互联网证券资讯监测系统①

莫 倩[1], 苑 峥[1], 张华平[2]

[1](北京工商大学 计算机与信息工程学院, 北京 100037)
[2](北京理工大学 计算机学院, 北京 100081)

摘 要: 目前, 网络资讯监测技术大都是用于政治和商务领域. 鉴于此, 提出一个将网络资讯监测技术应用于证券领域, 实现针对证券类资讯进行监测的互联网证券资讯监测系统. 系统能够根据不同的证券资讯策略, 动态监测互联网证券资讯信息源, 结合证券领域本体, 自动采集相关的证券资讯信息, 对证券资讯进行多空判别. 将从相关工作, 体系结构、主要功能和关键技术对互联网证券资讯监测系统进行描述.

关键词: 证券资讯; 资讯监测; 资讯采集; 证券本体; 多空判别

Internet Securities of Information Monitoring System

MO Qian[1], YUAN Zheng[1], ZHANG Hua-Ping[2]

[1](Institute of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100037, China)
[2](Department of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Currently, the technology of Internet information monitoring mostly uses for political and business fields. In view of this, this paper proposes a system for the Internet information monitoring in the securities field, achieve for monitoring of Internet securities information. Systems can monitor the information sources of Internet securities information dynamicly based on different strategies of securities information. Combining with the domain ontology of securities, it can collect the information related to securities information. The system also can distinguish long and short securities' information. This paper proposes its related work and architecture, specifies its major functions and key technologies.

Key words: securities of information; monitoring information; collection of information; domain ontology of securities; distinguish long and short

1 引言

证券市场是一个信息传播快、市场化程度高的市场[1], 证券市场上虚假消息、舆论操控、证券黑嘴等违法现象层出不穷却又稍纵即逝, 证券资讯安全问题越来越突出. 面对日渐泛滥、不断误导投资者的各类网络财经消息, 无论是证券市场管理机构、财经媒体、上市公司还是股民, 都需要及时了解和严密监控互联网上有关证券市场的资讯, 包括新闻、公告、股评、股民评论等, 以期适时化解因信息误传而带来的投资风险.

互联网证券资讯监测的关键在于: 综合海量的网络消息, 剔除各类干扰因素, 准确地判别出网络资讯对特定证券对象的看空或者看多的走势, 这直接关系到证券价值评估和股价的涨跌. 如何从互联网海量的证券消息中挖掘并分析出社会民众群体对特定证券对象的观点、态度、意见和看法, 如何防止资讯干扰, 如何依据微观的倾向性数据综合计算出资讯看空看多的走势, 并如何与实际交易价格或者指数的涨跌进行验证, 这已成为当前证券领域与计算机领域需要共同面对的一个挑战性研究问题.

① 基金项目:国家自然科学基金(61170112);北京市教委科技创新平台建设项目(PXM 2011_014213_113631);新疆自治区高新技术发展计划(201212124)
收稿时间:2012-05-27;收到修改稿时间:2012-06-30

本文介绍了一个将网络资讯监测技术应用于证券领域的互联网资讯监测系统. 第一部分为引言, 第二部分阐述了网络资讯监测的相关工作, 第三部分描述了证券资讯监测系统的体系结构和主要功能, 第四部分展示了证券资讯采集的关键技术以及对证券资讯进行多空判别的关键技术, 第五部分是结束语.

2 相关工作

最近几年来, 我国的一些政府管理部门联合了部分国立科研机构和公司, 开展了针对网络舆论、网络军事情报等相关信息进行监测管理关键技术研发和应用系统部署工作. 国家 863 计划以探索导向的形式资助资讯分析预警相关技术研究, 如: 面向资讯的话题发现[2-4]、文本倾向性分析[5-7]等. 国家 973 计划支持有监督的网络信息内容分析计算的基础理论方面研究.

在资讯监测系统的开发和应用方面, 比较有影响的单位包括英国 Autonomy 公司、北大方正、托尔思公司(TRS)[8]等. 这些单位能够通过提供通用性的产品, 对从网络中收集过来的信息进行聚合分析、全文检索, 从而在一定程度上满足网络资讯信息的分析与挖掘应用. 然而, 网络资讯分析挖掘是一个比较复杂的任务, 需要与实际需求紧密结合, 通过一两类所谓成熟的算法来达到资讯监测的深度分析实际上是不现实的.

以中科院计算所、北京大学[9]、复旦大学、北京理工大学[10]为代表的一些国内科研机构在网络资讯分析、监测与预警方面进行了大量的研发工作, 他们的部分成果已经形成了应用系统并在需求单位进行了试用. 此类研究工作融合了搜索引擎、文本挖掘和资讯计算的相关技术成果, 从技术创新性上来看优于前述公司. 但是, 目前的不足是这些科研机构各自为战, 在资讯监测与情报挖掘的一些棘手问题上没有形成合力, 从而在资讯观点分析、资讯态势分析、话题发现与趋势预测等相关问题上还没有实际可用的成果.

本文在前述已有的工作基础上, 提出了我国第一个基于证券领域的网络资讯监测系统. 系统能够根据不同的证券资讯策略, 动态监测互联网证券资讯信息源, 结合证券领域本体, 自动采集证券资讯信息, 综合海量的网络消息, 剔除各类干扰因素, 准确地判别出网络资讯对证券对象的看空或者看多的走势等.

3 体系结构设计

证券资讯监测系统的体系结构如图 1 所示.

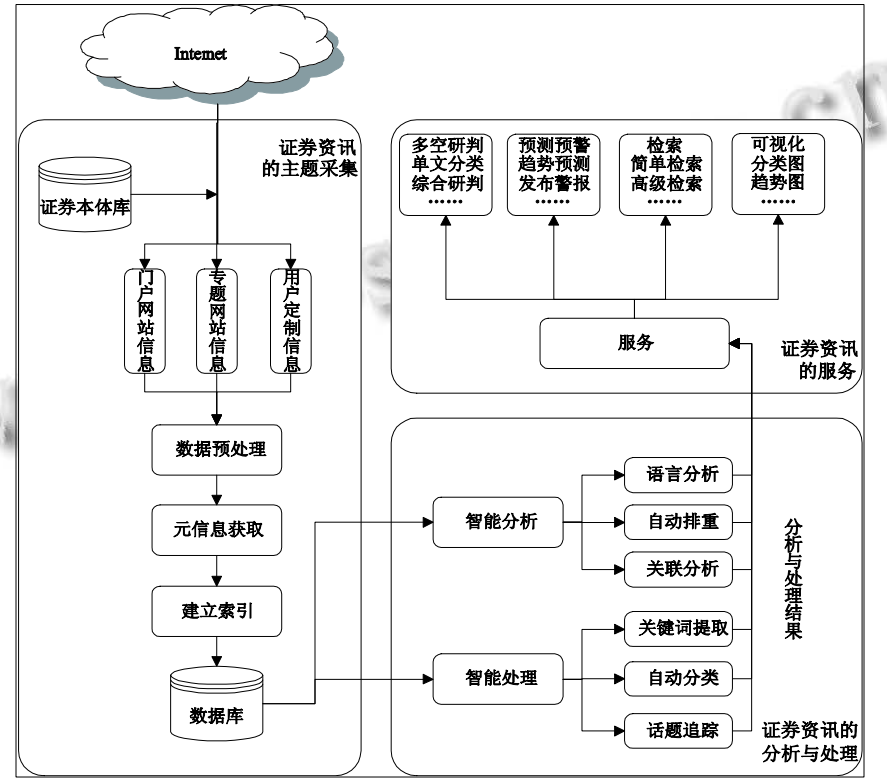

图 1 证券资讯监测系统体系结构

互联网证券资讯监测系统可以分为三个主要的模块: 证券资讯的主题采集模块、证券资讯的分析与处理模块和证券资讯的服务模块.

3.1 证券资讯的主题采集

系统设计采用数据采集机器人(crawler)与证券本体相结合的方式对互联网上各种交互式数据源中的证券资讯信息进行数据的采集. 在经过数据预处理(如编码识别、简繁体转化、格式转换等)后, 获取相关的元信息, 并进行内容提取, 然后对文本内容创建必要的索引, 最后将原始信息、各种元信息和索引数据存储到相应的数据库中, 并通过统一的数据访问接口为证券资讯分析与处理模块提供数据访问服务.

3.2 证券资讯的分析与处理

此模块利用各种自然语言处理技术与文本挖掘技术通过数据访问接口对采集到的证券资讯信息进行智能的分析与处理, 包括语言分析(如词法分析、命名实体识别、浅层语法分析等)、自动消重、自动摘要、关键词提取、自动分类、关联分析、话题跟踪等. 系统会自动的将分析与处理的结果提交给证券资讯服务模块作相应的处理.

3.3 证券资讯的服务

证券资讯的服务模块是用户与系统之间的直接接口, 它利用证券资讯分析与处理模块的各种技术支撑起应用功能, 为用户提供相应的资讯服务, 如资讯的多空研判、预测预警、检索和可视化等等. 资讯的多空综合研判将会在文章第四部分进行详细的说明. 各项服务结果通过统计报告、可视化图表等直观的形式, 借助互联网发布. 用户也可以定制 SMS 短信提示服务. 系统会自动将最新的证券资讯信息以短信发送的方式通知给用户, 从而为用户提供辅助决策支持.

4 关键技术

本系统区别于其它网络资讯监测系统的关键技术在于, 如何针对证券资讯信息进行数据采集以及如何对海量的证券资讯消息进行多空综合研判.

4.1 基于证券本体的数据采集

本系统设计利用数据采集机器人(crawler)与证券本体相结合的方式对互联网上各种交互式数据源中的证券资讯信息进行数据的采集.

首先是要建立证券领域的本体[11-13]. 证券领域本体库主要包含证券实体对象库、证券信息点库、证券多空属性库.

证券实体对象库主要包括在上交所、深交所、港交所、纳斯达克、纽约股市等上市的中国公司、股票代码、高管、行业等数据, 以及行业、大盘乃至交易所与各监管机构的基本属性信息.

证券信息点库主要包括上市公司、证券公司、基金、债券、监管机构、行业等不同证券实体对象的分类信息点. 如对上市公司, 我们一般要从高管、财务状况、股权结构、经营管理 、突发事件、 股价异动等角度对其资讯消息分类, 单就财务状况而言, 同样还需要考虑的层次有多种情况, 见下表.

表 1 财务状况本体构建范例

财务状况	业绩问题	巨亏巨盈、业绩造假、坏账计提、粉饰报表、隐瞒利润、财报造假
	资金链问题	资金链断裂、资金链恶化、资金链出现问题、负债
	挪用掏空	资金挪用、掏空上市公司
	资产转移	资产转移、热钱、IPO、金融危机、金融海啸
	债权债务	债权、债务、转让、收购、 认购、债务沉重
	资产抵押	资产抵押、资不抵债

证券多空属性库包含表示多空属性、多空极性的词语.

将证券领域的本体与数据采集机器人相结合, 通过网页关键字与证券领域本体匹配的方式, 对采集到的网页进行一个二值(是/否)的判定, 为了节省采集时间, 提高采集效率, 本系统设置了几条规则用来辅助生成判定结果: 设置正则表达式来判定网页 P 是否为网站主页, 若网页 P 为网站主页且被判定为属于证券领域, 则将该网站主页以下各级网页均视为证券领域网页, 不需再次进行匹配; 若网页 P 是某网站下第 n 级网页, 属于证券领域, 则将该网站下自网页 P 开始的所有网页均视为证券领域的网页, 不需再次进行判定. 根据网页判定结果, 过滤掉非证券领域的网页, 保留属于证券领域的网页, 从而进行数据采集.

4.2 证券资讯多空综合研判

证券资讯多空综合研判主要包含以下四个步骤:

① 将证券多空属性词与具体的证券信息点关联, 并映射到具体的证券实体对象, 统计多空数据结果, 得到单个消息的多空倾向性.

单个消息的多空倾向性判定方法如下. 首先进行本体扫描, 按照句子顺序, 找出每句话中的证券实体对象以及多空属性. 忽略与多空判别无关的句子. 将每句话中的证券实体对象以及多空属性词抽取出来组成关联词对, 以关联词对的数量及多空属性计算出单个消息的多空倾向性结果.

以“管窥海王生物”为例, 下划线的词为看多属性词, 斜体的词为看空属性词, 而加粗的词为证券实体对象.

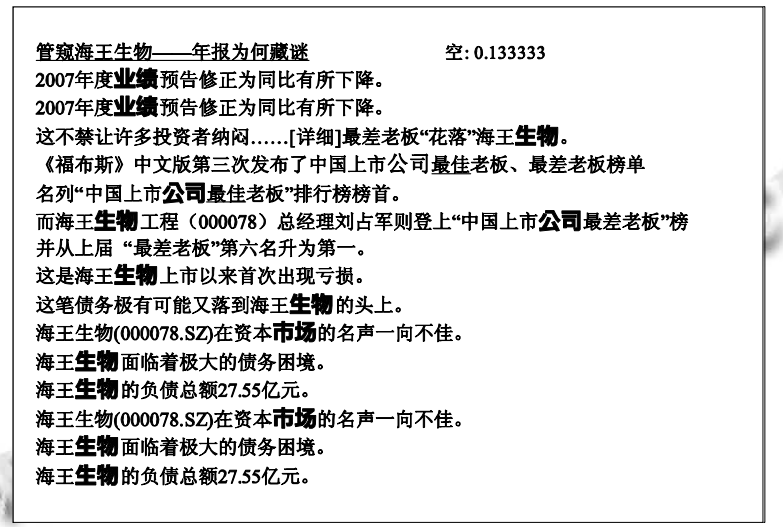

图 2 单篇消息本体扫描结果示例

<业绩,下降> <业绩,下降> <生物,差> <公司,最佳> <公司,最佳> <公司,差> <生物,差> <生物,亏损>
<生物,债务> <市场,佳,不> <生物,债务> <生物,负债> <市场,佳,不> <生物,债务> <生物,负债>

图 3 单篇消息多空词对结果示例

综合计算: 一共有 15 个消息词对, 其中看多的有 2 对, 看空的有 13 对. 最终得到结果, 该消息为看空, 概率为 0.133333.

② 将单个消息的多空倾向性与消息的可信度与影响力相结合, 加权计算, 其形式化表述如下:

$$Tend(object)=T[r(m),cred(m),\inf(m)]$$

其中, m 为单篇文章的信息, T 为多空计算函数, $Tend$ 为多空综合研判函数, $object$ 为特定证券实体对象, $r(m)$ 为单个消息的多空倾向性结果, $cred$ 为消息的可信度计算函数, inf 为消息的影响力计算函数. 消息的可信度计算主要依据消息的作者 author、发布的时间 release_time、媒介形式 source_type、发表的媒体 source_id 等情况综合计算. 消息的影响度计算, 主要依据消息的点击数 clicks 以及转载数 forwards 进行计算.

③ 将消息集上所有单个消息的加权结果求均值.

$$Tend(object)=\frac{\sum_{M=1}^{n}T[r(m),cred(m),\inf(m)]}{n}$$

④ 引进历史多空数据维度, 修正多空倾向性结果.

证券资讯存在“报喜不报忧”的传统, 看多消息的发布往往会得到证券实体对象的大力配合, 甚至有些消息本身就是证券实体对象发出的广告性质的通稿, 而看空信息在没有确凿证据支撑前, 往往话语空间有限. 因此, 除非有特别糟糕的情况出现, 比如老鼠仓或者是有重大失误外, 特定证券资讯往往都是看多信息为主.

为此, 我们还要进一步对综合研判模型进行修正, 具体思路为: 引进历史多空数据维度, 即相比过去一段时间来看, 看多消息总数量或者比例的下调, 均可以视为看空的特征. 也就是说, 对于某只股票在所有的消息集上, 看多的消息比例占优, 但是, 相比最近一段时间来说, 看多的实际数量或者多空对比都在减少, 那么, 很可能是公众看空的一个例证. 没有更多利好消息支持, 就是坏消息. 因此, 判别模型需要修正如下:

$$Tend_t(object)=A$$

$$A=\alpha\frac{\sum_{M=1}^{n}T[r(m),cred(m),\inf(m)]}{n}$$

t 为时间戳，α 为调节因子，主要由历史多空数据与当前数据比对计算. 采用上述公式计算出来的多空结果，如 1.0 并不一定就意味着 100%看多，该数据更大程度上是一种比较意义. 我们需要基于历史数据，进行机器学习，引入相对完善的调节因子，进行归一化，并定位出多空分界线.

以联想集团(00992.hk)2011.08.11 至 2011.08.18 的消息为例，综合研判的结果示意如下:

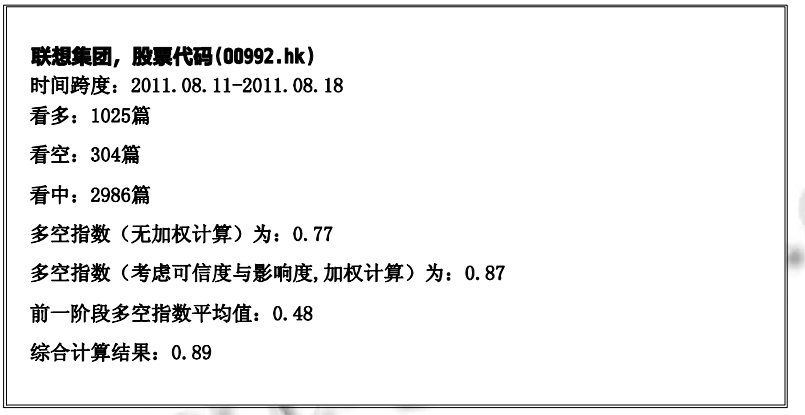

图 4 联想集团(00992.hk)综合研判结果示例

5 结语

互联网证券资讯监测系统是我国第一个基于证券领域的网络资讯监测系统. 在全面继承一般网络资讯监测系统的功能与技术的同时也充分体现了它的特点，即它的领域性. 不论是证券本体与数据采集的结合，还是证券资讯的多空判别，无不体现出它基于证券领域的特点.

在现有工作的基础上，今后将继续增加并改进系统的功能，如将证券领域本体引入到证券资讯的自动分类中，提高分类准确率，构建证券资讯与股票价格之间的关系模型，辅助决策支持.

参考文献

1 周斌.火爆的中国证券市场.上海经济,2007(11):60–62.

2 张华平,秦鹏.基于关键词提取的检索结果聚类研究,第五届全国信息检索学术会议(CCIR2009),上海,2009,11.

3 Allan J, Gupta R, Khandelwal V. Temporal Summaries of News Topics. Proc. of SIGIR. 2001:10–18.

4 Ku LW, Li LY, Wu TH, Chen HH. Major topic detection and its application to opinion summarization. SIGIR, 2005: 627–628.

5 Liu X, Mo Q, Zhang Z. Research on opinion classification of internet reviews. Journal of Beijing Technology and Business University (Natural Science Edition), 2008,26(3):61–65.

6 Yao TF, Chen XW, Xu FY, et al. A survey of opinion mining for texts. Journal of Chinese Information Processing, 2008,22(3):71280.

7 Pang B, Lee L. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2008, 2(1-2): 1–135.

8 Du YC, Wang HY, Wang HJ. Internet public opinion monitoring solution of TRS. Information Network Security, 2008, 6: 69–70.

9 Li XM, Zhu JJ, Yan HF. A collection, processing model and application of theme information on the Internet. Computer Research and Development, 2003,40(12): 1667–1671.

10 Qiu J, Liao LJ. Early warning technology research on public opinion and network culture security. Information Network Security, 2008,6:59–61.

11 李善平,尹奇,胡玉杰,等.本体论研究综述.计算机研究与发展,2004,41(7):1041-1052.

12 史一民,李冠宇,刘宁.语义网服务中的本体综述.计算机工程与设计,2008,29(23):5976–5979.

13 李宝敏,张娜.一种有效的本体创建方法——面向对象法.计算机技术与发展,2008,18(10):34–39.