Research on frontier topic detection based on probability outbreak and correlation analysis

Xu Min
Department of Information Management
Peking University, Beijing 100871, China
kingfoxing@pku.edu.cn

Supervisor: Li Guangjian
Department of Information Management
Peking University, Beijing 100871, China
ligj@pku.edu.cn

ABSTRACT: Aiming at the academic literature of no keyword or no citation network, we put forward a new method about frontier topic detection and topic representation. We use LDA to generate topics, calculate the probability of the annual theme, detect the frontier topic by the probability outbreak and use association rule mining for topic representation. In order to verify our method, we pick up the computer science's arXiv data (80020 abstracts) from 2010 to 2015 as the experimental data set to find every year's frontier topic.

KEYWORDS: Frontier Topic Detection; Subject Representation; LDA; Association Rules Mining

INTRODUCTION

In the rapid development of science and technology, the number of related documents appeared to explode. It has become increasingly difficult to rely on manual identification of frontier topics in the massive and multi-source scientific and technological data resources. The studying of frontier topics detection which has been paid more and more attention can help people understand the trend of knowledge development in related disciplines, and provide reference for researchers or research organizations to determine the studying path and integrate into the frontier studying.
LITERATURE REVIEW

The earliest relevant research "the author cited the scientific structure of the literature measurement method" is co-published by White and Griffith in 1981. In the paper, they put forward the "Author Co-citation Analysis" (ACA) method and used it to analyze the co-citation of 39 authors in the relevant research area to divide the branches and find structures of the disciplinary. After successful application of such metrological methods, researchers have further explored the new frontier topics detection methods based on content analysis, such as word frequency analysis, co-word analysis and clustering analysis, which have been the mainstream. Gerard Salton proposed a vector space model (VSM) in the 1970s. This model was quickly applied to the field of topics analysis, which result in the topics analysis method based on simple vector space. In 1990, Scott Deerwester and Susan T. Dumais proposed the latent semantic analysis model (LSA), which based on the vector space model and regard the mathematical statistics and matrix operations as key to greatly enhance the effect of subject analysis. Thomas Hoffmann published a paper titled "Probabilistic Latent Semantic Indexing" in 1999. In this paper, a probabilistic latent semantic analysis model is proposed, which is based on the LSA model and adds the probability calculation. It can solve the problem of dynamic data calculation in the latent semantic analysis model. In 2003, Blei et al. proposed a latent Dirichlet distribution model (LDA), and introduced the Dirichlet prior distribution on the basis of PLSA to better analyze the articles outside the training corpus. By 2008, the LDA almost completely replaced the PLSA, has become one of the best methods about text mining and topic probability calculation.

In the existing research, the detection of frontier topics need to use the citation network or keywords, which have the following problems: In the first place not every document has citation data and keywords, as a result, the types of documents that can be analyzed are limited; Secondly, In the citation network, the academic literatures with higher citation rate tend to be those "old" documents, and the new academic literatures’ citation rate are often have not high enough, which cannot guarantee the novelty of the frontier topic. Thirdly, most of the keywords are written by the authors themselves, so it takes a lot of time and effort to normalize its synonyms before the analysis. Finally, the interpretation of the detected frontier topics requires a thorough domain knowledge and may require extensive reading of the relevant literature. So all of these factors will affect the efficiency of the analytical process and the quality of the analytical results.

In this paper, we put forward a frontier topics detection method based on probabilistic burst and association rules mining, which can analyze the no keywords or no citation documents, such as preprint and Internet text. In addition, this method provides as much as possible the results of the processing that can be visually recognized by humans, to minimize the reader’s time cost and reduce the reader's background knowledge requirements.
ANALYSIS FRAMEWORK AND KEY TECHNOLOGIES

In the proposed method, the title, abstract, full text of the document are analytic data, LDA topic model and association rules mining are the analysis techniques. In addition to document clustering and topic generating for large-scale document sets at the semantic level, this method can also calculate the topic's probability change and display topic content. The analysis framework is shown in Fig 1.

The method proposed in this paper includes four stages: preprocessing, topics generation, probability calculation and topic representation.

Preprocessing stage: (1) The whole document set is divided according to a certain time span (such as year, month, day, etc.), and then extract the titles, abstracts or full texts of each time span’s documents. (2) All documents' abstracts or full-texts should be proceeded word-segmentation, stop-words removing and normalization, whose purpose is to remove punctuation and special characters and to normalize the words. I suggested that we can use the WordNet-like dictionary tools to combine the words based on semantic relationship. For example, "ran" and “run” have the same etyma and actually are the same word, while “confidently” and “confidential” also have the same etyma but different meaning.

Topics generation stage: (1) All documents are regarded as data sets that are trained by LDA model. Then we set the document-clusters' number to generate topics and to build a whole training model of "document-topic-probability". (2) The probability of each
document’s related topics in the document set can be calculated by previous established whole training model in different time span. In other words, we can find any document that has the relationship with which topics and get the probability value. But the detection of frontier topic should be focus on topic not document, so we need a conversion from “document-topics-probability” to “topic-documents-probability” in the next stage.

Probability calculation stage: (1) we use a data structure, dictionary, to achieve the transformation of the object from the document to the topic, which is to find out each topic related to which documents and their probability. (2) We calculate the average probability of each topic by probability sum dividing the number of relevant documents. The purpose is to avoid the interfering about chat topics, that is, some topics related to a large number of documents, the relevance is very low. (3) We sort the topics by average probability and give the corresponding weight value, and then calculate the difference of the topics’ weight in the two adjacent time span. In the process of comparing the difference, we define the frontier topic whose weight value of probability burst. Kleinbeng J[2] did reckon that outbreak-words can reveal the emerging trend of academic research, so I also deem that the frontier topics are not high-probability but probability outbreak. The burst of probability indicates that suddenly there are many people interested in this research over a period of time.

Take my experiment’s topic 0 to 4 as example (shown in table 1), it can be found that, in the adjacent year, most of the topics’ weight have little change, but a small number of topics' weight will have a sudden rise. For instance, the weight of topic 2 has risen from 17 to 39 in 2010-2011. The weight of topic 4 has increased from 3 to 19 in 2010-2011 and from 2 to 11 in 2013-2014. The anomaly of probability changing reflects the researchers’ attention to these topics during that time.

Table 1: The Distribution of Topic Weights

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0#</td>
<td>36</td>
<td>41</td>
<td>43</td>
<td>39</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>1#</td>
<td>30</td>
<td>33</td>
<td>39</td>
<td>35</td>
<td>41</td>
<td>42</td>
</tr>
<tr>
<td>2#</td>
<td>25</td>
<td>28</td>
<td>35</td>
<td>28</td>
<td>39</td>
<td>17</td>
</tr>
<tr>
<td>3#</td>
<td>66</td>
<td>59</td>
<td>59</td>
<td>58</td>
<td>56</td>
<td>57</td>
</tr>
<tr>
<td>4#</td>
<td>6</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>19</td>
<td>3</td>
</tr>
</tbody>
</table>

Topic representation stage: (1) Select topic that need to be displayed and extract the titles of all the documents related to the topic, which generate a titles-set. (2) The association rules mining method is used to calculate the frequent item sets of words in the titles-set. (3) Set the threshold to control the size of the frequent item sets and find the most frequently used and most frequently occurring words or phrases in the titles-set. (4) We calculate the word frequency of all the words in the frequent item sets, and then calculate the similarity between the words of the highest frequency and the words of the frequent items. The most similar
result is the most frequent topic phrases in the title, and they are the description of the topic. Because the title is the refining of the main content of a document, so we use association rules mining to analyze the titles-set to find the most frequently occurring and frequently used phrases which are often what most researchers are studying.

Take my experiment’s topic 92 in 2015 example (shown in table 2), through LDA model, it can be obtained the following result:

\{star(0.042), speed(0.039), skin(0.030), liquid(0.024), irregularity(0.023), locality(0.018), id (0.017), mail(0.014), cc(0.014), iris(0.014)\}.

From this, we can only know some words and their occurrence probability. Even if I am an expert in this area, I cannot understand what the topic is talking about. To understand the true meaning, I have to read some papers in the topic-related documents-set. But our method can find the frequent words set related with topic. As is shown in table 2, we obtain more new and accurate words, like wireless, performance, channel. So we have a profile about this topic, maybe the wireless network research.

Table 2: The maximum frequent words set of the topic 92 in 2015

<table>
<thead>
<tr>
<th>Multiword in frequent words set</th>
<th>Twice word in frequent words set</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized Double Gamma Performance</td>
<td>Networks Parallel</td>
<td>Wireless</td>
</tr>
<tr>
<td>Wireless Power Transfer Simultaneous</td>
<td>Generalized Fading</td>
<td>OFDM</td>
</tr>
<tr>
<td>Generalized Double FSO Channels Performance</td>
<td>Echo Networks</td>
<td>Algorithmic</td>
</tr>
<tr>
<td>Multilevel Coding Diversity</td>
<td>Communications Performance</td>
<td>modeling</td>
</tr>
<tr>
<td>Channels Performance Generalized FSO</td>
<td>Operating Performance</td>
<td>Uplink</td>
</tr>
<tr>
<td>Channel Codes Erasure</td>
<td>Channels Structured</td>
<td>Qubit</td>
</tr>
<tr>
<td>Channels Performance FSO</td>
<td>communities networks</td>
<td>Wiretapped</td>
</tr>
<tr>
<td>Generalized Double Gamma Fading Performance</td>
<td>Color Performance</td>
<td>Discrete</td>
</tr>
<tr>
<td>Gamma FSO Channels Performance</td>
<td>Channels Codes</td>
<td>Efficient</td>
</tr>
<tr>
<td>System Applications Performance</td>
<td>Channels Capacity</td>
<td>smartphone</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
<td>......</td>
</tr>
</tbody>
</table>

Then we compute the Euclidean distance between the high-frequency words and phrases in maximum frequent words set. The smaller the value of Euclidean distance, the higher the degree of similarity. So we find that the topic 92 in 2015 is actually generalized FSO (free-space optics) channels performance (shown in table 3).

Table 3: The representation of topic

<table>
<thead>
<tr>
<th>Euclidean distance(similarity)</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Generalized FSO Channels Performance</td>
</tr>
<tr>
<td>1.4142135623730951</td>
<td>FSO Channels Performance</td>
</tr>
<tr>
<td>1.4142135623730951</td>
<td>Generalized Double FSO Channels Performance</td>
</tr>
</tbody>
</table>
EXPERIMENTAL RESULTS AND ANALYSIS

Experimental data

Compared with the traditional academic resources, the preprint has the advantage of rapid publication and high reliability, and it can publish the most novel research results in the shortest time. So I take the arXiv’s CS data (80020 abstracts about computer science from 2010 to 2015) as a sample for detecting the frontier topic and verifying it.

As is shown in Fig 2, we have found that papers published in arXiv in the field of computer science are growing year by year, with the number of papers increasing fourfold over the six-year period from 2010 to 2015 and nearly 60 percent by 2015 compared to 2014. This amount of data rapid growth, indicating that more and more people are willing to publish the newest scientific ideas, which can be used as the best data for frontier topic detection.

In this experiment, we select the year, title, and abstract of arXiv’s computer science papers as the analysis attribution, take the space for word segmentation and use the WordNet dictionary’s Lemmatizer tool to normalize the words. The parameters of LDA model are set to the alpha =0.1, beta =0.2, the number of iterations =1000, the number of topics =100; The FP-Growth algorithm is used to calculate the frequent word set, and the threshold range of the support degree S and the confidence level C is between 17 and 33.

Experimental result analysis
In this study, we found out the frontier topic by calculating the change of probability. We selected the top three topics in the highest probability from 2010 to 2015, and showed the research contents. The results are shown in Table 4.

**Table 4: Top 3 frontier topic from 2010 to 2015**

<table>
<thead>
<tr>
<th>Year</th>
<th>Frontier Topic 1</th>
<th>Frontier Topic 2</th>
<th>Frontier Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-2015</td>
<td>Convolutional Neural Networks</td>
<td>Active Networks</td>
<td>Generalized FSO Channels Performance</td>
</tr>
<tr>
<td></td>
<td>Deep Neural Networks</td>
<td>Kernel Learning</td>
<td>FSO Channels Performance</td>
</tr>
<tr>
<td>2013-2014</td>
<td>Cellular Radio Networks</td>
<td>Wireless Sensor Clustering</td>
<td>Distributed Power Control</td>
</tr>
<tr>
<td></td>
<td>Cellular Networks MIMO</td>
<td>Efficient Sensor Wireless</td>
<td>Power Control Networks</td>
</tr>
<tr>
<td>2012-2013</td>
<td>Distributed Storage Systems</td>
<td>Energy Wireless Sensor Networks</td>
<td>Interference Channel Freedom Degrees</td>
</tr>
<tr>
<td></td>
<td>Distributed Systems</td>
<td>Adaptive Wireless Sensor Networks</td>
<td>Degrees CSIT</td>
</tr>
<tr>
<td>2011-2012</td>
<td>Bayesian Learning Networks</td>
<td>Loopy Belief Propagation</td>
<td>Social networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Online social networks</td>
</tr>
<tr>
<td>2010-2011</td>
<td>MIMO Freedom Channels</td>
<td>Wireless Routing Protocols</td>
<td>Distributed Memory</td>
</tr>
<tr>
<td></td>
<td>Degrees CSIT Broadcast</td>
<td>Wireless Sensor Networks</td>
<td>Network Memory</td>
</tr>
</tbody>
</table>

From Table 4, we select the frontier topic 3 in 2011-2012 to discuss and verify. Using the method proposed in this paper, we found that online social networks was one of the frontier topic during 2011-2012.

First of all, in the presentation of topic, the LDA model’s result is 
\{social(0.123),site(0.029),father(0.026),item(0.025),medium(0.025),networks(0.024),influence(0.021),online(0.021),contact(0.016),collaborative(0.016)\}.

**Table 5: Frontier Topic 3 in 2011-2012**

<table>
<thead>
<tr>
<th>Year</th>
<th>Topic</th>
<th>Topic terms and probability in LDA model</th>
<th>The result of representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-2012</td>
<td>Frontier Topic 3</td>
<td>social ( 0.123 )</td>
<td>Social networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>site ( 0.029 )</td>
<td>Online social networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>father ( 0.026 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>item ( 0.025 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>medium ( 0.025 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>networks ( 0.024 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>influence ( 0.021 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>online ( 0.021 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>contact ( 0.016 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>collaborative ( 0.016 )</td>
<td></td>
</tr>
</tbody>
</table>
From these words, it is easy to find that the LDA model’s results are consistent with our representation in experiment. Secondly, we try to make an investigation to prove that online social networks was the focus in that time. In July 6, 2011, founder and CEO of Facebook Mark Zuckerberg confirmed that the social network had reached the milestone, he added that sharing is growing at an “exponential rate.”[3] By contrast Networks Social in the Google trend of the attention distribution, as shown in Fig 3, it is found that the peak is about 2011 and 2012.

Fig 3: Google Trends[4]

CONCLUSION AND FUTURE WORK

The method proposed in this paper which has wide adaptability only needs the title, time and content of the document. This method solves the problem that only single object can be dealt with and that need to spend a lot of time and effort to interpret the results in the past topic detection research. So it can be applied to topic detection based on multi-source data. However, this approach does not work well when dealing with small sets of document sets and short text. This is because such documents are less informative, sparsely word-specific, and context-dependent, so we need to consider how to extend the amount of information in such documents in the future work.

REFERENCES

