KDD-2004 Workshop Report
Link Analysis and Group Detection (LinkKDD-2004)

Jafar Adibi, Hans Chalupsky
USC, Information Sciences Institute
Marina del Rey, CA
{adibi, hans}@isi.edu

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

Natasa Milic-Frayling
Microsoft Research Ltd,
Cambridge, United Kingdom
natasamf@microsoft.com

ABSTRACT

In this paper we provide a summary of the workshop on Link Analysis and Group Detection (LinkKDD-2004) held in conjunction with ACM SIGKDD 2004, on August 22, Seattle, Washington, USA. We report in details about the research issues addressed in the talks and the workshop.

Keywords

Data mining, Link Discovery, Link Analysis, Group Detection.

1. INTRODUCTION

Link Analysis and Link Discovery has been developed over the past 20 years in various fields including Discrete Mathematics (Graph Theory), Social Sciences (Social Network Analysis) and Computer Science (graph as a data structure). Recently this area has attracted a wider attention for its applicability in law enforcement investigations (e.g., terrorism), fraud detection (e.g., insurance, banking), WWW analysis (e.g., search engines, marketing), telecommunications (e.g., routers, traffic, connectivity), and similar. The common goal of this research is the development of techniques for mining large collections of data to extract valuable knowledge that may be present as hidden patterns or links among seemingly unrelated items. Successful applications will discover the hidden structure of organizations, relate groups, identify fraudulent behavior, model group activity and provide early detection of emerging threats.

In the departure from standard approaches is made clear in the following typical characteristics of the addressed problems and their representation:

- Data is heterogeneous, arriving from multiple sources.
- The data and patterns sought include representations of people, organizations, objects, actions and events. Each of these entities has its own set of attributes, and there are many types of relations that might exist between them.
- Unlike conventional data mining, in which nodes are variables and links are statistical relations among variables, nodes represent entities and links are relations amongst entities.
- Approaches assess the likelihood that an instance of a specific graph theoretic structure in the data matches a pattern of interest. The structure may include temporal, spatial, organizational, and/or transactional patterns. The addressed problems involve estimating a population based on a sample of data. Typically, 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.

Particularly interesting for the workshop were problems and issues that fall within the intersection of Link Analysis and fields such as Web and Text Mining, Relational Data Mining, and more general, Data Mining. Typical examples are in the area of trend analysis, community identification, Web user profiling, media clipping, marketing, etc., where Link Analysis complements other fields of research in order to achieve the higher information utilization. Another interesting scenario is extraction of information from unstructured data, representation of extracted data in the graphical form, and further analysis of the resulting graph structure to derive and discover new knowledge. The broader context of the workshop can be related in some respect to the areas of Data Mining, Machine Learning, Information Retrieval, Natural Language Processing, Social Networks Analysis, and the general Graph Theory. The purpose of this workshop was to provide a forum to foster such interactions, discuss the new achievements and identify future research directions. Hence, to highlight these efforts, we organized workshop on application of self-similarity and fractals in data mining held in conjunction with ACM SIGKDD 2002, the workshop on Link Analysis for Detecting Complex Behavior held in conjunction with ACM SIGKDD 2003, the workshop on Text Mining and Link Analysis at IJCAI-2003.

The program of the workshop included of introductory talk, an invited talk, twelve contributed papers and a panel at the end of the program. The on-line proceedings is available at: http://www.cs.cmu.edu/~dunja/LinkKDD2004.

The introduction by Jafar Adibi and Marko Grobelnik gave the definition of link analysis and group detection and at the end they opened the floor for the rest of the program.

2. INVITED TALK

In his invited talk Christos Faloutsos (CMU) talked about the interesting questions on graphs such as: What do real graphs look like? What properties of nodes, edges are important to model? and what local and global properties are important to measure?. In addition he also introduced some interesting questions regarding the real world applications such as: who is the best person/computer to immunize against a virus. Who is the best

customer to advertise? Who originated a raging rumor? and how to model and generate realistic graphs?

He also explained that we need to answer these questions to understand the following major issues in graphs. How will the Internet/Web look like next year? What is a realistic network topology, to try a new routing protocol or to study virus/rumor propagation, and immunization? How to get a 'good' sample of a network? Is this sub-graph / subcommunity / sub-network 'normal'? and what is normal ?

The talk continued with overview of some real world graphs. Real graphs are not random and we can see a power law distribution for in- and out- degree distributions. He explained how we can model a real world graph using such distribution. For instance virus propagation could be measured only by one factor which if each node infects its neighbors with more than factors a virus can survive and if it infects its neighbor with less than such factor it dies over time. He also mentioned that small-world navigation model to power-law distributions in which nodes know the degree of their neighbors; affects the sub-linear search time and facilitates the simultaneous search.

He concluded that real settings/graphs are skewed distributions and into 'mean' values is meaningless. We have to use slope of power law, instead. He also suggested to use of rank-frequency plot (a' la Zipf), NCDF, PDF in log-log and Correlation integral (= neighborhood function) as new tools for graph analysis. The talked was wrapped up by overview on open issues and possible future work in this field.

3. CONTRIBUTED PAPERS

The contributed papers spanned a series of interesting topics which is listed as following.

3.1 Graphs, Social Networks and Small World Phenomenon

Algorithms for semi-supervised and weakly supervised learning rely on the co-occurrence of features to propagate labels. Multiple-view algorithms, such as containing and coEM, use connectivity from the bipartite graph over two complementary feature sets to bootstrap models from a few initial labeled examples. In her talk Rosie Jones (Yahoo) showed how understanding algorithms and the data sets they operate on, in graph theoretic terms can explain some of the performance of these algorithms. She gave some empirical results over an application on learning semantic classes, as well as showing that Blum and Mitchell's data also displays small-world properties. She also showed that this enables them to predict the effectiveness of an algorithm on a task, based on the properties of the initial examples, as well as overall connectivity of the feature-co-occurrence graph.

Maitrayee Mukherjee (UTA) and Lawrence B. Holder (UTA) compared and contrast the salient features of illicit group information with legitimate group data. They discussed the graph-based knowledge discovery system SUBDUE, along with its unsupervised pattern discovery and supervised concept learning approaches. They described how SUBDUE, when run in unsupervised discovery mode, finds structural patterns embedded within social network data. They also illustrated how SUBDUE, in supervised mode, learns distinguishing patterns between legitimate and covert groups, based only on the communication activities of the group members. They also showed the effectiveness of SUBDUE in discovering patterns embedded within social network data and also in learning concepts to discriminate legitimate groups from illicit ones, based only on the communication patterns of the group members. In addition, they discussed how semi-structured data necessitated the advent of graph-based data mining. Since graphs lend a topological structure to data and since communication networks lend themselves to a graphical representation, they represented social networks as labeled graphs and then do graph-based data mining on them to discover novel and interesting concepts.

John Resig (RIT), Santosh Dawara (RIT), Christopher M. Homan (RIT) and Ankur Teredesai (RIT) illustrated how they extract social communities form instance massaging populations. In the analysis of large-scale social networks, a central problem is to discover how members of the network to be analyzed are related and how strongly any pair of members is connected. They introduce several such measures in their work which are obtained solely from the status logs of users where the status log of a user is a list of pairs of the form (time, state) and state is an element of a small set, such as {online, offline, busy, away}, and time is the time at which the member switched into that state. They believed status logs contain a great deal of structure since any pair of users can instant message each other only if they are both online at the same time. Hence, it seems reasonable to guess that any two users that are frequently online at the same time may in fact be frequently instant messaging each other. They built their work based on such hypothesis which forms the basis of each of their association measures.

A major challenge in link analysis is working with large real-world graphs from Internet connectivity to social networks. interesting way to compress network graphs. The size of these datasets presents a nearly insurmountable obstacle to understanding the essential character of the data. Anna C. Gilbert (AT&T , uMich) and Kirill Levchenko (UCSD) provided an interesting way to compress network graphs and to make it possible to understand \what the graph looks like. For instance which vertices and edges are important and what are the significant features in the graph. To do this they defined several compression schemes, including vertex similarity measures and vertex ranking. They presented a system for condensing large graphs using both auxiliary information such as geographic location and link type in the case of communication networks, as well as purely topological information. At the end they examineed the properties of these compression schemes, demonstrate their e ects on visualization, and explore what structural graph properties they preserve when applied to both synthetic and real-world networks.

3.2 Link Discovery Issues: Group Detection, Deduplication and Evaluation

Names of people, places, organizations, companies and other entities appear in the news all the time. To deduce information about these named entities is an interesting problem and has many practical applications. For example, if you are browsing the news it would be interesting to be able to click on a name and get information about the entity associated with that name. Hema Raghavan (uMass) , James Allan (uMass) and Andrew McCallum

(uMass) proposed a basic framework for representing entities. They explored the middle ground using a representation, in which questions about structured data may be posed and answered, but the complexities and task-specific restrictions of ontologies are avoided. An entity model is a language model or word distribution associated with an entity, such as a person, place or organization. Using these perentity language models, entities may be clustered, links may be detected or described with a short summary, entities may be collectively classified, and question answering may be performed.

One of the major concerns in link discovery and link analysis is similar entities presented by different ids. This problem is referred as alias detection, deduplication, record linkage and object consolidation. Deduplication also could be considered as a group detection task itself. Indrajit Bhattacharya (UMD) and Lise Getoor (UMD) proposed novel distance measures that take into account entity relationships for the purpose of clustering similar entities in linked environment. They proposed a distance measure based on a clearly defined generative probabilistic model which measures the chance in the probability of the observed data with regard to the generative model that occurs on merging two clusters and they showed how these measures can be useful for the important data mining tasks of data deduplication and group detection.

Shou-de Lin (USC/ISI) and Hand Chalupsky (USC/ISI) discussed another challenging problem within discovery systems; the verification of discovered results and evaluation of system performance. Considering that there is no direct way to verify the validity of a true discovery system, they addressed several issues of verifying machine discovery systems and proposed a set of indirect strategies one can adopt to assess the validity of discovered results. For instance they showed how recall and precision measures play a different role in machine discovery compared to machine learning. At the end and to test their idea they presented a set of novel link discovery tools to show how the proposed concepts can be applied to verify a real-world discovery system.

Many algorithms in link discovery and link analysis are deterministic and the constructed hypotheses are not qualified by probabilities. However, for nearly all applications the data available are sampled from a population. Hence, the discovered knowledge and implied hypotheses is probabilistic in nature and such uncertainty has to be measured. Due to the nature of the link discovery problems many of the current techniques and methods lack such measurement. Jafar Adibi (USC/ISI), Paul R. Cohen (USC/ISI) and Clayton T. Morrison (USC/ISI) addressed this methodological problem and provide a general method for measuring the confidence intervals of hypotheses generated by link discovery algorithms. They introduced bootstrap resampling method to measure the uncertainty in link discovery hypotheses.

3.3 Reasoning and Semantics

Many current systems for threat detection in use today provide only data visualization tools for manual link analysis leading to methods that do not scale to massive data sets. In their talk Nicholas J. Pioch (Alphatech), Daniel Hunter (Alphatech), James V. White (Alphatech), Amy Kao (Alphatech), Daniel Bostwick (Alphatech) and Eric K. Jones (Alphatech) presented the CADRE system (Continuous Analysis and Discovery from Relational Evidence) as a new alternative to addresses such deficiency by automating the link analysis process. CADRE combines an expressive knowledge representation of threat patterns with efficient, constraint-based abductive reasoning algorithms to automatically infer links and construct coherent threat hypotheses from structured data. A compact, factored representation of multiple hypotheses avoids redundant storage and enables scaling to large data sets. CADRE efficiently manages the growth of the hypotheses using probabilistic evaluation models and a consistency checking algorithm to prune unlikely hypotheses.

Jure Leskovec (Jozef Stefan Institute), Marko Grobelnik (Jozef Stefan Institute) and Natasa Milic-Frayling (Microsoft Research) presented an interesting work on summarizing document by creating a semantic graph of the original document and identifying the substructure of such a graph that can be used to extract sentences for a document summary. The whole techniques starts with syntactic analysis of the text and, for each sentence logical-form triples (subject–predicate–object) are extracted. At last step a cross-sentence pronoun resolution, co-reference resolution, and semantic normalization is applied to refine the set of triples and merge them into a semantic graph. They trained linear Support Vector Machine on the logical form triples to learn how to extract triples that belong to sentences in document summaries and the classifier is then used for automatic creation of document summaries of test data.

Link discovery is the process of identifying complex patterns from (multi)-relational data. The quality of link discovery outputs depends on the quality of the underlying data. Timothy Chklovski (USC/ISI) and Patrick Pantel (USC/ISI) discussed a method for refining multi-relational data in which they treat the data as a graph and apply global link analysis to refine the graph. They re-estimate the presence of a relation between a pair of nodes from the evidence provided by multiple indirect paths between the nodes. Despite each individual path being noisy, multiple indirect paths can provide sufficient evidence for adding, removing, or altering a relation between two nodes Our approach applies to a variety of relations: transitive symmetric, transitive asymmetric, and relations inducing equivalence classes. They presented their preliminary results on a semantic network called VERBOCEAN, which contains 22,306 relations between 3,477 verbs.

Recently information retrieval systems are evolving to question answering systems that return short text or find answers rather than just URLs. Most open-domain question answering systems depends strongly on elaborately designed hierarchies of question types. A question is first classified to a fixed type, and then hand-engineered rules associated with the type yield keywords and/or predictive annotations that are likely to match indexed answer passages. Soumen Chakrabarti (IIT Bombay) provided a data-driven approach, assisted by machine learning. He proposed a conditional exponential model over a pair of feature vectors, one derived from the question and the other derived from the candidate passage. In his model features are extracted using a lexical network and surface context as in named entity extraction, except that there is no direct supervision available in the form of fixed entity types and their examples. Candidate passages will be filtered using the exponential mode and substantial improvement in the mean rank at which the first answer is found is observed. The model parameters distill and reveal linguistic artifacts

coupling questions and their answers, which can be used for better annotation and indexing.

4. ACKNOWLEDGMENTS

We would like to thank workshop invited speaker Prof. Christos Faloutsos, authors of the workshop papers, panelists and participants for contributing to the success of the workshop. Special thanks are due to the program committee for their work on reviewing the papers, support and help. We are also thankful to the members of KDD 2004 workshop committee for providing the opportunity to stage this event.

About Authors:

Jafar Adibi is a Research Associate 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 and link discovery, fractals and self-similarity for advance application, and machine learning.

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.

Natasa Milic-Frayling is a researcher at the Microsoft Research lab in Cambridge, UK. She holds a PhD in Applied Mathematics from Carnegie Mellon University, Pittsburgh, PA. Her research interest ranges from identifying and solving specific information management problems that arise in highly distributed environments such as the WWW to answering various fundamental questions related to combining linguistic and statistical models in information management systems.

Program Committee