Journal of Computer Applications
计算机应用,2014,34(2):481-485,495
ISSN 1001-9081
CODEN JYIIDU
2014-02-10
http://www.joca.cn

文章编号:1001-9081(2014)02-0481-05
doi:10.11772/j.issn.1001-9081.2014.02.0481

基于语义规则的 Web 金融文本情感分析

吴　江[1*],唐常杰[2],李太勇[1],崔　亮[3]
(1. 西南财经大学 经济信息工程学院,成都 610074;　2. 四川大学 计算机学院,成都 610064;
3. 西南财经大学 统计学院,成都 610074)
(*通信作者电子邮箱 wuj_t@swufe.edu.cn)

摘　要:为有效提高非结构化 Web 金融文本情感倾向和强度分析的精度,提出了基于语义规则的 Web 金融文本情感分析算法(SAFT-SR)。该算法基于 Apriori 算法对金融文本进行属性抽取,构建金融情感词典和语义规则识别情感单元及强度,进而得到文本的情感倾向和强度。实验结果表明,与 Ku 提出的算法相比,在情感倾向分类方面,算法 SAFT-SR 情感分类性能良好,提高了分类器的 F 值、查全率和查准率;在情感强度计算方面,算法 SAFT-SR 的误差更小,更接近真实评分,证明了 SAFT-SR 是一种有效的金融文本情感分析算法。

关键词:Web 金融文本;情感词典;语义规则;情感分析;情感倾向
中图分类号:TP391　　**文献标志码**:A

Sentiment analysis on Web financial text based on semantic rules

WU Jiang[1*], TANG Changjie[2], LI Taiyong[1], CUI Liang[3]
(1. *School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu Sichuan* 610074, *China*;
2. *College of Computer Science, Sichuan University, Chengdu Sichuan* 610064, *China*;
3. *School of Statistics, Southwestern University of Finance and Economics, Chengdu Sichuan* 610074, *China*)

Abstract: In order to effectively improve the accuracy of sentiment orientation and intensity analysis of unstructured Web financial text, a sentiment analytical algorithm for Web financial text based on semantic rule (SAFT-SR) was proposed. The algorithm extracted features of financial text based on Apriori, constructed financial sentiment lexicon and semantic rules to recognize sentiment unit and intensity, and figured out the sentiment orientation and intensity of text. The experimental results demonstrate that SAFT-SR is a promising algorithm for sentiment analysis on financial text. Compared with Ku's algorithm, in sentiment orientation classification, SAFT-SR has better classification performance and increases F-measure, recall and precision; in sentiment intensity analysis, SAFT-SR reduces error and is closer to expert mark.

Key words: Web financial text; sentiment lexicon; semantic rule; sentiment analysis; sentiment orientation

0　引言

截至2012年末,我国已拥有超过2494家A股上市公司,然而随着全球金融市场的动荡,股票市场管理与优化及企业财务危机预测成为研究的热点。目前,大部分企业财务危机预测研究是基于财务报表数据来建立金融危机预测模型,但财务报表有以下缺点[1]:1)报表人为操作性强;2)基于静态数据,忽略了企业财务比率的时间序列特点;3)实效性较差;4)未考虑财务比率的历史累积值对现时的影响。因此,单纯利用财务报表进行判断,势必会造成预测结果的偏差。

财务报表和金融数据的局限性,使得人们寻求从其他角度着手于股票市场管理和企业财务危机的预测和研究。随着Internet的高速发展,Web信息量得到了前所未有的增长,公众在互联网上发布自己对企业的看法已司空见惯,普通投资者的情感倾向是联系投资者与股票市场、上市公司的桥梁,Web新闻或论坛对上市企业的评论可以直接反映出公众对该企业的看法。Web金融信息所具有的实时性、全面性和覆盖性等特点,不仅为财务危机预测研究提供了新的机遇,也为投资者情感分析提供了廉价且丰富的数据来源。由于Web金融信息是非结构化的文本信息,并具有领域知识,因此,如何对其进行深入挖掘,发现其中的情感倾向和强度,对文本挖掘提出了新的挑战。

本文基于语义规则,对Web金融文本进行情感分析,挖掘投资者的情感倾向和强度,并对投资者情感强度变化与股票市场之间的联动关系展开分析,可以为企业财务危机预测和股票市场管理与优化提供新的思路与选择。

1　相关工作

文本情感分析就是对带有情感色彩的词语、句子以及文本进行分析、处理、归纳和处置的过程[2]。文本情感分析可分为基于机器学习的分类方法和基于语义分析的方法两大类。运用机器学习的方法进行文本分类,先人工标注一些文本的情感倾向,作为训练语料,然后通过训练得到一个分类器,最后将测试语料用已训练好的分类器进行分类测试,得到

收稿日期:2013-08-22;**修回日期**:2013-10-23。　**基金项目**:教育部人文社会科学研究青年基金资助项目(11YJCZH084);贵州省自然科学基金资助项目(黔科合J字LKG[2013]45);西南财经大学"211工程"三期青年教师成长项目(2011QN2011050)。

作者简介:吴江(1980-),男,浙江衢州人,副教授,博士,主要研究方向:数据库、知识工程;　唐常杰(1946-),男,重庆人,教授,博士生导师,主要研究方向:数据库、知识工程;　李太勇(1979-),男,四川安岳人,副教授,博士,主要研究方向:数据挖掘;　崔亮(1981-),男,山东潍坊人,博士,主要研究方向:金融统计。

文本的情感倾向。Pang 等[3]运用朴素贝叶斯网络、最大熵模型和支持向量机三种分类器对于影评进行了分类研究。Cui 等[4]实验证明,当训练语料较少时,uni-gram 的效果最优,随着训练语料的增多,n-gram($n>3$)效果较好。

基于语义的情感倾向分析研究是对文本计算一个情感倾向值,值的符号表示其倾向性,而其绝对值的大小则反映其情感强度。基于语义的情感倾向分析又分为两类:基于情感词的文本倾向性分析和基于语义规则的文本倾向性分析。基于情感词的文本倾向性分析首先抽取出文本中的情感词,然后对情感词逐一进行情感倾向判断,得到各自的情感倾向值,最后通过累加这些倾向值获得文本最终的情感倾向和强度。代表性的研究有:Turney[5]运用点互信息和潜在语义分析方法计算目标词汇和种子词之间的关联度,进而得出目标词汇的倾向性;Yuen 等[6]在 Turney 研究的基础上,对中文极性词的自动获取进行了研究;朱嫣岚等[7]利用 HowNet 提供的语义相似度和语义相关场,计算目标词汇与已标注褒贬性的种子词之间的相似度,提出了词语倾向性判断方法。基于语义规则的文本倾向性分析首先建立一个情感倾向语义模式库,然后将文本按照这个语义模式库进行模式匹配,计算得到一系列情感倾向值,最后将这些倾向值进行累加,得到整个文本的情感倾向和强度。代表性的研究有:Wiebe 等[8]对语料库标注了级别(文档级、短语级和句子级),在此基础上,利用词语的搭配模式发现文本中的倾向性词语及其搭配关系;Wilson 等[9]研究证实了合并语言信息能显著地改进了细粒度情感分析的性能;Takamura 等[10]提出了 Latent Variable Models 用于短语的语义倾向性研究;Matsumoto 等[11]从组成或依赖结构抽取子串改善句子层模型的性能;Ku 等[12]对新闻和博客文本从词级、句子级和文档级进行了意见抽取,得出观点摘要,进而对文本进行情感倾向和强度分析。

从已有研究可以发现,文本倾向性分析已引起了学者们的普遍关注,但尚未被广泛应用于金融领域。Pang 的研究表明,在情感倾向性研究中,统计方法的准确率要高于机器学习方法。因此,本文立足于基于语义的情感倾向性研究,针对 Web 金融文本的特点,充分考虑句子中否定词和程度副词对文档极性转移的作用及其不同权重,提出了一种基于语义规则的 Web 金融证券域文本情感分析方法,并在此基础上,对投资者情感变化与股票市场之间的联动关系展开分析。

2 基于语义的 Web 金融文本情感分析

2.1 总体框架

Web 上的金融文本主要分为两大类:一类是各金融网站的新闻、专家评论、公告等;另一类是各股吧论坛中的用户评论帖子。其中,第二类信息,即普通投资者所发布的信息更能够反映投资者的情感倾向,因此本文选取股吧论坛中的用户评论帖子作为研究对象,在对 Web 金融文本进行预处理、分词和词性标注后,提取情感词以及可以影响情感倾向的副词和否定词等,通过定义语义规则进行匹配,计算文本情感倾向和强度。情感倾向值计算包含以下几步:1) 文本预处理;2)文本属性(特征)提取;3) 金融情感词典构建;4) 语义规则构建;5) 基于语义规则的情感单元识别和情感值计算;6) 整篇文档的情感倾向值计算。总体框架如图 1 所示。

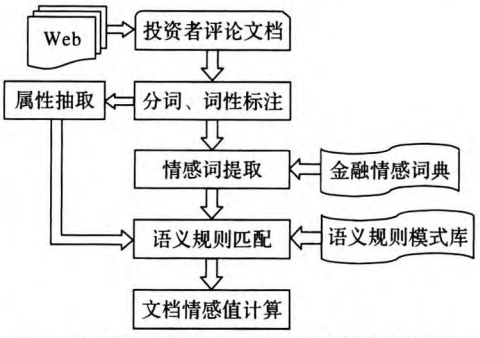

图1 基于语义规则的 Web 金融文本情感分析框架

2.2 Web 金融文本采集

投资者情感来源于网络舆论,包括股吧论坛、博客、微博、社交网站等多种媒体形态,因此数据的采集应考虑大型金融类论坛,研究选用东方财富网论坛投资者评论文本。在文本采集方面,采用网络蜘蛛进行收集。

2.3 文本预处理

2.3.1 网页解析及噪声消除

对下载生成的网页文件,编写 Java 程序,解析文件,将解析结果导入到数据库中以备后续查询与分析使用,保留的主要字段有股票代码、发表时间、标题、内容、点击量和回复量等信息。

去除噪声文本的方法如下:人工选取有效帖,统计有效帖文件大小、点击量、回复量等帖子特征,统计分析有效帖各特征的合理范围,依据该统计特征去除噪声文本,减小后续数据处理的压力。

2.3.2 中文分词和词性标注

研究以在线评论中的句子为单位,首先对得到的评论语料进行断句处理,按照评论中出现的标点符号(分号、句号、问号、叹号等),空格符等进行断句;同时应用中国科学院分词器 ICTCLAS3.0,对评论文本中的句子进行分词和词性标注。

2.4 属性提取

经过噪声处理的文本信息仍然不能满足对情感分析的需要,因为这里面依然含有大量的与上市公司属性无关的描述,这些描述可能含有情感,但是与主题无关,不能计入对上市公司的情感倾向计算。因此,首先应提取上市公司属性(特征),后续只对上市公司属性(特征)所在的句子进行情感分析,以此排除噪声信息的干扰。

设计了一个基于 Apriori 算法的属性提取算法 FEAA(Feature Extraction Algorithm based on Apriori),针对股吧金融文本信息特点,实现从海量文本信息中挖掘投资者情感关注的属性词,具体算法如下:

算法 1 基于 Apriori 算法的属性提取算法。

输入 所有评论文本;

输出 金融文本关键属性(特征)。

1) 对股吧评论信息进行中文分词及词性标注,创建关联规则事务文件 I;

2) 基于 Apriori 算法从事务文件 I 中,找到频繁项集作为候选特征集合 I_0;

3) 将 I_0 按照邻近规则修剪,成为候选特征集合 I_1;

4) 将候选特征集合 I_1 按照独立支持度规则继续修正,形成候选特征集 I_2;

5) 对 I_2 中频繁项名词进一步过滤,去掉非属性名词(如专有名词、时间名词、人称名词、口语化名词等) 和单字名词,过滤形成 I_3;

6) 对未包含在 I_3 中的非频繁项属性名词,人工补充形成 I_4,得到金融文本关键属性集合。

完成属性提取后,将重点对含有属性词的句子进行情感倾向分析,从而去除无关信息对投资者情感倾向分析的干扰。

2.5 情感词典构建

针对 Web 金融文本的特点,构建了一个包括基础词典、领域词典、网络词词典以及修饰词词典的情感词典。

1) 基础词典。

基础词典主要利用了《知网》、《国立台湾大学情感词典》和《学生褒贬义词典》提供的褒贬义情感词语,通过去重之后作为基础情感词典。

2) 领域词典。

某些极性词只在特定的领域才被使用,且具有情感倾向,如"涨停""利多";还有一些极性词在不同的领域修饰不同的特征时会表现出不同的情感,例如"升高"在描述工资收入时是褒义的,而在描述利率时对股票市场就是不利消息,可看成是贬义。本文利用常用的证券操作词汇表,提取具有情感倾向的词语进行人工筛选,构建了一部股票投资领域的情感词典。为了提高情感分析的准确性,还选取一定规模的股吧评论语料,抽取情感词进行人工标注,也加入领域词典。

3) 网络词词典。

大量涌现的网络用语,在一段时间内常被用来表达人们的情感倾向。因此,把使用频繁且带有情感倾向的网络用语加入所构建的情感词典中来,以满足对网络评论信息情感分析的需要。

4) 修饰词词典。

当程度副词或否定副词修饰情感词时,整个情感的情感极性和强度都可能发生变化,因此构建了一个包括否定副词和程度副词的修饰词词典。根据文献[13]中对否定副词范围的界定,选取 31 个否定副词,采用蔺璜等[14]对程度副词的分类,并结合《知网》中程度副词,选取了 212 个程度副词。

5) 情感词典扩展。

对于文本中的新词,即在以上构建的情感词典中检索不到的候选情感词,基于点互信息的算法对情感词典进行进一步扩展。

经过以上步骤,构建的情感词典含有 31 个否定副词,212 个程度副词,21 333 个情感词语,其中 7 779 个正面情感词语,13 554 个负面情感词语。

2.6 情感倾向和强度分析

基于语义规则的 Web 金融文本情感分析算法(Sentiment Analysis Algorithm for Web Financial Text Based on Semantic Rule, SAFT-SR)的基本思想是:对文本中的每个存在关注属性(特征)的句子,按照预设的语义规则,计算情感分析单元的情感强度,将这些情感分析单元的情感强度进行累加,求得平均值作为句子的情感倾向,然后对句子情感强度进行累加求平均,作为整个文本的情感倾向和强度。

基于极性累加判断句子的情感强度的算法流程如下:首先对待分析文本进行中文分词和词性标注,若文本中句子不包含属性词和其相关的情感词(正向情感词或者负向情感词),则认为这些句子是中性的,不进行分析,对文本中含有属性词及相关情感词的每一个句子 S:$sw_1, sw_2, \cdots, sw_m$,其中 sw_j 表示句子 S 中所包含的第 j 个属性词所在的情感分析单元,m 表示句子 S 中拥有的属性词的数量,则有:

$$E(S) = \frac{1}{m}\sum_{j=1}^{m} E(sw_j) \tag{1}$$

其中 $E(sw_j)$ 表示第 j 个属性词所在的情感分析单元的情感强度,语义规则和情感值见表 1 所示。$E(S)$ 表示句子的情感强度。

表 1　情感分析单元情感强度的计算规则

模式	情感强度计算公式	例句	情感值
$U = PW$	$E(PW)$	经营业绩好	0.8
$U = NW$	$E(NW)$	经营业绩差	-0.8
$U = NA + PW$	$E(NA) * E(PW)$	经营业绩不好	-0.64
$U = NA + NW$	$E(NA) * E(NW)$	经营业绩不差	0.64
$U = NA + NA + PW$	$E(NA) * E(NA) * E(PW)$	经营业绩不是不好	0.512
$U = NA + NA + NW$	$E(NA) * E(NA) * E(NW)$	经营业绩不是不差	-0.512
$U = DA + PW$	$E(PW) + (1 - E(PW)) * L(DA)$	经营业绩很好	0.94
$U = DA + NW$	$E(NW) + (-1 - E(NW)) * L(DA)$	经营业绩很差	-0.94
$U = NA + DA + PW$	$E(PW) + (1 - E(PW)) * (L(DA) - 0.2)$	经营业绩不很好	0.9
$U = NA + DA + NW$	$E(NW) + (-1 - E(NW)) * (L(DA) - 0.2)$	经营业绩不很差	-0.9
$U = DA + NA + PW$	$E(NA) * E(PW) + (1 - E(PW) * E(NA)) * E(PW)$	经营业绩很不好	-0.892
$U = DA + NA + NW$	$E(NA) * E(NW) + (-1 - E(NW) * E(NA)) * E(NW)$	经营业绩很不错	0.892

表 1 中,PW 代表正向情感词,NW 代表负向情感词,NA 代表否定副词,DA 代表程度副词,U 表示情感分析单元,$E(PW)$、$E(NW)$ 和 $E(NA)$ 分别代表正向情感词、负向情感词和否定副词的情感强度。根据程度不同,程度副词的情感强度 $L(DA)$ 分别设定为 0.9,0.7,0.5 和 -0.5。

若整篇文章包含 n 个情感句,则篇章情感强度计算可以通过篇章中每个句子的情感强度计算得到,如式(2)所示:

$$E(T) = \frac{1}{n}\sum_{i=1}^{n} E(S_i) \tag{2}$$

其中:$E(T)$ 代表篇章的情感值,由篇章中情感句的平均强度决定;$E(S_i)$ 是每个情感句的情感强度。基于语义规则的 Web 金融文本情感分析算法如算法 2 所示。

算法 2　基于语义规则的 Web 金融文本情感分析算法。

输入　金融评论文本;

输出 文本情感倾向和强度值。

1）文本预处理(分词和词性标注)；

2）调用基于 Apriori 算法的属性提取算法(FEAA)抽取出属性(特征)；

3）识别出文本中包含属性词和其相关的带有情感词的句子；

4）对每个情感句,按照表 1 识别出情感计算单元；

5）按照表 1 和式(1),计算情感句中的每个情感计算单元的情感值并求得每个情感句的情感值；

6）按照式(2),计算整篇文档的情感值,得出篇章情感倾向和强度。

3 实验及数据分析

3.1 实验数据集

实验数据选取国内最有影响力的财经金融论坛——东方财富网股吧作为文本来源,借助 MetaSeeker 的两个组件 MetaStudio 和 DataScraper 来实现网页的下载,采集 2010 年 10 月至 2012 年 5 月,沪深 300 成分股的股吧 1 000 多万个评论帖子作为原始信息数据。

预处理按照 2.3.1 节提出的方法进行统计分析,获取有效帖的统计特征,按照所获取的特征,将文件容量小于4 KB或大于 100 KB 的文件作为噪声帖排除掉；另外通过对股吧评论信息关注特征分析,确定把点击量小于 100 或者回复量等于零的帖子判定为噪声帖。由此,得到了 30 万有效帖,为了减少后续人工标注的工作量,随机从 30 万帖子中抽取 1 万帖子作为实验数据集。

3.2 数据集标注

选取熟悉领域知识的 3 个人作为文本情感标注者。将 3 人中多数人的标注结果作为最后的标注结果。标注完成后,进行标注者间信度分析,然后合并标注后的结果,确定最终结果。表 2 给出了 3 个标注者两两间标注相同的百分比和三者标注一致的百分比。

表 2 标注结果一致率

标注者	一致率/%
A vs. B	86.54
A vs. C	86.51
B vs. C	86.31
A vs. B vs. C	76.33

从表 2 可以看出,标注者间的一致率还是比较高的,主要是因为金融文本的情感倾向一般比较明显。然而随着标注者数目的增加,一致标注的相同率会有所下降。由于只考虑文本的情感倾向,所以剔除了中性的标注结果。最后,实验数据集中只包含了5172 条情感倾向为正的文本和3639 条情感倾向为负的文本。

3.3 情感倾向和强度评测

3.3.1 情感倾向评测指标

在情感倾向评测中选择了查全率(*Recall*)、查准率(*Precision*)和 *F* 值(*F-measure*)三个指标来进行评价。查全率反映了一个分类器的泛化能力,查全率高说明这个分类器能够把正确的类别识别出来。查准率反映了一个分类器对于类别的区分能力,查准率越高,表明分类器识别出的正确分类数与总分类数差距不大,即识别的错误率较低。*F* 值(*F-measure*)将查全率和查准率一并列入新的综合评价指标。参见表 3,正向文本查全率和查准率,负向文本查全率和查准率及相应的 *F* 值的定义如下。

$$Recall(\mathrm{P}) = \frac{A}{A+C} \tag{3}$$

$$Precision(\mathrm{P}) = \frac{A}{A+B} \tag{4}$$

$$F\text{-}measure(\mathrm{P}) = \frac{2 \times Recall(\mathrm{P}) \times Precision(\mathrm{P})}{Recall(\mathrm{P}) + Precision(\mathrm{P})} \tag{5}$$

$$Recall(\mathrm{N}) = \frac{D}{B+D} \tag{6}$$

$$Precision(\mathrm{N}) = \frac{D}{C+D} \tag{7}$$

$$F\text{-}measure(\mathrm{N}) = \frac{2 \times Recall(\mathrm{N}) \times Precision(\mathrm{N})}{Recall(\mathrm{N}) + Precision(\mathrm{N})} \tag{8}$$

其中:P 表示正向,N 表示负向。

表 3 情感倾向分类性能评价列表(文本数)

预测极性	实际正向极性	实际负向极性
正向极性	*A*	*B*
负向极性	*C*	*D*

3.3.2 情感强度评测指标

在情感强度评测方面,之前标注者在标注时不仅标注情感倾向,同时也标注情感强度,选择将情感倾向标注的多数(两位或三位)作为最后情感倾向,并将其标注(两位或者三位)情感值的平均值作为文本最后的情感强度。算法 SAFT-SR 的结果与标注结果间的误差计算如式(9)：

$$D(T) = \frac{1}{n}\sum_{i=1}^{n} | E(T_i) - C(T_i) | \tag{9}$$

其中:$D(T)$ 表示算法结果和专家标注之间差值的平均值,$D(T)$ 值越小,说明算法结果越接近专家标准,反之就越偏离专家标准；n 是总的文本个数；$E(T_i)$ 是算法 SAFT-SR 计算出来的第 i 个文本的情感强度；$C(T_i)$ 是专家标注的第 i 个文本的情感强度。由于 $D(T)$ 考虑文本中所有情感单元强度计算结果和专家标注结果之间差异的平均值,因此能较好地反映算法计算结果与标注结果之间的误差。

3.3.3 实验结果及分析

使用本文算法 SAFT-SR 和 Ku 算法[12]在上述数据集上分别进行实验。表4 和图2 分别给出了本文算法 SAFT-SR 和 Ku 算法对文本情感倾向判断的结果。

表 4 两种算法实验结果比较

指标	SAFT-SR/%	Ku 算法/%
Recall(P)	86.62	73.49
Recall(N)	76.45	57.68
Precision(P)	83.94	71.17
Precision(N)	80.08	60.49
F-measure(P)	85.26	72.31
F-measure(N)	78.22	59.05

从表 4 和图 2 可以得出,本文算法在正向文本上的 *F* 值是 85.26%,相对于 Ku 算法的 72.31% 提高了 12.95%,正向文本查全率 86.62% 和正向文本查准率 83.94%,相对于 Ku 算法的正向文本查全率 73.49% 和正向文本查准率 71.17%

分别提高了 13.13% 和 12.77%。本文算法在负向文本上的 F 值是 78.22%,相对于 Ku 算法的 59.05% 提高了 19.17%,负向文本查全率 76.45% 和负向文本查准率 80.08%,相对于 Ku 算法的负向文本查全率 57.68% 和负向文本查准率 60.49% 分别提高了 18.77% 和 19.59%。结果表明本文算法较 Ku 算法整体提高了情感倾向的识别精度,这是因为 Ku 算法在句子情感倾向计算时只进行简单的词汇情感统计或只是考虑到否定副词的修饰关系,并没有对其中的程度副词及句子的模式进行更深入的剖析,并且没有设计基于金融领域的情感词典。

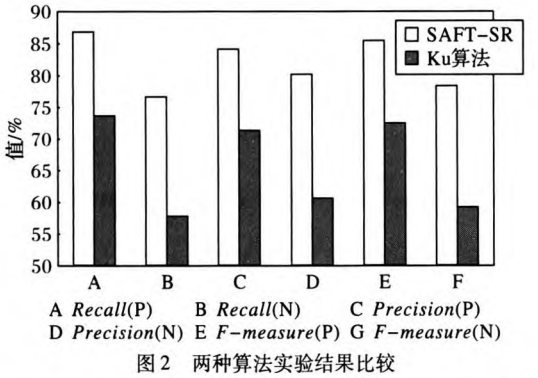

图 2　两种算法实验结果比较

由于沪深 300 包含了 300 支股票,限于篇幅,表 5 只列举了两个算法在前四支股票评论文本和所有股票评论文本情感强度上的 $D(T)$ 计算结果和专家的评分。通过对表 5 中结果比较可以发现,对于单支股票评论的情感强度,在前四支股票上,仅在南玻 A 一支股票上,Ku 算法略微好于本算法,而在其余三支股票上,本算法都好于 Ku 算法,情感强度误差值更小。对沪深 300 所有股票来说,在大多数情况下,Ku 算法比本文算法的误差大,在所有股票评论文本上 Ku 算法的平均误差较 SAFT-SR 高了 0.067,说明本文算法 SAFT-SR 计算的情感强度更接近专家评分,原因在于 SAFT-SR 算法考虑了文本金融领域的特点,且设计的语义模式更加符合人的理解模式。

表 5　两种算法实验结果比较

股票名称	专家标注	$D(T)$	
		本文算法	Ku 算法
中国平安	0.167	0.151	0.209
万科 A	0.292	0.198	0.225
中国宝安	0.205	0.095	0.124
南玻 A	0.195	0.198	0.155
沪深 300 所有股票文本	0.259	0.156	0.223

3.4　情感强度对股市影响效应分析

由于投资者情感强度的对数近似服从正态分布,因此对投资者情感强度的对数(ln(ISI))与股票市场特征变量的关系进行相关性分析,考察 2010 年 10 月至 2012 年 5 月,投资者情感强度与沪深 300 指数的对数(ln(price))、日成交量的对数(ln(volume))、日换手率(turnover)、日波动率(volatility)和日收益率(DR)等指标之间的相关系数,相关性分析结果如表 6 所示。

从表 6 可以看出,投资者情感强度(取对数)与沪深 300 指数(取对数)呈正相关,相关系数为 0.252;投资者情感强度(取对数)与日成交量(取对数)呈正相关,相关系数为 0.358;投资者情感强度(取对数)与日换手率呈正相关,相关系数为 0.319;投资者情感强度(取对数)与日波动率呈正相关,相关系数为 0.346;投资者情感强度(取对数)与日收益率呈显著正相关,相关系数为 0.432。在所有股市特征指标中,投资者情感强度与股市收益率的相关系数最大,也最为显著。因此可以认为,投资者情绪与股票市场价格和成交量呈正相关。

表 6　投资者指数与沪深 300 指数的相关性

指标	ln(ISI)	指标	ln(ISI)
turnover	0.319*	ln(price)	0.252*
volatility	0.346*	ln(volume)	0.358*
DR	0.432**		

注:**表示在 0.01 水平(双侧)上显著相关;
*表示在 0.05 水平(双侧)上显著相关。

4　结语

本文基于语义规则的文本倾向性分析技术,对非结构化的 Web 金融文本进行情感倾向和强度分析,构建了金融情感词典和语义规则,提出了基于 Apriori 的金融文本属性抽取算法(FEAA)和基于语义规则的 Web 金融文本情感分析算法(SAFT-SR)。实验结果表明,与 Ku 提出的算法比较,在情感倾向分类方面,本文算法 SAFT-SR 的 F 值、查全率和查准率均有较大提高;在情感强度计算方面,本文算法较 Ku 算法的误差更小,更接近真实评分。在今后的研究工作中,将进一步完善语义规则和情感词典,以进一步提高情感倾向和强度的计算精度。

参考文献:

[1] LI G. Sentiment computation of Web financial text based on semantic analysis[D]. Nanchang: Jiangxi University of Finance and Economics, 2012.(李国林. 基于语义分析的 Web 金融文本信息情感计算[D]. 南昌: 江西财经大学, 2012.)

[2] ZHAO Y, QIN B, LIU T. Sentiment analysis[J]. Journal of Software, 2010, 21(8): 1834 - 1848.(赵妍妍, 秦兵, 刘挺. 文本情感分析综述[J]. 软件学报, 2010, 21(8): 1834 - 1848.)

[3] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques[C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002, 10: 79 - 86.

[4] CUI H, MITTAL V, DATAR M. Comparative experiments on sentiment classification for online product reviews[C]// Proceedings of the 21st National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2006, 2: 1265 - 1270.

[5] TURNEY P D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 417 - 424.

[6] YUEN R W M, CHAN T Y W, LAI T B Y, *et al*. Morpheme-based derivation of bipolar semantic orientation of Chinese words[C]// Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004: 1008 - 1014.

[7] ZHU Y, MIN J, ZHOU Y, *et al*. Semantic orientation computing based on HowNet[J]. Journal of Chinese Information Processing, 2006, 20(1): 14 - 20.(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.)

(下转第 495 页)

果的态势决策能力。

3　结语

基于规则的态势感知主要采用证据理论实现规则融合。然而,由于证据理论自身在多源多证据合成时存在悖论问题,从而造成态势评估的结果并不准确。针对此问题,本文依据相异度矩阵和群体决策对证据理论进行改进,提出了一种基于改进证据理论的态势评估方法。该方法首先基于相异度计算对规则的重要性进行度量,并对少数组规则进行折扣,降低不可靠证据对评估结果的影响;然后,基于改进的证据理论进行规则融合;最后,基于最大信任进行态势决策。Matlab 仿真实验表明,本文方法能够避免利用 DS 证据理论进行态势融合过程中的悖论问题,且在态势融合效率和准确性等方面优于现有典型方法。

参考文献:

[1] CIMINO M G C A, LAZZERINI B, MARCELLONI F, *et al.* An adaptive rule-based approach for managing situation-awareness[J]. Expert Systems with Applications, 2012, 39(12): 10796 - 10811.

[2] ZENG F, LU M, ZHONG D. Using D-S evidence theory to evaluation of confidence in safety case[J]. Journal of Theoretical and Applied Information Technology, 2013, 47(1): 184 - 189.

[3] DEZERT J, WANG P, TCHAMOVA A. On the validity of Dempster-Shafer theory[C]// Proceedings of the 15th International Conference on Information Fusion. Piscataway: IEEE, 2012: 655 - 660.

[4] MASELENO A, HASAN M. The Dempster-Shafer theory algorithm and its application to insect diseases detection[J]. International Journal of Advanced Science and Technology, 2013, 50(1): 111 - 120.

[5] CHEN X, ZHAO C, LI Y, *et al.* Multi-feature suitability analysis of matching area based on D-S theory[J]. Journal of Computer Applications, 2013, 33(6): 1665 - 1669. (陈雪凌, 赵春晖, 李耀军, 等. 基于 Dempster-Shafer 证据理论的匹配多特征适配性分析方法[J]. 计算机应用, 2013, 33(6): 1665 - 1669.)

[6] LAI J, WANG H, ZHENG F, *et al.* Network security situation element extraction method based on DSimC and EWDS[J]. Computer Science, 2010, 37(11): 64 - 69. (赖积保, 王慧强, 郑逢斌, 等. 基于 DSimC 和 EWDS 的网络安全态势要素提取方法[J]. 计算机科学, 2010, 37(11): 64 - 69.)

[7] JIANG Y, LIU Y, LIN W, *et al.* Rough sets and evidence theory-based method to combine decision rules[J]. Journal of System Simulation, 2008, 20(4): 951 - 955. (姜元春, 刘业政, 林文龙, 等. 基于粗糙集与证据理论的决策规则合成方法[J]. 系统仿真学报, 2008, 20(4): 951 - 955.)

[8] YAGER R R. On the fusion of imprecise uncertainty measures using belief structures[J]. Information Sciences, 2011, 181(15): 3199 - 3209.

[9] LIANG C, YE C, ZHANG E, *et al.* An evidence combination method based on consistence of conflict[J]. Chinese Journal of Management Science, 2010, 18(4): 152 - 156. (梁昌勇, 叶春森, 张恩桥. 一种基于一致性证据冲突的证据合成方法[J]. 中国管理科学, 2010, 18(4): 152 - 156.)

[10] ALI T, DUTTA P, BORUAH H. A new combination rule for conflict problem of Dempster-Shafer evidence theory[J]. International Journal of Energy, Information and Communications, 2012, 3(1): 35 - 40.

[11] LEUNG Y, JI N, MA J. An integrated information fusion approach based on the theory of evidence and group decision-making[J]. Information Fusion, 2012, 8(2): 1 - 13.

[12] HU C, SI X, ZHOU Z, *et al.* An improved D-S algorithm under the new measure criteria of evidence conflict[J]. Acta Electronica Sinica, 2009, 37(7): 1578 - 1583. (胡昌华, 司小胜, 周志杰, 等. 新的证据冲突衡量标准下的 D-S 改进算法[J]. 电子学报, 2009, 37(7): 1578 - 1583.)

[13] CHAO F, YANG S. Group consensus based on evidential reasoning approach using interval-valued belief structures[J]. Knowledge-Based Systems, 2012, 35(1): 201 - 210.

[14] PASHA E, VMOSTAFAEI H R, KHALAJC M, *et al.* Fault diagnosis of engine using information fusion based on Dempster-Shafer theory[J]. Journal of Basic and Applied Scientific Research, 2012, 2(2): 1078 - 1085.

[15] YANG F, WANG X. Combination of conflict for D-S evidence theory[M]. Beijing: National Defense Industry Press, 2010. (杨风暴, 王肖霞. D-S 证据理论的冲突证据合成方法[M]. 北京: 国防工业出版社, 2010.)

(上接第 485 页)

[8] WIEBE J, BREUCE R, BELL M, *et al.* A corpus study of evaluative and speculative language[C]// Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue. Stroudsburg: Association for Computational Linguistics, 2001, 16: 1 - 10.

[9] WILSON T, WIEBE J, HOFFMANN P. Recognizing contextual polarity in phrase-level sentiment analysis[C]// Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2005: 347 - 354.

[10] TAKAMURA H, INUI T. Latent variables models for semantic orientation of phrases[C]// Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Tokyo: Fuji Press, 2006: 201 - 208.

[11] MATSUNOTO S, TAKAMURA H, OKUMURA M. Sentiment classification using word sub-sequences and dependency sub-trees [C]// Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2005: 301 - 310.

[12] KU L W, LIANG Y T, CHEN H H. Opinion extraction, summarization and tracking in news and blog corpora[C]// Proceedings of the 2006 AAAI Symposium on Computational Approaches to Analysing Weblogs. Menlo Park: AAAI Press, 2006: 100 - 107.

[13] HAO L. Study of modern Chinese negative adverb[D]. Beijing: Capital Normal University, 2003. (郝雷红. 现代汉语否定副词研究[D]. 北京:首都师范大学, 2003.)

[14] LIN H, GUO S. On the characteristics, range and classification of adverbs of degree[J]. Journal of Shanxi University: Philosophy and Social Sciences, 2003, 26(2): 71 - 74. (蔺璜, 郭姝慧. 程度副词的特点范围与分类[J]. 山西大学学报:哲学社会科学版, 2003, 26(2): 71 - 74.)

基于语义规则的Web金融文本情感分析

作者： 吴江， 唐常杰， 李太勇， 崔亮， WU Jiang， TANG Changjie， LI Taiyong， CUI Liang

作者单位： 吴江,李太勇,WU Jiang,LI Taiyong(西南财经大学经济信息工程学院,成都,610074)， 唐常杰,TANG Changjie(四川大学计算机学院,成都,610064)， 崔亮,CUI Liang(西南财经大学统计学院,成都,610074)

刊名： 计算机应用 ISTIC PKU

英文刊名： Journal of Computer Applications

年，卷(期)： 2014,34(2)

参考文献(14条)

1. 李国林 基于语义分析的Web金融文本信息情感计算 2012

2. 赵妍妍;秦兵;刘挺 文本情感分析综述 2010(8)

3. PANG B;LEE L;VAITHYANATHAN S Thumbs up? Sentiment classification using machine learning techniques 2002

4. CUI H;MITAL V;DATAR M Comparative experiments on sentiment classification for online product reviews 2006

5. TURNEY P D Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews 2002

6. YUEN R W M;CHAN T Y W;LAI T B Y Morpheme-based derivation of bipolar semantic orientation of Chinese words 2004

7. 朱嫣岚;闵锦;周雅倩 基于How-Net的词汇语义倾向计算 2006(1)

8. WIEBE J;BREUCE R;BELL M A corpus study of evaluative and speculative language 2001

9. WILSON T;WIEBE J;HOFFMANN P Recognizing contextual polarity in phrase-level sentiment analysis 2005

10. TAKAMURA H;INUI T Latent variables models for semantic orientation of phrases 2006

11. MATSUNOTO S;TAKAMURA H;OKUMURA M Sentiment classification using word sub-sequences and dependency sub-trees 2005

12. KU L W;LIANG Y T;CHEN H H Opinion extraction,summarization and tracking in news and blog corpora 2006

13. 郝雷红 现代汉语否定副词研究 2003

14. 蔺璜;郭姝慧 程度副词的特点范围与分类 2003(2)