A New Profile Learning Model for Recommendation System based on Machine Learning Technique

Shereen H Ali*, Ali I El Desouky and Ahmed I Saleh

Department of Computer Eng. & Systems, Faculty of Engineering, Mansoura University, Egypt

Abstract

Recommender systems (RSs) have been used to successfully address the information overload problem by providing personalized and targeted recommendations to the end users. RSs are software tools and techniques providing suggestions for items to be of use to a user, hence, they typically apply techniques and methodologies from Data Mining. The main contribution of this paper is to introduce a new user profile learning model to promote the recommendation accuracy of vertical recommendation systems. The proposed profile learning model employs the vertical classifier that has been used in multi classification module of the Intelligent. Adaptive Vertical Recommendation (IAVR) system to discover the user's area of interest, and then build the user's profile accordingly. Experimental results have proven the effectiveness of the proposed profile learning model, which accordingly will promote the recommendation accuracy.

Keywords: Recommendation systems; Machine learning; Classification; Profile learning model

Introduction

Recommender systems (RSs) have been used to successfully address the information overload problem by providing personalized and targeted recommendations to the end users [1]. Personalized information systems emerged as an answer to the problem of steadily growing amounts of information and constantly increasing complexity of navigation in the information space that overwhelms the user. These systems are able to learn about the needs of individual users and to tailor the content, appearance, and behaviour to the user needs [2].

Examples of personalization range from online shops recommending products identified based on the user's previous purchases to web search engines sorting search hits based on the user's browsing history. The aim of such adaptive behaviour is to help users to find relevant content easier and faster. To achieve such behaviour, the system needs a user model providing information about users, such as their interests, expertise, background, or traits. It also needs metadata of information resources and some logic or rules that govern how the resources must be delivered to users given their user model [3]. To build a user profile, the information needed can be obtained explicitly, which is provided directly by the user, or implicitly through the observation of the user's actions [4].

The application of Machine Learning techniques is a standard way to perform the task of learning user profiles in recommender systems [5], such as Clustering [6,7], Genetic Algorithms [8,9], Neural Networks [10,11], and Classification Techniques [12-15]. Unfortunately, these techniques suffer from vital drawbacks. In clustering techniques [6,7], they suffer from low precision, over fitting the training data, time consuming and the difficulty for evaluate a number of clusters automatically as in. In Genetic Algorithms [8,9], the computational requirements and high runtime were their greatest weakness. In Neural Networks [10,11], time to train NN is probably identified as the biggest disadvantage. In classification technique [12-14], they suffer from low accuracy and high computation cost respectively.

Since the classification technique is a common approach for user profile learning in recommendation systems. Therefore, it will be employed in the proposed profile learning model. The main contribution of this paper is to introduce: firstly, a brief survey of machine learning techniques. Secondly, new profile learning model to promote the recommendation accuracy of vertical recommendation systems. The proposed profile learning model employs the vertical classifier that have been used in Multi Classification Module of the Intelligent Adaptive Vertical Recommendation (IAVR) system [16], to discover the user's area of interest, and then build the user's profile accordingly.

The remainder of this paper is organized as follows; Section 2 introduces a review of Machine learning. In section 3, The Profile Learning Model (PLM) is introduced. The experimental results in PLM are presented in section 4. Finally, conclusions are drawn in Section 5.

Machine Learning

Machine learning is a branch of computer science that extracted from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the formation and study of algorithms that can learn from and create predictions on data. Such algorithms operate by building a model from the input data so as to form data-driven predictions or decisions, instead of following severely static program instructions [17].

Machine learning categories

Machine learning algorithms are divided into three broad categories which are [18]; (i) supervised learning: where the algorithm generates a function that maps inputs to desired outputs. The standard formulation of the supervised learning task is the classification problem where, the learner is needed to learn a function learn a function which maps a vector into one of several classes by looking at several input-output examples of the function. (ii) Unsupervised learning: a process that automatically detects structure in data and does not involve any steerage as the assignment of patterns to classes. (iii)Reinforcement learning: where the algorithm interacts with a dynamic environment

*Corresponding author: Shereen H Ali, Department of Computer Engineering and Systems, Faculty of Engineering, Mansoura University, Egypt, Tel: +20 50 2383781; E-mail: engshereen_2005@yahoo.com

within which it must perform a definite goal (such as driving a vehicle), without an instructor frankly telling it whether or not it has come close to its goal

Machine learning techniques

Four of the most common Machine Learning techniques, which perform the task of learning, are described below.

Clustering: Clustering, also called unsupervised classification, is the process of segmenting heterogeneous data objects into a number of homogenous clusters. Each cluster is a collection of data objects that are similar to one another and dissimilar to the data objects in other cluster [6].

Clustering has been employed for user profile learning in different recommendation systems. An approach to recommender systems based on clustering methods is introduced [6]. The clustering part identifies similar users, who then are taken to create clusters profiles. The profiles clarify the most common users' preferences in one cluster. An active user can be compared to the profiles instead of all data, which reduce computation time. The system was implemented in Apache Mahout Environment and tested on a movie database. Selected similarity measures are based on: Euclidean distance, cosine as well as correlation coefficient and loglikelihood function.

A model for dynamic recommendation based on a hybrid clustering algorithm is proposed [7]. This model analysis the users behaviours and depend on the interests of similar patterns provides appropriate recommendations for active user. This model performs clustering using fuzzy techniques for better dynamic recommendation process.

Genetic algorithms: In the field of artificial intelligence, a genetic algorithm is way that emulates the process of natural selection. Genetic algorithms belong to the larger class of Evolutionary Algorithms, which generate solutions to optimization problems, using techniques motivated from natural evolution such as: inheritance, mutation, selection, and crossover [8].

Various recommendation systems have been employed Genetic Algorithm for learning user profile. A recommender system based on the genetic algorithms is proposed by Athani et al. [8]. The content-based filtering technique is applied to generate the initial population of genetic algorithm. The interactive genetic algorithm is employed so that the users can directly evaluate fitness value of candidate solution themselves. The recommender system is partitioned into three stages which are; feature extractionstage, evolution stage, and interactive Genetic algorithm stage. The clam software is provided with music file which extracts unique properties of music like pitch, chord, and tempo. This extracted data is then stored on the database. Each stored data is resolved using content based filtering and interactive genetic algorithm. After analyzing records, the system recommends items relevant to users own preference.

The main idea is to tackle the problem of high dimensionality and sparsity typical to RS data [9]. A proposed genetic programming based feature extraction technique to transform the user-item preference space into a compact and dense user-feature preference space has been introduced. The proposed approach is able to merge the advantages of both memory-based and model-based techniques since the compact user profile is exploited for user similarity computation while the original training matrix is used for the rating prediction.

Neural networks: Neural network refers to the information processing systems or computer software system that can simulate the structure and function of the biological brain [10]. It is nonlinear complex network system consisting of a large number of processing units that are similar to neurons. The structure of neural network is determined by the basic processing unit and their inter-connection methods.

Neural network has been utilized in recommendation system for modelling user interests. A method is proposed by Rajabi et al. [10] to create a user profile using clustering and neural networks in order to predict the user's future requests and then generate a list of the user's preferred pages. Through this study, different user interactions on the web are tracked and then clusters are created based on user interests. By using neural networks and clustering, the navigation patterns are created, so as to predict the user future desires.

The extracted user navigation patterns are used to capture similar behaviours of users in order to increase the quality of recommendations. Based on patterns extracted from the same user navigation, recommendations are provided to the user to make it easier to navigate. Lately, web browsing techniques have been widely used for personalization.

In [11], a recommendation system based on collaborative filtering with k-separability approach to create a product bundling strategy is proposed. The proposed recommendation system applies Neural Network for customer clustering and Data Mining to find association rules; it achieves a high hit probability between actual bundling manner and recommended strategy.

Classification techniques: Existing recommender systems almost utilize different classification techniques in the field of profile learning such as Naive Bayes classifier, Support Vector Machine (SVM) classifier, K-Nearest Neighbour (KNN) classifier. In reference [12], a content-based book recommending has been discussed by stratifying automated text-categorization methods to semi-structured text extracted from the web. The prototype system, called LIBRA (Learning Intelligent Book Recommending Agent), uses a database of book information extracted from web pages at Amazon.com. Users give 1-10 ratings for a selected set of training books, and then the system learns a profile of the user using a Naive Bayes classifier and then generates a ranked list of the most recommended additional titles from the system's list. A Personalized News Filtering and Summarization (PNFS) system is proposed by Noia et al. [13]. The idea is to design a content based news recommender that automatically secures Word Wide Web news from the Google news website and recommends personalized news to users according to their preference. Two learning strategies are used to model the user interest preference including the k-nearest neighbor and Naive Bayes. Furthermore, a new keyword extraction method based on semantic relations has been presented in this paper.

In [14], a model-based approach for a content-based recommender system exploiting exclusively Linked Open Data cloud to represent both the information on the items and on the user profiles. The main idea is to show how a model-based approach can be easily adapted to cope with the semantic Web, and to use a Support Vector Machine (SVM) classifier for learning the user profile.

In [15], a personal news agent that uses synthesized speech to read news stories to a user is introduced. The main idea is to motivate the use of a multi-strategy machine learning approach that allows for the creation of user models that consist of separate models for long-term and short-term interests. The aim of the short-term model is; (i) it should contain information about recently rated events, so that stories which belong to the same threads of events can be identified. (ii) It should allow for identification of stories that the user already knows.

The nearest neighbor classifier (NN) is a natural choice to achieve the required functionality in short-term model. On the other hand, the aim of the long-term user model is to model a user's general preferences for news stories and compute predictions for stories that could not be classified by the short-term model. The naïve Bayesian classifier is selected to achieve the long-term user model.

Profile Learning Model (PLM)

The proposed Profile Learning Model is shown in figure1. This model starts with identifying the user status; if the user uses the system for the first time, he enters his personal data and this data is registered in the User Id List (UIL). As depicted in the Figure 1, after user login to the system successfully, he should enter a query that describes his preference. Then, a domain thesaurus is employed for mapping the domain keywords found in the query to the corresponding domain concepts. As illustrated in Figure 2, terms (keywords) of the proposed domain thesaurus [16] are arranged into separate clusters; each cluster consists of a set of synonyms and represents a specific concept. For each cluster, one preferred term (PT) is chosen to represent the underlying concept; the other terms are non-preferred terms (NPT) (e.g., synonyms of the concept). The proposed domain thesaurus considers only the Synonym relation, which specify terms that express the same concept.

Therefore, a Preference List (PL) is constructed. This list contains the extracted domain concept from Virtual Document entered by the user. The Multi-class Classifier is then used to classify the user's query into one of the domain hypotheses. Therefore, the user profile database will be constructed. Document classification has a good impact in the overall system performance as it simplifies the matchmaking among the user's preferences (i.e. target hypothesis) and those multi-classified documents stored in the system's database during the recommendation process. If the user accepts the recommendation results the process is then completed. Otherwise, the system merges the user feedback (i.e. preference concepts) with those in the Preference List (PL) and then updates the user profile data-table. Figure 3 illustrates the steps followed for the proposed Profile Learning Algorithm.

The Proposed Merged Multi-class Classifier (MMC)

A new classifier called Merged Multi-class Classifier (MMC) [16] where, we take the decision to merge both AC and ANB classifiers. To clarify the idea, consider a document $Doc=\{c_1, c_2,......, c_m\}$, where $c_i \; \forall i \in [1 \rightarrow m]$ are the concepts extracted from Doc. The conditional probability of c_i given hypothesis h_j, denoted as; $P_T(c_i|h_j)$ can be calculated using (1) as [16];

$$P_T\left(c_i|h_j\right) = P\left(c_i|h_j\right)_{AC} + P\left(c_i|h_j\right)_{INB} \tag{1}$$

Where $P(c_i|h_j)_{ANB}$ is the concept conditional probability considering the Accumulative Naive Bayes (ANB) classifier, which is calculated by the Private Probability Distribution of h_j (denoted as; $PPD(h_j)$). On the other hand, $P(c_i|h_j)_{AC}$ is the conditional probability of c_i given hypothesis h_j considering the association rules among c_i and other domain concepts extracted from Doc given hypothesis h_j. For calculating $P(c_i|h_j)_{AC}$, initially, all association rules that contain c_i given hypothesis is h_j are picked. Rules are then refined by only considering those rules whose concepts are included in both Doc and the rule set of h_j, which are denoted as the Refined Rule Set (RRS). Assuming that c_i may appear in several association rules in RRS, $P(c_i|h_j)_{AC}$ can be calculated in two different manners, which are illustrated in (2) and (3) [16].

$$P\left(c_i|h_j\right)_{AC} = MAX\left[Prob.\left(R_m, h_j\right)\right]_{\forall R_m \in RRS} \tag{2}$$
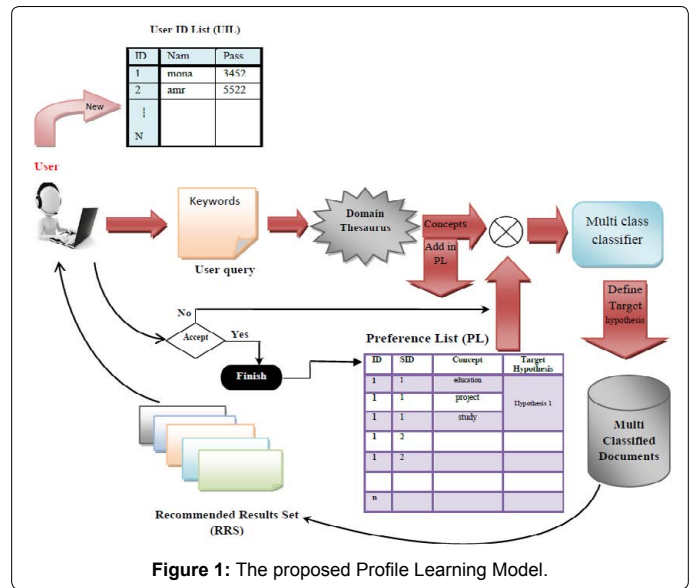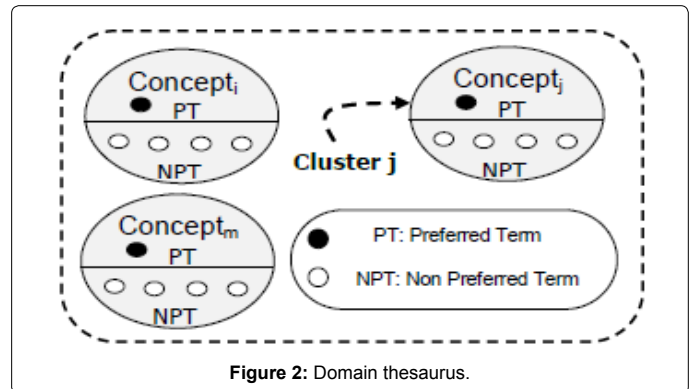

Figure 1: The proposed Profile Learning Model.
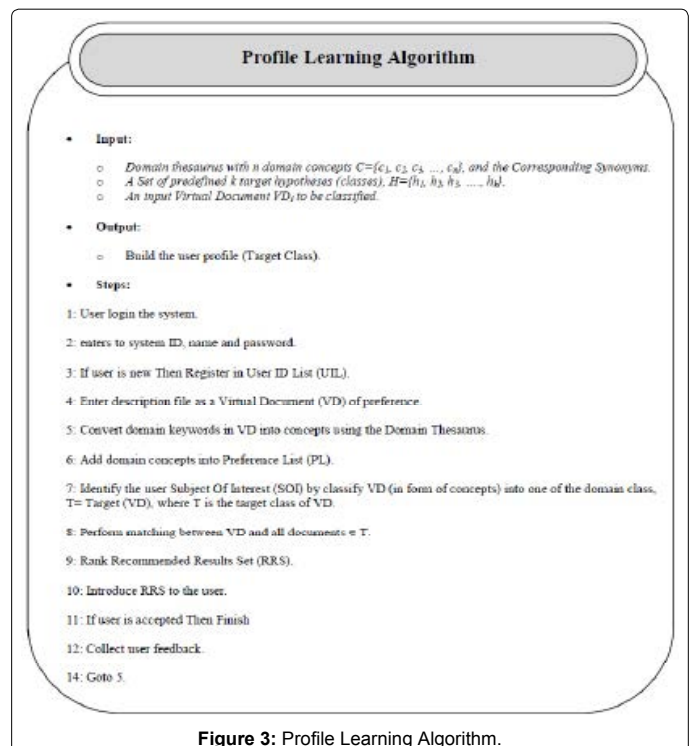

Figure 2: Domain thesaurus.

Figure 3: Profile Learning Algorithm.

$$P\left(c_i|h_j\right)_{AC} = \sum\nolimits_{\forall R_m \in RRS} Prob.\left(R_m, h_j\right) \qquad (3)$$

Where, $Prob(R_m,h_j)$ is the probability of the rule R_m given hypothesis h_j. Hence, the conditional probability for concept c_i can be calculated using (4) as [16];

$$P_T\left(c_i|h_j\right) = P\left(c_i|h_j\right)_{INB} + \begin{cases} MAX\left[Prob.(R_m,h_j\right]_{\forall R_m \in RRS} \\ OR \\ \sum\nolimits_{\forall R_m \in RRS} Prob.\left(R_m, h_j\right) \end{cases} \qquad (4)$$

It is also noted that, if RRS=Φ, the second term of (4) will be neglected and accordingly, the merged classifier will be equivalent to ANB classifier.

Testing the performance of the Profile Learning Model

The proposed Profile Learning Model (PLM) with its core the Merged Multi-class Classifier (MMC) will be examined. MMC will be tested against the LIBRA system [12] and PNA system [15], to test the validity of the proposed IAVR system [16]. Experimental results are shown in Figures (4-12). Different evaluation metrics (Precision, Recall, Accuracy, F1 and Error) will be measured against the number of Training Documents. Also, processing time metric is measured against the number of users query. To evaluate the average performance across domain hypotheses, "micro-average" and "macro-average" scores should be evaluated. WebKb [19] is the data set used in the experiments presented below.

Figure 4 illustrates the classification accuracy against the number of Training documents. It is noticed that MMC introduced significant improvements in accuracy over all other techniques. MMC accuracy reaches 89% when number of training documents = 2800, while accuracy in LIBRA and PNA is 0.79, 0.76 respectively when number of training documents = 2800. Figure 5 indicates that MMC introduces significant degradation in error compared all other techniques. MMC error reaches 11% number of training documents = 2800. While, error in LIBRA, and PNA is 0.21, 0.24 respectively when number of training documents = 2800.

Figures 6 and 7 illustrate the classification macro and micro average "precision" scores against the number of training documents. It is noticed that, MMC introduced significant improvements in both macro and micro average "precision" score compared with all other techniques. Their values are 0.89 and 0.85 respectively when number of training documents = 2800. While values of both micro and macro average "precision" score in LIBRA and PNA is (0.75, and 0.78), and (0.59, and 0.62) respectively when number of training documents = 2800.

Figures 8 and 9 illustrate the classification micro and macro average "recall" scores against the number of training documents. Again, MMC introduced significant improvements in both micro and macro average "recall" compared with other techniques. Their values are 0.88 and 0.91 respectively at number of training documents = 2800. However, average values of both micro and macro average "recall" scores in LIBRA, and PNA is (0.75, and 0.79), and (0.62, and 0.65) respectively at number of training documents = 2800.

Figures 10 and 11 illustrate the classification micro and macro average "F1-measure" scores against the number of training documents. It is noticed that, MMC introduced significant improvements in micro and macro average of "F1-measure" compared with other classification techniques. Their values reach 86.5% and 90% respectively at number
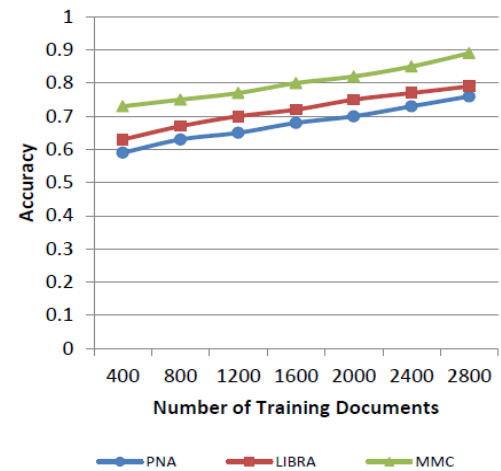

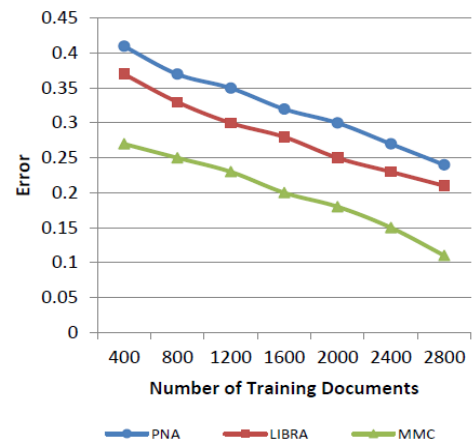**Figure 4:** Accuracy of the different classification techniques that can be used in PLL.

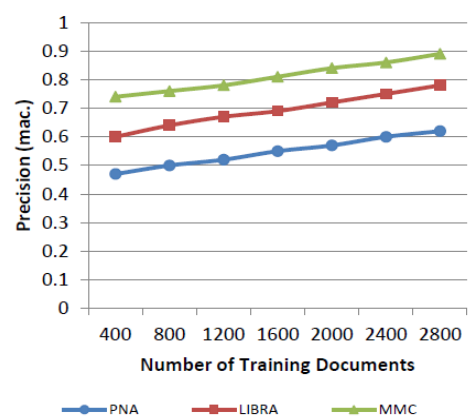**Figure 5:** Error of the different classification techniques that can be used in PLL.

Figure 6: Macro-Precision of the different classification techniques that can be used in PLL.

of training documents = 2800. While values of both micro and macro average of "F1-measure" in LIBRA and PNA is (0.75, and 0.78) and (0.6, and 0.63) respectively at number of training documents = 2800.

Figure 12 illustrates the processing time against the number of users query. It is noticed that, MMC introduced minimum processing time compared with other classification techniques. The average processing

time in MMC, LIBRA, and PNA reaches 700 msec., 1600 msec., and 3200 msec. respectively at number of users query=70.

Conclusion

In this paper, a new profile learning model has been proposed to promote the recommendation accuracy of vertical recommendation systems. The proposed profile learning model employs the vertical classifier that has been used in Multi Classification Module of IAVR
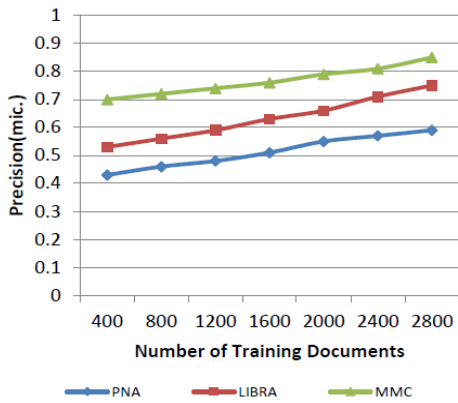

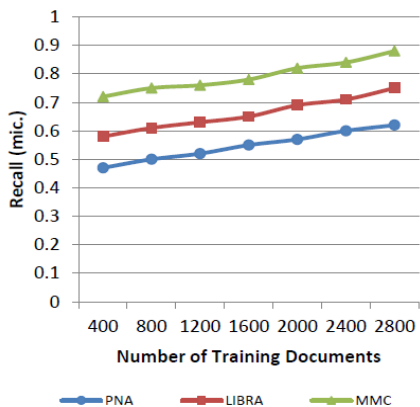**Figure 7:** Micro-Precision of the different classification techniques that can be used in PLL.

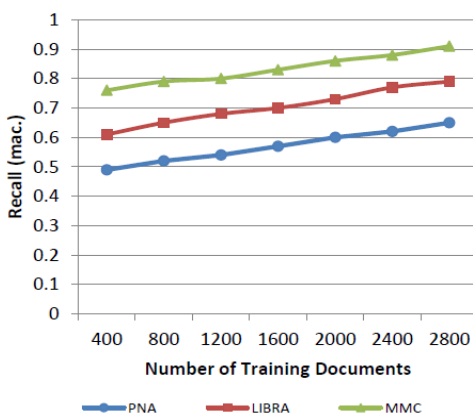**Figure 8:** Micro-Recall of the different classification techniques that can be used in PLL.

Figure 9: Macro-Recall of the different classification techniques that can be used in PLL.
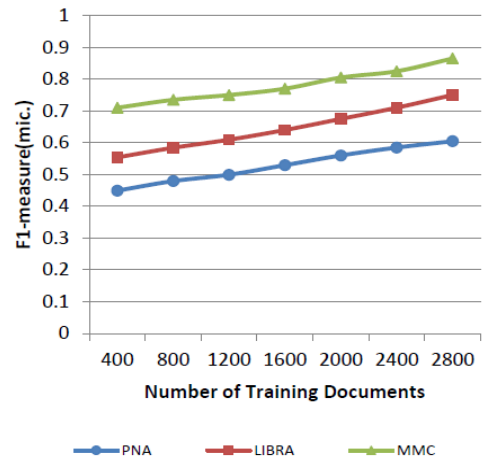

Figure 10: Micro-F1 of the different classification techniques that can be used in PLL.
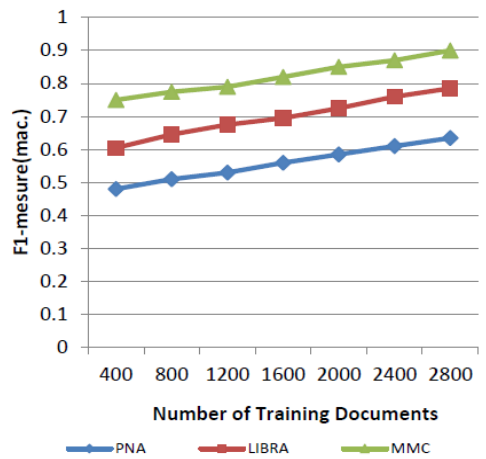

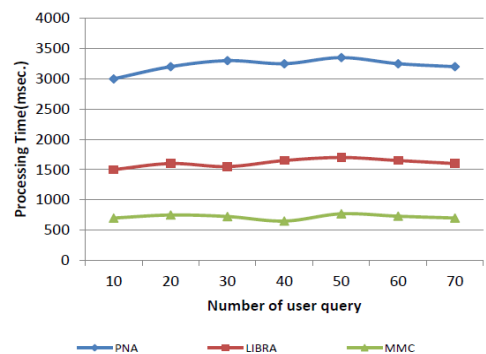**Figure 11:** Macro-F1 of the different classification techniques that can be used in PLL.

Figure 12: Processing Time of the different classification techniques that can be used in PLL.

System to discover the user's area of interest, and then build the user's profile accordingly. Experimental results have proven the effectiveness of the proposed profile learning model, which also will promote the recommendation accuracy. These results showed that MMC achieves the higher accuracy followed by LIBRA and finally comes the accuracy of PNA.

References

1. Shambour Q, Lu J (2012) A trust-semantic fusion-based recommendation approach for e-business applications. Decision Support Systems 54: 768-780.

2. Song G, Hu L, Zhao K, Xie Z (2014) Personalized Recommendation Algorithm Based on Preference Features. Tsinghua science and technology 19: 293-299.

3. Chandrashekhar H, Bhasker B (2011) Personalized Recommender System Using Entropy Based Collaborative Filtering Technique. Journal of Electronic Commerce Research 12: 214.

4. Jawaheer G, Szomszor M, Kostkova P (2010) Comparison of Implicit and Explicit Feedback from an Online Music Recommendation Service. In Proc. of 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems 47-51.

5. Pazzani M (1999) A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Review 13: 393-408.

6. Kużelewska U (2014) Clustering Algorithms in Hybrid Recommender System on MovieLens Data. Studies in logic, grammar and rhetoric 37: 125-139.

7. Nadi S, Saraee MH, Bagheri A (2011) A Hybrid Recommender System for Dynamic Web Users. International Journal Multimedia and Image Processing (IJMIP) 1: 3-8.

8. Athani M, Pathak N, Khan AU (2014) Dynamic Music Recommender System Using Genetic Algorithm. International Journal of Engineering and Advanced Technology (IJEAT) 3: 230-232.

9. Anand D (2012) Feature Extraction for Collaborative Filtering: A Genetic Programming Approach. IJCSI International Journal of Computer Science Issues. 9: 348-354.

10. Rajabi S, Harounabadi A, Aghazarian V (2014) A Recommender System for the Web: Using User Profiles and Machine Learning Method. International Journal of Computer Applications 96: 38-41.

11. Patil S, Mane Y, Dabre K, Dewan P, Kalbande DR (2012) An Efficient Recommender System using Collaborative Filtering Me-thods with K-separability Approach", International Journal of Engineering Research and Applications (IJERA) 30-35.

12. Moony RJ, Roy L (2000) Content-Based Book Recommending Using Learning for Text Categorization. Proceeding of 5th ACM Conference on Digital Libraries 195-204.

13. Wu X, Xie F, Wu G, Ding W (2011) Personalized News Filtering and Summarization on the Web. 23rd IEEE International Conference on Tools with Artificial Intelligence 414-421.

14. Noia T, Mirizzi R, Ostuni VC, Romito D (2012) Exploiting the Web of Data in Model-based Recommender Systems. Proceedings of sixth ACM conference on Recommender system 253-256.

15. Billsus D, Pazzani MJ (1999) A Personal News Agent that Talks, Learns and Explains. ACM Proceeding of the 3rd annual conference on Autonomous Agents 268-275.

16. Saleh I, Desouky AI El, Ali SH (2015) Promoting the performance of vertical recommendation systems by applying new classification techniques. Knowledge-Based Systems 75: 192-223.

17. Cios K, Kurgan L (2005) Trends in data mining and knowledge discovery. Springer.

18. Kurgan L, Musilek P (2006) A survey of knowledge discovery and data mining process models. Knowledge Engineering Review 21: 1-24.

19. Lin Y, Jiang J, Lee S (2014) A Similarity Measure for Text Classification and Clustering. IEEE Transactions on Knowledge & Data Engineering 26: 1575-1590.