Classification of Topic Evolutions in Scientific Conferences

Lin Zhang∗, Yao Guo∗, Xiangqun Chen∗, Weizhong Shao∗, and Lei Chen†

∗ Key Laboratory of High-Confidence Software Technologies (Ministry of Education), Peking University, Beijing, China
{zhanglin08, yaox, cherry, shaowzh}@sei.pku.edu.cn
† Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong
{csyuan, leichen}@ust.hk

Abstract—The number of scientific publications have been increasing explosively in recent years. Although scholar searching engines and recommendation systems help to find relevant papers, neither of them can build overviews of a certain scientific conference, which is more meaningful and important for researchers to keep up with academic trends. In this paper, we propose the concepts of topic trend and positive topic trend, and then declare the formal and quantitative definitions of topic categories, including hot topic, sunrise topic, sunset topic and emerging topic. We design an LDA-based framework to classify research topics of papers in a scientific conference into topic categories. Experiments are then constructed to study the best parameters for detecting different topic categories.

I. INTRODUCTION

Researchers are overwhelmed due to the dramatic growth of scientific papers. Take the field of artificial intelligence as an example, hundreds or even thousands of research papers are published every year, making it nearly impossible for an AI researcher to build a quick overview in a short time.

Scholar searching engines and scientific article recommendation systems [1] [2] [3] [4] [5] [6] help researchers to find relevant papers, however they cannot help to build general views over a certain research area. Kinds of works have been proposed to study topic evolutions in scientific literature [7] [8] [9] [10] [11] [12], however the answers to the following questions are still not clear:

• Which topics are hot in recent years?
• Which topics are becoming more and more popular?
• Which topics are disappearing?
• Which topics are emerging?

In this paper, we study the topic evolutions within a specific scientific conference, facing the following challenges:

• How to measure the impact of a paper on different topics, in case that it involves several topics simultaneously?
• How to measure the trends of topics and then classify them into categories, such as hot topics, sunrise topics, sunset topics or emerging topics?

We made the following contributions in this paper. 1) We proposed a novel method to measure the impact of a researcher on different topics; 2) We proposed a novel method to measure the impact of a paper on different topics; 3) We proposed the concepts of topic trend and positive topic trend, and made the formal and quantitative definitions topic categories, including hot topic, sunrise topic, sunset topic and emerging topic; 4) We studied the best parameters for detecting different topic categories.

Given a certain conference Con, and its frequency f, we calculate the topic vector and reputation vector of each paper in Con, and then associate these papers with their key topics to build Con’s topic rank. The same process is repeated until topic evolutions during the last decade(or even longer) can be generated. With the resulting topic trends and positive topic trends, topics can be dropped or classified into four categories: hot topic, sunrise topic, sunset topic or emerging topic.

II. PROBLEM DESCRIPTION

Since a paper p might be involved with several topics simultaneously, we proposed a vector topic(p) to indicate p’s topics and a vector reputation(p) to indicate p’s different reputations on different topics. We prune some criteria demonstrated by Matsatsinis in [13], and calculate p’s reputation vector reputation(p) as follows:

\[
\text{reputation}(p) = \text{author}(p) + \text{reference}(p)
\]

where author(p) indicates the author reputations on p’s different topics, and reference(p) indicates the reference reputations on p’s different topics. We refer rp, which is the highest reputation in reputation(p), as p’s key reputation, and the corresponding topic tp as p’s key topic.

Topic Order In the paper set Pyear of a conference Cyear, suppose that there are ni and nj papers related with the topic ti and tj, respectively. We declare that ti ≽ tj if and only if ni ≥ nj, which means that ti is more popular than tj since the number of papers related with ti is more than or at least equal to the number of papers related with tj.

Topic Rank In the paper set Pyear of a conference Cyear, we suppose that there are n1, n2, ..., and nK papers related with the topic t1, t2, ..., and tK, respectively(n1 ≥ n2 ≥ ... ≥ nK). Consequently, we define the topic rank TRyear as the

\[
TR_{\text{year}} = \sum_{i=1}^{K} \frac{n_i}{\text{Total Number of Papers}}
\]

The looking-back window size is defined as ten years to evaluate a certain conference, such as the VLDB 10-year best paper award, the AAAI outstanding paper award and so on.
topic sequence \( \{t_1, t_2, \ldots, t_K\} \), in which \( t_1 \succeq t_2 \succeq \ldots \succeq t_K \) can be guaranteed.

**Topic Rank Set** Given a specific conference \( c_{\text{year}} \), we consider the past ten conferences to build topic evolution, and define the topic rank set \( TR_{\text{year}} \) as \( \{ TR_{\text{year} - 9 * f}, TR_{\text{year} - 8 * f}, \ldots, TR_{\text{year} - 1 * f}, TR_{\text{year}} \} \).

**Topic Trend** With the topic rank set \( TR_{\text{year}} \), the topic trend of a certain topic \( t_i \) is defined as \( \{i - 9, i - 8, \ldots, i - 1, i\} \), in which \( i - 9 = \text{rank}(t_i, TR_{\text{year} - 9 * f}), i - 8 = \text{rank}(t_i, TR_{\text{year} - 8 * f}) \) and so on. If \( t_i \) is not contained in \( TR_{\text{year} - n * f} \), which means that no papers related with \( t_i \) were accepted by the conference \( c \) in the year \( \text{year} - n * f \), then \( i - n = \text{rank}(t_i, TR_{\text{year} - n * f}) = -1 \) (\( 1 \leq n \leq 9 \)).

**Positive Topic Trend** The positive topic trend of a certain topic \( t_i \) is defined as the subsequence of \( t_i \)'s topic trend without the ranks of \(-1\). For example, if \( t_i \)'s topic trend is \( \{8, 35, -1, -1, 9, 33, -1, 44, 37, 21\} \), its positive topic trend will be \( \{8, 35, 9, 33, 44, 37, 21\} \).

With the concepts of topic trends and positive topic trends, we define the following four categories:

**Definition 1:** **Hot Topics** Given a threshold \( \delta \), if the ratio of the length of \( t_i \)'s positive topic trend to the length of \( t_i \)'s topic trend is equal to or larger than \( \delta \), we classify the topic \( t \) into the topic category of hot topics.

**Definition 2:** **Sunrise Topics** Given a threshold \( \theta \), if the ratio of the length of the longest decreasing subsequence(LDS) of \( t_i \)'s positive topic trend to the length of \( t_i \)'s topic trend is equal to or larger than \( \theta \), we classify the topic \( t \) into the topic category of sunrise topics.

**Definition 3:** **Sunset Topics** Given a threshold \( \eta \), if the ratio of the length of the longest increasing subsequence(LIS) of \( t_i \)'s positive topic trend to the length of \( t_i \)'s topic trend is equal to or larger than \( \eta \), we classify the topic \( t \) into the topic category of sunset topics.

**Definition 4:** **Emerging Topics** If \( \text{rank}(t, TR_{\text{year} - n}) = -1 \) for every \( n \) (\( 1 \leq n \leq 9 \)), in other words, the length of \( t \)'s positive topic trend is equal to 1, we classify the topic \( t \) into the topic category of emerging topics.

**Problem** Given a conference \( c_{\text{year}} \), classify, rank and select the papers related with Emerging topics, Sunrise topics, Hot topics and Sunset topics.

**Requirements** First, Papers are classified with their key topics into the following four topic categories: Emerging topics, Sunrise topics, Hot topics and Sunset topics. Suppose that there are \( k_E, k_I, k_H, k_D \) topics in each topic category, and papers in the conference \( c_{\text{year}} \) are related with \( K \) topics all together, then \( k_E + k_I + k_H + k_D \leq K \) is guaranteed. Second, papers are ranked with their key topics within a certain topic category. Finally, papers related with the same topic are ranked with their key reputations.

**III. Solution**

As shown in Figure 1, given a conference \( c_{\text{year}} \) and its frequency \( f \), we collect papers in \( c \) from \( \text{year} - 9 * f \) to \( \text{year} \).

\(^2f\) represents the frequency of the conference \( c \).
B. Key Topic and Key Reputation

Papers are classified, ranked and selected with their key topics and key reputations. We determine the key topic and key reputation of a paper based on the following hypothesis. When writing an academic paper, it is most likely for the authors to choose the most relevant references, as well as the most influential references on corresponding topics. And a scientist is most likely to publish a new paper which is related with his/her most influential topic, instead of others.

For example, Table I(c) shows the topic vector, the corresponding author and reference reputation vectors of a paper. Since the maximum reference reputation is 177.3, which means the references of the paper are most valuable/influential on the topic 219, we will set the key topic of this paper as the topic 219. And its corresponding key reputation will be $415.7 + 177.3 = 593.0$.

C. Topic Trend and Topic Category

Papers are classified and ranked with their key topics and key reputations to build the topic rank set $TR_{year} = \{TR_{year-9+f}, TR_{year-8+f}, \ldots, TR_{year-1+f}, TR_{year}\}$. With which, the (positive) trend of a certain topic can be plotted. For example, the trend of a certain topic $t$ is $\{-1, 6, 25, -1, 22, 7, -1, -1, -1, 27\}$, so that its positive topic trend is $\{6, 25, 22, 7, 27\}$. Consequently, one longest increasing subsequence can be $\{6, 25, 22, 7\}$, and one longest decreasing subsequence can be $\{25, 22, 7\}$. According to our definitions, if $\theta$ is set to be 60%, $t$ can be classified either into sunset topics or sunrise topics. Those ambiguous topics are considered as sunset topics, so that the number of sunrise topics, which are more valuable for researchers, can be shrunk.

IV. Experiment Study

We retrieved the basic information of a paper from DBLP, such as title, authors, venue, and year, then applied an APPID from Microsoft Academic Search (MAS) to query the abstract, keywords, citations and references of each paper in the dataset. As a result, we built our dataset of 947,160 papers. We employed Mallet to calculate the topic vector of each paper, with papers published from 1995 to 2000 for training and others for inferring.

A. The Length of Topic Vector

Scanning “call-for-papers” of top conferences in computing science, the number of latent topics is set to be 300. For each abstract, Mallet calculates the probability distribution over topics, and sorts these topics in probabilities from high to low. We demonstrate the maximum, minimum and average probabilities on the top eight topic positions in Table IV and determine the length of topic vector as five.

B. $\delta$ for Hot Topics

We select several conferences in computer science to study the value of $\delta$ for Hot topic detection. For example, if $\delta$ is defined as 0.6, hot topics in ICML2005 include the following eight topics: 68, 116, 135, 154, 167, 188, 194 and 274; And if $\delta$ is strictly set as 1, there are only two hot topics in ICML2005, which are 116 and 188.

We define the topics, selected with $\delta$ as 0.6, as latent hot topics. Then we rank these topics by topic semi-partial order and regard the top five as target hot topics, since it is appropriate and acceptable to recommend users with 7 $\pm$ 2 items at a time. With the definition of target hot topics, we calculate the corresponding precision and recall for different values of $\delta$. It is shown that the best value of $\delta$ for hot topic detection is 0.8 to trade-off between precision and recall. Due to space limitation, we only demonstrate parts of experiment results here in Table III.

C. $\theta$ for Sunrise and Sunset Topics

With similar experiments, we figure out that the best value of $\theta$ for sunrise and sunset topic detection is 0.6 and 0.7, respectively. Due to space limitation, we can not report the experiment results and detailed analysis in this paper.
TABLE II. TOPICS AND KEYWORDS

<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords (Top 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>agent, agents, multi, multiple, intelligent, intelligence, coordination, distributed, system, systems</td>
</tr>
<tr>
<td>116</td>
<td>example, examples, training, learning, machine, artificial, inductive, reinforce, reinforcement, hypothesis</td>
</tr>
<tr>
<td>135</td>
<td>estimate, estimation, sample, parameter, method, distribution, probabilistic, statistical, density, entropy</td>
</tr>
<tr>
<td>154</td>
<td>fast, time, times, speed, efficient, problem, solution, algorithm, approximate, optimal</td>
</tr>
<tr>
<td>167</td>
<td>model, models, markov, chain, chains, continuous, discrete, stochastic, hmm, hmms</td>
</tr>
<tr>
<td>188</td>
<td>feature, features, classifier, classification, nearest, neighbour, extraction, accuracy, recognition, discriminate</td>
</tr>
<tr>
<td>194</td>
<td>match, matching, algorithm, algorithms, pattern, patterns, template, templates, feature, features</td>
</tr>
</tbody>
</table>

TABLE III. HOT TOPIC DETECTION

(a) Different values of delta

<table>
<thead>
<tr>
<th>Venue</th>
<th>Year</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICML</td>
<td>2005</td>
<td>194, 274</td>
<td>68, 135, 154</td>
<td>167</td>
<td>116, 188</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>274</td>
<td>135, 167</td>
<td>116, 188</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Precision and Recall

<table>
<thead>
<tr>
<th>Venue</th>
<th>Year</th>
<th>Top 5 Hot Topics</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICML</td>
<td>2005</td>
<td>116, 135, 167, 188, 274</td>
<td>0.6250</td>
<td>1.0000</td>
<td>0.6667</td>
<td>0.8000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>116, 135, 167, 188, 274</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.8000</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>116, 135, 154, 258, 274</td>
<td>0.5000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>116, 135, 154, 258, 274</td>
<td>0.3125</td>
<td>1.0000</td>
<td>0.6667</td>
<td>0.8000</td>
<td>0.8000</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>116, 135, 154, 258, 274</td>
<td>0.3571</td>
<td>1.0000</td>
<td>0.6667</td>
<td>0.8000</td>
<td>0.8000</td>
</tr>
</tbody>
</table>

Fig. 2. Topic Evolutions over Years (ICML, 1996-2010)

V. CONCLUSION

Scholar searching engines and scientific article recommendation systems help researchers to collect relevant papers, but cannot draw general views over a certain research field. Existing works study topic evolutions in scientific literature, however topic trends are not clear enough for researchers to catch up with academic frontiers. In this paper we proposed the concepts of topic trend and positive topic trend, and made the formal and quantitative definitions of topic categories, including hot topic, sunrise topic, sunset topic, and emerging topic. We proposed an LDA-based framework to classify topics of papers in scientific conferences into categories, and study the best parameters for detecting different topic categories.

ACKNOWLEDGEMENT

This work is supported in part by the High-Tech Research and Development Program of China under Grant No. 2013AA01A605, the National Basic Research Program of China (973) under Grant No. 2011CB302604, the National Natural Science Foundation of China under Grant No. 61103026, 61121063, U1201255, and the NSFC/RGC Joint Research Project (No.60931160444).

REFERENCES