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.

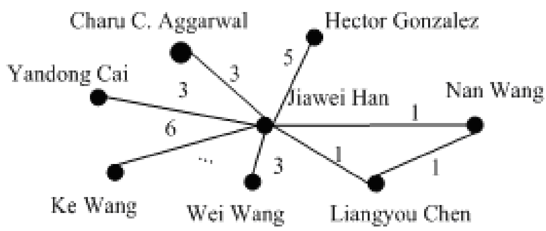

Figure 1. A coauthor network example with the frequency of the cooperation

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 "Jiawei Han" and "Nan Wang" is almost equal to the closeness between "Liangyou Chen" and "Nan Wang",* *because the distance from "Jiawei Han" to "Nan Wang" is almost equal to the distance from "Jiawei Han" to "Koperski". However, "Jiawei Han" is the famous scholar and has high prestige. "Liangyou Chen" is only the common scholars. So we can not accept the fact that "Jiawei Han" and "Liangyou Chen" are almost equal. In intuitive, we tend to accept the fact that the low prolific author "Liangyou Chen" is more closed to "Nan Wang" than "Jiawei Han" during measuring the similarity. So the relationship between "Nan Wang" and "Liangyou Chen" should be stronger than that between "Nan Wang" and "Jiawei Han".*

Through the example above, we notice that the each co-occurrence of the vertex of same type should be not equivalent, and the influence of vertex is very important for finding the similar vertices. For the given vertex, we should consider which vertex is more closed. When finding the more closed vertex, we need to measure the effective random walk distance by weakening the unnecessary possible path. This depends on the relationship strength between two vertices, which is used to measure effective connection strength between two vertices. So how to measure the RS is very significant since it can improve effectiveness of nearest neighbors search.

In this paper, we propose ERWD, an effective distance measure based on random walk, for finding nearest-neighbors in undirected graph. The factors, which can affect the effective distance, are analyzed theoretically. And a model, called Relationship Strength (RS), for measuring these factors is established according to their structural characteristics and influences of the vertices. Based on the RS and RWD, we formalized ERWD. Besides, we also extend the RS measure to other type graphs, such as directed graph, multi-type graph etc. Extensive experimental results demonstrate the effectiveness of ERWD through comparison with the RWD.

The rest of this paper is organized as follows. Section 2 introduces the preliminary concepts and proposed the conception of RS. Section 3 analyzed the factors which affect RS and then proposed our measure of it. We proposed the measure method of ERWD in Section 4. Section 5 presents extensive experimental results, followed by related work on similarity measure in directed graph, multi-type graph in Section 6. Finally, Section 7 concludes the paper.

II. Problem Statement

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, $v_k \in V, k = 1, 2, 3, ..., |V|$ 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, $v_k \in V, k = 1, 2, 3, ..., |V|$ 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 $G = \langle V, E, W \rangle$, we chose $\forall (v_u, v_v) \in E$, where $v_u, v_v \in V$, $w(v_u, v_v) \in W$. Combine with the coauthor network, we analyze the factors which affect $RS(v_u, v_v)$, 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_u and v_v, when the weight $w(v_u, v_v)$ is certain, the bigger the $\sum_{p \in N(v_u)} w(v_u, v_p)$ is, the weaker the connection strength $RS(v_u, v_v)$ is, and vice versa, where $N(v_u)$ is the set of the neighbor of vertex v_u. This circumstances is same for vertex v_v.

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 $\sum_{p\in N(v_u)} w(v_u, v_p)$ and $\sum_{q\in N(v_v)} w(v_v, v_q)$ is certain, the bigger $w(v_u, v_v)$ is, the stronger the connection strength $RS(v_u, v_v)$ is, and vice versa.

3.2. Measure of Relationship Strength

For given undirected weighted graph $G = \langle V, E, W\rangle$, $(v_u, v_v) \in E$, where $v_u, v_v \in V$, and $w(v_u, v_v) \in W$. We have analyzed that the Relationship Strength $RS(v_u, v_v)$ is not only depends on the weight $w(v_u, v_v)$ 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 $RS(v_u, v_v) \in RS$ is the RS of vertices v_u and v_v. Then we formalize the Relationship Strength of Definition 2 as follows:

$$RS(v_u, v_v) = \frac{w(v_u, v_v)^2}{\sum_{p\in N(v_u)} w(v_u, v_p) \sum_{q\in N(v_v)} w(v_v, v_q)}$$

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

$$d(v_i, v_j) = \sum_{\substack{\tau: v_i \leftrightarrow v_j \\ length(\tau)\le l}} p(\tau) c(1-c)^{length(\tau)}$$

where τ is a path from v_i to v_j whose length is $length(\tau)$ 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

$$d_s(v_i, v_j) = R^l(v_i, v_j)$$

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.

$$P(v_i, v_j) = \frac{w(v_i, v_j)}{\sum_{v_k\in N(v_i)} w(v_i, v_k)}$$

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.

$$Effec\_P(v_i, v_j) = \frac{RS(v_i, v_j)}{\sum_{v_k\in N(v_i)} RS(v_i, v_k)}$$

Let $Effec\_P = \left(Effec\_P(v_i, v_j)\right)$ be the transition probabilities matrix. The matrix form of the neighborhood random walk distance is

$$Effec\_R^l = \sum_{k=1}^{l} c(1-c)^k Effec\_P^k$$

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

$$Effec\_r^l(v_i, v_j) = Effec\_R^l(v_i, v_j)$$

Definition 5. Effective Distance (ERWD). Let $EffecDis$ be the effective distance, which is defined based on the above discussions as follows.

$$EffecDis(v_i, v_j) = r^l(v_i, v_j) = R^l(v_i, v_j)$$

$EffecDis(v_i, v_j)$ is used to measure the relevance of the vertex v_i and v_j. The higher $EffecDis(v_i, v_j)$, the more closed the two vertices. As far as we know, the relevance 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

Given author : Nan Wang		
Order	Author list of RWD	RWD Score
1	Hasan M. Jamil	0.106857
2	**Liangyou Chen**	0.0980396
3	**Jiawei Han**	0.0352372
4	Jian Pei	0.031s1464
5	Ying Lu	0.0254197
6	Yaqin Liao	0.0239645
7	Avigdor Gal	0.0158828
8	Giovanni A. Modica	0.0106741
9	Wei Wang	0.00450759
10	Anthony K. H. Tung	0.0035467
11	Philip S. Yu	0.00277789
12	Xin Xu	0.00271825
13	Laks V. S. Lakshmanan	0.00262149
14	Ke Wang	0.00244871
15	Guozhu Dong	0.00210961
16	Xifeng Yan	0.0020292
17	Dong Xin	0.00171564
18	Xiaolei Li	0.00160081
19	Louiqa Raschid	0.00158922
20	Danilo Montesi	0.00153194

Table 2. Authors similar to "**Nan Wang**" base on ERWD

Given author : Nan Wang		
Order	Author list of ERWD	ERWD Score
1	**Liangyou Chen**	0.142776
2	Hasan M. Jamil	0.13685
3	Yaqin Liao	0.0334393
4	Giovanni A. Modica	0.027001
5	Ying Lu	0.0247753
6	Avigdor Gal	0.0158442
7	Jian Pei	0.00308266
8	Haggai Roitman	0.00268965
9	Xin Xu	0.00208803
10	Gang Xu	0.00152345
11	Ateret Anaby-Tavor	0.00130588
12	**Jiawei Han**	0.0012486
13	Alberto Trombetta	0.00107109
14	Vijayalakshmi Atluri	0.000978057
15	Qiang Ye	0.000927326
16	Anthony K. H. Tung	0.000909882
17	Danilo Montesi	0.000908795
18	Vladimir Zadorozhny	0.000409823
19	Gao Cong	0.000380872
20	Wei Wang	0.000372509

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 matrixes 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

[1] C. C. Aggarwal, C. Procopiuc, J.L. Wolf, P.S. Yu, J.S. Park. Fast algorithms for projected clustering. In SIGMOD, 1999.

[2] S. Guha, R. Rastogi, K. Shim. CURE: an efficient clustering algorithm for large databases. In SIGMOD, 1998.

[3] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using page rank vectors. In FOCS, 2006.

[4] R.T. Ng, J. Han. Efficient and effective clustering methods for spatial data mining. In VLDB, 1994.

[5] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal, 16(1), 1973, pp. 30-34.

[6] T. Zhang, R. Ramakrishnan, M. Livny. BIRCH: An efficient data clustering method for very large databases. In SIGMOD, 1996.

[7] Y. Qiu and H.P. Frei. Concept Based Query Expansion. In SIGIR, 1993.

[8] R. Navigli, P. Velardi. An Analysis of Ontology-based Query Expansion Strategies. In ATEM, 2003.

[9] M. Brand. A Random Walks Perspective on Maximizing Satisfaction and Prot. In SIAM, 2005.

[10] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In SIGKDD, 2002.

[11] Y. Yin, J. Han and P. S. Yu. LinkClus: Efficient Clustering via Heterogeneous Semantic Links. In VLDB, 2006.

[12] Y. Sun, J. Han and P. Zhao etal. RankClus: Integrating Clustering with Ranking forHeterogeneous Information Network Analysis. In EDBT, 2009.

[13] Y. Sun, Y. Yu, J. Han. Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema. In SIGKDD, 2009.

[14] S. Deerwester, Indexing by Latent Semantic Analysis[J]. Journal of The American Society for Information Science, 1990(41), pp. 391-407.

[15] T. Hofmann. Probabilistic latent semantic analysis. In IUAI, 1999, pp.289- 296.

[16] Y. Zhou, H. Cheng, J. X. Yu. Graph Clustering Based on Structural/Attribute Similarities. In VLDB, 2009.

[17] G. Jeh and J. Widom. Scaling personalized web search. In Stanford University Technical Report, 2002.

[18] D. Aldous and J. A. Fill. Reversible Markov Chains. 2001.

[19] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In ICML, 2003.