Indonesian Journal of Electrical Engineering and Computer Science
Vol. 9, No. 3, March 2018, pp. 812~818
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v9.i3.pp812-818

Search Engine-inspired Ranking Algorithm for Trading Networks

Andri Mirzal
Department of Innovation and Technology Management, Arabian Gulf University, Bahrain

Article Info

Article history:

Received Jan 7, 2018
Revised Feb 16, 2018
Accepted Feb 22, 2018

Keywords:

HITS
Page rank
Ranking algorithms
Search engine
Trading networks

ABSTRACT

Ranking algorithms based on link structure of the network are well-known methods in web search engines to improve the quality of the searches. The most famous ones are PageRank and HITS. PageRank uses probability of random surfers to visit a page as the score of that page, and HITS instead of produces one score, proposes using two scores, authority and hub scores, where the authority scores describe the degree of popularity of pages and hub scores describe the quality of hyperlinks on pages. In this paper, we show the differences between WWW network and trading network, and use these differences to create a ranking algorithm for trading networks. We test our proposed method with international trading data from United Nations. The similarity measures between vectors of proposed algorithm and vector of standard measure give promising results.

Corresponding Author:

Andri Mirzal,
Department of Innovation and Technology Management,
Arabian Gulf University, Bahrain.
Email: andrim@agu.edu.bh

1. INTRODUCTION

With the fast growing of Internet use, trading activities that use Internet and WWW technology as the medium are flourish. Some works have tried to analyze online trading activities to find similarity among users so that recommendation schemes can be built. For example Kawachi et al. [1] use ranking algorithm to characterize users in online auction network and group the users based on their similarity in characteristic vectors. Then by using the most similar users' preferential, the recommendation to the users under consideration can be created. The recommendation schemes are not trivial tasks, for examples Netflix, an online DVD rental, on October 2006 offered one million USD prize to the first developer who can create recommendation algorithm that could beat its existing algorithm, Cinematch, at predicting customer ratings by more than 10%.

In [1], the authors use their modified HITS to group the users based on similarity on their labeled links. Online auction network is a labeled-link network, where the information about the activities cannot be represented only by nodes connected with weighted links, but to which category a link belong to must also be included. The method first calculates authority and hub vectors for each categories (in yahoo japan auction), and then uses these vectors as users' signature. Grouping the most similar users is done by inputting the characteristic vectors into self-organizing map (SOM) [2]. The closer the distance between two users on the map, the more similar they would. However, there is one question here, even though the trading activities are conducted on WWW network, is trading network really similar to WWW network so that a technique from WWW network research like HITS can be used there?

In this work, we try to explore the differences between these two networks and propose a new ranking algorithm based on their differences. To test the algorithm, instead of using online trading data, international trading data from United Nations [3, 4] is used. The reason for choosing the dataset is solid, as

shown in the next section, the trading networks are similar to each other. So by choosing much smaller networks with clear classification of goods and almost fixed price for the same product, the analysis of adjacency matrices of network becomes much easier. Not to mention that the matrices are not sparse, so some manipulation steps to ensure the convergence of the matrices can be avoided (even though the proposed model is written with assumption that the matrix is sparse). For the sake of algorithm testing it provides us with the best condition.

Actually there are also some issues in the online trading networks that prevent it from being good test datasets. For example in an auction network, which is the best example of trading networks where the users are free to buy and sell goods (so each users can have both inlinks and outlinks), the number of sold goods are too diverse in both type and price to allow any classification works. Consequently, it is difficult to infer that goods in the same class are more similar than goods from other class. And if it is the case, there is no point to use this classification as the base for users clustering. And if classification cannot be used, adjacency matrix for each kind of goods must be constructed, which is a really difficult task because there will be too many sparse matrices for one network with only one, or two nonzero entries (number of identical goods bought by a user). And for other type of trading networks like online shopping where there are two kind of nodes, buyers and sellers, the networks become bipartite graphs so ranking and clustering tasks become different problem, which is beyond the scope of this paper.

2. TRADING NETWORKS

The usual way to calculate the degree of importance of nodes in a trading network is by using total amount of export/import of particular goods. This method, however, fails to capture the link structure of the network; to which nodes a node connect to and being connected to. For example the same amount of export to an insignificant country and to an important country will give the same weight to ranking scores. This problem actually had ever occurred in WWW network, where the methods of only calculating content scores of web pages were no longer adequate to deal with users' satisfaction and accuracy of the queries response in the fast growing WWW network environment. The solutions of this problem were proposed independently by Brin and Page [5, 6] and Kleinberg [7]. Both solutions use link structure of WWW network to improve the quality of web search.

In PageRank, the important pages are the pages with many inlinks and a few or no outlinks [8, pp. 32]. And HITS, instead of producing only one score, proposes to use two scores; authority and hub scores. Good authorities are pointed to by good hubs and good hubs point to good authorities [8, pp. 115]. The final link structure scores are obtained by combining these scores (in web search purpose, usually only authority scores are used).

Even though there are already good ranking algorithms that deal with link structure of the networks, PageRank or HITS cannot simply be used because the nature of trading networks and WWW network is different. Each nodes in trading networks has at least one type of resource before any transaction can occur. The links addition happens when two nodes with different type of resources exchange their resources. Thus, the amount of resources limits number and weight of links that a node can have. In WWW network, links addition is simply by putting new hyperlinks on web pages, so there is no resource needs to be allocated in creating new links. Another important point that differentiates these networks is links addition in trading networks is mutual process, if the first node creates a new link to the second node, the second node also creates a new link to the first node. This is not the case in WWW network. Further, links attachment purpose in trading networks is to maximize the benefit of the transactions. Thus, in the export side, each nodes competes to get transactions from other nodes that lack of the resource it offers, and in import side, it competes to get resources from other nodes that have abundant resource it needs. In WWW network, the links attachment is to get inlinks from popular pages (pages with many inlinks) and the popular pages will likely to get more inlinks. Figure 1 shows the differences between trading network and WWW network where in trading network the process of links addition is mutual, and the links are different in type and weight, which describes the nature of transaction. In WWW network the links that connect page A and B are hyperlinks, which when A has a hyperlink to B, it doesn't necessary that B has a hyperlink to A also.

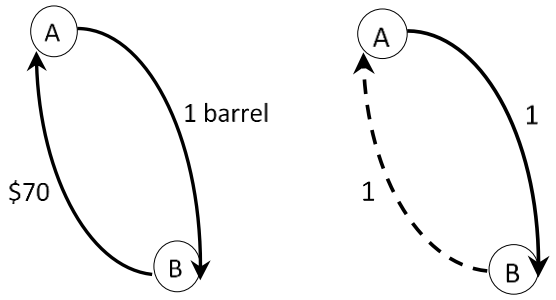

Figure 1. The Differences Between Trading Network (Left) and WWW Network (Right)

3. PROPOSED ALGORITHM

In trading networks, every nodes should be careful in making new inlinks and outlinks due to the needed resources. Each nodes competes to get inlinks from other important nodes (nodes with abundant number of resources that competing nodes need) and competes to make outlinks to the other important ones (nodes that need resources from competing nodes) by considering the cost.

Due to the nature of trading networks, none of the previous discussed web ranking algorithms are suitable. PageRank which focuses on inlinks clearly cannot be used in the environment where inlinks as well as outlinks are highly regarded. HITS is more interesting than PageRank, because it accommodates both inlinks and outlinks. But by definition, in HITS's a node should establish new outlinks to others with many inlinks, and should receive inlinks from others with many outlinks. In the context of our problem, where making and receiving new links can be expensive, it is more appropriate to make new outlinks to nodes that have many outlinks and receiving inlinks from nodes that has many inlinks, because receiving inlinks means getting resources and creating outlinks means giving up resources. Figure 2 shows the links addition process where in trading network, A prefers B (node with many outlinks therefore lack of resource) when making a new outlink and C (node with many inlinks therefore full of resource) when looking for a new inlink. This preferential is opposite to WWW network.

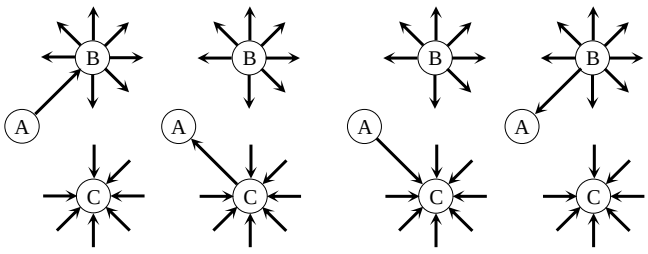

Figure 2. Links Addition Process in Trading Network (Left) and WWW Network (Right).

Proposed algorithm is defined with the following statement: *a node becomes more important if being pointed to by others that have many inlinks and points to others with many outlinks*. And further, by comparing the process of links addition as shown in fig. 2 and HITS model [8, pp. 115], this definition can be written into following equation.

$$r(n_i) = \beta \sum_{j \to i} \left(r(n_j) \times ca \right) + (1-\beta) \sum_{i \to j} \left(r(n_j) \times ch \right) \quad (1)$$

Where

$$ca = \frac{\sum n_j\ inlinks}{\sum n_j\ links} \times \left| \sum n_j\ inlinks - \sum n_j\ outlinks \right|^{p_j},$$

$$ch = \frac{\sum n_j\ outlinks}{\sum n_j\ links} \times \left| \sum n_j\ inlinks - \sum n_j\ outlinks \right|^{-p_j},$$

and

$$p_j = \begin{cases} 1 & \text{if } \sum n_j\ inlinks > \sum n_j\ outlinks \\ -1 & \text{if } \sum n_j\ inlinks < \sum n_j\ outlinks \\ 0 & \text{if } \sum n_j\ inlinks = \sum n_j\ outlinks \end{cases}$$

$r(n_i)$ is the ranking score of node i, $|*|$ denotes absolute value of $*$, $i \rightarrow j$ denotes that node i links to node j, and $\sum n_j$ inlinks / outlinks / links denotes the number of inlinks / outlinks / links node j has. The first term of right hand part is defined as authority part and the second one as hub part of corresponding node. Parameter β ($0 \leq \beta \leq 1$) is used to determine which links are more important. If outllinks and inlinks are equal set $\beta = 0.5$, if outlinks are more important set $\beta < 0.5$, and $\beta > 0.5$ otherwise.

The logic behind above equation is: ranking score of a node, $r(n_i)$, depends on the ranking scores of others that point to it ($r(n_j)$ where $j \rightarrow i$, the first term of the right hand part) and the nodes that it points to ($r(n_j)$ where $i \rightarrow j$, the second term of the right hand part). The rests of the right hand part function as the constants that depend on the number of inlinks and outlinks of each nodes, where for authority / hub part the bigger the number of inlinks / outlinks and the smaller the number of outlinks / inlinks, the larger the constants become. So, the above equation agrees with proposed algorithm definition.

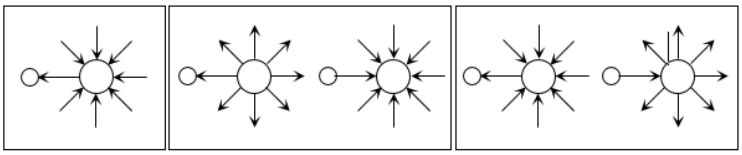

Figure 3. A Schematic Explanation of the Differences Among Algorithms; Pagerank (Left), HITS (Center) and Proposed Algorithm (Right).

The calculation of nodes' ranking scores can be done in two different ways, the first is by using direct method by inspecting the linear system property of the equation [8 pp. 71-74] and the second is by using iteration process (power method), a common method in calculating ranking vector for web pages. For small network the first method is preferable because it much faster than power method. As the network getting bigger, only second method is viable. In this paper, however, second method is used because we want to compare convergence property of proposed algorithm to PageRank and HITS.

We will modify eq. (1) into matrix form not only to allow property of network being seen from linear algebra perspective but also to ensure power method applied to the adjacency matrix converges by adjusting it into a stochastic and primitive matrix. Let $\mathbf{M} = \beta\mathbf{F} + (1-\beta)\mathbf{G}$, where $\mathbf{F} = \mathbf{K}\mathbf{D}^{-1}\mathbf{D_i}\mathbf{L}$ is the authority part which describes fraction of scores a node receives from its inlinks, and $\mathbf{G} = \mathbf{K}^{-1}\mathbf{D}^{-1}\mathbf{D_o}\mathbf{L}^T$ is the hub part which describes fraction of scores a node receives from its outlinks. And $\mathbf{L}$ is $N \times N$ the adjacency matrix of the network. Thus, eq. (1) can be rewritten as:

$$\begin{aligned} \mathbf{r}^T(k+1) &= \mathbf{r}^T(k)\left(\beta\mathbf{F} + (1-\beta)\mathbf{G}\right) \\ &= \mathbf{r}^T(k)\left(\beta\mathbf{K}\mathbf{D}^{-1}\mathbf{D_i}\mathbf{L} + (1-\beta)\mathbf{K}^{-1}\mathbf{D}^{-1}\mathbf{D_o}\mathbf{L}^T\right) \end{aligned} \quad (2)$$

where $k = 0, 1, 2, \ldots$ denotes the iteration process of the algorithm, diagonal matrices $\mathbf{D_i}$, $\mathbf{D_o}$ and $\mathbf{D}$ are defined as:

$$\mathbf{D_i} = \text{diag}(\mathbf{d_i}),\ \mathbf{D_o} = \text{diag}(\mathbf{d_o}),\ \text{and}\ \mathbf{D} = \mathbf{D_i} + \mathbf{D_o} \quad (3)$$

and $\mathbf{K}$ is a diagonal matrix with diagonal entries defined as:

$$K_{ii} = \left|\left(\mathbf{D_i} - \mathbf{D_o}\right)_{ii}\right|^{p_i} \quad (4)$$

given $i_i = \sum_j L_{ji}$ is the set of inlinks of node i, $o_i = \sum_k L_{ik}$ is the set of outlinks of node i, $\mathbf{d_i} = (i_1, i_2, \ldots, i_N)^T$ is inlink vector, and $\mathbf{d_o} = (o_1, o_2, \ldots, o_N)^T$ is outlink vector of node i.

To ensure the power method [9] converges to a positive and unique dominant eigenvector of matrix **M**, two adjustments are needed. The first is *stochasticity adjustment*; normalizes all nonzero rows of **M** and then fills zero rows by $1\times N$ positive real vectors which have 1-norm equals to one. Usually, each entry of these vectors is set to $1/N$. Let $\mathbf{e}^T$ is a $1\times N$ row vector which each of its entries is one and **c** is a $N\times 1$ column vector which its i^{th} row is set to 1 if row i of **M** is zero row, and 0 otherwise. Then stochastic version of matrix **M** is: $\mathbf{S} = \mathbf{M} + (1/N)\mathbf{ce}^T$. And the second, *primitivity adjustment* is done by replacing each zero entries of **S** with a small positive number; $\mathbf{P} = \alpha\mathbf{S} + (1/N)(1-\alpha)\mathbf{ee}^T$, where $0 < \alpha < 1$ is a parameter that control the amount of error ($\mathbf{ee}^T$) introduced to matrix **P**. Thus, eq. (1) can be written in more compact form as:

$$\mathbf{r}^T(k+1) = \mathbf{r}^T(k)\mathbf{P} \quad (5)$$

for initial condition $\mathbf{r}^T(0) = (1/n)\mathbf{e}^T$, until error of the process $\|\mathbf{r}^T(k+1) - \mathbf{r}^T(k)\|_1$ is smaller than desired error. Note that instead of using 1-norm termination criterion, the comparison between previous rank and current rank order can also be used to terminate the iteration process [10-15]

4. NUMERICAL RESULTS

Because **P** is stochastic and primitive, the power method applied to it converges to a unique positive vector called stationary vector for any starting vector [88, pp. 36]. So the problem left is "*will it converge to something that makes sense in the context of trading networks?*". We try to answer this question by measuring the similarity between vector of proposed algorithm with standard measure, vector of total export import.

The data used in the experiments is international trading data from United Nations [3, 4] where the nodes are the countries that involved in the export and import activities, and the links are the flow of the products. The computation performance of the proposed algorithm is measured by comparing the number of iterations it needs to achieve a desired error to the results of HITS and PageRank (note that it is only used for performance comparison, not for results comparison). In the experiments termination criterion is set to 10^{-8} and β is set to 0.5. The number of iterations is chosen instead of computational time because the size of trading networks is very small, so power method applied to the data produces negligible computational time. Then similarity measures, (1) cosine of the angle between ranking vector of proposed algorithm (**u**) and vector of total export import (**v**),

$$\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \quad (6)$$

and (2) Spearman rank order correlation coefficient,

$$\rho = 1 - \frac{6\sum_{i=1}^{N}\left(r(u_i) - r(v_i)\right)^2}{N\left(N^2-1\right)} \quad (7)$$

are used to measure the ranking quality, where $r(\mathbf{u})$ is ranking order of vector **u**. For example if **u** = [0.3397 0.1819 0.3328] then $r(\mathbf{u})$ = [1 3 2]. Table 1 gives summary of the results.

Table 1. The Performance of the Proposed Algorithm

Data	#Nodes	#Nonzero	Number of iteration			cos θ	ρ
			HITS	PageRank	Proposed algorithm		
Steel Products	97	2627	26	54	42	0.862	0.874
Ethylene	43	169	7	44	54	0.849	0.916
Propylene	38	144	10	40	143	0.974	0.905
Sodium	49	268	11	53	143	0.808	0.850
Hydrogen Peroxide	47	261	51	61	99	0.752	0.902
Carbon	51	535	22	37	65	0.912	0.929
Radio-active	53	717	25	23	26	0.884	0.927
Plastics	53	1410	20	37	39	0.985	0.968
Medicinal Products	53	1504	9	18	14	0.989	0.965
Average	54	848	20	41	69	0.891	0.915

As shown in table 1 the proposed algorithm takes more iterations to converge, but because the trading networks are usually much smaller than WWW network, this will not become a problem. And the similarity measures in both criterions give promising results, 89.1% and 91.5% in average respectively.

Instead of relying on only link structure ranking, we propose to combine both scores to get overall scores. This is because we believe that our algorithm is not to replace the conventional method, but to refine it. Table 2 and 3 show the ranking scores of two data, hydrogen peroxide (similarity measure 0.752, the least similar to standard measure in cosine criterion) and medicinal products (similarity measure 0.989, the most similar to standard measure in cosine criterion) for top ten countries.

Table 2. The Top Ten Countries in Hydrogen Peroxide Trading

Ranking by total export and import (normalized)		Ranking by link structure		Overall ranking	
Country	Score	Country	Score	Country	Score
Netherlands	0.132290	Japan	0.172970	Netherlands	0.123250
Canada	0.095014	Norway	0.123360	Japan	0.110820
US	0.088694	Netherlands	0.114200	Canada	0.088638
Moldova	0.065088	Canada	0.082261	Norway	0.071148
Austria	0.059850	Turkey	0.053170	US	0.067877
China	0.054194	US	0.047059	Moldova	0.051716
Japan	0.048676	Rep. Korea	0.043684	China	0.045555
Italy	0.045744	Moldova	0.038344	Turkey	0.045261
Colombia	0.037772	China	0.036916	Austria	0.043115
Turkey	0.037353	Thailand	0.034545	Italy	0.033318

Table 3. The Top Ten Countries in Medicinal Products Trading

Ranking by total export and import (normalized)		Ranking by link structure		Overall ranking	
Country	Score	Country	Score	Country	Score
Germany	0.133530	Germany	0.139490	Germany	0.136510
US	0.114520	UK	0.107270	US	0.106520
UK	0.096001	US	0.098509	UK	0.101640
France	0.092408	Switzerland	0.095938	Switzerland	0.089591
Switzerland	0.083244	France	0.085463	France	0.088935
Italy	0.067707	Italy	0.064711	Italy	0.066209
Belg-Luxemb.	0.056696	Belg-Luxemb.	0.051169	Belg-Luxemb.	0.053932
Netherlands	0.051564	Netherlands	0.047270	Netherlands	0.049417
Japan	0.049308	Ireland	0.043663	Japan	0.039471
Sweden	0.033573	Sweden	0.041134	Sweden	0.037353

5. CONCLUSION

Due to the different nature of trading networks and WWW network, the link structure ranking algorithm for WWW network that previously applied in (Kawachi et al., 2006) to analyze trading network cannot be used. Consequently special ranking algorithm must be developed to deal with the different link addition process. By considering the needed cost in link addition process, we propose new ranking algorithm for class of networks that requires resources to be exchanged in link addition process. The proposed algorithm not only considers the total volumes of export/import, as usually used to rank the most important countries, but also takes into account the network structure in the trading networks.

REFERENCES

[1]. Kawachi, Y., Yoshii, S. and Furukawa, M. *Labeled link analysis for extracting user characteristics in e-commerce activity network*. IEEE/WIC/ACM International Conference on Web Intelligence. 2006. 73-80.

[2]. Kohonen, T. The self-organizing map. *Neurocomputing* 21(1). 1998, 1-6.

[3]. United Nations. Annual Bulletin of Statistics of World Trade in Steel 1998. Destination of Export of Steel Products (Total). United Nation, 1999.

[4]. United Nations. Annual Bulletin of Trade in Chemical Products 1996. United Nations, 1996.

[5]. Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 30(1). 1998, 107-117.

[6]. Page, L., Brin, S., Motwani, R. and Winograd, T. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-0120, Computer Science Department, Stanford University. 1999.

[7]. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5). 1999, 604-632.

[8]. Langville, A. N. and Meyer, C. D. Google's PageRank and beyond: The science of search engine rankings. Princeton University Press, 2011.

[9]. Meyer, C. D. Matrix analysis and applied linear algebra. SIAM, Philadelphia, 2000, pp. 534.

[10]. Dwork, C., Kumar, R., Naor, M. and Sivakumar, D. *Rank aggregation methods for the web*. Proceedings of the 10th international conference on World Wide Web, 2001, 613-622.

[11]. Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A. and Williamson, D. P. *Searching the workplace web*. Proceedings of the 12th international conference on World Wide Web, 2003, 366-375.

[12]. Fagin, R., Kumar, R. and Sivakumar, D. Comparing top k lists. *SIAM Journal on Discrete Mathematics* 17(1), 2003, 134-160.

[13]. Haveliwala, T. Efficient computation of PageRank. Technical Report 1999-31, Computer Science Department, Stanford University, 1999.

[14]. Haveliwala, T. H. *Topic-sensitive pagerank*. Proceedings of the 11th international conference on World Wide Web, 2002, 517-526.

[15]. Mendelzon, A. O. and Rafiei, D. *An autonomous page ranking method for metasearch engines*. The Eleventh International WWW Conference, New York, 2002.