Expertise Search in a Time-varying Social Network

Yize Li and Jie Tang
Computer Science and Technology, Tsinghua University
FIT 1-308, Tsinghua University, Beijing 100084, China
liyize@gmail.com, jietang@tsinghua.edu.cn

Abstract

This paper is concerned with the problem of expertise search in a time-varying social network. Previous research work on expertise search, aiming at finding the most important/authoritative objects, usually ignores an important factor - temporal information, which reveals a huge amount of information contained in large document collections. Many real-world applications, for example reviewers matching for academic papers and hot-topic finding from newsgroup posts need to consider the evolution of information over times. In this work, we propose a unified model by integrating the temporal information into a random walk model. Specifically, the time information is modelled in a forward-and-backward propagation process in the random walk. The proposed model has been applied to expertise search in an academic social network. Experimental results show that the proposed approach can significantly outperform the baseline methods of using the language model (2.0% in terms of MAP) and the traditional PageRank algorithm (17.2% in terms of MAP).

1. Introduction

Expertise search is aimed at finding not only relevant but also authoritative objects. It has become one of the most important tasks from the emergence of the Web. A variety of expertise search techniques have been proposed for addressing this problem.

With the rapid development of the internet, characteristics of information on the Web (in particular Web 2.0) have been changing dramatically in several dimensions: from homogeneously to heterogeneously; from statically to dynamically; and from separately to intensively socially. An ideal solution to expertise search in the new setting should consider all the quickly changing characteristics. Unfortunately, most of existing approaches for expertise search (partly) ignore the change trends of different information. We here use an example of expertise search in academic research area to demonstrate the problem.

In the field of academic research, several search engines have been developed such as Citeseer, Google Scholar, Libra, and Arnenminer.org. The most important services in these search engines include finding experts, authoritative papers, and authoritative conferences. Figure 1 shows an example result of expert finding by one of the mentioned search engines. The user wants to find a reviewer for an academic paper. Thus the query is the title of the paper "The boosting approach to machine learning: An overview" and the first returned expert is "J. Ross Quinlan". By carefully analyzing the research career of "J. Ross Quinlan", we found that most of his research papers have been published before 2000 and after that he shifted his interest to business, which implies that "J. Ross Quinlan" may not be the most appropriate reviewer to the paper.

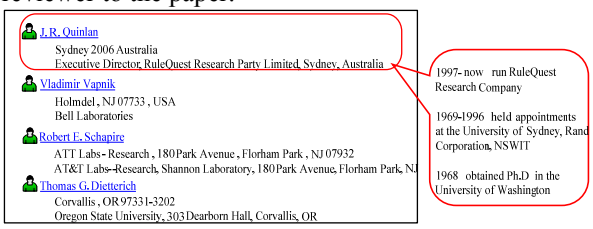


Figure 1. **Example of reviewer finding**

From this example, we can also see that it is highly ineffective to use traditional methods for expertise search in the fast changing social networks due to the natural disadvantages of the methods: (1) most of the existing methods view the different types of objects (e.g., papers, conferences, and authors) as a homogeneous object. The relationships between different types of objects are thus also viewed with the same weight, which would make the random walk style methods (e.g., PageRank [19]) result in unsatisfactory results; (2) most of the methods focus on finding the "general" expertise objects and ignore the temporal information and trends. In fact, the ranking list always changes over time, not only for a specific person or paper, but also at a more global level. The temporal information (even strongly) affects the results of expertise search (as shown in Figure 1).

In this paper, we aim to conduct a thorough investigation on the problem. First, we formalize the heterogeneous social network within a random walk

model. In the model, different types of objects and different types of relationships are modelled as different types of nodes and different types of links. Next, for modelling the time-varying information, we conduct a temporal forward-and-backward propagation in the random walk model. Experimental results indicate that our method significantly outperforms the baseline methods of using the language model (2.0% in terms of MAP) and the PageRank (17.2%).

The rest of the paper is organized as follows. In Sec. 2, we introduce notations and preliminary knowledge about graph model and random walk. In Sec. 3, we propose two methods to combine temporal information into the random walk model. In Sec. 4, we discuss the experimental results and in Sec. 5, we review related work. Finally, we conclude the paper.

2. Preliminary

We first briefly introduce the heterogeneous graph model and a basic random walk model.

2.1. Graph model

A heterogeneous network can consist of two components: time-intra social network and time-inter social network. For simplicity, we will use an academic social network as the example in the following explanation. We will describe the time-intra social network in this section and describe the time-inter social network in the next section.

In the academic network, the time-intra network is composed of three composite networks. At the centre is the directed graph of paper citations $G_c = (V_c, E_c)$, where V_c is the set of all the papers and a directed edge $(d_i, d_j) \in E_c$ suggests the paper d_i cites paper d_j. To model the author-paper publication relationship, we have a bipartite graph $G_{ac} = (V_a \cup V_c, E_{ac})$, where V_a is the set of authors, V_c is the set of papers, and the author-paper relationship is recorded in the edge-set E_{ac}. The relationship between papers and publication locations is modelled by the bipartite graph $G_{lc} = (V_l \cup V_c, E_{lc})$, where V_l is the set of publication locations, V_c is the set of papers, and the publication location-paper relationship is recorded in E_{lc}. Essentially, if a_i is the author of paper d_j, then there is an edge $\{a_i, d_j\} \in E_{ac}$, similarly for a paper d_j published at a location c_k we have an edge $\{c_k, d_j\} \in E_{lc}$.

We combine these different graphs to form a heterogeneous graph centred by the citation network: $G = (V_c \cup V_a \cup V_l, E_c \cup E_{ac} \cup E_{lc})$. In addition, for the sake of random walk, we represent each undirected edge in the bipartite graph as two directed edges, i.e. $\{v_i, v_d\} = (v_i, v_d) \cup (v_d, v_i)$. Further, we define a graph which describes the transition probabilities between different types of nodes (cf. Figure 2). Clearly, we need $\lambda_{cc} + \lambda_{ca} + \lambda_{cl} = 1$. We also define $\lambda_{ac} = \lambda_{lc} = 1$. This transition graph formalizes a random surfer's behaviour as follows. A random surfer will have the λ_{cc} probability to stay in the paper citation network, and will have λ_{ca} and λ_{cl} probabilities to find authors and conferences related to the paper. Since the paper citation network provides the objective measure for other information sources, our main focus is on this network. Thus, our model assumes that users will directly jump back to the paper citation network from the other two networks with probability $\lambda_{ac} = \lambda_{lc} = 1$.

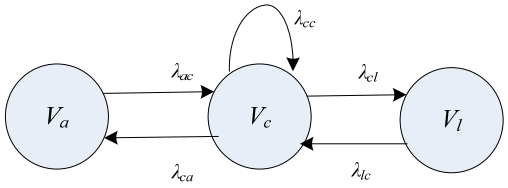


Figure 2. **Transition probability**

2.2. Random walk

We conduct a random walk over the time-intra network. The random walk transforms the entire network to a stochastic matrix (M), defined as follows:

Let $d_i, d_j \in V_c$:

$$P(d_j \mid d_i) = \lambda_{cc} \times \frac{1}{Out\_Degree(d_i \to V_c)} \quad (1)$$

Let $a_m \in V_a$:

$$P(a_m \mid d_i) = \lambda_{ca} \times \frac{1}{Out\_Degree(d_i \to V_a)} \quad (2)$$

$$P(d_i \mid a_m) = \lambda_{ac} \times \frac{1}{Out\_Degree(a_m \to V_c)} \quad (3)$$

where Out_Degree($d_i \to V_c$) is the number of directed edges from d_i to the nodes in V_c. Out_Degree($d_i \to V_a$) and Out_Degree($a_m \to V_c$) are similarly defined. Further, we can define $P(c_k \mid d_i)$ and $P(d_i \mid c_k)$ similarly where $c_k \in V_l$.

DEFINITION 1. *The rank vector r is a stationary distribution of the matrix M*:

$$r = Ar, A = M^T$$

Similar to the page rank algorithm, we introduce a random jump parameter β, which allows a surfer to randomly jump to any node in the network:

$$M' = (1-\beta)M + \beta E, E = (1/n, \ldots, 1/n)^T(1, \ldots, 1)$$

where n is the number of nodes in the network G, that is $n = |V_c| + |V_a| + |V_l|$. Given this, we can use an iterative method to find the rank vector r. The transition probability between two paper d_i and d_j becomes:

$$P'(d_j \mid d_i) = \lambda_{cc} \times \frac{(1-\beta)}{Out\_Degree(d_i \to V_c)} + \frac{\beta}{n} \quad (4)$$

Other probabilities can be similarly defined as:

$$P'(a_j \mid d_i) = \lambda_{ca} \times \frac{(1-\beta)}{Out\_Degree(d_i \to V_a)} + \frac{\beta}{n} \quad (5)$$

$$P'(c_k \mid d_i) = \lambda_{cl} \times \frac{(1-\beta)}{Out\_Degree(d_i \to V_l)} + \frac{\beta}{n} \quad (6)$$

$$P'(d_i \mid a_m) = \lambda_{ac} \times \frac{(1-\beta)}{Out\_Degree(a_m \to V_c)} + \frac{\beta}{n} \quad (7)$$

$$P'(d_i \mid c_k) = \lambda_{lc} \times \frac{(1-\beta)}{Out\_Degree(c_k \to V_c)} + \frac{\beta}{n} \quad (8)$$

Note that this random walk corresponds to the following authority ranking schema where the ranking of each node is determined by its neighbours, i.e.

$$r(d_i) = \sum_{(d_j, d_i) \in E_c} P'(d_i \mid d_j) r(d_j) + \sum_{(a_m, d_i) \in E_{ac}} P'(d_i \mid a_m) r(a_m) + \sum_{(c_k, d_i) \in E_{lc}} P'(d_i \mid c_k) r(c_k) \quad (9)$$

$$r(a_m) = \sum_{(d_i, a_m) \in E_{ac}} P'(a_m \mid d_i) r(d_i) \quad (10)$$

$$r(c_k) = \sum_{(d_i, c_k) \in E_{lc}} P'(c_k \mid d_i) r(d_i) \quad (11)$$

3. Temporal academic search

We propose two ranking methods for academic search for incorporating the temporal information into the random walk model.

3.1. Aggregation method

The basic idea of the first method is straightforward. We separate the heterogeneous social network G into several networks on different time slices $\{G_s\}$. The superscript s denotes the number of the time slice. Next we conduct a random walk on each time-window graph G_s and then combine the results of ranking scores obtained from different time-window graphs.

In this method, we can easily adjust the weights of networks on different time-windows. For example if we are interested in the recent active objects, we can set a higher weight to objects occurring in the current time-window, while a lower weight to objects in previous time-windows. We use a decay factor to control the scores obtained from different windows. Thus, the score of paper d_j can be defined as follows:

$$r'(d_j) = \sum_{s=1}^{t} \alpha^{t-s} \times r(d_j)^s \quad (12)$$

where t is the timestamps of the last time-window; α is the decay factor and α^{t-s} is the factor to penalize score that is obtained from previous time-window s. We can similarly define the final ranking score for author a_i and conference c_k:

$$r'(a_i) = \sum_{s=1}^{t} \alpha^{t-s} \times r(a_i)^s \quad (13)$$

$$r'(c_k) = \sum_{s=1}^{t} \alpha^{t-s} \times r(c_k)^s \quad (14)$$

The disadvantage of the method is that the dependencies between different time-windows are modeled with the same weight, as the parameter α is the only factor used to control the dependency.

3.2. Temporal random walk

The second method is to integrate the time-varying dependencies directly into the random walk. Assuming that we have timestamps $\{1, 2, \ldots, t\}$, we define G_s as the time-intra social network and $G_{(s-1)s}$ as the time-inter network, where $G_{(s-1)s}$ is a bipartite graph between neighbouring time-intra networks $G_{(s-1)}$ and G_s. Again, we introduce two transition probabilities: forward-transition probability λ_f and backward-transition probability λ_b, respectively representing that a surfer in $G_{(s-1)}$ has probability λ_f to random walk to objects in G_s and has probability λ_b to random walk back to objects in a previous time-intra network. Figure 3 shows the graphical representation of the temporal random walk.

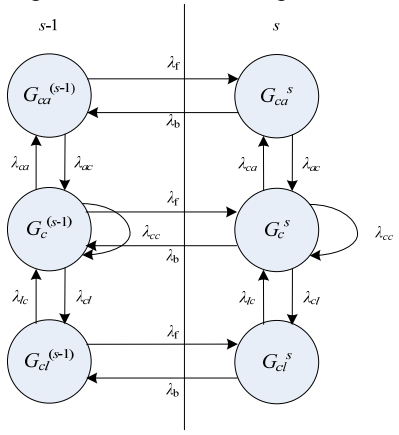


Figure 3. **Temporal random walk**

We see that there is a bipartite temporal-edge between objects when it appears both in (s-1) and s time-windows. Note that an object in this model can appear in more than one time-window networks, e.g., a paper was published at time-window (s-1) and was cited in time-window s. Now for papers, we have $\lambda_{cc} + \lambda_{ca} + \lambda_{cl} + \lambda_f + \lambda_b = 1$. For authors and conferences, we have $\lambda_{ac} + \lambda_f + \lambda_b = 1$ and $\lambda_{lc} + \lambda_f + \lambda_b = 1$, respectively.

Intuitively, we use $(\lambda_f + \lambda_b)$ to control how important the temporal information will affect the final expertise score with the two parameters, λ_f and λ_b, controlling the bias of the temporal information.

In the temporal random walk, the probabilities of a random surfer moving between authors, papers, and conferences are defined in the same way as those in equations (4)-(8). Here, the interpretation of the random jump parameter β is the probability that a surfer randomly jumps into any node in any time-window graph. For the transition probabilities of a random surfer moving forward or backward between time-inter networks, the definitions are as follows:

$$P(u_i^{s-1} \mid u_i^s) = \lambda_b \times \left(\frac{(1-\beta)}{Out\_Degree(u_i^s \to V^{s-1})} + \frac{\beta}{n}\right) \quad (15)$$

where u_i denotes an object in the time-varying network, for example, an author, a paper or a conference; n is the total number of the nodes in the network; Out_Degree($u_i^s \rightarrow V_{s\text{-}1}$) is the number of directed edges from u_i^s to nodes in the time-window (s-1). In our setting, temporal edges only exist between the same objects in the neighbouring time-windows. Thus, Out_Degree($u_i^s \rightarrow V_{s\text{-}1}$) equals to 1, or does not exit (the object u_i has never appeared before time-window s). Similarly, we can define $P(u_i^{s+1} | u_i^s)$ as follows:

$$P(u_i^{s+1} | u^s{}_i) = \lambda_f \times (\frac{(1-\beta)}{Out\_Degree(u^s{}_i \rightarrow V^{s+1})} + \frac{\beta}{n}) \quad (16)$$

Finally, the ranking score for a paper d_i in time-window s is defined as follows:

$$\begin{aligned} r^s(d_i) = & \sum_{(d_j,d_i)\in E_c^s} P'(d_i | d_j) r^s(d_j) + \sum_{(a_m,d_i)\in E_{ac}^s} P'(d_i | a_m) r^s(a_m) \\ & + \sum_{(c_k,d_i)\in E_{ic}^s} P'(d_i | c_k) r^s(c_k) + P(d_i^s | d_i^{s-1}) r^{s-1}(d_i) \\ & + P(d_i^s | d_i^{s+1}) r^{s+1}(d_i) \end{aligned} \quad (17)$$

For an author a_m and a conference c_k we similarly have:

$$\begin{aligned} r^s(a_m) = & \sum_{(d_j,a_m)\in E_{ac}^s} P'(a_m | d_j) r^s(d_j) \\ & + P(a_m^s | a_m^{s-1}) r^{s-1}(a_m) + P(a_m^s | a_m^{s+1}) r^{s+1}(a_m) \end{aligned} \quad (18)$$

$$\begin{aligned} r^s(c_k) = & \sum_{(d_j,c_k)\in E_{cl}^s} P'(c_k | d_j) r^s(d_j) \\ & + P(c_k^s | c_k^{s-1}) r^{s-1}(c_k) + P(c_k^s | c_k^{s+1}) r^{s+1}(c_k) \end{aligned} \quad (19)$$

In the temporal random walk model, we introduce a virtual node. That is, if a node starts appearing in time-window s and is not active in time-window (s+1), we then create a virtual node for it in time-window (s+1). This means that once a node appear in a time-window, it will continue existing in all of the following time-windows. The virtual node idea has an intuitive explanation: if a person once published papers in several years and does not publish any paper after that, she/he will be not active any more in the recent time-windows. By introducing the virtual node to represent the person in recent time-windows, we can obtain a positive expertise score for her/him in the recent time-windows. But, the score will decrease with time if this person continues to be not active. This virtual node method can be also viewed as a smoothing technique.

4. Experiment

4.1. Preparation

We evaluated the proposed methods in the context of Arnetminer (http://arnetminer.org) [21]. Arnetminer has been in operation online for two years. It gathered information of 448,365 researchers and 880,522 publications from the Web databases, pages, and files.

We evaluated our methods and the baseline methods on a subset of academic social network in Arnetminer. In total, the data set contains 15,169 citation relationships, 29,293 bi-directional authorship, 10,619 bi-directional paper-publish_at relationships and 26,368 bi-directional temporal relationships when setting time interval as 5 years. The timestamps of the papers span 33 years (from 1975 till present).

We collected seven queries for evaluation purpose. Specifically, we selected the most frequent queries from the log of Arnetminer (by removing overlap specific or lengthy queries, e.g., "A Convergent Solution to Tensor Subspace Learning"). We also normalized similar queries (e.g., "Web Service" and "Web Services" to "Web Service").

It is difficult to find a standard data set with ground truth. We employed the method of pooled relevance judgement [5] together with human judgements.

Specifically, for each query, we gathered the top 30 results from several similar academic search engines: Libra author search, Rexa authors search, and Arnetminer. We merged all the results together by removing ambiguous names (e.g., person with name "L. Liu") and names that do not exist in Arnetminer.

Then five senior graduates and two faculties were asked to provide judgements. Four grade scores (top expertise, expertise, marginal expertise, and not expertise) were asked to assign to each author, paper, and conference respectively. A specification was provided to guide the annotation process. For example, for paper ranking, the relevance and importance of each paper was evaluated based on the content relevance, cited number, published year, and impact to the field. The final golden ground truth was obtained by using "majority voting" for the judgements.

We evaluated performances of expertise search in each time-window (i.e., 1975-1979, 1980-1984, 1985-1989, 1990-1994, 1995-1999, 2000-2004, 2005-now).

In all experiments, we conducted evaluation in terms of P@5, P@10, P@20, R-pre, Recall and mean average precision (MAP). Readers are referred to [5, 11] for details of the measures.

4.2. Experimental setting

We use language model (LM) [3] and PageRank [19] as baseline methods. When calculating the ranking results in different periods, language model only utilized the data available before that period. Specifically, to get the author results in 1990, language model only used the papers published before 1990. We calculated the relevance between a document d and a query q as follows [3]:

$$p(q | d) = \prod_{t_i \in q} \lambda \cdot \frac{tf(t_i,d)}{|d|} + (1-\lambda) \cdot \frac{tf(t_i,D)}{|D|}, \qquad \lambda = \frac{|d|}{|d|+\mu} \quad (20)$$

where t_i is the i-th term in the query q; $|d|$ is the length of document d; $tf(t_i, d)$ is the term frequency of term t_i in document d; $|D|$ is total number of word tokens in the document collection D; $tf(t_i, D)$ is the term frequency of term t_i in D; λ is a parameter ranging in [0, 1] and is often set based on the length of document d; μ is another parameter and is commonly set as the average document length in D.

For authors or conferences, we merge all documents of a candidate (i.e., an author candidate or a conference candidate) together and treat as a virtual document [26]. We can also use equation (20) to calculate the relevance between a candidate and a query.

Another baseline method is PageRank, a popular method to calculate the importance of Web pages. The PagrRank $r(u_i)$ of an object u_i is defined as [19]:

$$r(u_i) = \sum_{(u_i,u_j)\in E} (1-\beta)\frac{r(u_j)}{Out\_Degree(u_j)} + \frac{\beta}{n} \qquad (21)$$

where n is the total number of the objects; Out_Degree(u_j) is the number of directed edges from u_j to others; β is a random jump parameter, which denotes the probability of a random surfer randomly jumps to another node in the network. In PageRank, we view the different types of objects as a unique homogeneous object and set the weights of different types of relationships equally.

4.3. Experimental results

4.3.1. Comparison with baseline methods. We performed experiments using the language model (LM), PageRank and the two proposed methods (shortly Aggregation and T-Random). For the parameter λ_{cc} , λ_{ca} , and λ_{cl} in the proposed methods, we set them based on analysis of the logs in the Arnetminer system from October, 2006 to July, 2007. Specifically, λ_{cc} represents how likely a user would click a cited paper when viewing the current paper; λ_{ca} denotes how likely the user would click its authors of the current paper; and λ_{cl} denotes how likely the user would click the published venue of the current paper. Finally, we obtained the ratio of three parameter, that is, λ_{cc}:λ_{ca}:λ_{cl} = 0.7:0.29:0.01. In addition, we set the random jump parameter β = 0.15. For the parameter λ_f and λ_b, we range the ratio of these two parameters from 0 to 1 with 0.1 as interval. Finally, to analyse the effect of the weight of time-intra network and time-inter network, we similarly range the ratio of the two set of parameters $((\lambda_{cc} + \lambda_{ca} + \lambda_{cl}):(\lambda_f + \lambda_b))$ from 0 to 1 with 0.1 as interval.

Table 1 shows the experimental results by evaluating the ranking lists of the current period. The results of Aggregation and T-Random are the best results obtained by tuning parameters. We can see that both of the proposed methods perform better than the baseline methods of using language model and PageRank. For example, in terms of P@5, the improvements are respectively +4.3% and +9.1% compared to language model. The baseline methods treat information located at different time-windows with the same weight and thus cannot distinguish current authoritative objects from those only active in a specific period.

Table 1. **Performances of four approaches (%)**

Method	Object	P@5	P@10	P@20	R-pre	MAP
LM	Paper	28.6	30.0	32.9	13.0	34.6
	Author	65.7	44.2	27.1	58.0	71.9
	Conference	54.2	34.3	22.1	46.7	58.3
	Average	**43.3**	**36.2**	**27.4**	**39.2**	**54.9**
Page-Rank	Paper	12.7	15.8	20.1	6.1	23.1
	Author	40.0	32.9	22.9	82.6	47.1
	Conference	48.6	31.4	22.9	89.1	48.8
	Average	**33.8**	**26.7**	**22.0**	**45.8**	**39.7**
Aggreg-ation	Paper	25.7	24.3	36.4	14.5	36.5
	Author	62.9	42.9	24.3	56.0	72.2
	Conference	54.3	35.7	22.1	46.7	58.7
	Average	**47.6**	**34.3**	**27.6**	**39.0**	**55.8**
T-Random	Paper	37.1	32.9	29.3	14.6	32.8
	Author	71.4	42.9	25.0	61.6	80.0
	Conference	48.6	35.7	22.1	44.8	58.1
	Average	**52.4**	**37.1**	**25.5**	**40.3**	**56.9**

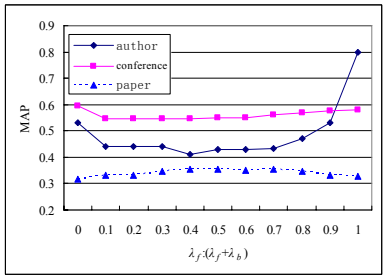


Figure 4. **MAP for different types of objects**

4.3.2. Time sensitivity. We conducted sensitivity analysis to the time information of different objects ranking. Figure 4 shows MAP of authors, papers, and conferences with different ratio of λ_f and λ_b in the temporal random walk when setting $\lambda_{cc} + \lambda_{ca} + \lambda_{cl} = 0.8$. We can see that author ranking is more sensitive than paper and conference. This is because: (1) one's research interest may change a lot over time. For example, "Raymond J. Mooney" focused on machine learning before 1997. Afterwards, his research switched to machine learning applications such as Natural Language Processing and Data Mining; (2) one's research work may be interrupted or terminated due to some accidents, e.g., switch to business; and (3) there are often rising 'stars' in a research field (a junior researcher quickly grows up). Compared with authors, conferences and papers seem to be more stable. They do change with time but the changes are smooth.

Table 2 shows the expertise search performances by different methods in each time-window. We can see that T-Random outperforms baseline methods in terms

of almost all evaluation measures. Figure 5 shows MAP measure of three methods in each time-window respectively. We see an interesting pattern: in the earlier period (say from 1984 to 1990's), the language model and PageRank based methods obtain very unsatisfactory results as they only use the information occurred before that period. In contrast, T-Random takes into account the temporal information and thus results into better performances.

Table 2. **Performances in each time-window (%)**

Method	Period	P@5	P@10	P@20	Recall	R-pre	MAP
LM	1980-1984	—	—	—	22.2	—	70.0
	1985-1989	46.7	13.3	—	18.3	—	61.2
	1990-1994	60.0	50.0	33.3	57.2	47.2	65.5
	1995-1999	33.3	46.7	30.0	52.2	38.3	49.1
	2000-2004	33.3	30.0	23.3	42.8	32.8	38.7
	2005- now	40.0	23.3	1.7	24.0	18.9	48.1
	Average	**42.7**	**32.7**	**20.1**	**36.1**	**34.3**	**55.4**
PageRank	1980-1984	—	—	—	22.2	—	70.0
	1985-1989	40.5	10.6	—	18.3	—	48.8
	1990-1994	50.0	31.4	23.9	47.4	54.7	45.4
	1995-1999	47.5	31.1	25.4	41.5	33.5	40.8
	2000-2004	33.1	30.0	24.0	37.4	35.3	36.2
	2005- now	35.0	27.4	22.6	25.8	45.8	38.8
	Average	**41.2**	**26.1**	**24.0**	**32.1**	**42.3**	**46.7**
Aggregation	1980-1984	53.3	30.0	15.0	100	63.5	75.1
	1985-1989	66.7	50.0	33.3	55.9	42.1	63.7
	1990-1994	66.7	53.3	45.0	67.2	48.9	64.7
	1995-1999	33.3	43.3	30.0	52.2	38.3	49.1
	2000-2004	33.3	30.0	21.7	42.8	31.1	38.6
	2005- now	40.0	23.3	15.0	28.8	23.2	53.5
	Average	**48.8**	**38.3**	**26.7**	**57.8**	**41.2**	**57.4**
T-Random	1980-1984	53.3	30.0	10.0	74.6	63.5	94.4
	1985-1989	73.3	60.0	45.0	79.7	51.7	65.4
	1990-1994	66.7	53.3	45.0	67.2	48.9	64.7
	1995-1999	33.3	43.3	30.0	52.2	38.3	49.1
	2000-2004	33.3	30.0	23.3	42.8	32.8	39.1
	2005- now	33.3	23.3	15.0	28.8	21.8	61.3
	Average	**48.9**	**40.0**	**28.1**	**57.3**	**42.8**	**62.3**

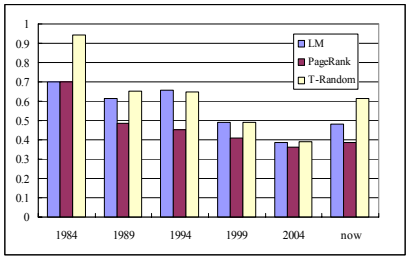


Figure 5. **Baseline methods vs. T-Random (MAP)**

4.3.3. Example analysis. Table 3 shows an example of ranking lists for the query "Machine Learning" obtained by the temporal random walk in different periods and ranking lists in the last time-window obtained by language model and Pagerank. Each paper is labeled with the time when this paper was published. From the column of papers, we can see that the temporal random walk, with the current parameter setting, tends to retrieve the "fresh" papers which may not the most important papers in the whole history but has great impact (e.g., hot topics) in the current period. From the column of conferences, we found that at the beginning of machine learning, papers on Machine Learning were primarily published on the conferences related to Artificial Intelligence. With the development of machine learning, several journals and conferences focusing on the machine learning issues came up, such as Machine Learning, ICML. From the column of authors, the top five authors in each period just match the active period of there authors.

4.3.4. Changes analysis. PageRank is used to evaluate the general importance of the objects whereas does not consider the dynamic changes of the links and objects themselves. Similarly, language model cannot distinguish the change trends either. We conducted analysis for the change trend of authors' interests and conferences obtained by the temporal random walk. Here, we use two examples to present our observations.

As a case study, we plot the ranking results of "Raymond Mooney" on three topics ("Natural Language Processing(NLP)", "Theory Refinement(TR)" and "Text Mining(TM)") obtained from the temporal random walk method by setting time-window size as one year. Figure 6 shows Raymond Mooney's ranking score changes with time. We can find that "Raymond Mooney" was retrieved as an expert in the field of natural language processing by T-Random in the entire period, though the ranking score changes over time. For another two topics, "Raymond Mooney" was considered as an expert only in a specific period. After checking his publications and research activities from 1985 to 2006, we found the evolution of his research interests match well his ranking changes over time. This confirms that the proposed model can effectively detect the trend of people's research interests.

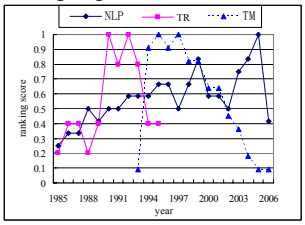
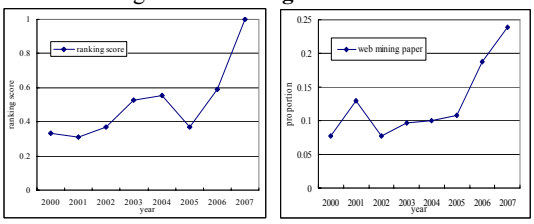


Figure 6. **Ranking score evolution**



(a) (b)

Figure 7. **The evolution of SIGKDD on "Web Mining"**

We also made an analysis on the changes of conference ranking. Similarly, in Figure 7(a), we show

the evolution of SIGKDD's ranking score on the topic of "Web Mining" from 2000 to 2007. Through the trend, we see that SIGKDD gradually placed more emphasis on web mining. Figure 7(b) displays the statistics obtained by counting the number of papers on web mining in SIGKDD 2000-2007. We see that the trend statistics is quite similar to our ranking evolution.

5. Related work

Random walk theory gained popularity in computer science with the emergence of the large number of Web-based networks. Considerable papers on link analysis have appeared. For example, HITS is a well-known link analysis algorithm [14], which divides the notion of importance of Web pages into two related attributes: hub and authority, and calculates two scores respectively for each page via the linkage between pages. PageRank is another state-of-the-art algorithm proposed by Brin and Page for estimating the importance of a Web page [19]. The basic idea in PageRank is to calculate the importance of each Web page based on the scores of the pages pointing to the page and thus Web pages pointed by many high quality pages become more important.

Based on PageRank, numerous extensions were proposed to special environments. For example, Xi et al [23] propose a unified link analysis frame work called link fusion to consider both inter- and intra- type link structure among multi-type data objects. Nie at al [17] propose an object-level link analysis model, called PopRank, to rank the objects within a specific domain. Liu et al [15] study a weighted, directed co-authorship network in digital libraries, and propose an AuthorRank algorithm to rank authors. A recent work also looks into random walk for learning on the subgraph its relation with the complement of it [13]. See also [7, 9, 10, 20, 25, 27].

Recently, Yu et al. [24] argue that the current famous search algorithms, such as PageRank and HITS, miss an important dimension of the Web, the temporal dimension. Berberich et al [6] indicated that temporal aspects should be taken into account in link analysis when computing the importance of a page and introduced T-Rank, a link analysis approach considering the freshness and activity of both pages and links. Amitay et al. [2] argue that if a page's last modification date is available then search engines will be able to provide more timely results and better reflect current real-list trends. Alonso et al. [1] propose that current information retrieval systems and applications do not take advantage of all the time information. They show some of the areas that can benefit from exploiting such temporal information. Nunes [16] indicated that the web is very active, exhibiting both high decay rates and high creation rates, and summarized two kinds of sources of temporal information on the web, namely document-based evidence and web-based evidence. On the other hand, temporal analysis has aroused many attentions these years. Berberich et al. [4] proposed the BuzzRank method that quantifies trends in time series of importance scores. Nallapati et al. [18] proposed a Multiscale Topic Tomography Model (MTTM) to model the evolution of topics with the time. In [21], each topic is associated with a continuous distribution over word co-occurrences and document's timestamp. Chi et al. [12] utilized the temporal and structural information to extract the communities. Backstrom et al. [8] studied the ways of communities in social networks growing over time at the level of individuals and their decisions to join communities.

6. Conclusion

In this paper, we have investigated the problem of expertise search in a heterogeneous social network. We have formalized the heterogeneous social network using a random walk model. For modelling the time-varying information, we have proposed a temporal random walk model by integrate a time-forward and -backward propagation in the random walk. Experimental results show that improvements can be obtained by comparison of the baseline methods. Another advantage of the proposed model is that it is easy to control the balance between information located at different time-windows.

7. Acknowledgment

The work is supported by the National Natural Science Foundation of China (90604025, 60703059), Chinese National Key Foundation Research and Development Plan (2007CB310803), and Chinese Young Faculty Research Funding (20070003093).

References

[1] O. Alonso, M. Gertz and R. Baeza-Yates, "On the Value of Temporal Information in Information Retrieval," in SIGIR Forum, vol. 41, Dec. 2007.

[2]E. Amitay, D. Carmel, M. Herscovici, R. Lempel, A. Soffer, and U. Weiss, "Temporal link analysis," IBM Research Lab Haifa, Tech. Rep., 2000.

[3]R. Baeza-Yates and B. Ribeiron-Net, Modern Information Retrieval, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[4]K. Berberich and S. Bedathur, "BuzzRank... and the trend is your friend," in Proc. WWW'06, 2006.

[5]C. Buckley, and E.M. Voorhees, "Retrieval evaluation with incomplete information," in Proc. SIGIR'04, 2004.

[6]K. Berberich, M. Vazirgiannis, and G. Weikum, "T-Rank: Time-Aware Authority Ranking," in Proc. WAW'04, 2004, LNCS 3243, pp.131-142.

[7]D. Byron, E. Iris, C. Alex, and Z. Yi, "Graph-based ranking for e-mail expertise analysis," in Proc. DMKD'03.

[8]L. Backstrom, D. P. Huttenlocher, J. M. Kleinberg, and X. Lan. "Group formation in large social networks: membership, growth, and evolution," in Proc. KDD 2006: 44-54
[9]C.S. Campbell, P.P. Maglio, A. Cozzi, and B. Dom, "Expertise identification using email communications," in Proc. CIKM'03, 2003.
[10]E. Garfied, Citation Indexing-Its Theory and Application in Science, Technology, and Humanities, 1979.
[11]N. Craswell, A. de Vries, and I. Soboroff, "Overview of the trec-2005 enterprise track," in TREC'05 Conference Notebook, 2005, p. 199.
[12]Y. Chi, S. Zhu, X. Song, J. Tatemura, and B. L. Tseng, "Structural and temporal analysis of the blogosphere through community factorization," in Proc. KDD'07, 2007, p. 163.
[13]J. Huang, T. Zhu, R. Greiner, D. Zhou and D. Schuurmans, "Information Marginalization on Subgraphs," in Proc. PKDD'06, 2006.
[14]Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, 46(5):604-632, 1999.
[15]Xiaoming Liu, Johan Bollen, Micheal L. Nelson, and Herbert Van de Sompl, "Co-authorship networks in the digital library research community," Information Processing and Management, 41(6): 681-682, 2005.
[16]S.Nunes, "Exploring Temporal Evidence in Web Information Retrieval," in Proc. FDIA'07, 2007, p.44
[17]Z. Nie, Y. Zhang, J. Wen, and W. Ma, "Object-level ranking: bringing order to web objects," in Proc. WWW'05, 2005, p. 567.
[18]R. Nallapati, W. Cohen, S. Ditmore, J. Lafferty, and K. Ung, "Multiscale topic tomography," in Proc. KDD'07.
[19]L. Page, S. Brin, M. Rajeev, and W. Terry, "The pagerank citation ranking: Bringing order to the web," Technical Report SIDL-WP-1999-0120, Stanford University.
[20]G. Pinski and Francis Narim, "Citation influence for journal aggregates for scientific publications: Theory, with application to the literature of physics," Information Processing and Management, vol. 12, pp. 297–312, 1976.
[21]J. Tang, D. Zhang, and L. Yao. "Social network extraction of academic researchers." in Proc. ICDM'2007. pp. 292-301
[22]X. Wang and A. McCallum, "Topics over time: a Non-Markov-Continuous-Time model for topical trends," in Proc. KDD'06, 2006.
[23]X W. Xi, B. Zhang, Y. Lu, Z. Chen, S. Yan, H. Zeng, W.Y. Ma, and E. A., "Link fusion: a unified link analysis framework for multi-type interrelated data objects," in Proc. WWW'04, 2004, p. 319.
[24]P. Yu, X. Li, and B. Liu, "On the temporal dimension of search," in Proc. WWW'04, 2004, p. 448.
[25]J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in Proc. WWW'07.
[26]J. Zhang, J. Tang, L. Liu, and J. Li, "A Mixture Model for Expert Finding," in Proc. PAKDD'2008.
[27]D. Zhou, S. Orshanskiy, H. Zha, and C. Lee Giles, "Co-Ranking Authors and Documents in a Heterogeneous Network," in Proc. ICDM' 07.

Table 3. The top ranking results in different periods obtained by T-Random for "Machine Learning"

Period	Authors	Conferences	Papers
1979	Tom M. Mitchell	IJCAI	A Model for Learning Systems (1977)
1984	Pat Langley(new) Tom M. Mitchell(1) Dana S. Nau(new)	AI Magazine IJCAI COLING	Learning from Solution Paths: An Approach to the Credit Assignment Problem (1982) A Model for Learning Systems (1977)
1989	Thomas G. Dietterich Nicholas S. Flann Tom M. Mitchell Richard M. Keller	Machine Learning ML IJMMS Cognitive Science	Learning at the Knowledge Level (1986) A Study of Explanation-Based Methods for Inductive Learning (1989) Explanation-Based Generalization: A Unifying View (1986) Limitations on Inductive Learning (1989)
1994	Pat Langley Ryszard S. Michalski Raymond J. Mooney Jude W. Shavlik	Machine Learning AAAI JCAMD NIPS	Elements of Machine Learning (1994) Inferential Theory of Learning as a Conceptual Basis for Multistrategy Learning (1993) Symbolic and Neural Learning Algorithms: An Experimental Comparison (1991) An Integrated Framework for Empirical Discovery (1993)
1999	Thomas G. Dietterich Vladimir Vapnik Pat Langley Raymond J. Mooney	Machine Learning Commun. ACM AI Magazine ACM Comput. Surv.	Encouraging Experimental Results on Learning CNF (1995) Support-Vector Networks (1995) Machine-Learning Research (1997) Machine Learning (1996)
2004	Olvi L. Mangasarian Ryszard S. Michalski Vladimir Vapnik Dayne Freitag	Machine Learning JMLR ICML SSPR/SPR	Machine Learning for Information Extraction in Informal Domains (2000) Selecting Examples for Partial Memory Learning (2000) Learnable Evolution Model: Evolutionary Processes Guided by Machine Learning (2000) Choosing Multiple Parameters for Support Vector Machines (2002)
now	Glenn Fung Olvi L. Mangasarian Nicholas Kushmerick Fabio Ciravegna	Machine Learning ICML KDD JASIST	Multicategory Proximal Support Vector Machine Classifiers (2005) Evaluating machine learning for information extraction (2005) Additive regularization trade-off: fusion of training and validation levels in kernel methods (06) Supervised clustering with support vector machines (2005)
LM (Now)	Thomas G. Dietterich Pat Langley Vladimir Vapnik Raymond J. Mooney	Machine Learning JMLR ICML Commun. ACM	Learning at the Knowledge Level (1986) Machine Learning for Information Extraction in Informal Domains (2000) Elements of Machine Learning (1994) Encouraging Experimental Results on Learning CNF (1995)
PageRank (Now)	Thomas G. Dietterich Raymond J. Mooney Daphne Koller Vladimir Vapnik	ICML AAAI Machine Learning NIPS	C4.5: Programs for Machine Learning (1993) Lessons Learned from Applying AI to the Web (2000) PRODIGY: An Integrated Architecture for Planning and Learning (1991) Comparative experiments on learning information extractors for proteins and their interactions (05)