Introduction to the Special Issue on Link Mining

Lise Getoor
Department of Computer Science/UMIACS
University of Maryland
College Park, MD 20742
getoor@cs.umd.edu

Christopher P. Diehl
Applied Physics Laboratory
Johns Hopkins University
Laurel, MD 20723
Chris.Diehl@jhuapl.edu

1. INTRODUCTION

An emerging challenge for data mining is the problem of mining richly structured datasets, where the objects are linked in some way. Many real-world datasets describe a variety of entity types linked via multiple types of relations. These links provide additional context that can be helpful for many data mining tasks. Yet multi-relational data violates the traditional assumption of independent, identically distributed data instances that provides the basis for many statistical machine learning algorithms. Therefore, new approaches are needed that can exploit the dependencies across the attribute and link structure.

With linked data, new challenges emerge in traditional data mining tasks and new inference tasks are introduced. Classification of data objects, for example, is more complex with linked data, since the classification decisions are now dependent. This leads to the new challenge of collective classification, in which the classification decisions are resolved jointly. The introduction of links implies a host of new descriptive and predictive inference tasks based on the link structure. These range from capturing descriptive patterns in the link structure and predicting emerging links to discovery of clusters or communities based on attributes and link structure.

In recent years, there has been a surge of interest in this area, fueled largely by developments on the Internet. The Internet has provided a medium for hundreds of millions of people to connect with others, communicate ideas, express value judgements through the organization of those ideas, and engage in commerce. This has led to significant interest in challenges such as mining the web for information retrieval, online social networks for fostering relationships, and transaction patterns for recommending items of potential interest. Rapid advances in sensing, computation, and communication have spawned increasing volumes of multi-relational data. In domains such as biosurveillance, law enforcement, homeland security, and e-commerce, there is a continuing need for algorithms that can identify notable patterns in real-time transactional and event data. We expect that the demands for algorithms and processes to help understand and make predictions based on multi-relational data will only increase.

2. OVERVIEW OF THE ISSUE

This special issue brings together a range of articles covering a number of challenges in link mining. We begin with a survey article describing various link mining issues, including data processing and statistical models, and reviewing work in this area.

To support robust inference in a range of link mining tasks, there is a need for sound generative models of real-world networks. Airoldi and Carley examine the available sampling algorithms for a set of pure network topology types, such as scale-free, core-periphery and cellular networks, to assess their stability and separability. Different sampling algorithms for a given pure type should be stable in the sense that they yield ensembles with equivalent topological properties. For many types, Airoldi and Carley show that stability is not a given. They also investigate whether it is possible to reliably discriminate the pure types based on observed topological properties. Such separability is desirable in order to apply type-specific results from the literature about what phenomena to expect in a given network. They find that for certain pure types, it is difficult to discriminate between them, even with a rich set of topological descriptors.

An important dimension for many link mining domains is the dynamic temporal nature of the link structure. Several of the articles in this special issue address the dynamic nature of linked data. O'Madadhain et al. describe approaches for temporal link prediction and node ranking for event-based network data such as email traffic, telephone call data, and co-authorship events in publication networks. Pairwise link prediction is framed as a binary classification problem, whereby features extracted from past event history are used to train a probabilistic classifier via logistic regression. For event-based node ranking, the authors posit a series of properties that a ranking algorithm should satisfy, and argue that PageRank and related algorithms fail to provide several of these properties. They propose a new algorithm, EventRank, that exhibits the desired properties, and qualitatively assess the algorithm's performance by comparing individual ranks derived from email traffic to relative positions in the organizational hierarchy and anecdotal knowledge of employee activity. Sarkar and Moore present an approach for modeling dynamic social networks that is a generalization of a static latent space model presented in the literature. Assuming the latent space transitions are Markovian, the authors present an approach for performing link prediction during model estimation that is scalable to large link datasets. The authors also demonstrate how the modeling approach can be utilized to produce low-dimensional representations of the underlying relationships for visualization.

Link prediction and link discovery remain core challenges

in link mining. Rattigan and Jensen propose a new link mining task—anomalous link discovery (ALD)—as an alternative to link prediction. In contrast to predicting the emergence of new links in a dynamic graph, ALD focuses on the identification of observed links that are anomalous. Rattigan and Jensen's advocacy for ALD is based on the challenges of performing link prediction in dynamic graphs in which most vertex pairs are not linked. Such sparsely linked graphs complicate learning because of the massive skew in the ratio of linked versus unlinked vertex pairs. The authors qualitatively demonstrate the utility of simple link prediction models for the ALD task.

Several of the papers describe specific network structures, such as temporal event networks. The next paper by Sun et al. investigates the problems of relevance search and anomaly detection in bipartite graphs. Many link datasets can be described as bipartite graphs, such as actor-event networks. This paper presents algorithms that compute relevance scores for nodes in bipartite graphs. The relevance score is computed by defining a simple Markov transition model on this graph. Sun et al. then use these relevance scores to rank the other nodes, and show how to find anomalous connections between nodes in the same partition.

An interesting aspect of some linked data is the availability of additional semantic information. Ramakrishnan et al. address the challenge of identifying informative paths between two entities in an RDF graph. The general goal is to transform a given RDF graph into a weighted, undirected graph that allows one to leverage algorithms specifically designed to extract relevant paths from these graphs. The authors propose several heuristics for generating edge weights from the known semantics associated with the RDF data store. The authors then evaluate the quality of various combinations of the proposed heuristics.

The next article analyzes existing graph mining approaches. Ketkar et al. compare two representative approaches for graph-based and logic-based link mining: Subdue and CProgol. They define two types of concept complexity—structural and semantic complexity—and compare the systems' abilities to learn structurally and semantically complex concepts and to utilize background knowledge. Through the use of a weaker representation, Subdue is able to significantly outperform CProgol when learning structurally complex concepts. However, CProgol is more effective than Subdue at learning semantically complex concepts and utilizing background knowledge. The authors discuss the representational issues that lead to these outcomes and offer thoughts on how the strengths of both systems may be leveraged in future link mining systems.

As link mining capabilities improve, there is increasing concern about the misapplication of these techniques, leading to unwarranted violations of privacy. Sweeney discusses this challenge and proposes the concept of privacy-enhanced linking as a means to address these concerns. Since individuals are not aware of ongoing link mining activities examining their personal data, the burden is on the technology and ultimately by the technology developer to ensure that link mining algorithms provide an appropriate degree of privacy protection. Sweeney argues that algorithm developers should routinely assess the privacy protection offered by their algorithms and publish those assessments, along with traditional task-specific performance measures. Given that the success of such an approach is ultimately dependent on its adoption by the research community, it is important to commence with this discussion as the technology matures.

We end this issue with an article by Senator that examines general link mining challenges posed by various application domains. He provides a unique retrospective look at the development of the field over the last decade, based on his experience in the domains of fraud detection and intelligence analysis. As the research community continues to tackle the numerous outstanding problems, he argues that it is equally important to begin developing a comprehensive and flexible framework for composing link mining algorithms. Such a framework will be key to integrating point solutions into knowledge discovery applications tailored to specific analysis tasks.

3. SUMMARY

In recent years, significant progress has been made in defining and addressing the core link mining challenges, yet much work remains to be done in refining and combining various approaches and solutions. Through this collection of articles, we have attempted to highlight important aspects of the link mining landscape. We hope these articles provide interesting insights and catalyze new research directions.