International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161
©2014 INPRESSCO®, All Rights Reserved
Available at http://inpressco.com/category/ijcet

Research Article

Role of Web Mining Algorithms for Ranking Web Pages

Anuradha[A*], G.Lavanya Devi[B] and M.S Prasad Babu[B]

[A]Department of CSE, GMRIT, RAJAM, india-532127
[B]Dept. of CS &SE, Andhra University, Visakhapatnam, India-532003

Accepted 20 May 2014, Available online 01 June 2014, **Vol.4, No.3 (June 2014)**

Abstract

The rate of the growth the web has been exponential in the past two decades. The advances that took place in the software and hardware particularly in the past decade are one of the vital factors for the advent growth of data on web. Web mining is application of data mining techniques to discover patterns from the web. Search engine is a software system/tool used to retrieve relevant data for a given query from the web. Page ranking algorithms play a chief role in the search engines. Hence the study of web mining, particularly search engines used in web mining has gained major interest amongst the researchers around the globe. This paper gives an overview of web mining and a distinctive survey of various web mining algorithms that are used in search engines for ranking web pages

Keywords: *Web mining, search engine, page ranking algorithms, link mining, content mining and usage mining.*

1. Introduction

The World Wide Web is a huge, widely distributed, global source for information services (*MPS Bhatia et al* 2005). By all measures, the Web is enormous and growing at a staggering rate, which has made it increasingly intricate and crucial for both people and programs to have quick and accurate access to Web information and services. It is imperative to provide users with tools for efficient and effective resource and knowledge discovery. Search engines have played a key role in the World Wide Web's infrastructure as its scale and impact have escalated. Although search engines are important tools for knowledge discovery on the Web, they are far from perfect(Lian- Wang *et al* 2009) but some of the challenges faced during the interaction with web are outlined below:

To find relevant information*:* Whenever user poses a query (simple keyword) to the search engine, It gives list of resultant web pages based on ranking, however these web search results usually have low precision and recall.

Personalization of web: People differ in contents and presentations of information while interacting with web(R. Kosala *et al* 2009). Example: Jaguar could be an animal, car, sports team or computer. This type of overloading keyword semantics can return many low quality answers.

The Web constitutes a highly dynamic information source: Not only does the Web continue to grow rapidly, the information it holds also receives constant updates. News, stock market, service center, and corporate sites revise their Web pages regularly. Linkage information and access records also undergo frequent updates (R. Kosala et al 2009). The freedom for anyone to publish information on the Web at anytime and anywhere implies that information on the Web is constantly changing. It is a dynamic information environment whereas traditional systems are typically based on static document collection (Lian- Wang et al 2009).

The above mentioned issues can be solved by using web mining techniques direct or indirect. The architecture of search engine is discussed below, It consists of the following three major components (G.Hanumantha Rao et al 2011).

1. The crawler and the indexer: It collects pages from the web, creates and maintains the index.
2. The user interface: It allows submitting queries and enables result presentation.
3. The database and the query server: It stores information about the Web pages and processes the query and returns results.

All the search engines essentially include a crawler, an indexer and a query server although the algorithms used in these components and quality of the algorithms may vary significantly. Search engines that are based on web crawling framework also used in web mining to find the interacted web pages.

2. Web mining

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following tasks (R. Kosala et al 2009)

Resource finding*:* It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on web.

*Corresponding author: **Anuradha**

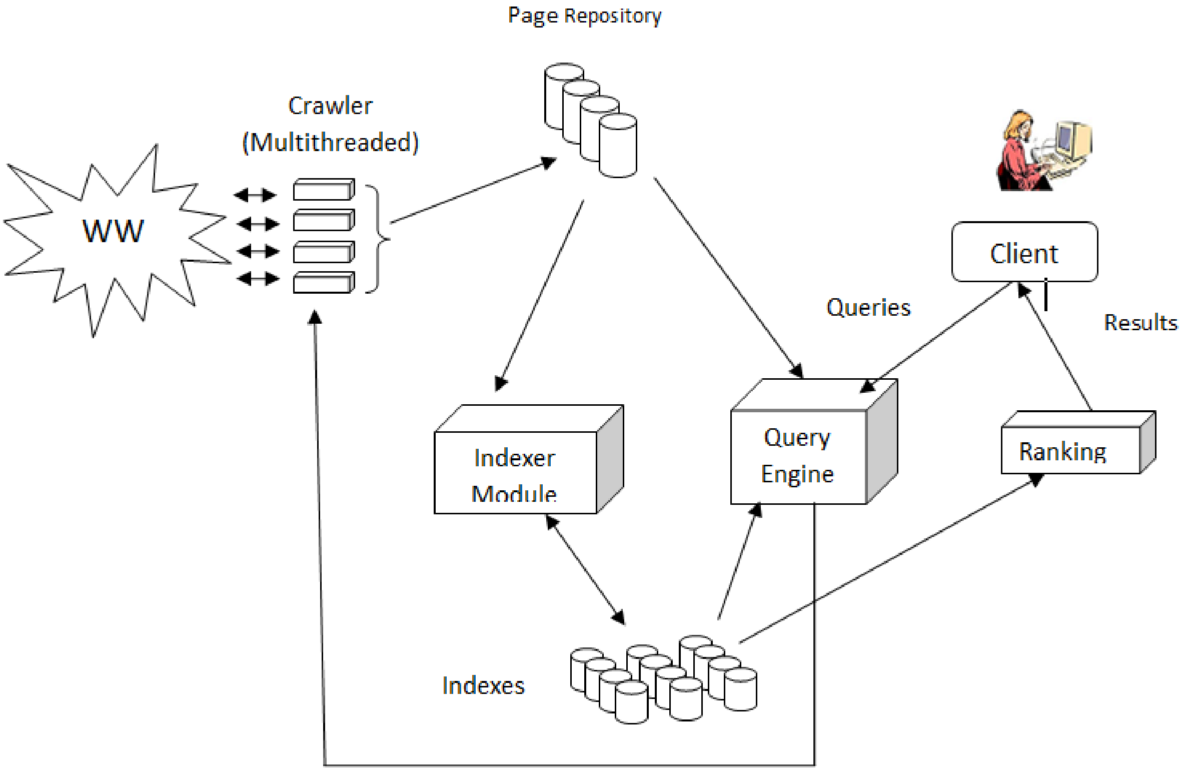


Fig1: Basic web search engine architecture

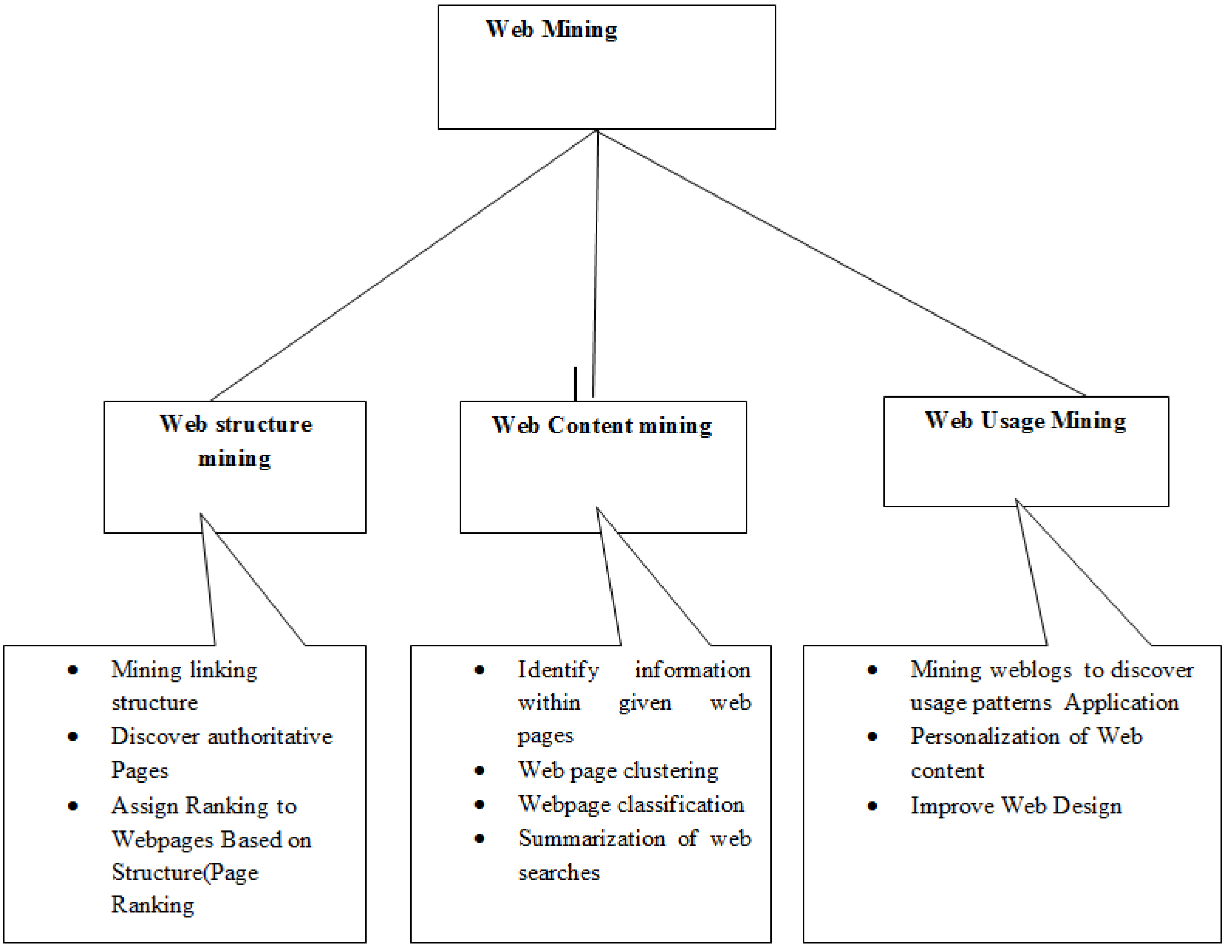

Web Mining
Web structure mining
Web Content mining
Web Usage Mining
Mining linking structure
Discover authoritative Pages
Assign Ranking to Webpages Based on Structure(Page Ranking
Identify information within given web pages
Web page clustering
Webpage classification
Summarization of web searches
Mining weblogs to discover usage patterns Application
Personalization of Web content
Improve Web Design

Fig2: Web mining types and tasks

Information selection and pre-processing: It involves the automatic selection and preprocessing of specific information from retrieved web resources.

This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in training corpus

Generalization: It is a technique used for retrieve patterns from different web pages or multiple web pages. Data Mining techniques and machine learning are used in generalization

Analysis: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

2.1 Web Structure Mining

Web structure mining generates structured summaries about information on web pages/web. It shows the links from one web page to the other web page, known as hyper link (T.Munibalaji et al 2012). A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. Intra-document web pages connects different parts of web pages using hyper link. Inter-document web pages connects two different web pages using hyperlink

In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts have focused on automatically extracting document object model (DOM) structures out of documents, which is also known as Document structure. The following are the link mining tasks that are used in web structure mining (Miguel Gomes da *et al 2005*)

Link-based Classification: Link-based classification is the most recent upgrade of a classic data mining task to linked domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

Link-based Cluster Analysis: The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

Link Strength: Links could be associated with weights.

Link Cardinality: The main task is to predict the number of links between objects. The following section gives a detailed review of the algorithms that use link mining tasks:

Brin, S, and Page,L. et al developed a Page rank algorithm which is being utilized by Google, Here, numerical weight is assigned to each element of hyperlink set of document such as World Wide Web Page. Rank considered randomly clicking on links based on probability distribution. The probability is expressed as a numeric value between 0 and 1. That numerical value is defined as damping factor. It is represented as d and usually its value set to be 0.85.

Kleinberg, John In, et al proposed HITS algorithm. It identifies two kinds of pages from the Web hyperlink structure; authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs. According to their paper, a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs.

Wenpu Xing and Ali Ghorbani et al. suggested Weighted Page Rank Algorithm that assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its out link pages. Each out link page gets a value proportional to its popularity (its number of in links and out links). The popularity from the number of in links and out links is recorded as Win (v, u) and Wout (v, u), respectively. Win (v, u) is the weight of link (v, u) calculated based on the number of in links of page u and the number of in links of all reference pages of page v. In this algorithm weight is assigned to both back link and forward link. Incoming link is defined as number of link points to that particular page and outgoing link is defined as number of links goes out.

Taher H. Haveliwala et al, generated Topic Sensitive Page Rank algorithm, that computes a set of Page Rank vectors, using a set of representative topics, to capture more accurately the notion of importance with respect to a particular topic. By using these pre computed based Page Rank vectors to generate query-specific importance scores for pages at query time, we show that generate more accurate rankings than with a single, generic Page Rank vector.

Ali Mohammad Zareh Bidoki et al.gave new ranking method Distance Rank based on novel recursive method, this is based on reinforcement learning which considers distance between pages as punishment, called ‘‘Distance Rank’’ to compute ranks of web pages. The distance is defined as the number of ‘‘average clicks’’ between two pages. Results indicate that Distance Rank outperforms other ranking algorithms in page ranking and crawling scheduling.

Fabrizio Lamberti et al, proposed Relation Based page rank on semantic search engine that would take into account keywords and return page only if both keywords are present within the page and they are related to the associated concept as described in to the relational note associated with each page.

2.2 Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents.

The two common tasks through which useful information can be mined from Web are Clustering and Classification.

The following are the various clustering techniques that are used to extract the information from the web.

Hierarchical Clustering*(* Niki R. Kapadiaet *et al 2012*) is a method of cluster analysis which builds hierarchy of clusters. It is the collection of objects arranged in hierarchical fashion.

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Partition Clustering Algorithm (Niki R. Kapadia et al 2012) the partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroid can be further clustered to produces hierarchy within a dataset.

Graph Based clustering (Miguel Gomes da et al 2005)] the documents to be clustered can be viewed as a set of nodes and the edges between the nodes represent the relationship between them. The edges bare a weight, which denotes the degree of that relationship. Graph based algorithms rely on graph partitioning, that is, they identify the clusters by cutting edges from the graph such that the edge- cut, i.e. the sum of the weights of the edges that are cut, is minimized. Since each edge in the graph represents the similarity between the documents, by cutting the edges with the minimum sum of weights the algorithm minimizes the similarity between documents in different clusters.

Neural Network based Clustering: Organizing feature Maps (SOM) (Takafumi Inoue et al 2005) is a widely used unsupervised neural network model. It consists of two layers: the input layer with n input nodes, which correspond to the n documents, and an output layer with k output nodes, which correspond to k decision regions. The input units receive the input data and propagate them onto the output units. Each of the k output units is assigned a weight vector. During each learning step, a document from the collection is associated with the output node, which has the most similar that it will become even more similar to the vector that represents that document.

Fuzzy Clustering: All the above approaches produce clusters in such a way that each document is assigned to one and only one cluster. Fuzzy clustering approaches, on the other hand, are non- exclusive, in the sense that each document can belong to more than one cluster. Fuzzy algorithms usually try to find the best clustering by optimizing a certain criterion function. The fact that a document can belong to more than one cluster is described by a membership function. The membership function calculates for each document a membership vector, in which the i -th element indicates the degree of membership of the document in the i-th cluster.

Probabilistic Clustering: Another way of dealing with uncertainty is to use probabilistic clustering algorithms. These algorithms use statistical models to calculate the similarity between the data instead of some predefined measures. The basic idea is the assignment of probabilities for the membership of a document in a cluster. Each document can belong to more than one cluster according to the probability of belonging to each cluster.

The following are various classification techniques that are used to extract the information from the web.

Decision Tree (Darshna Navadiya *et al 2012*)***:*** Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5. The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. The split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. A new example is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that example. The leaf node reached is considered the class label for that example.

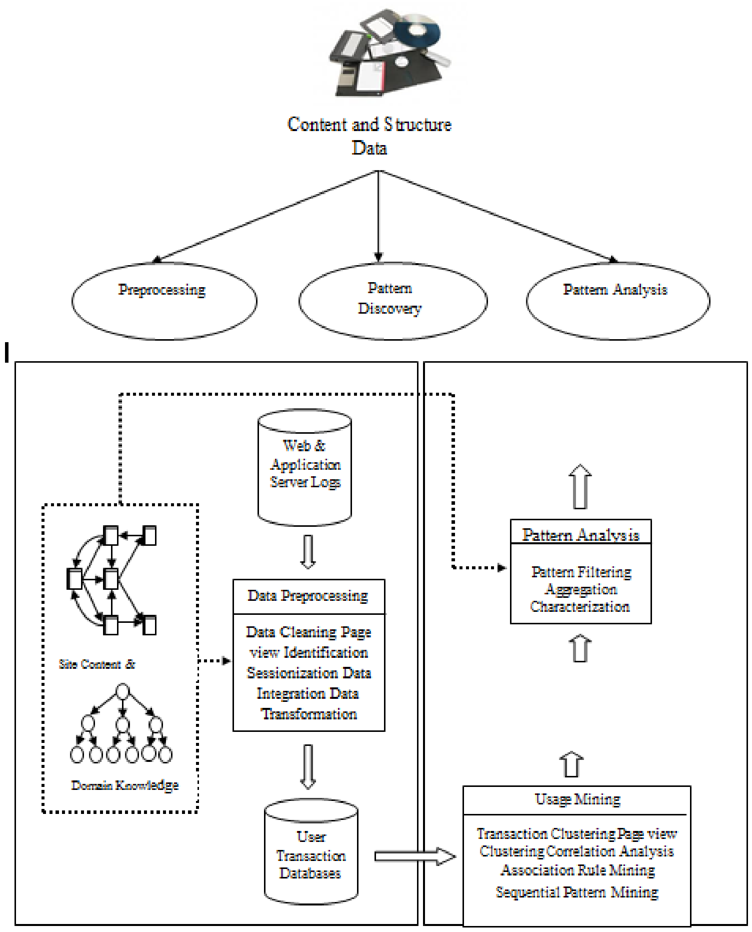
k-Nearest Neighbor (Darshna Navadiya *et al 2012*) KNN is considered among the oldest nonparametric classification algorithms. To classify an unknown example, the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation

Naive Bayes (Magdalini Eirinaki et al) Naive Bayes is a probabilistic classifier. It is constructed from the training data to estimate the probability of each class given the document feature values (words) of a new instance. Bayes theorem is used to estimate these probabilities. It works well even when the feature independence assumed by Naïve Bayes does not hold. However, it is based on simplifying assumptions (conditional independence)

The following section describes the algorithms used for assigning ranks to pages for both classifications and clustering methodologies that are used in web mining most frequently.

Yangzhou, et al, gives Tag Rank Algorithm it calculates the heat of the tags by using time factor of the new data source tag and the annotations behavior of the web users. This algorithm provides a better authentication method for ranking the web pages. The results of this algorithm are very accurate and this algorithm index new information resources in a better way.

Takafumi Inoue, et al, proposed Eigen Rumor Algorithm it scores each blog entry by weighting the hub and authority scores of the bloggers based on Eigen vector calculations. This algorithm enables a higher score to be


Content and Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Web & Application Server Logs
Data Preprocessing
Data Cleaning Page view Identification Sessionization Data Integration Data Transformation
Site Content &
Domain Knowledge
User Transaction Databases
Pattern Analysis
Pattern Filtering Aggregation Characterization
Usage Mining
Transaction Clustering Page view Clustering Correlation Analysis Association Rule Mining Sequential Pattern Mining

assigned to the blog entries submitted by a good blogger but not yet linked to by any other blogs based on acceptance of the blogger's prior work.

Lian- Wang, et al, generated Query Dependent Ranking Algorithm this is used to measure the similarities between the queries. A single model for ranking is made for every training query with corresponding document. Whenever a query arises, then documents are extracted and ranked depending on the rank scores calculated by the ranking model. The ranking model in this algorithm is the combination of various models of the similar training queries

Web structured mining helps calculating the importance of page whereas Web Content Mining calculates relevancy of the page to the query.

2.3 Web Usage Mining

Web Usage Mining process (Chintandeep Kaur *et al 2012*) is divided into 3 phases:
# Pre processing
# Pattern Discovery
#Pattern Analysis

3: Web usage mining process

The process of analyzing the user's browsing behavior is called Web usage mining, Web Usage Mining is used to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users.

Web usage mining extracts visitor's traffic information based on Web server log files, application server data and application level data. Web

2.3.1 Web usage mining process

The process of analyzing the user's browsing behavior is called Web usage mining, Web Usage Mining (B.SanthoshKumar et al 2010)is used to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining itself can be classified further depending on the kind of usage data considered. They are web server data, application server data and application level data. Web server data correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users. Web usage mining(Jinguang Liu et al 2008) extracts visitor's traffic information based on Web server log files, application server data and application level data. Web server log files were used initially by the webmasters and system administrators for the purposes of how much traffic they are getting, how many requests fail, and what kind of errors are being generated, etc. However, Web server log files can also record and trace the visitors' on-line behaviors. Search engine logs not only keep navigation information, but also the queries made by their users.

Queries and related clicks can be used to improve the search engine itself in different aspects: user Interface, index performance, and answer ranking

Hua Jiang, et al, Developed Time rank , This algorithm is used to improve the precision of ranking Web pages, after analyzing the original Page Rank and its improved versions, record the visited time of the page to check the degree of importance to the users. It means use the time factor to improve the precision of the ranking, Which is called Time Rank. It can be treated as the combination of content and link structure in another way.

Conclusion

We provide a survey about the different web mining algorithms in search engine for ranking web pages. Web structure mining is used to rank web pages based on link structure it gives relevant pages, web content mining uses clustering and classification for ranking web pages, this gives popular (important) web pages.

Web usage mining uses visitor's traffic information, Queries and related clicks can be used to rank web pages for improve the search engine results. This paper also gives a broad idea to opt a relevant web mining algorithms according to user requirement and also gives brief description on various challenges faced in web mining when interact with web.

References

Lian- Wang Lee, Jung- Yi Jiang, Chun Der Wuand Shie-Jue Lee,(2009),Query Dependent Ranking Algorithm, Computer Science and Engineering, International Workshop on 01/2009;1:259-263.

R. Kosala and H. Blockeel,(2000) Web mining research a survey, ACM SIGKDD Explorations, 2(1):1–15.

Jiawei Han Kevin Chen-Chuan Chang,(2002) University of Illinois at Urbana-Champaign Data Mining for Web Intelligence0018-9162 /02

O.tizioni,(1996)The world wide web; Quagmire or gold mine, Communications of the ACM, 39(11):65-68.

T.Munibalaji, C.Balamurugan(2012)Analysis of Link Algorithms for Web Mining, International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 2.81-86

AliMohammad Zareh Bidoki, Nasser Yazdani,(2008) ,Distance Rank: An intelligent ranking algorithm for web pages, Information Processing and Management Pp:877–892.

Fabrizio Lamberti, Andrea Sanna, and Claudio Demartini, (2009)A Relation-Based Page Rank Algorithm for Semantic Web Search Engines IEEE transactions on knowledge and data engineering, vol. 21, no. 1, january

Taher H. Haveliwala, (2002) Topic Sensitive PageRank, WWW2002,Honolulu, Hawaii, USA.ACM 1-58113-449-5/02/0005. Stanford, CA 94305.1.

Yangzhou, Chen Zhang Hui, Sun Rong-Shuang, (2008) Tag Rank: A New Rank Algorithm for Webpage Based on Social Web, Computer Science and Information Technology, ICCSIT '08, 254 – 258, 978-0-7695-3308-7.

Takafumi Inoue, Masayuki Sugisaki, (2005)The EigenRumor Algorithm for Ranking Blogs WWW 2005, May 10--14, 2005, Chiba, Japan.

http://www.webology.org/2008/v5n2/a55.html

Niki R. Kapadia , Kinjal Patel (2012)Web content mining techniques – a comprehensive survey, IJREAS Volume 2, Issue 2 (February 2012) ISSN: 2249-3905.

Miguel Gomes da Costa Júnior ZhiguoGong(2005) Web Structure Mining: An Introduction, Proceedings of the 2005 IEEE International Conference on Information Acquisition,June 27 - July 3, Hong Kong and Macau, China.

Balan, Ponnuthurumalingam (2013)A study of content mining of various research issues and tools IJIRS, ISSN2319-9725 May, vol2 Issue5.

Darshna Navadiya, Roshni Patel (2012) Web Content Mining Techniques-A Comprehensive Survey International Journal of Engineering Research & Technology (IJERT),Vol. 1 Issue 10, ISSN: 2278-0181.

Magdalini Eirinaki web mining: roadmapDept. of Informatics Athens University of Economics and Business.

G.Hanumantha Rao, G.Narender, B.Srinivasa Rao, M.Srilatha,(2011) Web Search Engine,*International Journal of Scientific & Engineering Research* Volume 2, Issue 12, December

J.Kleinberg,(1998)Authoritative sources in a hyperlinked environment, *ACM-SIAM Symp. Discrete Algorithms.*

L.Page, S. Brin, R. Motwani, and T. Winograd (1991)The pagerank citation ranking: Bringing order to the web Technical report,Stanford Digital Libraries SIDL-WP-1999-0120.

Wenpu Xing and Ali Ghorbani,(2004)Weighted PageRank Algorithm IEEE communication Networks and service research ,PP305-314

B.SanthoshKumar,K.V.Rukmani,(2010) Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms, Int. J. of Advanced Networking and Applications Volume:01, Issue:06, Pages: 400-404.

Jinguang Liu & Roopa Datla, Web Usage Mining, Pattern Discovery and its applications, Final Project: Research Paper

Hua Jiang ; Yong-Xing Ge ; Dan Zuo ; Bing Han(2008) TimeRank: A method of improving ranking scores by visited time Machine Learning and Cybernetics, 2008 International Conference on (Volume:3)pp1654-1657

Chintandeep Kaur, Rinkle Rani Aggarwal,(2012) Web mining tasks and types: a survey *IJRIM* Volume 2, Issue(ISSN 2231-4334).

Sanjay Bapu Thakare(2010) A Effective and Complete Preprocessing for Web Usage Mining,*International Journal on Computer Science and Engineering* Vol. 02, No. 03, 2010, 848-851

MPS Bhatia, Akshi kumar(2008) *http://www.webology.org/2008/v5n2/a55.html* Webology, Volume 5, Number 2

Darshna Navadiya , Roshni Patel(2012)Web Content Mining Techniques-A Comprehensive Survey *International Journal of Engineering Research & Technology (IJERT)*Vol. 1 Issue 10, December,ISSN: 2278-0181