A Personalized Literature Recommendation Method Based on the Domain-Driven User Interest Model

Wenjin Sheng[1,a,*], Jianzhuo Yan[1,b] , Jianhui Chen[1,c] , Ruying Lv[1,d] and Hongzhi Kuai[1,e]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

[a] 932767262@emails.bjut.edu.cn, [b] yanjianzhuo@bjut.edu.cn,
[c] chenjianhui@bjut.edu.cn , [d] lvruying@emails.bjut.edu.cn,
[e] kuaihongzhi@emails.bjut.edu.cn
*Corresponding author

Keywords: literature recommendation, cognitive science, spread activation theory, domain-driven user interest model

Abstract. At present, literature is the most important source of scientific knowledge. However, in the face of the information overload caused by the explosive growth of scientific literatures, it is difficult for researchers to find the literature that is really needed quickly. This is particularly acute in the current universal concern field of cognitive science. This paper forms a personalized literature recommendation method based on domain–driven user interest model. By adopting the recommendation method based on BI provenances, initial literatures can be acquired from the first recommendation module. Furthermore, spread activation theory is added into the second recommendation module for obtaining user interest model. Results of experiments show that the proposed method can make full use of the advantages of the two modules, which can not only recommend literatures that relevant to the user's research interests but also recommend literatures that in other relevant heat research domain.

1. Introduction

With the rapid development of Internet in information age, the number of literatures has been increased rapidly. According to the statistical data of literatures released by the public biomedical literature dataset PubMed, from 2012 to 2015, researchers had published about 4376440 literatures, while the number in 2015 was 9.61% higher than in 2012. And it is predictable that the growth trend won't change in the short term. This is the so-called information overload problem and the best way to solve the problem is through the recommendation system. Personalized literature recommendation system [1][2] is a subclass of the recommendation system. In personalized literature recommendation system, through analyzing user's personal information and historical behavior to find the user's potential interest preferences is the most important, so that it can actively

recommend the literature information to user corresponding to their research interests. Therefore, it is very important to design a good personalized literature recommendation method for scientific research.

In this paper, we will use the literature of cognitive science [3][4] field as an example to explore a personalized literature recommendation method. Cognitive science is a cutting-edge discipline that explores and transforms information in the human brain. As a highly interdisciplinary subject, the research covers a wide range of research fields such as psychology, artificial intelligence, philosophy, sociology and so on. In addition, it uses behavioral experiments, brain imaging, computational modeling, various neurobiology methods and other diversified research means. As an important branch of brain science, cognitive science is also a hot topic of current scientific research. So there are a large number of relevant scientific and technical literatures published. For example, when we want to retrieve the literatures published about "cognitive" from PubMed, the number of results is 287754. In face of such a large number of scientific and technological literatures, researchers need to spend much time to find suitable information. So In this paper, we develop a personalized literature recommendation method for cognitive science researchers.

The rest of this paper is organized as follows. Section 2 discusses background and related work, mainly about recommendation algorithms. Section 3 illustrates the details of the proposed method. Experiments are presented in Section 4. Finally, Section 5 gives conclusion and future work.

2. Background and Related Work

At present, the main recommendation algorithms are mainly categorized into the following three: collaborative filtering recommendation, content-based recommendation and knowledge-based recommendation. The process of collaborative filtering recommendation is to form a system by analyzing user's interest to find similar users and integrating similar user's evaluation to certain information. The system can predict user's preference degree for this information and recommend useful information to user. However, this method does not go deep into the content of the literatures. Although it is conducive to discover new interests, it always tends to over-enlarge the long tail effect of interests [5] and recommends the results that do not match the user's scientific research objectives. The content-based method solves the problem that does not analyze the content of the literatures. The core of this method is to calculate the similarity between the content features of the item and the interest features in the user model. However, when constructing the user interest model and the literature model, such method usually regards words as isolated elements and there is no semantic association between words. This kind of processing is unreasonable, because text is through a certain form of language to represent the entity, the concept and the relationship. Only the logical word arrangement can clearly express the semantic content of the text and the simple stack of keywords can't express the clear meaning.

In recent years, taking into account the importance of knowledge in the field of literature inquiry process, knowledge-based literature recommendation method has been paid more and more attention. It uses knowledge about users and items to pursue a knowledge-based approach to generating a recommendation, reasoning about what items meet the user's requirements. Pretschner and Gauch [6] is the pioneer of using ontology to establish user interest model and providing personalized literature access. They used domain ontology to organize literatures and according to the user's browsing history to establish user interest model. Finally, personalized literature access was achieved according to the current value of the ontology concept corresponding to literatures. Chen Yifeng [7] aimed at the difficulties and shortcomings of ontology construction and model updating in modeling user interest, they proposed a method of constructing user interest model based on ontology. Ontosearch [8] and Weidong Zhao [9] proposed that the spread activation theory

should be applied to the process of literature retrieval, in this way the initial concept set can be expanded. Jianhui Chen [10] adopted Data-Brain model for user interest filtering, so as to accurately capture user's interest. Through the use of knowledge-based method for literature recommendation, we can effectively use the semantic association to solve the problem of the isolation of words based on content recommendation, so as to provide users with more accurate recommendation results.

However, most of the current knowledge-based personalized recommendation methods can't both recommend literatures that meet user's interest and domain-related. In this paper, we propose a recommendation method based on domain-driven user interest model (DDUIM). We first acquire the initial recommendation literatures by using the recommendation method based on BI provenances [10]. In order to reason and expand the user's research field, fully exploit the user's potential interest preferences outside the current filed, recommend more diverse literature, broaden the user's research horizons, we propose a second recommendation module that adopts the Data-Brain as domain ontology to optimize recommendation results based on the spread activation theory.

3. Literature Recommendation Method Based on the Domain-Driven User Interest Model

As illustrated in Figure 1, there are two parts in the proposed literature recommendation method: the first recommendation module and the second recommendation module. The first recommendation module uses the recommendation method based on BI provenances. As this paper mainly realizes improving the lack of the first recommendation module, that is the second recommendation module. So we introduce the second recommendation module in detail. The block diagram of the method is as follows:

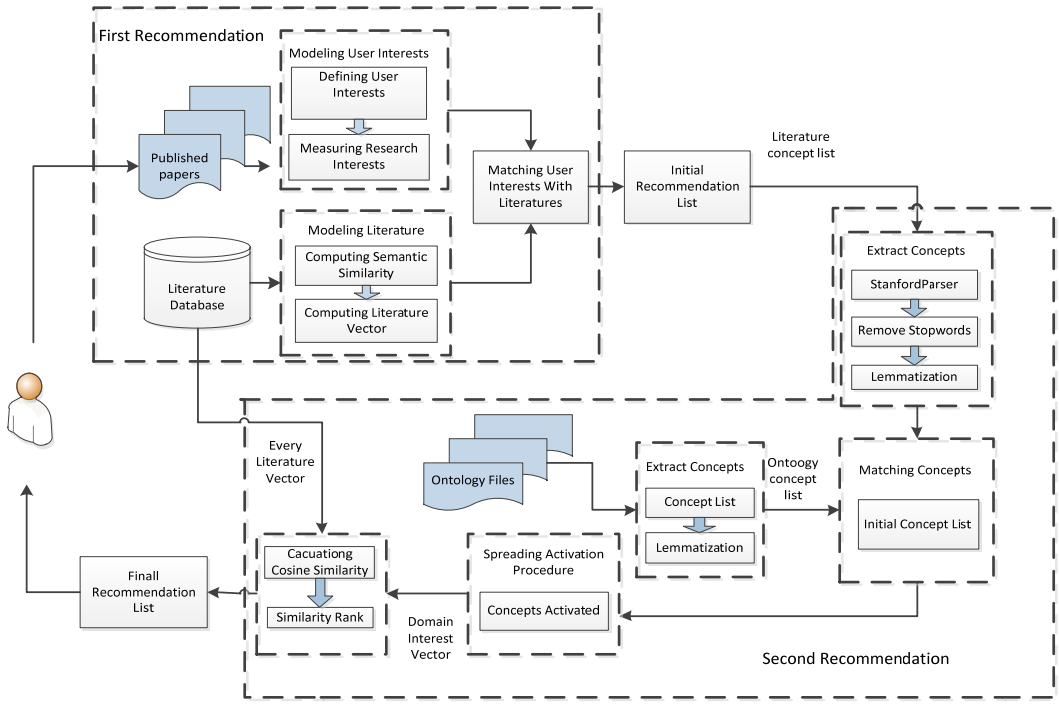

Figure 1 The framework of recommendation method based on DDUIM.

3.1. Theory of the first recommendation

The first recommendation module is based on BI provenances. It takes into account user's previous research preferences and short-term interest migration to reflect user real-time preferences exactly and efficiently. In addition, semantic similarity is added into traditional literature vector modeling for obtaining literature models. Results of experiments show that this method can improve the data sparseness caused by the vector space model and can accurately obtain the user's research interest so as to recommend needed information to users.

3.2. Theory of the second recommendation

As the recommendation results of the first recommendation module are too singular to find the user's potential interest in other research area. Therefore, we introduce the second recommendation module based on the result of the first recommendation module. The second recommendation module consists of three parts: modeling user interests, modeling literatures and matching user interests with literatures. The detailed will be discussed as follows.

3.2.1. Modeling User Interests

(1) Spreading Activation Model

In 1994 Bock and Levelt proposed spreading activation model [11] which regards concepts as nodes and nodes are connected with each other by using semantic relationship. Activation is the process that from one node to other surrounded nodes. It means that once a node is activated, its effect will spread to other nodes that connected with it and the activation energy will diffusion as the increase of spreading distance. Spread activation model was often used in concepts searching. The process of Spread activation theory can be described using the following formula [8]:

$$O = [\varepsilon - (1 - \alpha) \times w^{T}]^{-1} \times I \qquad (1)$$

Where $I = [I_1, I_2, \ldots, I_n]^T$ is the initial input of network; w represents the relationship matrix of network; the element in the matrix w_{ij} represents the relationship between the concept c_i and the concept c_j; α is the decay vector and we set it as 0.2 [12]; ε is an n× n identity matrix of order n; $O = [O_1, O_2, \ldots, O_n]^T$ is the final output vector, O_i represents the activation value of concept c_i。

(2) DDUIM

Spread activation model is often used in the domain of computer science, information retrieval as one of the common applications. When users input a keyword in the retrieval system of Ontosearch, the method will extract the relevant concepts as the initial concept list and then reason semantically related concepts using spread activation theory in computer domain ontology. Different from the method in Ontosearch, in this paper we combine spread activation theory with Data-Brain and set different weight for different relationship type in the Data-Brain. In this study, we set weight 1, 0.8, 0.6, 0.4 to the relationship isEquivalentOf, subClassOf, isSiblingClassOf, and the others respectively. In addition, as the change of researcher's interests is a gradual process of transition. So, different from the Ontosearch, we set a threshold value to the concepts after the spread activation process. And we consider the concept that has larger value than the threshold value is more relevant to domain research interest.

The input of the spread activation model I is the mapping of initial recommended literatures and Data-Brain, which represents the current research interest and current research heat of users. I is an

$n \times 1$ matrix that constituted by initial recommendation list and the element I_i in it represents the initial activation value of the concept c_i, the calculate formula is as follows:

$$I_i = \begin{cases} \frac{freq(c_i)}{\sum_{c_j} freq(c_j)}, & c_i, c_j \in LS_{initial} \cap KG_{BI} \\ 0 & , c_i \notin LS_{initial} \cap KG_{BI} \end{cases} \quad (2)$$

Where $freq(c_i)$ represents the frequency of concept c_i in initial recommendation literature; $LS_{initial}$ represents the initial recommendation list, KG_{BI} represents the Data-Brain.

When we acquire the input vector I, the process of spread activation theory is start. The relevant nodes will be activated and finally reach a steady state. In the literature recommendation method, w represents the relationship matrix of Data-Brain and the element w_{ij} represents the proportion of r_{ij} in Data-Brain, the calculate formula is as follows:

$$w_{i,j} = \begin{cases} \frac{freq(r_{i,j})}{\sum_j freq(r_{i,j})}, & rel \in isEquivalentClassOf \\ \frac{0.8freq(r_{i,j})}{\sum_j freq(r_{i,j})}, & rel \in subClassOf \\ \frac{0.6*freq(r_{i,j})}{\sum_j freq(r_{i,j})}, & rel \in isSiblingClassOf \\ \frac{0.4*freq(r_{i,j})}{\sum_j freq(r_{i,j})}, & rel \in others \end{cases} \quad (3)$$

Where the molecular $freq(r_{ij})$ represents the frequency of r_{ij} in Data-Brain, denominator represents the frequency of all the relationship that include concept c_i in the Data-Brain. When the spread activation process is over, we set a threshold value ε to the concepts. In this paper, the threshold is set to 0.01. The calculate formula is as follows:

$$O_i = \begin{cases} O_i & O_i \geq \varepsilon \\ 0 & \end{cases} \quad (4)$$

The output matrix $O = [O_1, O_2, \ldots, O_n]^T$ corresponds to the expanded user domain preference in the Data-Brain. As can be seen in the formula (3), the activation value depends on the structure level and semantic association in Data-Brain between the activated concepts with initial input concepts. The activation value will be larger if one concept has a stronger association with initial input concepts. On the one hand, the activation value describes the relation level between the current concepts with user research interest. On the other hand, it reflects the heat in the field of cognitive science.

3.2.2. Modeling Literatures

In this process, we use VSM [13] to define each literature. For each literature d_j, it is represented as $\vec{d_j} = (c_{1,j}, c_{2,j}, \ldots, c_{n,j})$, in which n is the number of non-repetitive concepts of Data-Brain and in the process of acquiring concepts we need to do duplicate removal and lemmatization, $c_{i,j}$ represents the weight of c_i in literature d_j. Different from the OntoSearch based method, in the process of calculating $c_{i,j}$ we use the traditional tf/idf [14] measure $c_{i,j} = freq_{i,j} \log \frac{N}{n_i}$,, where $freq_{i,j}$ is the frequency of c_i in literature d_j, N represents the total number of literatures in the dataset, n_i represents the number of literatures including concept c_i.

3.2.3. Calculating Similarity

Through the above steps we can get the literature vector d_j and the output vector O. And we use cosine similarity method to measure the correlation between the two vectors so as to calculate the correlation between the literature vector and the output vector. The similarity measure is computed as follows:

$$\mathrm{Sim}(d_j, q) = \frac{\overrightarrow{|d_j|} \cdot \overrightarrow{|q|}}{\overrightarrow{|d_j|} \times \overrightarrow{|q|}} \qquad (5)$$

Cosine similarity refers to measuring the angel of two vectors to measure the similarity between them. It is usually used in the field of textual mining and information retrieval. The two vector's angle varies from 0 to 1, 0 is completely dissimilar, 1 represents exactly similar. By using cosine similarity we can acquire papers that have different degrees of domain interests, and then sort the literatures according to the similarity value in descending order. The larger the value is, the more similar they are. And finally the sorted literature list will be returned to the users.

4. Experiments and Evaluation

4.1.Experiments Data

Our literature recommendation method is oriented to the cognitive science research staff. The main purpose of the method is to provide the service for the majority of cognitive science researchers. So we choose experiment data that relevant to cognitive science. Pubmed dataset provides publication information, title, abstract, keywords, author and so on. The core of pubmed dataset is biomedical, so we choose Pubmed as our experiment data source. We extract the literatures from Pubmed that published from 2005 to 2007 in three of top ten neuroscience impact factor including Trends in neurosciences, Nature neuroscience and Neuron.

We use Data-Brain as the domain ontology in which contains concepts and relationships [15][16]. The Data-Brain is a domain-driven conceptual model of brain data, which represents multi-aspect relationships among multiple human brain data. As Data-Brain was constructed by adopting ontology, so the Data-Brain can be used as a field of cognitive science ontology. It includes 297 cognitive science field concepts and it includes multiple relationship types, including isEquivalentClassOf, subClassOf, isSiblingClassOf and others. Therefore, this paper uses the Data-Brain as the domain ontology, on which the activation diffusion operation is performed.

4.2.Experiment Design

We treat Dr. Liang as a user, who is a cognitive science researcher. And we then recommend literatures to him by the proposed method. From the first recommendation module, the N literature list can be acquired. We extract the concepts in them and then match with the concepts in Data-Brain to form the initial input vector. Then we use UUIDM to activate relevant concepts to form the output vector. The output vector is the domain interest vector. Using TFIDF, we can get the literature vector. And we use cosine similarity to calculate the relevance between literature vector and domain interest vector, finally the sorted literature list will be returned to the users. The specific algorithm flow is as follows:

Table 1: Algorithm of the recommendation method based on DDUIM

```
Begin
  Input data：
    X：User published literatures
    Y: All literature in Pubmed
  Part 1:The first recommendation module:
      {
       Output: N literatures , represents by Z
      }
  Part 2:The second recommendation module:
      {
        Step 1: Model User Interests：
          For (all literatures in Z)
              {
                Extract title, abstract;
              }
          Calculate output vector by using formula(1)(2)(3)；
            Calculate user interest model by using formula(4),represents by O；
          END
        Step 2: Model Literatures：
          For (one literature in Y)
              {
                Calculate literature vector by using TFIDF， represents by d_j；
              }
          END
        Step 3: Calculate similarity：
          For(d_j)
              {
                Calculate similarity by using formula(5);
              }
        END
      Return the sorted literature list to the user
      }
END
```

4.3.Evaluation results and discussion

The results of first the recommendation module and the second recommendation module were analyzed. The results in Table 2 are acquired from the first recommendation module based on BI provenances. The results in Table 3 are acquired from the second recommendation module based on DDUIM. As can be seen from the table, there are some similar results between the top ten results of second recommendation module and the initial recommend results. This is because the input of DDUIM comes from the initial recommendation literatures and the interests retained after spread activation process. In addition, changes have taken place in most recommended literature's sorting. This is because after the spread activation process in Data-Brain, some concepts that have strong domain relevant with initial recommend literatures can be acquired.

Table 2 Literatures recommended based on BI provenances.

Sort	Title
Top-1	Separate modulations of human V1 associated with spatial attention and task structure.
Top-2	Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders.
Top-3	Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex.
Top-4	Two cortical systems for reaching in central and peripheral vision.
Top-5	Strange feelings: do amygdala abnormalities dysregulate the emotional brain in schizophrenia?
Top-6	Breakdown of functional connectivity in frontoparietal networks underlies behavioral deficits in spatial neglect.
Top-7	Reward-related FMRI activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation."
Top-8	Shift of activity from attention to motor-related brain areas during visual learning.
Top-9	Integration of touch and sound in auditory cortex.
Top-10	The effects of vestibular lesions on hippocampal function in rats.

Table 3 Literatures recommended based on DDUIM.

Sort	Title	sort in the first recommendation
Top-1	Strange feelings: do amygdala abnormalities dysregulate the emotional brain in schizophrenia?	5
Top-2	Network reset: a simplified overarching theory of locus coeruleus noradrenaline function.	2478
Top-3	Parietal lobe contributions to episodic memory retrieval.	28
Top-4	Language outside the focus of attention: the mismatch negativity as a tool for studying higher cognitive processes.	126
Top-5	Over-inhibition: a model for developmental intellectual disability.	77
Top-6	Theta burst stimulation of the human motor cortex.	698
Top-7	Cerebellar circuitry as a neuronal machine.	387
Top-8	Towards a neural basis of music perception.	798
Top-9	Beyond mind-reading: multi-voxel pattern analysis of fMRI data.	132
Top-10	Regret and its avoidance: a neuroimaging study of choice behavior.	12

Part of the spread activation process in DDUIM is shown in Figure 2. The white part of in the figure is the initial concepts with an initial value and the concepts with value 0 are not listed. The blue part in the figure is the activated concepts and their values. The activation value describes the

association degree in Data-Brain between the activated concepts and the initial concepts. It also reflects the current research heat in the field of cognitive science.

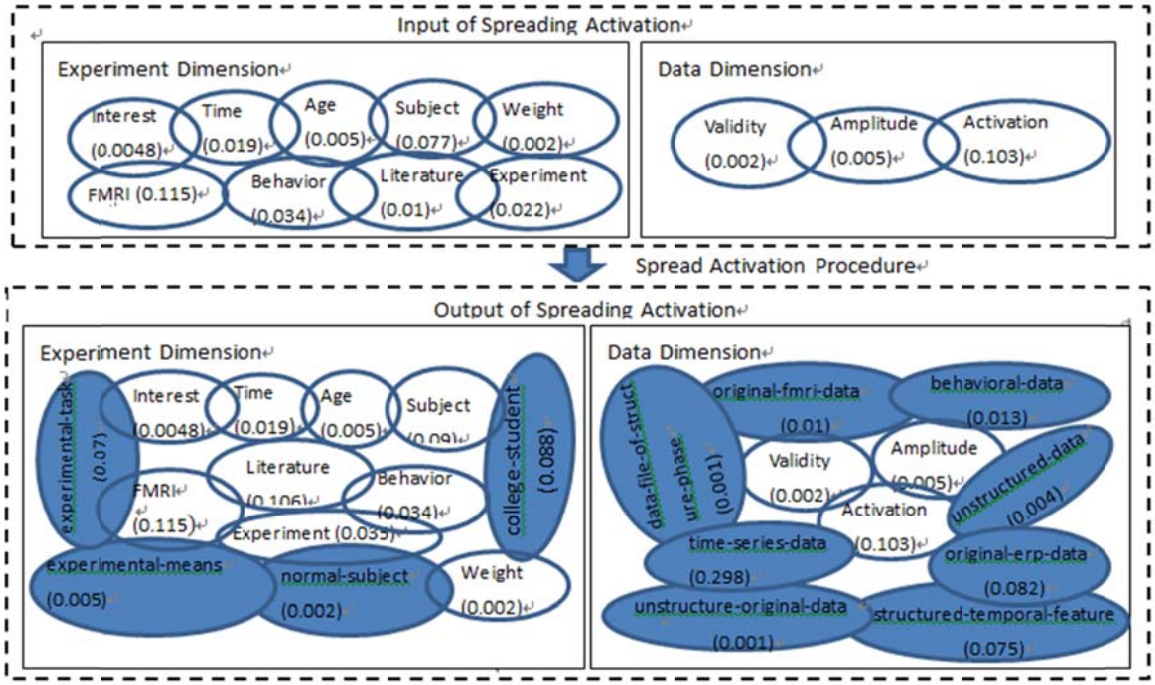

Figure 2 Spreading activation procedure of UUDIM.

DDUIM is to reason user's current research field. It breaks the limitation of user interest aspects in the recommendation method based on BI provenances. It fully mines user's potential interests and recommends literatures in other heat research domain. The diverse recommendation results can expand the researcher's perspective. These literatures meet the domain hotspot and can provide guidance in terms of users who want to expand the field of research. In addition, we further analyze the literatures that have big sorting contrast: Top 2 proposes a simplified overarching theory that according to the activation time and cognitive change of locus coeruleus to noradrenaline, Top 4 uses MMN for studying a higher cognitive process, Top 6 proposes using rTMS to study human motor cortex. Top 7 studies the cognitive and autonomic function of cerebellar. Top 9 uses multi-voxel pattern analysis of fMRI data to study human cognitive function. As can be seen from the Table 4, most of these literatures focus on cognitive function and the application of FMRI which is different from the result acquired from the first recommendation model. From the recent research we can see the user research interests changed, which pays more attention to the cognitive function in human brain and the application of FMRI. This is matching with our recommendation result to some degree. This explains that our recommendation results close to domain research hotspot and can help to expand the research interest. In addition, most of the recommended results are relevant to the cognitive function or application-related fMRI data, which further demonstrates the effectiveness of our domain relevance model.

Table 4 Papers published by the user in recent years.

Serial number	Title	Publication time	sort of author
1	Prefrontal and parietal activity is modulated by the rule complexity of inductive reasoning and can be predicted by a cognitive model.	2015	2
2	Abnormal regional homogeneity in Parkinson's disease: a resting state fMRI study.	2014	2
3	Different strategies in solving series completion inductive reasoning problems: An fMRI and computational study.	2014	1
4	Three Subsystems of the Inferior Parietal Cortex are Differently Affected in Mild Cognitive Impairment.	2012	1
5	Changes in thalamus connectivity in mild cognitive impairment: evidence from resting state fMRI.	2012	3
6	The baseline and longitudinal changes of PCC connectivity in mild cognitive impairment: a combined structure and resting-state fMRI study.	2012	2
7	Baseline and longitudinal patterns of hippocampal connectivity in mild cognitive impairment: evidence from resting state fMRI.	2011	2

5. Conclusion

In this paper, we develop a personalized method of literature recommendation method for cognitive science researchers. We combine the literature recommendation method based on BI provenances with the literature method based on the DDUIM. In addition, we consider the relation type of Data-Brain and we set threshold value to remove the unimportant concepts. Our method makes it possible for researchers to acquire literatures they are interested or relevant to their current domain interests. In the future, we need to improve the existing ontology so as to offer more accurate results to users.

Acknowledgements

The work is supported by National Key Basic Research Program of China (2014CB744605), National Natural Science Foundation of China (61272345), Research Supported by the CAS/SAFEA International Partnership Program for Creative Research Teams, the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (25330270).

References

[1] Liu, Y.Y., Zhang, X.M.: Document Recommender Systems: Approaches to Increasing Information Retrieval Effectiveness [J]. Library and Information Service, 2007, pp.11-19.

[2] Wang G. Survey of personalized recommendation system [J]. Computer Science, 2012, pp.66-76.

[3] Norman D A. Twelve issues for cognitive science. Cognitive Science, 1980, 4(1): 1—32.

[4] Simon H A. Cognitive science: The newest science of the artificial. Cognitive Science, 1980, 4(1): 33—46

[5] Hervas-Draney A. Word of Mouth and Recommender Systems: A Theory of the Long Tail [J]. 2014.

[6] PretschnerA, GauchS. Ontology based personalized search [C]//Proceedingsof the 11th IEEE International Conference on Tools with ArtificialIntelligence, 1999: 391 398.

[7] Chen Yifeng，Zhao Hengkai，Yu Xiaoqing，et al. Research on Ontology-based User Interest Model Construction [J].Computer Science, 2010, 36(21) : 46-48.

[8] Jiang X, Tan A H. OntoSearch: A Full-Text Search Engine for the Semantic Web [C]. National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, Usa.2006:1325-1330.

[9] Zhao W, Wu R, Dai W, et al. Research Paper Recommendation Based on the Knowledge Gap[C]. IEEE International Conference on Data Mining Workshop. IEEE, 2015:373-380.

[10] Ningning Wang, Ning Zhong, Jian Han, Jianhui Chen, Han Zhong, Taihei Kotake, Dongsheng Wang and Jianzhuo Yan. A personalized method of literature recommendation based on brain informatics provenances. Proceedings of Brain Informatics and Health-8th International Conference (BIH 2015), Springer Verlag, LNAI9250, 2015:167-178.

[11] Zhao W, Wu R, Dai W, et al. Research Paper Recommendation Based on the Knowledge Gap[C]. IEEE International Conference on Data Mining Workshop. IEEE, 2015:373-380.

[12] Ke Sheng.: Design and Implementation Strategy based on Semantic Network User Ontology Model. University of Electronic Science and Technology of China, (2014)

[13] Crouch, C.J., Crouch, D.B., Nareddy, K.: Connectionist Model for Information Retrievalbased on the Vector Space Model. International Journal of Expert Systems, 1994, pp.139-163.

[14] Rijsbergen, C. J. V.: Information Retrieval. London: Butterworths, 2nd edition, 1979.

[15] Zhong, N., Chen, J.H.: Constructing a New-style Conceptual Model of Brain Datafor Systematic Brain Informatics [J]. IEEE Transactions on Knowledge and DataEngineering, 2012, 24(12):2127-2142.

[16] Chen J, Zhong N. Data-Brain Modeling for Systematic Brain Informatics[C]// Brain Informatics, International Conference, BI 2009, Beijing, China, October 22-24, 2009, Proceedings. 2009:182-193.