Personalized Search Ranking Based on Semantic Tag

Lei Huang[1,a], Chanle Wu[1,2,b]

[1]Computer School of Wuhan University, Hubei, Wuhan 430072, China

[2]National Engineering Research Center for Multimedia Software, Wuhan, 430072, China

[a]llei_h@163.com [b]chanle.wu@gmail.com

Keywords: Semantic Tag, Personalized Search Ranking, Personalization, Search Engine

Abstract: The resource getting core of knowledge Service System is the search engine, but the most studies only put attention to improve efficiency, so as to mass resources retrieval results still allows the user to face "cognitive overload" problem when the user to use searcher to get knowledge, how to provide personalized search results become a research focus. This paper provide a new personalized search ranking method, which use semantic tag and user profile to personalized the search results. The experimental results indicate that the method is effective.

Introduction

The resource getting core of knowledge Service System is the search engine[1], the existing retrieval technology is mainly dependent on the encoding process, it includes classification model to describe information resources and full-text search to find the words in the text two categories. The applications include the classified directory-based search engine and full-text search engine[2]. Category-based search is manual processing and higher accuracy, suitable for browsing and navigation of the network of information resources. The realization of full text search is more convenient, which adapt to the need of the rapid growth of the amount of the network information resources automatic processing, it is the main mode of the network information retrieval. But the user retrieval always face the retrieval results of overload and low precision, user burden is heavy, the retrieval results and sort inconsistent, etc. problems. The main reason is the statistic matching of keywords is difficult to support the effective retrieval utilization of network information resources. So, the researchers to turn their attention to the mining of the meaning behind the words, to explore and realize the retrieval techniques and methods based on the concept matched. In the 1980s, the information retrieval international conference papers appear the discussion of semantic retrieval, but semantic retrieval research is always limited by the semantic information processing level of development. Along with the development of natural language processing and artificial intelligent technology, especially the rise and development of semantic web technology to promote the semantic retrieval research development rapidly. Although the concept of semantic retrieval is still no uniform definition, but different research had in common, which is based on the information resources of the semantic processing to realize efficient retrieval. In fact, it is because of the appearance and development of the semantic web to make the semantic research more clearly and quickly[3].However only improve efficiency is not enough, mass resources retrieval results still allows the user to face "cognitive overload" problem, how to provide personalized search results become a research focus.

SHOE search engines[4] are the one of the first semantic search engine based on form, it provides sophisticated WEB form, let the user to specify they query demand. However, this forms only suitable for those users who familiar with the system ontology and knowledge base, ordinary users is difficult to understand the meaning of these forms. Moreover, users is difficult to clearly express they want from their perspective. A typical example of search engine based on RDF query language is Corese[5]. Other include CS AKTiv Space [6] and Semantic Web portal, etc. These search engines provide a complex query language to support semantic data query. However, in order to use these search engine, end users have to be familiar with ontology and the query language. The search engine described in [6] and [7] is semantic keyword search engine. Through the related

Published by Atlantis Press, Paris, France.

analysis found that the mainstream of the search method by the use of available data and their semantic ontology can improve the performance of search engine, but is not suitable for ordinary users. The main problem is the burden of the knowledge. But the above studies rarely pay attention to personalization realization method, which makes the knowledge service system can not to present different results for different users.

Document Ranking

In the semantic web applications, on the one hand, personalized sort algorithm is according to the user's query semantic tagging to sort the documents; On the other hand, it is according to the user's relevant background to filter the content, so as to provide different search results for different users.

In document sorting, this paper use the concept vector space model, through the vector cosine angle to calculate the semantic similarity between document and query words, as show in formula(1):

$$Sim(D_i, Q) = \frac{\vec{d_i} \cdot \vec{q}}{|\vec{d_i}| \times |\vec{q}|} \quad (1)$$

$\vec{q} = (q_1, \cdots, q_m)$ is the relevant concept set through query processing, q_l express the weight of the lth concept; $\vec{d_i} = (d_{i,1}, \cdots, d_{i,m})$ express the concept set included in document d_i, $d_{i,m}$ is the weight of the mth concept in document i. Throught query extented, there are a problem of this processing method is the number of concept may be far beyond the number of the concept in the document. The reason is the document sets including polysemant, pronoun, etc. which cannot match the concept. So, this paper user the tag system to improve this problem.

2.2 Ranking Based on Semantic Tag

For the resource in pool, using the Symbiotic distribution feature to calculate the probability of concept belong to resource. Figure 1 is the relationship matrix A, which express the relationship between resource and tag, the tag is ontology concept, $a(i,\mathrm{j})$ is the tagging number of the resource i be tagged by j, such as resource 2 be tagged tag2 3th.

$$\begin{array}{c} \\ \mathrm{R1} \\ \mathrm{R2} \\ \cdots \\ \mathrm{Rm} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{C1} & \mathrm{C2} & \cdots & \mathrm{Cn} \end{array} \\ \begin{pmatrix} 1 & 0 & \cdots & 17 \\ 5 & 3 & \cdots & 7 \\ \cdots & \cdots & \cdots & \cdots \\ 8 & 2 & \cdots & 1 \end{pmatrix} \end{array}$$

Figure 1 resource-concept relationship matrix

A resource may correspond to a number of different concepts. When computing probability, to judge whether the user is already specifies the subordinate concept for the resource, if it is, you do not need to calculate the corresponding probability value, directly set to 1; Otherwise, it need to calculate the probability $\mathrm{P}(\mathrm{c_j}|\mathrm{r_i})$ of corresponding concepts, as is show in formula (2).

$$P(c_j|r_i) = \begin{cases} 1 & \text{User Assign} \\ \frac{a(i,j)}{n_i} & \text{Others} \end{cases} \quad (2)$$

$\mathrm{n_i}$ is the tagging total of tags which tagged the resource i. We can get a new matrix B after computing the $\mathrm{n_i}$, $b(i,\mathrm{j}) = P(c_j|r_i)$.

According to the resource-concept probability relationship matrix, Bringing it in the $\overrightarrow{\mathrm{d_i}} = (\mathrm{d_{i,1}}, \cdots, \mathrm{d_{i,m}})$ to get the new document vector $\overrightarrow{\mathrm{ed_i}} = (\mathrm{d_{i,1}} + \mathrm{P(1|i)}, \cdots, \mathrm{d_{i,m}} + \mathrm{P(m|i)})$, and then bringing it in formula (1) to get the formula (3).

$$Sim(D_i, Q) = \frac{\overrightarrow{ed_i} \cdot \vec{q}}{|\overrightarrow{ed_i}| \times |\vec{q}|} \quad (3)$$

To get the Top-N resources by formula (3), and then personalize these resources.

Resources Personalized Selection

① User similarity calculation, When fusing user-tag relationship with user similarity from rating matrix, we need to get a user similarity based on the user-tag relationship firstly. We represent the user-tag relationship in the form of a user-Gtag matrix according to section 2.2. If the user u_i used a tag which belong to $Gtag_j$, then the value of the $u_{i,j}$ plus 1, otherwise 0. Based on this user-Gtag matrix, we calculate a user similarity by adopting Pearson Correlation Coefficient, namely$SimUG(x,y)$. Next, when calculating the final user similarity between the user x and the user y through combining the $SimUG(x,y)$ with $SimUI(x,y)$.

$$Sim(x,y) = \alpha SimaUI(x,y) + (1-\alpha)SimUG(x,y) \quad (4)$$

The parameter α is used to adjust the weight of the SimUG and SimaUI , the bigger the α is, the rating matrix plays a more important role in the combined similarity. UI is user-item matrix, UG is user-Gtag matrix.

② To find the most similarity Top-N users, getting the Symbiotic distribution through formula (3), and to take the arithmetic average value $avg(R_i)$.

③ According to formula (5) to get the last score, and getting the Top-N resources recommend to the end user after ranking.

$$Rank(R_i) = t + d + avg(R_i) \quad (5)$$

t is the conformity of medium type, d is the conformity of difficulty level, the value of t and d is 0, 0.5 and 1 respectively, namely coincidence, partially coincidence, inconformity.

Experiment and Analysis

Experiment Data

The dataset contains 2982 users, 27563 tracks and 10023 tags.

Experiment Metrics

We adopted standard metrics in the area of information retrieval to evaluate our recommenders. During each round of cross-validation, we recommend and rank a set of potential tracks for each user. We then compare the predicted recommendation list with true preferences on tracks in the test set, and compute precision, recall, and F-measure scores.

Experiment Procedure and Results

We firstly evaluated the fusion approach via weighted-similarity. A user-based collaborative filtering recommender was run on the user-track rating matrix, resulting in the baseline represented by CF_{UI}. And then we tried to fuse user-tag relationship with the rating matrix. In the process of fusing, we used α to adjust the weight of rating matrix and user-tag relationship while computing user similarity. It achieves the peak when $\alpha = 0.56$, which means that the rating matrix contributes to 56% percent of the weight and the user-tag relationship contributes to 44% in calculating the user similarity. Specifically, results in the table 1, represented by CF_{UI+UG} , give the precision, recall and F- measure scores when $\alpha = 0.56$. It can be seen that it is better than the baseline, the improvement of F-measure achieved up to 7.33% when returning top 10 recommendations.

Top-n	Precision		Recall		F-measure	
	CF(UI)	CF(UI&UG)	CF(UI)	CF(UI&UG)	CF(UI)	CF(UI&UG)
5	0.256	0.259	0.068	0.072	0.107	0.113
10	0.231	0.245	0.113	0.122	0.152	0.163
15	0.201	0.211	0.135	0.145	0.162	0.172
20	0.164	0.172	0.172	0.181	0.168	0.176

Table 1 Precision、Recall、F-measure

Summary

This paper provide a new personalized search ranking method, which use semantic tag to get the

search results and to use the user profile to personalized the search results. The experimental results indicate that the method is effective. In the future, we are interested in further exploring the factors which impact the quality of ranking. Such as the potential relationships between items and associated groups, the friendship and membership, the user context.

Reference

[1] Demidova, E., Zhou, X., Nejdl, W.: Iq p : incremental query construction, a probabilistic approach. ICDE 2010, pp. 349-352 , 2010.

[2] Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the Semantic web with Corese Search Engine. In Proceedings of 15th ECAI/PAIS, Valencia (ES), 2008.

[3] R. Guha, R. McCool, and E. Miller. Semantic Search. In Proceedings of the 12th international conference on World Wide Web, pages 700-709, 2003.

[4] J. Heflin and J. Hendler. Searching the Web with SHOE. In Proceedings of the AAAI Workshop on AI for Web Search, pages 35-40. AAAI Press, 2000.

[5] C. Rocha, D. Schwabe, and M. de Aragao. A Hybrid Approach for Searching in the Semantic Web. In Proceedings of the 13th International World Wide Web Conference, 2004.

[6] M.C. Schraefel, N.R. Shadbolt, N. Gibbins, H. Glaser, and S. Harris. CS AKTive Space: Representing Computer Science in the Semantic Web. In Proceedings of the 13th International World Wide Web Conference, 2004.

[7] S. E. Sim, M. Umarji, S. Ratanotayanon, and C. V. Lopes. How well do search engines support code retrieval on the web? ACM Trans. Softw. Eng. Methodol., 21(1):4:1--4:25, 2011.

Published by Atlantis Press, Paris, France.