文本倾向性分析用于金融市场波动率与金融信息相互关系的研究

王超，李楠，李欣丽，梁循

（北京大学计算机科学技术研究所，北京，100871）

{wangchao,linan,lixinli,liangxun}@icst.pku.edu.cn

摘 要：互联网金融信息对于金融市场的影响在当代已经越来越不可忽视。面对海量的信息，其中大部分为非结构化的文本数据，本论文结合目前已有的文本倾向性算法，把信息的褒贬值作为外部变量加入到针对股价波动率建立的时间序列模型中去，对金融市场的股价波动率进行预测。实验揭示出金融市场波动率与互联网上金融新闻的相关性，并且提出了一种有效的股市预测方法。

关键词：文本倾向性，波动率，SVM，金融预测，金融信息

Associating Financial Volatility and Financial Information with Sentiment Analysis

Chao Wang ,Nan Li, Xinli Li, Xun Liang

Institute of Computer Science and Technology, Peking University, Beijing, China, 100871

{wangchao,linan,lixinli,liangxun}@icst.pku.edu.cn

Abstract: Within stock markets, trading volume and asset price of a financial commodity or instrument are highly changeable and unpredictable. Financial information is one of the facts that impacts on the movements especially in this new information era nowadays. In this paper, we utilize semantic techniques to probe into the correlations between information sentiment and asset price volatility.

Keywords: sentiment analysis, volatility , support vector machine ,Financial forecast, financial information,

1. 介绍

随着信息通讯技术和互联网的发展，互联网金融信息对金融市场的影响已经越来越不容忽视。信息的数量和内容都在很大程度上左右着金融实践者们的行为，同时进一步影响着股市变化的趋势。在金融市场中，股价和交易量的波动率是股票市场行为的真实反映，它与金融风险有着密不可分的关系，而波动的变化趋势更是体现着金融风险的走势。由此可见，如果我们可以有效地借助互联网金融信息对股票市场的波动率进行分析和预测，其将带来的学术和工业价值都是非常明显的。

面对海量的互联网金融信息，其中大部分数据是非结构化的文本数据。如何利用现有的自然语言处理技术对其进行处理，对于寻求金融信息与金融市场波动的关系有着重大意义。

金融市场非常容易受到信息的影响，而互联网凭借实时性、丰富性以及覆盖性的特点，逐步取代传统媒体成为人们获取信息的主要渠道。在金融市场，互联网金融信息扮演着重要的角色，相比于传统的媒体，从互联网上获得的金融信息将最具有实时性和全面性。因此，对互联网金融信息的挖掘工作非常具有实践价值。正像前面所提到的，充斥在互联网上的信息多半是非结构化

的文本，这对于金融信息也是一样。金融信息的主要形式有各大金融门户网站的金融新闻、金融分析师们的评论、国家政策和公告、网络社区讨论等等。这些文本信息对金融市场会产生什么样的影响呢？这些文本本身的褒贬倾向性，和金融市场究竟是一种什么样的关系？我们希望通过我们进一步的研究可以对这个问题给出定性甚至定量的回答。

2. 相关工作

文本倾向性分析目前已经成为自然语言处理和机器学习领域一个非常引人注目的课题。通过文本倾向性进行分析，我们可以掌握文本作者的感情倾向。在金融领域，新闻舆情被作为体现投资者、交易者以及调整者观点和情绪的重要指标，它们和金融市场其他的实际指标，包括交易数据以及各种宏观经济指标等具有同样重要的意义[1, 2]。

从目前的相关文献来看，文本倾向性分析可以主要从两个方面展开。第一是为每一个文本计算出一个数值，该数值的符号代表了该文本的正负取向，该数值的绝对值大小代表了该文本的影响程度。第二是将每一个文本归到一个类别里面，即积极或者消极。从这两个方面来看，文本倾向性分析可以是一个回归问题，也可以是一个分类问题。目前比较多的研究者们还是从分类的角度来分析文本倾向性[3-9]。目前，文本挖掘和分类已经成很多研究者共同的兴趣，但是主要的分类工作针对文本的主题，对文本按照倾向性进行分类仍然没有引起更多学者的关注。为对文本倾向性的分类又可以主要从两个方面展开，一是利用机器学习的方法[3, 8]，一是利用基于语义分析的自然语言处理的方法[3, 4, 5, 6, 8, 9]。目前，文本倾向性分类已经被运用于英文[5-7]、中文[8-9]以及阿拉伯文[3]。

从目前的研究成果来看，文本倾向性分类已经引得了很多学者的关注，但是仍然尚未被广泛地运用于金融领域。金融文本倾向性分析对着金融市场有着举足轻重的作用，但是却还没有被众多学者涉足。

在挖掘金融信息与金融市场波动率关系的研究方面，目前主要是立足在新闻信息量与波动率的关系[13]，即主要以互联网上新闻的数量作为研究对象，而对信息的具体内容、褒贬性研究尚未开展。

本论文将立足于研究金融文本信息的倾向性与金融市场公司的股价波动率的关系。使用SVM，通过对纽约证券交易所177个公司在2007全年的股价表现及该年度的金融新闻的数据挖掘，试图对文本本身的褒贬倾向性，和金融市场究竟是一种什么样的关系，这个问题给出一个解答。

3 我们的方法

研究金融市场波动率与金融信息之间的关系通过两步。第一步，对收集到的新闻进行文本倾向性处理，为每篇文章打出一个分数来表示其褒贬性及其强度。第二步，对得到的文本倾向性与金融市场的交易数据建立模型，从而定量地建立起新闻舆论与股市波动率之间的关系。

实验中，我们需要对金融市场的数据进行建模。众所周知，金融市场之所以难以被被人们所掌握，就是因为其表现出来的随机性。而这点又突出表现在两个方面：一是其价格或交易量的具体走势，即其具体的价位或大小；二是其价格或交易量的波动率。本实验我们着重对其波动率变

化进行预测。

3.1 信息的褒贬值计算

互联网上与金融相关的信息，主要有两类。一类是各大网站金融板块的新闻，另一类是针对新闻的回帖以及各个论坛中的跟贴。这两类不同的信息，有着各自不同的特点。网站金融板块的新闻，其褒贬倾向性不是非常明显，但其内容真实，言之有物。各种各样的跟帖往往情绪化现象比较严重，虽然这对于做倾向判断比较容易，但其包含的信息量不高，而且比较凌乱。所以，本次实验中，我们选取的是美国各大主流网站的金融新闻作为研究对象。

这里的一个关键步骤就是为每篇新闻求一个褒贬倾向值。我们期望对每一条新闻得到一个实数值，其符号代表作者的褒贬倾向，其绝对值代表作者倾向的程度。Hownet[17]是一个以汉语和英语的词语所代表的概念为描述对象，以提示概念与概念之间以及概念所具有的属性之间在的关系为基本内容的常识知识库。在实验中，我们利用 Hownet，把新闻主体划分成一个由一系列关键词组成的矩阵，其中每一个元素被赋予一个特定的褒贬值，通过对这些褒贬值进行算法求解，得到整篇文章的褒贬倾向。

新闻的褒贬值取决于其生成的关键词矩阵，这些关键词往往具有较明显的褒贬倾向。实验中，我们使用 Hownet 作为关键词词典，把词分为八类，分别是 POSITIVE, NEGATIVE, PRIVATIVE, $MODIFIER_i$(i=1,2,3,4,5)，每一类 MODIFIER 都有一个权重 WEIGHT，代表其倾向程度。表 1 定义了这八类词。

表 1: 用于计算褒贬值的八类词

词集 s	描述
POSITIVE	表示正向的英文词汇，包括 4363 个词
NEGATIVE	表示负向的英文词汇，包括 4574 个词
PRIVATIVE	表示否定的词汇，包括 14 个词 {no, not, none, neither, never, hardly, seldom, barely, scarcely, ain't, aren't, isn't, hasn't, haven't}
下面是五类修饰词及它们的权重	
$MODIFIER_1$	64 个词，$WEIGHT_1$=2
$MODIFIER_2$	25 个词，$WEIGHT_2$=1.8
$MODIFIER_3$	22 个词，$WEIGHT_3$=1.6
$MODIFIER_4$	15 个词，$WEIGHT_4$=1.4
$MODIFIER_5$	11 个词，$WEIGHT_5$=0.8

新闻的褒贬值通过在文本中查找这八类词及其匹配程度来计算。然后，我们将使用以上的结果，综合金融市场股票交易量的波动率，利用一定的数据挖掘方法，实验中我们使用 SVM，对互联网上金融信息与金融市场的波动进行挖掘，来找出两者数量上的非线性关系，从而对金融市场的预测提供一定的帮助。

本实验使用的金融新闻来源于互联网上的众多的金融网站，实验的主体是美国股票市场上 177 家上市公司。足够大的样本是实证性研究是否有价值的关键因素，对于能否揭示出互联网金融新闻与金融市场的波动率之间的非线性关系有着重要的意义。下表是计算出褒贬倾向性后的整个新闻体的一个截图。

表 2： 新闻整体的一张截图，计算了 ADCT 公司以及它的褒贬倾向值

News ID	News Title	Time Window	Company Symbol	News Body	News Sentiment
2007010102057465	An Insecure Future for McAfee	2007-1	ADCT	Perhaps it"s fatigue with the options scandal that has now spread to more than 100 technology companies. Perhaps it"s …	45.399997999999997

3.2 预测模型的搭建

波动率一般认为是金融市场中在某一个时间段内的价格或交易量的方差，它的大小往往意味着不确定性和风险的大小。如果我们能够对金融市场的波动率进行预测，则对于投资者做出决策有着重要的意义。我们知道金融市场的波动率可以表现为一时间序列，其当前的值与其之前的一系列值以及外界变量有着密切的关系。我们就将利用这些因素通过 SVM 对其进行预测。

1）对波动率的 SVM 预测模型

我们知道波动率是一时间序列，第 t+1 天的值会受到其前面若干天数据的影响，而且其相互关系很可能是非线性的。据此，我们假设其在第 t 天的值为 σ_t^2，则其第 t+1 天的波动率 σ_{t+1}^2 的函数为：

$$\sigma_{t+1}^2 = \alpha_0 + \sum_{i=0}^{p} \alpha_i \ \pi_i \ (\sigma_{t-i}^2) \tag{1}$$

其中 p 是向前回溯的天数，α 是系数，$\pi_i \ (\sigma_{t-i}^2)$ 是关于 σ^2 的非线性函数。此模型的局限是只考虑波动率自身的时间序列，而忽略了外部因素对其的影响。我们知道金融市场是一个非常复杂的系统，影响其波动的因素很多。因为本实验主要研究互联网上的新闻褒贬对其的影响，所以我们把互联网上的新闻褒贬值作为一个外部扰动项加入到模型中去，得到新的预测模型：

$$\sigma_{t+1}^2 = \alpha_0 + \sum_{i=0}^{p} \alpha_i \ \pi_i \ (\sigma_{t-i}^2) + \sum_{j=0}^{q} \beta_j \theta_j (\varepsilon_{t-j}^2) \tag{2}$$

其中 ε 表示了新闻的褒贬值，$\sum_{j=0}^{q} \beta_j \theta_j (\varepsilon_{t-j}^2)$ 表示了前 j 天的褒贬值的非线性影响。

跟据上述公式，我们就可以搭建 SVM 的输入与输出了。但是公式还有一个缺点，即其是针对于某一支股票进行建模的，这样的结果缺乏对整个金融市场的总体性认识。所以，实验针对金融市场的整体概念已经了改进，把 177 家上市公司作为一个整体进行研究，从而使结果更有意义。

具体来说，我们把计算单位变成了一个长度为 L 的时间窗口，而不再是交易日。对于预测某支股票的股价波动率来说，假设在窗口 W_i 中，所有在 W_i 内当前公司的金融新闻的舆情之和表示为 ε_i^p。

所以原来的公式(2)修改如下：

$$\sigma_i^{p2} = \alpha_0{}^p + \alpha_i{}^p \pi_i{}^p (\sigma_{i-1}^{p2}) + \beta_i{}^p \theta_i{}^p (\varepsilon_{t-i}^2). \tag{3}$$

假设当前窗口的股价波动率σ_i^{p2}和上一个时间窗口的波动率以及上一个窗口的舆情值之和是非线性的关系，并且用二个未知的函数$\pi_i{}^p$和$\theta_i{}^p$表示。Miller(1979)认为自回归平均移动模型的残差之间并没有体现出明显的自相关性，但是其平方值之间却有着明显的相关性，因此(3)中的输入项全部取了平方值。

2）波动率的计算方法

因为交易数据要按窗口进行划分，所以波动率的计算就是在这些已经划分好的窗口之上进行的。实验中，我们主要研究股价的波动率变化。

假定一个时间窗口W_i所跨越的交易日为R天，每天收盘价格为y_t（t=t_i，…，t_{i+R-1}），则波动率就是股价在W_i窗口内的方差值。表示为：

$$\sigma^2 = \frac{\sum_{t=t_i}^{t_i+R-1} (y_t - \overline{y_t})^2}{R-1} \tag{4}$$

其中$\overline{y_t}$表示一个窗口内的股价的均值。

3 实验结果及讨论

我们使用SVM的回归功能来挖掘金融新闻的褒贬值与金融市场的股价波动率之间的关系。舆情值与股价波动率都被分割成一个个的时间窗口。在实验中，我们的选取了美国股票市场上177家公司，在2007年全年的交易数据和新闻数据。其中的交易数据是从yahoo finance[16]上得到的，新闻数据是来源于超过200个以上的英文金融网站。

SVM工具选取LIBSVM［14,15］，这是目前使用比较广泛的一个SVM实现版本。新闻数据横跨2007.1.1-2007.12.3，以周为单位进行时间窗口的分割，共有49个时间窗口。新闻总量为153,468条，对应着177家公司，这些新闻都被赋予了一定的舆情值。

我们使用相关系数平方（SCC）和波动率趋势预测准确度（VTFA）2个主要的性能指标来描述实验的结果。

1） 复相关系数的平方（SCC）

这个指标显示两个变量之间的密切程度。该值越接近于1，则相关度越好。

2）波动率趋势预测准确度（VTFA）

这个度量指标的含义在于考察变化趋势的预测准确率。每个公司都会预测得到下一个窗口的波动率，记录这个预测波动率相比上一个窗口波动率是上升还是下降，同时观察真实波动率相比上一个窗口波动率是上升还是下降，如果二者得到的结论一致，则这个公司对趋势的预测是正确

的。这个指标将统计在一个窗口中趋势预测正确的公司数目总和，计算所占比例。

我们对实验结果进行统计发现，在股价波动率的预测上，达到了 60.3225%的预测准确率，71.2593%的复相关系数平方值。图 1 描述了其中一家公司 MDT 在股价预测上的效果。图 2 描述了实验结果中 VTFA 值的变化情况。

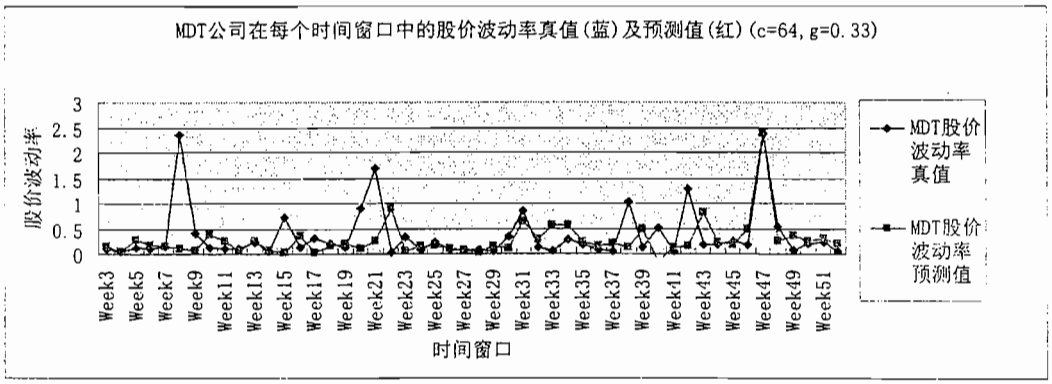

图 1 MDT 公司的股价波动率预测

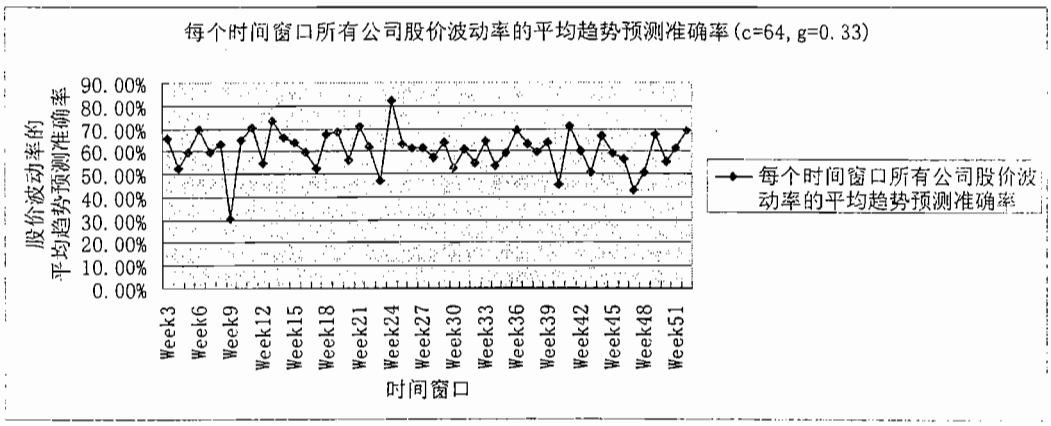

图 2 所有公司的 VTFA 情况

实验显示出金融市场上价格的波动与互联网上金融信息存在着比较密切的关系。参看图 1，预测值在绝大部分情况下与真实值相吻合。但当有重大的波动时，其预测性还是不能让人满意。

对于股价波动率和交易量波动率，其 VTFA 值都达到了 60%以上，说明在信息量与波动率之间确实存在着某种关联。我们的实验是在一个大的公司集上展开的，使结果更具有说服力。SCC 指标达到 70%以上，也支持了上述观点。

4 结论

本文着重研究了金融市场的波动与互联网上金融信息的相互关系，通过使用自然语言处理中的文本倾向性分析，结合金融市场本身的特点，共同来构建二者之间的模型，并通过就美国股票市场上的数据进行研究，论证了两者之间确实存在着比较密切的影响。虽然我们的实验是就美国股票市场进行的，但其模型和思想完全可以适用于任何一个金融市场。在未来的工作中，我们还

将考虑金融新闻褒贬性对单一股票的影响，设计对比实验，好进一步研究互联网上的新闻的褒贬值的重要作用。

参考文献

[1] Kindleberger, C. (2001), "Manias, Panics, and Crashes: A History of Financial Crises", Wiley Investment Classic, New York, John Wiley & Sons, 2001.

[2] Lakonishk, J., Lee, I. and Poteshman, A. (2004), "Investor behavior in the option market", NBER Working Papers 10264, Cambridge, Mass: National Bureau of Economic Research, 2004.

[3] Ahmad, K. and Almas, Y. (2005), "Visualising sentiments in financial texts?", Proceedings of the Ninth International Conference on Information Visualisation, Vol. 00, pp. 363-368.

[4] Chaovalit, P. and Zhou, L. (2005), "Movie review mining: a comparison between supervised and unsupervised classification approaches", Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[5] Turney, P.D. (2001), "Mining the web for synonyms: PMI-IR versus LSA on TOEFL", Proceedings of the Twelfth European Conference on Machine Learning, Berlin: Springer-Verlag, 2001, pp. 491-502.

[6] Turney, P.D. (2002), "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", presented at the Association for Computational Linguistics 40th Anniversary Meeting, New Brunswick, N.J., 2002.

[7] Turney, P.D. and Littman, M.L. (2003), "Measuring praise and criticism: inference of semantic orientation from association", ACM Transactions on Information Systems, Vol.21, pp. 315-346.

[8] Ye, Q., Lin, B. and Li, Y.J. (2005), "Sentiment classification for Chinese reviews: a comparison between SVM and semantic approaches", Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Aug. 18-21 2005, Vol. 4, pp. 2341-2346.

[9] Ye, Q., Shi, W. and Li, Y. (2006), "Sentiment classification for movie reviews in Chinese by improved semantic oriented approach", HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Jan. 04-07 2006, Vol. 3, pp. 53b-53b.

[10] H. Zheng, L. Xie, and L. Z. Zhang, "Electricity price forecasting based on GARCH model in deregulated market", *The 7th International Power Engineering Conference*, 29 Nov-2 Dec 2005.

[11] Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasticity", Journal of Econometrics, Vol. 31, no. 3, pp. 307-327, 1986.

[12] Engle, R.F. (1982), "Autoregressive conditional Heteroscedasticity with estimation of the variance of united kingdom inflation", Econometrica, Vol. 50, no. 4, pp. 987-1007.

[13] Nan Li, Chao Wang (2007) "Financial volatility forecasting based on inter-company connections and support vector machine", Proceedings of 2007 Journal Publication Meeting,Pre-Conference Meeting on Risk Management and Engineering Management, Toronto, Canada,September 24, 2007, pp. 112-118

[14] D. Thanh-Nghi and J. D. Fekete, "Large Scale Classification with Support Vector Machine Algorithms", *ICMLA 2007, Sixth International Conference on Machine Learning and Applications*, pp. 7-12, 13-15 Dec. 2007.

[15] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[16] http://www.finance.yahoo.com

[17] http://www.keenage.com/html/e_index.html