

中文微博情感分析研究综述

周胜臣 瞿文婷 石英子 施询之 孙韵辰
(上海大学悉尼工商学院 上海 201800)

摘 要 随着微博的风靡,与之相关的研究得到学术界和工商界的广泛关注。针对中文微博情感分析的研究进行综述。将中文微博文本情感分析分为三类任务:文本预处理、情感信息抽取和情感分类,对各自的研究方法和进展进行总结。其中情感信息抽取分为情感词、主题和关系的抽取,将微博主观文本情感分类方法归结为基于语义词典的情感计算和基于机器学习的情感分类。此外,从微博网站数据构成的角度出发,对情感分析做了延伸分析。最后总结微博情感分析的研究现状,并提出今后的研究方向。

关键词 中文微博 情感分析 情感信息抽取 情感分类

中图分类号 TP391 文献标识码 A DOI:10.3969/j.issn.1000-386x.2013.03.043

OVERVIEW ON SENTIMENT ANALYSIS OF CHINESE MICROBLOGGING

Zhou Shengchen Qu Wenting Shi Yingzi Shi Xunzhi Sun Yunchen
(Sydney Institute of Language and Commerce, Shanghai University, Shanghai 201800, China)

Abstract With the sweeping of microblogging, the associated research achieves widespread concerns from academia to business sector. In the paper, we summarise the studies in light of Chinese microblogging sentiment analysis. We divide the Chinese microblogging text sentiment analysis into three categories of tasks: the text pre-processing, the emotional information extraction and the sentiment classification, and present the conclusion in regard to every research method and its progress. Among them, the emotional information extraction is divided into emotional words, themes and relationships extraction, the sentiment classification method of microblogging subjective text is attributed to the semantic dictionary-based affective computing and machine learning-based sentiment classification. In addition, from the perspective of microblogging site data formation, we extend the analysis on sentiment analysis. In the end, the status quo of microblogging sentiment analysis is summarised, and the future research directions are proposed as well.

Keywords Chinese microblogging Sentiment analysis Emotional information extraction Sentiment classification

0 引 言

微博,即微博客的简称,是一个基于用户关系的信息分享、传播以及获取平台,用户可以通过 Web、WAP 以及各种客户端组件,以 140 字左右的文字更新信息,并实现即时分享。微博给予网络用户更自由、更快捷的方式来沟通信息、表达观点、记录心情,已经成为国内最为热门的互联网应用之一。以国内的新浪微博为例,目前其注册用户已突破 3 亿,用户每日发博量超过 1 亿条。

情感分析是指分析说话者在传达信息时所隐含的情绪状态,对说话者的态度、意见进行判断或者评估。情感分析在海量数据上的应用,将有助于完善互联网的舆情监控系统;丰富和拓展企业的营销能力;通过波动分析,实现对物理世界异常或突发事件的检测;此外,还可以应用于心理学、社会学、金融预测等领域的研究。故对于微博情感分析的研究有着很重要的现实意义。

目前关于中文微博情感分析方面的研究工作尚处于起步阶

段。国外对于微博的情感分析进行了一些探索,但是具体应用到中文领域,存在一定局限性,如 140 字的中文所蕴含的信息要比英文更为丰富;中文微博引入了表情、图片、视频等多媒体表现形式,而英文微博只是文字的表达,在产品设计上存在不同;中文与英文的语法规则和语言习惯存在很大的不同,更强调上下文语境,而英文意义的表达更为直接。所以本文主要就中文微博的情感分析做相关介绍和分析。

本文综合文本情感分析领域的研究成果以及已有的中文微博情感分析方法,将中文微博的情感分析归纳为三个部分:文本预处理、情感信息提取和情感分类,一般将情感信息提取视为情感分类的基础。文本预处理包括对文本进行分词、词性标注、停用词成立等,情感信息提取是根据一定的规则抽取微博中带有倾向性特征的单元要素;情感分类则利用底层情感信息抽取的结果将情感文本单元分为若干类别,对主观性文本极性以及强度进行分类。

收稿日期:2012-05-23。周胜臣,本科生,主研领域:信息管理与信息系统。瞿文婷,本科生。石英子,本科生。施询之,本科生。孙韵辰,本科生。

1 文本预处理

文本预处理技术包括分词、词性标注、句法分析等自然语言处理技术,这些技术相对比较成熟,国内也有若干软件及语言开放平台供研究人员使用。如中国科学院计算技术研究所研制的基于多层隐马模型的汉语词法分析系统 ICTCLAS^[2] (Institute of Computing Technology, Chinese Lexical Analysis System), 系统的功能有:中文分词、词性标注、命名实体识别和未登录词识别,分词正确率高达 97.58%;哈尔滨工业大学社会计算与信息检索研究中心研制的 LTP^[3] (Language Technology Platform) 开源语言技术平台具有分词、词性标注、命名实体识别、依存句法分析在内的一整套基于 XML 的中文语言处理模块。这些成熟技术的应用为文本情感分析奠定了良好的基础。

根据微博文本的特性,还需要对微博文本中的链接地址、“@”字符(用于回应或沟通其他用户)以及“#”字符(用于话题的归类)进行过滤工作。

2 微博情感信息抽取

情感信息抽取旨在抽取文本中有价值的情感信息,是根据预先给定的倾向性单元定义,抽取出文本中其所表达的倾向性单元的要素^[1]。本节结合微博的特点,总结情感词、主题和关系三类情感信息抽取的研究进展。情感词的抽取和判别是微博情感分析的基础;主题判断可以定位微博文本的领域,引入更多的情感判断标准,如专业领域词库;关系抽取可以辅助提炼微博文本的主体思想。

2.1 情感词抽取和判别

情感词,是指带有情感倾向的词语。微博中情感词的判别方法可分为利用词典计算相似度方法、基于大规模语料库的统计方法和表情符号的情感抽取方法(表情符号可以视为情感词的一种)。

2.1.1 基于情感词典的判别方法

语义知识库即词典的建立,给极性分析工作的开展以极大支持。利用词典能够直接判断相应词语的极性,同时也可以通过语义相似度、建立同义词典等方式对于未登录词进行极性的判断。文献[4]提出基于 HowNet 词汇语义相似度和语义相关场的情感词极性计算方法,其中基于语义相似度的方法收到了较理想的效果,词频加权后的判别准确率可达 80% 以上。文献[5]考虑目标词语其同义词的关系,提出了基于同义词的词汇情感倾向判别方法,取得了不错的效果。文献[6]基于 PMI_IR 和 HowNet 词汇提出了 Ontology 模型,将极性词图像化为“地球仪”,越近两极倾向性越明显,通过计算词汇相似度和相似度最大词汇在 Ontology 模型中的映射确定情感词的极性。

基于情感词典的判别难度在于情感词典的构建。中文里有较多的一词多义现象,且在不同的语境下表达的意义可能相反。此外,微博作为一款互联网应用,各式各样的网络用语不断涌现,现缺乏相应的网络语言情感词库。

2.1.2 基于大规模语料库的统计方法

基于统计的方法主要利用在大规模语料中挖掘出的语言学规则特征,以机器学习模型对词汇的情感极性进行判别。如文献[7]以情感词间的连接关系(如 and、but)作为特征来推断情

感词在某领域内的极性,提取了大量含情感特征的形容词。文献[8]预先挑选若干具有较明确极性信息的形容词构成种子词集合,通过计算待测词与褒、贬种子极性词(如“好”、“坏”)之间的点态互信息差值(SO-PMI)来确定待测词汇的极性特征。文献[9]建立基于二元语法以来关系的情感倾向互信息特征模型,通过机器学习方法自动判别词语的情感极性。

2.1.3 表情符号的情感抽取

中文微博网站给用户提供了丰富的表情符号,如图 1 所示。

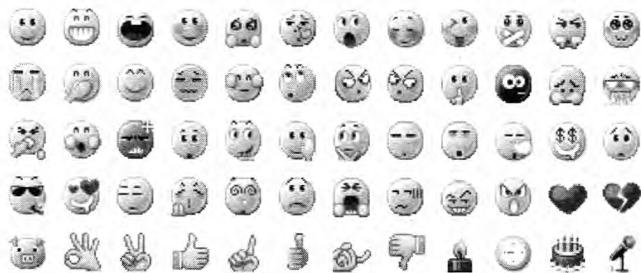



图1 新浪微博平台提供的表情符号

用户选择的表情符号通常反映了用户的心情,即表情符号蕴含了情感信息。文献[10]指出表情符号在微博文本中以“[. * ?]]”的正则表达式出现,如,对应的文本元素为[哈哈]。

对常用的表情进行统计,按情感分类项进行预先分类,可快速确定其在倾向性单元内的极性和强度。

2.2 主题抽取

主题又称为评价对象,多数为名词。

常规方法有利用预定义的本体树和层次化学习来分析复杂情况^[11]。为了实现自动构建层次化结构,可以利用领域特征,从而免去人工构建本体库的困难^[12]。文献[13]提出了利用句法结构相似度以及启发式的方法来查找可能的评价对象的方法。文献[14]通过结合《知网》本体库,提出基于语义的微博短信息主题分类方法。

结合微博的特性,还有以下两种提取主题的方法:微博在产品设计上通常用“#”来表示主题,如“#股市#”代表这是一条以股市为主题的微博,故可以借助“#”定位主题。文献[15]利用微博数据富含链接信息的优势,将文档解析成文档链。然后利用文本表示模型对文档链进行模型表示,根据预先设定的共现度阈值采用聚类方法抽取不同主题。

2.3 关系抽取

关系抽取指的是获取句子中评价词和评价对象对应的修饰关系。比如,“股价炒得好高啊”一句中,评价对象和修饰词间的关系为“股价—高”。

常用的方法有根据词性建立语言模版。如文献[16]先寻找语句中的目标词,通过对文本句法进行分析,得出与目标词有修饰关系的词,根据 7 条规则模版对关系进行分类。这些模版大多只能提取出浅层的情感关系,有许多局限性,精确度不高。文献[17]在此基础上将程度副词加入了特征,一定程度上提高了精确程度。

此外,也可根据语义结构进行句法分析,找到评价词和对应的修饰关系。如文献[18]通过句法路径的生成(如图 2 所示)和句法路径的两步泛化之后,建立情感评价单元的句法路径库。

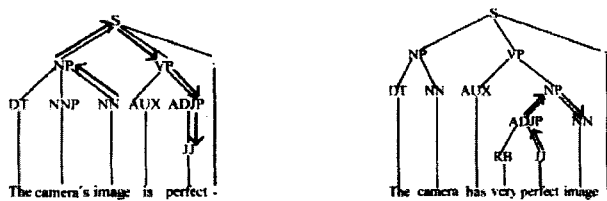


图 2 句法路径示例

3 微博情感分类

文献[19]指出只有带有主观色彩的文本才会蕴含着作者的情感。故对微博文本进行情感识别, 首先需要从原素材进行主客观的文本分类预处理。

3.1 主客观文本分类

主观性文本的甄别方法主要是联系语言表达习惯, 提取主观性特征, 再应用机器学习算法进行分类。如文献[20]通过实验选择了包括人称代词、不规范和含有感情色彩的标点符号及感叹号等六种稳定特征, 采用了 10 折交叉验证(Ten-folder Cross Validation)方法作为分类算法的测试方法。实验表明, 分类算法的 F 度量最高时可以达到 93.8%, 平均 F 度量也达到了 88.4%。此外, 文献[21]提出一种根据连续双词词类组合模式(2-POS), 利用 CHI 统计方法提取主观文本词类组合方式, 自动判别句子主客观性程度的方法。实验中, 主观文本的分类查准率和查全率达 76%。

考虑到个体的语言习惯存在规律性, 而多用户整体能够丰富判断特征, 文献[22]提出采用合作在线学习算法来对含有情感和不含情感的微博(即主客观微博)进行学习分类。一方面学习单个用户的微博数据, 另一方面对多个用户整体的微博数据进行学习, 如图 3 所示。

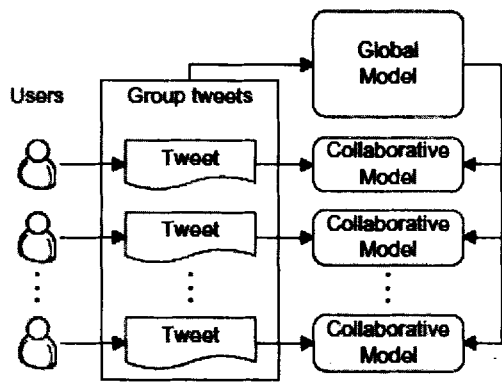


图 3 合作在线学习方法

3.2 主观微博情感分类

目前针对中文微博的情感分类方法可以分为两类: 基于语义词典的情感计算法和基于机器学习的情感分类法。

3.2.1 基于语义词典的微博情感计算

通过对语义词典进行情感的规则性研究, 可以找到情感判断的一定规律性。如文献[23]通过定义态度词典、权重词典、否定词典、程度词典以及感叹词词典来计算每条微博的情感指数。具体算法如图 4 所示。

$$F = \begin{cases} \sum_{i=1}^n ExcWeight(a_i) \left| \sum_{j=1}^m A(a_i, v_j) Deg(a_i) Neg(a_i) \right| & (\text{If there are interjection words}) \\ \sum_{i=1}^n \sum_{j=1}^m A(a_i, v_j) Deg(a_i) Neg(a_i) & (\text{otherwise}) \end{cases} \quad (III)^{+}$$
$$Neg(a_i) = \begin{cases} 1 & (\text{when number of negative words in } a_i \text{ is even}) \\ -1 & (\text{when number of negative words in } a_i \text{ is odd}) \end{cases} \quad (IV)^{+}$$
$$ExcWeight(a_i) = \begin{cases} 1 & (\text{positive interject on word}) \\ -1 & (\text{negative interject on word}) \end{cases} \quad (V)^{+}$$

图 4 基于语义词典的情感分析算法

基于语义词典的方法主要存在以下两个问题:

- (1) 尽管考虑到了感叹词对评价词强度的增加, 中文否定词、多重否定对情绪表达的影响, 但当一句话中辩证出现正反两方面阐述、或者句子中重复出现某情绪词的时候, 仅靠权重计算, 会产生很大的误差;
- (2) 所选的情感词往往是情感特征比较显著的词语, 在词典的构建上存在很多歧义的情况, 又忽略了许多能代表情感的词语。

3.2.2 基于机器学习的微博情感分类

机器学习分类方法主要是应用机器学习模型, 如支持向量机、朴素贝叶斯、最大熵等, 通过对训练集的特征进行学习, 构造模型, 从而应用于对测试集的分类判断。

如文献[24]使用三种机器学习算法、三种特征选取算法以及三种特征项权重计算方法对微博进行了情感分类的实证研究。研究发现, 综合考虑三种因素, 采用支持向量机(SVM)和信息增益(IG), 以及 TF-IDF(Term Frequency-Inverse Document Frequency)作为特征项权重, 三者结合对微博的情感分类效果最好。

在对机器学习分类效果的优化上, 可以考虑将微博分为不同的类别, 根据每种类别的特性采用效果最佳的学习策略。如文献[10]在对于 SVM 的训练方法进行探讨时, 依据是否对微博进行分句划分为两大策略, 然后再在每类策略下细分, 共包括四种方法。采取第二类分句策略主要是考虑到中文 140 字的微博可能包含多句子内容, 句与句的情感极性可能不同。其策略及方法如图 5 所示。

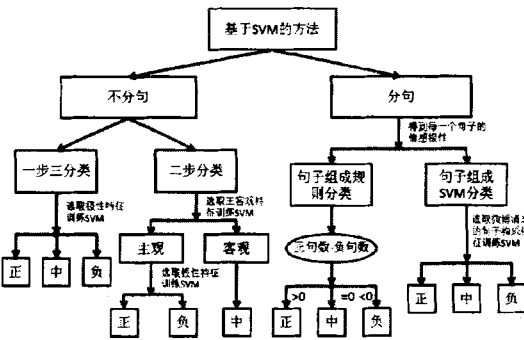


图 5 文献[10]基于 SVM 的情感分析的不同策略及方法

基于机器学习的微博情感分类方法主要有以下关键点:

- (1) 特征提取的方法。特征的提取方式多样, 如文献[25]根据表情符号结合词语 PMI 进行微博情绪特征提取; 文献[26]根据词语的属性和语义类别建立情绪关联规则, 用这些关联规则作为学习模型训练过程中的微博情绪特征。

此外, 文献[10]提出的策略考虑到了中文微博 140 字包含的句子情感极性并不单一的情况。所以在对微博情感特征提取的时候, 还需要综合考虑上一节所述的关系特点, 明确句子主要

表达的情感色彩,如“虽然……但是……”,其主要情感色彩趋于“但是”后面的句子。

(2) 学习语料库的建设。现有的微博语料库资源如由北京理工大学网络搜索挖掘与安全实验室^[27]整理的 NLPiR 微博内容语料库 23 万条。语料库可以按照专业领域建设,和微博的主题结合以进一步提升机器学习判断的正确率。

4 延伸研究

前文主要介绍了基于微博文本内容的情感分析。事实上,对于微博的情感分析可以通过其产品的其他数据特征做延伸挖掘。如文献[28]在研究英文微博 Twitter 时提出了基于图形的 Hashtag(Twitter 中用以类似“#obama”用以强调或概括主题名的标签)情感分类,他们发现带有情绪色彩的 hashtag 常伴随客观主题的 hashtag 出现,他们提出的方法可以对微博文本的情感分析给予补充和强化。

微博上与情感相关的数据还包含关键词发言量(如:统计微博文本中出现关键词“高兴”的日发言量)、微博的评论数和转发量(可以反映大众对于某一事件的情绪反应效果,如北京大学 PKUVIS 微博可视分析工具^[29]可以用视图清晰地呈现出一个事件中微博转发的过程,迅速地发现事件中的关键人物、关键微博、重要观点,从而把握微博平台上主要的事件情感倾向)、“微心情”数据(新浪微博平台上上线不久的应用,用户可以通过微心情窗口快速记录自己的心情状态,包含开心、得意、无聊、抓狂等 10 个心情类别)、微博投票数据(可以用来样本采集大众对于某一事件的情感倾向)等。

通过对这些数据的挖掘和提取,不仅可以对微博文本的情感分析进行改善和提升,并可以从一定程度上代替判断微博文本情感倾向本身。

5 结 语

微博的情感分析与传统的文本情感分析相比,有其相似性和特殊性。特殊性主要体现在微博的情感分析并不局限于微博原文本身,还可以从微博的产品特征出发,对微博的情感进行多样化的分析。中文微博情感分析作为数据挖掘的一个新的研究方向还有很多值得深入研究的课题。本文只是基于现有的研究成果,主要针对微博文本的情绪分析作了相关的介绍。未来对微博情感分析应该有以下几个方向:

- 1) 微博上存在许多垃圾信息。这些信息的存在无疑会干扰基于情感分析的相关研究,目前尚且没有成型的过滤技术。
- 2) 在对网络语言的过滤和情感挖掘上较为缺乏,没有相关的词典或语料库供使用,需要重点开发。
- 3) 微博的情绪较为丰富,不应只局限于正、负两个方面,并可以进一步延伸对于情绪强弱的判别。
- 4) 基于主题的不同,做特定的情感分析研究,如挖掘和股市相关的微博判断情感倾向。不同的主题,其语言规则,和词库判断标准都存在不同,可以对情感分析效果进一步提升。
- 5) 可以将视角拓展到整个微博社区的范畴,分析社区整体情感走势,如挖掘分析特定的关键词日发言量。

参 考 文 献

- [1] 黄萱菁,张奇,吴苑斌. 文本情感倾向分析[J]. 中文信息学报, 2011, 25(6): 118-126.
- [2] 张华平. 基于多层隐马尔科夫模型的中文词法分析[C]//第 41 届 ACL 会议暨第二届 SIGHAN 研讨会, 札幌, 日本, 2003: 63-70.
- [3] 李正华,车万翔,刘挺. 基于 XML 的语言技术平台[C]//第五届全国过青年计算语言学研讨会(YWCL), 武汉, 中国, 2010.
- [4] 朱嫣岚,闵锦,周雅倩,等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.
- [5] 王素格,李德玉,魏英杰,等. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报, 2009, 23(5): 68-74.
- [6] 王晓东,刘倩,陶县俊. 情感 Ontology 构建与文本倾向性分析[J]. 计算机工程与应用, 2010, 46(30): 117-120.
- [7] Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives[C]//Proc. Of the EACL' 97. Morristown: ACL, 1997: 174-181.
- [8] Turney P, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Trans. On Information Systems, 2003, 21(4): 315-346.
- [9] 张靖,金浩. 汉语词语情感倾向自动判断研究[J]. 计算机工程, 2010, 36(23): 194-196.
- [10] 谢丽星. 基于 SVM 的中文微博情感分析的研究[D]. 清华大学, 2011.
- [11] Wei Wei, Jon Atle Gulla. Sentiment Learning on Product Reviews via Sentiment Ontology Tree[C]//Proceedings of the Association for Computational Linguistics (ACL), 2010.
- [12] Yu Jianxing, Zha Zhengjun, Wang Meng, et al. Tat-Seng Chua, Domain-Assisted Product Aspect Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.
- [13] Zhao Yanyan, Qin Bing, Hu Shen, et al. Generalizing Syntactic Structures for Product Attribute Candidate Extraction[C]//Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL 2010), 2010.
- [14] 崔争艳. 基于语义的微博短信息分类[J]. 现代计算机: 专业版, 2010, 8(8): 18-24.
- [15] 王岩. 基于共现链的微博情感分析技术[D]. 国防科学技术大学, 2011.
- [16] Long Jiang, Mo Yu, Ming Zhou, et al. Target-dependent Twitter Sentiment Classification[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (AMACL), 2011.
- [17] 章剑锋,张奇,吴立德,等. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报, 2008, 22(2): 55-59.
- [18] 赵研研,秦兵,车万翔,等. 基于句法路径的情感评价单元识别[J]. 软件学报, 2011, 22(5): 887-898.
- [19] 王洪伟,刘懿,尹裴,等. Web 文本情感分类研究综述[J]. 情报学报, 2010, 29(5): 931-938.
- [20] 姚天昉,彭思威. 汉语主客观文本分类方法的研究. 中文短句之情绪分类[C]//第三届全国信息检索与内容安全学术会议, 2008.
- [21] 叶强,张紫琼,罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法[J]. 信息系统学报, 2007, 1(1): 79-91.
- [22] Li Guangxia, Steven C H Hoi, Chang Kuiyu, et al. Micro-blogging Sentiment Detection by Collaborative Online Learning[C]//Proceeding of IEEE International Conference on Data Mining, 2010.
- [23] Shen Yang, Li Shuchen, Zheng Ling, et al. Emotion Mining Research on Micro-blog[C]//Web Society, 2009. SWS'09. 1st IEEE Symposium, 2009.

(下转第 181 页)

表 3 事件序列数据库表结构

列名	时间戳	事件 ID	事件描述
长度(字节)	4	4	100

用 SQL 访问数据库,扫描事件序列时用时间戳索引相应事件。

实验比较三种算法:(1) SC DFA,基于 Apriori 逐层计算框架,每一层为候选集中每个情节维护一个 SC DFA 来计算支持度;(2) SCTree,基于 Apriori 逐层计算框架,每一层用 SCTree 计算候选集中所有情节的支持度;(3) SCTree ++,采用 SCTree 主动扩展技术的挖掘算法。固定 $\rho = 0.5$,比较在最小支持度 minsupp 取不同值时的运行时间。

首先比较 SC DFA 和 SCTree 算法的运行时间,如图 4 所示。

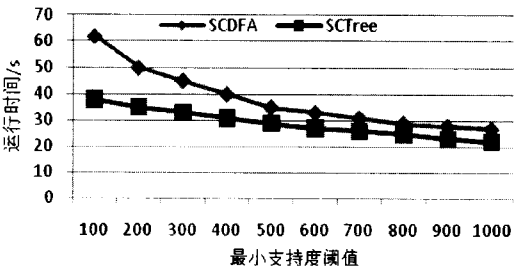


图 4 SC DFA 与 SCTree 效率比较

可以看出:(1) SC DFA 算法的运行时间比 SCTree 大;(2) 随着 minsupp 增大,两种算法运行时间均降低。分析原因:随着 minsupp 增大,频繁情节减少,裁剪效果增大,因此运行时间减少;(3) 随着 minsupp 增大,SC DFA 与 SCTree 差距变小。分析原因:随着 minsupp 增大,频繁情节减少,因此 SCTree 的压缩效果降低,从而运行时间差距变小。

然后比较 SCTree 和 SCTree ++ 算法的运行时间。实验结果如图 5 所示。

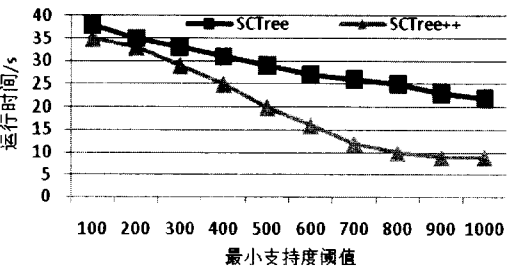


图 5 SCTree 与 SCTree ++ 效率比较

可以看出:(1) SCTree 算法运行时间比 SCTree ++ 大;(2) 随着 minsupp 增大,SCTree ++ 与 SCTree 运行时间差距增大。分析原因:随着 minsupp 增大,频繁情节减少,因此一次可主动扩展的层数增多,从而减少 SCTree ++ 的运行时间。表 4 显示 SCTree 和 SCTree ++ 扫描数据库的次数。

表 4 SCTree 和 SCTree ++ 扫描数据库的次数

算法\minsupp	100	300	500	700	900
SCTree	10	10	10	8	7
SCTree ++	9	9	7	5	4

6 结 语

本文提出了一种互异情节模式挖掘算法,引入带状态计数
万方数据

的前缀树(SCTree)结构来生成互异情节模式候选集,进行互异计数和裁剪。为减少数据库扫描次数和重复支持度计算,提出 SCTree 的主动扩展技术,基于此提出高效情节模式挖掘算法。实验表明了算法的有效性和高效性。

参 考 文 献

[1] Agrawal R, Srikant R. Mining Sequential Patterns[C]//Proc. of Intl. Conf. on Data Engineering, Taipei, Taiwan, 1995: 3 - 14.

[2] Seno M, Kartpis G. SLPMiner: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint [C]//Proc. of the 14th Intl. Conf. on Data Engineering, Maebashi City, Japan, December, 2002: 418 - 425.

[3] Mannila H, Toivonen H, Verkamo A I. Discovery of Frequent Episodes in Event Sequences[J]. Data Mining and Knowledge Discovery, 1997, 1(3): 259 - 289.

[4] Casas-Garriga G. Discovering unbounded episodes in sequential data [C]//Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD ' 03), Cavtat-Dubrovnik, Croatia, 2003: 83 - 94.

[5] Meger N, Rigotti C. Constraint-based mining of episode rules and optimal window sizes[C]//Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD ' 04), Pisa, Italy, 2004.

[6] Joshi M V, Karypis G, Kumar V. A Universal Formulation of Sequential Patterns[C]//Proc. of the KDD ' 2001 workshop on Temporal Data Mining, San Francisco, CA, August, 2001.

[7] Laxman S, Sastry P S, Unnikrishnan K P. Discovering frequent episodes and learning Hidden Markov Models: A formal connection[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1505 - 1517.

[8] Laxman S, Sastry P S, Unnikrishnan K P. A Fast Algorithm For Finding Frequent Episodes In Event Streams [C]//Proc. of the KDD ' 2007, San Jose, California, USA. August 12 - 15, 2007.

[9] 黄鹏. 面向事件流的频繁情节片段计数算法[J]. 计算机科学与探索, 2010, 4(10): 909 - 917.

(上接第 164 页)

[24] 刘志明, 刘鲁. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1): 1 - 4.

[25] Sun Yingts, Chen Chienliang, Liu Chunchieh, et al. Emotional Classification of Chinese short sentences[C]//Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), Puli, Nantou, Taiwan, September 2010.

[26] Wu C H, Chuang Z J, Lin Y C. Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models[C]//ACM Transactions on Asian Language Information Processing, 2006.

[27] 北京理工大学网络搜索挖掘与安全实验室博客[OL]. www. nlpir. org.

[28] Wang Xiaolong, Wei Furu, Liu Xiaohua, et al. Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach [C]//Proceeding of 20th ACM Conference on Information and Knowledge Management(CIKM), Glasgow, 2011.

[29] 北京大学 PKU VIS 微博可视分析工具[OL]. http://vis. pku. edu. cn/weibova/about/intro/.