推荐系统研究综述

王改芬

(宜昌教育学院 计算机系,湖北 宜昌 443000)

摘　要:推荐系统已经成为用户和网络应用软件交互的一个重要部分,特别是推动了电子商务的发展。商家已经意识到为了增加销售额和防止客户流失需要开发出具有个性化和可适应的推荐系统。同样,网络上的用户依靠这样的推荐系统在巨大的信息空间中可更有效地找到自己感兴趣的项目。提供了推荐问题的简洁描述,并概述了产生推荐的各种方法及发展趋势。

关键词:推荐系统;电子商务;用户概貌; 本体领域知识

中图分类号: TP311　　**文献标识码**: A　　**文章编号**: 1672- 7800- (2007)12- 0068- 02

0　前言

推荐系统根据用户以前或当前与其他用户或与其类似的其他用户的交互能推断出用户的需求并输出给用户。因此推荐任务被认为是一个预测问题:系统必须努力地预测对一个特定的用户来说有用的特殊范畴的内容、页面或项目,然后依照被预测出来的效用值引导这些预测出来的内容分成不同的等级。用一个分(数)来反映用户对某项目感兴趣级别,依此来描述一个项目的有效值。通常推荐系统输出是对一个活动用户具有最高预测兴趣值和最有用的一组项目。因此,推荐系统被认为是用户和一组有用值(或兴趣分)项目之间的一个映射。一般来说,把推荐作为预测任务的观点来自于映射从事实上来看并不是被定义在用户—项目对的整个域上,因而需要推荐系统为域的某些部分估计兴趣值。

形式上,把 $U=\{u_1,u_2,\cdots,u_m\}$ 看作是存在的一组用户,把 $I=\{i_1,i_2,\cdots,i_n\}$ 看作是一组项目。对一个 $u\in U$ 用户的概貌(profile)可看作是一个 n 维有序偶的向量,$u^{(n)}=\langle(i_1,s_u(i_1)),(i_2,s_u(i_2)),\cdots,(i_m,s_u(i_n))\rangle$。这里 $i_j's\in I$ 和 s_u 是用户 u 的偏好函数,把有用值赋给 I 中的项目。因此,$s_u:I\to R$ 函数表示了用户 u 的概貌(Profile),映射项目到一个有序有用值集合 R 中。系统里包括一些对项目的最终评价,函数 S_U 被称为一个评价等级函数,它映射项目到一个离散的评价等级集合里。要注意的是因为对一个给定的用户 u 的映射函数 S_U 一般并不定义在项目的全域上,推荐系统必须估计或预测一个给定用户对域中部分项目的兴趣分值。

在一个典型的推荐系统里,随着时间的推移,所有用户的概貌被收集并被存储。从概念上讲,所有用户的概貌数据库表示为 $m\times n$ 的矩阵,$UP=[S_{uk}(i_j)]_{m\times n}$,这里 $S_{uk}(i_j)$ 表示用户 u_k 对项目 i_j 感兴趣的程度,或者也可以把 UP 看作 n 维用户概貌向量的一个集合。

因此推荐系统可以被看成是 $REC:P(UP)\times U\to P(I)$ 的一个映射,每一个用户被映射到一个基于该用户的概貌和(可能为空)其它用户概貌集的一个推荐项目集中。这种映射的范围不是用户以前评价的项目 $I'\subseteq I$ 的子集。

不失一般性,假设对一个给定目标用户 $u_k\in U$ 将返回具有最高预测值的项目,那么 REC 映射定义如下:$REC(up,u_k)=\{\arg\max_{ij\in I}s_{uk}(ij)\}$。这里,up 是(可能为空)用户概貌 UP 集合的子集合,$S_{uk}(i_j)$ 是用户 u_k 对项目 i_j 的预测兴趣值。推荐系统返回按预测兴趣值排序推荐出来的 top N 项目集合。依靠使用的预测算法,算法也许不能够得到对一个特定项目的兴趣分值,在这种情况下,REC 映射将产生一个空的或不明确的值。

传统个性化推荐包括基于内容、协同过滤和基于知识系统。每一种方法以收集的数据特定类型来产生用户概貌,并由所使用的算法来产生推荐。

1　推荐系统的类型

从体系的观点来看,产生推荐的方法可分为两类:基于内存和基于模型。普通的使用懒惰学习算法的基于用户的协同过滤和大多数基于内容的过滤系统是基于内存方法的特例,然而在部署之前获得模型的基于项目和其它协同的过滤方法是基于模型系统的特例。基于内存的系统只是在产生推荐的时候存储所有的数据并由它来产生,所以它们更易受到可伸缩问题的影响。基于模型的方法,在离线阶段需要执行大量的计算。另一方面,通常收集的数据越多,基于内存的系统越适合于改变用户的兴趣,相对于基于模型的技术,它的模型对于可能是渐增的也可能是被重建并解释新的数据。

从算法的观点来看,推荐系统被分为 3 类:基于知识的系统、基于内容过滤系统和协同过滤系统、基于知识的推荐系

作者简介:王改芬(1973-),女,湖北钟祥人,宜昌教育学院讲师,研究方向为计算机网络。

统或依赖于关于项目的清晰的域知识或依赖于关于用户(例如基于人口统计学特征)的知识来得到相关的推荐。许多依靠手工或自动产生基于知识的决策规则系统推荐那些由排序规则约束的指定的用户满意项目。通常通过和用户交互来获得用户概貌。一些研究人员致力于研究基于用户的人口统计学,采用机器学习技术将用户分为若干类,因此,个性化系统中可以使用自动的推知决策规则。

在基于内容过滤系统里,用户概貌获得用户以前感兴趣项目的内容描述。项目内容描述由表征项目特性的特征集或属性集来组成。系统里产生的推荐任务通常包括从在用户概貌中用内容描述未被发现的或未评价特征里萃取出来的一些特征的对比。推荐给用户被认为是与用户概貌十分相似的项目。换句话来说,对用户 u 的目标项目 i_j 的兴趣分值 $S_{uk}(i_j)$ 是基于其它项目 $i_l \in I$ 的兴趣分值,属于 u's 概貌的则被认为是与 i_j "相似的"。

在基于内容(和一些基于知识的方法)的系统里,用户概貌数据库 UP 只包括个别概貌,有目标用户 u 和基于与用户概貌相似或基于人口统计学或其他用户的个性特性的目标项目 i_j 的 $S_{uk}(i_j)$。在大多数基于内容的过滤系统中,内容描述是从 Web 页或产品描述中抽取出来的文本特征。同样地,这些系统经常依赖于来源于信息搜索和信息过滤研究中清晰的文档模型技术。用户概貌和项目本身都用权重术语来描述(例如,基于 TF.IDF 权重术语模型)。可以通过向量相似性计算(例如,使用 Cosine 相似性计算)或使用贝叶斯分类概率方法来预测一个用户感兴趣的项目。和协同过滤不同的是,基于内容的过滤本质上被分为单个的,仅由活动用户以前评分过的和项目相关联的特征来建立的。基于内容的过滤系统的主要缺点是它们倾向于过于专业化的选择项目,因为概貌只基于用户以前评价过的项目。

协同过滤试图处理上述所提到的其它方法中的一些缺点。在标准的协同过滤系统中,兴趣分通常只表示从一个有序但离散范围中的等级值,UP 包括系统所有用户过去的评价。在这种情况下,基于用户的概貌和在 UP 中的其它概貌的相似来预测或评估目标用户的兴趣分。通常这些技术会影响当前用户对具有与其相似用户(最近邻居)目标(例如,电影或产品)的评价等级的匹配,这是为没有被活动用户评价或看到过的目标对象产生推荐。传统上,使用该技术来实现此任务是基于内存的 k 最近邻居(kNN)分类方法,为了找到具有相似的口味或兴趣前 k 个用户,此方法比较目标用户的概貌和其它用户的历史概貌。因此,在一个典型的协同过滤推荐系统中,一个用户 u 对一个目标项目 i_j 的兴趣分 $S_{uk}(i_j)$ 基于被认为和用户 u 十分相似概貌的用户 $u \in U$ 的兴趣分被估计出来。

基于内存协同过滤的最大局限是它的可伸缩性差。随着用户数和项目数的增加,这种方法为了在和用户交互期间推荐动态内容,可能导致无法接受的等待时间。kNN 技术的另外一个局限来源于数据集的稀疏性。随着数据库中,项目的增加,关于这些项目在每个用户概貌中的评价密度降低。依次,这将会降低用户被访问的或已被评价的项目的有意义的重复部分,降低了计算相关性的可靠性。

基于模型协同过滤被认为是基于项目的协同过滤,基于项目的协同过滤从相同用户评分概貌数据库开始,离线建立一个基于项目的相似矩阵,并在预测阶段产生推荐。不是以项目的内容描述的项目相似性,而是在典型的基于内容的过滤,项目之间的相似性基于用户对这些项目的评价。每一个项目用一个向量来表示,用余弦相似和相关性相似的度量来计算相似度。推荐过程预测利用项目中目标项目的邻居中的用户项目评价的权重总和出活动用户以前没有看到或评价过的项目的评价等级。

为了提高协同过滤的可伸缩性,研究人员提出了许多基于数据挖掘的其它方法,特别是利用单击流或其他类型的行为数据的 Web 个性化系统中的语境(上下文)方法。这些方法中的模式识别算法有关联规则挖掘、序列模式发现和聚类,为了产生聚类用户的模型,将这些算法应用到包含有以前用户的历史评价或基本概貌的 UP 上。相反,用户模型可以和活动用户概貌一起协力预测未来用户的行为或产生推荐。在这样的系统里,通常通过对网站上用户活动,例如在一个页面上的停留的时间、购买的或选择的产品等的隐性观测可得到用户对项目的兴趣分。一些推荐系统产生了较有效的推荐,另一些推荐系统证明在面对 shilling 攻击时比协同过滤有更具有鲁棒性[5]。

不论那一种个性化推荐方法都需要隐性或显性地收集用户概貌的数据。显性收集需要用户积极参与进来。需要人口统计学或与用户交互来获得个人信息的系统也可以采取在注册参与在线调查或在购买的时候提供个人或商业信息(它可结合各种数据聚合服务来获得有效的在线人口数据)的形式。传统的协同过滤和一些基于内容的过滤系统应用中以对单个项目评价的形式来反馈用户显性数据。日益个性化的系统需要用户行为数据和尝试用启发式方法(例如不管是否购买该项目,但他关注该项目的时间)来处理用户兴趣。使用用户概貌的隐性反馈去除了需要用户提供个性化信息相关的一些负担。

2 推荐系统发展新趋势

在协同和基于内容的过滤系统里,研究人员已经探索出整合本体领域知识和用户概貌的许多方法[6]。在领域本体中,用户概貌实际上能反映出领域的结构,因此可能需要一个比用在一般方法里单调的表示更复杂的表示。另外,整合本体知识将允许这样的系统来向用户推理出相关的推荐或更好地向用户解释产生的推荐。

尽管这种被表示成接受的领域知识整合可能会产生更智能化的推荐,特别是语义网络,但最近出现的随机(ad hoc)社区,从社群标签到其它社群导航系统,也为使用推荐技术提出了新的良机。在社群标签中,许多用户以关键字的形式添加元数据来共享内容,这导致了语义结构的松散即 Folksonomies。然而协同标签提供了新的机制,用户可以在巨大的芜杂信息空间里浏览并找到自己感兴趣的内容,这种系统的开放特性也将导致潜在的用户数据中的相当多的噪音,这些噪音导致用户不知道那些是他感兴趣

考试信息语音服务系统的开发和应用

朱其文

(中煤能源集团大屯公司 信息中心,江苏 徐州 221611)

摘 要:主要叙述了自主开发的计算机软件"考试信息语音服务系统"的语音查询、自动通知功能,以及软件在教育部门的应用。

关键词:D160A PCI 语音卡;信息查询;自动通知

中图分类号:TP393.09 文献标识码:A 文章编号:1672-7800(2007)12-00-0070-02

1 系统概述

针对每年高考、中考、成人高考成绩和录取揭晓后,考生迫切想得知考试成绩和录取情况。笔者开发了考试信息语音服务系统,从而实现考试信息自动查询、语音自动通知功能。

考生呼入查询热线后,系统自动接收主叫电话号码,根据考生输入的准考证号码和识别到的主叫电话号码来查询高考、中考、成考的考试成绩,录取学校以及通知信息,并实时记录各考生的查询次数,对查询结果分类统计。为了保密需要,系统自动识别主叫电话号码并与考生预先登记电话号码进行比较。每个考生只有用自己登记的电话,才允许查询本人的考试成绩、录取信息。该系统使用 TTS 文字语转换功能,实现文字的自动朗读,使系统应用更加灵活方便。

该系统具有语音自动通知功能,根据设置的自动通知类别、语音通知文件名、时间段、通知次数,对各类考生进行自动通知,并实时记录呼叫结果、呼叫时间、通知次数、无应答次数等,以及对呼叫结果分类统计。

该系统还具备了一些辅助功能,例如实现文本放音、文本语音转换、电话录音、语音文件试听等功能,可以用电话播放文本对应的语音,将输入的文本转换为语音文件,通过电话进行录音并且存为语音文件,用电话试听语音文件。

硬件环境:计算机一台(P4 CPU,1G SDRAM 内存,80G 硬盘),D160A PCI

的项目。因此,最新研究工作主要集中于帮助用户找到相关的标签、感兴趣的项目或具有公共兴趣的社群的在用户——项目——标签空间中的数据挖掘和推荐技术的应用[7]。社群标签系统的推荐技术有很大发展前途,它可以在保留这些新出现的环境民主和开放的特性的情况下,使得这些技术更容易被使用。

参考文献:

[1]M. J. Pazzani and D. Billsus. Content-based recommendation systems [A]. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, The Adaptive Web: Methods and Strategies of Web Personalization,volume 4321 of Lecture Notes in Computer Science [C]. Springer-Verlag, Berlin Heidelberg New York, 2007.

[2]J. B. Schafer, D. Frankowski, J. L. Herlocker, and S. Sen. Collaborative filtering recommender systems [A]. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, The Adaptive Web: Methods and Strategies of Web Personalization[C], volume 4321 of Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg New York, 2007.

[3]B. Mobasher. Data mining for web personalization [A]. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, The Adaptive Web: Methods and Strategies of Web Personalization [C], volume 4321 of Lecture Notes in Computer Science. Springer-Verlag,Berlin Heidelberg New York, 2007.

[4]R.Burke. Hybrid web recommender systems [A]. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, The Adaptive Web: Methods and Strategies of Web Personalization [C], volume 4321 of Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg New York, 2007.

[5]B. Mobasher, R. Burke, R. Bhaumik, and C. Williams. Towards trustworthy recommender systems: An analysis of attack models and algorithm robustness [J]. ACM Transactions on Internet Technology, 2007,(4).

[6]S. S. Anand, P. Kearney, and M. Shapcott. Generating semantically enriched user profiles for web personalization[J]. ACM Transactions on Internet Technology, 2007,(4).

[7]S.Niwa, T.Doi, and S.Honiden. Web page recommender system based on folksonomy mining [A]. In Proc.of the 3rd Int'l Conference on Information Technology: New Generations (ITNG'06)[C]. Las Vegas, 2006.

(责任编辑:刘 君)

作者简介:朱其文(1968-),男,安徽怀远人,中煤能源集团大屯公司信息中心工程师,研究方向为信息工程。