情报检索语言与智能信息处理丛书

丛书主编／侯汉清

文本自动标引与自动分类研究

章成志
白振田／著

東南大學出版社
SOUTHEAST UNIVERSITY PRESS

情报检索语言与智能信息处理丛书(丛书主编　侯汉清)

文本自动标引与自动分类研究

章成志　白振田　著

东南大学出版社
·南京·

图书在版编目(CIP)数据

文本自动标引与自动分类研究 / 章成志,白振田著.
南京:东南大学出版社,2009.12
(情报检索语言与智能信息处理丛书/侯汉清主编)
ISBN 978-7-5641-1913-3

Ⅰ.文… Ⅱ.①章…②白… Ⅲ.计算机应用—分类标引—研究 Ⅳ.G254-39

中国版本图书馆 CIP 数据核字(2009)第 200911 号

情报检索语言与智能信息处理丛书(侯汉清主编)
文本自动标引与自动分类研究

出版发行	东南大学出版社
出 版 人	江 汉
社 址	南京市四牌楼 2 号(邮编:210096)
印 刷	南京玉河印刷厂
责任编辑	李 正
	(电话:025-83790887;E-mail:leezheng1978@sina.com)
经 销	新华书店
开 本	880 mm×1 230 mm 1/32
总 印 张	50.625(本册 6.625 印张)
总 字 数	1 310 千字(本册 173 千字)
版 次	2009 年 12 月第 1 版 2009 年 12 月第 1 次印刷
总 定 价	200.00 元(共 8 本)

* 东大版图书若有印装质量问题,请与读者服务部联系,电话:025-83792328

丛书总序

这部丛书包括下列八本专著：

(1) 薛春香著《网络环境中知识组织系统构建与应用研究》；

(2) 陆勇著《面向信息检索的汉语同义词自动识别》；

(3) 杜慧平、仲云云著《自然语言叙词表自动构建研究》；

(4) 章成志、白振田著《文本自动标引与自动分类研究》；

(5) 张雪英著《情报检索语言的兼容转换》；

(6) 刘华梅、戴剑波著《受控词表的互操作研究》；

(7) 何琳著《领域本体的半自动构建及检索研究》；

(8) 李运景著《基于引文分析可视化的知识图谱构建研究》。

这八本专著是侯汉清教授多年来指导博士生、硕士生们进行科学研究(有些是同他们合作研究)的具体成果的一部分。这些著作的主题内容,可以归结为“情报检索语言的自动化”和“自然语言检索”两个相关的问题,或者更概括地说,就是“信息检索自动化的升级问题”,属于当前信息检索学术研究的前沿课题。

这些专著,如果将其分散来看,或许不觉得分量之重;但如果把八本专著放到一起,就可以看出其成果之丰硕。侯汉清教授在带研究生中看准一个方向不断开拓、持之以恒的精神,可以出大成果,值得我们效法。南京农业大学在侯汉清教授领导下进行的有益的研究工作,我想一定会成为我国信息检索自动化发展史册之中浓浓的一笔。

这一类项目，本质上都是情报语言学的研究课题。所以，在研究中必须遵循情报语言学的理论，吸取情报语言学的已有成果，其结论应切合情报语言学的要求。它们只是利用计算机技术作为方法手段来达到研究目的而已，不能过分强调网络环境的特殊性而置情报语言学关于检索效率的基本要求于不顾。计算机技术应当与情报语言学密切结合。侯汉清教授和他的弟子们同时具备这两方面的知识，是顺利地较好地完成这些研究项目的关键。

这八个研究项目，大多采取实验研究法，故其成果具有较大的可信度和易理解性。其中有些项目，难度较大，甚至极难，专著只是作了认真、有益的探索；有些项目，虽然尚有一些不足，但作为中间成果，可在当前信息检索工作中推广应用，在应用中进一步完善。

信息检索自动化的初级阶段已在我国普遍实现。但要晋升一级，扩大自动化过程的范围和提高自动化的水平，当前的研究还属起步，发表的科研成果尚少见，学术研究有待扩大和深入。这部丛书起了很好的开拓作用，为继续研究打下了基础，是研究者很好的学习和参考用书，希望对此感兴趣的读者能从中获益。

张琪玉

2009 年 7 月

序　言

随着因特网及其相关技术的不断普及深入，信息检索和文本挖掘成为人们获取知识的不可或缺的手段。但是面对异构的、动态的、海量的网络信息，如何快速找到用户感兴趣的信息并有效地加以利用是摆在我们面前亟待解决的重大问题。

近十年来信息检索技术得到了迅猛的发展，成为解决海量信息查找的有利工具。然而由于网络信息资源的更新速度和文本的非规范性，信息检索后的结果也往往数量巨大，“噪音”很多。所以需要快速有效的信息组织方式将这些信息分门别类加以分析处理，使用户能够更加有效地利用信息。文本的自动标引和自动分类是信息资源加工与组织中的关键技术。

《文本自动标引与自动分类研究》一书对文本挖掘的研究和应用给出了一个全新的视角。他们从情报检索的角度，以厚实的专业基础和丰富的领域经验，对于文本自动标引和自动分类技术进行了有益的探索和周密的实验，取得了丰硕的研究成果。

他们研究的特点是以数字图书馆、网络信息处理等为主要的应用背景，结合目前中文信息处理和文本处理的最新发展状况，在自动标引和自动分类方面有所继承，更有创新，主要研究特色在于：

一是自动标引技术的研究与应用，自动标引包括抽词标引与赋词标引两种类型。从信息组织的角度，自动标引相当于词汇级的摘要，而文本摘要则是句子级的摘要。而前者由于短小精悍、代表性强、使用效率高，成为信息检索、信息抽取、问答系统、主题识别和跟踪的有利工具。作者对自动标引中抽词词典构造、基于多特征选择及权值计算、标引源权重设置等问题进行了深入的研究，给出了高效的自动标引算法。

二是采用基于知识库的文本自动分类方法，以大规模的人工标引关键词词串与分类号对应记录为基础，生成关键词词串-分类号对应的分类知识库，并将其用于新文本的自动分类；对于文本分类难点问题——多层次分类问题，进行了深入的研究，给出了高效的层次分类算法。分类体系为《中国图书馆分类法》和行业分类表，分类体系庞大，总的分类类目数量多达到成千上万的规模。

很荣幸，能够先睹为快，拜读大作，掩卷之余，感受颇多。

目前在情报检索领域，专门讨论自动标引和自动分类的专著很少，这本书将情报检索技术和海量文本处理技术相结合，将自然语言处理和机器学习等引入情报检索领域，系统全面地介绍了文本自动标引和自动分类的研究工作，为情报检索的研究探索出有价值的研究方向，为文本挖掘的研究提供了很好的应用平台。

林鸿飞

2009 年 7 月 25 日于大连理工大学创新园大厦

目　次

第一部分

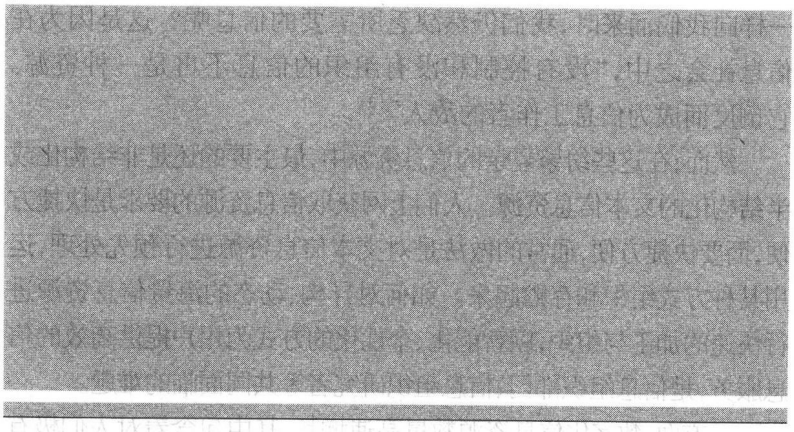
第二部分

第三部分

第四部分

图表目次

第一部分



第1章 引言

1.1 研究背景

随着计算机及网络的普及，数字化载体逐渐融入人们的生产、生活中，成为人们获取信息资源不可或缺的途径、方法和手段。根据第23次中国互联网络发展状况统计报告显示，目前我国上网网民已经突破1亿，网民用于上网的时间每周平均在14个小时以上[1]。我国网民人数的增加、上网时间的增长，从一个侧面说明，数字化载体十分具有吸引力，能够方便、快捷地为人们提供所需要的信息资源。在过去的20年中，万维网的迅速发展使其成为世界上规模最大的公共数据源。万维网数据量巨大且不断增长、数据类型丰富、信息异构、信息包含噪音等特点，使得挖掘有用的信息和知识的任务变得十分有趣，并富有挑战[2]。

我们正处于“信息爆炸”的时代，但为什么当各类信息像洪水

一样向我们涌来时，我们仍然缺乏所需要的信息呢？这是因为在信息社会之中，“没有控制和没有组织的信息不再是一种资源。它倒反而成为信息工作者的敌人”[3]。

然而，在这些纷繁复杂的信息资源中，最主要的还是非结构化或半结构化的文本信息资源。人们上网获取信息资源的要求是快捷方便，而要快捷方便，通常的做法是对文本信息资源进行预先处理，运用某种方式组织和存贮起来。如何对异构、动态的海量信息资源进行快速的加工与组织，以智能化、个性化的方式为用户提供高效的信息服务，是信息组织部门、信息组织研究者等共同面临的难题。

一方面，数字化信息资源数量高速增长，其中包含着对人们极有潜在价值的知识和信息；另一方面，人们运用网络获取信息资源的数量也在高速增长，而人们获得的有效信息资源的比例却在下降。其原因在于互联网的政策——任何人可以发布任何未经加工的信息，这些未经加工的信息难以获取效率，而经过加工的信息能够提高人们的获取效率。未经加工的信息越多，人们的信息资源检索效率就越低。

然而，要解决信息资源无限增长和检索效率低下的矛盾，究其原因是多方面的，非某一种技术所能解决，存在的困难也是多方面的。但主要的原因还是信息资源的多样化和海量化，且没有经过规范的加工处理。在这些杂乱无序的信息海洋中，用户要迅速准确地找到自己所需要的信息，困难重重。

关于信息资源的加工与组织方法比较多，其中文本的自动标引和自动分类是比较关键的技术，并且有广泛的应用。文本自动标引（本书是指狭义上的自动标引，即文本的关键词抽取或主题词获取）是对文本根据其表达的内容或主题，进行关键词或主题词自动获取的过程。分类是人们浏览和查找信息的主要手段之一。文本自动分类是根据某一特定的分类体系，将文本资源分门别类地进行自动组织的方法。

本书从文本的自动标引和自动分类角度出发，调研文本自动标引和自动分类技术的相关理论研究，结合实际应用，以中文文本