基于加权两层图的混合推荐方法

陈 泽 王国胤 胡 峰
(重庆邮电大学计算机科学与技术研究所 重庆 400065)

摘 要 结合用户-项目评分矩阵和项目-类别关联矩阵,提出了一种新的混合推荐模型。首先,利用用户-项目评分矩阵和项目-类别矩阵,提出一种新的项目关联度度量方法,该方法根据项目的特征信息和当前评分数据的稀疏情况,动态调节关联度的计算值,真实地反映彼此之间的关联度;其次,分别以项目关联度和用户-项目评分信息为权值,构建一个基于用户-项目的加权两层图模型;在此基础上,从两层图的全局结构出发,结合随机游走算法给出了基于加权两层图的推荐算法,以为用户提供个性化的项目推荐和用户推荐。实验结果表明,该算法相比文献中的其他推荐方法具有更高的准确度。

关键词 随机游走,混合推荐,项目类别,两层图
中图法分类号 TP311 **文献标识码** A

Hybrid Recommendation Filtering Method Based on Weighted Two-layer Graph

CHEN Ze WANG Guo-yin HU Feng
(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract Combined with the rating matrix of user-item and the correlation matrix of item-category, a new hybrid recommended model was proposed. First, a new correlation degree measuring algorithm was presented by using these two matrixes. This algorithm takes into account the feature information and dynamically adjusts the result based on the sparse situation of the rating data, truly reflects the degree of association with each other. Then, a new weighted two-layer graph model was constructed by using the item-item correlation degree and the user-item correlation degree as the weight. On this basis, starting from the global structure of the two-layer graph, the recommendation algorithm based on weighted two-layer graph was given by the random walk algorithm, to provide users with personalized item recommendations and user recommendations. The experiments show that the algorithm compared to other recommended models in the references has higher accuracy.

Keywords Random walk, Hybrid recommendation filtering, Item category information, Two-Layer graph

1 引言

推荐系统根据用户与系统的交互历史、项目的基本信息以及用户的个人信息等构建用户的兴趣模型,去预测用户可能感兴趣的产品或项目。推荐系统大致可以分为3类:基于协同过滤的方法(Collaborative Filtering)[1]、基于内容的方法(Content-based Filtering)和将二者结合的混合推荐算法(Hybrid Recommendation)[2]。

基于协同过滤的方法是目前应用最广泛的推荐技术,它首先利用用户的历史评分等信息来计算用户之间的相似性,然后使用与目标用户相似性较高的邻居用户对项目的评价信息,来预测目标用户对还未接触过项目的喜好程度,系统就根据这种喜好程度对其进行项目推荐。这种推荐技术不必对项目进行内容特征分析,并能发现潜在的项目,但它存在新用户推荐和评价信息稀疏的问题。基于内容的推荐是根据用户已经选择的项目,对其进行内容分析,以内容特征上的相似性来对目标用户进行相应推荐,不依赖于用户对项目的评价。但是这种推荐技术的问题在于,内容分析的方法并不能适用于所有类型的产品,一些电影或音乐项目的内容特征就很难被有效地提取出来,同时它无法帮助用户发现潜在的项目。基于协同过滤和内容的混合推荐模型融合了基于协同过滤和基于内容过滤的优点,它综合考虑用户和项目的各种影响因素,充分利用系统提供的各种用户信息和项目信息,可以有效地提高推荐质量,产生较好的推荐效果[2]。

混合推荐算法虽然能够结合另外两种方法的优点来提高推荐质量,但是在很多方面都还存在问题,其中最大的问题就是怎么样把用户和项目基本信息集成到协同过滤方法中[3]。当前主要有两种方案:一种是利用用户和项目的特征信息,从内容的角度计算用户和项目的关联度,然后通过加权的方法将其集成到协同过滤的关联度中,通过参数的训练权值使得

到稿日期:2012-02-12 返修日期:2012-06-25 本文受国家自然科学基金(61073146),中国与波兰政府间科技合作项目(国科外字[2010]179号),重庆市教委科学技术研究项目(KJ110522)资助。

陈 泽(1988—),男,硕士,主要研究领域为智能信息处理,E-mail:471519678@qq.com;王国胤(1970—),男,博士,教授,博士生导师,主要研究领域为Rough集理论、粒计算、数据挖掘、知识技术等;胡 峰(1978—),男,博士,副教授,主要研究领域为智能信息处理。

推荐精度达到最优；另一种是由 Huang Zan 等人提出的基于概念语义空间可视化的两层图(Two-Layer Graph Model)模型[3]，该模型包含用户层和项目层两层，两层内部用户或项目之间通过基于内容的方法计算出用户或项目的相似度(关联度)，两层之间通过用户-项目之间的显式或隐式历史评分数据进行关联。

本文结合上述两种集成方案，提出了一种新的混合推荐模型。首先将项目-类别的关联矩阵和用户-项目评分矩阵，分别从基于内容和基于协同过滤两个角度计算两项目之间的关联度，并利用两个项目共同评分用户个数对该关联度进行优化处理，获得最优的项目-项目关联度。然后利用上一步获得的项目-项目关联度和用户-项目的评分矩阵作为权值，构建一个基于用户-项目的加权两层图模型。在此基础上，从两层图的全局结构出发，结合随机游走算法，给出基于加权两层图的推荐算法，以为用户提供个性化的项目推荐和用户推荐。实验结果表明，该算法相比文献中其他推荐模型具有更高的准确度。

2 结合项目类别的项目关联度度量方法

在推荐系统中，项目关联度主要表现在两个方面：项目本身特征信息的相似性所带来的关联度和因相同用户喜好所带来的潜在关联度。基于特征信息的项目关联度的度量方法普遍应用在基于内容的推荐系统中，它提取项目的基本特征(如项目类别、项目名称等)构成项目特征向量，利用特征向量之间的相似性来度量项目之间的关联度。潜在的项目关联度主要是通过基于协同过滤的项目相似度度量方法计算得到，它的关联度主要表现为一个项目集合被同一个用户所喜欢，这个集合内的所有项目之间也具有一定的关联度。本节首先简单概述一些经典的基于协同过滤的项目相似性度量的方法，并阐述了这类度量方法的缺陷；然后结合项目-类别关联矩阵和用户-项目的评分矩阵提出一种新的项目关联度度量方法。

2.1 传统的项目相似度度量方法及缺陷

传统的项目相似度度量主要是从用户-项目的评分矩阵出发，利用协同过滤的思想来度量项目的相似度，主要的方法有余弦相似性、相关相似性、修正的余弦相似性和条件概率等[1]。

1. 余弦相似性(cosine-based similarity)

将项目 i 和项目 j 作为 m 维用户空间中的两个矢量，项目之间的相似程度用这两个矢量之间的夹角余弦来衡量。设项目 i 和项目 j 在 m 维用户空间上的评分分别表示为向量 a,b，则项目 i 和项目 j 之间的相似性为

$$\mathrm{sim}(i,j)=\cos(a,b)=\frac{a\times b}{\|a\|\times\|b\|}$$

式中，分子为两个项目评分矢量的内积，分母为两个矢量模的乘积，夹角越小，相似度越高。

2. 相关相似性(correlation-based similarity)

集合 U 表示同时对项目 i 和项目 j 评分过的用户，将两个项目 i 和 j 之间的 Pearson 系数作为它们之间的相似系数。

$$\mathrm{sim}(i,j)=\frac{\sum_{u\in U}(r_{u,i}-\overline{r_i})(r_{u,j}-\overline{r_j})}{\sqrt{\sum_{u\in U}(r_{u,i}-\overline{r_i})^2}\sqrt{\sum_{u\in U}(r_{u,j}-\overline{r_j})^2}}$$

式中，$r_{u,i}$ 表示用户 u 对项目 i 的评分，$\overline{r_i}$ 和 $\overline{r_j}$ 分别表示集合 U 内用户对项目 i 和项目 j 评分的均值。

3. 修正的余弦相似性(adjusted cosine-based similarity)

余弦相似性有一个严重的缺陷，即它没有考虑不同用户评分的尺度的不同。修正的余弦法通过从评分中减去相应用户的评分均值来克服这个缺陷，即

$$\mathrm{sim}(i,j)=\frac{\sum_{u\in U}(r_{u,i}-\overline{r_u})(r_{u,j}-\overline{r_u})}{\sqrt{\sum_{u\in U}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{u\in U}(r_{u,j}-\overline{r_u})^2}}$$

式中，$\overline{r_u}$ 是用户 $u\in U$ 对全部项目评分的平均值。

基于用户-项目评分矩阵的协同过滤推荐算法度量项目之间的关联度。最近邻居的选择是否合理直接影响推荐的准确率，然而随着电子商务系统规模扩大，用户数目和项目数据急剧增加，导致用户评分数据呈现极端稀疏性；同时对两个不同项目评分过的用户数很少，从而造成传统的相似性度量方法得到的目标用户的最近邻居不准确，算法质量低下，特别是对于从未被评分的项目，无法利用这种方法进行度量它与其它项目的相似性[4]。

此外，上述的项目关联度度量算法忽略了项目之间本身存在的特征相似性所带来的关联度。以项目类别关系为例，同属于某一类别的项目之间应该有更高的关联度。如果项目 i 和项目 j 属于共同类别的数目越多，这两者就应该有越高的关联度，这种考虑在数据比较稀疏的情况下会有更现实的意义[4,5]。

2.2 结合项目类别的项目关联度度量方法

结合项目类别的项目关联度度量方法的主要思路为：首先根据用户-项目评分矩阵，利用协同过滤的算法计算项目之间的潜在关联度，并利用本文提出的关联度收缩方法对由于评分数据稀疏所带来的缺陷进行修正；然后再从项目-类别的关联矩阵的角度计算特征关联度；最后利用权值把两种关联度进行结合，得到综合的项目关联度。

2.2.1 项目的潜在关联度的修正

在项目评分数据较多的情况下，传统的项目关联度度量方法一般都取得不错的度量效果。但在实际应用中，项目和用户评分数据往往极度稀疏，同时对两个不同项目评分过的用户数很少，此时传统的项目关联度度量方法不能准确地反映数据稀疏对关联度度量的影响。例如，当 $\mathrm{sim}(i,j)$ 一定时，如果项目 i 和项目 j 的共同评分的用户很少，则在不考虑项目特征相似性的情况下，可以认为两者之间的关联度也相对较小，此时利用 $\mathrm{sim}(i,j)$ 作为项目 i 和项目 j 的关联度是不合理的。

针对项目评分数据稀疏性问题，本文通过引入收缩参数[6]对项目的潜在关联度进行修正。

$$\mathrm{sim}_a(i,j)=\frac{|U_i\cap U_j|}{|U_i\cap U_j|+a}\mathrm{sim}(i,j)$$

式中，$\mathrm{sim}(i,j)$ 和 $sim_a(i,j)$ 分别为修正前后的项目 i 和项目 j 的关联度，$|U_i\cap U_j|$ 为项目 i 和项目 j 共同评分的用户个数，a 为修正参数。通过参数 a 的设定，可以控制项目 i 和项目 j 共同评分的用户个数 $|U_i\cap U_j|$ 对关联度的影响。如果 $|U_i\cap U_j|$ 远远大于 a，此时 $\frac{|U_i\cap U_j|}{|U_i\cap U_j|+a}\approx 1$，即通过基于协同过滤计算而来的项目潜在的关联度是可信的。如果 $|U_i\cap U_j|$ 与 a 相差不大或远远小于 a，此时 $0<\frac{|U_i\cap U_j|}{|U_i\cap U_j|+a}<1$，可实现对 $\mathrm{sim}(i,j)$ 的收缩，减小两者的关联度。因此该收缩方法能根

据数据的稀疏状况自适应地调节关联度的大小，更准确地反映在数据稀疏条件下项目间的潜在关联度。

2.2.2 结合项目类别的项目关联度度量方法

在实际的推荐系统中，所有项目都可以被划分到不同的项目类别中。例如，Movie Lens 数据中每一个项目(Movie)可以根据电影的类别(Genres)划分为动作片(Action)、战争片(War)、儿童片(Children's)等 18 个类别。并且一部电影可以属于一个或者多个类别，比如一部电影可以同时属于动作片(Action)和战争片(War)。直观上看，如果两个项目属于相同的类别越多，它们之间的关联度就越大。

设集合 C 为推荐系统所有的类别集合 $C=\{c_1,c_2,\cdots,c_k\}$，k 为项目类别的个数。对于项目 i 来说，项目 i 所属于的项目集合可以表示为位向量 $SC_i=\{sc_{i1},sc_{i2},\cdots,sc_{ik}\}$，如果项目 i 属于类别 h，$sc_{ih}=1$；否则 $sc_{ih}=0$，此时整个项目空间所有的位向量构成一个项目类别位图。对于项目 i 和项目 j，可以通过两位图变量的 Pearson 相关系数定义其基于项目类别的特征关联度 $\mathrm{sim}_c(i,j)$ 为：

$$\mathrm{sim}_c(i,j)=\frac{\sum_{h=1}^{K}(sc_{ih}-\overline{sc_i})(sc_{jh}-\overline{sc_j})}{\sqrt{\sum_{h=1}^{K}(sc_{ih}-\overline{sc_i})^2}\sqrt{\sum_{h=1}^{K}(sc_{jh}-\overline{sc_j})^2}}$$

对于 m 维的项目空间 I 中的任意的项目 i 和项目 j，通过基于协同过滤方法和修正得到它的潜在关联度 $\mathrm{sim}_a(i,j)$，并通过项目类别得到它们的特征关联度 $\mathrm{sim}_c(i,j)$。此时通过加权组合，可以得到项目 i 和项目 j 的综合关联度 $\mathrm{sim}_s(i,j)$，即

$$\mathrm{sim}_s(i,j)=(1-\beta)\mathrm{sim}_c(i,j)+\beta\mathrm{sim}_a(i,j)$$

式中，β 为权值系数，$\beta\in[0,1]$。

3 基于加权两层图的推荐算法

传统的推荐系统主要是从用户或项目的共同关联的链接节点个数出发，度量两者之间的关联度，从而为用户推荐其可能感兴趣的用户或者项目，这种度量只考虑两个用户或者项目之间的局部结构。然而对于一个实际的推荐系统来说，用户和用户、项目和项目、用户和项目之间的关联性是可以传递的。即需要从用户项目的全局结构出发，获取用户项目之间的全局关联性。文献[9]提出一种基于用户-项目的二分图的推荐模型，该模型从二分图的全局角度来度量两节点之间的全局关联性，但是该模型没有考虑用户节点或项目节点内部的关联性，并且对用户和项目之间的关联只考虑 0，1 关联。文献[6]提出一种最优化的全局邻居模型，该模型利用历史评分数据，通过全局最优化方法来计算两个节点之间的关联性。该方法过于依赖评分数据，并且同样没有考虑用户或项目本身的关联性。因此本文充分利用用户的历史评分信息和项目类别关联信息，结合用户-项目、项目-项目之间的关联性构建一个基于用户-项目的图模型，利用随机游走算法获取全局关联性，并通过实验证明该方法的有效性。

3.1 加权两层图模型(Weighted Two-layer Graph Model)

设 $G=\{V,E,W\}$ 为一个加权混合图，其中 V 表示为顶点集合，顶点包括两种类型顶点：V_{user} 和 V_{item}，V_{user} 为用户顶点集合，V_{item} 为项目顶点集合，即 $V=V_{user}\cup V_{item}$。E 表示为边集合，包含两种类型的边集合：E_{II} 和 E_{UI}，E_{II} 为项目之间的关联边，E_{UI} 为用户项目之间的关联边，即 $E=E_{II}\cup E_{UI}$。W 为边权值集合。为了能够清晰表示用户-项目图的构造，本文采用了由 Zan Huang 等人提出的一种基于概念语义空间可视化的两层图(Two-layer Graph Model)模型[3]，如图 1 所示，其中上面一层为用户层(User Layer)，下面一层为项目层(Item Layer)。

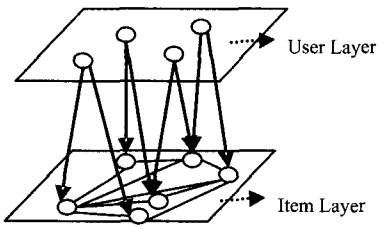

图 1 用户-项目两层图模型

边集合 E_{UI} 反映了用户节点和项目节点之间的关联。设集合 W_{UI} 为两节点之间的关联度集合，对于用户和项目节点来说，两者之间的关联度主要表现用户对项目的喜爱程度，即用户对项目的评分，如果评分越高，那么关联度就越大，设 w_{ui} 表示用户 u 节点和项目 i 之间的关联度，则有 $w_{ui}=r_{ui}$。此时存在一个问题，即原始评分数据存在着用户对电影的评分数据本身所体现出来的各种趋势性数据，例如有些用户对所有电影的评分都偏高，他(她)们可能以 4 分为基准，对不喜欢的电影给予 3 分的评分，而对喜欢的电影给予 5 分的评分；而另外一些用户可能对所有电影的评分都偏低，这些用户可能以 2 分为基准，对自己喜欢的电影只给予 3 分的评分。同样，有些电影较其他电影偏向于获得较高的评分(可能是电影的剧情、画面、音乐、演员或上映时间等因素造成的)。我们把这些趋势性数据统称为全局效应(Global Effects，GEs)[6]。因此我们必须从原始评分数据中剔除用户和项目的全局效应，从而得到用户和项目之间的真实的喜爱程度或关联程度。用户和项目评分数据全局效应很多，比如用户和项目本身基准评分、时间季节对项目和用户的评分的影响等，其中基准评分影响最大。本文仅考虑用户和项目的基准评分两种全局效应。设用户 u 和项目 i 的基准评分全局效应分别为 GE_u 节点和项目 GE_i，则有

$$GE_u=\frac{\sum_{(u,i)\in\rho}r_{ui}}{k(u)},GE_i=\frac{\sum_{(u,i)\in\rho}r_{ui}}{k(i)}$$

其中，ρ 为整个评分系统的评分集合，$k(u)$ 为用户 u 的评分数目，$k(i)$ 为项目 i 的评分次数。此时用户 u 节点和项目 i 之间的关联度 w_{ui} 可以表示为 $w_{ui}=r_{ui}-GE_u-GE_i$。最后采用线性归一化方法对 w_{ui} 进行归一化处理，即

$$w_{ui}=\frac{w_{ui}-w_{\min}}{w_{\max}-w_{\min}}$$

边集合 E_{II} 反映了项目节点之间的关联。设集合 W_{II} 为两节点之间的关联度集合，对于项目节点来说，两项目节点之间的关联度主要表现项目与项目之间的相似度，如果相似度越高，那么关联度就越大。设 w_{ij} 表示项目 i 节点和项目 j 之间的关联度，此时可以利用上一节的综合关联度 $\mathrm{sim}_s(i,j)$ 来表示两者之间的关联度，即 $w_{ij}=\mathrm{sim}_s(i,j)$。

3.2 基于加权两层图的随机游走算法

随机游走(random walk)是一种不规则的运动形式[8]，在运动过程中，每一步的转移都是随机的且不依赖于前面所做的运动。在用户-项目两层图模型的结构下，用户和项目之间的关联过程可以看作一个随机过程$\{X_n\}$，其状态空间是关联

图中的节点。设当前停留在节点 i(用户或项目节点),i 节点与其他节点之间的关联的概念是随机的,如果设下一步与节点 j 关联的概率为 $P_{i,j}$,其只与最近游走的 m 个节点有关,那么$\{X_n\}$被称作 m 阶马尔可夫链。当 $m=1$ 时,下一步的状态值 X_{n+1} 只与当前状态值 X_n 有关,那么

$$P_{i,j}=P\{X_{n+1}|X_n=i\}$$

式中,$\{X_n\}$被称作 1 阶马尔可夫链,简称马尔可夫链条;$P_{i,j}$ 表示状态值 i 到状态值 j 的一步转移概率。

马尔可夫链主要是用来根据当前的状态和当前的一步概率矩阵来预测下一步的状态。假设随机游走模型是平衡的马尔可夫过程,那么随机游走最后可以得到一个平衡的状态,这个平衡状态为两个节点的全局关联度。本文主要利用随机游走算法,从一个用户节点出发,得到系统中其他所有节点(包括用户节点和项目节点)与该用户节点的关联度,最后可以根据关联度大小选取与该用户关联度最大的 N 个用户或者项目,即 TOP-N 用户推荐和项目推荐[9]。

在加权图 G 随机游走模型中,P_{ij} 表示节点 i 与节点 j 之间的直接转移概率,此时 P_{ij} 可以利用加权图中两点之间的权值 w_{ij} 表示,即

$$P_{ij}=\frac{w_{ij}}{\sum_{k=1}^{n_i} w_{ik}}$$

$P_{ij}(k)$表示从节点 i 通过节点 k 随机游走到节点 j 的概率,即两步转移概率,它可以表示从节点 i 直接转移到节点 k 的概率 P_{ik} 与从节点 k 直接转移到节点 j 的概率 P_{kj} 之积,即 $P_{ij}(k)=P_{ik}\times P_{kj}$。同理,可以利用 $P_{i\to j}^{t}$ 表示通过 t 步从节点 i 转移节点 j 的转移概率,即 $P_{i\to j}^{t}=P_{i\to k}^{t-1}\times P_{kj}$。

通过 t 步随机游走以后,图中任意两节点之间的关联度将达到一个稳定平衡状态。此时可以利用从节点 i 随机游走到点 j 的转移概率来表示两节点之间的全局关联度,即 $RD(i,j)=\sum_{k=1}^{t}\delta^{k}P_{i\to j}^{k}$,其中 δ 用于平衡每步转移概率的权重,δ 值越大,邻居信息对最终的关联度描述的重要性越大;t 为随机游走的步数,t 值越大说明考虑的邻居信息越多[10]。

3.3 基于加权两层图的推荐算法

通过综合项目的类别信息,结合用户-项目评分矩阵,可以得到一个基于用户-项目的两层图模型;利用随机游走算法,可以得到两层图中任意两节点之间的全局结构关联度,该关联度从综合全局角度反映了两节点之间的关系。下一步根据随机游走的起始节点和终止节点的不同,可以得到两种不同的推荐算法。

1. 如果从用户节点 u 出发,以项目节点 i 为终止节点,可以得到用户节点和所有的项目节点之间的全局关联度,利用关联度的大小进行排序,向用户进行 TOP-N 推荐[9]。

2. 如果从用户节点 u 出发,以另一用户节点 i 为终止节点,可以得到用户节点与其他用户节点之间的全局关联度或相似度,利用传统 User-based 推荐方法进行项目评分和项目推荐[11]。

4 实验与分析

4.1 数据集

本文的实验采用 MovieLens 站点提供的数据集。MovieLens 是一个基于 Web 的研究型推荐系统,用于接收用户对电影的评分并提供相应的电影推荐列表。目前,该 Web 站点的用户已经超过 43000 人,可供用户评分的电影超过3500部,其评分尺度是从 1 到 5 的整数,数值越高,表明用户对该电影的偏爱程度越高。实验中我们用到该数据集中的一百万条评分数据,包括 6000 个用户和 3900 部电影,其中每个用户至少对 20 部电影进行了评分。在实验中还用到描述电影(项目)的类别文件,即每一个电影属于哪一个或哪几个类别,共有 18 个不同的电影(项目)类别。

4.2 评价标准

评价推荐系统质量的度量标准主要包括统计精度度量方法和决策支持精度度量方法[12],常见的指标有:准确度、覆盖度等。本文主要从准确度(平均绝对误差(MAE)或者均方误差(RMSE))的角度分析本文算法,主要采用平均绝对误差(MAE)来度量准确度:

$$MAE=\frac{\sum_{(u,i)\in\Gamma}|\hat{r}_{ui}-r_{ui}|}{|\Gamma|}$$

式中,Γ 是测试数据集,$\hat{r}_{ui}$ 是算法预测的用户 u 对物品 i 的打分,r_{ui} 是数据集中用户 u 对 i 的实际打分。MAE 越小,误差越小,算法准确度越高。

4.3 实验一:项目潜在关联度的修正的验证及其参数调整

本文针对利用用户-项目的评分矩阵计算项目之间的关联度受到两项目共同评分用户个数的限制的问题,提出了一种关联度修正的方法,其通过引入参数 a 的设定,可以控制项目 i 和项目 j 共同评分的用户个数 $|U_i\cap U_j|$ 对关联度的影响。实验采用相关相似性(Correlation-based Similarity)来度量两项目关联度,利用传统的 Item-based 协同过滤算法[1]对本文提出的关联度修正算法进行验证,并选取最优的参数 a,实验结果如图 2 所示。

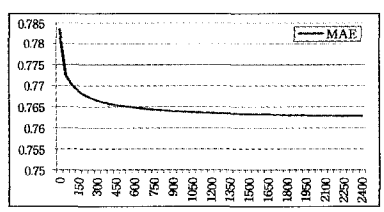

图 2 项目潜在关联度的修正

如图 2 可以看出,在参数 $a=0$ 时,不对关联度进行修正,此时平均误差 MAE≈0.783。通过参数 a 的递增,MAE 极速下降,表明了关联度修正的有效性。在参数 a 的大小超过一定值时,MAE 基本保持不变,当参数 $a=1900$ 时,MAE≈0.763。因此从上面的实验结果可以证明项目潜在关联度修正的有效性。

4.4 实验二:结合项目类别的项目关联度度量方法的验证及其参数调整

本文结合项目类别计算出项目之间的特征关联度,然后通过引入权值参数 β,将基于协同过滤方法和修正得到的潜在关联度 $sim_a(i,j)$,以及通过项目类别得到的特征关联度 $sim_c(i,j)$进行加权组合。本实验是在实验一的最优化的参数 $a=1900$ 的基础上,通过改变参数 β,得到最优化的混合关联度和最高的推荐精度。实验结果如图 3 所示。

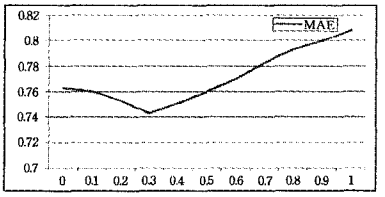

图 3 混合关联度及其参数的选择

从图 3 可以看出，在参数 $\beta=0$ 时，不利用基于项目类别的特征相似度，此时平均误差为实验一最优化的实验结果，即 MAE≈0.763。随着参数 β 的递增，MAE 有所下降，证明了关联度组合的有效性，在参数 β 的大小超过一定值时，此时 MAE 快速上升。当参数 $\beta=1$ 时，完全采用项目类别特征来计算项目关联度，此时 MAE≈0.807。当参数 $\beta=0.3$ 时，MAE≈0.743 为最优化的组合参数。因此，上面的实验结果可以证明结合项目类别的混合关联度计算的有效性。

4.5 实验三：随机游走的步数的选择

随机游走算法的游走步数反映了两节点之间的关联度的深度，由于随机游走是一个平衡的马尔可夫过程，因此在随机游走的步数达到一定的时候，整个模型达到一个平衡。本实验通过迭代游走的步数 t，得到整个模型达到平衡时随机游走的步数。实验通过固定最优化参数 $a=1900$，$\beta=0.3$，构建一个基于用户-项目的两层图模型，其中随机游走每步权重选择一个中间值 $\delta=0.5$。利用 3.3 节提出的推荐算法 1，构建一个推荐系统。实验结果如图 4 所示。

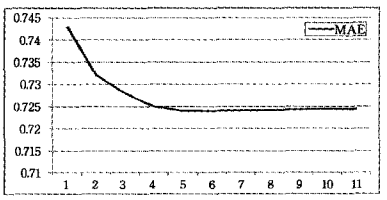

图 4 随机游走的步数选择

从实验结果图 4 可得，$t=1$ 时，只游走一步，表示直接用两个节点之间的直接转移概率来度量两节点之间的关联度（采用传统的 Item-based 协同过滤算法[1]），此时 MAE≈0.743。从图可以看到，$t>5$ 以后，MAE≈0.725，即随着游走的步数的增加，推荐精度基本保持不变。

4.6 实验四：随机游走每步权重参数选择

在随机游走算法中，不同的游走步数所产生的两点之间的关联度的权重不同，游走步数越少，产生的关联度的权重就应该越大。本文引入参数 δ 对游走所产生的关联度进行修正和组合，得到随机游走最终的关联度，其中 $0<\delta\leqslant 1$。如果 $0<\delta<1$，此时相当于对节点之间的关联度进行收缩处理，其中，收缩程度和游走步数成正比关系。当 $\delta=1$ 时，即不进行收缩处理，等价对待每步游走所产生的关联度。本实验通过固定最优化参数 $a=1900$，$\beta=0.3$，构建一个基于用户-项目的两层图模型，随机游走的步数选为 $t=5$。利用 3.3 节提出的推荐算法 1，构建一个推荐系统。实验结果如图 5 所示。

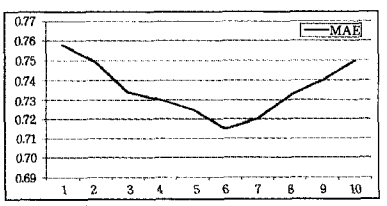

图 5 随机游走每步权重参数选择

从实验结果图 5 可得，当 $0<\delta<0.6$ 时，随着 δ 的增大，MAE 下降。当 $\delta>0.6$ 时，随着 δ 的增大，MAE 变大，此时收缩程度不足。当 $\delta=0.6$ 时，MAE≈0.715 为整个随机游走算法中最低的 MAE。

4.7 实验五：基于随机游走的混合推荐算法的有效性

通过上述 4 个实验，我们得到一个结合项目类别信息的最优两层图模型，模型的最优化参数如下：$a=1900$，$\beta=0.3$，$t=5$，$\delta=0.6$。本节将基于随机游走的传统的两分图算法[9]（算法 1）、基于传统 Item-Based 的协同过滤算法[1]（算法 2）、基于传统 User-Based 的协同过滤算法[11]（算法 3）、基于 Slope-One 的协同过滤算法[13]（算法 4），以及由 Bell & Koren 等人提出的基于全局邻居模型的协同过滤模型算法[6]（算法 5）同本文的基于项目类别信息的混合推荐算法进行比较。其中，算法 2—算法 4 为经典的协同过滤算法，都只仅仅考虑项目-项目之间或者用户-用户之间的局部相关性。而算法 1、算法 5，以及本文的算法都是从用户项目之间的全局角度出发来考虑用户项目之间的关联性，其中算法 1 从用户-项目评分矩阵出发，利用随机游走算法来得到用户项目之间的全局关联度。算法 5 也同样从用户-项目评分矩阵出发，利用全局最优化的方法来考虑项目之间的关联性。算法 1 和算法 5 都仅仅从用户-项目评分矩阵出发，没有考虑项目本身由于项目类别等因素带来的关联性。本文对上述的 5 个算法，通过调整每一个算法的相关参数，得到每一个算法的平均绝对误差。实验结果如图 6 所示。

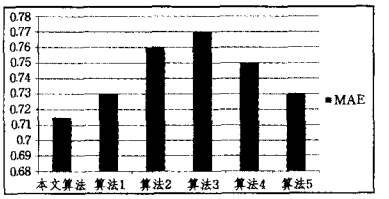

图 6 几种算法的平均绝对误差比较结果

由实验结果可以得到以下结论：

1. 本文算法、算法 1、算法 5 所得到的推荐精度明显高于算法 2、算法 3、算法 4，这验证了从全局角度出发来考虑用户项目之间的关联性的有效性。

2. 本文算法和算法 1 都是从用户项目之间的图模型出发，利用随机游走算法得到用户项目之间的全局关联性，但是本文算法从项目类别信息出发考虑项目-项目之间的关联性，把算法 1 的二分图模型提升为本文提出的两层图模型。从实验室结果可以看出，这种处理方法更有效。

结束语 针对传统的协同过滤算法仅仅考虑用户项目之间的局部关联关系，本文提出一种新的协同过滤推荐模型。该模型充分利用用户-项目评分矩阵以及项目-类别关联数据，通过参数收缩、关联度组合等方法，得到项目-项目之间的混合关联度，进而构建一个基于用户项目的两层图模型；利用随机游走算法，从全局的角度来考虑用户项目之间的关联关系。通过实验证明，这个模型是可行的。

另外，本文提出的两层图模型具有很强的可扩展性，下一步可以从用户的特征信息出发考虑用户本身之间的关联性，扩展两层图的结构，进而提高模型的推荐精度。

参 考 文 献

[1] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative

filtering recommendation algorithms[C]// Proceedings of the 10th international conference on World Wide Web. New York, N. Y., USA: ACM, 2001: 285-295

[2] Pazzani M J. A framework for collaborative, content-based, and demographic filtering [J]. Artificial Intelligence Review, 1999, 13(5/6): 393-400

[3] Huang Zan, Chung Wing-yan. A graph model for e-commerce recommender systems [J]. J. ASIST, 2003, 55(3): 259-274

[4] Zhang Guang-wei, Kang Jian-chu. Context based collaborative filtering recommendation algorithm [J]. Journal of system Simulation, 2006, 18(S1): 595-602

[5] 姚忠，吴跃，常娜. 集成项目类别与语境信息的协同过滤推荐算法[J]. 计算机集成制造系统，2008，14(7)：1449-1455

[6] Bell R M, Koren Y. Improved neighborhood-based collaborative filtering[C]// KDDCup'07. San Jose, California, USA, August 2007

[7] Bell R, Koren Y, Volinsky C. The BellKor Solution to the Netflix Prize[R]. 2009

[8] Baluja S, Seth R. Video Suggestion and Discovery for YouTube: Taking Random Walks through the View Graph[C]// Proc. 17th Int'l World Wide Web Conf. 2008

[9] Fouss F, Pirotte A, Renders J-M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation [J]. IEEE Trans. Knowl. Data Eng., 2007, 19: 355-369

[10] 周军军，王明文，何世柱. 基于随机游走和聚类平滑的协同过滤推荐算法[J]. 广西师范大学学报：自然科学版，2011，29(1)：173-178

[11] Sarwar B M, Karypis G, Konstan J A. Analysis of Recommendation Algorithms for E-Commerce [J]. Proceedings of the ACM EC'00 Conference, 2000, 40(3): 158- 167

[12] Ha V, Haddawy P. Toward Case-Based Preference Elicitation: Similarity Measures on Preference Structures[C]// Proceedings of 14th Conference on Uncertainty in Artificial Intelligence. 1998: 193-201

[13] Lemire D, Maclachlan A. Slope One Predictors for Online Rating-Based Collaborative Filtering[C]// Proceedings of SIAM Data Mining (SDM'05). 2005

(上接第 161 页)

(3)索引树更新开销

在两种情况下需要对索引顺序树进行更新操作。第一，当文件拥有者添加文件到 CSP 时，CSP 只需将该文件所对应的关键词集合插入到索引顺序树中。找到插入位置的平均时间为 $O(\log_2 n)$(其中 n 为索引树结点的个数，即关键词的个数)。第二，当文件拥有者向 CSP 申请删除某文件时，CSP 除了将文件库中的对应文件删除之外，只需将该文件 *ID* 从索引顺序树中与该文件相关联的所有关键词所对应的结点的文件 *ID* 链表中删除。查找删除结点位置的平均时间也为 $O(\log_2 n)$。由此可见，更新索引树也非常高效。

结束语 随着用户要求的提高，单关键词查找已经不能满足用户的需求。本文针对云计算环境的特点以及安全性，提出了一个适用于云计算平台的高效、安全的多关键词查找方案，即 PPMKS 方案。该方案在保证用户数据隐私的前提下，能够实现授权用户多关键词查找云服务器端的加密文件；此外，还引入了与或逻辑表达式，满足用户自定义查找的各种需求。在该结构下，用户删除、新增文件都相当便捷。PPMKS 方案充分利用了云服务器的特性，既保证了云数据的安全和隐私，又实现了高效地查找云端密文数据。

参 考 文 献

[1] 高巍. 2010 年中国通信产业十大关键词点评文章(三)—云计算[J]. 数据通信，2011(1)：5-6

[2] 李乔，郑啸. 云计算现状综述[J]. 计算机科学，2011(4)：32-37

[3] 史美林，姜进磊，孙瑞志. 云计算[M]. 向勇，译. 北京：机械工业出版社，2009

[4] Li C, Lu J, Lu Y. Efficient merging and filtering algorithms for approximate string searches[C]// Proceedings of ICDE. 2008

[5] Behm A, Ji S, Li C, et al. Space-constrained gram-based indexing for efficient approximate string search[C]// Proceedings of ICDE. 2009

[6] Ji S, Li G, Li C, et al. Efficient interactive fuzzy keyword search [C]// Proceedings of ACM WWW. 2009

[7] Bellare M, Boldyreva A, Neil A O. Deterministic and efficiently searchable encryption[C]// Proceedings of Crypto 2007. LCNS, Vol 4622, 2007

[8] Song D, Wagner D, Perrig A. Practical techniques for searches on encrypted data[C]// Proceedings of IEEE Security and Privacy. 2000

[9] Goh E-J. Secure indexes[R]. 2003

[10] Boneh D, Crescenzo G D, Ostrovsky R, et al. Public key encryption with keyword search[C]// Proceedings of EUROCRYP. 2004

[11] Waters B, Balfanz D, Durfee G, et al. Building an encrypted and searchable audit log[C]// Proceeding of 11th Annual network and Distributed System. 2004

[12] Cutmola R, Garay J A, Kamara S, et al. Searchable symmetric encryption: improved definition and efficient constructions[C]// Proceedings of ACM CCS. 2006

[13] Li J, et al. Fuzzy Search over Encrypted Data in Cloud Computing[C]// Proceedings of IEEE INFOCOM. 2010

[14] Chang Y C, Mitzenmacher M. Privacy preserving keyword searches on remote encrypted data[C]// Proceedings of ACNS. 2005

[15] Shi E, Bethencourt J, Chan T H H, et al. Multidimensional range query over encrypted data[C]// IEEE Symposium on Security and Privacy. 2007

[16] Cao N, et al. Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data[C]// Proceedings IEEE INFOCOM. 2011

[17] Chuah M, Hu W. Privacy-aware Bed Tree Based Solution for Fuzzy Multi-keyword Search over Encrypted Data[C]// Proceedings of ICDCS Workshops. 2011: 273-281

[18] Zerr S, et al. Zerber: r-confidential indexing for distributed documents[C]// Proceedings of EDBT. 2008: 287-298