An Intelligent Method for Expert Finding Based on Knowledge Organization Systems: Taking the Example of Oncology

Song Pei-yan¹,a, Wang Dong-fang¹ and Cao Yan¹
¹Institute of Scientific and Technical of Information of China, Beijing, China
²Corresponding author: songpy@istic.ac.cn

Abstract. In this paper, the experts are mapped and found through the concept and semantic relations of knowledge organization systems. First, a hierarchical model of expert description is designed based on the literature data and knowledge organization systems, as foundation of expert rapid finding, intelligent reasoning and dynamic updating. Then, 5 key steps are introduced based on knowledge organization systems, by which the research directions of experts can be identified accurately and visually in TF-IDF and LDA model. Finally, taking the example of Oncology, the experiment indicates that experts with the same or near research direction can be aggregated fast and effectively by semantic linkage compared with VSM agritourism. This method is expected to find out the most suitable experts by calculation and reasoning automatically.

1. INTRODUCTION
Experts with professional skill or technology are the most valuable human resource in the society. They usually play an important role in scientific research, project evaluation, achievement transformation, decision-making consultation and so on. Especially, the scientific research projects highly rely on the real suitable experts with crossing-discipline background, highly influence and outstanding innovative in their domains. So, it is urgent to figure out a method on how to find out the experts who have higher academic authority, more relevant discipline and stronger foundation in their own research to ensure the peer review experts quickly and objectively.

The current method of science and technology experts finding covers the following two ways generally: one is the experts themselves upload their academic information into database and then the experts will be selected by some constrained conditions such as age, affiliation or research direction etc. This traditional way is popular because it is easy to carry out. However, it is still facing some challenges. For example, it is not only difficult to verify the experts’ information one by one due to lack of independent control, but also very hard to update those information timely, which will greatly weaken the fairness and feasibility of expert finding. Furthermore, because all the information is stored in database statically, it is rather hard to link and reason by deep semantics to find out more and accurate experts to meet the real requirements. The other way is scientific metrology based on the literature statistics. In this way, experts will be evaluated by some common index according to the number or quality of literature, such as h-index, co-cited, cited-chain. This is an effective way to find out research paper and experts’ academic influence quantifiably. But it is still very hard to identify who
is the most related expert objectively, because some top experts didn’t have any paper published and no one can collect all the papers. Obviously, these methods mentioned above should be further improved in the mechanism of finding and updating. But the overall information of an expert should be described in a unified model and linked to multi resourcing data, by which the expert’s information could be changed on time. Then, based on semantics, the selected academic experts could be monitored more accurately including authority, professional skills and research direction. Fortunately, knowledge organization systems (KOS) have provided an effective semantic tools, such as thesaurus, ontology and knowledge graph, and it is expected to resolve the challenges above.

2. RELATED WORK
Researchers have proposed many effective methods for experts finding [1-2]. Based on the semantic relation of knowledge organizations, it is promising to solve the problems existing in the prior art. The knowledge organization tool with semantic and orderly knowledge representation model is suitable for effective aggregation and expansion of expert information, and it is also closely related to massive literature, so it facilitates monitor and verification of expert’ information objectively and dynamically. This provides a good theoretical basis and technical support for the accurate discovery and rapid updating the information of science and technology experts.

Fortunately, KOS has made great progress in recent years. Library scientists have developed classification tools [4-5], such as DDC, UDC and CLC, by which the knowledge could be organized hierarchically. Besides, thesaurus is another effective tool. The concepts and relations are linked by network framework, such as synonym, narrow term, broader term or related term. So all the terms will be arranged in the center of concept. Due to the semantic advantages, classification and thesaurus have been widely used in knowledge navigation, knowledge retrieval and knowledge reasoning [4-5]. In recent years, computer scientists also have developed more advanced knowledge organization system driven by knowledge engineering, for example, DBPedia, Yago and SUMO. Experts in professional domain can be described by terms of KOS, and the KOS semantic advantage also could be activated or calculated in expert finding. In a word, knowledge organization has made great progress, and it is possible to support the expert finding.

In this paper, a new model based on KOS semantic relation was designed, and then Oncology experts’ academic relations were constructed automatically. It is hopeful to find out the expert information quickly and accurately, and will be applied to the management and service of science and technology experts in other fields.

3. MODEL
The hierarchical model in this paper is defined as a kind of semantic aggregation framework for science and technology experts. Supported by the semantic framework of KOS, the hidden information of experts will be observed from mapping expert information to knowledge organization tools under the unified framework. The 3 level of the model are as follows.

A. Knowledge organization tools. It provides an open concept space, such as classification, thesaurus or ontology. It is very adaptive to different KOS. Meanwhile, multi-KOS also could be adapted by semantic mapping. So, under the concept framework, the expert’s information could be linked by terms.

B. Expert information aggregation. Each expert has his own research direction labelled by terms, and each term could be mapped to level A. Then, it is feasible to identify whether expert A and B belong to the same filed or not by semantic relevance. Especially, it is also possible to find experts hidden in different affiliations, even though they didn’t know each other before.

C. Expert evaluation and information update. The last level is literature, such as patent, research paper and reports. Therefore, it is helpful to evaluate information of an expert including academic influence, social network and recent research direction. Furthermore, it becomes feasible to update the information of experts in real-time and monitor their latest information containing working organization, title or post, and it provides a good foundation for the expert database construction.
Scientific projects usually need higher precision and dynamic characteristics to support the evaluation, management and innovation. This model covers expert information as an important part of knowledge base, in which semantic information will be open to share. The 3 levels are linked closely, and it is convenient to update the information of experts.

The model has taken advantage of the semantic features of knowledge organization tools, and mapped to the semantic relation between expert instances and topics. By means of literature keywords, the author and the knowledge organization tools are conceptualized to activate, associate and expand, and finally to build the knowledge relationship between expert individual and multidimensional semantic association. It has changed the traditional expert finding into intelligent automatic calculation. Therefore, it provides an open model to support the assessment and finding of experts in science and technology.

The main idea of the model is shown in Figure 1, which includes knowledge aggregation layer, expert aggregation layer and literature base layer.

![Figure 1. Sketch map of the main idea of knowledge map building model.](image)

As the research objects of this paper are the keywords, author, and research topic contained in science and technology literature, the network model is defined as follows:

1. Keywords, authors and research topic models

Keywords refer to knowledge units with complete knowledge expression, the network model constructed with keywords as nodes can be described as: $G=(V,E)$, $V(G) = \{k_1, k_2, k_3, ..., k_n\}$ is a collection of nodes, $E(G) = \{e_{12}, e_{13}, ..., e_{(n-1)n}\}$ is a collection of edges, $e_{ij}$ indicates the relationship between nodes, $e_{ij}=1$ means that there is a relationship between nodes $k_i$ and $k_j$. $\{(a_i, a_j), i,j = 1, 2, ..., m\}$ similarly, the author network and the research topic network are respectively described as: $G=(A,E_{aa})$, in which $A = \{a_1, a_2, ..., a_m\}$ is the collection of authors, $E_{aa}$ is a collection of edges, $\{(a_i, a_j)\}$ means there is a relationship between $a_i$ and $a_j$, which indicates that there is a cooperative relationship or the research fields of the two are similar. Where $G=(O,E_{o-o})$, $O = \{o_1, o_2, ..., o_m\}$ is the collection of research topics. $E_{o-o} = \{(o_i, o_j)\}$, $i,j = 1, 2, ..., m$ is a collection of edges, $\{(o_i, o_j)\}$ means there is a relationship between $o_i$ and $o_j$.

2. Interrelationships framework among networks

The definition of network in this paper covered the relationship between the author network and the research topic network, the relationship between author network and keyword network, and the relationship between keyword network and research topic network, defined as follows:

① Mapping keywords to author: Indicates which keywords the author has: $K(a_i) = \{k_j \mid k_j \in K, \theta(a_i, k_j) = 1\}$, $\delta(a_i, k_j) = 1$. indicates that the author $a_i$ has the keyword $k_j$.

② Mapping of keywords to research topics: Indicates which research topics the author has: $A(o_i) = \{a_j \mid a_j \in A, \theta(a_j, o_i) = 1\}$, $\delta(a_j, o_i) = 1$. indicates that the author $a_j$ has a research topic $o_i$.

③ Mapping of research topics to keywords: Indicates which keywords the research topic has: $O(k_j) = \{o_j \mid o_j \in O, \theta(o_j, k_j) = 1\}$, $\delta(o_j, k_j) = 1$. Indicates that the research topic $o_j$ has the keyword $k_j$. 

3
According to the definitions of the subnet network models of keywords, authors and research topics as well as the relationships among networks, through the relationship between nodes in a subnet and nodes in other subnets, construct the edges between nodes in the subnet and nodes in other subnets, so the mapping relationship between the subnet and other subnets is obtained. In this way, different mapping relationships is possible to be established according to different relations in their own subnets. A 1:m mapping of one node in each subnet and another heterogeneous node subnet is constructed and described. Finally establish the relationship between all sub-networks. The use of a hypergraph structure not only describes any nodes, but also characterizes hidden the relationships between networks, by which hyperedges become an obvious linkage.

4. METHODOLOGY

The main steps of constructing semantic linkage are as follows:

A. The candidate with greater influence will be selected.

The more the candidate’s paper be cited, the greater the influence the candidate will have;

B. the key words used to describe the candidate experts are closely related with the knowledge organization tool term, and the candidate expert is mapped by the knowledge organization tool to terms;

C. according to the constraint conditions of the expert information, the expert relation map is optimized to the most related experts with same topics;

D. According to the relationship between experts’ information, knowledge organization tools and literature resources, experts’ information will be updated automatically;

E. Based on the expert relation map, the graph tools are used to display and monitor the multi-dimensional semantic visualization.

Through the above steps will be looped to construct expert knowledge map. This paper provides a specific implementation technology based on RDF technology for expert discovery and query, which can effectively realize multi-source data integration and data reasoning based on the semantic labels. The suggestions of expert discovery and query based on RDF are as follows:

First, the experts semantic could be labelled by terms in the form of three tuples. Through semantic relations, it is possible to judge the expert's expertise, including the provisions of the static semantic relations and dynamic information from the literature. The cooperative relationship among experts could be detected by co-occurrence calculation. Then, through the semantic category, experts in the same filed could be clustered accurately. All the information above will be stored in the forms of three tuples or XML formats.

Then, the RDF resource description framework is established. In the framework of the system, each expert can be viewed as a resource or entity with its own URI (Uniform Resource Identifier). Using this URI to identify an expert, a detailed description of the attributes of the expert will be displayed. Jena will be used to build expert RDF semantic framework model. Jena is an open source project of Apache, which is used to construct the semantic Web program. It provides a set of tools and Java libraries to help develop semantic web, RDF model. Therefore, it is easy to read and generate RDF files.

Finally, the RDF semantic query was built. SparQL query in any of the three tuple of the information can be replaced by the first letter of the variable. For example the variable telephone information can be defined as: vcard:telephone? Telephone statement can be used to avoid redundant information prefix in the where statement; at the same time it can be carried out with the application of RDF logic reasoning. According to the RDF three tuple information semantic reasoning, such as expert A working in C units and expert B also working in C units, can specify A and B to the relationship between colleagues: < A, is Staff of, C unit >; < B, staffed, C unit >; < ?y, is Staff of, ?y >; < ?z, is Staff of, ?y >; < ?x, is colleague, ?z >; < A, is colleague, B >.

The cooperative group of expert and the research topics are presented in a variety of visual way. From dimensions such as time, theme and relationship, it will be monitored the mapping of the expert knowledge dynamically. The research tendency of an expert can be visualized by the metadata of expert’s literature. The expert's cooperative relationship can be well represented by the line graph,
which means that the nodes in the graph stand for the experts themselves, and the linkage between the nodes indicates the cooperation relationship of the experts. The research domain of experts is essentially a kind of attribute which can be described by color, position and shape in the same canvas. Animation can dynamically display the data in time series to reveal the change of research interests of an expert with time. For all the experts, a large number of visual elements can be described to display these attributes. According to the different needs, the appropriate form of visual elements will provide users with visual methods to ensure the readability graphics.

Term frequency - inverse document frequency (TF-IDF) is a statistical method. It can evaluate the importance of a word in a document in the corpus, the frequency of occurrence of keywords in each author’s literature collections, and the distribution in all authors' collections are also considered. TF indicates the frequency of occurrence of keywords in all authors' papers. The calculation formula is:

$$tf_{iA} = \frac{n_{iA} \cdot \sum n_{iA}}{n}$$

$n_{iA}$ indicates the relative frequency or absolute frequency of the occurrence of the keyword i in the author A. In this paper, the relative frequency is used as the value of TF. $\sum n_{iA}$ indicates the sum of all the keywords of author A. IDF is named inverse document frequency, which is used to measure the distribution of keywords in all papers. The calculation formula is:

$$IDF = \log \left( \frac{N}{N_i} \right)$$

N indicates the number of all papers published by the author, $N_i$ indicates the number of keywords i contained in the literature published by the author, the value of TF-IDF is TF*IDF. In general, the TF-IDF method identifies the similarity between authors based on the frequency of occurrence of keywords in each author's literature collection and the distribution in all authors' literature collections. The author’s keyword strength calculation method based on TF-IDF mainly identifies the similarity between authors by the word frequency of statistical keywords.

LDA is a common topic mining model. It can obtain the probability of authors under certain research topic based on the distribution of authors and research topics. The high probability indicates that the author's research direction matches the research theme of the topic. The LDA model consists of a three-tier structure of words, topics, and literature. The implicit subject of the literature can be extracted automatically from the corpus. LDA indicates the literature as the probability distribution of the topic and the topic as the probability distribution of the words. An author is randomly composed of multiple research topics, and each research topic can be represented by multiple keywords in the author's literature. Therefore, LDA represent an author as the probability distribution of the research topic (author-topic), then each research topic can be used as the probability distribution of keywords, and the author's relationship will be mapped to the level of research topic. LDA topic model is a probabilistic model which is capable of mining implicit topics, make up for the shortcomings of TF-IDF, and can automatically extract the research topic of the literature (the author in this article).

By applying the LDA topic mining method, JS(Jensen-Shannon) divergence is used to calculate the probability distance after obtaining the probability distribution of the author on each research topic, then the similarity between two authors can be obtained. The "S" divergence is a variation based on the KL divergence(Kullback - Leibler divergence), which solves the problem of KL divergence asymmetry. In general, JS divergence is symmetrical, its value is between 0 and 1. Which is defined as follows:

$$JS(P1||P2)=$$
The KL divergence formula is:

$$S_{w_i w_j} (w_i \bullet w_j) = p(w_j) \log \frac{p(w_i)}{p(w_j)}$$

(\(w_i, w_j\) indicate the probability of author i and author j under a topic).

Considering that authors can usually cover one research topic of the LDA and TF-IDF methods, this paper will combine the two methods and carry out linear fusion of similarity matrices, which calculates the author's potential collaboration weight as:

$$PCW = (1-S(A1)) + S(A2)$$

5. Experiment/Test: Taking Oncology Science for example

In this paper, an empirical study was carried out in the field of oncology, which is used to analyzing the reliability and value of the test model.

First of all, all the data was collected from the library literature. Selecting the "Chinese Library Classification" "Oncology" category of literature, the metadata was extracted. The types of literature include journal articles, conference papers, dissertations, science and technology reports, patents, etc. Meta-data fields contained the title, author, organization, key word, classification number, citation, h index, source, etc.. The key word frequency more than 10 will be selected a candidate expert. Then, the candidate whose co-occurrence times more than 10 also will be supposed as a candidate expert. In this way, the relevant experts of the candidate experts are obtained through the calculation of the co-authors of the literature, and it will provide the original data set for the optimization of the relationship. The key words of all the top 5 journals were selected to ensure the data quality.

Then, the key words are mapped into the existing knowledge organization tools such as "medical thesaurus" and the SUMO ontology knowledge base by means of the synonym calculation tool. Taking the semantic categories, each term will be linked to the nearest concept of the KOS. Then, the RDF format is used to represent and reason, by which whether the candidates belong to the same subdivision field or not will be judged.

Then, experts relationship will be adjusted according to KOS. If two experts belong to the same semantic group, they will be recommended even though there is no co-occurrence relationship. If it can't find out the direct peer experts, then it will expand the scope of the term by the semantic relations of the knowledge organization, and get the experts in the most relevant fields. In addition, the information of candidate expert will be associated with the existing expert base, if the two experts belong to the same organization or the same project, it is supposed to be avoided or removed from the list of candidates.

All the 3 levels are working dynamically. Based on the real-time changing of literature, it is possible to update the information respectively after determining the changes in research field and activity of experts. So, any level could be working separately and all the concept crossing level could be linked in semantic.

Finally, the visualization of the expert aggregation is realized. The experimental data are shown by co-occurrence graph, subject classification, expert influence thermodynamic diagram, making the expert aggregation process more accurate and dynamic. Statistical analysis of the co-occurrence of experts indicate the expert cooperation strength. Viewed from the time or the theme, the hidden relationship of the expert’s activity and their influence will be described visually. In order to meet the needs of the crossing-professional fields and expert database updating, the expert database is optimized by threshold. The most related experts (Figure 3 Part A) were found with the back of cancer graph database (Figure 3 Part B).
Figure 3  part A——The results of expert finding part B——The tool of knowledge organization

Table 1. Expert finding experiment test.

| topic               | VSM Cosine similarity | Fusion similarity |
|---------------------|-----------------------|-------------------|
|                     | Recall ratio (%)      | Accurate ratio (%)| F value (%)     |
| Breast cancer       | 41.7%                 | 57.6%             | 48.4%          | 60.6% | 78.8% | 68.5% |
| Endocrine gland cancer | 30%                 | 20%               | 24%            | 50%   | 75%   | 60%   |
| Liver cancer        | 57.1%                 | 61.9%             | 59.4%          | 70.6% | 76.2% | 73.3% |
| Lung cancer         | 41.9%                 | 32.4%             | 36.5%          | 60.6% | 64.9% | 62.7% |

Compared with VSM(vector space model), this method has proved the technical feasibility obviously. VSM is a comparable reference method, and its models represent literatures by vectors, by which a literature is described as a vector of series of keywords. Firstly, the word frequency vector of each literature was valued based on word frequency, then the weight of each word was calculated by TF-IDF, finally, the cosine similarity is used to calculate the similarity value between two literatures.

It can be observed that the fusion similarity matrix obtained by the method in this paper has better clustering effect. F value is also significantly improved. The clustering results obtained by combining LDA and keyword coupling are also improved. For example, the accuracy rate of breast tumors increased by 44.5% above baseline. It indicates that the method has certain practical significance and application value.

6. Conclusion

In this paper, the method of experts’ information aggregation based on knowledge organization semantic relation is introduced. By means of the semantic association between the expert information and the knowledge organization tool, the semantic aggregation, automatic discovery and dynamic updating of the expert information can be realized. The experiment indicates that the combination of LDA and TF-IDF is efficient to identify expert academic authority, professional relevance and research activity influence and it is a promising method for expert finding.
Acknowledgment
2016 National Social Science Fund of China "Scientific Experts Finding based on Knowledge Organization" (16BTQ079)

REFERENCES
[1] Katja Hofmann, Krisztian Balog, Toine Bogers, Maarten de Rijke. Contextual Factors for Finding Similar Experts[J]. Journal of the American Society for Information Science and Technology. 2010, 61(5): 994-1014.
[2] Jianhan Zhu, Dawei Song, Stefan Rueger. Integrating Multiple Windows and Document Features For Expert Finding[J]. Journal of the American Society for Information Science and Technology. 2009, 60(4): 694-715.
[3] Alon Friedman, Richard P. Smiraglia. Nodes and arcs: concept map, semiotics, and knowledge organization[J]. The Journal of Documentation. 2013, 69(1): 27-48.
[4] Silvana Drumond Monteiro, Maria Aparecida Moura. Knowledge Graph and 'Semantization' in Cyberspace: A Study of Contemporary Indexes[J]. Knowledge Organization. 2014, 41(6): 429-439.
[5] Fausto Giunchiglia, Biswanath Dutta, Vincenzo Maltese. From Knowledge Organization to Knowledge Representation[J]. Knowledge Organization. 2014, 41(1): 44-56.