Journal of Applied Science and Engineering Innovation, Vol.6, No.1, 2019, pp.26-30
ISSN (Print): 2331-9062
ISSN (Online): 2331-9070

Application of Individualized Service System for Scientific and Technical Literature In Colleges and Universities

Jiang Xiao-Ling, Zhang Hui, Xu JiaMing

School of Computer and Software Engineering Huaiyin Institute of Technology, Huaian 223001, China.

Abstract: In order to solve the issue of information overload caused by massive scientific and technical literature, the accuracy rate of pushing the literature to the users is improved. A personalized service system for teachers and students based on Word2Vec is proposed. The preparation of reptile digging knowledge, the Chinese patent database technology literature and the use of Lucene-based full-text retrieval program for data retrieval. Through the word vector model, the title of the literature is abstracted, and the similarity between the full word segmentation and the user interest ontology is used to extend the user interest preference word pocket model, which solves the problem of system data sparseness and cold start. On the basis of 1.2 million scientific and technological literature achieving a good effect. The experiment has obtained the recommended accuracy rate of 82.3% of the scientific literature, and provided the reference for the science and technology literature recommendation algorithm. The system teachers' and students' demand for the efficiency of literature retrieval.

Keywords High-speed Ethernet, Electromagnetic performance, Crosstalk

INTRODUCTION

According to the latest information, wanfang paper database can be retrieved from the literature including journals, dissertations, conferences, foreign literature database. The problem of "information overload" caused by massive data makes information searchers spend a lot of time and energy in searching for information that is valuable to them. Personalized recommendation service system is an effective way to solve the problem of "information overload". Literature personalized service of science and technology are the key technologies using precise fetching user use the system behavior, according to some users browsing history or tag project in extraction, build document for each user preference model is used to predict might like literature, at the same time can make preference model is adaptive, due to the time advance user interest of drift can update to the user of the system model. The individualized service technology of researching scientific and technological literature can improve the efficiency of researchers in searching literature and make breakthrough in scientific research.

RECOMMENDATION ALGORITHM FOR SCIENTIFIC LITERATURE

This research adopts document preprocessing algorithm, feature extraction algorithm of scientific and technological literature, user preference model establishment algorithm, personalized recommendation model of scientific and technological literature. The following focuses on the personalized recommendation model of scientific and technological literature.

The literature recommendation module is the most core module of the literature personalized service system. This module firstly extracts the literature browsed by the user, and then quantifies the document according to the user's preference word bag model. At the same time, according to the time the user browses each literature, it serves as the degree of the user's preference for literature. Under certain threshold conditions, literatures are valid data records, while too short or too long are regarded as invalid data. Users who browse literatures for too short a time are generally not interested in them, and those who stay for too long may forget to quit due to page browsing.

In the module, the literatures that have been vectorized are put into the classifier for training, and the newly added literatures are vectorized according to the users' preference word bags after preprocessing, and then the vectorized literatures are put into the classifier for analysis, and the literatures to be recommended are added to the list of literatures after meeting certain threshold conditions.

Finally, the literatures read by users with similar reading interests to the current users in the database are added to the list to be recommended. At the same time, the users who are most similar to the current users' interest word bag model are calculated in the database, and the literatures browsed by these users are added into the alternative literature set.

Finally, according to the sorting algorithm, the most suitable literature list for the current user is

presented to the user in order according to the degree of user interest.

Step 1. Input the set D of documents of interest to the user, N documents in total;

Step 2. Define the random variable I, I =1;

Step 3. If m<=N, execute step 4; otherwise, execute step 10;

Step 4: word segmentation is performed for all word items in document Di, and the character set CDi={w1,w2... Where n is the total number of word items after the word segmentation of CDi;

Step 5. Check CDi={w1,w1... Filter all the words in wn} to get the new character set WDi={w1,w1... ,wm}, where m is the total number of new words after the word "stop" is removed

Step 6. Use the regular pair CDi={w1,w1... All words in wn} are matched to remove illegal characters, and a new character set NDi={w1,w1... ,wp}, where p is the total number of new words after removing illegal characters;

Step 7: NDi={w1,w1... , all character encodings in wp} are transferred to utf-8 encoding uniformly, and UNDi={w1,w1... ,wp}, and the manual random verification of the extraction effect to adjust the threshold;

Step 8. According to the professional field of the document, get ZDi={w1,w1... ,ws}, where s is the total number of word items;

Step 9. LD={{w1,w2... }, {w1, w2,... },... , {w1, w2,... }};

Step 10. Define the random variable I, I =1;

Step 11. If I <=N, execute step 11; otherwise, execute step 13.

Step 12: vectorize the first document according to the user's reading preference word bag model;

Step 13. Get vector set DC={d1,d2... ,dN}, where N is the total number of vectors;

Step 14: train the preference model with classifier;

Step 15. Input R of other document sets in the database, A document in total;

Step 16. Define the random variable j,j=1;

Step 17. If m<=A, execute step 18; otherwise, execute step 27;

Step 18: word segmentation is performed on all the word items in document Rj according to the word segmentation dictionary, and the character set CRj={w1,w1... Where n is the total number of word items after the word segmentation of CRj;

Step 18. Check CRj={w1,w1... All the words in wn} are filtered to get the new character set WRj={w1,w1... ,wm}, where m is the total number of new words after removing the stopped word;

Step 19. Use the regular pair CRj={w1,w1... All words in wn} are matched to remove illegal characters, and a new character set NRj={w1,w1... ,wp}, where p is the total number of new words after removing illegal characters;

Step 20: NRj={w1,w1... , all character encodings in wp} are transferred to utf-8 encoding uniformly, and UNRj={w1,w1... Wp};

Step 21:According to the professional field of the document, get rid of the non-professional words and get ZRj={w1,w1... ,ws}, where s is the total number of word items;

Step 22. Vectorize the document Ri according to the user interest word bag model, and get the vector Ci={a1,a2... , an};

Step 23: get the document set QL that users are interested in according to the trained classifier model;

Step 24. Generate part of literature SL to be filtered according to the selection of similar users;

Step 25: sort the documents in QL and SL according to weight to get the final recommended document content;

Step 26. According to the item sorting algorithm, the final recommended literature items RC={item1, item2... ,itemx}, where x is the total number of recommended literatures;

Step 27. According to the item sorting algorithm, the final recommended literature items RC={item1, item2... ,itemx}, where x is the total number of recommended literatures;

THE SYSTEM DESIGN

The overall architecture

Figure1 shows the overall structure of the scientific literature service system for teachers and students in colleges and universities. The analysis layer mainly consists of four modules: literature preprocessing module, literature extraction module, user preference model building module and user push module. The extraction module is responsible for extracting document features. The preference building module builds a preference model by analyzing the documents that the mobile phone users have browsed in their history. The push module is responsible for presenting the final recommendation results of the system to users.

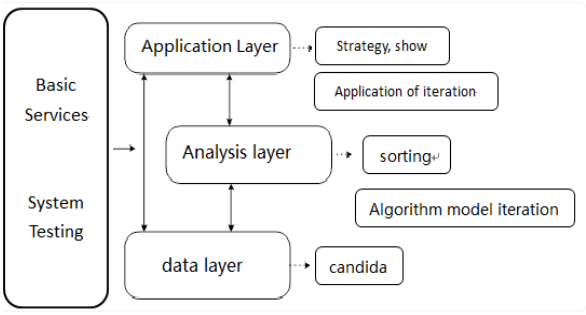

Figure1 The overall framework of scientific literature service system for teachers and students

System overall architecture is divided into three layers, respectively for the analysis of the application layer, layer, data layer, application layer is responsible for the system as a whole business process control, according to a certain strategy

iterative application function, the results presented to the user, this system through the micro letter public number provide recommendation service to users, the user specified by focusing on micro letter after the public order, input needs to retrieve literature subject keywords, system application module filter data from the database through the analysis of the algorithm, the algorithm analysis results by micro public letter presented to the user.

The data layer is responsible for storing all the data of scientific and technological literature in the literature recommendation system, keeping it updated, and adding the latest literature to the library dynamically through crawlers.

In the analysis layer, the candidate set literatures are sorted by the iterative algorithm model, and the literatures that users are most interested in are presented to users in order. Among them, the analysis layer consists of three modules: literature preprocessing module, user reading preference building module and literature recommendation module.

1) literature preprocessing module: this module firstly divides the full text of the literature extracted from the Web by crawlers, removes stop words and codes them uniformly. The literature was standardized according to the data input requirements of the analysis module.

2) user preferences to establish module: user preferences set up a module to collect all users to read the literature records, literature reading time and search keywords, as the user to read the literature of history keywords reading preference ontology, read through the history of literature title, keywords, word processing, the deep learning model is used to calculate each word after word similarity with the words in the user preference ontology, words give different weights according to the browsing time, and will meet a certain threshold condition words words are added to the user preferences in the collection of ontology, to achieve the user preference collection development, mining the user's interest in reading.

3) literature recommendation module

The literature recommendation module is the core module of the recommendation system, which is responsible for matching the user preference model with the features extracted from the literature, and generating part of candidate document sets. At the same time, users who are similar to the current user preferences are retrieved from the reading preferences of other users in the database, part of the historical browsing literature records of these users are extracted and added into the candidate literature set of users, and the most interesting literature result set is selected according to the sorting algorithm.

Module design

Text preprocessing is the most basic and important step in the field of natural language processing. The preprocessing effect directly affects the effect of post-processing. The module structure is shown in figure 2 below.

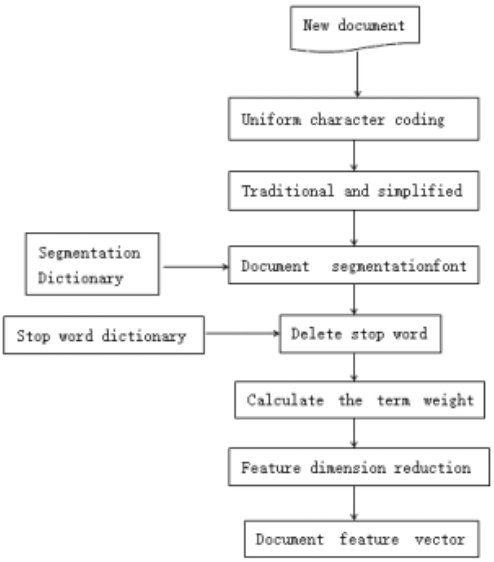

Figure2 Text preprocessing flowchart

First, the text characters are converted to utf-8 encoding using uniform characters, and then the Python OpenCC is called to convert the full-text traditional characters into simplified characters. Then, the text segmentation processing, text participle is using a mechanical model of word segmentation, according to all the words in the dictionary will be treated as text participle, this system adopts the jieba word segmentation module to word processing document, in addition to word segmentation results stop words, then calculate weights of each term and sorting, eliminate weight is lower than the threshold value of word document feature vector model.

The document depth representation model is based on a large corpus. This system USES a three-layer neural network for training model. The corpus is a Chinese corpus published by wikipedia, with the size of 1.2gb.

1) since the corpus is a raw XML data, it needs to process the data;

2) convert the traditional Chinese into simplified Chinese in the processed XML file;

3) calculate the synonyms of the word "neural network" for the word vectors trained by the neural network.

Users' reading preferences can be accurately acquired through the user preference model system, so as to meet the different reading needs of various users.

Step 1: input the document collection that users are interested in;

Step 2: get the title, key words and abstract of the literature;

Step 3: collect the browsing records of users, add keywords of all the literature that users are interested in into the ontology collection reflecting user preferences, and assign keywords with different

weights according to the browsing time of users;

Step 4: word segmentation of the title, key words and abstract of the historical browsing documents according to the dictionary;

Step 5: according to the stop word list to remove the stop word in the title, keywords, abstract segmentation results;

Step 6: Use the word vector model to calculate the most similar five words of each element in the user preference set and add them to the temporary user preference set;

Step 7: Calculate the similarity between the words obtained from the title, keywords and abstract and each element in the temporary user preference set.

Step 8:Add the word to the user's final preference model according to the final weight pair obtained for each word; otherwise, remove the word;

SYSTEM DEPLOYMENT AND TESTING

System deployment

The system is deployed on the Windows Server. The operating system version used is Windows Server 2008R2. The specific installed software is as follows:

Elasticsearch, full-text search engine, version 2.4.3

Django Web development framework, version 1.11.1;

MySQL database, community edition, version 5.7;

Python development environment, version number, 3.4.1;

JDK/JRE Java development and runtime environment, version 1.8;

PHP development environment, version number, 7.1.1

Elasticsearch Retrieve the test

Through the use of crawlers, import the literature in the paper database from the relational database MySQL into Elasticsearch for full-text retrieval. With a total data of 300,000 pieces of data, the required time to retrieve qualified literature entries based on keywords is less than 1 second.

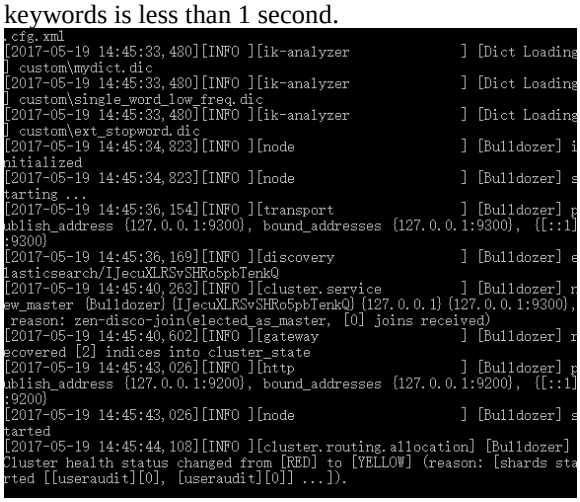

Figure 3 Start the Elasticsearch

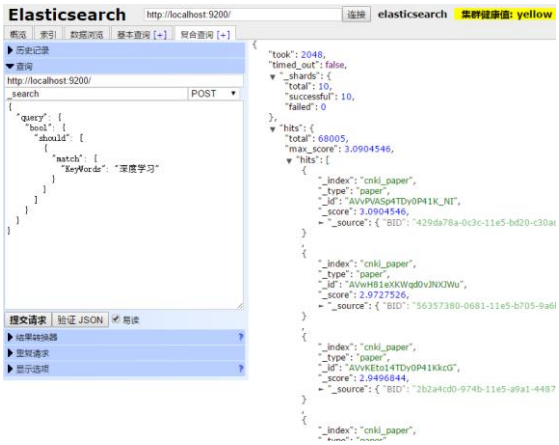

Figure 4 Use Elasticsearch to retrieve

Figure 5 FIG Start the Web service

CONCLUSIONS

In order to improve the efficiency of scientific literature retrieval and the accuracy of pushing scientific literature to teachers and students, a personalized recommendation algorithm based on Word2Vec is designed and implemented. In the design and implementation of the algorithm, the paper respectively studies document preprocessing, document depth representation, user preference model building module and personalized recommendation module of scientific and technological literature. Document preprocessing is a process of document format standardization, which can be directly used to analyze the input of the classifier and improve the efficiency of the system analysis module. The document depth representation model calculates the word vector of the words through the algorithm module and is used to calculate the similarity between words, which solves the problem of low analysis accuracy caused by semantics to some extent. Because with sparse characteristic literature database, the user preference modeling module through the user browsing history document of title, keywords, interest points after the extraction processing to produce alternative word, to read the literature of history keywords as a user reading interest ontology, after processing by calculation of alternative term interest points and the

degree of similarity between user interest ontology vocabulary, to meet a certain threshold of key words added to the user in the reading interest model. By expanding the user interest model, the problem of data sparsity is solved to a certain extent, the accuracy of classifier is improved and the problem of data sparsity is solved to a certain extent. For the retrieval and storage of massive data, this system adopts the distributed database architecture and adopts the full-text retrieval scheme based on Lucene to improve the efficiency of mass data retrieval. In the process of research, many other improvements have been found. For example, how to further improve the recall rate of literature recommendation through the optimized literature extraction algorithm on the basis of ensuring the accuracy of literature recommendation is one of the future research directions.

REFERENCES

Ren xingyi, Song meina, Song junde. Recommendation of interest points based on user check-in behavior [J]. Journal of computer science, 2017(1):28-51.

Kumara B T G S, Paik I, Siriweera T H A S, et al. Cluster-Based Web Service Recommendation, Proceedings of IEEE International Conference on Services Computing. IEEE, 2016:348-355.

Jiang xiaoling (1981-), female (han nationality), native of huai 'an, jiangsu province, master, lecturer, main research fields are image processing, data mining, etc.