An Ontology-based Knowledge Management System
- Integrated System of Web Information Extraction and Structuring Knowledge -

Hideki Mima
School of Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan
mima@biz-model.t.u-tokyo.ac.jp

Katsumori Matsushima
School of Engineering
University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033,Japan
matsushima@naoe.t.u-tokyo.ac.jp

Abstract

We will introduce a new web-based knowledge management system in progress, in which XML-based web information extraction and our structuring knowledge technologies are combined using ontology-based natural language processing. Our aim is to provide efficient access to heterogeneous information on the web, enabling users to use a wide range of textual and non textual resources, such as newspapers and databases, effortlessly to accelerate knowledge acquisition from such knowledge sources. In order to achieve the efficient knowledge management, we propose at first an XML-based Web information extraction which contains a sophisticated control language to extract data from Web pages. With using standard XML Technologies in the system, our approach can make extracting information easy because of a) detaching rules from processing, b) restricting target for processing, c) Interactive operations for developing extracting rules. Then we propose a structuring knowledge system which includes, 1) automatic term recognition, 2) domain oriented automatic term clustering, 3) similarity-based document retrieval, 4) real-time document clustering, and 5) visualization. The system supports integrating different types of databases (textual and non textual) and retrieving different types of information simultaneously. Through further explanation to the specification and the implementation technique of the system, we will demonstrate how the system can accelerate knowledge acquisition on the Web even for novice users of the field.

Key Words: information extraction, ontology, terminology, visualization, structuring knowledge, natural language processing, automatic term recognition.

1. Introduction

With the recent dramatic increase of importance of electronic communication and data-sharing over the internet, there exists an increasingly growing number and amount of publicly accessible knowledge sources, both in the form of documents and fact databases.

These knowledge sources available over the Web are intrinsically heterogeneous and dynamic. They are heterogeneous since they are autonomously developed and maintained by independent organizations with different purposes in mind. They are dynamic since constantly new information is being revised, added and removed. Such heterogeneous and dynamic nature of knowledge sources (KSs) in the Web imposes challenges on systems that help users to locate information relevant to their needs. Namely, the growing number of electronically available KSs emphasizes the importance of developing flexible and efficient tools for automatic information extraction (IE) and knowledge management in terms of structuring knowledge.

In this paper, we describe our knowledge management system in progress and we discuss how information extraction from the Web and natural language processing (NLP) based knowledge structuring can be integrated in the system to facilitate effective knowledge mining through the Web.

Conventional approaches on information extraction from the web generally required to write specialized extraction rules or programs because of different formula to each site. Shortcomings of these approaches are: 1) low reusability, 2) extraction must be processed just from whole of a document, and 3) requiring special skills in writing rules / programs. Hence, to resolve the problems, we propose at first an XML-based Web information extraction system. It contains a sophisticated control language to extract data from Web pages. With using standard XML Technologies in the system, our approach can make extracting information easy because of a) detaching rules from processing, b) restricting target for processing, and c) Interactive operations

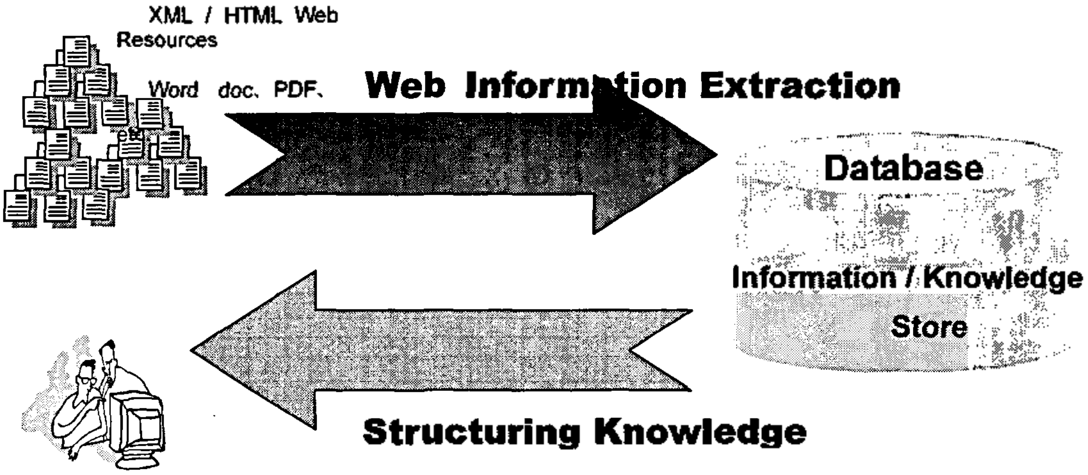

Figure 1: Information extraction and structuring knowledge

for developing extracting rules.

Regarding structuring knowledge, different text mining techniques have been developed recently in order to facilitate efficient discovery of knowledge contained in large textual collections. The main goal of text mining is to retrieve knowledge that is 뱉uried? in a text and to present the distilled knowledge to users in a concise form. Its advantage, compared to 백 anual? knowledge discovery, is based on the assumption that automatic methods are able to process an enormous amount of texts. It is doubtful that any users could process such huge amount of information, especially if the knowledge spans across domains / sites. For these reasons, text mining should aim at helping users in collecting, maintaining, interpreting, curating and discovering knowledge they need in a more efficient and systematic way. Currently, the system includes 1) automatic term recognition and 2) automatic term clustering as automatic ontology[1] development, 3) similarity-based document retrieval, 4) ontology-based real-time document clustering, and 5) visualization. The system supports integrating different types of databases (textual and non textual) and retrieving different types of information simultaneously. Through further explanation to the specification and the implementation technique of the system, we will demonstrate how the system can accelerate knowledge acquisition on the Web even for novice users of the field.

This paper is organized as follows: In section 2, we discuss with related work on Web information extraction and knowledge management, in section 3, we present our knowledge management system including information extraction and ontology-based knowledge mining, in section 4, *C/NC-value*-based terminological processing as an ontology development. In section 5, we explain our visualization scheme for automatic generation of a knowledge map with presenting a preliminary experiment for analyzing news paper articles automatically extracted using the IE system, section 6, we finish with conclusion and future work.

2. Related Work

Since the URLs are often too coarse to locate relevant pieces of information, users have to go through several stages of information seeking activities. After identifying the URLs of the KSs that possibly contain relevant information, they have to locate the relevant pieces of information inside the KSs by using their own navigation functions. This process is often compounded by the fact that users' retrieval requirements can only be met by combining pieces of information in separate databases (or documents). The user has to navigate through different systems that provide their own navigation methods, and has to integrate the results by herself / himself. An ideal knowledge-mining

[1] Although, definition to ontology is domain-specific, our definition to ontology is that the collection and classification of (technical) terms to recognize their semantic relevance.

aid system should provide a seamless transition between the separate stages of information seeking activities, namely, Web information extraction, information / knowledge store and structuring knowledge (Figure. 1).

2.1. Web information extraction

TSIMMIS [16] and TAMBIS [17] share several important properties for improving the transparency in dealing with Web resources. TAMBIS aims to provide a filter from biological information services by building a homogenizing layer on top of the different sources. This layer uses a mediator and many source wrappers to create the illusion of one all encompassing data source. The mediator uses a conceptual knowledge base of molecular biology to describe the universal model and to help users form queries against this universal model expressed in a modeling language. Also it mediates between the various sources to translate the mediators' model to the sources' models. Although TAMBIS like any knowledge rich approach has the possibility to obtain high benefits in certain domain, it is also true that it takes high cost in maintaining the knowledge.

2.2 Terminology management in knowledge mining

Knowledge encoded in textual documents is organized around sets of specialized *terms*. Hence, knowledge acquisition (KA) relies heavily on the recognition of terms. Obviously, a scheme to integrate terminology management as a key prerequisite for knowledge mining is needed.

There are several approaches to automatic term recognition (ATR), especially, in recent biomedicine and molecular biology domain. Some of them rely mainly on linguistic information, namely on morpho-syntactic features of domain terms. For instance, LaSIE [2], an adapted newswire name recognizer, uses a case-sensitive terminology lexicon of component terms, set of morphological cues (biochemical suffixes) and hand-constructed grammar rules in order to recognize terms belonging to specific terminological classes (e.g. enzymes, proteins, etc.). Another example of a rule-based system is PROPER [3], which uses 뱕ore? and 뱕eature? terms to identify strings that correspond to proteins. 뭏 ore? terms are domain-characteristic words (containing capitals, numerals etc.) and 뱕eature? terms are keywords that describe function and characteristic of a term (e.g. protein, receptor, etc.). Recently, hybrid approaches combining linguistic and statistical knowledge are increasingly used ([4, 5]). In order to assess the relevance of extracted term candidates, such methods calculate weights (i.e. termhoods) according to specific statistical measures. Machine learning techniques can be applied as well: for example, [6] presents a statistically based, unsupervised technique to acquire and disambiguate names of proteins, genes, and RNSs.

However, ATR is not the ultimate goal itself. The large number of new terms calls for a systematic way of accessing and retrieving the knowledge represented by them. Accordingly, the extracted terms need to be placed in an appropriate knowledge framework by discovering relations between them, and by establishing links between the terms and different factual databases.

In order to implement terminology-based knowledge structuring, several ontologies have been developed (e.g. MeSH terms, Gene Ontology, GENIA ontology, etc.). Each of them provides a top-down controlled framework, which aims to organize and describe the terminology in the domain. Ontologies implement a pre-defined classification system for terms and their relationships, as well as inference rules that are used to derive knowledge represented by them. However, ontology construction and maintenance are time-consuming activities, as terms are usually manually integrated into an ontology. This is one of the reasons why ontologies typically contain just a subset of existing terminology. In addition, no solution to the well-known difficulties in manual ontology development, such as ontology conflictions / mismatches [7], is provided. Therefore, techniques for automated ontology management [8] are required for efficient and consistent KA.

2.2. Integration of knowledge sources towards structuring knowledge

Different approaches to linking, integrating and interpreting relevant resources have also been suggested. For example, the Semantic Web framework [9] strives to link relevant XML-based resources in a bottom-up manner using the Resource Description Framework (RDF) and ontology

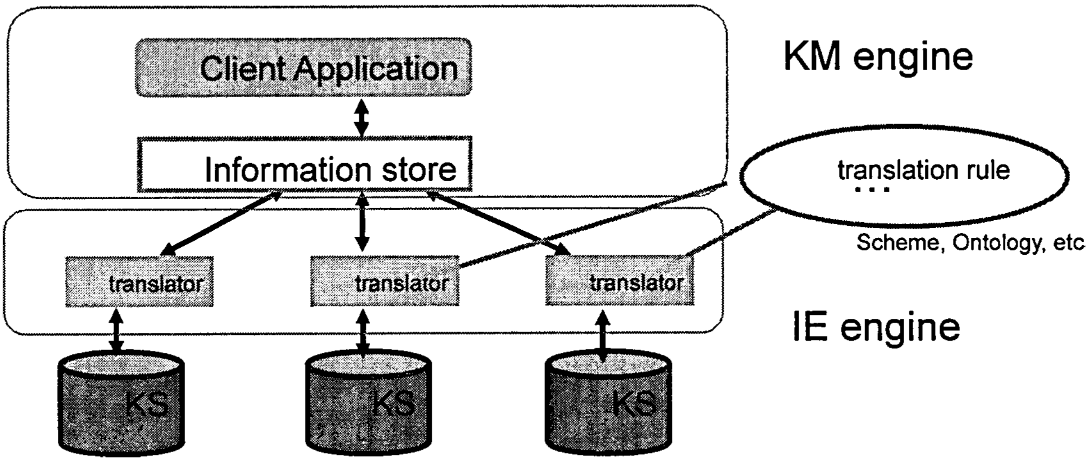

Figure 2: Common Web IE model: translator/wrapper-based approach

information. Since XML allows introduction of new domain- and/or application-specific tags, RDF [10] is used to define their 빜 eanings? and relationships to one another, while the corresponding ontology is used to combine and derive additional information (e.g. synonyms, hyponyms, etc.). In this sense, ontologies are used as a key domain knowledge repository. However, though the Semantic Web framework is powerful when it comes to expressing the content of resources to be semantically retrieved, manual description is needed when defining RDF descriptions and ontologies. If we, however, endeavor to process huge collections of new documents (which cover new knowledge), we need systems that do not rely solely on manual descriptions.

In this paper, we present our approach to Web information extraction, terminology management and mining of knowledge sources adopted in the structuring knowledge system.

3. The System Structure

3.1 Web information extraction

As discussed in the previous section, Web IE requires to have translators or source wrappers to create the illusion of one all encompassing data source, due to the differences in form of multiple, disparate knowledge sources, such as HTML/XML documents and databases, in general (Figure 2).

However, conventional approaches for extracting information from Web resources need to write specialized extraction rules or programs as a translator / wrapper, because of different accesses to each site.

Shortcomings of these approaches are:

1) low reusability due to the difficulty in writing and maintaining them.

2) extraction just from whole of a document.

3) requiring special skills to develop extraction rules / programs..

To resolve the problem, our XML-based Web information extraction system contains a language for controlling processes (flow rules) and a language for extracting data from Web pages (extraction rules). With standard XML Technology, our system can make extracting information easy because of a) detaching rules from processing, b) restricting target for processing, c) operating interactively.

The language for flow that we developed is one of XML subset languages. If you put up tags such as in HTML Documents, you can control many processes, for instance, crawling Web sites, tuning HTML sources to well-formatted ones, extracting data interested and exporting by other format). The language for extraction use XSL Transformations (XSLT).XSLT usually needs to restrict target for processing, so results on the way need to be expected each time. On the contrarily,

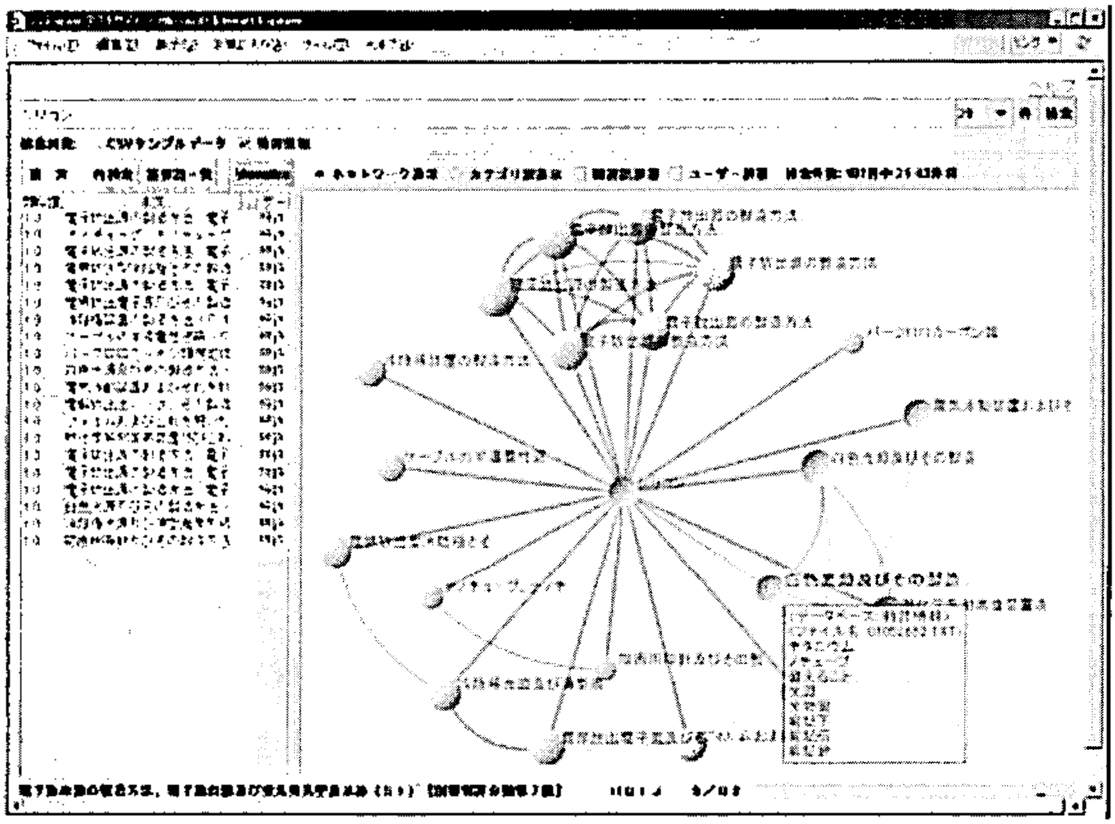

Figure 5: Visualization sample

innovation and engineering. In order to structure knowledge, the system draws a graph in which nodes indicate relevant KSs to keywords specified by the user and each links among KSs indicates semantic similarities calculated using ontology information developed by our ATR / ATC components, which is based on comparing ontological information extracted from each KSs, whereas conventional similarity calculation is generally based on nouns extracted from each KSs. Also, in drawing the graph, locations of each node are calculated and optimized based on the condition: the closer they are in meaning, the closer they are in location.

Also, Knowledge map generation is achieved by 1) cluster recognition, 2) terminology-based categorization. Thesaurus and SVM -based categorizer is provided to categorize clustered KSs automatically. Figure 6 shows an experimental result of automatic knowledge map generation for the latest news paper articles. In the experiment, the articles were extracted automatically from Yomiuri and Mainichi news paper Web sites (both English and Japanese) using our Web IE system and were stored into the database. Specified keywords for structuring knowledge were 뱑raq? and 뱚allujah?

As can be seen in the figure, seven clusters were recognized and the assigned topics (concepts) ware (1) Bin Ladin, (2) secretary of state Powell, (3) dispatch of the Japanese self-defense forces, (4) presidential election, (5) Samawah, (6) the prime minister Koizumi, and (7) the prime minister Allawi, and they were thought to be all important topics regarding the specified keywords. Categorization and mapping concepts are generally recognized as a basic method to accelerate understanding of information (knowledge acquisition). Thus, we can expect our proposed scheme is feasible enough for accelerating knowledge acquisition. Furthermore, the method is able to provide possibility to disambiguate semantic ambiguity (polysemy) of specified keywords. For example, keyword 뱎pple? includes at least two meanings as 1) fruit, and 2) computer company. However, if you obtain clustered and categorized IR results, you can find intended information more easily.

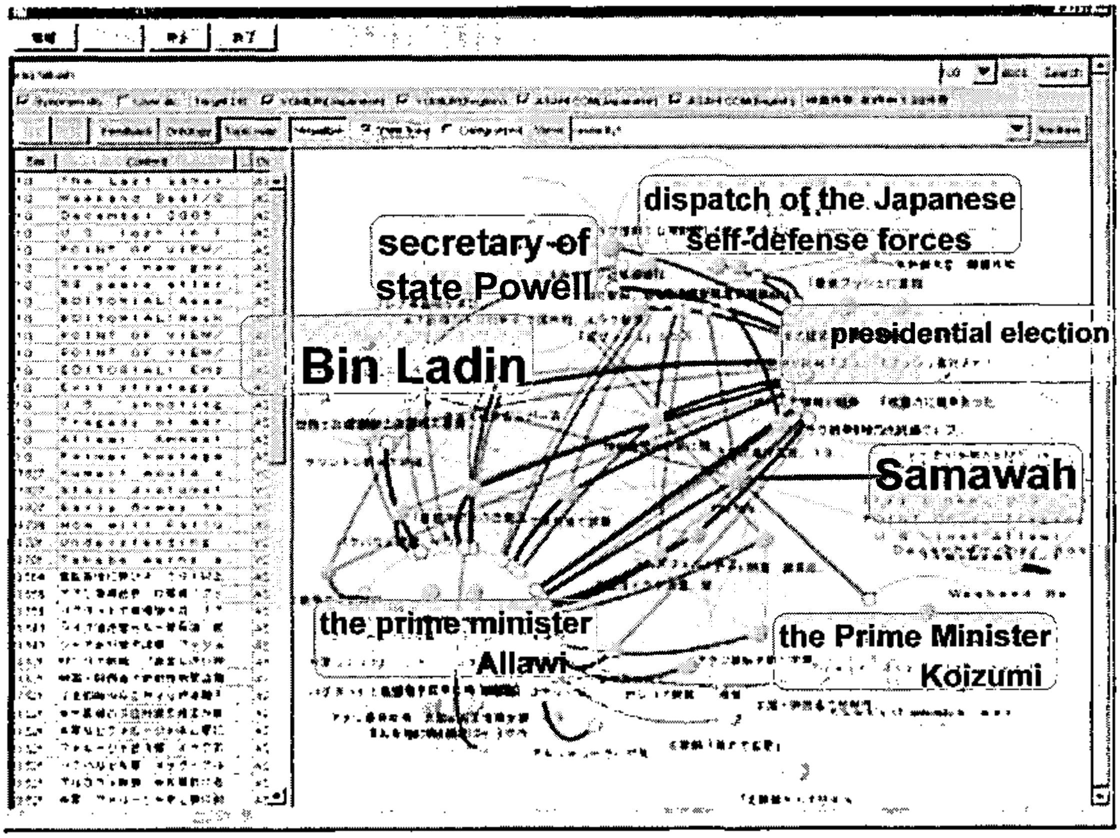
dispatch of the Japanese Self-defense forces
secretary of state Powell
presidential election
Bin Ladin
Samawah
the prime minister Allawi
the Prime Minister Koizumi

Figure 6: Knowledge map generation result

6. Conclusion

In this paper, we presented an integrated knowledge mining system, in which we integrate Web information extraction, automatic term recognition, term clustering, information retrieval, and visualization. The main objective of the system is to facilitate knowledge acquisition from Web documents and new knowledge discovery through a terminology-based similarity calculation and a visualization of automatically structured knowledge. Also, to accelerate knowledge discovery, we presented a visualization method for similarity-based knowledge map generation. The method is based on real-time ontology-based knowledge clustering and categorization and allows users to show automatically generated knowledge map in real-time. An experimentation we conducted using news paper sites show that we can expect the method is practical enough for accelerating knowledge acquisition / new knowledge discovery from existing knowledge sources.

Important areas of future research will involve usability evaluation for the system. Further, we will investigate the possibility of using a term classification system as an alternative structuring model for knowledge deduction and inference (instead of an ontology).

References

[1] National Library of Medicine, MEDLINE, www.ncbi.nlm.nih.gov/PubMed/, 2002.

[2] R. Gaizauskas, G. Demetriou, K. Humphreys, Term recognition and classification in biological science journal articles, Proc. of Workshop on Computational Terminology for Medical and Biological Applications, NLP-2000, Patras, Greece, 2000, pp. 37-44.

[3] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, Toward information extraction: identifying protein names from biological papers, Proc. of PSB-98, Hawaii, 1998, pp. 3:705-716.

[4] H. Mima, S. Ananiadou, G. Nenadic, ATRACT workbench: an automatic term recognition and clustering of terms, in: V. Matoušek, P. Mautner, R. Mouček, K. Taušer (Eds.) Text, Speech and Dialogue, LNAI 2166, Springer Verlag, 2001, pp. 126-133.

[5] H. Mima, S. Ananiadou, An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese, Int. J. on Terminology 6/2 (2001), pp. 175-194.

[6] V. Hatzivassiloglou, P. Duboue, A. Rzhetsky, Disambiguating proteins, genes, and RNA in text: a machine learning approach, in: BIOINFORMATICS 17/1 (2001), pp. S97-S106.

[7] P.R.S. Visser, D.M. Jones, T.J.M. Bench-Capon, M.J.R. Shave, An analysis of ontology mismatches - heterogeneity versus interoperability, Proc. of AAAI 1997 Spring Symposium on Ontological Engineering, Stanford University, California, USA, 1997, pp. 164-172.

[8] J. Gamper, W. Nejdl, M. Wolpers, Combining Ontologies and Terminologies in Information Systems, Proc. of the 5th International Congress on Terminology and Knowledge Engineering, Innsbruck, Austria, 1999, pp. 152-168.

[9] T. Berners-Lee, The semantic Web as a language of logic, available at: www.w3.org/DesignIssues/Logic.html, 1998.

[10] D. Brickle, R. Guha, Resource description framework (RDF) schema specification 1.0, W3C Candidate Recommendation, available at http://www.w3.org/TR/rdf-schema, 2000.

[11] P.G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, R. Stevens, TAMBIS: transparent access to multiple bioinformatics information sources - an overview, Proc. of 6th International Conference on Intelligent Systems for Molecular Biology - ISMB98, Montreal, 1998, pp. 25-34.

[12] A. Voutilainen, J. Heikkila, An English Constraint Grammar (ENGCG) a surface-syntactic parser of English, in: U. Fries et al. (Eds.) Creating and Using English language corpora, Rodopi, Amsterdam, Atlanta, 1993, pp. 189-199.

[13] M. Krauthammer, A. Rzhetsky, P. Morozov, C. Friedman, Using BLAST for identifying gene and protein names in journal articles, in: Gene 259 (2000), pp. 245-252.

[14] C. Jacquemin, Spotting and discovering terms through NLP, MIT Press, Cambridge MA, 2001, p. 378.

[15] A. Ushioda, Hierarchical clustering of words, Proc. of COLING '96, Copenhagen, Denmark, 1996, pp. 1159-1162.

[16] H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and Jennifer Widom. "Integrating and Accessing Heterogeneous Information Sources in TSIMMIS". In *Proceedings of the AAAI Symposium on Information Gathering*, pp. 61-64, Stanford, California, March 1995.

[17] Baker P. G., Brass A., Bechhofer S., Goble C., Paton N. and Stevens R. 1998. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview in Proc. of the Sixth International Conference on Intelligent Systems for Molecular Biology, ISMB98, Montreal.

Biography:

Hideki Mima, Ph.D., has worked in the area of Natural Language Interface, Machine Translation, Information Retrieval and Automatic Term Recognition. He was a researcher at the ATR Interpreting Telecommunications Research Laboratories, a lecturer at the Department of Computing and Mathematics, Manchester Metropolitan University, and a research associate at the Department of Information Science, University of Tokyo, Japan. Currently, he is a research associate at the School of Engineering, University of Tokyo and is working on Knowledge Acquisition and Knowledge Structuring from various databases/documents in the genome / nano-technology domains. E-mail: mima@biz-model.t.u-tokyo.ac.jp, URL: http://www.biz-model.t.u-tokyo.ac.jp/users/mima/.