Journal of Southeast University (English Edition) Vol. 23, No. 3, pp. 439 – 442 Sept. 2007 ISSN 1003—7985

Fuzzy *c*-means text clustering based on topic concept sub-space

Ji Xianghua[1] Chen Chao[2] Shao Zhengrong[2] Yu Nenghai[1]

([1]MOE-MS Key Laboratory of Multimedia Computing and Communication,
University of Science and Technology of China, Hefei 230027, China)

([2]Library, University of Science and Technology of China, Hefei 230027, China)

Abstract: To improve the accuracy of text clustering, fuzzy *c*-means clustering based on topic concept sub-space (TCS2FCM) is introduced for classifying texts. Five evaluation functions are combined to extract key phrases. Concept phrases, as well as the descriptions of final clusters, are presented using WordNet® origin from key phrases. Initial centers and membership matrix are the most important factors affecting clustering performance. Orthogonal concept topic sub-spaces are built with the topic concept phrases representing topics of the texts and the initialization of centers and the membership matrix depend on the concept vectors in sub-spaces. The results show that, different from random initialization of traditional fuzzy *c*-means clustering, the initialization related to text content contributions can improve clustering precision.

Key words: TCS2FCM; topic concept space; fuzzy *c*-means clustering; text clustering

Clustering[1–8] analysis is a method of exploratory identification, which is automatically classified based on the similarity between patterns. The objective is to make the similarity between patterns inside the cluster as much as possible, and the similarity between patterns outside the cluster as little as possible.

The fuzzy *c*-means clustering algorithm is a kind of unsupervised learning in pattern recognition. It does not need training and can be automatically classified by machine learning. Actually, the FCM algorithm is the mapping from the initial clustering center to the final clustering result. Once initialization[9–10] is determined, the clustering result is exactly identified.

Considering the characteristics of text clustering, this paper first extracts the key phrases, and then abstracts the topic concept to build topic concept sub-space. Finally, with the mapping from the text set to the topic concept sub-space, we can have each initial center and membership matrix.

1 Related Work

Hearst and Pedersen[2] found that the related documents should usually be similar, in other words, they should share the same topic. If a phrase is the key phrase that represents the topic of a document, the document set sharing this phrase should be similar to the topic.

Through sorting the order of parsed phrases and then extracting key phrases, Zeng et al. [1] proposed a novel method to cluster web search results. In this approach titles and snippets are used for analysis and key phrase extraction. Because each result set has a phrase as its topic, Zeng' s method is considered to be visually effective and practical. The experiments also indicate that it obtains a satisfactory effect.

1.1 Text parsing and candidate key phrase extraction

Clustering dealt with the whole text in previous versions, which caused too many feature vectors and was very time-consuming. Therefore, in TCS2FCM titles and snippets are used to parse documents.

1) Perform punctuation analysis for titles and snippets so as to divide them into single sentences.

2) Perform lexical analysis for each word of all sentences using the sentence parsing tool (minipar), then generate all the phrases with three or fewer words using *n*-gram method($n = 3$ in this paper)

3) Hande generated phrases:

① Delete phrases containing words in the stopword list;

② Deal with the remained phrases using Porter' s stemming algorithm to merge words with the same stem such as cluster and clustering.

By the above steps, we have the candidate key phrase pool.

1.2 Key phrase extraction

Five attributes are listed below, which represent five aspects of the relationship among documents or among patterns. During the period of text analysis, each candidate key phrase is calculated for these five attrib-

Received 2007-05-18.

Foundation items: The National Natural Science Foundation of China (No. 60672056), Open Fund of MOE-MS Key Laboratory of Multimedia Computing and Communication(No. 06120809).

Biographies: Ji Xianghua(1982—), male, graduate; Yu Nenghai(corresponding author), male, doctor, professor, ynh@ ustc. edu. cn.

utes to receive a final weighted linear combination eigenvalue. The extent to which how much a phrase can represent the topic of a document depends on the eigenvalue:

1) Phrase frequency/inverted document frequency (TFIDF);

2) Phrase length (LEN);

3) Intra-cluster similarity (ICS);

4) Cluster entropy (CE);

5) Phrase independence (PI).

We can sort candidate key phrases by weighted linear eigenvalues in five attributes and select the top N phrases as final key phrases, here we choose N as 75%.

Finally, each key phrase is treated as the topic of each cluster, and each cluster is just a document set sharing each phrase. Now we receive clustering result of corresponding topic.

Although the algorithm in Ref. [1] has the advantage of visual effectiveness, there are still some problems. It just simply classifies texts sharing key phrases into the same cluster, but does not actually use the clustering algorithm. Although there is little effect on the recall, it leads to relatively low clustering precision.

2 Text Clustering using TCS2FCM Algorithm

2.1 Introduction of FCM algorithm

The fuzzy c-means algorithm is an automated classification method of data samples. Through the optimization of the fuzzy objective function, membership of every sample point to the cluster center can be obtained, thus determining the ownership of sample point.

Initialization: choose a constant $\varepsilon > 0$, set the number of iterations $k = 0$, and provide random cluster center $V(0)$.

① Fix cluster number c and fuzzy coefficient m;

② Randomly initialize fuzzy cluster center $C_i (1 \leqslant i \leqslant c)$;

③ Compute membership matrix $\boldsymbol{\mu}$;

④ Update the fuzzy cluster center to C^*, compare C^* and C with an appropriate norm; if $\|C^* - C\| < \varepsilon$, then stop, or return to step ③.

2.2 TCS2FCM algorithm

Only rarely is there a text set containing only one topic. A text set usually has multiple topics. The concept phrases extracted in the above algorithm embody the contents of the text, so they can be seen as a set of topic phrases.

However, previous experiments have shown that the number of above topic phrases was very great. Moreover, due to non-orthogonality of concept phrases in the vector space, there is huge redundancy among them. So much redundancy results in much more clustering time; even worse, it can lower clustering precision.

WordNet[11-12] is a large lexical database of words, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Therefore, we can abstract common concepts from synonymous key phrases as topic concepts.

Meanwhile, we define the topic concept set of those texts similar in content as topic concept sub-spaces. Among them, each topic concept is one dimension of a sub-space and the value of a text in this dimension is exactly the projective value. The higher the value is, the more weight this topic concept occupies in the whole central meaning. Here, it must be noted that how "texts similar in content" can be defined. There are many ways to judge whether texts are similar in content, such as the Euclidean distance, cosine similarity and so on. But in this paper, the conclusion of Ref. [2] has been adopted: the texts sharing topic phrases are similar in content. Therefore, we consider text set sharing a certain amount of topic concept phrases as texts similar in content.

After obtaining topic concept sub-spaces, we can use the center of each sub-space as initial clustering centers. The specific clustering steps are as follows:

① Assume a total number of M topic concept phrases, and with each phrase the initial cluster types have been determined: $C = \{C_1, C_2 \ldots, C_M\}$.

② Assume a total number of N texts and each phrase has its corresponding TF value t_i, so the membership of a phrase can be represented by the proportion TF value of this phrase that occupies the whole TF value

$$\mu_{ik} = \frac{t_{ik}}{\sum_{i=1}^{M} t_{ik}} \qquad k = 1, 2, \ldots, N; i = 1, 2, \ldots, M$$

③ Compute centroid o_i for each initial cluster type C_i using the weighted formula:

$$o_i = \sum_{k=1}^{N} (\mu_{ik})^m \boldsymbol{x}_k \Big/ \sum_{k=1}^{N} (\mu_{ik})^m$$

where μ_{ik} is the membership of the k-th text in C_i and $\boldsymbol{x}_k$ is the k-th text vector, $i = 1, 2, \ldots, M$. If we set the value of C in the FCM algorithm to M, the initial M centers are just initial clustering centers $c_i = o_i$, $i = 1, 2, \ldots, M$.

④ Update membership matrix $\boldsymbol{\mu}^*$:

$$\mu_{ik}^{*} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \qquad \text{if } d_{ij} \neq 0;\ j = 1, 2, \ldots, c$$

$$\mu_{ik}^{*} = 1 \qquad \text{if } d_{ij} = 0$$

$$\mu_{ik}^{*} = 0 \qquad \text{if } d_{tj} = 0;\ t \neq j$$

where d_{jk} represents the distance between the k-th text and the j-th cluster center, $i = 1, 2, \ldots, c$

⑤ Update cluster center C^* according to $\boldsymbol{\mu}^*$:

$$C^{*} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{*})^{m} \boldsymbol{x}_{k}}{\sum_{k=1}^{N} (\mu_{ik}^{*})^{m}}$$

⑥ Compute criterion function $\|C^{*} - C\| < \varepsilon$, if satisfied, clustering is finished, or return to step ④, until criterion function is satisfied.

⑦ The initial cluster topics are just final cluster topics.

3 Evaluation Measurement

It is difficult to measure their final effects of traditional clustering methods, but with this method, we can make use of traditional IR measurements: recall and precision because each cluster has its own topic. This paper has re-defined recall and precision referring to Refs. [1 - 2].

First, the concept phrases are extracted as initial topics; then use manual tagging to distribute every text to its corresponding cluster. In this way, we obtain the reference clustering which is so-called soft clustering. In other words, one text may be able to have multiple topics. Finally, TCS2FCM is used to generate actual clustering. By comparison between the reference clustering and the actual clustering the evaluation measurement is presented.

$$P_i@N = \frac{|N_i \cap F_i|}{|N_i|}, \quad R_i@N = \frac{|N_i \cap F_i|}{|F_i|}$$

where $P_i@N$, $R_i@N$ represent the precision and the recall of the i-th cluster in result cluster set N, respectively; N_i and F_i represent the i-th actual cluster and the i-th reference cluster, respectively.

The total recall R and total precision P are the average of every recall and every precision.

$$P@N = \frac{\sum_{i=1}^{C} P_i@N}{M}, \quad R@N = \frac{\sum_{i=1}^{C} R_i@N}{M}$$

4 Experiments and Discussion

In this paper, experiment data set is about 3×10^5 IEEE/IEE articles downloaded from IEEE databases. In the experiment, we use the coefficients of five attributes in Ref. [1], but it should be noted that it can be obtained expediently by the well-known simulated annealing algorithm.

Pal[4] has drawn from the effectiveness of clustering that the best range of m should be from 1.5 to 2.5. Therefore, this paper chooses different m values from 1.5 to 2.5 in order to compare corresponding cluster precision and recall. Experiments show that 1.9 is a proper value of m.

Fig. 1 illustrates the comparison between Ref. [1] and TCS2FCM, which indicates that while corresponding recalls have minor differences, TCS2FCM has an obvious advantage over the algorithm of Ref. [1] in precision and promotes precision considerably.

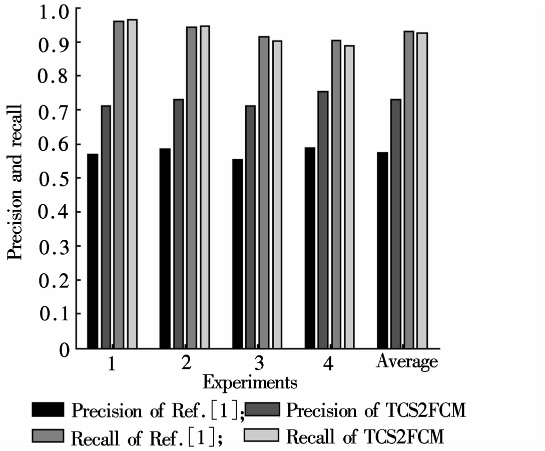

Fig. 1 Comparison of effect between the algorithm of Ref. [1] and TCS2FCM

What is the distinction between FCM of the random initial center and TCS2FCM? Fig. 2 gives the answer. Combining with Fig. 1 we can find that FCM of the random initial center performs unsatisfactorily, sometimes even worse than the algorithm of Ref. [1] in both precision and recall.

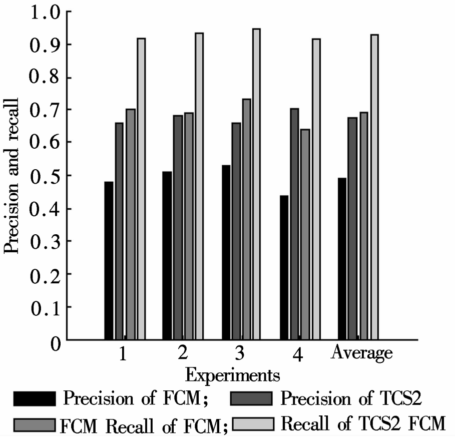

Fig. 2 Comparison of effect between FCM and TCS2FCM

Besides, to make a more comprehensive evaluation of the results, we employed the well-known F-score method to evaluate clustering effects:

$$F_{\text{score}} = \frac{2PR}{P + R}$$

As shown in Fig. 3, the common FCM algorithm will cause retrogression in clustering effects to be even worse than the algorithm of Ref. [1] on the one hand; on the other hand, TCS2FCM performs well in both criterions.

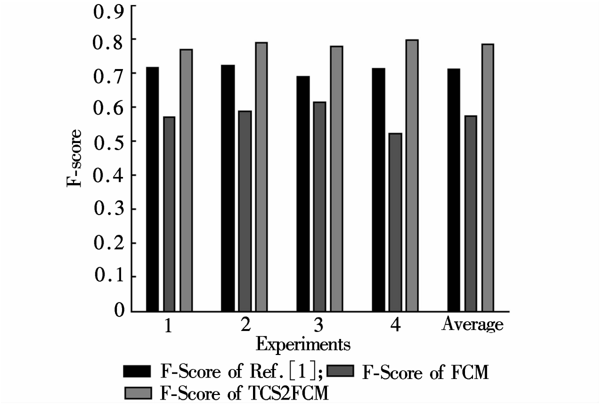

Fig. 3 Comparison of F-Score among Ref. [1], FCM and TCS2FCM

5 Conclusion and Future Work

We conclude that TCS2FCM is really beneficial for text clustering. In this paper, we propose a new method to effectively fuse the topic concept into fuzzy *c*-means clustering, and achieve good results on the performance of text clustering.

Considering the time-consuming characteristics of fuzzy *c*-means clustering, once its efficiency is well promoted, it is more valuable for applications such as article search result clustering, document clustering, etc.

References

[1] Zeng Huajun, He Qicai, Chen Zheng, et al. Learning to cluster web search results[C]//*Proc of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, 2004: 210 – 217.

[2] Hearst M A, Pedersen J O. Reexamining the cluster hypothesis: scatter/gather on retrieval results [C]//*Proc of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Zurich, Switzerland, 1996: 76 – 84.

[3] Jain A K, Murty M N, Flynn P J. Data clustering: a review [J]. *ACM Computing Surveys*, 1999, **31**(3): 264 – 323.

[4] Aiello Marco, Pegoretti Andrea. Textual article clustering in newspaper pages[R]. Trento: University of Trento, 2004.

[5] Ferragina Paolo, Gulli Antonio. A personalized search engine based on web-snippet hierarchical clustering[C]//*Special Interest Tracks and Posters of the 14th International Conference on World Wide Web*. Chiba, Japan, 2005: 801 – 810.

[6] Leouski Anton V, Croft W Bruce. An evaluation of techniques for clustering search results[R]. Amherst: Computer Science Department of University of Massachusetts, 1996.

[7] Pal N R, Bezdek J C. On cluster validity for the fuzzy *c*-means model [J]. *IEEE Trans on Fuzzy Systems*, 1995, **3** (3): 370 – 379.

[8] Chai Shengsan. Application of content words and co-citation clustering analysis to science structure studies [J]. *Journal of the China Society for Scientific and Technical Information*, 1997, **16**(1): 69 – 74. (in Chinese)

[9] Fan Jiulun, Wu Chengmao. The new explanation of membership degree in FCM and its applications[J]. *Journal of Electronics*, 2004, **32**(2): 350 – 352. (in Chinese)

[10] Xue Zhong, Xie Weixin. A initialization method of the fuzzy *C*-means clustering algorithm[J]. *Systems Engineering and Electronics*, 1995, **17**(11): 64 – 69. (in Chinese)

[11] Hotho A, Staab S, Stumme G. Wordnet improves text document clustering[C]//*Proc of the SIGIR 2003 Semantic Web Workshop*. Toronto, Canada, 2003.

[12] Shehata Shady, Karray Fakhri, Kamel Mohamed. Enhancing text clustering using concept-based mining model [C]//*Proc of the Sixth International Conference on Data Mining*. Washington, DC: IEEE Computer Society, 2006: 1043 – 1048.

基于主题概念空间的文本模糊 *c*-均值聚类方法

吉翔华[1] 陈 超[2] 邵正荣[2] 俞能海[1]

([1] 中国科学技术大学多媒体计算与通信教育部-微软重点实验,合肥 230027)
([2] 中国科学技术大学图书馆,合肥 230027)

摘要:为了改善文本聚类的准确度,提出用基于主题概念子空间的模糊 *c*-均值聚类(TCS2FCM)方法来分类文本. 采用 5 个评估函数的加权值来提取关键短语;利用 WordNet®对相应的关键短语提取概念短语并生成最后的类别描述. 初始中心和初始隶属度矩阵的建立是决定模糊 *c*-均值聚类效果的关键,使用能够代表文本主题的概念短语来建立相互正交的主题概念子空间,利用主题子空间中的概念向量来初始化聚类中心和隶属度矩阵. 实验结果表明:不同于传统模糊 *c*-均值聚类的随机化初始,与文本内容相关的初始化有助于改进最后的聚类结果,提高聚类精度.

关键词:TCS2FCM;主题概念空间;模糊 *c*-均值聚类;文本聚类

中图分类号:TP391