Design and Implementation of an Evolutional Data Collecting System for the Atomic and Molecular Databases

SASAKI Akira, JOE Kazuki[1], KASHIWAGI Hiroe, WATANABE Chiemi[1], SUZUKI Manabu[2], PICHL Lukas[2], OHISHI Masatoshi[3], KATO Daiji[4], KATO Masatoshi[4] and KATO Takako[4]

Advanced Photon Research Center, Japan Atomic Energy Research Institute,*

[1] *Nara Women's University,* [2] *The University of Aizu,*

[3] *National Astronomical Observatory of Japan,* [4] *National Institute for Fusion Science*

（Received: 21 October 2004 / Accepted: 21 September 2005）

Abstract

We study an evolutional system to assist an individual to collect articles, which contain atomic and molecular data, and to develop the database automatically. A text classification technique based on LVQ (Learning Vector Quantization) is proposed and its performance is evaluated using abstracts from atomic and molecular bibliographic databases as training and test samples.

Keywords:

atomic data, database, neural computing, learning vector quantization

1. Introduction

Increasing demand from basic science and industrial application requires collection and evaluation of a large amount of atomic and molecular data [1]. Present database development relies on manual collection of articles including the atomic and molecular data, and extraction of data. In recent times, papers in major journals have been published electronically and have been made available on-line. Therefore, it would be possible to develop computer software to collect articles automatically. The text classification approach will be useful to decide whether atomic data is included in the target article [2]. In this paper, we present the preliminary results of LVQ (Learning Vector Quantization [3,4]), which categorizes the paper in terms of the existence of atomic and molecular data.

2. Learning vector quantization (LVQ)

LVQ is a well-known supervised learning algorithm for pattern recognition. The learning is performed with pre-categorized training samples, and the resultant LVQ based system recognizes (classifies) unknown samples. The learning uses a set of reference vectors generated from several feature vectors. A feature vector represents the characteristics of a training sample. The initialization method for the reference vectors does not affect the learning process significantly.

Given a training sample, the nearest reference vector (II) to a feature vector (I) is selected. When the categories of (I) and (II) are equivalent, (II) is updated so that it moves closer to (I). On the other hand, when the categories are different, (II) is updated so that it is kept away from (I). The update is performed as follows.

In the case of identical category,

$$W(new) = W(old) + \alpha(X - W(old)).$$

In the case of different categories,

$$W(new) = W(old) - \alpha(X - W(old)),$$

where W, X, and α are reference vector, feature vector of training samples, and learning factor, respectively. The learning of LVQ is carried out by repeating the update of the reference vectors with training samples.

3. Paper classification by LVQ

3.1 Process overview

Our paper classification process consists of following steps.

Step 1）Preparation of training data (papers and their abstracts) and test data (abstracts, (3)). Classification of the training data into (1) the abstracts containing atomic and molecular data and (2) the others. We refer to category-1 as (1) and category-0 as (2).

Step 2）Pre-process (1), (2) and (3). The pre-process generates feature vectors with term frequency information.

Corresponding author's e-mail: sasaki.akira@jaea.go.jp

* *Present: Japan Atomic Energy Agency*

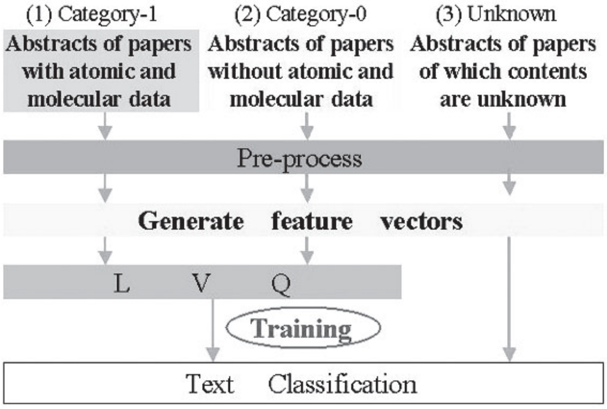
(1) Category-1
Abstracts of papers with atomic and molecular data
(2) Category-0
Abstracts of papers without atomic and molecular data
(3) Unknown
Abstracts of papers of which contents are unknown
Pre-process
Generate feature vectors
L V Q
Training
Text Classification

Fig. 1 System overview

Step 3) Apply the LVQ algorithm with the feature vectors of (1) and (2).

Step 4) Apply the learned reference vectors with the feature vectors of (3) for the recognition of the abstracts of atomic data and molecular data related papers.

3.2 Pre-process

3.2.1 Extract abstracts

Abstracts downloaded from on-line journals are a form of html files. We need to extract the text data from the html files for the LVQ learning process. In this paper, we use abstracts of six on-line journals, each of which have a different format. Hence, we develop a parser that understands the six different formats in order to detect and extract the text regions. In order to develop the parser, we make use of a freeware tool Flex [5] for lexical analysis.

3.2.2 Chemical symbols and mathematical expressions

Chemical symbols and mathematical expressions are commonly used in physics papers. Since they often include unexpected blanks or special fonts, detecting meaningful regions of the text using a parser is at times very difficult. For such chemical and mathematical expressions, we adopt two types of tags to be inserted by using Flex: ⟨math⟩⟨/math⟩ for a mathematical expression and ⟨chem⟩⟨/chem⟩ for a chemical symbol. With the help of the tags, the parser can understand the abstract text. In this paper, we do not use chemical symbols and mathematical expressions for constructing feature vectors because we focus on the text classification and not on the special keyword detection. All the other words are processed with stopword removing and stemming, which are commonly used for text processing.

3.3 Feature vector

We use the frequency of each word in the abstracts of the target samples for the feature vector of LVQ. The frequency is calculated by the TF/IDF (Term Frequency/Inverse Document Frequency) method [6]. The TF/IDF method is employed to measure the importance of a word used in a document on the basis of its appearance frequency, and in general assigns a weight to the word.

3.4 Learning with LVQ

The learning with LVQ is performed until the recognition rate for the training samples reaches 97 %. In the case that the recognition rate does not reach 97 % within twenty epochs, the learning is forced to stop. This is because the recognition rate converges within 20 epochs for the most cases. In our experiments, the learning factor is valid unless it exceeds 0.5. We adopt a conservative value of 0.2 through all the experiments. Each reference vector is obtained as an average vector of randomly selected five feature vectors for training samples on the basis of the category. This method of defining reference vectors is employed to reduce the learning cost.

4. Evaluation

In order to validate our paper classification system, we evaluate the prototype using several experiments. The test data for the experiments is selected from abstracts of 364 physics papers [7]. Each experiment shows the rates of correctness, reproduction, and accuracy.

The correctness rate is the probability that a given test sample is classified into the correct category. The reproduction rate shows the probability that a given test sample, which is an abstract of a paper for atomic and molecular data, is classified into category-1. The accuracy rate is the ratio of the number of test samples, which are known to be in category-1, to the number of test samples, which are classified into category-1 by our prototype.

A trade-off relation exists between the accuracy of the reproduction rates. When the reproduction rate is improved, the number of test samples that are classified into category-1 also increases. This leads to a deterioration in the accuracy rate. On the other hand, improvement in the accuracy rate causes a degradation of the reproduction rate because the number of test samples, which are removed from category-1, increases. Therefore, we should clearly determine which rate we should focus on.

At this point, we are interested in the robust classification wherein any test sample of category-1 is classified as belonging to category-1 with certainty rather than an inaccurate classification wherein any test sample of category-0 is classified into category-1. At this point, we are interested in the robust classification where any test sample of category-1 is surely classified into category-1 rather than the accurate classification where any test sample of category-0 should not be classified into category-1. Hence, we focus on the reproduction rate.

In this experiment, we use 364 abstracts: 182 as training samples and 182 as test samples. The feature vector consists of 1,140 elements.

4.1 Experiment-1

In order to investigate the change of the correctness, reproduction, and accuracy rates on the basis of the selection of the training samples, the LVQ learning is carried out with various sets of training samples (Experiment-1). The number of reference vectors is 120, where 60 % are for category-1, and 40 % are for category-0. Figures 2-4 show the correctness, reproduction, and accuracy rates, respectively.

Figure 2 shows that the correctness rate is in the range of 65 % – 75 %. In particular, the selection of training samples does not affect the variation of the correctness rate significantly. In the meantime, Fig. 3 and 4 show the reproduction and the accuracy rates are in the range of 55 % – 80 %. The amount of the variation cannot be ignored. We expect that the large variation can be reduced by increasing the number of training samples.

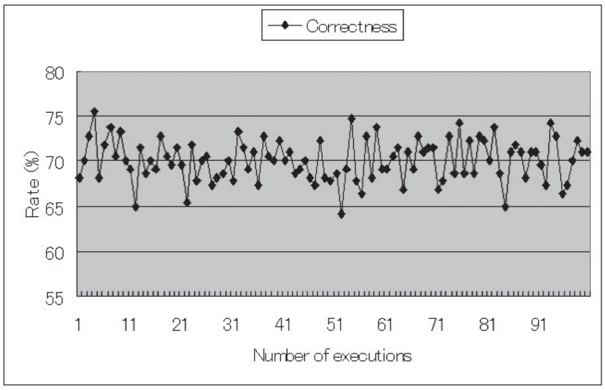

Fig. 2 Correctness rate and learning epochs

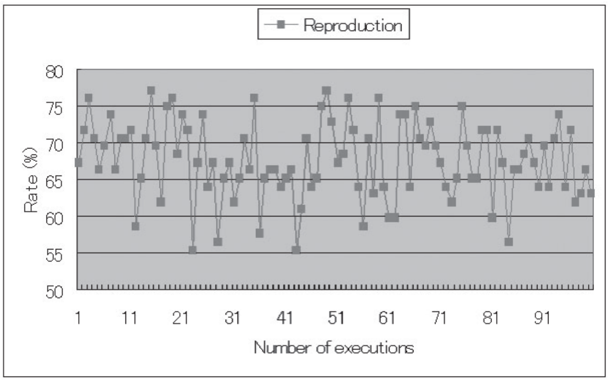

Fig. 3 Reproduction rate and learning epochs

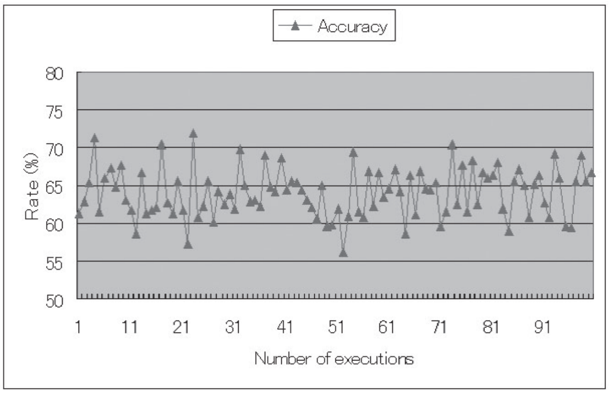

Fig. 4 Accuracy rate and learning epochs

4.2 Experiment-2

Experiment-2 is conducted to study the optimal ratio of the number of reference vectors for category-1 to that of all the reference vectors. In this experiment, the number of reference vectors is 120, and the ratio of the number of reference vectors for category-1 to that of all the reference vectors varies from 10 % to 90 %. Figure 5 indicates the correctness, the reproduction, and the accuracy rates for each ratio. Each rate is an average of one hundred recognition results with different LVQ learning tasks.

We observe a higher reproduction rate as the number of category-1 reference vectors increases. The correctness and accuracy rates reach the maximum when the ratio of category-1 reference vectors is 30 %. With

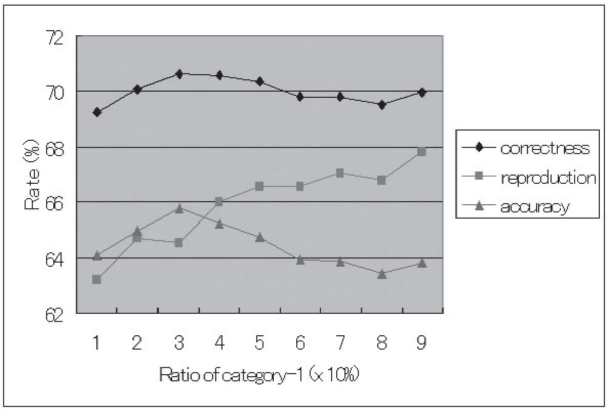

Fig. 5 Correctness, reproduction, and accuracy rates obtained by varying the ratio of the number of reference vectors for category-1 to all the reference vectors

more category-1 reference vectors, both the correctness and the accuracy rates decrease gradually.

As discussed above, we are of the opinion that the reproduction rate is more important than the accuracy rate. When the reproduction rate is at its highest, i.e., when the ratio of category-1 reference vectors is 90 %, the correctness and the accuracy rates are 69.94 % and 63.82 %, respectively. Both rates are too low to be accepted. Considering the combination of the three rates, we adopt 50 % as the optimal ratio of category-1 reference vectors.

4.3 Experiment - 3

Experiment-3 is conducted to investigate the optimal number of reference vectors. The correctness, reproduction, and accuracy rates are examined by varying the number of reference vectors from 10 to 350. Figure 6 shows those rates and the various number of reference vectors. Each rate is calculated as an average of one hundred recognition results with different LVQ learning tasks. It should be noted that the ratio of category-1 reference vectors is 50 %.

As the number of reference vectors increases, the correctness and accuracy rates decrease while the reproduction rate increases gradually. After the number of reference vectors reaches 270, each rate does not change significantly. When the number of reference vectors is 270, the correctness and accuracy rates are 68.99 % and 68.39 %, respectively. Since we are of the opinion that the correctness rate should at least be 70 %, the correctness rate of 68.99 % is not acceptable. With these restrictions, we observe that the number of reference vectors should have an optimal value of 110.

4.4 Evaluation summary

Performing several experiments, we observe that 110 as the optimal number of reference vectors, and 50 % is the optimal ratio of the number of category-1 reference vectors to that of all the reference vectors. Furthermore, we observe that the selection of training samples does not affect the variation of the correctness rate significantly while the reproduction and accuracy rates, between which a trade-off relation exists, vary relatively.

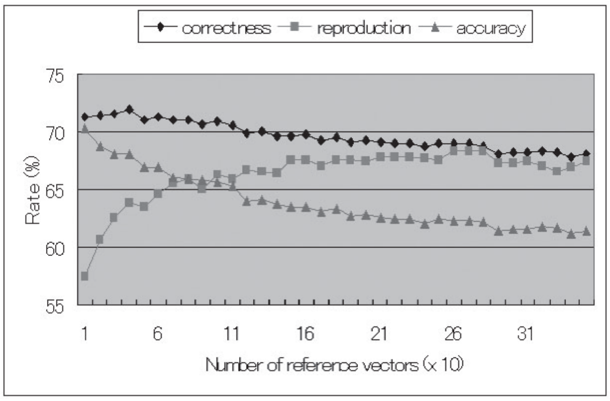

Fig. 6 Correctness, reproduction, and accuracy rates obtained by varying the ratio of the number of reference vectors

Another experiment with an optimal number of reference vectors and optimal ratio of catgegory-1 reference vectors results in the correctness rate of 70.52 %, reproduction rate of 65.88 %, and the accuracy rate of 65.21 %. Although the system appears to be unsuitable for practical application on the basis of its performance, we have identified a number of areas for improvement. First, we do not take into account the chemical and mathematical expressions, which are typically considered to be the most important information to classify the given abstracts, for our prototype. Second, we merely adopt the basic LVQ for our prototype. Other extensions of LVQ algorithms can be used for our prototype. Finally, we may reconstruct the architecture of feature vectors. The feature vector should be built from a larger database pertaining to atomic and molecular data.

Acknowledgements

The author (A.S.) acknowledges useful discussions with Drs. H. Kubo, H. Ihara, Y. Ueshima, and K. Moribayashi from JAERI and Dr. M. Murata from the National Institute of Information and Communications Technology.

References

[1] Atomic and Molecular Data Research Center, NIFS, http://dpc.nifs.ac.jp/admrc/index-j.html.

[2] S. Gao, W. Wu, Chin-Hui Lee, Tat-Seng Chua, ”Maximal figure-of-merit learning approach to text categorization”, ACM SIGIR, p.174-181. (2003).

[3] T. Kohonen ; The Self-Organizing Maps (3rd edition), Springer, 2001.

[4] HUT - CIS - Research - SOM_PAK, LVQ_PAK, http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml

[5] Flex - GNU Project - Free Software Foundation (FSF), http://www.gnu.org/software/flex/.

[6] G. Salton and M.J. McGill : *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, 1983.

[7] Positive samples are taken from the list created by Prof. Y. Itikawa, a part of the list is published in ADANDT, **63**, 315 (1996). Negative samples are taken from Phys. Rev. A-E, vol. 69, No.1.