THE OVERFITTING IN PROBABILISTIC LATENT SEMANTIC MODELS[1]

V. A. Leksin[2], K. V. Vorontsov[3]

[2] CC RAS, Moscow, voron@ccas.ru
[3] MIPT, Moscow, vleksin@mail.ru

The symmetric EM algorithm is proposed for probabilistic latent semantic analysis in collaborative filtering. The algorithm allows to reveal the latent interest profiles of both users and items, then to easily construct high-quality similarity measures of all required types: user–user, item–item, and item–user. The advantage of the proposed approach is that different profiles are consistent to each other due to symmetry of the algorithm. To estimate the quality of profiles and similarity measures empirically we use a sample of labeled items and the *k*NN classifier. Experiment show that the excessive optimization is redundant and can lead to overfitting.

The revealing of user preferences and tastes based on behavior data (purchases, visits, queries, clicks, etc.) is a very important intermediate task in business intelligence. The final applications are recommender systems, direct marketing, personalized advertising, similarity search, similar minded people search in social networks, etc.

The data are represented by the user activities records such as "the user *u* chose the item *r*». The similarity measures both between users and between items are very convenient to solve a broad variety of applied problems.

Simple item-based and user-based techniques of web usage mining (WUM) [2] and *collaborative filtering* (CF) [3] exploit either users or item similarities. A drawback is that these approaches lead to different inconsistent results if applied to the same task, e.g. recommendation.

The main idea of *clients environment analysis*, CEA [1,7,9] is to use the consistent similarity measures. The *consistency* property implies that "items are similar if they are used by similar (not obligatory the same) users; on the other hand, users are similar if they use similar (not obligatory coincident) items". This definition is recursive by its nature and requires an iterative procedure. The straightforward implementation of such a reconciliation of all pair-wise similarities is very inefficient.

The *latent semantic analysis* (LSA) also used in CF is much more efficient [4], but it has nothing to do with consistency.

In this work we consider the *probabilistic LSA* (PLSA), which has more sound statistical grounds [5,6]. The main idea of PLSA is to reveal the latent properties of users and items, that can be interpreted as interests or topics. An efficient EM algorithm can be used to do this. The advantage is that similarity measure can be defined as a distance between vectors of latent properties. These vectors are called *profiles* in this paper. To ensure the consistency of users and items profiles we propose a symmetric variant of the EM algorithm.

To estimate quantitatively the goodness of profiles and similarity measures we use a sample of labeled items and a simple *k* nearest neighbors (*k*NN) classifier.

We study empirically the dependence of the quality of profiles on the structure parameters of our symmetric EM algorithm. It turns out that the excessive optimization is redundant and can lead to overfitting. The experiment on both real and model datasets show that the robust Euclidean distance between item profiles is a much more adequate (dis)similarity measure that the standard measures based on correlations or other statistical tools applied explicitly to the initial users × items co-occurrence matrix.

[1] RFBR grants 07-01-12076-офи and 08-07-00422.

The symmetric EM algorithm

Let U be a set of users, R be a set of items, and $D=(u_i,r_i)_{i=1}^{l}\subset U\times R$ is a given sample of *co-occurrence* observations. The goal is to induce similarity functions on users $\rho_U(u,u')$ and items $\rho_R(r,r')$ that will be helpful for user behavior prediction, recommendations, items catalogization, similarity search, etc.

Each user is assumed to be interested in a subset of topics from the set of topics T.

We call *a latent profile of the user* $u\in U$ a vector of (unknown) conditional probabilities $p_{tu}=p(t\mid u)$ that the user u is interested in the topic $t\in T$, where $\sum_{t\in T}p_{tu}=1$. By analogy, a *latent profile of the item* $r\in R$ is a vector of (unknown) conditional probabilities $q_{tr}=q(t\mid r)$ that the item r can satisfy an interest in the topic $t\in T$, where $\sum_{t\in T}q_{tr}=1$.

The intermediate goal is to learn the latent profiles $\{p_{tu},t\in T\}$ for all $u\in U$ and $\{q_{tr},t\in T\}$ for all $r\in R$ from data D.

The probability of co-occurrence (u,r) can be alternatively represented by two different generative models:

$$\begin{aligned}p(u,r)&=\sum_{t\in T}p_u p_{tu}q(r\mid t,u)=\\&=\sum_{t\in T}\Big(p_u p_{tu}q_{tr}q_r\Big/\sum_{r'\in R}q_{tr'}q_{r'}\Big);\end{aligned}\qquad(1)$$

$$\begin{aligned}p(u,r)&=\sum_{t\in T}q_r q_{tr}p(u\mid t,r)=\\&=\sum_{t\in T}\Big(q_r q_{tr}p_{tu}p_u\Big/\sum_{u'\in U}p_{tu'}p_{u'}\Big);\end{aligned}\qquad(2)$$

where p_u and q_r are prior probabilities of the occurrence of a user u and an item r respectively. Posterior probabilities $q(r\mid t,u)$ and $p(u\mid t,r)$ are expressed from latent profiles through the Bayes' theorem. Note that both equations (1) and (2) must hold but none of them can not be reduced to another one.

We use the log-likelihood maximization to learn the latent profiles:

$$\sum_{i=1}^{l}\ln p(u_i,r_i)\to\max,\qquad(3)$$

where maximum is over profiles $\{p_{tu}\}$ and $\{q_{tr}\}$ normalized so that $\sum_{t\in T}p_{tu}=1$ for all $u\in U$ and $\sum_{t\in T}q_{tr}=1$ for all $r\in R$.

The Expectation Maximization (EM) algorithm is a standard tool that is used to approximately maximize the log-likelihood in mixture models like (3). We use a modified symmetric version of EM algorithm [7] that takes into account both (1) and (2) representations. The outer loop of iterations consists of two steps:

1) optimize $\{p_{tu}\}$ for fixed $\{q_{tr}\}$;
2) optimize $\{q_{tr}\}$ for fixed $\{p_{tu}\}$.

Each of these two steps is realized by the inner loop of EM iterations. Each iteration consists of E-step and M-step. The E-step evaluates posterior probabilities that a user u uses an item r being interested in a topic t. This allows to decompose the log-likelihood (3) into a sum of weighted log-likelihoods corresponding to individual users or items and then to optimize them separately. The M-step performs this optimization purely analytically and very effectively. Technicalities will be discussed in the full version of the paper.

The main peculiarity of the algorithm is its symmetry with respect to the alternative representations (1) и (2). Both are used in EM iterations assuring that users and items latent profiles will be consistent.

Experiments and conclusions

The algorithm was tested on real data of Yandex search machine (www.yandex.ru). The raw dataset was a one week log file of clicks on documents returned by the search machine. Neither query strings nor document content were used in the experiments.

After the data preprocessing stage we retained 1024 most visited web sites as items and 7292 most active users. The latent profile size has been fixed as $|T|=12$. The meaning of topics has not been fixed a priory. Nevertheless the latent profiles estimated by the algorithm turned out to be very well interpretable. Each position of the profile has been attributed to a concrete topic corresponding to a sufficiently large subset of items with pronounced maximum in this position, see Table 1.

There are a lot of ways to define a distance function (metric) on users $\rho_U(u,u')$ and on items $\rho_R(r,r')$.

[1] RFBR grants 07-01-12076-офи and 08-07-00422.

Table 1. The profiles of 14 items separates into 3 well-interpretable topics.

Music												
www.mp3real.ru	0	0.01	0.86	0	0.02	0.04	0.01	0	0.03	0	0.01	0.01
mp3.musicfind.ru	0	0	0.96	0	0	0	0	0	0	0.02	0	0.01
akkordi.ru	0	0.01	0.85	0.02	0.03	0.02	0.01	0	0.01	0.02	0.01	0.03
www.muzzone.com	0.01	0	0.94	0	0	0	0.02	0	0	0.01	0	0.02
mp3forum.ru	0.01	0.01	0.85	0.02	0	0.01	0.04	0.01	0.01	0.03	0	0.01
Mobile												
mindmix.ru/mobile	0.01	0.83	0.02	0	0.01	0.01	0.04	0	0.01	0.05	0	0
www.sotoman.ru	0.01	0.78	0.01	0.02	0.04	0.01	0.04	0.02	0.01	0.03	0.01	0.02
www.mobyline.ru	0.02	0.74	0.02	0.01	0.02	0.01	0.03	0.03	0.07	0.02	0.02	0.01
www.eurotel.ru	0.01	0.87	0.04	0	0.01	0.01	0.01	0	0	0.01	0.02	0.03
www.sota1.ru	0.01	0.91	0.01	0.01	0.01	0	0.02	0	0	0.01	0.01	0
Games												
gameguru.ru	0.01	0.01	0	0.01	0.02	0.03	0.77	0.01	0.02	0.09	0.01	0.02
www.gameland.ru	0.08	0.01	0.02	0.02	0	0	0.73	0.05	0.02	0.05	0.01	0
www.ag.ru	0	0.02	0.04	0.01	0.01	0.02	0.84	0.01	0	0.01	0.01	0.04
www.neogame.ru	0.02	0.01	0	0	0.04	0.01	0.81	0.04	0.01	0.04	0.01	0.02

The most obvious is the Euclidean distance between latent profiles. Better results were obtained by robust Euclidean distance that put to zero all profile components except several highest values.

To give a visual check-up of metrics quality we used the multidimensional scaling (MDS) representing a finite set of points with a given pairwise distances as a two-dimensional scatter plot also called a *similarity map* [1,7,9]. Fig. 1 shows the representation of items (web sites). The similarity map of users can also be drawn although it is less interpretable.

The main result is that MDS groups web sites of similar subject matter into clusters. The sites belonging to the same cluster usually have the maximal profile component in the same position, as shown in Table 1.

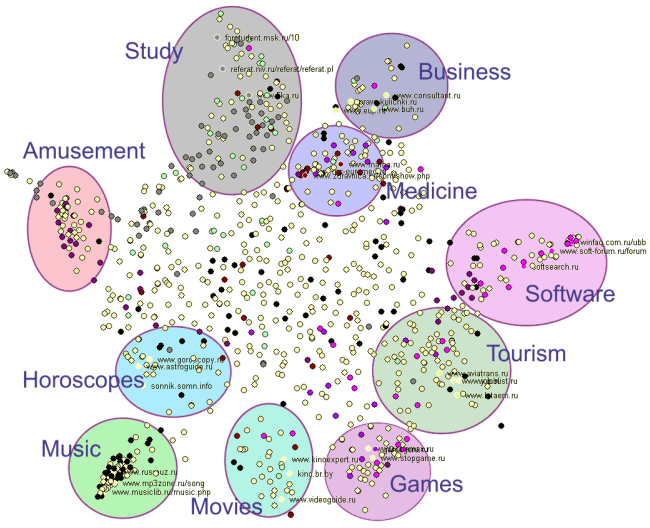

Fig. 1. Similarity map (the result of MDS).

To estimate the quality of profiles and metrics quantitatively we used a subsample of 400 web sites classified into 12 classes.

The profile quality criterion was defined as a number (in percents) of labeled sites such that the position of the maximum in their profile coincides with the most frequent one over the class. Fig. 2 shows the result of coordinate-wise optimization of three integer parameters of the algorithm. The value of the quality criterion is along Y axis. The number of inner (EM) loop iterations, the number of outer loop iterations, and the size of latent profiles $|T|$ are along X axis. The optimum was obtained for 8 outer iteration, 2 inner EM iterations and profile size 12, which is equal to the number of classes. The main conclusion is that the excessive optimization turns out to be redundant and leads to overfitting.

The metric quality criterion was defined as a number (in percents) of classification errors made by the k nearest neighbors (kNN) algorithm with optimal choice of k. Three metrics were compared: the robust Euclidean distance between profiles; the Pearson's correlation of visits [8]; the probability of co-occurrence of visits expressed through Fisher's exact test [9]. Results were 11%, 38%, and 25% of errors respectively. Fig. 3 shows the dependence of the quality criterion on k for these three types of metric. Thus we conclude that the robust Euclidean distance between profiles is a much more adequate distance measure if compared with standard techniques.

[1] RFBR grants 07-01-12076-офи and 08-07-00422.

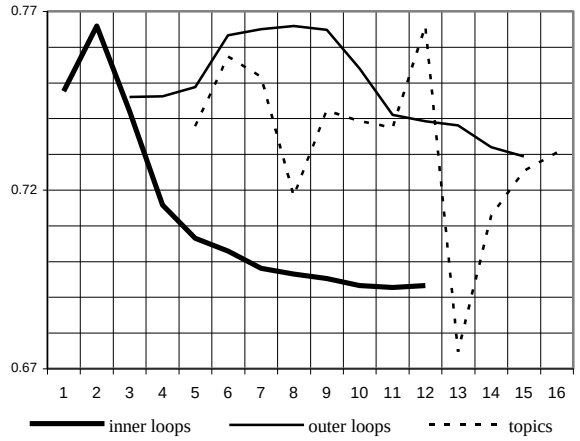

Fig. 2. The dependence of the number (in percents) of correctly reconstructed item profiles on three parameters of the algorithm (Yandex dataset).

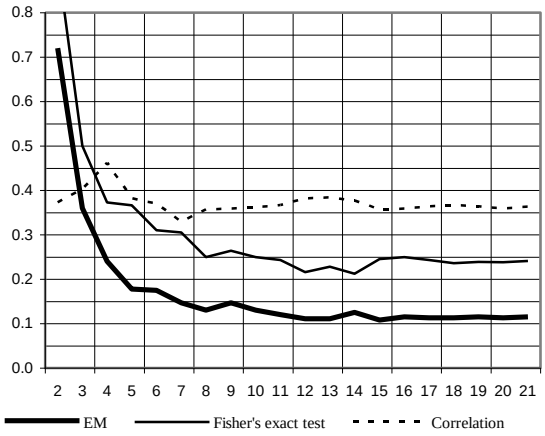

Fig. 3. The dependence of the number (in percents) of misclassified items on the parameter *k* in *k*NN algorithm for three types of metrics (Yandex dataset).

Also the algorithm was tested on artificial data generated by the probabilistic model (1). Data size was fixed at $|R|=500$, $|U|=1000$. The true profile of each user and each item was generated by putting 1/2 in two random positions, other positions was zero. The profile quality criterion was defined as a robust distance between the profile reconstructed by the algorithm and the true profile.

Parameters optimization gave again the optimal number of inner EM iterations 2 and the optimal number of outer iterations 6.

Also we investigated how the convergence of the algorithm depends on the data size l and the size of latent profiles $|T|$. The following result was obtained: if $|T|<10$ or $l<700$ then the algorithm diverges, that is the quality criterion worsen monotonically with the increase of the number of inner and/or outer iterations.

This work was supported by Russian Foundation of Basic Research, grants 07-01-12076-офи and 08-07-00422.

References

1. Customer Environment Analysis // Forecsys. — 2005. http://www.forecsys.ru/english/cea.php.
2. J. Fürnkranz. Web Mining // The Data Mining and Knowledge Discovery Handbook. 2005. P.899–920.
3. J. Breese, D Heckerman, C Kadie. Empirical analysis of predictive algorithms for collaborative filtering // 14th annual conference on Uncertainty in Artificial Intelligence. — 1998. — P. 43–52.
4. M. Grčar. User Profiling: Collaborative Filtering // SIKDD 2004 at multiconference IS 12-15 Oct 2004, Ljubljana, Slovenia.
5. T. Hofmann. Latent Semantic Models for Collaborative Filtering // ACM Transactions on Information Systems, Vol. 22, No. 1, 2004, Pp. 89–115.
6. X. Jin, Y. Zhou, B. Mobasher. Web Usage Mining Based on Probabilistic Latent Semantic Analysis // Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2004. — P. 197–205.
7. V. A. Leksin, K. V. Vorontsov. The client environment analysis: the reconstruction of latent profiles and similarity estimation of users and items // Russian conf. Mathematical Methods of Pattern Recognition (MMPO-13). — Moscow: MAKS Press, 2007. — Pp. 488–491 (in russian).
8. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews // In CSCW '94: Conference on Computer Supported Coorperative Work, Chapel Hill, ACM, P. 175–186.
9. K. V. Vorontsov, K. V. Rudakov, V. A. Leksin, A. N. Efimov // Web Usage Mining based on web users and web sites similarity measures // Artificial Intelligence. — Donetsk, 2006. — Pp. 285–288 (in russian).

[1] RFBR grants 07-01-12076-офи and 08-07-00422.