解读大数据

邓莎莎 著

支持决策研讨的文本分析方法研究

中西書局

解读大数据

邓莎莎 著

支持决策研讨的文本分析方法研究

中西書局

图书在版编目(CIP)数据

解读大数据：支持决策研讨的文本分析方法研究 / 邓莎莎著.—上海：中西书局，2017.10
ISBN 978-7-5475-1326-2

Ⅰ.①解… Ⅱ.①邓… Ⅲ.①数据处理—研究 Ⅳ.①TP274

中国版本图书馆CIP数据核字（2017）第215486号

本书出版得到上海汽车工业教育基金会资助

解读大数据

——支持决策研讨的文本分析方法研究

邓莎莎 著

责任编辑	伍珺涵
装帧设计	梁业礼
出版发行	上海世纪出版集团 中西書局(www.zxpress.com.cn)
地　　址	上海市陕西北路457号(200040)
印　　刷	上海世纪嘉晋数字信息技术有限公司
开　　本	700×1000毫米　1/16
印　　张	14
字　　数	156 000
版　　次	2017年10月第1版　2017年10月第1次印刷
书　　号	ISBN 978-7-5475-1326-2/T·014
定　　价	49.00元

本书如有质量问题，请与承印厂联系。T：021-69214196

表格目录

插图目录

目 录

第 2 章 支持在线研讨意义构建的文本分析方法研究

第 3 章　LTAS 系统的实验与评估

第4章 在线研讨过程中欺骗识别研究

第5章 面向研讨问题的网络民意分析研究

第6章 结论

第1章 绪论

1.1 研究背景

时至今日，互联网、物联网和云计算等信息和通信技术(Information and Communication Technologies，ICTs)在社会中的渗透和应用是如此深广。中国互联网络信息中心(CNNIC)于2013年1月15日在京发布的《第31次中国互联网络发展状况统计报告》(以下简称为《CNNIC第31次调查报告》)显示，截至2012年12月底，中国网民数量达到5.64亿，互联网普及率为42.1%，较2011年底，网民增量为5 090万，普及率提升3.8个百分点，手机超越台式电脑成为中国网民第一大上网终端。到2012年底，通过台式电脑，包括笔记本上网的网民占上网网民的70.6%，这个数字比2011年下半年的统计数字低了近3个百分点。相比而言，手机上网人数则是增加了74.5%，而这一数字已经超过了使用台式机器的上网人数，截至2012年底，我国利用手机上网的网民规模已经达到了4.2亿。智能手机的普及与移动互联的不断发展为中国广阔的农村地区以及庞大的移动人口提供了更为廉价和简便的接入互联网的方式。

伴随着Web2.0技术的迅速发展，各种提供交流与沟通的社会化媒体(Social Media)平台不断涌现。无论是协同社区(例如：Wikipedia、博客、微博和网络论坛等)和社交网络社区(例如：Facebook、人人网、开心网等)，还是各种特定主题的内容社区(例

如：Youtube、豆瓣、点评网等），用户已经从被动地接收信息转换成主动地输出信息，用户的参与度与贡献度已经显著提高。此外，随着移动互联的普及，利用手机即时通信的移动性、碎片性以及随时在线的特点，各种社会化媒体平台纷纷植入了移动互联技术。这种方式使得个人与社区之间的共享、共同创造以及修改用户产生内容变得更加频繁和容易。

社会化媒体的快速增长带来了人们日常交流、决策讨论方式的巨大改变。面对面的会议方式不再是组织内部讨论问题、制定决策的唯一沟通手段。使用社会化媒体技术来支持与商业相关的决策研讨的方式已经被越来越多的组织所采用（Mann 2011a）。根据最近麦肯锡季刊（*McKinsey Quarterly*）的报道，在超过 1 700 个受访的组织中，50% 的组织正在使用社交网络（Social Networking），41% 的组织在使用博客（Blogs），25% 的组织正在使用维基（Wikis），还有 23% 的组织在使用微博（Micro Blogs）。在过去的 4 年中，这些数字正在以超过 2 倍的速度增长。（Bughin et al. 2010）

无论是在组织内部还是组织外部，社会化媒体平台支撑的在线研讨方式从各个方面为组织带来了巨大的收益。在组织内部，社会化媒体平台能够更加灵活地支持组织或者部门内部的决策研讨，提高决策效率，从而加快组织获取知识的速度，降低沟通和操作成本，提高内部专家的识别率并增加员工的满意度。另外，社会化媒体已经逐渐成为与组织外部的客户、商业合作伙伴进行沟通的重要手段。支持组织内部成员与外部顾客的沟通能够有效地提高顾客对组织的知晓度以及忠诚度，降低营销费用和客户支持成本，并且增加新产品或者服务的创新成功率。（Bughin et al. 2010）

尽管社会化媒体技术支持的在线研讨给组织带来的价值巨大，并且已经有很多组织从这种交流沟通方式中得到了实实在在的收益，然而，在互联网环境中研讨文本分析面临着以下四个方面的挑战。

第一，数据量过大且内容复杂，而这超出了人们处理信息的能力。

社会化媒体平台支撑的在线研讨，常常随着事件的发生在短时间内发言数激增，一时间相关信息铺天盖地地出现在互联网中，各种累积的发言数量越来越庞大。因此，仅靠人工阅读所有研讨文本内容，继而综合分析并形成判断进而提供决策建议是一件耗时耗力同时也是非常困难的任务。根据 Hiltz 和 Turoff 的研究，当信息量达到某一信息熵(information entropy)时，群体或者个人就无法有效地组织和理解输入信息。(Hiltz et al. 1985)这时，人们就会抛弃所有信息而完全依靠对前面部分信息产生的主观印象和感受进行判断和决策。因此，仅靠人工处理这些海量信息不仅给处理人员带来极大负担，更重要的是，在很多时候这还是一件不可能完成的工作。

第二，如何快速有效地弄清发言之间的关系，并获得语言文字所要表达的真正意图？

首先，社会化媒体平台支撑的在线研讨产生的数据都是非结构的语言文字，并且这些内容并不是由一个人独立完成，而是由许多人协作完成的。譬如，对某一个问题的讨论常常是由微博博主与成千上万的粉丝互动完成的，论坛中对于热点问题的讨论常常会吸引成千上万的人参与，这时，网络文本已经被大量发言撕扯得支离破碎，上下文之间毫无逻辑关系，难以读懂。其次，语言文字