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

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

We have been studying an evolutional 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 since FY2003 [1]. We had working group meetings on 21-Jun-2005 and 27-Feb-2006, to understand the present status and direction of the project, and discuss about development and improvement of software to classify the papers on the atomic data.

We study the computational method to collect scientific articles, which contain atomic data, because this part can be realized using data mining methods and language processing technologies, and may relax a large amount of present effort of staff scientists in the data centers.

Firstly, we study a software program [2] which classify articles from the analysis of its abstract, based on a LVQ [3], a machine learning algorithm. We model abstracts using the term frequency and inverse term frequency (TF/IDF) of words in the abstracts. We use a set of papers concerning to the electron collisional ionization and excitation of atom and molecules collected by Itikawa [4] as the training samples, to decide the existence of the atomic data in unknown abstracts. Furthermore, we use approximately 300 abstracts from Phys. Rev. A-E, as well as 16000 abstracts from Phys. Rev. A as test examples to investigate characteristic feature of articles from a variety of fields of physics.

It is found that the possibility to find articles, which contain the atomic data, is approximately 1% in approximately 10^4 articles/year from Phys. Rev. A., which consists of the largest part of the source of the atomic data. That is, the accuracy and reproduction must be more than 90% to realize the application of the present software. Although, initial trial software shows almost sufficient reproduction >85%, the accuracy is found to be around 50%. Secondly, we study improvement of the accuracy of classification of articles by several ways. Initial classification software uses only independent words in the articles to characterize its content. The accuracy is expected to improve using better keywords as well as additional scientific and bibliographic information attached to the articles.

For instance, we investigate the PACS number in papers from American journals. It is found that 90% of the articles, which contains atomic data have a few common PACS numbers. Information of authors may also be useful to specify their research area.

On the other hand, we investigate technical terms in the papers in two different ways. Firstly, we take technical terms from the scientific dictionary [6]. We find approximately 800 keywords in the papers concerning to the atomic data, from the total 26000 English keywords in the dictionary. Using these keywords, accuracy of the classification is found to improve slightly.

Secondly, we also use another software termex [7], to see number of technical terms specific to the papers on the atomic data is relatively small (<100). Moreover, it is found that the terms are sometimes represented by two or more consecutive words and form a graph structure, which consists of a stem and additional words such as "excitation", "impact excitation", and " excitation cross section". Extracted technical terms will be also applied to the classification software.

Furthermore, we develop a software to recognize and extract atomic and molecular symbols in the abstract [5]. Information of atoms, electron configurations, and spectral terms are represented in special forms in scientific articles such as, Al XI, $1s^22s^22p^2$, and $^1S_0$. In the electronic form of articles on the web sites, these expressions are represented using the syntax of HTML. We develop a rule-based software using perl regular expression to recognize description of atomic speicies.

Initial program classifies 5 categories, however, after investigation of articles from Phys. Rev. A-E, the characteristic feature of papers for atomic data would become clearer if we would increase the number of categories to 7, where atomic expression with left super script is classified as nuclear species, and repeated expression of atoms is classified as molecules. It is found that 90% of atomic data papers contain expressions of either ions, configurations, and spectral terms. Application of these scientific symbols will also be useful for the classification of the papers.

In summary, we develop and evaluate computer programs to retrieve papers concerning to the atomic data. In FY2006, we are planning to apply above techniques to find better keywords and specific physical expressions to the text classification software. We are also planning to apply the software to newly published papers to demonstrate the usefulness of the present programs for the development of the atomic and molecular database.

References

[1] M. Suzuki, et al., Proc. ITC-14 & ICAMDATA 2004, JPFR series, vol. 7 pp. 343 (2006).
[2] A. Sasaki, et al., Proc. ITC-14 & ICAMDATA 2004, JPFR series, vol. 7 pp. 348 (2006).
[3] Kohonen, T.; The Self-Organizing Maps (3rd edition), Springer, 2001.
[4] Y.Itikawa, ADANDT, **63**, 315 (1996).
[5] A. Sasaki, et al., J. Plasma Fusion Res. , **81**, 717 (2005).
[6] 理化学事典第 6 版（岩波書店）.
[7] http://gensen.dl.itc.u-tokyo.ac.jp/