文章编号:1674-7070(2010)04-0337-03

# 一种基于三部图网络的协同过滤算法

陈超[1]　张颖超[1]　缪进[1]

**摘要**

推荐系统是电子商务领域最重要的技术之一,而协同过滤算法又是推荐系统用得最广泛的.提出了一种基于加权三部图网络的协同过滤算法,用户、产品及标签都被考虑到算法中,并且研究了标签结点的度对用户相似性计算的影响.实验结果表明,此算法在解决用户冷启动问题的同时,还具有较高的推荐准确性.

**关键词**

推荐系统;协同过滤;二部图网络;三部图网络;相似性

**中图分类号** TP182

**文献标志码** A

## 0　引言

Introduction

为了解决信息过载这一日益突出的问题,20 世纪 90 年代初,学者们开发了最早的 GroupLens 推荐系统,帮助人们从众多的信息中得到自己想要的那部分信息. 推荐系统,是一种收集各个用户对产品的推荐(反馈)意见、产品内容、用户特征等信息,用特定的知识表示方法进行处理存储;然后利用推荐算法分析所获得的知识,针对特定用户的需求偏好为其推荐相应产品,帮助用户做出决策的智能决策支持系统. 20 世纪 90 年代中后期,推荐系统已经应用于音乐、电影、书籍等各种产品的推荐,现在,推荐系统已经成功融入于电子商务领域,成为电子商务系统不可或缺的一部分. 学者们提出了协同过滤算法[1]、基于内容的推荐算法、混合算法[2]等不同的算法来提供更好的推荐,并不断吸收数据挖掘领域、机器学习领域的新方法,将其应用到对推荐算法的改进上来. 一个好的电子商务推荐系统,可以极大地促进销售,并且对提高品牌形象也可起到积极的效果. 但目前的推荐算法仍然存在很多问题,特别在冷启动、稀疏性与延展性问题上处理不力,使得推荐精度不高以及个性化程度不够. 文献[3]从复杂网络的角度,用全新的视角研究了推荐算法,取得了很好的效果. 本文从复杂网络的角度提出一种基于加权三部图网络的推荐算法,不仅从用户与产品的角度计算用户间的相似度,而且把标签(关键字)也考虑进计算用户相似度的算法中,解决了用户冷启动问题,使得用户相似度的计算更加准确,推荐精度更高.

## 1　理论与方法

Theories and methods

### 1.1　二部图网络的设计

若无向图 $G=(V,E)$ 的顶点集合 $V$ 可以划分成两个子集 $X$ 和 $Y$,使 $G$ 中的每一条边 $e$ 的一个端点在 $X$ 中,另一个端点在 $Y$ 中,则称 $G$ 为二部图,可记为 $G=(X,E,Y)$,$X$ 和 $Y$ 称为互补结点子集,$E$ 为二部图 $G$ 中的边,如图 1 所示. 在推荐系统中,包含了用户和产品,且每个用户对应了一定量的产品,定义产品集 $O=\{o_1,o_2,\cdots,o_n\}$,用户集 $U=\{u_1,u_2,\cdots,u_m\}$ 为二部图中两个顶点的集合,如果用户被允许与产品进行对应,则这个推荐系统可以以一个 $n\times m$ 的邻接矩阵 $\{a_{ij}\}$ 来充分

**收稿日期** 2010-03-15

**资助项目** 江苏省"六大人才高峰"项目(06-A-027)

**作者简介**

陈超,男,硕士生,主要研究智能推荐系统. chenchao051@126.com

1　南京信息工程大学　信息与控制学院,南京,210044

表示. 若用户 $u_j$ 购买了产品 $o_i$,则 $a_{ij}=1$,否则 $a_{ij}=0$.

$$a_{ij}=\begin{cases}1, & x_i,y_j\in E;\\ 0, & \text{其他}.\end{cases} \tag{1}$$

这其中有个合理的假设,即用户所购买的产品一定是用户所喜爱的.

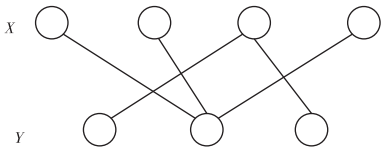


图 1　二部图网络

Fig. 1　Bipartite network

### 1.2　用户相似度计算

协同过滤算法的思想侧重于相似度的计算,通常对于一个用户行为的预测主要来自于另一些与其相似的用户,因此,可以使用下式来计算用户 $u_i$ 和用户 $u_j$ 之间的相似度:

$$s_{ij}=\frac{1}{\min\{k(u_i),k(u_j)\}}\sum_{l=1}^{n}\frac{a_{li}a_{lj}}{k(o_l)}. \tag{2}$$

其中 $k(u_i)$ 为用户结点 $u_i$ 的度.

对于任意未知的用户-产品配对且 $a_{ji}=0$,即用户 $u_i$ 没有购买过产品 $o_j$,在协同过滤算法中,预测评分(即用户 $u_i$ 对产品 $o_j$ 的喜爱程度)$v_{ij}$ 可通过公式(3)求得:

$$v_{ij}=\sum_{l=1,l\neq i}^{m}s_{li}a_{jl}\Big/\sum_{l=1,l\neq i}^{m}s_{li}. \tag{3}$$

因此,对用户 $u_i$,即可按 $v_{ij}$ 高低的降序排列向其推荐产品 $o_j$.

### 1.3　三部图网络与标签对算法的改进

考虑一种经常出现的情况,有个新用户进入系统,由于之前该用户在系统中没有任何的购买记录,所以应用式(2)无法计算该用户和其他用户的相似性,此即推荐系统中的冷启动问题,也是目前推荐系统面临的关键的几个问题之一. 所以本文在用户-产品的二部图基础上引入标签组成用户-产品-标签的三部图网络,如图 2 所示.

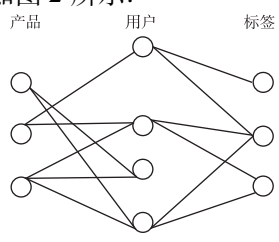


图 2　三部图网络

Fig. 2　Tripartite network

标签已经被越来越多的推荐系统所使用,当新用户进入系统的时候,他可以使用一些标签来描述自己的爱好或者购物倾向,这样系统就可以应用这些标签来计算新用户与其余用户之间的相似性,从而在新用户没有购买任何产品的情况下向新用户提供推荐.

定义标签集合 $T=\{t_1,t_2,\cdots,t_k\}$,如果用户 $u_i$ 使用了标签 $t_j$,则 $b_{ij}=1$,反之 $b_{ij}=0$. 这样,就可以由用户-标签构成另一个二部图网络,并计算用户间的相似度,在这个网络下,用户 $u_i$ 和用户 $u_j$ 之间的相似度 $s_{ij}'$ 由下式定义:

$$s_{ij}'=\sum_{t=1}^{k}b_{ti}b_{tj}\Big/\min\{k(u_i),k(u_j)\}. \tag{4}$$

其中 $k(u_i)$ 和 $k(u_j)$ 分别是用户 $u_i$ 和用户 $u_j$ 的度. 在现实情况中,用户如果选择了那些大部分人都选择过的标签,那么对计算相似性有着负面的影响. 比如在购物系统中,有个标签叫做“毛巾”,因为毛巾是日常必需用品,所以几乎每个人都会选,而购买毛巾的人很可能不都具有很大的相似性,所以当某个用户选择标签“毛巾”来代表自己的购物爱好时,就可能对最后的相似性计算有很大的负面影响. 基于以上观点,可以认为抑制大度标签对用户相似性计算的影响可以使推荐结果更加准确. 所以可以对式(4)进行改进得到新的基于标签的用户相似度计算公式:

$$s_{ij}'=\frac{1}{\min\{k(u_i),k(u_j)\}}\sum_{l=1}^{k}\frac{b_{li}b_{lj}}{k^{\alpha}(t_l)}. \tag{5}$$

其中:$k(t_l)$ 是标签 $l$ 的度;$\alpha$ 是自由变量,当 $\alpha=0$ 时,式(5)与式(4)相同. 为了抑制大度标签对用户相似性计算结果的影响,自然希望 $\alpha$ 能取到一个合适的正数值,下文的实验结果也很好地验证了这一想法.

由之上所讨论的用户-产品及用户-标签两个二部图网络,可以组成产品-用户-标签三部图网络. 采用线性形式组合上面的两个相似性[4]:

$$s_{ij}^{*}=\lambda s_{ij}+(1-\lambda)s_{ij}'. \tag{6}$$

其中 $\lambda\in[0,1]$,是一个可变参数,结合式(3),可以得到用户 $u_i$ 对产品 $o_j$ 的预测评分:

$$v_{ij}=\sum_{l=1,l\neq i}^{m}s_{li}^{*}a_{jl}\Big/\sum_{l=1,l\neq i}^{m}s_{li}^{*}. \tag{7}$$

因此,对用户 $u_i$,即可按 $v_{ij}$ 高低的降序排列推荐那些该用户还没有购买过的产品 $o_j$.

## 2　实验及分析

Experiment and analysis

使用 MovieLens 数据集(www. grouplens. org)来测试本文的算法. MovieLens 是一个评分系统,每个

用户可以为每部电影打分(1~5 分),从 2006 年开始,MovieLens 加入了标签功能. 只考虑那些至少被两个用户选择过的电影和标签,并且也只考虑至少选择了一部电影和一个标签的用户. 经过以上条件筛选以后,数据集中有 3 710 个用户,5 724 部电影以及 5 228 个标签. 用户-产品关系有 53 091 条边,用户-标签关系有 33 065 条边. 为了测试算法的表现,本文把数据集随机分为训练集和测试集,训练集包含 90% 的数据,而测试集包含 10% 的数据.

使用文献[3]中的评判标准来测试算法. 给定任意的用户 $u_i$,如果 $u_i$-$o_j$在测试集中,则相对于训练集来说,$o_j$就是一个没被选择过的产品,便可以度量 $o_j$ 在测试集列表中的位置. 例如,如果有 1 500 部影片是用户 $u_i$没选择过的,并且 $o_j$出现在推荐列表的第 30 位,可认为 $o_j$的位置为 30/1 500,$r_{ij}=0.2$. 注意到一个事实,测试集中的产品实际上是用户选择过的,所以 $r_{ij}$越小,表明算法越优秀.

在此主要测试参数 $\alpha$ 对整个算法表现的影响. 根据文献[4]的描述,当 $\lambda=0.8$ 左右时,推荐精度最高,因此为了测试方便,设定 $\lambda=0.8$,并且,设定推荐列表长度为 20. 通过图 3 可以很清楚地看到,当 $\alpha=1.98$ 时,$r=0.171$,推荐准确度最高,并且推荐精度要比 $\alpha=0$ 时(此时,$r=0.198$),即非加权三部图网络的情况下更高.

因此,抑制那些度大的标签的作用对最后的推荐结果有着比较大的影响,取一个合适的 α 值能显著地提高推荐精度.

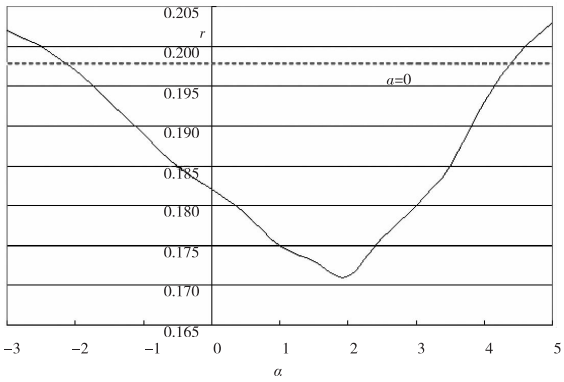


图 3 数值实验结果

Fig. 3 Numerical experiment results

## 3 结论

Conclusion

本文提出了一个基于加权三部图网络的推荐算法,通过引入标签来解决新用户的冷启动问题,并且也考虑到了拥有不同度的标签对于计算用户相似度的影响,通过抑制那些度比较大的标签结点的作用,来提高用户相似度计算的准确性. 实验结果表明,本文算法在解决用户冷启动问题的同时,也有着较高的推荐精度.

## 参考文献

References

[1] Sarwar B,Karypis G,Konstan J,et al. Item-based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International Conference on WWW. Hong Kong: IEEE Press, 2001:285-295

[2] Pazzani M J. A framework for collaborative,content-based,and demographic filtering[J]. Artificial Intelligence Review,1999,13(5/6):393-408

[3] Zhou T,Ren J,Medo M,et al. Bipartite network projection and personal recommendation [J]. Physical Review E,Statistical, Nonlinear,and Soft Matter Physics,2007,76(4):046115

[4] Zhang Z K,Zhou T,Zhang Y C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs[J]. Physica A,2010,389:179-186

# A collaborative filtering recommender algorithm based on tripartite network

CHEN Chao[1] ZHANG Yingchao[1] MIAO Jin[1]

1 School of Information & Cybernetics,Nanjing University of Information Science & Technology 210044

**Abstract** Recommender system is one of the most important technologies in E-commerce,and the collaborative filtering algorithm is the most widely used technique in recommender system. In this paper,we proposed a collaborative filtering algorithm based on weighted tripartite network,which takes users,items and tags into account,and we also studied the degree of tags which may affect the user-user similarity computation. The experimental results demonstrate that the algorithm can solve the cold start problem with high recommendation accuracy.

**Key words** recommender system;collaborative filtering;bipartite network;tripartite network;similarity