ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Issue 03, March 2018

Survey Paper on Link Mining

C.Uma[1] S.Krithika2 N.Rajasekaran[3]
[1]Assistant Professor, Department of CT & IT, umachinnsamy@gmail.com
[2]Assistant Professor, Department of Computer Science (P.G.), krithitup86@gmail.com
[3]Assistant Professor, Department of Computer Applications, rajasekarandpm@gmail.com
Kongu Arts and Science College (Autonomous), Erode-638 107.Tamil Nadu.

Abstract: Link mining refers to data mining techniques that explicitly consider these links when building predictive or descriptive models of the linked data. 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 is an ex-citing, rapidly expanding area. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models.
Keywords: object ranking, group detection, link prediction, graph discovery

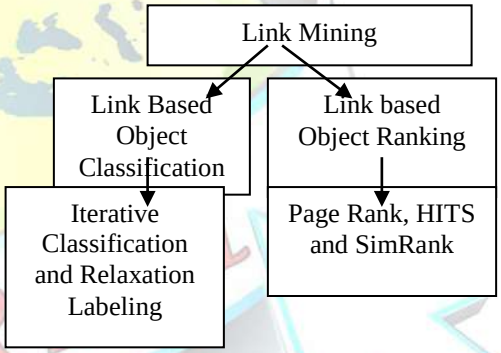
I. INTRODUCTION

Link mining is a newly emerging research area that is at the intersection of the work in link analysis [1; 2], hypertext and web mining [3], relational learning and inductive logic programming [4], and graph mining [5]. We use the term link mining to put a special emphasis on the links moving them up to rest-class citizens in the data analysis endeavor. In recent years, there have been several workshop series devoted to topics related to link mining.

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 subgraph patterns.

II. LINK MINING TASKS

The domain of link analysis encompasses several distinct tasks. These are essentially determined by the different possible outcome of analyzing link data. Link analysis tasks can usually be grouped into a small set of overall categories.



[Fig 1.1] Link based mining activities

2.1 Link-Based Object Classification (LOC)

Link based Object Classification is a technique used to assign class labels to nodes according to their link characteristics. One simplified example is to classify nodes as strongly connected and weakly connected depending solely on their degree.

A slightly more complex process would be find the average distance of each node to all other nodes, and classify them according to that quality. The distance of one node to another is number of edges that needed to be traverses along the shortest path between them. Assuming that all nodes are connected to each other, this average distance would be indicator of how central a node is within a network. Thus, nodes can be classified as belonging to the core of a network or not, based on a suitable threshold.

LOC can also incorporate information about a node's properties for classification. For instance, if task is to create

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Issue 03, March 2018

compatible teams from a pool of personnel, and have generic preference data from everyone, then build up a graph, where each node represents a person and each edge represents a common preference between two persons. After that manually assign different group labels to a select set of individuals and then assign groups to everyone else based on the number of edges share with people who have already have been labeled. A few iteration of this process should result in an amicable classification of team members. Such classification efforts that create groups of nodes are sometimes referred to as Group Detection tasks.

2.2 Link Based Object Ranking (LOR)

Link Based Object Ranking ranks objects in a graph based on several factors affecting their importance in the graph structure, whereas LOC assigns labels specifically belonging to a closed set of finite values to an object [6]. The purpose of LOR is not to assign distinctive labels to the nodes usually, all nodes in such networks are understood to be of the same type the goal is to associative a relative quantitative assessment with each node.

LOR can sometimes be more fine-grained version of LOC. If desire to mark each node with the precise number representing its degree of connectivity, then it can be one form of ranking nodes. Ranking nodes are usually much more complex than that, and take into account a large part of the graph when coming up with a figure for each node.

One of the most well- known ranking tasks is ranking web pages according to their relevance to a search query. Research and practical use have shown that the relevance of a search result not only depends upon the content of the document but also on how it is linked to other similar documents. There are algorithms that try to identify research papers that have the most comprehensive knowledge about a given topic by analyzing how many other relevant papers have cited them. Some social network games include a notion of popularity that is defined by how well connected each person is with others and what this person's respective popularity figure is.

2.3 Link Prediction

Being able to see the future is usually a nice capability, although it is quite hard. Predicting how things may turn out, within some proven bounds of approximation, is not bad either. Prediction has always been a basic for development of a many artificial intelligence techniques.

Not that while LOC and LOR are analysis of links to talk the nodes in a network, Link prediction actually deals with links themselves.

2.3.1 Link Prediction Algorithm

(a) Graph Data Processing:

The line numbers at the end of each step correspond to the line numbers of that step in Algorithm 3.

1. Accept raw data representation of a collaboration or co-authorship network, in the form of edge list and a year attribute for each edge at the least.
2. Spilt this data into training and test sets. For maximum accuracy, the prediction process should depend only on attributes intrinsic to the network. Hence the newer vertices in the test graph that are not in the training graph are pruned.

Algorithm: Graph Data Processing

1. Input: D- Duration of test data
   IG- Input graph
   Output: $GT_{training}$ – The training graph
   $GT1_{test}$ – The test graph
   $G`T_{test}$- The pruned test graph.

/*Let $year_{start}$ denote begin year of data
/* Let $year_{end}$ denote end of data
/* Let pruned denote vertices to be pruned from the test data
/* Let V(G) denote vertices of graph G

2. Extract the $year_{start}$ and $year_{end}$ from the year attribute of the edges.
3. $GT1_{test} = IG[year_{end} –D+1: year_{end}]$
4. $GT_{training} = IG- GT1_{test}$
5. $pruned = V(GT1_{test}) – V(GT_{training})$
6. $G`T_{test} =V(GT_{test}) – prund$
7. return $GT_{training}$, $GT1_{test,}$ $G`T_{test}$

(b) Computing Most Portable Links:

After having processed the graph data, the steps involved in computing probable links are quite straightforward.

1. Compute the score of all possible edges using the chosen proximity measure.
2. Select the proximity values above the threshold and return the edges associated with these values as a graph.

Algorithm: Compute Most Portable Links

1. Input: G2 – Input Graph
   T1 – Threshold for prediction
   M1- Proximity measure to be used in link prediction
   Output: $G1_{priticited}$ – A graph containing predicted scores.

/* Let predicted denote a matrix of proximity values for each pair of vertices
/* Let Output denotes a matrix of Boolean values
/* compute the proximity values by applying the measure on G2

2. Predicted:= M1(G2)

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Issue 03, March 2018

3. Output: = (Predicted >= T1)
4. Generate graph $G_{predicted}$ from adjacency matrix represented by output.
5. Return $G1_{predicted}$

2.4 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.

Three main approaches to graph classification have been explored. These are based on feature mining on graphs, inductive logic programming (ILP), and defining graph kernels. Feature mining on graphs is usually performed by finding all frequent or informative substructures in the graph instances. These substructures are used for transforming the graph data into data represented as a single table, and then traditional classifiers are used for classifying the instances. As an example of an ILP approach, King et al. [7] rst map the graph data describing mutagenesis into a relational representation. Their logical representation uses relations such as vertex (graphId, VertexId, VertexLabel, VertexAttributes) and edge (graphId, vertexId1, vertexId2, BondLabel), and then uses an ILP system to and a hypothesis in this space. Finding all frequent substructures is usually computationally prohibitive. An alternative approach makes use of kernel methods.

III. CONCLUSION

More and more domains of interest today are best described as a linked collection or network of interrelated heterogeneous objects. Data mining algorithms have typically addressed the discovery of patterns in collections of IID in-stances. 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. We have surveyed several of the better studied link mining tasks: link-based object ranking, link-based object classification, group detection, entity resolution, link prediction, subgraph discovery, graph classification, and generative models for graphs.

REFERENCES

[1]. D. Jensen and H. Goldberg. **AAAI Fall Symposium on AI and Link Analysis**. AAAI Press, 1998.

[2]. R. Feldman. Link analysis: Current state of the art, 2002.

[3]. S. Chakrabarti. Mining the Web. Morgan Kaufman, 2002.

[4]. S. Dzeroski and N. Lavrac, editors. **Relational Data Mining**. Kluwer, Berlin, 2001

[5]. D. J. Cook and L. B. Holder. Graph-based data mining. **IEEE Intelligent Systems**, 15(2):32{41, 2000

[6]. L.Getoor. Link Mining: A new data mining challenge. ACM SIGKDD Explorations Newsletter, 5:84-89, 2003.

[7]. R. D. King, S. H. Muggleton, A. Srinivasan, and M. J. E. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. National Academy of Sciences, 93(1):438{442, January 1996.

Ms. C.Uma, received M.Sc degree from Anna University, Chennai and M.Phil degree from Bharathiar University, Coimbatore, TN, India. At currently working as an Assistant Professor in Department of Computer Technology and Information Technology, Kongu Arts and Science College (Autonomous), Erode. I have 9 years of teaching and 7 years of research experience. I had published more number of papers in various international and national journals. My research interest in the area of data mining.

Ms. S. Krithika, received M.C.A and M.Phil degree from Bharathiar University, Coimbatore, TN, India. At currently working as an Assistant Professor in Department of Computer Science (PG), Kongu Arts and Science College (Autonomous), Erode. I have 9 years of teaching and 7 years of research experience. I had published papers in national and international conferences. My area of interest is Data Mining.

Mr. N.Rajasekaran, received M.C.A and M.Phil degree from Bharathiar University, Coimbatore, TN, India. At currently working as an Assistant Professor in Department of Computer Applications, Kongu Arts and Science College (Autonomous), Erode. I have 8 years of teaching and 7 years of research experience. I had published papers in international jourals. My area of interest is software testing.