INTELLIGENT SYSTEM FOR MEDLINE RECORD SELECTION ENHANCED WITH A LEARNING SYSTEM FOR CHARACTERISTICS OF ABSTRACT TEXTS AND A KEYWORD RECOMMENDATION SYSTEM

MIYAKO TANAKA
Department of Business Administration, Ube National College of Technology, 2-14-1, Tokiwadai, Ube 755-8555 Japan

SANAE NAKAZONO, HIROSHI MATSUNO
Faculty of Science, Yamaguchi University, 1677-1 Yoshida, Yamaguchi 753-8512, Japan

HIDEKI TSUJIMOTO, YASUHIKO KITAMURA
Faculty of Engineering, Osaka City University, 3-3-138 Sugimoto, Sumiyoshiku, Osaka 558-8585, Japan

SATORU MIYANO
Human Genome Center, Insititute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

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.

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[1]. 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 [2,3] 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*[4]. 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 45,000. 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 [4] 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[5].

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. Section 5 concludes this paper with further problems.

2 Literature Selection Method by Machine Learning Technique

This section briefly explains the method [4] 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 [6]. 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.

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

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 the indexing and the decision tree which may reduce the work of experts in selecting the records of interest.

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

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' non-coding 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	AooxooooooooooooAooooooyooooooooooAooooooooooyooooooooooo AzoAyooAzoozyooAzoooyoooAooooooAooooooyozy

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

1. Choose the set of records S according to the interests of experts and let the rest of records be the set K. Experts read abstracts in the set S, classifying the set S to the set of records POS of interest and the set of other records NEG.

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. Experts terminate the procedure, if they decide that the set *POS candi* might have no records which they want. It not, experts iterate the steps

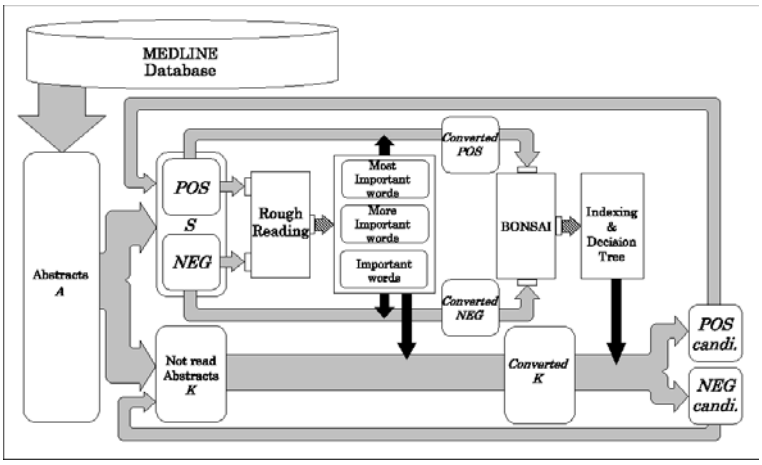

Figure 1: Procedure for selecting MEDLINE records of interest with the help of BONSAI.

2 and 3 above by substituting the set S for the set *POS candi* and substituting the set K for the set *NEG candi*.

2.2 Experimental results[4]

In order to evaluate the performance of the strategy, we prepared 758 MEDLINE records from Journal *Cell*. Experts classified them into two sets, namely, the set of 210 records of interest and the set of 548 records not of interest, by careful reading of the records. Then, by using these two classified sets as oracle, we made experiments to evaluate the performance of the strategy. The set of these 758 abstracts corresponds to the set A in Figure 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. Figure 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 Figure 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[4].

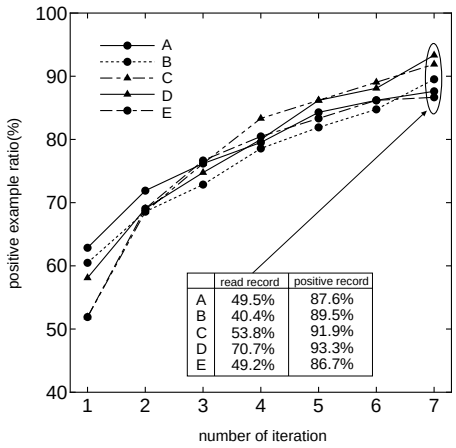

Figure 2: 90% of abstracts of interest can be obtained leaving half amount unread.[4]

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

Figure 3 shows a snapshot of a tool for choosing and ranking words in the set of abstracts A in Figure 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[4] for the formal description of the property. The words listed in Figure 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 Figure 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 Figure 3.

Experts classify MEDLINE records into three sets, YES, NO, UNKNOWN according to their decision. Figure 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 Figure 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 3 and 4, BONSAI starts the job in order to pick up the next YES candidate abstracts.

3.2 Keyword recommendation system helping experts to classify abstracts

Recall that *POS candi* in Figure 1 becomes the set S which is the set of abstracts to be read by experts in the next step. However, the number of abstracts sometimes reaches hundreds. We may well prefer to read them in order of our interest rather than to do in a random order. To reduce the number of MEDLINE abstracts to read, we can use the following clues:

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 key word recommendation system initiatively proposes a set of terms according to the above categories to reduce the number of records. The snapshot of the tool is shown in Figure 5. This is useful because (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, (3) it is difficult to specify a journal name in an abbreviated form which is used in MEDLINE (for example, Eur J Biochem). By using the tool,

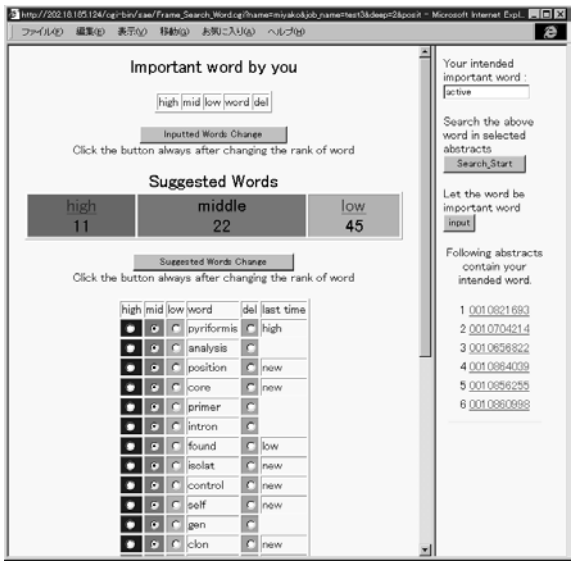

Figure 3: A tool for choosing and ranking words in the selected abstracts.

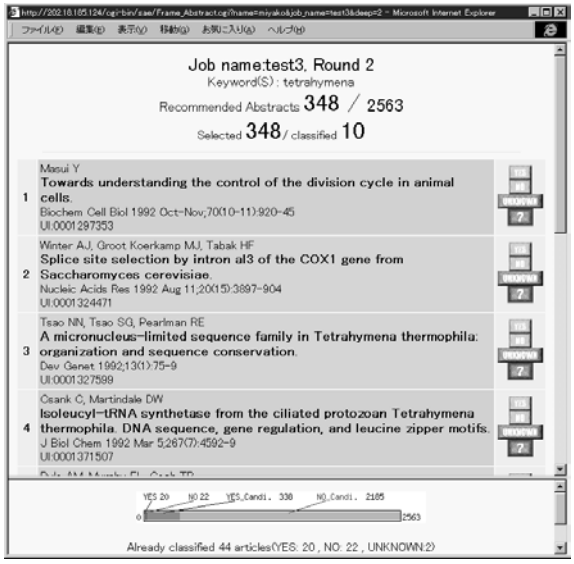

Figure 4: A tool for classifying abstracts into YES, NO, or UNKNOWN.

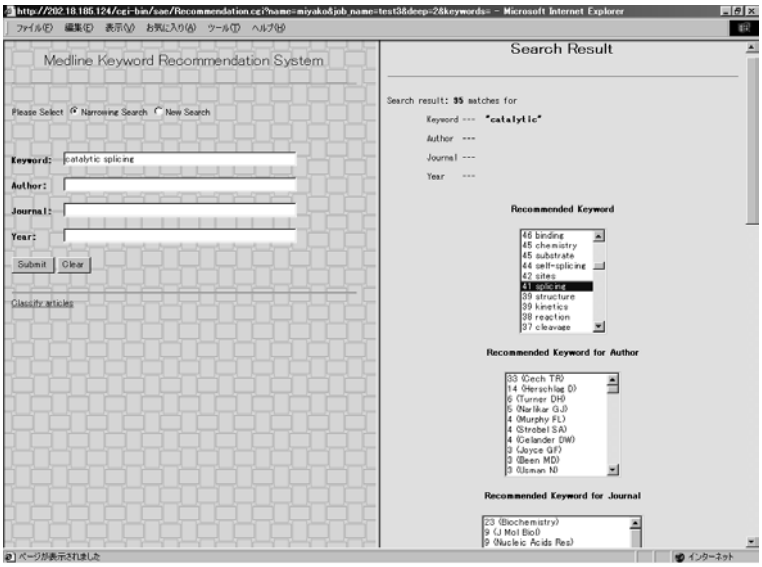

Figure 5: Keyword Recommendation system for MEDLINE.

we can reduce a large number of records of the set S in Figure 1 to be a small set which is readable at once.

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 S in Figure 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.

Tools described in this section are available on the site [7].

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 [8,9,10].

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 [4] 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.

This paper concretely demonstrates the next step of the paper [4] 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 enables biologists to read abstracts in the order of their interest. 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 [12]. 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.

5 Conclusion

We developed an intelligent system for selecting MEDLINE records aiming at the situation that keywords and their combinations do not promise efficient and high quality results. The system consists of the following two systems. The first is a tool which chooses records of interest with the help of machine learning technique. The efficiency of the tool had been guaranteed by the experiments in the paper [4]. The second is a tool which recommends a set of terms categorized according to keyword, author, journal, or publication year. We can reduce a large number of records suggested by the above tool to be a small set which is readable by experts at once. Both tools have the capability for experts to interact dynamically with the GUIs.

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. 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.

1. K.L. Sim, T. Uchida and S. Miyano, PDB-DO: Database of Disorder, *Genome Informatics* **10**, 284–285, 1999.
   *http : //bonsai.ims.u − tokyo.ac.jp/˜klsim/GIW*99_*disorder.html*
2. C.R. Kissinger, A.K. Dunker, and E. Shakhnovich, Disorder in Protein Structure and Function, *Pacific Symposium on Biocomputing'99*, 517–519, 1999.
3. C. Bracken, A.K. Dunker, and M. Young, PSB2001 Session: Disorder and Flexibility in Protein Structure and Function, *Pacific Symposium on Biocomputing*, 2001.
4. S. Usuzaka, K.L. Sim, M. Tanaka, H. Matsuno, and S. Miyano, A machine learning approach to reducing the work of experts in article selection from database: A case study for regulatory relations of *S.cerevisiae* genes in MEDLINE, *Genome Informatics* **9**, 91-101, 1998.
   *http : //www.jsbi.org/journal/GI*09.*html*
5. Y. Kitamura, T. Nanbu, and S. Tatsumi, A keyword recommendation system for GenBank, *Genome Informatics* **10**, 206–207, 1999.
6. S. Shimozono, A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara, and S. Arikawa, Knowledge acquisition from amino acid sequence by machine learning system BONSAI, *Trans. IPS Japan*, **35(10)**, 2009-2018, 1994.
7. *http : //moon.business.ube-k.ac.jp/˜miyako/sae.html*

8. T. Sekimizu, H.S. Park, J. Tsujii, Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts, *Genome Informatics* **9**, 62–63, 1998.
9. S.K. Ng and M. Wong, Toward routine automatic pathway discovery from on-line scientific text abstract, *Genome Informatics* **10**, 104–112, 1999.
10. T.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter, EDGAR: Extraction of drugs, genes and relations from the biomedical literature, *Pacific Symposium on Biocomputing 2000*, 517–528, 2000.
11. B.J. Stapley and G. Benoit, Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts, *Pacific Symposium on Biocomputing 2000*, 529–540, 2000.
12. Y. Kitamura, H. Nakanishi, T. Nozaki, T. Miura, T. Ishida, MetaViewer and MetaCommander: WWW tools for genome informatics, *Genome Informatics* **7**, 137–146, 1996.