Design and Implementation of an Ontology-based Knowledge Management System

Hideki Mima* Taesung Yoon** Katsumori Matsushima*
*School of Engineering, University of Tokyo
**Open Knowledge Corp.

Abstract

The purpose of the study is to develop an integrated knowledge management system for the domains of genome and nano-technology, in which terminology-based literature mining, knowledge acquisition, knowledge structuring, and knowledge retrieval are combined. The system supports integrating different types of databases (papers and patents, technologies and innovations) and retrieving different types of knowledge simultaneously. The main objective of the system is to facilitate knowledge acquisition from documents and new knowledge discovery through a terminology-based similarity calculation and a visualization of automatically structured knowledge. Implementation issue of the system is also mentioned.

Key Words: Knowledge structuring, knowledge management, information extraction, natural language processing, automatic term recognition, terminology

1. Introduction

New scientific discoveries result in an abundance of documents, such as scientific papers and patents, verbalizing these discoveries. These documents are created in an attempt to share new knowledge with other scientists. They are often reproduced in electronic form and placed on the Internet or other types of shared resources in order to make the new information widely and easily available. Electronically available texts are continually being created and updated, and, thus, the knowledge represented in such texts is more up-to-date than in any other knowledge media.

The sheer amount of published papers[1] makes it difficult for a human to efficiently localize the information of interest not only in a collection of documents, but also within a single document. The growing number of electronically available knowledge sources (KSs) emphasizes the importance of developing flexible and efficient tools for automatic knowledge acquisition and structuring in terms of knowledge integration. Different text and literature mining techniques have been developed recently in order to facilitate efficient discovery of knowledge contained in large textual collections. The main goal of literature mining is to retrieve knowledge that is "buried" in a text and to present the distilled knowledge to users in a concise form. Its advantage, compared to "manual" knowledge discovery, is based on the assumption that automatic methods are able to process an enormous amount of texts. It is doubtful that any researcher could process such huge amount of information, especially if the knowledge spans across domains. For these reasons, literature mining aims at helping scientists in collecting, maintaining, interpreting and curating information.

One of the main problems when processing a collection of KSs is their heterogeneity and dynamic nature. Even when confined to a single domain, the KSs are autonomously developed and maintained by independent organizations for different purposes, hence resulting in a *heterogeneous* set of KSs. Moreover, this set is *dynamic* as a result of continuous attempts to synchronize its content with up-to-date knowledge. New information is being added and existing information is revised and often removed from the KSs. These two facts, heterogeneity and constant evolution of KSs, set a challenge to systems designed to assist users in locating and integrating knowledge relevant to their needs.

In this study, we develop an integrated knowledge structuring (KS) system, in which terminology-based literature mining, terminology-driven knowledge acquisition (KA), knowledge integration (KI), and knowledge retrieval (KR) are combined using automatic term recognition, automatic term clustering and terminology-based similarity calculation. The system incorporates automatic term recognition / clustering and a visualization of retrieved knowledge based on the terminology, which allow users to access KSs visually though sophisticated GUIs.

[1] For example, the MEDLINE database [1] currently contains over 12 million abstracts in the domains of molecular biology, biomedicine and medicine, growing by more than 40.000 abstracts each month.

2. An overview of the system

The KS system has been developed with the intention to address the problems of the ontology-driven literature mining and KA. Similarly to the Semantic Web framework, our system deals with XML documents by using domain-specific RDF descriptions and ontology-based inference. However, it facilitates KA tasks not only by using manually defined resource

descriptions, but also by exploiting natural language processing techniques such as ATR and automatic term clustering (ATC), which are used for automatic population of the underlying ontology. Additionally, the system integrates an

information retrieval engine and a similarity calculation engine that allow users to show not only relevant KSs to keywords but also relevance between KSs.

The system acts as an information extraction engine, which is based on managing XML tag information obtained from its subfunctional components. Typically, IE-based KA process within the system has the following course: first, a collection of documents is linguistically processed (part-of-speech (POS) tagging, shallow parsing, etc.). Further, the collection is terminologically analyzed, i.e. relevant domain-specific terms are automatically recognized and structured (classified or incorporated into an ontology).

The system architecture is modular, and it integrates the following components (Figure 1):

- *Ontology Development Engine(s) (ODE)* – components that carry out the automatic ontology development which includes recognition and structuring of domain terminology;
- *Tag Data Manager (TDM)* – stores index of KSs and tag information in a tag information database (TID) and provides the corresponding interface;
- *Knowledge Retriever (KR)* – retrieves KSs from TID and calculates similarities between keywords and KSs. Currently, we adopt *tf*idf* based similarity calculation;
- *Similarity Calculation Engine(s) (SCE)* – calculate similarities between KSs provided from KR component in order to show semantic similarities between each KSs.
- *Graph Visualizer* – visualizes knowledge structures based on graph expression in

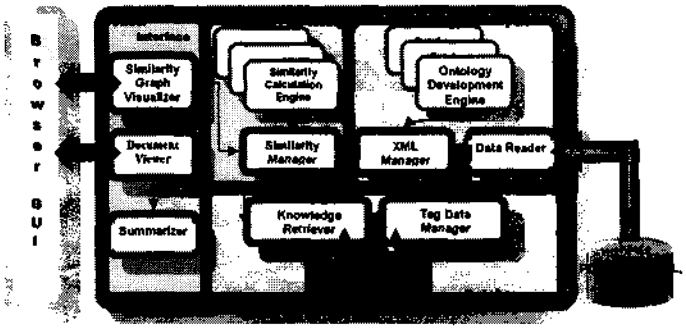

Figure 1: The system architecture

which relevance links between provided keywords and KSs, and relevance links between the KSs themselves can be shown.

Linguistic pre-processing within the system is performed in two steps. In the first step, POS tagging[2], i.e. the assignment of basic parts of speech (e.g. noun, verb, etc.) to words, is performed. In the second step, an ontology development engine is used to perform ATR and ATC. We also used feature structure-based parsing for English and Japanese for linguistic filter of the ATR.

3. Terminological processing as an ontology development

The lack of clear naming standards in a domain (e.g. biomedicine) makes ATR a non-trivial problem [2]. Also, it typically gives rise to many-to-many relationships between terms and concepts. In practice, two problems stem from this fact: the same term may denote a number of concepts, and, conversely, the same concept may be denoted by more than one term. In other words, there are terms that have multiple meanings (*term ambiguity*), and, conversely, there are terms that refer to the same concept (*term variation*). Generally, term ambiguity has negative effects on IE precision, while term variation decreases IE recall.

These problems point out the impropriety of using simple keyword-based IE techniques. Obviously, more sophisticated techniques are needed. Such techniques should identify groups of different terms referring to the same (or similar) concept(s), and, therefore, could benefit from relying on efficient and consistent ATR/ATC and term variation management methods. These methods are also important for organising

[2] We use EngCG tagger in English and JUMAN / Chasen morphological analyzers in Japanese.

domain specific knowledge, as terms should not be treated isolated from other terms. They should rather be related to one another so that the relations existing between the corresponding concepts are at least partly reflected in a terminology.

Terminological processing in our system is carried out based on C / NC-value method [4] for ATR, and average mutual information based ATC. Its main purpose is to help domain experts in gathering and managing domain-specific terminology. It is used to automatically retrieve and cluster terms.

3.1. Term recognition

The ATR method used in the system is based on the C- and NC-value methods [3]. The *C-value* method recognizes terms by combining linguistic knowledge and statistical analysis. The method extracts multi-word terms[3] and is not limited to a specific class of concepts. It is implemented as a two-step procedure. In the first step, term candidates are extracted by using a set of linguistic constraints, implemented using a LFG-based GLR parser, which describe general term formation patterns. In the second step, the term candidates are assigned termhoods according to a statistical measure. The measure amalgamates four numerical corpus-based characteristic of a candidate term, namely the frequency of occurrence, the frequency of occurrence as a substring of other candidate terms, the number of candidate terms containing the given candidate term as a substring, and the number of words contained in the candidate term.

The *NC-value method* further improves the C-value results by taking into account the context of candidate terms. The relevant context words are extracted and assigned weights based on how frequently they appear with top-ranked term candidates extracted by the C-value method. Subsequently, context factors are assigned to candidate terms according to their co-occurrence with top-ranked context words. Finally, new termhood estimations, referred to as NC-values, are calculated as a linear combination of the C-values and context factors for the respective terms. Evaluation of the C/NC-methods has shown that contextual information improves term distribution in the extracted list by placing real terms closer to the top of the list[3][4].

3.2. Term variation management

Term variation and ambiguity are causing problems not only for ATR but for human experts as well. Several methods for term variation management have been developed. For example, the BLAST system [6] used approximate text string matching techniques and dictionaries to recognize spelling variations in gene and protein names. FASTR [7] handles morphological and syntactic variations by means of meta-rules used to describe term normalization, while semantic variants are handled via WordNet.

The basic C-value method has been enhanced by term variation management [3]. We consider a variety of sources from which term variation problems originate. In particular, we deal with orthographical, morphological, syntactic, lexico-semantic and pragmatic phenomena. Our approach to term variation management is based on term normalization as an integral part of the ATR process. Term variants (i.e. synonymous terms) are dealt with in the initial phase of ATR when term candidates are singled out, as opposed to other approaches (e.g. FASTR handles variants subsequently by applying transformation rules to extracted terms). Each term variant is normalized (see table 1 as a simple example) and term variants having the same normalized form are then grouped into classes in order to link each term candidate to all of its variants. This way, a list of normalized term candidate classes, rather than a list of single terms is statistically processed. The termhood is then calculated for a whole class of term variants, not for each term variant separately.

Table 1: Term normalisation example

Term variants		Normalised term
human cancers cancer in humans human's cancer human carcinoma	→	human cancer

3.3. Term clustering

Beside term recognition, term clustering is an indispensable component of the literature mining process. Since terminological opacity and polysemy are very common in molecular biology and biomedicine, term clustering is essential for the semantic integration of terms, the construction of domain ontologies and semantic tagging.

[3] More than 85% of domain-specific terms are multi-word terms [4].

ATC in our system is performed using a hierarchical clustering method in which clusters are merged based on average mutual information measuring how strongly terms are related to one another [8]. Terms automatically recognized by the NC-value method and their co-occurrences are used as input, and a dendrogram of terms is produced as output. Parallel symmetric processing is used for high-speed clustering. The calculated term cluster information is encoded and used for calculating semantic similarities in SCE component. More precisely, the similarity between two individual terms is determined according to their position in a dendrogram. Also a commonality measure is defined as the number of shared ancestors between two terms in the dendrogram, and a positional measure as a sum of their distances from the root. Similarity between two terms corresponds to a ratio between commonality and positional measure.

Further details of the methods and their evaluations can be referred in [3][4].

4. Knowledge Management and Structuring Knowledge

Literature mining can be regarded as a broader approach to IE/KA. IE and KA in our system are implemented through the integration of tag- and ontology-based IE and semantic similarity calculation. Graph-based visualization for globally structuring knowledge is also provided to facilitate KR and KA from documents. Additionally, the system supports combining different types of databases (papers and patents, technologies and innovations) and retrieves different types of knowledge simultaneously and crossly. This feature can accelerates knowledge discovery by combining existing knowledge. For example, discovering new knowledge on industrial innovation by structuring knowledge of trendy scientific paper database and past industrial innovation report database can be expected. Figure 2 shows an example of visualization of knowledge structures in the domain of innovation and engineering. In order to structure knowledge, the system draws a graph in which nodes indicate relevant KSs to keywords given and each links between KSs indicates semantic similarities dynamically calculated using ontology information developed by our ATR / ATC components.

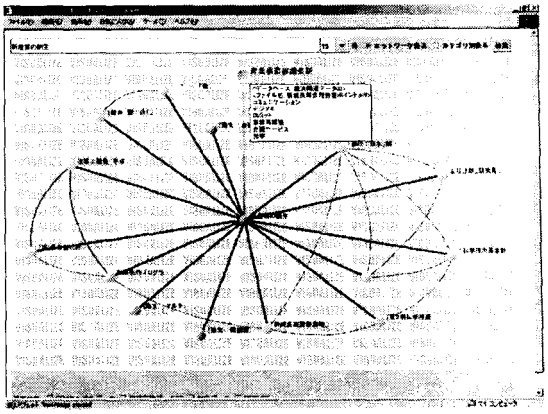

Figure 2: Visualization sample

5. Conclusion

In this paper, we presented a system for knowledge management over large KSs. The system is a terminology-based integrated KA system, in which we have integrated ATR, ATC, IR, similarity calculation, and visualization for knowledge structuring. It allows users to search and combine information from various sources. KA within the system is terminology-driven, with terminology information provided automatically. Similarity based knowledge retrieval is implemented through various semantic similarity calculations, which, in combination with hierarchical, ontology-based matching, offers powerful means for KA through visualization-based literature mining.

Important areas of future research will involve integration of a manually curated ontology with the results of automatically performed term clustering. 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

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