研究计划书

一、研究方向

1.1 研究问题

开发一种类似于 AZFinText（Arizona Financial Text System）的中文金融新闻信息文本挖掘系统，实现对海量中文财经新闻的即时文本挖掘，由此对股价进行预测，并结合一些常见的量化投资策略进行模拟程序化交易。根据不同策略组合可以得到不同的投资回报率，而从这些回报率的分析上可以进一步得到市场中投资者的行为模式信息。这一类研究的模板是 AZFinText 的开发者 Robert P. Schumaker 近年所发表的一系列论文，本研究亦可视为是其思路在中文环境下的实现（差异不仅在于市场和投资人，更多的在于因语言的不同导致文本挖掘技术的不同）。

1.2 研究现状

国外使用文本挖掘技术来预测舆情、股价的文章有很多，但将文本挖掘和量化投资结合起来的寥寥无几。尽管 AZFinText 很有名气[1]，但并没有引起金融学学者的关注（在谷歌学术搜索中有 16 次引用，无一来自金融学论文，这是令人难以相信的忽视），更主要影响仅在信息管理领域。Schumaker R P, Chen H（2008）研究了结合金融新闻文本挖掘预测系统（AZFinText）和量化投资策略进行选股的回报率。数据和新闻的获取时间段选在 2005 年 10 月 26 日-2005 年 11 月 28 日 5 周共 23 个交易日（选择该时间段的原因在于无重大市场冲击发生），持股期也选在该阶段（即没有卖出），期末计算回报率。使用的两种基本的投资策略是顺势和逆势投资策略，分别是将时段长度为 f 的股票回报率进行排序，购买或卖空排名在前 20%的股票称为顺势，而购买后 20%的则称为逆势（为了使结果具有稳健性，f 分为 1-5 周五种情况）。投资的结果是结合使用全部金融新闻挖掘和量化投资策略的投资在顺势、逆势等十种情况中回报率都远高于仅用部分金融新闻挖掘或仅使用量化投资策略的投资回报（作为对比，顺势投资策略且 f=1 时，三者的回报率分别是 20.79%，0.33%，-5.54%）。

国内学界对该领域使用文本挖掘技术来研究新闻对股市影响的文献更少，目

[1] 相关新闻可见于 http://www.technologyreview.com/view/419341/ai-that-picks-stocks-better-than-the-pros/

前使用谷歌学术搜索所能检索到的仅有 Xiangyu Tang 等（2009）、赵丽丽等（2012）、黄诒蓉（2013），甘甜甜（2012）几篇文献，并且无一和量化投资相关。在不进行程序化交易的情况下，金融新闻文本挖掘和人工阅读新闻相比并没有多大的优势。

至于国内业界，我使用谷歌搜索关键词 intitle:“文本挖掘”“量化投资”（表示搜索以“文本挖掘”，“量化投资”作为标题的网页），返回的搜索结果仅为两页，和业界相关的仅有“广发证券”的三个结果，其中真正涉及文本挖掘的仅有一个，“网络文本挖掘方法介绍”，“量化投资专题”。该文档形成时间为 2012 年 4 月，但内容也没有涉及到量化投资的应用。由此可以断定，即使有机构在从事这一方向的研究，那也仍是处于起步阶段。

1.3 研究意义

从研究现状来看，这是一项具有强烈市场需求但无论从学界还是业界来看都仅仅处于起步阶段的研究。如果文本挖掘技术有所突破，预计未来几年将会对业界有重大影响。例如业界对股评人、新闻分析师、行研报告撰写员等人才的需求将会大幅下降。

文本挖掘的优势在于，在短时间内分析大量的信息，从而能快速判断市场期望，而使用量化投资方法则能抢在期望实现之前完成交易。“扫描股票价格和金融新闻，买入哪些它相信在 20 分钟之内会上升超过 1%的股票，然后在 20 分钟之后卖出。”、“在进行长期预测时有很多因素需要考虑（因此这样的投资策略可能不管用），但在短期，例如五分钟，十分钟之内，（新闻对股票价格的影响是最主要的因素），这样做就很有优势。”[2]。

二、研究方案

2.1 新闻信息来源和提取

新闻有多种来源，按权威性排序可分为：（1）证监会，银监会，央行等政策发布的信息；（2）各上市公司发布的财务报告、著名评级公司评级；（3）各大财经门户网站新闻；（4）股评人评论，知名博客自媒体文章等等。

为了能实时抓取这些信息，可以编写网页爬虫（web crawler）来抓取网页信息。

[2] 翻译自 http://blogs.wsj.com/digits/2010/06/21/using-artificial-intelligence-to-digest-financial-news/

2.2 新闻信息的文本挖掘

对于本项研究而言，文本挖掘的目的在于，经过分类或聚类技术，从每一个文本中得到对某一项投资正面、负面或中性的意见（将评价数字化，例如正面评价设为 1，中性评价设为 0，负面评价设为-1）。而最后是否实施该投资则取决于所有文本的综合分析结果。例如某个上市公司财报上利润下降、著名评级公司或分析师下调某上市公司的信用评级等新闻将对投资该公司的股票给出负面的评价。不同的新闻有可能给出相反的评价，经过赋予每个新闻评价各自的权重（可以按照新闻来源的可靠性来决定权重大小），是否投资、如何投资则取决于全部评价的加权平均和。

权重设定对最终的加权平均和有重大影响，因此如何设置合适的权重是投资策略决定中非常重要的问题。Shah V H（2007）给出了一种通过机器学习来改进权重设置的方法。例如，在初始阶段每位股票分析师的股价预测赋予相同的权重，然后根据预测和实际结果偏离情况的历史数据来调整下一轮赋予分析师股价预测的权重。例如上期预测较为准确的分析师的预测权重在这期翻倍，而预测不准的则权重减半。可以预见，随着学习的阶段逐渐增加，权重会渐趋合理。

作为文本挖掘的初学者，在如何将非结构化的新闻文本信息转化为可以被计算机处理的结构化向量信息等技术细节上，我大量参考了周君（2009）及甘甜甜（2012）两篇文章。图示及例子也是来源于两篇文章。本节也可视为上述两篇文章的精简版本。

文本挖掘可分为以下步骤：文本预处理，文本特征表示，文本特征选择，文本分类和聚类。下面将讲解各个步骤的具体实现。

2.2.1 文本预处理

文本预处理是为了有效提炼出文本内容，消除无实义语词而对文本进行的处理。由于本研究仅对中文新闻进行文本挖掘，因此只讨论中文文本的预处理。中文文本与英文文本不同，词与词之间没有间隔，因此首先需要将文本进行分词处理，即拆分为离散的单词序列。

中文分词处理主要有三类方法：基于字符串匹配的分词方法、基于统计的分词方法和基于理解的分词方法。为简便起见，这里仅介绍基于字符串匹配的

正向最大匹配法。当词库中的最大词长为 n 时，将待处理的字符串序列前 n 个字作为匹配字段，遍历词库，若词库中存在，则匹配成功，该匹配字段被切分下来；若匹配不成功，则去掉匹配字段的最后一个字，重新遍历；重复以上两个步骤，直到文本中的所有词都被切分出来为止。

目前比较实用的中文分词系统有北大计算语言所的分词系统、清华 SEGATG 系统、中科院 ICTCLAS 系统、复旦分词系统和哈工大统计分词系统等。以下是使用北大计算语言所分词系统对一篇财经新闻分词后的效果（引自甘甜甜（2012））：

示例新闻：

[季报]建设银行首季净利降 18%资产减值损失急增 2009 年 04 月 24 日 22:47

建设银行(4.56,0.00,0.00%) (601939)周五晚间发布一季报,受央行选续降息及资产减值损失急增影响,公司一季度净利润同比下降接近一成。一季度建行实现营业收入 654.58 亿元,同比增长 0.84%,净利润 262.76 亿元,同比下降 18.22%,每股收益 0.11 元,同比下降 21.43%。数据显示,建行一季度利息收入 508.7 亿元,较上年同期下降 6.55%,因央行从 08 年 9 月开始连续降利影响,同时因资产规模增长并审慎足额计提减值损失准备,资产减值损失较上年同期增加 74.4 亿元。

截止 2009 年 3 月 31 日,建设银行资产总额为 86746.33 亿元,较上年末增加 11191.81 亿元,增长 14.81%,资本充足率 12.37%,较上年末上升 0.12 个白分点,核心资本充足率 9.60%,较上年末下降 0.57 个百分点。

分词后结果：

[/w 季报/n]/w 建设银行/n 首/d 季/Ng 净利/n 降/v 18%/n 资产/n 减/v 值/v 损失/v 急/a 增/v 2009 年/t 04 月/t 24 日/t 22:47 /m

/x 建设银行/n (/w 4.56/m ,/w 0.00/m ,/w 0.00%/n)/w (/w 601939/m)/w 周/nr 五/nr 晚间/t 发布/v 一/m 季报/n ，/w 受/v 央/Ng 行/a 连续/a 降/v 息/Ng 及/c 资产/n 减/v 值/v 损失/v 急/a 增/v 影响/n ，/w 公司/n 一/m 季度/n 净利/n 润/v 同/d 比/v 下降/v 接近/v 二/m 成/a 。/w

/x 一/m 季度/n 建行/j 实现/v 营业收入/n 654.58/m 亿/m 元/q ，/w 同/d 比/v 增长/v 0.84%/n ，/w 净利/n 润/v 262.76/m 亿/m 元/q ，/w 同/d 比/v 下降/v 18.22%/n ，/w 每/r 股/q 收益/n 0.11/m 元/q ，/w 同/d 比/v 下降/v 21.43%/n 。/w

/x 数据/n 显示/v ，/w 建行/j 一/m 季度/n 利息/n 收入/v 508.7/m 亿/m 元/q ，/w 较/v 上年/t 同期/d 下降/v 6.55%/n ，/w 因/c 央/Ng 行/a 从/p 08/m 年/q 9 月/t 开始/v 连续/a 降/v 利/n 影响/v ，/w 同时/n 因/p 资产/n 规模/n 增长/v 并/v 审慎/a 足额/d 计/v 提/v 减/v 值/v 损失/v 准备/v ，/w 资产/n 减/v 值/v 损失/v 较/v 上年/t 同期/d 增加/v 74.4/m 亿/m 元/q 。/w

/x 截止/v 2009 年/t 3 月/t 31 日/t ，/w 建设银行/n 资产/n 总额/n 为/u 86746.33/m 亿/m 元/q ，/w 较/v 上/v 年末/t 增加/v 11191.81/m 亿/m 元/q ，/w 增长/v 14.81%/n ，/w 资本/n 充足/a 率/v 12.37%/n ，/w 较/v 上/v 年末/t 上升/v 0.12/m 个/q 百分点/n ，/w 核心/n 资本/n 充足/a 率/v 9.60%/n ，/w 较/v 上/v 年末/t 下降/v 0.57/m 个/q 百分点/n 。/w

可以看到，连续的句子被切割成离散的词语，而词语后面的 n、v、t 等则是代表名词、动词、时间词的词性标注。

文本中出现频率高而意义不大的词，像副词，语气词，连词等统称为停用词（stop words）。这些词对文本分类和聚类没有用处，只会增加干扰，因此在分词结束后应该将它们去除掉。去除方法为建立一个停用词词表，然后遍历分词后的文档，发现匹配的词则将其从分词文档中去除。

经过上述两步之后，可以继续作仅保留名词和动词的处理（一般来说这些是最重要的词语），剩下的分词可以称为该文本的特征词。上述新闻可精炼如下图所示：

2.2.2 文本特征表示

经过中文分词和去除停用词后得到的分词文本，仍然是非结构化信息。文本的特征表示则是将文本用模型转变为某个特征向量，以便计算机识别及运算。新闻文本中的分词繁多，但其中各个分词对该篇新闻的反映能力是有区别的。因此，我们可以将一份文档看作是由拥有不同反映权重的分词组成的序列。换言之，文档可以表示成一组由特征项组成的特征向量$d_i=(w_1,w_2,...,w_k,...,w_n)$，其中$w_{ki}$为第 k 个分词对该文档的反映权重，当该值越大，表示该分词反映该文档的能力越强。计算反映权重的常见方法有：布尔加权法，词频法，TF-IDF 加权法（Term Frequency-Inverse Document Frequency）等。这里仅介绍三种中最复杂但也是最常用的的 TF-IDF 加权法。

TF-IDF 加权法的基本思想是：当某一个分词在某文本中出现的次数越多，表示其越重要；当某个分词在越多的文本中出现时，则表明其越不重要（这里有些问题：像“增长”、“利润”、“负债”等词语几乎是每篇财务报告中都会出现的，难道我们能认为它们不重要？还有仅看次数不看顺序是否也是有问题，毕

竟“下降”出现在“利润”后面还是出现在“负债”后面差别不小。姑且存疑）。

令词频 TF 表示该分词在某特定文本中出现的次数，文档频率 DF 则表示全部文本集中包含该分词的文本的比例，而逆文档频率 IDF 则是 DF 的倒数的对数。

TF-IDF 的基本公式：$w_{ki}=tf_{ki}\times\log(\frac{n}{n_{ki}})$ 。其中 tf_{ki} 为词频，n 为文本集中文本的总数量，n_{ki} 则是含有该分词的文本的数量。如果某个分词出现的频率非常高，以至于 $n=n_{ki}$，那么由公式计算出的权重则为 0.为了避免出现这样的情况，可以修正公式为 $w_{ki}=tf_{ki}\times\log(\frac{n}{n_{ki}}+0.01)$ 。另外，当某文本越长，该分词在文本中的权重就会越小。为了避免这种单纯因为文本长度而造成的反映权重偏差，公式规范化为 $w_{ki}=\frac{tf_{ki}\times\log(\frac{n}{n_{ki}}+0.01)}{\sqrt{\sum_{k=1}^{n}[tf_{ki}\times\log(\frac{n}{n_{ki}}+0.01)]^2}}$，由此公式得到每一个分词的权重值都在（0,1）区间。

2.2.3 文本特征选择

通过文本的特征表示，我们可以得到某一文本的特征向量。但通常特征向量的维数都会很高。将特征向量的维数降下来有利于提高后续分类、聚类的效率。特征选择是从原始的特征空间（所有文本特征词的集合）挑选出部分较为重要的特征重新组成一个低维的特征空间，其实现思想为：通过某个评估函数对每个特征词进行计算，选取计算结果满足预先设定的阈值的特征词作为最终的特征词。常见的方法有：文档频率(Document Frequency)、信息增益(Information Gain)、互信息(Mutual Information)、期望交叉熵(Cross Entropy)、x2 统计(CHI)等。

2.2.4 文本分类和聚类

分类是有监督的学习，需要事先分好类别作为训练集。分类常用的方法有 K 近邻算法、朴素贝叶斯法和支持向量机算法等、而聚类则是无监督的学习，不需要先人工分好类别。在分类/聚类结束后需要使用准确率、召回率和 F-

Measure 等指标进行评价。周君（2009）介绍了一种 K 均值和遗传算法相结合的聚类方法，在 F-Measure 评价上优于单纯使用 K 均值法。文本分类或聚类的结果是对某一项投资正面、负面或中性的意见。

2.3 文本挖掘云计算平台的搭建

从上面所介绍的文本挖掘的步骤，可以判断计算量将会随着需要分析的新闻数量增加而急速增加。然而，为了能够在新闻发出来的第一时间就进行文本挖掘并形成投资建议（在市场还没完全反应之前才能进行套利），我们需要实时监听各大财经门户和政府网站，尽量抓取任何有关的新闻页面。因此，进行文本挖掘所需要的计算能力就会急剧增加。这样的计算能力是个人 PC 难以实现的。

解决这一瓶颈的方法有以下两种；（1）租用云服务器。亚马逊，微软，阿里等云服务提供商提供个性化的计算服务：需要多少计算能力则购买多少。（2）搭建个人的云计算平台。刘智勇（2011），周姚（2011）分别介绍了如何搭建 Hadoop 云计算平台进行文本挖掘的过程。

2.4 文本挖掘结合量化投资

具体实现过程如图（引自 Schumaker R P, Chen H（2008））：

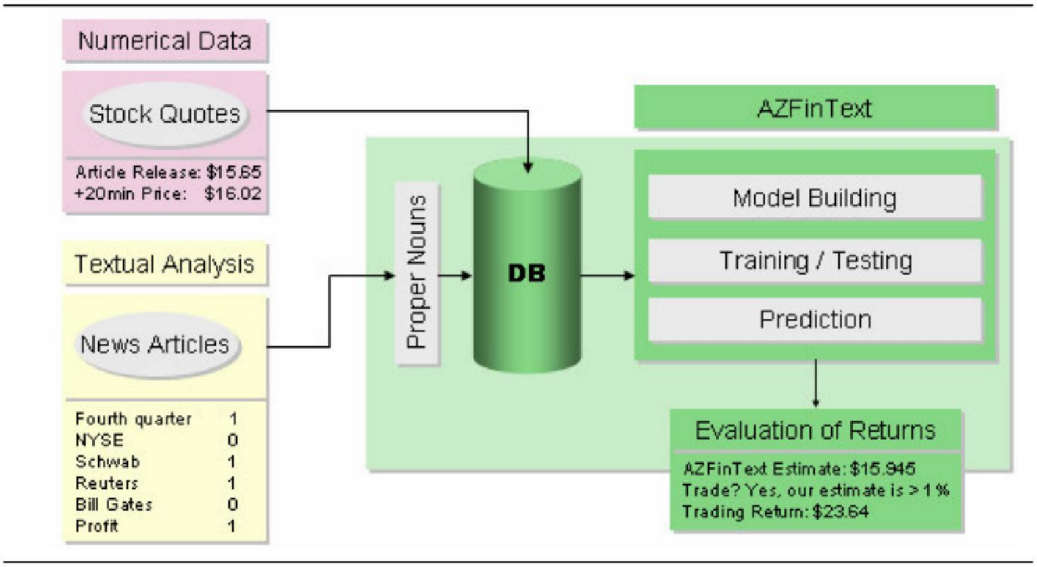

Figure 4. Example of AZFinText Processing

参考文献

[1] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market[J]. Journal of Computational Science, 2011, 2(1): 1-8.

[2] Robert P. Schumaker. Textual analysis of stock market prediction using breaking

financial news: the AZFinText system. ACM Transactions on Information Systems, 2009,2(2).

[3] Robert P. Schumaker,Hsinchun Chen. A quantitative stock prediction system based on financial news. Information Processing and Management, 2009, 45(5): 571-583.

[4] Shah V H. Machine learning techniques for stock prediction[J]. Courant Institute of Mathematical Science, New York University, 2007.

[5] Schumaker R P, Chen H. Evaluating a news‐aware quantitative trader: The effect of momentum and contrarian stock selection strategies[J]. Journal of the American Society for Information Science and technology, 2008, 59(2): 247-255.

[6] Xiangyu Tang, Chunyu Yang, Jie Zhou. Stock price forecasting by combining news mining and time series analysis, in Proceedings of 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2009, 279-282.

[7] 邓自立.云计算中的网络拓扑设计和 Hadoop 平台研究. 中国科学技术大学，2009

[8] 甘甜甜. 基于文本挖掘的财经领域趋势分析技术研究[D]. 北方工业大学, 2012.

[9] 黄诒蓉. 金融研究中的新闻分析框架及应用[J]. 证券市场导报, 2013 (001): 37-44.

[10] 刘智勇. 基于云计算的文本挖掘算法研究[D]. 电子科技大学, 2011.

[11] 赵丽丽, 赵茜倩, 杨娟, 等. 财经新闻对中国股市影响的定量分析[J]. 山东大学学报 (理学版), 2012, 47(7).

[12] 周君. Web 文本挖掘关键技术的研究与实现[D]. 西安电子科技大学, 2009.

[13] 周姚. 基于云计算的文本挖掘技术研究[D]. 国防科学技术大学, 2011.