Journal of Informetrics 8 (2014) 776–790

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Time gap analysis by the topic model-based temporal technique

Do-Heon Jeong[a,1], Min Song[b,*]

[a] Korea Institute of Science and Technology Information (KISTI), 245 Daehak-ro, Yuseong-gu, Daejeon 305-806, South Korea
[b] Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, South Korea

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form 10 July 2014
Accepted 14 July 2014

Keywords:
Text mining
Topic modeling
Latent Dirichlet Allocation (LDA)
Content analysis
Temporal analysis
Multiple resources

ABSTRACT

This study proposes a temporal analysis method to utilize heterogeneous resources such as papers, patents, and web news articles in an integrated manner. We analyzed the time gap phenomena between three resources and two academic areas by conducting text mining-based content analysis. To this end, a topic modeling technique, Latent Dirichlet Allocation (LDA) was used to estimate the optimal time gaps among three resources (papers, patents, and web news articles) in two research domains. The contributions of this study are summarized as follows: firstly, we propose a new temporal analysis method to understand the content characteristics and trends of heterogeneous multiple resources in an integrated manner. We applied it to measure the exact time intervals between academic areas by understanding the time gap phenomena. The results of temporal analysis showed that the resources of the medical field had more up-to-date property than those of the computer field, and thus prompter disclosure to the public. Secondly, we adopted a power-law exponent measurement and content analysis to evaluate the proposed method. With the proposed method, we demonstrate how to analyze heterogeneous resources more precisely and comprehensively.

1. Introduction

Research on citation analysis in the bibliometrics area started in the mid-1950s, when co-citation relationships between academic journals were studied (Garfield, 1955). Traditionally, main information resource utilized in bibliometrics was a scholarly paper (Ball & Tunger, 2006), whereas patent information was another main information resource used for the technology-competitive analysis and promising technology finding (Daim, Rueda, Martin, & Gerdsri, 2006). As the importance of World Wide Web has been highlighted since the late 1990s, the concept of webometrics which involved the intellectual structure of the Web has also emerged as a mainstream research (Björneborn & Ingwerson, 2004). Meanwhile, through scientometrics studies, the quantitative analysis area expanded into scientific technologies, research policies, and the social studies of science (Schoepflin & Glänzel, 2001). Currently, informetrics which covers various information resources and methodologies is widely used from the quantitative perspective (Egghe, 2005).

Several limitations have existed in the various fields of trend analysis. Firstly, most analysis-related studies only analyzed subjects using a single information resource. Xu, Zhu, Qiao, Shi, and Gui (2012) and Sajjad et al. (2013) argued that such

* Corresponding author. Tel.: +82 2 2123 2416.
E-mail addresses: heon@kisti.re.kr (D.-H. Jeong), min.song@yonsei.ac.kr (M. Song).
[1] Tel.: +82 42 869 1792.

http://dx.doi.org/10.1016/j.joi.2014.07.005
1751-1577/

an approach narrowly led to one-sided results. Secondly, although a few attempts have been made to identify meaningful connections between heterogeneous resources, most studies were still limited to connecting resources through citation information between patents and papers (Finardi, 2011; Shibata, Kajikawa, & Sakata, 2010). Thirdly, no major studies have been made on content-focused interlinking Web news articles with other resources such as papers or patents. A large number of studies have attempted to analyze recent trends based on Web news articles (Amitay, Carmel, Herscovici, Lempel, & Soffer, 2004; Kim & Oh, 2011; Vaughan & You, 2008), but integrative studies considering various resources were still insufficient.

This paper aims to achieve the following objectives by providing an integrative method for analyzing multiple resources such as papers, patents, and Web news articles. Firstly, the proposed method of the time gap analysis aims to shed light on how a specific resource precedes others. If a consistent time gap phenomenon is found between resources, then these resources can be arranged according to the time taken from creation to publication. Secondly, this study aims to identify whether the coherent time gap phenomenon occurs between academic areas. If a time gap between resources in a specific academic area is consistently shorter than other areas, topics of the academic area change faster than ones of other academic areas.

This paper is organized as follows. In the related work section, previous studies of recent trend and temporal analysis are discussed. In addition, quantitative research and text mining-based studies are reviewed from the perspective of the trend analysis and multiple resources. In the Section 3, we propose the topic model-based method for analyzing the time gap phenomenon, which is the core part of this study. In the Section 4, multiple resources are plotted on one time axis; the time gap phenomenon between Web news articles, papers, and patents is revealed, and trends between academic fields are identified through the experiments based on the optimization model. In the Section 5, we propose two measurements to interpret experimental results with the statistical method and interpretation of significant topic changes. Finally, in the conclusion section, the proposed method and various experimental results are summarized. Study limitations and future research directions are also discussed.

2. Related work

2.1. Trend analysis with various resources

To examine trends or find competition strategies, quantitative analysis method of informetrics was mainly utilized to determine co-authorship and citation relationships (Chua & Yang, 2008) and correlations of Web-link information (Vaughan & You, 2008). Other researchers made use of a combined research method using both quantitative analysis and content analysis based on text mining (Song, Kim, Zhang, Ding, & Chambers, 2014). Whereas, computer science field has been mainly interested in terminological trend analysis based on natural language processing (NLP) technique. Topic Detection and Tracking (TDT) is conducted on the basis of corpora built mainly through Web news articles and papers (Mei & Zhai, 2005).

Analysis studies by the type of resource can be classified as follows; firstly, research based on paper analysis primarily looks at citation relationships between papers. Therefore, many researchers analyzed the intellectual structures of scholarly fields using citation information (Chua & Yang, 2008), or revealed co-authorship in specific academic disciplines (Levitt & Thelwall, 2008); and still others found core articles on promising technologies (Shibata et al., 2010). Secondly, studies that focus on patents tend to analyze technological innovations or R&D trends. Lee and Lee (2013) attempted to detect emerging technologies in the energy-engineering field with patent data collected from the U.S. Patent and Trademark Office (USPTO). Guan and Zhao (2013) tried to select university-industry collaboration partner based on patent data in the field of nanobiopharmaceuticals. Thirdly, Web data-based research can be classified into Web structure analysis using Web links and Web content analysis based on text mining. With regard to the Web-link-associated traditional informetrics, which also means webometrics, Amitay et al. (2004) conducted trend detection by analyzing temporal link for Web site search and event detection. Furthermore, Vaughan and You (2008) conducted a study to find business competition using the co-link information using the content analysis of Web resource.

The aforementioned studies have the limitation of only addressing a single information resource. In the mid-1980s, Lancaster and Lee (1985) conducted topic analysis by examining topic diffusion of various journals. They also analyzed the time gap phenomenon in a domain, but still dealt with a single resource. Experimental research has recently attempted to investigate these problems and enable the use of heterogeneous resources. The mainstream of studies on heterogeneous data attempts to coordinate research papers and patents as the fundamental indicators for scientific research and development. Narin, Hamilton, and Olivastro (1997) investigated correlations between terms co-occurring in both papers and patents. Kim, Hwang, Jeong, and Jung (2012) categorized all technical terms into five development stages using academic papers and patents based on decision tree algorithm. Xu et al. (2012) constructed semantic linkages between patents and papers based on LDA topic modeling. Sajjad et al. (2013) determined the time differences in technological trends in the fields of bio and medical sciences based on papers, patents, and Web news articles. The results yielded the findings that the terms appearing in Web news articles were the slowest in start point but the fastest in growth and extinction, whereas papers marked the fastest start points and the most sustainable graph curves.

From the review of the related studies, it can be concluded that research efforts using multiple resources have not yet begun in full force. This study aims to propose a new temporal analysis method for performing integrative analysis of

multiple resources so that future studies will be able to perform more precise trend analyses in big and heterogeneous data environments.

2.2. Temporal analysis using topic modeling

The probabilistic topic modeling is an unsupervised learning-based text analysis method and provides both the estimation model for similar documents and semantic latent topics. LDA topic modeling involves methods conceptualizing literature in an unsupervised learning environment, but it does not cluster literature itself but model concepts through the word-to-word relationships within literature (Blei, Ng, & Jordan, 2003). This modeling method has been widely used in recent studies. There are an increasing number of research cases in which attempts are made to combine conventional theories and methods, such as information retrieval (Vulić, Smet, & Moens, 2013), Web search (Ralf & Peter, 2012), classification (Lu, Okada, & Nitta, 2013), image and video search (Vretos, Nikolaidis, & Pitas, 2012) with LDA topic modeling.

In the following, studies are introduced in which topic modeling is applied to time-series data processing and quantitative analysis. Song et al. (2014) carried out a quantitative study to estimate the productivity and influence in the field of bioinformatics on the basis of papers and citations stored in PubMed Central from 2000 to 2011. Griffiths and Steyvers (2004) applied topic modeling to identify the research topics studied most intensively (hot topics) and those studied less and less (cold topics) from the Proceedings of the National Academy of Sciences of the United States of America (PNAS) from 1991 to 2001. Blei and Lafferty (2006) suggested dynamic topic models (DTM) extended from the existing LDA model as an algorithm suitable for temporal information. Wang and McCallum (2006) suggested that a LDA-based Topics Over Time (TOT) model analyzed well topic-occurring time points by combining topic and time-stamp information. Kim and Oh (2011) attempted to detect the time-dependent changes of corpora contained in Web news articles using topic modeling and suggested the use of indirect assessment via topic chain analysis.

In Section 2, quantitative research and text mining-based studies were reviewed from the trend analysis viewpoint, and issues in studies with multiple resources were also discussed. Then, the topic modeling techniques were explained with a number of LDA-based studies. In Section 3, the main method of the time gap analysis will be introduced in detail.

3. LDA-based time gap analysis method

3.1. Overview and data collection

3.1.1. Overview

To discover the time gap phenomenon between various resources and between representative academic fields, the time gap analysis method consists of the following six steps:

(1) *Word extraction*: The four types of resources collected for word extraction are papers and proceedings, patents, and Web news articles. All the resources are stored by time-series unit.
(2) *Topic modeling*: The extracted words can be conceptualized through LDA topic modeling. LDA reorganizes resources not into units of documents but into topics. This study sets the time slice as the basic concept unit for analyzing the time gap.
(3) *Measuring similarity between topics*: This step is a process that infers the similarity between two resources. The similarity between two resources can be obtained by repeatedly measuring the similarity between a series of time slices.
(4) *Analyzing the time gap*: This study compares the time gap phenomenon of various cases such as proceedings and papers, patents and papers, Web news articles and papers, and Web news articles and patents.
(5) *Selecting a relevant model*: A model that interprets well the time gap phenomenon is selected by diversifying the number of topics K and the topic-linking method. The selected model is used to optimize the time gap of multiple resources in the final process.
(6) *Optimizing the time gaps between multiple resources*: Optimizing time gap means finding the point where the sum of similarities between multiple resources is maximized. This can be identified by sequentially moving the timeline of multiple resources. Finally, the time gap estimation for multiple resources are calculated through this optimization process (Fig. 1).

3.1.2. Data collection

For data collection, papers and patents were collected from S&T portal sites, NDSL (http://www.ndsl.kr/) for more than 14 and 12 years, respectively. The document surrogate such as abstract and title was collected from papers, proceedings, and patents whereas the article body was collected from Web news articles. All the resources were built into time-series data divided quarterly. Papers were selected by using a self-classification scheme of NDSL based on authoritative journal indexes such as Science Citation Index (SCI), Social Science Citation Index (SSCI), and Science Citation Index Expanded (SCIE) by Thomson Reuter, and SCOPUS by Elsevier. US patents collected from USPTO were used for patent data. International Patent Classification (IPC), a patent classification scheme, was used to divide them into the two subject areas. The entire "G06*" and "A61*" code series were used as the codes for computer science and medical science respectively. Web news

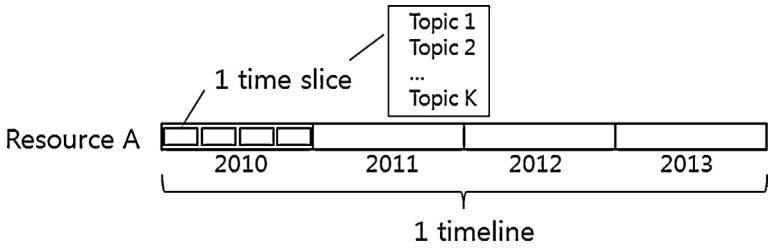

Fig. 1. Concept of the relationship between topic, time slice, timeline, and resource.

Table 1
The most frequent terms extracted from paper resource.

Computer science		Medical science	
Terms	Frequency	Terms	Frequency
System	64,515	Patient	366,861
Model	54,216	Group	129,715
Problem	51,639	Cell	125,250
Algorithm	51,467	Treatment	88,237
Method	48,676	Level	83,456
Result	40,872	Conclusion	82,036
Data	36,901	Effect	75,998
Approach	34,261	Disease	71,377
Network	32,562	Factor	60,219
Time	28,356	Analysis	60,188
New	24,227	Risk	59,792
Application	23,404	Rate	58,882
Information	23,038	Control	58,040
Performance	21,685	Clinical	56,154
Solution	21,248	Significant	54,358
Technique	20,315	Age	51,777
Function	19,964	Activity	50,964
Process	19,833	Time	50,920
Analysis	19,301	Expression	47,228
Image	18,887	Cancer	46,638

Table 2
Data collection statistics for academic subjects and resources.

Resource	Papers	Proceedings	Patents	Web news	Total counts
Computer science	86,922	37,977	607,434	11,232	743,565
Medical science	201,674	15,384	468,268	14,713	700,039
Total count	288,596	53,361	1,075,702	25,945	1,443,604

articles for more than 12 years were also collected from eight representative newspapers[2] using a Web crawler. As for Web news, categories and subcategories were used to identify the subject fields as much as possible. For unclassified data or data with broad concepts such as "tech" or "technology," the term-frequency dictionary extracted from previously classified papers was additionally used for subject classification of Web resources (Table 1). Apart from three representative resources, proceedings for about three years were collected. Proceedings were used only for pre-validation to evaluate performance of the proposed methods. Because there was no standard classification scheme such as International Standard Serial Number (ISSN) or Dewey Decimal Classification (DDC) codes available for proceedings, they were manually classified by reviewing the title and the keyword fields referring to the term-frequency dictionary extracted from papers. We attempted to maintain the field coherence between the various resources; (1) by making dictionary with the frequent terms from the already classified paper resources, (2) by applying this dictionary to the other resources (Table 2).

To compare trends between academic subject fields, a total of 1.443 million documents were collected from the fields of computer science and medical science. In computer science, ICT-related issues and social interest have increased, mainly with the introduction of smart devices such as smart phones and tablet PCs and their application technologies. In medical

[2] IDC Press Release (http://www.idc.org/idc/news/releases), The New York Times (http://www.nytimes.com/), BBC (http://www.bbc.co.uk/), Fox News (http://www.foxnews.com/), USA Today (http://www.usatoday.com/), CNN (http://edition.cnn.com/), Korea IT News (http://english.etnews.com/), and TechNewsWorld (http://www.technewsworld.com/).

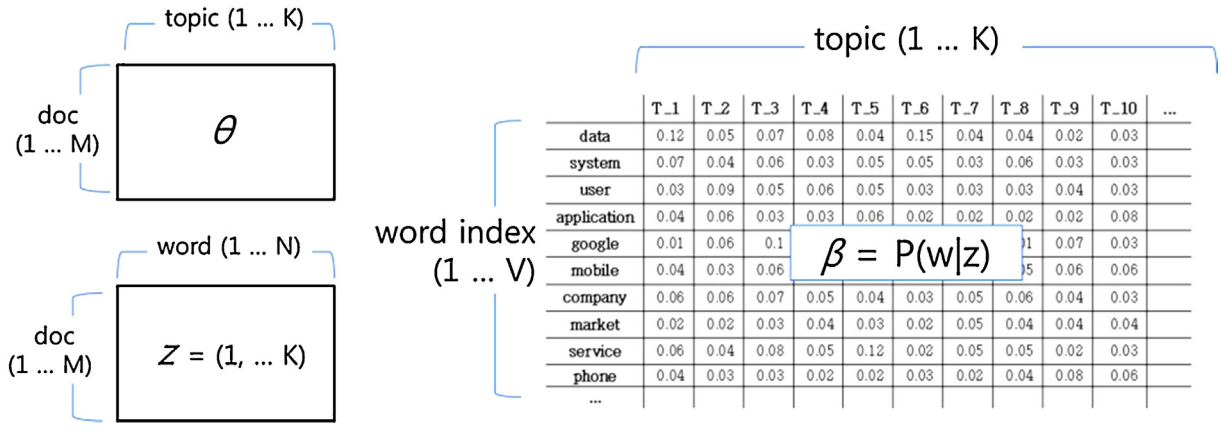

Fig. 2. Relationship of elements (parameter β, distribution θ, topic index Z, word, topic, and doc) for LDA.

Table 3
A comparison of the length of document by data resource and academic area.

Academic area	Avg. # of words of papers (abstract)	Avg. # of words of patents (abstract)	Avg. # of words of Web news articles (body)
Computer science	75.5	62.1	287.3
Medical science	112.9	49.7	366.4

science, social interest in biomedical and healthcare issues has increased recently. Thus, these are the representative S&T fields that generate numerous Web news articles and exhibit recent academic and commercial R&D.

3.2. LDA topic modeling

3.2.1. Latent Dirichlet Allocation

LDA is a widely used probabilistic generative topic modeling technique that models documents using mixtures of topics. In LDA, word w is generated by the probability distribution $p(z)$ of topic index z for word w. If one knows the probability $p(w|z)$ that a certain word will be generated by topic, the probability $p(w)$ of generating a document can be obtained (Blei et al., 2003).

When a document (a set of observable words w) is given and the Dirichlet parameters α and β are known, calculating the posterior distribution for the hidden variable θ and z is the key process in the LDA model. Topic index z related to individual word w is determined by the multinomial distribution θ for the topic ($z|\theta \sim Multinomial(\theta)$), and θ is determined by the Dirichlet distribution with parameter α ($\theta|\alpha \sim Dirichlet(\alpha)$). In other words, the probability $p(\theta_d|\alpha)$ of θ for each document determined by α is expressed with the K-dimensional Dirichlet distribution, where K is the number of topics.

The Dirichlet distribution is a continuous probability distribution and the conjugate prior for a multinomial distribution in Bayesian statistics; thus, it is frequently used to obtain prior distributions in Bayesian statistics. Therefore, to infer θ in the LDA method, the Dirichlet distribution is selected as the prior distribution for easier computation of the posterior. In addition to α, the Dirichlet parameter β is a $k \times V$ matrix containing the generative probability of words w for topic index z, where V is an index that shows the word position (Fig. 2).

Hence, LDA topic modeling is divided largely into two processes: (1) choosing z from the θ distribution as explained earlier, and (2) choosing w from the Dirichlet distribution with parameter β. The multinomial distribution φ_z of word w for topic z is inferred from the Dirichlet distribution with parameter β ($\varphi|\beta \sim Dirichlet(\beta)$), and word w can be obtained from φ_z ($w|\varphi \sim Multinomial(\varphi)$). However, because the coupling problem occurs between θ and β, the parameters γ and φ are added and the model needs to be modified. A process of estimation is also needed to obtain θ and z using γ and φ (Fig. 6, right side). By explaining the process of obtaining γ and φ as the expectation step (E-step) and the process of inferring α and β as the maximization step (M-step), Blei et al. (2003) suggested using the variational Expectation–Maximization (EM) algorithm to infer the LDA parameters. This study used the variational EM algorithm to infer the parameters. Gibbs sampling which is easy to implement has also been widely attempted recently (Wang & Sun, 2013; Xu et al., 2012).

3.2.2. Word extraction

Word extraction is the most basic task in the topic modeling process. We extract noun phrases with NLP to collect technological terms more precisely. The NLP package software NLTK for Python (http://www.nltk.org/) and the Penn Treebank Tags (http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html) were used for Part-of-Speech (POS) tagging. The Regular Expressions (RE) were also used to extract all types of noun phrases including prepositions or composed of nouns and adjectives to extract all the noun phrases as well as single noun. The statistics of extracted words are shown in Table 3.

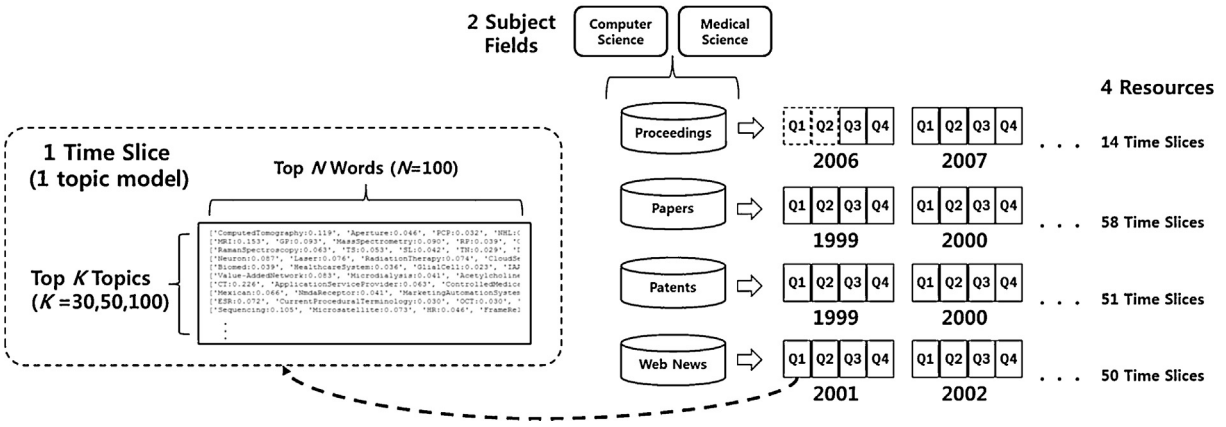

Fig. 3. Time slices constructed from multiple resources and inner structure of the time slice.

3.2.3. Generating time slices with LDA

A document composed of words was conceptualized by LDA topic modeling. All the resources were constructed for the temporal analysis experiment. Time-series information was also considered for the two academic fields; the entire data structure of topic modeling is shown in Fig. 3. All the time-series data were built quarterly, and each year had four time slices (Quarters 1–4). The time slices generated by LDA were composed of multiple topics. Each topic was expressed by the words and their probability value. In this study, a topic was composed of the top 100 words. For the initial experiment, three models such as T-30, T-50, and T-100 were generated with 30, 50, and 100 topics respectively. In case that more than 100 topics were generated, over-generation of topics led to redundancy of topics within a single time slice; therefore, modeling that exceeded $K = 100$ was not conducted. A total of 1038 topic slices were generated (348 for papers, 84 for proceedings, 306 for patents, and 300 for Web news articles) by using the open source toolkit, Gensim (http://radimrehurek.com/gensim/).

3.3. Method of the time gap analysis

3.3.1. Measuring topic similarity

This section explains how to generate semantic linkages between two heterogeneous resources. A time slice, the unit of time-series information that makes up the timelines of resources, is composed of many topics. The similarity between two times slices needs to be identified to determine the time gap. To realize this, the similarity of two topics must be measured. The Cosine similarity and KL (Kullback–Leibler) divergence are the main methods widely used to measure the topic similarity (He et al., 2009; Newman, Asuncion, Smyth, & Welling, 2009). We select the Cosine similarity coefficient to measure the similarity between topics. Fig. 4 illustrates the linking method of generating topic linkages between time slices, A and B. The proposed many-to-many (M-to-M) linking method allows all connections between topics whose similarity exceeds a certain threshold. We divide the connections into three groups depending on the topic similarity. Quartile, which is a widely used type of quantile, is used for the linking method. When the minimum value of a sample group is 0 and the maximum value is 1, Min (minimum), $Q1$ (1st quartile), Median (2nd quartile), $Q3$ (3rd quartile), and Max (maximum) refer to 0, 0.25, 0.5, 0.75, and 1, respectively. The first group with the highest similarity is called 'highly similar (*HS*)' group. After creating the $k \times k$ links, all the linkages in the range of $Q3 \sim$ Max are included in this group. The second group has medium level of similarity and is called the 'similar (*S*)' group. This group corresponds to the range of $Q1 \sim Q3$. Finally, the group with less than $Q1$ is classified as the 'not similar' group and is excluded when the sum of similarity is calculated. Therefore, the processing time requires $k \times k$ calculations for all the possible linkages $f(x)$, and it has O(n2) complexity. Algorithm 1 is a pseudo code explaining the process of the linking method.

Algorithm 1 'M-to-M' linking method-based similarity measurement algorithm between two time slices

input: Time_Slice_A, Time_Slice_B #with many topics

main:
1 #step 1: calculating all topic combination between input Time Slices
2 for Topic_i in Time_Slice_A:
3 put Topic_i in Stack_List_1
4 for Topic_j in Time_Slice_B: #iteration, i x j times
5 put Topic_j in Stack_List_2
6 calculate cosine sim (Topic_i, Topic_j) #topic consists of vectors
7 put calc result with Topic pairs in Stack_List_3 #for selecting proper pairs
8
9 #step 2: creating topic links and final similarity value

10 calculating 1st Quartile, 3rd Quartile from Stack_List_3 #calculating quartiles
11 for Topic_Pair_k in Stack_List_3 #iteration, i x j times
12 if sim value of Topic_Pair_k > = threshold #threshold: 1st Qu. or 3rd Qu.
13 add sim value of Topic_Pair_k to Final_Sim_Value
14 put Topic_Pair_k in Topic_Link_List
15 else
16 pass
17 #end of algorithm

output: Final_Sim_Value, Topic_Link_List between Time_Slice_A, Time_Slice_B

When analyzing the time gaps of resources A and B, the *i*th time slice of resource A is expressed as TS_i^A, and the *j*th time slice of resource B is represented by TS_j^B. Here, the similarity $TimeSliceSim(TS_i^A, TS_j^B)$ between the specific time section TS_i^A of resource A and another specific time section TS_j^B at a certain distance from TS_i^A can be measured as in Formula (1). Formula (1) shows how to measure the similarity between the time slices explained in Algorithm 1. Formula (1) yields *Topic_Pair_k*, the final *K* topic pairs stored in *Topic_Link_List*.

$$TimeSliceSim(TS_i^A, TS_j^B) = \sum_{k=1}^{K} TopicSim(Topic^A, Topic^B) \tag{1}$$

The sum of *k* similarity is the maximum similarity value that can be obtained from two time slices. Therefore, this maximum output value becomes the final similarity value between two time slices and corresponds to the *Final_Sim_Value* of the *output* lines of Algorithm 1. Here, $TopicSim(Topic^A, Topic^B)$ is the core function that measures the similarity between two topics. M-to-M connection allows all connections with a similarity above a certain threshold (Formula (2)).

$$TopicSim(Topic^A, Topic^B) = \sum_{k=1}^{K} f(s)^{if.f(s) \geq threshold}, \quad f(s) = \sum_{k=1}^{K} \cos(topic_k^A, topic_{all}^B) \tag{2}$$

3.3.2. Optimizing and estimating the time gap

The similarity of the timelines can be measured by using the previously calculating method of measuring the similarity between time slices. In other words, the method repeatedly measures the similarity between time slices while moving the timeline to calculate the time gap between two resources. Formula (3) explains that $TimelineSim(R^A, R^B)$ calculates the similarity between the two timelines. $TimelineSim(R^A, R^B)$ is expressed as the sum of the similarities between the time slices;

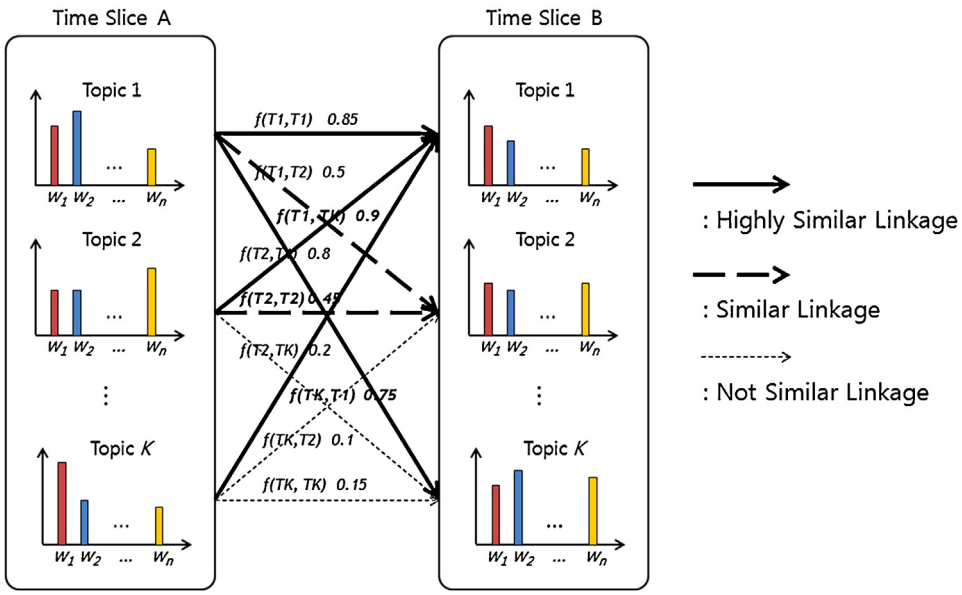

Fig. 4. An example of measuring similarity between two time slices.
Adapted from Xu et al. (Xu et al., 2012).

in other words, the total sum of $TimeSliceSim(TS_i^A, TS_j^B)$ in Formula (1). The value that is normalized by the total number of calculated pairs N is set as the final value.

$$TimelineSim(R^A, R^B) = \sum^{N} \frac{TimeSliceSim(R^A, R^B)}{N} \tag{3}$$

The function $TimeGapDecision(R^A, R^B)$ in Formula (4) ultimately determines the time gap between two resources. The time point with the highest value among all the values of $TimelineSim(R^A, R^B)$ is repeatedly measured while moving sequentially along the time axis within the time range. Then the selected point is determined as the final time gap, and the value of that point is returned. In this experiment, the time range is set to $i-14 \le j \le i+14$. Here, cases having time slices that do not match at the front and back by the time shifts are not used, and only the results where the entire section exists are used for analysis. The trend line of the time gap can be drawn using all the measured time gap values, which are obtained from $TimelineSim(R^A, R^B)$. A time gap optimization is conducted using the Formula (4). $TimeGapDecision(R^A, R^B)$ finds the one point as an optimized the time gaps between resources. All the time gap analyses belong to one of the three following types.

$$TimeGapDecision(R^A, R^B) = argmax_{i,j \in Range} TimelineSim(R_i^A, R_j^B) \tag{4}$$

The optimization formula *MultiResourceSim* (*R*1,*R*2,*R*3) for inferring the time gap of multiple resources is composed of three independent parts. Because the final similarity value is the result of a linear combination, the similarity values for *R*1–*R*2, *R*2–*R*3, and *R*1–*R*3 need to be calculated respectively. Among the sum of the three parts obtained by voting multiple values, the maximum value is determined as the optimized result of the time gap estimation processing. It can be expressed simply as in Formula (5). The time distance of the longest axis for *R*1–*R*3, is equal to the sum of the time distance *R*1–*R*2, and the time distance *R*2–*R*3. For example, if the time distance of *R*1–*R*3 is 3*Q*, because *R*2 moves in between them, four cases are measured: *R*1–*R*2 is 0*Q* and *R*2–*R*3 is 3*Q* (3*Q* = 0*Q* + 3*Q*); *R*1–*R*2 is 1*Q* and *R*2–*R*3 is 2*Q*; *R*1–*R*2 is 2*Q* and *R*2–*R*3 is 1*Q*; and *R*1–*R*2 is 3*Q* and *R*2–*R*3 is 0*Q*. As the longest distance *R*1–*R*3 increases from 1*Q* to 14*Q*, the number of cases where the *R*2 moves exponentially increases. Because each part of Formula (5) has a different range of similarity according to the resource, normalized values are used. The value that is divided by the maximum of measured values by resource can be used to normalize the range of similarity values. In addition, weights of α, β, and γ can be given to parts according to the confidence. In this experiment, all the weights are simply set to 1 ($\alpha = \beta = \gamma = 1$) assuming that all the parts have the same weight.

$$MultiResourceSim(R1, R2, R3) = argmax \sum \left(\alpha * \frac{TimelineSim(R1, R3)}{argmax\, TimelineSim(R1, R3)} + \beta * \frac{TimelineSim(R1, R2)}{argmax\, TimelineSim(R1, R2)} + \gamma * \frac{TimelineSIm(R1, R3)}{argmax\, TimelineSim(R2, R3)} \right) \tag{5}$$

4. Experiments for the time gap analysis

4.1. Interpretation of the time gap phenomenon

In this chapter, the time differences between main resources are analyzed using topic modeling. First, it is necessary to choose a standard modeling method with consistent performance. In this research, appropriate methods are chosen by comparing proceedings and papers. Proceedings have the characteristics presenting a research result through a conference and quickly introducing the research to the public. Therefore, with the premise that proceedings precede papers, the best modeling result can be chosen. The timeline for proceedings comprises 14 time slices from *Q*3 in 2006 to *Q*4 in 2009, and the timeline for papers comprises 58 time slices from *Q*1 in 1991 to *Q*2 in 2013. Continuously move the proceedings timeline +1*Q* to the right and adversely to −1*Q* to the left, centering on *T* + 0*Q* which is the standard time. Then, measure the similarity between *T* + 0*Q* of proceedings and the shifted timeline of papers, and describe the series of results as a trend line. Consequently, it is explained that proceedings precede papers by the fact that the trend line has a high degree of similarity in the right area of the vertical standard line (blue vertical line) based on the horizontal standard line (blue horizontal line) that crosses the time point, *T* + 0*Q* (Fig. 5).

Analyzing two topic-linking methods (HS− and HS+S-based) and three models according to the topic number *k* (*T*-30, *T*-50, and *T*-100) by two academic fields (computer and medical sciences) separately, a total of 12 result models are gained to compare with each other, and the final topic-linking method and topic number k are decided. As a result, two things are shown. Firstly, for the left standard graph of Fig. 5, we selected the HS-based T-50 model showing a steady analysis performance. We observed that the graphs of all models with the linking method and the topic number are very similar to each other, and this means all models show the similar performance. Secondly, in every graph, the similarity variance is bigger in the medical science than in computer science, but it is also well explained that proceedings precede papers (Fig. 6).

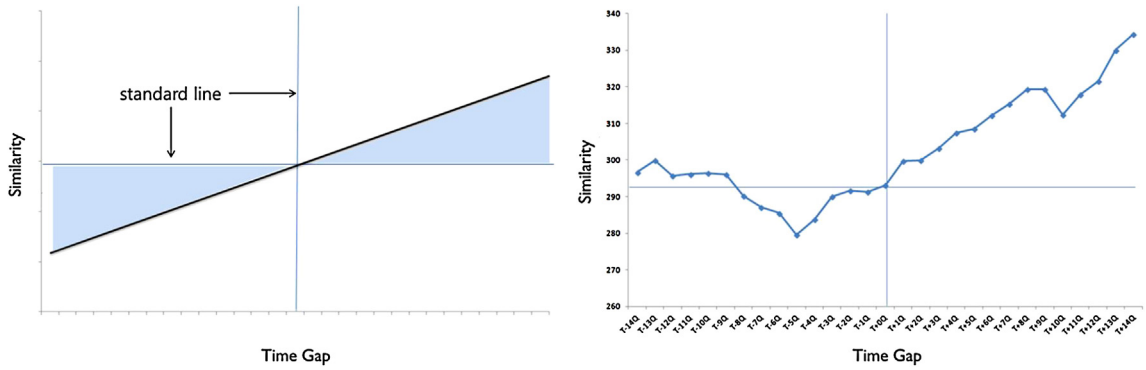

Fig. 5. An explanatory (left) and a measured (right) graph explaining phenomenon that proceedings precede papers (using linking method based on HS with T-50).

4.2. Time gap analysis between two resources

Considering the preliminary tests described above, a comprehensive result of the analysis between two resources is as follows. The comparison between Web-PP and between Web-PT shows that Web resources are the most current. It is also shown that patents are more current than papers, and the same tendency commonly appears in the computer science and medical science. Thus, if patents and papers are written at the same time, it is suggested that patents will be opened to the public earlier than papers. The results from all experiments are combined, and therefore, three kinds of resources can be arranged from most to least current as "*Web news (Web)* → *Patents (PT)* → *Papers (PP)*" (Fig. 7).

4.3. Optimizing the time gaps between multiple resources

Using the optimization formula, Formula (5) (*MultiResourceSim* (*Web*,*PP*,*PT*)), interpretation of the time gap phenomenon for three kinds of resources in two academic fields is conducted separately. Table 4 is the measured result of all possible movement of patents between Web news articles and papers. The shaded area represents the maximum value decided by optimizing and estimating the time differences.

The final results are shown in Fig. 8. As a result, it is found that there are 11Q of difference between Web news articles and papers, 7Q between Web news articles and patents, and 4Q between patents and papers in the computer science field. Further, in the medical science field, it is found that there are 7Q of difference between Web news articles and papers, 4Q between Web news articles and patents, and 3Q between papers and patents. Finally, as the time axis is shorter in medical science than in computer science, it is demonstrated that both patents and papers are more up-to-date in the medical science field compared with the computer science field. Through the experiments discovering the time gap phenomenon, we have obtained the exact time intervals between heterogeneous multiple resources by academic area.

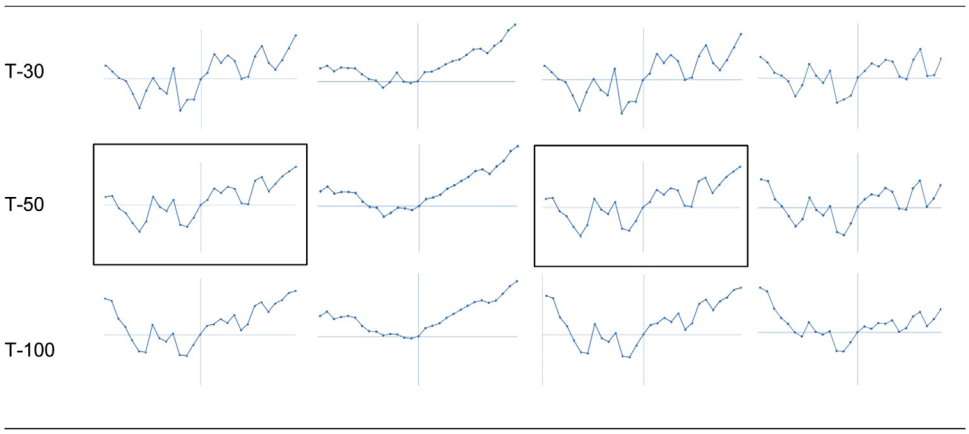

Fig. 6. The comprehensive comparison of analysis results by the combination of various modeling methods (*X* and *Y*-axis of each graph mean 'Time Gap' and 'Similarity' between two resources respectively as shown in Fig. 5.).

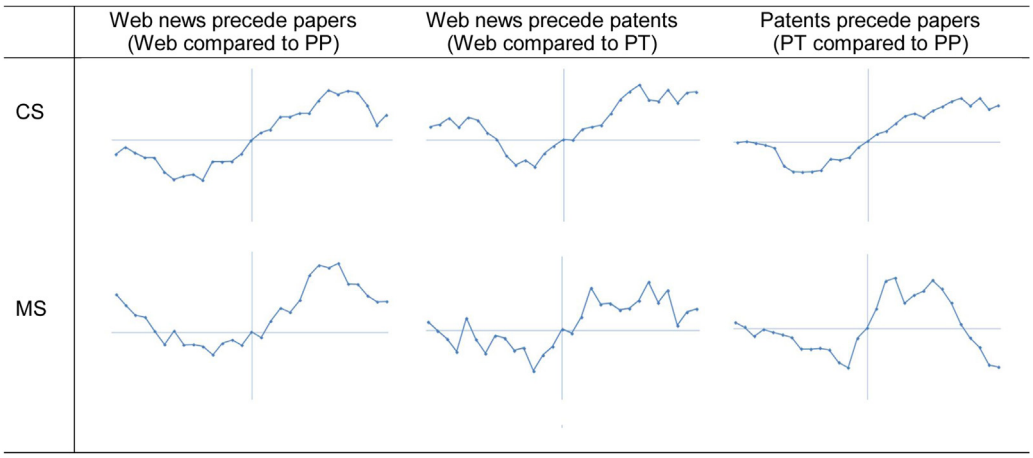

Fig. 7. The analyzed time gap phenomenon of multiple resources by academic area (*X* and *Y*-axis of each graph mean 'Time Gap' and 'Similarity' between two resources respectively as shown in Fig. 5.).

Table 4
The results of the time gap analysis and optimization of multiple resources (showing only estimation part).

Web news-Papers (Longest)		Web news-Patents		Patents-Papers		Final Sim. value ($X+Y+Z$)
Time gap (a)	Sim. value (X)	Time gap (b)	Sim. value (Y)	Time gap ($a-b$)	Sim. value (Z)	
(a) Computer science						
+10Q	1.000	+8Q	1.000	+2Q	0.965	2.965
		+9Q	0.985	+1Q	0.962	2.947
		+10Q	0.983	+0Q	0.955	2.938
+11Q	0.998	+0Q	0.946	+11Q	0.992	2.936
		+1Q	0.945	+10Q	1.000	2.943
		+2Q	0.956	+9Q	0.997	2.951
		+3Q	0.958	+8Q	0.991	2.947
		+4Q	0.960	+7Q	0.987	2.945
		+5Q	0.972	+6Q	0.980	2.950
		+6Q	0.985	+5Q	0.984	2.967
		+7Q	0.993	+4Q	0.981	2.972
		+8Q	1.000	+3Q	0.973	2.971
		+9Q	0.985	+2Q	0.965	2.948
		+10Q	0.983	+1Q	0.962	2.943
		+11Q	0.995	+0Q	0.955	2.948
+12Q	0.989	+0Q	0.946	+12Q	1.000	2.935
		+1Q	0.945	+11Q	0.992	2.926
		+2Q	0.956	+10Q	1.000	2.945
(b) Medical science						
+6Q	0.990	+5Q	1.000	+1Q	0.973	2.963
		+6Q	0.996	+0Q	0.956	2.942
		+10Q	0.983	+0Q	0.955	2.938
+7Q	0.998	+0Q	0.973	+7Q	0.998	2.969
		+1Q	0.970	+6Q	0.989	2.957
		+2Q	0.979	+5Q	0.985	2.962
		+3Q	0.996	+4Q	0.978	2.972
		+4Q	0.999	+3Q	1.000	2.997
		+5Q	1.000	+2Q	0.997	2.995
		+6Q	0.996	+1Q	0.973	2.967
		+7Q	0.985	+0Q	0.956	2.939
		+8Q	1.000	+3Q	0.973	2.971
		+9Q	0.985	+2Q	0.965	2.948
		+10Q	0.983	+1Q	0.962	2.943
		+11Q	0.995	+0Q	0.955	2.948
+8Q	0.996	+0Q	0.973	+8Q	0.990	2.959
		+1Q	0.970	+7Q	0.998	2.964
		+2Q	0.979	+6Q	0.989	2.964

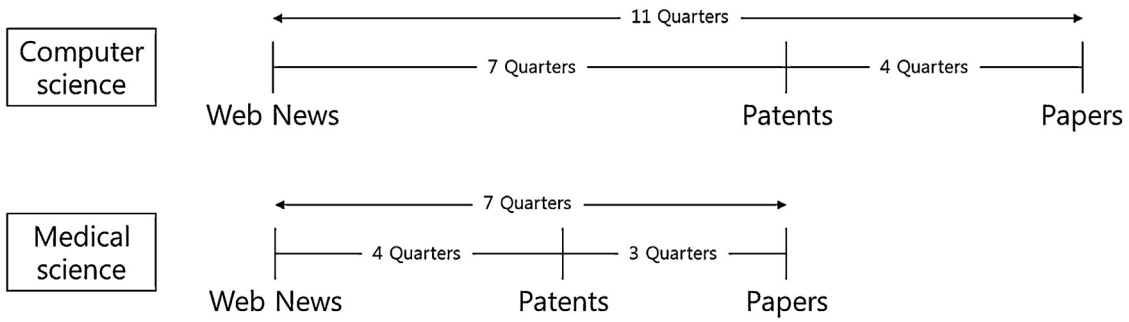

Fig. 8. The results of optimizing and estimating the time gap between multiple resources.

5. Evaluation methods for the temporal analysis

5.1. Statistical evaluation based on power-law exponent

With the previous experiments, it was confirmed that there exists the time difference between resources. For the results of the time gap analysis, comparison of the changes in term distribution before and after the time gap estimation is a statistical evaluation method used in this chapter. Changes in term distribution can be estimated by measuring changes of term distribution in multiple resources.

As a method for measuring changes in term distribution in documents, Lee and Jeong (2008) analyzed the distribution characteristics of index terms through the study of usage patterns of folksonomy tags with a power-law exponent. The exponent value can be calculated by comparing frequency of index terms with the number of each unique term. Therefore, the usage of the exponent can confirm the purpose of this research, which is to gather similar topics through estimation of the time differences of multiple resources, and its effectiveness. Statistics of all the measurements are explained in the form of a box plot chart in Fig. 9. *Y*-axis in Fig. 9 denotes the calculated value of the power-law exponent used as an evaluation method. At first, it can be seen that average, median, 1st and 3rd quartiles, and the minimum and maximum values are reduced after the time gap estimation in the both fields. There are no statistical outliers in the medical science field, and the outliers in the computer science field show that the deviation is remarkably reduced after the estimation (Fig. 9).

In conclusion, the result of measuring the term distribution based on the power-law exponent shows that even if the distribution of the terms was widely distributed before estimating the optimal time gap, the number of unique terms in the integrated resource increased after the time gap estimation.

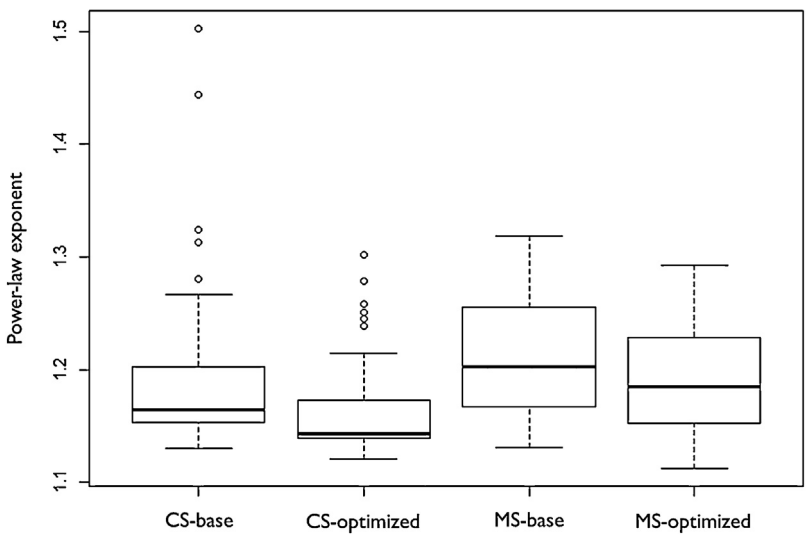

Fig. 9. The time gap optimization with multiple resources represented by box plot chart (low values are better).

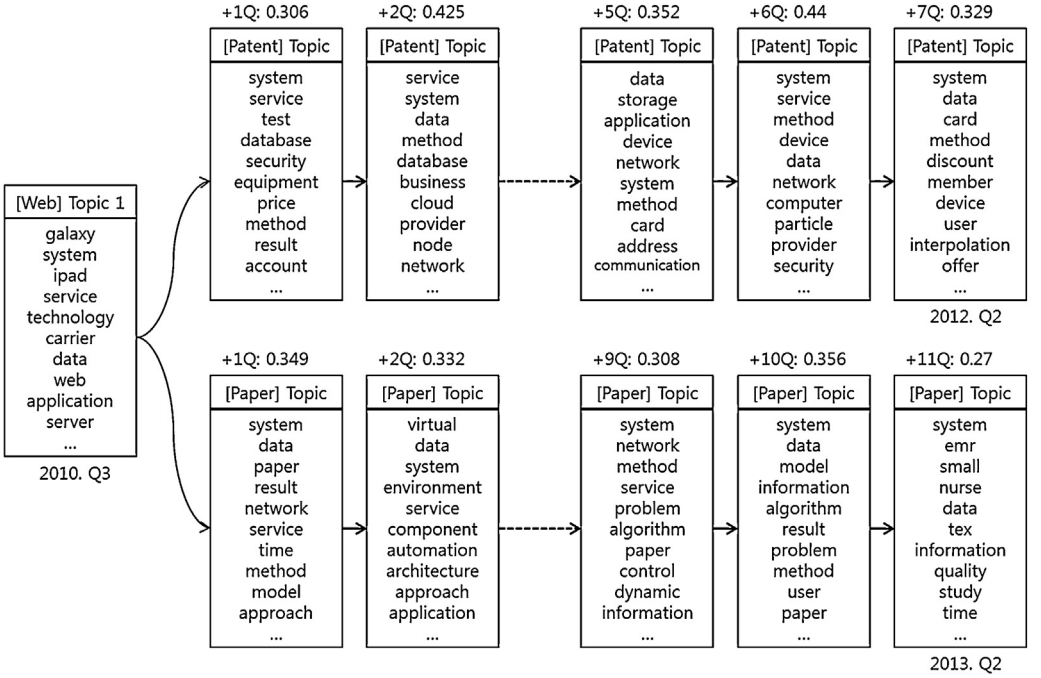

Fig. 10. Temporal flow and similarity changes of topics along with patents and papers in computer science.

5.2. Evaluation based on the content analysis

Evaluation of topic modeling has been a matter of debate in research. Chang, Boyd-graber, Gerrish, Wang, and Blei (2009) noted that the human-based practical assessment for the topic extraction was relatively absence in the topic modeling method. Kim and Oh (2011) conducted research on the technique of generating topic chains, and suggested an analysis method for evaluating the topics by interpreting topic changes over time. In this chapter, in order to interpret the meaning of the topic content, we select the first topics from representative topics of Web news articles. Next, the most similar topics of patents and papers are chosen respectively by comparing to the Web resource, and then the similarity changes are examined according to time flow.

Fig. 10 shows the changes in content of topics in computer science according to the resource-specific differential. First, we selected the most representative topic, which is related to "mobile advertising service," of Web news articles in the third quarter of 2010. Next, similarities between adjacent topics are measured along with the time flow of patents and papers separately. As a result, topic similarity of patents sharply decreases to 0.329 at the position of +7*Q*. In other words, this is a point where the meaning of the topic changes most. The words, "system," "service," "database," "data," etc. that begin with Web news articles persist to +2*Q* of patents, but change to "device," "network," "data," "system," "method," etc. However, major changes do not appear in the main words. In the topic of +7*Q*, a number of new words which are not seen in previous topics appear including "card," "discount," "member," "user," "interpolation," and "offer." The similarity between Web news articles and patents drops in +7*Q*. The major topics addressed in Web news articles and papers are described as follows; there is a shift of content from "system," "data," "service," "application," "approach," etc. to "system," "method," "algorithm," "paper," etc., and then followed by the surge of new words such as "EMR," "nurse," "tex," "quality," and "time." Through the flow of topics with the highest similarity, papers show abrupt semantic changes at the +11*Q* spot. This is close to the result of the time gap analysis of Fig. 8 in Section 4.

Fig. 11 shows the comparison result with the adjacent topic similarities between Web news articles and other resources. Taking average and median of the similarity values for 10 major topics addressed in patents and papers, we obtained that patents and papers incurred the largest shifts in content at +8*Q* and +11*Q* respectively.

In medical science, the most representative topic is related to "stem cell research for the treatment of patients." The major words of the topic of Web new articles, "cell," "method," "invention," "acid," and "cancer" continually appear for a while, and then words such as "wearable," "embodiment," "base portion," etc. newly appear in +5*Q* point. At the time point, similarity is shown in the lowest state. In the flow of papers, a topic "carotid disease" associated with the words "mouse," "treatment," "carotid," "disease," and "control" newly emerges in +8*Q*. Changes in the topic started from the Web news article prominently appear in +8*Q* of papers and in +5*Q* of patents (Fig. 12). On average of 10 topics, changes in the largest sense are also shown in +5*Q* of patents and in +8*Q* or +9*Q* of papers (Fig. 13).

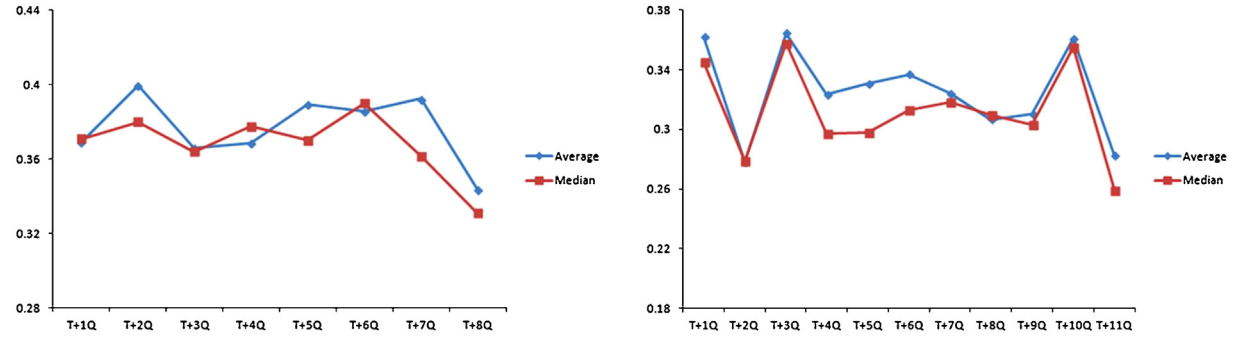

Fig. 11. Temporal trends of the adjacent similar topics between Web news articles and papers in computer science using average and median of 10 major topics (left: patents, right: papers).

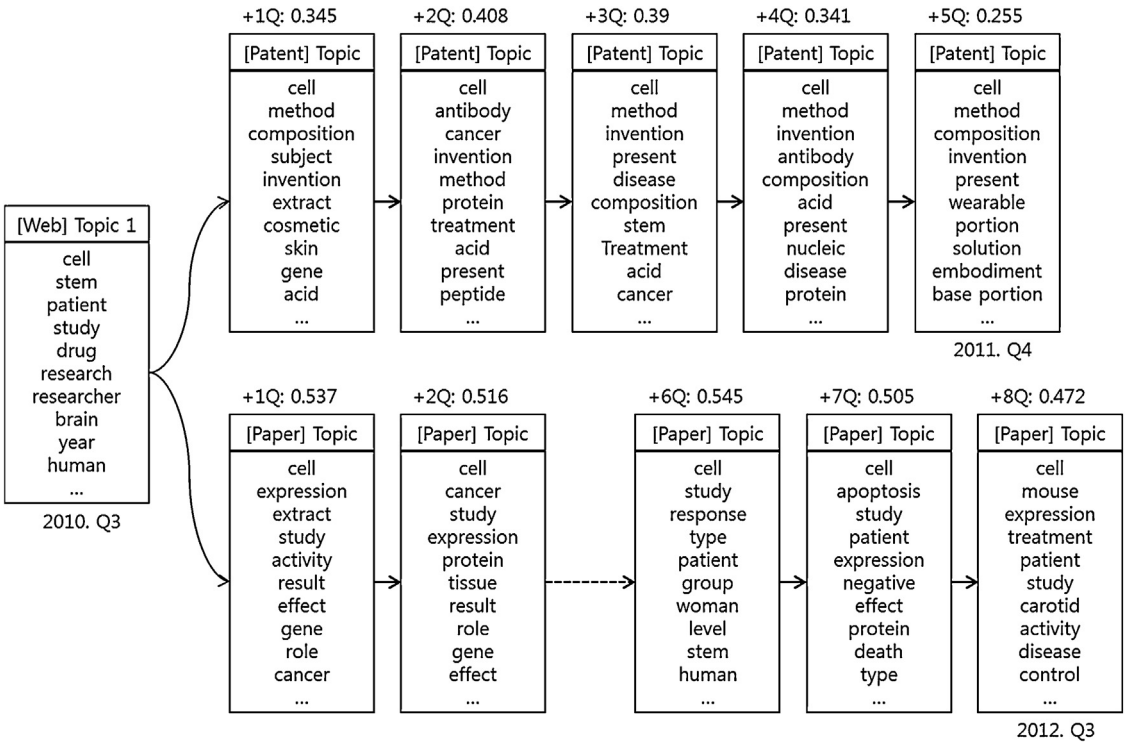

Fig. 12. Temporal flow and similarity changes of topics along with patents and papers in medical science.

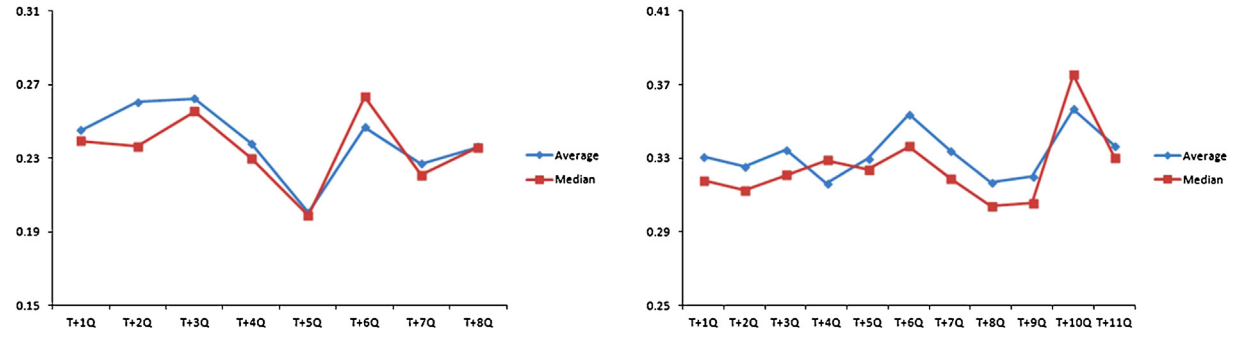

Fig. 13. Temporal trends of the adjacent similar topics between Web news articles and papers in medical science using average and median of 10 major topics (left: patents, right: papers).

As above, comparing the results through the content analysis of topics in the fields of computer and medical sciences, Web news articles and papers show that a significant change takes place in the meaning of the topic in +11Q of the computer science and in +8Q or +9Q of the medical science. In addition, in the case of Web news articles and patents, a sudden change of topic appears in +8Q of computer science and in +5Q of medical science. This confirms that the time gap interval in the medical science field is shorter than that of the computer science field, which is consistent with the result in Section 4, and human-based interpretation about topics explains it well.

6. Conclusion

In this paper, we aimed to reveal the time gap phenomenon of academic areas by estimating the time intervals between multiple resources. To this end, Latent Dirichlet Allocation (LDA), a topic modeling technique highlighted in the text mining area as of late, served the base for the new analysis approach. All the resources were conceptualized using LDA, and a time gap analysis was conducted by measuring and comparing topic similarity between heterogeneous resources. Furthermore, the results of the time gap optimization and estimation were derived via the experiment.

The interpreted results were then validated using two evaluation methods; first, through measuring the clustering tendency of terms based on statistical methods, and second, through content-centered analysis on the topic modeling technique. An evaluation method of topic modeling was employed to identify how the topic's subject changed according to time flow.

The results of the time gap analysis are summarized as follows. First, topic modeling was effective in determining the content characteristics and time-series trends of papers, patents, and Web news articles; we found that the noteworthy time gap phenomenon was revealed using the time gap analysis based on topic modeling. As revealing the precise time gap phenomenon, we obtained how much major topics of the medical science field were shorter than those of the computer science field. This finding showed that the resources of the medical science field had more up-to-date property than those of the computer science field, and thus prompter disclosure to the public (papers are faster one year, patents three quarters). Second, to evaluate the effectiveness of the proposed method, we measured the effect of the analysis via computational methods by using a power-law exponent measurement, and analyzed how the representative topics of Web news articles explained well the time gap phenomenon as the time flows through papers and patents.

The proposed method contributes to an in-depth understanding of various resources by interpreting content characteristics and time-series trends. It also enables to measure the exact time differences between academic areas by understanding the time gap phenomena in an integrated manner. In addition, by proposing a power-law exponent measurement and content analysis of representative topics, this study demonstrates these two methods were effectively applied to temporal analysis.

In the future, we plan to conduct multifaceted analyses by combining quantitative analysis methods and text mining-based approaches based on the proposed method. Furthermore, we will utilize this precise temporal analysis method to improve the performance of trend analysis and future prediction for the practical studies.

Acknowledgements

This study was supported by National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-2012S1A3A2033291).

References

Amitay, E., Carmel, D., Herscovici, M., Lempel, R., & Soffer, A. (2004). Trend detection through temporal link analysis. *American Society for Information Science and Technology*, *55*(14), 1270–1281.

Ball, R., & Tunger, D. (2006). Bibliometric analysis – A new business area for information professionals in libraries? *Scientometrics*, *66*(3), 561–577.

Björneborn, L., & Ingwerson, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, *55*(14), 1216–1227.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning (ICML)* (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chang, J., Boyd-graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 1–9.

Chua, A. Y. K., & Yang, C. C. (2008). The shift towards multi-disciplinarity in information science. *Journal of the American Society for Information Science and Technology*, *59*(13), 2156–2170.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, *73*(8), 981–1012.

Egghe, L. (2005). Expansion of the field of informetrics: Origins and consequences. *Information Processing and Management*, *41*(6), 1311–1316.

Finardi, U. (2011). Time relations between scientific production and patenting of knowledge: The case of nanotechnologies. *Scientometrics*, *89*(1), 37–50.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*, 108–111.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *1*(101), 5228–5235.

Guan, J., & Zhao, Q. (2013). The impact of university-industry collaboration networks on innovation in nanobiopharmaceuticals. *Technological Forecasting and Social Change*, *80*(7), 1271–1286.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: How can citations help? In *Proceedings of the 18th ACM conference on information and knowledge management (CIKM '09)* (pp. 957–966).

Kim, D., & Oh, A. (2011). Topic chains for understanding a news corpus. *Computational Linguistics and Intelligent Text Processing*, 163–176.

Kim, J., Hwang, M., Jeong, D. H., & Jung, H. (2012). Technology trends analysis and forecasting application based on decision tree and statistical feature analysis. *Expert Systems with Applications*, *39*(12), 618–12625.

Lancaster, F. W., & Lee, J.-L. (1985). Bibliometric techniques applied to issues management: A case study. *Journal of the American Society for Information Science and Technology*, *36*(6), 389–397.
Lee, J. Y., & Jeong, D. H. (2008). An Analysis on the Tag Usage Statistics in Folksonomy: Considering Controlled and Uncontrolled Vocabularies. In *Proceedings of the 15th conference of korean society for information management* (pp. 21–26).
Lee, K., & Lee, S. (2013). Patterns of technological innovation and evolution in the energy sector: A patent-based approach. *Energy Policy*, *59*, 415–432.
Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology*, *59*(12), 1973–1984.
Lu, Y., Okada, S., & Nitta, K. (2013). Semi-supervised latent dirichlet allocation for multi-label text classification. In *Proceedings of the recent trends in applied artificial intelligence: 26th international conference on industrial, engineering and other applications of applied intelligent systems* (pp. 351–360).
Mei, Q., & Zhai, C. X. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 198–207).
Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. *Research Policy*, *26*(3), 317–330.
Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research*, *10*, 1801–1828.
Ralf, K., & Peter, F. (2012). Reranking Websearch results for diversity. *Information Retrieval*, *15*(5), 458–477.
Sajjad, M., Hwang, M., Kim, J., Gim, J., Song, S. K., Jeong, D. H., et al. (2013). Discovering time difference in technology trends by heterogeneous resources. In *Proceedings of the 12th international conference on business innovation and technology management* (pp. 1–5).
Schoepflin, U., & Glänzel, W. (2001). Two decades of "Scientometrics" – An interdisciplinary field represented by its leading journal. *Scientometrics*, *50*(2), 301–312.
Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Opportunity discovery by assessing the difference between science and technology – Case study of secondary batteries. In *Proceedings of the 5th IEE international conference on management of innovation and technology (ICMIT2010)*
Song, M., Kim, S. Y., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. *Journal of the American Society for Information Science and Technology*, *65*(2), 352–371.
Vaughan, L., & You, J. (2008). Content assisted Web co-link analysis for competitive intelligence. *Scientometrics*, *77*(3), 433–444.
Vretos, N., Nikolaidis, N., & Pitas, I. (2012). Video fingerprinting using Latent Dirichlet Allocation and facial images. *Pattern Recognition*, *45*(7), 2489–2498.
Vulić, I., Smet, W. D., & Moens, M. F. (2013). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, *16*(3), 331–368.
Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 424–433).
Wang, J., & Sun, X. (2013). Unsupervised mining of long time-series based on latent topic model. *Neurocomputing*, *103*, 99–103.
Xu, S., Zhu, L., Qiao, X., Shi, Q., & Gui, J. (2012). Topic linkages between papers and patents. In *Proceedings on advanced computer science and technology (AST)* (pp. 176–183).