Terminology-Based Knowledge Mining for New Knowledge Discovery

HIDEKI MIMA
School of Engineering, University of Tokyo
SOPHIA ANANIADOU
School of Informatics, University of Manchester
and
KATSUMORI MATSUSHIMA
School of Engineering, University of Tokyo

In this article we present an integrated knowledge-mining system for the domain of biomedicine, in which automatic term recognition, term clustering, information retrieval, and visualization are combined. The primary objective of this system is to facilitate knowledge acquisition from documents and aid knowledge discovery through terminology-based similarity calculation and visualization of automatically structured knowledge. This system also supports the integration of different types of databases and simultaneous retrieval of different types of knowledge. In order to accelerate knowledge discovery, we also propose a visualization method for generating similarity-based knowledge maps. The method is based on real-time terminology-based knowledge clustering and categorization and allows users to observe real-time generated knowledge maps, graphically. Lastly, we discuss experiments using the GENIA corpus to assess the practicality and applicability of the system.

Categories and Subject Descriptors: I.2 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis*; H.5 [**Information Interfaces and Presentation**]: User Interfaces - *Natural language*; H.3.1 [**Information Systems**]: Content Analysis and Indexing - *Linguistic processing*; I.7 [**Document and Systems**]: Document Capture - *Document analysis*
General Terms: Design, Experimentation, Algorithms
Additional Key Words and Phrases: Automatic term recognition, biomedicine, natural language processing, structuring knowledge, terminology, visualization

1. INTRODUCTION

New scientific discoveries result in the creation of an abundance of documents (textual and non-textual), such as scientific papers, patents, and fact databases, that verbalize these discoveries, and are created to share new knowledge with other scientists. However, such a large volume of published documents[1] makes it difficult for a person to efficiently

Hideki Mima, School of Engineering University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan, Sophia Ananiadou, School of Informatics, University of Manchester, PO Box 88, M60 1QD, Manchester, UK, and National Centre for Text Mining, Katsumori Matsushima, School of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

[1] For example, the MEDLINE database [MEDLINE 2002] currently contains over 14 million references in the domains of molecular biology, biomedicine, and medicine that are increasing at a rate of more than 40.000 abstracts per month.

localize 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 order to integrate knowledge.

Different text mining techniques have been recently developed in order to facilitate knowledge discovery from large textual collections. The primary goal of text mining is to retrieve knowledge that is "buried" in text and to present the distilled knowledge to users in a concise form. As compared to "manual" knowledge discovery, the advantage of this technique is the assumption that automatic methods will enable the processing of enormous amounts of text. It is impossible for any researcher to process such huge amounts of information, particularly when the knowledge spans domains. Text-mining enables scientists to efficiently and systematically collect, maintain, interpret, curate, and discover knowledge for research or education.

The central challenges in processing a collection of KSs are its heterogeneity and dynamic nature. Even when confined to a single domain, the KSs are developed autonomously and maintained by independent organizations for different purposes, resulting in a *heterogeneous* set of KSs. One of the main challenges when text-mining in highly dynamic areas like biomedicine is identifying the terms that are key in accessing the information stored in KSs. The terms and their associations will convey knowledge across scientific domains. Terms (e.g., gene names, proteins, gene products, organisms, drugs, chemical compounds, etc.) enable scientific communication. Additionally, they are the linguistic realization of specialized concepts. It is not possible to "understand" an article and extract information from it without precise identification and association of terms. New terms are introduced in the domain vocabulary on a daily basis, and given the number of new names introduced around the world, it is practically impossible to maintain up-to-date terminologies that are manually produced, maintained, and standardized. For example, various curatorial teams had to identify terminologies in order to integrate them into special databases (such as Swiss-Prot,[2] SGD,[3] FlyBase,[4] and UniProt[5]). Curatorial teams maintain terminological resources; however, the integration of new terms is difficult and is not based on systematic extraction and collection of terminology from literature. In addition, since some terms appear frequently and some of them do not last for long, existing terms are frequently altered or discarded (obsolete terms).

For the domains of biomedicine, we introduce an integrated knowledge-mining system, which combines automatic term recognition, term clustering, information retrieval, and visualization. The main objective of this system is to facilitate the acquisition of knowledge from documents and discover new knowledge by calculating the similarities based on terminology and by the visualization (graphically drawn knowledge maps) of automatically structured knowledge. This system also supports the integration of different types of databases (textual and non-textual) and the simultaneous retrieval of different types of knowledge. In order to accelerate knowledge discovery, we propose a visualization method for generating similarity-based knowledge maps. This method is based on real-time terminology-based knowledge clustering and categorization, and it allows users to *graphically observe* knowledge maps generated in real time. Lastly,

[2] http://www.ebi.ac.uk/swissprot
[3] http://www.yeastgenome.org/
[4] http://www.flybase.org
[5] http://www.ebi.ac.uk/GOA/

we discuss experiments that were conducted using the GENIA corpus [GENIA Project 2002] in order to assess the practicality of the system.

2. RELATED WORK

2.1 Terminology Management

The knowledge encoded in textual documents is organized around sets of specialized *terms*. Hence, knowledge acquisition (KA) relies heavily on the recognition of terms. A prerequisite for knowledge mining is terminology management, which includes automatic term extraction, clustering, and classification.

Recently, several approaches for automatic term recognition (ATR) applicable to biomedicine have been introduced. Rule-based approaches primarily rely on linguistic information, namely, morpho-syntactic features of terms. For example, LaSIE [Gaizauskas et al. 2000], an adaptive newswire name recognizer, uses a case sensitive terminology lexicon of component terms, a 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 is PROPER [Fukuda et al. 1998], which relies on simple lexical patterns and orthographic features for protein name recognition. PROPER (PROtein Proper noun phrase Extracting Rules)[6] uses "core" and "feature" terms to identify strings that correspond to proteins.

Core terms are domain-characteristic words that reflect the core meaning (containing, e.g., capitals, numerals); *feature terms* are keywords that describe the terms' function and characteristics (e.g., protein, receptor). Fukuda et al. [1998] reported results with a precision of 94.7% at a recall of 98.8%.

A variety of machine learning and statistical techniques are used for ATR. Machine-learning systems rely on the existence of training data to learn features that can be used for term recognition. However, the main problem is that there are not many reliable and widely used training resources. An exception is the GENIA corpus that is one of the few terminologically-tagged corpora and is now widely used by the bio-text mining community.

Hatzivassiloglou et al. [2001] present a statistically-based unsupervised technique to acquire and disambiguate the names of proteins, genes, and RNSs. Collier et al. [2000] used HMMs and specific orthographic features (e.g., "consisting of letters and digits," "having the initial letter in upper case," etc.) to discover terms (belonging to a set of ten classes).

The use of hybrid approaches that combine linguistic and statistical knowledge is also increasing [Mima et al. 2001a; Mima and Ananiadou 2001b]. In order to assess the relevance of extracted candidate terms, these methods calculate the weights (i.e., term-hoods) using specific statistical measures. For an extensive overview of approaches to term-extraction in biomedicine, we refer the reader to Ananiadou and Nenadic [2006], Ananiadou et al. [2004], and Krauthammer and Nenadic [2004].

However, ATR is not the ultimate goal. The large number of new terms necessitates a systematic method for accessing and retrieving the knowledge that they represent. Accordingly, the extracted terms must be placed in an appropriate knowledge framework by identifying relationships between the terms and by establishing links between them and different factual databases.

Several ontologies (e.g., MeSH terms, gene ontology, GENIA ontology) have been developed to support knowledge structuring. An ontology is not concerned with lexical

[6] Available at: http://www.hgc.ims.u-tokyo.ac.jp/service/tooldoc/KeX/intro.html

realizations; so it cannot be termed a terminology. On the other hand, it offers a shared and structured view over a concept space. However, a terminology necessarily incorporates an ontology.

Ontologies implement a predefined classification system for concepts and their inter-relationships, as well as inference rules that are used to derive knowledge represented by the concepts. UMLS (unified medical language system) [UMLS 2004], GO (gene ontology), and GENIA are some of the existing biomedical ontologies. However, ontology construction and maintenance are time-consuming activities, since concepts are usually manually integrated into an ontology. This is one reason why ontologies typically contain only a subset of the existing terminology occurring in texts. In addition, solutions for the well-known difficulties of ontology development, ontology conflicts, mismatches [Visser et al. 1997], and a method to map ontological concepts to term-forms in text remain to be found. So techniques for (semi)-automated ontology management [Gamper et al. 1999; Spasic et al. 2005] are urgently required for efficient and consistent KA.

2.2 Integration of Knowledge Sources

Many different approaches to link, integrate, and interpret relevant resources have been suggested. For example, Semantic Web [Berners-Lee 1998] strives to link relevant XML-based resources in a bottom-up manner using the Resource Description Framework (RDF) and ontological information. XML facilitates the introduction of new domain and/or application-specific tags, while RDF [Brickle and Guha 2000] is employed to define their "meanings" and inter-relationships. Corresponding ontologies are used to combine and derive additional information (e.g., synonyms, hyponyms, etc.). In this sense, ontologies are used as key domain knowledge repositories. The Semantic Web is efficient in semantically retrieving the content of resources; however, manual description is still required when defining RDF descriptions and ontologies. However, if we endeavor to process large collections of new documents (including new knowledge), we require systems that do not rely solely on manual descriptions.

In this article, we present our approach toward terminology management and the mining of knowledge sources in a knowledge structuring system [KSS].

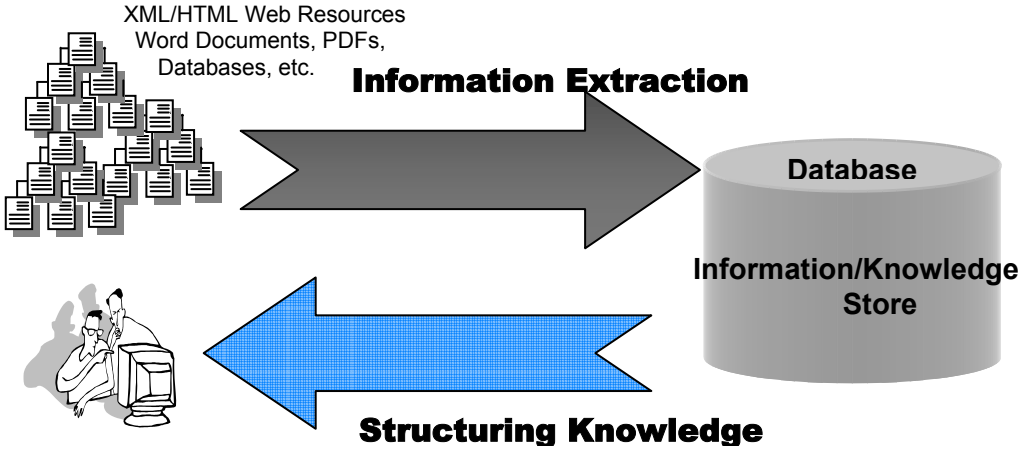

Fig. 1. Information extraction and structuring knowledge.

3. KNOWLEDGE STRUCTURING SYSTEM ARCHITECTURE

The KSS was developed to address the problems of ontology-driven literature mining and KA. This system, similar to the Semantic Web, deals with documents and ontology-based inference. However, it also facilitates KA tasks by using manually-defined resource descriptions and by exploiting natural-language processing techniques like automatic term recognition (ATR) and automatic term clustering (ATC). Both techniques are used for the (semi)-automatic management of the underlying ontology. The KSS system also integrates an information-retrieval engine and a similarity-calculation engine; these tools allow users to locate relevant knowledge sources based on keywords and the relationship between them.

As shown in Figure 1, the system acts as an information-extraction (IE) engine that is based on textual data obtained from its various components, which allow it to deal with documents that are generally available in different formats (e.g., pdf, Word, HTML, and XML) and databases (e.g., SQL server, Oracle, and POSTGRES). Typically, as shown in Figure 2, the KA process takes place via the following steps: first, a collection of documents is linguistically processed (segmentation, part-of-speech (POS) tagging, shallow parsing.) and the resulting texts are indexed for subsequent retrieval; second, the collection is terminologically analyzed, i.e., relevant domain-specific terms are automatically recognized and structured (classified or incorporated into an ontology). The indexing and the ontology development processes described above are performed offline; the index data and ontological information are stored in the corresponding database; third, and last, relevant information is retrieved and structured by using the ontological information. The structured information is then visualized to reveal the result of the knowledge structuring to the user. Details regarding the structuring of knowledge and its visualization are explained in Section 5.

The system architecture is modular, and integrates the following components (Figure 3):

Data Reader (DR) – It extracts textual data from target KSs.

Ontology Development Engine(s) (ODE) *and* ***Ontology Data Manager (ODM)*** – They carry out (semi)-automatic ontology development that includes automatic recognition and clustering of domain terminology and provides the corresponding interface to the other components.

Text Data Manager (TDM) – It stores the index of KSs and ontological information obtained by the ODE in the database.

Information Retriever (IR) – It retrieves the KSs from the TDM and calculates the similarities between keywords and KSs. We adopted tf*idf for similarity calculation.

Similarity Calculation Engine(s) (SCE) *and* ***Similarity Manager (SM)*** – They calculate similarities between the KSs by using their ontology information provided by the IR component in order to calculate semantic similarities among KSs.

Graph Visualizer – visualizes knowledge structures that are based on simple undirected graph expressions in which relevant links between the keywords and KSs; the relevant links among the KSs are also shown.

Linguistic preprocessing within the system is performed in two steps. In the first step, POS tagging[7], i.e., the assignment of basic parts of speech (e.g., nouns, verbs, etc.) to

[7] We use EngCG tagger [Voutilainen et al. 1993] for English and JUMAN/Chasen morphological analyzers for Japanese.

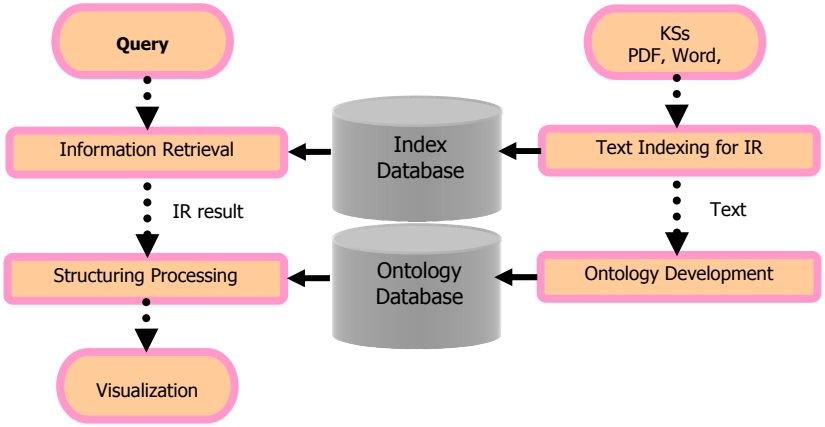

Fig. 2. Data flow diagram.

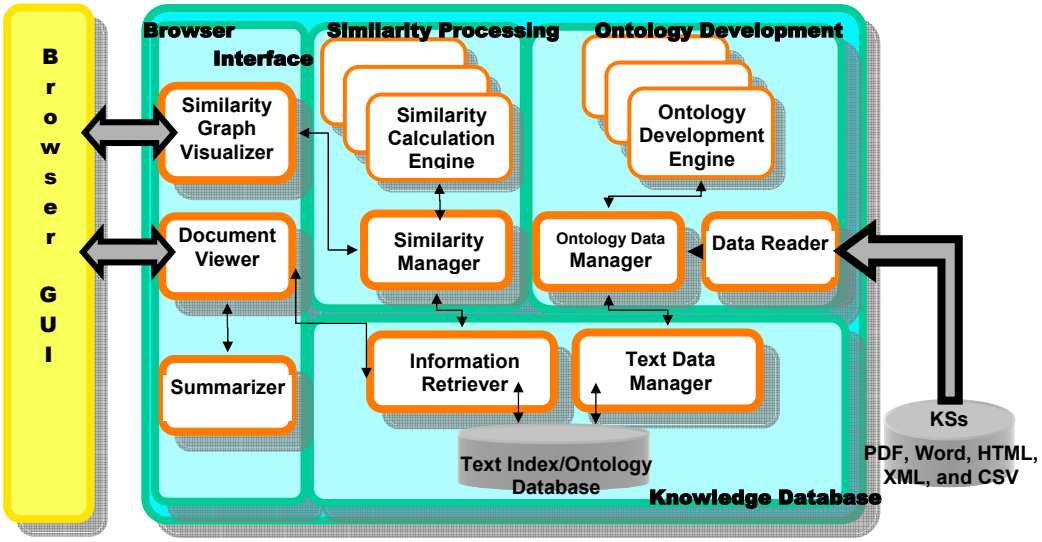

Fig. 3. System architecture.

words, is carried out. In the second step, an ontology development engine is used to perform parsing, i.e., the recognition of basic syntactic structures (e.g., noun phrases). An LFG-based parser that is implemented as a unification-based Generalized-LR (GLR) parser with feature structures is employed for English and Japanese.

4. PROCESSING TERMINOLOGY AS ONTOLOGY DEVELOPMENT

Automatic term recognition (ATR) denotes processes to systematically extract pertinent terms and their variants from a collection of documents. The primary aim of an ATR system is to highlight and extract lexical units that may be related to relevant domain

concepts, i.e., to extract sequences that may be of potential terminological relevance. In other words, ATR is the process of distinguishing terms that belong to a particular subject field from those that are not. ATR systems are typically concerned only with spotting term occurrences in texts and not with their "identification," i.e., mapping the extracted terms into the corresponding concepts. However, the lack of clear naming standards in some domains such as biomedicine often makes ATR a non-trivial problem [Fukuda et al. 1998]. Additionally, it typically gives rise to many-to-many relationships between terms and concepts. In practice, two problems arise due to this—a particular term may represent a number of concepts, while a particular concept may be denoted by more than one term. In other words, some terms may have multiple meanings (*term ambiguity*), while a group of terms may refer to the same concept (*term variation*). Generally, term ambiguity has negative effects on IE precision while term variation decreases the IE recall. These problems point out the disadvantages of using simple keyword-based IE techniques. Evidently, more sophisticated techniques are required. Such techniques should be able to identify groups of different terms that refer to the same (or similar) concept(s). Thus, users could benefit from techniques employing ATR/ATC and term variation management methods that perform efficiently and consistently. These methods are also important for organizing domain-specific knowledge as terms should not be treated in isolation from other terms. Rather, relevant relationships between the terms existing among the corresponding concepts should be formed in order to be at least partly reflected in a terminology.

In our system, term processing is based on the C/NC-value automatic term-recognition method [Mima and Ananiadou 2001b], while ATC is carried out using average mutual information (Figure 4). Its primary purpose is to help domain experts to gather and manage domain-specific terminology. ATC also automatically recognizes and clusters terms offline and transfers the results to the database.

4.1 Recognizing Biomedical Terms in Text

ATR is faced with many challenges, particularly in biomedicine. One of the main challenges is to recognize ad-hoc names (e.g., names of genes, names like *bride of sevenless*, *boss*, *yatio*, *for*) as domain-specific terms. Biomedical terms are generally multiword units (85% – 90%). Thus, term boundaries (e.g., whether the word *possible* is part of the term *possible T and natural killer cells*) and nested terms (e.g., terms *cell line* and *T cell line* are embedded, among others, in *leukemic T cell line Kit225*) must be recognized – but these are typically non-trivial tasks. And, due to their complexity, variations in terms pose another challenge. In addition, biological names are very complex; the literature contains huge numbers of synonyms and variant term-forms [BioCreAtIvE 2004]. Most terms are used along with synonyms and other variants such as acronyms, morphological and derivational variations, and so on (e.g., *TIF2, TIF-2, transcription intermediary factor-2, transcriptional intermediate factor 2*). Thus, term variations are an integral part of automatic term recognition. Further, many biological terms and their variants are ambiguous, as they share lexical representation with either common English words (gene names/abbreviations like *an*, *by*, *can*, and *for*) or other terms (systematic ambiguities that have to be resolved with ontological considerations).

To extract biomedical terms we developed and tuned the *C/NC-value* method [Mima et al. 2001a; Ananiadou et al. 2004] that recognizes primarily multiword terms by combining linguistic knowledge and statistical analysis. The C/NC-value method is an ATR approach that is independent of domain and language. This method enhances the commonly used baseline approach (frequency of occurrence) by making term extraction

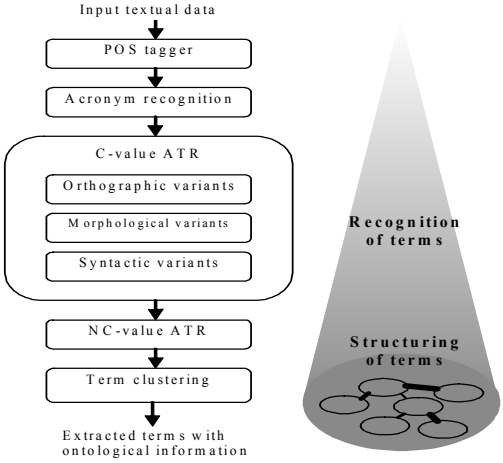

Fig. 4. Processing terminology.

sensitive to particular types of terms, namely, nested terms and multiword terms, which cause most of the problems. In addition, we incorporated term variations in order to enhance the performance of the C/NC-value method [Nenadic et al. 2004].

The C/NC-value method is implemented as a two-step procedure. In the first step, candidate terms are extracted by using a set of linguistic filters and implemented using an LFG-based GLR parser that describes general patterns of term formation. In the second step, the candidate terms are assigned term-hoods (C-values) according to a statistical measure. The measure combines four numerical corpus-based characteristics of a candidate term, namely, frequency of occurrence, frequency of occurrence as a substring of other candidate terms, 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* improves the C-value results further by taking the context of the candidate terms into account. The relevant context words are extracted and assigned weights based on how frequently they appear with top-ranked candidate terms that are 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 term-hood estimates, referred to as NC-values, are, for the respective terms, calculated as a linear combination of the C-values and the context factors. Evaluation of the C/NC-method (see Section 6) shows that contextual information improves term distribution in the extracted list by placing real terms closer to the top of the list; term variation further enhances the performance of the system.

4.2 Managing Term Variation

Term variation and ambiguity cause problems not only for ATR but also for human experts. Several methods for managing term variation have been developed, e.g., the BLAST system [Krauthammer et al. 2000] employs approximate text string-matching techniques and dictionaries to recognize spelling variations in the names of genes and proteins. FASTR [Jacquemin 2001] handles morphological and syntactic variations by employing meta-rules to describe term normalization, while semantic variants are dealt with by WordNet.

Table I. Term Normalization Examples

Synterms	Canonical representative
human cancers *cancer in humans* *human's cancer* *human carcinoma*	*human cancer*

Table II. Examples of Recognized Terms and Their Variants

Synterms	Canonical representative
All trans retionic acid, all-trans-retinoic acids, ATRA, at-RA	*All trans retionic acid*
Nuclear receptor, nuclear receptors, NR, NRs	*Nuclear receptor*
9-c-RA, 9cRA, 9-cis-retinoic acid, 9-cis retinoic acid	*9-cis-retinoic acid*
RAR alpha, RAR-alpha, RA receptor alph, retinoic acid receptor alpha	*Retinoic acid receptor a*
DNA, DNAs, deoxyribonucleic acid	*Deoxyribonucleic acid*
NF-KB, NF-kb, nuclear factor kappa B, NF-kappaB	*Nuclear factor kappa B*

The basic C-value method is enhanced by term-variation management [Mima et al. 2001a; Nenadic et al. 2004]. We consider various sources from which problems regarding term variation originate. In particular, we deal with orthographical, morphological, syntactic, lexico-semantic, and pragmatic phenomena. Our approach to managing term variations considers 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 candidate terms are separated, as opposed to other approaches (e.g., FASTR handles variants by subsequently applying transformation rules to the extracted terms). In order to conflate equivalent surface expressions, linguistic *normalization* of individual candidate terms (examples in Table I) is carried out. Firstly, each candidate term is mapped to a specific *canonical representative* (CR) by semantically isomorphic transformations. Thereafter, we establish an equivalence relationship wherein two candidate terms are related *iff* they share the same CR. The partitions of this relationship are denoted as *synterms*. A synterm comprises surface term representations sharing the same CR. Our aim is to form synterms before the syntactic estimation of term-hoods for candidate terms [Nenadic et al. 2004].

Examples of extracted terms are presented in Table II.

4.3 Term Clustering

In the literature, term clustering is an indispensable component of the mining process, in addition to term recognition. Since term opacity and polysemy are extremely common in molecular biology and biomedicine, term clustering is essential in order to integrate semantic terms and to construct domain ontologies and semantic tagging.

In our system the ATC is performed using a hierarchical clustering method that merges clusters based on average mutual information that measures how strongly terms are related to each other [Ushioda 1996]. The input consists of terms and their co-occurrences, recognized automatically by the NC-value method; a dendrogram of terms is produced as the output. Parallel symmetric processing is used for high-speed clustering.

The calculated term cluster information is encoded and used for calculating semantic similarities in the similarity calculation engine (SCE).

5. VISUALIZATION TO GENERATE KNOWLEDGE MAPS

As compared to IE/KA, knowledge mining can be regarded as the broader approach The IE and KA in our system are implemented through the integration of terminology-based ontology development and calculation of semantic similarities. Graph-based visualization for the automatic generation of knowledge maps is also provided to help in retrieving knowledge and KA from documents. The system also supports combining the different types of databases (papers, patents, technologies, and innovations) and retrieves different types of knowledge simultaneously across documents. This feature can accelerate the discovery of knowledge by combining existing types of knowledge. The basic idea behind the discovery of new knowledge by using ontological information follows: If we find knowledge based on the condition "if *A* then *B*" and "if *C* then *D*" and *iff the* ontological relationship "*B* = *C*" is provided, then we can obtain new knowledge "if *A* then *C*" by syllogism; whereas new knowledge cannot be discovered if the relationship between *B* and *C* (Figure 5) is not known. For example, we can expect to discover new knowledge about industrial innovation by structuring the knowledge of up-to-date collections of scientific papers and reports on past industrial innovation. Figure 6 shows an example of the visualization of knowledge structures from paper abstracts relevant to the term "receptor" in the GENIA corpus [GENIA 2002]. In order to structure knowledge, the system constructs a graph in which the nodes indicate relevant KSs for the keywords specified by the user. Links among the KSs indicate semantic similarities that are calculated using ontology information developed by our ATR/ATC components. Semantic similarity is based on comparing ontological information extracted from each KS, whereas conventional similarity calculation is generally based on nouns extracted from each KS. Additionally, the locations of each node are calculated and optimized when drawing the graph. The distance between nodes depends on how close they are in meaning. The complete algorithm of this knowledge-structuring method follows:

```
begin
  Q ← query specified to IR
  R ← IR(Q)  // retrieving relevant KSs to Q and putting them into R
  for every x in R do
    w(Q, x) ← IRscore(Q, x)  // calculate IR score between Q and x
     for every y in R do
       if x ≠ y then
             p ← Ont(x)  // retrieving ontology information of x
             q ← Ont(y)  //          ″                  y
             w(x,y) ← Sim(p,q)  // calculate similarity using p and q
      fi
       end
 end
  Visualize graph based on every {w(i,j)|i=Q or i∈R, j∈R, i≠ j}
end.
```

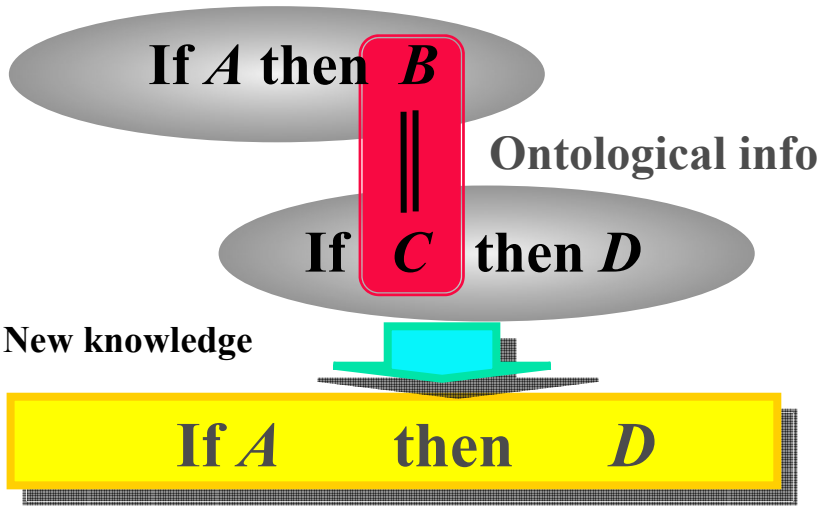

Fig. 5. New knowledge discovery.

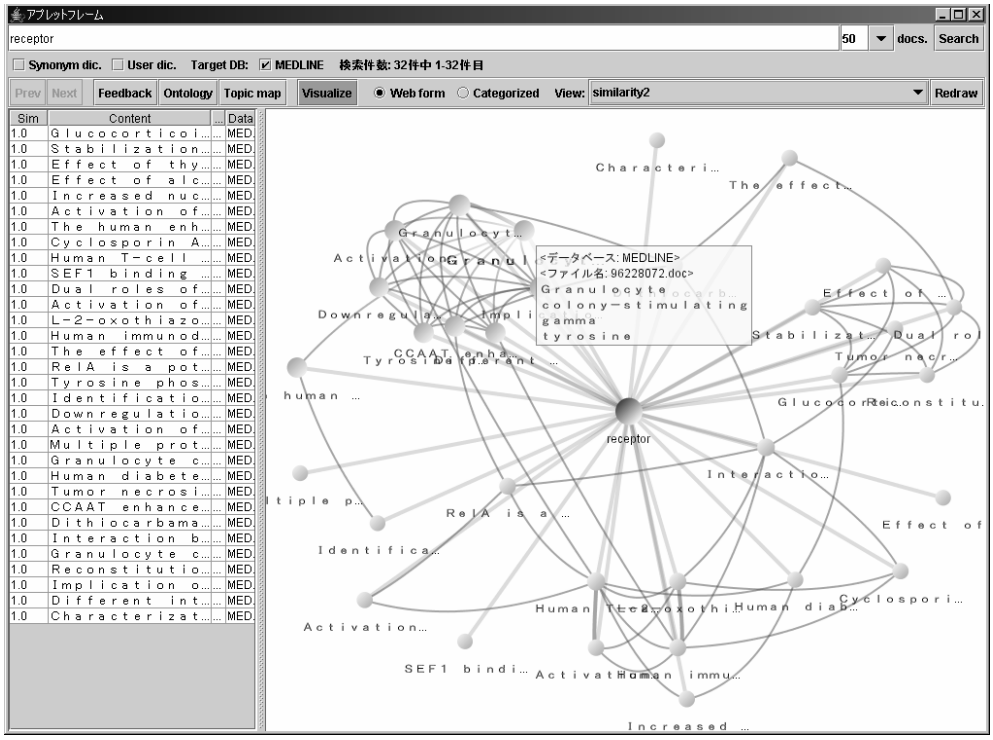

Fig. 6. Visualization sample.

We generate a knowledge map by means of (i) cluster recognition and (ii) terminology-based categorization. Cluster recognition is carried out by detecting groups of nodes in which every combination of included nodes is strongly linked (i.e., their similarity exceeds a threshold). Automatic categorization is done by using a thesaurus and an SVM-based categorizer. Figure 7 shows a knowledge map that was generated from news articles. The target information was extracted from online articles in *Yomiuri* and *Mainichi* (newspapers in both English and Japanese), where the keywords specified for IR are "Iraq" and "Fallujah."

As shown in Figure 7, seven clusters are recognized and category names assigned: (1) Bin Laden, (2) Secretary of State Powell, (3) Dispatch of the Japanese self-defense forces, (4) Presidential election, (5) Samawah, (6) Prime Minister Koizumi, and (7) Prime Minister Allawi. The basic method includes categorizing and mapping concepts in order to help to understand the information. Furthermore, this method can also be used to disambiguate semantically specified keywords. For example, the keyword "apple" includes at least two meanings, namely, "fruit" and "computer company." However, by using clustered and categorized IR results, we can find the information we want more easily.

6. EXPERIMENTS AND EVALUATION

In this section we report on the experiments conducted using an AI domain corpus for ATR, and the GENIA corpus for terminology-based categorization, to demonstrate how ontology development and knowledge map generation perform in practice.

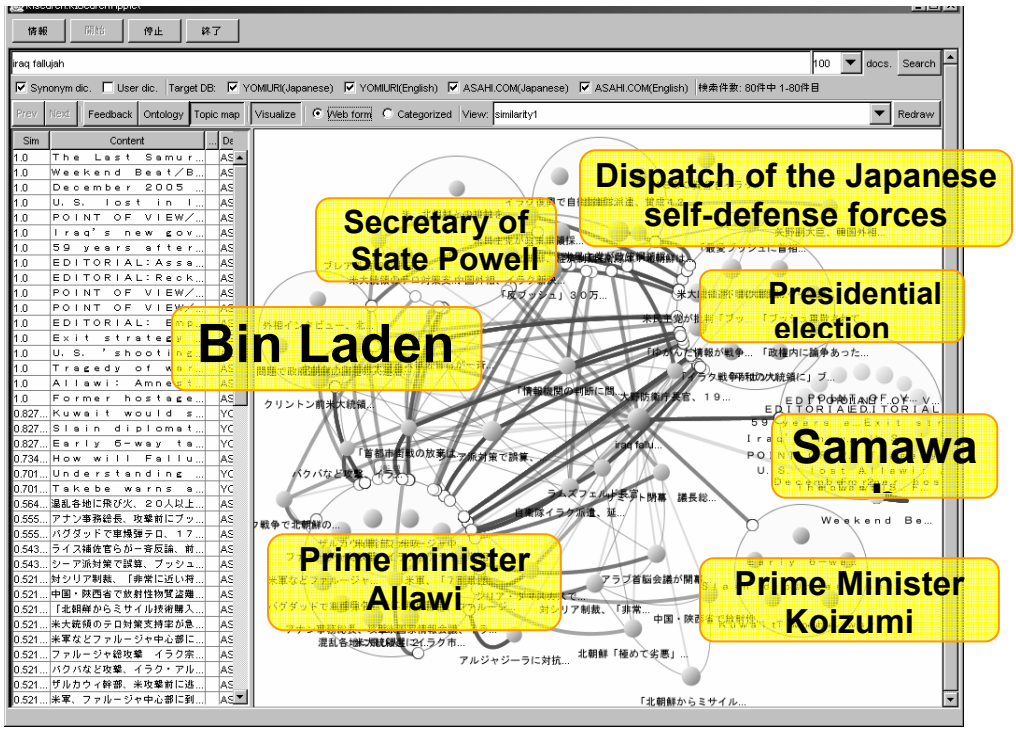

Fig. 7. Knowledge map generation sample.

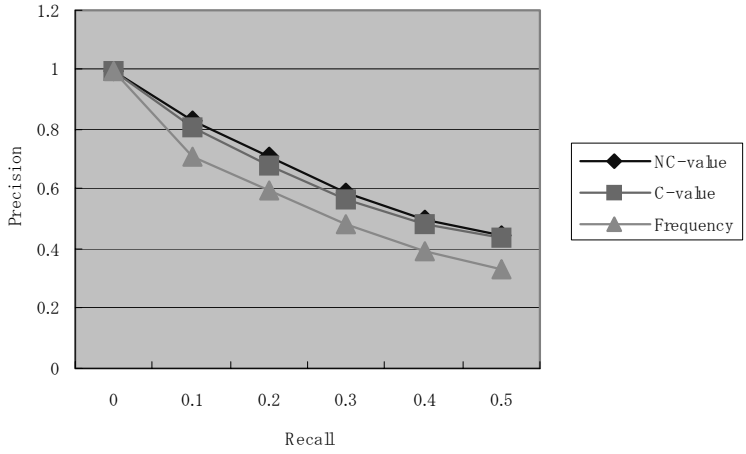

Fig. 8. Precision and recall of C/NC-value methods.

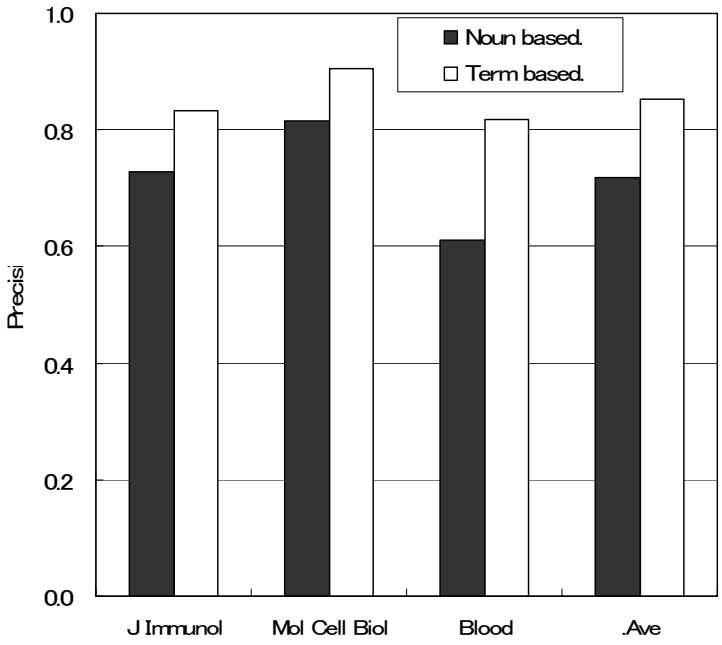

Fig. 9. Term-based vs. noun-based.

The experiments with term-variation management in ATR were conducted on the GENIA corpus containing 100,000 nouns (2082 abstracts) from the MEDLINE database [MEDLINE 2002]. Figure 8 shows the recall against precision graph for the C/NC-value method compared to the frequency of occurrence. It can also be seen that the NC-value

method is slightly more precise than the C-value method, and substantially more precise than conventional pure frequency-based methods.

The GENIA corpus [GENIA Project 2002] was also used for the categorization experiment. The test set contained approximately 10,000 terms across the three major GENIA classes (*nucleic acid*, *amino acid, and source*) and approximately 100,000 nouns. The terms and nouns were used for term-based and noun-based categorizations, respectively. It was found that low-frequency terms (nouns) play an important role in the categorization of texts. However, learning with low-frequency terms produces an increasing number of features. Hence, the calculation costs also increase. On the other hand, our method compresses features that reduce calculation costs by using terminology information -- resulting in an efficient text-categorization technique. This experiment allowed us to categorize GENIA abstracts into three major categories, namely (1) Immunol, (2) Mol Cell, and (3) Blood. We used TinySVM [2004] to learn the categorization model, and 50 abstracts to learn and create the three categories.

Figure 9 shows the precision for both term-based and noun-based categorizations. As shown in the figure, although the sample size was not large enough, term-based categorization precision was better than noun-based categorization. So we expect the method to be practical and efficient enough to generate knowledge maps to make new knowledge discoveries.

7. CONCLUSION

This article presents an integrated knowledge-mining system for the biomedicine domain, which integrates automatic term recognition, term clustering, information retrieval, and visualization. Its main objective is to facilitate knowledge acquisition from documents and the discovery of new knowledge by calculating terminology-based similarities and visualizing automatically-structured knowledge. Additionally, to accelerate knowledge discovery, we proposed a visualization method for generating similarity-based knowledge maps. This method is based on real-time terminology-based knowledge clustering and categorization, and allows users to observe knowledge maps being generated graphically in real time. Experiments on the GENIA corpus shows that this method is practical enough to use for enhancing new knowledge discovery from existing knowledge sources.

An area for future research includes evaluating the usability of the system. We also intend to investigate the possibility of using a system for classifying terms as an alternative structuring model for knowledge deduction and inference, instead of an ontology.

ACKNOWLEDGMENTS

Hideki Mima is grateful to the Artificial Intelligence Research Promotion Foundation for promoting this study, in part under the AI research grant scheme. Hideki Mima and Sophia Ananiadou thank the Daiwa Foundation for enabling their collaboration with the support of the Daiwa Adrian Prize scheme.

REFERENCES

ANANIADOU, S. AND NENADIC, G. 2006. Automatic terminology management in biomedicine. In *Text Mining for Biology and Biomedicine*, S. Ananiadou and J. McNaught (eds), Artech House, Norwood, MA, Ch.4, 67-98.

ANANIADOU, S., FRIEDMAN, C., AND TSUJII, J. (EDS). 2004 Named entity recognition in biomedicine. *J. Biomedical Informatics 37,* 6. Special issue.

BERNERS-LEE, T. 1998 The Semantic Web as a language of logic. Available at: www.w3.org/DesignIssues/Logic.html.

BRICKLE, D. AND GUHA, R. 2000. Resource description framework (RDF) schema specification 1.0, W3C Candidate Recommendation. Available at: http://www.w3.org/TR/rdf-schema.

COLLIER, N., NOBATA, C., AND TSUJII, J. 2000. Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the International Conference on Computational Linguistics* (COLING 2000, Saarbrücken, Germany), 201-207.

FRANTZI, K., ANANIADOU, S., AND MIMA, H. 2000. Automatic recognition of multi-word terms. *Int. J. Digital Libraries 3,* 2, 117–132. Special issue.

FUKUDA, K., TSUNODA, T., TAMURA, A., AND TAKAGI, T. 1998. Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the PSB-98* (Hawaii), 705-716.

GAIZAUSKAS, R., DEMETRIOU, G., AND HUMPHREYS, K. 2000. Term recognition and classification in biological science journal articles. In *Proceedings of the Workshop on Computational Terminology for Medical and Biological Applications* (NLP-2000, Patras, Greece), 37-44.

GAMPER, J., NEJDL, W., AND WOLPERS, M. 1999. Combining ontologies and terminologies in information systems. In *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering*, (Innsbruck, Austria), 152-168.

GENIA PROJECT. 2002. Genia project home page. www-tsujii.is.s.u-tokyo.ac.jp/GENIA/.

HATZIVASSILOGLOU, V., DUBOUE, P.,AND RZHETSKY, A. 2001. Disambiguating proteins, genes, and RNA in text: A machine learning approach. *Bioinformatics 17*,1, S97-S106.

JACQUEMIN, C. 2001. *Spotting and Discovering Terms through NLP*. MIT Press, Cambridge, MA, 378.

KRAUTHAMMER, M., RZHETSKY, A., MOROZOV, P., AND FRIEDMAN, C. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene 259*, 245-252.

KRAUTHAMMER, M. AND NENADIC, G. 2004. Term identification in the biomedical literature. *J. Biomedical Informatics*. Special issue on named entity recognition in biomedicine.

MEDLINE (NATIONAL LIBRARY OF MEDICINE). 2002. http://www.ncbi.nlm.nih.gov/PubMed/.

MIMA, H., ANANIADOU, S., AND NENADIC, G. 2001a. ATRACT workbench: An automatic term recognition and clustering of terms. In *Text, Speech and Dialogue*, V. Matoušek et al. (eds.), LNAI 2166, Springer Verlag, 126-133.

MIMA, H. AND ANANIADOU, S. 2001b. An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *Int. J. Terminology 6*/2, 175-194.

NENADIC, G., ANANIADOU, S., AND MCNAUGHT, J. 2004. Enhancing automatic term recognition through term variation, In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004, Geneva, Switzerland).

SPASIC, I., ANANIADOU, S., AND TSUJII, J. 2005a. Masterclass: A case-based reasoning system for the classification of biomedical terms. *Bioinformatics 21*, 11, 2748-2758.

SPASIC, I., ANANIADOU, S., MCNAUGHT, J., AND KUMAR, A. 2005b. Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics 6*, 3, 239-251.

TinySVM. 2004. http://chasen.org/~taku/software/TinySVM/.

UMLS. 2004. http://www.nlm.nih.gov/research/umls/.

USHIODA, A. 1996. Hierarchical clustering of words. In *Proceedings of the International Conference on Computational Linguistics* (COLING 1996, Copenhagen, Denmark), 1159-1162.

VISSER, P.R.S., JONES, D.M., BENCH-CAPON, T.J.M., AND SHAVE, M.J.R. 1997. An analysis of ontology mismatches -- Heterogeneity versus interoperability. In *Proceedings of the AAAI 1997 Spring Symposium on Ontological Engineering* (Stanford University, Stanford, CA), 164-172.

VOUTILAINEN, A. AND HEIKKILA, J. 1993. An English constraint grammar (ENGCG), a surface-syntactic parser of English. In *Creating and Using English Language Corpora*, U. Fries et al. (eds.), Rodopi, Amsterdam, 189-199.

Received December 2004; revised July 2005; accepted July 2005