An Analysis of User Influence Ranking Algorithms on Dark Web Forums

Christopher C. Yang
College of Information Science and Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
Chris.Yang@drexel.edu

Xuning Tang
College of Information Science and Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104

Bhavani M. Thuraisingham
School of Engineering and Computer Science
University of Texas, Dallas

ABSTRACT
Social media is actively utilized by extremists to spread out their ideologies. While the Internet provides a platform for any users around the world to share their opinions, some opinions in social media can be related to the national security and threatening to others. Given the large volume and exponential growing rate of messages on the social media platforms, it is impossible to analyze the messages by manual effort. An effective way to identify the threat through social media is detecting the influential users automatically. By identifying the influential users, we can determine the impact and the neighborhood of these users. In this work, we develop weights to incorporate message content similarity and response immediacy features with link analysis techniques. In our experiment, we investigate the impact of weights and the basic algorithms (iterative or prestige) on the user influence ranking. The experiment is conducted on the Dark Web forum provided in the ISI-KDD Challenge. The result shows that the weights make substantial impact on the ranking results, especially on the in-degree algorithm.

Categories and Subject Descriptors
H3.3 [Information Storage and Retrieval]: Information Search and Retrieval – information filtering, selection process; H3.5 [Information Storage and Retrieval: Online Information Services – web-based services]; H5.4 [Information Interfaces and Presentations] Hypertext/Hypermedia - Navigation

General Terms
Algorithms, Management, Performance, Experimentation.

Keywords
Social networking analysis, user influence ranking, social media.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.


1. INTRODUCTION
The Dark Web forum data was provided by the Challenge of the ACM SIGKDD Workshop on Intelligence Security and Informatics (ISI-KDD) 2010 to develop techniques and algorithms to analyze over one million of messages in multiple extremist forums. In this work, we investigate the techniques in ranking the influential users in the Dark Web forums. In particular, we conduct our experiment on the Ansar AlJihad Network (Ansar1) data.

Social media is becoming so popular that almost half of the Internet users are participating in some kinds of online social networking sites such as Face book, Twitter, and MySpace. Other than these popular social networking sites, there are also many other social networking sites that are focusing on specific topics. The Dark Web portal provides data of several extremist forums, including seventeen forums in Arabic, seven forums in English, three forums in French, one forum in German, and one forum in Russian. In these forums, extremist ideologies are spreading rapidly and influencing the society. The activities in a Web forum are usually following a power law. That means a relatively fewer number of users are actively participating in the forum and leading the discussion in the community. On the other hand, a large number of users are relatively less active in their participations and mainly following the messages led by the forum leaders or the influential users. However, these influential users cannot be simply detected by the frequency of participations. In this work, we are incorporating the message similarity and response immediacy features with link analysis to identify and rank the influential users in the Dark Web forum.

1.1 Related Work
In the literature, a number of works related to influence spreading, influence optimization, and expert ranking had been conducted. Some earlier works on the adoption of medical and agricultural innovations [2][10] and viral marketing [1][7][8] had focused on discovering how the adoptions of new products or ideas were spread out through the structure of social networks. A few works had focused on influence optimization in which the problem was defined as identifying $n$ nodes in a social network such that $n$ copies of an item can be spreading out optimally from these $n$ nodes. Probabilistic model [4][5], greedy approximation algorithms [8], and heuristics [2] had been adopted to solve the optimization problem. However, influential users ranking are not
investigated in the influence spreading and influence optimization problems although they are closely related.

Recently, Zhang, Ackeman, and Adamic [11] had investigated several link analysis algorithms for expert ranking in a forum setting. However, these algorithms only focus on the link structure but have not considered other features that reflect user influence. In this work, we propose to incorporate message content similarity and response immediacy with link analysis to measure user influence in a web forum environment.

2. Weighted Social Network

In order to incorporate other features with link analysis, we introduce weights to the links in social networks. A weighted social network is denoted as \( G = (V, E, W) \), where \( V \) is a set of actors, \( E \) is a set of directed edges representing the interaction between actors, and \( W \) is the set of weights assigned to the directed edges in \( E \). Figure 1 presents a typical thread in Web forum, the transformed hierarchical structure and social network.

![Figure 1](image)

Figure 1. (a) a forum thread, (b) transformed hierarchical tree representation, (c) social network.

Typical link analysis techniques such as PageRank and Hits are applying on a non-weighted social network. That means the weights on the directed edges are uniform, which is equal to 1. However, when we are considering the influence between actors in a web forum, we cannot assume that the influence between any two actors is uniform over the social network as long as there is at least one interaction between the two actors. Some actors have more influence while some others have less influence. The purpose of the weighted social network is integrating the features that reflect influence to the social network structure.

In this work, we propose incorporating two features: content similarity and response immediacy. When a user is making influence on other users in a forum thread, the discussion should be carried on. Other users are following the topic of discussion and spreading the influence to others. On the other hand, if the topic of discussion is shifted to a different topic by a user message within a thread, the user is trying to bring up a new topic or draw attention for other means. This message is not spreading the influence but trying to raise another issue. As a result, the influence is not carried on. In some cases, these messages can be spamming or advertisements. We propose to measure the content similarity between the two messages that one user is posting and another one is replying as a feature to measure the degree of influence. The content similarity is measured by cosine similarity which is a typical measure of two document vectors in information retrieval.

We also adopt the response immediacy as another feature to measure the degree of influence. A thread usually carries on for a period of time. However, as discussion goes on, the influence will become less because some users are no longer following the thread towards the end of the discussion period. If a response is made after a long delay, it reflects that a less impact is being made to other users. On the contrary, the most influence is usually made by an immediate response. In this work, we propose to use the time difference between any two messages within a thread to measure the response immediacy.

Given a forum \( F \), there are \( N \) threads, \( T_1, T_2, ..., T_N \) that consist of messages posted by \( v_1, v_2, ..., v_n \). Let \( M_{k,j} \) be the \( j \)-th message in \( T_k \), \( V(M_{k,j}) \) be the user who posts the message \( M_{k,j} \), and \( \text{Time}(M_{k,j}) \) be the timestamp of the message \( M_{k,j} \). We compute the content similarity between two messages \( M_{k,a} \) and \( M_{k,b} \) within the same thread \( T_k \) by cosine similarity, denoted as \( \text{Similarity}(M_{k,a}, M_{k,b}) \). The response immediacy is measured by the time difference between the timestamps of two messages, \( \text{Time}(M_{k,b}) - \text{Time}(M_{k,a}) \), assuming \( M_{k,b} \) is posted after \( M_{k,a} \) in \( T_k \) (i.e. \( \text{Time}(M_{k,b}) > \text{Time}(M_{k,a}) \)). The weight between \( v_i \) and \( v_j \) is computed by the following formulation:

\[
W(v_i, v_j) = \frac{1}{M} \sum_{k=1}^{N} \sum_{a,b : \text{Time}(M_{k,a}) = v_i} \sum_{\text{Time}(M_{k,b}) > \text{Time}(M_{k,a})} \left( \beta \frac{\text{Time}(M_{k,b}) - \text{Time}(M_{k,a})}{\text{Time}(M_{k,b})} \right) \text{Similarity}(M_{k,a}, M_{k,b})
\]

where \( M \) is the number of threads that \( v_i \) has replied to \( v_j \), \( \alpha \) and \( \beta \) are parameters with values between 0 and 1.

2.1 UserRank

In this work, we propose UserRank developed on the basis of PageRank. PageRank is an algorithm developed for Web page
ranking. PageRank score is representing the authority conferring by the authors of other Web pages which is explicitly presented by the hyperlinks. If a Web page receives a hyperlink from another Web page with a high PageRank score, it is receiving authority from a high authoritative page. The higher authority a Web page received from other Web pages, the higher the ranking of this Web page is. In PageRank algorithm, every Web page is initialized with a small value and computed iteratively until convergence. PageRank has also been investigated to rank the expert level of users in forum settings [11]. The interactions between users in a forum are represented as a social network (Figure 1(c)). The nodes are corresponding to the actors who participate in the forum and the directed edges are corresponding the interaction that one actor is offering an answer to the question raised by another actor in the forum. The experts are those who are offering advice to others. The more advice a user offers, especially offering to other expert users, the higher the PageRank score this user receives. The formulation of computing PageRank score, \( PR(\cdot) \), in each iteration is presented as follow:

\[
PR(v_j) = (1 - d) + d \sum_{v_i \in E} P(v_i | v_j) PR(v_i)
\]

where \( d \) is a constant, typically between 0.8 and 1.0, \( P(v_i | v_j) \) is the transition probability from \( v_i \) to \( v_j \) which is equal to \( 1 \) / out-degree(\( v_i \))

We propose UserRank by incorporating the content similarity and response immediacy to PageRank. As a result, the transition probability is a function of the weights presented in Section 2.

\[
P(v_j | v_i) = \frac{W(v_i, v_j)}{\sum_{v_k \in E} W(v_i, v_k)}
\]

By integrating the transition probability with the PageRank function, the UserRank score, \( UR(\cdot) \) is computed as follow:

\[
UR(v_j) = (1 - d) + d \sum_{v_i \in E} \frac{W(v_i, v_j)}{\sum_{v_k \in E} W(v_i, v_k)} UR(v_i)
\]

2.2 Weighted in-degree

Centrality and Prestige measures have been widely used in social network analysis. They describe an individual’s connection with other members in a social network. Degree centrality of a node \( v_j \) measures the ratio of the out-degree of \( v_j \) and the number of nodes other than \( v_j \) in the social network. Degree prestige of a node \( v_j \) measures the ratio of the in-degree of \( v_j \) and the number of nodes other than \( v_j \) in the social network. Since we are measuring the influence of a user, we use the degree prestige that reflects the ratio of the influences conferring from other users in a social network and the total number of users in a social network except the user himself. Therefore, the higher the degree prestige is, the more influence are conferred form others in a social network. Since the size of social network is static in our experiment, we can simply use the in-degree of a user to measure the relative value of prestige.

In order to compare with the proposed UserRank algorithm, we enhance the in-degree measure by incorporating the weights corresponding the content similarity and response immediacy. The weighted in-degree is therefore computed as follow:

\[
\text{weighted in-degree}(v_j) = \sum_{v_i \in E} W(v_i, v_j)
\]

3. Experiment

In this experiment, we intent to analyze the impact of two factors in the user influence ranking algorithms. The four algorithms in this analysis are PageRank, UserRank, in-degree, and weighted in-degree. The two factors are the basic algorithms and weights. The basic algorithms are iterative algorithm and prestige measure, in which PageRank and User Rank are using an iterative algorithm while the in-degree and weighted in-degree algorithms are using prestige measure. Regarding the second factor, PageRank and in-degree do not involve any weights while UserRank and weighted in-degree involves weights that incorporate content similarity and response immediacy.

3.1 Dataset

The dataset in this experiment was extracted from the ISI-KDD Challenge of the Dark Web forums. We downloaded the English data file, ansar1.txt, at http://128.196.40.222:8080/CRI_Indexed_new/datasets/ansar1.txt

The followings present the statistics of the forums data before pre-processing.

Start Date: 12/08/2008
End Date: 01/02/2010
Number of Members: 377
Number of Threads: 11133
Number of Messages: 29056

In this dataset, some messages are purely in Arabic although the language of the data file is English. In Order to measure content similarity without any machine translation processes, we remove the pure Arabic messages in the forums. The reason of avoiding machine translation is reducing the effect of translation ambiguity since the purpose of this work is ranking user influence rather than cross-lingual information retrieval. Indeed, the number of Arabic messages is relatively small. The followings present the statistics of the forums data after removing Arabic messages.

Start Date: 12/08/2008
End Date: 01/02/2010
Number of Members: 362
Number of Threads: 11036
Number of Messages: 28745
Average Number of Message per Thread: 2.6
Average Number of Message per User: 79.4
Average Number of Thread per User: 30.5

3.2 Distance Measures

We employed two distance measures [4] of the top-k lists generated by any two user influence ranking algorithms to measure how close the two top-k lists are to each other and. These two measures are Kendall’s tau and Spearman’s footrule measures. By measuring the distances of the top-k lists, the objective is analyzing the impact of the weights and basic algorithms on the user influence ranking results.

Given two top k lists, \( \Gamma_1 \) and \( \Gamma_2 \), \( D_{\Gamma_1} \) and \( D_{\Gamma_2} \) denote the domains of \( \Gamma_1 \) and \( \Gamma_2 \), respectively. \( \Gamma_1(i) \) and \( \Gamma_2(i) \) denote the ranking of \( i \) in \( \Gamma_1 \) and \( \Gamma_2 \). \( P(\Gamma_1, \Gamma_2) \) is defined as the set of all unordered pairs of distinct elements in \( D_{\Gamma_1} \cup D_{\Gamma_2} \).

Kendall distance with penalty parameter \( p \) is defined as follow:
where $p$ is a constant value between 0 and 1.

When $p$ equals to 0, it is considered as the optimistic approach because a non-zero penalty score is assigned to the pair $(i,j)$ only if we have enough information to know that $i$ and $j$ are in the opposite order in the two top $k$ lists.

The value of $R^p_{ij}(\Gamma_1, \Gamma_2)$ is either 0 or 1 depending on the following six conditions.

a) $R^p_{ij}(\Gamma_1, \Gamma_2) = 0$ if $i$ and $j$ exist in both top $k$ lists with the same relative order;

b) $R^p_{ij}(\Gamma_1, \Gamma_2) = 1$ if $i$ and $j$ exist in both top $k$ lists while the relative order of $i$ and $j$ in one top $k$ list is different from another;

c) $R^p_{ij}(\Gamma_1, \Gamma_2) = 1$ if $i$, but not $j$, exists in one top $k$ list and $j$, but not $i$, exists in another top $k$ list;

d) $R^p_{ij}(\Gamma_1, \Gamma_2) = 1$ if both $i$ and $j$ exist in one top $k$ list where $i$ is ahead of $j$, and only $j$, but not $i$, exists in another top $k$ list; or if both $i$ and $j$ exist in one top $k$ list where $j$ is ahead of $i$, and only $i$, but not $j$, exists in another top $k$ list.

e) $R^p_{ij}(\Gamma_1, \Gamma_2) = 0$ if both $i$ and $j$ exist in one top $k$ list where $i$ is ahead of $j$, and only $i$, but not $j$, exists in another top $k$ list; or if both $i$ and $j$ exist in one top $k$ list where $j$ is ahead of $i$, and only $j$, but not $i$, exists in another top $k$ list.

f) $R^p_{ij}(\Gamma_1, \Gamma_2) = p$ if both $i$ and $j$ exist in one top $k$ list while neither $i$ nor $j$ exist in another top $k$ list; $p$ is a constant value between 0 and 1.

Spearman’s footrule measure $F^{k+1}(\Gamma_1, \Gamma_2)$ is defined as:

$$F^{k+1}(\Gamma_1, \Gamma_2) = \sum_{i \in D_1 \cap D_2} |\Gamma_1'(i) - \Gamma_2'(i)|$$

where $\Gamma_1'(i) = \Gamma_1(i)$ if $i \in \Gamma_1$; otherwise, $\Gamma_1'(i) = k + 1$.

$\Gamma_2'(i) = \Gamma_2(i)$ if $i \in \Gamma_2$; otherwise, $\Gamma_2'(i) = k + 1$.

Kendall’s tau measure takes the relative ranking orders of any two elements in the union of two top $k$ lists into account while Spearman’s footrule measure the absolute distance between the rankings of the same element in the union of two top $k$ lists into consideration. In both measures, a smaller value corresponds to a closer pair of top $k$ lists.

Ideally, to compare the performance a number of ranking algorithms, we measure the distance between the top $k$ list generated by each ranking algorithm with the top $k$ list of the gold standard. The gold standard is usually created by human annotators. As discussed in [1], in an information retrieval task on a very large corpus or in the context of the World Wide Web, there is no clear notion of what the gold standard or ground truth is. It is almost impossible to have human annotators to go through a large number of items and rank their relevance, importance, or influence. It is similar to our experiment on comparing and analyzing the user influence rankings generated by a few algorithms. There is no clear gold standard that we can compare the top $k$ list generated by each algorithm with.

In this experiment, the objective is analyzing the impact of different factors in these algorithms by comparing the distance measures of the top $k$ lists created by all possible pairs of algorithms. The two factors are the basic algorithm and weights. In this work, we focus on analyzing the top $k$ lists generated by four algorithms, including PageRank, UserRank, in-degree, and weighted in-degree ($W$-in-degree). PageRank and UserRank are using the same basic algorithm, which is an iterative algorithm, while in-degree and $W$-in-degree are using another basic algorithm, which is a prestige measurement. PageRank and in-degree do not involve weights in the algorithm while UserRank and $W$-in-degree involve weights in the algorithm. Tables 1 and 2 present the commons and differences among the four algorithms. This analysis helps us to understand how weights affect the ranking generated by an algorithm. (In the future, we may consider randomly selecting $k$ users and recruiting human annotators to create the ranking of these $k$ users as the gold standard. This gold standard can be used in the ISI-KDD Challenge for any interested participants to measure the performance of their proposed techniques. The drawback of this approach is comparing the randomly selected $k$ users rather than the top $k$ users.)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>in-degree</th>
<th>UserRank</th>
<th>W-in-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>w</td>
<td>b</td>
<td>×</td>
</tr>
<tr>
<td>UserRank</td>
<td></td>
<td></td>
<td>w</td>
</tr>
</tbody>
</table>

Note: b denotes basic algorithms, w denotes weights, and × denotes none.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>in-degree</th>
<th>UserRank</th>
<th>W-in-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>PageRank</td>
<td>b</td>
<td>w</td>
<td>b, w</td>
</tr>
<tr>
<td>UserRank</td>
<td></td>
<td></td>
<td>b</td>
</tr>
</tbody>
</table>

Note: b denotes basic algorithms and w denotes weights.

### 3.3 Experimental Result

Tables 3 and 4 present the results of the Kendall’s Tau measure and the Spearman’s Footnote measure of the top 10, 20, 30, 40, and 50 lists generated by PageRank, UserRank, in-degree, and weighted in-degree algorithms. Figures 2 and 3 show the plots of the Kendall’s Tau measure and the Spearman’s Footnote measure. As shown in Figures 2 and 3, both measures present similar result. That means, the orders of distance measures between any two pairs of algorithms is approximately the same no matter if the distance measure is using the relative ranking or the absolute distance in the measurement.

Table 5 presents the order of the distance measure between the top $k$ lists of the six possible pairs of algorithms and the difference and common of these pairs of algorithms. It shows that the distance between the top $k$ lists of PageRank and in-degree is the shortest. The distances between the top $k$ lists of PageRank and UserRank and between the top $k$ lists of UserRank and in-degree are approximately the same. Their distances are the second shortest. The distance between the top $k$ lists of PageRank and weighted in-degree and the distance between the top $k$ lists of in-degree and weighted in-degree are approximately the same.
distances are the third shortest. The distance between the top k lists of UserRank and weighted in-degree are the furthest.

Table 3. Kendall’s Tau measures

<table>
<thead>
<tr>
<th>$k$</th>
<th>$K^k(P_k, U_k)$</th>
<th>$K^k(I_k, W_Ik)$</th>
<th>$K^k(U_k, W_Ik)$</th>
<th>$K^k(P_k, I_k)$</th>
<th>$K^k(P_k, W_Ik)$</th>
<th>$K^k(U_k, I_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>55</td>
<td>71</td>
<td>70</td>
<td>34</td>
<td>61</td>
<td>37</td>
</tr>
<tr>
<td>20</td>
<td>175</td>
<td>237</td>
<td>280</td>
<td>105</td>
<td>216</td>
<td>132</td>
</tr>
<tr>
<td>30</td>
<td>315</td>
<td>528</td>
<td>580</td>
<td>195</td>
<td>531</td>
<td>310</td>
</tr>
<tr>
<td>40</td>
<td>555</td>
<td>873</td>
<td>982</td>
<td>329</td>
<td>864</td>
<td>594</td>
</tr>
<tr>
<td>50</td>
<td>838</td>
<td>1318</td>
<td>1537</td>
<td>451</td>
<td>1313</td>
<td>850</td>
</tr>
</tbody>
</table>

Note: $P_k$ denotes the top $k$ list created by PageRank
$U_k$ denotes the top $k$ list created by UserRank
$I_k$ denotes the top $k$ list created by in-degree
$W_Ik$ denotes the top $k$ list created by W-in-degree

Table 4. Spearman’s Footnote

<table>
<thead>
<tr>
<th>$k$</th>
<th>$F^{k+1}(P_k, U_k)$</th>
<th>$F^{k+1}(I_k, W_Ik)$</th>
<th>$F^{k+1}(U_k, W_Ik)$</th>
<th>$F^{k+1}(P_k, I_k)$</th>
<th>$F^{k+1}(P_k, W_Ik)$</th>
<th>$F^{k+1}(U_k, I_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>74</td>
<td>88</td>
<td>86</td>
<td>52</td>
<td>76</td>
<td>46</td>
</tr>
<tr>
<td>20</td>
<td>240</td>
<td>286</td>
<td>322</td>
<td>154</td>
<td>254</td>
<td>174</td>
</tr>
<tr>
<td>30</td>
<td>444</td>
<td>626</td>
<td>684</td>
<td>290</td>
<td>584</td>
<td>396</td>
</tr>
<tr>
<td>40</td>
<td>728</td>
<td>1040</td>
<td>1130</td>
<td>458</td>
<td>1022</td>
<td>722</td>
</tr>
<tr>
<td>50</td>
<td>1086</td>
<td>1560</td>
<td>1724</td>
<td>622</td>
<td>1548</td>
<td>1104</td>
</tr>
</tbody>
</table>

Note: $P_k$ denotes the top $k$ list created by PageRank
$U_k$ denotes the top $k$ list created by UserRank
$I_k$ denotes the top $k$ list created by in-degree
$W_Ik$ denotes the top $k$ list created by W-in-degree

Figure 2. Kendall’s tau measures of top 10 to top 50 lists

Figure 3. Spearman’s footnote measures of top 10 to top 50 lists

It is important to note that the distances between the top k lists of two algorithms are the shortest and furthest when the basic algorithms are different. However, it is the shortest when weights are not involved and it is the furthest when weights are involved. First, the difference in the basic algorithms makes the least impact on their top k lists. Second, the weights are making the rankings of
two basic algorithms substantially different from each other although we are applying the same weights on two basic algorithms.

When the basic algorithm is iterative (i.e. PageRank and UserRank) but one is not applying weights and one is applying weights, the distance of their top \( k \) lists are the second shortest. When the basic algorithm is prestige measure (i.e. in-degree and weighted in-degree) but one is not applying weights and one is applying weights, the distances of their top \( k \) lists are the third shortest. It shows that the weights have more impact on the prestige algorithm (in-degree) than the iterative algorithm (PageRank).

The distances between the top \( k \) lists of two algorithms are not the furthest when they are different in both basic algorithms and weights. Indeed, they are the second (UserRank and in-degree) and third shortest (PageRank and weighted in-degree). Similarly, when weights are applying to the iterative algorithm (UserRank), the impact is less.

Table 5. Order of closeness between the top \( k \) lists generated by the six possible pairs of algorithms

<table>
<thead>
<tr>
<th>Distance Measure</th>
<th>Pair of Algorithms</th>
<th>Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (shortest)</td>
<td>basic algorithms</td>
<td>PageRank and indegree</td>
</tr>
<tr>
<td>2</td>
<td>weights</td>
<td>PageRank and UserRank</td>
</tr>
<tr>
<td></td>
<td>basic algorithms, weights</td>
<td>UserRank and indegree</td>
</tr>
<tr>
<td>3</td>
<td>basic algorithms, weights</td>
<td>PageRank and weighted indegree</td>
</tr>
<tr>
<td></td>
<td>weights</td>
<td>indegree and weighted indegree</td>
</tr>
<tr>
<td>4 (furthest)</td>
<td>basic algorithm</td>
<td>PageRank and weighted indegree</td>
</tr>
</tbody>
</table>

As a reference from another experiment we have conducted on MedHelp, which is a forum focused on health topics, the order of closeness between the top \( k \) list generated by the algorithm and the top \( k \) list of the gold standard is: weighted in-degree, UserRank, in-degree, and PageRank (i.e. weighted in-degree is the closest to the gold standard). This result concurs with the expert ranking result [11] that PageRank does not produce better ranking than in-degree. In MedHelp, the size of forum is relatively smaller. The three forums we have investigated are Swine Flu, Smoking Addiction, and Alcoholism forums. These forums have the number of messages ranging from a few hundreds to a few thousands. Therefore, we managed to have human annotators ranking the influential users in these forums. The order of closeness to the gold standard is consistent across the three MedHelp forums. The results in the MedHelp experiment are the distance measure from the top \( k \) list of each algorithm to the top \( k \) list of the gold standard rather than the distance measure between the top \( k \) lists of any two algorithms. The result in the MedHelp experiment cannot be directly used to interpret the result of the Dark Web experiment because the datasets are not the same. However, it may give an implication of the impact of the two factors. As we observe in the Dark Web experiment, the weights are making substantial impacts, especially on the in-degree algorithm. We are also observing from the MedHelp experiment, the weights are improving the ranking.

4. Conclusion

Dark Web portal is a large collection of extremist forums. The goal of ISI-KDD Challenge is finding radical and infectious threads, members, postings, ideas and ideologies. In this work, we propose incorporating content similarity and response immediacy with link analysis in ranking influential users in the Dark Web forums. We introduce the weights in forum social network to reflect the degree of influence which are computed by the cosine similarity and time difference of two messages. In our experiment, we find that the weights make a substantial impact on the ranking result, especially on the degree prestige algorithm. Although we cannot conclude whether the weights improve the ranking performance and whether the iterative algorithm (PageRank and UserRank) or the degree prestige algorithm (in-degree and weighted in-degree) perform in this experiment, the experiment conducted on another health related dataset shows consistently that the weighted in-degree produced the best performance and followed by UserRank, in-degree and PageRank. We cannot use the result on another dataset to imply that the weights can improve the ranking result in this experiment but the two algorithms with weights make the most difference in the ranking results. In the future, we may develop the ground truth to measure the relative performance of these algorithms.

5. REFERENCES

[4] P. Domingos and M. Richardson, Mining knowledge-sharing sites for viral marketing, Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, FL, USA, 2002, pp. 61-70
[7] J. Goldenberg, B. Libai and E. Muller, Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth


