改进的聚类邻居协同过滤推荐算法

何光辉[1] 魏曙光[1] 王蔚韬[2]

(重庆大学数理学院 重庆400044)[1] (重庆大学计算机学院 重庆400044)[2]

摘 要 推荐系统运用统计和知识发现技术在实时交互系统中提供产品推荐,并且已经在电子商务中取得了较广泛的应用。本文中我们介绍了一种不同于以往的推荐产生算法,称之为改进的聚类邻居协同过滤推荐算法,试验表明我们的算法比k-邻近点算法和聚类邻居算法具有更好的效果。

关键词 推荐系统,协同过滤,聚类邻居算法,电子商务

An Adaptive Algorithm of Collaborative Filtering Recommender Based on Aggregate-Neighborhood

HE Guang-Hui[1] WEI Shu-Guang[1] WANG Wei-Tao[2]

(College of Science, Chongqing University, Chongqing 400044)[1]

(College of Computer Science, Chongqing University, Chongqing 400044)[2]

Abstract Recommender systems apply statistical and knowledge discovery techniques to the problem of making product recommendations during a live interaction and they are achieving widespread success in E-commerce nowadays. In this paper, we introduce a different recommendation generation algorithm, we name it adaptive-aggregate-neighborhood collaborative filtering algorithm. Our experiments suggest that we prove that our algorithm has better performance than k-nearest neighbor and aggregate-neighborhood approach.

Keywords Recommender system, Collaborative filtering(CF), Aggregate-neighborhood algorithm, E-commence

1 引言

电子商务网站中的推荐系统从顾客处获取信息,获取顾客的兴趣、爱好并推荐相关的商品给顾客。由于推荐系统应用潜力非常大,因此针对推荐系统的研究也比较多。协同过滤算法CF(collaborative filtering)是最早最成功的推荐技术之一。目前,CF算法在实践和研究中都比较成功。然而,随着应用的深入和广泛,也暴露了一些缺点,最主要的就是以下两点:

(1) CF算法的伸缩性(scalability)

CF类算法能够实时(live)查找上万个的"邻居(与目标顾客具有类似爱好、背景的顾客)"。但对庞大的网上顾客群体来说,查找邻居顾客的速度还有待提高,进而提高CF算法的伸缩性。

(2) CF算法的推荐质量

推荐系统与其他查询系统一样易犯两种错误:弃真(错误否定),即虽然顾客喜爱物品 A 但推荐系统没有推荐;纳伪(错误肯定),即顾客不喜爱物品 A 推荐系统却推荐了。在电子商务中,我们应当尽量避免纳伪这种错误。因为,纳伪错误会引起顾客的不满情绪,从而破坏推荐系统的权威性。

随着电子商务系统规模的不断扩大,顾客数量和商品的数目急剧增长,导致计算得到的目标用户的最近邻居不准确,进而降低系统的推荐质量。本文主要从减少推荐系统的计算量角度提出了一种新的顾客-商品数据库的构造方法,进而给出了一种新的推荐算法,我们称之为改进的聚类邻居算法。实验结果和理论研究都表明本方法能有效地降低计算量,提高系统的推荐质量。

2 相关工作

随着电子商务的出现,个性化服务成为商家研究的重点。商家提倡建立"一对一"的市场销售模式。通过有关技术帮助商家将每个顾客单独看待,在互联网市场领域提高竞争力。因此,推荐系统也就应运而生。最早出现的是为团体顾客服务的Tapestry[5]推荐工具,后来出现了以评价为基础的自动推荐系统(ACF)。GroupLens研究组提出了一种匿名的CF方法用于Usenet新闻和电影。Kingo和Video是电子邮件和基于网页分别为音乐和电影而设计的推荐系统。目前提出的推荐算法有CF推荐系统、基于项目的推荐系统、Bayesian网络技术、聚类技术、关联规则技术及Horting的图技术。尽管这些应用这些算法的系统已经比较成熟,但其广泛应用却暴露了一些算法的缺点,如稀疏数据集的问题、高维度相关的问题等。稀疏问题在文[7,8]中已经讨论过,而有关降维技术在文[9]中已有所探讨。这里我们不再赘述。

3 基于CF算法的推荐系统

推荐系统主要应用于网页交互环境。对某一顾客,通过数据分析技术产生Top-N推荐产品,帮助顾客在网站中发现他们想要购买的产品。推荐的产生过程可以分为三个步骤:描述、邻居关系构造、产生推荐,如图1所示。

描述的任务是将顾客已经购买的产品用图表表示出来;邻居关系构造着重解决怎样确定邻居顾客;产生推荐主要是从邻居关系中寻找Top-N个推荐产品。下面我们将分别讨论。

3.1 描述

与以往典型的CF的推荐系统相同,我们输入的数据是 n 个顾客购买 m 种商品的的历史数据,但我们用一个 $m\times(n+1)$ 阶的顾客-商品矩阵 R_1 描述,其中元素 $r_{i,j}$ 是指第 i 个顾客购买第 $j(j>0)$ 种产品的数量,否则为0;当 $j=0$ 时,代表的是第 i 顾客所过购买的商品种类之和,如表1所示。

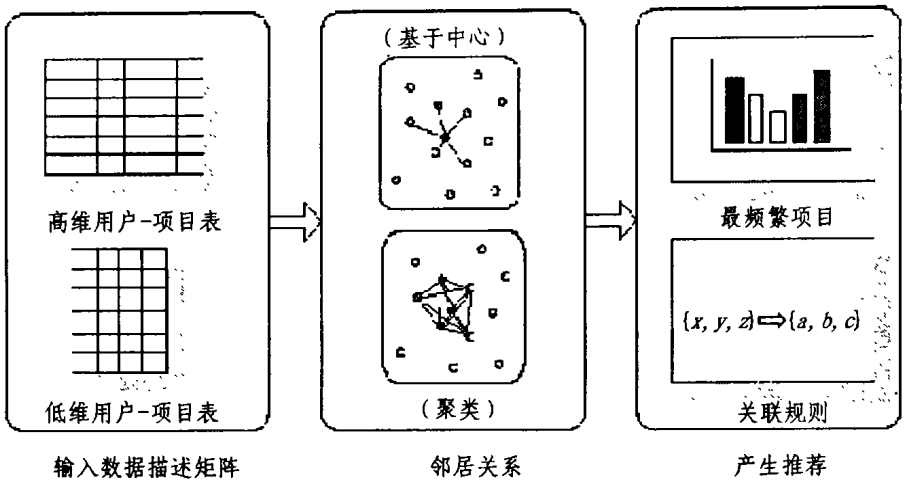

图1　推荐系统的三个主要步骤[10]

表1　初始描述矩阵 R_1

	商品种类和	商品1	商品2	商品3	商品4	商品5
顾客1	4	5	2	4	1	0
顾客2	3	0	7	3	0	1
顾客3	2	3	6	0	0	0

与以往描述不同的原因我们下文将进行解释。我们把这个输入的数据集 $m\times(n+1)$ 阶矩阵称为初始描述。

3.2　邻居关系构造

基于 CF 算法的推荐系统的最重要步骤就是计算顾客间的类似性，因为类似性可以确定目标顾客与具有类似爱好、背景的顾客之间的邻居关系。邻居关系的构造过程实际上是建立的模型或推荐系统的学习过程。邻居关系构造的主要目标是对每个顾客 u 寻找一个有序的 l 个顾客的序列，$N=\{N_1, N_2,\cdots,N_l\}$，使得 $u\notin N$，并有 $sim(u,N_1)>sim(u,N_2)>\cdots>sim(u,N_l)$。两个顾客之间的相似度我们用余弦公式(1)计算。

$$sim(\vec{a},\vec{b})=\cos(\vec{a},\vec{b})\frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|*\|\vec{b}\|} \tag{1}$$

应用相似度计算公式(1)，对 n 个顾客我们可以得到一个 $n\times n$ 的相似性矩阵 S_1。之所以称为 S_1 是与以往的相似性矩阵相区别。对构造相似性矩阵 S_1 我们作如下说明：

因为我们构造的初始描述 R_1 中的第一列已经给出了所有顾客以往所购买的商品种类和 $Q_i(1\leqslant i\leqslant m)$，因此我们由定理1(见第4部分)的结论可知，对目标顾客的候选邻居集选择范围可以进行缩小。设目标顾客 A 已购商品种类和为 Q_a，则我们候选邻居集的选择条件可以设为 $Q_a-K\leqslant Q\leqslant Q_a-K,(K\in Z^+)$，其中 K 为常数。一般情况下我们可以将 K 值取得稍微大一点，如 $K=30,40$，或者50，当然推荐系统用户也可以根据实际情况确定。之所以能这样选择的原因见第4部分的证明。这样既可以大大缩小候选邻居的范围，同时又能够使推荐的质量得到保证。

下面我们利用相似性矩阵 S_1 形成邻居关系，采用聚类邻居方法构造邻居关系，具体过程如下：

对给定顾客 C，形成长度为 l 的邻居关系，起初我们选择距离 C 最近的邻居，然后选择剩下的 $(l-1)$ 个邻居，步骤是：设在邻居关系 N 中，对某一固定点处有 j 个邻居 $(j<l)$，计算出邻居关系的中心(centroid)，N 的中心 $\bar{C}$ 定义如下：$\bar{C}=\frac{1}{j}\sum_{\forall\in N}\forall$，顾客 $W(W\notin N)$ 被选为第 $(j+1)$ 个邻居，当且仅当 W 与中心 $\bar{C}$ 最近。然后，对 $(j+1)$ 个邻居重新计算中心，直到 $|N|=l$。

3.3　产生推荐

邻居关系构造工作完成后，下一步工作就是从邻居关系中找出 Top-N 个推荐，我们采用最频繁项目推荐法。其方法是从邻居关系 N 中对每个邻居进行扫描，从其购买的商品数据中形成商品的频繁度。在所有的邻居计算后，系统按照降序排列商品频繁度，并返回最频繁的商品作为顾客仍未购买的推荐商品。由于在实际实验中，少量的邻居产生的频繁项可能效果不好，因此我们可以适当的增加邻居的数量 l。

4　相关理论证明

引理　对任意向量 $\vec{a}=\{a_1,a_2,\cdots,a_n\}$，$(a_i>0,1\leqslant i\leqslant n)$ 有以下不等式成立：

$\sum_{i=1}^{n}a_i^2\geqslant n(\bar{a})^2$，其中 $\bar{a}=\frac{\sum_{i=1}^{n}a_i}{n}$。

证明：我们用归纳法进行证明：

当 $n=1$ 时，有 $a_1^2=a_1^2$，命题显然成立；

假设当 $n=k$ 时命题成立，则有 $\sum_{i=1}^{k}a_i^2\geqslant k(\bar{a})^2$，其中 $\bar{a}=\frac{\sum_{i=1}^{k}a_i}{k}$，即有不等式 $a_1^2+a_2^2+\cdots+a_k^2\geqslant\frac{(a_1+a_2+\cdots+a_k)^2}{k}$ 成立，亦即 $k(a_1^2+a_2^2+\cdots+a_k^2)\geqslant(a_1+a_2+\cdots+a_k)^2$ 当 $n=k+1$ 时，我们考虑 $(k+1)(a_1^2+a_2^2+\cdots+a_k^2+a_{k+1}^2)-(a_1+a_2+\cdots+a_k+a_{k+1})^2=(k+1)(a_1^2+a_2^2+\cdots+a_k^2)+(k+1)a_{k+1}^2-[(a_1+a_2+\cdots+a_k)^2+2a_{k+1}(a_1+a_2+\cdots+a_k)+a_{k+1}^2]\geqslant a_1^2+a_2^2+\cdots+a_k^2+ka_{k+1}^2-2a_{k+1}(a_1+a_2+\cdots+a_k)=(a_1-a_{k+1})^2+(a_2-a_{k+1})^2+\cdots+(a_k-a_{k+1})^2\geqslant 0$，即有 $(k+1)(a_1^2+a_2^2+\cdots+a_k^2+a_{k+1}^2)\geqslant(a_1+a_2+\cdots+a_k+a_{k+1})^2$，亦即 $\sum_{i=1}^{k+1}a_i^2\geqslant(k+1)(\bar{a})^2$，所以当 $n=k+1$ 时命题也成立。

综上，引理得证。

定理1　对任意两个顾客 A,B，如果其所购商品种类和 Q_a,Q_b 之间满足 $Q_a-Q_b=\pm K$，则当 K 充分大时，顾客 A 与顾客 B 之间的相似度 $sim(\vec{a},\vec{b})$ 趋于零，其中 $sim(\vec{a},\vec{b})=\cos(\vec{a},\vec{b})=\frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|*\|\vec{b}\|}$。

证明：设顾客 A,B 的购物分别为记录 $\vec{a}=\{a_1,a_2,\cdots,a_N\}$，$\vec{b}=\{b_1,b_2,\cdots,b_N\}$，顾客 A,B 的购物种类分别为 Q_a,Q_b，则 $Q_a-Q_b=K$。根据记录 $\vec{a},\vec{b}$ 的非零位置，可以分为以下两种

情况：

1. 记录 $\vec{a},\vec{b}$ 的非零位置比较分散

由常识我们知道，顾客只有在比较多的商品上具有相似的认同(表现为都购买了该商品)，我们才能确定用户间的相似度较高。如果在数据比较稀疏的情况下，两个顾客共同购买的商品种类就很少，此时 $sim(\vec{a},\vec{b})=\frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|*\|\vec{b}\|}$ 中的分子 $\vec{a}\cdot\vec{b}$ 显然比较小，所以两个顾客之间的相似度近似为零。

2. 记录 $\vec{a},\vec{b}$ 的非零位置比较集中

当两个顾客所购买的商品分布比较集中时，我们不妨考虑如下形式的两个向量：$\vec{a}=\{a_1,a_2,\cdots,a_{Q_a},0,\cdots,0\}$，$\vec{b}=\{b_1,b_2,\cdots,b_{Q_b},0,\cdots,0\}$，其中 $Q_a=Q_b+K(K>0,K\in N)$，则相似度 $=sim(\vec{a},\vec{b})=\cos(\vec{a},\vec{b})=\frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|*\|\vec{b}\|}=\frac{\sum_{i=1}^{Q_b}a_ib_i}{\sqrt{\sum_{i=1}^{Q_a}a_i^2}\sqrt{\sum_{j=1}^{Q_b}b_j^2}}$，设 $\bar{a}=\frac{\sum_{i=1}^{Q_a}a_i}{Q_a}$，$\bar{b}=\frac{\sum_{j=1}^{Q_b}a_j}{Q_b}$，$M_a=\max\{a_1,\cdots,a_{Q_a}\}$，$M_b=\max\{b_1,\cdots,b_{Q_b}\}$，由引理得相似度 $sim(\vec{a},\vec{b})=\frac{\sum_{i=1}^{Q_b}a_ib_i}{\sqrt{\sum_{i=1}^{Q_a}a_i^2}\sqrt{\sum_{j=1}^{Q_b}b_j^2}}\leqslant\frac{\sum_{i=1}^{Q_b}a_ib_i}{\sqrt{Q_a}\bar{a}\times\sqrt{Q_b}\bar{b}}\leqslant\frac{Q_bM_aM_b}{\sqrt{Q_aQ_b}\times\bar{a}\ \bar{b}}=\sqrt{\frac{Q_b}{Q_a}}\times\frac{M_aM_b}{\bar{a}\bar{b}}=\sqrt{\frac{Q_a-K}{Q_a}}\times\frac{M_aM_b}{\bar{a}\bar{b}}=\sqrt{1-\frac{K}{Q_a}}\times\frac{M_aM_b}{\bar{a}\ \bar{b}}$，即有 $sim(\vec{a},\vec{b})=\sqrt{1-\frac{K}{Q_a}}\times\frac{M_aM_b}{\bar{a}\bar{b}}$ (2)

因为对确定的两个顾客，$M_a,M_b,\bar{a},\bar{b}$ 都是确定(即为常数)的，所以相关性与 K 成反比，在(2)式中当 K 充分大时，顾客 A,B 的相似性 $sim(\vec{a},\vec{b})$ 会接近于零。同理可以证明 $Q_a=Q_b-K$ 的情形。

综上，定理得证。

5 实验结果及其分析

5.1 实验数据集

我们采用实验的数据集来自 MovieLens 站点(http://movielens.umn.edu/)，MovieLens 是基于网页的研究型推荐系统。目前，该网站的用户(顾客)已经超过了70000人，提供的电影(提供的商品)超过5000部。我们从顾客-商品数据库中抽取8000条记录，实验数据集中包含200个用户(顾客)和1000部电影(商品)。根据实验需要我们将数据集划分为训练集和测试集。我们此处规定训练集占数据集的80%，测试集占20%。

5.2 度量标准

由于统计精度度量方法中平均绝对误差 MAE(Mean Absolute Error)便于理解，而且可以直观地对推荐质量进行度量，因此该方法是我们最常用的推荐质量评价方法。本文采用 MAE 作为度量标准。MAE 通过计算预测的顾客所购商品与顾客实际的所购商品之间的偏差来度量预测的准确性。MAE 值越小，推荐质量越高。MAE 的计算公式如下：

$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N}$，其中 $\{p_1,p_2,\cdots,p_N\}$ 代表预测的顾客所购商品，而 $\{q_1,q_2,\cdots,q_N\}$ 代表实际的顾客所购商品。

5.3 实验结果及分析

为了验证本文提出的算法的有效性，我们将改进的聚类邻居算法与传统的协同过滤(CF)算法进行了比较。通过计算不同的邻居数量时的 MAE 值，我们可以看出改进的聚类邻居算法能有效地提高系统的推荐质量，如图2。

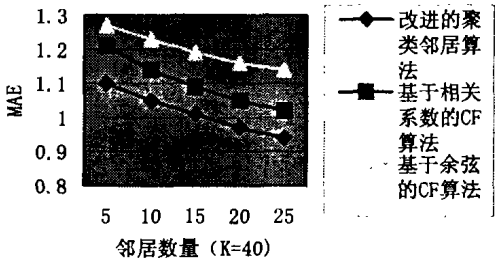

图2 算法推荐精度比较

我们提出的改进的聚类邻居算法在运行时所花费的时间相对于其他基于 CF 的推荐算法来说是非常短的，因为我们采用了有选择的候选邻居集方法，从而大大加快了计算速度。同时注意到我们构造的顾客-商品数据库的第一列是随着顾客购买商品种类而变化的，但不可能是所有的顾客同时在网上购物，所以我们只需要实时更新一小部分的顾客记录。这样我们就可以避免大规模的实时运算，从而提高推荐系统的运行速度。

总结 本文针对以往的推荐系统的算法存在的实时计算速度较慢的缺陷，提出一种新的构造顾客-商品矩阵方法，进而给出了一种改进的聚类邻居算法。通过构造一种新的顾客-商品矩阵，从而大大降低了寻找候选邻居集的数量。通过本文以上的分析和研究表明，我们提出的改进的聚类邻居算法能够有效地降低推荐系统计算量，从而提高系统的伸缩性，在应用时可以采取增加邻居候选集数量的方法来提高推荐质量。因为我们构造的顾客-商品数据库的优越性，即使适当增加候选邻居集的数量也不会影响推荐系统的灵活性。以后我们将继续研究推荐系统，以进一步提高推荐系统的性能和质量。

参考文献

1 Billsus D, Pazzani M J. Learning Collaborative Information Filters. In: Proc. of ICML'98. 1998. 46～53

2 Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence, 1998. 43～52

3 Goldberg D, Nichols D, Oki B M, Terry D. Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM. Dec. 1992

4 Herlocker J, Pei J, Yin Y. Mining Frequent Patterns Without Candidate Generation: [Technical Report CMPT99-12]. School of Computing Science, Simon Fraser University, 1999

5 Herlocker J, Konstan J, Borchers A, Riedl J. An Algorithmic Framework for Performing Collaborative Filtering. In: Proc. of ACM SIGIR'99, ACM press, 1999

6 Konston J, et al. GroupLens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, 1997, 40(3): 77～87

7 Peppers D, Rogers M. The One to One Future: Building Relationships One customer at a Time. Bantam Doubleday Dell Publishing, 1997

8 Resnick P, Iacovou N, Suchak M, Bergstorm P, Riedl J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In: Proc. of CSCW'94, Chapel Hill, NC, 1994

9 Wolf J, Aggarwal C, Wu K-L, Yu P. Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering. In: Proc. of ACM SIGMOD Intl. Conf. on Knowledge Discovery & Data

10 Badrul S, Goerge K, Joseph K, John R. Analysis of Recommendation Algorithms for E-Commerce EC'00, Minneapolis Minnesota, 2000