信息管理专业前沿论丛

中文新闻网页处理与舆情分析

钱爱兵 著

南京大学出版社

本课题研究受到以下项目资助：

◎江苏省教育厅高校哲学社会科学研究基金资助项目：
基于本体的高校突发事件网络舆情监控预警模式研究（项目编号：2010SJB870003）
◎江苏省社会科学基金项目立项资助项目：
网络舆情监控预警模式研究（项目编号：10TQC008）
◎江苏省教育厅“青蓝工程”资助项目。

信息管理专业前沿论丛

中文新闻网页处理与舆情分析

钱爱兵 著

南京大学出版社

图书在版编目(CIP)数据

中文新闻网页处理与舆情分析 / 钱爱兵著. -- 南京
：南京大学出版社，2012.12
ISBN 978-7-305-11046-7

Ⅰ. ①中… Ⅱ. ①钱… Ⅲ. ①互联网络—新闻—舆论
—分析 Ⅳ. ①G210.7

中国版本图书馆 CIP 数据核字(2013)第 007439 号

出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　　编 210093
网　　址 http://www.NjupCo.com
出 版 人 左　健

书　　名 中文新闻网页处理与舆情分析
著　　者 钱爱兵
责任编辑 王泰理　吴　汀　　编辑热线 025-83686531

照　　排 南京南琳图文制作有限公司
印　　刷 南京大众新科技印刷有限公司
开　　本 787×960 1/16　印张 10.25　字数 183 千
版　　次 2012 年 12 月第 1 版　　2012 年 12 月第 1 次印刷
ISBN 978-7-305-11046-7
定　　价 30.00 元

发行热线 025-83594756　83686452
电子邮箱 Press@NjupCo.com
Sales@NjupCo.com(市场部)

前　言

随着互联网络在全球范围内的飞速发展，网络媒体已被公认为是继报纸、广播、电视之后的“第四媒体”。在反映和引导社会舆论方面，网络媒体具有与传统媒体同样的功能。然而，网络媒体与传统媒体相比在传播载体和传播方式上又有着本质的不同：一方面，任何人都可以在 BBS 论坛、留言板或者自建站点上发布言论和观点，并且发布者往往不必考虑发布言论的真实性以及由此带来的社会影响，这在传统媒体领域通常是不可想象的；另一方面，网络媒体信息的正确性及传播范围均无法得到有效控制，因此，网络舆情热点、焦点层出不穷。

伴随着信息化建设的高速发展，政府决策者和相关职能部门开始对网络媒体的舆论导向提出更高的要求，如何加强网络信息的管理已成为迫切需要解决的问题。众所周知，网络媒体的传播载体是网页，对网络媒体的监督与管理实际上就是对网页信息的分析与处理，而网页的海量性、动态性和不可控性为信息处理、信息检索和信息使用带来新的挑战，也使得传统的手工方式难以胜任对网页的一系列处理工作。

本书正是以新闻网页为例，结合新闻的专有特性，面向舆情分析，从 6 个方面对中文新闻网页处理过程中涉及的关键技术问题进行深入系统的研究，即新闻网页正文抽取、新闻重复网页识别、新闻网页关键词抽取、新闻网页自动分类、新闻网页主题聚合、网络舆情分析，并给出相应的解决方法。通过对网络舆情信息进行及时、全面、准确地分析与处理，最终达到随时关注社会动态，为决策者进行正确选择与科学决策提供支持的目标。

本书关于面向舆情分析的中文新闻网页处理关键技术的研究内容概括如下：

（1）新闻网页正文自动抽取：该部分主要解决中文新闻网页中导航、广告、版权声明、相关链接等信息的过滤问题。本章针对抽取中文网页正文的传统方法的不足，提出一种基于统计的中文网页正文抽取方法。该方法首先利用 DOM 树计算文本结点的文本密度，即文本长度与 HTML 源码长度之比，再利用贝叶斯判别准则计算密度区分阈值，最后根据文本密度与密度区分阈

值的比较结果抽取正文：大于密度区分阈值的结点就判定为正文文本结点，小于或等于密度区分阈值的结点则判定为非正文文本结点，将所有判定为正文文本结点的文本连接起来即为要抽取的网页正文。通过使用中文新闻网页对该方法的有效性进行验证，结果表明：该方法是一种易于实现、抽取准确的通用性方法。

（2）新闻重复网页自动识别：该部分主要解决中文新闻重复网页自动识别的问题。本章提出一种基于后缀树的中文新闻重复网页识别算法，以后缀树作为基本数据结构，依据新闻网页的标题性和时间性，构建中文新闻重复网页识别算法。该算法以 Ukkonen 算法和 Matching Statistics 算法为基础，并对其具体实现进行优化。实验结果表明：该算法识别重复新闻网页的有效性，对计算字符串相似度也有启发意义。

（3）新闻网页关键词自动抽取：该部分主要解决中文新闻网页自动标引的问题。本章结合新闻的内容特征对中文新闻网页关键词的构成特点进行深入分析，在经典的 TF-IDF 加权公式基础上构建一个综合考虑多种影响因素的候选关键词评分加权公式。选择评分较高的词语作为候选关键词，利用词语的位置标注数据进行关键词抽取优化操作，将“切碎”的候选关键词进行组配，形成正式抽取的关键词。实验结果表明：该方法明显优于基准方法，能够抽取到令人满意的关键词。

（4）新闻网页自动分类：该部分主要解决中文新闻网页的自动分类问题。文档标题通常代表文章的中心和主旨，这一特点在新闻中体现得尤其明显。本章借鉴 TF-IDF 的思想，利用新闻标题来做中文新闻网页自动分类的依据，构建基于标题的中文新闻自动分类方法。通过设计多个实验对各种基于标题的中文新闻网页自动分类方法进行评测，结果表明：用标题来做中文新闻网页分类可以大大缩短判断处理的时间，也可以节省很多的存储空间，且准确率较高，特别是改进的类目加权法的分类效果较好。

（5）新闻网页主题自动聚合：该部分主要解决主题新闻网页自动聚合的问题。中文新闻主题网页聚合是信息处理领域内的一个新兴且有实用价值的方向。本章通过分析主题新闻网页聚合的基本问题，指出聚合技术的难点，在原有技术基础上，设计出许多独具特色的新算法，比如，将新闻的 RSS 元数据和内容结合起来判断新闻主题相关性的算法，进而提出利用 RSS 技术实现主题网页自动聚合，并给出详细的聚合系统设计方案。实验结果表明：基于 RSS 技术对中文新闻网页进行主题聚合的准确率较高，优势明显。

（6）网络舆情自动分析：该部分主要解决网络舆情自动分析问题。在网络环境下，舆情的表现形式就是网络舆情，它表达快捷、信息多元、方式互动，具备传统媒体无法比拟的优势，仅仅依靠传统的手工方法难以胜任舆情信息的采集、分析和处理工作，从而舆情信息的正确性及传播范围都无法得到有效控制，舆情形成迅速，热点、焦点层出不穷，对社会影响巨大。本章针对传统舆情分析方法的不足，提出基于主题进行网络舆情分析的思想，并构建一个基于主题的网络舆情分析模型。实验结果表明利用该模型进行网络舆情分析的有效性。

著者

2012 年 10 月

目　录

第1章 绪 论

1.1 研究背景

目前，在互联网络的各种应用中，以 Web 应用最为普及，发展最为迅速。Web 已经成为政治、经济、文化、教育等各个领域的重要组成部分，并从多个角度对当前全球化、多元化社会进行全方位反映。

Web 与传统媒体相结合形成网络媒体，并被公认为是继报纸、广播、电视之后的“第四媒体”。传统媒体作为社会舆论的工具，具有反映和引导社会舆论的功能，网络媒体在反映和引导舆论方面也具有同样的功能，然而，网络媒体与传统媒体相比在传播载体和传播方式上又有着本质的不同：一方面，任何人都可以在 BBS 论坛、留言板或者自建站点上发布言论和观点，并且发布者往往不必考虑发布言论的真实性以及由此带来的社会影响，这在传统媒体领域通常是不可想象的；另一方面，由于网络媒体信息的正确性及传播范围均无法得到有效控制，从而导致舆论热点、焦点层出不穷。

随着信息化建设的高速发展，政府决策者和相关职能部门也开始对网络媒体的舆论导向提出更高的要求，如何加强网络信息的管理已经成为迫切需要解决的问题。众所周知，网络媒体的传播载体是网页，因此，对网络媒体的监督与管理实际上就是对网页信息的分析与处理，而网页的海量性、动态性和不可控性为信息处理、信息检索和信息使用带来新的挑战，也使得传统的手工方式难以胜任对网页的一系列处理工作。

1.2 研究对象

舆情也即舆论，是指公众关于现实社会以及社会中的各种现象、问题所表达的信念、态度、意见和情绪表现的总和，具有相对的一致性、强烈程度和持续性，对社会发展及有关事态的进程产生影响，其中混杂着理智和非理智的成分[1]。以网络为平台，通过新闻、评论、博客、BBS 论坛、微博等为载体表现出

来的舆情就是网络舆情。网络舆情是舆情的一种具体表现形式，它既有舆情的共性，又有自己的特点。目前，我国对于网络舆情分析的研究还处于探索阶段，缺乏行之有效的方法，而各种中文新闻网页处理技术的出现和逐步成熟，为解决这一问题提供了良好的技术手段。

本书面向舆情分析，对中文新闻网页处理的关键技术进行研究。本书的研究对象是中文新闻网页，即来源于中文新闻网站，其主要目的在于提供最新信息的网页，其他类型的网页不在本文的研究范围之内。

1.3 研究现状

目前关于中文网页处理技术的研究已经很全面且很深入，包括主题网页采集[2~19]、网页正文抽取[20~25]、网页主题特征抽取[26~31]、网页搜索[32~33]、网页自动分类[34~35]等，但是专门针对新闻网页处理的研究[36~41]还很少，而且缺乏系统性，尤其是面向网络舆情分析[42~44]领域的新闻网页处理技术的研究更是少之又少。

本书正是以新闻网页为例，结合新闻的专有特性，面向舆情分析，对中文新闻网页处理过程中涉及的6个方面的关键技术进行深入系统地研究，即新闻网页正文抽取、新闻重复网页识别、新闻网页关键词抽取、新闻网页自动分类、新闻网页主题聚合、网络舆情分析，并给出相应的解决方法。最终实现：(1) 及时、全面、准确地分析与处理中文新闻网页信息；(2) 随时关注社会动态，对网络舆情的变化、发展趋势进行科学预测；(3) 为决策者和政府相关职能部门进行正确选择与科学决策提供支持。

1. 新闻网页正文抽取

在自然语言处理领域，利用自然语言处理技术对网页文本进行处理时，网页正文文本的准确抽取具有重要意义：(1) 与自动分词和命名实体识别系统集成，可以对网页进行自动分词和命名实体识别；(2) 与搜索引擎系统集成，可以过滤网页中导航、广告、版权声明、相关链接等内容对检索结果的干扰；(3) 与自动文摘系统集成，可以对网页进行自动摘要；(4) 与文本分类系统集成，可以对网页进行自动分类；(5) 与文本聚类系统集成，可以对网页进行自动聚类。

对网页正文进行抽取的传统方法主要是使用包装器(Wrapper)[45~46]，即从待抽取正文的网页中归纳总结出抽取规则创建包装器，然后根据包装器抽取需要的数据。一般来讲，包装器是最好、最准确的方式，但是由于网页结构

的复杂性和不规范性，一个包装器的实现只能针对一个站点，因此，在大多数条件下制作包装器是个费时费力的工作，而且如果抓取的种子站点过多，人工难以满足需求。于是自动生成包装器就成为一个折衷的替代方案[47]，但实际上为了保证正文抽取的准确性，自动生成的包装器还是需要人工确定具体的对应数据项。如果从语法分析去考虑，这个部分不参与人工，目前还没看到有成功的案例。文献[48]提出了一种专门针对新闻网页正文的抽取方法，但该方法只适合于网页正文文本存放在 Table 中的情况，对于其他情况难以处理，因而，不具有一般通用性。

综上所述，寻找到一种易于实现、抽取准确的通用型方法成为中文新闻网页处理的关键。

2. 新闻重复网页识别

在信息处理领域中，利用计算机处理文本信息时，重复及相似内容的识别是个比较重要的研究课题，它广泛应用于防抄袭识别、网络舆情分析、自动分类、搜索引擎等系统中，在基因序列分析领域也有较好的应用。

国际上对重复及相似文档识别的研究起源于大型文件系统，后来又被扩展到数字图书馆项目[49~50]。目前，代表性的方法主要有：基于聚类的方法[51~52]、排除相同 URL 的方法[53]、基于特征码的方法[54]等。聚类方法是以 6 763 个汉字作为向量的基，通过对网页中的汉字进行词频统计构建代表网页的向量，再计算向量的夹角余弦决定是否是相同网页。该方法操作简单，易于实现，但对于大规模网页，计算时间长，实时性差。排除相同 URL 的方法是根据网页的 URL 进行重复性识别，相同的 URL 认为是相同的网页。该方法操作简单，易于实现，但不能识别因转载造成的重复网页。基于特征码的方法是根据标点符号多数出现在网页正文中的特点，以句号两边各五个汉字作为特征码来唯一标识网页。该方法识别效率较高，但是特征码的精确匹配依然不能抵抗网页转载时产生的噪声，从而影响识别的准确率。

新闻网页具有转载率高、发生时间集中等特点，而中文也是一门语义纷繁复杂、语法结构灵活多变的语言。因此，利用上述通用方法识别中文新闻重复网页难以达到理想的识别效果，从而影响分析结果的准确性。

3. 新闻网页关键词抽取

关键词抽取是指如何从一篇文档(或多篇相关文档)中自动抽取出能很好地代表文档主题的若干个(一般不少于 5 个)词或短语。关键词抽取技术广泛应用于文本分类/聚类、信息过滤、信息检索、自动摘要等各种智能文本信息处

理领域，具有很好的应用价值。

目前，国内外的许多学者已经在关键词抽取领域做了大量研究工作，并且提出诸多有代表性的方法。1997 年，简立峰[55]采用 PAT 树结构，同时利用词之间的互信息来抽取中文关键词。实验结果表明：该方法抽取关键词的效果较佳，但是构建 PAT 树的时间和空间成本太大，抽取效率相对较低。2002 年，杨文峰[56]对 PAT 树的构建算法进行改进，提出利用文档中的最大重复串抽取关键词。2006 年，张阔[57]等人提出基于 SVM 模型进行关键词抽取。杨文峰、张阔等人提出的方法均是基于机器学习的方法，需要用带有关键词标注的语料文本训练抽取模型，而这正是中文关键词抽取领域所缺乏的。他们给出的解决方法是：手工标注一定数量文档的关键词，然后用标引好的文档语料训练抽取模型，再用训练好的模型抽取关键词。这就导致一个问题：由于没有统一的训练语料库，从而无法对各种抽取方法的优劣进行客观评价。此外，中文分词的质量也是影响关键词抽取效果的一个重要因素。英文文档的词之间存在天然的分隔符，即空格符，但中文文档的词之间不存在这样的分隔符，因此，在抽取关键词之前必须先使用分词器进行分词，然后才能进行关键词抽取操作，从而分词的质量将直接影响关键词抽取的结果。

综上所述，与英文关键词抽取研究相比，中文关键词抽取研究主要面临两方面的挑战：(1) 缺乏标准语料库；(2) 依赖分词。

4. 新闻网页自动分类

近年来，随着 Web 技术的飞速发展与普及，各种电子文本在数量和类别上不断累积，造成了有效管理与利用的难题，从而，电子文本分类的需求应运而生。作为电子文本信息重要组成部分的中文新闻网页也面临着同样的挑战，迫切需要通过标准化的分类加以规范，实现新闻行业之间、新闻行业和广大用户之间的新闻信息互换、存储、处理和共享。

传统意义上的文本分类均是由人工完成，即根据文本管理者的需求与期望，事先定义或选定类别，再由人工阅读文本，根据其主题大意给予适当类别标示。由此可以看出：人工分类的周期太长，成本又高且效率低下，难以适应中文新闻网页迅猛增长的实际情况，因此，实现自动分类是中文新闻网页分类工作的必由之路。众所周知，标准化是自动化的基础和前提，但长期以来，中文新闻信息没有统一的分类标准，这一瓶颈严重制约了中文新闻信息自动分类技术的研究与开发。直到 2006 年 1 月 5 日，我国第一部中文新闻信息分类国家标准——《中文新闻信息分类与代码》[58]正式颁布实施，才填补了中文新

闻信息分类法的空白，也从此打破了中文新闻信息自动分类研究的僵局。

目前，关于自动分类的研究，虽然所采用的分类算法不尽相同，包括 Naive Bayes[59~60]、SVM[61~63]、kNN[64]、Rocchio[65] 等等，但绝大部分都是依据文件内容进行分类，因此需要对整篇文章的文本作相关的预处理，包括分词、停用词过滤、关键词抽取等。如果处理的文本是网页，那么在分词之前还要进行正文抽取操作，因此，处理过程相当麻烦、耗时，计算量也十分庞大，且需要大量的存储空间，而且正文抽取、文本分词、关键词抽取的质量也将直接影响自动分类的精度。

5. 新闻网页主题聚合

我们每天都要花费相当多的时间在大量的 Web 站点上阅读新闻，了解世界正在发生的事情。不仅如此，为了及时了解最新新闻，我们还可能每隔一段时间（比如一小时或两个小时）跟踪浏览这些 Web 站点。这种跟踪浏览往往会占用我们大量的时间。当我们无暇分身的时候，有没有一种比较好的解决方案来帮助我们？是否可以创建这样一个系统，它能在新内容出现在我们感兴趣的 Web 站点时，为我们整合相关的信息内容并通知我们？随着 RSS[66] 技术的出现，这一设想变成了现实。

从本质上讲，RSS 不是内容，而是一种渠道。RSS 快速而准确地沟通内容提供商和用户之间的联系，缩短了信息延迟。但是，目前的 RSS 阅读器只是简单地解决了 RSS 订阅的问题，却未对订阅内容进行任何智能处理[67]，即仅将各个 RSS Feed 进行简单聚合，目标是尽可能多地聚合新闻页面，而较少考虑聚合页面的准确性，更没有对网络新闻进行准确而系统的主题分类。用户只是根据自己的个人喜好手动添加，结果导致大量不相关新闻的出现，分散了用户的注意力，浪费了用户的时间。因此，新闻网页按主题实现自动聚合具有重要意义。

6. 网络舆情分析

在网络环境下，舆情的表现形式就是网络舆情，它表达快捷、信息多元、方式互动，具备传统媒体无法比拟的优势，仅仅依靠传统的手工方法难以胜任舆情信息的采集、分析和处理工作，从而舆情信息的正确性及传播范围都无法得到有效控制，舆情形成迅速，热点、焦点层出不穷，对社会影响巨大。如何加强网络舆情信息的管理已成为网络舆情监督部门和决策者迫切需要解决的问题。

当前，一些学者和研究机构已经取得一定进展[68~69]，但同时也存在一些

不足：(1) 分析深度不够，仅停留在相关数据的统计层面，没有剖析数据背后的深层含义；(2) 虽然提出一些分析模式和判据[70]，但缺乏利用计算机自动化处理的可操作性，难以在实际工作中加以应用。

1.4 研究内容

1. 新闻网页正文抽取

针对中文新闻网页正文抽取问题，本章提出一种新的网页正文抽取方法——文本密度判别法。该方法基于网页中的一种普遍现象：网页正文文本结点的文本密度通常都要远远大于非正文部分。基本原理是：利用贝叶斯判别准则[71]求出文本结点的密度区分阈值，该阈值能够使得发生文本结点误判的平均损失达到最小，然后将各文本结点的文本密度与该密度区分阈值进行比较，大于密度区分阈值的结点就判定为正文文本结点，小于或等于密度区分阈值的结点则判定为非正文文本结点，将所有判定为正文文本结点的文本连接起来就是需要抽取的网页正文。实验结果表明：该方法简洁明了，抽取正文的准确率高，且易于实现。

2. 新闻重复网页识别

针对中文新闻重复网页识别问题，本章提出利用后缀树在字符串处理领域的卓越性能，依据新闻网页的标题性和时间性，构建基于后缀树的中文新闻重复网页识别算法。该算法基于 Ukkonen 算法[72]和匹配统计算法[73]，并针对中文独有的特征对其具体实现进行优化。由于 Ukkonen 算法和匹配统计算法的时间复杂度均为 $O(n)$，因此在本章设计的算法中，每次比较的时间复杂度也为 $O(n)$，计算时间少，识别效率高，从而能够解决在大规模中文新闻网页集合中识别重复网页实时性差的问题。此外，该算法在确认新闻网页标题和发布时间的基础上进行重复性识别，即使网页存在因转载而产生的噪音，也能够有效识别，大大提高识别的准确率。

3. 新闻网页关键词抽取

针对中文新闻网页自动标引问题，本章提出一种基于改进 TF-IDF 的关键词抽取方法。尽管抽取中文网页关键词同样面临缺乏标准语料库和依赖分词这两个挑战，并且朴素贝叶斯[74]、决策树[75]和支持向量机[57]等基于机器学习的方法在抽取中文关键词方面表现良好，但是在没有标准的训练语料库可供利用，而分词的准确率已达到实用化水平的条件下，有理由认为：基于改进