Extracting Semantic Information from Chinese Language Patent Claims

Yong Tang[1], Shihan Yang[2], Jiwen Chai[1], and Shanmei Liu[1]

[1] Sichuan Electric Power Research Institute, Chengdu, China
[2] Chengdu Documentation and Information Center, CAS, Chengdu, China
dr.yangsh@gmail.com

Abstract. A big challenge for automatically analyzing patent claims written in Chinese language is how to obtain semantically information from claims which are usually a non-structured, but free text. In this paper, a set of techniques is been provided to extract some valuable semantically information from claims in Chinese language. The method could automatically discovery some usable semantically information from the patent claim texts by means of regular expression pattern and predefined ontology model. It can extract not only surface semantic information but also deeper semantic information. Furthermore, the extracted semantically information is automatically translated into web ontology language (OWL), a machine readable semantic structure specification. The work proposes a potential semantic solution in Chinese language patent analyzing based on patent claims, such as semantic searching, legal status and invalidation checking, and discovering new technique trends. A case study on electronic engineering domain patent claims is also provided.

Keywords: Semantic Information Extraction, Patent Claim; Chinese Language; Patent Automatically Analyzing; Ontology.

1 Introduction

A patent is a set of exclusive rights granted by a national government to an inventor or their assignee for a limited period of time in exchange for a public disclosure of an invention. With the size of patents rapidly growing, and with the written style of them being free, it is a big challenge to conduct analysis, comparison, classification and retrieval of patents at semantically level. Patent legal status checking, patent examination and patent invalidation checking need to analyze claims semantically. These usually can be done by domain experts retrieving and reading all potential relevant patents. It is very tedious and costly. Our purpose of this work is to develop an automatic technique to extract some valuable semantic information from Chinese language patent claim and facilitate the semantically analyzing of that.

FASTUS [1] is a system using cascaded finite state automata to extract some semantic relations from free texts. Three steps are included: (a) phrase recognition; (b) pattern recognition in a phrase; and (c) merge the information found. But it can just find "desired" information, some patterns defining the requirement. On the other words, it can not "discovery" implicit semantic information.

Z. Du (Ed.): Intelligence Computation and Evolutionary Computation, AISC 180, pp. 547–556.
springerlink.com

[2] Proposes a set of techniques to extract the semantic relations from patent claims in English. A domain specific regular expression is employed. The method can find some explicit relations among patent claims, but can not find any implicit relations. Furthermore, English language is very different with Chinese language, so the method can not be facilitated to analyzing of Chinese patent claims.

[3] Introduces regular expressions to extract information from Chinese patent, but it obtains information based on the structure of the whole patent document. It does not consider the meaning of the free text in the patent document, i.e., it is not involved in Chinese text information processing. [4] Employs text-mining techniques to analyze Chinese patents, in order to qualify the novelty degree of a patent. It is also involved in chew on patent claims and abstract, but it just evaluates words (concepts or terms) mined from there. It does not extract and evaluate relations hidden in claims. Our method focus on analyzing patent claims in Chinese, furthermore on extracting semantic relations hidden in that.

2 Chinese Word Segmentation

In this paper, we substitute ICTCLAS [5], the first and the best international Chinese word segmentation tool, for challenging the complex problem of Chinese word segmentation. The lastly version is ICTCLAS2011 [6]. ICTCLAS is a competitive Chinese lexical analyzer, and its frame is the unified HHMM-based (Hierarchical Hidden Markov Model [7] based).

ICTCLAS provides an interface to configure user-defined dictionary. The tool could do better for the special domain word segmentation by means of defining some proper terms in user dictionary. The following is an example, the free Chinese text coming from the claim of Chinese patent No. CN1046065.

Example 1: yi zhong yan shi zi dong guan bi la xian kai guan, you xian la qi gang ti, gui wei tan huang, gu ding huo sai, dan xiang mi feng quan, dong chu dian, jing chu dian, qi guang qian la heng gan, chang kai ka feng deng gou cheng. (The delayed auto power-off electronic pull switch consists of pull cylinder block, return spring, stationary piston, one-way seal ring, traveling contact, stationary contact, cylinder draw-off crossbar and normal open slit.)

The following is result of Chinese word segmentation without user-defined dictionary. The result is trivial, and does not express the precise meaning of fundamental words. The letter(s) behind slash is the mark of part of speech, or user-defined mark in user dictionary. The result shows the segmentation as general morphemes, which can not correctly express the specific concept in the text. For example, the concept cylinder draw-off crossbar is split into cylinder / noun, draw / verb, pull / verb, crossbar / noun, but not treated as a single domain specific noun, so do normal open slit, traveling contact, stationary contact, etc.

yi/m zhong/q yan/vg shi/ng zidong/d guanbi/v laxian/n kaiguan/n , /wd you/p xian/n la/v qiguang/n ti/ng ,/wn guiwei/vi tanhuang/n ,/wn guding/a huosai/n ,/wn danxiang/b mifengquan/n ,/wn dong/v chu/v dian/qt ,/wn jing/ad chu/v dian/qt ,/wn qigang/n qian/v la/v hengguan/n ,/wn chang/d kai/v ka/n feng/n deng/udeng goucheng/n ./wj (one/m kind/q /over/vg time/ng automatic/d close/v pull/n switch/n comma/wd by/p line/n

pull/n cylinder/n body/ng comma/wn return/vi spring/n comma/wn stationary/a pistion/n comma/wn one-way/b seal ring/n comma/wn moving/v touch/v point/qt comma/wn stationary/ad touch/v point/qt comma/wn cylinder/n draw/v pull/v crossbar/n comma/wn normal/d open/v card/n slit/n etc./udeng consist/n period/wj)

A user dictionary like following is defined. One item locates one line, and each item is a pair, which consists of a word and a mark, separated by a blank. Explanations are included in the following parentheses.

yizhong (a kind of) mq (numeral and quantifier)
yanshi (delayed) a (adjective)
zidongguanbi (auto power-off) d (adverb)
xianlaqigangti (pull cylinder block) n (noun)
dongchudian (traveling contact) n
jingchudian (stationary contact) n
qianlahenggan (draw-off crossbar) n
changkai (normal open) d
kafeng (slit) n

With auxiliary of user defined dictionary, the tool gives the result as following, which is closer to the real meaning of the text content than before. And the segmentation provides better lexical expression for the sequential semantically analysis.

Yizhong/mq yanshi/a zidongguanbi/d laxian/n kaiguan/n ,/wd you/p xianlaqigangti/n ,/wn guiwei/vi tanhuang/n ,/wn guding/a huosai/n ,/wn danxiang/b mifengquan/n ,/wn dongchudian/n ,/wn jingchudian/n ,/wn qigang/n qianlahenggan/n ,/wn changkai/d kafeng/n deng/udeng goucheng/n ./wj (a kind of/mq delayed/a auto power-off/d pull/n switch/n comma/wd by/p pull cylinder block/n comma/wn return/vi spring/n comma/wn stationary/a piston/n comma/wn one-way/b seal ring/n comma/wn traveling contact/n comma/wn stationary contact/n comma/wn cylinder/n draw-off crossbar/n comma/wn normal open/d slit/n)

In the work, Chinese patent claims semantically analyzing, we always provide user defined dictionary for the special domain. The preparation work provided by domain experts or came from sharing domain ontology is really needed, because the text of the patent claims might contain a lot of domain specific terminology.

3 Semantically Pattern

We employ two semantically extracting patterns to discovery semantically information after Chinese patent claims lexical analyzed with the ICTCLAS tool. Firstly, regular expression pattern can find the explicit semantic information, which has precise semantically structure, such as subject-predicateobject structure, adjective-noun structure. Moreover, desired semantically pattern can also be made with regular expression. Secondly, ontology schema pattern can the find implicit semantic information, which can not be obtained directly from visual text, but which can be reasoned out. For example, from the free text “wo jiao li qiang, wo ba ba shi li gang

(My name is LiQiang, and my father's name is LiGang.)", we can discovery a "fuzi" (father-son) relationship between "ligang" and "liqiang", which can not be found by regular expression. The implicit semantic information like this can be discovered by ontology and reasoning.

3.1 Regular Expression Pattern

For getting useful explicit semantic information, we utilize domain specific regular expressions. Java regular expression specification (a part of JSR 51 [8]) is adopted in this paper, and the part of speech tags in ICTCLAS [6] is also included in the work. Five kinds of regular expressions have been identified to extract semantic information in Chinese patent claims.

- Type 0 : Claim type

Patent claims are the part of a patent or patent application that defines the scope of protection granted by the patent. The claims define, in technical terms, the extent of the protection conferred by a patent, or the protection sought in a patent application. The claims are of the utmost importance both during prosecution and litigation. There are two basic types of claims: independent and dependent claims. Type 0 regular expression pattern can distinguish them.

The independent claims stand on themselves and do not depend on other claims. Example 2 shows an independent claim from the Chinese language patent CN2852371.

*Example 2 (*independent *claim)*: 1.yizhong zhineng fanghuo kaiguan, ta baokuo duanluqi, tuokou zhuangzhi, dianliu huganqi, qi tezheng zaiyu:…(Claim 1. A kind of intelligent-controlled fireproofing switch consists of disconnector, release device, and current transformer. It features: …)

The dependent claims refer back to another claim and generally express particular embodiments as fall-back positions. They incorporate by reference to prior claims. Example 3 and 4 show two dependent claims in Chinese patent CN2852371.

*Example 3 (*dependent *claim)*: 2. **genju quanli yaoqiu 1 suosu** de zhineng fanghuo kaiguan, **qi tezheng zaiyu**: hai baokuo guangdian chuanganqi, suosu de shuju chuliji (7) de shuru he guangdian chuanganqi de shuchu lianjie.(Claim 2. According to claim 1 with intelligent-controlled fireproofing switch, characterized in that: …)

Example *4 (*dependent *claim)*: 3. **genju quanli yaoqiu 1 suosu** de zhineng fanghuo kaiguan, **qi tezheng zaiyu**: hai baokuo wendu chuanganqi, suosu de shuju chuliji (7) de shuru he wendu chuanganqi de shuchu xiang lianjie. (Claim 3. According to claim 1 with intelligent-controlled fireproofing switch, characterized in that:…)

Some domain specific regular expression for claim types are as follows, where /n is a noun mark, /m a numeral, and /w a Chinese punctuation. Other symbols are standard marks from Java regular expression specification.

regClaimType_Independent = *"yizhong"/n\S* "qi tezheng zaiyu"\n/S*| ("n (d+))"\n/S*;*

*regClaimType _Dependent =/m(/w?)"*genju *quanli yaoqiu" \(d+) "suosu"\S*|\s*;*

- Type 1 : Components

The type of regular expressions is to extract components in the claims. The invention components (concepts) are the most important element in an invention. They are similar to general concept domain term, because applicants wish to have more right protection. Therefore they should be extracted by more flexible expression.

A domain specific regular expression for components of Chinese patent claims is as follows.

```
term = "DT";
regComponents = \S* (/n+ |term+)\nS*;
```

In the regular expressions above, DT is some one domain special term, included in user defined dictionary. Example 5 shows components extraction from Chinese patent CN2852371 in Example 2, where bold-face words represent the extracted information, (here are concepts of electronic domain) similarly hereinafter.

Example 5 (components): 1. yizhong zhineng fanghuo kaiguan, ta baokuo duanluqi, ta baokuo duanluqi, tuokouzhuangzhi, dianliu huganqi, qi tezheng zaiyu: hai baokuo kaiguan dianyuan(10), dianya bianhuan danyuan(1), baojing danyuan(9), tiaojie danyuan(8), qianzhi danyuan(3), shujuchuliji(7), zhixing danyuan(6), dianya bianhuan danyuan(1) de shuchu tongguo qianzhi danyuan(3) yu shujuchuliji(7) shuru xiang lianjie, shujuchuliji(7) de shuchu fenbie yu zhixing danyuan(6), baojing danyuan(9) de shuru xiang lianjie, shujuchuliji(7) yu tiaojie danyuan(8) xiang lianjie, zhixing danyuan(6) de shuchu yu tuokou zhuangzhi(5) xiang lianjie, dianliu huganqi(2) tongguo qianzhi chuli danyuan(3) yu shujuchuliji(7) de shuru xiang lianjie. (switch, disconnector, release device, current transformer, power supply, preposition unit, voltage change unit, warning unit, accommodation unit, data processor, executing unit etc.)

- Type 2 : Attribute

The type of regular expressions can obtain the attribute (has-a and is) relationship. Following shows three regular expression for finding attributes in Chinese patent claims, where /a is adjective mark.

```
regAttribute_1 = /a+ \S*/n;
regAttribute_2 = /n "shi"\S* (/n+|/a+);
regAttribute_3 = /n "you"\S*/n+;
```

Example 6 shows result of the patent CN2852371 that "kaiguan de shuxing shi fanghuo de, shi zhineng de" (the switch is fireproof and intelligent controlled).

Example 6: 1. yizhong **zhineng fanghuo** kaiguan, … (a kind of fireproof and intelligent switch, …)

- Type 3 : Containment relationship

The type of regular expressions can extract the containment (part-of) relationship between components. Regular expressions belong to this type like the following, and its result is shown in Example 7.

regContainment = (/n\S "baokuo"(/n/w+)*)|(/n\S* "you"nS* "zucheng"| "goucheng"/w+);*

Example 7: 1. yizhong zhineng fanghuo kaiguan, ta baokuo duanluqi, ta baokuo duanluqi, tuokouzhuangzhi, dianliu huganqi, qi tezheng zaiyu: hai baokuo kaiguan dianyuan(10), dianya bianhuan danyuan(1), baojing danyuan(9), tiaojie danyuan(8), qianzhi danyuan(3), shujuchuliji(7), zhixing danyuan(6),… (Includes / consists of)

- Type 4 : Spatial relationship

The type of regular expressions is to be used to extract spatial relationships among components from a claim. The following is a sample regular expression belongs to the type, and Example 8 tells the result.

regSpatial = (/n + ["tongguo"]\S "yu"/n + "xiang lianjie"| "jiehe") | (/n + "zai"| "chuyu"\S*/n + "shang"| "zhong"| "xia"| "nei"| "wai");*

Example 8: … dianya bianhuan danyuan(1) de shuchu tongguo qianzhi danyuan(3) yu shujuchuliji(7) shuru xiang lianjie, shujuchuliji(7) de shuchu fenbie yu zhixing danyuan(6), baojing danyuan(9) de shuru xiang lianjie, shujuchuliji(7) yu tiaojie danyuan(8) xiang lianjie, zhixing danyuan(6) de shuchu yu tuokou zhuangzhi(5) xiang lianjie, dianliu huganqi(2) tongguo qianzhi chuli danyuan(3) yu shujuchuliji(7) de shuru xiang lianjie. (voltage change unit is connected with data *processor* by preposition unit, data processor is connected with inputting interfaces of executing unit and warning data, executing unit's outputting is connected with release device, and current transformer is connected with input of data processor also by preposition unit.)

4 Ontology Schema

For discovering valuable implicit semantic information, domain specific ontology has been employed. With reference to the domain specific ontology, we can get other semantic information, which can't be extracted only by regular expression pattern. The main idea is that we apply concepts and facts found by regular expression pattern into the reasoning on domain specific ontology to discovery implicit semantic information of patent claims. For the purpose, three types of semantic expansion operator are defined as follows:

*Definition 1 (*Inheritance *expansion)*: If concept *A* is a child of concept *B (A* $\subseteq$ *B)*, and *x* is an individual of *A (x* $\in$ *A)*, then *x* is also an individual of *B (x* $\in$ *B)*.

In the definition 1, *x* is the fact that been found by regular expression pattern. By the inheritance expansion operator, new concepts from the ontology and new assertions (relations) could be introduced into extracted semantic result. This could increase accuracy and completeness about semantics of patent claims.

*Definition 2 (*Attribute *expansion)*: If concept *C* has *n* attributes, where an attribute a_1 with weight $w_1 = 0.6$, an attribute a_2 with weight $w_2 = 0.2$, an attribute a_3 with weight $w_3 = 0.1$, … , etc., and where the sum of all weights is 1,and *x* has attributes a_i, … , a_j ,*(i, j*$\in$ *[1, n])*, then *x* is an individual of *C* with weight $w_i + w_{i+1} + \ldots + w_j$.

In fact, definition 2 provides a method to decide how much an extracted fact or concept is associated with another concept among domain specific ontology, and the weight of attributes to its concept is used. The weight of an attribute expresses how much it can distinguish its owner with other concepts, a value between *0* and *1* given by experts or computed by machine learning.

*Definition 3 (*Combination *expansion)*: If concept *C* consists of components $C_1, \ldots, C_n$, and *X* is a child of concept C_i, $i \in [1, n]$ (i.e. $X \subseteq C_i$), then *X* is part of *C*.

Combination expansion can introduce new *part-of* relations into the extracted semantic result. All three expansion operators can be compounded.

5 Semantic Information Normalization

The extracted semantic information should be shared as knowledge base, so normalization of semantic information is needed. In the paper, OWL [9] is adopted, i.e., the extracted information is expressed as OWL, which is the most popular standard in the semantic web world.

Extracted and expanded semantic information should be translated into OWL specification for knowledge sharing. Fortunately, the task is not too tough. The result of analyzing semantically Chinese patent claims is represented as a Wgraph [10], which is a simple formal semantic graph, a directed graph. [10] Developed a WGraph editing tool that can automatically translate WGraph files into OWL files effectively. In this research, we use the tool to translate extracted and expanded semantic information into OWL.

6 A Case Study

The process of extracting semantic information from Chinese language patent claims is shown in figure 1.

There are five main steps in the process:

1) Chinese word segmentation. It is difficult to split Chinese sentence up to semantic words, because there is no blank among words like English. The ICTCLAS2011[6] does this job better in general Chinese free text than that in domain specific free text, due to lots of technical terms. So user defined dictionary for analyzing Chinese patent claims is needed. Fortunately, ICTCLAS provides the interface to define user special dictionary for segmentation. In this step, ICTCLAS gets Chinese patent claims free text and user defined dictionary as its inputs, and its outputs is Chinese text annotated by predefined or user-defined lexical marks, just like the example in Section 2.

2) Parsing regular expression. With developed domain specific regular expression pattern, the parser can extract some required semantic information from the annotated patent claim, the output of Chinese word segmentation. The extracted information is found on command, and is usually some facts and terms occurred in the claim text in the view of ontology.

3) Reasoning on the domain specific ontology. To discovery deep semantic information in the claim text, ontology reasoner is a good choice. With the domain specific ontology and facts and terms found from the claim, the reasoner can get more semantic information, such as synonymy relations, inheritance relations, and containment relations.

4) Semantic checking and merging. A weak checking for semantic correctness and repetitiveness is done in this step by analyst manually or machine automatically. The result of semantic information is visualized as directed graph by means of WGraph editing tool [10], and it is easy to be adjusted.

5) Translating into OWL. In the step, the semantic information represented by WGraph is automatically translated into OWL file, which then be syntactically and partsemanticallyly validated by OWL validator. Finally the semantics information of the Chinese patent claim is put into Chinese patent knowledge base for public sharing.

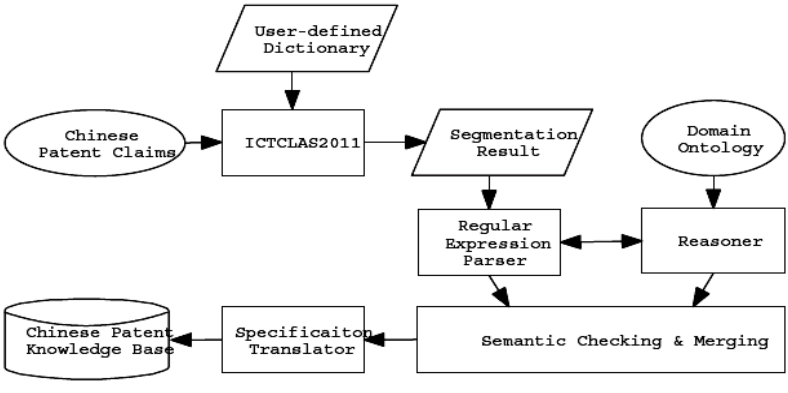

Fig. 1. The process of extracting semantic information

Figure 2 shows the result of extracted semantic information from Chinese patent CN2852371 as WGraph.

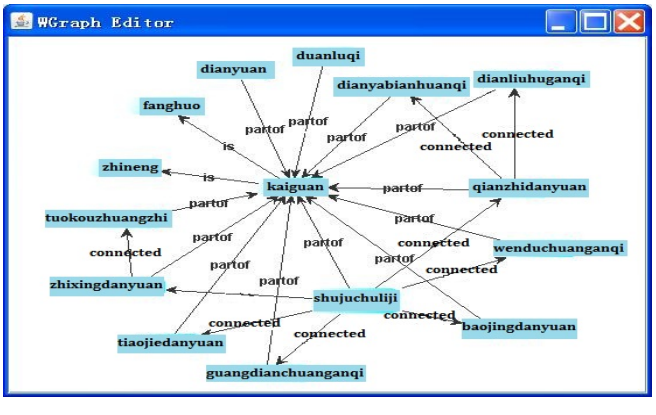

Fig. 2. Visulization of extracted semantic information without expanded

And Figure 3 shows the result after merged and expanded of extracted semantic information, which is the last result of analyzing the patent CN2852371, represented as WGraph file. Some concepts and relations are increased, and shown in the figure with red circles.

The extracted semantic information are confirmed in a sense by electronic domain experts. Then the extracted and expanded semantic information are translated into OWL file by means of the WGraph editing tool. The created OWL file is validated by the online OWL Validator[1].

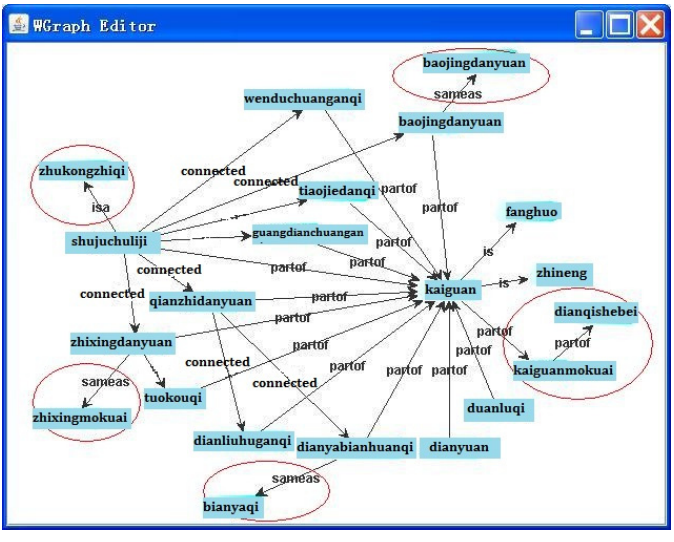

Fig. 3. Visulization of extracted semantic information with domain specific ontology expanded

7 Conclusion and Future works

A set of techniques is proposed to facilitate Chinese patent claims analyzing semantically. The method can not only find surface semantic information by means of domain specific regular expression pattern, in which we introduce 5 types regular expressions, but also discovery deeper semantic information by means of domain specific ontology pattern, in which we define 3 types semantic expansion operators. The result of extracted and expanded semantic information from claims is visually represented as a simple semantic graph (WGraph), which can be effectively translated into ontology language specification (OWL) for public knowledge sharing. An example has also been studied.

Concurrently, the regular expression patterns were devised manually based only on Chinese patent claims in electronic engineering domain, so were the ontology patterns. For other domains, the new regular expressions and ontology should be developed, and this is very tough work for claim analysts. In the near future, some machine learning

[1] http://owl.cs.manchester.ac.uk/validator/

approaches should be employed to building regular expression patterns and domain ontology patterns. Moreover, a better integration tool for analyzing Chinese patent claims will be developed. Based on extracted semantic information from claims, the researches on semantically searching and legal status invalidation checking is also our interest.

Acknowledgements. This research is partially supported by the West Light Foundation of the Chinese Academy of Sciences on the project: Automatically extracting semantically meta-data of digital document resources on science and technology, and by the Talent Training Plan of Chinese Academy of Sciences on the project: Research on patent claims searching and analyzing tools based on formal semantically reasoning.

References

[1] Hobbs, J.R., Appelt, D.E., Bear, J., Israel, D.J., Kameyama, M., Stickel, M.E., Tyson, M.: Fastus: A cascaded finite-state transducer for extracting information from natural-language text. CoRR, vol. cmplg/ 9705013 (1997)

[2] Yang, S.-Y., Lin, S.-Y., Lin, S.-N., Lee, C.-F., Cheng, S.-L., Soo, V.-W.: Automatic extraction of semantic relations from patent claims (2008)

[3] Qingying, Q., Guomin, Z., Pei'en, F., Jianwei, W.: Extraction approach of patent information based on regular expression. Journal of Chinese Machine Engineering 18(19), 2326–2329 (2007) (in Chinese)

[4] Yu-qin, L., Xue-feng, W., Xiao-ping, L.: Quality estimation of patent based on text mining and its empirical research. Computer Engineering and Applications 43(33), 12–14 (2007) (in Chinese)

[5] Zhang, H.-P., Yu, H.-K., Xiong, D.-Y., Liu, Q.: Hhmm-based chinese lexical analyzer ictclas. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, SIGHAN 2003, vol. 17, pp. 184–187. Association for Computational Linguistics, Stroudsburg (2003)

[6] Zhang, H.P.: Ictclas (version 2011), http://cid-51de2738d3ea0fdd.office.live.com/self.aspx/.Public/ictclas2011/ICTCLAS2011-SDK-release.rar (2011)

[7] Fine, S., Singer, Y., Tishby, N.: The hierarchical hidden markov model: Analysis and applications. Mach. Learn. 32, 41–62 (1998), http://portal.acm.org/citation.cfm?id=325865.325879

[8] Reinhold, M., et al.: Jsr 51: New i/o apis for the java platform (2002), http://www.jcp.org/en/jsr/detail?id=51

[9] Motik, B., Grau, B.C., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C.: Owl2 web ontology language profiles, http://www.w3.org/TR/2009/CRowl2-profiles-20090611/ (2009)

[10] Yang, S., Wu, J.: Mapping relational databases into ontologies through a graph-based formal model. In: International Conference on Semantics, Knowledge and Grid, pp. 219–226 (2010)