Expert Systems with Applications 37 (2010) 2247–2254

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Generic title labeling for clustered documents

Yuen-Hsien Tseng

National Taiwan Normal University, No. 162, Section 1, Heping East Road, Taipei 106, Taiwan, ROC

ARTICLE INFO

Keywords:
Hypernym search
Clustering labeling
WordNet
Correlation coefficient
Topic identification

ABSTRACT

Document clustering is a powerful technique to detect topics and their relations for information browsing, analysis, and organization. However, clustered documents require post-assignment of descriptive titles to help users interpret the results. Existing techniques often assign labels to clusters based only on the terms that the clustered documents contain, which may not be sufficient for some applications. To solve this problem, a cluster labeling algorithm for creating generic titles, based on external resources such as WordNet, is proposed. Our method first extracts category-specific terms as cluster descriptors. These descriptors are then mapped to generic terms based on a hypernym search algorithm. The proposed method has been evaluated on a patent document collection and a subset of the Reuters-21578 collection. Experimental results revealed that our method performs as anticipated. Real-case applications of these generic terms show promising in assisting humans in interpreting the clustered topics. Our method is general enough such that it can be easily extended to use other hierarchical resources for adaptable label generation.

1. Introduction

Document clustering is a powerful text mining technique to detect topics and their relations for information browsing, analysis, and organization. However, unlike document categorization where a set of labels or terms is predefined for each category, the documents sets partitioned, either hierarchically or non-hierarchically, by a clustering algorithm have virtually no such predefined labels for convenient recognition of the content of each set. Each of them requires post-assignment of a concise and descriptive title to help analysts to interpret the result. Although good clustering algorithms are widely available, good solutions for labeling the clustered results to meet analysts' needs are rare.

Most existing work selects the title words from the terms contained in the clustered documents themselves. Although this is justifiable, this may not be sufficient. It would be desirable to further suggest generic topic terms for ease of analysis, especially in the applications where documents cover a wide range of domain knowledge. Examples of this need are often found in topic analysis for patent or scientific publications (National Institute of Science & Technology Policy, 2004; Noyons & Van Raan, 1998; Uchida et al., 2004).

In this work, we attempt to automatically create generic labels which do not necessarily exist in the clustered documents for easier cluster interpretation. As an example, if the documents in a cluster were talking about *tables*, *chairs*, and *beds*, then a title labeled "*furniture*" would be perfect for this cluster, especially when this hypernym does not occur in it (or occurs rarely). This kind of problem was often solved by human experts, such as those in Glenisson, Glanzel, Janssens, and De Moor (2005), Lai and Wu (2005), where cluster titles were given manually. To make our automatic approach feasible, external resources such as WordNet or other hierarchical knowledge structures are used. Our method first selects content-indicative terms for each cluster. A proposed hypernym search algorithm is then applied to map these terms into their generic title.

The rest of the paper is organized as follows: Section 2 reviews some related work. Section 3 introduces our method for content-indicative term extraction. Section 4 describes the hypernym search algorithm based on WordNet. Section 5 details the experiments that evaluate our method. Section 6 discusses the results and suggests possible improvement. Section 7 concludes this work and shows its implications.

2. Related work

Labeling a clustered set of documents is an inevitable task in text clustering applications. Automatic labeling methods mainly rely on extracting significant terms from clustered documents, where the term significance can be calculated very differently from clustering algorithms to algorithms.

For example, in the vector space model, where clusters are represented as weighted sums or centroids of the document vectors, terms with heaviest weights in the cluster vectors are extracted as the cluster labels. In Cutting, Karger, Pedersen, and Tukey

E-mail address: samtseng@ntnu.edu.tw

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.07.048

(1992), Hearst and Pedersen (1996), the term weight in a document vector is the normalized term frequency (TF), while in Yang et al. (2000), it is a version of the TF multiplied by IDF (Inverse Document Frequency) weighting scheme. As to its effectiveness, Sahami, Mehran, Yusufali, Salim, and Baldonaldo (1998) pointed out (although without experiments) that the simple centroid-based approach outperformed the probabilistic odds scheme which computes the ratio of the conditional probability of a term appearing in a cluster over the sum of the conditional probabilities of the term in other clusters.

Treeratpituk and Callan (2006) proposed a linear model to combined at least 10 features including TF, TF × IDF, and statistical information from the surrounding clusters (sibling clusters or parent cluster) to rank the terms in the clustered documents. However, it requires human-labeled titles to train the model for setting optimal model parameters.

In the Self-Organization Map (SOM) method (Lagus, Kaski, & Kohonen, 2004), where clusters are organized in a 2-D map, the label of a cluster is the term having the highest goodness measure. This goodness measure is the square of the relative term frequency in the cluster normalized by the sum of the relative term frequencies in other distant clusters.

In an application to group terms for detecting events over time (Swan & Allan, 2000), the cluster title consists of the highest ranked named entity followed by the highest ranked noun phrase. The ranks of these terms were obtained by sorting the maximum chi-square values of the terms occurring in a time interval.

In clustering web search results, the longest phrases occurred in most documents in a cluster were used as its title (Zamir & Etzioni, 1998).

In other related fields, such as document summarization and translation, there were tasks in the Document Understanding Conference (DUC) (Document Understanding Conferences, 2004) to generate very short summaries. These short 10-words summaries have the potential to serve as cluster titles. However, most participants use extraction-based methods (Banko, Michele, Mittal, & Witbrock, 2000). Even though there were studies that generate document titles not from the document themselves, a large corpus of documents with human-assigned titles is required to train the "translation model" so as to map document words into human-assigned titles (Kennedy & Hauptmann, 2000). Besides, these summaries tend to be event-descriptive rather than topic-indicative for a set of documents.

As can be seen, despite there are various techniques to label a set of documents, there are few studies that attempted to deal with the problem that we propose here.

3. Cluster descriptor selection

The methods to extract cluster descriptors in the above-mentioned studies are mostly related to their clustering algorithms. In our application, we would like to have a general approach that is independent of the clustering process. To this end, we seek solutions from text categorization where selecting the best content-revealing or category-predictive features has been widely studied.

Yang and Pedersen (1997) had compared five different methods for selecting category-specific terms. They found that *chi-square* is among the best that lead to highest categorization performance. The chi-square method computes the relatedness of term T with respect to category C in the manner:

$$\chi^2(T,C) = \frac{(\mathrm{TP}\times\mathrm{TN} - \mathrm{FN}\times\mathrm{FP})^2}{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})(\mathrm{FN}+\mathrm{TN})}, \quad (1)$$

where TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) denote the number of documents that belong or not belong to C while containing or not containing T, respectively. With content terms (in the same cluster) sorted by their chi-square values in descending order, top-ranked terms can be selected as the cluster descriptors.

However, chi-square does not distinguish negatively related terms from positively ones. A remedy to this problem is the use of the *correlation coefficient* (CC), which is just the square root of the chi-square:

$$Co(T,C) = \frac{(\mathrm{TP}\times\mathrm{TN} - \mathrm{FN}\times\mathrm{FP})}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{FP}+\mathrm{TN})(\mathrm{TP}+\mathrm{FP})(\mathrm{FN}+\mathrm{TN})}}. \quad (2)$$

As pointed out by Ng et al. (1997), correlation coefficient selects exactly those words that are highly indicative of membership in a category, whereas the chi-square method not only picks out this set of terms but also those terms that are indicative of non-membership in that category. This is especially true when the selected terms are in small number. As an example, in a small real-world collection of 116 short documents (containing only the titles) with only two exclusive categories: construction vs. non-construction in civil engineering tasks, some of the best and worst terms that are computed by both methods are shown in Table 1. As can be seen, "engineering" is a lowest-ranked term (−0.7880) based on correlation coefficient in the non-construction category, while it is a top-ranked term in both categories based on chi-square (0.6210 is the square of −0.7880). Therefore, instead of chi-square, correlation coefficient is used as our basic descriptor selection method.

A further analysis of the correlation coefficient method reveals that it may be effective only for the case where the collection to be clustered contains large number of short documents. For the collections containing long documents, it tends to select category-specific terms that are too specific due to the ignorance of the term frequency in each document (i.e., TF). Therefore, we choose only those terms whose document frequency in a cluster exceeds a ratio r of the number of documents in that cluster. We denote this revised method as CC_r, where r is a tunable parameter and is 0.5 in our implementation. Another remedy is to multiply the term's CC with its total term frequency in the cluster (TFC), denoted as CC × TFC, where TFC is the sum of a term's TF over all documents in the cluster.

4. Generic title generation

The cluster descriptors generated in the above may not be topic-indicative enough to well summarize the contents of the clusters (e.g., see the examples in later sections). One might need to map the identified clusters into some predefined categories for supporting other data mining tasks (e.g. Feldman, Dagan, & Hirsh, 1998). If the categories have existing data for training, this mapping can be recast into a standard text categorization problem, to which many solutions can be applied. Another need arises from the situation that there is no suitable classification system at hand, but some generic labels are still desired for quick interpretations. This case is often solved by human experts, where cluster titles are given manually. Below we propose an automatic solution by use of an extra resource, i.e., WordNet.

WordNet is a digital lexical reference system (WordNet, 2004). English nouns, verbs, adjectives and adverbs are organized into synonym sets. Different relations, such as hypernym, hyponym, meronym, or holonym, are defined to link the synonym sets. With these structures, one can look up in WordNet all the hypernyms of a set of given terms and then choose the best among them as the generic title with some heuristic rules. Since the hypernyms were organized hierarchically, the higher is the level, the more generic are the hypernyms. To maintain the specificity of the set of terms while revealing their general topics, the heuristics have to choose