Book Information Retrieval Method Research Based on Word Sense Disambiguation

Yahong Li*
School of computer and Information Engineering, Nan Yang Institute of technology, Henan Nan Yang, China

*Corresponding author e-mail: ny6688@qq.com

Abstract. In this paper, formal description and data extraction of Book information is the core problem to solve the user query and information access. Along with the information resources diversification and rapid expansion, the existing description and extraction method had some defects, for example recall rate accuracy rate. In order to solve these problems, a new information retrieval method based on word sense disambiguation is provided in this paper, according to the features of book information, the domain concept set and the domain knowledge set is redefined. Based on it, construction process of the domain concept set and domain knowledge set is given, including domain concept automatic extraction, domain knowledge sets automatic construction and the similarity algorithm description. The experimental results show that the proposed method has higher recall and precision at this paper than the existing method, and has feasibility and good validity.

1. Introduction

Book information retrieval is based on the needs of users, using search tools and some techniques to inquire about ordering books, to get information that is consistent with the retrieval requirements. With the flourishing of the first online bookstore - Amazon online bookstore, the appearing of Dangdang online bookstore, Shanghai bookstore and other online bookstores, most of these books in the bookstore is mainly use the method of retrieving the directory of favorite books, especially favorite information. In addition, in each of these bookstores, the expression of message may be different. For example, the introduction of Press and publisher, some are expressed in the tabular form, while others have been descripted as characters. And the added new data is nor structural constrained in pre-defined, thus a lot of useful information cannot allow users to access books, resulting in the decrease of accuracy [1, 2].

The description of library information and data extraction is the premise of the integration and retrieval of books information, which has become a hot research topic and scholars at home and abroad have done a lot of beneficial attempt. Such as information extraction based on natural language processing, wrapper induction, and ontology, etc [3, 4]. These methods have made some breakthrough in some aspects, but the results are general, and the degree of satisfaction can not be reached.

In this paper, it combined some of the advantages of data extraction methods, such as web information extraction method based ontology [5], automatically acquire concepts method based on linear concept map [6], and add context features, grammatical structure rules, and the weights of
static-al information, provided a new information retrieval method based on word sense disambiguation. Compared with other methods, the method can not only improve the precision, but also have good feasibility and validity.

2. Formal Description of book information
As book information has a clear structure, we first need to extract the contents of the book information, and then give the integrated representation and stated problem. In this foundation, we can be rational and effective method to organize, manage and use data. This paper introduced domain ontology. Rovides unified and formalized description for the book information, and gives a new definition of domain ontology.

Domain ontology is that organized the concept in the field efficiently (Domain knowledge is the collection of, the concept and the relationship and constraint between concepts.), and makes the book information better sharing and reuse. According to the application and needs of research field, this paper gives the description concept of the domain ontology that includes the relation, constraints, hierarchical classification between concept and concept derives the new concepts and relations by new rules automatically. In this article, the important features of domain ontology are embodied in "Facing the compute" and "a normal person is very easy to obtain". "Facing the compute" refers to the domain ontology can use a computer to deal with problems. "a normal person is very easy to obtain" is that we can quickly obtain the required information according to our accumulated vocabulary of daily life or add new concepts to the field of ontology.

In conclusion, any kind of things can be defined as a domain ontology, which includes domain concepts set and domain knowledge set.

2.1. The formal method of domain concepts set (Di)
The Domain Concept Set is a formal description of the key words that abstract by fully understanding concepts, attributes, relationships etc, in the field.

Domain concepts set is composed of domain objects set and relationship set between objects.

Definition 1: Assuming that D denote the set of all areas, D = {D1, D2,……, Dk},

Di (1<=i<=k) represents some domain.

Definition 2: Di = <Oi, Roi >, Oi (1<=i<=k) represents all the concepts set of Di, Roi represents the relationships set between the concepts in Oi.

Definition3: Assuming that O denote the concepts set in the whole fields, Represents the relationship set between concepts in O, so:

O = O1 ∪ O2 ∪ O3 ∪……. ∪ Ok

R = R1 ∪ R2 ∪ R3 ∪……. ∪ Rk

D = <O,Ro >

2.2. The formal method of domain knowledge sets (Dk)
Domain knowledge set is the entity that is used to store the knowledge and is knowledge set which is structured, easy to operate, easy to use, comprehensive and organized, and is directed to problem-solving needs of certain specific areas. At present, knowledge set is divided into three levels, as shown in figure 1:

![Figure 1. The hierarchical structure of the domain knowledge set.](image-url)
Domain knowledge set is based on domain concepts set, and then bring new facts and rules into multiple domain concepts (as shown in figure 2), in order to deal with the original problem that single field cannot solve, while the domain concepts set can eliminate separate semantics, which can make knowledge more sharing and reuse. Such as field Dn said the concept of information about the school books’ information set, field Dm said about the concept set of network books, if one want to buy textbooks by refer the information that is provided in the web, we need to establish a relationship between the two fields, to solve the problem that is complex to queries, to provide convenience to users, to realize reuse information.

Building the relationship between the multiple domain concepts sets, which is based on the concept set of domain concepts and multi-level description method to realize the link of various fields

Definition 4: Assuming that x, y are the two elements that is Included in Di and Dj, H represents the relation of the two elements, so formal description as follows:

\[ \exists x \exists y (D(x) \land D(y)) \rightarrow H(x, y) \]

In short, the formal description of book information is "computer-oriented" premise that shows the world is constituted by the objects and relationships. And the extractive terminology is the first step of establishing domain ontology, defining the concept of lexical semantic unity is aimed to ensure the consistent understanding, so made the reasoning process explicit.

3. Domain Ontology Construction

Domain ontology is extracted by the expert’s concepts and relations in the field, and get the formal description of the field. It is the foundation that computer can understand, and also provides the condition that the computer can add the concept automatically.

3.1. Automatic extraction of domain concepts

The field concept extraction is a process that translating into computer can understand from external source or Web resource areas, and stored in a specific way. Extracting domain concepts include tasks:

1. Converting the existing concepts into a unique form by understanding, choosing, extracting, compiling, classifying and organizing.
2. Generating new concepts by learning and reasoning.
3. Checking and eliminating the concept of contradiction and redundancy to keep the concept of consistency and integrity constraints.

Using the artificial acquisition mode to extract the existing concepts, then the process is shown in Figure 3. For example, Dublin Core that is initiated by the OCLC has become the international standards, including 15 elements like Title, Creator, Date, Subject, Publisher, Rights, Relation, Coverage, etc. It is standard that covers ETF RFC 2413, ISO 15836, CEN / CWA 13874, Z39.85, Australia, Denmark, Finland, the United Kingdom, etc.
For new emerging concept, this paper provides a fully automatic extraction method, the implementation process as shown in figure 4.

In Figure 4, machine learning is the most important part, and it is the core of the domain concepts acquisition. It first divides the sentences into words that are based on semantic elements in "HowNet". Then adopts similar algorithm is applied to calculate similar word meaning, to increase the accuracy of concept extraction, eliminate redundancy according to Context feature, Semantic structure rule, Weights of statistical information. As shown in Figure 5.

1) Speech tagging. The main functions are word segmentation and part-of-speech tagging in the original text of the Web resources. They are based on "HowNet" Semantic original basis.

2) The keyword extraction. Select the text in the concept of a representative to determine the text of field that they belong to.

3) Check the keywords similarity in the field of concept. If the similarity were more than 85%, it means that it is the synonymous relationship; If the similarity is greater than 50% and less than 85%, it needs to be matched against its corresponding based on the secondary synonyms, to determine whether to add to the domain concepts set; If the similarities are less than 50%, then as a new concept, added to domain concepts set.

3.2. The automatic construction relation between the domain knowledge set
Construction of relational set in the different domain concepts set, an algorithm is proposed that is based semantic similarity and relevancy computation in HowNet. The algorithm evaluates the semantic similarity between the terms, and establishes the relation between the concepts by using...
cooccurrence relationship, semantic association, and appositive relationship of original meaning (Figure 6). There are 16 kinds of relations, including: synonymy, antonym, part-whole relationships, attributes-host relationships, material-product relations, tool-tool relationship, event-tool relationship, value-property relations, event-role relationships, etc.

![Diagram of concept relationship](image)

**Figure 6.** The establishment of relations between concepts.

Due to space limitations, the next example the calculation steps of word similarity algorithm:

1. Select concept 1 and concept 2 of the original expression dict1 dict2.
2. Judge part of speech con con1 and con2 by the dict1 and dict2.
   - If these two concepts are different parts of speech, such as one concept is function words, one is content words; it should match Similarity Sim with 0
   - If both are content words, move into (3)
   - If both are function words, move into (4)
3. The calculation of similarity between two content words:
   a) Calculate their similarity Sim1 from the dict1 and dict2 were taken out of the first sense element,
   b) The rest of the dict1 and dict2 were divided into three groups: independent sememe group, the original group and symbols original group.
   c) Calculate the similarity Sim2 about independent sememe group.
   d) Calculate the similarity Sim3 about the original group.
   e) Calculate the similarity Sim4 about the symbols original group.
   f) Get the similar degree Sim of two content words by the following formula:
   
   $$\text{Sim}(Con_{1}, Con_{2}) = \sum_{i=1}^{4} \beta_{i} \prod_{j=1}^{4} \text{Sim}_{j}(Con_{1i}, Con_{2j})$$

   Among them, $\beta_{i} (1 \leq i \leq 4)$ is adjustable parameter, and there is: $\beta_{1}+\beta_{2}+\beta_{3}+\beta_{4}=1$, $\beta_{1} \geq \beta_{2} \geq \beta_{3} \geq \beta_{4}$. It reflected from Sim1 to Sim4 with descending to the overall similarity of the role.
4. The similarity calculation between two function words:
   a) Get first original element of dict2 and dict1, respectively;
   b) If one is syntactic semantic element, one is relational semantic element, then Sim = 0;
   c) If the two are either syntactic or relational, then check the meaning of the original table dict in "HowNet", and the path length d of two, which is identified in semantic level system, it is a positive integer;
   d) By the formula:
   
   $$\text{Sim}(\text{semdict}_{1}, \text{semdict}_{2}) = \frac{\alpha}{d + \alpha}$$

   Calculate the similarity of these two concepts. The $\alpha$ is an adjustable parameters. $\alpha$ is the path distance when the semantic distance between two sememes is 0.5.
4. The instance validation and analysis

4.1. The experimental environment and case verification

The experimental data of the paper comes from the documents in the field of books from the Web resource. Then the resources in this field will be randomly selected to test the correctness of the proposed method. In the development environment, first of all, we construct the model of the book Domain by adopting Protégé. Then extract context feature, semantic feature, relation feature of concepts from term relation.

In the test environment, the concept set is improved according to the extracted domain information. That is, it will automatically add the concepts and the relationship between the concept and the concept in concept domain. Due to the limit of space, the following process describes the building process of parts relationship.

The description of the concept set in book domain:

\[ D_1 = \langle (\text{books}, \text{title}, \text{author}, \text{publisher}, \text{price}) \rangle \]

Where in \( E \) represents the equivalent relationship, \( P \) represents property relations.

The implementation steps are described as follows:

Step 1: Analyze web resources; improve the concept of centralized domain concepts;

In this paper, we extracted shopping site Books information and got the part of the concept as follows: Pricing, Dangdang, and publishing time.

Step 2: Calculating the similarity between concepts to determine the relationship between concepts.

According to the word similarity algorithm given in this paper, we obtain the following information:

\[ E (\text{price}, \text{pricing}), P (\text{books}, \text{pricing}), P (\text{books}, \text{Dangdang}), P (\text{books}, \text{publishing time}) \]

Step 3: we remove the redundant concept and improve the concept of the domain by using the similarity algorithm that is given in this paper. What we described is as follows:

\[ D_1 = \langle (\text{books}, \text{title}, \text{author}, \text{publisher}, \text{price}, \text{Dangdang}, \text{publishing time}), (E (\text{book}, \text{title}), E (\text{price}, \text{pricing}), P (\text{books}, \text{author}), P (\text{books}, \text{price}), P (\text{books}, \text{Dangdang}), P (\text{books}, \text{publishing time})) \rangle. \]

4.2. Test Analysis

In this paper, we extract some data of some Shopping websites to verify the validity of the algorithm. Firstly, we manually select the concepts and relationships in experimental data, which are including 1500 concepts (PAllConcept) and 2000 relationships (PAllRaship). Then for these concepts and relationships, we respectively adopt Semi structured data extraction method which is provided by this paper, natural language processing method, information extraction method based on ontology and based on wrapper induction to obtain concepts and relationships, including The concept of actual sampling (EPCConcept) and relationship of actual sampling (EPRship); Finally, the extracted content is analyzed and compared, we obtained the correct concepts (EAConcept) and relationships (EARship). The specific data are shown in table 1.

In order to verify the validity of the formal description of the book information and the correctness of the algorithm, in this paper, we aim at the date that is done by artificial means, which is based on the methods of natural language processing, ontology, and wrapper induction. The acquired data is shown as table 1. In this paper, we use recall and precision rate index as a measure of the method. Define it as follows:
Table 1. The data amount acquired by various extraction methods.

| Method                                        | Concept (1500) | Relationship (2000) |
|-----------------------------------------------|----------------|---------------------|
|                                               | concept of actual sampling | correct concepts | relationship of actual sampling | correct relationships |
| The method of this paper                      | 1427           | 1415                | 1975                          | 1770                  |
| Information extraction method based on natural language processing | 1330           | 1180                | 1732                          | 1270                  |
| Information extraction method based on ontology | 1350           | 1210                | 1760                          | 1345                  |
| Information extraction method based on wrapper induction | 1420           | 1325                | 1917                          | 1545                  |

Accuracy Rate = Number of correct information extracted / Number of all information extracted
Recall Rate = Number of correct information extracted / All number of correct information
Among them: Recall and precision rates in the range of [0, 1].
Number of correct information extracted = EAConcept + EARship;
Number of all information extracted = EPConcept + EPRship;
All number of correct information = PAIIConcept + PAllRaship.

Table 2. Experimental results contrast.

| Method                                         | Accuracy Rate | Recall Rate |
|------------------------------------------------|---------------|-------------|
| Information extraction method based on natural language processing | 0.8           | 0.7         |
| Information extraction method based on ontology [3, 4] | 0.82          | 0.73        |
| Information extraction method based on wrapper induction [5] | 0.86          | 0.82        |
| The algorithm of this paper                    | 0.94          | 0.91        |

The experimental results show that (see table 2), the method described in this paper can achieve a good result, and the accuracy rate and recall rate are higher than that based on natural language processing and that based on ontology method. Based on natural language processing and information extraction based on ontology, this paper adds context features, syntax structure rule and weight of statistical information, it can not only extract the concept of semi structure data, but also build the relationship between them, which is the foundation for the semi structured data integration.

5. Summary
In this paper, the method of data description and data extraction based on word sense disambiguation is combined with the context features, syntax structure and statistical information, which has raised the amount of information extraction. It improves the recall rate and accuracy. However, the implementation process of the proposed method is based on the data description, data extraction and storage from Web documents, which reduces the speed of data extraction, which needs to be studied in the future research.

References
[1] MARTNEZ-COSTA C, SCHULZ S, Ontology Content Patterns as Bridge for the Semantic Representation of Clinical Information. Applied Clinical Informatics. 2014.
[2] Jupp S, Malone J, Bolleman J, et al. The EBI RDF Platform: Linked Open Data for the Life Sciences. Bioinformatics. 2014.
[3] Abello A, Romero O, Bach Pedersen T, et al. Using Semantic Web Technologies for
Exploratory OLAP: A Survey. IEEE Transactions on Knowledge and Data Engineering. 2015.

[4] Galperin MY, Rigden DJ, Fernandez-Suarez XM. The 2015 Nucleic Acids Research Database Issue and molecular biology database collection. Nucleic Acids Research. 2015.

[5] Wang Yuanzhuo, Jiayan Tao. Information retrieval and data mining based on open network knowledge. Journal of Computer Research and Development. 2014.

[6] EVANGELISTA A T, HASSANIEN A E. Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft Computer. 2014.