International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

A Survey on Link Analysis Extension of Correspondence Analysis to Mine Web Pages

Dipali Rajput[1], Pooja Jardosh[2]

Department of Informatin & Technology, Silver Oak College of Engineering and Technology, Ahmedabad, India

Abstract: ***Link Analysis is a Procedure of Discovering Relationships in Web Database. It Generalizing Simple & Multiple Correspondence analysis. The Web is Growing Very fast, Sometimes the Users lost in the web's hyper structure. Users don't get that kind of results what actually they want. The main goal of Link Analysis to give the accurate result to the users. . The aim of this paper is providing the users accurate and relevant results.***

Keywords: Web mining, Link Analysis, HITS (Hypertext Induced Topic Selection), Web pages

1. Introduction

A Real world data coming from too many different fields. Link analysis technique used to find relationships between different databases. This technique can be applied on one or more relational databases. Lots of databases can be analysed by using this technique. It is used to find out different relation in the databases. With taking a random walk, different states are studied in all the databases.

The aim of this research is to discover an efficient and better system for mining the web topology to identify authoritive web pages.

1. 1 Web Mining

Web Mining means mining of data in World Wide Web database. It presented in the form of set of web pages. There are different types of web data. It can Contain Web pages as text & images. and also as data usage, which explain how web pages are visited by different users on the internet.

- **Web Content Mining**: web content mining is used to examine the both content and results of web searching. It uses data mining techniques for better efficiency, effectiveness and scalability.

  Web content mining divided into: agent-based approach & database based approach

  In agent-based approach, it contain software systems that perform the content mining. Such as, search engines.

  In database-based approach: it views the web data as belonging with a database. It has many query languages that target the web.

- **Crawler**: it is also called spider or robot. It traverses the hypertext structure. The set of pages where the crawler starts are referred as the seed URLs.by starting from that one page, all links from that are saved and recorded in a queue. There are four types of crawler: periodic crawler, incremental crawler, focused crawler, context focused crawler.

- **Virtual Web View**: MLDB is used here to handle large amounts of unstructured data which are on web.MLDB stand for Multiple Layered Database. In this database is massive and also distributed. Every layer is more generalized than the layer beneath it.

  MLDB provides a condensed and abstracted view of part of the web.

  VWV (Virtual Web View)is a view of the MLDB.to provide data mining,WebML a web data mining query language is used. It is an extension of DMQL.

- **Web Structure mining**: Here, information is obtained from the actual organization of pages on the web. It can creating a model of the web organization or a portion of it.

It is used to classify web pages to create similarity measures between the documents. There are two techniques of web structure mining: Page rank and HITS algorithm.

1.2 Mining Techniques

- **Page rank Algorithm:** it is used to increase the effectiveness. It also improves the efficiency of the search engines. It is used to measure the importance of a page and also to prioritize the pages which are returned from a traditional search engine using the keyword searching."GOOGLE" the biggest search engine use this technique.
- **HITS Algorithm:** It stands for Hyperlink-induced topic search. It finds hubs an authorative page. This technique contains two components: Based on a given set of keywords (found in a query), a set of relevant pages is found. Hub an authority measures are associated with these pages. Pages with the highest values are returned. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content.
- **Constraints with HITS algorithm:** 1.Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.

Licensed Under Creative Commons Attribution CC BY

1) Topic drift: HITS may not produce the most relevant results to the user queries because of equivalent of their weights.
2) Automatically generated links: HITS gives equal kind of importance for automatically generated links. It may not have that relevant topics for the user's query.
3) Efficiency: HITS algorithm is not efficient in real time. HITS was used in a prototype search engine called Clever. It is used for an IBM research project. Because of the above drawbacks HITS could not be implemented in a real time search engine.

2. Literature Review

On the base of all this paper, Link Analysis is very important. It also needs to be accurate and efficient. The results must be proper and well sorted. In link analysis the major drawback is that we don't get any proper accurate relevant results.

In paper [6] the author presents that the web results needs to be accurate. The World Wide Web is a rich source of information and it continues to expand in size and complexity. Retrieving of the required web page on the web, efficiently and effectively, is becoming a Challenge. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. The bulk amount of information becomes very difficult for the users to find, extract, filter or evaluate the relevant information. This issue raises the necessity of some technique that can solve these challenges. Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), and Machine Learning etc

In paper [5] the author gives the information about the working process of page rank and HITS both. The most important feature of HITS is the mutual reinforcement between hubs and authorities, while the most important feature of Page Rank is the hyperlink weight normalization. They can be generalized and combined. We also clarify and formalize weight propagation and random surfing as two different but related method to compute ranking scores. All these form a unified framework for link analysis. In this framework, one can easily design new ranking algorithms. We study three new ranking algorithms: the AuthRank, the Hub-Rank and the Sym-Rank. All these three rankings combine both features of HITS and Page Rank, thus they are expected to be somewhere between the rankings produced by HITS and Page Rank. The most important results are: all three rankings can be solved in closed-form. The authority rankings of Auth-Rank and Sym-Rank are identical to the ranking by in degrees. The hub rankings of Hub-Rank and Sym-Rank are identical to the ranking by out degrees.

In paper [4] it gives the certain information as below. It will be devoted to the application of this methodology to fuzzy SQL queries or fuzzy information retrieval. The objective is to retrieve not only the elements strictly complying with the constraints of the SQL query, but also the elements that almost comply with these constraints and are therefore close to the target elements. We will also evaluate the proposed methodology on real relational databases.

In paper [3] this paper contains the detailed information about the web mining. It gives me details about the web content mining and parts, crawler, web structure mining, Page Rank, HITS etc. The problem we all face when we search a topic in the web using a search engine like Google is that we are presented with millions of search results. First of all it not practically feasible to visit all these millions of web pages to find the required information.

In paper [2] it proposed a general method for semantic post processing of search results which is based on entity mining and Linked Data. For selecting the entities that better characterize the search results and their context, we proposed a Link Analysis-based method. The produced top-K semantic graphs allow the users to instantly inspect information that may lie in different places and that may be laborious and time consuming to locate (avoiding thereby the disengagement of the users from their initial task). In addition, they provide useful information about the context of the identified entities and allow the users to get a more sophisticated overview and to make better sense of the results.

In paper [1] the author gives the normal basic information about the link analysis. How it works, which kind of techniques it required etc,. Link analysis technique can be applied on large number of databases. It shows the relationships between multiple databases. This technique can be used for extracting relational databases or graph. This technique makes analysis between small instances of relational databases. This kind of data mining techniques discovers new relations in relational databases.

3. Conclusion

HITS does not produce relevant results, because of the equivalent weights and it's also not efficient in real time(real scenario).Our aim is to solve issues where authorities: pages that are relevant and are linked to different pages. And hubs: pages that are linked to several related authorities. We have improved HITS algorithm with novel alrothim's idea to achieve more relevant results. We have taken benefit of simplicity of HITS and added some novel steps that causes better results than the earlier algorithms.

References

[1] Luh Yen , MarcoSaerens, Member, IEEE and Franois Fous "A Link Analysis Extension of Correspondence Analysis for Mining Relational Databases.
[2] P. Kroonenberg and M. Greenacre, "Correspondence Analysis," Encyclopedia of Statistical Sciences, S. Kotz, ed., second ed., pp. 13941403, John Wiley & Sons, 2006.
[3] "A Link Analysis Extension Of Correspondence Analysis For Mining Relational Databases" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011
[4] Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time 2014 IEEE International Conference on Semantic Computing.

Licensed Under Creative Commons Attribution CC BY

[5] N. Duhan, A.K. Sharma and K.K. Bhatia, PageRanking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009.
[6] Chakrabarti,B.Dom,D.Gibson,J.Kleinberg,R.Kumar,P.Raghavan,S.Rajagopalan, and A.Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.
[7] F. Fouss, A. Pirotte, J.-M. Renders, M. Saerens,"Randomwalk computation of similarities between nodes of a graph, with application to collaborative recommendation", IEEE Transactions on Knowledge and Data Engineering, 19(3), pp. 355–369, 2007.
[8] I. Fellegi, A. Sunter,"Theory of record linkage", Journal of the American Statistical Association, 64, pp. 1183–1210, 1969.
[9] F. Greets, H. Manila, E. Terzi,"Relational link-based ranking", Proceedings of the 30th Very Large Data Bases Conference (VLDB), pp. 552–563, 2004.
[10] S. Lafon, A. B. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization", IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(9), pp. 1393–1403, 2006
[11] M. Thelwall,"Link Analysis: An Information Science Approach", Elsevier, 2004.
[12] J. M. Klienberg, Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, September 1999
[13] Volume 3 Issue 12 December 2014, Page No. 9391-9394 Smita Shinde1IJECS Volume 3 Issue 12 December, 2014 Page No.9391-9394 Page 9391 Link Analysis in Relational Databases using Data Mining Techniques.

Licensed Under Creative Commons Attribution CC BY