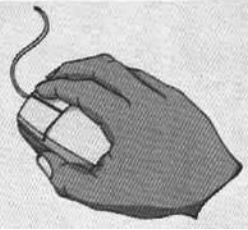
面向网络舆情分析的实用关键技术概述



■中国科学院计算技术研究所 戴媛 程学旗

【摘要】快速发展的互联网以其交流便捷和传播迅速的显著特征，成为反映社情民意最重要的窗口。越来越多的民众通过网络表达真实的想法和观点；同时，互联网涌现性特性使得大量网络舆情的爆发，某种程度上成为舆情安全的重要因素。因此，对网络舆情安全的监测与分析迫在眉睫，亟需舆情掌控的针对性关键技术。本文对网络舆情相关核心技术做了概述性的总结，并介绍了我们在实际的研发工作中遇到的问题及解决思路。

【关键词】网络舆情 信息采集 信息提取 话题发现 倾向性分析 自动文摘

1 引言

截至2008年2月底，我国网民总人数已达2.21亿，超过美国位居全球第一。作为继报纸、无线广播和电视这三大传统的传播媒体之后出现的新兴“第四媒体”，互联网已逐渐确立起其社会信息传播的主导地位，成为庞大的公共信息集散地和民众参政议政最常用的平台。社会民众通过网络所表达的群体性的情绪、态度、意见与要求等形成了网络舆情，是社情民意中最活跃、最尖锐的一部分，最直接、最快速地反映了社会各个层面的舆情状况与发展态势，对社会产生的影响面和影响力越来越大。

然而，互联网普及是一把双刃剑，对社会产生着正负两方面的影响。一方面，它对于信息的传播，特别是一些重要的新闻事件和社会突发公共事件的报道（如抵制家乐福、3.14西藏拉萨打砸抢杀事件等）表现出传统舆论无法与之相比的优势：民众可以实时参与对事件的讨论，进而引导和影响事态的进程和发展；另一方面，由于绕过了传统舆论管理的“把关人”等程序，使得一些民众通过网络散布谣言与虚假、低级庸俗的信息与灰色的言论，而且西方敌对势力也借助网络串联对我国日益发起“和平演变”攻势。

我国当前所面临的网络舆情分析与预警的形势极其严峻，政府及相关管理者亟需对处于“未然态”的舆情信息进行挖掘与分析，把握处理危机事件的最佳时机。然而仅依靠人工方法难以应对海量网络信息的收集和处理，需要多种信息分析技术，分析当前网络的舆情动态，对网络的热点、焦点与敏感话题及时做出反应，从而提高处置网络突发事件的能力和监管能力。

在网络舆情监控分析与预警方面，网络舆情信息获取的快与准、内容分析的确定性、舆情研判的准确性、舆情响应及时性、信息跟踪的及时性等目标的实现是网络舆情分析研究的重中之重。舆情监测分析的核心技术在于舆情分析引擎，涉及的最主要的技术包括文本分类与聚类、主题检测与跟踪、观点倾向性识别、自动摘要等计算技术。这些技术一向是国内外信息工作者关注的领域。下文主要阐述网络舆情分析中的四个实用关键性技术。

2 网络信息采集与提取

在真实网络环境下，由于网页设计的灵活性所造成的复杂性、网页结构的更新频率非常高所带来的动态性、待抽取内容的多态性、网页技术屏障以及网页的不完整性等特点，使得通过链接网页浏览或关键词检索等方式来获取信息的手段显示出缺陷。为了解决这个问题，出现了高效的、具有一定智能的网络信息采集与提取技术，来整合纷繁的网络信息资源。网络信息抽取属于网络内容挖掘（Web content mining）研究的一部分，主要包括结构化数据抽取（Structured Data Extraction）、信息集成（Information integration）和观点挖掘（Opinion mining）等[1]。

国内外在网络信息采集与抽取领域的研究主要集中在：如何建立针对各类网站的全自动化信息抽取工具，并将这些信息按照一定的格式进行整合，支持各类计算机应用：传统的网络数据抽取方法是针对抽取对象手工编写一段专门的抽取程序，这个程序称为wrapper。近年来，越来越多的网络数据抽取工具被开发出来，替代了传统的手工编写wrapper的方法。目前较为流行的网络数据抽取工具可分为以下六大类[2]：

（1）开发wrapper的专用语言：用户可用这些专用语言方便地编写wrapper。例如Minerva，TSIMMIS，Web-OQL，FLORID，Jedi等。

（2）以HTML为中间件的工具：这些工具在抽取时主要依赖HTML文档的内在结构特征。在抽取过程之前，这些工具先把文档转换成标签树；之后工具根据标签树自动或半自动地抽取数据。代表工具有W4F，XWRAP，RoadRunner，MDR。

（3）基于NLP的工具：这些工具通常利用filtering、part-of-speech tagging、lexical semantic tagging等NLP技术建立短语和橘子元素之间的关系，从而推导出抽取规则。这些工具比较适合抽取那些包含符合文法的页面，比如工作列表等。代表工具有RAPIER，SRV，WHISK。

（4）Wrapper的推导工具：Wrapper的推导工具从一组训练样例中推导出基于分隔符的抽取规则。这些工具和基于NLP的工具之间最大的差别在于：这些工具不依赖于语言约束，而是依赖于数据的格式化特征。这个特点决定了这些工具比基于NLP的工具更适合于抽取HTML文档。代表工具有WIEN，SoftMealy，STALKER。

（5）基于模型的工具：这些工具让用户通过图形界面，建立文档中其感兴趣的对象的结构模型，“教”工具学会如何识别文档中的对象，从而抽取出对象。代表工具有NoDoSE，DEByE。

（6）基于本体的工具：这些工具首先需要专家参与，人工建立某领域的知识库，然后工具基于知识库去做抽取操作。如果知识库具有足够的表达能力，那么抽取操作可以做到完全自动。而且由这些工具生成的wrapper具有比较好的灵活性和适应性。代表工具有：BYU，X-tract。

在实际工程应用中，对于网络信息采集与提取工具的定性分析必须考虑到以下六个指标：（1）自动化程度：这是个非常重要的指标。它意味着在生成wrapper的同时，需要用户参与的工作量；（2）是否支持复杂结构对象的处理；（3）是否支持页面的文本分析；（4）是否提供图形用户界面；（5）是否支持非HTML文档；（6）灵活性和适应性。

目前尚没有一个工具可以适应所有的数据抽取需求。近年来研发高度自动化的抽取工具成为不少研究者关注的热点，例如比较流行的全自动的抽取工具MDR、RoadRunner，但这些工具仍然存在不足之处：在线数据抽取、在线数据集成的速度不够快，且处理的准确率不高；抽取方法的通用性有限，对结构化程度较松散的网页处理不好；另外，对于大多数工具过多地考虑了抽取的自动化程度，但较少考虑到抽取的效率等问题，而该问题则是实际工程应用中至关重要的问题。

在此特别需要提出的是对于承载着松散的结构化信息的动态网页，例如博客、论坛信息的提取与采集技术。博客作为一个巨大的知识库，如何从博文中获取重要的信息成为目前信息检索领域一个新的研究课题。与新闻网页不同，博客正文分成文章和评论两部分。目前现有的一些正文抽取算法和主题划分算法都很难对其进行精确的定位和切分。因此，在博客信息抽取方面，需要研究新的定位和切分算法以适应博客检索的需求[12]。除了博客之外，论坛信息的抽取也是一项非常复杂的工作。在论坛中有一种非常重要的页面，称为版面页面，对其信息的抽取也是一大研究重点[13]。

3 话题发现与跟踪

话题发现与跟踪是一项旨在依据事件对语言文本信息流进行组织、利用的研究，也是为应对信息过载问题而提出的一项应用研究。目前研究中采用的主要表示方法多种多样，但主流的模型有两种：基于向量的模型和基于概率的模型。基于向量的表示就是把所有待处理数据表示为向量，判断两个文档是否讨论同一个话题是通过计算两个向量之间的相似度来完成，而基于概率的表示则是把文档表示为词的概率模型或N元语言模型，通过计算话题T与文档d的生成概率P(d|T)来判断两者之间的关系。

主流的话题发现算法都采用文本聚类技术来实现，该类算法的主要问题就是准确率低、大类现象比较严重。在早期的网络话题相关研究中，为了简化问题，一般假定所有的话题没有层次之分，而且一个文档只能与一个话题相关[3]。但随着研究的深入，从2003年开始，层次化话题发现作为话题发现与跟踪领域一个全新的研究问题被提了出来，它突破了传统的话题组织忽略话题多粒度现象的不合理之处，采用层次化的结构对话题进行组织。如图1所示

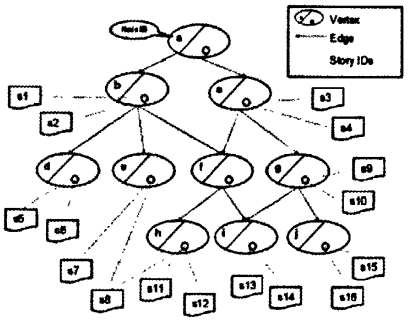


图1 层次话题发现的有向无环图示例

从应用角度来看，层次化话题识别突破了传统的“一类一团”的结果呈现方式，用户可以通过图形化的层次结构进行浏览，有利于减少阅读的工作量。正是基于这样的特点，使得层次化话题识别更能体现事件信息组织的本质，因而可以采用更加有意义的系统评测方式。

参考文献[14]提出了一种基于话题特征统计的热点话题发现方法，这一研究思路实现将文本聚类问题转换为话题特征聚类问题，也大大提高了话题的准确率和可读性。该方法分为文本预处理和话题发现与分析两个步骤。在文本预处理阶段，首先对文本进行分词，建立索引，将文本存入我们的索引库中。在分词的基础上，提取文本中的关键词列表，用一定数量的关键词来表示该文本信息。需要注意的是，文本的关键词是文本的特征，但不是话题的特征。如“中国”、“足球”等不能构成一个话题特征，因为它们对话题的刻画都过于空泛。不妨采取两两关键词组合的方式构建话题特征。比如每个每篇文本提取5个关键词，就可以得到10个话题特征（5个关键词两两组合，共有10种），然后将该话题特征存储到话题特征库中。话题特征库是话题特征构成的一个图结构。在该图中，每个话题特征构成一个图节点，节点之间的边表示两个话题之间的关系。在边和节点上都有一个权重，分别表示两个话题特征共现的频率和话题特征出现的频率。对于每篇报文，统计出其话题特征后，就修改话题特征库中的图结构，分别修改

节点的权重和边的权重。在统计的过程中就可以发现当前的热点话题特征，再然后利用检索系统检索出各话题特征对应的文本列表，然后将这些列表合并在一起就构成了该话题的所有文本列表，分析文本的时间信息，就可以知道该话题的源头信息，以及话题的演化信息。

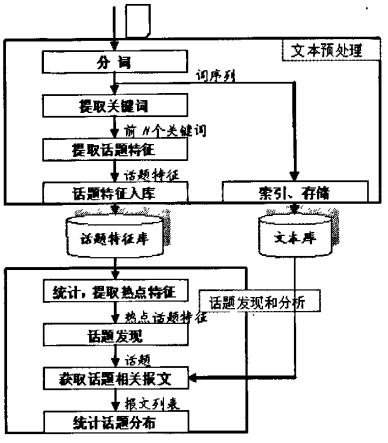


图2 热点话题发现的处理的流程图

4 网络文本的倾向性分析

由于网络的虚拟性和匿名性，使得网络文本内容在大多数情况下真实地表达出了民众的态度和情绪，通过倾向性分析可以明确网络传播者的意图和倾向。通俗地说，文本舆情描述的是文本所传递的情感。对文本舆情进行分析，实际上就是试图根据文本的内容提炼出作者的情感方向，但是我们希望这项工作可以由计算机帮助我们实现。网络文本的倾向性分析就是挖掘网络文本内容蕴含的各种观点、喜好、态度、情感等非内容或非事实信息。我们不仅需要掌握网络文本的影响强度，同时还需要对文本的感情取向有一个正确的把握；如果我们需要对每一个文本赋予一个值，那么影响强度可以看成是其绝对值的大小，而舆情可以看成是其正负号[15]。对大规模评论页面进行有效的倾向性分析具有很好的现实意义。

迄今为止，国内外所从事的网络文本倾向性分析研究工作可归纳为以下几个方面：（1）客观性分类：从Web上获取的评论文档按照类型和风格的不同区分为主观和客观两类，这类工作以Finn等人为代表，其结论是基于词性标注的特征选择方法比词袋方法效果好。Wiebe等人对人工标注的语料从短语、句子和篇章层次进行研究，发现对于不同的标注者，其主观性的判别有较大差异；（2）词的极性判别：即通过分析带有语气词的特征来判断词的极性。Hatzivassiloglou和McKeown使用关联词（如公平并合法，简单却受欢迎）来区分含义相近或相反的词。Turney和Littman提出了一种方法，他们使用AltaVista中的NEAR运算从Web上搜索得到两个词同时出现的次数，以此来决定两个词的相似程度，一个新词归属于正面语气还是负面语气，取决于它和手工选择的正面（或负面）种子词集合中所有词的关系，这类工作和常规的词聚类问题有一定的关联。Lin和Pereira等人使用语言学同位关系把用法和意义相似的词进行了归类；（3）语气分类：①基于语气标注的方法：加拿大Ottawa大学的Kennedy、加拿大国家研究委员会的Turney等提出语气词标注方法，对常用词汇进行语气标注，如（“好”标为正面，“坏”标为负面）。分类时直接统计一篇评论中的正面与负面语气词的个数，正面语气词多则判为正面，负面语气词多则判为负面，相等则判为客观。②基于语义模式分析的方法：Tetsuya Nasukawa和Jeonghee Yi等通过识别特定主题词和语气表达式之间的语义关系进行倾向性分析。Jeonghee Yi等人采用自然语言处理技术分析特定主题和语气词之间的语义关联。③基于机器学习的方法：其思想是直接利用传统的机器学习方法来训练语气分类器。康奈尔大学的Lillian Lee和Pang Bo等人以Usenet上的电影评论作为语料进行了研究，采用了不同的特征选择方法和机器学习方法。其实验结果显示，基于presence-based frequency模型选择UniGrams的方法，并采用Support Vector Machine（SVM）进行分类，能取得最好的分类结果，其准确率为82.9%。

倾向性分析面临的主要问题是目前的大部分方法和技术都和领域或话题相关，局限在某个特定领域或者关联于某个话题下进行倾向性的分析，缺乏一般性的通用技术。基于语气词标注的方法严重依赖于标注专家且不利用训练样本，其分类精度往往不如基于机器学习的方法。而基于机器学习的倾向性分析方法又取决于训练集的大小与质量，同时具有很强的领域或主题依赖性，由于已有的标注语料库的规模都很小，因而这类有监督的语气分析方法的效果仍然难以保证。基于语义模式分析的方法则受限于自然语言处理技术的不够成熟而很难实用。中文倾向性分析方面的情况则更加突出，一些基本问题尚未得到圆满的解决：（1）各种有监督的机器学习方法在中文数据集上的语气分类效果孰优孰劣；（2）文本特征表示方法和特征选择机制等因素对中文语气分类的性能将产生什么影响；（3）文档集的哪些语气特征对语气分类的精度具有决定性影响等。

因此，为解决上述问题，应着重研究倾向性主客观过滤技术和观点极性、强度、情感分析判别技术：研究网络环境下倾向性特征词的特点和类型，并进行语气极性判别和标注，从而构建一个面向互联网的倾向性语气词典，建设一定规模的标准数据集，为中文倾向性分析的深入研究和公开评测提供支持。

5 多文档自动文摘

多文档自动文摘技术是一种提炼概要信息的有效手段，已经被进行了广泛的研究。传统的多文档文摘技术是一种静态文摘，即针对某个封闭的静态文档集生成摘要，不考虑文档集的对外联系。但是在Web2.0时代，网络信息内容的动态演化性越来越明显，出现在BBS论坛、Blog、在线评论等新媒体中的网络信息（如网络话题、热点事件等，表现为一系列相关文章的集合）是动态演化的，它们随着时间的变化而出现、发展直至消亡，一个话题在不同的时刻具有不同的侧重点，而不同时刻的话题内容之间具有关联性。因此，如何对动态演化的网络信息进行文摘成为一个新的研究课题。动态文摘是传统静态文摘的延伸和扩展，除了需要保证文摘信息的主题相关性和内容的低冗余性外，还需要针对内容的动态演化性分析已出现信

息和新出现信息的关系，消除旧信息，摘要新信息，使文摘随话题的演化而动态更新。

动态内容的时序划分是动态文摘的基础，相关研究在新闻事件检测[4]和TDT[5][6][7][8]等领域得到了较多关注，Mani等人使用时域分析方法对新闻事件的内容进行分析[9]，James Allen等人借用图形学领域的时间线的构建来进行内容划分，并在TDT研究的基础上，探讨了基于内容有用性(useful)与新颖性(novel)的时域文摘研究方法[10][11]，其提出的时序文摘不同于本文的动态文摘，本质上是一种基于句子排列策略改进的静态文摘。DUC 2007国际评测的先导任务Update Task实际上就是一个动态文摘问题。这一任务主要来源于信息检索系统、问答系统和文摘系统中对用户行为的模拟。该任务假定用户对某个场景已出现的内容有了足够的了解，在后续文摘中重点关注那些新出现的内容。因此，当得到与此场景相关的新信息时，需要分析新信息与旧信息的关系并生成更新文摘。

参考文献[16]针对内容的动态演化性提出了三种动态文摘模型：(1)文档过滤模型(Document Filtering Model，DFM)，即从文档内容过滤的角度提取动态信息以生成文摘；(2)文摘过滤模型，即首先利用静态文摘方法对当前信息生成候选文摘，然后再从候选文摘中过滤掉与历史信息的重叠内容，从而得到所需的动态文摘；(3)合并过滤模型，它对前两种模型做出了改进，强调了当前信息与历史信息二者之间的关联性，首先对历史信息和当前信息合并的全文档空间生成文摘，再从中进行历史信息的过滤，从而生成动态文摘。针对以上三种动态文摘模型，基于模糊隶属度给出了具体的动态文摘生成方法。在DUC 2007测试数据上的实验证明了本文所提出动态文摘模型及生成方法的有效性。

如何分析不同的动态内容时域分析方法，将演化信息的内容差异性和主题相关性进行更好的结合是后续的主要研究工作。

6 总结

本文主要阐述了网络舆情分析中的四个使用关键技术：首先对相应的关键性技术以及国内外相关研究现状做了介绍，然后阐述了在实际的研发工作中遇到的问题及有效的解决思路包括工具和方法等。那么，如何将这些关键技术整合成一套自动化的网络舆情分析与预警系统，从技术上保障不同的业务部门之间形成统一协调的舆情处理和联动机制，为国家相关管理者提供一个基础性平台，是网络舆情研究工作的下一步重点。（责编 张岩）

参考文献：

[1] B. Liu. ACM SIGKDD Inaugural Webcast: Web Content Mining, Nov 29, 2006.

[2] A. Laender, B. Ribeiro-Neto, A. Silva, and J. Teixeira. A brief survey of web data extraction tools. ACM SIGMOD Record, 31(2):84-93, 2002.

[3] Allan James, Ron Papka & Victor Lavrenko, 1998. On-line New Event Detection and Tracking. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 37-45. Melboune, Australia.

[4] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection [A]. summer workshop at CLSP [C], 1999.

[5] http://www.nist.gov/speech/tests/tdt/

[6] J. Allan, R. Papka, and V. Lavrenko. On-line New Event Detection and Tracking [A]. In Proceedings of SIGIR [C], pp. 37-45, 1998.

[7] Bin Cao, Dou Shen, Jian-Tao Sun, Xuanhui Wang, Qiang Yang, and Zheng Chen. Latent Factor Detection and Tracking with Online Nonnegative Matrix Factorization [A]. In Proceedings IJCAI [C], pp. 2689-2694, 2007.

[8] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining [A]. In Proceedings KDD [C], pp. 198-207, 2005.

[9] I. Mani and G. Wilson. Robust temporal processing of news [A]. In Proceedings of ACL [C], pp. 69-76, 2000.

[10] James Allan, Rahul Gupta, Vikas Khandelwal. Temporal Summaries of News Topics. In Proceedings of SIGIR [C], pp. 10-18, 2001.

[11] R. Swan and J. Allan. Automatic generation of overview timelines [A]. In Proceedings of SIGIR [C], pp. 49-56, 2000.

[12] 曹冬林，王宇，郭岩．博客正文提取．

[13] 郭岩．论坛信息抽取工具EMIB —— Extract Meta Information from Board page

[14] 段建国，丁国栋，程学旗，基于话题特征统计的互联网热点分析技术．

[15] 陈华，梁循，阮进．网络舆情关联分析系统的设计实现

[16] 张瑾，许洪波，程学旗．面向网络演化信息的动态文摘方法研究

作者简介：戴媛(1984-)，女，硕士。程学旗(1971-)，男，研究员，博士生导师，中国科学院计算技术研究所，主要研究领域为网信息安全、大规模信息检索与信息挖掘等方面的研究。

启明星辰公司推出Web业务安全解决方案

随着应用业务的Web化进程日渐深入，针对Web业务的攻击增长迅猛，利用SQL注入等攻击手段而对网站进行网页篡改、窃取数据等安全事件频频发生。针对这一现象，国内著名网络安全公司启明星辰推出web业务安全解决方案，提供对SQL注入、XSS(跨站脚本攻击)、DOS/DDOS等多种web业务威胁的防御措施，保障企业web业务系统的安全。

根据世界上最知名的Web安全与数据库安全研究组织OWASP提供的报告，目前对Web业务系统威胁最严重的两种攻击方式是SQL注入攻击和跨站脚本攻击。对于这两种攻击手段，利用传统基于攻击特征匹配的方法进行检测收效甚微，存在大量的漏报和误报。启明星辰公司作为国内最专业的入侵检测研究安全厂商，在网络入侵研究方面有深厚的技术积累，采用融合基于原理和基于特征的柔性化检测机制来解决Web攻击的防御问题，独创出基于攻击手法的VXID专利检测算法，并将这一技术专利应用于天清入侵防御系统(IPS)，推出Web业务安全解决方案，有效解决了当前困扰广大用户的Web业务安全问题。

(编辑 杨晨)