第33卷 第24期 计 算 机 工 程 2007年12月
Vol.33 No.24 Computer Engineering December 2007
·软件技术与数据库· 文章编号：1000—3428(2007)24—0057—03 文献标识码：A 中图分类号：TP311.1

基于概念层次树的个性化推荐算法

张晓敏，王 茜
(重庆大学计算机科学与技术学院，重庆 400030)

摘 要：改进了传统的协同过滤算法，提出了基于概念层次树的用户模型，利用该模型进行协同运算，使系统在用户共同评分项极其稀疏时也能产生推荐。在相似性计算和产生推荐阶段引入了概念分层思想，分别在商品种类上产生推荐，避免了推荐的单一现象。MovieLens数据集实验表明，改进后的算法在推荐质量上有了明显的提高。

关键词：个性化推荐；协同过滤；概念层次树

Personalized Recommendation Algorithm Based on Concept Hierarchy Tree

ZHANG Xiao-min, WANG Qian
(School of Computer Science & Technology, Chongqing University, Chongqing 400030)

【Abstract】 This paper improves traditional collaborative filtering algorithm, proposes a new user profile based on concept hierarchy tree, which can make recommender systems still work even when users have no common rating items. In the process of similarity calculation and recommendation formation, it also uses concept hierarchy thought to generate recommendation lists by different categories, avoiding recommendation lack of diversity. Experimental results on MovieLens dataset show that the improved algorithm can provide better prediction in either accuracy or diversity aspect.

【Key words】 personalized recommendation; collaborative filtering; concept hierarchy tree

在目前的电子商务中，产品信息呈指数级增长，个性化推荐技术应运而生，它通过分析消费者的历史交易记录，获取消费者的兴趣偏好，并推荐产品或服务，节省了消费者寻找合适商品的时间。在当前的个性化推荐系统中，协同过滤及其改进算法被大多数电子商务网站所采用。协同过滤算法主要分为：基于用户的协同过滤和基于项目的协同过滤。在个性化推荐的发展过程中，Ringo 和 GroupLens 是较早使用个性化推荐的一批系统，初期主要采用基于用户的协同过滤技术。最近 5 年内，基于项目的协同过滤系统得到了快速而广泛的发展。目前的个性化推荐系统将多种技术相结合，如将协同过滤与基于内容的技术相结合，多种技术间相互扬长避短，提高系统性能。

1 传统协同过滤中的问题及分析

传统协同过滤技术是指基于用户的协同过滤，其主要存在的问题如下：

(1)稀疏性问题

稀疏的用户评分，使用户间的相似性计算缺乏数据依据，难于找到准确的邻居，从而影响系统的推荐效果。传统的相似性计算基于用户对商品的评分，忽略了不同商品在商品种类上的关联性。

如图 1 所示，用户 A 评分的电影为$\{A_1, A_2, F_1, F_2\}$，用户 B 评分的电影为$\{A_3, M_1, F_4\}$。若按照传统方法计算，稀疏的评分使用户 A 和用户 B 的相似性为 0，他们没有对任何相同电影进行评分。但从分类关系分析，用户 A 和用户 B 应该具有一定相似性，因为他们都喜欢 Adventure 和 Fantasy 类电影。另外，虽然用户 B 喜好的 M_1 属于 Military 类，但是它和 Adventure 类同属于 Action 类，在传统计算中，原本是无用项甚至破坏项的 M_1，实际上对用户 A 和用户 B 的相似性是有一定贡献的，它使用户 A 和用户 B 在上层的 Action 类上具有了一定的相似性。

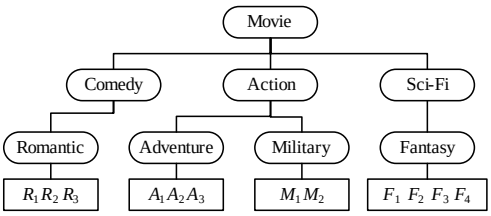

图 1 ebay 网部分电影分类

(2)多样性问题

用户的兴趣是多样而多变的，“accuracy does not tell the whole story”。目前的大多数推荐系统精确性良好，但缺乏推荐的多样性。例如，用户 A 喜好的电影中包含了 60%的 Romantic 类，20%的 Adventure 类和 20%的 Fantasy 类，那么通常情况下推荐结果会包含大量 Romantic 类电影，却很少甚至不出现其他两类推荐。

产生这个问题的主要原因为：传统协同过滤只关注用户对商品的评价，忽视了商品种类的区别。由于寻找的是对所有商品的评价与当前用户相似的用户，因此找到的多为主要偏好与当前用户相似的邻居，利用这些邻居产生推荐势必造成推荐项大多是与用户主要偏好类似的商品，推荐较为

基金项目：重庆市信息产业部基金资助项目(200502009)
作者简介：张晓敏(1981－)，女，硕士研究生，主研方向：个性化推荐，电子商务，数据挖掘；王 茜，副教授、博士
收稿日期：2006-12-25 **E-mail：**lizzia_zhang@yahoo.com.cn

单一。

2 基于概念层次树的协同过滤算法

在 Pazzani 所提出的“ collaboration via content ”[1]思想上，利用数据挖掘中常用的概念层次树[2]这一结构，本文提出了一种新的用户模型(user profile)，并对传统协同过滤算法进行了改进。在改进算法中，将用户对商品的评分转化为对每个商品种类的评分，建立起基于概念层次树的新用户模型；利用新的用户模型计算用户在不同商品种类上的相似性，寻找该商品种类上的最近邻居；综合各商品种类上的邻居推荐产生 top-N 推荐。

2.1 相关符号定义

$C=\{c_1,c_2,\cdots,c_n\}$：用户集。

$B=\{b_1,b_2,\cdots,b_m\}$：项目集，即商品集。

$R=\{R_1,R_2,\cdots,R_n\}$：用户评价集，$R_i\subseteq B$ 为根据隐式访问记录发现的用户 c_i 所喜好项目的集合。

$D=\{d_1,d_2,\cdots,d_l\}$：项目种类集。

T：根据领域分类知识由 D 中所有元素组成的概念层次树，T 中各结点互异。

$f:B\to 2^D$：项目种类分配描述，即把项目 $b_j\in B$ 分配到项目种类 $d_j\in D_k$ 下，$D_K\subseteq D$ 由 T 中的叶子结点组成，通常具体分配由网站或领域专家给出。

2.2 基于概念层次树的用户模型的生成

为了避免传统方法中使用“ 用户项目评分向量 ”表示用户模型的缺陷，本文使用“ 基于概念层次树的用户项目种类评分向量 ”，表示用户 c_i 的偏好模型为

$$\boldsymbol{v}_i=\left(v_{i1},v_{i2},\cdots,v_{i|D|}\right)$$

其中，维数为概念层次树中结点的总数 $|D|$。

v_{ik} 表示用户 c_i 对树中结点 $d_k\in D$ 的评价，以下介绍该用户模型的计算方法。

设总分 S 为用户 c_i 对各个项目种类的评分之和，即

$$\sum_{k=1}^{|D|}v_{ik}=S \tag{1}$$

将 S 平均分配给用户 c_i 喜好的项目，再将各个项目分得的分值平均分配给该项目所属的各个项目种类，由此得到用户 c_i 喜好项目 $b_j\in R_i$ 所属的项目种类 $d_{jk}\in f(b_j)$ 的初始分值 $t(d_{jk})$，即

$$t(d_{jk})=\frac{S}{|f(b_j)|\cdot|R_i|} \tag{2}$$

其中，$|R_i|$ 为用户 c_i 喜好的项目数目；$|f(b_j)|$ 为 b_j 所属的项目种类数目。

将 d_{jk} 分得的初始分值按一定规则分配给它的上层项目类别。用 $(p_0,p_1,\cdots,p_r)$ 表示从概念层次树的顶层结点 P_0 到底层结点 $p_r=d_{jk}$ 的路径，则该路径中各级结点 P_t 的分得的分值 $s(p_t)$ 为

$$\sum_{t=0}^{r}s(p_t)=t(d_{jk}) \tag{3}$$

$$s(p_t)=\frac{s(p_{t+1})}{b(p_{t+1})+1} \tag{4}$$

其中，$b(p_t)$ 表示结点 P_t 的兄弟结点个数。

将计算得到的 $s(p_t)$ 值加到 $\boldsymbol{v}_i$ 中 P_t 所对应的向量分量上。重复计算每个项目 $b_j\in R_i$ 的每个项目种类 $d_{jk}\in f(b_j)$ 所属路径上各结点的分值，由此建立用户 c_i 的偏好模型。

2.3 寻找最近邻居

为避免用户主要偏好“ 遮盖 ”次要偏好，使推荐单一的现象，受“ community-based recommendation ”思想[3]的启发，本文计算用户在每个喜好的项目大类上的相似性，寻找用户在各喜好的项目大类上的最近邻居。

2.3.1 划分项目种类子集

从概念层次树中的结点属性分析，根的各直接子树包含的项目种类的属性相对独立，相互间关联最小，如图 1 所示，根 Movie 下的 3 棵子树——Comedy 类电影、Action 类电影、Sci-Fi 类电影，它们各自包含的子类相互间差异都较大。因此，对项目种类集合作如下划分，即 $D=\{Root,D_1,D_2,\cdots,D_w\}$。*Root* 为 T 的根结点，$D_h(h=1,2,\cdots,w)$ 为 *Root* 的子树 T_h 中各结点(项目种类)组成的集合，w 为 *Root* 的子树棵数。

2.3.2 寻找喜好种类的邻居

本文需要判定用户喜好的项目种类子集。对于用户关注较少或从未关注过的项目种类，本算法将其视为用户“ 不感兴趣的种类 ”。根据实验经验值，若

$$\frac{\text{访问种类子集中的项目数}}{\text{访问项目总数}}<10\%$$

则视该项目种类子集为用户不喜好的项目种类，算法将不再在其上寻找邻居进行推荐，以节省系统的资源。

判定用户喜好的项目种类后，在各喜好的项目种类子集 D_h 上，依次计算用户 c_i 和 c_j 的在每个集合中的 Pearson 相关度，即

$$sim_h(c_i,c_j)=\frac{\sum_{k=1}^{|D_h|}(v_{ik}-\overline{v}_i)\cdot(v_{jk}-\overline{v}_j)}{\sqrt{\sum_{k=1}^{|D_h|}(v_{ik}-\overline{v}_i)^2\cdot\sum_{k=1}^{|D_h|}(v_{jk}-\overline{v}_j)^2}} \tag{5}$$

其中，v_{ik} 和 v_{jk} 为用户 c_i 和 c_j 对项目种类 $d_k\in D_h$ 的评分；$\overline{v}_i$ 和 $\overline{v}_j$ 为 c_i 和 c_j 对 D 中所有项目种类的平均评分，$\overline{v}_i=\overline{v}_j=S/|D|$。

根据计算出的 Pearson 相关度，选择与当前用户 c_i 最相似的 K 个用户作为 c_i 在项目种类子集合 D_h 上的邻居集 $neighbor_h(c_i)$。重复以上运算，找出当前用户 c_i 在所有喜好的项目种类子集上的邻居集。

2.4 产生推荐

2.4.1 评估候选项目

在喜好的项目种类子集 D_h 上，选取属于该集合中的项目种类、被当前用户 c_i 的邻居所喜好、且未被 c_i 访问过的项目，构成当前用户 c_i 的候选推荐项目集，即

$$CB_{ih}=\{b_k\mid\exists c_j\in neighbor_h(c_i):b_k\in R_j,b_k\notin R_i,f(b_k)\subseteq D_h\}$$

估算每个候选项目 $b_k\in CB_{ih}$ 受当前用户 c_i 关注的程度，用权重 $w_{ih}(b_k)$ 来表示。计算 $w_{ih}(b_k)$ 时，重点考虑的因素如下：

(1)喜好 b_k 的邻居 c_j 与当前用户 c_i 的相似程度 $sim_h(c_i,c_j)$。c_j 的偏好与 c_i 的越相似，c_j 的推荐可信度就越高。

(2)c_j 对 b_k 的喜好程度 $pref_h(c_j,b_k)$。c_j 对 b_k 越喜好，b_k 所获得的推荐权重就越高。

在计算 $pref_h(c_j,b_k)$ 时可以进行如下处理：假定一个虚拟用户 c_θ，$R_\theta=\{b_k\}$，则 $pref_h(c_j,b_k):=sim_h(c_j,c_\theta)$。

当 c_j 访问过较多与 b_k 同类的项目时，c_j 对 b_k 表现出较高的喜好程度。

根据以上因素，定义候选项目权重计算公式如下：

$$w_{ih}(b_k)=\frac{\sum_{c_j\in C_{ih}(b_k)}sim_h(c_i,c_j)\cdot pref_h(c_j,b_k)}{|C_{ih}(b_k)|} \tag{6}$$

其中，$C_{ih}(b_k)$ 为用户 c_i 的喜好 b_k 的邻居组成的集合；b_k 为其项目种类属于 D_h 的项目。

2.4.2　产生最终推荐

评估所有项目种类属于 D_h(用户喜好项目种类子集)的候选项目 $b_k \in CB_{ih}$ 后，按权重 $w_{ih}(b_k)$ 对 b_k 进行降序排列，得到 D_h 上的候选项目的推荐列表，即 $P_{ih}=\left\{b_1,b_2,\cdots,b_{|CB_{ih}|}\right\},1\le h\le w$。其中，$\forall b_j,b_k\in P_{ih}, j<k,\ w_{ih}(b_j)\ge w_{ih}(b_k)$。

根据当前用户 c_i 对不同项目种类的偏好，计算各喜好项目种类的侯选推荐项在最终推荐列表中所占比例，即

$$num_{ih}=N\cdot\frac{|R_{ih}|}{|R_i|} \tag{7}$$

其中，R_{ih} 为喜好种类子集 D_h 上 c_i 访问项目组成的集合；N 即为产生的 top-N 推荐的推荐项目数。

从各 P_{ih} 中抽取前 num_{ih} 个项目，将这些项目按 $w_{ih}(b_k)$ 值降序排列，形成对用户 c_i 的最终 top-N 推荐 $P_i=\{b_1,b_2,\cdots,b_N\}$。

3　实验评估

3.1　实验数据集

本文采用 MovieLens 站点提供的数据集对文中提出的算法与传统协同过滤算法进行比较。MovieLens 是明尼苏达州立大学计算机科学系 GroupLens 研究小组搜集的用于研究协同过滤算法的数据集，它包括 943 个用户对 1 682 部电影的 100 000 个评分(评分值为 1~5)记录，每个用户至少评价了 20 部电影，并且包含了简单的用户信息和电影分类信息。本文利用 ebay 网的电影分类结构和该数据集提供的电影分类描述构建概念层次树。

3.2　评估标准

3.2.1　精确性评估标准

本文使用 Sarwar 提出的查全率(recall)和查准率(precision)[4]作为算法在精确性方面的评估标准，即

$$Recall=100\cdot\frac{|P_i\cap TE_i|}{|TE_i|} \tag{8}$$

$$Precision=100\cdot\frac{|P_i\cap TE_i|}{|P_i|} \tag{9}$$

其中，P_i 为用户 c_i 的最终 top-N 推荐；TE_i 为用户 c_i 的测试集，它是用户喜好的项目集合的一个子集。

3.2.2　多样性评估标准

本文使用了 Ziegler 提出的 ILS(intra-list similarity)[5]作为推荐多样性的评估标准，即

$$ILS(P_i)=\frac{1}{2}\sum_{b_f\in P_i}\sum_{b_e\in P_i,b_f\ne b_e}g(b_f,b_e) \tag{10}$$

其中，b_f 和 b_e 为推荐列表 P_i 中的推荐项目；$g(b_f,b_e)$ 在本文中定义为 b_f 和 b_e 的语义贴近度，即

$$g(b_f,b_e)=\begin{cases}1, b_f=b_e\\ 1-l\left(p(b_f,b_e)\right)\big/H, b_f\ne b_e\end{cases} \tag{11}$$

其中，$p(b_f,b_e)$ 返回 b_f 和 b_e 对应叶结点种类 d_f 和 d_e 在树中的公共祖先结点；$l(d)$ 返回 d 结点所在的层数(叶结点层数为 0，每向上一级层数加 1)；H 为概念层次树的树高。

$ILS(P_i)$ 的值越小，表明推荐列表 p_i 中的项目的种类相似性越小，推荐的多样性越好。

3.3　实验方案与结果

提取用户评分≥4 的电影作为用户喜好的电影。根据用户喜好电影的数目，将用户集合划分为 10 个子集，每个子集中用户喜好电影的数目 $\le 20\cdot x$ (x=1,2,⋯,10)。对每个用户子集中的用户，将其用户的喜好项目集合等分成 5 份，随即抽取 4 份形成训练集 TR_i 用以产生推荐，剩余 1 份形成测试集 TE_i 用以检测推荐质量，随机抽取 5 次，由此在每个用户子集中形成 5 组训练集和测试集，产生 5 次 top -10 推荐。

本文对以下 3 种算法进行比较：(1)传统协同过滤算法(user-based CF)；(2)改进的协同过滤算法(CF-1)，使用基于概念层次树的用户模型，但仍沿用传统的相似性计算和推荐方法，未将推荐多样性的改进考虑在内；(3)本文的改进算法(CF-2)。通过在 10 个用户数据集上进行实验，3 种算法的性能比较见图 2~图 4。

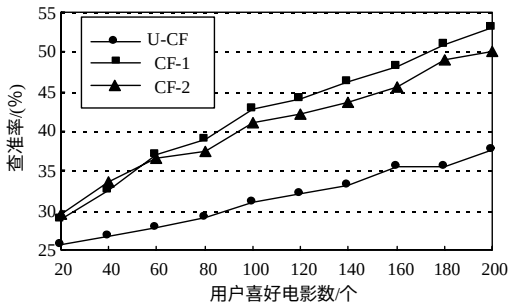

图 2　算法精确度-查准率比较

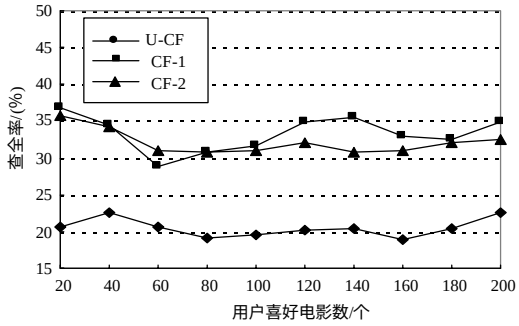

图 3　算法精确度-查全率的比较

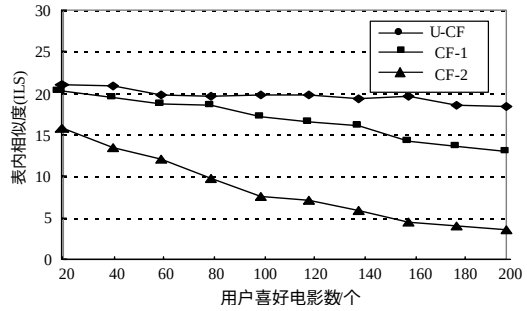

图 4　算法多样性比较

实验结果表明，使用新用户模型的改进算法 CF-1 和 CF-2 在推荐的准确性方面都明显优于传统的协同过滤算法，即使在评价数据较稀疏时这种优势也非常明显。在推荐的多样性方面，本算法也比传统的 CF 和部分改进的 CF 有明显的改善，尤其随着用户评价数据的密集，用户喜好电影的种类相对增多，推荐的多样性表现得更加明显。但随着用户评价项目的增多，不论是查准率还是查全率，CF-2 都逐渐略劣于 CF-1，可见，推荐的多样性对推荐的准确性产生一定的负面影响。

4　结束语

在评价数据稀疏的情况下，本文分析了传统协同过滤算法中用户模型和用户相似性计算存在的问题，提出了基于概念层次树的用户模型，基于该模型所改进的寻找邻居和产生推荐的新方法。实验结果表明，改进后的算法在推荐准确性

(下转第 62 页)

(9)计算 $S+S''$ 中所有个体的适应度，并淘汰掉适应度小的 M 个个体，形成新一代群体 S。

(10)终止操作：如果新一代个体的最大的适应度与上一代个体的最大适应度的差值小于 ε (ε =0.005)，则解码。否则转到步骤(5)。

(11)对解码后的二维坐标值应用 FCM 算法，并将得到的聚类结果对应回原始的高维样本中。

3.4　算法可行性分析

根据 3.1 节、3.2 节可知，$r_{ij}\in[0,1]$，$r_{ij}^{'}\in[0,1]$。若 $r_{ij}\approx 0$，即高维样本 i 与样本 j 的非相似性几乎为 0，说明样本 i 与样本 j 为一类。又 $r_{ij}^{'}$ 趋近于 r_{ij}，所以，$r_{ij}^{'}\approx 0$，即这两个高维样本映射到二维平面上的二维样本间的欧氏距离几乎为 0，根据类内距离小,类间距离大可得该二维样本经 FCM 聚类后应为一类。

同理，任何两个高维样本，若它们的模糊非相似性越大，那么它们对应的二维样本间的欧氏距离越大。而欧氏距离越大，则相似性越小，即二维样本之间的非相似性越大，这样就将高维样本间的差异程度转化为二维样本间的差异程度，因此，对映射后的二维样本聚类就相当于对原始的高维样本聚类，具有可行性。

4　实验仿真

实验选取了一部分 IRIS 数据作为样本，样本总数为 21，样本属性为 4，聚类类别为 3，其中，每类包括的样本数都为 7。

利用本文提出的方法对表 5 中的数据进行聚类时取种群规模 N=100，迭代次数 G=60，变异概率 Pm=0.5，交叉概率 Pc=0.2，则聚类结果如图 2 所示。

表 5　部分 IRIS 数据

数据编号	属性 1	属性 2	属性 3	属性 4
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
…	…	…	…	…
21	4.9	2.5	4.5	1.7

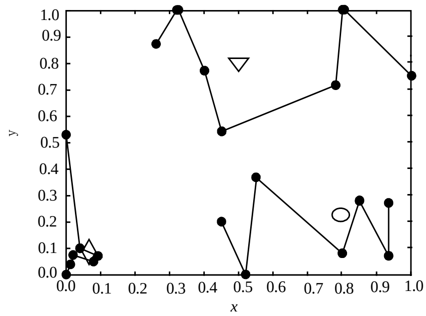

图 2　基于遗传算法的模糊聚类

实验结果表明，该方法有较好的聚类效果，即证明了将高维降为二维聚类的有效性。

5　结束语

本文提出了一种基于分治法的高维大数据集模糊聚类算法。以二维样本为例进行了实验，结果表明利用分治法在大多数情况下与一次性聚类结果一致，并且通过分析可得该方法能极大的提高聚类的速度。对于高维样本通过实验证明了将高维降为二维去聚类的有效性，并分析了可行性，因此，本文提出的方法适合对高维大数据集聚类，能提高聚类的效率，且具有有效性。

参考文献

[1] Zhang Yuanquan, Rueda L. A Geometric Framework to Visualize Fuzzy-clustered Data[C]//Proceedings of IEEE 25th International Conference of the Chilean Computer Science Society. Valdivia, Chile: [s. n.], 2005: 8-13.

[2] Davidson I, Satyanarayana A. Speeding Up *K*-means Clustering by Bootstrap Averaging[C]//Proc. of IEEE Data Mining Workshop on Clustering Large Data Sets. Brighton, UK: [s. n.], 2004: 98-102.

[3] Aggarwal C, Yu P. Finding Generalized Projected Clusters in Dimensional Spaces[C]//Proc. of ACM SIGMOD Conference on Management Data. Dallas, Texas, U.S.A: [s. n.], 2000: 78-52.

[4] Tsai Chengfa, Tsai Chunwei, Chen Chiping. A Novel Multiple Searching Genetic Algorithm for Multimedia Multicast Routing[C]//Proc. of IEEE Congress on Evolutionary Computation. Piscataway, NJ: [s. n.], 2002: 7065-7068.

~~~~~~

(上接第 59 页)

上比传统算法有了明显提高，同时在推荐的多样性上也有了明显的改善。下一步的工作为：将该算法部署到实际的推荐系统中，通过在线测试获得用户对推荐准确性和多样性的满意度的反馈，进一步改进算法。

参考文献

[1] Pazzani M. A Framework for Collaborative, Content-based and Demographic Filtering[J]. Artificial Intelligence Review, 1999, 13(5): 393-408.

[2] 王丽珍. 一种基于语义贴近度的抽象归纳法[J]. 计算机学报, 2000, 23(10): 1114-1121.

[3] Kamahara J, Asakawa T, Shimojo S, et al. A Community-based Recommendation System to Reveal Unexpected Interests[C]//Proc. of the 11th International Multimedia Modeling Conference. Tokyo, Japan: [s. n.], 2005: 433-438.

[4] Sarwar B, Karypis G, Konstan J, et al. Analysis of Recommender Algorithms for E-commerce[C]//Proc. of the 2nd ACM E-commerce Conference. Minneapolis, America: Minnesota Press, 2000: 135-141.

[5] Ziegler C, Mcnee S, Konstan J, et al. Improving Recommendation Lists Through Topic Diversification[C]//Proc. of International World Wide Web Conference. Chiba, Japan: [s. n.], 2005: 22-32.
~~~~~~