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 Univeristy, Beijing, China
{zhanglin08, yaoguo, cherry, shaowzh}@sei.pku.edu.cn
†Department of Computer Science and Engineering, Hongkong 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 is 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 c_{year} and its frequency f, we calculate the topic vector and reputation vector of each paper in c_{year}, and then associate these papers with their key topics to build c_year's topic rank. The same process is repeated until topic evolutions during the last decade(or even longer) can be generated[1]. 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 $\overrightarrow{topic}(p)$ to indicate p's topics and a vector $\overrightarrow{reputation}(p)$ to indicate p's different reputations on different topics. We prune some criterias demonstrated by Matsatsinis in [13], and calculate p's reputation vector $\overrightarrow{reputation}(p)$ as follows:

$$\overrightarrow{reputation}(p) = \overrightarrow{author}(p) + \overrightarrow{reference}(p)$$

where $\overrightarrow{author}(p)$ indicates the author reputations on p's different topics, and $\overrightarrow{reference}(p)$ indicates the reference reputations on p's different topics. We refer r_p, which is the highest reputation in $\overrightarrow{reputation}(p)$, as p's key reputation, and the corresponding topic t_p as p's key topic.

Topic Order In the paper set P^c_{year} of a conference c_{year}, suppose that there are n_i and n_j papers related with the topic t_i and t_j, respectively. We declare that $t_i \succcurlyeq t_j$ if and only if $n_i \geqslant n_j$, which means that t_i is more popular than t_j since the number of papers related with t_i is more than or at least equal to the number of papers related with t_j.

Topic Rank In the paper set P^c_{year} of a conference c_{year}, we suppose that there are n_1, n_2, ..., and n_K papers related with the topic t_1, t_2, ..., and t_K, respectively($n_1 \geqslant n_2 \geqslant ... \geqslant n_K$). Consequently, we define the topic rank TR^c_{year} as the

[1]The looking-back window size is defined as ten years to evaluate a certain topic with time. Similar idea has been widely employed by academic awards , such as the VLDB 10-year best paper award, the AAAI outstanding paper award and so on.

topic sequence $\{t_1, t_2, ..., t_K\}$, in which $t_1 \succcurlyeq t_2 \succcurlyeq ... \succcurlyeq t_K$ can be guaranteed.

Topic Rank Set Given a specific conference c_{year}, we consider the past ten conferences to build topic evolutions, and define the topic rank set $\mathbf{TR}^c_{year}$ as $\{ TR^c_{year-9*f}, TR^c_{year-8*f}, ... , TR^c_{year-1*f}, TR^c_{year}\}$[2].

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

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 δ, if the ratio of the length of t's positive topic trend to the length of t's topic trend is equal to or larger than δ, we classify the topic t into the topic category of hot topics.

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

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

Definition 4: **Emerging Topics** If $rank(t, TR^c_{year-n}) = -1$ for every n $(1 \leqslant n \leqslant 9)$, 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_{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_{year} are related with K topics all together, then $k_E + k_I + k_H + k_D \leqslant K$ 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_{year} and its frequency f, we collect papers in c from $year - 9 * f$ to $year$, and then build the corresponding topic rank $TR^c_{year-n*f} (0 \leq n \leq 9)$. With the resulting topic rank set $\mathbf{TR}^c_{year}$, we study the topic trend and the positive topic trend of each topic t in TR^c_{year}, and then classify t into a topic category or dropped directly.

[2] f represents the frequency of the conference c.

TABLE I. Example

(a) The Papers of a Researcher r

	1st	2nd	3rd	4th	5th	Citation Count
1	70	132	245	**206**	249	8
2	27	236	159	**206**	45	6
3	**206**	258	208	262	93	2
4	**206**	71	260	251	244	417
5	***111***	296	27	246	**206**	31
6	181	296	208	***111***	27	10
7	70	245	**206**	132	227	8
8	**206**	208	182	108	258	90

(b) The References of a Paper p

	1st	2nd	3rd	4th	5th	Citation Count
1	208	8	145	220	232	200
2	117	276	208	293	**206**	29
3	209	195	208	**206**	289	140
4	23	44	245	295	181	978
5	**206**	89	19	256	285	122
6	232	258	299	156	43	53

(c) Vectors of a Paper p

	1st	2nd	3rd	4th	5th
Topic Index	175	219	225	145	227
A. Reputation	234.8	415.7	354.1	25.4	9.7
R. Reputation	33	177.3	142.5	0	51

A. Topic and Reputation Vector

We employ Latent Dirichlet Allocation(LDA) to detect latent topics of a certain paper p to determine p's topic vector $\overrightarrow{topic}(p)$. For each topic t in $\overrightarrow{topic}(p)$, we calculate p's author reputation and reference reputation on t, and add them together to indicate p's reputation on t.

In this paper, we propose a novel method to evaluate the impact of a certain researcher, which is called scientist reputation, with a set of (topic, average citations) pairs to indicate his/her different impacts on different topics. We collect all of the topics that a certain scientist might have been involved with, and then calculate the average citations from other papers on each topic. For example, the eight papers of a researcher r are listed in Table I(a), and there will be two pairs (111, 20.5) and (206, 80.29) in his scientist reputation, indicating that his reputation on the topic 206 is more influential than that on the topic 111. We scan p's author list, and calculate the scientist reputation of each author. After that, for each topic t in $\overrightarrow{topic}(p)$, we add scientist reputations on t of all authors together to determine p's author reputation $\overrightarrow{author}(p)$.

In this paper we introduce a vector $\overrightarrow{reputation}(p)$ to indicate how the corresponding references of a paper p impact on its involved topics. For example, the six references of a certain paper p are listed in Table I(b), and p's topic vector is <206, 71, 260, 251, 244>. Consequently, p's reference reputation on the topic 206 is defined as $\frac{29+140+122}{3} = 97$, while its reference reputation on other topics are all zero since no references are available.

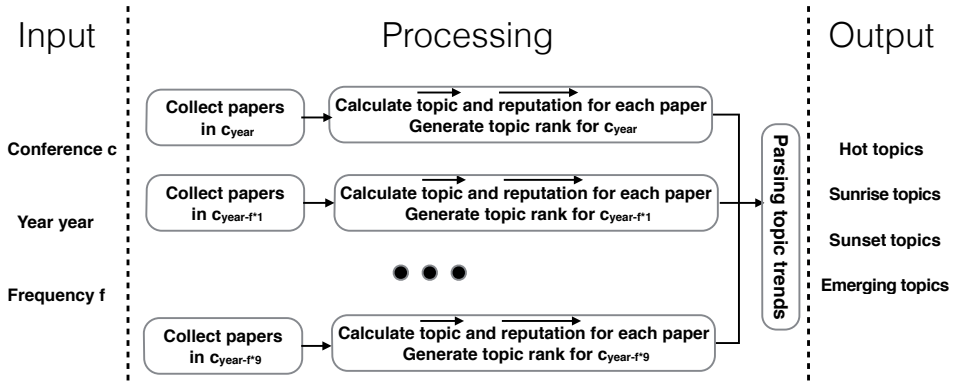


Fig. 1. Framework

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 $\mathbf{TR}^c_{year} = \{TR^c_{year-9*f}, TR^c_{year-8*f}, \ldots, TR^c_{year-1*f}, TR^c_{year}\}$. 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, 7, 27}, and one longest decreasing subsequence can be {25, 22, 7}. According to our definitions, if θ 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 shrank.

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.

TABLE IV. Probabilities

Topic Position	Maximum	Minimum	Average
1st	0.6442	0.0033	0.0856
2nd	0.2387	0.0017	0.0286
3rd	0.1400	0.0015	0.0360
4th	0.1140	0.0015	0.0291
5th	0.0839	0.0015	0.0245
6th	0.0648	0.0014	0.0212
7th	0.0570	0.0014	0.0186
8th	0.0463	0.0014	0.0186

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. δ for Hot Topics

We selecte several conferences in computer science to study the value of δ for Hot topic detection. For example, if δ 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 δ is strictly set as 1, there are only two hot topics in ICML2005, which are 116 and 188.

We define the topics, selected with δ 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 ± 2 items at a time. With the definition of *target hot topics*, we calculate the corresponding precision and recall for different values of δ. It is shown that the best value of δ 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. θ for Sunrise and Sunset Topics

With similar experiments, we figure out that the best value of θ 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

Topic	Keywords(Top 10)
68	agent, agents, multi, multiple, intelligent, intelligence, coordination, distributed, system, systems
116	example, examples, training, learning, machine, artificial, inductive, re-inforcement, hypothesis
135	estimate, estimation, sample, parameter, method, distribution, probabilistic, statistical, density, entropy
154	fast, time, times, speed, efficient, problem, solution, algorithm, approximate, optimal
167	model, models, markov, chain, chains, continuous, discrete, stochastic, hmm, hmms
188	feature, features, classifier, classification, nearest, neighbour, extraction, accuracy, recognition, discriminate
194	match, matching, algorithm, algorithms, pattern, patterns, template, templates, feature, features
274	partition, partitions, decomposition, present, presented, space, spaces, sub-space, dimension, dimensional

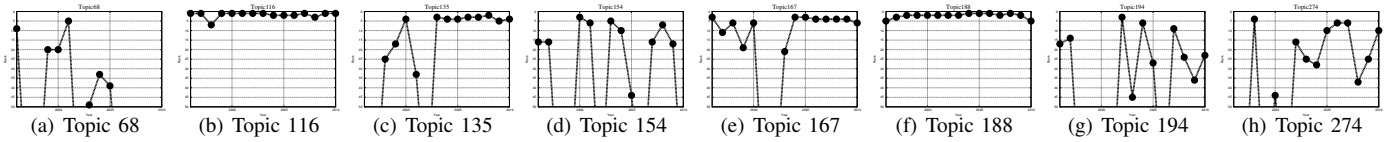
TABLE III. HOT TOPIC DETECTION

(a) Different values of $delta$

Venue	Year	δ				
		0.6	0.7	0.8	0.9	1
ICML	2005	194, 274	68, 135, 154	167		116, 188
	2006		274	135, 167		116, 188
	2007	16, 119, 154, 258, 271		167, 274	135	116, 188
	2008	16, 27, 29, 119, 160, 194, 201, 215, 238, 258	154	167, 274	135	116, 188
	2009	16, 27, 29, 119, 194, 201, 215, 258		154, 167	135, 274	116, 188

(b) Precision and Recall

Venue	Year	Top 5 Hot Topics	δ									
			0.6		0.7		0.8		0.9		1	
			Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
ICML	2005	116, 135, 167, 188, 274	0.6250	1.0000	0.6667	0.8000	1.0000	0.6000	1.0000	0.4000	1.0000	0.4000
	2006	116, 135, 167, 188, 274	1.0000	1.0000	1.0000	1.0000	1.0000	0.8000	1.0000	0.4000	1.0000	0.4000
	2007	116, 135, 167, 188, 274	0.5000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6000	1.0000	0.4000
	2008	116, 135, 160, 167, 188	0.3125	1.0000	0.6667	0.8000	0.8000	0.8000	1.0000	0.6000	1.0000	0.4000
	2009	116, 135, 167, 188, 215	0.3571	1.0000	0.6667	0.8000	0.6667	0.8000	0.7500	0.6000	1.0000	0.4000



(a) Topic 68 (b) Topic 116 (c) Topic 135 (d) Topic 154 (e) Topic 167 (f) Topic 188 (g) Topic 194 (h) Topic 274

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

[1] T. Strohman, W. B. Croft, and D. Jensen, "Recommending Citations for Academic Papers," in *SIGIR*, 2007, pp. 705–706.

[2] K. Sugiyama and M.-Y. Kan, "Scholarly Paper Recommendation via User's Recent Research Interests," in *JCDL*, 2010, pp. 29–38.

[3] J. Tang and J. Zhang, "A Discriminative Approach to Topic-Based Citation Recommendation," in *PAKDD*, 2009, pp. 572–579.

[4] Q. He, J. Pei, D. Kifer, P. Mitra, and C. L. Giles, "Context-Aware Citation Recommendation," in *WWW*, 2010, pp. 421–430.

[5] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl, "On the Recommending of Citations for Research Papers," in *CSCW*, 2002, pp. 116–125.

[6] C. Wang and D. M. Blei, "Collaborative Topic Modeling for Recommending Scientific Articles," in *KDD*, 2011, pp. 448–456.

[7] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the History of Ideas Using Topic Models," in *EMNLP*, 2008, pp. 363–371.

[8] T. L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl. 1, pp. 5228–5235, April 2004.

[9] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths, "Probabilistic Author-Topic Models for Information Discovery," in *KDD*, 2004, pp. 306–315.

[10] D. Zhou, X. Ji, H. Zha, and C. L. Giles, "Topic evolution and social interactions: How authors effect research," in *CIKM*, 2006, pp. 248–257.

[11] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: How can citations help?" in *CIKM*, 2009, pp. 957–966.

[12] T. Masada and A. Takasu, "Extraction of topic evolutions from references in scientific articles and its gpu acceleration," in *CIKM*, 2012, pp. 1522–1526.

[13] N. F. Matsatsinis, K. Lakiotaki, and P. Delias, "A System based on Multiple Criteria Analysis for Scientific Paper Recommendation," 2007.