Study on the Keywords Indexing System Based on Linked Triple Technology for Chinese Scientific and Technical Literature*

Li Hui[1, a], Jin Xiaohong[1,b], Liu Yanjun[1, c], Zhang Yunliang[2, d]

[1] Beijing Institute of Science and Technology Information, Beijing 100048, China;

[2] The Institute of Scientific and Technical Information of China, Beijing 100038, China

[a]lisa-lh@126.com, [b]jinxh@bjstinfo.com.cn, [c]liuyj@bjstinfo.com.cn, [d]88642191@qq.com

Keywords: Linked triples; web services; text annotation; keywords indexing

Abstract: For Chinese scientific and technical literature, keywords indexing and relevance construction is one of the most important links in literature processing. Based on the ontologies that registered in the ontology management platform and the use of web services techniques, the triples of concepts and relation from the knowledge base are assigned for the keywords recognition and indexing for scientific and technical literature. With domain dictionary the system can recognize the domain terminology and with word segmentation and grouping technology to recognize the unlisted words. Build links between the triples from ontologies and the indexing information of the keywords and form conceptual relation networks that deep inside the literature. The indexing results can support the deep analysis and mining work of literature texts. According to the corpus test, the system can carry out literature indexing and words linking at a relatively high speed of 86 pieces/s, with recall of 65%, precision of 69% and comprehensive performance of about 66%.

Introduction

Keywords indexing work is an important link in literature processing. For traditional keywords indexing, words reflecting the subject content are usually selected from the thesaurus for artificial words indexing, by which the subject content of literature can be accurately reflected. However, with refinement of disciplines and deepening of researches, there has been increasingly emerging scientific and technical literature, which makes it more and more difficult to acquire the required information from literature. Some overseas institutes, organizations and universities have carried out extensive research work to develop useful annotation tools, such as Magpie [1] developed by Open University in UK, CREAM [2] developed by the Karlsruhe Institute of Technology in Germany, Annotea [3] researched and developed by W3C and other systems that facilitate artificial indexing, but these systems with low automation degree can only assist indexing personnel to complete the selection of some tags. In face of extensive massive web-based texts, the keywords indexing system cannot meet the application requirements anymore, because it is significantly affected by domain knowledge structures of indexing personnel themselves. In recent years, as the prosperous development of semantic web and data link technology brings new ideas to the study on the literature annotation system, automatic and semi-automatic semantic annotation tools have been developed in succession, such as Semantic Platform [4] system developed by Ontotext Laboratory, SemTag [5] system developed by IBM Company and Armadillo [6] system developed by AKT Project. These systems can complete automatic annotation of domain texts by ontology and machine learning. In addition, some tools, such as Gate [7] developed by the University of Sheffield in UK and Text2Onto developed by the Karlsruhe Institute of Technology, integrating word discovery, relation mining, ontology construction, semantic annotation and various other functions to provide technical measures for English domain ontology construction and knowledge discovery, has been widely used in deep mining and content analysis of English texts. However, these annotation tools cannot meet the requirements for annotation of Chinese literature due to

Fund project: This is one of the research results of Beijing municipal finance projects "Information Processing and Analysis Ability Building Against Text Information " (PXM2016-178214-000006) and "Design and Implementation of Specific Entity Relation Extraction and Data Mining Tools" (PXM2016_178214_000007) in 2016.

differences between Chinese and English in sentence structure and word usage. In order to complete the annotation work of Chinese domain literature and provide resource services for users, this paper tries to implement the indexing system based on the web services technology for Chinese scientific and technical literature, integrate the existing dictionary and knowledge resources into the annotation module, and link various concepts and words describing domain knowledge to triples in the knowledge base by the methods of dictionary matching and word segmentation combination, and form a domain conceptual relation network, so as to provide underlying support for deep analysis and knowledge extraction of literature contents.

Framework of Keywords Indexing and Linking System

The keywords indexing and linking system for scientific and technical literature implemented in this paper mainly consists of two modules, including domain knowledge resource organization and keywords recognition. Keywords indexing and linking in this paper cannot be separated from the support of knowledge bases and domain dictionaries. In order to achieve association and organization of domain knowledge, we have developed an ontology management and service platform integrating ontology management, release, retrieval and service with the purpose of providing a set of standardized management modes for various created ontology resources so that these resources can be used by external users via the uniform service interface [8]. With the service function provided by the management platform, the keywords indexing module acquires the lexical and semantic resources required according to indexing types, so as to achieve keywords indexing and linking in literature. The structure of the whole indexing system is shown in Figure 1 below:

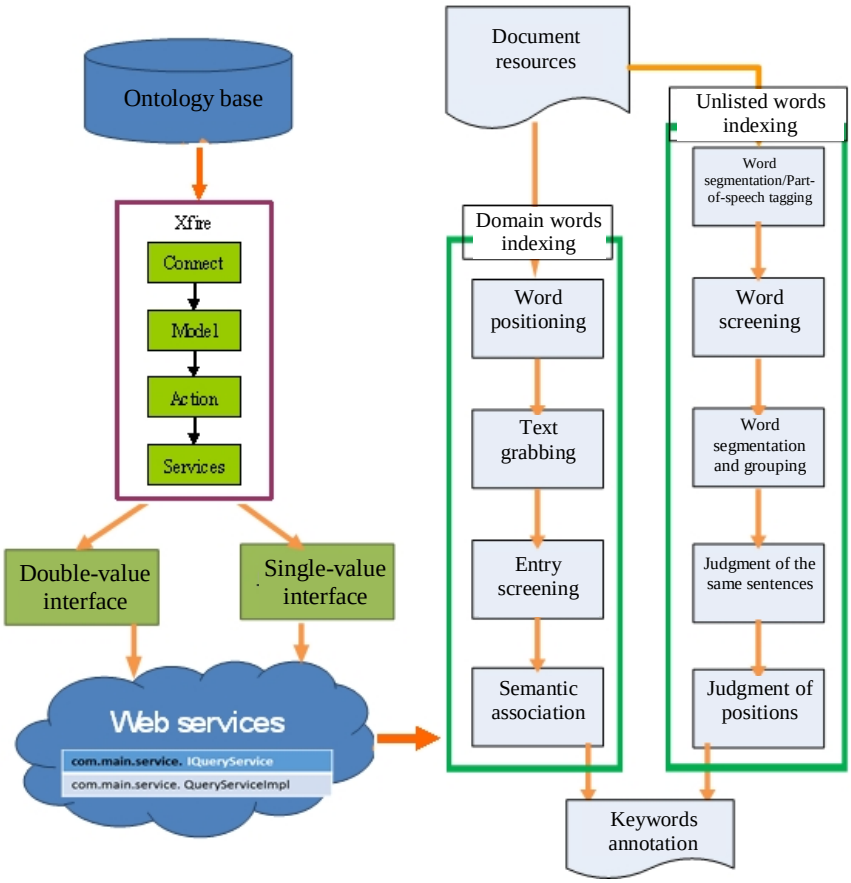

Figure 1 Process of Keywords Indexing and Linking System

There are multiple domain ontologies registered in the ontology management platform, such as

"new energy vehicles", "a new generation of industrial biotechnology", "smart material and structure technology", "major natural disasters monitoring and defense" and "new energy". Each document includes about 50,000 domain concepts. Considering the associated information provided by a single ontology is limited, the management platform creates semantic links for the same or similar concepts in several ontologies by analyzing structure and format information of concepts. In addition to ontology information registration and retrieval functions, the management platform is also designed with the web services interface for convenience of users in need of secondary development and embedded calling. The keywords recognition module achieves unlisted words indexing by means of domain words in the lexical resource identification literature provided by the ontology management platform through the word segmentation and part-of-speech grouping technologies.

In the ontology knowledge base, the domain knowledge framework consists of concepts, attributes, relations and constraints. Though these elements have different semantic features, all of their internal knowledge structures can be transformed into relation triples and attribute triples. The relation triple describes semantic links among various concepts, while the attribute triple describes individual features of concepts. After the knowledge framework is filled with a large number of instances, links are generated among concept instances via triples of "subject - relation - object" and "subject - attribute - value", so as to form an interconnected knowledge network. These associated triples are minimum knowledge units in the ontology knowledge base, serving as basic units for semantic reasoning, relation discovery and similarity computation. Since indexed domain words are sourced from the ontology knowledge base, and concepts in the ontology knowledge base generate various associations via triples, these domain concepts are linked to keywords in scientific and technical literature, so as to form a conceptual network of literature with semantic relations, which provides support for deep mining of subject content in literature.

This ontology management platform integrates ontology management, release, retrieval and service with the purpose of providing a set of standardized management modes for various created ontology resources so that these resources can be used by external users via the uniform service interface [9]. In addition to ontology information registration and retrieval functions, the management platform also provides the web services interface for convenience of users in need of secondary development and embedded calling.

Acquisition of Ontology Triple

For the ontology management platform serving as a service provider, its internal operating methods are released in the form of web services interfaces via the Xfire services proxy. The whole process of service processing involves Action, Model, Connect and Services modules, among which the Action module is responsible for control and forwarding of the retrieval process; the Model module is responsible for encapsulation of background data and generation of retrieval models; the Connect module is responsible for initialization, connection and other operations of the knowledge base; the Services module is responsible for mapping various retrieval conditions input by users to SPARQL expressions, and reassembling retrieval results to form relation models. The Services module consists of 2 internal interfaces and 1 services interface. Considering that users have different query

habits to construct fields with the SPARQL sentence, internal interfaces provided are divided into single-value query interface and double-value query interface. A single-value query interface is encapsulated with 4 functions, including “findAllProperty”, “findAllRelation”, “findAllClass” and “findAllIndividuals”; a double-value query interface is encapsulated with the functions of “findSubjectAndPredicate” and “findPredicateAndObject”. Details of interface service functions are shown in Table 1 below:

Table 1 Functions of Web Services Interfaces

Interface name	**Acceptance value**	**Return value**	**Function**
findAllProperty	String query, String ontology	List<String>	Input SPARQL sentence and ontology file to get all attributes in the ontology
findAllRelation	String query, String ontology	List<String>	Input SPARQL sentence and ontology file to get all relations in the ontology
findAllClass	String query, String ontology	List<String>	Input SPARQL sentence and ontology file to get all conceptual classes in the ontology
findAllIndividuals	String query, String ontology	List<String>	Input SPARQL sentence and ontology file to get all instances in the ontology
findSubjectAndPredicate	String term, String ontology	Map<String, List<String>>	Input conceptual words and ontology file to get subjects and relations in the triple
findPredicateAndObject	String term, String ontology	Map<String, List<String>>	Input conceptual words and ontology file to get relations and objects in the triple

If the front-end interface generates the query sentence in the type of "select ?x where {?x ?p ?y}", the Action module will transfer the retrieval needs to the single-value services interface, and select the corresponding function to call according to the retrieval conditions. The ontology management platform uses sesame local knowledge base to store triples. Prior to retrieval matching, the Connect module completes connection operations of sesame knowledge base. Next, the Model module reads these knowledge triples in the memory to develop a complete conceptual network model. Finally, the retrieval expression is pushed into the model to complete matching of needs. If the front-end

interface generates the expression in form of “select ?x ?p where {?x ?p ?y}” or “ select ?p ?y where{ ?x ?p ?y}”, the Action module will call the double-value services interface, and complete matching of needs with the assistance of Connect, Model and other modules.

During service calling and secondary development, users only need to interact with the external services interface of the ontology management platform. In this paper, various functions provided by the services interface are embedded during the implementation of text annotation and display of lexical relation.

Keywords Recognition and Annotation Methods

The annotation module is intended for extracting concept units that reflect the document theme by means of domain dictionary and term recognition technology, so as to provide support for the modules such as information retrieval, knowledge navigation and knowledge evolution. Text word annotation is divided into two modules, including dictionary matching and recognition unlisted words.

Domain Words Recognition

Domain words recognition cannot be separated from support of external dictionaries. The web services function provided by the ontology management platform can help complete this work. Content description of an ontology file is usually related to a specific domain, of which internal concept nodes and relations construct a domain knowledge framework. There are multiple domain ontologies registered in the ontology management platform, such as "new energy vehicles”, "a new generation of industrial biotechnology", "smart material and structure technology", "major natural disasters monitoring and defense" and "new energy". Each document includes about 50,000 domain concepts. Considering the associated information provided by a single ontology is limited, the management platform creates semantic links for the same or similar concepts in several ontologies by analyzing structure and format information of concepts. After a new ontology file is added, the knowledge base generates the storage file by using the ontology URL as namespace, and meanwhile, stores structural information such as concept classes, inter-class relations and attributes in the domain knowledge base in the form of triple. After the annotation function is activated, the system firstly acquire all conceptual words in the registered ontology with the web services interface, forms a domain dictionaryList, and complete annotation of domain words by means of the domain dictionary. Steps of domain words recognition are given in detail below:

(1) Word positioning: Traverse dictionaryList of the dictionary, look up the starting position of word in text, store its position information in the termList and mark the type of "DomainTerm".

(2) Text grabbing: Take the "word starting position + word length" as the starting point to grab the remaining text as the matching text. Traverse the dictionary again to complete word matching and position computation, and ensure all words in the dictionary have been traversed.

(3) Entry screening: Traverse the termList. If the starting position of any word is within that of another longer word, information of the shorter one in the termList will be removed, while

information of the longer one will be retained.

(4) Semantic association: Access to the ontology management platform via the web services interface, associate relevant concepts in the ontology triple with domain words, and store all annotation information in the database.

(5) Entry marking: Search in the database to acquire the starting position of word, transmit the position information to the foreground page and display it in highlighting.

Unlisted Words Recognition

Since the scope of words listed in the domain dictionary is significantly affected by selection criteria and timeliness, the words listed usually cannot comprehensively cover and reflect new trends of the whole domain. Thus, the unlisted words need to be screened by the keywords recognition algorithm. The keywords recognition module is mainly responsible for operations such as part-of-speech (POS) tagging, sentence segmentation, lexical collocation, word positioning, stop words filtering. Its main operational processes and functions are described in the table below:

Table 2 Unlisted Words Recognition Method

Step	Operation	Function
1	Part-of-speech tagging	The text becomes a word list attached with POS information through word segmentation.
2	Word screening	By means of the stop word list, some common words used a lot in the text and some words forming sentences but without actual meanings can be removed.
3	Word grouping	The word formation rule is an important basis in term recognition. The recognized term can be more meaningful if the rule is formulated according to the lexical collocations.
4	Judgment of the same sentences	The text is segmented to independent sentences according to the segmentation marks of sentences.
5	Judgment of positions	The position where each candidate word occurs in each document is recorded.

POS tagging is the primary task of text processing. After processing, the input text is split into string collections composed of individual words and symbols. Most of word segmentation software adopts the general dictionary, in which the segmentation granularity is relatively small for many words. However, domain words used in scientific and technical literature have strong specificity and large word length. Therefore, after word segmentation, the text needs to be combined according to POS collocation rules, so as to form relatively long keywords with domain characteristics. A large number of candidate words with strong specificity can be extracted from articles by means of word collocation and formation rules. Through analysis, we find that most of domain keywords are formed by collocations of noun, adjective, adverb, preposition, etc. In addition, a few keywords consist of verb and other words. During the test processing, sentences are segmented by some

punctuation marks with special meanings, such as ", . ? ! ", so as to ensure the scope of combination in the word sequence never spans two different sentences. After word segmentation and combination, a word list stored in the form of vector is generated for the text. Each word vector includes information of the word, such as POS, document ID and position. Through preliminary screening, the word list usually includes a part of domain words, and needs to be re-matched with the dictionary. After the domain words are removed, the remaining words are marked as unlisted words in the type of "New Term", and all annotation information is stored in the database.

Testing of text annotation effect

In order to the operating effect of the system, the authors annotated 64,121 domain literature documents of new energy vehicle. The annotation results are marked by different colors in the page; domain words are green and unlisted words are blue. The annotation effect is shown in Figure 2 below.

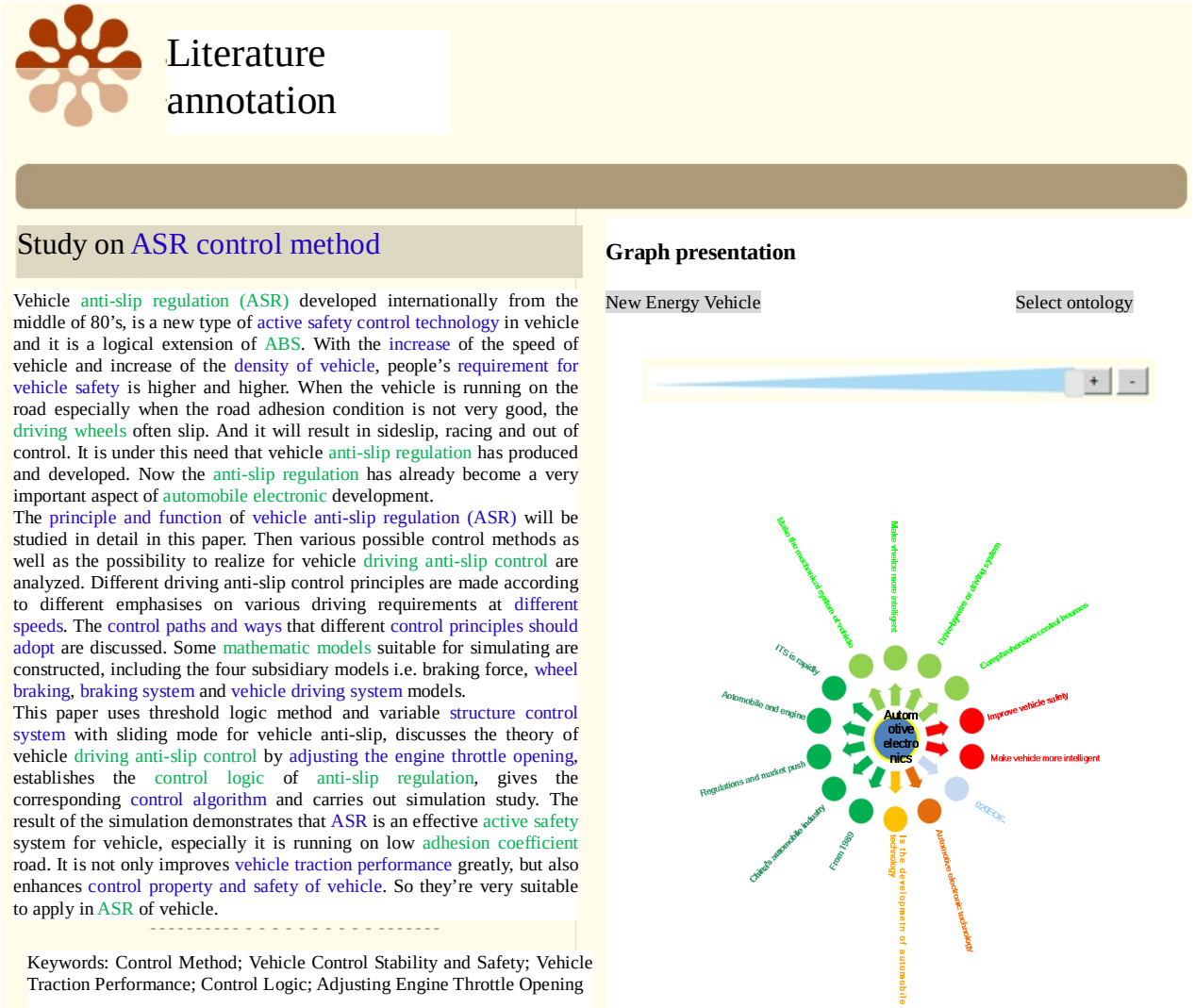

Figure 2 Keywords annotation

Green domain words are related with ontology triples through hyperlink. When users click a domain concept, the system will search it within the registered domain ontologies, and return the targeted ontology files to users. Upon users' confirmation of the domain which the concept belongs

to, the system will send the retrieval demand to the ontology base via web services interface and obtain a semantic triple, and generate a relational model by taking the entered concept as the core. Where there is a clear relation between two concepts, its structure can be represented by the triple. The entered word and relation between words are respectively mapped into the node and relation arc, and the interface will radiate outward from this entered word serving as central element so as to form a relational network. When users click a selected outside node, the processing program will regard the current node as the central node and search various relevant relations, and convert the tree graph into a circular structure through circular layout algorithm. In the circular structure, the root node is the central node, and the relation nodes are distributed on the outside circular ring. The relations of words in the graphic interface are classified by different colors, and the relations between words in the same color belong to the same category. The whole documents are integrated in the server where indexing testing is completed. The basic configurations of the service are: AMD Athlon™ II×2 B26 Processor 3.20GHz CPU, memory 4G, 32-bit Windows7 OS. The operation of indexing system is shown in Table 3 below.

Table 3 Indexing efficiency

	Quantity	**Time**	**Efficiency**
Literature	64,121 pieces	741 seconds	86 pieces/second
Domain words	78,727 entries	493 seconds	159 entries/second
Unlisted words	500,042 entries	248 seconds	2016 entries/second

The average processing speed of the annotation system is 86 pieces/second. It is found by monitoring the program that the matching computation is time-consuming in the whole indexing process. The system acquires more than 20,000 domain words from the ontology management platform. In the indexing process, 64,121 domain literature documents are required to match with each word in the dictionary in order and the location of each word is computed, leading to the result that the time is long for annotation of some literature documents.

In addition, the annotation module provides a subject word recommendation function, which gives different weights to the annotated words in accordance with the location of the word in the document. The words in the headline are provided with high subject word weight, and words in the text are comprehensively computed by the location importance, whether they come from the thesaurus, or other indicators. All words serve as candidate words after being ordered by weight values, and return to the page according to the screening quantity provided by users. In order to test the annotation performance of the system, the authors chose 50 literature documents from the corpus, carried out machine annotation and manual annotation respectively, and analyzed the annotation effect by the fullness rate, accuracy rate and F harmonic mean. The details are shown in Table 4.

Table 4 Annotation performance

Literature (piece)	Manually-annotated words	Machine-annotated words	Accurate words for machine annotation	Fullness rate - (machine-annotated words/manually-annotated words)*100%	Accuracy rate – (accurate words for machine annotation/machine-annotated words)*100%	F harmonic mean - (2*fullness rate * accuracy rate)/(fullness rate + accuracy rate)*100%
50	381	250	174	65%	69%	66%

250 keywords were screened out by the annotation system, and 381 keywords were screened out in a manual way on the basis of processing the keywords in 50 literature documents. Taking the manually-annotated words as a reference, contrastive analysis shows that the annotation system has a fullness rate of 65% and an accuracy rate of 69%, resulting in a comprehensive performance of about 66%.

Conclusion

Keywords indexing and relevance construction is one of the most important links in literature processing, and the basis for the deep analysis and mining of text contents. Based on the services techniques of the ontology management platform, the automatic indexing of domain words and unlisted words in the scientific and technical literature is realized in this paper. Use of web services techniques brings convenience to the system development. By accessing the services interface, the indexing system can not only use the dictionary resources from the ontology management platform, but also establish a semantic link with the triples of the ontology knowledge base. Various relations and attribute descriptions of keywords are shown in this paper, which facilitates users to know the knowledge structure of the whole domain.

The sources of literature in the corpus are diversified. A large number of literature documents are converted from pdf and html files, with a certain number of incomplete sentences, spaces, line segmentations and special symbols existing in the texts. Those noisy data have an impact on the word segmentation and part of speed judgment. Later, we consider constructing index for the domain dictionary and triple list that are loaded into the memory, and use more powerful pdf and web text extraction tool to complete the corpus cleaning. The accuracy rate and operating speed of the system can be increased by improving the matching efficiency of dictionary and triples as well as the text extraction quality. Meanwhile, we shall also note that it is difficult to provide an absolute word screening standard for analysis and statistics of machine annotation effect due to the influence of the individual difference in subjective cognizance and knowledge structure. In addition, the unordered discrete arrangement of domain words in the word level cannot represent the purpose of the literature although the scientific and technical literature is composed of a large number of domain words. In the future work, we shall organize the concepts in the sentence level, and form semantic chunks by semantic combination between concepts, so as to grasp the literature theme from the more macro level.

References:

[1] John Domingue, Martin Dzbor, Enrico Motta. Magpie: Browsing and Navigating on the Semantic Web [C].In: Proceeding of the Conference on Intelligent User Interfaces, Portugal, January 2004

[2] Siegfried Handschuh, Steffen Staab. Authoring and Annotation of Web Pages in CREAM [C]. In: Proceeding of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA, 2002

[3] Annotea [EB/OL].[2014-10-13].http://www.w3.org/2001/Annotea/

[4] Ontotext Semantic Platform [EB/OL]. [2014-10-13]. http://www.ontotext.com/products/ontotext-semantic-platform

[5] Stephen Dill, Nadav Eiron, David Gibson, Semtag and seeker Bootstrapping the semantic web via automated semantic annotation[C].In: Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 2003: 178-186

[6] Armadillo [EB/OL]. [2014-10-13]. http://www.hrionline.ac.uk/armadillo/links.html

[7] Gate [EB/OL]. [2014-10-13] https://gate.ac.uk/overview.html

[8] Xu Deshan, Zhang Yunliang. Design and Implementation of Integrated Ontology Management Platform [J]. Digital Library Forum, 2013, 11 (114): 15-20