Student Evaluation by Graph Based Data Mining of Administrational Systems of Education

András London, Tamás Németh

Abstract: *In this paper, we define a modified PageRank algorithm as a data mining technique with the aim of evaluating the achievements of students and generate a ranking between them. We applied our method to the data set of a complex administrational system that contains numerous well detailed data of several schools in public education. In order that the method can be applied, we constructed a directed, weighted network of students, where edges of the network represent the comparability of the students in an appropriate way. We compared the results of our method with the standard statistical techniques that are used for rating and ranking the students and observed that our method gives a clearer picture about their educational achievements. Further advantages of graph based data mining techniques in educational systems are also highlighted.*

Key words: *Computer systems, data mining, education, complex systems, networks, PageRank*

INTRODUCTION

Graph based data mining and network analysis have become very popular in the last decade since these tools have been proved to be extremely applicable and useful in a wide range of areas including biology, economy and sociology, among others. The large amount of available data allowed us to analyse and study of such large-scale systems. Usually, these complex systems can be mathematically represented by graphs or networks, where vertices (or nodes) stand for individuals, while edges (or links) represent the interaction between pairs of these individuals. For some excellent review and a book in the field, see e.g. [1, 2, 3]. The network approach is not only useful for simplifying and visualising this enormous amount of data, but it is also very effective to pick up the most important elements and find their most important interactions. Furthermore, several techniques have been developed for rating and ranking the individuals according to their position and role in the underlying graph that represents the bilateral relationships between them. Ranking of individuals based on the modelling interaction graph of them has also become a central ingredient of the widely-known Google's PageRank algorithm [4], where the vertices of the graph are ranked on the basis of their centrality in a diffusion process on the graph. Diffusion process or random walk based graph algorithms, which were originally developed for ranking web pages, have recently been applied in various topics, including for instance citation network analysis [5, 6], sport competitions [7, 8, 9], or evaluating the skills of wine tasters [10, 11].

In this paper, we continue this direction, and introduce a novel example of a real social system, taken from the world of public education which is suitable for network representation. For a good survey on educational data mining we refer to [12]. The Simple Network Workflow for Schools system (SNW) which we investigated is a complex administrational software system of more than sixty institutes of public education (including elementary schools, technical colleges, secondary schools and educational institutes of arts) with electronic diaries, quality management, measurement, and evaluation systems. The database of the system contains the administration of the daily work of more than 10 thousand teacher and 100 thousand students.

Description of the lessons including the equipment and educational methods were used, the competence areas were developed, the students were participated and the marks of the students are all contained in the database. Since a huge amount of data (which is more accurate as well) is produced by such a system, instead of the classical statistical methods, a new type of data processing is required to handle with this data in order to information extraction. The maintainers, the leaders and the teachers of the institute, the students and their parents justifiable need is to ask new type of questions about the educational work and quality of the institute and achievement of the students. Such questions, among others, can be the following:

- Can we say something more useful about the efficiency of certain teachers' work by using this data than by using the classical questionnaire system?
- Can we make spectacular statements that are easier to understand, e.g. with data visualisation techniques?
- By applying new type of models, can we get results that are modelling the reality better, than the simple statistical statements, e.g. for comparing the achievements of the students in the same year?
- Can we "measure" the improvement of the students and by using these measures, can we detect accidentally occurring problems, like drug use, alcoholic problems, crisis in the family, etc.?

In this paper, our goal is to define a directed and weighted network of students in the same school and same year in an appropriate way and calculate a rating and ranking of them according to their positions in the network. We apply a ranking algorithm to this network in order to quantify the achievements of the students. However, students have different classes (e.g. due to their specialisation), and the same subjects are not necessarily taught by the same teachers, we point out that our method performs well in general and gives a realistic picture of the students achievements relative to each other.

METHODS AND RESULTS

Definitions and preliminaries

Network analysis is not just a research model, but also an interpretation model based on graph theory that is an independent field of combinatorics. Formally, the pair $G = (V, E)$ is a *graph*, where $V = (1,2, \dots, n)$ denotes the set of vertices and $E \subseteq V \times V$ denotes the set of edges. The graph is *directed*, if the edges have a direction. In this work, we deal with directed graphs. The *adjacency matrix* of G is the $n \times n$ matrix $A = [a_{i,j}]_{i,j}$ with entries $a_{ij} = 1$, if (i, j) an edge, and $a_{ij} = 0$, otherwise. The *out-degree* and *in-degree* of a vertex i is defined as $d_i^+ = \sum_{j=1}^{n} a_{ij}$ and $d_i^- = \sum_{i=1}^{n} a_{ij}$, respectively.

Let $D = diag(d_i^+, \dots, d_i^+)$ be the diagonal matrix of the out-degrees, i.e. $d_{ii} = d_i^+$ $(i = 1,2, \dots, n)$ and $d_{ij} = 0$, if $i \neq j$. A *random walk* on the vertices of G can be defined by the *transition probability matrix* $P = D^{-1}A$ with entries $p_{ij} = a_{ij}/d_i^+$ and by an initial probability distribution p^0 on the set of vertices. This random walk defines a *Markov chain* (more description about Markov chains, see e.g. [13]) $M = (G, P)$, where the set of vertices of G stands for the states of M and the probabilities p_{ij} associated with the edges (i, j) stand for the transition probabilities of the chain. The assumption that graph G is *strongly connected* (means that there exist a directed path between any two nodes of G) is equivalent the *irreducibility* of the Markov chain and it is well-known that in this case, that M has a

unique *stationary distribution* $\pi = (\pi_1, \dots, \pi_n)$ satisfies $\pi_i > 0 \quad (i = 1,2, \dots, n)$ and the equation

$$\pi P = \pi. \tag{1}$$

Intuitively, if we consider a random walk on the points of the graph starting from a vertex i, the walker randomly chooses one of the out-going edges incident to i with uniform probability and steps to the end point j of that edge and repeat this process from j, and so on. Since the graph is strongly connected, it is guaranteed that the process never stops. The stationary distribution π_i of vertex i is interpreted as the long-term fraction of the number of visits in i during the random walk, as the number of steps goes to infinity. However, the assumption that the graph is strongly connected is very strong, and, in general, this is not true for "real-world networks", where usually there are nodes with zero out-degree. In this case, it can happen that a random walk stops just after a few steps by reaching such a node, moreover, there is no unique solution of Eq. (1). The PageRank method [4], developed by Sergey Brin and Larry Page (originally for quantifying the importance of web pages with respect to their positions in the "web graph") resolves this problem. The idea behind their algorithm is to start the walk in any vertex of the graph, and "restart" the walk in a randomly chosen vertex after a certain number of steps. The PageRank of a vertex i is defined by the recursion formula

$$PR(i) = \frac{1-d}{n} + d \sum_{j:j\to i} \frac{PR(j)}{d_j^+}, \tag{2}$$

where d is a free parameter between 0 and 1, usually called *damping factor.* To see the similarity with Eq. (1), the PageRank formula can be written in vector equation form and it can be shown, that the PageRank is the stationary distribution of that random walk for which the transition probability matrix is defined as $J(1-d)/n - dP$, where J is the $n \times n$ matrix with all entries equal to 1.

Data set and mathematical model

A crucial point of modelling a complex system by a network is the definition of the underlying graph that mathematically represents the network. As the first step, we collected the data from the database of the SNW-system. We used a dataset of 283 students in the same (secondary) school in the same year and considered all the end-year school reports of them. We defined the network of these students, such that each vertex of the network represents a student, and we defined the connection between two students in the following way: first we assumed, that two student can be compared, if they had at least one common subject taught by the same teacher; then we calculated the sum of differences between the end-year marks of such subjects (see Figure 1-2). Formally, if the common subjects (with the same teacher) and the end-year marks of student i and student j are $s_1, \dots, s_t$ and $(m_1^1, \dots, m_t^1)$, $(m_1^2, \dots, m_t^2)$, respectively, than we calculate

$$w_{ij} = \sum_{k=1}^{t} (m_k^1 - m_k^2). \tag{3}$$

If $w_{ij} > 0$ ($w_{ij} < 0$), we define a directed edge from vertex j to vertex i (from vertex i to vertex j, respectively) with the weight $|w_{ij}|$. If $w_{ij} = 0$, there is no edge between i and j. Finally, we considered a random walk of this graph with respect to the edge weights defined in (3), such that the transition probability from vertex i to vertex j is defined as

$$p_{ij} = \frac{w_{ij}}{\sum_{j:i\to j} w_{ij}}. \tag{4}$$

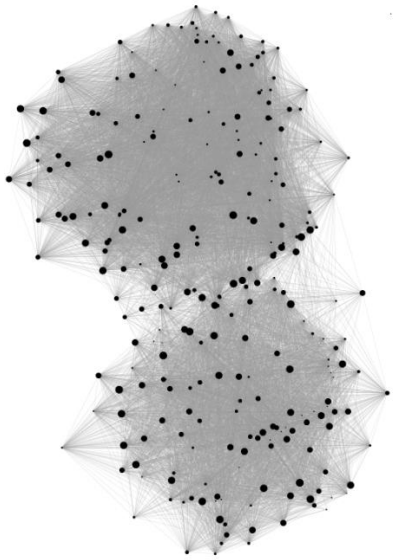

Figure 1. Network of 293 students in a secondary school in the same year. The bigger is the point, the higher is the PR value of the student.

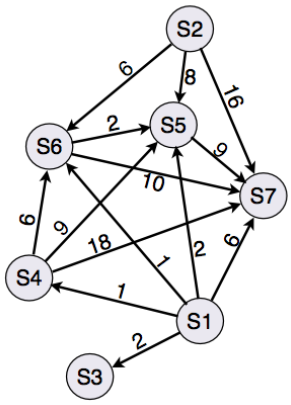

Figure 2. A subnetwork of the original network as a small example for the construction of the contact graph.

In this way, we get a probability distribution over the out-going edges of each node. The random walk on the vertices of the graph goes with respect to these distributions, and the PageRank value of vertex (i.e. student) i is obtained by the recursion formula

$$PR(i) = \frac{1-d}{n} + d \sum_{j:j\to i} p_{ij}\, PR(j). \tag{5}$$

The stationary values of the scores $PR(i)$ can be computed recursively. We set $PR(i) = 1/n$ $(i = 1, \ldots, n)$ at the beginning (but results do not depend on the initial values), and we iterated Eq. (5) until the PageRank values were converged to stable values. We stopped the iteration, when the Euclidean distance between the last two PageRank vectors were calculated became less the 0.001.

Experimental results

We applied the modified PageRank algorithm described before to the graph that we constructed. First, we checked the sensitivity of our method for different values of the damping factor d of Eq. (5). We observed that the method is robust against the choice of d (Figure 3) that was confirmed by the high correlation of the resulting PageRank vectors contains the PR scores of the students (the Pearson correlation was more than 0.9 for each pair of result vectors). For further studies, we chose $d = 0.8$, which is traditional in PageRank computation.

We used *Kendall's* τ *correlation* method [14] to quantify the rank correlation between the rankings obtained by the PageRank and the average method (which is simply gives a ranking of students by comparing the average of the end year marks of them). Although, the correlation coefficient was 0.68, which shows positive correlation, many differences can be seen between the two methods (as in Figure 4). We normalised the PR values, such that the obtained values of the two methods are in the same scale (the normalisation function was $10^{-28} + 1.035^{x}/\,9 \cdot 10^{6}$). We observed in general, that if the end-year average of a student is high, but the PR value of her is relatively small, than the student has only a few subjects, where "easy" to get mark five. The statement was justified by checking all the marks of those subjects. On the other hand, if a student i has low average, but her PR value is high compared to the others, it is usually true that most of the

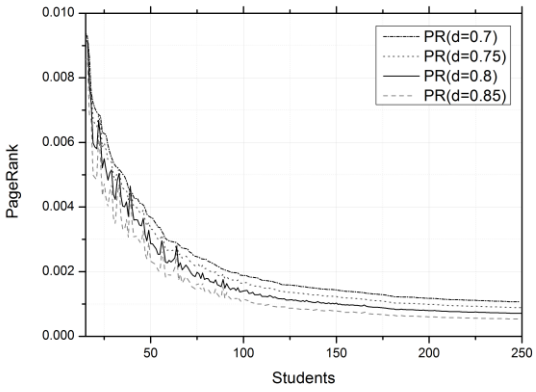

Figure 3. PageRank values of the students for different values of d. The sensitivity verified, that the method behaves robustly in the network with respect to d

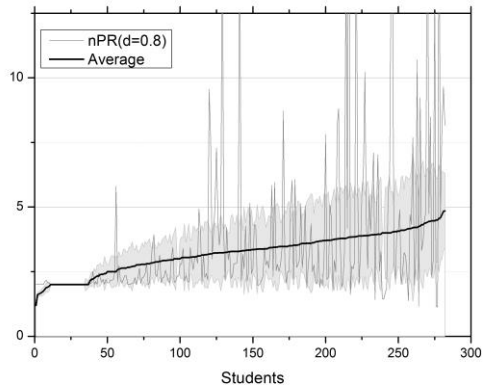

Figure 4. Relation between normalised PR scores and the average of the end-year marks.analysis The light grey area shows the domain of the variance of the average .

students taught by the same teachers also have low marks such as i, but high PR value of i means that she exceeds the performance of their schoolmates.

We checked the relationship between the PR values (and ranks resulting from that; the relation between the PR ranks and average ranks is shown in Figure 5) and the variance of the end-year marks. It can be observed in Figure 4, that outstanding PR values occur when the variance is high. Generally, it was verified, that if the variance of the marks of a student was high and the PR value of her was also high, than the student is talented in at least one subject. The high variance of the marks caused by large variety of the marks, but the large PR value must be caused by just one (or only a few) subject (of the same teacher), where the student was significantly better than her classmates.

It is also interesting to analyse the relationship between the difference of the normalised PR scores and average values, and the variance of the end-year marks (Figure 6). It can be checked that the difference between the PR score and the average of the marks is small in general regardless to the variance is low or high. It suggests that it should be paid much attention for that students for whom this difference is high. In these cases, it also should be checked, what causes this high difference.

By using this network based method, the talents, the problematic students, the martinet or too lenient teachers can be filtered out. By considering students in the same class, uniformly high or uniformly low PR scores (for instance) can also give a good picture about the difficulty of the different subjects and/or the personality of the teacher of a certain subject as well as the achievements of the a class of students in a global point of view. The PageRank method is also very effective in finding the best students in the same year. After filtering out the "outliers" (e.g. student who has just a very few marks because, for example, she left the school), PR scores give a good relative order of the students with respect to their achievements. Such rankings can also be useful to decide that which students need to be awarded in the end of the year, and also good for the teachers and parents to follow the educational progression of the students.

CONCLUSIONS AND FUTURE WORK

Graph based algorithms in data mining have been proved to be efficiently applicable in many different areas. In this paper we defined a random walk based graph algorithm and applied it to a network of students in a secondary school. According to our method, the achievements and ranking of the students are not only analysed and determined by simple statistical techniques, but the pairwise comparisons of the students and the complex system representation of them were also considered.

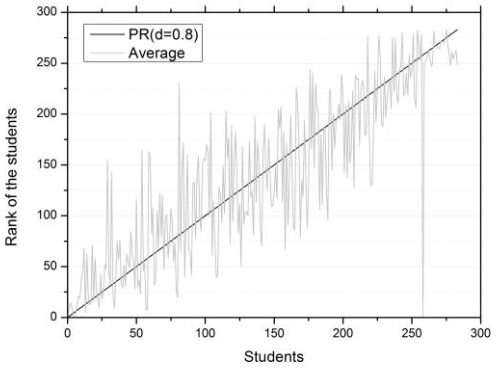

Figure 5. Ranks of the students obtained by the different methods. Ranks were ordered by the PR scores.

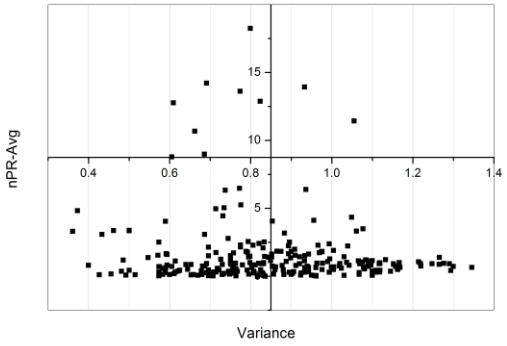

Figure 6. Relation between the variance of the end-year marks and difference of the normalized PR scores and average values.

We observed that our method gives a better picture about the students' relative performance, and it can also find outstanding and relatively weak students. In our experiments, the PageRank method especially gives a good picture about the students in that case, if we do not want to evaluate their average performance, but we want to investigate whether the student is outstandingly better than her schoolmates. Our method is able to eliminate the common drawback of the standard statistical methods, that is, they are not sensitive to the evaluation habits of the different teachers. From the achievements and rankings of the students, we can also conclude to the efficiency and quality of the teachers' work.

In the future, it would also be interesting to define the network of the students in various ways and investigate them by different graph based algorithms. One plausible idea is to define a bipartite graph of students and teachers with the aim of evaluating the personal properties of the teachers in the educational point of view, such as rigorism, excessive benevolence or consistency in the evaluation of the students. *Kleinberg's HITS* algorithm [15] was used for similar problems before [11], and the *SALSA algorithm* [16] by Moran and Lempel can also handle with these types of problems.

Another possible direction of studies is to define the network of students based on various similarity measures (i.e. stronger edge weights represent high level similarities between the students). Such similarities can be defined based on the end-year reports or the similarities of the students' marks, such that the subjects are differentiated to more general groups (such as natural sciences, arts, and humanities), etc. If such network is defined, a relevant problem can be the investigation of the topological properties of it, mainly the examination of its clustering and modularity structure and finding whether which particular reasons produced this structure.

Acknowledgement

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: **TAMOP-4.2.2.C-11/1/KONV-2012-0013**).

András London was supported by the **European Union** and the **State of Hungary, co-financed by the European Social Fund** in the framework of **TAMOP-4.2.4.A/2-11-1-2012-0001** 'National Excellence Program'.

REFERENCES

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. U. Hwang, "Complex networks: Structure and dynamics", *Physics reports,* vol. 424, no. 4, pp. 175-308, 2006

[2] E. Estrada. *The structure of complex networks: theory and applications.* New York, Oxford University Press, 2011

[3] M. E. J. Newman, "The structure and function of complex networks", *SIAM Review* vol. 45, no.2, pp. 167-256, 2003

[4] 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

[5] P. Chen, H. Xie, S. Maslov and S. Redner, "Finding scientific gems with Google's PageRank algorithm", *Journal of Informetrics*, vol. 1, no. 1, pp. 8-15, 2007

[6] D. Walker, H. Xie, K.K. Yan and S. Maslov, "Ranking scientific publications using a model of network traffic", *Journal of Statistical Mechanics: Theory and Experiment*, P06010, 2007

[7] A. London, J. Németh and T. Németh, "Time-dependent network algorithm for ranking in sports", *accepted for publication in Acta Cybernetica*, 2014

[8] S. Motegi and N. Masuda, "A network-based dynamical ranking system for competitive sports", *Scientific Reports*, vol. 2, 2012

[9] F. Radicchi, "Who is the best player ever? A complex network analysis of the history of professional tennis", *PLoS One*, vol. 6, no. 2, e17249, 2011

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

[11] A. London and T. Csendes, "HITS based network algorithm for evaluating the professional skills of wine tasters", *in Proc. of the IEEE 8th International Symposium on Applied Computational Intelligence and Informatics*, 2013, pp. 197-200.

[12] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005", *Expert Systems with Applications*, *vol. 33, no.1, pp. 135-146, 2007*

[13] J. R. Norris. *Markov chains*. New York, Cambridge University Press, 1998

[14] M.G. Kendall, "A new measure of rank correlation", *Biometrika*, vol. 30, pp. 81–93, 1938

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

[16] 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

ABOUT THE AUTHORS

András London, PhD student, Institute of Informatics, University of Szeged,
phone: +36 62 543444, email: london@inf.u-szeged,hu
web: http://www.inf.u-szeged.hu/~london/

Tamás Németh, Assistant professor, Institute of Informatics, University of Szeged,
phone:+36 62 543435, email: tnemeth@inf.u-szeged.hu
web: http://www.inf.u-szeged.hu/~tnemeth/