Data Min Knowl Disc (2012) 25:415–419
DOI 10.1007/s10618-012-0271-0

Guest editorial: special issue on data mining technologies for computational social science

Fei Wang · Hanghang Tong · Phillip Yu · Charu Aggarwal

Received: 8 May 2012 / Accepted: 14 May 2012 / Published online: 19 June 2012

1 Data mining and computational social science

Social Science has long been used as an umbrella term to refer to a plurality of fields outside natural sciences, such as economics, linguistics, education and even psychology. It can be any discipline or branch of science that deals with the sociocultural aspects of human behavior. There are many fundamental problems in social sciences, such as detecting underlying communities, analyzing the mechanism of a specific behavior (social activity) and discovering the evolutionary patterns in a community. The emergence of online social network sites and web 2.0 applications generate a large volume of valuable data. This greatly stimulates the development of computational social science, which tries to solve the research problems in traditional social science with the help of computational technologies.

In recent years, the computing technologies, such as data mining, machine learning and statistics, has been developing rapidly. More and more sophisticated methods, such as Support Vector Machine (Cristianini and John 2000), Matrix and Graph based

Responsible editor: Geoffrey I. Webb.

F. Wang (✉) · H. Tong · C. Aggarwal
IBM T. J. Watson Research Center, Hawthorne, NY 10532, USA
e-mail: feiwang03@gmail.com; fwang@us.ibm.com

H. Tong
e-mail: htong@us.ibm.com

C. Aggarwal
e-mail: charu@us.ibm.com

P. Yu
Department of Computer Science, University of Illinois at Chicago, Chicago, IL USA
e-mail: psyu@cs.uic.edu

models (Li et al. 2011), Probabilistic Networks (Cowell et al. 1999), and Parallel Computation (Almasi and Gottlieb 1989), have been proposed and many of them have already been deployed in the analysis of social sciences.

Computational social science was first introduced in (Lazer et al. 2009), which aims to understand the individual and group behavior by analyzing large-scale data sets. At the macro-level, many important patterns of the global statistics of the underlying social networks have been discovered in the past, e.g., power-law distribution and small diameter (Albert et al. 1999; Faloutsos et al. 1999; Newman 2003), the frequent substructure (Xin et al. 2005), the evolution and dynamics of social networks (Leskovec et al. 2005; Kumar et al. 2006), the dynamics of the on-line conversation (Kumar et al. 2010), the connected and disconnected components of social networks (Kang et al. 2010; McGlohon et al. 2008), influence propagation (Kempe et al. 2003; Leskovec et al. 2007; Wang et al. 2011b; Cui et al. 2011), the group and community structure (Girvan and Newman; Backstrom et al. 2006; Leskovec et al. 2010), human mobility (González et al. 2008; Wang et al. 2011a), etc. There are also extensive work to study the social networks at the micro-level, e.g., ranking the importance of nodes (e.g., people) (Page et al. 1998); proximity measure in social networks (Tong et al. 2006), link prediction (Liben-Nowell and Kleinberg 2003), triangle counting (Leskovec et al. 2008; Tsourakakis et al. 2009), radius estimation and characterization (Kang et al. 2011), etc.

The major goal of this special issue is to bring together the researchers in the intersection of data mining and computational social sciences to illustrate pressing needs, demonstrate challenging research issues, and showcase the state-of-the-art research and development.

2 The special issue

Recently, we started some activities to establish the platform for exchanging the ideas on data mining for computational social science. We have successfully organized the workshop on *Data Mining Technologies for Computational Collective Intelligence* (DMCCI) held in conjunction with IEEE International Conference on Data Mining (ICDM) 2011. This special issue provides a leading focused forum for timely, in-depth presentation of recent advances in algorithms, theory and applications on data mining technologies for computational social science. The selected papers underwent a rigorous extra refereeing and revision process.

There are some classic problems in computational social science, such as classification and community detection. The paper by Prakash Mandayam Comar, Pang-Ning Tan and Anil K. Jain considers the problem of multi-task learning on heterogeneous network data. Specifically, they present a matrix factorization based framework that enables one to perform classification on one network and community detection in another related network. The authors also show that the framework can incorporate prior information about the correspondences between the clusters and classes in different networks.

The paper by Hong Cheng, Yang Zhou, Xin Huang and Jeffrey Xu Yu considers the problem of clustering on large attributed information network. Here social network

can be viewed as a specific type of information network. The authors' algorithm is built upon the *SA-Cluster* approach, which performs matrix multiplication to calculate the random walk distances between network vertices, and the network edge weights are iteratively adjusted to balance the relative importance between structural and attribute similarities. The authors propose an efficient method to incrementally update the random walk distances given the edge weight increments. Moreover, a parallel matrix computation design is also presented in the multicore environment.

In addition to novel computational models, it would also be interesting to see the application of data mining techniques in real social problems. The paper by Shan Jiang, Joseph Ferreira and Marta C. González analyzes an activity-based travel survey conducted in the Chicago metropolitan area over a demographic representative sample of its population. They examine the cluster structure of the data set via Principal Component Analysis and K-means clustering techniques and find out the population can be clustered into 8 and 7 representative groups according to their activities during weekdays and weekends. This provides a new perspective for urban and transportation planning as well as for emergency response and spreading dynamics, by addressing when, where, and how individuals interact with places in metropolitan areas.

Besides clustering and classification, there are some newly emerged problems in computational social science that have attracted considerable interests from the researchers in relevant fields. One such typical example is social influence analysis. The paper by Lu Liu, Jie Tang, Jiawei Han and Shiqiang Yang studies the problem of quantitatively learning influence between users from heterogeneous networks. The authors propose a generative model which leverages both heterogeneous link information and textual content associated with each user in the network to mine topic-level influence strength. The authors also study the influence propagation and aggregation mechanisms: conservative and non-conservative propagations to derive the indirect influence. The proposed model is validated on Twitter, Digg, Renren, and Citation.

The paper by Chi Wang, Wei Chen and Yajun Wang considers the problem of influence maximization, which aims at finding a small set of seed nodes in a social network that maximizes the spread of influence under certain influence cascade models. The scalability of influence maximization is a key factor for enabling prevalent viral marketing in large-scale online social networks. They design a new heuristic algorithm that is easily scalable to millions of nodes and edges in their experiments, and their algorithm has a simple tunable parameter for users to control the balance between the running time and the influence spread of the algorithm. The authors also conduct extensive simulations on several real-world and synthetic networks demonstrating their algorithm is currently the best scalable solution to the influence maximization problem.

The paper by Zhongmou Li, Hui Xiong and Yanchi Liu studies the problem of mining blackhole and volcano patterns in directed graphs, where a blackhole pattern refers to the group of nodes whose average in-weight of is significantly larger than the average out-weight of the same group, while finding volcano patterns is a dual problem of mining blackhole patterns. The authors develop two pruning schemes, gBlackhole and approxBlackhole, to reduce the search computational cost by reducing both the number of candidate patterns and the average computation cost for each candidate pattern. The authors did experimental results on real-world data demonstrating that

the performance of approxBlackhole can be several orders of magnitude faster than gBlackhole, and both of them have huge computational advantages over the brute-force approach.

With the burst of various online social network, people are also interested in mining the people emotions from those online data. The paper by Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco and André Montoyo focuses on the creation of EmotiBlog, a fine-grained annotation scheme for labelling subjectivity in non-traditional textual genres. They also present the EmotiBlog corpus, a collection of blog posts composed by 270,000 tokens about 3 topics and in Spanish, English and Italian. A series of experiments are carried out on checking the robustness of the model and its applicability to Natural Language Processing tasks with regards to the 3 languages. The authors also apply EmotiBlog to Opinion Mining and show that their resource allows an improvement the performance of systems built for this task.

In the last paper of this special issue, Ian Davidson, Sean Gilpin and Peter B. Walker explore the analysis of human behavior encoded as a trail of their events over time and space, which they refer to as behavioral event data. Specifically, by using the adversarial event behavior of blue and red forces, the authors show three core problems and solutions in event behavior analysis: (1) Decomposing behavior to identify areas of intense activity, (2) Predicting what groups of events are likely to occur, and (3) Analysis to identify interacting behavior given a known template.

Acknowledgements We are very indebted to the reviewers who reviewed the papers very carefully. We would also like to thank all the authors who submitted their papers to the special issue. Special thanks to Prof. Geoff Webb for his great help and support in organizing the issue.

References

Albert R, Jeong H, Barabasi A-L (1999) Diameter of the world wide web. Nature 401:130–131

Almasi GS, Gottlieb A (1989) Highly parallel computing. Benjamin-Cummings, Redwood City

Backstrom L, Huttenlocher DP, Kleinberg JM, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: KDD, pp 44–54

Cowell RG, Dawid AP, Lauritzen SL, David JS (1999) Probabilistic networks and expert systems. Springer, Berlin

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge

Cui P,Wang F, Liu S, Ou M, Yang S, Sun L (2011) Who should share what?: Item-level social influence prediction for users and posts ranking. In: Proceeding of the 34th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2011, pp 185–194

Faloutsos M, Petros P, Faloutsos C (1999) On power-law relationships of the internet topology. In: SIGCOMM, Aug–Sept, pp 251–262

Girvan M, Newman MEJ (2002) Community structure is social and biological networks. Proc Natl Acad Sci USA 99(12):7821–7826

González MC, Hidalgo CA, Albert-László B (2008) Understanding individual human mobility patterns. Nature 453:779–782

Kang U, McGlohon M, Akoglu L, Faloutsos C (2010) Patterns on the connected components of terabyte-scale graphs. In: ICDM, pp 875–880

Kang U, Tsourakakis CE, Ana Paula A, Christos F, Jure L (2011) Hadi: Mining radii of large graphs. TKDD 5(2):8

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: KDD

Kumar R, Mahdian M, McGlohon M (2010) Dynamics of conversations. In: KDD, pp 553–562

Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: KDD, pp 611–617
Lazer D, Pentland A, Adamic L, Aral S, Barabási A, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T (2009) Computational social science. Science 323:721–723
Leskovec J, Kleinberg JM, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: KDD, pp 177–187
Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. In: TWEB. ACM Trans. Web 1(1):5
Leskovec J, Backstrom L, Kumar R, Tomkins A (2008) Microscopic evolution of social networks. In: KDD, pp 462–470
Leskovec J, Lang KJ, Mahoney MW (2010) Empirical comparison of algorithms for network community detection. In: WWW, pp 631–640
Li T, Ding C, Wang F (2011) Guest editorial: Special issue on data mining with matrices, graphs and tensors. Data Min Knowl Disc 22(3):337–339
Liben-Nowell D, Kleinberg JM (2003) The link prediction problem for social networks. In: CIKM, pp 556–559
McGlohon M, Akoglu L, Faloutsos C (2008) Weighted graphs and disconnected components: patterns and a generator. In: KDD, pp 524–532
Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. Paper SIDL-WP-1999-0120 (version of 11/11/1999)
Tong H, Faloutsos C, Pan J-Y (2006) Fast random walk with restart and its applications. In ICDM, pp 613–622
Tsourakakis CE, Kang U, Miller GL, Faloutsos C (2009) Doulion: counting triangles in massive graphs with a coin. In: KDD, pp 837–846
Wang D, Pedreschi D, Song C, Giannotti F, Barabási A (2011) Human mobility, social ties, and link prediction. In: KDD, pp 1100–1108
Wang D, Wen Z, Tong H, Lin C-Y, Song C, Barabási A (2011) Information spreading in context. In: WWW, pp 735–744
Xin D, Han J, Yan X, Cheng H (2005) Mining compressed frequent-pattern sets. In: VLDB, pp 709–720