A Nonparametric Hierarchical Bayesian Framework for Information Filtering

Kai Yu[†], Volker Tresp[†], Shipeng Yu[‡]
[†]Corporate Technology, Siemens AG, Munich, Germany
[‡]Institute for Computer Science, University of Munich, Germany
kai.yu@siemens.com, volker.tresp@siemens.com,
spyu@dbs.informatik.uni-muenchen.de

ABSTRACT

Information filtering has made considerable progress in recent years.The predominant approaches are content-based methods and collaborative methods. Researchers have largely concentrated on either of the two approaches since a principled unifying framework is still lacking. This paper suggests that both approaches can be combined under a *hierarchical Bayesian framework*. Individual content-based user profiles are generated and collaboration between various user models is achieved via a common learned prior distribution. However, it turns out that a parametric distribution (e.g. Gaussian) is too restrictive to describe such a common learned prior distribution. We thus introduce a *nonparametric common prior*, which is a sample generated from a *Dirichlet process* which assumes the role of a *hyper prior*. We describe effective means to learn this nonparametric distribution, and apply it to learn users' information needs. The resultant algorithm is simple and understandable, and offers a principled solution to combine content-based filtering and collaborative filtering. Within our framework, we are now able to interpret various existing techniques from a unifying point of view. Finally we demonstrate the empirical success of the proposed information filtering methods.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, retrieval models*

General Terms

Algorithms, Theory, Human Factors

Keywords

Collaborative Filtering, Content-Based Filtering, Dirichlet Process, Nonparametric Bayesian Modelling

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

1. INTRODUCTION

Information filtering denotes a family of techniques that help users to find the right information items while filtering out undesired ones. In a wide range of applications, such as spam email filtering, news filtering, and recommender systems for products (e.g. books), information filtering is playing an increasingly important role. Content-based filtering (CBF) and collaborative filtering (CF) represent the two major information filtering technologies. CBF [19, 2, 10, 13] has its root in the concept of *relevance feedback* in the information retrieval literature (e.g. Rocchio's algorithm [19]). CBF explores the similarity of contents between information items (e.g. articles, images, music), to infer which of the yet unseen items might be of interest for the active user, based on some annotated examples previously given by the user. In contrast, collaborative filtering methods [18, 20, 4] typically accumulate a database of item ratings—explicitly or implicitly—cast by a large set of users. The prediction of ratings for the active user is solely based on the ratings provided by all other users, under the assumption that like-minded users are sharing similar information needs. The method does not rely on a description of item content.

One major difficulty in designing CBF systems lies in extracting content features that are sufficiently indicative. There is often a large gap between low-level content features (visual, auditory, or others) and high-level user interests (like or dislike a painting or a CD). In some other circumstances, the features are not available at all. Fortunately, the information on personal preferences and interests are all carried in (explicit or implicit) user ratings. Thus CF systems can make use of these high level features rather easily, by combining the ratings of other like-minded users.

Pure CF only relies solely on user preferences, without incorporating the actual content of items. CF often suffers from the extreme sparsity of available data, in the sense that users typically rate only very few items, thus making it difficult to compare the interests of two users. Furthermore, pure CF can not handle items for which no user has previously given a rating. Such cases are easily handled in CBF systems, which can make predictions based on the content of the new item.

Naturally, previous researchers have worked on compensating the drawbacks of each particular approach. Many approaches have focused on *hybrid filtering* to unify both approaches [12, 7, 2, 3, 16]. However, due to the lack of a unifying framework for information filtering, existing solutions were developed mostly in heuristic or ad-hoc ways. The challenge is therefore to find a *principled* approach to combine CBF and CF. A recent publication [21] made first attempts to unify CF and CBF under a hierarchical Bayesian framework. The work is extended in this paper and the theoretical basis is further developed and clarified.

1.1 Overview of Our Work

This paper introduces the idea of *nonparametric hierarchical Bayesian modelling* to information filtering. This framework provides an understanding on the *structure* of information filtering and leads to a principled hybrid filtering algorithm. The framework assumes that each user's observed preferences data (i.e. annotations) are generated based on the user's own profile model, which itself is a random sample from a prior distribution of user profiles, shared by all the users and thus called the *common prior* in this paper. In this hierarchical Bayesian model each user's model is constrained by the common prior, through which the user is "communicating" with others.

The common prior is "learned" based on the observed annotations from a population of users. One may assume a parametric form (e.g. a Gaussian) for the common prior, and then estimate the associated parameters (e.g. the mean and variance in the Gaussian case). However, due to the complexity of the functional form of the learned prior, the true distribution of profiling models cannot adequately be described by any known parametric distributions (e.g. a Gaussian). As a solution, this paper relaxes the parametric limitation on the common prior and adopts a nonparametric form— a distribution generated from a *Dirichlet Process* [8]. This model encompasses that, *a priori*, a new user may follow other users' interests, but may also have his/her own unique interests. The process enables a learning session for a new user to inherit knowledge from the sessions of other users, which leads to quite meaningful results for information filtering. In the learning phase, typical Bayesian inference requires MCMC (Monte Carlo Markov Chain) sampling that is computationally expensive. This paper instead introduces novel approximations to learn the common prior effectively and efficiently. For a new user, the learned common prior represents the knowledge inherited from the previous models leading to good predictions, even with only few data points available for the new user.

Finally, we emphasize that the proposed work not only presents a novel and *principled* hybrid information filtering algorithm, but is also a quite *general framework* for information filtering and retrieval, since: (1) It unifies the CF and CBF in a single framework, where pure CF and pure CBF are special cases under certain circumstances; (2) Various existing algorithms combining CF and CBF can now be interpreted from a unified point of view, and their further improvements are also suggested; (3) The proposed work is also applicable to information retrieval, enabling retrieval sessions to inherit knowledge from each other. (4) The framework makes no requirements for the form of profiling models, it is thus applicable to a very wide range of user modelling applications in information filtering and retrieval (e.g. hidden Markov model for modelling user web browsing, and support vector machines for image retrieval).

1.2 Structure of This Paper

The rest of this paper is organized as follows. In Sec. 2, we will expand on the the idea of modelling information needs of users in a nonparametric hierarchical Bayesian framework. In Sec. 3 we derive effective means to learn the model. Its connections to related work are discussed in Sec. 4. Then a realization of the derived solution with SVMs is given in Sec. 5. Finally we present our empirical study in Sec. 6 and conclude the whole paper in Sec. 7.

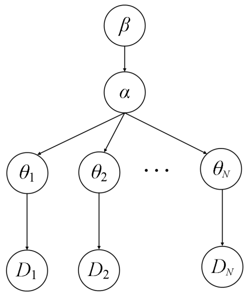


Figure 1: An illustration of the described hierarchical Bayesian model for information filtering

2. BAYESIAN MODELS FOR INFORMATION FILTERING

In the following, we assume that the filtering system is working on a set of items (e.g. articles or images), each one being represented by a vector of features $x \in \mathbb{R}^d$. Also, we have annotation data from n different users. Annotation data for user i consists of a set of rated items $\mathcal{R}_i$, together with a set of ratings $\{y_{i,j}\}, j \in \mathcal{R}_i$, where each rating $y_{i,j}$ is either $+1$ (liked that particular item) or -1 (disliked)[1]. The overall annotation data for user i, $i = 1, \ldots, n$ is denoted by $\mathcal{D}_i = \{(x_j, y_{i,j}) \mid j \in \mathcal{R}_i,\ y_{i,j} \in \{+1, -1\}\}$.

2.1 Non-Hierarchical Models

Given observations $\mathcal{D}_i$ from user i, a statistical content-based approach normally learns a predictive model, represented by parameters θ_i, and then applies it to make predictions via $p(y|x, \theta_i)$, which describes the predictive distribution of user i's rating y on some item, based on a vector of features x. Furthermore, a Bayesian approach takes into account the uncertainty of the models and makes predictions as follows:

$$p(y|x, \mathcal{D}_i) = \int_\theta p(y|x, \theta) p(\theta|\mathcal{D}_i) d\theta, \qquad (1)$$

where

$$p(\theta|\mathcal{D}_i) = \frac{p(\mathcal{D}_i|\theta)p(\theta)}{p(\mathcal{D}_i)}.$$

Here the prior $p(\theta)$ reflects our prior knowledge of the domain, and often prefers low-complexity models. In conventional information filtering, $p(\theta)$ is specified before seeing any data and remains fixed in the learning phase. The knowledge gained from some users (e.g. two web pages are often co-visited) can not be propagated to other users. In this way users are treated *independently*.

2.2 Hierarchical Models

[1] We restrict the discussion here to models for binary annotation data. But this restriction can be released.

However, people's information needs should be *related* in some way. Their connections can be characterized by exploring the structure of the data generating process. We introduce the common prior distribution. The θ_i for user i is a random sample from this common prior distribution. Then the overall observations $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\}$ are hierarchically generated as follows,

1. For the whole population of users, generate a sample of parameters $\alpha \sim p(\alpha|\beta)$, which defines the common prior $p(\theta|\alpha)$,
2. For each user $i = 1, \dots, n$, generate a sample $\theta_i \sim p(\theta|\alpha)$,
3. Given θ_i and a set of randomly chosen items $\mathcal{R}_i$, generate ratings $y_{i,j} \sim p(y|x_j, \theta_i)$, for $j \in \mathcal{R}_i$.

This hierarchical data generative process is described in Fig. 1. Then the likelihood of the overall observed annotations can be written as

$$p(\mathcal{D}|\alpha) = \prod_i \int_{\theta_i} p(\mathcal{D}_i|\theta_i)p(\theta_i|\alpha)d\theta_i, \qquad (2)$$

where

$$p(\mathcal{D}_i|\theta_i) = \prod_{j\in\mathcal{R}_i} p(y_{i,j}, x_j|\theta_i) = \prod_{j\in\mathcal{R}_i} p(y_{i,j}|x_j, \theta_i)p(x_j), \quad (3)$$

where we assume that the selection of items is random, independent of user profiles, and that the ratings for selected items depend on the user profiles. We assume that β is specified by the user and is a fixed parameter.

Though samples of θ are not directly visible, we can learn the parameters α of the common prior by computing its *a posteriori* distribution,

$$p(\alpha|\mathcal{D}, \beta) = \frac{p(\mathcal{D}|\alpha)p(\alpha|\beta)}{\int_\alpha p(\mathcal{D}|\alpha)p(\alpha|\beta)d\alpha}, \qquad (4)$$

where $p(\alpha|\beta)$ is the prior distribution of α, which is called the *hyper prior distribution* and β the corresponding parameters. By marginalizing out α, one can learn the common prior distribution as

$$p(\theta|\mathcal{D}, \beta) = \int p(\theta|\alpha)p(\alpha|\mathcal{D}, \beta)d\alpha. \qquad (5)$$

Then predictions for a new active user a with data $\mathcal{D}_a$ are made by

$$p(y|x, \mathcal{D}_a; \mathcal{D}, \beta) = \int_\theta p(y|x, \theta)p(\theta|\mathcal{D}_a; \mathcal{D}, \beta)d\theta \qquad (6)$$

where

$$p(\theta|\mathcal{D}_a; \mathcal{D}, \beta) = \frac{p(\mathcal{D}_a|\theta)p(\theta|\mathcal{D}, \beta)}{\int_\theta p(\mathcal{D}_a|\theta)p(\theta|\mathcal{D}, \beta)d\theta}.$$

Comparing Eq. (1) and Eq. (6), we can see that now the predictions for user a do not only depend on his/her own existing data $\mathcal{D}_a$, but also depend on other users data $\mathcal{D}$.

Non-hierarchical models described in Sec. 2.1 treat users separately, which is a conventional way of CBF (and also relevant feedback in information retrieval). Since each model is typically associated with a small sample set (as users do not bother to give many annotations), the method often suffers from overfitting. In contrast, a hierarchical Bayesian model puts dependence between these models by using the common prior $p(\theta|\alpha)$. Hence individuals are able to inherit knowledge from each other, and, as additional benefit, overfitting is avoided.

2.3 Nonparametric Hierarchical Models

One may first specify some parametric form for the common prior $p(\theta|\alpha)$ and then learn the parameters α (or their posterior distribution using Eq. (4)). However, due to the nature of the problem[2], the common prior is normally complex and cannot be adequately described by any known parametric form (like a Gaussian). Thus we replace the parametric common prior $p(\theta|\alpha)$ by a nonparametric prior distribution $G(\theta)$, which is generated from a *Dirichlet process* (DP) [1, 8]:

$$G|\beta \sim \mathrm{DP}(\tau, G_0), \qquad (7)$$

where $\beta = \{\tau, G_0\}$ specifies a DP, τ is a nonnegative scalar, called *concentration parameter*, and $G_0(\theta)$ is called the *base distribution*. Similar to the finite-dimensional case where a sample from a Dirichlet distribution is a multinomial distribution, we can consider G to be an infinite-dimensional multinomial distribution that can be written as

$$G = \sum_{l=1}^{\infty} \gamma_l \delta_{\theta_l}, \qquad (8)$$

where θ_l are independently drawn from the base distribution G_0, γ_l are probability weights depending on τ, and δ_{θ_l} the distribution concentrated on points θ_l. The concentration parameter controls how far away a randomly drawn distribution G is from the base distribution G_0. Though the expectation of G is equal to G_0, the smaller τ is, the more likely it is that an individual G deviates from G_0.

If we have directly observed the realizations of profile models $\theta_1, \dots, \theta_n$ (which is typically not the case), we can integrate over G, and the common prior (i.e. the distribution of the next coming θ) becomes

$$\theta|\{\theta_i\}_{i=1}^n, \beta \sim \frac{\tau G_0 + \sum_{i=1}^n \delta_{\theta_i}}{\tau + n}. \qquad (9)$$

From the above equation we can see a very important property of samples θ sampled from a distribution G that is generated by a DP. For an intuitive understanding, let us imagine a process of assigning persons into interest clubs. Suppose that there is potentially an infinite number of clubs, then

- The first person comes and creates a club θ_1 based on the base distribution G_0;
- The second person may either follow the first person to join in the same club with probability $1/(\tau + 1)$, or creates a new club from the distribution G_0 with probability $\tau/(\tau + 1)$;
- As the process is going on, n persons have chosen their own clubs $\{\theta_i\}_{i=1}^n$ (which might not all be distinct). Then a new person will join in a club by either following previous persons based on the distribution $\frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}(\theta)$ with probability $n/(\tau+n)$, or might create a new club from distribution G_0 with probability $\tau/(\tau + n)$.

[2] (1) Profiling models must be tailored to applications, like hidden Markov model for web browsing or support vector machines for image retrieval; (2) The distribution of people's interests are very complex.

The process is also known as a case of the *Chinese restaurant process* in the statistical literature (see [5]). In the process, a new user has a large chance to create a new club if τ is very large, or to follow previous users when τ is small. Thus in this paper, the hyper prior—a Dirichlet process—reflects our prior knowledge about how strongly users are influenced by each other.

3. LEARNING THE NONPARAMETRIC HIERARCHICAL MODEL

However, in the application of information filtering, user profile models $\{\theta_i\}_{i=1}^n$ are not directly visible, and only associated annotations $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^n$ are observed. Then the dependence among users, described by the common prior, should not only reflect our prior knowledge (i.e. the Dirichlet process), but should also be adapted to empirical data $\mathcal{D}$.

Replacing α by G in Eq. (5), one has to integrate out G to compute the common prior. Since the integral over $p(G|\mathcal{D}, \beta)$ is infeasible, a Bayesian approximation requires Markov Chain Monte Carlo (MCMC) to sample $p(G|\mathcal{D}, \beta)$. Instead, Eq. (9) indicates that Gibbs sampling can directly sample $p(\theta|\mathcal{D}, \beta)$, avoiding the difficulty of generating samples of the infinite dimensional G (see details in [8, 11]). Unfortunately, MCMC methods like Gibbs sampling is computationally expensive and sometimes technically difficult. In this paper we deviate from a fully Bayesian treatment and use the less expensive expectation-maximization (EM) to find the *posterior mode* of G as an estimate of the common prior.

During the E-step, the *a posteriori* $p(\theta|\mathcal{D}_j)$ for each individual j is computed. The M-step updates the common distribution $p(\theta)$ for the population by smoothing the ensemble of individuals $p(\theta|\mathcal{D}_j)$ under the constraint of the DP prior. Due to the infinite degrees of freedoms in the assumed model, we cannot explicitly represent $p(\theta|\mathcal{D}_j)$.

Thus we rely on a variational approximation:

$$\hat{p}(\theta|\mathcal{D}_j) \approx \sum_{i=1}^{n} \xi_{i,j}\delta_{\theta_i^*}, \tag{10}$$

where $\theta_i^* = \arg\max_\theta p(\mathcal{D}_i|\theta, G_0)$ correspond to the maximum *a posteriori* (MAP) estimates of each individual model[3], and $\xi_{i,j}$ are variational parameters to be specified. The iteration proceeds by repeating the following two steps:

1. E-step: Based on $\hat{p}(\theta)$ derived from the last M-step, we can calculate $\xi_{i,j}$

$$\xi_{i,j} = \frac{p(\mathcal{D}_j|\theta_i^*)\hat{p}(\theta_i^*)}{\sum_{i=1}^n p(\mathcal{D}_j|\theta_i^*)\hat{p}(\theta_i^*)} \tag{11}$$

2. M-step: Then we update the common prior

$$\hat{G}(\theta) = \frac{\tau G_0 + \sum_{i=1}^n \xi_i \delta_{\theta_i^*}}{\tau + n}. \tag{12}$$

where $\xi_i = \sum_j^n \xi_{i,j}$.

At the beginning of the iterations, we initialize $\xi_i = 1$ for each i. Finally, replacing $p(\theta|\mathcal{D}, \beta)$ by $\hat{G}(\theta)$ in Eq. (6), we make predictions for the active user as follows

$$p(y|x, \mathcal{D}_a) = w_0 p(y|x, \mathcal{D}_a; G_0) + \sum_{i=1}^n w_i p(y|x, \theta_i^*) \tag{13}$$

where

$$w_0 = \frac{1}{Z}\tau p(\mathcal{D}_a|G_0), \qquad w_i = \frac{1}{Z}\xi_i p(\mathcal{D}_a|\theta_i^*) \tag{14}$$

where Z is the normalizer to make $w_0 + \sum_{i=1}^n w_i = 1$, and $p(y|x, \mathcal{D}_a; G_0)$ is calculated as Eq. (1), but with the prior $p(\theta)$ replaced by the base distribution of DP, i.e. $G_0(\theta)$.

The final predictive model Eq. (13) averages other users' predictive models and the active user's own predictor (a Bayesian content-based filter). The first term on the right side corresponds to the predictor based solely on the data of the active user $\mathcal{D}_a$. This is particularly useful when other existing profile models do not fit the active user a's interests very well. In the following subsections, we further study the properties of predictor Eq. (13) under various conditions.

3.1 Content-Based Filtering

The extreme case $\tau \to \infty$ imposes a very strong hyper prior. The base distribution G_0 thus dominates the learned common prior, no matter if we adopt the fully Bayesian solution or the EM estimate $\hat{G}$. This gives $G = G_0$, which implies that observations $\mathcal{D}$ does not change our knowledge about the distribution of profile models. Then the predictions are given by solely the first term in the predictor Eq. (13)

$$p(y|x, \mathcal{D}_a) = p(y|x, \mathcal{D}_a; G_0) \tag{15}$$

In this case the hierarchical model actually degenerates to the non-hierarchical Bayesian model, which is the Bayesian version of pure content-based filtering. The model treats different users separately in two perspectives. First, since the distribution of θ dose not adapt to the observations from other users, the predictions for a new user are independent to other users' opinions. Second, usually a rather wide G_0 is assumed, which implies *a priori* that random samples (i.e. profiling models) are likely to be different from each other.

3.2 Content-Enhanced Collaborative Filtering

In the other extreme case, let $\tau \to 0$, then predictions are given by

$$p(y|x, \mathcal{D}_a) = \sum_{i=1}^n w_i p(y|x, \theta_i^*). \tag{16}$$

Let's take a closer look at the weighting terms in equation Eq. (16) (also in Eq. (13)). ξ_i represents the *typicalness* of model θ_i^*. $p(\mathcal{D}_a|\theta_i^*)$ indicates how well profile model θ_i^* can explain the active user a's interests; Then w_i models how likely the active user has user i's profiling model. Therefore persons like-minded to the active user have more impacts in predicting a's interests, which is essentially also the idea of CF, but expressed in a probabilistic way. However, the derived algorithm Eq. (16) is not simply CF, but *content-enhanced* CF, in the sense that many content-based predictors are combined to make predictions. However, the predictor Eq. (13) indicates that in a more general case we should also include the active user's own model.

4. CONNECTIONS TO RELATED WORK

Our work not only offers a principled hybrid information filtering approach, but also generalizes a bunch of existing information filtering algorithms. We already know that

[3] Alternatively, maximum likelihood estimates can be used.

when we impose a very strong hyper prior, the algorithm degenerates to the pure CBF.

Now let us examine its connections to pure CF which — in contrast to our content enhanced CF— would also give valid predictions without useful features. When content features are absent, we can rely on the fact that if a user would be required to re-rate an item the user had already rated, the user would be consistent in that both ratings would be (nearly) identical. This fact can be implemented by using the previous rating of an already rated item instead of using the prediction of the user model. Then the Eq. (16) becomes very similar to memory-based CF [18, 20, 14]. Our methods differs in that we treat cases (i.e. users) with different typicalness (indicated by ξ_i) while other CF methods assume cases are equally typical. Interestingly, a similar effect can be mimicked by simply overfitting the model!

Furthermore, our work also generalizes or improves on many hybrid filtering algorithms. Melville et al [9] suggest to build content-based model for each user and then generate *pseudo ratings* for non-rated items. The augmented data are used to feed a memory-based CF algorithm. Since pseudo ratings may be not accurate, heuristics like harmonic mean weighting are developed to incorporate the confidence of pseudo ratings. Our algorithm Eq. (16) essentially shares the same idea, but is derived in a principled way, in which the confidence of pseudo ratings are smoothly handled by the *predictive distribution* of y. Moreover, Eq. (13) further suggests that the predictor conditioned on the active user's own data should also be included.

A big family of hybrid filtering algorithms (e.g. [12, 7]) firstly treat CBF and CF separately and then average both results to make final predictions. Eq. (13) improves them in two aspects: (1) the weighting terms can explicitly be computed; (2) the CF part can be content-enhanced. As another example, Fab [2] maintains user profiles based on content analysis. An item is recommended to a user both when it scores highly against the user's own profile, and when it is also rated highly by users with similar profile. Eq. (13) resembles the Fab system in the way that it combines the active user's own model and other's models to make predictions.

5. REALIZATION WITH SUPPORT VECTOR MACHINES

So far we have studied the general theoretical framework of nonparametric hierarchical Bayesian solutions to information filtering, but we have not yet specified the detailed model $p(y|x, \theta)$. In principle, $p(y|x, \theta)$ can be any kind of a probabilistic predictive model. In the following we will discuss support vector machines (SVMs) models.

We consider SVM models for the preferences of user i, based on the ratings $\mathcal{D}_i$ this user has provided. A standard SVM would predict user i's binary rating y on some item x as $y = \mathrm{sign}(f^i(x))$. We follow the idea of [15], and compute the probability of membership in class y, $y \in \{+1, -1\}$ as

$$p(y|x, \theta_i) = \frac{1}{1 + \exp(yA_i f^i(x))} \tag{17}$$

A_i is the parameter to determine the slope of the sigmoid function. This modified SVM retains the decision boundary $f^i(x) = 0$, yet allows an easy approximation of posterior class probabilities.

Content-based methods using SVMs suffer from the problem of high variance of profiling models (due to the insufficient amount of training data from each individual). Now, we improve the performance of information filtering systems by applying Eq. (13). We maintain a set of profile models θ_i from users who have used the system[4]. Then we use the EM learning to estimate ξ_i, as discussed in Sec. 3(We found that, when n is very large, simply setting $\xi_i = 1$ also gives reasonable results.). Now suppose a new active user a with annotated $\mathcal{D}_a$. We first learn his/her profile model θ_a and then make predictions by

$$p(y|x, \mathcal{D}_a) = w_0 p(y|x, \theta_a) + \sum_{i=1}^{n} w_i \cdot p(y|x, \theta_i) \tag{18}$$

where the weights w_0 and w_i are computed as Eq. (14), the concentration parameter τ is set by cross-validation. To compute $p(\mathcal{D}_a|G_0)$ in Eq. (14), we assume that G_0 specifies a flat distribution where each item has equal chance to be liked or disliked. Eq. (18) realizes Eq. (13) via substituting $p(y|x, \mathcal{D}_a, G_0)$ by $p(y|x, \theta_a)$ and θ_i^* by parameters of the learned SVMs θ_i. The next section will demonstrate the success of this approach on two data sets.

6. EMPIRICAL STUDY

Empirical evaluations of our learning method are conducted in the following two experimental settings: (1) *Simulation on 4533 painting images*—From *Meisterwerke der Malerei* CDs we collected 4533 painting images. To enable an extensive objective measure of performance, we categorized them into 58 categories, mainly according to their respective artists. One artist corresponds to one category; (2) *Online survey on 642 painting images*—We collected 642 painting images from 30 artists. A web-based online survey[5] is built to gather user ratings. In the survey, each user gave ratings, i.e. "like", "dislike", or "not sure", to a randomly selected set of painting images. Finally we got a total of $L = 190$ users' ratings. On average, each of them had rated 89 images.

For all the images, we extract and combine *color histogram* (216-dim.), *correlagram* (256-dim.), *first and second color moments* (9-dim.) and *Pyramid wavelet texture* (10-dim.) to form 491-dimensional feature vectors to represent the images. The SVMs employed RBF (radius basis function) kernels.

We will mainly examine the performance of various algorithms in terms of their accuracy in predicting users' interests in painting images. These algorithms are: (1) *Hybrid filtering 1* that applies SVMs to implement the suggested algorithm Eq. (18), which essentially realizes the spirit of the derived nonparametric hierarchical Bayesian solution; (2)*Hybrid filtering 2* that applies SVMs to implement the suggested content-enhanced collaborative filtering. As Eq. (16), it differs from Eq. (18) by setting $\tau = 0$; (3) *Collaborative filtering* (CF) that combines a society of advisory users' preferences to predict an active user's preferences. The combination is weighted by *Pearson correlation* between test user and other advisory users' preferences. The algorithm

[4] One may select a subset of users to get a compact model. Our current work dose not discuss this issue

[5] The survey can be found on `http://honolulu.dbs.informatik.uni-muenchen.de:8080/paintings/index.jsp`.

applied here is described in [6]; (4) *SVM Content-Based filtering* (CBF) that trains a SVM model on a set of examples given by an active user, and then applies the model to predict the active user's preferences, which represents a typical content-based approach.

These algorithms are evaluated by two metrics. One is *Top-N accuracy*, i.e. the proportion of truly liked images among N top ranked images. Since normal users only care about the quality of first returned items, this quantity reflects the *subjective* quality of an information filter system. The other is ROC (receiver operating characteristics) curve, which plots *sensitivity* versus 1-*specificity*. Sensitivity is defined as the probability that a good image is recommended by the system; and specificity is the probability that a disliked image is rejected by the system. By changing the cut point (e.g. return top 10 or 20 images), a curve can be plotted. The ROC curve is insensitive to the prior distribution of liked (or disliked) images. The area under the curve, called *ROC sensitivity*, measures the *objective* quality of ranking. A higher ROC sensitivity indicates a better ranking.

6.1 Simulation with 4533 Painting Images

To enable an objective evaluation, we need to "mimic" many users' preferences for the images. We assume that each user is interested in m out of the 58 categories: For each user, we randomly choose m categories from the image database and assume the user is interested in the images in the selected categories. Then we randomly select up to 10 liked images and 10 disliked images according to the users' interests. We repeat the procedure 1000 times and thus generate 1000 user preference data. Since painting images from the same artist (e.g. one category) typically share similar painting styles, the simulation reflects real-world cases to some extent. By increasing the value of m, the users' interests are getting more different from each other. We are particularly interested in how this *heterogenous* effect influences the performances of various algorithms.

Our experiments assumes a 10-fold cross validation scheme, in which users are divided into 10 groups and where each group is selected as test users (i.e. active users) and the remaining ones serve as the database $\mathcal{D}$. For each test user, we predict her/his interests for the remaining images based on observed 20 annotations (10 positive and 10 negative). We compute the mean and standard deviation of the mean of Top-N accuracy, by firstly averaging over all the test users within each test group, and then over all test groups. We also plot the ROC curve by varying N from 1 to 4533[6].

We at first study the case that each user is interested in only one of the 58 categories. The situation indicates that users are likely to be related to each other since 1000 users are assigned to just 58 possibilities. The results have been shown in Fig. 2 (1-a,b). Content-based approach shows a poor accuracy, which indicates that the non-hierarchical model suffers from the small sample problem. CF significantly outperforms CBF. This is because in this simple case, with high probability, there will be a user very similar to the test user in the data base. However, due to the sparsity of user ratings (only 20 ratings observed for each user), CF dose not sufficiently takes advantage of this fact. That is why a large improvement is achieved by the two hybrid filtering algorithms. For the hybrid filtering 1, cross validation suggests $\tau = 0$, representing the prior belief (indicated by a Dirichlet process) that users are closely related. Thus hybrid filtering 1 degenerates to hybrid filtering 2 and the two curves in Fig. 2 (1-a,b) overlap completely. In addition, we have seen the bumps of the ROC curve for CF, currently we are yet not able to explain the reason.

We slightly increase the complexity of the experiment by allowing each user to be interested in 2 categories, indicating totally $58 \times 58 = 3364$ profile possibilities. The performances of four algorithms are demonstrated in Fig. 2 (2-a,b). Since users are less likely to be related, we can find that CBF does much better than CF. However, the two hybrid algorithms again achieve the best performance. For hybrid filtering 1, the concentration parameter τ is set to be 6000, implementing a situation where, with high probability, user profiles are different. Fig. 2 (2-a,b) indicates that hybrid filtering 1 is slightly better than hybrid filtering 2.

Finally, we assume that each user is interested in 3 out of 58 categories. In this case, a user's profile can be one of totally $58 \times 58 \times 58 = 195112$ types with equal chance, making users very *unlikely* to be related. The experimental results are summarized in Fig. 2 (3-a,b). We observed that, unlike in the former two cases, hybrid filtering 2 does much worse than content-based algorithm. Instead, hybrid filtering 1 ($\tau = 500000$) is slightly better than CBF. The concentration parameters in hybrid filtering 1 are set to be very large, reflecting the assumption that each user is very likely to be different from the other users. With such a hyper prior, the hierarchical model is analog to non-hierarchical models, i.e. CBF.

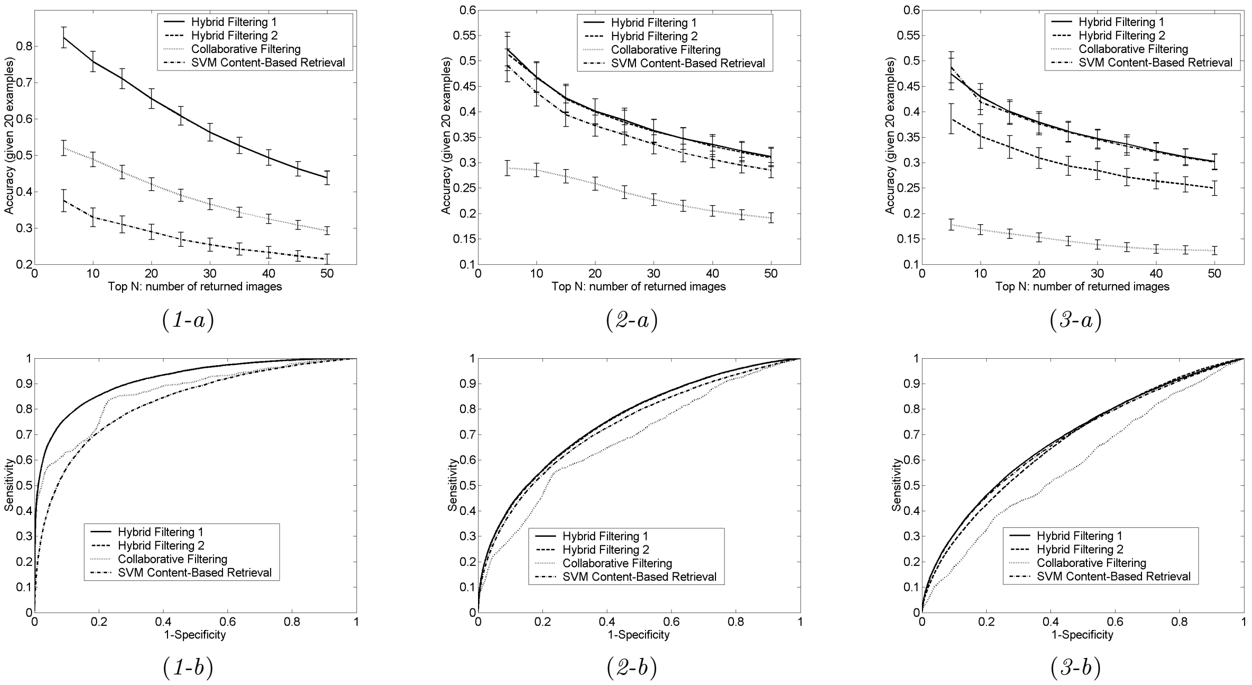
The empirical study demonstrates that, though different algorithms demonstrate each own strengths under various conditions, hybrid filtering 1, derived from the nonparametric hierarchical Bayesian framework, has the flexibilities to be adapted to different situations and always gives superior performance. In common cases hybrid filtering 2 (i.e. content-enhanced CF) shows good results too.

6.2 Experiments with the Online Survey Data

In this section, we will examine the performance of the four approaches based on 190 users' preference data on 642 painting images, which are gathered from the on-line survey. Again, we use top-N accuracy and ROC curve to evaluate the performance. Since we can not require a user to rate all of the 642 painting images in the survey, for each user we just partially know the "ground truth" of preferences. As a result, the true top-N accuracy cannot be computed. We thus adopt the accuracy measure that is the fraction of *known* liked images in the top ranked N images. The quantity is smaller than the true accuracy because *unknown* liked images are missing in the measurement. However, in our survey, the presenting of images to users is completely random, thus the distributions of rated/unrated images in both unranked and ranked lists are also random. This randomness dose not change the relative values of compared methods but just the absolute values. Thus in our following experiment it still makes sense to use the adopted accuracy measurement to compare the three retrieval methods. On the other hand, ROC is insensitive to this problem.

Our experiment again assumes the 10-fold cross validation scheme, in which we select each fold as a set of active users and treat the rest as users who have been previously seen. We fix the number of given examples for each active user to

[6] In the case $N = 4533$, the system recommends all the images to users.



(*1-a*) (*2-a*) (*3-a*)

(*1-b*) (*2-b*) (*3-b*)

Figure 2: Performance of 4 algorithms under three different conditions where each user is assumed to be interested in 1 category (1-a,1-b), 2 categories (2-a,2-b) or 3 categories (3-a,3-b). The left is Top-N accuracy and the right are ROC curves.

be 20 (10 positive and 10 negative), and predict the user's interests in the remaining painting images. For each active user, we repeat the experiment 10 times to randomize the 20 seen ratings and predict the rest. At the end, the overall average performance and its error bar are computed. Fig. 3 (a, b) shows the results. Both Top-N accuracy and ROC curve clearly indicate that two hybrid algorithms outperform CF and CBF. We found that the extracted image features are poor indicators of human interests. The content-based approach thus suffers from this problem and results in the worst predictions.

The ROC curve dose not show much difference between the two hybrid filtering algorithms. However, Top-N accuracy suggests that hybrid filtering 1 is slightly better. It is interesting to see the changes of Top-N accuracy curve with different values of concentration parameter τ, which is shown in Fig. 3 (c). When $\tau = 10000$, the hybrid filtering 1 reaches the optimal performance. As τ deviates away the performance degrades.

7. CONCLUSIONS AND FUTURE WORK

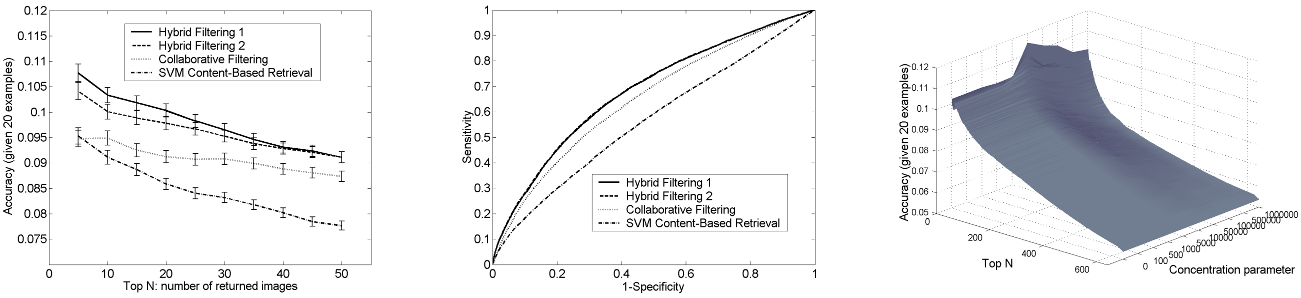
This paper describes a *nonparametric hierarchical Bayesian framework* to information filtering. In the framework, each user is modelled by a parametric content-based profile model, whose parameters θ are generated from a common prior distribution $p(\theta)$, which is shared by all the users. Then the model essentially allows different models to inherit knowledge from each other. We adopt a nonparametric form for the common prior, which is generated from a Dirichlet process (i.e. the hyper prior). We have shown that this nonparametric Bayesian treatment offers great advantages, including (1) flexible modelling ability and (2) meaningful insights into the problem. The hyper prior reflects the prior knowledge that users are related. We derive effective ways to learn the model from data. The resultant formulism is generally applicable to a wide range of probabilistic profiling models. The connections of this work to other related work are also clarified. Many independently proposed methods can now be interpreted from a unified view. Our empirical study initially demonstrates the excellent performance of this work.

Nonparametric Bayesian modelling has a long tradition in statistics and was recently introduced into the machine learning community. Its applications to information retrieval, like text modelling, document clustering, and user grouping, is still at the earliest stage. One recent related work is the hierarchical topic models proposed by Blei et al.[5], which applies Dirichlet process using MCMC to cluster documents by using MCMC. It has been known that hierarchical modelling using Dirichlet process with MCMC can discover the *clustering structure* of data and naturally find the number of clusters [8, 17], while conventional clustering methods suffer from deciding the number of clusters. Our current work aimed at realizing efficient learning and prediction based on an EM algorithm. However, as part of future work, it would also be interesting to implement the full MCMC solution to find groups of related users.

8. REFERENCES

[1] C. E. Antoniak. Mixtures of dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6), Nov. 1974.

[2] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.



(*b*) (*a*) (*c*)

Figure 3: (a)Top-N accuracy; (b) ROC curves; (c) Top-N accuracy of hybrid filtering 1 for different values of the concentration parameter

[3] C. Basu, H. Hirsh, and W. W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the Fifteenth National Conference on Artificial Intelligencen AAAI/IAAI*, pages 714–720, 1998.

[4] D. Billsus and M. J. Pazzani. Learning collaborative information filters. In *Proceedings of the 15th International Conference on Machine Learning*, pages 46–54. Morgan Kaufmann, San Francisco, CA, 1998.

[5] D. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarichical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[6] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52, 1998.

[7] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin. Combining content-based and collaborative filtering in an online newspaper. In *Proceedings of ACM SIGIR Workshop on Recommender Systems*, August 1999.

[8] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), June 1995.

[9] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002)*, pages 187–192, Edmonton, Canada, 2002.

[10] R. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference on Digital Libaries*, pages 195–204, San Antonio, US, 2000. ACM Press, New York, US.

[11] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. Technical Report 9815, Dept. of Statistics, University of Toronto.

[12] M. Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5–6):393–408, 1999.

[13] M. Pazzani, J. Muramastsu, and D. Billsus. Syskill and webert: Identifying interesting web sites. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 54–61, Portland, OR, August 1996.

[14] D. M. Pennock, E. Horvitz, S. Lawrence, and C. Giles. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 473–480, 2000.

[15] J. C. Platt. Probabilities for SV machines. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74, Cambridge, MA, 1999. MIT Press.

[16] A. Popescul, L. Ungar, D. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *17th Conference on Uncertainty in Artificial Intelligence*, pages 437–444, Seattle, Washington, August 2–5 2001.

[17] C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of gaussian process experts. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, 2002.

[18] P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 Computer Supported Collaborative Work Conference*, pages 175–186. ACM, 1994.

[19] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, 1971.

[20] U. Shardanand and P. Maes. Social information filtering algorithms for automating 'word of mouth'. In *Proceedings of ACM CHI'95 Conference*, 1995.

[21] K. Yu, A. Schwaighofer, V. Tresp, W.-Y. Ma, and H. Zhang. Collaborative ensemble learning: Combining collaborative and content-based information filtering via hierarchical Bayes. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.