International Journal of Database Theory and Application
Vol.10, No.6 (2017), pp.47-60
http://dx.doi.org/10.14257/ijdta.2017.10.6.04

Analysis of Criminal Profiling Utilizing Structured and Unstructured Data

Yonghoon Kim and Mokdong Chung[21]

Dept. of Computer Engineering Pukyong National University, Busan, Korea
[1]kimyhjava@pukyong.ac.kr, [2]mdchung@pknu.ac.kr

Abstract

In general, the structured data knows the meaning of the sentence and unstructured data refers to an unknown means. Although the quantity of structured information in the entire data and within organizations is increasing, the majority of information remains available only in unstructured data. While different in form, both unstructured and structured information sources provide information about entities in the world and their properties and relations. Due to the recent rapid changes in society and wide spread of information devices, diverse digital information is utilized in a variety of economic and social analysis. Information related to the crime statistics by type of crime has been used as a major factor in crime. However, statistical analysis using only the structured data has the difficulty in the investigation by providing limited information to investigators and users. In this paper, structured data and unstructured data are analyzed by applying Korean Natural Language Processing (Ko-NLP) and the Latent Semantic Analysis (LSA) technique. It will provide a crime profile optimum system that can be applied to the crime profiling system or statistical analysis [1].

Keywords*: Forensic, Unstructured-Data, Profiling, Digital-Forensic, Crime-Profiling*

1. Introduction

Utilizing for information is an indispensable component of our lives. But we have been using the most of structured data in the entire data. Recent social and economic analysis of digital information has expanded into unstructured data and SNS and also has led big data analysis to be helpful in understanding the various phenomena in the society.

This paper suggests the system of collecting accurate and professional information that could be used as a straightforward investigation, namely deductive investigation and inductive investigation to provide more information to the professional investigators and profilers to formalize the informal data which was not be formalized during a criminal investigation materials and analysis. The system first extracts structured and unstructured data (criminal investigation data, interview data and social resources) by using the Korean Natural Language Processing (Ko-NLP) and generate data pattern in the extracted data. The resulting pattern is re-used for extracting data by saving it separately.

Extracted structured and unstructured data are stored in an array applied to the index by using the Hash Map, and suggest the analysis system for providing a high affinity between the words and sentences in the real time mode on the basis of a variety of conditions and categories by invoking data using the Latent Singular Analysis (LSA) statistical techniques based on linear algebra when consumers request and search data [1].

The remainder of this paper is organized as follows. Section 2 surveys the related work and the theoretical backgrounds. Section 3 introduces the proposed crime data analysis

*Corresponding Author : Mokdong Chung (mdchung@pknu.ac.kr)
Dept. of Computer Engineering, Pukyong National University, Busan, Korea

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2017 SERSC

system model. In Section 4, we apply the proposed model to a comparative analysis of the experimental result using structured data, using unstructured data and using both data. Finally, Section 5 presents our conclusion and the direction of our future research.

2. Related Works

2.1. Korean Natural Language Processing

Tagging is a part of speech given to the corpus or the only selected parts of speech from the ambiguous results of morphological analysis. This research presents the morphological analysis utilizing HanNanum Ko-NLP Java version which is Semantic Web Research Center (SWRC) at Korea Advanced Institute of Science and Technology.

2.2. Finite Automata

Finite Automata (FA) is the simplest form of the recognition. In general, The Recognition of the language is determined whether the input to receive the string, the string of the language that the sentence or not. FA is used in the method to design and implement the lexical analyzer. FA is defined as the component of five (5-tuple). And in the form of a mapping function is shown in equation (1).

$$
\begin{aligned}
&FA(Finite\ Automata)\\
&M = (Q, \Sigma, \delta, q_0, F)\\
&Q : Finite\ set\ of\ state\\
&\Sigma : Finite\ set\ of\ input\ symbol\\
&\delta : Mapping\ function\\
&q_0, : start\ or\ initial\ stat\\
&F : set\ of\ final\ state
\end{aligned}
\tag{1}
$$

The mapping function is called the stat transition function, because it indicates that the mobile reports the input symbol from one state to another. Depending on type is divided into Deterministic Finite Automata (DFA) and Nondeterministic Finite Automata (NFA).

The state transition is represented by the Nondeterministic Finite Automata (NFA) as shown in Figure 1 where pattern was created by the lexical analysis based on the data analyzed by Ko-NLP, and presented to classify unstructured and structured data by analyzing the relationship based on the state transition [1].

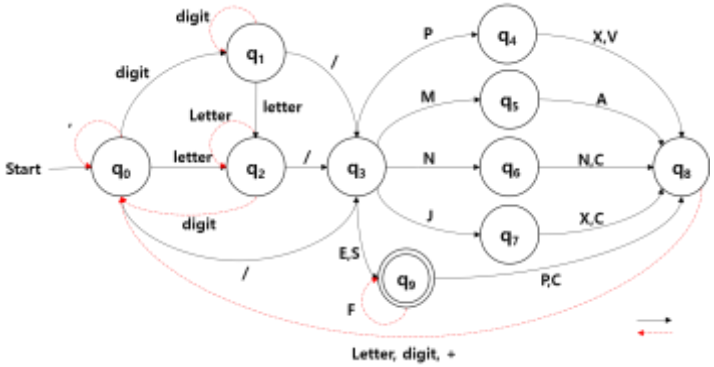

Figure 1. Lexical Analysis Algorithm

Copyright ⓒ 2017 SERSC

2.3. Hash Map

Hash by using a hash function is a function that outputs a hash value of a fixed way for receiving a message of arbitrary length. It ensures the integrity of the input and especially if the query is run frequently it can get quick retrieval effectiveness by using the Hash Map [2]. The map is managed by a key and a pair of value data and does not allow duplicate keys.

2.4. Latent Semantic Analysis

The mathematical foundation of LSA is the Vector Space Model (VSM), an algebraic model for representing documents as vectors in a space where dictionary terms are used as dimensions [3].

LSA is related to the analysis of the hidden meaning in a sentence. Deerwester *et al.* (1990) proposed an elegant way to improve similarity measurements with a mathematical operation on the term-document matrix, X based on Linear algebra. Deerwester, Landauer and Dumais (1997) focused on the document similarity than the word similarity. Document similarity is treated a main issue in Latent Semantic Indexing (LSI) while the words of similarity a main issue of LSA. They use a statistical method for the Vector Space Models (VSMs) based on Singular Value Decomposition (SVD) [4].

SVD can do diagonalization by utilizing eigenvalue decomposition to an orthogonal matrix $m \times n$, unlike the diagonalization of general square matrix, in a way that diagonalization as a Eigenvalue Decomposition method

SVD calculates X composed of the product of three matrices $U\Sigma V^T$, it is shown in equation (2).

$$X = U\Sigma V^T \tag{2}$$

$$U : m \times m \; Orthogonal \; matrix \; (AA^T = U(\Sigma\Sigma^T)U^T)$$
$$V : n \times n \; Orthogonal \; matrix \; (A^TA = V(\Sigma^T\Sigma)V^T)$$
$$\Sigma : m \times n \; Diagonal \; matrix$$

Wherein $U = AA^T$ and A^T means a transposed matrix of the matrix A and column vector of the eigenvectors x_i corresponding to the AA^T of eigenvalue λ_i and satisfying equation (3), Also A^TA means column vector of the eigenvectors x_i corresponding to the eigenvalue λ_i. And Σ is a diagonal matrix to A of eigenvalues as diagonal elements.

$$From \; Wx = \lambda x \; to \; (W - \lambda I)x = 0 \tag{3}$$
$$if \; W = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \; then$$
$$(W - \lambda I) = \begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix} = 0$$

Wherein, the similarity between the words is calculated as the similarity between the row of the $U \times \Sigma$ matrix, and similarity between the document use the $\Sigma \times V^T$ matrix. This research is analyzed by using the $\Sigma \times V^T$ matrix and the $U \times \Sigma$ matrix. The classification matrix was applied to the cosine similarity as shown in equation (4).

$$X = x_1, x_2, x_3, \ldots, x_n$$
$$Y = y_1, y_2, y_3, \ldots, y_n$$
$$\cos(X, Y) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{4}$$

2.5. Tf-idf Weight

Term Frequency (TF) represents the number of times the word appears, document Frequency (DF) represents the frequency of the article. Inverse document frequency (IDF) denotes the inverse document frequency. Equation of IDF is shown (5). In general, if the importance of a word to express a numerical value shall apply tf-idf.

Copyright ⓒ 2017 SERSC

$$idf_t = \log(\frac{N}{df_t}) \tag{5}$$

$$W_{t,d} = \log(1 + tf_{t,d}) \times \log(\frac{N}{df_t}) \tag{6}$$

3. Crime Data Analysis System Model

3.1. The Structure of the Proposed System

The proposed system in this paper is to extract the SNS data and unstructured data related to digital forensics and to provide the location and other related information of a criminal. From the overall architecture, Data Extractor and Algorithm module are shown in Figure 2 [1].

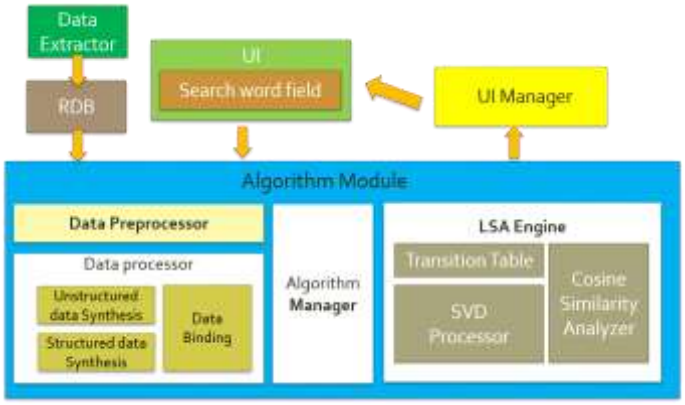

Figure 2. The Structure of the Proposed Model

This system consists of Data Extractor, RDB, UI, UI Manager, Algorithm module. And the propose system is divided into two phases.

Phase 1 is shown in Figure 3. Similarity among documents, this is to do document similarity between each, using the data stored in the Data Extractor when a new document is added.

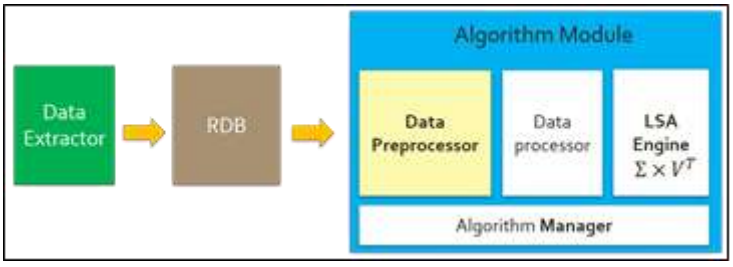

Figure 3. Similarity among Documents

Phase 2 is shown in Figure 4, Similarity between words and documents, this is to obtain a similarity between words and documents when a new word is added.

Copyright ⓒ 2017 SERSC

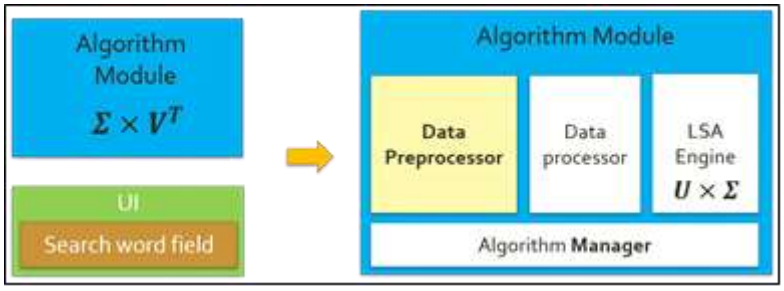

Figure 4. Similarity between Words and Documents

3.2. Data Extractor

This extracts the data in the pattern of Figure 1, in section 2, which is extracted by first using the Ko-NLP. The Algorithm is shown Figure 5.

Step	***Contents***
[Step 1]	*Final Ending divided into blocks(EF)*
[Step 2]	***Loop***
[Step 3]	***if*** *NP/NC+ Postposition Particle(Nominative) or NP/NC+ Postposition Particle(Objective) Output NP/NC (The Structured term of related Victims or Offenders)*
[Step 4]	***else*** *Output NP/NC (The Unstructured term of related Victims or Offenders)*
[Step 5]	***if*** *NP/NC followed by MA Output NP/NC (The sub-Unstructured term)*
[Step 6]	***until*** *PV/PX/PA comes up to* ***End Loop***

Figure 5. Similarity between Words and Documents

3.3. Algorithm Module

1) Algorithm Manager: When professional investigator -s and profilers input related data into the Search word field of the UI, Algorithm Manager sends the criminal data stored in the DB and the input data of the UI to preprocessor. When data processor completes entering and processing of the data, it delivers the classified data and UI input data to the LSA Engine.

Copyright ⓒ 2017 SERSC

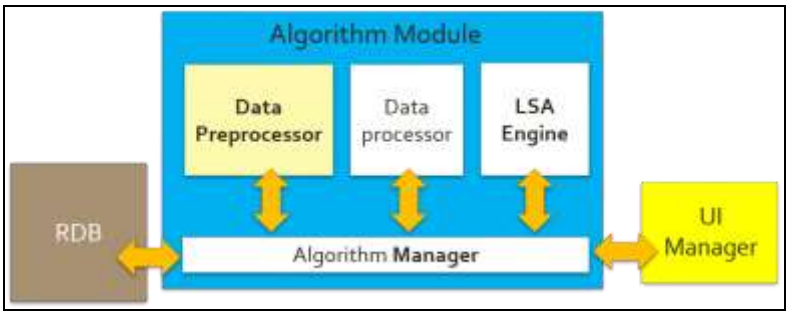

Figure 6. Algorithm Manager

2) Data Preprocessor: The data received from the DB is loaded into memory as the basic array type.

This system consists of unstructured data integration and structured data integration of Data processor integrated and connected as object-oriented structure. Binding Data combines the integrated and connected data to the vertical and horizontal, as structured data and unstructured data by using Hash Map.

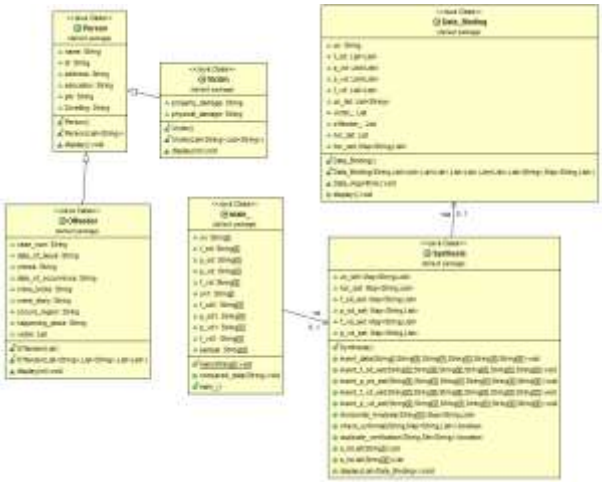

Figure 7. System Architecture

In Figure 7, the Victim and Offender inherit properties from Person class, and unstructured and structured data are in index structure as shown in Figure 8.

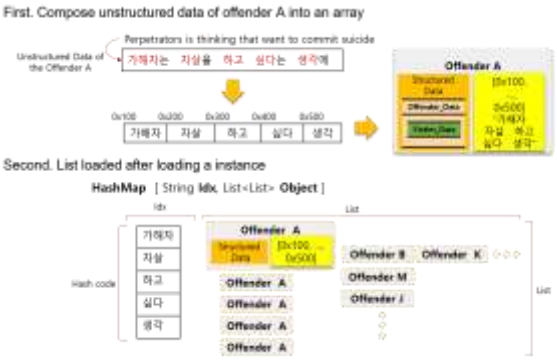

Figure 8. Structured·Unstructured Data Loading

Copyright ⓒ 2017 SERSC

3) LSA Engine: LSA Engine is sequentially recorded in Table Transition for the data input of UI by converting the matrix structure and the SVD Processor proceeds a diagonalization by the SVD Processor in accordance with the result conducted Cosine Similarity Analysis, determines the degree of similarity, and provides to the end user.

4. LSA-based Test System Implementation and Evaluation

4.1. LSA-Based Test System Implementation

We used eight incidents consisted of structured 131 words and 294 unstructured words based on the proposed model in this paper.

Figure 9. Example of Data Processor

Figure 9 shows the victim and offender's information when the user enters the "BangWha"(The meaning of words is arson). This can be found the motive that revenge for the penalty for no-show of reservists training in the form of unstructured data. The result is added into the last column of the Transition Table in addition to the weight [1].

		#1	#2	#3	#4	#5	#6	#7	#8	df
(36 years)	36세	0	0	0	0	0	1	1	0	2
(10 to 12:00)	10~12시	0	0	0	0	1	0	0	1	2
(October)	10월	0	0	0	0	0	1	1	0	2
(off-campus housing)	고시원	1	0	1	0	0	0	0	0	2
(pick)	곡괭이와		0	0	0	0	0	0	0	1
(Fracture)	골절	0	0	0	0	1	1	0	0	2
(Golf clubs)	골프채	0	0	0	0	0	0	0	1	1
(Fruit knife)	과도	1	0	0	0	0	0	0	0	1
(Professor)	교수로	0	0	0	0	1	0	0	0	1
(Guui-dong)	구기동	0	0	0	0	0	0	0	1	1
(Shoes)	구두	0	0	0	0	1	0	0	0	1
(Congress)	국회	1	0	0	0	0	0	0	0	1
(Precious metal)	귀금속은		0	0	0	0	0	0	1	0
(Coffers)	금고를	0	0	0	0	0	0	0	1	1

Figure 10. Example of Transition Table

In this paper, we calculate the SVD in Matlab, as shown in Figure 10, Figure 11, Figure 12, Figure 13 and Figure 14. Figure 15 shows the results of the Cosine Similarity.

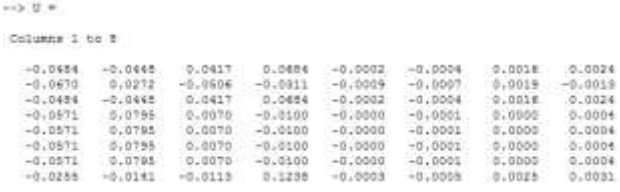

Figure 11. Example of U for SVD

Figure 12. Example of Σ for SVD

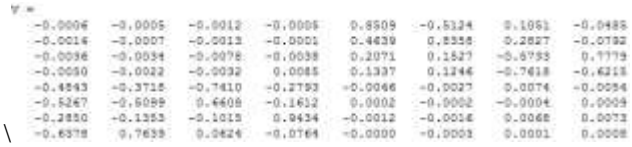

Figure 13. Example of V for SVD

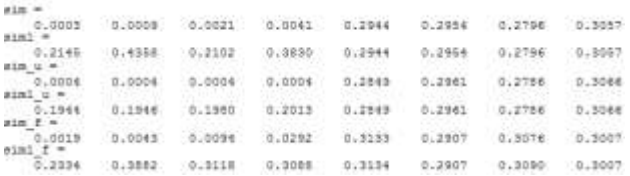

Figure 14. Example of $\Sigma \times V^T$ for SVD

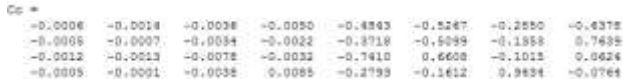

Figure 15. Example of Cosine Similarity

4.2. Evaluation

In this paper, we used Ko-NLP to extract structured and unstructured data management by storing the extracted data to the DB, and carried the LSA analysis based on the data stored in the Transition Table when retrieved data for investigators to enter. The final similarity analysis to implement an effective profiling will be enhanced by using clustering methods such as unstructured data as well as using only the existing structured data [1].

Figure 16 is applied structured data only, Figure 17 is the case of applying the structured data and unstructured data, shown in Figure 18 is the case of applying the structured data, the unstructured data and the both of structured data and unstructured data.

 Copyright ⓒ 2017 SERSC

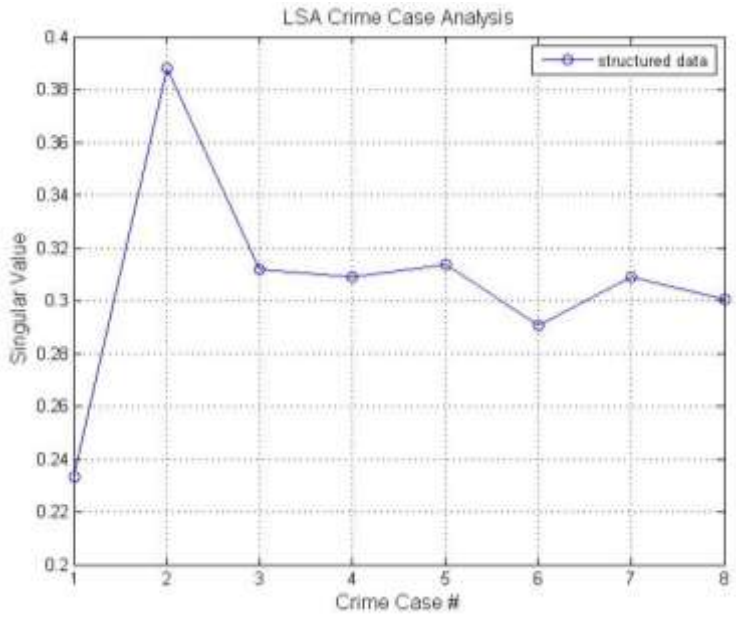

Figure 16. Using LSA Comparative Analysis of the Structured Data

In the case of Figure 16, we measured the Singular Value using only structured data. It is shown the three group (#1, #2, #3~#8).

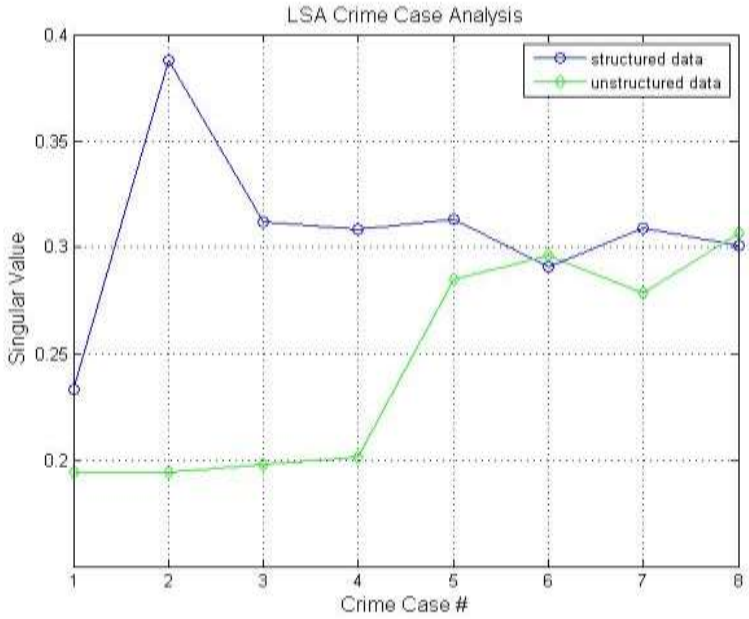

Figure 17. Using LSA Comparative Analysis of the Structured Data and the Unstructured Data

In the case of Figure 17, we measured the Singular Value using only unstructured data. It is shown the two group (#1~#4, #5~#8).

Copyright ⓒ 2017 SERSC

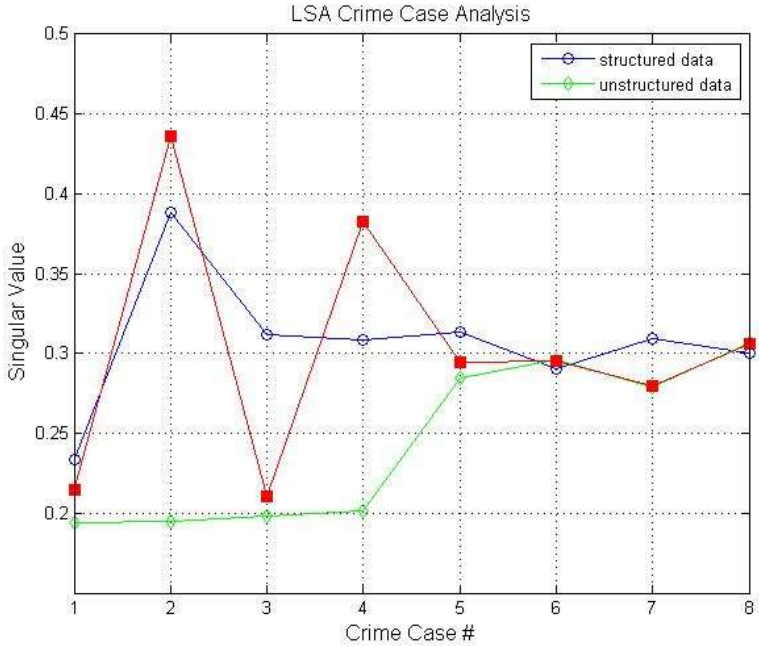

Figure 18. Using LSA Comparative Analysis of Structured·Unstructured Data, Unstructured Data, and Structured Data

Figure 18 is a comparison between composite method including unstructured data and only structured data. Event # 1 to # 4 is different injuries and general criminal homicide, and we enter the serial murders (the same offender) from Event # 5 to # 9. Case of structured and unstructured data shows that Event #1 to #4 are a different affinity and event # 5 to # 9 appears to be the same size. But if we apply only the structured data, it shows the failure to differentiate event # 3 to # 8. If we use unstructured data only, it is classified into two group events # 1 to # 4 and # 5 to #9 [1].

Figure 19 shows two words "SangHae(injury)" and "Gwigumsok(jewelry)", were reviewed from Figure 20. "SangHae" is applied to all data in structured events # 1 to # 4 and "Gwigumsok" is unstructured data to include only event # 6.

		#1	#2	#3	#4	#5	#6	#7	#8	df
(Shoes)	구두	0	0	0	0	1	0	0	0	1
(Precious metal)	귀금속은	0	0	0	0	0	0	0	1	0
(Samseong-dong)	삼성동	0	0	0	0	0	0	1	1	2
(Detriment)	상해	1	1	1	1	0	0	0	0	4
(Seoul)	서울	0	0	0	0	1	1	1	0	3

Figure 19. Data for Word Search

If we review only the structured data, it may not provide event # 6. In addition, if we apply only unstructured data, we can only provide event #6. If we, however, review of applying a structured data and unstructured data, the system can check for every case, event #1 to #4 and #6 [1].

Copyright ⓒ 2017 SERSC

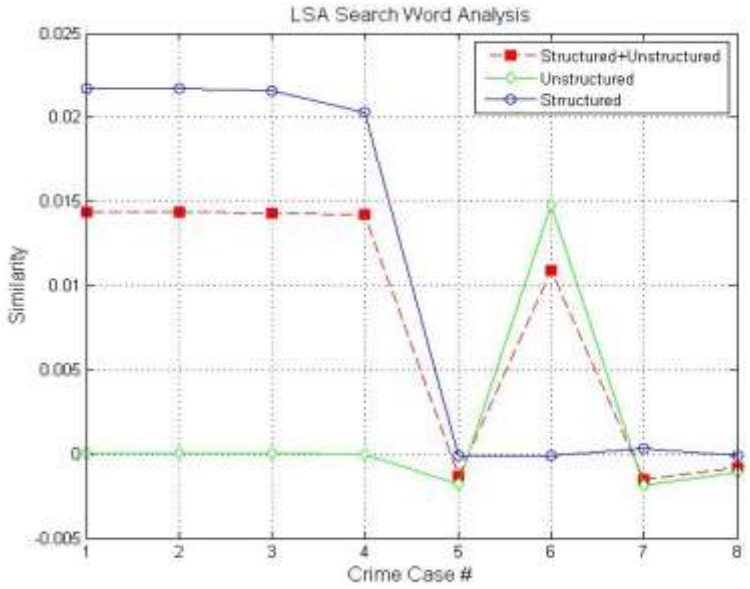

Figure 20. Compare to Find the Words (2 words)

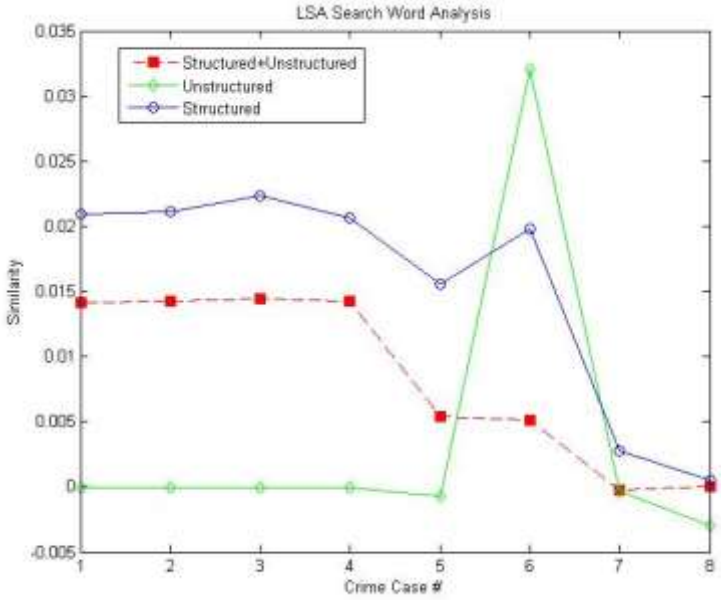

Figure 21. Compare to Find the Words (3 words)

This experiment searches for the three words "family", "blunt", "unemployed". "Blunt" is a word for a serial killer (#5~#8), "family" and "unemployed" is the word for the general crime (#1~#4). Although we do not completely separate the serial killer, we were able to check between serial killer and general crime in Figure 21.

In this paper, we used the matrix to the truncated SVD, and applied dimension (The means that the scale in the vector environment). The dimension in the experiment was 4. Although the higher dimension can increase the degree of match between the selected words and document, the weights could be reduced for the unselected word. We found that this disadvantage is the difficulty to provide associative word, and this is expected to be improved by utilizing machine learning [1].

Copyright ⓒ 2017 SERSC

5. Conclusion

The existing criminal analysis and statistics were based on the only structured data or a personal database of investigators, and a statistic data is written by using Boolean Search. This is a lot of lacking of information as a point of discussing the precise criminal investigation or criminal correct policy.

In this paper, we propose a system that can be applied to the statistical data and investigation data. This system responds to the demand of users and can be used in criminal investigations. It is because using the structured data and unstructured data. We were able to expand he system used in domestic Korea Information system of Criminal-justice Services (KICS) of the current [1]. These results can be of assistance to investigator or profilers need information to provide more accurate information. The LSA proposed in this paper is to calculate the affinity for the whole word as a Data Mining Techniques.

In the future, it plans to research by applying fuzzy theory to the analysis of unstructured data.

Acknowledgments

This paper is a revised and expanded version of a paper entitled Construction of Criminal Profiling System utilizing Structured and Unstructured Data presented in 2016 at The 12th International Conference on Multimedia Information Technology and Applications July4-6, 2016, Souphanouvong University (SU), Luang Prabang, Lao PDR [1].

References

[1] Y. Kim and M. Chung, "Construction of Criminal Profiling System utilizing Structured and Unstructured Data", Proceeding of the International Conference Multimedia Information Technology and Application (MITA 2016), Souphanouvong University (SU), Luang Parabang, Lao PDR, **(2016)**, pp 38-41.
[2] D. Willis, D. J. Pearce and J. Noble, "Efficient object querying for java", Journal of European Conference on Object-Oriented, **(2006)**, pp. 28-49.
[3] N. Evangelopoulos, "Latent semantic analysis", Wiley Interdisc. Rev, vol. 4, **(2013)**, pp. 683-692.
[4] P. D. Turney and P. Pantel, "2010 from frequency to meaning: Vector space models of semantics", Journal of Artificial Intelligence Research, **(2010)**, pp 159-164.
[5] T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsch, "Handbook of Latent Semantic Analysis", Lawrence Erlbaum Associates, Inc., **(2007)**.
[6] G. A. Saltron, "vector space model for automatic indexing", Commun ACM, **(1975)**, pp. 613-620.
[7] K. S. Choi, "HanNanum", KAIST GPL v3, **(1999)**.
[8] A. Gennady, "Thematic patterns in georeferenced tweets through space-time visual analytics", Computing in Science and Engineering, vol. 15, no. 3, **(2013)**, pp. 72-82.
[9] S. R. Boba, "Crime analysis with crime mapping", Sage, **(2012)**.
[10] S. Gerard, A. Wong and C. S. Yang, "A vector space model for automatic indexing", Communications of the ACM, vol. 18, no. 11, **(1975)**, pp. 613-620.
[11] H. J. Ho, S. Otgonchimeg and K. Seo, "Advanced metering infrastructure design and test bed experiment using intelligent agents: focusing on the PLC network base technology for Smart Grid system", The Journal of Supercomputing, vol. 72, no. 5, **(2016)**, pp. 1862-1877.
[12] F. David and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment", Natural Language Engineering, vol. 10, no 3-4, **(2004)**, pp. 327-348.
[13] G. Andrienko, "Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics", Computing in Science & Engineering, vol. 15, no. 3, **(2013)**, pp. 72-82.
[14] Microsoft, "Microsoft and NYPD Announce Partnership Providing Real-Time Counterterrorism Solution Globally", press release, **(2012)**.
[15] W. Petherick, "Serial crime: Theoretical and practical issues in behavioral profiling", 2nd Edition California, Elsevier, **(2006)**.
[16] S. Sukhpal, and I. Chana, "Introducing agility in cloud based software development through ASD", International Journal of u-and e-Service, Science and Technology, vol. 6, no. 5, **(2013)**, pp. 191-202.

 Copyright ⓒ 2017 SERSC

[17] J. H. Huh and K. Seo, "Design and test bed experiments of server operation system using virtualization technology", Human-centric Computing and Information Sciences, Springer, Berlin Heidelberg, vol. 6, no. 1, **(2016)**, pp. 1-21.
[18] T. Nguyen, T. Phuong and N. T. Thang, "Environmental economic load dispatch with quadratic fuel cost function using cuckoo search algorithm", International Journal of u-and e-Service, Science and Technology, vol. 7, no. 2, **(2014)**, pp. 199-210.

Authors

Yonghoon Kim, received Diploma in School of Tax Accounting, Busan Kyungsang College, Busan, Republic of Korea in Feb. 1998. Received B.S. in Engineering Degree from Department of Computer Information and Communication, Seoul Digital University, Seoul, Republic of Korea in Feb. 2014. Student in Master Course Major of Computer Engineering, Graduate School, Pukyong National University, Busan, Republic of Korea. His research interests are intelligent agent, computer security for application, and context aware computing.

Mokdong Chung, received B.S. in Engineering Degree from Department of Major of Computer Engineering, Kyungpook National University, Daegu, Republic of Korea in Feb. 1981. M.S. in Engineering Degree from Department of Computer Engineering from Seoul National University in 1983. Received the Ph.D. Degree in Computer Engineering from Seoul National University in 1990. He was a professor at Pusan University of Foreign Studies from 1985 to 1996. And he has been a professor at Pukyong National University since 1996. His research interests are OOP technology, computer security for application, intelligent agent, and context aware computing, He is a member of IEEE, KIISE, KIPS, KIISC, and KMMS.

Copyright ⓒ 2017 SERSC

 Copyright ⓒ 2017 SERSC