Comparative Study of Web Mining Algorithms

Rachna Singh Bhullar
Computer Science Department,
Guru Nanak Dev University,
Amritsar, Punjab, India

Dr. Praveen Dhyani
Banasthali University – Jaipur Campus, Jaipur
Rajasthan, India.

Abstract: Every second data is shared between web servers and web users either in the form of uploading or downloading to or from the web sites. In both the cases web servers play a very important role of providing the most efficient and relevant data from the web databases. To accomplish this task various web mining algorithms are used by web servers to satisfy the need of web users. In this paper we are presenting an overview of web mining algorithms and a comparative study among them. In the last section of this paper, we are suggesting some questions which cannot be answered by any of the web content or web usage mining algorithms which motivated us to implement the web structure mining algorithms.

Keywords- Web Usage Mining, Web Content Mining, Web Structure Mining, Web Content Outlier Mining, Web Mining Algorithms.

1. INTRODUCTION

In today's web scenario, web servers are behaving like the interface between the web databases and web users. The job of web server is to upload or download the data from databases. As a result, demand of most relevant and efficient data from the web databases and storage of data on the web databases is increasing exponentially. In such a situation, task of retrieving the efficient data is extremely challenging. Web mining is the collection of those data mining techniques which are best suited to perform such a challenging task.Web mining can be categorized into Web Usage Mining, Web Structure Mining, and Web Content Mining. These are described as follows:-

Web Usage Mining (WUM):-

WUM consists of various web mining techniques to be applied on user access patterns and their profiles stored in web server logs [1]. The results obtained from WUM are used by E-commerce agents to improve their interaction with the customers by analyzing their usage behavior. The web usage mining process is explained as below [2].

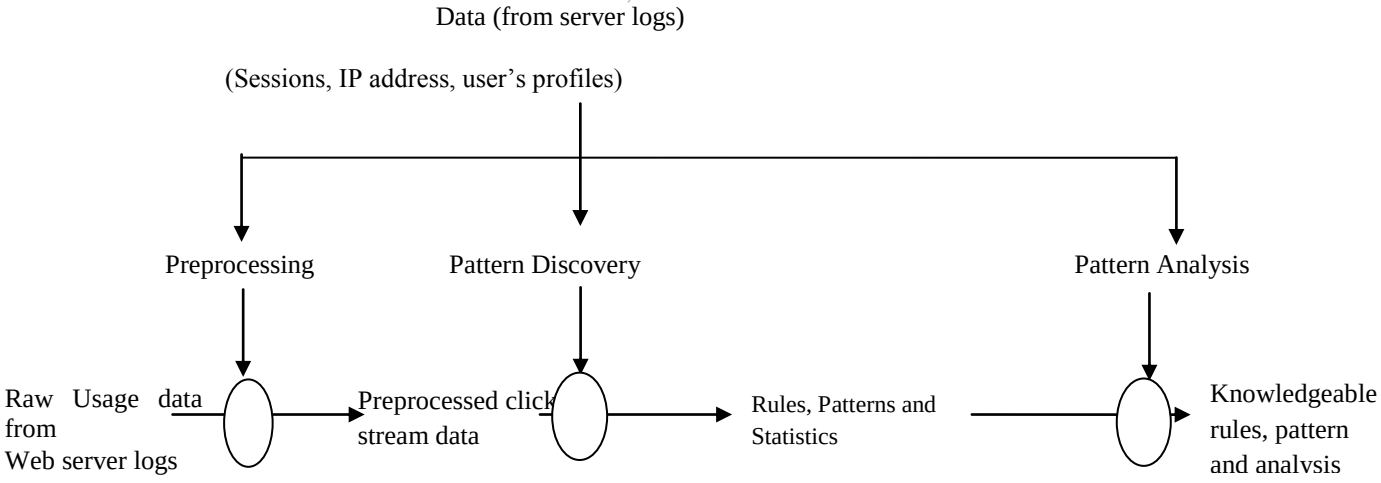

Fig 1

Applications of Web Usage Mining:-

WUM mainly involves Association rules, Clustering & Classification as sub-processes which are useful in:

(a) Protection of next event.

(b) Discovery of web page visitors.

(c) Grouping of visitors on the basis of common properties, interests, & common behavior.

Page recommendation is provided to web user only with the help of WUM techniques.

WUM is also helpful in fraud and intrusion detection.

Dynamic recommendations for friends in social network sites or for purchases on E-commerce sites are one of the mostly used applications of WUM.

WUM mining techniques can be applied to encourage personalized marketing in E-commerce site.

Web usage Mining Algorithms:-

Apriori algorithm is a common data mining technique for association based analysis. WUM phase includes data selection, data preprocessing, pattern discovery and analysis. Apriori algorithm is the part of pattern discovery sub function.

An average linear time algorithm for WUM was suggested by Mark Levene and Jose Borges to show the linear proportionality between the behavior of algorithm and no. of web pages accessed.

Web Content Mining (WCM):-

WCM is one of the specialized fields of web mining which deals with text, image, audios, videos, patterns, hyperlinks, as the content of the web page. Further, mining techniques specifically applied on text are known as text mining. Pattern analysis and spatial mining deals with the images and patterns. Mining techniques applied on video & audio signals fall into the category of Multimedia mining. If the content of a web page contains hyperlinks to the related web pages then WCM is categorized as Web Structure Mining. The following figure summarizes the WCM as a whole:

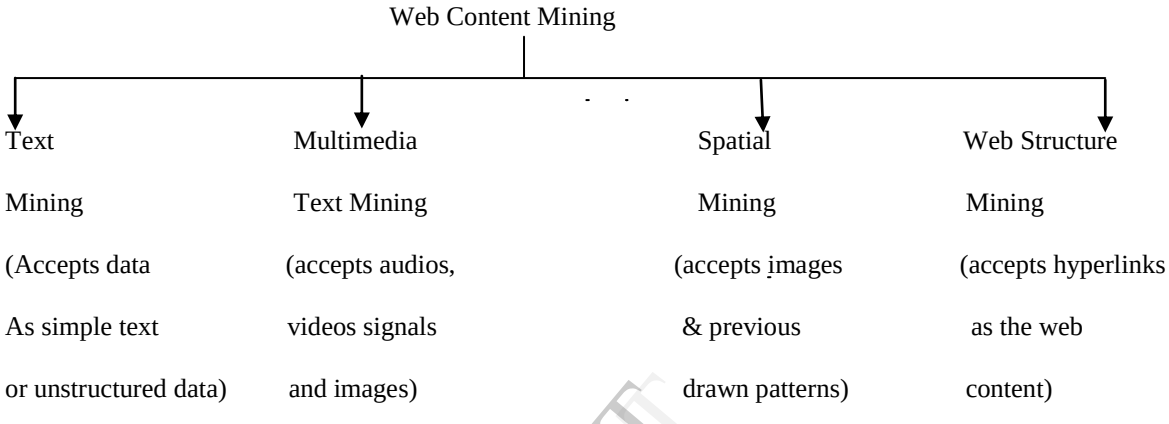

Figure2

In brief, WCM discovers new inferences or useful information. It is to be noted that simple text mining, Multimedia Text Mining and Spatial Mining are three different fields, except the web page because the data available through web is either unstructured or semi-structured. Generally, text and spatial mining deals with structured text and images. Web content mining deals with the semi-structured or unstructured data.

Applications of Web Content Mining:-

WCM is a vast research field but algorithms and methods are applicable in various fields as explained below:

Ranking:-

WWW is considered as a graph of connected web pages. To retrieve the related ones is a challenging task; Search engines make use of Ranking algorithms to rank the list of related web pages on the basis of content to be searched and their popularity.

E-marketing:-

E-commerce is getting popular day by day so in order to improve the interaction between the customer and E-company and to recommend related options for their products, they make extensive use of content mining algorithms as discussed ahead in this chapter.

Fraud Detection:-

ATM cards, Debit-Shopping cards, Credit cards etc. obtain their authority after being checked through content mining algorithms embedded in their respective machines.

Communities:-

Social networking sites like Twitter, Facebook, Google and LinkedIn all make use of WCM algorithms to give authenticity to their users and their frauds.

Web Structure Mining:-

WWW is considered as a web or directed graph having millions of nodes as the web pages. The hyperlinks to the related pages are the edges in the directed graph. This graphical information is used by various Ranking algorithms like Page Rank, Weighted Page Rank and Hypertext induced topic search algorithms to find the related web pages for a search query. The research on WSM can be carried out in two ways. First one deal with the structure of the web document, that is, structure of the HTML program to construct a web page.

Second one considers the structure of hyperlinks in the address part of a web page. In our research, we mainly focus on the hyperlinks and their structures. Hyperlinks in a web page, to another web page denote the relevancy between the two pages and are used to rank the importance of the referred web page. In the statistics or numerical value analysis, values lying beyond some fixed range are known as outliers whereas in web mining, unauthentic, unauthorized and non related web pages are treated as outliers. In traditional web outlier mining, web content mining algorithms are specially focused on the text in web pages. There exist various algorithms like Signed approach for Web Content Outlier Mining [3] [4], Classic outlier detection from web clusters using dissimilarity measure [5], Chinese Web text outlier mining based on domain knowledge[6] and Web content outlier-N-gram based algorithm [7].

Outline of work
Section II provides the brief review of related work in web content mining. Section III explains the proposed study. Section IV provides the results of the study while in section V future work is summarized.

2. RELATED WORK

Web Content Outlier Mining Algorithms

1.Signed approach for Web Content Outlier Mining [4].

G.Poonkuzhali, K.Thaigrajan, K,Sarukesi and G.V.Uma proposed a system in which user provides a search query to search engine. As a result, 'D' web documents are extracted from the web which in turn is preprocessed, that is the Stop word, Stun word and data like hyperlinks, sound, images etc. are removed. The output set of documents containing white-space is represented in two-dimensional format as (i, j), where 'i' represents the web page and 'j' represents word. The domain is related by taking into account such that all 1-letter word will be indexed first, followed by 2-letter words, then 3-letter words similarly upto 15-letter words which are a reasonable upper bound for no. of characters in a word. Second step is to mine each page individually to detect relevant and irrelevant documents using the Signed approach of full word matching in organized domain dictionary. Finally, a relevant web document is obtained which contains the required information.

2.Classic outlier detection from web clusters using dissimilarity measure [5].

E.Sateesh and M.L.Prasanthi suggest an improved version of Signed approach for web content outlier detection in which a dissimilarity measure is completed to determine the difference among the pages within the same category [8].

$$DM_i=\sum_{i,j}[e_j(0.5+(0.5*f(t_j,d_i)/maxfreq(d_i))*\log_{10}(N/K)]/e_i$$

where,

F (t_j,d_i)=frequency of term $t_j$in d_i, document

Maxfreq (d_i)=maximum frequency of a word in the document.

N=total number of documents.

K=number of documents with term t_j.

e_i=shows the words in the document that exist in the domain dictionary.

Finally, this paper identifies the outliers as the irrelevant documents from the web clusters. The documents with minimum dissimilarity are the relevant ones.

3. WCOND-mine algorithm (Web Content Outliers-ngram) mine algorithm :-

Malik Agyemang proposes this algorithm for mining web content outliers without using the domain dictionary. In this paper, he is making use of N-gram technique to find the outliers. The total no. of possible N-grams resulting from a string of length K is (K-N+1), where N=size of N-gram e.g. a string 'sciences' of length 8 has (8-5+1)=4, 5,-grams (where K=8 and N=5) as 'scien', 'cience' and 'ences'. The N-gram technique is used to find the related but different words. According to this N-gram technique, if two words having N-grams of length greater than half the length of string in common then these different words are syntactically similar words. To illustrate the above statement, take an example of 'commodity' and 'community' in which eight 2-grams are common but no common 5-gram, which concludes that they are not related words. N-gram frequency profile for each document in the set of documents obtained from a web search engine is generated separately. Dissimilarity between the documents of a set is computed and ranked and then the top-N documents with highest dissimilarities are declared as Outliers. There are three different phases of this algorithm.

Document Preprocessing:-

This phase generates the set of documents of the same domain and removes the tags other than <Title>, <Meta>, <Body> toys plus the stop words. The output obtained at the end of preprocessing phase is the set of documents

containing white spaced separated words under various HTML tags.

Generate N-gram frequency profile:-

1-gram and 2-gram are not used here as they are used to show the distribution of letters, common suffixes and prefixes in the alphabet of the language, so only 4-gram & 5-gram are used as the desired N-gram sizes. The N-gram frequency profile is computed as follows:

(i) Text obtained from preprocessing phase is tokenized.

(ii) Tokens obtained in Step (1) are used to make 4-grams and 5-grams.

(iii) Frequency of each 4-gram and 5-gram is maintained with its original word into a Hash table.

Finally, sorted N-grams and their frequencies are maintained into an output file.

Computation of dissimilarity:-

N-grams obtained in step (ii) are associated with weights W_{ik} representing the weight of N-gram N_k in document Di, computed by Salton's term-weighting formula [9].

$$W_{ik}=tf_{ik}*\log (N/w_k)/ \sqrt{(\sum_i (tfik)^2 * (\log (N/N_k))^2)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where,

$tf_{ik=}$ frequency of n-gram N_k in the document D_i

N = size of D_i

n_k = number of documents having n-grams N_k

The denominator in the above formula is used for content normalization to avoid the documents with large number of n-grams to be assigned with high similarities. N-grams of D_i appearing in the metadata (i.e. <Title>, <Meta>) are assigned larger weights than the n-grams of body tag. The weights are assigned as follows:

$$w(N_{kit}) = \begin{cases} 1 \text{ if } N_k \text{ € metadata tag} \\ \text{otherwise, } 0< \quad < 1 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where $w(N_{kit})$ is the weight assigned to n-gram N_k in D_i in HTML tag H_t. C_{ik} is the number of times N_k appears in H_t.

Thus, the term frequency tf_{ik} in equation (1) is computed as follows:

$$tf_{ik}=\sum C_{ik}.w(N_{kit}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

where, $w(N_{kit})$ is the weight of n-gram N_k computed as above in (2).

For example, we have two documents d_i and d_j in the set of related documents obtained earlier. These d_i and dj are represented in vector space model

Then the dissimilarity between $\vec{d_i}$ and $\vec{d_j}$ is computed as:

$$DIS(di,dj)=1-\left((\sum_k w_{ik} w_{jk})/ (\sqrt{\sum} w_{ik}^2)(\sqrt{\sum} w_{jk}^2) \right) \quad \ldots\ldots\ldots\ldots\ldots (4)$$

It states that:-

Two vectors are highly similar if d_i and d_j are orthogonal vectors.

If DIS=0 or near to zero, then lower is the dissimilarity and they are the related one.

If DIS=1 or closer to one, then the probability of their dissimilarity is highest.

Experimental results obtained in [7] show that this algorithm is capable of finding out the outlier from large web datasets.

3. PROPOSED STUDY:

COMPARATIVE STUDY OF WEB CONTENT MINING ALGORITHMS IS AS SHOWN BELOW:

Algorithms	Signed Approach	Classic Outlier Detection using Dissimilarity	WCOND mine Algorithm
Web datasets contains	White spaced text only	Measure white spaced text only	White spaced text & HTML tags <Title> <Head> <Body>
Data structures used	Domain Dictionary	Domain Dictionary	Frequency profile is created for N-grams
Vector space mode	No, datasets only	Datasets only	Yes, documents containing n-grams are represented as vectors.
Output Parameters	Related documents in Output file.	Related & Non related both	Angle value as measurement of dissimilarity
Techniques used	WCM	WCM	WCM

Table1

4. RESULTS

Issues and Problems in Web Content Outlier Mining Algorithms:-

The comparative study of all the above explained algorithms concludes that these algorithms are applied to web text (like white spaced words) only whether they are contained in body part or in meta tags like <Title>,<Head>,<Meta> tags.

But,

(1) What about the web content like images, patterns and hyperlinks etc. as the web search engine results into the set of hyperlinks to the related web pages first.

(2) What about the authority of the hyperlinks i.e. whether they are coming from a good hub or not.

(3) Whether they actually exist or are the fake ones.

Future Work:

All these questions can be answered by implementing the web structure outlier mining algorithms which is a very vast research area to be focused.

REFERENCES:

Power point presentations

1. Web Usage Mining by Margaret H. Dunham.
2. Web Mining by Margaret H. Dunham.

Publications:

3. Semantic Web Mining of Unstructured Data: Challenges and Opportunities published in International Journal of Engineering volume5, issue3, 2011.
4. Signed Approach for mining web content Outliers by G.Poonkuzhali in World Academy of Science, Engineering and Technology.
5. Web Usage Mining-Languages and Algorithms by John R. Punin, Mukkai S. Krishnamoorthy, Mohammed J. Zaki.
6. Chineese Web text Outlier Mining based on Domain Knowledge by Xia Theosong in 2010 Second WRI Global Congress on Intelligent Systems.
7. WCOND-MINE: Algorithm for detecting Web Content Outliers from Web Documents by Malik Agyemang in proceeding sof 10 th IEEE Sympossium computers and communications.
8. Web Mining - Concepts,Applications & ResearchDirections by *Jaideep Srivastava, Prasanna Desikan, Vipin Kum*