A Framework for Guiding the Museum Tours Personalization

Mykola Pechenizkiy and Toon Calders

Department of Computer Science,
Eindhoven University of Technology,
P.O. Box 513, 5600MB Eindhoven, the Netherlands
{m.pechenizkiy,t.calders}@tue.nl

Abstract. The importance of web access personalization to cultural heritage has been recognized by many museums. Recent research in recommender systems, information retrieval and data mining has facilitated the development of personalized applications. This research provides a number of intelligent technologies for supporting user navigation, information filtering, and other important processes of a user-centered interactive information exchange between museum websites and their visitors. In this paper we study some of the challenges of personalizing museum tours. We focus on (i) the efficient learning of user models in offline and online settings, and (ii) the scientific evaluation of a personalization effect. Given few examples of the preferences of a user, the system must learn to suggest the most relevant artworks. This setting is particularly important for the development of an interactive museum tour. An additional challenge here is to learn user preferences quickly to be able to start recommending relevant artworks as soon as possible. We review the basic approaches that have been (or potentially can be) used for personalizing the access to the cultural heritage. We also present a formal framework for evaluating online and off-line learning user models. In the off-line setting, the user provides the system with some labeled examples from which the system must generalize. In the online setting, however, the system can interactively query the user's interest. Then, we discuss the challenges of scientific evaluation of personalization techniques and personalization process. Finally, we present a methodological framework for guiding a museum tour personalization and discuss the potential utility of this framework.

Keywords: personalization, adaptation, user modeling, recommender systems, data mining, scientific evaluation, web-access to cultural heritage

1 Introduction

Recent advances in IT affect our everyday life in various aspects, providing access to different educational entertainment, scientific and other information resources in a new manner. Traditional museums and galleries are not an exception and an increasing number of artworks are becoming available in the digital form via the Internet to the potential visitors and simply interested people.

It rapidly became evident that this new form of access and presentation of traditional resources opens up new possibilities for providing more personalized and thus more effective service. The development of a personalized access to the cultural heritage resources has become an increasingly significant trend in the museum world too [6], [14], [15] where many national and smaller-size museums create virtual environments providing access to their collections. An obvious enhancement of such service is providing personalized tours and suggestions of additional artworks and collections to browse, after having observed the virtual visitor's reaction to what (s)he has already seen.

In general, cultural heritage personalization aims first at assisting visitors in the selection and filtering of material (like artworks and corresponding information) without spending too much time to looking for that material. This assistance may also assume the improved usability of a virtual museum's navigation.

Such web access personalization to museums helps addressing the growing amount of digital data and the growing diversity of visitors, which may differ in age, level of education, learning style and prior knowledge. Considering the different interests and preferences of each visitor individually, personalization addresses the real challenge of turning the museum monologue ("talking *to* the visitors") into a dialogue ("talking *with* the visitors") [6].

Overall, providing a more tailored service, scientifically justified personalization helps to respond to the educational and marketing needs of museums if and only if the systems are implemented in a clear and easy manner and is not overly intrusive towards the visitors. In this way, satisfied visitors are stimulated more to come back to reuse the system and to encourage other people to try it as well [6].

The basic challenge behind providing a personalized access is to tailor it to a visitor's (potentially changing) interests and preferences[1] without demanding them to express them explicitly. One straightforward approach to model the interests of the user, is to ask him or her to rate a collection of selected artworks. As a result the user also learns what aspects of art interest them without a prior understanding of how experts describe them. It is, however, usually more desirable to start offering the recommendations to the visitors as soon as possible, hence minimizing intrusiveness to the users.

In this paper we focus on the efficient learning of user preferences in offline and online settings and the scientific evaluation and further enhancement of the learning of user preferences and recommendations.

The organization of the paper is as follows. In Section 2, we introduce the problem of personalization and adaptation of cultural heritage content and we describe the related work. In Section 3, the problem of the museum tour personalization is formally defined; we discuss also issues related to the learning of the user model. In Section 4, we discuss the challenges of scientific evaluation of a personalization effect, illustrate the basic personalization evaluation methodologies, and present a methodological framework for guiding the evaluation of cultural heritage personalization evaluation and enhancement. We conclude with a brief summary and discussion of our further research plans in Section 5.

2 Personalization and Adaptation of Cultural Heritage Content

In this section we present an overview of recommendation systems and related approaches that are used in cultural heritage applications for facilitating the personalized access to a collection of artworks. We discuss the basic principles behind these approaches and their inherent limitations.

2.1 Personalization process

According to the type of adaptation, different systems are usually categorized into customizable (adaptable) and personalized (or adaptive). Although the border is not always clear, in general, customization (or adaptability) assumes active user participation (a visitor has a possibility to configure the application) and explicit input (manually creating and/or editing an own profile). Thus, a visitor has explicit ways of controlling the outlook and content of a virtual museum. In personalized applications, on the contrary, not a visitor, but the system (we will use personalized systems and recommender system as synonyms here) is responsible for automatic personalization of structure, content and its outlook according to a visitor's preferences (often called user model or user profile), which can either also be learnt automatically by the system, or, alternatively, the necessary information can be explicitly provided by the visitor. Here, we consider the situation of a personalized system that is aimed at learning user preferences and providing recommendations automatically.

The key issues in providing personalized access to cultural heritage content are (1) understanding who is the user and what kind of content is of his or her interest, through a user modeling process that often consists of some relevant data collections, its analysis and the transformation to actionable knowledge; (2) delivering the personalized content, and (3) measuring and evaluating the impact of personalization on the visitor's satisfaction, in particular, and on achieving goals defined by the resources provider in general. Adomavicius and Tuzhilin in [3] consider personalization as an iterative process defined by the three stages of the *understand-deliver-measure* cycle. (We present a simplified view of providing personalized access to cultural heritage in Figure 1.)

[1] We focus here on satisfying visitors *interests* and *preferences* rather than a single interest or preference.

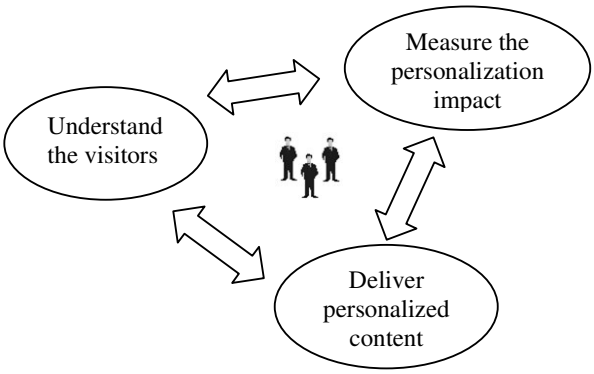

Fig. 1. A simplified view of providing personalized access to cultural heritage.

Personalized information may be presented in several forms, including narratives, ordered lists, etc. According to the policy of delivering personalized content, different methods are categorized into pull (notifying user that personalized information is available but display it only after an explicit request), push (such as sending e-mail), and passive (displaying personalized information as by-products) methods, later being the most frequent way of delivering personalization in virtual museums.

The performance of these delivery and presentation methods depends on various factors. The quality of underlying matchmaking (for example rule-based or statistics-based) technologies, however, is of prior importance, and therefore the development of the matchmaking technologies for recommendations has been under active research.

In the next section we briefly review the major types of techniques that facilitate the learning of visitors' preferences.

Adomavicius and Tuzhilin in [3] stress also the importance of vertical personalization research and the need for solid design principles for integrating all stages of the personalization process. The three most important issues in developing these principles are: (1) develop good metrics to determine personalization impact; (2) study the feedback-integration problem and develop novel methods to address it; and (3) investigate the goal-driven design process in order to achieve better personalization solutions. We will touch on these issues later on, in Sections 3 and 4.

2.2 Basic approaches for personalization

Personalization methods are often classified into broad categories, according to their recommendation approach and algorithmic techniques.

Recommender systems enhance user access to relevant information by using techniques such as collaborative filtering, content-based filtering, and hybrid approaches (see for example [2] for the recent survey of the state-of-the-art).

Collaborative-based methods search for peers of a visitor that have similar known preferences and then recommend those items that were most liked by the peers. Because collaborative systems rely solely on preferences of the visitors to make recommendations, new artworks added into the museum's collection will not be recommended until a substantial number of users have rated it. Another major problem with applying collaborative-based methods is the sparsity problem; i.e., data that reflects preferences of the users is sparse and insufficient to identify similarities. Huang *et al.* in [9] propose the associative retrieval framework to addresses this problem.

Content-based methods analyze the common features among the items a visitor liked and recommend those items that have similar features. Acquiring the preferences of a user in an efficient way has been recognized to be a bottleneck with content-based methods.

The main problem of applying content-based techniques for museum tour personalization is that both automatic feature extraction from graphical images and manual assignment of these features are difficult. Because of these difficulties, a recommender system usually has a rather limited set of features, which are explicitly associated with the artworks through semantic annotation (using

for example RDF format), and which therefore may not tell about the quality, originality, uniqueness, etc. of some artwork. An additional problem here is that, two different artworks with the same set of features (with similar values), are indistinguishable for the recommender system.

Besides, a recommender system that employs a content-based technique suggesting artworks that score highly against a visitor's profile tends to be biased towards showing only those artworks that are similar to those already rated by that visitor. However, the diversity of recommendations is often a desirable feature in recommender systems. Thus, if a visitor liked *Nightwatch*, it should not lead the recommender system to recommend all the artworks by Rembrandt one-by-one, even if the visitor rates each of them highly, since in a limited amount of time he or she might like to enjoy a range of different artworks.

Content-based methods are usually classified into two major categories according to the employed algorithmic technique: memory-based (heuristic-based), and model based [5].

Memory-based algorithms, predict ratings based on memorizing (therefore, also known as *lazy-learners*) and searching (therefore, also known as *heuristic-based algorithms*) the entire collection of previously rated artworks by the visitors. They compute an aggregate of the ratings of several other most similar visitors for the same item. Beside traditional correlation-based and cosine-based techniques for measuring similarity, recently more sophisticated approaches were employed. For example, [2] refer to default voting, inverse user frequency, case amplification, and weighted-majority prediction as extensions to the traditional techniques.

Model-based algorithms on the contrary use the collection of ratings to learn a model (employing for example some statistical and machine learning techniques), which is then used to make predictions. For example, a probabilistic approach to collaborative filtering calculates unknown ratings as conditional probabilities (given the ratings of previously rated items). In [10] it was demonstrated with a simple probabilistic model that collaborative filtering is valuable, also when relatively little data for each user is available. It is a common approach to perform clustering of visitors to facilitate some group-based personalization (see for example [17][20]).

Some empirical studies suggest that model-based approaches may produce more accurate recommendations than memory-based approaches. However, a combination approach may be beneficial here as well. For example, Pennock and Horvitz in [11] demonstrated empirically that combining both memory-based and model-based approaches can result in better recommendations.

Hybrid approaches combine collaborative and content-based methods in order to overcome the limitations of these methods. Three major ways of combining collaborative and content-based methods are: (i) combining recommendations (for example with some voting or selecting mechanism [7]) separately produced with a content base and a collaborative-based technique; (ii) introducing some elements of one type of techniques to another (for example reducing the dimensionality and sparseness of the ratings matrix [7]), and (iii) developing a generic model that includes elements of both types of techniques (for example probabilistic latent semantic analysis [12]). In general, a lot of results achieved in statistical, machine learning, information retrieval, and other related research communities recently have been used with success in recommender system research.

For a more detailed discussion of different approaches used in recommender systems for personalization we suggest [2].

2.3 Non-intrusiveness: Efficient Learning of User Preferences

The minimization of feedback requests is usually desirable in personalization systems, since this reduces intrusiveness and as a consequence helps to avoid the visitors' irritation. Several truly non-intrusive feedback-determination methods have been proposed in the literature. However, such techniques are often inaccurate. Therefore, usually the problem is formulated as to minimize intrusiveness while learning visitor's preferences accurately and thus maintaining personalization quality.

Recently, techniques that utilize information about item popularity (well- vs. poorly known artworks) and item controversy (similarly vs. differently rated artworks) have been applied to increase the efficiency of leaning users' interests. Popularity and controversy can serve as a measure of item entropy [16], as well as their balanced combinations [13]. Yet, *non-intrusiveness*

is still recognized as one of the central problems in enhancing the state-of-the-art in the recommender systems research and development [2].

Analogous problems have been tackled in the machine learning community. Some of the approaches based on the active learning paradigm recently have been adapted to the user modeling area. For example, the ActiveCP approach, introduced in [19] and then further developed in [16] utilizes information about items controversy and popularity, based on the assumptions that (i) rating the most popular items first will result in a much greater information gain (when a user evaluates a popular item, the system becomes able to determine his similarity with a greater number of other people), and (ii) rating an item, which users have assigned widely varying ratings, will more likely provide the system with more discriminative information.

Another interesting approach was proposed in [7]; their VC-WMP algorithm clusters items by categories in order to reduce the dimensionality and sparseness of the score matrix. VC-WMP applies a majority vote learner in which the selection of votes is based on the correlation of user profiles.

Weber and Pollack in [21] introduce an entropy-driven active learning algorithm (for interactive calendar management application) that allows to better balance learning efficiency and user satisfaction.

For content-based image retrieval, Xie and Ortega in [22] propose to employ an empirical method to capture the probabilistic information of the user's preference. This probabilistic information consists of positive and negative samples for their SVM kernel called *User Preference Information Divergence (UPID)*.

Zhang *et al.* in [25] address the problem of extending an adaptive information filtering system to make decisions about the novelty and redundancy of relevant items in the context of document retrieval. The idea is to suggest items that are similar to previously recommended items (sharing the same topic), but also dissimilar to the previously recommended items in the sense of containing new information.

Yu *et al.* in [24] introduce a *transductive experimental design* that explores available unrated items and selects such items that are on the one side hard-to-predict and on the other side representative for the rest of the items.

We also develop a framework that enables leaning user preferences efficiently by trying to minimize the number of requests to the visitor and accounting for the coverage problem. We describe it formally in the next section connecting it to the CHIP project[2] that is one typical example of research and development effort directed to enhancing personalized access to cultural heritage. The CHIP project aims at improving the user's interactive experience with the Rijksmuseum repository interface by predicting the interest of the user in the repository artworks. CHIP currently employs a simple content-based algorithm for deducing art preferences in topics from the collection which then produce content recommendations. Based on a short questionnaire in which a set of artifacts needs to be scored, the virtual museum application has to design an optimal museum tour. We call this setting the off-line setting. We will also consider an online setting, where the user does not have to fill out a questionnaire, but, instead, during the tour, the personalization system constantly asks and receives feedback of the user regarding the presented artifacts.

3 A Generic Framework for the Museum Tour Problem

The goal of this section is to present a formal description of the museum tour personalization problem and a quality measure. In this quality measure, we assume that a user is not per se interested in seeing the N artifacts (s)he likes most, but rather in seeing the set of N artifacts that best covers his or her interest. Consider, for example, a visitor of the Rijksmuseum that likes landscapes, and also, but to a lesser extent, portraits. This visitor would probably not be satisfied with a tour consisting solely of landscapes. To deal with this last assumption, we introduce the notion of *coverage*. Intuitively, good coverage is obtained by not only rewarding good recommendations, but also penalizing categories of artworks of interest to the user that are not recommended. We also take into account that a too large number of artifacts in the questionnaire

[2] http://www.chip-project.org/

needs to be penalized. In this, we consider an online and an offline setting. In the offline setting, the user has to complete a questionnaire first, after which the tour is constructed. In the online setting, the feedback of the user and the tour are interleaved; for example, by the user giving scores to the presented artworks.

3.1 Coverage

First, we assume the existence of a set of objects *O*. In the context of the Rijksmuseum, these objects are the artifacts (artworks). Second, with every object $o \in O$, a set of characteristics *c(o)*, also called features, is associated. For example, for the painting *the Nightwatch*, the set of characteristics could be {*rembrandt*, *17th century*, *oil paint*, *militias*}. Notice that this encoding is an over-simplification; in fact, in reality for the Rijksmuseum collection, the artifacts are semantically annotated using the RDF format. The framework described here, however, is sufficiently general to capture the main characteristics of the museum tour problem. Third, we assume that there is a user *u* which has a preference *u(o)* for every object $o \in O$. Obviously, this preference function *u* is not known to the system, but needs to be deduced from individual scores *u(o)*. In concordance with the tool developed in the CHIP project to score the artifacts of the Rijksmuseum, the score *u(o)* is an integer ranging from -2; "*I hate it*" to 2; "*I love it*".

A *museum tour* of size *N* is defined as a set *T* of *N* artifacts. The quality of a tour is measured as follows. First, we assume that there exists a similarity function that assigns to every pair of artifacts *o*1; *o*2, a number $0 \leq sim(o_1, o_2) \leq 1$ pairs artifacts for which this similarity is high, are considered to be very alike, and pairs of very different artifacts have low similarity score. Finding an appropriate similarity measure is a research question on itself. It can, for example, be based on information of ratings of previous visitors. A good candidate, when lots of ratings are available, might be based on a combination of similarities between characteristics and the correlation between the scores for the artifacts for users that rated both artifacts. We do not go into detail here on how to find a good similarity function, as already an enormous amount of research effort has been spent on this, in the context of, for example, content-based methods.

Given a tour *T*, for a certain artifact *o*, we will define its coverage as the maximal similarity of o with the artifacts in the tour:

$$coverage(o, T) = \max_{t \in T} sim(t, o) . \quad (1)$$

Thus, an artwork can have a large coverage without being in the tour at all; if a closely related artwork is in the tour, intuitively, we assume that the artwork itself is being covered to a large extent in the tour. The quality of the tour can now be computed as follows:

$$q(T) := \sum_{o \in \mathcal{O}} coverage(o, T) \cdot u(o) . \quad (2)$$

Hence, desirable artifacts being covered will result in a higher score, while undesirable artifacts being covered are penalized with a negative contribution to the overall score. Also, tours that focus on only a limited number of categories of artworks will have a large disadvantage over tours that cover all higly desirable categories and subspaces.

3.2 Online and offline setting

The second assumption in our framework is that a user does not like to complete overly long questionnaires. To deal with this assumption, we introduce an online and offline setting.

In the off-line setting, the user is required to first take a short test in order to determine his' or hers preferences. Ideally, this *questionnaire* is as short as possible, while still allowing the learning system to acquire enough information to design a museum tour. To this end, for each artifact the visitor has to rate, a cost *c* has to be paid. After the off-line test, the system has to come up with a set *T* of *N* artifacts that will be visited during the tour. The goal in this setting is, given *N*, to find a tour *T* that optimizes the following objective function *B* (of benefit), where *n* is the number of artifacts in the questionnaire the learning algorithm uses:

$$B(T,n) := q(T) - n \cdot c\,. \tag{3}$$

Thus, the more artifacts there are in the questionnaire, the better the algorithm will be able to select an optimal tour *T*. As a consequence, however, the cost of the questionnaire increases. Therefore, the learning algorithm has to deal with the trade-off of having a better user model versus a shorter, and hence cheaper, questionnaire. The determination of the cost must be directly related to the nuisance of having to go through these questions experienced by the user. As such, also cost models where the cost increases monotonically with the number of questions asked can be highly useful.

In the online setting, the situation is quite different, but the objective function is the same. In this setting, the visitor does not have to take a test, but the tour starts without any information about preferences at all. During the tour, however, the user gives feedback after every artifact visited. The goal is now, to adaptively provide a tour *T*, visiting a fixed number of *N* artifacts that optimizes the objective function:

$$B(T) = q(T) - nc\,. \tag{4}$$

In this setting there is a tension between offering the user a ”safe” artifact with optimal score, or a “risky” artifact optimizing the user model. Clearly, a successful algorithm has to balance between exploring and building the user preference model on the one hand, and optimizing the scores on the other hand.

3.3 Some notes on implementation of the framework

The proposed framework allows for a nice theoretical exploration of optimal strategies for the museum tour problem. Quite essential in the whole discussion, however, is the existence of a similarity function, and the fact that for every artifact part of the score relies on external factors; i.e., not its neighbors. Therefore, for the successful implementation of this framework in a practical situation, it is of utmost importance that these assumptions are thoroughly verified. For the similarity function, for example, a machine learning technique based on available historical data can be used to learn parameters.

When learning the user model, it is important to have a good coverage of the whole space with as little examples as possible. Important here is the notion of certainty we have about the score of an artefact. Obviously, when it has already been scored, we may assume that we are perfectly certain about its score. On the other hand, if some very similar artifacts have been scored quite consistently (i.e., either all high, all medium, or all low), the certainty of the score is higher than when there is no similar artifact scored at all, or the scoring is quite inconsistent.

The decision whether or not to ask the user to rate an extra artifact, and what artifact can then be based on the expected gain presented by knowing the score for the artifact. The decision whether or not to stop harassing the user with the questionnaire can be based on whether or not the cost is higher than the gain.

In general, finding the optimal tour $T^{opt}(Q)$ is provable a hard computational problem. We do not go into detail here about the complexity, as a full proof of NP-completeness is beyond the scope of the paper. The problem, however, can be approximated reasonably well with a greedy method (always selecting the element with the largest improvement), which often leads to a (reasonably close to) globally optimal solution, or a gradient descent-type of approach, which finds coordinates of virtual artifacts (once we have found the optimal virtual artifacts, we can each of them replace with the nearest-by real artifact) that optimally cover the space.

4 Evaluation methodologies for personalization

In this section we discuss some methodological issues of scientific evaluation of personalization, corresponding challenges, and the ways of addressing them. Besides, we overview the basic types of metrics used for the evaluation of a personalization effect. Afterwards, we introduce the methodological framework for guiding the cultural heritage content personalization evaluation and enhancement.

4.1 Challenge of Scientific Evaluation of Personalization

The scientific evaluation of personalization impact is of the highest importance in recommender systems. If the performance metrics indicate the personalization strategy does not result in any gain of some utility function, there is an obvious need to understand if the cause is poor data collection, inaccurate modelling of visitor preferences, poorly chosen techniques for matchmaking, or ineffective content delivery [2].

In the data mining (DM) community it is common to see evaluation as an integral part of the whole DM process. In fact it can be considered in different stages – (i) when a new technique is being developed and (re)learned on some simulated or benchmark dataset, (ii) when its performance (for example predictive accuracy) in real settings is estimated, and (iii) when the developed techniques are compared to alternative (existing) techniques. Yet, traditional DM research rarely concentrates on deployment of developed techniques and evaluation of their use in real settings. Therefore it is not so easy to estimate their real utility for a certain type of application. This problem becomes even more severe in personalized applications where DM is aimed to serve as an intelligent tool to discover actionable knowledge (to be used for example by a recommendation system) from the data being collected during interactions between the users and the applications.

One problem is that evaluation (in the uncontrolled experiment that is illustrated in Figure 2) is performed on some *test data* that the *users chose and decide* (that may naturally result in a skewed sample) to rate and we cannot estimate the ability of the system to properly rate a random item [23]. Therefore, they conclude that it is exceptionally hard to guarantee the scientific evaluation of personalization in the natural setting. In a normal operational setting, the historical data does not reflect the true effect of personalization since it has no information what would the performance estimates be if personalization did not take place. Therefore, such evaluation can provide only weak evidence of a positive effect.

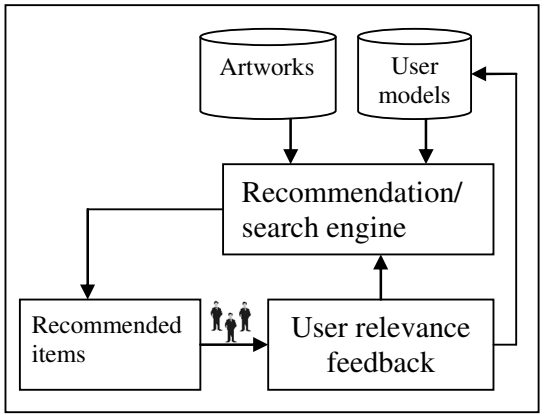

Fig. 2. Evaluation framework with relevance feedback loop.

Yang and Padmanabhan in [23] argue that good prior knowledge can partially overcome this problem. More specifically, knowledge about *expected outcomes* (such as increase/decrease of 'profit' that application gains due to personalization) under different personalization scenarios can potentially be used to evaluate personalization. Some ideas of employing utility-based and economics-oriented measures like return on investments (ROI), customer lifetime value (LTV) measures, and others can be adopted from e-commerce applications [18] where this problem has been addressed to some extent.

Still, the most frequently used performance metrics are accuracy-based metrics, reflecting how much the consumer liked or disliked a specific personalized offering and metrics that incorporate both some coverage and accuracy. Coverage accounts for the percentage of items for which a recommender system is capable of making predictions, and an accuracy (or error) measure reflects the similarity (correlation, difference, etc.) between the estimated (predicted) ratings against the actual ratings. Besides, classical information retrieval measures like precision and recall (and their

mean known as *F*-measure) can be used for performance estimation. Precision accounts for the percentage of correctly predicted "high" ratings among all rating that were predicted to be "high", and recall for the percentage of correctly predicted "high" ratings among all the ratings known to be "high". ROC-curves are another common technique that allow to see the trade-off between true positive and false positive rates in recommender systems.

In the next section we suggest one personalization evaluation framework and argue why it can be easier accepted in cultural heritage applications in particular, and in educational and entertainment applications in general, as compared to the classical e-commerce applications (like e-shops).

4.2 The Methodological Framework for Evaluating and Guiding Personalization Process

In general, during the personalization evaluation process we can be interested to answer questions of two types; (1) whether a personalization technique has positive effect, i.e. a performance metric should reflect that personalizing is better than not personalizing, and (2) whether this personalization technique is better than some other alternative approach for personalization. These questions can be answered with classic controlled experiments called AB and multivariable tests.

In the AB type of experimental design, users are randomly allocated into a treatment group that assumes personalization and a control group with no personalization. In both groups, the desired metric is observed before and after the treatment group intervention.

For example, using this method to evaluate personalization technique used in the CHIP project would mean randomly selecting two groups of users (see Figure 3), then measuring the number of browsed artworks (or some other metrics) in the first period characterized by no personalization for both groups, followed by measuring the number of browsed artworks in the second period where one of the groups has a personalized content. Following such scenario, the analysis of performance estimates in the two periods for both groups indicates whether personalization was successful or not. The results of such guided personalization can be stored for further analysis and the application of inferencing (meta-learning) techniques. This leads to discovery of knowledge that can be used for further improvement of personalization strategies and directing the continuation of the controlled experiments.

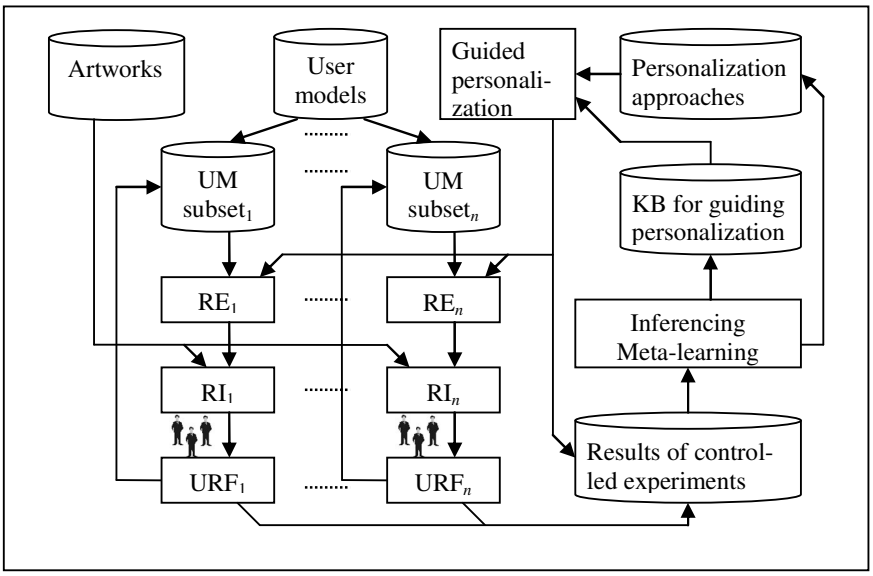

Fig. 3. Evaluation framework of 'controlled experiment'-based guided personalization (RE – recommendation engine, RI – recommended items, URF – user relevance feedback).

Such experimental design in e-commerce applications in the natural settings is less acceptable, since having a period without personalization implies the risk of losing customers. In educational and entertainment applications, however, it can be more freely used, especially if the visitors are aware of such procedures and can treat them as a natural part of the personalization process during which the recommendation system learns to improve a personalization technique or to select the most appropriate personalization techniques among available.

5 Discussions and Further Research

The importance and potential utility of web access personalization to cultural heritage has been recognized by many museums. From the other hand, the vast amount of research in recommender systems, information retrieval and data mining has facilitated the development of such personalized applications.

In this paper, we studied some of the challenges of museum tour personalization, focusing on the problems of efficient learning visitor preferences in offline and online settings. We reviewed the basic approaches that can be used for personalizing the access to the cultural heritage content and we presented a formal framework for evaluating online and off-line learning of these preferences. In the future we plan to implement this framework on top of real data from the Rijksmuseum.

Our further research direction is to develop methods that utilize some of the more advanced profiling techniques based on data mining (finding actionable rules, sequential patterns, and signatures) in addition to using traditional features such as keywords and simple user demographics. Some related work that has been accomplished in the area of web usage mining (like discovering the navigational patterns) can be adapted to rating-based tour personalization systems.

Another interesting research direction is to adjust recommendations to the context in which it is offered. For example, it was shown in [1] for a movie recommendation application that, by extending the traditional memory-based collaborative filtering approach to take into consideration the *when*, *where*, and *with whom* a movie is seen, the resulting recommender system was able to show better results in comparison to the purely traditional collaborative filtering method. Naturally, similar contexts can be recognized in the museum tour personalization application.

However, let's not forget that there exist sound arguments not in favor of personalization. For example, there exists an opinion that personalization is often over-rated, claiming that good basic Web navigation is much more important. In general, good personalization can provide some benefits, but another possibility is that certain techniques can erroneously personalize and incur associated costs.

And one way or another, there is a need for high-quality controlled experiments to faithfully estimate the benefits and limitations of certain personalization techniques in the context of certain application areas. Such experiments can be expensive and time-consuming indeed, but are necessary to provide sound conclusions concerning the usefulness of personalization and intelligent techniques that contribute to the personalization procedure.

Evaluation and feedback integration in the personalization process has not been studied extensively in the personalization literature. Furthermore, most state-of-the-art recommender systems do not implement sound methods for adjusting personalization strategies, although this also constitutes an urgent task of personalization. We would like to attract the readers' attention again and again to the problem of scientific evaluation of personalization.

In this paper we overviewed methodological issues of personalization adaptation and came to the conclusion that: (i) evaluation procedures should constitute the integral part of recommender system and be applied at every stage of the interactive and iterative processes of personalization; (ii) the scientific evaluation of the personalization impact and, furthermore, the evaluation of DM techniques that are utilized to produce (discover) actionable knowledge which facilitates personalization, is impossible in traditional view of "on-line" settings (remember Figure 2) due to the lack of an adequate control group (that is essential). Therefore, controlled experiments must regularly take place in personalization research. In this paper we tried to argue that it is possible to employ such controlled experiments also within a *normal* operational setting of cultural heritage applications. The corresponding methodological framework was presented.

References

1. Adomavicius G., R. Sankaranarayanan, S. Sen, and A. Tuzhilin, Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach, ACM Trans. Information Systems Vol. 23(1), (2005)
2. Adomavicius G. and Tuzhilin A. 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(6), (2005) 734-749
3. Adomavicius, G. and Tuzhilin, A. 2005. Personalization technologies: a process-oriented perspective. Commun. ACM Vol. 48(10), (2005) 83-90
4. Ansari A., Essegaier S., and Kohli R. Internet Recommendations Systems, J. Marketing Research, Vol. 37, Aug. (2000), 363-375
5. Billsus, D. and Pazzani, M.J. User modeling for adaptive news access. User Modelelling and User-Adaptive Interaction, Vol. 10(2-3), (2000) 147–180
6. Bowen, J. P. and Filippini-Fantoni, S. Personalization and the web from a museum perspective. In: D. Bearman and J. Trant (Eds.), Museums and the Web 2004: Selected Papers from an Int. Conference, Arlington, Virginia, USA, (2004) 63–78
7. Delgado J. and Ishii N. On-line learning of user preferences in recommender systems. In: Proc. of International Joint Conference on Artificial Intelligence (IJCAI-99), Workshop on Machine Learning for Information Filtering, Stockholm, Sweden, July 1999
8. Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T. Evaluating collaborative filtering recommender systems. ACM Transactions of Information Systems, Vol. 22(1), (2004) 5–53
9. Huang, Z., Chen, H., and Zeng, D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions of Information Systems, Vol. 22(1), (2004) 116–142
10. Kumar R., Raghavan P., Rajagopalan S., and Tomkins A. Recommendation Systems: A Probabilistic Analysis, J. Computer and System Sciences Vol. 63(1), (2001) 42-61
11. Pennock D.M. and E. Horvitz, Collaborative Filtering by Personality Diagnosis: A Hybrid Memory And Model-Based Approach, In: Proc. Int. Joint Conf. Artificial Intelligence Workshop: Machine Learning for Information Filtering, Aug. 1999
12. Popescul A., L.H. Ungar, D.M. Pennock, and S. Lawrence, Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments, In: Proc. 17th Conf. Uncertainty in Artificial Intelligence, 2001.
13. Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., and Riedl, J. Getting to know you: learning new user preferences in recommender systems. In Proc. of the 7th Int. Conference on intelligent User interfaces. IUI'02. ACM Press, New York, (2002) 127-134
14. Rutledge, L., Aroyo, L., and Stash, N. Determining user interests about museum collections. In Proceedings of the 15th international Conference on World Wide Web WWW '06. ACM Press, New York, (2006) 855-856
15. Rutledge, L., Aroyo, L., and Stash, N. 2006. Interactive User Profiling in Semantically Annotated Museum Collections. In: Proc. 5th Int. Semantic Web Conference, Athens, GA, USA, November 5-9, 2006, LNCS 4273, (2006)
16. Sampaio, I., Ramalho, G., Corruble, V. and Prudencio R. 2006. Acquiring the Preferences of New Users in Recommender Systems: The Role of Item Controversy, ECAI 2006 Workshop on Recommender Systems, (2006) 107-110
17. Savia, E., Puolamäki, K., Sinkkonen, J., Kaski, S. Two-Way Latent Grouping Model for User Preference Prediction. In Proceedings of the UAI'05, (2005) 518-525
18. Schafer, J., Konstan, J., and Riedl, J. E-commerce recommendation applications. Data Mining and Knowledge Discovery Vol. 5, 1–2 (2001), 115–153
19. Teixeira, I. R., Carvalho, F. d., Ramalho, G., and Corruble, V. 2002. ActiveCP: A Method for Speeding up User Preferences Acquisition in Collaborative Filtering Systems. In: G. Bittencourt and G. Ramalho, (Eds.) Proceedings of the 16th Brazilian Symposium on Artificial intelligence: Advances in Artificial intelligence, LNCS, vol. 2507. Springer-Verlag, London, (2002) 237-247
20. Ungar, L.H. and Foster, D.P. Clustering Methods for Collaborative Filtering, In. Proc. Workshop Recommender Systems, AAAI Press, (1998)

21. Weber, J. S. and Pollack, M. E. Entropy-Driven online active learning for interactive calendar management. In Proc. of the 12th Int. Conference on intelligent User interfaces. IUI'07. ACM Press, New York, (2007) 141-150
22. Xie, H. and Ortega, A. An user preference information based kernel for SVM active learning in content-based image retrieval. In Proc. of the 6th ACM SIGMM Int. Workshop on Multimedia Information Retrieval MIR '04. ACM Press, New York (2004)
23. Yang Y. and Padmanabhan B., On Evaluating Online Personalization, Proc. Workshop Information Technology and Systems, (2001) 35-41
24. Yu, K., Bi, J., and Tresp, V. Active learning via transductive experimental design. In: Proceedings of the 23rd Int. Conference on Machine Learning ICML '06, vol. 148. ACM Press, New York, (2006) 1081-1088
25. Zhang, Y., Callan, J.P., and Minka, T.P. Novelty and redundancy detection in adaptive filtering. In Proc. of the 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, (2002) 81–88