A Survey on Link Mining Applications

Zaved Akhtar*, Ravindra Kumar** and Umesh Chandra Jaiswal***
*Department of CSE , Dr. APJ AKTU, Lucknow (U.P.), India
javed.gkp@rediffmail.com
**Department of CSE , N. I. E. T., Greater Noida, U. P., India
rkumar_viet@rediffmail.com
***Department of CSE , MMMUT, Gorakhpur (U.P.), India
ucj_jaiswal@yahoo.com

Abstract: Now a day it is an emerging challenge in data mining to mining richly structures datasets where the objects are linked. Links between objects may demonstrate certain patterns which can be helpful for many types of data mining tasks and are usually very hard to capture with traditional statistical models. Several datasets of interest today are best described as a link collection of inter related objects. These may be represent homogeneous networks in which there are single-object type and link type (eg. people connected by friends links or the World Wide Web, a collection of linked web pages) or richer heterogeneous networks in which there may be multiple object and link types and possibly other semantic information. Examples of heterogeneous networks include those in medical domain describing patients, diseases, treatments and contacts, or in bibliographic domains such as describing publications, authors and venues. Link mining refers to data mining techniques that explicitly consider these set of links when building descriptive or predictive models of the linked data. Commonly link mining tasks include object ranking, collective classification, link prediction, group detection and sub-graph discovery. It is an exciting and rapidly expanding area. In this article we review link mining tasks, some of the common emerging themes and discuss ongoing link mining challenges, open issues and suggest ideas that could be opportunities for solutions. The most conclusion of this article is that providing an idea to usage link mining techniques from link mining to help to construct the Semantic Web as well as providing future scope in Link Mining.

Keywords: Link Mining, Data Representation, Semantic Web, Link Prediction, Graph Classification.

Introduction

Link mining refers to a data mining techniques that explicitly consider a link or collection of links for building predictive or descriptive models of the linked data. Generally Links mining tasks include, group detection, object ranking, link predictions, collective classification, graph and sub-graph discovery. It also represents an important and essential set of techniques for constructing useful applications of data mining in a wide variety of real and important domains, especially those involving complex event detection from highly structured data.

There are several emerging challenge for data mining is tackling the problems of mining richly and highly structured heterogeneous datasets. These types of datasets are best described as networks, graphs or sub-graphs. These domains often consist of a variety of object types; the objects can be linked in a different ways. Thus, the graph may have different node and edge (or hyper-edge) types. Attention must be taken that potential correlations due to links are handled appropriately. In fact object linkage is knowledge that should be exploited. This type of information 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. In addition, the graph structure itself may be an important element to include in the model. Structural properties such as degree and connectivity can be important indicators.

Links or 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, all links will not 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 an account more complex patterns arise as well. This leads to another challenge focused on discovering substructures, such as communities, groups, or common sub graphs.

Link mining is a newly emerging research area which is the intersection of link analysis [1; 2], hypertext and web mining [3], relational learning, inductive logic programming [4] and graph mining [5]. We use the term link mining is to put a special emphasis on the links –moving them up to first class citizens in the data analysis endeavor. In recent years, there have been several workshop series devoted to topics related to link miningin which one of the workshops was in the 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis [1]. Other workshop series include the workshops on Statistical Relational Learning, Multi- Relational Data Mining [6; 7], Link KDD [8; 9], Link Analysis, Counter-terrorism and Security

[10], Mining Graphs, Trees and Sequences [11].
The goal of this survey is to provide a perspective on research within the relevant communities that are addressing current link mining challenges. Link mining having a wide range of tasks; therefore our review will cover the core challenges addressed by a majority of ongoing research in the field. Beginning by describing some of the challenges in data representation for link mining. Then we progress through survey on link mining tasks. Finally, we close with a discussion of areas that we believe have not yet received sufficient attention and suggest ideas that could be opportunities for solutions.

Challenges in Data Representation

Feature selection and data representation are significant issues for traditional machine learning algorithms; data representation for linked data is even more complex. Consider a simple example of a social network describing actors and their participation in events. Such social networks are so commonly called affiliation networks and are easily represented by three tables representing the actors, the events and the participation relationships. Even this simple structure can be represented as several distinct graphs. The most natural representation is a bipartite graph, with a set of actor nodes, a set of event nodes, and edges that represent an actor's participation in an event. Other representations may enable different insights and analysis. For example, we may construct a network in which the actors are nodes and edges correspond to actors who have participated in an event together. This representation allows us to perform a more actor-centric analysis. Alternatively, for more easily we may represent these relations as a graph in which the events are nodes, and events are linked if they have an actor in common. Therefore the choice of an appropriate data representation is actually a major issue in link mining, and is often more complex than in the case when we have IID data instances.

Literature Review on Link Mining Tasks

There are several Link Mining tasks, some of them are as follow: -

Link Prediction

Link prediction is used to predict future possible links in the network or predict missing links due to incomplete data (E.g., Food-webs – this is related to sampling that Olivia spoke of earlier). Prediction is very difficult, especially when if it's about the future social networks are dynamic [38, 39]. New Links appear indicating new interaction between objects. Link prediction is a problem of predicting the existence of a link between two entities, based on attributes of the objects and other observed links. A. Popescul and L. H. Ungar [24] introduce a structured logistic regression model that can make use of relational features to predict the existence of links.
Examples include predicting links among actors in social networks (E.g., Facebook) such as predicting friendships; predicting the participation of actors in events. Other examples are such as email, co-authorship and telephone calls; and predicting semantic relationships such as "advisor-of" based on web page links and content [31].

Sub-graph Discovery

Sub-graph identification finds characteristics sub-graphs within networks. An area of data mining which is related to link mining is the work on sub-graph discovery. This work attempts to find commonly or interesting occurrence of sub-graphs in a set of graphs. Discovery of these patterns may be used for graph classification or the sole purpose of the systems.

Graph Classification

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.
Three main approaches to graph classification are based on feature mining on graphs, inductive logic programming (ILP), and defining graph kernels. Feature mining on graphs uses methods related to those described in the previous section on sub-graph discovery, mining of features on graphs are usually performed by finding all frequent or informative sub structures in the graph instances. These sub structures are usually used for transforming the graph data into data represented as a single table and then traditional classifiers are used for classifying the instances.

Link-Based Object Ranking

Another Link Mining task is that of link-based object ranking. Primary focus of link-based object ranking is the link analysis community. The objective of Link Based Object Ranking is to exploit the link structure of a graph to order or prioritize the set of objects within the graph. Page Ranking and HITS algorithms are most notable approaches for Link Based Object Ranking.
In Context of web information retrieval, the PageRank [12] and HITS [13] algorithms are the very most notable approaches to Link Based Object Ranking. PageRank models web surfing as a random walk where the surfer randomly selects and

follows links and occasionally jumps to a new web page to start another traversal of the link structure. The rank of a given web page in this context is the fraction of time that the random web surfer would spend at the page if the random process were iterated ad infinitum. This can be determined by computing the steady-state distribution of the random process.
Bharat and Henzinger [23] and Chakrabarti et al. [29] propose modifications to HITS that exploit web page content to weight pages and links based on relevance. Ng et al. analyze the stability of PageRank and HITS to small perturbations in the link structure and present modifications to HITS that yield more stable rankings. Haveliwala [27], Jeh and Widom [28] propose topic-sensitive PageRank algorithms that identify topically authoritative web pages efficiently at query time.
In the domain of social network analysis (SNA), Link Bases Object Ranking 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.

Link-Based Object Classification
In traditional classification method, objects are classified based on the their 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. Traditional machine learning has focused on the classification of data consisting of identically structured objects that are typically assumed to be IID. Many real-world datasets, however, lack this homogeneity of structure. In Link Based Object Classification problem, a data graph G = (0;L) is composed of a set objects 0 connected to each G = (0;L) is combined of a set objects 0 connected to each other via a set of links L. The tasks are to label the members of 0 from a finite set of categorical values. Features of Link Based Object Classification s makes it different from traditional classifications in many cases, the labels of related objects tend to be correlated.
Link Based Object Classification has received considerable attention recently. S. Chakrabarti, B. Dom, and P. Indyk [36] consider the problem of classifying related news items in the Reuters dataset. They were the first who focused in this research paper that on exploiting class labels of related objects aids classification, where exploiting features of objects which are related to them can actually harm classification accuracy. Oh reported similar results on a collection of encyclopedia articles: simply incorporating words from neighboring documents was not helpful, while making use of the predicted class of neighboring documents was helpful. Lafferty et al. introduce conditional random fields (CRF), which extend traditional maximum entropy models for LBC in the restricted case where the data graphs are chains. B. Taskar, P. Abbeel, and D. Koller [30] extend Latterly et al.'s approach to the case where the data graphs are arbitrary graphs. Another researcher Neville proposed a simple Link Based Object Classification algorithm for classifying corporate datasets with richly structured schemas that produces graphs with heterogeneous objects in which everyone with its own distinct set of features. Lu and Getoor [2, 25] extends simple machine learning classifiers to perform Link Based Object Classifications by introducing new features that measures the distribution of class labels in the Markov blanket of the object which has to be classified.

Link-Based Cluster Analysis
The goal of Link-Based cluster analysis is to find occurring subclasses. This is done by segmenting the data into groups, where objects in a group are similar to each other and are very dissimilar from objects in different groups. Unlike classification, clustering is unsupervised and can be applied to discover *hidden patterns* from data. This makes it an ideal 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 [14; 15]. All of these algorithms assume that each object is described by a fixed length attribute-value vector.

Group Detection
Group detection is another clustering task. It predicts when a set of objects belong to the same group or cluster, based on their set of attributes as well as their link structure. For example group detection is to cluster the nodes of a graph into group or groups which share common characteristics. In recent years, there was a central challenge to develop scalable methods that can exploit increasingly complex graphs to aid the knowledge discovery process.
Consider first the case where a graph contains several objects and links of a single type, without attributes. Many of the techniques for identifying groups in this scenario can be classified as either divisive or agglomerative clustering methods.
The task of create a block modeling of social network analysis (SNA) involves partitioning social networks into the sets of individuals which is called positions, that exhibit similar sets of links to others in the network. A similarity measures are defined between agglomerative clustering and are used to identify the positions. A Spectral Graph partition 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 [16]. In a related vein, D. Gibson, J. Kleinberg, and P. Raghavan [17] have shown that the dominant eigenvectors of the HITS authority matrix provide a natural decomposition of web community structure. Other recent approaches for group detection use a measure of edge betweenness, derived from Freeman's notion of betweenness centrality,

to identify links connecting groups [18]. Links with high edge betweenness are incrementally removed to partition the graph.

Entity Resolution

Entity Resolution is anther object-centric task, which involves to identifying the set of objects in a domain. The main objective of entity resolution is to determine which references in the data refer to the same real-world entity.

The use of links for resolution was first explored in databases. R. Ananthakrishna, S. Chaudhuri, and V. Ganti [37] introduced a methodology for deduplication using links in data warehouse applications where there are dimensional hierarchy over the link relations. D. V. Kalashnikov, S. Mehrotra, and Z [19] enhanced feature-based similarity between an ambiguous reference and the many entity choices for it with linkage analysis between the entities, such as affiliation and co-authorship. In database [20; 21] collective entity resolution approaches have also been proposed where one resolution decision affects another if they are linked. Bhattacharya and Getoor [20; 22**]** propose different measures for linkage similarity in graphs and show how these can be combined with attribute similarity for collective entity resolution in collaboration graphs. X. Dong, A. Halevy, and J. Madhavan [21] collectively resolve entities of multiple types by propagating evidence over links in a dependency graph.

Open Issues and Future Research Areas

In core pattern discovery and pattern detection, the existing real applications typically have many requirements. Requirements arises from characteristics of the environment; e.g., the need to combine data from multiple sources, the need to compute certain derived attributes for use by the pattern discovery and pattern detection algorithms, , the need to record and audit system activities, the need to support multiple analysts, the need to sort data into distinct threads of interest, the need to visualize patterns (in both senses of the term), , the need to support detection of patterns that may occur over long time periods, the need to both discover and detect patterns that are continually changing, the need to protect the identity of entities until a particular level of belief in their interestingness is supported and, perhaps, proper approvals are obtained, the need to allow for multiple competing hypotheses, the need to allow for frequent tuning between false positives and false negatives, the need to allow for refutation of previously asserted evidence, the need to support both comprehensive review of all data based on approved patterns and ad-hoc review based on particular external indicators, the need to support organizational workflow processes, the need to support what-if analyses by individuals, the need to operate continuously and perhaps autonomously on incrementally arriving massive data streams, and perhaps others. Apart from these there are several other research issues. We suggested some ideas for future work which are as follow:

Usage of Links for Query-Based Classification

Using links, Query-Based Classification is an interesting research direction area. When a user is interested in specific area or set of contents, it is worthless to provide huge amount of information to him. In classification approach to consider the dataset in its whole as one linked instance of object that may perform prediction or classification for all of these linked objects jointly. When a user is interested in classifying only a small subset of these objects, it is worthwhile to classify other objects only if they are helpful in correctly classifying the objects of interest via the link structure.

Probabilistic Relational Models for Feature Construction

In multi-relational setting feature construction is a great challenge. The attributes of an object provides a basic description of the object. Traditional classification algorithms are based on these types of object features. When we use link-based approach, it may also make sense to use attributes of linked objects. Furthermore if the links themselves having set of attributes then these may also be used. This is the idea behind propositionalization [32]. However, as others have noted, simply flattening 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 [33]. We have found this works well for learning probabilistic relational models [34], but this approach may not always be appropriate so there is a need to modify and upgrade probabilistic relational models for feature selection.

Web Search & Retrieval

Web Information Retrieval evaluation suggest that ‘user effort’ can be approximated to the search length method where search length corresponds to the number of irrelevant documents encountered before arriving at a relevant document. A more elaborate version of the metric is ‘expected search length for n’, which is defined as the number of documents it is necessary to traverse before finding ‘n’ relevant documents. At the time of retrieval of information from the web, majority of web search engines return a linear ranked list of results and even when there is an attempt to convey that several documents share the same rank – users are oblivious to it. Some research has gone into how best to present search results, but as so many search engines opt for the same linear ranked list presentation – it would be impossible to differentiate them in that regard. We suggest link mining techniques used for web search and retrieval the text as well as information efficiently.

Link Mining Techniques to discovering patterns and build useful prediction system

Now a day's amount of data grows and the number of sources expands very rapidly, techniques from link mining can help us discover patterns and build useful prediction system. We suggest applying link mining techniques for developing a pattern discovery and prediction system to predict associations among objects. This is useful to identify web user surfing patterns, user groups etc.

Usage of Link Mining Techniques to Construct Semantic Web

Now a day's Web search results are high recall, low precision as well as results are highly sensitive to vocabulary. The Semantic web is a web of data which is an extension of the current web in which information is in well-defined meaning, exact or approx as well as provide better enabling computers and users to work in co-operation.

The Semantic Web is about two things, describes the relationships between these things (like A is a part of B and Y is a member of Z) and the properties of things (like age, weight, size and price).There are common formats for combination and integration of data drawn from diverse sources, where on the original web mainly concentrated on the interchange of documents. It is also about the used languages for recording how the data relates to real world objects as well as allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing. The Semantic Web is not about links between pages, it describes the relationships and properties between things.

We suggest to discovering interesting sub-graph based on semantic information associated with the edges and combining information extraction with techniques from Link Mining to help to construct the Semantic Web.

Conclusion

To defining and addressing the core link mining challenges, there are significant progress has been made in recent years, yet much work remains to be done in refining and combining various approaches and solutions. The most notable conclusion of this paper conclusion of this that there are several link mining techniques that works well for individual or specific link mining tasks but not yet a comprehensive framework or tool that can support a combination of link mining tasks which are needed in many applications. To construction of successful and useful link mining applications is still remain a challenge. Link mining tasks and challenges provide interesting insights and catalyze new research directions. To designing or developing an effective architecture to support all necessary functions of an integrated application and providing a solution to usage a link mining for Semantic Web is also a key to success.

References

[1] D. Jensen and H. Goldberg, "AAAI Fall Symposium on AI and Link Analysis", .AAAI Press, 1998.
[2] L. Getoor, "Link Mining a New Data Mining Challenge", SIGKDD Explorations,2003, 5 (1), pages 84-89.
[3] Peng Wang, Wen XUBao, Yurong WU and XIO Yu Zhou, " Link Prediction in Social Networks: The State-Of-the-Art", Cornel University library, Information Sciences, China, Vol 58, January 2015, pages 011101:1-011101:38.
[4] S. Dzeroski and N. Lavrac,. "Relational Data Mining", Kluwer, Berlin, 2001.
[5] D. J. Cook and L. B. Holder, "Graph-based Data Mining", IEEE Intelligent Systems, 2000, 5(2), pages 32-41..
[6] T. Dietterich, L. Getoor, and K. Murphy, "ICML Work-shop on Statistical Relational Learning and its Connections to Other Fields", 2004.
[7] S. Dzeroski and H. Blockeel, KDD Workshop on "Multi-Relational Data Mining", 2004.
[8] Anshu Zhang, Wenzhung Shi, Geoffery, I. Webb, "Mining Significant Association Rules From uncertain data", Data Mining and Knowledge Discovery, Vol 30, ISSN: 1348-5810, July 2016, pages 928-963.
[9] J. Adibi, H. Chalupsky, M. Grobelnik, N. Milic-Frayling, and D. Mladenic, KDD Workshop on "Link Analysis and Group Detection", 2004.
[10] D. Skillicorn and K. Carley, SIAM Workshop on "Link Analysis, Counterterrorism and Security". 2005.
[11] P. Bhattacharya, A. Garg and F. S. Wig, "Analysis of User Keyword Similarities in Online Social Networks", Social Network Analysis and Mining, 2011, Issue 1, pages 143-158,..
[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web", Technical report, Stanford University, 1998.
[13] J. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 46(5),1999, pages 604-632.
[14] B. Taskar, E. Segal, and D. Koller, "Probabilistic classification and clustering In relational data", IJCAI-01, 2001.
[15] J. Kubica, A. Moore, J. Schneider and Y. Yang, "Stochastic link and group detection", AAAI-02, 2002.
[16] M. E. J. Newman, "Detecting community structure in networks", European Physical Journal B, 38, 2004, pages 321-330.
[17] D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring web communities from link topology", In ACM Conference on Hypertext and Hypermedia, 1998, pages 225-234.
[18] J. R. Tyler, D. M. Wilkinson and B. A. Huberman, "Email as Spectroscopy: Automated Discovery of Community Structure within Organization", Kluwer, B.V., Deventer, The Netherlands, 2003.
[19] D. V. Kalashnikov, S. Mehrotra, and Z. Chen, "Exploiting relationships for domain-independent data cleaning", In SIAM International Conference on Data Mining, April 21-23, 2005.

[20] I. Bhattacharya and L. Getoor, “Iterative record linkage for cleaning and integration”, In SIGMOD 2004 Workshop on Research Issues on Data Mining and Knowledge Discovery, June 2004.
[21] X. Dong, A. Halevy, and J. Madhavan, “Reference reconciliation in complex information spaces”, In ACM SIGMOD International Conference on Management of Data, 2005, pages 85-96.
[22] I. Bhattacharya and L. Getoor, “Entity resolution in graphs”, Technical Report 4758, Computer Science Department, University of Maryland, 2005.
[23] K. Bharat and M. R. Henzinger, “Improved algorithms for topic distillation in a hyperlinked environment”, In ACM SIGIR International Conference on Research and Development in Information Retrieval, 1998, pages. 104-111.
[24] A. Popescul and L. H. Ungar, “Statistical relational learning for link prediction”, In IJCAI Workshop on Learning Statistical Models from Relational Data, 2003.
[25] Lu, Q, L. Gooter, “Link-based Classification”, ICML'03, Washington DC, 2003.
[26] Sen, P, L. Gooter, “Link-based Classification”, University of Maryland CS-TR, 2007, pages 48-58.
[27] T.H.Haveliwala,“Topic-sensitive PageRank. In International Conference on the World Wide Web (WWW),2002, pages. 517-526.
[28] G. Jeh and J. Widom, “Scaling personalized web search”, In International Conference on the World Wide Web (WWW), pages. 271-279, 2003.
[29] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan, “Automatic resource list compilation by analyzing hyperlink structure and associated text”, In International World Wide Web Conference (WWW), 1998.
[30] B. Taskar, P. Abbeel, and D. Koller, “Discriminative probabilistic models for relational data”, In UAI, Edmonton, Canada, 2002, pages 485-492.
[31] B.Taskar, M.. F.Wong, P.Abeel, and D. Koller, “Link prediction in relational data”, “In Neural Information Processing Systems Conference, Vancouver, Canada, December 2003.
[32] S. Kramer, N. Lavrac, and P. Flach, “Propositionalization approaches to relational data mining”, In S. Dzeroski and N. Lavrac, editors, Relational Data Mining, Kluwer, 2001, pages 262 - 291.
[33] H.J. Oh, S. H. Myaeng, and M.-H. Lee, “A practical hypertext categorization method using links and incrementally available class information”, SIGIR-00, 2000.
[34] L. Getoor, N. Friedman, D. Koller and A. Pfeffer, “Learning probabilistic relational models”, In S. Dzeroski and N. Lavrac, editors, Relational Data Mining, Kluwer, 2001, pages 307– 335.
[35] D. Jensen, “Statistical challenges to inductive inference in linked data”, In Seventh International Workshop on Artificial Intelligence and Statistics, 1999.
[36] S. Chakrabarti, B. Dom and P. Indyk, “Enhanced hypertext categorization using hyperlinks”, In SIGMOD International Conference on Management of Data, 1998, pages 307- 318.
[37] R. Ananthakrishna, S. Chaudhuri and V. Ganti, “Eliminating fuzzy duplicates in data warehouses”, In International Conference on Very Large Databases (VLDB), Hong Kong, China, 2002.
[38] Tarique Anwar and Muhammad Abulaish, “Ranking Radically Influential Web Forum Users”, IEEE Transactions on Information Forensics and Security, Vol. 10, Issue 6, 2015, pages. 1289-1298.
[39] P. Symeonidis and N. Manta, "Spectral clustering for link prediction in social networks with positive and negative links", Social Network Analysis and Mining, 2013, pages 1433-1447.