Article

JIS

Journal of Information Science
2014, Vol. 40(2) 167–179
© The Author(s) 2013
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551513508500
jis.sagepub.com
SAGE

An unsupervised cascade learning scheme for 'cluster-theme keywords' structure extraction from scientific papers

Feiliang Ren
Northeastern University, People's Republic of China

Abstract

The large amount of scientific papers provides a convenient way for users to know the latest research progress of a specific research topic. However, the large volume and the diverse research themes hiding among these papers usually hinder users from conveniently locating the specific papers that they are interested in. To tackle this problem, we propose a novel unsupervised cascade learning scheme that aims to extract a '*cluster-theme keywords*' structure from the related papers of a research topic so as to help users locate their research interests quickly. Our approach first selects some representative papers for a research topic. It then clusters these selected papers into several small clusters with the help of a domain ontology. It finally extracts some theme keywords for each cluster. Our approach not only greatly reduces the time-consuming and labour-intensive paper-seeking process for users, but also comprehensively displays the diverse themes of a research topic. We conducted extensive experiments to evaluate our proposed approach. The experimental results demonstrate the effectiveness of this approach, which produces promising results.

Keywords

'cluster-theme keywords' structure; information extraction; paper selection; text clustering

1. Introduction

Scientific papers are a kind of high-volume and high-value information resource. With them, users can conveniently know the latest research progress and research trends on research topics. However, users often have to browse through lots of papers to find the ones that they are interested in. This is due to the following two reasons. First of all, there are always many related papers for a research topic. Secondly, there are always diverse research themes hiding among these related papers. For example, for the research topic '*statistical machine translation*', there are more than *3600* related papers published *only* in Chinese periodicals in recent years,[1] and this number is still increasing rapidly. In addition, the hiding research themes among these papers are so diverse that they potentially cover following themes: *IBM model, phrase*-based *model, tree-to-string model, tree-to-tree model, string-to-tree model*, and so on. While one may think that employing clustering or multidocument summarization approaches can help users to quickly understand the main content of a research topic, some factors, including the large volume of related papers, the diverse themes and the diverse paper vocabularies, usually prevent these approaches from being employed to their full potential. For example, in the multidocument summarization task at TAC, the number of documents to be summarized is 10, as introduced in Owczarzak and Dang [1], which is far less than the number of related papers for most research topics. Moreover, the research themes of these related papers are usually so diverse that it is difficult to generate a proper summary within several hundred words, which is often required in the multidocument summarization tasks like TAC [1]. On the other hand, although there are diverse themes for a research topic, users' interests are often focused on only a few of them.

Corresponding author:
Feiliang Ren, Northeastern University, Shenyang, Liaoning Province, 110819, People's Republic of China.
Email: renfeiliang@ise.neu.edu.cn

In such cases, we think that users would benefit from a '*cluster-theme keywords*' structure to locate their research interests: grasping the main research themes of a research topic by glancing at the theme keywords of the clusters covered by this research topic and then quickly locating their research interests from these clusters. This is especially useful when users want to have comprehensive information for some research topics quickly.

In this paper, we propose an unsupervised cascade learning scheme to extract the '*cluster-theme keywords*' structure from the related scientific papers of a research topic. Our work aims to help users know the diverse themes of a research topic quickly and further to locate their research interests rapidly. We tackle three issues in this paper. First, to tackle the papers' large volume issue, our approach will select some representative papers for a research topic. Second, to tackle the research theme diversity issue, we cluster the selected representative papers into several small clusters. Third, to tackle the issue of rapidly locating users' interests, we extract some theme keywords for each cluster.

The main contributions of our work can be summarized as follows:

- To the best of our knowledge, this work is the first attempt at extracting the '*cluster-theme keyword*s' structures from scientific papers, which aims to provide a convenient way for users to know the diverse research themes of a research topic and further to locate their research interests quickly.
- Our approach adopts a paper selection strategy that has not been discussed by most clustering or summarization researchers.
- Many measures are taken during the paper clustering process, including feature weight adjustment and feature extension based on a domain ontology, among others.

The reminder of this paper is organized as follows: in Section 2, we briefly review the related work. We introduce the details of our proposed approach and the parameter tuning method in Sections 3 and 4 respectively. We present the experimental results in Section 5. We discuss the proposed method in Section 6, followed by concluding remarks in Section 7.

2. Related work

There are many kinds of methods that can be used to help users to know the diverse themes of a research topic and further help users to rapidly locate their research interests. Among these methods, clustering [2–4], multidocument summarization [5–15] and keyword extraction [16–22] are three often used methods.

For clustering, there are usually two types of basic algorithms: *partitioning*-based algorithms and *hierarchical* algorithms. *Partitioning*-based algorithms construct a partition of a dataset into a set of *k* clusters, and *k* is an input parameter for these algorithms. These *partitioning*-based algorithms typically start with an initial partition of original dataset and then use an iterative strategy to optimize an objective function. Each cluster is represented by the gravity centre of the cluster (*k-means* algorithms) or by one of the objects of the cluster located nearest centre (*k-medoid* algorithms). Generally, *partitioning*-based algorithms use a two-step procedure. First, determine *k* original nodes for each cluster. Second, assign each data object to the '*closest*' cluster. *Hierarchical* algorithms create a hierarchical decomposition of a dataset, and the *hierarchical* decomposition is represented by a tree structure that iteratively splits original dataset into smaller subsets until each subset consists of only one data object. In such a hierarchy, each node of the tree represents a cluster of original dataset. The tree can either be created from the leaves up to the root or from the root down to the leaves by merging or dividing clusters at each step. In *hierarchical* algorithms, a termination condition has to be defined to indicate when the merge or division process should be terminated.

For multidocument summarization, most previous work [5, 6] has focused on the fact that frequency of words is an important factor. Thus some feature-based learning approaches are proposed to discover salient features by measuring similarity between candidate sentences and summary sentences. Recent summarization studies focused on the discovery of latent topics of documents in extracting summaries [7].

For keyword extraction, there are usually three kinds of approaches: rule-based approaches, statistical approaches and hybrid approaches. In a rule-based approach, keywords are extracted by some well-defined linguistic rules. For example, Houngbo and Mercer [8] used a regular expression to extract some method mention from scientific papers. More complex keywords can be extracted using shallow parsing and dependency analysis between words in the sentence. Statistical approaches are based on statistical information, such as the frequency of keywords appearing in the corpus. Rule-based methods and statistical approaches are often integrated to extract terminology. There are usually two ways to combine them. One is to extract candidate terms with rule-based methods and then if no terms are found then the statistical method is applied. The other is to obtain candidate terms using statistical methods first and then use rule-based methods to discard those terms that are inconsistent with linguistic patterns [8].

3. Our method

Our work aims to provide a convenient way for users to know the diverse research themes of a research topic and further to locate their research interests quickly. To do this, our basic ideas are as follows.

First, for a research topic, there are usually many related papers. However, in our opinion, it is not necessary to provide all of them to users. On the contrary, it would not hinder users from comprehensively understanding the research themes of a research topic if we carefully select some representative papers and then provide them to users. The advantage of this paper selection is obvious: both the performance and the efficiency of most clustering or summarization methods are expected to increase.

Second, there are usually diverse research themes for a research topic. So it is also helpful to users if we can divide the related papers of a research topic into several clusters according to their focused research themes. This has the advantage of helping users to know the diverse research themes for a research topic and then further helping them to locate their research interests quickly.

Third, for each cluster, we can further help users to know its main research themes quickly if we can provide users with several theme keywords about this cluster.

Integrating its above ideas together, for each research topic, we think there is a '*cluster-theme keywords*' structure between users' intentions and related papers. Obviously such a structure can provide users with a rapid way to know the diverse research themes of a research topic, and then allow them to conveniently locate their research interests.

3.1. The general framework

The general workflow of our method is shown in Figure 1. It is a cascade learning scheme that consists of three phases: representative paper selection, paper clustering and theme keyword extraction. Here for a research topic, we take all the scientific papers whose titles or keyword lists contain this research topic as its related scientific papers.

3.2. Representative paper selection

The aim of this phase is to select a small paper collection from the original related paper collection. This selected paper set should have a smaller size but should not have much information loss compared with the original paper set. To do this, we must find a proper metric that can be used to evaluate the importance of a paper to the input research topic. The more important a paper is, the more likely it should be selected as a representative paper.

In our opinion, whether a paper should be selected as a representative paper depends on the following factors:

- Authors – if a paper is written by the authors that come from some stable research groups, this paper is more likely to be selected as a representative paper. Here a stable research group refers to the group that has stable researchers and stable publications in the past few years for a research topic.
- Citation count – it is obvious that the higher the citation count of a paper, especially the citation count from other researchers, the more likely it should be selected as a representative paper.
- Citation extent – if a paper is cited by a large number of different periodicals, it is more likely to be selected as a representative paper.
- Publication year – most of the time, an old paper cannot provide much useful information to users, so the earlier a paper is published, the less likely it is to be selected as a representative paper.

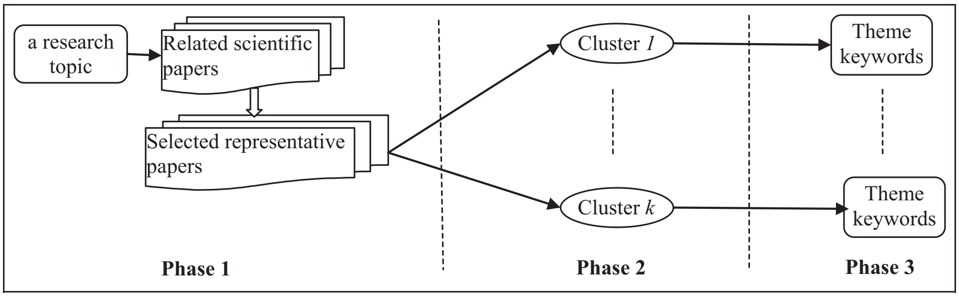


Figure 1. The general framework.

In addition, if we view each paper as a node in a net and view the citation relationships and the co-author relationships among papers as directed edges linking the related papers, all the papers and their linking information will compose a directed graph. Thus the process of representative paper selection is somewhat like the process of relevance web page ranking, whose aim is to evaluate which web pages are more important and should be ranked in the previous positions in the returned web page list. Therefore, those ranking algorithms used in search engines can help us to design our representative paper selection algorithm. Finally, inspired by the famous *PageRank* algorithm [23], we design the following metrics to evaluate the importance of a specific paper A:

$$IP(A) = \alpha G(A) + \beta Y(A) + \lambda E(A) + (1-d) + \frac{d * IP(P_1)}{C(P_1)} + \ldots + \frac{d * IP(P_n)}{C(P_n)} \tag{1}$$

$$G(A) = \frac{|rs(A)|}{\max(|rs(P_i)|)} * \frac{|pub(A)|}{\max(|pub(P_i)|)} \tag{2}$$

$$Y(A) = \frac{1}{1 + (CurYea - PubYea(A))} \tag{3}$$

$$E(A) = \frac{|cp(A)|}{\max(|cp(A_i)|)} \tag{4}$$

In the above equations, $IP(A)$ is the importance value of a paper A; $G(A)$, $Y(A)$ and $E(A)$ are metrics that are used to evaluate the author contribution, the publication year contribution and the citation extent contribution to paper A, respectively. Parameters α, β and λ are used to denote the contributions of $G(A)$,$Y(A)$ and $E(A)$, respectively.

In eqn (1), P_i refers to the ith paper that cites current paper A.[2] $C(P_i)$ refers to the number of total cited papers in P_i. d is a damping coefficient that often belongs to (0, 1). It is obvious that there are two main components in eqn (1). One is the evaluation factor coming from itself. The other is the evaluation factor coming from the papers that cite it.

In eqn (2), $|rs(A)|$refers to the number of researchers in the research group that the authors of paper A come from. $|pub(A)|$refers to the number of papers that are published by the researchers in the group that the authors of paper A come from. The two maximum values in the denominator parts are used as normalization factors. Here we use a simple method to compute $|rs(A)|$and $|pub(A)|$ as shown in Figure 2, in which *RelPapLst* refers to the initial related paper list, and *AutLst*(A) refers to the author list of paper A.

In eqn (3), *CurYea* and *PubYea*(A) refer to the current year and the publication year of paper A respectively. In eqn (4), $|cp(A)|$ refers to the number of periodicals that paper A is cited in. The maximum value in the denominator part is also used as a normalization factor.

To obtain each paper's importance value, we conduct following iterative process as shown in Figure 3. In Figure 3, parameters δ, ε and θ are used to terminate the iteration process.

3.3. Paper clustering

Traditional clustering algorithms such as *k-means* cannot offer robust solution to this paper clustering task directly for the following reasons. First, although all of the selected papers are related to one common research topic, their research themes are diverse. Second, to make one's research work looks more novel, most researchers prefer to use diverse words and diverse description styles in their papers. Both of these two factors will lead to more diverse vocabularies used in these papers. On the other hand, traditional similarity measures used in clustering algorithms often depend on the *tf* values and the *idf* values of some feature words. However, in our paper clustering task, for most of the feature words, both their *tf* values and their *idf* values approximate to the same values for the reasons mentioned above. Thus it is difficult for these similarity-based clustering algorithms to tackle these diversities well.

```
Input: RelPapLst and AutLst(A).
Output: $|rs(A)|$ and $|pub(A)|$
1. Init Step: $|rs(A)| = |AutLst(A)|$, $|pub(A)|=1$;
2. For each paper $p_i$ in RelPapLst Do
       If $AutLst(A) \cap AutLst(p_i) \neq \Phi$ Then $|rs(A)| += |AutLst(p_i)|$;  $|pub(A)| += 1$;
```

Figure 2. The computation of $|rs(A)|$ and $|pub(A)|$.

```
Input: 1. The initial related paper list RelPapLst
       2. Counting parameters $m$ and $n$;
       3. Every paper's citation list
Output: every paper's importance value
1.  For each paper $p_i$ in RelPapLst Do
        $IP_0(p_i)=\alpha G(p_i)+\beta Y(p_i)+\lambda E(p_i)$
2.  For each paper $p_i$ in RelPapLst Do
        $IP_{i+1}(p_i)=\alpha G(p_i)+\beta Y(p_i)+\lambda E(p_i)+(1-d)+\frac{d*IP_i(P_1)}{C(P_1)}+\cdots+\frac{d*IP_i(P_n)}{C(P_n)}$
        If $|IP_{i+1}(p_i)-IP_i(p_i)|\leqslant\delta$ Then $m$++;
        If $m/|RelPapLst|\geqslant\varepsilon$ or $n\geqslant\theta$ Then Break
        Else $n$++;
3.  Return
```

Figure 3. The algorithm for computing papers' importance values.

From the above analysis we can see that the diversity issue must be well tackled so as to obtain a good paper clustering performance. To tackle this issue, two strategies are taken here.

The first strategy is to reduce the length of each paper. In our opinion, it is unnecessary to use the whole paper text during paper clustering. Instead, we remove the *content* part of a paper during paper clustering. This will not lead to worse performance for paper clustering for the following two reasons. First, the *abstract* part in a paper has summarized the whole paper well. Second, the *content* part is usually too *specialized*, which is the main reason for the diversity problem. On the other hand, it is also worth noting that, in the *abstract* part, not all sentences make the same contribution to paper clustering. Those sentences that express the methods or techniques described in these papers will be more important. Therefore we use a rule-based method as reported by Houngbo and Mercer [8] to extract these method sentences. The words in these sentences will be assigned higher weights during paper clustering. The rule used can be represented by following regular expression:

$$(\text{Adjective}|\text{Noun})+(\text{method}|\text{analysis}|\text{algorithm}|\text{approach}|\text{model}) \quad (5)$$

Here it is worth noting that the above regular expression is originally designed for processing English, and when it is used for processing other language data, one can simply translate the words in the second part of the regular expression into corresponding language words. Taking Chinese as example, the above regular expression would be changed to: '(*Adjective* | *Noun*) + (*方法* | *分析* | *算法* | *技术* | *模型*)', in which '*方法*' is the translation of 'method', '*分析*' is the translation of '*analysis*', and so on. Obviously, a part of speech tagging tool is necessary here.

The second strategy is to introduce a domain ontology to alleviate the word diversity problem in paper clustering. We have pointed out that many researchers prefer to use different words to describe their methods. So it is difficult to obtain a good paper clustering result if only the words in papers are used. On the other hand, there are rich concepts and relations in a domain ontology, and these relations link different concepts in a semantic fashion. So it is natural to think that a paper's original vocabulary can be extended with a domain ontology. With this extension, more related words will be added into the original vocabulary of a paper, and the word diversity problem will be alleviated accordingly. Although the *keywords* in a paper can express the rough research theme of the paper, some of these keywords are somewhat too *generalized* so that many of them cannot play much of a role during paper clustering. Thus introducing a domain ontology is necessary. Fortunately, Ren [24, 25] has reported a domain ontology construction method that takes academic papers as the data source. Obviously the domain ontology constructed by his work is suitable for our task. During extension, for each feature word, if it can be found in the introduced domain ontology, some of its adjacent *upper* layer concepts and adjacent *lower* layer concepts will be added into the feature word collection. Figure 4 is an example of this word extension, in which each circle denotes a concept in domain ontology, and each directed edge denotes a *hierarchical* relation.

Integrating above two strategies, we use a *k-means* clustering algorithm to conduct the paper clustering task. In this clustering algorithm, the used similarity computation method is shown in eqn (6). In this equation, the co-author factor is also taken into consideration along with the common cosine similarity, because, for a researcher, their research work is

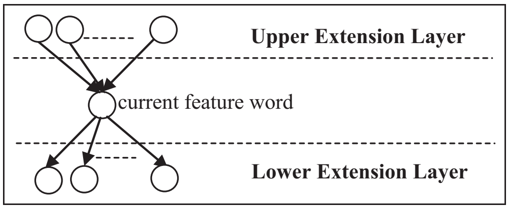


Figure 4. Feature extension based on domain ontology.

```
Input: a text fragment's word list $L=w_1w_2,...,w_n$
Output: the extended feature word list $L'$
For $i=1$ to $n-2$ Do
  If $w_iw_{i+1}$ can be found in the introduced domain ontology Then
      1. Take ($w_iw_{i+1}$) as a center concept
      2. Select the top-3 of its upper layer concepts and the top-3 of its lower layer concepts
      3. Add these selected concepts into the feature word list
  Else If $w_iw_{i+1}w_{i+2}$ can be found in the introduced domain ontology Then
      1. Take ($w_iw_{i+1}$ $w_{i+2}$) as a center concept
      2. Select the top-3 of its upper layer concepts and the top-3 of its lower layer concepts
      3. Add these selected concepts into the feature word list
Return
```

Figure 5. The maximum forward matching feature word extension algorithm.

continuous and stable. Therefore, papers written by the same authors are more likely to be in the same cluster and their similarity should be higher accordingly. Parameter k_1 is used to leverage the contributions of these two factors:

$$\mathrm{sim}(p_i, p_j) = k_1 \cos(p_i, p_j) + (1 - k_1)\delta(p_i, p_j) \tag{6}$$

$$\cos(p_i, p_j) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \times \sqrt{\sum_{k=1}^{n} w_{jk}^2}} \tag{7}$$

$$\delta(p_i, p_j) = \begin{cases} 1 & \text{if } p_i \text{ and } p_j \text{ have common authors} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

In addition, pre-processing on feature selection and feature weight assignment is conducted before paper clustering. We briefly introduce this pre-processing as follows.

First, for each paper, we combine its *title, abstract* and *keyword list* as a new text fragment for clustering. Then we do part of speech tagging for this text fragment, after which all of the *auxiliary words, punctuations, numbers, prepositions* and *pronouns* are removed and the remaining words are taken as feature words.

Second, for each feature word that can be found in the introduced domain ontology directly, we select the *top-3* of its *upper* layer concepts and the *top-3* of its *lower* layer concepts according to their co-occurrences with current feature word and add them into the feature list. In addition, as some concepts in the introduced domain ontology are multiword phrases, we use a *maximum forward matching* algorithm to extend such kinds of feature words when they cannot be found in the introduced domain ontology directly. This algorithm is described in Figure 5, in which the maximum length of a phrase is empirically set as 3. The input of this algorithm is a word list of the text fragment mentioned previously.

Third, we assign a basic *tf*idf* weight for each feature word w_i. Then this basic weight will be adjusted according to following two rules:

- If w_i is in the method sentences, its final weight will be adjusted to *tf*idf* + δ_1.
- If w_i is extended from the domain ontology, its final weight will be adjusted to *tf*idf**δ_2, where δ_2 usually belongs to (0, 1). Here δ_2 is used to denote our trust degree to the introduced domain ontology.

```
Input: 1. a test topic set $T=\{t_1, t_2, ..., t_n\}$
       2. learning rate $\eta$
Output: weight vector $w$ for eqn (1)
For each test topic $t_i$ Do
    Search $t_i$'s related scientific papers list $lst(t_i)$
    Set initial weight vector $w_0$
    Rank $lst(t_i)$ with current vector $w_i$
    Output the top-k papers to $m$ human experts
    For each paper $p_i$ in these top-k results Do
        If the majority of experts make a position adjustment: $p_j \leftarrow p_i$, Then $w_i \leftarrow w_i + \eta(x_j - x_i)$
$w \leftarrow \frac{1}{|T|}\sum_{i=1}^{|T|} w_i$
```

Figure 6. Weight tuning method for eqn (1).

In addition, when all the papers have been clustered, we can further rank the papers in each cluster according to their importance values. This is also a way to help users quickly locate their research interests.

3.4. Theme keyword extraction

The aim of this phase is to extract some theme keywords from each cluster so that users can rapidly catch the main research theme of a cluster. Here we use a very simple solution for this task: for each cluster, taking all of its paper's keywords together as the theme keywords of this cluster. In addition, we also noticed that there are usually more expressive theme keywords in the method sentences. Therefore it is necessary to extract such kinds of theme keywords. In our method, when the method sentences are extracted with eqn (5), all of the words in the regular expression will be removed and the remained words are taken as part of the needed theme words. Of course, we use a *df*-based metric to make the number of each cluster's theme keywords within a reasonable scale.

4. Parameter tuning

There are four parameters used in eqn (1) that need to be tuned. However, this tuning task has the following difficulties. First of all, it is impossible for us to manually label a reasonable score for a paper to represent its real importance value. That is to say, we cannot construct a proper training set for traditional parameter tuning methods like the least-squares method or the perceptron-based weight tuning method. Second, we also cannot convert this task into a classification task. This is because these importance values will be used as a ranking metric for the paper selection procedure other than some simple '*select/not select*' labels.

To overcome these difficulties we designed an expert feedback-based weight tuning method, which is shown in Figure 6. In our method, it combines the relevance feedback technology that is often used in the *information retrieval* field and the perceptron learning strategy. During this weight-tuning procedure, every paper can be easily denoted as a four-dimensional vector according to eqn (1). Thus the vector x_i in Figure 6 denotes a paper's vector. The weight vector (α, β, λ, d) is denoted w. In Figure 6, the learning rate η is a step parameter and here we set it to 0.01. The number of involved human experts is 3, that is to say, m is set to 3 here.

There are also other three parameters that need to be tuned: the parameter k_1 in eqn (6) and the parameters δ_1 and δ_2 for weight adjustment during paper clustering. For them, a simple 10-fold cross-validation method is used to select the best parameter values.

5. Experiments

5.1. Basic experimental settings

(1) *Data source* – we used Chinese academic papers as our dataset because there is a very easy way to obtain these data: almost all of Chinese published academic papers can be downloaded from a website (http://www.cnki.net/).

Table 1. Brief statistics on the returned related papers for test topics.

Maximum related paper number	Minimum related paper number	Total related paper number	Average related paper number
2789	423	433,517	867

From this website, we download more than 400,000 academic papers in the information technology domain, spanning the past 30 years (1979–2010). These data are taken both as the data source to construct a domain ontology and as the data source to evaluate our method.

(2) *Test set* – we manually selected 50 keywords in the information technology domain as test research topics to evaluate the proposed method. For these test topics, the total number of related papers is 433,517. It should be noted that there is a big difference between the number of related papers for each of these topics; brief statistics about their related papers are shown in Table 1.

(3) *Dev set* – we constructed a dev set for parameter tuning. Specifically, we select another 50 keywords in the information technology domain for tuning the parameters in eqn (1). For these dev topics, the total number of related papers is 475,219. For these topics in dev set, they have similar related papers statistics to the test set.

5.2. Evaluation of paper clustering

The paper clustering procedure is a key component in our whole system and we evaluated its performance first. Specifically, when evaluating every clustering method, three human experts were asked to label the correct clustering result separately for each test topic. The average *MacroF* score was used as an evaluation metric and the experimental results are shown in Table 2. In these experiments, the *k-means* clustering algorithm that does not use paper selection was taken as the baseline method. In all the experiments, the parameter *k* in *k-means* clustering was taken as 10.

It can be observed from Table 2 that the baseline method obtained a very poor clustering performance. The main reason for this has already been analysed. It can also be observed from Table 2 that the paper selection procedure plays an important role in improving the performance of paper clustering. The paper selection is somewhat similar to a data cleaning procedure. Many of the low-quality papers are removed after it. It is expected that the themes of the remaining papers will be more concentrated, which is obviously helpful for the *k-means* clustering because the parameter *k* is always limited to a small value.

Of course, the paper selection procedure may also bring the risk of losing some valuable papers. To assess the effect of this procedure on users' information seeking, we conducted another experiment, whose process is shown in Figure 7.

In this experiment, we used the same test topic set used in the clustering evaluation. Five human experts were used to find the interesting papers in a returned paper list. The experimental results are shown in Figure 8.

It can be observed from Figure 8 that there is a large information loss with a small number of selected papers. However, the coverage ratio usually tends to be stable when we select more than the *top-400* papers. Thus we selected the *top-400* papers for every test topic in most of our experiments.

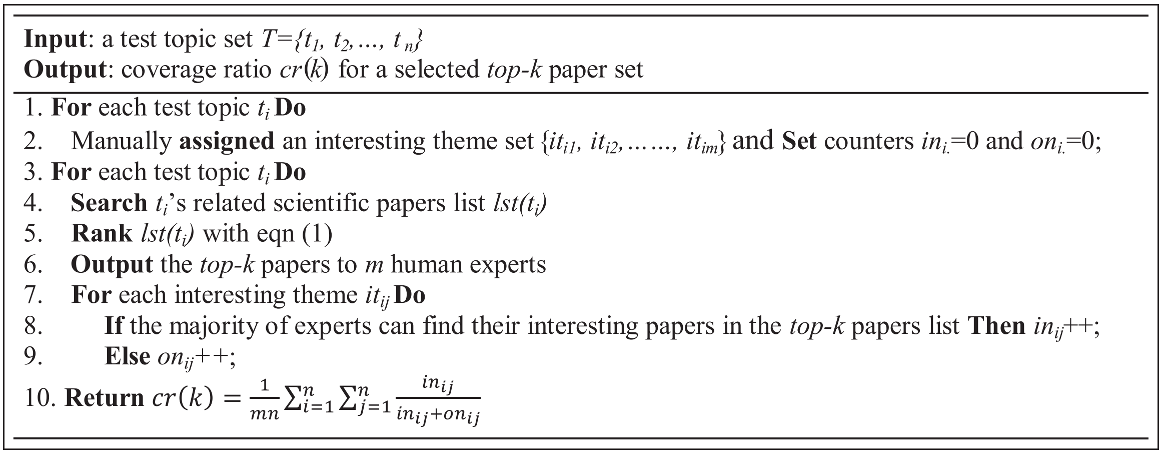
It can be observed from Table 2 that the domain ontology extension method is also useful for improving the performance of paper clustering. The reason for this is that the papers' similarity distribution has been changed when the ontology extension method is used. We can use the test topic '*machine translation*' as an example to demonstrate this. For this test topic, we selected *top-200* representative papers from its related papers. We computed the similarities for every two selected papers and showed the different distributions of these similarities *with/without* domain ontology extension in Figure 9, in which the *x-axis* represents the similarity zones and the *y-axis* represents the number of papers.

It can be observed from Figure 9 that, when the domain ontology extension method is not used, most of the similarities are concentrated in a relatively narrow band. Thus it is very difficult for the original clustering method to distinguish them properly. On the other hand, when the ontology extension strategy is added, the original similarity distribution is changed in two aspects. First, the distribution becomes less steep, which means that the similarity values are distributed in a broader range. Second, the similarity curve shift towards the right, which means there is an obvious improvement in distinguishing papers' similarities. Obviously, both of these changes make the papers easier to cluster compared with the direct clustering method.

We also noticed that the performance improvement brought by domain ontology extension is less than the performance improvement brought by paper selection. Part of the reason for this is that the used domain ontology itself has many errors. These errors will prevent the domain ontology reaching its full potential.

Table 2. Paper clustering results.

Method	Average *MacroF*
Baseline	0.34
Baseline + paper selection	0.57
Baseline + ontology extension	0.49
Baseline + weight adjustment	0.36
Baseline + paper selection + ontology extension + weight adjustment	0.65



```
Input: a test topic set $T=\{t_1, t_2, \ldots, t_n\}$
Output: coverage ratio $cr(k)$ for a selected top-k paper set
For each test topic $t_i$ Do
    Manually assigned an interesting theme set $\{it_{i1}, it_{i2}, \ldots\ldots, it_{im}\}$ and Set counters $in_{i.}=0$ and $on_{i.}=0$;
For each test topic $t_i$ Do
    Search $t_i$'s related scientific papers list $lst(t_i)$
    Rank $lst(t_i)$ with eqn (1)
    Output the top-k papers to $m$ human experts
    For each interesting theme $it_{ij}$ Do
        If the majority of experts can find their interesting papers in the top-k papers list Then $in_{ij}$++;
        Else $on_{ij}$++;
Return $cr(k) = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{in_{ij}}{in_{ij}+on_{ij}}$
```

Figure 7. Assessment method for paper selection.

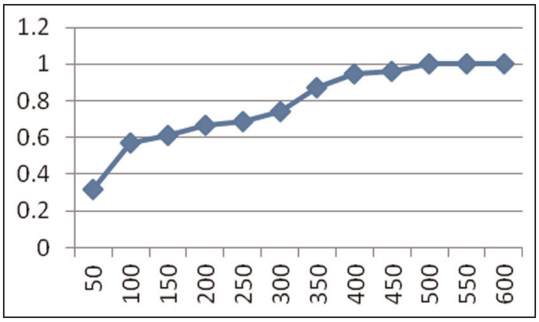


Figure 8. Coverage ratio with different *top-k* in paper selection.

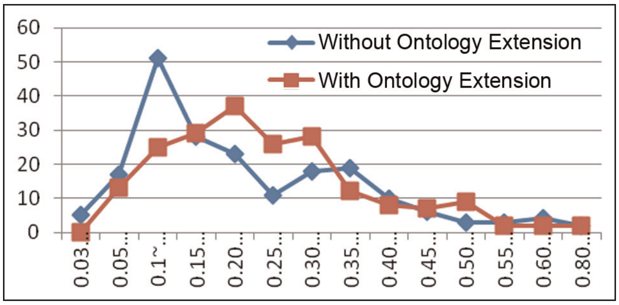


Figure 9. Similarity distribution comparison for 'machine translation'.

Table 3. Comparison of different feature selection methods and feature numbers.

		Feature number				
		500	700	900	1200	1400
Feature Selection Methods	DF	0.630	0.646	0.629	0.624	0.629
	TC	0.618	0.681	0.666	0.632	0.590
	TS	0.639	0.665	0.633	0.632	0.593

Table 4. Results of the first aspect.

	Level 1	Level 2	Level 3	Level 4
Our method	18%	36%	30%	16%

In addition to the above experiments, we also compared the performances of paper clustering when using different feature selection methods with different feature numbers. The experimental results are shown in Table 3. The used feature selection methods are listed as follows:

- DF – document frequency;
- TC – term contribution [10];
- TS – term strength [18];

In this experiment, all of the improvement methods were used, including *paper selection, ontology extension* and *weight adjustment*. It can be observed from Table 3 that both the feature selection methods and the feature numbers affect the final clustering performance. The *term contribution* (*TC*) method obtains the best clustering performance when 700 features are used.

5.3. Effectiveness of the whole method

We evaluated the effectiveness of the proposed method from two aspects. The first aspect is whether our method can provide enough information for users to quickly catch the diverse themes of a research topic. The second aspect is to determine to what extent our method can save users' time in locating their research interests.

To evaluate the first aspect, we asked five human experts to check the output results for each test topic, the '*cluster-theme keywords*' pairs, to see whether these results clearly and correctly addressed the main research themes of this test topic. Five human evaluation metrics are used, which are defined as follows:

- *Level 1* – one can completely catch the diverse research themes with the returned results.
- *Level 2* – one can catch most of the diverse research themes with the returned results.
- *Level 3* – one can catch part of the diverse research themes with the returned results.
- *Level 4* – one cannot catch any of the diverse research themes with the returned results.

To evaluate the second aspect, we first manually assigned an interested paper set for each test topic. Then five human experts were asked to locate these interested papers based on the '*cluster-theme keywords*' pairs for this test topic. The average time for locating these papers was used as the evaluation metric. The experimental results are illustrated in Tables 4 and 5, respectively.

It can observed from these tables that our method is not only effective for helping users to quickly catch the main research themes of a topic, but also for helping them further to locate their research interests. In fact, our method is tolerant of clustering errors for the following reason: even if there are some clustering errors, users can still quickly locate their research interests by checking only several clusters with the help of theme keywords other than the whole related paper list.

We also conducted other experiments to evaluate the contributions of our method's different components to helping users to quickly catch the diverse research themes for a research topic and further to quickly locate their research

Table 5. Results of the second aspect.

	Average time
Our method	~3 min

Table 6. Contribution comparisons for the first aspect.

Method	Level 1	Level 2	Level 3	Level 4
Baseline	6%	30%	40%	24%
Baseline + paper selection	14%	32%	36%	18%
Baseline + paper selection + keyword	14%	32%	36%	18%
Baseline + paper selection + methodword	16%	30%	36%	18%
Baseline + paper selection + methodword + keyword	18%	36%	30%	16%

Table 7. Contribution comparisons for the second aspect.

Method	Average time
Baseline	~9 min
Baseline + paper selection	~5 min
Baseline + paper selection + keyword	~5 min
Baseline + paper selection + methodword	~4 min
Baseline + paper selection + methodword + keyword	~3 min

interests. In these experiments, the direct *k-means* clustering method was taken as the baseline method. The experimental results are shown in Tables 6 and 7, in which the *methodword* and *keyword* refer to the method of extracting theme keywords from method sentences and the method of extracting theme keyword from papers' keyword lists.

It can be observed from these tables that the theme keyword extraction procedure is an important supplement to the paper clustering procedure, which can further help users to quickly locate their research interests.

6. Discussion

In this paper, we focus our method on the application scenario of scientific papers. In fact, the proposed method is not confined to scientific papers. On the contrary, it can be widely used in other scenarios. For example, in market surveys, product reviews and patent prior search, and so on What is different is that, in these application scenarios, there may be no available keywords that can be directly used as the theme keywords. Thus an extra theme keyword extraction procedure should be well designed. As for the needed domain ontology, as long as one can collect large amounts of data in the application scenario, the needed domain ontology can be easily constructed by the method proposed by Ren [24, 25]. Thus we can say that the proposed method provides a general framework that can be widely used in such application scenarios: users want to quickly catch the diverse aspects of a given topic, and then quickly locate their interesting aspects.

In addition, our method is also language-independent. That is to say one can easily transplant the proposed method to similar application scenarios in another language.

Some researchers may think that a '*cluster-summarization-theme keywords*' structure or a '*cluster-summarization*' structure may also be an alternative structure to reach our goals. In our opinion, there are no essential differences between these alternative structures and our structure. For these alternative structures, the '*summarization*' component is a beneficial supplement for the '*theme keywords*' part, but cannot substitute for it. This is because, first, the size of a '*summarization*' would be larger than a 'theme keywords' set, and second, the quality of '*summarization*' is not so good. In other words, the used '*cluster-theme keywords*' is a concise and effective structure for our application.

7. Conclusions

In this paper, we propose an unsupervised learning scheme to extract a '*cluster-theme keywords*' structure from the related scientific papers of a research topic. Our method aims to help users know the diverse research themes of a research topic quickly and further help them to rapidly locate their research interests. A series of measures were taken to accomplish this goal, including representative paper selection, paper clustering and theme keyword extraction. In addition, a domain ontology was introduced to alleviated the word diversity issue during paper clustering. The experimental results demonstrate its effectiveness: compared with the baseline method, our method can save more time for users in determining the diverse research themes of a research topic and locating their research interests.

Funding

This paper was supported by the 'Fundamental Research Funds for the Central Universities' and the National Natural Science Foundation of China (grant no. 61003159, 61100089, 61073140 and 61272376).

Notes

1. One can test this conclusion from the website http://www.cnki.net/ by taking '统计机器翻译' (statistical machine translation) as input research topic.
2. One can use the website http://www.cnki.net/ for Chinese papers' citations and use the website http://scholar.google.com/ for English papers' citations.

References

[1] Owczarzak K and Dang HT. Overview of the TAC 2011 summarization track: Guided task and AESOP task. In: *Proceedings of the fourth text analysis conference*, Gaithersburg, MD, 2011, http://clef2012.org/resources/slides/clef2012.tac.pdf

[2] Johnson SC. Hierarchical clustering schemes. *Psychometrika* 1967; 32(3): 241–254.

[3] Ester M, Kriegel HP, Sander J and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of 2nd international conference on knowledge discovery and data mining*, 1996, pp. 226–231.

[4] Hu X, Sun N, Zhang C and Chua TS. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: *Proceedings of the 18th ACM conference on information and knowledge management*, 2009, pp. 919–928.

[5] Nenkova A and Vanderwende L. The impact of frequency on summarization. Technical Report MSR-TR-2005-101. *Microsoft Research, Redwood, WA*, 2005.

[6] Conroy JM, Schlesinger JD and O'Leary DP. Topic-focused multi-document summarization using an approximate oracle score. In: *Proceedings of ACL2006*, Stroudsburg, PA, 2006, pp. 152–159.

[7] Haghighi A and Vanderwende L. Exploring content models for multi-document summarization. In: *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 362–370.

[8] Houngbo H and Mercer RE. Method mention extraction from scientific research papers. In: *Proceedings of COLING* 2012, 2012, pp. 1211–1222.

[9] Lin H and Bilmes J. A class of sub modular functions for document summarization. In: *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies*, 2011, pp. 510–520.

[10] Liu T, Liu S and Chen Z. An evaluation on feature selection for text clustering. In: *Proceedings of the 20th international conference on machine learning*, 2003, pp. 488–495.

[11] Abu-Jbara A and Radev D. Coherent citation-based summarization of scientific papers. In: *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies*, 2011, pp. 500–509.

[12] Wang D and Liu Y. A pilot study of opinion summarization in conversations. In: *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies*, 2011, pp. 331–339.

[13] Celikyilmaz A and Hakkani-Tur D. A hybrid hierarchical model for multi-document summarization. In: *Proceedings of the 48th annual meeting of Association for Computational Linguistics*, 2010, pp. 815–824.

[14] Genest PE and Lapalme G. Fully abstractive approach to guided summarization. In: *Proceedings of the 50th annual meeting of the Association for Computational Linguistics: Short paper*, 2012, pp. 354–358.

[15] Qazvinian V and Radev DR. Identifying non-explicit citing sentences for citation-based summarization. In: *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, 2010, pp. 555–564.

[16] Frantzi K, Ananiadou S and Mima H. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries* 2000; 3(2): 115–130.

[17] Wilbur WJ and Sirotkin K. The automatic identification of stop words. *Journal of Information Science* 1992; 18(1): 45–55.

[18] Kageura K and Umino B. Methods of automatic term recognition: A review. *Terminology* 1996; 3(2): 259–289.

[19] Patry A and Langlais P. Corpus-based terminology extraction. In: *Proceedings of the 7th international conference on terminology and knowledge engineering*, 2005, pp. 313–321.

[20] Vivaldi J, Màrquez L and Rodríguez H. Improving term extraction by system combination using boosting. In: *Proceedings of the 12th European conference on machine learning*, 2001, pp. 515–526.

[21] Vivaldi J and Rodriguez H. Improving term extraction by combining different techniques. *Terminology* 2001; 7(1): 31–48.

[22] Zhang C, Wang H, Liu Y, Wu D, Liao Y and Wang B. Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems* 2008; 4(3): 1169–1180.

[23] Larry P. PageRank: Bringing Order to the Web. *Stanford Digital Library Project, talk*, 18 August 1997.

[24] Ren FL. A demo for constructing domain ontology from academic papers. In: *Proceedings of COLING 2012: Demonstration papers*, 2012, pp. 369–376.

[25] Ren FL. A cheap domain ontology construction method based on graph generation and conversion method. *Journal of Information and Computational Science* 2012; 9(18): 5823–5830.