Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 9/2 (2016), 88-95
DOI: http://dx.doi.org/10.21609/jiki.v9i2.383

DOCUMENT CLUSTERING BY DYNAMIC HIERARCHICAL ALGORITHM BASED ON FUZZY SET TYPE-II FROM FREQUENT ITEMSET

Saiful Bahri Musa, Andi Baso Kaswar, Supria, and Susiana Sari

Department of Informatics Engineering, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Jl. Teknik Kimia, Gedung Teknik Informatika, Surabaya, 60111, Indonesia

E-mail: saiful14@mhs.if.its.ac.id

Abstract

One of ways to facilitate process of information retrieval is by performing clustering toward collection of the existing documents. The existing text documents are often unstructured. The forms are varied and their groupings are ambiguous. This cases cause difficulty on information retrieval process. Moreover, every second new documents emerge and need to be clustered. Generally, static document clustering method performs clustering of document after whole documents are collected. However, performing re-clustering toward whole documents when new document arrives causes inefficient clustering process. In this paper, we proposed a new method for document clustering with dynamic hierarchy algorithm based on fuzzy set type-II from frequent item set. To achieve the goals, there are three main phases, namely: determination of keyterm, the extraction of candidates clusters and cluster hierarchical construction. Based on the experiment, it resulted the value of F-measure 0.40 for Newsgroup, 0.62 for Classic and 0.38 for Reuters. Meanwhile, time of computation when addition of new document is lower than to the previous static method. The result shows that this method is suitable to produce solution of clustering with hierarchy in dynamical environment effectively and efficiently. This method also gives accurate clustering result.

Key Words: *Dynamic Hierarchical Algorithm, Fuzzy Set Type-II, Document Clustering.*

Abstrak

Salah satu cara untuk mempermudah proses information retieval adalah dengan melakukan pengklasteran terhadap koleksi dokumen yang ada. Dokumen teks yang ada seringkali tidak terstruktur, formatnya bervariasi, dan pengelompokannya ambigu. Hal ini menimbulkan kesulitan dalam proses information retrieval. Selain itu, setiap detik dokumen baru bartambah dan perlu untuk dikelompokkan. Pada umumnya, metode pengklasteran dokumen statis melakukan pengklasteran dokumen setelah keseluruhan dokumen terkumpul. Namun, melakukan pengklasteran ulang terhadap keseluruhan dokumen ketika dokumen baru tiba mengakibatkan proses pengklasteran menjadi tidak efisien. Penelitian ini mengusulkan metode baru untuk pengklasteran dokumen dengan algoritma hierarki dinamis berbasis fuzzy set type-II dari frequent itemset. Untuk mencapai tujuan tersebut, terdapat 3 tahapan utama yang akan dilakukan, yaitu; ekstraksi keyterm, ekstraksi kandidat klaster dan pembangunan hirarki klaster. Berdasarkan eksperimen yang telah dilakukan diperoleh nilai F-Measure 0,40 untuk Newsgroup, 0,62 untuk Classic, dan 0,38 untuk Reuters. Sedangkan waktu komputasi pada saat penambahan dokumen dapat direduksi dibanding dengan metode statis sebelumnya. Hasil percobaan terhadap beberapa dataset koleksi dokumen menunjukkan bahwa metode ini tidak hanya sesuai untuk menghasilkan solusi pengklasteran secara hirarki dalam lingkungan yang dinamis secara efektif dan efisien, tetapi juga memberikan hasil pengklasteran yang akurat.

Kata Kunci: *Algoritma Hirarki Dinamis, Fuzzy Set Tipe-II, Pengklasteran Dokumen.*

1. Introduction

Plentiful information as a result from development of information technology is a big advantage for the seeker of information. However, at the same time the big problem appears as a result of the increase of exist data where it's difficult to determine needed information from large quantity of unnecessary/unimportant data. So, information retrieval (IR) and information extraction (IE) are present to handle the problem. Managing, accessing, searching, and big browsing repository from text document need efficient organizing from the information. In that case, document clustering have important role as tool that organize document collection to be meaningful cluster collection to increase information retrieval efficiency and document management [1].

However, the increasing of text document explosively on the internet and must be clustered generally have unstructured form and their groupings are ambiguous. So they cause the difficulties in seeking and managing document. One of methods that functions to organize document collection is document clustering hierarchically. Document clustering into structure of tree hierarchically is able to increase efficiency of IR [2–4]. However, there are some challenges in hierarchy document clustering, namely; high dimensionality, scalability, accuracy, easy of browsing and meaningful cluster label [2-5].

Some researchers [2–4,6] use frequent itemset from association rule for document management. The method is able to solve the problem like reduction of dimension, input of cluster amount and the ease of seeking by meaningful label. Next, Chen et al [4] show that Fuzzy Frequent Item-set-based on Hierarchical Clustering method (F2IHC) can avoid overlapping cluster and increase the accuracy of document clustering result. However the method uses fuzzy set type-1. Fuzzy set type-1 that has function of distinct membership is not able to model uncertainty directly [7].

On the other hand, method of fuzzy set type-2 has interval membership function which is able to model uncertainty in defining membership function on fuzzy set type-1 [7-8]. Then, Sari et al [9] build a document clustering method hierarchically based on fuzzy set type-2 from frequent item-set to increase the quality of clustering result. Fuzzy set type-2 trapezoidal as upper membership function and fuzzy type-II triangular as bottom membership function toward frequent item-set that gotten from association rule mining to increase the accuracy of document clustering. The proposed method is able to produce qualified cluster and to solve the ambiguity. However the recommended method is only implemented in the static document. In environment of dynamic information like World Wide Web is necessary to apply adaptive method for organizing of document.

Static clustering methods generally do clustering where the entire document had been ready before applying clustering algorithms. When adding a new document, it is necessary to do reclustering toward the whole documents. However, performing re-clustering toward the whole documents when new documents come afterward, caused the process of clustering are not efficient.

In this paper, we proposed a new method for document clustering with Dynamic Hierarchical Algorithm based on Fuzzy Type-II from the Frequent Item-set. Type-II fuzzy sets used to overcome the problem of ambiguity and dynamic clustering method with dynamic hierarchical algorithm can process documents added or eliminated to or from the collection. Dynamic algorithm has the capability to renew clustering when data is added or

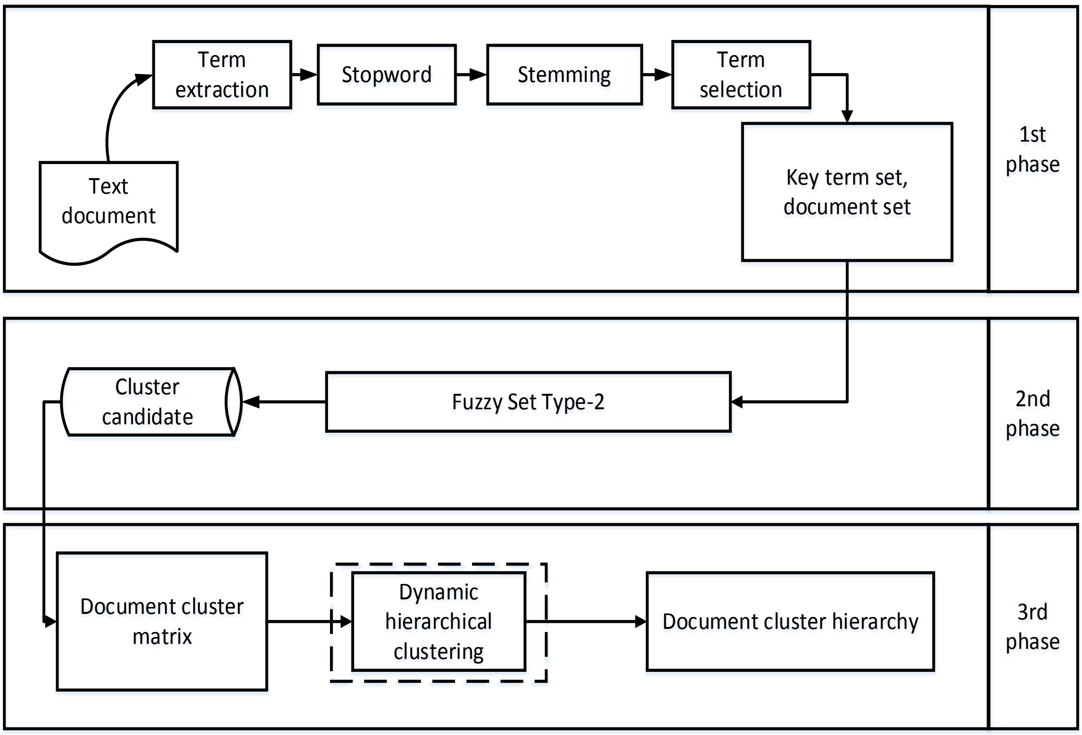

Figure 1. Document clustering phase.

TABLE 1
FUZZY MEMBERSHIP FUNCTION TYPE-II TRAPEZOIDAL AND TRIANGULAR DEFENITION.

Upper Membership Function (UMF)		*Lower Membership Function (LMF)*	
$W_{ij}^{L.U}(f_{ij}) = \begin{cases} 1, a \le f_{ij} \le b \\ \frac{(f_{ij}-c)}{(b-c)}, b \le f_{ij} \le c \\ 0,\ f_{ij} \ge c \end{cases}$	$a, d, b_1 = \min(f_{ij})$ $b = \frac{(\min(f_{ij}) + c_1)}{2}$	$W_{ij}^{L.L}(f_{ij}) = \begin{cases} \frac{(f_{ij}-a_1)}{(b_1-a_1)},\ a_1 \le f_{ij} \le b_1 \\ \frac{(f_{ij}-c_1)}{(b_1-c_1)}, b_1 \le f_{ij} \le c_1 \end{cases}$	$a_1 = 0$
$W_{ij}^{M.U}(f_{ij}) = \begin{cases} \frac{(f_{ij}-d)}{(e-d)}, d \le f_{ij} \le e \\ 1, \quad e \le f_{ij} \le f \\ \frac{(f_{ij}-g)}{(f-g)}, f \le f_{ij} \le g \end{cases}$	$c, h = avg(f_{ij})$ $e = \frac{(c_1 + avg(f_{ij}))}{2}$ $f = \frac{(avg(f_{ij}) + f_1)}{2}$	$W_{ij}^{M.L}(f_{ij}) = \begin{cases} \frac{(f_{ij}-d_1)}{(e_1-d_1)}, d_1 \le f_{ij} \le e_1 \\ \frac{(f_{ij}-f_1)}{(e_1-f_1)}, e_1 \le f_{ij} \le f_1 \end{cases}$	$c_1, d_1 = \frac{(\min(f_{ij}) + avg(f_{ij})}{2}$ $e_1 = avg(f_{ij})$
$W_{ij}^{H.U}(f_{ij}) = \begin{cases} \frac{(f_{ij}-h)}{(i-h)}, h \le f_{ij} \le i \\ 1,\ i \le f_{ij} \le j \end{cases}$	$g, j, h_1 = max(f_{ij})$ $i = \frac{(g_1 + max(f_{ij}))}{2}$	$W_{ij}^{H.L}(f_{ij}) = \begin{cases} 0, \quad f_{ij} \le g_1 \\ \frac{(f_{ij}-g_1)}{(h_1-g_1)}, g_1 \le f_{ij} \le h_1 \end{cases}$	$f_1, g_1 = \frac{(avg(f_{ij}) + max(f_{ij}}{2}$

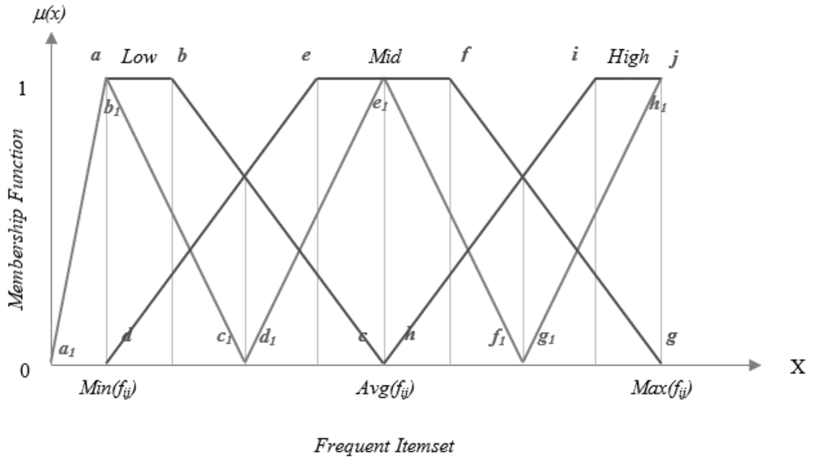

Figure. 2. Fuzzy membership function type-II trapezoidal and triangular.

eliminated from the collection. This algorithm enables us dynamically to track the large-scale of information constantly changing, either inserted to the web every day, without having to perform a complete clustering. The proposed method is expected to perform dynamical documents clustering so re-clustering toward all existing documents are not necessary to do as well as overcoming the problem of ambiguity so it can provide accurate clustering results at the same time.

2. Methods

There are several phases are done for document clustering with dynamic hierarchical clustering algorithm, namely: keyterm extraction, extracion of candidate clusters and cluster construction (Figure 1). Keyterm extraction is a process to obtain the most representative terms for the document. Extraction of cluster candidate is a process done to give value of fuzzy for frequent item-set and to get 1-itemset candidate. Cluster construction is a process to build a cluster hierarchy for frequent item-set.

Keyterm Extraction

The first phase is the process of keyterm extraction. The aim of keyterm extraction is to get the most representative term to be input for cluster candidate extraction phase. The input of this phase is the collection of document that will be clustered, while the output of this phase is in the keyterms. The procedure that must be done in this phase is the extraction of term, stop word removal, stemming, and term selection.

Term extraction is the process of term extraction from a document collection that is denoted by

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

Figure. 3. *Document Term Matrix.*

$$G = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_p \end{array} \begin{array}{c} \begin{array}{cccccc} \tilde{c}_1 & \cdots & \tilde{c}_2 & \tilde{c}_3 & \cdots & \tilde{c}_k \end{array} \\ \left[\begin{array}{cccccc} g_{11}^{max-R_j} & \cdots & g_{12}^{max-R_j} & g_{13}^{max-R_j} & \cdots & g_{1k}^{max-R_j} \\ g_{21}^{max-R_j} & \cdots & g_{22}^{max-R_j} & g_{23}^{max-R_j} & \cdots & g_{2k}^{max-R_j} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ g_{p1}^{max-R_j} & \cdots & g_{p2}^{max-R_j} & g_{p3}^{max-R_j} & \cdots & g_{pk}^{max-R_j} \end{array}\right]_{pxk} \end{array}$$

Figure. 4. *Term-Cluster Matrix.*

($\boldsymbol{D = \{d_1, d_2, d_3, \ldots, d_n\}}$) in order to obtain the set of term $\boldsymbol{T_D} = \{\boldsymbol{t_1, t_2, t_3, \ldots, t_n}\}$). The set of term $\boldsymbol{T_D}$ still contains common words (stop word) such as "and, with, what, etc.", so it`s needed the stop word-removal process to eliminated the stop word. Stop word removal is used contains 571 words in English.

Stemming process is performed then on the remaining term by stemming WordNet 2.0. This process aims to return the exist word into their basic form. To get the most representative keyterm, the term selection is based on calculations $\boldsymbol{tf.idf}$, $\boldsymbol{tf.df}$ and $\boldsymbol{tf^2}$. Term which has value more than the minimum threshold value $\boldsymbol{tf.idf}$ (α), minimum threshold $\boldsymbol{tf.df}$ (β), and minimum threshold $\boldsymbol{tf^2}$ (γ) are defended as a set of keyterms. To get the value of keyterms set, we use equation(1) for $\boldsymbol{tf.idf}$, equation(2) for $\boldsymbol{tf.df}$ and equation(3) for $\boldsymbol{tf^2}$.

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

$$tfdf_{ij} = TF * DF, \qquad (2)$$

where $TF = \frac{f_{ij}}{\sum_{j=1}^{m} f_{ij}}$, $DF = \frac{|\{d_i|\, t_j \in d_i, d_i \in D\}|}{|D|}$

$$tf_{ij}^2 = tfidf_{ij} * \ tfdf_{ij} \qquad (3)$$

Cluster Candidate Extraction

The second phase is performing the cluster candidate extraction to provide value of fuzzy for frequent item set and get 1-itemset candidate. In giving fuzzy value for frequent item-set, using membership functions of fuzzy set type-II as the upper trapezoidal membership function (UMF) and triangular membership function as lower membership function (LMF) will be optimal (Figure 2).

Term-frequency ($\boldsymbol{tf}$) fuzzy set of documents $\boldsymbol{d_i}$ denoted as ($\boldsymbol{F_{ij}, W_{ij}^{r.z}}$). $\boldsymbol{F}$ value has the range $[\mathbf{0, 1}]$. In $\boldsymbol{F_{ij}}$ consists of three regionals, namely Low ($\boldsymbol{L}$), MID ($\boldsymbol{M}$) and High ($\boldsymbol{H}$). $\boldsymbol{F_{ij}}$ denoted as $\{\mathbf{W_{ij}^{L.U}}\boldsymbol{(F_{ij})/t_j{\bullet}L.U, W_{ij}^{M.U}(F_{ij})/ t_j{\bullet}M.U, W_{ij}^{H.U}(F_{ij})/t_j \bullet H.U}\}$, $\{\boldsymbol{W_{ij}^{L.L}(F_{ij})/t_j \bullet L.L, W_{ij}^{M.L}(F_{ij})/t_j \bullet M.L, W_{ij}^{H.L}(F_{ij})/t_j \bullet H.L}\}$. $\boldsymbol{t_j \bullet r.z}$ are regional fuzzy $\boldsymbol{t_j}$. $\boldsymbol{Z}$ can serve as UMF ($\boldsymbol{U}$) or LMF ($\boldsymbol{L}$). For a term ($\boldsymbol{t_j, f_{ij}}$) in the document ($\boldsymbol{d_i}$), $\mathbf{W_{ij}^{r.z}}(\boldsymbol{f_{ij}})$ is the membership de-gree $\boldsymbol{t_j}$ in $\boldsymbol{d_i}$ which is defined in Table 1.

Where $\boldsymbol{min(f_{ij})}$ is the minimum frequency term in $\boldsymbol{D}$, $\boldsymbol{max(f_{ij})}$ is the term maximum frequency of the term in $\boldsymbol{D}$ and $\boldsymbol{avg(f_{ij}) = (min(f_{ij}) + max(f_{ij}))/2}$.

Fuzzy frequent item-set which has higher support value than minimum support value will take into account as candidate 1-itemset. The support value calculation of a term derived from the ration between the value of the fuzzy frequent itemset and the amount document. Candidate cluster ($\tilde{\mathbf{c}}$) can be obtained from the collection of documents ($\mathbf{D}$). $\tilde{\mathbf{c}}$ may also be denoted as $\tilde{\mathbf{c}}_{(\boldsymbol{t_1, t2, \ldots tq})}$ or $\tilde{\mathbf{c}}_{(\tau)}$. Candidate cluster has 2-tupples denoted as $\tilde{\mathbf{c}} = (\widetilde{\mathbf{D}}_{\mathbf{c}}, \boldsymbol{\tau})$, where $\widetilde{\mathbf{D}}_{\mathbf{c}}$ is part of $\boldsymbol{D}$ and $\boldsymbol{\tau}$ is the fuzzy frequent item-set to describe $\tilde{\mathbf{c}}$. $\boldsymbol{\tau}$ is denoted as $\boldsymbol{\tau} = \{\mathbf{t_1, t_2, \ldots, t_q}\} \subseteq \mathbf{K_D}$. $\mathbf{K_D}$ is collection of keyterms

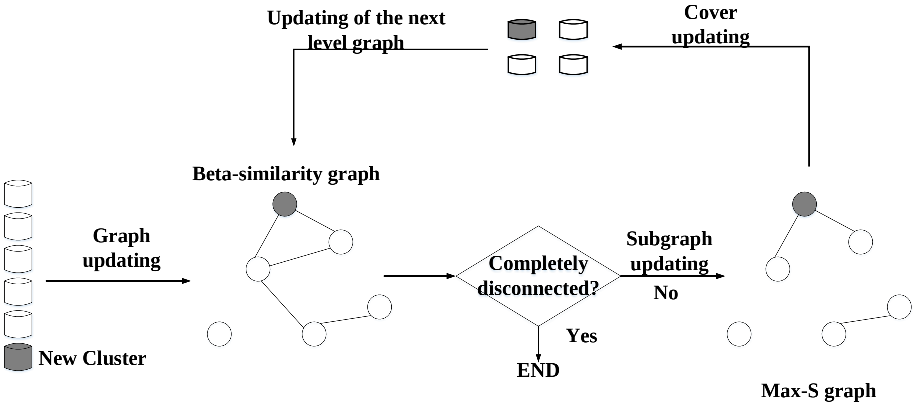

Figure. 5. Dynamic hierarchical algorithm.

D and q are the number of keyterms that are included in **τ**. Collection of cluster candidate is denoted as $\mathbf{\tilde{C}_D} = \{\,\mathbf{\tilde{c}_1}, \mathbf{\tilde{c}_2}, \mathbf{\tilde{c}_3}, \ldots, \mathbf{\tilde{c}_k}\}$, where k is the total amount of cluster candidate.

Cluster Hierarchy Construction

The third phase is the tree cluster construction. The process consists of two steps, namely: Document-Cluster Matrix (DCM) and construct a cluster hierarchy using dynamic hierarchical clustering algorithm. DCM functions to place each document to the proper cluster, so each c_i^q contains subset of the document. To achieve this goal, de-fined two matrices, that is Document-Term Matrix (DTM) and Term-Cluster Matrix (TCM). DTM (W) is the weight of term t_j in the document d_i and $t_j \in L_1$. The first step is to consider each candidate cluster $\tilde{c}_{(\tau)} = \tilde{c}_{(t_1, t2, \ldots t_q)}$ by fuzzy frequent item-set τ. τ is considered as a reference to generate target of cluster. To present the importance of document in the candidate cluster $\tilde{c}_l$, then calculated the similarity of the terms in d_i and $\tilde{c}_l$ (Figure 3).

$W_{ij}^{max-R_j}$ is the weight of *term* t_j in the document $d_i \in \tilde{c}_l$ and λ is the minimum value of the *confidence*. TCM (G) is a matrix *p x k* (Figure 4). TCM for $1 \le j \le p$, $1 \le l \le k$ calculated by using equation (4) where score calculated by using equation (5).

$$g_{jl}^{max-R_j} = \frac{score\,(\tilde{C}_l^q)}{\sum_{i=1}^{n} W_{ij}^{max-R_j}}, \qquad (4)$$

Each $g_{jl}^{max-R_j}$ in TCM presents the degree of importance of keyterms t_j in a *candidate cluster* $\tilde{c}_{(\tau)}$ by referring to all documents which have τ.

In cluster construction, the proposed framework is an agglomerative method based on Figure 5 [10] where consist of two graphs. In the framework, we use the vertex as the cluster of document.

$$\boldsymbol{score\,(\tilde{C}_l^q) = \left\{ \begin{matrix} \sum_{d_i \in \tilde{c}_l^1, t_j \in L_1} W_{ij}^{max-R_j} \; if\ q = 1, \\ \dfrac{\sum_{d_i \in \tilde{c}_l^q, t_j \in L_1} w_{ij}^{max-R_j}}{\lambda}, else \end{matrix} \right\}} \qquad (5)$$

$$\boldsymbol{\beta}_{\mathbf{Sim}}(\mathbf{c_x}, \mathbf{c_y}) = \frac{\sum_{\mathbf{d_1} \in \mathbf{c_x}, \mathbf{c_y}}^{\mathbf{n}} \mathbf{v_{ix}\ x\ v_{iy}}}{\sqrt{\sum_{\mathbf{d_1} \in \mathbf{c_x}}^{\mathbf{n}} (\mathbf{v_{ix}})^2\ \mathbf{x} \sum_{\mathbf{d_1} \in \mathbf{c_y}}^{\mathbf{n}} (\mathbf{v_{iy}})^2}} \qquad (6)$$

The first graph is an undirected graph, where the vertices are the cluster and there is an edge between node ***i*** and ***j***. Furthermore, the graph is called β-similarity graph. In this graph, an edge is formed between vertex ***i*** and ***j***, if the vertex j is β-similar to the vertex ***i***. Two clusters are ***β***-similar if the similarity of both is greater than or equal to ***β***, where β is determined parameter by the user who presents a minimum similarity threshold. ***β***-cluster similarity between two target clusters $\mathbf{c_x}$ and $\mathbf{c_y}$, $\mathbf{c_x} \neq \mathbf{c_y}$, is defined by equation(6).

The second graph is called max-***S*** graph. Max-***S*** graph relies on the maximum β-similarity relationship and it is a sub-graph of the first one. Vertices of the graph is the same as vertices in the graph ***β***-similarity. Vertices ***i*** and ***j*** given the edge, if cluster ***i*** is the most ***β***-similiar to cluster ***j***. The use of Max-S graph not only reduce time and room utilize (because it has a little edges) but also produce dense cluster.

Being given a cluster hierarchy that previously was created using the algorithm. If there is a new document addition of the cluster, the cluster at all levels of the hierarchy should be revised. When a new document arrives, singleton will be created and ***β*** -similarity graph at the bottom level is updated.

Then update the max-***S*** graph, where this process can produce or remove a vertex and can also produce a new edge and remove the others. Let ***N*** be the set of cluster to add to max-S graph. Add all

TABLE 2
F-MEASURE RESULT

Addition process	Classic		Reuters		NewsGroup	
	Proposed	Static method	Proposed	Static method	Proposed	Static method
1	0.62	0.62	0.39	0.39	0.39	0.39
2	0.62	0.62	0.39	0.39	0.40	0.39
3	0.62	0.62	0.38	0.38	0.40	0.39

TABLE 3
TIME EXECUTION

Addition process	Classic		Reuters		NewsGroup	
	Proposed (s)	Static method (s)	Proposed (s)	Static method (s)	Proposed (s)	Static method (s)
1	98	149	1229	1316	979	940
2	36	231	403	1608	357	1303
3	54	252	573	2207	615	1877

vertices of $\boldsymbol{N}$ to max-$\boldsymbol{S}$ graph. Find the most $\boldsymbol{\beta}$ –similar vertices of each vertex in $\boldsymbol{N}$ and add the corresponding edges to max-S graph. Find all vertices for which a vertex in $\boldsymbol{N}$ is its most $\boldsymbol{\beta}$-similar and update the corresponding edges. The value of $\boldsymbol{\beta}$-similarity the same for all levels of hierarchy levels.

A cover routine applied to the max-S graph to renew cluster. Let $\boldsymbol{N}$ be the set of vertices added to the max-$\boldsymbol{S}$ graph. Let, also, $\boldsymbol{NE}$ be the set of edges added to the max-S graph. Let $\boldsymbol{Q}$ be a queue with the vertices to be processed, $\boldsymbol{Q} \neq \emptyset$. Put the remaining vertices of the clusters into $\boldsymbol{Q}$. Remove these clusters from the list of the existing clusters. Put all vertices of $\boldsymbol{N}$ into the queue $\boldsymbol{Q}$. Build the connected components from the vertices in $\boldsymbol{Q}$ and add them to the list of existing clusters. For each edge of $\boldsymbol{NE}$, merge the clusters to which its vertices belong. When a cluster is created on the hierarchy level, $\boldsymbol{\beta}$-similarity graph next level should be updated. This process is repeated until the graph completely disconnected.

3. Results and Analysis

The implementation of the method is supported by Intel ® Core ™ i3-2120 CPU @ 3.30GHz (4CPUs) processor, with a RAM of 4 GB and Java Development Kit 6 update 31 with Netbeans IDE 8.0.1. Regarding the evaluation of the method, we used 1000 documents from Classic, 1930 documents from Reuters and 1000 document from Newsgroup. Therefore the evaluation of the proposed method is based on the scalability and F-Measure of the hierarchical cluster.

Overall F-Measure will be measured using equation(7).

$$F(C) = \sum_{l_j \in L} \frac{|l_j|}{|D|} \max_{c_{i\in C}}\{F\}, \tag{7}$$

where $|D|$ is the amount of all document in dataset D. C is a cluster obtained from system. L is the class label that obtained from dataset. $|ci|$ is the amount of document in cluster C. $|lj|$ is the amount of document in class L. F is the F-measure, P is the precision and R is recall obtained from equation(8) to equation(10) respectively.

$$F = \frac{2PR}{P+R} \tag{8}$$

$$P = \frac{|c_i \cap l_j|}{|c_i|} \tag{9}$$

$$R = \frac{|c_i \cap l_j|}{|l_j|} \tag{10}$$

In the keyterm extraction process, the amount of keyterms depends on the minimum threshold. The keyterms that has *tf. idf, tf. df, and* tf^2 more than threshold value will be consider as the keyterm. The obtained keyterms are supposed to be the representative keyterm toward document collection.

In the candidate extraction phase of the clusters using fuzzy set type-II and with $minsup$ higher than 20%, we obtained 4 clusters from the dataset Classic, 9 clusters from Reuters and 15 from Newsgroup. If the $minsup$ used is too low, it will obtained many cluster candidate. Therefore, it is possible to results a low clustering accuracy. On the other hand, if we use high $minsup$ value, it is possible that it will obtained a litle cluster candidate which doesn`t represent the whole documents within the document collection.

In the cluster construction phase, the graph that was formed was a vertex that was obtained using type-II fuzzy sets. First of all, all documents that were placed in the various cluster, uses the DCM method. Thus, from the above process, each of the cluster candidates have got their own members. Based on the result of clustering and construction of the hierarchy structure, we obtained the values of F-Measure as mention in Table 2. The experiment was done by doing the process of documents addition periodically. Based on the experiments that were done, we obtained value F-Measure as much as 0.40 from Newsgroup, 0.62 from classic and 0.38 from Reuters after clustering the whole documents. The F-Measure value decreased in each additional document process because the more documents, the more possibility of cluster formed. Which means, each document will occupy a cluster that is not supposed.

To show the dynamic and efficiency of the proposed method, documents clustering was intially carried out. Next, with the assumption that there are new documents, these new documents are clustered immediatelly. From Table 3, we can see that in every clustering process, the proposed method can reduce document clustering processing time consideredly when compared with the static method which means, the proposed method can improve the efficiency of the document clustering method in term of time execution if there are new documents.

The proposed method can performs clustering process faster than the previous method because the proposed method only apply the whole processes toward the document comes afterward then find the similarity value toward the existing clus-ters (vertex). Therefore, the document that has just been added can join in cluster that has been formed before or form new hierarchy. While the previous method performs clustering toward the whole documents either the document comes after-ward or the document that has been clustered before.

Beside that we can also know, the proposed method can give value F-measure as good as the method before. This can happen because of the obtained result from cluster candidate extraction using fuzzy set type-II that overcome ambiguity problem and each exist document has been joined in its own cluster. Therefore, we can know that the proposed method is effective and more efficient in clustering the document.

For addition, proposed method can improve the efficiency in terms of RAM usage. In every clustering process/when the the addition of new documents, RAM usage increases only in small amounts. Whereas conventional method require larger RAM allocation. For example, in the first iteration of the clustering process for 500 documents Classic, it is requires allocation of 18.98 MB RAM by using the proposed method and 19.09 MB by using conventional method. Then we added 200 new documents in the second iteration. RAM usage using the proposed method only increased by 4 MB. While conventional method increased about 20.19 MB.

4. Conclusion

In this paper we proposed a new document clustering method by dynamic hierarchical algorithm based on fuzzy set type II from frequent item-set. Dynamic hierarchical algorithm used to perform dynamic document clustering and fuzzy set type II used to solve ambiguity problems when clustering. From the obtained results, the proposed method can improve efficiency in terms of time of clustering and RAM usage because the proposed method only apply the whole processes toward the document comes afterward then find the similarity value toward the existing clusters (vertex). Besides that, it can be seen that the proposed method gives good accuracy of the clustering.

Reference

[1] T. M. Nogueira, H. A. Camargo, S. O. Rezende, R. W. Luís, and S. Carlos-sp, “Fuzzy Rules for Document Classification to Improve Information Retrieval,” *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 3, pp. 210–217, 2011.

[2] F. Beil, M. Ester, B. Bc, and C. Va, “Frequent Term-Based Text Clustering,” 2002.

[3] B. C. M. Fung, “Hierarchical Document Clustering using Frequent Itemsets,” 2002.

[4] C.-L. Chen, F. S. C. Tseng, and T. Liang, “Mining Fuzzy Frequent Itemsets for Hierarchical Document Clustering,” *Inf. Process. Manag.*, vol. 46, no. 2, pp. 193–211, 2010.

[5] J. Han and M. Kamber, *Data Mining Concept and Techniques*. 2006.

[6] T. Hong, K. Lin, and S. Wang, “Fuzzy Data Mining for Interesting Generalized Association Rules,” *Fuzzy Sets Syst.*, vol. 138, pp. 255–269, 2003.

[7] J. M. Mendel and R. I. B. John, “Type-2 Fuzzy Sets Made Simple,” *Fuzzy Syst. IEEE Trans.*, vol. 10, no. 2, pp. 117–127, 2002.

[8] J. T. Starczewski, “Efficient Triangular Type-2 Fuzzy Logic Systems,” *Int. J. Approx. Reason.*, vol. 50, no. 5, pp. 799–811, 2009.

[9] S. Sari and A. Z. Arifin, “Clustering Dokumen Secara Hierarki Berbasis Fuzzy Set Tipe-2 Trapezoidal dan Triangular dari Frequent Itemset,” in *Prosiding Seminar Nasio-*

nal Manajmemen Teknologi XVI, 2012, pp. 1–8.

[10] R. Gil-garcía and A. Pons-porrata, “Dynamic Hierarchical Algorithms for Document Clustering,” *Pattern Recognit. Lett.*, vol. 31, no. 6, pp. 469–477, 2010.