International Journal of Database Theory and Application
Vol.8, No.3 (2015), pp.259-268
http://dx.doi.org/10.14257/ijdta.2015.8.3.22

An Approach for Optimizing Library Digital Resource Based on Semantic Information Retrieval

Chaoyang Ji

Xuchang University, Xuchang, China
E-mail:Pjcy@xcu.edu.cn

Abstract

In order to enrich and improve the collection of library digital information resources, this paper puts forward an optimization method of library digital resources based on semantic information retrieval. The method collects related information automatically from the Internet using semantic information retrieval, and selects the relevance value meeting the preset threshold of the network information to expand and update library digital resources. The experiment result shows that these methods gets a good expectant performance and dramatically optimize the library digital resources and improve the efficiency of resource retrieval and utilization.

Keywords: *semantic information retrieval; digital resource; digital library; resource optimization*

1. Introduction

Library digital resources as an important support condition of knowledge production and innovation, covering variety of subject knowledge and social knowledge, is the main pattern of manifestation and existing form of the traditional literature resource in the digital library. Further development and utilization of library digital resources, to provide targeted and comprehensive service of knowledge for all types of users, has a very important significance for enhancing the digital library's knowledge service ability and level, promoting the creation of knowledge, the construction of national knowledge innovation system and economic development.

However, with constantly improving of network infrastructure and continuously growth and update of the network spatial information capacity, the frequency of allowing users to access to academic information and knowledge to use the public digital resources representatived by Google, Google Scholar is significantly higher than the utilization of University Digital Library to obtain information, public digital information resources system is becoming the first choice platform for users to obtain professional knowledge and information. The main reason is that although the construction of the digital library has already begun to take shape in many of our colleges and universities, the total amount of digital resources has even exceed the total gross of the traditional literature resources, but still follows the traditional relatively extensive resource procurement scheme in the resources construction.

In the face of the rapid growth of digital resources, diversification of publishing mode, professional and marginalization of social knowledge needs, the traditional relatively extensive resource procurement scheme becoming more difficult to meet the needs of information and knowledge for the vast number of users. While the public digital resource system cannot get beyond the university digital library in the organizational structure and the total amount of resources, but because these resources system use a network engine automatic search to get the latest information source for optimization and perfection of its own knowledge system, to realize fine and comprehensive resource allocation, so the

ISSN: 2005-4270 IJDTA

efficiency of resource use and the size of the user are superior to digital library of university.

Semantic information retrieval as a kind of intelligent resource acquisition method , applied to the construction and optimization of library resources, to make the digital library resources from buying changing into the combination of purchase, leasing, real time network retrieval diversified resources establishment, is the effective way to improve the efficiency of utilization of digital resources of library, and is the inevitable requirements to optimize and improve library digital resources service system and the user satisfaction.

At present, research on semantic information retrieval mainly focus on three aspects: one is the query technology based on ontology, namely in the use of the hierarchical structure relationship of ontology and concepts collection to disambiguation semantic and query expansion for the queried contents; two is the semantic annotation problem, namely, select the appropriate annotation model for semantic annotation of text content to realize the resources semantic construction, such as use cascaded hidden Markov model to extract semantic information and annotation, use the vector space model to weight and sort for massive keywords; three is the semantic relationship retrieval, namely take semantic relationship between concept and concept, text and text, webpage and webpage as retrieves content, such as the method and technology involved in the semantic relationship retrieval, Barnaghi and Aleman-Meza have carried out a systemic research.

These studies contribute to the expansion of the network information resources collection and database resources in a certain extent; provide methods and technical support for the intelligent acquisition of the network information resources. Therefore, this paper on the basis of these studies is combined with semantic information technology to optimize Library Digital Resources. The research work is mainly reflected in: blend the ideas and methods of the semantic information retrieval in the construction and optimization of the digital library resources, study how to use the semantic information retrieval technology and method to search open source resources highly related to digital resources in cyberspace and use search results to optimize and improve digital resources automatically, thus to provide fine and comprehensive digital information resource service for users, to enhance the efficiency and satisfaction of using digital resources; finally, use the experimental analysis to verify the feasibility and efficiency of the method.

2. Resource Optimization Method based on Semantic Information Retrieval

The whole structure of optimization method of library digital resources based on semantic information retrieval is shown in Figure 1.

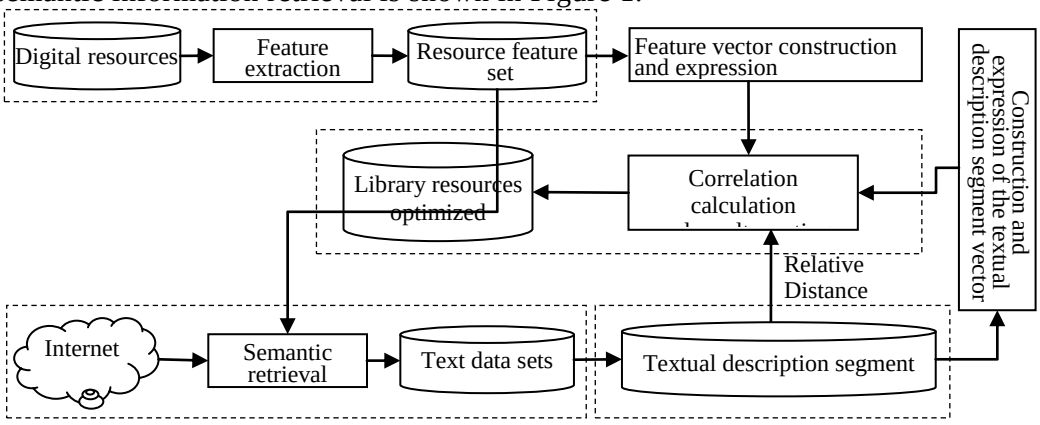

Figure1. The Flow Chart of Digital Resources Optimization Method based on Semantic Information Retrieval

The method aims to enrich and perfect the library digital information resources, with the semantic information retrieval as the core, according to the resources characteristics of library digital resources, using semantic information retrieval technology to collect associated network information automatically, through correlation calculation, select the correlation value to meet the preset threshold network information to expand and update library digital resources, to realize automatic optimization of library digital resources. Specifically, the method includes 4 steps: the extraction of digital resources characteristic, semantic information retrieval, the feature vector construction and expression, correlation calculation and results sorting.

2.1 Digital Resources Characteristics Extraction

Library digital resources characteristics extraction is to extract resource characteristics information of the target resource from digital resources standard described format, including resource title, author, subject of the belonged field, classification number, brief introduction, network abstract etc. These features information help to identify resource types, to provide basis to retrieve relevant Webpage information, used to calculate relevance of the text information fragment contained in library digital resources and Webpage. At present, the technology and method of extracting digital resources characteristic is comparative maturity, this paper takes example by multi-modal semantic relatedness extraction method and ideas presented in Ref. 12 to extract features of library digital resources, and store the extracting results to the set of library resources characteristics.

2.2 Semantic Information Retrieval

Semantic information retrieval is to collect webpage information related to target information automatically by using the characteristic information of library digital resources from the Internet. The process mainly includes the expansion and optimization of semantic query, the retrieval process and search result re-ranking.

(1) The expansion and optimization of semantic query. Semantic query expansion and optimization is mainly based on the characteristic information of library digital resources to get concept related to its semantic, and then use the concept to expand characteristic information of library digital resources. This query expansion, can not only narrow down the scope of the query expansion effectively, but also can ensure consistent of subject of characteristics information before and after expansion, make sure the final results improve the retrieval recall and ensure the recall ratio. In the process of semantic query expansion, word sense disambiguation (Characteristics information to the concept of semantic mapping process) is the key to impact accuracy of expansion query.

This paper uses query disambiguation method based on local context to realize semantic query expansion. Local context query disambiguation is to determine the semantic features of library digital resources based on the meaning of ambiguous characteristics information. The method uses HowNet as the disambiguation data source, at the same time use word sense definition, examples, structured semantic relations as a number of knowledge sources in HowNet to construct the vocabulary set and the domains description set, for the similarity $Sim(w, C, S_i)$ of each sense of ambiguous words and local context vocabulary is:

$$Sim(w,C,S_i)=\frac{\sum_{j=1}^{|C|}W(C_j)\times Gauss(dis(C_j,w))}{\sum_{i=1}^{n}\sum_{j=1}^{|C|}W(C_j)\times Gauss(dis(C_j,w))} \quad (1)$$

$$W(C_j)=\begin{cases}0\\ weight(S_i^k), if\ C_j=S_i^k\cap S_i^k\in RG(S_i)\end{cases} \quad (2)$$

Among them, n represents the total number of ambiguous word W in resources characteristics information, |C| represents the total number of words contained in the context C, C_j represents the j vocabulary in C, RG (S_i) represents the vocabulary meaning S_i, weight (C_j) represents the weight of C_j, the weight is determined according to the extended level of C_j in HowNet. Dis (C_j, t) represents the number of intervals characters from C_j to W, Gauss (dis (C_j, t)) is distance weighting factor determined by using the Gauss formula.

As the field property and the using frequency to make Sim(w, C, S_i) calculations appear deviation in the actual word sense disambiguation process, therefore, need to optimize formula Sim(w, C, S_i) by using domain attributes and word frequency information.

$$Sim^*(w,C,S_i)=M^a\times(Sim(w,C,S_i))^{\log f}+\sum_{j=1}^{|C|}DomSim(S_i,C_j) \quad (3)$$

$$DomSim(S_i,C_j)=\max\sum_{|DD(S_i)\cap DD(C_j^k)|}\frac{1}{|DD(S_i)|}\times\frac{1}{|DD(C_j^k)|} \quad (4)$$

Among them, M represents the number of semantic overlap words between RG (S_i) and C, a represents the number of semantic overlap lexical between S_i and C, f represents word frequency sort value of S_i in HowNet. |DD (S_i)| represents the total of field label in DD(S_i), C_j^k represents the Kth meaning of C_j. After using the formula calculation similarity of each meaning of ambiguity word and local context lexical, can realize and optimize the process of mapping from characteristics information to the semantic concept according to the similarity calculation results, then realize the semantic query expansion of characteristics information.

(2) The process of retrieval and search result re-ranking. After disambiguation expansion and optimization of library digital resources characteristics information, constructing the resource feature semantic vector Q_C= (qc_1, qc_2, qc_m). Among them, m is the representation of dimension of semantic vector space; qc_i represents the Ith element weights of query semantic vector. Then based on the vector constructing semantic vector D_C= (dc_1, dc_2… dc_m) of retrieval returned documents. Among them, dc_i represents the Ith element weight of document semantic vector; the calculation method is using the cf-idf method:

$$dc_i=cf_{D,i}\times idf_i,\ cf_{D,i}=\sum_{t_k\in T_i}tf_{D,k},\ idf_i=1-\frac{\log(1+hypo(c_i))}{\log(CS)} \quad (5)$$

Among them, $cf_{D,i}$ represents the appear frequency of characteristics concept c_i in the document D, idf_i represents the inverse document frequency of the characteristics concept c_i, T_i represents the vocabulary set of meaning c_i in the document D, the $tf_{D,k}$ represents the appear frequency of words t_k in the D, hypo(c_i) represents the total number of under concept of characteristics concept c_i, CS is the total number of concept in HowNet. Then use Cosine included angle cosine method to calculate semantic correlation Sim (Q_C, D_C) between library digital resources features and network text information.

$$Sim(Q_C, D_C) = \frac{\sum_{i=1}^{m} qc_i \times dc_i}{\sqrt{\sum_{i=1}^{m} qc_i^2} \times \sqrt{\sum_{i=1}^{m} dc_i^2}} \quad (6)$$

2.3 The Construction and Expression of the Feature Vector

The construction and expression of the feature vector is mainly contains two parts: The construction and expression of the textual description segment (TDS) vector, the construction and expression of library resources features vector.

(1) The construction and expression of TDS vector. After storage network information accessed by semantic information retrieval to text data sets , take each text as an TDS, after extraction of word segmentation, semantic annotation, features concept for each fragment, construct the vector for each TDS using the formula ⑤.

(2) The construction and expression of library resources features vector. The process of constructing feature vectors by using the CF-IDF method, the main concern is the appearing frequency of features information in the text, but for the library resources features information, the relative position of characteristics information appearing in the description of the resource file is also important, for example, the emergence of the vocabulary in the title of thesis, its importance is obvious higher than other position. Therefore, in the construction of resources characteristics information vector, should amend weight according to the different position of features information. Based on this, this paper improves the cf in the cf-idf method:

$$cf_{D,i}^{*} = \frac{\sum_{j=1}^{num_{i,d}} pos(t_i, j)}{\sum_{k=1}^{n} \left(\sum_{j=1}^{num_{k,d}} pos(t_k, j) \right)} \quad (7)$$

Among them, $num_{i,d}$ represents the number of times of occurrence of the feature concept t_i in text information fragment D, pos(t_k, j) represents the weight of the jth position of feature concept. Complete the constructed and expressed of library resources features vector by using formula ⑦.

2.4 The Correlation Degree Calculation and Results Sorting

Considering the particularity of library digital resources, this paper calculates the correlation degree between TDS and library digital resources from two aspects: text semantic similarity, the relative distance of TDS and library digital resources characteristic information in the place of webpage.

(1) Text semantic similarity between TDS and library digital resources characteristic information. It is mainly calculated by using formula ⑥ in 2.2 sections.

(2) The relative distance of TDS and library digital resources characteristic information in the place of webpage. Generally speaking, the closer the distance between the text content and library digital resources characteristics information , the bigger the possibility of the existence of association between the webpage content and library digital resources, therefore, to perfect and expand library digital resources is through the network text content extracted near the characteristic information.

Through compositing the calculation results between the two, get the overall relevancy value.

$$\text{Relevance}(R_C, TDS_C) = \frac{\text{Sim}(R_C, TDS_C)}{\log(\text{dis}(R_C, TDS_C) + 1) + 1} \quad (8)$$

Among them, Sim (R_C, TDS_C) represents text semantic similarity between the library digital resources R_C and network text information fragments TDS_C, dis(R_C, TDS_C) represents the distance between the library digital resources R_C and network text information fragments TDS_C.

3. Experiment and Result Analysis

This paper take library digital resources in the library of author's University as the target data source, to expand and optimize library digital resources by using the method proposed in this paper for computer science, economics, library and information science, agriculture, medicine . The experiment content of this paper is mainly divided into two parts:

(1) The correlation between expanded text information fragments and library digital resources. Select 200 published scientific literatures for each of 5 areas randomly, extract characteristic information of related field literature, and then use the method of section 2.2 retrieval relevant webpage text by using these characteristics information from the network space, and storage the text of more than 0.70 similarity as extended text, data size of each extension text as shown in table 1.

Table1. The Number of Original Text and Retrieved Text

selected fields	the number of original text	the number of retrieved text
computer science	200	3593
economics	200	2958
library and information science	200	2313
agricultural informatization	200	1768
medicine	200	4216

Calculate the correlation of the original text and the extend text respectively by using the method of section 2.3 and section 2.4, and take the internet text related degree more than 0.70 as the eventual expansion information, the experimental results as shown in table 2.

Table 2. The Average Correlation Degree and Expansion Degree

selected fields	computer science	economics	library and information science	agricultural informatization	medicine
average correlation degree	0.7382	0.7454	0.7621	0.7597	0.7239
expansion degree	13.26	11.02	8.81	6.72	15.26

Through tables 1 and 2, in the process of optimizing library digital resources by using the method designed in this article, has achieved the ideal results in the average correlation degree and expansion degree. Among them, the expansion degree of computer science, economics, medicine science is better than that of library and information science and agricultural fields, but the average correlation is lower than that of Library and information science and agriculture, the main reason is that the extension of computer science, economics, medicine is quite broad, lead to larger fluctuation in the process of calculating the correlation, resulting low average degree; and the library and information science, agricultural informatization is relatively specific, in the process of constructing the feature vector and retrieval, the degree of correlation obtain the text is average, and feature vector constructed can well reflect the specific information in this field, so the average correlation degree is higher, but because of the extension is narrower, the total number of network text gained is less, expansion level is lower.

(2) User retrieval efficiency comparison before and after optimization. Retrieve data set for each field before and after the extended, and utilize evaluation index commonly used in retrieval field --- precision (Precision, P), recall (Recall, R) and F1 value to test the experimental result, wherein, calculation formula of P, R, F1 is:

$$P = \frac{A}{A+B}, R = \frac{A}{A+C}, F_1 = \frac{2PR}{P+R}$$

Among them, a represents the number of related text searched in retrieval, B represents the number of related text searched in retrieval, C represents the number of related text not retrieved. The experimental results as shown in Table 3, table 4, and table 5.

Table 3. Comparison of Precision before and after Extension

selected fields	computer science	economics	library and information science	agricultural informatization	medicine
before the expansion	0.7529	0.7284	0.7832	0.8356	0.7179
after the expansion	0.8497	0.8179	0.8926	0.8875	0.8083

Table 4. Comparison of the Recall Rate before and after Expansion

selected fields	computer science	economics	library and information science	agricultural informatization	medicine
before the expansion	0.6852	0.6294	0.7035	0.7371	0.6288
after the expansion	0.7835	0.7584	0.8261	0.8357	0.7518

Table 5. Comparison of the F1 Value before and after Expansion

selected fields	computer science	economics	library and information science	agricultural informatization	medicine
before the expansion	0.7175	0.6753	0.7621	0.7833	0.6704
after the expansion	0.8153	0.7870	0.8581	0.8608	0.7790

Through table 3, table 4 and table 5 it can be shown that the retrieval efficiency after extension in the recall, the recall rate, F1 index is superior to before. The main reason lies in the digital resources of library information expanded more comprehensively and fully, the calculation results is more close to the reality, so the performance in searching efficiency is better.

4. Conclusion

Aiming at the existing defects of the library digital resources such as low utilization efficiency, low user satisfaction, lack of comprehensive resources, this paper studies and puts forward the optimization method of digital resources based on semantic information retrieval. The methods put semantic information retrieval into construction and optimization of digital library collection resources, in the use of semantic information retrieval technology and method to search open resources highly interconnected collection of digital resources on the Internet, by using the correlation of search results and the characteristics information of digital resources to optimization and perfection of library digital resources collection automatic. The experimental results show that, the method can obtain a large number of network open source information highly correlated with the library digital resources, can greatly enrich and perfect the library digital resources and improve the retrieval efficiency of library digital resources.

Acknowledgements

Supported by Science and Technology Research Project of Education Department of Henan Province, Project Number: 13A520746.

References

[1] K. Changbo, H. Zhiqiu, “Self-adaptive semantic web service matching method”, Knowledge-Based Systems, vol. 35, no. 11, **(2012)**, pp. 41-48.
[2] H. Xiaoling, L. Wen, “Optimizing the Digital Resources Construction of Academic Libraries on the Data Analysis: Case Study of Hunan University Library”, Library Work in Colleges and Universities, (2012), vol. 32, no. 1, pp. 58-60.
[3] M. Sussna, “Word sense disambiguation for free-text indexing using a massive semantic network”, Proceedings of the second international conference on Information and knowledge management, **(1993)**; Washington, D.C., USA.
[4] C. Aijun, Z. Zhaozhong, M. Lu, “The Tendency of Electronic Resources Development by Analyzing the Statistics of Three Consecutive Years of Electronic Resources Expenditures in Chinese and American Academic Libraries”, Journal of Academic Libraries, vol. 30, no. 1, **(2012)**, pp. 55-58.
[5] C. Lijia, “Development and Enlightenment of the British Library Digital Collection”, Library Theory and Practice, vol. 2, **(2012)**, pp. 93-96.
[6] W. Hongjuan, “Research on the Construction and Maintenance of University Digital Library Information Resources”, Journal of Suihua University, vol. 33, no. 2, **(2013)**, pp. 132-134.
[7] M. Albanese, P. Capasso, A. Picarcllo, A. M. Rinaldi, “Information Retrieval from the Web: An Interactive Paradigm”, Proceedings of the International Conference of Multimedia Information Systems, **(2005)**; Sorrento, Italy.
[8] D. Hui, Y. Chuanming, J. Ying, Y. Ning, X. Guohu, Z. Hua, “Research on the Ontology-based Retrieval Model of Digital Library (Ⅱ)---Semantic Information Acquisition”, Journal of the China Society for Scientific and Technical Information, vol. 25, no. 4, **(2006)**, pp. 451-461.

[9] P. Castells, M. FernaNdez, D. Vallet, "An Adaptation of the Vector Space Model for Ontology-based Information Retrieval", IEEE Transactions on Knowledge and Data Engineering, **(2007)**, vol. 6, pp. 161-272.
[10] P. Barnghi, W. Wei, J. Kurian, "Semantic Association Analysis in Ontology-based Information Retrieval", Handbook of Research on Digital Libraries Design Development and Impact, **(2009)**, pp. 131-141.
[11] B. Aleman-Meza, C. Halaschek, I. B. Arpiner, "Context-aware Semantic Association Ranking", Proceedings of The first International Workshop on Semantic Web and Databases, **(2003)**; Berlin, Germany.
[12] W. Ruijia, L. Yao, "Study on the Feature Extraction and Expression System of Multi Modal Semantic Information for Scientific and Technical Literature", Journal of Academic Libraries, vol. 30, no. 5, **(2012)**, pp. 71-76.
[13] W. Ruiqin, "Information Retrieval Model Based on Semantic Processing Technology", Journal of the China Society for Scientific and Technical Information, vol. 31, no. 1, **(2012)**, pp. 9-17.
[14] L. Wang, M. Li, S. Cai, G. Li, X. Bing, Y. Fuqing, "Internet Information Search Based Approach to Enriching Textual Descriptions for Public Web Services", Journal of Software, vol. 23, no. 6, **(2012)**, pp. 1335-1349.
[15] C. Deng, H. Xiaofei, L. Zhiwei, M. Weiying, W. Jirong, "Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information", Proceedings of the 12th Annual ACM Int'1 Conference on Multimedia, **(2004)**; New York, USA.
[16] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G.M. Petrakis, E. E. Milios, "Semantic similarity methods in wordNet and their application to information retrieval on the web", Proceedings of the 7th annual ACM international workshop on Web information and data management, **(2005)**; Bremen, Germany.
[17] H. Defang, Z. Jianxun, "Study on In-depth Integration of Library Collections Based on Semantics", Journal of Library Science in China, **(2012)**, vol. 38, no. 7, pp. 79-87.
[18] C. Whitelaw, N. Garg, S. Argamon, "Using appraisal groups for sentiment analysis", Proceedings of the 14th ACM international conference on Information and knowledge management, **(2005)**; Bremen, Germany.
[19] S. Bratus, Anna Rumshisky, Rajendra Magar, Paul Thompson. Using domain knowledge for ontology-guided entity extraction from noisy, unstructured text data. Proceedings of the Third Workshop on Analytics for Noisy Unstructured Text Data, (2009) July 23-24; Barcelona, Spain.
[20] D. Yoo, "Hybrid query processing for personalized information retrieval on the Semantic Web", Knowledge-Based Systems, vol. 27, no. 3, **(2012)**, pp. 211-218.
[21] Z. Huaiguo, L. Guangda, T. Cuiping, Z. Jingjuan, Q. Lin, "Problems and Solutions on Acquisition Quality of Digital Resources in Libraries", Library and Information Service, **(2012)**, vol. 56, no.1, pp. 112-115.

Authors

Chaoyang Ji, he received his Master of Engineering in Computer Science from University. Now he is associate professor of computer science at Information Engineer Department, University. Since 2011 he is Member of IEEE. His current research interests include different aspects of Artificial Intelligence and Information Retrieval.