Using Text Mining and Clustering to Group Research Proposals for Research Project Selection

Yibo Wang  
School of Information,  
Renmin University of China,  
Beijing, 100872, China  
wangyibo90@yeah.net

Wei Xu  
School of Information,  
Renmin University of China,  
Beijing, 100872, China  
weixu@ruc.edu.cn

Hongxun Jiang  
School of Information,  
Renmin University of China,  
Beijing, 100872, China  
jianghx@ruc.edu.cn

Abstract

Great efforts have been taken to guarantee the efficiency of research project selection. Before research proposals are assigned to appropriate experts for evaluation, research proposals needed to be grouped according to their similarities in research topics. This paper applies an ensemble method to cluster research proposals to support research project selection. Several cluster algorithms are applied to group the proposals and the merits of each algorithm in text clustering are made full use of in the form of ensemble method. Based on the real data from Beijing Natural Science Foundation (BNSF), empirical experiments are conducted to validate the effectiveness of the proposed method. Meanwhile, the proposed method promotes the efficiency in the proposal grouping process.

1. Introduction

Research project selection plays a regular and important role in many organizations (e.g. research funding agencies). To support research project selection task efficiently and effectively, several steps should be carried out. In general, an organization that needs new technologies announces the call for proposals first, and then waits submission before deadline. Second, the submitted proposals then are assigned to external experts in the related field for peer review. Thirdly, external experts review the proposals with their specialized knowledge according to the rules from the organization, and return the evaluation results. Eventually, the organization collects and aggregates the evaluation results and makes the final decision.

Among aforementioned four steps for research project selection, reviewer assignment is critically important because external experts cannot review the assigned proposals or give valuable comments when they deal with proposals which are not familiar. So, research on reviewer assignment has been carried out in recent years, especially in research funding agencies. For example, Sun et al. [1] proposed a hybrid knowledge and model method for reviewer assignment. Similarly, Xu et al. [2] provided a decision support approach for assigning reviewers to proposals. When the number of proposals is extremely large, they are required to be grouped according their similarities first and then the grouped proposals are sent to experts for peer review [3]. Therefore, how to group proposals effectively and properly based on knowledge structure of experts and content similarities of proposals is one of the important issues in research project selection. Several formal methods have been proposed to support research proposal grouping. For example, Fan et al. [4] developed a hybrid method using genetic algorithm and knowledge rule to balance knowledge structure of experts, while Ma et al [3] proposed an ontology-based text mining method using the content similarities of research proposals.

A number of methods for research proposal grouping have been applied to National Science Foundation of China (NSFC) [3, 5]. For example, Xu et al. [5] proposed an ontology-based frequent itemset method to solve research proposal grouping problem to enhance the effectiveness of reviewer assignment in the context of NSFC. The requirement of that demand also has happened in other research funding agencies, such as Beijing Natural Science Foundation (BNSF). However, the organizational manners of proposal submission are different among various research funding agencies. Taking NSFC and BNSF for example, NSFC has eight departments, and each department includes some divisions. Every division contains several discipline codes and the discipline codes nearly fixed each year. When an applicant submits research proposals to NSFC, he/she
may first propose a research propose, and then select the proper discipline codes for submission. Different from NSFC, BNSF only has nine departments. Every year, each department gives several research directions when call for proposal begins, and the research directions are dynamic each year. So, an applicant should select proper research direction first, and then submit the research proposal to that direction. With the different organizational manners, the methods for research proposal grouping are different. In NSFC, a research ontology can be built and renewed by collecting the keywords of the proposals, and research proposal grouping process can be carried out by incorporating the research ontology to get a better performance. However, In BNSF, due to the dynamics and uncertain of research directions, the research ontology constructed by BNSF may be less effective for research proposal grouping. In BNSF, manual matching is firstly introduced for research proposal grouping. However, proposals cannot often be accurately grouped by project managers because of the huge number and their understanding deviation. Therefore, an accurate and efficient method to sort the research proposals with the support of computers is in great need. As commonly used cluster algorithms cannot group research proposals effectively, we propose a cluster ensemble method to support research proposal grouping process.

The rest of this article is organized as follows. Section 2 introduces the relevant literature and theoretical background of this research. The method is proposed with details in Section 3, followed by Section 4 that gives empirical analysis and shows the experimental results. The conclusion, limitations and future work are provided in Section 5.

2. Literature review

Great efforts have been taken by research funding agencies to guarantee the order and efficiency of research project selection. Previous studies solve this problem from different aspects. For example, Machacha and Bhattacharya [6] presented fuzzy logic method, while Huang et al. [7] employed fuzzy AHP and crisp judgment matrix. Similarily, Amiri [8] developed the method by adding fuzzy TOPSIS method. Chang and Lee [9] used DEA method with knapsack formulation and fuzzy set theory. Chiang and Che [10] presented fuzzy AHP and fuzzy DEA. Eilat et al. [11] developed an extended DEA method. Chen and Gorla [12] established a fuzzy logic model as decision tool to support incomplete information processing. Henriksen and Traynor [13] proposed an improved scoring and ranking tool for research and development (R&D) project selection and evaluation. Butler et al. [14] developed a procedure combining statistical ranking and selection with multiple attribute utility theory. Greiner et al. [15] proposed a hybrid decision support methodology combining a 0-1 integer portfolio optimization model with the analytic hierarchy process (AHP) to support project selection. Tian et al. [16] designed an organizational decision support method to select R&D projects. Cook et al. [17] offered an approach of optimal allocation of proposals to experts in two ways: a heuristic procedure and an integer-programming set-covering model. Besides general research project selection process, to support research project selection, some specific methods and systems are proposed, such as reviewer assignment [1-2], and especially research proposal grouping [3-5].

The aim of research proposal grouping is categorizing proposals into different groups such that proposals in the same group are more similar to each other than to proposals in any other group. Clustering algorithms share the same aim. A large number of clustering algorithms have been promoted to search for a better solution of categorization without label; however, no single clustering algorithm performs best for all datasets [18]. Cluster ensemble appears as an effective method to overcome the limitation. A lot of researches have been promoted on clustering ensembles. For instance, Dimitriadou et al. [19] proposed a Voting-Merging method (V-M) for cluster ensemble. Dimitriadou et al. [20] promoted voting for fuzzy clustering method. Ayad and Kamel [21] used a voting consensus method with cumulative property. Zhou and Tang [22] proposed an ensemble comprising several clusters which are trained by k-means algorithm with different initial parameter. What’s more, graph and hypergraph based cluster ensemble method is also promoted. Strehl and Ghosh [23] proposed three heuristics based on hypergraph partitioning. In the method, clustering ensemble is represented by hypergraph and each partition is represented by hyper-edge. Topchy et al. [24] offered a probabilistic model of consensus function to cluster ensemble. Yoon et al. [25-26] applied genetic algorithms in cluster ensemble and promoted heterogeneous clustering ensemble. Luo et al. [27] used information based genetic algorithm in cluster ensemble. Vega-Pons et al. [28] proposed a cluster ensemble method based on kernel functions. Mahmood et al. [29] proposed a semi-supervised cluster ensemble method for web video categorization.

Document clustering is an important and challenging research issue in which cluster ensemble

However, research aforementioned has not applied to research proposal grouping, such as BNSF. Therefore, we need to extract features hidden in proposals and convert them from implicit knowledge to explicit knowledge. The extraction and conversion can be supported by text mining methods. What’s more, research proposals consist of different kinds of data types, such as numeric data and nominal data. And no single cluster algorithm performs best for all kinds of data type. Hence, we use ensemble method to make full use of merits of different types of cluster algorithms. Clustering ensemble performs as an effective method for improving robustness and stability of cluster solutions.

This paper proposes a text mining-based cluster ensemble method for Chinese research proposal grouping. It adopts three different kinds of clustering algorithms and ensemble the results of them to cluster the proposals. Empirical experiments have been conducted to test the validity of the proposed approach. The result indicated that the cluster ensemble method improves the accuracy and stability of single clustering algorithms in Chinese research proposal grouping process.

3. Cluster ensemble for research proposal grouping

3.1. The framework of the proposed method

Several steps are taken by BNSF on the submitted proposals for research proposal grouping. Processing capacity of manual grouping cannot match the large scale and complexity of the task. When a large number of proposals of a same topic wait to be segregated into groups, subjective views and possible misinterpretations may lead to inefficient grouping. Though several text-mining methods have been proposed to classify documents in English, the results of those methods on Chinese text are not effective and robust enough to process research proposals. Ma et al. [3] proposed a method to support research proposals grouping in Chinese text. However, the proposals mentioned in [3] are organized in the form of discipline codes while the proposals submitted to BNSF are organized in the form of research directions. Hence we propose an ensemble method of text-mining to group Chinese research proposals.

Cluster algorithms classify proposals into different groups based on distance, density or grid, etc. Each of them has its own merits in the use of the features of proposals. Ensemble methods make it possible to make full use of merits of several cluster algorithms. The framework of the proposed method is shown in Figure 1.

As shown in Figure 1, after research proposals are collected, they are preprocessed by Chinese word segmentation. And then various features in segmented proposes, such as TF-IDF, are extracted. Next, cluster algorithms make it possible to reduce the uncertainty in the basis for grouping. After processed in single algorithm, each proposal is sorted into a cluster according to the similarity. Then, we ensemble the result of each cluster algorithm and generate the final grouping result which will be sent to support the organization to assign research proposals to the external experts.

Briefly, the cluster ensemble method for research proposal grouping could be divided into three stages. The first stage includes proposal preprocessing and word segmentation. The second stage focuses on the extraction of text features, and changes segmented research proposals into text vectors. The third stage groups the processed proposals with cluster ensemble method. Details of the three stages are shown in the following three subsections.
3.2. Proposal preprocessing

A proposal consists of several parts, such as Title, Abstract, Keywords and Main body. Different from English and other Latin languages, Chinese text doesn’t have space as natural delimiter. Besides, strings of Chinese characters have a stronger power of expression than a single word; consequently word segmentation is necessary in Chinese text proposal preprocessing. During the word segmentation, Chinese text is split into strings of Chinese characters with a specific meaning in some field. Meanwhile, stop words could be filtered out to prevent interference.

3.3. Text feature extraction

As research proposals are grouped by cluster algorithms in terms of the similarity of features, feature extraction from segmented proposals is a critically important phase. Taking feature extraction in abstract of research proposes for example, TF-IDF is a widely used method for text feature extraction, so, in this paper, we employ TF-IDF for extracting features in abstract of research proposals. The frequency of words emerge more in a proposal, the more importance of the words are. The frequency of words emerge more in all proposes, the less importance of the words are. After feature extraction, each split proposal can be changed into a text vector $V = (v_1, v_2, \ldots, v_m)$, where $m$ is the number of features and $v_i (i = 1, 2, \ldots, m)$ represents the $i$th feature in text vector.

3.4. Proposal clustering with cluster ensemble

Considering the feature of the task, clustering is an appropriate method of grouping a set of proposals in such a way that proposals in the same cluster are more similar to each other than proposals in other clusters. However, features of processed proposals are in the form of different data types, such as nominal, numeric and binary. An individual cluster algorithm can hardly be good at processing all kinds of data types. Therefore we integrated several cluster algorithms to make full use of the merits of different cluster algorithms. For example, some proposals needed to be separated may have similar numeric attributes but their nominal attributes are different and a cluster algorithm good at nominal data can easily group them. Some cluster algorithms perform well with numeric data, such as Self Organizing Map (SOM) algorithm, a type of artificial neural network which can be used for clustering.

A proposal may cover several topics but only one group it should be sorted to. The Expectation-Maximization (EM) algorithm solves such problems by computing fuzzy clustering and probabilistic model-based clustering. A proposal is sorted to a cluster to which it is mostly like by EM with the text vector.
The proposals provided by BNSF are complex data of different types of attributes with high dimension. It requires the grouping method to discover clusters with arbitrary shape and deal with noisy data. Therefore, besides SOM and EM, we also employ a density-based method called Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which can discover arbitrary shape clusters and perform well in dealing with noisy data, for research proposal grouping purpose.

Based on several cluster algorithms, a novel cluster ensemble method is carried out to group research proposals in the following two steps.

Step 1: Initial clustering using individual algorithms.

The split proposals are processed by each of the three cluster algorithms.

SOM is a type of artificial neural network that implements Kohonen’s Self Organizing Map algorithm for unsupervised clustering. The SOM creates a mapping from a high dimensional data space to a low dimensional map space. It clusters data in two steps. Firstly, the algorithm trains input samples to build the map. Then it automatically classifies a new input sample using the mapping.

EM can be divided into two steps. Firstly, EM sorts the proposals to clusters according to the calculation of the cluster probabilities, which are the expected class values; secondly, EM calculates the distribution parameters to maximize the likelihood of the distributions given the proposals available. The algorithm repeats the two steps until the clustering cannot be improved.

DBSCAN assigns proposals as follows: it randomly selects p proposals. A new cluster is created for proposal d which has more than t proposals in the e-neighborhood of d, where t is the minimum proposal number required and e is the radius required of the neighborhood. It repeats the step above and add unselected to the clusters. The algorithm stops until clusters cannot expended.

Step 2: Proposal Clustering using cluster ensemble method.

The results of three individual cluster algorithms are added to the dataset. The original dataset can be expressed as the Matrix \((n, m)\), where \(n\) denotes the number of proposals and \(m\) denotes the number of features. After assigned by cluster algorithms \((C_1, C_2, \ldots, C_h)\), where \(h\) is the number of cluster algorithms, proposal \((i = 1, 2, \ldots, n)\) has several candidate clusters which can be viewed as new features in cluster ensemble. Matrix \((n, m)\) is extended to Matrix \((n, m+h)\). Then clustering on the new dataset with \((m+h)\) features is implemented again.

From the description above, the process of cluster ensemble can be summarized in Table 1.

### Table 1. The cluster ensemble algorithm

<table>
<thead>
<tr>
<th>For i=1 to n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster proposals with cluster algorithm i</td>
</tr>
<tr>
<td>Save cluster i</td>
</tr>
<tr>
<td>End</td>
</tr>
<tr>
<td>Add cluster i=1 to n to the dataset</td>
</tr>
<tr>
<td>For j=1 to n</td>
</tr>
<tr>
<td>Cluster the proposals</td>
</tr>
<tr>
<td>Output final result</td>
</tr>
<tr>
<td>End</td>
</tr>
</tbody>
</table>

#### 3.5. Proposals grouping and reviewer assignment

When research proposes are classified into clusters, it can form research groups automatically. If a cluster is still large, the cluster can further divided into proper groups randomly. So, the grouped proposals could be assigned to external experts for peer review.

### 4. Empirical analysis

#### 4.1. Data description and evaluation criteria

To validate the effectiveness of the proposed approach, several experiments are conducted using the previous granted research projects. The corpus, which consists of 1064 funded research project with 9 different research topics, have been carried out in our empirical analysis.

Furthermore, to evaluate the performance of the clustering, F measure is used, because F measure is the typical criterion for measuring the quality of text clustering [3]. For predefined research topic \(t\) and cluster \(c\), the Precision and Recall are defined as follows:

\[
\text{Precision}(c, t) = \frac{n(c, t)}{n_c} \quad (1)
\]

\[
\text{Recall}(c, t) = \frac{n(c, t)}{n_t} \quad (2)
\]

where \(n(c, t)\) be the number of research proposals in both topic \(t\) and cluster \(c\), \(n_c\) be the number of
research proposals in topic $c$, and $n_c$ be the number of research proposals in cluster $c$.

The F measurement can be denoted as follows:

$$F = \sum_{i}^{n} \max \{ F(i,j) \}$$  \hspace{1cm} (3)

where $F(c,t) = \frac{2 \cdot \text{Recall}(c,t) \cdot \text{Precision}(c,t)}{\text{Recall}(c,t) + \text{Precision}(c,t)}$.

4.2. Experimental results

Research projects are randomly selected, and the number of research projects in our experiments is from 100 to 1000 for the evaluation purpose. The performance of the proposed method is compared with ones of other baseline methods. The empirical results can be found in Table 2. In Table 2, NRP stands for the number of research projects.

<table>
<thead>
<tr>
<th>Methods</th>
<th>NRP 100</th>
<th>NRP 300</th>
<th>NRP 500</th>
<th>NRP 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM</td>
<td>0.43</td>
<td>0.53</td>
<td>0.61</td>
<td>0.70</td>
</tr>
<tr>
<td>EM</td>
<td>0.59</td>
<td>0.47</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>BDSCAN</td>
<td>0.39</td>
<td>0.47</td>
<td>0.39</td>
<td>0.61</td>
</tr>
<tr>
<td>Cluster Ensemble (SOM)</td>
<td>0.64</td>
<td>0.63</td>
<td>0.68</td>
<td>0.83</td>
</tr>
<tr>
<td>Cluster Ensemble (EM)</td>
<td>0.51</td>
<td>0.64</td>
<td>0.67</td>
<td>0.80</td>
</tr>
<tr>
<td>Cluster Ensemble (BDSCAN)</td>
<td>0.44</td>
<td>0.39</td>
<td>0.53</td>
<td>0.66</td>
</tr>
</tbody>
</table>

As can be seen from Table 2, with the increase of the number of research projects, the F values increase in each cluster algorithm. Meanwhile, the cluster ensemble models can improve the clustering performance compared with baseline models.

5. Conclusions and future work

This paper presents a cluster ensemble-based text mining method for grouping of research proposals. It uses text mining and cluster ensemble techniques to
cluster research proposals based on their similarities. The experimental results showed that the proposed method improved the similarity in proposal groups and promotes the efficiency in the proposal grouping process at BNSF.

Future work is needed to assign grouped research proposals to external reviewers systematically. Also, there is a need to empirically compare the results of manual classification to automatic classification. The speed of the manual system can also be compared with the machine-human proposed methods. Finally, the method can be expanded to cluster documents in other fields such as web document clustering.

Acknowledgments

This work was supported in part by 973 Project (Grant No. 2012CB316205), National Natural Science Foundation of China (No. 71001103, 91224008, 91324015), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No. Z13110100041306), and Program for Excellent Talents in Beijing.

References