USING CATEGORY-BASED COLLABORATIVE FILTERING IN THE ACTIVE WEBMUSEUM

Arnd Kohrs and Bernard Merialdo

Institut EURECOM – Department of Multimedia Communications
BP 193 – 06904 Sophia-Antipolis – France
{Arnd.Kohrs,Bernard.Merialdo}@eurecom.fr

ABSTRACT

Collaborative filtering is an important technology for creating user-adapting Web sites. In general the efforts of improving filtering algorithms and using the predictions for the presentation of filtered objects are decoupled. Therefore, common measures (or metrics) for evaluating collaborative filtering (recommender) systems focus mainly on the prediction algorithm. It is hard to relate the classic measurements to actual user satisfaction because of the way the user interacts with the recommendations, determined by their representation, influences the benefits for the user. We propose an abstract access paradigm, which can be applied to the design of filtering systems, and at the same time formalizes the access to filtering results via multi-corridors (based on content-based categories). This leads to new measures which better relate to the user satisfaction. We use these measures to evaluate the use of various kinds of multi-corridors for our prototype user-adapting Web site the: Active WebMuseum.

1. INTRODUCTION

Information filtering techniques can be used to create user-adapting Web sites. User-adapting Web sites adapt their presentation to the user's preferences. We evaluate the use of information filtering for user-adapting Web sites in our prototype the Active WebMuseum.

Techniques, which have been proposed for information filtering fall in two classes: Content-based filtering and collaborative filtering. In recent research we studied the combination of both approaches in the context of the Active WebMuseum[3, 4]. Filtering techniques, content-based as well as collaborative filtering generally create predictions of user ratings for objects. Usually the performance of filtering algorithms (or prediction algorithms) is then evaluated using measures which asses the prediction error. Other measures consider the order of all predicted ratings and compare it to the order of ratings that the user would assign.

While standard measures are commonly used to evaluate and tune the performance of filtering algorithms, they are hard to relate to the utility of the recommender system and therefore to the user satisfaction. The utility of predictions depends strongly on how the user accesses the recommendations, on available choices during the exploration, and on which choices the user picks. Current measures do not consider arrangements and user choices. Recommender systems which are tuned to the wrong measures might not be useful for the users.

We propose a new access paradigm for the application of filtering techniques which is based on multi-corridors. Multi-corridors are constructed as follows:

- The objects are divided in categories, possibly by using features obtained through automatic indexing.
- For each category the objects are sorted according to the filtering algorithm and arranged in a single *corridor*.

The user is assumed to behave as follows:

- choose a corridor
- enjoy the objects of a corridor
- exit the corridor when disappointed

This paradigm gives an abstract definition of arrangement and usage of filtering results. This paradigm is obvious and simple. It allows a reliable formalization so that performance measures can be deduced, that are based on the interaction between the user and the system and therefore relate more to the user satisfaction. When this paradigm is applied to user-adapting Web sites or other types of recommender systems, more choice for the users can be provided via various multi-corridors and therefore increase performance from the perspective of the user.

In this paper we first describe the prototype Active WebMuseum in section 2, which implements the multi-corridor paradigm. Further in section 3 we introduce collaborative filtering and other prediction algorithms used in this paper. Then we describe in more detail the multi-corridor access paradigm, how multi-corridors can be obtained automatically from content indexing, and the $count$ and $score$ measures which are used to assess the performance of multi-corridors and prediction algorithms. Later we apply these measures to various types of multi-corridors resulting from different combinations of filtering algorithms and categorizations.

2. THE ACTIVE WEBMUSEUM

In an ideal world a visitor of a museum would enter a museum and then find in the first corridor exactly those items, which he would find most interesting. Given that real museums serve many people at the same time, it is not feasible to rearrange the collection for individual visitors. When a museum's art collection is presented through the Web, it becomes feasible to rearrange the collection for each individual visitor. Our Active WebMuseum[1]has a dynamic topology which is adapting to the museum visitor's taste and choices. The dynamic topology is achieved by dynamic corridors, virtual corridors which contain paintings of a chosen category sorted according to personalized predictions produced by collaborative filtering. The user may choose from several dynamic corridors which are interconnected, so that users keep the ability

[1] The Active WebMuseum (accessed through `http://www.eurecom.fr/~kohrs/museum.html`) uses the collection of paintings from the *WebMuseum, Paris* (accessed through `http://metalab.unc.edu/wm/`), which has been created by Nicolas Pioch and contains roughly 1200 paintings by about 170 painters.

to choose (among paintings and corridors) while at the same time benefitting from recommendations. See figures 1 and 2 for examples.

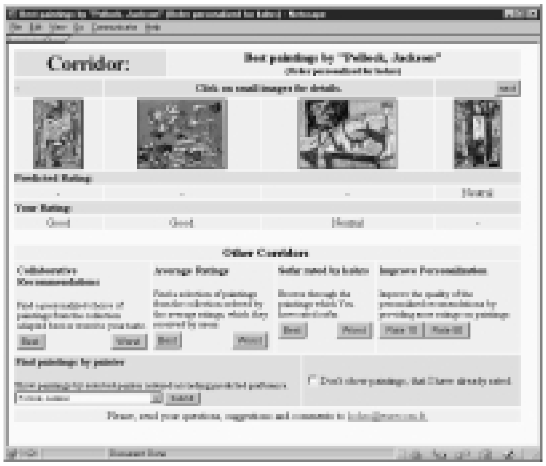

Corridor:
Best paintings by "Pollock, Jackson"
Click on small images for details.
next
Predicted Rating:
Neutral
Your Rating:
Good
Good
Neutral
Other Corridors
Collaborative Recommendations
Average Ratings
Improve Personalization
Best
Worst
Find paintings by painter
Submit
Don't show paintings, that I have already rated.

Figure 1: Browsing dynamic corridors: When a user has entered a dynamic corridor (in this example a corridor containing paintings by Jackson Pollock), he is presented iconized paintings ordered according to his preference. From here the visitor continues in the same corridor until he looses interest, or may choose to see more details of one painting, or enter a new corridor.

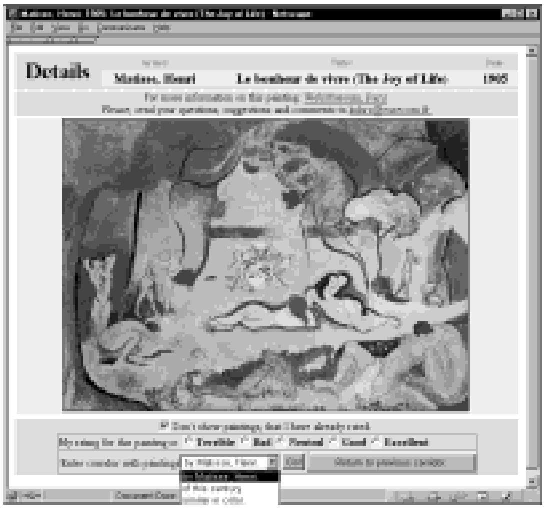

Details
Matisse, Henri
Le bonheur de vivre (The Joy of Life)
1905
Don't show paintings, that I have already rated.
Terrible
Bad
Neutral
Good
Excellent
Go!
Return to previous corridor
of this century

Figure 2: A single painting in detail close-up: When the user choses an iconized painting from a corridor it is presented in more detail (artist, title, creation date). From here the user may return to current corridor or enter new corridors related to this painting.

While visiting the Active WebMuseum, users may express preferences by giving symbolic ratings to paintings (*excellent*, *good*, *neutral*, *bad*, *terrible*). For historic reasons, the symbolic ratings are then mapped on numerical ratings in the interval $[0..10]$. For paintings which have not been rated by the visitor, the ratings are predicted using other users ratings and collaborative filtering technology.

3. COLLABORATIVE FILTERING

Collaborative filtering systems select items for a user based on the opinions of other users. Generally, collaborative filtering systems do not rely on content-based information about the items, considering only human judgments on the value of items. Collaborative filtering systems consider every user as an expert for his taste, so that personalized recommendations can be provided based on the expertises of taste-related users. Collaborative filtering has been applied to several domains of information: News articles, GroupLens [5, 6, 7]. Music, Ringo [8]. Movies, MovieCritic[2].

Most collaborative filtering systems collect the users' opinions as ratings on a numerical scale, leading to a sparse matrix $r(user, item)$. Collaborative filtering systems then use this rating matrix in order to derive predictions. Several algorithms have been proposed on how to use the rating matrix to predict ratings [2, 8, 1]. For the Active WebMuseum we derived a collaborative filtering algorithm from a commonly used technique, also used in the GroupLens project and in Ringo, which is based on Pearson vector correlation. The predictions are weighted sums of other users ratings, and the weights are determined by correlation coefficients between the users' ratings vectors. Please refer to [3] for the detailed formulas. In this paper we refer to this algorithm, when we mention collaborative filtering algorithm, or *collab* for short.

In order to compare the performance of collaborative filtering we use in experimental context also the following filtering algorithms:

Base: The base algorithm uses the mean rating of all users for an object as a prediction. This is also a collaborative filtering algorithm but it does not produce personalized results and is solely used for comparison and as a backup algorithm in rare cases when collaborative filtering fails.

Random: The random algorithm uses random numbers within the rating range as prediction. This algorithm is only useful for experimental purpose to estimate baseline results.

4. CATEGORIES

Filtering techniques are used to create predictions of a user's ratings for objects. The simplest way of using these predictions, is to provide the user with a ranked list of objects which the user has to follow in best-first order to benefit from the filtering algorithm. While this represents the way how filtering systems are usually evaluated, this access-paradigm is rather awkward for the user and might not be percepted as beneficial. We therefore propose the multi-corridor access paradigm, which formalizes the arrangement of filtered objects into corridors according to a categorization scheme. The user may choose a corridor while still benefiting from the predictions of the filtering algorithm.

Here we provide a formal model for multi-corridors. This model leads to two measures $count$ and $score$, which more closely capture the benefit that users get from filtering algorithms within the context of multi-corridor. Then we explore the multi-corridor paradigm by applying the proposed measures to various combinations of categories and prediction algorithms.

4.1. Multi-Corridor Model and Metrics

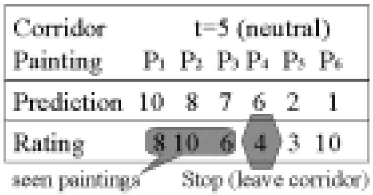
When presented as multi-corridors the objects are grouped according to a categorization scheme. Each category contains objects, which when ordered by a filtering algorithm are presented in a

[2] http://www.moviecritic.com

corridor-like fashion. The user chooses a corridor and sequentially sees as many objects as he likes. When done with one category the user switches to the next. A performance measure for the whole system should relate to the satisfaction the user experiences. The users satisfaction is maximized, when the system shows him many objects which he likes and few objects which he does not like enough. For our evaluations we use metrics, which capture the user's satisfaction resulting from the combination of a categorization and a prediction algorithm. In our metric it is assumed that the user stops using one corridor as soon as he is presented an object that he does not like, i.e. an object which he would rate with a rating below a threshold t. The value gained by visiting the corridor is determined in terms of the sum of the ratings which the user would have assigned to the seen objects or simply the number of objects that the user sees.

In the following we provide a framework for assessing the value of a multi-corridors. When assigned to corridors (ordered categories) the objects can be referenced by $o_{c,i}$, which refers to the ith objects in the cth corridor. The value $\hat{r}_{c,j}$ refers to the predicted rating of $o_{c,j}$. The value $r_{c,j}$ refers to the rating the user would assign to the object $o_{c,j}$. The absolute prediction error $|\hat{r}_{c,j} - r_{c,j}|$ which is commonly used for assessing the performance of recommender systems is not important for the following considerations.Within a corridor c the objects are ordered according to the predicted rating, so that $\hat{r}_{c,j} \geq \hat{r}_{c,j+1}$ holds for all objects. In our model we assume that the user stops using a corridor as soon as he gets disappointed when he sees an object with a rating ($r_{c,j}$) below a threshold t, which is the *neutral* rating. So that $stop_c = min\{j : r_{c,j} < t\}$ is the index of the object in corridor c which causes the user to exit corridor c. Until the user sees $o_{c,stop}$ he sees the objects $\{o_{c,1}, \ldots, o_{c,stop-1}\}$ and the ratings that the user would assign to those objects or the number of objects seen can be assumed to relate to the user's satisfaction gained by visiting this corridor. If the prediction was perfect, the user would see all objects which are important to him in this corridor. This leads us to the measures which estimate the percentage of experienced ratings ($score$) and the number of objects seen in one corridor ($count$):

$$score_c = \frac{\sum_{j<stop_c} r_{c,j}}{\sum_{j\in\{l:r_{c,l}\geq t\}} r_{c,j}} \quad count_c = \frac{stop_c - 1}{|\{l:r_{c,l}\geq t\}|} \qquad (1)$$



Corridor	t=5 (neutral)					
Painting	P1	P2	P3	P4	P5	P6
Prediction	10	8	7	6	2	1
Rating	8	10	6	4	3	10

Figure 3: When a user enters a corridor he keeps on seeing paintings as long as he likes the paintings, i.e. he would rate them higher than t. As soon as he gets disappointed he leaves the corridor and might miss paintings which he would also like.

In a typical visit it can be assumed that the user visits one or more corridors. In order to assess the utility of a multi-corridor for the user we model the access in a simplified way: The user is expected to choose only one corridor. Each corridor is chosen with a probability $w_c = P(\text{corridor } c \text{ is chosen})$. So that the mean experienced $score$ and $count$ can be estimated as follows.

$$score = \sum_c w_c \cdot score_c \quad count = \textstyle\sum_c w_c \cdot count_c \qquad (2)$$

The distribution of w_c is assumed to be uniform if not otherwise indicated for the following experiments.

4.2. Datasets

In this study we examine three generally different categorizations:

Automatic: In previous work we discovered a correlation between colors of paintings and the users ratings for this paintings[3, 4]. This leads us to assuming that users might profit from a color categorized presentation of the paintings. This categorization is based on automatically generated color histograms, which have been clustered in a fixed number of categories using the K-Mean algorithm. This categorization scheme is referred to as an *automatic* categorization, since no manual work is needed to create the categorization.

Manual: The paintings of the Active WebMuseum were extracted from Web pages. These pages usually contained descriptions about the painter and the style. In these descriptions the mentioning of certain keywords were counted, which allowed the assumption that the paintings belong to a certain style. By this technique all the paintings (≥ 1200) were categorized in ten different styles, i.e *Baroque, Cubism, Expressionism etc*. This categorization is based on the *manual* work of the author of the descriptions and his expert knowledge and could not be automatically generated without the descriptions.

Random: As a third categorization we created random categories. The paintings are randomly distributed over all categories. This categorization scheme serves solely as a basis for comparison.

In order to allow comparisons between different categorization schemes, each categorization has the same number of categories ($C_n = 10$) and each painting is assigned exactly once to one category within each categorization.

For the user rating based prediction we use 8780 ratings assigned to paintings by 320 user during the ongoing trial of the prototype Active WebMuseum.

4.3. Experimental Framework

We ran several sets of off-line experiments in order to resolve questions concerning the application of multi-corridors in the Active WebMuseum. For the experiments the user ratings are split in a test-set and a training-set. For 16 users with more than 100 ratings a subset of 80 ratings is sampled in the test-set (1280 ratings). The remaining ratings are used as a training-set (7500 ratings). In order to eliminate the effects of biased splits the splitting is repeated 50 times using random samples for the test-set. For each split the performance is measured (using the measure proposed in section 4.1) and accumulated to a mean performance over all splits and all users.

4.4. Results

We will now approach certain aspects of the use of multi-corridors in recommender systems. For the following results we use the $score$ measure which is strongly correlated to the $count$ measure.

Categorization Algorithm	Random	Base	Collab
Random	0.324	0.495	0.500
Automatic	0.379	0.525	0.528
Manual	**0.406**	**0.555**	**0.558**

Table 1: Performance comparison of categorization schemes and prediction algorithms.

4.4.1. Comparison of Categorization Schemes

Table 1 compares the performance using the $score$ measure for various prediction algorithms in combination with the three kinds of categorization schemes.

In general the *manual* categorization performs best, as would be expected. However, the *automatic* categorization performs notably better than the random categorization.

Surprisingly in the first column, the random predictor leads to different results for different categorizations. This can be explained with the fact that the categories are correlated with the ratings for the contained paintings (even though that is not intended), e.g. some categories contain more generally good objects so that even a random predictor can succeed. This result implies that the presentation in multi-corridors based on meaningful categorizations relating to style and even the automatic categorization improves the performance of the recommender system in terms of more perceived $score$, especially if the prediction algorithm is weak.

These results indicate that using *base* is better than *random* and *collaborative* is better than *base* prediction. Which is obvious since *base* takes into account general popularity and *collaborative* prediction personalizes the predictions. However, while generally remarkably outperforming *base* prediction, here *collaborative* only adds little performance over *base* prediction. This can be explained with the limited data-set which is available, i.e. a critical mass of users to have groups of users with differing and common tastes.

4.4.2. Category Weighting

Until now we assumed for the $score$ measure equal probabilities that a user chooses a particular corridor (Equation 2), i.e. the distribution of the weights w_i is uniform. In practice that would mean that the user would blindly choose one of the available corridors. Sometimes it might be reasonable to assume that the user favors one corridor over the others, i.e. because he usually likes objects in a particular corridor. We model this by adapting the weights w_i to the known ratings that a user had assigned to objects of one corridor. Low weights are assigned to corridors with low average ratings and high ratings are assigned to corridors with high average ratings. This weighting scheme is referred to as the *popular* weighting scheme as opposed to the *uniform* weighting scheme.

Categorization Algorithm	Uniform	Popular
Random	0.500	0.502
Automatic	0.528	**0.562**
Manual	0.558	**0.568**

Table 2: Comparison of uniform and popular weighting of categories.

Table 2 compares the results of uniform and popular weighting. Significant improvements can be observed for the *automatic* categorization with the *popular* weighting. For the *manual* categorization the improvement is not as significant. One could argue that the results are improved by changing the measurement, however, since the measurement is in the same quantity related to the user satisfaction, it can be concluded that the user is better served with a weighted categorization, i.e. particular categories are recommended.

5. CONCLUSION

This paper makes three contributions: First, the multi-corridor access paradigm is identified and defined. Second, we provide metrics for evaluating multi-corridors which are focused on the performance of information filtering prediction algorithm while at the same time considering the access patterns of the users. Third we evaluate multi-corridors based on various combinations of prediction algorithms and categorization schemes.

In the future we plan to refine the multi-corridor measures. Certainly a measure is needed for on-line experiments which can then lead to highly useful tools for observing the quality of a recommender system at run-time, e.g. when an on-line measure detects that a current user experiences bad performance a system operator could intervene and provide some incentive for the user to stay. Further we will explore the integration of categorizations and prediction algorithms with the hope that prediction results in general improve and become more reliable.

6. REFERENCES

[1] Jack Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, WI, July 1998. Morgan Kaufmann Publisher.

[2] Arnd Kohrs and Bernard Merialdo. Clustering for collaborative filtering applications. In *Computational Intelligence for Modelling, Control & Automation*. IOS Press, 1999.

[3] Arnd Kohrs and Bernard Merialdo. Improving collaborative filtering with multimedia indexing techniques to create user-adapting web sites. In *Proceedings of the 7th ACM Multimedia Conference*. ACM, 1999.

[4] Arnd Kohrs and Bernard Merialdo. Using color and texture indexing to improve collaborative filtering of art paintings. In *Proceedings of the European Workshop on Content-Based Multimedia Indexing (CBMI'99)*, 1999.

[5] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordan, and John Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, March 1997.

[6] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, Sharing Information and Creating Meaning, pages 175–186, 1994.

[7] Badrul M. Sarwar, Joseph A. Konstan, Al Bochers, Jon Herlocker, Brad Miller, and John Riedl. Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. In *Proceedings of ACM CSCW'98 Conference on Computer-Supported Cooperative Work*, 1998.

[8] Uprendra Shardanand. Social information filtering: Algorithms for automating ”word of mouth”. In *Proceedings of Human Factors in Computing Systems, CHI '95*, 1995.