An Implementation of Web Usage Mining on DSpace Website Structure

Ashish Jaiswal#1, Bireshwar Ganguly#2, Devashri Raich#3
#1Department of Computer Science & Engg, P.I.G.C.E, Nagpur, M.S, India
#2Department of Information Tech, R.C.E.R.T, Chandrapur, M.S, India
#3Department of Information Tech, R.C.E.R.T, Chandrapur, M.S, India

ashish.jaiswal2783@gmail.com
bireshwar.ganguly@gmail.com
devashrirai@gmail.com

Abstract --Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. Web usage mining consists of three phases, pre-processing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use these information for the specific needs. In this project, the DSpace log files have been pre-processed to convert the data stored in them into a structured format. Thereafter, the general procedures for bot-removal and session-identification from a web log file, have been written down with certain modifications pertaining to the DSpace log files, in an algorithmic form. Furthermore, analysis of these log files using a subjective interpretation of a recently proposed algorithm EIN-WUM has also been conducted. This algorithm is based on the artificial immune system model and uses this model to learn and extract information present in the web data i.e server logs. This algorithm has been duly modified according to DSpace Website structure.

Keywords: Web Mining, Dspace,

LITERATURE REVIEW

Web usage mining is divided into two tiers of tracking and analysis in order to find out the user access pattern. The resultant of web usage mining which is knowledge about user access pattern is useful in various applications. Due to information overload on the web its full economic potential has not been realized completely. Web personalization is introduced as a solution to this problem in [23]. A complete framework for web usage mining and FM model is introduced in order to analyze the access pattern of user. The framework has been capable in doing implicit tracking which handles sparsity problem in web personalization and capturing short term changes in user behaviour. Web usage mining is being used in various spheres. In [24] web usage mining is used for improving the scalability and answer time of search engines. A simple storage model is proposed in which main
memory is assigned with the function of doing dynamic caching of the answers and of inverted lists that are available in secondary memory. The possibility of web usage mining on proxy servers is discussed in [25]. Collection of data from server log does not include the page views generated as a result of cache. Because of this, behaviour of multiple users can be analyzed only on a single site. In order to analyze the access pattern of multiple users on multiple site proxy server log data has been used. Data from several different sectors such as TV, radio, newspaper and finance has been analyzed and compared by making use of e-metrics. The result shows that proxy servers contain a large amount of interesting information which has been used for finding the navigational pattern of users. In [26] pattern recognition strategy has been used for developing web personalization based on web usage mining. This strategy helps in analyzing both static and dynamic features. This strategy comprises of two phases: in first phase unsupervised clustering algorithms has been used for performing pattern analysis and classification. In second phase reclassification is performed to overcome the inaccuracy of registration information.

A framework for mining of client side web usage data called personal web usage mining is proposed in [31]. Author considers client side data because it will provide individualism and personalization of data whether server side considers a group of data for performing web usage technique. Another problem was that it is also possible that a group of user do not have same interest so server side mining proves to be inefficient in this respect. The framework includes four modules: logging, data warehousing, data mining and tool application. Logging stores the activities of user in activity recorder and cache pages. The logs are cleansed, transformed and stored in the data warehouse. In data mining useful information is extracted which is used in tool application.

In [32] author proposes a new technique of web usage mining called self organized map (SOM), a type of artificial neural network for determining the access patterns of users. The problems that are faced in performing web usage mining include processing and cleaning of logs, user sessions identification and user habit identification. The solution proposed is SOM that helps in identification of user habits. The input to SOM is 0 or 1 vector that shows presence or absence of a page and output is a map of N*N matrix. The SOM is also compared with another system called K-means for two different websites and result shows that SOM is better in terms of collection of users and pages.

Web usage mining proves to be useful in the area of enhancing the usage of web based learning environment [29].

III. IMPLEMENTATION

1. A brief pseudo Code of the algorithm is as follows (as taken from [2])
2. Initialize antibodies using some of input data (sessions)
3. Construct neighbourhoods using a simple clustering method.
4. Set the neighbourhood threshold to average dissimilarity between cluster prototypes.
5. For each antigen (in coming sessions)
   a. Calculate danger level of antigen (interestingness of session), if danger level of antigen is more than a threshold continue, else go to 4.
   b. Present antigens to cluster prototypes.
   c. Choose the most activated neighbourhood.
   d. If affinity between antigen and selected neighbourhood prototype is less than neighbourhood threshold add a new neighbourhood with a copy of antigen to the network, update neighbourhood information and go to step 4.
   e. Else calculate stimulation level of antibodies in the selected neighbourhood, update neighbourhood information and update antibodies’ vector according to flaw or supplementary state.
   f. Clone antibodies. After processing every T antigen, mutate antibodies; add the new antibodies to the network.
   g. Delete excess antibodies with least stimulation level, move stagnated antibodies to second memory.

Interpretation:-
Our Interpretation of the algorithm is as follows:-

Limit value of no. of antibodies to 6 (based on the category from Dspace Website. We define the category of each entry in the Server Log by assigning it a number

- (0 through 6). The numbers signify
- 0 – Default value
- 1 - Content searched by title
- 2 - Content searched by author
- 3 - Content searched by date
- 4 - Content searched by author
The antibodies are initialized from the first 10 sessions. For each session an entry goes to the corresponding number of antibody as its category is. So each antibody contains only one category of server log entry. For each in coming session

- Compare with each existing antibody
  - If (Similarity of antibody > threshold)
    - Replace old session with new session
  - Else if (similarity < threshold)
    - Update antibody with most similarity
    - Put a limit on the size of antibody
    - If (antibody crosses limit)
      - Delete old entries.
    Here Danger level[2] or interestingness is based on content.

**Utility:-**
Utility of the above Interpretation
a) At the end of the program, the ten most interesting antibodies will remain.
b) The contents accessed in the antibodies will be the most frequently accessed contents in the whole website.
c) Based on (b) the following changes can be brought to the concerned site
   i) Improvements on frequently accessed pages
   ii) Deletion or merging of unused pages
   iii) Improvement of content
   iv) Improve interaction with referral sites.

IV. RESULTS
The results obtained from our analysis include
A. Pre-processed log files information e.g
1 true 203.129.199.129 10/Jan/2010:04:04:26 GET 200 17013B
0 /dspace/browse-title?top=2080%2F905
2 true 203.129.199.129 10/Jan/2010:04:04:29 GET 200 14295B
0 /dspace/browse-author?bottom=Misra%2C+M

The data separated by spaces in the above examples denote the various details about the user hits. The second data item represents whether the hit has come from a bot or a human user.

B. Summary of the log file giving overall details as an example shown below

33
***************Summary**************
********
number of hits = 14274
number of visitor hits= 7923
number of spider hits= 6351
Number of days= 5
Average hits per day = 2854
Total Bandwidth used = 1494419892 Bytes
Average Bandwidth= 29883978 Bytes
***************
********
C. The sessions and the different log file entries that constitute of the sessions as shown below
session 1 182
session 1 183
session 2 193
session 2 219
session 2 215
session 2 216
session 2 217
session 2 218

D. The different frequently accessed contents in the Dspace website for example
16 0 1 /dspace/browse-author?starts_with=Das%2C+Atanau
92 0 1 /dspace/browse-author?bottom=Wai%2C+P+K+A
181 0 1 /dspace/browse-author?starts_with=Verghese%2C+L
227 1 1 /dspace/browse-author?top=Joshi%2C+Avinash
364 1 1 /dspace/browse-author?top=Joshi%2C+Avinash
527 5 1 /dspace/browse-author
530 5 1 /dspace/browse-author?starts_with=C
532 5 1 /dspace/browse-author?top=Chatterjee%2C+Saurav
536 5 1 /dspace/browse-author?starts_with=S
569 7 1 /dspace/browse-author?starts_with=S
571 7 1 /dspace/browse-author?top=Chatterjee%2C+Saurav
715 8 1 /dspace/browse-author?top=Bal%2C+S
748 8 1 /dspace/browse-author?starts_with=Das%2C+B+M
831 8 1 /dspace/browse-author?starts_with=Karanam%2C+U+M+R
This sample result has been taken from the final status of the antibody 2 (with 2nd category entries. For categories see section 4.2)

V. CONCLUSION AND FUTURE SCOPE

The research and implementation that we have presented in this thesis is in a nascent stage and is purely DSpace Website specific. However this subjective interpretation of the algorithm EIN-WUM [2] is very ingenious and proposes a lot of scope to be extended on to other problem domains. Furthermore, anyone interested in this field can take a similar approach and modify these methods to expand them to a general scenario or to shift their use to a different area. The proposed methods were successfully tested on the log files for bot removal and user sessions identification. The results which were obtained after the analysis were satisfactory and contained valuable information about the Log Files. The subjective interpretation and implementation of the EIN-WUM [2] algorithm produced results which depicted the usage patterns (frequently accessed contents) of the DSpace website users.

REFERENCES

[7]. I. Muslea,”Extraction Patterns for Information Extraction tasks: A Survey”, In AAAI-99 Workshop on Machine Learning for Information Extraction, 1999