International Journal of Computer Applications (0975 8887)
Volume 108 - No. 3, December 2014

User Profiling-A Short Review

Ayse Cufoglu
Anglia Ruskin University
Department of Computing and Technology
Cambridge, UK

ABSTRACT

In Today's technology driven world user profiles are the virtual representation of each user and they include a variety of user information such as personal, interest and preference data. These profiles are the outcome of the user profiling process and they are essential to service personalization. Different methods, techniques and algorithms have been proposed in the literature for the user profiling process. This paper aims to give an overview on the user profiling and its related concepts, and discuss the pros and cons of current methods for the future service personalization. Furthermore, it also give details about the simulations which have been carried out with well known classification and clustering algorithms with real world user profile dataset. This work is based on the doctoral thesis of the author.

General Terms:

User Profiling, Machine Learning

Keywords:

user profiling, user profile, personalization, classification, clustering

1. INTRODUCTION

Today there are numerous services available for the users across various electronic devices (e.g. smartphone, tablet computers). In this competitive market, user profiles become very important for service providers to achieve a successful service personalization. Personalized services aim to match users' requirements, preferences and needs with the service delivery. The success of these services relies on how well the service provider knows the user and how well this is reflected on the service. User profiles are the representation of the users and they are the outcome of the user profiling process.

There are two main challenges in user profiling process. These are the generation of an initial user profile for a new user and the continuous update of the profile information to adapt user's changing preferences, interests and needs. In literature two fundamental user profiling methods have been proposed to tackle these challenges. These are the content-based and the collaborative methods. Both of these methods have limitations and the hybrid user profiling has been proposed to overcome these limitations by combining these two. The user profiles that are created based on the traditional user profiling methods were not adequate to personalize different services. For this reason, various clustering and classification algorithms have been utilized to create more comprehensive user profiles. However, these profiles are lack in representing the multi-dimensionality of the user profiles and still not adequate to personalize different services.

This paper gives an explicit overview on the user profiling, presents simulation results that have been carried out with well known classification and clustering algorithms on real world user profile dataset, and discusses the ideal user profiling method for the future service personalization applications. The rest of this paper is organized as follows; In Section 2 background to the user profile, user profiling and personalization are given. Following this, in Section 3, user profiling methods are discussed. Section 4 presents the related works. Section 5 gives details about the well known classification and clustering algorithms and their simulation results with the user profile dataset. Section 6 presents a discussion for this paper. Finally, Section 7 concludes this paper.

2. BACKGROUND

A user profile is a set of information representing a user via user related rules, settings, needs, interests, behaviours and preferences [1]-[5]. This collection of personal information can either be represented as static data (e.g. native country) or dynamic data (e.g. needs). The content and amount of the information within a user profile can vary depending on the application area. However, regardless of the information, the accuracy of the user profile is based on how the user information is gathered and organized, and how accurately this information reflects the user. In other words, it depends on the user profiling process in which the information is gathered, organized and interpreted to create the summarization and the description of the user [5].

In the literature, there are two fundamental ways of retrieving information about the user. These are called explicit or implicit information gathering. In the explicit method, information regarding to the user's interest and preferences is provided explicitly by the user to the system. The downside of this method is that the explicit profiles have a static nature and are valid only until the user changes their interest and preferences parameters explicitly[6]. Explicit information gathering methods are used by the static profiling that analyzes the static and predictable characteristics of the user. Implicit information, on the other hand, is gathered dynamically by monitoring the users interactions with the system automatically. The implicitly created user profile is called implicit or dynamic user profile. Unlike static profiling, dynamic profiling uses the implicit method and analyzes user's behaviour pattern (e.g. usage history) to deter-

mine user's interests [7][8]. Here, the accuracy of the user profile depends on the amount of generated data through user-system interaction. It is also possible to produce a hybrid user profile which can be achieved in two ways [7]. The first way starts by using the explicit techniques to collect the initial data, followed by the implicit techniques to update the user profile. In the second way, on the other hand, implicit techniques are followed by the explicit techniques. In general, it has been cited that the hybrid methods are more efficient than both of the fundamental methods [7]. Table 1 [7] compares the aforementioned user profile types.

Personalization is a process to change the functionality, information content or distinctiveness of a system to increase its personal relevance to an individual [9]. Moreover, personalization is defined as the adaptation of the services in a way that they fit the user's interests, preferences and needs [9]-[15]. There are two types of personalization methods: implicit personalization and explicit personalization. In implicit personalization, information about the user for user profiles is gathered implicitly (e.g. click streams, scrolling, saving) [16]. Therefore, the user is unaware of the information gathering process. In explicit personalization, on the other hand, user profile information is gathered via direct involvement with the user (e.g. questionnaires, ratings and feedback forms) [16]. Here, the user is aware of the information gathering process. In implicit personalization, the accuracy improves with the continuous use of the system by the user. In explicit personalization, on the other hand, accuracy of the personalized information is based on manually provided information that is updated by the user.

3. USER PROFILING METHODS

Two fundamental user profiling methods are the content-based and the collaborative methods [2][17]. It is also possible to use the hybrid of these two methods. Following sub-sections will give detailed information about each of these methods and their techniques.

3.1 Content-Based Method

Content-based method, also referred as content-based filtering, assumes that the user show the same particular behaviour under the same circumstances [1][2][17]. Hence in this method, user's current behaviour is predicted from his past behaviour. In this scheme user profiles are represented similar with queries and the system selects the items that have high content correlation with the user profile. The content dependence is the main drawback of the content-based filtering. Therefore, this method performs badly if the item's content is very limited and cannot be analysed easily by the content-based filtering [7][17]. Furthermore, eclectic tastes and ad-hoc choices also cause bad performance as the provided recommendations are only based on the users previous choices [7]. For instance, a system that employs content-based approach can start recommending history books to a computer professional who usually buys IT books but happens to buy once a history book for his brother.

Following four sections will give information about the well known content-based techniques.

3.1.1 Vector-Space Model. Vector-Space Model (VSM) is a statistical-term based technique [18] and mostly used for the information retrieval. In this model, the contents of the documents are represented with vector/s of weighted terms [2]. Similar to the documents, the user profile is also represented as vector/s of weighted keywords/queries which reflects user's interests and preferences [2]. Here, weights indicate the importance of the term or keywords (i.e. how often the term appears in the particular document) [17]. The dimensions of the vectors are equal to the number of terms that are used to identify the content of the documents or the number of queries that are used to identify the user's interests and preferences [18]. User interests are represented either with a single vector that includes all the interest or with multiple vectors, which reflects interest in several domains [19]. In this model the effectiveness of the user profiles depends on the vectors degree of generalization. The VSM holds both synonym and polysemy issues which may cause unsuccessful detection of the relevant documents and incorrect selection of irrelevant documents. This model assumes that all terms and related concepts are orthogonal while in reality they are not as a result of synonym [18]. There are several methods to derive a weighted term representation of the documents or queries. Three of the main ones are Boolean, Term-Frequency (TF) and Term-Frequency Inverse Document Frequency (TF-IDF). Moreover, Cosine Similarity (CS) is commonly used to calculate the similarity between two weighted term vectors.

3.1.2 Latent Semantic Indexing. Latent Semantic Indexing (LSI) is also a statistical-term based technique. This method resolves the orthogonal problem of the VSM by examining the latent structure of a document and the terms within. Singular Value Decomposition (SVD) is one of the techniques that is used in LSI to identify patterns in the relationship between the terms and concepts within a document [18]. Unlike VSM, with the use of SVD, LSI retrieves relevant documents even though they do not have common terms with the user profile [18]. In this technique, the document is taken as a word by document matrix that is computed from the individual document vectors in the system which is obtained using the TF-IDF. This is followed by the reduction of the matrix's orthogonal dimensions to reduce the vector space [17][18]. Some of the well known methods to reduce the dimensions are the stop-word elimination, stemming and feature selection [17].

3.1.3 Learning Information Agents. Learning Information Agents (LIA) is one of the techniques that are used to incorporate Artificial Intelligence (AI) and Neural Networks (NNs) into the user profiling. In this technique, agents use the feedback of the user to update the user profile [18]. In general, agent technology provides an automated information gathering technique over the internet or any large information repositories (e.g. digital libraries) [3]. Application of the agent technology can be passive filtering of incoming messages (e.g. e-mail) or active information seeking (e.g. browsing assistant, digital libraries search). In LIA the normalised TF-IDF weighting is used to create the vector based representation of the document. In the user profile vector the weight of each keyword corresponds to the user preferences. The learning algorithm that is used by the information agent system uses the selection of documents and associated user evaluation (feedback (i.e. scoring)) to update the weights of the user preferences.

3.1.4 Neural Network Agents. Neural Network Agents (NNA), or Artificial Neural Networks (ANN), are used to incorporate the AI and NNs into the user profiling and like LIA, user profile updates are made based on the users feedback to the system. In this technique, user profile reflects the neural network that includes the nodes representing terms that are important for the user and edges representing the strength of association between the terms [17]. The terms in the network are the ones that occur within the documents that are accepted and rejected by the user. In NNA the terms are extracted by using the TF-IDF and they are used to create more comprehensive user profiles. Here, unlike LIA, the user does not have to score the document as the scoring is calculated by the system when

Table 1. Comparison of the User Profile Types

User Profile Type	Description	Techniques Used	Advantages	Disadvantages
Explicit User Profiles	User manually creates user profile	Questionnaires, Rating	Information gathered is usually of high quality	Requires a lot of efforts from user to update the profile information
Implicit User Profiles	System generates user profile from usage history of interactions between user and content	Machine learning algorithms	Minimal user effort is required and easily updatable by automatic methods	Initially requires a large amount of interaction between user and content before an accurate user profile is created
Hybrid User Profiles	Combination of explicit and implicit user profiles	Both explicit and implicit techniques	To reduce weak points and promote strong points of each of the techniques used	N/A

a user accepts or rejects the document. Terms are connected in the network if the same words are related throughout the documents [18].

3.2 Collaborative Method

Collaborative method, also referred as collaborative filtering method, assumes that the users who belong to the same group (e.g. of same age, sex or social class) behave similarly, and therefore have similar profiles [1][2][17]. The collaborative method is based on the rating patterns of similar users [2]. In this method people with similar rating patterns, or in other words people with similar taste, are referred to as 'like minded people' [2]. Unlike content-based method, collaborative method ignores the item's content and does recommendation of the items only based on the similar users' item ratings [17]. There are two main drawbacks of collaborative filtering which are the sparsity and the first-rater problem [7]. The sparsity is the situation when there is a lack of ratings available that is caused by an insufficient number of user or very few ratings per user [7][17]. Moreover, the first-rater problem, also referred as cold-start problem, can be observed when a new user has a deficient number of ratings [7][17]. Following section gives detailed information about the two well known collaborative techniques.

3.2.1 Memory-Based and Model-Based Techniques. Memory-based and model-based techniques enable users to filter the received information according to the ratings, which is the feedback given by the like minded users of the system [18]. Therefore, in these techniques the user can be provided recommendations from the categories which are not previously declared as interesting or relevant by the user but have received high ratings from the users with similar tastes. In these techniques, user's profile is a set of ratings that the user have given to a selection of items from the system database [2][18] (see Figure 1 [20]). As a result, the system's recommendation accuracy improves as the number of ratings increase in a user profile [18]. Systems based on memory-based technique estimate item's rating prediction for a particular user (active user/current user) based on the entire collection of previously given ratings by similar users [21]-[23]. There are number of algorithms applied to memory-based systems. The Mean Square Difference (MSD) is one of the popular algorithms where the MSD between the current user profile and all other profiles are calculated. If any user j of the system has MSD below the threshold then that user is considered to have similar taste with the current user i. Another popular algorithm to find the user similarity is the Pearson Correlation Coefficient (PCC). Different from memory-based systems, systems based on model-based technique use the collection of ratings to learn a model that will be used to estimate item

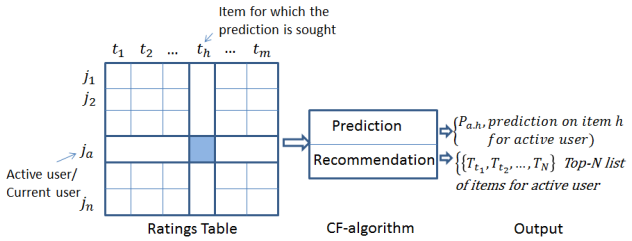

Fig. 1. Basic Principle of Collaborative Method

rating predictions [21][22]. Clustering and classification algorithms are commonly used to make item rating predictions in model-based systems [22][23]. These algorithms treat collaborative filtering as a classification or clustering problem.

3.3 Hybrid Method

A hybrid method, also referred as hybrid filtering method, uses content-based and collaborative methods to combine the advantages and overcome the limitations of both methods [7][17][24]. This method guaranties the immediate availability of a profile for each user. The system that employs the hybrid method provides a more accurate description of the user interests and preferences, as it continuously monitors and retrieves the user related information through the user-system interaction [1]. Generally, the hybrid method assigns the new user a default profile with the use of the collaborative method and further enhances the profile using the content-based method [1]. In the literature four hybrid user profiling techniques have been introduced [8]. These are called 'static content profiling', 'dynamic content profiling', 'static collaborative profiling', and 'dynamic collaborative profiling'. The static content profiling is the combination of static profiling and content based method. Here, the information about user's interests is gathered during registration. Consequently, in dynamic content profiling, information about user's interests are retrieved via monitoring user's behaviour. Moreover, in static collaborative profiling, information relating to user's interests is collected based on user's explicit requests. Here grouping of the users is done explicitly. In dynamic collaborative profiling, on the other hand, information gathering and grouping of users with similar behaviours is done based on the dynamic feedback from the users.
Table 2 [7] compares the aforementioned three main user profiling methods with respect to their techniques, advantages and disadvantages.

Table 2. Comparison of User Profiling Methods

User Profiling Method	Description	Techniques Used	Advantages	Disadvantages
Content-based Filtering	Filtering content from a data stream based on extracting content features that have been expressed in	Vector Space model, Latent semantic indexing, Learning information agents, Neural network agents	Objective analysis of large and/or complicated (e.g. multimedia) sources of digital material without much user involvement	1. Content dependent 2. Hard to introduce serendipitous recommendations as approach suffers from tunnel vision effect
Collaborative Filtering	Filtering items based on similarities between target users collaborative profile and peer user/group	Memory-based and Model-based	1. Content independent 2. Proves more accurate than content-based filtering for most domains of use enables introduction of serendipitous choices	1. Sparsity: poor prediction capabilities when new item is introduced to database due to lack of ratings 2. First-rater: poor recommendations made to new users until they have enough ratings in their profiles for accurate comparison to other users
Hybrid Filtering	Combines two filtering techniques	Collaborative Content based	To reduce weak points and promote strong points of each of the techniques used	Weak points can out-weight strong points if the hybrid is created naively

4. RELATED WORKS

Following subsections present a literature review of the user profiling research works.

4.1 Personalized Mobile Services

Personalized mobile services become very popular [25]-[30] and among those services, many systems have been developed for the tourist activities [25][26]. The moreTourism, which stands for mobile recommendations for tourism [25], is one of these systems and provides personalized tourist information (i.e. tourist attractions) for the users with similar interests. This hybrid system makes use of mashups along with social networks to enhance its users travelling experiences. To perform recommendation, the social content-based filtering compares the user tag cloud (collection of tags attached by the user) with the attraction tag cloud (collection of tags attached by the users to describe the attraction) and the social collaborative filtering creates one new tag cloud for each attraction using the tag clouds of the users who liked it. Hence, the recommendations are based on the user tag cloud, relationship among tags, location in time and space, and the nearby context. Similarly in [26], Fernandez *et al.* proposed a tourism recommender system that offers tourist packages (i.e. tourist attractions and activities) that best matches the user's social network profiles. Different from [25], the proposed hybrid system does recommendations based on both the user's viewing histories and the preferences in the social network.

4.2 Personalized Online Services

In the literature, various works has been carried out for the personalized online services [31]-[36]. In [32], Yeung *et al.* proposed a technique to analyse the personal data, personomies, within the folksonomies. Folksonomies are the user contributed data that are gathered via collaborative tagging system [32]. This work aimed to investigate how accurate the user profiles can be generated from the folksonomies and discusses how these profiles can be used for the web page recommendation. The proposed algorithm aimed to generate user profiles that were representing users multiple interests. The method was tested on the data which was taken from the 'www.delicious.com'. This data was the collection of bookmarks (documents) and tags that have been used by the users. Here, the VSM has been used for the term vector representation of tags, bookmarks and queries. In addition, the CS has been used to find the similarity between the bookmarks and the queries and the evaluation is done based on precision, recall and F15 (harmonic mean of precision and recall) measures. In another work, Park *et al.* [31] proposed a hybrid framework for online video recommendations where the recommendations are done according to the similar viewing patterns. In this work, user profiles are constructed as an aggregate of tag clouds, also known as global tag cloud, of videos. Here, both user profiles and videos were represented with tag cloud vectors. The cloud-based CS was employed to compute the user similarity. The user's profile is updated every time the user plays a video, by including the global tag cloud of the video into the user's tag cloud. Park *et al.* argued that different from the previous hybrid methods, this approach is based on the implicit users' view transaction data instead of the explicit ratings data. Another hybrid framework has been proposed in [33]. Different from the works described above, in [33] collaborative filtering was employed together with techniques from the Multi Criteria Decision Analysis (MCDA) for item recommendation. In this study user profiles were included with user's numerical ratings and ranking order, and represented as vectors. The user profile is updated with a feedback mechanism, which is activated by the user when he/she is willing to rate an item after a recommendation. In this system, the MCDA was used to find the similar users while collaborative filtering was used to recommend items. In [37] Lee *et al.* is focused on the use of social network data and utilizes content-based method for user profiling. In this work, users' twitter timelines has been used to create user profiles for news article recommendation. Here, user profiles are represented as normalized weighted keyword vectors where keywords extracted from users timelines information that included tweets, re-tweets and hashtags. Decision on which articles to be recommended in which order is decided based on the similarity between user profile vector and news article vector. Like in many aforementioned works, the CS has been used for this similarity calculation. In this work, the prediction accuracy of news recommendation was measured in terms of hit ratios.

4.3 Personalized Television Services

There has been a considerable amount of work for personalized program and advertisement recommendations for television (i.e. for Internet Protocol Television (IPTV) and Integrated Digital Televi-

sion (iDTV)) users [38]-[41]. In [38], a hybrid TV program recommender system, gueveo.tv, has been proposed. According to the Martinez *et al.*, the proposed system works well because both methods are complement with each other in a way that the content-based method recommends usual programs and collaborative method provides the discovery of new shows. In this study, each user represented with user's preference profile that contains two types of information that are domain preferences (i.e. list of available TV channels, preferred viewing times) and program preferences (i.e. subject keywords or tags). This information was gathered via implicit (i.e. monitoring viewing times) and explicit methods (i.e. filling questionnaire). In gueveo.tv, VSM has been employed to generate a vector representation of the user profile and programmes viewed. Here, cosine measure is used to calculate the similarity between the program vectors and the user profile vectors.

4.4 Classification and Clustering algorithms for User Profiling

Different from the traditional content-based and collaborative techniques, classification and clustering algorithms have also been used for the user profiling. In [42], Irani et al. focused on the social spam profiles in MySpace. Here, authors compared well known machine learning algorithms (AdaBoost algorithm, C4.5, Support Vector Machine (SVM), Neural Networks (NNs), Naive Bayesian (NB)) with respect to their abilities to distinguish spam profiles from legitimate profiles. In this study, each user was represented with a social network profile where each profile included two kinds of data which were categorical data (i.e. sex, age, relation- ship status) and free-from data (text information i.e. about me, interests). In another work [43], Paireekreng and Wong investigated the use of clustering and classification of user profile at the client-side mobile. Here, the authors focused on the content personalization to help mobile users retrieve information and services efficiently. In their proposed two phase framework, clustering was used to construct a user profile, while classification was used to classify user profile based on the class information from clustering. In this work, K-means, Two Step, Anomaly and Kohenen clustering algorithms were compared for clustering. Moreover, Locally Weighted Learning (LWL), RepTree, Decision Table and SVMReg classifiers were compared for classification. Previous works [44], [45], [46] and [47] have been the first in the literature to present the comparison of the classification and clustering accuracy performance of different algorithms with user profiles. In [45] NB, Instance Based Learner (IBL), Bayesian Network (BN) and Lazy Bayesian Rules (LBR) classifiers were compared using a user profile dataset. Furthermore in [46], Decision Tree (DT) algorithms to be used for user profiling (i.e. Classification and Regression Tree (SimpleCART), NBTree, Id3, J48 -a version of C4.5- and Sequential Minimal Optimization (SMO)) were included and compared with large user profile data. Following this, in [48] a variant of the IBL algorithm, namely Weighted Instance Based Learner (WIBL), for the user profiling has been proposed. The WIBL uses a modified version of the Per-Category Feature Weighting algorithm to assign weights to the features during the user profiling process. In [47] comparison of the performance of WIBL with other well known clustering algorithms have been done. In this study simulation results showed that WIBL outperforms the well known clustering algorithms.

5. SIMULATIONS

This section gives details about the simulations that have been carried out with well known classification and clustering algorithms in the literature. Here, user and instance, and also feature and attribute are used interchangeably.

5.1 Dataset

For these simulations user profile dataset within the "restaurant and consumer" [49] has been used. Different from the datasets that have been used in author's previous studies, this dataset originally created as a user profile dataset and includes the information of 138 instances. Here, each user is represented with 16 attributes that are 'activity', 'dress preference', 'personality', 'interest', 'smoker', 'drinking level', 'ambiance', 'transport', 'marital status', 'hijos', 'birth year', 'religion', 'color reference', 'weight', 'budget' and 'height'. For this study, longitude, latitude and user-id attributes have been removed from the original user profile dataset.

5.2 Algorithms

For this work nine classification algorithms and five clustering algorithms have been tested. Below paragraphs briefly describes each of these algorithms. Here test instance refers to a new unclassified/unclustered instance while training instance is the already classified/clustered instance.

The BN is one of the well known classification algorithms that is named after Thomas Bayes, founder of the Bayesian methods. BNs are probability values, which are based on and used for the reasoning and the decision making in uncertainty where such reasoning heavily relies on Bayes rule [50]. The NB classifier is one of the BN algorithms. However, unlike BN, NB classifier assumes that all attributes within the same class are independent, given the class label. The NBTree classifier is one of the hybrid classifier where it generates a DT with NB classifiers at the leaves. This classifier holds the advantages of both DT and NB classifiers. Another DT classification algorithm is the J48. The J48 classification algorithm is the enhanced version of C4.5 and has been developed to generate a pruned or un-pruned C4.5 DT [51][52]. The SVMs are the supervised learning methods that are used for the classification. These methods perform classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories [53][54]. The SMO classifier implements SMO algorithm for the training of a SVM [52]. The IBK is one of the well known IBL algorithm where K closest instances are retrieved and the label of the majority class among these instances is assigned as a class label for the test instance [55][56]. The LWL algorithm is a weighted IBL that assigns weights to instances using IBL and uses these locally weighted training instances for classification [57]. The Kstar (K*) is another IBL [58] which aims to provide a consistent approach to handle symbolic attributes, real valued attributes and missing attributes [59]. K* is based on the entropy distance measure where the distance between two instances is defined as the complexity of transforming one instance into another [58]. The Voting Feature Intervals (VFI) [61] algorithm considers each feature separately as each feature participates in the classification process by distributing real-valued votes among classes. The class receiving the highest vote is declared to be the predicted class.

The single-linkage, complete linkage and average-linkage are the well known hierarchical clustering algorithms. In single-linkage clustering, the resulted distance between two clusters is equal to the shortest distance from any member of one cluster, to any member of the other cluster [62]. In this algorithm the shortest distance reflects the maximum similarity between any two data objects in two different clusters. In complete-linkage clustering, on the other hand, the distance between two clusters is the maximum distance from any data object of one cluster to any data object of the other cluster [62].

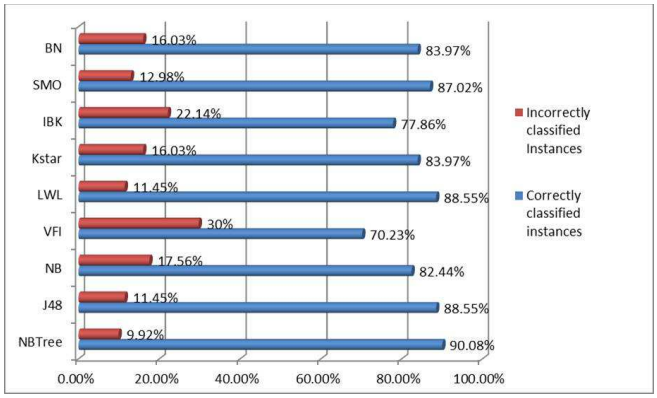

Fig. 2. Classification Performance

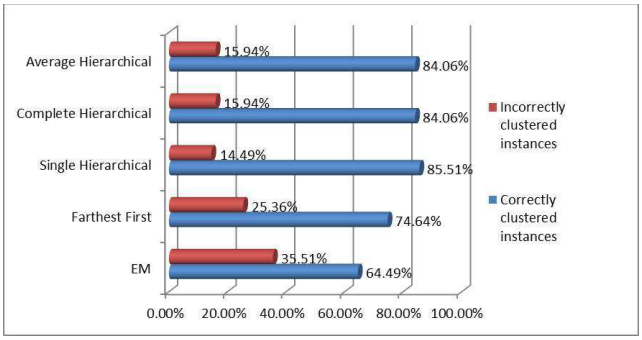

Fig. 3. Clustering Performance

In average-linkage clustering, the distance between two clusters is equal to the average greatest distance of all paired data objects of these clusters. The farthest first [63] proposed by the Hochbaum and Shmoys where, in turn, each cluster center is put at the furthest from the existing cluster center. This furthest point has to be within the data area. Expectation Maximization (EM) [64] is a iterative method that assigns a probability distribution to each instance which indicates the probability of the instance belonging to each of the clusters.

5.3 Simulation Results

All simulations have been carried out on WEKA [60] platform and above given dataset (see Section 5.1.) has been used as both training and test dataset. All algorithms have been trained and tested on the same user profile dataset. Test mode has been set as '10 fold cross validation' for the classification and 'classes to clusters' for the clustering simulations. The simulations have been repeated for all above mentioned classification and clustering algorithms (see Section 5.2.). Figure 2 shows the simulation results for the classification algorithms. From this figure it can be seen that the best result is achieved by the NBTree (90.08%) algorithm. This result is followed by the J48 (88.55%), LWL (88.55%) and SMO (87.97%) classifiers. Although the lowest result is achieved by the VFI (70.23%), overall performance of all classification algorithms were above 70%. For these simulations 'activity' attribute has been chosen to be the class attribute. Figure 3 shows the simulation results for the clustering algorithms. From this figure it can be seen that the best result is performed by the single-linkage clustering algorithm where 85.51% of the instances clustered correctly. The results also showed that, for this dataset, both complete and average linkage algorithms clustered the same amount of users correctly. The worst performance is achieved by the EM clustering algorithm where 35.51% of the users are clustered incorrectly. In general classification algorithms achieved better accuracy than the clustering algorithms. However, the best clustering accuracy results are very close to the best classification results. Referring back to author's previous studies, it can be said that NBTree classifier still archives one of the best results among the well known classifiers with user profile dataset.

6. DISCUSSIONS

From Section 4 it can be seen that collaborative and content-based methods have been widely used for the various personalization applications. In these applications, the content-based systems have mostly been designed to recommend text-based items (i.e. articles) via predicting ratings or the relative preferences of the user. In these systems, user profiles are mostly described with keywords obtained by analysing the items which have been previously seen, used (i.e. tweets) or rated by the user. These applications also showed that the user profile can be represented as a vector of weighted keywords, where the CS is commonly used to find the similarity between two vectors (i.e. user profile vector and news article vector). The collaborative systems, on the other hand, are mostly used for e-commerce websites. These systems consider similar buying behaviour of the customers to estimate users' preferences on items. In these systems, user profiles are the collection of the ratings of items which other users have already rated. Here, CS and PCC similarity measurement techniques are widely used to identify the similarity between users. Both content-based and collaborative systems use the CS. In content-based it is used to find the similarity between the term vectors, while in collaborative systems it is used to find the similarity between the vectors of actual user ratings. As previously discussed in Section 3. (also see Table 2), both collaborative and content-based methods have limitations and hybrid systems have been proposed to overcome these limitations by combining both methods. However, it has been observed from the hybrid systems that the content of the user profiles are just maintained. In recent years, tag aggregation based personalization has received considerable attention. In these studies user profiles are represented with tag clouds which are the collection of tags attached by the users. It can be argued that this way of representation tackled the aforementioned sparsity and first-rater problems of collaborative method (see Section 3.) as similarities between users does not have to be calculated based on users common ratings. Moreover, tags make it unnecessary to analyse the content of the web page, image, video or advertisement, which can be a difficult process to build user profiles. It can be argued that this offers a solution to a content dependence limitation of the content-based method (see Subsection 3.1.). However, in these systems, the quality of the user profiles rely on the number of users participating in tagging, how accurately tags represent the content and the number of tags the user used that are produced by others. It can be said that, tag cloud based user profiles reflect the web content more than user itself.

Accurate user profiles are important to both the user and the service provider. From the user point of view, it is important for the personalized services not to be misrepresented. For the service providers, on the other hand, it is the way to achieve optimum user satisfaction by providing accurate personalized services. It can be seen that the literature on user profiling focused on the usage of features such as ratings, items, keywords and simple demographics to represent each user. Although this traditional way of profiling works well for

specific services, it lacks in representing the multi-dimensionality of the user profiles accurately. For example, user profiles that reflect the ratings which were given to music videos cannot be used to recommend restaurants for the same user. This constraint motivates the need to conduct more advance profiling to build a more comprehensive profiles which reflects different user information such as users interest, preferences and demographics. This way of profiling can provide user related information that can be used by various third party service providers to personalize different services. To be able to use the multi-dimensional profiles effectively, feature weighting should be taking into account. Feature weighting is essential for accurate user profiling because the relevancy of all user profile information is not the same for different service personalization. For instance, users book interest information may not be as relevant as income information of the user for personalized restaurant recommendations. Using weights to make the distinction between relevant and irrelevant information can provide a solution for this problem. In summary it can be said that weighted multi-dimensional user profiling could be the new profiling method for the future service personalization. Although in [48] authors proposed WIBL for this purpose, there are other potential feature weighting algorithms that could be used with IBL to achieve much better accuracy with multi-dimensional user profiles.

7. CONCLUSION

User profiles represent users and they reflect each user's preferences, needs, behaviours and interest. These profiles are the outcome of the user profiling process and they are essential for the service personalization. This paper presents a review on user profiling including its related concepts, methods, techniques, as well as the existing solutions in the literature. In addition, it discusses the pros and cons of the user profiling methods together with commonly used techniques and user profile information. The paper also presents simulations that were carried out with well known classification and clustering algorithms with real world user profile dataset and discusses how the traditional way of profiling lacks in representing the multi-dimensionality of the user profiles accurately. Finally, the paper talks about the ideal user profiling method which could be the solution for the future service personalization applications. As a future work, the author would like to continue working on this ideal method by focusing on other feature weighing algorithms to achieve more accurate multi-dimensional user profiles which could be used for the personalization of different services.

8. REFERENCES

[1] G. Araniti, P. D. Meo, A. Iera and D. Ursino (2003). Adaptive controlling the QoS of multimedia wireless applications through user profiling techniques, IEEE Journal on selected areas in communication, 21(10), pp. 1546-1556.

[2] T. Kuflik and P. Shoval (2000). Generation of user profiles for information filtering-research agenda, Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 313-315.

[3] M. J. Martin-Bautista, D. H. Kraft, M. A. Vila, J. Chen and J. Cruz (2002). User profiles and fuzzy logic for web retrieval issues, Soft Computing (Focus), 15(3-4), pp. 365-372.

[4] European Telecommunications Standards Institute (ETSI) (2005). Human Factors (HF); User Profile Management, pp.1-100, Available: http://www.etsi.org/

[5] S. Henczel (2004). Creating user profiles to improve information quality, Factiva, 28(3), p. 30.

[6] C. Gena (2005). Methods and techniques for the evaluation of user-adaptive systems, The Knowledge Engineering, 20(1), pp. 1-37.

[7] M. Khosrowpour (2005). Encyclopaedia of information science and technology, Electron. Book, Hershey, PA Idea Group Reference, pp. 2063-2067.

[8] D. Poo, B. Chng and J. M. Goh (2003). A hybrid approach for user profiling, Annual Hawaii International Conference on System Sciences, 4(4), pp. 1-9.

[9] J. Blom (2000). Personalization-a taxonomy, Conference on Human Factors in Computing Systems, pp. 313-314.

[10] I. Jorstad, D. V. Thanh and S. Dustdar (2004). Personalization of Future Mobile Services, International Conference on Intelligence in Service Delivery Networks.

[11] I. Jorstad, D. V. Thanh and S. Dustdar (2005). The personalization of mobile services, IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, 4, pp. 59-65.

[12] H. Stormer (2004). Personalized websites for mobile devices using dynamic cascading style sheets, International Conference on Advances in Mobile Multimedia, pp. 351-360.

[13] E. Lillevold and J. Noll (2004). Personalization in telecom business, European Institute for Research and Strategic Studies in Telecommunications, Available: http://archive.eurescom.eu.

[14] R. Guarneri, A.M. Sollund, D. Marston, E. Fossbak, B. Berntsen, G. Nygreen, G. Gylterud, R. Bars and A. Kerdraon (2004). Report of state of the art in personalisation, Common Framework, pp. 1-59, Available: http://www.ist-eperspace.org/deliverables/D5.1.pdf

[15] P. S. Yu (1999). Data mining and personalization technologies, International conference on database systems for advance applications, pp. 6-3.

[16] D. Kelly and J. Teevan (2003). Implicit feedback for inferring user preference: a bibliography, ACM Special Interest Group on Information Retrieval (SIGIR) forum, 37(2), pp. 18-28

[17] D. Godoy and A. Amandi (2005). User profiling in personal information agents: a survey, The Knowledge Engineering Review Journal, 20(4), pp. 329-361.

[18] S. Steward and J. Davies (1997). User profiling techniques: a critical review, British Computer Society, BCS-IRSG Annual Colloquium on IR Research, pp. 1-22.

[19] D. Godoy and A. Amandi (2005). User profiling for web page filtering, IEEE internet computing, 9(4), pp. 56-64.

[20] E. J. Neuhold (2003). Personalization and user profiling & recommender systems, WI/IM Information Management Proseminar, pp. 1-25.

[21] H. Luo, C. Niu, R. Shen and C. Ullrich (2008). A collaborative filtering framework based on both local user similarity and global user similarity, Springer Computer Science Machine Learning, 72(3), pp. 231-245.

[22] G. Adomavicius and A. Tuzhilin (2005). Towards the next generation of recommender systems:a survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering, 17(6), pp. 734-749.

[23] X. Su and T. M. Khoshgoftaar (2009). A survey of collaborative filtering techniques, Advances in Artificial Intelligence, p.p. 1-19.

[24] M. Khosrowpour (2006). Encyclopaedia of ecommerce, e-governments, and mobile commerce, Electron. Book, Hershey, PA Information science Reference, pp. 118-123.

[25] M. R. Lopez, A. B. B. Martinez, A. Peleteiro, F. A. M. Fonte and J. C. Burguillo (2011). moreTourism:mobile recommendations for tourism, IEEE International Conference on Consumer Electronics, pp. 347-348.

[26] Y. B. Fernandez, M. L. Nores, J. J. P. Arias, J. G. Duque, M.I.M. Vicente (2011). TripFromTV+:Exploiting social networks to arrange cutprice touristic packages, IEEE International Conference on Costumer electronics, pp. 223-224.

[27] C. K. Georgiadis and S. H. Stergiopoulou (2008). Mobile commerce applications development: implementing personalized services, International Conference on Mobile Business, pp. 201-210.

[28] W. Woerndl, C. Scheuller and R. Wojtec (2007). A hybrid recommender system for context-aware recommendations of mobile applications, IEEE International Conference on Data Engineering Workshop, pp. 871-878.

[29] H. Jeon, T. Kim and J. Choi (2008). Mobile semantic search personal preference filtering, International Conference on Networked Computing and Advanced Information Management, pp. 531-534.

[30] C. Biancalana, F. Gasparatti, A. Micarelli and G. Sansonetti (2011). Social tagging for personalized location-based services, International Workshop on Social Recommender Systems, pp.1-9

[31] J. Park, S. J. Lee, S. J. Lee, K. Kim, B. S. Chung and Y. K. Lee (2011). Online video recommendation through tag-cloud aggregation, IEEE Multimedia, 18(1), pp. 78-87.

[32] C. A. Yeung, N. Gibbins and N. Shadbolt (2008). A study of user profile generation from folksonomies, Workshop on Social Web and Knowledge Management, pp. 1-8.

[33] K. Lakiotaki, N. F. Matsatsinis and A. Tsoukias (2011). Multicriteria user modelling in recommender systems, IEEE Intelligence Systems, 26 (2), pp. 64-76.

[34] R. V. Meteren and M. V. Someren (2000). Using content-based filtering for recommendation, Workshop on Machine Learning in the New Information Age, pp. 312-321.

[35] Y.W. Park and E.S. Lee (1998). A new generation method of a user profile for information filtering on the internet, International Conference on Information Networking, pp. 261-264.

[36] G. Specht and T. Kahabka (2000). Information filtering and personalization in databases using gaussian curves, International Symposium on Database Engineering and Applications, pp. 16-24.

[37] W. J. Lee, K. J. Oh, C. G. Lim and H. J. Choi (2014). User profile extraction from twitter for personalized news recommendation, International Conference on Advanced Communication Technology, pp. 779-783.

[38] A. B. B. Martinez, M. R. Lopez, E. C. Mantenegro, J. C. Burguillo, F. A. M. Fonte and A. Peleteiro (2010). A hybrid content-based and item-based collaborative filtering to recommend TV programs enhanced with singular value decomposition, Elsevier Information Sciences: an International Journal, 180(22), pp. 4290-4311.

[39] Z. S. Shibeshi, S. Ndakunda, A. Terzoli and K. Brandshow (2011). Delivering a personalized video service using IPTV, International Conference on Advanced Communication Technology, pp. 1489-1494.

[40] M. Kodialam, T.V. Lakshman, S. Mukherjee and L. Wang (2011). Online scheduling of targeted advertisements for IPTV, IEEE/ACM Transactions on Networking, 19(6), pp. 1825-1834.

[41] T. Pessemier, T. Deryckere, K. Vanhecke and L. Martens (2008). Proposed architecture and algorithm for personalized advertising on iDTV and mobile devices, IEEE Transactions on Consumer Electronics, 54(2), pp. 709-713.

[42] D. Irani, S. Webb and C. Pu (2010). Study of static classification of social spam profiles in MySpace, International Conference on Weblogs and Social Media, pp. 82-89.

[43] W. Paireekreng and K. W. Wong (2009). Client-side mobile user profile for content management using data mining techniques, International Symposium on Natural Language Processing, pp. 96-100.

[44] A. Cufoglu, M. Lohi and K. Madani (2008). A comparative study of selected classification accuracy in user profiling, International Conference on Machine Learning and Applications, pp. 787-791.

[45] A. Cufoglu, M. Lohi and K. Madani (2008). classification accuracy performance of Nave Bayesian (NB), Bayesian Networks (BN), Lazy Learning of Bayesian Rules (LBR) and Instance-Based Learner (IB1) - comparative study, International Conference on Computer Engineering and Systems, pp. 210-215.

[46] A. Cufoglu, M. Lohi and K. Madani (2009). A comparative study of selected classifiers with classification accuracy in user profiling, World Congress on Computer Science and Information Engineering, pp. 708-712.

[47] A. Cufoglu, M. Lohi and C. Everiss (2013). Clustering Algorithms and Weighted Instance Based Learner for User Profiling, International Conference on Advances in Information Mining and Management, pp.7-11.

[48] A. Cufoglu, M. Lohi and C. Everiss (2012). Weighted Instance Based Learner (WIBL) for user profiling, International Symposium on Applied Machine Intelligence and Informatics, pp. 201-205.

[49] B. V. Govea, J. G. G. Serna, R. P. Medellan (2011). Effects of relevant contextual features in the performance of a restaurant recommender system. In :ACM RecSys 2011: Workshop on Context Aware Recommender Systems.

[50] F.V. Jensen (1993). "Introduction to Bayesian network". Denmark, Hugin Expert A/S.

[51] M. Panda and R. M. Patra (2008). A comperative study of data mining algorithms for network intrusion detection. International Conference on Emerging Trens in Engineering and Technology, pp. 504-507.

[52] H. W. Ian and E. Frank(2005). "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.

[53] L. Liu, Z. Liz and H. He (2008). The research of decision support vector machine in web information classification, International Conference on Computer Supported Cooperative Work in Design, pp. 196-200.

[54] K.P. Bennettand and J. A. Blue(1998). Support vector machine approach to decision trees. International Conference on Neural Networks, pp. 2396-2401.

[55] D. W. Aha, D. Kibler and M. K. Albert (1991). Instance-based learning algorithms, Machine Learning Journal, 1(6), pp. 37-66.

[56] O. Gomez, E F. Morales and J. A. Gonzales (2007). Weighted instance-based learning using representative intervals, Mexican International Conference on Advances in Artificial Intelligence, pp. 420-430.

[57] C. G. Atkeson, A. W. Moore and S. Schaal (1997). Locally weighted learning, Artificial Intelligence, pp.11-73.

[58] J. G. Clear and L. E. Trigg (1995). K*: An instance-based learner using an entropic distance measure, International Conference on Machine Learning, pp. 108-114.

[59] G. Demiroz and H. A. Guvenir (1996). Genetic algorithms to learn feature weights for nearest neighbour algorithm, Belgian-Dutch Conference on Machine Learning, pp. 117-126.

[60] I. H. Witten, E. Frank and M. A. Hall (2011). "Data mining practical machine learning tools and techniques" 3rd Edition, Morgan Kaufmann, USA pp. 472-550.

[61] G. Demiroz and H. A. Guvenir (1997). Classification by Voting Feature Intervals, Conference on Machine Learning, pp.85-92.

[62] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Cluster Analysis", 5th Edition, John Wiley and Sons, Ltd.(London), 2011, pp. 73-78.

[63] D. S. Hochbaum and D. B. Shmoys (1985). A best possible heuristic for the k-center problem, Mathematics of Operations Research, 10(2), pp. 180-184

[64] A. P. Dempster, N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM Algorithm, Journal of the Royal Statistical Society. Series B (Methodological),39(1), pp.1-38.