The Research of TF-IDF Recommendation Algorithm of Colleges and Universities' Patent System

He Liu [1,a] Ping Li [1,2,b], Chenxi Li [1,c]

1. Jilin Agricultural University, Changchun, 130118,China

2. Changchun Institute of Technology,Changchun,130600,China

[a]liuheliaoshi_7978@163.com, [b]307360899@qq.com, [c]455169130@qq.com

Keywords: Recommendation algorithm; Word segmentation system; Text recommendation algorithm

Abstracts. The users in search of patent achievements, its demand is also often vague and broad, in order to meet the user's search request, and at the same time in order to improve the conversion efficiency of patent information in colleges and universities. The research is based on NIPIR which is word segmentation system uses that to separate TF-IDF (term frequency–inverse document frequency) from patent's introduction. Then, use the same method to get TF (term frequency) of search log of user, and build data of user' preferences. According to the value of user's preferences and patent's TF-IDF, the system active send user information. The system that translates potential into actual demand increase conversion efficiency of scientific research of college. That realizes friendly docking between scientific research institutions and application institutions，improve technological innovation and promotion ability.

Introduction

The transformation of scientific and technological achievements in colleges and universities is a combination of technology and economy, it helps to translate scientific and technological achievements into actual productivity and promote national economic construction and social development and it is of great significance in improving scientific and technological innovation and promotion capabilities, promoting the capability of independent innovation of agricultural science and technology, consolidating and improving comprehensive production capacity of agricultural, etc[1]. However, it is difficult for enterprises to excavate and apply scientific research results of colleges and universities and to find a breakthrough through upgrading certain achievements and innovative technologies, so building the promote system of scientific research findings at colleges and universities is particularly important. For one thing, it provides the condition of practice value for the achievement technology in universities, for another, it promotes enterprise development and translate the achievements of scientific research into real productivity to improve the efficiency .Accordingly, it is necessary for us to study and discuss the scientific research achievements in the recommender system and research how to use better recommendation algorithm to achieve achievement transformation technology, which helps to transfer these technologies to the hands of those in need more conveniently and quickly.

Related Research Method

The requirements of users are usually vague and uncertain when browsing the site information. Therefore, if we can push information which meet user's vague demand from massive search information proactively, it helps to translate the potential requirements into actual requirements. In addition, it helps to improve the transformation efficiency of scientific research achievements in colleges and universities, which achieves to docking scientific research institution and application organization harmoniously.

Personalized recommendation, first proposed in early 1990s by Resnick[2], was used in the field of collaborative filtering recommendation research on news. It is based on the user's interest, essential information, operation behavior and a series of related information, to analyze users' preferences, and build a user or user group model, the ultimate aim is to push the information of interest to the users. All these practices reduce the time and vigor they waste on browsing the information without any purpose. At the same time, they can present the information of their publisher proactively for the first time, which has greatly changed the supply of Internet information services model. At present, there are various of recommendation system based on recommendation algorithm, which has been mainly used in e-commerce, entertainment, news and other aspects, such as Alibaba, Amazon, Douban and other network merchants, all of those are very successful applications.

Frequently-used recommendation algorithm. (1)Collaborative Filtering Recommendation Algorithm: Recommending the content of interest to the users can be recommended to the users by finding other users who are similar to the user's preferences. Collaborative filtering method is divided into memory-based and model-based approaches.

(2) Content-based recommendation algorithm: Constructing users preference documents based on the user's operating behavior, such as comments, sharing, collection and other operations, then perform similarity calculations with the proposed algorithm. Finally, recommend the most similar items to the users. This recommendation algorithm is widely used in the field of text (news, web pages, blogs), because it is relatively easy to extract the characteristics in the field of text.

(3) Graph-structure based recommendation algorithm: The method is a dynamic network resource allocation process and it is a recommendation algorithm based on structural analysis of bipartite graph. The bipartite graph is a model representing users and projects into vertex, user evaluation of the project into edge, then modeling with the vertex and the edge.

(4) Hybrid recommendation algorithm: To address the limitations of a single recommendation algorithm application itself, so the hybrid recommendation algorithm came into being, it can improve the effects of recommendation effectively.

TF-IDF text recommendation algorithm. TF-IDF is the main statistical method to calculate the weight of keywords in text, and a frequently-used weighted technology for information retrieval and exploration. Its effect is achieved by assessing the importance of a word for a file set or one document in a corpus. The importance of the word is represented by the number of times it appears in the documents. The more times it appears in the documents, the more important it is. The more times the word appears in the corpus, the less important it is.

The keywords are typically used to represent the characteristics of the user preference documents and the recommended project document, then we can use the TF-IDF (term frequency-inverse documents frequency) method to determine the weight of each feature.

The idea of TF-IDF method is embodied in two aspects: on the one hand, the number of occurrences in a document is directly proportional to its importance,the more the number of keywords appear in the documents, the higher the importance of the keyword to the documents, then the keyword can indicate the semantics of the documents; On the other hand, the number of occurrences of a keyword in a document set is inversely proportional to its importance.,the more frequent occurrence of the same keyword in different documents, the less the value of the keyword to each documents, thus, the lower the semantic possibility of the documents. Based on the above two points, we propose the method of setting the feature weight of TF-IDF. We supposed that documents set to be DS contains N documents d, and documents in d which contains keyword k_i to be n_i , and the number of k_i appears in documents d_j to be f_{ij} , and the word frequency of k_i appears in documents d_j to be TF_{ij} . TF_{ij} is defined as :

$$TF_{ij} = \frac{f_{ij}}{max_z f_{zj}} \tag{1}$$

In which z represents the keywords that appear in the documents d_j, and the inverse documents frequency of k_i appears in documents DS to be $IDF_{i.}$ IDF_i is defined as :

$$IDF_i=\log\frac{N}{n_i} \quad (2)$$

And calculate the weight of each word in the documents to be w_{ij}:

$$w_{ij}=TF_{ij}*IDF_i \quad (3)$$

Then calculate the word frequency of users search history adopted to the user's search history record, and then form a k-dimensional vector d_c=(w_{1c} ,w_{2c} , …,w_{kc}), then calculate the angle cosine between k-dimensional vector d_c and w_{ij}, which is calculated to form the similarity , then use the 0 to patch the default value of each word does not exist for user search terms, follow the users search terms in the main order. Thus deriving the k-dimensional vector of the article d_j=(w_{1j} , w_{2j} , ⋯, w_{kj}), and the similarity is calculated to be:

$$sim(c, d_j)=\cos(d_c, d_j)=\frac{\sum_{i=1}^{k} w_{ic}w_{ij}}{\sqrt{\sum_{i=1}^{k} w_{ic}^2}\sqrt{\sum_{i=1}^{k} w_{ij}^2}} \quad (4)$$

At last, we sort the similar values in order from high to low, which leads to a preference ranking

Research of TF-IDF text recommendation algorithm in University Patent Achievement Recommendation System

Brief introduction of ICTCLAS. ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System), is the reaserch result that the Chinese Academy of Sciences has accumulated many years to research and accumulate, The main functions include Chinese words segmentation, part-of-speech tagging, named entity recognition, unregistered words recognition, supporting of users dictionaries at the same time.

In recent years, 973 expert group evaluation results show that it has the accuracy of 98.45% in word segmentation. This text uses the analysis system to complete the word segmentation work.

Design of Data Structure. The data tables in the system's data model need to be created to be three: A search history table that describes the user's search behavior and the search word record table, the word segmentation table which complete the word frequency calculation, and patent information table to complete the entry of checklist, the table structure is shown in the fig 1:

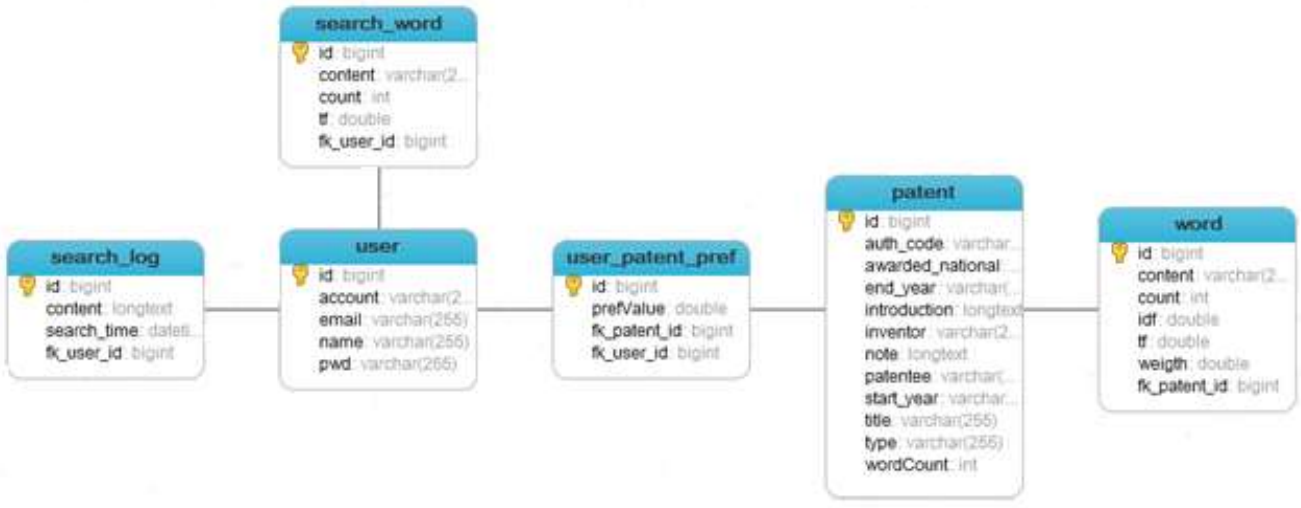

Figure 1. the structure between each datasheet

Implementation of recommendation algorithm

The system uses the Java language development, relies on the NIPIR word segmentation system to realize the participle of patent word and word frequency calculation, uses the Spring framework and the MVC design pattern to manage the system resources, and completes the development with the IDEA2017 integrated environment tool. The implementation process is shown as follows:

First enter the sample data in NIPIR word segmentation system, complete the word egmentation operation of the document, at the same time calculate the word frequency. Each sample must repeat this operation, the code example is shown as Fig 2:

```
Patent patent1 = new Patent.Builder().setTitle("一种加有大豆异黄酮的护肤剂及其制备方法 ").setType("发明").setAuthCode("CN102113977A").
        setInventor("徐慧").setAwardedNational("中国").setStartYear("2010").setEndYear("2031").
        setIntroduction("本发明公开了一种加有大豆异黄酮的护肤剂及其制备方法。以重量份计,组成为:大豆异黄酮粉:5~8,鲸蜡:1~3,蜂蜡:10~15,
        build();
patent1.setWords(nlpirUtils.getWordsList(patent1.getIntroduction()));//设置文档的words
patent1.setWordCount();//计算文档词数 并 计算 词频
```

Figure 2. the entry code of sample data

Then store the data in the document that completes the word segmentation. Since the entire document set needs to be calculated while calculating the inverse frequency, therefore, the calcul ation process needs to call the service layer code to operate, and calculate the weight of each word frequency at the same time. The result is shown in Fig 3 below:

id	content	count	idf	tf	weigth	patentId
1	搅拌	4	0.3010299956639812	0.03508771929824561	0.010562455988209866	1
2	将	3	0.6020599913279624	0.02631578947368421	0.0158436839823148	1
3	对	3	0.6020599913279624	0.02631578947368421	0.0158436839823148	1
4	大豆	3	0.3010299956639812	0.02631578947368421	0.0079218419911574	1
5	异	3	0.6020599913279624	0.02631578947368421	0.0158436839823148	1
6	黄	3	0.6020599913279624	0.02631578947368421	0.0158436839823148	1
7	酮	3	0.6020599913279624	0.02631578947368421	0.0158436839823148	1
8	的	3	0	0.02631578947368421	0	1
9	均匀	3	0.3010299956639812	0.02631578947368421	0.0079218419911574	1
10	发明	2	0	0.017543859649122806	0	1
11	石蜡	2	0.6020599913279624	0.017543859649122806	0.010562455988209866	1
12	醇	2	0.6020599913279624	0.017543859649122806	0.010562455988209866	1
13	硼砂	2	0.6020599913279624	0.017543859649122806	0.010562455988209866	1

Figure 3. The example data of word frequency decomposition of soybean patent information

The third step, simulate the search behavior of a single user, generate search data, then calculate the user's search word frequency.

The fourth step, calculate single-user preferences, if there’s not any same word between search keyword and its patent introduction while the calculating the similarity, the denominator is calculated as 0.0, and we will get NaN (not a number) error results, then the fault tolerance will be judged.

The fifth step, achieve the information push operation according to the ranking of calculation the user preference value, the implementation code and result are shown in Fig 4:

```
/*输出基于TD-IDF的用户搜索历史偏好排名 *未实现分页*/
@Test
public void outUpf() {
    User user = new User.Builder().setAccount("12345678").setPassword("123456").build();
    user = userService.login(user);
    for (UserPatentPref upf : user.getUserPatentPrefs()) {
        System.out.println("patent->content:"+
                upf.getPatent().getIntroduction()+
                "~user->account:"+
                user.getAccount()+"~upf->prefValue:"+
                upf.getPrefValue());
    }
}
```

```
Rank: 1, Patent name: a skin-protecting agent added with soybean isoflavone and its preparation method

Rank: 2, Patent name: Fermented bean curd with ginger juice and its preparation method
Rank: 3, Patent name: Preparation method of natto active capsule
Rank:4, Patent name: Preparation method of seasoning soybean milk
```

Figure 4 The example of output preference code and result

Conclusion

The research of user's patent search preference includes five parts: System data entry, system data calculation, data acquisition, acquisition data calculation, and preference ranking. Each section deals with three different aspects of data, operations, and user interfaces.

The raw data is generated by inputting data through user interface. The original data is processed by the NLPIR word segmentation system for word segmentation data processing, and then calculate TF and IDF with the results. Calculate the formation of patent frequency data and patent inverse frequency data, thus forming the weighting data of patent word segmentation Patented data. And collect users history search data for NLPIR word segmentation system classification, then the TF is calculated to form a single user search history word frequency (Preference weight). Then calculate the angle cosine between Patent segmentation weight data and Single-user search history word frequency, which is calculated to form the similarity. Finally make the preference ranking in accordance with the level of similarity, to form the user recommendations

After debugging, output the results in descending order according to prefValue. The most forward it ranking, the highest similarity it is, and then push its information to users. The experimental results show that it has certain feasibility and accuracy in pushing patent information by TF-IDF text-based recommendation algorithm, and helping users find valuable information quickly to some extent, which will play a certain role in the transformation of scientific research achievements in Colleges and Universities.

Acknowledgements

Jilin Provincial Science and Technology Department NaturalFund Project (201701051JC)

References

[1] Chen Dianfan. Design and implementation of scientific research management system of Zhuhai Institute of Jilin University based on UML modeling[D]. Jilin University, 2015

[2] Huang Zhenhua, Zhang Jiawen, Tian Chunqi, Sun Shengli, Xiang Yang. Survey on learning-to-rank based recommendation algorithms[J]. Journal of Software. 2016, 27(03):691-713. (2015-12-30)

[3] Yang Bo, Zhao Pengfei. Review of the art of recommendation algorithms [J]. Journal of Shanxi University (Natural Science Edition). 2011, 34(03):337-350.

[4] Zhao Chengling, Chen Zhihui, Huang Zhifang, Recommendation algorithm and application of

adaptive learning path [J]. China Educational Technology. 2015, (08):85-91
[5] Chen Jiemin, Tang Yong, Li Jianguo, Cai Yibin. Survey of personalized recommendation algorithm [J]. Journal of South China Normal University (Natural Science Edition). 2014, 46(05):8-15
[6] Zeng Siyan, Zhou Jin, Huang Guohua. A book recommendation algorithm based on term frequency-inverse document frequency and community partition[J]. Journal of Shaoyang University (Natural Science Edition). ,2017, 14(02):19-22+37
[7] Wu Sheng. Study and implementation of text recommendation system based on similarity of word sense [D]. University of Electronic Science and Technology of China. 2015
[8] Shen Yinan, Zhang Chao, Zhu Guofeng, Sun Dongbo. The existent problem, causes and solutions of the university's conversion of science and technology production [J]. Chinese University Technology Transfer. 2016, (03):8-11