§21. Study of Evolutional Data Collecting System for the Atomic and Molecular Databases

Sasaki, A., Kubo, H. (Japan Atomic Energy Research Institute)
Joe, K. (Nara Women's University)
Pichl, Lukas (The Aizu University)
Ohishi, M. (National Astronomical Observatory)
Kato, D., Kato, M., Murakami, I., Kato, T. (National Institute for Fusion Science)

We study atomic and molecular database based on computer and internet technology to collect and evaluate a large amount of data required for scientific as well as industrial applications [1]. The project has been started since FY2003, and during FY2004 we have developed several trial codes for the unified search over online journals [2] to find scientific abstracts concerning to the atomic data, and to identify and collect information on the atomic and molecular states in the abstracts [3].

We had a working group meeting on 17-Jun-2004 to understand the present status and direction of the project, as well as programming and network infrastructure applicable to the project.

In addition, we had a hearing from Prof. Itikawa for the experience of collecting articles concerning to the atomic data based on the analysis of titles, abstracts, and figures and figure captions in the articles, to learn the selection is rather made using limited information.

In order to find a scientific article on the atomic and molecular data using a computer program, we need to model articles to be understood by a computer. As a data mining systems in a variety of fields, we firstly used frequency of keywords appeared in abstracts of the papers. Abstracts are chosen because we learn after investigations of the conventional procedure of searching articles that they contain sufficient information of the content. More specifically TF/IDF (term frequency/inverse document frequency) of each keyword is calculated [3]. Using the TF/IDF, any abstract is represented by a point in the *n*-dimensional document space, having *n* determined from the total number keywords contained in the test collection.

The model allows us to evaluate similarity between two abstracts by calculating the Euclid distance in the *n*-dimensional space. Moreover, having a set of reference abstracts concerning to the atomic and molecular data, it may be possible to classify whether the data is likely to be contained or not in test abstract by calculating similarity to the reference set. Although, mathematical representation of the reference set in the document space is not clear, having a sufficient number of reliable positive and negative examples as training data [3], machine language algorithms such as LVQ (Learning Vector Quantization) can be applied to find the criteria for the decision [4].

The accuracy of the text classification is represented by two quantities, precision and recall. In the present system, recall, which corresponds to the fraction of the articles classified by the system as the atomic and molecular data is contained, to the total articles in the sample set, which contain the atomic and molecular data, is more important. Because, we estimate the number of articles, which contain atomic and molecular data is small (<100/year) compared to the total number of articles published in the field of physics and physical chemistry (>10^3/year).

Accuracy of the classification depends on the quatlity of the keywords. Useful keywords are those appear in the articles concerning to the atomic and molecular data, and do not appear in other articles, thus usefulness of each keywords can be estimated by measuring the probability of appearance in the positive and negative examples.

Using machine learning algorithms, pairs of keywords which frequently appear together can are also be recognized. In particular, technical terms are sometimes represented by two or more consecutive nouns, such as "cross section" and "impact ionization". We analyze such technical terms used in the field of physics from the abstracts of articles in Phys. Rev. A-E statistically to find keywords appear frequently in the articles concerning to the atomic and molecular data.

Furthermore, we investigate methods which can recognize and extract expressions of atomic and molecular states in the abstracts, which should appear frequently in the articles concerning to the atomic and molecular data.

For instance, an atomic state is represented hierarchically from the atomic and ionic spicies (Al XI, Li^{3+}, H-like Al etc.), electron configuration ($1s^2 2s^2 2p^2$), and to fine structure ($^1S_0$). In electronic documents, these representations of the atomic states are written according to the syntax rules such as using html as ¹S₀ for the fine structure level. These rules can be generalized and formulated using the regular expression, and the information of atomic states can be recognized and extracted from articles automatically. Interestingly, expression of atomic spicies can easily be extended to simple molecules by accepting the repetition of the rule for atoms.

The present project will be continued to FY2005. We are going to combine techniques to find good keywords for the text classification. Precision and recall of the system will be evaluated quantitatively using a larger set of reference and test sets to improve the accuracy.

References

[1] http://dpc.nifs.ac.jp/amdrc/index-j.html
[2] L. Pichl, et al., Proceedings of DNIS 2005, LNCS 3433, pp. 159, Springer, 2005.
[3] A. Sasaki et al.,paper presented at Joint ITC14 / IAMDATA 2004, October 5-8, 2004
[4] Y.Itikawa, ADANDT, **63**, 315 (1996).
[5] Salton, G. and McGill, M.J. : Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1983.
[6] Kohonen, T.; The Self-Organizing Maps (3rd edition), Springer, 2001.