Topic Diffusion Analysis of a Weighted Citation Network in Biomedical Literature

Munui Kim
Department of Library and Information Science, Yonsei University, Seoul, 03722, Republic of Korea.
E-mail: munui0822@gmail.com

Injun Baek
Department of Library and Information Science, Yonsei University, Seoul, 03722, Republic of Korea.
E-mail: dolja315@gmail.com

Min Song
Department of Library and Information Science, Yonsei University, Seoul, 03722, Republic of Korea.
E-mail: min.song@yonsei.ac.kr

In this study, we propose a framework for detecting topic evolutions in weighted citation networks. Citation networks are important in studying knowledge flows; however, citation network analysis has primarily focused on binary networks in which the individual citation influences of each cited paper in a citing paper are considered identical, even though not all cited papers have a significant influence on the cited publication. Accordingly, it is necessary to build and analyze a citation network comprising scholarly publications that notably impact one another, thus identifying topic evolution in a more precise manner. To measure the strength of citation influence and identify paper topics, we employ a citation influence topic model primarily based on topical inheritance between cited and citing papers. Using scholarly publications in the field of the protein p53 as a case study, we build a citation network, filter it using citation influence values, and examine the diffusion of topics not only in the field but also in the subfields of p53.

Introduction

The accumulated publication set for most fields in science is relatively large (Liu & Lu, 2012), and a large number of scientific publications are publishing at an enormous rate. As such, it often seems impossible to trace the development of knowledge within a given scientific field without using a quantitative method. Garfield, Sher, and Torpie (1964) suggested that it is feasible to "write the history of science" by analyzing citation relationships between science publications because when a citation occurs, a piece of knowledge in a cited paper is suggested to flow into the citing paper. Accordingly, knowledge evolution can be detectable and describable using citations as a key indicator of knowledge flow.

Many studies (Garfield et al., 1964; Hummon & Doreian, 1989; Liu & Kuan, 2016; Ma & Liu, 2016; Mina, Ramlogan, Tampubolon, & Metcalfe, 2007; Verspagen, 2007) have focused on identifying the development of knowledge in a given scientific field by analyzing citation networks, and many of them have utilized main path analysis, which enables us to find the most significant paths in a given citation network using traversal counts of each citation link, thereby interpreting these paths as knowledge developmental paths under the assumption that a citation link indicates diffusion of a piece of knowledge from a cited paper to a citing paper (Liu & Lu, 2012).

There are, however, two major limitations to the main path analysis approach in identifying knowledge evolution. First, this approach works with a binary citation network in which all citations are equally weighted. In other words, the degree of topical relatedness between the citing and cited publications is not directly reflected in the main path analysis (Liu & Lu, 2012), which thereby makes the results of the analysis less precise. In recent years, many attempts have been made to find ways to refine citation analysis such that all citations of a citing paper are not treated as equivalent (Kim, Jeong, & Song, 2016; Smith, 1981; Zhu, Turney, Lemire, & Vellino, 2015). In addition, it is reasonable to assume that a high degree of semantic similarity between the content of the citing paper and the content of the cited paper indicates that the cited paper has a significant impact on the citing paper (Zhu et al., 2015).

Received December 16, 2016; revised June 17, 2017; accepted August 7, 2017

 DOI: 10.1002/asi.23960

Dietz, Bickel, and Scheffer (2007) defined the strength of influence to be the similarity between the topic mixtures of the citing and cited publications.

The second limitation of the main path analysis is that it only shows the most significant path in the citation network. Furthermore, it does not show what topics evolve along the path. Because each paper in a path can be summarized by its topics (He et al., 2009), the topics of papers must be incorporated into identifying knowledge evolution. The current approaches based on main path analysis, however, have difficulty reflecting the topical impact of the cited paper on the citing paper in analyzing the backbone path of the citation network.

To overcome the limitations noted above, we incorporated a topic modeling technique to gauge the citation influence between a citing paper and a cited paper; we also identified the topics of each paper to determine how such topics evolve in a given scientific discipline. Using this integrated approach, we constructed a weighted network using papers in the field of p53, which is an important protein encoded by homologous genes in various organisms. The constructed network comprises nodes representing citing and cited papers and edges representing the degree of the citation influence between these papers; here, the degree is based on the topical similarity of a cited paper and a citing paper as well as the relative importance of the papers. In the constructed weighted citation network, we then examined how knowledge evolves in the field of p53.

To best present our work, the rest of our paper is organized as follows. In addition to this introductory section, we discuss related work in the next section. We present our proposed methods in the section that follows and then analyze our results. We conducted analysis from the macro level to the micro level, from community to path analysis to provide more details on how main topics in the field of p53 change over time. Specifically, we analyzed communities consisting of one or more dominant topics in the filtered citation network and identified significant topics based on cited count and citation influence weight, to explore when subfields focusing on specific topics of p53 research occur, and what the main topics of the subfields are. However, both community and topic analysis of the filtered citation network just show an overview of what the main topics are at a particular period of time and when the subfields related to a specific topic emerge. To provide more details on how topics diffuse, we extracted and analyzed the top 5 critical paths. The final section concludes our paper and presents avenues for future work.

Related Work

Knowledge Diffusion

Studies on knowledge diffusion using citations have primarily been conducted to identify technological knowledge diffusion patterns in patent citation data and to better understand knowledge evolution in a given scientific field using scholarly publications. Specifically, many studies have strived to trace technological knowledge flow patterns using patent citations to explore contribution patterns of multinationals to the technological progress of the United States semiconductor industry (Almeida, 1996), to identify technological knowledge patterns across institutional and national boundaries (Hu & Jaffe, 2003; Jaffe & Trajtenberg, 1996, 1999; Singh, 2005), and to estimate the process of knowledge diffusion and decay from patents of universities, public laboratories, and corporate entities in six countries, thereby examining the differences across countries and technological fields using patent citation data (Bacchiocchi & Montobbio, 2009). In other words, using patent citations, these studies focused on identifying the technological influence between institutions and nations as well as on comparing knowledge diffusion patterns.

Several studies have investigated knowledge diffusion among papers to identify knowledge evolution in a given scientific discipline using citation data. Garfield et al. (1964) traced the history of the DNA research field by analyzing a citation network under the assumption that "the history of science is regarded as a chronological sequence of events in which each new discovery is dependent upon earlier discoveries" (p. iii). As the number of scholarly publications grew, as noted above, a quantitative method called main path analysis was introduced by Hummon and Doreian (1989) to trace knowledge evolution in a large citation network.

Since then, many studies have applied main path analysis to investigate knowledge diffusion over time in a given scientific field. Mina et al. (2007) shed light on knowledge evolution connected to the emergence of coronary angioplasty by analyzing main paths in citation networks. Calero-Medina and Noyons (2008) combined bibliometric mapping and citation network techniques to investigate the process of knowledge creation and transfer through scientific publications. Moreover, Liu and Lu (2012) proposed a novel approach to explore the knowledge diffusion path of the resource-based theory (RBT) field.

Although studies on tracing knowledge evolution in a target field have, to date, primarily applied main path analysis to a citation network, such an approach has its limitations. First, main path analysis is conducted on binary citation networks (Liu & Lu, 2012). In other words, all citations are considered equal even though individual cited and citing pairs have different relevancies.

Many studies have been conducted to identify different relevancies between cited and citing publications. Specifically, Fujita, Kajikawa, Mori, and Sakata (2014) proved that a weighted citation network performs better than a binary citation network in terms of normalized size, publication year, text similarity, and density. In the study, content similarity was measured with references and author keywords to respectively obtain different relevance between citing and cited publications.

Peters, Braam, and van Raan (1995) mentioned that highly cited publications of a given scientific field form the paradigmatic base of the field, and thus citing papers show the current study of a research specialty based on a group of the highly cited publications. Accordingly, the citing papers

DOI: 10.1002/asi

should be cognitively and thus semantically related to one another as well as with the cited publications. In the study, it was proved that content relatedness of publications with a citation relationship is significantly higher than other publications. Moreover, to identify meaningful cited papers, Valenzuela, Ha, and Etzioni (2015, April) proposed a supervised classification method to address this issue with some features such as the similarity between abstracts of the cited and citing papers. This feature implies that the more similar the abstracts are, the more likely the citing paper extends the cited paper. Similarly, Moravcsik and Murugesan (1975) created relationship indicators to distinguish critical citations from all cited papers, including whether the citing paper extends previous ideas. Accordingly, in the present study we assumed that if citing publications that are strongly influenced by cited publications are more content-related to the cited papers.

Second, main path analysis only identifies the most significant paths; it does not show topical flows between papers along these paths. Thus, a content analysis should be conducted to identify knowledge development detected on the main paths. Finally, some highly cited papers are occasionally not included in any of the main paths (Liu & Lu, 2012) because the impact of each paper is not reflected in the construction of these main paths.

To address the above-mentioned problems, we propose a novel method to trace knowledge diffusion in a given discipline by constructing a citation network in which each link is weighted based on topical relevancies between a cited publication and its citing publication as well as the impact of each paper. Moreover, using our approach, topics of individual papers are automatically identified such that further work on topic detection will become increasingly unnecessary.

Topic Modeling in Citation Analysis

Latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003), a generative latent variable model, treats documents as a mixture of topics, wherein each topic is a multinomial distribution over the words. LDA is an unsupervised approach to topic discovery and is applicable to a large set of documents (Nallapati, Ahmed, Xing, & Cohen, 2008). Several studies have extended this model by integrating LDA with citations for link prediction, impact measurement of papers, and topic evolution detection.

Nallapati et al. (2008) devised two models: the Pairwise-Link-LDA and Link-PLSA-LDA models. These models are joint models for both text and citations. They can predict data that are not visible in the training sets by capturing topical similarities between the content of the cited publications and their citing publications.

In 2007, Dietz et al. proposed a citation influence model that captures the strength of citation influence among papers. They argued that not all cited papers have an equal contribution to a citing publication and that it is important to show articles that significantly impact one another in a citation network. Moreover, to evaluate the performance of the citation influence prediction model, citation influences scored manually by authors were compared with the scores generated by the model.

In 2006, Mann, Mimno, and Mccallum demonstrated the applicability of topic-model-based subfield discovery in measuring the impact of scholarly publications. Similarly, Gerrish and Blei (2010) presented a time-series topic model for measuring the impact of articles.

When it comes to topic evolution detection, He et al. (2009) applied the LDA model to a citation network and developed a novel inheritance topic model to improve the understanding of topic evolution in scientific literature. AlSumait, Barbará, and Domeniconi (2008) presented an online topic model (OLDA) that automatically captures topic patterns and identifies emerging topics in documents as well the changes in them over time.

Given the objective of our research, i.e., to detect topic evolution within a given scientific field in a citation network, the relevance between a cited paper and its citing paper should be identified as well as the topics of each paper. Accordingly, we employed a citation influence model proposed by Dietz et al. (2007) to measure the influence of cited papers on a citing paper. We also applied the PageRank algorithm (Page, Brin, Motwani, & Winograd, 1999) to each paper to reflect the importance of papers when measuring the citation influence based on the assumption that influential articles have a large impact on citing papers (Maslov & Redner, 2008). Using this combined model, we constructed a weighted citation network in which nodes represent individual scholarly publications and weighted edges represent the citation influences between articles.

Although several studies have applied main path analysis to identify the developmental trajectory in various research fields, they also have limitations in that the main path analysis methods were applied to a small citation network, and there was no effort to apply the developed methods to a weighted citation network. We address these limitations by using large datasets and by applying main path analysis to a weighted citation network in which relevancies between the cited and citing papers are properly reflected.

Methodology

The methodology and overall workflow of our study are illustrated in Figure 1. The subsections that follow provide further details.

Data Collection

As noted above, we selected the topic of p53 as a case study for topic diffusion analysis. The protein p53, also known as tumor protein p53, plays a pivotal role in multicellular organisms, where it helps prevent cancer formation. The web-focused October 2009 issue of the journal *Nature Reviews Cancer* primarily focused on p53. More specifically, the issue contained review papers that described the history of p53 research over the past 30 years; this enabled us to identify topical evolution in the research field of p53. Note that these review papers were written by researchers

DOI: 10.1002/asi

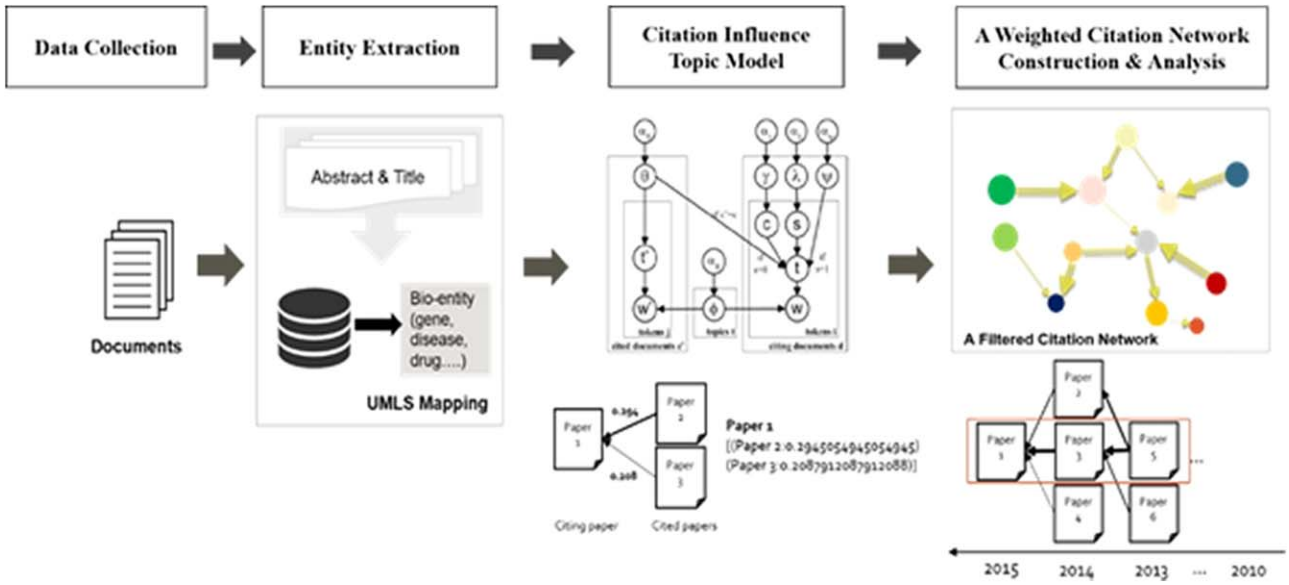

FIG. 1. Workflow of our proposed system. [Color figure can be viewed at wileyonlinelibrary.com]

and were therefore primarily based on expert subjective opinions and intuition. Thus, our proposed topic diffusion analysis should help p53 researchers grasp a more fine-grained overview of topic diffusion and evolution in their scholarly communication network.

We selected PubMed and PubMed Central as the data sources for p53-related articles. Search terms p53, tp53, P53, and TP53 were used to retrieve all papers in the field of p53 from PubMed Central, which is a full-text search engine covering the biomedical domain. Only articles that included these terms in their titles or abstracts were considered. The resulting number of articles was 20,589, all of which were published between 1910 and 2015. We then collected PMIDs (i.e., unique PubMed article identifiers) from the reference section of those retrieved articles. Using these collected PMIDs as input, we then downloaded records in XML from PubMed using the PubMed Elink service (Sayers, 2009) to get PMIDs of publications with the citation relationship to the 20,589 seed papers. The number of recollected PMIDs including the seed papers was 367,033, and those PMIDs were used as an input to retrieve EFetch XML records, including information such as author names, titles, abstracts, and publication dates, which were collected using the PubMed EFetch service (Sayers, 2009). Finally, we extracted titles, abstracts, publication dates, and author names from the collected EFetch XML records.

The basic statistics of the collected 367,033 papers is described in Appendixes A and B, which are found at http://informatics.yonsei.ac.kr/cim/appendix.docx. The numbers of papers in a given date range are presented in Appendix A, whereas the numbers of papers by cite count are presented in Appendix B. The majority of papers have low citation counts. Thus, we used publications with 30 or more citations, which accounts for the top 20% of total publications. As we mentioned in the Introduction, highly cited publications of a given scientific field form the paradigmatic base of the field and thus including papers that are highly cited by other publications in analyzing main paths is important. As such, we used the top 20% scientific publications with 30 or more citations based on the Pareto principle, the law of the vital few. Basile (1996) mentioned that 20% of well-known variables will account for 80% of the results. Accordingly, we assumed that 20% of publications with the highest citations have created meaningful results of the field. Thus, the number of papers that were finally selected was 77,570.

Entity Extraction

A major problem in topic modeling is that it is highly likely to generate general topics because the text includes many general terms such as human, cell, and patients. Therefore, we must first construct a list of stopwords to indicate the words that frequently occur but have no importance within a given field. Instead of creating this stopword list, we extracted biological entities from the titles and abstracts using a named entity recognition technique based on the assumption that stopwords would be excluded in the course of the mapping process. Here, we used PKDE4J (Song, Kim, Lee, Heo, & Kang, 2015), a biomedical text mining tool, for entity extraction. Using this tool, entities can be extracted by either a dictionary, via supervised learning, or both.

In our experiments, we used a combined hybrid approach to entity extraction. To extract the biological entities, entity candidates were extracted using supervised learning. Then, to decrease false positives, these candidates were mapped to Unified Medical Language System (UMLS) concepts. Developed by the National Library of Medicine, UMLS is a vocabulary database of biomedical concepts and corresponding relationships. To extract p53-related biological terms that are not general, exactly 103 UMLS semantic types out of a total of 143 were carefully selected with the help of a

DOI: 10.1002/asi

biologist. Although we extracted entities corresponding to specific semantic types, there were still many general words such as disease, therapy, and cell. These certainly are biological terms related to p53 but have more general meanings in the field. Therefore, we decided to exclude the terms with meanings that were too broad. For example, when the term cancer was detected to have many subterms, such as breast cancer, prostate cancer, gastric cancer, and liver cancer, we removed these terms from the extracted entities.

Citation Influence Topic Model

As noted above, LDA is a generative probabilistic model for collecting discrete data, for example, text corpora (Blei et al., 2003). LDA is based on the idea that documents are random mixtures of latent topics, where each topic is a probability distribution over the words. To infer topical influences from citations, Dietz et al. (2007) developed the citation influence model (CIM) based on LDA. CIM is designed to predict the citation influence between citing and cited publications.

The generative CIM assumes the following process:

- For each topic t
  - Draw word distribution $\phi_t = p(w|t) \sim dirichlet(\vec{\alpha}_\phi)$
- For each cited paper c′
  - Draw a topic mixture of cited paper $\theta_{c'} = p(t'|c') \sim dirichlet(\vec{\alpha}_\theta)$
    - For each entity j
      - Draw topic $t'_{c',j} \sim \theta_{c'}$ from the topic mixture
      - Draw word $w_{c',j} \sim \phi_{t'_{c',j}}$ from the topic-specific word distribution
- For each citing paper d
  - Draw citation mixture $\gamma_d = p(c|d)|_{L(d)} \sim dirichlet(\vec{\alpha}_\gamma)$, where c is directly cited by citing paper d
  - Draw innovation mixture $\psi_d = p(t|d) \sim dirichlet(\vec{\alpha}_\psi)$, where paper d is not affected by cited paper c
  - Draw the ratio between entities, i.e., whether they are related to citations or innovation topic mixture $\lambda_d = p(s=0|d) \sim beta(\alpha_{\lambda_0}, \alpha_{\lambda_\psi})$
  - For each entity *i*
    - Flip a coin, i.e., $s_{d,i} \sim bernoulli(\lambda_d)$
    - If $s_{d,i}=0$
      - Draw cited document $c_{d,i} \sim multi(\gamma_d)$
      - Draw topic $t_{d,i} \sim multi(\theta_{c_{d,i}})$ from the cited document's topic mixture
    - Else, if $s_{d,i}=1$
      - Draw topic $t_{d,i} \sim multi(\psi_d)$ from the innovation topic mixture
    - Draw word $w_{d,i} \sim multi(\phi_{t_{d,i}})$ from the topic-specific word distribution

The variables in the above process are described as follows. First, c' and d represent cited papers and citing papers, respectively; θ represents the topic mixture of the topical atmosphere of a cited publication, wherein all entities associated with cited papers are called the topical atmosphere of a cited paper. Next, ψ denotes the innovation topic mixture of a citing paper, ϕ represents the word distribution for each topic, γ indicates the distribution of citation influences, and λ represents the parameter for the coin flip, which is used to select whether to draw topics from θ or ψ. Furthermore, w' and w represent words in cited and citing papers, respectively, whereas t' and t indicate topic assignments of the entities in cited and citing papers, respectively. Next, c represents a cited publication to which an entity in a citing publication is related, s denotes whether the topic of a citing paper is drawn from inheritance or innovation, and α indicates Dirichlet, the beta parameters of the multinomial, and Bernoulli distributions.

Through the given process, the model predicts γ (i.e., the influence of a citation link) and λ (i.e., the innovativeness of the citing paper). To establish the absolute citation influence, we use $\gamma \cdot \lambda$ as the "measure for absolute strength of influence" (Dietz et al., 2007).

Using the extracted and refined entities, we then utilize the CIM to capture the strength of citation influence among papers and to show which topics of a cited paper primarily flow into a citing paper. Unfortunately, CIM does not reflect the impact of the citing paper into its topic modeling. Therefore, we extend the CIM by incorporating the PageRank (Page et al., 1999) value of each cited paper into the calculation of the strength of the citation influence between a cited paper and the citing paper because we assumed that influential cited papers have more impact on citing papers. We therefore use the following equation to measure citation influence:

$$\text{Citation Influence} = \gamma \cdot \lambda \cdot (\text{PageRank of Cited Paper}).$$

Our extended CIM can capture different citation influences among articles on the basis of topics and PageRank values. In other words, it enables us to calculate different relevancies between citing and cited pairs. Using the results obtained by the topic model, we constructed a citation network that shows publications that have significant influence on one another. We applied our model to p53-related papers, with the number of topics set to 30. We provided the explanations of why we chose 30 topics over other topic numbers in the Network Analysis section, below.

Overlay of Citation Influence With Component Analysis

To identify topic evolution in the field of p53, we filtered the citation network with a citation influence value greater than 0.030, which accounts for the top 20% of the total. The threshold of 0.03 was selected based on cited count, path average, and the number of unique nodes in the paths for the critical path analysis.

The top 5%, 10%, 20%, and 30% citation influence value produces a different ratio of articles, respectively with cited count below the average, the number of paths with length above 3, and unique nodes included in the paths, which are described in Table 1. Specifically, when the top 5% and 10% citation influence values are used, the ratio of articles with

DOI: 10.1002/asi

TABLE 1. Comparison between top 5%, 10%, 20%, and 30% citation influence value.

	Top 30%	Top 20%	Top 10%	Top 5%
The ratio including articles with cited count below the average	0.45	0.44	0.52	0.58
Paths having length above 3	8681	2981	685	7
Unique nodes included in paths	1002	615	202	19

cited count below average is high and path length above 3 is too small to analyze the development of the p53 field.

However, using the top 20%, the percentage of nodes with less than average cited count (88) is reduced, which can decrease the weight of the paper with low impact. In other words, when using a citation influence value over 0.03, the ratio of publications with citation count over average is highest. In addition, using the top 30% citation influence value generates many duplicate nodes in the critical paths as opposed to using the top 20% including more unique nodes in the paths. Therefore, we believe that using the top 20% is the most appropriate for the research goal of tracking the development of science.

Next, we organized papers by publication date to identify topic evolution. These four steps were followed to generate the filtered citation network.

1. Choose nodes with out-degree measures greater than three as the starting nodes.
2. Start with one starting node selected at random. Find all outgoing links for this starting node and then select the three links with the highest citation influence values. Repeat step 1 until there are no outgoing links for the starting node.
3. For the rest of the starting nodes, perform step 2.
4. Exclude the extracted paths with path lengths less than three.

Ranking Major Citation Influence Paths

We ranked the extracted paths using the normalized sum of edge weights, which we calculated via the following formula:

$$\text{The normalized sum of edge weights} = \frac{\textit{The sum of weights of each edge in the graph}}{\log_2 \textit{Path length}}$$

The reason we normalize the sum of the weights of each edge of the path is that the path length range is so diverse that adjusting values to a common scale is necessary for comparison. Using the normalized values, we selected the top 10 paths to analyze topical diffusion between the papers belonging to each path.

Results

Network Analysis

To have a better understanding of the overall topology of the collected data, we constructed a citation network comprising 77,570 nodes and 153,362 edges. The nodes represent the papers published in the field of p53, whereas the edges denote the citation influences between the various papers. The citation network was then filtered by retaining the papers with citation influence values greater than 0.030; thus, papers that have a significant influence on one another were retained. After this filtering, the number of remaining nodes and edges was 13,407 and 26,342, respectively. We then organized the papers by publication date and major topic groups to identify topic evolution in the field of p53. The assignment of the paper to a topic was determined by the document-topic probability distributions generated by the proposed approach.

We visualized the filtered network using Gephi (Bastian, Heymann, & Jacomy, 2009), a network visualization tool. The visualized citation network is presented in Figure 2. In the figure, the color of each node denotes the modularity class that it belongs to. The modularity class value for each node was generated using the community detection algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Furthermore, in the figure the size of each node is shown proportional to the value of its out-degree; this means that the papers with high citation frequencies are shown to be bigger than those with low citation frequencies. Finally, the thickness of each edge represents the degree of the citation influence between a cited paper and its citing paper.

To identify topical evolution in the field of p53, we placed the visualized network in the coordinate plane, with the x-axis indicating the 30 topics generated from the extended CIM and the y-axis denoting the publication date. By placing the network in the coordinate plane, we could detect which communities represented by one or two topics are dominant in a certain year; this enabled us to understand how topics evolve over time.

Specifically, studies on p53 started in 1979, and the number of papers investigating p53 has increased drastically from 1990 onward. Moreover, before 1990, various topics of papers on p53 belonged to only one community, which means diverse topics are densely connected and were studied together at that time. In other words, until 1990, studies on p53 covered broad topics and were not specialized; however, after 1990, several communities were detected focusing on a specific topic. In other words, we can assume that subfields of p53 research dealing with specific topics emerged. The list of resulting topics is provided in Appendix C, which is found at http://informatics.yonsei.ac.kr/cim/appendix.docx.

We compared the results for topics by setting the number of topics to 20, 30, 40, and 50 to identify the optimal number of topics for capturing the truly distinct topics. To this end, we calculated the cosine similarity between topics in each case. The cosine similarity values between topics were

DOI: 10.1002/asi

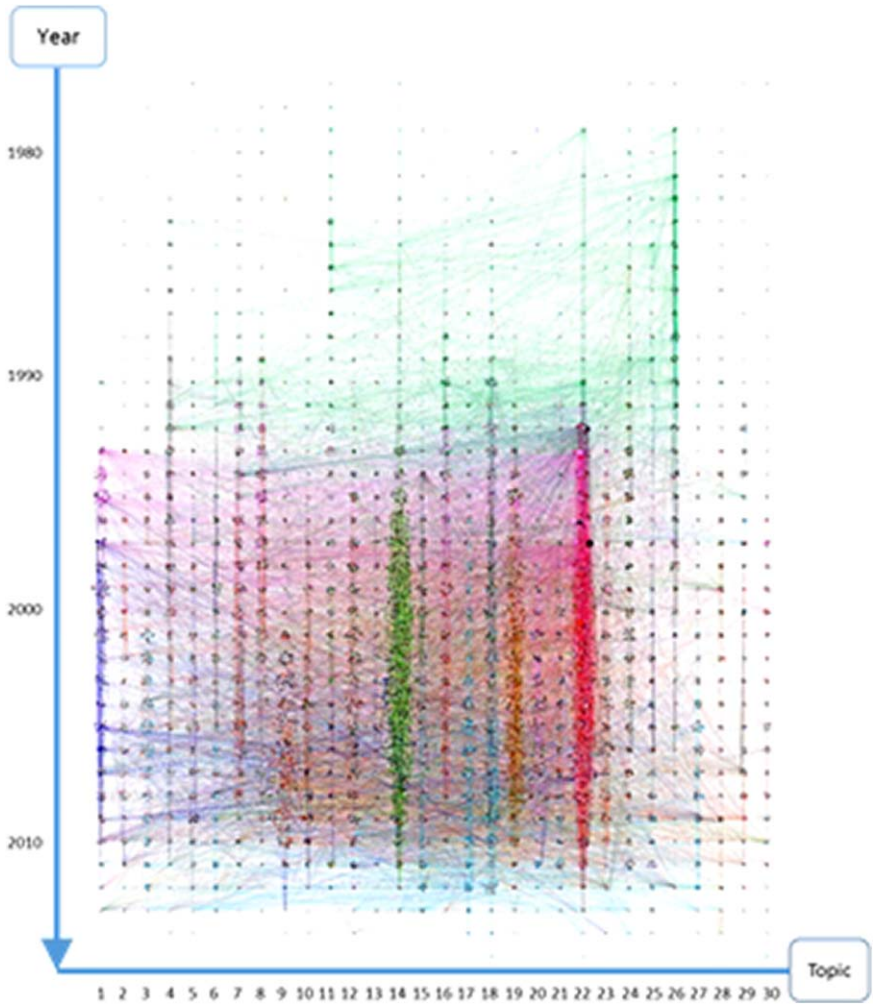

FIG. 2. Visualized filtered citation network. [Color figure can be viewed at wileyonlinelibrary.com]

0.1694, 0.1705, 0.2225, and 0.1765 for the number of topics set to 20, 30, 40, and 50, respectively. Although our results show that we were able to capture the most distinct topics when the number of topics was set to 20, we chose the number of topics to be 30 to show the diversity in the topics used in our topical diffusion analysis because it gave us a greater variety of topics and the difference between 20 and 30 topics here was insignificant.

To identify significant topics in the network, for each topic we summarize in Table 2 the total number of papers, the range of publication dates, the sum of out-degrees (i.e., cited counts) of the papers, the average out-degree per paper, the sum of citation influences of the cited papers, and the average citation influence weight per paper. As shown in the table, the primary topics of the filtered citation network were topics 22 and 26, given the average out-degree and average citation influence weight of each paper. Although more papers belonged to topic 14 than topics 22 and 26, each paper belonging to topic 14 had a lower out-degree and citation influence weight; this proves that topics 22 and 26 are the most significant topics in the network, as they include more influential nodes than topic 14.

For the evaluation purpose, we compared our results to a review paper that was published in the journal *Nature Reviews Cancer* by Levine and Oren (2009) to validate the similarity between the two quantitatively. We extracted the PMIDs from the reference page of the paper and compared them with our results. In the paper, there were 149 distinct PMIDs and 124 PMIDs among 149 were included in the filtered network, which covers 83.2% of the total papers, which proves that our results are quite credible.

We analyzed a weighted citation network to identify how topics diffuse and evolve over time in the research field of p53. We ranked paths using the formula of the normalized sum of edge weights to select influential paths for analysis. The top 10 paths are described in Table 3. We further analyzed the top 5 paths from this list to identify how topics that were diffused and evolved over time can be identified at the paper level.

To investigate the influence of self-citation on our results, we measured the number of citations and self-citations as well as the average of citation influence value in the filtered citation network. In addition, to explore if the citation influence value has an impact on the path extraction, we

DOI: 10.1002/asi

TABLE 2. Statistics of the filtered citation network.

Topic	Paper	Year (average)	Range of date (95%)	Out-degree (cited count)	Average out-degree per paper	Sum of citation influence weights	Average citation influence weight per paper
1	496	2001	1990–2011	1241	2.50	58.78	0.12
2	283	2002	1988–2012	265	0.94	10.43	0.04
3	432	2002	1988–2012	514	1.19	22.65	0.05
4	424	1998	1984–2010	743	1.75	31.47	0.07
5	322	2002	1986–2012	410	1.27	16.71	0.05
6	359	2002	1987–2012	468	1.30	18.36	0.05
7	531	1999	1982–2010	855	1.61	38.97	0.07
8	517	1999	1982–2009	753	1.46	30.66	0.06
9	607	2005	1994–2011	1144	1.88	48.67	0.08
10	301	2003	1990–2011	319	1.06	12.53	0.04
11	401	1996	1979–2011	615	1.53	26.62	0.07
12	516	2002	1990–2011	742	1.44	31.43	0.06
13	233	2002	1988–2012	231	0.99	8.74	0.04
14	1017	2001	1985–2010	2142	2.11	93.35	0.09
15	455	2003	1993–2012	863	1.90	43.39	0.10
16	448	1999	1987–2010	810	1.81	37.27	0.08
17	445	2003	1987–2013	570	1.28	24.4	0.05
18	722	2000	1983–2012	1234	1.71	56.79	0.08
19	731	2002	1992–2011	1473	2.02	67.96	0.09
20	413	2005	1996–2011	673	1.63	27.58	0.07
21	318	2002	1984–2012	492	1.55	25.3	0.08
22	877	2001	1989–2011	5998	6.84	391.47	0.45
23	413	2002	1990–2012	448	1.08	17.89	0.04
24	467	1999	1980–2011	570	1.22	24.68	0.05
25	314	2002	1982–2013	517	1.65	24.53	0.08
26	397	1993	1979–2009	1141	2.87	56.15	0.14
27	272	2004	1989–2012	336	1.24	13.27	0.05
28	245	2000	1981–2012	284	1.16	11.26	0.05
29	233	2000	1981–2012	239	1.03	10.36	0.04
30	218	2002	1984–2013	252	1.16	10.21	0.05

TABLE 3. Ranking of critical citation influence paths.

Rank	Path ID	Path length	Sum of weights of each edge in the path	Normalized sum of the edge weights of the path
1	3847	5	0.684231	0.294682
2	3840	5	0.677971	0.291986
3	2928	4	0.576262	0.288131
4	2934	5	0.634776	0.273383
5	3849	4	0.628549	0.270701
6	3842	5	0.622289	0.268005
7	1284	14	1.017363	0.26721
8	676	14	1.016045	0.266864
9	496	14	1.006249715	0.029584612
10	1269	13	0.972681204	0.029708442

compared the average citation influence value of all citations to that of self-citation in the extracted paths, which is described in Table 4.

The total citation relations used in the research is 133,553. We defined citation between two papers that have the same author names in the papers as self-citation. Among the total citations in the filtered citation network, 13,862 citations are self-citations. The average citation influence weight of self-citation is 0.02 and the standard deviation is 0.01, whereas the average weight of nonself-citation is 0.018 and the standard deviation is 0.009. After the path extraction process, among the 811 citations, 141 are self-citations. The average weight of self-citation is 0.062 and the standard deviation is 0.023. The average weight of nonself-citation is 0.068 and the standard deviation is 0.032. Because the citation influence value is not so different between the self-citation and the other, self-citation does not have any special impact on the filtering and path-making process.

In conclusion, after the path extraction process the portion of self-citation increased slightly, but the weight was not significantly higher. Given that the rate of self-citations in the previous study of the biology and biomedicine reached 21.74% (Costas, Van Leeuwen, & Bordons, 2010), the data used in our study are relatively small in the effect of self-citation.

DOI: 10.1002/asi

TABLE 4. Influence of self-citation on the results.

	Citation count	Average of citation influence value	Standard deviation
All citation relations of the filtered citation network	133553	0.018	0.01
Self-citation of the filtered citation network	13862	0.02	0.009
Citation of the extracted paths	811	0.068	0.032
Self-citation of the extracted paths	141	0.062	0.023

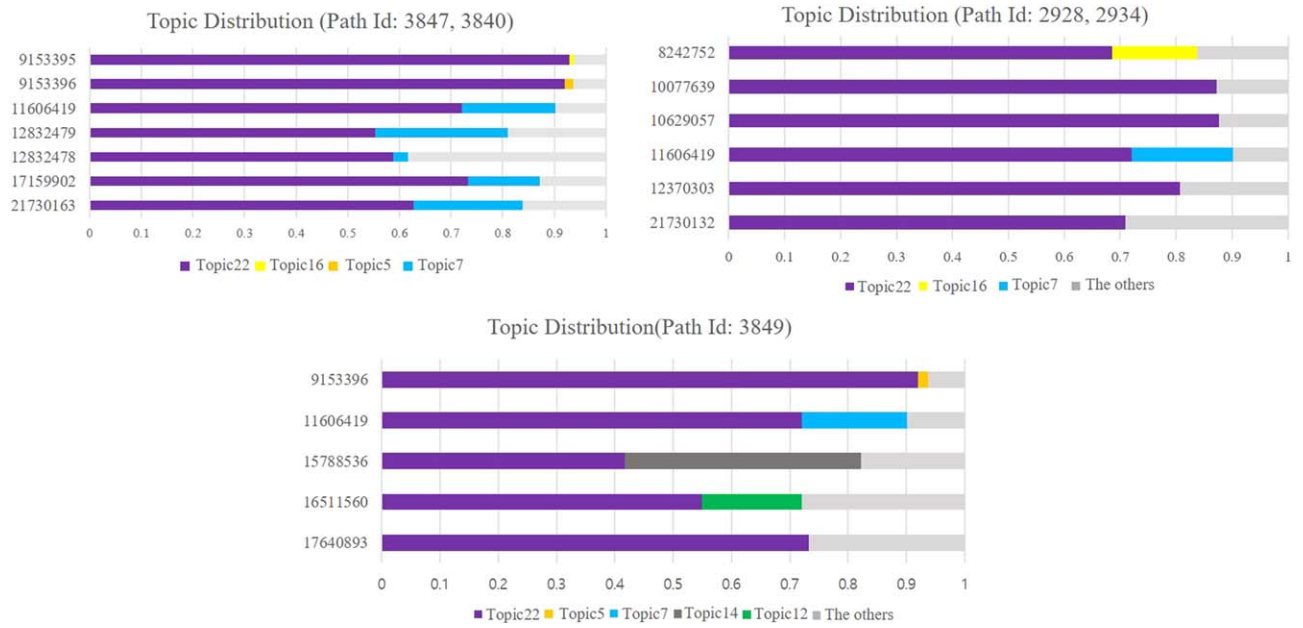

FIG. 3. Topics of the top 5 paths. [Color figure can be viewed at wileyonlinelibrary.com]

Topical Diffusion of the Top 5 Paths

We selected the top 5 paths to identify important topics and topical diffusion in the p53 citation network using our proposed ranking algorithm (described above). The main topic of each path can be detected by the topic distribution of papers belonging to each path.

Figure 3 shows the topic distribution of papers in the top 5 paths. As illustrated in Figure 4, the first-ranked and second-ranked paths (i.e., Path IDs 3847 and 3840, respectively) have the same papers but their starting nodes differ. More specifically, papers with PMIDs 9153395 and 9153396 are the starting nodes of paths 3847 and 3840, respectively; therefore, we present them in the same chart. As for the third-ranked and fourth-ranked paths (i.e., Path IDs 2928 and 2934, respectively), which are illustrated in Figure 5, the paper with PMID 10629057 is included only in the fourth-ranked path, and the rest of the papers in the paths are identical. Accordingly, the two paths are described in the same chart.

As shown in Figure 3, the major topic of the top 5 paths is topic 22. In other words, the paths comprise papers that primarily focus on the ARF-Mdm2-p53 tumor suppressor pathway, as indicated in Appendix C.

Topical Diffusion of the Top 5 Paths

Next, we visualized the top 5 paths to identify how topics diffuse in each path. Figures 4, 5, and 6 show each path; the corresponding descriptions that include the path ID, ranking, main topic group of the path, keywords, and topical diffusion are provided further below. Moreover, 30 topics were drawn using the extended CIM, where the topic proportions of each paper are represented by a unique color in the figure. Two major topic groups that each paper belongs to are listed below each circle, which represents each individual paper. A list of 30 labeled topics and word distributions corresponding to each topic are attached to the last page of our paper. Moreover, link thickness reflects the strength of citation influence between a citing paper and a cited paper.

The topics colored in the paths and the thematic terms of a given topic allow for understanding topical diffusion of the p53 field in an effective manner. Specifically, by visualizing top 5 critical paths, we would like to help researchers to identify topical diffusion with ease. In the path, each publication is colored by its topic proportions and it would help researchers to identify what main topics in the field of p53 are and how those topics diffuse over time based on color changes through the path. In other words, path visualization providing information on publication date, change of topic color, topic label, and words of each topic could provide an overview of development in the field of p53.

Path ID: 3840, 3847
Ranking: 1, 2
Main topic of the papers in the path: ARF-Mdm2-p53 tumor suppressor pathway

DOI: 10.1002/asi

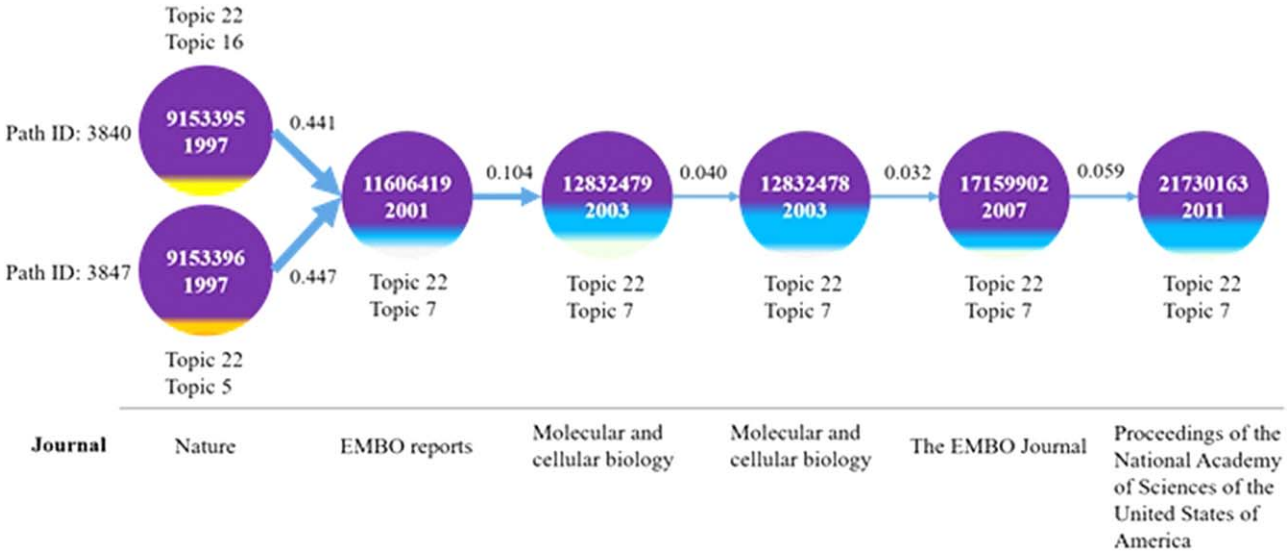

FIG. 4. First- and second-ranked paths. [Color figure can be viewed at wileyonlinelibrary.com]

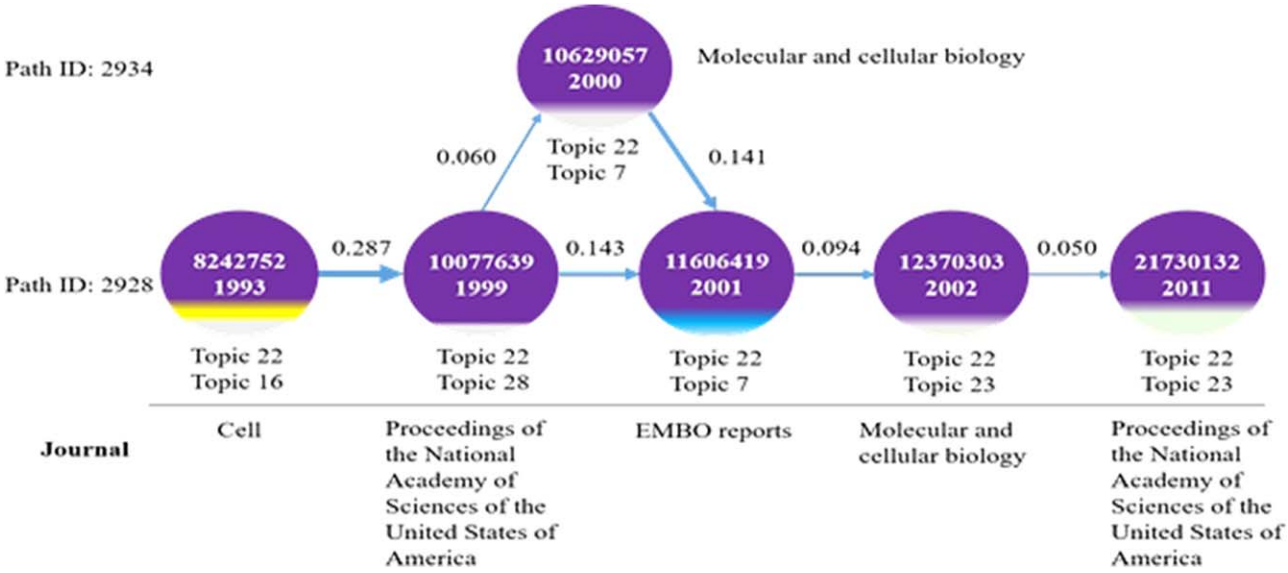

FIG. 5. Third- and fourth-ranked paths. [Color figure can be viewed at wileyonlinelibrary.com]

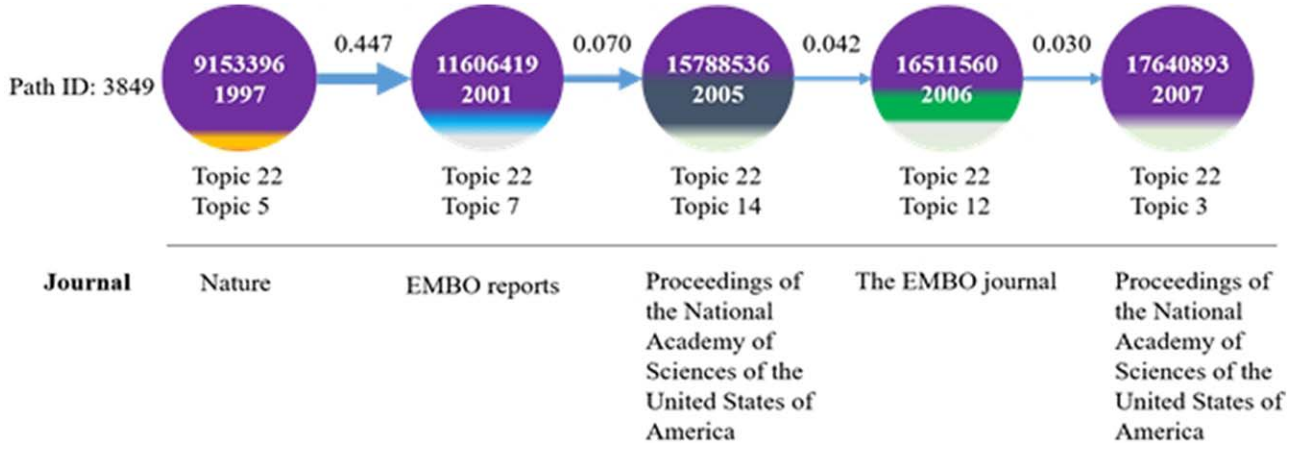

FIG. 6. Fifth-ranked path. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

Keywords of topics:

- Topic 22: Mdm2, Cell cycle arrest, Apoptosis, Tumor Suppressor, DNA damage
- Topic 7: Ubiquitination, Ubiquitin, Degradation, E3, Ligase, ARF

Main community group: 17

Topical diffusion in the path: The two paths show how the topic "ARF-Mdm2-p53 tumor suppressor pathway" diffused between papers from 1997 to 2011. More specifically, two paths with path IDs 3840 and 3847 comprise almost the same papers (i.e., nodes), except for the starting nodes. Staring papers 9153395 and 9153396 have similar content regarding the role of Mdm2 in p53 regulation; however, they do not cite one another because they were both published in the same journal in May 1995. In 2001, the two papers were cited by the paper with PMID 11606419. The citation influence between the citing paper and the two cited papers was found to be high; this indicates that the cited papers had significant influence on the citing paper. Moreover, the content of the paper is similar to those of the two cited papers; however, in the case of cited papers, they primarily focused on two proteins, Mdm2 and p53, whereas the citing paper studied one additional protein, MDMX, for regulation of Mdm2 and p53.

Path ID: 2928, 2934

Ranking: 3, 4

Main topic of papers in the path: ARF-Mdm2-p53 tumor suppressor pathway

Keywords of topics:

- Topic 22: Mdm2, Cell cycle arrest, Apoptosis, Tumor Suppressor, DNA damage
- Topic 7: Ubiquitination, Ubiquitin, Degradation, E3, Ligase, ARF

Main community group: 8242752 (5); the rest of the papers (17)

Topical diffusion in the path: Paths with path IDs 3928 and 2934 comprise the same papers, except for the 10629057 paper. In other words, path 2934 includes the 10629057 paper, whereas path 3928 does not. The main topic of the 10629057 paper was topic 22, which is the same as that of the papers in the two paths. Accordingly, these two were analyzed simultaneously to identify the topical diffusion between papers published during 1994–2011.

The main topic group of the papers in the path was topic 21, as in the case of the first- and second-ranked paths; however, topical diffusion between papers is expected to be different from that of the paths because, as Figure 2 shows, a small topic change occurs throughout the path.

Overall, the path shows how studies on Mdm2-related or p53-related proteins and genes involved in p53 regulation changed over time. In other words, we can identify how topic 21 diffused along the path.

Path ID: 3849

Ranking: 5

Main topic of papers in the path: ARF-Mdm2-p53 tumor suppressor pathway

Keywords of topics:

- Topic 22: Mdm2, Cell cycle arrest, Apoptosis, Tumor Suppressor, DNA damage
- Topic 7: Ubiquitination, Ubiquitin, Degradation, E3, Ligase, ARF
- Topic 14: ATM, DNA damage, DNA repair, DNA replication, Ionizing radiation
- Topic 12: Phosphorylation, Kinases, Activation, AKT, PTEN

Main community group: 21178072 (21); the rest of papers (17)

Topical diffusion in the path: The path shows how studies on Mdm2 changed from 1997 to 2011. The first paper, with PMID 9153396, primarily examined the role of Mdm2 in p53 regulation. This paper is cited by the 11606419 paper, and the citation influence between the two papers is higher than that between other papers in the path. Thus, we can assume that the content of the citing paper is similar to that of the cited paper. More specifically, based on the finding of the cited paper that Mdm2 also regulates p53 protein levels, the citing paper introduced the Mdmx protein, which is structurally homologous to Mdm2 and can regulate both Mdm2 and p53 by preventing the Mdm2-dependent degradation of p53 and the self-ubiquitylation by Mdm2.

Discussion

We further examined the top *N* critical paths to identify the major topics that characterize the paths. To this end, primary topics comprising the top 30, 50, 100, 200, and 300 nodes were analyzed to detect the main topics of each of the top *N* paths. The main topics of each top *N* path were identified by the number of nodes (papers) that were assigned to

TABLE 5. Topics in the top *N* paths.

	Topic											
Top N	4	10	11	16	18	22	24	25	26	27	28	30
30 paths	1	21	**48**	2	0	**223**	1	0	**56**	0	1	2
50 paths	2	40	**93**	3	0	**366**	1	0	**100**	0	2	2
100 paths	6	86	**202**	6	1	**701**	4	1	**204**	0	4	5
200 paths	8	185	**371**	11	5	**1382**	11	10	**436**	2	11	30
300 paths	12	282	**509**	14	9	**2057**	23	14	**668**	6	16	58

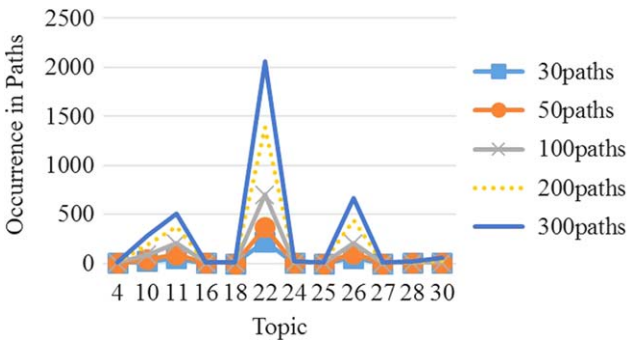

FIG. 7. Topic distribution of the top *N* paths. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

TABLE 6. Average betweenness centrality of the top 5 topic groups.

Topic	Citation network average betweenness centrality	Rank	Filtered citation network average betweenness centrality	Rank	Rank difference
10	140.42833	25	48.134966	9	16
11	240.58433	21	107.79104	4	17
22	6346.471	1	475.61225	1	0
26	829.58527	7	299.16852	2	5
30	286.25824	17	6.657612	21	−4

each topic in terms of the document-topic probability distribution. Table 5 shows the topics in the top 30, 50, 100, 200, and 300 paths, wherein the topics presented in boldface indicate that they are the most frequently engaged topics in paths.

To show which topics are most frequently engaged in the paths, Figure 7 shows the topic distribution of the top *N* paths, wherein the x-axis represents the list of topics and the y-axis represents the number of occurrences of a topic in a path.

From Table 5 and Figure 7, we observe that the main topics in the top 30, 50, 100, 200, and 300 paths are similar. Specifically, the major topics in the top 30 paths are 22, 26, and 11, which are similarly observed in the top 50, 100, 200, and 300 paths. As shown above as part of the discussion section, the major topics in the filtered citation network are topics 22 and 26. These two topics include many nodes for which the citation count and citation influence weight are high. In other words, a large number of influential papers that have a significant influence on one another belong to topics 22 and 26; this leads them to be selected and included in the significant paths.

Conversely, topic 11 includes papers that have relatively low cited counts and citation influence weights; however, we found that topic 11 has a relatively high average betweenness centrality in the filtered citation network, where betweenness centrality is an indicator that "measures the extent to which a vertex lies on paths between other vertices" (Newman, 2010, p. 185). As noted above, we filtered our citation network to leave critical paths. Through this filtering process, topic 11's average betweenness centrality drastically changed. Table 6 shows how average betweenness centrality of the five topic groups that appeared in the top 300 paths changed before and after the filtering process. Before the filtering process, the ranking of topic 11 was relatively low in comparison to the other topic groups; however, in the filtered network, the betweenness centrality ranking of topic 11 changed from 21 to 4. Therefore, we assumed that the nodes in which the topic is 11 are well-linked with the influential papers included in topics 26 and 22. A topic analysis of the top *N* paths also proves that the main topics on the critical paths are topics 22 and 26, as confirmed by the findings mentioned in the previous sections.

Conclusion

In this paper we presented a method for detecting topic evolution in a citation network in which links are weighted with citation influences between citing and cited papers. To measure citation influence between papers, we employed a citation topic model devised by Dietz et al. (2007) and extended it by reflecting the PageRank (Page et al., 1999) of the cited papers.

We assumed that analyzing knowledge diffusion among influential papers that have a significant impact on one another could yield more meaningful and accurate results than considering papers as being equal. Thus, we reflected the citation counts of each paper when constructing our citation network, filtered this citation network by considering citation influence weights between papers, and then further narrowed the network by leaving only critical paths.

Through community analysis and critical path analysis, our approach proved to be effective in discovering the most significant research topics and paths representing topic evolution in a given research field in an automated and precise manner.

Our approach has several benefits over the traditional main path analysis in that critical paths are identified in a weighted citation network consisting of papers that significantly influence each other by filtering out noncritical citations. Thus, the critical paths could include meaningful citation relationships among papers. In addition to that, with the proposed method, the additional task for topic identification of papers is not required for analyzing topic evolution of a given scientific field.

The history and evolution of a particular scientific field that is identified using the presented framework can provide scientists with the opportunity to enter into a new field by providing them with an overview of that field's influential events. Moreover, when researchers write a review paper, some knowledge in a given scientific domain is required. To this end, browsing and analyzing documents is an essential step, but a time-consuming and difficult task because of a large number of accumulated publications. In this situation, our proposed methodology can help researchers discover critical publications and write review papers.

Despite these benefits, there are several limitations. Specifically, a major limitation of our study is that topical influence was measured on the basis of only titles and abstracts, which may underrepresent the topical linkage between citing and cited papers. Another limitation is that our proposed approach was applied to the specific research topic of p53 protein. Therefore, as a follow-up study we plan to utilize full-text articles to measure topical influence and apply our proposed method to a broader domain, for example, to the field of information science, to examine whether it helps in

DOI: 10.1002/asi

detecting the topic evolution of citation networks in other fields in the same systematic, unbiased manner. Finally, we admit a shortcoming of the present work in that that we did not employ a statistical method such as survey or focused group interview to examine the results generated by the proposed approach more quantitatively and systematically. This is an interesting research topic, and we plan to investigate it as a follow-up study.

Acknowledgments

This work was supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT, and Future Planning through the National Research Foundation.

References

Almeida, P. (1996). Knowledge sourcing by foreign multinationals: Patent citation analysis in the US semiconductor industry. Strategic Management Journal, 17, 155–165.

AlSumait, L., Barbará, D., & Domeniconi, C. (2008, December). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (pp. 3–12). IEEE.

Bacchiocchi, E., & Montobbio, F. (2009). Knowledge diffusion from university and public research. A comparison between US, Japan and Europe using patent citations. The Journal of Technology Transfer, 34, 169–181.

Basile, F. (1996). Great management ideas can work for you. Indianapolis Business Journal, 16, 53–54.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. ICWSM, 8, 361–362.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10, P10008.

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

Calero-Medina, C., & Noyons, E.C. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. Journal of Informetrics, 2, 272–279.

Costas, R., Van Leeuwen, T.N., & Bordons, M. (2010). Self-citations at the meso and individual levels: Effects of different calculation methods. Scientometrics, 82, 517–537.

Dietz, L., Bickel, S., & Scheffer, T. (2007, June). Unsupervised prediction of citation influences. In Proceedings of the 24th International Conference on Machine Learning (pp. 233–240). New York: ACM.

Fujita, K., Kajikawa, Y., Mori, J., & Sakata, I. (2014). Detecting research fronts using different types of weighted citation networks. Journal of Engineering and Technology Management, 32, 129–146.

Garfield, E., Sher, I.H., & Torpie, R.J. (1964). The use of citation data in writing the history of science. Philadelphia, PA: Institute for Scientific Information.

Gerrish, S., & Blei, D.M. (2010). A language-based approach to measuring scholarly impact. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 375–382).

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009, November). Detecting topic evolution in scientific literature: How can citations help? In Proceedings of the 18th ACM Conference on Information and Knowledge Management (pp. 957–966). New York: ACM.

Hu, A.G., & Jaffe, A.B. (2003). Patent citations and international knowledge flow: The cases of Korea and Taiwan. International Journal of Industrial Organization, 21, 849–880.

Hummon, N.P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social Networks, 11, 39–63.

Jaffe, A.B., & Trajtenberg, M. (1996). Flows of knowledge from universities and federal laboratories: Modeling the flow of patent citations over time and across institutional and geographic boundaries. Proceedings of the National Academy of Sciences of the United States of America, 93, 12671–12677.

Jaffe, A.B., & Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. Economics of Innovation and New Technology, 8, 105–136.

Kim, H.J., Jeong, Y.K., & Song, M. (2016). Content-and proximity-based author co-citation analysis using citation sentences. Journal of Informetrics, 10, 954–966.

Levine, A.J., & Oren, M. (2009). The first 30 years of p53: Growing ever more complex. Nature Reviews Cancer, 9, 749–758.

Liu, J.S., & Lu, L.Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. Journal of the American Society for Information Science and Technology, 63, 528–542.

Liu, J.S., & Kuan, C.H. (2016). A new approach for main path analysis: Decay in knowledge diffusion. Journal of the Association for Information Science and Technology, 67, 465–476.

Ma, V.C., & Liu, J.S. (2016). Exploring the research fronts and main paths of literature: A case study of shareholder activism research. Scientometrics, 109, 33–52.

Mann, G.S., Mimno, D., & McCallum, A. (2006, June). Bibliometric impact measures leveraging topic analysis. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 65–74). New York: ACM.

Maslov, S., & Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. Journal of Neuroscience, 28, 11103–11105.

Mina, A., Ramlogan, R., Tampubolon, G., & Metcalfe, J.S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. Research Policy, 36, 789–806.

Moravcsik, M.J., & Murugesan, P. (1975). Some results on the function and quality of citations. Social Studies of Science, 5, 86–92.

Nallapati, R.M., Ahmed, A., Xing, E.P., & Cohen, W.W. (2008, August). Joint latent topic models for text and citations. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 542–550). New York: ACM.

Newman, M. (2010). Networks: An introduction. Oxford, UK: Oxford University Press.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web (Technical report). Stanford, CA: Stanford University.

Peters, H.P., Braam, R.R., & van Raan, A.F. (1995). Cognitive resemblance and citation relations in chemical engineering publications. Journal of the American Society for Information Science, 46, 9.

Sayers, E. (2009). Entrez programming utilities help. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK25499.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. Management Science, 51, 756–770.

Smith, L.C. (1981). Citation analysis. Library Trends, 30, 83–106.

Song, M., Kim, W.C., Lee, D., Heo, G.E., & Kang, K.Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. Journal of Biomedical Informatics, 57, 320–332.

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. In Proceedings of the 29th AAAI Workshop: Scholarly Big Data (pp. 21–26). Austin, TX: AAAI.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in Complex Systems, 10, 93–115.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology, 66, 408–427.

Appendix A

Number of papers matching the given criteria (by year range)

Year	Number of papers
1910–1969	154
1970–1979	2246
1980–1989	15604
1990–1999	74612
2000–2009	181264
2010–2016	93153

Appendix B

Number of papers matching the given criteria (by cited count)

Cited count (based on PubMed)	Number of papers
0–9	184327
10–19	68409
20–29	36727
30–39	20271
40–49	13567
50–59	9483
60–69	6331
70–79	4831
80–89	4116
90–99	2523
Over 100	16448

Appendix C

Results of Topic Modeling

Topic no.	Topic label	Keywords				
Topic 1	Cellular senescence	Senescence Cellular	p21	Cyclin D1	p16	p16INK4A
Topic 2	Insulin signaling	Insulin	Beta catenin	Kidney	Neurons	Brain
Topic 3	Lung cancer genetics	Lung Cancer	Polymorphism	Alleles	Genotype	Lung
Topic 4	Human Papilloma virus (HPV)	Human Papillomavirus	c-myc	E6	HPV	MYC
Topic 5	Oxidative stress and cell death	Oxidative Stress	Reactive Oxygen Species	Mitochondria	Autophagy	Antioxidants
Topic 6	Gene methylation in cancer	DNA Methylation	Colorectal Cancer	APC	Methylation	Cell Line
Topic 7	Ubiquitin mediated proteasome degradation	Ubiquitination	Degradation	Ubiquitin	Ligase	E3
Topic 8	Cell cycle regulation	Cell cycle	Phosphorylation	PKR	S Phase	pRb
Topic 9	Stem cell gene expression	Stem Cell	miRNA	Gene Expression	Differentiation	Population
Topic 10	Inflammation	Inflammation	NF-kappaB	cytokine	NRF2	Activation
Topic 11	Central dogma (Biology)	mRNA	cDNA	Transcript	Nucleotides	Polypeptides
Topic 12	PI3K/AKT/mTOR pathway	mTOR	Phosphorylation	AKT	Kinases	PTEN
Topic 13	Metastasis	Angiogenesis	Endothelial Cells	Melanoma	Metastasis	VEGF
Topic 14	DNA damage signaling path way	DNA Damage	ATM	DNA repair	Ionizing radiation	Irradiation
Topic 15	Tumorigenesis	Tumorigenesis	BRCA1	Ovarian cancer	Epithelial Cells	P63
Topic 16	Transcriptional/epigenetic regulator	p300	Promoter	Curcumin	Binding Sites	CBP
Topic 17	Metastatic cancer	Metastasis	Breast Cancer	Prostate Cancer	WT1	Lymph nodes
Topic 18	Tumor suppressor genes	LOH	Amplification	Alleles	Cell Line	Glioblastoma
Topic 19	Apoptosis	Apoptosis	Cell Death	Bcl-2	Cytochromes c	BAX
Topic 20	Chromatin modification	Histones	Chromatin	SIRT1	Lysine	H3
Topic 21	Epithelial tumors	Pancreatic Cancer	Neck	Adenocarcinoma	Carcinoma	Lesion
Topic 22	ARF-Mdm2-p53 tumor suppressor path way	Mdm2	Cell cycle arrest	DNA damage	Tumor Suppressor	Apoptosis
Topic 23	Notch signaling path way	Signaling	Embryo	NOTCH1	Differentiation	Proliferation
Topic 24	Lymphoma	T-Cells	B-Cells	Antibodies	Lymphocyte	Lymphoma
Topic 25	Leukemia	Oncogenes	leukemia	C abl	AML	NPM
Topic 26	Adenoviruses	T-antigens	Adenoviruses	Virus	E1A	SV40
Topic 27	Drug resistance mechanism in lung carcinoma	Cell Line	EGFR	Cancer Cell	Resistance	Doxorubicin
Topic 28	Ca2+ Dependent nuclear export	Nuclear Export	Microtubules	Cytoplasm	ER	localization
Topic 29	Epidermal growth factor	FAK	Receptor	E-Cadherin	EGF	Growth Factor
Topic 30	Hypoxia signaling pathways in cancer metabolism	Hypoxia	TGF-beta	HIF-1Alpha	Cancer cell	Nitric Oxide

DOI: 10.1002/asi