Application of graph based data mining techniques in administrational systems of education

András London and Tamás Németh

Graph based data mining and network analysis have become widespread 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 social sciences, among others. The large amount of available data allowed us to study of such large-scale systems that appears in the mentioned areas. Usually, these complex systems can be 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 an excellent review, see *e.g.* [1]). The network approach is not only useful for simplifying and visualising this enormous amount of data, but it also very effective to pick up the most important elements and find their most important interactions. Besides, several techniques have been developed to explore the deeper topological features of a network, such as community structure [2], core-periphery structure [3] or small-world property [4] and scale-freeness [5]. Ranking individuals based on their position in the model interaction graph has also become an important direction of studies in the last decade. Random walk based ranking algorithms (such as the widely-known PageRank algorithm by Google), that were originally developed for ranking web pages, have been used recently for such different purposes like citation network analysis [6], ranking in sports *e.g.* in [7, 8] or evaluating the quality of wines and skills of the tasters [9], etc.

In this work, by following the complex network approach, we introduce a novel example of a real social system taken from the world of public education. Since a huge amount of data (that is more accurate as well) is produced by a complex administrational software system of educational institutes, new type of data processing methods are required to handle with it in order to information extraction, instead of the classical statistical analysis. The maintainers, the leaders and the teachers of the institute, the students and their parents would have asked new type of questions about the educational work and quality of the institute, the teachers and students. Such questions, among others, can be the following:

- Can we say something more useful about the efficiency of the 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 visualization techniques?
- By applying new type of models, can we get results that are modeling 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 results, can we detect accidentally occurring problems, like drug use, alcoholic problems, crisis in the family, etc.?

We define several suitable network representations of the system with the goal of answering or partially answering these questions. Depending on the construction of the underlying graph we consider three different network models. First we construct an undirected, weighted graph based on various similarity measures between students (in the same school and year) and investigate the topological features, especially the community structure of it. Second we define a directed and weighted graph based on the same data set and apply a ranking algorithm to this network in order to quantify the achievements of the students. Third we construct a bipartite

graph of students and teachers that can be used to compare both the students and teachers with each other according to different aspects. Although we use different algorithms to handle with the different representations, we try to analyze the results simultaneously to get a clear and detailed picture about the achievement and quality of students and teachers and to answer such questions mentioned above.

Keywords Data mining, Educational evaluation, PageRank, Modularity

Acknowledgements

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] M.E.J Newman. The structure and function of complex networks. *SIAM review*, 45(2), 167-256, 2003

[2] M.E.J Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2), 026113, 2004

[3] P. Csermely, A. London, L.Y. Wu and B. Uzzi. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 1(2), 93-123, 2013

[4] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440-442, 1998

[5] A.L. Barabási. Scale-free networks: a decade and beyond. *Science* 325(5939), 412-413, 2009

[6] P. Chen, H. Xie, S. Maslov and S. Redner. Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, 1(1), 8-15, 2007

[7] S. Motegi and N. Masuda. A network-based dynamical ranking system for competitive sports. *Scientific reports*, 2, 904, 2012

[8] F. Radicchi. Who is the best player ever? A complex network analysis of the history of professional tennis. *PLoS One*, 6(2), e17249, 2011

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