基于用户行为分析的自适应新闻推荐模型*

高琳琦
天津师范大学管理学院 天津 300387

〔摘要〕针对新闻浏览者的偏好易变等特点，通过度量在线用户的点击和阅读行为，依据其不同的阅读策略类型，分析其页面偏好，并综合各页面偏好和新闻偏好，以关键字偏好表的形式表示；然后设计自适应的评分推荐机制，动态地分析用户兴趣及其转移；设计学习机制，根据用户实际阅读的新闻，调整其关键字偏好，并采用模糊相似度来分析用户偏好结构与新闻结构的相似性，从而产生推荐。实验表明，所构造的模型能够提供良好的个性化新闻推荐服务。

〔关键词〕用户行为 需求偏好 个性化推荐 学习策略
〔分类号〕TP311.5

Adaptive News Recommended Model Based on User s' Behaviors Analysis

Gao Linqi
Management School of Tianjin Normal University, Tianjin 300387

〔**Abstract**〕According to the characteristics of Web news browser, such as inconstancies of preference, user's online behaviors are measured. Firstly, user's page preference and news preference are analyzed at the basis of user's reading strategies to form a table of keywords-preferences. Then, the adaptive recommended mechanism is designed to deal with the changes of user's preference. Learning mechanism is also designed to adjust the keywords-preferences of user, based on news actually read by users. At last, fuzzy method is applied to analyze similarity between user's preference and news structure to produce recommendations. The proposed model has been approved to have higher capability than traditional method.

〔**Keywords**〕user's behavior requirement preference personalized recommendation learning strategy

1 引 言

目前，中国网民人数已达到1.23亿，且其对互联网的使用越来越频繁。调查显示，网民平均每周上网达16.5小时，网络已成为网民获取信息的主要途径之一，而浏览新闻是他们主要的网络活动之一[1]。

为了更好地为在线用户提供服务，个性化推荐系统(personalized recommendation system)成为网络信息检索领域的一项重要研究内容。它通过判断用户兴趣，为其选择并推荐适当的信息来解决用户的信息过载和迷失问题。其中，信息过滤是一种被广泛应用的策略。例如，智能新闻过滤组织系统(intelligent news filtering organizational system，INFOS)要求用户对所阅读的每篇新闻进行评价，结果包括“接受”、“拒绝”和“不确定”，以用户的主动反馈为基础来分析其偏好，从而预测新的文章是否符合其兴趣[2]；Anatagonomy系统从用户的操作行为中学习其偏好，并依据用户的评分建立用户轮廓(user profile)，作为推荐的依据[3]；GroupLens提供了一种协同过滤系统，用于收集用户评分数据并将用户分类，因为同类用户应具有相近需求，所以可依据近邻用户的态度为当前用户提供推荐[4]。

但是，用户获取信息的行为有两类：随意浏览(scanning)和特定搜索(focused search)[5]。网络环境中，大多数新闻浏览者的行为属于前者，他们仅仅要了解发生了什么新闻，再找出相关的新闻来阅读，而没有特定的、持续的信息搜索目标。这些用户的兴趣时常改变，要求其经常说明自己的兴趣是不可行的，为文章评分也会增加使用者负担，一旦评分不准确，将导致推荐系统的性能急剧下降。如何为这些占据多数的用户提供良好服务，成为提高新闻网站服务能力的一个重要方面。

2 基于用户行为的偏好分析

个性化推荐主要通过用户需求分析、信息搜索和结果表示等步骤，为用户主动地提供符合其偏好的信息。其中，识别用户需求是良好推荐的基础。目前，针对不同的应用领域，

*本文系国家自然科学基金项目“面向电子商务的顾客偏好分析与个性化推荐系统”(项目编号：70402009)研究成果之一。
收稿日期：2006-11-05 修回日期：2006-12-25 本文起止页码：77-80,71

人们已提出了多种需求分析方法[6]。在新闻网站中，大量的用户是浏览者，要求显式地表达其需求是困难的。但是，可以从用户的浏览行为中，隐式地获取其新闻兴趣的特征，作为推荐的依据。

分析过程包括两个步骤：①页面偏好分析：用户在线行为能够表示该用户的一般特征，部分表达其潜在的兴趣和偏好；②新闻偏好分析：综合用户的各页面偏好，依据新闻结构，以关键字兴趣表的形式表示用户的新闻偏好。

2.1 页面偏好分析

在互联网中，用户浏览新闻标题，如果符合其兴趣，他们将点击并进行详细阅读。因此，可以通过用户选择和阅读的新闻来推断其兴趣。在这个过程中，能够感知的用户动作是对某个链接的点击和在某个页面的停留时间，即用户阅读时间，这成为分析用户兴趣的依据。因此，分析的基础是以下假设[7]：①用户对其感兴趣的新闻进行点击，以详细了解其内容；②用户在某个新闻页面的停留时间越长，说明其阅读越仔细，对该新闻越关注。

用户对某一新闻页面所采取的动作，只有二种情况——点击与未点击，因此用户i对新闻页面j的点击关注度C_{ij}可定义为：

$$C_{ij}=\begin{cases}1 & \text{点击}\\ 0 & \text{未点击}\end{cases}$$

不同用户的阅读能力不同，但其阅读速度是稳定的，只要观察足够久，也是可计算的。在用户日志文件中，保存了用户对每则新闻的阅读时间，并且，每则新闻的长度也容易得到，因此，根据用户阅读的历史记录，可以计算出用户i的平均阅读速度S_i：

$$S_i=\frac{L}{T}$$

其中，L为所阅读新闻的总长度，T为总阅读时间。

对于用户的阅读行为，其关注度可描述为随阅读时间变化的函数，考虑到正常阅读速度和页面篇幅，函数取值不能随时间无限制增加，因而用户i对新闻页面j的阅读关注度R_{ij}定义为：

$$R_{ij}(t)=\begin{cases}0 & t<t_1,\ t>t_2\\ \dfrac{t\times S_i}{L_j} & t_1\leqslant t\leqslant t_2\end{cases}$$

其中，t_1表示最小阅读时间，低于t_1时，认为用户没有阅读该页面，从而减少误操作的影响；t_2是最大阅读时间，超出之后，关注度不再增加，避免用户由于处理其他事情而耽搁所带来的影响。L_j为第j个页面的长度。

综合这两个方面，可以得到在一个交互循环(包括新闻页面点击、进入和退出)中，用户i对该新闻页面j的关注度P_{ij}如下：

$$P_{ij}=\omega_c C_{ij}+\omega_r R_{ij}$$

其中，权重ω_c和ω_r表示这两类动作对评价结果的影响程度，且$\omega_c+\omega_r=1$。

一般地，用户对于感兴趣的新闻将花费较多的时间去阅读、思考等。依据用户阅读速度的不同，可以判断其态度，从而分析其偏好值。下式表示用户i对新闻页面j的阅读速度与其平均阅读速度的比例：

$$Rate_{ij}=\frac{S_{ij}}{S_i}$$

根据阅读相对速度$Rate_{ij}$，可以把用户对新闻的阅读策略分为5种，并为每种策略分配不同的偏好系数(preference coefficient)，如表1所示。

表1 页面偏好系数

阅读相对速度	阅读策略	偏好系数
$Rate_{ij}$>175%	快速浏览：阅读速度超过正常值的75%以上	0.2
125%< $Rate_{ij}$ <=175%	一般了解：快于正常速度25%—75%	0.4
75%< $Rate_{ij}$ <=125%	正常阅读：与正常速度相差±25%内	0.6
25%< $Rate_{ij}$ <=75%	仔细阅读：比正常速度减慢25%—75%	0.8
$Rate_{ij}$ <=25%	详细推敲：比正常速度减慢75%以上	1.0

使用偏好系数，可以修正用户对页面的关注度，得到用户对于该页面的偏好值P_{ij}：

$$P_{ij}=Rate_{ij}\times P_{ij}$$

2.2 新闻偏好分析

根据新闻结构，用户的页面偏好被分解为各关键字偏好，再综合其对各新闻页面的偏好，能够得到用户对新闻的总体偏好，用关键字偏好表的形式表示。

目前，已经存在多种文本结构分析方法，通过内容分析，筛选出文本中的各关键字及其比例，用于表示该文本的特征[8]。如WordNet作为一个著名的英文关键字库，包含了大量的名词及动词，并记载了各字词之间的关系[9]。由于中文的新闻关键字库较为缺乏，在本研究中，我们自行构造了基本的关键字库。

一般情况下，挑选关键字应包括下列几个方面：人物(who)、事件(what)、地点(when)、时间(time)和原因(why)[10]。为便于实现，选择新闻中的角色(role)与主题(topic)作为关键字，以表示新闻的人、事、物等基本要素。不同类型新闻具有不同的特性，其角色范围与主题范围各不相同。例如，对于政治新闻，其角色包括人物、机构、国家等，主题则包括经济、军事、外交、政策等。由于本文的实证研究主要针对体育新闻进行，因此选择球星、球队等为角色，选择运动项目(如足球、篮球等)和比赛名称(如世界杯、亚洲杯等)为主题。

新闻结构包括关键字及构成比例，构成比例以关键字的出现频率表示。设关键字K_i出现次数为N_i，则其构成比例为：

$$W_i=\frac{N_i}{\sum N_i}$$

以二元组集合表示新闻N_j的结构：

$$N_j=\{<K_1,\ W_1>,<K_2,\ W_2>,\cdots,<K_m,\ W_m>\}$$

对用户阅读的每则新闻，查询新闻结构，由页面偏好可以得到用户对于该新闻的关键字偏好结构。设新闻j中出现的

关键字集合为K，用户i对该新闻的偏好为P_{ij}，则用户i对于该新闻的偏好结构为：

$$\{<K_1, P_{ij}^1>, <K_2, P_{ij}^2>, \cdots, <K_m, P_{ij}^m>\}$$

$$P_{ij}^k = W_k \times P_{ij}$$

那么，由用户i所阅读的全部新闻，可以计算其总体的新闻偏好结构：

$$P_i^k = \frac{1}{m}\sum_{j=1}^{m} P_{ij}^k$$

实验中，以关键字表的形式存储用户i的偏好结构，如表2所示：

表2 关键字偏好结构示例

关键字	偏好值
球星A	0.6
球星B	0.2
足球队A	0.5
中超联赛	0.1
……	……

3 自适应推荐机制

一个好的推荐机制不仅要针对个性化的需求，同时应具有学习和自适应能力，以便随着用户兴趣的转移，更好地产生推荐。

3.1 兴趣转移

用户的兴趣是不断发生变化的，既受其个人、社交群体的影响，也受重大新闻事件的影响。因此，其偏好是不断发生转移的。并且，每个人的新闻关注点总是保持在一定的数量之内，不会无限制膨胀，即过去的偏好也需要不断地“遗忘”。

一般地，近期所阅读的新闻总是能够反映其最近偏好，因此我们设置了衰减系数(attenuation coefficient)，用于修正用户的新闻偏好，以适应用户兴趣的转移：

$$C_a = \frac{1}{e^{D/K}}$$

其中，D为当前日期与阅读日期的差值，D越大表明阅读时间越早，C_a就越小；反之，C_a就越大。对于当天阅读的新闻，C_a的取值为1(D=0)。K为选择的常数，其值大于1，且K取值越大，衰减的速度就越小。这样，可以定期对用户的偏好进行修正。

$$P_i^k = C_a \times P_i^k$$

3.2 学习策略

用户兴趣的改变还体现在对于新的新闻角色或主题产生偏好，例如学习过程就是不断调整用户的偏好结构。衰减系数将使用户不再感兴趣的关键字偏好趋向于0，而学习机制就是将新的偏好添加到关键字结构中，并使原有的、被频繁阅读的关键字偏好得到加强。

在系统推荐的新闻与用户真正阅读的新闻之间，存在着4种情况，分别采取不同的学习策略，如表3所示。

表3 学习策略

学习策略	系统推荐的新闻	系统未推荐的新闻
用户阅读的新闻	保持策略	加强策略
用户未阅读的新闻	改进策略	忽略策略

3.2.1 保持策略 如果用户阅读系统推荐的新闻，表明系统掌握了用户此偏好，则采用前述的方法，能够不断加强其偏好。

3.2.2 加强策略 如果系统没有推荐，而用户自行阅读，表明系统不掌握其偏好，或者对其偏好的估计过低而没有推荐。此时，将新增的关键字加入偏好结构中，并赋值为偏好结构表中的平均偏好值。

3.2.3 改进策略 如果系统推荐，而用户没有阅读，表示用户对于此关键字的兴趣已经发生改变，此时可将该偏好值减半。

3.2.4 忽略策略 对于那些未推荐也未阅读的关键字，系统将维持其原来权重，不做修正。

3.3 新闻推荐

推荐即挑选与用户偏好结构相似的新闻，可以视为对用户偏好结构与新闻结构进行相似性分析，采用模糊相似度进行处理。

对于用户尚未阅读的每条新闻，采用下面的公式分析其新闻结构与用户偏好的相似性：

$$\sigma(P_i, N_j) = \frac{\sum_{i=1}^{m}(P_i(K_i) \wedge N_j(K_i))}{\sum_{i=1}^{m}(P_i(K_i) \vee N_j(K_i))}$$

其中，$P_i(K_i)$表示用户i在关键字K_i上的偏好，$N_j(K_i)$表示新闻j中关键字K_i的构成比例。设置阈值t，选择相似度大于t的新闻进行推荐。

该公式满足相似度定义的有界性、单调性和交换律等要求，这一点已得到了证明[11]。

4 实验分析

应用Frontpage 2000和ASP技术开发实验网站，收集有关体育方面的新闻共4 693篇，在本学院研究生中进行实验，采用准确度(precision)和召回度(recall)作为评价标准。并且，用页面预测时常用的一次马尔可夫模型(First-order Markov Model)进行对比实验。

$$\text{precision} = \frac{\text{推荐并被阅读的新闻数}}{\text{所有推荐的新闻数}}$$

$$\text{recall} = \frac{\text{推荐并被阅读的新闻数}}{\text{所有阅读的新闻数}}$$

实验结果如图1和2所示。

在实验中，最小阅读时间t_1取值为3秒，最大阅读时间t_2取值为300秒，K取值为10，权重ω_c和ω_r分别取0.2和0.8。实验结果表示，本文所提出的模型，比一次马尔可夫模型具

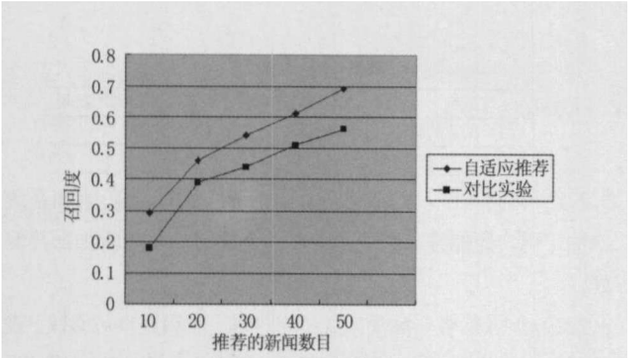

图1 召回度比较

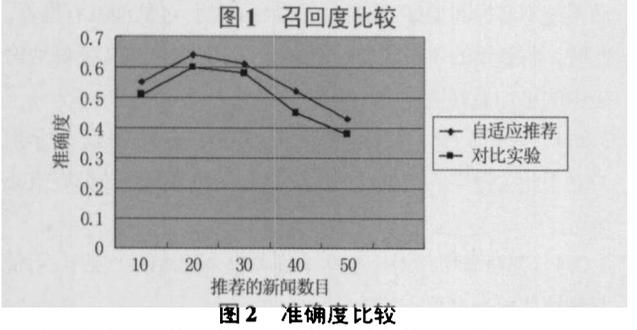

图2 准确度比较

有更高的推荐性能。并且，该算法过程简单，是一种较好的新闻推荐方法。

5 结 语

本文通过分析用户的在线行为——点击和阅读，分析其页面偏好和新闻偏好，以关键字偏好表的形式表示，并设计自适应机制，推荐结果能随着用户兴趣的变化而动态调整。实验表明，所提出的模型能够动态、实时地分析用户兴趣及其转移，提供良好的个性化推荐服务。

在实验中，一些实验者更关注角色，如对于某个球星，除关注其赛事、运动成绩外，对其生活、家庭、爱好等也表现出较大兴趣；而另外一些实验者则更关注主题，虽然他们可以通过学习过程，逐渐适应这些偏好。但这种对不同类别关键字具有显著偏好差异的现象，需要在后续的研究中深入分析，区分主偏好和次偏好，设计针对性更强的推荐机制，以进一步提高推荐的性能。

参考文献：

[1] 中国互联网信息中心. 第十八次中国互联网络发展状况统计报告.[2006-07-09].http://www.cnnic.net.cn.

[2] Mock K J, Vemuri Rao V. Information filtering via hill climbing, wordnet, and index patterns. Information Processing & Management, 1997,33(5):633-644.

(下转第71页)

(上接第47页)

根据单个机构的电子资源管理的工作流程而设计，缺乏对整个联盟工作环节的支持。如何将联盟管理功能扩充到电子资源管理系统中，并能够根据本地化情况进行定制和扩展，是电子资源管理系统需要发展和完善的一个重要环节。

3.2.2 电子书的处理 目前电子资源主要是电子期刊，电子资源管理系统无论是工作流程，还是数据元素，主要都是根据连续出版物的特点而设计。然而随着图书数字化出版比例的加大，其将成为庞大的电子资源管理中的一部分，如何将电子书纳入电子资源管理系统中进行管理，是需要进一步思考的问题。

4 结 语

国外电子资源管理系统的研究虽然还不完善，还有不少地方有待进一步的研究和发展，但已在逐步地走向成熟。而我国电子资源管理系统的研究现处于起步阶段，国内的厂商还没有推出相关的产品，国外商业性电子资源管理系统也刚进入国内市场，目前在国内推出的主要产品是 Ex Libris 的 Verde、Innovative的电子资源管理系统。国内一些图书馆逐步意识到电子资源管理系统的重要性，开始了此方面的尝试。清华大学图书馆是国内首家应用电子资源管理系统的图书馆，也是Ex Libris的合作伙伴，承担了Verde系统的汉化和在中文环境中的开发工作。在这种背景下，回顾和分析电子资源管理系统研究的进程和趋势，可以使我国在发展电子资源管理系统时借鉴国外研究的成果和经验，少走弯路，同时也能跟上电子资源管理系统研究和发展的主流趋势，避免重复建设。

〔作者简介〕 刘 峥，女，1979年生，博士研究生，发表论文8篇。

参考文献：

[1] 廖三三，李浩凌，张春红. 图书馆电子期刊管理与服务的现状及发展. 大学图书馆学报，2005(5):68-71.

[2] Jewell T, Anderson I, Chandler A, et al. Electronic resource management: report of the DLF electronic resource management initiative. Washington D C: Digital Library Federation Council on Library and Information Resources, 2004.[2006-05-20].http://www.diglib.org/pubs/dlfermi0408/dlfermi0408.htm.

[3] Alan R. Keeping track of electronic resources to keep them on track.[2006-06-07].http://www.library.cornell.edu/cts/elicensestudy/pennstate/PSUNASIGPresentation2002.ppt.

[4] Jewell T. Selection and presentation of commercially available electronic resources: issues and practices. Washington D C: Digital Library Federation, Council on Library and Information Resources, 2001.[2006-04-02].http://www.clir.org/pubs/reports/pub99/pub99.pdf.

[5] NISO/DLF workshop on standards for electronic resource management.[2006-05-23].http://www.niso.org/news/events_workshops/NISO-DLF-wkshp.html.

[6] DLF electronic resource management initiative.[2006-05-20]. http://www.diglib.org/standards/dlf-erm02.htm.

[8] Collins M. Electronic resource management systems: understanding the players and how to make the right choice for your library. Serials Review, 2005,31(2):125-140.

的过程进行了整理，建立案例，更新知识库，为以后进行同类工作提供了方便。

表1 方案形成过程

问题编号	劣构问题及其子问题
1	科技信息网的改版方案？(原问题)
1.1	采用什么后台系统？
1.1.1	使用原有的AC系统，还是独立开发一套新的系统？
1.1.1.1	针对目前的使用情况，AC系统有什么优劣势？
1.2	如何把科技信息网做成真正意义上的门户？
1.2.1	如何成倍加大首页信息量？
1.2.2	目前哪些是品牌栏目？
1.2.3	有哪些实用的应用系统可以集成到科技信息网中？
1.3	为了配合改版工作，需要有什么样的人力资源结构？
1.3.1	现有结构是否能基本满足改版和运营工作？还要在哪方面做改进完善？
1.3.2	如何分配人力资源来完成改版工作？

系统选择了B/S结构，采用了Asp.net技术开发，数据库管理系统使用了MS SQL Server 2000。基于Web的应用在界面上更容易让用户接受，同时可以很好地解决因物理工作位置不同带来的问题。图5是系统界面。

图5 系统界面

劣构问题求解是一个复杂的过程，需要大量的信息、知识来支持以及多人协作完成。如果单纯从知识管理的角度来构建知识管理系统，很难满足实际工作的需要，而通过将劣构问题分解为多阶段、多目标、多任务的方式，将知识管理的相关工作结合到解决劣构问题过程中，则可以灵活地处理各种不同的劣构问题。在面向劣构问题求解的知识管理系统中，问题求解和知识管理相辅相成，相互支撑。知识管理的功能辅助问题求解的过程，问题求解过程又反过来支撑知识管理系统的知识收集、创造等，两大模块相互促进，既实现了知识管理，又提高了问题求解的效率。

参考文献：

[1] Jonassen D H. Toward a design theory of problem solving. Educational Technology: Research and Development, 2000, 48(4):63-85.

[2] Schraw G, Dunkle M E, Bendixen L D. Cognitive processes in well-defined nd ill-defined problem solving. Applied Cognitive Psychology, 1998(9):523-538.

[3] 周迪勋. 人工智能——专家系统设计原理. 武汉：武汉工业大学出版社, 1992:45-48.

[4] 丁 崩. 知识管理系统——建立学习型组织的工具. 图书情报工作, 2001(6):5-8.

[5] 李欣苗, 张朋柱. 面向创新任务的知识结构及其动态可视化. 中国信息系统研究与应用前沿. 北京：清华大学出版社, 2005: 488-491.

[6] Tian Q J, Ma J, Liang J Z et al. An organizationaldecision support system for effective R&D project selection. Decision Support Systems, 2005,39(3):403-413.

[7] Jonassen D H. Instructional design model for well-structured and ill-structured problem-solving learning outcomes. Educational Technology: Research and Development, 1997,45(1): 65-95.

[8] 刘儒德. 论问题解决过程的模式. 北京师范大学学报(社会科学版), 1996(1):22-29.

〔作者简介〕 梁凯春，男，1979年生，博士研究生，发表论文9篇；蔡淑琴，女，1955年生，教授，博士研究生导师，发表论文100余篇。

(上接第80页)

[3] Sakagami H, Kamba T. Learning personal preferences on online newspaper articles from user behaviors. Computer Networks and ISDN Systems, 1997,29(8):1447-1455.

[4] Cheung Kwok-Wai. Learning user similarity and rating style for collaborative recommendation. Information Retrieval, 2004,7(6):395-410.

[5] Huber G P. Organizational learning: The contributing process and the literatures. Organization Science, 1997,2(1):88-115.

[6] 吴丽华, 刘 鲁. 个性化推荐系统用户建模技术综述. 情报学报, 2006,25(1):55-62.

[7] 高琳琦, 李龙洙. 基于顾客行为的产品推荐方法. 计算机工程与应用, 2005,41(3):188-191.

[8] Smith K A, Ng A. Web page clustering using a selef-organizing map of user navigation patterns. Decision Support Systems, 2003,35(2):245-256.

[9] Miller G A. WordNet: A lexical database for English. Communication of the ACM, 1995,38(11):39-41.

[10] Kilgour F G. Effectiveness of surname-title-word searches by scholars. JASIS, 46(2):146-151.

[11] 高琳琦. 多智能体营销信息系统的研究. [学位论文]. 西安：西安交通大学, 2001.

〔作者简介〕 高琳琦，男，1970年生，副教授，副院长，博士后，发表论文30余篇。