# 2015年TML系统详细设计

孟涛

2015/3/13

## 产品简介

2015年南京网感的第一款产品方向是“股市负面头条”。同时2B与2C。

### 系统架构

系统架构如下图所示：



其中红色背景的字体是有人工参与的部分，包括编辑、用户、管理员等三类角色。

### 主要模块

已经完成的模块包括：

TML服务器集群：进行股票负面因素匹配和股票价格预测的后台引擎；

TML流水线：给TML机群提交文档，将抽取出来的结果存入数据库和索引；

爬虫系统：获取新闻网站、元搜索、微博和微信的数据；

TML开发环境：编写股票负面因素相关的知识规则；

有待实现和改进的新增模块包括：

WEB知识库：整理和存储三千家沪深两市公司的信息，包括股票名称、企业名称、品牌名称、上下游企业、行业等；

后台管理系统：对每天新增的股票头条新闻进行人工审核，然后发布；

股票业务知识库：基于TML开发环境来开发，一方面，建立网站来整理所有股票相关的事件类型与新闻，另一方面，对每一个事件类型编写TML知识规则；

股票业务分析引擎：针对每天TML引擎新匹配的股票新闻导出来，分析每一篇对哪些股票会有影响，按照2B和2C的需求将结果存入数据库；

股票APP服务API：基于Django的APP数据服务接口，在TML-WEB中添加；

手机APP：IOS和ANDROID端的两款APP；

将TML应用于推荐负面新闻相关的分析流程为：

1. 整理股价负面消息种类（如附件1），并将其转化为股票相关的TML规则；
2. 整理股票和上市公司的知识库，确定每一只股票的所在行业和竞争对手、企业领导和企业品牌、产品；
3. 采集沪深两市股票与这些负面关键词相关的企业/品牌和行业新闻，与TML模型匹配，记住都有哪些股票都有哪些类别的负面新闻，并记录日期；
4. 将股票历史日期的负面新闻特征，与数据库中的股票当天/三天/一周/一月/三月/六月/一年价格涨幅进行关联，建立股票的负面新闻特征影响股价的语料库，使用TML建立分类模型，统计每种特征的贡献权重，并统计特征与每只股价下跌的相关性；
5. 新采集当天的股票新闻，对每一篇新闻都与TML模型相匹配，然后根据输出的负面新闻特征，运用上述分类模型，输出股票下跌的概率，并输出匹配的权重最高的所有负面特征；
6. 将新闻按照下跌概率（也就是属于下降类别）的大小排序，与重要负面特征一起传入数据库供审核；
7. 编辑审核之后的新闻，企业用户和股民都可以按照上述方式阅读操作，并查出该新闻的负面新闻特征最相关的已经跌过的股票。

## 数据存储设计

存储在倒排中的数据包括：

网页的TML匹配结果；

存储在关系数据库中的数据包括：

数据库1：用户信息与行为数据；

数据库2：每只股票的新闻及相关文本和传播信息（传播；评论等）；

数据库3：与每篇新闻相关的股票分析结果；

## 数据分析设计

对股票的离线分析包括以下子任务：

1. 获取某些特定的重大财经政策事件及其时间，比如最近三年的利率变化，准备金变化等，分析它们与股票板块和个股的相关性；
2. 获取每只股票最近三年的负面新闻，通过TML分析之后，把负面新闻的种类和量存在数据库中，统计它们与股票板块和个股的相关性；
3. 针对（2）中的历史数据结果建立分类语料库，将股票特征对是否价跌的影响建立分类模型，并输出最强的特征；

三类离线分析的结果都需要存在上一节的数据库3中。

对股票的在线分析包括以下子任务：

1. 来一篇新的新闻，抽取它的负面特征，找到具有类似特征的历史新闻；
2. 分析类似历史新闻与股票板块和个股的相关性统计结果，作为佐证；
3. 利用分类模型预测是否会跌，将最强的特征一块输出。

## 附1：影响股票的负面消息类型

整理影响股票的负面消息种类如下：

1. 企业的产品和市场宣传言论与实际市场年报长期不符合，例如号称的产品研发实际未能进行或是销售业绩实际未能达标，让人觉得董事会不靠谱；
2. 企业的销售业绩出现问题；
3. 企业的核心员工离职导致研发中断或是表现出相关的产品失败；
4. 企业分支机构关闭和业务收缩；
5. 经济增长率，利率，通胀率高低有负面影响；
6. 外贸顺差、逆差、汇率有负面影响；
7. 行业不景气；
8. 美联储、欧洲央行货币政策有负面影响；
9. 国家石油价格、黄金价格、天然气价格有负面影响；
10. 加息、降息有负面影响；
11. 存款准备金率上调下涨有负面影响；
12. 行业政策收紧有负面影响；
13. 国家政策对行业有特定负面影响；
14. 竞争对手太强导致市场占有率失去；
15. 公司净资产大幅减少；
16. 公司盈利水平大幅降低；
17. 公司董事会增资或减资；
18. 新产品开发能力不足，比如长期没有新产品入市；
19. 长期没有收购；
20. 竞争对手技术突破、人员加盟；
21. 庄家和机构占股的稳定性；
22. 自然灾害影响某些行业内的生产；
23. 国内外战争或政治纠纷影响某些行业或企业；
24. 反腐政策；
25. 与台湾、香港、日本、美国关系影响某些行业或企业；
26. 公司的财务是否有瑕疵或是造假，或是行业普遍有瑕疵；
27. 公司是否资不抵债；
28. 公司是否虚报收购、纵容内部交易；
29. 高估营收和利润；
30. 伪造销售数据；
31. 财务造假；
32. 高层变动；
33. 夸大业绩，，夸大营收数据；
34. 虚构客户；
35. 利润率造假；