HITS based network algorithm for evaluating the professional skills of wine tasters

András London[1,*] and Tibor Csendes[1]

***Abstract*—Two popular and widely used webpage ranking algorithms are PageRank and HITS. We considered the 2009 Szeged Wine Fest data and another reliable data set of wines from the famous region Villány, and, on basis of each data set, constructed a directed and weighted bipartite graph of wine tasters and wines. We applied an extended version of PageRank and HITS, the Co-HITS algorithm to wine tasting graph in order to rank tasters according to their ability and professional skill. The results of our technique were compared to other simple statistical methods. In general we observed that our ranking method performed better: it can filter out incompetent tasters, who, for example, gave the average score of some other tasters for the wines she or he tasted. Furthermore, our method gives a clearer picture about the competence of wine tasters.**

***Index Terms*—HITS, Co-HITS PageRank, wine tasting**

I. Introduction

PageRank is the algorithm used by the Google search engine, originally formulated by Sergey Brin and Larry Page [1]. It was designed to determine the importance of web pages and it was used to rank the pages found for search results. Independent of Brin and Page, Jon Kleinberg proposed an advanced concept for the same purpose [2]. While PageRank computes the ranks of the pages on the complete web graph, Kleinberg's HITS (Hypertext induced Topic Selection) makes the distinction between hubs and authorities, in other words between pages that links to many authorities (i.e hubs) and pages that many hubs link to (i.e. authorities). Both algorithms have been extended and improved in a number of ways [3]–[7].

Bipartite graphs have been widely used to represent the relationship between two sets of entities. Deng et al. proposed the Co-HITS algorithm [8] to incorporate the bipartite graph with the content information from both sides and determine the relevance of the nodes of the graph. We use Co-HITS algorithm to improve the results of earlier studies of Csendes and Antal [9] to determine the quality and expertise of wine tasters by using the wine competition data of the 2009 Szeged (Hungary) Wine Fest and other reliable wine tasting data for wines from the well-known wine region of Villány (Hungary).

This article is organized as follows: in Section II we discuss the mathematical background of the notions and algorithms mentioned above and their relationships with each other; in Section III we apply the algorithms to wine tasting and compare the results with different basic statistical approaches. Finally, in Section IV we discuss our results and highlight the possible applicability of the algorithms in different fields where ranking of the participants can be interpreted.

[1] University of Szeged, Institute of Informatics, 6701 Szeged, P. O. Box 652., Hungary

[*] Corresponding author: András London, University of Szeged, Institute of Informatics, 6701 Szeged, P. O. Box 652. Hungary; Tel: +36-62-543-444; E-mail: london@inf.u-szeged.hu

II. Methods

Consider a bipartite graph $G = (X \cup Y, E)$ whose vertices can be divided into two disjoint sets X and Y such that each edge in E connects a vertex in X to one in Y. Equivalently, there is no edge between two vertices in the same set; that is X and Y are each independent sets. Let $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be the two sets of m and n independent vertices. Consider that G is a weighted directed graph. Given $x_i \in X$ and $y_j \in Y$, if there is an edge between x_i and y_j let $w(\overrightarrow{x_i y_j}) > 0$, $w(\overleftarrow{x_i y_j}) > 0$ denote the weights of the directed edges (x_i, y_j) and (y_j, x_i), respectively, otherwise let $w(\overrightarrow{x_i y_j}) = w(\overleftarrow{x_i y_j}) = 0$. We assume, that each weight is normalized such that $\sum_{j \in Y} w(\overrightarrow{x_i y_j}) = 1$ and $\sum_{i \in X} w(\overleftarrow{x_i y_j}) = 1$ (this can be assumed without loss of generality, e.g let $w(\overrightarrow{x_i y_j}) = w'(\overrightarrow{x_i y_j}) / \sum_{j \in Y} w'(\overrightarrow{x_i y_j})$, where w' was the original weight of the edge without normalization). The weight $w(.)$ can be considered as the transition probability from a vertex in X (Y) to a vertex in Y (X). On this bipartite graph natural random walk can be defined, where $\overrightarrow{W} = W(\overrightarrow{XY}) = (w(\overrightarrow{x_i y_j}))_{ij} \in \mathbb{R}^{m \times n}$ denotes the transition matrix from X to Y and $\overleftarrow{W} = W(\overleftarrow{XY}) = (w(\overleftarrow{x_i y_j}))_{ij} \in \mathbb{R}^{n \times m}$ denotes the transition matrix from Y to X. For the vertices in one side, a hidden transition probability $w(\overrightarrow{x_i x_j})$ from x_i to x_j can be defined as

$$w(\overrightarrow{x_i x_j}) = \sum_{k \in Y} w(\overrightarrow{x_i y_k}) w(\overleftarrow{x_j y_k}), \tag{1}$$

and it is also obtained that $\sum_{j \in X} w(\overrightarrow{x_i x_j}) = 1$.
Note, that $W_X = W(\overrightarrow{XX}) = \overrightarrow{W}\overleftarrow{W} = (w(\overrightarrow{x_i x_j}))_{ij} \in \mathbb{R}^{m \times m}$ is the hidden transition probability matrix within X; W_Y within Y can be obtained similarly.

The algorithm we use works on such a directed weighted bipartite graph defined above. The idea of the algorithm is to assign scores for the vertices of the graph via an iterative process. Let p_i^0 and q_k^0 are the initial scores of the vertices x_i and y_k, respectively. The following equations describe the generalized Co-HITS algorithm:

$$p_i = (1 - \lambda_x) p_i^0 + \lambda_x \sum_{k \in Y} w(\overleftarrow{x_i y_k}) q_k, \tag{2}$$

and

$$q_k = (1-\lambda_y)q_k^0 + \lambda_y \sum_{j\in X} w(\overrightarrow{x_j y_k})p_j, \tag{3}$$

where $\lambda_x \in [0,1]$ and $\lambda_y \in [0,1]$. By substituting Eq. 3 for q_k in Eq. 2 it is obtained that

$$\begin{aligned} p_i &= (1-\lambda_x)p_i^0 + \lambda_x(1-\lambda_y)\sum_{k\in Y} w(\overleftarrow{x_i y_k})q_k^0 + \\ &+\lambda_x\lambda_y \sum_{j\in X} w(\overrightarrow{x_j x_i})p_j. \end{aligned} \tag{4}$$

It can be easily calculated, that HITS and personalized PageRank [10] are special cases of the Co-HITS algorithm. If $\lambda_x = \lambda_y = 1$, then Eq. 4 becomes

$$p_i = \sum_{j\in X} w(\overrightarrow{x_j x_i})p_j \tag{5}$$

which is the original HITS equation. It is worth noting here, that this is the stationary state of the Markov chain defined by a random walk on the weighted graph defined above [11]. If $\lambda_y = 1$, then

$$p_i = (1-\lambda_x)p_i^0 + \lambda_x \sum_{j\in X} w(\overrightarrow{x_j x_i})p_j, \tag{6}$$

which is the personalized PageRank algorithm.

III. Applications to wine tasting

Usually, wine tasting is a personal, subjective procedure to specify the quality of wines. Different wines are scored in an anonymous way called blind tasting (i.e. the tasters do not know which wine is being tasted). Each taster scores the wines she or he tasted and the wines would be ranked by using these obtained points. However, there are several accepted ranking methods for evaluating the quality of the wines, and it is still open, how to determine the quality of the tasters, which is also mentioned in the article of Csendes and Antal.

Before we apply the co-HITS algorithm to provide a ranking of the tasters according to their quality, it is needed to make the following assumptions:

- in the first step, the wines are sorted by the points they received (i.e. there is no reference value for them)
- tasters will be sorted by only considering the points that the wines received from the tasters
- there is no cheater among the tasters (i.e. they score more or less on the "same scale")

Now, we describe how the Co-HITS algorithm can be used for wine tasting. Let X and Y (defined previously) be the set of wine tasters and wines, respectively. We start from the same p^0 value for each $x_i \in X$ taster. Let $w'(\overrightarrow{x_i y_j})$ be the score that wine y_j obtained from taster x_i and let $w(\overrightarrow{x_i y_j}) = w'(\overrightarrow{x_i y_j}) / \sum_{j\in Y} w'(\overrightarrow{x_i y_j})$ be its normalization. Consistent to our first assumption, we define the q_j^0 value (for wine y_j) as the average of the points that the wine received. Then, we define the weight $w(\overleftarrow{x_i y_j})$ in the following way: suppose that wine y_j was tasted by ℓ different tasters and let

$$D = \sum_{i\in X} \left| q_j^0 - w'(\overrightarrow{x_i y_j}) \right|, \tag{7}$$

be the sum of differences from the average score received by wine y_j. Finally, let

$$w(\overleftarrow{x_i y_j}) = \frac{\left| D - \left| q_j^0 - w'(\overrightarrow{x_i y_j}) \right| \right|}{(\ell-1)D}. \tag{8}$$

Note, that $\sum_{i\in X} w(\overleftarrow{x_i y_j}) = 1$, i.e. each weight $w(\overleftarrow{x_i y_j})$ can be regarded as a transition probability from y_j to x_i. Figure 1 shows a concrete example for the calculation of the weights.

The weight between two tasters x_i and x_j can be defined as the hidden transition probability defined by Eq. 1. Then, the solution $p = (p_1, p_2, \ldots, p_m)$ of the HITS equation $p = W_X p$ provides the result for ranking the tasters.

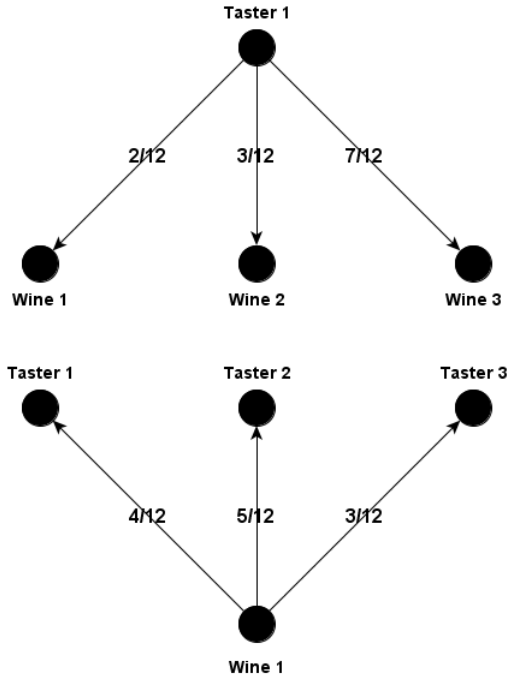

Fig. 1. Weights of the graph when taster 1 gives the scores 20, 30 and 70 to wine1, wine2, and wine3, respectively (left) and when wine 1 received the scores 20, 30, and 70 from taster 1, taster 2 and taster 3, respectively (right).

We tested our model in the selected data of two wine tasting events. The first event was the Szeged Wine Fest 2009, where 104 wines were blind tasted by four groups of five tasters, and each group tasted 33-34 different wines. The second data set is a bit more specific: just red wines from the wine region Villány were blind tasted by seven groups of six-person tasters. Each group tasted 40-48 different wines. In both events, each wine was scored in pursuance of the widely used and accepted international 100 point rating system. Table I and Table II show the detailed results obtained by Co-HITS algorithm on the Szeged Wine Fest data and wine tasting data of wines from region Villány, respectively. The calculated values can be interpreted as normalized merit values: the larger the better. The best taster of each group were highlighted.

TABLE I
THE RESULTS OBTAINED BY CO-HITS ALGORITHM ON THE 2009 SZEGED WINE FEST DATA

Taster	Team 1	2	3	4
1	**1.000**	**1.000**	1.000	0.987
2	0.963	0.824	1.000	**1.000**
3	0.960	0.917	1.000	0.999
4	0.938	0.925	1.000	0.992
5	0.948	0.977	1.000	0.992

TABLE II
THE RESULTS OBTAINED BY CO-HITS ALGORITHM ON THE WINE TASTING DATA OF THE REGION OF VILLÁNY

Taster	Team 1	2	3	4	5	6	7
1	0.843	0.908	0.890	**1.000**	**1.000**	0.969	0.958
2	0.970	0.941	0.980	0.985	0.994	0.984	0.961
3	0.894	0.986	0.941	0.967	0.955	**1.000**	0.933
4	0.957	**1.000**	0.977	0.946	0.944	0.982	**1.000**
5	0.966	0.899	**1.000**	0.978	0.938	0.944	0.932
6	**1.000**	0.901	0.870	0.950	0.966	0.945	0.944

We compared the results obtained by the Co-HITS algorithm with some simple statistical methods which seems natural to use for our purpose. The first statistics based ranking method (SM1) was to evaluate the sum of differences from the average point of each wine received for each taster, as S_i for taster i. Then, the increasing order of these S_i values gives the ranking of the tasters. Formally,

$$S_i = \sum_{j \in Y} |q_j^0 - w'(\overrightarrow{x_i y_j})|. \tag{9}$$

We consider the normalized points $\min_{i \in X} S_i / S_i$ for all i (thus, the score of the taster with minimal S_i value will be 1.000). Table III and Table IV show the results obtained by this method on the Szeged Wine Fest data, and data of wines from Villány, respectively.

The second statistical method (SM2) we used was the Pearson correlation coefficient between the scores that a taster gave for the wines and the average scores that those wine received. In other words, we were interested in how the scores of a taster correlate with the average scores of the wines reveived. The calculated values are normalized. Table V and VI show the results obtained by this method on the Szeged Wine Fest data, and data of wines from Villány, respectively.

For better illustration, Figure 2 shows the summarized results on the Szeged Wine Fest data. In the figures, axis-x shows the tasters, axix-y shows the ranks of the tasters such as longer bar refers to better rank. For each taster, the three different colored bars from the left to the right refers to the methods used for calculation, Co-HITS, SM1, and SM2, respectively.

The results show, that the Co-HITS algorithm is more sophisticated than SM1 and SM2. The stochastic process based HITS eventuated closer values between the tasters. Coherently with this fact, much larger differences that the statistical methods produced cannot be justified in the knowing of the

TABLE III
THE RESULTS OBTAINED BY SM1 METHOD ON THE 2009 SZEGED WINE FEST DATA

Taster	Team 1	2	3	4
1	**1.000**	**1.000**	**1.000**	0.709
2	0.870	0.489	0.470	0.856
3	0.753	0.677	0.496	0.713
4	0.743	0.687	0.475	0.735
5	0.743	0.940	0.510	**1.000**

TABLE IV
THE RESULTS OBTAINED BY SM1 METHOD ON THE WINE TASTING DATA OF THE REGION OF VILLÁNY

Taster	Team 1	2	3	4	5	6	7
1	0.491	0.556	0.495	0.991	0.901	0.891	0.746
2	0.779	0.672	0.932	**1.000**	**1.000**	0.932	0.760
3	0.478	0.794	0.625	0.872	0.839	**1.000**	0.870
4	0.638	**1.000**	0.892	0.665	0.644	0.919	**1.000**
5	0.781	0.613	**1.000**	0.781	0.651	0.822	0.705
6	**1.000**	0.492	0.505	0.856	0.829	0.739	0.678

TABLE V
THE RESULTS OBTAINED BY SM2 ON THE 2009 SZEGED WINE FEST DATA

Taster	Team 1	2	3	4
1	**1.000**	**1.000**	**1.000**	**1.000**
2	0.932	0.932	0.932	0.932
3	0.738	0.738	0.738	0.738
4	0.917	0.917	0.917	0.917
5	0.988	0.988	0.988	0.988

TABLE VI
THE RESULTS OBTAINED BY SM2 ON THE WINE TASTING DATA OF THE REGION OF VILLÁNY

Taster	Team 1	2	3	4	5	6	7
1	0.939	0.954	0.965	0.982	**1.000**	0.983	0.898
2	0.971	0.933	0.943	0.998	0.999	**1.000**	**1.000**
3	0.829	0.961	0.804	**1.000**	0.997	0.947	0.963
4	0.949	**1.000**	0.966	0.913	0.934	0.952	0.981
5	0.958	0.951	**1.000**	0.921	0.964	0.935	0.928
6	**1.000**	0.917	0.850	0.999	0.997	0.923	0.962

concrete data sets. It should be mentioned, that all the three methods produced the same results for the best taster in many cases and the differences appear in the rest of the ranking lists. It can be observed, that SM1 prefers the "closeness to the average" (due to its definition) and SM2 is better if the scores co-movement with the average. It follows from these that both statistical methods can offer an opportunity to cheat.

The described network based algorithm considers the wine tasting data not only as a database that contains the scores of individual tasters, but as a complex network shows each taster relationship to one another. The relation between the tasters can be defined well in respect of the purpose of investigation. Thus, the Co-HITS algorithm which works on networks can give a better picture about the quality of tasters.

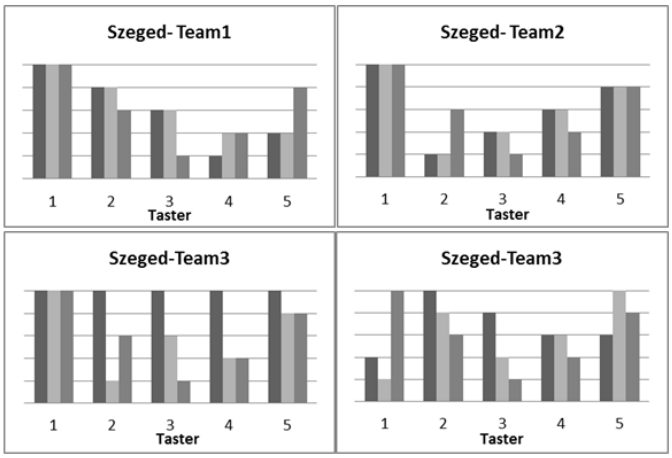

Fig. 2. Ranks of the tasters in different evaluations: dark grey, light grey and middle grey bar refers to Co-HITS, SM1 and SM2, respectively. The longest bar means the first rank while the shortest bar means the last (fifth) rank.

IV. Conclusions

In this paper we investigated how a HITS based algorithm can be used for determining the quality of wine tasters. However, there are several methods for the evaluation the quality of wines by using the points that a wine received in a wine tasting event, but it is still an open question, how to evaluate the professional skills of the tasters. We have applied the generalized Co-HITS algorithm for the data sets of two wine tasting events and compared the results with two simple statistical methods. Experimental results show that Co-HITS algorithm produced promising results, furthermore it is more sophisticated than the statistical methods: both of them produced unreasonably large differences between the tasters and ranked those tasters too high who (maybe due to the incompetence) gave the average of the points of other tasters for the wines. In future work, we plan to refine the HITS algorithm for application to wine tasting data: it would be interesting to use other modifications of the HITS, different rules for the weights of the network. With the analysis of appropriate null models and artificially generated data sets we can highlight further advantages and incidental drawbacks of the application of the ranking algorithms.

However, we used this HITS based algorithm only for evaluating the quality of wine tasters, it could be widely applicable in fields where people evaluate someone or something, such as sports like figure skating, diving, ski jumping, synchronized swimming; social events like beauty pageant, song contest and other tasting events like cooking competition or beer tasting, etc.

Acknowledgment

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013). The authors are grateful to András Pluhár and Márk Jelasity for their useful advises and to Melinda Braun for typing the wine tasting data.

References

[1] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.

[2] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[3] K. Bharat and M. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1998, pp. 104–111.

[4] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas, "Finding authorities and hubs from link structures on the world wide web," in *Proceedings of the 10th International Conference on World Wide Web*. ACM, 2001, pp. 415–429.

[5] D. Cohn and H. Chang, "Learning to probabilistically identify authoritative documents," in *Machine Learning International Workshop and Conference*, 2000, pp. 167–174.

[6] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (salsa) and the tkc effect," *Computer Networks*, vol. 33, no. 1, pp. 387–401, 2000.

[7] A. Ng, A. Zheng, and M. Jordan, "Stable algorithms for link analysis," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2001, pp. 258–266.

[8] H. Deng, M. Lyu, and I. King, "A generalized co-hits algorithm and its application to bipartite graphs," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2009, pp. 239–248.

[9] T. Csendes and E. Antal, "Pagerank based network algorithms for weighted graphs with applications to wine tasting and scientometrics," *Proceedings of the 8th International Conference on Applied Informatics*, pp. 209–216, 2010.

[10] T. Haveliwala, S. Kamvar, and G. Jeh, "An analytical comparison of approaches to personalizing pagerank," Stanford University, Tech. Rep., 2003.

[11] J. R. Norris, *Markov chains*. Cambridge University Press, 1998, no. 2008.