Acquisition of Data for Plasma Simulation by Automated Extraction of Terminology from Article Abstracts

Lukáš PICHL, Manabu SUZUKI[1)], Masaki MURATA[2)], Akira SASAKI[3)], Daiji KATO[4)], Izumi MURAKAMI[4)] and Yongjoo RHEE[5)]

Division of Natural Sciences, International Christian University, Osawa 3-10-2, Mitaka, Tokyo 181-8585 Japan
[1)]*School of Systems Science, Arkansas Technical University, 1811 N. Boulder Ave., Corley 201-C, AR 72801-2222, USA*
[2)]*National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho Kyoto 619-0289, Japan*
[3)]*Kansai Photon Science Institute, Japan Atomic Energy Agency, 8-1 Umemidai, Kizu-cho, Kyoto 619-0215, Japan*
[4)]*Coordination Research Center, National Institute for Fusion Science, Oroshi-cho, Toki, Gifu 509-5292, Japan*
[5)]*Korea Atomic Energy Research Institute, P.O. Box 105, Yuseong, Daejeon 305-600, Korea*

(Received 4 December 2006 / Accepted 18 April 2007)

Computer simulation of burning plasmas as well as computational plasma modeling in image processing requires a number of accurate data, in addition to a relevant model framework. To this aim, it is very important to recognize, obtain and evaluate data relevant for such a simulation from the literature. This work focuses on the simultaneous search of relevant data across various online databases, extraction of cataloguing and numerical information, and automatic recognition of specific terminology in the text retrieved. The concept is illustrated on the particular terminology of Atomic and Molecular data relevant to edge plasma simulation. The IAEA search engine GENIE and the NIFS search engine Joint Search 2 are compared and discussed. Accurate modeling of the imaged object is considered to be the ultimate challenge in improving the resolution limits of plasma imaging.

Keywords: plasma imaging data, simultaneous database search, automated abstract retrieval, atomic and molecular data, terminology extraction

DOI: 10.1585/pfr.2.S1118

1. Introduction

An accurate plasma simulation for fusion projects as complex as ITER, for instance, remains a challenging task. In order to understand the imaged system accurately, a number of complementary diagnostics and imaging methods are employed. In the case of ITER, these are for instance the infrared thermography, visible spectroscopy, X-ray imaging, gamma-ray and neutron spectroscopic imaging. In order to combine these imaging techniques into a unique picture of the studied system, accurate data within the framework of a relevant physical model are indispensable.

The data needs for imaging purposes are not specific to burning plasmas. For instance, the accurate description of atomic and molecular processes is important in fusion edge plasmas, but also in effective hadron therapy of brain tumors, positron emission tomography, and in other imaging fields; the corresponding data are the key factors, which determine the ultimate limits of the resolution of each particular imaging technique. Here we present a new method for automated retrieval of such data from online publisher databases, in particular the Joint Search 2 (JS2) tool developed at NIFS, which allows simultaneous queries to the Spinweb database of the American Institute of Physics (AIP), and the Electronic Journal database of the Institute of Physics in the United Kingdom (IOP).

The JS2 tool implements a module for the recognition of Atomic and Molecular (A+M) data terminology developed at the National Institute of Information and Communications Technology, which facilitates automated relevance judgment on journal article abstracts by means of machine learning algorithms [1], and forms the basis of automatic processing of INSPEC abstract articles for the NIFS database (http://dbshino.nifs.ac.jp/) in near future.

Finally, the advantages of the JS2 tool are compared to the IAEA GENIE tool, i.e. the General Internet Search Engine for Atomic Data, which supplies numerical data from various online databases. The present work reported herewith is a part of an ongoing A+M database project at the Atomic and Molecular Data Research Center at NIFS [2], and was performed with the support and under the auspices of the NIFS Collaborative Research Program.

2. Joint Search 2 with Terminology Extraction

Since the Joint Search tool was reported in previous publication [3], we only summarize its main features here and focus on the extraction of A+M terminology, which can also be applied in the same manner to the automated processing of article abstracts distributed within the INSPEC abstracting system. Figure 1 shows

author's e-mail: lukas@icu.ac.jp

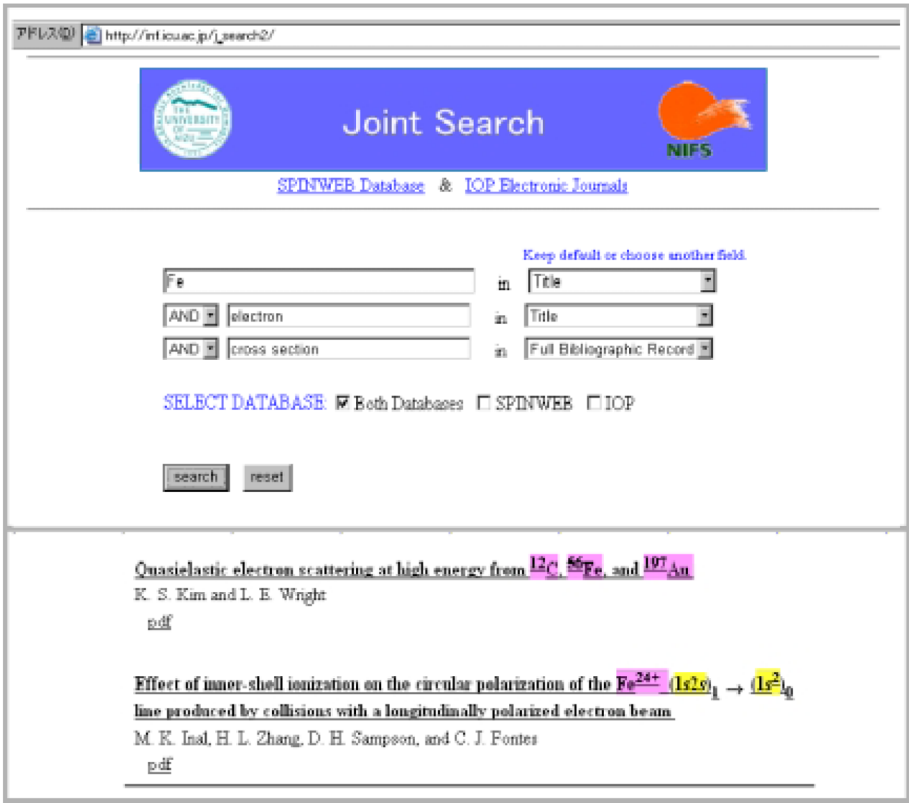

Fig. 1 The interface of the Joint search tool available at http://inf.icu.ac.jp/j_search2/. The upper panel shows the standard form for sending simultaneous queries to the Spinweb and IOP databases. The lower panel is a sample of several retrieved abstracts from both sites, which are extracted, bibliography-classified, and analyzed for A+M expressions. Note the different coloring for chemical elements and their charge states (violet) and electronic state symbols (yellow). The HTML output was edited to fit the format of this figure.

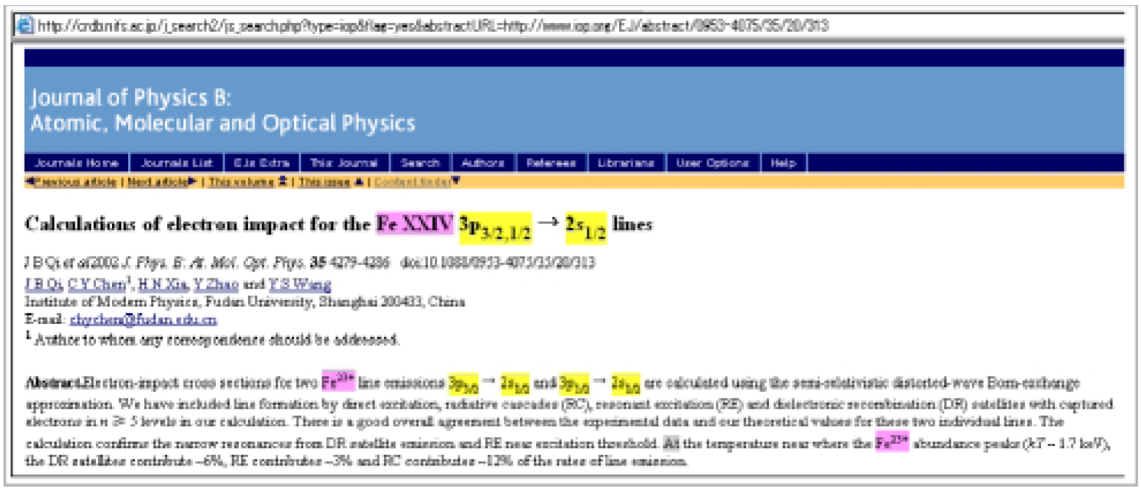

Fig. 2 The A+M textual output for one URL from Fig. 1 (JS2 service is mirrored at crdb.nifs.ac.jp and inf.icu.ac.jp). The HTML output was edited to fit the format of this figure.

the online query form, which is sent simultaneously to http://scitation.aip.org/vsearch/servlet/VerityServlet?KEY = ALL and http://www.iop.org/EJ/search/ using the standard form of parameter transfer to both search forms via HTTP requests. The retrieved answer pages from AIP and IOP are processed by PHP scripts, which discard redundant formatting features, extract bibliography information and apply a Perl-written CGI to colorize all matched A+M terms in the output, as shown in Fig. 2.

In order to judge whether each particular paper obtained in response to a pre-designed keyword query is relevant, i.e. it contains data relevant to fusion plasma simula-

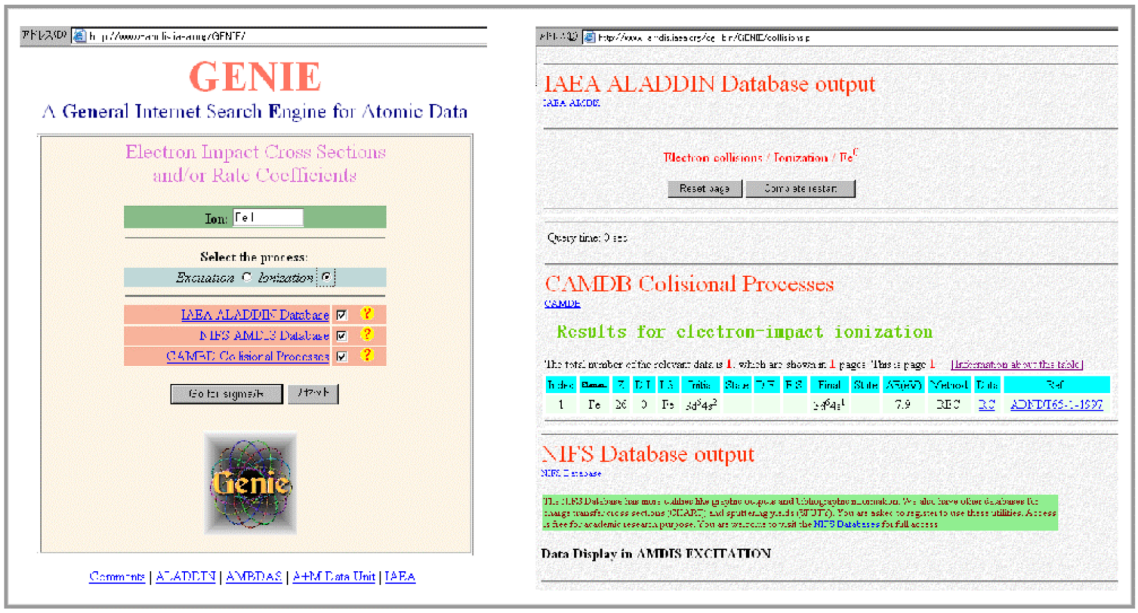

Fig. 3 The General Internet Search Engine for Atomic Data, GENIE, at http://www-amdis.iaea.org/GENIE/. The left panel shows the joint interface for sending queries to IAEA ALLADIN Database, Database of Basic Science Research Center in China, and the NIFS AMDIS Database. The output on the right is a mere merger of the three output pages (details may not resolve in the right panel). The HTML output was edited to fit the format of this figure.

tion or imaging, we have previously developed a two-step procedure. First, only the text of the abstract is analyzed for relevance, preferring recall to precision in this phase of information retrieval. By using the Learning Vector Quantization on simple text with the expressions of A+M terminology discarded (because of formatting ambiguities), it was possible to reach the recall of about 70 % [3]. The specific recognition of A+M terms, and its inclusion in the LVQ algorithm, is a step towards moving the recall of the first phase to values of 90 % and more.

The second step of the information retrieval concerns with the numerical data themselves, and may not be solely achieved based on the analysis of the abstracts [4]. Textual captions of all Figures and Tables in the manuscript are analyzed for simultaneous occurrence of (1) processes and (2) atomic/chemical species; the article is analyzed as relevant if such a simultaneous hit occurs at least once. The combination of the two stages in the automated retrieval of newly published data has been shown to reach 90 % in case of a sample data set, which consisted of ~ 300 articles from Physical Review A. The example of matched A+M expressions in Fig. 2 includes charge states of atomic ions and the angular momentum and spin designation of their electronic states. The Perl script deals with nuances such as iron, FERRUM, Fe for the chemical elements, or He-like Lithium, Li II or Li^+.

3. IAEA GENIE vs. Joint Search 2

In the public domain, there may exist only one cross-database search tool comparable to the Joint Search 2, in particular the IAEA The General Internet Search Engine for Atomic Data tool. Unlike from JS2, which crawls to its target databases over the HTML port 80, GENIE queries the three databases in a predefined way. The use of the standard HTML port by JS2 suffices for the availability of this service while abstaining from any requirements on database contents providers, and is therefore expected to translate into other applications in the field of database unification.

The search output from the IAEA Database, CAMDB and NIFS database collected by GENIE is not semantically analyzed; the resulting pages are divided by line delimiters and merged into a single output page. GENIE has no cataloguing features; the output, on the other hand, includes numerical data formatted according to each of the three databases (skipped in Fig. 3 for the sake of Figure space).

4. Conclusion

The Joint Search 2 tool for simultaneous querying of online databases of A+M abstracts was presented here, including the implementation of a module for A+M terminology extraction and paper relevance assessment. The analogy and complimentary features with the IAEA GENIE tool have also been assessed. It is worth noting that both search engines could in principle implement semantic modules coping with the occasional updates of HTML output pages, HTML query database forms, and even HTML/XML standards by means of backward compatibility. The present work may have added an insight into the database unification field from the viewpoint of the atomic and molecular data field.

Acknowledgements

We would like to thank Dr. Yuri Ralchenko, National Institute of Standards and Technology for useful discussions about the GENIE search engine. Partial support by the Academic Frontier Project and the Core University Program of MEXT for this work are also gratefully acknowledged.

[1] M. Murata, T. Kanamaru, H. Isahara, Proceedings of CICLing 293 (2005).
[2] A. Sasaki, K. Joe, H. Kashiwagi, C. Watanabe, M. Suzuki, L. Pichl, M.Ohishi, D. Kato, M. Kato and T. Kato, J. Plasma Fusion Res. Ser. **7**, 348 (2006).
[3] M. Suzuki, L. Pichl, I. Murakami, T. Kato, A. Sasaki, J. Plasma Fusion Res. Ser. **7**, 343 (2006).
[4] L. Pichl, M. Suzuki, K. Joe and A. Sasaki, Lecture Notes in Computer Science **3433**, 159 (2005).