A Hierarchical Document Clustering Approach with Frequent Itemsets

Cheng-Jhe Lee, Chiun-Chieh Hsu, and Da-Ren Chen

***Abstract*—In order to effectively retrieve required information from the large amount of information collected from the Internet, document clustering in text mining becomes a popular research topic. Clustering is the unsupervised classification of data items into groups without the need of training data. Many conventional document clustering methods perform inefficiently for large document of collected information and require special handling for high dimensionality and high volume. We propose the OCFI (Ontology and Closed Frequent Itemset-based Hierarchical Clustering) method, which is a hierarchical clustering method developed for document clustering. OCFI uses common words to cluster documents and builds hierarchical topic tree. In addition, OCFI utilizes ontology to solve the semantic problem and mine the meaning behind the words in documents. Furthermore, we use the closed frequent itemsets instead of only use frequent itemsets, which increases efficiency and scalability. The experimental results reveal that our method is more effective than the well-known document clustering algorithms. The clustering results can be used in the personalized search service to assist users to obtain the information they need.**

***Index Terms*—OCFI, documents clustering, ontology, closed frequent itemsets.**

I. Introduction

Due to the popularity of the Internet, a large number of documents, reports, e-mails, and web pages cause the information overload problem. Many enterprises spend lots of manpower on organizing these unstructured data into a logical structure for later use. In order to save the manpower and find interesting knowledge effectively, text mining becomes more and more important. Text mining also refers to as text data mining [1]–[4], roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. People usually use methods in data mining to find interesting information from text, such as the well-known data mining methods association rule mining [4]-[6], classification [7]-[8], clustering [9]-[10], etc.

The main difference between data mining and text mining is the type of input data. Data mining focuses on the logical and structured data, such as transactions in the large database; and text mining usually focuses on the unstructured documents. In order to use data mining methods to extract information from text, text mining usually involves the process of parsing the input text, mining valuable information through the analysis models, and finally evaluating and interpreting the output. However, most of the text mining algorithms do not have enough capability to solve the problems from text mining, because many methods are modifications of traditional data mining algorithms that were originally designed for relational database. Therefore, traditional clustering algorithms become impractical in real-world document clustering which requires special handling for high dimensionality, high volume, and ease of browsing.

Fung proposed Frequent Itemset-based Hierarchical Clustering (FIHC) [11] method to solve the problems from traditional algorithms. FIHC is a hierarchical clustering method developed for document clustering. Clustering or cluster analysis is the task of assigning a set of objects into groups, called clusters, so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main topic of data mining algorithm, and a common technique for statistical data analysis used in many fields, including machine learning, image analysis, information retrieval, and bioinformatics.

A major breakthrough of FIHC is that the clustering algorithm utilizes an important notion, frequent itemset, in association mining to cluster documents. The intuition of FIHC is that there exist some common words for each cluster, and FIHC use such words to cluster documents and build hierarchical topic tree. However, FIHC algorithm only used the keyword in document set to cluster documents, but ignored the semantic problem and the meaning behind the words. Therefore, the accuracy of FIHC in the Chinese document set is reduced.

In order to improve the accuracy in clustering Chinese documents, in this paper, we proposed an improved method named OCFI (Ontology and Closed Frequent Itemset-based Hierarchical Clustering). OCFI uses common words to cluster documents and builds hierarchical topic tree. In addition, OCFI utilizes ontology to solve the semantic problem and mine the meaning behind the words in documents. Furthermore, OCFI uses the closed frequent itemsets instead of only using frequent itemsets, which can increase efficiency and scalability. Because OCFI makes use of association rule mining to cluster documents, we will briefly explain association rule mining in the following.

In data mining, association rule mining is a popular and well known method for discovering interesting relations, associations, frequent patterns between objects or items in large databases. Market basket analysis is a typical example of association rule mining on transaction database. It finds out customer buying habits by extracting associations among

Manuscript was received October 10, 2015; revised March 1, 2016. This work was supported in part by the Ministry of Science and Technology of the Republic of China under grant MOST 104-2221-E-011-162.

Cheng-Jhe Lee and Chiun-Chieh Hsu are with the Department of Information Management, National Taiwan University of Science and Technology, Taipei 106 Taiwan, R.O.C. (e-mail: cchsu@mail.ntust.edu.tw).

Da-Ren Chen is with the Department of Information Management, National Taichung University of Science and Technology, Taichung, Taiwan, R.O.C. (e-mail: danny@nutc.edu.tw).

DOI: 10.7763/IJET.2017.V9.965

the different items that are purchased together. For example, we could say the set [onion, potatoes, burger] is frequent if the number of occurrences of this set is larger than a threshold, called minimum support; and the rule [onion, potatoes] ➔ [burger] found in the transaction database would show that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities. Many algorithms were proposed for computing frequent itemsets, and the most well-known methods are the Apriori and the FP-growth algorithms.

Because OCFI is a hierarchical clustering method that combines closed frequent itemset and ontology to cluster documents [12]-[14]. We will make a brief introduction to ontology. Ontology [15]-[17] is the structural frameworks for organizing information and is widely used in information analysis field, such as artificial intelligence (AI), Semantic analysis. In computer and information science, ontology structure formally represents a set of concepts and knowledge within a domain, where we could obtain additional information by the relationships between those concepts. Because the construction of ontology do not have a recognized standard, the creation of domain ontology is fundamental to the definition and use of an enterprise architecture framework.

This paper is divided into four sections: The proposed method OCFI is presented in section II. OCFI could be further divided into five steps: document preprocessing, translation by ontology, using association rule mine frequent terms, document clustering, and showing the clustering results by building cluster tree. In section III, we compared our algorithm with the well-known clustering methods by using experimental results and analysis. In section IV, we make a brief conclusion for this paper.

II. The Proposed Approach

The purpose and contribution of this paper is to construct an automated document clustering system. Most document clustering algorithms employ several document preprocessing steps to organize the unstructured data into a logical structure for later use. Those steps include parsing and removing stop words. Stop words are the most common words in a language, but they do not convey any significant information so they are stripped from the document set.

After fragmenting documents into independent terms and removing unused words, our method uses ontology to solve the semantic problem and mine the meaning behind the words in documents. Fig. 1 is the ontology structure used in this paper. To solve the semantic problem between words in document, we compare each word with the node in the ontology. The comparison steps of synonymous could be divided into several parts, which is shown in the following flow diagram of Fig. 2.

After capturing a word *T*, OCFI compares *T* with each node in the ontology. If *T* matches any one node in the ontology, then we keep this word and do the next step, mining the meaning behind the word *T*. On the other hand, if *T* is different from all nodes in the ontology, OCFI does a further check by comparing *T* with the synonymous field of each node in the ontology. Because the synonymous fields of each node may be more than one, word *T* must compare with the synonyms one by one. If *T* matches any one node in ontology, we replace *T* with the name of the node, and mine the meaning behind the word *T*.

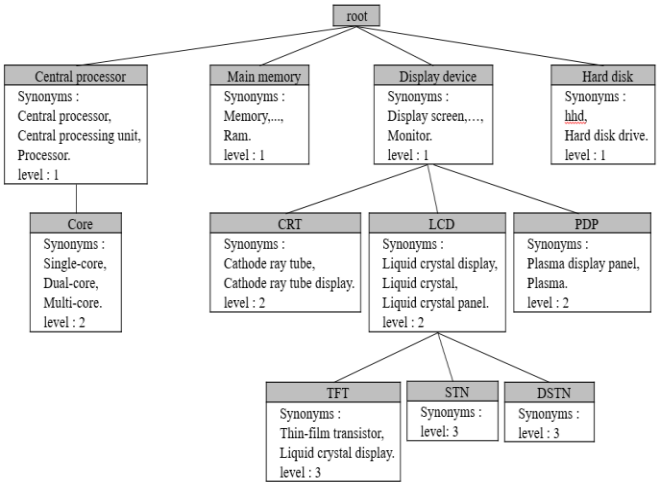

Fig. 1. The PCDIY ontology.

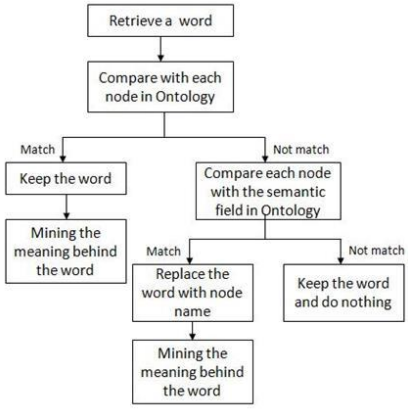

Fig. 2. Comparison steps of synonymous.

If word *T* matches any one node in the ontology, OCFI starts the other procedure to mine the meaning behind *T*. If the parent node *P* in the ontology of *T* is not the root, OCFI adds the parent node *P* into document, and checks the parent of *P* recursively until the parent of *P* is the root. Fig. 3 shows the process step by step.

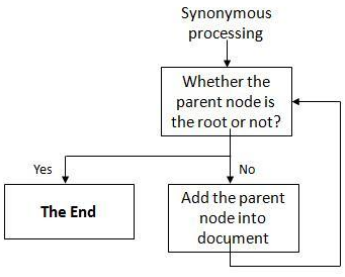

Fig. 3. Mining the meaning behind words.

Before document clustering, we first introduce some definitions. A global frequent "itemset" (common words) refers to a set of items (words) that appear together in more than a user-specified threshold of the document set; and a global frequent "item" refers to an item that belongs to some global frequent itemset.

After solving the semantic problem and mining the meaning behind words, the document set is organized into logical structure for association rule mining. The main idea of

our clustering algorithm is based on an observation: the documents under the same topic should share a set of common words. In association rule mining, the common words are regarded as frequent itemsets. OCFI builds FP-tree structure [18]-[20] and uses FP-growth method to find out all common words (closed frequent itemset) from document set. Fig.4 is a document set that contains twelve documents and there topics, and Fig. 5 shows the result of finding out common words.

No	Document ID	Monitor	Graphics card	LCD	Processor	Memory	Intel
1	C001	0	1	0	0	0	0
2	G001	1	1	1	0	0	0
3	G002	2	0	1	0	0	0
4	G003	2	1	2	0	3	0
5	G004	2	0	3	0	0	0
6	G005	1	0	2	0	0	0
7	Y001	0	0	0	8	1	2
8	Y002	0	1	0	4	3	1
9	Y003	0	0	0	3	0	2
10	Y004	0	0	0	6	3	3
11	Y005	0	1	0	4	0	0
12	Y006	0	0	0	9	1	1

Fig. 4. Document sets.

Global closed frequent itemset	support
{Graphics card}	0.42
{ Processor}	0.50
{ Memory}	0.42
{ Display device, LCD}	0.42
{ Processor, Intel}	0.42

Fig. 5. The result of finding out common words.

A. Constructing Initial Clusters

The initial clusters are constructed by each global closed frequent itemset (common words). All the documents containing this itemset (words) are included in the same cluster. Since a document usually contains more than one global closed frequent itemset (common word), the same document may appear in more than one initial cluster; in other words, initial clusters may be overlapped. The purpose of initial clusters is to ensure the property that all the documents in a cluster contain all the items in the global frequent itemsets that define the cluster. We use this global closed frequent itemsets (common words) as the cluster label to identify the cluster. The cluster label has two other functions. First, we use those cluster labels to build a hierarchical structure, called cluster tree, where cluster tree is the finally result of document clustering. Second, the meaningful cluster labels make user browse easier. We remove the overlapping of clusters in subsection B. Fig. 6 shows the result of the initial clustering.

Cluster	document
C(Graphics card)	C001, G001, G003, G004, G005
C(Processor)	Y001, Y002 ,Y003 ,Y004 , Y005, Y006
C(Memory)	G003, Y001, Y002, Y004, Y006
C(Display device, LCD)	G001, G002, G003, G004, G005
C(Processor, Intel)	Y001, Y002, Y003, Y004, Y006

Fig. 6. Initial clustering.

B. Removing the Overlapping Problem

In this step, we assign a document to the best-fit cluster so that each document belongs to exact one cluster. The measuring function used in OCFI is the same as used in FIHC, where Equation (1) shows the function in detail. Equation (1) measures the fitness of an initial cluster C_i for a document D_i. To make clusters non-overlapping, we assign each D_i to the initial cluster C_i of the highest Score ($C_i \leftarrow D_i$). After this assignment, each document belongs to only one cluster.

$$\text{Score}(C_i \leftarrow D_j) = \sum n(x) \times \text{cluster\_support}(x) - \sum_{x'}^{x} n(x') \times \text{global\_support}(x') \quad (1)$$

After using the measuring function to find the best-fit cluster for each document, the overlapping problem is removed. We could understand the topic of a document by the cluster label. Fig. 7 shows the final result of document clustering. We will build a hierarchical structure named cluster tree in the next subsection.

Cluster	Documents	Cluster frequent items
C (Graphics card)	C001	{Graphics card , LS=100%}
C(Processor)	Y005	{Processor , LS=100%}
C(Memory)	null	null
C(Display device, LCD)	G001, G002, G003, G004, G005	{Display device , LS=100%}, { LCD , LS=100%}
C(Processor , Intel)	Y001, Y002, Y003, Y004, Y006	{Processor , LS=100%}, { Intel , LS=100%}

Fig. 7. The Final clustering result.

C. Building Cluster Tree

OCFI builds the cluster tree bottom-up. The method finds out the best-fit parent cluster *P* for each cluster by the cluster label. We use Equation (1) to measure and choose the best parent cluster for each cluster. The cluster tree for the document set in Fig. 4 is shown in Fig. 8.

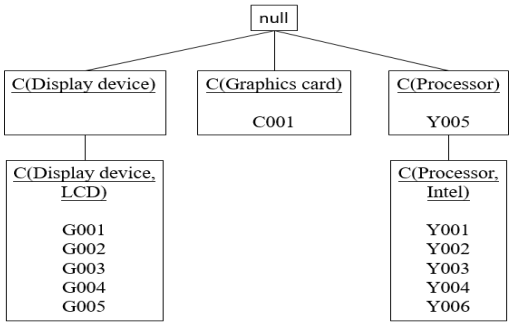

Fig. 8. Cluster tree.

III. Experimental Evaluation

This section shows the experimental evaluation of our method (OCFI) and compares its results with those of several popular document clustering algorithms, such as FIHC, k-means, and Bisecting k-means. The document sets are collected by manpower. Table I shows detail of the document sets.

TABLE I: Document Sets

Topic ID	Topic	Number of Document	Topic ID	Topic	Number of Document
G	LCD	83	O	Memory	89
L	Motherboard	76	Y	Processor	90
M	Hard Disk	89			

We use F-measure to measure the accuracy of our method. F-measure produces a balanced measure of precision and recall rates, and is also a standard evaluation method for both flat and hierarchical clustering structures. The accuracies of our method (OCFI) are shown in Table II and Table III, and

the comparison of our results with those of several popular clustering algorithms is shown in Table IV.

The experimental results reveal that the proposed approach has more accurate clustering results than those of other well-known methods.

TABLE II: THE RESULT OF DOCUMENT SET 1

ID	Ki	Ci	Nij	Nij/Ki	Nij/Ci
G	83	83	83	1	1
L	76	103	71	0.934211	0.68932
M	89	84	81	0.910112	0.964286
O	89	75	66	0.741573	0.88
Y	90	82	80	0.888889	0.97561
	427	427	381	0.894957	0.901843

ID	F-measure		
G	1	0.194379	0.194379
L	0.793296089	0.177986	0.141196
M	0.936416185	0.208431	0.195178
O	0.804878049	0.208431	0.167761
Y	0.930232558	0.210773	0.196068
Sum / Average	1		0.89458

TABLE III: THE RESULT OF DOCUMENT SET 2

ID	Ki	Ci	Nij	Nij/Ki	Nij/Ci
A	199	209	198	0.994975	0.947368
B	197	196	187	0.949239	0.954082
C	200	246	192	0.96	0.780488
G, Z	392	359	347	0.885204	0.966574
L	176	259	162	0.920455	0.625483
M	199	165	158	0.79397	0.957576
O	199	127	116	0.582915	0.913386
Y	90	88	74	0.822222	0.840909
	1652	1649	1434	0.863622	0.873233

ID	F-measure		
A	0.970588235	0.12046	0.116917
B	0.951653944	0.119249	0.113484
C	0.860986547	0.121065	0.104236
G, Z	0.924101198	0.237288	0.219278
L	0.744827586	0.16538	0.079352
M	0.868131868	0.12046	0.104575
O	0.711656442	0.12046	0.085726
Y	0.831460674	0.054479	0.045297
Sum / Average	1		0.86887

TABLE IV: COMPARING WITH OTHER METHODS

	K-means	Bisecting K-means	FIHC	OCFI
Dataset(I)	59.78%	64.22%	73.66%	89.45%
Dataset(II)	54.10%	62.18%	73.16%	86.88%
	56.94%	63.2%	73.41%	88.17%

IV. CONCLUSION

Most traditional clustering methods do not satisfy the special requirements for document clustering, such as high dimensionality, high volume, and ease of browsing with meaningful cluster labels. This paper proposes a method combining ontology and closed frequent itemsets for document clustering.

The experimental evaluation shows that the proposed method has more accurate clustering results than those of other well-known clustering algorithms, including FIHC algorithm, on various types of document sets, even when the number of clusters is unknown. We use the low-dimensional feature vector, which is composed of global closed frequent items, in place of the original high-dimensional document vector. Because the proposed method utilizes closed frequent items instead of frequent items, and low-dimensional feature vector instead of vector space model, our algorithm is more efficient and scalable than others. In addition, since the clustering result of OCFI is a cluster tree where the nodes can be treated as topics and subtopics, user can easily navigate different topics in the document set through the tree. Each topic has a label which concisely summarizes the members in the cluster and the method does not require additional processing to generate these cluster labels. Moreover, the clustering results can be used in the personalized search service to assist users to obtain the information they need.

REFERENCES

[1] C. Aggarwal and C. Zhai, *Mining Text Data*, 1st ed., Spinger, 2012.

[2] V. Bijalwanl, V. Kumar, P. Kumari, and J. Pascual, "KNN based machine learning approach for text and document mining," *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 61-70, 2014.

[3] S. Chakrabarti, "Data mining for hypertext: A tutorial survey," *ACM SIGKDD Explorations Newsletter*, vol. 1, pp. 1–11, 2000.

[4] J. Han and M. Kimber, *Data Mining: Concepts and Techniques*, Morgan-Kaufmann, August 2000.

[5] J. Hipp, U. Guntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining - a general survey and comparison," *ACM SIGKDD Explorations Newsletter*, vol. 2, pp. 58–64, 2000.

[6] B. Kamsu-Foguem, F. Rigal, and F. Mauget, "Mining association rules for the quality improvement of the production process," *Expert Systems with Applications*, vol. 40, pp.1034–1045, 2013.

[7] T. Botsis, M. Nguyen, E. Woo, M. Markatou, and R. Ball, "Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection," *Journal of the American Medical Informatics Association*, vol. 18, pp. 631-638, 2011.

[8] K. Wang, S. Zhou, and Y. He, "Hierarchical classification of real life documents," in *Proc. International Conference on Data Mining*, pp. 1-16, 2001.

[9] S. Krishna and S. Bhavani, "An efficient approach for text clustering based on frequent itemsets," *European Journal of Scientific Research*, vol. 42 no. 3, pp. 385-396, 2010.

[10] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD Workshop on Text Mining*, pp. 1-20, 2000.

[11] B. C. M. Fung, K. Wang, and M. Ester, "Hierarchical document clustering using frequent itemsets," in *Proc. International Conference on Data Mining*, pp. 59-70, 2003.

[12] D. Fensel and M. A. Musen, "The semantic web: A brain for humankind," *IEEE Intelligent Systems*, vol. 16, no. 2, pp. 24-25, March/April 2001.

[13] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, pp. 60-76, 2009.

[14] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proc. ACM-SIGMOD International Conference on Management of Data*, pp. 1–12, 2000.

[15] J. Ma, W. Xu, Y. Sun, E. Turban, S. Wang, and O. Liu, "An ontology-based text-mining method to cluster proposals for research project selection," *IEEE Trans. on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 42, no. 3, 2012.

[16] A. Maedche and S. Stabb, "Ontology learning for the semantic web," *IEEE Intelligent Systems*, vol.16, no.2, pp. 72-79, March 2001.

[17] S. Staab and A. Maedche, "Knowledge portals ontologies at work," *AI Magazine*, vol. 22, no. 2, pp. 63-75, 2001.

[18] T. Hu, S. Y. Sung, H. Xiong, and Q. Fu, "Discovery of maximum length frequent itemsets," *Information Sciences*, vol. 178, pp. 69–87, 2008.

[19] C. Lin, T. Hong, and W. Lu, "An effective tree structure for mining high utility itemsets," *Expert Systems with Applications*, vol. 38, pp. 7419–7424, 2011.

[20] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, *et al.* "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol.14, pp.1–37, 2008.

Cheng-Jhe Lee is currently a master's student in the Department of Information Management at National Taiwan University of Science and Technology. His current research interests include web technology, data mining, and graph mining.

Chiun-Chieh Hsu is currently a full professor in the Department of Information Management at National Taiwan University of Science and Technology. He received his B.E., M.E., and Ph.D. degrees all in electrical engineering from National Taiwan University in 1983, 1987, and 1990, respectively. He worked as a software and firmware design engineer in Acer Computer Company from 1983 to 1985. His current research interests include web technology and applications, data mining, information retrieval, parallel and distributed processing, and graph theory

Da-Ren Chen received his a B.E., a M.E. and a Ph.D. degrees in the Department of Information Management from National Taiwan University of Science and Technology, Taipei, Taiwan, in 1999, 2001 and 2006, respectively. Presently, he is an associate professor in the Department of Information Management at National Taichung University of Science and Technology. His research interests include real-time systems, cyber-physical systems, cloud computing, parallel and distributed computing and wireless sensor networks.