ERWD: A Measure for Nearest-Neighbor Search in Undirected Graph

Junyin Wei
School of Computer Science and Technology
Donghua University
Shanghai, China
E-mail: jywei@dhu.edu.cn

Binghui Qi
School of Computer Science and Technology
Zhengzhou Institute of Aeronautical Industry Management
Zhengzhou, China
E-mail: 787145210@qq.com

Mingxi Zhang
School of Computer Science
Fudan University
Shanghai, China
E-mail: WAXL7461@aliyun.com

Abstract—Finding nearest neighbors in graph plays an increasingly important role in various applications, such as graph clustering, query expansion, recommendation system, etc. To tackle this problem, we need compute the most “similar” k vertices for the given vertex. One popular class of similarity measures is based on random walk approach on graphs. However, these measures consider each co-occurrence frequency of two vertices is equivalent, means that each occurrence of two vertices is not differentiated, and the influence of the vertices have not been considered enough. In this paper, we proposed an effective distance measure based on random walk distance, called ERWD, for nearest-neighbor search in undirected graph. The Relationship Strength (RS) of two vertices, which affects ERWD, is proposed firstly, and a model for measuring RS is established according to their structural characteristics and influences of the vertices. Extensive experimental results demonstrate the effectiveness of ERWD through comparison with the common random walk distance.

Keywords—similarity measure; random walk distance; Relationship Strength

I. INTRODUCTION

Finding nearest neighbors in undirected graph plays an increasingly important role in various applications, e.g., graph clustering [1, 2, 3, 4, 5, 6], query expansion [7,8], recommendation system [9]. To tackle this problem, we need resolve the computation of the most “similar” k vertices for the given vertex. Measuring “similarity” via graph structure, which is used to find the desired similar objects, has received great attentions recently. Existing measures may often focus on topological structure of graph [10, 11, 12, 13], attributes of the vertex [14, 15], and structural/attribute [16]. Traditional topological structure-based measure methods focused on the connectivity and structural characteristics. For example, the number of possible paths between two vertices, the number of common neighbors of two vertices etc.

One popular class of similarity measures is based on Random Walks Distance (RWD) in undirected graphs [17, 18, 19]. There are two problems in these methods. Firstly, most of them consider each co-occurrence frequency of two vertices is equivalent, means that each occurrence of two vertices is not differentiated. Weight is measured only according to the co-occurrence frequency between two vertices. For example, in the co-author network, the weight between two authors is only the frequency of the cooperation, which is not differentiated. Secondly, the influence of the vertex has not been considered enough. For example, the score that a given author cooperated with a high prolific author once is equal to the score cooperated with a low prolific author, without considering the influence of the cooperated author.

However, in many applications, the co-occurrence between two vertices may not be equivalent, and the influence of vertex is very important for finding the similar vertices. For example, in the social network a handshake between two public figures are very normal, but a handshake between a public figure and a pipsqueak may give many people a surprising. Because we think that acquaintanceship between two public figures is normal, and hence we never think it is normal that there is an acquaintanceship between a public person and a pipsqueak. So it is significant to reconsider the influence of the vertex and the importance of the co-occurrence between two vertices.

Example 1 Figure 1 shows an illustrating example of a coauthor graph where a vertex represents an author, the edge represents the coauthor relationship of two authors and the weight is the frequency of the cooperation. In this figure, “Jianwei Han” is highly prolific author of data mining, so he has many coauthors, while others not. For a given author “Nan Wang”, we want to find the most nearest authors to this author. According to the traditional random-walk-based measure, we find that the closeness between “Jianwei Han” and “Nan Wang” is almost equal to the closeness between “Li Yangyou Chen” and “Nan Wang”.

Figure 1. A coauthor network example with the frequency of the cooperation

© 2013. The authors - Published by Atlantis Press
Definition 1. Undirected Weighted Graph. Let \( G = (V, E, W) \) be Undirected Weighted Graph, where \( V \) denotes the set of vertices, \(|V|\) denotes the number of vertices, \( v_i \in V, k = 1, 2, 3, ..., |V| \) is the vertex of \( G : E \) is the set of edges, and \(|E|\) is the number of edges, \( e(v_i, v_j) \in E \) is the edge between vertices \( v_i \in V \) and \( v_j \in V \), also denoted as \((v_i, v_j) \in E \). \( W \) is the set of weight, and \( w(v_i, v_j) \in W \) is the weight \((v_i, v_j)\).

Definition 2. Relationship Strength (RS). Let \( RS \) be the set of Relationship Strength, \( RS(v_i, v_j) \in RS \) be the RS of \( v_i \) and \( v_j \), where \( v_i, v_j \in V \).

RS is mainly to measure effective connection strength of the two vertices. More details will be discussed in the next section.

Definition 3. Undirected Weighted Graph with RS. Let \( G = (V, E, RS) \) be Undirected Weighted Graph with RS, where \( V \) is the set of vertex, \(|V|\) is the number of vertex, \( v_i \in V, k = 1, 2, 3, ..., |V| \) is the vertex of \( G \). \( E \) is the set of edges, and \(|E|\) is the number of edges, \( e \in E \) if the edge. \( RS \) is the set of Relationship Strength, and \( RS(v_i, v_j) \in RS \) is the Relationship Strength of \((v_i, v_j)\).

III. RELATIONSHIP STRENGTH

We have noticed that the more prolific the author, the weaker connection strength between this author and other author, and vice versa. The more influence the vertex, the weaker connection strength for other vertex, and vice versa. In this section we will analyze the factors which affect the strength of connection between two vertices, and we proposed RS for this strength measure, which is used to measure effective relationship strength of two connected vertices.

3.1. Intuitive Analysis of Factors which Affect Relationship Strength

For given undirected weighted graph \( G = \{V, E, W\} \), we chose \( \forall (v_i, v_j) \in E \), where \( v_i, v_j \in V \), \( w(v_i, v_j) \in W \). Combine with the coauthor network, we analyze the factors which affect \( RS(v_i, v_j) \), which are shown as followed two aspects.

1. The importance of connected vertices. We assume that an author is highly prolific and have cooperated with many other authors, so we would not surprise if there is a connection between this author and other author, then this connection strength should be relatively weak, and vice versa. If his or her collaborator is also a highly prolific author, then the connection strength will be weaker, this connection can not make much contribution for nearest vertices search, and vice versa. In a graph, the affluence of the vertex can be considered as the value of the sum of the weights between a given vertex and its neighbors.

Then we can infer that, for two vertices \( v_i \) and \( v_j \), when the weight \( w(v_i, v_j) \) is certain, the bigger the \( \sum_{j \in N(v_i)} w(v_i, v_j) \) is, the weaker the connection strength \( RS(v_i, v_j) \) is, and vice versa, where \( N(v_i) \) is the set of the neighbor of vertex \( v_i \). This circumstances is same for vertex \( v_j \).

2. The weight between two vertices. We assume that the two authors are certain, the cooperation frequency is
important for measure the strength. Intuitively, the higher cooperation frequency, the more closed the two authors, and vice versa. So we hold that the higher of frequency of the cooperation, the strength of the RS, and vice versa.

Then we can infer that, if \( \sum_{q \in N(v_i)} w(v_i, v_q) \) and \( \sum_{q \in N(v_j)} w(v_j, v_q) \) is certain, the bigger \( w(v_i, v_j) \) is, the stronger the connection strength \( RS(v_i, v_j) \) is, and vice versa.

3.2. Measure of Relationship Strength

For given undirected weighted graph \( G = (V, E, W) \), \((v_i, v_j) \in E\), where \( v_i, v_j \in V \), and \( w(v_i, v_j) \in W \). We have analyzed that the Relationship Strength \( RS(v_i, v_j) \) is not only depends on the weight \( w(v_i, v_j) \) of the two vertices, but also depends on the sum of weights of neighbors of each vertices.

Let \( RS \) be the set of RS, and \( RS(v_i, v_j) \in RS \) is the RS of vertices \( v_i \) and \( v_j \). Then we formalize the Relationship Strength of Definition 2 as follows:

\[
RS(v_i, v_j) = \frac{w(v_i, v_j)}{\sum_{q \in N(v_i)} w(v_i, v_q) \sum_{q \in N(v_j)} w(v_j, v_q)}
\]

IV. EFFECTIVE DISTANCE

Random walk process is well known method for the relativity measure. The number of possible paths between two vertices, the number of common neighbors of two vertices and the weight of edges are all considered in this measure. It is also works well in many fields to find the most similar objects. The most related vertex usually is the most similar vertex. In this section, we proposed a new measure based on the random walk distance.

4.1. Random Walk Distance

In a large graph \( G \), some vertices are close to each other while some other vertices are far apart based on connectivity. If there are multiple paths connecting two vertices \( v_i \) and \( v_j \), then they are close. On the other hand, if there are very few or no paths between \( v_i \) and \( v_j \), then they are far apart. In this paper, we use neighborhood random walk distances to measure vertex closeness.

Definition 4. Neighborhood Random Walk Distance.

Let \( P \) be the \( N \times N \) transition probability matrix of a graph \( G \). Given \( l \) as the length that a random walk can go, \( c \in (0,1) \) as the restart probability, the neighborhood random walk distance \( d(v_i, v_j) \) from \( v_i \) to \( v_j \) is defined as

\[
d(v_i, v_j) = \sum_{r, \tau \in \tau_i, \tau_j} p(\tau) c(1-c)^{\text{length}(r)}
\]

where \( \tau \) is a path from \( v_i \) to \( v_j \), whose length is \( \text{length}(\tau) \) with transition probability \( p(\tau) \).

The matrix form of the neighborhood random walk distance is

\[
R^l = \sum_{k=1}^{l} c(1-c)^{k} P^k
\]

Here, \( P \) is the transition probability matrix for graph \( G \), and \( R \) is the neighborhood random walk distance matrix.

According to the equation, the structural closeness between two vertices \( v_i \) and \( v_j \) is

\[
d_{erwd}(v_i, v_j) = R^l(v_i, v_j)
\]

4.2. ERWD: Connection to Absorption Random Walk

Given a Undirected Weighted Graph \( G = (V, E, W) \). \( w(v_i, v_j) \in W \) is the weight of edge \((v_i, v_j) \in E\), where \( v_i, v_j \in V \). The original transition probabilities in \( G \) based on weight can be described as follows.

\[
P(v_i, v_j) = \frac{w(v_i, v_j)}{\sum_{q \in N(v_i)} w(v_i, q) + \sum_{q \in N(v_j)} w(q, v_j)}
\]

where \( P(v_i, v_j) \) is the probability of transition from \( v_i \) to \( v_j \).

Given Undirected Weighted Graph with Relationship Strength \( G = (V, E, RS) \). \( RS(v_i, v_j) \in RS \) is the RS of \( v_i \in V \) and \( v_j \in V \). We define our improved transition probabilities in \( G \) based on RS as follows.

\[
E_{effecDis}(v_i, v_j) = \frac{RS(v_i, v_j)}{\sum_{q \in N(v_i)} RS(v_i, q) + \sum_{q \in N(v_j)} RS(q, v_j)}
\]

Let \( E_{effecDis} = E_{effecDis}(v_i, v_j) \) be the transition probability matrix. The matrix form of the neighborhood random walk distance is

\[
E_{effecDis}^l = \sum_{k=1}^{l} c(1-c)^{k} E_{effecDis}^k
\]

where \( l \) as the length that a random walk can go, \( c \in (0,1) \) as the restart probability, the neighborhood random walk distance \( E_{effecDis}^l(v_i, v_j) \) from \( v_i \) to \( v_j \) is defined as

\[
E_{effecDis}^l(v_i, v_j) = E_{effecDis}^l(v_i, v_j)
\]

Definition 5. Effective Distance (ERWD).

Let \( E_{effecDis}(v_i, v_j) \) be the effective distance, which is defined based on the above discussions as follows.

\[
E_{effecDis}(v_i, v_j) = d_{erwd}(v_i, v_j) = R^l(v_i, v_j)
\]

\( E_{effecDis}(v_i, v_j) \) is used to measure the relevance of the vertex \( v_i \) and \( v_j \). The higher \( E_{effecDis}(v_i, v_j) \), the more closed the two vertices. As far as we know, the relevance
The measure problem is transformed into the computation of ERWD. This computational complexity can be reduced with the recent research result on fast random walk computation.

Table 1. Authors similar to “Nan Wang” base on RWD

<table>
<thead>
<tr>
<th>Order</th>
<th>Author list of RWD</th>
<th>RWD Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hasan M. Jamil</td>
<td>0.106857</td>
</tr>
<tr>
<td>2</td>
<td>Liangyou Chen</td>
<td>0.0980396</td>
</tr>
<tr>
<td>3</td>
<td>Jiawei Han</td>
<td>0.0352372</td>
</tr>
<tr>
<td>4</td>
<td>Jian Pei</td>
<td>0.0314164</td>
</tr>
<tr>
<td>5</td>
<td>Ying Lu</td>
<td>0.0254197</td>
</tr>
<tr>
<td>6</td>
<td>Yaqin Liao</td>
<td>0.0239645</td>
</tr>
<tr>
<td>7</td>
<td>Avigdor Gal</td>
<td>0.0158815</td>
</tr>
<tr>
<td>8</td>
<td>Giovanni A. Modica</td>
<td>0.0106741</td>
</tr>
<tr>
<td>9</td>
<td>Wei Wang</td>
<td>0.0050759</td>
</tr>
<tr>
<td>10</td>
<td>Anthony K. H. Tung</td>
<td>0.0035467</td>
</tr>
<tr>
<td>11</td>
<td>Philip S. Yu</td>
<td>0.0027789</td>
</tr>
<tr>
<td>12</td>
<td>Xin Xu</td>
<td>0.0027185</td>
</tr>
<tr>
<td>13</td>
<td>Laks V. S. Lakshmanan</td>
<td>0.0026241</td>
</tr>
<tr>
<td>14</td>
<td>Ke Wang</td>
<td>0.0024487</td>
</tr>
<tr>
<td>15</td>
<td>Guozhu Dong</td>
<td>0.0021096</td>
</tr>
<tr>
<td>16</td>
<td>Xifeng Yan</td>
<td>0.0020292</td>
</tr>
<tr>
<td>17</td>
<td>Dong Xin</td>
<td>0.0017564</td>
</tr>
<tr>
<td>18</td>
<td>Xiaoqin Li</td>
<td>0.0016008</td>
</tr>
<tr>
<td>19</td>
<td>Louisa Raschid</td>
<td>0.0015892</td>
</tr>
<tr>
<td>20</td>
<td>Danilo Montesi</td>
<td>0.0015319</td>
</tr>
</tbody>
</table>

Table 2. Authors similar to “Nan Wang” base on ERWD

<table>
<thead>
<tr>
<th>Order</th>
<th>Author list of ERWD</th>
<th>ERWD Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Liangyou Chen</td>
<td>0.142776</td>
</tr>
<tr>
<td>2</td>
<td>Hasan M. Jamil</td>
<td>0.13685</td>
</tr>
<tr>
<td>3</td>
<td>Yaqin Liao</td>
<td>0.0334393</td>
</tr>
<tr>
<td>4</td>
<td>Giovanni A. Modica</td>
<td>0.027001</td>
</tr>
<tr>
<td>5</td>
<td>Ying Lu</td>
<td>0.0247753</td>
</tr>
<tr>
<td>6</td>
<td>Avigdor Gal</td>
<td>0.0158442</td>
</tr>
<tr>
<td>7</td>
<td>Jian Pei</td>
<td>0.0030826</td>
</tr>
<tr>
<td>8</td>
<td>Haggai Rostman</td>
<td>0.0026896</td>
</tr>
<tr>
<td>9</td>
<td>Xin Xu</td>
<td>0.0020880</td>
</tr>
<tr>
<td>10</td>
<td>Gang Xu</td>
<td>0.0015234</td>
</tr>
<tr>
<td>11</td>
<td>Ateret Anbush-Tavor</td>
<td>0.0013058</td>
</tr>
<tr>
<td>12</td>
<td>Jiawei Han</td>
<td>0.0012486</td>
</tr>
<tr>
<td>13</td>
<td>Alberto Trombetta</td>
<td>0.0010710</td>
</tr>
<tr>
<td>14</td>
<td>Vijayalakshmi Altaf</td>
<td>0.0009780</td>
</tr>
<tr>
<td>15</td>
<td>Qiang Ye</td>
<td>0.0009273</td>
</tr>
<tr>
<td>16</td>
<td>Anthony K. H. Tung</td>
<td>0.0009088</td>
</tr>
<tr>
<td>17</td>
<td>Danilo Montesi</td>
<td>0.0009088</td>
</tr>
<tr>
<td>18</td>
<td>Vladimir Zadorozhny</td>
<td>0.0004098</td>
</tr>
<tr>
<td>19</td>
<td>Gao Cong</td>
<td>0.0003808</td>
</tr>
<tr>
<td>20</td>
<td>Wei Wang</td>
<td>0.0003725</td>
</tr>
</tbody>
</table>

V. EXPERIMENTAL STUDY

All experiments were done on a 2.93GHz Intel(R) Core(TM)2 PC with 2GB main memory, running Windows XP. All algorithms were implemented in C++ and compiled using Visual C++ 6.0 compiler, except that matrices of RWD and ERWD was computed by Matlab.

We use the coauthor of DBLP data from four international conferences of SIGMOD, VLDB, ICED, EDBT before the year of 2010. We build a coauthor graph with all the 9489 authors and their coauthor relationships. The edges of the graph are the coauthor relationships. And the weight of each edge is the frequency of the coauthor relationship. We focused comparison of accuracy of RWD and the ERWD. Because the time complexity of them is in the same, so performance and scalability is not the focus of this paper of our experiment.

5.1. Effectiveness

Table 1 lists the authors closed to “Nan Wang” based on RWD, where the order of “Jiawei Han” is 3, and Liangyou Chen is 2. Table 2 lists the authors closed to “Nan Wang” based on EffecD, where the order of “Jiawei Han” is 12, and Liangyou Chen is 1. As our analysis in introduction, “Jiawei Han” is the high prolific author, the relationship strength between him and prolific or low prolific author should be more weak. Obviously, table 2 can reflect the closeness of the authors more accurately described in the example 1. The phenomenon demonstrates that ERWD weakened the connection between the vertex and the more affluent vertex.

VI. CONCLUSION

In the paper, we proposed a novel measure method, ERWD based on random walk to finding nearest-neighbor in undirected graph. We firstly studied the characteristic of the vertex and the edge, and established a model for RS, then proposed the formula for ERWD. The experimental results show that ERWD can improve order of the similar vertices to the given vertex.

REFERENCES


