The International Journal Of Engineering And Science (IJES)
||Volume||2 ||Issue|| 3 ||Pages|| 25-30 ||2013||
ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805

A SURVEY ON WEB MINNING ALGORITHMS

[1,]Sandhya(M.tech CSE), [2,]Mala chaturvedi(M.tech CSE)
1,Jayoti Vidyapeeth Women's University ,Jaipur
2,Manav Bharti University ,solan `

----------Abstract----------
Web mining is data mining from web. We all know that now days data on web exceeding day. Web mining algorithms are slightly different from data mining algorithms, because on data mining the content or data is store in one particular place and we have metadata for our data. But in web mining data is distributed on various places and data is not in only text form it could text, audio, video, images and many more. The aim of this paper to study the algorithms that we have been using for web mining in current scenario and compare them with each other according to their performance.

key words: *web mining, PageRank, Weighted PageRank, HITS, clustering .latent semantic analysis, k-means, relevance ranking , CHCA.*

Date Of Submission: 8 .March, 2013 Date Of Publication: 25 March 2013

I. INTRODUCTION

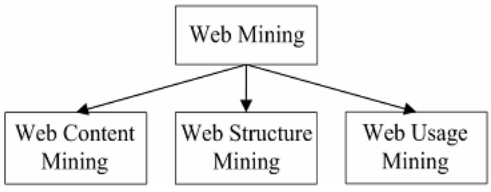
With the tremendous growth of the amount of data or information available on internet or web. World wide web is the collection of documents, images, text files and other forms of data in structured, semi structured and unstructured. The aim of web mining to extract and mine useful information and knowledge from web. Web mining is a multidisciplinary field it include: ***data mining, machine learning, natural language processing, statistics, databases, information*** *retrieval, multimedia, etc*. The data store on web is become main source of information for the users in many domain. This increase the users on web and most of the users are inexperienced. Due to heterogeneity and lack of structure of web data web mining becoming challenging task. The web is noisy it contain mixture of many kinds of information. The web is dynamic because the information on the web changes constantly keeping up with he changes and monitoring the changes are important issues on web data mining. The aim of this paper is review and analysis of the algorithms that we are using for web data mining.

II. WEB MINING

Web Mining Is The Data Mining Technique That Automatically Discovers Or Extracts The nformation From Web Documents. It Consists Of Following Tasks

[1] 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. Information selection and pre-processing: It involves the automatic selection and pre processing of specific information from retrieved web resources.
[2] 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.
[3] Generalization: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization.
[4] 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.

III. Web Mining Categories



Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the web pages. Data mining because many data mining techniques can be applied in web content mining. It is also related with text mining because much of the web contents are text, but is also quite different from these because web data is mainly semi structured in nature and text mining focuses on unstructured text. Web structure mining focuses on the hyperlink structure of the Web. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models. The goal of Web Structure Mining is to generate structured summary about the website and web page. It tries to discover the link structure of hyper links at inter document level. As it is very common that the web documents contain links and they use both the real or primary data on the web so it can be concluded that Web Structure Mining has a relation with Web Content Mining. It is quite often to combine these two mining tasks in an application. Web usage mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. It is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. Web usage mining collects the data from Web log records to discover user access patterns of Web pages. It focuses on techniques that can be used to predict the user behavior while the user interacts with the web. This activity involves the automatic discovery of user access patterns from one or more web servers. Through this mining technique we can ascertain what users are looking for on Internet. The applications generated from this analysis can be classified as personalization, system improvement, site modification, business intelligence and usage characterization.

3.1 WEB STRUCTURE MINNING:

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The link analysis algorithm contain pagerank, weighted pagerank and HITS.

3.2 LINK ANALYSIS ALGORITHMS

There are number of algorithms proposed based on link analysis. Three important algorithms PageRank, Weighted PageRank and HITS (Hyper-link Induced Topic Search are discussed below.

3.3 PAGE RANK

This algorithm was developed by Brin and Page at Stanford University which extends the idea of citation analysis. PageRank provides a better approach that can compute the importance of web page by simply counting the number of pages that are linking to it. These links are called as backlinks. If a backlink comes from an important page than this link is given higher weightage than those which are coming from non-important pages. The link from one page to another is considered as a vote. Not only the number of votes that a page receives is important but the importance of pages that casts the vote is also important.

Page and Brin proposed a formula to calculate the PageRank of a page A as stated below-

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

here PR(Ti) is the PageRank of the Pages Ti which links to page A, C(Ti) is number of outlinks 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.

The PageRank of a page can be calculated without knowing the final value of PageRank of other pages. It is an iterative algorithm which follows the principle of normalized link matrix of web. PageRank of a page depends on the number of pages pointing to a page.

3.4 Weighted Page Rank

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of PageRank algorithm.This Algorithm assigns rank values to pages according to their importance rather than dividing it evenly. The importance is assigned in terms of weight values to incoming and outgoing links.

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). It is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m.

$$\mathbf{w}^{in}_{(m,n)} = I_n / \sum_{p \in R(m)} I_p \qquad \text{.....(2)}$$

In 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.

$$W^{out}_{(m,n)} == O_n / \sum_{p \in R(m)} O_p \qquad \text{----(3)}$$

O_n is number of outgoing links of page n, O_p is number of outgoing links of page p,

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

$$WPR(n)=(1-d)+ d \sum_{m \in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)} \qquad \text{....(4)}$$

3.5 HITS (Hyper-link Induced Topic Search)

Klienberg gives two forms of web pages called as hubs and authorities. 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.The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Figure 1 shows the hubs and authorities in web.

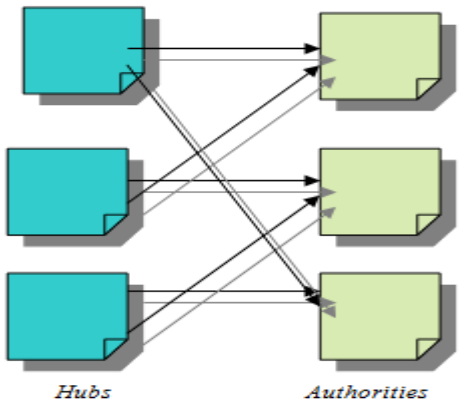


Figure 1. hubs and authorities

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. *Following expressions (7,8) are used to calculate the weight of Hub* (Hp) *and the weight of Authority* (Ap).

$$H_p = \sum_{q \in I(p)} \mathbf{A}_q \qquad \textbf{....(7)}$$

$$A_p = \sum_{q \in B(p)} H_q \qquad \text{.....(8)}$$

here Hq is Hub Score of a page, Aq is authority score of a page, I(p) is set of reference pages of page p and B(p) is set of referrer pages of page p, the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

IV. WEB USAGE MINNING

Discovery of meaningful patterns from data generated by client-server transactions on one or more Web servers. Typical Sources of Data:

1. Automatically generated data stored in server access logs, referrer logs, agent logs, and client-side cookies
2. E-commerce and product-oriented user events (e.g. shopping cart changes, ad or product click-throughs, etc.)

3. User profiles and/or user ratings
4. Meta-data, page attributes, page content, site structure.

4.1. Clustering

Clustering the process of partition a set of data in a set of meaning full subclasses known as clusters. It helps users understand the natural grouping or structure in a data set.Clustering is an unsupervised learning technique which aim is to find structure in a collection of unlabeled data. It is being used in many fields such as data mining, knowledge discovery, pattern recognition and classification.Central clustering algorithms are often more efficient than similarity-based clustering algorithms. We choosecentroid-based clustering over similarity-based clustering.We could not efficiently get a desired number of clusters, e.g., 100 as set by users. Similarity-based algorithms usually have a complexity of at least O (N2) (for computing the datapair wise proximity measures), where N is the number of data instances.In contrast, centroid-based algorithms are more scalable, with a complexity of O (NKM), where K is the number of clusters and M the number of batch iterations. In addition, all these centroid-based clustering techniques have an online version, which can be suitably used for adaptive attack detection in a data environment.

4.2. K-Mean Algorithm

The K-Means algorithm is one of a group of algorithms called partitioning clustering algorithm. The general objective is to obtain the partition that, for a fixed number of clusters, minimizes the total square errors. Suppose that the given set of N samples in an n-dimensional space has somehow been partitioned into K-clusters {C1, C2,C3... CK}. Each CK has nK samples and each sample is in exactly one cluster, so that Σ nK = N, where k=1… K. The mean vector Mk of cluster CK is defined as the centroid of the cluster.

$$M_K = (1/n_k)\sum_{i=1}^{n_k} x_{ik} \quad \text{------(1)}$$

Where xik is the ith sample belonging to cluster CK. The square-error for cluster CK is the sum of the squared. Euclidean distances between each sample in CK and its centroid. This error is also called the within-cluster variation

$$e_k = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2 \quad \text{---(2)}$$

The square-error for the entire clustering space containing K cluster is the sum of the within-cluster variations.

$$E^2_k = \sum_{k=1}^{k} e^2_k \quad \text{-----(3)}$$

The basic steps of the K-mean algorithm are:

- Select an initial partition with K clusters containing randomly chosen sample, and compute the centroids of the clusters.
- Generate a new partition by assigning each sample to the closest cluster center.
- Compute new cluster centre as the centroids of the clusters.
- Repeat steps 2and 3 until optimum value of the criterion function is found or until the cluster membership stabilizes.

4.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) model is an approach to capture the latent or hidden semantic relationships among co-occurrence activities and has been widely used in web information management. For example, based on LSA method, the latent semantic relationships among web pages can be discovered from linkage information, which will lead to find relevant web
pages and improve web searching efficiency and effectivity Factor analysis technique has become another LSA-based web usage mining approach recently,
for example, Principal Factor Analysis (PFA) model is proposed to extract web user session or web page clusters and reveal the latent factors associate with user access patterns.

V. WEB CONTENT MINNING

It is the process of retrieving the information from WWW into more structured forms and indexing the information to retrieve it quickly. These algorithms are use for web content mining.

5.1. Correlation algorithm for relevance ranking

Input : Web document D = {D1, D2,.....DN}
Method : Correlation method
Output : Relevant documents.
Step 1: Extract the input web documents D_i where $1 \leq i \leq N$.
Step 2: Pre-process the entire extracted documents.
Step 3: Initialize redundant document set RD = {};
Step 4: Initialize i = 1
Step 5: Initialize j = 1.
Step 6: Perform the correlation coefficient R_{ij} between D_i and D_j
If i = = j then $R_{ij} = 0$ Goto step 1
Else
Compute the following steps :
Extract the common words between D_i and D_j that matches with domain dictionary. Let T be the set of common words and the number of elements in the
T be m, i.e. $|T| = m$.
Compute the term frequency TF $(W_k)_i$ in D_i and TF $(W_k)_j$ in D_j where $1 \leq k \leq m$.
Determine:
X_k = TF $(W_k)_i$ for the words in document D_i and
Y_k = TF $(W_k)_j$ for the words in document D_j
Calculate: $\sum X_k$, $\sum X^2_k$, $\sum Y_k$, $\sum Y^2_k$, $\sum X_k Y_k$,
Compute: R1, R2 and R3

$R_1 = \sum X^2_k - \sum X^2_k / |T|$
$R_2 = \sum Y^2_k - \sum Y^2_k / |T|$
$R_3 = \sum X_k Y_k - \sum X_k \sum Y_k / |T|$ *perform:*

$R_{ij} = R_3 / \sqrt{R_1 x} \sqrt{R_2}$

Step 7: If (R_{ij} = = 1) then D_i and D_j are redundant;
Assign $RD_i = RD_i \cup D_j$ where $1 \leq i \leq N$.
else D_i and D_j are not redundant;
Step 8: Increment j, and repeat from step 6 to step 7 until $j \leq N$.
Step 9: Compute the total correlation coefficient:
$\sum R_{ij}$ where j=1 to N.
Step 10: Increment i, and repeat from step 5 to step 9 until $i \leq N$.
Step 11: Sort total correlation coefficient in descending order.
Step 12: Remove redundant data set (RD).
Step 13: Display the top 'n' relevant documents.

5.2 Cluster Hierarchy Construction Algorithm (CHCA)

The algorithm takes a binary matrix (a table) as input. The rows of the table correspond to the objects we are clustering. Here we describing this algorithm with web pages, but the method is applicable to other domains as well. The columns correspond to the possible attributes that the objects may have (terms appearing on the web pages for this particular application). When row i has a value of 1 at column j, it means that the web page corresponding to i contains term j. From this table, which is a binary representation of the presence or absence of terms for each web page, we create a reduced table containing only rows with unique attribute patterns (i.e., duplicate rows are removed). Using the reduced table, we create a cluster hierarchy by examining each row, starting with those with the fewest terms (fewest number of 1's); these will become the most general clusters in our hierarchy. The row becomes a new cluster in the hierarchy, and we determine where in the hierarchy the cluster belongs by checking if any of the clusters we have created so far could be parents of the new cluster. Potential parents of a cluster are those clusters which contain a subset of the terms of the child cluster. This comes from the notion of inheritance discussed above. If a cluster has no parent clusters, it becomes a base cluster. If it does have a parent or parents, it becomes a child cluster of those clusters which have the most terms in common with it. This process is repeated until all the rows in the reduced table have been examined or we create a user specified maximum number of clusters, at which point the initial cluster hierarchy has been created. The next step in the algorithm is to assign the web pages to clusters in the hierarchy. In general there will be some similarity comparison between the terms of each web page (rows in the original

table) and the terms associated with each cluster, to determine which cluster is most suitable for each web page. Once this has been accomplished, the web pages are clustered hierarchically. In the final step we remove any clusters with a number of web pages assigned to them that is below a user defined threshold and re-assign the web pages from those deleted clusters.

VI. CONCLUSIONS

In this paper we study the various algorithms that are used for web mining and its three categories. In web structure mining we are using PageRank algorithm, weighted PageRank algorithm. In web usage mining we are using clustering, k-means algorithm and latent semantic analysis. In web content mining we are using Correlation algorithm for relevance ranking and Cluster Hierarchy Construction Algorithm (CHCA). These all algorithm makes web mining more useful and easy for user. These algorithm made easy search for any document on internet. There are several more advance version of these algorithms are have been produced. In this paper we had reviewed the popular algorithms of web mining to have an idea about in their application and effectiveness. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful for identifying opportunities for further research.

REFERENCES

[1] Page Ranking Algorithms for Web Mining,Rekha Jain Department of Computer Science, Apaji Institute, Banasthali University C-62 Sarojini Marg, C-Scheme, Jaipur,Rajasthan Dr. G. N. Purohit Department of Computer Science, Apaji Institute, Banasthali University.

[2] Graph-theoretic techniques for web content mining,Adam Schenker, University of South Florida 3. ON TWO ALGORITHMS USED IN WEB STRUCTURE MINING,Claudia Elena Dinucă Ph. D Student,University of Craiova Faculty of Economics and Business Administration Craiova, Romania Dumitru Ciobanu Ph. D Student University of Craiova Faculty of Economics and Business Administration Craiova, Romania.

[3] Performance Improvement Of Web Usage Mining By Using Learning Based K-Mean Clustering , Ms. Vinita Shrivastava M.Tech (Information Technology) Technocrats Institute of Technology, Mr. Neetesh Gupta Head Of Department (Information technology) Technocrats Institute of Technology,Bhopal india.

[4] J. Hou and Y. Zhang, Effectively Finding relevant web Pages from Linkage Information, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003

[5] P Ravi Kumar, and Singh Ashutosh kumar, Web Structure Mining Exploring Hyperlinks an algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010. .

[6] Semantic Web Mining and the Representation, Analysis, and Evolution of Web Space Bettina Berendt1, Andreas Hotho2, and Gerd Stumme2 1Institute of Information Systems, Humboldt University Berlin, D-10178 Berlin, Germany, http://www.wiwi.hu-berlin.de/_berendt 2 Knowledge and Data Engineering Group, University of Kassel, D-34121 KasselGermany, http://www.kde.cs.uni-kassel.de/[hotho|stumme

[7] C.C. Aggarwal. Collaborative crawling: Mining user experiences for topical resource discovery In [42], pages 423–428, 2002.

[8] C.C. Aggarwal, F. Al-Garawi, and P.S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In Proceedings of the WWW Conference, 2001.

[9] C.C. Aggarwal, S.C. Gates, and P.S. Yu. On the merits of building categorization systems by supervised clustering. In KDD'1999 – Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 352–356, 1999.

[10] Web Content Mining Bing LiuDepartment of Computer Science University of Illinois at Chicago (UIC) liub@cs.uic.edu http://www.cs.uic.edu/~liub.

[11] Improved FCM algorithm for Clustering on Web Usage Mining K.Suresh1 R.Madana Mohana2 A.RamaMohanReddy3 1 Department of Software Engineering, East China University of Technology,ECIT Nanchang Campus, Nanchang, Jiangxi-330013, P.R.China. 2 Department of Information Technology, Vardhaman college of Engineering, Shamshabad, Hyderabad, A.P, India. 3 Department of Computer Science and Engineering , S.V.University College of Engineering, Tirupati, A.P, India.

[12] Web Usage Mining: A Research Area in Web Mining Rajni Pamnani, Pramila Chawan Department of computer technology, VJTI University, Mumbai.

[13] Relevance Ranking and Evaluation of Search Results through Web Content Mining G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma, K.Sarukesi. Using Probabilistic Latent Semantic Analysis for Web Page Grouping Guandong Xu1,3, Yanchun Zhang1, Xiaofang Zhou2 1School of Computer Science and Mathematics Victoria University, PO Box 14428, VIC 8001, Australia {xu,yzhang} @csm.vu.edu.au 2 School of Information Technology & Electrical Engineering University of Queensland, Brisbane QLD 4072, Australia zxf@itee.uq.edu.au 3School of Computer Science & Engineering Wenzhou University, Wenzhou 325003, China xgd@wznc.zj.cn.

[14] Web Structure Mining: An Introduction Miguel Gomes da Costa Júnior Zhiguo Gong Department of Computer and information Science Faculty of Science and Technology University of Macau Av. Padre Tomás, S.J., Taipa, Macao S.A.R., China {mcosta, fstzgg}@umac.mo.