基于舆情挖掘的策略分享

——2014夏季主动量化及期权会议研究之九

证券分析师
陈杰　A0230513080006
2014.6

SWS RESEARCH

主要内容

1. 背景介绍
2. 系统构架
3. 策略介绍
4. 结果与进度

1、背景介绍

- 目前为止，我们所进行策略开发利用的数据基本上都是那些以数字形式存放在标准数据库中的结构化数据，而这部分数据仅仅占了所有金融信息中的一小部分，而金融信息中绝大部分的数据均是以文本的形式存在的一种非结构化的数据。

- 众所周知，财经新闻、股票论坛、股票研究报告等媒介信息能够迅速地反应投资者的投资情绪，同时也会对投资者的投资策略起到潜移默化的影响。这些文本信息中蕴含了大量未被利用的信息，如果能够合理地利用这些信息，我们相信这将为我们理解市场提供一个全新的角度。

1、背景介绍

构架

数据获取与清洗

1. 网站获得文本信息（股吧、微博、研报）
2. 关键指标存储数据库（MySQL）
3. 全文索引的建立

选股模型

1. 情感择时模型
2. 关注度选股模型
3. 事件驱动选股模型

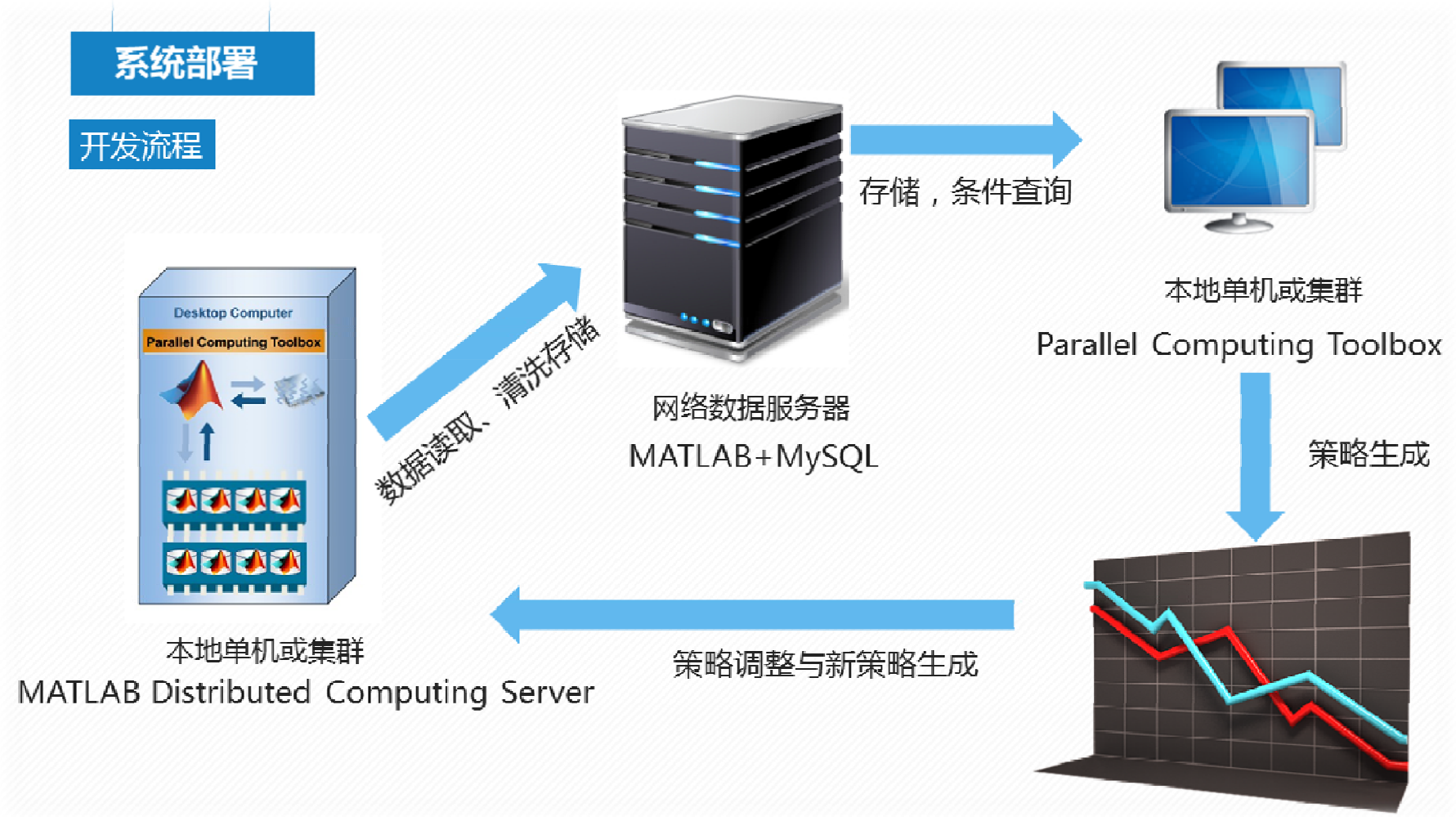
一般策略

1. 核心关键词的发现（全股票、单只股票）
2. 语义的判断（核心算法）
3. 指数生成（关注热度、情感、...）

主要内容

1. 背景介绍
2. 系统构架
3. 策略介绍
4. 结果与进度

2. 系统构架


系统部署
开发流程
Desktop Computer
Parallel Computing Toolbox
本地单机或集群
MATLAB Distributed Computing Server
数据读取、清洗存储
网络数据服务器
MATLAB+MySQL
存储，条件查询
本地单机或集群
Parallel Computing Toolbox
策略生成
策略调整与新策略生成

2. 系统构架

数据来源

- （1）股吧（已经实现）
- （2）论坛
- （3）新浪微博
- （4）研究报告
- （5）财经新闻

采集频率

- （1）当天可以循环读取整个股吧的全部股票评论
- （2）针对单只股票，可在分析前直接快速读取内容
- （3）对于历史数据，直接读取MySQL数据库上文件

代码编写

- 目前已经完成MATLAB代码编写，并在网络服务器中建立MySQL数据库，通过MATLAB读取数据库评论相关数据。

相关问题

- （1）网络数据库支持远程写入，但是采集需要计算机资源
- （2）采集代码用MATLAB编程，可以转变成C++
- （3）不同来源的结果如何综合，权重如何处理？

主要内容

1. 背景介绍
2. 系统构架
3. 策略介绍
4. 结果与进度

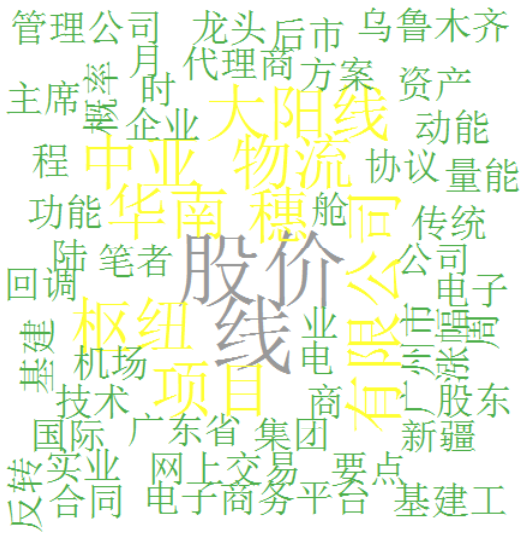
3.1.1关键词策略概念

策略1：指数策略

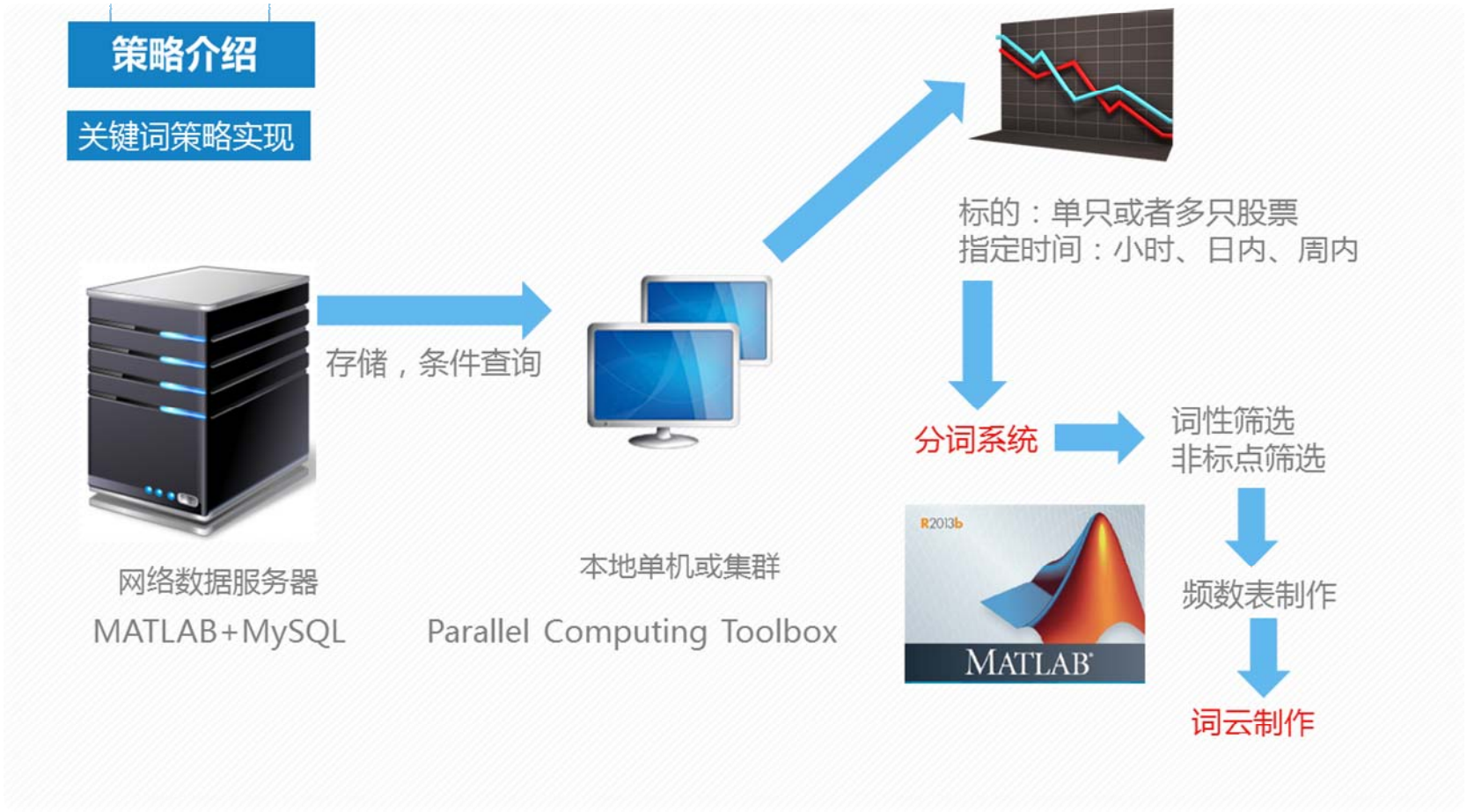
- 在关键词指数数据引擎中，我们可以从选定的文本资料源中搜索任意的关键词或者词语组合，提取包含有该关键词或者关键词组合的文本资料。然后按周统计搜索得到的文本资料数目，将其除以本周总的文本资料数目得到各周关键词指数。通过该数据引擎，我们可以对市场上对于该方面的关注的值进行量化，同时通过该数据引擎，我们也可以迅速的了解历史上市场对于该方面的关注程度值及目前所处的历史位置。

策略2：词云策略（关键词预警）

- 在数据库中搜索当周、当日、各小时的评论进行分析，制作出词云，可以分析全部股票或者单只股票的评论热点词，并对常见的拐点类词汇已经预警。



3.1.2关键词策略构架介绍



3.1.3关键词策略核心介绍

核心1：分词系统

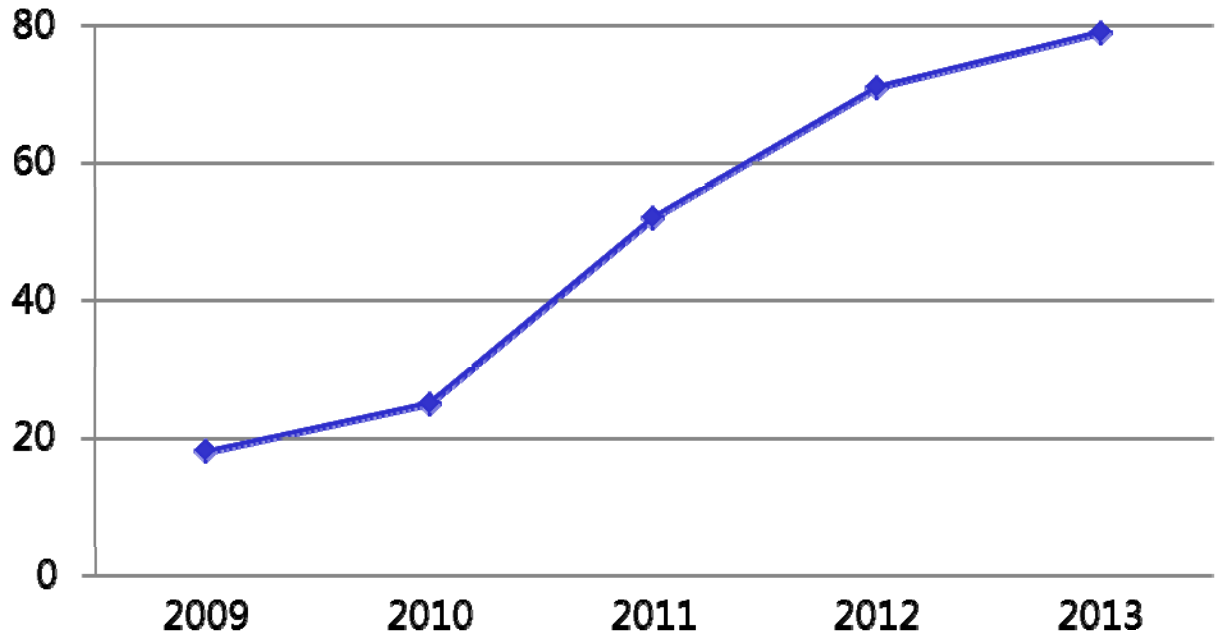
- 本策略分词系统是关键，其需要具备的特征为：
- 1 词库完整，支持金融词汇
- 2 支持自定义词汇库
- 3 支持高强度分词，分词速度快
- 4 能提供分词后的词性

核心2：词云系统

- 针对个股或者全部股票的词云系统，其需要具备的特征为：
- 1 能根据词频设置词云大小
- 2 可以改变词云形状
- 3 快速词云系统，能适应不同分词后数据请求

3.2.1评论热点策略概念介绍

- 在数据库中搜索指定时间内的个股评论数与讨论数，对突然评论增加的个股，但其股价没有巨大波动的引起注意。



3.3.1情感词策略概念

- 对文本资料进行情感判别并打分，其中情感判别分为两类：乐观与悲观，正的分值对应乐观情绪，分值越高情绪越乐观；负的分值对应悲观情绪，分值越低情绪越悲观。

策略1：普通投资者情绪

- 搜索所有评论和发帖，对每一条帖子进行情感识别得到其情感分值后，按交易日将所有的得分分类汇总，将所有乐观帖子和悲观帖子的分值分别加总，得到每一个交易日的乐观总分（定义为看多值）和悲观总分（定义为看空值）。将看多值除以看空值，然后按20个交易日平滑得到普通投资者多空指数。

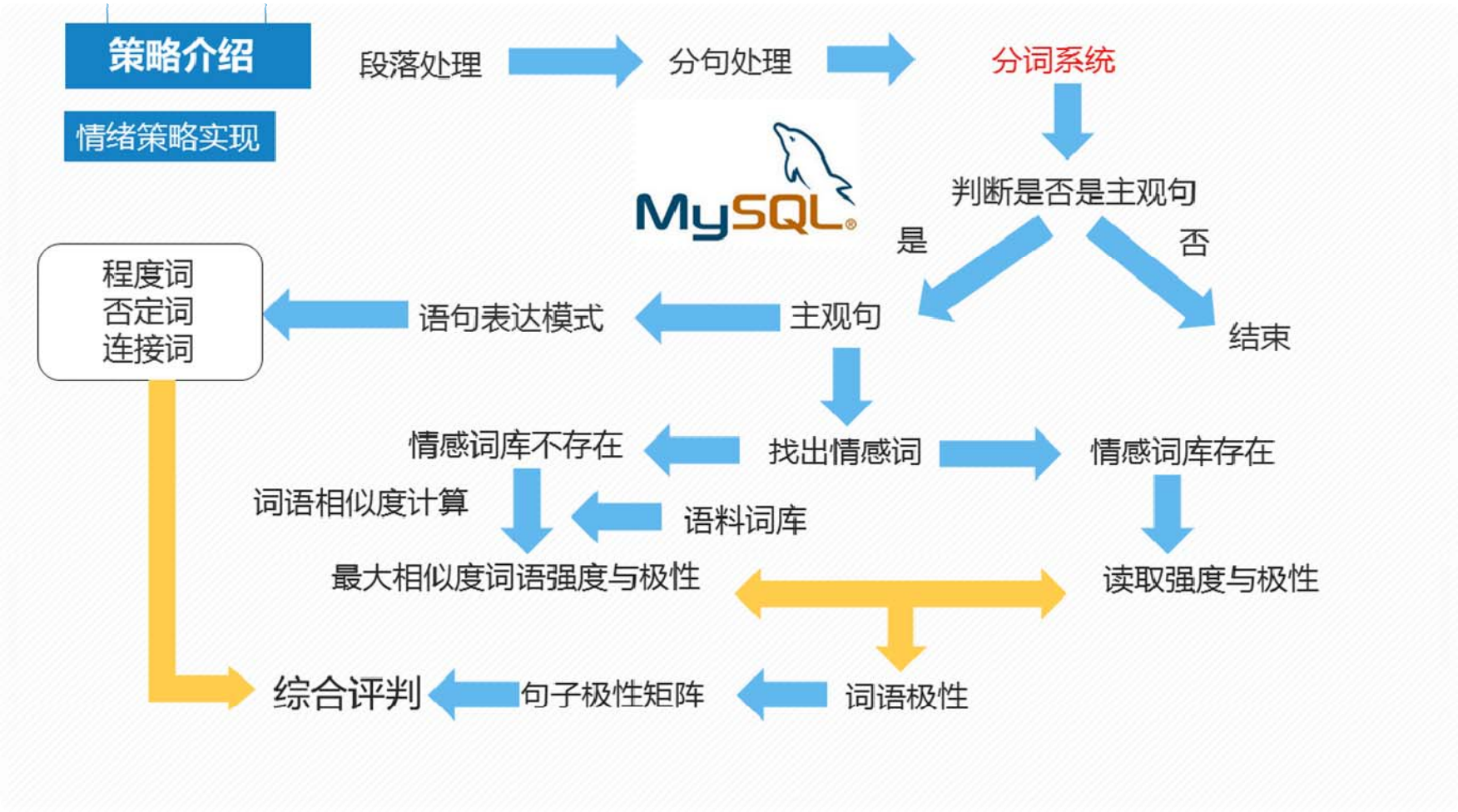
策略2：普通投资者偏好

- 针对个股，选择情感识别情感分数最为正向的股票作为股票池筛选指标，同时根据每一只股票情感强度走势与股价走势进行对比，建立非线性回归模型，探索其中的相关关系。

策略3：机构多空情绪

- 对每篇研究报告进行情感识别得到每篇报告的情感分值后，同样的，我们按交易日将所有的报告分类汇总，将所有乐观报告和悲观报告的分值分别加总，得到每一个交易日的乐观总分（定义为看多值）和悲观总分（定义为看空值）。将看多值除以看空值，然后对其求对数值，将结果按四周平滑得到机构多空指数。

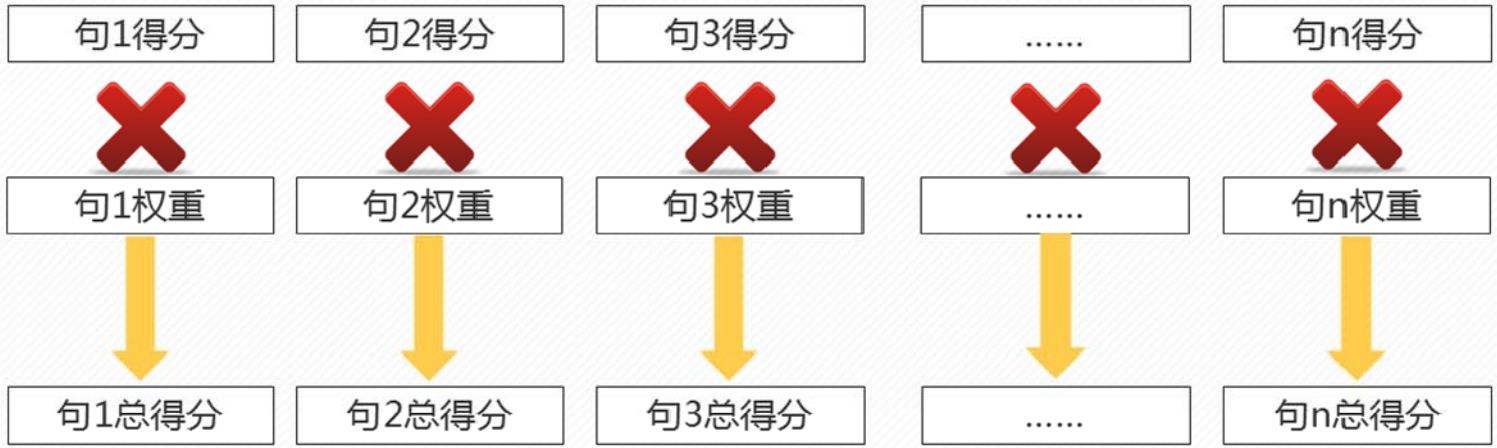
3.3.2情感词策略系统构架



3.3.3情感词策略策略实现

策略介绍

情绪策略实现

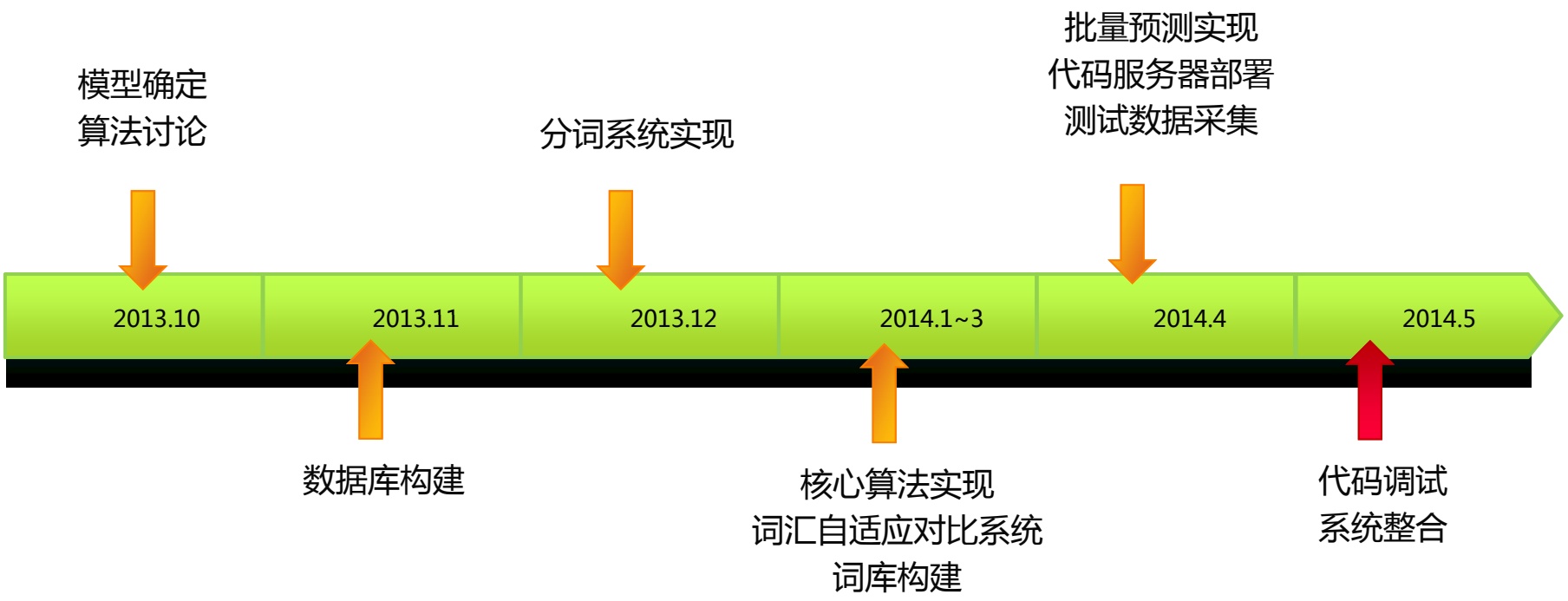


句子权重矩阵的设计：重点的话，后面说。

主要内容

1. 背景介绍
2. 系统构架
3. 策略介绍
4. 结果与进度

4 结果与进度


模型确定
算法讨论
分词系统实现
批量预测实现
代码服务器部署
测试数据采集
2013.10
2013.11
2013.12
2014.1~3
2014.4
2014.5
数据库构建
核心算法实现
词汇自适应对比系统
词库构建
代码调试
系统整合

4. 结果与进度

结果表明：

- 当前的词库和算法可以将语句的情感分析准确率达到近60%。

改进方向：

- 1 完善算法，包括主观句判断模式
- 2 改写成Java
- 3 代码部署与并行计算

后续任务：

- 1、建立策略进行历史回测，挖掘舆论与标的的非线性关系；
- 2、代码优化，如何快速部署与实现实时数据分析与结果呈现；
- 3、自动预警功能
- 4、词汇索引库管理

信息披露

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告,并对本报告的内容和观点负责。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

与公司有关的信息披露

本公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可，资格证书编号为：ZX0065。发布证券研究报告，是证券投资咨询业务的一种基本形式，本公司可以对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向本公司的客户发布。

本公司在知晓范围内履行披露义务。客户可通过compliance@swsresearch.com索取有关披露资料或登录www.swsresearch.com信息披露栏目查询从业人员资质情况、静默期安排及关联公司持股情况。

法律声明

本报告仅供上海申银万国证券研究所有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

本报告首页列示的联系人，除非另有说明，仅作为本公司就本报告与客户的联络人，承担联络工作，不从事任何证券投资咨询服务业务。

客户应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司http://www.swsresearch.com网站刊载的完整报告为准，本公司并接受客户的后续问询。

客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为作出投资决策的惟一因素。客户应自主作出投资决策并自行承担投资风险。本公司特别提示,本公司不会与任何客户以任何形式分享证券投资收益或分担证券投资损失，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。市场有风险，投资需谨慎。

若本报告的接收人非本公司的客户，应在基于本报告作出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。

申万研究 • 拓展您的价值

SWS Research • CHINA Value Revealed

上海申银万国证券研究所有限公司

陈杰
chenjie@swsresearch.com