A Technical Approach for Suggesting Research Directions in Telecommunications Policy

Junseok Oh[1] and Bong Gyou Lee[2]
[1] Communications Policy Research Center, Yonsei University, Seoul, Korea
[e-mail: jseok@yonsei.ac.kr]
[2] Graduate School of Information, Yonsei University, Seoul, Korea
[e-mail: bglee@yonsei.ac.kr]
*Corresponding author: Bong Gyou Lee

Received July 11, 2014; revised September 21, 2014; revised October 30, 2014; accepted November 25, 2014; published December 31, 2014

Abstract

The bibliometric analysis is widely used for understanding research domains, trends, and knowledge structures in a particular field. The analysis has majorly been used in the field of information science, and it is currently applied to other academic fields. This paper describes the analysis of academic literatures for classifying research domains and for suggesting empty research areas in the telecommunications policy. The application software is developed for retrieving Thomson Reuters' Web of Knowledge (WoK) data via web services. It also used for conducting text mining analysis from contents and citations of publications. We used three text mining techniques: the Keyword Extraction Algorithm (KEA) analysis, the co-occurrence analysis, and the citation analysis. Also, R software is used for visualizing the term frequencies and the co-occurrence network among publications. We found that policies related to social communication services, the distribution of telecommunications infrastructures, and more practical and data-driven analysis researches are conducted in a recent decade. The citation analysis results presented that the publications are generally received citations, but most of them did not receive high citations in the telecommunications policy. However, although recent publications did not receive high citations, the productivity of papers in terms of citations was increased in recent ten years compared to the researches before 2004. Also, the distribution methods of infrastructures, and the inequity and gap appeared as topics in important references. We proposed the necessity of new research domains since the analysis results implies that the decrease of political approaches for technical problems is an issue in past researches. Also, insufficient researches on policies for new technologies exist in the field of telecommunications. This research is significant in regard to the first bibliometric analysis with abstracts and citation data in telecommunications as well as the development of software which has functions of web services and text mining techniques. Further research will be conducted with Big Data techniques and more text mining techniques.

Keywords: Bibliometric Analysis, Text Mining, Telecommunications Policy, Web Service, NoSQL, Citation Analysis

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1042) supervised by the NIPA(National IT Industry Promotion Agency)

http://dx.doi.org/10.3837/tiis.2014.12.013

1. Introduction

The bibliometric analysis has used for analyzing academic literatures. The majority of methods in the analysis focus on the citation analysis and the classification of contents [1][2][3]. The traditional bibliometrics has been conducted with data from top journals or important literatures because of the computing resources and limited techniques [4]. Recently, the bibliometric analysis with massive literature data becomes possible due to various text mining techniques, distributed computing methods such as Hadoop, and management techniques of unstructured data [5]. The bibliometric analysis has majorly been used in the field of information science, and it is currently applied to other academic fields as well [2][4].

The bioinformatics is the representative field and the text mining techniques are recently used for the bibliometric analysis. An author co-citation analysis was conducted in medical informatics [6]. This research used four analysis methods: the author citation analysis, the cluster analysis, the factor analysis, and the multidimensional scaling. The author co-citation analysis is one of representative citation analysis methods. The authors conducted the analysis based on co-citation frequencies with data from ISI Science Citation Index database in Web of Science (WoS). The Pearson correlation matrix and the factor analysis are used for clustering author groups which are related in similar research area. The last method is to visualize the similarity of authors. This research provides comprehensive methods for citation analysis in a bioinformatics field. The bibliometric study based on the classification methods was conducted in the bioinformatics field [7]. The authors selected major journals from PubMed which is the publication database for bioinformatics and analyzed the trend of studies with the data from PubMed. They reported the quantity of publications by years, languages, publication types, countries, and institutes. The authorship patterns were analyzed in this research as well but the authors did not provide the citation analysis results based on the author information. Bansard et al. showed research domains in bioinformatics and medical informatics. The authors identified upcoming trends in two fields by bibliometric analysis [8]. They used the factor analysis and the bigram analysis methods for finding major research topics in two fields. Their results implied that both fields are seemingly correlated but they are still separated. For instance, ‘genetics’ and ‘proteomics’ terms are presented in bioinformatics but they are not shown in medical informatics, while the latter focused on hospital information and patient management but the former had few research for them. Song and Kim analyzed literatures to identify research domains in bioinformatics based on full-text articles from PubMed Central database [4]. The authors conducted the analysis of words relations and the citation analysis, and they found significant patterns in articles and citations for research in bioinformatics. They also classified the research by years, countries, and institutes by the Named Entity Recognition (NER) technique. The log-likelihood method of the author co-citation analysis was used for finding the strength of correlations between authors.

The researches on the text mining exist in the field of information and communications technology (ICT). Carbonell et al. used the bibliometric analysis for searching appropriate research papers in Internet, video games, and cell phone addiction [9]. The authors did not use traditional text mining techniques. Instead, they classified the research articles which are retrieved by a particular term, and checked the relevance of retrieved articles with the term. The research on the application of a text mining technique to telecommunication services was conducted for improving the services [10]. The authors analyzed the pattern of customer enquiries for constructing semantic contents. They used the co-occurrence technique for

efficiently classifying customer enquiries. Also, they suggested that telecommunication operators use this method in an operation support system. Tsui el al. collected 22,000 terms from six IT magazines, and they explored the relationship among 50 information technologies [11]. The co-occurrence of terms are computed and the matrix is constructed prior to the clustering step. The hierarchical clustering and multidimensional scaling methods are used for clustering terms and for visualizing seven topic clusters. The research on the cognitive network of consumbers by Big Data mining for the marketing strategies [12]. The authors collected words related to brand from blogs and analyzed the similarity of brands by mining techniques. They computed pairwise Jaccard similarity coefficietns and decided the inter-category brand preferences by chi-square test. Hong et al. used text mining techniques for building the SPAM mail detection system [13]. The K-mean and the Hcluster algorithms are used for clustering of terms. Also, the stemming and the stopword techniques are applied as the pre-processing of Spam corpus.

Although there are attempts to apply text mining techniques to ICT and computer science fields, few research for the bibliometric analysis exist in the fields. Thus, our research suggests the use of bibliometrics based on text mining techniques for analyzing research trends and for finding empty research area in ICT. The telecommunications is selected as the research field in our paper because its technologies are rapidly developed, and related policies have to be changed in accordance with fast development. Furthermore, the greater interests in the Big Data analysis techniques increase the movements in the applications of unstructured data from social network services, online news articles, and academic research articles as well as opinions of experts for policies in a particular field. Our research provides the fundamental framework based on unstructured database system for mining knowledge structures from massive dataset. We identify research trends in the telecommunications policy by text mining techniques and unstructured database system. Also, the insufficient researches and the future directions are discussed on the basis of the results. This paper consists of four sections. The entire analysis process is described and text mining techniques are introduced in section 2. Section 3 shows data collection results and bibliometric analysis results. Specifically, KEA and co-occurrence results are shown in sub-section 3.2, and the results of citation analysis are described in sub-section 3.3. Conclusions with limitations of this work are discussed in section 4.

2. Data Collection and Analysis Methodology

2.1 Analysis Process

Fig. 1 shows the entire text mining process with web services. The data for research articles is retrieved from Thomson's WoS database. The Extensible Markup Language (XML) data is received from the database by the web service and the data is collected by XML parser. The parsed data is classified into the abstract data and the citation data. They are stored into CouchDB via RESTful web service. The frequencies of author defined keywords are computed by the word count module in the developed software for evaluating top keywords in collected publications. Also, topics of the collected articles are extracted from the abstract database by Keyword Extraction Algorithm (KEA) which is suggested by [14]. The topics are compared with top author defined keywords. The word frequencies in the collected abstracts are calculated, and the terms in all abstracts are converted to term-document matrix form. The word frequency data is stored in CouchDB. The frequency data are used for computing word co-occurrences as well as for visualizing word frequency networks from the abstracts. The

word co-occurrence analysis results are visualized with KEA results for showing the relations of interesting research terms. Additionally, the important articles in the telecommunications policy are selected by PageRank algorithm with citation data.

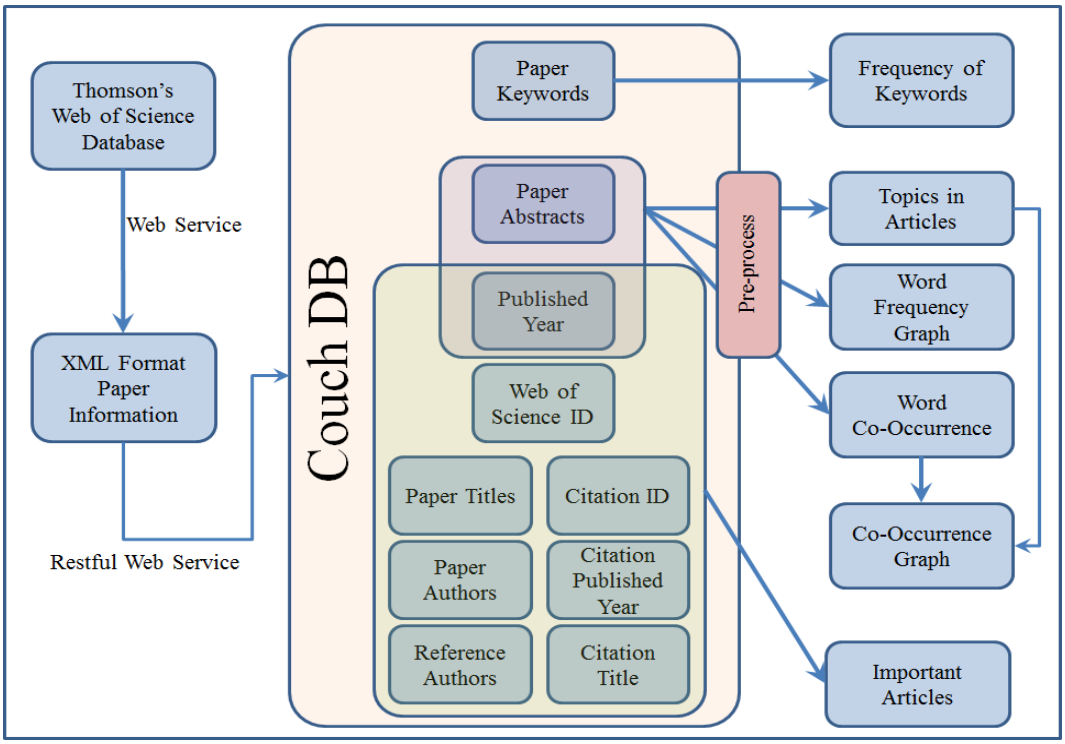


Fig. 1. Analysis Process Diagram

2.2 Data Collection

The data for our research is collected from WoS database. Thomson Reuter provides the WoK web services and users access the database by Web Service Definition Language (WSDL). The web services return search results in XML type. The commercial client software for web services can be used for consuming the WoK web services, but the web service client software on the basis of Java program language is developed for this research because it processes the consuming web services, the parsing data, and the converting data according to purposes of our research. Thomson Reuter provides three web services: the 'Authentication', 'WokSearch', and 'WokSearchLite' [15]. The 'Authentication' web service is to validate authorized users. The authentication is performed by user registration or Internet Protocol (IP) registration. We used IP registration and the authorization is completed by consuming the web service with its WSDL file. The session ID is randomly created by the web service, and it is used for consuming the data retrieve web service.

Two web services are provided for retrieving data, and only 'WokSearch' web service is used in this research. 'WokSearchLite' web service provides 'search', 'citingArticles', 'relatedRecords', and 'retrieveById' operations, while 'WoKSearch' web service provides 'citedReferences' operation in addition to above operations. The information for this research is not only the author defined keywords, abstracts, and publication years but also citing articles and cited references. The 'search', 'citingArticles', and 'citedReferences' operations are

required to obtain this information. Particularly, articles which cite the retrieved papers are generally used for the citation analysis, but PageRank analysis uses the articles which are cited by the retrieved papers as well. Therefore, the 'WokSearch' web service which includes 'citedReferences' operation is used at the data collection step of our research.

The query parameters and the retrieve parameters are required for consuming the 'WokSearch' web services. The query parameters are 'databaseId', 'timeSpan', 'queryLanguage', 'editions', and 'userQuery'. The 'databaseId' parameter indicates the database to search and the prefix 'WOS:' is attached to database Id in WoK web services. The 'timeSpan' parameter is a range of publication dates. We only substituted '2013-12-31' as the end publication date. Because this research focuses on papers which are written in English, the 'queryLanguage' parameter becomes 'en' which means English. Although only English articles are retrieved, some abstracts include other language abstracts if the article has more than one abstract. The abstract which are not written in English will be excluded by post-processing and it will be discussed in section 2.3.1. The 'editions' parameter means the paper collection types such as SCI, SSCI, AHCI. The 'editions' parameter string is defined as a blank since the papers in all collection types have to be collected in this research. The 'userQuery' parameter is used for defining the search field, and the field tags are in 26 types including topic (TS), author (AU), and title (TI). The topic is only used for the search field and its values are the combination of string 'telecommunications', 'Internet', 'broadband', 'policy', 'regulation', and 'economy'.

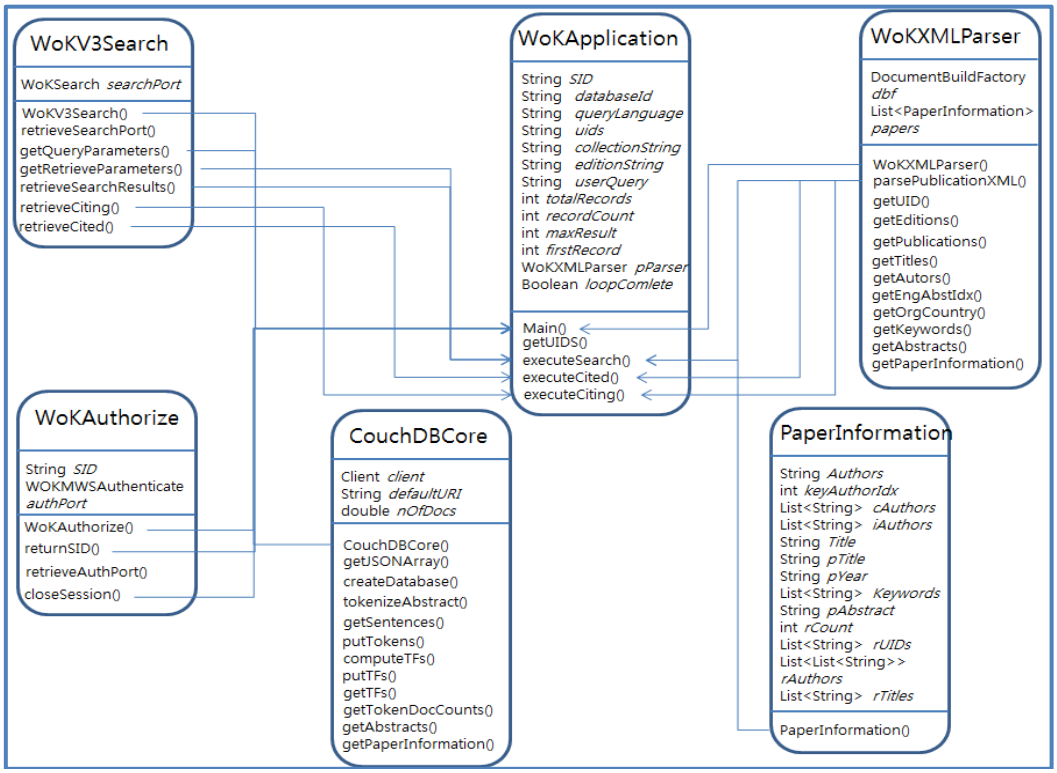


Fig. 2. Components of Data Collection Software

The retrieve parameters include ‘firstRecord’, ‘count’, and ‘sortField’. The ‘WokSearch’ web service returns up to 100 records for each search request. The multiple XML documents will be returned when the result includes more than 100 records. The ‘firstRecord’ parameter means first record number in each result (XML document) and this number must be greater than zero. The ‘count’ parameter means the number of records to display in each result and it cannot exceed 100. The ‘sortField’ parameter is used for sorting the results and the parameter has 14 types. The records are sorted by publication year and the value of this parameter is set as ‘CY’.

The return values of ‘search’ operation in ‘WokSearch’ web services consists of the number of total records, the number of records in each result, and the record information in each result. Since each result returns the maximum of 100 records, the Web Service module of the developed software adds the number of records to the ‘firstRecord’ value in each result (100) for increasing the ‘firstRecord’ value of the next result. Also, the software iterates this loop until the ‘firstRecord’ value meets the number of total records. The record information in each result is parsed by the ‘XMLParser’ module of the software. They are stored in CouchDB by the Database module of the software. The records related to citations of retrieved articles are collected by two operations in the same way. We will discuss the results of collected data in section 3.1.

CouchDB is used for storing retrieved data. It is an open source database which is developed by Apache Software Foundation. The unstructured data in the document is easily stored in CouchDB and the database captures a self-contained document [16]. Since CouchDB is not a relational database, it does not use Structured Query Language (SQL). Instead, the data is stored in JavaScript Object Notation (JSON) format by RESTful web services. CouchDB also includes MapReduce functions for improving the efficiency of data query [17]. The MapReduce plays a role of index in the database, and it is used for counting terms in an abstract and for creating elements of the term-document matrix in our research. The ‘PUT’ type RESTful web service creates database and stores tokenized terms of abstracts, keywords, reference documents information, and term frequencies in each document. The ‘GET’ method of the web service retrieves words, keywords, term frequencies, and citation information. The CouchDB for our research has the documents database and the references database. The parsed XML information is stored in each database. The documents database is a set of documents. Each document has collected paper information such as keywords, abstracts, titles, publication years, and authors. The references database has not only authors of reference papers, titles, publication years, and journal titles but also IDs of reference paper with IDs of collected articles which are in the documents database. The IDs will be used for the citation analysis to match a collected paper with its references. The information about the organization and country of authors is also gathered for our research but it will be used for the further research.

2.3 Analysis Methodology

Since this paper suggests applications of bibliometric approach to telecommunications policy area, the technical methodology for text mining with equations are introduced in this section. We describe text mining techniques for helping understands of readers who are not familiar with text mining.

2.3.1 Pre-processing and Keyword Extraction Algorithm

First, KEA is used for finding keywords of whole collected documents in this paper. The KEA automatically extracts keyphrases on the basis of the Naïve Bayes machine learning algorithm [14]. The algorithm consists of the training stage and the extraction stage. The training stage identifies keyphrases by a model which uses the training data. The keyphrases of collected documents are chosen on the basis of the model at the extraction stage. The algorithm identifies candidate phrases at both stages. The pre-processing techniques for cleaning up phrases need to be applied to the identification of candidate phrases. The non-alphabetic words are excluded, long phrases and stopwords are eliminated, and stemming technique is applied in this step. The stopword, the stemming, and the lemmatization are representative pre-processing techniques in text mining. The stopword technique is to remove common words which frequently appear but not valuable in the text mining analysis. The example stopwords are ‘a’, ‘an’, ‘be’, ‘that’, ‘was’, and ‘with’. The stemming technique is to make words as root forms. The example stemming is to reduce words ‘provide’, ‘provided’, ‘providing’, and ‘provider’ to the word ‘provid’ [18].

To extract new keyphrases, the model uses two features which are term frequency-inverse document frequency (tf-idf) and the first occurrence from candidate phrases [14][19]. The tf-idf is a popular weighting scheme in text mining. It is used for word classifications and ranking words. It is the method to find words which frequently appear in a document but rarely appear in a entire corpus. As shown in equation (1), tf-idf combines the term frequency and the document frequency. The term frequency indicates the number of a word (it presents the phrase in KEA) in a document, while the document frequency means the number of documents which contain a word in the collected corpus. If a term commonly appears in the entire collection, document frequencies divided by the number of total documents will be increased. Conversely, the inverse of this number will be high when the term rarely appears in the collection. Thus, the idf value is used in evaluating the importance of a word to a document in a global corpus.

$$\mathrm{tf-idf} = \frac{f(t,d)}{N(d)} * log\frac{N(d,w)}{f(d,t)} \tag{1}$$

where, *f(t,d)* is number of a particular word *t* in a document *d*

N(d) is number of words in a document *d*

f(d,t) is number of documents which contain a particular word *t*

N(d,w) is total number of documents in the global corpus *w*

The first occurrence is the value to present the distance of the first appearance of the phrase in a document. It is computed by the number of words which precede the first appearance of the phrase divided by the number of words in a document. This number is used in weighting the probability of a keyphrase.

The keyphrases of collected documents are determined by the model with the tf-idf and the distance features. The equation (2) presents the Naïve Bayes model with tf-idf and the first occurrence.

$$\mathrm{P(yes)} = \frac{Y}{Y+N} P_{tf-idf}(t|yes) * P_{distance}(d|yes) \tag{2}$$

where, Y is the number of author identified keyphrases,

N is the number of candidate phrases which is not keyphrases

t is a feature value for tf-idf

d is a feature value for distance

$P_{tf\text{-}idf}(t|yes)$ is the probability of a tf-idf value of a term in test documents when the term meets an author identified keyphrase in training documents

$P_{distance}(d|yes)$ is the probability of a distance value of a term in test documents when the term meets an author identified keyphrase in training documents

$P(yes)$ is the probability where a candidate keyphrase in test documents becomes an author identified keyphrase in training documents.

Therefore, $P(yes)$ is the ratio of identified keyphrases to total keyphrases with the probability of the tf-idf and the probability of the distance when the candidate keyphrase is author identified keyphrases. Similarly, $P(no)$ means the probability of no keyphrases. The overall probability of keyphrases becomes the ratio of the probability of identified keyphrases to the probability of total keyphrases and it is shown in the equation (3).

$$\mathrm{P}_{keyphrase} = \frac{P(yes)}{P(yes)+P(no)} \tag{3}$$

2.3.2 Word Co-occurrence Analysis

The word co-occurrence analysis is conducted for finding the major concepts in collected documents. The word co-occurrence is the number of occurrences of two words in a corpus and two words appear alongside each other. The frequency of co-occurrences has the meaning itself, but additional measure is applied to the word co-occurrence analysis. The pointwise mutual information (PMI) is the representative method to evaluate word association by Church and Hanks [20]. Statistically, the mutual information (MI) is the method to estimate the mutual dependence of two random variables. It is used for estimating the word association. The MI, $I(x,y)$, is computed as equation (4). $P(x)$ is the probability of x, $P(y)$ is the probability of y, and $P(x,y)$ is a joint probability of two variables x, y.

$$\mathrm{I(x,y)} = \log \frac{P(x,y)}{P(x)P(y)} \tag{4}$$

From here!!!!! The joint probability $P(x,y)$ corresponds to the co-occurrence in PMI. It is computed by the number of appearance of both word x and word y divided by the number of documents (N), where $P(x)$ is the frequency of word x divided by N and $P(y)$ is the frequency of word y divided by N. The higher MI indicates the more correlation between two words. According to Bouma (2009), MI and PMI have same theoretical backgrounds and both are generally used in the fields of computational linguistic and information retrieval, but the behaviors are not similar and PMI has a problematic issue. For example, the word pairs which

have low frequency are able to get high PMI values [21]. Other statistical algorithms to estimate the word co-occurrence are the chi-square test and the log-likelihood ratio [22]. The chi-square value is calculated by the existence of two words and the number of documents which have or do not have the words. The expected chi-square value is calculated as equation (5). The increase of chi-square means the stronger correlation between two words.

$$\chi^2(\mathrm{x},\mathrm{y}) = \frac{N(ad-bc)}{x_{yes}x_{no}y_{yes}y_{no}} \tag{5}$$

where, a is the number of documents which have both word x and y,
b is the number of documents which have word x but does not have word y
c is the number of documents which have word y but does not have word x
d is the number of documents which do not have either word x or y,
x_{yes} is $a+b$, x_{no} is $c+d$, y_{yes} is $a+c$, y_{no} is $b+d$, N is $a+b+c+d$

The statistical methods in text analyses are based on known distributions such as the normal distribution and the chi-square distribution. The chi-square test for the word co-occurrence is an example method based on the chi-square distribution. Although the measure is useful in the word co-occurrence analysis, it still has the problem by the assumption of the normality. The log-likelihood ratio test which is suggested by Dunning is based on the unknown parameters of model [23]. The likelihood ratio is computed as follows.

$$\lambda = \frac{max_p L(p,k_1,n_1)L(p,k_2,n_2)}{max_{p_1p_2} L(p_1,k_1,n_1)L(p_2,k_2,n_2)} \tag{6}$$

where, $L(p,k_1,n_1) = p^{k_1}(1-p)^{n_1-k_1}$, $p = \frac{k_1+k_2}{n_1+n_2}$, $p_1 = \frac{k_1}{n_1}$, $p_2 = \frac{k_2}{n_2}$
k_1 is the frequency where the word x appears and the word y follows
k_2 is the frequency where the word x appears and is followed by the word which is not the word y
n_1, n_2 are the number of word y and the number of words excepting the word y

The likelihood ratio is used in the form of logarithm. If two consequence words have a large log-likelihood ratio, they have a high correlation. We used the log-likelihood ratio for evaluating word co-occurrence because it provides more acceptable results than the chi-square test by solving the problem of the assumption of the normality.

2.3.3 Citation Analysis

PageRank is an algorithm to calculate the importance of web pages, which is suggested by Page and Brin. PageRank shows rankings of web pages based on the graph of the web [24]. The importance is basically computed by the count of links, but the number of links is not equally counted for all pages. Instead, PageRank is normalized by the number of links [25]. It is the probability of a random surfer's behavior which randomly visits pages when the surfer gets bored. The parameter called the damping factor is suggested in PageRank. It is the

probability that a random surfer gets bored and clicks a link to visit another page. The factor is set between 0 and 1, and generally set as 0.85. The equation (5) presents *PR(A)*, the PageRank of a page *A* which has pages *T1* …*Tn*. This method becomes the core of Google search engine [26].

$$\mathrm{PR(A)} = (1 - \mathrm{d}) + \mathrm{d}\left(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)}\right) \tag{7}$$

where, *PR(A)* is PageRank of page *A*,

PR(T1)…*PR(Tn)* are PageRanks of linking pages *T1*…*Tn*,

C(T1)…*C(Tn)* are the number of links going out of pages *T1*…*Tn*,

d is the damping factor

The PageRank is applied to the citation analysis for ranking authors of papers [27]. According to Ding et al., important factors of PageRank in citation analysis are the co-citation with important authors. A paper corresponds to a page, and a link becomes a citation of PageRank in co-citation networks. Song and Kim also suggested that PageRank is an appropriate algorithm to find important articles because it overcomes the problems in citation counts, which reflect the citations by highly cited publications but do not weight the citations by not highly cited papers [4]. Therefore, PageRank algorithm is used in our research and the results are compared with the results of citation counts.

3. Analysis Results

3.1 Data Collection Results

The collected raw data by WoK web services consists of the WoS IDs of articles, the publication years, the paper keywords, and the IDs of papers which are cited by the articles and cites them. Although WoK web services provide more information of papers including publication date, publication source, authors, and authors' organizations shown in **Fig. 3**. Only necessary information is filtered by 'XMLParser' module in the developed software. The ID of a article which has 15 digit number with 'WOS:' prefix is parsed from the 'UID' element, and then it will be stored as a document ID in CouchDB. The publication year information is collected by filtering the 'pubyear' attribute of the 'pub_info' element in a result XML file.

The element includes publication year, month, type, issue, and page information, but only the publication year attribute is parsed because our research focuses on the difference of research trends at the year level. The abstract information is collected by parsing the paragraph element ('p') in 'abstract' element. The 'abstract' element is nested in the 'abstracts' element because a couple of papers include more than one abstract in different languages. Thomson's WoS database stores an English abstract as first abstract (value of the abstract_text count element is 1). The 'XMLParser' module investigates the 'abstracts' element first, and it only parse 'p' element of first abstract if the number of abstracts is greater than one (the 'count' attribute of the 'abstracts' element is 1). The keywords of papers are collected by parsing the 'keyword' elements for comparing the paper keywords with selected keywords by KEA algorithm as well.

The abstract is converted to a term-document matrix which includes the number of terms in each document. The term-document matrix shows the frequencies of words which occur in

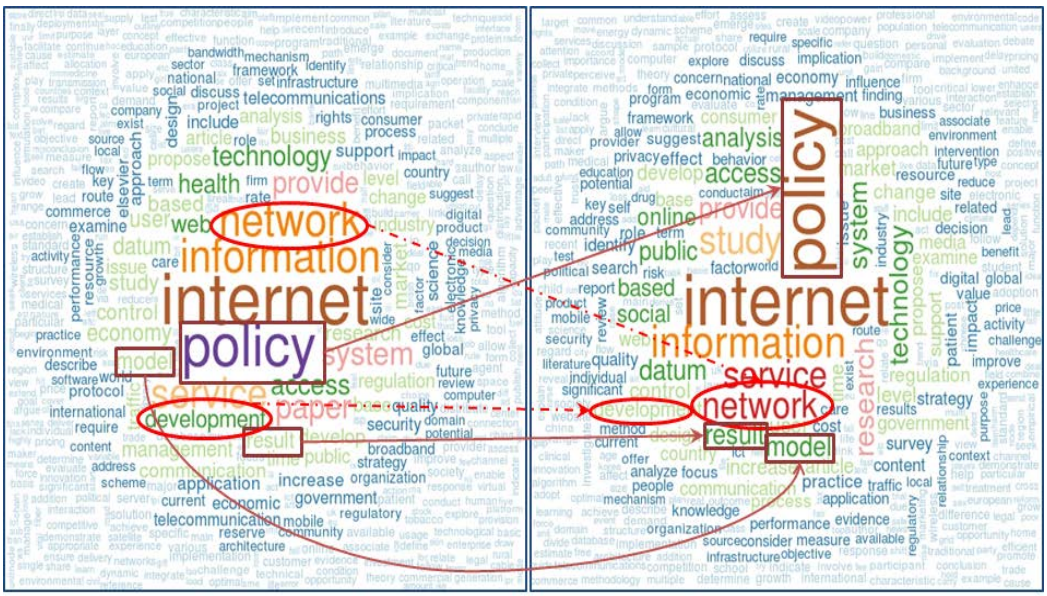
collected documents. The matrix is generally converted to another form such as tf-idf for the applications in natural language processing, but it is created for the input of R software to visualize term related frequencies in our research. The abstract sentences are break into terms by 'tokenizeAbstract' function and the tokens are stored in each document of CouchDB by the 'putTf' function of the developed software shown in **Fig. 2**. The software executes the text pre-processing module for excluding words which impedes the accuracy of the text mining analysis before counts the term frequencies. The stopword and the stemming techniques are applied at the pre-processing step. Some abstracts include unnecessary words related to 'Elsevier company', such as 'Elsevier Ltd. All rights reserved.' and 'Published by Elsevier Inc'. The 'XMLParser module of the software excludes these words in the abstracts as well. The 'computeTf' function computes frequencies of terms in each document of CouchDB, and the filtered words by pre-processing algorithms are stored as the word and frequency pair form in CouchDB.


```
    - <headings count="2">
        <heading>Social Sciences</heading>
        <heading>Science & Technology</heading>
      </headings>
    - <subheadings count="1">
        <subheading>Technology</subheading>
      </subheadings>
    - <subjects count="9">
        <subject ascatype="traditional">Communication</subject>
        <subject ascatype="traditional">Information Science & Library Science</subject>
        <subject ascatype="traditional">Telecommunications</subject>
        <subject ascatype="extended">Communication</subject>
        <subject ascatype="extended">Information Science & Library Science</subject>
        <subject ascatype="extended">Telecommunications</subject>
        <subject ascatype="traditional" jcr_rank="13/72" jcr_quartile="Q1">COMMUNICATION</subject>
        <subject ascatype="traditional" jcr_rank="17/85" jcr_quartile="Q1">INFORMATION SCIENCE & LIBRARY SCIENCE</subject>
        <subject ascatype="traditional" jcr_rank="18/78" jcr_quartile="Q1">TELECOMMUNICATIONS</subject>
      </subjects>
  </category_info>
- <keywords count="5">
    <keyword>Investment</keyword>
    <keyword>Incentive regulation</keyword>
    <keyword>Access regulation</keyword>
    <keyword>Local loop unbundling</keyword>
    <keyword>Telecommunications</keyword>
  </keywords>
- <abstracts count="1">
  - <abstract>
    - <abstract_text count="1">
        <p>Investment in broadband communications and its infrastructures (the so-called Next Generation Networks) is receiving
          from policy makers all over the world, due to the significant impact of high-speed Internet access on the whole economy
          even before the recent financial crises, a dramatic downward trend in telecommunications investment has occurred, mai
          to incumbent operators - to excessively intrusive regulatory intervention. The typical conflict between regulation, compe
          emerges. It is therefore important, for both future research and regulatory and practitioners' references, to review the s
          branch of the literature on this interesting and policy-relevant issue. The purpose of this paper is therefore to survey the
          empirical literature on the relationship between regulation, at both retail and wholesale level, and investment in telecom
          picture that emerges is not conclusive, and further research is still needed, both theoretically and empirically, to better u
          of regulatory incentives on investments. (C) 2009 Elsevier Ltd. All rights reserved.</p>
      </abstract_text>
    </abstract>
  </abstracts>
</fullrecord_metadata>
```


Fig. 3. Retrieved raw data by WOS Web services

The filtered data is stored into two databases which are the database for before the year of 2004 and the database after the year of 2003 in CouchDB. Total of 1035 documents are collected from the year of 1990 to the year of 2003, and 3668 documents are collected from the year of 2004 to the year of 2013 by WoK web services. Total of 9200 and 18793 distinct terms are aggregated by the developed software. The number of collected documents after 2003 is almost three times as many as them before 2004, but the documents after 2003 has almost twice terms as compared with before 2004. It implies that research topics before 2004 are more diverse despite of small amount of researches, whereas more analogous researches are conducted after 2003 in telecommunications policy.

The developed software retrieves the term frequencies with terms and documents information, and it creates the term-document matrix file. **Fig. 4** shows the wordcloud results which draw term frequencies from the term-document matrix. The wordcloud is a programming code package which is provided in R software. It shows the importance of words by term frequencies in all documents [28]. The higher frequencies of words in the collection are presented as the larger text size in the wordcloud. The left picture of **Fig. 4** is a result for the word frequencies in the abstract of papers before 2004 and the right one is the result after 2003. Both pictures show the word 'Internet' has the most frequencies in the abstracts. The word frequencies of 'network', 'system', 'web', 'development', and 'traffic' are relatively reduced in abstracts after 2003. Also, the words 'economy' and 'business' have the significant decrease of word frequencies. The other way, the frequencies of terms such as 'policy', 'study', 'research', 'datum', 'survey', 'model', 'public', and 'broadband' are slightly increased after 2003 and 'online', 'social', and 'mobile' terms are only shown in the wordcloud of research abstracts after 2003. These results present that the research domains in the telecommunications policy are changed from researches for infrastructures, businesses, and economy to data and survey-driven analysis as well as the public data analysis. Also, researches have been conducted in the respect of social, online, and mobile telecommunications policies or analyses for the new telecommunications technology in recent ten years.



(a) Term Frequencies 1990-2003 (b) Term Frequencies 2004-2013

Fig. 4. Comparison of Term Frequencies in Research Paper Abstracts by Wordcloud

3.2 Keyword Extraction and Co-occurrence Analysis Results

The top eleven most appeared keywords in collected papers are shown on the left of **Table 1**. The term 'Internet' most frequently appeared on both before 2004 and recent decade abstracts, but the importance of other keywords by the relative frequencies are changed between two groups. The importance of 'E-commerce', 'Telecommunications', and 'China' are decreased while the frequencies of 'Policy' and 'Regulation' are increased in recent ten years. Actually, 'Electronic commerce' and 'E-commerce' show fifteen times respectively in the 1990-2003 group, but we summed up the frequencies since two words have same meaning. The terms

related technologies and services in telecommunications are disappeared, whereas more policy related terms are presented in keywords of recent decade papers.

The right of **Table 1** shows keywords by KEA. The probabilities of keywords for 'policy' and 'technology' are decreased, while the probabilities of 'public', 'data' and 'analysis' are increased in abstracts of recent decade articles. The analysis results for author defined keywords in articles which are published after 2003 may imply that the authors more focus on the policy researches for human welfares than the researches for economics, the development of technologies, and services in the field of telecommunications policy. The KEA results also show that recent researches more emphasize data analyses, and they validate the analysis results of the term frequency.

Table 1. Most frequent author defined keywords in papers and KEA results

Paper Keyword Frequencies				**Keyphrase Extraction Analysis Results**			
1990-2003		**2004-2013**		**1990-2003**		**2004-2013**	
Keyword	**Freq.**	**Keyword**	**Freq.**	**Keyword**	**Prob.**	**Keyword**	**Prob.**
Internet	148	Internet	439	Policy	0.732	Public	0.483
E-commerce	30	Policy	104	Telecommunication	0.732	Analysis	0.483
Policy	17	Regulation	96	Business	0.483	Communication	0.483
Telecommunications	16	Broadband	84	Technology	0.483	Policy	0.390
Regulation	15	Digital Divide	75	Public	0.483	Service	0.390
Information Technology	13	Privacy	61	Analysis	0.483	Data	0.390
China	11	Telecommunications	58	Service	0.390	Technology	0.321
Quality of Service	11	China	54	Data	0.321	Time	0.319
World Wide Web	10	E-commerce	51	Industry	0.319	Management	0.319
Innovation	9	Public Policy	36	Infrastructure	0.319	Quality	0.319
Globalization	9	E-government	35	National	0.319	Applications	0.319

The co-occurrence analysis results show the important word pairs in collected documents. **Table 2** presents the top ten important word pairs with log-likelihood ratios. The importance of word co-occurrence is evaluated by log-likelihood ratios in our research since it is better method for the estimation of word co-occurrence than Pearson's Chi-square and PMI [4]. The 'web site' is the top ranked word pair in 1990-2003, whereas the 'policy maker' is the most important word pair in 2004-2013. Both top ten log-likelihood ratio ranking results present that Internet technology and policy making are associated with popular research trends in telecommunications policy. However, the researches related to policies of e- businesses and service quality in telecommunications are not highly ranked in recent ten years, whereas the

policies in social media and the differences in socioeconomic levels become more interesting research topics. Also, more practical researches are conducted than theoretical researches in a recent decade. Although the infringements of copyright, the security, and the privacy have significantly become important issues in ICT, the researches on the security and the privacy have not shown in telecommunications policy. Even the high ranked word pair for the copyright disappeared in recent decade researches.

Fig. 5 shows the co-occurrence graphs based on log-likelihood ratios and keywords which are extracted by KEA. The term-term adjacency matrix has to be created for visualizing the co-occurrences of terms in R software. The word frequency information in the term-document matrix file is converted to the term-term adjacency matrix by the 'tm' package in R software [29][30]. Here, we excluded low frequency terms in the term-document matrix because the large sparse matrix causes optimization problems at converting matrix step in R. The width of edges which means the log-likelihood ratio scores in graphs are the strength of association between two terms.

Table 2. Comparison of high ranked word pairs by LLR in two groups

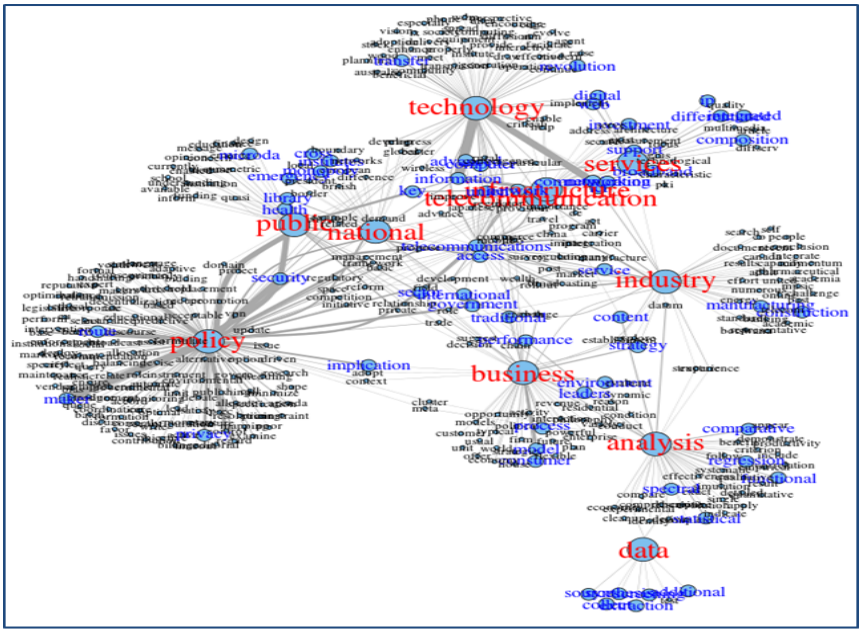
1990-2003			2004-2013		
Word Pair	**Co-Occurrence**	**LLR**	**Word Pair**	**Co-Occurrence**	**LLR**
Web site	90	753.98	Policy maker	267	2007.25
Policy maker	53	420.13	Digital divide	177	1956.32
Real time	43	414.12	Web site	192	1530.99
Electronic commerce	44	388.82	Original value	109	1335.19
World wide	41	381.5	Search engine	110	1235.89
Supply chain	27	346.44	Communication technology	185	1041.52
Service provider	41	264.97	Social media	160	958.75
Intellectual property	18	255.57	Real time	114	944.91
Communication technology	45	240.22	Service provider	152	899.01
Quality Service	45	225.94	Practical implication	91	888.92

The graphs not only visualize the results of co-occurrence matrix but also show additional relationships. First, the term 'technology' commonly has high correlation with the term 'communications' in both groups. The term 'public' is highly correlated with the term 'maker' via 'security' and 'policy' in the graph for the relationship before 2004. It implies that researches on policy making in public security may have been conducted in the field of telecommunications policy, and this presents the hidden relationship of terms which does not appear in the results for top ten word pairs. The connection between 'public', 'policy', 'service', and 'quality' is presented in the co-occurrence graph for researches in recent ten years. This shows that researches which are related to public services and their quality have importantly conducted in the telecommunications policy, and they did not appear in the results

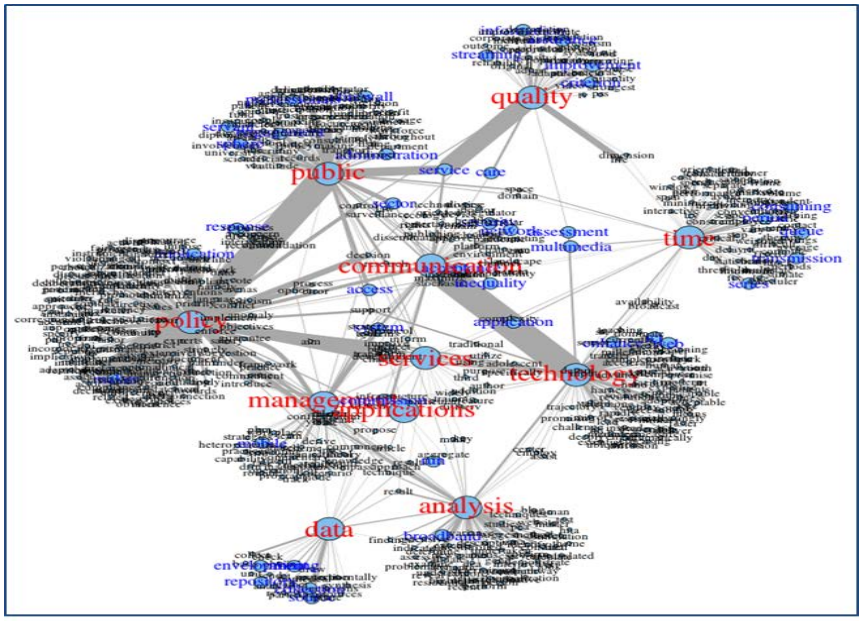
for top ten word pairs after 2003 as shown in **Table 2**.

3.3 Selection of Important Articles

Total 3162 of 4703 publications are cited and the number of citations is 31597 in collected articles. The proportion of not cited papers is only 33% and it seems to indicate that a lot of publications have the citations in the field of telecommunications policy.



(a) Co-occurrences by LLR 1990-2003



(b) Co-occurrences by LLR 2004-2013

Fig. 5. Word co-occurrence graphs by LLR

The upper graph of **Fig. 6** shows a trend in the number of citation and the number of cited papers. The number of cited papers is dramatically decreased when the number of citations is from 2 to 4. It indicates that a lot of papers have only one citations and it is very difficult to receive more than four citations in telecommunications policy. The proportion graph shows that almost half of total publications have low citations (one to four citations). Specifically, almost 70% of papers which received less than 4 citations (33% of total cited papers) got very small citations. Thus, above two results show that the papers are generally received citations, but most papers did not get high citations in telecommunications policy. Also, almost 90% of total papers which are published before 2004 had been cited and 60% of papers have been cited since 2004. This phenomenon is a usual trend in citations of publications because the old papers relatively received many citations over the time. However, papers which are published before 2004 received 9861 (31% of total citations) citations, while recent ten years' papers received 21736 (69%) citations. This shows that the proportion of cited papers is 35% decreased but the proportion of the number of citations are increased in 120% after 2003. This implies that the productivity of papers in terms of citations was increased in recent ten years comparing to the researches before 2004.

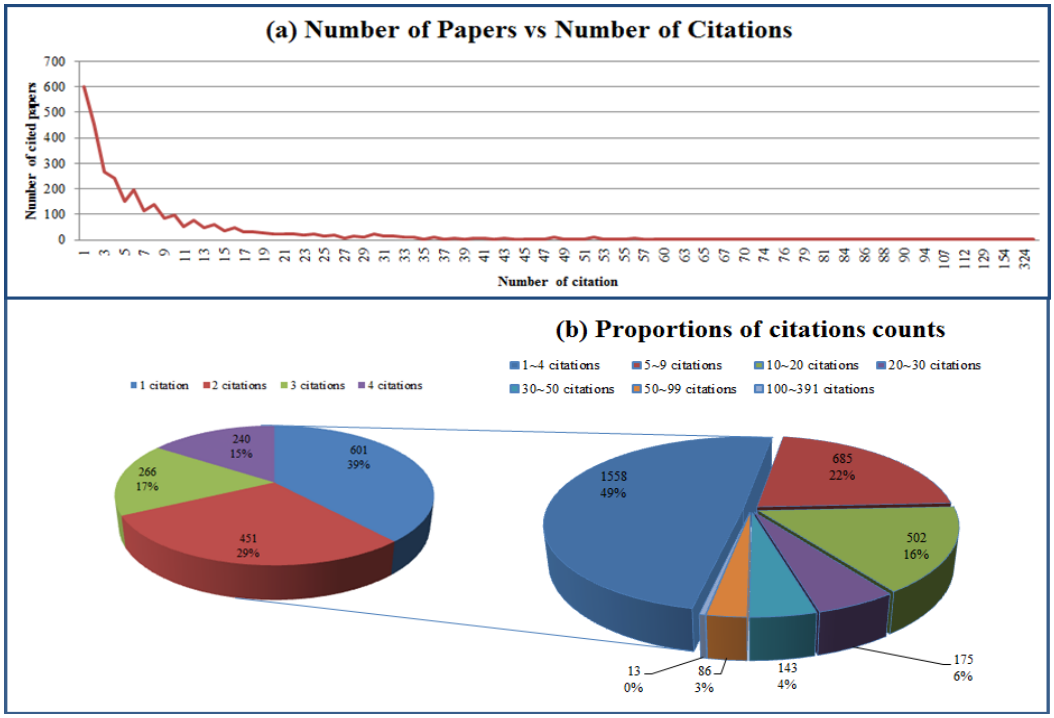


Fig. 6. Overall citation proportion

Table 3 shows top ranked papers in terms of the number of citations and PageRank results. Eight articles are shown in both analysis results and the most important article is about the role of Internet in the healthcare and the related policies [31]. Next frequently cited articles are related to economics and policies in the broadband distribution [32][33]. This research topic is shown in the result of PageRank analysis, but the importance is relatively decreased. On the other hand, research articles for the role of new technology such as cloud computing or grid computing in ICT has higher or same ranking in the PageRank result in comparison with the

results for citation orders [34][35]. Also, the research article for the digital inequality and the digital gap has fifth ranking in both the citation analysis and the importance analysis [36][37]. Next highly ranked articles are associated with the effect of ICT or Internet on the productivity in economy [38][39][40]. Also, the research on the application of Delone's IS success model on e-commerce is eighthly ranked in both methodologies [41]. Two references appear as an important article among publications before 2004 in top five ranking, while only one article is shown as a highly cited paper and an important article from sixth to tenth ranking.

Table 3. Top cited papers and top ranked papers by PageRank

Citation	Pub. Year	Paper Title	Journal Title	Authors
391	2003	Use of the Internet and e-mail for health care information - Results from a national survey	Jama-Jounal of The American Medical Association	Baker, L et al.
324	2009	Broadband investment and regulation: A literature review	Telecommunications Policy	Cambini, C, Jiang, Y
166	2000	A game theoretic framework for bandwidth allocation and pricing in broadband networks	IEEE-ACM Transactions on Networking	H. Yaiche et al
154	2009	Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility	Future Generation Computer Systems-the International Journal of Grid Computing and eScience	R. Buyya et al.
144	2008	Understanding digital inequality: Comparing continued use behavioral models of the socio-economically advantaged and disadvantaged	MIS Quarterly	J.J.P. Hsieh, A. Rai, M. Keil
114	2002	GridSim: a toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing	Concurrency and Computation-Practice & Experience	R. Buyya, M. Murshed
112	2002	Does the Internet make markets more competitive? Evidence from the life insurance industry	Journal of Political Economy	J.R. Brown, A. Goolsbee
112	2003	Information technology and economic performance: A critical review of the empirical evidence	ACM Computing Surveys	J. Dedrick, V. Gurbaxani, K.L. Kraemer
112	2004	Measuring e-commerce success: Applying the DeLone & McLean information systems success model	International Journal of Electronic Commerce	W.H. DeLone, E.R. McLean
112	2003	Consumer surplus in the digital economy: Estimating the value of increased product variety at Online booksellers	Management Science	E. Brynjolfsson, Y. Hu, M.D. Smith
PageRank Result				
Rank	**Pub. Year**	**Paper Title**	**Journal Title**	**Authors**
1	2003	Use of the Internet and e-mail for health care information - Results from a national survey	Jama-Jounal of The American Medical Association	L. Baker et al.
2	2009	Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility	Future Generation Computer Systems-the International Journal of Grid Computing and eScience	R. Buyya et al.
3	2007	Gradations in digital inclusion: children, young people and the digital divide	New Media & Society	S. Livingstone et al.

4	2000	A game theoretic framework for bandwidth allocation and pricing in broadband networks	IEEE-ACM Transactions on Networking	H. Yaiche et al
5	2007	Closing the rural broadband gap: Promoting adoption of the Internet in rural America	Telecommunications Policy	R. LaRose et al.
6	2002	GridSim: a toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing	Concurrency and Computation-Practice & Experience	R. Buyya, M. Murshed
7	2004	Measuring e-commerce success: Applying the DeLone & McLean information systems success model	International Journal of Electronic Commerce	W.H. DeLone, E.R. McLean
8	2003	Information technology and economic performance: A critical review of the empirical evidence	ACM Computing Surveys	J. Dedrick, V. Gurbaxani, K.L. Kraemer
9	2003	Consumer surplus in the digital economy: Estimating the value of increased product variety at Online booksellers	Management Science	E. Brynjolfsson, Y. Hu, M.D. Smith
10	2002	Does the Internet make markets more competitive? Evidence from the life insurance industry	Journal of Political Economy	J.R. Brown, A. Goolsbee

4. Conclusions

The bibliometric analysis is widely used for understanding research domains, research trends, and knowledge structures in a particular field. The analysis has been used in library and information science, but it is currently used in various science and engineering fields such as bioinformatics. We identified academic domains and trends in telecommunications policy by the bibliometric analysis. Instead of traditional bibliometric analysis, text mining techniques are used with the data in academic papers from Thomson Reuters' WoS database. The Java language based software is developed for gathering the publication data via web services and for performing text mining techniques. The R software is used for the visualization of analysis results in this paper as well. We used four text mining techniques such as the term frequency analysis, the KEA, and the co-occurrence analysis as the contents analysis and the PageRank and the citation counts as the citation analysis methods. The abstracts, authors, citations information are collected. They are classified into the dataset in 1990-2003 and the one in 2004-2013.

The analysis results for the term-frequency and the document-frequency implies that the scope of research topics become more narrow in recent ten years than the one before 2004. The researches on policies for the telecommunication infrastructures, public services, and QoS have been conducted in the field of telecommunications policy. The researches on technology issues such as security and copyright, and the business have appeared before 2004, but policies related to social communication services and the distribution of telecommunications infrastructures are shown in recent ten years' publications. The authors of publications have conducted more practical and data-driven analysis researches in a recent decade. The rational establishment of infrastructures could be a significant issue at the settlement stage of telecommunications network. Because a lot of hardware and software which are related to telecommunications are developed in this era, most enterprises and governments emphasized on securing their products. Thus, they could make efforts to establish laws and policies to protect their products and data. The academic outputs can be focused on this domain as well. On the other hand, due to high demands in telecommunications services after the infrastructures are stably settled, providers are more interested in QoS and governments also

want to provide technologies for public services. Therefore, there could be efforts to find political solutions which are difficult to handle in technical methods recently. Also, more political efforts could be made to improve welfares by public telecommunications services.

The decrease of researches on the political approaches for technical problems appears as an issue in the field of telecommunications policies. The interests in technical problems are not supposed to be decreased because more technologies are developed, more technical problems come up. Therefore, enterprises and academia have to more cooperate with government to find political solutions for the technical issues in telecommunications. In addition, the analysis results show that it is difficult to find the publications about policies of new technologies although new services, such as cloud services and Internet of Things, emerged and commercial services are already provided in telecommunications. For example, although mobile technologies and their services are big issues in telecommunications field, relatively a few researches on policies for mobile telecommunications appears in the publications. Thus, new researches on the balanced distribution of mobile infrastructures, mobile telecommunications price, and the revitalization of mobile telecommunications industry have to be conducted. Also, as mentioned above, the security and the privacy issues need to be reconsidered as important research topics in telecommunications policy since this is the long-term significant issues in ICT.

The citation analysis results indicate that the publications are generally received citations, but most of them did not get high citations in telecommunications policy. The results also show that the proportion of cited papers is decreased but the proportion of the number of citations is increased in recent ten years. This implies that although recent publications did not get high citations, the productivity of papers in terms of citations was increased in recent ten years comparing to the researches before 2004. The analysis results also show that the important articles before 2004 are about the applications of Internet to healthcare, while topics in important references are the distribution methods of infrastructures, and the inequity and the gap which arise in the distributions. Most important articles in recent decades are based on data analysis as well. This indicates that most researchers have considered on the data-centric analysis-based researches rather than review researches and this validates the results of contents analysis.

This paper has contributions in regard to the first bibliometric analysis with abstract and citation data in telecommunications as well as the development of software which has functions of web services and text mining techniques. In addition to existing studies for research summary by statistical techniques such as meta analysis, our research provides more comprehensive way to understand research trends and future direction of researches in tge telecommunication policy field. Also, the developed software will be used for collecting publication data from Thomson Reuter database and analyze the knowledge structure of publications in other ICT fields. However, the fundamental text mining techniques are used in this paper. Thus, further research will be conducted by more text mining techniques such as topic modeling and co-citation analysis. Also, the GN-algorithm is able to be considered for clustering terms in co-occurrence matrix and for clustering authors in co-citation network. Another limitation of our research is not using whole terms in abstracts because of computing resource problems. For example, we excluded terms which have low frequencies (frequencies less than 2 before 2005 and frequencies less than 5 after 2004) in order to create the term-term matrix. Since Hadoop has received attention as a solution to analyze large unstructured data, we will apply the matrix conversion algorithm by Hadoop for the analysis with whole dataset. Also, not only more text mining algorithms by Hadoop but also RHadoop which are libraries for using MapReduce computation scheme in R software is able to be considered for an

appropriate method to solve the problems.

References

[1] B. Ziegler, *Methods for Bibliometric Analysis of Research: Renewable Energy Case Study*, Working paper for Master Thesis, Massachusetts Institute of Technology, 2009. http://web.mit.edu/smadnick/www/wp/2009-10.pdf.

[2] F. Osareh, "Bibliometrics, Citation Analysis and Co-Citation Analysis: A Review of Literature I," *Libri*, vol.46, no.3, pp.149-158, October 1996. Article (CrossRef Link).

[3] N.D. Bellis, *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*, Scarecrow Press, 2009.

[4] M. Song and S.Y. Kim, "Detecting the Knowledge Structure of Bioinformatics by Mining Full-text Collections," *Scientometrics*, vol.96, no.1, pp.183-201, November 2013. Article (CrossRef Link).

[5] C. Lam, *Hadoop in Action*, Manning Publications, 2010.

[6] J.E. Andrews, "An Author Co-citation Analysis of Medical Informatics," *Journal of Medical Library Association*, vol.91, no.1, pp.47-56, January 2003. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC141187/.

[7] S.K. Patra and S. Mishra, "Bibliometric Study of Bioinformatics Literature," *Scientometrics*, vol.67, no.3, pp.477-489, August 2006. Article (CrossRef Link).

[8] J.Y. Bansard, D. Rebholz-Schuhman, G. Cameron, D. Clar, E.V. Mulligen, F. Beltrame, E.D.H. Barbolla, F. Martin-Sanchez, L. Millanesi, I. Tollis, J.V. Lei and J.L. Coatrieux, "Medical Informatics and Bioinformatics: A Bibliometric Study," *IEEE Transactions on Information Technology in Biomedicine*, vol.11, no.3, pp.237-243, May 2007. Article (CrossRef Link).

[9] X. Carbonell, E. Guardiola, M. Beranuy and A.Belles, "A Bibliometric Analysis of the Scientific Literature on Internet, Video games, Cell phone Addiction," *Journal of the Medical Library Association*, vol.97, no.2, pp.102-107, April 2009. Article (CrossRef Link).

[10] M. Iwashita, S. Shimogawa and K. Nishimatsu, "Text Mining for Customer Enquiries in Telecommunication Services," *Knowledge-Based and Intelligent Information and Engineering Systems Lecture Notes in Computer Science*, vol.5712, pp.228-235, October 2009. Article (CrossRef Link).

[11] C.J. Tsui, P. Wang, K.R. Fleischmann, A.B. Sayeed and A. Weinberg, "Building an IT Taxonomy with Co-Occurrence Analysis, Hierarchical Clustering, and Multidimensional Scaling," in *Proc. of the iConference*, pp.247-256, February 3-6, 2010. https://www.ideals.illinois.edu/bitstream/handle/2142/14918/building%20IT%20taxonomy%20iConference2010_Final.pdf?sequence=2.

[12] G.Y. Song, Y.J. Cheon, K.H. Lee, K.M. Park and H.C. Rimg, "Inter-category Map: Building Cognition Network of General Customers through Big Data Minig," *KSII Transactions on Internet and Information Systems*, vol.8, no.2, pp.583-600, February 2014. Article (CrossRef Link).

[13] S.S. Hong, J.H. Kong and M.M. Han "The Adaptive SPAM Mail Detection System using Clustering based on Text Mining," *KSII Transactions on Internet and Information Systems*, vol.8, no.6, pp.2186-2196, June 2014. Article (CrossRef Link).

[14] I.H. Witten, G.W. Paynter, E. Frank, C. Gutwin and C.G. Nevill-Manning, "KEA: Practical Automatic Keyphrase Extraction," in *Proc. of the 4th ACM conference on Digital libraries*, pp.254-255, August 11-14, 1999. Article (CrossRef Link).

[15] Thomson Reuters, *Web of Knowledge Web Services Lite v. 3.0*, Thomson Reuters, 2012.

[16] J.C. Anderson, J. Lehnardt and N. Slater, *CouchDB: The Definitive Guide*, O'Reilly Media, 2010.

[17] B. Holt, *Writing and Querying MapReduce Views in CouchDB*, O'Reilly Media, 2011.

[18] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2006.

[19] X. Hu and B. Wu, "Automatic Keyword Extraction using Linguistic Feature," in *Proc. of 6th IEEE International Conference on Data Mining*, pp.19-23, December 18-22, 2006.

Article (CrossRef Link).
[20] K.W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, vol.16, no.1, pp.22-29, March 1990. http://dl.acm.org/citation.cfm?id=89095.
[21] G. Bounma, "Normalized (Pointwise) Mutual Information in Collocation Extraction," in *Proc. of the Biennial GSCL Conference*, pp.31-40, September 30-October 2, 2009. https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf.
[22] C.D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, The MIT press, 2002.
[23] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol.19, no.1, pp.61-74, March 1993. http://dl.acm.org/citation.cfm?id=972454.
[24] L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank Citation Ranking: Brining Order to the Web*, Technical Report, Standord InfoLab, 1999.
[25] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol.30, no.1-7, pp.107-117, April 1998. Article (CrossRef Link).
[26] A.N. Langville and C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Ranking*, Princeton University Press, 2012.
[27] Y. Ding, E. Yan, A. Frazho and J. Caverlee, "PageRank for Ranking Authors in Co-Citation Networks," *Journal of the American Society for Information Science and Technology*, vol.60, no.11, pp.2229-2243, July 2009. Article (CrossRef Link).
[28] Y. Zhao, *R and Data Mining: Examples and Case Studies*, Academic Press, 2012.
[29] I. Feinerer, K. Hornik and D. Meyer, "Text Mining Infrastructure in R," *Journal of Statistical Software*, vol.25, no.5, pp.1-54, March 2008. http://www.jstatsoft.org/v25/i05/paper.
[30] I. Feinerer and K. Hornik, *Text Mining Package*, R reference Manual, R-project.org, 2014. http://cran.r-project.org/web/packages/tm/tm.pdf.
[31] L. Baker, T.H. Wagner, S. Singer and M.K. Bundorf, "Use of the Internet and E-mail for Health Care Information-Results from a National Survey," *The Journal of the American Medical Association*, vol.289, no.18, pp.2400-2406, May 2003. http://www.jmir.org/2007/3/e20/.
[32] C. Cambini and Y. Jiang, "Broadband Investment and Regulation: A Literature Review," *Telecommunications Policy*, vol.33, no.10-11, pp.559-574, September 2009. Article (CrossRef Link).
[33] H. Yaiche, R.R. Mazumdar and C. Rosenberg, "A Game Theoretic Framework for Bandwidth Allocation and Pricing in Broadband Networks," *IEEE/ACM Transactions on Networking*, vol.8, no.5, pp.667-678, October 2000. Article (CrossRef Link).
[34] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg and I. Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," *Future Generation Computer Systems*, vol.25, no.6, pp.599-616, June 2009. Article (CrossRef Link).
[35] R. Buyya and M. Murshed, "GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing," *Concurrency and Computation-Practice and Experience*, vol.14, no.13-15, pp.1175-1220, November-December 2002. Article (CrossRef Link).
[36] J.J.P. Hsieh, A. Rai and M. Keil, "Understanding Digital Inequality: Comparing Continued Use Behavioral Models of the Socio-Economically Advantaged and Disadvantaged," *MIS Querterly*, vol.32, no.1, pp.97-126, March 2008. http://dl.acm.org/citation.cfm?id=2017384.
[37] S. Livingstone and E. Helsper, "Gradations in Digital Inclusion: Children, Young People and the Digital Divide," *New Media & Society*, vol.9, no.4, pp.671-696, August 2007. Article (CrossRef Link).
[38] J. Dedrick, V. Gurbaxani and K.L. Kraemer, "Information Technology and Economic Performance: A Critical Review of the Empirical Evidence," *ACM Computing Systems*, vol.35, no.1, pp.1-28, March 2003. Article (CrossRef Link).
[39] E. Brynjolfsson, M.D. Smith and Y.J. Hu, "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science*, vol.49, pp.1580~1596, November 2003. Article (CrossRef Link).

[40] J.R. Brown and A. Goolsbee, "Does the Internet Make Markets More Competitive? Evidence from the Life Insurance Industry," *Journal of Political Economy*, vol.110, pp.481-507, June 2002. Article (CrossRef Link).

[41] W.H. DeLone and E.R. McLean, "Measuring e-Commerce Success: Applying the DeLone & McLean Information Systems Success Model," *International Journal of Electronic Commerce*, vol.9, no.1, pp.31-47, Fall 2004. http://dl.acm.org/citation.cfm?id=1278171.

Dr. Junseok Oh is a research professor at Communications Policy Research Center in Yonsei University. He received B.E degree from Information Engineering at Hansung University in 2002 and M.S degree from Computer Science at Chungbuk National University in 2004. He also received MSCE and PhD from the Pennsylvania State University in 2006 and 2010. His research interests are data mining, data science, Internet of Things, and the econometrics analysis.

Dr. Bong Gyou Lee who is a professor at Graduate School of Information has served as a director of Communications Policy Research Center(CPRC) in Yonsei University since 2009. Dr. Lee received a B.A. from the Department of Economics at Yonsei University and M.S, Ph.D. from Cornell University. During 2007 and 2008 he served as Commissioner of the Korea Communications Commission