ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

A PARTICULARIZED DELIBERATION ON HYPERLINK ANALYSIS

Tamanna Jain

M.Tech. (CSE) Student

Maharishi Markandeshwar University, Mullana, Haryana, India

ABSTRACT

The onset of World Wide Web has devastated the typical home computer user with an immense and prodigious flood of information. To be able to grapple with the abundance of available information, users of www need to lean on intelligent tools that comfort them in discovering, categorizing and purifying the available information. Web mining, an application of data mining, is used to find the content of the Web, the user's behaviour in the past, and the web pages that the users want to view in the future. Web structure mining is one of the mining techniques through which one can perform the extraction of the desired data by understanding the structure of web and the relationship between the links and web pages. This manuscript exemplifies theweb search via link analysis and depiction of Page Rank and HITS algorithmthat are commonly used in web structure mining.

Keywords: Link Analysis, Page rank, HITS, Web Search, Web Structure Mining.

1. INTRODUCTION

Web mining is the process of extracting the useful information from the huge accumulation of information. One can find anything on the World Wide Web within fraction of seconds by simply entering the query on the search engine but to find the exact and appropriate data from the search results may be time consuming and irritating for the user. Normally the user bothers the above four or five links of his interest and if he didn't get the relevant information then he simply switches himself to some other search engines. The order in which the search results are shown to the user depends on certain factors, like the ranking of the page or the latest updated information or the results having keywords that best match with the query submitted by the user. These factors decide the relevancy of the web pages and the associated links.

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

Web structure mining is used to classify the relationship between Web pages linked by information or direct link connection. This connection permits a search engine to bring the data that is related to the query directly to the linking web page from the web site where the data resides.The goal of the structure mining is to generate the previously unknown relationships between the web site and web pages. Many datasets are best described as a linked collection of inter related objects. These may represent:

- Homogeneous networks
    - Single object type
    - Single link type
    - Single model social networks (eg, friends)
    - www, a collection of linked web pages
- Heterogeneous networks
    - Multiple object types
    - Multiple link types
    - Medical network: patients, doctors, disease
    - Bibliographic network: publications, authors, venues

Structure mining minimizes two main problems of the World Wide Web.

a) Irrelevant search results

b) Inability to index the information

2. LINK ANALYSIS

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

Link mining is a newly emerging research area that is at the intersection of the work in the link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. The term **link mining** is used to put a special emphasis on the links.

Links or generally relationships, among data instances are omnipresent. These links usually reveal patterns that can indicate properties of the data instances such as the importance, rank or category of the object. In some cases, not all the links will be observed; therefore, we may be interested in predicting the existence of links between instances. In some domain, our goal may be to predict whether a link will exist in the future, given the previously observed links.

2.1 LINK MINING TASKS

There are some important link mining tasks that are shown in fig 1.1.

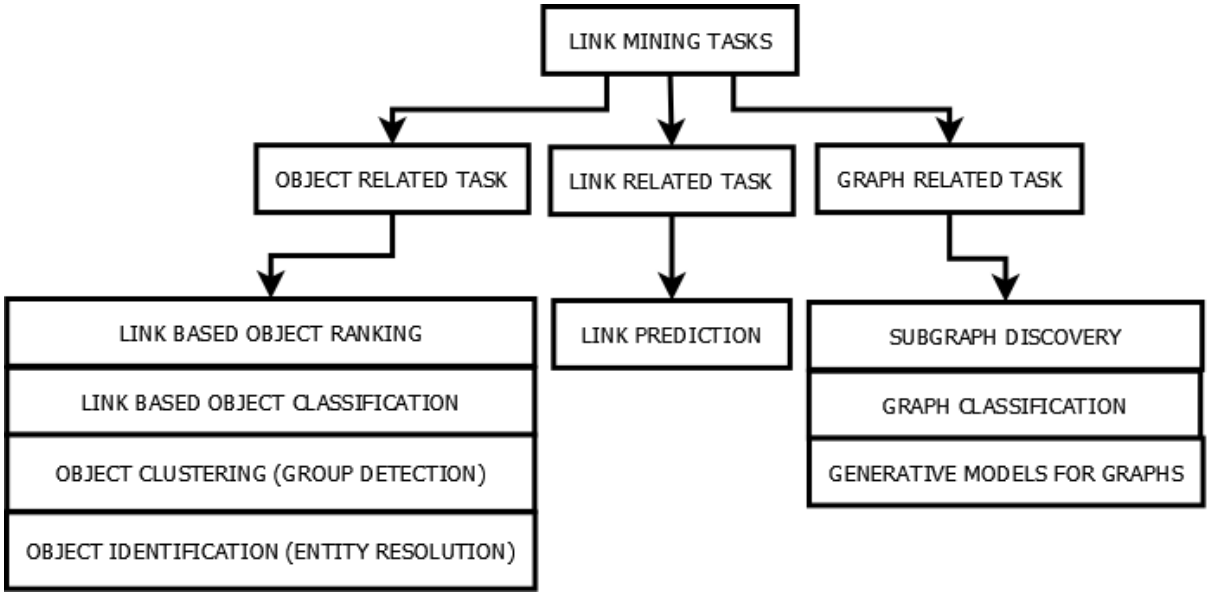

Fig 1.1 Link mining tasks

- ***LINK BASED OBJECT RANKING***

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

- It exploits the link structure of a graph to prioritize the set of objects within the graph.
- It is focused on the graphs with single object type and single link type.
- Typical link based object ranking LBR approaches are Page rank and HITS algorithm.
- In a social network analysis, the main objective of the LBR is to rank individuals in terms of 'centrality'.

- ***LINK BASED OBJECT CLASSIFICATION***
  - It predicts the category of an object based on its attributes, its links and the attributes of linked objects.
  - LBC have many application areas as web, citation, epidemics and communication.
  - In web, it predicts the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags etc.
  - In citation, it predicts the topic of a paper, based on word occurrence, citations and co-citations.
  - In epidemics, it predicts disease type based on characteristics of the patients infected by the disease.
  - In communication, it predicts whether a communication contact is by email, phone call or mail.

- ***OBJECT CLUSTERING (GROUP DETECTION)***
  - Group detection is the clustering of nodes in the graph into groups that share common characteristics.
  - It is used to identify communities in the web.
  - Methods for performing the task:
    - Hierarchical structure

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

- Block modelling of SNA
- Spectral graph partitioning
- Stochastic block modelling
- Multi-relational clustering

- ***OBJECT IDENTIFICATION (ENTITY RESOLUTION)***
  - Also called de-duplication, reference reconciliation, co-reference resolution, object consolidation.
  - It involves the predicting when two objects are same, based on their attributes.
  - Applications: web, citation, epidemics and biology.
- ***LINK PREDICTION***
  - It is the process of discovering whether a link exists between two entities, based on the attributes and other observed links.
  - Techniques used in the link prediction:
    - Often viewed as a binary classification problem.
    - Local conditional probability model, based on structural and attribute features
    - Collective prediction, eg, Markov random field model
- ***SUBGRAPH DISCOVERY***
  - This work attempts to find interesting or commonly occurring sub graphs in a set of graphs.
  - It basically finds the characteristic sub graphs.
- ***GRAPH CLASSIFICATION***

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

- 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 the 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.

- ***GENERATE MODELS FOR GRAPHS***
  - Generative models for a range of graph and dependency types have been studied extensively in the social network analysis community.
  - For directed graphs with homogeneous networks, there are several major classes of random graph distribution as Bernoulli graph distributions, conditional uniform graph distributions, p* models etc.
  - *Bernoulli graphs* are by far the simplest generative models. They assume that the random variables $\{l_{i,j}\}$ that indicate the existence of directed edges among the objects o_i and o_j are IID (independent, identically distributed).
  - *Conditional uniform graph distribution* defines uniform distributions over set of graphs with specified structural characteristics.
  - *P* models* assume that the links sharing at least one object in common are dependent.

3. WEB SEARCH

Web mining is the pertinence of the data mining techniques in search engines. It automatically discovers and extracts information from web documents. Web structure mining, a category of web mining process uncovers useful data from hyperlinks. The main aim of the mining is to find and extract relevant information that is hidden in the web-related data. It is required to convert web data into web

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

knowledge. Search engines use the concept of *web crawlers* to provide local access to the most recent versions of possibly all web pages. Web crawlers collectall web documents by browsing the web systematically and exhaustively. The region of the web to be crawled can be specified by using the URL structure.

3.1 **INDEXING AND KEYWORD MATCHING**

There are two types of data:

a) *Structured data* have keys associated with each data item that reflect its content.

b) Content-based access to *unstructureddata* without considering the meaning is the keyword search approach.

To facilitate the process of matching keywords and documents, some pre-processing steps are taken first:

i) Documents are tokenized.

ii) Characters are converted to upper or lower case.

iii) Words reduced to canonical form.

iv) Stop words are usually removed.

3.2 **DATA REPRESENTATION**

Data representation for linked data is complex than the traditional machine learning algorithms. Consider a simple example of social network epitomizing actors and their coalition in events. Such networks are commonly called *affiliation networks*, and are easily represented by three tables depicting the actors, events and participation relationships. This simpler structure can further be represented as the most natural representation, i.e., the bipartite graph where we have a set of actor nodes, a set of event nodes and edges representing an actor's participation in the event. Thus we may develop a network in which the actors are the *nodes* and the *edges* correspond to the actors who

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

have participated in an event together. As a result, it is concluded that the proper representation is really an important issue in effecting web structure mining.

3.3 HYPERLINK ANALYSIS

Many web pages do not include words that are explanatory of their basic purpose and their exist web pages which contain very little text such as image, music, video etc., making a text-based search techniques difficult. This type of characterization is included in the text that surrounds the hyperlink pointing to the page. There are a number of algorithms proposed on the basis of link analysis. These algorithms are simple and deeper relationships among the pages cannot be discovered.

Hyperlink analysis provides a means for judging the quality of pages. These algorithms make either one or both of the facilitating hypotheses:

A) A hyperlink from page 1 to page 2 is a reference of page 2 by the author of page 1.

B) If page 1 and page 2 are connected by a hyperlink, then they might be on the same topic.

The two main uses of hyperlink analysis in web information retrieval are: **crawling** and **ranking**.

Crawling is the process of collecting web pages. The process usually starts from a set of source web pages. The web crawler follows the source page hyperlinks to find more web pages. This process is repeated on each new set of pages and continues until no more new pages are discovered. The crawler has to decide in which order to collect hyperlinked pages that have not yet been crawled. ***Ranking*** is the process of ordering the returned documents in decreasing order of relevance, i.e, best answers are on the top. Ranking that uses hyperlink analysis is called *connectivity ranking*.

When a user sends a query to a search engine, the search engine returns the URLs of documents matching all or one of the terms, depending on both the query operator and the algorithm used by the search engine.

4. PAGE RANK ALGORITHM

S.Brin and L.Page, the developers of Page Rank algorithm at Stanford University extends the idea of citation analysis. In citation analysis, the incoming links are treated as citations but the idea was unsuccessful because it gives some approximation of importance of page.

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

DEFINITION: We assume that page A has T_1……..T_n pages which point to it (i.e, are citations). The parameter d is a damping factor, which can be set between 0 and 1 (usually set to 0.85). The page rank of a page A is given as follows:

$$PR(A) = (1-d)+d[PR(T_1)/C(T_1)+……..+PR(T_n)/C(T_n)]$$

Page rank forms a probability distribution over web pages so the sum of all pages' Page rank will be one.

4.1 Glossary

1. Page Rank: The actual, real, page rank for each page as calculated by Google.

2. Toolbar Page Rank: The page rank displayed in the Google toolbar in the browser. This ranges from 0 to 10.

3. Backlink: Page Rank counts the number of pages that are linking to it. These links are called backlinks.

4. Vote: The link from one page to another is considered as a vote.

5. Rank Sink: The rank scores of pages of a website could be calculated iteratively starting from any web page. Within a website, two or more pages might connect to each other to form a loop. If these pages did not referred to by other web pages outside the loop, they would accumulate rank but never distribute any rank. This scenario is called a rank sink.

6. PR (T_n): Each page has a notion of its own self-importance. That's PR(T_1) for the first page in the web all the wayupto PR(T_n) for the last page.

7. C (T_n): Each page spreads its vote out evenly amongst all of its outgoing links. The count, or number, of outgoing links for page 1 is C(T_1), C(T_n) for page n, and so on for all pages.

8. PR(T_n)/C(T_n): If a page (say page A) has a backlink from page then the share of the vote page A will get is PR(T_n)/C(T_n).

9. Damping factor d: The probability at each page the "random surfer" will get bored and request another random page. Usually this value is set to be 0.85.

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

5. HITS ALGORITHM

HITS algorithm – *Hyperlink-Induced Topic Search* – is a link analysis algorithm that rates web pages, developed by Jon Kleinberg. He suggested that there are two types of pages that could be pertinent for a query:

a) Authorities are the pages that contain useful information about the query topic. It is a valuable and informative webpage usually pointed to by a large number of hyperlinks.

b) Hubscontain pointers to good information sources.It is a webpage that points to many authority pages.

Both types of pages are typically connected through mutually reinforced relationship: good hubs contain pointers to many good authorities and good authorities are pointed to by many good hubs.It has two steps:

1) Sampling Step: in this step a set of relevant pages for the given query are collected.

2) Iterative Step: in this step hubs and authorities are found using the output of sampling step.

Numerically,

“If p points to many pages with large x-values, then it should receive a large y-value; if p is pointed to by many pages with large y-values, then it should receive a large x-value.”

Given weights $x^{<p>}$, $y^{<q>}$, then the x-weights and y-value are as follows:

$$X^{<p>} \leftarrow \sum_{q:\,(q,\,p)\,\epsilon\,E} y^{<q>} \qquad\qquad Y^{<p>} \leftarrow \sum_{q:\,(q,\,p)\,\epsilon\,E} x^{<q>}$$

An authority pointed to by several highly scored hubs should be a strong authority while a hub that points to several highly scored authorities should be a popular hub.

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

The algorithm calculates score without indexing and operates on small graphs representing a linkage between hub and authority web sites. It is used in multiple environments from institutes to search engine crawlers.The fig 5.1 shows some of the properties of HITS algorithm.

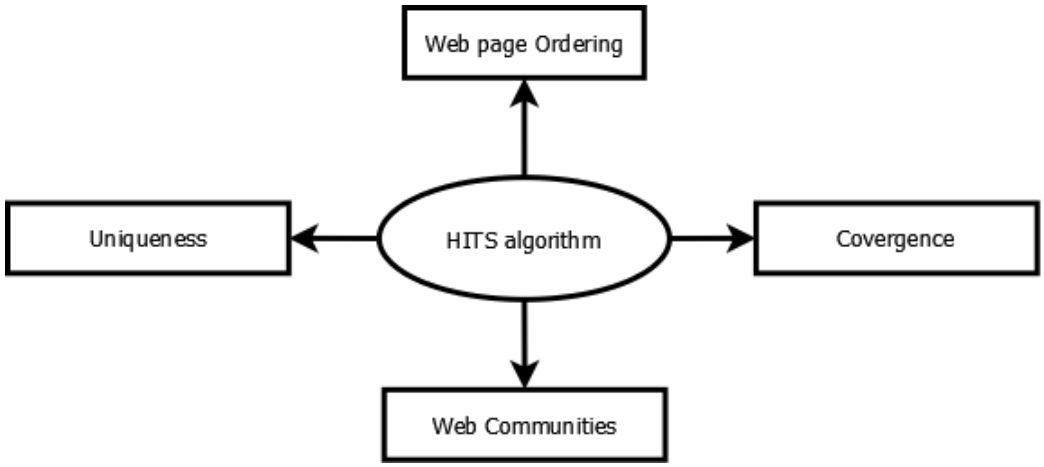

Fig 5.1 Properties of HITS

HITS provide good search results for a wide range of queries, but HITS did not work well in all cases due to the following reasons:

i) Mutually reinforced relationships between hosts.

ii) Automatically generated links.

iii) Non-relevant nodes.

iv) Topic drift occurs while the hub has multiple topics.

v) It assumes that all links pointing to a page are of equal weight.

6. CONCLUSION

In this paper, the concept of web mining, an application of data mining technique is studied. It is a powerful technique that is used to extract the information from past behaviour of users. Various

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

Volume 2 Issue 1 January 2013

International Manuscript ID : ISSN23194618-V2I1M3-012013

algorithms Page rank, HITS etc. is used in web structure mining to rank the relevant pages. The main focus of web structure mining is on link information.**HITS** algorithm is in the same spirit as **PageRank**. They both make use of the link structure of the Web graph in order to decide the relevance of the pages. Page rank algorithm calculates the score at indexing time and sort them according to the importance of page whereas HITS calculates the hub and authority score of n relevant pages. Many prominent pages are not self-descriptive. In HITS algorithm all links should be equally treated. Some links may be more meaningful than other links. For the future works, there are still many issues that need to be explored with the HITS algorithm.

REFERENCES

1. URL:http://www.di.unipi.it/~coscia/files/WMA-SNA4.pdf
2. URL:http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html
3. URL: http://en.wikipedia.org/wiki/Web_mining
4. URL: http://www.ijcaonline.org/volume13/number5/pxc3872448.pdf
5. URL: http://www.ijcst.com/vol22/2/tamanna.pdf
6. URL: http://www.web-datamining.net/structure/
7. URL:http://www.ke.tu-darmstadt.de/lehre/archiv/ss11/web-mining/wm-graph.pdf