改进的协同过滤推荐算法①

张 亮[1], 赵 娜[2]

[1](中国石油大学(华东)网络及教育技术中心, 青岛 266500)
[2](山东省青岛市黄岛区建筑工程质量监督站, 青岛 266500)

摘 要: 传统的选修课系统存在结构性的不足和缺憾, 为了避免高校学生盲目的选择选修课程, 本文利用改进的协同过滤算法对高校学生进行个性化的选课推荐. 本文首先介绍了两种推荐算法, 并着重介绍基于协同过滤的推荐算法, 并分析了两种算法的优缺点, 最后针对协同过滤算法的数据稀疏性问题, 提出了一种改进的协同过滤算法, 即在协同过滤中加入基于内容的因素来解决这个问题. 这种改进的协同过滤算法避免了传统协同过滤算法中存在的数据稀疏问题, 以学生为本推荐适合学生的课程, 满足学生学习的个性化要求.

关键词: 协同过滤; 相似度; 特征值; 推荐系统; 兴趣

Improved Collaborative Filtering Algorithm

ZHANG Liang[1], ZHAO Na[2]

[1](Network Information Center, China University of Petroleum (East China), Qingdao 266500, China)
[2](Construction Quality Supervision Station,HuangdaoDistrict, Qingdao 266500, China)

Abstract: Traditional elective system has structural deficiencies and defects. To avoid the fact that college students choose a course with blindness, therefore, with improved collaborative filtering algorithm, college students can get personalized elective course election. This paper first introduces two kinds of recommendation algorithms. Also the paper emphatically introduces recommendation algorithms based on collaborative filtering. It analyzes the advantages and disadvantages of the two algorithms. Finally, for data sparsity of collaborative filtering algorithm, it proposes an improved collaborative filtering algorithm, that adds factor in content-based collaborative filtering to solve this problem. Improved collaborative filtering algorithm avoids the traditional algorithms emerging data sparseness problem. Recommending appropriate courses for students on human-oriented, individual needs of students can be met.

Key words: collaborative filtering; similarity; characteristic value; recommended system; interest

随着教育改革的不断深入, 需要拓宽学生的知识, 培养个性化、多元化的创新型人才, 高校选修课程的开设是高校培养模式的有效补充[1-4]. 传统的高校选课系统虽然可以通过搜索功能帮助学生快速找到所要选修课程的信息, 但搜索的排序结果单一, 无法针对不同学生的自身的兴趣爱好提供个性化推荐, 导致学生选修课程带有一定的盲目性. 而推荐系统, 作为一种信息过滤的重要技术和手段, 它实现了学生与信息间的主动交互, 通过对学生选课行为和属性“猜测”学生感兴趣的内容并进行推荐, 被认为是可解决学生个性化选课的有效的工具. 协同过滤算法的研究开始于 20 世纪 90 年代, 是目前运用最广泛的推荐算法. 目前为止, 已经开发了许多协同过滤系统的应用, 例如 GroupLens[5]、Video Recommender[6]和 Ringo[7]被认为是第一批能够进行自动预测的系统过滤系统. 其他系统过滤系统的例子包括亚马逊的图书推荐, 笑话推荐的 Jester 系统[8]等等. 本文提出一种改进的协同过滤算法, 即在协同过滤中加入基于内容的因素来解决高校选修课系统目前存在的数据稀疏问题, 并重点介绍了这种算法的核心内容.

① 收稿时间:2015-11-06;收到修改稿时间:2015-12-10 [doi:10.15888/j.cnki.csa.005224]

1　推荐系统常用算法

1.1 基于内容的推荐算法

当前主流的推荐算法之一就是基于内容的推荐算法[9]-[11], 它首先把每个用户的信息需求表示成一个用户兴趣模型, 然后根据项目与某个用户感兴趣的项目之间的相似程度来预测该项目对该用户的效用, 从而推荐给用户.

大部分基于内容的是在应用文本推荐领域, 在文本领域, 推荐的输入是有用户感兴趣的一组文档和一组用于推荐的数量较大的文档(项目)组成. 而文档的特征集合一般采用向量空间模型VSM来表示, 给定一个文档 $D_v=(w_{1v},w_{2v},...w_{kv})$, w_{iv} 表示第 i 个此在文档 d_v 中的权重. 在向量空间模型中用 TD-IDF 方案, 每个文档被表示成由关键字形成的向量. 之后将用户资料与项目特征向量进行相似度计算, 余弦相似度是最典型的相似度计算标准, 公式如(1)所示.

$$sim(\mathrm{u},\mathrm{v})=\cos(\overrightarrow{w_u},\overrightarrow{w_v})=\frac{w_u\cdot w_v}{\|w_u\|_2\times\|w_v\|_2}=\frac{\sum_{i=1}^{K}w_{i,u}w_{i,v}}{\sqrt{\sum_{i=1}^{K}w_{i,u}^2}\sqrt{\sum_{i=1}^{K}w_{i,v}^2}} \tag{1}$$

1.2 基于协同过滤推荐算法

协同过滤算法[12-15]是根据其他志趣相投的用户过去打分或者购买过的项目来预测项目对某个用户的效用. 该方法通常指利用用户-产品交互数据而忽略用户和产品本身属性. 其利用整个用户-项目数据库直接进行预测, 即没有构建模型. 这种方法包括基于用户和基于项目的两种算法.

1.2.1 基于用户的协同过滤算法

一个典型的基于用户的协同过滤算法由两阶段组成: 邻居形成阶段和推荐阶段. 在第一个阶段, 该算法对目标用户(也称作访问者)的活动记录和其他用户的历史记录 T 进行比较从而找到与目标用户有相似风格或兴趣的前 k 个用户. 访问者记录与其邻居之间的映射可以依据项目评分相似度, 对相似内容或页面的访问或者对相似项目的购买来进行. 在大部分典型的协同过滤应用中, 用户记录或资料是该用户在一个项目子集上的一组评分.

1.2.2 基于项目的协同过滤算法

该方法可以预先计算所有项目对之间的相似度. 基于项目的方法根据用户对项目的评分模式来对项目进行比较. 同样的, 我们会采用 k 近邻方法寻找被不同用户打分相似的 k 个相似项目. 该算法的主要步骤和基于用户的协同过滤算法基本相同, 不同之处, 不是计算用户的 k 个近邻, 而是计算项目的 k 个近邻.

通过上面的介绍, 可以对这两种算法的优缺点如下总结, 具体内容见表 1.

表 1　两种算法的比较

推荐算法	优点	缺点
内容推荐	推荐结果直观，容易解释。 不需要领域知识。	稀疏问题； 新用户问题； 过于专门化； 要有足够数据构造分类器。
协同过滤	新异兴趣发现、不需要领域知识； 随着时间推移性能提高； 推荐个性化、自动化程度高； 能处理复杂的非结构化对象。	稀疏问题； 可靠扩展性问题； 新用户问题； 质量取决于历史数据集； 系统开始时推荐质量差。

2　选修课程推荐系统核心算法设计

本文所研究的内容是在协同过滤中加入内容的因素来进行学生选课. 学生在选修一门新的课程时, 对此课程的评分值等于零, 在协同过滤中加入基于内容的因素, 可用于计算学生之间的相似度, 在一定程度上可以有效缓解传统协同过滤算法所遇到的稀疏问题[16-19]. 这种方法的另一优点在于, 学生所看到的推荐课程不仅仅是被其相似的学生所感兴趣的课程, 而且是与目标学生的档案相似的课程. 选修系统的体系结构图和核心算法的流程图如下所示.

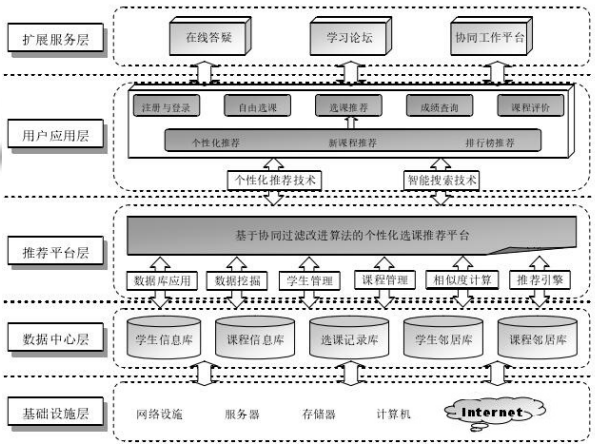

图 1　体系结构图

(1)根据特征的来源不同, 可以将学生的特征分为基本特征和操作特征. 学生的基本特征包括了性别、年龄、专业、爱好等信息. 学生的操作信息浏览课程、申请选课、收藏课程等信息. 本文研究的选课系统, 结合学生和及其特征各自的属性, 生成学生与特征的向

量矩阵, 如表 2 所示, 以便于计算它们间的相似度. 对学生历史选课记录挖掘其兴趣点、学生历史成绩判断其需求、根据学生已选课程进行相似度计算, 如果两个学生操作特征值相同布尔值设为 1, 不同布尔值设为 0.

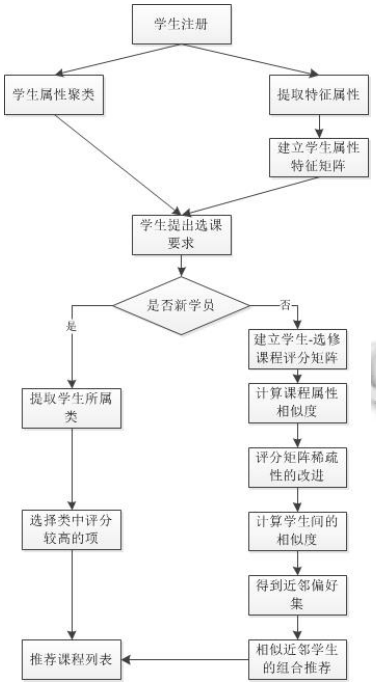

图 2 算法流程图

表 2 学生-特征矩阵

Item / Stu	I_1	I_2	.....	I_n
S_1	$R_{1,1}$	$R_{1,2}$	.....	$R_{1,n}$
S_2	$R_{2,1}$	$R_{2,2}$	.....	$R_{2,n}$
.....	.....	.....	.....	
S_m	$R_{m,1}$	$R_{m,2}$	.....	$R_{m,n}$

(2)学生选修课程的相似度计算

这部分主要是实现实时调整相似度, 主要是通过将基于内容的推荐与协同过滤的预测值进行加权求和来实现. 用下面所示的余弦相似度公式[20]来计算学生选修课程 u 和学生选修课程 v 之间的综合相似度.

$$sim(\mathrm{u},\mathrm{v})=\alpha*sim_{\mathrm{u}}(\mathrm{u},\mathrm{v})+\beta*sim_{\mathrm{DS}}(\mathrm{u},\mathrm{v})+\chi*sim_{\mathrm{JS}}(\mathrm{u},\mathrm{v}) \quad (2)$$

其中 α、β、γ 是可调节的权重参数, 且 $\alpha+\beta+\gamma=1(0\le\alpha\le1, 0\le\beta\le1, 0\le\gamma\le1)$. α 表示同属专业的学生对选修课程的兴趣模型相似度权重系数, β 表示对选修课内容兴趣的相似度权重系数, γ 表示学生间的合作类对选修课程兴趣的权重系数. 当 $\alpha=0$, 则为协同过滤的推荐模型; 当 $\beta=0, \gamma=0$, 则为基于内容的推荐模型.

建立选修课相似性列表$\{c_1, c_2, \ldots, c_m\}$, 根据列表选取课程 c 的最近邻集合, 并建立选修课的相似矩阵 C_{sim}, 当 $sim(\mathrm{u},\mathrm{v})=1$ 时, $sim(\mathrm{u},\mathrm{v})= sim(\mathrm{v},\mathrm{u})=1$. 推荐阶段用下面的公式来进行预测评分, 其中 J 是 k 个相似用户的集合, $r_{c,i}$ 是学生 C 对选修课 i 的评分, $\overline{r_{\mathrm{u}}}$ 和 $\overline{r_{\mathrm{v}}}$ 分别是 u 和 v 的平均打分.

$$P_{cour}=\overline{r_c}+\frac{\sum_{j\in J} sim(i,j)\times(r_{\mathrm{c},i}-\overline{r_j})}{\sum_{j\in J}\left|sim(\mathrm{u},\mathrm{v})\right|} \quad (3)$$

(3)最近邻的确定

每位学生对于自己选修课程的偏好程度是不相同的, 可以对学生偏好程度给予不同的频分(权重). 学生 u 和学生 v 之间的相似度可以用 Pearson 相关系数来计算:

$$sim(\mathrm{u},\mathrm{v})=\frac{\sum_{i\in C}(r_{\mathrm{u},i}-\overline{r_{\mathrm{u}}})(r_{\mathrm{v},i}-\overline{r_{\mathrm{v}}})}{\sqrt{\sum_{i\in C}(r_{\mathrm{u},i}-\overline{r_{\mathrm{u}}})^2}\sqrt{\sum_{i\in C}(r_{\mathrm{v},i}-\overline{r_{\mathrm{v}}})^2}} \quad (4)$$

其中, C 代表同时被学生 u 和学生 v 对某门选修课程的评过集合, $r_{u,i}$ 和 $r_{v,i}$ 分别是选课学生 u 和邻居 v 对特征 i 的评分(或权重), 而 $\overline{r_{\mathrm{u}}}$ 和 $\overline{r_{\mathrm{v}}}$ 分别是 u 和 v 的平均打分(或权重). 一旦计算出相似度, 最相似的学生就会被选择.

当最近邻确定后, 就可以在推荐阶段用下面的公式来进行预测评分, 其中 V 是 k 个相似用户的集合, $r_{\mathrm{v},i}$ 是学生 v 对项目 i 的评分, $\overline{r_{\mathrm{u}}}$ 和 $\overline{r_{\mathrm{v}}}$ 分别是 u 和 v 的平均打分.

$$P_{stu}=\overline{r_{stu}}+\frac{\sum_{\mathrm{v}\in \mathrm{V}} sim(\mathrm{u},\mathrm{v})\times(r_{\mathrm{v},i}-\overline{r_{\mathrm{v}}})}{\sum_{\mathrm{v}\in \mathrm{V}}\left|sim(\mathrm{u},\mathrm{v})\right|} \quad (5)$$

(4)评分组合预测

当最近邻的学生(或选修课程)确定后, 根据相似度加权的所有邻居的兴趣再加上选课学生的平均打分. 这个思想是基于不同学生的评分是围绕不同的基线分布的. 一旦评分预测出来, 我们就可以选择评分最高的项目用以推荐给学生, 具体评分公式如下所示.

$$P_{ui}=\lambda\times P_{stu}(u,i)+(1-\lambda)P_{cour}(u,i) \quad 0\le\lambda\le1 \quad (6)$$

3 实验分析及结果

改进的协同过滤算法实现时, 难点是 K 值的确定. 表 3 给出了不同 K 值下改进的协同过滤算法的不同实

验结果.

表 3 不同 K 值下算法的实验结果

K	准确率/%	召回率/%	覆盖率/%
5	17.57	8.59	49.60
10	25.86	10.06	42.57
15	33.61	11.20	35.52
20	35.29	11.88	30.06
25	35.60	11.27	28.87

从表 3 可以看出, *K* 值与选课系统的准确率和召回率不是成线性关系的, *K* 值选择为 20 左右时可以得到较为准确的推荐结果. 本实验主要以已有选课历史的学生样本进行推荐, 以 2014-2015 学年上半学期的选课为例, 最终选出学生可能感兴趣的 10 门课程进行推荐, 下图为选课系统的课程推荐界面.

图 3 选课系统的推荐界面

4 结语

针对传统选修课系统无法满足学生个性化需求的缺陷, 本文在基于协同过滤的算法中加入基于内容的推荐, 两者结合起来, 相互取长补短, 提出了一种改进的协同过滤推荐算法, 向学生推荐适合其个性化特征的选修课程, 避免学生选课的随意性、盲目性. 但还在一些问题, 比如用户的兴趣模型与课程的不同标志性特征所占权重的分配值得变化有待于进一步研究.

参考文献

1 陶小红.Web 数据挖掘在智能选课系统中的应用研究.办公自动化,2010,(2):27-29.

2 Borges J, Levene M. Data Mining: Concepts and Techniques. Proc. of Workshop Web Usage Analysis and User Profiling. San Diego. 2000. 31-36.

3 邹芳红.Web 数据挖掘与个性化搜索引擎综述.计算机与现代化,2007,(8):44-47.

4 孟凡荣,施蕾,胡继成.数据挖掘中分类技术的研究.计算机与现代化,2008,(3):29-31.

5 Konstan JA, Miller BN, Maltz D, Herlocker JL, Gordon LR, Riedl J. Grouplens: Applying collaborative filtering to usenet news. Communications of the ACM, 1997, 40(3): 77-87.

6 Stead WHL, Rosenstein M, Furnas G. Recommending and evaluating choices in a virtual community of use. Proc. of the SIGCHI Conference on Human Factors in Computing Systems. ACM. 1995. 194-201.

7 Shardanand U, Maes P. Social information filtering: Algorithms for automating "word of mouth". Proc. of the SIGCHI Conference on Human Factors in Computing Systems. ACM. 1995. 210-217.

8 Goldberg K, Roeder T, Gupta D, Perkins C. Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval, 2001, 4(2): 133-151.

9 Liu B. 俞勇,等译.Web 数据挖掘.北京:清华大学出版社,2013.

10 江周峰,杨俊,鄂海红.结合社会化标签的基于内容的推荐算法.软件,2015,36(1):1-5.

11 王全民,刘鑫,朱蓉等.一种新型的混合个性化推荐算法.计算机与现代化,2013,(8):64-67.

12 王景波,郑丽英.混合推荐技术在 Web 挖掘中的研究.科技信息,2010,(33):74-75.

13 李娜.基于混合协同过滤的用户在线学习资源系统个性化推荐方法研究.计算机光盘软件与应用,2015,(2):1-2.

14 翁星星.协同过滤算法研究综述.科技传媒,2013,(16):232-233.

15 曾志浩,张琼林,姚贝,孙琪.基于 Mahout 分布式协同过滤推荐算法分析与实现.计算技术与自动化,2015,34(3):67-72.

16 曹毅.基于内容和协同过滤的混合模式推荐技术研究[学位论文].长沙:中南大学,2007:10-38.

17 陈昊天,帅建梅,朱明.一种基于协作过滤的电影推荐方法.计算机工程,2014,40(1): 55-58,62.

18 陈彦萍,王赛.基于用户-项目的混合协同过滤算法.计算机技术与发展,2014,24(12):88-91,95.

19 焦晨斌,王世卿.基于模型填充的混合协同过滤算法.微计算机信息,2011,27(11):126-128.

20 余小高.大数据环境中微课程个性化学习的研究.中国教育信息,2015,(13):18-21,26.