International Journal of Computational Intelligence Systems, Vol. 4, No. 5 (September, 2011), 1012-1021

WebScore: An Effective Page Scoring Approach for Uncertain Web Social Networks*

Shaojie Qiao [1]†, Tianrui Li [1], Michael Chau [2], Jing Peng [3], Yan Zhu [1], Hong Li [1]

[1] School of Information Science and Technology, Southwest Jiaotong University
No. 111, Erhuanlu Beiyiduan, Chengdu, Sichuan 610031, China
E-mail: sjqiao@home.swjtu.edu.cn

[2] School of Business, The University of Hong Kong, Pokfulam, Hong Kong, China
E-mail: mchau@business.hku.hk

[3] Department of Science and Technology, Chengdu Public Security Bureau
No. 136, Wenwu Road, Chengdu, Sichuan 610017, China
E-mail: pjxxlpsj@hotmail.com

Abstract

To effectively score pages with uncertainty in web social networks, we first proposed a new concept called transition probability matrix and formally defined the uncertainty in web social networks. Second, we proposed a hybrid page scoring algorithm, called WebScore, based on the PageRank algorithm and three centrality measures including degree, betweenness, and closeness. Particularly, WebScore takes into a full consideration of the uncertainty of web social networks by computing the transition probability from one page to another. The basic idea of WebScore is to: (1) integrate uncertainty into PageRank in order to accurately rank pages, and (2) apply the centrality measures to calculate the importance of pages in web social networks. In order to verify the performance of WebScore, we developed a web social network analysis system which can partition web pages into distinct groups and score them in an effective fashion. Finally, we conducted extensive experiments on real data and the results show that WebScore is effective at scoring uncertain pages with less time deficiency than PageRank and centrality measures based page scoring algorithms.

Keywords: uncertainty, web social network, PageRank, centrality measures

1. Introduction

During the last two decades, several page scoring (or ranking) measures has been proposed to improve the accuracy and efficiency of ranking the query results obtained by search engines. Consequently, other research fields could benefit a lot from this promis-

*This work is partially supported the National Natural Science Foundation of China under Grant Nos. 61100045, 61165013, 61003142, 60902023, 61171096; the China Postdoctoral Science Foundation under Grant Nos. 201104697, 20090461346; the Specialized Research Fund for the Doctoral Program of Higher Education under Grant Nos. 20110184120008, 20090184120020; the Youth Foundation for Humanities and Social Sciences of Ministry of Education of China under Grant No. 10YJCZH117; the Fundamental Research Funds for the Central Universities under Grant Nos. SWJTU09CX035, SWJTU11ZT08.
†Corresponding author.

ing technology. For instance, we can explore key players from a social network described by criminals and their contracting e-mail relations[1]. In addition, we can find a sub-group after we assign each member a score based on link structure in a given social network[2]. Thus, we can see that page scoring is of great practical value and can be applied to web social networks (WSNs for short), e.g., Facebook, Twitter and Myspace, which are famous web social networks that possess a large number of users. The most important characteristic of WSNs is that: it is a social structure made of pages called "nodes" which are tied by one or more specific types of interdependency, such as similar themes or contents related to friendship, kinship, financial exchange[3]. In this study, we focus on addressing the problem of how to score pages in WSNs in an effective fashion with a guarantee of time efficiency.

With the increasing availability of wide-coverage online collaborative sources and web-based networking techniques, new requirements in the developments of knowledge discovery in WSNs have attracted much attention. Online social networks become very intricate as their inner structures become huge as well as change dynamically. Simultaneously, WSNs accumulate a huge number of useful data, which raises the urgent need to analyze complex network structures in order to help electronic service providers find important client base. Herein, an important and challenging problem is to score the importance of members and identify the essential players from complex social networks, especially in the research domain of criminal data mining that has become prevalent since the tragic events of September 11 and subsequent anthrax contamination of letters[4].

Since the clients from the Internet pay close attention to the accuracy of personalized services, the ranking of the query results should be improved to satisfy the given requirements. Normally, existing page ranking technologies are based on deterministic data and the query results are rarely changed in a short period of time. Recently, Web 2.0 applications become more prevalent and it can facilitate interactive information sharing, interoperability, and collaboration on the Web. We can obtain a large amount of data that contain useful information about web social networks from Web 2.0. However, these data change dynamically, which increases theirs uncertainty. So, we should take into account the uncertainty in WSNs in order to accurately score pages.

The uncertainty of web data is formed due to the following reasons:

- The Internet structures could change dynamically over time. For example, pages or links are created and disappeared every minute. Consequently, it is difficult for us to accurately analyze the dynamically evolving network structure.
- Unpredictable events may occur at some websites. A website could be under maintenance, updating. Some other uncertain factors, e.g., network congestion, may lead to occasional unavailability of web services. After a while, these web pages will reappear in the Internet.
- Another important factor is missing or imperfect data generated in the phase of ETL (extract, transform, and load) process. So, some information can be neglected and some void value might be assigned by default.

PageRank is a popular algorithm that has been mainly used to rank query results based on link structure among URLs. Particularly, it can be used to determine the importance of social actors in a proper social network where links imply similar "recommendation" relations as the hyperlinks in Web graph[5]. PageRank adapts the voting strategy to effectively rank pages by analyzing the themes of adjacent web pages, but it still has some drawbacks[6].

From the above discussions, we believe that it is important to develop an effective (i.e., accuracy of scoring) and efficient (i.e., time efficiency) page scoring approach for uncertain WSNs that change dynamically and contain complex link structures in forms of hyperlinks. To achieve this goal, we make the following original contributions in this study:

1. We formally define the uncertainty in WSNs and apply it to compute the transition probability of each page. Moreover, we transfer the problem of computing the transition probability of one page into calculating the probability of all possible routes passing through it.

2. We propose an effective page scoring algorithm, called WebScore, which is a hybrid approach based on PageRank, the three centrality measures (i.e., degree, betweenness and closeness) and uncertainty in WSNs, in order to evaluate the importance of pages in a continuously changing environment.

3. We develop a WSN analysis system, called WebMiner, which integrates several web mining techniques as well as visualizes WSNs.

4. Extensive experiments are conducted on WebScore to verify its effectiveness and efficiency. The results show that WebScore is better at accurately scoring pages than the previously proposed page ranking approaches with less time deficiency.

2. Related Work

Uncertain data on the Web become ubiquitous with the rapid development of Internet technologies, especially the Instant Messaging softwares, e.g, MSN Messenger, ICQ, which work in a friend-to-friend network, in which each node connects to the friends on the friends list and these nodes form a web social network. Uncertain web social networks appear in a wide range of fields including business, finance, logistics, engineering, and society. Uncertain data appear in different forms including imprecision, incompleteness, vagueness and inconsistency. Möller et al.[7] gave a mathematical description of uncertain data, which is essential for forecasting by means of fuzzy random process and neural networks. Zhou et al.[8] reviewed the uncertain data management techniques including data model, preprocessing, integrating, storage, indexing, and query processing. In addition, they introduced several famous uncertain data management system. Cheng[9] introduced a generic data model to capture uncertainty as well as provided query operators that return answers with statistical confidences after cleaning uncertain data.

In the real world, there exist various kinds of uncertain information on the Web, especially for WSNs where the link relationships between pages are complex and uncertain. Social network analysis (SNA) has played a critical role in determining the importance of individuals that act in an organization or network. Particularly, SNA technologies can be used to find key players and communities in criminal networks[1,10]. The centrality measures in SNA are appropriate to estimate the importance of members and can be applied to WSNs for scoring pages[3].

PageRank[11] is based on the idea that the pages linked by the ones with high authority are often treated as high quality pages. However, PageRank could cause the phenomena of subside which will affect the correctness of scoring. When the subside occurs, a group of pages are connected with each other without out-links to other pages[6]. In order to overcome the drawbacks of PageRank, Zhang et al.[12] proposed a new PageRank algorithm to improve the time performance of scoring by distributing the authority in a fast fashion. Haveliwala[13] proposed a topic-sensitive PageRank algorithm which can cope with the problem that some pages may not be treated as important in some fields though they have a high score by PageRank in other fields. Richardson et al.[14] proposed a PageRank algorithm based on URLs and the content of pages. The algorithm can handle the problem that when users browse the Internet from one page to another, their behavior will be greatly affected by the content of the current visiting page and query keywords.

The previously proposed PageRank algorithms aim to improve the accuracy of scoring a page without taking into account the structure characteristics of WSNs. There is a growing trend that the Web is partitioned into distinct WSNs, which motivates us to develop a new PageRank algorithm by considering the characteristics of WSNs or communities. SNA is an alternative solution to analyze patterns of interactions between social actors in order to discover an underlying social structure[15]. Several centrality measures in relational analysis of SNA can be used to identify key members who play important roles in a network, that is, degree, betweenness, and closeness. These centrality measures are particularly appropriate to analyze the importance of players and can be applied to WSNs for scoring pages as well. Based on SNA techniques, Qiao et al.[1] proposed a centrality measure based approach to iden-

tify the central players in criminal networks.

In the rest of this paper, we will introduce the preliminaries in Section 3, present an overview of the WebScore algorithm in Section 4, introduce a web social network analysis system in Section 5, describe the performance studies in Section 6, and finally conclude this study in Section 7.

3. Problem Statement and Preliminaries

In this section, we will first define a web social network, then introduce the concepts of transition probability matrix and uncertainty in WSNs, and address the theoretical foundations of computing the transition probability of pages. Finally, we will formalize three important centrality measures in SNA.

Definition 1. (Web social network)[3] Let $\mathscr{N}$ be a web social network. $\mathscr{N}$ is defined as a graph $\mathscr{G} = (V,E)$, where V is a set of web pages, E is a set of directed edges (p,q,w), $p,q \in V$, and w is a weight between p and q. The pages in $\mathscr{G}$ should satisfy that they have at least one in-link or out-link from other pages, that is, nodes in $\mathscr{G}$ directly (indirectly) connect to others.

Figure 1 is an illustrative example of a WSN that is extracted from our WSN analysis system. We can see that a WSN will become more complicated as the number of pages grows.

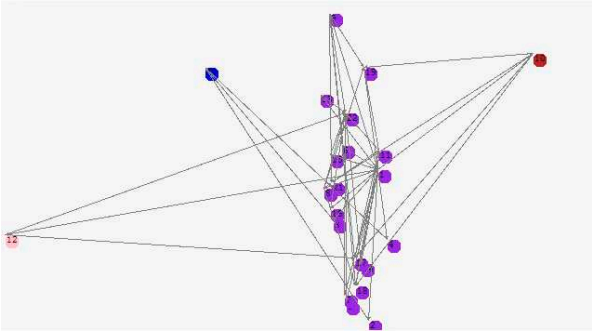

Fig. 1. Example of a web social network.

Based on the concept of web social network, the page scoring problem is defined as ranking web pages in WSNs by their importance.

Definition 2. (Binary matrix) Let $U = \{X_1, X_2, ..., X_n\}$ be a finite set of web pages. The pages are related by a binary relation[16]:

$$\mathscr{R} \subseteq U \times U \tag{1}$$

which determines a network $N = (U, \mathscr{R})$, the relation $\mathscr{R}$ is described by a binary matrix $\mathbf{R} = [r_{ij}]_{n\times n}$, where

$$r_{ij} = \begin{cases} 1 & X_i \mathscr{R} X_j \\ 0 & otherwise \end{cases} \tag{2}$$

Definition 3. (Transition probability matrix) A transition probability matrix is defined as:

$$\Phi = [\phi_{ij}]_{n\times n} = \begin{pmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{nn} \end{pmatrix} \tag{3}$$

where ϕ_{ij} represents the transition probability by which a web surfer jumps from $Page_i$ to $Page_j$, n is the number of pages, and ϕ_{ij} is calculated by the following equation.

$$\phi_{ij} = \begin{cases} 1/p_i, & p_i \neq 0 \text{ and } r_{ij} = 1; \\ 0, & p_i \neq 0 \text{ and } r_{ij} = 0; \\ 1/n, & p_i = 0. \end{cases} \tag{4}$$

where p_i is the number of out-links on $Page_i$.

According to Equation 4, we can see that $\sum_{i=1}^{n} \phi_{ij} = 1$, where j represents an arbitrary page number.

In order to describe the uncertainty of a page in WSNs, we present the formal definition of uncertainty in WSNs as follows.

Definition 4. (Uncertainty in web social networks) The uncertainty of a WSN is defined as a vector $U = (u_1, u_2, \ldots, u_n)$, and $u_i = t_i/T$, where t_i is the number of possible routes going from other pages to page i, and T is the number of all possible routes in a WSN. t_i is the number of possible routes after $n-1$ jump operations (i.e., surfers move from one page to another) and is calculated by the following formula, in which the real number 1 can guarantee $t_i \neq 0$.

$$t_i = \sum_{k=1}^{n-1} \sum_{m=1}^{n} r_{mi}^{(k)} + 1 \tag{5}$$

where $r_{mi}^{(k)}$ is the number of real routes going from page m to i after k jump operations and n is the number of pages. $r_{mi}^{(k)}$ is an element in a *k-order transition matrix* $M^{(k)}$ as given below:

$$M^{(k)} = [r_{ij}^{(k)}]_{n\times n} = \begin{pmatrix} r_{11}^{(k)} & r_{12}^{(k)} & \cdots & r_{1n}^{(k)} \\ r_{21}^{(k)} & r_{22}^{(k)} & \cdots & r_{2n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1}^{(k)} & r_{n2}^{(k)} & \cdots & r_{nn}^{(k)} \end{pmatrix} \quad (6)$$

Straightforwardly, T can be obtained by the following equation.

$$T = \sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \sum_{k=1}^{n-1} \sum_{j=1}^{n} r_{ji}^{(k)} + 1 \quad (7)$$

It is an essential task to compute the *k-order transition matrix* $M^{(k)}$. Here, we use the following theorem to prove how to calculate it.

Theorem 1. *Let $M^{(1)}$ be a 1-order transition matrix, $M^{(k)}$ be the k-order transition matrix, it does hold that:*

$$M^{(k)} = \prod^{k} M^{(1)} = \underbrace{M^{(1)} \times M^{(1)} \times \cdots \times M^{(1)}}_{k} \quad (8)$$

where

$$M^{(1)} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \quad (9)$$

each element r_{ij} in $M^{(1)}$ represents the number of routes going from page i to page j.

Proof: Mathematical induction can be used to prove this theorem. If $M^{(k)}$ holds for $k = m$, then we can also prove that $M^{(m+1)}$ holds for $k = m+1$. Finally, we can conclude that $M^{(k)}$ holds for all natural k. For space limitation, we omit the detail here. □

In this study, we use the following PageRank scoring formula by integrating the *uncertainty in WSNs* to compute the rank value of each page.

$$PR^*(p_i) = 1 - e + e \times \sum_{p_j \in R(p_i)} u_{p_j} \times \frac{PR^*(p_j)}{N(p_j)} \quad (10)$$

where $PR^*(\cdot)$ is the page scoring function, p_i represents a web page, e is a dampening factor between 0 to 1 and is set to 0.85 which has been verified to be a reasonable value by Qin[5], p_j is any page linking to p_i, $R(p_i)$ is the set of pages that link to p_i, u_{p_j} is the uncertainty of p_j, and $N(p_j)$ is the number of out-links on p_j.

According to Equation 10, we compute the rank value by multiplying $PR*(p_j)$ with u_{p_j}. u_{p_j} takes into a full consideration of the uncertainty in WSNs and can be obtained by Definition 4. Empirically, the experimental results given in Section 6 show the effectiveness of our proposed PageRank algorithm. Intuitively, the rank values can be obtained by recursively calculating the above formula.

Here, we will formally define the centrality measures used in the WebScore algorithm. First, we will introduce the distance measure called geodesic.

Definition 5. (Geodesic) Geodesic[3] is the shortest path between two web pages. It is defined as:

$$L(x_i, x_j) = \min_{0 \leqslant k < n} \{L(x_i, x_k) + L(x_k, x_j)\} \quad (11)$$

Geodesic can be obtained by iteratively computing the elements in an adjacent matrix which indicates the length of the shortest path from one page to another. Geodesic is used by the centrality measures to compute the importance of pages.

In this study, we use the three most popular centrality measures which were proposed by Freeman[17] as shown below.

Degree measures how active a particular web page is. It is defined as the number of direct links a page k has:

$$S_d(k) = \sum_{i=1}^{n} a(i,k) \quad (12)$$

where n is the total number of pages in a network, and $a(i,k)$ is a binary variable indicating whether a link exists between pages i and k. A page with a high degree could be the "hub" of a WSN.

Betweenness measures the extent to which a particular page lies between other pages in a network. The betweenness of a node k is defined as the number of geodesics passing through it:

$$S_b(k) = \sum_i^n \sum_j^n g_{ij}(k), i \neq j \tag{13}$$

where $g_{ij}(k)$ indicates whether the shortest path between two other pages i and j passes through page k. A page with a high betweenness may act as a gatekeeper or "broker" in a WSN.

Closeness is the sum of the length of geodesics between a particular page k and all other pages in a network. It actually measures how far away one page is from other pages and is called *farness* in a WSN.

$$S_c(k) = \sum_{i=1}^n d(i,k) \tag{14}$$

where $d(i,k)$ is the length of the shortest path connecting pages i and k.

4. WebScore Algorithm

In this section, we will introduce a hybrid page scoring algorithm, namely WebScore, based on PageRank and centrality measures. WebScore can help score pages from an uncertain WSN in an effective fashion. The detail is given in Algorithm 1.

WebScore contains four essential phases: (*a*) use Equations 10∼14 to compute the rank value, degree, betweenness, and closeness of each page, respectively (lines 2-7); (*b*) standardize these three centrality measures into values between 0 and 1, respectively (lines 9-11); (*c*) compute the scoring value by using the weighted average formula in line 12; finally (*d*) output the ranking score of each page in WSNs (line 13).

Algorithm 1: An effective page scoring approach for uncertain web social networks

Input: A WSN database $\mathbb{D}$, a weight vector $w =< w_1, w_2, w_3, w_4 >$, where the elements in w represent the weight of the rank value, degree, betweenness, and closeness, respectively.

Output: A scoring list of pages l_i, where i is the sequence number of pages.

1 $n = Number(\mathbb{D})$;
2 **for** *(each page* $p_i \in \mathbb{D}$*)* **do**
3 $PR_i = PR^*(p_i)$;
4 $degree_i = S_d(p_i)$;
5 $betweenness_i = S_b(p_i)$);
6 $closeness_i = S_c(p_i)$;
7 **end**
8 **for** $(i = 0;\ i < n;\ i{+}{+})$ **do**
9 $d'_i = \frac{degree_i}{\sum_{i=1}^n degree_i}$;
10 $b'_i = \frac{betweenness_i}{\sum_{i=1}^n betweenness_i}$;
11 $c'_i = \frac{min(closeness_1, \cdots, closeness_n)}{closeness_i}$;
12 $l_i = w_1 * PR_i + w_2 * d'_i + w_3 * b'_i + w_4 * c'_i$;
13 **output** l_i;
14 **end**

5. Web Social Network Analysis System

In this section, we will introduce a WSN analysis system as shown in Figure 2, called WebMiner, which is composed of five main components as introduced below. In addition, we use this system to estimate the performance of WebScore by comparing it with other algorithms.

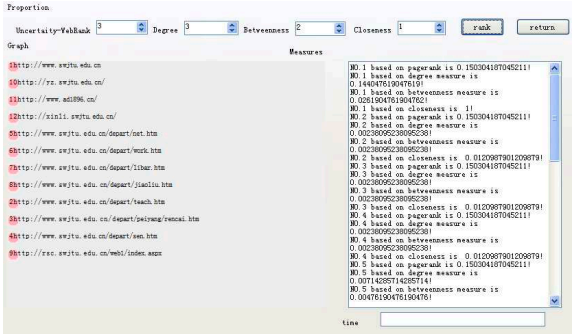

Fig. 2. Scenario of the WebMiner system.

1. **Web crawler**. It is an efficient web crawler to obtain data from the Web. After ETL processes, it stores useful pages into a web database.

2. **Web social network creation**. It employs a concept space approach[18] to automatically create networks.

3. **Web page score**. It uses the WebScore algorithm that integrates uncertainty based on PageRank and centrality measures in order to score pages in WSNs.

4. **Web cluster analysis**. We develop an improved *k*-means algorithm and a blockmodeling based hierarchical clustering algorithm[2] to partition WSNs in an effective and efficient fashion.

5. **Web social network visualization**. We use the multidimensional scaling (MDS)[19] approach to visualize web social networks.

Figure 2 is a snapshot of the web page score module which consists of three distinct parts. The user can specify the weight vector in the text fields following the labels "Uncertain-WebRank", "Degree", "Betweenness", and "Closeness", respectively. When clicking the tab labeled "rank", the scored URLs are listed in the canvas labeled as "Graph". In addition, the values of these four measures are illustrated in the right text area labeled "Measures", which can help the user understand the score sequence of URLs.

6. Experiments

In this section, we report the experimental studies which compare WebScore with PageRank, centrality measures based SNA algorithm (called CentralityRank), and the WebRank algorithm[3], respectively. Basically, CentralityRank uses three centrality measures to compute the importance of pages and rank them. WebRank is a hybrid algorithm by combining centrality measures into PageRank without taking into account the effect of uncertainty. All algorithms are implemented in the VS.net development platform and experiments are conducted on an Intel T2300, 1.66GHz CPU with 1.5GB of main memory, running on Windows XP operating system. All experiments are run on real data that are crawled by WebMiner.

6.1. Weight vector tuning

To find an appropriate proportion across the rank value, degree, betweenness, and closeness in a weight vector is important for accurately scoring pages in web social networks. In this section, we use fifteen pages as shown in Figure 1 to conduct experiments in order to find an appropriate weight vector among these four measures and estimate the performance of WebScore.

In this section, we perform four sets of experiments by using distinct proportion of the weights corresponding to the rank value, degree, betweenness and closeness, respectively. The results are given in Table 1.

Table 1. Page sequence after scoring by WebScore under distinct proportions of weights on these four measures.

Page No.	Proportion of weights on four measures			
	1:1:1:2	1:0:0:0	0:3:2:1	3:3:2:1
1	1	4	1	1
2	5	5	2	5
3	6	6	3	6
4	7	7	4	7
5	8	8	5	8
6	9	9	6	9
7	10	1	7	10
8	11	2	8	11
9	12	3	9	12
10	14	14	15	15
11	13	13	13	13
12	15	15	14	14
13	2	10	10	2
14	3	11	11	3
15	4	12	12	4

In Table 1, the first column is the sequence number of web pages, and the numbers from the second to the fifth columns represent the scored page sequence that is obtained by WebScore under different weight combinations. As we can see from the above table, the orders of pages in the second, the

fourth, and the fifth columns are similar and consistent with the real world situation that *Page 1 acts as an essential hub in Figure 1 since it has at least one link to other pages*. On the other hand, the results in the third column deviate from the real case. We can see that we cannot only use the PageRank measure without taking into account SNA measures. After a careful estimation by Web mining experts, we get the best proportion of the four measures is $< 1, 1, 1, 2 >$, i.e., the order of pages in the second column is the best scoring result. Particularly, this proportion of weights agrees with the conclusion from another work on SNA, as Qiao[1] addressed "*we should combine these three centrality measures (degree, betweenness, closeness) to find the key members in a social network*".

Note that the scoring sequence of 15 pages are similar for PageRank, CentralityRank, and WebRank as well. When the number of pages increases, the results are the same as the case of 15 web pages. In the following sections, we use a weight vector $< 1, 1, 1, 2 >$ to evaluate the performance of these four algorithms.

6.2. Accuracy comparison of page scoring

In this section, we compare the scoring accuracy of WebScore with PageRank, CentralityRank and WebRank, respectively. We observe the scoring accuracy as the number of pages grows from 10 to 1000, and the results are given in Figure 3. Note that we run each experiment ten times and use the average value to denote the accuracy.

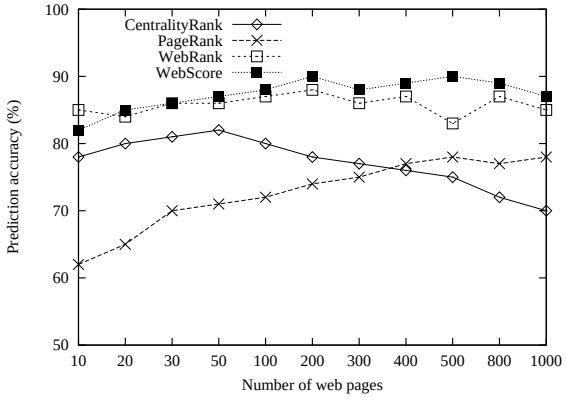

Fig. 3. Scoring accuracy comparison among algorithms.

By Figure 3, we can see that: (1) the scoring accuracy of CentralityRank grows as the cardinality of pages is small (less than 50 pages), but when the number of pages increases large, the accuracy falls drastically. This is because the prediction accuracy of centrality measures will decrease when the WSN becomes very complex; (2) the scoring accuracy of PageRank increases with the number of pages. Since PageRank adopts the authority transition strategy to rank pages, if the number of authoritative pages increases, the authority will transfer to other pages as well as improve the scoring accuracy; (3) for WebRank and WebScore, the accuracy keeps at a similar level and is better than the other two algorithms, which shows their stability and effectiveness. This is because both WebRank and WebScore combine the advantages of PageRank and CentralityRank; (4) WebScore outperforms PageRank and CentralityRank in all cases with a big gap and is a little better than WebRank. Whereas, the performance of PageRank is worse than the other three algorithms. Because PageRank is mainly used to rank pages, instead of uncertain WSNs. WebScore takes into account the uncertainty in WSNs, so it is more suitable to score pages than WebRank.

The relationship of pages in WSNs is an important factor for the scoring accuracy of uncertain web pages, so we have to further analyze the results as shown in Figure 4 under distinct WSNs with similar nodes and distinct link relations among pages, where the *x*-axes represent 5 difference WSNs with 500 nodes, and the *y*-axis represents the scoring accuracy.

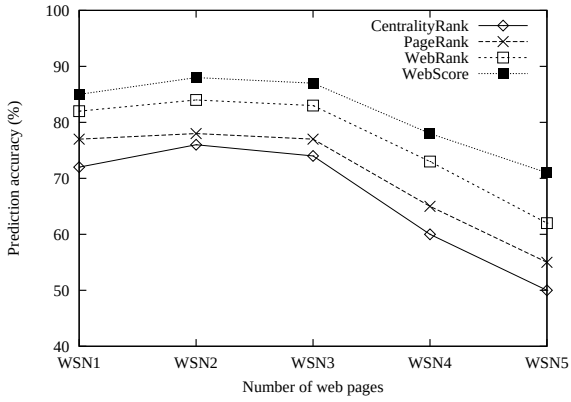

Fig. 4. Scoring accuracy comparison among algorithms under five WSNs with 500 pages.

The results in Figure 4 agree with Figure 3 under distinct WSNs, that is, WebScore performs better in all cases than other algorithms whether the WSN is complex (e.g., WSN5) or not (e.g., WSN2). The results further show the stability of WebScore.

6.3. Efficiency comparison of page scoring

We further compare the execution time of WebScore with PageRank, CentralityRank, and WebRank. The results are given in Figure 5, where the *x*-axis is the number of pages and the *y*-axis represents the execution time corresponding to each algorithm.

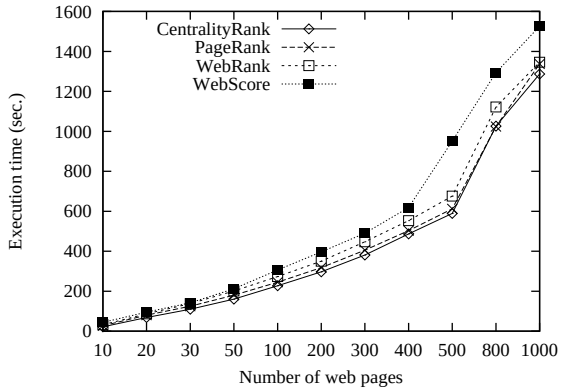

Fig. 5. Execution time comparison among algorithms.

As we can see from Figure 5, WebScore has comparable time efficiency to WebRank and PageRank, while CentralityRank has the most economical cost of time. As the number of pages increases, the time gap between WebScore (WebRank or PageRank) and CentralityRank is big. Because WebScore needs to recursively compute the score and uncertainty of pages in WSNs, which is time consuming.

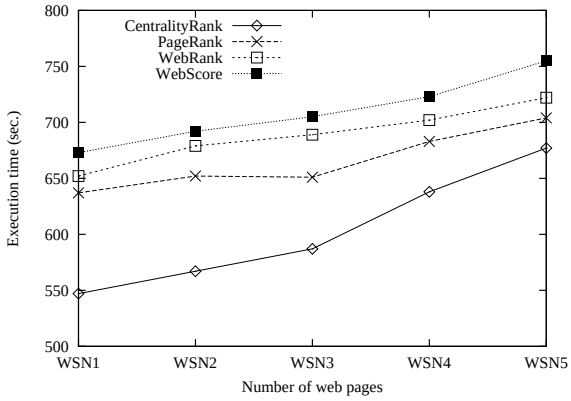

Fig. 6. Execution time comparison among algorithms under five WSNs with 500 pages.

Figure 6 illustrates the runtime comparison among these four algorithms under distinct WSNs with 500 similar nodes and distinct link relations among pages. We can see that the time deficiency of WebScore, WebRank, and PageRank is not obvious, but they have a big difference to the CentralityRank algorithm. However, as web social networks change complex from WSN1 to WSN5, the differences become small. This is because CentralityRank needs much time to compute the geodesic distance among pages as the relationships of web pages become intricate.

7. Conclusions and Future Work

There is inherent uncertainty in web data. How to manage the uncertainty of web social networks has becomes an important and challenging problem. In this study, we aim to compute the uncertainty of pages and apply uncertainty in WSNs to the PageRank algorithm in order to effectively score pages in WSNs. Existing PageRank algorithms focus on computing the authorities by analyzing the link structure of web pages, and often overlook the effect of uncertainty and the essential roles of pages in WSNs. So, we provide a solution called WebScore to score pages in WSNs by applying centrality measures and uncertainty in WSNs to the PageRank algorithm. Empirical studies show that WebScore is effective at scoring pages with good time performance by comparing it with PageRank, CentralityRank and WebRank algorithms. Furthermore, we develop a powerful web social network analysis system, called WebMiner, which can group and score pages in an effective and efficienty fashion.

Our future research includes: (1) consider other important uncertainty factors existing in WSNs, and integrate them into WebScore to effectively score pages; (2) improve the time efficiency of WebScore by optimizing the PageRank algorithm or using the parallel computing approach[20]; (3) improve the WebMiner system by developing other useful web mining and intelligent information processing approaches[21,22,23,24,25].

References

1. S. Qiao, C. Tang, J. Peng, W. Liu, F. Wen, J. Qiu, "Mining key members of crime networks based on personality trait simulation email analysis system," *Chinese Journal of Computers*, **31**(10), 1795–1803 (2008) (in Chinese).
2. S. Qiao, T. Li, H. Li, H. Chen, J. Peng, J. Qiu, "HCUBE: a hierarchical clustering algorithm using blockmodeling in web social networks," *Proceedings of the 9th International FLINS Conference on Foundations and Applications of Computational Intelligence*, **4**, 697–702 (2010).
3. S. Qiao, J. Peng, H. Li, T. Li, L. Liu, H. Li, "WebRank: a hybrid page scoring approach based on social network analysis," *Proceedings of the 5th International Conference on Rough Set and Konwledge Technology*, 475–482 (2010).
4. S. Qiao, C. Tang, H. Jin, S. Dai, X. Chen, "Constrained k-closest pairs query processing based on growing window in crime databases," *Proceedings of 2008 IEEE International Conference on Intelligence and Security Informatics*, 58–63 (2008).
5. J. Qin, J. J. Xu, D. Hu, M. Sageman, H. Chen, "Analyzing terrorist networks: a case study of the global salafi jihad network," *Proceedings of 2005 IEEE International Conference on Intelligence and Security Informatics*, 287–304 (2005).
6. S. Qiao, T. Li, H. Li, Y. Zhu, J. Peng, J. Qiu, "SimRank: a page rank approach based on similarity measure," *Proceedings of 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering*, 390–395 (2010).
7. B. Möller, U. Reuter, "Uncertainty forecasting in engineering," Springer-Verlag, Berlin, Heidelberg (2010).
8. A. Zhou, C. Jin, G. Wang, J. Li, "A survey on the management of uncertain data," *Chinese Journal of Computers*, **32**(1), 1–16 (2009) (in Chinese).
9. R. Cheng, "Querying and cleaning uncertain data," *Proceedings of the First International Workshop of Quality of Context*, LNCS **5786**, 41–52 (2009).
10. J. J. Xu, H. Chen, "CrimeNet explorer: a framework for criminal network knowledge discovery," *ACM Transactions on Information Systems*, **23**(2), 201–226 (2005).
11. L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank citation ranking: bringing order to the web," *Technical Report*, Stanford InfoLab (1999).
12. L. Zhang, F. Ma, "Accelerated ranking: a new method to improve web structure mining quality," *Journal of Computer Research and Development*, **41**(1), 98–103 (2004) (in Chinese).
13. T. H. Haveliwala, "Topic-sensitive PageRank: a context-sensitive ranking algorithm for web search," *IEEE Transactions on Knowledge and Data Engineering*, **15**(4), 784–796 (2003).
14. M. Richardson, P. Dominggos, "The intelligent surfer: probabilistic combination of link and content information in PageRank," MIT Press, Cambridge, MA (2002).
15. R. L. Breiger, "The analysis of social networks," *Handbook of Data Analysis*, Sage Publications, London, UK, 505–526 (2004).
16. V. Batagelj, A. Mrvar, A. Ferligoj, P. Doreian, "Generalized blockmodeling with pajek," *Metodoloski Zvezki*, **1**(2), 455–467 (2004).
17. L. Freeman, "Centrality in social networks: conceptual clarification," *Social Networks*, **1**(10), 215–239 (1979).
18. H. Chen, D. Zeng, H. Atabakhsh, W. Wyzga, J. Schroeder, "COPLINK: managing law enforcement data and knowledge," *Communications of the ACM*, **46**(1), 28–34 (2003).
19. S. Wasserman, K. Faust, "Social network analysis: methods and applications," Cambridge University Press, Cambridge, UK (1994).
20. S. Qiao, T. Li, J. Peng, J. Qiu. "Parallel sequential pattern mining of massive trajectory data," *International Journal of Computational Intelligence Systems*, **3**(3), 343–356 (2010).
21. D. Ruan, "Lessons learned from soft computing applications at SCK·CEN," *International Journal of Computational Intelligence Systems*, **3**(2), 249–255 (2010).
22. S. Qiao, C. Tang, H. Jin, T. Long, S. Dai, Y. Ku, M. Chau, "PutMode: prediction of uncertain trajectories in moving objects databases," *Applied Intelligence*, **33**(3), 370–386 (2010).
23. Y. Zhu, D. Ruan, X. Zeng, L. Koehl, C. Chaigneau, "Characterization of fashion themes using fuzzy techniques for designing new human centered products," *International Journal of Computational Intelligence Systems*, **3**(4), 452–460 (2010).
24. S. Qiao, C. Tang, H. Jin, J. Peng, D. Davis, N. Han, "KISTCM: knowledge discovery system for traditional Chinese medicine," *Applied Intelligence*, **32**(3), 346–363 (2010).
25. S. Qiao, J. Peng, T. Li, H. Li, T. Li, C. Wang, "Hybrid page scoring algorithm based on centrality and PageRank," *Journal of Southwest Jiaotong University*, **46**(3), 456–460 (2011).