基于领域本体的数字图书馆信息过滤模型研究*

易　明　王学东

摘　要　数字图书馆传统信息过滤技术有很大的局限性。基于领域本体的数字图书馆信息过滤模型最大的特点在于它保留了概念之间以及概念属性之间的关系,能够在复杂语义层次进行逻辑推理。该模型实现的关键问题在于基于领域本体的资源评价值转化和基于领域本体的匹配。图1。表2。参考文献9。

关键词　数字图书馆　信息过滤　领域本体

分类号　G250.76

ABSTRACT　After analyzing the limitations of traditional information filtering technologies of digital libraries, the authors propose an information filtering model of digital libraries based on domain ontology, and then discuss its advantages and the key problems concerning its implementation. 1 figs. 2 tabs. 9 refs.

KEY WORDS　Digital library.　Information filtering.　Domain ontology.

CLASS NUMBER　G250.76

1　数字图书馆传统信息过滤技术的局限性

针对数字图书馆“信息过载”的问题,如何帮助用户滤除与兴趣无关的资源已成为当前研究的重点课题。近几年,在国外兴起的信息过滤技术成为解决这一问题的重要手段。目前,信息过滤技术主要分为两类:一类是基于内容的过滤;另一类是协作过滤。

基于内容的过滤假定每个用户是相互独立操作的,因此,过滤的结果只取决于资源与用户兴趣模型的匹配程度,即利用资源与用户兴趣的相似性来过滤资源[1]。系统通过学习用户评价过的资源特征来获得对用户兴趣的描述。这种技术的优点是简单、有效,缺点是难以发现用户新的兴趣,只能发现和用户已有兴趣相似的资源。另外,从实现方法来看,基于内容的过滤通常利用关键词来表征资源,进而基于关键词来描述用户兴趣。然而,关键词无法深层次地揭示资源所涉及的各种对象之间的复杂关系,如数字图书馆中的图书、作者和出版社之间的关系就会被丢失。由此,这种方法所描述的用户兴趣模型存在很多盲区,而一些有价值的资源就可能被错误过滤。

协作过滤的出发点在于任何人的兴趣不是孤立的,而是处于某个群体中[2]。这种技术的关键是根据用户对资源的评价进行用户聚类,进而依据与用户兴趣最为相似的用户组的共同兴趣来判断该用户的兴趣。其最大优点是能够发现用户新的兴趣,而且由于不依赖于资源内容,不仅适用于文本资源,而且还可以广泛应用于多媒体资源。但是,从方法层次来看,协作过滤技术并没有对资源作更为细致的表征,始终是基于资源层次来描述用户兴趣,最终所生成的用户兴趣模型的盲区会更多。同时,也导致该种技术面临一些难以解决的问题[3]:①“稀疏性”问题,即如果用户一般都只对很少的资源进行评价,那么整个数据阵将变得非常稀疏,这种情况带来的问题就是用户间相似性的比较不

*　本文系教育部人文社会科学研究青年基金项目“面向用户的点击流信息资源开发与利用研究”(项目批准号:08JC870005)的研究成果之一。

准确;②“冷开始”问题,又称新资源问题,即如果一个新资源没有用户评价,那么这个资源就往往被系统过滤了,无论它对当前用户是否有价值;③“灰色绵羊”问题,即一位用户游离于不同用户组之间,无法对该用户的兴趣进行准确定位;④“可扩展性”问题,即随着用户和资源的增多,系统性能会越来越低。

同时,基于内容的过滤和协作过滤都不能实现领域之间的相似性比较。主要原因是,在不同的领域,资源的表示方法很可能是不同的,从而增加了跨领域相似性比较的难度。比如,在描述图书时,就可能不会采用与电影相同的表示方法。然而,用户的兴趣在某个范围内是基本一致的,比如喜欢科幻图书的用户可能对科幻电影也感兴趣。

此外,有学者综合基于内容的过滤和协作过滤两种技术的优点,提出了在数字图书馆中采用基于混合模式的信息过滤模型[1]。虽然这种混合模式在一定程度上能够提高信息过滤系统的性能,但是两种技术自身所存在的问题还是没有得到根本解决。

2 基于领域本体的数字图书馆信息过滤模型

领域本体是用于描述指定领域知识的一种专门本体,它给出了领域实体概念及相互关系、领域活动以及该领域所具有的特性和规律的一种形式化描述[4]。领域本体确定了该领域内共同认可的概念的明确定义,通过概念之间的关系描述了概念的语义,这使得用户之间以及用户与机器之间的交互不仅能够基于语法层次,而且还可以基于复杂的语义层次。将领域本体应用到信息过滤中,可以有效弥补传统过滤技术的诸多不足。图 1 描述了基于领域本体的数字图书馆信息过滤模型,该模型揭示了三种过滤函数的生成方法。其中,过滤函数①和②的产生,类似于协作过滤,过滤函数③的生成,类似于基于内容的过滤。

首先,利用本体理论构建数字图书馆领域本体。这一环节是模型的基础。其次,依据不同用户对资源库中相关资源的评价值(如果对某种资源的评价为空,即用户没有评价,则需要进行一定的技术处理),对用户进行聚类,形成 k 个用户组,使得用户兴趣的相似性在同一用户组之间最大化,而在不同用户组之间最小化,并利用每个聚类的质心点矢量来表征该用户组对资源库中相关资源的评价值[5]。再次,利用数字图书馆领域本体,将不同用户组对资源库中相关资源的评价值转化为对概念集中不同概念的评价值。同样,用户 A 对资源库中相关资源的评价值也可以转化为对概念集中不同概念的评价值。最后,基于对不同概念的评价,将用户 A 与不同的用户组进行匹配,找到与用户 A 兴趣最为相似的用户组,并利用该用户组对各种概念的评价值来判断用户 A 的兴趣,进而生成过滤函数①。

当然,还可以考虑利用数字图书馆领域本体,直接将不同用户对资源库中相关资源的评价值转化为对概念集中不同概念的评价值,然后再对用户进行聚类,形成 k 个用户组,并利用聚类的质心点矢量来表征该用户组对概念集中相关概念的评价值。同样,基于对不同概念的评价,将用户 A 与不同的用户组进行匹配,找到与用户 A 兴趣最为相似的用户组,并利用该用户组对各种概念的评价值来判断用户 A 的兴趣,进而生成过滤函数②。

此外,还可以依据用户 A 对概念集中不同概念的评价值,直接在数字图书馆领域本体中寻找相似或相关的概念及其实例,以形成对用户 A 兴趣的判断,进而生成过滤函数③。

过滤函数产生之后,数字图书馆信息过滤系统便可以帮助用户 A 滤掉没有价值的资源。同时,该模型还会利用反馈功能进行学习,不断优化过滤函数。

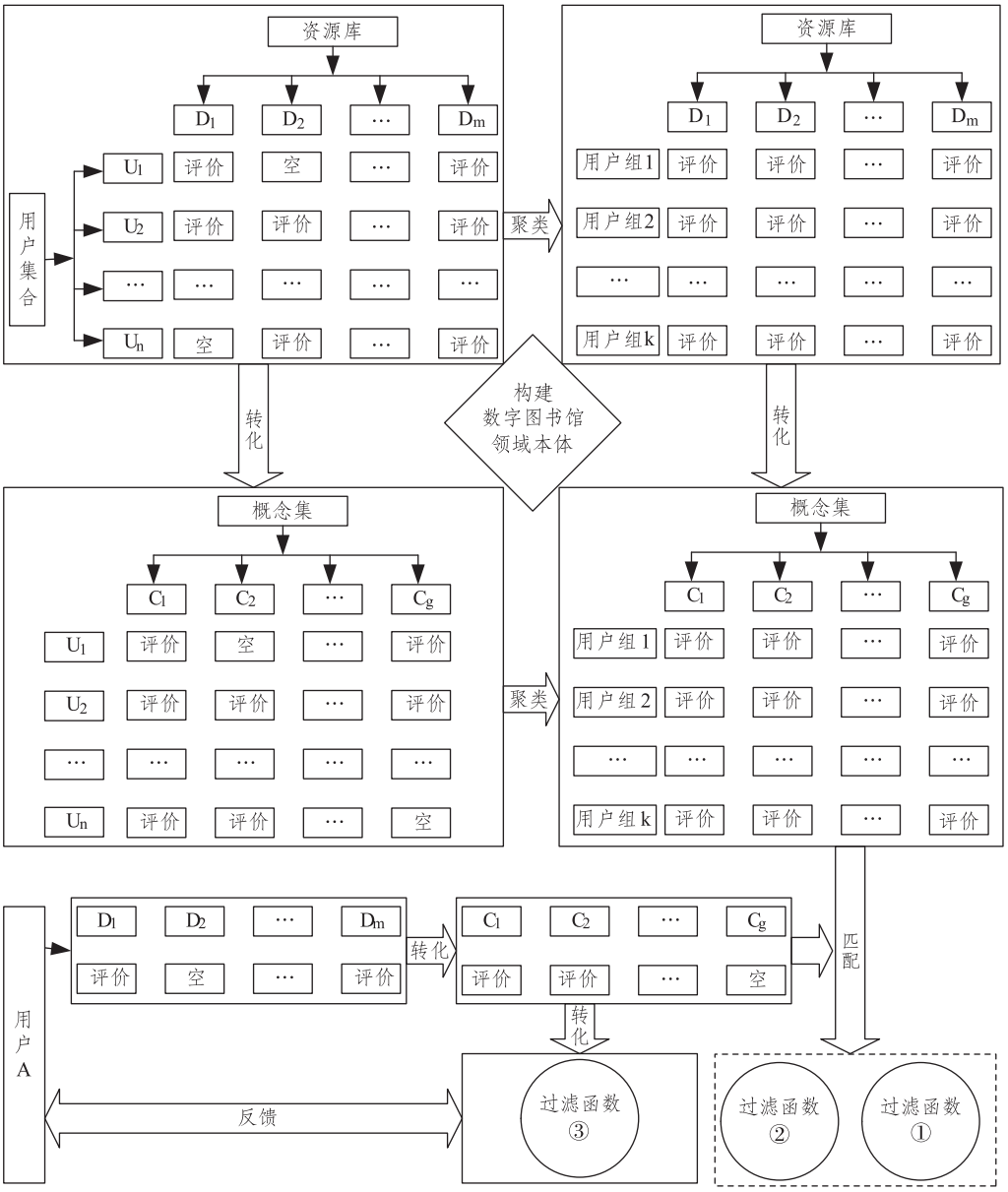

图 1　基于领域本体的数字图书馆信息过滤模型

3　基于领域本体的数字图书馆信息过滤模型的优势

基于领域本体的数字图书馆信息过滤模型最大的特点,在于它保留了概念之间以及概念属性之间的关系,能够在复杂语义层次进行逻辑推理。由此,相对于传统的信息过滤技术而言,它具有以下优势:

第一,基于内容的过滤主要是利用关键词来揭示资源库中资源的特征,协作过滤中没有对资源作任何层次的细分,而基于领域本体的

过滤将资源以概念及其关系的形式进行表示，由此形成的用户兴趣模型能够更深刻地描述用户兴趣，进而减少过滤函数的盲区。比如，在一个包含网络课程资源的数字图书馆中，用户的历史评价已经表明他对 Java 感兴趣，如果基于领域本体进行逻辑推理，那么用户的兴趣应当包含“学习 Java 课程所必需的先行课程”和“最好的 Java 课程老师”等方面，而不是简单地理解为“Java 这门课程”。

第二，基于领域本体的过滤可以减轻协作过滤中的“稀疏性”问题，因为它允许用户之间、用户和用户组之间进行模糊匹配，提供了更大的灵活性。用户之间、用户和用户组之间相似性的比较可以是基于有相似属性的不同对象（例如，同一作者的不同小说，同一类型的电影和图书）。由此，在原始评价数据比较稀少的情形下，利用协作过滤方法计算用户之间、用户和用户组之间的相似性，其准确度相对较低，如果能够利用部分资源的语义属性进行相似性比较，无疑可以提高其准确度，进而减轻“稀疏性”问题。

第三，基于领域本体的过滤可以依据用户对概念集中不同概念的评价值，直接在数字图书馆领域本体中寻找相似或相关的概念及其实例，由此可以解决协作过滤中的“冷开始”问题。比如，当用户对某本图书做出了较高的评价后，基于领域本体的过滤模型可以利用该图书的类型、作者以及出版社等属性来推导该用户的兴趣（例如，同一出版社的相关图书、同一类型的其他图书）。由此，基于领域本体的过滤模型可以直接分析当前用户的兴趣，即使某本图书没有被任何用户评价过，只要它属于当前用户的兴趣范围，就不会被过滤掉，进而避免“冷开始”问题。

第四，协作过滤中主要是依据用户对资源的评价来进行相似性的比较，在实际应用过程中，可能会产生“灰色绵羊”问题。而基于领域本体的过滤，将不同用户（组）对资源库中相关资源的评价值转化为对概念集中不同概念的评价值之后，能够利用概念之间以及概念属性之间的关系全面深刻地揭示用户（组）的兴趣，以此为基础进行的相似性比较也将更为准确，进而可以在很大程度上避免“灰色绵羊”问题的出现。

第五，基于领域本体的过滤能够实现跨领域的匹配，有效弥补了传统过滤技术在该方面的不足。当用户关注属于数字图书馆领域本体的一个子领域（次本体）时，通常由于兴趣的相关性，该用户可能对该领域本体下的其他子领域同样感兴趣。例如，用户当前关注的分支领域为图书时，由于图书、电影同属于一个数字图书馆领域本体，基于领域本体的过滤模型也可以推断该用户在电影分支领域的潜在兴趣：当用户对战争题材的图书感兴趣时，可以认为该用户也可能会对战争题材的电影感兴趣。从技术角度来看，正是由于两个子领域同属于一个领域本体，它们在表示方法上就存在某种程度的相似性，这就为跨领域的匹配提供了可能[6]。

4 基于领域本体的数字图书馆信息过滤模型实现的关键问题

考虑到过滤函数①、②、③的生成方法在本质上是一致的，为了便于研究，本文重点分析过滤函数①的相关问题。

4.1 基于领域本体的资源评价值转化

假设数字图书馆有 n 个用户，资源库中有 m 种资源，其领域本体的概念集中有 g 个概念。依据基于领域本体的数字图书馆信息过滤模型，过滤函数①的生成，需要首先对用户进行聚类以形成 k 个用户组，并利用每个聚类的质心点矢量来表征该用户组对资源库中相关资源的评价值。由此，给定了用户组质心点矢量之后，利用数字图书馆领域本体，通过在每种资源中抽取实例，可以将 k 个用户组对资源库中相关资源的评价值进行转化。经过转化后，将形成 k 个 pr，$pr = \{ <o_1, w_1>, <o_2, w_2>, \cdots, <o_x, w_x> \}$。其中，$o$ 是领域本体中的各个概念的实例，w 为转化后的评价值。

需要强调的是，此时并没有完成资源库中相关资源的评价值向概念集中不同概念的评价

值的转化,因为在每个 pr 中,一个概念可能包含多个概念实例。所以,需要将属于同一概念的实例进行整合,才能将资源库中相关资源的评价值转化为概念集中不同概念的评价值。事实上,这样处理的目的是为了提高信息过滤模型的计算效率。因为对于数字图书馆而言,每个 pr 可能包含成千上万的实例。

由此,问题就转化为如何整合每个概念 c_i 中的各个实例。在实现过程中,可以为概念 c_i 的每个属性 a 赋予一个整合函数 φ_a,利用 φ_a 来完成整合任务[7]。经过整合之后,每个 pr 将转化为对应的 spr, $spr = \{ < so_1, sw_1 >, < so_2, sw_2 >, \cdots, < so_g, sw_g > \}$。其中,$so$ 为概念 c_i 的整合实例,sw 为对应的评价值。同样,可以将用户 A 对资源库中相关资源的评价值转化为对概念集中不同概念的评价值,$UserA = \{ < uo_1, uw_1 >, < uo_2, uw_2 >, \cdots, < uo_g, uw_g > \}$。

表 1 给出了概念“Book”的一个实例集,共有 4 个关于概念“Book”的实例,w_o^{cBook} 为每个实例在概念“Book”中的权重。针对概念“Book”拥有的不同属性,需要构造不同的整合函数 φ_a,进行整合之后,概念“Book”将只有 1 个实例,表 2 提供了一个示例[8]。

表 1　概念“Book”实例集

o^{Book}	w_o^{cBook}	title	author	publisher	year	genre	…
Book1	1	{t1}	{A:1}	{p1}	{2001}	图书→社会科学→文学→中国文学→小说→武侠小说	…
Book2	0.8	{t2}	{B:0.6;A:0.4}	{p2}	{2002}	图书→社会科学→文学→中国文学→小说→军事小说	…
Book3	0.6	{t3}	{C:0.5;B:0.3;D:0.2}	{p3}	{2002}	图书→社会科学→文学→中国文学→小说→社会、言情小说	…
Book4	0.3	{t4}	{D:0.6;A:0.4}	{p2}	{2003}	图书→社会科学→文学→中国文学→小说→史传小说	…

表 2　概念“Book”整合实例

o^{Book}	title	author	publisher	year	genre	…
Book	{<t1,1>, <t2,0.8>, <t3,0.6>, <t4,0.3>}	{<A,0.53>, <B,0.24>, <C,0.11>, <D,0.11>}	{<p1,1>, <p2,1.1>, <p3,0.6>}	{<2001,1>, <2002,1.4>, <2003,0.3>}	图书→社会科学→文学→中国文学→小说	…

4.2　基于领域本体的匹配

将用户组和用户 A 对资源库中相关资源的评价值转化为概念集中不同概念的评价值之后,剩下的一个重要任务就是基于对相关概念的评价,将用户 A 与不同用户组进行匹配以找到与用户 A 兴趣最为相似的用户组,即比较 $UserA = \{ < uo_1, uw_1 >, < uo_2, uw_2 >, \cdots, < uo_g, uw_g > \}$ 和 k 个 $spr = \{ < so_1, sw_1 >, < so_2, sw_2 >, \cdots, < so_g, sw_g > \}$ 的相似性。考虑到语义对象的复杂性,传统的匹配方法不能直接应用于两者相似性的计算,需要进行一定的处理。比如,语

义矢量 *UserA* 和 *spr* 的相似性 $Sim(UserA, spr)$ 应该取决于各概念实例对的语义相似性 $SemSim(uo_i, so_i)$[9]：

$$Sim(UserA, spr) = \sum_{i=1}^{g} \alpha_i \times SemSim(uo_i, so_i)$$

其中，α_i 为概念 c_i 在数字图书馆领域本体中的重要性。同理，各概念实例对的语义相似性 $SemSim(uo_i, o_i)$ 应该取决于每个概念中各属性实例对的语义相似性 $Simlarity(uo_i.a_j, so_i.a_j)$：

$$SemSim(uo_i, so_i) = \sum_{j=1}^{l} \beta_j \times Simlarity(uo_i.a_j, so_i.a_j)$$

其中，l 为概念 c_i 中属性的总数，$uo_i.a_j$ 表示 uo_i 的属性 a_j 的实例，$so_i.a_j$ 表示 so_i 的属性 a_j 的实例，β_j 为属性 a_j 在概念 c_i 中的权重。显然，$Simlarity(uo_i.a_j, so_i.a_j)$ 的计算需要根据各概念属性的不同而采取不同的方法。

参考文献：

[1] 焦玉英，王娜. 信息过滤技术在数字图书馆的应用[J]. 中国图书馆学报，2006(3)：46－49.

[2] 曾春，邢春晓，周立柱. 个性化服务技术综述[J]. 软件学报，2002，13(10)：1952－1961.

[3] 余力，刘鲁. 电子商务个性化推荐研究. 计算机集成制造系统[J]. 2004，10(10)：1306－1313.

[4] 陈刚，陆汝钤，金芝. 基于领域知识重用的虚拟领域本体构造[J]. 软件科学，2003，14(3)：10－13.

[5] Perkowitz M.，Etzioni O. Adaptive web sites：Automatically synthesizing web pages[C]. In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI98). Wisconsin，1998：727－732.

[6] 陈晋进. 基于本体论的个性化信息服务的研究[D]. 湘潭：湘潭大学信息工程学院，2004.

[7] H. K. Dai，B. Mobasher. Using ontologies to discover domain-level web usage profiles[C]. In Second Workshop on Semantic Web Mining at the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02). Helsinki，Finland，2002：61－82.

[8] 易明. 基于 Web 挖掘的电子商务个性化推荐机理与方法研究[D]. 武汉：华中科技大学管理学院，2006.

[9] H. Dai，B. Mobasher. Integrating semantic knowledge with web usage mining for personalization. In Web Mining：Applications and Techniques. [2008-06-11]. http://maya.cs.depaul.edu/~mobasher/papers/DM04-WM-Book.pdf.

易 明 华中师范大学信息管理系副教授，博士。通讯地址：武汉。邮编 430079。

王学东 华中师范大学信息管理系主任，教授，博士生导师。通讯地址同上。

（收稿日期：2008-10-14；修回日期：2008-11-09）

（上接第 84 页）

[12] 唐琼等. 海峡两岸共话创新——海峡两岸图书资讯学研究生论坛召开[OL]. [2008-07-22]. http://sim.whu.edu.cn/newsim/board/show_board_news.php?board_news_id=1202.

[13] 吴钢等. 海峡两岸专家就图书资讯学专业教育达成共识——记海峡两岸图书资讯学系主任联席交流[OL]. [2008-07-22]. http://sim.whu.edu.cn/newsim/board/show_board_news.php?board_news_id=1200.

[14] 徐忆农. 第四次中文文献资源共建共享合作会议综述[J]. 新世纪图书馆，2005(1)：24－25.

[15] 中文文献资源共建共享合作会议模式[OL]. [2008-07-22]. http://hi.baidu.com/wanhh/blog/item/ce5e7981ae6c69d8bd3e1e23.html.

[16] 第六次中文文献资源共建共享合作会议综述[OL]. [2008-07-22]. http://www.gslib.com.cn/lswx/共建共享会议 index.htm.

[17] 沈宝环. 我们要手牵手心连心，一步接一步向国家统一的目标迈进[J]. 图书馆论坛，1995(3)：3－4.

张 正 广州大学图书馆副研究馆员。通讯地址：广州市广州大学城外环西路 230 号。邮编 510006。

（收稿日期：2008-08-19；修回日期：2008-11-23）