# A Study on Web usage Data Mining in Online Sales and SASF Crawler in Online Advertisement

M. Anisha
Assistant Professor
KPR Institute of Engineering and Technology
Coimbatore, India

P. Joyce Beryl Princess
Assistant Professor
KPR Institute of Engineering and Technology
Coimbatore, India

## ABSTRACT
In the world of Internet, users always expects the services, provided by the service providers must be easy to access. When online users need to retrieve information while doing online sales, then the ease of use depends on the frequency of the items that are available in the product recommendations. To achieve this, the crawlers are used to retrieve and download the required information from the web pages. To improve the performance of product recommendations Self Organizing Maps are used. A study has been made about Crawlers and Self Organizing Map.

## General Terms
Crawler, Online Sales, Ontology

## Keywords
SASF crawler, WebUsage Mining, pattern discovery, FAB

## 1. INTRODUCTION
Traditionally in Web Mining, crawlers has been used for information retrieval from the World Wide Web. In the earlier stages of crawler evolution, crawlers like the purely semantic focused crawler, does not have an ontology-learning function to automatically evolve the utilized ontology. Rather it utilized the service ontologies and the service metadata formats from the transportation, service domain and the health care service domain. With the combination of technologies like semantic focused crawling and ontology learning, its possible to solve internet issues i.e whereby semantic focused crawling technology solves the issues of heterogeneity, ubiquity and ambiguity of mining service information, and ontology learning technology maintains the high performance of crawling in the uncontrolled Web environment. A Self Organizing Map has been used to generate product recommendations which is so personalized, when a user is in need of retrieving online services [2] from an enormous amount of documents and hyperlinked documents. The information from the webpages will be shared based on the user′s browsing interest.The surfing through these documents considers the user query and retrieves documents that are related to the keywords in queries.

The authenticated search engine includes three operations such as searching, indexing and downloading. Thus to avoid problem of downloading lot of web pages several crawlers are designed that improves the efficiency of crawling specific documents.

## 2. BROWSING PATTERNS IN ONLINE
A toolset exploiting web usage mining techniques to identify customer internet browsing patterns. In online sales, these patterns will be used to underpin a personalized recommended product. A Kohnen neural network, the Self Organizing Map has been trained for use both offline, to discover the profiles of the user group and in real time to examine active user click stream data.

## 3. ISSUES IN ONLINE SALES
Too much of choice in recommendations, hence time consuming in finding the right product. In web personalization, the user′s navigation behavior will be watched over and then it make product recommendations on return visit of a user to the site. But this is not advisable, since the user may look for different type of product on his return. In this paper work, web personalization means, the users current navigation behavior will be monitored and the product search will be enabled based on their interest[1]. Even though Data Mining has been successful in identifying the Customer navigation behavior, it is also computer intensive and cannot provide real time responses with the current times processing capability.There are two modules in the architecture of the existing system, which includes:

- Offline Usage Pattern Discovery Module
- Online Real-Time Recommendation Module

### 3.1 Offline Usage Pattern Discovery Module
It uses data mining techniques to identify a customer′s browsing pattern changes in a multiple anonymous website visits. From the web log, it is possible to find the trail of the user′s requirement. By applying data mining technique it is possible to automate the discovery of the knowledge. From these logs, which in turn enables the predictive modelling of the current customer′s navigation behavior.

#### *3.1.1 Information Filtering*
There are two information filtering techniques including:

- Content filtering
- Collaborative filtering

Content filtering recommends items by analyzing the contents of a products supporting information. Collaborative or social based filtering recommends the items based on the review of others user, usually by user′s ratings [3, 4].

## 4. PHASES IN USAGE MINING PROCESS
- Data Preprocessing
- Pattern Discovery
- Pattern Analysis

In Web Topology and content data, a valuable domain knowledge that helps in constructing set of Web Pages, which have some similarities in contents. These sets are called Belief Sets. The patterns identified in Data Mining will be considered as more interesting patterns to the belief set, since it is unknown to the user′s expectation list.

## 5. FAB

From the way the users access web pages, the multiple agents learns the user′s preferences. To measure the performance, a new parameter called NDPM (Normalised Distance Based Performance) measure has been adopted [5, 6]. NDPM is the distance between the user′s ranking of a set of documents and the system′s ranking of the same documents. This lies between 0 and 1. The main disadvantage of its performance is the users needs to select their desired area of interest on registration. Then the information from registration will be used as the initial user profile to which items in the databases are matching and recommendations will be given.

Content filtering has been applied in online book stores. The text categorization approach has been applied over by information extraction and a machine learning algorithm. In this approach, it recommends previously unrated items to users with unique interests and provides explanation for its recommendations. Frias-Martinez, et.al[13] presented an approach through E-libraries. In this approach, the content filtering system can automatically learn user preferences and goals create an adaptive interface which delivers a tailored service to the users. In this system the supervised and unsupervised data mining techniques have been employed.

### 5.1 Demerits

Cannot understand the temporal nature of a customer′s behaviour, hence failed to track their concept drift.

To improve the performance of the collaborative filtering, clustering techniques are used. In a high-dimensional format, it is difficult to define the distance between user sessions and also applying distance-based clustering technique is difficult. Association rule hypergraph partitioning based on frequent item set generates the Association rule mining & (PACT) Profile aggregations is based on clustering transaction. To group and summarize user transactions these AHRP is used.

## 6. PRIVACY FOR A COLLABORATIVE FILTERING

In the protocol, it has adopted encryption of data to protect people′s privacy. To address the problem of poor relationship among user′s data, latent semantic indexing has set up to use. For reducing the dimension in collaboration filtering systems the latent indexing is incorporated with singular value decomposition, which acts as a matrix factorization algorithm. To overcome the problems of two little information or too much unrelated information a system called Vzpro is used in collaborative filtering technique. In this, to address the problem of sparse binary data, a recommendation tool is used with association mining algorithm. There are 2-Phases in the system include: (i) Preprocessing and analyzing of customer historical data through association mining algorithm, which generates the rule sets. (ii) To rank the online recommendations, a Scoring algorithm is used. In the performance measure, Vzpro has outperform the other dependency network and item-based algorithms. This has not recommended based on the current behavior of the user. For the study of web-searching, a three step-methodology is involved. This includes

- Data collection
- Preparation
- Analysis of web server access logs

To support the analysis methodology, a web based application is used which records client side user′s interaction. This improves scalability by non-intrusive deal on transaction logs. The limitations are there will be an incomplete log data, due to cache of server data on the client machine or proxy server.

## 7. COMBINING CONTENT & COLLECTIVE FILTERING

Yoda[12] supports large scale web based applications. In collaborative filtering techniques it improves the traditional nearest- neighbor algorithm. To extend the locality sensitive hashing technique, a novel filtering mechanism incorporated with a novel distance measure is used.For each class of user, a predefined recommendation list called a cluster – wish list is generated in the offline process by these aggregative functions.

### 7.1 Demerits of Collaborative Filtering

Scalability has not met the requirement. Curse of scalability is due to combining offline pattern discovery with online pattern matching. The engine that has been used in the implementation is a neural network, which unsupervised learning model. The unsupervised learning model is used to enable the predictive modeling of the customer′s navigation behavior.

## 8. SELF- ORGANIZING MAP

When the input data is of high dimensionality and complexity, the self-organizing map is highly useful [7, 8]. Representing the high dimensional data, in the low dimension will not let the data to lose its essence in organizing the data, the similarity entities will be geometrically close to each other [10, 11].

## 9. PERSONALIZED RECOMMENDATION SYSTEM

Recommendations must be based on the user′s current behavior, thus avoiding irrelevant recommendations based on previous visits to the site. The recommended objects must include dynamic links, promotional advertisements or services based on user′s preference. The recommendation engine collects the active user′s visit trail characteristics behavior and compares it to known patterns of previously classified user group behavior.

## 10. PATTERN USAGE DISCOVERY MODEL

The components of the offline module include:

- Data pre-processing
- Data cleaning & selection
- Data Integration
- Discovering patterns
- Data Post-processing & recommendation
- Data Pre-Processing

In raw web log files, it may contain a considerable amount of incomplete and irrelevant information. So in data pre-processing, the raw data sets will be cleaned and transformed into a form suitable for mining in web. The source of these data sets are server access logs. In these logs, the each and every query given to the web server will be recorded. The recorded details include:

- IP address
- User ID & Password
- Requisition of date and time
- Request type, query strings and the Protocol
- Cookies
- Data Cleaning Selection

The log files used to contain a large amount of erroneous, misleading and incomplete attributes including error requests, request resets, web agent's requests, proxies requests.

**a) Data Integration**
By integrating the data, the log entries and the semantically related activities will be mapped together. Data Integration is a difficult task to deploy in search-driven websites, where a client communicates with the server by encrypted query strings. To the web usage log, the significant page views will be kept hidden. The products which are related to search are the significant one, which will be kept hidden and is of no use.

**b) Tasks of data Transformation**
By series of clicks by user, the navigation behavior can be depicted. The click stream can be separated as sessions, with meaningful metadata.

**c) Pattern Discovery**
In the phase of pattern discovery, the patterns based on user behavior will be determined. These patterns used to be available in pre-processed web logs. The techniques like clustering, association rules, navigational pattern mining can be applied over the web logs to group the similar user behavior. In neural network, it has the advantages over K-means clustering and K-nearest neighbour methods. In the Kohonen neural network, it has the capacity to provide the relationship between the source data and the discovered clusters. It can also cluster data sets into user profile groups. The neural network is trained using web logs, thus it provides information about the visitor's browsing activity in retrieving the web pages. This trained network is used in offline, to examine the active user data and provides matching to a specific user group. The Self Organizing Map used in modeling the customer's navigational behavior. This unsupervised model clusters queries that are related to user sessions from a web log. SOM neural network also needs input vectors.

## 11. NEURON UNITS

Neuron units holds the input vectors that are used by the SOM modeling neural network. The user session is given by a set S={q1,q2,...qn,t1,t2,...tn-1}, where q represents the queries and t represents the time interval between the queries. From the two dimensional SOM, the units in SOM competes with all the other records to win. When a unit wins a record, then the weight of the unit will be adjusted so that it matches with the predicted value patterns for that record.

WeightN(After output updation)=WeightM(Before output updation) + K[Xn- WeightM]

WeightN(After output updation)= K*Xn + (1-K) WeightM

where K indicates the consistent change in weight. This range lies between 0 and 1.

## 12. MOSTFREQUENT RECOMMENDATION (MFR)

In MFR method, it looks for the common interest between a groups of users. It scans through the product retrieval data information and hence calculates the frequency count. Based on the frequency count, it will return the N-most frequent products as the product recommendation list. This frequency count used to be a greater value than the pre-specified threshold. The only constraint lies in the method is that it should not contain the products, which are actually browsed by the user in current.

### 12.1 Study on System Evaluation

In system evaluation two models namely Hypothesis Model and Comparison Model has been used. The system evaluation with Hypothesis Model is as follows.

### 12.2 Hypothesis Model

Hypothesis specifies the distribution of data completely. In this the sample size of the data will be considered as a function of the sampling function.Hypothesis Model includes three metrics in performance measure

$$Recall = \frac{Size\ of\ correctly\ Identified\ Set}{Total\ Data\ Set}$$

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

the usage of F1 score, is to avoid the drawbacks that may arise while using Recall and precision alone in calculation. The drawback is when N is increasing, Recall will increase, whereas Precision will decrease. Thus F1 score helps in alleviating the problem through Harmonic Average.

### 12.3 Mean Absolute Error

The Performance Evaluation measure is calculated as

$$MAE = Actual\ Numeric\ Output - Predicted\ Numeric\ Output$$

It also maintains the error dimensionality, such that they will not deviate from large difference in values.

### 12.4 Correlation Coefficient

It measures the statistical correlation between the actual and predicted numeric output and explores the consistent Correlation between the actual and predicted testing results.

$$C_i = \frac{Cov(T,P)}{\sigma_t \sigma_p}$$

### 12.5 Statistical Accuracy Metrics

**Table 1: Performance Metrics**

| S.No | Metrics | K-Means Model | SOM Model |
|---|---|---|---|
| 1 | MAE | 3.316 | 3.15 |
| 2 | Correlation Coefficient | 0.218 | 0.485 |

## 13. CRAWLERS

A crawler is a software program used to create search engine index entries by visiting Web sites. It automatically reads the web pages and retrieves information from those pages for web indexing. Web crawling application software is used in web search engines updates the web content of the web sites. Web search engine indexes the pages that has been downloaded pages by crawler for later usage for the user, making the search

quicker. Web Crawler is incessant running programs that download pages at regular intervals from internet [15]. For assembling the Web content in local, crawlers are used as a tool. Web crawlers are used in applications where large number of pages is quickly fetched into a local repository and is indexed based on keywords in user query . Since crawlers extract information from web sites, they are used in Web Scrapping.

## 14. SELF-ADAPTIVESEMANTIC FOCUSED CRAWLER

Self-Adaptive semantic focused crawler [SASF] framework [14], used to discover format and index mining service information by considering for the three issues by using ontology learning for maintain the performance of the crawler. The new concept involved in the SASF crawler are Vocabulary based ontology learning and hybrid algorithm for matching semantically relevant concepts and metadata. Discovering the service or service information in particular environment is done automatically or semi-automatically. SASF framework uses Semantic Focused Crawling to solve the above the problem of heterogeneity, ubiquity, ambiguity in service discovery. Heterogeneity refers to the problem of classification of service advertisement. Ambiguity refers to the problem of identifying the service information that doesn't have a consistent format and standard. Ubiquity refers to the problem of discovering the registries of services that are geographically distributed. SASF crawler is used for helping search engines to precise the mining of service information. Supervised ontology learning method is used to maintain the harvest rate of the crawler. The input given to the supervised learning method is the domain and the topic represented by a concept. It may work with in an uncontrolled web environment.

Unsupervised ontology learning method where the input is topic and relevance score of the topic. Metadata crawler contains information such as content, length and length variation, value based analyses, frequencies, patterns, domains, dependencies, relationships. Determines the semantic relatedness between concepts and metadata concept metadata using semantic similarity algorithm. In this algorithm the semantic similarity between concept description and service description is measured. It follows a hybrid pattern by aggregating two algorithms namely semantic based string matching algorithm and statistics based string matching algorithm.

**Table 2: Comparison of Various Crawler's Harvest Rate**

| S.No | Methods | Harvest Rate |
|---|---|---|
| 1 | Breadth First Crawler | 0.546 |
| 2 | SSRM Crawler | 0.604 |
| 3 | VSM Crawler | 0.661 |
| 4 | CMCFC | 0.763 |
| 5 | SASF Crawler | 0.6 |

## 15. CONCLUSION

From the performance measure it has been shown that the personality product recommendation system using SOM, has produced useful recommendation. In the test cases it was also proved to be equal to or greater than 50% in the rate of precision or recommendation quality. It has been also found that a supervised ontology learning crawler enhances the harvest rate of crawling without considering the classification. It may not even work in an uncontrolled web environment when new unpredicted term appears. Hence using Ontology learning based focused crawlers helps in improving the performance.

## 16. REFERENCES

[1] M.Balabanovic (1997), "An Adaptive Web Page Recommendation Service", Proceedings of the 1st International Conference on Autonomous Agents, pp:378-385.

[2] C. Basu,H.Hirsh,V.Cohen (1998)," Recommendation as classification: using social and content-based information in recommendation", Proceedings of the 15th National Conference on Artificial Intelligence,pp:714-720.

[3] E.Frias-Martinez, G. Magoulas,S.Chen, R. Macredie (2006), "Automated User Learning for Text Categorization", Proceedings of the International Journal of Information Management, pp: 19-25.

[4] Hai Dong, Member, IEEE, and Farookh Khadeer Hussain (2014)," Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery", IEEE Transactions on Industrial Informatics, vol. 10, no.2.

[5] Jaytrilok Choudhary, Devshri Roy,( 2013),"Priority based Semantic Web Crawler ", International Journal of Computer Applications ,Vol. 81, No 15.

[6] T.Kohonen (1981), "Construction of similarity diagrams for phonemes by a Self Organizing Algorithm", Technical Report TKK- FA463,Helsiniki University of Technology,Espoo,Finland.

[7] T.Kohonen (1982), " Self Organized Formation of Topologically Correct Feature Maps", Biological Cybernetics. Pp: 59-69.

[8] H.Lieberman, N.W.Van Dyke, A.S.Vivacqua, "Let's Browse: A Collaborative Browsing Agent",pp: 378-385.

[9] M.D.Mulvenna, S.Anand, A.G. Buchner (2000), "Personalization on the net using Web Mining", Communications of the ACM. pp:123-125.

[10] P.Resnick, N.Iacovou, M.Sushak, P.Bergstrom, J.Reidl (1994), "Grouplens: An Open Architecture for Collaborative Filtering of Netnews", Proceedings of ACM Conference on Computer Supported Cooperative Work. pp:175-186.

[11] C.Shahabi, F.Banaei-Kashani, Y.S.Chen, D.McLeod (2001), "Yoda: An Accurate and Scalable Web Based Recommendation System", Proceedings of the 9th International Conference on Cooperative Information Systems, pp:418-432.

[12] Y.Shih, R.Liu (2005), "Hybrid Recommendation Approaches: Collaborative Filtering via Valuable Content Information", Proceedings of the 38th Hawaii International Conference on System Sciences, pp: 217b.

[13] UCL,http://www.ucl.ac.uk/ontology/Microcore/HTML_resource/SOM_Intro.htm.

[14] Willamette, http://www.willamette.edu/~gorr/classes/cs449/Unsupervised/SOM .html.

[15] Xuejun Zhang, John Edwards, Jenny Harding (2007), " Personalised Online Sales using Web Usage Data Mining", Computers in Industry. pp:772-782.