开发商业智能应用小心大数据“陷阱”

黄锦辉
香港中文大学工程学院，香港 999077

摘要
大数据的应用和研究是信息爆炸时代的热点话题。就如何更智能地发现大数据中的有用信息展开讨论，探讨了大数据中的“陷阱”和其引发的社会危害，提出一种面向社交文本的智能应用系统，以有效规避大数据中的“陷阱”并自动提取有用信息；基于提到的框架，展示了笔者研究组近些年在社交媒体上的事件检测、自动摘要和谣言检测方面的研究成果。

关键词
大数据；自然语言处理；社交媒体；数据处理

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017016

Beware of traps of big data analytics in business

WONG Kam Fai
Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China

Abstract
In the era of data explosion, research and application of big data has become a hot topic. How to automatically discover useful information from big data was focused. The organization is as following: examples of big data “traps” and their influences were discussed. The framework of an intelligent system to process social media texts that avoids traps and extracts useful information from big data was described. The research works proposed by our team and based on the framework about event detection, summarization and rumor detection were covered.

Key words
big data, natural language processing, social media, data processing

1 大数据中的“陷阱”

自从前美国总统奥巴马2012年3月推出2亿美元的“大数据研究和发展计划”后，世界各大小经济体陆续效仿，大力投资相关领域。全球信息科技企业亦不敢怠慢，积极推出适合的大数据信息科技方案及产品，更大洒金钱地推广大数据分析的优点及其所能带来的商机。据观察，近期不少从事金融、医疗、社会工作、工商业、政务等范畴的主管都已被潜移默化，鼓吹大数据的功能及效益。然而，大数据真的是万能的吗？目前为止，大数据的处理和应用仍存在如下问题。

（1）采集数据时“南辕北辙”

采集的数据应与目标相符，并不是越多越好，如果南辕北辙将得出错误的结果。题为“谷歌流感的寓言：大数据分析的陷阱”的报告描述了谷歌公司曾利用大数据分析推算2011—2012年度美国流感的趋势[1]，但结果却不如人意，估计的流感个案数目远超过实际数目。而谷歌公司利用的数据来自用户使用的关键词（如“禽流感”）次数及分布，进而进行推算分析。专家认为构成严重误差的主要原因是谷歌公司盲目地广泛收集关键词，以为越多越好，却没有了解用户查询时的出发点，结果收集的数据大部分来自非流感病患者，因此在数据采集阶段已严重犯错，自然推算失准。若数据分析全力集中在流感病患者，结果便会截然不同。

（2）计算模型选择错误

数据量过大时质量也有一定程度的下降，面对海量规模的数据，人工去噪已经很难实现，这时候如果不能选择正确的计算模型，将在分析结果中引入噪声。美国加州大学伯克利分校的国际知名学者米高·佐敦（Michael Jordon）教授，最近接受美国IEEE学会杂志访问时指出，“大数据”在现今商业市场被过分炒作，它最后可能只是一场空欢喜，米高·佐敦教授更预测“大数据”的“冬天”即将来临[2]。他认为“大数据”用户进行假设的速度将会超越“大数据”的统计范畴，在这种情况下数据分析结果难免会出现错误，造成大量噪音，影响推算的可靠性。

（3）数据处理不能“与时俱进”

大数据用户往往忽略数据的“动力（dynamics）”。在大数据时代，数据和信息的更新换代十分迅猛，随着时间的推移，需求也在变化，过时的数据对当下的需求来说可能没有任何作用。例如在变幻无常的商务环境中，用户的需求不停在变，那么昨天的大数据分析结果能有效地应用于今天的商务环境吗？能够满足用户今天的需求吗？若不能，则需要重新进行分析，但昨天采集商务数据的方法能满足用户今天的新需求吗？归根究底，什么时候开始和停止分析既是统计学应用的老问题，亦是大数据分析必须严肃面对的问题，但在千变万化的应用及数据环境下，要应对这个问题更是难上加难。因此米高·佐敦教授进一步指出，大数据分析服务提供商有责任清楚说明分析推算法的质量标准及其误差度，做好用户的“期望管理（expectation management）”。

（4）大数据环境下的人才荒

今天的信息科技业界急切需要解决“人才荒”问题。缺乏专业的数据科学家（data scientist），大数据行业将难以起飞。以“瞎子摸象”比喻，多个瞎子（非专家）触摸大象（大数据），然后判断象的形状，结果以误判收场。在商业应用中，错误的大数据分析推断的商务智能（business intelligence，BI），可能会弄巧成拙，严重影响生意。笔者认为大学应考虑开设更

多的相关课程，校方在课程设计上必须加入更多的体验学习（experience-based learning）活动，避免就读“数据科学”的学生过于纸上谈兵。

综上，大数据虽然是块诱人的蛋糕，却也充满“陷阱”，一步踏错，满盘皆输。下文将探讨如何规避陷阱，从大数据中提取有用信息。

2 基于社交媒体的大数据智能处理系统

大数据处理的研究不应是空中楼阁，一个好的大数据处理系统应有效地帮助用户过滤无用或是错误信息，从而最大化地获得有用信息。笔者研究组致力于开发规避“陷阱”的大数据智能处理系统，该系统主要分为事件检测（event detection）、自动摘要（summarization）和谣言检测（rumor detection）3个模块，如**图**1所示。

社交媒体上，数据的更新换代迅猛，信息极易过时，而过时的信息并非用户所需要的，例如，关于“德国欧洲杯”的讨论已非当今大众所关心，人们更加在意“法国欧洲杯”上的比武论剑谁将笑到最后。自动事件检测模块将从海量原始数据中抽取大众最为关心的时下热点，淘汰过时信息。然而，诸如“法国欧洲杯”之类的热点事件通常能吸引成千上万条的讨论，其中充满了重复和无用的信息，此时，自动摘要信息模块将取其精华，去其糟粕，自动生成关于事件的简练描述，让用户能够在有限的时间之内快速了解事件的来龙去脉。有了热门事件和有关它们的描述，还需要判断眼前的事件是否真实，避免被虚假信息所扰。这时候，谣言检测模块能智能地给予事件真实与否的判断，过滤掉其中的不实信息，从而推送用户所需的信息，智能地避开大数据中的“陷阱”。

3 自然语言处理与社交媒体中的大数据研究

以微博为代表的社交网络的迅速崛起和流行引发了信息的爆炸，利用自然语言处理技术自动提取最有用、最可信的内容，让用户能够在有限的时间内获得最需要的信息是大数据时代火热的研究课题之一。基于第2节中提出的框架，本节将详细介绍笔者研究组在事件检测、自动摘要、谣言检测方面的研究工作。

（1）事件检测

微博已经超越报纸等传统媒体，成为

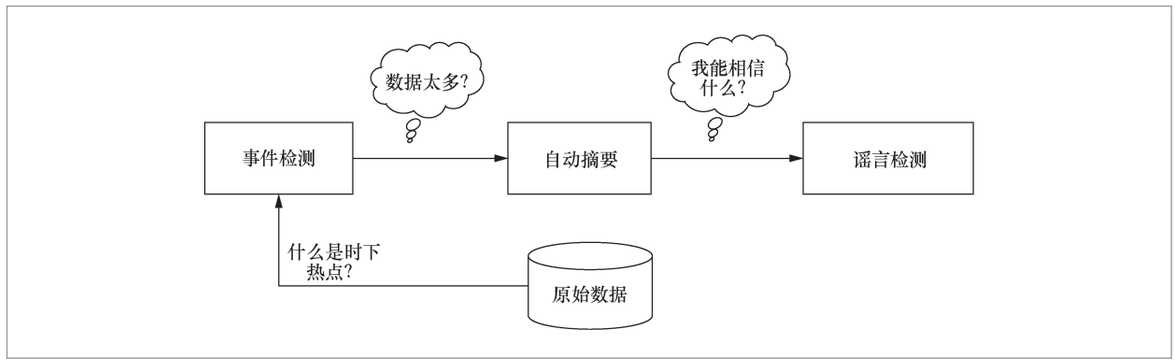

图1 基于社交媒体的大数据智能处理系统框架

热门事件播报的最快、最重要的传媒，如何从微博中自动地发现用户需要和感兴趣的事件成为如今数据挖掘领域的一个重要研究方向。利用自然语言处理技术对语言和语义的分析，对热门事件的自动发现有着重要的作用，Ou等人[3]和Peng等人[4]对微博中的情感进行建模，通过检测情感的爆发来发现热门的事件。

（2）自动摘要

自动摘要技术能够自动地从海量文本中挖掘重要信息，这对用户快速了解社交媒体事件有着十分重要的作用。面向传统文本的自动摘要方面的研究已经进行多年，并且在方方面面都相当成熟，然而在面对微博文本时却面临了严峻的挑战，主要原因是词语的共现模式在短小而不规范的微博文本中表现并不明显，这又恰恰是自动摘要技术的技术基础。笔者研究组致力于从微博的转发和评论内容寻找上下文信息（context information）的补充，并挖掘其中的语义信息，从而发现更多更有用的词语共现特征，以提高自然语言处理模型的性能。在Li等人[5, 6]的研究中，以回复和转发关系为基础构建微博对话树（microblog conversation tree），借助树的结构信息将其中的微博分为领导者（leader）和追随者（follower），并对两类微博文本之间的关系进行建模，为微博自动摘要的研究提供了新的思路。

（3）谣言检测

虽然热门事件在大多数情况下为用户所需，然而，在许多情况下并非如此。谣言是热门事件中特殊的一类，作为虚假信息，谣言传播迅速，易于引起大众恐慌，给社会带来了严重的危害。如果谣言能在传播初期就被系统检测到，就能够有效缩小其恶劣的社会影响范围。Ma等人[7, 8]发现谣言和真实事件在许多方面时序特征都呈现出不同，并首次采用深度学习（deep learning）的方法建模谣言和真实事件的时序变化，从而从热门事件中自动发现谣言，避免对社会造成危害。

4 结束语

本文详细介绍了大数据的处理和应用中存在的问题，并提出了基于社交媒体的大数据智能处理系统。随着大数据行业的不断发展，笔者在自然语言处理和社交媒体中的大数据研究方面还会进行更深一步的探索。

参考文献:

[1] LAZER D, KENNEDY R, KING G, et al. The parable of Google flu: traps in big data analysis[J]. Science, 2014, 343(6176): 1203-1205.

[2] LEE G. Machine-learning maestro Michael Jordan on the delusions of big data and other huge engineering efforts[J]. Communications of the ACM, 2011.

[3] OU G, CHEN W, WANG T, et al. Exploiting community emotion for microblog event detection[C]//EMNLP, October 25-29, 2014, Doha, Qatar. [S.l.:s.n.], 2014: 1159-1168.

[4] PENG B, LI J, CHEN J, et al. Trending sentiment-topic detection on twitter[M]. [S.l.]: Spring International Publishing, 2015: 66-77.

[5] LI J, GAO W, WEI Z, et al. Using content-level structures for summarizing microblog repost trees[C]//EMNLP, September 17-21, 2015, Lisbon, Portugal. [S.l.:s.n.], 2015: 2168-2178.

[6] LI J, LIAO M, GAO W, et al. Topic extraction from microblog posts using conversation structures[C]//The Meeting of the Association for Computational

Linguistics, August 7–12, 2016, Berlin, Germany. [S.l.:s.n.], 2016: 2114–2123.

[7] MA J, GAO W, WEI Z, et al. Detect rumors using time series of social context information on microblogging websites[C]//CIKM,October 19–23, 2015, Melbourne, Australia. New York:ACM Press, 2015: 1751–1754.

[8] MA J, GAO W, MITRA P, et al. Detecting rumors from microblogs with recurrent neural networks[C]//IJCAI, July 9–15, 2016, New York, USA. [S.l.:s.n.], 2016.

作者简介

黄锦辉(1960–),男,博士,香港中文大学工程学院副院长(外务)、系统工程与工程管理学系教授及创新科技中心主任,并担任哈尔滨工业大学深圳研究院特聘教授、北京大学信息科学技术学院客座教授以及东北大学兼职教授。香港信息科技联会、香港工程师学会、英国工程技术学会及英国计算机学会会员。主要研究方向为数据库及中文信息处理,并在多份国际刊物、会议及书籍中发表过超过250份技术研究报告。是《ACM Transactions on Asian Language Information Processing》的创办人及总编辑,也是《Computer Processing of Oriental Languages》及《Computational Linguistics and Chinese Computer Processing》等国际学术刊物的编辑小组成员。是亚洲自然语言处理协会(AFNLP)创会成员及现届会长(2015—2016年)。同时也是亚洲信息检索会议(AIRS)系列的创办人。

收稿日期: 2016-08-22

2017016-5