A Study of Design and Implementation of Korean Plagiarism Detection System

WonKyun Joo, KiSeok Choi, YongJu Shin and JaeSoo Kim

Department of R&D System development, KISTI, 245 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{joo, choi, yjshin, jaesoo}@kisti.re.kr

Abstract

Recently, there has been the issue of plagiarism on Korean contents in Korea. Even though a computer program to detect plagiarism on contents is available, there is no overall system to protect plagiarism including Korean contents. Therefore, this research is aimed to design and implement Hangul Plagiarism Detection System (HPDS). From the system perspective, key functions for plagiarism detection are defined, and the structure is designed to allow users to be able to use other plagiarism similarity search engines. In addition, the contents aspect is also significantly considered during the design process. Before the building of contents, about 10,010 units of Korean contents are processed and added to the plagiarism similarity search indexing database. Finally, this research evaluates plagiarism similarity search speed, changes in similarity levels by conversion of original texts, and use of citation information.

Keywords: *Plagiarism Detection, Plagiarism Similarity Search, Hangul Plagiarism Detection System, iThenticate Connection, Reference & Information Service*

1. Introduction

In advanced countries, research misconduct in science emerged as a social problem in the 1980s [1]. For the recent years, research ethics, research misconduct and plagiarism [2] have been in the spotlight in Korea as well. Unlike advanced countries which already own a lot of intellectual property rights, Korea has been passive in preventing plagiarism because of growth-oriented modernization policy so far. As plagiarism committed by high-ranking officials recently emerged as a hot potato in Korea, people's interest in plagiarism has skyrocketed. For example, the number of news articles on plagiarism increased about 8 times from 11 in 2006 to 82 in 2010 [3]. At present, a plagiarism prevention system is under development, but it is still limited to prevent plagiarism from the ethical and systematic aspects. Considering current situations that new knowledge is piling up dramatically, the ground to produce new knowledge and culture could collapse unless intellectual property rights are guaranteed. Even though ethical and systematic culture settled, unintended plagiarism could take place because there is too much knowledge to cover. After all, it appears that the biggest synergy effect may occur in Korea when ethics and systems (programs) are operated in an inter-complementary manner.

This study has investigated how plagiarism has been prevented in advanced countries. In the U.K, a plagiarism prevention system was established by public organizations. The Joint Information System Committee (JISC) launched a pilot project for the introduction and distribution of a plagiarism detection system in 2000 after recognizing its necessities [4]. To maximize the effect of prevention of plagiarism and promote cost efficiency, JISC promoted the project in strategic alliance with iParadigms Europe. As a result, the plagiarism detection

system titled 'Turnitin [5]' was distributed to almost all universities (over 97%) until 2009, and the ground for detection of plagiarism was established in universities. Since 2010, plagiarism prevention-related activities have been actively promoted primarily in high schools [3].

In case of the U.S., the government's role was separated from private/public sector's role. The Office of Research Integrity (ORI) which is under a direct control of the Department of Health and Human Services (DHHS) [6] was launched 1992. Since then, it has established basic policies for prevention of plagiarism and supported related education and research activities. Because customized private and public councils are well organized, and a research ethics-related team is organized in most organizations and universities, plagiarism prevention-related policies have been established. In addition, Turnitin/iThenticate system has been effectively introduced. The councils which have been active in the U.S. include CrossRef [7], Committee on Publication Ethics (COPE) [8] and university council.

In case of China, all processes from system development to distribution were led by the government / public organization. Chinese government and Tsinghua University jointed promoted a national project titled 'China National Knowledge Infrastructure (CNKI)' to develop Chinese digital resources database. The plagiarism detection system named 'CNKI AMLC [9]' which was developed based on extensive contents which have been developed through the CNKI was launched in December 2008 [10].

Unlike the three major countries mentioned above, Japan and Europe were less active in preventing plagiarism. In case of Japan, the government established and promulgated plagiarism prevention-related basic policies only. European countries also focused on coming up with guidelines only to encourage research activities instead of inspection and punishment even though they already had an advanced research ethics system [11].

As stated above, many countries have realized the importance of preventing plagiarism and promoted various plagiarism prevention policies. Thus the fields of study are clearly divided into government-led topics and other plagiarism prevention issues. This study targets to focus on a way of effectively supporting plagiarism prevention using the plagiarism detection system. Referring to iThenticate/Turnitin in the U.K. and the U.S. and AMLC in China, it is expected that the efficiency of plagiarism detection would be enhanced when both contents are programs are provided together. Even though English or Chinese-language plagiarism detection system is established, no Korean-language plagiarism detection system is available.

In this study, how the Hangul Plagiarism Detection System (HPDS) was developed is introduced. HPDS is a practical system which provides Korean contents as well. It has been designed in consideration of extensibility to connect the system with diverse contents. It is likely that HPDS would be useful in detecting and preventing plagiarism in Korean and English-language papers. This paper is structured as follows: In Chapter 2, previous studies on plagiarism similarity detection program and related algorithm are mentioned. In Chapter 3, the design of an extensible plagiarism detection system is explained, focusing on the functional diagram of the system. In Chapter 4, implementation of the system based on the design and development of Korean-language contents are stated. In Chapter 5, the results of the evaluation on plagiarism similarity detection performance are described. In Chapter 6, the conclusion is made.

2. Related Works

2.1. Plagiarism Similarity Detection Program

According to analysis on plagiarism similarity detection programs, a lot of programs have been developed and used especially in advanced countries (ex: The U.K, the U.S., Europe,

China, Korea, etc.). In terms of development of a plagiarism development system, 8 cases (including iThenticate [12]) were found in the U.S. with 2 cases each in the U.K. and Germany, 1 case each in the Netherlands, Japan and China and five cases (ex: MemeCheker [13], etc.) in Korea [3]. The characteristics of the plagiarism similarity detection program are summarized in Table 1. As shown in this figure, the program is led by some systems, and it was not easy to find a system which provides both the program and contents.

Table 1. Analysis of Plagiarism Similarity Detection Program (Summary)

Country	C / P	Providers	Similarity Measure Target	Features	Program Name (Service Name)
USA	Commercial	iParadigms	Report Paper Web	Widely used in universities and organizations * Used by 97% of all universities in the U.K. and the world's renowned universities The greatest database (24 billion web pages, 120 million academic resources)	**iThenticate Turnitin CrossCheck**
		CaNexus	Internet works	Online plagiarism detection service provided since 2000	EVE2
		EduTie.com	Paper	Online plagiarism detection service provided since 2000 Database (about 250,000 papers)	-
		Glatt Plagiarism Services, Inc.	Paper	Small-scale plagiarism detection program	GPSP, GPSD
		Indigo Stream T echnologies	Web	Inspection on plagiarism against online contents	Copyscape
		SkyLine	Web	Search engine(Google/Bing)-based quick online plagiarism detection service	Plagiarism-Detector
	Public	Villanova University	Short Paragraph	Google search engine based online plagiarism detection service Plagiarism detection on short paragraph	SNITCH
		University of Virginia	General Works	Program distributed to small groups since 2002	WCopyfind
				Similarity search-based search engine used	eTBLAST 3.0
UK	Public	University of Leeds	Report	Program used in University of Leeds (small-scale)	CopyChecker
		CEBE Centre	Report	Program distributed to small groups since 2003 Renewed in 2010 using Semantic technology	CopyCatch Gold
Germany	Public	karlsruhe	S/W	Plagiarism detection service on software since 2005	jPlag
	Commercial	Docoloc KG	General Works	Data search service (1 billion data)	**Docoloc**
Netherlands	Commercial	Ephorus	Student Works	4,000 schools and universities in 28 countries using the Ephorus plagiarism prevention system	**Ephorus**
Japan	Commercial	Kanazawa Institute of Technology	General Works	Commercialized through ANK in December 2009 Distributed in an offline program format	Copypelna
China	Commercial	The Chinese government Tsinghua University	Thesis Paper	82 million papers in China / 1billion online academic data Database (science technology, social science/literature, thesis)	**AMLC TMLC SMLC**
Korea	Commercial	Konan Tech	General Works	Online plagiarism detection service (contents NOT included)	**MemeChecker**

Among the said programs, the characteristics of two leading systems – iThenticate and CNKI AMLC – are described. In terms of the number of users and size of database, iThenticate has been dominant. At present, iThenticate has extensive index information for plagiarism similarity detection (24 billion web pages, 90 million publications, 32 million theses/books/proceeding data) [14]. CNKI includes a great number of academic resources including Chinese theses. The number of theses which has been counted since 1915 reaches up to 82 million. It increases by 20,000 papers every year. In addition, it includes 1 billion online academic references and 20 million foreign academic resources. CNKI features three plagiarism detection programs; AMLC for science technology references, TMLC for academic papers and SMLC for social science/literature. CNKI AMLC is developed based on China Knowledge Resource Integrated Database (CKRID), which has taken 5 years to build. CNKI integrates internet resources of Chinese and foreign languages and utilizes its own core technology to make CNKI AMLC. The system is so fast that it just takes 0.2 second to detect a plagiarized paper. As of July 2010, a list of 5,000 users was acquired. Then,

plagiarism detection was conducted on more than 3 million papers since 2008 [10]. TMLC has been installed and operated in 360 universities in China to detect and prevent plagiarism [15].

2.2. Plagiarism Similarity Detection Algorithm

Regarding text-based plagiarism similarity detection, a lot of studies have been conducted, and their topics have been classified from the two perspective. First, they can be classified into external and intrinsic plagiarism detections depending on whether not external knowledge is used [16]. The external plagiarism detection method has high accuracy provided that external knowledge is sufficiently provided. In fact, it has been applied to most studies. Second, the topics can be classified depending on how the judgment for plagiarism is approached. Stein classified them into global and local similarity strategies depending on the type of plagiarism analysis approach [17, 18]. A global similarity analysis is a method to find similarities by using commonly matched information based on global term vectors. This method is advantageous in detecting similar documents with fast search speed, but it has difficulty in detecting similar areas in the documents. On the contrary, a local similarity analysis is a method to check if particular parts are matched based on consecutive phrases and overlapped paragraphs. It has a relatively high accuracy in finding similar parts in documents, but it needs to enhance detection speed and increase storage space.

The global similarity method is developed based on Vector Space Model (VSM) [19]. Because of the effect of information search, it has a firm ground. Zechner [16] proposed a method to detect plagiarism by configuring the conventional inverted indexing technique in a unit of paragraph while Rehurek [20] suggested a latent Semantic indexing method. Regarding related studies with local similarity analysis, there are sentence comparison technique and fingerprint matching technique. Campbell [21] analyzed six plagiarism detection systems and proposed sentence-based hash algorithm and system. Yerra [22] examined sentence-based plagiarism detection techniques against web data using 4-gram or fuzzy-set technique. The fingerprint-based method shortens search time by substituting conventional words with connotative code. A variety of studies have been conducted focusing on creation of hash code and the comparison method. Schleimer [23] proposed a method to detect similarities in certain length by reducing calculation cost. Stein [17] proposed a way to easily identify candidate sets, using a concept of neighbor similarity based on fuzzy. Among the commercial system, iThenticate, provides plagiarism similarity search service by using its own undisclosed fingerprint-based plagiarism similarity search [12]. AMLC provides multi-level fingerprint-based plagiarism similarity search services [10]. It is unknown how the two systems are operated, but they have been widely used in practice.

3. Design of Korean Plagiarism Detection System (HPDS)

To design the HPDS considering future extensibility, the core functions to support plagiarism detection on Korean papers are derived at first. While the conventional plagiarism detection systems provide plagiarism similarity detection services only, the main functions of HPDS are divided into plagiarism similarity search service, reference & information service and case study support service as shown in Figure 1 below.

Plagiarism similarity search service, the key function of HPDS, provides basic information for plagiarism detection. As shown in related works, various techniques or system could be applied depending on the target of inspection. As a result, there could be big difference in plagiarism inspection performance. Therefore, a plagiarism similarity search service is designed to make it easy to apply and use multiple

plagiarism detection engines properly and effectively support the processing of Korean texts.

To use various plagiarism search engines, Plug-in API concept is applied. It is common technology which allows an external API developer to expand system functions at development of a particular system. It depends on the communication or input/output specifications. In HPDS, this concept was applied in the opposite. The plagiarism similarity detection services were modified into a manner in which information on external plagiarism similarity detection engine is examined and saved in advance. Under this method, the plagiarism similarity search service calls API which meets the goal among multiple similarity search engines, using API mapping tables pre-defined.

The plagiarism similarity search engine with its own algorithm could be developed and applied. Because it is urgent to develop and provide a plagiarism detection system, however, a commercial engine is adopted. Since fingerprint matching is known as the most efficient plagiarism similarity search technique, this research decided to use iThenticate which has adopted the fingerprint matching technique for its plagiarism similarity search engine. The fingerprint matching-based similarity search procedure is explained under the Figure 1. Under this method, index tokens are created using fingerprints and used as search index. In addition, it provides the results of plagiarism similarity detection in four stages; acquisition of documents, extraction & indexing fingerprints, plagiarism similarity detection and report of results. The final result of plagiarism similarity detection is reported (0~100%) by summing partial plagiarism similarities.

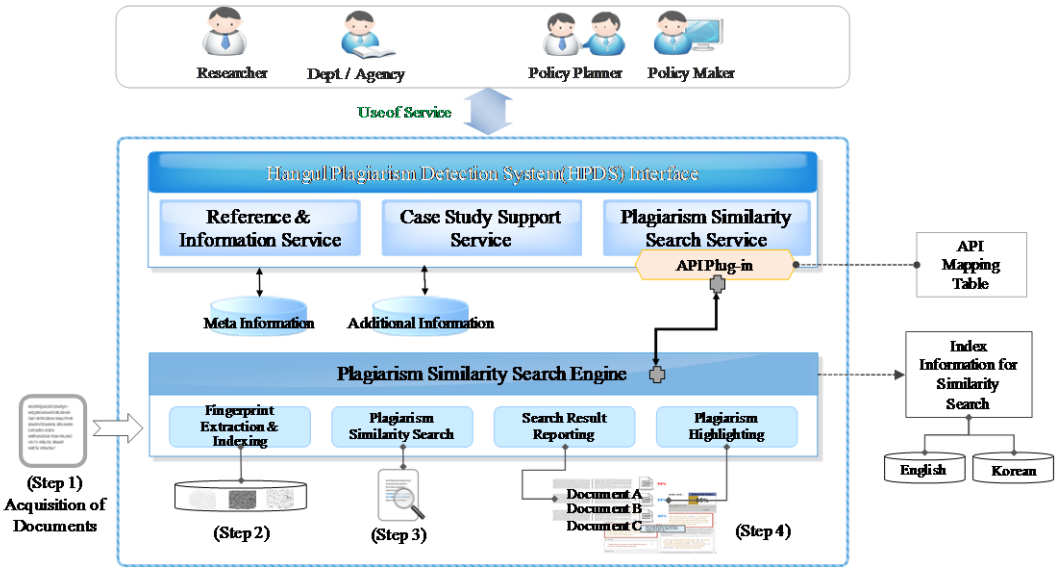

Figure. 1. Schematic Diagram of Korean Plagiarism Detection System

The experts from the Committee on Research Integrity decide whether or not the contents are plagiarized after getting help from the plagiarism similarity detection system. Moreover, iThenticate can be independently operated with a particular language including Korean language. The results of the evaluation on the performance of the Korean-language plagiarism similarity detection of iThenticate are stated in Chapter 6.

The reference & information service and case study support service are additional services to enhance the effect of plagiarism detection by complementing the negative aspect of plagiarism similarity search. The reference & information service is aimed to provide information on the field of research to researchers. It provides reference information associated with the fields of research such as research trend and thesis & patent analysis information. The case study support service is designed for researchers who start or review a certain field of research. If a keyword related with the research topic is entered, a list of related papers and experts is suggested to users through similarity analysis. It is expected that the two functions above would be available as core functions to support plagiarism similarity detection in the future.

4. Implementation of Korean Plagiarism Similarity Search System

4.1. Implementation of HPDS

Among the three major functions of HPDS, the implementation of plagiarism similarity search service is only explained here. For interlocking between HPDS and iThenticate, it is necessary to improve certain functions of iThenticate. To use an API plug-in method, linking API was defined in iThenticate. The linking API provided in iThenticate consists of 17 major APIs including login, group & folder management, user management, submission of documents and report search. The functions from iThenticate web pages are not all provided in API, but most functions except for lookup of the resulting report are available in API. To make this API information available in the similarity search service, API mapping table-based management functions were implemented.

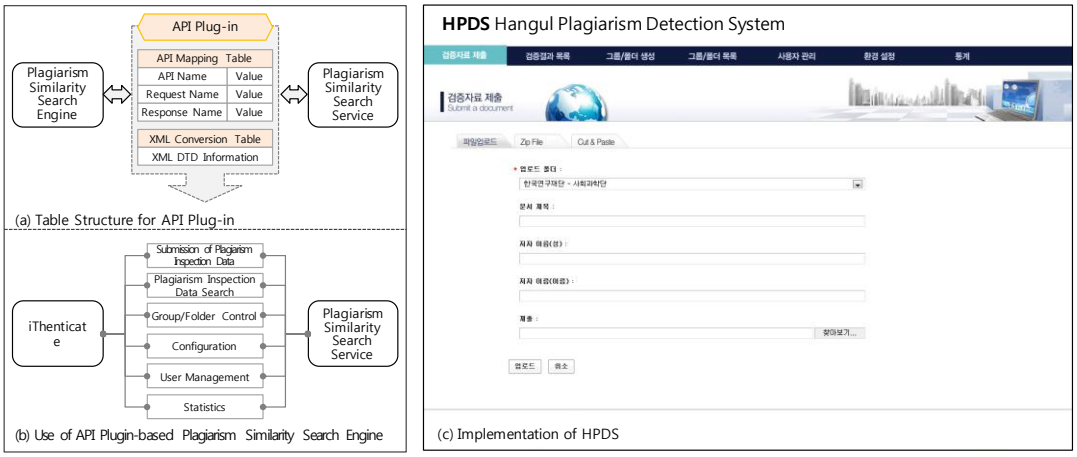

Figure 2. Implementation of HPDS-iThenticate Linkage

The linking system between HPDS and iThenticate using API plug-in is overviewed in (a, b) in Figure 2, and the Korean version is shown in (c). As shown in Figure 2, API plug-in makes it easy for both iThenticate and HPDS which provide and use API each to use the function for connection. As shown in Figure (2-a), API name, request value and reply value are managed using the API mapping table. In addition, information to send related data is managed using an XML conversion table. When API plug-in is used, the major API functions which are applicable to both plagiarism similarity search service and iThenticate are shown in Figure (2-b).

4.2. Generation of Korean article DB for the plagiarism detection

According to an analysis on collection of Korean contents in iThenticate, even though a great number of Korean web pages are included, the contents on academic paper, research report and journals are hardly found.

In this research, a plagiarism similarity search system was developed, and at the same time database on Korean contents was established to provide the ground for plagiarism similarity search on Korean contents. We seriously considered copyright issues. Among 54 different types of journals which have no problem in intellectual property right issues among academic papers in science technology, 10,010 papers are targeted to establish indexed database. The journals use either Korean or English. For this collection, 50.8% of the selected papers were written in Korean, and individual papers are available in a PDF format.

In designing HPDS, the support on uploading of contents in various document formats and verification on the impact of cited parts on plagiarism similarity were considered. Additional works were performed at development of database to find out the effect of change in similarity levels after automatic conversion of document format and processing of cited information on plagiarism detection. For this, data were developed in two different formats (original PDF, manually processed XML). The development of database in an XML format was limited to exact paragraph processing and cited parts. It was designed to be able to clearly classify a paragraph after a manual processing on the texts extracted from the original PDF, and citation was identified. The process and related information to handle the data in an XML format are shown in Figure 3.

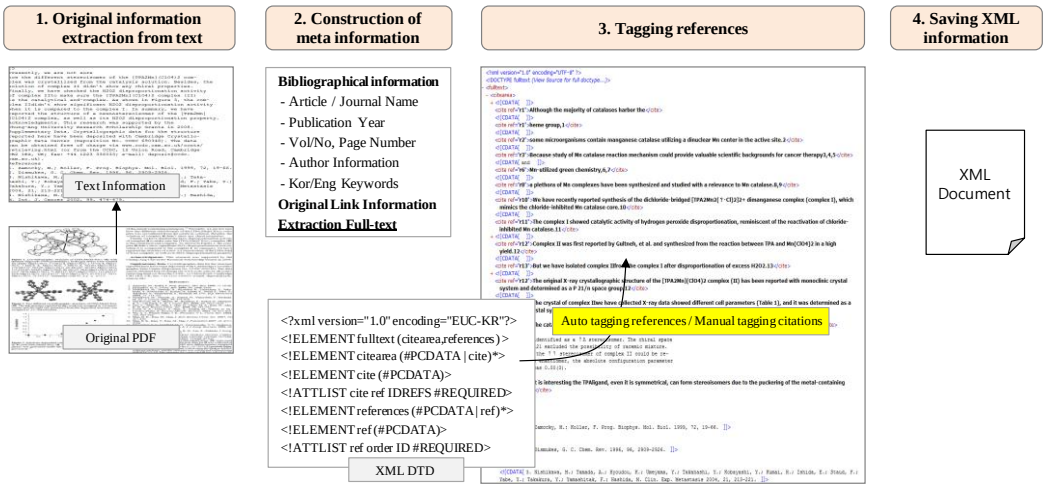

Figure 3. XML Data Development Procedure and Related Information

According to analysis on the collection of cited information, 75.2% were cited in a single paragraph while 10.6%, 7.8%, 5.3% and 1.3% were cited in phrase, clause, multiple sentences and others respectively, which show the unit of citation in science technology papers in Korea.

Korean language-only repository has been formed in iThenticate for its independent configuration. The prepared data were divided into PDF and XML formats and stored in the Korean language-only repository. Either repository can be selected at detection of plagiarism similarity.

5. Experimental Results

In terms of performance evaluation, plagiarism similarity search service among various functions of HPDS is intensively investigated. The performance is evaluated based on the following criteria; search speed, conversion of original texts, use of quoted information and Korean-language plagiarism similarity search performance. Among the total of 54 different types of Korean-language papers, a total 100 papers (2 papers per type) were selected, and an experiment data was set. According to analysis on search speed against the experiment data, it ranged from 20 seconds to 2 minutes. Then, according to investigation on changes in similarity by automatic conversion of document format, there was a slight change (less than 10%) in similarity levels because of the occurrence of a forced paragraph in the process of extracting texts when they were automatically sampled. However, the effect was minor. When the quoted part are excluded from the plagiarism similarity test which is performed using the data in XML format, it is likely that plagiarism inspection time could be somewhat reduced by narrowing down the scope which should be checked by a user.

Considering the significant efforts and costs spent for the construction of XML contents, it could be effective to use original data as they are without a separate processing. Finally, it has been identified that the performance of plagiarism similarity search shows similar results regardless of the length of plagiarized contents. Because the whole Korean contents are not indexed yet, however, plagiarism similarity search levels differ by field. In case of Korean language, in addition, search function could not be available due to spacing words and relocation of words.

6. Conclusion and Future Works

In this research, HPDS which enables plagiarism similarity search on both Korean and English contents has been designed and implemented. For effective linkage of external similarity search engines during implementation, API plug-in technique is used. For plagiarism similarity search engine, iThenticate, a fingerprint matching technique-based commercial system, was chosen. To make it possible to perform plagiarism similarity search on Korean-language theses, about then thousand of Korean-language papers were gathered and stored in HPDS. According to performance evaluation on HPDS, some parts which are specialized in Korean language need to be improved. In overall, however, it has produced satisfying results.

In the future, it would be possible to examine the plagiarism similarity search functions of HPDS more meticulously by suggesting analysis results in detail from diverse perspectives such as the size & length of plagiarized sentence and a type of plagiarized contents. Moreover, it is planned to develop a plagiarism similarity search engine which reflects the characteristics of Korean language more effectively and have it linked with the HPDS.

Acknowledgements

This work was supported by Development of a Prototype System for Plagiarism Detection on Academic Papers Project funded by the Ministry of Education, Science and Technology (MEST) (N-11-NM-09-00R-1).

References

[1] Y. M. Lim and C. E. Yoon, "Research Integrity in Science", Issue Paper of Samsung Economic Research Institute, **(2006)** March.
[2] D. C. Kwack, "A Study on the Types of Plagiarism and Appropriate Citation Practices of Writing Research Papers", Korean Journal of Library and Information, vol. 3, no. 41, **(2007)**.
[3] K. S. Choi and W. K. Joo, "A Study of Plagiarism Prevention System", **(2010)**.
[4] Plagiarism Projects of JISC, http://www.jisc.ac.uk/whatwedo/topics/plagiarism.aspx.
[5] Turnitin, http://www.turnitin.com.
[6] ORI, http://ori.hhs.gov/.
[7] CrossRef, http://www.crossref.org.
[8] COPE (Committee on Publication Ethics), http://publicationethics.org.
[9] CNKI AMLC, http://check.cnki.net/amlc2/.
[10] X. Sun, "CNKI AMLC System's Advances in Development and Application and the Plan for International Cooperation", Proceedings of Second WCRI (World Conference on Research Integrity), **(2010)** July 21-24; Singapore.
[11] MEST, "Establishing and Promoting Research Ethics: at National and Transnational Levels", **(2008)**.
[12] iThenticate, http://www.ithenticate.com.
[13] MemeChecker, http://www.memechecker.com.
[14] iThenticate database contents, http://www.ithenticate.com/plagiarism-detection-database.
[15] Z. Xiaojun, S. Hongli and Z. Fan, "Preventing Plagiarism and Academic Misconduct: A Case Study of Chinese Universities", Proceedings of Fourth International Plagiarism Conference, **(2010)** July 05; Newcastle upon Tyne, UK.
[16] M. Zechner, M. Muhr, R. Kern and M. Granitzer, "External and Intrinsic Plagiarism Detection Using Vector Space Models", Proceedings of third Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, **(2009)** September 10; San Sebastian, Spain.
[17] B. Stein and S. M. Z. Eissen, "Near Similarity Search and Plagiarism Analysis", Proceedings of 29th Annual Conference of the German Classification Society, **(2005)** March 9-11; Magdeburg, Germany.
[18] B. Stein, "Fuzzy-Fingerprints for Text-based Information Retrieval", Proceedings of 5th International Conference on Knowledge Management, **(2005)** June 29-July 1; Graz, Austria.
[19] G. Salton and M. J. McGill, Editors, "Introduction to Modern Information Retrieval", McGraw Hill, New York, **(1983)**.
[20] R. Rehurek, "Plagiarism Detection through Vector Space Models Applied to a Digital Library", Second Workshop on Recent Advances in Slavonic Natural Language Processing, **(2008)** December 5-7.
[21] D. M. Campbell, W. R. Chen and R. D. Smith, "Copy Detection Systems for Digital Documents", Proceedings of Advances in Digital Libraries, **(2000)** May 22-24.
[22] R. Yerra and Y. K. Ng, "A Sentence-Based Copy Detection Approach for Web Documents", Lecture Notes in Computer Science, vol. 3613, **(2005)**, pp. 557-570.
[23] S. Schleimer, D. S. Wilkerson and A. Aiken, "Winnowing: local algorithms for document fingerprinting", Proceedings of ACM SIGMOD, **(2003)**.

Authors

Won-Kyun Joo

He received the B.S. and M.S. degrees in Computer Science from Chungnam National University, Korea in 1997 and 1999 respectively. Currently, he is a senior researcher at KISTI.

KiSeok Choi

He received the B.S. degree in Computer Science and Statistics from Seoul National University, Korea in 1988 and M.S. degree in Computer Science from KAIST in 1997. Currently, he is a leader of R&D System Development office at KISTI.

YongJu Shin

She received the B.S. degree in Library and Information Science from Hannam University, Korea in 2012. Currently, she is a senior researcher at KISTI.

JaeSoo Kim

He received the B.S. degree in Computer Science from Hongik University, Korea in 1985, M.S. degree in Computer Science from Hankuk University of Foreign Studies, Korea in 1987, and Ph.D. degree in Electronic and Computer Engineering from Hongik University, Korea in 2009. Currently, he is a director of NTIS center at KISTI.