网络信息过滤原理与应用

WANGLUO XINXI GUOLÜ YUANLI YU YINGYONG

黄晓斌 等 编著

北京图书馆出版社

网络信息过滤原理与应用

黄晓斌等　编著

北京图书馆出版社

图书在版编目(CIP)数据

网络信息过滤原理与应用/黄晓斌等编著. —北京:北京图书馆出版社,2005.7

ISBN 7-5013-2786-6

Ⅰ.网… Ⅱ.黄… Ⅲ.计算机网络—信息管理 Ⅳ.①G203②TP393

中国版本图书馆 CIP 数据核字(2005)第 037044 号

书名 网络信息过滤原理与应用

著者 黄晓斌等 编著

出版 北京图书馆出版社 (100034 北京西城区文津街 7 号)

发行 010-66139745 66175620 66126153
66174391(传真) 66126156(门市部)

E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)

Website www.nlcpress.com

经销 新华书店

印刷 北京市广内印刷厂

开本 880×1230 毫米 1/32

印张 11.5

版次 2005 年 7 月第 1 版 2005 年 7 月第 1 次印刷

字数 280(千字)

书号 ISBN 7-5013-2786-6/G·620

定价 25.00 元

前　言

随着网络技术的发展和普及利用，如何有效地对网络信息资源进行管理与控制成为比较突出的一个问题。网络信息过滤是根据一定的标准，利用一些手段和方法对某些网络信息进行阻挡，使其不能在一定的范围内传播，而其他的信息仍然可以传播利用。过滤是对网络信息进行有效管理的途径之一，目前国外在这方面的研究取得了一些进展。概括起来，国外有关研究的内容主要包括如下几个方面：首先是技术问题，如各种过滤方法和系统设计等；其次是法律问题，如过滤的法律依据和信息自由关系等；再就是应用性问题，如各种过滤软件的性能特点比较和使用技巧等。随着网络技术的进一步发展，国外对这方面的研究将进一步深入：在技术上更重视自动化和智能化，如研究开发各种自适应的智能代理系统以及自动评价和分级系统等；在法律方面重视研究各种法规的制定和运用；在应用方面着重探讨各种可行性办法，并注意对过滤的实际效果进行分析评估等。应该指出的是，此前我国在这方面的研究不多，缺乏系统性的研究成果。为了适应网络化发展的需要，进一步提高网络信息的服务质量，保护青少年用户的身心健康，有必要加强这方面的研究，尤其是结合我国实际情况探索相应的对策。

本书是在国家社会科学基金项目“图书馆对网络信息过滤问题的研究”和教育部人文社会科学“十五”规划研究项目“网络信息过滤及应用”的部分成果基础上改编与补充而成。这两个项目主要是对网络信息过滤有关问题进行综合研究，采取从一般到具体、从理论方法到技术应用的方式，先对网络信息过滤基本问题进行较为全面的研究，然后再对具体的应用问题进行分析，并在与国外比较分析的基础上寻求切实可行的对策。在研究的指导思想上力求做到：(1)

求新。站在学科发展的前沿,对网络信息服务中的新问题进行新的认识,争取在理论、技术和方法等方面有所创新。(2)求实。联系实际,尤其是针对我国目前的具体情况有的放矢地研究。(3)求深。对一些关键问题进行深入研究,使成果具有一定的深度和较大的参考和利用价值。本课题采用了如下研究方法:(1)文献调查分析法。通过对有关本课题研究的文献调查分析,掌握目前此课题的研究发展状况和动态。(2)问卷调查分析法。设计问卷,抽样调查有关图书馆网络信息服务的情况、采用过滤系统的效果与质量等。(3)案例研究分析法。对公共图书馆、学校图书馆、网吧等场所网络信息过滤的一些问题进行深入研究,重点分析,从而了解有关这方面的问题和特点。(4)实验观察法分析。通过对网络信息过滤系统进行实验观察和可用性分析,发现其不足,提出有关的改进方案。(5)比较分析法。通过对国内外有关问题进行比较研究,探索适合我国信息服务的网络信息过滤有效途径。

本课题研究的实际意义在于可以为有关管理部门提供决策参考,为实际工作提供有关的理论、技术和方法指导。网络信息过滤可以使人们获取特定需求的信息,减少因为信息超载而造成的认知过度,从而提高工作效率,集中精力进行知识创新和管理活动;可以在一定程度上防止不良信息的污染,净化网络信息环境,保护青少年的身心健康;可以减少不必要的信息传播,有效地调节网络信息流,从而提高网络传输效率。理论意义在于探讨网络环境下信息管理的规律。网络信息过滤是网络信息资源管理、网络信息计量学等计算机应用领域研究的重要内容之一,有关的原理、技术、方法的研究有助于形成、完善这些新学科的体系;有关成果可为网络信息服务机构在提供服务、制定有关政策规定时参考,也可以作为实际工作人员在了解此专题的发展动态、采取有关措施时提供依据;有关理论方法和技术等方面的内容可作为信息管理等专业教学内容的补充。

本项目的主持人为黄晓斌(中山大学信息管理系),其他主要成员有:邱明辉(海军兵种指挥学院图书馆)、黄少宽(中山大学社会学

系）、王玉（广东商学院信息管理系）、夏明春（浙江师范大学图书馆）、叶楚旋、贺亚锋（中山大学图书馆）、吴红（中山大学计算科学与计算机应用系）等。本书由黄晓斌拟定大纲和完成主要的内容，邱明辉提供了本书的第1、2、4、9、10章部分内容的初稿，黄少宽、王玉、夏明春分别撰写了第7、9章部分内容的初稿，最后由黄晓斌修改和统稿。

本书在编写过程中参考了国内外一些有关资料，特向作者表示感谢。北京图书馆出版社对本书的出版给予大力支持：宋安莉老师给予极大的鼓励和帮助，并提出了许多很好的意见；其他有关同志也为本书的出版付出了辛勤劳动，在此表示衷心感谢。

由于网络信息过滤是一个新的研究课题，许多问题正在不断发展和变化之中，加上编著者的水平有限，本书难免会有错漏之处，敬请广大读者批评指正。

编著者

2005年2月28日

目　录

前言 …… (1)

第 1 部分　原理篇

1　网络信息过滤概述 …… (3)
1.1　网络信息过滤的基本概念 …… (3)
1.2　网络信息过滤的原理 …… (4)
1.3　网络信息过滤的种类 …… (5)
1.4　网络信息过滤与网络信息检索 …… (7)
1.5　网络信息过滤的意义与局限性 …… (9)
2　网络信息过滤的主要方法 …… (12)
2.1　分级法 …… (12)
2.2　URL 地址列表法 …… (28)
2.3　动态文本分析法 …… (30)
2.4　网络信息过滤的其他方法 …… (36)
2.5　网络信息过滤方法存在的主要问题及其发展方向 … (38)
3　网络信息过滤的主要技术 …… (41)
3.1　网络文本信息过滤技术 …… (41)
3.2　网络多媒体信息的过滤技术 …… (60)
3.3　防火墙过滤技术 …… (78)
3.4　基于 XML 的网络信息过滤技术 …… (86)
3.5　基于智能代理的网络信息过滤技术 …… (92)
3.6　语义网与网络信息过滤 …… (104)

第 2 部分　工具篇

4　网络信息过滤系统 …… (115)
4.1　网络信息过滤系统研究综述 …… (116)
4.2　网络信息过滤系统的结构和功能 …… (121)
4.3　网络信息过滤系统的特点 …… (125)
4.4　网络信息过滤软件的评价与选择 …… (126)
5　常见的网络信息过滤软件述评 …… (131)
5.1　客户端浏览器的过滤装置 …… (131)
5.2　儿童搜索引擎 …… (134)
5.3　国外常见的过滤软件 …… (135)
5.4　常见中文网络过滤软件 …… (140)
5.5　电子邮件过滤软件 …… (159)

第 3 部分　应用篇

6　网络信息过滤在反垃圾电子邮件中的应用 …… (177)
6.1　垃圾电子邮件的类型 …… (177)
6.2　垃圾电子邮件的危害 …… (178)
6.3　垃圾电子邮件产生的原因 …… (181)
6.4　电子邮件信息的过滤原理 …… (182)
6.5　电子邮件信息过滤的主要方式 …… (188)
6.6　电子邮件信息过滤的主要方法 …… (194)
6.7　电子邮件过滤语言 …… (196)
6.8　反垃圾邮件过滤技术的发展趋势 …… (198)
7　网络信息过滤在虚拟社区管理中的应用 …… (201)
7.1　虚拟社区的概念 …… (201)
7.2　虚拟社区的主要特征 …… (201)

7.3　虚拟社区的主要功能 …… (206)
7.4　虚拟社区的影响 …… (209)
7.5　BBS 的信息过滤 …… (211)
7.6　博客的信息过滤 …… (219)
8　网络信息过滤在电子商务推荐系统的应用 …… (231)
8.1　电子商务推荐系统的概念 …… (231)
8.2　电子商务推荐系统的作用 …… (232)
8.3　电子商务推荐技术的种类 …… (233)
8.4　电子商务推荐技术的比较 …… (237)
8.5　网络信息过滤在电子商务推荐系统中的应用 …… (241)
8.6　推荐系统的实例 …… (243)
8.7　电子商务推荐系统的发展方向 …… (247)
9　网络信息过滤在图书馆的应用 …… (252)
9.1　网络不良信息的影响 …… (252)
9.2　国外图书馆过滤网络不良信息的情况 …… (259)
9.3　我国图书馆处理网络不良信息问题的现状 …… (276)
9.4　我国图书馆过滤网络不良信息的对策 …… (307)
10　网络信息过滤在网吧管理中的应用 …… (327)
10.1　网吧管理的政策法规 …… (327)
10.2　网吧信息的监管软件 …… (330)
10.3　网吧信息过滤应处理好的问题 …… (333)
参考文献 …… (336)
附录 1　RSACi 分级体系 …… (342)
附录 2　ICRA 分级体系 …… (343)
附录 3　N2H2 MS Proxy Server 2.0 分级体系 …… (346)
附录 4　图书馆对网络信息过滤调查问卷
（对象:图书馆工作人员） …… (348)
附录 5　图书馆对网络信息过滤调查问卷
（对象:图书馆网络用户） …… (350)

第 1 部分　原理篇

1 网络信息过滤概述

1.1 网络信息过滤的基本概念

目前有关信息过滤的定义比较多，概括起来，比较有代表性的主要有如下几种：

(1)信息过滤指的是把信息传递给需求者的一系列过程的总称[①]。

(2)信息过滤主要是运用一定的工具和根据一定的标准设置过滤条件，在网络运行过程中一旦触发条件则将不需要的信息拒之门外，而其他信息仍可以通过网络传播进来[②]。

(3)信息过滤是一种系统化的方法，用来从大量的信息流中抽取某个人希望得到的信息[③]。

(4) 信息过滤是从网络输入的数据流中选取相关的信息或剔除不相关的信息过程[④]。

(5)信息过滤就是信息的选择性传播[⑤]。

从上述不同角度的定义，可以看出信息过滤的内涵主要包括如

---

① Belkin, Nicholas J., et al. Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM, 1992,12

② 黄晓斌. 因特网络信息过滤研究. 图书情报工作,2000(7):42~45

③ 何军,周明天. 信息网络中的信息过滤技术. 系统工程与电子技术,2001(11):76~79

④ Donna Harman, et al. Information Filtering and Retrieval: Overview, Issues and Directions. http://citeseer.nj.nec.com/38885.html

⑤ 梅海燕. 信息过滤问题的研究. 现代图书情报技术,2002(2):44~47

下几个方面：

信息过滤的目的是要满足特定用户的信息需求；

信息过滤的依据是信息与特定用户信息需求的相关性；

信息过滤的对象是动态的网络信息流；

信息过滤不仅从信息流中选取需要的信息，也从中剔除不需要的信息。

综上所述，我们认为：网络信息过滤就是根据一定的标准和运用一定的工具从动态的网络信息流中选取用户需要的信息或剔除用户不需要的信息的方法或过程。

1.2 网络信息过滤的原理

图 1－1 是信息过滤系统的一般模型。图中，一个或一组用户由于工作、学习、生活的需要产生了信息需求，这种需求在较长的一段时间里保持相对的稳定。用户对某种信息的选择需求是信息过滤的前提。用户的信息需求以计算机能够识别的形式揭示出来，就是用户需求模板（profile，也叫过滤模板），它是进行信息过滤的预定条件。对于用户需求模板，可以是正向的，也可以是反向的，也就是说既可以揭示用户需要的感兴趣的信息，也可以描述用户不需要的不感兴趣的信息。

在系统中，对动态的网络信息集不作预处理，只是当信息流经过系统时才运用一定的算法把信息揭示出来。匹配算法与用户需求模板的描述方法、信息的揭示方法是相互联系的，常用的匹配模型有布尔模型、向量空间模型、概率模型、基于知识的表示模型以及混合模型等，主要任务是剔除不需要的信息、选取需要的信息并按相关性的大小程度提供给用户。

为了提高信息过滤的效率，系统还根据用户对过滤结果的评价，通过反馈机制作用于用户和用户需求模板，使用户逐渐清晰表示自己的信息需求，使用户需求模板的描述变得越来越具体和明确。

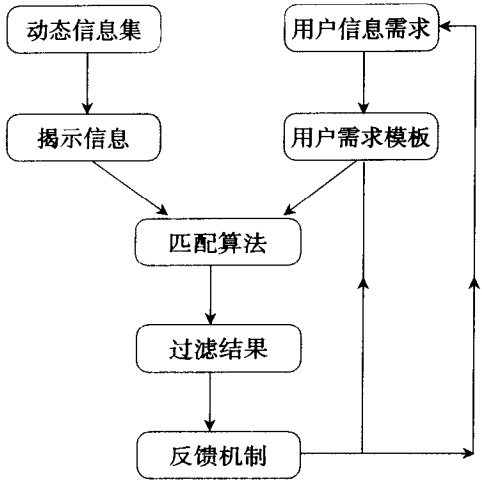

图 1-1 信息过滤系统一般模型

在整个信息过滤系统中,用户需求模板的构建、信息的揭示、匹配算法和反馈机制是最为关键的部分。在现有技术条件下,全自动的信息过滤系统还处于试验阶段,为了提高实用性,往往会在这些关键部分进行必要的人工干预,如对动态的信息集先作预处理、人工修改用户需求模板等。

1.3 网络信息过滤的种类

(1)根据信息过滤系统的结构,可以分为基于内容的过滤和协作过滤两种。

基于内容的过滤(content-based filtering)又叫认知过滤,是利用用户需求模板与信息的相似程度进行的过滤,能够为用户提供已感兴趣的相似的信息,但不能为用户发现新的感兴趣的信息,在反馈机制的作用下,用户的信息需求处于循序渐进的变化过程中。

协作过滤(collaborative filtering)又叫社会过滤,是利用用户信息需求之间的相似性或用户对信息的评价进行的过滤。价值观念、思想观点、知识水平或需求偏好相同或相似的用户,他们的信息需求往往也具有相似性。基于这一思路,通过比较用户需求模板的相似程度或者根据用户对信息的评价而进行的过滤,既可以为用户提供正感兴趣的信息,又可以提供新的感兴趣的信息。在这种系统中,用户的信息需求有可能呈现跃进式的变化。

(2)根据过滤操作的不同位置,又可以分为服务器过滤和客户端过滤。或依据用户需求模板所在的位置可以分为上游过滤和下游过滤两种。

上游过滤(upstream filtering),是将用户需求模板存放在服务器端或代理端上。服务器过滤可以在将信息下载到客户端之前进行过滤,外来的信息要经过它才能进入本地局域网,而内部信息也要经过它才能传送出去。因而可以设置一些相应的限制,对一些网址和信息进行控制。这种方式的优点是被过滤掉的信息决不会被下载到客户端。优点是支持基于内容的过滤,也支持协作过滤,缺点是模板不能用于不同的网络应用中,用户和过滤系统之间的交流变得越来越复杂,适用于大型单位。

下游过滤(downstream filtering),其过滤软件也可以是客户端的一部分,用户需求模板存放在客户端上。优点是模板可用于不同的网络应用,缺点是只能实现基于内容的过滤。在软件下载到客户端再进行过滤,需要大量时间进行维护,适用于个人或小的单位。

(3)根据信息过滤的作用,可以分为推荐系统、阻挡系统和一般过滤系统3种。

推荐系统(recommended system):根据用户对信息的需求特点把信息推荐给合适的接收者,属于协作过滤系统的一部分。

阻挡系统(blocking system):通过设置一定的条件,剔除用户不需要的信息,而对其他信息用户仍然可以获取,它主要用于过滤网络不良信息或者是用户不需要的信息。我们主要研究这种系统。其

中,在下面论述的过滤方法中,分级法、URL 地址列表法主要用于协作过滤系统,而自动文本分析法既可用于基于内容的过滤系统,也可用于协作过滤系统。

一般过滤系统是除推荐系统、阻挡系统之外的其他过滤系统或兼有多种功能的系统。

(4)根据网络信息的应用工具,可以分 Web 的信息过滤、email 的信息过滤、FTP 的信息过滤、BBS 的信息过滤、Blog 的信息过滤等。

(5)根据网络信息过滤的实现方式,可以分基于专门过滤软件的信息过滤、基于浏览器的分级审查系统和专门过滤引擎的信息过滤、基于防火墙的信息过滤和基于硬件系统的信息过滤等。

(6)根据网络信息过滤的目的,可以分为两个方面:一是过滤不良信息,主要目的在于维护网络信息的健康,净化网络环境,可称之为不良信息过滤;二是过滤掉无用、不相关的信息,主要目的在于获取与用户需求密切相关的信息,在此可称之为获取相关信息过滤。

1.4 网络信息过滤与网络信息检索

网络信息检索的目的是使用户从经过专门的组织整理的网络信息集合中获取特定需要的信息,它包括了通过对信息的揭示、存储、查找和传递以满足特定用户信息需求的方法和过程。相对来说,信息检索比较重视静态信息的揭示和用户需求的匹配问题,对于动态信息的处理和用户需求的变化则重视不够。随着网络的日益普及和网络信息总量的激增,信息过载、信息污染的问题正严重困扰着用户,单靠信息检索的方法,要从动态的庞大信息流中获取相关的信息日益困难,信息过滤作为一种能选取用户需要的信息、剔除用户不需要的信息的有效方法应运而生。比较上面所述的信息过滤的概念和信息过滤系统的基本原理,从中可以看出信息过滤与信息检索有许多共同之处。

(1)目的相同。信息过滤与信息检索都是根据用户的信息需求,从

信息集中获取信息，目的是要满足用户的信息需求。在信息检索中，描述用户信息需求的是检索式，而在信息过滤中则是用户需求模板。

（2）信息过滤系统和信息检索系统的原理基本相同，只是在信息检索系统中必须预先对信息集进行收集、揭示，揭示的结果保存在数据库中，而信息过滤系统不对信息集进行预处理。信息检索的技术方法广泛应用于信息过滤系统中。信息检索长期发展积累起来的技术方法，如用户需求的描述方法、信息的描述方法和匹配算法等都在信息过滤系统中得到了应用，但信息过滤更重视反馈机制和用户需求模板的构建和维护。

（3）效率的评价基本相同。对于信息过滤的效率评价，不同领域有不同的评价指标。过滤软件开发商往往用过滤正确率（应该过滤的信息被正确过滤的比例）、过滤错误率（不应该过滤的信息被错误过滤的比例）来评价过滤软件的效率，计算公式为：

过滤正确率 = 被正确过滤的信息/应该过滤的信息 ×100%

过滤错误率 = 被错误过滤的信息/不应该过滤的信息 ×100%

而计算机信息管理学者 Donna Harman 则认为应该把过滤效率和检索效率的评价指标统一起来，像信息检索分为检准率和检全率一样，信息过滤也分为滤准率和滤全率[①]。计算公式如下：

滤准率 = 过滤得到的用户需要的信息/过滤得到的信息 ×100%

滤全率 = 过滤得到的用户需要的信息/用户需要的所有信息 ×100%

这样，信息过滤和信息检索的效率就有了可比性，而且像检准率和检全率一样，滤准率和滤全率在一定情况下也存在互逆的关系。

信息过滤和信息检索也有许多不同之处（见表 1 - 1）。信息过滤更关注一个或一组用户的长期的信息需求，同时跟踪用户需求的变化；过滤的是动态的半结构化或非结构化的数据，而且不对信息集进行预处理；过滤的结果着重于剔除不需要的信息，带有即时性。

① Donna Harman, et al. Information Filtering and Retrieval: Overview, Issues and Directions. http://citeseer.nj.nec.com/38885.html