International Journal of Multimedia and Ubiquitous Engineering
Vol. 9, No. 10 (2014), pp. 33-44
http://dx.doi.org/10.14257/ijmue.2014.9.10.04

Implementation of Plagiarism Analysis System through Digital Conversion Processing of Sound Source

Mi-Hae Shin[1], Eui-Jeong Kim[1], Su-Seok Seo[2] and Young-Chul Kim[3]

[1]Department of Computer Education, Kongju National University, Gongju 314-701, Korea
[2]Department of E-Business, Yuhan College, Bucheon 422-749, Korea
[3]Department of Smart Communications, Yuhan College, Bucheon 422-749, Korea
tlsalgo@nate.com, ejkim@kongju.ac.kr, ssseo@yuhan.ac.kr, kim0725@yuhan.ac.kr

Abstract

The melody in music can only be assessed by the assessor's emotions so it can only be subjective and cannot be expressed in standardized numbers. Therefore nobody can know for sure if two songs are plagiarized or not. In this paper, we attempt to realize a plagiarism analysis system of musical content that can provide judging grounds for plagiarism – usually analyzed in consideration of human emotions - using IT technology. To protect music copyright, both knowledge of music and IT technology is compulsory to create a music plagiarism analysis system. First, by understanding factors of the musical content and analyzing plagiarism between two separate but similar voices. We tried to invent a system that measures the level of similarity. For this we considered the digital music factors for the analysis of voices and searched for methods to use IT technology for judging plagiarism. For an efficient analysis of plagiarism, we utilized the music string provided by JFugue, and designed a plagiarism analysis system for musical content and proposed an abstract syntax tree (AST).

Keywords: *digital musical contents, digital conversion, plagiarism analysis system, digital conversion processing*

1. Introduction

With the development of information communication technology and a sudden increase in data use, much interest has been shown in searching various contents such as image, video, audio, *etc.* [1], and there has been industrial development that satisfies the emotional desires of humans and creates a profit system. Especially in the case of popular culture that deals with a lot of people, there has been rapid advancement along with content digitalization and it is yielding a lot of influence on the development of a country and is also boosting the national image [2, 3].

Due to new technology, compared to other cultural content the comparatively short-lived popular music has developed while experiencing many sudden changes. In the case of musical content, accessibility to diverse information has become infinite through the internet its importance has increased as well as its demand and with it copyright issues have become a cause for international conflict and economic problems. Infringement of copyright due to plagiarism of popular music is especially becoming a bigger problem day by day because of a lack of rights consciousness and standards for legal judgment.

ISSN: 1975-0080 IJMUE

In this paper, we attempt to realize a plagiarism analysis system of musical content that can provide judging grounds for plagiarism – usually analyzed in consideration of human emotions - using IT technology. To protect music copyright, both knowledge of music and IT technology is compulsory to make a music plagiarism analysis system. In other words, without knowledge of music, one cannot properly understand the properties of the voice and without IT technology, one cannot create a plagiarism analysis system.

The melody in music can only be assessed by the assessor's emotions so it can only be subjective and it cannot be expressed in standardized numbers. So nobody can know for sure if two songs are plagiarized or not. We cannot be sure whether the knowledge was used unconsciously due to development of the internet, but there must be a systematic standard of judgment to protect the rights of the artist [4-7]. Based on many studies, this paper excludes the involvement of emotion from the assessor and focused instead on a plagiarism analysis system that can be approached systematically.

The plagiarism analysis system for musical content introduced in this paper and its implementation method will provide a new solution for plagiarism in the field of music that is a subject for debate in the past, present and future. Also, by providing standardized numbers for plagiarism issues that could only be judged by emotional standards in the past, it will suggest grounds for a new solution. Also, it will enable artists to protect their individual creative work from plagiarism of musical content, give a cost-cutting effect in plagiarism analysis, save time, and secure the evolving information searching and analysis technology. Lastly, we can expect effects such as securing basic technology regarding plagiarism domestically and gaining international competitiveness.

2. Musical Content Search Technology

The necessity of searching information of voices both on and offline is increasing because of the development in the digital voice market due to network advancement through the internet.

Musical content search technology is utilized in searching for similar music by analyzing the music recorded through digital signs or finding music information that the user needs. Thanks to the recent rapid development of computer functions in transferring large amounts of data in a short time along with the advent of a high-speed communication network, musical content search technology is making steady development.

Musical content search technology can be distinguished between symbol-based analysis and signal-spectrum based analysis, depending on the standard of analysis [8, 9].

The symbol-based analysis is analyzing the notes on the music score, so the analysis focus is on how exact the visual description of the music score is, and whether it can be applied in real performance. When dealing with musical information in symbol-based analysis, either a graphic interface in the form of a music score or the method of using script language is used, and it was mostly used in analyzing western classical music in which the instrumental sounds are clearly defined, but recently the range of study has been limited to classical music data. Because professional knowledge of music is required for symbol-based analysis, the fact that only professionals with a deep understanding of classical or vocal music can participate in the study is its distinct characteristic. The signal-spectrum based analysis is a technique that analyses the sampling signals of recorded music, and takes up a big part of musical content searching technology. Music consumed by the public recently is less classical music played according to a music score, and more popular music that is recorded with the human voice and synthesizer. And because this kind of popular music mostly uses sampling audio signals such as the CD and MP3, so the importance and necessity of signal-spectrum based analysis

is increasing to analyze this kind of music. Because signal-spectrum-based analysis does not require as much professional knowledge, a lot of research is being made by electronic engineers and computer engineers with experience in signal processing or voice recognition technology.

3. Design of Plagiarism Analysis System for Musical Content

3.1. Research Method

It is difficult to decide the criteria adjustable to plagiarism that determines how much of the plagiarized work is expropriated. A certain amount needs to be appropriated in order to acknowledge actual likeness, but even small amounts appropriated can be considered plagiarism if it is a significant part that contains the creative efforts of the artist [10-13].

Basic knowledge about music is required in order to determine whether musical content was plagiarized or not. Therefore, in order to diagnose the plagiarism problem and set a standard, there must be a preceding operation of understanding the various elements of musical content. Then an analytic technique that infers the level of similarity from the various characteristics of musical content elements is used [14-17].

Musical content is generally composed of the melody, rhythm and chord, and these three elements constitute a musical format by being arranged and chosen while following a certain order. However, many of the existing research on musical content plagiarism has only a part of the musical content elements as the research subject [18-20].

Table 1. Methods of Plagiarism Analysis System of Musical Contents

<Digital conversion of the sound source> - Sound source (melody, rhythm, chord, etc.) - Tie, length, triplets - Bar, key signatures, musical instruments
↓
<Sound analysis> - MIDI percussion tracks, tempo, pitch, channel programs, etc. - The pitch of the sound source matrix output format - Music String create and produce BNF grammar transformation
↓
<Implementation of plagiarism analysis system of musical contents> - Appropriate similarity analysis algorithm running on the sound source expression - Note values, note length optimization - Plagiarism checking algorithm verification

The aim of this paper is to create a system that can analyze various voices by applying technology related to musical content. For this, we tried to design a system that measures the similarity in two different voices by analyzing plagiarism on the basis of detailed understanding of musical content elements, and the basic research methods are as shown in Table 1.

3.2. Digital Transformation Process of Voices

For the digital transformation process of voices, there needs to be an analysis of not only the melody, rhythm and chord but also notes, rests, chords, *etc.*, and also their digital transformation in order to analyze plagiarism.

In this paper, the voices were transformed into digital signals through the method of converting the midi into Music String, by using the open source library 'JFugue' and 'Smile v.5.2 for Csound 5'. JFugue farces Music String and plays each note. So Music String has a syntax that focuses on the performance, and so is appropriate for plagiarism analysis by converting to AST or partial performance. Voices can be widely circulated because it can easily be transferred through the internet or other media. This entails another side effect of music that simplifies plagiarism and leaves room for the use of similar notes on an emotional basis. In order to formally analyze these side effects we conducted a plagiarism analysis. Figure 1 shows a screen of an example composition using a music system called 'Smile V.5.2. for Csound 5'. Based on this, we can conduct plagiarism analysis by streaming voices.

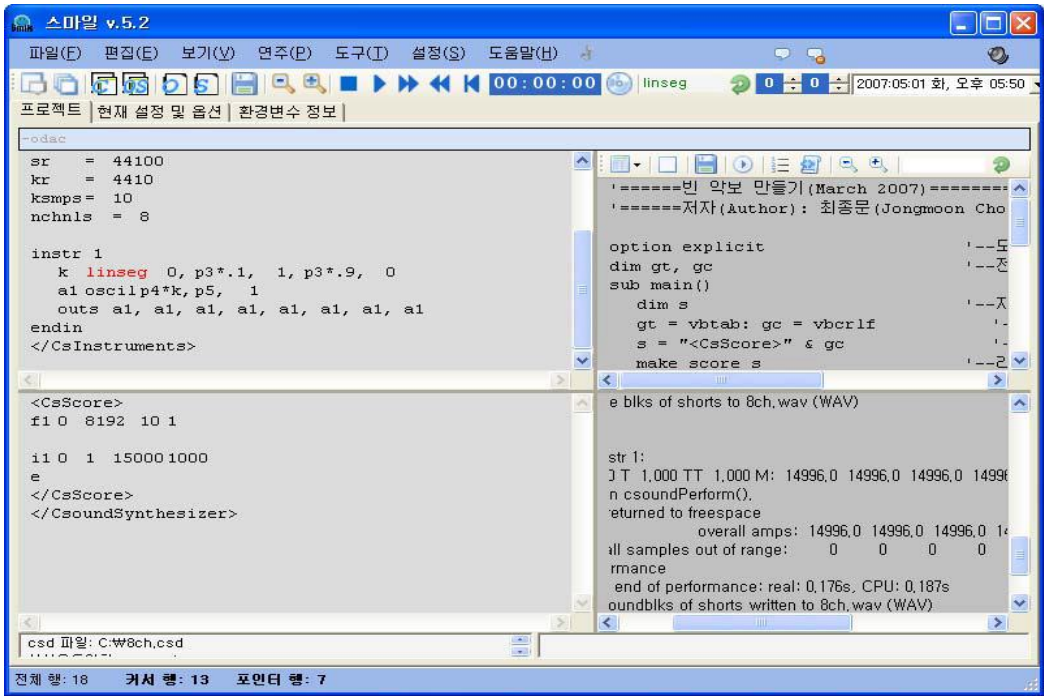

Figure 1. Smile V5.2 for Csound 5

A voice is also called a channel or a track. Midi supports 16 channels simultaneously. Each voice has its own melody and uses a specific instrument. JFugue uses a value between 0 and 15. In the case of operations in Midi, because there are only 16 channels, a number of sources are used alternately.

It is also very important to note that there has been almost no research that digitize voice elements, and that this is a breaking of new ground. Also, these kinds of numbers give grounds for analyzing the level of similarity between two songs whether they are plagiarized or not, and there can be said to be great meaning in measuring the voices itself.

3.3. Voice Analysis

While creating Music String that will be used for pitch matrix algorithm for voice analysis with the digitally transformed voice as the base, we created material that will be used for the plagiarism analysis system for musical content through carrying out research on AST(Abstract Syntax Tree) transformation and Music String BNF(Backus Naur Form) transformation.

First, in order to develop a plagiarism analysis module for digital music, a standardizing method that can discriminate plagiarism between two midi files was used. For this, by streaming notes related to music we transformed them into character strings and used the method of determining plagiarism between the two songs through a similarity algorithm. A more detailed method is using the pitch matrix algorithm [7]. The pitch matrix algorithm is an algorithm that transforms a tone row constituting more than 1 pitch into a 'set complex(magic square)'. If the proposed tone row is 'C C# D Eb E F F# G Ab A Bb B', the prime tone row becomes 'C D E G A C D E G A C D E G A C D E G'.

C	D	E	G	A	C	D	E	G	A	C	D	E	G	A	C	D	E	G
Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F
Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb
F	G	A	C	D	F	G	A	C	D	F	G	A	C	D	F	G	A	C
Eb	F	G	Bb	C	Eb	F	G	Bb	C	Eb	F	G	Bb	C	Eb	F	G	Bb
C	D	E	G	A	C	D	E	G	A	C	D	E	G	A	C	D	E	G
Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F
Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb
F	G	A	C	D	F	G	A	C	D	F	G	A	C	D	F	G	A	C
Eb	F	G	Bb	C	Eb	F	G	Bb	C	Eb	F	G	Bb	C	Eb	F	G	Bb
C	D	E	G	A	C	D	E	G	A	C	D	E	G	A	C	D	E	G
Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F
Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb
F	G	A	C	D	F	G	A	C	D	F	G	A	C	D	F	G	A	C
Eb	F	G	Bb	C	Eb	F	G	Bb	C	Eb	F	G	Bb	C	Eb	F	G	Bb
C	D	E	G	A	C	D	E	G	A	C	D	E	G	A	C	D	E	G
Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F	G	Bb	C	D	F
Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb	F	Ab	Bb	C	Eb
F	G	A	C	D	F	G	A	C	D	F	G	A	C	D	F	G	A	C

Figure 2. Pitch Matrix

Music String is writing music performance information in the form of a character string, so the C note played would be expressed as 'C'. Each note is played at a token unit that is distinguished by blank characters, and in order to analyze the level of similarity each token was built into an AST node. A token constitutes the note, chord, rest, instrument change, beat indication, controller event, constant definition, *etc*.

The notes C, D, E, F, G, A, B, R(rest), and half-note raise(#), half-note lower(b), octave(0~10, default 5) were used, and while comparing the level of similarity the two elements note character string and note figure were used. The chord was indicated after the note, and Cmaj and C+E+G are the same yet the latter was used for notation in order to maintain consistency and measure similarity. The default duration was 1/4 notes, and duration was used right after the octave or right after the note when it was omitted, and it was shown in alphabet letters or numbers.

Next, AST transformation is defined as abstract syntax, which is a collection of Music String syntax. Each music file, depending on the syntax, is expressed as a derivation tree, or in other words an AST. AST is a data structure that was made to be able to analyze music plagiarism by farcing files. The analysis results among music files are all reflected in the AST instance.

The final step of voice analysis is making the BNF and using it in the plagiarism analysis system, and the following shows part of the BNF that we are attempting to create while transforming characters in this research for the syntax study of Music String BNF.

musicstring := (element whitespace)+ element?

element:= voice | tempo | instrument | layer | key | controller | time | poly_pressure | channel_pressure | pitch_bend | measure | expression | system_exclusive | collected_note

- omit -

voice := "V" int_or_const
tempo := "T" int_or_const
instrument := "I" int_or_const
layer := "L" int_or_const
time := "@" int_or_const
poly_pressure := "*" int_or_const
channel_pressure := "+" int_or_const

pitch_bend := "&" int_or_const

measure := "|"

controller := "X" int_or_const "=" controller_value
controller_value := int | symbol

- omit -

key := "K" root scale
root := letter_note | int_note
scale := "MAJ" | "MIN"
letter_note := ["A"-"G"] modifier?
modifier := "#" "#"? | "B" "B"? | "N"
int_note := "[" int "]"

note := ("R" | letter_note | int_note) octave? (chord_scale chord_inversion?)? duration? velocity* ("+" note)*
octave := "10" | digit
chord_scale := scale | "AUG" | "DIM" | "DOM7" | "MAJ7" | "MIN7" | "SUS4" | "SUS2" | "MAJ6" | "MIN6" | "DOM9" | "MAJ9" | "MIN9" | "DIM7" | "ADD9" | "DAVE" | "MIN11" | "DOM11" | "DOM13" | "MIN13" | "MAJ13" | "DOM7<5" | "DOM7>5" | "MAJ7<5" | "MAJ7>5" | "MINMAJ7" | "DOM7<5<9" | "DOM7<5>9" | "DOM7>5<9" | "DOM7>5>9"

- The last part omitted –

4. Realization of the Plagiarism Analysis System of Musical Content

In order to determine whether or not there was plagiarism between two midi files, there needs to be a standardized method. To realize this the notes related to music need to be streamed and transformed into character strings, and by using the similarity algorithm the plagiarism analysis system of musical content determines whether there was plagiarism or not. This is shown in the following Figure 2.

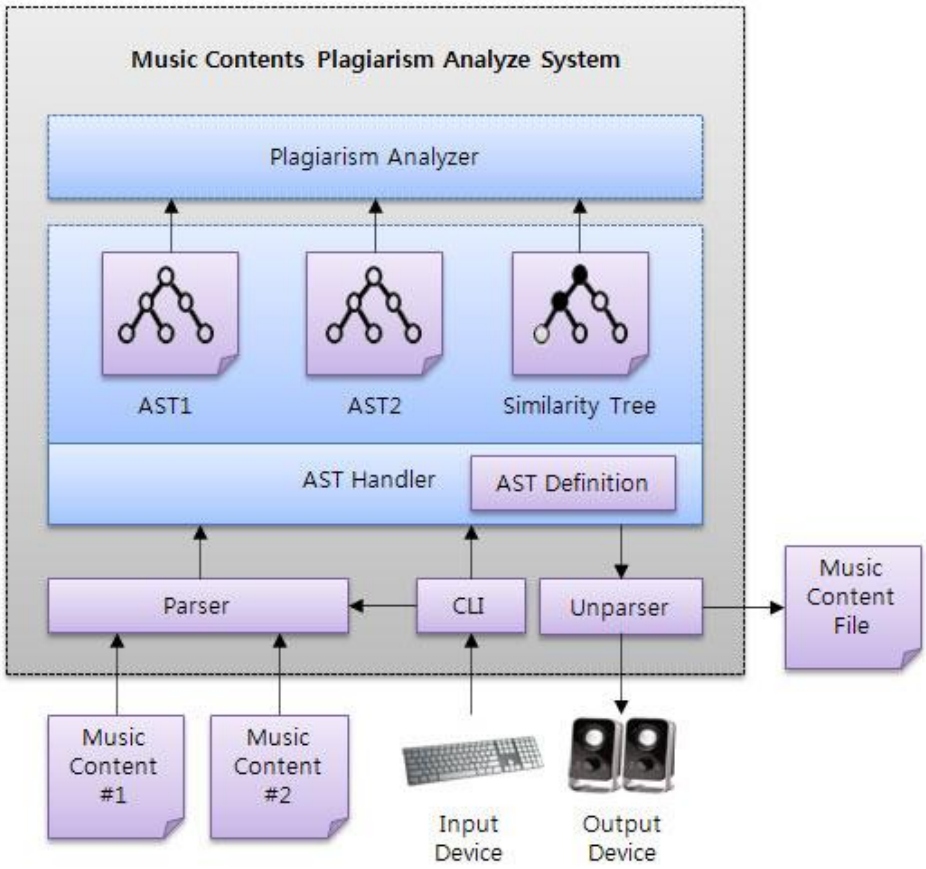

Figure 3. Scheme of Plagiarism Analysis System of Musical Contents

The parser performs a syntax analysis of the music file and examines its suitability. After being verified by the parser, the AST handler creates the AST by referring to the ASTD (Abstract Syntax Tree Definition), which is a syntax tree made to process music files in an internal plagiarism analysis instrument or reverse parser. The user receives the input command that farces files, analyzes plagiarism or outputs results from the CLI (Command Line Interpreter). The plagiarism analyzer, after comparing the plagiarism in 2 ASTs, creates a similarity tree with the results in the form of AST. This similarity tree outputs the plagiarized part through the speaker or as a music file through the Unparser.

The similarity analysis algorithm used for the realization of the plagiarism analysis system for musical content is as such.

```
double Sim(NodeString A, NodeString B, long int minlength) {
```

```
String matchstring, totalmatchstring; /* Match the string */
int maxmatch = 0; /* Initialize the number of strings matched */
long int matchlength = 0; /* Initialize the total number of strings matched */
Set(totalmatchstring) = {}; /* Full set of matching strings */

/* String matching algorithm until it finds one, two repeat */
do {
matchstring = ""; /* Match the string */
matchstring = MatchString(A, B); /* Algorithm 1 calls */
Set(totalmatchstring) = Set(totalmatchstring) + matchstring;
} while (maxmatch > minlength);

/* Calculate the total number of the string to match */
for each matchstring in Set(totalmatchstring)
matchlength = matchlength + Length(matchstring);
end for

/* Calculated and returns of similarity value */
return ();
}
```

In detail, like shown in Table 2, it was designed so that the level of similarity between 2 different voices would be analyzed, and then the output results would show the same bar and similarity value.

A similarity algorithm of this system returns a value which is greater or equal to 0 and less or equal to 1, and the meanings of this value is as follows.

If a value of a similarity is 0, one of two node strings which will be checked are empty node strings, or the length of the identical string between two programs is less than a minlength. If a value of a similarity is 1, it means that node strings A and B which will be checked are identical completely. In case a value of similarity is greater than 0 and less than 1, it means that two programs which will be performed a plagiarism check are identical partially.

Table 2. Similarity Evaluation Engine

<Sound similarity evaluation engine>
- Algorithm 1 : Music String match search, save, delete, - Algorithm 2 : Similarity measure
↓
<Results of plagiarism analysis>
- Musical content analysis of plagiarism : 0 <Sim (p1, p2) 1 - Output : 24 words in four word matching, similarity 0.17 : Matching part: Dm-C-Bb-Am ...

For the verification of the plagiarism analysis system of musical content realized through this research, we applied the song "Alone" by the group 'CNBLUE' in 2010 with the song "Bluebird" released in 2008. 4. 26 by the indie band 'ynot?' songs which were actually under plagiarism disputes in 2011. As a result, we easily found that 4 out of 24 bars accorded and also the level of similarity (0.17) and resulting value of the according parts (Dm-C-Bb-Am). This affirms the Copyright Commission's ruling in April 27th, 2011 that CNBLUE's 'Alone' was not plagiarized, and by standardizing the time and effort spent by the judges into an objective system, it has great meaning in that it secured an information searching and analysis technology for musical content that is constantly evolving.

Table 3. Implementation and Verification Environment

Systems	Intel Core i5 3.0GHz
Operating systems	MS Windows 7
Programming language	C++
Programming tools	Visual Studio 2010
Verify the source	CNBLUE - Alone ynot? - Bluebird

5. Conclusion

This paper offers a new construction and implementation method of a plagiarism analysis system for musical content which can provide grounds for judging music plagiarism – usually analyzed in consideration of human emotions - through IT technology. For this, first the voice was digitally transformed in order to analyze diverse voices and a system determining plagiarism between two musical content using the similar algorithm was realized.

The factor evaluation system related to voice plagiarism examines and analyzes the structural similarity between two voices, with IT technology as the base. This technology enables a quantitative comparison of structural characteristics of musical content, and when plagiarism is quantified formally through this system, plagiarism analysis formerly depending on professional emotions can be carried out in a quantitative, subjective method. Henceforth, for the commercialization of the system, supplementation of research methods and improving system portability will result in more generalized and detailed research results.

Acknowledgements

This research was conducted out with the support of National Research Foundation of Korea (NRF), with the financial support of Ministry of Science, ICT and Future Planning (MSIP) (No. NRF-2012R1A2A2A03045162).

References

[1] J. L. Park, and S. W. Kim, "Development of a System for Music Plagiarism Detection Using Melody Databases," Journal of Multimedia Society, vol. 8, no. 1, **(2005)**, pp. 1-8.
[2] J. W. Jo, M. H. Shin, A. R. Park, and Y. C. Kim, "An analysis of Empirical Studies of Musical Literary Work Plagiarism Standard : The Popular Music," The Journal of The Korea Contents Association, vol. 14, no. 3, **(2014)**, pp. 176-185.
[3] M. H. Shin, J. W. Jo, H. S. Lee, and Y. C. Kim, "A Study of Digital Music Element for Music Plagiarism Analysis," Journal of the Korea society of computer and information, vol. 18, no. 8, **(2013)**, pp.43-52.

[4] P. J. Larkham, and S. Manns, "Plagiarism and its Treatment in Higher Education," Journal of Further and Higher Education, vol. 26, no. 4, **(2010)**, pp. 339-349.
[5] D. L. McCabe, L. K. Trevino, and K. D. Butterfield, "Cheating in academic institutions: A decade of research," Ethics and Behavior, vol. 11, no. 3, **(2010)**, pp. 219-232.
[6] C. W. Kim, and S. Park, "Enhancing Text Document Clustering Using Non-negative Matrix Factorization and WordNet," Journal of information and communication convergence engineering, vol. 11, no. 4, **(2013)**, pp. 241-246.
[7] H. I. Lim, "Comparing Binary Programs using Approximate Matching of k-grams," Journal of KISS: computing practices, vol. 18, no. 4, **(2012)**, pp. 288-299.
[8] Y. Y. Jang, and H. K. Kim, "A Case Study of Animation Plagiarism through the Case of Music Video "Sonata of Temptation"," Journal of The Korea Contents Association, vol. 11, no. 6, **(2011)**, pp. 144-154.
[9] W. A. Arentz, M. L. Hetland, and B. Olstad, "Retrieving Musical Information Based on Rhythm and Pitch Correlations," Journal of New Music Research, vol. 34, no. 2, **(2007)**, pp. 151-159.
[10] L. Alexandral, U. Bogerd, and J. Zobel, "An Architecture of Effective Music Information Retrieval," Journal of the American Society for Information Science and Technology, vol.55, no.12, **(2004)**, pp. 1053-1057.
[11] S. W. Park, J. Y. Kim, T. H. Lee, S. B. Hong, J. S. Lim, and W. S. Kang, "Development of Document Plagiarism Detection Algorithm using Syntactic Analysis Method," Proceeding of The Korean Association of Computer Education, vol. 17, no. 1, **(2013)**, pp. 89-93.
[12] J. Y. Kuo, F. C. Huang, C. Hung, L. Hong, and Z. Yang, "The Study of Plagiarism Detection for Object-Oriented Programming," 2012 Sixth International Conference on Genetic and Evolutionary Computing (ICGEC), Aug, **(2012)**, pp. 188-191.
[13] W. S. Kang, and D. S. Hwang, "Discriminator of Similar Documents Using Syntactic and Semantic Analysis," The Journal of the Korea Contents Association, vol. 14, no. 3, **(2014)**, pp. 40-51.
[14] D. Bhattacharjee, and S. Dutta, "Plagiarism Detection by Identifying the Equations," Procedia Technology, vol. 10, **(2013)**, pp. 715-723.
[15] M. S. Arefin, Y. Morimoto, and M. A. Sharif, "Bilingual plagiarism detector," 2011 14th International Conference on Computer and Information Technology (ICCIT), Dec, **(2011)**, pp.451-456.
[16] M. Drlik, M. Munk, and J. Skalka, "Usage analysis of system for theses acquisition and plagiarism detection," Procedia computer science, vol. 3, **(2011)**, pp. 866-871.
[17] J. H. Ji, G. Woo, and H. G. Cho, "A Plagiarism Detection Technique for Java Program Using Bytecode Analysis," Third International Conference on Convergence and Hybrid Information Technology (ICCIT), **(2008)**, pp. 1092-1098.
[18] I. S. Hwang, "Development of A Plagiarism Detection System Using Web Search and Morpheme Analysis," Journal of information technology applications & management, vol. 16, no. 1, **(2009)**, pp. 21-36.
[19] R. E. Roxas, N. R. Lim, and N. Bautista, "Automatic Generation of Plagiarism Detection Among Student Programs," 7th International Conference on Information Technology Based Higher Education and Training (ITHET), July, **(2006)**, pp. 226-235.
[20] D. W. Cho, "A Survey of Plagiarism Inspection Method for Efficient Protecting of Intellectual Properties and Proposal of Art works Plagiarism Inspection," 2003 Conference on The Korea Contents Association, Nov. vol. 1, no. 2, **(2003)**, pp. 72-78.

Authors

Mi-Hae Shin, she is studying in Ph.D. degree in Computer Science from Kongju National University of South Korea. She is interested in plagiarism analysis system, multimedia contents, multimedia education and virtual reality.

Eui-Jeong Kim, He received the M.S. degree and Ph.D. degree in Computer Science from Chungnam National University, Korea. He works at Kongju National University of South Korea as an associate professor. His research interests are pattern recognition, computer vision, multimedia and medical image processing.

Su-Seok, Seo, He received the M.S. degree in Electronic Commerce from Kongju National University, Korea, in 2003, and the Ph.D. degree in Electronic Commerce from Kongju National University, Korea, in 2007. He works at Yuhan College of South Korea as an assistant professor. His research interests are electronic commerce, electronic business models and social network services.

Young-Chul Kim, He received the M.S. degree and Ph.D. degree in Computer Science from Soongsil University, Korea. He works at Yuhan College of South Korea as an associate professor. His research interests are plagiarism analysis system, programming languages, network management system, compiler, XML and computer communication.