信息推荐与信息过滤的概念辨析

陈定权 / 中山大学资讯管理系 广州 510275

摘要：文章分析了信息推荐和信息过滤的概念，并通过几个具体的应用实例来展示它们的不同功能和各自的应用范围。文章最后指出，在中文语境里，应该明确区分信息推荐与信息过滤，并且这种区分是有助于那些没有技术背景的中文读者理解的。

关键词：信息推荐, 信息过滤, 信息检索, 协同过滤

图书情报学领域的很多论著阐述了信息过滤和信息推荐的内容，但不同领域的专家对它们的不同认识造成这些领域的学生很难弄清两者的联系和区别。笔者曾经给多届本科生讲授过信息检索类课程，在授课过程中作者就发现，学生很难理解这两个概念。尽管信息过滤与信息推荐的技术原理几乎如出同辙，但因为学生缺乏相应的技术知识，再加上教材编写过程中引用不同学者的观点却没有合理糅合，甚至是编者自己也没有吃透它们的区别和联系，从而造成学生理解上的困难。本文通过分析信息过滤和信息推荐的共同理论基础，然后再分别介绍各自的思想和应用系统，希望为图书情报及相关专业的学生在学习信息检索相关课程时有所帮助。

1. 信息检索基础

有学者将信息检索最本质的部分概括为：对信息集合与需求集合的匹配与选择[1][2]。通俗一点来说，就是用户提出信息需求，需求表达方式可能是一个或组检索词、也可能是一幅图像、甚至是哼出的一段旋律。信息的集合是经过整理或序化的，信息可以是文本型，也可以是音频、视频、图像、图形，甚至是上述种类信息综合而成的多媒体信息。信息既可以是结构化的信息，如各种文献数据库的记录或 XML 文件，也可以是非结构的信息，如 HTML 页面或图像。经典的信息检索技术主要是检索文本信息，后来为了检索音频、视频、图像等多媒体信息，就通过如下方法来实现：①先将这些非文本信息进行著录，也就是文字性描述，然后再利用文本信息检索的技术来实现检索；②直接基于音频、视频、图像的内容来实现检索。

经典意义上的信息检索所处理的信息集合在一段时间内保持相对稳定不变，用户的信息需求则是不断变化的。例如，搜索引擎系统的有序信息集合在一段时间内保持不变（变化情况取决于系统的更新频率，可能是一个星期也可能是几天），而这些信息集合在网络上接受成千上万的来自不同用户的不同的检索请求。下面将要讲到的信息推荐和信息过滤所面对的信息集合则相对是动态的，而来自用户的信息需求则相对不变或变化甚小，它们都可以看成是经典信息检索的一个的检索任务。

2. 信息推荐

所谓信息推荐（Information Recommendation，以下简称 IR）就是将满足信息需求的信息通过某种方式推荐给相关用户，尤其是将最新的满足需求的信息推荐给用户。推荐的方式

可以是系统主动通知用户，也可以是用户主动去获取。

在信息推荐系统中，信息需求一般被表述为“用户兴趣”。用户如何表达自己的兴趣抑或信息推荐系统如何获取用户的兴趣呢？信息推荐的关键在于如何恰当地描述用户的兴趣，也即是如何构建能够代表用户信息需求的用户需求档（User profile）[3]，用计算机专业术语来说，就是用户建模（User Modeling），用户需求档就是建模的结果或者是用户模型的另一种称呼。用户兴趣可以是由用户显性设定，也可以系统隐性设定。用户显性设定可以是一组关键词，也可以系统内部定义好的用户原型。系统隐性设定主要是根据用户的各种特征或使用行为来学习用户现在的兴趣以及兴趣的转移情况等，这个学习过程是一个渐进的、不断完善的过程。除非用户的兴趣发生变化，否则用户需求档则保持相对固定。

用户需求档既可用于信息推荐，也可用于个性化的信息检索。例如，搜索引擎 Google 提供了个性化信息检索服务，根据用户的检索历史来学习用户的兴趣，从而提供个性化搜索结果。登陆用户起初可能感觉不到有多大差异，但随着搜索历史的积累，个性化搜索结果的效果将会不断得到改善。

传统的基于内容(content-based)的信息推荐的技术与经典的信息检索原理相似，都是信息需求与信息集合的匹配与选择，将满足条件的信息推荐给用户。还有一种技术是协同过滤（Collaborative Filtering）或社会过滤（Social Filtering），它通过分析用户的兴趣，在用户群中找到特定用户的相似用户，综合这些相似用户对某一信息（商品）的评价（可以显性评价也可以时候隐性评价）来形成对指定用户对此信息（商品）的喜好程度的预测[4]。这样，相似用户群构成了一个具有共同兴趣的群体，可以称之为兴趣共同体。当兴趣共同体的某个用户或某几个用户对某信息（商品）很感兴趣的时候，可以预测共同体的其他成员也感兴趣，从而将该信息（商品）推荐给其他成员。尽管“协同过滤”名称上有“过滤”一词，但却用来信息推荐，也就是说，协同过滤是实现信息推荐的一种技术。

基于协同过滤的技术和基于内容的技术可以分开使用，也可以结合在一起使用。下面通过几个信息推荐的实例来说明信息推荐的功能和应用领域。

- 基于协同过滤的推荐系统：美国俄列冈州立大图书馆开发的 SERF 推荐系统、Connotea 社会书签系统、亚马逊网络书店、当当网上商城、卓越亚马逊网上商城网等。
- 基于内容的推荐系统：①Google 快讯：它将符合用户指定主题（一组关键词）的在线新闻通过电子邮件发送；②百度邮件新闻订阅：它将符合用户指定主题（一组关键词）或某类新闻（预先分类好）的新闻通过邮件定时发送；③图书馆提供的新书推荐功能。
- 信息推荐功能可集成在信息检索过程中。例如 Google 在向用户呈现检索结果的时候，会在每个结果后附上“相似网页”来推荐相关信息（一种综合了网页内容和网络评价的复杂算法），而前面提到的各种网上商城则是在用户浏览某条检索结果的时候，向用户推荐相关商品。

3. 信息过滤

在分析信息过滤之前，先看看什么是过滤。《新牛津英语词典》(上海外语教育出版社，2001 年第 1 版)对 filter 的解释为“pass(a liquid, gas, light, or sound)through a device to remove unwanted material”。《现代汉语词典（2002 年增补本）》(外语教学与研究出版社，2002 年第 1 版）对“过滤”的解释为“使流体通过过滤纸或其他多孔材料把包含的固体颗粒或有害成分分离出去”。无论是英文的“filter”还是中文的“过滤”都有“把某物质提纯”或“把不想要的东西或有害的东西去除”的含义。在信息检索上下文中，什么是信息

过滤（Information Filtering）？信息过滤有两种解释：一种是对检索结果提纯（去除用户不需要的结果），使得检索结果更满足用户需求；另一种是将检索结果中包含有各种暴力、色情等有害结果去除。这就出现了分歧。例如文献[5]认为，网络过滤是防止用户访问内容“不合适”的网页，而文献[3]认为，信息过滤就是构建一个能够反应用户信息需求的用户需求档，然后将满足用户需求档的信息反馈给用户，将不满足用户需求档的信息屏蔽掉。所以，国内文献[6]就总结出信息过滤有两种理解，一是软件按照用户的参数设置将不希望进入到本地机的信息阻挡在外面；二是软件系统将大量的动态信息排序，并根据用户需求档的内容将其提供给需要它们或是符合他们要求的用户。如果说前者强调的是不良信息的阻挡，那么后者则侧重于信息推荐服务。所以，很多论文将“信息推荐”纳入到“信息过滤”的范畴也就不足为怪了。

那么哪一种理解更大众化些呢？作者通过如下分析认为前一种解释更符合大众思维习惯。理由如下：我国公安部计算机信息系统安全产品质量监督检验中心 2003 年颁布的《信息技术信息过滤产品安全检验规范》[7]将信息过滤（Information Filtrate）定义为“对网络上的信息内容进行实时分析，对预先定义的非法信息内容进行过滤和拦截”；中国电信为宽带接入用户所提供的绿色上网业务对过滤服务的解释就是“拦截互联网不良信息如黄、赌、毒、邪教等内容”；国内外的各种信息过滤软件也是强调屏蔽一些不良信息。所以本文的信息过滤就是通过某种技术将不良信息屏蔽掉，不呈现给最终用户的过程。

在因特网十分普及的今天，信息交流的渠道和速度已经不可同日而语了，不良信息和垃圾信息对普通网络用户尤其是青少年的危害日益凸现，引起了全社会的重视，都期望在全社会营造一个健康的绿色空间。这样，网络过滤系统就应运而生，但是否安装过滤系统以及怎样设置过滤条件一般由第三方的监管人员来决定，例如孩子的家长、公共机构的负责人、网络论坛的管理员、企业管理者、网络运营商等。实现信息过滤系统首要问题是过滤标准的是什么，也就是什么信息过滤什么信息不过滤的问题。对一个具体过滤系统而言，过滤条件一般会随着过滤标准的变化而发生，但在一定时期内是不变的。

实现信息过滤的技术一般有基于内容的过滤、站点过滤和协议过滤。

- 基于内容的过滤：基于内容过滤方法最常见的是关键词过滤，基于视频、图像内容的信息过滤，因为技术不成熟，大规模应用还很少见。设置者列出一系列需要隔离词语的清单, 然后将所有包括这些词语的信息，如邮件、网页等，不呈现给最终用户。现在有些智能信息过滤系统可以针对主题进行过滤，使得含敏感词汇但不是不良信息的也不会遭到过滤，在一定程度上避免了“误判”。
- 站点过滤：过滤条件的设置者列出可以访问的站点或不可以访问的站点，前者称为白名单，后者称为黑名单。只有在白名单之中的站点才可以访问，条件十分苛刻。除了黑名单中的站点不可访问之外，其余都可以访问，条件比较宽松。
- 协议过滤：设置者将通过某些协议或某些端口号提供的服务或信息彻底屏蔽掉。这种方法不分信息的好坏，一般应用在企业或政府机构当中。严格上讲，协议过滤不属于信息过滤的范畴，防火墙就具备这种功能。

上述三种技术都各有利弊，在实际应用中, 它们一般是结合起来使用。例如，对于图书馆、学校、家庭用户而言，期望为学生和孩子过滤掉不良信息，可能倾向于使用站点过滤和关键词过滤；对于公司用户而言，则可能更倾向于采用协议过滤，期望屏蔽像 MSN、QQ、FTP 等网络功能给公司带来的不利影响。

国内外商家都意识到网络信息过滤的必要性和它的市场前景，纷纷开发出各种不同的过滤软件。例如 CyberPatrol[8]、CYBERsitter[9]、SafeSurf[10]几款软件就可以用在图书馆、企业和学校或家庭，为用户网络冲浪保驾护航。

下面给出几个信息过滤的应用实例来说明信息过滤系统的功能和应用领域。

- 因特网运营商在为用户提供网络接入服务的同时也提供过滤功能。中国最典型的信息过滤应用应该是由中国电信为宽带接入用户所提供的绿色上网业务。该业务可以为用户提供拦截互联网不良信息如黄、赌、毒、邪教等内容的过滤服务，与其他功能一道共同为青少年营造绿色的上网空间[11]。该项业务让家长有权过滤黄、赌、毒、邪教等内容、禁止访问网站的列表等，从而为孩子创造了一个绿色的网络空间，避免受到不良信息的侵蚀。
- 搜索引擎也开始引入了信息过滤功能。世界上最大的搜索引擎 Google 在其系统中也加入了过滤系统，这样用户在检索信息的时候，它会过滤掉一些不良信息。例如，在 Goolge 中输入“色情”检索词，Google 会在结果页面提示“***据当地法律法规和政策，部分搜索结果未予显示***”。这是一种典型的信息检索与信息过滤综合运用的实例。

4. 结束语

基于内容的信息推荐和信息过滤，在本质上与经典的信息检索在技术原理上极为相似，它们的关系犹如一枚硬币的正反面，只不过信息推荐是“取真”的过程，而信息过滤是“取否”的过程。两者之间的异同如表 1 所示。

表 1: 信息推荐与信息过滤比较

	信息推荐	信息过滤
条件提出者	用户自行提出推荐条件。推荐条件有时称为用户兴趣。	一般由监管者提出过滤条件，最终用户一般无权干涉。
主要任务	向用户推荐新的相关信息，可以扩大用户的知识范围。	过滤掉不良信息，为用户净化网络信息空间。
需求变化率	根据用户的兴趣变化而变化，变化率低。	根据不良信息的定义标准来设置过滤条件，变化率低。
功能描述	将满足推荐条件的信息推荐给用户。	将满足过滤条件的信息屏蔽掉，不呈现给最终用户
与信息检索（浏览）的集成	可以实现在检索和（或）浏览的过程中，向最终用户推荐相关信息	在检索和（或）浏览过程中，屏蔽掉不良结果，哪怕该结果满足了最终用户信息需求

尽管信息推荐与信息过滤在技术上有很多相似的地方，但作者认为，在中文语境中我们应该严格区分信息过滤与信息推荐的概念和明确各自的功能，这样严格区分将会有助于那些没有技术知识背景的中文读者来理解。另外，因为信息过滤和信息推荐也是近几年才流行的词汇，在诸如情报学、计算机科学之类的词典中还尚未收录该专业术语，明确信息推荐和信息过滤的概念和各自功能对将来这类词典的编纂也是大有裨益的。

参考文献

[1] 赖茂生, 等. 计算机情报检索[M]. 北京:北京大学出版社, 1993: 3

[2] 叶鹰. 信息检索:理论与方法[M]. 北京:高等教育出版社,2004:4
[3]Ricardo Baeaz-Yates 等著，王知津 等译. 现代信息检索[M]. 北京:机械工业出版社,2005 :17
[4]黄晓斌. 基于协同过滤的数字图书馆推荐系统研究[J]. 大学图书馆馆学报,2006, (1):53-57
[5]Elisa Bertino, et al. Content-based Filtering of Web Documents: The MaX System and the EUFORBIA Project[J]. International Journal of Information Security, 2003, 2 (1): 45-58
[6]黄晓斌. 数字图书馆信息过滤系统初探[J]. 现代图书情报技术, 2004, (6):6-10.
[7]公安部计算机信息系统安全产品质量监督检验中心. 信息技术信息过滤产品安全检验规范[EB/OL].[2007-06-12]. www.mctc.gov.cn/bz/MSCTC-GFJ-06.pdf
[8]CyberPatrol Internet Security Software. [EB/OL]. [2007-7-10] http://www.cyberpatrol.com/
[9]CYBERsitter Official Website[EB/OL]. [2007-7-10]. http://www.cybersitter.com/
[10]SafeSurf [EB/OL]. [2007-7-10]. http://www.safesurf.com/
[11] 广东省电信有限公司. 绿色上网业务介绍 [EB/OL]. [2007-7-10] .http://green.gd.vnet.cn/main.html

作者简介

陈定权，(1974 年－)，男，中山大学资讯管理系，副教授。
通讯地址：广州市新港西路 135 号。
联系方式：13380085107，chendq@mail.sysu.edu.cn

The Differentiation between Information Recommendation and Information Filtering

Chen Dingquan / Department of Information Management，Sun Yat-sen University，Guangzhou,510275

Abstract：This paper analyses the concept of information recommendation(IR) and information filtering(IF), and demonstrates their difference and respective application domain with some application instance. Lastly, this paper argues that IR and IF should be differentiated clearly in the context of Chinese language and this differentiation is helpful to understand these concepts for those Chinese readers who have no technical background.

Key words：Information Recommendation, Information Filtering, Information Retrieval, Collaborative Filtering