Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/7446
DOI URL: http://dx.doi.org/10.21474/IJAR01/7446
ISSN NO. 2320-5407

RESEARCH ARTICLE

SIMILARITY MEASURE FOR SOCIAL NETWORKS.

Roland T H Siagian, Sawaluddin and Opim Salim Sitompul.
Department of Mathematics, University of Sumatera Utara, Medan, Indonesia.

Manuscript Info

Manuscript History

Received: 20 May 2018
Final Accepted: 22 June 2018
Published: July 2018

Keywords:-
Social Networks, DFS, BFS, SimRank, P-Rank, PageSim, Vertex Similarity.

Abstract

Along with the growth in the use of social networks, the measurement of social parameters (e.g., centrality and similarity) becomes more important. Many algorithms have been proposed to measure the graph similarity as a representation of social networks. The purpose of this paper is to analyze the characteristics of similarity algorithms, and compare the performance of each algorithms to determine its strengths and weaknesses.

Copy Right, IJAR, 2018,. All rights reserved.

Introduction:-

Over the past few decades, major revolutions in the mass media field have been marked or more precisely created by new media such as computers, telephone networks, communication networks, the internet and multimedia technologies (Mahmoud and Auter, 2009). Mahmoud and Auter also explained further that online communication is the most important format in computer-mediated communication, that is with the internet as the medium.

The internet is a great many-to-many communication networking the form of emails, news portals, chat rooms, groups, and webpages, which can applied in online journalism, e-commerce and online advertising. Along with this, social users increasing aggressively utilize the internet as a medium of communication.

The use of the internet as a medium has consequences on the need for various changes and adjustments in communication strategies. Rampant social networking, virtual community, viral communication and user-generated content, citing the terms Wertime and Fenwick (2010), become gamechangers that bring change from traditional marketing to digital marketing. Of course, this trend towards digital changes is directly directing strategic planning of current marketing communications. This can be seen clearly from the number of corporate websites, Facebook Fanpage brand, twitter brand accounts, interactive games embedded brand, and many other ways that almost all seek contact and interaction with consumers.

Previously, a social network could be represented as a graph (Diaz and Ralescu, 2012), which is a collection of nodes or profiles. Similarities between nodes can be based on node (textual) and/or side/ link (structure) attributes.

Some similarity sizes consider the same neighbors of the nodes, while others allow nodes to be similar even when nodes do not have the same neighbors. Several similarity sizes consider only the similarity of links of length 2, others specify similarity based on longer paths, while others are defined as the number of paths that vary the path length between them (Leicht et al., 2011).

Corresponding Author:-Roland T H Siagian.
Address:-Department of Mathematics, University of Sumatera Utara, Medan, Indonesia.

Literature review:-

Social Network:-

Social networking is a network that connects a group of people in the context of social relationships. Social relations here can be various interrelations. Examples are family relations, friend relationships, relationships, business relationships, organizational relationships and so on. This social network is very useful in analyzing how the interactions that occur in an environment.

Nodes are individual actors in the network, and ties are the relationship between actors. The resulting graph structure is often very complex. There will be many types of ties between nodes. Research in a number of academic areas has shown that social networks operate on many levels, from family to country level, and play an important role in determining the way in which problems are solved, the organizations being run, and the extent to which individuals succeed in achieving their goals.

In plain language, social networks are graphs with specific relational ties. The relationship between nodes / individuals is called social contact. There are several terms in the measurement of social networks, ie (Ahmad Syuhada, 2010) :

Bridge:-

The node where if omitted will break the relation from the ends of the side of the node.

Centrality:-

The 'rough' measure gives an indication of how well the nodes are in social network relationships with other properties.

Degree Centrality / Degrees of Degrees:-

Number of connections or relationships connected to one actor / node. There is an indegree term for the relation that leads to the node tesebut and outdegree for the relation that leads out the node.

$$C_D(v) = \frac{\deg(v)}{n-1}$$

Betweenness:-

A measure that states how many events pass through a node in a shortest path.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $\sigma_{st}(v)$ is the number of shortest paths from s to t through node v. And $\sigma_{st}(v)$ is the number of shortest paths from s to t. The sum of these calculations is called Betweenness.

Closeness:-

The degree of proximity of a node to another node is averaged by the amount of distance between nodes.

$$\frac{\sum_{t \in V/v} d_G(v,t)}{n-1}$$

Eigenvector:-

This measure gives higher weight to nodes connected to nodes that also have high connectivity.

PageRank:-

This measure is used by Google to determine the quality of a page. Can be used for network of directed graph. The principle used is the more important a node, the more nodes are referenced by other nodes.

Similarity Graph:-

The graph can be represented as follows:

Graph G = (V, E)
V = the set of nodes - can not be empty
= {v_1, v_2, , v_n}
E = set of sides - each connecting 2 nodes
= {e_1, e_2, , e_n}

For example given: 2 graphs are G_1 (n_1, e_1) and G_2 (n_2, e_2) with different number and sides and mapping between the nodes in the graph.

Graf G_1 and G_2 are assumed to have correspondence between the 2 graphs in which each side is weighted or not. Similarity graphs require the degree of similarity between 2 graphs with the scale range of 0 and 1. The rationale of the graph similarity states that the nodes of a graph are similar to other graph nodes if their neighbors are similar. The method is based on counting repeatedly and there is a score-passing between connected nodes.

Matching Graph:-
Matching Graf is a subgraph of a graph where there is no adjacent side to each other. Simply put, there is no same node between the two sides.
Let G = (V, E) be a graph. A subgraph is said to match M (G), if each node G incident is at most one side in M, ie :

deg(V) $\leq 1 \forall V \in$ G

which means that in the M (G) graph matching, the node must have a degree of 1 or 0, where the sides should be incidents of graph G. The number of matching on a graph is called cardinality matching and is denoted by $|M|$.

Discussion:-

SimRank:-
SimRank is an algorithm that measures the similarity of in-links on each pair of nodes in either node a or node b. Given that the rationale behind finding this similarity measure is "2 nodes are said to be similar when the node is referenced by similar nodes as well". It is assumed that the similarity s (a, b) is between the nodes a and b with s (a, b) $\in [0,1]$. If a = b then s (a, b) is assumed to be 1. Thus, SimRank similarity equation is expressed as follows (Jeh and Widom, 2002) :

$$s(a,b) = \frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}s\left(I_i(a), I_j(b)\right) \quad (1)$$

where C is a constant between 0 and 1.
There are several records found in the SimRank algorithm:

1. The most basic method that measures similarity with the same number of in-links between two nodes that are then fixed by many researchers afterwards.
2. Fixed one of them with P-Rank which also consider out-link.
3. SimRank utilizes in-link links only for similarity calculations while paying little attention to the similarity beliefs of out-link directions.

P-Rank:-
P-Rank is an algorithm that measures the similarity of in-link and out-link on each node pair both in node a and node b in the same graph. Basically in finding the P-Rank algorithm can be expressed as "two entities (nodes) in an Information Network (IN) are similar when they relate to similar entities". More specifically, the meaning of P-Rank is doubled as follows :

1. "two nodes are similar when they are referenced by similar nodes"
2. "two nodes are similar when they reference a similar node"

For each pair of different nodes, P-Rank considers both in-link and out-link relationships for similarity calculations. As mentioned earlier, P-Rank can be formulated in equation (2), where a ≠ b.

$$s(a,b) = \lambda x \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s\left(I_i(a), I_j(b)\right) + (1-\lambda) x \frac{C}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} s\left(O_i(a), O_j(b)\right) \quad (2)$$

in addition, P-Rank also defined as :

$$s(A, B) = 1 \quad (3)$$

equation (2) is written for each pair of nodes $a, b \in G$, yielding the n^2 equation for size n graph. To solve the problem of equation n2, the P-Rank equation is rewritten to iteration as follows (Zhao et al., 2011):

$$R_0(a,b) = \begin{cases} 0, & (\text{if a} \neq \text{b}) \\ 1, & (\text{if a} = \text{b}) \end{cases} \quad (4)$$

and

$$R_{k+1}(a,b) = \lambda x \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} R_k\left(I_i(a), I_j(b)\right) + (1-\lambda) x \frac{C}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} R_k\left(O_i(a), O_j(b)\right) \quad (5)$$

Some of the things that are noted in the P-Rank algorithm are :

1. here all users / nodes are considered equal (no PageRank / Introduction).
2. all friendships are considered the same even though both friends of a, can be a and b relationships stronger than a and c.

PageSim:-

Mathematical calculations stated that the PageRank value first sought on the PageSim algorithm. The PageRank is a numerical value that states how important a web page is on the internet. PageSim can be considered an extension of the cocitation algorithm, where the similarity value between 2 web pages is defined by the number of incoming neighbors they have (Lin et al., 2006).

Definition 1. Suppose PR (v) denotes the PageRank value of page v, for $v \in$ V. Suppose PG (u, v) denotes the PageRank value where page u spreads parameter C ϵ to page v through PATH (u, v), ie PG (u , V) = $\sum_{p \in PATH(u,v)} C \frac{PR(u)}{\Pi_{w \epsilon p, w \neq v} |O(w)|}$; u, v V,

where C = adjustable parameter values from 0 to 1.

$$(PG(v_i, v))^T, i = 1, \ldots, n \qquad v, v_i \in V.$$
$$PS(u,v) = \sum_{i=1}^{n} \min\left(PG(v_i, u), PG(v_i, v)\right)$$

DEFINITION 2. Suppose that the vector similarity of page v, then get $\overrightarrow{SV}(v)$ = ; where Suppose PS (u, v) means the PageSim value of page u and v, , where $u, v \in$ V.

Some of the things that are recorded in the PageSim algorithm are:

1. PageRank must be available first.
2. Suitable for social networks that allow one user to follow other users.
3. All friendships are equal.

Vertex Similarity:-

Similarity of vertices is an algorithm that uses adjacency matrix / eigen value and takes into account the difference in path length. Alternatively, this measure can be seen as a weighted value of the number of paths of all path lengths between related nodes. This similarity measure yields the following equation

$$S = 2m\lambda_1 \mathrm{D}^{-1}\left(\mathrm{I} - \frac{\alpha}{\lambda_1}\mathrm{A}\right)^{-1}\mathrm{D}^{-1} \quad (6)$$

where D is a diagonal matrix that has a degree of node in its diagonal element: $D_{ij} = k_i\delta_{ij}$.

As a practical step, the simplest matrix similarity calculations can be achieved by direct multiplication. By eliminating the constant factor, therefore the vertex similarity equation can be rewritten as (Leicht et al., 2011) :

$$\mathbf{DSD} = \frac{\alpha}{\lambda_1}\mathbf{A}(\mathbf{DSD}) + \mathbf{I} \quad (7)$$

Some of the things found in the Vertex Similarity algorithm are :

1. Consider long-term side relationships.
2. Have not considered the node's introduction.
3. Have not considered the quality of the 2 node relationship.

Conclusion:-

In determining the similarity measure for a good social network, it is necessary to consider the parameter factor used in propagating the parameters to each node. In almost every parameter selection algorithm is very important to support the ability of the algorithm, for example the parameters used by the algorithm PageSim in measuring similarity that C. As well as PageRank used by Google, as well as with PageSim that have similarities with PageRank in determining the level of importance of web pages . So like the previous example of PageSim calculations, the greater the value of a Yahoo web page for example page a or page b then the higher the importance of the web. Thus the PageSim algorithm is superior when compared with other algorithms in measuring the size of similarity, so this algorithm is more suitable for social networking.

Where in practice in social networking, one user follows another user.
Compared with:

1. P-Rank: PageSim considers a longer path number.
2. Vertex Similarity: PageSim considers vertex centrality while vertex similarity does not.

References:-

1. Diaz, I., and Ralescu, A.2012. Privacy issues in social networks: a brief survey. In Advances in Computational Intelligence. Springer. 509-518.
2. Jeh, G., and Widom, J. 2002. SimRank: a measure of structural-context similarity. In Proceedings of the eight ACM SIGKDD international conference on Knowledge discovery and data mining, 538-543. ACM.
3. Leicht, E.; Holme, P.; and Newman, M. E. 2011. Vertex similarity in networks. Physical Review E 73(2):026120.
4. Lin, Z.; King, I.; and Lyu, M. R. 2006. Pagesim: A novel link-based similarity measure for the world wide web. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 687-693. IEEE Computer Society.
5. Mahmoud and Auter. 2009. The interactive Nature of Computer-Mediated Communication. American Communication Journal. Vol. 11, No.4.
6. Syuhada, Ahmad. 2013. Teori Graf dalam Social Network Analysis dan Aplikasinya pada Situs Jejaring Sosial. Program Studi Teknik Informatika, Institut Teknologi Bandung.
7. Wertime K. and Fenwick, I. 2010. DigiMarketing: The EssentialGuide to New Media and igital Marketing,

Singapore: John Wiley and Sons (Asia) Pte. Ltd, pp.3-10.
8. Z.Lin, Allen, 2012. PageSim: A Link-based Measure of Web Page Similarity. Research Group Presentation.
9. Zhao, P.; Han J.;and Sun, Y.2011. P-rank: a comprehensive structural similarity measure over information networks. In Proceedings of the 18th ACM conference on Information and knowledge management, 553-562. ACM