基于群体智能优化算法的文本过滤关键技术研究

朱振方　刘培玉　尉永清 ⊙ 著

中国水利水电出版社
www.waterpub.com.cn

基于群体智能优化算法的文本过滤关键技术研究

朱振方　刘培玉　尉永清　著

中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

计算机技术和互联网技术的迅速发展，使得网络上的网站、网页等各种信息以爆炸性的趋势增长，随之而来的还有大量的冗余信息和垃圾信息，并由此带来了信息泛滥、信息迷航以及信息疾病等一系列问题。这些冗余信息、垃圾信息不但影响着用户对Internet 的使用效率和质量，同样影响着网络的健康发展。因此，基于此而产生的网络信息过滤技术相关研究具有巨大的社会效益和经济效益。

网络信息过滤，就是根据用户的信息需求，利用一定的工具从大规模的动态信息流中自动筛选出满足用户需求的信息，同时屏蔽掉无用的信息的过程。广义的信息过滤包括对文本、音频、图像、视频等多种信息存在形式的过滤处理，狭义的信息过滤是特指对文本信息的过滤处理。本书相关研究就是针对文本信息过滤特别是中文文本信息过滤中存在的问题而提出的。

本书面向从事自然处理、网络信息、网络舆情分析等领域研究的高年级本科生、研究生和研究人员。

图书在版编目（CIP）数据

基于群体智能优化算法的文本过滤关键技术研究 / 朱振方，刘培玉，尉永清著. -- 北京 : 中国水利水电出版社，2019.11（2020.1 重印）
ISBN 978-7-5170-8228-6

Ⅰ. ①基… Ⅱ. ①朱… ②刘… ③尉… Ⅲ. ①计算机算法－最优化算法－研究 Ⅳ. ①TP301.6

中国版本图书馆CIP数据核字(2019)第254494号

策划编辑：石永峰　责任编辑：张玉玲　加工编辑：武兴华　封面设计：李　佳

书　名	基于群体智能优化算法的文本过滤关键技术研究 JIYU QUNTI ZHINENG YOUHUA SUANFA DE WENBEN GUOLÜ GUANJIAN JISHU YANJIU
作　者	朱振方　刘培玉　尉永清　著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路 1 号 D 座　100038） 网址：www.waterpub.com.cn E-mail：mchannel@263.net（万水） sales@waterpub.com.cn 电话：（010）68367658（营销中心）、82562819（万水）
经　售	全国各地新华书店和相关出版物销售网点
排　版	北京万水电子信息有限公司
印　刷	三河市元兴印务有限公司
规　格	170mm×240mm　16 开本　12.5 印张　203 千字
版　次	2019 年 11 月第 1 版　2020 年 1 月第 2 次印刷
定　价	58.00 元

前　言

计算机技术和互联网技术的迅速发展，使得网络上的网站、网页等各种信息以爆炸性的趋势增长，随之而来的还有大量的冗余信息和垃圾信息，并由此带来了信息泛滥、信息迷航以及信息疾病等一系列问题。这些冗余信息、垃圾信息不但影响着用户对 Internet 的使用效率和质量，同样影响网络的健康发展。因此，基于此而产生的网络信息过滤技术相关研究具有巨大的社会效益和经济效益。网络信息过滤，就是根据用户的信息需求，利用一定的工具从大规模的动态信息流中自动筛选出满足用户需求的信息，同时屏蔽掉无用的信息的过程。广义的信息过滤包括对文本、音频、图像、视频等多种信息存在形式的过滤处理，狭义的信息过滤是特指对文本信息的过滤处理。本书相关研究就是针对文本信息过滤特别是中文文本信息过滤中存在的问题而提出的。

本书在介绍文本信息过滤涉及的关键技术的基础上，通过提出基于统计与规则的特征项联合权重文本权重计算方法、融合段落特性的文档权重计算方法、基于自适应惯性权重混沌粒子群的特征子集优化方法，优化用于过滤的特征集合；通过提出基于模糊遗传算法的文本信息过滤模板生成算法，生成文本信息过滤模板；通过一种基于概念的逻辑段落匹配方法，解决使用传统自然段落进行匹配造成的匹配率较低的问题；通过构建一种基于微粒群的协作过滤模板动态调整和基于反馈增量学习的过滤模板更新机制，提高模板的准确性；最后设计实现了一个文本信息过滤原型系统。

感谢课题组历届毕业生为本书撰写做出的贡献，2008 级硕士研究生杨玉珍为本书第三章撰写做了大量工作，2010 级硕士研究生周燕为本书第五章撰写做了大量工作，2007 级硕士研究生张立伟和 2009 级硕士研究生许明英为本书第九章撰写做了大量工作，山东管理学院王培培副教授为本书做了大量的校正工作，山东交通学院信息科学与电气工程学院卢强、国强强、武文擎、张殿元等硕士研究生为本书校正和文字修改做了大量工作，感谢课题组 2004 级到 2018 级历届硕士生为本书撰写做的大量基础性研究工作。

本书出版得到了国家社科基金年度项目（19BYY076）、教育部人文社会科学研究一般项目（14YJC860042）和山东省社会科学规划研究项目（19BJCJ51，18CXWJ01，18BJYJ04）的资助。

作者

2019 年 7 月

目　录

第一章　绪论

随着计算机技术和网络技术的迅速发展，计算机和网络走进千家万户，成为人们生产和生活中不可或缺的组成部分，在人们享受计算机和网络技术带给大家巨大便捷的同时，也给人们带来了很多负面影响。

第一节　研究背景及意义

一、中国互联网迅速发展

2018 年 7 月 12 日，2018（第十七届）中国互联网大会在北京国家会议中心落下帷幕。同时，中国互联网协会发布《中国互联网发展报告 2018》。数据显示[1]，2017 年网民数量接近 7.72 亿，第三方互联网支付达到 143 万亿元。电子商务和网络零售以及网络购物分别达到了 29.16 亿元、7.18 亿元和 5.33 亿元。网络游戏与网络广告达到了 2354.9 亿元和 3828.7 亿元。

截至 2017 年底，中国网页数目达到了 2604 亿个，年增长率 10.3%，其中静态网页数量为 1969 亿个，占网页总数的 75.6%；动态网页数量达 635 亿个，占网页总数量的 24.4%。中国域名总数达到了 3848 万个，同比减少 9%，但.CN 域名增长 1.2%，达到 2085 万个。与此同时，2017 年手机已成为最主要的移动上网设备，从 2016 年的 95.1%提升至 97.5%；人均上网时长达到了 27 个小时；家庭成为主要的上网场所；网民主要以 10～39 岁群体为主，男女网民比例为 52.6:47.4。图 1-1 为连续五年（2013.6—2018.6）中国网民规模和互联网普及率发展状况图。

二、互联网迅速发展带来的负面影响

伴随着互联网业务的不断扩大和普及，在互联网上流传的网络信息也日益庞杂，并由此带来了信息泛滥、信息超载、信息浪费和信息疾病等一系列问题。

图 1-1 中国网民规模和互联网普及率

1. 信息泛滥

据预测，到 2020 年，全球数字信息总量将达到 35ZB。为形象地表达当前数据量之大，如果用 DVD 记录，一张张叠加起来的长度可以往返地球与月球之间。当前，全球数据存储量每年以 60%的速度递增。

网络信息的不断膨胀，导致信息泛滥和信息洪水现象不断涌现，威胁着信息安全，影响着人类应用计算机解决问题。据日本《信息流通调查报告》估计，人类标准供给信息量每 10 年约增加 4 倍，而个人消费量几乎没有大的变化。如此日积月累，过剩的信息必然堆积如山，最终会造成信息“雪崩”、信息洪水，危害社会和人类自身。

2. 垃圾信息

而与此同时，大量无用信息、虚假信息和违法信息，充斥了网民的眼球，污染了网络环境。中国互联网违法和不良信息举报中心数据显示[2]，2018 年 12 月，全国各级网络举报部门受理有效举报 860.4 万件，环比和同比分别增长 2.6%和 95.9%。其中，中央网信办（国家互联网信息办公室）违法和不良信息举报中心受理 11.1 万件，环比和同比分别增长 78.6%和 10.6%；各地网信办举报部门受理 164.1 万件，环比增长 16.1%，同比下降 10.6%；全国主要网站受理 685.2 万件，环比下降 0.9%，同比增长 1.8 倍。

垃圾信息不但影响用户对 Internet 的使用效率和质量，而且影响网络的健康发展。特别是，中国有数以亿计的大、中、小学生，他们在通过计算机网络获取

信息、了解外面世界的同时却遭受着不良信息的侵蚀，这也是教育主管部门、学校和家长共同面临的问题。网络信息过滤的研究，正是基于以上问题的解决提出来的，具有巨大的社会效益和经济效益。

三、信息过滤研究的意义

面对庞大的信息源以及芜杂的网络信息，如何有效地对这些信息进行处理，实现对有用/有益信息的获取，并且自动屏蔽无用/有害信息至关重要，势必需要一款高效的过滤工具，给网络一方净土。

网络信息过滤[3,4]，就是根据用户的信息需求，利用一定的工具从大规模的动态信息流中自动筛选出满足用户需求的信息，同时屏蔽掉无用的信息的过程。广义的信息过滤包括对文本、音频、图像、视频等多种信息存在形式的过滤处理，狭义的信息过滤特指对文本信息的过滤处理。本书正是基于解决文本信息过滤中存在的问题而展开的。

英文信息过滤的研究开展较早，人们在用户模板、信息的比较和选择、自适应学习、共享评注和文档的可视化等方面都进行了一定的研究[5,6]，但仍有较大的提升空间。中文信息过滤的研究起步较晚，目前中文信息过滤和推送系统主要还是基于关键词规则的过滤，真正的文本过滤特别是自适应的过滤的研究很少。这一方面是限于中文文本的表示和处理的难度，另一方面也是因为缺少适当的有说服力的评测集和评测标准。

中文语言的特殊性和其特有的复杂性、灵活性，给中文信息过滤技术的研究工作带来了较大的困难。在借鉴国外信息过滤技术成果的基础上，对中文信息过滤技术进行深入的研究并开发出适合我国国情的中文信息过滤系统，成为我国信息化进程的一种迫切需要。

1. 理论意义

在理论意义方面，信息过滤技术是著名的国际文本检索会议 TREC 以及主题检测和跟踪会议的主要研究内容之一，对信息过滤技术的研究具有较高的学术价值。中文是我国信息的主要载体，面向中文信息的过滤技术的研究是中文信息处理的一个重要研究方向，对中文信息过滤技术的研究也对中文信息处理的研究有较大的促进作用。

2. 改善 Internet 信息查询技术的需要

随着用户对信息利用效率要求的提高，以搜索引擎为主的现有网络查询技术受到了挑战，网络用户的信息需求与现有的信息查询技术之间的矛盾日益尖锐，其矛盾主要有以下方面：

（1）在使用搜索引擎时，只要使用的关键词相同，所得到的结果就相同，它并不考虑用户的信息偏好和用户的不同，对专家和初学者一视同仁，同时返回的结果成千上万、良莠不齐，使得用户在寻找自己喜欢的信息时如大海捞针。

（2）网络信息是动态变化的，用户时常关心这种变化。而在搜索引擎中，用户只能不断地在网络上查询同样的内容，以获得变化的信息，这花费了用户大量的时间。

因此，在现有情况下，传统的信息查询技术已经难以满足用户的信息需求，对信息过滤技术的研究日益受到重视，把信息过滤技术用于 Internet 信息查询已成为一个重要的研究方向。

3. 个性化服务的基础

个性化的实质是针对性，即对不同的用户采取不同的服务策略，提供不同的服务内容。个性化服务将使用户以最小的代价获得最好的服务。在信息服务领域，就是实现“信息找人，按需要服务”的目标。既然是“信息找人”，那什么信息找什么人就是关键。每个用户都有自己特定的、长期起作用的信息需求。用这些信息需求组成过滤条件，对资源流进行过滤，就可以把资源流中符合需求的内容提取出来进行服务，这种做法就叫作“信息过滤”。信息过滤是个性化主动服务的基础。

4. 维护我国信息安全的迫切需要

网络为信息的传递带来了极大的方便，也为机密信息的流出和对我国政治、经济、文化等有害信息的流入带来了便利。发达国家通过网络进行政治渗透和价值观、生活方式的推介，一些不法分子利用计算机网络复制、传播和查阅一些色情的、种族主义的、暴力的、封建迷信或有明显意识形态倾向的信息。我国 80% 的网民在 35 岁以下，80%的网民具有大专以上文化学历，而这两个 80%正是我们国家建设发展的主力军。因此，我国的信息安全问题的处理已迫在眉睫，必须引起我们的高度警惕和重视，而信息过滤是行之有效的防范手段。

5. 信息中介（信息服务供应商）开展网络增值服务的手段

信息中介行业的发展要经过建立最初的客户资料库、建立标准丰富档案内容和利用客户档案获取价值三个阶段。其中第一阶段和第三阶段的主要服务重点都

涉及信息过滤服务。过滤服务过滤掉客户不想要的信息，信息中介将建立一个过滤器以检查流入的带有商业性的电子邮件，然后自动剔除与客户的需要和偏好不相符的不受欢迎的信息。客户可提前指定他们想经过过滤服务得到的信息或经过过滤服务排除出去的任何种类的经销商或产品。对于不受欢迎的垃圾信息，信息中介将会在客户得到之前把它们过滤掉。在网络环境下，尽量减少无效数据的传输对于节省网络资源、提高网络传输效率具有十分重要的意义。通过信息过滤，可减少不必要的信息传输，节省费用，提高经济效益。

综上所述，对中文信息过滤技术的研究无论是在学术理论上还是在具体应用方面都具有较高的价值。

第二节　文本信息过滤面临的问题

自从 1982 年 Denning 首次提出信息过滤（Information Filtering，IF）的概念[7,8]以来，信息过滤相关技术和产品从设想成为现实，并且不断地进步和完善。在此期间，国内外众多的研究机构和个人做了大量工作。

一、国外相关研究

1982 年，Denning 提出信息过滤的概念。在此后的 10 年间，关于信息过滤的应用研究逐渐开展起来，研究领域也从最初的电子邮件延伸到其他相关领域，出现了许多研究成果。1989 年，美国国防高级研究项目署（Defense Advanced Research Project Agency，DARPA）资助了第一届 Message Understanding Conference，极大地推动了信息过滤的发展。1992 年，NIST（美国国家标准和技术研究所）与 DARPA 联合赞助了每年 1 次的文本检索会议（Text Retrieval，TREC），对文本检索和文本过滤倾注了极大的热忱。

随着互联网的迅速发展、需求的不断增加，文本过滤以及相关技术方面取得了长足的进展，成为信息产业新的增长点，取得了很多研究成果。Nanas 等人通过类似于生物免疫系统的机能，构造具有动态性和自适应性的信息防御体系[9]；Yokoi 等人使用奇异值分解移除文档噪声，提高主题抽取的准确率，并利用独立成分分析方法分析文档的潜在语义来描述用户需求，提高了过滤准确率[10]。Nanas 等人使用滑动窗口捕获网络中的项之间的依赖性，在对文档进行评价时，使用激

活扩散方法将项依赖性考虑在内，提高了过滤性能[9]；Zhou 等人提出了一种使用模式分类挖掘技术的新的信息过滤模型[11]。Acilar 等人提出了一种基于人工免疫网络的协作过滤算法，利用人工免疫网络降低数据稀疏性，并通过描述数据结构提供数据集的可扩展性[12]；Chen 等人提出了一种应用正交非负矩阵分解的协作过滤框架，通过矩阵分解减轻稀疏性问题，通过同时聚类用户评分矩阵的行和列来解决可扩展问题[13]。Damankesh 等人使用人类合情推理理论构建多语种过滤框架[14]；Liu 等人利用社交网络信息来加强推荐效果从而提高协作过滤的性能[15]。

二、国内研究进展

国内对于文本信息过滤特别是中文文本信息过滤的研究相对较晚，但是发展很快，特别是 1996 年以来，国内很多机构对信息过滤进行了大量研究。

国内的微软亚洲研究院、清华大学、复旦大学、中科院软件所、哈尔滨工业大学以及东北大学等机构相继开展了信息过滤技术，特别是面向中文的信息过滤技术的研究，其间积累了很多宝贵的经验，也取得了一些不错的成绩，这也为本书的研究提供了大量有益的借鉴。

近年来，仍有一些研究机构和个人为此做了大量工作。例如，曾春等人提出利用领域分类模型上的概率分布表达用户的兴趣模型，给出相似性计算和用户兴趣模型更新方法[16]；而洪宇等[17]提出了一种建立信息流二元近似关系模型，辅助信息过滤系统识别和屏蔽反馈中的噪声，在众多基于语义技术的信息过滤研究中，文献[18]提出了一种基于本体的信息检索技术，利用本体概念的语义描述能力实现信息准确检索；文献[19]则提出利用 OWL 描述信息语义，进而在语义网环境中实现信息过滤；文献[20]则给出了一种通过奇异值分解以及独立分量分析获取的潜在语义描述方法实现信息过滤，文献[21-23]则着重研究了协同过滤算法及其在推荐系统中的应用。

近年来，还有一些值得借鉴的研究成果。文献[24]针对社交网络 Facebook 上出现的不需要的消息以及令人反感的文字，提出了一个自动化的框架对网页上没意义的内容进行过滤。文献[25]则是对社交网络 Facebook 上出现的政治性的词汇进行分类识别。文献[26]针对非典型文本进行分类，改进后使得分类结果更加准确，在安全性和动态适应性方面也有着较好的表现。文献[27]针对垃圾邮件过滤采用了贝叶斯这类决策方法，在对文本特征选择时，对特征词的类别条件熵计算做出了改进。文献[28]提出了一种基于贝叶斯过滤和近似字符串匹配技术相混合