International Journal of Bio-Science and Bio-Technology
Vol.6, No.2 (2014), pp.117-128
http://dx.doi.org/10.14257/ijbsbt.2014.6.2.12

A Summary and Recommendation System based on Bio-Text Analysis

Ki-Young Lee[1], Jeong-Jin Kang[2], Ji-Won Yoon[1], Yong-Gyu Jung[1,*], Gyoo-Seok Choi[3], Sang-Bong Park[4], Eun-Young Kang[5]

[1]Department of Medical IT and Marketing, Eulji University, Seongnam, Korea
(kylee@eulji.ac.kr, yjw8623@nate.com, ygjung@eulji.ac.kr)
[2]Department of Information and Comminication, Dong Seoul University, Seongnam, Korea
jjkang@du.ac.kr
[3]Department of Computer Science, ChungWoon University, Incheon, Korea
lionel@chungwoon.ac.kr
[4]Department of Information and Communication, Semyung University, Jecheon, Korea
psbcom@semyung.ac.kr
[5]Department of Information and Communication, Dongyang Mirae University, Seoul, Korea
eykang@dongyang.ac.kr

**Corresponding Author: ygjung@eulji.ac.kr*

Abstract

Text mining is a technique to find meaningful information from unstructured text data. In this paper, the study is processed about the structure of the summary system centered around the keyword of bio-related text by using various a part of speech DB such as concept word, relationship word, and conjunctions. Also, it compares one or more documents, find the concept of common. To allow more accurate summary, it updates the DB continuously by forming the ontology about concept of common. This system can be used to recommend the books and other papers through extracted keywords. As a result, it gives a detailed research and convenience on bio technology.

Keywords: *Bio-Text Mining, Apriori Algorithm, TF-IDF, Speech DB*

1. Introduction

Currently, since its rapid growth rate and vast knowledge in the bio technology field, the need to gather information that is distributed in a text and a database by using a automated method have increased. Thus, many studies that extract specific information shown in literature by Text Mining technology have been processed [1, 2].

However, the current study only has proceeded the relation extraction between the objects by analyzing the related text. Therefore, making an accurate summary of the text is too difficult. It is also inappropriate to compare and analyze more than one text because only one text can be analyzed [3-5].

ISSN: 2233-7849 IJBSBT

In this paper, the study is processed about the structure of the summary system centered around the keyword of bio-related text by using various part of speech DB such as concept word, relationship word, and conjunctions. Also, it compares one or more documents, find the concept of common. To allow more accurate summary, it updates the DB continuously by forming the ontology about concept of common. This system can be used to recommend books and other papers through extracted keywords. As a result, it gives detailed research and convenience on bio technology.

In this paper, related research is described in Chapter 2, proposition system in Chapter 3, system implementation in Chapter 4, performance evaluation in Chapter 5 and conclusion in Chapter 6.

2. Related Research

2.1. Bio-Text Mining

Bio-text mining goes through the process Automatic analysis, Named entity recognition and Relation extraction. First of all, Automatic analysis process parses sentence using a word class attacher, a base phrase recognition and a sentence analyzer [6]. Named entity recognition is a process that understands term built in the database. Finally, after recognizing the named entity, relation extraction creates the pairs of objects.

In this paper, by using the large DB related bio, the accuracy of bio text mining increased. Through Named Entity Recognition Algorithm (word extraction algorithm) using PgSQL, information of bio is extracted.

2.2. TF-IDF Algorithm

To extract keyword of bio-related text, it can be judged by considering the frequency of words in the document. Further, in order to find a keyword in common for multiple documents, investigate the word frequency in a document set as a whole, then it is necessary to select a word which is infrequent. This is because, if the frequency is high, the possibility of auxiliary word to support the concept word is high, so it's hard to see as a keyword.

Technique used in the study of word frequency is TF-IDF. It is used to extract a keyword of the document through the weighting of words contained in the document [7,8]. The TF (word frequency) is a frequency value of the words that appear in the document. If the value of TF is high, that word in the document has high importance. The DF (document frequency) indicates the frequency of words that appear in the set of documents. It is used to investigate the word frequency against multiple documents. If the DF value is higher, that word is a word that is usually used in the set of documents. The word that is usually used in a full set of documents have a small chance of being a keyword. Therefore, the keyword is determined when the IDF (the value of the reciprocal of DF, Inverse Document Frequency) is high.

2.3. Apriori Algorithm

Association rules can find out if the relationship between the items are forming. It can calculate the relationship by using the concept of support and confidence [9].

Table 1. Examples of association rules

Consumer	Purchase items			
1	Soap		Cookie	
2	Soap	Clothes		
3			Cookie	
4	Soap	Clothes		Cap
5	Soap	Clothes	Cookie	Cap

Table 1 shows an example of the association rules. It is data divided by the buyer of items purchased. For example, let's assume that People buying a soap is a high possibility of buying cookie. Since support is a percentage of people who bought the cookie and soap in all transactions, it is the number of people who bought the cookie and soap divided by number of all transactions which equals 40%. Confidence is a percentage of people who bought cookies from the people who bought soap. It is the number of people who bought the cookie and soap divided by number of people who bought soap which equals 50%. As a result, the support is that 40 percent of all transactions purchase at the same time the cookie and soap. Also, Confidence is that 50 percent of people who bought the soap buys cookie.

Apriori Algorithm is a method that Min-Support and Min-Confidence compare with the support and confidence which is obtained. A set of data is not a formed relationship when it does not satisfy the value more than value of Min-Support and Min-Confidence. Therefore, it is deleted. If this qualification satisfies, a set of data is called a formed relationship.

In this paper, after finding the frequency of word in bio document by using the TF-IDF algorithm, it builds data by calculating the support and the confidence between words by applying to the association rules. To extend the ontology about concept word continuously, it can be added to the existing ontology to the pair of words that satisfies the value of Min-support and Min-confidence of Apriori algorithm.

3. Proposed System

3.1. System Architecture

This program provides systems such as the summary system, the recommendation system of the other texts, and the composition of ontology for the convenience of the user. Figure 1 is a content obtained by diagramming the structure of the system.

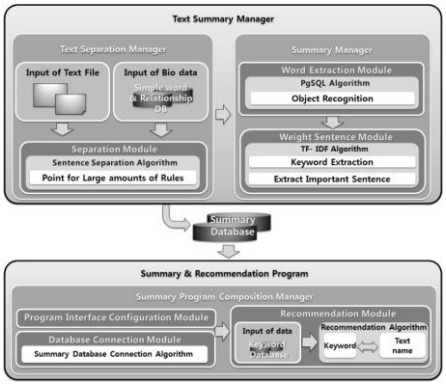

Figure 1. System Architecture

The overall system is done through Text Separation Manager, Summary & Recommendation Program. After a Text File inserted to the program, Text Summary Manager divides into each sentences in the Text File by using the separation module. After separation in the sentences, it extracts objects by comparing Bio-DB with sentences by PgSQL algorithm of Word Extraction Module[10]. And delete words that do not exist in the DB. Weight Sentence Module selects a keyword by investigating the frequency of word by using TF-IDF algorithm to target the words that remain.

It expects a relationship between words by using Apriori Algorithm by this keyword and adds it to the ontology which exists. Extract Important Sentence extracts important sentences when the frequency of keyword is high and removes the sentence that the weight is low. It makes up a summary database by saving important sentences.

In the Summary & Recommendation Program, Program Interface Configuration Module gives convenient interface for the user. Database Connection Module receives the information by connecting to the Summary Database.

By comparing the text name and information of the extracted keywords in the recommended module(keyword database is built), the program recommend another text.

3.2. System Flow Chart

Figure 2 is a content obtained by diagramming the flow of the entire system of Figure 1.

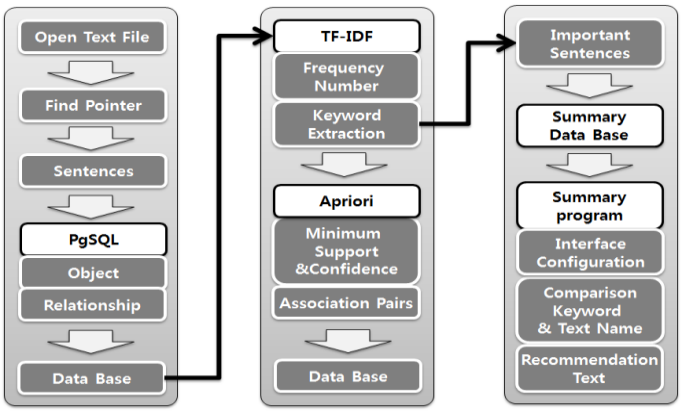

Figure 2. System Flow Chart

In the system flow chart, after importing the text file, it divides the sentence by finding the part with the pointer in the text. After distributing an object through a PgSQL, save it to the database. By implementing the TF-IDF to a database that was previously saved, extract the keyword of having the most frequency. The extracted keyword is applied in the Apriori Algorithm, then extracts the pairs of word, and saves to database. After extracting the keyword, distinguish the important sentences having the keyword, and then save it to the summary database. After constructing with a summary program by using summary database, recommend other texts that focuses on the keyword.

3.3. Apriori Algorithm

Figure 3 is a picture of Apriori Algorithm.

```
D : database over the set of items A,
P : the set of frequent itemsets
k = 1; Ck = A
while Ck ≠ 0 do
support_count(D, Ck)
for all candidates c ∈ Ck do
if c.support >= minsup then
Pk = c
end if
end for
Ck+1 = candidate_generation(Pk)
k = k+1
end while
```

Figure 3. Applying Apriori Algorithm

Through Apriori Algorithm, the association relationships between concept words can be determined. Also these association relationships configure the ontology [11]. Ontology is built by DB and it can have a variety of uses. Further, the user is able to learn not only the user's text via the ontology but also new information.

4. System Implementation

This system is a configured based on the DB associated with the genome in the Java language. It was implemented in accordance with the system flow chart of Figure 2.

Figure 4 is a screen that imports a text file that the user wants to summarize. Then the program gets the text from a text file imported.

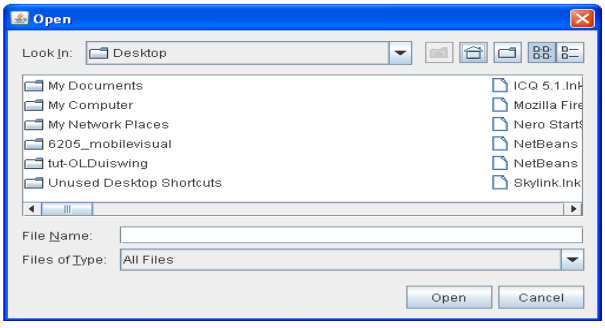

Figure 4. Screen Loaded File

Through the genome DB, the word is extracted from the imported text. Frequency is investigated with the word that has been extracted by the frequency investigation. Apriori Algorithm shown in Figure 4 is performed. In addition, the summary is performed around the keyword.

Figure 5 is the screen that has been processed summary around the keyword.

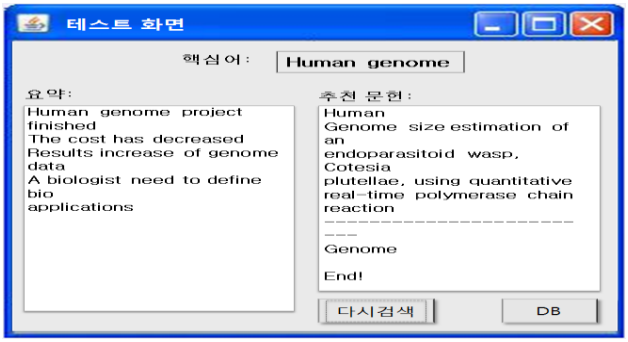

Figure 5. Screen Completed Summary

On the left text window, the summary was made. On the right, text window recommended texts about the keyword was made. In this example, the obtained keyword was human and genome. Human occupied the support rating of 73% of the sentence and genome occupied the support rating of 65% of the sentence. In multiple tests, when two or three keywords are extracted, it is found that most accurate summary is possible. The summary was processed by the sentence containing the most keywords.

Figure 6 shows the output of ontology DB of genome.

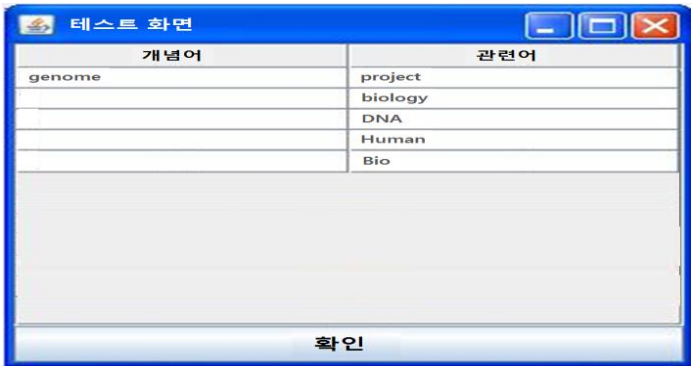

Figure 6. Ontology DB for Genome

Table 2 is the obtainments of the confidence in the genome and other objects.

Table 2. Confidence in the genome and other objects

A pair of relationship	*Confidence*
(genome, project)	10/16 = 0.625
(genome, biology)	9/16 = 0.563
(genome, DNA)	14/16 = 0.625
(genome, human)	10/16 = 0.625
(genome, bio)	13/16 = 0.813
(genome, biologist)	3/16 = 0.188

In the test program, it was based on 60% min-support and 55% min-confidence. If a set of data is below the standard removed it. As a result, the ontology DB in Figure 6 can be seen as DB of data related to the genome.

5. Performance Evaluation

This program implemented the performance evaluation with summary accuracy and configuration ontology. Hardware specifications of the experimental environment used are Intel (R) Core (TM) i3 CPU, 2.93GHz, 4GB RAM, and Samsung Electronics. The operating system was Windows7.

The experiment was performed with data as paper of "A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going". Experiment proceeded over three times before evaluating the summary accuracy and the configuration ontology.

Table 3 shows the test results of the summary accuracy and the configuration ontology of existing bio-related programs.

Table 3. Existing Program Related Bio

Division	First	Second	Third	Fourth	Average
Summary Accuracy	56.6	54.1	55.4	53.0	54.8
Configuration Ontology	39.5	41.0	40.2	42.0	40.7

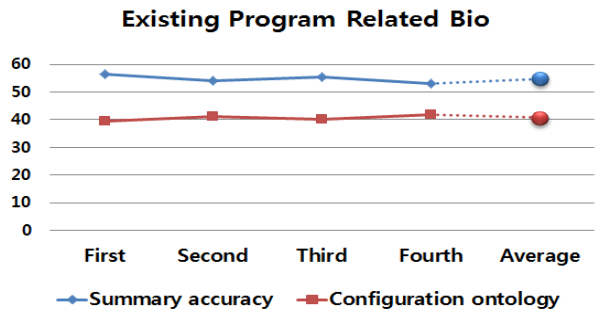

Figure 7. Existing Program Related Bio

Until the third trial, summary accuracy recorded 55.0% average and configuration ontology recorded 40.2% average. Since the configuration of the ontology is carried out without comparing several papers, it can be seen as a decrease.

Table 4 shows the test results of the summary accuracy and the configuration ontology of the proposed program.

Table 4. Proposed Program

Division	First	Second	Third	Fourth	Average
Summary Accuracy	57.9	59.0	58.2	59.7	58.7
Configuration Ontology	41.1	44.3	43.2	45.0	43.4

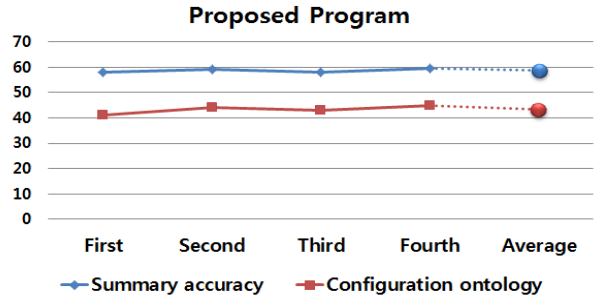

Figure 8. Proposed Program

Summary accuracy recorded 58.3% average and configuration ontology recorded 42.9% average. Results of comparing summary accuracy and configuration ontology of the proposed program with those of existing program is that the proposed program has a higher value than the existing program.

6. Conclusion

By improving the disadvantages of the existing bio program that do not consider the user, users can save time to analyze literature and learn new information from the proposed system.

In future research, I want to be able to make more exact summary by building the DB of the concept word, relationship word, conjunctions, and postposition. In addition, research on determining the standard of min-support and min-confidence of Apriori Algorithm is required.

Acknowledgements

This article is a revised and expanded version of a paper entitled [A Study on the Efficient Bio Summary System using Text Mining] presented at International Symposium on Advanced and Applied Convergence held on November 14-16, 2013 at Seoul, Korea.

References

[1] H. J. Lee, J. C. Park, "Probabilistic Filtering for a Biological Knowledge Discovery System with Text Mining and Automatic Inference," Journal of Korea Society of Computer & Information(JKSCI), vol. 17, no. 2, **(2012)**, pp. 139-147.
[2] K. Y. Lee, J. J. Kang, J. W. Yoon, Y. G. Jung, G. S. Choi, S. B. Park, and E. Y. Kang, "A Study on the Efficient Bio Summary System using Text Mining," Advanced and Applied Convergence Letters(AACL), vol. 1, **(2013)**, pp. 7-8.
[3] H. J. Kim and J. Y. Chang, "Discovering News Keyword Associations Using Association Rule Mining," Journal of the Institute of Internet, Broadcasting and Communication (JIIBC), vol. 11, no. 6, **(2011)**, pp. 63-71.
[4] R. D. Caytiles and H.J. Kim, "U-Learning: An Interactive Social Learning Model", International Journal of Internet, Broadcasting and Communication (IJIBC), vol. 5, no. 1, **(2013)**, pp. 9-13.
[5] K. J. Park, J. H. Kwon and Y. K. Kim, "Design of Hard Partition-based Non-Fuzzy Neural Networks," International Journal of Advanced Smart Convergence (IJASC), vol. 1, no. 2, **(2012)**, pp.30-33.
[6] K. M. Park and K. B. Hwang, "A Bio-Text Mining System Based on Natural Language Processing," Journal of Computing Science and Engineering (JCSE), vol. 17, no. 4, **(2011)**, pp. 205-213.
[7] G. S. Go, W. K. Jung, Y. G. Shin, S. S. Park and D. S. Jang, "A Study on Development of Patent Information Retrieval Using Textmining," Journal of the Korea Academia-Industrial Cooperation Society (JKAICS), vol. 12, no. 8, **(2011)**, pp. 3677-3688.
[8] http: // en.wikipedia.org/wiki/Vector_space_model.
[9] Y. Kim, "A Study on Design and Implementation of Personalized Information Recommendation System based on Apriori Algorithm," Journal of the Korean BIBLIA Society for library and Information Science, vol. 23, no. 4, **(2012)**, pp. 283-308.
[10] S. Y. Sim, H. M. Kang and Y. H. Lee, "Access to Databases through the R-Language," Journal of the Korean Statistical Society (JKSS), vol. 15, no. 1, **(2008)**, pp. 51-64.
[11] D. Wang, *et al.*, "Association Rules Mining on Concept Lattice using Domain Knowedge," in Proc. 1st International Conference on Ma-chine Learning and Cybermetrics, vol. 4, **(2005)**, pp. 2151-2154.

Authors

Ki-Young Lee

He received his B.S. degree in Computer Science at Soongsil University in 1984. In 1988 and 2005, he received M.S. and Ph.D. degrees in Databases at Konkuk University, respectively. From 1984 until 1991, he worked for Korea Institute of Ocean Science & Technology (KIOST) as a researcher in Data Information & Processing department. He is currently a professor at the department of Medical IT and Marketing at Eulji University. He is also the head of department of S/W development in Bio-Meditech Regional Innovation Center at Eulji University. He is the director of the Korea Institute of Internet, Broadcasting and Communication (IIBC), and the director of the Korea Electronics Engineers (IEEK). His research interests include spatial databases, geographic information systems (GIS), location-based services (LBS), u-Healthcare, ubiquitous sensor network (USN), moving objects databases, and telematics, *etc.*

Jeong-Jin Kang

He is currently the faculty of the Department of Information and Communication at Dong Seoul University in SeongNam, Korea since 1991, and currently the President of the Korea Institute of Internet, Broadcasting & Communication (IIBC). During 3 years from Feb. 2007 to Feb. 2010, he worked as a Visiting Professor at the Department of Electrical and Computer Engineering, The Michigan State University. He was a lecturer of the Department of Electronic Engineering at (Under)Graduate School(1991-2005), The Konkuk University. Dr. Kang is a member of the IEEE Antennas and Propagation Society(IEEE AP-S), the IEEE Microwave Theory and Techniques Society (IEEE MTT-S), and a member of the Korea Institute of Internet, Broadcasting & Communication(IIBC), Korea. His research interests involve Smart Mobile Electronics, RF Mobile Communication, Smart Convergence of Science and Technology, RFID/USN, u-Healthcare and ultrafast microwave photonics, as well as GIS, LBS, moving objects databases, and telematics, *etc.*

Ji-Won Yoon

He is currently a student in the department of Medical IT and Marketing at Eulji University. Her research interests include spatial databases, embedded systems, location-based services (LBS), u-Healthcare, ubiquitous sensor network (USN), moving objects databases, and telematics, *etc.*

Yong-Gyu Jung

He received the B.S. in Physics from Seoul National University in 1981. And then he got the M.S. and ph.D. degree of Computer Science from Yonsei and Kyonggi University in 1994 and 2003 respectively. Since 1999, he has been a Faculty of Medical IT marketing Dept. in Eulji University. His research interests are in the areas of medical information analysis of Baysian Networks Based Data Mining for Clinical Assisted Reproduct Technology. He has been a leading member and Head of Delegation at UN/CEFACT and ISO/TC154 with National Body of Standard Development Organization.

Gyoo-Seok Choi

He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Yonsei University, Seoul Korea, in 1982, 1987, and 1997, respectively. He worked at the laboratory of DACOM Company as a researcher from 1987 to 1990. He also worked at the laboratory of SK Telecom Company as a senior researcher from 1991 to 1996. He is currently a professor at the Dept. of Computer

Science in Chungwoon University. He is a vice-president of the Korea Institute of Internet, Broadcasting & Communication(IIBC). His current research interests include Artificial Intelligence, Telematics, Mobile Computing, *etc.*

Sang-Bong Park received the M.S. degree in Electronic Engineering from Korea University in 1987. He received the Ph.D. in Electronic Engineering from Korea University in 1992. He managed ASIC team that designed the embedded 16M DRAM for graphic controller in Samsung Electronics. He is currently the faculty of the Department of Information and Communication at Semyung University in Jecheon, Korea since 1991, and also with ATLab that develops touch screen controller and optical mouse. His research interests include digital IC design with strong emphasis in mobile product such as touch screen controller.

Eun-Young Kang

She received her B.S. degree in Computer Science at Sookmyeong Women's University in 1987. In 1999 and 2009, she received M.S. and Ph.D. degrees at Sungkyunkwan University, respectively. From 1987 until 2002, she worked for Kyobo as a researcher in Computer Information department. She is currently a professor at the department of Information & Communication at Dongyang Mirae University. She is the director of the Korea Institute of Internet, Broadcasting & Communication(IIBC). Her research interests include u-Healthcare , mobile computing, *etc.*