ATLANTIS PRESS
Advances in Social Science, Education and Humanities Research (ASSEHR), volume 130
2nd International Forum on Management, Education and Information Technology Application (IFMEITA 2017)

Research and design of education vertical search engine based on Ontology

Xiaozheng Wang[1] , Xiaoqiu Xia[2], Xiaoxiao Wei[3]
[1]Nanjng Xiaozhuang University, Jiangning, Nanjing, China
[2]Nanjng Xiaozhuang University, Jiangning, Nanjing, China
[3]Nanjng Xiaozhuang University, Jiangning, Nanjing, China
[a]xz_wang@163.com, [b] 1254644615@qq.com, [c]409269777@qq.com

Keywords: ontology; semantic; search engine
Abstract. In this paper, considering the needs of educational network resource search, a framework of ontology based semantic search engine system is designed based on open source technology Hadoop and Nutch. Several key technologies related to semantic retrieval based on tology are explored, including the construction of educational resource ontology and the storage of ontology data.Finally, this paper focuses on the key technologies of semantic search, which has some guiding significance and reference value.

Introduction

In recent years, with the continuous development of digital education, China has made great achievements in the construction of educational resources. The number of educational resources is huge and is increasing with the geometric progression. With the development of search engine technology, the function of general search engine becomes more and more powerful, and has achieved great success, but it still has limitations, such as the depth of search is not enough, and the precision is low, timeliness is poor. In particular, the existing search engine is implemented in the form of keywords, which does not satisfy the user's individual needs according to the individual differences of users, and the return results are often not satisfactory[1].

Ontology-based search engine means that the search engine's work is no longer rigidly dependent on the key words that users input, but it can be used for semantic reasoning of these keywords[2]. By linking the keywords and the concepts of their mapping in the semantic level, the problem of semantic comprehension can be solved in part. Semantic search fully tap the semantic information contained in the web page document information ,at the same time the user's retrieval requirements are converted to the corresponding semantic representation. The domain ontology is used for identifying and reasoning, and the user queries are understood at the semantic level, and the results based on ontology reasoning are returned to the user.

Based on Hadoop and Nutch, this paper designs a vertical search engine based on ontology for basic education, and focuses on how to realize the key technology of semantic search.

Basic concepts

In the broad sense, the semantic information includes semantic entities, grammatical relations, entity contextual features, textual structural features, etc. The semantic search engine not only gives the relevant network documents as the query results, but also in the ontology query related resources can also be given, semantic search engine as a hot spot of the new network informatics research , has been launched at domestic and abroad quickly[3].

Ontology has a good concept of hierarchy and logical reasoning support, use the relationship between the strict definition and concept of the concept of the concept to determine the precise meaning, represent the common recognized and shared knowledge, its wood is the field of knowledge sharing and reuse[4].Therefore, using ontology technology can solve the retrieval problem of educational resources at semantic level. The semantic search based on ontology in educational field

can accurately express the user query semantic and as a semantic document object query, can greatly improve the retrieval accuracy and retrieval efficiency.

System framework design

Using the open source software Nutch built on the Hadoop distributed system to crawl the basic resources of the basic education, and filter out irrelevant information of the basic education, and parse the crawled content, and save results into the distributed database Hbase.And then use the manual construction and automatic extraction technology to achieve the educational resources ontology library, retrieval of semantic query content using domain ontology, which makes the collection engine has the characteristic of "specialized, refined, deep, accurate, reliable and fast retrieval and update timely. The framework design of the system is shown in Fig. 1.

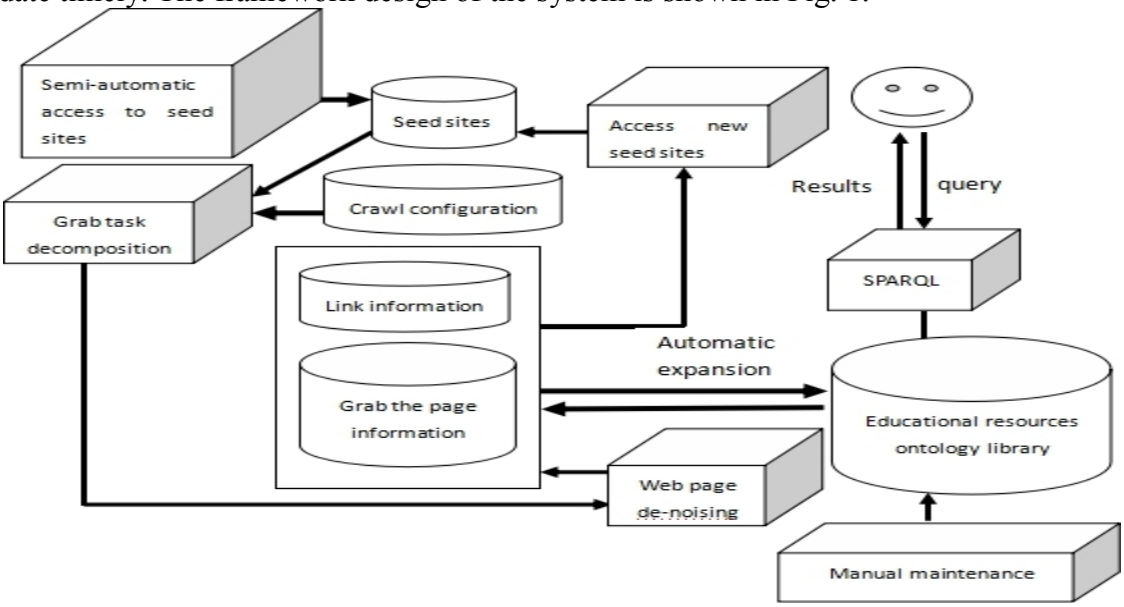

Fig. 1 Educational Resources Semantic Search Engine System Framework diagram

Key technology research

(1)Seed site selection and filtering

Because the search engine designed in this paper is aimed at the specific theme of basic education, in order to make the search site scope more representative , and only to crawl and the topic-related URL, and filter URL by using a certain strategy algorithm on the "web spider" .This process uses Web- Harvest open source software to extract the list of domain specific sites in open categories such as Intute, DMOZ, and generates site descriptions of XML files.In order to get more seed sites, you can use the words that represent domain features to retrieve more candidate sites URL through Yahoo, Search, and API.The candidate sites obtained by the way need to be filtered according to the PageRank value, the connectivity situation index and the subject correlation prediction algorithm, and the sites with little influence, difficulty in access and low correlation are excluded,and further by verification and classification, finally obtain high quality seed site.

(2)Automatic de-noising and de-duplication of web pages

Pages captured by Nutch, in addition to containing effective text content, also carry advertising information, client running code, copyright statements, column settings and other noise information.In order to provide high quality data, reduce the interference of noise information in the stage of information extraction and analysis, it is necessary to set the web page de-noising module according to the general characteristics of the noise information in the web crawling phase to filter the web content.The web page de-noising feature ensures that the pages stored in the content database are not duplicated, and new pages can also be identified.

(3)Distributed system construction

Based on the open source cloud platform Hadoop, a distributed system is constructed to improve the efficiency of information capture and information retrieval by using distributed platform. Make full use of the Nutch oriented interface plug-in technology, encapsulate the key modules, and make the system highly reusable, so as to lay a good foundation for the expansion of the system in the future.

(4)Construction of educational resource ontology

With the continuous development of Semantic Web research and the continuous development of practice, with the maturity of XML and RDF technology and the recognition of OWL language by W3C , ontology based on semantic web provide a solution to effectively develop, manage and use of education resources.

There is not a recognized standard framework of ontology construction, the current widely accepted is the five principles of Constructing Ontology proposed by Gruber in 1995 : clarity, consistency and scalability, the minimum constraint and integrity.These five principles are the basic ideas of domain ontology building, but the disadvantage is that they reflect the abstract content and difficult to grasp in practice, the researchers put forward many methods of ontology from different angles.This paper proposes a semi-automatic construction of educational resources ontology library model. The specific process is as follows:

First of all, according to the various authoritative vocabularies in the field of education, the domain ontology can be constructed, and the lightweight ontology can be constructed according to the simple semantic relations existing in the concept of vocabulary. The ontology constructed by this method is simple and easy to realize the large quantities of automatic conversion with writing a program. This method chooses the "Education Resource Construction Technical Specification" (CELTS-41) as a metadata scheme. Based on this, the core class of educational resource ontology is defined, and the Protégé software can be used to construct the ontology.After establishing the core class of the ontology of the education field, it is necessary to determine the attribute relationship between the ontology concepts. The attribute relationships in Protégé include two attributes: relationship attributes and numeric attributes.After the completion of the design of the class, relational attribute and digital attribute of the ontology, the overall architecture design of the educational resource domain ontology is completed, and then the educational resource information is input, that is, the instance information data.Based on the existing ontology library, this paper designs a model view of the ontology-based adaptive Web information extraction platform. The model view is shown in Fig. 2.

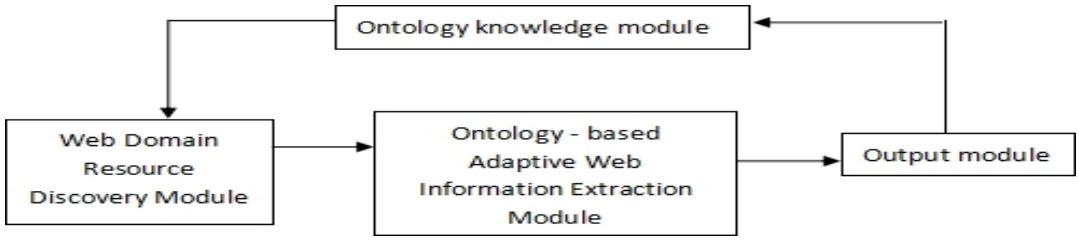

Fig. 2 A Model View of Adaptive Web Information Extraction Platform Based on Ontology

Among them, the Web domain resource discovery module: This module is to achieve the function of web crawler for different page types, directional get relevant education resources, through machine learning techniques to filter web data, will be submitted to the relevant education data obtained to adaptive information extraction module based on ontology.

Adaptive Web information extraction module based on Ontology: the module receives information from the Web information extraction module, and describes information combined with the corresponding information extraction task, calls the appropriate method to complete the extraction work of different types of data.

Output module: This module outputs the verified results to a specific database or knowledge base, and establishes the relation between the extraction result and the corresponding ontology, so as to realize the expansion of the ontology.

Ontology knowledge module: This module contains ontology knowledge related to the object to be extracted, which involves different education domain ontology, database description ontology, interactive relation ontology and various knowledge base resources.

Ontology data and instance data storage model

RDF Schema is a simple language, but it is too simple and it's descriptive ability is weak, difficult to represent complex knowledge, so it needs to be extended, the Web ontology description language OWL is an extension of RDF Schema.Based on relational OWL data storage model and distributed system platform, this paper proposes a OWL data storage model based on distributed database HBase.The ontology information of the OWL course designed in this paper is shown in Fig. 3.

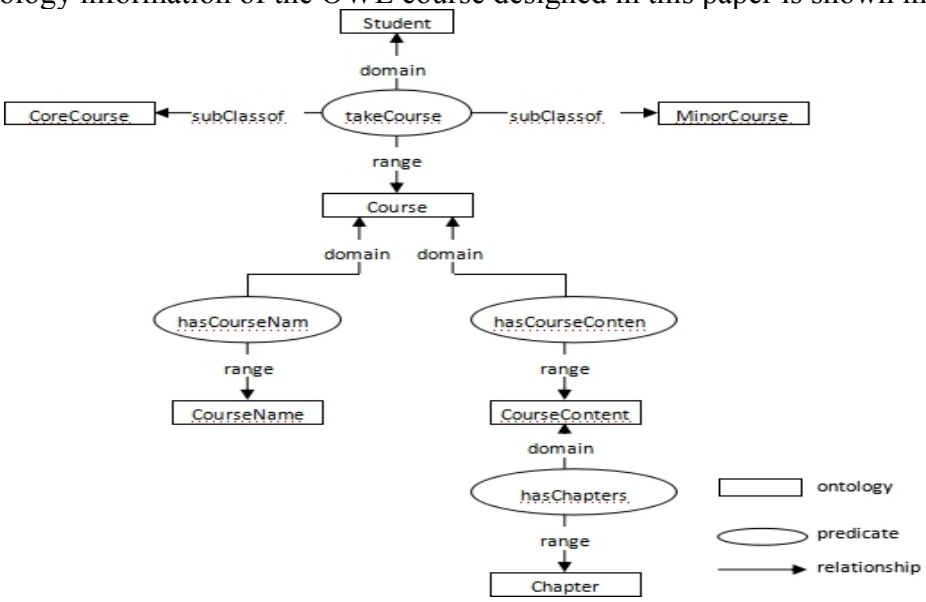

Fig. 3 OWL Course Ontology Information

This article uses the HBClass table to store the class information in the ontology definition. The class name is row-key, which contains two column families: properties, subClass, which store the attribute information and subclass information. The HBClass table stores values with the column labels of the column family, and the cell values are "1" by dynamically incrementing the columns to store multiple values. The specific storage structure is shown in Table 1.

Table 1 HBClass table storage structure

Row-Key	Column Family	
	properties	subClass
Course	properties :hasCourseName:"1"	subClass:CoreCourse:"1"
	properties :hasCourseContent:"1 "	subClass:minorCourse:"1"

The HBProperty table stores the attribute information defined by the ontology. The attribute name is row-key, which contains four column families: subProperty, inverseProperty, domain and range, which store the sub-attributes, reverse attributes, domain and range information of the attributes, such as the HBClass table ,the value is stored as a column label, and the cell value is "1" by dynamically incrementing the column to store the multi-value. The specific storage structure as shown in Table 2.

Table 2 HBProperty table storage structure

Row-Key	Column Family			
	domain	range	subProperty	inverseProperty
hasCourseName	domain:Course:"1"	range:CourseName:"1"		
hasCourseContent	domain:Course:"1"	range:CourseName:"1"	subProperty:hasChapters:"1"	

Then, we need to design corresponding tables to store RDF instance data. In this paper, we use RDF instance data division method to create two HTable tables for each class in the ontology, which is used to store the instance data of the class.

The table names for each of the two storage tables corresponding to each class are class name _S_PO and class name _O_PS. The column family of a table consists of the attributes of that class, including the attributes of the class itself and the attributes inherited from the parent class. Each attribute corresponds to a column family, column name is the attribute name. The data in the two table is the three tuple with the instance of that class.

The class name _S_PO table takes the subject of the triplet as the row-key, stores the triplet object with the column name. by dynamically adding columns to store multiple values, the value of cell is "1", as shown in the following table structure definition.

```
Row-key Subject: {
 Column Family Predicate1: {
   Column Object1: {tl: "1"}
   Column Object2: {t2: "1"}
   ...
 }
 Column Family Predicate2:{
  Column Object3: {t3; "1"}
   ...
 }
 ...
}
```

The class name _O_PS structure definition is shown below.

```
Row-key Object: {
 Column Family Predicate1: {
   Column Subject1: {tl: "1"}
   Column Subject2: {t2: "1"}
   ...
 }
 Column Family Predicate2:{
  Column Subject3: {t3; "1"}
   ...
 }
 ...
}
```

We create table HBType to record the class for each instance.the table uses the instance URI as the row-key, only includes one column family named type, which uses the column labels to store class name of the instances, by dynamically adding columns to store multiple values, the cell value is “1".

We create table HBInstance to record the instance for each class .the table uses the class name as the row-key, only includes one column family named instances, which uses the column labels to store the instances of the class, by dynamically adding columns to store multiple values, the cell value is “1".

Conclusion

This paper only has initially designed a semantic search engine framework based on ontology, which needs to be further explored and improved in many key technologies. We believe that in the near future, especially with the continuous improvement of ontology construction technology, semantic retrieval will become more and more widely used.

Acknowledgements

This work was financially supported by Key Laboratory of Trusted Cloud Computing and Big Data Analysis(Nanjing XiaoZhuang University), Jiangsu Engineering Research Center for Networking of Elementary Education Resources(BM2013123).

References

[1] Cun Zhou: submitted to Journal of knowledge economy(2011)

[2] Guier Feng: submitted to Journal of Computer Education(2007)

[3] Jing Zhang,Jie Tang; submitted to Journal of Chinese Computer Society Newsletter(2013)

[4] Information on https://baike.baidu.com/item