International Journal of u- and e- Service, Science and Technology
Vol.8, No.5 (2015), pp.241-252
http://dx.doi.org/10.14257/ijunesst.2015.8.5.22

Design and Implementation of Retrieval System of Network Learning Resources Based on Semantic Web

Haibin Hu[1] and Shipin Chen[2]

[1]*Experiment Center, China West Normal University, Nanchong Sichuan 637002, China*
[2]*School of Education, China West Normal University, Nanchong Sichuan 637002, China*

Abstract

The gradual finding of large quantities of low-relevant or even non-relevant information is returned by the traditional keyword-based query. To further improve the efficiency and quality of network learning resources retrieval, this article designs a generic Semantic Web-based e-learning resources retrieval system model and achieves B/S mode of semantic retrieval prototype system, considering the influence of various factors for the similarity and correlation, such as degree of semantic overlap, semantic distance, semantic hierarchy, and property relation. By comparing with the traditional query, the retrieval results showed that the semantic retrieval methods to ensure precision in the case of a higher recall rate.

Keywords: *Semantic Web, Ontology, Semantic Expansion, Similarity, Correlation*

1. Introduction

With the development of internet technology, the amount of information increases exponentially. Facing the enormous amount of information, people become increasingly dependent on search engine. According to the latest three statistics reports (the report is accomplished every half year) of CNNIC (China Internet Network Information Center) between December 2010 and January 2012, over 75% of all users of the Internet choose search engines as the main way of searching information resources. Moreover, people find that as the amount of information increases, it is increasingly more difficult to obtain their target knowledge. Facing the enormous amount of information, a new problem is arising. That is, the scale of amount of information unprecedentedly expands. On the other hand, information is too much to be organized. It is difficult to find the information that people really want. The phenomenon [1] of 'overloaded information and deficient knowledge' occurs. It is urgent to solve the problem of how to find the required information and knowledge accurately and efficiently from the overwhelming information. Semantic Web technology provides a new way for solving the problem.

In the study, traditional keyword-based information retrieval is transferred to the semantic expansion-based search with the support of semantic Web technology. Moreover, the retrieval efficiency and quality of network learning resources are promoted through adjusting similarity and correlation parameters and controlling the threshold values.

2. Related Theories

2.1. Semantic Web

Semantic Web is an internet that contains documents or segments of documents. It describes an obvious relationship between things, and contains semantic information,

ISSN: 2005-4246 IJUNESST

which is in favor of the automatic processing of machines [2]. The goal of semantic Web is to make the information on Web accessible to computers, and then the automatic processing of Web information can be realized to adapt to the rapid growth of Web information resources.

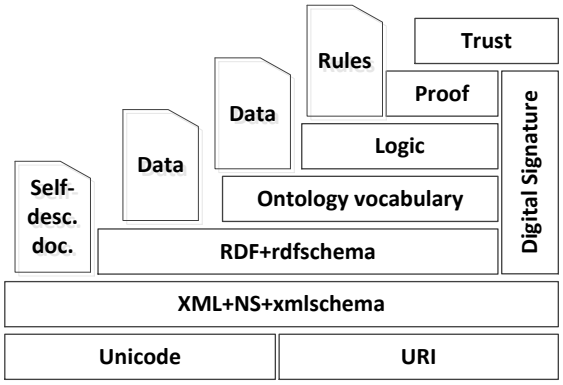

Figure 1. Structure of Semantic Web System

The technological bases of semantic Web are XML, RDF and Ontology. The basic method of implementation is to develop the formal specification language for information with a gradually increased function to determine the unique meaning of information. In 2000, Tim Berners-Lee proposed the seven-layer structure of semantic Web system, as shown in Figure 1. In the structure, the function gradually increases from bottom to top. The function includes character encoding and marking, expression of data contents and structure, description of resources, as well as the abstract description, argumentation and inference of concepts and the relationship between them.

2.2. Extensive Markup Language (XML)

XML, like HTML, is a type of markup language, with its standard recommended by the world wide web consortium (W3C). The difference between them is that tags in HTML are predefined with limited number, while no predefined tags exist in XML. Users can define their own tags and document structure. For example, '<name> </name>' is used to represent someone's name. Hence XML has good expandability. Moreover, HTML is used to define data and focused on the appearance effect of data presentation, while XML is used to store data and focused on describing information structure and data.

Since XML has the characteristics of supporting the separation between contents and presentation, convenient data exchange and sharing, strong cross-platform and expandable tags, many other languages have been defined with XML as the meta-language, such as xHTML,WSDL,RDF,DAML and OWL (Ontology Language).

2.3. Resource Description Framework(RDF)

RDF is usually used to express various resource information. It is a standard that is recommended by W3C based on XML. Three basic concepts of resource, attribute and statement are defined in RDF. The RDF provides a simple data type, which is made up of connecting arcs with tags between nodes. Nodes are used to represent the resources on Web, and arcs are used to represent the attribute of resources. Objects (or resources) and the relationship between them can be described with the data model conveniently without the limitation of data type. The data model can be used to describe any type of information. Therefore, the data model of RDF can serve as the basic model of any other complicated relation model. This is because essentially, the data model of RDF is a type

of binary relation, and any complicated relations can be separated into multiple simple binary relations.

2.4. Ontology

In philosophy, ontology is defined as 'systematically describing any objective existence'. The concept is a systematic explanation and illustration of objective existence, and is focused on the abstract essence of objective reality [3]. As the artificial intelligence develops, the concept of ontology has expanded to artificial intelligence, information science and library science from philosophy, and has attracted great attention from specialists and scholars [4]. In the field of artificial intelligence, Neches et al. firstly proposed the definition of ontology. That is, basic terms involved in the relevant field and their relation are provided. Then the rule of the vocabulary expansion is defined with the standard determined by using the basic concepts and their relation. [3] Neches considered that ontology defines the basic terms of subject domain and their relation, and then the rule of vocabulary expansion was defined by combining the terms to their relation. [5] Gruber defined the ontology as ' a clear specification on a conceptual model' [6]. Another definition was given by Borst as 'ontology is a clear specification on a shared conceptual model' [7]. This definition is the most widely used.

From the perspective of content analysis, ontology has a high ability in expressing knowledge, and plays an important role in knowledge sharing and knowledge reuse on the semantic level required for realizing semantic Web. The ontology extracts uniform concepts (terms) and their relation from complicated domain knowledge, making the knowledge sharing available. Formal language is used to describe the ontology. Construction of ontology is independent on tasks. The ontology can be reused in different application systems to avoid repetitive analysis of domain knowledge.

The ontology provides a common understanding and description of domain knowledge. The certainty of concept description guarantees the effectiveness of inference from data level, and supports the logical inference that ensures the computational completeness and decidability, providing a base for solving the logical inference and verification on semantic level. Moreover, the ontology plays a role as bridge between various types of application system of computers and knowledge that we have. It is the key to realize the semantic Web.

3. Design of System Model

The design framework of retrieval model of semantic system is shown in Figure 2. The framework is made up of two parts. The first part is to extract resources from the learning websites involving education technology, and to establish a local library of network learning resources with retrieval indexes for users. The second part is to process the user's search. The first involved step is to obtain user's search request. Then the keywords input by users are processed by using the ontology library of education technology. The processing results will be fed back to the users.

3.1. Achievement of Network Learning Resources

In this study, electronic resources including electronic teaching plan, self-edited handout, teaching courseware, reference material, expert's lecture, exercise books and examination database of the excellent courses such as 'Introduction to Education Technology' and 'Education Technology' are obtained by using the web crawler tool. The document formats contain PPT,DOC,TXT,PDF,HTML and MHT.

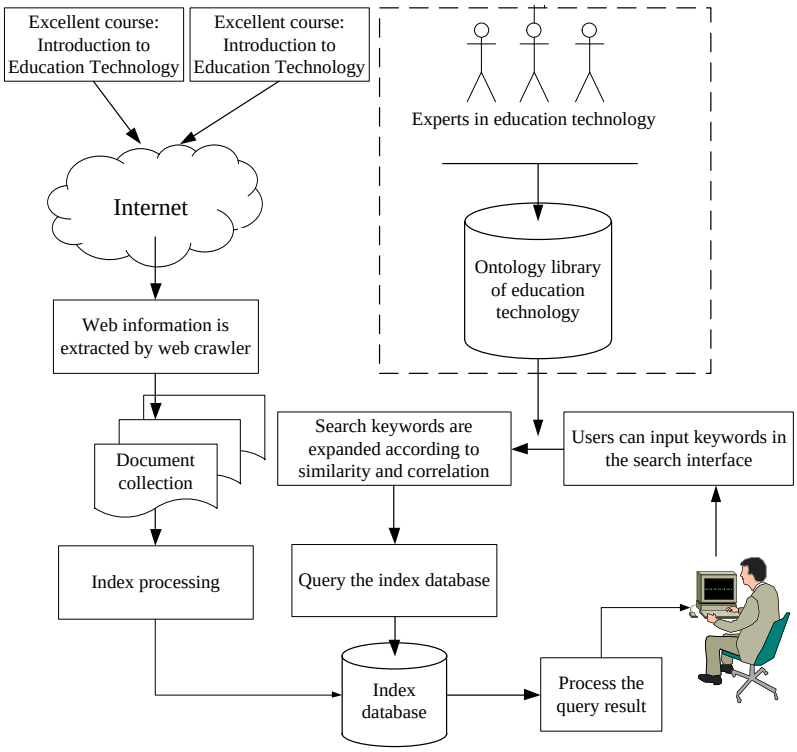

Figure 2. Structure Model of Semantic Retrieval System

3.2. Establishment of Index for Resource Libraries

Currently, full-text index is supported by Lucene which is supported and provided by the Apache Software Foundation. Lucene supports an SDK of full-text index developed by Java, and provides powerful APIs such as search engine, indexing engine and textual analysis engine.

The key technologies involved in the establishment of index for resource libraries are Chinese character segmentation technology, document transformation technology and the use of SDK of Lucene.

3.3. Semantic Expansion of Search Keywords

For the keywords input by users, search terms that are related to the original semanteme are obtained through inference engine under the support of ontology. According to comprehensive weights (threshold of semantic expansion) of similarity and correlation, the keywords are added to the original search after the expanded vocabulary with small similarity and correlation. Then a set of search terms with high similarity and correlation are constituted. The words in the set of search terms are ordered through the comprehensive weights to provide a base for the next search. The flow of semantic expansion of search keywords is shown in Figure 3.

The key technologies involved in the semantic expansion in addressing the search keywords contain reading, analysis and inference of ontology proposed by Jena [8], search and computation of similarity and correlation of ontology with the ontology query language Sparql [9].

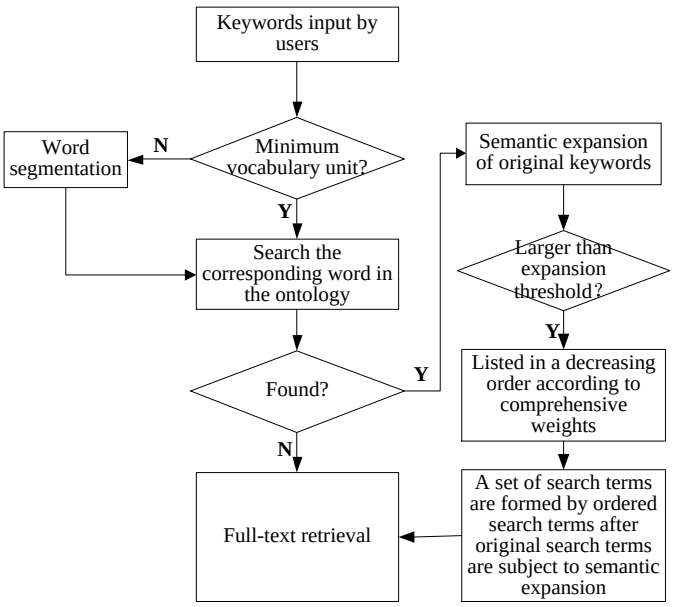

Figure 3. Flow Diagram of Semantic Expansion of Keywords

4. Similarity and Correlation

The semantic similarity of ontology concept contains three key words: 'ontology', 'semantic' and 'similarity'. 'ontology concept' indicates the object of similarity measurement; 'Semantic' provides a basis of similarity measurement; 'Similarity' means a psychological reaction [10] generated with the concept. Liu et al. defined the similarity of concept by the dependences between concepts. The correlation of two concepts reflects the degree of correlation between them, which can be measured by the possibility of co-occurrence of two concepts in the same context [11]. Lin proposed a generalized definition of similarity based on information science and four intuitions of similarity required for the definition of similarity [12].

The concept of similarity is different from that of the correlation. For example, the similarity between two concepts, 'Dale' and 'cone of experience', is very low in education technology, but the correlation between them is very high. In literature [13], the difference between the two concepts has also been clearly described by using the example of vehicle, gasoline and bicycle, suggesting that similarity is not equivalent to correlation. The premise of the existence of similarity is if there are any common characteristics.

The similarity is closely related to correlation. If the two concepts are very similar, the correlation degree between the concepts will also be very high. That is, similar concepts are usually correlated. On the contrary, if two concepts are correlated, they are not necessarily similar.

At present, influence factors of concept similarity are being gradually refined by researchers. Factors such as the degree of semantic overlap, semantic distance and structure of ontology (depth, width and density of concept) are being considered comprehensively for calculation, and to form a reasonable computation method of similarity. In this study, comprehensive weights of similarity and correlation between concepts of ontology are calculated. The influence on the similarity and correlation brought by multiple factors such as degree of semantic overlap, semantic distance, semantic hierarchy and attribute relationship are considered comprehensively.

4.1. Computation of Similarity

On the premise of no multiple inheritance, ontology can be abstracted as a hierarchical tree with superordinate relationship and subordinate relationship, as shown in Figure 4. Concepts are shown by the nodes, and parent-child relationships are shown by the links.

(1) Degree of semantic overlap

The number of semantic overlap represents the number of the same superordinate concept contained between the concepts in the ontology. The degree of semantic overlap indicates the ratio of the number of the same superordinate concept to the number of overall nodes. It represents the degree of similarity between two concepts. In practical applications, the ratio can be transferred to the number of common nodes and that of the overall nodes [14, 15].

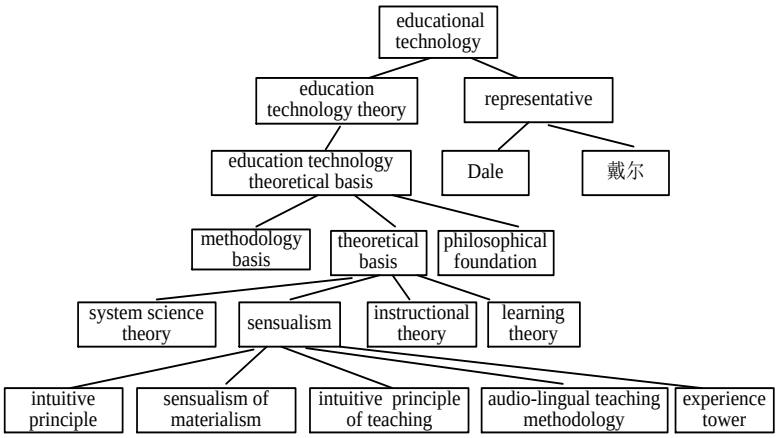

Figure 4. (Partial) Structure of the Ontology of Education Technology

Suppose that the degree of semantic overlap between concept 1 and concept 2 in the same ontology, which is denoted by *Overlapratio(Concept1, Concept2)*, is calculated. $Count(Concept1 \cap Concept2)$ represents the number of common nodes of the paths from concept 1 to the root node and from concept 2 to the root node. $Count(Concept1) \cup Count(Concept2)$ represents the number of overall nodes of the paths from concept 1 to the root node and from concept 2 to the root node. The calculation formula of degree of semantic overlap is shown as follows:

$$Overlapratio(Concept1, Concept2) = \frac{Count(Concept1 \cap Concept2) \times b}{Count(Concept1) \cup Count(Concept2)}$$

where b is an adjustable parameter for controlling the influence on similarity brought by the degree of semantic overlap. The value falls in the range of $1 \leq b \leq$ MaxDepth/(MaxDepth-1), where MaxDepth is the maximum depth of the tree.

(2) Semantic Distance

Semantic distance is the length of the shortest path between two concepts (nodes) of the tree. The larger the distance between two concepts, the lower the similarity will be [11]. The semantic distance is denoted by *SemDistance(Concept1, Concept2)*. The distance between two nodes is denoted by *Distance(Concept1, Concept2)*. The formula of semantic distance is shown as follows:

$$SemDistance(Concept1, Concept2) = \frac{a}{Distance(Concept1, Concept2) + a}$$

where a is an adjustable parameter for controlling the influence on similarity brought by the value of semantic distance. It is a positive real number.

(3) Semantic Hierarchy

In the hierarchical tree, the classification of concepts becomes finer from the root of tree to the leaves. The deeper the hierarchy of concept is, the finer the classification will be. The similarity between the concepts which are farther from the root is larger than that close to the root. Moreover, the similarity between the concepts in the same hierarchy is larger than that in different hierarchies. For two concepts with the same semantic distance, concept similarity increases as the sum of the hierarchy where they locate increases, and decreases as the difference between the hierarchies where they locate increases [16]. That is to say, the larger the difference between the two hierarchies where the two concepts locate, the smaller the similarity between them is [14]. Suppose that Concept1 and Concept2 are any two nodes in the ontology tree. Depth (Concept1) represents the depth of Concept1 in the ontology tree. The value of Depth (Root) is assumed as 1, then the formula of semantic hierarchy is shown as follows:

$$SemDepth(Concept1, Concept2) = \frac{c}{\left|Depth(Concept1) - Depth(Concept2)\right| + c}$$

where c is an adjustable parameter for controlling the effect on similarity brought by the difference between semantic hierarchies. It is a positive real number larger than 1.

The computation formula of semantic similarity of any two concepts (Concept1 and Concept2) in the ontology is obtained by synthesizing the impact on similarity brought by the degree of semantic overlap, semantic distance and semantic depth. It is shown as follows:

$$Sim(Concept1, Concept2) = Overlapratio(Concept1, Concept2) \times SemDistance(Concept1, Concept2) \times SemDepth(Concept1, Concept2)$$

4.2 Computation of Correlation

In the computation of semantic similarity of concepts in ontology, only the hierarchy of concepts (superordinate relationship and subordinate relationship) is considered, while the attribute relationship is ignored. If the hierarchy is removed, a structure (as shown in Figure 5) is constituted according to the attribute relationship. The similarity is defined by the distance between attributes. The larger the distance between attributes, the larger the semantic similarity is. If this distance is 0 and similarity is 1, two concepts are equivalent, namely, different concept names refer to the same meaning. Specific formula is shown as follows:

$$Correlation(Concept1, Concep2) = \frac{d}{PropertyDistance(Concept1, Concept2) + d}$$

where d is an adjustable parameter for controlling the weight between similarity and comprehensive factors of similarity and correlation. It is a positive real number.

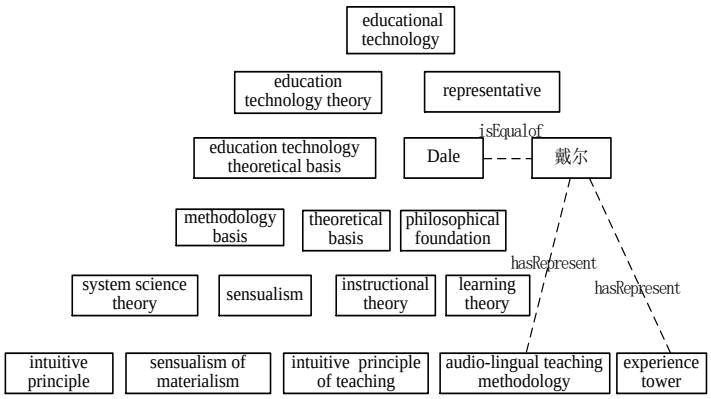

Figure 5. (Partial) Attribute Relationship of Education Technology

The weight of any two concepts in the ontology is obtained by synthesizing the semantic similarity and correlation. The formula is shown as follows:

$$TotalWeight\ (Concept1,Concept2) = Sim(Concept1,Concept2) + Correlation\ (Concept1,Concept2)$$

5. System Implementation and Result Analysis

5.1 System's Developing Platform and Tool

The B/S mode is used for implementing the system. The developing language is JAVA, and database is MySQL Server. The main toolkits include Tika 1.2,Protégé 4.2,JDK1.7.0_01,Jena2.5.6,Lucene3.6.1 and mmseg4j 1.8.

5.2 Homepage of Retrieval

In order to create a friendly operation interface for users, word segmentation methods such as intelligent word segmentation and artificial word segmentation with whitespace are provided on the interface. Operations such as the retrieval range, control of threshold of semantic expansion and setting of number of returned results are provided. For the convenience of research, semantic retrieval and traditional retrieval are designed. Specific interface is shown in Figure 6.

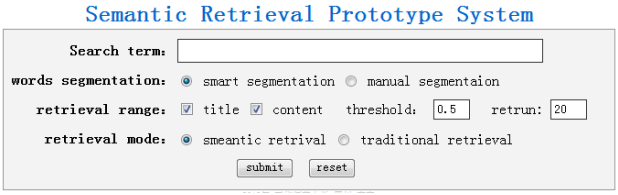

Figure 6. Interface of Semantic Retrieval System

5.3 Implementation of Computation of Semantic Similarity and Correlation

There are a number of sets of keywords after the keywords input by users are subject to semantic expansion. The comprehensive weight is computed through the semantic similarity and correlation. Precision and recall ratio can be promoted by adjusting the threshold.

Before the semantic similarity and correlation are computed, adjustable parameters of α, β, γ and δ should be set up. The operation is completed by domain experts. They are used for adjusting the semantic distance (α), degree of semantic overlap (β), semantic hierarchy (γ) and the weight of correlation (δ) in the computation of similarity, respectively. They are adjusted according to practical condition of applications. The pannel of adjustable parameters is showed in Figure 7.

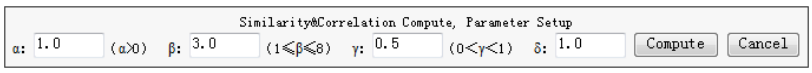

Figure 7. The Pannel of Adjustable Parameters

The computation results of similarity, correlation and comprehensive weight between concepts are shown in Table 1.

Table 1. Similarity, Correlation and Comprehensive Weight

Concept 1	Concept 2	similarity	correlation	comprehensive weight
Multipurpose board	Visual aided board	0.75	1.0	1.0
Computer	计算机	0.714286	1.0	1.0
CD	laser disc	0.714286	1.0	1.0
video disc	laser disc	0.714286	1.0	1.0
laser disc	VCD	0.714286	1.0	1.0
培根	Bacon	0.5	1.0	1.0
戴尔	Dale	0.5	1.0	1.0
杜威	Dewey	0.5	1.0	1.0
Connectionist	stimulus-response	0.75	1.0	1.0
Picture	Photo	0.875	0.0	0.875
Graphic material	cartogram	0.875	0.0	0.875

5.4 Retrieval Results and Feedbacks

Figure 8. Result Page of Semantic Retrieval

According to the options such as keyword information and word segmentation method, retrieval range and retrieval method submitted by users on homepage, search results are finally presented to users in browsers with HTML after the submissions are processed through the application programs on servers. The page of retrieval results contains result of word segmentation, retrieval method, a set of semantically extended keywords aligned sequentially and highlighted query results. See Figure 8.

5.5 Test of Experimental System and Result Analysis

(1) Evaluation index of performance

The recall ratio T_{recall},precision ratio $T_{precision}$ and F1-Measure F_1 are the concepts in information retrieval domain. T_{recall} refers to proportion of retrieved related documents I_{rs} out of all related documents I_{sum} in system. $T_{precision}$ is defined as proportion of retrieved related documents I_{rs} relative to all retrieved documents I_{ss}. They are the important indexes reflecting the retrieval effect.

$$T_{recall} = \frac{retrieved\ related\ \ inf ormation\ I_{rs}}{retrieved\ \ inf ormation\ I_{ss}}$$

$$T_{\Pr ecision} = \frac{retrieved\ related\ \ \inf ormation\ I_{rs}}{all\ related\ \ \inf ormation\ I_{sum}\ in\ system}$$

$$F_1 = \frac{2 \times T_{recall} \times T_{precision}}{T_{recall} + T_{precision}}$$

(2) Analysis of retrieval results

Seven representative words, i.e. 'Dale', 'experience tower', ' audio-lingual teaching methodology', 'communication theory', 'behaviorism', 'learning theory' and 'system theory' are extracted from the course of 'Introduction to Education Technology'. They are represented respectively by {*W1*, *W2*, *W3*, *W4*, *W5*, *W6*,*W7*}. The recall ratio and precision ratio are compared by using traditional and semantic retrieval methods. Parameters involving the semantic similarity and correlation are set up as α= 1.0, β= 3.0, γ= 0.5 and δ= 1.0. The threshold of semantic expansion is 0.5. The retrieval result is shown in Table 2. We can calculate mean search efficiency shown in the following Table 3.

Table 2. Comparison between Results of Two Retrieval Methods

keyword	retrieval method	T_{recall}	$T_{precision}$
W1	traditional	18/32=0.56	18/23=0.78
	semantic	22/32=0.68	22/29=0.75
W2	traditional	19/32=0.60	19/26=0.74
	semantic	19/32=0.61	19/29=0.67
W3	traditional	4/31=0.12	4/5=0.73
	semantic	24/31=0.77	24/29=0.82
W4	traditional	43/133=0.33	43/55=0.79
	semantic	92/133=0.69	92/124=0.74
W5	traditional	24/132=0.18	24/35=0.69
	semantic	85/132=0.64	85/116=0.73
W6	traditional	62/143=0.44	62/96=0.65
	semantic	83/143=0.58	83/127=0.65
W7	traditional	13/38=0.35	13/19=0.69
	semantic	22/38=0.58	22/33=0.67

Table 3. Comparison of Mean Efficiency of Two Search Ways

	Traditional	Semantic
Mean T_{recall}	36.86%	65.00%
Mean $T_{precision}$	72.43%	71.86%
F_1	48.86%	68.26%

The results by two retrieval methods are showed with a histogram. The comparison of recall ratios and precision are shown in Figure 9 and Figure 10, respectively. The retrieval result shows that good effect has been obtained by using the semantic retrieval implemented in the system. On the premise of high precision, the recall ratio obtained by the semantic search is obviously superior to that obtained by traditional retrieval.

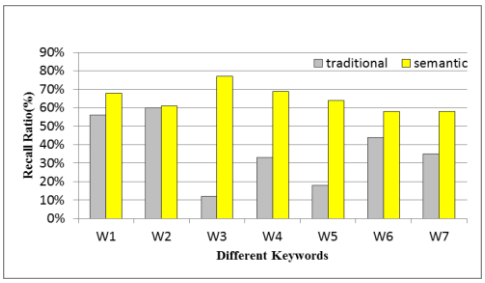

Figure 9. Comparison of Recall Ratio

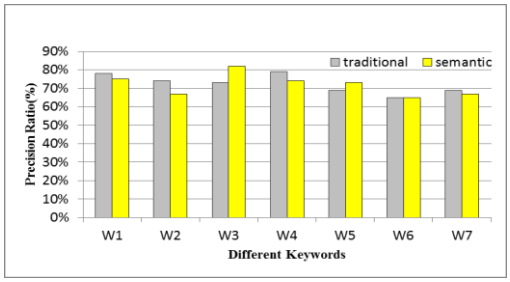

Figure 10. Comparison of Precision Ratio

6. Conclusion

The retrieval of network learning resources in the curriculum of 'Introduction to Education Technology' is taken as an example. The retrieval model of network learning resource system based on semantic Web is designed through establishing the domain ontology of education technology. Function modules such as full-text index, semantic inference and expansion, adjustment of similarity and correlation, threshold control, full-text retrieval and feedback are realized. Through the comparison with the traditional search method, it has been proved that semantic retrieval has higher recall ratio under the condition of guaranteeing precision with the support of ontology technology of semantic Web.

Acknowledgements

'Research of Structure, Process and Mechanism of Information Construction in Universities' supported by National Social Science Fund; the youth program of humanistic and social science of Ministry of Education 'Theoretical Model and Operating Mechanism of Remote Education System' supported by the Youth Program of Humanistic and Social Science of Ministry of Education; 'Reconstruction of Remote Education System with New Network Technology' supported by the Key Project of National Education Science Planning of Ministry of Education.

References

[1] D. Weimin, "Information Organization Technology and Methods of Semantic Web", Xuelin Publishing House, **(2008)**.
[2] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web, Scientific American", vol. 284, no. 5, **(2001)**, pp. 34-43.
[3] D. Zhihong, T. Shiwei, Z. Ming, Y. Dongqing and C. Jie, "Research Overview of Ontology", Acta Scientiarum Naturalium Universitatis Pekinensis, vol. 38, no. 5, **(2002)**, pp. 730-738.
[4] Y. Ting, "Research and Implementation of Semantic Search Technology Based on Ontology", Hangzhou Normal University, **(2011)**.
[5] R. E. Neches, R. T. Finin, T. Gruber, R. Patil, T. Senator, W.R. Swartout, "Enabling Technology for Knowledge Sharing", AI Magazine, vol. 12, no. 3, **(1991)**, pp. 16-36.
[6] T. R. Gruber, "A Translation Approach to Portable ontology Specifications", Knowledge Acquisition, vol. 5, no. 2, **(1993)**, pp. 199-220.
[7] W. N. Borst, "Construction of Engineering ontologies for Knowledge Sharing and Reuse", University of Twente, **(1997)**.
[8] http://jena.apache.org/index.html.
[9] http://jena.apache.org/tutorials/sparql.html.
[10] D. Jinxiang, "Service-Oriented Knowledge Management and Processing Based on Semanteme", Zhejiang University Press, Zhejiang, no. 8, **(2009)**, pp. 137-139.
[11] L. Qun and L. Sujian, "Computation of Semantic Similarity of Words Based on How-net", Computational Linguistics and Chinese Language Processing, vol. 7, no. 2, **(2002)**, pp. 59-76.
[12] D. Lin, "An information-theoretic definition of similarity", "Proceedings of the 15th International Conference on Machine Learning", vol. 98, **(1998)**, pp. 296-304.
[13] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", Proceedings of the International Joint Conference on Artificial Intelligence, **(1995)**, pp. 448-453; Montreal,Canada,.
[14] G. Jianhou, J. Yue and X. Youming, "Method and Application of Ontology", Science Press, vol. 6, **(2011)**, pp. 36-36.
[15] L. Wenjie and Z. Yan, "Algorithm of Semantic Similarity Between Concepts Based on Ontology Structure, Computer Engineering", vol. 36, no. 23, **(2010)**, pp. 4-6.
[16] W. Jian, "Discovery of Web Service Based on Ontology and Semantic Similarity of Words", Chinese Journal of Computers, vol. 28, no. 4, **(2005)**, pp. 595-602.

Authors

HaiBin Hu, he was born in 1980. He works as lecturer of China West Normal University. His main research area includes information retrieval, network technology and application of computer in education.

Shipin Chen, he was born in 1977 and his Ph.D. degree. He works as professor of China West Normal University. His main research area includes educational informationization and distance education.

