DOI: 10.13663/j.cnki.lj.2007.10.008

·海外眺望·

Melvyl 推荐项目

——发展中的图书馆推荐服务

王艳翠 编译

(聊城大学图书馆　山东 252059)

1　导言

在过去的 10 年里，OPAC(图书馆联机检索目录)少有改变并难以为用户所用。图书馆读者期望的信息检索系统应提供的功能和图书馆信息检索系统实际提供的功能之间存在着差距，Melvyl 推荐项目正在探寻一种减少这种差距的方法并验证减少这种差距的可行性。项目小组在 5 个主要方面展开了探索性的工作，这 5 个方面是：相关性排序、自动更正、基于文本发现系统的使用、用户界面策略以及推荐。本文特别关注项目的推荐部分以及潜在延伸的推荐工作。Melvyl 推荐项目有三个显著的特点：a. 它用历史借阅数据产生推荐。b. 项目探索了在联合编目设置中作出推荐的方法。c. 除了集中于学术读者的需求外，项目也探索了专业技术水平如何使用户对推荐系统感到满意做出贡献的问题。

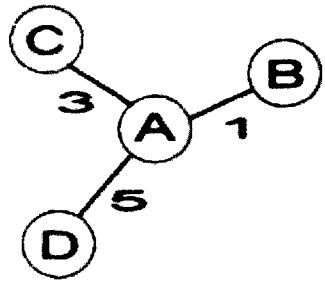
2　探索的推荐方法

Meivyl 推荐项目小组探索了两种生成推荐的方法。第一种方法使用来自 UCLA(加利福尼亚大学洛杉矶分校)的借阅数据来确定条目之间的链接；第二种基于内容的策略是用来自书目记录的术语开发相似条目的询问。

2.1　基于借阅的推荐方法

基于借阅的方法取决于 UCLA 借阅数据的两个重要集合的可获取性。一个集合由 770 万个借阅记录组成，它的时间跨度是 1999 年 7 月到 2004 年 6 月。另一个集合从 2004 年 6 月到 2005 年 5 月由 160 万个借阅处理记录组成。由于 2004 年夏天的系统转换(由 Taos 系统转换到 Voyager 系统)，数据被分成了两个集合。这些数据集合适合于推荐方法探索的原因有三：首先，它们保持了匿名但持续的用户标识号。虽然我们不能识别某一个具体的用户个人，但我们能看到随着时间的推移个人检验的条目之间的链接。其次，处理记录的容量非常大。这点非常重要，因为相对少量的条目被非常频繁的使用，其他多数条目形成了一条偶尔运用的“长尾巴”；相对少量的用户极其活跃，其他大多数用户形成了一条非频繁活动的“长尾巴”，随着时间的推移，大容量数据的积聚使我们更有可能观察我们感兴趣的模式。最后，能把借阅记录与从 UC 联合编目中抽取的实验记录中的书目记录联系起来。

需要注意的是，数据中有一些不能克服的弱点：用户标识号在每个数据集合中是连续的，但在两个集合之间是不连续的；在使用借阅数据的 OPAC 中应用标准联合过滤技术时会出现各种各样的问题，包括数据分布、数据稀疏、客户隐私涉及；借阅数据仅反映物理流通活动，省略了关于使用数字化选择方法的信息。因此，项目没有采用合作过滤的方法，项目小组采用了一种基于加权图模型的简单方法：以书作为结点，以用户共同检索出来的书为边；书被共同检索出来的次数越多，在图表中的边的权重越大；通过沿着同一用户检索出来的到其他条目的边，可以为图表中的任一节点产生推荐；推荐结果可以根据边上的权重迅速加以排序，如下图：



图：为项目 A 做推荐，按次序依次排列是 D、C、B，因为 5 个人也同样检索出了条目 D，3 个人也同样检索出了条目 C，1 个人检索出了条目 B。

这种方法产生了混杂的结果：有一些诱人的、好的推荐，也有一些不切实际、偏离主题的推荐。可能由于大量大学生课程的需求，一些条目被经常的、不

恰当的推荐。项目小组排除了那些借阅频率分布最末端的条目。但采用数值截断忽略了一个事实:一个条目经常借阅可能有许多原因,它可能是一个学科中的基础性工作而被需求,或者它在某一特殊领域内非常有用而受欢迎。项目小组通过限制相同一般内容范围内条目的推荐转而寻求一种过滤策略。

粗略过滤的第一步是使用了索书号种类的首字母。如果显示的索书号以"P"开头,就只推荐以"P"开头的条目。这种过滤方法的结果更有内聚力,但有一定的成本。这种方法的确在大多数情况下排除了不和谐的、不佳的推荐,还有降低交叉学科链接可能性的作用。但在某些领域内索书号种类范畴的编组过粗,其他领域的编组过细;推荐集中于一些尖端领域,其他领域内粗劣的推荐并未被过滤掉。

第二步:通过一般学科范围使用整个索书号创造了编组。在哥伦比亚大学数字图书馆项目已有工作基础上的编组适合于UCLA纪录的空白,并用绘图补充了国会图书馆数据,绘图包括从医学国家图书馆索书号到一般学科范围的相同主题。这些绘图产生了一个内容过滤器,它导致了更加均势的推荐:在早期编组极度细微的地方增加了随意性;对早期粗略的编组增加了限制。

2.2 基于内容的推荐方法

项目小组也试验了第二种产生推荐的方法。这种方法为目标条目分析了书目元数据的内容,选择记录中最重要的项,从而形成一个新的询问。排在最前面的项目是由于新询问作为推荐而被提出。这种方法产生的推荐与基于借阅的推荐方法在性质上有极大的不同。这种方法理论上简单,但变更数量庞大并且程度复杂。有许多对最高项进行选择和排序的方法,还有许多对新询问进行公式化的方法。此外,书目记录是不连贯的,有的记录编目详尽,有的则编目粗略。特别是在编目粗略的记录里,唯一一个主标题的选择可能极大的影响选择和条目的权重,并可能导致意想不到的结果。如:一本书的第二版编目粗略并在主标题上有轻微的区别,就可能产生不同的推荐。假定项目的时间安排加速了,就选择基于借阅的推荐方法来进行产生推荐的用户测试,放弃目前基于内容推荐方法的测试。它可以从多个分析结果中排除大量有标记推荐的复杂性。

3 主要的观察结果

项目小组开展了基于借阅推荐方法的小规模评估。他们与加州大学伯克利分校联合招聘了人文科学和历史学专业的10名大学生和研究生。他们在Rewyl(Melvyl推荐原型)中进行了搜寻,用稍做修改的界面来获取数据。当参与者评价推荐时,他们对定量数据进行了分析,并从收集的调查结果和观察报告中的定性数据里揭示出了几个关键主题:

(1)用户想看到编目中的推荐支持他们的学术研究工作。即使是那些持怀疑态度的人也有兴趣弄明白和尝试推荐服务,直到或除非他们认为推荐质量不佳。

(2)表述是至关重要的。用户有必要理解为何要做推荐服务,他们需要看到充足的元数据来评价一个条目的潜在用途。但是,书目记录不提供在其它设定中极其有用的线索:书的概要或摘录、内容目录和索引目录。

(3)推荐的优先来源是教员(faculty)、书目和脚注。

(4)在支撑学术任务方面,推荐是成功的。大约1/3推荐的条目被认为是建设性的。这些条目在研究过程中如同是媒介资源,用户对他们很感兴趣。推荐帮助用户从一个新的、有力的点来思考学术工作。

(5)推荐可以作为一个询问扩展有效的方法,推荐有效的帮助建立询问的框架。

(6)被推荐的条目通常不具有新颖或令人惊奇的性质。用户偏好倾向于他们描述为"权威性的"或"特殊化的"条目,知识不够渊博的用户也偏好那些概述或全面评述的条目。

(7)用户非常倚重标题和出版时间来评价被推荐条目的有用性。

(8)结果条目和推荐集合可以为不同的角色服务。结果集合与推荐集合的结构之间存在着某种张力。通常,当用户在结果集合中评价条目时,他们寻找新的、未经阅读的条目,并忽视已知的好条目。然而,由于已知有用的条目作为具有相关性"被评审"并可能有与其他有用的或高质量的条目相联系的更多的机会,它们会是潜在的推荐的优秀来源。为了鼓励作为用户产生推荐来源的已知的好条目的使用,有效的用户界面应当允许用户根据最终目的的不同选择条目。

(9)对一个已知的推荐集合和一个已知的推荐条目来说,用户对前者更满意。

4 下一步的工作

推荐系统持续的技术开发能寻求两种非常不同的策略,它们并不互相排斥。一种策略是"用户中立"策略——沿着用户评价推荐而用户又不须具备

持续的知识的发展道路继续前进。第二种策略可以称之为“用户档案(兴趣爱好)”策略，调查持久用户档案的使用以促进合作过滤及大众服务(诸如列表、注释和标记)。除了这些技术开发路径外，另外重要的下一步是隐私权保护调查及在大学信息环境中的实施，因他们与推荐系统的发展及类似的改进有关。

4.1 隐私权与个性化

在图书馆体系中推荐服务发展的一些选择取决于一系列难题，围绕这些难题是为保护用户隐私而设计的图书馆政策。如果我们象 Amazon 和 Netflix 那样，我们能开发更丰富和更个性化的服务，也能做复杂的浏览挖掘和基于个人档案(兴趣爱好)购买习惯的挖掘。但是智力自由原则深深的根植于图书馆文化和整个文化中，并被写在了各级系统的法律和政策中。为了保护图书馆用户、用户风险和由于个性化服务而摆在大学面前的风险、转移这些风险潜在的方法以及对那些可供选择的办法中用户直觉的分析应当伴随给基于用户档案(兴趣爱好)推荐策略的开发，设计了一次大学隐私权保护政策的深入调查：现有政策考虑到隐私权保护和服务条款之间的适当平衡了吗？如果给用户提供了多个可供选择的方法，怎样清楚地传送它们能使用户充分估计风险和潜在的利益？如果用户面临着接受改进的服务就需要泄漏一些隐私的选择，他们会怎样选择？

4.2 用户中立策略的持续开发

现存的基于借阅的方法是用户中立策略：它合并了先前用法模式的信息，但缺少允许回溯检索单个用户的具体数据。虽然能产生一些有用的推荐，但还有明显的、足够的改善余地。小组考虑当前策略的持续改进(包括潜在的改进)，如：A. 包括 FRBR 在内另外编组方法的应用。可以应用许多不同的方法，如：工作层面的聚集借阅统计法用于推荐；或减少推荐集合中重复的出现。B. 为了最大化交叉学科边界容易产生推荐的可能性，在多个内容领域里进行了复合“存储箱”推荐方法的实验。C. 对算法进行修改而不是应用过滤器产生每个结点的主体领域的“权重”；为了平衡人为引起配合的信号与内容相似性的区别，允许推荐集合更加复杂的操作。

然而，未来恰当的、匿名借阅数据的获取必须在开始任何改进之前加以确认。除非图书馆政策和数据处理实践有变化，否则我们继续获取匿名借阅数据库是不可能的。另外，由于越来越多的内容可在线获取，传统物理借阅数据将会逐步的越来越少的代表图书馆用法模式，并且从这些数据中产生的推荐的相关性将会逐步减少。我们能有选择性地把已知的知识应用到新数据的潜在来源。包括网络上可自由获取的课程读物表、论文和书中的参考书目、在线使用统计，每个都提出了不同程度的挑战。考虑的另一个用户中立策略将是匿名、基于时间的个性化的推荐集合。它能持续的改进推荐的一些轻量级临时定制，而不要求持续档案的使用。提供一种基于时间的“书袋子”功能也会为推荐提供数据而不必要求个人可识别的信息。

4.3 基于档案(兴趣爱好)的策略

最后考虑基于持续用户档案维护的方法。一个策略包括真实合作过滤的应用(虽然数据分布和数据稀疏问题仍然可能存在)。探索的第二个领域是存储档案的使用。目的是考虑到更丰富和更持续的推荐集合的定制，用户能根据来自于可构型的选择权选择能想象得到的优化服务。第三个策略是能支持个人和共享资源名单、标记和注释。

5 结论

Melvyl 推荐项目的结果显示了一个有力的事实：图书馆用户对支持学术需求和个人信息需求的推荐感兴趣。在用户测试中，系统产生的推荐结果仅有 1/3 对参与者是有用的。但是参与者几乎一致同意支持发展推荐服务。

进一步发展的可能性大致被分成了两种趋势：用户中立(不要求进入系统用户持续信息的存储)和基于档案(兴趣爱好)(要求一些关于用户模型一定水平的持续的知识)。由于借阅数据的长期可获取性和可用性是不确定的，用户中立策略的进一步发展将可能要求我们把已获得的知识应用到数据的新来源。这些可能包括网上书单、来自于全文论文和书的参考书目、或者记录数据。

基于档案策略必须要包括：围绕改进了的服务和图书馆用户隐私的权衡对图书馆政策和用户观念的分析。技术探索包含合作过滤技术、定置推荐设置的用户控制、个人及共享资源表和标记及注释的存储和挖掘。

支持任何一种选择，特别是那些要求持续用户兴趣爱好的选择，是对于保护用户隐私权政策、用户需求和用户对政策的态度、以及怎样影响推荐服务的发展等是一个详尽的检验。

(编译文献来源：http://www.dlib.org/dlib/december06/whitney/12whitney.html)

王艳翠 女 馆员，山东大学管理学硕士在读。

(收稿日期：2007-01-11 编发：方子丽)