ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, Supplement, December 2010, pp.349−362

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向股票新闻的情感分类方法*

高 旸[1], 周 莉[1], 张 勇[1,2,3+], 邢春晓[1,2,3], 孙一钢[4], 朱先忠[4]

[1](清华大学 计算机科学与技术系,北京 100084)

[2](清华大学 信息技术研究院,北京 100084)

[3](清华大学 清华信息科学与技术国家实验室(筹),北京 100084)

[4](国家图书馆,北京 100084)

Sentiment Classification for Stock News

GAO Yang[1], ZHOU Li[1], ZHANG Yong[1,2,3+], XING Chun-Xiao[1,2,3], SUN Yi-Gang[4], ZHU Xian-Zhong[4]

[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2](Research Institute of Information Technology, Tsinghua University, Beijing 100084, China)

[3](Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China)

[4](National Library of China, Beijing 100084, China)

+ Corresponding author: E-mail: zhangyong05@tsinghua.edu.cn, http://wiki.riit.tsinghua.edu.cn/west/ZhangYong/CV?action=print

Gao Y, Zhou L, Zhang Y, Xing CX, Sun YG, Zhu XZ. Sentiment classification for stock news. *Journal of Software*, 2010,21(Suppl.):349−362. http://www.jos.org.cn/1000-9825/10036.htm

Abstract: Web news articles play an important role in stock market. Sentiment classification of news articles can help the investors make investment decisions more efficiently. This paper implements an approach of Chinese new words detection by using *N*-gram model and applied the result for Chinese word segmentation and sentiment classification. Appraisal theory is introduced into sentiment analysis and Naïve Bayes, *K*-nearest Neighbor and Support Vector Machine are used as classification algorithms. This method is used for a Chinese stock news data set. The best accuracy reaches 82.9% in all experiments. Additionally, it develops a prototype system to demonstrate this work.

Key words: Chinese new word detection; *N*-gram model; sentiment classification

摘 要: 互联网新闻资讯对证券市场和投资者有举足轻重的影响,新闻进行情感分类后再展示给用户,可以帮助投资者迅速做出投资决定.从文本分类的基本方法出发,实现了基于 *N*-gram 统计模型的新词发现方法,并将所得结果用于构建中文分词词典和情感词典.同时引入评价理论,并用朴素贝叶斯、*K* 近邻和支持向量机 3 种方法进行股票新闻标题的情感分类实验.所用实验数据来自 2009 年“新浪财经”共计 23 万余条的新闻标题,结果表明二分类的准确率最高可达 82.9%.此外,还实现了一个原型系统用于展示股票新闻的分类结果.

关键词: 中文新词发现;*N*-gram 模型;情感分类

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2009AA01Z143 (国家高技术研究发展计划(863)); the Research Foundation of the National Railway Ministry of China under Grant No.20091111068 (铁道部研究基金)

Received 2010-07-01; Accepted 2010-12-10

随着经济的发展和人民生活水平的提高,当今社会人们通过购买股票进行投资已逐渐为大势所趋,如何准确地购买股票成为投资者非常关心的问题.炒股是一种实时性很强的行为,因为股票的价格、涨跌、买卖盘等信息每时每刻都在发生变化.这些信息不仅真实地反应了股市的走向,而且会对股民的投资决策起到关键性影响[1].对广大股票投资者而言,如果看到某些板块的正面新闻比较多,那么投资相应板块股票的意愿就会比较强烈.反之,如果看到的负面新闻比较多,那么就不太愿意投资甚至抛售相应板块的股票.

因此,新闻观点的正负倾向或情感倾向,在很大程度上会影响投资者的投资行为.如今,股民往往通过网络获取证券市场有关的各种资讯.然而面对网上海量的信息,传统基于手工的信息处理模式显得捉襟见肘,甚至不能满足需求.这就对如何快速而准确地发现有用信息提出了一定的要求.

文本分类是一种处理大规模文本数据的技术,可以在较大程度上解决信息繁杂的问题.文本分类是对给定的文档,按照其文本内容,将其分配为事先规定若干个文本类别中的一类或多类.传统的文本分类方法大多是基于文本外延的,其基本思想是用特征向量表示文档,通过计算特征向量的相似度达到文本分类的目的.这种方法是根据没有意义的字符串处理文本,虽然简单却没有从实际意义出发,可能对分类效果造成一定的影响[2,3].由于新闻资讯一般都包含积极或消极的倾向以及个人观点等信息,所以可以在文本分类过程中考虑词本身的意义,从表达的情感角度区分文本[4].

现有的证券工具一般只给出了股票的相关信息,却没有对信息的情感倾向给出提示.如果能通过对股票新闻进行情感分类,挖掘和分析文本中的立场、观点、看法、情绪等主观信息,对文本的情感倾向做出类别判断,为用户给出的新闻的正负性提示,那么可以帮助投资者迅速做出投资决定.

本文第 1 节介绍相关研究工作.第 2 节~第 4 节介绍具体的研究内容、方法以及相关实验.第 5 节介绍实现的原型系统,用于展示股票新闻的分类结果.第 6 节总结全文,并对今后的研究方向提供意见.

1 相关研究

1.1 有效市场理论

有效市场理论(efficient markets hypothesis,简称 EMH),又被称为有效市场假说或有效市场假设,始于美国芝加哥大学著名教授 Eugene Fama 在 1965 年发表在《商业学刊》的一篇名为《股票市场价格走势》的论文,而后 Eugene Fama 在 1970 年发表于《金融》的论文《有效资本市场:理论与实践研究回顾》中深化并提出的.有效市场理论假定所有公开的信息都会反映到市场价格之中,相关的信息如果不受扭曲且在证券价格中得到充分反映,市场就是有效的[5].既然证券价格能充分反映一切可获得的信息,那么,可获得的相关信息就成为价格能否有效的决定因素.

按照可获得的信息分类的不同,有效市场理论在有效率的资本市场分为以下 3 种表现形态:弱式有效市场、半强式有效市场和强式有效市场.从中国的现实情况来看,国内多数学者支持中国股市是弱式有效的[6].

在弱有效市场中,信息发布后需要一段时间才能反应到股价中,也就是说信息发布后,股票会经过一段时间才能调整到合适的价位.因此不能忽视股票新闻对于股市的影响,新闻的数量以及内容的倾向性在很大程度上也会左右投资者的购买行为.例如,“国务院将于 2010 年 4 月 24 日把印花税税率由 3‰下调为 1‰”的消息一出,沪指暴涨 304 点,千余个股涨停;又如,在 2010 年年初的“两会”上,政府工作报告提出要发展“低碳经济”,之后“新能源板块”引来利好,逐渐走强.因此研究股市新闻的倾向性,对辅助投资者做出投资决策具有一定的实用意义.

1.2 中文分词

在自然语言中,词语一般是最小的具有独立语法或语义的单位.因此,对词的处理自然是文本处理的基础[4].在西方语言中,词与词之间由空格分开,一目了然.但对于中文,最小的语素是单个文字,词与词之间没有明显的界限区别,所以如何分词是中文文本处理首先要解决的一项关键技术.中文分词的任务就是让计算机能按正确的意思识别不同的词.

1.2.1 新词发现

目前主要的分词方法分为 3 类:基于字典或词库的匹配分词方法,基于统计的分词方法和基于理解的分词方法,而前两种方法目前使用比较广泛.对于第 1 种方法,词典的好坏在很大程度上影响分词结果的准确率,尤其是对某些专业领域的文本.对于第 2 种方法,其实分词的过程也就是新词发现的过程.对于第 3 种方法,语法分词等过程也需要及时发现新词.

在中文分词领域,新词大致分为 4 种:1) 缩略词,如中石油、国投、中金等;2) 专有名词,如股份有限公司、证券投资基金等;3) 派生词,如黑马股、领涨、利空等;4) 复合词,如冲高回落、分红派息等[7,8].

目前,新词发现通常有以下两种做法[7,8]:1) 基于规则的方法,即由专家归纳出某些新词的构成规则或特点,猜测可能的新词并给出置信度,之后再做进一步的鉴定;2) 基于统计的方法,即利用一些统计策略和相关度,寻找那些出现可能性最大词,该方法适用于发现较短的新词.

因为本文针对的对象主要是股票新闻标题,其中的词多以简短的形式出现,所以可以采用基于统计的方法,例如 *N*-gram 语言模型作为新词发现的方法.

1.2.2 *N*-gram 语言模型

在自然语言中,一个句子可以由任意的字符串组成,但它们出现的概率 $P(s)$有很大差别.例如:s_1="A 股股民失望告别 2008 年",s_2="2008 年告别失望股民 A 股",显然前者作为一句话出现的概率更大,即 $P(s_1)>P(s_2)$.

对于给定的字符串,$P(s)$通常是未知的.假设用 $W=w_1w_2\ldots w_n$ 表示文本中的一个字符串序列,则 W 在文本中出现的概率 $P(w)$为[9]

$$P(w)=\prod_{i=1}^{n}P(w_i \mid w_1w_2...w_{i-1}) \tag{1}$$

通常为了简化模型和便于计算,一般只考虑 $n-1$ 次构成的历史,即认为任意一个词出现的概率只与它前面 $n-1$ 个词有关,这时的语言模型叫做 *N*-gram 语言模型,也被称为一阶马尔科夫链.常用最大似然估计(maximum likelihood estimation)的参数估计方法计算 $P(w)$[9−12]:

$$P_{MLE}(w_n \mid w_1w_2...w_{n-1})=\frac{C(w_1w_2...w_n)}{C(w_1w_2...w_{n-1})} \tag{2}$$

其中,$C(w_1w_2\ldots w_n)$表示该字符串在文本中出现的次数.

在实际应用中,最大似然估计无法避免由于训练样本不足而导致的数据稀疏问题.Zipf 定律描述了词频以及词在词频表中的位置之间的关系,说明在自然语言中,常用词只占很少的一部分,大部分词都是低频词[13].最大似然估计在计算时对样本中未出现事件的概率按零概率处理,而 Zipf 定律表明,由于不可能存在一个足够大的训练预料,包含了所有的词序列,因此数据稀疏问题不可避免.

1.2.3 数据平滑方法

由于存在数据稀疏问题,为了解决最大似然估计的零概率问题,需要引入平滑技术.本文主要采用以下几种平滑算法,包括 Add-one、Add-delta、留存估计和删除估计[11,14].

Add-one 平滑方法假定任何一个 *n*-gram 的统计次数是其在训练语料中实际出现次数加 1,认为那些未出现过的 *n*-gram 也在训练预料中出现了一次,即 $C(n\text{-}gram)_{new}=C(n\text{-}gram)_{old}+1$.采用 Add-one 平滑方法,公式(2)的参数估计结果可以表示为

$$P_{Add\text{-}one}(w_1w_2...w_n)=\frac{C(w_1w_2...w_n)+1}{C(w_1w_2...w_{n-1})+N} \tag{3}$$

如果有大量的 *n*-gram 没有出现在训练语料中,用 Add-one 方法平滑后这些没有出现的 *n*-gram 将会在整个概率分布中占据较大比例,这是不太合理的.一种改进方法是出现次数不加 1,而是加上一个小于 1 的数,即用 Δ 替换公式(3)中的 1,用 $\Delta\cdot N$ 替换 N,这就是 Add-delta 平滑方法.

留存估计(held-out estimation)的基本思想是,把全体语料分为训练语料和留存语料两个部分,其中训练语料作为最初的频率估计,而留存语料用于改善最初的频率估计.具体做法是首先对于每个 *n*-gram,分别计算其在训

练语料和留存语料中出现的频率,即 $C_{tr}(w_1w_2...w_n)$ 和 $C_{ho}(w_1w_2...w_n)$. 并且设 T 是留存语料中所有的 n-gram 个数,用 r 表示某个 n-gram 在训练语料中出现的频率,即 $r=C_{tr}(w_1w_2...w_n)$, 同时设 Nr 表示在训练语料中出现了 r 次的不同的 n-gram 的个数,T_r 表示所有在训练语料中出现了 r 次的 n-gram 在留存语料中出现的频率之和,即 $T_r=\sum_{\{(w_1w_2...w_n)|C_{tr}(w_1w_2...w_n)\}}C_{ho}(w_1w_2...w_n)$. 因此,采用留存估计方法的参数估计结果是

$$P_{ho}(w_1w_2...w_n)=\frac{T_r}{T}\times\frac{1}{N_r} \tag{4}$$

删除估计(deleted estimation)是把训练语料分为两部分,分别以其中一部分做训练语料和留存语料,按留存估计方法计算一次后,训练语料和留存语料互换后,再按留存估计方法计算一次,最后求两者的加权平均,即

$$P_{del}(w_1w_2...w_n)=\frac{T_r^{01}+T_r^{10}}{N(N_r^0+N_r^1)} \tag{5}$$

1.3 中文文本表示

文本经过分词等预处理后,还需要对其进行数学建模,从而方便计算机进行计算.目前最常用的是向量空间模型(vector space model,简称 VSM)[15].VSM 的基本方法是用特征空间中一组正交的特征词向量表示文本,其中每个不同的词条就作为特征空间中独立的一个维度.

在向量空间模型中,不同的特征对文本的区分力度是不一样的,需要对不同特征进行加权处理,其目的是提高区分力度强的特征的权重,而减弱区分力度弱的特征的权重.

布尔权重是最简单的加权方法,用公式表示为

$$w_{kj}=w(t_k,d_j)=\begin{cases}0, & \#(t_k,d_j)=0\\ 1, & \#(t_k,d_j)>0\end{cases} \tag{6}$$

其中,$\#(t_k,d_j)$表示特征词 t_k 的在文本 d_j 中出现的频率.

布尔权重不能区分不同特征词之间的重要性.在文本分类中,往往认为出现次数多的词比出现次数少的词对分类有更大的作用,所以出现次数不一样的特征词的权重应该是不一样的.词频权重是直接以特征词在文档中出现的频率作为权重.

考虑到文本的长度可能差别很大,为了让权重落入[0,1]的区间内,可以用余弦标准化的方法对词频权重做归一化处理,公式如下[2,16]:

$$w_{kj}=w(t_k,d_j)=\frac{\#(t_s,d_j)}{\sqrt{\sum_{s=1}^{|T|}(\#(t_s,d_j))^2}} \tag{7}$$

1.4 文本情感分类

文本分类可以定义为每个文档 d_j 和类别 c_i 组成的元素对$\langle d_j,c_i\rangle\in D\times C$ 确定一个布尔值,其中 $D=\{d_1,\ldots,d_{|D|}\}$ 表示文档的集合,$C=\{c_1,\ldots,c_{|c|}\}$表示预先定义的类别集合.如果$\langle d_j,c_i\rangle$的值是 T(表示真),表明文档 d_j 属于 c_i 类;如果$\langle d_j,c_i\rangle$的值是 F(表示假),表明文档 d_j 不属于 c_i 类[2].

文本分类的过程一般分为一下几个步骤:1) 文本预处理:包括分词或词干抽取、去停用词等;2) 文本表示:包括特征选择等,将文本处理成计算机可以“理解”的形式;3) 文本分类方法:使用合适的方法构建分类器,对文本分类;4) 分类结果的评价:通过测试样例,评价分类器的结果和性能[2].

文本分类的方法有很多,其中朴素贝叶斯、k 近邻(k-NN)和支持向量机(SVM)等方法比较成熟[2,16,17].朴素贝叶斯的基本思想是先根据贝叶斯条件概率公式,计算在知道文档特征向量的条件下该文档属于某一类别的条件概率;然后根据极大似然原理可知,该文档应该属于使后验概率取最大值的那一类[2,16,17].k-NN 方法的基本原理是,对于一篇待分类的文档 d,在文档集合 D 中找到与其最相似的 k 篇文档,然后根据这最相似的 k 篇文档的类别来判断待分类的文档 d 的类别[2,16].支持向量机应用于二分类问题,基本思想是寻找一个最优超平面(曲面)作为决策平面,使得两类样本之间的分类距离达到最大[2,17].

文本情感分类主要研究如何对文本所表达的情感等主观内容进行分类,其目标是判断给定文本片段所体

现的说话者的情感倾向,并且根据作者的情感偏向给出其正面或者负面的评价[18].文本情感分类不同于传统的基于主题自动文本分类.文本分类是指按照预先定义的类别来决定一篇文本的归属的过程,其关注的焦点是文本的主题.而情感分类主要通过挖掘和分析文本中的立场、观点、看法、情绪等主观信息,来判别自然语言文字中表达的观点、喜好以及与感受和态度等相关的信息,有时候还需要考察词的含义和语法结构等.文本情感分类是进行倾向判断的很好的方法,在个性化推荐、个性化观点检索、用户兴趣挖掘、信息过滤、邮件过滤、社会舆论分析等方面得到很好的应用.

目前对于文本情感分类的主要研究内容包括文章的情感倾向识别(即粗粒度的识别出文本表现的情感是积极还是消极情感)、词语的语义倾向性识别、文本的主观性分析、文本的情感极性分类、观点提取等,这些研究工作可以归纳为:1) 主客观分类,即将文档按照类型和风格的不同,划分为主观和客观两类;2) 词的极性分类,即将文档划分为积极和消极两类,或者按照强度进行更细致的分类;3) 情感分类,即将文档按照作者感情的喜怒哀乐分类[4,19].

文本情感分类需要有效的语言特征来表征文本的情感特,如何有效地表示和获取语言特征显得尤为重要.因此在本文中需要构建一个合适的情感词典,这也是情感分类中需要重点解决的问题.

2 中文股票新闻分词

2.1 ICTCLAS

中文文本预处理的第一步是分词.本文所涉及的实际的分词过程,采用 ICTCLAS[20]作为分词工具.该系统的主要功能包括中文分词,词性标注,同时支持用户词典.ICTCLAS 对一般文本的分类准确率可达 98.45%,但由于缺少相关的词典,不能直接用于股票领域.因此可以采用第 1.2 节所述方法,构建股票领域的分词词典.

2.2 数据集

通常股票新闻标题带有情感倾向,可以反映出股票的涨跌趋势.本文选取 2009 年"新浪财经"股票新闻的标题,共计 233 282 条作为原始语料数据集,用于构建股票词典.

2.3 子串过滤

首先在原始语料上建立 *N*-gram 模型.理论上,*n* 较大时,提供的语境信息较多,语境更具区别性,但计算量也较大,参数估计较不可靠;而 *n* 较小时提供的语境信息较少,语境区别性较小,但计算量也较小,参数估计较可靠.因此,在实际应用中需要合理地选择 *n* 的大小.如果一个长度为 *L* 的字符串,建立模型后可得 *L*–*n*+1 个子串.考虑到新闻标题中的词多以简称、缩略语等短语形式出现,因此 *n* 的取值范围是 $2 \leqslant n \leqslant 6$,即 *n*=2,3,4,5,6.同时采用的平滑方法包括 Add-one、Add-delta(实验中取 *delta*=0.5)、留存估计和删除估计.

通过实验可以发现,Add-delta 和留存估计的效果较好,但采用这两种方法的结果仍不能直接利用.这是因为在实际应用中存在很多常用词,组成这些词的若干子串往往只能在这些词中出现,很难单独作为一个词出现.例如,"股份有限公司"是一个父串,"份有限公"则是它的一个子串.通常即使一些子串不能单独成词,它们的出现的频率却与其父串基本相同.在一般的统计语言模型中,这样的子串和父串的参数估计结果很接近,但子串往往是没有用的,因而成为了干扰项,需要对它们进行过滤.

观察发现,那些没有用的子串与其父串的概率之差往往很小.所以过滤无用子串的基本思想是,对于一个字符串的所有父串,如果该字符串与其父串的概率之差小于某个值,并且该字符串与其父串的长度之差小于某个值时,可以将该子串过滤.本文正是采用这种方法进行过滤,其中概率差值取 0.000 1,而长度差值取 3.

2.4 分词结果

原始语料经过上述处理过程,取出高频词即可以作为股票领域的新词,其结果可以分为 4 种(见表 1),即用来构建股票分词词典:

表 1 股票领域新词发现结果举例

专有名词	证券投资基金、黑马股、权重股、中小板
缩略词	中金、国投、沪指、深指、股改
情感词	谨慎推荐、利空、利多、看高
描述股市行为	冲高回落、定向增发、分红派息、震荡上行、减持

股票分词词典的另一部分由股票的名称、代码、主营业务、相应的公司和板块名称等组成.最终实验所用的用户词典包含 21 262 个词,当然词典条目也可以由用户手工添加.ICTCLAS 系统根据该词典对原始文本进行分词,效果比没有使用用户词典的分词效果好很多,见表 2.

表 2 股票新闻分词结果

使用前	权/n 重/a 股/n 萎靡不振/v 两/m 市/n 早/a 盘/qv 冲/v 高/a 回落/v
使用后	权重股/n 萎靡不振/v 两市/n 早盘/n 冲高回落/n

2.5 停用词过滤

为方便文本表示及后续处理,还需要对分词后的结果进行去停用词的处理.停用词是指一些出现频率比较高,却没有太多实际意义的词,对文本处理几乎没有实用价值[21].去除停用词对于提高文本处理的效率是非常必要的,股票新闻文本中的停用词表主要有两种:第 1 种是介词、冠词、助词、连词和标点符号;第 2 种是股票新闻标题前提示性的词,如快讯、锐点、大盘、市场等.

去除停用词的方法可以结合分词结果的标注信息和停用词表.

3 中文股票新闻文本表示

对于情感分类可以引入评价理论(appraisal theory),即通过从文本中提取形容词及其修饰语构成的短语作为特征,进行语义倾向分析,这种形容词短语被称为评价组(appraisal group),实验表明,利用评价组作为特征集能够提高情感分类的精确度[22].Whitelaw 根据 Martin 的评价理论,为评价设置了 4 个属性:态度(attitude)、倾向(orientation)、等级(graduation)和极性(polarity).

中文股票新闻文本表示可以利用类似评价组的方法,但有所不同的是,提取的不仅是形容词短语,还应包括带有情感色彩的形容词、动词以及修饰词,将这些词统称为情感词.因此,可以初步将股票情感词划分为正面、负面、程度、否定词及不确定词五类:正面词就是描述股票价格上涨,股票上市公司业绩好等词汇;负面词则是描述股票价格下跌以及上市公司业绩差等词汇;程度词是指描述正负程度的;另外由于一个正面或负面词之前加上否定词后就是相反的意思,因此需要考虑否定词的影响;最后,不确定词决定着正面及负面的可信度.这 5 种情感词举例见表 3.

表 3 股票情感词

正面词	负面词	程度词	否定词	不确定词
佳	减少	明显	非	不确定
优	降	大幅	不	可能
增	锐减	小幅	并非	或
好	补跌	快速	没有	或将
增长	下降	快	没	或是
盈	亏损	适当	别	是否
涨	赔	强烈	无	有望
补涨	亏	更	未	疑
赚	跌停	最	未能	约
涨停	减持	相当	没	缘何
飙升盈利	降低	谨慎	难	能否

根据上面的方法,可以为股票设倾向性、等级、极性、确定性这 4 个属性.倾向性表示评价是正面还是负

面;等级表示强度,比如“大大”、“快速”可以加强强度,“小幅”则减弱强度;极性表示是否包含否定词;确定性表示是否包含不确定词.具体结构如图 1 所示.

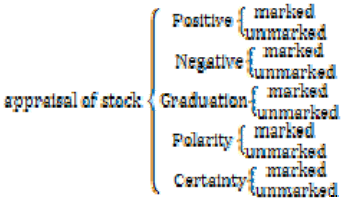

图 1 股票评价组

按照上述方法进行中文股票新闻文本表示,可以建立一个相应的情感词典.情感词典中的词主要通过 3 种途径组成:一部分由原始新闻标题新词发现的结果结合人工判断挑选;另一部分根据已有股票领域词典,参考领域专家意见,直接由手工构建;还有一部分则来自“知网 HowNet”的情感词库.

4 股票新闻情感分类

4.1 数据集

分类实验的数据来源和第 2.2 节一样,也选取自新浪财经网 2009 年全年的股票新闻标题,并且人工标记其中 15 830 条新闻的情感倾向,作为分类实验的数据集,其中情感倾向统计结果见表 4.

表 4 股票新闻情感倾向统计

总计	样本倾向		
	Positive	Neutral	Negative
15 830	9 060	3 757	3 013

4.2 实验工具

文本情感分类所用工具是新西兰开发的一个的数据挖掘工作平台——Weka,其中集成了一些机器学习方法,可以对数据进行预处理和分类、聚类等.

4.3 评价方法

文本情感分类中还有一项重要任务,就是对分类方法进行评估.目前,一般采用的评价方法是准确率(precision)和召回率(recall).总结一个分类系统的分类结果,可以得到表 5[2].

表 5 分类结果统计

类别的集合为 $C=\{c_1,\ldots,c_{\|C\|}\}$		实际情况	
		属于该类别文档	不属于该类别文档
分类结果	属于该类别文档	$TP=\sum_{i=1}^{\|C\|}TP_i$	$FP=\sum_{i=1}^{\|C\|}FP_i$
	不属于该类别文档	$TN=\sum_{i=1}^{\|C\|}TN_i$	$FN=\sum_{i=1}^{\|C\|}FN_i$

此时,准确率和召回率分别定义为[2]

$$Precision=\frac{TP}{TP+TN} \tag{8}$$

$$Recall=\frac{TP}{TP+FP} \tag{9}$$

准确率和召回率分别从效果(effectiveness)和效率(efficiency)两方面考察分类结果.为了综合考虑两种因

素,可以得出一个新的指标 *F*-Measure,即

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

4.4 实验结果及分析

原始语料,经过文本预处理,并采用评价组方法进行文本特征表示,结合 Weka 工具,重点对 Naïve Bayes, *k*-NN 和 SVM 这 3 种方法进行实验.

首先使用标记的全部数据集,也就是 15 830 条新闻标题,分为 Positive,Negative 和 Neutral 这 3 类,利用布尔及词频特征加权法,使用 Naïve Bayes,*k*-NN 和 SVM 这 3 种方法,实验结果的 *F*-Measure 值见表 6.

表 6 三分类结果

特征加权	算法		
	Naïve Bayes	*k*-NN	SVM
布尔	0.632	0.630	0.627
词频	0.553	0.650	0.649

将 Neutral 的新闻标题去除,也就是只保留具有情感的新闻标题.此时是一个二分类问题,重复上面的实验过程,结果见表 7.

表 7 二分类结果

特征加权	算法		
	Naïve Bayes	*k*-NN	SVM
布尔	0.801	0.800	0.796
词频	0.829	0.828	0.825

由以上两个实验结果可以看出,去除不含情感的数据,能够大大提高分类精度.此时采用词频特征加权法时,采用 Naïve Bayes 分类方法效果最好.

第 3 个实验在第 2 个实验的基础上,主要考察不同数据规模对分类结果的影响.可以随机选取全部训练集的 100%,1/2,1/3,1/4,1/5 分别进行实验,选择布尔权重,利用 Naïve Bayes,*k*-NN 和 SVM 这 3 种方法,实验结果的 *F*-Measure 值见表 8.

表 8 不同规模的分类结果(布尔权重)

规模	算法		
	Naïve Bayes	*k*-NN	SVM
100%	0.801	0.8	0.796
1/2	0.799	0.797	0.792
1/3	0.799	0.803	0.789
1/4	0.797	0.795	0.798
1/5	0.809	0.803	0.799

采用词频加权,重复上面的实验过程,结果见表 9.

表 9 不同规模的分类结果(词频权重)

规模	算法		
	Naïve Bayes	*k*-NN	SVM
100%	0.829	0.828	0.825
1/2	0.826	0.821	0.819
1/3	0.824	0.821	0.819
1/4	0.825	0.819	0.828
1/5	0.835	0.830	0.827

由以上两个实验结果可以看出,总体上来看,训练集越大,分类算法的精度有提高的趋势,也就是说足够大的训练集对于提高分类精度有很大的作用.同时,一般情况下词频权重的分类结果要好于布尔权重.

本体可以用来描述某领域中的概念以及概念之间的关系,通过构建股票本体,可以将所有的新闻自动区分

为不同的股票或板块的相关新闻.对新闻区分后,可以进行热点板块或股票的研究.

如果假设一只股票的涨幅或一个板块内所有股票的平均涨幅高于一个具体的数值,如 n%,那么可以认为与这支股票相关的新闻都是正面的;反之,如果跌幅低于某具体数值,如–n%,那么,可以认为与这支股票相关的新闻都是负面的.

因此,可以取 n=3,4,5,…,9,选择词频权重,利用 Naïve Bayes,k-NN 和 SVM 这 3 种方法,对股票新闻进行分类实验,结果的 F-Measure 值见表 10.

表 10 不同 n 取值的个股新闻分类结果

N	算法		
	Naïve Bayes	KNN	SVM
3	0.562	0.587	0.545
4	0.584	0.610	0.575
5	0.607	0.620	0.606
6	0.635	0.647	0.638
7	0.682	0.671	0.674
8	0.725	0.713	0.723
9	0.776	0.760	0.781

由该实验结果可以看出,各个算法的效果都是随着 n 的增大而提高,横向比较看来,随着 n 取值的不断增大,3 种算法的效果区别不是很大,但 Naïve Bayes 算法快,KNN 其次,SVM 最慢,因此想比较而言,还是 Naïve Bayes 比较适合应用于实际应用.因为从上图中可以看到,当 n=9 的时候各类算法的效果最好,因此可以采用 n=9 进行接下来的实验.

表 11 个股新闻情感分类 3 种算法对比

特征加权	算法		
	Naïve Bayes	KNN	SVM
布尔词频	0.593	0.597 (K=3)	0.594
绝对词频	0.776	0.779 (K=5)	0.781
归一化词频	0.783	0.786 (K=5)	0.781

表 11 的结果表明,情感词与股票实际交易涨幅具有一定的相关性,利用涨幅标记的数据集,进行情感分类最好的 F-Measure 值能够达到 78.6%,这就表明,利用情感倾向来判断股票是否是热点具有一定的可行性.

对与表 10 所示第 5 个实验,如果对板块新闻重复该实验,则其结果的 F-Measure 值见表 12.

表 12 不同 n 取值的板块新闻分类结果

n	算法		
	Naïve Bayes	KNN	SVM
2	0.460	0.520	0.509
3	0.427	0.508	0.488
4	0.380	0.565	0.529
5	0.673	0.668	0.675

由该实验结果与表 10 的结果类似.另外从上图中可以看到,当 n=5 的时候各类算法的效果最好,因此可以采用 n=5 重复表 11 所示第 6 个实验,结果见表 13.

表 13 板块新闻情感分类 3 种算法对比

特征加权	算法		
	Naïve Bayes	KNN	SVM
布尔词频	0.674	0.674	0.674
绝对词频	0.673	0.706	0.675
归一化词频	0.665	0.689	0.674

表 13 的结果表明,情感词与板块实际交易涨幅具有一定的相关性,利用涨幅标记的数据集,进行情感分类最好的 F-Measure 值能够达到 70.6%,这也表明,利用情感倾向来判断股票是否是热点具有一定的可行性,但效果还不是很好,需要进一步的研究.

5 原型系统

为了展示本文的工作成果,开发了原型系统.该系统主要包括 3 个模块:数据预处理模块,主要包括对新浪财经中证券相关新闻的爬取和数据类型选取;中文文本分词模块,主要包括文本分词,文件分词和用户词典反馈等功能;我的财经模块,能够自动爬取股票新闻,实现个股新闻查询、热点排行统计,具体的查询使用了本体扩展查询,能够得到跟股票相关的各种新闻,包括个股、上市公司、所属板块新闻,热点排行包括热点股票以及热点板块排行,同时能够对相关新闻进行情感分类展示分类结果.

5.1 开发工具

开发环境:Windows,JDK 1.6,Jena 2.5.6

开发语言:Java,JSP,JS

Web 框架:Struts+Spring+Mule

Web 服务器:Tomcat 5.5.17

数据库:MySQL 5.0

辅助工具:

1) 中文文本分词工具:ICTCLAS2009 共享版
2) 文本分类工具:Weka3.7
3) 本体构建工具:Protégé3.4.1

5.2 数据预处理模块

该模块主要进行数据的准备和预处理工作,主要包括数据抓取和预选数据类型两个功能.

5.2.1 数据抓取

如图 2 所示,该模块实现了从新浪财经抓取新闻的功能,为用户提供按照时间批量抓取新闻的界面,用户可自由选择抓取的时间段,抓取后的新闻将批量插入数据库,抓取股票新闻的链接、标题、时间以及新闻类型存入数据库.

图 2 股票新闻爬取模块

5.2.2 预选数据类型

实际股票相关的新闻有很多种类,有关于个股的新闻,还有关于上市公司的新闻,还有行业相关的新闻,而且股票新闻具有很强的实时性,选取什么类型的数据、什么时间范围的数据对于实验的精度会有相应的影响,预选数据类型模块用于方便地生成不同类型数据集,如图 3 所示.

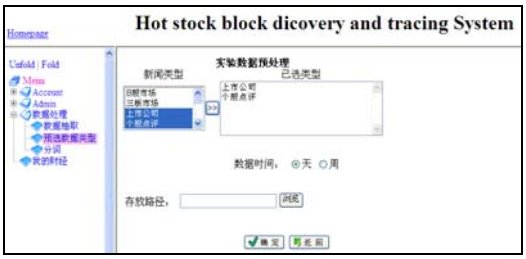

图 3　股票新闻类型选取模块

5.3　中文文本分词模块

如图 4 所示,分词模块主要由 3 部分组成:用户词典反馈、文本分词和文件分词.由于在股票新闻中经常出现一些“新词”,导致不能完全分词的现象.对此,在本模块的解决办法是为用户增加股票分词词典的反馈功能,包括添加单个词条和文本文件中的所有词条.通过该模块,用户可以向股票分词词典中添加条目,这些新加入的词将会直接影响分词结果.为了直接展示分词效果,还在页面中提供了一个文本框,用户可以输入带分词的文本,系统能够将分词结果显示出来.另外,为了方便处理实验数据集,还提供了为批量文件分词的功能.

图 4　分词模块

5.4　我的财经模块

“我的财经”模块主要展示热点发现功能.系统首先自动爬取当天的新浪财经新闻,对于爬取的新闻,经过文本预处理,利用股票本体新闻分发至个股及板块,并且还区分新闻的语义倾向,用户可以通过查询找到某只股票的相关新闻,不同倾向的新闻讲以不同颜色显示给用户.同时还能够进行热点排行,通过热点计算得到热点股票及热点板块显示给用户.关于处理新闻的后台流程如图 5 所示.

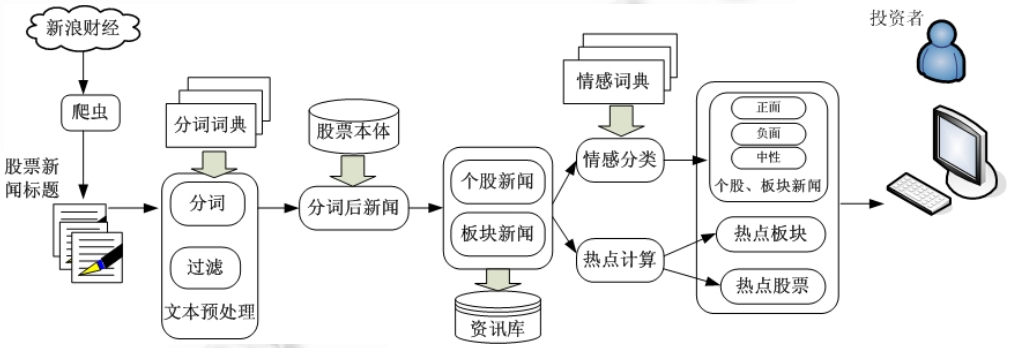

图 5　“我的财经”实现流程图

图 6 是该模块的初始化界面,系统能够自动爬取新浪财经新闻,显示当天的大盘日线图以及股票相关新闻,同时还能够进行热点排行,包括热点股票以及热点板块.

图 6　我的财经

对于每一只股票,可以根据该股票的代码或名称,查询到该股票,系统将该股票的实时交易信息显示给用户,并且同时返回当天该股票的相关信息,包括个股新闻、上市公司以及该股票所属的板块的信息,并且还能够区分新闻语义,将不同语义的新闻以不同的颜色显示出来.对于每一个板块,也有同样的功能.个股查询结果如图 7 所示.

图 7　个股查询结果

6 总结和展望

本文的工作主要是研究面向股票领域的文本情感分类的一些相关方法,包括:构建股票本体并实例化、爬取股票新闻并人工标记情感倾向、构建股票分词词典、采用评价组方法构建情感词典并用于情感分类、验证不同分类方法的分类结果、实现原型系统.

本文只是针对股票新闻文本,做了一些前期的探索工作,未来的还有很多值得关注的研究点.今后的工作主要可以针对以下几个方面:构建更详细的本体并提高应用范围;改进数据集质量;实现更完美的分词方案;建立更适合的特征选择方法;提供更全面的情感分类方法并考虑时间、地域等因素对新闻情感倾向的影响.

References:

[1] Huang TH, Li ZY. Stock prediction using Web news articles. Technical Report, Graduate Institute of Networking and Multimedia, Taiwan University, 2008.

[2] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002,34(1):1−47.

[3] Aas K, Eikvil L. Text categorisation: A survey. Technical Report, Oslo: Norwegian Computing Center, 1999.

[4] Chen B. Research on key problems in Web text sentiment classification [Ph.D. Thesis]. Beijing: Beijing University of Posts and Telecommunications, 2008.

[5] Wang C, Li N, Li XL, Liang X. The research on financial volatility with sentiment analysis. Journal of Chinese Information Processing, 2009,23(1):95−99.

[6] Zhang B, Li XM. An evolving market efficiency test on Chinese stock market. Economic Research Journal, 2003,1:54−61.

[7] Sun MS, Zou JY. A critical appraisal of the research on Chinese word segmentation. Contemporary Linguistics, 2001,3(1):22−32.

[8] Liao XT. Overview of Chinese new word detection. Technical Report, Harbin: Center of Information Retrieve, Harbin Institute of Technology, 2004.

[9] Brown PF, de Souza PV, Mercer RL, Pietra VJD, Lai JC. Class-Based *N*-gram models of natural language. Computational Linguistics, 1992,18(4):467−479.

[10] Heinrich G. Parameter estimation for text analysis. Technical Note, University of Leipzig, 2005.

[11] Zhai CX, Lafferty J. A study of smoothing methods for language models applied to information retrieval. ACM Trans. on Information System, 2002,22(2):179−214.

[12] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 1977,39(1):1−38.

[13] Adamic LA, Huberman BA. Zipf's law and the Internet. Glottometrics, 2002,3:143−150.

[14] Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. In: Proc. of the 34th Annual Meeting on Association for Computational Linguistics. 1996. 310−318.

[15] Salton G, Lest ME. Computer evaluation of indexing and text processing. Journal of the ACM, 1968,15(1):8−36.

[16] Sebastiani F. A tutorial on automated text categorization. In: Amandi A, Zunino A, eds. Proc. of the 1st Argentinian Symp. on Artificial Intelligence (ASAI'99). 1999. 7−35.

[17] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proc. of the ACL 2002 Conf. on Empirical Methods in Natural Language Processing. 2002. 79−86.

[18] Lai HY, Liu GS. Research on semantic orientation of Chinese texts based on topic correlation [MS. Thesis]. Shanghai: School of Information Security Engineering, Shanghai Jiaotong University, 2009.

[19] Ma YZ, Wang C. Research on phrase pattern based sentiment extraction for reviews [MS. Thesis]. Beijing: University of Posts and Telecommunication, 2008.

[20] Zhang HP, Yu HK, Liu Q. HHMM-Base Chinese lexical analyzer ICTCLAS. In: Proc. of the 2nd SIGHAN Workshop on Chinese. 2003. 184−187.

[21] Hua BL. Stop-Word processing technique in knowledge extraction. New Technology of Library and Information Service, 2007,2(8): 48−51.

[22] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis. In: Proc. of the 14th ACM Conf. on Information and Knowledge Management. 2005. 625−631.

高旸(1988－),男,陕西临潼人,硕士生,主要研究领域为数据库.

周莉(1985－),女,硕士生,主要研究领域为数据库.

张勇(1973－),男,博士,副研究员,主要研究领域为数据库,数字图书馆.

邢春晓(1967－),男,博士,研究员,博士生导师,主要研究领域为数据库,数字图书馆.

孙一钢(1961－),男,研究馆员,主要研究领域为数字图书馆,计算机应用.

朱先忠(1965－),男,高级工程师,主要研究领域为数字图书馆,网络新媒体展现技术.