Measuring standardized-concepts relationship for knowledge-oriented standard information service

Wei Lu
School of Information Management, Wuhan University, Wuhan 430072. Email: weilu@whu.edu.cn

Heng Ding
School of Information Management, Wuhan University, Wuhan 430072.

Abstract: *Standard literature is a kind of important information resources, and it plays an important role in guiding the economic activities. However, standard literature information management system is still staying in the literature retrieval level and is not able to provide knowledge-oriented standard information service. Ontology and semantic technologies are widely regarded as powerful approach for organization and management of knowledge. However, constructing the relationship between concepts in standard literatures manually is very time-consuming and expensive. In this paper, we unified natural language processing, semantic computing, and made full use of the existing knowledge bases to build a standardized-concepts relation network automatically. Based on the concept network, we designed a knowledge-oriented standard information service system to provide standard knowledge service such as standard knowledge map and knowledge-oriented standard retrieval.*

Keywords: *knowledge organization, standard literature, concept network*

Introduction

The development of information technology has changed the way of information transmission, and it also has great influence on contents, methods of information science. Specially, the variety of digital information resources enriches the research objects of information science. Nowadays, studies are not only limited to academic literatures, but also focused on webpages, books, patents and standard literatures (Z. Li & Zhang, 2012; Luo, Yu, Zheng, & Jin, 2012; Ma & Gao, 2010). From the perspectives of both information chains (Ma, 2013) and basic concepts of information science (Zheng & Hua, 2011), knowledge management and service are top-level goals of information science study. Standard literature is a kind of important information source and knowledge carrier in digital age, and the standard literature service is facing with the transformation from information service to knowledge service (Guo, 2011). However, the recent standard literature service still remains in document-level (Deng, 2008; W. Li, Wang, & Gu, 2010; Liu & Zhong, 2011), and it is not able to fulfill the intellectual hunger. Many research indicated that ontology and semantic technologies are very useful for standard knowledge service (Alani et al., 2003; Berners-Lee & Hendler, 2001; Ghoula, Khelif, & Dieng-Kuntz, 2007; Müller, Kenny, & Sternberg, 2004). But constructing the relationship among concepts in standard literatures manually is very time-consuming and expensive. Based on above issues, we unified natural language processing, semantic computing, and made full use of the existing knowledge bases to build a standardized-concepts relation network automatically. Based on the concept network, we designed a knowledge-oriented standard information service system to provide standard knowledge services such as standard knowledge map and knowledge-oriented standard retrieval.

Knowledge-Oriented Standard Information Service System

Problem and Solution

According to our investigation, the main difficulties in knowledge-oriented standard information service are the following aspects:

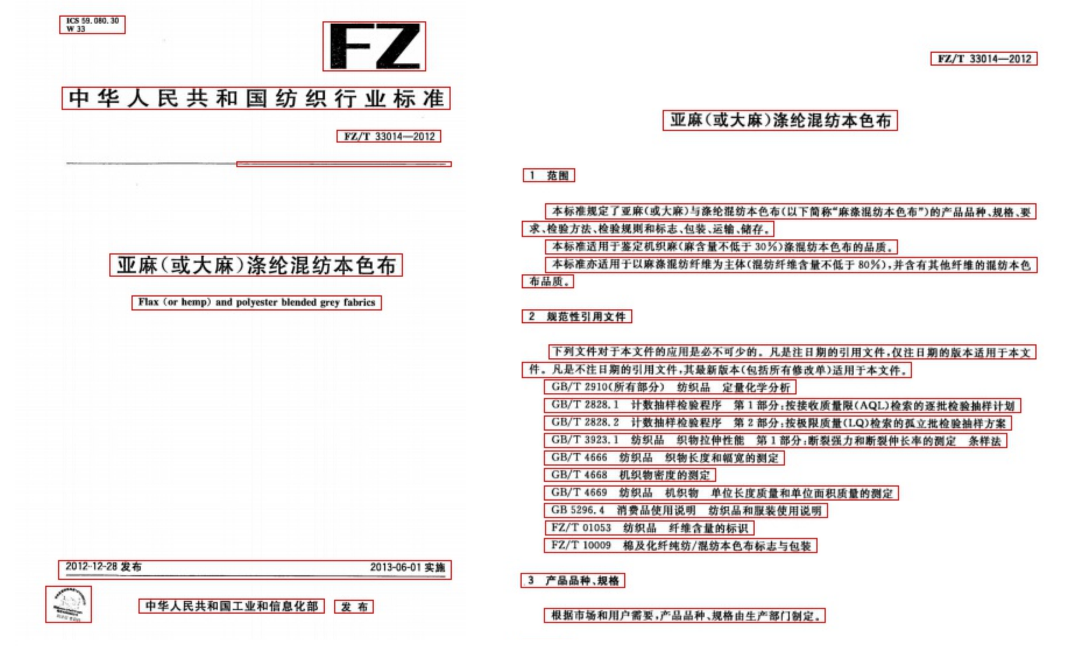
a) The deficiency of semantic data. Due to the copyright issue, most of available Chinese standard literatures are scanned-documents in Portable Document Format (PDF). Although it is easy to extract plain text from these files by using general OCR software, the semantic structure information such as standard serial number, title, reference standard, alternate standard, publisher, implementation dates and standard terms lost in the process of OCR. The plain text is not conducive to semantic processing and knowledge organization.

b) The excessive cost of ontology construction. The size and complexity of standard-knowledge-ontology make it very hard to build it manually.

In light of the above questions, we proposed a solution containing two critical steps:

a) The semantic re-structuralization of standard literatures. In this step, we convert the original scanning documents of Chinese standard literatures to a XML file with rich semantic information. Figure 1 shows an example.

b) Automatic building a concept network for the expression of relationship between standardized-concepts. A standard literature is a document that provides requirements, specifications, guidelines or characteristics for some materials, products, processes and services. Thus, we think that the core concepts in standard literature are terms of materials, products, processes and services, and the relationship between these terms has formed a network revealing structure of knowledge in standard literatures. In this paper, we called these terms "standardized-concepts". In this step, we unified natural language processing, semantic computing, and made full use of the existing knowledge bases to measure the relationship between standardized-concepts and to build a concept network automatically.



ICS 59.080.30
W 33

FZ

中华人民共和国纺织行业标准

FZ/T 33014—2012

亚麻(或大麻)涤纶混纺本色布

Flax (or hemp) and polyester blended grey fabrics

2012-12-28 发布　　2013-06-01 实施

中华人民共和国工业和信息化部　发布

FZ/T 33014—2012

亚麻(或大麻)涤纶混纺本色布

1　范围

本标准规定了亚麻(或大麻)与涤纶混纺本色布(以下简称"麻涤混纺本色布")的产品品种、规格、要求、检验方法、检验规则和标志、包装、运输、储存。

本标准适用于鉴定机织麻(麻含量不低于30%)涤混纺本色布的品质。

本标准亦适用于以麻涤混纺纤维为主体(混纺纤维含量不低于80%),并含有其他纤维的混纺本色布品质。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是不注日期的引用文件,其最新版本(包括所有修改单)适用于本文件。

GB/T 2910(所有部分)　纺织品　定量化学分析

GB/T 2828.1　计数抽样检验程序　第1部分:按接收质量限(AQL)检索的逐批检验抽样计划

GB/T 2828.2　计数抽样检验程序　第2部分:按极限质量(LQ)检索的孤立批检验抽样方案

GB/T 3923.1　纺织品　织物拉伸性能　第1部分:断裂强力和断裂伸长率的测定　条样法

GB/T 4666　纺织品　织物长度和幅宽的测定

GB/T 4668　机织物密度的测定

GB/T 4669　纺织品　机织物　单位长度质量和单位面积质量的测定

GB 5296.4　消费品使用说明　纺织品和服装使用说明

FZ/T 01053　纺织品　纤维含量的标识

FZ/T 10009　棉及化纤纯纺/混纺本色布标志与包装

3　产品品种、规格

根据市场和用户需要,产品品种、规格由生产部门制定。

```
<standard>
<standard_serial_number>FZ/T 33014-2012</standard_serial_number>
<title_in_chinese>亚麻(或大麻)涤纶混纺本色布</title_in_chinese>
<title_in_english>Flax (or hemp) and polyester blended grey fabrics</title_in_english>
<implementation_dates>2013-06-01</implementation_dates>
<publisher>中华人民共和国工业和信息化部</publisher>
<reference_standard>
<item>GB/T 2910(所有部分) 纺织品 定量化学分析</item>
<item>…</item>
</reference_standard>
…
</standard>
```

Figure 1 An example of the semantic re-structuralization of standard literatures

System Architecture

Figure 2 shows the entire architecture of our knowledge-oriented standard information service system. Follow the service-oriented architecture (SOA), there are three layers in our system including data layer, basic service layer and application layer.

- ***Data layer*** provides the service of data organization and storage, and it also defines the standard of data interaction between different modules of system.
- ***Basic service layer*** provides the basic service of data processing and calculation. In this layer, we integrate some basic service modules that we used in the two critical steps, including “word segmentation”, “part-of-speech tagging”, “web crawler”, “semantic computing”, “layout analyzer”, and so on. We also provide some applying-oriented interfaces, such as “data indexing”, “information visualization”.
- ***Application layer*** provides the knowledge-oriented standard information service, such as knowledge map, knowledge-oriented standard retrieval and some of other user-oriented services.

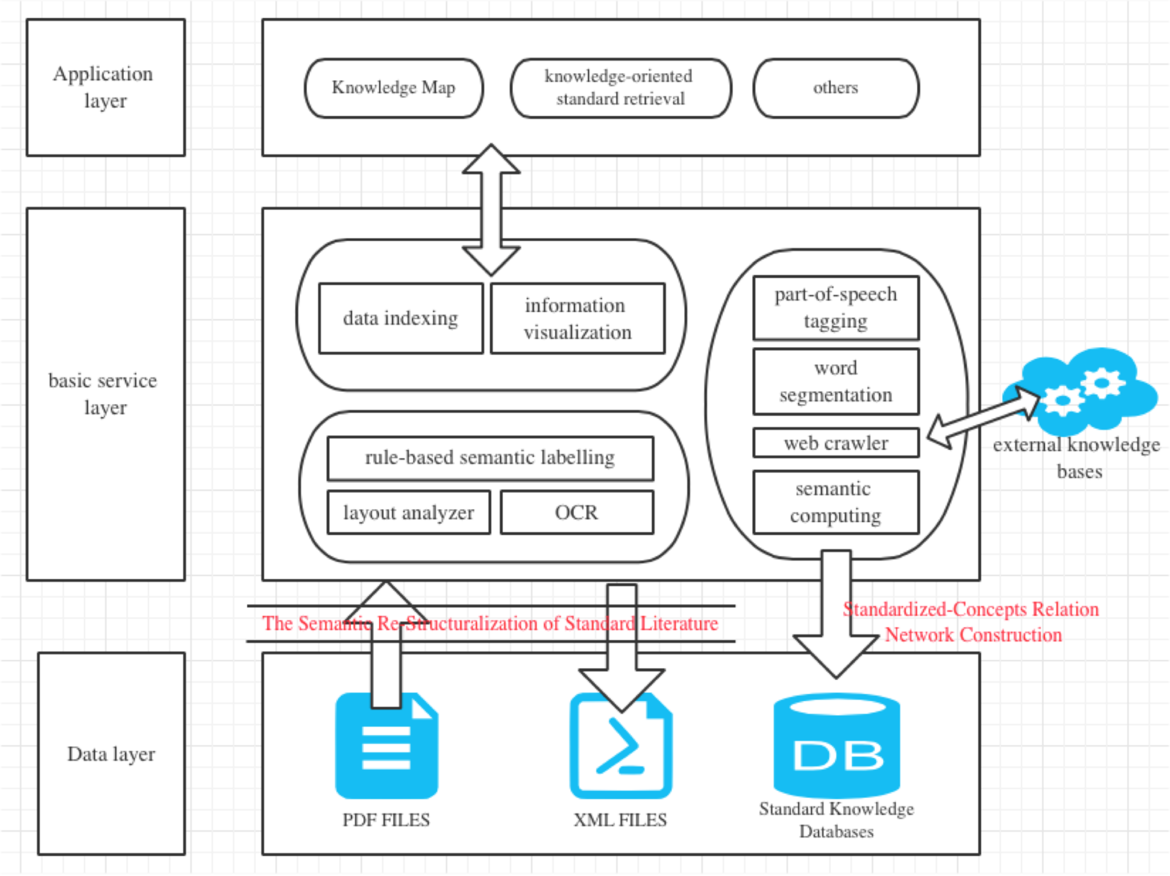

Application layer
Knowledge Map
knowledge-oriented standard retrieval
others
basic service layer
data indexing
information visualization
rule-based semantic labelling
layout analyzer
OCR
part-of-speech tagging
word segmentation
web crawler
semantic computing
external knowledge bases
The Semantic Re-Structuralization of Standard Literature
Standardized-Concepts Relation Network Construction
Data layer
PDF FILES
XML FILES
DB
Standard Knowledge Databases

Figure 2 architecture of knowledge-oriented standard information service system

The Semantic Re-Structuralization of Standard Literature

As we mentioned above, plain text obtained by general OCR software is lack of semantic structure information such as standard serial number, title, reference standard, alternate standard, publisher, implementation dates and standard terms, and it is not conducive to semantic processing and knowledge organization. The aim of the semantic re-structuralization of standard literature is to capture semantic structure information during the process of text extraction. In our knowledge-oriented standard information service system, we designed a sub-module to achieve this function. For each standard document uploaded by users, we firstly used a layout analyzer to detect text blocks and corresponding information including page number, bounding box coordinates, font size. Secondly, we used an open source OCR software to obtain the content (characters) of text blocks, and we designed a rule-based algorithm to annotated these text blocks with semantic structure labels. Table 1 lists part of semantic structure labels. Due to the formal textual structure of Chinese standard literature, it is easy to design a rule-based method to extract semantic information. For example, the section heading always locates in single text line and starts with a number and punctuation, and the reference standards are always listed in the section of “Normative References”. Nevertheless, sometimes there is a little different in the text of the section heading, Levenshtein Distance (Heeringa, 2004), also termed Edit Distance, is a good way to measure the similarity of text string, and provide an automatic method to optimize the result obtained by rule-based character matching.

Table 1 Part of semantic structure labels

Xml Labels	Details
standard_serial_number	unique identifier of standard document
title_in_Chinese	Chinese title of the document
title_in_English	English title of the document
publish_dates	the publish date of the document
implementation_dates	the effective date of the document
reference_standards	reference standard documents
standard_terms	terms defined in the document
Provisions	describing the main content of the document

Standardized-Concepts Relation Network Construction

Figure 3 shows the procedure of standardized-concepts relation network construction. At first, we extracted concept terms from different sections of standard XML files via word segmentation and part-of-speech tagging. Considering that a standard document often provides requirements, specifications, guidelines or characteristics for a special object (such as a material, a product, a process or a service), and the special object (standardized-concept) always locates in document title. Therefore, most of standardized-concepts can be extracted from titles. Secondly, we obtained relevant documents and explanatory texts from existing knowledge bases (wiki encyclopedia, Baidu encyclopedia, standard glossary, etc.), and we also extracted relevant concept terms in this step. Thirdly, we used a semantic computing method for measuring the relationship among concepts. Figure 4 describes the semantic computing method in detail. For each standardized-concept E, there is a relevant document set D = {d_1, d_2, …, d_n} that we obtained from existing knowledge bases. S = {s_1, s_2, …, s_t} are the sentences in document d_i. t defines the position of sentence. Following this way, the important concepts in standard literatures are well organized and linked in a concept-term-network. In the network, each node is a unique standardized-concept. The edge represents the relevance between two adjacent concept terms. And the relevance between any concept terms can be calculated with the network path.

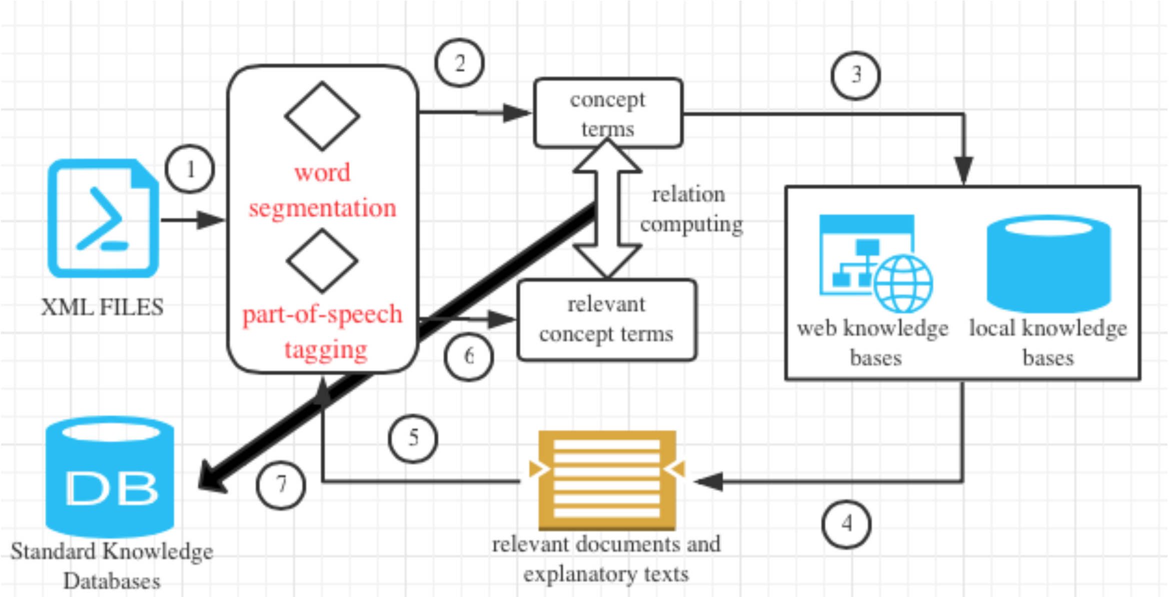


Figure 3 The procedure of standardized-concepts relation network construction

```
rel(E, $w_k$) = 0
n = len(D)
for d in D:
    co-occurrence = 0
    S = get_sentence(d)
    m = len(S)
    for s in S:
        t
```

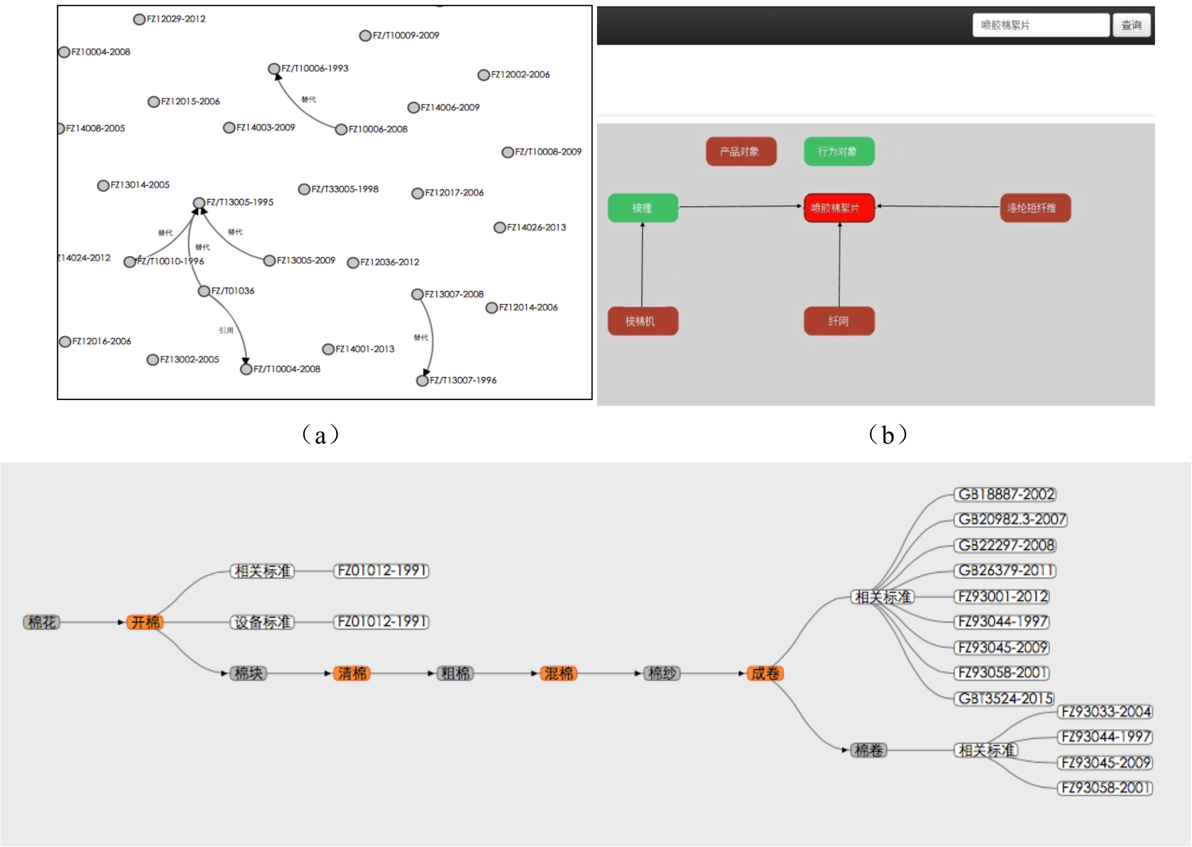
```
            W = pos_cut(s)
            W = remove(s)
            if $w_k$ in W and E in $w_k$:
                co-occurrence += 1
            elif $w_k$ in W and E not in $w_k$:
                co-occurrence += 1/sqrt(t)
        rel(E, $w_k$) += co-occurrence/m
rel(E, $w_k$) = rel(E, $w_k$)/n
```

Figure 4 co-occurrence based standardized-concept-relation algorithm

Application and Conclusion

By cooperating with standard service institution, we apply our system in the standard information service of textile industries. In total, we obtained 2268 Chinese standard documents and 5965 English standard documents about textile industries from the cooperative organization. We have attempted to provide knowledge map and knowledge-oriented standard retrieval for users.

- ***Knowledge Map.*** We provide a good approach to understand the relationship between knowledge units via information visualization. Since we have extracted the semantic information by semantic re-structuralization and organized it in XML files, it is very easy to discover the relationship between standard literatures. Figure 5 (a) shows an example, we can see that "FZ/T 10006-1993" is replaced by "FZ 1006-2008" and "FZ/T 01036" cites "FZ/T 10004-2008". In addition, our system also has the ability to display the relationship between standardized-concepts (Figure 5(b)). We can find that "喷胶棉絮片" is related with "涤纶短纤维", "梳棉机" and "纤网". We also provide a visualized interface for users to browse the relevant standard documents in production process (Figure 5(c)).



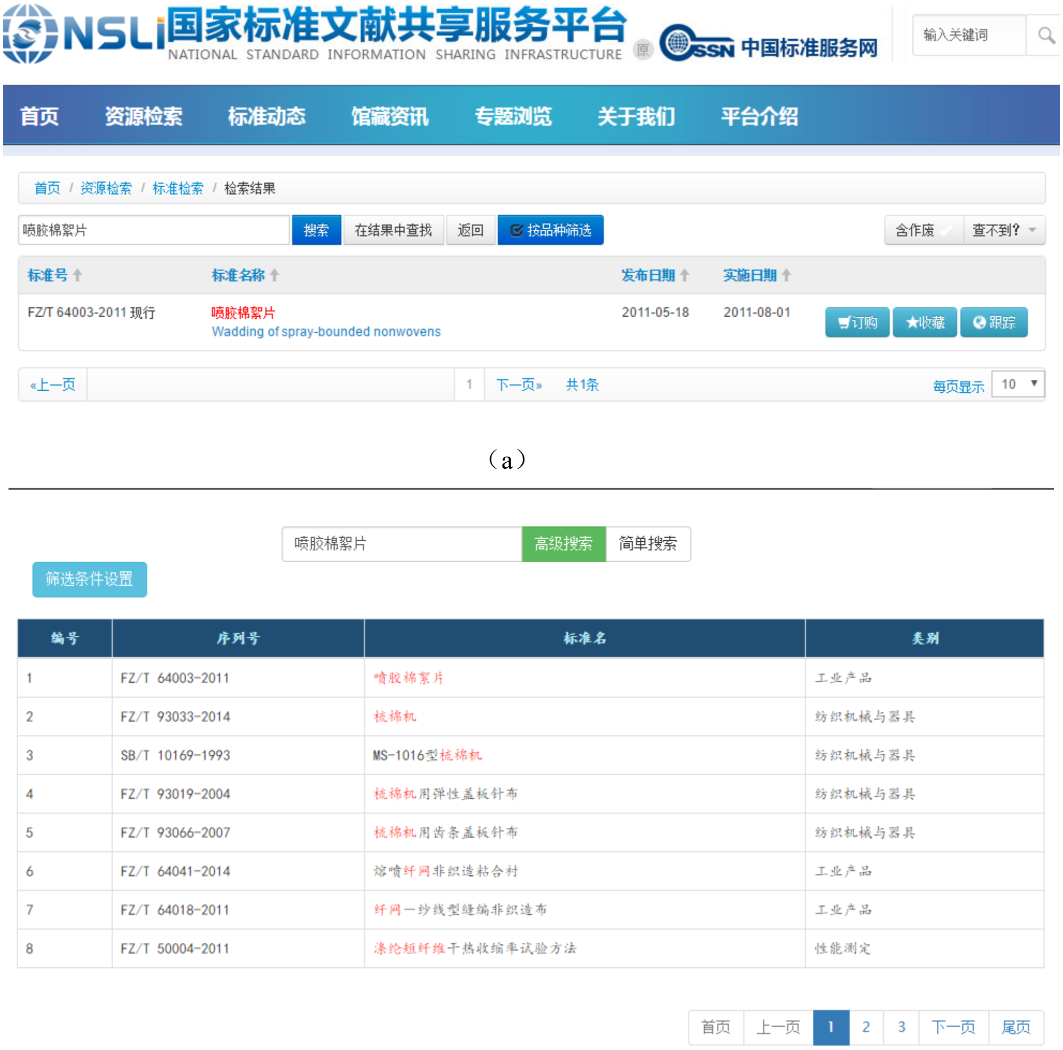
（a）

（b）

(c)

Figure 5 Knowledge map service

- ***Knowledge-oriented standard retrieval.*** By using query expansion based on standardized-concepts relation network, we provide a knowledge-oriented standard retrieval service. Unlike the former standard retrieval system (National Standard Information Sharing Infrastructure, Figure 6(a)), our system has the ability to capture the knowledge connectedness, and returns relevant standard documents in knowledge level. Figure 6(b) shows an example.



喷胶棉絮片

标准号	标准名称	发布日期	实施日期
FZ/T 64003-2011 现行	喷胶棉絮片 Wadding of spray-bounded nonwovens	2011-05-18	2011-08-01

（a）

喷胶棉絮片

编号	序列号	标准名	类别
1	FZ/T 64003-2011	喷胶棉絮片	工业产品
2	FZ/T 93033-2014	梳棉机	纺织机械与器具
3	SB/T 10169-1993	MS-1016型梳棉机	纺织机械与器具
4	FZ/T 93019-2004	梳棉机用弹性盖板针布	纺织机械与器具
5	FZ/T 93066-2007	梳棉机用齿条盖板针布	纺织机械与器具
6	FZ/T 64041-2014	熔喷纤网非织造粘合衬	工业产品
7	FZ/T 64018-2011	纤网一纱线型缝编非织造布	工业产品
8	FZ/T 50004-2011	涤纶短纤维干热收缩率试验方法	性能测定

（b）

Figure 6 Knowledge-oriented standard retrieval service

Knowledge-oriented standard information service is the way forward. However, there is little research about this topic recently. The "deficiency of semantic data" and the "excessive cost of ontology construction" are impediments to the development of knowledge-oriented standard information service. In this paper, we proposed a solution containing two critical steps "semantic re-structuralization of standard literature" and "standardized-concepts relation network construction". We also built an information service system, and attempted to provide knowledge-oriented standard information service such as knowledge map and knowledge-oriented standard retrieval. Certainly, it is just an elementary attempt in the field. The research of knowledge structure of standard literature, knowledge organization and knowledge service form need to go further.

References

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., & Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, *18*(1), 14–21. http://doi.org/10.1109/MIS.2003.1179189

Berners-Lee, T., & Hendler, J. (2001). Publishing on the semantic web. *Nature*, *410*(7382), 409. http://doi.org/10.1038/481409a

Deng, Y. (2008). The Retrieval and Applications of Scientific and Technical Report, Patent Document and Standard Literature Resources. *Library Work and Study*, *7*, 71–74.

Ghoula, N., Khelif, K., & Dieng-Kuntz, R. (2007). Supporting patent mining by using ontology-based semantic annotations. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007* (pp. 435–438). http://doi.org/10.1109/WI.2007.98

Guo, D. (2011). Study on Knowledge Link Service Mode of Standard Literatures. *Library and Information Service*, *55*(9), 76–79.

Heeringa, W. (2004). Measuring dialect pronunciation differences using Levenshtein distance. *Dissertations.Ub.Rug.Nl*. Retrieved from http://dissertations.ub.rug.nl/FILES/faculties/arts/2004/w.j.heeringa/titlecon.pdf

Li, W., Wang, T., & Gu, Y. (2010). The Retrieval and Applications of Scientific and Technical Report, Patent Document and Standard Literature Resources. *Information Research*, *12*, 74–77.

Li, Z., & Zhang, L. (2012). Research of Books Retrieval System Under Thinking of Hybrid System. *New Technology of Library and Information Service*, *21*, 54–58.

Liu, J., & Zhong, Y. (2011). Comparison and Enlightenment of the Retrieval Systems of International Standard Document. *Researches in Library Science*, *20*, 60–64.

Luo, L., Yu, X., Zheng, W., & Jin, Y. (2012). Comparative Study on Patent Retrieval Websites. *Journal of Intelligence*, *03*, 163–167.

Ma, F. (2013). Historical Review of the Development of Information Science with Proposing Frontier Topics. *Document,Informaiton & Knowledge*, *2*, 4–12.

Ma, F., & Gao, J. (2010). Research on Web2. 0 information half life measurement and its impact factors —— Taking social bookmark website as an example. *Journal of Information Studies: Theory and Application*, *11*, 1–6.

Müller, H. M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology*, *2*(11). http://doi.org/10.1371/journal.pbio.0020309

Zheng, Y., & Hua, B. (2011). Discussion on Transforming Relationship of Data, Information, Knowledge and Intelligence. *Journal of Information Studies: Theory and Application*, *34*(7), 1–4.