

## THE PRAGMATIC REVIEW ON WEB MINING BASED WEB PAGE RANKING APPROACHES

<sup>1</sup>Divjot, <sup>2</sup>Dr. Jaiteg Singh

<sup>1</sup>Chitkara School of Engineering and Technology;Rajpura,Punjab,India

<sup>2</sup>Chitkara School of Engineering and Technology;Rajpura,Punjab,India

### ABSTRACT

Hyperlink-Induced Topic Search (HITS) is a link ranking and scoring algorithm used in web mining which helps in rating Web pages. These web pages which are ranked are also known as Hubs. HITS mechanism for web crawling and retrieval still remains as a challenging issue and is under research in the academic as well as the corporate world. Web mining technique is used to categorize users and pages by analyzing the content of pages and sequence of URLs accessed. In this research work, an empirical algorithmic approach has been developed that makes use of the deep investigation of the hubs and authorities. In the proposed approach, the external links are verified as well as validated for scoring and execution time comparison that is finally compared to the existing approach. In the classical approach, the scoring criteria of the hubs and authorities are obsolete and do not fetch the external links which are mandatory as per the current paradigm. In this work, the algorithmic approach and implementation has been performed to prove the performance of the proposed work.

Keywords – Link Popularity, Page Rank, Web Mining, Web Rank

### INTRODUCTION

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is freely available for users to access. Data has to be managed and organized in such a way that it can be accessed by different users effectively and efficiently. Several data mining methods are used to discover the hidden information in the Web. Web mining technique is used to categorize users and pages by analyzing the content of pages, and the way how URLs are accessed. Artificial intelligence, information retrieval and natural language processing techniques can be used efficiently with datamining. Thus, Web mining is developing as an autonomous research area.

### WEB SITE RANKING ALGORITHM

With the increasing use of academic search engines, it becomes more important for academic authors to have their articles well ranked in these search engines in order to reach their audience. Ranking is an integral component of any information retrieval system. In the case of Web search, the role of ranking is very crucial due to the web size and different users type. Web users only want pages which are relevant and authoritative in nature. It is important for the ranking function to output the accurate desired results.

Ranking algorithms are a well-guarded black box proprietary mathematical formula that search engines utilize to calculate site ranking in the SERPs (Search Engine Result Page) for each search query [9].

An efficient ranking algorithm is important in any information retrieval system. In a web search engine, due to the dimensions of the current web, and the special needs of the users, its role becomes critical. Recent studies estimated the existence of more than 11.5 billion pages on the web. [10]

### SOURCES OF DATA FOR WEB USAGE MINING

The data for Web Usage Mining can be gathered from:

- Web Server Logs
- Proxy Server Logs
- Browser Logs

### HITS ALGORITHM

In the same time that Page Rank was being developed, Jon Kleinberg a professor in the Department of Computer Science at Cornell came up with his own solution to the Web Search problem. He developed an algorithm that made use of the link structure of the web in order to discover and rank pages relevant for a particular

topic. **HITS (hyperlink-induced topic search)** is now part of the **Ask** search engine (www.Ask.com).

One of the interesting points that he brought up was that the human perspective on how a search process should go is more complex than just compare a list of query words against a list of documents and return the matches. Suppose we want to buy a car and type in a general query phrase like "the best automobile makers in the last 4 years", perhaps with the intention to get back a list of top car brands and their official web sites. When you ask this question to your friends, you expect them to be able to understand that automobile means car, vehicle, and that automobile is a general concept that includes vans, trucks, and other type of cars. When you ask this question to a computer that is running a text based ranking algorithm, things might be very different. That computer will count all occurrences of the given words in a given set of documents, but will not do intelligent rephrasing for you. The list of top pages we get back, while algorithmically correct, might be very different than what expected. One problem is that most official web sites are not enough self-descriptive. They might not advertise themselves the way general public perceives them. Top companies like Hyundai, Toyota, might not even use the terms "automobile makers" on their web sites. They might use the term "car manufacturer" instead, or just describe their products and their business. What is to be done in this case? It would be of course great if computers could have a dictionary or ontology, such that for any query, they could figure out synonyms, equivalent meanings of phrases. This might improve the quality of search, nevertheless, in the end; we would still have a text based ranking system for the web pages. We would still be left with the initial problem of sorting the huge number of pages that are relevant to the different meanings of the query phrase. We can easily convince ourselves that this is the case. Just remember one of our first examples, about a page that repeats the phrase "automobile makers = cars manufacturers = vehicle designers" a billion times. This web page would be the first one displayed by the query engine. Nevertheless, this page contains practically no usable information.

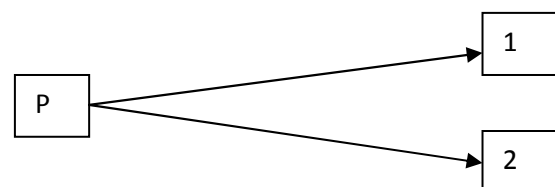
The conclusion is that even if trying to find pages that contain the query words should be the starting point, a different ranking system is needed in order to find those pages that are **authoritative** for a given query. Page  $i$  is called an **authority** for the query "automobile makers" if it contains valuable information on the subject. Official web sites of car manufacturers, such as www.bmw.com, HyundaiUSA.com, www.mercedes-benz.com would be authorities for this search. Commercial web sites selling cars might be authorities on the subject as well. These are the ones truly relevant to the given query. These are the ones that the user expects back from the query engine. However, there is a second category of pages relevant to the process of finding the authoritative pages, called **hubs**. Their role is to advertise the authoritative pages. They contain useful links towards the authoritative pages. In other words, hubs point the search engine in the

"right direction". In real life, when you buy a car, you are more inclined to purchase it from a certain dealer that your friend recommends. Following the analogy, the authority in this case would be the car dealer, and the hub would be your friend. When you trust your friend then you trust what your friend recommends. In the World Wide Web, hubs for our query about automobiles might be pages that contain rankings of the cars, blogs where people discuss about the cars that they purchased, and so on.

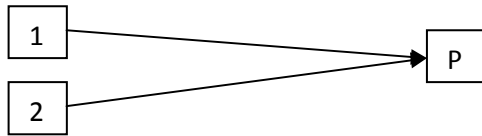
In order to get a set rich in both hubs and authorities for a query  $Q$ , we first collect the top 200 documents that contain the highest number of occurrences of the search phrase  $Q$ . These, as pointed out before may not be of tremendous practical relevance, but one has to start somewhere. Kleinberg points out that the pages from this set called root ( $R_Q$ ) are essentially very heterogeneous and in general contain only a few (if any) links to each other. So the web sub graph determined by these nodes is almost totally disconnected; in particular, we cannot enforce Page Rank techniques on  $R_Q$ .

Authorities for the query  $Q$  are not extremely likely to be in the root set  $R_Q$ . However, they are likely to be pointed out by at least one page in  $R_Q$ . So it makes sense to extend the sub graph  $R_Q$  by including all edges coming from or pointing to nodes from  $R_Q$ . We denote by  $S_Q$  the resulting sub graph and call it the seed of our search. Notice that  $S_Q$  we have constructed is a reasonably small graph (it is certainly much smaller than the 30 billion nodes web graph!). It is also likely to contain a lot of authoritative sources for  $Q$ . The question that remains is how to recognize and rate them? Heuristically, authorities on the same topic should have a lot of common pages from  $S_Q$  pointing to them. Using our previous terminology, there should be a great overlap in the set of hubs that point to them.

From here on, we translate everything into mathematical language. We associate to each page  $i$  two numbers: an authority weight  $a_i$ , and a hub weight  $h_i$ . We consider pages with a higher  $a_i$  number as being better authorities, and pages with a higher  $h_i$  number as being better hubs. Given the weights  $\{a_i\}$  and  $\{h_i\}$  of all the nodes in  $S_Q$ , we dynamically update the weights as follows:



**Figure 1.1:** Association of Pages with other links in the website



**Figure 1.2:** Representation of Pages as Nodes in HITS Algorithm

A good hub increases the authority weight of the pages it points. A good authority increases the hub weight of the pages that point to it. The idea is then to apply the two operations above alternatively until equilibrium values for the hub and authority weights are reached.[24]

## LITERATURE REVIEW

A number of International Research Papers has been studied to analyze different page ranking algorithms and mining techniques with a search engine as a prediction system. They are as follow:

(Sowmya.M, V.S.Sreelaxmi, Muneshwara M.S, Anil G.N, 2013 [19]) - Pagerank vector for ranking the search-query results, which made use of link structure of the Web, to get the importance of Web pages, particularly independent of any search query. To get correct search results, the authors proposed calculating a set of PageRank vectors, biased with a set of archetypical topics, to capture more correctly the notion of prominence with respect to a particular topic. By taking these biased PageRank vectors we compute query-specific rank scores for web pages at query time, it is shown that the authors can compute more accurate importance score than with a single PageRank vector. This work computes the topic-sensitive PageRank scores for pages for normal keyword search queries, sufficing the query using the topic of the keywords. PageRank scores using context in which the query appeared. For better ordering of web pages compute an associated PageRank algorithm for search engines to get quality results by scoring based on relevance between web documents. The modified PageRank algorithm created certain ordering using relevance than the original one, and reduces the query time overhead of topic-sensitive PageRank. PageRank is a global ranking of all web pages based on their locations in the web graph structure.

(Xiuxia Tian et. Al. 2014 [22]) - HITS algorithm developed by Jon Kleinberg made use of the link structure of the web network in order to discover and rank pages relevant to a particular topic. But it only took account of the hyperlink structure and completely excluded contents of web pages. Moreover, it ignored the fact that degrees of the importance of many links may be different. Therefore, this algorithm will lead to topic drifts. In this paper, we propose an improved HITS

algorithm based on the theory of triadic closure and VSM. This method firstly computes the relevance between arbitrary two pages based on page topic similarity and common reference degree. Then, by using the relevance, a new adjacency matrix is constructed to iteratively calculate authorities and hubs. Preliminary experiments show the new algorithm improves the efficiency and quality of query, reduce the theme drifts.

(Miguel Gomes da Costa JúniorZhiguo Gong 2005 [20]) - Due to the increasing amount of data available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing. In this paper, the authors provide an introduction of Web mining as well as a review of the Web mining categories. Then we focus on one of these categories: the Web structure mining. Within this category, we introduce link mining and review two popular methods applied in Web structure mining: HITS and PageRank. In this paper we survey the research area of Web mining, focusing on the category of Web structure mining.

(Dr.Miji C. Kan Beijing 2014 [21]) Hits Algorithm - With the rapid increase in internet technology,users get easily confused in large hyper textstructure. Web mining is used to categorize users andpages by analysing the users behaviour, the contentof the pages, and the order of the URLs that tend tobe accessed in order. Web structure mining playsvery important role in this approach. Its defined asthe process of analysing the structure of hyperlinkusing graph theory. There are many proposedalgorithms for web structure mining such asPageRank (PR), Weighted PageRank (WPR), andHyperlink-Induced Topic Search (HITS) etc.Hyperlink-Induced Topic Search or simply HITS isa link analysis algorithm that rates Web pages,developed by Jon Kleinberg. It was a precursor toPageRank.. In this manuscript, various aspects of the HITS algorithm isanalyzed. In this manuscript, the major factor and aspect that the classical HITS Algorithm do not access the external links is addressed.

(Anshuman Sharma 2012 [2]) explained that the web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the web. The study of ant colonies behavior and their self-organizing capabilities is of interest to knowledge retrieval/ management and decision support systems sciences, because it provides models of distributed adaptive organizations, which are useful to solve difficult optimization, classification, and distributed control problems, among others. In this paper, we studied the possible use of the neural networks learning capabilities to classify the web traffic data

mining set. The discovery of useful knowledge, user information and server access patterns allows web based organizations to mine user access patterns and helps in future developments, maintenance planning and also to target more rigorous advertising campaigns aimed at groups of users.

(R. Cooley et al 1997 [14]) - In this research work the term Web mining has been used in two distinct ways. The first called Web content mining is the process of information discovery from sources across the World Wide Web. The second called Web usage mining is the process of mining for user browsing and access patterns. The manuscript defines Web mining and presents an overview of the various research issues, techniques, and development efforts. They briefly describe WEBMINER, a system for Web usage mining, and conclude the paper by listing research issues.

(Han Jiawei 2002 [5]) In this research work to supplement keyword-based indexing, researchers have applied data mining to Web-page ranking. In this context, data mining helps Web search engines find high-quality Web pages and enhances Web click stream analysis. As researchers continue to develop data mining techniques, the authors believe this technology will play an increasingly important role in meeting the challenges of developing the intelligent Web. Ultimately, data mining for Web intelligence will make the Web a richer, friendlier, and more intelligent resource that we can all share and explore. The paper considers how data mining holds the key to uncovering and cataloguing the authoritative links, traversal patterns, and semantic structures that will bring intelligence and direction to our Web interactions.

(Andrew McCallum et al 1999 [1]) This paper proposes the use of machine learning techniques to greatly automate the creation and maintenance of domain-specific search engines. This manuscript describes new research in reinforcement learning, text classification and information extraction that automates efficient spidering, populating topic hierarchies, and identifying informative text segments. Using these segments, a demonstration system is built.

(Fuchun Peng, et al [4]) This paper employs Conditional Random Fields (CRFs) for the task of extracting various common fields from the headers and citation of research papers. The basic theory of CRFs is becoming well-understood, but best-practices for applying them to real-world data requires additional exploration. This paper makes an empirical exploration of several factors, including variations on Gaussian, exponential and hyperbolic-L1 priors for improved regularization, and several classes of features and Markov order.

(B.Santhosh Kumar et al 2010 [3]) implements three phases of Web usage mining namely reprocessing,

pattern discovery, and pattern analysis. Apriori algorithm is used to generate an association rule that associates the usage pattern of the clients for a particular website. The output of the system was in terms of memory usage and speed of producing association rules.

(Pooja Sharma et al 2011 [12]) proposed a clustering algorithm to find out data clusters for both numerical and nominal data by calculating the average and log values of data set. This algorithm improves the techniques of Web Usage Mining by first discover the log files of individual users at one place.

(Romo Martinez et al 2010 [17]) have analyzed different information retrieval methods for both, the selection of terms used to construct the queries submitted to the search engine, and the ranking of the candidate pages that it provides, in order to help the user to find the best replacement for a broken link. To test the sources, they have also defined an evaluation methodology which does not require the user judgments, what increases the objectivity of the results.

(MahendraPratap Singh Dohare et al 2012 [10]) proposed a new reactive session reconstruction method. This algorithm is better than previously developed both time and navigation oriented heuristics as it does not allow page sequences with any unrelated consecutive requests to be in the same session. They have also implemented agent simulator for generating real user sessions.

(Resul Das et al 2007 [16]) analyzed the web server user access logs of Firat University to help system administrator and Web designer to improve their system by determining occurred system errors, corrupted and broken links by using web using mining.

(Priyanka Patil et al 2012 [13]) have focused on web log file format, its type and location. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log file. Data preprocessing is an important step to filter and organize appropriate information before using to web mining algorithm. They have also proposed two algorithms for field extraction and data cleaning. Preprocessing web log file is used in data mining techniques, also used in intrusion detection system as input to detect intrusion.

## **PROBLEMS IN EXISTING SYSTEM**

### **RANKING OF EXTERNAL LINKS**

The classical HITS Algorithm does not analyse External Links. The existing approach analyse the links available on the same domain. The HITS algorithm is based on the ranking of current domain. It does not rank and score external entities.

### **DEEP LINKS NAVIGATION AND SCANNING OF LINKS**

The deep link analysis and multiway penetration of a web page for the deep link analysis is not available. Deep link analysis is done to check the various HTML error codes. Classical approach only investigate broken links which are available on the same host.

The page ranking algorithms are used to rank and score the web pages existing and uploading on the same server. The pages which are external in nature are not accessed and indexed by the classical approach. There is a need to develop a novel algorithmic approach and then to perform the implementation to pragmatically analyze the performance of both approaches.

## IMPLEMENTATION OF PROPOSED ALGORITHM

### Anticipated Technical Implementation and Approach

- The proportional analysis on various rules and patterns mining using on the clustering and its applications in the web log files
- Design of a new algorithmic approach towards the web performance algorithm in log files analysis and forensic information
- Implementation of the improved and efficient algorithm on the sample log record fetched from the live server
- Relative study of the results with the existing techniques and generation of the detailed reports
- Acceptance and trust level of the hypothesis and objectives specified in the research proposal
- Framing out the conclusion and future work from the implementation performed and results fetched

### Methodology Used

- Collection of the Training Data Set for Analysis
- The Training Data Set Consists of the URLs for investigation
- Generation of the Authorities and Hubs
- Deep Analysis on each parameter
- Applying the proposed model on the Training data set
- Fetch Results
- Data Interpretation

### Objectives

- *To investigate the drawbacks and shortcomings in the HITS Algorithm for Page Analysis*

- *To propose an empirical model and algorithmic approach for the improved case of HITS Algorithm*
- *To simulate the training set on the classical as well as improved HITS Algorithm*
- *To perform the empirical and pragmatic analysis on the improved HITS Algorithm*
- *To propose a novel algorithm for link analysis*
- *The Results will be analyzed on multiple parameters*

### Algorithmic Approach

- *Read/ scan URL*
- *Associate mandatory parameters : keywords, title, description, sub links*
- *Generate and investigate authority and hub parameters*
- *Perform deep association and complexity removal technique on the training data set*
- *If final results obtained*
- *Then Stop and terminate with success*
- *Go to step 3*
- *Plot charts and interpretation*

### Mathematical Formulation of the Problem

$$URL_i (T, K, D, S_i) \Rightarrow AUTH (URL_i \rightarrow URL_{ij}) = > HUBS (URL_i \rightarrow URL_{ij} | AUTH) \Rightarrow PE_{Aij} \Rightarrow GEN (AUTH, HUB | FINAL) \Rightarrow S(URL_i) \Rightarrow CR(URL_i) \Rightarrow T(S) URL_i$$

*URL – Universal Resource Locator*

*PEA – Performance Evaluation Algorithm*

*CR – Cumulative Result*

*AUTH – Authority*

*PEA – Performance Evaluation Function*

## CONCLUSION

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that performs ranking of the Web pages. It was developed after PageRank. The concept behind Hubs and Authorities was, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information. They were used as broad catalogue of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. The classical HITS algorithm is having number of problems including redundancy and

interdependency of links ,delay and total turn aroundtime.The proposed algorithmic approach will remove the shortcomings of classical approach and will output the more accurate set of data that is more reliable in terms of execution time,complexity and indexing by the crawlers.

## SCOPE OF FUTURE WORK

The classical techniques does not use the tree based navigation inside each url for investigation and ranking as the proposed approach will use. In the future action, an empirical algorithm will be developed and simulated for better results in terms of ranking and executiontime.To improve the existing techniques, the use of metaheuristics can be implemented for optimal results. Additionally, the usage of ontologies based parameterscan be used to improve the algorithmic approach and results. The future work of the proposed technique can be extended towards making use of the swarm intelligence techniques. The swarm intelligence techniques make use of number of iterations for the investigation and analysis of the given pseudocode or algorithmic flow. However, there are number of algorithms for link ranking, there is the scope of improvements and further work. Additionally, there are hybrid approaches to improve the current research work in which two or more algorithms can be joined together and then results can be fetched using parallel techniques

## REFERENCES

- [1] Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore, "Building Domain-Specific Search Engines with Machine Learning Techniques", Digital Library, AAAI Spring Symposium on Intelligent Agents in Cyberspace, 1999.
- [2] Anshuman Sharma, "Web Usage Mining Using Neural Network", International Journal of Reviews in Computing, Vol. 9, pp. 72-78, 2012.
- [3] B. Santhosh Kumar, K. V. Rukmani, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms", International Journal of Advanced Networking and Applications, Vol.1, Issue 6, pp.400-404, 2010.
- [4] Fuchun Peng, Andrew McCallum, "Accurate Information Extraction from Research Papers using Conditional Random Fields", Digital Library, The Pennsylvania State University, 2004.
- [5] Han Jiawei, "Data Mining for Web Intelligence", IEEE Computer Society, pp. 64-70, 2002.
- [6] Jaideep Srivastava, PrasannaDesikan, Vipin Kumar, "Web Mining – Concepts, Applications and Research Directions, Springer Studies in Fuzziness and Soft Computing Volume 180, 2005, pp 275-307.
- [7] K. R. Suneetha, R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File", International Journal of Computer Science and Network Security, Vol.9 No. 4, pp.327-332, 2009.
- [8] Kavita Sharma, GulshanShrivastava, Vikas Kumar, "Web Mining: Today and Tomorrow", International Conference on Electronics Computer Technology, pp. 399-403, 2011.
- [9] Li Pingxiang, Chen Jiangpingn, BianFuling, "A Developed Algorithm Of Apriori Based On Association Analysis", Geo-spatial Information Science (Quarterly), Vol.7, Issue 2, pp. 108-112, 2004.
- [10] MahendraPratap Singh Dohare, Premnarayan Arya, ArunaBajpai, "Novel Web Usage Mining for Web Mining Techniques", International Journal of Emerging Technology and Advanced Engineering, Vol.2, Issue 1, pp.253-262, 2012.
- [11] Navin Kumar Tyagi, A.K. Solanki and Manoj Wadhwa, "Analysis of Server Log by Web Usage Mining for Website Improvement", Vol. 7, Issue 4, pp. 17-21, 2010.
- [12] Pooja Sharma, RupaliBhartiya, "An efficient Algorithm for Improved Web Usage Mining", International Journal of Computer Technology and Applications, Vol.3 (2), pp.766-769, 2011.
- [13] Priyanka Patil, UjwalaPatil, "Preprocessing of Web Server Log File for Web Mining", World Journal of Science and Technology, pp.14-18, 2012.
- [14] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings ninth IEEE International Conference, pp. 558-567, 1997.
- [15] Rahul Mishra, AbhaChoubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth Algorithm for Web Usage Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.2, pp. 311-318, 2012.
- [16] Resul Das, Ibrahim Turkoglu, Mustafa Poyraz, "Analyzing of System Errors for Increasing A Web Server Performance by Using Web Usage

- Mining”, Journal of Electrical and Electronics Engineering, Vol.7 No. 2, pp. 379-386, 2007.
- [17] Romo Juan Martinez, Araujo Lourdes, “Analyzing Information Retrieval Methods to Recover Broken Web Links”, ECIR, pp. 26-37, 2010.
- [18] VidhuSinghal, Gopal Pandey, “A Web Based Recommendation Using Association Rule and Clustering”, International Journal of Computer and Communication Engineering Research, Vol. 1, Issue 1, pp. 1-5, 2013.
- [19] Survey On Pagerank Algorithms Using Web-Link Structure Sowmya.M, V.S.Sreelaxmi, Muneshwara M.S, Anil G.N, International Journal of Engineering Research-Online, Vol.1.Issue.1:2013
- [20] Miguel Gomes da Costa JúniorZhiguo Gong, Web Structure Mining: An Introduction, Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China
- [21] Dr. Miji C. Kan, Hits Algorithm - An Effective Link Investigation Approach, Beijing University of Technology China, ISSN (Online) : 2348 - 2001 International Refereed Journal of Reviews and Research Volume 2 Issue 1 January-February 2014
- [22] Journal of Information & Computational Science 11:6 (2014) 1861–1868 April 10, 2014 Available at <http://www.joics.com> Improvements of HITS Algorithm Based on Triadic Closure \*Xiuxia Tian a, Yajun Du a;\*, Wen Song a, Wenjun Liu a, YingminXie [22] School of Mathematics and Computer Engineering, Xihua University, Chengdu 610039, China b109 Geological Brigade of Sichuan Bureau of Geology and Mineral Resources, Chengdu 610100, China
- [23] Journal of Information & Computational Science 11:6 (2014) 1861–1868 April 10, 2014, Improvements of HITS Algorithm Based on Triadic Closure, Xiuxia Tian et.al
- [24] ]<http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture4/lecture4.html>