Eighth Workshop on Mining and Learning with Graphs

Ulf Brefeld
Yahoo! Research
Barcelona, Spain
brefeld@yahoo-inc.com

Lise Getoor
Computer Science
University of Maryland
getoor@cs.umd.edu

Sofus A. Macskassy
Fetch Technologies
El Segundo, CA
sofmac@fetch.com

ABSTRACT

The Eighth Workshop on Mining and Learning with Graphs (MLG)[1]was held at KDD 2010 in Washington DC. It brought together a variete of researchers interested in analyzing data that is best represented as a graph. Examples include the WWW, social networks, biological networks, communication networks, and many others. The importance of being able to effectively mine and learn from such data is growing, as more and more structured and semi-structured data is becoming available. This is a problem across widely different fields such as economics, statistics, social science, physics and computer science, and is studied within a variety of sub-disciplines of machine learning and data mining including graph mining, graphical models, kernel theory, statistical relational learning, etc. The objective of this workshop was to bring together practitioners from these various fields and areas to foster a rich discussion of which problems we work on, how we frame them in the context of graphs, which tools and algorithms we apply and our general findings and lessons learned. This year's workshop was very successful with well over 100 attendees, excellent keynote speakers and papers. This is a rapidly growing area and we believe that this community is only in its infancy. We hope that the readers will join us next year for MLG 2011!

Keywords

Graph mining, dynamic network analysis, link mining, data mining, machine learning, relational learning, network analysis, statistical relational learning, pattern recognition, kernel methods, scalable graph mining.

1. INTRODUCTION

The analysis of graphs is a problem which spans widely different fields such as economics, statistics, social science, physics and computer science, and is studied within a variety of sub-disciplines of machine learning and data mining including graph mining, graphical models, kernel theory, statistical relational learning, etc. Many contributions developed in one area can have a direct impact on others once a common abstraction of the underlying problems is established.

The objective of the MLG workshop was therefore to bring together researchers from a variety of these areas, and discuss commonality and differences in challenges faced, survey some of the different approaches, and provide a forum to present and learn about some of the most cutting edge research in this area. One of our desired outcomes was to have participants walk away with a better sense of the variety of different tools available for graph mining and learning, and an appreciation for some of the interesting emerging applications for mining and learning from graphs. One of the key challenges we addressed in this workshop was how to efficiently analyze large data sets that are relational in nature and hence easily represented as graphs. Such data are becoming ubiquitous in a plethora of application and research domains and now was an opportune time to bring together people from these various fields to exchange ideas about how we can mine and learn from these large data sets. As such, one of the primary goals of this workshop was to explore the state-of-the-art algorithms and methods, leveraging existing knowledge from other sub-disciplines, to examine graph-based models in the context of real-world applications, and to identify future challenges and issues. In particular we were interested in exploring following topics:

- Graph mining
- Kernel methods for structured data
- Probabilistic models for structured data
- (Multi-)relational data mining
- Methods for structured outputs
- Network analysis
- Large-scale learning and applications
- Sampling issues in graph algorithms
- Evaluation of graph algorithms
- Relationships between mining and learning with graphs and statistical relational learning
- Relationships between mining and learning with graphs and inductive logic programming
- Semi-supervised learning
- Active learning
- Transductive inference
- Transfer learning

The remainder of our report will discuss the format of the workshop, the themes and feedback we received from participants, and details about the keynote speakers and accepted papers. We finish by acknowledging everybody whom we felt had a large part in the overall success of this workshop.

2. WORKSHOP FORMAT

This workshop was the eighth in a series of workshops organized either as standalone workshops, or in conjunction with machine learning conferences (ICML and ECML). This is the first time that it has collocated with KDD, and we were very excited about the opportunity to bring MLG researchers together with KDD researchers. This was also the first time the workshop was held in the US. MLG has historically been a two-day workshop and we

[1] http://www.cs.umd.edu/mlg2010/

were very pleased when SIGKDD accepted our proposal and granted us a two-day workshop. We believe that the nature of this helped improve participation significantly.

We decided on a format where we would have a large number of keynote speakers to introduce the scope of MLG to this audience. We had only a small number of paper presentations, a large poster session, a final wrap-up discussion and many long breaks, all in order to facilitate in-depth discussions and fostering face-to-face interaction. We believe that this format was very successful, and the breaks and the poster session were filled with discussion much as we had hoped. Our final wrap-up session was well-attended and we received much useful feedback from the audience and participants.

3. THEMES AND FEEDBACK

We felt that this workshop was a great success all around. We had an unexpectedly large number of attendees and participants and all attendees whom we spoke with (presenters, keynote speakers and general audience alike) were universally very positive about the workshop. In particular, the keynote sessions and the poster session were two parts of the workshop which received the most positive feedback and we believe this workshop format really helped with the participation.

There were a number of themes that emerged from the workshop. Certainly the statistical challenges in dealing with graph data and the scaling challenges in dealing with large graph data were common. Other themes where the need to operate with dynamic graph data, handling missing and noisy data (missing attributes, nodes or edges), handling heterogeneous graph data, and graph databases. At the same time, there were a diverse range of techniques leveraged, including graphical models, kernel methods, spectral methods and frequent pattern algorithms.

There was a general agreement that identifying and working on concrete applications of graph modeling would be a useful way to advance this field and that having more diverse and rich social network data sets would also be useful.

4. ATTENDANCE

While we have not received the final number of registrants to the workshop, we counted well over 100 attendees during the first day. The second day, our workshop was competing with two other network-centric workshops at SIGKDD and we saw the attendance drop to about 80. The number of attendees was quite astonishing to us, and the rooms we had been assigned were too small to accommodate all participants. We were lucky on the first day to have one wall taken down to double the room-size. This helped in enabling us to have the poster session in our assigned room without having to spill into the hallways. However, we believe that having a small room the second day also hurt attendance, but it was good to know that those attendees at least had other network-centric workshops they could attend instead. The number of attendees was also unexpected as we only had 65 official registrants a week before SIGKDD was held. It also speaks to the timeliness, popularity and general interest of the topics covered in MLG. We believe that the MLG community, despite this being the eight annual workshop, is still in its infancy and we believe that the community will only grow and the area and field will only become more mature in coming years. We look forward to being part of this growing community.

5. INVITED SPEAKERS

As mentioned above, this was the first time that MLG was at SIGKDD and in the US. We decided early on that the best way to introduce MLG to the US and data mining audience was to invite senior high-quality speakers which could community the breadth of topics, problem formulations and algorithms which are at the core of MLG. We settled on seven speakers which would achieve this goal for us and we believe their presence was one critical aspect of making MLG the success it was.

The seven keynote speakers were:

Stephen E. Fienberg, CMU: Graphs for Machine Learning: Useful Metaphor or Statistical Reality

Aristides Gionis, Yahoo! Research: Efficient tools for mining large graphs: Indexing, sampling, counting, and predicting

Thomas Gärtner, University of Bonn and Fraunhofer IAIS: Kernel Methods for Structured Inputs and Outputs

Jennifer Neville, Purdue University: Evaluation Strategies for Network Classification

Padhraic Smyth, UC Irvine: Network Event Data over Time: Prediction and Latent Variable Modeling

Chris Volinsky, AT&T Labs: Mining Massive Graphs for Telecommunication Applications

Eric Xing, CMU: Dynamic Network Analysis: Model, Algorithm, Theory, and Application

6. ACCEPTED PAPERS

We had a large number of high quality submissions, and given that this was a workshop we ended up accepting most of the submissions. We accepted 27 submissions; all 27 submissions were presented at the poster session and a selection of six papers has additionally been presented in contributed talks. The accepted papers are as follows:

Time-Based Sampling of Social Network Activity Graphs, Nesreen Ahmed, Fredrick Berchmans, Jennifer Neville and Ramana Kompella

Structure, Tie Persistence and Event Detection in Large Phone and SMS Networks, Leman Akoglu and Bhavana Dalvi

SVM Optimization for Lattice Kernels, Cyril Allauzen, Corinna Cortes and Mehryar Mohri

A Compact Representation of Graph Databases, Sandra Álvarez, Nieves R. Brisaboa, Susana Ladra and Óscar Pedreira

Binary Bit String Representation for Networks based on Exchangeable Graph Modeling, Hossein Azari, Edoardo Airoldi and Vahid Tarokh

A Community-Based Model of Online Social Networks, Leendert Botha and Steve Kroon

Enhancing Link-Based Similarity Through the Use of Non-Numerical Labels and Prior Information, Christian Desrosiers and George Karypis

Network Community Discovery: Solving Modularity Clustering via Normalized Cut, Chris Ding and Linbin Yu

Analyzing Graph Databases by Aggregate Queries, Anton Dries and Siegfried Nijssen

Multi-Network Fusion for Collective Inference, Hoda Eldardiry and Jennifer Neville

Bayesian Block Modeling for Weighted Networks, Ian Gallagher

An Efficient Block Model for Clustering Sparse Graphs, Adam Gyenge, Janne Sinkkonen and Andras A. Benczur

Centrality Metric for Dynamic Networks, Kristina Lerman, Rumi Ghosh and Jeon Hyung Kang

Design Patterns for Efficient Graph Algorithms in MapReduce, Jimmy Lin and Michael Schatz

Prioritizing Candidate Genes by Network Analysis of Differential Expression using Machine Learning Approaches, Daniela Nitsch

Document Classification Utilising Ontologies and Relations between Documents, Katariina Nyberg, Tapani Raiko, Teemu Tiinanen and Eero Hyvönen

Graph Visualization with Latent Variable Models, Juuso Parkkinen, Kristian Nybo, Jaakko Peltonen and Samuel Kaski

Relational Motif Discovery via Graph Spectral Ranking, Alberto Pinto

Symmetrizations for Clustering Directed Graphs, Venu Satuluri and Srinivasan Parthasarathy

Pruthak- mining and analyzing graph substructures, Swapnil Shrivastava, Kriti Kulshrestha, Pratibha Singh and Supriya N Pal

Structural Correlation Pattern Mining for Large Graphs, Arlei Silva, Wagner Meira Jr. and Mohammed J. Zaki

Meaningful Selection of Temporal Resolution for Dynamic Networks, Rajmonda Sulo, Tanya Berger-Wolf and Robert Grossman

Community Evolution Detection in Dynamic Heterogeneous Information Networks, Yizhou Sun, Jie Tang, Jiawei Han, Manish Gupta and Bo Zhao

Network Quantification Despite Biased Labels, Lei Tang, Huiji Gao and Huan Liu

Frequent Subgraph Discovery in Dynamic Networks, Bianca Wackersreuther, Peter Wackersreuther, Annahita Oswald, Christian Böhm and Karsten Borgwardt

Querying Graphs with Uncertain Predicates, Hao Zhou, Anna Shaverdian, H. V. Jagadish and George Michailidis

Frequent Subgraph Mining on a Single Large Graph Using Sampling Techniques, Ruoyu Zou and Lawrence Holder

7. ACKNOWLEDGMENTS

We would like to foremost thank the authors and keynote speakers without whom this workshop would not have been the success it was. We would like to take the opportunity to thank the large number of attendees which made the workshop vibrant and filled with energy and enthusiasm. Further, we would like to thank our program committee who worked within a very tight schedule to provide thoughtful and clear reviews which were integral in shaping the final workshop program. Lastly, we would like to thank our sponsors: Yahoo! Labs, PASCAL 2, Microsoft Research, and Google. These sponsors made it possible to support and bring in the outstanding keynote speakers.

About the authors:

Ulf Brefeld is a researcher at Yahoo! Research in Barcelona. Prior to joining Yahoo!, he worked for Technische Universität Berlin, Max Planck Institute for Computer Science in Saarbrücken, and at Humboldt-Universität zu Berlin. He received a Diploma in Computer Science in 2003 from Technische Universität Berlin and a Ph.D. (Dr. rer. nat.) in 2008 from Humboldt-Universität zu Berlin. He is interested in statistical machine learning and data mining. This includes learning from structured data, kernel methods, semi-supervised techniques, information extraction/retrieval, and applications in natural language processing and computational biology.

Lise Getoor is an associate professor in the Computer Science Department at the University of Maryland, College Park. She received her PhD from Stanford University in 2001. She is well-known for her work in statistical relational learning. Her current work includes research on social network analysis, link mining, data and metadata alignment, and representing uncertainty in structured and semi-structured data. She has published numerous articles in machine learning, data mining, database, and artificial intelligence forums. She has an NSF Career Award, is an action editor for the Machine Learning Journal, a TKDD associate editor, has been a JAIR associate editor and a member of AAAI Executive council, and has served on a variety of program committees including AAAI, ICML, IJCAI, KDD, SIGMOD, UAI, VLDB, and WWW.

Sofus A. Macskassy is the Director of Fetch Labs at Fetch Technologies and an Assistant Adjunct Professor of Computer Science at the University of Southern California. He is the primary developer of the open-source Network Learning Toolkit for Statistical Relational Learning. He has published numerous articles in statistical relational learning, machine learning, (social) network analytics, and information extraction and integration. The problems he currently focuses on include social media analytics and mining and automatic extraction of relational data from open source data such as the Web. He received his Ph.D. from Rutgers University in 2003 focusing on text mining and information filtering. He has served on a variety of organizing and programming committees, including AAAI, ICML, ILP, KDD, WSDM, and WWW.