KDD-2005 Workshop Report
Link Discovery: Issues, Approaches and Application (LinkKDD-2005)

Jafar Adibi, Patrick Pantel
USC Information Sciences Institute
Marina del Rey, CA, USA
{adibi, pantel}@isi.edu

Marko Grobelnik, Dunja Mladenic
J. Stefan Institute
Ljubljana, Slovenia
{dunja.mladenic, marko.grobelnik}@ijs.si

ABSTRACT

In this paper we provide a summary of the workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005) held in conjunction with ACM SIGKDD 2005, on August 21st in Chicago, Illinois, USA. We report in detail about the research issues addressed in the talks at the workshop.

Keywords

Data Mining, Link Discovery, Link Analysis

1. INTRODUCTION

The LinkKDD-2005 workshop aimed to bring together a diverse group of researchers and industry practitioners to advance the state of the art in link discovery. Recently, there has been increasing interest in developing information technology for Link Discovery (LD). LD research studies and develops data mining techniques for extracting valuable patterns linking together seemingly unrelated items. LD, rooted in fields such as discreet mathematics, graph theory, social science, pattern analysis, link analysis and spatial databases, is relevant to a wide range of research topics that have been developed in past decades. Successful LD systems will discover the hidden structure of organizations, relate groups, identify fraudulent behavior, model group activity and provide early detection of emerging threats. The broader context of this workshop invites both theoretical and applied contributions to LD spanning techniques from Data Mining, Machine Learning, Information Retrieval, Natural Language Processing, Social Network Analysis, and general Graph Theory.

Typical characteristics of link discovery problems are:

- Data is heterogeneous, arriving from multiple sources;
- Data and patterns sought include representations of people, organizations, objects, actions and events, each of which has its own set of attributes, and particular types of relations linking them;
- The structure may include temporal, spatial, organizational, and/or transactional patterns;
- A relatively low number of observations for each entity can be recorded and the overall sample is typically small relative to the size of the population;
- The data becomes available over time, so the timing of when to make a decision based on the analysis is a central issue.

LD problems are found in various areas such as homeland security, social network analysis, user modeling, fraud detection, and recommendation systems. The interdisciplinary nature of LD promotes a concerted effort from various researchers. The purpose of this workshop was to provide a forum to foster such interactions, discuss the new achievements and identify future research directions in link discovery.

The program of the workshop included an introductory talk by Jafar Adibi on link discovery issues, an invited talk by Ted Senator, thirteen contributed long and short papers and a discussion panel at the end of the program. The on-line proceedings of the workshop is available at: http://www.isi.edu/LinkKDD-05/ .

2. INVITED TALK

In his invited talk Ted Senator (DARPA) addressed two major questions in this field: what is LD and how difficult it is. He explained his point of view on LD and particularly he talked about LD role in anomaly detection, link representation and link analysis for alias resolution. In addition, he described some of the major operations such as path-finding, finding objects that are connected by path with specified properties, higher-order entity/pattern detection, discovering interesting/significant subgraphs instances, exploiting connections to classify/score entities according to specific attributes, identifying commonly occurring subgraph patterns and learning Interesting subgraph templates.

In the second part of his talk he provided a mathematical combinatorial model of LD to show the dimensions of LD problem. He used a set of jigsaw puzzles as a metaphor for LD problems by modeling the LD problem as 1 billion jigsaw puzzles, each with 1000 pieces and all painted over in a uniform color (until they are "detected"). Assume we see only 5% of the pieces (so we have only 50 billion pieces) . Hence the question is how hard is it to find "bad" puzzles? What if you had 1000 analysts? How would you divide up the pieces? Would it make it easier or harder? And finally how would they collaborate? He used such metaphor and its mathematical properties to study the LD problem .

He concluded his talk with the following points. 1) More data is not the answer since it makes the problem worse. 2) Collaboration has its limits and it depends strongly on parameters and assumptions. 3) Most useful technologies will be those that sort

data into categories with the highest likelihood of being linked, enable analysts to consider more combinations of data and facilitates collaboration between analysts. 4) Model can be used to examine alternative ways of organizing the analytical process. The talk ended with an overview of open issues and possible future works in this field.

3. Sessions

The contributed papers spanned a series of interesting topics related to link analysis and group detection. They are summarized below.

3.1 Session 1

The main focus of the morning session was on graph models for link detection and applications of textual analysis. The session consisted of eight papers.

The session kicked off with a paper from CMU which sparked an interesting discussion on analyzing a social network by using the structure of a learned Bayesian network. They applied their idea to co-authorship networks using Medline abstracts. The dependency links between authors learned by the Bayesian model gave light to new conclusions that were impossible to see by simply connecting co-authorships.

A group from UC Irvine followed with a very interesting framework for node-ranking time-varying social networks. With recent applications such as recovering social networks by analyzing emails or other related communications, important information is lost by ignoring the temporal relations of the input. In their framework, called EventRank, the authors model event sequences and provide a way of tracking rankings that change over time.

Badia and Kantardzic, from the University of Louisville, stepped back and reflected on the parameters and methodologies one should use in order to build a social network from raw data. This step is often assumed given (mostly because of its application-dependent nature) and consequently there has been a lack of discussion on the general steps one should take to build a network.

To close out the first part of the morning session, an extended abstract was presented from NYU on dynamic networks, which is a special type of graph where nodes have repeated evolving interactions. The authors proposed a framework for efficiently tuning and evaluating these graphs and showed empirical results on a fraud detection task that outperformed state of the art approaches.

After the break, we resumed with a fantastic talk by a group at U Mass who presented their Group-Topic model that jointly discovers groups in a network along with clusters of event topics that influence the interaction between entities. They applied their text-based approach to 16 years of U.S. Senate bills and 43 years of United Nations bills and outperformed current methods on both group and topic detection.

Next, Adamic and Glance presented their text analysis work on measuring the degree of interaction between political bloggers. Interestingly, political blogs were well-divided by party lines, as measured by their co-reference links, with conservatives being much more densely cross-linked. The study was performed on two months of postings on 40 “A-list” blogs as well as a one day snapshot of over 1000 blogs.

A University of Maryland group followed with an analysis of what they call *friendship-event* networks. These networks aim to capture the interaction between both a traditional friendship network between entities as well as a network of events (including event organizers and participants) interrelating entities (while capturing the temporal sequence of events). Although the work is in its preliminary stages, the authors proposed quantitative definitions for social capital, benefit received and benefit given to compare different event series.

Finally, the morning session ended with another CMU paper which looked at consolidating email addresses using a social network of co-occurring email-addresses extracted from web pages. On a real-world test set, preliminary results showed an accuracy of 15% when returning the top-10 most likely aliases for a given address.

3.2 Session 2

The main focus of the afternoon session was on discovering hidden groups, identifying important nodes and finding missing links in graphs. The session started with keynote talk delivered by Ted Senator described in Section 2.

The first paper of this session was a paper from a group at University of Illinois at Urbana-Champaign and University of Chicago on mining hidden communities in heterogeneous social networks. Their approach is founded on the fact that different relations have different importance with respect to a certain query. The authors proposed an optimal linear combination of these relations to find hidden communities.

A group from Indiana University followed with introducing GiveALink, a web service that mine semantic network of bookmarks for web search and recommendation. GiveALink uses semantic similarity measure for URLs that takes advantage both of the hierarchical structure of the bookmark files of individual users, and of collaborative filtering across users.

The close out the first part of the afternoon session, a work was presented by Shetty and Adibi from USC Information Sciences Institute on discovering important nodes in graphs. The authors used graph entropy to measure the find most influential nodes in a graph and evaluated their technique on publicly available Enron email database.

The second part of afternoon session started with another work on group detection by a group from CMU. They used a Bayesian model that uses a hierarchy of probabilistic assumptions about the way objects interact with one another in order to learn latent groups and the degree of membership of objects to discovered groups.

The last paper of the workshop was on discovering missing links in Wikipedia by Sisay Fissaha Adafre and Maarten de Rijke from University of Amsterdam. In their method first they compute a cluster of highly similar pages around a given page, and then they identify candidate links from those similar pages that might be missing on the given page.

Finally the workshop ended with a discussion among participants, authors and audience led by Jafar Adibi and Marko Grobelnik.

4. ACKNOWLEDGMENTS

We would like to thank our invited speaker Ted Senator, the authors of the workshop papers and all participants for contributing to the success of the workshop. Special thanks are due to the program committee for their work on reviewing the papers and all their support and help. We are also thankful to the members of KDD 2005 workshop committee for providing the opportunity to stage this event.

About the Authors:

Jafar Adibi is a Research Scientist at the University of Southern California's Information Sciences Institute. He has published over 25 refereed articles, achieved 3 international medals, and he holds two pending patents. His research interests include data mining for pattern recognition, social network analysis, link discovery mining blog space and study of networks dynamic behavior.

Patrick Pantel is currently an Assistant Research Professor and Research Scientist in the Natural Language Group at the USC Information Sciences Institute where he does research in semi-automatic ontology construction, text mining, knowledge acquisition, and machine learning. He is the recipient of various prestigious awards, including two national scholarships from the Natural Sciences and Engineering Research Council of Canada and the Izaak Walton Killam Memorial scholarship.

Marko Grobelnik has been associated with the Department of Knowledge Technologies of the Jozef Stefan Institute, Ljubljana, Slovenia since 1984. Most of his research work is connected with the study and development of Data Mining techniques and their application to different problems in economy, medicine, manufacturing, and game theory. His current research focuses on text and Web mining with particular interest in learning from text applied on large text data sets and the semantic Web.

Dunja Mladenic has been associated with the Department of Knowledge Technologies of the Jozef Stefan Institute, Ljubljana, Slovenia since 1987. She was at School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, as a visiting researcher in 1996-1997 and as a visiting faculty in 2000-2001. Most of her research work is connected with the study and development of machine learning and data mining techniques and their application on real-world problems from different areas recently focusing on text data, link analysis and semantic Web.

Program Committee

Lada Adamic, Hewlett Packard Laboratories
Jim Blythe, USC Information Sciences Institute
Hans Chalupsky, USC Information Sciences Institute
Tim Chklovski, USC Information Sciences Institute
Diane Cook, University of Texas at Arlington
Lise Getoor, University of Maryland
Antonio Gulli, AskJeeves/Teoma and University of Pisa
Jiawei Han, University of Illinois at Urbana-Champaign
Larry Holder, University of Texas at Arlington
David Jensen, University of Massachusetts Amherst
George Karypis, University of Minnesota
David Kempe, University of Southern California
Filippo Menczer, Indiana University
Rada Mihalcea, University of North Texas
Natasa Milic-Frayling, Microsoft Research
Michael Mitzenmacher, Harvard University
Andrew Moore, Carnegie Mellon University
Dragomir Radev, University of Michigan