内部资料
仅供参考

传　导

（2019 年第 34 期，总第 511 期）

中国证券业协会　　2019 年 4 月 12 日

“推进新时代资本市场建设和证券业高质量发展”重点课题研究系列报告之十八

深度学习和知识图谱在智能公司监管中的应用研究

摘 要

随着近年来我国上市公司数量快速增加，证券市场规模持续扩大。但与此同时，市场新问题层出不穷，资本脱实向虚，杠杆高企，部分上市公司成为乱象的制造者，风险积聚。对此，上市公司一线监管迫切需要加强监控力度和敏锐度，及时发现市场主体的潜在风险，维护市场稳定，切实保护中小投资者利益。这就需要在传统监管模式的基础上，借助金融科技，寻求监管突破。

本文基于深度学习和知识图谱技术开展上市公司智能监管研究，主要技术路径和研究内容是：首先对上市公司进行多维度、

全上市生命周期画像，涵盖公司历史沿革、股东情况、关键人员、关联关系、财务运营、同业比较、重要交易、舆情股价、诚信档案和监管评价等全景信息。在此基础上，我们研究构建了上市公司知识图谱的风控数据维度体系，运用知识图谱技术构建企业知识图谱风险监控模型和初始的应用测试系统，并对近期发生的典型案例进行了模型测试，验证挖掘风险源和潜在的风险传导路径的效果。其次，通过总结证券市场建立以来上市公司风险分类、风险特征及其分布，并结合股东行为、合规运作、经营状况等关键信息，构建了基于人工神经网络和深度学习技术的上市公司财务失败风险监测预警模型，并遴选部分上市公司作为训练样本数据集与测试样本数据集进行模型训练和测试，测试结果表明模型具有较好效果。最后，由于上市公司负样本数量有限，我们结合专家经验继续进行了上市公司财务风险评价模型构建，并用近 5 年来除金融和房地产行业的上市公司做样本进行了财务风险评测，而后根据国际和国内公司风险分级标准，制定上市公司风险评级标准，对测试的上市公司近 5 年的风险进行了评级，形成了上市公司风险波动图谱。

本文研究表明，利用深度学习和知识图谱技术可以改进监管模式，提升监管穿透性，缓解监管时滞性，提高监管发现问题和防范风险的能力。

正 文

一、引言

现有公司业务管理系统在服务一线监管工作方面发挥了重要作用，但在支持科技监管方面仍有不足，主要体现在三个方面：一是数据基础薄弱带来监管乏力。一方面，系统中上市公司及其股东相关信息、历史数据等不全面、不完整，无法满足监管需要；另一方面，对上市公司的监管评价、监管过程数据及处分记录等散落在各处，缺乏有效整合，且与其他监管部门的信息共享程度不高。二是由于缺乏新技术支撑造成信息分析能力有限。现行监管模式大多采用统计报表、监管问询、手动搜索处理信息、人工识别处置等传统方式，面对数量巨大、来源分散、格式多样的公司数据，已逐渐出现不适应，机器处理分析数据的能力有待提高。同时，部分上市公司股权结构复杂、违规行为隐蔽，对监管人员准确识别风险隐患以及穿透核查带来了挑战，需要借助监管科技工具提升监管能力。三是欠缺实时监控和动态预警。目前，公司业务管理系统的智能化程度仍处于较低水平，在实时数据采集、实时数据计算、动态监测风险态势、及时发现预警问题等方面存在不足，技术上无法满足复杂的监管需求，不能充分发挥辅助监管的效用。

为了对上市公司的智能监管进行赋能，本文建立了相关的研究框架、监管技术选型以及理论研究，确定了本次研究的重点方向。

（一）技术发展趋势现状

1.财务监管

财务风险预测模型的发展经历了漫长的过程，主流的预测方法分为两大类别：一个是基于统计学和概率论的传统预测模型，另一类是基于人工智能的神经网络无参数预测模型。主要方法如下:

（1）单变量预警模型。该模型是最早运用在财务风险的相关研究中的方法，采用某一个财务指标，来对公司的财务风险状况进行评判和预警。最早利用该模型进行财务风险预警模型的研究是 FitZpatriCk(1932)，研究结果表明，市净资产收益率和企业杠杆两个财务指标是在财务风险的预测上，具有最高的准确率。而后,Beaver(1966)在自己的研究之中认为,现金流量/债务总额、净资产收益率（ROE）和资产负债率这三个财务指标，在财务风险的预测上具有很高的准确率。并且，距离公司的破产日越近，其判断的准确率越高。但是由于单变量模型逻辑上过于简单，难以取得实用效果，在之后的研究中逐渐被淘汰，成为了后续研究的基础。后来开始针对单变量研究之中寻找出的有效财务指标，进行多变量研究。

（2）多元线性模型。为了克服单变量模型单薄的缺陷，Altman(1968)开始使用多变量判别的方法，也就是著名的 Altman 奥特曼 Z 值统计法。具体就是建立一个多元变量模型，模型中涉及多个财务指标，采用多财务指标进行判别，并根据多个指标的

评价积分，得出一个最终的 Z 值作为总的判别标准，根据 Z 值的大小作为判断公司财务危机的阈值。在模型中，Altman 认为企业的变现、获利、资产规模、财务结构、运营等方面可以综合地反映一个企业的财务状况。但 Z 值模型也有一定的缺陷，在构建模型的财务指标当中，并没有包含现金流量的考量，而现金流量对于公司的财务安全又是至关重要的部分 。之后，Altman 再一次更新了 Z 值模型，更新后的 ZETA 模型包含了所选样本公司的利息保障倍数、资产收益率（ROA）、股东权益比、留存收益总资产比、流动比率等 7 项财务指标。通过对于研究结果的比较分析，表明 ZETA 模型相对于 Z 值模型具有更高的预测准确率。而在之后不断的应用过程中，Altman 又进行了两次修正。

总体来看，Z 值统计法由于其简便性和有效性，成为了应用领域中最为广泛的一种，被用于各样环境下的财务预警。

（3）多元逻辑回归模型。这类模型现在广泛运用在学术研究领域，最为典型的是 Probit 回归模型和 Logistic 回归模型。第一次利用 Logistic 模型来进行财务风险预测的是 Martin(1977)，当时 Martin 利用该模型对银行的财务失败进行了预测，最终结果表明 Logistic 模型具有很好的预测准确性。而后，Ohlson(1980)在这一领域进行了进一步的研究，在具体的指标上，他提出利用公司规模的大小、公司的股权结构、公司现有资产的变现能力以及公司的盈利和增长来对财务危机进行预警，可以获得更高的准确率，研究结果表明准确率达到了 96.12%。在 1985 年，BartCZak、

Norman 在自己的研究中对现金流量对于财务风险预测的作用进行了研究，结果显示经营现金流量的信息引入到财务风险的预测当中并没有提高模型预测的准确度，而还是应当以应计制的财务指标来进行模型构建。

（4）生存分析模型。根据以上的研究，财务失败的预警模型大部分都是建立在静态的截面数据之上的，忽略了时间序列维度上的变化对于财务风险的影响，因此实际上传统模型都属于静态预警模型。而现实中，财务失败的发生是一个渐进的过程，因而上市公司在不同的时间点上发生财务失败的概率是不一样的。生存分析模型的引入，就能够描述一个事件发生的概率是如何随时间变化的。在近 20 年来，该模型也被广泛地运用在关于风险评估的学术研究当中。过新伟在关于财务预警的研究中，比较了离散时间风险模型和 Prbit 回归以及 Logistic 回归三种方法的预测准确度，结果也证明是离散时间风险模型具有更好的预测效果。

（5）人工神经网络模型。二十世纪九十年代，Odom、Sharda 在财务风险研究中开始采用人工神经网络的技术，他们将样本分为为训练集与测试集，将 Z 值模型使用的五个财务比率作为模型变量，运用类神经网络的模型进行预测，结果证明有很高的正确率。由于人工智能的不断发展，开始出现了与以前不同的预测方法，如通过 bp 神经网络的模型来预测一家上市公司财务失败的概率，打破了统计学方法里众多假设的限制。而在国内，刘洪、何光军也在研究中比较了 bp 神经网络、Fisher 判别和 Logistic 回

归模型，结果表明，bp 神经网络模型摆脱了以前需要各种统计假设和线性回归的限制，具有更好的优越性和更准确的预测能力。

2.知识图谱

知识图谱技术近几年得到了快速发展，如今知识图谱已经成为了互联网结构化信息系统发展的一块重要部分。知识图谱的雏形来自于 1980 年提出的智能系统，为了满足智能搜索的需求，它将知识整合成块提供给用户。近些年来，随着开放关联数据集类似 DBpedia 的出现，以及搜索巨头谷歌公司在 2012 年提出了知识图谱的概念，知识图谱的发展越来越受到如今科学界的重视。

如今有多种方式构建知识图谱，人工方式比如 Cyc，通过 Freebase 以及维基数据，或者从大规模、半结构化的数据集类似维基百科、DBpedia、YAGO 抽取得来。此外，更多的学者提出了基于结构化或者半结构化的信息抽取系统，在这个基础上产生了 NELL、PROSPERA 和 KnowledgeVault 模型。

近些年来构建知识图谱的方法种类很多，但是没有一种方法可以称得上完美。作为真实世界或者理论概念的模型，知识图谱不可能达到完全百分之百的覆盖率，不可能覆盖整个宇宙所有的信息和实体。不过随着现在深度学习算法的提出和发展，构建一个相当准确的知识图谱可能性大大提高。

3.深度学习

深度学习是一种特殊的人工神经网络，深度学习最早的模型就是具有深度网络结构的人工神经网络。 最早的循环神经网

（Recurrent Neural Network，简称 RNN）就是由 John Hopfield 在 1982 年提出的 Hopfield 网络。由于 Hopfield 实现难度大，同时没有找到合适的应用场景，之后逐渐被前向神经网络代替。十年后又出现了 Elman&Jordan SRN 两种新的循环神经网络，也因为没有合适的场景来应用导致没有得到研究领域的重视。后来论文《THE VANISHING GRADIENT PROBLEM DURING recurrent neural networks and problem solutions》的作者 Dalle Molle 人工智能研究所的主任 Jurgen Schmidhuber 提出了长短时记忆网络，才推动了 RNN 的发展，尤其在深度学习得到广泛应用的现在，长短时记忆网络在自然语言处理领域，特别是情感分析、机器翻译、智能聊天等领域取得了令人惊异的效果。

（二）研究内容

随着知识图谱、专家系统以及深度学习技术的飞速发展，以科技手段对上市公司进行智能监管成为可能。本文以上市公司智能监管的需求作为应用场景，主要研究了下面三个方面的问题:

1.建立一套上市公司的风险评价体系，配合人工智能技术在监管方面的实践，具有很大的实用性要求。本文结合我国证券市场已发生风险的相关情况，运用全面风险管理理论和已有财务预测的应用研究成果，选择了可行性、实用性和可靠性最广泛的奥特曼 Z 值统计模型，来构建关于我国上市公司风险的专家系统。

2.根据公司一个周期的财务状况判断其财务风险的问题，采用深度学习当中的循环神经网络（RNN）进行研究。RNN 是一类用

于处理序列数据的神经网络。根据公司历次财务报表，分析公司当前财务状况，采用深度学习模型来进行财务风险的判断。神经网络可以当做是能够拟合任意函数的黑盒子，只要训练数据足够，给定特定的 x，就能得到希望的 y。

3. 参考知识图谱的通用构建框架，基于深度学习技术，以长短时记忆网络为语料特征学习模型，建立了命名实体识别方案。在实体关系抽取中，从公开的上市公司研报和公告中提取素材，结合领域知识和业务需求找到实体之间的关系，为知识图谱提供理论及数据支撑。最后在上市企业知识图谱基础上提出一些规则模型和概率模型等分析方法来起到对业务层监管的支撑作用。

二、总体研究框架

图 1：总体研究框架

（一）系统逻辑分层模型

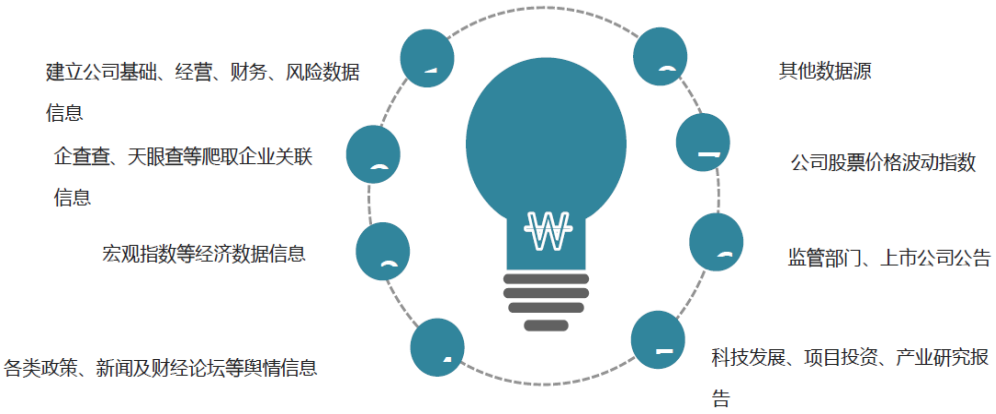
构建一个面向上市公司的智能监管分层模型，需要一个数据采集系统以及建立在数据之上的金融知识图谱。在此基础上，本文通过企业财务数据为上市公司建立财务评级体系标准，以此对企业的正常运营做出健康度模型，并通过深度学习技术来构建企业财务风险预测模型。另外，以知识图谱的图分析技术为基础，结合一系列的规则和概率模型来对公司存在的风险进行方法探索。

图 2：系统分层模型

（二）数据获取

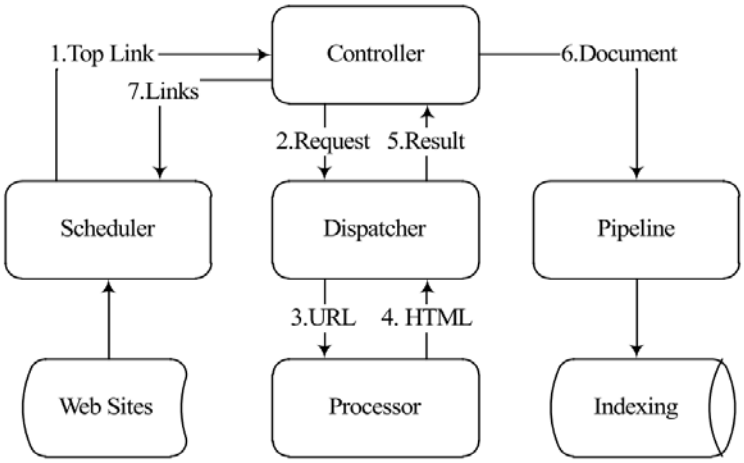
智能监管需要建立在庞大的数据基础上，图 3 是智能监管需要的数据维度。

图 3：智能监管数据维度



为了解决数据问题，首先需要开发一个能得到实时资讯数据的大规模爬虫。该爬虫由 5 个部分组成，包括调度器（Scheduler）、控制器（Controller）、分发器（Dispatcher）、处理器（Processor）及传输器（Pipeline）。

图 4：爬虫架构



调度器控制下载状态以及未下载链接的队列。当新的链接加入到调度器时，它会计算该链接的权重并插入正确的位置。此外，当新的网页被下载下来时，调度器还会更新与该网页相关的链接的权重，并重排队列。控制器控制整个系统。它从调度器获取新

的 URL，发送给分发器，从分发器接收下载结果，将新的链接加入调度器并且通知传输器处理下载下来的文档。分发器负责将下载请求分发给处理器。处理器负责下载网页。分发器和处理器可以部署在不同的机器上，因而爬虫可以通过用不同的网络（宽带）来提示下载速度。最后，传输器负责索引下载下来的文档。因为爬虫是为实时追踪最新资讯设计的，所以新发布的信息比旧信息更重要。在实践中，发现新发布的高质量信息往往位于网站的首页或者首页链接的页面。因而，调度器根据网址所处页面对网址进行排序，为每个链接都分配了一个权重，其中首页的权重是固定的，其他链接的权重根据下面公式计算:

$$pri(l) = \max_{pg \in PG} pri(pg) - 1$$

其中 l 是待计算的链接，PG 是包含 l 的网页的集合。在爬虫中，网页被下载的时候同步更新相关链接的权重。最后，爬虫会定时访问网站的首页。有了大规模的爬虫，除了爬取实时资讯数据，还考虑爬取百科类数据。百科类数据对于知识图谱非常重要。

1.百科类数据

原始数据的来源多种多样，根据领域的不同也存在不同的数据接口，权威的基础数据以百科网站为主。作为最大的在线百科全书，维基百科是通过协同编辑的方式来完成的。为了获取维基百科的数据，可以采取下面的方式：在概念页面发现各种概念以

及其上下位关系；以歧义页面和内链锚文本得到同音异义词并在文章页面上获得实体；在重定向页面抽取实体的同义词；以页面关联的开放分类解析实体所对应的类别;以信息框解析实体的“关系-实体对”和“属性-值对”。同样以互动百科和百度百科中的数据作为维基百科的补充数据。另外，Freebase 也是重要基础数据源,Freebase 的一个数据源就覆盖了谷歌知识图谱一半的规模。Freebase 是直接编辑知识，包含实体及其属性关系，以及实体所属的类型等结构化数据，所以不需要通过任何抽取规则即可得到高质量的数据；而维基百科是以各种词条的形式进行编辑的，词条又以文章的形式展现出来，其中包括半结构化信息，还要根据事先制定的规则来进行知识抽取。Freebase 目前是独立运行的开放知识管理平台，搜狗和百度在自己的知识图谱中也将 Freebase 数据加入进去。

大规模知识库的基本组织单位是词条，现实世界的每个概念和某个词条有对应关系，协同编纂内容的工作由世界各地的编辑者义务参与。随着 Web 2.0 理念的普及，传统的专家百科全书由于其生成理念和实现手段的落后已经不能和协同编辑的知识库在总量、质量以及效率方面相比较。现如今,维基百科已有超过 2200 万词条被收录，英文版超过 400 万条，而大英百科全书也才 50 万条收录，成为全球流量排名第 6 的站点。

2.结构化数据

结构化数据经过了复杂的数据加工以及清洗处理过程，因此是其他形态的数据无法比拟的，不需要太多的过程处理及技术门槛就可以轻易地进行知识融合。而对于结构化数据以外的数据类型（例如互联网的网页数据），需要利用自然语言处理里面的实体抽取、实体属性以及实体关系等技术进行获取。

除了百科类的数据，知识图谱的建设还需要考虑结构化数据，例如在金融领域采用公开的商业数据。此外，LOD 项目通过 owl:sameAs 将新发布的语义实体和 LOD 已包含的潜在同一实体进行关联进行实体对齐。LOD 包括 DBpedia 和 YAGO 等通用语义数据集，特定领域的知识库还有 MusicBrainz 和 DrugBank 等。此外，通过收购 Web 上存在大量高质量的垂直领域站点或者购买其数据来进一步扩充其知识图谱在特定领域的知识。

（三）知识图谱建设

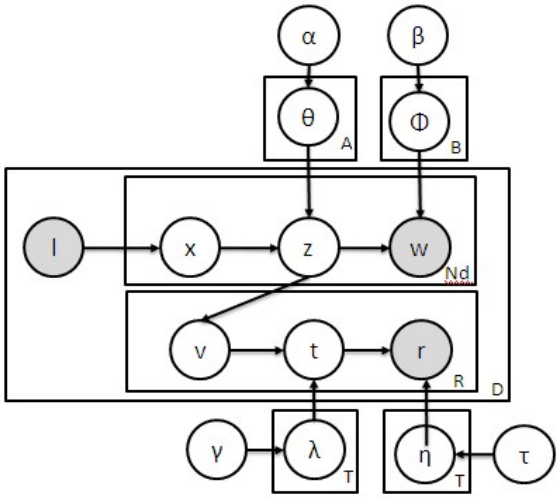
知识图谱建设需要先构建庞大的数据基础，在此基础上进行金融知识图谱的建设才能水到渠成。

1.社交数据建模

利用话题这个隐含的变量来表示出大数据环境下社交数据中不同领域信息之间存在这种依赖关系。基于领域的社交网络话题模型 DSNT (Domain-based Social Network Topic Model)是从人们通过社交网络沟通这个过程出发的。具体来说，社交网络中的某一篇文章，可能由作者根据其社交关系及其所属域 x，围绕主

题 z 撰写出来的。文章中作者的想法能决定这个作者写出什么内容，也代表本模型中的隐含变量话题。模型中用话题将文章的内容和关联文章联系在一起，这些关联文章即作者的非原创部分，比如部分引用社交关联的文章内容。在话题和关联文章之间加入了一个子话题层来表达这种多对多的关联关系。从概率分布来看这个模型，基于领域生成话题是一个多项式分布；同样，话题生成单词、关联文章等也都是多项式分布。根据以上分析，构建的 DSNT 模型示意图如下:

图 6：基于领域的社交网络话题模型



在 DSNT 模型中，α是话题先验分布的参数，β是单词先验分布的参数，θ是文档生成话题的概率，Φ是话题生成单词的概率，z 是文档中每个单词的话题，w 是文档中的单词，l 是有关话题的各个领域，x 是生成话题的领域，v 是决定关联文章选择的潜在话题，t 是文档中每个关联文章的子话题，r 是文档中的关联文章，λ是话题生成子话题的概率，η是子话题生成关联文章的概率，

γ是子话题先验分布的参数，τ是关联文章先验分布的参数，Nd是文档中单词的个数，R是关联文章的个数，D是文档的个数，A是潜在话题的个数，B是词典中单词的个数，T是潜在的子话题个数。

整个模型求解的关键是迭代中怎么样为单词根据其领域选择合适的话题和在社交网络中引用的子话题，并且如何为每个引用的文章选择合适的话题和子话题，根据模型的特征推导如下:

$$P(z_{di},x_{di}|\vec{z}_{-di},\vec{x}_{-di},\vec{w},\alpha,\beta)=\frac{P(\vec{z},\vec{x},\vec{w}|\alpha,\beta)}{P(\vec{z}_{-di},\vec{x}_{-di},\vec{w}|\alpha,\beta)}\propto\frac{n_{x_{di}z_{di}}^{-di}+\alpha}{\sum_z(n_{x_{di}z}^{-di}+\alpha)}\frac{n_{z_{di}w_{di}}^{-di}+\beta}{\sum_v(n_{z_{di}v}^{-di}+\beta)}$$

$$P(v_{di},t_{di}|\vec{t}_{-di},\vec{r}_{-di},\vec{z},\gamma,\tau)=\frac{P(\vec{t},\vec{r},\vec{z}|\gamma,\tau)}{P(\vec{t}_{-di},\vec{r}_{-di},\vec{z}|\gamma,\tau)}\propto\frac{n_{t_{di}r_{di}}^{-di}+\tau}{\sum_r(n_{t_{di}r}^{-di}+\tau)}\frac{n_{v_{di}t_{di}}^{-di}+\gamma}{\sum_t(n_{v_{di}t}^{-di}+\gamma)}\frac{n_{dv_{di}}}{N_d}$$

其中z_{di}是文档 d 中第 i 个单词被指派到的话题；x_{di}是文档 d 中第 i 个单词被指派的领域；v_{di}是文档 d 中第 i 个关联文章被指派到的话题；t_{di}是文档 d 中第 i 个关联文章被指派到的子话题；$n_{x_{di}z}^{-di}$表示话题z_{di}属于领域x_{di}的次数，不包含当前这次；$n_{x_{di}z}^{-di}$表示话题 z 属于领域x_{di}的次数，不包含当前这次；$n_{z_{di}w_{di}}^{-di}$表示文档 d 中第 i 个单词被指派给话题z_{di}的次数，不包含当前这次；$n_{z_{di}v}^{-di}$表示单词 v 被指派到话题z_{di}的次数，不包含当前的这次；$n_{t_{di}r_{di}}^{-di}$表示文档 d 中的第 i 个关联文章被指派到话题t_{di}的次数，不包含当前这次；$n_{t_{di}r}^{-di}$表示关联文章 r 被指派到话题t_{di}的次数，不包含当前这次；$n_{v_{di}t_{di}}^{-di}$表示子话题t_{di}被指派到话题v_{di}的次数，不包含当前这次；$n_{v_{di}t}^{-di}$表示子话题 t 被指派到话题v_{di}的次数，不包含当前这次；$n_{dv_{di}}$表示文档 d

中被指派到话题v_{di}的单词的个数。在实际应用中，词变量 w 及其所属领域 l 是可观察，而变量 z，x，v 和 t 是需要估算的潜在变量。通过设定 Dirichlet 分布的先验参数 α、β、γ和 τ，通过期望最大化（Expectation Maximization）或者 Gibbs Sampling 等方法计算变量 θ，Φ，λ和 η的概率分布，从而确定文档在不同领域的关键主题。另外，对该模型在复杂度上的优化，提高算法的收敛速度和准确率是研究中还需要考虑的重要因素。

2.基于迁移学习的关键概念抽取

因为社交数据通常包含不同的领域，采用传统的机器学习的方法来提取关键概念需要投入大量的人力进行多次开发，所以可以采用迁移学习的方法来进行。本研究采用的迁移学习的方法主要针对相同的任务，即关键概念提取，在不同领域的实现。通常在机器学习的过程，想学习得到模型的最佳参数 θ^*来降低期望风险（Expected Risk）E,即:

$$\theta^* = arg\min_{\theta\in\Theta} \mathrm{E}_{(x,y)\in P}[l(x,y,\theta)]$$

其中$l(x,y,\theta)$是依赖于参数 θ 的损失函数(Loss Function)，P 是样本的概率分布。为了从源领域 D_S的数据样本的概率分布迁移学习到目标领域 D_T，可定义并推导出损失函数的如下:

$$\theta^* = arg\min_{\theta\in\Theta}\sum_{(x,y)\in D_S}\frac{P(D_T)}{P(D_S)}P(D_S)l(x,y,\theta) \approx arg\min_{\theta\in\Theta}\sum_{i=1}^{n_S}\frac{P_T(x_{Ti},y_{Ti})}{P_S(x_{Si},y_{Si})}l(x_{Si},y_{Si},\theta)$$

为了计算损失函数中的$\frac{P_T(x_{Ti},y_{Ti})}{P_S(x_{Si},y_{Si})}$，采用 kernel-mean matching(KMM)算法对源领域数据和目标领域样本数据在新生成

的核希尔伯特空间（Reproducing Kernel Hilbert Space）中进行匹配。在该问题中，KMM 算法是以下多项式的优化问题:

$$\min_{\beta} \frac{1}{2}\beta^T K\beta - k^T\beta$$

$$s.t.\beta_i \in [0,B] \text{ and } |\sum_{i=1}^{n_S}\beta_i - n_S| \le n_S e$$

其中 $K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix}$和 $K_{ij} = k(x_i, x_j)$ 。$K_{S,S}$ 和 $K_{T,T}$ 分别是源领域数据和目标领域数据的核矩阵；$k_i = \frac{n_S}{n_T}\sum_{j=1}^{n_T} k(x_i, x_{Ti})$，其中 $x_i \in X_S \cup X_T$ 和 $x_{Ti} \in X_T$；最终求解的 $\beta_i = \frac{P_T(x_T \quad \partial_{Ti})}{P_S(x_{Si}y_{Si})}$，从而构建目标领域的参数估算模型。采用 KMM 算法的一个好处是直接计算了 $\frac{P_T(x_{Ti}\partial_{Ti})}{P_S(x_{Si}\partial_{Si})}$，从而避免了单独计算概率函数 $P_T(x_{Ti}, y_{Ti})$和 $P_S(x_{Si}, y_{Si})$，这两个概率函数在数据样本比较小的时候是非常难计算的。

3.基于 NID 的语义关系计算

对关键概念的语义关系计算一直是语义 Web 领域的一个难点，因为语义关系的计算存在计算量大和不确定性等问题。在研究了大量文献的基础上，本项目创新性地提出了利用 NID 理论模型来计算语义关系的方法。NID 主要是基于 Kolmogorov 复杂性的理论对对象之间的关联度进行衡量。Kolmogorov 复杂性可用于衡量独立个体的绝对语义信息，基于 Kolmogorov 复杂性，两个对象 x 和 y 之间的 NID 可以被定义为:

$$e(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}$$

其中,K(x) 和 K(y)分别是对象 x 和 y 的 Kolmogorov 复杂性。

但是，现实中 Kolmogorov 复杂性是不可算的，所以必须采用某种方法来估算 Kolmogorov 复杂性。可以采用两种方法来达到这个目的:

一是利用压缩算法，可以计算规范压缩距离(Normalized Compression Distance)，通过规范压缩距离来估算 NID。因为有现成的各种压缩算法可以利用，例如 gzip、 bzip2、 PPMZ 等，所以可以通过试验来选择效果最好的压缩算法实现对 NID 更加精确的估算。对利用压缩算法估算 NID 的尝试，有关研究进行了剽窃检测、音乐文件聚类、异源数据聚类等实验，证明其估算的有效性。

二是利用互联网的海量数据，可以计算规范网络距离(Normalized Web Distance)。基于搜索引擎，先计算各个概念的网页关联数，以及和某个概念集合所关联的网页数，从而计算出规范网络距离，通过规范网络距离来估算 NID。对利用网络来估算 NID，有关研究进行了分层次的聚类、中文名字的识别、问答系统的开发等多个实验，从而广泛地证明了规范网络距离能有效地估算 NID。

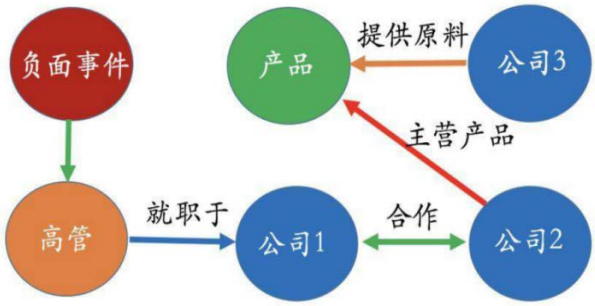
实验证明，以上两种方法都可以在特定场景下达到相当高的精确度。所以在实践中，可以把 NID 作为一个理论基础来研究如何最大限度准确计算概念之间的语义关系，并通过数据集验证所计算的语义关系的可靠性。

三、上市公司监管模型

（一）基于知识图谱的智能监管模型

在证券行业，时间对于公司股价的影响一直是关注的焦点，例如公司 1 的高管出现了负面新闻，而且公司 1 和公司 2 之间存在密切合作关系，公司 3 是公司 2 主营产品的原材料供应商，之间关系就可以如图 7 所展示：

图 7：知识图谱在公司监管中的应用



知识图谱具备展示这种关联关系的能力，公司 1 高管的负面事件的影响范围就能够清晰地确定。但是，相关性的强度必须要得到数据验证。所以知识图谱的优势是快速圈定某一事件的关注范围。

搭建金融风控的知识图谱，其核心在于对业务的理解及对知识图谱本身的设计，其包含如下几个完整的步骤：1. 定义具体的业务问题；2. 数据收集&预处理；3. 知识图谱设计；4. 知识图谱实现；5. 上层应用开发。这里仅对具体业务定义问题进行详述。

1. 具体业务定义

具体的业务定义决定了自身业务对于知识图谱系统的需要程

度。在一些实际应用当中，即便需要进行一定的关系分析，传统的数据库也可以完成，并不需要建立新的知识图谱系统。因此，为了确定知识图谱系统的现实必要性，以及实现更好的技术选型，可以通过表 1 进行参考选择。

表 1：知识图谱的使用场景对比

简单方式	知识图谱
对可视化要求不高	有强烈的可视化需求
很少涉及到关系的深度搜索	经常涉及到关系的深度搜索
关系查询效率要求不高	对关系查询效率有实时性要求
数据缺乏多样性	数据多样化、解决数据孤岛问题
暂时没有人力或者成本不够	有能力、有成本搭建系统

以下是对于上市公司欺诈的风险控制来进行业务定义，即如何对一个公司的欺诈行为进行判断。诸多的欺诈风险是隐藏在复杂的关系网络之中的，而知识图谱系统的核心功能就是梳理复杂的关系网络，因此知识图谱在反欺诈这个领域拥有巨大的实用价值。

对于反欺诈，公司的基本信息、行为数据、运营商数据、网络上的公开信息等数据源是较易获取的。假设已经建立了一个数据源的列表清单，下一步则是对哪些数据需要进一步处理做出判断，例如非结构化数据大多情况都需要自然语言技术进行处理后才可以正常使用。公司的基本信息主要存储在业务表里，除了个别字段需要进一步处理外，大部分字段是可以直接用于建模，或者添加到知识图谱系统当中的。

表 2：上市公司知识图谱数据维度

资本结构关系	股权	最终控制人（自然人或法人）
		最终控制人至本公司间的控制路径
		母公司（可能与实际控制人重合）
		股东
	债权	短期借款方
		长期借款方
		债券
	投资	子公司
		孙公司
		子孙公司及其他被实际控制的公司（通过多层控股的方式实现控制）
		参股公司
	并购重组	对手方
公司经营关系	供应商	前五大供应商
	短期闲置资金运营	交易性金融资产项目
		持有至到期投资项目
		可供出售金融资产项目
	客户	前五大客户
	应收账款项目	按欠款方归集的期末余额前五名的应收账款
		应收账款计提坏账（按欠款方）
	其他应收款项目	按欠款方归集的期末余额前五名的其他应收款
		应收账款计提坏账（按欠款方）
	商誉项目	商誉减值准备项目
	同业竞争	按行业分类的龙头公司
	关联交易	关联方
利益相关关系	董监高	董事
		监事
		高级管理人员
		独立董事
	未决诉讼	原告
		被告
	担保	担保方
		被担保方
	合作	重大战略合作
		重大技术合作
		会计师事务所
		历史会计师事务所
	机构持股	机构名称
		机构类型
		持股比例
	股权行为	股权增持/减持
		股权质押
	违规处罚/劣迹记录	发出单位
		违规涉及的当事方
		处罚类型
		处罚事由

2.业务设计

对于风控知识图谱来说，首要任务就是挖掘关系网络中隐藏的风险。从算法角度来讲有两种不同的场景：一种是基于规则的，另一种是基于概率的。鉴于目前 AI 技术的现状，基于规则的方法论还是在垂直领域的应用中占据主导地位，但随着数据量的增加以及方法论的提升，基于概率的模型也将会逐步带来更大的价值。

（1）基于规则的方法

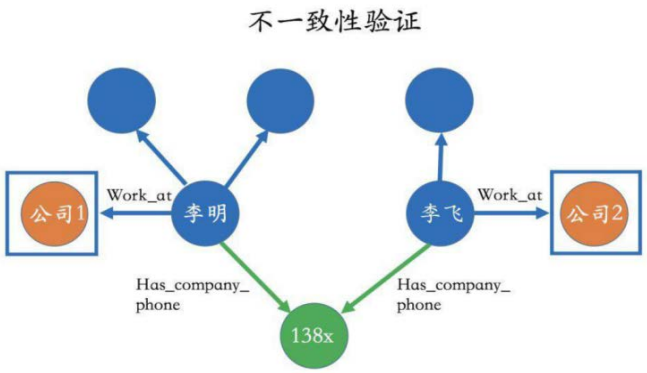
首先来看几个基于规则的应用，分别是不一致性验证、基于

规则的特征提取、基于模式的判断。

①不一致性验证

为了判断关系网络中存在的风险，一种简单的方法就是做不一致性验证，也就是通过一些规则去找出潜在的矛盾点。这些规则是以人为的方式提前定义好的，所以在设计规则时需要一些业务知识。比如图 8 中，李明和李飞两人都注明了同样的公司电话，但实际上从数据库中判断这两人其实不在同一个公司上班，这就是一个矛盾点。类似的规则可以有很多，这里不一一列出。

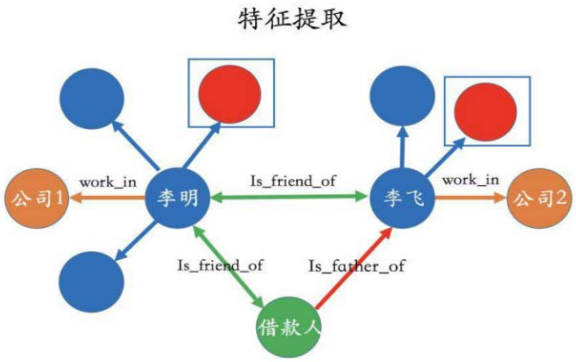
图 8：知识图谱不一致性监管场景模型



②基于规则提取特征

可以基于规则从知识图谱中提取一些特征，而且这些特征一般基于深度搜索比如二度、三度甚至更高维度。比如可以问一个这样的问题:“上市公司的股东二度关系里有多少个实体触碰了黑名单？”，图 9 中很容易观察到二度关系中有两个实体触碰了黑名单（黑名单由红色来标记）。等这些特征被提取之后，一般作为风险模型的输入。另外，如果特征不涉及到深度的关系，其实传统的关系型数据库就可以满足。

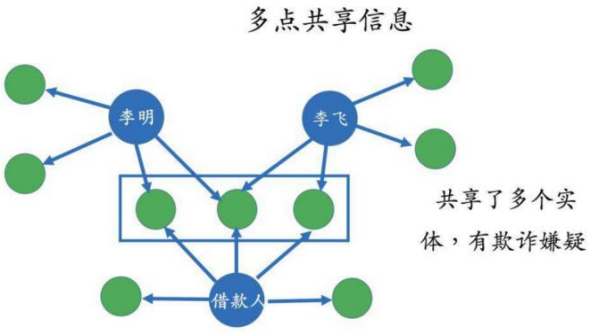
图 9：知识图谱基于特征提取监管场景模型



③基于模式的判断

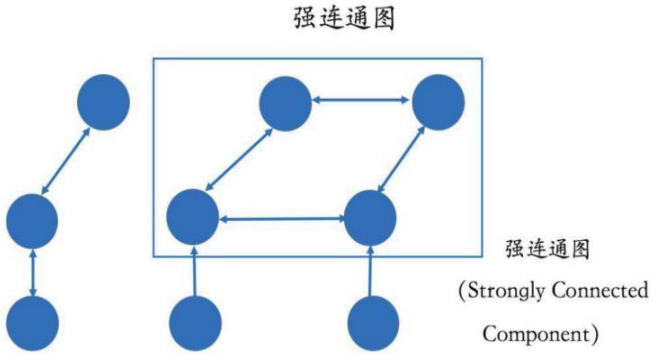
基于模式的判断方法更适用于找出团体欺诈，它的核心在于通过一些模式来找到有可能存在风险的团体或者子图，然后对这部分子图做进一步的分析。这种模式有很多种，比如图 10 中，三个实体共享了很多其他信息，可以看做是一个团体，并对其做进一步的分析。

图 10：知识图谱模式判断监管场景模型



再比如，也可以从知识图谱中找出强连通图，并把它标记出来，然后做进一步的风险分析。强连通图意味着每一个节点都可以通过某种路径达到其他的点，也就说明这些节点之间有很强的关系。

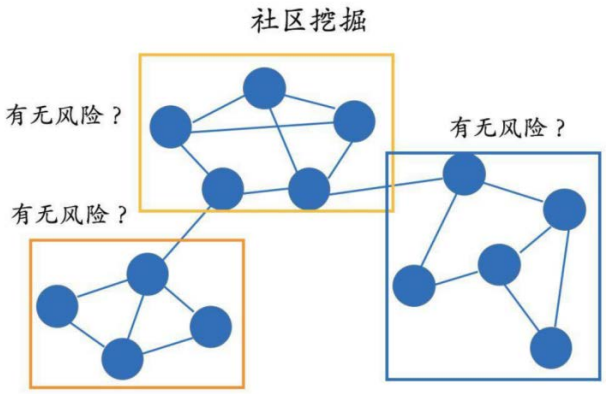
图 11：知识图谱强连通监管场景模型



（2）基于概率的方法

挖掘关系网络中的风险还可以使用概率统计的方法，比如社区挖掘、标签传播、聚类等技术都属于这个范畴。社区挖掘算法的目的在于从图中找出一些社区。对于社区可以有多种定义，但直观上可以理解为社区内节点之间关系的密度要明显大于社区之间的关系密度。图 12 表示社区发现之后的结果，图中总共标记了三个不同的社区。一旦得到这些社区之后，就可以做进一步的风险分析。由于社区挖掘是基于概率的方法论，好处在于不需要人为去定义规则，特别是对于一个庞大的关系网络来说，定义规则本身是一件复杂的事情。

图 12：知识图谱概率方法监管场景模型

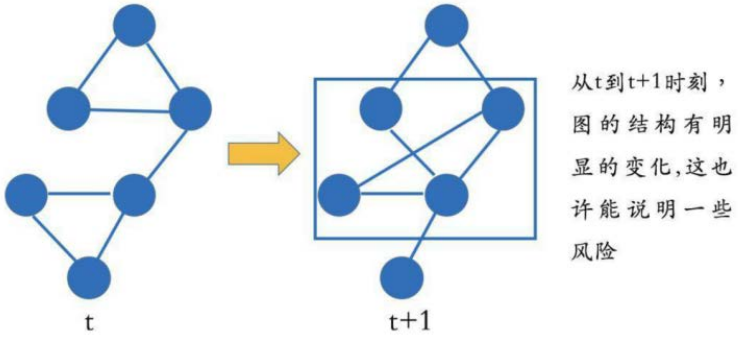


（3）基于动态网络的分析

以上所有的分析都是基于静态的关系图谱。所谓的静态关系图谱，意味着不考虑图谱结构本身随时间的变化，只是聚焦在当前知识图谱的结构上。然而知识图谱的结构是随时间变化的，而且这些变化本身也可能跟风险有所关联。

图 13 给出了一个知识图谱 T 时刻和 T+1 时刻的结构，很容易看出这两个时刻中间，图谱结构（或者部分结构）发生了很明显的变化，这其实暗示着潜在的风险。

图 13：知识图谱基于动态分析监管场景模型



从以上分析可知，基于知识图谱的上市公司监管方法需要基于场景需求采取不同的方法进行分析，在监管的应用层面会基于业务场景做不同的监管模型来满足不同维度的监管需求。

（二）基于深度学习的财务智能监管模型

根据公司历次财务报表，分析公司当前财务状况，将采用深度学习模型来进行财务风险的判断。神经网络可以当做是能够拟合任意函数的黑盒子，只要训练数据充足，给定特定的 x，就能得到希望的 y。

为了根据公司一定市场内的财务状况，对其财务风险状况进行判断，选择采用深度学习当中的循环神经网络（Recurrent Neural Network，简称 RNN）进行研究。RNN 要处理的是序列数据。首先要明确什么是序列数据：在不同时间点上收集到的数据，这类数据反映了某一事物、现象等随时间的变化状态或程度叫做时间序列数据。这是时间序列数据的定义，当然这里也可以不是时间，比如文字序列，但总归序列数据有一个特点，即后面的数据跟前面的数据有关系。

1.模型构建

从基础的神经网络中可知，神经网络包含输入层、隐层、输出层，通过激活函数控制输出，层与层之间通过权值连接。激活函数是事先确定好的，那么神经网络模型通过训练“学”到的东西就蕴含在“权值”中。基础的神经网络只在层与层之间建立了权连接，RNN 最大的不同之处就是在层之间的神经元之间也建立的权值连接，具体如图 14 所示。

图 14：标准的循环神经网络结构图

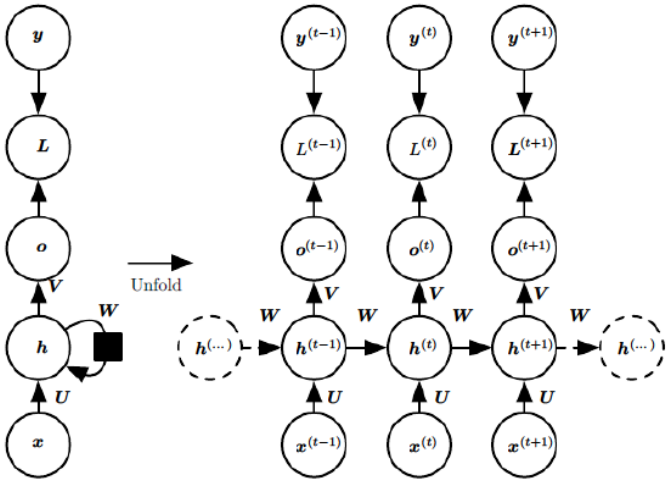


图 14 是一个标准的 RNN 结构图，图中每个箭头代表一次变换，即箭头连接带有权值。右侧展开图显示了左侧图的“循环”体现在隐层中。

在展开结构中可以观察到，在标准的 RNN 结构中，隐层的神经元之间也是带有权值的。也就是说，随着序列的不断推进，前面的隐层将会影响后面的隐层。图中 o 代表输出，y 代表样本给出的确定值，L 代表损失函数。可以看到，“损失”也是随着序列的推进而不断积累的。

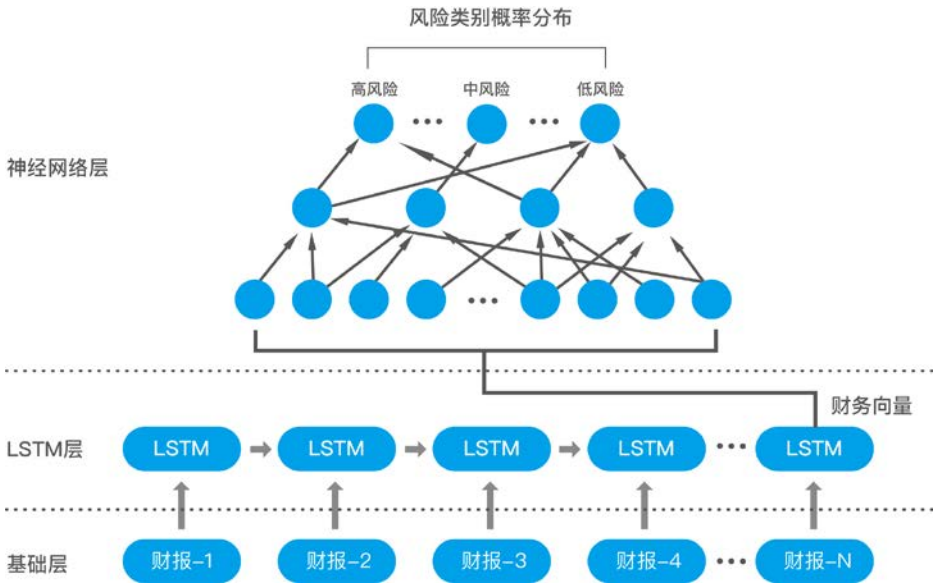
除上述特点之外，标准 RNN 的还有以下特点：

（1）权值共享，图中的 W 全是相同的，U 和 V 也一样。

（2）每一个输入值都只与它本身的那条路线建立权值连接，不会和别的神经元连接。

基于上述循环神经网络的研究分析，结合财务风险判断逻辑，构建了基于深度学习的财务风险 AI 模型，如图 15 所示：

图 15：基于深度学习的财务风险 AI 模型



构建的财务风险模型是基于RNN模型，其中引入了LSTM机制更好地对公司财务在时间序列上进行记忆和分析。模型的输入为公司的各个时间段的财务状况，以最后一层的输出作为神经网络层的输入，通过机器的学习，模型逐步收敛后，机器就可以计算每个风险类别的评分。

在财务风险判断模型当中，输入公司三年的所有财报（2015年、2016年、2017年），每个财报包含的指标都在财务预警指标体系当中。每份财报中，将基于语义识别技术，构建财务指标向量X。根据沪深两市2015、2016、2017年所有上市公司的财务指标值X，人工专家首先需要对每个公司的财务风险状况进行标签，记录为Y值。每家公司至少要根据最近一次的财报，进行标签的录入。最理想的状况是所有的财报都可以生成一个标签，这样模型能拥有更多的训练数据。

根据每个公司目前已经打好的标签，基于构建的财务风险AI模型进行训练，通过训练调参，让整个模型的输出结果达到收敛，即机器学习的结果具有一致性。

基于训练好的模型，输入公司的财务指标，对输出的财务风险判断进行审核，如果出现误差，检查原因并对模型参数或者结构调整，直到输入的财务风险判断和人工专家的判断一致。

2.构建步骤

（1）LSTM 初始化，设定各个层节点个数，将权值和阈值初

始值设为比较小的随机数。

（2）输入样本及对应的输出，对样本进行逐一学习，也就是对每个样本进行（3）到（5）的过程。

（3）根据输入的样本进行计算结果并输出，这个结果也包含了隐含层的输出。

（4）算出期望差值，这个期望差即是输出层还有隐含层的误差。

（5）根据（4）得出的结果来重新迭代每一层节点之间的连接权值。

（6）求误差函数，判断其是否收敛到期望的学习精度以内，如果满足学习就结束，否则转向（2）继续进行。

3.采集样本及指标选择

（1）采集样本

选取120家上市公司，应用LSTM模型建立并预测上市公司是否面临财务危机。再以60家公司作为检验样本，对模型预测结果进行检测。具体选择上市总体 90 家股票存在终止上市风险的公司，然后考虑到终止上市风险公司和非终止上市风险公司存在总体数量上的差异，以及上市公司的版块分布情况，因此抽样的90家非终止上市风险公司里，以农业板块和房地产板块为主。基于上面的考虑，对于样本的选择，终止上市风险公司几乎全部采纳，而用随机抽取方法来对非终止上市风险公司进行样本的选取，这

样才能对两种类型的样本规模取得一致。所以样本选择满足了客观性和科学性，为后续的科学分析提供了坚实的数据基础。

（2）指标选择

结合现有研究成果，最终选择短期和长期偿债能力、赚钱能力、主营业务突出度、业绩增长等 6 个方面中 15 个备选指标。

表 3：备选预测指标

财务特征	财务比率指标	财务特征	财务比率指标
A. 短期偿债能力	X_1：流动比率	D. 资产管理能力	X_9：存货周转率
	X_2：速动比率		X_{10}：应收账款周转率
	X_3：现金比率		X_{11}：总资产周转率
B. 长期偿债能力	X_4：产权比率	E. 主营业务鲜明程度	X_{12}：主营业务鲜明率
	X_5：利息保障倍数		
C. 盈利能力	X_6：盈利现金比率	F. 公司增长能力	X_{13}：资本保值增值
	X_7：总资产报酬率		X_{14}：净利润增长率
	X_8：净资产收益率		X_{15}：累积盈利能力

在样本足够大的情况下，财务指标的筛选有很多种选择，我们认为 T 假设方法可以用来进行检验，通过财务比率指标差异的 T 检验分析来对终止上市风险公司和非终止上市风险公司进行检验，结果发现:

①终止上市风险公司在短期偿债方面不如非终止上市风险公司。终止上市风险公司的速动比率均值也低于非终止上市风险公司，尽管流动比率及现金比率的置信度都大于 5%。

②5%显著性尺度上，终止上市风险公司的产权比率低于非终止上市风险公司，表明终止上市风险公司债务负担比非终止上市风险公司轻。在负债利息率大于总资产报酬率时，比率太高加速了财务恶化速度，从而造成成为终止上市风险公司的可能性加大。

③非终止上市风险公司的盈利能力明显大于终止上市风险公司，总资产报酬率和净资产收益率两个财务比率指标方面非终止上市风险均高于终止上市风险公司，这两个指标的综合性比较强。

④虽然存货周转率和应收账款周转率不存在太大的差异，但是非终止上市风险公司应收账款周转率明显高于终止上市风险公司，而且总资产周转率的差别特别明显，说明非终止上市风险公司的信用政策和资产管理方面能力较强。

以上发现说明，在业务上比较专注的公司往往可以远离财务困境。

4.样本及指标

表 4：样本描述和 T 检验结果

变量	Means		T 检验		Wilcoxon 秩检验	
	非 ST	ST	t-值	p-值	Z-值	p-值
X_1: 流动比率	2.054	1.536	-4.266	0.000	-6.027	0.000
X_2: 速动比率	1.528	1.208	-0.553	0.582	-5.488	0.000
X_3: 现金比率	0.553	0.340	-5.183	0.000	-5.953	0.000
X_4: 产权比率	0.590	0.293	-9.220	0.000	-6.785	0.000
X_5: 利息保障倍数	-34.820	-8.60	1.148	0.254	-3.398	0.001
X_6: 盈利现金比率	3.985	-0.316	2.015	0.017	-3.074	0.002
X_7: 总资产报酬率	0.042	-0.164	-10.532	0.000	-7.679	0.000
X_8: 净资产收益率	0.069	-0.207	-4.283	0.000	-6.680	0.000
X_9: 存货周转率	6.018	6.593	0.202	0.840	-2.460	0.014
X_{10}: 应收账款周转率	41.303	14.802	-1.487	0.141	-4.055	0.000
X_{11}: 总资产周转率	0.578	0.363	-2.916	0.005	-3.600	0.000
X_{12}: 主营业务鲜明率	5.047	0.356	-2.879	0.005	-6.652	0.000
X_{13}: 资本保值增值率	1.291	0.356	-4.024	0.000	-7.653	0.000
X_{14}: 净利润增长率	0.440	-11.442	-4.727	0.000	-6.636	0.000
X_{15}: 累积盈利能力	0.091	-0.235	-10.870	0.000	-7.692	0.000

5.预测结果检验

按照模型检验，将样本数据带入 LSTM 神经网络，对样本上市公司进行返回判定，结果如表 5 所示。

表 5：预测结果

组　别	建模样本		检验样本	
	实际个数	正确判定个数	实际个数	正确判定个数
ST	60	51	30	25
非 ST	60	58	30	29
正确判定率	90 8%		90%	

结果表明，基于 LSTM 的深度神经网络方法对企业是否会进入财务困境进行预判是一种非常可靠的方法。

（三）基于专家系统的财务智能监管模型

1.模型选择

结合我国证券市场已发生风险的相关情况，运用全面风险管理理论和已有财务预测的应用研究成果，选择了可行性、实用性和可靠性最广泛的奥特曼 Z 值统计模型，来构建关于我国上市公司风险的专家系统。实用性的风险模型建立之后，可以根据模型判断上市公司财务风险的水平，结合人工智能技术的实现，以知识图谱为基础，向监管部门立体地展示上市公司的风险状况。

选择 Z 值统计模型有以下两个原因：首先 Z 值统计法由于其简便性和有效性，成为了风控应用领域中最为广泛的一种，被用于各样环境下的财务预警，而前面提到的几类回归模型应用于学术探索研究较多，在现实应用层面上，这几类方法在实现上难度较大，评价层次不够丰富。其次，Z 值统计模型可以最大程度地结合原有的研究，优化指标和维度的选择，并针对行业和特殊情况等，加入特定的判断条件，完善丰富整个预警系统。

2.财务风险评估模型构建

（1）评价维度和指标设计逻辑

上市公司的财务失败，是指公司经营失败而显露在财务数据上的结果。从内因上看，企业发生经营失败导致财务失败，主要有三个方面，分别是公司的财务状况、经营质量和治理效能，除此之外，企业作为社会经济的主体，必然深受宏观环境和行业发展的影响。因此，财务状况、经营质量、治理效能和宏观及行业经济环境，也成为分析企业发生财务失败危机的四个维度，这四个维度也包含了从宏观到中观到微观的整个企业经营全部内容。

（2）评价维度的指标设定

①财务状况

根据对于财务失败的公司的调查研究，财务失败作为一种表象，其背后的根本原因来自于企业作为经济主体所承受的支付压力和他的支付能力之间的差距，也就是流动性与举债偿债能力之间的矛盾。

A.流动性

公司经营之中，当支付压力和支付能力不匹配，就会出现偶发性财务危机。这种财务危机来自于流动性和现金流，此时公司的价值也难以得到公允评估，最终也必然会陷入财务危机。

B.举债和偿债能力

在公司的财务评估当中，对于公司的债务结构是否健康的评

判，依据是企业是否可以得到充足的现金流入来支付必须的债务和利息。有大量研究证明，企业的杠杆率和财务危机有非常大的相关性。除此之外，还需将企业的杠杆率作为一个维度，进行仔细计量。

②企业经营质量

企业的经营质量，在于三个能力的动态平衡，即可持续的创新能力、可持续的发展能力和可持续的盈利能力。

A.盈利与成长能力

上市公司的盈利和增长是企业生存与发展的基础，对于所有的利益相关者都至关重要，特别是关系到企业偿债能力。对于企业盈利能力的评估，关系着企业不断获得资源，来保障自己财务安全的能力，是企业作为资源主体的活水，因此是企业财务风险预警评估的重要维度，也是评价财务安全水平的重要依据。

B.市场价值表现

上市公司作为公众公司，股价包含丰富的信息，经营状况较差的公司、失效的管理团队和不被看好的投资行为通过市场表现得到反应，因此对于风险的评估和预测来说，也是对于公司财务风险评估的一个重要维度。

③公司治理

在实证研究下，发现上市公司的经营失败和上市的低治理成效有着很大的关系。从非财务视角下，将公司治理相关指标引入

到对上市财务危机的研究，其变量都会深刻影响着企业经营状况和最终的财务风险水平，并且这些影响难以直观地在现有的财务指标中体现出来。因此对公司治理的相关指标进行评判，有利于提高模型对公司财务危机预警的逻辑完整性和准确度。

④宏观和行业经济环境

公司作为整个经济活动的主题，必然处在一定的宏观环境当中。就经济环境影响来说，宏观经济的周期性波动和厂商对于周期的滞后反应，让公司更容易陷入财务困境。就行业环境而言，在整个行业快速增长期，行业内的公司生存环境良好，不容易出现财务危机。而当行业发展速度下降，内部竞争加剧，更容易导致财务危机的出现。

基于此，对于宏观经济和行业环境的判断，也必须加入到预警模型当中，以提升整个预警模型的层次性和严密性。

（3）模型体系

基于以上分析，本文对具体指标进行了初选，依据业已形成的对于单指标和多指标效果研究，挑选出了备选指标。再经过统计学的回归验证，剔除了具有截面数据共线性的贡献指标，对于指标体系进行了精简。最后再根据研究基础和实践经验，对于指标体系进行了修正和进一步完善，形成了以下指标体系，共计 30 个指标。为了完成评价功能，在充分研究和广泛采纳相关专业人士的意见之后，对于各个维度和具体的指标权重及阈值进行了赋

值。权重和阈值体系的构建，主要综合了经典财务理论、行业经验和统计分布三种方法，并且还加入了宏观和行业状况对于参数的调整项，以完善整个权重和阈值系统。这套系统独立开发，区别于当下所有成型的风险评价体系，涵盖了大量对于企业风险的独到理解和专家经验。具体见表 6:

表 6：指标选择及量化标准

维度	指标	0	25	35	50	65	75	100
杠杆结构 10%	股东权益比率 1	0-0.15	0.15-0.3		0.3-0.4		0.4-0.7	0.7-1
流动性 20%	流动比率 25%	0-1		1-2		2--4		4-无穷
	存货周转率 10%	0-1		1-3		3-6		6-无穷
	应收账款周转率 20%	0-2		2-3		3-7		7-无穷
	总资产周转率 10%	0-0.3	0.3-0.5		0.5-0.8		0.8-1	1-无穷
	流动负债比 25%	<0.3 或>0.98		0.9-0.98		0.7-0.9		0.3-0.7
偿债能力 25%	现金流动负债比率 20%	无穷小-0	0-0.1		0.1-0.2		0.2-0.5	0.5-无穷
	现金负债比率 20%	0-0.1		0.1-0.25		0.25-0.5		0.5-无穷
	资产负债率 20%	0.85-1	0.7-0.85		0.6-0.7		0.3-0.6	0-0.3
	利息保障倍数 20%	无穷小-0	0-1		1-2.5		2.5-10	10-无穷
	资产流动率 20%	0-0.2	0.2-0.4		0.4-0.6		0.6-0.7	0.7-1
	净资产收益率（营业利润)10%	无穷小-0	0-3%		3%-7%		7%-14%	14%-无穷

公司盈利能力指标25%	净资产收益率（净利润)15%	无穷小-0	0-3%		3%-7%		7%-14%	14%-无穷
	资产收益率10%	无穷小-0	0-2%		2%-5%		5%-8%	8%-无穷
	净利润率15%	无穷小-0	0-5%		5%-8%		8%-11%	11%-无穷
	总资产增长率5%	无穷小-0	0-4%		4%-10%		10%-25%	25%-无穷
	营业利润增长率5%	无穷小-0	0-4%		4%-10%		10%-25%	25%-无穷
	净利润利润增长率10%	无穷小-0	0-4%		4%-10%		10%-25%	25%-无穷
	现金获利指数20%	<0	0-0.1或1.2-2		0.1-0.5或1.2-2		0.5-0.8	0.8-1.2
	留存收益总资产比5%	无穷小-0	0-5%		5%-10%		10%-25%	25%-无穷
市场表现10%	市盈率30%	无穷小-0	100-无穷		50-100		15-50	0-15
	每股收益（营业利润）10%	无穷小-0	0-0.2		0.2-0。6		0.6-1	1-无穷
	每股收益（净利润）10%	无穷小-0	0-0.2		0.2-0。6		0.6-1	1-无穷
	每股净资产20%	无穷小-1	1-3		3-4		4-6	6-无穷
	每股经营现金流量30%	无穷小-0	0-0.2		0.2-0.6		0.6-1	1-无穷
治理结构10%	Z指数20%	1-1.2	1.2-1.6		1.6-3		3-6	6-无穷
	会计师事务所20%	否						是
	董事会规模20%	人数<7			人数7-9			人数>9
	独立董事比例20%	<0.329			0.33-0.669			大于0.67
	机构持股占流通股比例20%	0-5%	5%-15%		15%-30%		30%-50%	50%—1

①适用范围

风险评价体系的适用范围去除金融、房地产企业和上市不满两年的次新股。金融公司由于其特殊的经营模式，财务指标和一般企业不同，评估方法也不同，因此需要另外开发风险评估模型，并不覆盖在本评估体系当中。地产企业由于宏观环境和中国国情的特殊情况，其财务风险大部分可以列为重点风险关注对象，且财务指标特殊性也较强，因此也不列为评估体系的正常评估范围。而由于上市两年以内的次新股，其从非公众公司到成为公众公司，对于经营是一个重大影响事件，导致上市初期财务指标波动较大，因此难以通过统一的风险评估体系进行衡量，建议对于次新股进行重点关注。

总之，金融公司需要另外开发评价体系，而次新股和地产企业都应重点做出风险监控。

②调节处理

所谓调节处理，是在指标体系的权重基础上，对宏观和行业情况、指标异动和特殊事件的发生，在评估结果中进行处理，将这三个方面的内容包含在评估体系当中，以保证整个体系严密性和完整性，提升结果的准确率。

A.宏观和行业状况

由于宏观和行业状况的多变性和行业划分复杂性，这部分信息难以通过自动系统进行更新和计量。因此选择通过专家对于宏

观经济环境和行业状况（行业划分越细效果越好）进行评估，将这部分信息划分为四个风险等级。每个被评估对象根据自己所处的行业所在的风险等级，对企业杠杆、偿债能力和盈利能力三个维度赋予不同调节系数，来进行最终分数的调节，从而将宏观和行业状况包含进评估体系里。风险等级如表 7 所示。

表 7：宏观风险调节参数

风险程度	含义	调节系数
正面	当前宏观的经济环境和行业发展有利于行业内主要竞争者信用水平的提高，宏观和行业的风险较低	1.15
稳定	当前宏观的经济环境和行业发展使行业内主要竞争者信用水平保持稳定，宏观和行业的风险适中	1
负面	当前宏观的经济环境和行业发展不利于行业内主要竞争者信用水平的提高，宏观和行业的风险较高	0.85
发展中	当前宏观环境和行业发展对于行业内主要竞争者的影响多样，难以得到准确结论。或者未来发展趋势和影响尚待观察。	0.95

B.指标异动

为了完善预警体系，在评分体系值外加入一套异动预警体系，将财务指标异动加入到预警当中。通过比较 N 期和 N-2 期的体系内财务指标，如果财务指标的变动超过评分体系内的两个区间，或者指标超过最高风险区间的临界值，则要以预警进行标识，并自动将本参数评分降至 0 分。采用 N 与 N-2 期的跨期指标对比，是为了屏蔽粉饰性财务技巧对于财务数据的影响，得到更为真实的信息。

C.舆情异动

利用 AI 技术收集市场的舆情信息，当上市公司出现重大负面消息，导致舆情预期下降，则做出风险预警提示，以便监管部门

进行进一步调查。由于上市公司的股价很大程度是由市场参与者预期所决定的，重大的负面消息容易引起连锁反应，导致上市公司出现巨大风险。

（4）风险计量及评级标准使用方法

根据上述风险预警指标体系，进行 Z 值统计评分。最终上市公司的风险评分对应到不同的分数区间，每个分数区间对应了不同的风险等级，对于低风险等级的上市公司进行风险预警。具体的评级标准如表 8 所示。

表 8：风险划分

风险程度	等级	对应评分	结论
安全级	A	80-100	被评主体具有极高的信用质量、极低的预期信用风险、极强的财务安全质量和偿付能力
	B+	70-80	被评主体具有高的信用质量、低的预期信用风险、强健的财务安全质量和偿付能力，投融资价值高，抵御风险的能力较强
	B-	60-70	被评主体具的中性偏好信用质量、偏低预期信用风险、稳定的财务安全质量和偿付能力，具有一定的投融资价值，拥有抵御风险的能力
关注级	C	50-60	被评主体具有一定的风险防范能力，但防范能力比较脆弱，信用风险正在形成，就个别问题需要关注，应定期察看其趋势
风险级	D	25-50	被评主体财务脆弱，违约可能性很高，受经济环境和经济条件的影响较大，存在很高财务风险，上市公司有 ST 可能性
	E	0-25	被评主体短期内存在巨大的财务风险，出现风险的可能性非常高

四、上市公司知识图谱的构建方法

智能监管涉及到对上市公司及其法人、股东潜在问题的研究，借助知识图谱技术可以发现实体背后二度甚至是三度关系及异常问题，从而需要对本体进行研究。本体定义了知识图谱中的数据

模式，所以知识图谱的构建很大程度上依赖于本体研究的成果。

本体意味着知识的共享和重用，在计算机科学领域中是指对领域知识的建模*。传统的人工生成本体的方法因为需要大量人力的投入并且效率不高，已经不能满足大量生成本体的需求，如何有效地生成可用性高的本体已经成为一个迫切需要解决的问题。为此，科学家们研究并开发了一系列本体学习的系统和方法。本体学习可以理解为从网络上的半结构化或者非结构化信息中自动或者半自动地生成本体，是提取语义信息的一个过程。

本项目在语义 Web 领域的研究中，已经进行了本体学习的深入研究并取得了一系列成果。郑海涛等基于大规模的领域本体采用潜在语义分析(Latent Semantic Analysis)方法，学习生成和文本集密切相关的小本体，实验表明该方法比其他的聚类算法有更好的 F-Measure 值，并能较准确地抽取关键的领域概念；郑海涛等利用基因本体(Gene Ontology)学习生成本体来对医疗文本进行聚类；郑海涛等还开发了 KABiCo 系统，该系统允许用户导入自己关心的领域知识来学习生成与文本集关联的本体。

本体学习的模型从学习方法的角度可以归结为以下几类:

一是基于语法分析的本体学习模型: 该模型主要采用语法分析器，对自然文本中的名词和动词进行标注，利用聚类的方法来构建领域内的概念，基于自然语言表达中的模式来分析概念之间

* 在计算机科学与信息科学领域，理论上本体是指一种“形式化的，对于共享概念体系的明确而又详细的说明”，实际上本体是对特定领域之中某套概念及其相互之间关系的形式化表达。本体一般可以用来针对某一领域的属性进行推理，亦可用于对该领域进行建模。

的关系，从而生成该领域的本体。该模型代表性的系统是 ASIUM 和 SVETLAN,此模型生成的本体后期需要一个用户来验证其有效性的过程，因此效率相对较低。

二是基于统计分析的本体学习模型：该模型主要利用统计的方法来计算各个概念在文本集中出现的频率和多个概念共同出现的频率，从概率的角度分析概念的重要性和概念之间的关系。该模型的代表性系统是 DODDLE II，因为概率计算的不确定性，所以该模型生成本体的准确率相对较低。

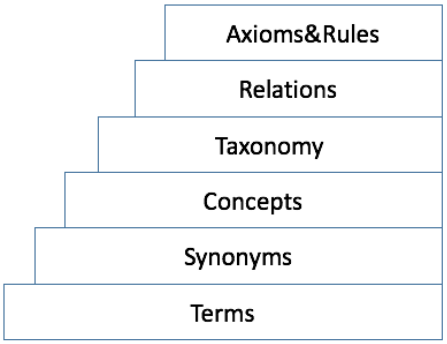
三是基于逻辑分析的本体学习模型：以一个已有的小本体为核心，HASTI 和 SYNDIKATE 等系统逐步地从自然文本中学习概念以及他们之间的关系，从而扩展该本体成为某个领域的知识载体。该模型采用的学习方法包括：基于符号的方法(Symbolic Approach)、语言分析方法(Linguistic Approach)、模板驱动的方法(Template Driven Approach)和启发式方法(Heuristic Method)。该模型需要用户预先构建一个有效小本体，否则生成的本体准确性不高，因此前期需要较大的时间投入。

四是基于混合策略分析的本体学习模型：TEXT-TO-ONTO 采用关联规则(Associate Rules)、规范概念分析(Formal Concept Analysis)和聚类(Clustering)等学习方法，从结构化、半结构化和非结构化数据中提取概念和语义关系；WEB→KB 结合贝叶斯学习(Bayesian Learning)和一阶逻辑规则(First Order Logic

Rules)两种方法，从网络文本中学习本体中的实例(Instance)及实例的提取规则；OntoLearn 结合 WordNet 和词频统计的方法对文本中的概念进行识别；David 等采用点互间信息(Pointwise Mutual Information)方法和基于模式的知识获取方法提取非分类的语义关系(Non-taxonomic Relationship)；OntoBuilder 使用词频统计和模式匹配的方法学习生成本体。该模型是目前本体学习的发展主流，但是这些方法都是静态的本体学习方法，即和时间无关。当文本信息随着时间变化的时候，这些方法没有相应的机制动态地从文本中提取语义信息，导致生成的本体质量不高。

本项目拟采用层次的结构来构建本体，如图 16 所示，自底向上依次包含术语(Terms)、同义词(Synonyms)、概念(Concepts)、分类(Taxonomy)、关系(Relations)以及公理与规则(Axions&Rules)。

图 16：本体的构建层次图



（一）关系图谱构建方案设计

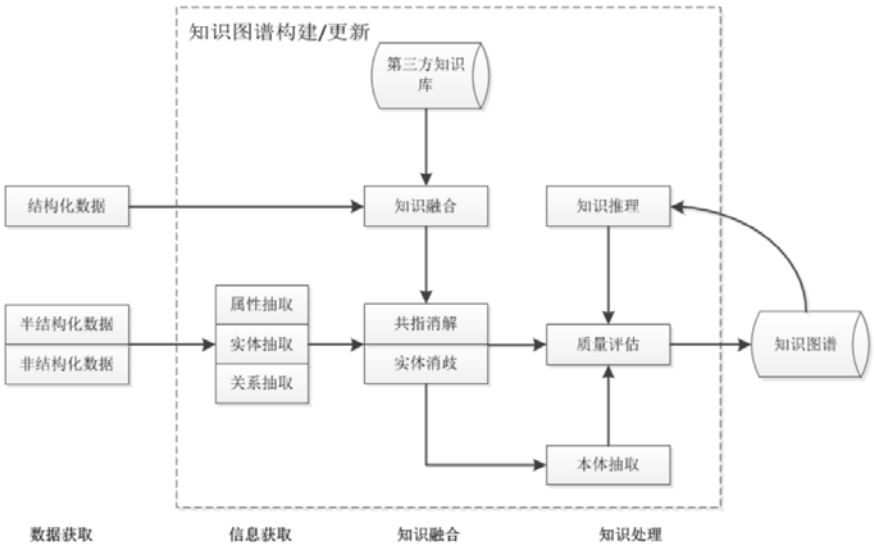
构建一个金融领域的知识图谱的主要原理和其他开放领域的图谱构建并没有太大的区别，本质上也是收集不同来源的数据，

并抽取出其中的关系集合，我们称这种集合为二元关系集合，一般用（实体 1，关系，实体 2）三元组的方式来表达，这些三元组通常称为元数据。参考上述通用方法，下面会详述对于上市公司的关系图谱如何建立的。

1.关系图谱构建框架

知识图谱的构建理论很多，一般来说都有一些通行的方法，我们会先做一下阐述，后面会提到怎么去搭建一个图谱的基本架构。

图 17：图谱架构



图谱的建立需要下面的 5 个步骤:

（1）非结构化数据。数据来源存在于不同的架构体系而且表现形式也很难统一，接口随领域不同而不同。

（2）结构化数据本身经过严格的数据清洗加工，其质量已经很高并且冗余度很低，可以直接使用。而非结构化数据可以用实

体抽取技术将数据里面的实体关系抽取出来。

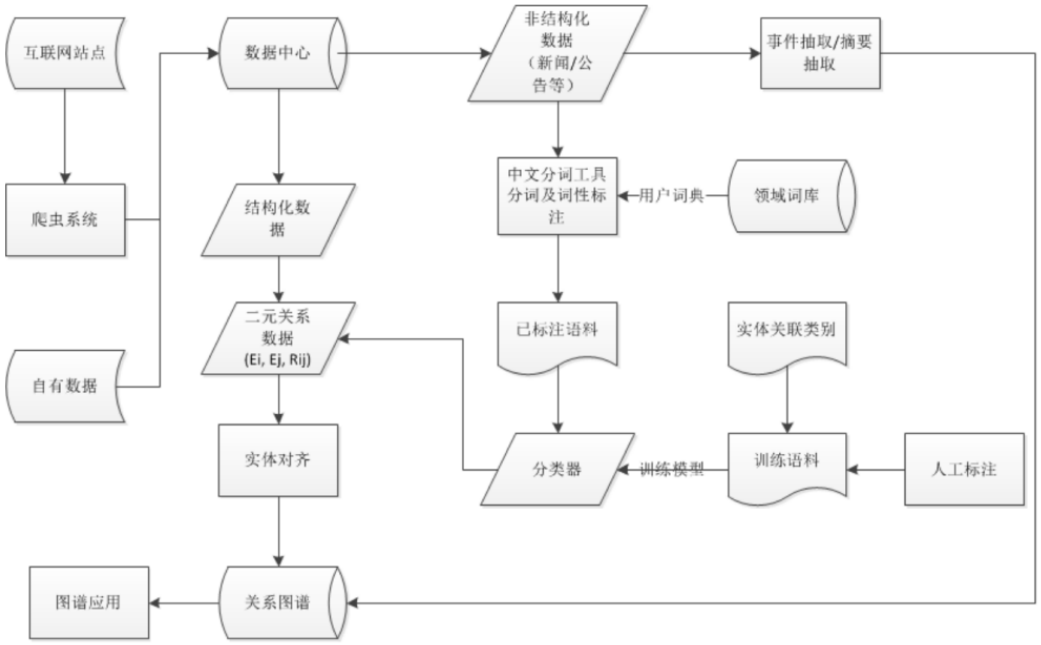
（3）知识提取完成后，从基础数据中得到了海量具备实体关系表征的元数据，这些数据还需要进行下一步的处理，因为这些数据还是缺乏层次以及存在错误重复的问题，也没有进行有效的组织。为了得到高质量的结构化数据，还需要对数据进行清洗和合并。

（4）通过上面的步骤基本能得到海量高质量数据。数据处理的阶段，需要将数据用人类逻辑进行抽象并组织，进行知识的模型构建，让数据的组织符合人类的认知。这个阶段的工作需要人的高度参与。

（5）知识图谱的构建需要不停地升级迭代，随着知识的不断更新，上面的数据获取以及知识体系的构建也要不停地更新。

综上，对于关系图谱的建设，除了获取结构化的数据外，更重要的是知识体系对海量数据支撑的设计。这个体系设计含有实体、概念、层次以及逻辑，也即本体抽取过程。而将现存的知识升维和抽象是一个非常困难的过程。另外，关系图谱的建设复杂的原因就在于其把建立所有实体和概念及其关联关系的库作为建设目的，这在人的认知及行业区分上没有明显的边界。对于财经领域，由于行业具备很垂直的属性，边界清晰，在对领域的知识描述方面不存在复杂的本体抽取。

图 18：知识图谱建设框架



2.识别命名实体

（1）知识图谱建设第一要务是识别命名实体。在进行中文命名实体识别前对文本进行分词是中文语言的特殊性决定的，本文是中文财经领域的图谱构建，所以主要解决 3 个工作：分词、标注和命名实体。

本文以长短时记忆神经网络模型（LSTM）来做命名实体的识别技术。LSTM 是深度学习的中 RNN 的一种特殊形式，是一种特征学习的方法，在文中由于需要对复杂结构的语言和序列进行处理，这是选择 LSTM 的首要原因。LSTM 把输入时间进行序列化处理，这样就和传统的神经网络有了质的区别。图 19 是传统 RNN（循环神经网络）结构。

图 19：神经网络结构

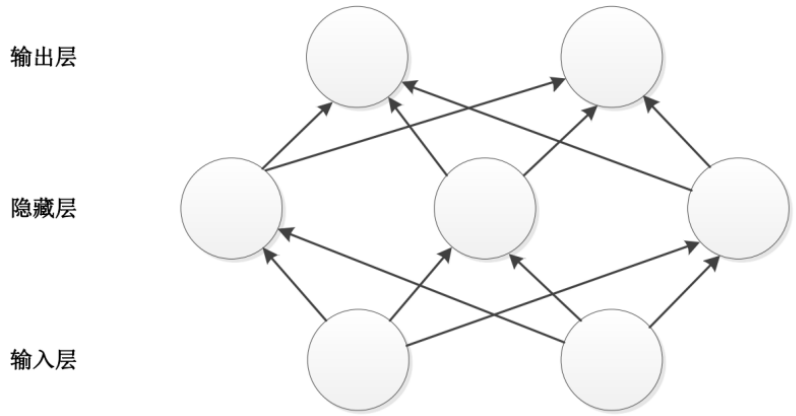
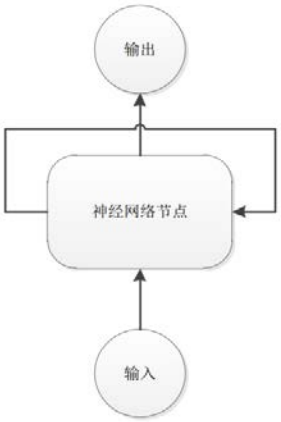


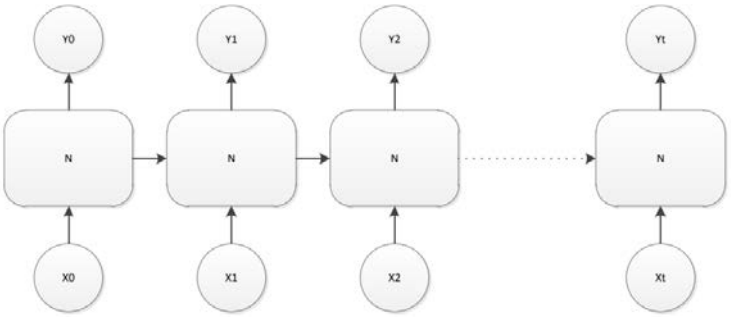
图 19 中，输入层节点与隐藏层节点相互连通，隐藏层节点与输出层节点相互连通，层和层中间的连线是一个权重值。模型需要训练出一个矩阵，这个矩阵是输入到输出的线性方程，最终输出采用激活函数来进行非线性激活。这就是前向传播的过程。训练时，为了得到最小错误率的矩阵，需要定义损失函数来做最优解，梯度下降法通常可以达到这个目的。梯度下降法的原理是计算参数项在损失函数中的偏导数，学习速率慢慢接近损失函数的最优结果，而这个学习速率是预先设定好的。在处理相互独立的输入数据方面，传统神经网络结构的结果是令人满意的，但具有时间序列关系的数据方面，传统神经网络却体现出了不足。本文中对中文语言局部文本的理解往往和上下文本之间存在联系。作为输入序列的句子，每个序列项之间总存在着非独立的关联性。考虑到自然语言句子作为输入序列的这种场景，对传统的神经网络结构做了一些改进后就成为循环神经网络(RNN)，具体如图 20 所示。

图 20：RNN 网络结构



传统神经网络仅仅解决了输出只与输入有相关性的问题，而 RNN 设计考虑的是隐藏层节点与自我上一时刻状态的相关性，输出不光与当前输入相关，也与上一刻的输出相关。将图 20 按时间序列展开后如图 21 所示。

图 21：RNN 展开示意图



在展开的 RNN 结构中，i 时刻的输出y_i 与输入值x_i相关，同时又与上一刻输出y_i-2相关，y_i-1则与上一刻y_i-2 相关，这种链式传递让隐藏层节点能够将前面所有时序的状态参数进行前向传递，这样 RNN 具备了时间序列处理能力。这种处理时间序列的能力让 RNN 在机器翻译、语音识别等多个应用场景发挥了巨大的作用。 但是在实际应用中，这种简单的神经网络并不能真正地实际落地。因为学习长距离的依赖关系仅仅在理论上可行，在真实情

况下，“梯度消失”成为反向传播难以克服的重大问题，模型的学习往往不能传递距离较远的依赖关系而倾向于较近的关系。所以真实场景下，简单的 RNN 并不能输出好的效果，特别是输入时间序列和长期依赖有较多相关时更为突出。本文通过改变 RNN 的结构可以解决长依赖关系无法传递信息的问题，改进后的 RNN 称做长短时依赖循环神经网络（Long Short-Term Memory Networks）。LSTM 用记忆细胞中的几个门来控制输入，其可以控制当前输入、上一刻输入和细胞上一刻的状态值，它通过这种方式来选择或抛弃传递的参数，即卷积或者遗忘，以此实现传递长依赖信息的结果。LSTM 的结构如图 22 所示。

图 22：长短时依赖循环神经网络结构

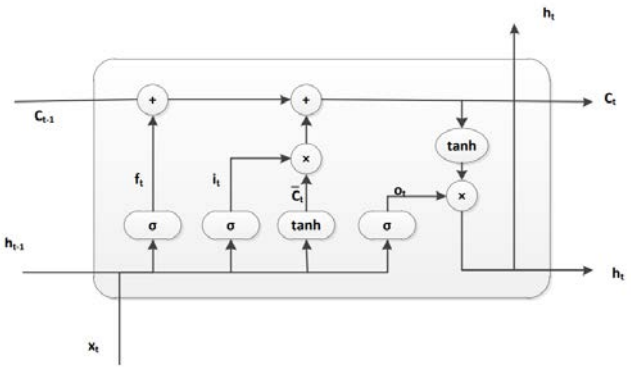


图 22 的神经元结构拥有一个细胞的状态参数 C_t，这个参数每次状态更新时会有条件地“遗忘”和“卷积”信息。每个细胞的输入有当前输入x_t以及上一刻输出的h_t-1。而 LSTM 的重点在于 x_t 和 h_t-1 再次建模并将有关参数存储在 C 中，可以保证在模型的训练过程中能够将长依赖信息保留，并对无用的信息进行“遗忘”。 本文采用了图 23 所示的不同结构。

图 23：采用的依赖循环神经网络结构

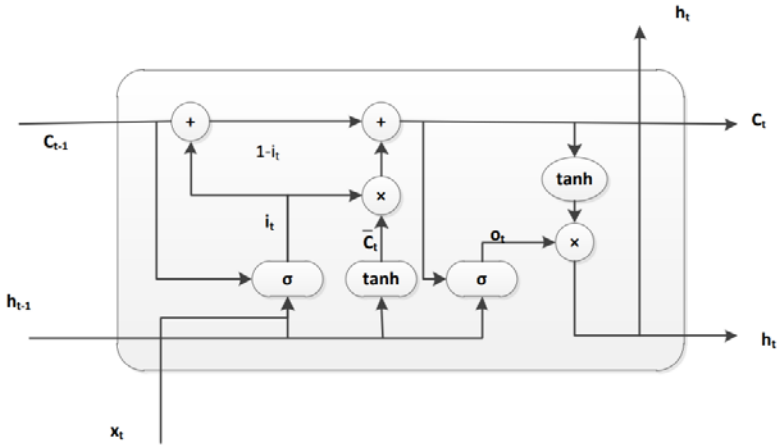


图 23 所示的机构与图 22 中的结构有明显不同，上一刻的状态C_t-1作为输入与 x_t和h_t-1一起计算。新的状态C_t在最终结果出来时需要加入到计算中。在文中采用双向长短时记忆网络（Bi-LSTMs）来做特性学习模型，因为在处理文本时不光要考虑当前文本前的依赖信息，也要考虑文后的依赖信息，这是因为自然语言需要考虑上下文信息。一个句子包含了 n 个输入词，网络首先计算每个单词 t 的左侧输出h_{t1}，将输入序列逆转再进行一次相同计算并得出右侧的输出h_{tr}，这样便是双向 LSTM。最后，两个方向的输出便可以很好地表达文章中的每个单词。

（2）在一个有限的标注集合内对于给定的输入序列标注标签并输出，即序列标注可以解决命名实体识别问题。用序列向量将文本表达出来并采用 LSTM 模型表达一段输入序列的特征，然后进行序列标注。直接使用 LSTM 模型的输出 h_t来作为特征进行标注是一个十分简单却有效的方法，这样的标注模型可以认为全部输入

序列之间具有相互独立性。将输入文本当作非独立的序列而构建概率模型，为了实现对序列进行模型搭建和实体标注的需求，本文采用条件随机场模型。表 9 是本课题采用的标签集合。

表 9：命名实体识别标注方案标签集

标签	说明
S-label	独立命名实体
B-label	命名实体头部
I-label	命名实体前部
O-label	命名实体后部
E-label	命名实体尾部

3. 训练模型参数

结合上述特征学习模型和序列标注模型，本文采用的识别方案整体结构如图 24 所示。

图 24：总体模型结构

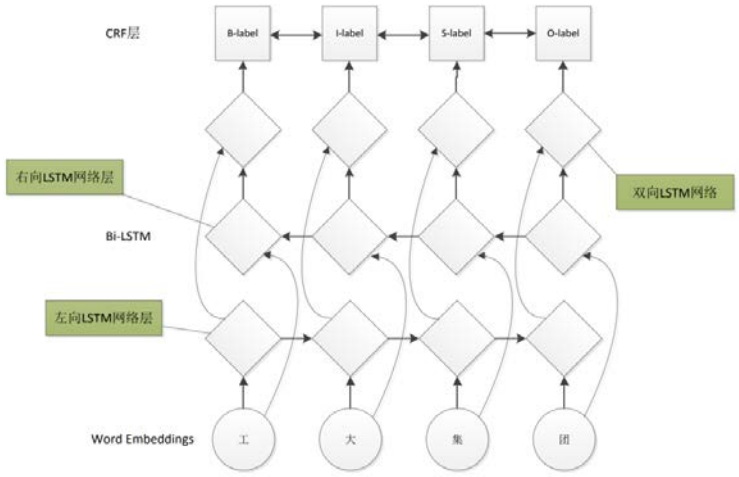


图 24 中圆为输入层，菱形节点代表 LSTM 模型的双向结构，正向（左侧）学习与逆向（右侧）学习组合可以作为输入的整体上下文表达，此结果当作条件随机场的输入参数，正方形为 CRF 模型的随机变量。此模型要训练的参数有条件随机场模型的序列

转移矩阵和 LSTM 模型中的参数，以及每层节点与节点的权值。

（二）关系图谱实现效果

实际使用关系图谱是可以通过一度或者二度查询来发现潜在的实体关系。展示效果如下:

1. 关键词搜索

在搜索框中输入企业名，系统会自动联想出相关结果集以备候选，选择候选集后，选择“看一下”搜索按钮。搜索结果用列表输出（见图 25 和图 26）。

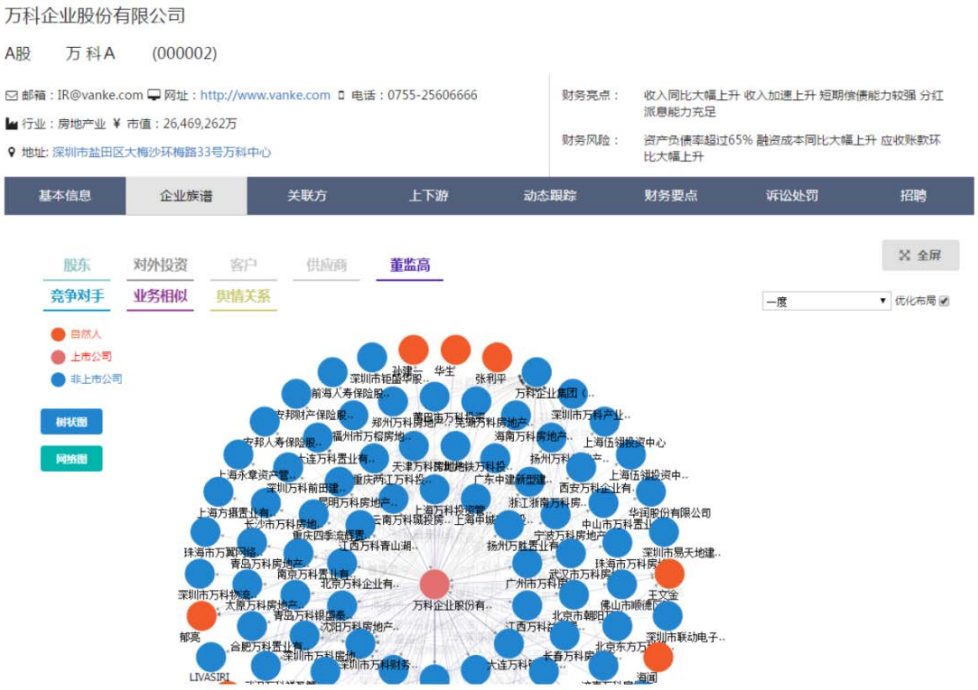
图 25：关键词搜索

图 26：搜索结果

2. 企业信息

公司 Tab 以不同的卡片展示功能区，默认显示的是一度关系。一度关系展示了和公司相关的企业名称（见图 27）。

图 27：公司信息页面



3. 实体颜色标注

图 28：关联图谱

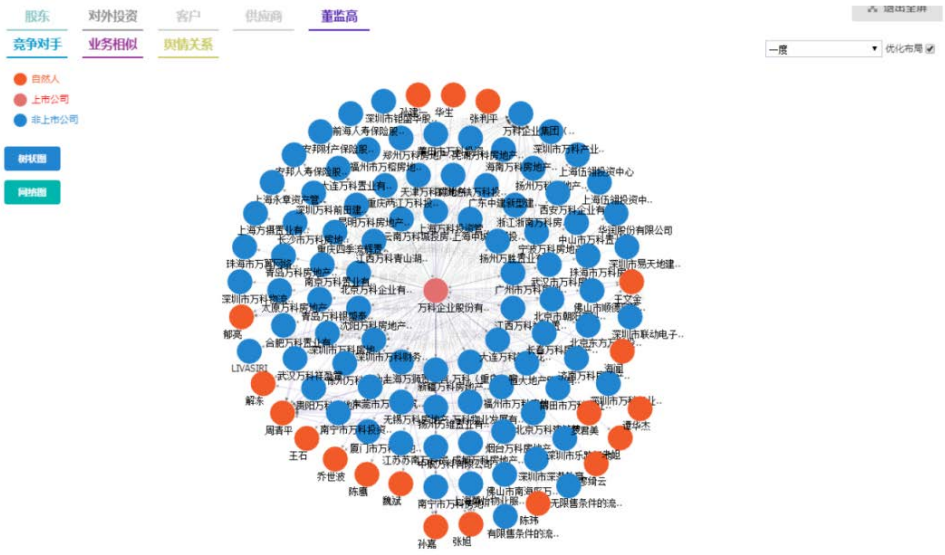
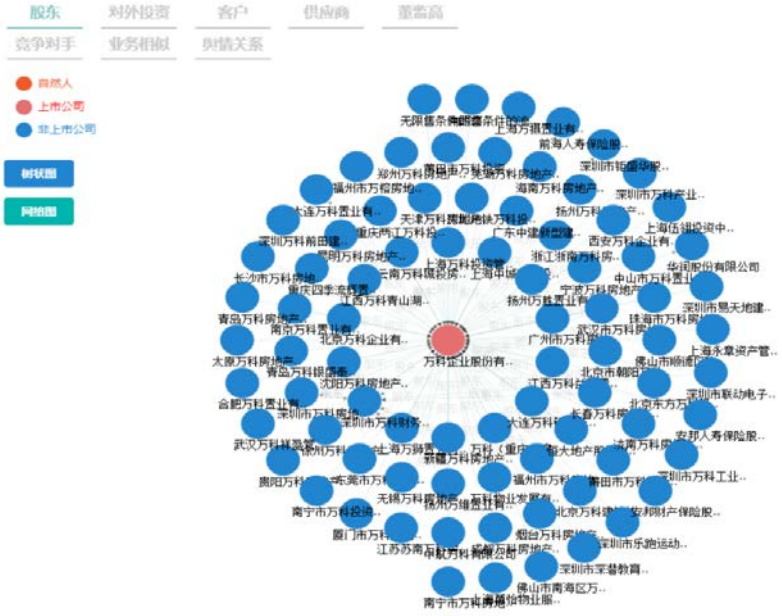


图 28 是公司的一度关系，实体类型用不同的颜色进行区分，蓝色的为非上市企业，红色为上市企业，橙色为自然人。

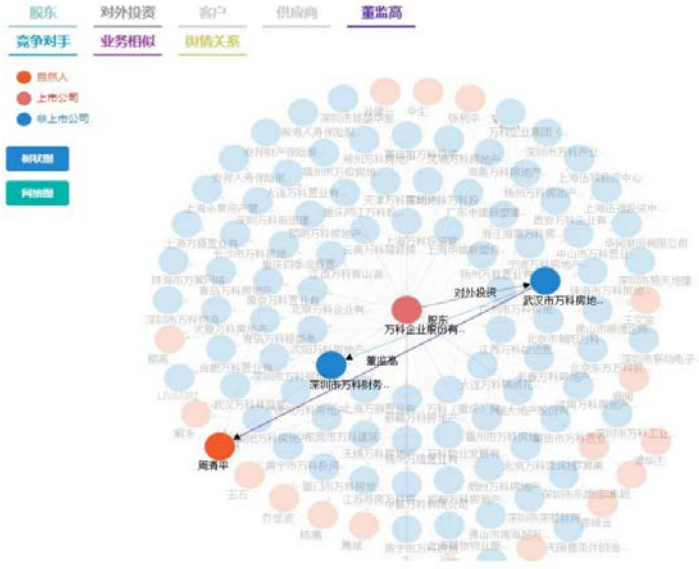
图 29：筛选关联图谱信息



可以用不同颜色的线条来表示不同的实体关系类型，图 29 是股东关系的效果。

4. 关联关系

图 30：指定节点到中心节点的关系

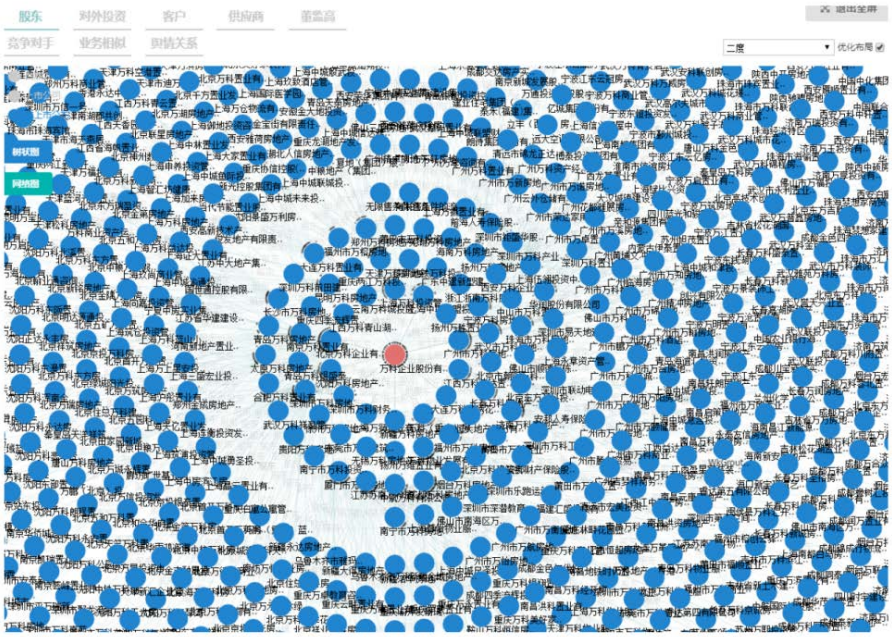


在图 30 中，点击指定节点，将返回该节点到中心节点的关系子图，展示了该公司与中心节点之间包含了 4 个节点的关联关系。

5. 二度关联关系

选择二度关联关系将返回中心节点和间接相关的公司与人的关系图谱，如图 31 所示。

图 31：二度关系查询



（三）知识图谱存储架构

知识图谱是一种图结构，因此需要采用图形数据库。图形数据库将点、线、面等基本元素按一定数据结构（通常为拓扑数据结构）建立起数据集合。Neo4j 是一个 NoSQL 图形数据库，它是用 Java 实现的，具备高性能特点。图数据库可以更好地存储由节点关系和属性构成的网络。

高可用性只能在企业版中可用，Neo4j 企业版高可用提供以

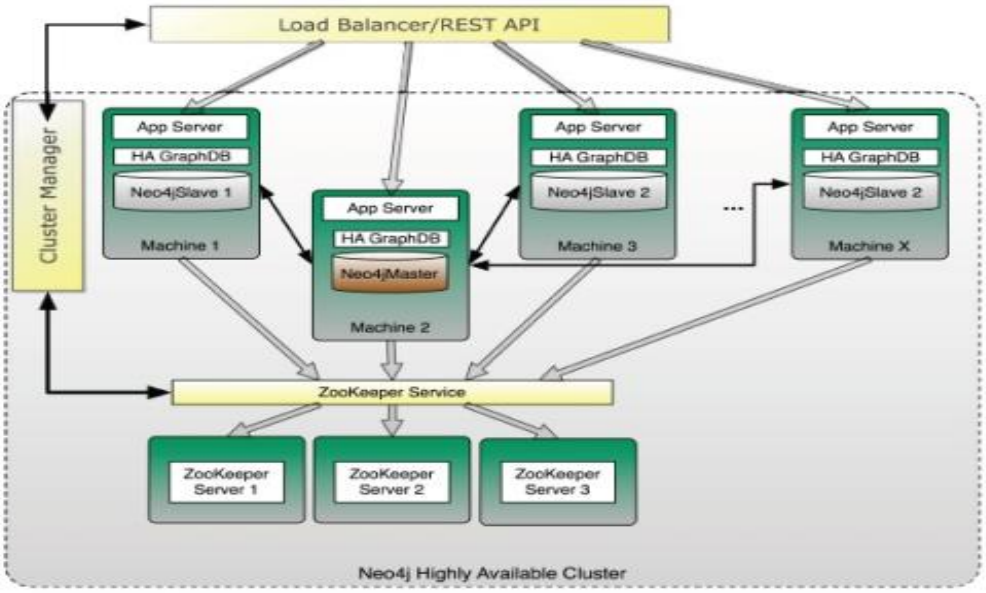
下两个主要特征:

1. 使用多台从数据库设置可以替代单台主数据库的容错，可以在硬件设备出问题时使数据库具备完善的功能和读写操作的能力。

2. 具有比单台数据库处理更多的读取负载处理的横向扫描主读架构。

Neo4j 高可用设计的目的是为了从一台到多台机器事务的操作简单，已存在的应用中并不需要做任何更改。

图 32：Neo4j 高可用架构



五、研究总结

总体而言，基于深度学习和知识图谱的上市公司的智能监管研究主要涉及公司基本信息及监管记录的集中展示、动态调整、系统预警，使分管人员更迅速、全面地了解公司，更及时、有效地发现公司的潜在风险。具体目标包括以下三个层面:

一是提升监管人员对公司情况掌握的深度和广度。利用大数据等技术手段，对上市公司进行多维度、全历史的画像，全面展示公司历史沿革、股东情况、关键人员、关系图谱、财务运营、同业比较、重要交易、舆情股价、诚信档案和监管评价等信息。在此基础上，以上市公司潜在风险为导向，针对性地结合股东行为、合规运作、经营状况等关键信息，对上市公司进行实时动态的风险评价打分。根据风险得分对上市公司进行分级，确定哪类公司或行为需要采取额外的监督，使有限的监管力量可以投入需要加强监管的高风险公司中。

二是提升监管人员发现问题和识别风险的能力。利用云计算、人工智能、机器学习等技术手段，增强上市公司监管智能化水平，实现对违规行为的辅助识别和对异常风险点的发现预警。例如，对简单违规事实，如业绩预告违规、董监高窗口期买卖股票等，由系统进行实时运算和识别提醒，通过增加机器的智能把关，提高监管效率和反应速度。

三是提升上市公司一线监管的实时性和有效性。伴随上市公司数量的增加，以及违规行为与潜在风险的日趋隐蔽和复杂，监管人员事中监管压力显著加大。一旦发生较大的风险事件，监管识别、判断和处置所需要的时间窗口越长，对投资者造成的影响往往也越大，对市场的破坏性越严重。鉴于此，监管的及时性就显得尤为重要，也由此决定了监管的有效性。在这一背景下，只

有通过发展监管科技，采用人机互补以及跨市场监控信息共享的模式，对上市公司信息披露和微观行为进行实时、全貌监控，才能提高监管识别和处置风险的能力，增强监管实效，维护市场稳定运行。

（上交所技术有限责任公司、同济大学、
深圳市智搜信息技术有限公司供稿*）

* 研究单位：上交所技术有限责任公司、同济大学、深圳市智搜信息技术有限公司；课题负责人：陶睿，上交所技术有限责任公司高级经理；课题组成员：吴继春、郑海涛、谢胜强、毛子舒、徐丹、范鸿燕。

（此页无正文）

报送：证监会主席、纪检组长、副主席。

证监会各部门，各派出机构，各会管单位。

分送：协会领导，理、监事，专业委员会主任、副主任委员，各部门，存档。