LINK ANALYSIS IN DATABASES

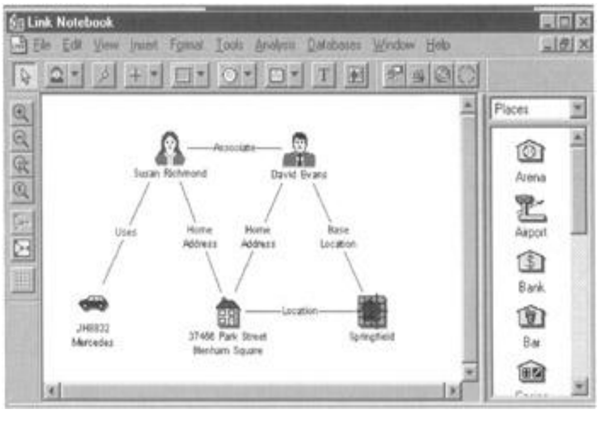
A. Yeshvini, Kanika Arora
Btech, Dronacharya College Of Engineering, Gurgaon, India

Abstract- **The data keep in info has several forms and relationships with one another. Link analysis is best technique utilized in data processing. It identifies the connection pattern in knowledge. during this paper we wish to indicate the connection between the databases and the way they're shaped with totally different varieties of association and multiplicity rules. we've got taken the survey of various algorithms and their operating.**

***Index Terms*— Data mining, Association rules, Multiplicity, Localised Links.**

I. INTRODUCTION

The process of information mining derived helpful, valid, authentic and connected info from massivedatabases. It's a very important tool for looking massive store of information to depict specific relationships, patterns and trends in existing information. Data processing is additionally referred to as information Discovery in information (KDD). The patterns can't be discovered just byinformation exploration as a result of relations between information square measure terriblyadvanced. Thus, we have a tendency to need some extra analytic tool to perform information analysis.



Link Analysis is a technique that identifies the relationships as well as connections between data groups. It is derived from *graph theory[1]*. Graphs are used to represent relationships. These are useful in both computer science and mathematics for developing algorithms that explore these relations. It consists of two components: *nodes* (data variables that possess relation with others) and *edges* (pair of nodes connected by a relation). Based on graph theory, Link analysis have proved to be efficient in solving problems like analyzing the pattern of telephone calls, state transition diagrams, analyzing links on web (search engines, site maps),traffic monitoring, decision support, fraud systems.

II. LINK ANALYSIS APPROACHES

the goal of this technique is to detect relationships between items.[2] It enable the analysts and researchers to uncover hidden patterns in large data sets, such as "customers who order product A often also order product B or C" or "employees who said positive things about initiative X also frequently complain about issue Y but are happy with issue Z."

A. Association Rules

The utility of this method is to handle distinctive data processing issues.[2] Suppose you'recollection knowledge at a check-out till during a giant book store. every client group action is logged during a info, and consists of the titles of the books purchased by the client, magazine titles, gift items, etc. Hence, every record within the info can represent one client (transaction), might|and should|and will} accommodates one book purchased by that client or may accommodates many alternative things (perhaps hundreds) that were purchased, organized in AN arbitrary ordercounting on the order within which the various things fell the transporter at the till.

The purpose of the analysis is to seek out associations between the things that were purchased, i.e., to derive association rules that establish things|the things} and co-occurrences of various items thatseem with greatest frequencies. for instance, you wish to find out that books square measurepossible to be purchased by a client World Health Organization you recognize already purchased (or is getting ready to purchase) a specific title. this sort of data might then quickly be accustomedcounsel to the client those extra titles. you will already be at home with the results of those forms of analyses if you're a client of varied on-line (Web-based) retail businesses; again and again oncecreating a buying deal on-line, the seller can counsel similar things at the time of "check-out,"supported rules like "customers World Health

Organization purchase book title A are possible to get book title B," so on.

B. Association Rule

We use Association Rules to seek out rules of the sort If X then (likely) Y wherever X and Y may besingle values, items, words, etc., or conjunctions of values, items, words, etc. (e.g., if (Car=Porsche and Gender=Male and Age<20) then (Risk=High and Insurance=High)).[3] The program may beaccustomed analyze easy categorical variables, divided variables, and/or multiple response variables. The algorithmic rule can verify association rules while not requiring the user to specify the quantityof distinct classes gift within the knowledge, or any previous data concerning the utmost factorial degree or complexness of the necessary associations. In a sense, the algorithmic rule canconstruct cross-tabulation tables while not the necessity to specify the quantity of dimensions for the tables, or the quantity of classes for every dimension. Hence, this method is especially compatiblefor knowledge and text mining of big databases.

C. Multilevel Association Rules

For many applications, it's troublesome to search out robust associations among knowledgethings at low level of abstraction because of the scantiness of knowledge in dimensional area.[4]data processing systems ought to give capabilities to mine association rules at multiple levels of abstraction and traverse simply among completely different abstraction areas.

Using uniform minimum support for all levels (referred to as uniform support): a similar minimum support threshold is employed once mining at every level of abstraction. once a homogenous minimum support threshold is employed, Difficulties in Uniform Support

- It is unlikely that items at lower levels of abstraction will occur as frequently as those at higher levels of abstraction.
- If the minimum support threshold is set too high, it could miss several meaningful associations occurring at low abstraction levels.
- If the threshold is set too low, it may generate many uninteresting associations occurring at high abstraction levels.

Using reduced minimum support at lower levels (referred to as reduced support): Each level of abstraction has its own minimum support threshold. The lower the abstraction level is, the smaller the corresponding threshold is. There are a number of alternative search strategies for this approach.

1. *Level-by-level independent:* This is a full breadth search, where no background knowledge of frequent item sets is used for pruning. Each node is examined, regardless of whether or not its parent node is found to be frequent.
2. *Level-cross filtering by single item:* An item at the i^{th} level is examined if and only if its parent node at the $(i-1)^{th}$ level is frequent. If a node is frequent, its children will be examined; otherwise, its descendents are pruned from the search.
3. *Level-cross filtering by k-item set*: A k-item set at the i^{th} level is examined if and only if its corresponding parent k-item set at the $(i-1)^{th}$ level is frequent.

III. LINK ANALYSIS ALGORITHM

There are number of algorithms proposed based on link analysis. Three important algorithms Page Rank[5], Weighted PageRank [6] and HITS (Hyper-link Induced Topic Search)[7] are discussed below as follow.

A. Page Rank

This algorithmic program was developed by Brin University that extends the concept of citation analysis [5]. Page Rank provides a higher approachthat may figure the importance of online page by merely count the amount of pages that area unit linking thereto. These links area unit referred to as as back-links. If a back link comes from a very important page than this link is given higher weight age than those that area unitcoming back from non-important pages. The link from one page to a different is taken into accountas a vote. Page and Brin planned a formula to calculate the PageRank of a page A as explicit below

$$PR(A)= (1-d)+d(PR(T1)/C(T1)+.....+PR(Tn/C(Tn)) \quad ....(1)$$

Where *PR (Ti)* is the PageRank of the Pages Ti which links to page A, *C (Ti)* is number of out links on page Ti and *d* is Damping factor. It is used to stop other pages having too much influence. The total vote is “damped down” by multiplying it to 0.85.

Secrets of Page Rank:

- The PageRank forms a probability distribution over the web pages so the sum of Page Ranks of all web pages will be one.
- The PageRank of a page can be calculated without knowing the final value of Page Rank of other pages.
- It is an iterative algorithm which follows the principle of normalized link matrix of web.

B.Strenghts of page rank algorithm

The strengths of Page Rank algorithm are as follows:

1. **Less Query time cost:** Page Rank has a clear advantage over the HITS algorithm, as the query-time cost of incorporating the pre-computed Page Rank importance score for a page is low.

2. **Less susceptibility to localized links:** Furthermore, as Page Rank is generated using the entire Web graph, rather than a small subset, it is less susceptible to localized link spam [6].

3. **More Efficient:** In contrast, Page Rank computes a single measure of quality for a page at crawl time. This measure is then combined with a traditional information retrieval score at query time. Compared with HITS, this has the advantage of much greater efficiency [10].

4. **Feasibility:** As compared to Hits algorithm the Page Rank algorithm is more feasible in today's scenario since it performs computations at crawl time rather than query time.

Limitations of Page Rank algorithm

The following are the problems or disadvantages of Page Rank:

1. **Rank Sinks**: The Rank sinks problem occurs when in a network pages get in infinite link cycles.

2. **Spider Traps**: Another problem in PageRank is Spider Traps. A group of pages is a spider trap if there are no links from within the group to outside the group.

3. **Dangling Links**: This occurs when a page contains a link such that the hypertext points to a page with no outgoing links. Such a link is known as Dangling Link.

4. **Dead Ends**: Dead Ends are simply pages with no outgoing links.

5. Page Rank doesn't handle pages with no out edges very well, because they decrease the Page Rank overall.

6. **Circular References:** If you have circle references in your website, then it will reduce your front page's Page Rank [11].

IV. WEIGHTED PAGE RANK

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of Page Rank algorithm[7]. It assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among it's outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks).This is denoted as $W^{in}(m,n)$
and $W^{out}(m,n)$ respectively. $W^{in}(m,n)$ is the weight of link(m,n) as given in (2).

$$W^{in}_{(m,n)}=I_n / \Sigma_{p\varepsilon R(m)} I_p \qquad \ldots(2)$$

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p}$$

I_n is number of incoming links of page n, Ip is number of incoming links of page p, R(m) is the reference page list of page m. $W^{out}(m,n)$ is the weight of link(m,n)as given in (3). It is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m.

$$W^{out}_{(m,n)} = O_n / \Sigma_{q\varepsilon R(m)} O_p \quad \ldots(3)$$

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p}$$

On is number of outgoing links of page n, Op is number of outgoing links of page p,

Then the weighted PageRank is given by formula in (4)

$$WPR(n) =(1-d)+d\ \Sigma_{m\varepsilon B(n)} WPR(m)\ W^{in}_{(m,n)} W^{out}_{(m,n)} \ldots(4)$$

A. HITS Algorithm

Kleinberg [7] developed a WSM based algorithm named Hyperlink-Induced Topic Search (HITS) which presumes that for every query given by the user, there is a set of authority pages that are relevant and accepted focusing on the query and a set of hub pages that contain useful links to relevant pages/sites including links to many authorities. Kleinberg states that a page may be a good hub and a good authority at the same time. This spherical relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Search).

[8]Steps in HITS algorithm:

- Sampling Step:- In this step a set of relevant pages for the given query are collected.

- Iterative Step:- In this step Hubs and Authorities are found using the output of sampling step.

The expressions (7,8) are used to calculate the weight of HUB(H_p) and weight of Authority (A_p).

$H_{p=} \Sigma_{q\varepsilon I(p)} A_q$ ….(7)

$A_{p\,=} \Sigma_{q\varepsilon B(p)} H_q$ ...(8)

Where Hp is the hub weight, Ap is the Authority weight, I(p) and B(p) denotes the set of reference and referrer pages of page p. The page's authority weight is proportional to the sum of the hub weights of pages that it links to it; similarly, a page's hub weight is proportional to the sum of the influence weights of pages that it links to.

Advantages of HITS

We list below a few considerable advantages of HOTS:
1. HITS scores due to its ability to rank pages according to the query string, resulting in relevant authority and hub pages.
2. The ranking may also be combined with other information retrieval based rankings.
3. HITS is sensitive to user query (as compared to PageRank).
4. Important pages are obtained on basis of calculated authority and hubs value.
5. HITS is a general algorithm for calculating authority and hubs in order to rank the retrieved data.
6. HITS induces Web graph by finding set of pages with a search on a given query string.
7. Results demonstrate that HITS calculates authority nodes and hub ness correctly.

Drawbacks of HITS algorithm

Some notable drawbacks of HITS[10] algorithm are:

1. **Query Time cost:** The query time evaluation is expensive. This is a major drawback since HITS is a query dependent algorithm.

2. **Irrelevant authorities:** The rating or scores of authorities and hubs could rise due to flaws done by the web page designer. HITS assumes that when a user creates a web page he links a hyperlink from his page to another authority page, as he honestly believes that the authority page is in some way related to his page (hub).

3. **Irrelevant Hubs:** A situation may occur when a page that contains links to a large number of separate topics may receive a high hub rank which is not relevant to the given query. Though this page is not the most relevant source for any information, it still has a very high hub rank if it points to highly ranked authorities.

4. **Mutually reinforcing relationships between hosts:** HITS emphasizes mutual reinforcement between authority and hub webpages. A good hub is a page that points to many good authorities and a good authority is a page that is pointed to by many good hubs.

5. **Topic Drift:** Topic drift occurs when there are irrelevant pages in the root set and they are strongly connected. Since the root set itself contains non-relevant pages, this will reflect on to the pages in the base set. Also, the web graph constructed from the pages in the base set, will not have the most relevant nodes and as a result the algorithm will not be able to find the highest ranked authorities and hubs for a given query.

6. **Less Feasibility:** HITS invokes a traditional search engine to obtain a set of pages relevant to it, expands this set with its inlinks and outlinks, and then attempts to find two types of pages, *hubs* (pages that point to many pages of high quality) and *authorities* (pages of high quality)[10]. Because this computation is carried out at query time, it is not feasible for today's search engines, which need to handle tens of millions of queries per day [10].

V. COMPARISON OF DIFFERENT ALGORITHM[13][14][15][16]

Algorithm	Page Rank	HITS	Weighted Page Rank
Mining Technique	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Structure Mining
Methodology	It computes the number of pages at the time of indexing.	It calculates the hubs and authority of relevant pages.	Weight of web pages is calculated from input and output links.
Application	It is applied to the entire	The basic aim is to find the set of	It is also applied to the entire

	web.	web pages relevant to our search topic.	web.
Query Dependency	It does not depend upon query.	It depends upon query.	It also depends upon query.
Quality of result	Medium	Less than Page Rank	More than Page Rank
Preference	High, as it consider the back links	Moderate, Hubs and authorities are considered.	High, as it sort the pages according to its importance.
Complexity	O(log N)	<O(log N)	<O(log N)
Limitation	Results come at the time of indexing and not at the query time.	Topic drift and efficiency problem.	Relevancy is ignored.

A. Link analysis application

Link analysis can also be used for deciding which web pages to add to the collection of web
pages, i.e., which pages to crawl. A *crawler* (or *robot* or *spider*) performs a traversal of the web graph with the goal of fetching high-quality pages. After fetching a page, it needs to decide which page out of the set of uncrawled pages to fetch next. One approach is to crawl the pages with highest number of links from the crawled
pages first.[17]

B. Advantages /strengths of link analysis

- It capitalizes on relationships
- It is useful for visualization
- It creates derived characteristics that can be used for further mining
- Links are a very natural way to visualize some types of data.
- Link analysis can lead to new and useful data attributes. Examples include calculating an authority score for a page on the World Wide Web and calculating the sphere of influence for a telephone user.
- It uses mathematical and statistical calculations to find new patterns, depict future trends and therefore assist in decision making.

B. Disadvantages of Link Analysis

- It is not appropriate for all types of problems.
- Unlike neural network it input the data and produces the result.
- Many types of data are simply not appropriate for link analysis. Its strongest use is probably in finding specific patterns, such as the types of outgoing calls, which can then be applied to data. These patterns can be turned into new features of the data, for use in conjunction with other directed data mining techniques.

REFERENCES

1) http://engineering inventions.blogspot.in/2011/03/data-mining-techniques-link-analysis.html, last visited on 05August,2013

2) https://sites.google.com/site/assignmentssolved/mca/semester6/mc0088/20 last visited 05 August, 2013.

3) http://www.statsoft.com/textbook/association-rules/

4) J Han and M. Kamber,1998 Link analysis book page 14

5) S. Brin, and L. Page, The Anatomy of a Large Scale, Hypertextual Web Search Engine,, Computer Network
and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117,
1998.

6) Wenpu Xing and Ali Ghorbani, Weighted PageRank
Algorithm, Proceedings of the Second Annual
Conference on Communication Networks and Services
Research (CNSR ’04), IEEE, 2004.

7) J. Kleinberg, Authoritative Sources in a Hyper-Linked
Environment, Journal of the ACM 46(5), pp. 604-632,
1999.

8) *International Journal of Computer Applications (0975 – 8887)*
Volume 13– No.5, January 2011

9) International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, ISSN: 2278-0181October - 2012

10) The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank Matthew Richardson Pedro Domingos Department of Computer Science and Engineering University of Washington Box 352350 Seattle, WA 98195-2350, USA *{mattr, pedrod}@cs.washington.edu*

11) International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181

12) Google and the Page Rank Algorithm, slides by Székely Endre 2007. 01. 18.

13) *International Journal of Computer Applications (0975 – 8887)*

Volume 13– No.5, January 2011

14) Dilip Kumar Sharma et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676

15) International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-1, June 2012

16) International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181

17) 8.a - J. Cho, H. Garc´ıa-Molina, and L. Page. Efficient crawling through URL ordering. In *Proceedings of the Seventh International World Wide Web Conference* 1998, pages 161–172.

18) http://engineering-inventions.blogspot.in/2011/03/data-mining-techniques-link-analysis.html