A bio-text mining system using keywords and patterns in a grid environment

Hyuk-Ryul Kwon1, Tae-Sung Jung2, Kyoung-Ran Kim3,
Hye-Kyoung Jahng3, Wan-Sup Cho3, and Jae-Soo Yoo4
[1]*Dept. of Bio-Informatics, Chungbuk National University,*
361-763 Cheongju, Chungbuk, Korea
Tel: 043-276-3258, Fax: 043-266-8865, E-mail:Khl80@naver.com

[2]*Dept. of Information Industrial Engineering, Chungbuk National University,*
361-763 Cheongju, Chungbuk, Korea
Tel: 043-276-3258, Fax: 043-266-8865, E-mail : E-mail :{mispro,
tkkim}@chungbuk.ac.kr

[3]*Dept. of Management Information Systems, Chungbuk National University,*
361-763 Cheongju, Chungbuk, Korea
Tel: 043-276-3258, Fax: 043-266-8865, E-mail:{ lodestone, shira07,
wscho }@chungbuk.ac.kr

[4]*Dept. of Computer and Communication Engineering, Chungbuk National University,*
361-763 Cheongju, Chungbuk, Korea
E-mail:yjs@chungbuk.ac.kr

Abstract: *As huge amount of literature including biological data is being generated after post genome era, it becomes difficult for researcher to find useful knowledge from the biological databases. Bio-text mining and related natural language processing technique are the key issues in the intelligent knowledge retrieval from the biological databases. We propose a bio-text mining technique for the biologists who find knowledge from the huge literature. At first, web robot is used to extract and transform related literature from remote databases. To improve retrieval speed, we generate an inverted file for keywords in the literature. Then, text mining system is used for extracting given knowledge patterns and keywords. Finally, we construct a grid computing environment to guarantee processing speed in the text mining even for huge literature databases. In the real experiment for 10,000 bio-literatures, the system shows 95% precision and 98% recall.*
Keywords: *Biological database, grid computing, bio-text mining*

1. INTRODUCTION

After Human Genome Project, bio-data of various forms such as sequences are created in large quantities [21]. Furthermore, literatures in the bio-related areas are published constantly. It has been gradually difficult for scholars to search desired information in huge data. Now, importance of the text-mining technique for the bio-related documents is increasing remarkably because of analogizing the significant information rapidly and sophisticatedly. PUBMED maintains about 15,000,000 documents and updates a dozens of documents in a day [21]. If a scientist examines these data one by one, it takes a lot of time and cost to process them. Usually, it is well-known that an ordinary person can read about 60 pages in an hour. It means that

it takes about 285 years to read 15,000,000 documents completely. In this paper, we propose the bio-text mining technique for efficiently retrieving the significant data from the huge bio-data efficiently.

2. Related Work

2.1 Trend of Bio-Text Mining

As huge data are published in the biology area, many people are recognizing the importance of text-mining and developing text-mining tools based on the natural language processing. There are several typical text-mining systems. MedStract[13] pulls out protein-protein interaction information and GENIES[05] extracts molecular pathway from the biology literatures. BIOBIBLIOMETRICS retrieve the related information from the large biology literatures based on the gene name. Two-Hybrid system from Ito and Uetz finds protein-protein interactions related to the germinative yeast [09, 17]. Two-hybrid system has put out protein interaction. Agrawl, Satou and et. al. have extracted the association rules among heterogeneous information of the sequences, protein structure, and their functions with the data mining technique[01, 15, 16].

There are remarkable characteristics in MedStract, GENIES and BIOBIBLIOMETRICS. MedStract recognizes the object names and part of speeches based on UMLS thesaurus dictionary by considering the characteristic keyword in the bio-literatures [06, 07, 12]. And then, it identifies the noun and verb phrases by applying the automata in stages and continues to extract interaction information according to the defined patterns.

MedStract visualizes the results from the keyword searches as the forms of the graphs and tables.

GENIES consists of the term tagger, the preprocessor, the parser, and error recoverer. Term tagger identifies gene names and the preprocessor determines sentences, phrases, and worlds. The parser recognizes the interaction relationships with the restriction rules and semantic patterns. The error recoverer processes the errors of the sentence analysis with several heuristic. GENIES constructs the sophisticated pathway based on extracted interactions .

BIOBIBLOMETRICS understands that the documents are closely related if obtaining the similarity from the coexistence between two genes and it exceeding the critical value.

Actually, if researcher associated with biology search special genes, another gene related with them was searched and visualized. It is helpful to study a relation gene and gene from a result.

All programs uses text-mining but are solving the problems through other methodology and access. Now, many scientists are studying text-mining, will make an effort to search fast and exacted.

2.2 Limitation of the Related Work

Existing text-mining tools can not accommodate the large quantity of the data and extract information about interesting keyword and pattern. It is difficult to collect the life science thesis once consisting of 15,000,000 data. If a man should search websites and collect, store, and group a thesis, these works need a lot of time, manpower and have low accuracy. This work spends a lot of time to collect, store, and group the web sites by oneself and low accuracy. To do it, we should make use of a tool like the web agent for users to get specific information from the collected data. Text-mining tools with keyword/pattern system and inverted index system would be good alternatives

for it. We suggest how to advance mining speed which is a weak point of existing text-mining systems using the natural language processing.

3. Architecture of Proposed Text-mining System

SBML We proposed the text-mining system like Fig.1. It is consisted of 5 components. First, web-robot collects a lot of the documents from the web sites. Second, parser refines the collected documents. Third, a database stores them. Fourth, grid computing system is used to speedup the query processing. Finally, searching system retrieves information which users want to find with the keywords or patterns.

The procedures of <Fig.1> are like follows.

Step 1: Data collection

① Web robot collects biological literatures and they are stored as forms of the files in the local place.

② Parser loads biological documents collected from web sites into the memory.

③ The Parser refines or converts biology documents and stores them in the database.

Step 2: The search using text-mining

④ Input a keyword/pattern which users want to find.

⑤ Use the inverted index to search existing keywords.

⑥ Utilize the computing grid to speedup query processing for the huge inverted index.

⑦ Extract the candidate sentences and literature numbers matching with keywords from the database.

⑧ When the keyword and pattern exist inside one sentence, it search abstract thesis name and sentence including keyword pattern using grid-computing

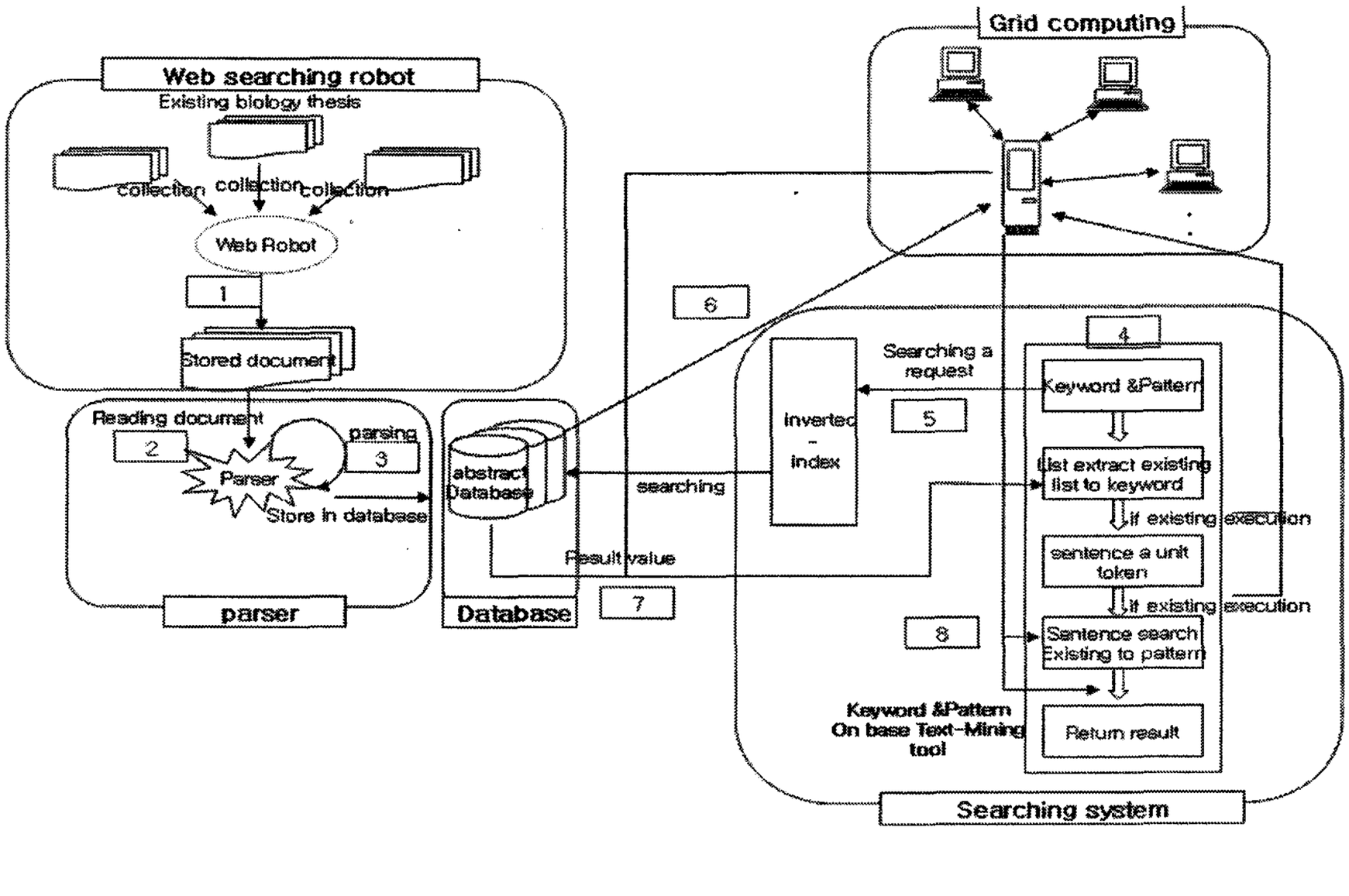

Figure 1. Architecture of the Proposed Text-Ming System

4. Evaluation

Experiment environment of the proposed text-mining system is like Table.1.

Table 1. System Test Environment

All document number	Document number all contain keyword & pattern	Keyword number	Pattern number
10,000 documents	100 documents	12 number	7 number

The number of documented experiment is 10000 data. 100 documents of them contain 12 user inputting keywords and 7 patterns. The essential points in this experiment are processing time, its accuracy, and reappearance ratio.

Table 2. - Test Result

Accuracy	Reappearance radio	Process time
95%	98%	34 seconds

Experimental result, reappearance ratio, and processing time are 95%, 98%, and 34 seconds respectively. The results are very remarkable and promising. Proposed text-mining system is developed for the very large data processing. To do this, we adopted grid-computing system.

5. Conclusion and Future Work

As huge amount of the biological literatures are being generated, it is difficult for researchers to find the useful knowledge from them. Bio-text mining systems based on natural language processing can be good alternatives for the intelligent knowledge retrieval. We proposed a bio-text mining system which finds significant knowledge from the huge literatures. The proposed text-mining system constructs the inverted index consisting of keyword & pattern to speedup. The text-mining system has 95% accuracy and 98% reappearance ratio for 10,000 data. In future, we will provide the web services for all users to use text-mining system widely.

Acknowledgments

This work was supported by the Chungbuk BIT Research-Oriented University Consortium and the 2nd Brain Korea Project.

References:

[1] Agrawal, R. et al., "Mining association rules between sets of items in large databases," Proceedings of ACM SIGMOD, pp. 207-216, 1993.
[2] BackwellPublising,http://www.blackwellsynergy.com/?cookieSet=1
[3] B.Stapley and G.Benoit, "BIOBIBLIOMETRICS : Information Retrieval and Visualization form Co-occurrences of Gene Name in MEDLINE Abstracts," PSB 2000.
[4] BIOBIBLIOMETRICS, http://bbm.icnet.uk/~stapleyb/biobib
[5] C.Friedman et al., "GENIES : A Natural language Processing System for the Extraction of Molecular Pathways form Journal Artcles," Bioinformatics 2001.
[6] D.Hindle, "DeteMinistic parsing of sysntactic non-fluencies," In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, 1983.
[7] D.McDonald "Robust partial parsing through incremetal multi-algorithm processing," In P.Jacobs, editor, Text-based Inteligent Systems, 1992.
[8] I.Foster, C.Kesselman, and s.Tuecke, "The Anatomy of the Grid : Enabling Scalable Virtual Organizations," International J.Supercomputer Application, 15(3), 2001.

[9] Ito, T. el al., "A comprehensive analysis to explore the yeast protein interactome," Preceedings of Natl Acad. Sci. USA, pp.4569-4574, 1998.
[10] Ito, T. el al., "A comprehensive analysis to explore the yeast protein interactome," Preceedings of Natl Acad. Sci. USA, pp.4569-4574, 1998.
[11] J. Pustejovsky et al., "MedStract: Creating large-scale information servers for biomedical libraries," Proceedings of the Association for Computation Linguistics(the Workshop on Natural Language Processing in the Biomedical Domain), pp.85-92, 2002
[12] J. Pustejovsky et al., "Semantic indexing and typed hyperlingking," In AAAi Symposium on Language and the Web,Stanford, CA, 1997.
[13] MedStract, http://www.medstract.org
[14] Pubmed, http://www.ncbi.nlm.nih.gov/PubMed
[15] Satou, k. et al., "Extraction of substructures of proteins essential to their biological fuctions by a data mining technique," Proceeding of Int. Conf. Intell. Syst. Mol. Biol., Vol. 5, pp.397-408, 1997.
[16] Satou, k. et al., "Finding association rules on heterogeneous genome data," Proceedings of Pacific Symposium on Biocomputing, pp.397-408, 1997
[17] Uetz, P. et al., "A comprehensive analysis of protein-protein interaction in Sacchromyces cerevisiae," Nature, Vol. 403, pp.623-627, 2000.
[18] UMLS, http://www.nlm.nih.gov/research/umls
[19] Wiliam B. Frakes, "Imforamtion Retrieval : Data Structures & Algorithms" 1995.