●焦玉英　王娜

信息过滤技术在数字图书馆的应用*

摘　要　信息过滤技术将在数字图书馆个性化服务中起重要作用。可以构建一个基于信息过滤技术的数字图书馆模型。它主要包括信息检索模块和信息过滤模块。图 1。参考文献 4。

关键词　数字图书馆　信息过滤　协作过滤　基于内容的过滤

分类号　G250.76

ABSTRACT　*Information filtering technologies will play important roles in the personalized services of digital libraries. In this paper, the authors propose a model of digital library based on information filtering technologies, which includes an information search module and an information filtering module.* 1 figs. 4 refs.

KEY WORDS　Digital library.　Information filtering.　Collaborative filtering.　Content-based filtering.

CLASS NUMBER　G250.76

1　信息过滤技术

随着网络的快速发展,数字信息的爆炸使得用户获取和使用信息非常困难。为了克服信息过载问题,为用户提供个性化和实用化的信息,信息过滤技术应运而生。

1.1　信息过滤的特征

信息过滤技术为区分相关数据和不相关数据提供了工具。信息过滤技术支持用户通过用户文档表达的长期信息需求。在用户文档的基础上,信息过滤系统可以从输入的信息流中过滤出不相关的信息,并且可以减少呈现至用户界面上的信息结果[1]。

信息过滤技术的特征有:(1)信息过滤系统是一个为非结构化和半结构化数据设计的信息系统。(2)信息过滤系统主要用于处理文本信息,并且也处理多媒体信息,例如图像、音频和视频。(3)信息过滤系统涉及大量的数据。(4)信息过滤的应用主要涉及动态的信息流。(5)信息过滤基于个人或群组的信息偏好的描述,这些描述通常被称作文档,这样的文档主要是表述用户长期的兴趣[2]。(6)信息过滤通常意味着一个引入的信息流中的数据移动,而不是在那个信息流中发现数据。

1.2　信息过滤与信息检索间的区别

Belkin 和 Croft 将信息过滤看做是一种特殊类型的信息检索,并且因而指出信息过滤继承了信息检索研究中的问题和研究结果[3]。信息过滤与信息检索的共同目标是选择相关的信息,而且信息过滤系统与信息检索有很多共同的特征。然而,在信息过滤与信息检索间也有一些不同之处。

信息检索主要关注即时用户的特别使用,满足即时的信息需求。而信息过滤关注有重复需求的长期使用的用户的长期信息需求。

在信息检索系统中,固有的问题是提问的频率被认为是信息需求的表述。而在信息过滤系统中,长期的用户需求被描述为用户文档。

信息检索关注一个相对静态的数据库中的文件选择。信息过滤则是从动态的数据流中过滤掉不相关的数据,或是根据用户文档从指定的资源中选择和传递相关的数据。

信息检索系统关注于用户对其同一个单个信息检索过程中的文件之间的交互行为所做出的反应。信息过滤系统则关注一系列的信息检索过程所反映出的长期变化。

信息检索系统服务不了解系统,但可以通过系统提出提问的用户。信息过滤系统的用户则必须熟悉系统,系统存有通常由用户文档所表示的用户模型。

1.3　信息过滤的分类

Malone 认为信息过滤可以划分为两种主要的类型:基于内容的过滤和协作过滤。在实际的系统中,

* 本文系国家自然科学基金项目(70473067)的研究成果之一。

两种类型可以单独使用也可以联合使用。

基于内容的过滤,也被称为认知过滤,是指过滤是基于内容的。基于内容的过滤首先要将信息的内容和潜在用户的信息需求特征化,然后再使用这些表述,职能化地将用户需求同信息相匹配[4]。基于内容的过滤系统在可机读的数据项的基础上容易实现,因此大部分商用的过滤系统都是在基于内容的过滤类型的基础上实现的。

协作过滤,也被称为社会过滤。这种类型支持社会上个人间和组织间的相互联系。协作过滤将人们之间的推荐过程自动化。一个数据条款被推荐给用户,是基于它同其他有相似兴趣用户的需求相关。这种过滤类型对那些不是很清楚自己的信息需求或者表达信息需求非常困难的用户非常有效。

2　构建数字图书馆的信息过滤模型

基于因特网的数字图书馆集中了大量来自多个数据仓库的信息,并将这些在因特网上分布的异构信息资源以一种统一的形式呈现在用户面前。用户对信息的需求是一个长期的过程,并且他的兴趣也会随着时间的改变而改变。怎样更为有效和正确地发现用户的信息需求,怎样满足用户的新的信息需求成了亟待解决的问题。信息过滤技术可以过滤掉和用户无关的信息,成为解决用户需求问题和实现个性化服务的重要工具。通过将信息同用户文档相比较,信息过滤系统可以根据比较结果选择出用户所需要的信息。

将基于内容的过滤同协作过滤比较,可以发现:(1)前者的好处是简单有效,但缺点是难以区分资源的特征与形式,而且它只能发现那些与用户曾经感兴趣的资源相似的资源,而不能发现新的、用户可能感兴趣的资源。(2)后者的优势是能够发现新的、用户可能感兴趣的资源;但是缺陷有两个方面,一个是在使用系统之初表述兴趣的文档不是很有价值,另一个是随着用户和信息资源的逐渐增长,其可行性将会降低。协作过滤可以克服基于内容的过滤的缺陷,因为协作过滤可以更好地理解从不同部分的用户中所特征化提取的用户文档。(3)协作过滤不能取代基于内容的过滤,因为在决定信息的相关性中,用户兴趣所起的作用是主要的。

鉴于两种信息过滤类型各有利弊,本文试图构建一个混合型的过滤模型(将基于内容的过滤同协作过滤联合起来),从而综合两种类型的优势。这个信息过滤模型可以被分为两个主要的模块:基于内容的过滤模块和协作过滤模块。

2.1　基于内容的过滤模块

这个模块主要包括4个基本的组成部分:信息表述子模块、用户需求分析子模块、算法匹配子模块、用户反馈子模块。

信息表述子模块可以从动态的信息集合中分析数据条款,并以一种合适的格式来表述数据条款。这些表述将被传入算法匹配子模块。

用户需求分析子模块可以直接或间接地收集用户的信息和其需求信息,并构造用户文档。用户的文档也将被传入算法匹配子模块。

算法匹配子模块是这个模块的核心成分。它可以根据被选择的算法,将用户文档同被表达的数据条款进行匹配处理,并且决定数据条款是否同用户相关。算法的种类有很多。有些算法是二元的,过滤的结果要么是相关的数据,要么是不相关的数据;但是有些算法是基于概率论的,过滤结果是根据相关性进行排列的数据。在得到结果后,用户可以对结果进行评价选择,并可以进一步反馈其意见。

用户反馈子模块可以通过用户文档,进一步地进行过滤。这个模块首先会收集用户的反馈;然后它将会增加、修改或删除用户的某些信息。影响过滤结果的用户文档不会发生错误。

这个模块的运作流程主要是:将动态信息集合中的被表述的数据,同来自用户需求分析子模块的用户文档进行匹配,并将过滤后的结果推送给用户。

2.2　协作过滤模块

这个模块主要包括6个基本的组成部分:信息表述子模块、用户需求分析子模块、相似用户文档的提取子模块、预测计算子模块、算法匹配子模块、用户反馈子模块。与基于内容的过滤模块比较,在这个模块中有以下两个独有的子模块。

(1)相似用户文档的提取子模块可以分析用户的文档,并且发现相似的用户文档。这个子模块通常用相似性的程度来选择一个小的用户组,在这个用户组中,用户的兴趣都与当前用户相似。然后子模块可以从用户组中提取相似的用户文档。最常用的相似性程度的计算方法是皮尔森相关系数(Pearson correlation 的 coefficient),即通过用户兴趣函数同向量的距离(或是数量的乘积)来反映用户间或各信息条目之间的相似度。这个公式如下:

$$S_{xy}=\frac{\sum_{j\in i_{xy}}(r_{xj}-\bar{r}_x)(r_{yj}-\bar{r}_y)}{\sqrt{\sum_{j\in i_{xy}}(r_{xj}-\bar{r}_x)^2}\sqrt{\sum_{j\in i_{xy}}(r_{yj}-\bar{r}_y)^2}}$$

公式中，i_{xy}表示用户 x 和用户 y 都评价过的信息集合；$\bar{r}_x$ 是用户 x 对信息的评价平均值；r_{yj}是用户 y 对信息 j 的评价值；S_{xy}是用户之间或信息条目间的相似度。

(2)预测计算子模块可以通过对比当前用户的文档与被选择用户组的文档，并通过相应计算来提供推荐列表。预测的方法是：通过结果的兴趣函数来提供推荐列表，这个列表可以反映尚未被当前用户所表达出的兴趣。在使用兴趣函数之前，小用户组的兴趣的平均值必须要先被计算出来。预测喜爱程度的计算公式是：

$$r_{xj}=\bar{r}_x+\Lambda\sum_{y=1}^{n}S_{xy}(r_{yj}-\bar{r}_y)$$

公式中，r_{xj}是系统预测用户 x 对信息 j 的喜好及评价；Λ 是规范因子，它可以是所有的 S_{xy}的和。

这个模块的运作流程是：将用户文档输入用户文档的数据库，然后提取相似用户文档，并对比这些文档从而形成推荐列表，模块最后将会将列表同动态信息集合中被表述的数据项进行匹配，形成最终的过滤结果。

3 信息过滤技术在数字图书馆的应用分析

随着数字图书馆规模的扩张，信息过载变得越来越为严重。信息过滤技术可以根据用户文档，将大量的动态信息进行排列，并提供符合用户需求的信息。应该说，信息过滤技术更有利于进行信息推荐服务，也符合个性化信息服务的思想。因此信息过滤技术可以作为数字图书馆信息推荐服务的一种解决方法。信息过滤可以满足不同背景、不同目的和不同时代用户的信息需求，信息过滤技术将会在数字图书馆的个性化服务中起到越来越重要的作用。

数字图书馆的信息过滤系统可以通过内在的过滤机制，从大量的资源中选择符合用户需求的文档，并且通过友好的方式及时地将文档传送给用户。这样，用户通过信息过滤系统就可以节约宝贵的时间和精力。数字图书馆的信息过滤系统关注用户的长期需求，并高度关注用户的信息服务。信息过滤服务是一种自动的、主动的服务形式，它的目的是使用户更为便利地使用信息、减少用户的操作时间、满足用户的信息需求。信息过滤系统和信息检索系统一样，都是数字图书馆的子系统。将信息过滤系统应用于数字图书馆，我们可以试图构建一个基于信息过滤技术的数字图书馆模型(见图1)。

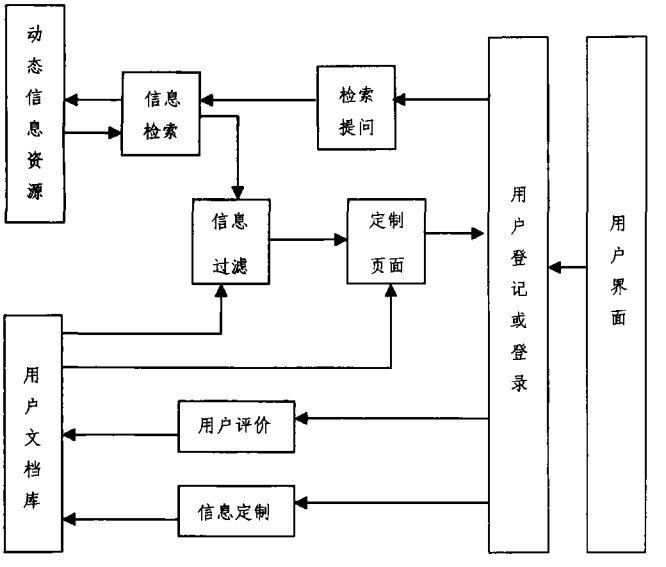

图1 基于信息过滤技术的数字图书馆模型

在这个模型中主要包括信息检索模块和信息过滤模块。信息检索模块的功能是通过用户界面获得用户的信息需求，并使用搜索引擎从信息索引数据库中搜索符合用户提问的信息，最后将已选择的信息推

送给信息过滤模块。搜索引擎中的机器人可以在固定的时间搜索动态的信息资源并抽取新的信息资源,然后对这些新的信息资源建立索引,最后将这些索引输入索引库。这个模型的运作过程如下:

(1)用户通过用户界面和登录(新用户)来登记自己的信息,或是直接登录数字图书馆(老用户)。

(2)用户可以定制自己的信息需求和个性化的页面,或者作为用户反馈评价所获得的信息。这些定制信息和评价信息将被送入用户文档数据库来作为用户文档。这一步可以跳过继续。

(3)用户输入检索提问,这些问题将会被传入信息检索模块。信息检索模块会从动态信息资源中检索出同用户提问相匹配的信息。

(4)被检索出的信息被传送至信息过滤模块。信息过滤模块将会根据在用户文档库中由评价信息和定制信息形成的用户文档,推送个性化的信息到用户的个性化定制页面。

在这个运作过程中,信息过滤模块可以根据用户定制信息将定制化的信息推送给用户。因此,数字图书馆可以通过应用信息过滤技术实现个性化的信息服务。

参考文献

1 Gao Wen, Huang Tiejun. China-US Million Book Digital Library Project. The Proceedings of Digital Library-IT Opportunities and Challenges in the New Millennium, Beijing, China, 2002

2 Nicholas J. Belkin, W. Bruce Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin?. Communications of ACM, vol. 35. no. 12. 1992

3, 4 URI HANANI, BRACHA SHAPIRA, PERETZ SHOVAL. Information Filtering: Overview of Issues, Research and Systems. User Modeling and User-Adapted Interaction 2001(11):203 - 259

焦玉英　武汉大学信息管理学院教授,博士生导师。通信地址:武汉。邮编430072。

王　娜　武汉大学信息管理学院2005级博士生。通信地址同上。

(来稿时间:2005 - 10 - 20)

(上接第9页)

22 Novak, J., Learning, creating and using knowledge: Concept mapsTM as facilitaive tools in schools and^e orporation. 1998, Mahwah, NJ: LEA.

24 Ruiz - Primo, M. and R. J. shavelson, Problems and issues in the use of concept maps in science assessment. Journal of Research in Science Teaching, 1996. 33(6):p. 569 - 600

25 Ford, K. M., et al., ICONKAT: An integrated constructivist knowledge acquisition tool. Knowledge Acquisition, 1991.3(215 - 236)

26 Ford, K. M., et al., Diagnosis and explanation by a nuclear cardiology expert system. International Journal of Expert Systems, 1996.9: p.499 - 506

27 Canas, A. J., D. B. Leake, and D. C. Wilson, Managing, Mapping and Manipulating Conceptual Knowledge: Exploring the Synergies of Knowledge Management and Case-based Reasoning. 1999, Menlo Park, CA: AAAI Press.

28 Zanting, A., N. Verloop, and J. D. Vermunt, Using interviews and concept maps to access mentor teachers' practical knowledge. Higher Education, 2003.46: p.195 - 214

30 Nelson, K. M., et al., Understanding software operations support expertise: A revealed causal mapping approach. MIS Quarterly, 2000.24(3): p.475 - 507

32 Eppler, M. J. Making knowledge visible through intranet knowledge maps: Concepts, Elements, Cases. in Proceedings of the 34th Hawaii International Conference on System Sciences. 2001. Hawaii.

34 陈锐等.知识、知识经济、知识管理.图书情报知识,1999(3)

35 Stoddart, T., et al., Concept maps as assessment in science inquiry learning: A report of methodolody. International Journal of Science Education, 2000.22(12): p. 1221 - 1246

马费成　武汉大学信息资源研究中心主任、教授、博士生导师。通信地址:武汉大学信息管理学院。邮编430072。

郝金星　武汉大学和香港城市大学博士生。主要从事信息经济与信息资源管理、信息系统的学习和研究。通信地址同上。

(来稿时间:2006 - 01 - 17)