A Technical Approach on Large Data That Is Distributed Over a Network Using Link Mining

Suhasini Gadala

***Abstract*— Data Mining is nontrivial extraction of implicit, previously unknown and potential useful information from the databases. For a database with number of records and for a set of classes such that each record belongs to one of the given classes, the problem of classification is to decide the class to which the given record belongs. The classification problem is also to generate a model for each class from given data set. We are going to make use of supervised classification in which we have training dataset of record, and for each record the class to which it belongs is known. There are many approaches to supervised classification. Decision tree is attractive in data mining environment as they represent rules. Rules can readily expressed in natural languages and they can be even mapped through database access languages.**

Now a day's classification based on decision trees is one of the important problems in data mining which has applications in many areas. Database system has become highly distributed, and we are using many paradigms. We consider the problem of inducing decision tree in a large distributed network of highly distributed databases. The classification based on decision tree can be done on the existence of distributed databases in healthcare and in bioinformatics, human computer interaction and by the view that these databases are soon to contain large amounts of data, characterized by its high dimensionality. Current decision tree algorithms would require high communication bandwidth, memory, and they are less efficient and scalability reduces when executed on such large volume of data. So there are some approaches being developed to improve the scalability and even approaches to analyse the data distributed over a network.

A key challenge for data mining is tackling the problem of mining richly structured datasets, that is distributed and links between objects in some way. 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. Recently there 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. Data on the web is huge and distributed across various sites. The traditional approach is to integrate all data into one site and perform required analysis. The problem with this is its time consuming and not scalable, so we need to find more efficient algorithms to mine data that is distributed over the network.

***Index Terms*— Datamining, distributed databases,web structure mining,link mining.**

Suhasini Gadala,

I. Reason and Purpose of undertaking this project

Extracting efficient data is most important task these days, however there are many algorithms used to mine the data. But these algorithms are not sufficient to meet the scalability if the data is distributed over the network. Another important aspect lies wether datasets distributed over the network is structured or unstructured. So there are challenges in increasing the scalability and tackling the problem of mining richly structured datasets where objects are linked in some way. So we need to provide classification and clustering in linked relational domains requires new data mining algorithms. But with the introduction of links, new tasks also came 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. Link mining is a promising new area where relational learning meets statistical modelling. We believe many new and interesting machines learning research problems lie at the intersection, and it is a research area.

II. RESEARCH AIM

My research aims to know what are the algorithms used for link mining and how they can be applied on distributed data over the network and how to improve the scalability of these algorithms for efficient information retrieval.

III.LITERATURE REVIEW

According to **P.K Chan and S.J stolfo**, a **Distributed Database** is a database in which storage devices are not all attached to a common CPU. It may be stored in multiple computer located in the same physical location, or may be dispersed over a network of interconnected computers. Collections of data (e.g. in a database) can be distributed across multiple physical locations. A distributed database can reside on network servers on the internet, on corporate internet or on other company networks. The replication and distribution of databases improves database performance at end users worksites. Besides distributed database replication and fragmentation, there are many other distributed database design technologies. For example, local autonomy, synchronous and asynchronous distributed database technologies. These technologies' depend on the needs of the business and the sensitivity/confidentiality of the data to be stored in the database, and hence the price the business is willing to spend on ensuring data security and consistency [1].

In the view of **H.Karguptha, B.Park, D.Hershberger and E.Johnson, Distributed computing** plays an important role in the Data Mining process for several reasons. First, Data Mining often requires huge amounts of resources in storage space and computation time. To make systems scalable, it is important to develop mechanisms that distribute the work load among several sites in a flexible way. Second, data is often inherently distributed into several databases, making a centralized processing of this data very inefficient and prone to security risks. Distributed Data Mining explores techniques of how to apply Data Mining in a non-centralized way [2].

According to **JaiweiHan and Micheline Kamber** one interesting aspect of **Large Databases** is that they are often distributed over many locations. The main reason for this is that they are produced by a variety of independent institutions. While these institutions often allow a second party to browse their databases, they will rarely allow this party to copy them. There could be a number of reasons for this: the need to retain the privacy of personal data recorded in the database, through questions regarding its ownership, or even because the sheer size of the data makes copying non permissively costly in CPU, disk I/O or network bandwidth [3].

As said by **Amir Bar-or, Daniel keren Assaf schuter and Ran Wolff**, We require **Distributed Algorithms** for data mining over a distributed network as the data is distributed on various locations. A distributed decision tree induction algorithm is one that executes on several computers, each with its own database partition. The outcome of the distributed algorithm is a decision tree which is the same as, or at least comparable with, a tree that would be induced were the different partitions collected to a central place and processed using a sequential decision tree induction algorithm.

DHDT (Distributed hierarchical decision tree) focuses on reducing the volume of data sent from each level to the next while preserving perfect accuracy .The common approach to reduce the communication overhead would be to sample the distributed data set and transfer these decision trees to the central sites this approach is called as Meta learning. This suffers from scalability limitations. So a different Meta learning algorithm was suggested this algorithm turns each decision tree to set of rules and then merge the rules into single superset of rules. As numbers of sites increase the accuracy of Meta learning classifiers drop.

Then much attention was given to parallel algorithm .There are three parallel algorithms

1. Synchronous tree construction
2. Partitioned tree algorithm
3. Hybrid approach

These algorithms cannot be used in large scale distributed system because data movement is often impractical in distributed network. In order to split the attribute list the hash table must be available on all computing nodes which make certain algorithms highly unscalable [4].

According to **J.Kleinberg**, to improve the information retrieval system we need to exploit the link structure which makes use of links. There are well known algorithms for the link structure they are page ranker and hubs and authority score. These algorithms are based on citation relationship between pages. There are algorithm based on co- citation to find relation between web pages and finger grained representation of web pages [5].

According to **Jaideep Srivastava**, Web is collection of inter related files on one or more web servers. Web mining is an application of data mining techniques to extract knowledge from web data or to find patterns on data downloaded from web page. Web data may be **web content, web structure and web usage.** Based on the main source of data type use, these techniques can be broadly classified as Web content mining, weblink mining and web usage mining.

Web content:

- Extract useful information from the content of web document.

Web structure:

- It is the process of discovering structure information from web page.

Web usage:

- User identification, session creation, robot detection and filtering, and extracting usage path

IV. WEB STRUCTURE MINING

This type of mining can be performed at (intra-page) document level or at the (inter-page) hyperlink level. Hyperlinks serve for two purposes pure navigation and point to pages with a set of idea or statements supporting the same topic of the page containing the links. This can be used to retrieve useful information from the web such as quality of web page, interesting web structure, web page classification, which pages to crawl, finding related pages, detection of duplicated pages [6].

In the view of **S.Chakrabarthi, B.Dom and P.Indyk**, a closely related area to this is hypertext and web page classification. A hypertext is a collection of highly structured data which should be exploited to improve the accuracy of classification. Hypertext has both incoming and outgoing links. The traditional information retrieval document model do not make full use of hypertext link structure, so in the web page classification problem a web page is viewed as large directed graph. So the objective here is to label the category of web page based on the features of current page .By this we can achieve better categorization of results. A number of algorithms were introduced but they could improve the efficiency of linked documents but they were not effective for linked neighbours. These algorithms were based on words [7].

According to **S.Slattery and M. Carven**, they have suggested another approach to hypertext and link mining combines the techniques from logic programming with statistical learning algorithm to construct features from related documents. There are numbers of algorithms developed to improve learning. These algorithms went beyond the words. They considered the anchor text, neighbouring text, capitalised words and alphanumeric words [8]. They introduced regression logistic model to improve the accuracy of document mining.

As said by **Y.Yang, S.Slattery, and R.Ghani**, there is another approach to link mining that is finding regularities in the hyper links i.e. encyclopaedic regularities which mean identifying links that belong to same class and co-citation regularities which means to identify linked objects that do not belong to same class but objects sited by the same object belong to same class. A number of investigations have been done on the data sets, but they concluded that usefulness of regularities varied depending on the datasets and classifies being used [9].

In the view of **J.Kubica, A.Moore, J.Schneider and Y.Yang**, there is another area that took the impact in link mining is identifying communities and groups. For this a probabilistic model for link detection and modelling groups was proposed which makes use of demographic information and linkage information to infer group membership [10].

According to **S.Dzeroski, N.Lavrac and D.Cook and L.Holder**, Tradition machine and data mining approaches assume random sampling of homogeneous objects from single relation. But real world data sets are multi relational, heterogeneous and semi structured. Link mining is an intersection of social networks, link analysis, hypertext and web mining, relational learning and inductive logic programming [8] and graph mining [11].

According to **Miguel Gomes da Costa Junior Zhiguo Gong**, Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents .The objects in the WWW are web pages, and links are in, out and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks.

V. LINKED DATA

It is a collection of heterogeneous data . heterogeneous data is the combination of objects and links .Object may have multiple relationships. Nodes represent objects [12].

An Object may

- Be of different kinds.
- Have attributes.
- Be a label or a class.

Edges are represented as link.

- They may be directed
- They can also have attributes.

Link mining tasks which are applicable in web structure mining are summarised as follows:

- ➢ Link based object classification.
- ➢ Link type prediction.
- ➢ Predicting link existence.
- ➢ Link cardinality estimation.
- ➢ Object identification.
- ➢ Identification of sub graph.

There are two ranking algorithms HITS concept and page rank method both focus on link structure of the web to find importance of web page.

Although HITS provides good search results for a wide range of queries, HITS did not work well in all cases [13].

Page rank: **L. Page and S. Brin** proposed that the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. The limitation with this is Page Rank algorithm needs a few hours to calculate the rank of millions of pages [14].

ZivBarYossef and Sridhar Rajagopalan categorized the algorithms, which use links, as follows:

- Relevant Linkage Principle: Links points to relevant resources.
- Topical Unity Principle: Documents often co-cited are related, as are those with extensive bibliographic overlap. This idea is previous addressed by Kessler for bibliographic information retrieval.
- Lexical Affinity Principle: Proximity of text and links within a page is a measure of the relevance of one to another [15].

According to **Ziv Bar-Yossef, Deng Cai, Shian-Hua Lin**, Even though those link algorithms can always provide a good support for Web information retrievals, clustering and knowledge discoveries on the Web, authors also find problems associated with those technologies. There are many techniques invoved in improving the efficiency on information retrieval from web document but have their own limitations [16].

CONCLUSION

Data on the web is huge and distributed across various sites. The traditional approach is to integrate all data into one site and perform required analysis. The problem with this is its time consuming and not scalable, so we need to find more efficient algorithms to mine data that is distributed over the network.

REFERENCES

[1]P.K Chan and S.j stolfo,"towards parallel and distributed learning by Meta learning," Working notes AAAI works, Knowledge discovery in data bases".

[2]H.Karguptha, B Park, D Hershbereger and E.johnson, collective data mining: A new perspective towards distributed data mining.

[3]JiweiHan and Micheline Kamber:"Data mining concepts and techniques".

[4]Amir Bar-or, Daniel keren, Assaf schuter and Ran wolff" Hierarchical decision Tree induction In distribute genomic data base": IEEE transaction on knowledge and data engineering.

[5]J.Kleinberg.Authoritative source in hyperlinked environment. Journal of ACM, 40(3):(63-65), 1997.

[6]Jaideep srivastava, Accomplishment and future direction.

[7]S.Chakrabarthi, B.Dom and P.Indyk. Enhanced hypertext categorization using hyperlinks. In Proc .of SIGMOD-98, 1998.

[8]S.Slattery and M. Carven.Combining statistical and relational methods for learning in hypertext domains. In Proc. Of ILP-98,1998

[9]Y.Yang, S.Slattery, and R.Ghani.A study of approaches to hypertext categirization. Journal of intelligentInformation system, 18(2-3):219-241, 2002.

[10]J.Kubica,A.Moore,J.Schneider, an Y.Yang. Stochastic link and group detection. In proc .of AAAI-02, 2002.

[11].Dzeroski and N.Lavrac.,editors. Relational datamining.Kluwer,Berlin,2001.

[12]L. Getoor, Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003.

[13]http://www.research.com Last accessed 15/04/2005.

[14]http://www.google.com Last accessed 15/04/2005.

[15]Ziv Bar-Yossef and Sridhar Rajagopalan.Template Detection via DataMining and its applications. In proceedings of WWW2002, May7-11, 2002, Honolulu,Hawaii,USA.580-591.

[16]Sian-Hua Lin and Jan-Ming Ho. Discovering Information Content Blocks from Web Documents. In: proceedings of ACM SIGKDD' 02,July 23-26,2002, Edmonton, Alberta,Canada. 588-593.

[17]Deng Cai, XiaofeiHe,Ji-RongWen, and Wei-Ying Ma,Block-level Link Analysis,In Proceedings of ACMSIGIR'04, July25-29, 2004, Sheffield,SouthYorshire,UK. 440-447.

[18]T.Hofmann. Probabilistic latent semantic indexing. In proceedings of the 22nd Annual ACM conference onresearch and development in Information retrieval, pages (50- 57),Berkley California,August 1999.

[19]S.Sarawagi and A.Bahmidipaty. Interactive duplication using active learning.In proceedings of the eigth ACM SIGKDD international conference on knowledge discovery and data mining.

[20]Eirinaki M., Vazirgiannis M. (2003). Web mining for web personalization. ACM On Internet Technology (TOIT), 3(1), 1-27.