智能图书系统中的个性化推荐①

乔亚飞[1], 张 霞[1], 张文博[2]

[1](武汉大学 印刷与包装系, 武汉 430079)
[2](广州博冠信息技术有限公司, 广州 510000)

摘 要: 互联网时代的到来造成知识的“过度”传播, 知识点的分散和无组织使得有系统学习要求的用户无从下手, 用户对专业知识的查找、学习变得困难, 如何面向用户实现个性化的知识推荐是智能图书系统中需要解决的关键问题之一. 本文利用上下文偏好提取技术, 获取用户的兴趣度. 在基于用户的协同过滤推荐算法和基于项目的协同过滤推荐算法的基础上引入时间加权因子, 较好地解决了对新用户推荐时产生的“冷启动”问题, 实现了服务推荐结果的个性化.

关键词: 智能图书; 协同过滤推荐算法; 时间加权; 上下文信息

Personalized Recommendation in Smart Books Sysmtem

QIAO Ya-Fei[1], ZHANG Xia[1], ZHANG Wen-Bo[2]

[1](School of Printing and Packaging, Wuhan University, Wuhan 430079, China)
[2](Guangzhou BOSMA IND Co.LTD, Guangzhou 510000, China)

Abstract: The advent of the era of the Internet causes the “excessive” of knowledge dissemination, knowledge fragment and disorganization makes users who have requirements of a systematic study difficult to know how to start. How to realize personalized knowledge recommendation for users is one of the key problems to be solved in the intelligent library system. This paper uses the context preference extraction technology to obtain the users’ interest. And it introduces weighted factor based on the intelligent library system as an example in the user-based collaborative filtering recommendation algorithm and the collaborative filtering recommendation algorithm, it better solves “cold start” in the recommendation to new users as, implements the service of personalized recommendation results.

Key words: smart books; collaborative filtering; time weighted; the context information

在网络快速发展的背景下, 如何将知识以一种更有效率、更全面的方式组织起来, 供不同需求的人进行学习是知识学习领域的重要研究课题. 智能图书系统是基于上下文感知计算的智能化知识学习的系统, 能够根据用户的不同特点和需求, 提取异构系统中的信息, 自动生成基于用户水平层次与兴趣方向的知识结构体系, 以供用户使用[1-3]. 针对不同的用户进行个性化的知识推荐是智能图书系统的主要功能之一. 本文以传统的 Item-based 和 User-based 协同过滤算法为基础, 对用户进行上下文偏好的提取, 获取用户的兴趣特征, 再将上下文信息引入到协同过滤推荐算法中去, 自动生成适应用户知识水平层次与兴趣方向的知识结构体系. 传统的协同过滤算法由于没有相关的用户数据和上下文信息, 无法利用上下文感知推荐系统进行相关推荐, 针对这种问题, 本文引入了时间加权因子, 使智能图书系统能够更好地向新用户推荐知识内容.

1 用户上下文偏好的提取

“上下文相关偏好(Context-dependent Preference)”指在不同的上下文条件下, 用户对项目及其属性的偏好有可能不同. 上下文用户偏好提取技术(Contextual User Preference Elicitation)已成为上下文感知推荐系统的先决条件和关键技术[4,5]. 本文采用本体建立用户偏

① 收稿时间:2016-01-13;收到修改稿时间:2016-03-10 [doi:10.15888/j.cnki.csa.005374]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

好模型, 对用户偏好进行量化. 用户个性化模型定义为一个三元组, 如式(1)所示.

Model= (Information, Inquiry, Interest) (1)

Information 表示用户的个人信息, Inquiry 表示用户请求查询的集合, Interest 表示用户的兴趣程度. 偏好的提取分为两个步骤:

1) 用户静态信息

Information 是用户静态信息, 由用户在注册时输入, 可以直接提取, 数据项内容如表 1 所示.

表 1 Information 用户信息数据表

名称	说明	数据类型	约束
UID	用户 ID	int	主键
UPassword	用户登录密码	varchar	非空
UEmail	注册邮箱	varchar	
USignDate	注册日期	varchar	
UIDNo	身份证号	varchar	
UName	姓名	varchar	
USex	性别	enum	
Ujob	职业	varchar	
Udegree	最高学历	varchar	
Umajor	最高学历专业	varchar	
Uschool	毕业院校	varchar	
Upreference	用户素材偏好	enum	

其中 Upreference 这一字段, 知识资源包括网页, 文献, 书籍等. 对于用户注册时提取的偏好, 则在算法初始化时给予较高的兴趣度, 推荐时首先按照用户选择的类型来推荐.

2) 用户个性化信息

用户的个性化信息包括 Inquiry 与 Interest, 是用户动态信息. 在用户完成注册后, 系统会对用户的相关操作进行记录, 用隐式的方法来获取用户的需求和兴趣度, 完成对用户模型中 Inquiry 和 Interest 数据项的填充. Inquiry 用三元组表示如式(2)所示.

Inquiry= (Knowledge, History, Relation) (2)

其中, Knowledge 表示用户需求知识点的集合, 表示用户希望获得的信息; History 表示用户已经阅览过的知识点的集合, Relation 表示已阅览知识点之间的联系.

Interest 用三元组表示如式(3)所示.

Interest= (Preference, Bookmark, Duration, Frequency) (3)

在 Interest 数据项中 Preference 表示对知识资源类型的偏好, Bookmark 表示用户收藏的感兴趣的知识资源, Frequency 表示用户访问知识资源的频率或者次数, Duration 表示用户在当前资源停留的时间.

用户模型中 Inquiry 数据项和 Interest 项是动态变化的, 对用户最终个性化推荐结果生成影响的偏重点不同. Inquiry 集合主要对知识点(指具体的知识内容)的推荐结果产生影响. 通过对本体模型的使用, 系统根据 Inquiry 集合中的 History 和 Relation 生成基于语义的概念检索和关联, 生成用户偏好知识点的推荐结果; Interest 集合主要对知识资源(指承载知识内容的资源, 如期刊论文、网页等)的推荐产生影响. 根据用户阅读知识资源并对其产生的操作, 例如标记、长时间阅览、多次查看, 对集合进行填充, 然后使用推荐算法对用户下次阅览的知识资源进行优化推荐.

用户动态数据的完善过程是一个迭代进行、螺旋上升的过程, 随着用户使用图书历史的增加, 用户偏好模型的实例信息对用户的描述更加准确.

2 协同过滤推荐算法

协同过滤推荐(CFRS)一般有两种方式来实现, 一种是基于用户(User-based, 简称 UBCF-RS)的协同过滤推荐系统, 一种是基于项目(Item-based, 简称 IBCF-RS)的协同过滤推荐系统[6-8].

2.1 基于用户上下文的协同过滤推荐算法

UBCF-RS 运算步骤如下:

1) 用户发起请求, 在判别用户已注册并且用户使用过系统后(拥有历史上下文信息), 通过引入用户显式提供的信息和历史上下文信息, 建立“用户-知识点”空间向量模型, 利用余弦相似度计算法计算当前用户 Ui 与其他目标用户之间的相似度.

2) 对与当前 Ui 而言, 排列与 Ui 比较类似相关的 n 名目标读者, 然后将结果程度最高的 N 名读者筛选出来. 之后进行从高到低的排序, 得到的结果就是与读者 Ui 最类似的读者集合 S.

3) 针对读者集合 S, 遍历 S 中所有读者对知识点阅览情况, 统计 S 中读者阅览次数最多, 但读者 Ui 没有阅读过的项目, 然后对这些知识资源进行高到低的排序, 寻找合适的 Top-n 知识点集合推荐给读者 Ui.

2.2 基于项目的协同过滤推荐算法

与 UBCF-RS 相类似的, 在 IBCF-RS 中, 主要通过分析用户的行为记录来计算知识点之间的相关度. IBCF-RS 主要步骤如下:

© 中国科学院软件研究所 http://www.c-s-a.org.cn

1) 计算项目之间的相似程度. 可以采用计算用户之间相似度的空间向量模型, 但在计算中略去了调用次数, 采用对知识点的调用行为来表示知识点的特征.

2) 对比项目的类似程度, 与读者的调用记录比对.

3) 对相应的用户进行个性化的推荐.

本文建立了颜色科学领域的智能图书系统, 对于已经注册并使用过系统的用户, 可以通过上述两种推荐算法生成推荐. 在脱机时主动计算注册用户之间的相似度, 用户相似度的计算采用公式(4).

$$Sim(T_1,T_2)=\frac{\overrightarrow{w_1}\times\overrightarrow{w_2}}{|\overrightarrow{w_1}|\times|w_2|} \tag{4}$$

如果存在与当前用户相似的用户, 选择基于用户的协同过滤推荐算法进行推荐, 如果没有相似用户, 则选择基于项目的协同过滤推荐算法进行推荐. 具体步骤如图 1.

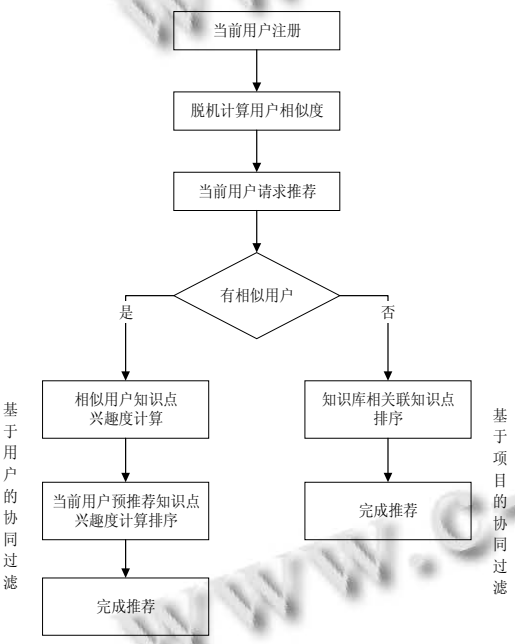

图 1 基于上下文的协同过滤推荐算法

2.3 推荐算法的改进

当新用户对系统进行使用时, 由于没有相关的用户数据和上下文信息, 无法利用上下文感知推荐系统进行相关推荐, 这通常被称之为“冷启动”问题. 为了解决这个问题, 除了在最初用户注册时显式获取的用户相关信息进行推荐之外, 还需要让用户选择自己感兴趣的知识点进行相关初步推荐, 以隐式获取用户上下文信息. 在进行初步推荐时, 需要计算现有知识点内容的热门程度, 将热门程度较高的知识点推荐给用户.

一般来讲知识点内容的热门程度由调用次数来进行衡量, 但是需要考虑到领域本体模型也是在迭代更新的, 一个构建时间较久远的知识模型的调用次数一般比一个新迭代的知识模型调用次数要多, 但从热门程度而言, 新迭代的知识模型很可能更加受到用户的关注, 因此为了削弱迭代时间这个因素对热门程度的影响, 本文对推荐算法进行了补充, 提出了基于时间加权的知识点热门程度的算法, 如公式(5)所示.

$$p_i=\frac{\sum_{k\in N(i)}\frac{1}{1+\alpha|t_{now}-t_k|}}{N(i)} \tag{5}$$

公式的核心思想是加入了一个时间衰减因子, 其中 t_{now} 表示档期的时间, t_k 表示知识点在第 k 次调用的档期. N(i)表示知识点的调用次数. α 是常量因子, 可以看出, 随着迭代时间的推移, $|t_{now}-t_k|$ 的值就会越大, 这样对知识点的总调用记录就会对知识点的热门程度影响变小. 这种补充算法能够让新用户在第一次使用智能图书系统时, 就对系统内的知识点有一个大致且直观的了解, 同样的, 也让新迭代的知识模型和以前更新的知识同样的调用机会, 避免了被调用次数越多的知识一直被推荐的情况. 这种算法也可以在脱机的情况下进行, 节约用户时间.

在对 CFRS 的优化中, 也可以在对知识点进行排名时, 加入时间衰减因子, 让推荐结果更加准确, 如公式(6)所示.

$$p(u,i)=\sum_{v\in S(u,K)\cap N(i)} w_{uv}r_{vi}\,p_i \tag{6}$$

其中, w_{uv} 表示用户 u 和用户 v 的兴趣相似度(由算法第一步余弦相似度计算而来), N(i)表示对知识的 i 有调用行为的所有用户. Pi 为时间加权因子.

2.4 实例

本文以颜色科学领域的知识为例构建知识资源数据库. 该实例就是通过颜色科学知识本体, 使用者的用户模型, 根据改进后的推荐算法把资源推荐给学习者.

根据上述的数据收集方法, 我们对日志数据预处理之后得到的分析数据如表 2.1 所示, 其中 UserID 为学习者的标识, Res_ID 为资源标识, KP_ID 为知识点标识, Starttime 为开始时间, Endtime 为离开时间, Duration 为在该资源上停留的学习时间, Action 为用

© 中国科学院软件研究所 http://www.c-s-a.org.cn

户对资源的操作(保存 S, 浏览 R).

表 2 日志数据处理后的分析数据

UserID	Res ID	KP-ID	StarttimeEndtime	Duration	Action	
S11111	R1004	1299	2014-5-27 8:10:11	2014-5-27 8:12:40	3	S
S11114	R1136	1326	2014-5-27 9:20:10	2014-5-27 9:24:12	5	R
S11116	R2008	1023	2014-5-27 10:10:03	2014-5-27 10:13:43	4	R
S11118	R2345	1402	2014-5-27 10:25:57	2014-5-27 10:30: 20	6	R
S11110	R1009	1818	2014-5-27 11:11:12	2014-5-27 11:15:40	5	R
…	…	…	…	…	…	…

从表 2 中, 我们可以很容易的看出学习者的学习情况, 例如 UserID 为 S11111 的学生 2014-05-27 KP-ID1299 的知识资源, 持续时间为 3 分钟. 通过用户日志数据预处理得到能够体现用户兴趣度的有效数据, 按照用户兴趣度计算算法得到用户—知识点兴趣模型. 根据用户模型, 我们得出用户×资源的评分矩阵, 其中每行代表用户对资源的真是兴趣度, 每列代表一个学习资源. 因此该矩阵中的每项就代表该行对应的用户对该列对应资源的评估分值. 如下表 3 所示, 每个资源表示一个知识点, 每个分值表示用户对该知识点的学习评估值.

表 3 用户-知识资源偏好量

	A1	A2	A3	A4	A5	Sim(i,1)
U1	2.5	0.8	0	0	2.1	
U2	0.8	0	1.7	2.1	2.7	1.05
U3	0	1.6	0	2.5	0	0.2
U4	0	2.3	0.8	2.2	1.2	0.63
U5	0.6	0	1.2	2	0	0.25

在这里我们以 U1 为学习用户, 使用改进后的推荐算法, 首先要比较用户 U1 和用户 U2、U3、U4、U5 的相似度, 利用公式 2.2.1 计算得出他们之间的相似度分别为: sim(U2,U1)=1.05, sim(U3, U1)=0.2, sim(U4, U1)=0.63, sim(U5, U1)=0.25 从中可以看出用户 U1 和用户 U2 最为相似, 因此我们把用户 U2 作为用户 U1 的邻居, 通过公式 2.3.2 计算得把 A4 推荐给用户 U1 的推荐度为 0.519.

2.5 实验及结果分析

加入了时间加权因子的协同过滤推荐算法能够精确的计算出用户感兴趣的资源, 然后对其进行准确的推荐, 算法在 Matlab 平台上仿真实现.

精确度是评价推荐算法的一个重要指标. 平均绝对偏差 MAE(Mean Absolute Error)是最常用的一种推荐质量度量方法, 通过计算预测用户评分与实际用户评分之间的偏差来度量预测的准确性, MAE 越小, 推荐质量就越高. MAE 定义为:

$$MAE=\frac{\sum_{i=1}^{N}|pi-qi|}{N}$$

N 为测试集大小, p_i 为预测评分, q_i 为实际评分. 实验分别对基于时间加权的推荐算法和传统的协同过滤推荐算法进行了比较和分析, 取最近邻居数分别为 10、20、30、40、50、60, 分别进行实验测试, 实验结果如图 2.

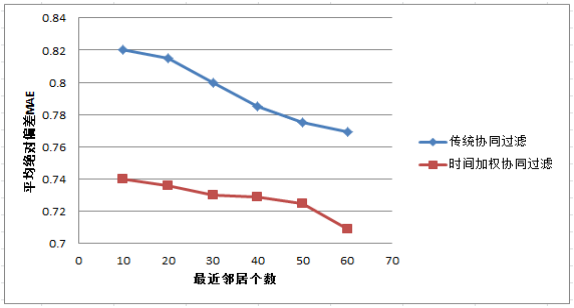

图 2 MAE 值

© 中国科学院软件研究所 http://www.c-s-a.org.cn

从图中可以看出加入了时间加权的协同过滤算法比传统的协同过滤算法的 MAE 值要小很多，推荐效果要好，并且随着最近邻居个数的增多推荐效果越来越好.

3 智能图书系统原型

本文开发了颜色科学领域智能图书系统原型，用户的个人主页中如图 3 所示.

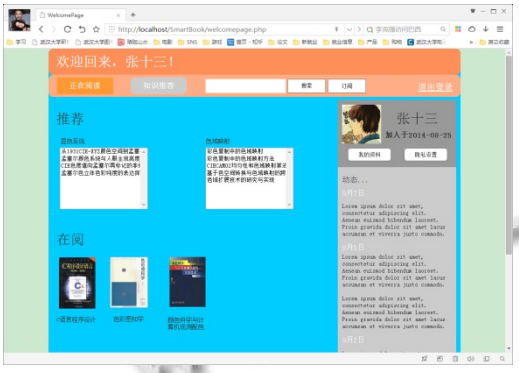

图 3 用户个人主页

图 4 知识点检索结果

其中，菜单栏主要有两个功能，一个是检索功能，另一个是相关知识推荐功能. 假设用户想要对显色系统进行更进一步的了解，在对显色系统中的“奥斯特瓦尔德系统”进行检索学习时，系统会根据本体模型推荐相近的“孟塞尔表色系统”和“自然色系统”给用户，这是基于颜色科学领域本体模型的知识点相关性推荐，做到基于语义的匹配，给用户推荐有相关性的知识点. 同时，系统还会根据相近用户的浏览，根据基于用户的协同过滤推荐算法，推荐例如“同色异谱”这一相关知识点. 用户点击相关内容就可以进行阅读，阅读时可以对文章进行收藏等操作，阅读界面如图 4 在积累了一定的检索量和阅读量后，系统就可以通过对用户上下文的提取，依照相应的规则和推荐算法进行智能推荐.

4 总结

综上所述，本文研究了智能图书系统中的个性化推荐算法，将时间加权引入协同过滤推荐算法基本满足了不同知识层次用户的推荐需求，也一定程度上的解决了推荐系统的“冷启动”问题，初步实现了智能图书系统中的个性化推荐功能.

参考文献

1 WBeer AW. Smart books - Adding context-awareness and interaction to electronicbooks. International Conference on Advances in Mobile. 2011. 218–222.

2 张文博,张霞.基于本体与上下文感知的智能图书系统.计算机系统应用,2015,24(5):220–226.

3 张文博.基于领域本体与上下文感知计算的智能图书系统[硕士学位论文].武汉:武汉大学,2015.

4 章诗杰.移动环境上下文感知的协同过滤推荐模型研究[硕士学位论文].杭州:杭州电子科技大学,2013.

5 张德英.基于上下文和用户行为的移动偏好获取系统的设计与实现[硕士学位论文].北京:北京邮电大学,2012.

6 王柱,周兴社,王海鹏,倪红波,武瑞娟.智能博物馆个性化导航技术的研究与应用.计算机工程,2009,35(15):280–283.

7 Lee SK, Cho YH, Kim SH. Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations. Information Sciences, 2010, 180(11): 2142 –2155.

8 Park DH, Kim HK, Choi IY, Kim JK. A literature review and classification of recommender systems research. Expert Systems with Applications, 2012, 39(11): 10059–10072.

© 中国科学院软件研究所 http://www.c-s-a.org.cn