使用主题建模进行博客文本分析，命名实体识别和情感分级机

普拉纳夫·瓦伊拉* ，VK Singh**和 MK Singh* DST-CIMS,印
* 度瓦拉纳西贝拿勒斯印度教大学 南亚大学计算机科学系,印度新德
** 里pranav.waila@gmail.com、 vivek@cs.sau.ac.in、 mks_kjist@yahoo.co.in

摘要- 本文介绍了我们通过新颖的语言处理和可视化技术组合对社会政治博客数据进行计算分析的实验工作。我们利用主题建模、实体提取和情感分析设计了一个集成框架,从非结构化的自由形式博客圈数据中得出与社会相关的推论。该数据集包含 9290 多篇与阿拉伯之春有关的社会政治事件的博客文章。我们试图从数据集中提取重要的推论,例如文本中围绕不同实体的内容的关键主题、人物、地点、组织和整体情感取向。我们试图验证通过手动和 Google 搜索趋势统计获得的推论。获得的结果非常相关,并证明了我们的方法对社交媒体数据的计算分析的实用性。

关键词博客圈;信息抽取;情感分析;社交媒体分析;社交计算。

一、引言

几年前,博客圈和社交媒体是人们热议的话题。随着社交媒体技术的飞速发展和普通人的参与,博客圈如今已成为了解客户偏好和态度的丰富数据库。顶级商业公司正在分析这些博客和用户互动,以了解客户的选择、经济行为并确定市场趋势。最近披露的有争议的棱镜项目表明,政府正在利用网络内容进行电子监控或间谍活动。博客圈的形态和结构不断变化;尤其是在发生重大政治或社会运动时,博客数据会大幅增长。

博客是当今最强大、最值得信赖的文本媒体渠道之一,它为普通人提供了平台,允许他们不受歧视地独立表达自己的观点。博客平台（如 blogger、blogspot、worldpress）具有易于维护的内容管理系统,为用户提供了表达自己的自由。这种免费且易于发布的博客服务导致博客圈中创建和发布了大量内容。博客跟踪公司Technorati在 2004 年 9 月跟踪了大约 400 万个博客,到 2011 年 7 月,博客数量已增长到大约 1.64 亿个[1],在短短七年内增长了 41 倍。Wordpress [2]（排名第一的博客网站)最近的统计数据显示,其用户群超过 3.83 亿人,每月博客页面浏览量超过 35 亿次。仅 Wordpress 的用户就产生了 3390 万个博客页面。每月有新的博客文章和约 4090 万条新评论。人们在博客上撰写各种主题,从技术讨论到社会政治事件和机构。 2011年和2012年,社交媒体被用来

协调和管理许多国家针对政府的社会政治抗议活动。人们使用博客平台来控制和协调历史上从未见过的运动。正是这种动机促使我们承担了一项关于一系列社会政治事件的博客数据的计算分析任务。现在,博客被认为是最值得信赖和最原始的数据源,它充满了情感和人们观点的无拘无束的表达。

显而易见的是,超过 60% 的博客作者都是业余爱好者,他们写文章没有任何商业动机[3]。此外,博客圈内容创作速度快,使其成为最新的内容,也是人们对事件即时反应的真实存储库。

因此,博客圈分析现在是一个重要且有趣的跨学科研究领域。对于博客分析,语言处理、机器学习和模式识别技术可用于发现隐藏的信息模式、识别实体和其他重要方面。本文介绍了我们对社会政治事件博客数据的计算分析的实验工作。我们为进行此分析设计了一个综合框架。本文的其余部分组织如下。第二部分介绍背景和动机。第三部分介绍计算公式。第四部分解释数据集,第五部分介绍实验设置和结果。本文最后总结了第六部分中所示的观察结果。本文的主要贡献是提出了一种新颖的综合文本分析框架,用于基于内容的社会政治博客数据分析。该框架已使用足够多的博客数据进行了测试,从而证明了我们工作的适用性和实用性。

二.动机

大多数博客圈分析的早期工作都是为了商业利用。无论是广为人知的策略　通过在 2007 年初联系有影响力的博主并说服他们分享自己的经验来获得有关微软新发布的操作系统 Vista 的早期反馈,还是通过免费赠送笔记本电脑作为回报的默契营销策略多家公司;博客圈现已成为广泛认可的商业开发平台。然而,直到最近,博客圈仍然相对未被探索的一个方面是,它也是跨文化心理和社会学分析的丰富而独特的宝库。博客的增长速度前所未有,规模巨大

978-1-4673-6217-7/13/$31.00 c 2013 IEEE

博客圈中包含的大量数据不仅对于商业利用而且对于社会政治分析来说都是独特的财富。博客网站现在是博客作者对各种问题和事件进行跨文化和多样化社会政治描述的非常丰富的来源。

过去几年,来自不同领域的研究人员开始探索博客圈的非商业方面。这种分析工作有两种广泛的类型。一种更偏向计算机科学的类型包括寻找有影响力的博主 [4] 和关于某个事件的博客网站 [5]、社区发现、过滤垃圾博客等任务。

[6]、[7]。另一种风格更倾向于博客文章的社会政治分析 [8]、[9]、[10]。这包括围绕特定社会政治事件绘制博客圈的任务[11],分析与重要事件/个性/组织或过程相关的博客文章[12],[13],[14],[15]。我们的分析方法是一种面向社会政治推理的方法。我们重点探讨了博客文章中讨论的主要实体（个人、组织等）;确定有关主题的关键问题并了解博主如何看待该主题。选择博客文本是因为它是获取世界各地人们不受约束、第一手、未经编辑的表达、想法和观点的最佳来源。

三.计算公式

我们使用集成的计算公式评估了社会政治博客数据,该公式结合了主题建模、命名实体识别和情感分类器。

以下段落对每项任务进行了简要描述。

主题发现可识别文档集合中固有的主题,换句话说,它尝试用某个特定主题注释大量文档。大多数基本主题建模器都应用聚类算法进行主题检测。我们的主题建模器采用一组统计方法来分析集合中文本文档的单词,并使用单词使用模式的信息将所有具有相似模式的文档联系起来。它使用基于文本文档的分层贝叶斯分析的概率模型 [16]。现在,主题建模不仅用于发现主题,还用于找出这些主题如何相互联系以及它们如何随时间变化。基于潜在狄利克雷分配 (LDA) 的主题建模器 [17] 假设文档是多个主题的混合。更具体地说,它假设一些 k 个主题与文档集合相关联,并且每个文档以不同的比例展示这些主题。因此,集合中的所有文档共享同一组主题,但每个文档以不同的比例展示这些主题。贝叶斯非参数主题模型、动态主题模型和相关主题模型是主题模型的其他一些变体 [18]。

命名实体识别 (NER) 是一种流行的信息检索技术,可以自动注释文本文档的实体。这可以看作是一种

分类任务或众所周知的标记问题。通用命名实体识别器是监督分类器。有多种成熟的方法可用于命名实体识别,例如基于训练语料库的统计技术,该技术根据其相邻单词来识别单词

关键字,或严格按照定义的规则运行的基于规则的方法[19],[20]。一些标准的众所周知的 NERC 库可以非常准确地识别实体,例如人员、组织、日期时间、位置等。我们利用 Alchemy Web 服务 [21] 来识别每个帖子中的重要实体。

情感分类是一项识别文本情感分数的任务。在文本中,情感分类可以在不同层面进行,即文档级、方面级和实体级。情感分类有多种方法,从机器学习分类器到基于词典的方法。基于

SentiWordNet 的方法是基于 SentiWordNet 词典 [22] 的著名半监督方法之一。这种方法针对选定的特征出现,并根据其在词典中的分数关联情感极性。我们设计了一种基于 SentiWordNet 的算法公式,在 [23] 和 [24] 中有详细描述,用于计算博客数据中出现的重要实体的情感极性。对于这项任务,我们首先计算博客文章中相关实体的频率和极性,然后通过添加情感分数来汇总这些实体的极性。

为了以易于理解的格式描述获得的结果,我们使用了可视化工具。

信息可视化是一项将一些原始数据以信息丰富的视觉形式表示的任务。在这项工作中,我们使用了各种技术,例如使用 Gephi [25] 绘制标签云图和绘制图形来表示主题比例信息。

实体表示为标签,而实体的频率决定了其在绘制的标签云图中的大小。通过用两种不同的颜色描绘实体来表示围绕实体的情绪,以将它们标记为具有“积极”或“消极”取向的实体。我们使用了“红色”和“绿色”颜色的全色谱范围来显示博客数据中描述的各种实体周围的情绪极性的强度。

四.数据集

我们已将我们的算法公式应用于最初为[5]中报告的工作收集的足够大的数据集。该数据集包含总共 9290 篇关于“阿拉伯之春”这一更广泛主题的博客文章,其中包含关于 2011 年和 2012 年及其前后举行的各种革命的博客。

阿拉伯地区纪念民主斗争。大部分抗议活动都是通过社交媒体进行管理和协调的。该数据集总共包含 7343883 个单词,博客的平均单词数为 790.5148547。数据集包含 3 类博客文章,并具有以下统计数据。使用该数据集的主要动机是识别和衡量社交媒体数据对现实世界中发生的实际社会政治事件的代表性。

表一数据集

	博客文章数量	单词数数	平均字数数数
埃及革命	5799	4228360 729.1533	
利比亚革命	2271	2032937	895.1726
突尼斯革命	1220	1082586	887.3656

五、实验工作及结果

我们提取了整个数据集中的主题,以识别整个博客数据集合中隐藏的底层主题。为了执行此操作,我们使用了基于条件随机场方法分类器的斯坦福主题建模工具包。我们进行了基于 LDA 的主题建模,迭代了 1000 次。最初,我们使用不同数量的主题执行主题建模器来查找主题困惑度。困惑度高的主题数量更适合主题选择。表 II 显示了困惑度值。

表二。主题困惑

数量 话题	困惑
5 主题	4877.675154561823
10个主题	4582.594127526477
15 主题	4354.035194905415
20 主题	4175.618282787014

根据从主题困惑度值获得的输入,我们执行了 5 个主题的主题模型,并记录了 05 个主题中每个主题的前 50 个单词作为代表性术语概况。下表 III 给出了观察到的 05 个主题主题的关键字示例列表。

表三。代表关键词的热门主题

串行	主题关键词
主题 0。	世界、政治、抗议、民主、运动
主题 1.	卡扎菲, 塞里安, 利比亚, 战争, 政府
主题 2.	军队、开罗、抗议者、权力、公民。
主题 3.	时间、她、她、生活、革命、背后、战争。
主题 4.	死亡、死亡、战争、2011、2012、革命、人。

在追踪了 05 个主题中记录的前 50 个关键词后,我们为 05 个主题中的每一个分配了主题名称。下表 IV 显示了分配的主题标签。由于博客数据集是关于埃及、利比亚和突尼斯的社会政治运动,因此确定的主题与主要问题、角色扮演者和与被称为“阿拉伯之春”的三场革命相关的事件密切相关。主题包括抗议民主、军队对抗公民、战争反对政府;所有这些都代表了该问题的社会政治情景。

表四。组合数据集中的主题

连续剧	话题
主题 0。	为了民主而抗议。
主题一。	反对政府的战争。
主题 2。	军队对抗公民。
主题 3。	妇女与战争。
主题 4。	社会抗议中的死亡事件（2011 年和 2012 年）。

在图1中,我们显示了文档中的主题分布,即三个事件对应的博客数据中主题的相关程度。我们可以清楚地看到，“针对公民的军事”和“妇女与战争”主题在三组博客数据中很常见。对于“埃及革命”的数据,这些是最普遍的主题,这也是其他媒体广泛报道的主题。对于“利比亚革命”数据集，“军事反对公民”是一个普遍的主题,这是从其他媒体来源的有关利比亚实际发生事件的新闻报道中得知的正确描述。

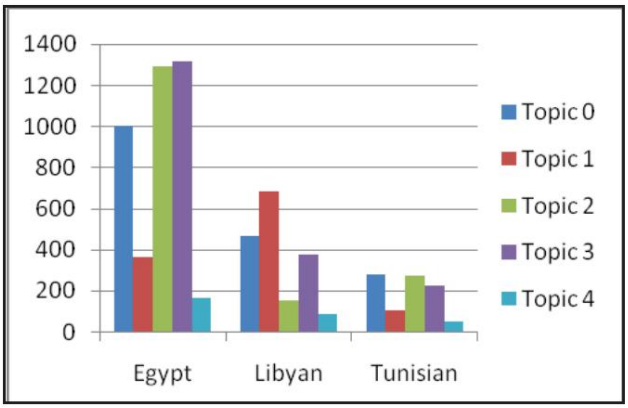

图 1文档主题分布

为了对主题建模练习获得的结果更有信心并验证结果,我们利用给定主题探索了 Google 趋势来验证我们的结果。谷歌趋势可视化人们随时间的搜索兴趣。下图 2 绘制了 Google Trend 记录的搜索趋势图。该图显示了这些事件相关期间的高搜索趋势。从 2010 年底开始,它的价值开始上升,并持续到最近。

图中显示，“反政府战争”似乎拥有最高的搜索趋势值,其次是“军队反公民”。在此期间，“抗议民主”和“妇女与战争”等话题也出现了多次激增。

因此,搜索趋势数据增强了我们对所获得的主题建模结果的信心。

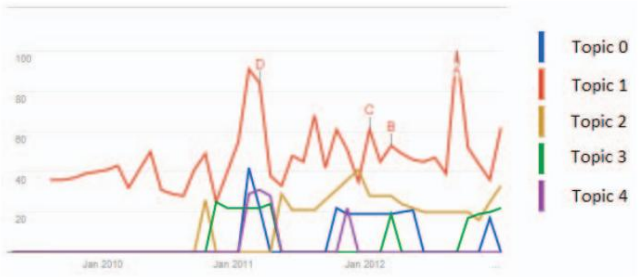

图2主题搜索趋势

我们执行的下一个分析任务是识别博客数据中引用的命名实体。这项任务的主要目标是更深入地了解与三个社会政治事件相关的角色扮演者、地点和组织,为此我们

有博客数据集。在完整博客数据集中观察到的关键实体被提取并识别为属于不同的类别，例如人物、地点和组织。我们还获得了关于博客数据集中特定实体的文本的情感极性。来自不同博客文章的特定命名实体的情感极性被聚合以具有围绕该实体的聚合情感极性值。标签云图中的一些单词比其他单词以更大的尺寸表示。标签云图中单词的大小是它们在博客数据集中出现的相对频率的度量。所识别的实体以红色或绿色的阴影显示，具体取决于实体周围的情感极性是“负面”还是“正面”。下图 3 绘制了数据集中观察到的人员类别的实体。

图3全数据集中人物类实体标签云

如上图 3 所示，人物类实体“穆阿迈尔·卡扎菲”、“胡斯尼·穆巴拉克”和“总统巴沙尔·阿萨德”表现得十分突出，且带有负面情绪。

这是一个很好理解的结果，因为他们是三场革命的主要人物，起到了负面的压迫作用。我们还看到“ban ki-moon”和“president obama”被标记为红色。这可能是因为博客作者认为他们可以帮助这些国家的被压迫人民，但在他们看来，他们的行为可能没有达到他们的预期。像互联网活动家和计算机工程师“wael ghonim”和“mohamed albaradei”这样的人以绿色显示，表达了博主们对他们的“积极”情绪。人员类实体的大小和颜色分配非常准确。

图 4 显示了博主在其帖子中提到的重要组织的标签云图。这些组织也用颜色编码来描绘博主表达的与其相关的情绪极性。

许多社交网络和媒体网站也被视为标签云图中博主提到的组织。

我们可以清楚地看到，大多数社交网站如“facebook”、“youtube”、“flickr”和“google”都显示出“积极”的情绪。这主要是因为社交网站是人们用来提高声音、抗议和表达对该地区压迫政权的愤怒的唯一平台。其他形式的媒体在很大程度上受到这些政府的审查。大多数其他实体被显示为接近“中性”，一些电视频道被显示为“负面”类别。

图4全数据集中组织类实体标签云

图 5 和图 6 显示了实体的地点类别，其中图 5 描绘了完整数据集的地点，图 6 描绘了“埃及革命”数据集中的地点。该地区的首府城市表现出负面情绪，这是衡量这些城市人民与压迫性政府的联系并表达了围绕这些城市的负面情绪的指标。

大多数其他城市的情绪极性接近“中性”类别。一些以“灰色”阴影显示的城市是人们在很大程度上大声疾呼并支持该地区人民权利的城市。在图 6 中，“开罗”和“亚历山大”以红色突出显示，这是真实的描述，因为这是受影响最严重且发生暴力活动的两个城市。我们还记录了数据集中出现的大陆引用，如表 V 所示。有趣的是，与其他地区相比，“欧洲”和“北美”被视为具有负极性，这是衡量这些地区受影响人群的期望的指标。

图5全数据集地点类实体标签云

图 6埃及革命数据集中地点类实体的标签云

表 V.各大洲出现频率和情绪得分

大陆	频率	情绪评分
亚洲	101	-.99845
南极洲	2	-0.02651
欧洲	660	-3.29104
拉美	73	.392666
北美	78	-2.68954
南美洲	四十六	.325686
撒哈拉以南非洲	2	.118754

六观察

我们为探索性分析特定主题的博客数据而设计的计算框架已经能够获得非常有趣且相关的结果。我们承担的分析任务是针对当代世界非常重要的一组社会政治事件,我们展示了一种算法设置,通过结合使用主题建模、NER 和情感分类的计算公式来完成这项分析任务。

通过主题建模实现,我们能够从整个博客数据集中识别主要主题关键词。这些以主题为导向的关键词描述了与博客数据相关的主要问题、角色扮演者、地点和实体,主题包括“抗议民主”、“反对政府的战争”、“军事反对公民”、“妇女与战争”和“死亡”在社会抗议中”。 NER 实现允许提取数据集中提到的人员、位置和组织,从而有助于进一步识别实体。我们能够确定在有关该主题的所有著作中经常被谈论或与该问题密切相关的主要人物、地点和组织。情感分析的结果进一步深入了解博客数据。它显示了博客作者在完整和个人博客数据集中关于不同实体（人、地点和组织)的观点。因此,我们可以观察并推断哪些实体与博主的“积极”情绪相关,哪些实体与博主的“消极”情绪相关。基于实体的情感分析获得数据集中讨论的所有主要实体的情感倾向。情感极性结果使用易于理解的颜色编码方案显示,描绘了情感极性的强度。

我们提出的计算分析方法与传统的主观分析相比有许多优势,尽管它并非旨在取代传统的主观分析。首先,我们的计算公式会自动收集世界各地人们撰写的相关文本,从而允许从跨文化和人口统计学的角度来看待问题。其次,我们可以快速分析的数据量不受限制。

以传统的手工方式分析这种规模的数据需要付出更多的精力和时间。第三,这种公式可以识别贯穿整个文本集合的主要主题,并衡量它们的相对强度。

此外,我们能够捕捉社会（以博主为代表)对各种问题和方面的整体情绪围绕社会政治事件。因此,这种计算公式为文本文档的自动分析提供了一个独特的框架,与传统的主观方法相比,它需要更少的精力和时间,并且本质上提供了跨文化社会学和社会政治视角,可以对任何重要的主题/感兴趣的问题进行分析。研究结果还可以为围绕主题进行详细的主观分析提供一个初始起点（或思考的素材）。

我们寻求通过将其转变为社交媒体文本的通用计算分析器来进一步扩展这项工作。我们正在努力建立一个完整的集成设置,该设置易于社会科学家使用,无需用户干预即可执行所有分析任务,并且允许社会科学家调整各种分析参数。我们的目标是设计一个系统,可以作为对分析任何社会政治事件或现象感兴趣的社会科学家的探索工具。社会科学家将决定她/他必须收集哪些社会政治事件或现象的博客数据,然后决定她/他对什么样的分析任务感兴趣。我们希望在系统中引入时间跟踪功能,其中可以在所需的时间段内映射事件或现象分析。毫无疑问,这样的工具对于在很短的时间内并且不需要太多努力来执行社交媒体数据的快速分析非常有用。然后可以通过详细分析对探索性结果进行和/或验证。

参考

[1] Technorati 和 Blogpulse 博客统计数据,2013 年 1 月 15 日取自 http://www.socialmediaexaminer.com/tag/blogging-statistics/。

[2] Wordpress 博客统计数据,2013 年 1 月 15 日取自 en.wordpress.com/stats/。

[3] 博客统计（信息图）,2013 年 1 月 17 日取自 http://blogging.org/blog/blogging-stats-2012-infographic/。2012

[4] N. Agarwal、H. Liu、L. Tang 和 PS Yu，“识别社区中有影响力的博主”,网络搜索和网络数据挖掘国际会议记录,第 207--218 页。 ACM 出版社,美国帕洛阿尔托。 2008 [5] D. Mahata 和 N. Agarwal，“每个人都知道什么?识别社交媒体中特定事件的来源”,第四届社交网络计算方面国际会议 (CASoN

2012) 的论文集。 2012 年 11 月 21 日至 23 日。巴西圣卡洛斯。

[6] H. Liu、PS Yu、N.Agarwal 和 T. Suel，“社会计算博客圈”,IEEE 互联网计算,2010 年 4 月,第 12-14 页。

[7] N. Agarwal 和 H. Liu，“博客圈:研究问题、工具和应用”,SIGKDD Explorations,卷。10,第 1 期,第 18-31 页,2008 年。

[8]五. K. 辛格,M.穆克吉, G。 K。 Mehta、N. Tiwari 和 S. Garg，“从网络日志中挖掘观点及其与社会政治研究的相关性”,载于 M. Natarajan、C. Nabendu 和 N. Dhinaharan (Eds.)《计算机科学与信息技术的发展》。计算机科学与工程,第二部分,2012 年 1 月,LNICST 85,Springer,第 134-145 页

[9] VK Singh、D. Mahata 和 R. Adhikari，“从社会政治角度挖掘博客圈”,《第六届计算机信息系统和工业管理国际会议论文集》,IEEE 出版社,2010 年 11 月,第 365-370 页。

[10] VK Singh，“挖掘博客圈中的社会学推论”,载于 S. Ranka 等编:《当代计算》,CCIS 第 94 卷,Springer-Verlag,海德堡,第 547-558 页,2010 年。

[11] Y. Mehrav、F. Mesquita、D. Barbosa、WG Yee 和 O. Fireder，“从博客圈中提取信息网络”,ACM Transactions on the Web,卷。 6,第 3 期,2012 年 9 月。

[12] H. Moe，“绘制挪威博客圈:互联网研究国际化的方法论挑战”,社会科学计算机评论 29(3) 313-326,2011。

[13] Y. Suhara、H. Toda 和 A. Sakurai，“使用主题词从博客圈进行事件挖掘”,ICWSM 会议记录，2007 年 [14] L. Adamic 和 N. Glanse，“政治博洛圈和 2004 年美国选举:分裂了
他们的博客”,第三届链接发现国际研讨会论文集,ACM,2005 年。

[15] J. Lin 和 A. Halavais，“绘制美国博客圈”,WWW 2004 年网络日志生态系统研讨会:聚合、分析和动态,2004 年 [16] D. Blei，“概率主题模型”,Communications of美国计算机协会，55(4),第77-84页,2012年。

[17] D. Blei、A. Ng 和 M. Jordan，“潜在狄利克雷分配”,《机器学习研究杂志》， 3,第 993–1022 页,2003 年 1 月 [18] D. Blei 和 J. Lafferty，“主题模型”,A. Srivastava
和 M.
Sahami（编辑)文本挖掘:分类、聚类和应用,Chapman & Hall/CRC 数据挖掘和知识发现系列,2009 年

[19] D. Nadeau 和 S. Sekine，“命名实体识别和分类的调查”,Lingvisticae Investigationes 30.1,第 3-26 页,2007 年。

[20] 斯坦福命名实体识别器,http://nlp.stanford.edu/software/ 检索 从 CRF-NER.shtml,2013 年 1 月 15 日。

[21] Alchemy API,1 月 15 日取自 http://www.alchemyapi.com/，2013年。

[22] SentiWordNet,2017 年 1 月 15 日取自 http://sentiwordnet.isti.cnr.it/，2013年。

[23] VK Singh、R. Piryani、A. Uddin 和 P. Waila，“电影评论和博客文章的情感分析:使用不同的语言特征和评分方案评估 SentiWordNet”,印度第三届IEEE 国际高级计算会议论文集,2013 年 2 月。

[24] VK Singh、R. Piryani、A. Uddin 和 P. Waila，“电影评论的情感分析:一种新的基于特征的方面级情感分类启发式”,国际自动化、通信会议记录,计算、控制和压缩感知,印度喀拉拉邦,2013 年 3 月

[25] Gephi:开放图可视化平台,取自 https://gephi.org/ 2013年1月15日。