DOI:CNKI:11-2746/G2. 20111028. 1419. 001

科技文献跨语言推荐模型研究

赖院根

摘　要　信息超载和语言障碍影响我国科研人员对外文科技文献的有效获取,如何提高获取效率成为亟待解决的问题。个性化推荐能很好地处理信息超载现象,但当前国内外相关研究都基于单一语种进行,多语种环境下的推荐研究非常缺乏。本文提出网络环境和海量数据背景下的科技文献跨语言推荐模型,并论证用户兴趣特征抽取、语言转换和混合推荐等步骤。利用 Web 日志挖掘技术,分析基于多种信息行为的整合分析方法抽取用户兴趣特征,以分类表作为参考体系建立用户兴趣表示模型,在用户—特征词转化为用户—类目矩阵的基础上开展推荐研究。图 3。参考文献 17。

关键词　科技文献　个性化推荐　跨语言　用户分析

分类号　G250. 76

Study on Cross-language Personalized Recommendation of Academic Literatures

Lai Yuangen

ABSTRACT　Information overload and language barrier seriously affect the efficiency of acquiring academic literatures in foreign language and how to help users obtain their targeted literatures becomes an urgent problem for digital libraries. It is reported that personalized recommendation systems can well deal with information overload problem, but most of the current researches are based on single language and seldom discusses the methods of recommending academic literatures under a multilingual environment. The paper proposes a framework of cross-language recommendation system of academic articles, and its relevant modules are described in details, including extraction of users' interest feature, language translation and hybrid recommendation etc. This research is expected to expand the research content of personalized recommendation systems and provide technical solutions for improving the efficiency of obtaining academic articles in other languages. 3 figs. 18 refs.

KEY WORDS　Academic literatures.　Personalized recommendation.　Cross-language.　User analysis.

1　引言

外文科技文献对促进科技创新、实现技术跨越发展有着重要意义,为此我国每年花费大量经费用于外文文献资源建设,但当前我国科研人员对外文文献的获取却费时费力。其原因除了语言障碍外,信息超载(Information Overload)是另一重要因素,动辄成百上千个检索结果严重影响了科技文献的获取效率。如何协助科研人员有效获取其感兴趣的外文文献成为我国图书情报界亟待解决的问题。

跨语言信息检索[1](Cross Language Information Retrieval, CLIR)能在一定程度上帮助用户克服语言障碍,但对解决信息超载问题并无太大作用;个性化推荐作为处理信息超载现象的重要手段[2],已在多个领域得到成功应用,但当前研究都基于同种语言进行。迄今为止国内外

通讯作者:赖院根,Email:laiyg@ istic. ac. cn

鲜有与跨语言个性化推荐相关的文献研究。

基于现实需求与研究现状,本文试图通过个性化推荐与跨语言技术的结合来解决我国外文科技文献获取效率低的问题,构建科技文献的跨语言推荐模型并对相关内容进行详细论述。本文对多语种环境下的推荐研究在理论上有望拓展个性化推荐研究方向,在实践中有助于更好地满足我国科研人员的文献需求。

2 研究现状

推荐系统(Recommendation System)是一种为了减少使用者在信息搜寻过程中所附加的额外成本而提出的信息过滤(Information Filtering)机制,依据用户偏好、兴趣和行为等向用户推荐可能需要的信息、服务或产品[3]。推荐系统作为一个独立的研究方向出现在20世纪90年代[4],随后迅速成为学术界和企业界关注的热点,目前已广泛应用于包括网页、音乐[5]、电影[6]、旅游[7]等在内的多个领域。

推荐系统算法有多种类型,包括基于内容(Content-based)、协同过滤(Collaborative filtering)、基于知识(Knowledge-based)、基于人口统计学特征(Demographic-based)、基于效用(Utility-based)和混合推荐(Hybrid recommendation)等,其中前两者应用更为广泛[8]。基于内容的推荐根据用户偏好和项目内容信息之间的相似性进行推荐,适合于机器自动进行内容分析的信息;协同过滤推荐利用用户兴趣的相似性进行推荐,是当前应用最为成功的推荐技术,但却存在数据稀疏性[9]、可扩展性[10]和冷启动[11]等问题。

当前与文本推荐相关的研究大多采用基于内容推荐或混合推荐算法,步骤一般包括[2]:①使用关键词及其权重表示文本,得到文本特征向量;②从用户浏览/评分过的文本中抽取特征词并使用向量空间模型表示,得到用户兴趣特征向量;③根据文本特征向量与用户兴趣特征向量的相似度进行推荐,或采用协同过滤思想利用用户兴趣特征向量计算用户相似度来完成。在这些研究中,基本不考虑语言差异问题,其背后隐藏的假设前提是:推荐文本特征词与用户兴趣特征词语种相同,用户兴趣特征向量中所有特征词属于同一语种。

当用户只对一种语言书写的科技文献感兴趣时,不考虑语言差异的推荐系统(以下简称单语种推荐系统)能很好地工作,但当用户想要获取多种语言的科技文献时,就存在以下问题:①系统无法向用户推荐不同语种的文献;②如果用户在文献检索时使用了不同语种的关键词,或浏览/评分的科技文献涉及多个语种,通过特征词抽取得到的用户兴趣特征向量中可能出现语言混杂现象。举例说明,假设某用户在一段时间内分别浏览了关于"个性化推荐"的中文和英文文献,那么基于这些文献进行特征词抽取建立的用户兴趣特征向量中就会同时出现"个性化推荐"和"Personalized recommendation"等关键词。很显然,如果使用这种语种混合的兴趣特征向量进行个性化推荐,其精度不会太高。

随着国际化进程的发展,对多个语种的科技文献存在需求的现象越来越普遍,这一点对非英语国家的用户来说尤其如此,毕竟当前国际上重要学术期刊大多以英文为载体。对我国来说,有效获取与利用外文文献是促进科技创新、实现技术跨越发展的重要前提条件之一。但在信息超载现象日益严重的现实背景下,外文文献获取费时费力,语言障碍又进一步加大了获取成本。为此,本文将语言因素考虑进推荐模型中,试图通过推荐技术与跨语言技术的结合来应对多语种环境下的个性化推荐问题,以更好地满足用户需求,提高我国科研人员对外文文献的获取效率。

3 跨语言个性化推荐模型

3.1 难点分析

一个完整的推荐系统通常包括用户数据收集、用户兴趣建模和推荐算法匹配等模块[12]。其中,用户数据收集模块负责记录、整理用户数据;用户兴趣建模模块采用合适的模型描述用户偏好,常见的表示模型有向量空间模型、用户—项目评分矩阵等;推荐算法匹配模块根据

用户偏好从资源中筛选出最能满足用户需求的项目并向其推荐。相应地,要实现科技文献的跨语言推荐,存在以下难点:

(1)用户数据的获取。了解用户需求是实现个性化推荐的前提,对用户了解越多越有利于推荐工作的开展。获取用户信息包括显性和隐性两种方式,前者由用户提供信息,能相对准确地反映用户需求,但在实际中很难实施[13];后者由系统自动对用户信息进行分析挖掘,不需要用户参与,能在不增加用户负担的情况下动态分析用户需求,但过程相对复杂[14]。随着网络成为科技文献获取的重要渠道,在虚拟环境中如何获取用户信息并分析其文献需求是个难点。

(2)语言差异问题。从上一节分析可以看出,跨语言推荐在语言差异上需要处理两个问题:用户兴趣模型表示语种(为方便起见,以下简称用户语种)和推荐文献语种(以下简称推荐语种)的不一致,用户兴趣特征词中可能存在的语种混合现象。

(3)用户兴趣表示。利用特征词及其权重表示用户兴趣是文本推荐中的常用方法,受特征词多义性、同义性的影响,用户模型的准确度有限[15]。此时如果采用基于内容推荐算法,不仅维数和计算量过大[16],推荐精度也不高;如果采用协同推荐算法,在海量数据背景下,无论使用用户—文档矩阵还是用户—特征词矩阵来计算用户相似度都存在高维稀疏性问题[17]。采用哪种方式来表示用户兴趣直接关系着推荐系统的实用性和推荐精度。

此外,推荐系统中还存在新用户[2]、可扩展性[10]等问题。本文主要关注如何在网络环境和海量数据背景下实现科技文献的跨语言推荐,对其他问题暂不作探讨。

3.2 模型构建

针对以上问题,本文提出科技文献跨语言推荐模型(见图1),包括特征词抽取、语种统一和混合推荐三个模块,可以简单理解成在常见推荐系统的用户兴趣建模前增加一个语言转换功能,目的是实现用户语种与推荐语种的统一。在推荐算法匹配前实现语种统一有利于充分利用已有的推荐技术。具体步骤包括:①特征词抽取模块。网络环境下,用户通过访问数字图书馆网站来获取科技文献,其在网站上的每一次操作都被服务器自动记录,数据详尽且完备,有利于采用隐性方式了解用户需求。模型中采用日志挖掘技术,通过对用户获取科技文献过程中多种信息行为(包括文献检索、文献浏览和文献下载等)的整合分析来了解用户偏好。

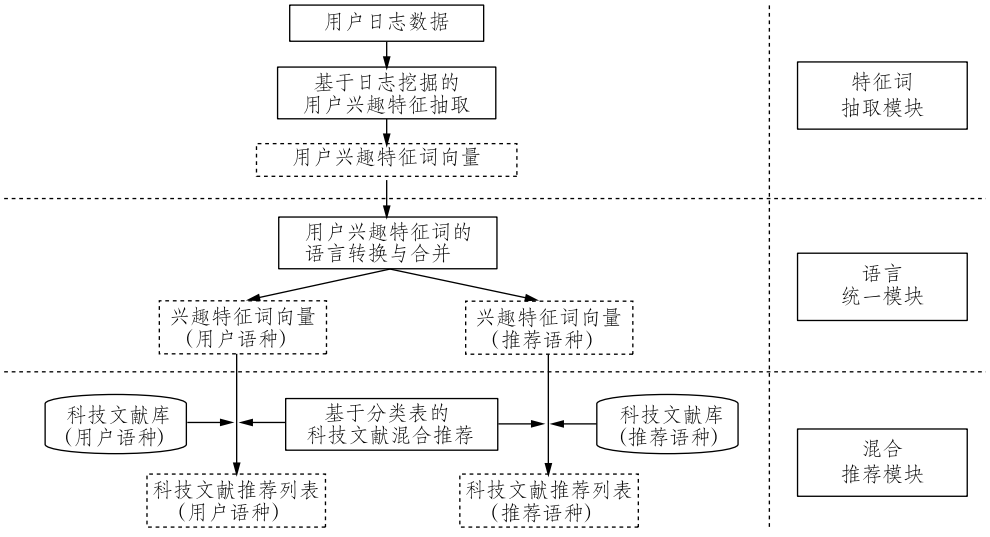


图1 跨语言推荐模型

这样既有利于实现数据获取与用户兴趣模型构建的自动化,也有利于全面了解用户需求及其变化。跟其他文本推荐研究类似,在特征词抽取完毕后使用兴趣特征词向量来表示用户兴趣。需要说明的是,用户兴趣存在多种类型,定义范围可宽可窄,在此仅关注用户对科技文献内容的兴趣。②语种统一模块。虽然可以先将推荐文献翻译成用户语种再进行推荐,但其计算量和成本太大。模型中选择对用户兴趣特征词向量进行语言转换来实现用户语种与推荐语种的统一。该模块的主要功能是将上一模块得到的用户兴趣特征词向量分别用用户语种和推荐语种表示。也就是说,在完成转换后每个用户将拥有两个特征词向量。③混合推荐模块。利用转换后的特征词向量,分别与相对应语种的科技文献库结合来开展推荐工作。因为两者语种相同,相关工作就转化成单一语种下的推荐问题。针对兴趣特征词存在多义性和同义性、推荐计算中数据稀疏性严重等问题,本文利用国内外科技文献都具有相对完善的知识组织体系这一特点,采用基于分类表的混合推荐方法来提高推荐精度。

4 关键技术分析

4.1 基于日志挖掘的用户兴趣特征抽取

4.1.1 兴趣特征词提取

随着信息技术的发展,网络成为用户获取科技文献的主要渠道。在登录数字图书馆网站后,用户利用检索工具查找自己感兴趣的文献,并在缴纳相关费用后将文献下载到本地计算机。整个过程由用户与服务器交互完成,一般不需要服务人员的介入,步骤主要包括登录、检索、下载等,其中能反映用户文献需求的有文献检索、二次文献浏览与全文下载行为。前两者属于信息查寻行为,文献下载属于信息存储行为。

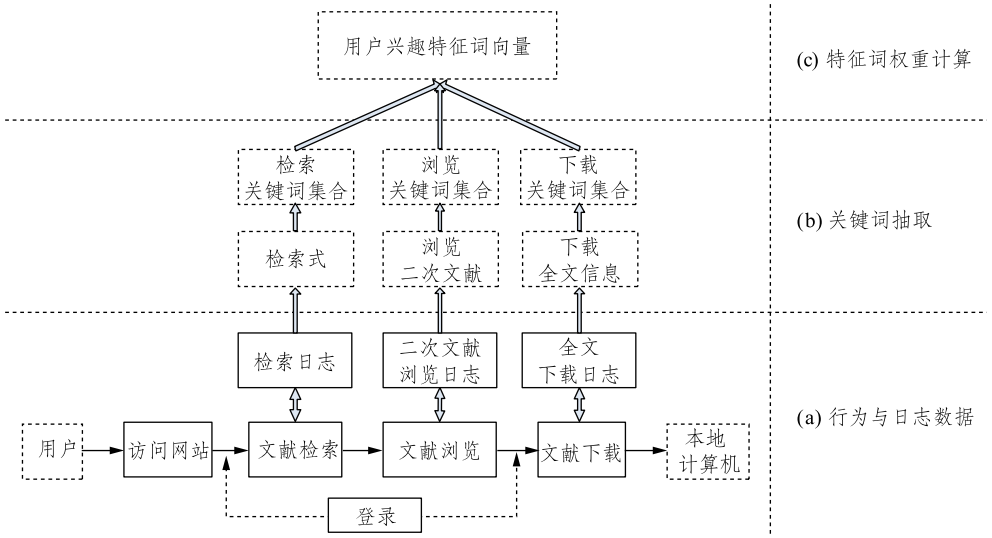


图 2 网络环境下文献获取步骤与兴趣特征抽取

由于用户在数字图书馆网站上的每一次操作都被自动记录,因此可以利用日志数据来分析用户需求。对检索行为,对应的是检索日志,从中可以抽取检索式。对浏览行为,对应的是二次文献浏览日志,其中记录了浏览时间和二次文献 ID 号,通过 ID 号可得到相关联的二次文献信息。同样,对下载行为可得到用户下载的全文信息。虽然检索式、浏览二次文献信息和下载全文信息在一定程度上都能反映用户文献兴趣,本文认为只有将它们整合起来才能更全面地了解用户真实需求。

图 2 展示了网络环境下文献获取步骤与用

户兴趣特征抽取过程。由于涉及多个数据源，本文采取先分别提取关键词然后加权合并的方式来得到用户兴趣特征词向量。关键词抽取过程包括：①根据日志数据得到相应的文献信息（检索式、浏览二次文献信息和下载全文信息）；②从检索式、二次文献信息和全文信息中抽取特征词，分别得到检索关键词集合、浏览关键词集合和下载关键词集合。由于三者在格式、记录字段上差异很大，需要采用不同的方法来提取关键词。对检索式，可以利用正则表达式通过匹配的手段来实现；对二次文献信息和全文信息，有必要结合文献著录项采用自然语言处理和文本分析技术来完成。为了解用户兴趣的动态变化，提取过程中同时记录相应信息行为的发生时间。

4.1.2 兴趣特征词权重计算

文献检索、二次文献浏览和文献下载等行为都能反映用户文献需求，但在程度上应有所差别，因此对关键词进行整合分析时有必要进行加权计算而不是简单汇总。从逻辑上讲，如果一个特征词在检索、浏览和下载关键词集合中都出现，可以认为其能很好地反映用户兴趣，从而需要赋予更高的权值。

在对用户兴趣特征词进行加权合并时，需要考虑以下因素：①检索、浏览与下载行为的权值分配；②信息行为发生的时间。越靠近分析时间点的行为，与其相关的关键词权值越高；③检索式、二次文献信息和全文信息中的特征词个数等。权值计算完毕后，用户需求用兴趣特征词向量 $V(U)$ 表示：

$$V(U) = \{(K_1,\omega_1),(K_2,\omega_2)\cdots,(K_n,\omega_n)\}$$

其中，K 表示特征词，ω 表示特征词权重，n 表示特征词个数。

4.2 用户兴趣特征词的语言转换与合并

4.2.1 兴趣特征词翻译与扩展

借鉴跨语言信息检索技术[1]，可以利用双语词典、多语种叙词表和平行语料来实现用户兴趣特征词向量的语言转换和扩展。由于用户在检索时可能使用多个语种的关键词，浏览与下载的科技文献也可能不属于同种语言，因此向量 $V(U)$ 中的特征词会出现三种情况（在这里仅考虑用户语种和推荐语种两种语言的情形）：①全是用户语种特征词；②全是推荐语种特征词；③用户语种与推荐语种混合。语言转换的目的就是将每个用户的文献需求用两个兴趣特征词表示，其语言分别为用户语种（$V_c(U)$）和推荐语种（$V_f(U)$），比如：

$$V_c(U) = \{(K_{c1},\omega_{c1}),(K_{c2},\omega_{c2})\cdots,(K_{cm},\omega_{cm})\}$$

$$V_f(U) = \{(K_{f1},\omega_{f1}),(K_{f2},\omega_{f2})\cdots,(K_{fp},\omega_{fp})\}$$

其中，K 与 ω 表示特征词及其权重，m 和 p 是特征词个数（在经过转换扩展等过程后，两个向量的特征词个数未必相同）。因此，本文的语言转换实际上是一个互译的过程，这一点与跨语言信息检索有所不同。

跟跨语言信息检索类似，语义消歧和未登录词处理是语言转换中的技术难点。对此可以利用用户检索、浏览和下载的文献内容来提高转换精度，例如：①用户兴趣特征词之间会存在一定的语义关系，利用特征词共现统计来协助翻译项的选择；②对于特征词语种混合的情况，把特征词的相互对照、翻译后的特征词是否出现在用户浏览的二次文献或下载的全文信息中作为判定条件进行消歧。此外，还可以利用平行语料（比如中文科技文献中普遍存在的中英文摘要）来作为翻译项的限定条件。

4.2.2 同义兴趣特征词合并

在语种混合的用户特征词向量 $V(U)$ 中，可能存在语义相同语言不同的特征词，比如同时存在“个性化推荐”与“Personalized recommendation”两个特征词。在经过特征词互译后，单一语种的向量 $V_c(U)$ 和 $V_f(U)$ 中可能出现完全相同的特征词，此时需要对它们及其权重进行合并，必要时进行归一化处理。

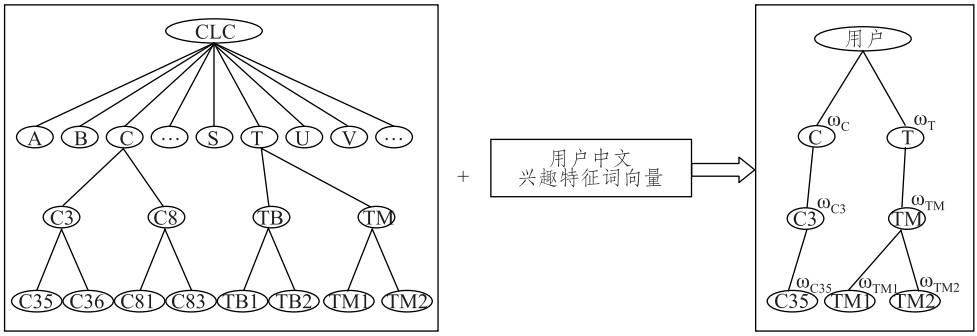
4.3 基于分类表的混合推荐

跟网页、新闻等文本资源相比，科技文献有着相对完善的知识组织体系。本文利用分类表来开展科技文献的个性化推荐，目的是消除特征词同义性和多义性影响，降低数据稀疏程度，提高推荐精度。

4.3.1 基于分类表的用户兴趣表示

利用分类表来表示用户兴趣主要基于以下考虑:①直接应用用户兴趣特征词向量进行推荐,无论是采用基于内容还是协同过滤推荐技术,都存在计算量大、推荐精度不高的问题;②分类表在中文、外文科技期刊中应用非常普遍,其中《中国图书馆分类法》(以下简称 *CLC*)是中文期刊中应用最普遍的分类体系,杜威十进分类法(*Dewey Decimal Classification*,以下简称 *DDC*)是世界上使用最广泛的分类法;③分类表以概念逻辑和知识分类为基础,经过特征词与类目之间的映射后有助于消除特征词同义性和多义性的影响;④分类表类目数量远小于科技文献和兴趣特征词数量,以类目表示用户兴趣等同于将用户—特征词矩阵转换为用户—类目矩阵,能达到降维的效果。

基于分类表的用户兴趣表示示例见图 3,其实质是把分类表作为参考体系,通过兴趣特征词跟类目之间的映射,用类目及其隶属度来表示用户需求。具体步骤包括:①特征词类目隶属度计算,特征词与分类表类目之间关系的强弱用隶属度表示。由于分类表为树形结构,计算时需要把类目层次、类目上下位关系等因素考虑在内;②用户兴趣类目隶属度计算。用户兴趣特征词向量中含有多个特征词,将每个特征词的权值 ω 与其类目隶属度结合起来可得到用户感兴趣的分类表类目及其隶属度。将用户的中文兴趣特征词向量 $V_c(U)$ 和 CLC、外文兴趣特征词向量 $V_f(U)$ 和 DDC 分别进行映射,得到用户的中文、外文类目表示模型 $C_c(U)$ 和 $C_f(U)$(示例见图 3 中的"用户类目表示模型样例",类目右上角的字符代表用户兴趣类目隶属度大小)。



CLC树形结构样例　　用户类目表示模型样例

图 3　基于 CLC 的用户兴趣表示示例

4.3.2 混合推荐

利用用户类目表示模型开展科技文献的个性化推荐,步骤包括:①基于类目表示模型的用户相似度计算,可以采用经典的最近邻(K-Nearest Neighbor,KNN)方法,也可利用表示模型的树形结构,在同时考虑类目层次、类目分散度、类目重合度等因素的情况下借鉴基于距离的概念相似度计算方法来进行;②将相似度大于某一阈值的用户视为近邻用户,根据这些近邻用户的文献浏览、下载记录进行推荐。

在计算用户相似度时,如果用户希望获得中文科技文献推荐,利用中文类目表示模型 $C_c(U)$ 进行;如果想获得外文推荐,则利用外文类目表示模型 $C_f(U)$ 来完成。在对推荐文献排序时,除了考虑近邻用户的浏览、下载时间及次数外,必要时结合科技文献著录项(如文献发表时间、收录影响因子、被引次数等)以更好地满足用户需求。分类表、文献著录项的引入还有助于改善当前推荐模式单一的状况。

对推荐过程中可能出现的几种现象:①没有相似性用户或用户相似度都小于设定阈值;②近邻用户浏览或下载的文献全部属于同一语

种,可以将用户类目表示模型和用户兴趣特征词向量结合起来实现推荐。举例说明,假设特定用户的近邻用户浏览与下载的文献语种都是中文,那么协同过滤算法就无法向其推荐外文文献。对此先利用其外文类目表示模型 $C_f(U)$ 对推荐文献范围加以限定以减少计算量,然后计算用户外文兴趣特征词向量 $V_f(U)$ 与外文文献特征词向量的相似度,根据相似度高低排序进行推荐。

5 结束语

随着数据资源的急剧增长,信息超载现象日益严重。对我国科研人员来说,获取外文科技文献不仅要克服语言障碍,还经常被淹没于成百上千个检索结果中。作为处理信息超载现象的重要手段,个性化推荐研究近十年来受到学界和企业界的强烈关注,但当前研究都基于同种语言进行。本文对多语言环境下的个性化推荐展开了研究,提出了有针对性的研究框架和技术方案,试图通过推荐技术与跨语言技术的结合来弥补当前研究的不足,并为提高我国科研人员对外文科技文献的获取效率提供研究思路。

本研究在理论上有望拓展个性化推荐的研究方向,但对许多具体问题并未开展深入讨论,比如基于多种信息行为的用户兴趣特征整合模型、特征词的转换方法、特征词的分类表类目隶属度计算等。这些问题的解决还需要开展大量的实际工作,将在未来的研究中逐一进行探讨。

参考文献:

[1] 刘伟成,孙吉红. 跨语言信息检索进展研究[J]. 中国图书馆学报,2008(1):88-92. (Liu Weicheng, Sun Jihong. Cross-language information retrieval[J]. Journal of Library Science in China, 2008(1):88-92.)

[2] Adomavicius G, Tuzhilin A. Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6):734-749.

[3] Rashid A M, Albert I, Cosley D, et al. Getting to know you: Learning new user preferences in recommender system[C]. Proceedings of the International Conference on Intelligent User Interfaces, 2002:127-134.

[4] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12):61-70.

[5] Yoshii K, Goto M, Komatani K, et al. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(2):435-447.

[6] Miller B N, Albert I, Lam S K, et al. MovieLens unplugged: Experiences with an occasionally connected recommender system[C]. Proceedings of the 8th International Conference on Intelligent User Interfaces, 2003:263-266.

[7] Sebastia L, Garcia I, Onaindia E, et al. e-Tourism: A tourist recommendation and planning application[C]. Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence, 2008:89-96.

[8] Rodriguez R M, Espinilla M, Sanchez P J, et al. Using linguistic incomplete preference relations to cold start recommendations[J]. Internet Research, 2010, 20(3):296-315.

[9] Sarwar B M, Karypis G, Konstan J A, et al. Application of dimensionality reduction in recommender system—a case study[C]. Proceedings of ACM 2000 KDD Workshop on Web Mining for e-commerce-Challenges and Opportunities, Boston, MA, 2000.

[10] Gábor Taká cs, István Pilászy, Bottyán Németh, et al. Scalable collaborative filtering approaches for large recommender systems[J]. Journal of Machine Learning Research, 2009(10): 623-656.

[11] Rosa M. Rodríguez, Macarena Espinilla, Pedro J.

Sánchez, et al. Using linguistic incomplete preference relations to cold start recommendations[J]. Internet Research: Electronic Networking Applications and Policy, 2010, 20(3): 296 - 315.

[12] 刘建国,周涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1 - 15. (Liu Jianguo, Zhou Tao, Wang Binghong. Review on personalized recommendation system [J]. Progress in Natural Science, 2009, 19(1): 1 - 15.)

[13] Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites[J]. Machine Learning, 1997(27): 313 - 331.

[14] 宋丽哲,牛振东,宋翰涛,等. 数字图书馆个性化服务用户模型研究[J]. 北京理工大学学报,2005,25(1):58 - 62. (Song Lizhe, Niu Zhendong, Song Hantao, et al. Study on the user profile of personalized service in digital library [J]. Transactions of Beijing Institute of Technology, 2005, 25(1): 58 - 62.)

[15] 许欢庆,王永成. 基于加权概念网络的用户兴趣建模[J]. 上海交通大学学报,2004,38(1): 34 - 38. (Xu Huanqing, Wang Yongcheng. User modeling based on weighted concept network[J]. Journal of Shanghai Jiaotong University, 2004, 38(1): 34 - 38.)

[16] 陈基漓,牛秦洲. 用户兴趣模型在图书馆个性化推荐服务中的应用[J]. 情报杂志,2009,28(5):190 - 193. (Chen Jili, Niu Qinzhou. Application of library's personal recommending based on user's interest model[J]. Journal of Intelligence, 2009, 28(5): 190 - 193.)

[17] 颜端武,罗胜阳,成晓. 协同推荐中基于用户—文档矩阵的用户聚类研究[J]. 现代图书情报技术,2007(3):25 - 28. (Yan Duanwu, Luo Shengyang, Cheng Xiao. Toward user-document matrix based user clustering for collaborative recommendation[J]. New Technology of Library and Information Service, 2007(3): 25 - 28.)

赖院根 中国科学技术信息研究所信息资源中心助理研究员,博士。通讯地址:北京市复兴路15号。邮编:100038。

(收稿日期:2011-04-28)