ADDIS ABABA UNIVERSITY

School of Graduate Studies

DEPARTMENT OF INFORMATION SCIENCE

E-learning material Recommender System Using Learner Interest Modeling

Tamirat Sisay

November, 2015

Declaration

I declare that this thesis is my original work, has not been presented for a degree in any other university and all sources of materials used for the thesis has been well acknowledged.

_____

Tamirat Sisay

This thesis has been submitted for examination with my approval as university advisor.

Dr. Solomon Tefera
Advisor

November, 2015

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Title

E-learning material Recommender System Using Learner Interest Modeling

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By:
Tamirat Sisay

November, 2015

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

Title of the thesis

E-learning material Recommender System Using Learner Interest Modeling

By:
Tamirat Sisay

Members of the Examining Board

Name	Title	Signature	Date
_____	Examiner	_____	_____
_____	Examiner	_____	_____
_____	Advisor,	_____	

ACKNOWLEDGEMENTS

First and for most I would like to give a special gratitude to the Glory of God who provided me the courage to finish my program of study.

I am deeply indebted to my advisor Dr. Solomon Teferra for his dedication, encouragement, inspiring guidance, and valuable comments throughout my thesis work.

My greatest gratitude is extended to all my family members for their encouragement and supports throughout my study. Finally, I would also like to thank sincerely all my friends who helped me with their valuable support during the entire process of this thesis.

TABLE OF CONTENTS

Contents

LIST OF TABLES

LIST OF FIGURES

Abstract

Today recommender systems are widely used not only in e-commerce but in e-learning as well. They are actually used in the latter environment to suggest resources and learning materials to learners and, thus, contribute in improving the quality of both teaching and learning processes. As a result, predicting the needs of a learner and recommend e-learning resources in e-learning system has gained attention. The requirement for predicting user needs in order to recommend the user of e-learning system and improve the usability of the system can be addressed by recommending pages(resources) to the learner that are related to the interest of the user at that time.

The aim of this research is to assess how effective the uses of the visiting time and visiting frequencies of pages in web based e-learning system to learner interest modeling for recommending e-learning resource.

The main data source used is log file of Moodle e-learning Management System from Mekele University of which data set size of 267 sessions. And the approach we used that learner sessions are clustered according to the similar amount of time that is spent on common e-learning resources among sessions. Accordingly, if there is a similarity between the new learner session page time and the existing clustered sessions, the system uses two Methods to assigns the solution (recommended e-learning page links).

The performance of the approach is measured using developed prototype system by the standard measure of relevance (IR system) precision for the two methods, where the system registers 46.3%, 51.4% precision for popularity information (Method 1) and popularity information and the Poisson parameter (Method 2) respectively. Finally, conclusion and future research directions are forwarded.

1 Chapter One

1.1 Study background

With the rapid increase of Information Communication Technology (ICT) infrastructures, every educational institution has the opportunity to make use of the Internet as a communication medium with the students. For an effective and efficient access to learning materials, the concepts and methodologies of technology-based learning are increasing in importance with web based e-learning [6] becoming a crucial resource for institutions.

The advantages of web based e-learning as opposed to traditional learning are instantly evident with e-learning making education independent of time and location. More importantly, it opens up fresh possibilities for implementing pedagogical innovations in an environment where students are expected to function as active, independent, self-reflected and collaborative participants [38].

Web based e-learning environments are becoming increasingly popular educational establishments. The rapid growth of e-learning has changed traditional learning behavior and presented a new situation to both educators (lecturers) and learners (students). Educators are finding it harder to guide Students to select suitable learning materials due to more and more learning materials online. Learners are finding it difficult to make a decision about which of learning materials best meet his / her situation and need to read. Therefore, on the educator's side, educators need an automatic way to get feedback from learners in order to better guide their learning process. On the Learner’s side, it would be very useful an e-learning system could automatically guide the learner's activities and recommend learning materials that would improve the learning [11]. Locating suitable learning material is a big challenge. One of the possible ways to overcome this problem is e-learning resource recommender system.

1.2 Problem Statement

Along the lines of Education and Training Policy of April 1994made by Ethiopia Federal Ministry of education, rolling Education Sector Development Program (ESDP) was launched in 1997/98 to meet the Education for All (EFA) and Millennium Development Goals (MDGs) by 2015 (Ministry of Education, 2008) in Ethiopia[7]. Various phases has been under going through this program with the interval of five years starting from 1997/98; Specifically, ESDP III and ESDP IV (the Ethiopian National action plans on education), emphasize the integration of ICT infrastructures to support the country's education system with ICT. In view of this, ICT infrastructures are provided to schools to receive satellite education transmission (plasma instruction) to enhance the quality of education at secondary level since September 2004. The ICT for education policy which extends from these education action plans also recognizes ICT as an enabler for widening access to education for the Ethiopian population, for supporting literacy education, and for facilitating delivery and training at all levels.

Nowadays more and more people have benefited from the various e-learning programs. However, the high diversity of the learners, having different knowledge and learning interest poses new challenges to the traditional "one-size-fit-all" learning model, in which a single set of learning resource is provided to all learners. In fact, the learners could have various interests; and hence they cannot be treated in a uniform way [70].

Content placed within an e-learning platform, easily accessible from any place at any time, seems to fulfill individual needs of learners. However, easy access to learning content does not ensure better teaching and learning results. Recommend learning materials based on analysis of interest of learners satisfies the learner's need. And there is no research conducted on designing learning system based on the above assumption.

It is of great importance to provide learning material recommender system which can automatically recommend learning resources based on learners' interest to provide best learning materials to students.

Therefore, the research focused on answering the following questions.

1. How to model learner interest using the visiting duration and visiting frequencies of pages(Resources)
2. How can we use their interest to group learners?
3. How to develop e-learning recommender system?
4. How to evaluate and improve the performance of the recommender system

1.3 Objective of the study

The following general and specific objectives are formulated towards solving the research problems.

1.3.1 General objectives

The general objective of the study is to design a prototype Recommender System that can provide possible recommendation on the selection of e-learning material in e-learning management system (LMS).

1.3.2 Specific objectives

- To identify the main criteria that influences the learning interest of learners in the selection e-learning material (Recourses) for learners while using Learning management system (LMS).
- To develop a prototype of recommender system to learners on the selection of e-learning material that best matches with their interest.
- To determine the performance of the proposed recommender system using different evaluation techniques.
- To recommend further research areas for future work.

1.4 Scope of the study

Even though the study is wide-ranging and complex that needs to cover the whole Ethiopia, due to limited time it is bounded to study only Ethiopian higher institutes. Whereas effective e-learning recommender system includes resource (Learning object) model, and user (Learner) model due to shortage of time the researcher focuses on recommendation based on learner modeling. Besides availability and relying on of e-learning system in Ethiopia higher institution is limited, we are limited to log file from Mekel University Moodle learning management system(LMS) only. Again, due the same problem of getting enough log file, the user interest model in this research is based on this 267 data set.

1.5 Ethical Consideration

In the process of the study, the following ethical issues were considered. In order to obtain an informed consent from the log data provider, Mekele University, the purpose of the study was

explained clearly. Information obtained from the log file particularly user's personal data was promised to be kept confidential.

1.6 Significance of the study

From this study, primarily learners and educators, specifically learners who are in remote with no people to help are the immediate beneficiaries to enhance their day to day learning activities. The model has a great significance to design e-learning material recommender system for knowledge transfer in institute through e-learning management system plate form and it also help institution to analyze their students' interest so that they can design the content of the course to address the learners need. Besides since the study focused on one type of course, the result of the study can be applied in all types of courses given .Consequently, those remotely reside learners can use the model in recommender system in recommending e-learning resources based on their interest. Moreover, it is an academic exercise to fulfill the requirement of masters program that the researcher is enrolled in.

1.7 Organization of the Thesis

This research is organized into six chapters. Chapter one consists of background, statement of the problem and its justification, objective of the study, the scope of study followed in the course of the study and the significance. In chapter two literature reviews on e-learning material recommender system approaches were discussed. Further literature reviews on learner modeling, e-learning material recommender system components, and design process were performed. Chapter three present methodologies, the data preparation and cleaning processes we used for e-learning material recommender system. Chapter four states how our users (learner) of e-learning are modeled. Experiment and results are stated in chapter five. Finally conclusion and recommendation presented in chapter six.

2 Chapter Two

Literature Review

This literature review is divided into five main parts. Part one describes discusses e-learning, its challenges and opportunities as well as the limitation that lead to starting of e-learning recommender system. Part two also discusses recommender system in general. Part three discusses approaches how to recommend e-learning resources to different users based on the user profile. The last part deals with recommender system, the different techniques we are using on how recommender systems provide recommendation of e-learning resources that users might appreciate or be interested in.

2.1 E-Learning

There are still discussions about the definition of the term e-learning According to [23]; e-learning is defined as follows:

"*E-learning is mostly associated with activities involving computers and interactive networks simultaneously. The computer does not need to be the central element of the activity or provide learning content. However, the computer and the network must hold a significant involvement in the learning activity*."

A number of other terms are also used to describe this mode of teaching and learning. They include online learning, virtual learning, distributed learning, network, and web-based learning. Fundamentally, they all refer to educational processes that utilize information and communications technology to mediate asynchronous as well as synchronous learning and teaching activities. The term e-learning also comprises a lot more than online learning, virtual learning, distributed learning, networked or web-based learning. As the letter 'e' in e-learning stands for the word ''electronic'', e-learning would incorporate all educational activities that

are carried out by individuals or groups working online or offline, and synchronously or asynchronously via networked or standalone computers and other electronic devices [24].

2.1.1 TRENDS IN E-LEARNING

The growing interest in e-learning seems to be coming from several directions. These include organizations that have traditionally offered distance education programs either in a single, dual, or mixed mode setting. They see the incorporation of online learning in their repertoire as a logical extension of their distance education activities. The corporate sector, on the other hand, is interested in e-learning as a way of rationalizing the costs of their in-house staff training activities. E-learning is of interest to residential campus-based educational organizations as well [32] see e-learning as a way of improving access to their programs and also as a way of tapping into growing niche markets. The growth of e-learning is directly related to the increasing access to information and communications technology, as well it's decreasing cost. The capacity of information and communications technology to support multimedia resource-based learning and teaching is also relevant to the growing interest in e-learning [25].

2.1.2 OPPORTUNITIES OF E-LEARNING

For the e-learning industry to be popular, the following are some of the opportunities:

The flexibility that e-learning technology affords: Flexible access which refers to access and use of information and resources at a time, place, and pace that is suitable and convenient to individual learners rather than the teacher and/or the educational organization. The concept of distance education was founded on the principles of flexible access [26]. It aimed to allow distance learners, who were generally adult learners in full or part-time employment to be able to study at a time, place, and pace that suited their convenience [27].

Electronic access to hypermedia and multimedia based resources: with the growing of information and communication technology, the capture, and storage of information of various types including print, audio, and video is possible. Networked information and communications technologies enable access to this content in a manner that is not possible within the spatial and temporal constraints of conventional educational settings such as the classroom or the print mode [27]. In the context of this distributed setting, users have access to a wide variety of educational resources in a former that is amenable to individual approaches to learning and accessible at a time, place and pace that is convenient to them typically, these educational resources could include hyper-linked material, incorporating text, pictures, graphics, animation, multimedia elements such as videos and simulations and also links to electronic databases, search engines, and online libraries.

2.1.3 CHALLENGES OF E-LEARNING

Despite this level of interest in e-learning, it is not without constraints and limitations. The fundamental obstacle to the growth of e-learning is lack of access to the necessary technology infrastructure, for without it there can be no e-learning. Poor or insufficient technology infrastructure is just as bad, as it can lead to unsavory experiences that can cause more damage than good to teachers, students, and the learning experience. While the costs of the hardware and software are falling, often there are other costs that have often not been factored into the deployment of e-learning ventures. The most important of these include the costs of infrastructure support and its maintenance, and appropriate training of staff to enable them to make the most of the technology [27]

As e-learning matures as an industry and a research stream, the focus is shifting from developing infrastructures and delivering information online to improving learning and performance [28]. Examples of e-learning on the internet today are, too often, little more than lecture notes and some associated links posted in HTML format. However, the true power of

e-learning comes from the exploitation of the wide range of capabilities that technologies afford. One of the most obvious is to provide assessment and instructional content that adapt to learners interest.

There are many e-learning systems, but they provide only the same material to all students regardless of individual interest. The material is still oriented for on-campus homogenous, will prepared and motivated students. However, the students are heterogeneous having very different interest, goals, backgrounds, and knowledge levels. The traditional e-learning systems have got problems to achieve its learning goals.

2.2 Recommender system

Learning resources available in the web are heterogeneous and in various media formats as well. The probability of learners accessing the relevant items is of greater concern and is intensively researched. Recommender Systems are software tools and techniques that provide suggestions to a user in various decision making processes [5]. In the e-learning context, a recommender system is a software agent that recommends useful and interesting learning resources to a learner by accounting the ratings, preferences and expertise of other learners.

The basic elements which constitute a recommender system are event, session and recommendation process [7]. An event is a call to the system provoked by an action performed by the user. For instance, every click on a hyperlink generates a new event session s(u) is a set of close events provoked by a user u. A recommendation process is the sequence of actions that a recommender executes to produce a set of recommendations. An item denotes what the system recommends to users. A recommendation event can have one or more sessions. The basic units of a recommendation event are the set of items available to be recommended, a recommendation window created for each event, a filter for creating and filling a window, and a guide to wrap and display the items to be recommended [7].

2.3 Recommender system Filtering Approaches

In general, e-learning recommender systems have three types of filtering approaches these are content-based filtering, collaborative filtering and knowledge-based filtering [16]. To improve the accuracy of performance and result of filtering, researchers devised hybrid-filtering approach by combining the other approaches [15].

Figure 2-1 shows Recommendation system filtering Technique tree

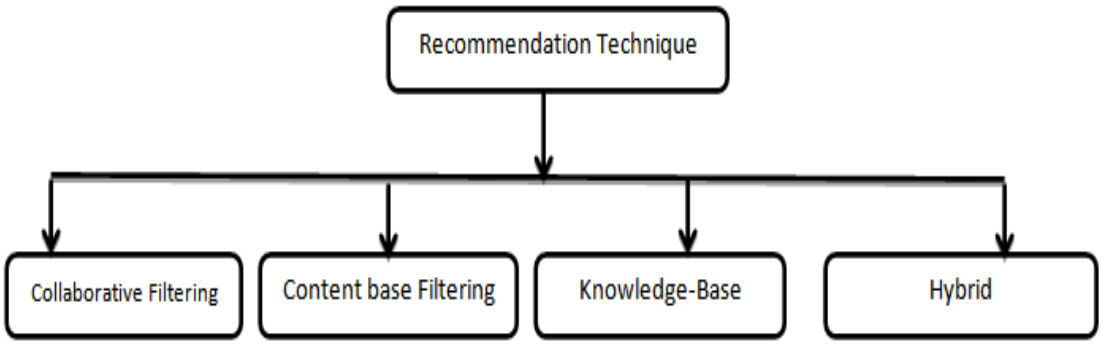

Figure 2-1: Recommender System Filtering Approach

Content-Based Filtering (CBF)

In CBF, the users/learners are recommended relevant items/learning contents that are similar to the ones they preferred in the past [15]. This type of filtering relies on the of user / item profiles that assigns consequence to these characteristics. Sometimes, there is not enough information in the items' profile or the user did not access the item before and rate it before, so the system is unable to conclude any recommendation for the users / learners [20]. This problem is called cold-start in the term of recommender systems. Cold start problem occurs in both the user and the item. These problems result when the domain system does not have enough information on both items (learning content) and users / learners' profiles [15]. Consequently the system is unable to acclaim the users/learners interest and unable to

recommend the relevant item accurately. In both (user and item) cases the cold start problem occur because of ratings. Item cold start problem occurs when the item(learning content) has not been rated by any user / learner or it haven't enough keywords and tags information are not available in its profile. If the user / learner has not rated any item (learning content) before and does not have sufficient information (item-ratings) regarding required interest / goals, the domain system is unable to recommend any item (learning content) to user/learner. This is called user cold start problem [15].

Collaborative Filtering

One of the most successful and widely used technologies for building recommendation system is Collaborative Filtering (CF) [23].The system relied on the explicit opinions of people from a small community, such as an office work group.CF systems collect visitor opinions on a set of objects, using ratings provided by the users or implicitly computed, to form peer groups and that establishes the basis of a learning system to predict a particular user's interest in an item [22]. It is often based on matching, in real-time, the current user's profile against similar records (nearest neighbors) obtained by the system over time from other users. The ratings collected by the system may be both implicit and explicit [22].

Explicit voting refers to a user consciously expressing her preference for a title, usually on a discrete numerical scale. The lack of explicit user ratings as well as the sparseness and the large volume of data pose limitations to standard CF. As a result; it becomes hard to scale CF techniques to a large number of items, while maintaining reasonable prediction performance and accuracy [50].

CF techniques can be an important part of the recommender systems. One key advantage of CF based on explicit voting is that it does not consider the content of the items being recommended, rather than map user to items through user ratings. In addition to the

limitations mentioned above, another difficult, though common problem of CF systems is the cold-start problem, where recommendations are required for items that no one in the data set has yet rated [69].

Knowledge-Based Filtering (KBF)

Knowledge-based filtering (RBF) approach does not seek to build long-term generalization of their users/learners but they prefer to generate a relevant recommendation based on matching users / learner's needs, interests and preferences [16]. With this approach, the relationship between users' needs and relevant recommended items can be explicitly modulated in a knowledge base on underlying [19]. Generally, these types of systems attempt to solve three types of knowledge questions that are based on user profiling, point profiling and comparison between the user and the point corresponding to the user and binding targets / interest / needs [15]. Gradually, the knowledge profile of the user plays an essential role in this filtering approach.

Hybrid Filtering (HF)

The HF generally combines the content-based and collaborative filtering methods [15]. These combined methods borrow both content-based and collaborative (some time knowledge-based and collaborative or combination of all) features to get the user's interest and recommend him / her required relevant items (learning content) more closely related to learner goal / interest and preferences. Hybrid filtering technique improves the user element of the cold start problem more than both content-based filtering and collaborative [51]. In hybrid systems, however; the main problem is the complexity of time data. Time complexity occurs when the size of the same dataset increases and the recommender system performs slowly when the system uses more than one but different dataset. These multiple datasets slow down the recommendation performance and decrease the learner interests [17].

CLUSTERING APPROACHES

Clustering approaches divide the user base into segments, or groups of users who have very similar preferences, treating the task as a classification problem and predictions for a user are then calculated based on averaging the opinions of other users in the cluster [47]. In some clustering approaches, a user can have partial participation in numerous clusters, and the predictions are then based on the average across the clusters of participation, weighted by degree of participation. Clustering techniques, once clustering is done, can have a good online scalability and performance, as the size of the group to be analyzed is smaller than e.g. in collaborative filtering. However, the recommendations generated are not very well personalized and the quality can be low [54] .By using its strengths, clustering can be used as a first step method that shrinks the candidate set.

DEMOGRAPHIC FILTERING

Demographic information, such as gender, age, and country of residence, can be used to generate somewhat personalized recommendations [36]. The idea is that users with common demographic attributes also have common preferences [36]. Demographic data is gathered through various means, e.g. by surveys or through machine learning [21]. The demographic information is first matched to a stereotype and then the items connected to it are recommended or the ratings in a demographic place to which the user belongs are combined to produce recommendations [21]. Personalization is naturally limited when users are generalized in this manner [36]. The good side of demographic filtering is that it does not require a history of user ratings and, consequently, does not suffer from new user problems [21]. Consequently, demographic filtering has been combined e.g. with collaborative filtering to deal with cold-start issues [36]. Demographic data has also been suggested to be used to lessen cold-start problems outside of pure demographic filtering approaches. However, demographic recommenders also

have to gather the demographic information somehow, so they are not free from data gathering challenges altogether [21].

KNOWLEDGE-BASED, UTILITY-BASED, AND CRITIQUING SYSTEMS

Consider both utility-based and knowledge-based systems as variants of case-based systems [21]. Case-based systems are based on case-based reasoning techniques that solve new problems using a case database of past problem solving experiences, retrieving a similar case and adapting its solution to the current problem. In case-based recommender systems, items are represented as cases and recommendations are generated by picking the case, i.e. items, that correspond most closely to the user query or profile. In comparison to content-based filtering, case based systems rely on a more structured representation of item content and they use various similarity assessment approaches for identifying similar cases [21]. Utility-based systems generate recommendations by computing the utility of each object for the active user [32]. Even non-product attributes, e.g. product availability, can be factored into the utility computations. However, this flexibility is also a challenge in utility-based systems, as creating a utility function for each user means that each user has to construct a complete preference function, considering the significance of every possible feature. In most cases, this constitutes a significant burden for the user, at least in the case of more complex and subjective domains [21]. Utility-based approach can be seen as a special case of knowledge-based approach [20]. Knowledge-based systems also require the user to input their preferences [21]. After the user has input their preferences, the system presents them with recommendations based on the knowledge contained in the system [20]. After a few iterations of the process, the recommendations are tailored to the user. In learning knowledge-based systems, feedback from the user is used to add to the knowledge [20]. The problem for all knowledge-based systems, including recommenders, is that there is a need for knowledge acquisition that is categorized in

three types of knowledge necessary in a knowledge-based recommender system[21]: 1) Catalog knowledge: Knowledge about the items to be recommended and their attributes; 2) Functional knowledge: The system has to "have knowledge about how a particular item meets a particular user need" so that they can "reason about the relationship between a need and a possible recommendation 3) User knowledge: The system also needs to have some knowledge about the user, perhaps "general demographic information or specific information about the need for which a recommendation is sought. Overall, knowledge-based systems are more suited to casual browsing than utility-based ones, as they demand less effort and knowledge from the user [21]. Knowledge-based and utility-based approaches do not need to accrue user profiles; instead, they "base their advice on an evaluation of the match between a user's need and the set of options available". Consequently, they do not suffer from cold-start or sparsity problems, since the recommendations are not based on accumulated statistical data. Also, there is no plasticity problem, a problem that occurs in other approaches that base recommendations on user modeling when the user preferences change. Utility-based and knowledge-based systems that do not rely on historical profile data do not need any re-training and can respond immediately to the current user need. In addition to utility-based and knowledge-based systems critiquing based systems was also considered, as a third type of case-based systems [20, 21]. Critiquing systems represent a form of case-based systems that operate in a reactive fashion. The critiques from the user allow the recommender to refine its recommendations until a satisfactory item is found [21]. In this way, critique-based systems allow users to construct and refine their preference model incrementally [20].

2.4 Automated Recommender Systems

Most of the techniques for automated recommender systems are based on data mining methods, which attempt to discover patterns or trends from a variety of sources. Web usage mining is an obvious and popular one of these techniques. Recently, a number of approaches

have been developed dealing with specific aspects of Web usage mining like automatically discovering user profiles, web based recommender systems, Web perfecting, design of adaptive Web sites, etc. In all these applications the goal is the development of an effective prediction algorithm. The core issue in prediction is the development of an effective algorithm that deduces the future user requests. The most successful approach towards this goal has been the exploitation of the user's access history to derive prediction. This section describes pattern discovery methods that have been applied to Web domain.

It is very difficult to classify the studies according to the methods they use for Web usage mining. In most of the works mentioned below, methods are combined together in discovering usage patterns in Web domain.

Statistical Analysis

Statistical techniques are the most common methods to extract knowledge about visitors to a Web based system. By different kinds of statistical analysis (frequency, median, mean, etc.) of the session file, one can extract statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This type of knowledge can be potentially useful for improving the system performance, and enhancing the security of the system [69].

An example for the application of statistical methods to web mining is Page Gather [52]. This algorithm processes the access logs by using a statistical approach to find pages (resources) that are often visited together. It then creates a graph in which each node represents a page at the Web and finds maximal cliques in the graph in order to discover user profiles. While the generated profiles were not integrated as part of a recommender system, they were used to automatically synthesize alternative static index pages for a site.

Association Rules

Association rules capture the relationships among items based on their patterns of co-occurrence across transactions. The problem of discovering association rules was introduced in [53]. Given a set of transactions, where each transaction is a set of items, an association rule is an expression of the form X → Y, where X (defined as the left-hand-side (LHS) of the association rule) and Y (defined as the right-hand-side (RHS) of the association rule) are sets of items such that no item appears more than once in XUY. The intuitive meaning of such a rule is that transactions in the database which contain the items in X tend to also contain the items in Y. Two common numeric measures assigned to each association rule are "support" and "confidence". Support quantifies how often the items in X and Y occur together in the same transaction as a fraction of the total number of transactions, or |X U Y |/D where |D| denotes the total number of transactions. Confidence quantifies how often X and Y occurs together as a fraction of the number of transactions in which X occurs, or |X U Y |/X. In the context of Web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold [54]. These pages may not be directly connected to one another via hyperlinks. For example, using association rule discovery techniques, we can find correlations such as following:

Clustering

Clustering is a technique to group together a set of items having similar characteristics. In the Web usage domain, there are three kinds of interesting clusters to be discovered: 1: Session clusters; 2: User clusters; and 3: Page clusters. Session clustering implementation allows clustering of user sessions in which users have similar access patterns. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Page clustering can be partitioned into two methods. The first is to cluster pages according to their contents. For this method an analysis of the content of Web site is needed. The second method computes clusters

of page references based on how often they occur together [54].

Classification

Classification is the task of mapping a data item into one of several predefined classes. In the context of Web usage mining, one is interested in developing a usage profile belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. In Web usage mining, classification techniques allow one to develop a profile for users who access particular server files based on their demographic information available on those users, or based on their access patterns.

Sequential Patterns

Sequential pattern mining, which discovers frequent subsequences as patterns in a sequence database, is an important data mining problem with broad applications, including e-learning recommender system, the analysis of customer purchase behavior, Web access patterns, scientific experiments, disease treatments, natural disasters, DNA sequences, and so on[41]. The sequential pattern mining problem was first introduced by Agrawal and Srikant in [55]. Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user specified min_support threshold, sequential pattern mining finds all of the frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences is no less than min_support.

2.5 E-learning recommender systems:

Today, e-business applications and e-services are commonly taking advantage of advanced information and communication technologies and methodologies to personalize their interactions with users. Personalization aims to tailor services to individual needs, and its immediate objectives are to understand and to deliver highly focused, relevant content, services

and products matched to users' needs and contexts [63]. E-services personalization and web adaptation have been employed in many different ways:

(i) the personalization service can be designed and used as an advice-giving system to provide recommendations to each individual and to generate up-sell and cross-sell opportunities (ii) personalization services are used to (dynamically) structure the index of information, product pages based on click-stream analysis to minimize the users' search efforts, where personalized content based on the user's profile is generated. The users can personalize not only the content but also the interface of the application used [64]. Applications of personalization technology are found to be useful in different domains. These include information dissemination, entertainment recommendations, search engines, medicine, tourism, financial services, consumer goods and e-learning [63].

All these methods are based on the analysis of log files as our methodology does. Especially clustering in web usage domain allows clustering of learner session in which learners have spent similar amount of time on resources assuming that visiting duration shows interest on that specific resource. And learners who are in the same group are recommended resources as they have similar interest.

3 Chapter Three:

Methodology

3.1 Overview

Research methodology is the core of the research process to carry out the research (Yin, 2003). It is an overall roadmap of the research process. It includes the research standard, procedure, data collection instruments, data analysis methods, and data interpretation. As a result, this step needs much attention on choosing the appropriate method which can provide the desired output.

3.2 Research design

Research design is the blue print or plans of procedure that cover the decision from a wide assumption to detailed methods of data collections (Creswell, 2009). Previous scholars used different types of research methodology depending on the kinds of the problem situations, the existing knowledge and the resource availability. Accordingly, in this study we used quantitative research methods. Quantitative research methodology was selected basically to gather all information in Log file that can be transformed into useable statistics.

To conduct this research, we used extracted log file of web based e-learning system to collect quantitative data from target population.

3.3 Data Collection and Analysis

The research aim is to develop learner interest model for e-learning resource recommender system for the case of Ethiopian higher learning institution by extracting log file of their e-learning management system. Log files contain information about User Name, IP Address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent. The log files are maintained by the web servers. By analyzing these log files

gives idea about what a user interest is [22]. Preliminary study was conducted to identify universities which use e-learning system in their teaching and learning processes. The study shows almost no university integrates in their curriculum system. As a result we could not find enough e-learning system's log files for our user interest modeling purpose. We tried to alleviate lack of log file by in communicating University of North Texas (UNT) to find their students e-learning activities' log file. But the U.S. privacy law and institution's policies are very strict about sharing such data. Further efforts were taken in developing and lunching web based e-learning system prototype in St. Mary University though not successful. At end, the researcher was forced to use the log file of e-learning system of Mekele University (Model version 2.9.1) only.

3.3.1 Data collection procedure

Data collection for this study began on a first week of May, 2015, and ended in the fourth week of August, 2015. The primary data for the research was gathered by extracting log file from web based e-learning system for only one course this is because it is the only course that learners are daily access e-learning resources through the e-learning management system. The research aim is to develop learner interest model for e-learning resource recommender system.

3.3.2 Data Analysis Procedures

Data analysis involves critical thinking, expertise, and inside understanding about the concept. It is performed after data collection from the data source and some primary data processing activities. The analysis of data is done according to the research objective. Log file analysis can be either qualitative or quantitative in nature. It can also be manual or automated. Qualitative log file analysis is generally conducted manually (by a person interpreting logs); however, computerized tools can help the human reader. The human reader in a qualitative analysis may use a number of theoretical frameworks to organize that analysis such as grounded theory activity theory, distributed cognition etc [34]. Quantitative log file analysis can be either manual

(with a human translating log entries into specified metrics), or automated (with a computer program performing that translation). Some kinds of log file data lend themselves to automated analysis more readily than others. The nature of the research question affects whether manual or automated analysis is preferable. Quantitative analysis often simply measures amounts of activity over time.

Data analysis is the process of examining the trend of log file data in web based e-learning system. In this research the analysis of log file data was based on manual quantitative log file analysis methods and In order to improve the validity of our data analysis, we eliminate bias, outliers, and errors.

Grouping Learners

Clustering is a standard procedure in multivariate data analysis. It is designed to explore an inherent natural structure of the data objects, where objects in the same cluster are as similar as possible and objects in different clusters are as dissimilar as possible [79]. The equivalence classes induced by the clusters provide a means for generalizing over the data objects and their features.

Clustering is an exploratory data analysis. Therefore, the explorer might have no or little information about the parameters of the resulting cluster analysis. In typical uses of clustering the goal is to determine all of the following:

- The number of clusters,
- The absolute and relative positions of the clusters,
- The size of the clusters

Clustering Methods

We do have many clustering algorithms. The main reason for having many clustering methods is the fact that the notion of “cluster” is not precisely defined [79]. Consequently many clustering methods have been developed, each of which uses a different induction principle. Categorizing of clustering methods into three main categories [56]:

A. Density-based methods

It assumes that the points that belong to each cluster are drawn from a specific probability distribution [80].The overall distribution of the data is assumed to be a mixture of several distributions. The aim of these methods is to identify the clusters and their distribution parameters. These methods are designed for discovering clusters of arbitrary shape which are not necessarily convex

The idea is to continue growing the given cluster as long as the density (number of objects or data points) in the neighborhood exceeds some threshold. Namely, the neighborhood of a given radius has to contain at least a minimum number of objects. When each cluster is characterized by local mode or maxima of the density function, these methods are called mode-seeking. Clusters are formed by connecting neighboring ’core’ objects and those ’non-core’ objects either serve as the boundaries of clusters or become outliers[80]. Since the noises of the data set are typically randomly distributed, the density within a cluster should be significantly higher than that of the noises. Therefore, density-based approaches have the advantage of extracting clusters from a highly noisy environment. Much work in this field has been based on the underlying assumption that the component densities are multivariate Gaussian (in case of numeric data) or multinomial (in case of nominal data). An acceptable solution in this case is to use the maximum likelihood principle. According to this principle, one should choose the clustering structure and parameters such that the probability of the data being generated by such clustering structure and parameters is maximized. The expectation maximization algorithm(EM) which is a general-purpose maximum likelihood algorithm for missing-data problems, has been

applied to the problem of parameter estimation[80]. This algorithm begins with an initial estimate of the parameter vector and then alternates between two steps: an "E-step", in which the conditional expectation of the complete data likelihood given the observed data and the current parameter estimates is computed, and an "M-step", in which parameters that maximize the expected likelihood from the E-step are determined[19]. This algorithm was shown to converge to a local maximum of the observed data likelihood.

B. Model-based Clustering Methods

These methods attempt to optimize the fit between the given data and some mathematical models. Unlike conventional clustering, which identifies groups of objects; model-based clustering methods also find characteristic descriptions for each group, where each group represents a concept or class. The most frequently used induction methods are decision trees and neural networks.

C. Grid-based Methods

These methods partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The main advantage of the approach is its fast processing time [79].

In this research the clustering of learners based on their activity in the log file is made by using of density based model clustering since density-based approaches have the advantage of extracting clusters from a highly noisy environment.

3.3.3 Tools

In order to do the experiment the researcher have adapted XLSTAT (version 2015), statistical analysis software, add-in offers a wide variety of functions to enhance the analytical capabilities of Excel, making it the ideal tool for everyday data analysis and statistics

requirements [62]. The rational for using the XLSTAT statistical analysis software from other software is described as follows. The XLSTAT system contains some visualization, pre and post processing tools as well as suite of statistical analysis for classification, and clustering. XLSTAT relies on Excel for the input of data and the display of results, but the computations are done using autonomous software components. The use of Excel as an interface makes XLSTAT user-friendly and multivariate data analysis package.

3.4 System Diagram for Automated e-Learning Recommender System

As shown in Figure 3-1, the overall process of automated recommendation can be divided into four components, namely: 1: Data collection, 2: Data preparation and cleaning, 3: Pattern extraction, and 4: Prediction.

The first step is data collection from data sources. Web usage mining can potentially use data from the several sources but for our case the data is collected from server level collection of e-learning system:

- Server Level Collection: A web based e-learning server log is an important source for performing Web usage mining because it explicitly records the browsing behavior of web based e-learning user. Server log file provide details about file requests to a Web server and the server response to those requests. In the access log, which is the main log file, each line describes the source of a request, the file requested, the date and time of the request, the user identification, and other data such as errors and the identity of referring pages.

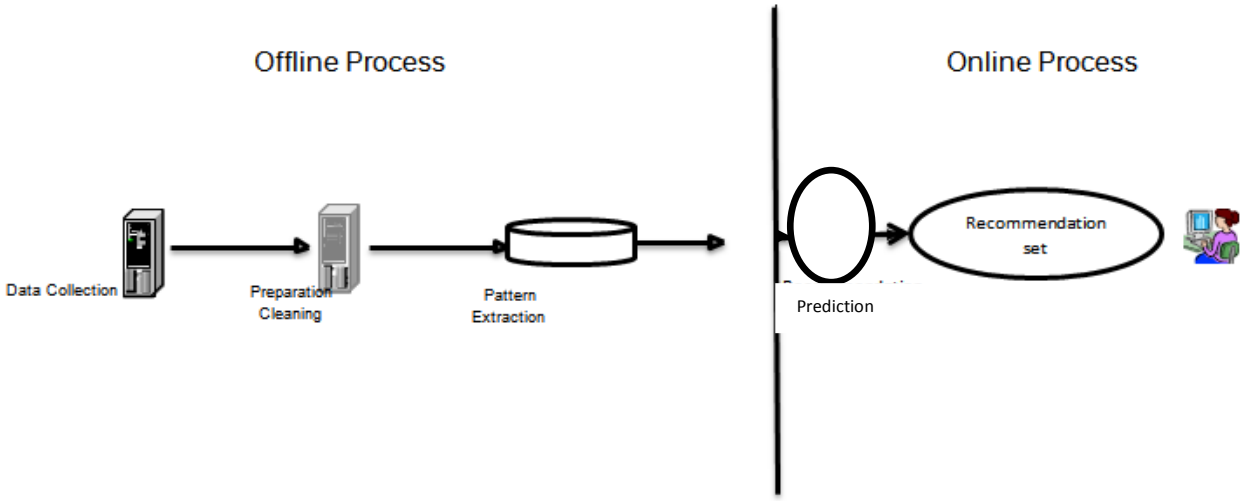

Figure 0-1: Simple diagram of recommender system

The second step is to clean the data and prepare forming the usage patterns. Fundamental methods of data cleaning and preparation were applied in this thesis. And details are given in this chapter. The third step is to extract usage patterns. Partitioning Learner session in to groups such that session that represent similar aggregate interest of learner are placed in the same group. The fourth step is to build a predictive model based on the extracted usage patterns. The prediction step is the real-time processing of the model, which considers the active user session and makes recommendations. Once the mining tasks are accomplished, the discovered patterns are used by the online component of the model to provide dynamic recommendations to users based on their current navigational activity. Finally the produced recommendation set is then added to the last request page as a set of links before the page is sent to the client browser.

3.5 Data preparation and Cleaning

3.5.1 Data Source

In this research, we use log files of web based e-learning system, Moodle (Version 2.9.1) from Mekele University As shown in Table 3-1. The details of the log files are given below. A web based e-learning system's log file is an important source for performing web based e-learning system resource usage mining because it explicitly records the browsing behavior of web based e-learning system's users(learner).The log file data base records the time and date of the transaction. It records the IP address to which the e-learning resource was sent. If the user goes to a page by clicking a link on some other page, the server records the address of the page with that link. It also records some details about how the e-learning resource is utilized (View, update, delete) by users and any errors that may have occurred as well as information about the different activities of a learner or the teachers make on the system like forum, submit of assignment and made enrolment of users so that the user can access resources. The data that is recorded in the log file database reflects the (possibly concurrent) access of a web based e-learning system's by multiple users.

In order to understand the user behavior, the following information should be extracted from log files:

- Who is visiting the web based e-learning systems? One of the major steps in web based e-learning system usage mining is to identify unique users in order to obtain what the user view;
- The path users take through the web based e-learning system's pages. With knowledge of each page that a user viewed, updated or deleted and the order, one can identify how users navigate through the web based e-learning system's pages.

- How much time users spend on each page? A pattern of lengthy viewing time on a page (Resource) might lead one to deduce that the page is interesting; and how often (frequently) a page is visited can also how important that page for users who are in the same group.
- Where visitors are leaving the web based e-learning system's? The last page a user viewed before leaving the web based e-learning system might be a logical place to end a server session.

	Course	Time	IP Address	Full name	Action	Resourse
2	IENG 5911	2015 August 14 19:59	10.128.130.50	User 4	logout	
3	IENG 5912	2015 August 14 18:51	10.128.130.50	User 4	course view	Executive Guide to Improvement and Change,
4	IENG 5912	2015 August 14 17:46	10.128.130.50	User 4	resource view	Ch 5. Kaizen
5	IENG 5912	2015 August 14 16:58	10.128.130.50	User 4	resource view	Ch 5. Kaizen
6	IENG 5912	2015 August 14 16:52	10.128.130.50	User 4	resource view	Ch 3. Quality Measurement
7	IENG 5912	2015 August 14 15:47	10.128.130.50	User 4	resource view	Quality Measurement
8	IENG 5915	2015 May 29 23:17	213.55.104.235	User 4	logout	
9	IENG 5916	2015 May 29 22:01	213.55.104.235	User 4	resource view	Executive Guide to Improvement and Change,
10	IENG 5916	2015 May 29 21:06	213.55.104.235	User 4	resource view	Ch 3. Quality Measurement
11	IENG 5915	2015 May 29 20:10	213.55.104.235	User 4	resource view	Sigma For Financial Senices New York, NY
12	IENG 5915	2015 May 29 19:14	213.55.104.235	User 4	resource view	Quality Standards, Milwaukee, Wisconsin
13	IENG 5912	2015 May 29 18:01	213.55.104.235	User 4	resource view	Ch 4. Quality Measurement
14	IENG 5912	2015 May 29 18:00	213.55.104.235	User 4	course view	Quality Managment
15	IENG 5914	2015 June 3 20:41	213.55.104.235	User 4	Logout	
16	IENG 5915	2015 June 3 19:25	213.55.104.235	User 4	resource view	Sigma For Financial Services New York, NY
17	IENG 5912	2015 June 3 19:00	213.55.104.235	User 4	resource view	Manufacturing
18	IENG 5912	2015 June 3 18:12	213.55.104.235	User 4	course view	Quality Managment
19	IENG 5915	2015 May 29 22:00	10.128.130.105	User 4	logout	
20	IENG 5916	2015 May 29 21:49	10.128.130.105	User 4	resource view	Executive Guide to Improvement and Change,
21	IENG 5916	2015 May 29 20:58	10.128.130.105	User 4	resource view	Quality Standards, Milwaukee, Wisconsin
22	IENG 5915	2015 May 29 20:29	10.128.130.105	User 4	resource view	Ch 2. Quality Measurement
23	IENG 5915	2015 May 29 19:35	10.128.130.105	User 4	resource view	Sigma For Financial Senices New York, NY
24	IENG 5912	2015 May 29 18:01	10.128.130.105	User 4	resource view	Ch 3. Quality Measurement
25	IENG 5912	2015 May 29 18:00	10.128.130.105	User 4	course view	Quality Managment
26	IENG 5915	2015 May 29 23:55	213.55.104.235	User 4	logout	
27	IENG 5916	2015 May 29 23:49	213.55.104.235	User 4	resource view	Executive Guide to Improvement and Change,
28	IENG 5916	2015 May 29 23:42	213.55.104.235	User 4	resource view	Ch 3. Quality Measurement

Table 0-1: Sample Moodle log file form Mekele Universityple

For the sake of easy management and simplicities of the resource in Moodle e-learning system, we changed manually the name of the topic title and reference of the course that are used as link to get the respective content as P_0,P_1,P_2……..$P_{n;}$ where P_0 is My Home page in Moodel, it is a customizable page for providing users with links to their courses and activities within them; and the name of the user as U_1,U_2…..U_n, As shown in Table 3-1-1 which is what users will see once logged in, for the course , P_1 is for resource of "chapter 1" whose title is quality in design and manufacturing and so on. The following table shows the representation of each resource with respective pages.

1	Course	Time	IP Address	Full name	Action	Resourse	Equivalent page
2	IENG 5911	2015 August 14 19:59	10.128.130.50	U4	logout		
3	IENG 5912	2015 August 14 18:51	10.128.130.50	U4	course view	Executive Guide to Improvement and Change,	P9
4	IENG 5912	2015 August 14 17:46	10.128.130.50	U4	resource view	Ch 5. Kaizen	P5
5	IENG 5912	2015 August 14 16:58	10.128.130.50	U4	resource view	Ch 5. Kaizen	P5
6	IENG 5912	2015 August 14 16:52	10.128.130.50	U4	resource view	Ch 3. Quality Measurement	P3
7	IENG 5912	2015 August 14 15:47	10.128.130.50	U4	resource view	Quality Measurement	P0
8	IENG 5915	2015 May 29 23:17	213.55.104.235	U4	logout		
9	IENG 5916	2015 May 29 22:01	213.55.104.235	U4	resource view	Executive Guide to Improvement and Change,	P9
10	IENG 5916	2015 May 29 21:06	213.55.104.235	U4	resource view	Ch 3. Quality Measurement	P3
11	IENG 5915	2015 May 29 20:10	213.55.104.235	U4	resource view	Sigma For Financial Services New York, NY	P6
12	IENG 5915	2015 May 29 19:14	213.55.104.235	U4	resource view	Quality Standards, Milwaukee, Wisconsin:	P7
13	IENG 5912	2015 May 29 18:01	213.55.104.235	U4	resource view	Ch 4. Quality Measurement	P4
14	IENG 5912	2015 May 29 18:00	213.55.104.235	U4	course view	Quality Managment	P0
15	IENG 5914	2015 June 3 20:41	213.55.104.235	U4	Logout		
16	IENG 5915	2015 June 3 19:25	213.55.104.235	U4	resource view	Sigma For Financial Services New York, NY	P6
17	IENG 5912	2015 June 3 19:00	213.55.104.235	U4	resource view	Manufacturing	P1
18	IENG 5912	2015 June 3 18:12	213.55.104.235	U4	course view	Quality Managment	P0
19	IENG 5915	2015 May 29 22:00	10.128.130.105	U2	logout		
20	IENG 5916	2015 May 29 21:49	10.128.130.105	U2	resource view	Executive Guide to Improvement and Change,	P9
21	IENG 5916	2015 May 29 20:58	10.128.130.105	U2	resource view	Quality Standards, Milwaukee, Wisconsin:	P7
22	IENG 5915	2015 May 29 20:29	10.128.130.105	U2	resource view	Ch 2. Quality Measurement	P2
23	IENG 5915	2015 May 29 19:35	10.128.130.105	U2	resource view	Sigma For Financial Services New York, NY	p3
24	IENG 5912	2015 May 29 18:01	10.128.130.105	U2	resource view	Ch 3. Quality Measurement	P3
25	IENG 5912	2015 May 29 18:00	10.128.130.105	U2	course view	Quality Managment	P0
26	IENG 5915	2015 May 29 23:55	213.55.104.235	U4	logout		
27	IENG 5916	2015 May 29 23:49	213.55.104.235	U4	resource view	Executive Guide to Improvement and Change,	P9
28	IENG 5916	2015 May 29 23:42	213.55.104.235	U4	resource view	Ch 3. Quality Measurement	P3

Table 0-2: Sample Course substituted with respective page number

However, a log file does not contain all of the information required for web based e-learning system usage mining. Even if it contains other data, that make it difficult to interpret. As a result of this, irrelevant information has to be removed and some manipulations on relevant data should be done on raw dataset. Regard less of the application, data preparation and cleaning steps should be completed in order to create server sessions. Data preparation and cleaning tasks performed in this study consist of the following steps: 1: User Identification; 2: Session Identification; 3: Page Time Calculation; and 4: Data Cleaning. This chapter also presents the methods applied in these steps.

3.5.2 User Identification

In order to know who is visiting the web based e-learning system, the log file must contain a person ID such as log into the system or to the user's own computer. In this regard each web based e-learning system's learner has its own user name and password. Since this user name and password can identify the user uniquely in the system thus we take it as user identification attribute or user ID. Our system converts a set of serve logs expressed as:

$$L = L_1, L_2, \ldots,.., L_{|L|}$$

$$L_i = (UI_i, Course, Time_i, Action_i, ResID_i, IP_i)$$

$$L_i \in L,\ i \in [1,|L|]$$

Into a set of user transactions T such that transactions are grouped by users:

$$T = T_1, T_2 , .,.,., .,T_{|L|}$$

$$T_i = (UID_i, Course_i , TIME_i , Action_i, ResID_i, IP_i)$$

$$T_i \in T, i \in [1,|L|]$$

Where $|L|$ is the number of logs in L and UID_i is the User Identification Number. For every identical user name, we simply assign a unique User Identification Number for each unique user. Thus, some of the user requests in T have the same User Identification Number, $Course_i$ represent course identification number, $TIME_i$ represent at what time a user start visiting a page, $Action_i$ represents what kind of action a user doing when he visits the page(update or viewing etc), $ResID_i$ stands for e-learning resource Identification number.

1	Course	Time	IP Address	Full name	Action	Resourse	Equivalent page
2	IENG 5911	2015 August 14 19:59	10.128.130.50	U4	logout		
3	IENG 5912	2015 August 14 18:51	10.128.130.50	U4	course view	Executive Guide to Improvement and Change,	P9
4	IENG 5912	2015 August 14 17:46	10.128.130.50	U4	resource view	Ch 5. Kaizen	P5
5	IENG 5912	2015 August 14 16:58	10.128.130.50	U4	resource view	Ch 5. Kaizen	P5
6	IENG 5912	2015 August 14 16:52	10.128.130.50	U4	resource view	Ch 3. Quality Measurement	P3
7	IENG 5912	2015 August 14 15:47	10.128.130.50	U4	resource view	Quality Measurement	P0
8	IENG 5915	2015 May 29 23:17	213.55.104.235	U4	logout		
9	IENG 5916	2015 May 29 22:01	213.55.104.235	U4	resource view	Executive Guide to Improvement and Change,	P9
10	IENG 5916	2015 May 29 21:06	213.55.104.235	U4	resource view	Ch 3. Quality Measurement	P3
11	IENG 5915	2015 May 29 20:10	213.55.104.235	U4	resource view	Sigma For Financial Services New York, NY	P6
12	IENG 5915	2015 May 29 19:14	213.55.104.235	U4	resource view	Quality Standards, Milwaukee, Wisconsin:	P7
13	IENG 5912	2015 May 29 18:01	213.55.104.235	U4	resource view	Ch 4. Quality Measurement	P4
14	IENG 5912	2015 May 29 18:00	213.55.104.235	U4	course view	Quality Managment	P0
15	IENG 5914	2015 June 3 20:41	213.55.104.235	U4	Logout		
16	IENG 5915	2015 June 3 19:25	213.55.104.235	U4	resource view	Sigma For Financial Services New York, NY	P6
17	IENG 5912	2015 June 3 19:00	213.55.104.235	U4	resource view	Manufacturing	P1
18	IENG 5912	2015 June 3 18:12	213.55.104.235	U4	course view	Quality Managment	P0
19	IENG 5915	2015 May 29 22:00	10.128.130.105	U2	logout		
20	IENG 5916	2015 May 29 21:49	10.128.130.105	U2	resource view	Executive Guide to Improvement and Change,	P9
21	IENG 5916	2015 May 29 20:58	10.128.130.105	U2	resource view	Quality Standards, Milwaukee, Wisconsin:	P7
22	IENG 5915	2015 May 29 20:29	10.128.130.105	U2	resource view	Ch 2. Quality Measurement	P2
23	IENG 5915	2015 May 29 19:35	10.128.130.105	U2	resource view	Sigma For Financial Services New York, NY	p3
24	IENG 5912	2015 May 29 18:01	10.128.130.105	U2	resource view	Ch 3. Quality Measurement	P3
25	IENG 5912	2015 May 29 18:00	10.128.130.105	U2	course view	Quality Managment	P0
26	IENG 5915	2015 May 29 23:55	213.55.104.235	U4	logout		
27	IENG 5916	2015 May 29 23:49	213.55.104.235	U4	resource view	Executive Guide to Improvement and Change,	P9
28	IENG 5916	2015 May 29 23:42	213.55.104.235	U4	resource view	Ch 3. Quality Measurement	P3

Table 0-3:Transactions after user identification step

The Table 3-2 shows assignment of use identification number (**UID**) for each user in each transaction T_i.

3.5.3 Session Identification

Once users have been identified, the click-stream for each user must be divided into sessions. A click stream is the recording of what a user clicks on while using the web. Every time he or she clicks on a link, an image, or another object on the page, that information is recorded and stored. This information can help to find out the habits of one individual [38].

A session can be defined as the time period of an activity from its beginning until its end. The activity may end for a variety of reasons: The user reaches her goal, the user finds the activity not interesting anymore, or there is a time constraint involved. A session has a clean meaning in an online system with user login and logout facilities [53]. A session, in this case, starts from the time when a user performs login, and finishes upon logout in Moodle. AccordingtoW3C, a session is the group of activities performed by a user from the moment he/she enters the site to the moment she leaves it [54]. Since there is official login and logout to access and use most of web based e-learning system, it is very clear when a session begins and ends. Since page (Resource) request from other servers are not typically available in our case and a user may visit a web based e-learning system more than once, the log files records multiple sessions for each user. The goal of session identification is to divide the page accesses of each user into individual sessions. The simplest method of achieving this is through looking an attribute value for "Action" for every L_i if the entry value of Attribute Action is "log in" or "log out" to the system.

A new field is added to the user transactions created in the previous step. Thus, every user transaction in T takes the form:

T_i= (UID_i, $Course_i$, $TIME_i$, $Action_i$, $ResID_i$, IP_i ,SID_i)

Where SID_i is unique session identification number;

It is given every time a new session is created for the same user. Table 3.3 shows user transactions with **UID** and **SID** extracted from the server logs in Table 3.1

1	Course	Time	IP Address	Full name	Action	Resourse	Equivalent page	SID
2	IENG 5911	2015 August 14 19:59	10.128.130.50	U4	logout			
3	IENG 5912	2015 August 14 18:51	10.128.130.50	U4	course view	Executive Guide to Improvement and Change,	P9	7
4	IENG 5912	2015 August 14 17:46	10.128.130.50	U4	resource view	Ch 5. Kaizen	P5	7
5	IENG 5912	2015 August 14 16:58	10.128.130.50	U4	resource view	Ch 5. Kaizen	P5	7
6	IENG 5912	2015 August 14 16:52	10.128.130.50	U4	resource view	Ch 3. Quality Measurement	P3	7
7	IENG 5912	2015 August 14 15:47	10.128.130.50	U4	resource view	Quality Measurement	P0	7
8	IENG 5915	2015 May 29 23:17	213.55.104.235	U4	logout			6
9	IENG 5916	2015 May 29 22:01	213.55.104.235	U4	resource view	Executive Guide to Improvement and Change,	P9	6
10	IENG 5916	2015 May 29 21:06	213.55.104.235	U4	resource view	Ch 3. Quality Measurement	P3	6
11	IENG 5915	2015 May 29 20:10	213.55.104.235	U4	resource view	Sigma For Financial Services New York, NY	P6	6
12	IENG 5915	2015 May 29 19:14	213.55.104.235	U4	resource view	Quality Standards, Milwaukee, Wisconsin:	P7	6
13	IENG 5912	2015 May 29 18:01	213.55.104.235	U4	resource view	Ch 4. Quality Measurement	P4	6
14	IENG 5912	2015 May 29 18:00	213.55.104.235	U4	course view	Quality Managment	P0	6
15	IENG 5914	2015 June 3 20:41	213.55.104.235	U4	Logout			5
16	IENG 5915	2015 June 3 19:25	213.55.104.235	U4	resource view	Sigma For Financial Services New York, NY	P6	5
17	IENG 5912	2015 June 3 19:00	213.55.104.235	U4	resource view	Manufacturing	P1	5
18	IENG 5912	2015 June 3 18:12	213.55.104.235	U4	course view	Quality Managment	P0	5
19	IENG 5915	2015 May 29 22:00	10.128.130.105	U2	logout			1
20	IENG 5916	2015 May 29 21:49	10.128.130.105	U2	resource view	Executive Guide to Improvement and Change,	P9	1
21	IENG 5916	2015 May 29 20:58	10.128.130.105	U2	resource view	Quality Standards, Milwaukee, Wisconsin:	P7	1
22	IENG 5915	2015 May 29 20:29	10.128.130.105	U2	resource view	Ch 2. Quality Measurement	P2	1
23	IENG 5915	2015 May 29 19:35	10.128.130.105	U2	resource view	Sigma For Financial Services New York, NY	p3	1
24	IENG 5912	2015 May 29 18:01	10.128.130.105	U2	resource view	Ch 3. Quality Measurement	P3	1
25	IENG 5912	2015 May 29 18:00	10.128.130.105	U2	course view	Quality Managment	P0	1
26	IENG 5915	2015 May 29 23:55	213.55.104.235	U4	logout			3

Table 0-4: Transactions after user's Session identification step

3.5.4 Page Time Calculation

In this step, as shown in Figure 3-2, we calculate visiting web based e-learning system page time for each page (Resources) which we define as the time difference between consecutive page requests for a user in the same session.

***Input:T*;**
***Output:T* with page visiting time**

1:Sort *T* by *UID* and *SID*
2: **for** all unique UID_i and SID_i pair do
3:**For** all T_j with UID_i and SID_i do
If(j>= 1 and j < maximum click stream number)
4: visiting time ($ResID_j$) = $TIME_{j+1}$ - $TIME_j$
5: **End if**
6: Jump the loop
7: **End for**
8: **End for**

Figure 0-2: Algorithm for calculating visiting page times

Where T: transaction, UID: User Identification, SID: Session Identification, ResID: Resource Identification number: $TIME_j$: time at which a user visit a page, J is integer as counter.

However, the raw time durations may not be an appropriate measure for the interest of a Learner in that page. This is because a variety of factors, such as structure of the page, content length of the page, and the speed of network connection, as well as the Learner's interests in a particular item, may affect the amount of time spent on that page.

User	Action	Pages	SID	Time(Min	Normalized Time
U1	resource view	P8	1	**51**	**6**
U1	resource view	P7	1	85	10
U1	resource view	P2	1	71	9
U1	course view	P0	1	7	2
U1	course view		1	0	1
U2	resource view	P9	1	11	2
U2	resource view	P7	1	51	6
U2	resource view	P2	1	29	4
U2	resource view	P3	1	54	6
U2	resource view	P3	1	**40**	**5**

Table 0-5: Visiting and normalized page times for Moodle e-learning system pages

Appropriate normalization of the time can play an essential role in correcting for these factors [69]. Since we want to capture the relative importance of a page (resource) to a particular user relative to other pages visited by that user in the same session, we normalized the visiting times across the visiting times of pages in the same session S_k:

$$norm_{ResIDi} = [(TIME_i - min(T(S_k)))/(max(T(S_k))-min(T(S_k)))] * (max(norm)-min(norm)) + min(norm) \quad (3.1)$$

Where *max(T(S_k))* and min*(T(S_k))* are the maximum and minimum visiting page times respectively taken across the visiting page times spent by a user in the same session S_k. max(norm) and *min(norm)* are the maximum and minimum values for normalized page times respectively. For evaluating the effect of the normalization values, we try five different maximum values: 1,2,3,5 and 10. The minimum value of normalized time is set to 1 in order to differentiate the existence or non-existence of a page in a session. Table 3-5 shows the normalized time of pages in the sample log file. The maximum value for the normalized times in this case is 10.At the end of this step, the log entries in the data sets are converted to the form:

$$T_i = (UID_i\ ,\ norm_{\ ResIDi}, Action_i, ResID_i, IP_i, SID_i)$$

Where T_i : Every user transaction in T, UID_i is the User Identification Number. For every identical user name, we simply assign a unique User Identification Number for each unique user. Thus, some of the user requests in *T* have the same User Identification Number, $Course_i$ represent course identification number, $norm_{\ ResIDi}$ represents The normalized visiting times across the visiting times of pages in the same session S_k, $Action_i$ represents what kind of action a user doing when he visits the page (update or viewing etc), $ResID_i$ stands for e-learning resource Identification number.

3.5.5 Data Cleaning

In this step filtering methods are applied in order to remove irrelevant log entries. A user's request to view, update, delete a particular page (Resource)often results in several log entries while students or teachers are doing their day to day activities for teaching and learning activities. Since the main intent

of Moodle's log file mining is to extract a pattern from the user's behavior, it does not make sense to include file requests that the users did common activities that perform in each transaction like login and log out activities entry.

The *ResID* is returned by the server as a response to the user request. *ResID* value of Errors means a failure for log in to Moodle and for *ResID* value 'log in' and 'logout' entries are also removed since they are common activities that each user performs in their day to day activity for accessing learning resource. The next step is removing log entries of users who have administrator privileges because they left entries in log file for their course adding, removing, user enrollment etc. We also removed entries for forum, submission of assignment etc. since they are not our focus on them. And their importance for determine learner behavior is insignificant.

In the last step of the data cleaning, the Dashboard that provides links to guide users to the content pages is removed. For the datasets, we identify the dashboard as *ResID* values are $\mathbf{P_0}$. Since our objective in this study is to recommend resources that contain a portion of the resource content that the web based e-learning system provides not the dashboard which uses as a user guide for the e-learning resources. The system should recommend pages (resources) that the user may find interesting.

3.6 Evaluation

Evaluating recommender systems is innately difficult and it has been approached in many, often dissimilar ways [61]. In addition to algorithms tending to perform better or worse depending on data set characteristics, the goals and purposes for which recommender systems are developed and evaluated differ, meaning that no one evaluation technique is going to suit them all [71]

In general, perspectives to evaluating recommender systems can be divided into system-centric and user-centric, although the two approaches can also be combined. System-centric approaches evaluate recommenders against a pre-built or pre-collected dataset of user preferences using such quality measures as precision and recall without users interacting with the system during the test; the user opinions on the items have been gathered beforehand and testing is done against values that are withheld from the dataset available to the recommender. [70]

In contrast, user-centric approaches have users interact with a running recommender—or recommenders if two or more variations are being compared—and the data is collected during or based on the interaction; users are either asked, e.g. through interviews or surveys, or their behavior is observed during the use or their interactions are recorded and then analyzed, [43].

In this research we use system centric evaluation so as to evaluate the performance of the recommender system in using the user's interest modeling, and grouping of learner based on their interest

Below we discuss System-centric evaluation in recommender systems in further detail but leave user-centric measures outside of the discussion.

3.6.1 Steps in Evaluation

There are three steps to successfully measuring the performance of recommender system [73].

1. Identify the high-level goals of the system
2. Identify the specific tasks towards those goals that the system will enable

3. Identify system-level metrics and perform system evaluation

Identify High Level Goals

Before measuring the performance of an information recommender system, it must be determined exactly the goals of the system as well as the exact tasks the users will be performing with the system [73].

Information recommender systems are not valuable by themselves. Rather they are valuable because they help people to perform tasks better than those people could without assistance from the recommender system. Therefore, at the highest level, the goal of e-learning recommender system is to recommend e-learning material based on user interest.

If people are currently engaged in the e-learning activity, then it is not the part of the recommender system builder to justify that activity. The valuable contribution that can be made is to improve significantly the efficiency, quality, or speed of such a learning activity [73].

Identify Specific Tasks

Having specified what the high-level goals are, the next step is to specify specific tasks that the users will perform, aided by the recommender system. These tasks will describe explicitly the nature of interaction between the user and the system [73]. The choice of the appropriate metric to use in evaluating a system will depend on the specific activities that are identified for the system.

Performing System-Level Analysis

System-level evaluation is performed in cases where researchers can identify measurable indicators of the system that will significantly correlate with the effectiveness of a system independent of the user interaction. Researchers who use system-level evaluation assume that differences in the given indicators will result in better task performance given any reasonable user interface [73]. System-level evaluation has been the most prevalent form of evaluation in recommender system, because it offers inexpensive, easily repeatable analysis [73] The data are collected from users once, and then many different systems can be evaluated on the collected data without further expensive user sessions.

3.6.2 Evaluation Metrics

For measuring the performance of recommender algorithms measures originating from statistics, machine learning and information retrieval are used. Given the goal of e-learning recommender systems, helping learner more effectively identify the content they want, the utility of the system is defined to include two dimensions: coverage and accuracy [75].

Coverage is a measure of the percentage of items for which a recommendation system can provide recommendations [72]. A low coverage value indicates that the user must either forego a large number of items, or evaluate them based on criteria other than recommendations. A high coverage value indicates that the recommendation system provides assistance in selecting among most of the items. Coverage is usually computed as a percentage of items for which the system was able to provide a recommendation [76].

$$Coverage = \frac{\text{number of items recommended}}{\text{total number o f items}} \qquad \text{Eq 3.1}$$

Coverage may be appropriate for certain ranking-based tasks However, for tasks in which a user can request a prediction for any item in the database, not be able to produce a prediction is generally inappropriate [73]. In any case, coverage should be reported, and system accuracy should only be compared on items for which both systems can produce predictions [73].

Accuracy is a measure of the correctness of the recommendations generated by the system [74]. It is the fraction of correct recommendations to total possible recommendations. The metrics for evaluating the accuracy of a prediction algorithm can be divided into two main categories: statistical accuracy metrics and decision-support metrics [74]. Statistical accuracy metrics evaluate the accuracy of a predictor by comparing predicted values with user-provided values. Decision-support accuracy measures how well predictions help users to select high-quality items.

Statistical accuracy metrics: - statistical accuracy measures the closeness between the numerical recommendations provided by the system and the numerical ratings entered by the user for the same items [75] Common metrics used include:

Mean Absolute Error (MAE) and Related Measures

Mean absolute error measures the average absolute deviation between a predicted rating and the user's true rating. It has been used to evaluate recommender systems in several cases [74]

$$\text{MAE } |\overline{E}| = \frac{\sum_{j=1}^{N} |p_i - r_i|}{N} \qquad \text{Eq 3.2}$$

where $|p_i-r_i|$ is the absolute error of each component and N is total number of items we produce recommendations for.

MAE may be less appropriate for tasks where a ranked result is returned to the user, who then only views items at the top of the ranking. However, he maintained that the mean absolute error metric should not be discounted as a potential metric for ranking-based tasks. Intuitively, it seems clear that as mean absolute error decreases, all other metrics must eventually show improvements. There are two other advantages to mean absolute error. First, the mechanics of the computation are simple and easily recognized by all. Second, mean absolute error has well studied statistical properties that provide a means for testing the significance of difference between the mean absolute errors of two systems.

Two related measures are the mean squared error and the root mean squared error. These variations square the error before summing it. The result is more emphasis on large errors. For example, an error of one point increases the sum by one, but an error of two points increases the sum by four.

Precision and Recall

Precision and recall are the most popular metrics for evaluating information retrieval systems [77].Precision and recall are computed from a contingency table, such as the one shown in Table 3-6 below. The item set must be separated into two classes – relevant or not relevant. Recall measures the ability of the system to present all relevant documents. Precision, on the other hand, measures the ability of the system to withhold non-relevant documents [76].

	Selected	Not Selected	Total
Relevant	Nrs	Nrn	Nr
Irrelevant	Nis	Nin	Ni
Total	Ns	Nn	N

Table 0-6: Contingency table showing the categorization of items in the document set

If an item meets an information need, then it is a successful recommendation (i.e. relevant). If we measure how likely the system is to return relevant documents, then we are measuring how likely the system meets the user's information need.

Likewise, we need to separate the item set into the set that was returned to the user (selected), and the set that was not. We assume that the user will consider all items that are retrieved. Precision is defined as the ratio of relevant documents selected to number of documents selected, shown in Equation 3.3.

$$P = \frac{N_{rs}}{N_s} \qquad \text{Eq 3.3}$$

Precision represents the probability that a selected document is relevant. Recall is defined as the ratio of relevant documents selected to total number of relevant documents available. Recall represents the probability that a relevant document will be selected.

$$P = \frac{N_{rs}}{N_s} \qquad \text{Eq 3.4}$$

Precision and recall depend on the separation of relevant and non-relevant items. The definition of "relevance" and the proper way to compute has been a significant source of argument within the field of information retrieval [76]. Most information retrieval evaluation has focused on an objective version of relevance, where relevance is defined with respect to the query, and is independent of user. Teams of experts can compare documents to queries and determine which documents are relevant to which queries. However, objective relevance makes no sense in recommender system. Recommender system is recommending items based the likelihood that they will meet a specific user's taste or interest. That user is the only person who can determine if an item meets his taste requirements. Thus, relevance in recommender system is inherently subjective. To compute precision on a non-binary scale, a rating threshold must be selected such that items rated higher than the threshold are relevant and items rated below the threshold are not relevant. Because of variance in both rating distributions and information need, the appropriate threshold may be different for each individual user. Furthermore, it may be extremely hard to determine. Recall is even more impractical to measure in recommender system than it is in IR systems. To truly compute recall, we must determine how many relevant items are contained in the entire database. Since the user is only person who can determine relevance, we must have every user examine every item in the database. While this problem existing in information retrieval systems, IR researchers approximated the value by taking the union of relevant documents that all users found. This was possible because relevance was defined to be objective and global, which is not the case with recommender system. To compute recall in recommender system, a large enough set must be selected Precision and recall can be linked to probabilities that directly affect the user. This makes them more understandable to users and managers than metrics such as mean absolute error. Users can more intuitively comprehend the meaning of a 10% difference in precision than they can a 0.5-point

difference in mean absolute error [78] As a result, precision and recall may be more appropriate than other metrics for arguing cost benefits.

Another weakness of precision and recall is that there is not a single number for comparisons of systems. Rather, two numbers must be presented to describe the performance of the system. In addition, it has been observed that precision and recall are inversely related and are dependent on the length of the result list returned to the user. If more documents are returned, then the recall increases and precision decreases. Therefore, if the information filtering system doesn't always return a fixed number of documents, we must provide a vector of precision/recall pairs to fully describe the performance of the system. While such an analysis may provide a good amount of information about a single system, it makes comparison of more than two systems complicated, tedious, and variable between different observers [73].

It should be noted that in certain cases, the task for which we are evaluating is not concerned with recall, only precision. This is true for filtering many entertainment domains. For example, a person who is looking to find a video to rent for the weekend doesn't need to see all the videos in existence that he/she will like. Rather he/she cares primarily that the movies recommended accurately match his/her tastes. In this case, it becomes easier to compare systems, although we must still be concerned about how different sizes of retrieval sets will affect the precision [73].

4 Chapter Four

Learner Interest Modeling

The process of a recommendation system was shown in Figure 3.1 in Chapter 3.The first two components of this process; Data collection and Data preparation and cleaning are implemented using the methods which were detailed in Chapter 3. For the last component, Prediction, of the recommendation process we use Learner interest model.

The Learner Interest Model (LIM) uses only the visiting time and visiting frequencies of pages without considering the access order of page requests in learner session. The discovered patterns do not depend on any personal data about the learners. Each section in this chapter presents a model .First, brief background information of the model is given and next, the details of the model are discussed.

4.1 Learner Interest Model

Making a recommendation requires predicting what is of interest to a learner at a specific time. Even the same learner may have different desires at different times. It is important to extract the aggregate interest of a user from his/her navigational path through the web based e-learning system in a session. This chapter concentrates on the discovery and modeling of the user's aggregate interest in a session.

Background

The LIM relies on the premise that the normalized visiting time of a page in web based e-learning system is an indicator of the learner's interest in that page. As the user stays long proportion time on a page with in a session, the page can show interest of the learner. The normalized proportion of times

spent in a set of pages requested by the learner within a single session forms the aggregate interest of that learner in that session. We first partition user sessions, after the time is normalized, in to clusters such that only sessions which represent similar aggregate interest of learner are placed in the same cluster. The key idea behind this work is that learner sessions can be clustered according to the similar amount of time that is spent on common web based e-learning system's pages among learner sessions. In particular, we model learner sessions in log data as being generated in the following manner:

1. When a learner arrives to the web based e-learning system, its current session is assigned to one of the clusters;
2. The behavior of that learner in this session, in terms of visiting time, is then generated from a Poisson model of visiting times of that cluster.

Since we do not have the actual cluster assignments, we use a standard learning algorithm, the Expectation-Maximization (EM) algorithm [58], to learn the cluster assignments of sessions as well as the parameters of each Poisson distribution. The resulting clusters consist of sessions in which users have similar interests and each cluster has its own parameters representing these interests.

The next page request of an active user is predicted using parameters of the cluster to which the active user is assigned. The model produces a set of recommendations based on this prediction. The detailed model is given in this chapter.

Model-Based Cluster Analysis

Model-based clustering methods optimize the fit between the given data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions [59].

Given a data set of K observations $D=\{x_1,x_2,\ldots x_k\}$,every observation x_i,($i \in [1, k]$) is generated according to a probability distribution defined by a set of parameters, denoted by Θ. The probability distribution consists of a mixture model of components $c_j \in C$, $c_j=\{c_1,c_2,\ldots c_G\}$.The parameters of each component Θ_g, is a disjoint subset of Θ where Θ_g ($g \in [1..G]$) is a vector specifying the probability distribution function (pdf) of the g^{th} component[60]. An observation, x_i, is created by first selecting a mixture component according to the mixture weights (or cluster prior probabilities), $P(c_g|\Theta)=\tau_g$, where $\sum_{g=1}^{G} Tg = 1$, then having this selected mixture component generate τ_g an observation according to it shown parameters, with distribution $P(x_i|c_g; \Theta_g)$ Thus, the likelihood of a data point, x_i, can be characterized with a sum of total probabilities over all mixture components[59]:

$$p(\mathbf{x}_i|\mathbf{\Theta}) = \sum_{g=1}^{G} p(c_g|\mathbf{\Theta}) p(\mathbf{x}_i|c_g, \mathbf{\Theta}_g) = \sum_{g=1}^{G} \tau_g p(\mathbf{x}_i|c_g, \mathbf{\Theta}_g) \qquad 4.1$$

Statisticians refer to such a model as *mixture model with G components* [52].Thus, the model-based clustering problem consists of finding the model, i.e. the model structure and parameters for that structure that best fit the data. The parameters are chosen in two ways [52]. The *maximum likelihood* (ML estimation) approach maximizes:

$$\ell_{ML}(\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_G; \tau_1, \ldots, \tau_G|D) = \prod_{i=1}^{K} \sum_{g=1}^{G} \tau_g p(\mathbf{x}_i|c_g, \mathbf{\Theta}_g) \qquad (4.2)$$

The second approach, Maximum Aposteriori (MAP estimation), maximizes the *posterior probability* of Θ given the data:

$$\ell_{MAP}(\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_G; \tau_1, ..., \tau_G|D) = \prod_{i=1}^{K} \sum_{g=1}^{G} \frac{\tau_g p(\mathbf{x}_i|c_g, \boldsymbol{\Theta}_g) p(\boldsymbol{\Theta})}{p(D)} \tag{4.3}$$

The term p(D) can be ignored in Equation 5.3,since it is not a function of Θ.

In practice, the log of these expressions is often used. Thus, the log likelihood of Equation 5.2 and 5.3 are respectively:

$$L(\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_G; \tau_1, ..., \tau_G|D) = \sum_{i=1}^{K} \ln \left(\sum_{g=1}^{G} \tau_g p(\mathbf{x}_i|c_g, \boldsymbol{\Theta}_g) \right) \tag{4.4}$$

$$L(\boldsymbol{\Theta}_1, ..., \boldsymbol{\Theta}_G; \tau_1, ..., \tau_G|D) = \sum_{i=1}^{K} \ln \left(\sum_{g=1}^{G} \tau_g p(\mathbf{x}_i|c_g, \boldsymbol{\Theta}_g) \right) + \ln p(\boldsymbol{\Theta}) \tag{4.5}$$

The set of parameters of the model (Θ) include mixture weights representing cluster prior probabilities (τ_g), which indicate the probability of selecting different mixture components and the set of the parameters of the probability distribution assumed for the data:

$$\Theta = \{\Theta_1, ..., \Theta_G, \tau_1, ..., \tau_G\}, \ \sum_{g=1}^{G} \tau_g = 1$$

(4.6)

EM Algorithm for Clustering

The model parameters can be trained using the Expectation Maximization (EM) algorithm. The EM algorithm is a very general iterative algorithm for parameter estimation by maximum likelihood when some of the random variables involved are not observed (i.e., considered missing or incomplete).In the Expectation step (E-step), the values of the unobserved variables are essentially "filled in", where the filling-in is achieved by calculating the probability of the missing variables, given the observed variables and the current values of parameters. In the Maximization step (M-step), the parameters are adjusted based on the filled-in variables [59].

Let D =$\{x_1,\ldots x_k\}$,be a set of K observed variables, and H = $\{z_1,\ldots z_k\}$,represent a set of K values of hidden variables Z, such that each z_i is in the form of $z_i = \{z_{1i},\ldots z_{G\,i}\}$,and corresponds to a data point x_i. It can be assumed that Z is discrete and represents the class (or cluster) labels for the data with the following possible values:

$$z_{ji} = \begin{cases} 1 & \text{if } x_i \text{ belongs to cluster } j; \\ 0 & \text{otherwise.} \end{cases}$$

If Z could be observed, then the ML estimation problem would be based on the maximization of the quantity:

$$L_c(\Theta; D, H) \triangleq \ln p(D, H|\Theta) \tag{4.7}$$

In the presence of missing data, we calculate conditional expectation of the complete data likelihood given the observed data and the current parameter estimate as follows:

$$Q(\Theta, \Theta') = E\left[L_c(D, H|\Theta)|D, \Theta'\right] \tag{4.8}$$

where the term $L_c(D, H|\Theta)$ is:

$$L_c(D, H|\Theta) = \sum_{i=1}^{K} \ln p(\mathbf{x}_i, \mathbf{z}_i|\Theta) \tag{4.9}$$

Equation 4.8 involves Θ, which is the parameter of the complete likelihood and Θ', which is the parameter of the conditional distribution of complete data.

The Q-function in Equation 4.8 can be expanded as follows:

$$
\begin{aligned}
E\left[L_c(D,H|\Theta)|D,\Theta'\right] &= E\left[\sum_{i=1}^{K}\ln p(\mathbf{x}_i,\mathbf{z}_i|\Theta)|D,\Theta'\right] \\
&= \sum_{l=1}^{G}\sum_{i=1}^{K}\ln p(\mathbf{x}_i,\mathbf{z}_i|\Theta)\prod_{j=1}^{K}p(z_{lj}|\mathbf{x}_j,\Theta') \\
&= \sum_{i=1}^{K}\sum_{l=1}^{G}(\ln p(\mathbf{x}_i,\mathbf{z}_i|\Theta)p(z_{li}|\mathbf{x}_i,\Theta'))\prod_{j\neq i}\sum_{l=1}^{G}p(z_{lj}|\mathbf{x}_j,\Theta') \\
&= \sum_{i=1}^{K}\sum_{l=1}^{G}\ln p(\mathbf{x}_i,\mathbf{z}_i|\Theta)p(z_{li}|\mathbf{x}_i,\Theta') \\
&= \sum_{i=1}^{K}\sum_{\mathbf{z}_i}\ln p(\mathbf{x}_i,\mathbf{z}_i|\Theta)p(\mathbf{z}_i|\mathbf{x}_i,\Theta') \\
&= \sum_{i=1}^{K}\sum_{\mathbf{z}_i}p(\mathbf{z}_i|\mathbf{x}_i,\Theta')\ln\left[p(\mathbf{x}_i|\mathbf{z}_i,\Theta)p(\mathbf{z}_i|\Theta)\right] \\
&= \sum_{i=1}^{K}\sum_{\mathbf{z}_i}p(\mathbf{z}_i|\mathbf{x}_i,\Theta')\left[\ln p(\mathbf{x}_i|\mathbf{z}_i,\Theta)+\ln p(\mathbf{z}_i|\Theta)\right]
\end{aligned}
\tag{4.10}
$$

At each EM iteration, the Q-function is maximized with respect to the parameters Θ using the current parameters Θ'. At the end of each iteration, a set of new optimal parameters Θ becomes the current parameters Θ' for the next iteration. Given these steps, the EM algorithm can be implemented as follows:

1. Choose an initial estimate for parameter set Θ'(0), and set n=0.

2. **(E)xpectation Step:** For n ,compute Q (Θ,Θ'(n)) using Equation 4.8.

3. **(M)aximization Step:** Replace the current estimate (n) with the new estimate Θ'(n+1)where,

$$\Theta'(n+1) = argmax_{\Theta} Q(\Theta, \Theta'(n))$$

4. Set n = n+1 and iterate steps 2 and 3 until convergence.

By iteratively applying the E-step and M-step, the parameters Θ will converge to at least a local maximum of the log likelihood function.

S	PAGES	NORMS
1	{p_1, p_3, p_6, p_5, p_{10}}	{1, 0,8,0, 3,10,0,0, 0,1}
2	{p_4, p_9, p_6, p_{10}, p_7}	{0, 0,0,2, 0,1,10,0, 8,10}
3	{p_7, p_6, p_5, p_4, p_{10}, p_9}	{0, 0,0,2, 1,1,9, 0,8, 10}
4	{p_1, p_3, p_6, p_5, p_{10}, p_9, p_2}	{10,10,3, 0,1,6, 0,0, 1, 4}
5	{p_1, p_{10}, p_8, p_2, p_5, p_3, p_6}	{1, 1,7,0, 3,10,0,1, 0,1}
6	{p_9, p_{10}, p_6;p_3, p_2, p_1}	{10,9, 2,0,0, 6,0,0, 1,4}
7	{p_1, p_5, p_6, p_3}	{1, 0,8,0, 4,10,0,0, 0,0}

Table 4-1: A set of user sessions as running example

4.2 Model Implementation

Once the data cleaning and preprocessing tasks described in the Chapter 3 are performed, Web server logs have been converted in to a set of user sessions. User sessions can be clustered according to the similar amount of time spent on common pages.

A sample set of user sessions, for a web based e-learning system with ten pages, P= {p_1,p_2,….,p_{10}} is shown in Table 4-1. PAGES corresponds to a subset of pages in P and NORMS corresponds to the normalized visiting times of pages in P.

Clustering User Sessions in Web based e-learning Log Data

In this section, we first describe the specific mixture model that we use for clustering the learner sessions in Web log data. Next, the update parameters for training the mixture model of Poisson distributions with the Expectation Maximization algorithm are given. We use a model-based technique to group the learner sessions according to the interests of learner in each session. We assume the data to be generated in the following fashion:

1. When a learner arrives at Web based e-leaning system, its session is assigned to one of *G* clusters with some probability.

2. Given that a learner's session is in a cluster, its next request in that session is generated according to a probability distribution specific to that cluster.

Since it is assumed that the data are produced by a mixture model, every learner session is generated according to the probability distribution defined by a subset of model parameters, denoted Θ_g. Let X $=\{x_1,x_2,\ldots x_k\}$, be a set of K learner sessions and C be a discrete valued variable taking values$c_1,..c_G$, which corresponds to an unknown cluster assignment of a user session. Then the mixture model for a user session is:

$$p(\mathbf{X}=\mathbf{x}_i|\mathbf{\Theta}) = \sum_{g=1}^{G} p(\mathbf{C}=c_g|\mathbf{\Theta})p(\mathbf{X}=\mathbf{x}_i|\mathbf{C}=c_g,\mathbf{\Theta}_g)$$
$$= \sum_{g=1}^{G} \tau_g p(\mathbf{X}=\mathbf{x}_i|c_g,\mathbf{\Theta}_g) \quad (4.11)$$

Where τ_g is the probability of selecting cluster c_g.

A user session, x_i ,is considered to be an n-dimensional vector of visiting page times,(x_{i1}, x_{i2},……x_{in}), where x_{ij} corresponds to $norm_{pj}$ in NORMS field of a user session; each p_j is a page in the set of pages(in a given web-based e-learning system) P={p_1,p_2,….p_{10}}. Each page in the set of pages P corresponds to a dimension in the model. Then-dimensional vector represents the aggregate interest of the user.

In our case, the mixture model can be regarded as a distribution in which the class labels are missing. There is still a problem of how to estimate the probabilities. One of the key ideas to handle this problem is to impose a structure on the underlying distribution, for example by assuming the independence of dimensions:

$$p(\mathbf{x}_i) = \prod_{j=1}^{n} p_j(x_{ij}) \quad (4.12)$$

Since a user session is an n-dimensional vector of normalized visiting times, we can easily adapt this assumption to our model. Even the order of visiting pages may be different in two user sessions, each session can be represented by the equal vectors if the normalized page times corresponding to the same

page in each session are equal.

To illustrate the independence assumption for our model, consider the user sessions 1, 4 and 7 in Table 4-1.The order of page requests in sessions 1 and 7 are different. However, the aggregate interests of the sessions are very similar, because the normalized page times of each page are similar. Although the first 5 pages in sessions 1 and 4 are requested in the same order, the aggregate interests of these sessions are not similar. According to our clustering criteria, sessions 1 and 7 would be in the same cluster, where as session 4 would be in a different cluster. Thus, the value of the m^{th} dimension of a session, where $m \in [1 \ldots n]$, is independent of the values in the preceding dimensions.

The independence assumption enables us to use *n* separate probability distributions to model each dimension of a user session. To model this data, we assume that the data at each dimension have been generated by a mixture of Poisson distributions. A random variable X has a *Poisson distribution* with parameter *m* if for some $m > 0$ [60]:

$$p(X = k) = \frac{m^k e^{-m}}{k!} \qquad k = 0, 1, \ldots \tag{4.13}$$

Figure 4-1 shows the shape of posisson distribution as the parameter,m changes. [60]

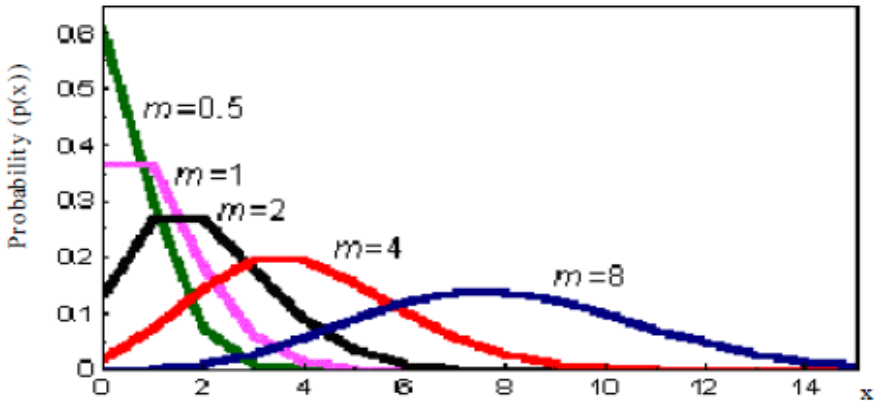

Figure 4-1: shape of the Poisson distribution for different parameters, m. (Source:[60])

In Figure 4-1 ([60],p. 73) described *"As m increases, the shape of the Poisson distribution begins to resemble a bell shaped distribution"*. The Poisson model can be used to model the rate at which individual events occur [63], for example the rate at which a user session has the value1for a particular page. To confirm our assumption, that the data in each dimension have been generated by a Poisson distribution, the histogram of the occurrence of each of the ten possible values at each dimension has been plotted. Most of the histograms verify our assumption. Figure 4.2 presents one of these histograms. As can be seen, the histogram has the shape of the Poisson distribution with a low parameter *m*.

According to the independence assumption, a user session x_i is generated in a cluster g by a Poisson model as follows [54]:

$$p(\mathbf{x}_i|c_g, \boldsymbol{\Theta}_g) = \prod_{j=1}^{n} \frac{(\theta_{gj})^{x_{ij}}\ e^{-\theta_{gj}}}{x_{ij}!} \tag{4.14}$$

Where θ_{gj} is the parameter of the Poisson distribution for a dimension j in cluster *g*

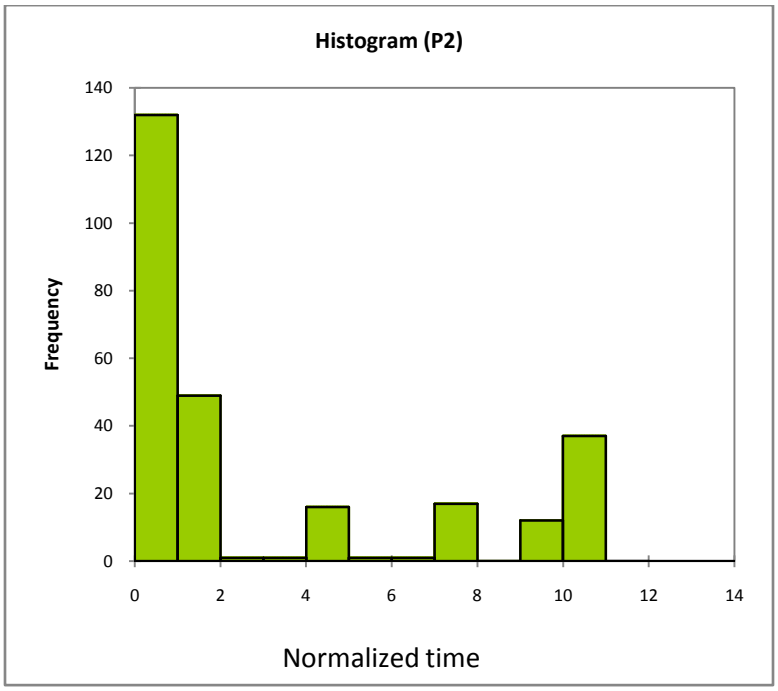

Figure 4-2: Mekele university Moodle log file

By combining Equation 4.11 and Equation 4.14 we obtain:

$$p(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_{g=1}^{G} \tau_g \left(\prod_{j=1}^{n} \frac{(\theta_{gj})^{x_{ij}}\ e^{-\theta_{gj}}}{x_{ij}!} \right) \tag{4.16}$$

Where $\theta_{g\,j}$ ($g \in [1 \ldots G]$, $j \in [1 \ldots n]$) is the Poisson parameter of cluster c_g at dimension *j*.

For the sample set of user sessions in Table 4-1, there are 10 Poisson parameters for each cluster, where the number of unique pages is 10 in that data set.

The model parameters to be learned are then:

$$\Theta = \{\Theta_1, ..., \Theta_G, \tau_1, ..., \tau_G\},\ \Theta_g = (\theta_{g1}, ..., \theta_{gn}),\ \sum_{g=1}^{G} \tau_g = 1 \tag{4.17}$$

Learning the Model Parameters

We can train the model parameters of the mixture model, developed in the previous subsection, using EM algorithm where the conditional independence assumption is enforced during Maximization step. The learning algorithm is carried out for each component of the model. There are several reasons for using the EM algorithm:

- We want to represent the behavior of the user in one session using Poisson distribution;
- Its performance is linear to the number of sessions;
- It is robust to noisy data;
- It provides a cluster membership probability per session;
- It can handle high dimensionality;

In order to implement the EM algorithm we should pick the number of clusters (G), an initial starting point ($\Theta'(0)$),a convergence criteria and prior probabilities for Θ in case of MAP estimate of model parameters[54]. To determine the number of clusters, we run the algorithm with several numbers of clusters. We initialize the parameters of our components, Θ_g,($g \in [1....G]$) by estimating the Poisson parameters for a single component model and then randomly perturbing the parameter values by a small

amount to obtain G sets of parameters. We determine the convergence criteria such that the algorithm converges when the log likelihoods of two consecutive iterations on the training data differ less than 0.001%.There is a trade-off between the estimation accuracy of parameters and the number of iterations. With a smaller value the number of iterations required for convergence will increase so that the algorithm converges in a longer period of time. If it is greater than the selected value, then the estimation for the parameters would be less precise. Finally, to assign prior probabilities to Θ for MAP estimate we use a prior distribution for the Poisson distribution.

ML Estimate of Model Parameters One approach to learning parameters from data is to find those parameter values that maximize the likelihood of data:

$$\Theta^{ML} = argmax_{\Theta}\{p(D|\Theta_1, ..., \Theta_G, \tau_1, ..., \tau_G)\} \quad (4.17)$$

Pages	Cluster 1	Cluster2	Cluster3
P1	0.400	6.891	7.526
p2	2.371	1.727	7.605
p3	6.257	1.473	0.079
p4	0.100	1.564	0.579
p5	2.957	2.073	0.474
p6	3.386	8.236	2.237
p7	3.529	1.818	4.868
p8	5.686	0.055	0.632
p9	3.629	1.436	0.395
P10	0.514	0.018	0.632

Table 4-2: Poisson parameters for three clusters

These parameters in Table 4-2 shown are often referred to as a maximum likelihood or ML estimate.

The E-step of ML estimate of parameters involves an update of the conditional probability of missing class labels given the current parameter set Θ'[52]. We define this probability as *cluster-posterior* probability, $P_{ig}(\Theta')$, that the transaction x_i arose from the g^{th} cluster.

$$P_{ig}(\boldsymbol{\Theta}') = \frac{\tau_g p(\mathbf{x}_i|c_g, \boldsymbol{\Theta}'_g)}{\sum_{j=1}^{G} \tau_j p(\mathbf{x}_i|c_j, \boldsymbol{\Theta}'_j)} \tag{4.18}$$

In the M-step, the Q function in Equation 4.10ismaximizedandthisstepconsistsof the update of cluster priors and Poisson parameters:

$$\widehat{\tau}_g = \frac{1}{K} \sum_{i=1}^{K} P_{ig}(\boldsymbol{\Theta}') \tag{4.19}$$

$$\widehat{\theta}_{gm} = \frac{\sum_{i=1}^{K} (P_{ig}(\boldsymbol{\Theta}') x_{im})}{\sum_{i=1}^{K} P_{ig}(\boldsymbol{\Theta}')} \tag{4.20}$$

At the end of the EM algorithm each cluster has its own set of parameters such that:

$$pc_g = \{\tau_g, (\theta_{g1}, ..., \theta_{gn})\}$$

MAP Estimate of Model Parameters One difficulty associated with using the maximum likelihood approach relates to zero probabilities [50]. For example, if there is no request for a page p_i in the data set, then our estimate of the Poisson parameter for that page will be zero. That is, according to our model, the probability of requesting page p_i is zero. To address this difficulty, we can assign prior probabilities to Θ and use the maximum of the posterior distribution over Θ as our estimate for the parameters. Thus, the MAP parameters that correspond to the maximum of posterior distribution of Θ can be found by maximizing the posterior probability of Θ given the data:

$$\Theta^{MAP} = argmax_{\Theta} = \{p(D|\Theta_1, ..., \Theta_G, \tau_1, ..., \tau_G)p(\Theta)\} \quad (4.21)$$

Where the second identity follows by Bayes 'rule and is the prior distrib ution of the model parameters.

To perform MAP estimate of parameters we first need to choose a functional form for the prior p(Θ).The parameter set Θ consists of a set of Poisson parameters and the class weights. An often used prior distribution for Poisson distribution is Gamma distribution with two parameters of α and β[61]. The distribution of selecting a class can be regarded as a multinomial distribution. The conjugate prior

distribution for the multinomial distribution is Dirichlet distribution with the parameter [61].

The choice of the parameters of the conjugate prior distributions is to be determined by one's prior beliefs based on the knowledge of the problem [54]. In general, however, such prior knowledge is difficult to obtain. In the absence of such knowledge one usually uses a "non-informative" prior, typically a uniform prior. In this work several combinations of the parameters are tested.

The E-step of the MAP parameter estimation consists of an update of the conditional probability of missing class labels given the current parameter set Θ' as in Equation 4.17.The Q-function for the log-posterior (MAP) function is defined as:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}') = \sum_{i=1}^{K} \sum_{g=1}^{G} P_{ig}(\boldsymbol{\Theta}') \left[\ln p(\mathbf{x}_i | c_g, \boldsymbol{\Theta}_g) + \ln \tau_g\right] + \ln p(\boldsymbol{\Theta}) \tag{4.22}$$

If we maximize the Q-function with respect to each subset of parameters Θ one can show that the following update rules for mixture weights and Poisson parameters can be inferred for the M-step of the EM algorithm:

$$\widehat{\tau}_g = \frac{\sum_{i=1}^{K} P_{ig}(\boldsymbol{\Theta}') + \gamma_g}{\sum_{j=1}^{G} \left[\sum_{i=1}^{K} P_{ij}(\boldsymbol{\Theta}') + \gamma_j\right]} \tag{4.23}$$

$$\widehat{\theta}_{gm} = \frac{\sum_{i=1}^{K} P_{ig}(\Theta')x_{im} + \alpha_{gm}}{\sum_{i=1}^{K} P_{ig}(\Theta') + \beta_{gm}} \quad (4.24)$$

Where γ_i is the hyper parameter associated with τ_i, $i \in [1.....G]$, α_{im} and β_{im} are the hyper parameters associated with θ_{im}; $i \in [1.....G]$; $m \in [1.....n]$.

The output of EM algorithm with MAP estimates is a set of cluster parameters such that each cluster has its own parameters:

$$pc_g = \{\tau_g, (\theta_{g1}, ..., \theta_{gn})\}$$

For the data set in Table 4-1 we compute in the E-step the cluster posterior probabilities using Equation 4.17.In the M-step we update the model parameters using Equation 4.19 and Equation 4.20.Thus, the parameters in Table 4-2 and the cluster priors are updated in each M-step. The E and M-steps are applied until the convergence criteria are obtained. The output of this algorithm is the set of cluster parameters. For example, {0.3; (0.02, 1.2, ., .)} tells us that a cluster has a prior probability of 0.3 and the Poisson parameter of the first page is 0.02, of the second page is 1.2 and soon. In case of using the MAP estimate, there will be a small difference in the parameters according to the hyper parameters of conjugate priors .The clusters in Table 4-3 are obtained by assigning each session to the cluster that has the highest closest cluster.

Cluster No.	S	PAGES	NORMS
1	1	$\{p_1,p_3,p_6,p_5,p_{10}\}$	{1,0,8,0, 3,10,0,0, 0,1}
	5	$\{p_1,p_{10},p_8,p_2,p_5, p_3,p_6\}$	{1,1,7,0, 3,10,0,1, 0,1}
	7	$\{p_1,p_5,p_6, p_3\}$	{1,0,8,0, 4,10,0,0, 0,0}
2	2	$\{p_4,p_9,p_6, p_{10}, p_7\}$	{0,0,0,2, 0,1,10,0, 8,10}
	3	$\{ p_7,p_6,p_5, p_4, p_{10},p_9\}$	{0,0,0,2, 1,1,9, 0,8,10}
3	4	$\{p_1, p_3, p_6, p_5, p_{10}, p_9, p_2\}$	{10,10,3, 0,1,6, 0,0,1, 4}
	6	$\{p_9;p_{10};p_6;p_3;p_2;p_1\}$	{10,9, 2,0,0, 6,0,0, 1,4}

Table 4-3: Sample clusters built by using EM algorithm

Cluster Profiles

In order to obtain a set of pages for recommending and rank those in this set *recommendation scores* are calculated for every page in each cluster using the Poisson parameters of that cluster. Thus, each cluster has a set of recommendation scores additional to its parameter set created in the previous subsection. We modify the cluster parameters such that each cluster has a recommendation score set, $RS_g = \{ rs_{g1}, \ldots . rs_{gm} \}$ where rs_{gi} , $i \in [1 \ldots n]$ is the recommendation score for page p_i in cluster c_g . The updated cluster parameters are then in the form $pc_g=\{ \tau_g;(\theta_{g1},\ldots..,\theta_{gn});(rs_{g1},\ldots rs_{gn})\}$Those are the only parameters that the system needs in order to produce a set of pages for recommendation.

We use one of the methods for calculating recommendation scores for every page. The recommendation scores are then normalized such that the maximum score has a value of 1. These methods are as follow:

Method 1 : We count the number of requests for every page in each cluster. We define this number as popularity, (f_{gi}); where i ∈ [1....n] and g∈ [1....G]]. For example, if R_{pi} is the total number of page requests for page p_i and R_p is the total request for all pages in a cluster c_g,then the popularity of page p_i in that cluster is $f_{gi}=R_{pi}/R_p$.

In this method we use only the popularity information for recommending pages. The intuition behind this is to recommend pages that are most likely visited in a cluster. The recommendation score for page p_i in active cluster c_g is

$$rs_{gi}=f_{gi}$$

Method 2 : In this method also we use the popularity information (f_{gi}) and the Poisson parameter. We calculate the recommendation scores by multiplying the popularity by the Poisson distribution parameter.

$$rs_{gi}=\theta_{gi} * f_{gi}$$

Recommendation Engine

The real-time component of the model calculates closest cluster for every cluster $c_g \in C = \{c_1,\ldots c_G\}$where s_i is the portion of a session in test set that is used to find the most similar cluster. The active session is assigned to the cluster that has the highest probability. We define this cluster as *Active cluster*. A *recommendation set*, which is the set of predicted pages by the model, is then produced ranking the recommendation scores of the active cluster in descending order. The recommendation set consists of pages which have a recommendation score greater than a threshold ε (or

top N items with the highest recommendation scores where N is a fixed number) in the active cluster and that the user has not yet visited.

5 Chapter Five

Experimental Results

In order to test the web based recommender system using recommendation models discussed in Chapter 4, Log file is cleaned and converted in to user sessions. This chapter discusses the results from several experiments run to test the performance and effectiveness of the models.

5.1 Data Sets

The data set is from Mekele University from Moodle e-learning system's web server over the months of May and August 2015.

The server log is in the form as shown in Table 4-1 format. Data preprocessing and cleansing as stated in Chapter 3 produce results of how long a user spent (in minutes) on each page in each session .Before filtering the data at the last step of the cleaning and preprocessing procedure, 23% of sessions in the log have shown an activity of administrator and course manager who did different administrative activities like generating reports, adding or deleting course materials etc. and since those do not represent learner behavior, they are removed. After cleaning the data set, the number of sessions is decreased significantly in these logs. Table 5-1 shows the number of remaining Pages and the number of sessions for each data set.

SN	Types of Actions	No of course	Number of Pages	Number of session	Total of Action
1	1(only view)	1	10	267	1563

Table 5-1: Characteristics of cleaned log data set

Randomly a total of 5%, 10% and15% from the clustered data are selected. Random selection is made for each percentage from each cluster as the test set, and the remaining part as the training set. And the selected test data is deleted from the clustered data.

Figure 5.1 shows the interface of the recommendation module for testing the test data and lists of functions of components of prototype recommender system's interface.

Pages: Represent the pages that we considered in this study

Time(min): Data entry field that show for how much normalized time each user spent on the page

No. of Recommended Paged: Data entry field that show how many unique page do we need to be recommended

Popularity and Popularly and Poisson Dist: are option the help us to select the recommendation score matrix mentioned as Method 1 and Method 2 respectively in chapter 4.

Recommended Pages: Result display box that lists the recommended page

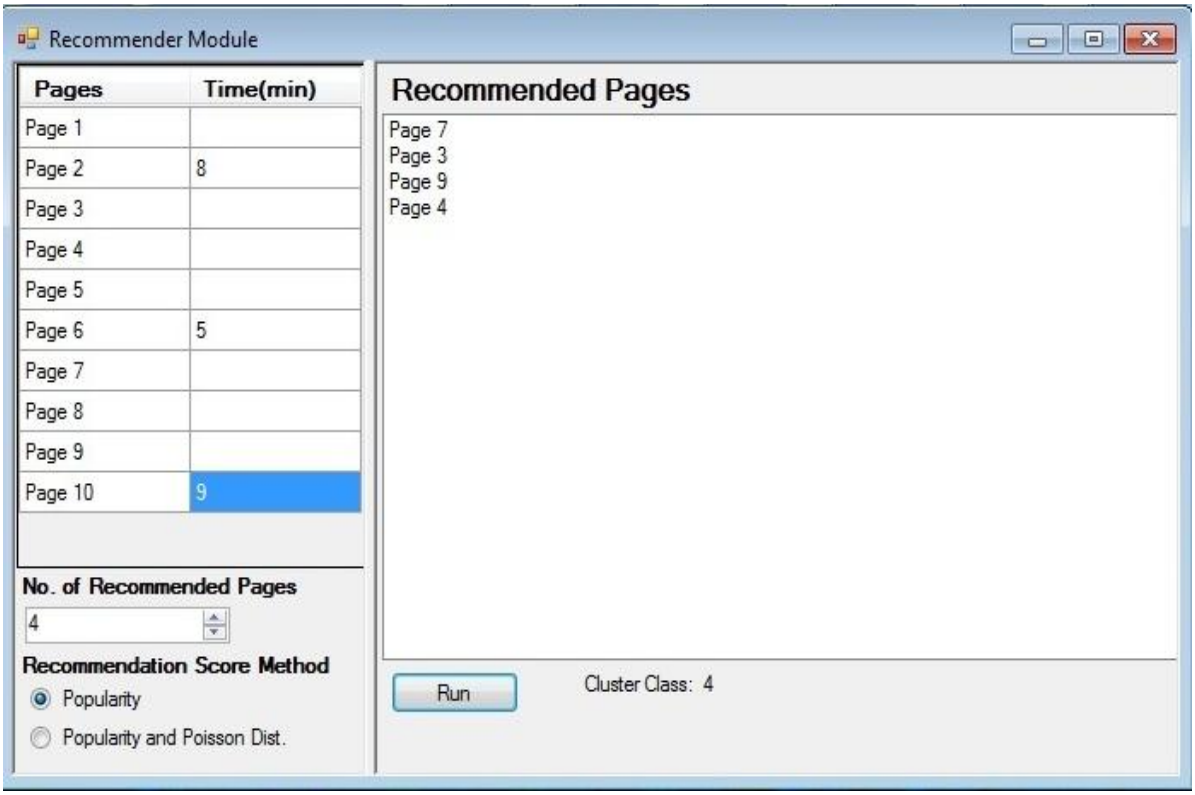

Figure 5-1: Interface for recommendation Module

5.2 Evaluation Metrics

The main focus in evaluating the performance of the models is to determine the extent to which the recommended pages match the actual user session. We define the following metric to evaluate the model:

Precision: For each session S_i in the test set we select the first *w* requests in S_i chronologically. These *w* requests are used to calculate the active cluster and produce the recommendation set. The recommendation set contains all the pages that have a recommendation score greater than threshold; the first N pages and that are not in the first *w* requests .We denote this set as PS(*w*,N) and the number of pages in this set that match with the remaining part of active session as *m*. Then the precision for a

session is defined as:

$$precision(S) = \frac{m}{|PS(w, \text{N}}$$

(5.1)

5.3 Results of the Learner Interest Model

To evaluate the Learner Interest Model (LIM) proposed in Chapter 4, we run several experiments with different sets of initial parameters for EM algorithm.

In our experiments, we tried different values for the threshold, the first N pages of recommendation scores ranging from 1to (10-|*w*|).If the threshold, N is small, then fewer recommendations are produced. If it is large, then irrelevant pages are recommended with a low recommendation score. The maximum number of pages a learner access in a session is 7.

Recommended No of page,N	No of the first given page,*w*	**Precision**	
		Method1	Method2
6	1	15.2	15.7
5	2	13.3	22.3
4	3	33.2	38.4
3	4	21.5	11.4
2	5	16.3	14.6

Table 5-2:Avg. Precision in (%) of Test data of 5%.Visiting time is normalized between 1 and 10

Recommended No of page,N	No of the first given page,*w*	**Precision**	
		Method1	Method2
6	1	17.2	14.7
5	2	19.3	24.0
4	3	31.2	39.2
3	4	13.4	19.6
2	5	10.7	15.6

Table 5-3:Precision in (%) of Test data of 10%.Visiting time is normalized between 1 and 10

Recommended No of page, N	No of the first given page, *w*	**Precision**	
		Method1	Method2
6	1	18.7	23.7
5	2	11.5	25.0
4	3	46.3	51.4
3	4	14.3	22.7
2	5	18.8	26.5

Table 5-4:Avg.Precision in (%) of Test data of 15%.Visiting time is normalized between 1 and 10

Recommended No. of page, N	No of the first given page, *w*	**Precision**	
		Method1	Method2
6	1	10.2	8.7
5	2	11.4	14.0
4	3	4.3	30.2
3	4	10.3	19.8
2	5	10.7	16.6

Table 5-5:Avg. Precision in (%) of Test data of 15%. Visiting time is normalized between 1 and 2

Our experiments show that setting N to 4 and *w* to 3 and test data set size is 15% produces few but highly relevant recommendations as shown in Table 5-4 comparing from Table 5-2 and Table 5-3. This measurement is made for each session in the selected test data and average of precision is taken for each 5, 10 and 15% test data set.

These experiments show that normalization of time between 1 and 10 improves the prediction accuracy. For a comparison we only give the results of experiments that are run with the normalization values of 1-2 and 1-10. Table 5-2 and Table 5-3 and Table 5.4 present the results of the experiments for the data set 5, 10 and 15% respectively and Table 5-5 present the results of the experiments for the data set 15% that is taken as sample for comparison of normalized Visiting time between 1 and 2 with 1and 10. At the same time calculating the recommendation score in method 2 improves the prediction accuracy of the model comparing method 1, which only uses popularity for calculating recommendation.

These results prove that modeling the user transaction with a mixture of Poisson distributions produces promising prediction and recommendation rates when page (resource) time is normalized between 1 and 10 and w is 3 and N is 5 with method 2 comparing to Method 1 and Method 2 with w and N different from 3 and 5 respectively.

6 Chapter Six

Conclusion and Recommendation

6.1 Conclusion

Web based e-learning environments are becoming increasingly popular educational establishments. The rapid growth of e-learning has changed traditional learning behavior and presented a new situation to both educators (lecturers) and learners (students). Educators are finding it harder to guide Students to select suitable learning materials due to more and more learning materials online. Learners are finding it difficult to make a decision about which of learning materials best meet his / her situation and need to read.

To tackle the above stated problems, the researcher initiated to conduct a research having the main goal of developing learner interest model using log file and, cluster learners who have the same interest Finally test its performance using a prototype e-learning recommender module system for the education sector that can assist the students in recommending e-learning material.

Pertinent knowledge required for the development of learner interest model in web based e-learning recommender system was acquainted in extracting valuable attributes and information from log files and cluster learners by using XLSTAT statistical analysis software.

Regarding to the evaluation process of the model, the prototype e-learning recommender system registers promising retrieval performance which is an average value of 46.3%, 51.4% precision for the

methods using popularity information of pages and combination of popularity information and the Poisson parameter respectively .

Furthermore, the following conclusions are drawn from the findings with regard to the research questions:

- The applicability of learner interest model using time in e-learning material recommender system for recommending e-learning resources is promising.

- The proposed model in e-learning material recommender system contributes a lot in especially for those having less experienced educators or in the absence of educator.

6.2 Recommendations

Even though, promising results are observed under this study, there are problem areas that need further investigation for future work. Therefore, the researcher recommends the following issues as a future research direction based on the findings of this study.

- For designing learner interest model in web based e-learning recommender system using log file, one has to consider and analyze in detail about the access order of resources of learner, duration spent on each learning resource, frequency of visited learning resources and others. However, because of lack of enough log file, the designing learner interest model in this research is made based on duration of spent and frequency of visiting on each learning resource. So future study can be made integrating above all method.

- For designing learner interest model in web based e-learning recommender system using log file, one has to consider and analyze big enough log file to cluster learners very well .However only 267 sessions log file was used in this study. Therefore the researcher recommends making further study by using big enough log file to measure the performance of the learning interest.

- E-learning recommender system includes resource (Learning object) model, and user (Learner) model however, in this study the researcher focused only using of Learner model. So further research can be done to integrate learning object model to evaluate the performance of the recommender system

- In general, perspectives to evaluate performance of LIM and recommender systems we use system-centric, user-centric or combination of both approaches. In this study the designed learner interest model's performance was measured from system side .Therefore the researcher recommends further study to evaluate from user side.

7 Reference:

[1] **Dordrecht Heidelberg,(2011)**. *Recommender Systems Handbook,* Springer New York London Springer © Science + Business Media, LLC.

[2] **Lamia Berkani1, Omar Nouali and AzeddineChikh,(2009).** *Recommendation-based Approach for Communities of Practice of E-learning*, Department of Computer Science, USTHB University, Bab-Ezzouar, Algiers, Algeria.

[3] **Khairil Imran Ghauth , Nor Aniza Abdullah(2010).***Measuring learner's performance in e-learning recommender systems* Multimedia University and University of Malaya.

[4]**Khairil Imran Bin Ghauth, Nor Aniza Abdullah,(2009).***Building an E-Learning Recommender System using Vector Space Model and Good Learners Average Ratingin Multimedia*, University of Malaysia.

[5]**Reema Sikka, Amita Dhankhar, Chaavi Rana,(2012)** *.A Survey Paper on E-Learning Recommender System, Journal Article published 30 Jun 2012 in International Journal of Computer Applications volume 47 issue 9 on pages 27 to 30*India.

[6] **Waterhouse, S, (2005):***The Power of E-learning: The Essential Guide for Teaching in the Digital Age, Upper Saddle River, NJ: Pearson Education .Inc*

[7] **The Federal Democratic Republic of Ethiopia**, **Federal Ministry of Education**, **(2010).***Education Sector Development Program IV (ESDP IV) in Ethiopia, Program Action Plan.*

[8] **Reema Sikka ,Amita Dhankhar Chaavi Rana,(2012).** *A Survey Paper on E-Learning Recommender System, Journal Article published 30 Jun 2012 in International Journal of Computer Applications volume 47 issue 9 on pages 27 to 30.*

[9]**JOHN SSEBUWUFU, TERALYNN LUDWICK**, **(2012):** *STRENGTHENING UNIVERSITY-INDUSTRY LINKAGES IN AFRICA, A Study on Institutional Capacities and Gaps,* Association of African Universities (AAU).

[10]**A.S. KANNAN, (2012).***EXISTENCE OF AND BENEFITS FROM LINKAGES BETWEEN UNIVERSITY AND INDUSTRY IN ETHIOPIA*, Dilla University.

[11] **Jie Lu(2004)** *A Personalized e-Learning Material Recommender System, Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004).*

[12] **JOHN W. CRESWELL (2009)**.*RESEARCH DESIGN Qualitative, Quantitative, and Mixed Methods Approaches, 3 RD EDITION, UNIVERSITY OF NEBRASKA-LINCOLN.*

[13] **Kassahun M. and Zelalem T**,**(2008*)*.** *Assessing the impact of plasma Television in Ethiopia*, *Ethiop. J. Educ. &Sc. Vol. 7 No. 2 ,2012*

[14] **Radoslav Pavlov, DesislavaPaneva(2009).***PERSONALIZED AND ADAPTIVE e-LEARNING - APPROACHES AND SOLUTIONS*, Institute of Mathematics and Informatics - BAS

[15]**Khribi, M. K., Jemni, M., &Nasraoui.O**,**(2009).** *Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval, Educational Technology & Society, 12 (4), 30–42.*

[16]**Professor Kate Ashcroft**, **(2006).** *New Higher Education Institutions for Ethiopia, Analysis and discussion of curriculum, resource and organizational issue*s, Ethiopia higher Eduction strategy center.

[17]**ME Herselman, HR Hay**,**(2003)**. *Challenges Posed by Information and Communication Technologies (ICT) for South African Higher Education Institutions, informing Science, University of the Free State, Bloemfontein, South Africa.*

[18]**A. S. Sife, E.T. Lwoga and C. SangaSokoine**,**(2007).***New technologies for teaching and learning: Challenges for higher learning institutions in developing countries*, University of Agriculture, Tanzania

[19] **Chris Fraley, Adrian E.(2007**): *Model-based Methods of Classification: Using the mclust Software in Chemometrics, University of Washington, Journal of Statistical Software January 2007, Volume 18, Issue 6.*

[20] **Robin Burke (2002).** *Hybrid Recommender Systems: Survey and Experiments, California State University, Fullerton.*

[21] **P Pu, L Chen, R Hu (2012).** *User Modeling and User-Adapted Interaction, Evaluating recommender systems from the user's perspective: survey of the state of the art.*

[22] **L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai (2011),** *ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011*

[23]**Tsai and Machado(2002),**Susanna Tsai and Paulo and Machado. Essay: *E-learning, current terminology,e_Learning*,no.7 p.p.3,

[24]**Romiszowski, A**. **(2004).***How is the e-learning baby? Factor leading to success or failure of education technology innovation, Educational Technology*.

[25]**Willems,J**. **(2005).***Flexible learning: Implication of "when-ever," "where-ever" .Distance Education*,26(3),429-435.

[26]**Dede,C**. **(2000)**.*Emerging technologies and distributed learning in higher education. In D.Hanna (Ed.), Higher education in an era of digital competition: Choices and challenges*. New York: Atwood.

[27]**Spiro,R.J.,Feltovich,P.J.,Jacobson,M.J.,&Coulson,R.L.(1991).***Cognitive flexible, constructivism and hypertext: Random access instruction for advanced knowledge aqustion in ill-structured domains*. Educational Technology.31(5),24-33.

[28]**Naidu,S**. **,(2003).***Designing instruction for e-learning,* In M.G.Moore& *B.G, the University of Melbourne, Australia, 3010.*

[29]**Satchidananda,Dehuri, ManasRanjanPatra, BijanBihariMisra&Alok Kumar Jagadev**: **(2012).***Intelligent Techniques in Recommendation Systems: Contextual Advancements,* 1st *IGI Publishing Hershey, PA, USA ©2012 ,ISBN:1466625422 9781466625426.*

[30]**Huizhi Liang**,**(2011)**:*USER PROFILING BASED ON FOLKSONOMY INFORMATION IN WEB 2.0 FOR PERSONALISED RECOMMENDER SYSTEMS*, Queensland University of Technology.

[31]**THORSTEN JOACHIMS,LAURA GRANKA,BING PAN,HELENE HEMBROOKE,FILIP RADLINSKI,GERI GAY**, **(2007).**: *Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search, Journal, ACM Transactions on Information Systems (TOIS) , Volume 25, Article No. 7.*

[32]**Brusilovsky and Millan**, **(2007).***User Models for Adaptive Hyper Media and adaptive Educational System, Book: the adaptive web page p.p 3-53.*

[33]**Daniel Billsus and Michael J. Pazzani,(1999).**:*A Hybrid User Model for News Story Classification, Proceeding UM '99 Proceedings of the seventh international conference on User modeling pp 99-108*

[34]**Joseph P.Forgas,Kipling D. Williams and William Von Hipple,(2003)**: *Social Judgments Implicitly and explicitly process, USENIX Symposium on Internet Technologies and Systems (USITS'99), Boulder, CO, USA, October 11-14, pp. 139-150s.*

[35]**Eibe Frank and Ian H.Witten**, **(2000)**. *Data mining: Practical Machine Learning Tools and Techniques with Java Implementation*, From the University of Waikato, New Zealand.

[36] **Jesús Bobadilla (2013).***Recommender systems: survey and experiment, Yale University, Fullerton.*

[37] **Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., (1996).***Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, MA.

[38] **Etzioni, O.,(1996).***The World Wide Web*: Quagmire or gold mine, Communications of the ACM, 39, 65-68.

[39] **Borges, J. and Levene, M.,(1999).***Data mining of user navigation patterns, International WEBKDD Workshop - Web Usage Analysis and User Profiling*, San Diego, CA, USA, August 15, pp. 31-36.

[40] **Madria, S. K., Bhowmick, S. S., Ng, W.K. and Lim E. P**., **(1999).***Research issues in Web data mining, 1st International Conference on Data Warehousing and Knowledge Discovery (DaWaK'99)*, Florence, Italy, August 30 -September 1, pp. 303-312.

[41] NETCRAFT, http://www.netcraft.com/.

[42] **Cooley, R., Mobasher, B. and Srivastava, J**., **(1997).***Web mining: Information and pattern discovery on the World Wide Web, 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, Newport Beach, CA, USA, November 03-08, pp. 558-567.

[43]**Kosala, R. and Blockeel, H.,(2000).***Web mining research*: A survey, ACM SIGKDD Explorations, 2, 1-15.

[44] **Aggarwal, C. C, Wolf, J. L. and Yu, P. S., (1999***). Caching on the World Wide Web, IEEE Transactions on Knowledge and Data Engineering*, 11, 95-107.

[45] **Shim, J., Scheuermann, P. and Vingralek, R.,(1999).***Proxy cache algorithms: Design, implementation and performance, IEEE Transactions on Knowledge and Data Engineering*, 11, 549-562.

[46] **Pitkow, J. and Pirolli, P**., **(1999).** *Mining longest repeating subsequences to predict World Wide Web surfing, USENIX Symposium on Internet Technologies and Systems (USITS'99)*, Boulder, CO, USA, October 11-14, pp. 139-150.

[47]**Sarukkai, R. R.,(2000).***Link prediction and path analysis using markov chains, Computer Networks*, 33, 377-386.

[48] **Pirolli, P. Pitkow, J. and Rao, R.,(1996**). *Silk from a sow's ear: Extracting usable structures from the Web*, ACM Conference on Human Factors in Computing Systems.

[49] **Goldberg, D., Nichols, D., Oki, B. and Terry, D.,(1992).***Using collaborative filtering to weave an information tapestry*, Communications of the ACM, 35, 61-70.

[50]**Yan, T. W. and Molina, H. G.,(1999).***The SIFT information dissemination system, ACM Transactions on Database Systems*, 24, 529-565.

[51]**Rucker, J. and Polano, M. J.,(1997).***Siteseer: Personalized navigation for the Web, Communications of the ACM*, 40, 73-75.

[52] **Shardanand, U. and Maes, P.,(1995).***Social information filtering : Algorithms for automating word of mouth., ACM Conference on Human Factors in Computing Systems(CHI'95)*, Denver, CO,USA,May7-11, pp. 210-217.

[53] **Agrawal, R., Imielinski, T. and Swami, A.,(2003).***Mining association rules between sets of items in large databases, ACM SIGMOD Conference on Management of Data*, Washington, D.C, USA, May 26-28, pp. 207–216.

[54]**Mohamed KoutheaïrKhribi, Mohamed Jemniand OlfaNasraoui,(2009).***Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval*, *University of Tunis*, *Tunisia.*

[55] **Catledge, L. D. and Pitkow, J. E.,(1995).***Characterizing browsing strategies in the World-Wide Web, Computer Networks and ISDN Systems*, 27, 1065-1073.

[56] **Jiawei Han and Micheline Kamber (2000)** *Data Mining: Concepts and Techniques (2000), 2nd Edition University of Illinois at Urbana-Champaign*

[57] **Han, J., Pei, J. and Yin, Y.,(2000).***Mining frequent patterns without candidate generation, ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, USA, May 16-18, pp. 1-12.

[58]**Dempster, A. P., Laird, N. M. and Rubin, D. B**., **(2007).***Maximum likelihood from incomplete data via the EM algorithm*, Journal of Royal Statistical Society, 39, 1-38.

[59] **Hand, D., Mannila, H. and Smyth, P.,(2001).** Principles of Data Mining, A Bradford Book The MIT Press Cambridge, Massachusetts London England .

[60]**Bartoszynski.R.and Niewiadomska-Bugaj. M**., **(1996).** *Probability and Statistical Inference*, John Ẃiley & Sons, Inc.

[61] **DeGroot, M.H. and Schervish, M.J.,(2002)**.*Probability and Statistics*, Addison-Wesley Publishing.

[62]**XlStat (n.d).** *Retrieved August 05, 2015, from http://www.xlstat.com/en/*

[63]**Adomavicius, G., Tuzhilin, A.(2005).***Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions*. IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No 6, pp. 734–749.

[64]**Brusilovsky, P.,(2002).***Adaptive and Intelligent Technologies for Web-Based Education, in: Rollinger, C. &Peylo, C.Eds., Special issue on intelligent systems and tele-teaching, KunstlicheIntelligenz*, , pp. 19–25.

[65]**Ho, S.Y.(2006).***The Attraction of Internet Personalization to Web Users*. Electronic Markets, Vol.16, No 1, pp.41–50.

[66]**Kim, H.K., Kim, J.K., Ryu, Y.U.(2009).***Personalized Recommendation over a Customer Network for Ubiquitous Shopping*. IEEE Transactions on Services Computing, Vol. 2, No. 2, pp. 140-151.

[67]**Kim, H.K., Cho, Y.H., Kim, W.J., Kim J.R., Suh, J.H**. **(2002)**. *A personalized recommendation procedure for Internet shopping. Electronic Commerce Research and Applications*,Vol.1, pp.301–313.

[68] **Ricci, F., Werthner, H.(2006).***Recommender systems. International Journal of Electronic Commerce*, Vol. 11, No 2, pp. 5–9.

[69]**S¸ule GÜNDÜZ(2003)** RECOMMENDATION MODELS FOR WEB *USERS*, ISTANBUL TECHNICAL UNIVERSITY INSTITUTE OF SCIENCE AND TECHNOLOGY

[70]**Melkamu Beyene**,**(2010).***Adaptive E-learning Model Design By Artificial Neural Network Techniques: A case of Ethiopian Higher Learning Institution*.

[71]**User Factors in Recommender Systems (2007):** *Case Studies in e-Commerce, News Recommending, and e-Learning, Dissertations in Interactive Technology, Number 17 Tampere 2014*

[72]**Griffith, Josephine and Colm O'Riordan. (2000).** *Collaborative Filtering, ACM Conference on Human Factors in Computing Systems(CHI'95)*, Denver, CO,USA,May7-11, pp. 210-217.

[73]**Herlocker, J.L.(2000).** *Understanding and Improving Automated collaborative Filtering Systems. University of Minnesota: Faculty of Graduate School.*

*[74]***Herlocker, J.L., and et al. (1999).** *"An Algorithmic Framework for Performing Collaborative Filtering." In Proceedings of the 1999 Conference on Research and Development in Information Retrieval.*

*[75]***Koubarakis, Manolis et al. (2001)**. *Efficient Agent-Based Dissemination of Textual Information, Journal, ACM Transactions on Information Systems (TOIS) , Volume 25, Article No. 7.*

.

[76] **Geyer-Schulz, Andreas and Michael Hahsler. (2002)**. *"Evaluation of Recommender Algorithms foran Internet Information Broker Based on Simple Association Rules and on The Repeat-Buying Theory." In Proceedings WEBKDD.*

[77] **Sarwar, Badrul and et al. (1998)**. *Using Filtering Agents to Improve Prediction Quality in the Group Lens Research Collaborative Filtering System, In Proc. Of ACM CSCW'98.*

[78] **Breese, J. S., D. Heckerman, and C. Kadie. 1998**. *"Empirical Analysis of Predictive Algorithms for Collaborative Filtering." In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp. 43-52.*

[79] **Vladimir Estivill-Castro :(2002):** *Why so many clustering algorithms: a position paper, ACM SIGKDD Explorations Newsletter V .2, P.P 65-67.*

[80] **Jeffrey D. Banfield;Adrian E.Raftery(1993):***Model-based Gaussian and Non Gaussian Clustering, Biometrics, Vol.49,803-821*

7.1 Appendix: Source Code for Module of recommendation

```
ListView1.Items.Clear()
Dim mTempRecommendedPages As New Dictionary(Of String, Integer)
Dim mRecommendedPages As New List(Of String)
Dim mMinDifferenceClass As Integer = 0
Dim mViewdPages As New List(Of String)
For i As Integer = 0 To Me.dgPagesList.Rows.Count - 1
    If Not String.IsNullOrEmpty(Me.dgPagesList.Rows(i).Cells(1).Value) Then
        If Me.dgPagesList.Rows(i).Cells(1).Value <> 0 Then
            mViewdPages.Add(Me.dgPagesList.Rows(i).Cells(0).Value)
        End If
    End If
Next

mMinDifferenceClass = mClusterPageMean.Get_ClusterClass(Me.dgPagesList)
Me.lblClusterClass.Text = mMinDifferenceClass
Dim mFrequencyDataSet As DataSet = mClusterPageMean.ReadFrequency(mMinDifferenceClass)
Dim mPoissonDataSet As DataSet = mClusterPageMean.ReadCluterPagesPoisson(mMinDifferenceClass)
If mFrequencyDataSet.Tables.Count > 0 Then
    Dim mFreqTable As DataTable = mFrequencyDataSet.Tables(0)
    Dim mPoissonTable As DataTable = mPoissonDataSet.Tables(0)
    For i As Integer = 0 To mFreqTable.Columns.Count - 1
        If Me.rbFrequency.Checked Then
            mTempRecommendedPages.Add(mFreqTable.Columns(i).ColumnName, mFreqTable.Rows(0).Item(i))
        Else
            mTempRecommendedPages.Add(mFreqTable.Columns(i).ColumnName, mFreqTable.Rows(0).Item(i) * mPoissonTable.Rows(0).
        End If
    Next


    Dim mma As New Dictionary(Of String, Integer)
    For i As Integer = 0 To mTempRecommendedPages.Count - 1
        mma.Clear()
        mma.Add(mTempRecommendedPages.Keys(i), 0)
        For Each item As KeyValuePair(Of String, Integer) In mTempRecommendedPages

            If mma.Values(0) < item.Value Then
                If Not mRecommendedPages.Contains(item.Key) Then
                    mma.Clear()
                    mma.Add(item.Key, item.Value)
                End If
            End If
        Next

        If Not mRecommendedPages.Contains(mma.Keys(0)) Then
            mRecommendedPages.Add(mma.Keys(0))
        End If

    Next
```

```
        '

    For Each item As String In mRecommendedPages
        If Not mViewdPages.Contains(item) And ListView1.Items.Count < Me.nudRecommendedPages.Value Then
            ListView1.Items.Add(item)
        End If

    Next
End If
```

7.2 Appendix: The EM algorithm for Mixture of Poison Distribution

7.2.1 The ML Optimization Frame Work

To compute the necessary equations used for obtaining ML parameters in the E-step, we should compute the conditional probability of missing class labels given the current parameter set Θ'. We define this probability as *cluster-posterior* probability, $P_{ig}(\Theta')$, that the session $\mathbf{x}_i$ arose from the g^{th} cluster. We can write the cluster-posterior probability using Bayes' rule as:

$$\begin{aligned} P_{ig}(\Theta') &= p(\mathbf{C} = c_g|\mathbf{x}_i) \\ &= \frac{p(\mathbf{C} = c_g)p(\mathbf{x}_i|c_g, \Theta'_g)}{p(\mathbf{x}_i)} \\ &= \frac{\tau_g p(\mathbf{x}_i|c_g, \Theta'_g)}{\sum_{j=1}^{G} \tau_j p(\mathbf{x}_i|c_j, \Theta'_j)} \end{aligned} \tag{B.1}$$

The Q-function can be written as:

$$Q(\Theta, \Theta') = \sum_{i=1}^{K} \sum_{g=1}^{G} P_{ig}(\Theta') \left[\ln p(\mathbf{x}_i|c_g, \Theta_g) + \ln \tau_g\right] \tag{B.2}$$

In M-step, keeping the cluster-posterior probabilities fixed, we reassign a new set of parameters $\Theta'(n+1)$ so as to maximize the expected log likelihood of the training data. The Q-function is maximized subject to the constraint that the cluster priors sum to 1. In order to perform constrained maximization, a Lagrange multiplier is used. The estimating equations for cluster priors are as follows:

$$\begin{aligned} \frac{\partial}{\partial \tau_g}\left[Q(\Theta, \Theta') - \lambda \sum_{j=1}^{G} \tau_j\right] &= 0 \\ \sum_{i=1}^{K} P_{ig}(\Theta')\left[\frac{1}{\tau_g}\right] - \lambda &= 0 \end{aligned} \tag{B.3}$$

from which it follows:

$$\lambda\tau_g = \sum_{i=1}^{K} P_{ig}(\Theta') \tag{B.4}$$

If we sum Equation (B.4) over g we obtain:

$$\lambda = \sum_{i=1}^{K}\sum_{g=1}^{G} P_{ig}(\Theta') = K \tag{B.5}$$

Last equation follows from the fact that $\sum_{g=1}^{G} P_{ig}(\Theta') = 1$. By combining Equation (B.4) and Equation (B.5), we obtain the equation for updating the cluster probabilities:

$$\hat{\tau}_g = \frac{1}{K}\sum_{i=1}^{K} P_{ig}(\Theta') \tag{B.6}$$

Similarly we can maximize the Q-function with respect to the parameters of Poisson model, Θ_g, under the independence assumption:

$$\begin{aligned}
\frac{\partial}{\partial\theta_{gm}}[Q(\Theta,\Theta')] &= 0 \\
\frac{\partial}{\partial\theta_{gm}}\left[\sum_{i=1}^{K} P_{ig}(\Theta'_g)\left(\ln\prod_{j=1}^{n}\frac{(\theta_{gj})^{x_{ij}}\, e^{-\theta_{gj}}}{x_{ij}!} + \ln\tau_g\right)\right] &= 0 \\
\sum_{i=1}^{K} P_{ig}(\Theta')\left[\frac{x_{im}}{\theta_{gm}} - 1\right] &= 0
\end{aligned} \tag{B.7}$$

which yields the following update equation for Poisson parameters:

$$\hat{\theta}_{gm} = \frac{\sum_{i=1}^{K}(P_{ig}(\Theta')x_{im})}{\sum_{i=1}^{K} P_{ig}(\Theta')} \tag{B.8}$$

7.2.2 The MAP Optimization Framework

The E-step of the MAP estimation of parameters is the same as the ML estimation problem where the conditional probability of missing class labels given the current parameter set Θ' is computes as in Equation B.1. The Q function of the EM algorithm for the log posterior function is defined as:

$$Q(\Theta,\Theta') = \sum_{i=1}^{K}\sum_{g=1}^{G} P_{ig}(\Theta')\left[\ln p(\mathbf{x}_i|c_g,\Theta_g) + \ln\tau_g\right] + \ln p(\Theta) \tag{B.9}$$

where the parameters Θ consists of all mixture model parameters:

$$\begin{aligned} \Theta &= \{\Theta_1, ..., \Theta_G, \tau\} \\ \Theta_g &= \{\theta_{g1}, ..., \theta_{gn}\} \\ \tau &= \{\tau_1, ..., \tau_G\}, \sum_{g=1}^{G} \tau_g = 1 \end{aligned} \tag{B.10}$$

The prior term $p(\Theta)$ in Equation B.9 consists of Poisson parameter prior and cluster priors. This can be decomposed as:

$$p(\Theta) = \prod_{g=1}^{G} \prod_{l=1}^{n} p(\theta_{gl}|\alpha_{gl}, \beta_{gl}) p(\tau|\gamma) \tag{B.11}$$

where we use gamma priors with parameters α and β for Poisson parameters and Dirichlet priors for cluster weights :

$$\begin{aligned} p(\theta_{gl}|\alpha_{gl}, \beta_{gl}) &\propto \theta_{gl}^{\alpha_{gl}} e^{-\beta_{gl}\theta_{gl}} \\ p(\tau|\gamma) &\propto \prod_{g=1}^{G} \tau_g^{\gamma_g} \end{aligned} \tag{B.12}$$

Then, the Q function in Equation B.9 can be written as:

$$\begin{aligned} Q(\Theta, \Theta') &= \sum_{i=1}^{K} \sum_{g=1}^{G} P_{ig}(\Theta') \left[\ln p(\mathbf{x}_i|c_g, \Theta_g) + \ln \tau_g\right] \\ &+ \sum_{g=1}^{G} \sum_{l=1}^{n} (\alpha_{gl} \ln \theta_{gl} - \beta_{gl}\theta_{gl}) + \sum_{g=1}^{G} \gamma_g \ln \tau_g \end{aligned} \tag{B.13}$$

To calculate the optimal parameters we maximize the Q function in Equation B.13 subject to the constraint that cluster priors sum to 1:

$$\begin{aligned} \frac{\partial}{\partial \tau_g}\left[Q(\Theta, \Theta') - \lambda \sum_{j=1}^{G} \tau_j\right] &= 0 \\ \sum_{i=1}^{K} P_{ig}(\Theta')\left[\frac{1}{\tau_g}\right] + \frac{\gamma_g}{\tau_g} - \lambda &= 0 \end{aligned} \tag{B.14}$$

from which it follows:

$$\lambda\tau_g = \sum_{i=1}^{K} P_{ig} + \gamma_g \tag{B.15}$$

Summing Equation B.15 over g we obtain:

$$\lambda = \sum_{j=1}^{G} \left[\sum_{i=1}^{K} P_{ij} + \gamma_j \right] \tag{B.16}$$

Upon substituting Equation B.16 into Equation B.15 and solving cluster priors, we obtain the update equation for cluster weights:

$$\hat{\tau}_g = \frac{\sum_{i=1}^{K} P_{ig}(\Theta') + \gamma_g}{\sum_{j=1}^{G} \left[\sum_{i=1}^{K} P_{ij}(\Theta') + \gamma_j \right]} \tag{B.17}$$

Optimizing the Q-function with respect to Poisson parameters we obtain:

$$\begin{aligned} \frac{\partial}{\partial\theta_{gm}} [Q(\Theta, \Theta')] &= 0 \\ \sum_{i=1}^{K} P_{ig}(\Theta') \left[\frac{x_{im}}{\theta_{gm}} - 1 \right] + \frac{\alpha_{gm}}{\theta_{gm}} - \beta_{gm} &= 0 \end{aligned} \tag{B.18}$$

Equation B.18 can be solved for θ_{gm} to obtain update equation for Poisson parameters as follows:

$$\hat{\theta}_{gm} = \frac{\sum_{i=1}^{K} P_{ig}(\Theta') x_{im} + \alpha_{gm}}{\sum_{i=1}^{K} P_{ig}(\Theta') + \beta_{gm}} \tag{B.19}$$