Latent Class Models for Collaborative Filtering

Thomas Hofmann
CS Division, UC Berkeley
and International CS Institute
Berkeley, CA, USA
hofmann@cs.berkeley.edu

Jan Puzieha
Institut fur Informatik
University of Bonn
Bonn, Germany
jan@cs.uni-bonn.de

Abstract

This paper presents a statistical approach to collaborative filtering and investigates the use of latent class models for predicting individual choices and preferences based on observed preference behavior. Two models are discussed and compared: the aspect model, a probabilistic latent space model which models individual preferences as a convex combination of preference factors, and the two-sided clustering model, which simultaneously partitions persons and objects into clusters. We present EM algorithms for different variants of the aspect model and derive an approximate EM algorithm based on a variational principle for the two-sided clustering model. The benefits of the different models are experimentally investigated on a large movie data set.

1 Introduction

The rapid growth of digital data repositories and the overwhelming supply of on-line information provided by today's communication networks bears the risk of constant information overload. *Information filtering* refers to the general problem of separating useful and important information from nuisance data. In order to support individuals with possibly different preferences, opinions, judgments, and taste, in their quest for information, an automated filtering system has to take into account the diversity of preferences and the relativity of information value. One commonly distinguishes between (at least) two major approaches [Resnik *et* a/., 1994]: (i) *content-based* filtering organizes information based on properties of the object of preference or the carrier of information such as a text document, while (ii) *collaborative filtering* [Goldberg *et* al., 1992] (or *social filtering)* aims at exploiting preference behavior and qualities of other persons in speculating about the preferences of a particular individual.

1.1 Information Filtering

Most information filtering systems have been designed for a particular application domain, and a large fraction of the research in this area deals with problems of system architecture and interface design. In contrast, this paper will take a more abstract viewpoint in order to clarify some of the *statistical foundations* of collaborative filtering. In particular, the presupposition is made that no external knowledge beyond the observed preference or selection behavior is available, neither about properties of the objects (such as documents, books, messages, CDs, movies, etc.) nor about the involved persons (such as computer users, customers, cineasts, etc.). This working hypothesis is not as unrealistic as it may seem on first sight since, for example, many computer systems which interact with humans over the Web do not collect much personal data for reasons of privacy or to avoid time-consuming questionnaires. The same is often true for properties of objects where it is sometimes difficult to explicitly determine those properties that make it relevant to a particular person. Moreover, one might integrate information from both sources in a second step, e.g., by deriving prior probabilities from person/object features and then updating predictions in the light of observed choices and preferences.

1.2 Dyadic Data

We thus consider the following formal setting: Given are a set of persons $\mathcal{X} = \{x_1, \ldots, x_N\}$ and a set of objects $\mathcal{Y} = \{y_1, \ldots, y_M\}$. We assume that observations are available for person/object pairs *(x*,y), where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$; this setting has been called *dyadic data* in [Hofmann *et al,* 1999]. In the simplest case, an observation will just be the co-occurrence of *x* and y, representing events like "person *x* buys product y" or "person *x* participates in *y".* Other cases may also provide some additional preference value *v* with an observation. Here, we will only consider the simplest case, where $v \in \{-1, +1\}$ corresponds to either a negative or a positive example of preference, modeling events like "person *x* likes/dislikes object y".

Two fundamental learning problems have to be addressed: (i) probabilistic modeling and (ii) structure dis-

covery. As we will argue, different statistical models are suitable for either task. The aspect model presented in Section 2 is most appropriate for prediction and recommendation, while the two-sided clustering model introduced in Section 3 pursues the goal of identifying meaningful groups or *clusters* of persons and objects. All discussed models belong to the family of *mixture models,* i.e., they can be represented as *latent variable models* with discrete latent variables. The main motivation behind the introduction of latent variables in the context of filtering is to explain the observed preferences by some smaller number of (typical) *preference patterns* which are assumed to underly the data generation process. In probabilistic modeling, this is mainly an attempt to overcome the omnipresent problem of data sparseness. Models with a reduced number of parameters will in general require less data to achieve a given accuracy and are less sensitive to overfitting. In addition, one might also be interested in the structural information captured by the latent variables, for example, about groups of people and clusters of objects.

2 The Aspect Model

2.1 Model Specification

In the aspect model [Hofmann *et* al., 1999], a latent class variable $z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$ is associated with each observation (x, *y).* The key assumption made is that *x* and *y* are independent, conditioned on z. The probability model can thus simply be written as

$$P(x,y) = \sum_{z \in \mathcal{Z}} P(z)P(x|z)P(y|z), \qquad (1)$$

where $P(x|z)$ and $P(y|z)$ are class-conditional multinominal distributions and *P(z)* are the class prior probabilities. Notice that the model is perfectly symmetric with respect to the entities *x* and *y.* Yet, one may also re-parameterize the model in an asymmetric manner, e.g., by using the identity $P(z)P(x|z) = P(x,z) = P(x)P(z|x)$ which yields

$$P(x,y) = P(x)P(y|x), \text{ where} \qquad (2)$$

$$P(y|x) = \sum_{z \in \mathcal{Z}} P(z|x)P(y|z). \qquad (3)$$

A dual formulation can be obtained by reversing the role of *x* and *y.* Eq. (3) is intuitively more appealing than (1) since it explicitly states that conditional probabilities $P(y|x)$ are modeled as a convex combination of *aspects* or *factors* $P(y|z)$. In the case of collaborative filtering, this implies that the preference or selection behavior of a person is modeled by a combination of *typical preference patterns,* represented by a distribution over objects. Notice that it is neither assumed that persons form 'groups', nor is stipulated that objects can be partitioned into 'clusters'. This offers a high degree of flexibility in modeling preference behavior: Persons may have a multitude of different interests, some of which they might share with some people, some with others, a fact which can be expressed perfectly well in the aspect model. It is also often the case that objects are selected by different people for different reasons. In this case, one might have a number of aspects with high probability *P(y\z)* for a particular object *y.*

2-2 Model Fitting by EM

The standard procedure for maximum likelihood estimation in latent variable models is the Expectation Maximization (EM) algorithm [Dempster *et* al., 1977]. EM alternates two steps: (i) an expectation (E) step where posterior probabilities are computed for the latent variables z, based on the current estimates of the parameters, (ii) an maximization (M) step, where parameters are updated for given posterior probabilities computed in the previous E-step.

For the aspect model in the symmetric parameterization Bayes' rule yields the E-step

$$P(z|x,y) = \frac{P(z)P(x|z)P(y|z)}{\sum_{z'} P(z')P(x|z')P(y|z')}. \qquad (4)$$

By standard calculations one arrives at the following M-step re-estimation equations

$$P(y|z) = \frac{\sum_x n(x,y)P(z|x,y)}{\sum_{x,y'} n(x,y')P(z|x,y')}, \qquad (5)$$

$$P(x|z) = \frac{\sum_y n(x,y)P(z|x,y)}{\sum_{x',y} n(x',y)P(z|x',y)}, \qquad (6)$$

where n(x, *y)* denotes the number of times the pair *(x, y)* has been observed. Alternating (4) with (5) and (6) defines a convergent procedure that approaches a local maximum of the log-likelihood.

Implicit in the above derivation is a multinomial sampling model, which in particular implies the possibility of multiple observations. This may or may not be appropriate and one might also consider hypergeometric sampling without replacement, although according to statistical wisdom both models are expected to yield very similar results for large populations.

2.3 Extension to Preference Values

Let us now focus on extending the aspect model to capture additional binary preferences $v \in \{-1, +1\}$.[1] We distinguish two different cases: (I.) situations where the selection of an object is performed by the person, which then announces her or his preference in retrospect, (II) problems where the selection of *y* is not part of the behavior to be modeled, for instance because it is controlled or triggered by some other external process.

[1]The presented models can be further generalized to handle arbitrary preference values, but this requires to specify an appropriate likelihood function based on assumptions on the preference *scale.*

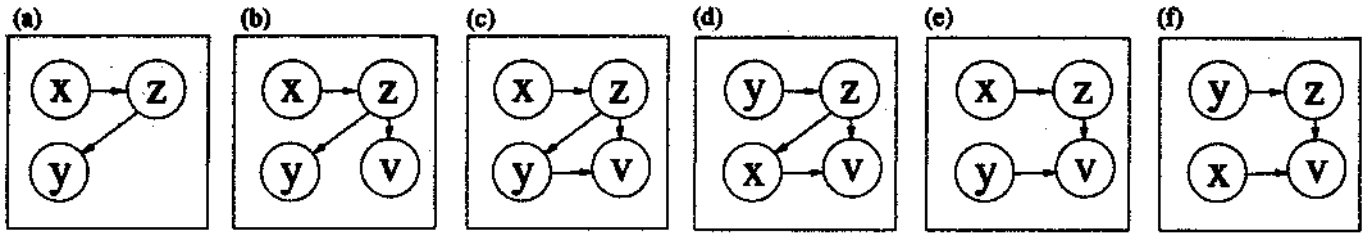

Figure 1: Graphical model representation of the aspect model (a) and its extensions to model preference values (b)-(d) (case I) and (e),(f) (case II).

Case I. In the first case, there are three different ways to integrate the additional random variable v into the model, as shown by Figure 1 (b)-(d). In (b) v is conditionally independent of *x* and y given *z,* which is a very strong assumption. One implication is that aspects are typically employed to either model positive or negative preferences. In variant (c) and (d), v also depends on either *x* or y which offers considerably more flexibility, but also requires to estimate more parameters. It is straightforward to modify the EM equations appropriately. We show the equations for model (c), the other variants require only minor changes. For the E-step one obtains

$$P(z|x,y,v)=\frac{P(z)P(x|z)P(y,v|z)}{\sum_{z'}P(z')P(x|z')P(y,v|z')} \qquad (7)$$

while the M-step equations can be summarized into

$$P(y,v|z)=\frac{\sum_x n(x,y,v)P(z|x,y,v)}{\sum_{x,y',v'} n(x,y',v')P(z|x,y',v')}, \qquad (8)$$

where *n(x,* y, *v)* denotes the number of times a particular preference has been observed (typically n(x,y, *v)* $\in$ {0,1}). From *P(y, v\z)* one may also derive $P(y|z)$ and *P(v\y,* z), if necessary. The M-step equation for *P{x\z)* does not change. Effectively the state space of *y* has been enlarged to $\mathcal{Y}' = \mathcal{Y} \times \{-1,+1\}$.

Notice that one might also consider to combine the model variants in Figure 1 by making different conditional independence assumptions for different values of *z.* The resulting combined model corresponds to a Bayesian multinet [Geiger and Heckerman, 1996].

Case II. In the second case, the multinomial sampling model of selecting *y* or a (y, *v)* pair conditioned on *z* is no longer adequate. We thus propose a modification of the aspect model starting from (3) and replace multinomials *P(y\z)* with Bernoulli probabilities *P(v\y,* z), assuming that y is always conditioned on (cf. Figure 1 (e)). This modification results in the E-step

$$P(z|x,y,v)=\frac{P(z)P(x|z)P(v|y,z)}{\sum_{z'}P(z')P(x|z')P(v|y,z)}. \qquad (9)$$

and a M-step re-estimation formula

$$P(v|y,z)=\frac{\sum_x P(z|x,y,v)}{\sum_{v=\pm 1}\sum_x P(z|x,y,v)}. \qquad (10)$$

Comparing (9) with (7) one observes that $P(y,v|z)$ is now replaced by $P(v|y,z)$ since *y* is treated as a fixed (observation-dependent) conditioning variable. Note that by conditioning on both, *x* and y, one gets $P(v|x,y)=\sum_z P(v|y,z)P(z|x)$ which reveals the asymmetry introduced into the aspect model by replacing one of the class-conditional multinomials with a vector of Bernoulli probabilities. The presented version is the "collaborative" model. Reversing the role of *x* and *y* yields the dual counterpart in Figure 1 (f), where the prediction of *v* depends directly on *x* and only indirectly on *y* (through z). Again combining both type of dependency structures in a multinet might be worth considering.

3 The Two-Sided Clustering Model

3.1 Model Specification

In the two-sided clustering model the strong assumption is made that each person belongs to exactly one group of persons and that each object belongs to exactly one group of objects. Hence we have latent mappings $c(x) \in \mathcal{C} = \{c_1,\ldots,c_K\}$ and $d(y) \in \mathcal{D} = \{d_1,\ldots,d_L\}$ which partition *X* into *K* groups and *y* into *L* groups, respectively. This is very different in spirit from the aspect model, where the leitmotif was to use convex combinations of prototypical factors. While we expect the clustering model to be less flexible in modeling preferences and less accurate in prediction (a fact which could be verified empirically), it might nevertheless be a valuable model for structure discovery which has applications of its own right. To formalize the model, let us introduce the following sets of parameters: *P(x)* and *P(y)* for the marginal.probabilities of persons and objects, *P(c)* and *P(d)* for the prior probabilities of assigning persons/objects to the different clusters, and, most importantly, cluster association parameters $\phi(c,d) \in \mathbb{R}_0^+$ between pairs of cluster *(c,d).* Now we may define a probabilistic model by

$$P(x,y|c(x)=c,d(y)=d;\phi)=P(x)P(y)\phi(c,d). \qquad (11)$$

A factorial prior on the latent class variables

$$P(c(x)=c)=P(c), \quad P(d(y)=d)=P(d) \qquad (12)$$

Star *Trek* IV 0.024 Star Trek.II 0.023 Star Trek VI 0.023 Star Trek III 0.021 The Fifth Element 0.018	Dr. Strangeiove 0.029 A Clockwork Orange 0.020 Delicatessen 0.018 Cinema Paradiso 0.018 Brazil 0.017717	Pinocchio 0.281 The Aristocats 0.213 Snow White and the Seven Dwarfs 0.211 The Jungle Book 0.049 The Lion King 0.020	Richard III 0.160 Les Miserables 0.124 The Madness of King George 0.113 In the Name of the Father 0.076 The Visitors (Les Visiteurs) 0.043
The Rock 0.553 Eraser 0.232 Independence Day (ID4) 0.089 Mission: Impossible 0.077 Trainspotting 0.021	The Piano 0.288 The Remains of the Day 0.077 In the Name of the Father 0.067 Forrest Gump 0.052 Shadowlands 0.047	Ready to Wear 0.097 What's Love Got To Do With It? 0.091 Circle of Friends 0.070 Dolores Claiborne 0.037 When a Man Loves a Woman: 0.030	Como Agua Para Chocolate 0.132 Three Colors: Red: 0.086 Three Colors: Blue: 0.079 Three Colors: White: 0.068 The Piano: 0.064

Figure 2: Movie aspects extracted from EachMovie along with the probabilities $P(y|z)$.

completes the specification of the model. The association parameters ϕ increase or decrease the probability of observing a person/object pair *(x,y)* with associated cluster pair (c, *d)* relative to the unconditional independence model P(x,y) = *P(x)P(y).* In order for (11) to define a proper probabilistic model, we have to ensure a correct global normalization, which constrains the choice of admissible values for the association parameters ϕ.

3.2 Variational EM for Model Fitting

The main difficulty in the two-sided clustering model is the coupling between the latent mappings *c(z)* and *d(y)* via the cluster association parameters $\phi(c,d)$. An additional problem is that the admissible range of ϕ also depends on c(x) and *d(y).* Since an exact EM algorithm seems to be out of reach, we propose an approximate EM procedure. First, since *c(x)* and *d(y)* are random variables we define the admissible range of ϕ to be the set of values for which

$$\mathbf{E}\left[\sum_{x,y} P(x)P(y)\phi(c(x),d(y))\right] = 1, \qquad (13)$$

where the expectation is taken w.r.t. the posterior class probabilities

$$P^{x,y}_{c,d}(\phi) \equiv P(c(x)=c, d(y)=d|\mathbf{n},\phi). \qquad (14)$$

Secondly, the posteriors $P^{x,y}_{c,d}(\phi)$ are approximated by a variational distribution of factorial form,

$$P^{x,y}_{c,d}(\phi) \approx Q(x,c)Q(y,d), \qquad (15)$$

where the *Q* distributions are free parameters to be determined. In the (approximate) M-step one has to maximize [Hofmann and Puzicha, 1998]

$$\mathcal{L}(\phi|\phi') = \sum_{x,y} n(x,y) \sum_{c,d} P^{x,y}_{c,d}(\phi') \log \phi(c,d), \qquad (16)$$

with respect to ϕ. Technically, one introduces a Lagrange multiplier to enforce (13) and after some rather lengthy calculations arrives at the equation

$$\phi(c,d) = \frac{\sum_{x,y} P^{x,y}_{c,d}(\phi')n(x,y)}{\left(\sum_x P^x_c(\phi')n(x)\right)\left(\sum_y P^y_d(\phi')n(y)\right)}, \qquad (17)$$

where P^x_c and P^y_d are marginals of the posteriors $P^{x,y}_{c,d}$ and n(x), *n(y)* are marginal counts. Eq. (16) can be given a very intuitive interpretation by considering the hard clustering case of $P^{x,y}_{c,d} \in \{0,1\}$, where it reduces to the expected mutual information between pairs of classes *c* and *d* in either spaces: the numerator in (17) then simply counts the number of times a person *x* belonging to a particular cluster *c* has been observed in conjunction with an object *y* from cluster *d,* while the denominator reduces to the product of the probabilities to (independently) observe a person from cluster c and an object from *d.*

It remains to perform the variational approximation and to determine values for the Q-distributions by choosing *Q* in order to minimize the KL-divergence to the true posterior distribution. Details on this method - also known as *mean-field approximation* - can be found in [Jordan *et a*/., 1998; Hofmann and Puzicha, 1998]. For brevity, we report only the final form of the variational E-step equations:

$$Q(x,c) \propto P(c)\exp\left[\sum_y n(x,y)\sum_d Q(y,d)\log\phi(c,d)\right], \qquad (18)$$

$$Q(y,d) \propto P(d)\exp\left[\sum_x n(x,y)\sum_c Q(x,c)\log\phi(c,d)\right]. \qquad (19)$$

Notice that these equations form a highly non-linear, coupled system of transcendental equations. A solution is found by a fixed-point iteration which alternates the computation of the latent variables in one space (or more precisely their approximate posterior probabilities) based on the intermediate solution in the other space, and vice versa. However, the alternating computation

Four Weddings and a Funeral	Apollo 13	E.T.: The Extraterrestrial	M*A*S*H	Kalifornia !
Home Alone	Batman	Alice in Wonderland	Full Metal Jacket	Short Cuts
Sleepless in Seattle	Batman Forever	Cinderella	The Bridge on the River Kwai	Smoke
Dave	Star Trek: Generations	Old Yeller	Apocalypse Now	Red Rock West
Pretty Woman	Stargate	Mary Poppins	Chinatown	Romeo is Bleeding
The Piano	Goldeneye	The Fox and the Hound	The Shining	Crumb

Figure 3: Movie clusters extracted from EachMovie.

has to be interleaved with a re-computation of the ϕ-parameters, because certain term cancelations have been exploited in the derivation of (18,19). The resulting alternation scheme optimizes a common objective function and always maintains a valid probability distribution. To initialize the model we propose to perform one-sided clustering, either in the *X* or the *y* space.

3.3 Clustering with Preference Values

Like the basic aspect model, the two-sided clustering model is based on multinomial sampling, i.e., it models independently generated occurrences of (x, *y)* pairs. To model preference values *v conditioned* on (x, *y)* pairs, we modify the model by replacing the association parameters ϕ with Bernoulli parameters $P(v|c,d)$,

$$P(v|x,y,c(x)=c,d(y)=d) = P(v|c,d)\,. \qquad (20)$$

The assumption is that v is independent of *x* and y, given their respective cluster memberships.[2] Although the latent mappings c and *d* are coupled, this model is somewhat simpler than the model in (11), since there is no normalization constraint one needs to take care of. The conditional log-likelihood is thus simply

$$\mathcal{L} = \sum_{x,y}\sum_{v=\pm 1} n(x,y,v)\log P(v|c(x),d(y)), \qquad (21)$$

where of course $P(-1|c,d) = 1 - P(+1|c,d)$. In the M-step we have to maximize $\mathbf{E}[\mathcal{L}]$, the expected log-likelihood under the posterior distribution of the latent mappings *c(x)* and *d{y)* which yields the formula

$$P(v|c,d) = \frac{\sum_{x,y} P^{x,y}_{c,d}(\phi)\, n(x,y,v)}{\sum_{v'=\pm 1}\sum_{x,y} P^{x,y}_{c,d}(\phi)\, n(x,y,v')}. \qquad (22)$$

In the hard clustering limit, this simplifies to counts of how many persons in cluster c like $(v = +1)$ or dislike (v = —1) objects from cluster *d* (ignoring missing values). The denominator then corresponds to the total number of votes available between *x's* belonging to c and y's belonging to *d.* A factorial approximation of the posterior probabilities $P^{x,y}_{c,d}$ along the same lines as discussed above, yields

$$Q(x,c) = P(c)\exp\left[\sum_{y,v,d} n(x,y,v)Q(y,d)\log P(v|c,d)\right], \qquad (23)$$

$$Q(y,d) = P(d)\exp\left[\sum_{x,v,c} n(x,y,v)Q(x,c)\log P(v|c,d)\right]. \qquad (24)$$

These equations are very similar to the ones derived in [Ungar and Foster, 1998]. The clustering model they present is identical to the Bernoulli model in (20), but the authors propose Gibbs sampling for model fitting, while we have voted for the computationally much faster variational EM algorithm.[3]

K(L)	Aspect Co-occ. (a)	Cluster Co-occ.	Aspect Pref. (d)
1(1)	442	442	827
8(8)	255	349	475
16 (16)	241	335	442
32 (32)	237 (228)	308	434 (401)
64(64)	234 (224)	341(301)	425 (395)
128(128)	231 (219)	380(298)	418 (388)

Table 1: Perplexity results on EachMovie for different model types (columns) and different model complexities (rows).

4 Experimental Results

To demonstrate the utility of latent class models for collaborative filtering, we have performed a series of experiments with the EachMovie dataset which consists of data collected on the internet (almost 3 million preference votes on a 0-5 scale which we have converted to $-1/+1$ preferences by thresholding).[4] Table 1. summa-

[2] Refined models may also consider additional weights to account for individual preference averages.

[3] For example, on the EachMovie database used in the experiments we were not able to train models with Gibbs sampling because of the immense computational complexity.

[4] For more information on this dataset see www/research.digital.com/SRC/EachMovie.

Data #1	Recommendations	Data #2	Recommendations
Star Trek: The Motion Picture	The Empire Strikes Back	Pulp Fiction	The Silence of the Lambs
Star Trek: Generations	Star Trek: First Contact	Fargo	Toy Story
Star Trek II	Raiders of the Lost Ark	Smoke	Dead Man Walking
Star Trek III	Stargate	Three Colors: Blue	Batman
Star Trek V	The Terminator	Four Weddings and a Funeral	Leaving Las Vegas
Star Trek VI	Return of the Jedi	A2001: A Space Odyssey	The Piano

Figure 4: Two exemplary recommendations computed with an aspect model *(K* = 128).

rizes perplexity results[5] obtained with the aspect model and the two-sided clustering model for different number of latent classes. As expected the performance of the aspect model is significantly better than the one obtained with the clustering model. The aspect model achieves a reduction of roughly a factor 2 over the marginal independence model (baseline at $K = 1$). By using *annealing* techniques (cf. [Hofmann and Puzicha, 1998]) slightly better results can be obtained (numbers in brackets).

To give an impression of what the extracted movie aspects and movie clusters look like, we have displayed some aspects of a $K = 128$ model in Figure 2 and clusters of a $K = L = 32$ solution represented by their members with highest posterior probability in Figure 3. Notice that some movies appear more than once in the aspects (e.g. 'The Piano'). Both authors have also been subjected to a test run with the recommendation system. The result - which was perfectly satisfying from our point of view - is shown in Figure 4. We hope the reader might also spot one or another valuable recommendation.

5 Conclusion

We have systematically discussed two different types of latent class models which can be utilized for collaborative filtering. Several variants corresponding to different sampling scenarios and/or different modeling goals have been presented, emphasizing the flexibility and richness of latent class models for both, prediction and structure discovery. Future work will address alternative loss functions and will have to deal with a more detailed performance evaluation.

Acknowledgments

This work has been supported by a DAAD postdoctoral fellowship (TH) and the German Research Foundation (DFG) under grant #BU 914/3-1 (JP). The EachMovie preference data is by courtesy of Digital Equipment Corporation and was generously provided by Paul McJones.

[5]The perplexity V is the log-scale average of the inverse probability $1/P(y|x)$ on test data.

References

[Dempster *et ai,* 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. *Royal Statist. Soc. B,* 39:1-38, 1977.

[Geiger and Heckerman, 1996] D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. *Artifical Intelligence,* 82(I):45-74, 1996.

[Goldberg *et al,* 1992] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry. Using collabrorative filtering to weave an information tapestry. *Communications of the ACM,* 35(12):61-70, 1992.

[Hofmann and Puzicha, 1998] T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. Technical report, Artifical Intelligence Laboratory Memo 1625, M.I.T., 1998.

[Hofmann *et al,* 1999] T. Hofmann, J. Puzicha, and M. I. Jordan. Learning from dyadic data. In *Advances in Neural Information Processing Systems 11,* 1999.

[Jordan *et al,* 1998] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models,* pages 105-161. Kluwer Academic Publishers, 1998.

[Resnik *et al,* 1994] P. Resnik, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM, Conference on Computer Supported Cooperative Work,* pages 175-186, 1994.

[Ungar and Foster, 1998] L. Ungar and D. Foster. A formal statistical approach to collaborative filtering. In *Conference on Automated Learning and Dicovery, CONALD'98, CMU,* 1998.