一种基于主题的文本聚类方法

赵世奇，刘挺，李生

（哈尔滨工业大学信息检索实验室，哈尔滨 150001）

摘 要： 现有的文本聚类方法难以正确识别和描述文本的主题，从而难以实现按照主题对文本进行聚类。本文提出了一种新的基于主题的文本聚类方法：LFIC。该方法能够准确识别文本主题并对文本进行聚类。本方法定义和抽取了“主题元素”，并利用其进行基本类索引。同时还整合利用了语言学特征。实验表明，LFIC 的聚类准确率达到 94.66%，优于几种传统聚类方法。

关键词： 基于主题文本聚类；基本类索引；语言学特征

A Topical Document Clustering Method

Shiqi Zhao, Ting Liu, Sheng Li

(Information Retrieval Lab, Harbin Institute of Technology, Harbin 150001)

Abstract: Few of the existing document clustering methods can detect or describe document topics properly, which makes it difficult to conduct clustering based on topics. In this paper, we introduce a novel topical document clustering method called Linguistic Features Indexing Clustering (LFIC), which can identify topics accurately and cluster documents according to these topics. In LFIC, “topic elements” are defined and extracted for indexing base clusters. Additionally, linguistic features are investigated and exploited. Experimental results show that LFIC can gain a higher precision (94.66%) than some widely used traditional clustering methods.

Keywords: topical document clustering; base clusters indexing; linguistic features

1 Introduction

With the continuing growth of online information, it becomes more and more important to provide improved mechanisms to organize the web documents and facilitate users' access to the information they need. Therefore, document clustering has been widely researched. Although a variety of document clustering methods have been proposed, few of them can cluster documents based on topics. In this paper, a new clustering method is presented, which is named Linguistic Features Indexing Clustering (LFIC) method. The main question in a topical clustering method is how to describe a topic. We believe that a topic, which consists of a series of closely related events [1], is represented by a set of “topic elements”, such as participants, locations, dates, properties, and activities. According to the above principle, we set up “topic elements indexes”, whereby documents on the same topic can be indexed and form base clusters.

We conduct a number of evaluations that compare LFIC with some traditional clustering methods, including Agglomerative Hierarchical Clustering (AHC), K-means Clustering (KMC) and Suffix Tree Clustering (STC). Results show

基金资助：受自然科学基金资助（60435020，60575042，60503072）；受腾讯基金项目资助

作者简介：赵世奇(1981-)，男，辽宁抚顺，博士研究生，zhaosq@ir.hit.edu.cn

that the LFIC method can achieve a higher precision while maintaining an acceptable level of recall. Considering the "information overload" on the web, precision is more important than recall. Thus we can conclude that LFIC is effective.

2 Related Work

Document clustering methods can be classified into two categories: hierarchical and partitional. Agglomerative hierarchical clustering (AHC) methods aim to build a hierarchy of document classes. AHC methods start with the documents as individual clusters and, at each step, merge the most similar pair of clusters. The result of AHC methods can be displayed as a tree which could show the merging process and intermediate clusters clearly. K-means clustering is one kind of partitional clustering methods. It starts by defining k cluster centroids and then compares every document with every centroid. Each document is assigned to the cluster with the closest centroid. After that, the method recomputes the centroid of each cluster as the average of the cluster's elements. The process of assigning and recomputing repeats until the centroids don't change.

Though the methods are widely accepted, they share some common weaknesses. For instance, both of them need a preset halting criterion that is usually difficult to get in practice. Besides, neither of them can describe the clustering results properly. What's more, these methods confine each document to only one cluster regardless of the fact that a document may be on multiple topics.

To overcome the above shortcomings, Zamir and Etzioni have introduced a method named Suffix Tree Clustering (STC) [2]. STC utilizes a suffix tree [3] to efficiently detect documents sharing common phrases and uses this information to construct base clusters. We can view the shared common phrases as indexes of each cluster. In order to avoid the proliferation of identical clusters, STC merges base clusters with a high overlap. STC doesn't require users to specify the number of clusters and it can make use of the shared phrases to describe the resultant clusters. Moreover, STC allows a single document to appear in more than one cluster.

LFIC borrows the idea of using indexes to create base clusters. The indexes in LFIC are shared "topic elements". Generally, documents indexed by identical topic elements are on the same topic. Some topic elements, such as participants, locations, and dates are expressed as named entities (NEs). The rest, including properties and activities correspond to nouns and verbs. Therefore, we need to exploit linguistic features such as part-of-speech (POS) and NE to extract topic elements.

3 LFIC Method

3.1 Indexing Base Clusters

As introduced before, STC method has some obvious advantages. Actually, these good qualities should all be ascribed to the smart way of forming base clusters by creating an inverted index of phrases for the documents. For the sake of convenience, we call this kind of way indexing base clusters hereafter. The precise meaning of indexing base clusters can be formulated as below:

Let $D = \{D_1, D_2, ..., D_n\}$ be a document collection and $I = \{I_1, I_2, ..., I_m\}$ be a set of chosen indexes. Document D_i will be placed in the cluster indexed by I_j if and only if the number of times that I_j occurs in D_i exceeds a predetermined threshold T.

In indexing base clusters, no halting constraint is needed. That's deserving commendation because it is impossible to determine the optimal number of clusters beforehand with the scales of documents and the numbers of actual topics varying a lot in practice. Besides, a single document may be on two or more topics simultaneity. Thus, it seems reasonable to allow a document to appear in more than one cluster. This need is met by indexing base clusters as a document can be indexed by several indexes. Furthermore, the clusters can be described by indexing base clusters since each base cluster created in this way has an index which represents the cluster's content.

3.2 Exploiting Linguistic Features

We hope to benefit from indexing base clusters just as STC. However, the indexes used in STC are merely phrases or rather n-grams which have the following shortcomings. On the one hand, lots of n-grams make no sense or hardly represent any topics. On the other hand, truly valuable information often can't be confined in certain n-grams, since a same meaning can be expressed differently.

We take advantage of linguistic features in forming indexes. In natural language processing, a document is generally represented as a vector of words [4]. The weights of words are usually calculated using statistical techniques (such as "tf · idf"). Nevertheless, the linguistic features of words themselves should also be given enough attention. Take POS for example, intuitively, words with different POSes should have different contributions to characterizing a document [5]. Usually, nouns and verbs are the most indicative. Adjectives and adverbs are less valuable. Function words have little or no influence and should be excluded as stopwords. In addition, NEs are more discriminative than normal words and should be assigned higher weight. In LFIC, we form the indexes by utilizing NEs coupled with important nouns and verbs.

3.3 Main Steps

Document preprocessing and representing: First, documents are submitted to sequential preprocessing modules including word segmentation, POS tagging, and NE recognition. In this step, stopwords are removed. Here, a stoplist containing punctuations, common used words and some news specific words (e.g. names of newspaper offices and some journalistic terms) is maintained. In the stage of document representation, Vector Space Model (VSM) is employed. The vector terms here contain only NEs, nouns, and verbs. The tf · idf is used for weighing the vector terms.

Forming indexes and creating base clusters: An index used in LFIC consists of two parts: an NE-part and a keyword-part:

Definition 1: Let D be a document, $X = \{x_1, x_2, ..., x_m\}$ be a set of NEs occuring at least twice in D, $Y = \{y_1, y_2, ..., y_n\}$ be a set of keywords (nouns and verbs) in D whose tf · idf weights exceed a preset threshold T. $\forall x \in X$ and $y \in Y$, the two-tuple of (x, y) is defined as one of D's indexes.

If the size of X and Y are m and n, then document D has $m \times n$ chances to be indexed by any of its indexes. This makes it possible that a single document can be indexed on different topics and put into several base clusters. With the well designed indexes combining NEs and keywords, LFIC constructs base clusters by merging documents that share common indexes.

Combining base clusters into clusters: The base clusters formed in the last section overlap a lot. Hence we combine base clusters to reduce duplication and form more complete clusters. Let c_i, c_j be two base clusters. If their distance is less than a preset threshold *Thre*, then they will be combined. In order to measure the distance between two base clusters, the centroids of them have to be calculated. The distance measure used in the combination algorithm is the cosine measure:

$$\cos(c_i, c_j) = c_i \cdot c_j / \|c_i\| \|c_j\| \quad (1)$$

4 Experiments

4.1 Data and Metrics

The method is evaluated using a collection of 2021 news documents collected from the web. From these documents, we have manually identified 266 topics, whose maximum size is 24 documents and minimum is 3. In this paper, the precision and recall are computed respectively in evaluation.

Given a particular topic T_i of size n_i and a particular cluster C_j of size n_j, suppose n_{ij} documents in the cluster C_j belong to T_i, then the precision of this topic and cluster is defined to be:

$$precision(T_i, C_j) = n_{ij} / n_j \quad (2)$$

The precision of T_i is the maximum precision value attained at any cluster in the cluster set C:

$$precision(T_i) = \max_{C_j \in C} precision(T_i, C_j) \tag{3}$$

The overall precision is computed by taking the weighted average of the individual precision:

$$precision = \sum_{i=1}^{N_T} \frac{n_i}{N} precision(T_i) \tag{4}$$

where N is the total number of documents and N_T is the number of topics.

Similarly, the recall of the entire clustering results can be defined as

$$recall = \sum_{i=1}^{N_T} \frac{n_i}{N} recall(T_i) \tag{5}$$

where $recall(T_i) = \max_{C_j \in C} recall(T_i, C_j)$ and $recall(T_i, C_j) = n_{ij} / n_i$

4.2 Comparisons with Other Methods

We conduct two sets of experiments. In the first set, the LFIC is compared with AHC, KMC and STC. The stopping criterion for AHC and KMC is set to 266, which is the factual number of topics.

First of all, the precision of the above four methods are computed and compared (Fig.1). As expected, the LFIC method scores highest. We believe that this positive result is mainly due to LFIC's well designed indexes which can identify topics more accurately.

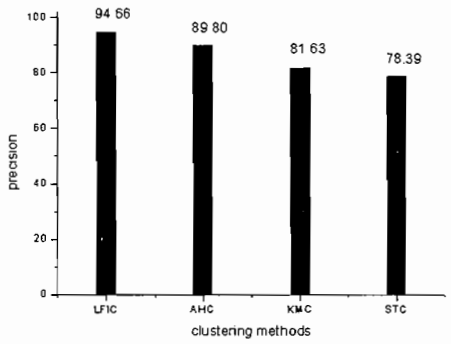

图 1 四种聚类方法的准确率

Fig.2 Precision of four clustering methods

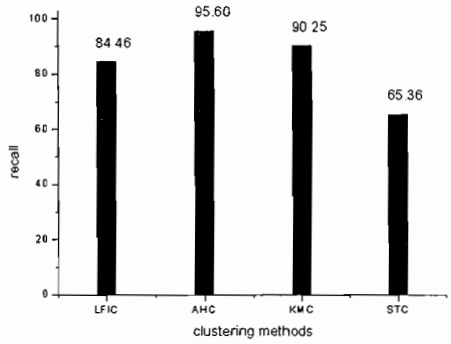

图 2 四种聚类方法的召回率

Fig.3 Recall of four clustering methods

Fig.2 compares the recall of the methods. We can see that the recall of LFIC is 84.46%, which is lower than AHC's 95.60% and KMC's 90.25%. This is partially because that some documents are difficult to be indexed in LFIC, since they share barely any common NEs or keywords with other documents. However, 84.46% is high enough if we consider the information overload of the web.

4.3 Comparisons of Different Parts

In the second set of experiments, we evaluate the contributions of LFIC's different parts. Firstly, we try to find out whether the indexes involving both NEs and keywords work better than using only NEs or keywords. Tab.1 compares three runs of LFICs which use different kinds of indexes. It is obvious that the LFIC indexed by keywords plus NEs performs much better than that indexed by keywords in both precision and recall. This indicates that it is not enough to describe topics using keywords alone. We can also see that the run using NEs alone achieves the highest recall but the lowest precision. This is because, in the experimental data, more than one topic may be related to the same NE entity. Thus a single NE may index several topics. This results in many "large" base clusters which should be responsible for the high recall and low precision.

表 1 **LFIC 中不同的基本类索引方法的比较**

Tab.1 Comparison of LFICs using different kinds of indexes for creating base clusters

different kinds of indexes	precision(%)	recall(%)
keywords + NEs	94.66	84.46
keywords only	88.19	78.08
NEs only	78.31	86.94

Secondly, we try to prove the necessity of combining base clusters. To this end, we compare the resultant clusters produced by LFIC with its uncombined base clusters. The comparison result is presented in Tab.2. Notice that, the recall achieved by the uncombined base clusters is lower than 50%, which is unacceptable. This is not surprising, as the documents on a certain topic could be "partitioned" into several base clusters, and the combination is necessary for producing larger, more complete clusters. We can conclude that combining base clusters can improve recall significantly.

表 2　　LFIC 中基本类合并与基本类未合并的比较

Tab.2 Comparison of LFICs having base clusters combined or not

base clusters combined or not	precision(%)	recall(%)
combined	94.66	84.46
uncombined	96.90	49.63

5 Conclusions

In this paper, we propose a novel clustering method LFIC. The LFIC method has three main points. First, LFIC is a topical document clustering method that can identify topics by indexing base clusters. Second, in LFIC, linguistic features are investigated. NEs, nouns and verbs are used to form indexes and document vectors. Third, base clusters are combined to enhance the recall.

Our experiments reveal that the LFIC method attains a higher precision than some widely used clustering methods, and the recall is fairly acceptable. Experiments also indicate that the techniques we introduce in LFIC, which include indexing base clusters, exploiting linguistic features, and combining base clusters, are all critical to the success of LFIC.

Admittedly, there are still some limitations in LFIC. First and foremost, LFIC relies on the performance of the underlying modules, especially the NE recognizer. What's more, in LFIC some heuristic thresholds are needed which may influence its effectiveness.

In future work, we will employ more and richer NE features in LFIC, such as dates, time etc. These features will be helpful in topic identifying. In addition, since the present used algorithm for combining base clusters is simple, we aim to improve the combination algorithm.

References

[1] Hatzivassiloglou V, Gravano L and Maganti A. An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering [A]. In Proceedings of the 23rd ACM SIGIR Conference, Athens [C]. 2000. pp 224-231.

[2] Zamir O and Etzioni O. Web Document Clustering: A Feasibility Demonstration [A]. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. 1998. pp 46-54.

[3] Gusfield D. Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology [M]. Cambridge, UK: Cambridge University Press, 1997.

[4] Lee D-L, Chuang H and Seamons K. Document Ranking and the Vector-Space Model [J]. IEEE Software, 1997, Vol.14 (2): 67-75.

[5] Kummamuru K, Lotlikar R, Roy S, et al. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results [A]. In Proceedings of the 13th International Conference on World Wide Web [C]. 2004. 658-665.