Research Topics Variation Analysis and Prediction Based on FARO and Neural Networks

Hongyin Zhu *, Yi Zeng *†, Yiping Yang*
*Institute of Automation, Chinese Academy of Sciences, Beijing, China
†Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China
Email: {zhuhongyin2014, yi.zeng}@ia.ac.cn

***Abstract*—Given the explosive growth of scientific information and the fast advancement of research fields, researchers may not be able to find the most promising topics to combine with their current research and may be trapped in a few familiar research topics without creative ideas. Many studies of recommendation system make the effort to address the above problem, but they ignore the different styles of users and generate the recommendation results based on a common strategy. In this paper, we propose a framework to generate the adaptive recommendation results according to the research styles of users. Our framework contains 3 main parts, the research topic ontology construction, trend prediction and recommendation. First of all, the Fun of Academic Research Ontology (FARO), which has the capacity of describing dynamic and static research features and building a social network, is constructed to organize entities about academic research. Secondly, this paper predicts the popularity variation of research topics with the neural network model. Finally, some adaptive topics are recommended to specific researchers according to the evaluation of their research styles. Basically, this paper is inspired by the associative thinking of human brain to combine the advantages of Web knowledge representation language and the neural network to execute the prediction and recommendation. We test our results based on the publication data of IEEE and Springer. The experimental results demonstrate that our prediction model has a good generalization performance. A questionnaire survey is carried out to assess the recommendation results, and the result shows the feasibility of our method.**

I. INTRODUCTION

Researchers may change their research topics or combine with other creative ideas, given the fast advancement of science. If researchers can combine promising ideas with their current research topics, they may make further excellent contributions. However, it is hard to discover the best topics, which their current work can combine with, in manual screening. Generally, their horizons may be limited to the classical or familiar ones due to lacking of the knowledge about other fields. It is also difficult to predict which research topics will be the next public focus. Some studies of recommendation system make the effort to address the above problems, but they ignore the different styles of users, leading to generating the classical recommendation results for everyone. Inspired by the associative thinking of the human brain [1], this paper provides the objective evaluation of research styles of researchers and predict the development trend of research fields. Our framework also recommends some adaptive research topics to users.

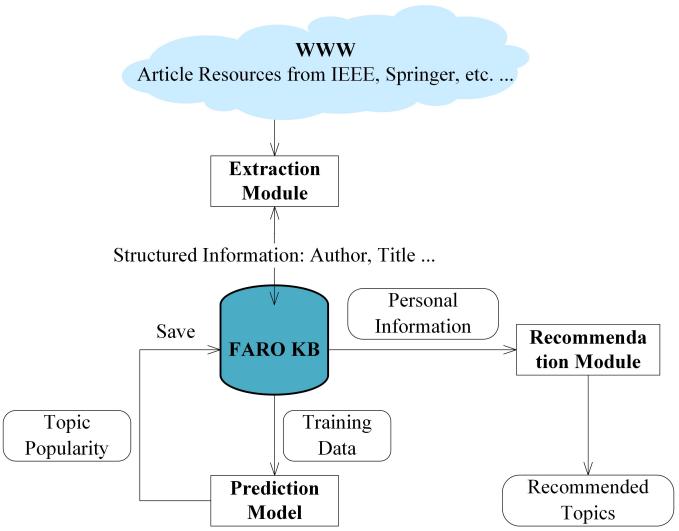

Fig. 1. The architecture for Research Topic Recommendation

Rungworawut et al. adopt ontology to build user profile for product search system [2]. However, this method cannot generate the user profile automatically, because it is based on manual inputs from users. Nakamura et al. generate the recommendation results based on the Web log of users [3]. Many recommendation systems adopt the history information of users to execute the recommendation, but they cannot carry out the recommendation based on the prediction information. As for the research topics, there is a long time delay between articles publication and the beginning of the research work, because the process of research and publication is time-consuming. The work corresponding to the publication data may be started years ago. We do not suggest to directly generate the recommendation from the current publication data since the prospects of topics may be changed during this period. So we generate the recommendation results based on the prediction information. In this paper, we propose a novel method to evaluate the research styles of researchers to generate the adaptive recommendation results. In addition, we add the semantic relevance factor to social community construction to describe and organize research communities in a more complete way.

The Fun of Academic Research Ontology (FARO) is pro-

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

posed and constructed to be used in managing the academic research related resources and generate training data for the prediction model. As shown in Fig. 1, the extraction module extracts information from the World Web Web and saves them into the FARO Knowledge Base (FARO KB). The FARO KB generates the training data for the prediction model, and the prediction values are stored back to the FARO KB. Predicting the variation trend of research topics is challenging due to its uncertain pattern. The neural network models are adopted to predict the popularity degree of any research topic. The reason for choosing these models is their fitting capacity to linear and nonlinear problems [4]. Incorporating with the resources in FARO KB, the prediction model can recommend some adaptive research topics for a specific person.

II. RELATED WORKS

In this section, we will discuss related works and our method on dynamic research topics description and prediction. We analyze the challenges in the description of interests and discuss the works on the prediction of research topics.

A. Research topics description

Many personalized search and recommendation systems are built based on the user profile generated in the registration stage or through Web browsing logs [3]. They cannot build an appropriate model for the weight of user interests. In the recommendation level, some clustering algorithms are adopted to calculate the similarities of users. For example, the personalized book list is recommended according to the different groups of university members clustered by K-means algorithm [5]. As demonstrated in [6], the dynamics and evolution of research interests also need to be considered, because they are dynamically changing all the time.

Understanding the current interests of users is a challenging task, especially when facing the dynamics of the personal information description. In this paper, FARO is designed to describe and support understanding of research topics and interests of scientific researchers and their relationships in the academic network. The experimental data is extracted from the IEEE and Springer digital library.

B. Research topics prediction

After obtaining the related information on published articles, we organize the information into training data format for the prediction model. Kang et al. use Hidden Markov model to predict the behavior of users [8]. Li et al. adopt the RBF neural network to improve the mining strategy [9]. But the above methods did not take full consideration on the influence of previous inputs. We adopt the dynamic neural network to take both the current input and the history of the input sequence into consideration, since we believe the variation process of the research topic is highly relevant with the historical line of changes.

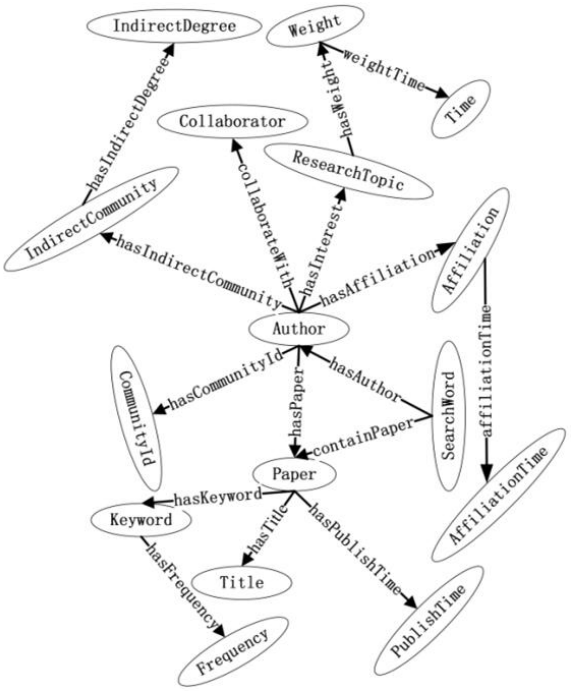

Fig. 2. Entities and Relations in the FARO Ontology

III. FARO

The Fun of Academic Research Ontology (FARO) is designed and used to describe the academic research related resources, as Fig. 2 shows. It is author information centric, and is with the capacity of describing dynamic progress and static resource related to scientific research. social relations of scientific community and dynamic description characteristics are the most notable features of FARO.

In FARO, various research topics are labeled with weight ("hasWeight") according to the importance degree to a specific person. The property "weightTime" can describe the time when the research topic has the corresponding weight. Given that the weight of the research topic is dynamic to most researchers, every recommendation needs to consider topic weights of all the members of their community. If related researchers were calculated every time, it would be a huge computation burden for every recommendation request. But the "hasWeight" and "weightTime" are used to save the weight and the time according to the first calculation, and these can be done offline. So the proper use of FARO can reduce the computational cost to make the real-time recommendation possible. Every keyword is linked to the published articles which are connected with publication time by "hasPublishTime". So FARO has the capacity of managing every resource in time dimension respectively.

In addition to the previously mentioned function, FARO is designed in this structure due to the following demands of calculation and description. For example, it takes social network factor into consideration, because lots of groups are clustered according to the research interests. We use FARO

to generate the social network of academic research from the perspective of authors. For example, we find out the researchers who are directly and indirectly correlated. It is hard to calculate all the possible communities every person belongs to in the real time. The "hasCommunityId" can save the communities that are clustered offline so that we can reproduce the social network according to the "CommunityId" easily. The indirect communities, which are generated based on some patterns, are used for recommending creative topics. From the "Keyword" perspective, the keyword network can be constructed to analyze the direct and indirect relations about research topics. The construction of the keyword network is based on the semantic relevance which will be discussed in section 5. This paper is inspired by the associative thinking, which is an approach to creative ideas because disparate concepts can be associated together in new combinations for specific solution or purpose [1]. Based on the above work, the structured information is saved in the FARO KB through an extraction module.

IV. Model Construction and Prediction

Research topics reflect the trend for scientific advancement. If researchers can predict the trend, they will have the higher probability to make more achievements. There are many factors in the world which influence the development of research fields. We find the number of relevant publications in a specific time interval can reflect the popularity of a specific research topic. This paper assumes the variation number of relevant articles follows a certain pattern. Based on the above hypothesis, the neural network is adopted to reveal the pattern of popularity variation.

Artificial neural network models have the capacity to fit linear and nonlinear problems [4]. In this paper, several methods are used to build models, and we compare their generalization performances with their MSE and bias distribution by 10-fold cross validation (10-CV). Given their performances, our training dataset is constructed as 10-dimension vector, which is better than other dimension settings. Each input dimension represents the number of relevant articles in a specific year of the past 9 years and the output is the number of relevant articles in the tenth year, as shown in Equation (1). According to [4], the neural network models are divided into two categories, the static and dynamic neural networks, on the basis of their mechanism.

$$f((x_t^i, x_t^{i+1}, x_t^{i+2}, x_t^{i+3}, ..., x_t^{i+7}, x_t^{i+8})^T) = x_t^{i+9} \quad (1)$$

where i is an arbitrary year, x_i^t denotes the number of articles on the topic t in the ith year. The x_{i+9}^t represents the number of articles on the topic t in the $(i+9)$th year.

We adopt the static neural networks. We compare 3 kinds of feed-forward neural networks, which consist of the back-propagation neural network (BPNN), Radial basis neural network (RBFNN) and the Cascade-forward neural network. The RBFNN earned the least MSE value (0.2219) compared to the BPNN (0.2782) and the Cascade-forward neural network (0.2933). The results show that most of the bias are under 4%, as shown in Fig. 3, where the y-axis represents the prediction bias and x-axis represents the test samples. The RBFNN achieves the best performance with the least MSE value and the fastest training speed.

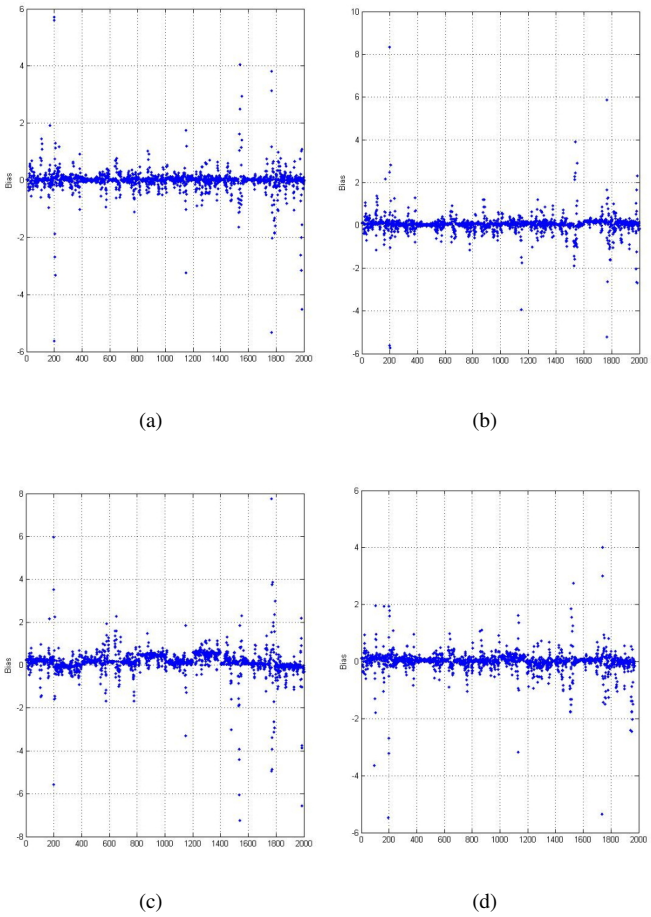

Fig. 3. (a) RBF neural network prediction bias distribution (b) BP neural network prediction bias distribution (c) Cascade-forward neural network prediction bias distribution (d) NARX prediction bias distribution

The output of the dynamic neural networks not only depends on the current input but also is influenced by the previous input [4]. From the perspective of dynamic input, our training dataset consists of time series in different years. We assume the popularity degree of a specific topic in the last year still influences the next year. We adopt the dynamic neural network to learn 2 models. The prediction result shows the least MSE (0.1781) of nonlinear autoregressive with external input (NARX) network and the MSE (0.1910) of Layer recurrent neural network, as shown in Fig. 3. It means the dynamic neural network may have good generalization ability to predict the research topics variation. And the popularity of research topics may be not only reflected but also influenced by the past several years. We believe these prediction results can be improved if we can set better model structure and use more training data [4].

V. EVALUATION

In this section, we propose a method to evaluate a researcher from the perspectives of Popularity, and Potential. Then we recommend research topics to the researcher as a service. The "Popularity" means the popularity degree of all the research topics of a person, and it reflects the research style. The "Potential" denotes the prediction of the research status of a researcher.

$$K(t(i),n) = \sum_{j=1}^{n} y_{t(i),j} \times AT_{t(i)}^{-b} \tag{2}$$

$$P(t(i),n,o) = \sum_{i=1}^{n} [K(t(i)) \times \omega_{t(i)}] \tag{3}$$

$$P(t(i),n,o) = \sum_{i=1}^{n} [K(t(i)) \times \frac{K(t(i),n)}{\sum_{l=1}^{o} K(t(l),n)}] \tag{4}$$

$K(t(i))$ is the number of relevant articles of keyword $t(i)$ in the last year in our knowledge base. For every person, $K(t(i),n)$ is the cumulative interest of keyword $t(i)$ to him [6], n is the total number of time interval, $y_{t(i),j}$ is the number of relevant articles of keyword $t(i)$ in the interval j. $T_{t(i)}$ is the duration of $t(i)$. According to [6], we use $A = 0.855$ and $b = 1.295$ in Equation (2). As is shown in Equation (3) and (4), $P(t(i),n,o)$ is the "Popularity" of the whole status of the person. $\omega_{t(i)}$ is the weight of the keyword $t(i)$ which represents the author's interests. o is the total number of keywords to a specific author and l is the index of the corresponding keyword.

$$P'(t(i),n,o) = \sum_{i=1}^{o} [K'(t(i)) \times \omega_{t(i)}] \tag{5}$$

The "Potential" is calculated by the prediction model. $K'(t(i))$ is the prediction number of relevant articles of keyword $t(i)$. $P'(t(i),n,o)$ is the "Potential" value of this author, as illustrated in Equation (5). For example, we use the data of one researcher to predict the trend of every keyword in next year. The positive "P" value means the keyword will be more popular while the negative value means the opposite. And the value also represents the degree of increase or decrease. The "C" means the concrete variation number of every keyword. The "R" means the variation rate compared with the last year, as shown in illustrative examples in Table 1.

During the above efforts, we did not take semantic relatedness of research topics into consideration. In order to obtain more accurate results, topics need to be reranked according to their semantic relatedness [7]. Firstly, according to [6], [7], when we rank the dynamic interests of a specific person, some of the interests are relevant to each other, such as "inference" and "reasoning". This problem has some influence in the characterization of user interests, so we adopt the following method to calculate the semantic relatedness and build the communities from the semantic perspective. According to [11],

TABLE I
TOPICS POPULARITY VARIATION PREDICTION

K	P	C	R(%)
Lexical	0.01483	15.96	1.48
Syntactic	0.01421	11.21	1.44
Parsing	0.01402	21.24	1.43
Semantic Web	0.01102	22.67	1.14
XML	0.01045	18.69	1.07
Trust	0.00802	117	1.02
web engineering	0.00829	50.88	0.97
ontology	0.00876	35.56	0.93
Semantic	0.00572	34.97	0.66
Attacks	0.0041	54.43	0.55
Ranking	0.00013	3.636	0.02
Similarity	-0.0096	-249	-1.9
security	-0.0133	-297	-2.2
process models	-0.0388	-3989	-3.9
software engineering	-0.0327	-795	-6.6

TABLE II
SEMANTIC RELATEDNESS OF TOPICS

T1	T2	NGD
Lexical	Syntactic	0.05337
ontology	Semantic	0.097347
Lexical	Semantic Web	0.112202
Semantic Web	Semantic	0.13002
ontology	Similarity	0.143103
Syntactic	Semantic	0.148108
Trust	security	0.182942
Trust	Attacks	0.250265
Attacks	security	0.25466

the semantic relatedness of any topic pair can be calculated by Equation (6).

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \tag{6}$$

The $f(x)$ and $f(y)$ represent the number of Web pages containing topic x or y respectively, and the $f(x,y)$ means the number of Web pages containing topic x and y at the same time. The M represents the total number of Google indexed Web pages and we take 1×10^{12} in this task. According to [6], if the NGD value is closer to 0, the topic pair is more relevant, vice versa. After obtaining the semantic relatedness, we also take the order of the topics into consideration, so the ranking of topics may be adjusted partially. We adopt the method in [6] to find the most relevant topic pairs and take the top-ranked topics to represent the ranking of other relevant topics. When the NGD is less than 0.3, the topic pair can be considered as relevant. Table 2 shows the highly relevant topics according to the statistical data from Google and Table 3 presents the adjustment of ranking according to the above method.

As for the community construction, the semantic relatedness can extend communities to add more relevant members and get the more complete result. First of all, we calculate the relevance of topics to get the network of directly and indirectly relevant topics. For example, the two topics, "Lexical" and "Syntactic", have the relevance value 0.05337, which means they are closely relevant. As for the indirectly relevant topics, they are irrelevant directly but connected by some indirect paths. For example, the two topics, "Synaptic web" and

TABLE III
ADJUSTMENT OF TOPICS RANKING BASED ON SEMANTIC RELATEDNESS

R1	R2
Ontology	Ontology
Security	Semantic
Parsing	Semantic web
Attacks	Lexical
Syntactic	Syntactic
Semantic	Similarity
Trust	Security
Semantic web	Trust
Soft engineering	Attacks
Web engineering	Parsing
Lexical	Soft engineering
Ranking	Web engineering
Similarity	Ranking
XML	XML
Process model	Process model

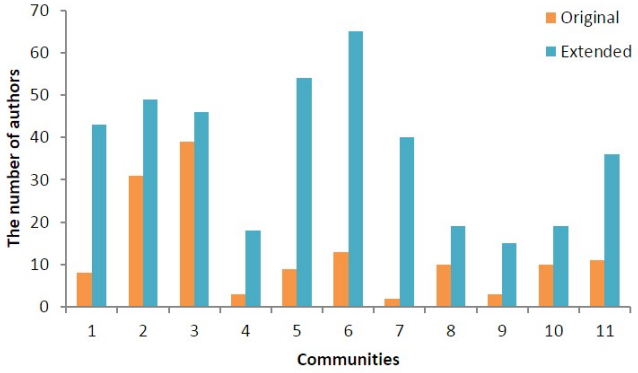

Fig. 4. The comparison of original and extended communities

"Recommendation of contents", which are connected through the path "Synaptic web, Inductive logic programming, Recommendation of contents". After we calculate the relevance of topics, the community can be extended at the same time. As is shown in Fig. 4, we select 11 communities randomly to analyze the number of authors who belong to these communities. The "Original" means the original communities, while the "Extended" means the extended communities with semantic relatedness. The results show that our method extends the communities in varying degrees.

Finally, we calculate the score of "Popularity" and "Potential" and rank the users in communities. The first 20% will get 5-star, and 20%∼40% will get 4-star, etc. If some researchers are accustomed to some unpopular topics, our system will change the recommendation strategy. Users with the score less than 2 stars will be recommended with innovative topics first instead of popular topics, because they are looking for something different compared to the major trend. The "Recommendation" is used to save the topics that are thriving or unique in a community. This paper ranks the prediction keywords, according to their C value, to take the first 10 as the recommendation results, as Table 4 shows. We removed the words that are already contained in the interests list. The variation trend of every topic is evaluated, as is shown in

TABLE IV
THE RESEARCH TOPICS RECOMMENDATION OF DIRECT SOCIAL COMMUNITIES

K	P	C	R(%)
Social Web	0.009	105	1.1
recommendation of contents	0.01	98.04	1.18
collaborative learning	0.013	73.13	1.38
Semantic Publishing	0.018	69.62	1.91
Collaboration	0.004	67.77	0.59
Knowledge retrieval	0.012	63.52	1.39
Agent Technology	0.004	61.69	0.54
service composition	0.008	55.5	0.95
Entropy	0.009	53.2	1.03
knowledge sharing	0.002	47.81	0.31

TABLE V
THE RESEARCH TOPICS RECOMMENDATION OF INDIRECT SOCIAL COMMUNITIES

K	P	C	R(%)
3D content	0.007	44.9	0.9
Domain and Range Identifier	0.016	21	1.7
3D web	0.011	18.7	1.2
Web Usage Mining	0.01	12.2	1.1
Web Content Mining	0.005	11.9	0.6
Semantic Annotated Data	0.011	11.6	1.1
Semantic Web Mining	0.007	10.9	0.7
semantic description	0.004	10.2	0.5
fuzzy inference system	0.012	8.46	1.2
Semantic Web Documents	0.009	8.13	1

Table 4. For example, "Social web" is predicted to have a fast growth compared with others. Sometimes, we cannot judge the popularity of topics merely based on the variation number, because research topics have the different cardinal number. So we can also take the variation rate into consideration.

In our system, users can find the popularity variation progress of topics in the last 25 years. For example, if users search the keyword "ontology construction", they will find the trend in the past 25 years as shown in Fig. 5. Since the dataset only contains data by the end of April, 2015, the value of 2014-2015 had been less than 2013-2014.

For a specific user, as shown in Fig. 6, the "Popularity" and "Potential" items present the 4-star score respectively, and the "Recommendation" item lists the recommended topics. Furthermore, the associative thinking can be used to find more creative research topics. For example, if the indirect distance to the community of an user is one, the system will provide some research topics from indirect communities, as shown in Table 5. If some more innovative topics are expected, further distance can be set.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a framework for research topic variation analysis and prediction. Firstly, the FARO ontology is proposed and designed to describe research related resource, and the FARO KB is used to manage the extracted resources which are automatically extracted from the Web. Secondly, we propose a method to predict the variation trend of research topics with dynamic neural network models. Our experimental results indicate that our prediction model has good generaliza-

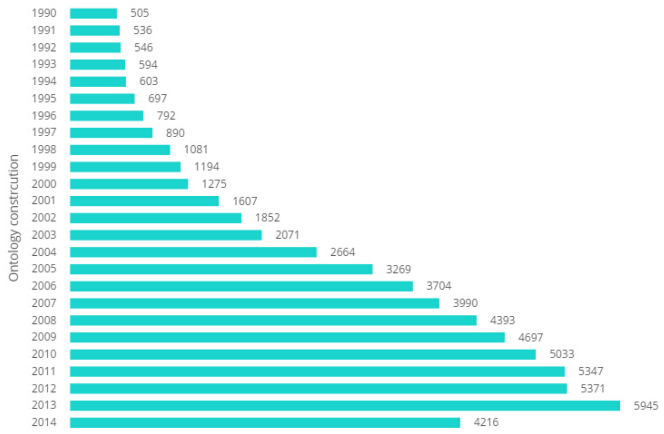

Fig. 5. The variation trend of the topic "ontology construction"

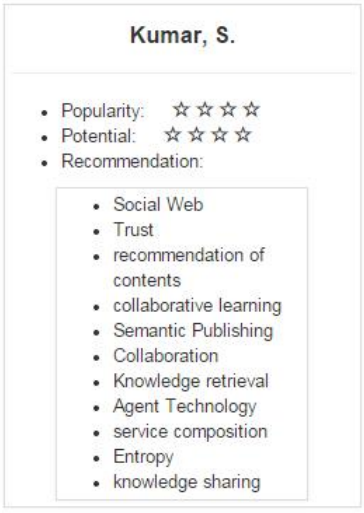

Fig. 6. The recommended research topics of one person

tion performance. Finally, we propose a method to evaluate the research style of a researcher and recommend some adaptive topics to them. We combine the advantages of knowledge representation and the neural network models to perform the prediction and recommendation.

In the future, the system will get the higher degree of integration based on many other multiple data resources. For example, we can add the interest analysis based on social media such as Facebook and Twitter, to extend the social network potential of our system [12], [13]. And we can also use more training data to train the neural network models. Finally, we might also use the proposed framework to enhance the functionality of the Linked Brain Data platform [14].

ACKNOWLEDGMENT

This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB02060007), and Beijing Municipal Commission of Science and Technology (Z151100000915070, Z161100000216124).

REFERENCES

[1] Sandra W. Russ, and Jessica Dillon. Associative Theory. in Encyclopedia of Creativity (Second Edition), Mark A. Runco, Steven R. Pritzker (Eds.) San Diego: Academic Press, 66-71, 2011.

[2] Rungworawut Wararat, and Surachet Kachonsri. Applying ontology-based personal profile for product search system. Proceedings of the 2012 International Conference on Information Science and Applications, 1-5, IEEE Press, 2012.

[3] Nakamura Akinori, and Nobuhiko Nishio. User profile generation reflecting user's temporal preference through web life-log. Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 615-616, ACM Press, 2012.

[4] M. H. Beale, M. T. Hagan, and H. B. Demuth, Neural Network ToolBox 7 User's Guide. The MathWorks, Inc., http://www.mathworks.com, Sept. 2010.

[5] Maneewongvatana Suthathip, and Songrit Maneewongvatana. A recommendation model for personalized book lists. Proceedings of the 2010 International Symposium on Communications and Information Technologies (ISCIT), 389-394, IEEE Press, 2010.

[6] Yi Zeng, Erzhong Zhou, Yan Wang, Xu Ren, Yulin Qin, Zhisheng Huang, and Ning Zhong. Research Interests : Their Dynamics, Structures and Applications in Unifying Search and Reasoning. Journal of Intelligent Information Systems, 37(1): 65-88, Springer, 2011.

[7] Yan Wang, Cong Wang, Yi Zeng, Zhisheng Huang, Vassil Momtchev, Bo Andersson, Xu Ren, and Ning Zhong. Normalized Medline Distance and Its Utilization in Context-aware Life Science Literature Search. Tsinghua Science and Technology, 15(6): 709-715, Elsevier, 2010.

[8] Kang Wonjoon, Dongkyoo Shine, and Doingil Shin. Prediction of state of user's behavior using Hidden Markov Model in ubiquitous home network. Proceedings of the 2010 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM 2010), 1752-1756, IEEE Press, 2010.

[9] Li Xiang, Ningjiang Chen, Qiqi Xie, Shilong Dong, Lirong Zhu, and Ying Tan. An Improved Mining Strategy of Preferred Paths in Web Applications Based on RBF Neural Network. Proceedings of the 2013 International Conference on Computer Sciences and Applications (CSA 2013), 300-304. IEEE Press, 2013.

[10] Sergio de Cesare, Damir Juric, and Mark Lycett. Toward the automation of business process ontology generation. Proceedings of the 16th IEEE Conference on Business Informatics, 1, 70-77, IEEE Press, 2014.

[11] Rudi L. Cilibrasi, and Paul MB Vitanyi. The google similarity distance. IEEE Transactions on knowledge and data engineering, 19(3): 370-383. IEEE Press, 2007.

[12] Yi Zeng, Ning Zhong, Xu Ren, and Yan Wang. User Interests Driven Web Personalization Based on Multiple Social Networks. Proceedings of the 4th International Workshop on Web Intelligence & Communities, colocated with the 2012 World Wide Web Conference (WWW 2012), Lyon, France, April 16th, 2012.

[13] Paula Penas, Rafael Del Hoyo, Jorge Vea-Murguła, Carlos Gonzlez, and Sergio Mayo. Collective Knowledge Ontology User Profiling for Twitter–Automatic User Profiling. Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 1, 439-444. IEEE Press, 2013.

[14] Yi Zeng, Dongsheng Wang, Tielin Zhang and Bo Xu. Linked Neuron Data (LND): A Platform for Integrating and Semantically Linking Neuroscience Data and Knowledge. Frontiers in Neuroinformatics. Conference Abstract: The 7th Neuroinformatics Congress (Neuroinformatics 2014), Leiden, the Netherlands, August 25-27, 2014.