文章编号:1671 - 6833(2012)05 - 0118 - 03

基于加权信息熵相似性的协同过滤算法

刘文龙[1],张桂芸[1],陈 喆[2],朱蔷蔷[1]

(1. 天津师范大学 计算机与信息工程学院,天津 300387; 2. 天津师范大学 城市与环境科学学院,天津 300387)

摘 要:协同过滤算法是推荐系统中最为成功的技术之一,相似性计算是协同过滤算法的核心.针对传统的相似度计算方法在数据稀疏的情况下推荐不准确问题,提出了基于项目间差异信息熵的相似度计算方法,先通过差异值和共同评价数目对信息熵进行加权,再归一化处理来计算项目间的相似度.用基于项目(Item - based)相似性的协同过滤算法进行了实验验证,实验结果表明,该算法提高了个性化推荐精度.

关键词:信息熵加权;相似度计算;协同过滤;个性化推荐

中图分类号:TP391 **文献标志码**:A doi:10.3969/j.issn.1671 - 6833.2012.05.026

0 引言

互联网技术的快速发展使我们进入了信息爆炸的时代[1],用户需要处理大量毫无意义的信息和垃圾数据.个性化推荐系统是一种解决信息过载问题的工具,而协同过滤技术是推荐系统中最为成功的技术之一,尤其是在电子商务领域里的应用[2].它是基于这样一种假设:兴趣爱好相似的用户对相同项目的评价相似.实现协同过滤技术时,依据所建立模型的种类,可以分为基于用户的协同过滤和基于项目的协同过滤[3].由于在实际应用中,项目数量更加稳定,并往往远低于用户数量,因此,基于项目的协同过滤方法更为常用[4].它的大体步骤如下:①收集项目信息,如用户的浏览购买和评价记录;②根据收集的信息计算项目的 K 邻近集合;③通过 K 邻近集合进行分析计算产生对目标用户的推荐.作者选择基于项目的协同过滤算法对实验结果进行分析验证.

由上面介绍的协同过滤技术步骤可以看出,相似性计算是协同过滤技术的核心.传统的相似度计算方法有余弦相似性(Cosine)[5]、Pearson 相似相关系数[5]、修正的余弦相似性[5]、Spearman 相似性.其中,Pearson 相似相关系数是最为常用的相似度计算方法,Pearson 相关系数用于衡量两个向量之间的线性关系.设项目 i 和项目 j 共同评分的用户集合为 U_{ij},利用 Pearson 相关系数得到两者相似性为 $Sim(i,j)$

$$Sim(i,j)=\frac{\sum_{u\in Uij}(R_{u,i}-\overline{R_i})(R_{u,j}-\overline{R_j})}{\sqrt{\sum_{u\in Uij}(R_{u,i}-\overline{R_i})^2}\sqrt{\sum_{u\in Uij}(R_{u,j}-\overline{R_j})^2}},\tag{1}$$

式中:$R_{u,i}$,$R_{u,j}$ 分别为用户 u 对项目 i 和 j 的评分;$\overline{R_i}$ 和 $\overline{R_j}$ 分别表示项目 i 和 j 的平均得分.

1 基于加权信息熵的相似度计算方法 NNWD

1.1 算法的提出

传统的相似度计算方法在协同过滤技术中存在一定弊端,如:①在数据高维稀疏的情况下,用户之间关注圈交集(共同评分项目)的规模大多偏小且不一致,传统的相似性度量方法容易过分地夸大或者缩小用户间的真实相似性[6];②受数据稀疏等影响,推荐精度较低[6];③Pearson 相关系数必须满足数据之间的线性关系以及残差相互独立且均值为 0 等假设[6].当这些条件不满足

收稿日期:2012 - 04 - 15;**修订日期**:2012 -

基金项目:国家自然科学基金资助项目(60970060);天津市教委资助项目(20071328);天津市科技支撑计划重点项目(09ZCKFGX00500);天津师大博士基金项目(52LX17)

作者简介:张桂芸(1965—),女,天津蓟县人,教授,博士后,硕士生导师,主要从事人工智能和数据挖掘研究,E-mail:dyxy1999@126.com

时，其计算准确度将会降低.

例如对于项目 I_1 和 I_2，首先找出 I_1 和 I_2 共同评分的用户评分，$I_1(2,1,2,1)$ 和 $I_2(5,4,5,4)$，用 Pearson 相关系数计算 I_1 与 I_2 的相似性 $\mathrm{Sim}(I_1,I_2)=1$，完全正相关，相似度最高，而实际上 I_1 的评分普遍偏低，I_2 的评分普遍偏高，他们的相似度没有那么高. 对于 $I_3(4,5,4,5)$ 和 $I_2(5,4,5,4)$，$\mathrm{Sim}(I_2,I_3)=-1$，完全负相关，相似度最低，而 I_3 与 I_2 的普遍评分都比较高，他们的相似度没有那么低. 对于判断 $I_1(2,1,2,1)$ 与 $I_4(2,1,2,2)$，$I_1(2,1,2,1)$ 与 $I_5(2)$ 谁更相似时，由于 I_1 与 I_5 只有一个项目评分一样，用 Pearson 相关系数计算 $\mathrm{Sim}(I_1,I_5)=1$，$\mathrm{Sim}(I_1,I_4)=0.5774$，而 I_1 与 I_4 有 3 个项目评分一致，它们相似度应该更高. 对于某些项目的评分，像 $I(1,1,1,1)$ 和 $I(5,5,5,5)$，用传统的相似度计算方法无法准确计算它们之间的相似度.

1.2　NNWD 算法设计

信息熵是信息论中用于度量信息混乱程度的一个概念. 信息越混乱，信息熵越大. 对于给定的样本集 X，它的信息熵公式为

$$H(X)=-\sum_{i=1}^{N}p(x_i)\log_2 p(x_i),\qquad(2)$$

式中：N 为 X 中分类的数量；$p(x_i)$ 为 X 中第 i 类元素出现的概率. 将信息熵用于项目之间相似度的计算，两个项目之间评分差异的信息熵越大，表示两个项目差异越混乱，相似度也就越低. 基于信息熵的相似度计算步骤如下：

(1)假设项目 I_1 和 I_2 共同评分的用户集合为 $U=\{u_1,u_2,\ldots,u_n\}$，I_1 和 I_2 的共同评分为 $I_1=(R_{u_1,I_1},R_{u_2,I_1},R_{u_3,I_1},\cdots,R_{u_n,I_1})$ 和 $I_2=(R_{u_1,I_2},R_{u_2,I_2},R_{u_3,I_2},\cdots,R_{u_n,I_2})$，$I_1$ 和 I_2 的评分差异度 $D(I_1,I_2)$ 定义为

$$\begin{aligned}D(I_1,I_2)&=(R_{u_1,I_1}-R_{u_1,I_2},R_{u_2,I_1}-R_{u_2,I_2},\\&\quad R_{u_3,I_1}-R_{uI,I_2},\cdots,R_{u_n,I_1}-R_{u_n,I_2})\\&=(d_1,d_2,d_3,\cdots,d_n).\end{aligned}\qquad(3)$$

(2)根据公式(2)，计算差异度的信息熵为

$$H[D(I_1,I_2)]=-\sum_{i=1}^{N}p(d_i)\log_2 p(d_i).\qquad(4)$$

这里 N 表示 d_i 的种类数，极端情况下若 d_i 全都相同，则 $N=1$. 考虑到评分差异 $|d_i|$ 对相似度的影响，$|d_i|$ 越大，相似度越低. 所以计算信息熵时，加入权重 $|d_i|$ 更加合理. 同时两个项目拥有的共同评价数 n 也会对相似度产生影响，n 越大，相似度越大，所以加入 $1/n$ 作为权重. 新的加权差异信息熵的计算公式为

$$NWD(I_1,I_2)=-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{p(d_i)}{N_i}\times\log_2 p(d_i)\times|d_i|\right]\qquad(5)$$

式中：n 为项目 I_1 和 I_2 的共同评分集合大小；d_i 为第 i 项评分的差值；N_i 为 d_i 在评分差异度集合 D 中出现的次数. 由公式可知，$NWD(I_1,I_2)$ 取值范围为 0 到 $+\infty$，$NWD(I_1,I_2)$ 越大相似度越低.

(3)将 $NWD(I_1,I_2)$ 归一化到 [0,1]

由于 $NWD(I_1,I_2)$ 越大相似度越低，所以采用如下归一化方法[6]

$$NNWD_{I_a}[i]=\frac{\mathrm{Max}(NWD_{I_a})-NWD_{I_a}[i]}{\mathrm{Max}(NWD_{I_a})-\mathrm{Min}(NWD_{I_a})},\qquad(6)$$

其中 $\mathrm{Max}(NWD_{I_a})$ 表示 $NNWD_{I_a}$ 集合中最大值；$\mathrm{Min}(NWD_{I_a})$ 表示 $NNWD_{I_a}$ 集合中最小值；$NNWD_{I_a}$ 就是归一化之后的相似度，取值范围为 0 到 1，值越大，项目间的相似度越高.

NNWD (Normalized New Weighted Differences)算法是利用两个项目之间的差异，将项目间共同评分的交集大小和差异大小作为权值加入到差异信息熵公式去，最后进行归一化处理，形成了归一化的新加权差异信息熵(NNWD)算法.

2　数据实验及结果分析

2.1　实验数据集

实验采用 MovieLens 站点(http://movielens.umn.edu)的实验数据，共汇总了用户 943 个，项目(影片)1 682 个，以及用户对影片产生的 100 000条评分记录，数据集稀疏度为 $1-100\,000/(943\times1\,682)\approx0.93\,695$[7]，非常稀疏. 用户评分从 1 到 5 五个等级. 数据集按 80% 和 20% 划分成训练集和测试集.

2.2　预测评分和度量方法

将相似性最高的若干项目作为目标项目 I_a 的邻居集合 $M=\{I_1,I_2,\cdots,I_k\}$，其中 $I_a\notin M$，集合 M 中的项目按照与 I_a 相似度从高到低排列. 根据 K 个最相似邻居预测目标用户 u 对项目 I_a 的评分，公式为[8]：

$$P_{u,I_a}=\overline{|R_{I_a}|}+\frac{\sum_{I\in M}sim(I_a,I)(R_{u,I}-\overline{R_I})}{\sum_{I\in M}|sim(I_a,I)|},\qquad(7)$$

式中：$R_{u,I}$ 为用户 u 对 I 的评分；$\overline{|R_{I_a}|}$ 和 $\overline{R_I}$ 为 I_a 和 I 的平均评分；$sim(I_a,I)$ 为 I_a 和 I 的相似度.

平均绝对误差(MAE)是最常用的用于统计测试集精准度的度量方法[9]. 设用户 u 对项目的

预测值集合为$\{p_1,p_2,\cdots,p_n\}$,用户u的实际评分集合为$\{q_1,q_2,\cdots,q_n\}$,平均绝对误差MAE定义为[10]

$$MAE=\sum_{i=1}^{n}|q_i-p_i|/n. \quad (8)$$

2.3 实验结果及分析

取测试集中10个项目来预测目标用户对它们的评分.分别取最邻近集合大小K为10到60,步长为10,在同一数据环境下,与基于余弦相似性的协同过滤、基于Pearson相似性的协同过滤、基于Spearman相似性的协同过滤进行比较.最终结果如图1所示,可以看出基于信息熵的相似度计算方法一定程度上优于其它方法.

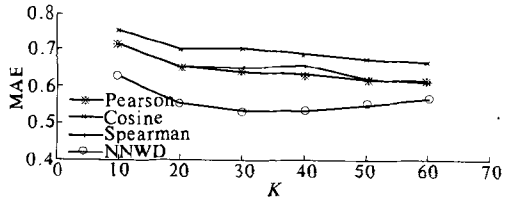

图1 不同的相似度计算方法产生的结果

Fig.1 The result of different similarity calculation methods

进而计算当$K=70,80,90$时,用NNWD方法的MAE值分别为0.5741,0.5712和0.5665.

3 结论

作者将信息论中的信息熵理论应用到协同过滤算法的相似度计算当中,又考虑到不同的差异度对相似性的影响,对信息熵计算方法进行相应的加权.运用基于项目相似性的协同过滤算法进行试验比较,相对于传统的方法提高了预测精度.

参考文献:

[1] 刘建国,周涛,王秉宏.个性化推荐系统的研究进[J].自然科学进展,2009,19(1):1-14.

[2] 许海玲,吴潇,李晓东,等.互联网推荐系统比较研究[J].软件学报,2009,20(2):350-362.

[3] 李涛.推荐系统中若干关键问题研究[D].南京:南京航空航天大学,2009.

[4] 罗辛,欧阳元新,熊璋,等.通过相似度支持度优化基于K近邻的协同过滤算法[J].计算机学报,2010,33(8):1437-1445.

[5] PANG Huan-li,ZHOU Lian-zhe,LIU Hai-mei. Personalization Portal System Based on Collaborative Filtering Algorithm[A]. International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE)[C]. Changchun,JL,China:IEEE Industrial Electronics Society,2010:383-386.

[6] 夏培勇.个性化推荐技术中的协同过滤算法研究[D].青岛:中国海洋大学,2011.

[7] 邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法[J].软件学报,2003,14(9):1621-1628.

[8] 吴月萍,郑建国.协同过滤推荐算法[J].计算机工程与设计,2011,32(09):3019-3021.

[9] 黄国言,李有超,高建培,等.基于项目属性的用户聚类协同过滤推荐算法[J].计算机工程与设计,2010,31(5):1038-1041.

[10] 孙小华.协同过滤系统的稀疏性与冷启动问题研究[D].浙江:浙江大学,2005.

Collaborative Filtering Algorithm Based on Weighted Information Entropy Similarity

LIU Wen-long[1], ZHANG Gui-yun[1], CHEN Zhe[2], ZHU Qiang-qiang[1]

(1. College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China; 2. College of Urban and Environmental Science, Tianjin Normal University, Tianjin 300387, China)

Abstract: Collaborative filtering algorithm is one of the most successful recommender system technology. The similarity calculation is the core of the collaborative filtering algorithm. In view of the poor predication quality existing in traditional similarity calculation with sparse data, we propose a similarity calculation method based on the information entropy between differences of items. First, we weight the entropy by the difference and common evaluation and then normalized it to measure the similarity between items. Verified by experiments with item-based collaborative filtering algorithm, the results show that it improves accuracy of personalized recommendation.

Key words: weighted information entropy; similarity calculation; collaborative filtering; personalized recommendation