IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 50-55

Link Mining: Information Mining Using Links

Mr. AkshayJaveri
(Masters of Computer Applications, Navinchandra Mehta Institute of Technology and Development, Dadar(W),Mumbai, India)

Abstract: *The Challenging task for data mining is solving the problem of mining proper structured databases, for creating this structured databases various objects are linked together in some manner. These links among the objects describes certain patterns, these pattern identification can help for doing many data mining tasks easily. Recently there is a sudden strong feeling of interest in this area, because of web and hypertext mining due to interest in social networks mining, security and law enforcement data, Reference to a book, article or other published item, epidemiology.*

Keywords: *Data mining, Link mining, mutual information, epidemiology, bibliography, KDD, CRISPDM, XML, HITS, IR.*

I. Introduction

For interrelated data, knowledge can be considered as power. In various links of heterogeneous information networks the knowledge is hidden. Varying on the degrees of success, there are many data mining methodologies such as KDD and CRISPDM. They are also used in practice. As per our knowledge there is no methodology which is developed to support link mining. However, for data mining and knowledge discovery in databases well know methodology such as Cross Industry Standard Process for Data Mining (CRISPDM), developed by a group of companies that work closely together for a particular purpose which can be connected with what is happening to link mining study. To detect something that is different from usual in the field of link mining the study of CRISPDM has been adapted. To infer the links which are not known in a network is the most important goal in link mining. Through the use of case study which aims to use the information to interpret connected to with the meaning of words or sentences something that is different from usual identified co-citations this approach was implemented, this can help in gaining the insights by determining the nature of a particular link and also identifying future links and how they are related to each other.

To find patterns in a dataset characterized by a collection of independent instances of a single relation the traditional data mining tasks such as association rule mining, market basket analysis and clustering analysis were commonly used. This is consistent with statistical inference problem of trying to identify a model given a random sample from a commonly underlying distribution.

The mining of richly structured, heterogeneous datasets is the problem and for the same is the key challenge for data mining. These are multi-relational type of datasets. XML is a semi-structured representation, first-order logic or using relations these may be described by a relational database. A variety of object types and the linking of objects in some manner are the commonalities of the domain which consists of such objects. In this case, a constructed link using the join operation between the tables stored in database, explicit link such as URL, the instances in our datasets are linked in some manner.

Inappropriate conclusions can be led by naively applying of traditional statistical inference procedures which assumes that instances are independent [1]. Potential correlation due to the links should be handled appropriately this care must be taken. The knowledge of record linkage should be used. To improve predictive accuracy of learned models this information can be used: attributes of linked objects are often correlated and links are more likely to exist between objects that have some commonality.

For classification and clustering new data algorithms are required in linked relational domains and now also the new tasks come into presence due to the introduction of the links. Examples include predicting the number of links and its type, inferring object identity, finding co-references and discovering patterns.

The analysis must be undertaken on efficient, reliable and robust data also by identifying the outliers is to be ensured which is the crucial step for data and link mining. Unless the quality of the underlying links makes it infeasible high data delivery reliability should be achieved by reliability of detection anomaly. Even if topology changes, dynamic networks, huge and complex social network failures robustness should be robust. Efficiency in communication often applies both complex as well as different types of anomalies this provides an opportunity to make method detection anomalies more efficient. Even if outliers are considered as a noise or an error in data mining, are referred as something that is different from usual as they can carry important information. Often data tends to be contained noise and they must be examined context of domain is considered for doing so after that people may come across some true errors which are to be removed[2].

It is believed that if people apply mutual information by better understanding of the anomalies to data entities and links and objects to reveal their semantic relationship. Based on mutual information approach to anomaly detection in link mining is all about this paper presents.

II. History Of Link Mining

The major example of making the use of the link mining exploring link structure is information retrieval results improvisation. Based on the web link structure is the well-known page hub and authority scores [3]. There are various such algorithms such as HITS (Hypertext Induced Topic Search) which are between web pages citation related. There are many algorithms which are examining the various relations such as based on co-citations of the web pages to find the other related webpages proposed by Henzinger and Dean[4].

Another Important aspect to be worked on is related to working on hypertext as well as the classification of the web pages. For the classification of the web pages hypertext can be used. Earlier the IR(Information Retrieval)Community did not make the use of the link structures. The webpage also have the classification problem because the categorizations of the web pages based on the information between the features of the current and neighbor webpage. But now with the use of the linking information labeling category wise of the web pages can be done which helps in achieving the results better so this is one of the major objectivewhich is to be taken care. Only naively incorporating words from the neighbor page can reduce the performance sometimes while prefix category of web page can improve performance.

III. Method Of Link Mining

As discussed earlier the CRISPDM was being developed and used in data mining and knowledge discovery and the same is being adapted in the rising field of link mining. As data mining procedure is used to find out certain patterns in the data entities while in link mining the case is finding patterns as well as modeling the links among the objects. The adopted CRISPDM methodology consists of six stages as follows:

Diagram:

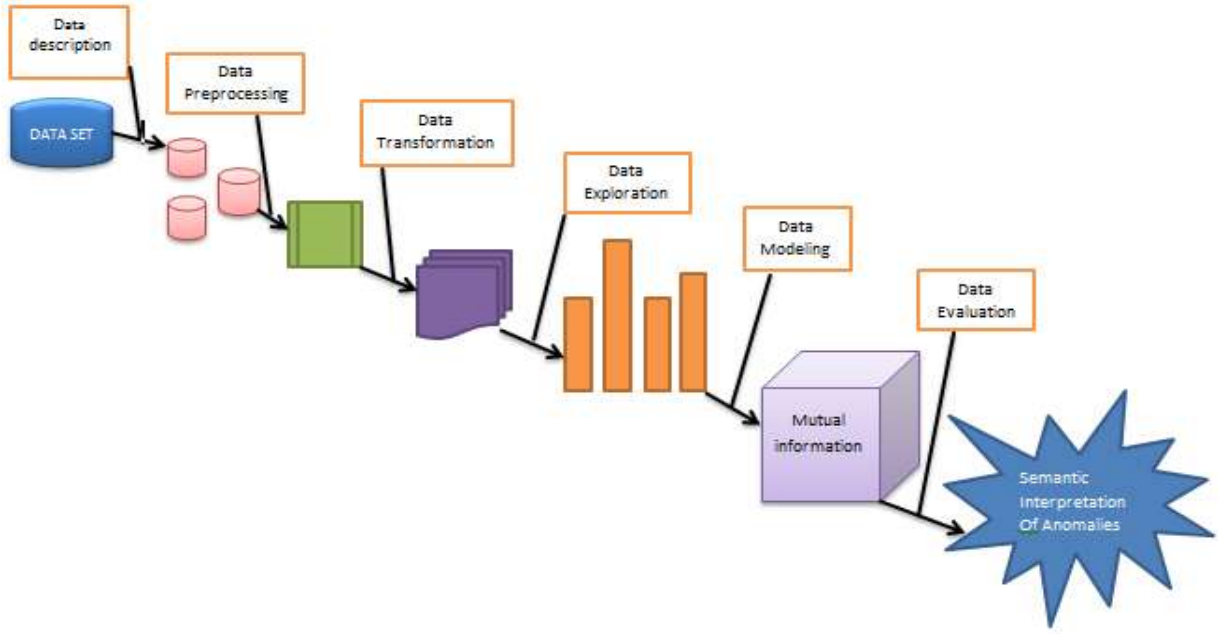

Figure 1: Link mining methodology

Explanation:

This method helped in understanding the various tasks and objective in regards with link mining:

1) **Data description:** Firstly in this stage the raw data is being collected and the user is allowed to get familiar with the data. It focuses more on quality of the data being provided to the user and to get insights and check if there are any problems associated with it as well as identifying hidden information.
2) **Data preprocessing:** In this stage the raw data is cleaned and also being integrated to finally construct a proper database. In data mining while cleansing the data the noise in the raw data is eliminated by smoothing the outliers as they were considered to be as errors in the data mining but in the case of link mining the these outliers prove to be of major advantage as they can provide important knowledge of this study.

3) **Data transformation:** The modeling tool may be required to modify the data along with the features of the objects for their proper representations which is important in link mining. Objects can be anything such as people, countries. Once these objects are transformed they can help in knowing the strength of the links among the objects.
4) **Data exploration:** In this stage the data is distributed and given as an input to graphical tools to get the diagrammatic representation of the data which will be easy to observe and understand as well as it will help in getting the knowledge of the objects structure and their links.
5) **Data modeling:** In this stage based on the previous stage identification of all entities and their relationship is carried out and accordingly the type of algorithm is decided such as nearest neighbor or classification based on information for various huge and complex data.
6) **Data evaluation:** This stage describes that based on mutual information provided early by the researcher with the described contents in it has helped the user with the more important knowledge which may be hidden or according to ranking of the links more useful data is being provided by also removing the redundant data.

IV. Link Mining Workflow

As described in the introduction link mining gives a new direction to the traditional data mining tasks and possesses new problem with it. There is a list of possible tasks which will be illustrated below:
Web page collection: In a web page collection , there are many web pages which are considered as objects and links such as In-link, Out-link as well as co-citation link and these are used to link the pages and this is possible by using the attributes such as HTML tags, anchor text and word appearances.
Bibliographic domain: In this domain, papers, authors, institutions, journals and conferences are the objectives. Paper citations, author and co-authorships, affiliations, and the appear-in relation between a paper and a journal are the links.
Epidemiological Studies: In this domain objects include patients, people they have come in contact with and disease strains. The person infected with which disease strain is represented by the links.

4.1 Classification Based On Links

By using the links for upgrading in the old traditional data mining task is link-based classification. Here people are interested in predicting object with its category, attributes, links as well as objects linked by some path of edges.

One of the example based on the classification of links and which has received a fair attention is web page collection and in this domain the problem is to identify the web pages with its category, attributes, links between the web pages, anchor text. In bibliographic domain, the example based on link classification is predicting the category of paper, citation, papers that are cited with this paper. In epidemiological domain, the example based on link classification is predicting the disease type based on the characteristics of the people or predicting the persons age based on the disease they are infected with.

4.2 Cluster Analysis Based On Links

Here the data is segmented and divided into the groups based on theirs characteristics by finding their naturally occurring sub classes. Here the group of similar objects are placed in one group and the dissimilar ones in another group. Clustering can help in discovering the hidden patterns of data. This technique can help many applications such as information retrieval, criminal analysis and many others as well.

K-means is the most commonly used clustering algorithm. More research work going on in areas of clustering such as pattern recognition.People can cluster objects, links of the objects as well. Frequently occurring pattern finding is been more focused which is based on apriori algorithm which is used for the same purpose [5].

In web page collection the example of clustering is ranged from pages that point to a lot of pages of the same category to identifying similar cites. Example of clustering in bibliographic domain includes finding authors that publish together as a group, discovering research areas, based on citations. Example related to epidemiological study which includes finding patients with diseases with similar transmission patterns.

4.3 Link Type Identification

To check whether the link exists or not there are various tasks amongst them the simple one is to determine the type of the link. For example people may be trying to predict the relation between the two people who know each other such as family member or co-authors.

The link type can be described in a number of ways. In some cases the type may be an attribute of the link. In such situations person know that there is an existence of the link but person may be interested in

knowing the type of link. In first example described above person might know there is some connection between two people but the person is interested in finding that whether it is familial or co-author or coworkers relation. In other cases there may be various different kinds of links. There can also be many different relationships such as for example there are two possible relationships: a co-author relationship or adviser-advisee relationship. Person may completely un-observe any of the type of link in this case.

Another important task related to it is to get known of link purpose. Links between the web pages may occur for different reasons in web collections. Links can be for any purpose for example advertisement or even for navigation.

4.4 Predicting Link Strength

Links considers with weights associated with it. In web page collection, page ranking authority access of incoming link is interpreted by the weights of links.

4.5 Link Cardinality

Predicting the number of links involves many practical inferences. Depending on the particular domain number of links is often a proxy for some more meaningful property as described below:
In bibliographic domain, predicting the number of citations of a paper is an indication of impact of a paper. In other words papers with more citations are likely to be seminal.
In web collection predicting the number of links to a page indicates authority. Number of links from a page indicates that the particular page is a hub. The measure of page ranks is also related to number of links.
Similarly in epidemiology study, predicting the patient and the people with whom he has been is contact with is an indication of potential transmission of the disease.

4.6 Record Linkage

Another important concept is link mining is object identity uncertainty [6; 7; 8]. In many of the cases the can be redundant or duplicated and these are the practical problems which are to be eliminated and is not so easy to do so because the object may sometimes not have its unique identification for example the two different looking items may refer to the same object.

In link mining setting it is important to take into account the similarity of the objects based on their links and attributes. For example in bibliographic study, this means taking into account paper citations, note matches identified, new matches may also become apparent.

V. Case Study

Link mining technique can be applied to the real world applications and this case study shows the same how mutual information can help in exploring and detecting anomalies of real data. Data representation is a difficult task in this technique, for example graphs to visualize dataset and clustering approach. This case study focuses on the set of co-citation data as described in the below figure. The link mining technique applied to the case study goes through the various stages: data description, preprocessing, transformation, exploration, modeling, graphing, tree structure and its visualization and finally evaluation of data.
Three important link mining tasks are covered in this case study. It helps in identifying the objects, studying links between the objects, objects clustering, graph representation.

The link mining technique shows that with the data set where the anomalies detection is unknown the application of it is valid so it is necessary to demonstrate that the same technique when applied to the real world data which is inconsistent noisy and with huge volumes does it work fine or not . A person have to check whether it works well with the other clustering algorithms or not. In this case study hierarchical clustering is used to address the issue. Using bibliographic data, 5 clusters were created where the clusters in the ascending order were ranked from starting with the strongest link to the weakest link depending on the contained data. Clusters were created based on the mutual information applied to it the algorithm used the semantics of data. Cluster 5 consists of very low value of data based on mutual information.
The table shows clusters with its co-citation values link strength based on mutual information.

Clusters	Items	Mutual information
1	58	0.93
2	49	0.82
3	38	0.63
4	29	0.43
5	19	0.00

Table no 1: Result of mutual information[9]

Having clustered and visualized as well as examined the data, it was clear that the proposed approach is proved to valid when applied on data where anomalies detections are unknown. The results produced with

approach were valid. Analyzing each cluster and the elements in those clusters and their relationships were time consuming. The data may contains noise such as misspellings of authors name or journal title so because of this reason some additional information is to be added and due to this direct analysis cannot be applied and there is a need of preprocessing the data first so that no issue would arise in the later stages. Here clustering approach was used to cluster the data based on the characteristics, graph visualization and validate approach.

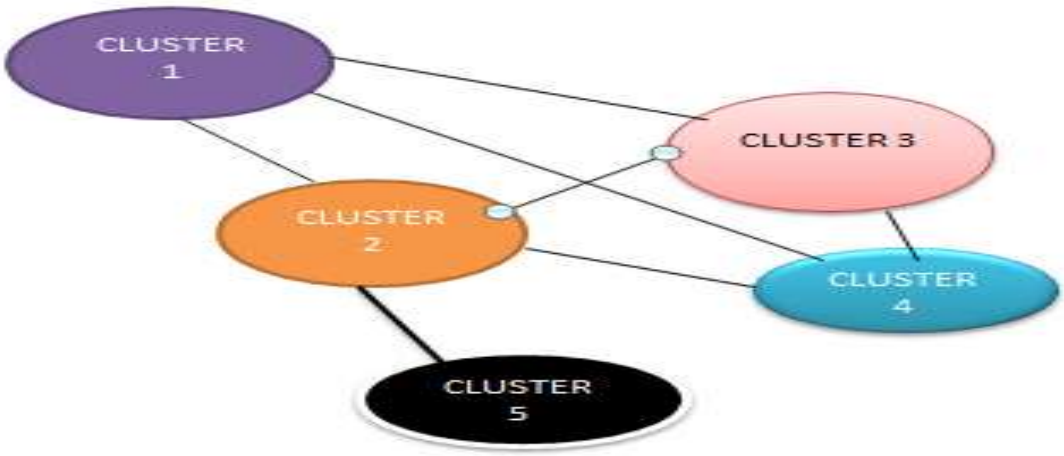

Figure 2: Mapping Nodes

Clusters were designed to classify the normal observation. The anomalies used in this case study are collective anomalies, defined by Chandola et al. as follows ”The individual data instances in a collective anomaly may not be the anomalies by themselves, but their occurrence together, as a collection is anomalous.”[1] Mutual information can be used to interpret collective anomalies. In this technique it considers data sets as a linked collection of interrelated objects so it focuses on explicit links.

The co-citation data applied hierarchical clustering where the nodes represented the authors and the link is represented by the edges between the nodes. The cluster 5 shares no link as the variables are not dependent on each other and there was zero mutual information as there was no strong links between the elements in that cluster.

VI. Applications Of Link Mining

1. Classification of Query using links: Sometimes the user may be interested in small subset of information or some particular content, so for these users providing the huge amount of data is worthless. In the traditional classification approach the whole dataset was used to be considered as a single linked instance of an object but as in some cases person may only be interested in particular research area then link mining can help in classifying the small subsets of the objects is worthy and can also help in correctly classifying the objects via structure of that particular link.
2. Web Search & Retrieval: Many times when the user searches for information on web the results retrieved is based on the evaluation of the user efforts also the length encountered and responds user with the documents now the user may come across various irrelevant documents or the list of documents retrieved after the search according to web ranking may have duplication of the same document over various sites so to overcome this problem link mining can be used to getting the information efficiently and more precisely based on user demands by the solution of constructing a semantic web where in which the information is well defined or exact in this web relationship and properties about things are two important concepts because here in this web links are not between the pages as it describes the relationships and properties between things.
3. Social bookmarking tools enables the users to save URL, share URL as well as create label for reference of web pages in this the use of link mining is to investigate users bookmarking and tagging behavior which helps is describing various approaches in finding patterns in data.

VII. Conclusion

The link mining study helped in learning the data which is linked like in artificial intelligence here the graph diagram representation of the link structure is being described where the objects are nodes of the graph and the edges as well the hyper or sub edges in the hierarchical structure are the links connected to the nodes or relation between objects. Various tasks included in this study were classification of the hypertext labeling and segmenting the data as per the mutual information and based on the same searching for more useful data by

ranking links with its discovery for information retrieval. World-wide web as well as citation such as reference to a book, article, web page or other published item are few of its including domains. Link mining challenges and some of its working area are being summarized.

The data is represented using the graphical representation and hierarchical clustering was applied in the case study. For evaluating the cluster content and its validity mutual information can be applied. The data used was the real world databases and therefore it was important to preprocess the data. For large scaled volumes of data including of noisy, redundant as well as inconsistent data and a combination of similar preprocessing this approach was developed.

Anomalies detection can be difficult problem if the relationship exists from different data points.

References

[1]. D. Jsensen. Statistical challenges to inductive inference in linked data. In Seventh International Workshop on Artificial Intelligence and Statistics,1999.
[2]. G. Chandola V., Banerjee A., and Kumar V.(2009) Anomaly Detection A. Survey, ACM. Computing Survey.
[3]. J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999.
[4]. J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. Computer Networks, 1999.
[5]. M. Bilenko and R. J. Mooney. On evaluation and training-set construction for duplicate detection.Under review.
[6]. S. Russell. Identity uncertainty.In Proc. of IFSA-01, Vancouver, 2001.
[7]. H. Pasula, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching.In Advances in Neural Information Processing Systems.MIT Press, 2003.
[8]. S. Chakrabarti. Mining the Web. Morgan Kaufman, 2002.
[9]. Dr. Zakea II-Agure and Mr. HinchamNoureddineItani. Higher Colleges of Technology, United Arab Emirates.
[10]. Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2009).