A Study on Developing and Refining a Large Citation Service System

Kwang-Young Kim*, Hwan-Min Kim**

ARTICLE INFO

Article history:
Received 20 May 2013
Revised 28 May 2013
Accepted 5 June 2013

Keywords:
Citation System,
Citation Index,
Citation Service,
NoSQL

ABSTRACT

Today, citation index information is used as an outcome scale of spreading technology and encouraging research. Article citation information is an important factor to determine the authority of the relevant author. Google Scholar uses the article citation information to organize academic article search results with a rank algorithm. For an accurate analysis of such important citation index information, large amounts of bibliographic data are required. Therefore, this study aims to build a fast and efficient system for large amounts of bibliographic data, and to design and develop a system for quickly analyzing cited information for that data. This study also aims to use and analyze citation data to be a basic element for providing various advanced services to the academic article search system.

1. Introduction

The volume of information in the field of science and technology, including academic journals, has continued to increase during the last 30 years. E-publication and the developed Internet contribute to continuously increasing digital academic literature. The method of accessing the information is changing significantly as well. In particular, researchers share articles published or not published in academic journals through specific websites. Although various changing environments contributed to improving the performance of search systems, it is still not easy to find ideal literature.

The number of articles cited by researchers online is gradually increasing, more so than offline. The trend of using highly accessible academic articles is increasing, and it is now possible to access most articles of high quality online. Kim (2003) says that online academic literature which provides a hyperlink is accessible from cited articles. This online access is referred to as 'reference linking' or citation linking.

The citation index has been and is currently used in the field of information search for finding,

* Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information, Korea (kykim@kisti.re.kr)
** Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information, Korea (mrkim@kisti.re.kr) (Corresponding Author)
International Journal of Knowledge Content Development & Technology, 3(1): 65-80, 2013.
http://dx.doi.org/10.5865/IJKCT.2013.3.1.065

evaluating, and analyzing academic articles. The citation index, such as the SCI (Science Citation Index), collects citation information to be used as a scale for evaluating academic journals or articles. Web of Science, Scopus, and the like are citation-based online database systems, which most often use citation information. Google Scholar and the Naver Search System use citation information in academic articles as an important algorithm element for deciding the search result rank. Beel and Gipp (2009) studied the search system of Google Scholar to find that it provides article search service for each author, title, and subject, and how to track it through empirical studies, although the rank algorithm is not open . For the citation information in academic literature, the SCI database has been built for articles listed in academic journals in the global fields of science and technology since 1961, and was then integrated with Thomson ISI in 1992.

PLos ONE, a journal published by a public science library in the US, provides citation information via CrossRef. As of May 2013, the data from the Cited-by linking website CrossRef reveals that the number of articles published with bibliography accounts for roughly 40% of the entire DOI given to CrossRef, which includes 19,360,000 articles. The number of articles cited at least once is 24,490,000, and the total number of links is more than 336 million. As described above, because the citation information of CrossRef covers academic journals all over the world, and the number of articles and journals is huge, it is thus possible to examine a global tendency and the information is valuable bibliographic data.

Therefore, this study aims to develop a large bibliography system for efficiently processing large amounts of bibliographic data. This study also aims to suggest a method of various citation services by using the system to analyze citation information.

2. Related Studies

Ko and Choi (2005) wrote that bibliographies have been analyzed and studied since 1970 in Korea, in relation to electricity, electronics, mechanical engineering, medicine, and computing, and various studies have been made about bibliographical analysis in relation to the science of agriculture, mathematics, physics, chemistry, engineering, electricity, machines, veterinary study, economics, sociology, and pedagogy in other countries.

WoS (Web of Science) is a citation database system provided by Thomson Reuters, which allows access to databases, and provides the service of searching a particular sub-field in education and science studies to support the selection of the most influential literature in a particular field.

The DBLP (DataBase systems and Logic Programming) Computer Science Bibliography List of the Trier University of Germany is one of the index systems used by many researchers involved in computers. The DBLP provides author and article search service, has a journal list for articles, academic meetings, and books, and the author search includes a co-author search service. It also provides information about cooperation with authors for the relevant study.

An, Janssen, and Milios (2004) tried to plot computer science literature information in directional graphs, in which nodes represent articles, and arrows represent the articles which reference the relevant articles. This kind of literature graph is used to search for research areas or to measure

the relationship between research areas and to track how studies have temporarily developed.

The KISTI (Korea Institute of Science and Technology Information) provides the KSCI (Korea Science Citation Index) as citation index service and academic journal citation index service in the field of science and technology. It covers articles published in Korea's 760 major science and technology journals, and provides related indexes, including the influence index for 400,000 articles and 7.10 million bibliographies published since 2002, the number of citing articles, and the number of articles cited.

Jeong (2011) studied the CrossRef/DOI deposit project by KISTI since 2007. This deposit project is for supporting overseas distribution networks to go to international academic journals through citation outcome improvement of academic societies or institutions, the academic information producers. This is intended to manage participating academic societies and academic journals, DOI number, taking out CrossRef, deposit result and errors through the system for supporting academic article DOI registration. The Cited-by linking service provided by CrossRef on the basis of DOI is also available from the Science and Technology Society since 2009.

The National Research Foundation of Korea has the KCI (Korea Citation Index) for analyzing the citation relation of journals, which is similar to the SCI of ISI. The KCI was suggested to provide article information for Korea's academic journals and to complement the quality evaluation system. To this end, the KCI provides bibliographies, statistics, and citation information of Korea's articles, and uses the information collected through article evaluation and support projects to establish a citation index database. On the basis of the outcome, the KCI provides the citation and article information service.

3. Large Citation Systems

Jeong (2011) studied the available large data technology for prediction and analysis, to store a huge volume of data, to search and visualize meaningful data therein, and the technology to implement and apply the data to the business process.

This study intends to develop and build up large bibliographic data with a NoSQL-based system in order to analyze them fast by using various citation information, such as, for example, index information, author navigation, journal report, and citation map. It is also intended to visualize the service related to bibliographies in order to provide user-friendly services with the analyzed data.

3.1. Analyzing CrossRef XML and designing schema

Table 1 shows the bibliographic metadata acquired from CrossRef. The bibliographic metadata is composed of journal, journal_issue, journal_article, title, and citation list. The journal includes journal title, the abbreviation of the journal title, ISSN, DOI, and resources. The journal_issue includes issue information, DOI, and date of publication. The article includes article title, author, the first page, and the last page of the article.

Table 1. CrossRef Citation MetaData

```
<crossrefxmlns="http://www.crossref.org/..."
<journal> ...
<journal_issue> ...
<journal_article> ...
<titles><title>...
<citation_list> ...
...
</journal>
<journal>
...
</journal>
```

Therefore, the schema is designed to focus on journal, journal_issue, article, reference information, and co-authors to configure the bibliography system in this study. In particular, because of the large amount of articles, bibliographic information, and co-authors, the system is designed in consideration of variation.

Table 2. Architecture of journal schema

Column	Description
lang	language code value
full_title	journal title
abbrev_title	journal title abbreviation
issn_p/issn_e	e-publication ISSN number
doi	journal DOI identification number
resource	journal URL
journal_key	journal ID key value
…	…

As shown in Table 2, it is essential that the journal schema uses journal title, the abbreviation of the journal title, ISSN, and DOI (Digital Object Identifier). In particular, because abbreviations of the journal titles are generally used in bibliographic data, it is essential to have them for future journal matching. Because authors generally use journal title abbreviations to write their bibliography, they must be managed together. Journals can be found more accurately with ISSN or DOI, rather than with a journal title, because a journal title can have variations resulting from errors or journal title changes.

As shown in Table 3, the article schema information is composed of article title, year and month of publication, the first page of the relevant article, and DOI. It also has the journal key value for the relevant article and the key value of the issue information, in order to determine the relevant journal and issue information quickly. A key value is assigned within the architecture to enable journals, issues, and bibliographic information to be found easily, focusing on articles, as described above.

Table 3. Architecture of article schema

Column	Description
title	article title
year/month	year/month of publication
first_page	first page of article
last_page	last page of article
doi	article DOI identification number
resource	article URL
journal_key	journal ID key number
issue_key	issue ID key number
article_key	article ID key number
…	…

Table 4 shows the architecture of bibliographic data schema. That is, it is composed of bibliographic data values of the relevant article, for example, bibliography number, journal title, ISSN, author, and the first page of the article. If there is a doi value in the bibliographic data provided by CrossRef, other field values are provided blank. Therefore, if there is a doi value only to provide the bibliography service, it is necessary to find the relevant article to obtain information values including journal, article title, and year of publication and then to enter data values. It is also necessary to have an article key value for linking the relevant article to the bibliography list.

Table 4. Architecture of citation schema

Column	Description
key	bibliography No.
journal_title	journal title
volume_title	volume title
Issn	ISSN No.
author	author
volume/issue	volume value
first_page	first page value of the article
cYear	year of publication
unstructured	Original bibliography
doi	article DOI identification No.
resource	URL to the article
article_title	article title
article_key	article ID key value
…	…

3.2. Designing a Large Citation system

The recent data boom contributes to a great number of systems (150) for supporting NoSQL

(Not Only SQL). For example, exemplary systems include the column-oriented storage system, Hadoop/HBase and Cassandra, the document-oriented storage system, MongoDB and CouchDB, and the tuple-oriented storage system, DynamoDB and Redis. The aforementioned NoSQL systems have their own weaknesses and strengths. Therefore, it is very important to select an ideal system for the intended purpose.

In this study, MongoDB, a document-oriented storage system, is used for the bibliographic data, to manage data in document-wise manner while handling large amounts. MongoDB provides data sharding, called auto-sharding. Data sharding is essential to handle large amounts of bibliographic data. Sharding adds a device for addressing increased load and data without an impact on applications, in the case of frequent writing or if more disk space is required. MongoDB supports the replica-set to recover stored data safely.

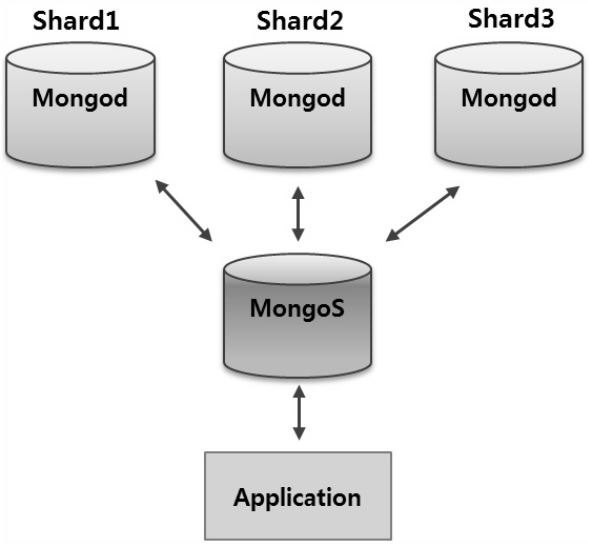

Fig. 1. Architecture of sharding system

In Figure 1, the bibliography system is composed of 3 Mongod servers, 3 Config servers, and one MongoS server. The MongoDB developers recommend adding the replica-set server in order to implement stable system services. First, the cited analyzer reads article data to examine DOI values. If there are DOI values, it uses the DOI identifier to search for citation data, and then to test cited data matching. It then stores only the matching information in the Cited DB.

Table 5 shows the architecture of DB collection in the bibliography system. Therefore, the system is composed of the JOURNAL collection with the information for journals, the ISSUE collection with the issue information for the journals, the ARTICLE collection with article information, the CITATION collection with bibliography information of the articles, and the CITED collection with cited article information.

Table 5. DB Collections and keys

Collection		ID Key
journal	JOURNAL	journal_key
Issue title	ISSUE	journal_key issue_key
Article	ARTICLE	journal_key issue_key article_key
Citation	CITATION	article_key citation_key
Cited	CITED	journal_key article_key citation_key cited_journal_key cited_article_key cited_citation_key
Co-author	CO_AUTHOR	article_key author_key
...	...	...

In this study, ID keys are essential for connecting and identifying collections. For example, it is essential that an article has more journal and issue information, and connection information. It is designed to have connection information for citing articles, and connection information for cited articles. Therefore, Figure 2 shows the relation between citing and cited. That is, the citing information means the reference data cited by an article, and the cited information means the articles which cite the relevant article.

In Figure 2, the cited collection has both citing information and cited information. This capability is used to provide navigation between authors or articles in the design.

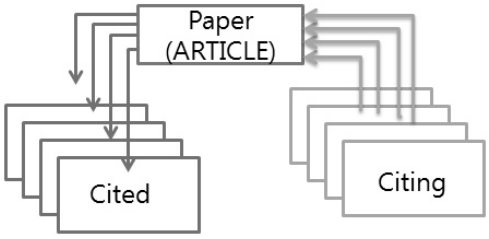

Fig. 2. Relation between article citing and cited

Table 6. Auto sharding collection information

Collection	Sharding ID key	Remarks
ARTICLE	journal_key	journal key-centered sharding
CITATION	article_key	article key-centered sharding
CITED	article_key	article key-centered sharding

The constructed CrossRef bibliographic data includes 22,000 journals, 28 million pieces of article collection data, 266 million pieces of citing collection data, 28 million pieces of cited collection data, and other analysis-type data, which makes 300 million data in total.

An analysis of the bibliographic data provided by CrossRef reveals that an article has approximately 10 bibliographic entries on the average. In this study, the system is designed to do auto-sharding for the collections of large articles, citing and cited, as shown in Table 6. General MySQL and data loading took 7 days and 9 hours, but MongoDB data loading took 2 days and 2 hours. The system designed as described above implements faster data loading than general RDBMS, and data search speed is also faster.

3.3. Algorithm for analyzing bibliography for being cited

Figure 2 shows how to analyze cited bibliography information. First, the cited analyzer reads article data to examine DOI values. If there are DOI values, it uses the DOI identifier to search for citation data and then to test cited data matching to store only the matching information in the Cited DB.

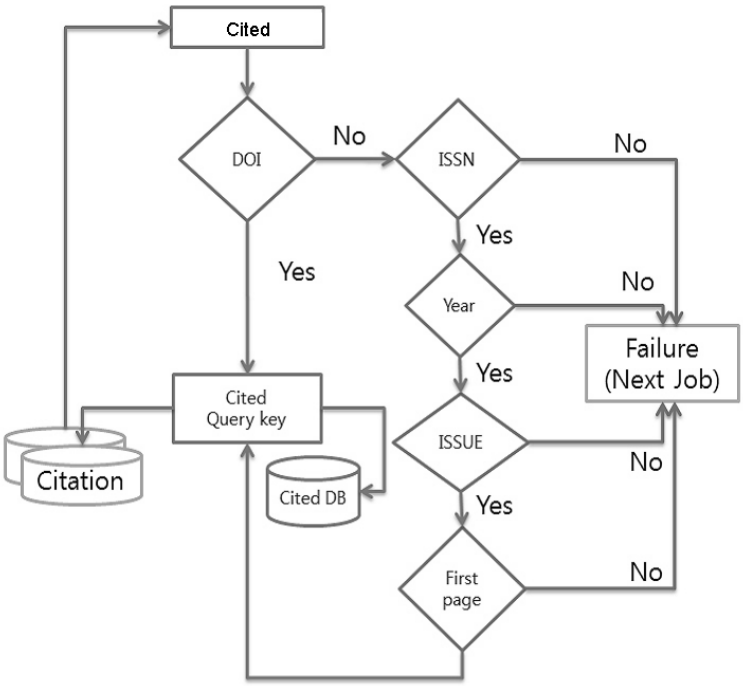

Fig. 3. Cited analyzer flow chart

Second, if there is no DOI value, it uses data including journal ISSN, the year of issue, and the first page of the article to carry out cited data matching, and stores only the matching information in the cited DB. Stored information includes journal key, article key, cited journal key, and cited

article key.

In this study, it is possible to create a journal report to calculate the Impact factor by using the cited mapping information. The system is also enabled to create networks between journals and article authors by using the cited information, and also to create a map between cited and citing articles. Therefore, the cited information is the most important factor to use the bibliographic data.

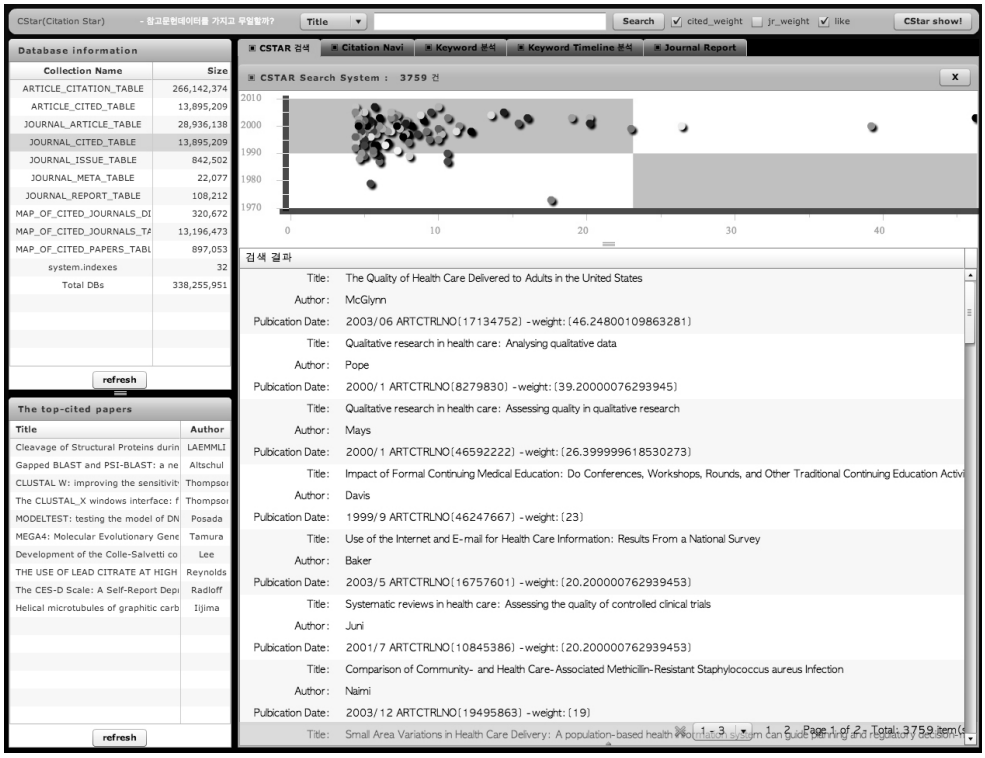

Fig. 4. Bibliography system user interface

In this study, 300 million pieces of CrossRef bibliographic data are used to build a NoSQL-based system as shown in Figure 4, and to develop an algorithm for analyzing data quickly. A method of various citation services is studied by using bibliography citing and cited information.

4. Method of refining citation service

Various services can be provided, including article authority, search ranks, journal reports, and bibliography maps if the citation data are used. Therefore, it is essential to establish bibliographic data in order to provide high value-added services related to academic articles. A method of refining various services has been studied and developed, in relation to bibliographies by using the established bibliography system.

4.1. Reflection on search rank algorithm

Figure 5 shows the greatest strength of using bibliographic data by using cited bibliographic data to reflect on the search rank algorithm. The result is shown in the higher rank of cited information for the searched articles by users. However, because articles of higher recognition in the past have more cited information, it is necessary to consider the latest years to reflect them on the search rank algorithm.

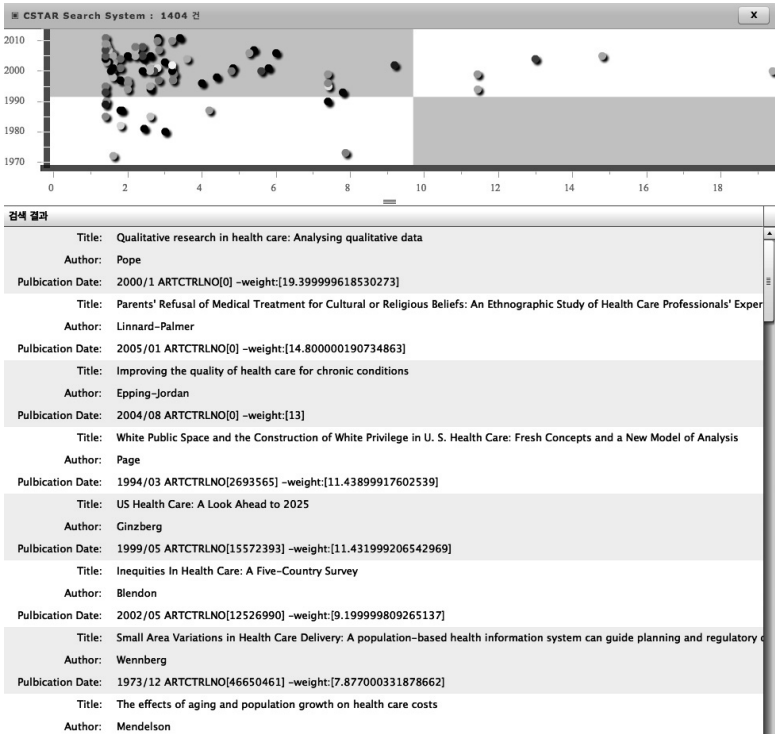

Fig. 5. Search rank algorithm

Figure 5 shows the visualized ranks for cited information and years with the axis X for document weight and the axis Y for years, to design a system which enables users to select search results in consideration of weight and years.

4.2. Citation service

When a user selects View for the searched article, the citing bibliographic information of the relevant article is provided, and the information of articles cited by other authors for the relevant article is also viewed, as shown in Figure 6. The trend of relevant cited article is viewed by showing the cited information for each year. Therefore, it is easy to examine how many times the relevant article has recently been cited, and was cited in the past.

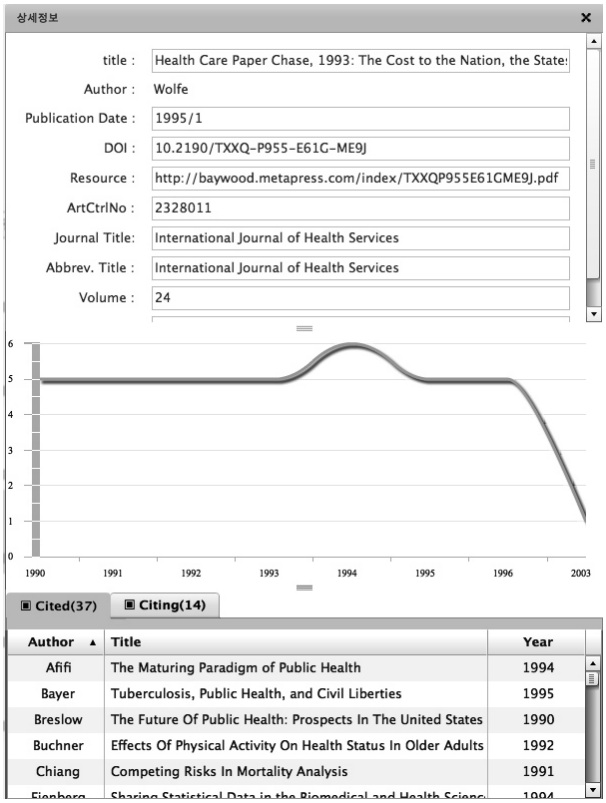

Fig. 6. Trend of cited and citing/cited information

With respect to the above article, it is shown that it continued to be cited in the past, but the level of citing the article has recently dropped.

4.3. Citing and cited information navigation service

Figure 7 shows navigation through article authors. All articles have citing information. However, they may not have cited information. Therefore, when a user searches for articles, they can be visualized in network nodes focusing on citing and cited information.

In this study, it is assumed that "an author more frequently cited has better recognition in the relevant field" in order to examine an author's impact. This service is used to examine the relationship between authors or the impact of the relevant author.

Figure 7 shows the result of citing and cited navigation among authors for the exemplary articles searched with "health care". The words in red are cited information and those in blue are citing information, with respect to implemented navigation through citing and cited authors.

For real-time search, the schema is configured to keep connection information focusing on articles, the ID key values of citing and cited collections, shown in Table 5. The user interface is designed so that users can control the number of citing and cited articles in real time.

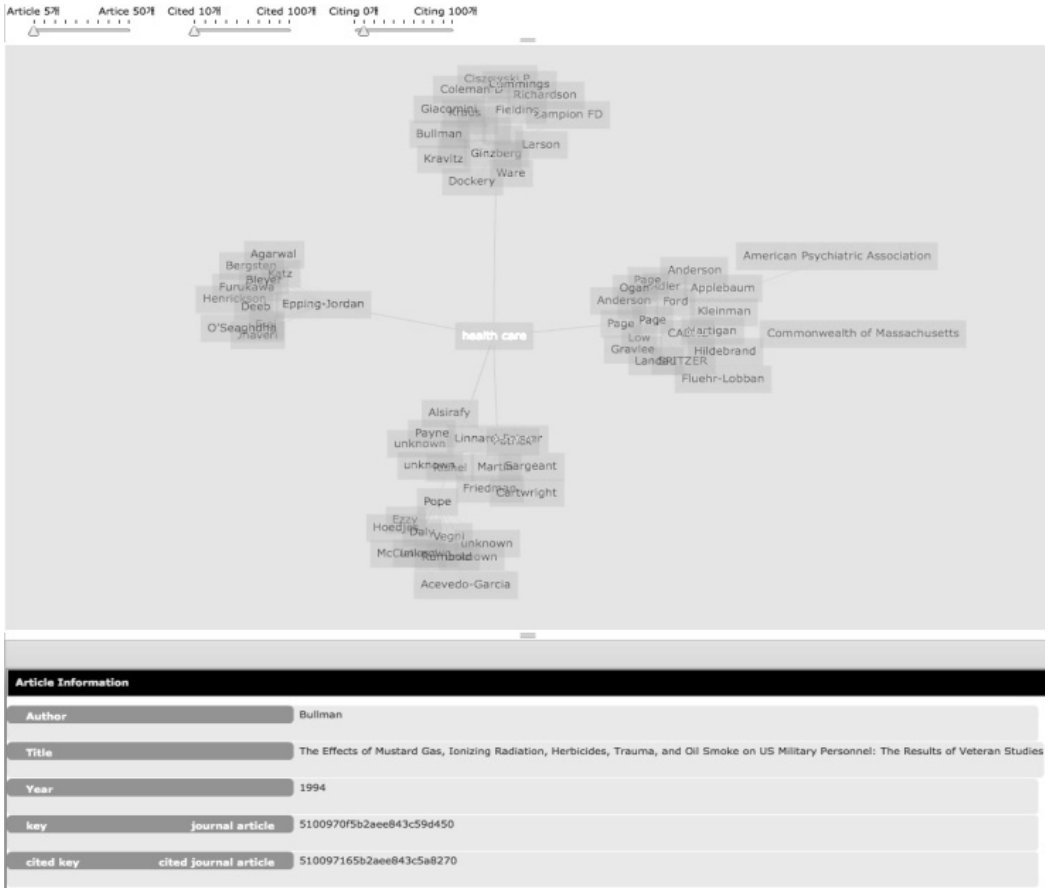

Fig. 7. Network among authors by using cited information

4.4. Service for searching relevant journals

Most users want to find articles and journals they want with a few keywords. Therefore, the most ideal journals can be recommended by using user's keywords to know which journal publishes articles focusing on them and to use the cited index of the relevant journal.

Figure 8 shows how to group and visualize the search results focusing on journals of higher cited indexes and which journals publish articles related to the relevant keywords by using the sample keywords "Health Care". The result is that the Health Affairs (in blue) has recently published many articles for the relevant key word, and has a higher citation index.

In this study, the user interface is designed to allow sampling of search results to be controlled between highest 1 to 100%. The 10 highest journals are grouped to visualize the result as shown in Figure 8. The result shows analyzed result for each year with respect to Health Affairs, American Journal of Public Health, and Community Health.

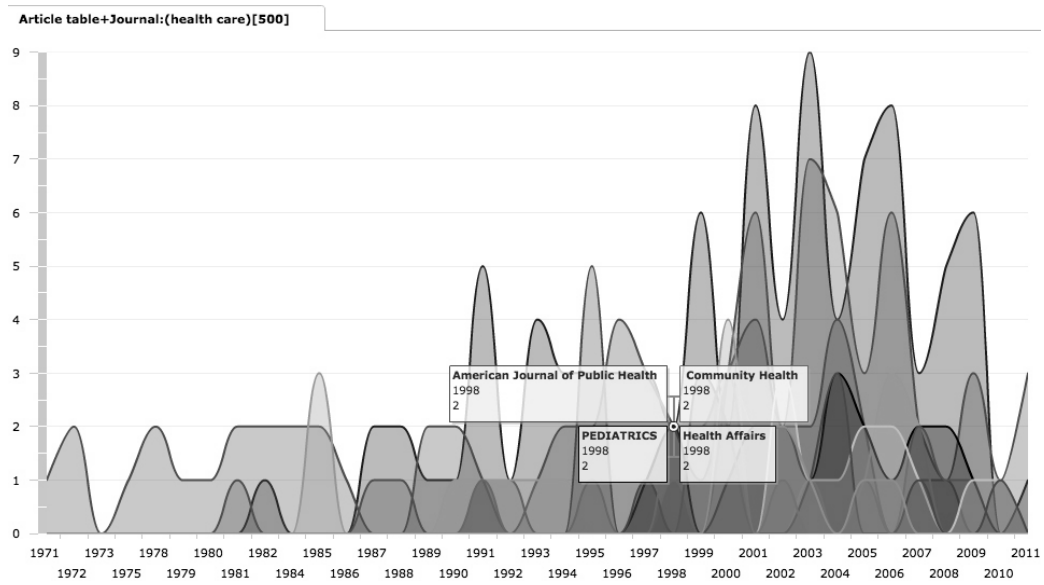

Fig. 8. Finding a relevant journal

4.5. Citation analysis service

The citation analysis service is the same as the citation analysis report in that they suggest bibliometric analysis and citation indexes for analyzed articles, but different from each other depending on whether the analysis is carried out in real time, or in a particular time range. That is, the citation analysis service is for bibliometric analysis of academic literature searched by using keywords. The citation analysis report is for major academic journals published in a given period, bibliometric analysis reports and indexes in the unit of institutions and countries. In this respect, Web of Science and SCOPUS correspond to the citation analysis service. JCR and SCIMAGO correspond to the citation analysis report. The most important service for using the bibliographic data is the citation analysis service and report.

Figure 9 shows how to calculate the Impact factor by using the cited frequency for the last 2 years. It is enabled to calculate the Immediacy index and also the Impact factor for the time period from minimum one to 5 years.

$$\text{Impact Factor} = \frac{\textit{Number of total citations for the latest 2 years}}{\textit{Number of total essays for the latest 2 years}}$$

In this study, the Impact factor for journals for each year is obtained for 22,000 journals by establishing the 300 million bibliographic data. Therefore, more bibliographic data would enable more accurate Impact factors to be obtained.

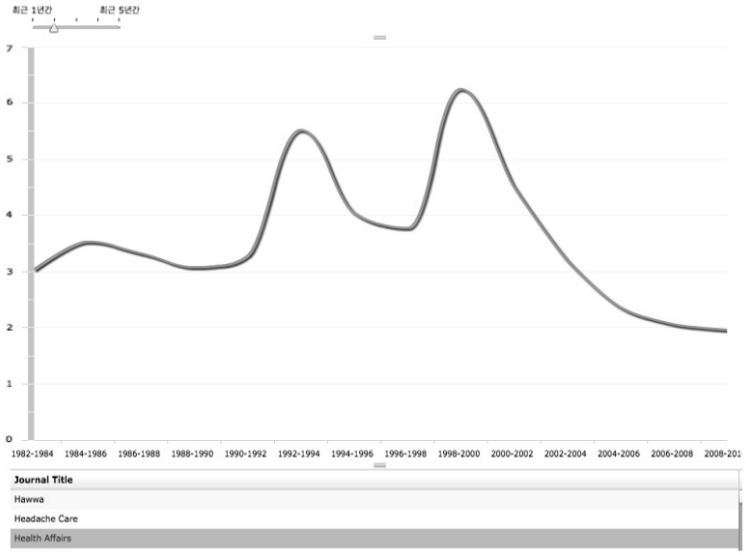

Fig. 9. Impact factor for journals

4.6. Citation map

In this study, the cited information is used to make a network among article authors and a network among journals. Figure 10 shows the results of using the cited information to make a journal map. This uses the citation index information to visualize the network relationship among journals. Therefore, the relation information among journals is known.

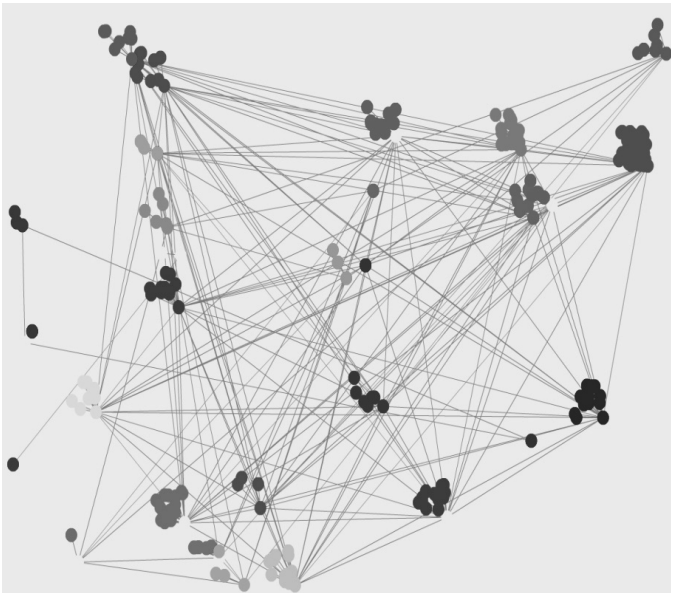

Fig. 10. Relationship among journals by using cited information

For example, the allergy journals have information interconnected with expert opinions on therapeutic patents, hematologic disorder journals, blood journals, AIDS, and the like.

In this way, users can know which journals are interconnected for studies by using the citation information among journals. It is also possible to classify topics of the above journals to analyze interconnected studies for which subject areas interact.

5. Discussion and Conclusion

Today, the citation index data are used as a scale for the outcome of technology diffusion and expanding studies. The citation index has been used in the field of information search for finding, evaluating, and analyzing academic articles. They are used as an important factor in deciding search ranks by using the citation information in Google Scholar and the Naver Search System.

Therefore, in this study, the NoSQL-based system is used to implement large data sharding in order to establish large amounts of bibliographic data, and a system is designed and developed for quick analysis of citation information.

To this end, collections of journals, issues, articles, and bibliographic data are built, and a unique identification system is configured to make connections fast. The system is configured to search cross-citation information among articles focusing on DOI. Without DOI values, information including journal ISSN, article title, author, issues, year of publication, and the first page of the relevant article is used to develop an algorithm for matching citation information quickly.

Various services were implemented and studied, including search rank algorithm, citation service, navigation, citation analysis service, and citation map, implemented by using the citation data on the basis of the development. In particular, the analysis of the relationship among journals by using the citation map reveals which academic fields interact for studies. Therefore, the citation data are a very important factor in the academic article system, and it is essential to establish and use the bibliographic data to refine various academic article services. It is necessary to analyze the influence of journals by establishing a wide variety of large citation databases. The citation data are basic, but used to provide various high value-added services to users, including studies among authors, the relationship between study areas, and analysis of temporal progress of studies on the basis of the above influence analysis.

It is necessary to further study bibliography extraction algorithms and matching algorithms from PDF files or XML files in order to continue to get bibliographic data. It is also necessary to build an identification and cross-connection system including metadata, author authority, bibliography, and e-original text in order to refine the academic article search service. An academic database focusing on the full-text XML data must also be established. The full-text XML data is composed of abstract, full-text data, figures, tables, and bibliographic data. Therefore, it is essential to implement high value-added services of the academic article data.

References

An, Y., Janssen, J., & Milios, E. E. (2004). Characterizing and mining the citation graph of the computer science literature. *Knowledge and Information Systems*, *6*(6), 664-678.

Beel, J., & Gipp, B. (2009). Google Scholar's Ranking Algorithm: The Impact of Citation Counts. *Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference on*, 439-446. doi:10.1109/RCIS.2009.5089308

Jeong, B. K. (2011). The future society and big data, 13-15, NIPA.

Jung, E. K., Kim, B. K., & Park, J. W. (2010). Application and effect analysis of DOI based service for national academic information, *Proceedings of the Korean Society for Information Management Conference Proceedings of the Korean Society for Information Management Conference*, 29-32.

Jung, E. K., Kim, B. K., & Choi, S. H. (2011). Application system plan for scholarly information producer using the result of cited-by linking based on CrossRef DOI. *Proceedings of the Korea Information Processing Society Conference*, *18*(1), 1573-1575.

Kim, J. H. (2003). A study on automatic extraction of citation information for reference linking. *Journal of the Korean Society for Library and Information Science*, *37*(1), 238-268. doi:10.4275/KSLIS.2003.37.1.247

Ko, S. S., & Choi, S. K. (2005). Analysis on foreign journals seeking behaviors through citation analysis. *Korea Library and Information Science Society Magazine*, *36*(1), 441-457.

Lee, J. Y., Yu, S. Y., & Lee, J. Y. (2010). Strategies for improving scholarly information service with citation information. *Journal of information Management*, *41*(1), 43-67. doi:10.1633/JIM.2010.41.1.043

Lee, S. H., Kim, H. M., & Choe, H. S. (2012). A preliminary study on the multiple mapping structure of classification systems for heterogeneous databases. *International Journal of Knowledge Content Development & Technology*, *2*(1), 51-65.

The CrossRef. (2013, March 9). Retrieved from http://www.crossref.org

The DBLP. (2013, May 6). Retrieved from http://dblp.uni-trier.de

The KSCI. (2013, May 6). Retrieved from http://ksci.kisti.re.kr/main/about.ksci

The Mongo DB. (2013, May 7). Retrieved from http://www.mongodb.org

The NOSQL. (2013, May 7). Retrieved from http://www.nosql-database.org

The Plos One. (2013, March 9). Retrieved from http://www.plosone.org/home.action

The Thomson returers. (2013, May 6). Retrieved from http://thomsonreuters.com

The KCI. (2013, May 14). Retrieved from https://www.kci.go.kr