Design and Implementation of a Terminology-based Literature Mining and Knowledge Structuring System

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

Sophia Ananiadou

School of Computing, Science and Engineering, University of Salford, Salford M5 4WT, UK

National Centre for Text Mining

S.Ananiadou@salford.ac.uk

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

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 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 of automatically visualization knowledge. Implementation issues of the system are also mentioned.

Key Words: Structuring knowledge, knowledge acquisition, information extraction, natural language processing, automatic term recognition, terminology

1. Introduction

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.

In this paper, we introduce a knowledge

integration and structuring system (KISS) we designed, in which terminology-driven knowledge acquisition (KA), knowledge retrieval (KR) and knowledge visualization (KV) are combined using automatic term recognition, automatic term clustering and terminology-based similarity calculation is explained. The system incorporates our proposed automatic term recognition / clustering and a visualization of retrieved knowledge based on the terminology, which allow users to access KSs visually though sophisticated GUIs.

2. Overview of the system

The main purpose of the knowledge structuring system is 1) accumulating knowledge in order to develop huge knowledge bases, 2) exploiting the accumulated knowledge efficiently. Our approach to structuring knowledge is based on:

- automatic term recognition (ATR)
- automatic term clustering (ATC) as an ontology¹ development
- ontology-based similarity calculation
- visualization of relationships among documents (KSs)

of our definitions to structuring knowledge is discovery of relevance between documents (KSs) and its visualization. In order to achieve real time processing for structuring knowledge, we terminology / ontology-based similarity calculation, because knowledge can also be represented as textual documents passages (e.g. sentences, subsections) which are efficiently characterized by sets of specialized (technical) terms. Further details of our visualization scheme mentioned in Section 4.

¹ Although, definition of ontology is domainspecific, our definition of ontology is the collection and classification of (technical) terms to recognize their semantic relevance.

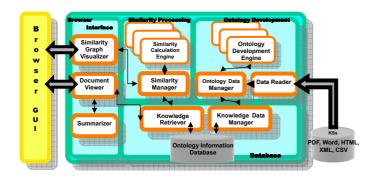


Figure 1: The system architecture

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;
- Knowledge Data Manager (KDM) stores index of KSs and ontology in a ontology information database (OID) 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 using ontology developed by ODE in order to show semantic similarities between each KSs. Semantic clusters of KSs are also provided.
- Graph Visualizer visualizes knowledge structures based on graph expression in 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 tagging², 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 а domain (e.g. biomedicine) makes ATR a nontrivial problem [1]. Also, it typically to many-to-many rise relationships between terms and concepts. In practice, two problems stem from this fact: 1) there are terms that have multiple meanings (term ambiguity), and, conversely, 2) 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 These problems point out the recall. impropriety of using simple keyword-based techniques. Obviously, more sophisticated techniques, identifying groups of different terms referring to the same (or similar) concept(s), and, therefore, could benefit from relying on efficient consistent ATR/ATC and term variation management methods are required. These methods are also important for organising 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 [2,3] for ATR, and average mutual information based ATC (Figure 2).

3.1. Term recognition

The ATR method used in the system is based on the C / NC-value methods [2,3]. The C-value method recognizes terms by combining linguistic knowledge statistical analysis. The method extracts multi-word terms³ 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 filters, implemented using a LFG-based GLR parser, which describe general term formation patterns. In the second step, the term candidates assigned termhoods (referred to as C-values) according to a statistical measure. The measure amalgamates four numerical corpus-based characteristics of a candidate

² We use EngCG tagger[4] in English and JUMAN / Chasen morphological analyzers in Japanese.

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

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 cooccurrence 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 [3] has shown that contextual information improves term distribution in the extracted list by placing real terms closer to the top of the list.

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 [5] used approximate text string matching techniques and dictionaries to recognize spelling variations in gene and protein names. FASTR [6] handles morphological and syntactic variations by means of metarules used to describe term normalization, while semantic variants are handled via WordNet.

The basic C-value method has been enhanced by term variation management [2]. 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 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 an example) and term variants having the same normalized form are then grouped into classes in order to link each term candidate

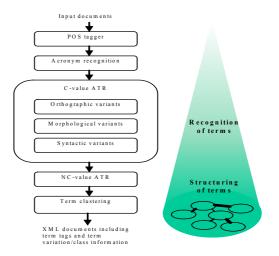


Figure 2: Ontology development

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: Automatic term normalization

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.

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 [7]. 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. precisely, the similarity between 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 [2,3].

4. 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 ATR, ATC, and ontologybased 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 databases (papers and patents, technologies and innovations) and retrieves different types of knowledge simultaneously and crossly. This feature can accelerate 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 3 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. Since characterization for KSs using terminology is thought to be ultimate most efficient and summarization to KSs, achieving a fast and just-in-time processing for structuring knowledge can be expected.

5. Conclusion

In this paper, we presented a system for literature mining and knowledge structuring 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 structuring knowledge. 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, combination with hierarchical, ontology-

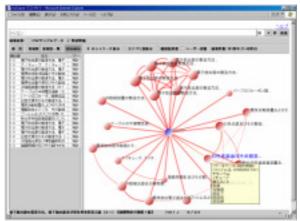


Figure 3: Visualization

based matching, offers powerful means for KA through visualization-based literature mining.

Preliminary experiments we conducted show that the system's knowledge management scheme is an efficient methodology to facilitate KA and new knowledge discovery in the field of genome and nano-technology[2].

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] 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.
- [2] 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.
- [3] 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.
- [4] A. Voutilainen, J. Heikkila, An English Constraint Grammar (ENGCG) a surfacesyntactic parser of English, in: U. Fries et al. (Eds.) Creating and Using English language corpora, Rodopi, Amsterdam, Atlanta, 1993, pp. 189-199.
- [5] 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.
- [6] C. Jacquemin, Spotting and discovering terms through NLP, MIT Press, Cambridge MA, 2001, p. 378.
- [7] A. Ushioda, Hierarchical clustering of words, Proc. of COLING '96, Copenhagen, Denmark, 1996, pp. 1159-1162.