

网络评论情感语料库的构建研究^{*}

崔大志 李媛

【提要】随着因特网服务的不断深入,情感信息的检索和利用日益受到重视。消费者在线产品评论蕴含大量的情感信息,在意见挖掘、个性化推荐等方面具有重要应用价值。针对手机、家电、化妆品等三个不同领域的产品在线评论进行语料收集和初步整理;在语料加工阶段,建立在线产品评论情感XML标注体系,并完成第一期语料的标注,实现在线评论情感语料库的初步构建,为文本情感计算、情感推理奠定基础。

【关键词】网络评论 语义网 情感语料库 情感标注

〔中图分类号〕 H0 **〔文献标识码〕** A **〔文章编号〕** 1000-2952 (2010) 04-0119-05

一、引言

从信息检索的角度看,当前各种搜索引擎技术主要是对客观性信息的检索,搜索与用户查询相关的网页。随着因特网服务的不断深入,近些年对于主观性信息的检索和利用日益受到重视,这项技术的关键是如何识别人们的主观意见,如何了解人们的情感表达,这是现有的搜索引擎的薄弱环节。

情感计算也因此成为人机交互领域中一个热门的研究领域,它的目标是使计算机拥有情感,这主要包含两个方面的含义,即计算机能识别人类的情感和表达自身的情感。^①互联网上最便利最丰富的文字资源日益成为人们情感分析的重要资源,这些情感信息在许多方面都有广阔的应用前景,如声誉分析、舆情监控、意见挖掘、产品评论、个性化推荐等。目前对于文本的情感计算研究处在起步阶段。究其原因是目前文本情感计算中缺乏有针对性的大规模情感语义资源的支撑,使之不能深入研究情感

的表达方式。而情感语义资源的基础是情感语料库。

已有的语料库的建设在收集语料、制定标注规范和质量监控等方面积累了宝贵的经验。语料库的发展从以语言研究为导向的第一代语料库(1970年代~1980年代),如Brown、Lob语料库等;到第二代以应用为导向的词典编纂(1980年代~1990年代),如Cobuild、Longman语料库等;再到第三代采用标准编码体系的超大规模语料库(1990年代以后),如ACL/DCI语料库等;目前进入到以互联网作为语料库的第四代研究阶段。^②汉语方面有国家现代汉

^{*} 本文受国家自然科学基金面上项目(60972090)和辽宁省教育厅人文社会科学研究项目(2010129)资助。

^① R. W. Picard, Affective Computing, Cambridge, MA: MIT Press, 1997.

^② 刘颖:《计算语言学》,清华大学出版社2003年版,第157~158页。

语语料库、^① 中港台汉语语料库、^② 《人民日报》语料库等。上述语料库的建设专注于语料的句法语义标注和分析。在情感语料标注方面, 目前已有的语料源有 Whissel 语料库,^③ Berardinelli 电影评论等;^④ 清华大学标注了对部分旅游景点描述的情感语料, 徐琳宏、林鸿飞等对一部分文学色彩较浓的作品、期刊等进行了情感类别标注与分析。^⑤ 如何构建适应互联网发展趋势和便于情感计算的语料库是目前需要解决的问题之一。

语义网是对未来网络的一个设想, 是一种能理解人类语言的智能网络, 是能够根据语义进行判断的网络, 它不但能够理解人类的语言, 而且还可以使人与电脑之间的交流变得像人与人之间交流一样轻松。语义网 (Semantic Web) 的研究目的是“开发一系列计算机可理解和处理的表达语义信息的语言和技术, 以支持网络环境下广泛有效的自动推理”,^⑥ 使人类从搜索相关网页的繁重劳动中解放出来。在这样的网络中, 信息都被赋予了明确的含义, 机器能够自动地处理和集成网上可用的信息。

语义网使用 XML 来定义定制的标签格式, 语义数据库是标注型语义网应用的一个发展方向。在情感研究领域, 实现情感语义的 XML 标注, 进行情感数据库的构建, 将为文本情感计算、情感推理打好基础。

在信息技术和网络通讯技术的发展推动下, 网络购物日益受到消费者的青睐, 数以亿计的网民消费者通过参与和自治, 形成以论坛、博客为代表的网络社区。网络社区以在线评论为主要形式, 情感互动性强、反馈及时, 极大地影响着网络消费者的心理情感和购买行为。因此, 构建消费者网络评论的情感语料库不仅具有重要现实意义, 而且在文本情感分析方面也具有典型性和示范性。同时, 由于产品评论所体现的情感较单一, 所用的情感词汇总体规模较小, 因此相应的语料库相对较易研究与分析。

本文针对消费者网络产品评论, 介绍了情感模糊语料库的构建与分析, 对手机、家电、化妆品等三个不同领域的产品系列进行语料收

集和初步整理。在语料加工阶段, 建立了网络产品评论情感 XML 标注体系, 并完成了语料的初步标注工作。最后介绍了该语料库的应用领域, 并总结了经验与不足。

二、在线语料收集

构建语料库, 首先便是语料本身, 即语料收集。在提供优质训练集和提供统计规律以及常识性知识方面, 语料库的优势是巨大的, 同时也可以进一步提高情感词汇本体的质量和规模。语料的收集工作是指选择合适的语料、做预处理、为语料的标注提前做好准备。

语料选择要考虑到语料库的覆盖率, 从来源上包括各大论坛、贴吧、博客以及电子商务网站; 从对象上看, 有手机、家电、化妆品等典型的产品评论语句; 在风格方面, 尽可能选择用词比较规范、严谨的评论, 但消费者在线评论的口语特征比较明显, 口语化情感将作为今后研究的一个方面。总的来说, 语料的选择偏重情感表达丰富多彩的评论。我们选择诺基亚手机 N97、惠普笔记本 4411s 和资生堂化妆品 Za 系列作为研究对象, 详见表 1。

表 1 数据形式 d_1/d_2 中, d_2 为从各个资料来源采集到的原始数据信息, 经无关性筛选以及上述的选择条件, 得到 d_1 条最终评论语料。采集工具选用的是神采软件工作室出品的《网络神采》共享版, 虽然采集条件有一定限制 (单次最多 500 条), 但经过循环采集, 仍

① 刘连元:《现代汉语语料库研制》,《语言文字应用》1996 年第 3 期。

② 胡百华、李行得、汤志祥:《香港的语料库和相关研究概况》,《语言文字应用》1997 年第 2 期。

③ http://www.icl.pku.edu.cn/icl_groups/corpus tagging.asp [DB/OL]。

④ T. Athanaselis, S. Bakamidis and I. Dologlou, Recognizing Verbal Content of Emotionally Colored Speech, European Signal Processing Conference, 2006.

⑤ 徐琳宏、林鸿飞、赵晶:《情感语料库的构建和分析》,《中文信息学报》2008 年第 22 期 (1)。

⑥ 张晓林:《Semantic Web 与基于语义的网络信息检索》,《情报学报》2002 年第 21 期 (4)。

然可以达到上述效果。采集结果自动发布到数据库，在此过程中去掉重复行，即相同评论，最终得到 9229 条有效评论，占总评论数的 48.7%。

表 1 评论语料详情

对象 来源	Nokia N97	HP 4411s	资生堂 Za	总计
中关村 在线	1038/ 2279	946/ 2395	\	1984/ 4674
百度 贴吧	1109/ 3426	1208/ 3375	1827/ 3481	4144/ 10282
新浪 博客	241/ 500	273/ 491	316/ 500	830/ 1491
美容 社区*	\	\	2271/ 2500	2271/ 2500
总评 论数	2388/ 6205	2427/ 6261	4414/ 6481	9229/ 18947

注*：闺蜜网、YOKA 时尚网。

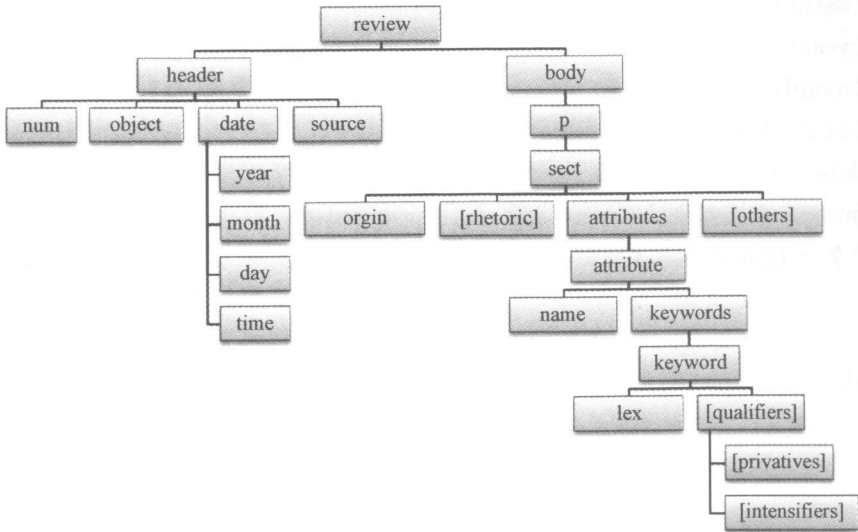
三、网络评论语料的 XML 标注

语料库是语言材料的集合，具有三个特点：必须是真实语言环境中出现过的语言材料；必须以电子计算机为载体；必须经过一定的分析、加工和处理。语料标注是语料库的第

二个方面。标注（Annotation）是一种将隐含的（语言学）知识形式表示出来的过程；是将非结构化的文本转换为半结构化文本的重要环节；是将文本变为知识的过程。针对每条/篇评论语料，我们沿用标注体系从词汇、语句和篇章三个加工层次进行。当今语料库语言学的代表人物之一 Leech 认为语料的标注应该遵循七个原则：（1）标注附码可以删除，可以恢复到原始语料；（2）所作的标注可以单独抽出，另外储存；（3）语料的最终使用者应该清楚标注的原则和附码的意义；（4）在语料的使用和说明文件中，应该说明标注者以及标注所使用的方法；（5）应向用户表明，语料的标注并非完美无缺，它只是一种可能有用的工具；（6）标注应该尽量采用被人们普遍接受的中立的模式；（7）任何标注模式都不能作为第一标准。

编码体系即标注附码的选择，采用 TEI（Text Encoding Initiative，机读语篇的国际信息编码规范）标准与自定义编码体系（模糊情感特殊需求）相结合的方法。在 TEI 标准中，每个语言单位由起始标记和结束标记构成。结合上述标注原则，我们制定了在线评论的情感标注体系，如图 1 所示，方括号内的为可选部分。

图 1 在线产品评论情感标注体系



一篇评论由 header 和 body 两部分组成。header 表示评论的基本信息,包括编号(num)、评论对象(object)、评论日期(date)以及评论来源(source)。body 部分是情感标注体系的主体,p 标记段落,sect 标记语句。语句标注由原始语句(orgin)、修辞(rhetoric)、产品属性(attributes)和涉及的其他品牌(others,为品牌竞争分析提供依据)组成。其中产品属性又由属性名(name)和与其对应的关键词(keywords)来描述。我们规定一个特殊的产品属性,即 name 的一个特定值“overall”,用来表示在没有指定产品属性名的时候的总体评价和情感。情感词汇由词汇本身(lex)和限定词修饰前缀(qualifiers)构成,限定前缀包括否定前缀(privatives)和程度副词前缀(intensifiers),它们将改变情感词汇模糊本体的情感类别和情感强度。

我们以 Nokia N97 的一篇评论为例,截取部分示例标注如下:

```
<? xml version = "1.0" encoding =
"gb2312"?>
  - < review s>
  - < object> Nokia N97</object>
  - < review>
    - < header>
      < num> 1</num>
      - < date>
        < year> 2010</year>
        < month> 04</month>
        < time> 18: 56: 53</time>
      </date>
      < source> 【诺基亚 N97i 精英版 入
手 7 天 使用感受】 - 诺基亚 N97 论坛 - ZOL 中
关村在线
    </source>
    </header>
    - < body>
      - < p>
      - < sect>
        < orgin> 转眼间小 N 已经到手四
```

五天了,对她爱不释手。

```
</orgin>
  - < attributes>
    < name> overall</name>
  - < keywords>
    - < keyword>
      < lex> 爱不释手</lex>
    </keyword>
  </keywords>
</attribute>
</attributes>
</sect>
- < sect>
```

< orgin> 优点: 屏幕大, 触摸灵敏, 菜单滑动设计很人性化, 外表大气, 用蓝牙听歌音质非常好, 电池待机时间长, 无线局域网流畅浏览非常爽。

```
</orgin>
```

利用这一标注形式, 我们已完成对所有 9229 条有效评论的标注, 构建起了一个针对相关产品的小规模的网络评论情感语料库。

四、结语

本研究针对消费者的网络产品评论, 对三个不同领域的产品系列进行语料收集, 在语料加工阶段, 设计了在线评论的情感标注体系, 主要标注出了情感对象(产品属性)、修辞手法、情感词汇和相应的限定词前缀等信息。

语料库的标注内容和标注形式决定了它的应用范围。以语料库构建为基础, 可以训练文本情感识别模型, 从事情感词汇本体的自动学习和统计情感迁移规律等研究。另外, 情感语料库的构建可以为各种产品推荐系统、评价信息挖掘系统、智能决策系统、检索工具等提供第一手数据资料和实践基础。

语料库的建设本身就是一项长期而复杂的任务, 从收集语料, 制定标注规范, 到完成语料加工, 每一步都要既确保速度, 又确保标注

质量。通过语料库的初步构建，我们发现，部分网上评论口语化色彩比较强烈、书写随意性较大、句式结构相对混乱，因此在进行语料标注时，对段落和句式的划分以及属性情感的标注有时无法达到统一。因此针对这种评论文本，还需要我们从语言学角度分析语言规律，在标

注方法和标注体系上做进一步的改进和探讨。

本文作者：崔大志是英语语言文学硕士、大连理工大学外国语学院讲师；李媛是大连理工大学管理学院硕士研究生
责任编辑：王姣娜

Research on Affective Corpus Construction of Online Reviews

Cui Dazhi Li Yuan

Abstract: Affective information retrieval and utilization are attached growing importance with the development of online service. Loaded with a large amount of affective information, consumer online product reviews contain great application value in opinion mining and personalized recommendation. In this paper, online product reviews in three different areas such as mobile phones, appliances and cosmetics are collected and initially filtered for corpus construction. In the corpus processing section, an affective tagging system XML for online product reviews is established, and the tagging work of the first phase is basically completed. The affective corpus of online review is primarily constructed, which lays the foundation for text emotion calculating and reasoning.

Key words: online review; semantic network; affective corpus; affective tagging

观点选萃

军事“软实力”视野下我军形象力建设探析

田 祥

南京陆军指挥学院军队政工系硕士研究生田祥认为：军事“软实力”视野下我军形象力建设必须坚持“听党指挥”的形象力建设内核，筑牢“服务人民”的形象力建设宗旨，遵循“保障打赢”的形象力建设目标，确立“形象制胜”的形象力建设战略。要遵循形象力生成规律，拓宽我军形象力建设的路径，形成形象塑造与形象传播“一体互动”的形象力生成格局。

军事“软实力”是指将物质和人力资源转化为有效战斗力的能力，是用非强制性手段影响和塑造别国的能力，是间接形成威慑与打击能力的实力。军队形象力是军事“软实力”的重要组成部分，是指以军队形象为中介，以形象传播为手段，使军队的良好形象被军内外公众所感知，进而影响公众对军队的评判和态度的能力。在要素构成上，军队形象力包括凝聚力、向心力、吸引力、辐射力、威慑力、瓦解力等要素。在战略应用上，军队形象力表现为用非强制性手段影响和征服对方，从而实现其军事效益。

(马光 摘编)