International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:12, 2015

A Methodology for Investigating Public Opinion Using Multilevel Text Analysis

William Xiu Shun Wong, Myungsu Lim, Yoonjin Hyun, Chen Liu, Seongi Choi, Dasom Kim, Kee-Young Kwahk, Namgyu Kim

***Abstract*—**Recently, many users have begun to frequently share their opinions on diverse issues using various social media. Therefore, numerous governments have attempted to establish or improve national policies according to the public opinions captured from various social media. In this paper, we indicate several limitations of the traditional approaches to analyze public opinion on science and technology and provide an alternative methodology to overcome these limitations. First, we distinguish between the science and technology analysis phase and the social issue analysis phase to reflect the fact that public opinion can be formed only when a certain science and technology is applied to a specific social issue. Next, we successively apply a start list and a stop list to acquire clarified and interesting results. Finally, to identify the most appropriate documents that fit with a given subject, we develop a new logical filter concept that consists of not only mere keywords but also a logical relationship among the keywords. This study then analyzes the possibilities for the practical use of the proposed methodology thorough its application to discover core issues and public opinions from 1,700,886 documents comprising SNS, blogs, news, and discussions.

***Keywords*—**Big data, social network analysis, text mining, topic modeling.

I. Introduction

RECENTLY, the rapid increase in the dissemination and utilization of social media has caused growth in the communication space that allows many users to discuss and post their personal opinions. In other words, we are able to form public opinions by gathering similar opinions on social issues from different individuals through various social media. Thus, the formed public opinions can be passed to the government to complement policies and to develop new policies. Paradoxically, various public opinions have already been exposed to a variety of social media. However, more citizens cite dissatisfaction with the government because such public opinions do not fully converge with the government's policies.

Thus, various recent attempts have been made to establish or improve the policies to reflect public opinions on science and technology at the national level. However, overlooking a feature that is formed when each scientific technique is applied to a particular social issue is better than making an earnest attempt to form a public opinion about science and technology itself. We use Fig. 1 to further explain this concept. In Fig. 1, a three-dimensional diagram indicates the three different axes of issue, subject, and period. When a subject is applied to an issue, it can form a new opinion. Therefore, because the main focus of this study is on science and technology (S&T), all subjects in this study are related to S&T. Therefore, applying an S&T-related subject to a social issue may form a public opinion. For example, in Fig. 1, when the S&T keyword "intelligent unmanned vehicle"—typically known as a "drone"—is applied to the issues "surveillance patrol" and "unmanned delivery," the keywords of the formed public opinions are displayed differently. Public opinions corresponding to the issue of "surveillance patrol" are displayed as "security," "human rights," and "privacy." For the issue "unmanned delivery," the public opinions "efficiency," "damage," and "remote place" are displayed.

To support the establishment of a national policy, previous studies have used text analysis but have failed to carefully handle the keyword selection process when identifying a document. The process of selecting the keyword to identify the target document has a significant impact on the quality of the overall results of the analysis. However, in previous studies, the keyword selection process appeared less important compared with other analytical procedures, such as topic modeling, issue tracking, and visualization.

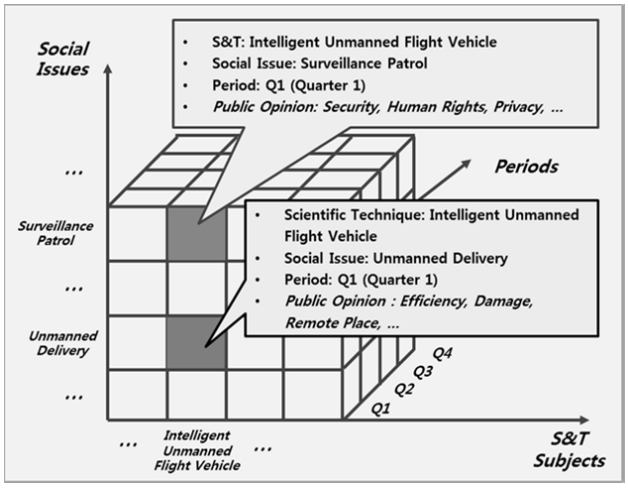

Fig. 1 Different Social Issues and Public Opinions for the Same Science and Technology

In general, most studies addressing text analysis have experienced a dilemma in setting the refinement level of vocabulary. In the analysis using a word dictionary—also referred to as a start list—if the dictionary does not include a

William Xiu Shun Wong, Myungsu Lim, Yoonjin Hyun, Chen Liu, Seongi Choi, Dasom Kim, Kee-Young Kwahk, and Namgyu Kim are with the Graduate School of Business IT, Kookmin university, Seoul, 136-702 Republic of Korea (e-mail: {williamwong , amr2001, yoonjin0630, liuchen, csy0000, dskim1225, kykwahk, ngkim} @kookmin.ac.kr, respectively).

sufficient number of terms, a common analysis result is displayed. In the analysis using a stop word dictionary—also referred to as a stop list—if the dictionary does not include a large number of stop words, the inappropriate words are displayed in the analysis result.

In this study, a methodology is then proposed to overcome the three limitations previously mentioned in the public opinion analysis process on S&T. Specifically, the two-stage analysis methodology is proposed to reflect the phenomenon of public opinion formed on the basis of combinations of each S&T with the specific social issues. In first stage, the word dictionary and the stop word dictionary are proposed for use accordingly to solve the dilemma of selecting the refinement level for the vocabulary. In second stage, instead of using a simple keyword list, a method is newly presented to utilize the logical filter to identify the document that matches the S&T subject, for which the logical filter is composed of a logical combination of S&T keywords.

The remainder of this paper is organized as follows. The next section introduces the related work on text mining, which is the main core of our proposed methodology. Section III describes the proposed methodology in detail, together with the experimental results. Finally, Section IV presents the contributions and limitations of this study.

II. Related Work

Many types of information in the real world are usually expressed and communicated in text form [1]. A series of processes to extract valuable information from voluminous text is called text mining [2]–[5]. Currently, many attempts have been made to solve the complicated problems of various domains using text mining techniques. Identifying the original document for some documents [6], discovering new crimes by analyzing patterns in previous crimes, and structuring unstructured storage using text categorization are recent examples of text mining applications.

Text mining utilizes various techniques, such as natural language processing, information retrieval, issue tracking, and text categorization areas [3], [7], [8], as well as association, classification, and clustering techniques that have been used in traditional data mining applications. Among the techniques, natural language processing is regarded as the core technique used by text mining applications. A large gap exists between text data and traditional data in that text data are presented in the form of unstructured documents, whereas traditional structured data are presented in the form of two-dimensional tables [9]. Therefore, many techniques for structuring text data into matrices, hierarchies, and vectors have been proposed in the literature [10]. The most fundamental and widely used technique is the vector space model [11], [12], through which frequencies of terms in each document are summarized.

Among the various contemporary text-related applications, topic modeling draws the most attention from researchers and practitioners. The vector space model and a TF-IDF (term frequency–inverse document frequency) measure [13] form two main theoretical foundations of topic modeling. The main process of topic analysis is usually performed directly after parsing and filtering. In the parsing stage, sentences in the documents are separated into tokens. In the filtering stage, some tokens are eliminated according to predefined restrictions. Topic analysis is similar to traditional clustering techniques in that the goal is to group similar objects and separate the dissimilar ones. However, they are also significantly different in that topic analysis can map each document to multiple topics, whereas each element can belong to only one specific cluster in traditional clustering algorithms.

III. A Methodology for Investigating Public Opinion Using Multilevel Text Analysis

A. Proposed Model and Research Scope

In this section, the overview of the public opinion analysis methodology on S&T issues is presented through Fig. 2.

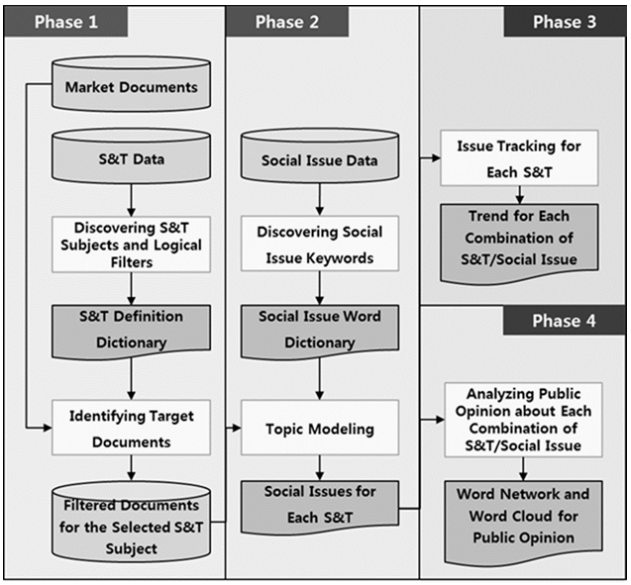

Fig. 2 Research Overview

In Fig. 2, Phase 1 indicates the process of extracting only the documents related to a particular S&T from a set of documents containing public opinions, such as news, blogs, twitters, and discussions. First, we select the S&T subjects from a variety of S&T documents as targets of the analysis. We then derive a logical filter that can be defined for each S&T subject. A logical filter is a new concept proposed in this study that illustrates a logical combination of S&T keywords for each subject. By applying the logical filter that includes the S&T definition dictionary, we are able to extract the documents related to a particular S&T from the public opinion data. Next, Phase 2 indicates the process of extracting the major social issues related to a particular S&T. In Phase 2, we first derive the major keywords from various materials related to social issues and save them as a social issue word dictionary. In the next step, this dictionary is used as a start list. By using the particular S&T documents derived in Phase 1 as target data, we derive the major social issues for each S&T through the topic modeling stage. In Phase 3, we analyze the periodic flow of major social

issues for each S&T through issue tracking. Finally, we perform a public opinion analysis for each social issue related to a particular S&T. The public opinion analysis is carried out in detail by using the word cloud and word network, enabling us to observe the word patterns that appear frequently and simultaneously. To derive a better result from the analysis that includes a wide range of public opinions, a stop word dictionary is applied in this process. Therefore, we are able to determine the public opinion from combinations of each S&T with the specific social issues by using the results derived in the previously described processes, such as social issues for each S&T, issue tracking, word cloud, and a word network of the derived social issues.

B. Data Description

In this section, we introduce the experimental data applied in the proposed methodology. In the case of S&T documents used to define a logical filter for each S&T, we use the "2014 Technology Level Evaluation" report provided by the Korea Institute of S&T Evaluation and Planning (KISTEP). Next, to refine the issue keywords derived from topic modeling, we use social issue documents to construct the word dictionary. In this study, we use the report "General Practice Plan on Solving the Social Issue based on S&T" provided by the National Science & Technology Council as the social issue document. Finally, we collected a vast amount of recent articles from various social media to use as public opinion data. The recent year—June 11, 2014 June 10, 2015—was set as the analysis period. A total of 1,700,886 text documents were collected from Twitter, Naver Blog, Daum Agora, KBS news, and Yonghap news. The number of documents collected from each social media platform is displayed in Fig. 3.

	Twitter	NAVER Blog	DAUM Agora	KBS News	YONHAP News
Volume	931,000	365,000	77,486	128,055	199,345
Date	2014.06.11 ~ 2015.06.10	2014.06.11 ~ 2015.06.10	2014.07.06 ~ 2015.06.10	2014.06.11 ~ 2015.06.10	2014.06.11 ~ 2015.06.10

Fig. 3 Experimental Data Description

C. Document Identification by Utilizing the Logical Filter

In this section, we further explain the concept of a logical filter used in Phase 1 to identify only the documents related to a particular S&T among the vast amount of documents. Most existing studies identify public opinions through text analysis and typically perform the search using specific keywords. The documents that match the specific keywords are return as the search result. Finally, all of the documents are used as target documents. However, this approach has two limitations. If this approach presents a large number of keywords to be included in a document, then some potential documents may be missing from the target document. In contrast, if the approach presents a low number of keywords, then some unrelated documents might be included in the analysis. Therefore, in this study, we proposed the new concept of a logical filter, which contains a logical combination of keywords rather than a simple list of keywords to more precisely identify the target document. For example, consider the case of identifying the documents related to "infectious disease control." This case requires at least two concepts to describe the technology—"infectious" and "control." In this case, "infectious disease control" can be defined by a logical filter, such as in (1). In the expression, the "*" symbol is the logical operation AND, whereas the "+" symbol represents the logical operation OR.

$$\text{infectious disease control} \leftarrow (\text{prevention} + \text{immune} + \text{antibiotic} + \text{vaccine} + \text{inoculation} + \text{isolation} + \text{control} + \text{epidemiologic investigation}) \quad (1)$$

Using the same method, "intelligent unmanned aerial vehicle," widely known as "drone," can be defined by a logical filter such as (2):

$$\text{intelligent unmanned aerial vehicle} \leftarrow (drone) + (unmanned + intelligent) * (flight\ vehicle + load + photograph) \quad (2)$$

The logical filter for each of the two subjects previously listed is applied to summarize the number of documents extracted from different social media in Fig. 4. A large number of documents are extracted from the news for the two subjects, whereas the number of documents obtained from Twitter appears to be relatively small. This result occurs because the length of a twitter is relatively short compared with the length of a news article, making it difficult to match up with the logical filter that contains a variety of keywords. Furthermore, compared with the news article, a twitter post contains a lot of slang and neologism that makes satisfying the expression of the logical filter difficult. A total of 3,404 documents exist for "infectious disease control" and 450 documents exist only for "intelligent unmanned aerial vehicle" in the same period.

Subject	Twitter	NAVER Blog	DAUM Agora	KBS News	YONHAP News
Total Documents	931,000	365,000	77,486	128,055	199,345
Infectious Disease Control	77	335	217	1786	989
Intelligent Unmanned Flight Vehicle	38	163	16	109	124

Fig. 4 Number of Logically Filtered Documents

Among the data on public opinion, the documents corresponding to a particular S&T can be identified through the previously described process. Thus, the identified document set is used as initial data in the process of deriving the major social issues.

D. Major Social Issues Extraction & Issue Tracking

This section introduces the task that executed in Phase 2 and Phase 3. First, in Phase 2, we extract the frequent terms from a variety of social issues material, and construct the social issue word dictionary after examine by experts. Next, we perform topic modeling on S&T related documents derived in Phase 1,

so we can extract the major social issues related to corresponding S&T. After that, we construct a social issue word dictionary containing all the issue keywords that describe each social issue.

Fig. 5 displayed the result of topic modeling for 3,404 documents related to "infectious disease control" and 450 documents related "intelligent unmanned aerial vehicle." We performed topic modeling using the Topic Analysis module of Text Miner provided by SAS Enterprise Miner 13.1. The number of topic (which also refers as issue) is set as five for each subjects. Fig. 5 displayed the keywords of five derived issues for each subject, corresponding terms number and documents number. As topic modeling is not only applied by many other existing studies, it is also possible to be performed through various commercial applications, so the detail process of topic modeling will be omitted in this section.

Subject	Issue ID	Keywords	Num. Terms	Num. Docs
Infectious Disease Control	Issue1	Patient, Check, Car, Respiratory Apparatus, Seoul	81	588
	Issue2	Foot-and-Mouth Disease, Farm, Stockbreeding, Vehicle, Agriculture and Forestry	36	226
	Issue3	Government, President, USA, Nation, North Korea	100	494
	Issue4	Treatment, Virus, USA, Dyscrasia, Disease	98	498
	Issue5	School, Student, Ministry of Education, Parents of Students, Elementary School	46	214
Intelligent Unmanned Flight Vehicle	Issue6	USA, President, Region, Country, Government	35	55
	Issue7	IT, Internet, Service, Smart, Market	51	59
	Issue8	Japan, China, World, Aviation, Accident	52	62
	Issue9	Game, Vision, Vehicle, Location, Car	53	60
	Issue10	North Korea, Against North Korea, Human Rights, North Korean Defector, North Korean Defect	21	25

Fig. 5 Results of Topic Modeling for Two Subjects

Because the particular S&T-related social issues derived through topic modeling appear in static form, the limitation is that we cannot discover meaningful insights through the social issues that changed dynamically. Therefore, in Phase 3, we analyze the distribution of issues by period through issue tracking of the major social issues derived in Phase 2. We also provide visualization about the trends in the growth, continuous, and shrinking stages of each issue.

To explore the periodic transition of social issues derived in Fig. 5, the issue tracking results are displayed in Figs. 6 and 7. The trend of the issue related to "infectious disease control" is displayed in Fig. 6, whereas Fig. 7 shows the trend of the issue related to "intelligent unmanned aerial vehicle." In addition, the horizontal axis of the two graphs represents the period up to one year from June 11, 2014 to July 10, 2015, which is divided into Period 1 to Period 4, respectively. Further, the vertical axis represents the number of documents corresponding to each issue in that period.

In Fig. 6, we see that the recent attention paid to Issue 1—"infectious disease control"—is increasing rapidly relative to other issues. On the basis of the number of documents in each period, the corresponding issues such as "patient," "check," "tea," "respiratory apparatus," and "Seoul" received little attention from Period 1 to Period 3, but this attention suddenly showed a sharp increase in Period 4. Therefore, this issue was selected for further analysis to discover the cause and public opinion regarding the corresponding issues. The analysis result is presented in the next section.

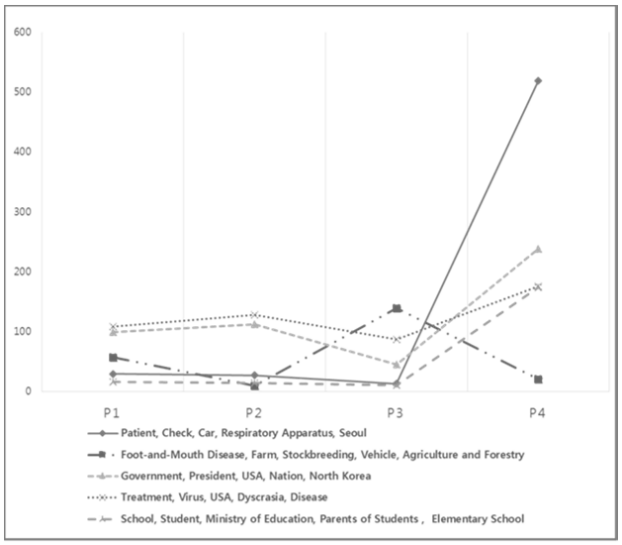

Fig. 6 A Result of Issue Tracking for "Infectious Disease Control"

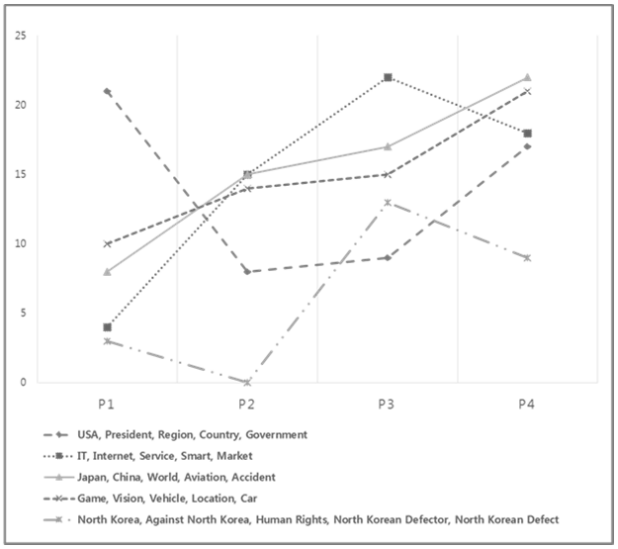

Fig. 7 Example: Issue Tracking for "Intellectual Unmanned Aerial Vehicle"

E. Public Opinion Analysis on S&T-related Social Issues

In this section, we introduce the process of creating word clouds and word networks. Using the results from the topic modeling as in Fig. 2 and the issue tracking as in Fig. 6, we select some interesting social issues for further analysis. Then, keyword frequency and associations related to the selected issues are represented in word cloud and word network forms. In the previous section, we discovered that the "infectious disease control" issue indicates a very extraordinary and interesting trend. In particular, the number of documents on the "patient," "check," "car," "respiratory apparatus," and "Seoul" keywords increased rapidly in the last period. Therefore, we

present the word cloud and word network for 588 documents related to the issue. We used the word cloud package R 3.1.2 for word cloud analysis and the word networks are drawn using NetDraw of UCINET6.5.7.6.

Frequent terms related to the issue appear in the word cloud in Fig. 8. Note that the major keywords consisting of the issue are "MERS," "patient," "infectious disease," "hospital," "quarantine," and "government." Therefore, the outbreak of the MERS infectious disease and the response of the government against this outbreak primarily form the issue. In this analysis, we use a stop list instead of a start list to preserve less refined but vivid terms. In Fig. 8, "this day," "most," and "corresponding" are examples of less refined but vivid terms. Additionally, co-occurrence patterns between frequent terms can be schematized in the word network. But as the words appear in the word network are written in Korean ad the pattern of the network is complicated, so we did not include the word network in this study.

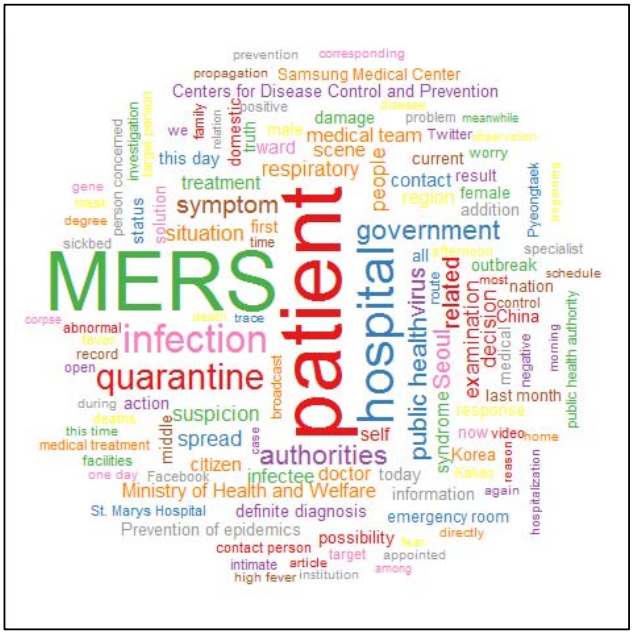

Fig. 8 Result of Word Cloud Analysis for the Selected Issue

In this section, we introduced the process of analyzing public opinions on S&T-related social issues. By applying our methodology to analyze 1,700,886 market documents, we discovered trends in issues related to the S&T subjects of "infectious disease control" and "intellectual unmanned aerial vehicle." Finally, we presented public opinions on some interesting issues in the form of word clouds and word networks.

IV. Conclusion

In this study, we proposed a novel methodology for discovering science- and technology-related social issues from various market data, such as news articles, tweets, blog postings, and online discussions. Additionally, our methodology shows trends in the main social issues and provides various public opinions regarding the combination of a specific S&T subject and a specific social issue related to the subject. By applying our methodology to analyze 1,700,886 real market documents, we detect an interesting trend in S&T-related social issues and indicate the public opinion on the issue in a word network and a word cloud.

We devised the concept of a logical filter to identify only the documents related to a particular S&T subject. Additionally, our methodology attempted to investigate public opinions not on a sole subject or issue but on the combination of a subject and an issue. Finally, we used both a start list and a stop list to resolve the dilemma of selecting the refinement level for the vocabulary. We expect that this approach can satisfy the need to acquire a refined topic lists and to discover vivid public opinions.

Despite the contributions, the current status of our work still reveals some limitations. Primarily, too much time and effort were spent identifying documents using logical filters. This finding implies that the process of designing logical filters and identifying documents corresponding to the filters should be automated in future studies. Additionally, we presented word networks and word clouds as the final outcomes of our methodology. However, the diagrams are too complicated to capture meaningful insights directly from them. Therefore, we need to summarize the information in word networks and word clouds using various social network analysis algorithms.

References

[1] I. H. Witten, *Text Mining, Practical Handbook of Internet Computing*, CRC Press, 2004.
[2] J. Hong, H. Choi, H. Han, J. Kim, E. Yu, S. Lim, and N. Kim, "A Data Analysis-based Hybrid Methodology for Selecting Pending National Issue Keywords," *Entrue Journal of Information Technology*, vol. 13, pp. 97-111, Jun. 2014.
[3] R. J. Mooney, and R. Bunescu, "Mining Knowledge from Text Using Information Extraction," *ACM SIGKDD Explorations*, vol. 7, pp. 3-10, Jun. 2006.
[4] S. Song, J. Yu, and E. Kim, "Offering System for Major Article Using Text Mining and Data Mining," *Proceedings of the 32th annual conference on Korea Information Processing Society*, pp. 733-734, 2009.
[5] E. Yu, J. Kim, C. Lee, and N. Kim, "Using Ontologies for Semantic Text Mining," *The Journal of Information Systems*, vol. 21, pp. 137-161, Sep. 2012.
[6] D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel, "Similarity Measures for Tracking Information Flow," *Proceedings of CIKM, Bremen*, Germany, 2005.
[7] C. J. V. Rijsbergen, *Information Retrieval*, 2nd edition, Butterworth, 1979.
[8] F. Sebastiani, Classification of Text, Automatic, *The Encyclopedia of Language and Linguistics 14*, 2nd edition, Elsevier Science Pub, 2006.
[9] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the Power of Text Mining," *Communications of the ACM*, vol. 49, pp. 76-82, Sep. 2006.
[10] S. M. Weiss, N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining*, Springer, 2010.
[11] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, pp. 613-620, Nov. 1975.
[12] R. Albright, *Taming Text with the SVD*, SAS Institute Inc., 2006.
[13] G. Salton, and M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.