

用于细颗粒度挖掘的产品评论语料库构建技术

刘远超, 宋明凯, 刘 铭, 张 想

(哈尔滨工业大学 计算机科学与技术学院, 150001 哈尔滨, lyc@insun.hit.edu.cn)

摘 要: 为了辅助进行产品评论中特征-观点对识别的挖掘工作, 对细颗粒度产品评论语料库的构建技术进行了研究. 介绍了用于产品评论细颗粒度挖掘的语料库构建方法, 以及目前初步进行的语料标注工作. 标注数据可以数据库形式存储, 从而实现了无结构化到结构化的转变, 为自动查询等处理提供了极大方便. 实验结果表明: 虽然文中的标注方法以手机产品为例, 但具有良好的移植性, 可以应用到其他产品评论的细颗粒度语料库构建. 相应的语料库构建对于高性能机器学习方法的应用、特征-观点对识别算法的性能提高以及自动评价等具有重要意义.

关键词: 产品意见挖掘; 细颗粒度语料库构建; 语料标注

中图分类号: TP391 **文献标志码:** A **文章编号:** 0367-6234(2012)03-0064-05

Research of product review corpus constructing technology for fine-granularity mining

LIU Yuan-chao, SONG Ming-kai, LIU Ming, ZHANG Xiang

(School of Computer Science And Technology, Harbin Institute of Technology, 150001 Harbin, China, lyc@insun.hit.edu.cn)

Abstract: Quantitative analysis and mining of product reviews posted by users are helpful for both manufacturers and consumers. During the work of fine-granularity product review mining, extracting feature-opinion pair is one of the core works. The corresponding corpus construction is of great significance for the application of high performance machine learning methods, improving the performance of feature-opinion extraction algorithm and automatic evaluation. This article introduces corpus constructing technology for fine-granularity product review mining and the initial corpus labeling work, thus realizing non-structured to structural changes. The corpus can be stored in database and thus provide great convenience for automatic query processing. Although current labeling work was performed in mobile phone products, it can be applied also to other product types for fine granularity corpus construction. So our work has good transplantation ability.

Key words: product review mining; fine-granularity corpus construction; corpus annotation

产品评论挖掘是文本挖掘研究领域中最近几年兴起的研究热点之一^[1-4]. 其主要做法是以 Web 上发表的用户产品评论作为挖掘对象, 从大量文本数据中发现用于对该产品各方面性能的评价^[5-7]. 由于产品评论挖掘是针对某类产品进行

挖掘, 因此有必要建立语料库. 国外比较知名的大规模语料库包括 Brown 语料库、Lob 语料库^[8]、Cobuild 语料库等. 汉语现有的大规模语料有《人民日报》语料库、现代汉语语料库、北京大学双语语料库等等. 目前国内外已有的用于文本挖掘的语料库包括 Pang 语料库、Berardinelli 电影评论语料库^[9]、MPQA 语料库^[10]等.

本文要研究的手机产品评论语料库的来源是用户在不同手机网站上的产品使用评论(称之为原始评论). 对语料库的标注意味着让它能够机读化, 能够方便的存入数据库中, 并且能够快速的

收稿日期: 2010-03-07.

基金项目: 教育部人文社会科学研究青年基金资助项目(10YJCZH099); 中央高校基本科研业务费专项资金资助项目(HIT.NSRIF.2009065); 语言语音教育部-微软重点实验室开放基金资助项目(HIT.KLOF.2009022).

作者简介: 刘远超(1971—), 男, 博士, 副教授.

得到手机产品特征 - 观点对 (Feature-Opinion Pair),方便以后进行基于统计的方法研究.

1 语料标注方法

相关术语:

复句:定义由 4 个标点符号结束的句子成为复句,包括“.”,“;”,“!”和“ ”. 一个复句由一个或多个单句组成.

单句:定义在一个复句里,由两个标点符号结束的句子成为单句,包括“,”,“?”.

特征 - 观点对 (Feature-Opinion Pair):在用户的一条评论中出现的相对应的产品特征和观点对,一条评论可能包含多个特征 - 观点对,表示为 (Feature, Opinion).

语料库的一个重要标志是机器可读,语料标注是实现原始语料机读化的关键环节,对语料库信息进行结构化的标注对辅助意见挖掘是非常重要的一步,目前语料库大都采用结构化标注,大多依据某种特点的标注方法和标准. 语料标注通常分为自动标注和手工标注,前者通常采用某种标注模式来对语料库进行标注,如卫乃兴等^[11]使用 XML 语言对 Colsec 语料库进行标注,吴琴霞等^[12]使用 XML/Schema 对甲骨文语料进行标注;后者则是根据一定的标注规则,采用特定的方法,人为手工的对语料库进行标注. 本文研究的意见挖掘语料库就是采用手工标注的方法,通过编写一个 MFC 程序,把规则嵌入到程序中,然后手工完成语料库的标注.

1.1 预处理

在意见挖掘语料库中,研究的目的在于快速的定位 Feature-Opinion Pair (简称 Pair),由于这样的 Pair 通常会整体出现在一个复句中,而用户的一条评论可能包含多个复句,举例如下:

“我刚买,屏幕很大. 朋友都说太薄不好看,我觉得挺喜欢的,就是电池很不耐用这是真的,想去买块电池备用.”

上述评论中有两个复句:1) Pair 为 (屏幕,很大);2) Pair 为 (电池,不耐用). 另外一个复句可能有多个单句组成,在一个复句中出现的 Pair,也有可能来自于不同的单句. 例如:“说到它的音质,那是非常不错的!”. 这条评论的 Pair 为 (音质,非常不错),但它们分别出现在不同的单句中.

针对上述的几种情况,在语料库标注前先要对原始评论进行预处理,首先把原始评论分割成复句,划分的标准是“.”、“!”等标点符号;然后对复句再分割成单句,目前统一划分为 10 个单句,

划分的标准是“,”、“?”等标点符号,对不足 10 个单句的以 Null 表示,如图 1 所示.

“复句:朋友都说太薄不好看,我觉得挺喜欢的,就是电池很不耐用这是真的,想去买块电池备用.

单句 1:朋友都说太薄不好看,

单句 2:我觉得挺喜欢的,

单句 3:就是电池很不耐用这是真的,

单句 4:想去买块电池备用

单句 5:Null

单句 6:Null

单句 7:Null

单句 8:Null

单句 9:Null

单句 10:Null”

图 1 复句划分成单句示例

1.2 语料标注流程

在预处理完成之后,再添加上原始评论中存在的一些信息,构成了要进行语料标注的一个单元,称之为标注单元,如图 2 所示.

“(9022)

单句 1:朋友都说太薄不好看,

单句 2:我觉得挺喜欢的,

单句 3:就是电池很不耐用这是真的,

单句 4:想去买块电池备用

单句 5:Null

单句 6:Null

单句 7:Null

单句 8:Null

单句 9:Null

单句 10:Null

复句:朋友都说太薄不好看,我觉得挺喜欢的,就是电池很不耐用这是真的,想去买块电池备用

编号:144

型号:诺基亚 5310 (XpressMusic)”

图 2 标注单元内容示例

一个标注单元中所包含的信息如表 1 所示,所有的标注单元构成了评论文件. 针对上述的标注单元,设计的标注流程如图 3 所示.

表 1 标注单元中包含信息说明

包含信息	说明	举例
复句编号	被切分的复句的编号	(9002)
单句 1 ~ 10	切分一分复句为 10 个单句	
复句	被切分的复句	
编号	复句所在的原始评论的编号	144
型号	手机产品的型号	诺基亚 5310 (XpressMusic)

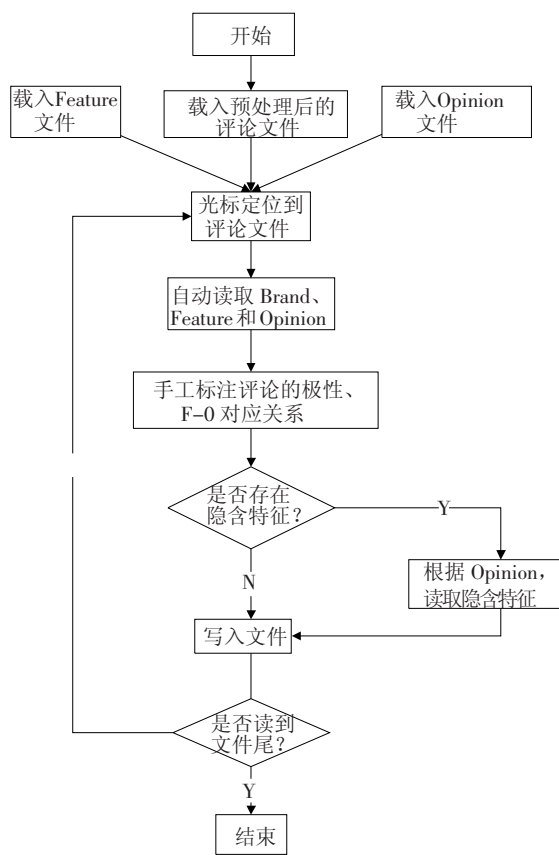


图 3 语料标注流程

1.3 语料标注具体操作

如图 3 所示,首先系统在载入预处理后的评论文件时,同时还要载入 Feature 库文件和 Opinion 库文件,Feature 库文件收录了可能在手机产品评论中出现的 349 个产品特征,Opinion 库文件收录了可能在手机产品评论中出现的 4 905 个情感词或短语.

载入文件后,把鼠标光标定位到标注单元的任意位置,可以通过点击“产品型号、Feature 和 Opinion”来依次得到标注单元中存在的相应信息;“极性”系统初始默认为“-”(实践结果中得出评论为负的比例较大),并可通过后面的按钮进行极性修改;上述只要用户点击,系统可自动完成,之后的“对应关系”,需要用户从 Feature 列表框和 Opinion 列表框中选择相对应的 Feature 和 Opinion,然后点击“对应关系”按钮,它们所在的单据编号就会自动的添加到对应关系列表框中,这一步(标注 Feature-Opinion Pair)是标注过程的核心.根据评论中的大多数情况,本文把对应关系的上限设为 3,即在一个标注单元中,最多能标注 3 对 Pair,不足的依然以“Null”补充.语料标注系统的主界面如图 4 所示.

在实践中会遇到的一种情况是一些评论句中并不显式的存在相对应的特征词,比如在上述举例的复句中存在如下单句“朋友都说太薄了”,可以发现“太薄”是观点词,但是在该复句中并不能找到相对应的特征词,于是本文提出了隐含特征的定义,即在一条评论中存在对其的情感评论词,但并没显式出现的特征词.

针对上述情况,专门为“隐含特征”添加了一个标注功能,用户选中 Opinion 列表框中的一个 Opinion,然后点击“隐含特征”,就会打开一个新的对话框,里面包含了可供选择的产品特征,用户在选择即可.“隐含特征”运行示意图如图 5 所示.

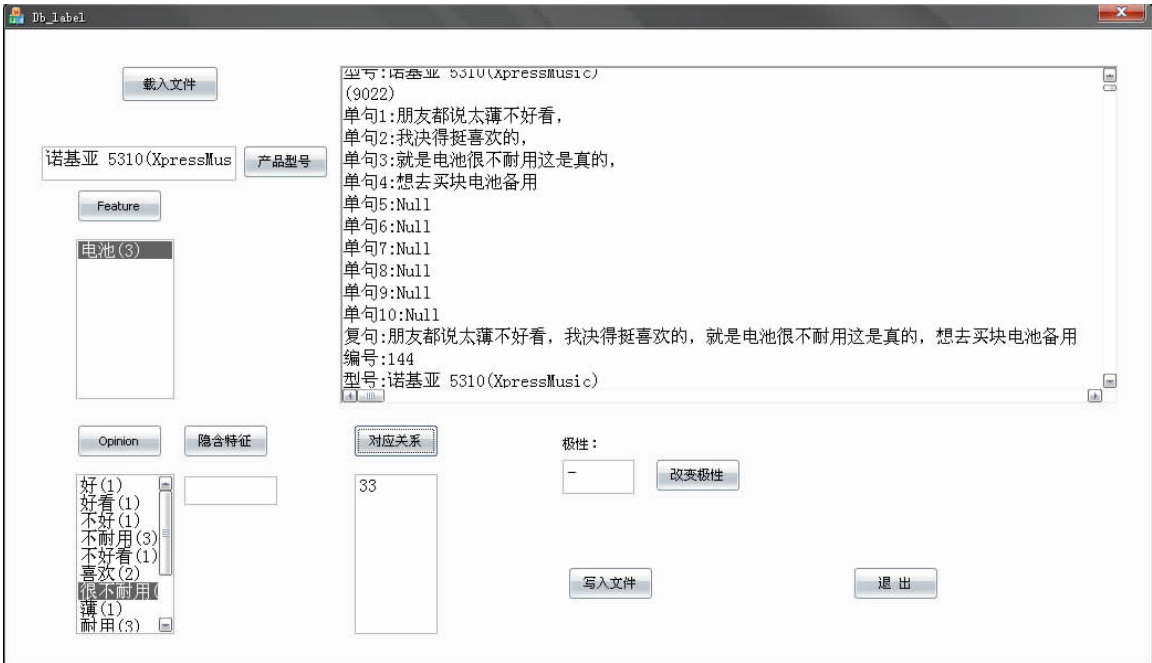


图 4 语料标注系统



图5 “隐含特征”运行示意图

进行完上述步骤之后,点击“写入文件”按钮,把标注好的一个标注单元暂时写入到一个文本文件中,全部完成之后,再方便的导入到数据库中.数据库的格式如图6所示.其中:ID为复句编号;Brand为产品型号;FO1-FO3为3个对应关系;Feature1-3,Fnum1-3分别为特征1-3和特征

1-3所在的单句编号;Opinion1-3,Onum1-3分别为观点1-3和观点1-3所在的单句编号;HFeature、HOpinion和Hnum分别为隐含特征、隐含特征的观点及隐含特征观点所在的单句编号;Sim1-10为10个单句;Com为复句;Num为复句所在的原始评论的编号;Polarity为该评论句的情感极性.

ID	Brand	FO1	FO2	FO3	Feature1	Fnum1	Feature2	Fnum2	Feature3	Fnum3	Opinion1	Onum1	Opinion2	Onum2	Opinion3	Onum3
(100)	诺...	12	11	null	键盘	1	键盘	1	null	null	不好...	2	硬	1	null	null
(101)	索...	null	null	null	质量	3	照相	2	拍摄	6	差	3	逊色	2	模糊	6
(102)	诺...	11	11	null	格式	1	多媒...	1	null	null	非常...	1	非常多	1	null	null
(103)	诺...	11	null	null	功能	1	null	null	null	null	很强	1	null	null	null	null
(104)	诺...	11	null	null	屏幕	1	null	null	null	null	暗	1	null	null	null	null
(105)	索...	22	33	66	电池	2	手机	3	null	null	很快	2	不成熟	3	null	null
(106)	诺...	22	null	null	功能	2	null	null	null	null	非常...	2	null	null	null	null
(107)	诺...	75	null	null	待机	7	null	null	null	null	好	5	null	null	null	null
(108)	诺...	11	null	null	感觉	1	null	null	null	null	还是...	1	null	null	null	null
(109)	诺...	11	null	null	设计	1	null	null	null	null	不够...	1	null	null	null	null
(110)	诺...	11	11	null	键盘	1	格式	1	null	null	太小	1	太少	1	null	null
(111)	诺...	11	null	null	格式	1	null	null	null	null	少	1	null	null	null	null
(112)	诺...	11	11	null	电池	1	时间	1	null	null	短	1	短	1	null	null
(113)	诺...	11	null	null	外形	1	null	null	null	null	经典	1	null	null	null	null
(114)	诺...	13	null	null	软件	1	null	null	null	null	好	3	null	null	null	null
(115)	诺...	12	42	null	软件	1	情景...	4	null	null	少	2	少	2	null	null

HFeature	HOpinion	Hnum	Sim1	Sim2	Sim3	Sim4	Sim5	Sim6	Sim7	Sim8	Sim9	Sim10	Com	Num	Polarity
null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null
屏幕	黑屏	1	...	...	单...	...	...	单...	...	...	单...	单...	...	27...	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	16...	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	32...	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	510	+
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	510	+
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	563	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	25...	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	12...	+
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	209	+
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	23...	+
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	12...	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	563	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	562	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	27...	-
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	41...	+
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	633	+
null	null	null	...	...	单...	...	...	单...	...	...	单...	单...	...	27...	-

图6 数据库的存储格式图

2 语料标注系统数据分析

截止目前,利用本文设计的语料构建方法和工具,初步标注完成 1 870 个标注单元,合计约 10 326 037 字.对总体标注的 Feature-Opinion Pair 进行统计,由于之前设定最多为 3 对 Pair,所以考虑以下 3 种情况:只有 1 对 Pair (FO1! = null, FO2 = null, FO3 = null),有 2 对 Pair (FO1! = null, FO2! = null, FO3 = null)和 3 对 Pair (FO1! = null, FO2! = null, FO3! = null).统计结果如表 2 所示.

表 2 Feature-Opinion Pair 总体统计表

Feature-Opinion Pair	FO1 - FO3 取值	统计结果
0 对	FO1 = null, FO2 = null, FO3 = null	78
1 对	FO1! = null, FO2 = null, FO3 = null	1 280
2 对	FO1! = null, FO2! = null, FO3 = null	376
3 对	FO1! = null, FO2! = null, FO3! = null	136

再对几个手机品牌进行统计,如表 3 所示.

表 3 部分品牌 Feature-Opinion Pair 统计表

品 牌	诺基亚 N70	MOTO ROKR E2	索尼爱立信 W810C
总数	236	227	198
Pair = 0 对	14	7	4
Pair = 1 对	162	161	130
Pair = 2 对	45	53	40
Pair = 3 对	15	6	24

通过上述数据可以看出,在确定以复句作为 F-O Pair 的抽取单元后,通常一个复句中包含一个 Pair,其次是 2 个 Pair,0 或者 3 个的占少数.

3 结 论

1)从实现过程上看,文中提出的方法对语料库中的原始评论进行加工,把表示产品特性的特征词和表示用户情感的观点词标注上相对应的关系,以便于计算机的识读和进行下一步统计比较工作.

2)利用 MFC 程序对原始评论进行结构化标注的方法,使数据存储在统一的数据库中,为数据查询、分类和共享等提供了方便.更重要的是,经过结构化处理,可以方便地引入机器学习和数据挖掘中的先进理论和技术,从而加速相应的挖掘和分析工作.

参考文献:

[1] LIU J, WU G, YAO J. Opinion searching in multi-product reviews[C]//Proceedings of the Sixth IEEE In-

ternational Conference on Computer And Information Technology. Washington, DC: IEEE Computer Society, 2006: 25 - 31.

[2] DAVE K, LAWRENCE S, PENNOCK D. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews[C]//Proceedings of the 12th International Conference on World Wide Web. New York, NY: ACM, 2003: 519 - 528.

[3] ACIAR S, ZHANG D, SIMOFF S, et al. Informed recommender: Basing recommendations on consumer product reviews[J]. Intelligent System, 2007, 22(3): 39 - 47.

[4] HU Mingqing, LIU Bing. Mining opinion features in customer reviews [C]//Proceedings of the 19th National Conference on Artificial Intelligence. San Jose: AAAI press, 2004: 755 - 760.

[5] LIU Bing, HU Mingqing, CHENG Junsheng. Opinion observer: Analyzing and comparing opinions on the Web [C]//Proceedings of the 14th International Conference on World Wide Web. New York, NY: ACM, 2005: 342 - 351.

[6] HU N, PAVLOU P A, ZHANG J. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication[C]//Proceedings of the 7th ACM Conference on Electronic Commerce. New York, NY: ACM, 2006: 324 - 330.

[7] GHOSE A, IPEIROTIS P G. Designing novel review ranking systems: Predicting the usefulness and impact of reviews[C]//Proceedings of the Ninth International Conference on Electronic Commerce. New York, NY: ACM, 2007: 303 - 310.

[8] JOHANSSON S, ATWELL E, GARSIDE R, et al. The tagged LOB corpus user's manual[J]. Norwegian Computing Centre for the Humanities, 1986:10 - 12.

[9] PANG Bo, LEE Lillian. Thumbs up? sentiment classification using machine learning techniques [C]//Proceedings of the ACL - 02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2002: 79 - 86.

[10] WIEBE J, WILSON T, CARDIE C. Annotating expressions of opinions and emotions in language[J]. Language Resources and Evaluation, 2005, 39(2/3):165 - 210.

[11] 卫乃兴,李文中,濮建忠. COLSEC 语料库的设计原则与标注方法[J]. 当代语言学, 2007, 9(3):235 - 246.

[12] 吴琴霞,刘永革. 基于 XML/Schema 甲骨文语料库语料标注的研究[J]. 科学技术与工程, 2009, 9(17):1671 - 1673.

(编辑 张 红)