**利用消息面关联股票简单版本**

**1、基本思路**

给定一则消息，找出其关联的股票。首先，利用分词工具对该消息分词，提取出关键词汇；这里需要自定义的词典。其次，将关键词汇扔进构建好的知识图谱找到关联股票。利用知识图谱找关联股票主要基于两种关联方式：结构关联与语义关联。

**1.1、知识图谱**

知识图谱组织成图的形式，节点是概念或者实体，边是实体之间的关系。

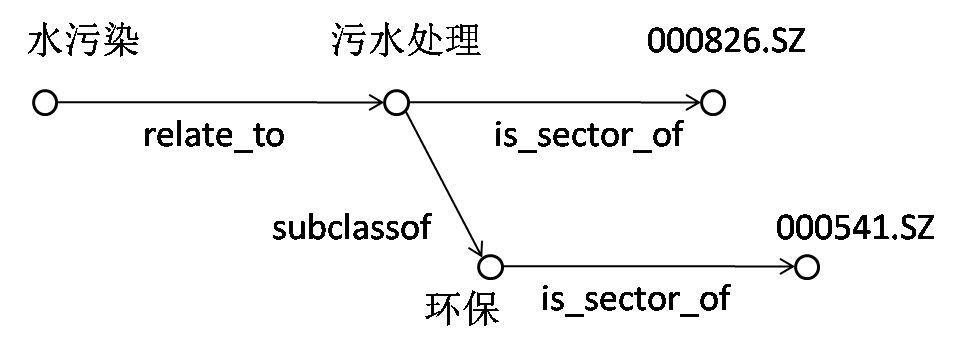


图1

如上图，是一个简单的知识图谱。节点代表概念或者实体，这里的实体有“水污染”，“污水处理”，“环保”，“000826.SZ”和“000541.SZ”。这些实体之间的关系由边来表示，比如“水污染”和“污水处理”之间具有关联性，这里采用标记为“relate\_to”的边；“污水处理”是“000826.SZ”的一个行业板块名称，这里用“is\_sector\_of”沟通“污水处理”和“000826.SZ”。文本表示上，图1的知识图谱可以表示成下面的主谓宾三元组形式：

表1

|  |  |  |
| --- | --- | --- |
| subject | predicate | object |
| 水污染 | relate\_to | 污水处理 |
| 污水处理 | is\_sector\_of | 000826.SZ |
| 污水处理 | subclassof | 环保 |
| 环保 | is\_sector\_of | 000541.SZ |

第一行subject，predicate与object作为schema方便加载器识别，首字母均小写。第二行开始，便是三元组格式，四条三元组等价地描述了图1的知识图谱。

**1.2、结构关联**

通过分词器与自定义词典对消息提取出了关键词，利用关键词通过知识图谱找到相关股票。因为知识图谱本身是一个有向图，所以遍历图结构，以找到标签是股票或者股票代码的那些节点。这就是结构关联。

如果关键词与关联股票的路径比较近，那么说明该关键词与该股票关联程度比较大。所以，利用结构关联，优先选择那些关联路径比较小的股票。因此，这里采用*广度优先*的算法搜索图结构。比如，消息中出现了“水污染”这样的词汇，通过图1，以及结构关联的算法，可以发现“水污染”链向“000826.SZ”与“000541.SZ”这两只股票，分别通过“污水处理”和“环保”这两个节点。

**1.3、语义关联**

结构关联的结果是所有有关联路径的股票都被关联了。这显然会带来一些问题。如果某一则消息只针对污水处理方面，那么最好只关联到和污水处理有关的股票。通过图1以及结构关联，000826.SZ这只股票被关联到是合理的；但是000541.SZ这只股票被关联到却是不合理的，因为它和污水处理没有关系。

首先，查看一下000541.SZ这只股票是怎么被关联到的：从“水污染”开始，到达“污水处理”，由于“污水处理”是“环保”的子板块，因此将“环保”也关联进去了。然而，环保是一个更广范围的概念。在实际处理消息中，我们希望概念范围越小越好。因此需要做一些特别处理。

这里主要针对“subclassof”这个谓词的语义进行特别的处理。通过subclassof实际上链接到了更广义的节点。实际中，可以这样设置，搜索知识图谱时，只往更狭义的范围去搜索，而不去看更广义的范围。这样一来，这里的“subclassof”这条边就被阻止访问了。因此000541.SZ这只股票就不会被关联到。

上述情况是最简单的一种语义关联。关于语义关联，还可以定义其它具有特殊语义以及处理方式的谓词。针对这些谓词进行预处理（比如添加更多的知识）被称作推理。这些工作都是为了方便最后进行关键词的关联。

**2、基本设计**

**2.1、内部定义词汇**

|  |  |  |
| --- | --- | --- |
| 关键词 | 位置 | 描述 |
| is\_a | predicate | 谓词，描述主语是宾语的成员：  例："600008.SH is\_a SCode",  描述600008.SH是一个证券代码 |
| has | predicate | 是is\_a的互反谓词 |
| scode | predicate | 谓词，描述主语的证券代码是宾语，  例："首创股份 scode 600008.SH "  描述首创股份的证券代码是600008.SH |
| is\_scode\_of | predicate | 是scode互反谓词 |
| sname | predicate | 谓词，描述主语的证券名称是宾语，  例："600008.SH sname 首创股份 "  描述600008.SH的证券名称是首创股份 |
| is\_name\_of | predicate | 是sname的互反谓词 |
| sector | predicate | 谓词，描述主语所属板块是宾语，  例："600008.SH sector 环保 "  描述600008.SH属于环保板块 |
| is\_sector\_of | predicate | 是sector的互反谓词 |
| theme | predicate | 谓词，描述主语所属题材是宾语，  例："600008.SH theme 美丽中国 "  描述600008.SH属于美丽中国这个题材 |
| is\_theme\_of | predicate | 是theme的互反谓词 |
| same\_as | predicate | 对称谓词，描述主语与宾语一样，  例："雄安 same\_as 雄安新区 "  描述雄安与雄安新区代表同一事物。 |
| subclassof | predicate | 谓词，描述术语描述的概念包含于宾语，  例："污水处理 subclassof 环保"  污水处理术语环保 |
| superclassof | predicate | 是subclassof的互反谓词 |
| relate\_to | predicate | 对称谓词，描述主语与宾语具有关联性，  例："污水处理 relate\_to 污水 "  描述污水处理与污水具有关联性 |
| SCode | object | 宾语，证券代码security code的总类，  例："600008.SH is\_a SCode"  描述600008.SH是一个证券代码 |
| SName | object | 宾语，证券名称security name的总类，  例："首创股份 is\_a SName"  描述首创股份是一个证券名称 |
| Sector | object | 宾语，行业，板块总类，  例："环保 is\_a Sector "  描述环保是一个板块或行业 |
| Theme | object | 宾语，题材，  例："美丽中国 is\_a Theme "  描述美丽中国是一个题材 |

**2.2、知识图谱格式**

采用简化的RDF三元组格式，不引入命名空间，如表1的格式，第一行subject，predicate与object作为schema方便加载器识别，首字母均小写。第二行开始，便是三元组格式，其中内部定义的关键词不得作为用户词汇来使用；必须严格按照上面内部词汇的定义来使用。

**3、例子**

现给定一则消息如下，针对下面的消息进行分析和关联：

“*生态环保显著成绩另一面：水污染形势严峻，治水思路举措亟需改善 王禹2017-07-18 19:16“目前全国31个省区市均已启动生态保护红线划定工作。2017年底，京津冀区域，长江经济带沿线等14个省市要完成生态保护红线划定工作。”在近期的“2017中国生态环保大会暨西宁绿色发展论坛”上，国家环保部副部长黄润秋透露。黄润秋表示，生态文明建设和环保工作，目前取得了明显成效。2016年，新环境空气质量标准第一阶段实施监测的74个城市，平均优良天数比例达到了74.2%，比2013年提高了13.7%。11个考核PM2.5的省份，PM2.5浓度均明显下降，地表水达到或好于三绿水质断面的比例，比2010年提高了17.9%，大江大河干流水质明显改善。同时，全国共建立各级自然保护区2740个，其中陆地自然保护区面积约占全国陆地面积的14.85%，高于世界平均水平。推进实施生物多样性保护重大工程，全国森林覆盖率由本世纪初的16.6%上升到了21.63%。以西宁和南京为例。青海省委常委、西宁市委书记王晓透露，西宁的绿色发展正在从末端治理向源头治理拓展，从静态保护向动态保护升级，在从需求要素向供给要素转型。南京市委副书记、市长缪瑞林介绍称，南京市先后出台实施了南京生态红线区域保护规划，生态红线区域总面积达到了1600平方公里，占全市国土面积的24.75%，另外建成绿地覆盖率达到44.75%，城市人均公园绿地面积达到了15.3平方米。此外，黄润秋还透露，中央环境保护督察已经覆盖23个省份，初步罚款约8.8亿元，立案侦查1183件，拘留110.3人，约谈13593人，问责11390人等。他表示，接下来还将重点做好大气、水、土壤三大污染防治行动计划，加强生态保护力度及生态监管等方面工作。不过，成绩之外，生态环保问题依然需予以更多重视和努力，比如水污染及其治理方面。环保部原总工程师、国家环境咨询委员会委员万本太认为，中国水污染较重，形势严峻。 比如地表水。万本太透露，全国地表水监测的断面1940个，满足3类水的断面比例为68%，劣5类水的比例为9%。7大水系只有珠江水系和长江水系总体上还好，珠江水系监测断面165个，能满足3类水的比例大约是90%，劣5类水的比例是4%。长江水系监测断面510个，满足3类水的比例是82%，劣5类水的比例是4%。地下水方面，国土资源部在31个省225个城市里面采了6214个点，得出的结论是较差和极差比例占60%。水利部的调查情况更可怕，他们在部分省里采了2104个点进行检测分析，最后得出的结论是较差和极差的比例占76%。万本太认为，水污染严重的原因包括污染物排放量大、治理不彻底、监管不到位。在他看来，治水也存在一些基本问题，主要有三点：一是现在是治理污水没“踩”到点上，没有下决心老老实实治工业企业的排放水，没有100%达标，没有100%治理等；二是河流综合整治，黑臭水体治理不尊重水生态规律，不从水环境的改善制度、水生态的健康管理角度去治理，光知道挖管子、搞清淤；三是农村污水治理，用湿地处理的时候效果不明显，管理不善没有达到当初设计的排放目的。于此，对“十三五”的治理污水，万本太建议，应遵循山水统一生命体的概念，与恢复流域水生态系统健康和功能完整性的目标，以生物治理，生态恢复技术为主流，以点源治理为重点，集中和分散治理相结合，点源和面源统筹防治，尽力减轻流域水环境的压力，努力使水环境质量得到显著的改善。其总方针是“节水优先，治理着力，富有为纲”。同时，根据接纳水体的功能来制订不同的标准，凡是直接向地表水排放的标准要加严。此外，要全面落实河长制，坚决执行一控双达标的制度，所有点源排放达标，所有地表水的环境质量达标，并执行流域上下游的水的界面水质的生态补偿制度等。*”

**3.1、构建知识图谱**

按照2中的方式构建环保方面股票的知识图谱，如下面形式（截取部分）：

|  |  |  |
| --- | --- | --- |
| subject | predicate | object |
| 600008.SH | is\_a | SCode |
| 600874.SH | is\_a | SCode |
| 000826.SZ | is\_a | SCode |
| 000544.SZ | is\_a | SCode |
| 600168.SH | is\_a | SCode |
| 600064.SH | is\_a | SCode |
| 600478.SH | sector | 节能建材 |
| 000541.SZ | sector | 节能建材 |
| 600481.SH | sector | 节能建材 |
| 600874.SH | sector | 污水处理 |
| 600388.SH | sector | 废气处理 |
| 600526.SH | sector | 废气处理 |
| 环保 | same\_as | 环境保护 |
| 环保 | same\_as | 环保工程 |
| 环保 | relate\_to | 环境 |
| 污水处理 | relate\_to | 污水 |
| 节能建材 | relate\_to | 节能建材 |
| 节能建材 | relate\_to | 建材 |
| 节能建材 | relate\_to | 建筑材料 |
| 节能建材 | relate\_to | 绿色材料 |
| 废气处理 | relate\_to | 废气 |
| 垃圾处理 | relate\_to | 垃圾 |
| 污水 | same\_as | 水污染 |
| 污水 | subclassof | 环保 |
| 污水 | subclassof | 环境 |

**3.2、自定义词典**

有了知识图谱，不需要专门构建词典。默认将知识图谱中出现的词汇集合作为自定义词典输进分词器。

**3.3、分词**

针对上述消息进行分词，得到前50个关键词，关键词的提取个数可以设置：

|  |
| --- |
| '生态', '治理', '类水', '比例', '万本', '地表水', '断面', '水污染', '点源', '水系', '达标', '黄润', '2017', '红线', '西宁', '保护', '监测', '达到', '排放', '环保', '污水', '水质', '个点', 'PM2', '31', '100', '透露', '全国', '采了', '环境质量', '形势严峻', '环保部', '环境', '极差', '流域', '治水', '长江', '满足', '南京', '改善', '划定', '水体', '珠江', '工作', '标准', '省份', '覆盖率', '较差', '制度', '实施' |

利用知识图谱筛选关键词，即保留出现在知识图谱中的关键词，得到如下结果：

|  |
| --- |
| '环保', '污水'，'水污染'，'环境' |

**3.4、结构关联**

通过以上关键词基于结构关联的方法到知识图谱中找到关联股票（截取部分）：

|  |
| --- |
| **相关股票:000939.SZ**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600874.SH'), ('600874.SH', 'sector', '环保'), ('环保', 'is\_sector\_of', '000939.SZ')]  关联路径1:[('环保', 'in', 'News'), ('环保', 'is\_sector\_of', '000939.SZ')]  关联路径2:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600874.SH'), ('600874.SH', 'sector', '环保'), ('环保', 'is\_sector\_of', '000939.SZ')]  关联路径3:[('环境', 'in', 'News'), ('环境', 'relate\_to', '环保'), ('环保', 'is\_sector\_of', '000939.SZ')]  **相关股票:000541.SZ**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600874.SH'), ('600874.SH', 'sector', '环保'), ('环保', 'is\_sector\_of', '000541.SZ')]  关联路径1:[('环保', 'in', 'News'), ('环保', 'is\_sector\_of', '000541.SZ')]  关联路径2:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600874.SH'), ('600874.SH', 'sector', '环保'), ('环保', 'is\_sector\_of', '000541.SZ')]  关联路径3:[('环境', 'in', 'News'), ('环境', 'relate\_to', '环保'), ('环保', 'is\_sector\_of', '000541.SZ')]  **相关股票:600008.SH**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600008.SH')]  关联路径1:[('环保', 'in', 'News'), ('环保', 'is\_sector\_of', '600008.SH')]  关联路径2:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600008.SH')]  关联路径3:[('环境', 'in', 'News'), ('环境', 'relate\_to', '环保'), ('环保', 'is\_sector\_of', '600008.SH')] |

上述返回结果中，不仅返回关联的股票，也同时将如何关联到它的路径给了出来。可以发现，关联路径可能不止一条。

**3.5、语义关联**

上述基于结构关联的方法中，股票000541.SZ也被关联了进来，观察其四条关联路径，最后都是从“环保”关联过去的。在此，利用语义关联做进一步处理：具体而言，在寻找关联性时，不考虑“subclassof”的关联。

同时，股票000541.SZ的关联路径2的长度为5，如果认为长度太长关联性便越差，因此可以限制关联路径的长度在4以及4以下。

根据以上设置，得到下属结果，这些结果都是和污水处理相关的：

|  |
| --- |
| **相关股票:000826.SZ**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '000826.SZ')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '000826.SZ')]  **相关股票:000753.SZ**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '000753.SZ')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '000753.SZ')]  **相关股票:600168.SH**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600168.SH')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600168.SH')]  **相关股票:600649.SH**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600649.SH')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600649.SH')]  **相关股票:600874.SH**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600874.SH')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600874.SH')]  **相关股票:000544.SZ**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '000544.SZ')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '000544.SZ')]  **相关股票:600064.SH**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600064.SH')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600064.SH')]  **相关股票:600008.SH**  关联路径0:[('水污染', 'in', 'News'), ('水污染', 'same\_as', '污水'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600008.SH')]  关联路径1:[('污水', 'in', 'News'), ('污水', 'relate\_to', '污水处理'), ('污水处理', 'is\_sector\_of', '600008.SH')] |