结题报告新闻信息分析处理及展示部分

**1. 深度学习**

深度学习的概念由Hinton等人于2006年提出，其前身为神经网络。2016年DeepMind公司所开发的AlphaGo击败围棋世界冠军李世石之后，深度学习这一概念逐渐被人认可。通过深度学习技术，可有效地处理机器翻译、语音识别、图像分类以及自然语言处理等机器学习的问题。深度学习的基本原理是通过神经网络结构进行低维到高维的映射，从而能够对数据的特征进行提取，最后能够得到数据的特征表示。典型的深度学习模型见下图1-1。

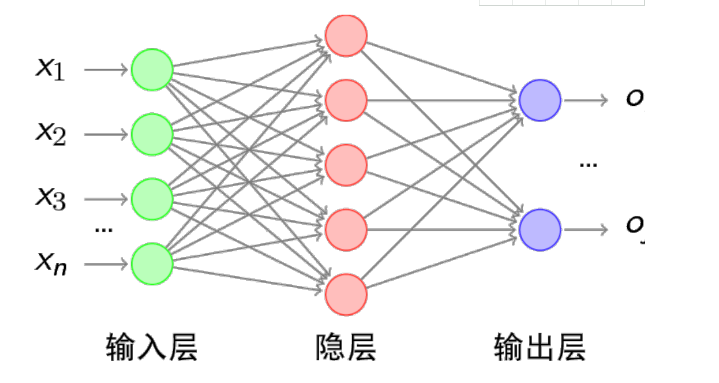


图1-1 深度学习模型示例图

深度学习本质上是构建含有多个隐含层的机器学习架构模型，通过大规模数据进行训练，得到大量更具代表性的特征信息。从而对样本进行分类和预测，提高分类和预测的精度。这个过程是通过深度学习模型的手段达到特征学习的目的。深度学习模型和传统浅层学习模型的区别在于：

1. 深度学习模型结构含有更多的层次，隐含层节点的层数通常在5层以上，有时甚至包含多达10层以上的隐藏节点；
2. 明确强调了特征学习对于深度模型的重要性，即通过逐层特征提取，将数据样本在原空间的特征变换到一个新的特征空间来表示初始数据，这使得分类或预测问题更加容易实现。
3. 和人工设计的特征提取方法相比，利用深度模型学习得到的数据特征对大数据的丰富内在信息更有代表性。

深度学习算法打破了传统神经网络对层数的限制，可根据设计者需要选择网络层数。它的训练方法与传统的神经网络相比有很大区别，传统神经网络随机设定参数初始值，采用BP算法利用梯度下降算法训练网络，直至收敛。但深度结构训练很困难，传统对浅层有效的方法对于深度结构并无太大作用，随机初始化权值极易使目标函数收敛到局部极小值，且由于层数较多，残差向前传播会丢失严重，导致梯度扩散，因此深度学习过程中采用贪婪无监督逐层训练方法。即在一个深度学习设计中，每层被分开对待并以一种贪婪方式进行训练，当前一层训练完后，新的一层将前一层的输出作为输入并编码以用于训练；最后每层参数训练完后，在整个网络中利用有监督学习进行参数微调。

深度神经网络的训练过程主要是先逐层预训练初始化深度神经网络每一层的参数，然后通过监督学习对整个深度结构神经网络进行微调。

训练过程如下：

1. 将第一层作为一个自编码模型将第一层作为一个自编码模型，采用无监督训练，使原始输入的重建误差最小；
2. 将自编码模型的隐单元输出作为另一层的输入；
3. 按步骤2迭代初始化每一层的参数；
4. 采用最后一个隐含层的输出作为输入施加于一个有监督的层（通常为输出层），并初始化该层的参数；
5. 根据监督准则调整深度结构神经网络的所有参数。

**2. 深度学习算法框架**

深度学习算法的结构通常非常复杂，包含上百兆的可训练参数。比如2016年由He K等人提出的ResNet网络，最大的一种模型包含了上千层。深度学习算法的复杂网络结构是由有限的基本算子通过各种复杂连接关系搭建起来的，常用的算子包括卷积、池化、全连接等。因此，可以提取出机器学习算法中共性的部分，将其抽象出来，便于反复调用。

**2.1 基于数据流图的搭建框架**

基于数据流图（data flow graph）的机器学习编程框架利用节点（node）和边（edge）构造的有向图来描述计算过程。节点可以表示一个运算操作，或者表示一块数据的输入起点或者输出终点，边则表示节点之间的输入/输出关系。数据被表示为多维数组（张量）的形式，可以在这些边上进行传输。通过一个节点时，数据就会作为该节点运算操作的输入被计算，计算的结果则顺着该节点的输出边流向后面的节点。一旦输入端的所有数据准备好，节点将被分配到各种计算设备，完成异步并行的执行运算。

**2.2 基于层的搭建框架**

基于层的编程框架为用户提供一组表示各种层（比如卷积层、池化层、全连接层等）的函数作为接口。用户通过反复调用这些层的函数接口构成网络。通过将计算单位限制到层上，库的开发者们可以对各种层进行充分的性能优化，因此可以提供更好的运行效率。Caffe是一种常用的卷积神经网络框架，它采用层的调用方式，用一个prototxt文件对每一个层进行定义和配置，之后程序分析这个文件，获得这个网络结构的信息。

**2.3 基于数据流图的搭建框架**

基于算法的编程库提供了大量的机器学习算法，涵盖各种任务和算法。和前两种框架不同，基于算法的编程库不需要构建复杂的网络结构，只需要通过设置库里面提供的算法函数接口中的参数即可完成任务。比如Scikit-Learn机器学习库，它是一种基于Python语言的机器学习库，提供了数据分析和数据挖掘中针对各种任务（包括分类、聚类、数据降维、模型选择、预处理等）的算法，提供了各种不同类别（如k-NN、k-means）的决策树（如C4.5、ID3等）的算法接口，开发者只需要设置接口中的参数，并且将数据传入，就可以得到训练和预测的结果。

**2.4 Google BERT框架**

BERT的全称是Bidirectional Encoder Representation from Transformers，即双向Transformer的Encoder。模型的主要创新点都在pre-train方法上，即用了Masked LM和Next Sentence Prediction两种方法分别捕捉词语和句子级别的representation。BERT模型的结构如下图所示：

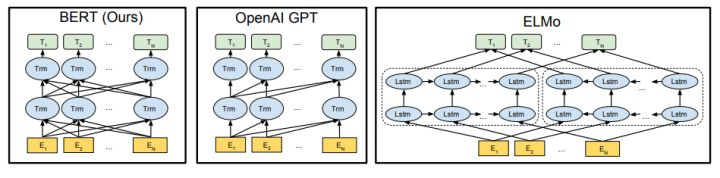


图2-1 BERT模型结构图

**2.4.1输入的表示**

针对不同的任务，模型能够在一个token序列中明确地表示单个文本句子或一对文本句子(比如[问题，答案])。对于每一个token, 其输入表示通过其对应的token embedding, 段表征(segment embedding)和位置表征(position embedding)相加产生。图2-2是输入表示的直观表示：

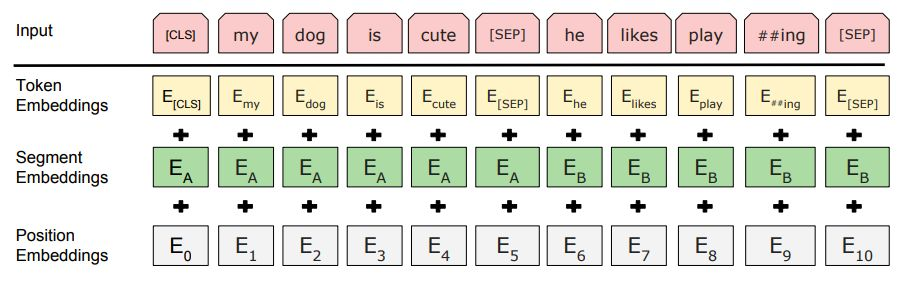


图2：BERT输入表示

具体实现细节有以下几个部分：

1. 论文使用了WordPiece embeddings含有30000个token，并且使用“##”来拆分词片段
2. 使用了positional embeddings， 长达512位，即句子的最大长度
3. 每句话的第一个token总是[CLS]。对应它的最终的hidden state(即Transformer的输出)用来表征整个句子，可以用于下游的分类任务
4. 模型能够处理句子对。为区别两个句子，用一个特殊token [SEP]隔开它们，另外针对不同的句子，把学习到的Segment embeddings 加到每个token的embedding上
5. 对于单个句子仅使用一个Segment embedding

**3. Flask web框架**

Flask 是一个轻量级的基于python的web框架。Flask建立于一系列的开源软件包之上，这其中 最主要的是WSGI应用开发库Werkzeug和模板引擎Jinja。Flask是一个WSGI应用框架，在进行Flask开发时，不需要关注网络方面的操作，Flask应用的入口是封装过的网络请求包，出口是网络响应，在使用flask框架时仅需要关注这个阶段内的处理逻辑。

本课题在网页展示方面采用了flask框架设计搭建网页。

**4. spark大数据处理**

Spark是基于内存计算的大数据并行计算框架，可用于构建大型的、低延迟的数据分析应用程序。本课题为了应对每天高流量的数据流采用了spark大数据计算框架来实现对于高密度数据的处理。Spark计算框架主要有如下几个特点：

* 运行速度快：Spark使用先进的DAG（Directed Acyclic Graph，有向无环图）执行引擎，以支持循环数据流与内存计算，基于内存的执行速度可比Hadoop MapReduce快上百倍，基于磁盘的执行速度也能快十倍；
* 容易使用：Spark支持使用Scala、Java、Python和R语言进行编程，简洁的API设计有助于用户轻松构建并行程序，并且可以通过Spark Shell进行交互式编程；
* 通用性：Spark提供了完整而强大的技术栈，包括SQL查询、流式计算、机器学习和图算法组件，这些组件可以无缝整合在同一个应用中，足以应对复杂的计算；
* 运行模式多样：Spark可运行于独立的集群模式中，或者运行于Hadoop中，也可运行于Amazon EC2等云环境中，并且可以访问HDFS、Cassandra、HBase、Hive等多种数据源。

Spark的基本运行流程如下图4-1所示：

1. 当一个Spark应用被提交时，首先需要为这个应用构建起基本的运行环境，即由任务控制节点（Driver）创建一个SparkContext，由SparkContext负责和资源管理器（Cluster Manager）的通信以及进行资源的申请、任务的分配和监控等。SparkContext会向资源管理器注册并申请运行Executor的资源；
2. 资源管理器为Executor分配资源，并启动Executor进程，Executor运行情况将随着“心跳”发送到资源管理器上；
3. SparkContext根据RDD的依赖关系构建DAG图，DAG图提交给DAG调度器（DAGScheduler）进行解析，将DAG图分解成多个“阶段”（每个阶段都是一个任务集），并且计算出各个阶段之间的依赖关系，然后把一个个“任务集”提交给底层的任务调度器（TaskScheduler）进行处理；Executor向SparkContext申请任务，任务调度器将任务分发给Executor运行，同时，SparkContext将应用程序代码发放给Executor；
4. 任务在Executor上运行，把执行结果反馈给任务调度器，然后反馈给DAG调度器，运行完毕后写入数据并释放所有资源。

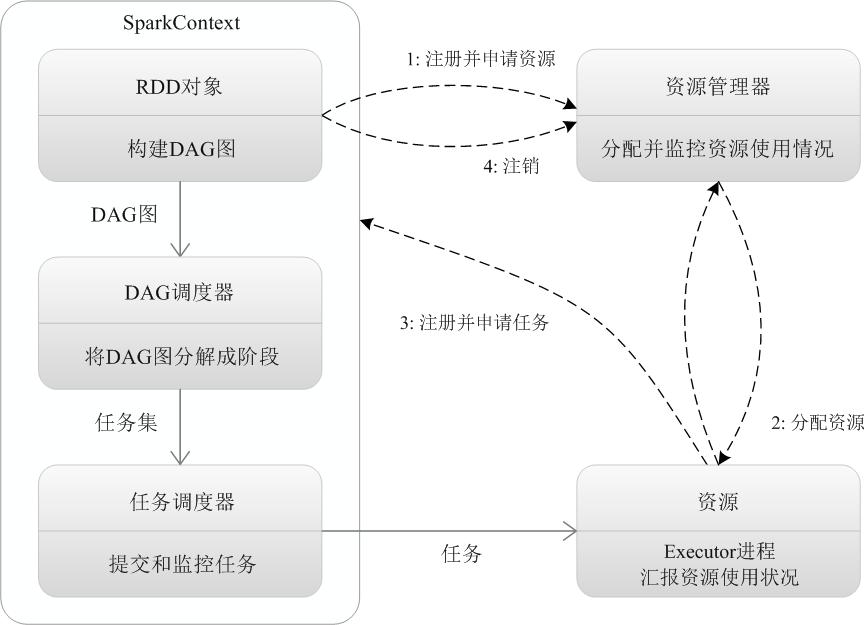


图4-1 spark运行流程图

**5系统功能实现**

系统功能实现主要分为如下图所示四个部分：

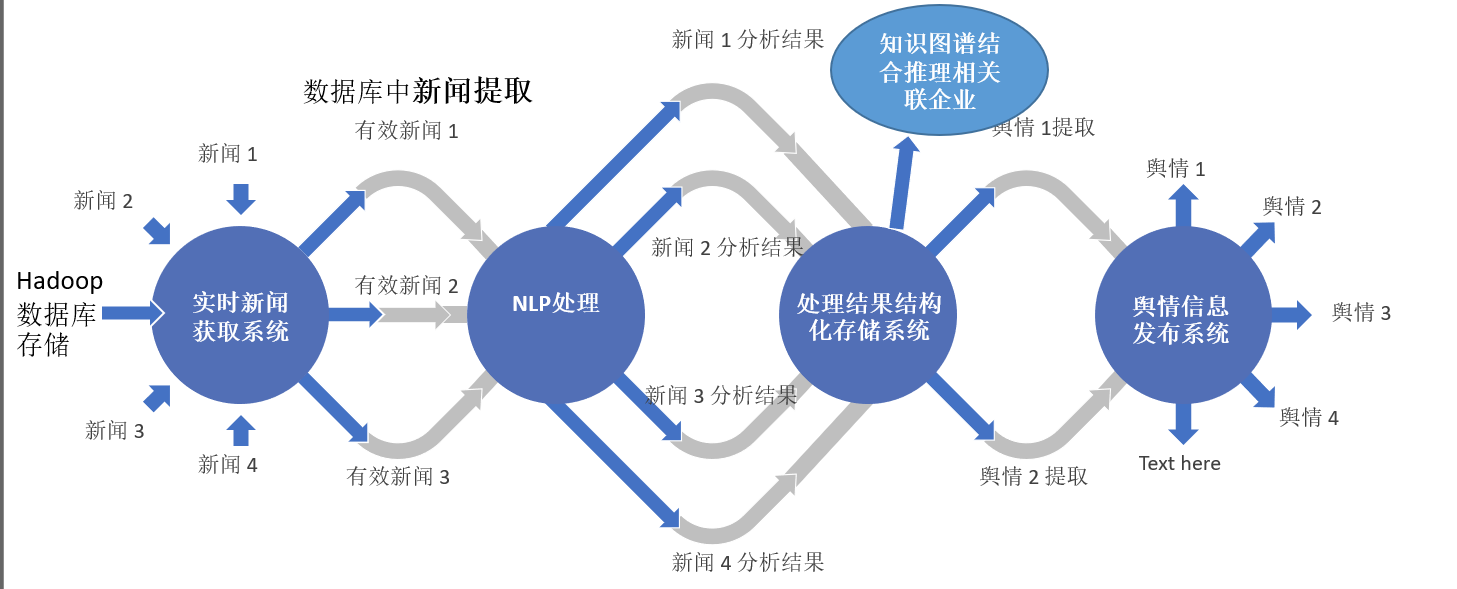


图5-1 系统主要模块图

1. 实时新闻获取系统：此系统通过获取外界新闻并对新闻信息进行存储分发筛选，提取其中有效的新闻分发到下一系统中进行NLP分析处理。
2. NLP处理阶段通过预先给出的数据新闻以及专家提供的风险类型标签，利用Bert深度神经网络训练得出判别模型；NLP处理阶段为核心技术对于新闻中出现的相关企业、证券、股票或行业进行采用关键词向量匹配算法提取；NLP处理阶段为核心技术对于新闻中出现的相关企业、证券、股票或行业进行采用关键词向量匹配算法提取
3. 存储系统将处理结果以信息流的方式存储到Hadoop数据系统中，等待下一步结果信息提取
4. 舆情信息发布系统将信息发布到网页端，在网页端工作的业务人员可以第一时间了解到风险舆情并作出相应的处理

本课题采用知识图谱结合深度学习来对舆情做预警判断。知识图谱与舆情分析的结合体现在下图5-2中：根据分析所得结果判断相关企业或行业并得出风险类型；在知识图谱中以所获的企业或行业进行扩展推理得到与之关联的其他上下游企业

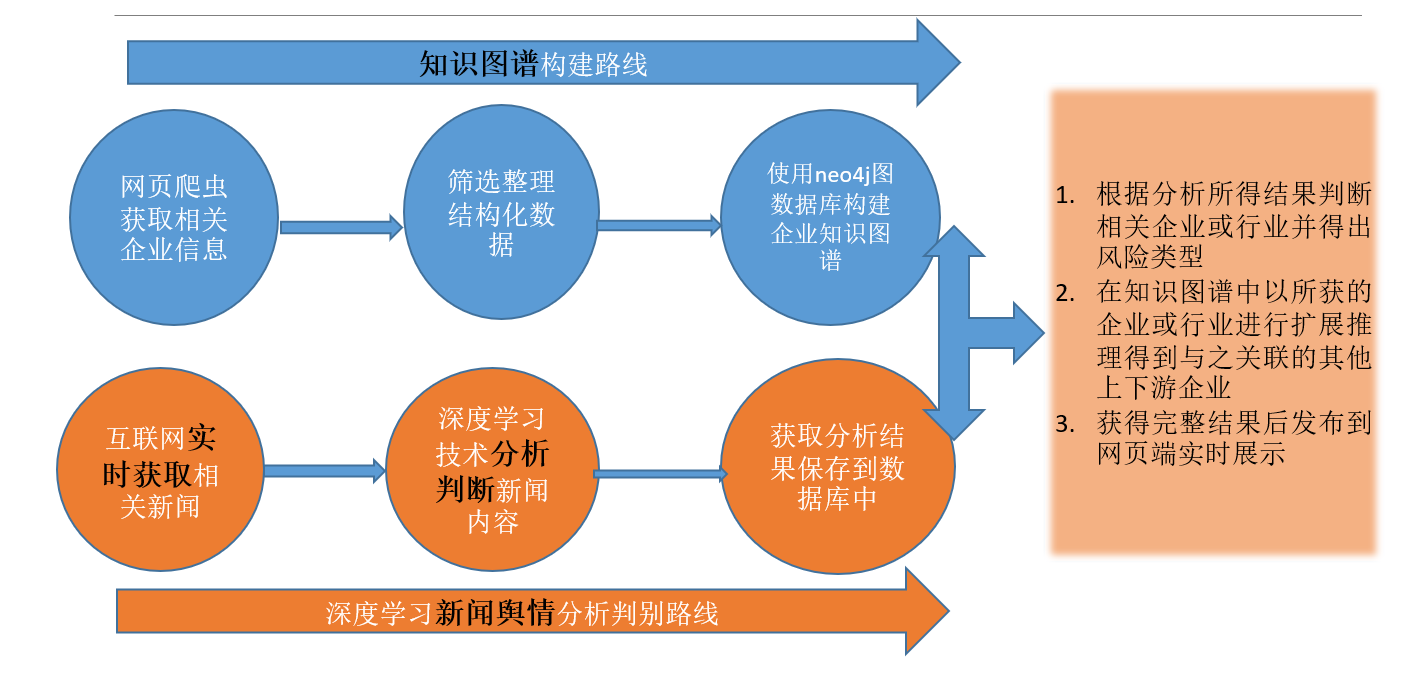


图5-2 知识图谱与舆情分析结合图

**6对于新闻的NLP处理分析**

**6.1企业主体识别**

企业主体识别采用关键词识别的方式进行分析提取。对于完整的标题做分词切词得到一列词列表，由企业知识图谱中导入的企业名称和企业别名等等信息对分词后的结果进行比对筛选最终获取到相对于的企业主体。

对于已经获取到的企业主体可通过知识图谱的推理计算得到相关联的行业、上下游企业、子母公司等。通过这些推理得到的内容可以由单个企业产生的舆情预警到多企业的预警。

**6.2新闻风险类型判断**

新闻风险类型判断是指通过对新闻标题及内容进行深度学习的模型计算得到相对于的风险类型。在本课题中，通过专家给出的建议，风险类型大致分为如下类型：

1. 信用风险；
2. 市场风险；
3. 经营风险；
4. 法律风险；
5. 高管风险；
6. 子母公司风险；

在前期BERT模型预训练时，通过对于给定新闻标题的风险类型标注得到的预训练数据进行训练可以得到分类精度足够准确的模型。后期在对于新闻标题的风险类型判断中。只需将新闻标题输入到模型中就可以在很短的时间内得到相对于的风险类型。

**6.3新闻风险等级判断**

新闻风险等级判断采用关键词分析的方法来处理，由专家针对于每种风险类型给出相对应的多个关键词，并对不同关键词出现给与不同等级权重。

在经过风险类型判断后得到相应的风险类型并根据指定的风险类型关键词来对新闻标题做匹配分析，根据不同关键词出现的频率和相应的权重做加权计算得到相对应的风险等级。

**7主要功能展示**

**7.1新闻滚动显示及分析结果显示**

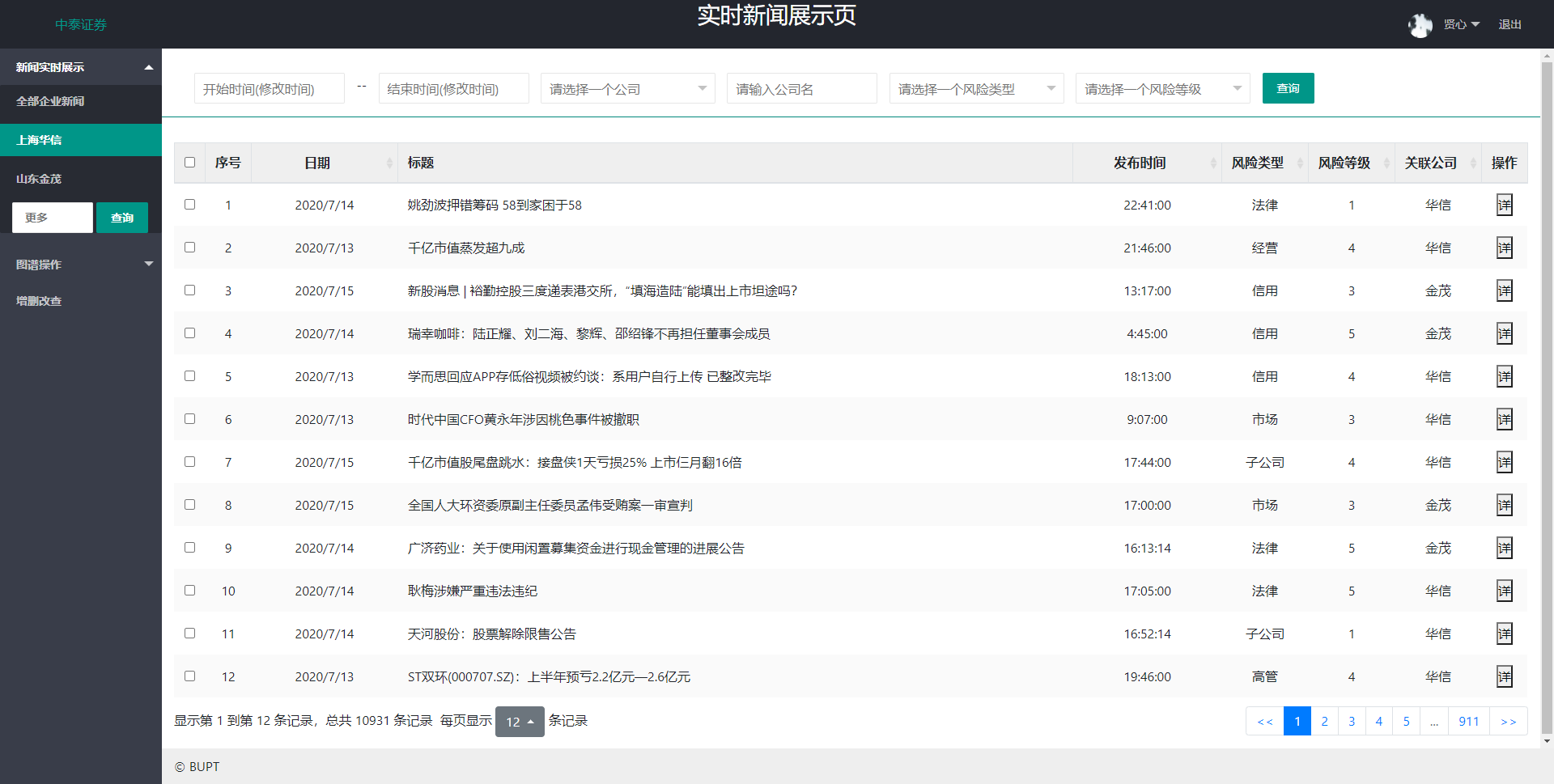


图7-1

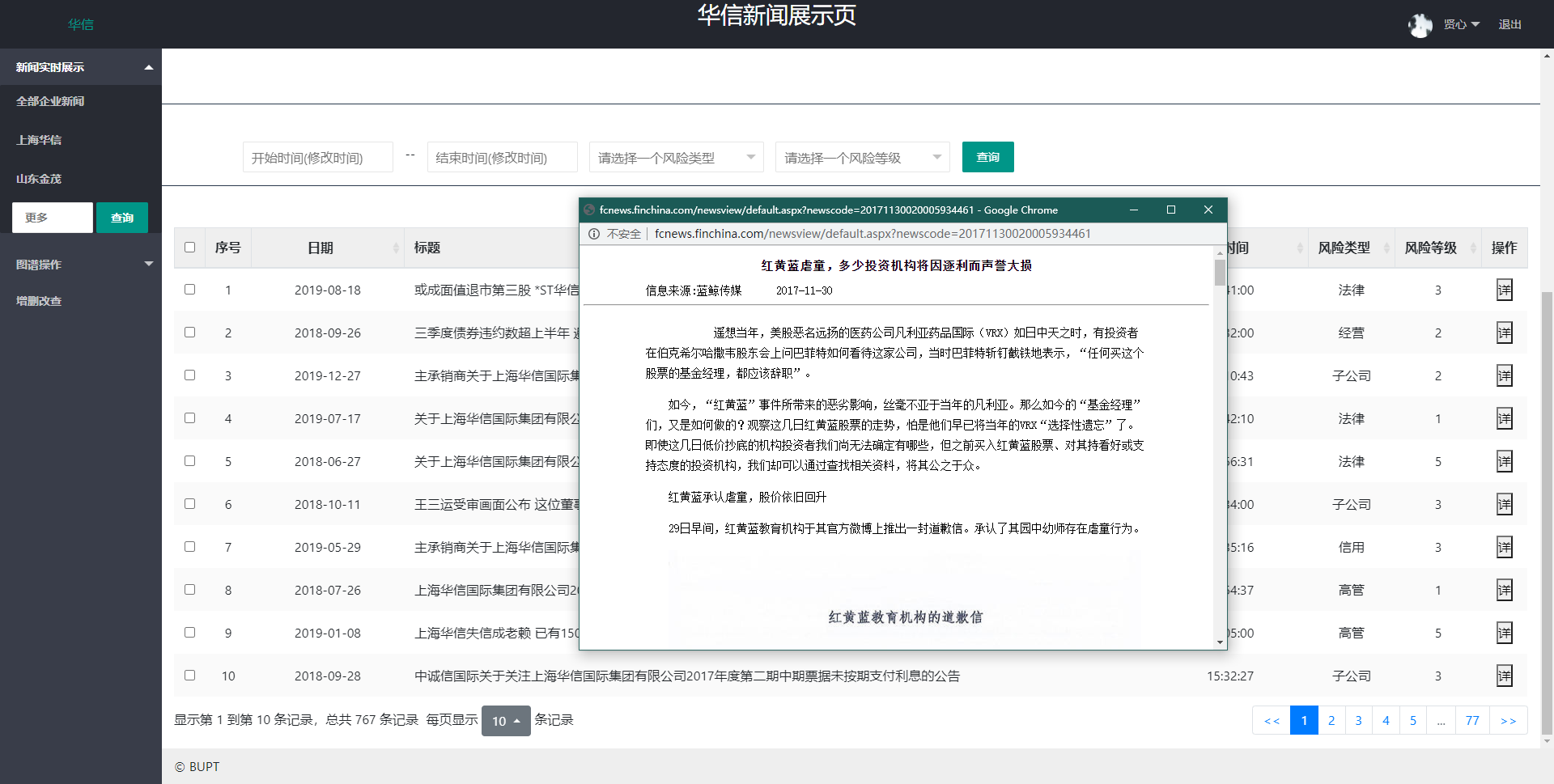


图7-2新闻标题实时滚动显示页面

新闻标题滚动页面图7-1和图7-2是综合显示系统所获取到的新闻，并实时更新，在新闻获取系统获取到了新的有效的新闻后会动态的插入到新闻列表的首条。此页面列表展示的新闻主要有新闻标题、新闻发生的日期及时间、通过BERT深度学习预测的对应新闻的风险类型、通过关键词匹配识别得到的相关企业主体以及当前风险所对应的风险等级。最后一列的详情按钮可小窗弹出完整的新闻内容。

滚动新闻展示的上方有多个可供选择的下拉框及输入框。通过下拉框及输入框选择的值可以对新闻列表进行筛选，找到所关心关注的新闻。可供选择的内容有日期选择范围、公司企业主体、风险类型以及风险等级。

**7.2单个企业主体新闻滚动显示**

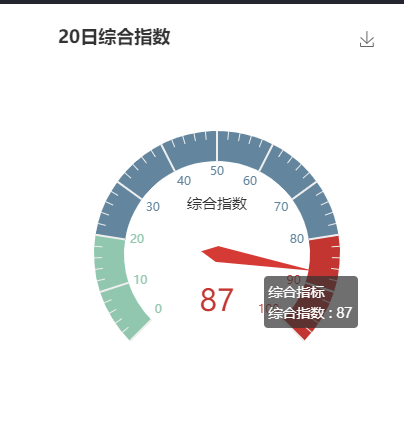


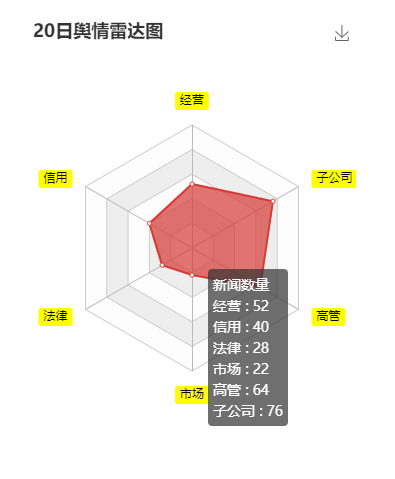
图7-3 单个企业主体新闻滚动显示

单个企业主体界面是在全部企业新闻信息滚动显示界面的基础之上提取其中所关注的当前企业的有关新闻进行展示。另外针对于单个企业系统给出了更多的分析结果及指标。

在页面底部是类似于全部企业新闻信息滚动展示界面的新闻滚动显示，区别在于取消了相关联企业的显示及筛选选择。在上方依次展示的是企业当天当前时刻所累积的舆情综合值数值，企业当天发生的不同风险类型时间的数量统计雷达图显示，以及企业当月每天的三种不同的舆情值柱状图显示。

下面三个图是针对于单个企业主体的实时分析结果图：

右图显示的是企业当天风险舆情综合值，这是对于企业的一个综合评价。对于对一个企业做综合评价至关重要。此数值是根据当天企业发生的风险事件由风险类型及风险等级做加权相乘求和得到的数值。用户可根据此值的大小来确定企业的风险大小。



左图显示的是当天企业不同风险类型新闻发生的数量统计，用户可以实时掌握企业主体在哪些方面出现了较多的舆情，在哪些方面需要重点关注和监控。

下图显示为每月企业每天的舆情指数值。将舆情指标转化为三种指标，分别是负面舆情、自身风险、关联风险。用户可以清楚的观察到企业主体每月的舆情分布趋势以便对与企业的了解和掌握。

