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

Sasaki, A., Kubo, H. (JAEA),
Joe, K. (Nara Womens's Univ.),
Pichl, L. (Aizu Univ.),
Ohishi, M. (National Astronomical Observatory),
Murata, M. (National Institute of Information and Communications Technology),
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], through a working group activity over the physics and the information science including language processing fields.

We study a software program [2], which classify articles from the analysis of its abstract using a machine learning algorithm [3], based on the frequency of appearance of words. We use a set of abstracts which contains atomic data collected by Itikawa [4], and approximately 300 abstracts from Phys. Rev. A-E, as well as 16000 abstracts from Phys. Rev. A as training and test samples.

The fraction of articles, which contain the atomic data, is approximately 1% among 10^4 articles/year from Phys. Rev. A. Therefore, the accuracy and reproduction must be more than 90% to realize the application of the software. Because initial trial software shows the reproduction of ≈85% and the accuracy of 50%, we have tried to improve the performance of the software by several ways.

We have tried to improve the model, feature vector to characterize the abstract. In the initial development, we used simple TF/IDF (Term frequency / inverse document frequency) of words, however, scientific abstracts usually contains technical terms, which characterize the subject of the article as well as corresponding research fields. We included the frequency of appearance of technical terms specific to atomic and molecular physics, as well as information of atomic and molecular states in symbolic forms.

We collected technical terms from the scientific dictionary [5], and using termex [6] software. In both methods not only simple words, but expressions, consisitng of combinations of multiple words such as "electron collisional excitation" are taken into account.

We also developed software to extract atomic and molecular symbols such as, Al XI, $1s^2 2s^2 2p^2$, and $^1S_0$ from the electronic form of articles on the web. Initially the software recognized expressions of atoms, electron configurations and spectral terms. Susequently, we have improved the software to distinguish nuclear and molecular expressions from atomic and ionic expressions. We classify test abstracts using the improved feature vectors. The reproduction of the classification is improved to nearly 90%.

During the study, we investigate methods to manipulate information of atomic and molecular states in computer programs. Using the regular expression, most of the atomic symbols can be recognized. Even some symbols correspond to more than one physical states such as S for sulfur and S term (L=0) in the LS coupling notation, the two expressions can be distinguished by checking the range of the accompanying physical quantities given in the sub and superscript to the symbol. Moreover using HTML for rendering, the atomic states used inside the computer programs can be displayed in familiar forms for scientists.

Manipulation of information of the atomic states inside the computer program becomes easier using structures and objects. Physical rules such as number of electrons in a shell. With such knowledge, validity of the atomic data given in the input file of the computer program can be checked automatically, which help doing simulation of atomic processes including a large number of states. Moreover, such computerized knowledge can be useful for evaluating theoretical and experimental atomic data and making an estimation of missing data from similar existing processes.

In order to improve the accuracy of the classification of the scientific articles, the quality of the set of the technical terms may be important. However, scientific dictionaries only include general terms, thus the technical terms, which characterize each individual research field should rather be collected using computer programs. Development of dictionaries and method for translation between different scientific fields may be useful for realizing collaboration over different societies.

The software to extract technical terms and atomic and molecular expressions can also be applied to develop software to assist scientific readers to read and analyze a large number of electronic abstracts by highlighting and color coding in the display.

In summary, we study computer programs to retrieve papers concerning to the atomic data. During the study, computational methods to manipulate information of atomic and molecular state are developed, and application of such methods to develop atomic and molecular databases as well as to develop a model for atomic and molecular processes are discussed.

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] Rikagakujiten (6th edition), Iwanamishoten (in Japanese)
[6] http://gensen.dl.itc.u-tokyo.ac.jp/