Available online at www.sciencedirect.com

SciVerse ScienceDirect

Energy Procedia 13 (2011) 790 – 800

Energy Procedia

ESEP 2011: 9-10 December 2011, Singapore

An Ontology-based Automatic Semantic Annotation Approach for Patent Document Retrieval in Product Innovation Design

Feng Wang[a], Lanfen Lin[a,]*, Zhou Yang[a]

[a] *College of Computer Science, Zhejiang University, Hangzhou 310027, China*

Abstract

Patent retrieval plays a very important role in product innovation design. However, current patent retrieval approaches lack semantic comprehension and association, and usually can't capture the implicit useful knowledge at a semantic level. In order to improve the traditional patent search, this paper proposes a novel ontology-based automatic semantic annotation approach based on the thorough analysis of patent documents, which combines both structure and content characteristics, and integrates multiple techniques from various aspects. Multilayer semantic model is established to realize unified semantic representation. The approach first utilizes template schemes to extract the structure information from patent documents, and then identifies semantics of entities and relations between entities from the content based on natural language processing techniques and domain knowledge, and at last employs a heuristic pattern learning method to abstract patent technical features. Case study is provided to show that our approach can acquire multi-level patent semantic knowledge from multiple perspectives, and discover semantic correlations between patent documents, which can further promote the accurate patent semantic retrieval effectively.

Keywords: patent document, semantic annotation, ontology, semantic retrieval

1. Introduction

Product innovation design is the lifeblood for today's enterprises to survive. Enterprises maintain competitiveness by introducing innovative products [1]. When engineers develop new products or

* Corresponding author. Tel.: 86-571-87952699; fax: 86-571-87951247.
E-mail address: llf@zju.edu.cn.

1876-6102 © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Singapore Institute of Electronics
doi:10.1016/j.egypro.2011.11.110

technologies, they usually have to refer to existing inventions described in patent documents. Patent documents contain a lot of important valuable knowledge, which can save time for new product development, increase success chance for market, and reduce potential patent infringement. But they are large in quantity, lengthy in space and rich in technique terminology such that they are difficult for human to sift through.

To solve the problem above, popular search engines like Google and Baidu, and professional patent retrieval tools like PatentCafe [2] and SooPAT [3], both provide patent search service. However, most of them rely on keywords-based methods, which take no account of some important patent semantic knowledge such as design aim, design principle and application effect, let alone discover semantic association between patents or capture implicit useful knowledge at a semantic level. Effective patent semantic retrieval need to be supported. For example, users may want to fast identify the relations between entities existing in one single patent, the concise technical features of patents and further the implicit association across patents.

The purpose of this research is to automate patent document semantic annotation as a key step to effective patent retrieval. Semantic annotation is intended to extract and annotate semantic information from patent documents, which can make patent documents machine-understandable and produce rich semantic knowledge for patent semantic retrieval.

In this paper, we propose a novel automatic semantic annotation approach that integrates ontology-based technique, structural template schemes, natural language processing, and pattern learning to annotate patent documents from various aspects according to the structure and content characteristics of patent documents. Multilayer semantic representation model is presented to represent patent document semantics with different granularities.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 describes semantic representation for patent documents. The detail semantic annotation approach is presented in sections 4. Case study is discussed in section 5. Finally, section 6 concludes the paper and discusses some future work.

2. Related work

Many projects that utilize different techniques for patent retrieval have been developed. PatentCafe [2] offers a latent semantic indexing method to patent information search in combination with keywords and Boolean operators, considered to be shallow semantic. BioPatentMiner [4] facilitates keyword search and queries linking the properties specified by RDF triples for biomedical patents. PATExpert [5] focuses on patent document processing based on several semantic web technologies. SooPAT [3] is a keywords-based and IPC-based search engine towards Chinese patent documents. However, available patent retrieval tools mostly depend on keywords-based statistics methods which have limited capability in semantic comprehension and association for patent documents. Although some studies like PATExpert have allowed for patent semantic retrieval to some degree, they almost do not take into account technical feature semantics of patent documents.

Technical features are the core of patent documents. The lack of design intention capture may lead to inaccurate understanding of patent documents. There have been several researches to obtain technical features. Cascini et al. [6] presented a TF-IDF ranking algorithm to identify peculiarities through patent functional analysis. Trappey et al. [1] proposed an ontology based and concept clustering approach for patent document summarization. Tseng et al. [7] gave a general methodology using a series of text mining techniques. But they focus on component or key phrases extraction, and can’t reflect technical feature semantics well.

In general, patent document retrieval in product innovation design still faces some challenges.

Semantic representation should be provided, and technical feature semantics should be included. Further, semantic association between concepts and patents or among patents should be considered. Realization of the above points depends on the support of rich semantic knowledge produced by desirable semantic annotation.

Semantic annotation aims to make documents machine-understandable [8]. It comes with ontology that employs rich modeling languages like OWL and provides an effective means for making implicit information explicit. The process of semantic annotation is to extract, recognize and annotate concepts and relations between concepts in documents, thereby linking formal semantic descriptions to documents.

Existing semantic annotation approaches differ with respect to automation degree, document format, processing techniques, semantic model, and application domain [8]. KIM [9] is a platform for automatic ontology-based annotation of named entities using NLP and pattern matching. Cerno [10] is a tool for semi-automatic semantic annotation of textual documents according to domain-specific semantic model founded on NLP and software code analysis. Tao and Embley [11] provided a structure pattern method for automatic hidden-web table semantic annotation. Li et al. [12] described an annotation approach using ontology engineering and NLP for unstructured engineering documents. Cui et al. [13] proposed an unsupervised algorithm employing bootstrapping procedures to semantic annotation for biosystematics literature. Ghoula et al. [14] investigated an initial ontology-based semantic annotation approach in biomedical patent domain. Most studies target at web pages, but little on patent documents. Hybrid method, combining with multiple strategies, is more likely to meet better semantic annotation needs.

3. Semantic representation

3.1. Characteristics of patent documents

Characteristics of patent documents are of great importance to semantic annotation. If exploited effectively, they can facilitate semantic annotation. Thus some key characteristics of patent documents are concluded as follows:

1) Unified document structure formats. Patent document structure refers to the standardized front page, the structuring of claims, the layout of description, the references to figures, etc. Patent standards determine document structure hierarchy as well as the layout of each document type. Furthermore, each unit has its own composition such as headings, paragraphs, subsections that should be taken into consideration, especially for patent document structure analysis. Taking advantages of the structure characteristics, it's easy to identify various chunks of structure information in patent documents.

2) Regular but domain-related content expression styles. In patent documents, writing styles or expression styles are mainly similar and regular. For example, “is constituted by” is one of the very common expressing ways that represents “part-of” relation between concepts. There are also some common expression terms such as “thing”, “time”, “space”, etc. The content expression characteristics are helpful for semantic recognition. Of course, patents belonging to different technical areas contain much exclusive technical terminology that requires the support of domain-specific knowledge. So, regular and domain-dependent content expression styles can be made use of to facilitate semantic knowledge acquisition.

3) Multi-level underlying semantic knowledge. Patent documents contain not only explicit but also implicit fine-grained information, apart from coarse-grained structure information. Explicit information is given in the first page of patent documents, including application no, title, inventor, etc. Implicit information has to be extracted by further content processing, including concepts and relations between concepts, and semantic association across patents. But the multi-level information has no semantics itself and can only be understood by humans not by machine. Thus it’s necessary to link the information to the corresponding semantic descriptions by semantic annotation, enabling information to become the accessible semantic knowledge. Above all, each patent has its own technical features that are distinguishable from others.

3.2. Semantic representation

As analyzed in the previous section, there are both shared and domain-specific knowledge, macro and minor knowledge in patent documents. These resources are often located implicitly in documents, making effective capture and retrieval a critical issue. In order to manage semantic knowledge in a manner that is explicit, flexible and extensible, multilayer semantic representation model is presented as shown in Fig.1. Ontology is employed, providing rich semantic modeling and inference support.

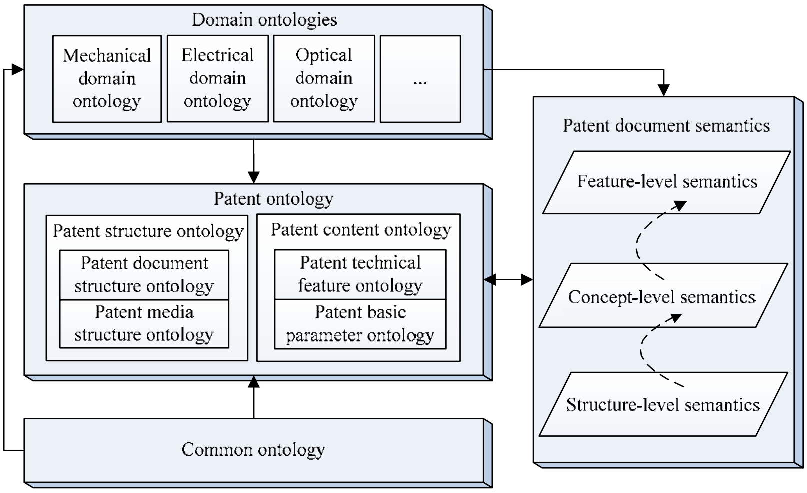

Domain ontologies
Mechanical domain ontology
Electrical domain ontology
Optical domain ontology
...
Patent ontology
Patent structure ontology
Patent document structure ontology
Patent media structure ontology
Patent content ontology
Patent technical feature ontology
Patent basic parameter ontology
Common ontology
Patent document semantics
Feature-level semantics
Concept-level semantics
Structure-level semantics

Fig. 1. Ontology-based multilayer semantic representation model

1) Common ontology. Common ontology provides common concepts and relations among concepts that are the basis for other ontologies. It typically contains specifications of domain-independent concepts and relations derived from linguistics. Other ontologies build upon this common ontology, which can facilitate share and reuse, and ease integration and alignment.

2) Patent ontology. Patent ontology, extending basic concepts and relations from common ontology, defines patent-specific fundamental concepts and relations. In order to describe patent semantics at different granularities, it is further divided into two ontologies: patent structure ontology and patent content ontology. Because technical terminology stems from different technical areas, patent ontology is closely related to domain ontologies.

Patent structure ontology models patent-specific document structure and media objects that appear in patent documents. Document structure ontology describes document structure layouts, for example, first page, claims, description, and each part has their own subparts as well. Media ontology represents figures such as flowcharts and images whose tokens are corresponding to the concrete content.

Patent content ontology models patent-specific content including basic parameters and technical features. Basic parameter ontology refers to explicit metadata like patent no, title and inventor, and basic relations between metadata. Technical feature ontology describes technical features, such as design aim, design principle, and application effect. Patent content ontology is related to patent structure ontology, for instance, basic parameters of patent content have to be extracted from the front page of patent structure, and technical features of patent content are usually contained in claims and description of patent structure.

3) Domain ontologies. Product innovation design is a very complex process that involves extensive design knowledge from mechanical, electrical, optical, and other domains. Domain ontologies model knowledge for different technical fields. Different domains also contain sub-domains. Domain ontologies are designed to be loosely-coupled with other ontologies according to the actual demands.

4) Patent document semantics. Based on these ontologies, patent document semantics are equal to instantiations of ontologies. In other words, ontologies are the generalization of patent document semantics, and patent document semantics is the specialization of ontologies. Patent document semantics are embodied in three levels: structure-level, concept-level and technical feature-level. Structure-level semantics refer to patent structure semantics and patent basic parameter semantics. Concept-level semantics refer to instances of concepts and relations. Technical feature-level semantics refer to concise semantic descriptions of patent technical features. Patent ontology and domain ontologies are exploited to obtain patent semantics from different angles.

The detail method about ontology construction has been presented in our previous work [15]. In summary, the semantic model systematizes multiple aspects of patent document semantics, and reflects patent documents in a semantic level.

4. Ontology-based automatic semantic annotation approach

According to typical characteristics of patent documents and patent semantic representation model presented above, we propose an automatic semantic annotation approach for patent documents, as shown in Fig.2.

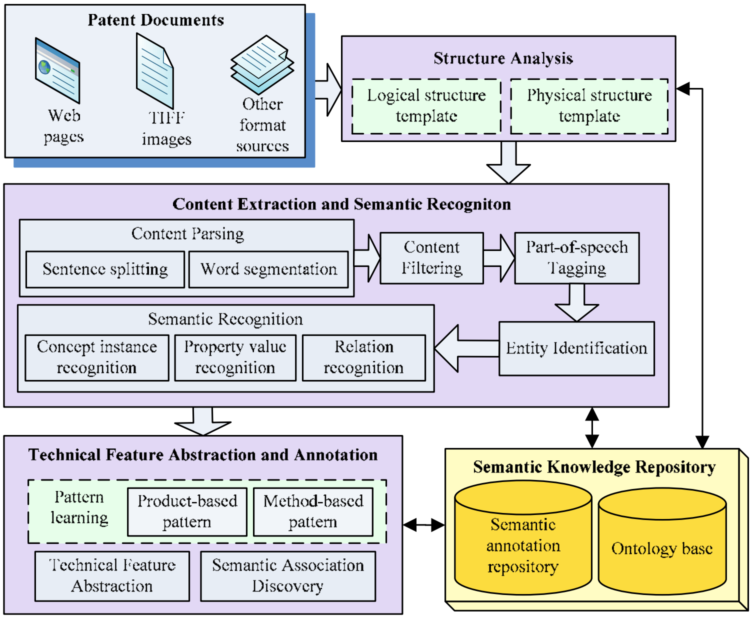

Patent Documents
Web pages
TIFF images
Other format sources
Structure Analysis
Logical structure template
Physical structure template
Content Extraction and Semantic Recogniton
Content Parsing
Sentence splitting
Word segmentation
Content Filtering
Part-of-speech Tagging
Semantic Recognition
Concept instance recognition
Property value recognition
Relation recognition
Entity Identification
Technical Feature Abstraction and Annotation
Pattern learning
Product-based pattern
Method-based pattern
Technical Feature Abstraction
Semantic Association Discovery
Semantic Knowledge Repository
Semantic annotation repository
Ontology base

Fig. 2. Ontology-based automatic semantic annotation framework

Patent documents are usually open in some formats like web pages or TIFF images. Ontology base, consisting of multiple ontologies mentioned in the preceding section, is utilized to obtain semantics. Semantic annotation repository is expected to store extracted semantic descriptions. We call ontology base and semantic annotation repository as semantic knowledge repository, which sets the basis for patent knowledge retrieval. Semantic knowledge repository is not only used to assist the automatic knowledge extraction and semantic recognition, but also provide a dynamic concept space to annotate and index the documents. The framework can be divided into three major modules, and the next three subsections will elaborate on these modules.

4.1. Structure analysis

Document structure defines the elements in a document and how they are organized, and those with well-defined structures can facilitate the extraction of document elements [16]. As for patent documents, their standard structure styles deserve special attention. Because of the fixed structure of patent documents, we can predefine some template schemes from logical structure view and physical structure view. Logical structure templates are used to get logical structure content. Take Chinese patent document as example. Logical structure of Chinese patent documents contains five parts: metadata, including several subparts like application number, application date, title, inventors, etc.; abstract; claims; description, including several subparts, namely, technical field, background art, content of the invention, description of figures, and mode of carrying out the invention; figures. Physical structure templates are used to obtain physical structure content like heading, body, section, subsection, paragraph, and figure. Physical structure templates are used to obtain media structure content, i.e., figures, titles of figures, and number tokens in figures. Each part has its own composition such as subsections and paragraphs that are also taken into consideration. The templates contain tag styles and locations of all parts of structure content. On the basis of the templates, patent document page codes are parsed by template matching, and then document elements are identified. Consequently, basic parameters of patent content are obtained and linked with semantic descriptions corresponding to patent basic parameter ontology, and various parts of patent structure including document structure and media structure are also acquired corresponding to patent structure ontology. These solved results set foundation for the following procedure.

4.2. Content extraction and semantic recognition

Upon the completion of structure analysis, each part of content in patent documents has been acquired, which then turns to the content extraction and semantic recognition module. The purpose of this module is to extract and recognize semantics from each part of patent content.

Linguistic analysis such as content parsing, content filtering and POS tagging are firstly executed to transform structure content to segmented content. Due to the language differences, the processes are associated with linguistics. We employ HowNet [17] as Chinese lexicon and WordNet [18] as English lexicon. Based on them, content parsing separates the content into reasonable words and phrases, transforming the content into a set of terms. Sentences are split in accordance with punctuation marks or blanks in the content, and the forward maximum match algorithm is used for word segmentation according to lexicon. Content filtering then filters out terms those are semantically irrelevant like interjections and auxiliary words. These meaningless words are regarded as noise data in the course of content processing, and are predefined in stop word list. Accordingly, the noise data is eliminated from the content. POS tagging is the identification of words as nouns, verbs, adjectives, etc. It is done in light of the linguistic context, associating separated terms with corresponding parts of speech. Since semantic items extracted are composed of basic elements like nouns and verbs, this phase produces meaningful tagged terms to be used for the subsequent procedure.

After linguistic analysis, important terms of the content have been segmented, and the next procedure is to identify and determine semantics of the segmented terms. First of all, entities are recognized in view of part of speech. Nouns and noun phrases are usually denote entities. Then, semantic recognition is performed to identify entities' semantics and relation semantics between entities in accordance with domain ontologies. Choice of the concrete domain ontology is based on patent class numbers, i.e., international class numbers and category class numbers, which are obtained in previous structure analysis module. Semantic recognition determines whether the entities are thought as the instances of concepts or the property values of instances or new concept, and also recognizes relations between them. Semantic items are basically represented in subjects, predicates, and objects. For example, in a sentence ''Window cover could facilitate electrical equipment case", the subject is window cover, the predicate is facilitate, and the object is electrical equipment case. Subjects and objects are often expressed in nouns, and predicates in verbs. Hence the foremost task in semantic recognition is to analyze nouns and verbs in the segmented content and determine the concepts or relations they belong to.

Nouns are often corresponding to concepts or instances of concepts in ontology when they are subjects or objects. And nouns with numeric values usually refer to the value of the concept's attribute. For example, a sentence "This utility model disclosed one kind of lamp, which is constituted by lamp holder and lamp body…Its rotational angle can reach 180 degrees" where "lamp" and "lamp holder" as well as "lamp body" are domain concepts, "rotational angle" is an attribute of the concept "lamp", "180 degrees" is a value of the attribute "rotational angle".

Verbs usually represent relations when they are located between subjects and objects. For example, in the same sentence used in previous paragraph, "lamp" is constituted by "lamp holder" and "lamp body". "is_constituted_by" is opposite to "is_part_of". In spite of the difference, they both represent composing meaning. That A is constituted by B and C is equal to that B and C are parts of A. Other default relations can be expressed as the corresponding verb attribute relations. Besides, compound phrases such as prepositional phrases can also denote relations, which are taken into account according to their compositions.

Semantic matching is important in semantic recognition. We utilize structure matching and name matching, and they are assigned different weights to ensure accuracy and stability. Simultaneously, last semantic annotation results in semantic annotation repository are also applied to assist semantic recognition. New concepts discovered are added to ontology base. Semantic terms like nouns and verbs as well as their semantics recognized in this procedure will be exploited for the technical feature abstraction and annotation module.

4.3. Technical feature abstraction and annotation

Technical features determine the scope of patent protection. We consider technical features of patent as six facets: technical field, design aim, composition, core technique, design principle, and application effect. When people search patents, they usually desire to have a fast comparison on technical features between patents. Technical feature abstraction is to abstract technical features from patent documents and provide the most important content to users in shortened and refined form.

Patent documents follow special patterns, which is helpful for technical feature abstraction. In particular, pattern-based method is more efficient than other statistics-based methods [11]. Yet writing accurate patterns covering all conditions requires laborious work. Therefore, we use pattern learning method. It makes use of the context, learns patterns from the training corpus and applies them to new documents. First, we predefine some regular patterns. Then, with pattern application, new triples (prefixTerms, pattern, postfixTerms) are generated. Both prefixTerms and postfixTerms are segmented and filtered in the previous module, remaining concise and useful words. PrefixTerms and postfixTerms refer to prefix and postfix terms located in the pattern respectively. For example, (A, [Subject] is attached to [Object], B). On the basis of new triples, new patterns are discovered. When multiple candidate patterns appear in the meantime, heuristic constrained rules are used to reduce hypothesis space and eliminate uncertainty. The pattern-learning algorithm is described as follows:

Input: annotated training corpus
Output: set of patterns
Step1: Initialize patterns according to the former pattern set;
Step2:Generate all triples(prefixTerms, pattern, postfixTerms) for each pattern through pattern matching while traversing annotated training corpus;
Step3: Base on two-tuples(prefixTerms, postfixTerms), learn new patterns from corpus;
Step4: When multiple candidate new patterns appear in the meantime, make use of heuristic constrained rules to remove unreasonable patterns and keep reasonable patterns;
Step5: Reasonable new patterns are added into the pattern set.

In this way, new patterns can be applied to generate new triples, and new triples can learn new patterns. Repeating these processes, patterns will be self-expanding by learning.

We first divide patterns into two main categories: product-based patterns and method-based patterns. Product-based patterns are applied to invention patents and method-based patterns are applied to utility model patents. Further, different patterns are grouped and used to abstract technical features in different facets. Initial patterns are constructed in advance.

Technical features are embodied in abstract, claims and description of the patent document. So pattern learning is applied to realize feature abstraction automatically from these sections, based on the results produced in the previous modules. This pattern-learning method can extract technical features fast and precisely, though it requires some initial patterns. Especially, through feature abstraction, semantic association among patents is discovered. For one aspect, explicit association is acquired from priority and background art of patent documents. For the other aspect, implicit association on technical features is known in view of technical feature relation mechanism. The technical feature semantics are then recorded corresponding to patent technical feature ontology. So far, patent semantics have been gotten from various aspects, and then these semantic descriptions are associated with patent documents. In the end, the results of semantic annotation are evaluated, and updated into semantic annotation repository.

5. Case study and discuss

To demonstrate the feasibility of our proposed approach, an illustrative example is provided to explain semantic annotation from the raw document to the semantic representation results. Take a Chinese patent document named “elevator operation plate” (application number: 200820002091) as an example.

1) Structure analysis. Structure information is acquired from open web pages, and then divided into several parts and subparts. For example, application number is “200820002091”; abstract is “The utility model disclosed one kind of elevator operation plate…” , and so on. Each part and its subparts are all obtained. Then, basic parameters of patent content are linked with semantic descriptions, and various parts of patent structure are acquired corresponding to patent structure ontology.

2) Content extraction and semantic recognition. The text content of each part is segmented, filtered and tagged with part of speech. Consider a sentence “The elevator operation plate is constituted by integrated circuit plate, keystroke, spangle, display module, and support plate” in the abstract of the patent as an example. Words are segmented and tagged: “The elevator operation plate/n is constituted by/v integrated circuit plate/n keystroke/n spangle/n display module/n and/conj support plate/n”. Then entities are identified. Semantic recognition analyzes their surroundings to recognize semantics of entities corresponding to concepts and relations in ontology. Concept-level semantics are gotten in this phrase, as depicted in Fig.3.

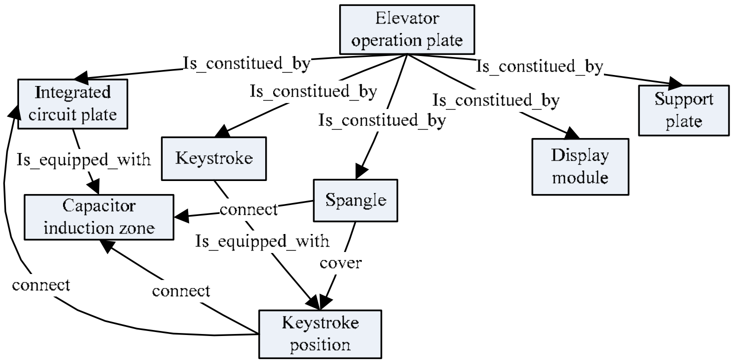
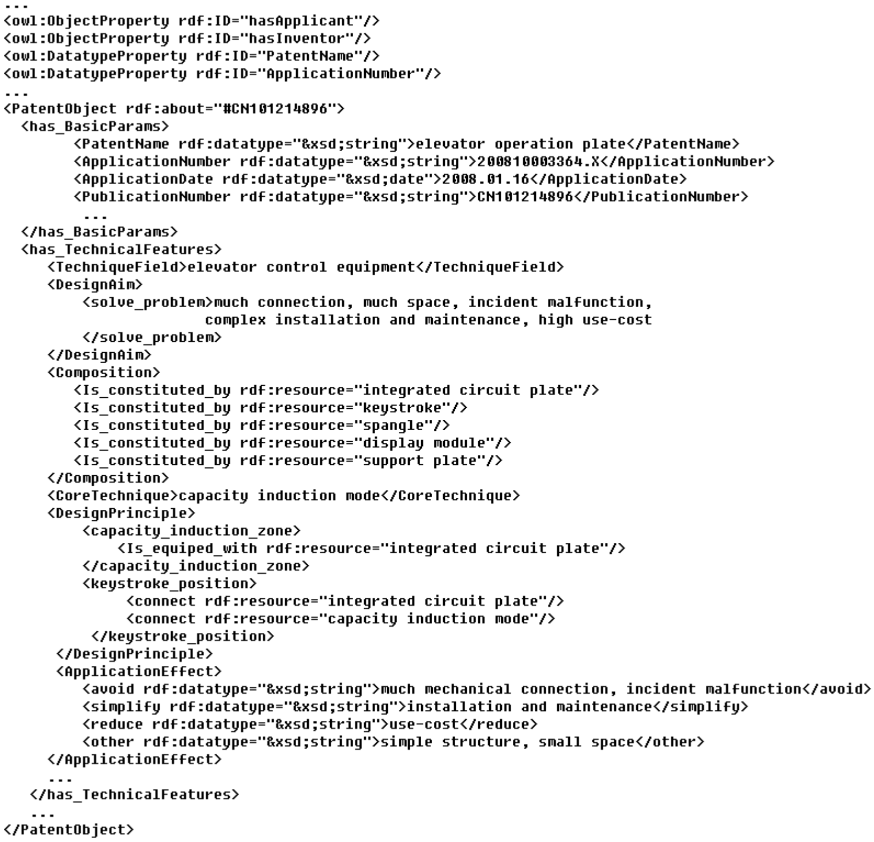


Fig. 3. A portion of concept-level semantics

3) Technical feature abstraction and annotation. This procedure further analyzes the preceding content by pattern learning. Six facets of patent technical feature semantics are abstracted as shown in Fig.4. From this, it’s easy to find relations underlying in one patent, and also across patents. For instance, concepts in one patent are connected by relations, and multiple patents can compare their technical features fast and directly, which are of great interest to users.



```
...
<owl:ObjectProperty rdf:ID="hasApplicant"/>
<owl:ObjectProperty rdf:ID="hasInventor"/>
<owl:DatatypeProperty rdf:ID="PatentName"/>
<owl:DatatypeProperty rdf:ID="ApplicationNumber"/>
...
<PatentObject rdf:about="#CN101214896">
  <has_BasicParams>
        <PatentName rdf:datatype="&xsd;string">elevator operation plate</PatentName>
        <ApplicationNumber rdf:datatype="&xsd;string">200810003364.X</ApplicationNumber>
        <ApplicationDate rdf:datatype="&xsd;date">2008.01.16</ApplicationDate>
        <PublicationNumber rdf:datatype="&xsd;string">CN101214896</PublicationNumber>
         ...
  </has_BasicParams>
  <has_TechnicalFeatures>
     <TechniqueField>elevator control equipment</TechniqueField>
     <DesignAim>
         <solve_problem>much connection, much space, incident malfunction,
                       complex installation and maintenance, high use-cost
         </solve_problem>
     </DesignAim>
     <Composition>
        <Is_constituted_by rdf:resource="integrated circuit plate"/>
        <Is_constituted_by rdf:resource="keystroke"/>
        <Is_constituted_by rdf:resource="spangle"/>
        <Is_constituted_by rdf:resource="display module"/>
        <Is_constituted_by rdf:resource="support plate"/>
     </Composition>
     <CoreTechnique>capacity induction mode</CoreTechnique>
     <DesignPrinciple>
         <capacity_induction_zone>
             <Is_equiped_with rdf:resource="integrated circuit plate"/>
         </capacity_induction_zone>
         <keystroke_position>
              <connect rdf:resource="integrated circuit plate"/>
              <connect rdf:resource="capacity induction mode"/>
          </keystroke_position>
      </DesignPrinciple>
      <ApplicationEffect>
         <avoid rdf:datatype="&xsd;string">much mechanical connection, incident malfunction</avoid>
         <simplify rdf:datatype="&xsd;string">installation and maintenance</simplify>
         <reduce rdf:datatype="&xsd;string">use-cost</reduce>
         <other rdf:datatype="&xsd;string">simple structure, small space</other>
     </ApplicationEffect>
     ...
    </has_TechnicalFeatures>
    ...
</PatentObject>
```

Fig. 4. Sample of OWL-based semantic annotation results

From the above case study, we can find that multi-level semantic knowledge has been obtained from the patent document with our approach, compared with other related research [1-7, 14]. At the macro level, structure-level semantics are acquired. At the micro level, concept-level semantics are recognized. Most importantly, technical feature semantics are abstracted. Using this semantic annotation, users can effectively retrieve required information based on available semantic knowledge. Relations between concepts in patent documents and correlations on technical features across patent documents can also be utilized to facilitate further knowledge visual navigation. In addition, the proposed approach is implemented in MyEclipse with Java language. Though we concentrate on Chinese patents, it can be easy to be adjusted to other language patents only needing little modification.

6. Conclusion and future work

This paper proposes a novel ontology-based automatic semantic annotation approach, which produces multi-level semantic knowledge to enhance patent semantic retrieval in product innovation design. Semantic model is established by multiple ontologies from different perspectives and various granularities, which realizes unified semantic representation and is also flexible and scalable. The approach takes advantage of both structure and content characteristics of patent documents to present corresponding methods: using template schemes to obtain structure information; utilizing domain knowledge and natural language processing techniques to identify semantics of entities and relations; incorporating a heuristic pattern learning method to abstract concise technical features and discover semantic association not only between concepts in single patent document but also across patents.

Our project is still going on. Future work will look into developing more diverse domain-specific ontologies in order to capture extensive semantics. Technical feature abstraction in this study mainly relies on text content, but media objects like figures containing valuable semantics should also be considered. As multilingual technical feature abstraction is demanded urgently with economic globalization, we will deal with the multilingual semantic abstraction other than Chinese patent documents. Additionally, we will take into account patent documents from different technical areas in order to improve our approach more efficiently.

Acknowledgements

This work is supported by the National High Technology R&D Program of China (No.2009AA04Z151), and the Zhejiang Natural Science Foundation of China (No.Z1090461).

References

[1] Trappey A, Trappey C, Wu C. Automatic patent document summarization for collaborative knowledge systems and services. *J Syst Sci Syst Eng* 2009, 18(1): 71-94.

[2] PatentCafe, http://www.patentcafe.com

[3] Wanner L, Baeza-Yates R, BruGmann SR, Codina J, Diallo B, et al.. Towards content-oriented patent document processing. *World Patent Information* 2008, 30(1): 21-33.

[4] SooPAT, http://www.soopat.com/

[5] Mukherjea S, Bamba B, Kankar P. Information retrieval and knowledge discovery utilizing a biomedical patent semantic web. IEEE T Knowl Data En 2005, 17(8): 1099-1110.

[6] Cascini G, Russo D, Zini M. Computer-Aided Patent Analysis: finding invention peculiarities. in *IFIP International Federation for Information Processing, Trends in Computer Aided Innovation*, Boston: Springer, 2007, 250: 167-178.

[7] Tseng Y, Lin C, Lin Y. Text mining techniques for patent analysis. *Inform Process Manag* 2007, 43(5): 1216-1247.

[8] Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, et al.. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web* 2006, 4(1): 14-28.

[9] Popov B, Kiryakov A, Kirilov A, Manov D, Ognyanoff D, et al.. KIM-Semantic Annotation Platform. *The SemanticWeb-ISWC* 2003: 834-849.

[10] Kiyavitskaya N, Zeni N, Cordy JR, Mich L, Mylopoulos J. Cerno: Light-weight tool support for semantic annotation of textual documents. *Data Knowl Eng* 2009, 68(12): 1470-1492.

[11] Tao C, Embley DW. Automatic hidden-web table interpretation, conceptualization, and semantic annotation. *Data Knowl Eng* 2009, 68(7): 683-703.

[12] Li Z, Liu M, Anderson DC, Ramani K. Semantics-Based Design Knowledge Annotation and Retrieval. *Proc.IDETC/CIE* 2005: 799-808.

[13] Cui H, Boufford D, Selden P. Semantic annotation of biosystematics literature without training examples. J *Am Soc Inf Sci Tec* 2010, 61(3): 522-542.

[14] Ghoula N, Khelif K, Dieng-Kuntz R. Supporting Patent Mining by using Ontology-based Semantic Annotations. *IEEE/WIC/ACM Int. Conf. Web Intelligence* 2007: 435-438.

[15] Lin LF, Zhang WY, Lou YC, Chu CY, Cai M. Developing manufacturing ontologies for knowledge reuse in distributed manufacturing environment. *Int J Prod Res* 2010, 49(2): 343-359.

[16] Liu S, McMahon CA, Darlington MJ, Culley SJ, Wild PJ. A computational framework for retrieval of document fragments based on decomposition schemes in engineering information management. *Advanced Engineering Infomatics* 2006, 20(4): 401-413.

[17] Dong Z, Dong Q. HowNet - A Hybrid Language and Knowledge Resource. *Int. Conf. on Natural Language Processing and Knowledge Engineering*, 2003.

[18] Miller GA. WordNet: a lexical database for English. *Commun Acm* 1995, 38(11): 39-41.