Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Using heterogeneous patent network features to rank and discover influential inventors*

Yong-ping DU†[1], Chang-qing YAO[2], Nan LI[1]

([1]*College of Computer Science, Beijing University of Technology, Beijing 100124, China*)

([2]*Institute of Scientific and Technical Information of China, Beijing 100038, China*)

†E-mail: ypdu@bjut.edu.cn

Received Nov. 16, 2014; Revision accepted May 26, 2015; Crosschecked June 23, 2015

Abstract: Most classic network entity sorting algorithms are implemented in a homogeneous network, and they are not applicable to a heterogeneous network. Registered patent history data denotes the innovations and the achievements in different research fields. In this paper, we present an iteration algorithm called inventor-ranking, to sort the influences of patent inventors in heterogeneous networks constructed based on their patent data. This approach is a flexible rule-based method, making full use of the features of network topology. We sort the inventors and patents by a set of rules, and the algorithm iterates continuously until it meets a certain convergence condition. We also give a detailed analysis of influential inventor's interesting topics using a latent Dirichlet allocation (LDA) model. Compared with the traditional methods such as PageRank, our approach takes full advantage of the information in the heterogeneous network, including the relationship between inventors and the relationship between the inventor and the patent. Experimental results show that our method can effectively identify the inventors with high influence in patent data, and that it converges faster than PageRank.

Key words: Heterogeneous patent network, Influence, Rule-based ranking
doi:10.1631/FITEE.1400394 **Document code:** A **CLC number:** TP391

1 Introduction

Patent data contains rich information and denotes the innovative technology which is being protected. It also represents the competitive picture of the enterprise. More and more research has been focused on patent data.

There are different factors that can be used to assess the importance of the inventors, such as the number of the patents owned. Most of the existing algorithms are based on this idea. One inventor often co-invents the patent and thus appears in many inventor lists, and then he/she will be ranked relatively high. Without considering the order of the inventors, this kind of method is not satisfactory in evaluating the influence of the inventor. Moreover, ranking all of the inventors in different fields would not only require a huge calculation, but also make the results meaningless.

Recently, a lot of research has been performed based on network topology (Baglioni *et al.*, 2012; Zhang *et al.*, 2012), such as ranking entities of networks and qualifying their importance. Most of the algorithms are used for the homogeneous network which has the same node type (Hirsch, 2005; Sun *et al.*, 2009; Sun and Han, 2012). However, the network structure in a real application is often complex and has more than one kind of node type. Moreover, there are different relationships between nodes and they are constructed as a heterogeneous network. A patent network is a typical heterogeneous network. It includes two main entity types, patent and inventor.

* Project supported by the National Science and Technology Support Plan (No. 2013BAH21B02-01), Beijing Natural Science Foundation (No. 4153058), and Shanghai Key Laboratory of Intelligent Information Processing (No. IIPL-2014-004)

We need to adopt an effective method to analyze this kind of network and explore the relationship between the different entity types.

The topic model is a type of statistical model for discovering the topics that occur in a collection of documents. It has been included in the context of natural language processing. A topic represents the generalization and abstraction on the semantic in which the text is contained. To obtain a deep analysis of the interested topics for an influential inventor, each inventor's patents can be expressed as a distribution of multiple topics and the topic model can be used to determine the distribution.

For the topic model probabilistic latent semantic analysis (PLSA), the probability distribution of the topic is a parameter rather than a random variable, and an increase in the parameters may lead to an overfitting problem. The latent Dirichlet allocation (LDA) model proposed by Blei *et al.* (2003) is one of the most perfect topic models. It introduces the super parameter using a Dirichlet prior distribution and transforms the probability distribution into random variables. It avoids the overfitting problem as the super parameter is the only required parameter.

We propose a new rule-based iteration approach to rank the inventors in a patent heterogeneous network and use the patent information and develop a set of rules based on the relationship between the nodes in the patent network. The influence of the inventor is measured and we obtain a ranked inventor list. In addition, we use the topic distribution to discover the interest of the top ranked inventors by the LDA model. Experiments on the real data set show that our method achieves a good performance and that the inventor-ranking method converges faster than the PageRank algorithm.

2 Related work

To analyze the network structure, the most classic ranking algorithm is PageRank (Brin and Page, 1998), which is used to evaluate the importance of a specified web page. PageRank applies link information to calculate the value of a web page and treats it as a ranking factor. The main idea is that the higher the value of a web page linked to, the more important the web page. Kleinberg (1999) first extended the sorting method from a homogeneous network to a heterogeneous network, and proposed the hyperlink-induced topic search (HITS) algorithm, which holds that the interaction between different types of nodes in a heterogeneous network contains rich information and can be used to improve the efficiency of the network sorting algorithm.

Recently, more work has appeared focusing on the graph theory. For instance, Chiang *et al.* (2012) used a social link and local information to find the top-*k* authors in a co-authorship network (Ahmedi *et al.*, 2011) based on a probabilistic model and using random walk. Liu *et al.* (2005) compared the author ranking results between different approaches and concluded that the method based on centrality measurement is effective.

We use the idea of PageRank and HITS to implement inventor ranking. The heterogeneous ranking network is built according to the patent information. There has been some research on the academic heterogeneous network for evaluating the importance of the author. In contrast, we put forward the iterative ranking approach to compute the importance of the inventor using both partnership between the inventors and inventorship with the patents. This is the first time that the relationship based ranking method has been applied in a patent heterogeneous network.

The topics that the influential inventor is interested in can be discovered using the topic model. The keyword distribution can be used to represent the topic. Latent semantic analysis (LSA) is based on a linear algebra proposed by Wang *et al.* (2006). It uses a dimension reduction method, singular value decomposition (SVD), to determine the semantic structure of the documents, and implements semantic analysis in the low-dimensional semantic space. LSA can simulate the human's comprehension when the semantic space dimension is similar to a human's cognition (Blei, 2012). In other words, it can translate surface text information into a deep representation (Zelikovitz and Hirsh, 2004).

The probability model PLSA proposed by Hofmann (1999) is based on the maximum likelihood method and the generative model. PLSA follows the idea of LSA on dimension reduction. The text is represented as high-dimensional data by a commonly

used method such as the TF-IDF (term frequency–inverse document frequency) approach. However, the number of topics is limited and it is denoted by the low-dimensional semantic space. The topic mining method maps the document from the high-dimensional space to the low-dimensional semantic space by dimension reduction. PLSA has many applications such as information retrieval and machine learning.

The LDA model introduces the Dirichlet prior distribution based on PLSA. Blei (2012) pointed out that no uniform probability model is used in the probability calculation in PLSA. At the same time, the large number of parameters will lead to the overfitting problem, and it is difficult to distribute probabilities to the documents that are not included in the training data set. LDA introduces the super parameters to build the Bayes model by three-layer document-topic-word (Tang and Yang, 2012). Each document may be viewed as a mixture of various topics. This is similar to PLSA except that the topic distribution is assumed to have a Dirichlet prior in LDA.

After discovering the important inventors by our ranking approach, we can analyze the patent data of these inventors to acquire the topics they are interested in by the topic model. The LDA model has been widely used in text analysis. The semantic information can be achieved and represented by keyword distribution. The topic distributions of the inventor are also analyzed by the LDA model and the keywords are extracted to denote the interest of the inventor, especially for the top ranked inventors.

3 Inventor-patent heterogeneous network

The data set depicting the heterogeneous network is built from the Chinese patent data and the time span is from Jan. 1, 2008 to Dec. 30, 2008. We selected the patents and the related inventors from the Medicine field and the statistics is shown in Table 1. The data model of the inventor-ranking framework is shown in Fig. 1.

The co-inventor network $G_I=(V_I, E_I)$ is the weighted undirected graph between inventors. V_I denotes the set of inventors and E_I denotes the set of edges, representing the partnership of inventors. The set of inventors can be represented as $V_I=(I_1, I_2, \ldots, I_n)$ and the inventor number is $n=|V_I|$. The weight of edge denotes the number of cooperations between inventors.

Table 1 Statistics for the patent data

Parameter	Value
Number of inventors	4787
Number of patents	2096
Number of co-inventorships	17807
Number of inventorships	5640
Time interval	Jan. 1, 2008 to Dec. 30, 2008

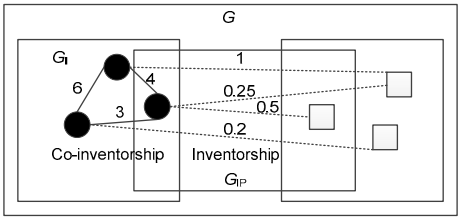

Fig. 1 Data model of the inventor-patent heterogeneous network

The inventor is represented by a black circle and the patent by a square. The network G is constituted by two sub-networks: co-inventor network G_I and inventor network G_{IP}

The inventor network $G_{IP}=(V_{IP}, E_{IP})$ is also the weighted graph representing the relationship between the patent and the inventor, where V_{IP} is the union set of inventors and patents, defined as $V_{IP}=V_I \cup V_P$. Edges in E_{IP} connect each patent with all of its inventors. The weight of the edge is decided by the inventor's order in the inventor list, and the value is $1/r$ when the order is r. For example, the weight value is 0.5 when the order of the inventor is 2.

G_I can be represented by matrix $\boldsymbol{M}_{II}$ where the element $m_{i,j}$ represents the number of cooperations between inventor i and inventor j. G_{IP} can be expressed as matrix $\boldsymbol{M}_{IP}$ and the element $m_{i,p}$ is computed as shown in Eq. (1). The matrices $\boldsymbol{M}_{II}$ and $\boldsymbol{M}_{IP}$ are illustrated in Fig. 2.

$$m_{i,p}=\frac{1}{\mathrm{order}_p(i)}, \tag{1}$$

where $\mathrm{order}_p(i)$ represents the order of inventor i in the inventor list of patent p, and its value is varied from 1 to the total inventor number n.

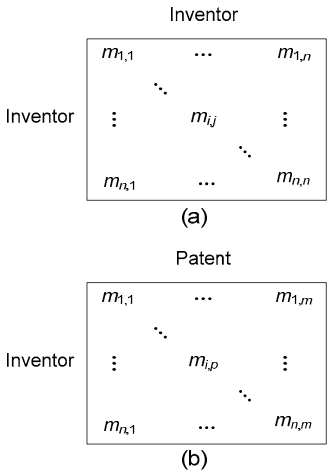

Fig. 2 Matrices M_{II} (a) and M_{IP} (b)

4 Algorithm of discovering the influential inventor

4.1 Rule-based ranking

Our method is based on a set of rules to rank the inventor entity, and the ranking result is updated by a recursive ranking process. The ranking value of the inventor is determined by two factors, the rank value of the co-inventor and the rank value of the patents owned by the inventor. The parameters (w_{II}, w_{IP}) are the weights for these two factors, ranging from 0 to 1. They are both set to be 0.5 in the experiment.

The three rules used are described as follows:

Rule 1 Highly ranked inventors tend to co-invent with other highly ranked inventors.

We obtain the rank value of inventor k by

$$\mathrm{RankInventor}_i(k)=\alpha\left[\sum_{r=1}^{n}M_{\mathrm{II}}(k,r)\mathrm{RankInventor}(r)\cdot w_{\mathrm{II}}+\mathrm{RankInventor}_{i-1}(k)\cdot(1-w_{\mathrm{II}})\right]. \tag{2}$$

The rank value of the inventor is determined by the co-inventor's rank value and his/her own rank value in the previous iteration, which is labeled as $\mathrm{RankInventor}_{i-1}(k)$. Here, r denotes the co-inventor of inventor k.

Rule 2 Highly ranked inventors generally invent highly ranked patents. The rank value of the patent is computed according to its inventor's rank value. We obtain the rank value of patent j by

$$\mathrm{RankPatent}(j)=\frac{\beta\sum_{k=1}^{n}M_{\mathrm{PI}}(j,k)\mathrm{RankInventor}(k)}{\mathrm{RankPatent}_{\max}(M)}, \tag{3}$$

where k denotes the inventor of patent j. The rank value of the patent is determined by all its inventors' rank scores. The matrices $\boldsymbol{M}_{\mathrm{PI}}$ and $\boldsymbol{M}_{\mathrm{IP}}$ are symmetric, M is the patent data set, and $\mathrm{RankPatent}_{\max}(M)$ is the maximum rank value on the set of total patents.

Rule 3 Highly ranked patents are invented by highly ranked inventors. The rank value of the inventor is determined by the rank value of his/her own patent. We update the rank value of inventor r by

$$\mathrm{RankInventor}_i(r)=\gamma\left[\sum_{j=1}^{m}M_{\mathrm{IP}}(r,j)\mathrm{RankPatent}(j)\cdot w_{\mathrm{IP}}+\mathrm{RankInventor}_{i-1}(r)\cdot(1-w_{\mathrm{IP}})\right], \tag{4}$$

where j denotes the patent of inventor r and m denotes the patent number of inventor r. Eq. (4) illustrates that the rank value of the inventor is updated according to the rank score of the patent the inventor owns and his/her rank value in the previous iteration, labeled as $\mathrm{RankInventor}_{i-1}(r)$.

There are three parameters α, β, and γ in the above three rules, which determine the importance of the different rules, and their values range from 0 to 1.

Fig. 3 shows our approach for inventor-ranking. During the initialization process, there are many approaches that we can choose to give the initial rank value of the inventor. However, the choice does not strongly affect the final results. We rank the inventors initially by the ratio of the patent number the inventor owns to the total patent number. The ranking rules will be implemented iteratively. Users can choose the number of iterations and the difference threshold between the current iteration and the previous iteration to judge when to stop. The difference computation is shown as follows:

$$\Delta(t,t+1)=\frac{1}{|V|}\sum_{i=1}^{|V|}|\,\mathrm{rank}(i,t+1)-\mathrm{rank}(i,t)\,|, \tag{5}$$

where $|V|$ is the number of vertices in network G and $\mathrm{rank}(i, t+1)$ is the ranking score of vertex i in the $(t+1)$th iteration.

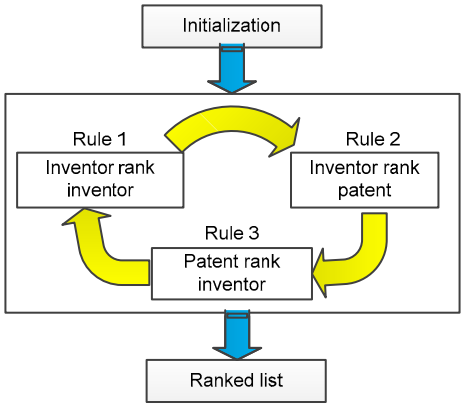

Fig. 3 The rule-based inventor-ranking process

The algorithm of inventor-ranking is described in Algorithm 1. After several iterative ranking experiments by setting different parameter values for α, β, and γ, we find that $\Delta(t, t+1)$ will no longer change when it reaches 0.04. Thus, we set the threshold δ to 0.04 in the final experiment.

4.2 Evaluation

Here we use the correlation metric to assess the different ranking algorithms. For two ranking algorithms a and b, $t_a(n)$ and $t_b(n)$ represent the two sets containing the top n entities ranked by algorithms a and b, respectively. The correlation of two ranking algorithms is defined as

$$S(a,b)=\frac{\left|t_a(n)\cap t_b(n)\right|}{\left|t_a(n)\cup t_b(n)\right|}. \tag{6}$$

We take the order of the entity element in $t_a(n)$ and $t_b(n)$ into consideration. The correlation of the two algorithms a and b can be calculated as

$$\mathrm{SP}(a,b)=\frac{\sum_{i=1}^{\left|t_a(n)\cap t_b(n)\right|}\left|F(t_a(n),i)-F(t_b(n),i)\right|}{\left|t_a(n)\cup t_b(n)\right|}, \tag{7}$$

where $F(t_a(n), i)$ denotes the order of element i in set $t_a(n)$.

This method can be used to evaluate the effectiveness of the ranking algorithm, and we apply it to compare our inventor-ranking method with PageRank in the final experiment.

Algorithm 1 Rule-based inventor-ranking

1. Build two matrices $\boldsymbol{M}_{\mathrm{II}}$ and $\boldsymbol{M}_{\mathrm{IP}}$.
   Inventor-inventor matrix $\boldsymbol{M}_{\mathrm{II}}$ (Fig. 2), in which each element $m_{i,j}$ represents the number of cooperations between inventors i and j.
   Inventor-patent matrix $\boldsymbol{M}_{\mathrm{IP}}$ (Fig. 2), in which each element $m_{i,p}$ represents the importance of patent p to inventor i and it is calculated using Eq. (1).
2. Rank the inventor using Eq. (2) to calculate the rank value of each inventor k and set the value of parameter α.
3. Rank the patent using Eq. (3) to calculate the rank value of each patent j and set the value of parameter β.
4. Rank the inventor using Eq. (4) to obtain the updated rank value of each inventor r and set the value of parameter γ.
5. Calculate the rank value difference $\Delta(t, t+1)$ between the two adjacent iterations t and $t+1$ using Eq. (5).
6. Repeat steps 2–5 until the difference value $\Delta(t, t+1)$ is less than threshold δ.

5 Topic distribution of the influential inventor

5.1 LDA topic model

The LDA model is a hierarchical Bayesian model and it has the following three layers (Blei *et al.*, 2003):

1. Word layer. Word collection $V=\{w_1, w_2, \ldots, w_v\}$ is the set of words excluding the stop word in the corpus.

2. Topic layer. Each topic z_i in topic collection $\varphi=\{z_1, z_2, \ldots, z_k\}$ is a probability distribution based on word set V. It can be represented as a vector $\boldsymbol{\varphi}_k=<p_{k,1}, p_{k,2}, \ldots, p_{k,v}>$ where $p_{k,j}$ denotes the generation probability of word w_j on topic z_k.

3. Document layer. Each document is represented as the frequency vector $\boldsymbol{d}_i=<\mathrm{tf}_{i,1}, \mathrm{tf}_{i,2}, \ldots, \mathrm{tf}_{i,V}>$, where $\mathrm{tf}_{i,j}$ denotes the occurrence of word j appearing in document i. The document set can be represented as $\boldsymbol{\theta}=<\theta_1, \theta_2, \ldots, \theta_D>$ and vector $\boldsymbol{\theta}_d=<p_{d,1}, p_{d,2}, \ldots, p_{d,k}>$ is represented as a topic. $p_{d,z}$ denotes the generation probability of topic z in document d.

The graph model of LDA is shown in Fig. 4. The LDA model uses the Dirichlet distribution as the prior distribution for the probability topic model.

5.2 Inventor model based on LDA

The standard LDA model is based on the three-layer document-topic-word Bayesian model (Fig. 5). The topic of the inventor can be defined as the degree of attention to different topics. Thus, the inventor-topic probability distributions in the topic model represent the inventor's interest on different topics.

We merged all of the patent data of the inventor into one document to construct the topic model. The probability distribution of the inventor to topic can be achieved to generate the topic model. Then the three-layer document-topic-word model is transferred to the inventor-topic-word model (Fig. 6).

For inventor collection $U=\{u_1, u_2, \ldots, u_m\}$, each inventor u_i is represented by the corresponding patent data $\{\boldsymbol{f}_{u_i,1}, \boldsymbol{f}_{u_i,2}, \ldots, \boldsymbol{f}_{u_i,n}\}$ and $\boldsymbol{f}_{u_i,j}$ denotes the word vector, which is shown as the matrix $\boldsymbol{U}$ in Fig. 6. The vector $\boldsymbol{\rho}_{u_i}=\{p_{u_i,1}, p_{u_i,2}, \ldots, p_{u_i,k}\}$ for inventor u_i denotes the topic probability distribution (Fig. 7a). $p_{u_i,z}$ is the generated probability of topic z to inventor u_i and it can be used to express the inventor's interest. The word w_i can also be represented by a vector $\boldsymbol{\theta}_{w_i}=\{q_{w_i,1}, q_{w_i,2}, \ldots, q_{w_i,k}\}$ in the LDA model (Fig. 7b). Here, $q_{w_i,z}$ is the generated probability of word w_i to topic z.

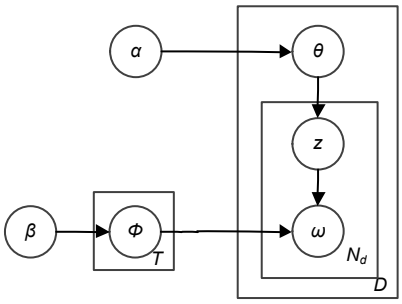

Fig. 4 Graph model of latent Dirichlet allocation (LDA)
D represents the whole document set; N_d is the word collection of document d; α and β denote the prior knowledge of document-topic probability distribution θ and document-word probability distribution Φ; θ and Φ satisfy the polynomial distribution; α and β satisfy the Dirichlet distribution

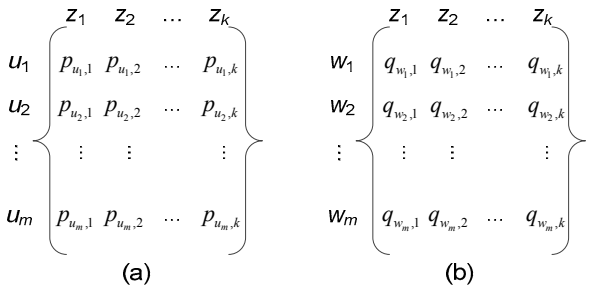

Fig. 7 Matrix representation in the latent Dirichlet allocation (LDA) model: (a) inventor-topic; (b) word-topic

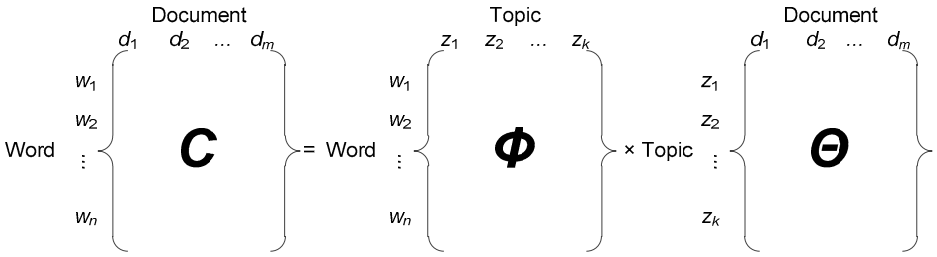

Fig. 5 Standard three-layer document-topic-word model of latent Dirichlet allocation (LDA)

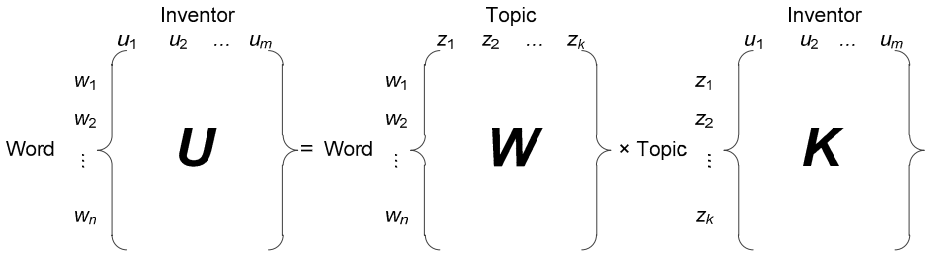

Fig. 6 Updated three-layer inventor-topic-word model of latent Dirichlet allocation (LDA)

6 Experiments and evaluation

6.1 Inventor-ranking result evaluation

The iterative process of inventor-ranking is shown in Fig. 8. During the initialization, we estimate the inventor importance by the percentage of patents owned. However, due to the limited number of patents owned, most inventors have a very low value, ranging from 0 to 0.01. After the first iteration, more than 50% of inventors are assigned a relatively high rank value. After the 7th iteration, the distribution of rank value becomes consistent and stable, and it almost follows the Gaussian distribution.

Table 2 shows the top 10 inventors ranked by our inventor-ranking method and PageRank. These two algorithms have a different focus in the list. The PageRank algorithm gives higher rank values to those inventors who appear in different inventor lists more frequently as the co-inventor. In contrast, our inventor-ranking approach considers not only the co-inventorship but also the number of patents the inventor owns as the primary inventor. It can be concluded that our method works better as it makes full use of the available features in the patent network.

We give more precedence to the first inventor in our ranking approach. The statistics of patent data of three typical inventors is shown in Table 3, which illustrates the effectiveness of the ranking algorithm.

The inventor Wei Zhu has the most owned patents and in most of them Wei Zhu is the first inventor. Our ranking algorithm and PageRank both rank Wei Zhu as the top inventor. In contrast, although Shuren Guo owned more patents than Lijuan Wang, Lijuan Wang ranked higher than Shuren Guo in our ranking list because in fewer patents Shuren Guo is the first inventor. This is due to the fact that our algorithm considers not only the patent number but also the order of the inventor, which is more truthful.

After obtaining the ranking list, we compared the results of inventor-ranking and the PageRank algorithm by choosing different N's for the top N inventors after 20 iterations. The parameter values of α, β, and γ also varied. Table 4 shows the results evaluated by metrics S and SP.

Our approach is flexible in assigning importance to different ranking rules. The larger value of α means more importance is assigned to Rule 1, so the co-inventorship is the primary factor for the final ranking, which is similar to the idea of PageRank. Therefore, the ranking results obtained by these two ranking algorithms are similar when α, β, and γ are set to be 0.8, 0.1, and 0.1 respectively, and S and SP are more than 0.64 with the highest S being 0.81. We can set larger values to parameters β and γ, taking inventor order into consideration. Meanwhile, our approach allows users to take a multi-typed entity and relationship into consideration, based on the given set of rules.

Since our experiment is applied on patent data from 2008, we give data analysis in the following years from 2009 to 2013 to verify the top ranked inventors. Fig. 9 shows the number of patents for the top three inventors Wei Zhu, Lijuan Wang, and Junbao He. They all obtained authorized patents continuously and most other inventors ranked lower had no or fewer patents in the following years, verifying that our ranking results are credible. In addition, we found that the patent reference frequencies of the top three inventors are much more than those of the other inventors who had no or fewer references.

Fig. 10 shows the convergence rate of these two ranking algorithms. Our inventor-ranking method converges faster than the PageRank algorithm.

6.2 Topic distribution results of the top ranked inventors

To analyze the patent topic of the influential inventors, we conducted the experiment using the LDA model. The patent data of the top 10 inventors ranked by our inventor-ranking algorithm was collected. The patent data was selected from the medicine field and the statistics of the data is as described in Section 3. After word segmentation, stop words removal, keyword extraction, and weight calculation, we obtained the word vector of each inventor. The LDA user model was applied to implement topic mining and keyword extraction.

The Gibbs sample method was used to deal with the LDA model and the parameters were set according to experience. α was set to $50/T$ and β was set to 0.01. Here, T denotes the topic number. We observed the differences of keyword distribution when the topic number varied from 3 to 10, and it was found that 4 is a reasonable number. The topic distribution probability of different inventors can be achieved after applying the LDA model on the patent data, and

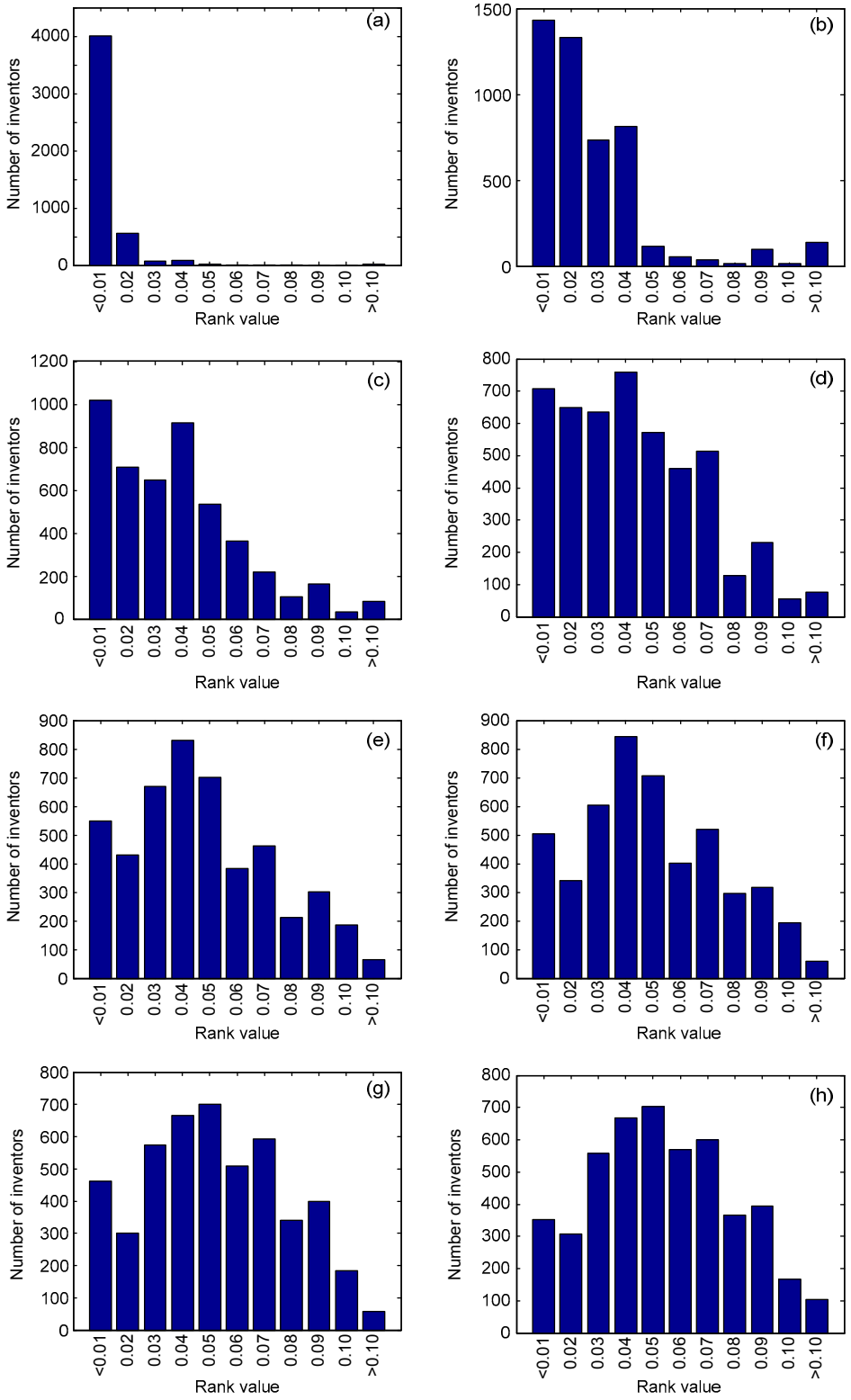

Fig. 8 Distribution of the rank values in each iteration: (a) initial phase; (b) 1st iteration; (c) 2nd iteration; (d) 3rd iteration; (e) 4th iteration; (f) 5th iteration; (g) 7th iteration; (h) 9th iteration

Table 2 The top 10 inventors ranked based on our inventor-ranking method and the PageRank algorithm

Rank	Inventor	
	Inventor-ranking	PageRank
1	Wei Zhu	Wei Zhu
2	Lijuan Wang	Lijun Yan
3	Junbao He	Shuren Guo
4	Lijun Yan	Yong Wang
5	Rongling Wu	Binwen Liang
6	Binwen Liang	Chao Wang
7	Junyong An	Marcus Dirk
8	Chao Wang	Wei Li
9	Zhenwen Duan	Lijuan Wang
10	Shuren Guo	Gang Zhao

Table 3 Statistical patent data of three inventors

Inventor	Number of co-invented patents	Number of patents as the first inventor
Wei Zhu	67	66
Lijuan Wang	42	37
Shuren Guo	51	10

Table 4 Evaluation results by different parameter values

N	S			SP		
	$\alpha=\beta=\gamma=1.0$	$\alpha=0.8$, $\beta=\gamma=0.1$	$\alpha=\beta=0.1$, $\gamma=0.8$	$\alpha=\beta=\gamma=1.0$	$\alpha=0.8$, $\beta=\gamma=0.1$	$\alpha=\beta=0.1$, $\gamma=0.8$
10	0.42	**0.68**	0.38	0.37	**0.64**	0.36
20	0.45	**0.74**	0.44	0.41	**0.71**	0.40
50	0.54	**0.81**	0.53	0.50	**0.77**	0.50
80	0.67	**0.78**	0.62	0.60	**0.72**	0.57
100	0.65	**0.79**	0.63	0.59	**0.75**	0.60

Bold values are the highest for each N

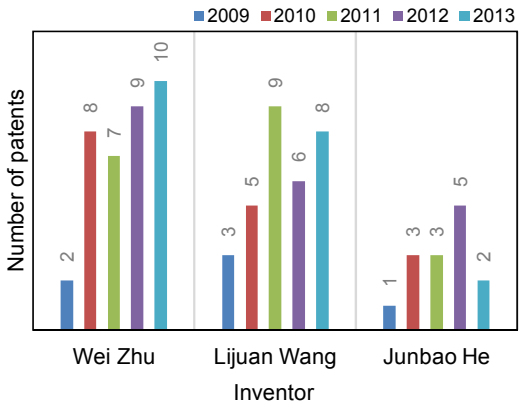

Fig. 9 Number of authorized patents for the top three inventors from 2009 to 2013

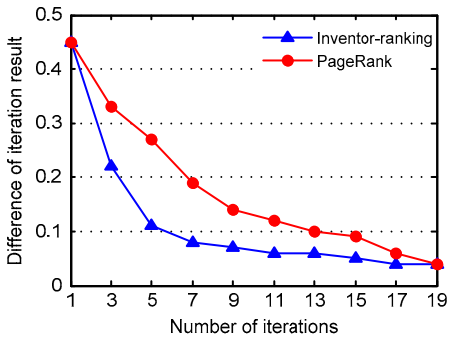

Fig. 10 Convergence of the ranking computation

then we obtained the topic-inventor probability distribution. The topic-inventor model was generated by the matrix as shown in Fig. 7. The distribution of the top 10 inventors is shown in Table 5. The probabilities contributed to the four topics of each inventor.

Figs. 11a–11c show the keyword distribution results for the top three inventors. There are four groups of keywords for each inventor on four topics. They denote the related interest of the inventor. For example, the four topics of Wei Zhu (Fig. 11a) are "孕妇微量元素补充" (microelement supplement for pregnant women), "妇女和孕妇的保健" (health care for gravida and women), "孕期补充维生素" (vitamin supplement of pregnancy), and "记忆缺陷" (memory defect). The keyword distribution results show the focused subjects of the inventor.

Table 5 Topic distribution probability of the top 10 inventors

Inventor	Probability			
	Topic 1	Topic 2	Topic 3	Topic 4
Wei Zhu	0.316	0.207	0.239	0.237
Lijuan Wang	0.277	0.213	0.248	0.261
Junbao He	0.248	0.258	0.253	0.241
Lijun Yan	0.278	0.23	0.22	0.271
Rongling Wu	0.245	0.259	0.264	0.231
Binwen Liang	0.252	0.305	0.231	0.212
Junyong An	0.175	0.296	0.256	0.272
Chao Wang	0.280	0.230	0.256	0.233
Zhenwen Duan	0.273	0.303	0.179	0.244
Shuren Guo	0.321	0.229	0.217	0.232

The more influential inventor generally owns the more patents. Among the top 100 ranked inventors, we used the statistics on the top three (Fig. 12a) and last three (Fig. 12b) inventors to verify this idea.

Fig. 11 Keyword distribution of the three inventors (in Chinese): (a) Wei Zhu; (b) Lijuan Wang; (c) Junbao He

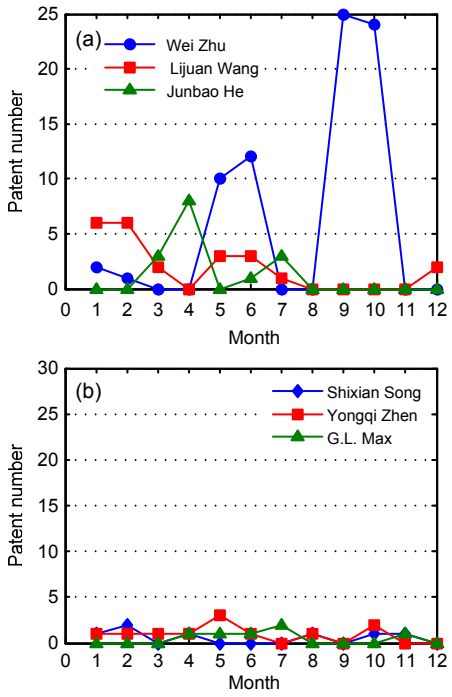

Fig. 12 The patent distribution for the top three (a) and last three (b) among the top 100 inventors

The top three inventors obtained new patents in many different months, especially the top inventor Wei Zhu. In contrast, the last three inventors had few patents within the same time period.

7 Conclusions

We propose a new inventor-ranking method based on the heterogeneous patent network. The rank value of the inventor is calculated iteratively by setting up rules on the relationship between inventors, as well as the relationship between the inventor and the patent. The experiment results show that our ranking approach is effective for the patent network, and it is more efficient than the PageRank algorithm. Also, the set of ranking rules can be adjusted flexibly for different applications. The topic distributions of the inventor are also analyzed by the LDA model, and they can denote the interest of the inventor, especially for the top ranked inventors.

In the future, the effect of the ranking method on larger data sets should be tested. Also, it is necessary to develop a reasonable and effective algorithm for evaluating and validating the results in practical applications. Currently, there is some research that focuses on patent evaluation. However, some of the technical parameters and economic information data are difficult to obtain. The current patent database lacks some of the statistical indicators for evaluating the patent, such as degree of improvement in product quality. Our present research is focused on inventor ranking and patent evaluation will be our future work.

References

Ahmedi, L., Abazi-Bexheti, L., Kadriu, A., 2011. A uniform semantic web framework for co-authorship networks. IEEE 9th Int. Conf. on Dependable, Autonomic and Secure Computing, p.958-965. [doi:10.1109/DASC.2011.159]

Baglioni, M., Geraci, F., Pellegrini, M., *et al.*, 2012. Fast exact computation of betweenness centrality in social networks. Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining, p.450-456. [doi:10.1109/ASONAM.2012.79]

Blei, D., 2012. Probabilistic topic models. *Commun. ACM*, **55**(4):77-84. [doi:10.1145/2133806.2133826]

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**(3):993-1022.

Brin, S., Page, L., 1998. The anatomy of a large-scale hyper textual web search engine. *Comput. Networks ISDN Syst.*, **30**(1-7):107-117. [doi:10.1016/S0169-7552(98)00110-X]

Chiang, M.F., Liou, J.J., Wang, J.L., *et al.*, 2012. Exploring heterogeneous information networks and random walk with restart for academic search. *Knowl. Inform. Syst.*, **36**(1):1-24. [doi:10.1007/s10115-012-0523-8]

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. *PNAS*, **102**(46):16569-16572. [doi:10.1073/pnas.0507655102]

Hofmann, T., 1999. Probabilistic latent semantic indexing. Proc. 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.50-57. [doi:10.1145/312624.312649]

Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. *J. ACM*, **46**(5):604-632. [doi:10.1145/324133.324140]

Liu, X., Bollen, J., Nelson, M.L., *et al.*, 2005. Co-authorship networks in the digital library research community. *Inform. Process. Manag.*, **41**(6):1462-1480. [doi:10.1016/j.ipm.2005.03.012]

Sun, Y., Han, J., 2012. Mining heterogeneous information networks: principles and methodologies. *Synth. Lect. Data Min. Knowl. Disc.*, **3**(2):46-89. [doi:10.2200/S00433ED1V01Y201207DMK005]

Sun, Y., Han, J., Zhao, P., *et al.*, 2009. RankClus: integrating clustering with ranking for heterogeneous information network analysis. Proc. 12th Int. Conf. on Extending Database Technology: Advances in Database Technology, p.565-576. [doi:10.1145/1516360.1516426]

Tang, X.N., Yang, C.C., 2012. TUT: a statistical model for detecting trends, topics and user interests in social media. Proc. 21st ACM Int. Conf. on Information and Knowledge Management, p.972-981. [doi:10.1145/2396761.2396884]

Wang, X.H., Sun, J.T., Chen, Z., *et al.*, 2006. Latent semantic analysis for multiple-type interrelated data objects. Proc. 29th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.236-243. [doi:10.1145/1148170.1148214]

Zelikovitz, S., Hirsh, H., 2004. Using LSI for text classification in the presence of background text. Proc. 10th Int. Conf. on Information and Knowledge Management, p.113-118.

Zhang, J., Ma, X., Liu, W., *et al.*, 2012. Inferring community members in social networks by closeness centrality examination. Proc. 9th Web Information Systems and Applications Conf., p.131-134. [doi:10.1109/WISA.2012.52]