基于领域本体的协同过滤推荐算法

A Collaborative Filtering Algorithm Using Domain Ontology

龚松杰　潘红艳　（浙江工商职业技术学院 信息工程学院 浙江宁波 315012）

摘　要： 为了解决协同过滤推荐系统中所存在的可扩展性、稀疏性和冷启动等问题带来的推荐性能底下，提出新的基于领域本体的协同过滤推荐算法，该算法综合考虑了项目的语义相似性和评分相似性的影响，改善基于项目的协同过滤算法性能。实验结果表明，基于领域本体的协同过滤算法不仅能很好的解决基于项目的协同过滤算法带来的问题，而且还提高了推荐系统的推荐质量。

关键词： 协同过滤 领域本体 语义相似性 评分相似性 稀疏性

随着互联网的快速发展和网上信息的增加，如何准确地获取信息已成为焦点。个性化服务是推荐系统根据用户访问网站或购买商品的历史中表现出的行为，动态调节站点的内容，自动适应用户的动作和向用户推荐可能感兴趣的内容的过程[1]。个性化推荐系统可使用户快速、准确地得到所需信息。在个性化推荐中，协同过滤是当前应用最成功的技术。

在基于项目的协同过滤推荐中，为了对目标用户产生推荐，需要搜索最近邻居，在此过程中，定义项目之间的相似性成为关键问题之一。目前有余弦相似性，相关相似性，最小平方差等。随着电子商务规模的扩大，用于产生推荐的数据将极端稀疏，使用上述度量的产生推荐效果将逐步减弱[2,3]。这样就导致项目之间的相似性不准确，产生的最邻近的邻居项目不可靠，对于新启动的项目，根本无力产生推荐。

针对传统相似性度量不足和新项目的冷启动问题，提出了一种新的基于领域本体的协同过滤算法，该方法将基于用户评分计算项目的相似性与语义相似性互相组合，从而在一定程度上缓解了数据稀疏性带来的问题，克服相似不相同问题，提高系统的推荐生成质量。实验也表明，基于领域本体的系统过滤算法优于传统的基于项目的协同过滤算法。

1　基于项目的协同过滤及相似性计算

1.1　问题描述

协同过滤推荐的出发点是任何人的兴趣都不是孤立的，应处于某个群体所关心的兴趣中。它基于这样一个假设 如果用户对一些项目的评分比较相似，则他们对其它项目的评分也比较相似。基本思想是采用某种技术找到目标用户的最近邻居，然后根据最近邻居对目标项目的评分，产生推荐。

用户评分数据可以用一个 m * n 阶矩阵 A(m * n) 来表示，m 行代表 m 个用户，n 列代表 n 个项目，第 i 行第 j 列元素 Ri,j 代表第 i 个用户对项目 j 的评分。如表 1 所示。

表 1　用户—项目矩阵

Item / User	Item1	Item2	… …	Itemn
User1	R1,1	R1,2	… …	R1,n
User2	R2,1	R2,2	… …	R2,n
… …	… …	… …	… …	… …
Userm	Rm,1	Rm,2	… …	Rm,n

1.2　传统的三种相似性度量方法

在计算邻居用户对目标用户的影响时，需要查找目标用户的最近邻居，这就需要度量。用户的兴趣可以用评分向量来表示，相当于表 1 中的某一行。有三种传统的度量方法：

余弦相似性 把用户评分看作是 n 维项目空间上的向量。通过计算两个向量之间的夹角余弦来度量两个用户之间的相似性。计算公式如下：

$$sim(i,j)=(\sum_{k=1}^{n}R_{i,k}*R_{j,k})/(\sqrt{\sum_{k=1}^{n}R_{i,k}^{2}*\sum_{k=1}^{n}R_{j,k}^{2}})$$

$R_{i,k}$,$R_{j,k}$:用户 i,j 对项目 k 的评分。

相关相似性 :通过 Peason 相关系数来度量两个用户的相似性。计算时,首先找到两个用户共同评分过的项目集 $I_{i,j}$,然后计算这两个向量的相关系数。计算公式如下:

$$sim(i,j)=\left(\sum_{c\in I_{i,j}}(R_{i,c}-A_i)(R_{j,c}-A_j)\right)/\sqrt{\sum_{c\in I_{i,j}}(R_{i,c}-A_i)^2*\sum_{c\in I_{i,j}}(R_{j,c}-A_j)^2}$$

$I_{i,j}$:用户 i 和 j 共同评分过的项目集,$R_{i,c}$:用户 i 对项目 c 的评分,Ai :用户 i 对资源的平均评分。

修正的余弦相似性 :在余弦相似性中没有考虑不同用户的评分尺度问题。修正的余弦相似性通过减去项目的平均评分来弥补这种不足,计算公式如下:

$$sim(i,j)=\left(\sum_{c\in I_{i,j}}(R_{i,c}-A_c)(R_{j,c}-A_c)\right)/\sqrt{\sum_{c\in I_{i,j}}(R_{i,c}-A_c)^2*\sum_{c\in I_{i,j}}(R_{j,c}-A_c)^2}$$

A_c :项目 c 的平均评分。

1.3 传统的三种相似性度量的不足

当两个用户评分的共同项较少或者根本没有共同的评分项目时,如果采用传统的相似度计算方法,用户间的相似度会很小或为零。这种情况在系统刚建立时最为突出,因为在系统刚建立时,只有很少的用户评价了很少的项目,数据的极端稀疏使得用户共同评分的项目很少,这时用户就无法找到最近邻进行协同推荐。

综上所述,传统的三种相似性度量方法在用户评分数据极端稀疏的情况下并不能有效地度量用户之间的相似性,从而使得计算出来的最近邻居不准确,导致整个推荐系统的推荐质量急剧下降。

2 基于领域本体的协同过滤推荐

考虑到项目的语义性,提出的协同过滤推荐算法建立在领域本体基础之上,该文首先构建汽车领域的本体知识库,然后提取文档项目的特征向量,最后按照此向量进行语义相似性的度量。

2.1 构建领域本体知识库

目前对构造领域本体的方法和方法的性能评估还没有统一的标准,不过在构造特定领域本体的过程中,有一点是得到业界和学术界认可的,那就是需要该领域专家的参与[4,5]。由领域专家和语言学家共同确定该领域的基本词汇和词汇间的关系。领域本体是很庞大的,如果手工构建,工作量是很大的。如果我们能够搜集足够多的领域训练文本,从这些文本中抽取出该领域的基本词汇,再利用某种技术得到这些词汇之间的关系,就可以实现领域本体的构建。这样做在理论上是行得通的,缺点就是大量训练文本的获取存在困难。

具体做法如下:

(1) 对训练文本进行预处理

对文本进行分词,对照停用词表去掉停用词。计算词的权重,并对其进行正规化。正规化公式如下:

$$W_{ik}=f_{ik}/\sum_{i=1}^{nk}f_{ik}$$

n_k 为文档 k 中不同术语的总数。

(2) 构建术语—文档矩阵

术语—文档矩阵如下:

$$W=\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix}$$

每一行表示一个术语在每篇文档中的权重,每一列表示一个文档向量。wik 表示第 i 个术语在第 k 篇文档中的权重。

(3) 对 W 进行 SVD 分解

$$w=u^k*s^k*v^k$$

u^k 为术语 – 概念矩阵,v^k 为文档 – 概念矩阵。

这样我们就可以得到一组概念和描述每一个概念的一组术语,这些概念之间的关系通过聚类来得到。

(4) 概念聚类

每个概念看作是由 m 个术语组成的向量,对其聚类。

采用此方法构建了一个汽车领域数据库,如图 1 所示。

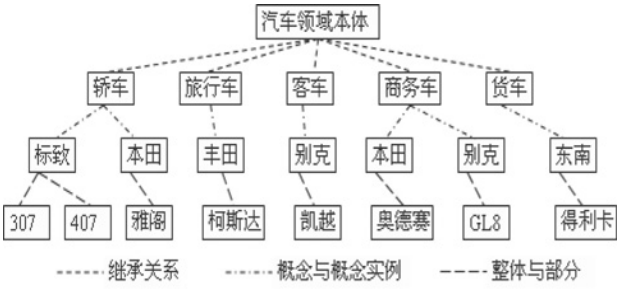

图 1 汽车领域数据库

目前概念之间的关系不能自动获取,是采用手工方法得到的。随着社会的发展,会不断的涌现出新的概念,所以需要对本体不断的更新,本体更新是一项繁重的工作,目前一般采用手工更新,可以采用一种半自动的更新方法。对一个新出现的概念,系统要求用户同时给出表达此概念的术语向量,然后按照上面介绍的构建本体的方法进行概念聚类。

2.2 语义相似性度量

文档项目的特征向量可以利用分类器对大量已标注文本进行训练,进而得到概念的特征向量。目前一般采用这种方法,这样做有一个缺点是工作量大,而且要获得大量已标注文本也是相当困难的。前面已经讲过领域本体是由该领域的基本词汇构成的,那么这些词的某种组合就能表征某一个概念,因此,我们提出了一种新的抽取特征向量的方法,其基本思想是,概念的特征项由概念本身及其所有子结点组成。确定特征项的权重时,需要综合考虑结点之间的距离对分类的作用因子,和结点本身对分类的贡献度。

根据文档的特征向量,就可以计算文档之间的语义相似性,计算公式如下:

$$sim(d_i,d_j) = (\sum_{k=1}^{n} ntd_{ki} * ntd_{kj}) / (\sqrt{\sum_{k=1}^{n} ntd_{kj}^{2} * \sum_{k=1}^{n} ntd_{ki}^{2}})$$

di,dj 是两个文档的特征向量,nt 是正规化权重算子。

2.3 融合语义相似性和评分相似性的推荐

基于领域本体的协同过滤推荐算法,综合考虑两个文档项目之间的语义相似性和评分相似性,计算公式如下:

$$Sim_{ij} = \omega Sim_{ij1} + (1-\omega) Sim_{ij2}$$

Sim_{ij} :混合相似度,Sim_{ij1} :语义相似度,ω :语义相似度的贡献权值,Sim_{ij2} :评分相似度,1 - ω :评分相似度的贡献权值。

计算目标用户 U 对未评分项目 I 的评分时,根据最近邻居对项目 I 的评分产生推荐。采用下式计算 U 对 I 的评分:

$$P_{ui} = (\sum_{j=1}^{c} sim(i,j) * R_{uj}) / \sum_{j=1}^{c} sim(i,j)$$

Ruj :用户 u 对项目 j 的评分, sim(i,j) :项目 i 和 j 的综合相似度。

3 实验及结果分析

设计并实现了一个基于汽车领域本体的个性化推荐系统,并采用系统数据来进行测试。该系统已有 800 个用户对 1500 个文档项目进行了评分,评分值为从 1 到 5 的整数,数值越高,表明用户对该文档的喜爱程度越高。系统中采用的文档是从搜狐(http ://www.sohu.com)的汽车频道上搜集的。内容涵盖产业新闻、车市新闻、国际新闻、最新科技和综合新闻五个版面的内容。

利用平均绝对误差 MAE 来衡量算法的预测精度。MAE 是测试集中所有用户对资源打分的实际值与预测值的偏差的绝对值的平均[6]。MAE 较早的在 Shardanand & Mases 及 Sarwar 等中用于评价系统预测的性能。MAE 值越小说明推荐算法的预测精度越高。在以下介绍的精度试验中,我们采用了 MAE 来衡量算法的预测精度。

考虑混合相似度计算中的 ω 权重参数对 MAE 的影响,在使用混合相似度算法预测时,找出最优 ω 值的范围,权重参数 ω 从 0 变化到 1,间隔为 0.1。实验的结果如图 2 所示。由此得出结论,α 的取值范围在 0.4 附近是最优的。

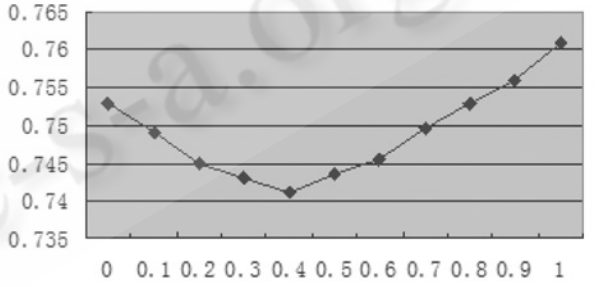

图 2　权重参数 ω 对推荐精度的影响

为检验基于领域本体的协同过滤算法的性能,该文将传统的基于项目的协同过滤算法与基于混合相似度计算的协同过滤算法进行比较,权重参数 ω 取 0.4。最近邻居数 N 从 10 增加到 100,间隔为 10,实验结果如图 3 所示。

从实验结果分析,通过语义相似性与评分相似性的融合,挖掘出项目之间的语义关系,抽取出项目的语义信息,不仅能很好的解决传统的基于项目的协同过滤算法的项目评分的稀疏性问题、新项目冷启动问题以及提高推荐精度,还能进一步解释说明用户对特定

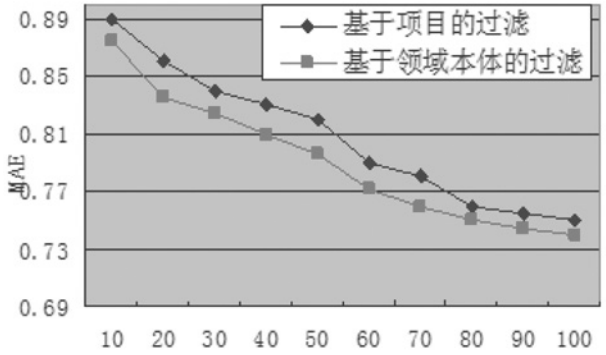

图 3 推荐算法推荐精度比较

领域的项目是否感兴趣。

4 总结

随着用户和项目数目的增加，在整个空间上用户评分数据极端稀疏，传统的相似性度量存在各自的弊端。因此，提出新的基于领域本体的协同过滤算法，综合考虑了项目的语义相似性和评分相似性的影响，改善基于项目的协同过滤算法性能。实验表明，该算法不仅能很好的解决基于项目的协同过滤带来的问题，而且还提高了推荐系统的推荐质量。

参考文献

1 Herlocker J，Konstan J，Terveen L，Riedl J. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems (TOIS)，2004，22(1): 5 – 53.

2 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. 软件学报，2003，14(9): 1621 – 1628.

3 Sarwar B，Karypis G. Item – based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference. 2001: 285 – 295.

4 潘红艳，林鸿飞，赵晶. 基于 Ontology 的个性化推送系统[J]. 计算机工程与应用，2005，20: 176 – 180.

5 Susan Cauch，Jason Chaffe，Alexander Pretschner. Ontology – Based Personalized Search and Browsing [C]. In: Web Intelligence and Agent System，2003: 219 – 234.

6 陈健，印鉴. 基于影响集的协作过滤推荐算法. 软件学报，2007，18(7): 1685 – 1694.