一种自适应的协作过滤图书推荐系统研究

Research on the Adaptive Collaborative Filtering Recommendation System

马 炎

（西北师范大学 兰州 730070）

摘 要 在知识大爆炸、信息高速发展的年代，如何快速地将用户感兴趣或是对用户有用的信息反馈给用户是本文要解决的问题。通过介绍传统的协作过滤方法，分析其特点以及存在的不足，并基于此提出一种自适应的协作过滤图书推荐系统，以期帮助用户快速地找到需要的书籍条目。

关键词 协作过滤 图书推荐系统

信息社会的快速发展使得数字图书馆信息资源数量日益庞大，以至于用户在面对海量的数字资源时，往往会“迷失”在知识的海洋，要找到自己需要的一种图书往往像海底捞针一样，难以快速有效地找到与用户需求相关的信息。为了解决这个问题，推荐系统应运而生。推荐系统是一种在特定类型的数据库中进行知识发现的应用技术，使用多种数据分析技术为用户更好的服务，向用户主动、及时、准确地提供所需信息，并能根据用户对推荐内容的反馈进一步改进推荐结果。

构建一个综合信息检索、信息过滤、数据挖掘等多种技术的个性化信息推荐系统，是数字图书馆实现信息服务的有效手段。本文将针对以往的协作过滤技术进行改进，分析其中存在的问题和不足，以及其特点，提出一种自适应的协作过滤图书推荐系统。

1 推荐系统实现的主要技术概述

目前，用户解决数字资源获取不便的技术主要有三种：信息检索、信息过滤、协同过滤技术。通常这些都需要有一定的基础，或者称为先验知识，常见的有：客户的浏览行为作为推荐系统的输入，但客户并不知情，称为隐式浏览输入；客户的浏览行为是有目的的向推荐系统提供自己的喜好，称为显式浏览输入；客户输入关键词或项目的有关属性以得到推荐系统有价值的推荐，称为关键词和项目属性输入等。

推荐方法模块中可以采用的推荐技术包括协同过滤推荐、基于内容的推荐、基于效用的推荐、基于知识的推荐和基于关联规则的推荐等。组合推荐也是可以采用的推荐方式，研究和应用最多的是协同过滤推荐和内容推荐的组合。

最近邻居协同过滤推荐是当前最成功的推荐技术之一，其基本思想就是基于评分相似的最近邻居的评分数据向目标用户产生推荐。由于最近邻居对项目的评分与目标用户非常相似，因此目标用户对未评分项目的评分可以通过最近邻居对该项目评分的加权平均值逼近。该算法一般分为表示、邻居生成和推荐产生三个主要阶段。

1.1 基于内容过滤与协同过滤 信息过滤重点在于将新的信息内容分类。这些系统需要建立基于用户兴趣爱好的描述文件，描述文件的作用相当于一个过滤器，使用它可以确保只有那些令用户感兴趣的信息被推荐给用户。信息推荐系统建立描述文件的途径有两种：用户直接输入能反映他们感兴趣领域的关键字或术语或者由信息推荐系统通过观察用户的浏览习惯，自动推出描述文件。基于内容过滤是按照信息的内容特性，采用向量空间法来选择信息。协同过滤技术是过滤技术中应用比较成功的一种技术，协同过滤又称为社会过滤（Social Filtering），是依据其它用户的评价来选择信息的一种十分有效的信息过滤技术，它不依赖于内容，仅依赖于用户之间的相互推荐。

1.2 信息检索与信息抽取 信息检索采取关键词匹配法，准确率不高；信息抽取面向特定领域。信息检索主要针对用户的查询，让用户输入自己感兴趣的某些项目，或者不希望看到的某些项目，按照这些条件进行检索，找到匹配的书目信息。只能根据用户输入的条件进行匹配，不能挖掘用户潜在的兴趣。

1.3 数据挖掘与知识发现 知识发现是从数据中发现有用知识的过程，数据挖掘则指的是整个知识发现过程中的一个特定步骤，是知识发现中最核心的部分。所以，知识发现和数据挖掘往往也是作为同义词使用，一般指运用关联分析、序列模式分析、分类分析、聚类分析以及 OLAP 等知识发现算法，对信息源进行智能处理和知识抽取，发现数据间隐藏的依赖关系，并以法则、规则、科学定律、方程或概念网等特定方式表示抽取的知识。

1.4 书籍推荐系统的特点分析 数字图书馆的推荐系统有其自身的特点，我们需要获得的信息是文本信息，是对我们

作者简介：马 炎，女，1956 年生，馆员。

有用的图书的题名、作者、概要或者出版社信息，所以推荐系统的信息服务对象限于单一的文本信息，不用考虑复杂的图像、声音等数据的推荐，实现起来复杂度要低，而且数字图书馆中资源一般限于各种文献，因此数据源本身具有确定的知识分类体系框架。

但是从用户角度来讲，使用图书推荐系统的用户是比较杂的，不同年龄、不同专业、不同研究方向、不同学历层次、不同文化背景的用户都有可能使用图书推荐系统，所以图书推荐系统所需要考虑的用户对象范围涉及面比较广。

2　常规的协同过滤技术分析

2.1　协同过滤的实现　协同过滤的基本思想是用户可以按照兴趣分类，同类用户具有非常相似的兴趣，因此可以由其他用户的资料协同过滤得到对目标的推荐。用户信息由项目及用户对该项目的评分组成的向量表示，即用户—项目矩阵，矩阵中的数据是用户对项目的评分。协同过滤系统的主要实现步骤一般分为三步：用户信息的表示、邻居的产生和产生推荐，如图 1 所示。

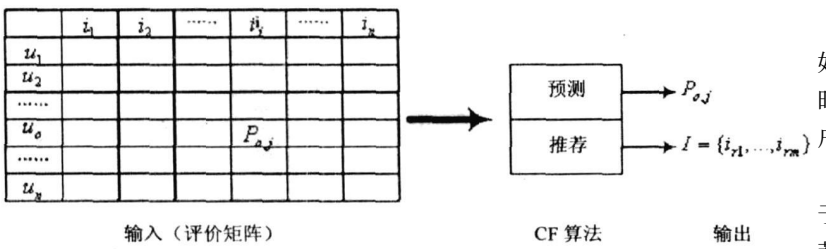

图 1　协同过滤过程

第一步，协作过滤算法把用户对某个信息项的兴趣评价当成初始的评价矩阵，也是算法的基础。用户对某一事物兴趣评价的一般方法前文也有描述。通常是用某一范围之内的数值大小表示兴趣评价级别的高低，数值 0 也通常用来表示用户尚未对此信息项做兴趣评价。

第二步，根据用户对信息的评价矩阵，计算出用户之间的相似度，建立相似用户组。进行用户相似度计算的方法很多，比如向量夹角余弦(Cosine— based Similarity)、用户相关相似度(Correlation— based Similarity)等。建立相似用户组后，根据当前用户在同组中每个成员对某信息的评价信息，预测当前用户对该信息的偏好度。

第三步，根据算法得到的结果，将信息输出给用户。

2.2　存在的问题分析

2.2.1　评价矩阵的稀疏性。由上面的分析，我们也可以看到，必须存在两个用户都对同一信息项做出了评价，CF 的算法才有基础，如果不同用户评价过的信息没有重叠的部分，则无法计算相似性，而且为了计算准确，重叠评价的信息项在数量上不能太少，否则，会导致推荐质量下降。

2.2.2　不具有时变性。因为我们是根据历史数据来进行推荐的，所以在协同过滤技术中，一般不考虑用户的兴趣会随时间的推移发生变化，即认为用户对项目的评分是不随时间改变的。而这在实际的生活中是不可能的，用户的兴趣不会是一成不变的。由于用户在评价信息项时，对其中意义可能不是太了解，在以后的使用中，发现了评价具有失真性，这些变化都要求评价矩阵是一个动态的矩阵，是一个需要时刻更新的矩阵。

3　自适应的协作过滤技术设计

针对第三部分中所述的动态性，我们在设计协作过滤技术时，需要考虑时间的因素，也就是需要设计时变的协作过滤技术。

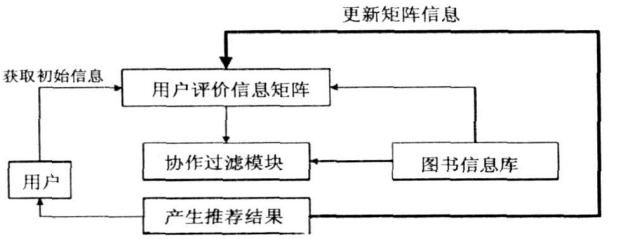

图 2　基于协作过滤技术的图书推荐系统结构

基于协作过滤技术的图书推荐系统结构如图 2 所示。为了体现用户评价信息矩阵的时变性，我们在产生推荐结果的同时还要对用户评价信息矩阵进行相应的更新。

具体更新的策略如下：比如该用户是属于第 m 类用户，即 u_m 用户组，当得到相应推荐书目的同时，需要对评价信息表中的第 m 行的信息项评价系数进行修正。

假设 I_m 是用户组 m 对 n 个信息项评价的向量，即：

$$I_m = [\, i_{(m,1)},\ i_{(m,2)},\ \cdots,\ i_{(m,n)}] \tag{1}$$

在 CF 的算法中，根据不同的加权系数对上面信息项进行加权，得到推荐结果。设这些加权系数用 w 表示：

$$w = [\, w_1,\ w_2,\ \cdots,\ w_n]^{T} \tag{2}$$

则对应得到的结果矢量用 y 表示：

$$y = I_m \circ w \tag{3}$$

这是系统对用户给出的推荐图书，用户根据个人情况会对这些图书进行选择，选出满足要求的图书，所选出来的结果矢量用 d 表示，推荐结果 y 和用户期望结果 d 一般来说是有差别的，这个差别用 e 表示：

$$e = d - y = d - I_m \circ w \tag{4}$$

我们就根据这个差别 e 来对更新用户信息评价矩阵，我们的目的就是使得 e 尽量小，让 e 等于 0，按照最小均方(LMS)准则，可以求得更新后的 I_m 用 I_{new} 表示：

$$I_{new} = I_m - 2\mu \circ e^{T} \circ w \tag{5}$$

4　小　结

本文提出的自适应协作过滤算法，很好地将用户评价矩阵对时间的变化性考虑进来，形成一种对评价矩阵的修正，而且评价矩阵是针对划入同一组人对信息项的评 （下转第 109 页）

含层及输出层的神经元数目、权值和误差因子等，这些相当于是设定分析系统的初始数据。接下来设定网络的初始权值，即确定神经网络的运算规模和复杂程度，生成数据的初步分类结果，并以该结果作为神经网络数据推演为条件，如果获取的结果用户不满意，还可以按照预先设定的规则集进行条件的修正，直到获取的结果为用户满意的结果为止。

表 1 读者偏好分析系统数据格式

读者姓名	性别	年级	专业	借阅书名	借阅地点	图书类目	借阅时期	借阅时间	归还时间	借阅方式
张×	男	2	信息管理	MBA管理	A校区	管理	学期初	*年*月	*年*月	刷卡

4 创新点总结

采用神经元网络的方式，分析并获取图书馆读者偏好的智能系统是一种行之有效的方法。通过对读者的个人信息历史数据的训练和学习，调整预测模型各组成神经单元之间的连接权重，确定输入输出之间的内在联系，从而使模型具备了对读者信息的预测分析能力。通过该模型进行读者偏好信息的分析，首先弱化了传统分析中的人为因素，提高了预测结果的准确性和权威性。其次应用神经网络超强的非线性处理能力，更加准确地体现了读者个人信息中的潜在规律和特点，该方法强大的自我学习能力能够在实践中不断调整结果的精确性和正确性。

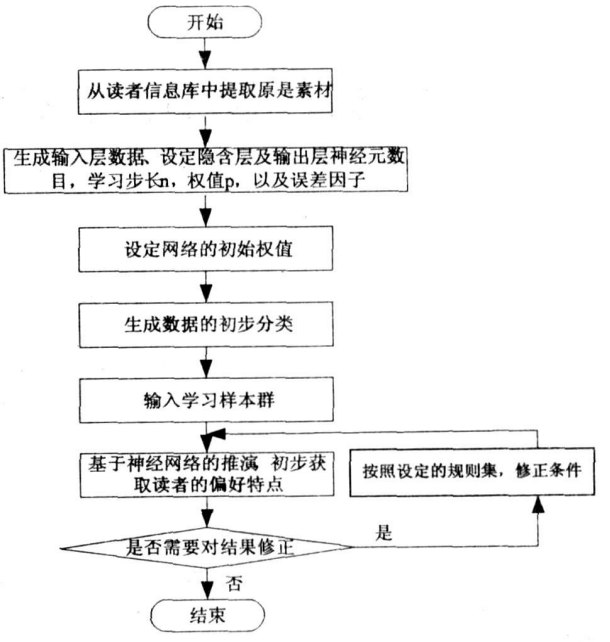

图 2 读者偏好分析流程

参 考 文 献

1 刘增良. 神经网络与模糊技术选编[M]. 北京航空航天大学出版社, 2006

2 鲍翠梅, 王尊新, 白如江. 数据挖掘技术及其在图书馆中的应用[J]. 情报杂志, 2004(9): 49－51

3 何少卓. 浅谈数据挖掘及其在图书馆的应用[J]. 图书馆界, 2004(3): 52－54

4 Kodratof Y. Rating the Interest of Rules Induced From Data and Within Texts, Database and Expert Systems Applications. Proceedings, 12th International Workshop on, 2001: 265－269

5 罗仕健, 朱光磊. 网络环境下数据挖掘技术在图书馆中的应用[J]. 情报杂志, 2004(6): 22－24

（责编：军阳）

（上接第106页）分，所以在客观上避免了个体用户对数据的影响，形成一种统计上的平均，更加接近实际。

随着时间的推移，搜索结果会越来越与用户期望结果逼近，用系统推荐结果与用户期望结果的匹配度来检验系统的性能，系统结果如图 3 所示。

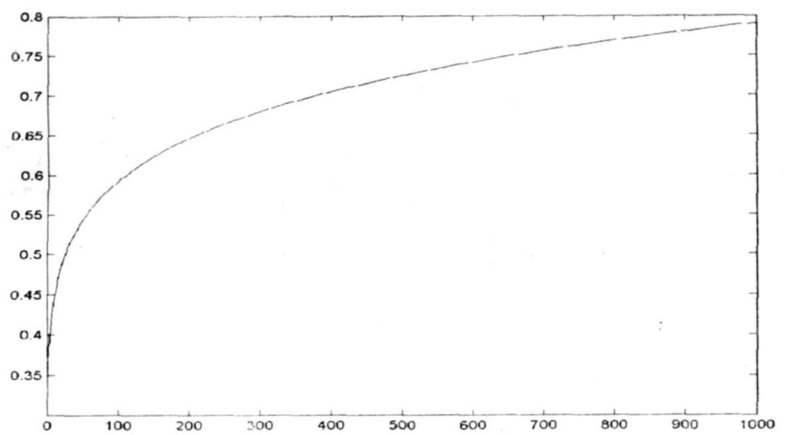

图 3 系统仿真结果

我们发现，在刚建成系统后，系统推荐的结果与用户期望的结果匹配性比较差，大概只有 0.35，这个初始值的高低取决于系统初始矩阵的设计值，或者引入专家推荐的水平的高低。随着用户使用时间的增长和同组不同用户的体验，匹配度会逐渐增加。

参 考 文 献

1 曾庆辉, 邱玉辉. 一种基于协作过滤的电子图书推荐系统[J]. 计算机科学, 2005

2 SarwarB, Karypis G, Konstan J, et al. Item－based Collaborative Filtering Recommendation Algorithms. WWW10, Hong Kong, 2001

3 Robin Burke. Hybrid Recommendation Systems: Survey and Experiments. Department of Information Systems and Decision Sciences, California State University, Fullerton.

4 Schafer, J B, Konstan, J, and Riedl, J. E－Commerce Recommendations Applications[J]. Journal of Data Mining and Knowledge Discovery, 2001, 5(1－2): 115－153

5 龚耀寰. 自适应滤波[M]. 北京: 电子工业出版社, 2003

（责编：军阳）