CHAPTER 3

USAGE CLASSIFICATION FOR PERSONALIZATION

3.1 INTRODUCTION

We list out our thoughts on how to classify users based on the topic of interest. A particular users’ usage data is a measure of his/her access behaviour on a particular site. This classified data could effectively be used for creating a personalized web page. The authors observed the fact that amid the many methods in statistical analysis, web mining and pattern matching, very few attempt at bringing the latent correlation between web pages and its users. The proposed work is an attempt to bring out the user behaviour based on classified usage data. The authors propose a probabilistic approach based on Bayssian classification for this. The authors focus on identification of users’ topic of interest and recommends content to a user based on past behaviour without major restructuring of the site. The authors suggest the novel idea of automatic personalized website without much modification of existing structure of the website.

The organization of this chapter is as follows: In section 3.2, the motivation behind this idea is briefed. In section 3.3, an outline of the methods/models used for classification is given. Related study leading to proposed method is explained in subsections. Section 3.4 gives an overview of the proposed methodology for usage classification and enhanced personalization. Section 3.5 is a comparison of our study with related methods. Conclusion and our further study is presented as Section 3.6.
3.2 MOTIVATION

Users of web often find it difficult to reach the requested page faster and effectively due to two main reasons. First due to the tremendous growth of pages related to a particular search and secondly the data is highly unstructured and unorganized in server. For enhancing user personalization and thus faster navigation the mined usage server logs alone are not sufficient. The idea of classifying access log data and studying patterns based on this may serve the purpose. We can follow a probabilistic approach on the behaviour of user in fetching data. This time bound classified usage log is a source to classify his pattern of access. This could be effectively used for personalization.

Hofmann (1999) tried to find hidden relationship between access behaviours of users and to draw relationship between co occurrences by his proposed Probabilistic latent semantic analysis (PLSA) model. Guandong Xu et al.(2005a) proposed a method for better web personalization system from PLSA model. Their model, based on a probabilistic analysis of latent factor, usage and linkage data derives a better presentation of user access patterns. Studying a specific users access behaviour and thus to personalize his next visit on the website without much reconstructing the site is a challenge.

Guandong Xu (2008) in his research proposed an algorithm named User Profiling Algorithm, which incorporates a filtering approach in mined usage patterns to create better user profiles. Session similarity measures are clustered and user profiles are extracted from sequences based on page weights. We propose a better method incorporating session analysis, probability based classification based on total number of hits on a page for better personalization. We put a threshold for number of hits on a page and tried to relate the lower and upper bounds from this threshold value. Our
model present better personalization in comparison with the above mentioned models.

Mostly studies focus on web personalization rather than user personalization. Thi Thanh Sang Nguyen et al. (2014) proposed a method based on domain ontology to represent domain knowledge of a website. Then based on the domain knowledge and usage knowledge they tried to do the web personalization. Their studies mainly focus on finding the access behaviour of various users of a website. Still the user personalization is of high demand than general web personalization based on access behaviour as the users are in need of faster and accurate navigation results. Specifically in some systems such as virtual learning environments users look for a personalized structure for their ease and continuity in studies. This work focuses on such systems and to automatic personalization of pages without much re structuring.

3.3 PROBABILISTIC CLASSIFICATION METHODS / MODELS

To select the categories mostly visited by user, whether to follow a deterministic or probabilistic approach is a matter of concern. Most of the text mining and machine learning applications follow deterministic as well as rely on Support Vector Machines which are non probability classifiers. Since we deal with the numeric weight values and base on threshold values we followed a probabilistic approach. Many probabilistic algorithms are in use for classification. The models for probabilistic classification are listed below. Our approach is based on Bayes classification.
3.3.1 Bayes Classifier

Bayesian Classifier is a statistical classifier. It is based on Bayes Theorem. It has high speed, accuracy and has a comparable performance. Let A be a sample of data whose class label is unknown. Let H be a hypothesis that A belongs to class C.

The classification is to determine \( P(H/A) \) ie, to prove A comes under or satisfies the hypothesis H. \( P(H) \) is the prior probability. \( P(A) \) is the probability that sample data is observed.

Bayes theorem states that

\[
P(H/A) = \frac{P(A/H) P(H)}{P(A)}
\]

ie, posterior = [likelihood*prior]/evidence.

3.3.2 The Probabilistic Analytical Model - PLSA

The Probabilistic Semantic Analysis Model (PLSA) is a statistical model which makes efficient classifications based on conditional probability. Hoffmann (1999) initially applied this probabilistic model for text mining applications. PLSA algorithm checks for the maximum likelihood of co-occurrences based on the usage patterns of samples. It maps the association between usage sessions and weight of web pages.

The PLSA algorithm puts forward the idea of a latent aspect factor \( z_r \), which considers the occurrence of an item in a web page from the usage data.

We can represent \( Z = \{z_1, z_2, z_3 \ldots z_r\} \) where \( Z \) is the aspect factor and \( z_r \in Z \).
z1, z2, z3… can be different topics of interest of users from the mined data of related categories.

According to the PLSA algorithm

The probability of finding a word in session

\[
P(w_n, s_m) = P(s_m) \cdot P(w_n | s_m)
\]  

(3.1)

Where

\[
P(w_n | s_m) = \sum_{z=1}^{Z_r} P(w_n | z)P(z | s_m)
\]  

(3.2)

and

\[
P(w_n | z) = \text{the probability of occurrence of word } w_n \text{ in session.}
\]

\[
P(w_n | z) = \text{probability of picking word from latent space.}
\]

\[
P(z | s_m) = \text{session specific probability in the associated latent space.}
\]

Applying Bayes Rule in Eqn 3.2, probability calculation spanning over latent space for user sessions is re written as

\[
P(z_r | s_m) = \frac{P(s_m | z_r)P(z_r)}{\sum_{z=1}^{Z_r} P(s_m | z_r)P(z_r)}
\]  

(3.3)

3.3.3 Related Model - LDA for Finding Usage Pattern Based on PLSA

Guandong Xu (2008) in his research study proposed a model namely Latent Dirichlet Allocation (LDA) model, which effectively does web recommendations from usage logs. LDA is the process where a group of words are taken together for possibility of occurrence and the single needed
word is selected from the group. In his proposed user profiling algorithm, the access patterns were effectively calculated. He proposed a threshold based filtering approach for finding the most suitable pages for page weight calculation. Based on Bayes probabilistic approach the work studied

- the posterior probability of pages in a session
- selected the probability values of occurrence of pages above threshold
- recommended pages for personalization based on score.

The score for recommending a specific page for personalization is calculated using equation

$$ps(p_i) = \sqrt{\sum_{z \in P_i} P(z_{r}, m). wt(p_i, z)}$$

(3.4)

where \( p_i \) is the \( i \)th page and contained in set of pages \( P \). Also latent factor \( z_r \) is contained in the filtered task sequence above threshold value.

3.4 PROPOSED CLASSIFICATION BASED ON USAGE LOGS

Each user will have his/her strategy of visit on web sites. The challenge is to find this pattern and classify user based on his style of browsing. Usage logs can contribute to this. The process of classifying the usage of the site and finally user personalization could be done by 1) Finding the associated weight on a web page from sessions for a particular period of time 2) calculate the top visited categories by user and 3) personalization based on his/her visit interest.
3.4.1 Discovering Relationship Between Sessions and Associated Weight on Web Page

The key factor that contributes to personalization is identifying the usage pattern of particular user. Sessions, which are a list of pages clicked by the user, play an important role in providing this information. A specific user with user id when starts browsing, the visited pages by him/her could be listed from the server logs which actually is generated from the recorded session data. Users' behavioural/access pattern could be easily figured from such log data. This visits or the number of clicks on a particular page gives weight to the specific page. Thus there is a direct relationship between the usage data, which is a reflection from session and server logs and the weight associated with each page. Our study is focused in this direction. The user preference of a particular page could be identified from its weight value and this could effectively used for enhancing personalization. Session of a particular site visit by user can be represented in terms of pages as

\[ S_e = \sum_{r=1}^{n} p_1 p_2 p_3 \ldots p_r \] (3.5)

Where a specific period of time is selected for \( n \), say 10 days or so on.

The category of the \( r^{th} \) page visited by user is represented as \( p_r \), where \( r = (1,2,3 \ldots k) \).

Once the session is identified the total weight on a particular page is calculated as the total number of hits on that page. When a page is browsed by user for the first time we can not say it is his preference. We set a threshold value for the number of visits. Lower and upper bounds of these threshold and distance from lower to upper bound is considered for finding the desired page accurately. We are not considering the time spent on a particular page for
personalizing as the time duration sometimes can be idle time after login. We calculate weight from the number of hits on a particular page and time consideration is only the duration of session.

Thus sessions can be written in terms of associated weight as

\[ S_e = \{w_1, w_2, w_3, \ldots, w_r\} \]

where \( w_r \) is the weight of \( r^{th} \) page.

A filtering mechanism could be employed for pre processing of data.

### 3.4.2 Proposed Probability Based Method to Calculate Weight and to Find User Priorities

Once the session is extracted it is segmented first to get the required fields only. To select the categories mostly visited by user, whether to follow a deterministic or probabilistic approach is a matter of concern. Most of the text mining and machine learning applications follow deterministic as well as rely on Support Vector Machines which are non probability classifiers. Since we deal with the numeric weight values and based on threshold upper bound and lower bound values a probabilistic approach was adopted.

In the proposed Easy User Navigation (EUN) algorithm we propose a methodology based on Bayes approach for classification of user preferences. This is a weight centric approach. When writing session as associated weights, for weight calculation in link and out link factors are considered. Number of mouse clicks (N) on a page is taken for weight calculation.

The weight \( W_k \) on page \( k \) can be expressed as
\[ W_k = \left( |N|_{in} \right) + \frac{|N|_{in}}{\sum_{out(pk)} |N|_{out}} \]  \hspace{1cm} (3.6)

Where

- \(|N|_{in}\) – total number of hits on in links of the page for the particular session
- \(\sum_{out(pk)} |N|_{out}\) - total number of hits considering all the out links from the page.

Applying Bayes rule in our method, the probability of finding a page of weight \(W_k\) in session \(s_{em}\) can be written as

\[
P(W_k | s_{em}) = \frac{P(s_{em} | W_k).P(W_k)}{\sum_{p=1}^{out(pk)} P(s_{em} | W_k).P(W_k)} \cdot UNC \hspace{1cm} (3.7)
\]

Where \(UNC\) - the Easy User Navigation constant, which is the user preference for page \(k\). The constant \(UNC\)'s value lies between 1 and 5. That is \(1 < UNC < 5\) based on the continuous interest shown by user in recent times. For example, for a session of 10 days, if there is continuous navigation of the particular page for the last 5 days \(UNC\) is assigned a value above 2.5. This is under the assumption that the user is highly interested in this topic of search.

If the values of \(P(W_k|s_{em})\) are high for many pages, the pages falling above the centroid could be filtered and suggested for personalization. Centeroid is calculated as average of maximum distance and minimum
distance from the pre set threshold. Measured distance between minimum value and maximum value above threshold is recorded for this purpose. The probability values are calculated from the usage log data for particular session.

An example of observed weights of our virtual learning environment VLSchools and the calculated associated probabilities is given in Table 3.1. Data taken for a particular user and for a particular session with 60 pages. Pages are numbered sequentially from left to right in level wise manner.

Table 3.1 Observed Weights and associated probability

| Page # | Weight ($W_k$) | $P(W_k|S_m)$ |
|--------|----------------|--------------|
| 3      | 75             | 0.713        |
| 9      | 36             | 0.052        |
| 12     | 28             | 0.011        |
| 23     | 40             | 0.326        |
| 27     | 110            | 0.873        |
| 32     | 15             | 0.008        |
| 39     | 67             | 0.47         |
| 41     | 94             | 0.621        |
| 53     | 5              | 0.002        |
| 60     | 11             | 0.009        |

Out of the ten page values generated, five can be selected for finding user preference. Threshold could be set as 0.3 and a first filtering is done. Setting of centroid and the probabilities related to desired page is done automatically in the system.
3.4.3 Recommendation Approach for Personalization

Many factors can be associated with navigational behaviour to draw relationship between access patterns. For example, in our virtual learning environment VL Systems, the user may be interested in different type of courses at different times. Suppose z1 represents users interest in a specific tutorial category named PHP, z2 can be represented as his/her interest in JSP course. From the usage space Z, whether to select z1 or z2 is the matter of concern. We propose a novel approach based on the usage log for automatic personalization. The user on his login will be getting a personalized page based on the factors of weight, user preference and UNC. If needed further filtering is also done. Algorithm for recommending personalization based on probability approach:

Algorithm

**Input** : User profile, time limit for session, Number of pages and list of pages

**Output** : Probability of selecting the particular page from list for input user profile.

Step 1 : Find the total category count for a particular category _id from the user category state data store

Step 2 : find the count of each subcategory from the user sub category data store.

Step 3 : Joined the user category _id from user category state stored and category id from web category stored.

Step 4 : list number of continuous visit of category for session for calculation of navgtn const. Assign values for constant based on list.
Step 5: Probability of selecting a category in session s will be

\[
\text{(navgtn const } \times \text{ count }) / \text{ total}.
\]

Step 6: The probability and subcategory id is stored in webcategory data store.

The calculated probability values enhance recommendation for personalization. Data analysis done for a sample educational website. It is observed that the category that is visited most will dynamically assigned as the first link to facilitate navigation. The category that is visited least will be the last link. When the user login not only the most visited category but also the subcategory and contents are loaded.

3.5 COMPARISON WITH RELATED WORK

In this section we draw a comparison of our method with two popular models. Our study is compared with the Probabilistic Semantic Analysis Model (PLSA) proposed by Hofmann (1999) and Latent Dirichlet Allocation (LDA) model, which effectively does web recommendations from usage logs proposed by Guandong Xu (2008). We also list the features of our system compared to an ontology based web personalization system proposed by Thi Thanh Sang Nguyen et al. (2014).

3.5.1 Comparison of Proposed Method for Classification with PLSA Model

The Probabilistic Semantic Analysis Model (PLSA) proposed by Hofmann (1999) is a statistical model which makes efficient classifications based on conditional probability. It mainly considers the latent factor for finding probability. In this model probability of occurrence of a particular
word in different aspects is mainly considered. It then clusters the aspects of similar usage and recommends for web personalization.

The proposed model is similar to PLSA in the ways that

1) Is a statistical method based Baye’s rule for finding probability of a page in a session.

2) To reflect access preference of user, both models partition access space

3) Both the models follow a threshold based approach for selection of pages.

Our model differs from PLSA in the following ways

1) Ours is a numeric model which finds probability of occurrence of a page based on the number of clicks/weight on a page.

2) Ours is a user personalization model which produces enhanced classification by considering recent activities of user.

3) In the proposed method session duration and proposed constant UNC is taken in to account which avoids older usage data contributing to probability of setting user priority.

3.5.2 Comparison of Proposed Method with LDA Model

Latent Dirichlet Allocation (LDA) model presented by Guandong Xu (2008) effectively does web recommendations from usage logs. He proposed a threshold based filtering approach for finding the most suitable pages for page weight calculation. Both LDA model and our proposed model follows probabilistic approach for page recommendation. Both methods
follow a filtering approach. But the proposed method follows a filtering approach focusing on a centroid from threshold. The work concentrates on in link and out link factors rather than the latent factor space of LDA.

The LDA method recommends pages for personalization based on a page score value. Our model effectively calculates/classifies pages based on the probability of occurrence in a session by considering constant UNC, the easy user navigation constant. Weight of a page is calculated from usage logs by considering the number of mouse clicks and the user’s behavioural pattern which may change over time. UNC is calculated based on the recent repeated visits of user which promotes better setting of user priorities. This also avoids very older pages visited being considered for probability of selection. Based on all these factors it is observed that our method does automatic user personalization more effectively in comparison with both PLSA and LDA methods.

3.6 CONCLUSION

Our study on usage classification for personalization is described in this chapter. We followed a probabilistic approach for setting priority for a page from the classified data. Our study also focuses on recent interest factor of user for setting priority. Thi Thanh Sang Nguyen et al.(2014) method represented domain knowledge of a website based on ontology. Then based on the domain knowledge and usage knowledge they tried to do the web personalization.

We focused our study on user personalization rather than web personalization for faster navigation. We focused on session wise segmented usage data for classification and thus personalization. We considered the inlink and outlink factors of a page for calculating weight. With all these
considered measures effective user personalization is achieved. As a continuation we have also done classification of different users based on usage data, clustering of this classified data to find similarity in navigation, and prediction of users future visit from this access pattern. The study on this is described in chapter 5.

We have done experimental evaluation on a virtual learning webpage. The main challenges were changing user behaviours over sessions and the unstructured data. In future we wish to elaborate our study to other websites with high web clutter.