Link Mining in Social Networks

Zow Afshan¹ Mr. Sauod Sarwar²
¹M.Tech Scholar, ²Senior Professor, H.O.D (Computer Science)
¹,² AL-FALAH School of Engineering and Technology, Dhauj, Faridabad, India

Abstract— In data mining a social network is a diversion content and multi-relational data set represented by a graph. A graph is very large, with nodes corresponding to objects and edges corresponding to links representing relationships or interactions between objects. Both nodes and links have attributes. Objects may have class labels. Links would be one-directional and are not required to be binary. A key challenge for data mining is tackling the problem of mining richly structured datasets, where the object is linked in some way. Links among the objects can find certain patterns, which may be helpful for many data mining tasks and are typically hard to capture with traditional statistical models. Recently it has been a surge of interest in this area, fueled largely by interest in web and hypertext mining, but also by interest in mining social networks, security and law enforcement data, bibliographic citations and epidemiological records. Traditional data mining tasks 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. This is consistent with the classical statistical inference problem of trying to identify a model given a random sample from a common underlying distribution. A key challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets. Naively applying traditional statistical inference procedures, which assume that instances are independent, can lead to inappropriate conclusions. Care must be taken that potential correlations due to links are handled appropriately. In fact, record linkage is knowledge that should be exploited. Clearly, this is information that can be used to improve the predictive accuracy of the learned models: attributes of linked objects are often correlated and links are more likely to exist between objects that have some commonality. 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 inductive logic programming and graph mining. Link mining is an instance of multi-relational data mining (in its broadest sense); however, we use the term link mining to put an additional emphasis on the links moving them up to first-class citizens in the data analysis endeavor. Link mining encompasses a range of tasks including descriptive and predictive modeling. Both classification and clustering in linked relational domains require new data mining algorithms. But with the introduction of links, new tasks also come to light. Examples include predicting the numbers of links, predicting the type of link between two objects, inferring the existence of a link, inferring the identity of an object, finding co references, and discovering sub graph patterns.

Keywords: Social Network, Mining On Social Network, Link Mining Tasks, Statistical Models For Link Mining, Link Cardinality Estimation.

I. INTRODUCTION

A. What is Social Network?

A social network is a diversion content and multi-relational data set represented by a graph. The graph is typically very large, with nodes similar to objects and edges similar to links representing relationships or influence between objects. Both nodes and links have features. These typically contain from tens of thousands to millions of nodes. Often, a great deal of information is able to be used at the level of individual nodes. The availability of powerful computers has made it possible to probe the structure of networks. Searching social networks can help us better understand how we can reach other people.

B. Densification power law

Earlier, it was believed that as a network develop gradually, the number of degrees grows linearly in the number of nodes. This was known as the constant average degree assumption. However, extensive experiments have shown that, on the contrary, networks become increasingly dense over time with the average degree increasing (from now, the number of edges growing super linearly in the number of nodes). The densification follows the densification power law (or growth power law), which states e(t) $\alpha$ n(t)^a; where e(t) and n(t), individually, represent the number of edges and nodes of the graph at time t, and the exponent a generally lies strictly between 1 and 2. Note that if a = 1, this corresponds to constant average degree over time, whereas a = 2 corresponds to an extremely dense graph where each node have edges to a constant fraction of all nodes.

C. Shrinking diameter:

There have been experimentally shown that the effective diameter behave particularly to decrease as the network grows. This contradicts an earlier belief that the diameter slowly increases as a function of network size. As an intuitive example, suppose a citation network, where nodes are papers and a citation from one paper to another is indicated by a directed edge. The out-links of a node, b (representing the papers cited by b), are —frozen at the moment it joins the graph. The decreasing distances between pairs of nodes as a result appears to be of follow after papers acting as bridges by citing earlier papers from other areas.

D. Heavy-tailed out-degree and in-degree share :

The number of out-degrees for a node tends to go along a
Heavy-tailed distribution by noticing the power law, $1 = n$, where $n$ is the rank of the node in the order of decreasing out-degrees and typically, $0 < a < 2$. The smaller the value of $a$, the heavier the tail. The fact observed to exist is represented in the involved preference attachment model, where each new node attaches to an existing network by a constant number of out links, following a —rich-get richer rule. There degrees follow a heavy-tailed distribution, in spite of the fact tends to be more skewed than the out-degrees distribution. A Forest Fire model for graph generation was proposed, which captures these characteristics of graph evolution over time. It is based on the notion that new nodes attach to the network by —burning through existing edges in epidemic fashion. It uses two parameters, forward burning probability, $k$, and backward burning ratio, $l$, which are described below. Suppose a new node, $b$, arrives at time $t$. It attaches to $Gt$, the graph constructed so far, in the following steps:

1. It chooses an ambassador node, $w$, at random, and forms a link to $w$.
2. It selects $x$ links incident to $w$, where $x$ is a random number that is binomially distributed with mean $(1 - p)^x$.
3. It chooses from out-links and in-links of $w$ but selects in-links with probability $r$ times lower than out-links. Let $w_1; w_2; \ldots; w_x$ denote the nodes at the other end of the selected edges.
4. Our new node, $v$, forms out-links to $w_1; w_2; \ldots; w_x$ and then applies step 2 recursively to each of $w_1; w_2; \ldots; w_x$. Nodes cannot be visited a second time so as to prevent the construction from cycling. The process continues until it dies out.

II. MINING ON SOCIAL NETWORKS:

Mining on social networks, namely, link prediction, mining customer networks for viral marketing, mining newsgroups using networks, and community mining from multirelational networks. Other example include characteristic sub graph detection and mining link structures on the Web. Pointers to research on link-based classification and clustering are given in the bibliographic notes and exercises.

A. Link Mining:

Traditional data mining tasks 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. The consistent with the classical statistical inference problem of trying to identify a model given a random sample from a common underlying distribution. A key challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets. Naturally applying traditional statistical inference procedures, which assumes that instances are independent, can lead to inappropriate conclusions. Care must be taken that potential correlations due to links are handled appropriately. In fact, record linkage is knowledge that should be exploited. Clearly, this is information that can be used to improve the predictive accuracy of the learned models: attributes of linked objects are often correlated and links are more likely to exist between objects that have some commonality. 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 inductive logic programming and graph mining. Link mining is an instance of multi-relational data mining (in its broadest sense); however, we use the term link mining to put an additional emphasis on the links moving them up to first-class citizens in the data analysis endeavor. Link mining encompasses a range of tasks including descriptive and predictive modeling. Both classification and clustering in linked relational domains require new data mining algorithms. But with the introduction of links, new tasks also come to light. Examples include predicting the numbers of links, stating the type of link between two objects, inferring the existence of a link, implying the identity of an object, finding co-appearances, and discovering sub graph patterns.

III. LINK MINING TASKS:

As discussed above, link mining puts a new turn to some classic data mining tasks, and also has new problems. Here we make available for use a (non exhaustive) list of possible tasks. We serve an example to each of them using the following domains as motivations:

Web page collection: In a web page collection, the objects are web pages, and links are in-links, out-links and co citation. Links (two pages that are both linked to by the same page). Attributes have HTML tags, word appearances and anchor text. Bibliographic domain: In a bibliographic domain, the objects include papers, authors, institutions, journals and conferences.

A. Link Based Categorization

The most straightforward upgrading of a classic data mining task to linked domains is link-based categorization. In link based categorized, we are interested in predicting the category of an object, based not just on its attributes, but on the links it involving in, and on attributes of objects linked by some path of edges. An example of link-based categorization that has accepted a good amount of attention is web-page categorization. In this problem, the goal predicts the category of a web page based on words on the page, links between pages, anchor text and other attributes of the pages and the links. In the bibliographic domain, an example of link-based classification is predicting the category of a paper, based on its citations.

B. Link based Cluster interpret:

The aim in cluster interpret is to find naturally occurring subclasses. Which is done by segmenting the data into groups, where objects in a group are of same kind to each other and are very dissimilar from objects in different groups. Unlike classification, clustering is not done and can be applied to discover hidden patterns from data. This makes it a perfect technique for applications such as scientific data exploration, information retrieval, computational biology, web log analysis, criminal analysis and many others. There has been extensive research work on clustering in areas such as pattern recognition, statistics and machine learning. Hierarchical agglomerative clustering (HAC) and k-means are two of the most common clustering algorithms. Probabilistic model based clustering is gaining increasing popularity. All of these algorithms assume that
each object is described by a fixed length attribute-value vector. In the case of clustering linked data, even the definition of an element in a cluster is open to interpretation. We can cluster individual objects, collections of linked objects, or some other sub graph of the original. How do we compare the similarity of two of these elements or sub graphs, with potentially different structures? As this may necessitate tests for graph isomorphism, things will quickly become intractable. There has been surprisingly little work done on this type of link mining. Examples of clustering in web page collections range from finding hubs (pages that point to lots of pages of the same category) to identifying mirror sites. Examples of clustering in the bibliographic domain include finding groups of authors that commonly publish together, and discovering research areas, based on common citations and common publication venues and discovering. An example of clustering in the epidemiology domain is finding patients with similar sets of contacts or diseases with similar transmission patterns. Next, we turn to some more specific tasks that arise in link mining. These can often be seen as special cases of link-based classification or link-based cluster analysis.

C. Identifying Link Type:
There is a wide range of tasks related to predicting the existence of links. One of the simplest is predicting the type of link between two entities. For example, we may be trying to predict whether two people who know each other are family members, coworkers, or acquaintances, or whether there is an adviser–advisee relationship between two coauthors. The link type may be modeled in different ways. In some instances, the link type may simply be an attribute of the link. In this case, we may know the existence of a link between two entities, and we are simply interested in predicting its type. In our first example, perhaps we know there is some connection between two people, and we must predict whether it is a familial relation, a coworker relation or acquaintance relation. In other instances, there may be different kinds of links. These may be different potential relationships between entities; in the second example, there are two possible relationships: a co-author relationship and an adviser–advisee relationship. We may want to make inferences about the existence of one kind of link, having observed another type of link. A closely related task is predicting the purpose of a link. In a web page collection, the links between pages occur for different reasons. At the coarsest grain, links may be for navigational purposes or for advertising; it may be quite useful to distinguish between the two. The links may also indicate different relationships; the purpose of a link may be to refer to a professor’s students, a student’s friends, or a course’s assignments.

D. Predicting Link Strength:
Links may also have weights deals with them. In a web page collection, the weight may be interpreted as the authoritativeness of the incoming link, or its page rank. In an epidemiological domain, the strength of a link between people may be an indication of the length of their exposure.

E. Link Cardinality:
There are many practical conclusions by reasoning that involve predicting the number of links between objects. The number of links is often a proxy for some more meaningful property whose semantics depend on the particular domain:
- In a bibliographic domain, predicting the number of citations of a paper is an indication of the impact of a paper—papers with more citations are more likely to be seminal.
- In a web collection, predicting the number of links to a page is an indication of its authoritativeness; predicting the number of links from a page is an indication that the page is a hub. The page rank measure is also clearly related to the number of links.
- In an epidemiological setting, predicting the number of links between a patient and people with whom they have been in contact (their contacts) is an indication of the potential for disease transmission; predicting the number of links between a particular disease strain and people infected by it is an indication of the strain’s virulence. Note that link counts can be generalized to paths. A count of the number of paths between two objects may be significant.

F. Record Linkage:
Another significant concept in link mining is to identify uncertainty. In many practical approaches, such as information extraction, duplication elimination and citation matching, objects may not have unique identifiers. The demanding task is to determine when two similar looking items in fact refer to the same object. Thus problem have been studied in statistics under the umbrella of record linkage; it has also been studied in the database community for the task of duplicate elimination. In the link mining setting, it is significant to take into account not just the similarity of objects based on their attributes, but also based on their links. In the bibliographic setting, this means taking into account the citations of a paper, note that as matches are identified, new matches may become obvious.

IV. STATISTICAL MODELS FOR LINK MINING:
Given the above collection of tasks, there are some unique challenges to applying statistical modeling techniques. Here, we identify several; see also other papers in this volume, and papers in several recent workshops on learning statistical models from relational data.

A. Logical vs. Statistical Dependencies
The first challenge in link mining and multi-relational data mining is coherently handling two different types of dependence structures:
- link structure — The logical relationships between objects
- probabilistic dependency — The statistical relationship between attributes of objects. Typically we limit the probabilistic dependence to be among objects that are logically related. In learning statistical models for multi-relational data, we must not only search over probabilistic dependencies, as is standard in any type of statistical model selection problem, but potentially we must search over the different possible logical relationships between objects. This search over logical relationships has been a focus of research in inductive logic programming, and the methods and machinery developed in this
community should be used to tackle this problem.

**B. Feature Construction:**

A second challenge is feature construction in the multi-relational setting. The attributes of an object provide a basic description of the object. Traditional classification algorithms are based on these types of object features. In a link-based approach, it may also make sense to use attributes of linked objects. Further, if the links themselves have attributes, these may also be used. This is the idea of the following proposition. However, as others have noted, simply flat modeling the relational neighborhood around an object can be problematic. Several have noted that in hypertext domains, simply including words from neighboring pages degrades classification performance. A further issue is how to deal appropriately with relationships that are not one-to-one. In this case, it may be appropriate to compute aggregate features over the set of related objects. We have found this works well for learning probabilistic relational models, but this approach may not always be appropriate.

**C. Collective Classification**

Using a link-based statistical model for classification. Two steps: Model construction and feature extraction using learned model a third challenge is feature construction using a learned model. A learned link-based model specifies a distribution over link and content attributes, which may be correlated based on the links between them. Intuitively, for linked objects, updating the category of one object can influence our inference about the categories of its linked neighbors. This requires a more complex classification algorithm than for a propositional learner. Iterative classification algorithms have been offered for hypertext categorization and for relational learning. The general approach of iterative classification has been studied in numerous fields, including relaxation-labeling in computer vision, inference in Markov random fields and loopy belief propagation in Bayesian networks. Some approaches make assumptions about the influence of the neighbor’s categories (such as that linked objects have similar categories); we believe it is important to learn how the link distribution affects the category. An example, this allows us to learn the notion of hubs — e.g., a computer science department homepage is likely to point to a lot of professor homepages.

Labeled & Unlabeled Data

In link-based domains, unlabeled data supply three sources of information: Helps us infer object attribute distribution Links between unlabeled data allow us to make use of attributes of linked objects Links between labeled data and unlabeled data (training data and test data) help us make more accurate inferences.

**D. Effective Uses of Unlabeled Data:**

Recently there’s increased interest in learning using a mix of labeled and unlabeled data. General approaches include semi-supervised learning, co-training and transductive inference. There are some the unique ways in which unlabeled data can be used to improve classification performance in relational domains:

- As in the case of the classical machine learning framework, where there are no links among the data, unlabeled data may help us learn the distribution over object descriptions.
- Links among the unlabeled data (or test set) may provide information that can help with classification.
- Links between the labeled training data and unlabeled (test) data induce dependencies that should not be ignored.

**E. Link Prediction:**

A fifth challenge is link discovery, or predicting the existence of links between objects. A range of the tasks that we have been discussed fall under the category of link prediction. A difficulty here is that the prior probability of a link among any set of individuals is typically quite low. While we have had some success with simple probabilistic models of link existence, we believe this is an area where there is much research to be done. A further challenge is the discovery of common relational patterns or sub graphs; some progress has been made in this area; however, this is an inherently difficult problem. The data comprising social networks tend to be heterogeneous, multirelational, and semi structured. As a result, a new field of research has emerged called link mining.

1. **Link-based object classification.** In traditional classification methods, objects are classified based on the attributes that describe them. Link-based classification predicts the category of an object based not only on its attributes, but also on its links, and on the attributes of linked objects. Web page classification is a well-recognized example of link-based classification. It predicts the category of a Web page based on word occurrence (words that occur on the page) and anchor text (the hyperlink words, that is, the words you click on when you click on a link), both of which serve as attributes. In addition, classification is based on links between pages and other attributes of the pages and links. In the bibliography domain, objects include papers, authors, institutions, journals, and conferences. A classification task is to predict the topic of a paper based on word occurrence, citations (other papers that cite the paper), and co-citations (other papers that are cited within the paper), where the citations act as links. An example from epidemiology is the task of predicting the disease type of a patient based on characteristics (e.g., symptoms) of the patient, and on characteristics of other people with whom the patient has been in contact. (These other people are referred to as the patients’ contacts.)

2. **Object type prediction.** This predicts the type of an object, based on its attributes and its links, and on the attributes of objects linked to it. In the bibliographic domain, we may want to predict the venue type of a publication as conference, journal, or workshop. In the communication domain, a similar task is to predict whether a communication contact is by e-mail, phone call, or mail.

3. **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.

4. Predicting link existence. Unlike link type prediction, where we know a connection exists between two objects and we want to predict its type, instead we may want to predict whether a link exists between two objects. Examples include predicting whether there will be a link between two Web pages, and whether a paper will cite another paper. In epidemiology, we can try to predict with whom a patient came in contact.

5. Link cardinality estimation. There are two forms of link cardinality estimation. First, we may predict the number of links to an object. This is useful, for instance, in predicting the authoritativeness of a Web page based on the number of links to it (in-links). Similarly, the number of out-links can be used to identify Web pages that act as hubs, where a hub is one or a set of Web pages that point to many authoritative pages of the same topic. In the bibliographic domain, the number of citations in a paper may indicate the impact of the paper—the more citations the paper has, the more influential it is likely to be. In epidemiology, predicting the number of links between a patient and his or her contacts is an indication of the potential for disease transmission. A more difficult form of link cardinality estimation predicts the number of objects reached along a path from an object. This is important in estimating the number of objects that will be returned by a query. In the Web page domain, we may predict the number of pages that would be retrieved by crawling a site (where crawling refers to a methodological, automated search through the Web, mainly to create a copy of all of the visited pages for later processing by a search engine). Regarding citations, we can also use link cardinality estimation to predict the number of citations of a specific author in a given journal.

6. Object compatibility. In object compatibility, the task is to predict whether two objects are, in fact, the same, based on their attributes and links. This task is common in information extraction, duplication elimination, object combining into single unit, and citation matching, and is also known as record linkage or identity uncertainty. Examples comprises predicting whether two websites are mirrors of each other, whether two citations actually refer to the same paper, and whether two apparent disease strains are really the same.

7. Group detection. Group detection are a clustering task. It predicts when a set of objects belong to the same group or cluster, based on their attributes as well as their link structure. An area of application is the recognition of Web communities, where a Web community is a collection of Web pages that focus on a particular theme or topic. A similar example in the bibliographic domain is the identification of research communities.

8. Sub graph detection. Sub graph identification finds characteristic sub graphs within networks. In chemistry, we can search for sub graphs representing chemical substructures.

9. Metadata mining. Metadata are data about data. Metadata provide semi-structured data about unstructured data, ranging from text and Web data to multimedia databases. It is useful for data integration tasks in many domains. Metadata mining can be used for schema mapping (where, say, the attribute customer id from one database is mapped to cust-number from another database because they both refer to the same entity); schema discovery, which generates schema from semi-structured data; and schema reformulation, which refines the schema based on the mined metadata. Examples include matching two bibliographic sources, discovering schema from unstructured or semi-structured data on the Web, and mapping between two medical anthologies.

F. Object Identity:

A final challenge is identity detection. How do we infer aliases, i.e., determine that two objects refer to the same individual? As mentioned earlier, some work has been done in this area by several research communities, but there is a great deal of room for additional work. Another aspect of this challenge is whether our statistical models refer explicitly to individuals, or only to classes or categories of objects. In many cases, we’d like to model that a connection to a particular object or individual is highly predictive; on the other hand, if we’d like to have our models generalize and be applicable to new, unseen objects, we also have to be able to model with and reason about generic collections of objects.

V. CONCLUSION

There have been a growing interest in learning from linked data, which is described by a graph in which the nodes in the graph are objects and the edges/hyper-edges in the graph are links—or relations—between objects. Tasks include hypertext classification, segmentation, information extraction, searching and information retrieval, discovery of authorities and link discovery. Domains include the world-wide web, bibliographic citations, criminology and bioinformatics, to name just a few. Learning tasks range from predictive tasks, such as classification, to descriptive tasks, such as the discovery of frequently occurring sub-patterns. We have given a brief summary of some of the work in this area, and some of the challenges in link mining. Link mining is showing a great potential in new area where relational learning meets statistical modeling; we believe many new and interesting machine learning research problems lie at the intersection, and it is a research area—which time has come.

REFERENCES

[1] Lise Getoor Dept. of Computer Science/UMIACS University of Maryland College Park, MD 20742, Link Mining: A New Data Mining Challenge
[2] Jiawei Han, Macheline Kambler, Jian Pei, Data Mining Concepts and Techniques, 2012.
2003.
[6] Link-based Classification for Text Classification and Mining, Q. Lu and L. Getoor. IJCAI workshop on Text Mining and Link Analysis