ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(26), DOI: 10.17485/ijst/2015/v8i26/80470, October 2015* ISSN (Online) : 0974-5645

Topic Clustering from Selected Area Papers

Ho-Hyun Park[1], Jaehwa Park[2] and Young-Bin Kwon[2*]

[1]Department of Electrical and Electronics Engineering, Chung-Ang University, Seoul, 156 – 756, Korea
[2]Department of Computer Science and Engineering, Chung-Ang University, Seoul, 156 – 756, Korea;
ybkwon@cau.ac.kr

Abstract

An extracting method of research trend in the field of a computer network which contained in the published papers of related conferences is presented. A topic is defined as a subset of vocabulary and the interest of the topic is represented as 'saliency'. The saliency, the degree of topic interest is measured as a product of joint distributions of vocabularies which consist the topic. To reduce the computational burden, clustering and selection procedures of vocabularies are applied before actual topic grouping. Two experiments: 1. Research trend analysis and 2. Topic correlation analysis of conferences has been performed. The leading 24 conferences related to the computer networks which are held during 2009-2010 are exploited. The experimental results show the validity of the presented method.

Keywords: Research Trend Extraction, Topic Correlation Analysis, Topic Grouping

1. Introduction

Information technology is one of the fundamental fields to support overall industries and a lot of new techniques are rapidly developed and applied for our daily life to social infrastructures. Many conferences related to information and computer technology have been held recently with a dramatic development of computer technology for last decades. Also, a lot of information related to the technology is provided through the conferences and the information is one of the sources of leading power on this field of industry as well as academia. However, to extract the major trends among the conferences are somewhat difficult and require a lot of time and elaborations.

Researchers on the field of information technology meet difficulties to find out proper information and research trends since technology changes too rapidly and related conferences are too many to attend. Many of researchers on this field meet difficulties to find out the proper information effectively and efficiently. One of the most important things is to figure out the essential requirements or necessities of technologies which are following the trend.

However, the trends are rapidly changing; it is not easy to catch up in proper time. Also the relativity between conference topics and his or her own research interests are hard to determine due to the wide spectrum of newly created topics. Methods for generating guidelines or trend analysis for related conferences are highly requested for researchers to refine research direction.

In this paper, an extracting method of research trend in the field of computer networks via analysis of the topics contained in the published papers of related conferences is presented. A topic grouping method is presented to extract research trend and to measure relativity among conferences. A computational model based on a joint distribution for a topic and a trend extraction is developed.

For the experiment, the leading 24 conferences of the IEEE (Institute of Electrical and Electronics Engineers) related to the computer networks which are held during 2009-2010 are exploited. The research trend analysis using the topic model is performed. The contents of papers which are presented in the proceedings of the conferences are used as a corpus for the analysis. Also the topic correlation between two conferences is measured using the topic vectors which are generated from extracted topics.

** Author for correspondence*

2. Background

A topic summary on the research field of Computer Vision and Image Analysis has been done based on analysis of 1600 papers which are extracted from related conferences for 1988[1]. The topic summary presented in[1] shows research trends on the field clearly and provides a direction of further research. In[2], research trends have been analyzed based on text mining. A combination of TF-IDF (term frequency-inverted document frequency) and RFM (Recent Frequency Monetary)[6–8] is used for the trend analysis.

Similar trend analyses on various research field have been exploited; Technology on Fuel Cell[3], Aquaculture of Canada[4], New Product Development or New Technology Creation[5] etc. In[9] a method enhancing efficiency of website clustering is introduced by detecting the flow structures of the hyperlink, webpage and directory structure. The flows are generated from the analysis of keywords extracted from the webpages. The approach provides more detailed and dynamic information compared to previous approaches.

Keyword analysis based approaches are one of the major streams to extracting the characteristics and research trends of specific journals or conferences and several works have been published recently[10,11].

In[10], to compare two conferences A and B, from the published papers (A_1, A_2, ... A_n) and (B_1, B_2, ... B_m) of the conferences, keywords are extracted from the papers and keyword groups ($\boldsymbol{W}_{A1}$, $\boldsymbol{W}_{A2}$, ... $\boldsymbol{W}_{An}$) and ($\boldsymbol{W}_{B1}$, $\boldsymbol{W}_{B2}$, ... $\boldsymbol{W}_{Bn}$) are generated respectively. Then the relativity between the two keyword groups is measured using the cosine similarity[12]. The cosine similarities between all pair of conferences are obtained and converted to 0 or 1 with predefined threshold. Finally, a relativity matrix is generated from the binary similarities and an overall relativity of all the conferences is displayed as a binary matrix of network structure. The characteristics of the conferences are analyzed using the matrix.

In[11], keyword analysis is adopted to define the research subfield of Engineering Education Research. The show-up frequencies of keywords are measured in conferences and journals related to Engineering Education Research. And the ratio among the frequencies is measured as a popularity metric of the subfield which is represented by the keyword. The ambiguity problem of the keyword which disturbs the purity of measurements is claimed and a filtering method is proposed to solve the ambiguity of keywords.

A similarity measuring method between two documents is presented in[13]. In this approach, the similarity is measured based on comparing the bibliographic information using the title and keywords. This method has advantages when a hierarchical similarity measure is required. However, the method is primarily for comparing two documents, and has limitations to extend conference topic comparison and trend extraction.

3. Topic Grouping

Topic Grouping is a method to summarize a large collection of text data with a set of keywords or smaller number of words with distributions. The set of keywords, ranked or not, is the major topics of the collected data set, which represent the salient themes crossing the data set. Finding these topics among related research papers allows for effective representation of research trends in the research field.

In this section, a computational model for a topic is represented and a method for research trend analysis is introduced using the topic model. A topic is defined as a vocabulary sequence and the degree of importance or interest of the topic is represented as 'saliency'. The saliency is measured as a product of joint distributions of vocabularies which consist the topic. To reduce the computational burden, clustering and selection procedures of vocabularies are applied before actual topic grouping. The selected vocabularies become keywords for next coming topic grouping.

3.1 Computational Topic Model

Let V be a dictionary of the collected text data and v be a vocabulary of V, $v \in V$. A topic t is defined as a sequence of vocabularies.

$$t = (v_0, v_1, v_2, \ldots, v_n) \qquad (1)$$

where n and N are the sizes of sequence and dictionary respectively $n << N$.

Let θ_t denote the saliency of the topic t. The saliency of the topic can be given by the joint distributions of vocabularies as,

$$\theta_t = d(v_0).d(v_1 \mid v_0).\ d(v_2 \mid v_0, v_1)\ldots\ d(v_n \mid v_0, v_1 \ldots\ v_{n-1}) \qquad (2)$$

where $d(v_i \mid v_j)$ denotes a distribution of v_i given

v_j. d becomes a joint probability if the distributions of vocabularies are given by probabilities. The higher saliency of a topic represents the more frequency to be presented in the data set. For example, the saliency of a topic ‘wireless network’ is obtained by the product of $d(wireless^1)$ and $d(wireless^{11} | network^1)$.

If n increases, the joint distributions are hard to be obtained, since the N is usually much larger than n. The number of joint distributions to find saliency of topic with length n becomes N^n theoretically.

To reduce the computational burden, the saliency can be approximated or redefined only considering the bigrams as,

$$\theta_t = d(v_0).d(v_1 \mid v_0).\ d(v_2 \mid v_1)\dots\ d(v_n \mid v_{n-1}) \qquad (3)$$

Then, the required joint distributions becomes $N \times N$. However, since N is usually large, the size of $N \times N$ is still very large.

Sometimes the bigrams of two vocabularies do not much represent the joint distributions of the topics. For example, the topic ‘wireless sensor network’ is much closer to the union of ‘wireless network’ and ‘sensor network’ rather than the union of ‘wireless sensor’ and ‘sensor network’.

To compensate above mentioned problem and to represent the saliency of topic with larger n more accurately, the bigram distribution is redefined. Let the joint distribution of a vocabulary v_1 of a topic t, denoted as $P_t(v_i)$ be defined as the maximum bigram distribution of $d(v_\mathrm{i} \mid v_\mathrm{j})$ as,

$$P(v_i) = \max\{d(v_\mathrm{i} \mid v_j) \mid \forall v_\mathrm{j}, 0 \le \mathrm{j} \le n_t \qquad (4)$$

where n_t is the size of t.

If the entire vocabulary sequence of a topic t is not important in point of the bigram distribution, then the saliency of topic $t$is able to be defined as a product of bigram vocabularies as,

$$\theta_t = p(v_0).p(v_1).p(v_2)\dots p(v_n) \qquad (5)$$

Since the ‘wireless sensor network’ is a more specific topic of ‘wireless network’ and ‘sensor network’, the topic ‘wireless sensor network’ could be a sub-topic of both higher level topics. It is the inverse concept of saliency metric definition, since the saliency of ‘wireless sensor network’ is calculated from the saliency product of sub-topics composed by subset terms of ‘wireless network’ and ‘sensor network’.

Following the definition, if a topic is a specific topic of general topics which can be composed by the words consisting the specific topic, the saliency value is always equal or smaller than the saliency values of the general topics. However, in practical cases, sometimes ‘wireless sensor networks’ is a hotter topic than ‘wireless network’ and ‘sensor network’, since ‘wireless network’ and ‘sensor network’ are the common and generalized topics that no more can represent the research trends.

If the saliency of a topic is represented only by bigram distributions, the saliency of ‘wireless sensor network’ could be calculated as a much smaller value than those of other topics. If ‘wireless network’ and ‘sensor network’ are too broad to represent latently interested topics, the two topics are seldom found in the data set and result a lower value of saliency of ‘wireless sensor network’.

To avoid such a problem, the joint distribution of a vocabulary $p_t(v_\mathrm{i})$ needs to be open to multi-grams as

$$p(v_\mathrm{i}) = \max\{d(v_\mathrm{i} \mid \omega) \mid \forall\ \omega \subset \{v_0, v_1, v_2, \dots, v_{\mathrm{n_t}}\}\} \qquad (6)$$

If the $p(v_\mathrm{i})$ of (6) is used to calculate θ_t, the sub-topic problem can be avoided. If ‘wireless sensor network: WSN’ is the hottest topic, then

$$\theta_{wsn} = p(w).p(s).p(n) = d(wsn) \qquad (7)$$

θ represents its own distribution of ‘wireless sensor network’ itself. However, this approach increases the computational burdens. Thus to reduce the computations, the size of vocabulary and length of topics should be reduced.

3.2 Topic Grouping

The eventual target for topic grouping is to generate a summary or a trend from the given a large collection of text data. Since the number of vocabularies on a dictionary of large collection of text data is relatively huge compared to other applications, it is almost impossible or impractical to generate all the possible joint distributions among vocabularies.

The role of summary is to represent the major research direction of trends of the given data, if the data is collected from research paper of a specific field. Thus the summary as a result of topic grouping does not have to be so accurate requiring analysis of relations among all the vocabularies.

To make the computational amount in a practical level, the size of vocabulary and length of topic need to be limited. The vocabularies shown more frequently in papers have more importance and more chance to represent the trends. However, the higher frequency of a vocabulary does not directly mean the more saliency to represent research trends. Of course the vocabularies should have an eligibility to be used as a topic. For example, though the word ‘the’ or ‘get’ have higher shown frequencies, they cannot represent a research trend. A pre-screening procedure of vocabularies should be preceded before actual topic grouping.

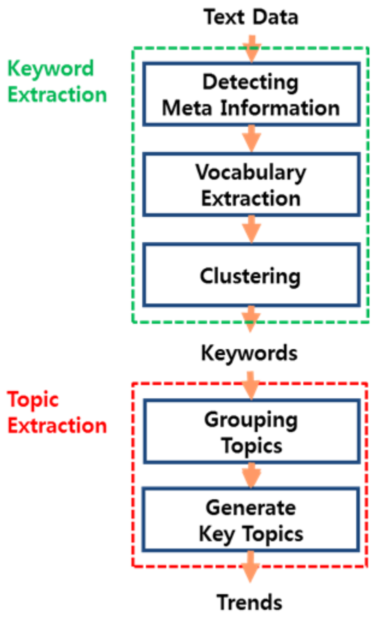

Figure 1. Block diagram of trend analysis.

Figure 1 shows the procedure of topic grouping for trend analysis of given text data. Two steps are applied for trend analysis: keyword extraction and topic extraction. Since the text data is collected from research papers of conference proceedings, the meta information about the conference is necessary to select the major vocabularies. The titles of papers, the number of sessions, papers, session topics and format of papers are detected for pre-screening of vocabulary.

In the vocabulary extraction, using the meta information of a conference, vocabularies which are assumed to have eligibility to represent the technical trends of the conferences are extracted. In general, keywords of each paper are the essential elements to extract in this process.

Then a clustering procedure is applied to the collection of vocabularies of all the input conferences. The clustering reduces the representation varieties of a word which are caused from the different format of the same word such as the difference between the plural and singular forms, for example ‘network’ and ‘networks’. For the distance metric of clustering the ‘edit distance’ is adopted to measure the similarities between vocabularies.

As the result of keyword extraction step, a list of keywords is generated. The list of keywords is the dictionary of major vocabularies in single words. Hyphenated words or abbreviations are allowed but multiple words or phrases are not allowed as a keyword.

Using the extracted keywords, the joint distributions are obtained from the text data. The whole collection text data can be used as well as a subset, for example abstraction of the paper, can be used as a corpus to extract the joint distributions. A combination of vocabularies is generated as a topic and the saliency value of the topic is determined using joint distributions. The maximum length of the topic can be predetermined before topic grouping or incrementally increases with a certain metric.

Select a group of top ranked m topics which has larger values of saliency, or a group of all topics which has larger saliency over a predetermined threshold. Among the selected topics grouping of topics is applied. Since the more specific topic usually includes the broader topics, for example, ‘wireless sensor network’ and ‘sensor network’, the relation of inclusion among the selected topics is checked.

As a result of inclusive checking, a hierarchy of topic is generated. A topic is able to be a hierarchy of topic as well as a group of topics. Also a topic is eligible for a member topic of multiple groups of topics. Top k hierarchies words of topics are selected to summary the trends of the data collection.

3.3 Correlation Analysis

Topic correlation between two conferences can be measured comparing the inclusiveness of the major topics. The topic correlation procedure is shown in Figure 2.

For topic correlation analysis among conferences, the target conferences are selected and the target conference group *C* is generated as,

$C = \{c_1, c_2, \ldots\ldots, c_n\}$

where c denotes a conference and n denotes the number of selected conferences. Then a topic group K is generated from the conference group C as,

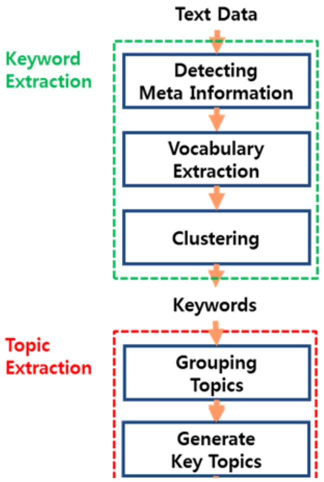

Figure 2. Block diagram of correlation analysis.

$K = \{k_1, k_2, \ldots\ldots, k_m\}$

where k denotes a topic and m denotes the size of the topic group.

Using the topic group, a topic vector Λ is generated for each conference c

$\Lambda_i = \{\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ij}, \ldots, \lambda_{im}\}$

If conference c_i includes the topic k_j then λ_{ij} the j^{th} bit field of topic vector Λ_i becomes '1', otherwise '0'. Then a topic vector with m bits for each conference is generated and a n by m topic relativity matrix can be generated as shown in Table 1. The binary vector Λ_i represents a research trend of the conference c_i projected on the topic group K.

Table 1. Topic relativity matrix

Conference	c_1	c_2	....	c_n
k_1	1	1		0
k_2	1	0		1
...	...	...	...	...
k_m	1	0		0

The topic correlations between two conferences are measured by cosine similarity[12]. The topic correlation η_{ij} between conference c_i and c_j is calculated as,

$$\eta_{ij} = \frac{\Lambda_i \cdot \Lambda_j}{\|\Lambda_i\|\|\Lambda_i\|} = \frac{\sum_{k=1}^{m} \lambda_{ik} \times \lambda_{jk}}{\sqrt{\sum_{k=1}^{m} \lambda_{ik}} \times \sqrt{\sum_{k=1}^{m} \lambda_{jk}}} \qquad (8)$$

4. Experimental Results

Two experiments: 1. Research trend analysis and 2. Correlation analysis of conferences have been done to test the performance of the proposed method and to check the usability of extracted topics.

In the first experiment, the performance of the proposed method as a tool to summary the large collection of paper data has been checked. And in the second experiment, the usages of extracted topics as a metric to measure the similarities or relativity between two conferences are tested.

The 24 conferences of the IEEE related to the computer networks which were held during 2009-2010 annually are exploited. Titles of 6 conferences among 24 used in the experiments are shown in Table 2.

Table 2. Titles of conferences used in the experiments (partial)

Index	Conference Title
1	IEEE International Symposium on World of Wireless Mobile and Multimedia Networks (WoWMoM).
2	IEEE International Symposium on Network Computing and Applications.
3	IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WIMOB).
4	IEEE Workshop on Secure Network Protocols (NPSec).
5	IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON).
6	International Conference on Wireless and Optical Communications Networks (WOCN).

4.1 Research Trend Analysis of Network Area

Before the actual trend analysis of conferences, the raw vocabulary extraction and filtering process have been applied. Since the scope of the trend analysis spans to entire 24 conferences related to the research field of computer networks, the size of the dictionary is huge. Extraction of a subset of vocabulary which can effectively represent the contents of a data set is necessary for the practical experiment.

Where TN: Telecommunication Network, QoS: Quality of Service, WSN: Wireless Sensor Networks, RP: Routing Protocols

The vocabularies are selected from paper titles, keywords listed in the papers and session names of all the conferences and the frequencies are measured. With ranked list of extracted vocabularies, general words, such as words 'the', 'of', 'study'...etc, are filtered out manually.

Table 3. Extracted Keywords

2009

Size of Group	100	200	300	Unlimited
Top 1	TN	TN	QoS	WSN
Top 2	QoS	RP	TN	QoS
Top 3	RP	QoS	WSN	internet

2010

Size of Group	100	200	300	Unlimited
Top 1	WSN	WSN	WSN	WSN
Top 2	TN	TN	RP	internet
Top 3	RP	RP	QoS	QoS

2009 - 2010

Size of Group	100	200	300	Unlimited
Top 1	TN	TN	WSN	WSN
Top 2	WSN	RP	TN	QoS
Top 3	QoS	WSN	QoS	internet

From the 24 conferences, total 12,188 words are extracted finally. The average size of extracted vocabularies is 872 per conference. With the 12,188 words, total 6,965 topics are generated with saliency threshold of 0.15. Top 3 topics are shown in Table 3 for 2009, 2010 and 2009-2010. The size of group means the maximum size of grouped topics when topic clustering is applied. The edit distance is used as a metric to measure the similarity between two topics and k-means algorithm is used for topic clustering.

The most commonly shown topics during 2009-2010 are wireless sensor network, QoS, Routing Protocols, a telecommunication network, and internet. As shown in the analysis results in Table 3, the hot interested topics in the research field of computer networks for 2009-2010 are extracted and represent the direction of research trends.

4.2 Topic Correlation Analysis of Conferences

Table 4. Topic vector for correlation analysis

Conference Index of Table 2	1	2	3	4	5	6
Wireless Sensor Networks	1	1	1	1	1	1
Routing Protocol	1	1	1	1	1	1
Energy Efficiency	1	1	1	1	1	1
Cognitive Radio	1	0	1	1	0	1
Ad Hoc Network	0	1	1	1	1	1
p2p	0	1	0	1	0	1
Resource Allocation	1	1	1	1	0	1
Reliable Transport Protocol	1	0	0	1	1	0

Topic correlation between two conferences can be measured comparing the inclusiveness of the major topics. From the first experiment, 6,965 topics with unlimited topic grouping are selected for topic vector generation. If a conference includes the topic then the bit field of a topic vector becomes '1', otherwise '0'. Then a topic vector with 6,965 bits for each conference is generated as shown in Table 4.

Table 5 shows a result of topic correlation analysis among the conferences which are shown in Table 2. The topic vectors of Table 4 are used for calculation of the topic correlation.

Table 5. Result of correlation analysis (in partial)

Conf. Index	1	2	3	4	5	6
1.	1.00					
2.	0.56	1.00				
3.	0.48	0.88	1.00			
4.	0.65	0.48	0.38	1.00		
5.	0.47	0.43	0.55	0.38	1.00	
6.	0.52	0.39	0.48	0.34	0.83	1.00

The most correlated conference pair is (2, 3) whereas the least correlated pair is (4, 6) with 0.88 and 0.34 topic similarity measurement respectively. The conferences 2 and 3 are highly related conferences on the research field of 'Advanced Information Networking and Application'. The conferences 4 and 6 are not much related on the topics since conference 4 is specialized in 'Mesh Networks' whereas conference 6 is focused on 'Social Networks'. As shown in the experimental result, the topic grouping and correlation analysis are working reasonably well.

5. Conclusion

An extracting method of research trend in the research field of computer network contained in the published paper of related conferences is presented.

A computational model for topic and a trend extraction method using topics are presented. The experimental results show the validity of the methods. However, more computationally efficient method for topic grouping and accurate metric to measure the similarity between two conferences are necessary for improvement.

6. Acknowledgment

This research was supported by National Research Foundation of Korea, Electronic and Information Research Information Center program, 2013, and Basic Science Research program, NRF-2014R1A1A2056266.

7. References

1. Rosenfield A. Image analysis and computer vision: 1988. Computer Vision, Graphics, and Image Processing. 1989; 46(2):196–250.
2. Terachi M, Saga R, Tsuji H. Trends recognition in journal papers by text mining. IEEE International Conference on Systems, Man and Cybernetics. 2006; 6:4784–9.
3. Maxwell JC. A treatise on electricity and magnetism. 3rd ed. Oxford: Clarendon. 1892; 2:68–73.
4. Yamada K, Komine H, Kinukawa H, Nakagawa H. Abstract of abstract: A new summarizing method based on document frequency and clause length. The 8th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI'2004); Orland. 2004.
5. Yamamoto H, Ohmi S, Tsuji H. Entropy-based indexing term selection for N-gram text search system. IEEE International Conference on Systems, Man and Cybernetics (IEEE/SMC'2003); 2003. p. 4852–7.
6. Salton G. Automatic Text Processing. Addison-Wesley Publishing Company. 1989.
7. Libey DR. Libey on RFM value. e-Book. Available from: http://www.e-rfm.com
8. Lee S, Yoon B, Park Y. An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation. 2009; 29(6):481–97.
9. Tonella P, Ricca F, Pianta E, Girardi C. Using keyword extraction for web site clustering. Proceeding of the Fifth IEEE Int Workshop on Web Site Evolution (WSE); 2003. p. 41–8.
10. Su HN, Lee PC, Chan TY. Bibliometric assessments of network formations by keyword-based vector space model. Technology Management for Global Economic Growth (PICMENT); 2010. p. 1–9.
11. Nawaz S, Rajan P, Yu JH, Yi L, Choi JH, Radcliffe DF, Strobel J. A keyword based scheme to define engineering education research as a field and its members. IEEE Global Engineering Education Conference (EDUCON); 2011. p. 201–9.
12. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Pearson Addison Wesley; 2005.
13. Maneewongvatana S. A similarity model for bibliographic records using subject headings. JCSSE. 2011: 263–8.