Analysis of Heterogeneous Data Integration and Document Clustering method in Digital Library

Fang Liu, Guoqiang Lu*

Mu Danjiang Medical University, Mu Danjiang 157011, China

Abstract

This paper proposes a hierarchical organization framework model of heterogeneous data integration and resource utilization of digital library. The model is divided into metadata layer, ontology layer, link data layer and application layer from top to bottom. Aiming at the drawbacks of the traditional document clustering methods, an improved LDA method is proposed to cluster the shared subject documents of the same category. The results show that the retrieval time of the proposed linked data organization framework model of digital library is much smaller than that of the traditional model, and the search response time is reduced by 56.3%. The establishment of the linked data organization framework model of digital library can greatly increase the interconnection of distributed heterogeneous information and make the data source more easily crawled by search engines. TC-LDA algorithm has the highest retrieval accuracy and the best stability. The time consumption of the five algorithms is almost the same, thus it hard to get the algorithm with advantage of time-consuming. The improved LDA text subject clustering algorithm has more advantages than the traditional text clustering method, which is obtained from the analysis of clustering accuracy and time consuming.

Keywords: digital library, heterogeneous data, text resource integration, improved lda model, document clustering.

1.INTRODUCTION

At present, the huge data resources in the digital library are composed of many scattered and independent information, which leads to the inconvenience of the information retrieval and query (Frommholz et al., 2003). Digital library urgently needs to reorganize, cluster and integrate the existing information resources so that the visitors can get all the resources by sending a query request, and realize the network information sharing, which is the current research hotspot (Lagoze & Sompel,2003).

Since the Linked Data has been proposed, it was quickly applied to heterogeneous data resources integration of the digital library (Foster,2001). The linked data method has wide application prospect in integration of isolated information and semantic interoperability of resources(Meßmer,2013). In recent years, with increasing of linked data of the related information resources, the original method of data integration is no longer applicable, how to more directly and effectively query the network heterogeneous data has become an urgent problem to be solving (Alemu et al.,2012), The visitor is able to benefit from the semantic network's excellent expression ability, while shielding its complexity (Alemu et al.,2012).

Due to the disadvantages of traditional document clustering methods, that is document clustering is only based on the co-occurrence rule of words, and can't explain whether the two documents have a common theme (Burke, 2002). Researchers have made relevant improvements, at present the widely used clustering algorithms including LDA
method, kmeans, FCM (fuzzy c-means) and AP (Affinity Propagation) algorithm (Koren et al., 2009).

Regarding to the difficulty of unified access and retrieval of distributed data in digital library (Oscar et al., 2009), this paper puts forward a hierarchical organization framework model of heterogeneous data integration and resource utilization of digital library, which makes the data source more easily crawled by search engine (Mollá and Vicedo, 2007). At the same time, Aiming at the drawbacks of the traditional document clustering methods, an improved LDA method is proposed to cluster the shared subject files in the same category to improve the existing shortcomings of the easy-to-shadow text catalogs and easy-to-distortions of the digital library (Bizer, 2009). So that retrieval of document information can be more accurate and convenient.

2. ANALYSIS OF INTEGRATION AND RESOURCE UTILIZATION OF HETEROGENEOUS DATA IN DIGITAL LIBRARY

2.1 The Construction of hierarchical organization framework of digital library

Construction of the linked data Digital library model needs to have the following conditions:

(1) To realize the semantic description (Lopez, 2011), retrieval and related data association of distributed heterogeneous data resources;

(2) To achieve resource association of different knowledge unit of the digital library, form text information multi-level network framework (Shehata, 2009);

(3) To achieve a seamless junction of network professional text message;

(4) To achieve unified retrieval and access of distributed heterogeneous data resources.

![Figure 1. The resource description and organization framework of digital library](image-url)
Based on the above principles, the hierarchical organization framework is adopted to resource description of the digital library (Shehata, 2009). The model is shown in figure 1. As can be seen from the figure, the model is divided into metadata layer, ontology layer, link data layer and application layer from top to bottom.

In the digital library, different metadata specifications are used to describe the network resources and text resources, there are similarities between the metadata but not fully compatible (Blei, 2003). Because of the lack of accurate and uniform definition of metadata, it is difficult to understand and deal with metadata in the way of computer technology, so metadata can't solve the problem of heterogeneity of digital library resources (Rubin, 2012). The integration of the concepts and attributes in different metadata and the construction of the core metadata ontology of text resource sharing can better solve the problem of heterogeneity of different metadata, that is the metadata ontology layer shown in figure 1, The construction of core metadata ontology can formalized describe the core elements of the metadata specification, in this way the flexibility of metadata is retained. The third layer of the model framework is the link data layer. The release of the linked data can reorganize the ontology metadata, strengthen the connection between the related resources and external information in the digital library, and make all the resources of the digital library constitute the organic whole. The last layer of the organizational framework is the application layer. After realizing the data and resource association of the digital library, it is necessary to carry out the unified browsing, searching and semantic interoperation of the linked data to improve the intelligence of the retrieval, so that access can be based on their needs to obtain accurate query results (Pfeiffer and Pfeiffer, 2007).

Based on the limitations of metadata, it is necessary to construct a mechanism to achieve semantic interoperability between heterogeneous data. In this paper, an ontology description language (RDF, OWL, DCTERMS, etc.) is proposed to describe the metadata based on metadata normalization. Finally, the metadata ontology is constructed to realize the metadata conversion and semantic interoperability of different formats and types. As shown in figure 2.
Construct core metadata ontology (CM), with the consideration of general metadata specification such as DC and DCTERMS, the paper divides resources into two categories, namely, Document and Collection, Document resource is divided into single resource and collection resource, among which single resource consists of books, literatures, papers and web pages; while collection resources consists of collection of compilation books, multi-volume books, journals and conference proceedings. The text set is general term of websites, magazines, newspapers and other related documents. CM through dcterms: is Part Of and dcterms: has Part to refer the relationship between the collection and the monomer in the literature. CM is a common ontology shared by digital library. It can extend the original CM by adding new attributes and definitions, and generate specialized metadata ontology for a specific document resource.

On the basis of construction of the core metadata ontology, further constructing the corresponding linked data. The text constructs linked data by using RDF format semantic metadata, and the RDF data of different data sets are correlated. At the same time, the internal resources of the single library are set to processing open data association with external resources so that the digital library resources can be more easily browsed and retrieved by the visitors.

2.2 Evaluation of model system performance

In order to evaluate the system performance of hierarchical organization framework model of digital library, a contrastive analysis is made between the constructed model and a traditional model. Sending a query request to the two models every 30min to and collecting the search response time of the models. In order to improve the accuracy of the calculation, the results of each time period were counted 10 times and averaged. The retrieval response time of the two models is shown in Figure 3.

![Figure 3. The curves of retrieval response time with two models](image)

From the figure we can see, the retrieval response time shows obvious time characteristics, that is, from 1 am to 10 am, search retrieval performance is better; and from 11 am to 12 pm the retrieval query performance is worse. As a whole, the retrieval time of the linked data organization framework model of digital library proposed in this paper is much smaller than that of the traditional model, and the retrieval response time
is reduced by 56.3%. According to the above analysis, it can be seen that the establishment of the linked data organization frame model of the digital library can greatly increase the interconnection of distributed heterogeneous information and make the data source more easily crawled by the search engine.

3. CLUSTERING ANALYSIS OF IMPROVED LDA MODEL

3.1 TC-LDA theory

According to the above analysis, it can be seen that the establishment of the linked data organization frame model of digital library can greatly increase the interconnection of distributed heterogeneous information, and make the data source more easily crawled by the search engine. Because there are many drawbacks of the traditional document clustering methods, this section proposes an improved LDA method cluster the shared subject files in the same category to improve the existing shortcomings of the easy-to-shadow text catalogs and easy-to-distortions of the digital library, in this way retrieval of document information can be more accurate and convenient.

At present, LDA model is the widely used model in the digital library and text-related areas of information retrieval and document classification. Characteristics of LDA-based generated model, which can better solve the over-fitting problem of the similar text and semantic analysis. The relations between text and semantics are fitted by the hidden variable $Z_m$ in the model, so that the two satisfy the Dirichlet distribution, and distribution feature of the subject in the target text and the related key word are estimated to be obtained through the probability distribution of $Z_m$. The subject similarity function of any two texts in the digital library is:

$$C_m(d_i,d_j) = \begin{cases} 
1 & n \in [1,2,...,K], \exists \{p(z=n)|d_i \& d_j > T \\
0 & \text{else}
\end{cases}$$

(1)

$K$ is the subject type; $T$ is the determine threshold; $z$ is the subject; $n$ is the probability that the subject is given in the text. When $C_m=1$, considered two text are cluster, otherwise, they are non-clustering text. The similarity of the text subject can be calculated by subject probability.

Carrying our improvement on original LDA model, the text catalog and the text are merged and modeled, namely TC-LDA model. Gibbs sampling is used to obtain the parameter estimation, the formula is as follows:

$$\delta_{t_{i}} = \frac{n_{t_{i}} + \alpha}{n_{t_{i}} + K \alpha}$$

(2)

$$\delta_{c_{i}} = \frac{n_{c_{i}} + \alpha}{n_{c_{i}} + L \alpha}$$

(3)

$$\tau_{w_{i}} = \frac{n_{w_{i}} + \beta}{n_{w_{i}} + N \beta}$$

(4)

$$\xi_{w_{i}} = \frac{n_{w_{i}} + \beta}{n_{w_{i}} + M \beta}$$

(5)
In the formula, $\delta_t$, $\delta_c$, $\tau_w$ and $\zeta_w$ are parameters approximate estimation function; $n^d_{it}$ and $n^d_{ic}$ denote the number of vocabularies allocated to the subject $z_t$ and $z_c$ in the $i$-th text and catalog information respectively; $n^d_{it}$ and $n^d_{ic}$ are the number of vocabularies in the $i$-th text and catalog information respectively; $n_{wt}$ and $n_{wc}$ are the number of times that vocabularies assigned to the subject $z_t$, $z_c$; $n^1_{it}$ and $n^1_{ic}$ are the total number of vocabularies that assigned to the subject $z_t$, $z_c$; $\alpha$ and $\beta$ are pending parameters of the Dirichlet distribution.

The subject similarity degree of the TC-LDA model is calculated based on the probabilities calculation of $p(z_t | d_i)$ and $p(z_c | d_i)$, and the aggregation results $P_{O1}$ and $P_{O2}$ of the text and catalog information are calculated according to formula 2 to 5. After combination, obtain $R = P_{O1} \cup P_{O2}$.

### 3.2 Analysis of instance

Select two types of text libraries for instance validation, the first class is a general document class, which has 10 categories of which includes environment, medicine, finance, art, computer and so on. The second class is the text library of digital books, and the source files are selected from the network, including economic, military, health, diet, education and others total eight categories. The two types of text libraries are preprocessed to extract the catalog information, and delete the vocabulary with frequency less than 7 times appearance in the text, and the noun and verb words are extracted. The processed text used as a test set for clustering calculation, and the improved clustering algorithm is compared with the traditional clustering algorithm. In LDA and TC-LDA, set parameters $\alpha = 50/K$, $\beta = 0.01$; set Gibbs sampling iteration times as 100; the best subject number $K$ is calculated as shown in formula 6.

\[
AvD = \frac{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \text{Dis}(z_i, z_j)}{K \cdot (K - 1)/2}
\]  

$AvD$ is the average similarity of $z_i$ and $z_j$; $\text{Dis}(z_i, z_j)$ is the distance between two subjects. According to formula 6, when the number of subjects $K$ is set between 180-250, $AvD$ has the maximum value, the similarity of the two subjects is the smallest, that is, the retrieval confusion rate can be reduced to minimum, therefore, the value of $K$ may be taken as 200.

The LDA method, TC-LDA method and three traditional text clustering methods (Kmeans, FCM and AP) were compared to calculate the retrieval accuracy of the five methods under different textual quantities. The results are shown in Figure 4, because each search results there is a certain degree of fluctuation, so for each algorithm are calculated 5 times and take its average.
As can be seen from the figure 4, On the whole, test accuracy of LDA and TC-LDA is significantly higher than that of the three traditional algorithms. The LDA algorithm namely TC-LDA algorithm has the highest accuracy, followed by the LDA algorithm, and then the AP algorithm, the FCM algorithm has the lowest relative accuracy. The stability of TC-LDA algorithm is the best, because the nature of LDA algorithm is based on the clustering analysis of subject model, which satisfies the essential characteristics of the shared subject documents of the same category. The TC-LDA algorithm combined with the catalog information and text information for subject clustering modeling, compared with the LDA algorithm can better reflect the accuracy and robustness of its retrieval.

Figure 4 The test precision of different document number with different methods

Figure 5 shows the time consumption by the five algorithms for text clustering analysis. In order to calculate the precision, each algorithm is still calculated 5 times and takes its average. It can be seen that TC-LDA algorithm and LDA algorithm take a relatively long time in clustering analysis compared with traditional clustering methods. The calculation time of TC-LDA is 63.75s and that of LDA is 61.78s; while in the traditional algorithm, AP algorithm takes 51.82s, is the least calculation time consuming. The time consumption of Kmeans algorithm and FCM algorithm is basically the same. From the above analysis, we can see that the analysis time consumption of the five algorithms have not big different, and it is impossible to obtain which algorithm has the obvious advantage in time consumption. Because TC-LDA calculation method needs to set the parameter value in advance, thus it does not affect the cluster analysis time. Therefore, the improved LDA text subject clustering algorithm has more advantages than the traditional text clustering method, from the analysis of the accuracy and time consumption of clustering analysis.
4. CONCLUSION

(1) This paper proposes a hierarchical organization framework model of heterogeneous data integration and resource utilization of digital library. The model is divided into metadata layer, ontology layer, link data layer and application layer from top to bottom. Using RDF format semantic metadata constructs the corresponding linked data. The experimental results show that the retrieval time of the proposed linked data organization framework model of digital library is much smaller than that of the traditional model, and the retrieval response time is reduced by 56.3%. According to the above analysis, it can be seen that the establishment of the linked data organization framework model of digital library can greatly increase the interconnection of distributed heterogeneous information and make the data source more easily crawled by search engines.

(2) TC-LDA algorithm has the highest retrieval accuracy, followed by LDA algorithm, the retrieval accuracy of traditional text clustering analysis method is obviously lower than that of subject model text clustering algorithm. The essence of LDA algorithm is based on the clustering analysis of subject model, which satisfies the essential characteristics of the shared subject documents of the same category. The TC-LDA algorithm combines the catalog information and the text information to subject clustering modeling. Compared with the LDA algorithm, the TC-LDA algorithm can reflect the accuracy and robustness of its retrieval, indicates the stability of the TC-LDA algorithm is the best. The time consumption of the five algorithms is without big different, it is hard to tell which algorithm has the advantage in time consumption. For consideration of clustering accuracy and time consumption, the improved LDA text subject clustering algorithm has more advantages than the traditional text clustering method.

ACKNOWLEDGMENTS

This work was supported by the Heilongjiang province Philosophy and social science planning project (grant number 15TQE02), and Heilongjiang province Art science planning project (grant number 2016C043).
REFERENCES


