

基于网络链接预测的推荐算法

华秋云¹, 陈 峻^{1,2}

(1.扬州大学信息学院计算机系,江苏 扬州 225009)
(2.南京大学软件新技术国家重点实验室,江苏 南京 210093)

[摘要] 提出基于二部网络连接预测的推荐算法. 将用户-项目的评分矩阵用带权的二部网络来表达,根据推荐问题和带权二部网络连接预测问题的相似性将推荐问题抽象为二部网络上的链接预测问题,采用基于相似度的连接预测算法进行项目推荐. 算法综合考虑了顶点间的拓扑关系,以及用户之间、项目之间的相似性,找出用户对其尚未表达的项目的潜在兴趣度,应用二部网络连接预测的算法来解决推荐问题. 实验结果表明,算法能够有效地提高推荐的精度.

[关键词] 带权二部网络,链接预测,相似性,推荐算法

[中图分类号] TP311 **[文献标志码]** A **[文章编号]** 1001-4616(2015)01-0075-06

A Recommendation Algorithm Based on Network Link Prediction

Hua Qiuyun¹, Chen Ling^{1,2}

(1.Department of Computer Science, Information Institute, Yangzhou University, Yangzhou 225009, China)
(2.State Key Lab of Novel Software Tech, Nanjing University, Nanjing 210093, China)

Abstract: An algorithm for recommendation based on link prediction in a bipartite network is presented. We use a weighted bipartite network to represent the user-item matrix. Due to the similarity between recommendation and link prediction in bipartite network, we transform the recommendation into the problem of link prediction in bipartite network. The similarity based method is used to predict the potential links considering the topological similarity between the nodes in the network, the similarity between users and the similarity between the items. The potential interests of the users to the items can be found by link prediction in the bipartite network. Our experimental results show that our algorithm can get high recommendation results.

Key words: weighted bipartite network, link prediction, similarity, recommendation algorithm

据 CNNIC 统计,截至 2013 年 6 月底,中国网民规模达到 5.91 亿,较 2012 年底增加 2 656 万人. 互联网普及率为 44.1%,较 2012 年底提升 2%. 中国手机网民规模达 4.64 亿,较 2012 年底增加 4 379 万人. 互联网已经深入人们生活的各个方面,越来越多的人愿意在互联网上阅读新闻和书籍.

伴随着人们对互联网的依赖,人们可以获取的信息量也呈直线上升. 没有人愿意花精力在淘宝中输入衣服然后去挑选数以亿计的商品,也没有人愿意花时间在亚马逊阅读全部书籍. 因为其中包含的信息量太大,这也就是所谓的信息量超载. 个性化推荐系统^[1]已经成为当前解决信息超载问题的最为有效的信息过滤的技术之一.

个性化推荐技术,是电子商务中引入的一种崭新的智能信息服务方式^[2],个性化推荐技术研究的主要内容是如何高效地从庞大的信息海洋中挖掘出有用的内容,如何动态准确地发现用户的兴趣偏好,如何主动并实时地向用户提供与其个性化偏好相符合的信息,如何有效提高推荐能力等问题. 个性化推荐技术采用统计分析和数据挖掘技术设计合理的算法,通过分析用户行为数据和用户的评价等发现用户的真实需求,进而为用户提供更加能满足其需求的以人为中心的一对一推荐服务^[3].

收稿日期:2014-08-16.

基金项目:国家自然科学基金(61379066、61070047、61379064)、江苏省自然科学基金(BK20130452、BK2012672、BK2012128)、江苏省教育厅自然科学基金(12KJB520019、13KJB520026、09KJB20013).

通讯联系人:陈峻,教授,博士生导师,研究方向:并行计算、数据挖掘、计算机软件、人工智能、蚁群算法等. E-mail: yzuchan@163.com

推荐算法是整个推荐系统的核心部分^[4],其性能的好坏关系到整个推荐系统的推荐效率和推荐质量.个性化推荐系统的快速发展,使得研究人员必须不断地根据不同的情况提出更符合当下情景的推荐模型.在现有的推荐系统中,目前采用最多的推荐算法主要有基于关联规则的推荐算法、基于内容的推荐算法和基于协同过滤的推荐算法,以及综合运用两种或两种以上算法的混合推荐算法.

基于关联规则的推荐^[5,6]是以关联规则为基础,把已购商品作为规则头,规则体为推荐对象.关联规则挖掘可以发现不同商品在销售过程中的相关性,在零售业中已经得到了成功的应用.该方法不需要领域知识就能够发现新兴趣点,但不容易发现规则,规则抽取困难且耗时,是算法的瓶颈,时间复杂度,个性化程度较低.

基于内容的推荐算法^[7,8]根据历史信息(如评价、分享、收藏过的文档)构造用户偏好文档,计算推荐项目与用户偏好文档的相似度,将最相似的项目推荐给用户.基于内容推荐只考虑用户的兴趣偏好,推荐结果直观容易理解,但是受数据格式影响,对于机器难以理解的数据格式,如图像、音乐等通常无法处理,不能为用户发现新的兴趣点.

协同过滤推荐算法^[9,10]的基本假设是:为用户推荐感兴趣的内容可通过找到与该用户偏好相似的其他用户,将他们感兴趣的内容推荐给该用户.然而协同过滤方法只考虑了用户评分数据,忽略了项目和用户本身的诸多特征.协同过滤推荐算法不受数据格式影响,能够推荐难以处理的复杂数据,能挖掘用户新的兴趣点,但也同时存在多个问题,最典型是稀疏问题和冷启动问题.

本文根据推荐问题和带权二部网络连接预测问题的相似性,将推荐问题抽象成二部网络上的链接预测,提出基于链接预测的推荐算法.我们将用户-项目的评分矩阵用带权的二部网络来表达,以采用基于相似度的连接预测算法进行项目推荐,找出用户对其尚未表达的项目的潜在兴趣度.算法综合考虑顶点间的拓扑关系,定义了用户之间、项目之间、以及用户和项目之间的邻居节点,再根据邻居节点定义用户之间、以及用户和项目之间的相似性,应用这种相似性作为用户对其尚未表达的项目的潜在兴趣度的估计.由于算法充分利用了二部网络的拓扑信息和用户-项目的评分信息,有效地提高了推荐结果的精度.

1 算法的基本思想

一个由 n 个用户和 m 个项目组成的推荐系统,可用一个 $n \times m$ 的用户-项目兴趣度矩阵 $P = [p_{ij}]$ 来表示,矩阵 P 的部分元素已被赋值,另一部分元素尚未被赋值,我们记其值为 0. 矩阵 P 的元素 p_{ij} 的值非 0,表示用户 i 对项目 j 的兴趣度, p_{ij} 值越大表示用户 i 越喜欢项目 j , p_{ij} 的值可以使任何实数,不过通常取某个区间上的整数.矩阵 P 的元素 p_{ij} 的值为 0 表示用户 i 对项目 j 尚未表达其兴趣度.推荐问题就是:已知的用户-项目兴趣度矩阵 $P = [p_{ij}]$,要求出对某些用户 i 尚未表达其兴趣度的项目 j ,估计其兴趣度的值,选取其中兴趣度最高的若干个推荐给该用户.

根据上述推荐问题和带权二部网络连接预测问题的相似性,我们将推荐问题转换为二部网络上的链接预测问题来解决.我们将用户-项目的评分矩阵用带权的二部网络^[11]来表达,这样可以综合考虑顶点间的拓扑关系,以及用户之间、项目之间的相似性,应用二部网络连接预测的算法来解决推荐问题,以提高推荐的精度.

我们对一个由 n 个用户和 m 个项目组成的推荐系统,通过如下的方法来构建一个带权二部网络 $G = (U, V, E, P)$:二部网络 G 中 U 和 V 为两部分的顶点, U 为用户顶点, V 为项目顶点, $|U| = n, |V| = m$. 在 U 、 V 各自内部的顶点之间不存在边相连.设如果用户 x 购买且评价过项目 a ,则用户 x 和项目 a 的相应的顶点就被一条边 (x, a) 连接,这里 $x \in U, a \in V$. 在边 (x, a) 上的权值表示用户 x 对项目 a 的偏好程度,设为用户 x 对项目 a 兴趣度值 p_{xa} . 这样就构成了用户-项目二部网络 G .

链接预测是指如何通过已知的网络结构等信息评估预测网络中尚未链接的两个节点之间存在或产生链接的可能性^[14].链接预测包括:(1)预测已存在但尚未被发现的链接,即预测未知链接;(2)预测现在未存在但未来可能新产生的链接,即预测未来链接^[15].链接预测最主要的方法是通过估计顶点之间的相似性来衡量潜在链接出现的可能性.

我们的算法首先根据用户-项目兴趣度矩阵构造一个带权的用户-项目二部网络,然后在这个用户-项目二部网络上进行用户-项目顶点之间进行连接预测.具体地,就是对尚未链接的用户-项目顶点对(即

尚未评分的用户-项目对)计算其相似度,并以此相似度作为用户对其尚未表达的项目的潜在兴趣度的估计.

2 带权二部网络中基于共同邻居的相似度

我们在二部网络中采用基于相似度的方法进行链接预测,但目前现有的链接预测的相似度大多是针对单部网络的,而且仅适用于无权图. 因此有必要针对带权的二部网络定义顶点间的相似度. 我们将常用的共同邻居(Common Neighbor)相似度度量扩展到带权二部网络上^[18]. 共同邻居算法是最简单的基于局部信息的相似性定义方法,该算法中假设与两个节点相连的公共节点数量越多,它们之间的关联性越强,即相似性程度越高. 换句话说,就是如果两个节点间的共同邻居节点个数越多,那么两者间存在链接的可能性越大.

设有带权二部网络 $G=(U,V,E,P)$, 其中 U 和 V 分别为用户、项目两部分的顶点, $|U|=n, |V|=m, E$ 为网络 G 中链接的集合, 在二部网络 G 中, 链接只在不同的二部网络的顶点之间产生. $n \times m$ 矩阵 $P=[p_{ij}]$ 为权矩阵. p_{ij} 为非 0 值时, 即为相应边上的权值, p_{ij} 为 0 值时, 表示相应的顶点间不存在链接.

由于在二部网络中, 链接只在不同的二部网络的顶点之间产生, 我的仅定义顶点对 $(x, a), x \in U, a \in V$ 之间的相似度. 但由于不同的二部网络的顶点之间不存在直接的相似关系, 我们只能通过他们所连接的顶点之间来寻求其相似性. 为此, 首先我们对两部分的顶点定义“邻居”的概念:

定义 1 用户顶点之间的邻居

用户顶点之间的邻居是指对相同的项目进行评分的不同用户顶点的集合, 设 $x \in U$, 定义其在 U 中的邻居集合:

$$N_{uu}(x) = \{y | y \in U, \exists a \in V, (x, a) \in E \text{ 且 } (y, a) \in E\}, \quad N_{uu} \subset U,$$

即与 x 至少有一个共同的 V 中连接顶点 a 的 U 中的顶点集合.

定义 2 项目顶点在用户顶点中的邻居

项目顶点在用户顶点中的邻居是指对同一项目都有评分的用户顶点的集合, 设 $a \in V$, 定义其在 U 中的邻居集合:

$$N_{uv}(a) = \{x | x \in U, (x, a) \in E\}, \quad N_{uv} \subset U,$$

即与 a 相连的 U 中的顶点集合.

定义 3 用户顶点在项目中的邻居

用户顶点在项目中的邻居是指同一用户在不同的项目中都有的评分的项目顶点的集合, 设 $x \in U$, 定义其在 V 中的邻居集合:

$$N_{uv}(x) = \{a | a \in V, (x, a) \in E\}, \quad N_{uv} \subset V.$$

定义 4 用户顶点与项目顶点的共同邻居

用户顶点与项目顶点的共同邻居是指用户顶点之间的邻居和用户顶点在项目顶点中的邻居的交集, 设 $x \in U, a \in V$, 它们的共同邻居集合定义为:

$$\Gamma_{uv}(x, a) = N_{uu}(x) \cap N_{uv}(a), \quad \Gamma_{uv}(x, a) \subset U.$$

定义 5 用户顶点之间的共同邻居

用户顶点之间的共同邻居是指用户 x 和 y 共同评分的项目集, 设 $x, y \in U$, 它们在 V 中的共同邻居集合定义为:

$$\Gamma_{uu}(x, y) = N_{uv}(x) \cap N_{uv}(y), \quad \Gamma_{uu}(x, y) \subset V.$$

在定义了共同邻居的基础上我们来定义用户顶点与项目顶点的相似性. 相似性是用来衡量节点的接近程度. 两个节点的相似性程度越高, 两者间链接存在的概率就越大, 这是基于相似性链接预测方法的重要前提条件^[19]. 不同算法的区别则在于节点相似性的计算方法上, 定义一个基于节点相似性, 只需要给出节点之间相似性的计算算法. 首先我们定义用户节点之间的相似度, 在此基础上定义用户与项目节点之间的相似度.

设 $\Gamma_{uu}(x, y)$ 表示用户 x 和 y 共同评分的项目集. 不同的评分值表示不同的兴趣度, 在比较两用户的喜好差异时, 没有同时评过的项目不具有可比性, 因此可以用 $\Gamma_{uu}(x, y)$ 中项目评分差来表示不同用户

之间的兴趣度的差异.

定义 6 用户顶点之间的距离

设 $x, y \in U, \Gamma_{uu}(x, y) \subseteq V$ 表示用户 x 和 y 共同评分的项目集, $p(x, a)$ 表示用户 x 对项目 a 的评分, 用户节点 x, y 之间的距离为

$$\text{Dis}(x, y) = \frac{1}{|\Gamma_{uu}(x, y)|} \sum_{a \in \Gamma_{uu}(x, y)} \frac{|p(x, a) - p(y, a)|}{\max(p(x, a), p(y, a))}. \quad (1)$$

式(1)表明, $\text{Dis}(x, y)$ 越大, 则表示 x 和 y 的喜好差异越大, 所以当 $\Gamma_{uu}(x, y)$ 一定时, $\text{Dis}(x, y)$ 越大, 它们之间的相似度就越小. 而如果 $\Gamma_{uu}(x, y)$ 越大, 则表示 x 和 y 相似的可能性越大. 因此,

定义 7 用户顶点之间的相似度

用户顶点 x, y 之间的相似度定义为:

$$\text{Sim}_{uu} = \frac{1}{\text{Dis}(x, y)}, \quad (2)$$

通过式(2)计算, 充分考虑了用户节点之间的相似度的属性, 根据用户节点 x 和 y 间的相似度, 再结合用户 y 对项目 a 的偏好程度可实现对用户 x 的项目推荐. 为了更好构成推荐结果的兴趣度估计值, 我们在此基础上定义用户与项目节点之间的相似度.

定义 8 用户与项目节点之间的相似度

设 $x \in U, a \in V, \Gamma_{uv}(x, a) \subseteq U$ 表示用户 x 和项目节点 a 的共同邻居, 用户 x 与项目节点 a 之间的相似度为

$$\text{Sim}_{uv}(x, a) = \frac{\sum_{y \in \Gamma_{uv}(x, a)} \text{Sim}_{uu}(x, y) p(y, a)}{\sum_{y \in \Gamma_{uv}(x, a)} \text{Sim}_{uu}(x, y)}. \quad (3)$$

以 $\text{Sim}_{uv}(x, a)$ 作为用户 x 对项目 a 的兴趣度估计值. 对于所要查询的用户 x , 我们设其所有尚未评分的项目为 a_1, a_2, \dots, a_k , 计算用户 x 对它们的相似度 $\text{Sim}_{uv}(x, a_i) (i=1, 2, \dots, k)$, 然后对这些相似度进行排序, 取其中较大的 $K (K < k)$ 个作为对用户 x 的推荐项目.

3 算法的框架

综上所述, 基于链接预测的推荐算法的总体框架如下:

算法 LPR(Link Prediction based Recommendation)

输入: P : 二部网络的用户-项目评分矩阵;

输出: $S: S = [\text{Sim}_{uv}(i, j)]$ 为表示推荐结果的评分矩阵;

begin

(1) 计算集合用户顶点之间的邻居 N_{uu} 、项目顶点在用户顶点中的邻居 N_{uv} 、用户顶点在项目中的邻居 N_{vu} 、用户顶点和项目顶点的个共同邻居 Γ_{uv} 、用户顶点之间的共同邻居 Γ_{uu} ;

(2) 对所有的用户 $i=1, \dots, n$, 计算用户结点之间的相似度 $\text{Sim}_{uu}(i, j)$;

(3) 对所有的用户 $i=1, \dots, n$ 和所有的项目 $j=1, \dots, m$, 计算用户与项目结点之间的相似度 $\text{Sim}_{uv}(i, j)$;

(4) 通过以上计算构成表示推荐结果的评分矩阵 $S = [\text{Sim}_{uv}(i, j)] = [S_{ij}]$, 根据构成的推荐结果评分矩阵进行推荐;

(5) 对于用户 i , 设其尚无评分的项目为 j_1, j_2, \dots, j_h , 在 $S_{ij_1}, S_{ij_2}, \dots, S_{ij_h}$ 中选比较大的 N 个作为对用户 i 的 TOP-N 个项目推荐.

End

4 实验结果及分析

为了测试上述算法 LPR 的性能, 我们用著名的电影网站 MovieLens 的实际数据对算法进行了实验, 并与其他推荐算法进行了分析比较. 实验在内存为 4 G、有关实验的所有程序使用 matlab 语言编写, 底层操作系统是 64 位的 windows7 操作系统.

4.1 数据集

本文采用的实验数据来自于电影网站 MovieLens. 该网站的数据集被普遍采用的推荐算法检验数据集. 我们选取数据集的 100 K 的数据,其中包括 943 个用户对 1682 部电影的评价信息,而且每个用户的评分电影数不少于 20 部,评分的结果从 1 分(最差)到 5 分(最好),并包含了电影的分类信息以及用户的基本信息. 用户评分表(u. data)包含了 943 个用户的 100 000 条评分信息,表中共有 4 个字段:用户编号(User)、项目编号(Movie)、评分(Rating)及评分时间(Timestamp).

为了进行实验比较,将数据集进一步划分为训练集和测试集,对此,还将引入变量 x ,表示训练集占整个数据集的百分比. 比如, $x=0.8$ 表示数据集中的 80% 作为训练集,20% 作为测试集. 在实验中,我们取 x 的值为 0.8.

4.2 评价方法

推荐算法的推荐质量评价标准中,目前最常用的是平均绝对差(MAE)标准,它是通过计算预测的用户对商品的兴趣度与用户对该商品的真实兴趣度之间的偏差来度量算法预测的准确性,MAE 值越小,则预测越精确,推荐精度也越高. 假设利用文中的算法预测用户对 k 个商品的评分值分别为 $\{p_1, p_2, \dots, p_k\}$, 用户真实的兴趣度值为 $\{r_1, r_2, \dots, r_k\}$, 则平均绝对偏差描述为

$$MAE = \frac{\sum_{i=1}^k |p_i - r_i|}{k}.$$

在实验结果的分析 and 比较中,我们采用平均绝对差(MAE)标准衡量推荐结果的质量.

4.3 实验结果

本文实验先将用户-项目的评分矩阵用带权的二部网络来表示,再根据相似度计算得出用户节点之间的相似度,将用户寻找其相似的 k 个近邻作为对用户的项目推荐的参考,然后计算得出用户与项目之间的相似度构成表示推荐结果的评分矩阵,对用户进行项目推荐,找出用户对其尚未表达的项目的潜在兴趣度.

本实验的目的是验证本文提出的基于链接预测的推荐算法(LPR)的有效性,同时分别与基于双向关联规则项目评分预测的推荐算法(BAR-CF)、基于项目评分预测的推荐算法(IR-CF)、基于链接预测的移动用户偏好预测方法(PLP)进行比较,比较它们推荐结果的平均绝对差(MAE). 实验选取的最近邻居个数 k 从 4 逐渐增加到 20,间隔为 4,实验结果如图 1 所示.

由图 1 可以看出,随着选取的最近邻居个数的增加,所比较的推荐算法的平均绝对差都有不同程度的降低,推荐质量不断提高. 但在选取同一个最近邻居数目的情况下,本文提出的基于链接预测的推荐算法 LPR 的平均绝对差明显小于其他所与之比较的推荐算法,推荐精度显著提高,说明本文提出的算法明显优于传统的推荐算法.

5 结束语

本文提出一个基于链接预测的推荐算法. 该算法把推荐问题和链接预测结合起来,将推荐问题抽象为二部网络上的链接预测,将用户-项目的评分矩阵用带权的二部网络来表达,根据推荐问题和带权二部网络连接预测问题的相似性,采用基于相似度的连接预测算法进行项目推荐,找出用户对其尚未表达的项目的潜在兴趣度. 算法综合考虑了顶点间的拓扑关系,以及用户之间、项目之间的相似性,应用二部网络连接预测的算法来解决推荐问题,以提高推荐的精度. 由于算法充分利用了二部网络的拓扑信息和用户-项目的评分信息,有效地提高了推荐结果的精度. 实验结果表明,该算法与其他类似算法相比,有效地提高了对用户的项目推荐精度.

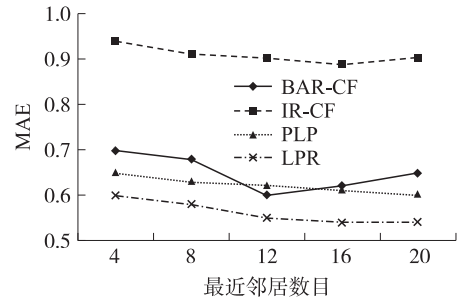


图 1 算法推荐精度比较

Fig. 1 Comparison of the accuracy of recommendation algorithm

[参考文献]

- [1] 刘建国,周涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1-15.
- [2] Schafer B,Konstan J A. E-commerce recommendation applications[EB/OL]. <http://citeseer.nj.nec.com>,2010,5(1):115-153.
- [3] Mobasher B,Dai H,Luo I,et al. Discovery of aggregate usage profiles for web personalization[C]//Proceedings of the Web KDD Workshop on Web Mining for E-commerce-Challenges and Opportunities.NY,2000,4(1):1-11.
- [4] 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):1-10.
- [5] LI J,XU Y,WANG Y F,et al. Strongest association rules mining for efficient applications[C]//Proceedings of the Fourth IEEE Conference on Service Systems and Service Management. Piscataway:IEEE Press,2007:502-507.
- [6] 李杰,徐勇,王云峰,等. 面向个性化推荐的强关联规则挖掘[J]. 系统工程理论与实践,2009,29(8):144-152.
- [7] Wang F H,Jian S Y. An effective content-based recommendation method for Web browsing based on keyword context matching[J]. Journal of Informatics and Electronics,2006,1(2):49-59.
- [8] Wartena C,Slakhorst W,Wibbels M,et al. Selecting keywords for content based recommendation[C]//CIKM'10:Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York:ACM Press,2010:1 533-1 536.
- [9] Herlocker J L,Konstan J A,Terveen L G,et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems,2004,22(1):5-53.
- [10] Chen Y L,Cheng L C. A novel collaborative filtering approach for recommending ranked items[J]. Expert Systems with Applications,2008,34(4):2 396-2 405.
- [11] 张新猛,蒋盛益. 基于加权二部图的个性化推荐算法[J]. 计算机应用,2012,32(3):654-657.
- [12] Zhou Tao,Ren Jie,Medo M,et al. Bipartite network projection and personal recommendation[J]. Physical Review E,2007,76(4):046115.
- [13] Liu Jianguo,Wang Binghong,Guo Qiang. Improved collaborative filtering algorithm via information transformation[J]. International Journal of Modern Physics C,2009,20(2):285-293.
- [14] 王茜,段双艳. 一种改进的基于二部图网络结构的推荐算法[J]. 计算机应用研究,2013,33(3):771-774.
- [15] Getoor L,Diefl C P. Link mining:a survey[J]. ACM SIGKDD Explorations Newsletter,2005,7(2):3-12.
- [16] Lü Linyuan. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China,2010(9):5-39.
- [17] 姜波,张晓筱,潘伟丰,等. 基于二部图的服务推荐算法研究[J]. 华中科技大学学报,2013,12:93-99.
- [18] Newman M. The structure and function of complex networks[J]. SIAM Review,2003,45(2):167-256.
- [19] Liben-Nowell D,Kleinberg J. The link-prediction problem for social networks[J]. J Am Soc Inform Sci Technol,2007,58(7):1 019-1 031.
- [20] He Yinhui,Chen Duanbing. Similarity algorithm based on users common neighbors and grade information[J]. Computer Science,2010(9):184-187.

[责任编辑:顾晓天]