Rule-Set Evaluation of Article Relevance for Autonomous Bibliography Databases

Lukáš Pichl[1], Manabu Suzuki[2], Daiji Kato[3], Kazuyuki Joe[4], Akira Sasaki[5]

[1] Division of Natural Sciences, International Christian University, Tokyo, 181-8585 Japan

[2]Department of Computer Software, University of Aizu, Ikki, Aizuwakamatsu, 965-8580 Japan

[3]Coordination Research Center, National Institute for Fusion Science, Oroshi-cho, 509-5292 Japan

[4]Department of Information and Computer Sciences, Nara Women University, Nara, 630-8506 Japan

[5]Kansai Research Establishment, Japan Atomic Energy Research Institute, Kyoto-shi, 600-8216, Japan

Abstract

Text relevance assessment in data mining is usually dealt with by various artificial intelligence based methods. This paper focuses on one important case - computerized decision whether a certain scientific article contains numerical data of interest for a specialized research database or not. The area of applications covers activities of most major research data centers, with applications ranging from proteomics to fusion plasma research. As a part of coordinated research on plasma-process bibliography databases at the National Institute for Fusion Science in Japan (NIFS), we have recently developed a linux system that automates the process of data collection, data extraction and database input, with a customizable interface to relevance-assessment software modules. Here we deal with the article relevance assessment from two major viewpoints: a rule-based decision making system based on the analysis of figure and table caption texts, and a machine learning system that analyzes html abstracts of the articles. It is demonstrated that a combination of the two methods may result in a highly specific and sensitive relevance assessment system. Features unique in atomic and molecular data mining are also discussed.

Keywords: article relevance assessment, data mining, learning vector quantization, rule-based methods, autonomous bibliography databases.

1 Introduction

Various data centers over the world, such as the National Institute for Fusion Science (NIFS) in Japan, Oak Ridge National Laboratory in the US or IAEA Data Center in Vienna maintain and expand databases of specialized research data. The production of new data in theory calculations or new measurements is minor in terms of quantity, and most of the new data are obtained from independent sources. To that aim, working groups of specialists are organized that search for, extract and evaluate the quality of new data. Journal articles assessed as relevant are input to bibliography databases and the extracted data are made available in numerical databases. Due to the considerable human labor involved, data extraction and input errors are likely to occur, making the above procedures inefficient and expensive.

To be specific, let use consider a fusion plasma process database as an example. It typically consists of bibliography data, two-dimensional numerical tables (sputtering yields, cross sections, rate constants) and electronic potential surface data (multi-dimensional numerical tables). The 2D-tables need to be displayed routinely on the fly as graphs in linear or logarithmic scales, and pdf, ps, or gif formatted picture download options are available. For obvious reasons, an automatic update of bibliography databases is much easier than that one of numerical databases, especially since the abstracts of all recent journal articles have become freely available online. Still, most articles published before 1990 are available only in hard-copies or as a bitmap-type pdf, subject to processing with optical character recognition software. This paper focuses on relevance assessment of articles for which html abstracts and text-based pdf fultext is available (i.e., a standard at present).

The major source of information is the online databases of scientific journal articles. Their data access is restricted to standard html search interface with pages generated on the fly. Therefore any automated data collection methods across journal publishers must deal with particular formats for web access and conversion of data to a format suitable for machine-based relevance assessment. The basic trade-off in relevance assessment of the text [1] is speed of processing (and data volume) vs. amount of information. If the articles are to be judged based on fultext source (e.g. in pdf files), a huge amount of data needs to be downloaded and processed. If the articles are to be judged based

on the abstract text (and/or keywords, cataloguing numbers etc.), some important information is expected to be missed. Here we discuss both approaches in a NIFS case study, having developed a prototype system of autonomous bibliography database before as a free software open source solution.

The paper is organized as follows. In Section 2, we briefly introduce NIFS data center activity and describe the prototype autonomous bibliography database. Section 3 describes the two major article relevance assessment algorithms along with results and computational benchmarks. Concluding remarks close the paper in Section 4.

2 Autonomous Database

The ultimate goal of computerized database making activities is an automatic systems that searches, downloads, assesses, extracts, evaluates, and inputs new data. Before moving onto article relevance assessment, we describe our prototype of such a system as well as the NIFS database system.

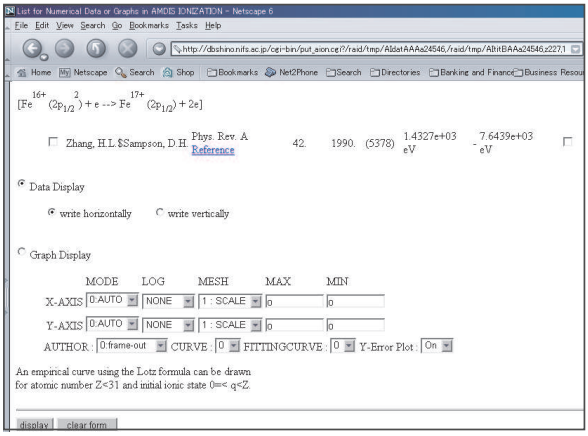

Figure 1: NIFS AMDIS database: bibliography, numerical data and display formats.

Figure 1 shows the database of electron impact processes for atoms and molecules. It displays the formula of an ionization process, $Fe^{16+}(I) + e \rightarrow Fe^{17+}(F) + 2e$ (I, F denote the initial and final electronic state, respectively), a link to the bibliographic entry, settings for possible data display, check-boxes for fitting and semi-empirical curves, and graph options for liner or logarithmic data scale. Because of such complexity in NIFS database entries, data mining of source journal articles is a nontrivial task.

Next, Fig. 2 displays a typical web interface of on-line article database (American Physical Society, APS), the major source for article data mining. Query form fields and the logical operators in this form impose the structure of the web interface for the initial article search.

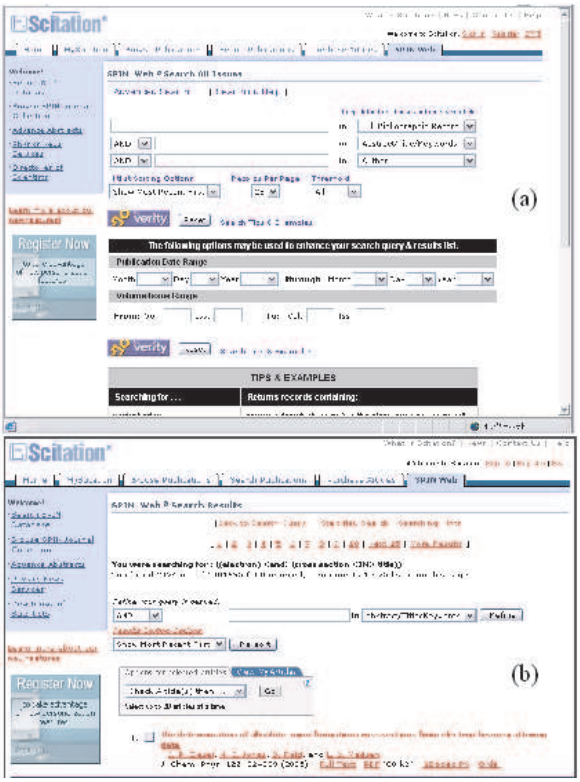

Figure 2: Online SPINWEB article database (American Physical Society) (a), and its web output format (b).

Downloading and scheduling The journal articles are downloaded from on-line publisher websites by using wget, UNIX command line tool. Excessively frequent accesses often results suspensions of the publiser's on-line systems by website protection software. To prevent such a problem, we schedule downloads using wait intervals following the histogram of normal internet connections to the provider. Articles are retrieved in sets corresponding to predesigned queries compatible with SPINWEB interface.

Data extraction and input Once the articles matching a predesigned query are downloaded, database entry items, such as journal name, article title, abstract text, etc. are extracted from the html-formatted text file by using PHP based string matching and input to the database. The extract scripts can be quite lengthy, depending on the web format of each online article repository (cf. Fig. 2b).

Removing duplicates It is also important to avoid duplicate downloads and duplicate inputs. Therefore new articles are searched for only in incremental time periods for each registered base query; accidental hits common to two or more registered queries are automatically removed by

the SQL database management system. In weekly time intervals, queries are automatically sent to the on-line publisher databases and the output is retrieved, analyzed and stored to the database.

Implementation The system described above was implemented using a custom-built PC as a dedicated server. The hardware specification is as follows: Pentium 4 (FSB 533) 2GHz CPU, 1GB RAM and 120GB HDD, which matches the needs of research database users (large data set, limited access rates). The operating system is Linux (Fedora Core 2) running Apache 2.0 HTTP server. The relational database management system is MySQL 3.23.54 with the connecting logic layer for web interface written in PHP 4.2.2. The above operating system, web and database server, and programming language are free-software open-source products.

System integration Figure 3 shows the Linux based database with automated data collection system, Evolutionary Database at NIFS [5], automatically updated once a week. It consists of bibliography classification entries, html-formatted abstracts, and a link to pdf full text of each article on-line. To prevent illegal interference with copyright issues, the pdf files are not stored to the local disk, but their URL location is linked at the external publisher site for download on subscription basis.

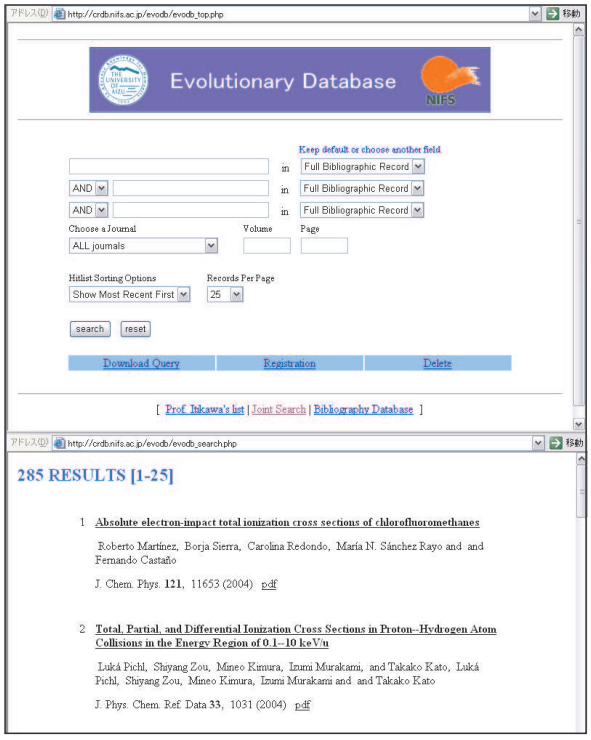

Figure 3: Autonomous database with automatic data retrieval, data extraction, and data input.

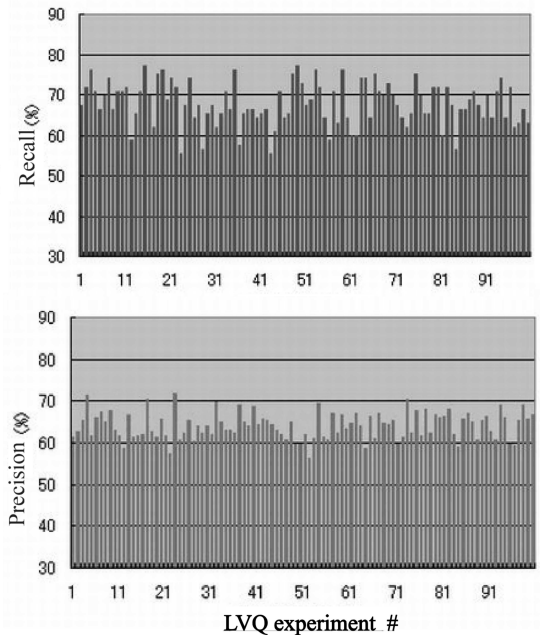

Figure 4: Precision and recall in abstract retrieval with the LVQ algorithm.

3 Relevance Assessment

Journal articles downloaded using the source pool of fixed queries are indexed by their abstracts (html format). Their relevance is assessable by various machine learning algorithms [2]. We adopt the Learning Vector Quantization (LVQ) [4] which is trained to develop feature vectors, vicinity of which represents relevant articles. Each coordinate represents normalized counts of special words or expressions in atomic and molecular physics terminology.

1. Prepare training (T) and evaluation data set (E) - papers and abstracts. Classify T into (1) relevant data (A) and (2) the others (B).

2. Pre-process all abstracts of (A), (B) and (T). The pre-processing phase generates feature vectors from frequency of word roots, technical terms, or high-relevance keyword format.

3. Apply the LVQ algorithm with the feature vectors of abstracts in the sets (A) and (B).

4. Apply the learned reference vectors with the feature vectors of (T) for the recognition of the abstracts in group (E). The result has a binary form: paper is relevant or not (based on the abstract analysis).

Since the LVQ algorithm starts by selecting feature vectors at random, we performed varios abstract retrieval experiments. Using 364 papers as a training set [3], even when removing all special notation of atomic and molecular physics, precision stays within 65% to 75% for different training sets,

Table 1: Rule based method

Total No. of pdf papers:	248
No. of text-formatted pdf files:	167
No. of relevant articles:	64
No. of irrelevant articles:	103
Number of keywords:	92
No. of "process" category keywords:	52
No. of "species" category keywords:	40
No. of retrieved articles:	58
Precision:	100%
Recall (58 out of 64):	90%

and the recall varies between 55% to 80%. The IR metrics is shown for 100 experiments in Fig. 4. Our preliminary results suggest that it is further possible to increase the accuracy of LVQ algorithm by developing a special dictionary of molecular physics to accentuate special coordinates or to combine the LVQ algorithm with rule-based information retrieval already at the stage of abstract assessment.

Next, we proceed to the article fultext. The pdf files are converted with the unix command

> ps2ascii source.pdf output.txt,

then the body text is removed, retaining only figure and table captions. Commercial pdf-to-html software usually fails on APS and other major publisher articles.

The text file extracted from the output of ps2ascii converter is analyzed caption-by-caption for simultaneous occurrence of (1) physical process name from DB-registered list P, and (2) physical species from DB-registered list S. Such a procedure does not require grammar processing and analysis of global meaning, because the captions merely label relevant data. In order to provide rigorous testbed for the rule-based method, we classify the journal articles based on their actual relevance (whether or not these were retrieved in the preceding LVQ step).

Table 1 demonstrates that in the set of 64 fultext articles relevant to NIFS AMIDIS database plus 103 irrelevant articles, the rule-based method achieved 100% precision 90% recall. The size of keyword set was 52 keywords in the "process" category and 40 keywords in the "species" category. We believe that by elaborating on the keyword lists of processes and species that are matched in article captions, and by using a larger set of calibration data, the level of recall can still increase. In addition, the analysis of the 6 papers which failed to be retrieved indicates that loss of formatting in pdf to text conversion was the reason, which is another source of possible improvements.

4 Conclusion

We have developed a 2-step procedure for relevance assessment of articles containing atomic and molecular data. First the abstracts are screened by machine learning algorithms (LVQ at present) with the reproduction rate calibrated as high as possible. Then the pdf fultext is downloaded, converted to ascii, and the text of figure and table caption analyzed for simultaneous occurrence of process and species names. Articles which do not pass this check are assessed as irrelevant. This algorithm is interfaced to a free-software open-source autonomous bibliography database system. The present work is a step towards automatic databases which will identify, collect and input both textual and numerical data into atomic and molecular process database without human intervention. Although the present system focuses on atomic and molecular data in APS journals, applications in related fields can also be expected [6, 7].

Acknowledgement The authors acknowledge a partial support by JSPS Grant-in-Aid. This work is also performed with the support and under the auspices of the NIFS Collaborative Research Program.

References

[1] G. Salton. Automatic Text Processing. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989.

[2] P. E. Utgoff: Incremental induction of decision trees. Machine Learning **4** (1989) 161–186.

[3] Y.Itikawa: ATOMIC DATA AND NUCLEAR DATA TABLES **63** 315-351 (1996).

[4] A.Sasaki, K.Joe, H.Kashiwagi, et al: Design and implementation of an evolutional data collection system for the atomic and molecular databases, Joint ITC14 and ICAMDATA2004 Conferance, Ceratopia Toki, Toki, Gifu, Japan, October 5-8, 2004.

[5] Autonomous DB system
http://crdb.nifs.ac.jp/evodb/

[6] G. Salton, A. Singhal, M. Mitra and C. Buckley: Automatic text structuring and summarization.341-355, Advances in Automatic Text Summarization, edited by I. Mani and M. Maybury, 1999.

[7] H. Chen, A. Lally, B. Zhu, M. Chau: HelpfulMed, Intelligent Searching for Medical Information over the Internet, Journal of the American Society for Information Science and Technology (JASIST), **54** (2003) 683–694.