

# 知识图谱构建与可视化系统展示

LevenKoko

## 一、实验原理

在2022年全国投资者保护宣传日，中证协正式发布《2021年度证券公司投资者服务与保护报告》显示，截至2021年底，我国个人股票投资者已超过1.97亿，基金投资者超过7.2亿。抛开炒股技术不讲，这么多的股票数据非常难找与统计，但我们能够使用网络爬虫爬取股票的各种详细数据，再将这些数据进行可视化整理与分析，形成知识图谱。

在本实验中，我们使用 *Python Selenium* 库以驱动浏览器执行特定的动作，爬取九方智投的“沪深京A股”栏中的股票相关信息数据并存储到本地。随后使用 *Neo4j* 工具对爬取后的数据组织为知识图谱。

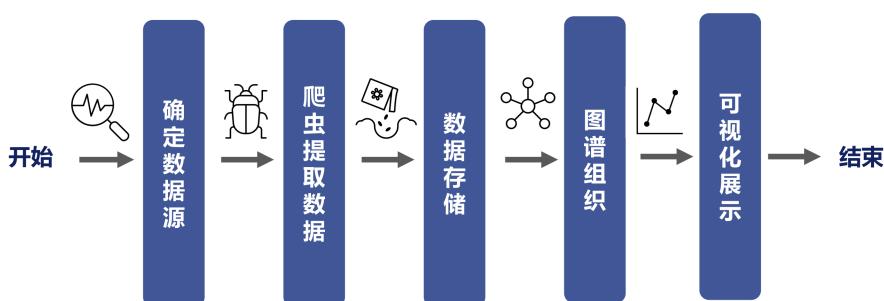
## 二、实验目的

在本实验中，我们的最终目的是爬取九方智投中的沪深京A股数据中的**股票概念、股东、基金、公告**等数据，并形成**参股、持股、发布公共、概念属于**等图形关系，以便后来数据分析者能够更加便捷的对这些数据进行**量化统计分析**。

## 三、实验内容

(1) **爬虫部分**: 爬取相关股票数据并以 *csv* 格式保存。表格包括《股票基础数据》《股东数据》《基金持股数据》《股票概念数据》《股票公告数据》。

(2) **知识图谱部分**: 使用 *Neo4j* 配合 *py2neo* 库对上一部分中爬取到的数据组织为知识图谱，并将其存储在 *Neo4j* 数据库中进行可视化展示。创建 5 类实体：**公告、概念、股东、股票、基金**，创建 4 类关系：**参股、持股（基金）、发布 公告、概念属于**，最后使用 *Cypher* 语句查询其中一些关系进行可视化展示。

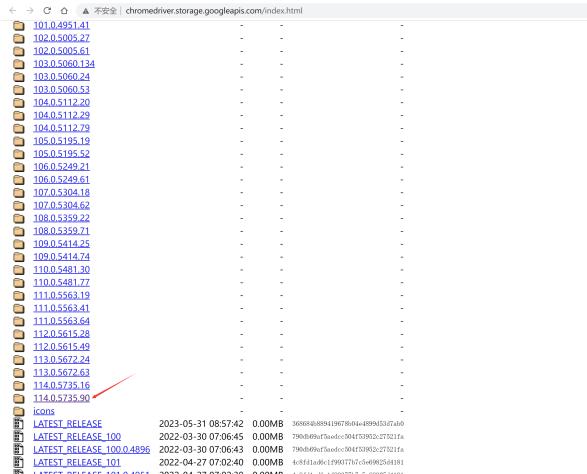


## 四、实验步骤

### (1) 爬虫部分

#### Step 1: 安装 selenium 库, ChromeDriver 等必要环境

使用 `pip install selenium` 安装 `selenium` 库, 在 [官方库](#) 中下载对应浏览器的 `ChromeDriver`。



#### Step 2: 声明浏览器对象并访问目标页面抓取股票列表

以下为声明浏览器对象的方法:

```
34 def get_browser():
35     option=ChromeOptions()
36     option.add_experimental_option('excludeSwitches',['enable-automation'])
37     option.add_experimental_option('useAutomationExtension',False)
38     # 设置不显式地显示浏览器
39     option.add_argument('--headless')
40     browser=webdriver.Chrome(options=option)
41     browser.execute_cdp_cmd("Page.addScriptToEvaluateOnNewDocument",
42                             {'source': 'Object.defineProperty(navigator,"webdriver",{get:()=>undefined})'})
43     browser.implicitly_wait(10)
44     return browser
```

以下为获取需要抓取股票列表的代码:

```
59 def scrape_stock_list(browser,page_num):
60     browser.get(BASE_URL)
61     urls=[]
62     WebDriverWait(browser,20,0.5).until(lambda browser:len(
63         browser.find_element(By.XPATH,'//*[@id="__next"]/div/div[3]/div[2]/ul[20]/li[13]/span').text)>0)
64     for page in range(page_num):
65         if(page >= 4):
66             # continue
67             urls+=(parse_page(browser.page_source))
68             time.sleep(0.2)
69             ac=ActionChains(browser)
70             # 鼠标移动到下一页按钮上
71             ac.move_to_element(browser.find_element(By.NAME, 'whj_nextPage')).perform()
72             # 点击确定跳转至下一页
73             ac.click(browser.find_element(By.NAME, 'whj_nextPage')).perform()
74             time.sleep(0.2)
75             WebDriverWait(browser,10,0.5).until(lambda browser: len(
76                 browser.find_element(By.XPATH,'//*[@id="__next"]/div/div[3]/div[2]/ul[20]/li[13]/span').text)>0)
77             print(urls)
78     return urls
```

其中 `urls` 为存储后续需要爬取的股票网址 `url` 的 `list` 对象, `WebDriverWait` 表示为当其中对应 `xpath` 文本内容加载出来后在进行爬取, 后续采用 `for` 循环进行切换页面, 对于每一个页面都使用 `parse_page` 函数对 `html` 进行解析, 提取出需要用到的 `url`。

### Step 3：解析需要的股票的基本数据

基本数据包括了股票代码、股票名称、公司名称、公司省份及城市、上市时间、行业、公司介绍、公司法人、经理、电话等信息。

对此，我们访问对应股票站点中的“公司资料”页面，找出以下图中框选的需要爬取的信息：

The screenshot shows a stock information page for 'N海科新源' (Stock Code: 301292). The top navigation bar includes tabs for '最新动态', '公司资料' (highlighted in blue), '股东研究', '经营分析', '股本结构', and '盈利预测'. Below the tabs are sub-sections: '公司高管', '财务分析', '分红融资', '公司大事', '行业对比', and '公告看点'. The main content area is titled '详细情况' (Detailed Information) and contains a table with company details. Key fields highlighted with red boxes include: '公司名称' (Company Name: 山东海科新源材料科技股份有限公司), '经营范围' (Business Scope: 一般项目:新兴能源技术研发;新材料技术推广服务;化工产品生产(不含许可类化工产品);化工产品销售(不含许可类化工产品);食品添加剂销售(除依法须经批准的项目外,凭营业执照依法自主开展经营活动)许可项目:危险化学品生产;危险化学品经营;药品生产;食品添加剂生产;进出口代理(依法须经批准的项目,经相关部门批准后方可开展经营活动,具体经营项目以审批结果为准)), '法人代表' (Legal Representative: 张生安), '联系电话' (Phone Number: 0546-7061006), and '网址' (Website: http://www.hi-techspring.com/).

公司名称	山东海科新源材料科技股份有限公司	英文名称	Shandong Hi-Tech Spring Material Technology Co., Ltd.
经营范围	一般项目:新兴能源技术研发;新材料技术推广服务;化工产品生产(不含许可类化工产品);化工产品销售(不含许可类化工产品);食品添加剂销售(除依法须经批准的项目外,凭营业执照依法自主开展经营活动)许可项目:危险化学品生产;危险化学品经营;药品生产;食品添加剂生产;进出口代理(依法须经批准的项目,经相关部门批准后方可开展经营活动,具体经营项目以审批结果为准)		
公司介绍	山东海科新源材料科技股份有限公司坐落于省级经济开发区——山东东营高新区,简称“海科新源”,成立于2002年,公司历程20年,合作客户遍布国内外,主营锂电池电解液溶剂、药用辅料丙二醇、食品添加剂异丙醇、化妆品级1,3丁二醇、二丙二醇、二异丙醚等精细化学品的生产和销售,广泛应用于锂电、医药、烟草、日化、食品、涂料等领域。发行人主要从事碳酸酯系列锂电池电解液溶剂和高端丙二醇、异丙醇等精细化学品的研发、生产和销售。公司的主要产品有碳酸酯系列锂电池电解液溶剂和高端丙二醇、异丙醇等精细化学品,可广泛应用于锂离子电池电解液、医药、化妆品、香精香料、烟草等行业。公司股票于2023年7月7日在深圳证券交易所创业板上市。		
公司成立时间	2002-10-30	所属地区	—
公司注册资本(元)	2.23亿	所属行业	电力设备
法人代表	张生安	总经理	张生安
董事会秘书	陈保华	证券事务代表	—
实际控制人	杨晓宏		
联系电话	0546-7061006	传真	0546-7061006
网址	http://www.hi-techspring.com/	E-Mali	dongban@hi-techspring.com
联系地址	山东省东营市东营高新技术产业开发区邹城路23号		
注册地址	山东省东营市东营高新技术产业开发区邹城路23号	办公地址	山东省东营市东营高新技术产业开发区邹城路23号
合计师事务所	立信会计师事务所(特殊普通合伙)	法律顾问	北京市中伦律师事务所

我们找到对应的 xpath 值：

```
stock_code_name_xpath='//*
[@id="app"]/div/section/div[1]/section/div[1]/div/div/div[1]/h3/i/text()'
company_name_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[1]/div[2]/div/span/text()'
company_time_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[5]/div[2]/div/span/text()'
company_business_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[23]/div[2]/div/span/text()'
company_representative_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[9]/div[2]/div/span/text()'
company_manager_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[10]/div[2]/div/span/text()'
company_hangye_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[8]/div[2]/div/span/text()'
company_secretary_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[11]/div[2]/div/span/text()'
company_phone_xpath='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[14]/div[2]/div/span/text()'
company_place_pattern='//*
[@id="xxqk"]/div[2]/div/div/section/div/ul/li[18]/div[2]/div/span/text()'
```

同时，我们要对联系地址进行地址到“省”“市”的转换，因此我们使用正则表达式来进行提取：

```
city_pattern=re.compile('(.?省)?(.?市)?',re.S)
```

解析其中某一个 html 代码的过程如下：

```
1 def parse_company_data(html):
2     html=etree.HTML(html)
3     code_name = html.xpath(stock_code_name_xpath)[0].split(' ')
4     stock_code = code_name[0]
5     stock_name = code_name[1]
6     company_place=html.xpath(company_place_pattern)[0]
7     province='--'
8     city='--'
9     if re.search(city_pattern,company_place).group(1):
10         province=re.search(city_pattern,company_place).group(1)
11     if re.search(city_pattern,company_place).group(2):
12         city=re.search(city_pattern,company_place).group(2)
13     company_name=html.xpath(company_name_xpath)[0]
14     company_time=html.xpath(company_time_xpath)[0]
15     company_hangye=html.xpath(company_hangye_xpath)[0]
16     company_business=html.xpath(company_business_xpath)[0]
17     companyRepresentative=html.xpath(company_representative_xpath)[0]
18     company_manager=html.xpath(company_manager_xpath)[0]
19     company_secretary=html.xpath(company_secretary_xpath)[0]
20     company_phone=html.xpath(company_phone_xpath)[0]
21
22     return [stock_code,stock_name,company_name,province,city,company_time,company_hangye,company_business,companyRepresentative,
23             company_manager,company_secretary,company_phone]
```

然后我们将提取到的结果存入一个 *DataFrame* 对象中，转换为 *csv* 进行本地保存。

```
1 data=[company[0] for company in companys]
2 company_data_table=get_table_company_data(data)
3 company_data_table.to_csv('company_data_table.csv',index=False,encoding='gbk')
```

如下图为部分结果展示：

stock_code	stock_name	company_name	province	city	company_time	行业	company_by companyRepresentative	company_manager	company_secretary	company_phone
2 300817 双飞股份	浙江双飞无油轴承股份有限公司	浙江省 嘉兴市	2000/8/15 机械设备	公司主要由周引春	周引春	浦助金	0573-84518146			
3 301192 泰祥股份	十堰市泰祥汽车零部件有限公司	湖北省 十堰市	1997/7/29 汽车	公司专注于王世斌	王世斌	姜雷	0719-8306877-8999			
4 300657 兆信电子	厦门弘信电子技术集团股份有限公司	--	2003/9/8 电子	公司是由李强、	李强	宋钦	0592-3160382			
5 301488 C豪恩汽电	深圳市豪恩汽车电子设备股份有限公司	--	2010/1/13 汽车	主要股东为罗小平	罗小平	李小娟	0755-28032222			
6 301141 中科融业	浙江中科融业股份有限公司	浙江省 东阳市	2010/3/22 有色金属	永磁材料的吴中平	吴中平	范明	0579-86099583			
7 300475 香农芯创	香农芯创科技股份有限公司	安徽省 --	1998/9/16 电子	公司主营业务李小红	李小红	曾柏林	0563-4186119			
8 301186 超达装备	南通超达装备股份有限公司	江苏省 南通市	2005/9/15 汽车	公司实际控制人冯建军	吴洁	郭焱焱	0513-87735878			
9 300270 中威电子	杭州中威电子股份有限公司	--	2000/3/14 电子	从事的主要李一策	何珊珊	孙玲	0571-88373153			
10 301007 德迈仕	大连德迈仕精密科技股份有限公司	辽宁省 --	2001/1/10 汽车	公司另一家何建平	何建平	孙百芸	0411-62187998-2066			
11 300008 天海防务	天津融合商务装备技术股份有限公司	--	2001/10/29 国防军工	主营业务涉及占金锋	占金锋	董文婕	021-60859800-9374,021-60859800-9837			
12 300928 华安鑫创	华安鑫创控股(北京)股份有限公司	--	2013/1/25 汽车	主营业务为何攀	何攀	易册	010-56940328			
13 2036 联创电子	联创电子有限公司	江西省 南昌市	1998/4/22 电子	专业从研曾吉勇	曾吉勇	黎红雷	0791-88161608			
14 2229 鸿博股份	鸿博股份有限公司	福建省 福州市	1999/6/15 轻工制造	公司业务潘锐辉(代)	黎红雷	王彬彬	0591-88070028			
15 600601 方正科技	方正科技集团股份有限公司	--	1993/1/15 电子	公司主要业务齐子鑫	陈宏良	黄飞照	021-58400030			
16 2567 唐人神	湖南唐人神集团有限公司	湖南省 株洲市	1992/9/11 农林牧渔	公司坚持陶以山	陶业	孙双烈	0731-28591247			
17 301045 天禄科技	苏州天禄光科技术股份有限公司	--	2010/11/9 电子	公司是一家梅坦	梅坦	佟晓刚	0512-66833339			
18 600230 沧州大化	沧州大化股份有限公司	--	1998/9/24 基础化工	公司是一家谢华生	杜森森	刘晓婧	0317-3556143			
19 300177 中海达	广州中海达星导航科技股份有限公司	广东省 广州市	2006/6/21 国防军工	主营业务所廖定海	李洪江	黄金矩	020-22889398			
20 600105 永鼎股份	江苏永鼎股份有限公司	江苏省 苏州市	1994/6/30 通信	光通信业务董庆海	路庆海	张国栋	0512-63272489			
21 2409 雅克科技	江苏雅克科技股份有限公司	--	1997/9/10/29 电子	公司业务包沈琦	沈琦	张晓宇	0510-87126509			
22 300127 银河磁体	成都银河磁铁股份有限公司	--	2001/3/23 有色金属	从事钐钴戴成	吴志坚	朱鹤文	028-61838299			
23 2115 三维通信	三维通信股份有限公司	浙江省 --	1993/5/13 通信	主要业务赵李越伦	李越伦	任锋	0571-88923377			
24 688668 鼎通科技	东莞市鼎通精密科技股份有限公司	广东省 东莞市	2003/6/11 通信	公司主要王成海	王成海	王晓兰	0769-85377166-609			
25 601127 龙立斯	龙立斯集团有限公司	--	2007/5/11 汽车	公司是一家张正萍	张正萍	申薇	023-65179666			
26 600353 光光电子	成都旭光电子股份有限公司	--	1994/2/28 电子	公司是一家刘卫东	张纯	熊尚荣	028-83967599			
27 2453 华软科技	金陵华软科技股份有限公司	--	1999/1/13 基础化工	主要业务包瞿辉	一	吕博	0512-66571019			
28 300243 瑞丰高材	山东瑞丰高分子材料股份有限公司	--	2001/10/26 基础化工	公司从事周仕斌	刘春佑	赵子阳	0533-3220711			
29 300445 康斯特	北京康斯特仪表科技股份有限公司	--	2004/9/20 机械设备	康斯特是一姜维利	何欣	刘楠楠	010-56973355			
30 809 铁岭新城	铁岭新城投资控股集团股份有限公司	--	1996/11/5 房地产	土地一级开隋景宝	张铁成	迟峰	024-74997822			
31 600975 新五丰	湖南新五丰股份有限公司	湖南省 长沙市	2001/6/26 农林牧渔	公司主要何军	刘艳书	罗丽飞	0731-8449588-811			
32 2403 爱仕达	爱仕达股份有限公司	浙江省 温州市	1993/5/13 家用电器	公司一家陈合林	陈合林	李培伊	0576-86199005			
33 600719 大连热电	大连热电股份有限公司	辽宁省 大连市	1993/9/1 公用事业	公司主要田鲁炜	张永军	郭晶	0411-84498196			
34 688629 华丰科技	四川华丰科技股份有限公司	四川省 绵阳市	1994/1/12 国防军工	光、电连接杨艳辉	刘太国	蒋海才	0816-2330358			
35 300128 铸富技术	苏州铸富技术股份有限公司	江苏省 苏州市	2004/3/28 电子	主要业务包顾清	顾清	张锐	0512-62820000			
36 961 中南建设	江苏中南建设集团股份有限公司	--	1998/7/22 房地产	公司业务陈锦石	陈益含	梁洁	021-61929799			
37 688620 安凯微	广州安凯微电子股份有限公司	--	2001/4/10 电子	物联网智能NORMAN SHENGFA HU (胡胜发)	胡胜发	李瑞懿	020-32219000			
38 838 财信发展	财信地产发展集团股份有限公司	--	1996/10/26 房地产	公司从事的贾森	贾森	熊秋伟	023-67675707			
39 301163 泰德特股份	江苏泰德特种部件股份有限公司	江苏省 南通市	1994/6/13 电力设备	高端装备制造杨金德	许玉松	李林(代)	0513-80600008			
40 300283 温州永丰	温州永丰电子合金有限公司	浙江省 --	1997/9/11 电力设备	公司是一家陈晓	陈晓	严学文	0577-85515911			
41 104240 泰森国际	泰森国际集团有限公司	浙江省 乐清市	2000/11/17/24 电子	公司阳江吴俊	中国香港	周丽华	18677-87717711			

## Step 4: 爬取股东数据

基本步骤和第三步中类似，需要爬取下图中的数据。

The screenshot shows a table titled "十大股东" (Top 10 Shareholders) with the following data:

名次	股东名称	股东性质	股份类型	持股数(股)	占总股本比例	增减(股)	变动比例
1	山东海科控股有限公司	一般企业	限售流通A股	1.36亿	60.90%	新进	—
2	国金证券-招商银行-国金证券...	券商集合资产管理...	限售流通A股	557.41万	2.50%	新进	—
3	中保投资有限责任公司-中国...	投资、咨询公司	限售流通A股	500.25万	2.24%	新进	—
4	杭州璟侑投资管理合伙企业(...	其他金融产品	限售流通A股	443.70万	1.99%	新进	—
5	上海合银投资管理有限公司-...	投资、咨询公司	限售流通A股	443.70万	1.99%	新进	—
6	上海辰韬资产管理有限公司-...	投资、咨询公司	限售流通A股	388.24万	1.74%	新进	—
7	赵洪修	个人	限售流通A股	332.78万	1.49%	新进	—
8	深圳市达晨财智创业投资管理...	投资、咨询公司	限售流通A股	305.05万	1.37%	新进	—
9	苏民投资管理无锡有限公司-...	投资、咨询公司	限售流通A股	277.32万	1.24%	新进	—
10	深圳市安鹏股权投资基金管理...	其他金融产品	限售流通A股	194.12万	0.87%	新进	—
总计		—	—	1.70亿	76.34%	—	—

但是由于股东在其中以**表格的方式存储**，一种方式是对表格中每一个对象进行单独的 *xpath* 处理，但是比较复杂繁琐，由于不知道是否存在相应的库函数进行优化，我在实验中采用的方法是对 *xpath* 进行**字符串分割**，对于行与列的数字额外进行**字符串加法合并**。

如下为 *html* 代码的解析部分：

```
52 def parse_sharehold(html, id):
53     company_id = company_id_pattern.search(html).group(1)[7:]
54     html = etree.HTML(html)
55     subfix0 = '"]/text()'
56     subfix1 = ']/span/text()'
57     name0 = '//*[@id="SHNameTopTenSH"]'
58     nature0 = '//*[@id="SHNatureTopTenSH"]'
59     type0 = '//*[@id="ShareTypeTopTenSH"]'
60     prefix0 = '//*[@id="sdgd"]/div[2]/div/div/section/div[2]/div/div/div/div/div/table/tbody/tr['
61     number1 = ']/td[5]/span/text()'
62     rate1 = ']/td[6]/span/text()'
63     change1 = ']/td[7]/span/text()'
64     name = html.xpath(name0 + str(id) + subfix0)[0]
65     nature = html.xpath(nature0 + str(id) + subfix0)[0]
66     type = html.xpath(type0 + str(id) + subfix0)[0]
67     number = html.xpath(prefix0 + str(id + 1) + number1)[0]
68     rate = html.xpath(prefix0 + str(id + 1) + rate1)[0]
69     change = html.xpath(prefix0 + str(id + 1) + change1)[0]
70     return [company_id, name, nature, type, number, rate, change]
```

其中 *subfix*、*prefix* 为**字符串分割部分**，传入参数 *id* 表示我们需要提取表格中的第 *id* 行。调用上述代码形成 *DataFrame* 的部分如下：

```
73 def get_table_sharehold_data(data):
74     table = pd.DataFrame(columns=['stock_code', '股东名称', '股东性质', '股份类型', '持股数(股)', '占总股本比例', '增减(股)'])
75     # , '股东性质'
76     for item in data:
77         for id in range(10):
78             table.loc[len(table)] = parse_sharehold(item, id)
79
80     return table
```

最后进行 csv 的本地存储，结果如下：

	stock_code	股东名称	股东性质	股份类型	持股数(股)	占总股本比例	增减(股)
1	300817	周引春	个人	无限售流通A股\限售流通A股	6073.92万	41.74%	不变
2	300817	嘉善顺飞股权投资管理有限公司	投资、咨询公司	无限售流通A股	881.28万	6.06%	不变
4	300817	浦志林	个人	无限售流通A股\限售流通A股	791.42万	5.44%	不变
5	300817	嘉善腾飞股权投资管理有限公司	投资、咨询公司	无限售流通A股	725.76万	4.99%	不变
6	300817	顾美娟	个人	无限售流通A股\限售流通A股	492.48万	3.38%	不变
7	300817	沈持正	个人	无限售流通A股\限售流通A股	397.44万	2.73%	不变
8	300817	周锦洪	个人	无限售流通A股	321.49万	2.21%	-68.83万
9	300817	单亚元	个人	无限售流通A股\限售流通A股	304.13万	2.09%	不变
10	300817	浦四金	个人	无限售流通A股\限售流通A股	304.13万	2.09%	不变
11	300817	吕良丰	个人	无限售流通A股	201.29万	1.38%	-30.58万
12	301192	王世斌	个人	限售流通A股	5385.00万	53.90%	不变
13	301192	姜雪	个人	限售流通A股	1387.50万	13.89%	不变
14	301192	十堰众远股权投资中心	投资、咨询公司	限售流通A股	555.00万	5.56%	不变
15	301192	蒋在春	个人	限售流通A股	82.50万	0.83%	不变
16	301192	何华强	个人	限售流通A股	82.50万	0.83%	不变
17	301192	张云	个人	无限售流通A股	78.30万	0.78%	-3700
18	301192	李彦彦	个人	无限售流通A股	34.74万	0.35%	不变
19	301192	国泰君安证券股份有限公司	金融机构—证券公司	无限售流通A股	12.94万	0.13%	新进
20	301192	李栋	个人	无限售流通A股	12.52万	0.13%	新进
21	301192	舒志兵	个人	无限售流通A股	11.76万	0.12%	新进
22	300657	弘信创业工场投资集团股份有限公司	投资、咨询公司	无限售流通A股	8418.53万	17.24%	不变
23	300657	厦门海翼投资有限公司	投资、咨询公司	无限售流通A股	2930.46万	6.00%	不变
24	300657	巫少峰	个人	限售流通A股	1130.90万	2.32%	不变
25	300657	吴放	个人	无限售流通A股	999.99万	2.05%	119.99万
26	300657	新余善思投资管理中心(有限合伙)-善思慧成玖号私募证券投资基金管理人	其他金融产品	无限售流通A股	979.94万	2.01%	不变
27	300657	朱小燕	个人	限售流通A股	969.35万	1.98%	不变
28	300657	李毅峰	个人	无限售流通A股	903.97万	1.85%	不变
29	300657	李奎	个人	无限售流通A股	872.43万	1.79%	不变
30	300657	张洪	个人	无限售流通A股	783.58万	1.60%	-488.40万
31	300657	施伟	个人	无限售流通A股	742.80万	1.52%	新进
32	301488	深圳市豪恩科技集团股份有限公司	一般企业	限售流通A股	3284.50万	35.70%	新进
33	301488	罗小平	个人	限售流通A股	700.00万	7.61%	新进
34	301488	陈金法	个人	限售流通A股	680.00万	7.39%	新进
35	301488	深圳市华恩泰科技有限公司	一般企业	限售流通A股	580.00万	6.30%	新进

## Step 5：基金投资数据爬取

同样是一个表格，爬取过程与上述过程完全一致，只需要修改相关 xpath 部分即可。

需要爬取的数据如下：

基金持股							个股主力持仓 >
2023-03-31		2022-12-31	2022-09-30	2022-06-30	2022-03-31	来源：基金季报	
名次	基金代码	基金名称	持股数(股)	持仓市值(元)	占总股本比例	占总流通股本比例	占净值比例
1	002363	华安安康灵活配置混合A	2769.91万	3.53亿	1.96%	2.00%	3.49%
2	010659	民生加银质量领先混合A	831.44万	1.06亿	0.59%	0.60%	5.97%
3	011738	华安兴安优选一年混合A	699.76万	8907.92万	0.50%	0.50%	4.90%
4	000913	农银医疗保健股票	607.81万	7737.45万	0.43%	0.44%	3.96%
5	001312	华安新优选A	521.53万	6639.08万	0.37%	0.38%	3.21%
6	160607	鹏华价值优势混合(LOF)	370.20万	4712.69万	0.26%	0.27%	3.19%
7	010619	华安添利6个月债券A	303.02万	3857.44万	0.21%	0.22%	2.04%
8	399011	中海医疗保健主题股票A	300.26万	3822.32万	0.21%	0.22%	3.72%
9	009596	泰康创新成长混合A	278.84万	3549.63万	0.20%	0.20%	3.07%
10	010795	民生加银价值发现一年…	251.27万	3198.66万	0.18%	0.18%	6.14%

代码如下：

```
In [1]:  
1 def parse_fund(html, id):  
2     company_id = company_id_pattern.search(html).group(1)[7:]  
3     html = etree.HTML(html)  
4     fund_id_xpath0 = '//*[@id="jjcg"]/div[2]/div/div/section/div[2]/div/div/div/div/table/tbody/tr['  
5     fund_id_xpath1 = ']/td[2]/text()'  
6     fund_name_xpath0 = '//*[@id="FundNameFundHold"  
7     fund_name_xpath1 = '"]]/text()'  
8     fund_number_xpath0 = '//*[@id="jjcg"]/div[2]/div/div/section/div[2]/div/div/div/div/table/tbody/tr['  
9     fund_number_xpath1 = ']/td[4]/span/text()'  
10    fund_rate_xpath0 = '//*[@id="jjcg"]/div[2]/div/div/section/div[2]/div/div/div/div/table/tbody/tr['  
11    fund_rate_xpath1 = ']/td[6]/span/text()'  
12    if len(html.xpath(fund_number_xpath0 + str(id + 1) + fund_number_xpath1)) == 0:  
13        return None  
14    fund_id = html.xpath(fund_id_xpath0 + str(id + 1) + fund_id_xpath1)[0]  
15    fund_name = html.xpath(fund_name_xpath0 + str(id) + fund_name_xpath1)[0]  
16    fund_number = html.xpath(fund_number_xpath0 + str(id + 1) + fund_number_xpath1)[0]  
17    fund_rate = html.xpath(fund_rate_xpath0 + str(id + 1) + fund_rate_xpath1)[0]  
18    return [company_id, fund_id, fund_name, fund_number, fund_rate]  
19 def get_table_fund_data(data):  
20     table = pd.DataFrame(  
21         columns=['stock_code', '基金代码', '基金名称', '持股数(股)', '占总股本比例'])  
22     for item in data:  
23         for id in range(10):  
24             tmp = parse_fund(item, id)  
25             if tmp != None:  
26                 table.loc[len(table)] = parse_fund(item, id)  
27     return table  
28 fund_data_table = get_table_fund_data(sharehold)  
29 fund_data_table.to_csv('fund_data_table.csv', index=False, encoding='gbk')
```

爬取结果如下：

1	stock_code	基金代码	基金名称	持股数(股)	占总股本比例
2	300817	2083	新华鑫动力灵活配置混合A	141.40万	0.97%
3	300657	362	国泰聚信价值优势灵活配置混合A	788.00万	1.61%
4	300657	12173	国泰兴泽优选一年持有期混合A	288.37万	0.59%
5	300657	739	平安新鑫先锋混合A	194.00万	0.40%
6	300657	8415	国泰大制造两年持有期混合	150.00万	0.31%
7	300657	5746	国泰聚利价值定期开放灵活配置混合	73.50万	0.15%
8	300657	512100	南方中证1000ETF	57.58万	0.12%
9	300657	320003	诺安先锋混合A	40.00万	0.08%
10	300657	161039	富国中证1000指数增强(LOF)A	39.01万	0.08%
11	300657	10651	平安双季增享6个月持有债券A	37.65万	0.08%
12	300657	11807	平安研究精选混合A	36.96万	0.08%
13	301141	160628	鹏华中证800地产指数(LOF)A	2617	0.00%
14	301141	656	前海开源沪深300指数A	2617	0.00%
15	301141	160629	鹏华中证传媒指数(LOF)A	2617	0.00%
16	301141	6321	中欧预见养老2035三年持有(FOF)A	2617	0.00%
17	301141	160631	鹏华银行A	2617	0.00%
18	301141	213010	宝盈中证100指数增强A	2617	0.00%
19	301141	481009	工银沪深300指数A	2617	0.00%
20	301141	13381	中欧甄选3个月持有混合(FOF)A	2617	0.00%
21	301141	501211	民生加银优享6个月定开混合(FOF-LOF)	2617	0.00%
22	301141	512220	景顺长城中证科技传媒通信150ETF	2617	0.00%
23	300475	14600	博时回报严选混合A	11.25万	0.03%
24	300475	12650	博时半导体主题混合A	11.03万	0.03%
25	301186	1410	信澳新能源产业股票	55.62万	0.76%
26	301186	14254	信澳智远三年持有期混合A	22.08万	0.30%
27	301186	12608	信澳领先智选混合	17.03万	0.23%
28	300270	360001	光大保德信量化股票A	161.59万	0.53%
29	300270	6195	国金量化多因子A	27.94万	0.09%
30	300270	14805	国金量化精选A	12.01万	0.04%
31	300270	11231	光大保德信锦弘混合A	10.44万	0.03%
32	300270	4457	光大保德信多策略智选18个月混合	1.22万	0.00%
33	300270	970041	国海量化优选一年持有股票A	9100	0.00%
34	301007	14232	博时专精特新主题混合A	26.84万	0.17%
35	301007	15148	华安中证1000指数增强A	9.85万	0.06%

## Step 6：股票概念数据爬取部分

这一部分不是表格数据，但是以逗号隔开，违背了数据库中 ACID 的原子性。

601567 三星医疗 14.37 ↑ 1.31 +10.03% 收起 ^

最新动态 公司资料 股东研究 经营分析 股本结构 盈利预测  
公司高管 财务分析 分红融资 公司大事 行业对比 公告看点

公司概要 最新指标 近期重要事件 公司公告 股东分析 龙虎榜单 大宗交易 融资融券

公司概要

所属行业	电网设备 (630800)
涉及概念	物联网,电力物联网,RCS富媒体通信,充电桩,换电概念,仪器仪表概念,智能电网,沪股通,民营医院,健康中国,互联网医疗,高压氧舱,融资融券,转融券标的,纳入富时罗素,标普道琼斯中国,浙江
主营业务	一般项目:以自有资金从事投资活动;医院管理;工程管理服务;仪器仪表制造;仪器仪表销售;仪器仪表修理;电工仪器仪表制造;电工仪器仪表销售;供应用仪器仪表制造;供应用仪器仪表销售;机械电气设备制造;机械电气设备销售;通用零部件制造;电力设施器材制造;电力设施器材销售;电子元器件制造;电子元器件批发;电子元器件零售;有色金属合金制造;有色金属合金销售;有色金属压延加工;五金产品制造;五金产品批发;五金产品研发;五金产品零售;智能仪器仪表制造;智能仪器仪表销售;电容器及其配套设备制造;电容器及其配套设备销售;软件开发;软件销售;技术服务、技术开发、技术咨询、技术交流、技术转让、技术推广;通信设备制造;通信设备销售。变压器、整流器和电感器制造;配电开关控制设备制造;配电开关控制设备销售;配电开关控制设备研发;输配电及控制设备制造;智能输配电及控制设备销 ... <span style="float: right;">展开</span>

对此，我们使用 `split` 函数对逗号隔开的数据进行分离，产生一个一对多的关系表。

爬取得代码如下：

```

In [1]: 1
        2 def get_table_cept_data(data):
        3     table = pd.DataFrame(columns=['stock_code', '概念'])
        4
        5     for item in data:
        6         company_id = company_id_pattern.search(item).group(1)[7:]
        7         html = etree.HTML(item)
        8         target = html.xpath('//*[@id="gsgy"]/div[2]/div/div/section/div/ul/li[2]/div[2]/div/span/text()')[0]
        9         op = target.split(',')
       10        for x in op:
       11            table.loc[len(table)] = [company_id, x]
       12
       13    return table
       14 concept = [company[2] for company in companys]
       15 for i in range(len(concept)):
       16     concept[i] = concept[i] + company_data_table.loc[i][0]
       17 fund_data_table = get_table_cept_data(concept)
       18 fund_data_table.to_csv('concept_data_table.csv', index=False, encoding='gbk')
       19 print(fund_data_table)

```

爬取结果如下：

	stock_code	概念
1	300817	特斯拉
2	300817	汽车零部件概念
3	300817	人民币贬值受益
4	300817	风电概念
5	300817	军工
6	300817	昨日涨停
7	300817	昨日首板
8	300817	浙江
9	301192	新能源车
10	301192	汽车零部件概念
11	301192	融资融券
12	301192	新股与次新股
13	301192	转融券标的
14	301192	昨日涨停
15	301192	专精特新
16	301192	昨日首板
17	301192	注册制次新股
18	301192	湖北
19	300657	华为概念
20	300657	东数西算/算力
21	300657	汽车电子概念
22	300657	小米概念
23	300657	元宇宙
24	300657	宁德时代概念
25	300657	OLED
26	300657	无线耳机
27	300657	军工
28	300657	深股通
29	300657	电子烟
30	300657	PCB概念
31	300657	柔性屏
32	300657	手机产业
33	300657	融资融券
34	300657	转融券标的
35	300657	

需要注意的是，这些概念是**实时更新的**，如包含了**昨日首板**等概念，因此这一类数据需要**每日进行更新**。

## Step 7: 公告数据爬取

由于在股票页面没有公告这一个直接的链接，我们需要先任意进入一个页面后进行行业内跳转进行模拟点击，因此爬取 html 代码的部分与以上部分有一些不同：

```
165     logging.info('scraping 公司公告...')  
166     browser.find_element(By.LINK_TEXT, '更多').click()  
167     browser.switch_to.window(browser.window_handles[1])  
168     WebDriverWait(browser, 10, 0.5).until(lambda browser: len(  
169         browser.find_element(By.XPATH, '//*[@id="app"]/div/section/div[1]/section/div[1]/div/div[2]/ul/li[12]/div'))>0)  
170     # 模拟移动点击跳转  
171     time.sleep(0.3)  
172     ac = ActionChains(browser)  
173     ac.move_to_element(browser.find_element(By.XPATH, '//*[@id="app"]/div/section/div[1]/section/div[1]/div/div[2]/ul/li[12]/div')).perform()  
174     ac.click(browser.find_element(By.XPATH, '//*[@id="app"]/div/section/div[1]/section/div[1]/div/div[2]/ul/li[12]/div')).perform()  
175     time.sleep(0.3)  
176     # 等待公告加载完成  
177     WebDriverWait(browser, 10, 2).until(lambda browser: len(  
178         browser.find_element(By.XPATH,  
179             '//*[@id="MediaGszx0"]').text)>0)  
180     news=browser.page_source  
181     browser.close()  
182     browser.switch_to.window(browser.window_handles[0])
```

获取到页面的 html 页面后，我们进行html 的解析，同样是表格数据：

```
In _ 1 def parse_news(html, id):  
2     company_id = company_id_pattern.search(html).group(1)[7:]  
3     html = etree.HTML(html)  
4     news_date_xpath0 = '//*[@id="gsggDetail"]/div[2]/section/div/div/section/div/div/div/div/div/div/table/tr['  
5     news_date_xpath1 = ']/td[1]/text()'  
6     news_name_xpath0 = '//*[@id="ggTitleGszxDetail'  
7     news_name_xpath1 = '"]/text()'  
8     if len(html.xpath(news_name_xpath0 + str(id) + news_name_xpath1)) == 0:  
9         return None  
10    news_date = html.xpath(news_date_xpath0 + str(id+1) + news_date_xpath1)[0]  
11    news_name = html.xpath(news_name_xpath0 + str(id) + news_name_xpath1)[0]  
12    print(company_id)  
13    return [company_id, news_date, news_name]  
14  
15 def get_table_news_data(data):  
16     table = pd.DataFrame(columns=['stock_code', '时间', '公告'])  
17  
18     for item in data:  
19         for id in range(10):  
20             tmp = parse_news(item, id)  
21             if tmp != None:  
22                 table.loc[len(table)] = parse_news(item, id)  
23  
24     return table  
25  
26 news = [company[3] for company in company]  
27 for i in range(len(news)):  
28     news[i] = news[i] + company_data_table.loc[i][0]  
29 news_data_table = get_table_news_data(news)  
30 news_data_table.to_csv('news_data_table.csv', index=False, encoding='gbk')  
31 print(news_data_table)
```

爬取到的结果如下：

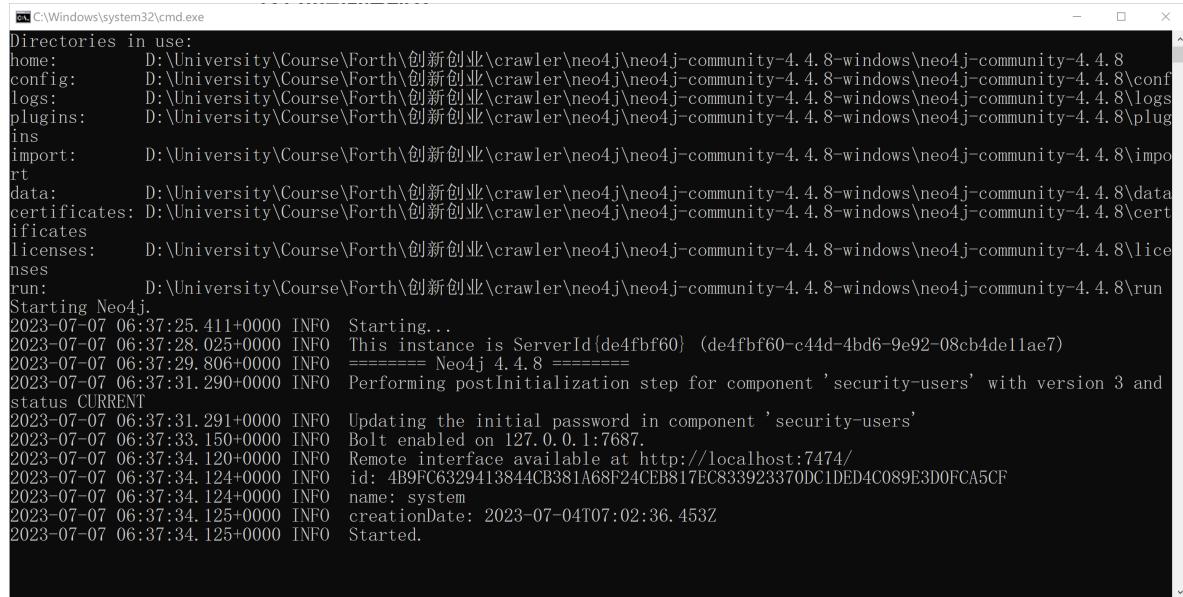
序号	stock_code	时间	公告
1		300817 2023/6/1	双飞股份-2022年年度权益分派实施公告
2		300817 2023/5/12	双飞股份-浙江六和律师事务所关于浙江双飞无油轴承股份有限公司2022年年度股东大会的法律意见书
3		300817 2023/5/12	双飞股份-2022年年度股东大会决议公告
4		300817 2023/5/5	双飞股份-关于召开2022年年度股东大会的提示性公告
5		300817 2023/4/26	双飞股份-2023年一季度报告
6		300817 2023/4/26	双飞股份-第四届监事会第十七次会议决议公告
7		300817 2023/4/26	双飞股份-第四届监事会第十七次会议决议公告
8		300817 2023/4/26	双飞股份-关于变更高级管理人员的公告
9		300817 2023/4/26	双飞股份-独立董事关于聘任公司副总经理的独立意见
10		300817 2023/4/26	双飞股份-第四届董事会第二十次会议决议公告
11		300817 2023/4/26	双飞股份-2023年第一季度报告披露的提示性公告
12		301192 2023/6/26	泰祥股份-关于控股子公司完成工商变更登记并换发营业执照的公告
13		301192 2023/6/20	泰祥股份-关于重大资产重组实施的进展公告
14		301192 2023/6/12	泰祥股份-股票交易异常波动公告
15		301192 2023/5/25	泰祥股份-2023年5月25日投资者关系活动记录表
16		301192 2023/5/22	泰祥股份-2022年年度权益分派实施公告
17		301192 2023/5/22	泰祥股份-关于参加湖北辖区上市公司2023年度投资者网上集体接待日活动的公告
18		301192 2023/5/8	泰祥股份-独立董事关于第三届董事会第二十二次会议相关事项的独立意见
19		301192 2023/5/8	泰祥股份-长江证券承销保荐有限公司关于十堰市泰祥实业股份有限公司向控股子公司提供财务资助的核查意见
20		301192 2023/5/8	泰祥股份-第三届董事会第二十二次会议决议公告
21		301192 2023/5/8	泰祥股份-第三届监事会第十九次会议决议公告
22		300657 2023/6/30	弘信电子-弘信电子投资者关系管理档案
23		300657 2023/6/19	弘信电子-与摩尔线程智能科技（北京）有限责任公司签订战略合作协议的公告
24		300657 2023/6/10	弘信电子-关于股东减持计划实施完成的公告
25		300657 2023/6/3	弘信电子-关于股东减持计划实施完成的公告
26		300657 2023/5/23	弘信电子-关于为子公司提供担保的进展公告
27		300657 2023/5/19	弘信电子-关于控股股东部分股份解除质押及再质押的公告
28		300657 2023/5/19	弘信电子-2022年年度股东大会决议公告
29		300657 2023/5/19	弘信电子-北京国枫律师事务所关于厦门弘信电子科技股份有限公司2022年年度股东大会的法律意见书
30		300657 2023/5/13	弘信电子-国信证券股份有限公司关于厦门弘信电子科技股份有限公司2020年度向不特定对象发行可转换公司债券的保荐总结报告书
31		300657 2023/5/10	弘信电子-关于为子公司提供担保的进展公告
32		301141 2023/7/6	中科磁业-股票交易异常波动的公告
33		301141 2023/7/4	中科磁业-关于公司扩能规划暨与东阳市高铁新城管委会签署投资协议的公告
34		301141 2023/7/4	中科磁业-关于调整公司组织架构的公告
35		301141 2023/6/6	中科磁业-浙江中科磁业股份有限公司2022年年度权益分派实施公告

到此为止，我们已经完成了所有数据的爬取，将进入下一个知识图谱展示步骤。

## (2) 知识图谱部分

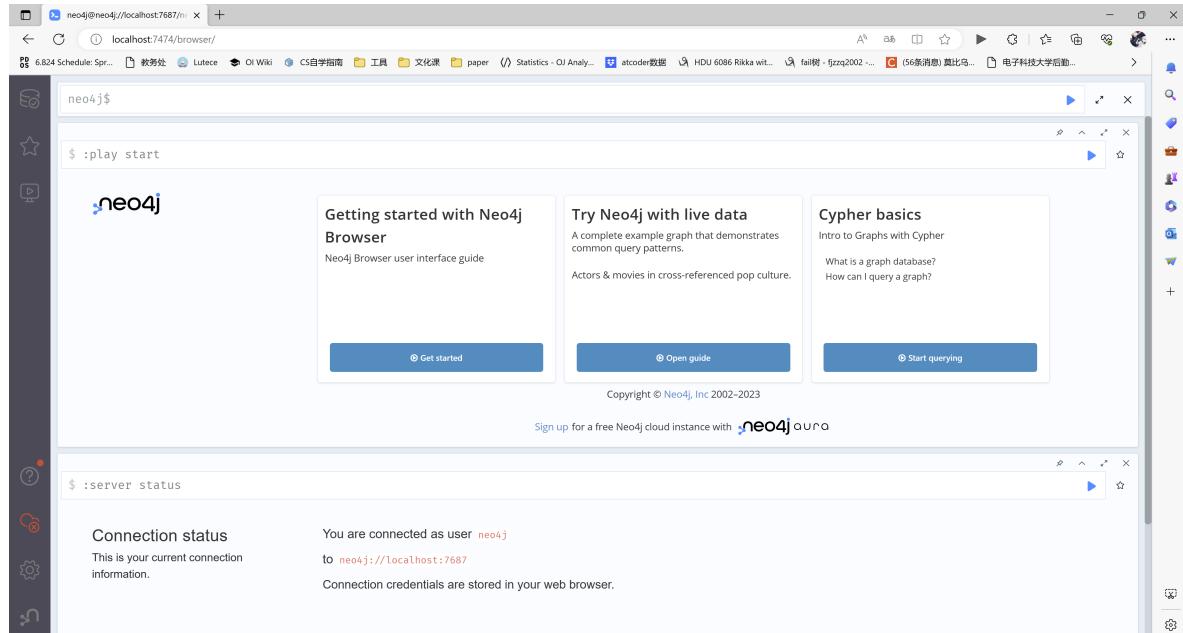
### Step 1: 安装 Neo4j 以及 py2neo 库

安装对应版本的 Java 环境以及 Neo4j 工具, 如图为后台启动 Neo4j 服务:



```
C:\Windows\system32\cmd.exe
Directories in use:
home:      D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8
config:    D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\conf
logs:      D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\logs
plugins:   D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\plug
ins:
import:   D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\imp
rt
data:     D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\data
certificates: D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\cert
ificates
licenses: D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\lice
nses
run:      D:\University\Course\Forth\创新创业\crawler\neo4j\neo4j-community-4.4.8-windows\neo4j-community-4.4.8\run
Starting Neo4j...
2023-07-07 06:37:25.411+0000 INFO Starting...
2023-07-07 06:37:28.025+0000 INFO This instance is ServerId{de4fbf60} (de4fbf60-c44d-4bd6-9e92-08cb4de11ae7)
2023-07-07 06:37:29.806+0000 INFO ===== Neo4j 4.4.8 =====
2023-07-07 06:37:31.290+0000 INFO Performing postinitialization step for component 'security-users' with version 3 and
status CURRENT
2023-07-07 06:37:31.291+0000 INFO Updating the initial password in component 'security-users'
2023-07-07 06:37:33.150+0000 INFO Bolt enabled on 127.0.0.1:7687.
2023-07-07 06:37:34.120+0000 INFO Remote interface available at http://localhost:7474/
2023-07-07 06:37:34.124+0000 INFO id: 4B9FC6329413844CB381A68F24CEB817EC833923370DC1DED4C089E3D0FCA5CF
2023-07-07 06:37:34.124+0000 INFO name: system
2023-07-07 06:37:34.125+0000 INFO creationDate: 2023-07-04T07:02:36.453Z
2023-07-07 06:37:34.125+0000 INFO Started.
```

如图为前端部分:



以下为使用 py2neo 库连接 neo4j 的部分:

```
1 from py2neo import Graph, Node, Relationship, NodeMatcher
2 import pandas as pd
3 graph = Graph('http://localhost:7474/db/data/', auth=("neo4j", "123"))
```

### Step 2: 进行股票基本信息的实体建立

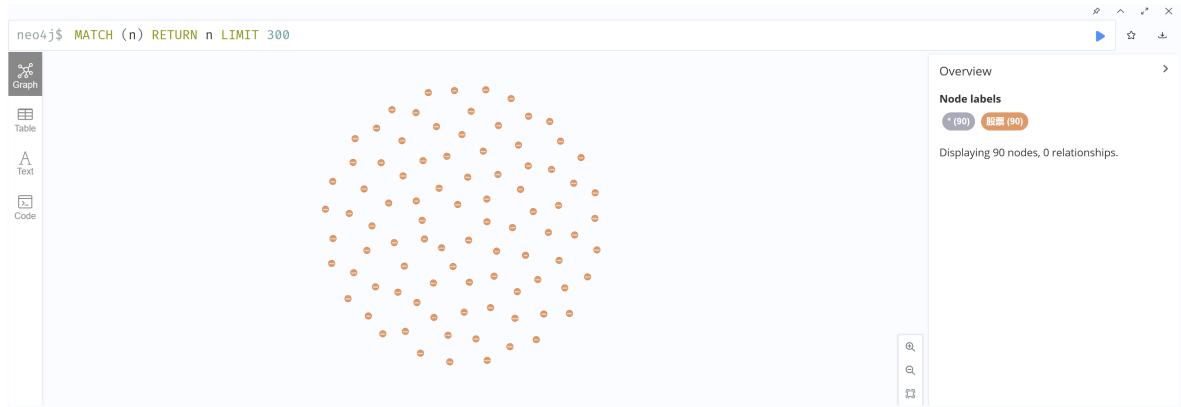
首先我们需要从爬虫部分的 csv 文件中读取数据, 然后进行数据的预处理部分, 最后对于每一个股票都建立一个相应的实体, 并在实体中存储相应的属性, 如**TS代码**, **股票名称**, **公司名称**, **公司成立时间**, **行业**, **公司介绍**, **公司联系电话**等。对应上面的步骤, 我们可以写出以下代码:

```

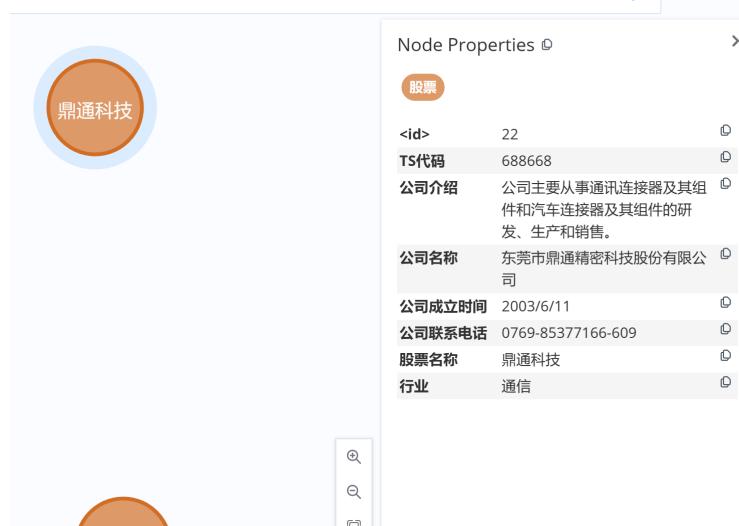
1 stock = pd.read_csv('1company_data_table.csv', encoding='gbk')
2 # print(stock)
3 # 空值填充为未知
4 stock['行业'] = stock['行业'].fillna('未知')
5 # 去重
6 stock = stock.drop_duplicates(subset=None, keep='first', inplace=False)
7 # 创建股票信息实体
8 for i in stock.values:
9     a = Node('股票', TS代码=i[0], 股票名称=i[1], 公司名称=i[2], 公司成立时间=i[5],
10             行业=i[6], 公司介绍=i[7], 公司联系电话=i[-1])
11 graph.create(a)

```

如图，建立完成后，我们进入前端查看可视化展示，以下为整体展示：



以下为某一个实体的属性的展示：



### Step 3：创建股东实体，并建立股东与股票的联系

我们定义**股东实体**，这个实体有“**股东名称**”“**股东性质**”两个属性。

同上一步骤类似，我们可以写出以下的代码：

```

1 # 创建股东实体
2 holder = pd.read_csv('1sharehold_data_table.csv', encoding='gbk')
3 # print(holder)
4 holder_iden = holder[['股东名称', '股东性质']].drop_duplicates(keep='first', inplace=False)
5 for i in holder_iden.values:
6     a = Node('股东', 股东名称=i[0], 股东性质=i[1])
7     graph.create(a)

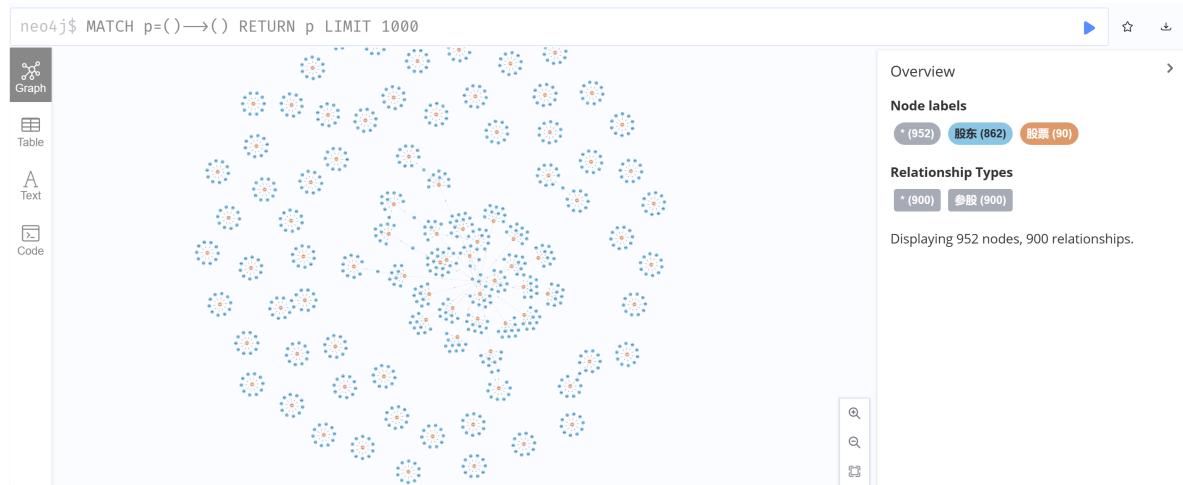
```

接下来，我们定义**参股关系**，这是一个**多对多关系**，即一个股东可以投资多个股票，一个股票可以被多个股东投资。同时我们定义参股关系中有**股份类型、持股数、占总股本比例**这些关系，我们可以写出以下代码：

```
1 # 创建参股关系
2 matcher = NodeMatcher(graph)
3 for i in holder.values:
4     a = matcher.match("股票", TS代码=i[0]).first()
5     b = matcher.match("股东", 股东名称=i[1]).first()
6     r = Relationship(b, '参股', a, 股份类型=i[3], 持股数=i[4], 占总股本比例=i[5])
7     graph.create(r)
```

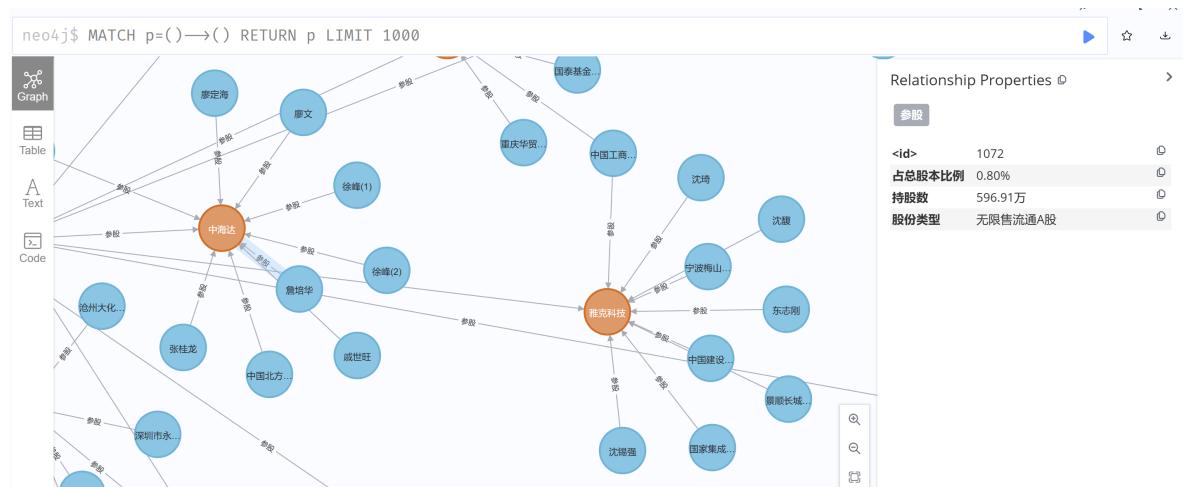
**tips:** 这里和PPT中的代码不太一样，因为我爬取的 csv 表格中是一个**多对多关系表**，每一行的一个元组就代表一个**关系**，所以不需要进行循环遍历 b 中的每一个元素。

运行结果可视化展示如下：



可以发现，中间有一些股东同时投资了多个股票，中间形成了一张较大的连通图。

放大查看关系模型如下，右侧为某个参股关系中的属性：



## Step 4: 基金实体与持股关系的建立

与上述步骤类似，我们定义基金实体拥有基金名称和基金代码两个属性，持股关系是多个多对多关系，关系中有属性持股数和占股本总比例。

同样的我们可以写出以下代码：

```
1 fund = pd.read_csv('1fund_data_table.csv', encoding='gbk')
2 fund_iden = fund[['基金名称', '基金代码']].drop_duplicates(keep='first', inplace=False)
3 for i in fund_iden.values:
4     a = Node('基金', 基金名称=i[0], 基金代码=i[1])
5     graph.create(a)
6 matcher = NodeMatcher(graph)
7 for i in fund.values:
8     a = matcher.match('股票', TS代码=i[0]).first()
9     b = matcher.match('基金', 基金名称=i[2]).first()
10    r = Relationship(b, '持股', a, 持股数=i[3], 占总股本比例=i[4])
11    graph.create(r)
```

由于加入的实体与关系过多，图形较为杂乱，这里我们只展示持股关系部分的展示，不进行参股部分的展示。

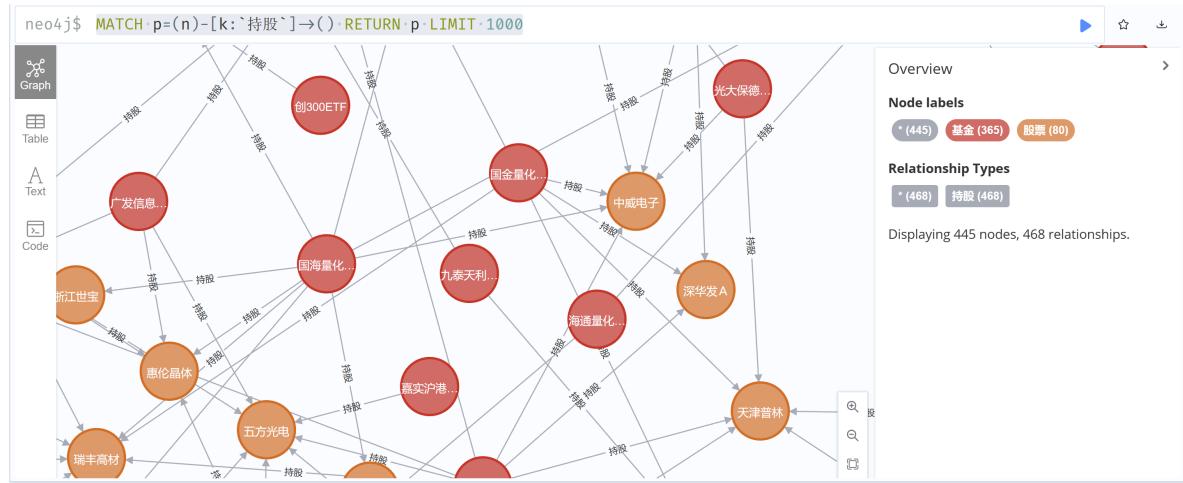
使用以下查询命令：

```
MATCH p=(n)-[k:`持股`]->() RETURN p LIMIT 1000
```

以下为运行上述代码与查询命令执行完成后的展示：



同样可以看到中间很大一部分被基金公司连成了连通图，以下为详细属性展示：



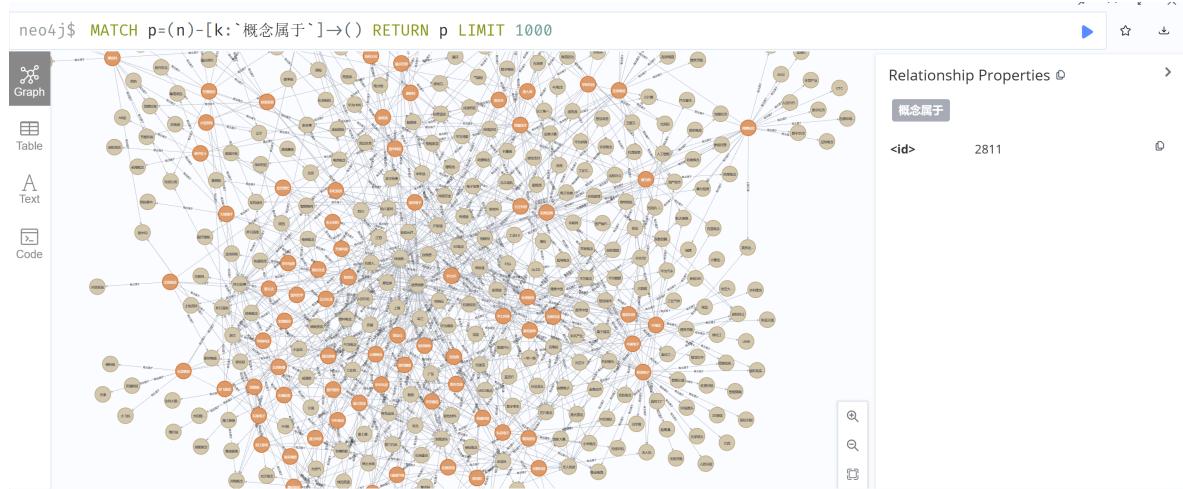
## Step 5：概念实体与概念属于关系的建立

与上面的过程类似，我们建立概念实体，实体只有一个属性**概念名称**。建立**概念属于**关系，关系中不存在属性。

同样可以写出以下代码：

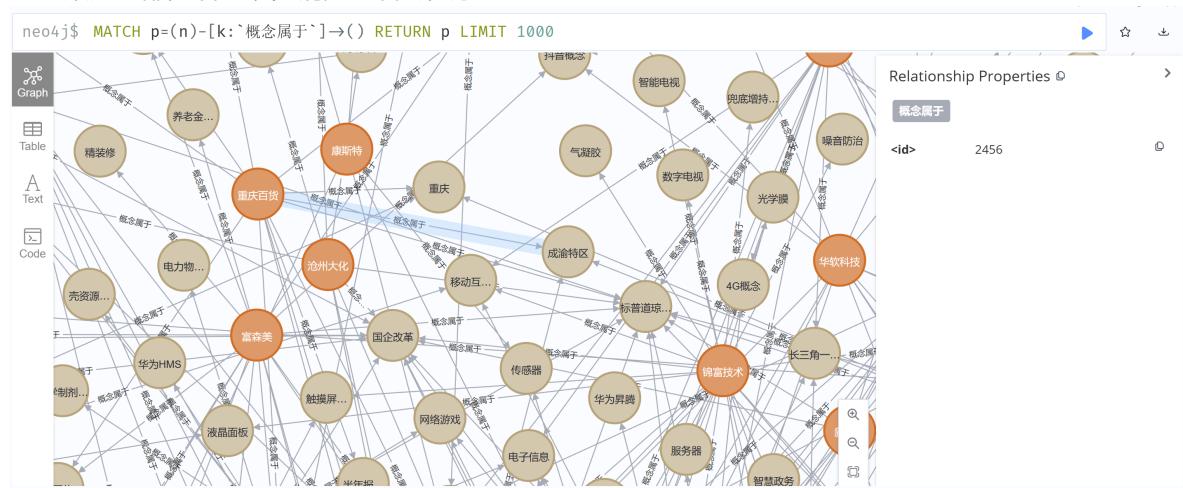
```
1 concept = pd.read_csv('concept_data_table.csv', encoding='gbk')
2
3 concept_iden = concept[['概念']].drop_duplicates(keep='first', inplace=False)
4 for i in concept_iden.values:
5     a = Node('概念', 概念名称=i[0])
6     graph.create(a)
7
8 matcher = NodeMatcher(graph)
9 for i in concept.values:
10    a = matcher.match('股票', TS代码=i[0]).first()
11    b = matcher.match('概念', 概念名称=i[1]).first()
12    r = Relationship(a, '概念属于', b)
13    graph.create(r)
```

最后可以产生这样的可视化结果：



由于一个股票对应的概念可能很多，且概念可以重合，所以上图中的关系错综复杂，形成了十分庞大的网络。

放大一部分后可以更清楚的看到关系网。



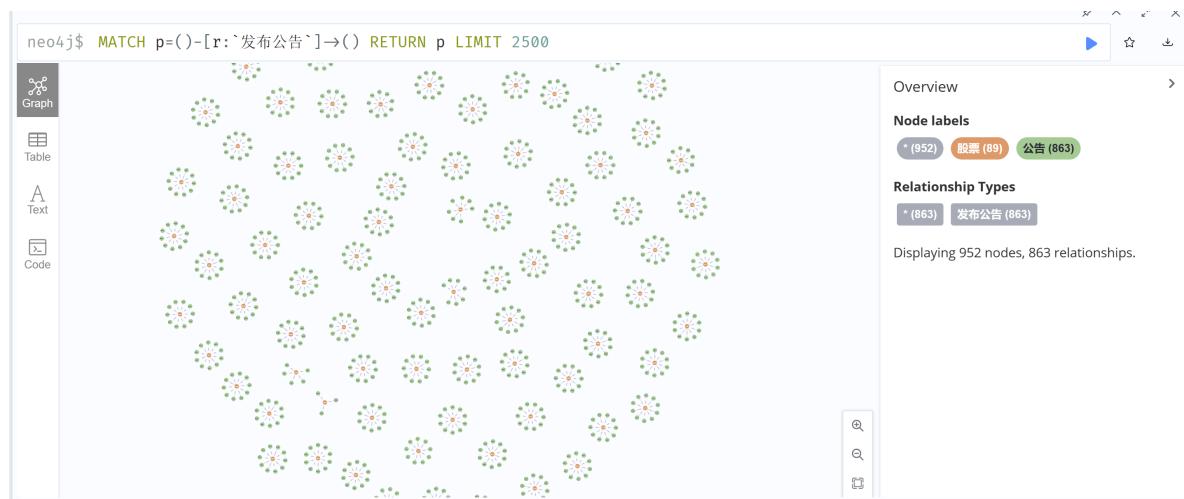
## Step 6: 公告实体和发布公告关系的建立

同样和以上步骤都类似。公告实体包含了**公告名称**属性，值得注意的是，由于公告理论上可以作为一个**弱实体**，因此公告发布时间既可以在关系中，也可以在公告实体中，这里我们将其作为属性**放在关系中**，这里的关系是一个**一对多关系**。

可以写出以下代码：

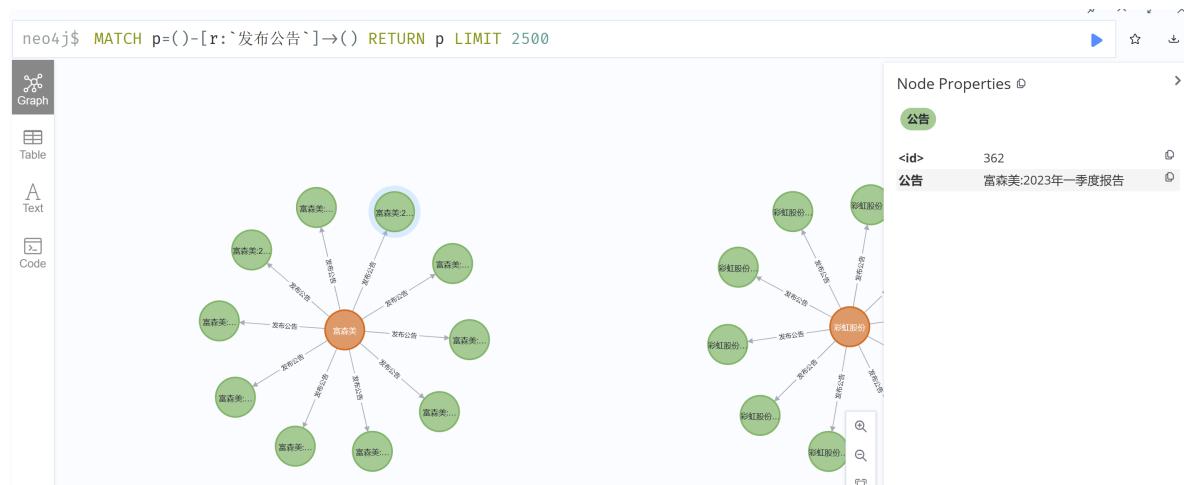
```
1 news = pd.read_csv('news_data_table.csv', encoding='gbk')
2
3 news_iden = news[['公告']].drop_duplicates(keep='first', inplace=False)
4 for i in news_iden.values:
5     a = Node('公告', 公告=i[0])
6     graph.create(a)
7
8 matcher = NodeMatcher(graph)
9 for i in news.values:
10    a = matcher.match('股票', TS代码=i[0]).first()
11    b = matcher.match('公告', 公告=i[2]).first()
12    r = Relationship(a, '发布公告', b, 时间=i[1])
13    graph.create(r)
```

运行结果如下：

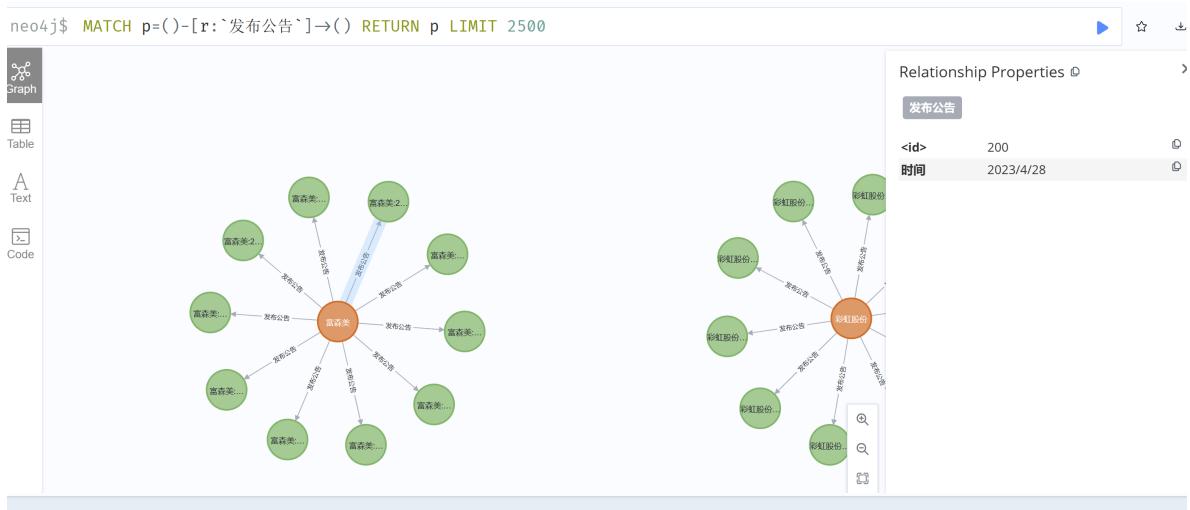


由于不同公司之间不可能发布同一个公告，因此应该是不相邻的多个“菊花图”产生的森林，符合预期。

放大后如下图：



其中时间属性在关系中，如下图：



到此为止，我们完成了所有的实体与关系的构建，接下来我们进行10个股票相关信息的查询。

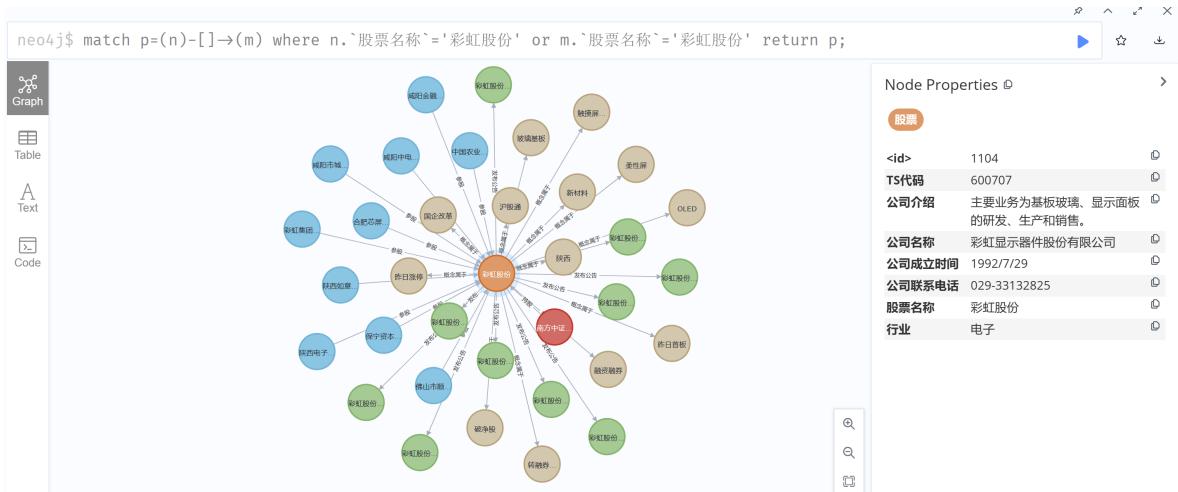
### (3) 股票相关信息查询

#### (a) 彩虹股份

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='彩虹股份' or m.`股票名称`='彩虹股份' return p;
```

查询结果如下：

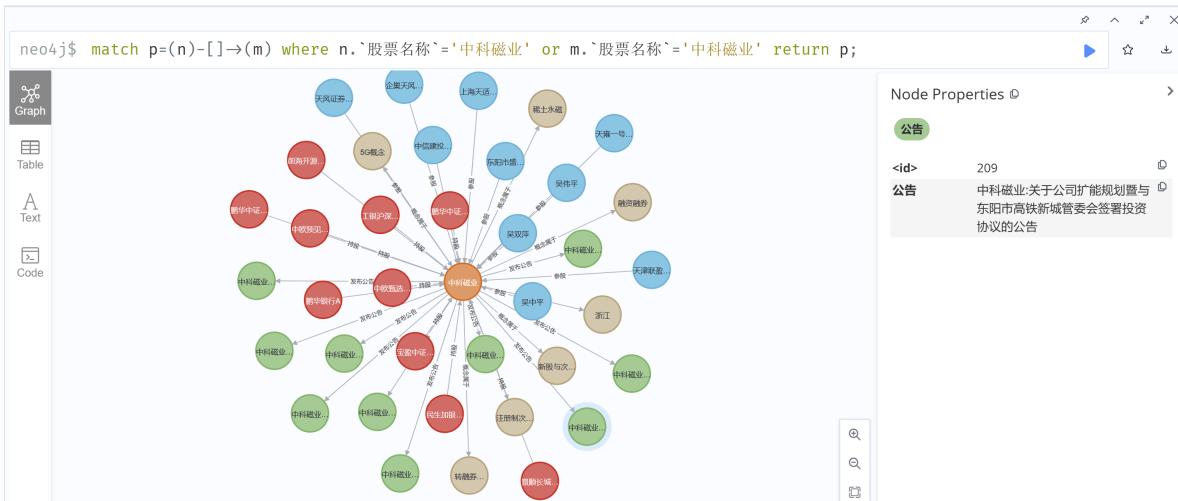


(b) 中科磁业

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='中科磁业' or m.`股票名称`='中科磁业' return p;
```

查询结果如下：

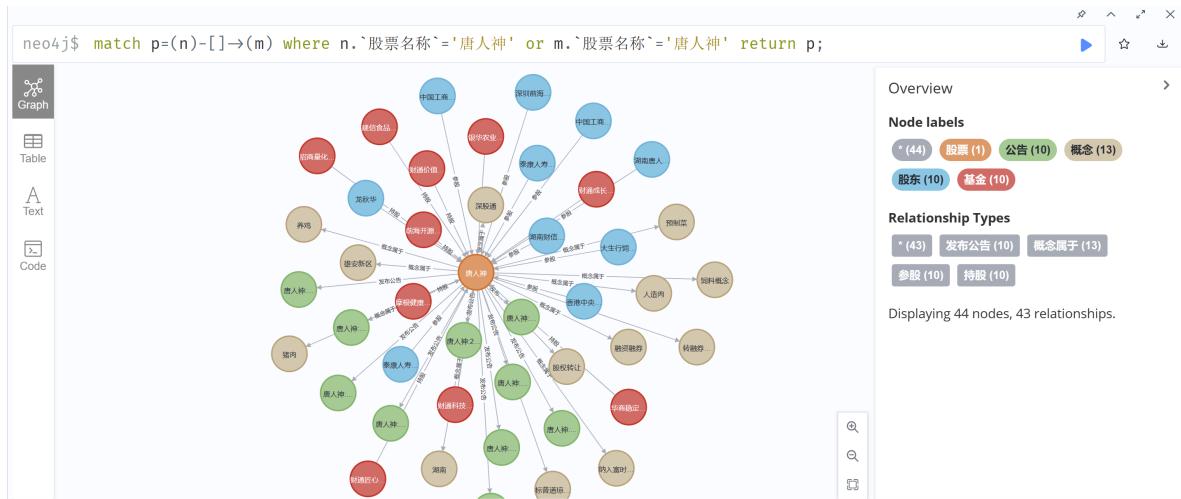


(c) 唐人神

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='唐人神' or m.`股票名称`='唐人神' return p;
```

查询结果如下：

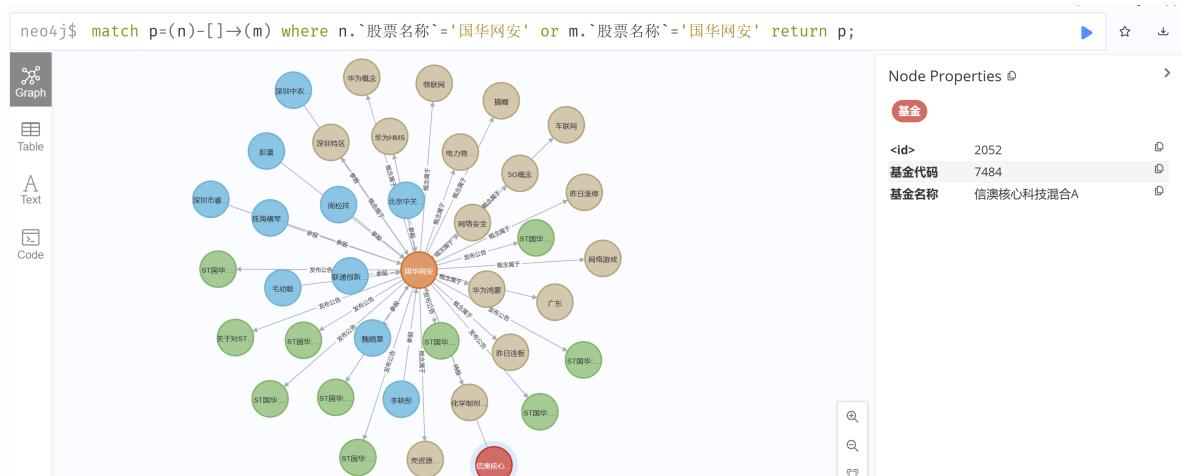


(d) 国华网安

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='国华网安' or m.`股票名称`='国华网安' return p;
```

查询结果如下：

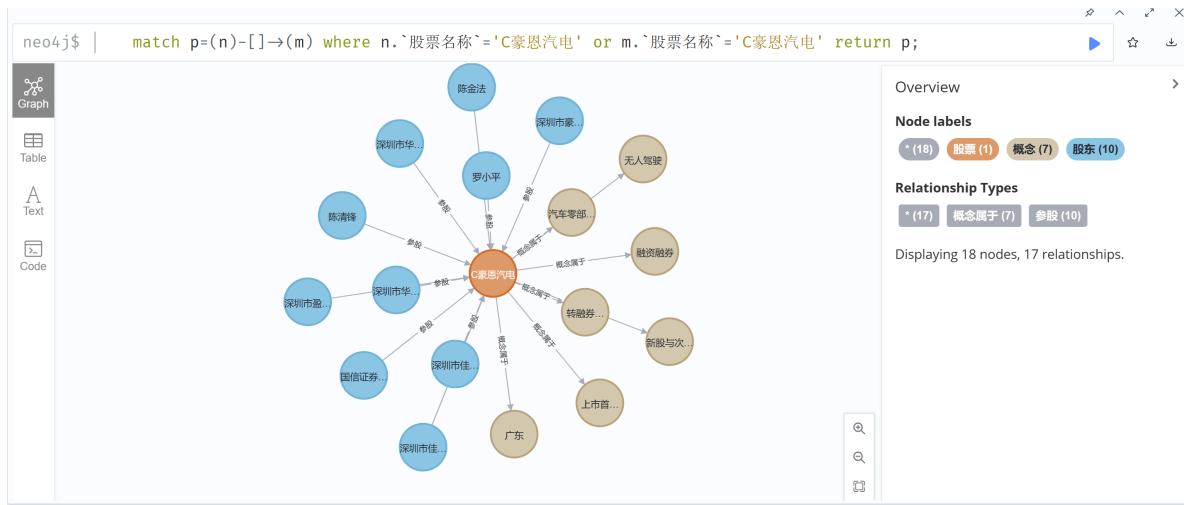


(e) C豪恩汽电

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='C豪恩汽电' or m.`股票名称`='C豪恩汽电' return p;
```

查询结果如下：



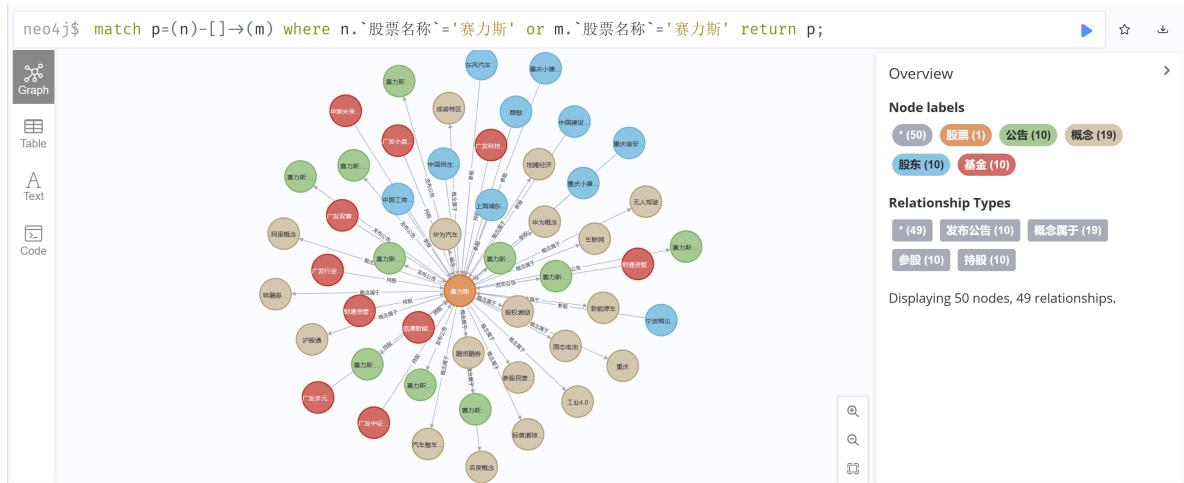
由于是新发股，相关信息较少。

(f) 赛力斯

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='赛力斯' or m.`股票名称`='赛力斯' return p;
```

查询结果如下：



(g) 银河磁体

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='银河磁体' or m.`股票名称`='银河磁体' return p;
```

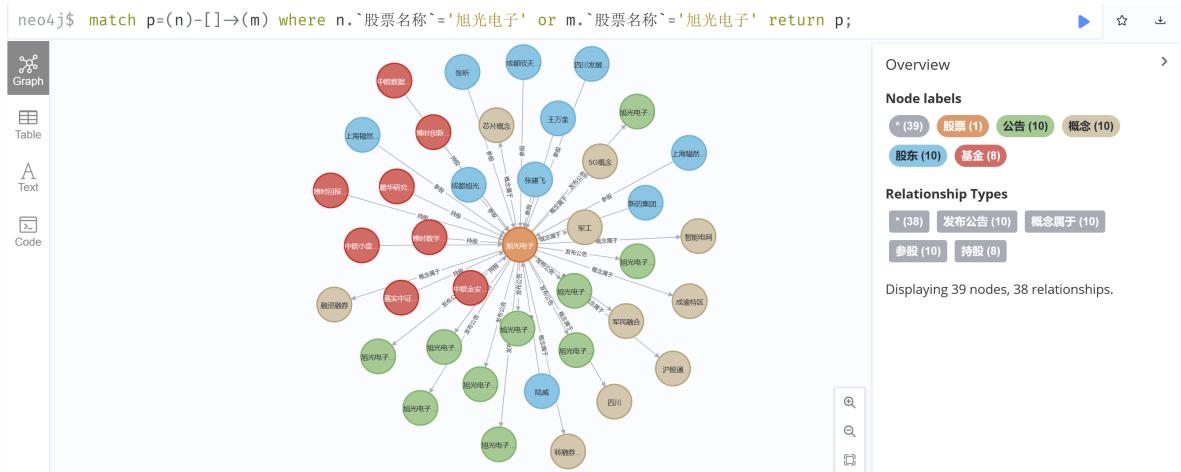
查询结果如下：

(h) 旭光电子

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='旭光电子' or m.`股票名称`='旭光电子' return p;
```

查询结果如下：

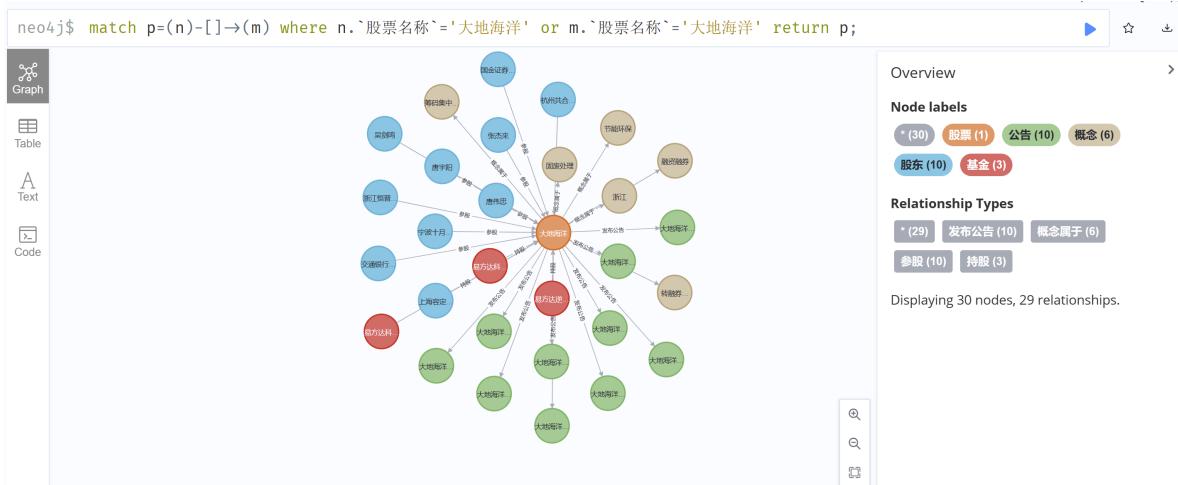


### (i) 大地海洋

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='大地海洋' or m.`股票名称`='大地海洋' return p;
```

查询结果如下：

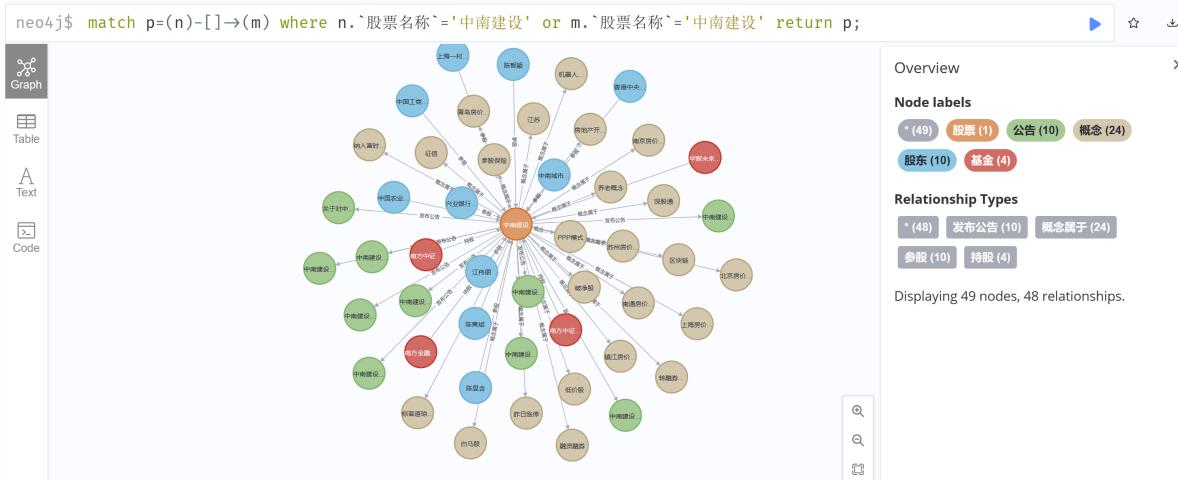


(j) 中南建设

查询语句如下：

```
match p=(n)-[]->(m) where n.`股票名称`='中南建设' or m.`股票名称`='中南建设' return p;
```

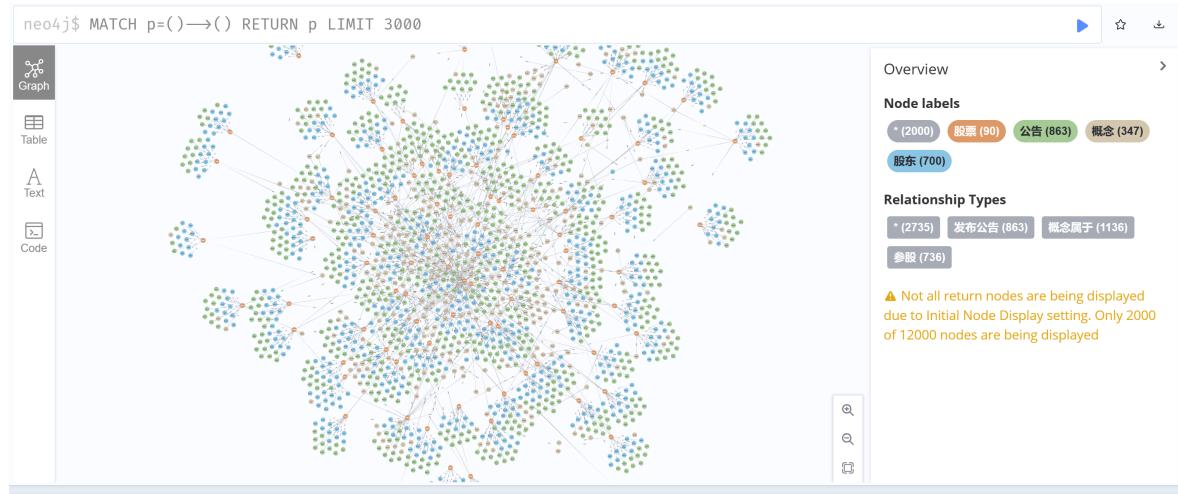
查询结果如下：



以上一共十个股票的查询结果。

## 五、实验数据以及结果分析

大部分实验结果在上图中已经展示，下面展示所有节点的图谱，由于节点过多导致系统卡顿，这里展示2000个节点：



可以发现，在仅仅爬取了以上一些信息时，形成的知识图谱关系已经十分复杂。

## 六、实验结论

在本实验中，我们的最终做到了是爬取九方智投中的沪深京A股数据中的**股票概念**、**股东**、**基金**、**公告**等数据，并形成**参股**、**持股**、**发布公共**、**概念属于**等图形关系，达成了为后来数据分析师能够更加便捷的对这些数据进行**量化统计分析**的目标。

同时在实验中有以下收获：

- 1、此实验让我学会利用 *Python* 的 *Selenium* 框架爬取数据，定位 *Xpath*、正则选择器、简单爬取数据的操作。
- 2、进一步加深对知识图谱的Neo4j图数据库的基本语法操作的运用。