ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 3, No.15 , May 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

The Application of Link Mining in Social Network Analysis

Zahra Zamani Alavijeh[1]
[1] **Department of engineering, University of Isfahan**
Isfahan, Iran
z.zamani@eng.ui.ac.ir

Abstract

Recently, link mining is becoming a very popular research area not only for data mining and web mining but also in the field of social network analysis. Many researches are focusing on developing new link mining techniques and algorithms, or devoting to improve traditional mining techniques for social network analysis. This paper discuss about challenges in the area of link mining and focusing on application requirements especially in the area of social network analysis. It discusses some ongoing challenges and suggests ideas that could be opportunities for solutions.

Keywords: *Link mining, Link Analysis, Social Network, Social Network Analysis.*

1. Introduction

The advent of online social networks has been one of the most exciting events in this decade. Many popular online social networks such as Twitter, LinkedIn, and Facebook have become increasingly popular. In addition, a number of multimedia networks such as Flickr have also seen an increasing level of popularity in recent years.

Many such social networks are extremely rich in content, and they typically contain a tremendous amount of content and linkage data which can be leveraged for analysis. The linkage data is essentially the graph structure of the social network and the communications between entities, whereas the content data contains the text, images and other multimedia data in the network [1]. The richness of this network provides unprecedented opportunities for data analytics in the context of social networks.

Data mining is a technique that has the ability to process and analyze large amount of data and by this to discover valuable information from the data. In recent year, due to the growth of social communications and social networking websites, data mining becomes a very important and powerful technique to process and analyze such large amount of data. But, Traditional data mining algorithms such as association rule mining, market basket analysis, and cluster analysis commonly attempt to find patterns in a dataset with independent instances of a single relation [2]. So applying traditional procedures, which assume that instances are independent, for heterogeneous datasets, like social network data, can lead to inappropriate conclusions because the data comprising social networks tend to be heterogeneous, multi relational, and semi-structured. As a result, a new field of research has emerged called link mining.

Link mining refers to data mining techniques that explicitly consider these links when building predictive or descriptive models of the linked data [2]. Commonly addressed link mining tasks include object ranking, group detection, collective classification, link prediction and sub graph discovery. While network analysis has been studied in depth in particular areas such as social network analysis, hypertext mining, and web analysis, only recently has there been a cross-fertilization of ideas among these different communities.

This paper discuss about challenges in the area of link mining, focusing on application requirements and how they have and have not yet been addressed, especially in the area of social network analysis. Also, we try to survey challenges and open research issues for each task.

The remainder of this paper proceeds as follows. In Section 2, we discuss about link mining and its tasks. Then we have a survey in social networks and their analysis in Section 3. We detail tasks of link mining on social networks in Section 4. Our conclusions are given in Section 6.

2. Link Mining

Link mining is a newly emerging research area that is at the intersection of the work in link analysis, hypertext and web mining, relational learning and graph mining [3]. Link mining can be seen as the task of applying data mining techniques on networks, while explicitly considering and emphasizing on links between social network actors.

Links have more generically relationships, among data instances are ubiquitous. These links often exhibit patterns that can indicate properties of the data instances such as the importance, rank, or category of the object. In some cases, not all links will be observed. Therefore, we may be interested in predicting the existence of links between instances. In other domains, where the links are evolving over time, our goal may be to predict whether a link will exist in the future, given the previously observed links. By taking links into account, more complex patterns arise as well. This leads to other challenges focused on discovering substructures, such as communities, groups, or common sub graphs. Traditional data mining algorithms such as association rule mining, market basket analysis, and cluster analysis commonly attempt to find patterns in a dataset characterized by a collection of independent instances of a single relation.

By considering links between objects, more information is made available to the mining process. Link mining tasks are broadly categorized into following tasks [4]. They are:

1. Object-Related Tasks
(a) Link-Based Object Ranking
(b) Link-Based Object Classification
(c) Object Clustering (Group Detection)
(d) Object Identification (Entity Resolution)

2. Link-Related Tasks
(a) Link Prediction
(b) Link type Prediction

3. Graph-Related Tasks
(a) Sub graph Discovery
(b) Graph Classification
(c) Generative Models for Graphs

In the rest of the paper, we will discuss about these tasks and their application in social network analysis.

3. Social network analysis

From the point of view of data mining, a social network is a heterogeneous and multi relational dataset represented by a graph. The graph is typically very large, with nodes corresponding to objects and edges corresponding to links representing relationships or interactions between objects. Both nodes and links have attributes. Links can be one directional and are not required to be binary [3].
In practice, social networks offer to web users new interesting means and ways to connect, communicate, and share information with other members within their platforms. In theory, these social networks are made of several components, can hold different types of data, and have various representations.

Social network analysis (SNA) is a process of quantitative and qualitative analysis of a social network. The focus of SNA is to study social relationships between actors rather than only using their attributes. Thus, the availability of social networks data, actors, and links used to model social networks is fundamental for analyzing these networks [5]. Recently, SNA has been used in different application domains such as mobile networks, co-authorship and citation networks [6] and online social networks [7].
In the following, we detail the main social network analysis measures. The definition of these measures can be important in understanding remind of this paper.

4. Link mining in social networks

In this section, we describe link mining tasks that addressed in section 2 and we discuss about their application on social networks.

4.1 Object-Related Tasks

4.1.1 Link-based object ranking

Perhaps the most well known link mining task is that of link-based object ranking (LBR), which is a primary focus of the link analysis community [3]. The objective of LBR is to exploit the link structure of a graph to order or prioritize the set of objects within the graph. Much of this research focuses on graphs with a single object type and a single link type.

The objective of link-based object ranking is to prioritize corresponding actors based on their importance with respect to a chosen measure. In link mining, centrality measures (degree, closeness, betweenness, eigenvector, etc.) that exploit the network structure are used to rank the actors. There is a wide range of real-world applications that can use link-based object ranking. In the area of web information retrieval, the most common approaches to LBR, are PageRank and HITS algorithms [3].

In the domain of social network analysis (SNA), LBR is a core analysis task. The objective is to rank order individuals in a given social network in terms of a measure of their importance, referred to as centrality. Measures of centrality have been the subject of research in the SNA community for decades. These measures characterize some aspect of the local or global network structure as seen from a given individual's position in the network. They range in complexity from local measures such as degree centrality, which is simply the vertex degree, to global measures such as eigenvector/power centrality, which use spectral

methods to characterize the importance of individuals based on their connectedness to other important individuals.

Dealing with dynamic networks where relational data can change over time, like social networks, and with large networks characterized by uncertainty is regarded as an important research direction in the context of dynamic networks [8]. One challenge in these large and dynamic networks is to maintain updated the dynamic relational data. Another challenge is to be able to detect interesting trends like nodes whose importance may increase rapidly. Most of the current algorithms are designed for static networks [9]. However, it is interesting and challenging to develop new algorithms for object ranking within dynamic networks.

4.1.2 Link-Based Object Classification

Among several object related tasks, the link-based object classification task classifies nodes of a network to a finite set of categories. This type of classification is not only based on nodes' attributes but also on their links to other nodes and on the attributes of these linked nodes [3].

In the social network, some of the nodes may be labeled, and it may be desirable to use the attribute and structural information in the social network in order to propagate these labels. For example, in a marketing application, certain nodes may be known to be interested in a particular product, and it may be desirable to use the attribute and structural information in the network in order to learn other nodes which may also be interested in the same product. Social networks also contain rich information about the content and structure of the network, which may be leveraged for this purpose. For example, when two nodes in a social network are linked together, it is likely that the node labels are correlated as Well. Therefore, the linkage structure can be used in order to propagate the labels among the different nodes. Content and attributes can be used in order to further improve the quality of classification.

A key challenge for link-based node classification algorithms is to exploit the correlation between the nodes. For instance, to be able to infer the category of a web page 1 (WP1), the category of web page 2 (WP2) must be used. Similarly, to assign a category to WP2, the category of WP1 must be known. In fact, most of the real-world data networks are mainly heterogeneous datasets with correlated linked nodes [10]. To provide a solution and improve the classification results, it is crucial to design new link-based node classification algorithms that jointly classify nodes using collective classification.

4.1.3 Link-based Object Clustering

Object clustering, also called group detection, is another well-studied link mining task. Its objective is to identify similar nodes and cluster them together without having predefined labeled categories [1]. Any two nodes, members of the same cluster, are more similar to each other than to any node in other cluster. They represent communities where the level of interaction or communication (emails, messages, collaborations, etc.) between actors of the same cluster is higher than with any actor in other clusters.

Object clustering in social networks, allow the identification of different communities where members Share the same type of activities, are interested in the same hobbies, or seek the same kind of services. Furthermore, link-based object clustering approaches enable to create clusters for persons who share different relationship types (family members, classmates, co-workers, etc.) [10] Partitioning social networks into sets of individuals, called positions, which exhibit similar sets of links to others in the network. A similarity measure is defined between link sets and agglomerative clustering is used to identify the positions. Spectral graph partitioning methods address the group detection problem by identifying an approximately minimal set of links to remove from the graph to achieve a given number of groups [11].

Typically, clustering similar nodes together takes into account both the attributes of nodes and the links among them. Missing attributes, noisy data, and dynamic networks can make this task more complex. It is therefore necessary to overcome these issues when clustering nodes. However, it is infeasible to infer the presence of implicit groups by analyzing whole networks since enormous amount of data is currently available on social networks. Consequently, proposing new approaches that scale up node clustering to large networks is an interesting research topic.

4.1.4 Link-based object Identification

Link-based object identification, or entity resolution, is a topic that has received a lot of attention in the literature. It aims to identify objects in datasets that have different identifiers while referring to the same real-world entity. In this case, these nodes form a matched entity pair.

Social network users publish different information on several social network sites either purposely or accidently. Each profile, aside from being created on different sites, may only contain a portion of the complete profile information of the user. Aggregating data of co-referent

users can be regarded as the task of combining and merging information in order to automatically create a complete representation of a user profile from her disparate profiles on different social sites.

Entity resolution is associated with many challenges: 1) it is computationally expensive since comparing all pairs of items is not always feasible, 2) its accuracy depends on the available attributes to use, and 3) it depends on the typographical errors, abbreviations, and inverted words order which can make the task more complex.

4.2 Link-Related Tasks

4.2.1 Link Prediction

Link prediction, or link existence prediction, is the task of inferring the existence of a link between two nodes, based on the properties of the nodes [3]. Examples include predicting whether there will be a link between two Web pages, and whether a paper will cite another paper. While link prediction in static networks aims at inferring missing links and facilitating the task of creating links, link prediction in dynamic networks consists of predicting the snapshot of links at a future time.

Social networks are dynamic. New links appear, indicating new interactions between objects. In the link prediction problem, we are given a snapshot of a social network at time t and wish to predict the edges that will be added to the network during the interval from time *t* to a given future time, *t+1*. In essence, we seek to uncover the extent to which the evolution of a social network can be modeled using features intrinsic to the model itself. As an example, consider a social network of coauthorship among scientists. Intuitively, we may predict that two scientists who are "close" in the network may be likely to collaborate in the future. Hence, link prediction can be thought of as a contribution to the study of social network evolution models.

Recently, many good works were done on this field. For example in [12], authors develop approaches to link prediction based on measures for analyzing the "proximity" of nodes in a network. In another work, Backstrom and Leskovec presented an algorithm based on *Supervised Random Walks* that naturally combines the information from the network structure with node and edge level attributes [13].

The link prediction and link recommendation problems are challenging from at least two points of view. First, real networks are extremely sparse, i.e., nodes have connections to only a very small fraction of all nodes in the network.

The second challenge is more subtle; to what extent can the links of the social network be modeled using the features intrinsic to the network itself? Similarly, how do characteristics of users (*e.g.*, age, gender, home town) interact with the creation of new edges? Thus the question is how network and node features interact in the creation of new links [13].

4.2.2 Link Type Prediction

This predicts the type or purpose of a link, based on properties of the objects involved. Given epidemiological data, for instance, we may try to predict whether two people who know each other are family members, coworkers, or acquaintances. In another example, we may want to predict whether there is an advisor-advisee relationship between two coauthors. Given Web page data, we can try to predict whether a link on a page is an advertising link or a navigational link.

Unlike link prediction, where the aim is to predict the existence of a link between two nodes at a particular time, link type prediction tries to identify the type of an existing link. Here, it is assumed that we know that a link already exists between the two nodes.

Today's social network users can be connected to different types of contacts such as colleagues, relatives, friends, etc. In fact, contacts on social networks and real-life contacts are increasingly interwoven [14]. Relationship discovery has found considerable interest recently due to the rowing number of social network users and the increasing need to analyze social interactions.

In fact, identifying social relationship types is useful in many situations. Nowadays, users get connected to a big number of contacts, managing large contact lists is a tedious task. Organizing contacts based on the relevant relationship types can be useful for different situations where targeted social content sharing and filtering are needed.

Applying link type prediction is of great importance to avoid the daunting task of manually labeling the links. In addition, knowing links' types could be critical toward better node classification for many tasks, and crucial for various privacy applications.

4.3 Graph-Related Tasks

4.3.1 Subgraph Discovery

Subgraph discovery is a link mining task that attempts to detect similar substructures in pairs of graphs. It is one of well-studied topics related to graph mining where graphs

are natural representations of diverse complex structures. Frequent subgraph discovery approaches are successfully used on graphs where each actor's label is used only onceper graph [3]. This type of graphs is called *relational graphs*.

Two key emerging challenges for subgraph discovery address the following problems: 1) the size of networks where an enormous amount of data is to be processed: dealing with the whole set of links and nodes is computationally expensive, and 2) networks that change dynamically over time, also known as streaming networks: as new links are incrementally received, subgraph discovery task is supposed to process the received data in real time.

4.3.2 Graph Classification

Unlike link-based object classification, which attempts to label nodes in a graph, graph classification is a supervised learning problem in which the goal is to categorize an entire graph as a positive or negative instance of a concept. This is one of the earliest tasks addressed within the context of applying machine learning and data mining techniques to graph data. Graph classification does not typically require collective inference, as is needed for classifying objects and edges, because the graphs are generally assumed to be independently generated [3].

4.3.3 Generative Models for Graphs

Generative models for graphs try to understand the characteristics of networks, to identify the mechanisms of networks growth and evolution, and to generate networks with realistic properties given few parameters. Studying generative models for graphs is becoming increasingly important, in particular when temporal metrics are considered i.e., the network changes over time. In particular, one of the main characteristics of social network is that social relationships evolve between users, which mean that new links can appear and others can disappear. Therefore, it is interesting to apply generative methods to understand and to reveal the future trend of network evolution.

5. Conclusions

Today most of domains, like social networks, are described as a linked collection or network of interrelated heterogeneous objects. Link mining is an emerging area within data mining that is focused on finding patterns in data by exploiting and explicitly modeling the links among the data instances.

On the other hand, in recent years the growth of social networking platforms has been phenomenal. Due to the growth of social communications and social networking websites, data mining becomes a very important and powerful technique to process and analyze such large amount of data.

Understanding the characteristics of social networks, or more generally networks, has been among the most studied topics in the social network analysis community. Two main research areas have focused on studying social networks: 1) social network analysis, and 2) link mining. In this paper, we provided a review of a number of link mining tasks and their applications on social network analysis. We find that each task has challenges yet, and there are many open research issues for future works.

References

[1] D.M. Boyd and N.B. Ellison, "Social Network Sites: Definition, History, and Scholarship," Journal of Computer-Mediated Communication, 2008, Vol.13, pp. 210–230.

[2] L. Getoor **,"** Link mining: a new data mining challenge," In ACM SIGKDD Explorations Newsletter ,2003, Vol. 5, Issue 1, pp. 84-89.

[3] L. Getoor and C. P. Diehl, "Link Mining: A Survey," In ACM SIGKDD Explorations Newsletter, 2005, Vol. 7, Issue 2, pp. 3-12.

[4] K.Srinivas, L.Kiran, and A.Govardhan," A Theoretical Approach to Link Mining for personalization," International Journal of Computer Science Issues, 2010, Vol. 7, Issue 3, pp.41-44.

[5] J. P. Scott, Social Network Analysis: A Handbook. SAGE Publications, Jan. 2000.

[6] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," in Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. pp.1055–1060, 2008.

[7] A. Mislove, M. Marcon, K. P. Gummadi, and B. Bhattacharjee,"Measurement and analysis of online social networks," Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007, pp. 29–42.

[8] T. Carpenter, G. Karakostas, and D. Shallcross, "Practical issues and algorithms for analyzing terrorist networks," Proceedings of the Western Simulation MultiConference, 2002.

[9] J. Scripps, R. Nussbaum, P.-N. Tan, and A.-H. Esfahanian, "Link-based network mining," in Structural Analysis of Complex Networks , 2011, pp. 403–419.

[10] G. Barbier and H. Liu, "Data mining in social media," in Social Network Data Analytics, Ed. Springer US, 2011, pp.327–352.

[11] M. E. J. Newman.,"Detecting community structure in networks." European Physical Journal, 2004, pp. 321-330.

[12] D. Liben and J. Kleinberg, " The Link Prediction Problem for Social Networks," Journal of the American Society for Information Science and Technology, 2007, Vol.58, Issue 7, pp. 1019–1031.

[13] L. Backstrom and J. Leskovec," Supervised Random Walks: Predicting and Recommending Links in Social Networks," In Proc. of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11), February 2011.

[14] E. Raad, R. Chbeir, and A. Dipanda, "Discovering relationship types between users using profiles and shared photos in a social network," Multimedia Tools and Applications, 2011, pp.1–30.

Zahra Zamani Alavijeh was born in Isfahan, Iran in 1986. She received a B.S. degree in Computer Engineering from Isfahan University of technology, Iran, and M.S. degree in Computer Engineering from University of Isfahan, Iran. Her research interests are in the areas of Data mining, Link mining and social networks.