentuk *cluster tree* dibutuhkan beberapa tahap, yaitu membentuk *Document-Term Matrix* (DTM), membentuk *Term-Cluster Matrix* (TCM), dan membentuk *Document-Cluster Matrix* (DCM). *Document-Term Matrix* (DTM) atau matriks $W = [w_{ij}^{max-R_j}]$, dimana $w_{ij}^{max-R_j}$ adalah bobot (nilai fungsi keanggotaan) dari *term* t_j dalam dokumen d_i. Matriks ini merupakan representasi dari kumpulan nilai maksimum fungsi keanggotaan dari tiap *term* t_j dalam dokumen d_i dengan ukuran $n \times p$, dengan n adalah jumlah dokumen d_i dalam koleksi dokumen D, dan p adalah jumlah *key term t* dari hasil ekstraksi *candidate* 1-*itemset*. Ilustrasi dari matriks DTM terdapat pada gambar 2.

Setelah terbentuk matriks DTM, selanjutnya adalah pembentukan *Term-Cluster Matrix* (TCM) atau matriks $G = [g_{jl}^{max-R_j}]$ dengan ukuran $p \times k$, dimana p adalah jumlah *key term t* dari hasil ekstraksi *candidate* 1-*itemset*, dan k adalah jumlah kandidat *cluster* $\tilde{c}_l^q$ dari ekstraksi candidat 1-*itemset* dan *candidate* 2-*itemset* yang diilustrasikan dalam Gambar 3. Variabel $g_{jl}^{max-R_j}$ menyatakan derajat tingkat kepentingan suatu *key term* t_j dalam suatu candidat *cluster* $\tilde{c}_l^q$ yang dijabarkan dalam persamaan (3).

$$g_{jl}^{max-R_j} = \frac{score\left(\tilde{c}_l^q\right)}{\sum_{i=1}^{n} W_{ij}^{max-R_j}}, dimana,\ score\left(\tilde{C}_l^q\right) = \begin{Bmatrix} \sum_{d_i \in \tilde{c}_l^1, t_j \in L_1} W_{ij}^{max-R_j}\ if\ q = 1, \\ \frac{\sum_{d_i \in \tilde{c}_l^q, t_j \in L_1} w_{ij}^{max-R_j}}{\lambda}, else \end{Bmatrix} \quad (3)$$

$$W = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_p \end{array} \\ \begin{bmatrix} w_{11}^{max-R_j} & w_{12}^{max-R_j} & \cdots & w_{1p}^{max-R_j} \\ w_{21}^{max-R_j} & w_{22}^{max-R_j} & \cdots & w_{2p}^{max-R_j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1}^{max-R_j} & w_{n2}^{max-R_j} & \cdots & w_{np}^{max-R_j} \end{bmatrix} \end{array} nxp$$

Gambar 2. Ilustrasi *Document-Term Matrix*

$$G = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_p \end{array} \begin{array}{c} \begin{array}{cccccc} \tilde{c}_1^1 & \cdots & \tilde{c}_{l-1}^1 & \tilde{c}_l^q & \cdots & \tilde{c}_k^q \end{array} \\ \begin{bmatrix} g_{11}^{max-R_j} & \cdots & g_{1l-1}^{max-R_j} & g_{1l}^{max-R_j} & \cdots & g_{1k}^{max-R_j} \\ g_{21}^{max-R_j} & \cdots & g_{2l-1}^{max-R_j} & g_{2l}^{max-R_j} & \cdots & g_{2k}^{max-R_j} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g_{p1}^{max-R_j} & \cdots & g_{pl-1}^{max-R_j} & g_{pl}^{max-R_j} & \cdots & g_{pk}^{max-R_j} \end{bmatrix} \end{array} pxk$$

Gambar 3. Ilustrasi *Term-Cluster Matrix*

$$V = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccccc} \tilde{c}_{11}^1 & \cdots & \tilde{c}_{1l-1}^2 & \tilde{c}_{1l}^q & \cdots & \tilde{c}_{1k}^q \end{array} \\ \begin{bmatrix} v_{11} & \cdots & v_{1l-1} & v_{1l} & \cdots & v_{lk} \\ v_{21} & \cdots & v_{2l-1} & v_{2l} & \cdots & v_{2k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ v_{n1} & \cdots & v_{nl-1} & v_{nl} & \cdots & v_{nk} \end{bmatrix} \end{array} nxk = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & \cdots & t_p \end{array} \\ \begin{bmatrix} \cdots & \cdots & \cdots \\ w_{21}^{max-R_j} & \cdots & w_{2p}^{max-R_j} \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots \end{bmatrix} \end{array} nxp \cdot \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_n \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{c}_1^1 & \tilde{c}_2^1 & \cdots & \tilde{c}_k^q \end{array} \\ \begin{bmatrix} \cdots & g_{12}^{max-R_j} & \cdots & \cdots \\ \cdots & g_{22}^{max-R_j} & \cdots & \cdots \\ \ddots & \vdots & \ddots & \ddots \\ \cdots & g_{p2}^{max-R_j} & \cdots & \cdots \end{bmatrix} \end{array} pxk$$

Gambar 4. Ilustrasi Document-Cluster Matrix

Pada persamaan (3), $w_{ij}^{max-R_j}$ adalah bobot (nilai fungsi keanggotaan) dari *term* t_j dalam dokumen d_i, dengan λ merupakan *minimum confidence*. Hasil dari terbentuknya matriks DTM dan matriks TCM digunakan untuk membangun matriks *Document-Cluster Matrix* (DCM). DCM memiliki ukuran matriks $n \times k$ yang merupakan turunan dari hasil perkalian antara matriks DTM dan matriks TCM. Matriks DCM secara keseluruhan dapat diilustrasikan dalam Gambar 4.

Setelah ditemukan matriks DCM, maka langkah selanjutnya adalah melakukan *tree pruning*. *Tree pruning* adalah melakukan suatu kegiatan untuk mengganti suatu *subtree* dengan suatu *leaf*. *Tree pruning* dalam Clustering dokumen bertujuan untuk menggabungkan beberapa *cluster* sejenis dan memiliki kemiripan sama yang berada di level 1 , sehingga akan menghasilkan cluster yang lebih baik. Setiap pasang cluster pada level 1 dapat

dihitung nilai kemiripannya dengan menggunakan ukuran *similiarity* yaitu inter_sim. Pasangan cluster yang memiliki nilai inter_sim tertinggi akan di gabung hingga nilai dari inter_sim dari seluruh pasangan cluster pada level 1 kurang dari nilai minimum threshold dari inter_sim. Pengukuran kemiripan antara cluster (c_x^1) dengan cluster (c_y^1) menggunakan inter_sim dapat didefinisikan dalam persamaan (4).

$$inter_{sim}(c_x^1, c_y^1) = \frac{\sum_{d_i \in c_x^1, c_y^1}^{n} v_{ix} \times v_{iy}}{\sqrt{\sum_{d_i \in c_x^1}^{n} (v_{ix})^2 \times \sum_{d_1 \in c_y^1}^{n} (v_{iy})^2}} \quad (4)$$

dimana v_{ix} dan v_{iy} adalah nilai yang diperoleh dari hasil perhitungan DCM (*Document Cluster Matrix*). Variabel x merupakan *term* pertama dan y merupakan *term* kedua. Nilai dari inter_sim memiliki rentang antara [0,1] yang didapat dari penjumlahan hasil perkalian antara v_{ix} dan v_{iy} sebanyak n dokumen dimana cluster (c_x^1) dan cluster (c_y^1) merupakan kandidat cluster dari dokumen d_i. Hasil penjumlahan tersebut akan dibagi dengan akar kuadrat dari penjumlahan kuadrat v_{ix} sebanyak n dokumen yang dikalikan dengan penjumlahan kuadrat v_{iy} sebanyak n dokumen.

III. HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan mengenai hasil uji coba serta evaluasi dari metode yang diusulkan dalam penelitian ini. Metode dalam penelitian ini diaplikasikan dengan didukung oleh hardware dan software dengan spesifikasi Processor Intel® Core™2 Duo T5750@2.00Ghz, memori 1014 MB, sistem operasi Windows 7, dan menggunakan Java Netbeans 6.9.1 dengan jdk1.6.0_18.

A. Dataset

Penelitian ini menggunakan 3 jenis dataset yang berbeda. Penjelasan mengenai dataset tersebut dijelaskan sebagai berikut :

- Classic : merupakan dataset dari abstract jurnal ilmiah yang terdiri atas kombinasi empat kelas CACM, CISI, CRANFIELD, dan MEDICAL. Jumlah data yang digunakan dalam dateset classic ini berjumlah 1000 data, dimana setiap kelas, yaitu : CACM, CISI, CRANFIELD dan MEDICAL berjumlah 250 data. CACM merupakan jurnal dengan topik akademis, CISI merupakan jurnal dengan topik informatian retreival, CRAN merupakan jurnal dengan topik sistem penerbangan, dan MED merupakan jurnal dengan topik medis.
- Reuters : merupakan dataset yang berasal dari koleksi Reuters newswire. Dalam dataset ini terdapat beberapa kelas, diantaranya reut2-001, reut2-002, reut2-003, dan reut2-004. Masing-masing kelas terdiri dari 250 data, sehingga total keseluruhan data adalah 1000 data.
- 20 Newsgroup : merupakan kumpulan dari dokumen Newsgroup yang terbagi kurang lebih 20 kelas berbeda. Dalam penelitian ini kelas yang digunakan dalam dataset 20 Newsgroup adalah 4 kelas yang terdiri dari : comp.sys.mac.hardware, rec.sport.baseball, sci.space, dan talk.politics.mideast. Masing – masing kelas terdiri dari 150 data, sehingga total terdapat 600 data.

B. Pengujian

Pengujian terhadap metode yang diusulkan dilakukan dengan tiga skenario berbeda, yaitu : pertama adalah pengujian tanpa menggunakan hipernim dalam ekstraksi kata kunci (hipernim level 0), kedua adalah pengujian dengan menggunakan hipernim level 1 atau hipernim dari *term* di level 0, dan yang ketiga adalah pengujian dengan menggunakan hipernim level 2 atau hipernim dari *term* di level 1 . Setiap skenario pengujian dilakukan untuk mengetahui pengaruh jumlah dataset terhadap nilai *overall f-measures*. Jumlah dataset yang digunakan adalah 200, 400. 600. 800, dan 1000. Hasil dari pengujian ini untuk dataset Classic terdapat pada Tabel I dan Gambar 5, untuk dataset Reuters terdapat pada II dan Gambar 6, dan untuk dataset 20 Newsgroup terdapat pada Tabel III dan Gambar 7.

Tabel I. Hasil Pengaruh Jumlah Data Terhadap *Overall F-Measure* pada Data Classic

Jumlah Data	*Overall F-Measure*		
	Hipernim Level 0	Hipernim Level 1	Hipernim Level 2
200	0.5694	0.6174	0.5809
400	0.5358	0.5759	0.5794
600	0.4992	0.5335	0.5583

800	0.5382	0.5627	0.5897
1000	0.5461	0.5510	0.5830

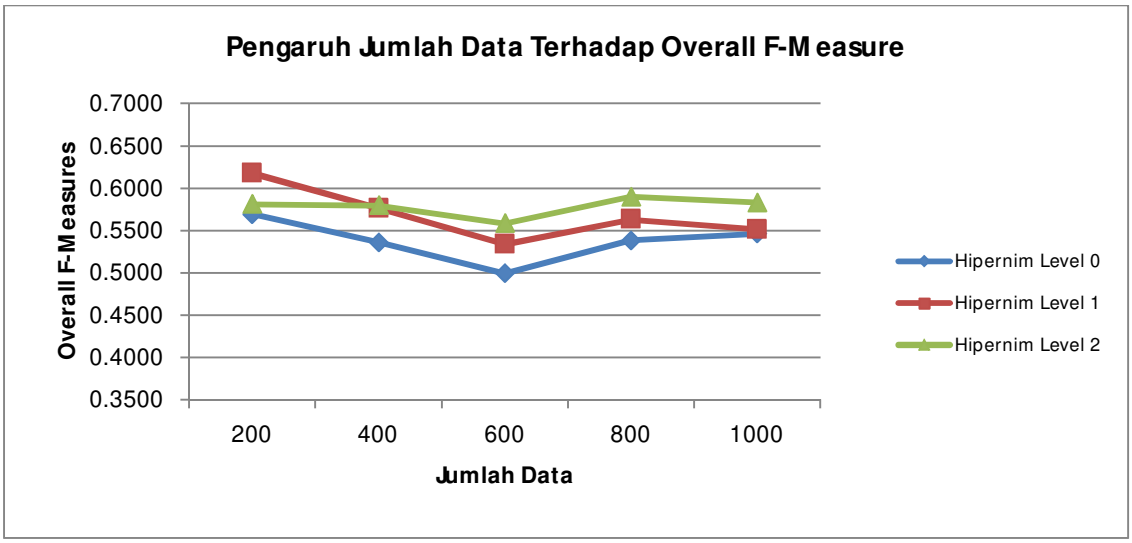
Tabel II. Hasil Pengaruh Jumlah Data Terhadap *Overall F-Measure* pada Data Reuters

Jumlah Data	***Overall F-Measure***		
	Hipernim Level 0	**Hipernim Level 1**	**Hipernim Level 2**
200	0.3780	0.3980	0.4016
400	0.3975	0.3992	0.3994
600	0.3964	0.3988	0.4000
800	0.3966	0.3988	0.3998
1000	0.3994	0.3993	0.3995

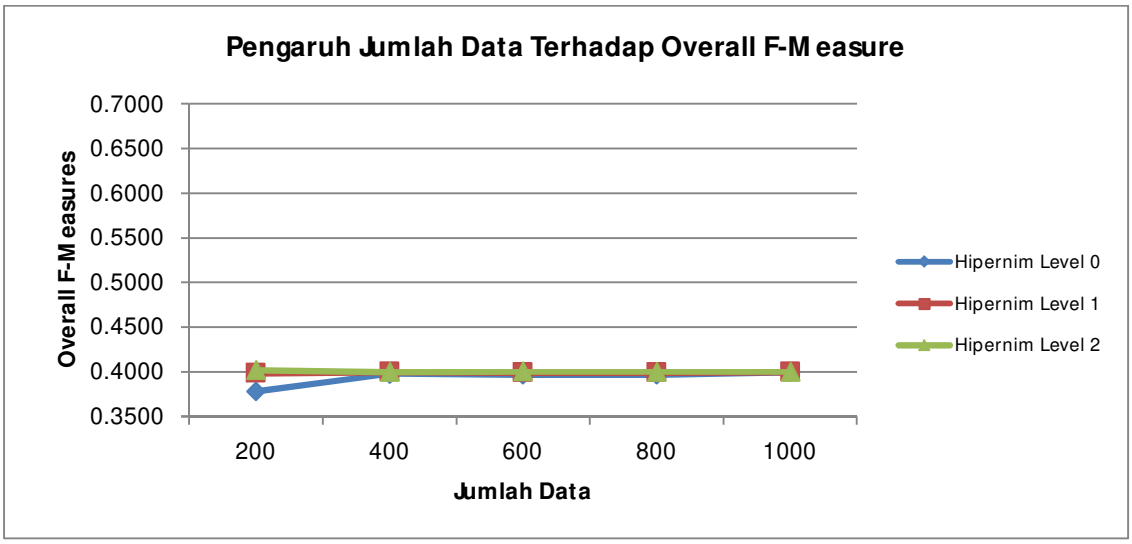
Tabel III. Hasil Pengaruh Jumlah Data Terhadap *Overall F-Measure* pada Data 20 Newsgroup

Jumlah Data	***Overall F-Measure***		
	Hipernim Level 0	**Hipernim Level 1**	**Hipernim Level 2**
200	0.5651	0.5286	0.4885
400	0.5343	0.4823	0.4322
600	0.4658	0.5565	0.4317
800	0.5387	0.5336	0.3914
1000	0.5542	0.5335	0.3955

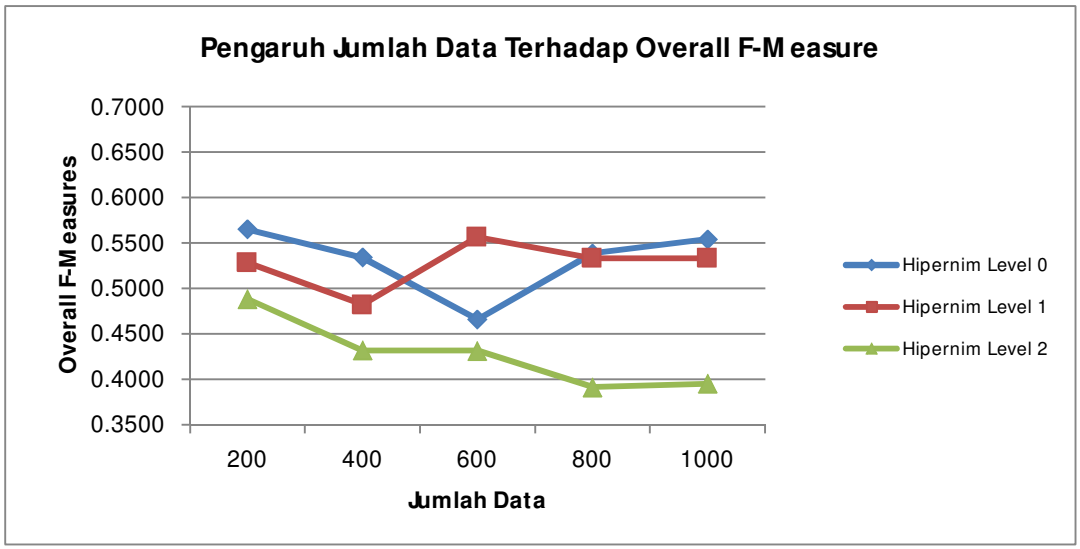
Berdasarkan Tabel I, Tabel II, dan Tabel III diketahui bahwa setiap pada dataset memiliki hasil *overall f-measures* yang berbeda. Pada dataset Classic penggunaan hipernim dengan level 1 dan level 2 pada metode yang diusulkan memiliki nilai overall f-measure yang lebih tinggi dibandingkan pada metode yang tidak menggunakan hipernim. Metode dengan menggunakan hipernim level 2 memiliki rata-rata *overall f-measure* lebih baik dibanding metode dengan hipernim level 1 dengan nilai rata-rata *overall f-measure* sebesar 0.5783. Hipernim pada *dataset* Classic mampu memperluas makna dari *term* sehingga dokumen-dokumen dengan karateristik yang sama namun tidak memiliki *term* yang sama dapat dikelompokkan menjadi satu kelompok yang sama karena memiliki *term* yang sama terhadap hipernim. Pada *dataset* Reuters memiliki hasil yang hampir sama dengan *dataset* Classic, yaitu penggunaan hipernim dengan level 1 dan level 2 pada metode yang diusulkan memiliki nilai *overall f-measures* yang lebih tinggi dibandingkan pada metode yang tidak menggunakan hipernim. Namun, penggunaan hipernim pada *dataset* Reuters memiliki nilai *overall f-measures* yang hampir sama dari ketiga jenis level hipernim, hal ini terlihat dari Gambar 6 dimana ketiga grafik dari hipernim memiliki grafik yang hampir sama. Metode dengan menggunakan hipernim level 2 memiliki rata-rata *overall f-measure* lebih baik dibanding metode dengan hipernim level 1 dengan nilai rata-rata *overall f-measure* sebesar 0.4001. Hipernim pada *dataset* Reuters memiliki pengaruh yang hampir sama seperti pada hipernim *dataset* Classic yang mampu memperluas makna dari *term* sehingga dokumen-dokumen dengan karateristik yang sama namun tidak memiliki *term* yang sama dapat dikelompokkan menjadi satu kelompok yang sama karena memiliki *term* yang sama terhadap hipernim. Sementara pada *dataset* 20 Newsgroup pengunaan hipernim level 1 dan level 2 pada metode yang diusulkan tidak memberikan nilai *overall f-measures* yang lebih tinggi dibanding metode tanpa menggunakan hipernim (level 0). Metode dengan menggunakan hipernim level 1 memiliki rata-rata *overall f-measure* lebih baik dibanding metode dengan hipernim level 2 dengan nilai rata-rata *overall f-measure* sebesar 0.5269. Rendahnya nilai *overall f-measures* pada penggunaan hipernim level 1 dan level 2 dibandingkan hipernim level 0 dapat diakibatkan *term* dari hipernim level 1 dan level 2 tidak memiliki hubungan (*semantic*) dengan *term* pada level dibawahnya yang menyebabkan *term* pada level 1 dan level 2 hanya menjadi *noise* saja.



Gambar 5. Grafik Pengaruh Jumlah Data Terhadap Overall F-Measure pada Data Classic



Gambar 6. Grafik Pengaruh Jumlah Data Terhadap Overall F-Measure pada Data Reuters



Gambar 7. Grafik Pengaruh Jumlah Data Terhadap Overall F-Measure pada Data 20 Newsgroup

Berdasarkan Gambar 5, Gambar 6, dan Gambar 7 juga dapat diketahui bahwa *dataset* Classic memiliki nilai *overall f-measure* tertinggi dibandingkan *dataset* Reuters dan 20 Newsgroup. *Dataset* Classic memiliki nilai *overall f-measure* tertinggi karena pada *dataset* Classic memiliki kumpulan dokumen yang seragam. Sementara

pada *dataset* Reuters memiliki nilai *overall f-measure*s terendah karena memiliki kumpulan dokumen yang tidak seragam. Begitu juga dengan *dataset* 20 Newsgroup, dari hasil pengujian metode dengan menggunakan hipernim (level 1 dan level 2) memiliki nilai yang lebih rendah dibandingkan metode tanpa hipernim (level 0), hal ini dapat diakibatkan karena *dataset* 20 Newsgroup memiliki kumpulan dokumen yang tidak seragam. Sehingga metode yang diusulkan tepat digunakan untuk mengelompokkan dokumen yang memiliki tingkat keseragaman yang tinggi.

IV. KESIMPULAN

Kesimpulan yang dapat diambil berdasarkan serangkaian hasil pengujian serta analisa yang telah dilakukan terhadap metode yang diusulkan. Beberapa kesimpulan yang dapat diambil sebagai berikut :

- Penggunaan hipernim mampu meningkatkan akurasi *clustering*. Namun, penggunaan hipernim dalam beberapa *dataset* hanya akan menambah jumlah *term* yang akan dianggap sebagai *noise* sehingga mengurangi akurasi.
- *Dataset* Classic merupakan data yang tepat digunakan dengan metode yang diusulkan karena memiliki nilai *overall f-measures* terbaik dibanding *dataset* Reuters dan 20 Newsgroup.
- Penggunaan hipernim pada metode yang diusulkan mampu menghasilkan rata-rata *overall f-measure* sebesar 0.5783 untuk data Classic, 0.4001 untuk data Reuters, dan 0.5269 untuk data 20 Newsgroup.

DAFTAR PUSTAKA

[1] Congnan Luo, Yanjun Li, and Soon M. Chung. (Juli 2009). Text document clustering based on neighbors. Data & Knowledge Engineering. [Online]. *68 (1)*. hal. 1271-1288. Tersedia : http://www.sciencedirect.com/science/article/pii/S0169023X09000974.

[2] Chun Lieng Chien, Frank S.C Tseng, and Tyne Liang. (September 2010). An Integration of *Wordnet* and fuzzy association rule mining for multi-label document clustering. Data & Knowledge Engineering. [Online]. *69 (1)*. hal. 1208-1226. Tersedia : http://www.sciencedirect.com/science/article/pii/S0169023X10000972.

[3] Ridvan Saracoglu, Kemal Tutuncu, and Novruz Allahverdi. (2008). A new approach on search for similiar documents with multiple categories using fuzzy clustering. Expert Systems with Applications. [Online]. hal. 2545-2554. Tersedia : http://www.sciencedirect.com/science/article/pii/S0957417407001467.

[4] Florian Beil, Martin Ester, and Xiaowei Xu. (2002) . Frequent *Term*-Based Text Clustering. Proc. of Int'l Conf. on knowledge Discovery and Data Mining. [Online]. hal. 436-442. Tersedia : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.7997&rank=1.

[5] B.C.M Fung, K. Wang, and M. Ester. (2002). Hierarchical document clustering using frequent itemset. Simon Fraser University. [Online]. Tersedia : http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.9326.

[6] Ling Chun Chen, Frank S.C Tseng, and Tyne Liang. (Oktober 2010) .Mining fuzzy frequent itemset for hierarchical document clustering. Information Processing and Management. [Online]. *46*. hal. 193-211. Tersedia : http://www.sciencedirect.com/science/article/pii/S0306457309001113.

[7] Susiana Sari. (2012) . Clustering berbasis dokumen secara hierarki barbasis fuzzy set tipe-2 trapezoidal dan triangular dari frequent itemset. Institut Teknologi Sepuluh Nopember . [Online]. Tersedia : http://digilib.its.ac.id/ITS-Article-5110512000027 4/21815.

[8] Yuen Hsien Tseng. (2010). Generic title labeling for clustered documents. Expert Systems with Applications.[Online]. hal. 2247-2254. Tersedia : http://www.sciencedirect.com/science/article/pii/S0957417409007167.

[9] Jerry M. Mendel and Robert I. Bob John. (2002). Type-2 Fuzzy Sets Made Simple. IEEE Transactions on Fuzzy System. [Online]. hal. 117-127. Tersedia : http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=995115.

[10] Janusz T. Starczewski. (Mei 2014). Centroid of triangular and Gaussian type-2 fuzzy sets. Information Sciences. [Online]. *280*. hal. 289-306. Tersedia : http://www.researchgate.net/publication/263092625_Centroid_of_triangular_and_Gaussian_type-2_fuzzy_sets.

[11] Cengiz Kahraman, Basar Oztaysi, Irem Ucal Sari, and Ebru Turanoglu. (Februari 2014). Fuzzy analytic hierarchy process with interval type-2 fuzzy sets. Knowledge-Based Systems. [Online]. *59*. hal. 48-57. Tersedia : http://www.sciencedirect.com/science/article/pii/S0950705114000410.

[12] Janusz T. Starczewski. (2009). Efficient triangular type-2 fuzzy logic systems. International Journal of Approximate Reasoning. [Online]. hal. 799-811. Tersedia : http://www.sciencedirect.com/science/article/pii/S0888613X09000565.