Intelligent System for Topic Survey in MEDLINE by Keyword Recommendation and Learning Text Characteristics

Miyako Tanaka[1] miyako@ube-k.ac.jp
Sanae Nakazono[2] b1835@sty.cc.yamaguchi-u.ac.jp
Hiroshi Matsuno[2] matsuno@sci.yamaguchi-u.ac.jp
Hideki Tsujimoto[3] ht@kdel.info.eng.osaka-cu.ac.jp
Yasuhiko Kitamura[3] kitamura@info.eng.osaka-cu.ac.jp
Satoru Miyano[4] miyano@ims.u-tokyo.ac.jp

[1] Department of Business Administration, Ube National College of Technology, 2-14-1, Tokiwadai, Ube 755-8555, Japan
[2] Faculty of Science, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8512, Japan
[3] Faculty of Engineering, Osaka City University, 3-3-138 Sugimoto, Sumiyoshiku, Osaka 558-8585, Japan
[4] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Abstract

We have implemented a system for assisting experts in selecting MEDLINE records for database construction purposes. This system has two specific features: The first is a learning mechanism which extracts characteristics in the abstracts of MEDLINE records of interest as patterns. These patterns reflect selection decisions by experts and are used for screening the records. The second is a keyword recommendation system which assists and supplements experts' knowledge in unexpected cases. Combined with a conventional keyword-based information retrieval system, this system may provide an efficient and comfortable environment for MEDLINE record selection by experts. Some computational experiments are provided to prove that this idea is useful.

Keywords: machine learning system, keyword recommendation, rough reading, MEDLINE, GUI

1 Introduction

Information retrieval from MEDLINE is a daily activity for individual researchers. A set of keywords together with their combinations may specify each individual research topic and this information plays an important role as a key knowledge to efficient and high quality information retrieval from MEDLINE.

On the other hand, people working for constructing new databases of specific biological interests are facing with a different situation. There are some cases that selection by keywords and their combinations does not achieve efficient and high quality results. It is sometimes hard to identify appropriate information for their specific selection interests.

For example, one of the authors has been involved in constructing a database of disordered proteins [8]. In that project, identification of a set of keywords and their combinations which may capture most of the PDB entries containing disordered proteins has exhausted a number of very precise text searches and analyses of the full PDB databases and thorough readings of related articles. In that project, we have given up to use MEDLINE data for finding candidate disordered proteins since the

"disorderedness" has recently received attention [1, 2] and MEDLINE records do not contain explicit descriptions about disorderedness. This may be an extreme case, but there might be a possibility to make use of MEDLINE abstracts for this purpose. Another example is the case of collecting all MEDLINE records which contain useful information for organizing the gene regulatory network of *S. cervisiae* [11, 12]. This rather ambiguous selection condition assumes experts to read the MEDLINE documents for selection. Even with specific keywords such as "Saccharomyces cerevisiae AND regulatory", the experts have to cope with more than 2,800 records in MEDLINE where the total number of MEDLINE records containing "Saccharomyces cerevisiae" is more than 51,000. High quality and high coverage selection is inevitable in order to reduce the cost for thorough reading of the full articles after selection. Thus it is required for experts to read the fewest possible candidates MEDLINE documents covering almost all records of interest with the help of information retrieval systems in an intelligent way.

For such purpose, any informatics tools which can assist experts efficiently in a comfortable way would be appreciated. This paper describes a system we have developed for assisting experts in selecting articles, especially for database construction purposes. The computational learning method [12] we have developed and tested is implemented in this system together with GUIs. This method learns the characteristics in abstracts of MEDLINE as patterns and uses the learnt patterns to select candidate MEDLINE records. This method has a theoretical foundation for guaranteeing the performance, and the experiment using the articles in *Cell* about *S. cerevisiae* showed that 90% correct records were selected by reading the abstracts of 50% records with the assistance of this learning method. Selection by keywords is, of course, implemented in this system as a well-established strategy for selection. But in order to cope with the case that appropriate keywords may not come out from experts due to the nature of selection topic, this system equips a system which recommends keywords for better selection which is implemented based on the system Keyword Recommendation System [4].

Section 2 describes a computational learning method and experimental results supporting this method. In Section 3, GUIs are described for realizing the learning method in our system. Section 4 firstly discusses the difference between natural language processing techniques and the technique proposed in the paper. Secondly, we briefly explain the method to collect MEDLINE records automatically.

2 Literature Selection Method by Machine Learning Technique

This section briefly explains the method [12] for selecting the records of experts' interest from MEDLINE. Our strategy is to express the knowledge of experts by a procedure called a *rough reading function* and to classify the set of abstracts converted by the rough reading function with the help of the machine learning system BONSAI [9]. For convenience of explanation, we describe a method for the purpose of selecting MEDLINE records which contain useful information for organizing the gene regulatory network of *S. cervisiae*.

2.1 Overview of Strategy

•*Basic Operation*

A rough reading function is defined by a mapping of words in texts to a small number of the specified characters. Table 1 shows an example for the above purpose of selecting MEDLINE records where words in the abstracts are converted to the specified characters A, x, y, z and o whose implicit meanings are as follows:

A: Gene Name, x: Very relevant, y: Relevant,
z: Weakly relevant o: Not relevant.

Table 1: Abstract conversion by rough reading function.

abstract	CYC7-H3 is a cis-dominant regulatory mutation that causes a 20-fold overproduction of yeast iso-2-cytochrome c. The CYC7-H3 mutation is an approximately 5 kb deletion with one breakpoint located in the 5' noncoding region of the CYC7 gene, approximately 200 base from the ATG initiation codon. The deletion apparently fuses a new regulatory region to the structural portion of the CYC7 locus. The CYC7-H3 deletion encompasses the RAD23 locus, which controls UV sensitivity and the ANP1 locus, which controls osmotic sensitivity. The gene cluster CYC7-RAD23-ANP1 displays striking similarity to the gene cluster CYC1-OSM1-RAD7, which controls, respectively, iso-1-cytochrome c, osmotic sensitivity and UV sensitivity.
conversion	AooxooooooooooooAoooooooyooooooooooAooooooooooyooooooooooo AzoAyooAzoozyooAzoooyoooAooooooooAooooooyozy

Note that these characters, each of which corresponds to a word in abstracts, will be determined by experts according to their interests and knowledge.

The notion of rough reading is based on an observation that when experts read the abstract of an article for such classification, they first read the abstract roughly and then, if it is considered as a candidate, they read the abstract carefully. Thus, the experts' knowledge is reflected onto the converted texts through the rough reading process.

For the sequences of characters A, x, y, z, and o converted from MEDLINE abstracts through the rough reading process, we employ the machine learning system BONSAI and apply it for discovering rules as which may reduce the work of experts in selecting the records of interest.

A rule consists of an indexing of an alphabet and a decision tree. The indexing of an alphabet Σ is a mapping which categories the symbols in Σ. The decision tree employs regular patterns on the categories as decision rules.

•*Procedure for collecting almost all records of interest by iterating Basic Operations*

Fig. 1 shows the procedure for selecting almost all MEDLINE records of interest. This procedure contains the Basic Operation above. Initially, experts need to define a set A by selecting MEDLINE records according to keywords such as *Saccharomyces cerevisiae*. Then the set A is examined for further selection. In the following, we present the procedurewhich iterates the Basic Operation until almost all relevant records are collected.

1. Choose a set S of records according to the interest of expert and let K be $A - S$. Expert reads abstracts in the set S and divides S into the set POS of records of interest and the set NEG of other records.

2. By applying Basic Operation (rough reading function and BONSAI) to the set S, the indexing and the decision tree are obtained as a rule for classifying the rest of abstracts. Also the set *Converted* K is obtained by applying only the rough reading function to the set K.

3. The set *Converted* K is classified into the sets *POS candi* and *NEG candi* by applying the rule obtained in 2.

4. Expert terminates the procedure, if he/she considers that the set *POScandi* might have no records which they want. It not, expert iterates steps 2 and 3 above by substituting the set S for the set *POS candi* and substituting the set K for the set *NEG candi*.

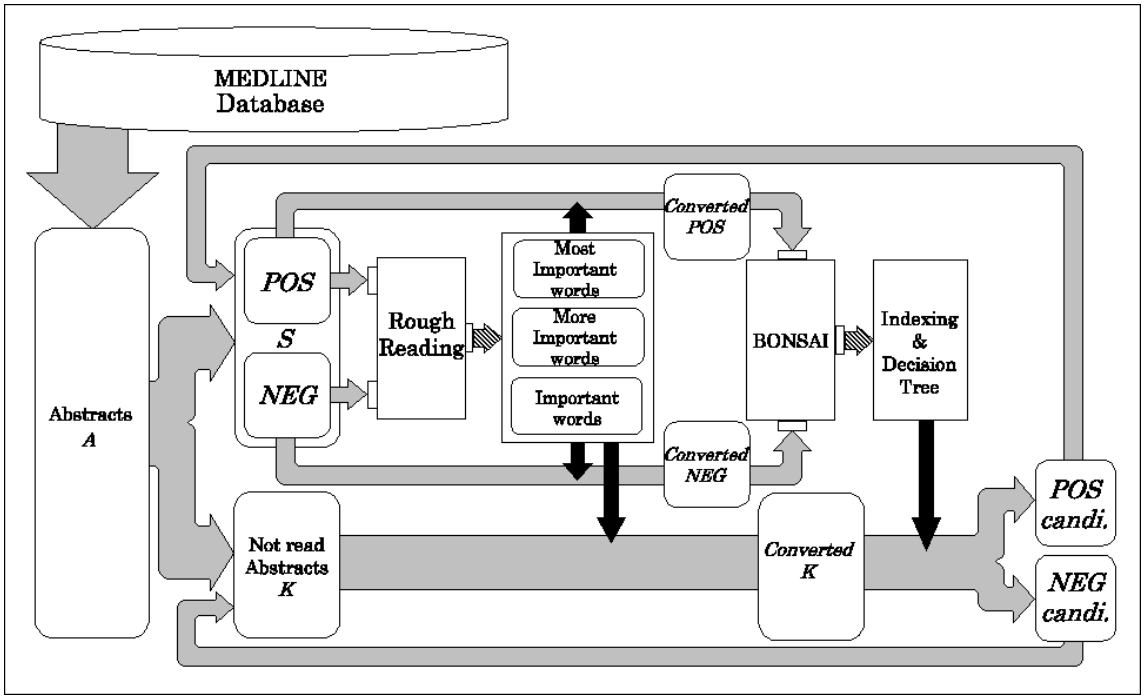

Figure 1: Procedure for selecting MEDLINE records of interest with the help of BONSAI.

2.2 Experimental results[12]

In order to evaluate the performance of the strategy, we prepared 758 MEDLINE records from Journal *Cell*. Prior to computational experiments, experts classified them into two sets, namely, the set of 210 records of interest and the set of 548 records not of interest by asing "*S.cerevisiae*" as a keyword and by reading the records carefully.

Thus we prepared

$$\begin{aligned} POS &: 210 \quad records \\ \underline{NEG} &: 548 \quad \underline{records} \\ A &: 758 \quad records \end{aligned}$$

where"positiveness" / "negativeness" is checked by experts for each record. Then, by using these two classified sets as oracle, we made experiments toevaluate the performance of the strategy. The set of these 758 abstracts corresponds to the set A in Fig. 1. From the set A, we chose randomly five pairs of sets A, B, C, D, and E each of which consists of 20 abstracts of interest POS and 50 abstracts not of interest NEG. We iterated Basic Operations seven times for each of these five pairs of sets. Fig. 2 shows how the ratio of abstracts of interest grows with the number of iterations on five pairs of sets A∼E. We should note that Fig. 2 shows that the method enables us to select nearly 90% of abstracts of interest while leaving half amount of abstracts unread. A mathematical performance evaluation in a simpler framework is also given [12].

3 Design of Tools and Graphical User Interfaces

3.1 Tools for defining rough reading function and classifying abstracts

From the discussions in Section 2, experts need to do the following two works; (i) define the rough reading function, that is, choose and rank words according to the experts' knowledge of the importance

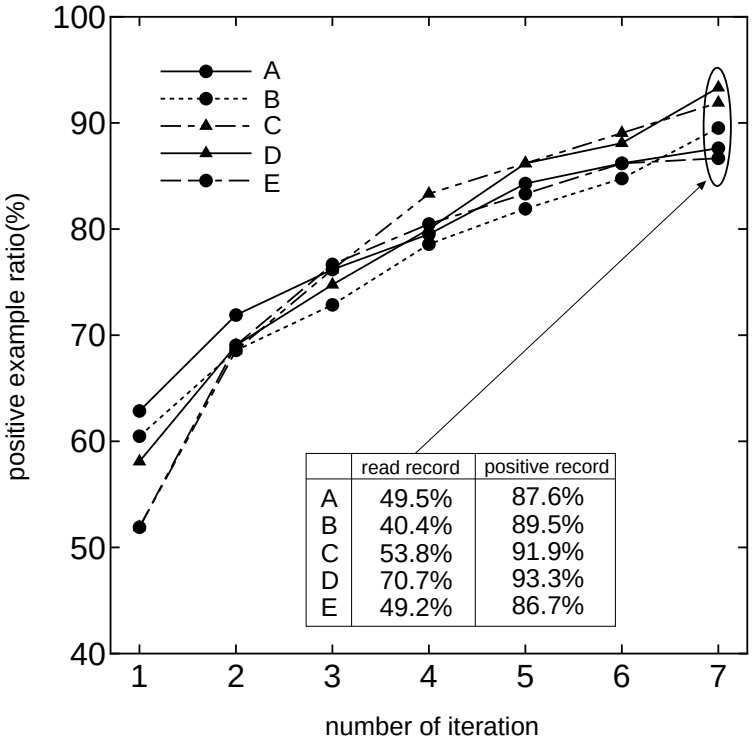

Figure 2: 90% of abstracts of interest can be obtained leaving half amount unread.[12]

of the words, (ii) classify abstracts to positive ones (ones of interest) and negative ones (not ones of interest).

Fig. 3 shows a snapshot of a tool for choosing and ranking words in the set of abstracts A in Fig. 1. The system chooses automatically the words from the set A which has the following features: the number of occurrences of the word is high and the word has a tendency to occur in positive abstracts. See [12] for the formal description of the property. The words listed in Fig. 3 are the words chosen by the system as important words. All words are classified to three ranks of importance, high, middle, and low in terms of the level of the property stated above, but details are omitted here. Experts can check these suggested words and change the importance of the words or delete the words from the list only by clicking the button, if necessary. Note that, since the ranks of importance are recalculated at the each stage of suggestions, some of them would be changed by the system. Thus, the previous rank of importance of the words is displayed in the list as is seen in Fig. 3. Of course, experts can input any words of their intentions to the tool. This tool also has a convenient function of picking up abstracts containing the intended word for helping experts to check the significance of the word as is seen in the right part of Fig. 3.

Experts classify MEDLINE records into three sets, YES, NO, UNKNOWN according to their decision. Fig. 4 shows a snapshot of a tool for that job. The bodies of all abstracts do not always need to be displayed, since experts can decide whether the abstract is needed or not only by seeing the title in many cases. The window describing the body of abstract will be displayed on the screen by pushing the button "?" in Fig. 4. The indicator at the lower part of the tool indicates the numbers of records, these are decided for YES or NO, suggested for YES candidate or NO candidate. This indicator can encourage experts by showing the current status in handling the records.

After accomplishing above two processes corresponding to Figures (Fig. 3 and Fig. 4), BONSAI starts the job in order to pick up the next YES candidate abstracts.

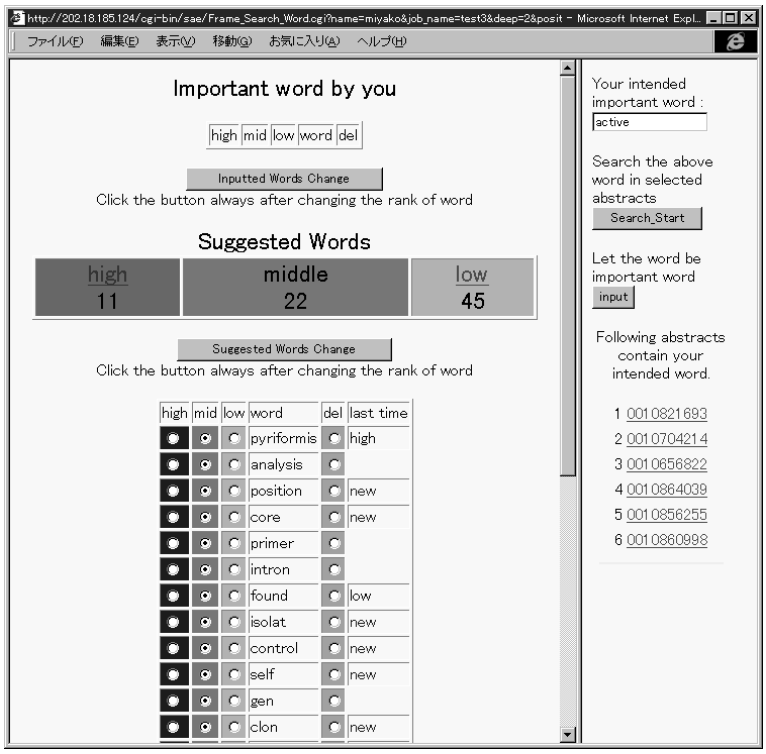

Figure 3: A tool for choosing and ranking words in the selected abstracts.

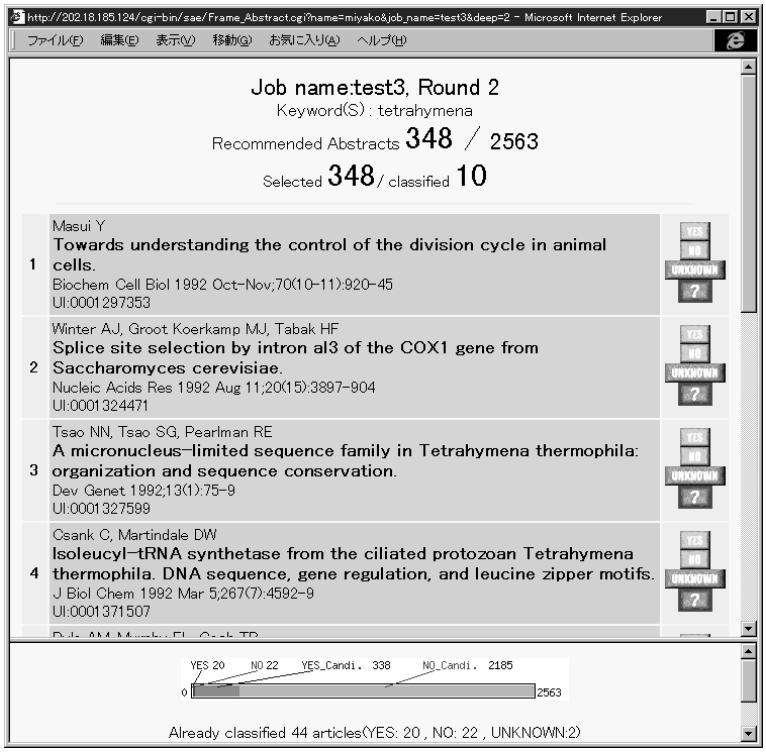

Figure 4: A tool for classifying abstracts into YES, NO, or UNKNOWN.

3.2 Keyword recommendation system and its effectiveness

Recall that *POS candi* in Fig. 1 becomes the set S which is the set of abstracts to be read by expert in the next step. However, the number of abstracts sometimes reaches hundreds and the set S may contain many non-interest abstracts for the expert.

Then, we consider to take a subset of *POS candi* as the set S instead of the set S itself. If this subset has larger ratio of records of interest than that ratio of *POS candi*, we may get more efficient classification rule by BONSAI.

For trying to verify the effect of this idea, we implement Keyword Recommendation System (KRS) [4] in our system. KRS can assist experts who have difficulty to think of proper keywords to narrow down the search basically by recommending appropriate keywords.

By noting the following facts, we design the KRS for our purpose; (1) we sometimes cannot think of proper keywords to specify a field or topic, (2) it is meaningless to specify an author or journal name which are not included in the original documents, and (3) it is difficult to specify a journal name in an abbreviated form which is used in MEDLINE (for example, Eur J Biochem). Then, we decide to take Keywords, Author, Journal, and Publication Year as the categories processed by the KRS.

Keywords We can specify keywords, which are to appear in the abstract, in the title, or in the MH term, to reduce the records to be the ones which are related to a more specific field or topic.

Author We can specify names of authors who published papers to reduce the records to be the ones that are related to a researcher or a research group.

Journal We can specify titles of journals to reduce the records to be the ones that are more credible.

Publication Year We can specify the year of publication to reduce the records to be the ones that are the latest.

The snapshot of the tool is shown in Fig. 5. The KRS initiatively proposes a set of terms according to the above categories to reduce the number of records.

Here we show the mechanism of the tool that proposes a set of terms. As we mentioned, it can proposes a set of terms categorized according to keyword, author, journal, or publication year, but the process is common. Hence, as an example, we show how the tool process the terms in the category of keyword.

At the initial stage, the tool performs the following preprocesses:

1. It extracts the fields of title, MH term, and abstract from MEDLINE records, and decomposes the fields into a set of keywords. It removes meaningless keywords such as “a,” “of,” “the,” and so on.

2. It produces two indices; K-D index and D-K index. K-D index contains pairs of a keyword and a list of record ID's. To an input of a keyword, it returns a set of records that the keyword hits. Conversely, D-K index contains pairs of a record ID and a list of keywords. To an input of a record ID, it returns a set of keywords that the record contains.

At the stage of keyword recommendation, an initial set of records is provided by the set *POS candi* in Fig. 1. The tool sums up the keywords that are contained in the initial set by using the D-K index and shows them in order of the number of appearance. When a keyword is chosen, the tool recalculates the set of records that the keyword hits by using the K-D index, and repeats the above process again.

In order to verify that the concept of KRS is effective or not, we made the experiment by the following strategy. At each stage of iterations, we take a set S from the set *POS candi* by three methods;

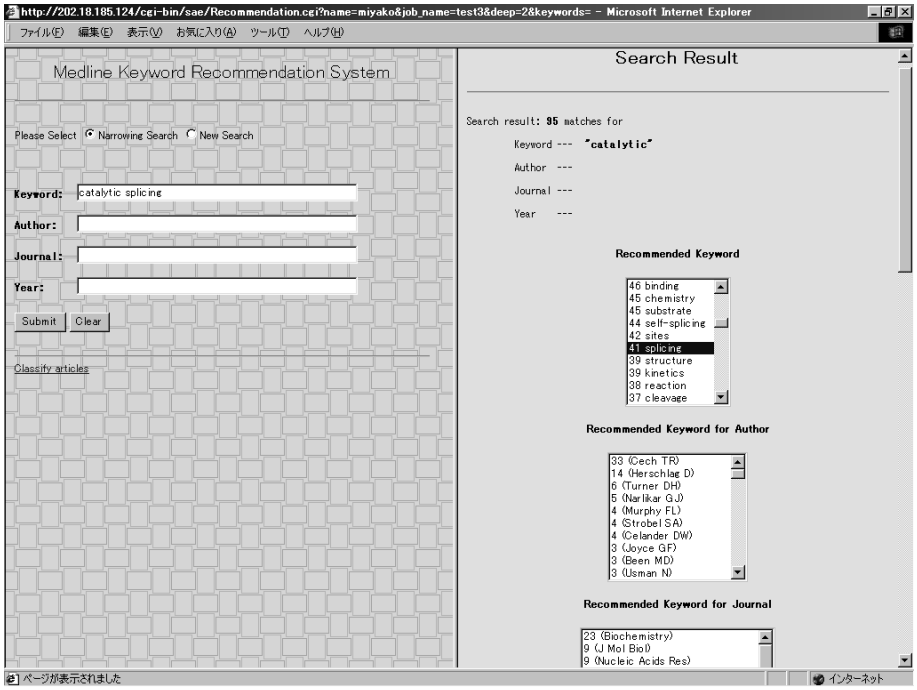

Figure 5: Keyword Recommendation System for MEDLINE.

(1) choose all records in *POS candi* (original method described in Fig.1),

(2) choose n (≥ 1) records in *POS candi* randomly, and

(3) choose n (≥ 1) records in *POS candi* by identifying keywords related to the interest.

Note that in cases (2) and (3), we choose same number ($n \geq 1$) of records in order to make these experiments under the same condition. We can say that the knowledge of expert is reflected in the case (3), but not reflected in the cases (1) and (2).

Fig. 6 shows the result of experiment. We can see that the method of case (3) is more effective than the methods of cases (1) and (2). This concludes that the concept of KRS makes our tool more powerful, since it has an ability to improve reliability of the set S in Fig. 1.

Tools described in this section are available on the site [13].

4 Discussion

Applying Natural Language Processing technique is surely the most effective method of extracting useful information from the biological literature database such as MEDLINE. In fact, many investigations on this direction are made in the literature [5, 6, 7].

Our approach is different. We believe that it is enough to provide a whole abstract to biologists of their interest, since they can identify the desired sentence precisely from the abstract by themselves without large effort. It leads us to that we can suppose the unit of processing is not a sentence but a whole abstract. Note that the technique proposed in the paper [12] is based on this assumption. That is, a whole abstract is converted to the sequence of characters in accordance with the knowledge of biologists, and the converted characters are processed at the same time for obtaining the useful information to biologists.

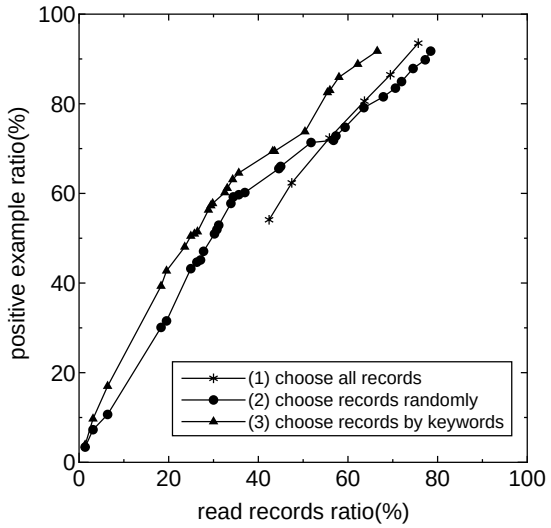

Figure 6: Effect Evaluation of Keyword Recommendation System

This paper concretely demonstrates the next step of the paper [12] by developing the tools for selecting MEDLINE records. In addition, in order to cope with the difficulties of reading the large amount of abstract at once, we also combine the keyword recommendation system with the tool which makes the record set read by biologist more small and reliable. The function of the tool is inherently needed, since, in many cases of our supposing situation, biologists must identify the abstracts without any ideas of the keywords.

In our system, we need to collect a large number of MEDLINE abstracts in advance for analyzing their abstracts by using our machine learning technique and for preprocessing keywords to be recommended by the keyword recommendation system. To assist them, we are utilizing the MetaCommander [3]. The MetaCommander is a generic software robot that automatically collects data from WWW information sources dispersed on the Internet by interpreting a script. It is compatible with the CGI, so, for example, it can submit a query with some parameters to the PubMed and can collect MEDLINE records that are returned as the answer. Hence, the MetaCommander can make an annoying task for collecting a large number of related records easier. Moreover, it can selectively collect the latest records only, so we may provide a service where we recommend a MEDLINE abstract to read as soon as it is submitted to the MEDLINE database.

Of course, we are recognizing that it is indispensable for us to verify the ability of the tool by the experiment in real works of biologists. We had already communicated with the biologists who can help us to verify our tool on the following topics of their interests;

1. Identifying Gene relations in mitochondria DNA,
2. Protein-protein and protein-DNA interactions in yeast, and
3. Genes and SNPs related to disease susceptibility and drug responsiveness.

Furthermore, our subsequent efforts need to be paid to make the tool open to the public. For the

realization, we are planning the experiment and considering the evaluation method of the currently available tool with the help of biologists.

Acknowledgment

This work is partly supported by the Grant-in-Aide (08283101: "Genome Science") for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Bracken, C., Dunker, A.K. and Young, M., PSB2001 Session: Disorder and Flexibility in Protein Structure and Function, *Proc. Pacific Symp. Biocomputing*, 2001.

[2] Kissinger, C.R., Dunker, A.K. and Shakhnovich, E., Disorder in Protein Structure and Function, *Proc. Pacific Symp. Biocomputing '99*, World Scientific, 517–519, 1999.

[3] Kitamura, Y., Nakanishi, H., Nozaki, T., Miura, T. and Ishida, T., MetaViewer and MetaCommander: WWW tools for genome informatics, *Genome Informatics* 7:137–146, 1996.

[4] Kitamura, Y., Nanbu, T. and Tatsumi, S., A keyword recommendation system for GenBank, *Genome Informatics*, 10:206–207, 1999.

[5] Ng, S.K. and Wong, M., Toward routine automatic pathway discovery from on-line scientific text abstract, *Genome Informatics* 10:104–112, 1999.

[6] Rindflesch, T.C., Tanabe, L., Weinstein, J.N. and Hunter, L., EDGAR: Extraction of drugs, genes and relations from the biomedical literature, *Proc. Pacific Symp. Biocomputing 2000*, World Scientific, 517–528, 2000.

[7] Sekimizu, T., Park, H.S. and Tsujii, J., Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts, *Genome Informatics* 9:62–63, 1998.

[8] Sim, K.L., Uchida, T., and Miyano, S., ProDDO: A database system for disordered proteins from the Protein Data Bank (PDB), to appear in Bioinformatics.

[9] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S. and Arikawa,S., Knowledge acquisition from amino acid sequence by machine learning system BONSAI, *Trans. IPS Japan*, 35(10):2009-2018, 1994.

[10] Stapley, B.J. and Benoit, G., Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts, *Proc. Pacific Symp. Biocomputing 2000*, World Scientific, 529–540, 2000.

[11] Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. and Mostafa, J., Detecting Gene Relations from MEDLINE Abstracts, to appear in *Proc. Pacific Symp. Biocomputing 2001.*

[12] Usuzaka, S., Sim, K.L., Tanaka, M., Matsuno, H. and Miyano, S., A machine learning approach to reducing the work of experts in article selectio from database: A case study for regulatory relations of Genes in MEDLINE, *Genome Informatics*, 9:91–101, 1998.

[13] *http : //moon.business.ube-k.ac.jp/˜miyako/sae.html*