Application of Intelligent Information Retrieval for Big Data Oriented Brain Science

Jia Baoxian

Liaocheng University, Shandong Province, 252059

Keywords: Brain Science; Big Data; Information Retrieval; Semantic Similarity Computation; Ontology

Abstract: With the development of brain and cognition technology, the research on brain science is more and more thorough. And big data related to brain science is emerging. How to acquire and share these data in time is a very urgent problem in the development of brain science. Traditional information retrieval technology is often based on keyword matching, which does not investigate the semantics of retrieval words or search phrases. It has great defects in the field of brain science. And it is difficult to meet the retrieval requirements of brain science. The information retrieval about brain science has characteristics of real-time and retrieval methods diversification. So combining Ontology technology and professional needs, we proposed intelligent retrieval model based on ontology, which is suitable for the field of brain science. We improved semantic similarity algorithm more suitable for brain science information query. And the experimental results show that the retrieval of semantic similarity based on the improved algorithm can improve the recall and precision of information retrieval.

Background

The study of brain science become the research focus in the world. The United States, the EU, Japan have developed brain research program. New technology and new methods, new discoveries continue to emerge the brain science. In document [1],Innovation and new technology research means creation objective conditions to solve the human brain the mystery.

In recent years, the theory of complex network, brain network visualization direction has achieved a breakthrough. It will play important significance for understanding the brain mechanisms and prevention of brain diseases. And it will promote the development of artificial intelligence systems. The progress of brain science is becoming more and more complex. At the same time, a large number of literature materials emerge in an endless stream. If we can index these resources in time and share them with the researchers in the most convenient way, we will greatly promote the progress of scientific research projects. Therefore, we need to design an efficient search system in the field of brain science to provide services for scientific research activities. The fundamental way to improve the quality of Brain Science information retrieval is to change disordered data into ordered knowledge, let computers understand the meaning of Brain Science information, and thus realize semantic retrieval.

Research status

In the field of modern brain science, information technology has become one of the ke y factors to promote the development of brain science. There are about more than 3290000 00 returned keywords in Google, including brain science, news and latest achievements. Fro m these data, we can see that at present, the information retrieval service in the field of fo reign brain science is mainly focused on the retrieval of the literature and information of periodicals and magazines.

The American National Biotechnology Information Center (National Center for Biotechnology Information, called NCBI) is very famous. It contains hundreds of internationally journals in the field of brain science and neuroscience, and provides all the documents, abstracts and part of the full text information. Without exception, they provide services for researchers by collecting information and latest achievements in related fields. Most of them only provide the most basic information retrieval service. However, this relatively simple way is becoming more and more difficult to meet the development of the field of brain science and the needs of scientific researchers. In document [2], the inspiration provided by the traditional information retrieval system is difficult to meet the needs of the users in the field of brain science. Therefore, it is very meaningful to establish a professional information retrieval system which can provide personalized services in the field of brain science. In document [3], semantic retrieval based on ontology can solve the problem of low precision and recall in information retrieval, and semantic similarity is the key technology that affects semantic retrieval.

The research on the semantic similarity calculation method of terminology by foreign researchers has formed a rich result. The focus is semantic similarity calculation based on corpus. Corpus is mainly concentrated in the WordNet, British National Corpus, Mihalcea pointed mutual information and LSA two kinds of corpus based methods in document[4]. Gabrilovich with Wikipedia, put forward ESA (display semantic analysis) method in document[5]. Marco SquaT++ is put forward and the balance of the maximum spanning tree method, mainly for Web search task in document[7]. The successful application system of semantic similarity includes the following.Onto Seek is a semantic retrieval system developed by Apple and IBM on the basis of retrieval content in document[8].

Ontology

Ontology is the key technology of semantic Web. It was first a philosophical concept. From the perspective of philosophy, Ontology is an objective existence of a systematic explanation or explanation. And it is concerned with the abstract nature of the objective reality. In the AI field, the document[13] defined Ontology is a clear formal specification specification of shared conceptual models, which contains four meanings: conceptual model (conceptualization), explicit (explicit), formalization (formal) and sharing (share).

The famous items of Ontology application in information retrieval include (Onto)^2 Agent[9], Ontobroker[10] and SKC[11]. The 3 projects also represent 3 directions, respectively. (Onto)^2 Agent helps the user to retrieve the existing Ontology on the required WWW, mainly using the reference Ontology. Reference Ontology is a Ontology built with the Ontology as an object on the WWW, which preserves the metadata of all kinds of Ontology. Ontobroker is oriented to web resources on WWW to retrieve the required web pages for users. And these pages contain the content of the user's concern.

Improved method of semantic similarity calculation

The improved calculation method of marking weight is improved. Considering the calculation of annotation weights, we combine the vocabulary frequency, location of words, and feedback from professional users as parameters to calculate the weight of annotation. TF-IDF is a word (concepts in ontology). The conclusion section is more important than text in the vocabulary, and feedback is used to adjust the weights of the artificial annotation, make up the automation possible defects brought by the semantic understanding.

The specific weight calculation formula is as follows.

Formula 4.1 $\mathrm{weight}_i = \mathrm{fren}_i + \mathrm{loc}_i + \delta_i$

In the formula, weighti refers to the concept of ontology. I is used as the weight value of annotation. Freni represents the word frequency parameter of ontology concept I, loci stands for the concept of standard, I is used as the location parameter of the annotation, delta I is the artificial adjustment factor.

First calculate the frequency parameters using the TF-IDF idea, because it can make the frequency parameter increases with word frequency and increased slowly, wherein n represents a specified number of words in the document, N said the number of specified document segmentation, C said the total number of documents in the document set, DF (WI) said the document set in the current wi document number. Formula 4.2 $\mathrm{freni} = \frac{n}{N} \times \log \frac{C}{\mathrm{df}(w_i)}$

Then determine the location parameters, according to the experience, most of the first paragraph of educational resources is often the introduction section, and the last section is the concluding paragraph often, in the title, introduction and conclusion in the article the author will generalize, apparently also a concept appeared in the three positions than some of the relative importance of the text. Here we will use the form of fine tuning word frequency parameters to show that if a concept appears on these 3 positions, it is equivalent to that the word appears more than m times in the text. After statistical m, the effect of 3 is better, otherwise loci takes 0, loci, so loci is calculated through the following formula.

Formula 4.3 $\mathrm{loci} = \frac{n+3}{N} \times \log \frac{C}{\mathrm{df}(w_i)} - \mathrm{fren}_i$

The last is the artificial regulation factor, taking into account the authority personnel feedback, high credibility, can correctly express the meaning of the paper, the weight calculation should occupy a larger proportion of the formula, to be considered with maximum weight of ideas, which is to find out the maximum value through the TF-IDF obtained by the method of the article, and then combined with the feedback the concept of delta I the original TF-IDF value, two were added as I.

Formula 4.4 $\delta_i = \sum_1^n \mathrm{Max}(\mathrm{fren}_j) + \mathrm{fren}_i$

The n is the number of participle of the document, and the frenj is the weight calculated by the TF-IDF of the word J. Finally, the word frequency parameters, position parameters and artificial adjustment factors corresponding to the concept I are filled into the formula 4.4, and the weight of the label can be obtained.

Based on distance semantic similarity, we use R (n, m) to express the semantic correlation between the two concepts, and set d1 and d2 as the two unrestricted concepts in the ontology. The semantic correlation between D1 and D2 can be calculated in the following way which is provided in literature[9].

Among them, L (d1, d2) means the length of the shortest path between two concepts in the same concept hierarchy on. ηis a regulator, this paper takes ηas 1.

Table 1 experimental results

Retrieval Mode	Recall Rate	Precision	QueryTime
Information retrieval based on keywords	63.7%	47.4%	0.065s
Information retrieval based on Ontology	98.3%	94.7%	0.062s

Using semantic similarity matching algorithm, we designed the intelligent retrieval system of the brain science. The system construction of Ontology, Web and the evolution of resource acquisition, annotation, semantic metadata extraction is performed offline, which does not affect the real-time response performance of the system. From the point of view of user query, we first need to transform and extend the retrieval form submitted by users, then we can inference the RDF model, and finally match the converted search mode with the inference model. Ontology data and document instance metadata are stored in the relational database. And the efficiency is very high. As can be seen from table 1, compared with the traditional information retrieval system based on keyword matching, the system has a great improvement in the recall and precision.

ACKNOWLEDGMENT

This research is partially supported by National Science Foundation of China 61603169, Basic education research foundation of Liaocheng University.

References

[1]. Liu Yadong, Hu Dewen, LIUYa-Dong, et al. High performance computing from the perspective of brain science [J]. computer journal, 2017 (9): 2148-2166.

[2]. Yang Xiaolong. (2014). Design and implementation of information retrieval system in the field of Brain Science (Doctoral dissertation, Zhejiang University).

[3]. Wang C G, Wang B, Yao W L, et al. Study on web information retrieval system based on ontology[J]. Computer Engineering & Design, 2008, 105(8):836–848.

[4]. Chandrasekar S，Dakshinamurthy R. A Novel Indexing Scheme for Efficient Handling of Small Files in Hadoop Distributed File System[C]. ICCCI，2013 International Conference on. Coimbatore，4-6 Jan. 2013，1-8

[5]. Mihalcea R, Corley C, Strapparava C. Corpus-based and Knowledge-based Measures of Text Semantic Similarity[C]// National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, Usa. DBLP, 2006:775--780.

[6]. Gabrilovich E, Markovitch S. Wikipedia-based Semantic Interpretation for Natural Language Processing[J]. Journal of Artificial Intelligence Research, 2014, 34(4):443-498.

[7]. Marco A D, Navigli R. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction[J]. Computational Linguistics, 2013, 39(3):709-754.

[8]. Guarino N, Masolo C, Vetere G. OntoSeek: Content-Based Access to the Web[J]. IEEE Intelligent Systems, 1999, 14(3):70-80.

[9]. Wei T, Jia Y, Zhang Z, et al. Improved hybrid semantic similarity algorithm for terminology application[C]// International Conference on Natural Computation and, Fuzzy Systems and Knowledge Discovery. 2016:1734-1738.

[10].ArpirezJ,PerezAG,LozanoA,etal.(Onto)2agent:An Ontology-based WWWBroker to Select Ontologies.In:Go-mez-PerezA,BenjaminsVR,eds. Proceedings of the Workshop on Application of Ontologies and Problem-SolvingMethods UK,1998,16~24

[11].Ontobroker.http://ontobroker.aifb.uni-karlsruhe.de
[12].SKC.http://www-db.stanford.edu/skc
[13].GruberT R.A Translation Approach to Portable Ontology Specifications.Knowledge Acquisition,1993,5:199~220