2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Data Mining with Constructing Database and Researching Trend Investigation Related with the Field of Nonlinear Problem

Ayahiko Niimi
School of Systems Information Science, Future University-Hakodate
116-2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041-8655 Japan
Email: niimi@fun.ac.jp

***Abstract*—In this paper, we propose an approach which contains with constructing a bibliography information database, extracting the fields of research, and researching trend of them, using data mining. To apply our approach to IEICE Technical Report (nonlinear problem society), the database was constructed based on its report, keywords were analyzed using the frequency analysis and the association analysis, and we discussed about the result. We could extract some field of research from the result.**

I. Introduction

In this paper, we discuss about the technique for investigating the research trend from bibliography information.

There are some methods of investigating the research trend from bibliography information. However, journals, academic society proceedings, and society reports, etc. were made online recently, and it becomes to be able to do the use for the retrieval on the network easily. Moreover, it becomes to be able to do the data analysis easily by improving the performance of personal computers. We thought whether it was possible to search for the research trend from document bibliography information by using data mining from such topics.

Then, we discuss the technique of the research trend investigation including construction of document database. We propose about the following flow as a procedure of the research trend investigation. First of all, the document bibliography information database is constructed with bibliography information. The text mining is applied to the constructed database, and keywords are extracted from bibliography information. And, we discuss about the research trend investigation based on analysis of keyword which is defined in bibliography information and is extracted from other bibliography information. In this paper, our proposed technique can become to apply to a wider field because proposed technique is considered not only using existing bibliography database but olso constructing of the document bibliography information database.

In this paper, we used document bibliography information database related with the field of chaos and the nonlinear problem as target database, chaos and nonlinear problem bibliography database is constructed, the research field related with chaos and its research trend was investigated based on keywords. The title of this paper contains "research trend investigation", but we have not finished investigation of the research trend yet. We discussed constructing and analyzing database, constructed bibliography database, and analyzed keywords by frequency and association. The research trend analysis every year is analyzing now.

II. Database Construction

To do data mining from bibliography information, the database is constructed. The full text was not registered in the database but only document bibliography information to be registered in the database.

Related DataBase (RDB) and XML (eXtensible Markup Language) is combined for the easiness of construction and extendibility after constructed. Input/output data becomes more flexible, its data can be read by both computer and human and high-speed data search with SQL become possible, because XML is used in input/output and RDB is used in data storage.

The system that retrieves paper information by the automatic operation is proposed for the bibliography database construction. [1] It becomes easy to cooperate with such a system by inputting and outputting with XML.

The interface part is constructed with Servlet of Java. It becomes a system that exchanges XML and RDB through Servlet. Figure 1 shows the outline of the system configuration.

We decide input bibliography information which is refer to 15 elements of Doublin Core, the society database constructed by the institute of electronics, information and communication engineers(IEICE), academic conference presentation database by the national informatics laboratory, and BiBTeX, etc. The data are described by RDF (Resource Description Framework) according to Input/Output. RDF is one of the standards to describe the relationships of the resource of the meta data.

The meta data set of Doublin Core referred the decision of document bibliography information. As for the meta data set of Doublin Core, 15 elements are proposed as meta data that describes the resource. [2] (Refer table I.)

There is an advantage that it is easy to treat with XML because 15 elements of Doublin Core are possible the description with RDF. In this paper, 15 elements which in with some elements were changed to be suitable for the target bibliography data were used.

III. Database Analyzing

When data is analyzed from the document bibliography information database, data mining is applicable. The typical

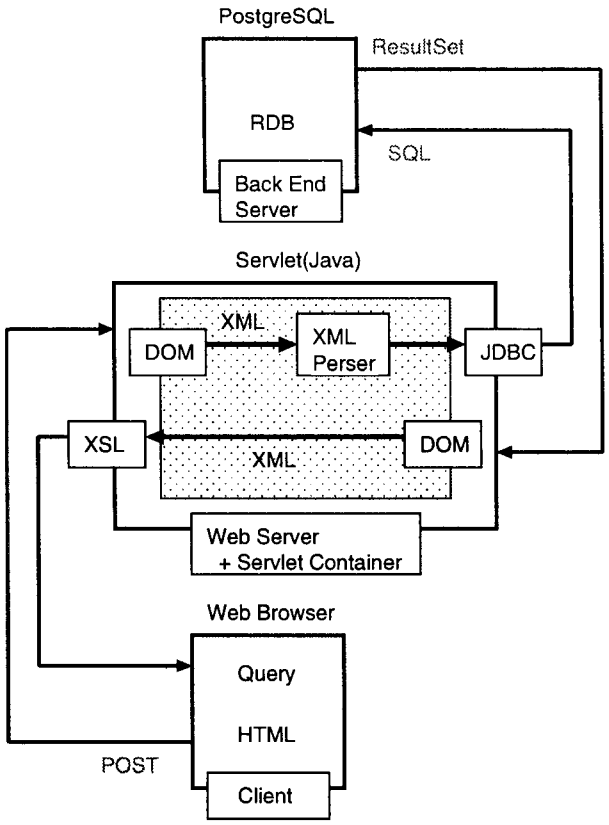

Fig. 1. System Configuration

analysis techniques of data mining are the frequency analysis, the association analysis, and clustering. Using these techniques, the extraction of the keyword of the research field and investigation of the extension of the research field can be possible. Moreover, it is possible to treat the keyword group in the time series as research trend. In addition, the term can be arranged by the association of keywords.

A Keyword Extraction

The title, the chapter title and keyword are written by the natural language. Then, the technique of the natural language analysis can be used for the title, the chapter title. It can be considered about the morphological analysis to the title, the extraction of keywords using the morphological analysis results, the analysis based on frequency and part of speech, the extraction of keywords using association of each keyword.

We have to consider about the extraction keywords from the title and the chapter title when a paper has no keyword. Various methods are proposed as a keyword extraction method from sentences of natural language. [3–8]

B Researching Trend Investigation

It is thought about the analysis using the appearance frequency of the keyword which is input as document bibliography information and is the keyword extracted from the title. It is possible to consider about the keyword often used as a center keyword of the research field related with its database.

TABLE I
DOUBLIN CORE METADATA ELEMENT SET

Title	name of the resource
Creator	responsible party of the resource
Subject	topic, keyword
Description	summary, index
Publisher	entity for making the resource
Contributor	entity for making contributions
Date	created date, availability date
Type	categories, functions, genres
Format	media type, amount, size
Identifier	reference to the resource (URI, ISBN)
Source	source resource
Language	language of the content
Relation	reference to a related resource
Coverage	extent or scope of the content
Rights	information about rights

Moreover, it is possible to consider extracting the word often used at the same time by the association rule for the keyword in a paper. For instance, it can be thought that there is a deep connection between these two keywords if the keyword of chaos and the keyword of neural network are often used at the same time.

In addition, it is consider that the research trend is investigated by using the extracted keywords. At the first, it is possible to discuss that it becomes reference that investigates the research trend in the society using the result of the frequency analysis every year. Moreover, it is possible to discuss that it becomes reference that investigates the research trend of the research field related with the keyword by analyzing frequency every year about a specific keyword.

By the keyword clustering, investigation of the extension of the research field of research becomes possible. And, It is thought that the extension of the research trend can be understood from the clustering result every year.

It is thought whether the support of a grasp of the present research trend and a new research field development can be facilitated by building these analyses into the system. It is thought that analysis tools of such a research trend are very useful as the tool to advance the research.

IV. DATABASE ABOUT BIBLIOGRAPHY INFORMATION RELATED WITH THE FIELD OF NONLINEAR PROBLEM

The chaos bibliography database was constructed as a database to be analyzed. Input/Output were made XML base as a document bibliography information database based on RDB. Input document bibliography information was decided referring to Doublin Core. The Input/Output of the database were developed by Servlet of Java. Technical reports of nonlinear problem society in The Institute of Electronics, Information and Communication Engineers(IEICE) from 1959 to 2001 were used as input data. IEICE already provided the public of the bibliography database on the Internet, but new database was constructed because IEICE database does not have enough data (a registered numbers of paper, a registered numbers of elements) to use technical reports with our proposed analysis.

The title of each chapter was decided to be input instead of abstract because there were a lot of old papers which abstract and the keyword were not attached. In the analysis, the keyword was extracted from the title and the chapter title, and it was analyzed together with the keyword in bibliography information. When registering in the database, the chapter title that do not clearly become keywords are not input. (ex. "Introduction" and "Conclusion") There were some papers without the keyword and the chapter title. The captions of figure and table were input for database instead of the chapter title.

Refer to Table II for the size of the database, Table III for the registered elements corresponding Doublin Core elements. In the table, "ja" means Japanese, and "en" means English.

TABLE II
DATABASE ABOUT NONLINEAR PROBLEM RESEARCH SOCIETY

Journal Source	The Institute of Electronics, Information and Communication Engineers, Nonlinear Problem Research Society
Year	1959 — 2001
Amount of Papers	2315 papers
Japanese Keywords	5881 words
English Keywords	5953 words
Chapter Titles	14395
Extracted Keywords	9439 words

TABLE III
BIBLIOGRAPHY INFORMATION

Elements	Corresponding Doublin Core Element
Title (ja, en)	Title
Keyword (ja, en)	Subject
Chapter title	Description
Authors(ja, en)	Creator
Affiliation (ja, en)	Creator
Document Number	Identifier
Journal	Source
Journal Number (Vol, No)	Source
Pages	Source
Research society	Contributor
Scientific society	Publisher, Rights
Publication date	Date
Publication language	Language
Classification(Proceedings)	Type

The method of arranging the paper with data mining is applied various fields such as bioinformatics field. However, it is difficult to analyze papers because there are extensions of the field in chaos and nonlinear, and it covers over two or more fields of electrics, mathematics, physics, neural system, image processing, and signal processing, etc. In this paper, we used a nonlinear problem society of IEICE in which a wide presentation was done. In this society, the researches of a lot of different fields are included, and it is possible to grasp it with various view, such as theory center or applied center, mechanism, view, usage, expression, the theory center or simulation center or real world oriented, etc.

V. RESULTS OF KEYWORD ANALYSIS

After database construction, words were extracted from the title and the chapter title by Japanese parser. The keywords of extracted keyword and bibliography information were mixed, we analyzed research trend with keywords which is total 9439 words. We made and analyzed each combination of keywords that is items of a Japanese keyword, an English keyword, and is extracted from Japanese titles and English titles and chapter titles. The keywords were filtered to remove words which do not become a good keyword.

In the frequency analysis, the keywords which were strongly related with chaos and nonlinear problem research field (chaos, neural, bifurcation, circuit, and model) have been extracted. (A part of an analytical result of English keywords by the frequency analysis is shown in Table IV.) The frequency is defined by $TF.IDF$.

TABLE IV
RESULT BY THE FREQUENCY ANALYSIS (TOP20)

$tf.idf$	keyword	$tf.idf$	keyword
557.46	chaos	288.08	equation
455.56	neural	285.04	oscillator
447.60	system	282.94	chaotic
415.57	bifurcation	279.46	phase
410.16	network	276.81	model
394.69	nonlinear	273.09	linear
383.93	method	266.54	analysis
362.38	circuit	254.16	control
313.36	coupled	251.38	networks
307.43	synchronization	248.25	function

In the N-gram method, the keywords were analyzed while increasing word of numbers to the maximum that could take N words. (In this paper, N-gram is an analysis of the combination of consecutive n words.) The extraction keywords were almost same as the result by the analysis of extraction with the continuous noun appearance frequency. (A part of an analytical result of English keywords by the n-gram analysis is shown in Table V.)

TABLE V
RESULT BY N-GRAM ANALYSIS (TOP20)

$tf.idf$	keywords	$tf.idf$	keywords
557.46	chaos	307.43	synchronization
455.56	neural	288.08	equation
445.73	system	285.04	oscillator
415.57	bifurcation	281.34	chaotic
410.16	network	279.46	phase
394.69	nonlinear	276.81	model
382.84	method	273.09	linear
363.14	neural network	266.54	analysis
358.83	circuit	252.43	control
313.36	coupled	251.38	networks

Table VI shows the result by the frequency analysis by the keyword that contains the keyword in bibliography, the extraction keyword from the title and the chapter title. In the table, "ja" means Japanese keyword, and "en" means

English keyword. Because Japanese and English are mixed and analyzed, the word that a Japanese term was paraphrased to English is comparatively extracted by a near score. When investigating for the field of research, it is necessary to filter the paraphrase of Japanese-English to bring it together.

TABLE VI
RESULT OF EXTRACTED KEYWORDS (TOP20)

$tf.idf$	keywords
564.07	chaos
554.89	chaos(ja)
377.79	neural network(ja)
359.74	bifurcation
318.01	bifurcation(ja)
315.79	simulation(ja)
294.19	neural network
227.84	circuit model(ja)
191.58	synchronization
187.92	bifurcation phenomenon(ja)
186.89	numerical example(ja)
183.20	circuit equation(ja)
180.09	neural networks
178.41	model(ja)
177.17	hysteresis
165.24	hysteresis(ja)
143.39	associative memory
140.16	coupled oscillator
138.96	nonlinear circuit(ja)
136.76	fundamental equation(ja)

Next, we analyzed keywords by the association rule. The words on data sets were filtered to remove the words which did not become a good keyword, and analyzed. Furthermore, because equations were difficult to analyze association, they were removed from data sets. Table VII shows the result of the association analysis of English keywords. (In the table, Sup was defined as a ratio including all keywords, $conf$ was defined as an average of confidence of the rules by combining keywords.)

Table ?? was an extraction of the keywords with a long rule among analytical results. In the result, "neural network" and "chaos" appeared with high frequency. Then, it is though that these research fields are centered in this society, and the research fields of this society are connected these keyword with other keywords.

The research trend analysis every year is analyzing now.

VI. CONCLUSION

In this paper, we discussed the technique of the research trend investigation including construction of document database, and applied it to document bibliography information database related with the field of chaos and the nonlinear problem.

The document bibliography information database was constructed. Next, the keywords were extracted from title, chapter title. Using the bibliography keywords and extracted keywords, we analyzed keywords by frequency analysis and association analysis, and discussed the results. The research trend analysis is analyzing now.

TABLE VII
RESULT OF THE ASSOCIATION ANALYSIS (TOP20)

sup,$conf$ (%),(%)	keywords
(1.0, 83.4)	neural,networks,optimization,problems,combinatorial
(1.2, 83.5)	neural,network,memory,associative
(1.0, 91.7)	neural,system,network,dynamical
(1.0, 83.6)	neural,networks,problems,combinatorial
(1.0, 83.6)	networks,optimization,problems,combinatorial
(1.5, 84.4)	chaos,system,linear,piecewise
(1.0, 78.9)	neural,optimization,problems,combinatorial
(1.2, 96.1)	neural,networks,optimization,problems
(1.2, 97.4)	neural,networks,optimization,combinatorial
(1.2, 86.7)	chaos,control,feedback,delayed
(1.2, 80.6)	network,memory,associative
(1.4, 49.0)	neural,network,associative
(1.2, 58.9)	bifurcation,nonlinear,circuit
(1.5, 32.4)	neural,system,network
(1.1, 58.5)	neural,network,model
(1.3, 86.1)	control,feedback,delayed
(1.0, 41.5)	neural,problem,optimization
(1.2, 59.6)	chaos,system,dynamical
(1.2, 62.4)	neural,networks,problems
(1.9, 60.5)	chaos,system,piecewise

It could be confirmed to be able to look for some field of research from results. Moreover, it could be confirmed to the analysis that the union of terms was important.

Currently, the database is cleaned, and the application of other data mining is considered in parallel. We discuss about cleaning in which the removing an improper words as keywords and a union of the term are important. We will discuss to consider not only the keyword level, but also the keyword group, the field of research, and the research trend.

REFERENCES

[1] N. Takada, T. Tamura, K. Osawa: Construction of the Papers Reference System on Web in XML, Transaction of The Institute of Electronics, Information and Communication Engineers, D-I, Vol.J84-D-I, No.6, pp.650–657 (2001). (In Japanese)

[2] Doublin Core Metadata Initiative (DCMI), http://doublincore.org/

[3] M. Nagata, H. Taira: Text Classification - Showcase of Learning Theories -, IPSJ Magazine, Vol.42 No.1,pp.32–37 (2001). (In Japanese)

[4] Y. Ichimura, T. Hasegawa, I. Watanabe, M. Sato: Text Mining: Case Studies, Journal of Japanese Society for Artificial Intelligence, Vol.16 No.2,pp.192–200 (2001). (In Japanese)

[5] T. Nasukawa, H. Kawano, H. Arimura: Base Technology for Text Mining, Journal of Japanese Society for Artificial Intelligence, Vol.16,No.2,pp.201–211 (2001). (In Japanese)

[6] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, M. Asahara: Morphological Analysis System ChaSen version 2.2.1 Manual (2000). [Online] Available: http://chasen.aist-nara.ac.jp/chasen/bib.html.en

[7] M. Nagao, S. Mori: A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, In Proceedings of the 15th International Conference on Computational Linguistics pp.611–615 (1994).

[8] R. Agrawal, R. Srikant: Fast Algorithms for Mining Association Rules, the 20th International Conference on Very Large Databases, Santiago, Chile, September 1994:32pages (1994).