**哈尔滨工业大学**

**国家示范性软件学院**

**本科毕业设计(论文)中期报告**

**题 目： 智能资讯平台上市公司舆情分析子系统的设计与实现**

**专 业 软件工程**

**学 生 姓 名 罗峻**

**学 号 1123710521**

**联 系 方 式 13392426025**

**年 级 2012级**

**实 习 基 地 深圳证券交易所**

**基地指导教师 许保勋**

**联 系 方 式 15805054250**

**校内指导教师 黄虎杰**

**联 系 方 式 hjhuang@hit.edu.cn**

**中 检 日 期 2016.03.28**

**哈尔滨工业大学软件学院**

目 录

[1. 毕业设计（论文）内容概述 1](#_Toc438538735)

[1.1项目来源及开发目的和意义 1](#_Toc438538736)

[1.2主要开发任务 2](#_Toc438538737)

[1.3本人所承担任务（模块）说明 2](#_Toc438538738)

[1.4开发环境和开发工具 3](#_Toc438538739)

[1.5项目原定进度安排 3](#_Toc438538740)

[2. 中期完成情况说明 6](#_Toc438538741)

[2.1预定计划的执行情况 6](#_Toc438538742)

[2.2中期工作说明及成果汇报 7](#_Toc438538743)

[2.2.1系统总体结构设计 7](#_Toc438538744)

[2.2.2系统数据库设计 9](#_Toc438538745)

[2.2.3系统详细设计 10](#_Toc438538746)

[2.2.3.1文件处理详细设计 10](#_Toc438538747)

[2.2.3.2公告处理详细设计 11](#_Toc438538748)

[2.2.3.3分类器详细设计 11](#_Toc438538749)

[2.2.3.4分类结果处理详细设计 17](#_Toc438538750)

[2.2.3.5工作日志详细设计 18](#_Toc438538751)

[2.2.4系统实现 18](#_Toc438538752)

[2.2.4.1文件处理模块实现 18](#_Toc438538753)

[2.2.4.2公告处理模块实现 19](#_Toc438538754)

[2.2.4.3分类器处理模块实现 21](#_Toc438538755)

[2.2.2.4分类结果处理模块 26](#_Toc438538756)

[2.2.2.5工作日志模块实现 26](#_Toc438538757)

[2.2.5系统实现效果 27](#_Toc438538758)

[2.3存在的困难与问题 29](#_Toc438538759)

[2.4如期完成预定任务的可能性分析 29](#_Toc438538760)

[2.5后期工作安排（或进度和计划调整） 29](#_Toc438538761)

# 1. 毕业设计（论文）内容概述

## 1.1项目来源及开发目的和意义

1. 来源

本项目来源于深圳证券交易所的最新业务需求。深交所提供证券交易的场所和设施，组织并监督证券交易。“智能资讯系统”是面向深交所内部的内网信息系统，通过信息管理部提供的网络资讯资源，形成的一个具有大量证券市场资讯的综合服务平台，为深交所各部门提供信息支持。“新华在线”系统是过去深交所的信息平台，多年来向深交所提供真实可靠的资讯。但是由于系统运行时间已经较长，信息来源、数据处理已经较为落后，而且各个部门对信息平台有更高的要求和更广泛的需求。出于以上原因，信息管理部开始开发“智能资讯系统”。“上市公司舆情分析系统”是“智能咨询系统”的一个重要的子系统。

1. 开发目的和意义

随着证券市场的快速发展，市场监管的需求也在逐渐复杂，多元化。与其他传统市场相比，证券市场有更大的不确定性，面临更大的风险，一旦失控，后果极其严重。

深圳证券交易所信息管理部通过调研，发现市场监察部和公司管理部，有迫切的获取上市公司舆情的需求。舆情是指民众对于个人、组织及其政治、道德等取向的社会态度。研究表明，舆情可以直接影响到上市公司的股价，如果舆情事件比较严重，就会影响到市场的稳定性。

现今有一些互联网公司，已经可以实现对社会舆情进行的研究和讨论，但是针对证券市场舆情的研究和应用较少。深交所作为重要的证券交易场所，需要在全方面有效监管市场，证券市场舆情分析在现今信息增长迅猛的互联网时代显得尤为重要。

上市公司舆情分析子系统开发的目的有以下几点：

（1） 数据采集

采集传统媒体站点，门户网站，博客，论坛，社区，股吧等特别是证券行业相关网站的大量数据，为舆情分析和其他文本挖掘工作做数据储备。

（2） 数据存储

通过数据库系统对原始数据，处理后数据进行存储，并采用索引技术提高查询效率。

（3） 数据挖掘与分析

分析互联网信息中和上市公司相关的舆情，通过结合舆情关注度，媒体关注度，网民关注度综合计算，获取针对上市公司的媒体关注度、网民关注度、公司负面舆情预警、以及板块，行业相关的舆情分析、溯源等。

（4） 数据可视化

将舆情分析结果，通过可视化技术，生成直观的报告，提供给市场监察、公司监管部门，通过对上市公司舆情和高热度资讯的跟踪，达到对市场信息的提前预知，实时监管，及时反应等目的。

## 1.2主要开发任务

本系统是深圳证券交易所的实际开发项目，是“智能资讯平台”的一个子系统。如图3-1，本系统可以对上市公司舆情信息进行数据采集，数据存储，数据挖掘与分析，以及数据可视化。其中，系统能够实时进行互联网数据采集，采集包括传统媒体站点，门户网站，博客，论坛，社区，股吧等。数据收集之后，将数据进行处理、存储。再经过数据挖掘与分析，挖掘出与上市公司相关的舆情状况，进行正面、负面、中性的情感分析，并分析舆情事件与重大公告披露、研究报告发表等市场信息是否相关。最终以网站的形式将结果进行展示，展示内容包括上市公司负面舆情预警，媒体关注度，网民关注度，地区分析，板块分析，行业分析。同时根据公司管理部门和市场监察部的需求，提供网站的个性化自定义配置服务，方便用户对特点的上市公司，板块或者行业进行监管。该项目通过文本挖掘的方法，大大减少了人工监察的工作，提供真实可靠的舆情统计情况，能为证券市场监察、监管提供便利。

## 1.3本人所承担任务（模块）说明

本人在整个项目中的具体任务包括数据采集模块，数据存储模块，数据挖掘与分析模块，数据可视化模块4个模块的设计开发以及测试。

（1） 数据采集模块

采集传统媒体站点，门户网站，博客，论坛，社区，股吧等特别是证券行业相关网站的大量数据，为舆情分析和其他文本挖掘工作做数据储备。使用java的HttpClient作为客户端进行开发，设计并实现高性能的并行算法对多个非结构化目标进行爬取和解析，处理出有用信息。

（2） 数据存储模块

通过数据库系统对原始数据，处理后数据进行存储，并采用索引技术提高查询效率，使用的索引为B-TREE索引。

（3） 数据挖掘与分析模块

本系统是针对上市公司的舆情分析系统，因此，首先需要筛选出与上市公司相关的信息。这就涉及到文本分类技术。使用机器学习技术中的决策树模型和随机森林模型，通过训练数据的训练，得到所需模型。重点需求在于，如何提高模型的准确性，这涉及到对决策树模型和随机森林模型算法上的适应性优化。

分类结束之后，得到的结果需要通过文本聚类技术来获取舆情热点。常见的聚类方法有K-means聚类算法，层次聚类算法，SOM聚类算法（神经网络），FCM聚类算法等。聚类算法通过权值和距离来进行无监督学习，最后得出聚类结果。

同时，本系统使用LDA(Latent Dirichlet Allocation)基于概率模型的主题模型算法对文本集进行整理归纳。聚类算法只是通过距离（文章中词语结构相似度）进行聚类归纳，而LDA可以挖掘词语的隐藏的含义，比如“ISIS”和“恐怖分子”是同义词。由于中文的同义词较多，因此语义分析中使用LDA和聚类结果结合，是非常有意义的。

由于监控重点在于负面舆情，因此情感分析是一项重要任务。结合证券相关知识库，才能准确判断舆情情感。情感词库的获取同样会使用LDA主题模型进行收集，根据以往的股价走势和舆论用词，总结出关键词做完情感词库。再根据情感词库进行情感分析。

关键词提取的结果将会用于展示，也是聚类结果的集中体现。通过分析TF-IDF（词频-逆文档频率）值，可以对关键词进行判断。这同样也可以结合LDA获得的主题词进行提取。

话题追踪是利用深交所现有资源如证券分析师研究报告，上市公司公告等，对舆情事件进行深入分析的一种重要方式。

（4） 数据可视化模块

将舆情分析结果，通过可视化技术，生成直观的报告，提供给市场监察、公司监管部门，通过对上市公司舆情和高热度资讯的跟踪，达到对市场信息的提前预知，实时监管，及时反应等目的。开发时使用SpringMVC架构，前台使用html5,css3,javascript.

## 1.4开发环境和开发工具

1. **开发语言**

主要使用java语言进行开发。

服务器结构与配置使用xml语言开发。

部分算法实现使用Matlab语言。

数据库部分使用SQL语言。

数据可视化与展示使用html5,css3,javascript进行开发。

1. **开发工具**

Java开发：Eclipse, MyEclipse

数据库：MySQL

前端：Brackets

数据处理：Matlab

版本控制：GitHub

服务器：Tomcat 7

1. **项目开发环境**

操作系统：Windows 7 专业版32位SPQ(DirectX11)

处理器： Intel(R) Core(TM) i5-2400 CPU @ 3.10GHz

安装内存：4.00GB

系统类型：32位操作系统

1. **未来系统运行环境**

操作系统：Windows Server 2008 R2 Standard

处理器： Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz (2 处理器)

安装内存：288GB

系统类型：64位操作系统

## 1.5项目原定进度安排

项目进度及毕业设计（论文）工作安排见表1-2。

表1-2 项目进度及毕业设计（论文）工作计划表

|  |  |  |  |
| --- | --- | --- | --- |
| 起始时间 | 完成时间 | 计划工作内容 | 备注 |
| 2015.07.20 | 2015.10.07 | 项目需求与可行性分析 | 已完成 |
| 2015.10.08 | 2015.11.23 | 概要设计 | 已完成 |
| 2015.11.23 | 2016.01.05 | 详细设计 | 已完成 |
| 2016.01.06 | 2016.03.20 | 编码实现 | 正进行 |
| 2016.03.21 | 2016.05.19 | 系统测试 | 未开始 |
| 2016.05.20 | 2016.06.20 | 撰写、修改论文，参加毕业答辩 | 未开始 |

# 2. 中期完成情况说明

## 2.1预定计划的执行情况

项目的需求调研以及可行性分析在10月已经完成。通过与系统用户部门的会议交流以及访谈，确定了系统的功能性需求。同时通过对现有的硬件资源进行评估，为满足初步确定的系统非功能性需求，计划对硬件设备进行更新。

在项目概要设计中，并没有发生变化，项目主要分为5大模块分别为文件处理模块，公告处理模块，分类器模块，分类结果处理模块以及工作日志模块。文件处理模块包括对文件格式处理，文件内容读取以及读取失败时的处理；

公告处理模块主要是进行标题正文的切分，以及日期格式及其后指定字数字符的提取；分类器模块主要由各个分类器参数设置，分类效果计算，以及相关业务类型的分类器组成；分类结果处理模块，主要是将分类结果导出到文件中，方便其他系统的使用；工作日志的模块主要是对分类器分类时所使用的参数进行记录，以便后期维护以及为系统进一步开发提供相关数据。现在5个模块的概要设计已经完成。

项目的详细设计部分，现在已完成了所有模块的详细设计。在文件处理模块部分，主要完成了doc、pdf文件的文件处理设计以及相关流程，文件内容的提取也已完成，包括对标题的提取以及正文的提取，失败处理部分主要使用vbs脚本对相关doc文件进行转换，现在相关的vbs脚本已经完成。公告处理模块完成了标题及正文的切分以及日期格式的提取。在分类器模块部分，分类器参数格式设计以及各个分类器的设计已经完成，此外还完成了各个分类器的分类效果计算的流程。分类结果处理模块的相关流程已经设计完成，主要是将信息打包整合导出到xml格式的文件中。在工作日志模块的详细设计中，主要是完成了日志的记录内容的设计，以及日志的相关查询统计功能设计，在日志查询统计中主要是对指定的参数使用情况进行统计查询。

上市公司公告分类项目完成进度表如表2-1所示。

表2-1 项目进度及毕业设计（论文）工作计划表

|  |  |  |  |
| --- | --- | --- | --- |
| 起始时间 | 完成时间 | 计划工作内容 | 备注 |
| 2014.07.09 | 2014.10.10 | 项目需求与可行性分析 | 已完成 |
| 2014.10.13 | 2014.11.07 | 项目概要设计 | 已完成 |
| 2014.11.10 | 2014.12.31 | 详细设计 | 已完成 |
| 2015.01.01 | 2015.03.20 | 编码实现 | 完成80% |
| 2015.03.23 | 2015.05.19 | 系统测试 | — |
| 2015.05.20 | 2015.06.20 | 撰写、修改论文，参加毕业答辩 | — |

## 2.2中期工作说明及成果汇报

### 2.2.1系统总体结构设计

1. 系统总体设计

整个系统主要是由5个功能模块组成，包括：文件处理模块、公告处理模块、分类器模块、分类结果处理模块、工作日志模块。系统的功能结构图如图2-2所示。

文件处理模块主要是对文件进行格式统一化处理以及标题、正文的读取，由于系统所需处理的文件格式不统一，所以在对文件进行分类之前需要将文件转换成统一格式并对文件进行初步的处理以方便下一阶段进行分类处理。

公告处理模块主要是对文件处理模块中的结果进行进一步处理以便于分类器进行分类，公告处理模块主要包括标题正文切分以及日期格式提取。标题正文的切分主要是将文件内容切分成标题及正文两部分，而日期格式提取是将文中出现的日期及其后指定字数的内容提取出来。

分类器模块主要包括参数设置、分类效果计算以及各个分类器，其中参数设置主要是对一些分类器的参数进行设置，分类效果的计算主要是在用户对参数进行设置之后对新参数下的分类器的分类效果进行计算方便用户作为参考，系统的分类器有股东大会分类器、重大资产重组分类器以及停复牌分类器。

分类结果处理模块主要提供一个将分类结果导出到文件的功能，导出的文件类型为xml格式。

工作日志模块主要是对分类器处理过程中一些中间信息进行记录。日志记录主要是提供给分类器使用的，对分类器在分类过程中的一些中间信息进行记录，主要是对一些判断依据（如使用的关键字、条件等）进行记录。

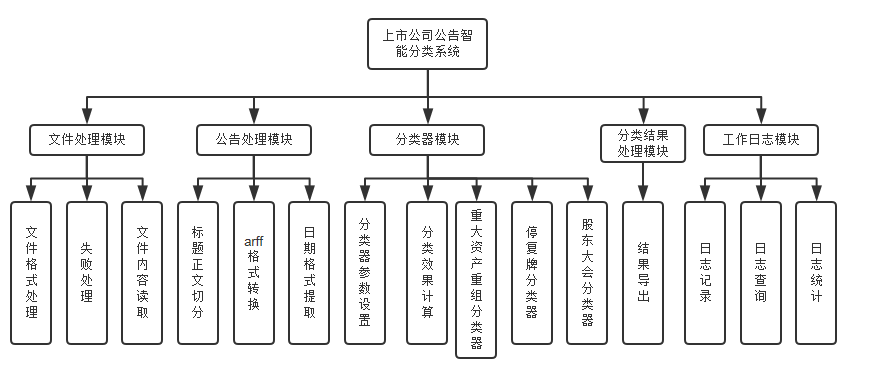


图2-2 上市公司公告智能分类系统功能结构图

1. 分类器设计方案

分类器的设计主要采用分层分类的方法，总共分成二层。

1. 一层分类

一层分类采用随机森林分类的方法来对重大资产重组以及股东大会大类进行分类。随机森林是一个包含多个决策树分类器的集成学习方法，它由多个决策树分类器组合投票进行决策，以达到一个相对公平的决策结果。它具有较强的鲁棒性，能够避免过度拟合而且对包含大量缺省值的数据集具有良好的效果，因此在一层分类中，采用随机森林的分类算法进行分类。其模型构建示意图如2-3所示。

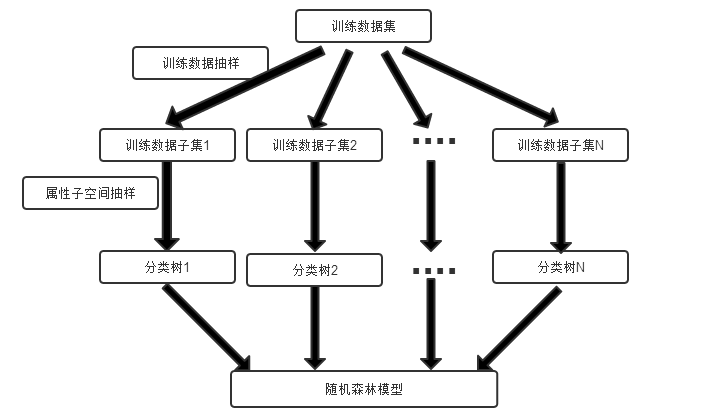


图2-3随机森林模型构建示意图

1. 二层分类：规则分类

由于公告文件中，不同类别的文件存在的存在相对较为明显的特征差异，比如重大资产重组停牌公告与重大资产重组预案中，就存在停牌与预案的差异，因此规则分类适用于对细类的分类，规则分类的实现较为简单，且能够设计成只需修改规则参数就能进行改进，因此不需要再重新建立模型，具有较强的可维护性。

### 2.2.2系统数据库设计

系统的数据库主要包括日志记录、分类结果、文件信息以及分类器参数数据。其中文件信息主要是用来存放用户录入的公告文件的相关信息，包括公告文件ID、公告事件ID、文件发布时间、文件对应股票的交易状态，文件的存放路径等信息。分类器的参数主要是用来存放分类器的配置参数的信息，主要有关键字条件、业务类别、启用状态等信息，关键字条件具有唯一性，用户做出删除操作时，不能删除该关键字条件的记录，而是将启用状态设置为否。分类结果主要是用来存放分类器识别完的结果（即未合并前的结果），主要包含文件的信息以及分类结果。而日志记录则主要用来存放分类器在分类过程中产生的中间信息，主要是分类器分类中所使用的关键字条件，日志记录包含分类结果以及分类器参数等信息。系统初步设计的数据库ER图如图2-4所示。可以根据需要进行进一步的设计。

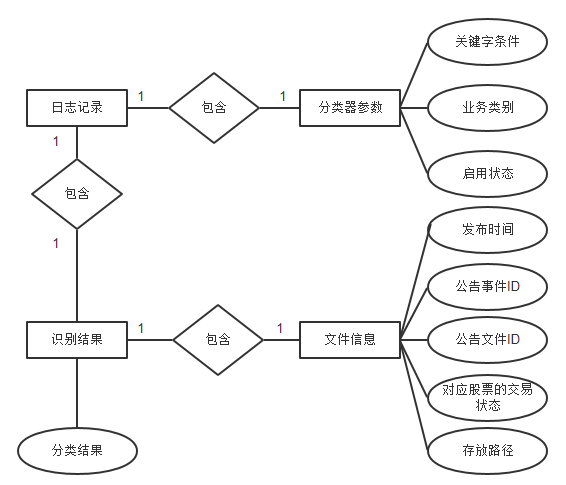


图2-4 上市公司公告智能分类系统数据库ER图

### 2.2.3系统详细设计

### 2.2.3.1文件处理详细设计

文件处理模块包含文件格式处理，文件内容提取以及文件失败处理。在文件处理时，文件处理器首先对文件的格式进行判断，如果不是pdf以及doc格式的文件则返回格式类型错误，否则开始提取文件内容，根据判断的格式采用相应的文件处理器进行读取，在对doc格式文件的读取中，如果出现错误，则调用失败处理，失败处理主要是采用VBS脚本对doc格式文件进行转换，将其转换为txt格式的文件然后在读取内容；在读取完内容后，对读取的内容进行标题和正文的提取。文件处理模块的类图如图2-5所示。



图2-5文件处理模块类图

### 2.2.3.2公告处理详细设计

公告处理主要包含标题arff格式文件转换、正文切分以及日期格式的抽取。

arff格式文件转换主要是为随机森林分类器对文件做的预处理，在进行随机森林分类前，arff格式转换器将文件转换成arff格式的文件，以供随机森林分类器进行分类。对于标题的切分，由于大部分公告标题及正文间有较明显的特征，所有标题主要取内容中指定格式前120个词，剩余的内容则为正文；对于日期格式的提取主要采用正则表达式，正则表达式中由3部分组成：日期字母表，分隔符以及中文日期，字母表主要是阿拉伯数字及中文数字，分隔符包含常见的日期分隔符，中文日期则主要有今日，明日等与日期相关的信息。

### 2.2.3.3分类器详细设计

1. 分类器模块整体设计

分类器中，重大资产重组分类器、股东大会分类器、停牌分类器以及复牌分类都实现分类器接口，其中重大资产重组分类器以及股东大会分类器由随机森林分类器以及规则分类器组成，重大资产重组分类器以及股东大会分类器负责分类的流程；而停复牌分类器则由规则分类器组成，停复牌分类器负责分类的流程，分类器模块的类图如图2-6所示。



图2-6分类器类图

1. 分类参数详细设计

分类器参数主要包含：业务类别，细类，关键词条件以及是否启用这几项。

其中业务类别包括重大资产重组，股东大会，停牌以及复牌。细类中，重大资产重组包括4个类：重大资产重组停牌公告，重大资产重组预案，重大资产重组报告书，停牌期间终止重大资产重组；股东大会主要包括2个类：股东大会决议公告，董事会决议召开股东大会

其字段的设计如表2-7所示。

表2-7 分类器参数字段

|  |  |  |  |
| --- | --- | --- | --- |
| 业务类别 | 细类 | 关键词条件 | 是否启用 |
| 停牌 | 停牌 | “停牌” | 是 |

针对关键词条件的设计，考虑到分类器需要区分正文及标题，以及在使用时需要各个关键词之间的逻辑关系，因此，设计的分类器参数主要包含以下符号：! , “” , ( , ) , and , or , T。其中“”（英文字符）用于表示关键词，引号内的内容为关键词，例如表达正文中包含停牌字样时可以使用 “停牌”；！（英文字符）用于表示逻辑关键中的非，当需要表达不包含某个关键词时可以使用！，例如表达正文中不包含停牌字样时可以使用：！“停牌”；and（小写英文字母，半角）用于表示逻辑关系与，当需要表达正文中既包含关键词停牌，又包含关键词公告时，可以使用：“停牌” and “公告”；or（小写英文字符，半角）用于表达逻辑关系或，当需要表达正文中包含关键词停牌，或者包含关键词公告时，可以使用：“停牌” or “公告”；T（大写英文字母，半角）用于表达该关键词是在标题中使用，当需要表达在标题中包含停牌字样时可以使用：“停牌T”来表示；()（英文括号，半角）用于与其他字符组合使用，例如需要表达标题不包含复牌或者正文包含资产 , 并且标题包含停牌时可以使用：“停牌T” and (!“复牌T” or “资产”)。关键词条件的设计及用例可参见表2-8。

表2-8 关键词条件设计及用例

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 符号 | 作用 | 用例 | 用例含义 | 备注 |
| “” | 用于表示关键词 | “停牌” | 正文包含停牌 | 英文字符 |
| ! | 逻辑非，表示不包含 | ! “停牌” | 正文不包含停牌 | 英文字符 |
| and | 逻辑与，既..又.. | “停牌” and “重组” | 正文包含停牌并且正文包含重组 | 小写英文字母，半角 |
| or | 逻辑或 | “重组” or “停牌” | 正文包含重组或者正文包含停牌 | 小写英文字母，半角 |
| T | 表示该关键词在标题中 | “停牌T” | 标题包含停牌 | 大写英文字母，半角 |
| () | 与其他符号配合使用 | “停牌T” and (!“复牌T” or “资产”) | 标题不包含复牌或者正文包含资产 , 并且标题包含停牌 | 英文字符 |

1. 股东大会分类详细设计

股东大会分类时，系统采用二层分类的方法进行分类。分类时，系统首先将文件按格式进行处理，接着对文件进行arff格式转换，然后进入一层分类，一层分类将转换完成后的文件用指定的随机森林模型进行判别，得出该文件为指定类的概率，依据设定好的阀值，得出该文件是否为该类，若为该类则进入第二层分类，否则返回非该业务类别。进入第二层分类后，二层分类器在初始化时将分类器参数读入内存进行初始化，在接收到传入的文件后，依据分类器参数中设定的条件进行判断得出文件的具体细类，若未匹配到细类则返回非该业务类别，若匹配到结果，则返回所有匹配到的细类。其分类流程如图2-9所示

1. 重大资产重组分类详细设计

在进行重大资产重组分类前，系统首先会对文件进行格式判断，文件内容处理，然后再对文件内容仅arff格式的转换。重大资产重组分类首先需要进行初始化，初始化时将从指定路径上读取重大资产重组分类的模型，模型读取完成后，重大资产重组分类器将arff格式文件载入到模型中，通过模型得出该文件为重大资产重组大类的概率值，通过与设置的阀值进行比较得出其是否为重大资产重组类别，若不为该类则返回空结果，否则进入二层分类进行判断。使用二层分类时，首先重大资产重组分类器将初始化一个规则分类器，在规则分类器中，将所有属于重大资产重组的参数导入到规则分类器中，并初始化参数，然后调用规则分类器对文件内容进行分类。最后返回分类结果。其分类流程可参见图2-9。

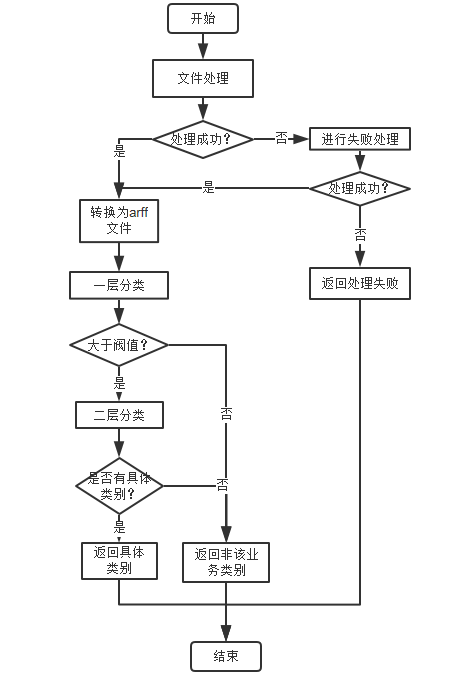


图2-9 重大资产重组及股东大会分类流程图

1. 停复牌分类器详细设计

停复牌分类主要使用二层分类中的关键词条件进行分类。但在进行停复牌分类前需要对文件进行额外的处理，主要是使用公告处理模块中的标题正文切分以及日期格式提取。在进行停复牌分类时，将二层规则分类中的正文替换为日期格式提取的内容，规则中对正文使用的规则在停复牌分类中对日期格式提取内容使用。其流程图如图2-10所示。

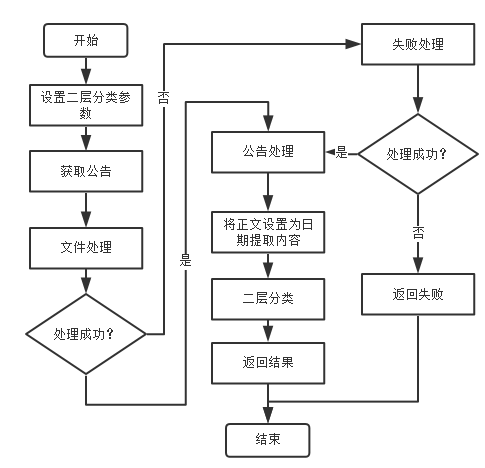


图2-10停复牌分类流程图

1. 分类效果计算详细设计

分类效果计算主要是在用户修改分类器参数之后，为用户提供一个新参数下的分类模拟效果，其类图如图2-11所示。分类效果计算主要包含训练数据选取以及分类效果统计2个部分。但用户修改了新的参数后，可以设置训练数据的情况包括正类样本数，总样本数，若用户未设置则系统启用默认参数进行数据的选取，然后分类效果计算器对该类别进行训练数据的选取，在完成训练数据准备后，使用该类别的分类器（针对该细类的分类器）进行分类，并将每个训练数据的分类结果记录下来，并与训练数据的真正类别进行比较，计算准确率和召回率，其公式如图2-12所示。



图2-11分类器效果计算类图

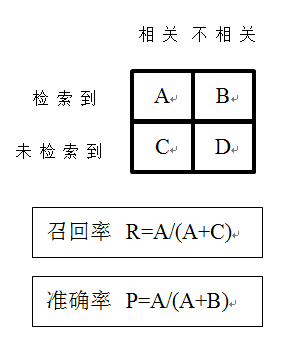


图2-12 准确率和召回率计算公式

### 2.2.3.4分类结果处理详细设计

分类结果处理主要是将分类完成的全部结果导出到文件中，文件的类型为xml格式文件，其结构如图2-13所示。

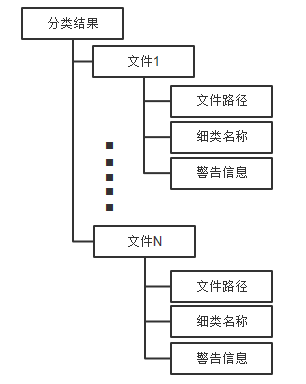


图2-13分类结果文件结构

### 2.2.3.5工作日志详细设计

工作日志主要是将系统的规则使用情况记录下来，当系统使用一条规则判断出该文件属于该类别时，工作日志将记录下文件的路径以及当前使用的规则，作为后续工作的以及系统升级的数据。工作日志提供对这些记录的查询，查询以规则或日期为检索项，检索结果可以是原始的未经过加工的记录结果，也可设置成系统统计后的结果。

### 2.2.4系统实现

### 2.2.4.1文件处理模块实现

1. 文件格式处理实现

文件个数处理主要是对文件格式进行判断。进行格式判断时，系统获取文件路径，并将文件路径统一为小写格式，然后再查找文件路径中最后一个“.”符号，并截取其后面的字符，进行匹配。若未匹配到结果则返回格式错误。

文件处理内容读取实现

1. doc文件处理实现

doc文件处理器在对文件进行处理时，首先调用其父类对文件是否存在进行检查，若不存在直接返回失败，若存在则进行下一步处理，读取文件时通

过调用导入的工具类对文件进行读取，若在读取时出现异常则进行文件失败处理

1. pdf文件处理实现

pdf文件处理器在对文件进行处理时，首先调用父类对文件是否存在进行检查，若不存在则直接返回失败，若存在则进行下一步处理，读取文件时通过调用导入的工具类对文件进行读取，若在读取时出现异常则返回文件不能读取。

1. 文件失败处理实现

在进行文件失败处理时，首先检查失败处理的vbs文件是否存在，若不存在则编辑vbs文件并保存，然后开始进行文件失败处理。处理时调用外部vbs程序进行处理，同时获取设置的超时时间，开启一个计时线程，若超过超时时间，则终止外部程序的调用，返回处理失败，否则调用txt文件读取器读取内容。文件失败处理的流程图如图2-14所示。



图2-14文件失败处理流程图

### 2.2.4.2公告处理模块实现

1. arff格式文件转换

在进行arff格式转换时，arff转换器首先读取相关文件，获取特征词信息，并将特征词写入文件中，并在特征词后写入{true,false}；然后转换器调用分词器对该传入的文件内容进行分词，完成分词后，转换器逐一检查分词中是否包含特征词，若包含特征词则在指定位置写入true，否则写入false。检查完所有分词，arff格式即转换完毕。arff格式如图2-15所示。

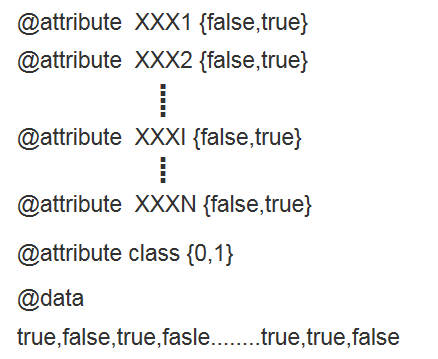


图2-15 arrf格式文件示意图

1. 标题及正文切分

在对文本内容进行标题正文切分时，由于一篇公告的标题通常与正文间存在一段提示性语言，标题正文间存在较明细的特征，所以在进行标题正文的切分时，采用通过定位这段提示性语句的方法来切分，若文中不存在提示性语句或者语句的位置超过文件内容的一半，则取文本内容前10%且不超过120词的语句作为标题，其余内容作为正文。

1. 日期格式提取

日期格式的提取，主要是通过使用正则表达式来对日期格式进行匹配。所以对日期的提取主要是对正则表达式的编写，为了便于对日期正则表达式进行拓展，将正则表达式分为字母表，分隔符以及中文日期这几部分。在正则表达式匹配到结果后，程序对匹配结果进行检查和转换，日期的检查主要是对年份的检查，正则表达式可能匹配到只含有月份和日期的日期格式即未包含年份信息的日期格式，对此程序将用获取到的文件日期进行补全；日期的转换部分，首先将日期转换成统一的格式:“xxxx-xx-xx”然后程序将转换后的日期转换成程序中日期的格式。在对所有日期提前完成后，程序将用文件日期对其进行比较，筛选掉早于文件日期的日期，留下比文件日期晚或者等于文件日期的日期，并将这些日期后指定范围内的文字进行截取。日期格式提取的流程图如图2-16所示。正则表达式的内容可参见表2-17。

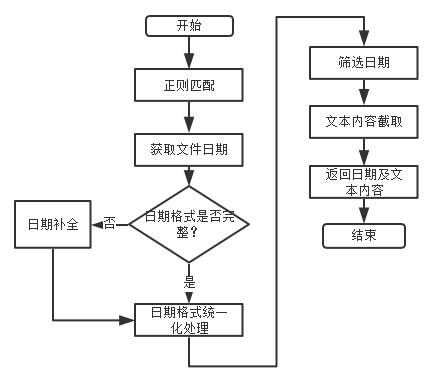


图2-16日期格式提取流程图

表2-17正则表达式细表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 名称 | 代表字符 | 含义 | 内容 | 备注 |
| 字母表 | alphabet | 日期的表达字符 | ０0１1２2３3４4５5６6７7８8９9〇○OΟ零零一二三四五六七八九十 | 包含全角及半角符 |
| 分隔符 | seperateChar | 年月日间的分割符 | |—|－|－|\ | 包含全角及半角符 |
| 年分隔 | year | 年与月间的分隔符 | \s\*[年|-|\|－|—|－]\s\* | 与分隔符一起使用 |
| 月分隔 | month | 月与日间的分隔符 | \s\*[年|-|\|－|—|－]\s\* | 与分隔符一起使用 |
| 中文日期 | regrexSpecial | 中文标示的日期 | 明日 明天 今日 今天 后天 后日 | 无 |
| 正则式（程序的表达式） | regrex="(((("+alphabet+"{4})|("+alphabet+"{2}))"+year+"){0,1}("+alphabet+"{1,2})"+month+"("+alphabet+"{1,2}))|"+"("+regrexSpecial+") | | | |

### 2.2.4.3分类器处理模块实现

1. 分类器参数处理实现

分类器参数处理主要对二层分类中使用的规则参数进行处。分类器参数

处理器首先将从xls文件读取全部参数，根据分类器指定的细类将对应的参数进行逻辑词与关键词的分离，并且分类器参数处理器将对每个关键词产生一个代替符号，将原来含有关键词的参数转换成含有符号以及逻辑关系的逻辑表达式。分类器参数处理器进行逻辑词语关键词的分离时，将会建立一个逻辑词到关键词的映射，除此之外，对参数中包含的业务类别，分类器参数处理器也将建立一个业务类别到细类的映射。分类器参数处理的流程图如图2-18所示。

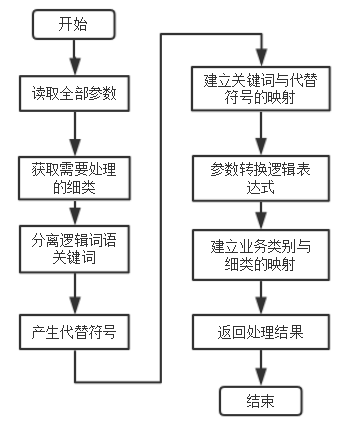


图2-18 分类器参数处理流程图

1. 一层分类器实现
2. 随机森林分类模型的训练

在一层分类中，分别对重大资产重组以及股东大会建立不同的模型。在建立模型前，需要分别对重大资产重组以及股东大会的所出现的词语进行词频统计，计算其区分度。区分度的计算公式如图2-19所示。分子表示该词在目标类文件中出现的次数，分母表示该词在非目标类文件中出现的次数，区分度由两者之商得出。得出特征词的区分度后，选取所有大于某一阀值的特征词作为该类的特征，然后从所有训练数据中对训练样本用weka软件进行训练，得出相应的模型。

1. 随机森林分类实现

一层分类器主要是实现随机森林分类。在项目中，随机森林的模型是利用Weka软件生成的。首先对训练数据进行特征词的选取，选取的依据为区分度的大小，然后将训练数据转换成一个arff文件作为weka软件的输入，weka将依据输入的文件生成一个模型文件。一层分类器中主要是对这份模型的文件的使用，当进行随机森林分类时，分类器将进行初始化，首先将模型文件读入到weka工具类中的一个分类器，然后随机森林分类器获取处理好的文件，并将文件转换为weka工具类中的一个实例类对象，weka提供的分类器对实例进行分类得出分类的结果。随机森林分类器根据设置好的阀值以及分类的得分返回分类结果，其流程如图2-20所示。

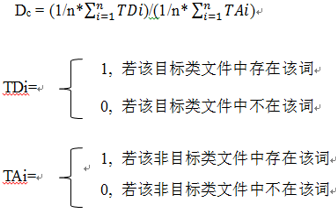


图2-19 区分度计算公式

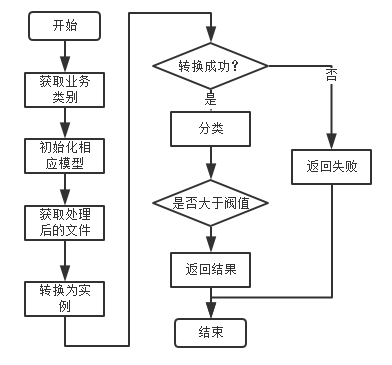


图2-20随机森林分类流程图

1. 二层分类器实现

二层分类器主要是实现规则分类。首先规则分类器获取需要初始化的参数细类名称，并用参数处理器对指定细类的参数进行处理，完成参数的初始化。完成初始化后，分类器逐一取出每个初始化后的参数并获取需要进行分类的文件的标题以及正文信息，然后，分类器获取参数中关键词到代替字符的映射，并对这些关键词进行判断，若相对位置中存在该关键词，则将该关键词对应的代替字符设置true否则置为false，并建立代替字符的真假值。在完成所有关键词的判断后，程序获取参数中的逻辑表达式，并用代替字符的真假值作为输入用js脚本进行最终结果的判断。若判断结果为真则分类器根据规则记录的设置，来进行规则使用情况的记录。在完成每一条规则的判断后，分类器将所有结果存在集合中返回。二层分类的流程图如图2-21示。

1. 分类器效果计算实现

在进行分类效果计算时，首先分类效果计算器获取需要计算效果的细类，

并根据设置好的训练数据的参数进行训练数据的选取，若为设置训练数据参数则使用默认参数选取训练数据。训练数据选取主要是选取指定数量的正类数据，然后选取足够数量的其余类别的数据组成指定数量的总样本。完成训练数据选取后，分类器效果计算根据细类初始化分类器并关闭日志记录进行分类，然后将每篇文件的分类结果与文件的真实类别进行比较，并统计准确率和召回率。分类效果计算流程图如图2-22示。

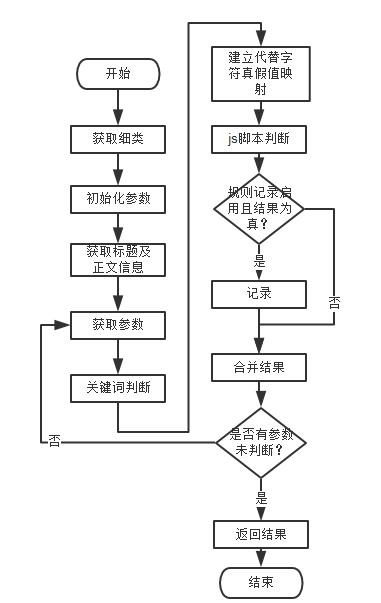


图2-21规则分类器分类流程图

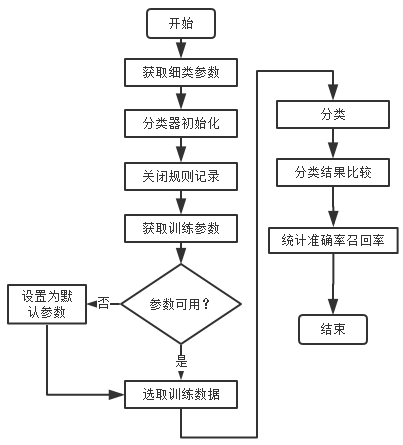


图2-22分类效果计算流程图

### 2.2.2.4分类结果处理模块

分类结果处理模块主要是将结果导出到文件中，导出时首先，获取指定的文件路径，然后取出分类结果中的各个字段的值，最后用导入的xml工具类将结果写到对应的节点上，保存文件。

### 2.2.2.5工作日志模块实现

工作日志模块包括日志记录，查询及统计。日志记录主要是在二层分类时进行记录，当二层分类器判断出结果为真时，日志记录将使用的规则以及文件的信息写入数据库。日志查询统计主要是对日志记录进行的一个查询统计的操作，用户输入检索项，系统将从数据库查找相应的记录显示出来，系统可以对一定日期范围的记录进行统计。

### 2.2.5系统实现效果

系统运行时首先进行初始化，初始化过程中主要是依据默认设定初始化相应的分类器。初始化界面如图2-23所示。

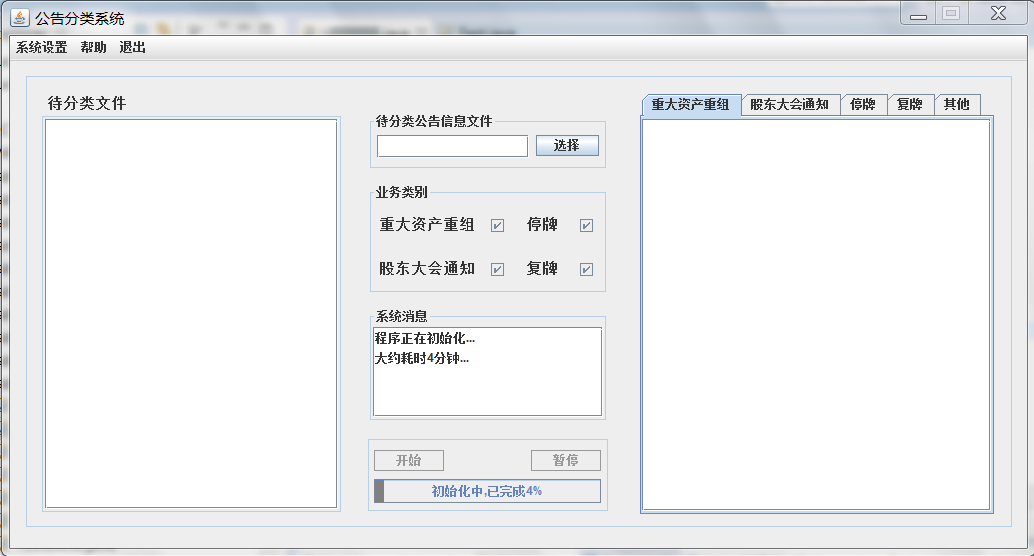


图2-23系统初始化

初始化完成后，开始进行分类。如图2-24所示。

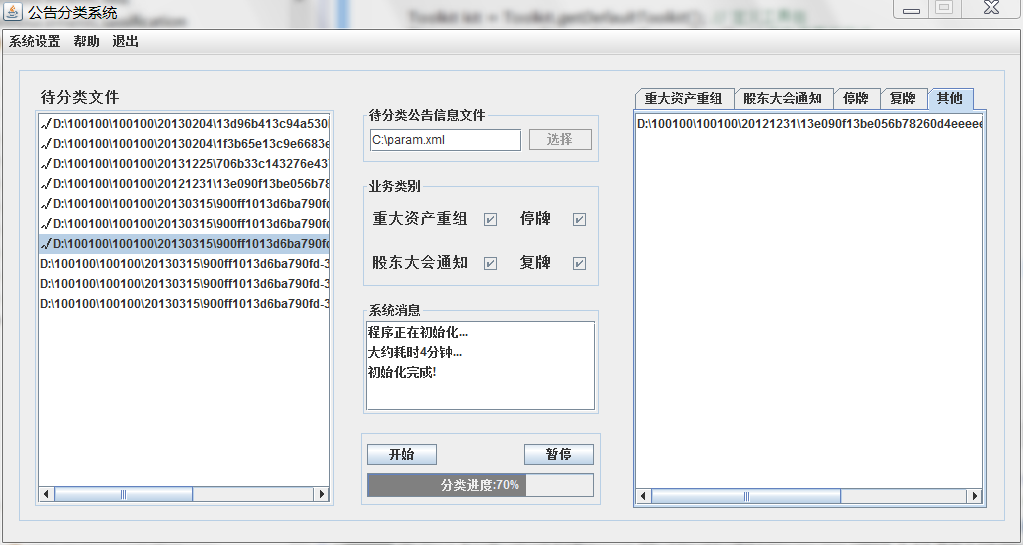


图2-24系统进行分类

继续分类，直至分类完成，如图2-25所示。

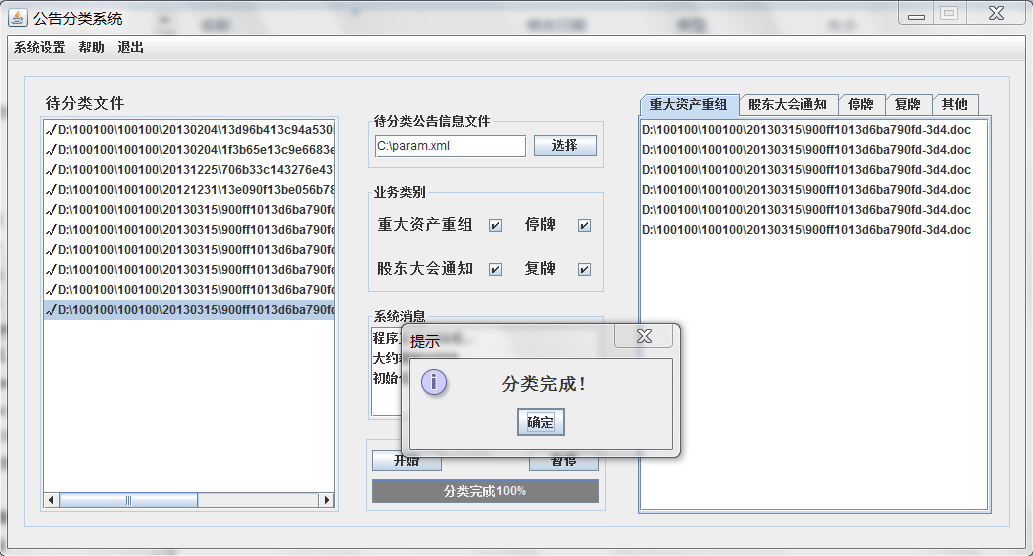


图2-25 分类完成

重新设置业务类别进行分类，如图2-26所示。

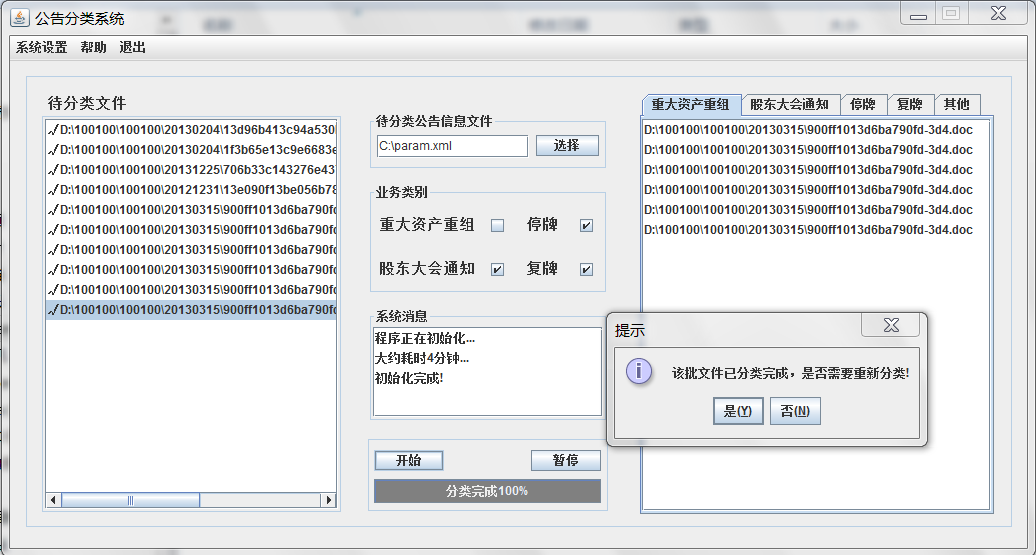


图2-26 设置业务类别重新分类

## 2.3存在的困难与问题

1. 初始化时间较久，系统在进行初始化时由于需要导入重大资产重组的随机森林模型以及股东大会的随机森林模型，所以将耗费较长的时间，平均初始化时间在120秒左右。
2. 分类时将占用较大的系统资源，由于需要对文件进行大量的处理，因此在系统进行分类时将占用较多的系统资源。

## 2.4如期完成预定任务的可能性分析

目前项目的实现已经完成了大部分，文件处理，公告处理，分类器，分类效果计算，日志记录以及文件导出等部分的实现及测试已经完成，还未完成的部分主要是界面，包括分类效果计算界面、参数设置界面、文件导出界面以及日志查询界面。因此项目能够在预定的时间内完成。

## 2.5后期工作安排（或进度和计划调整）

后半期的工作进度及时间安排如表2-27所示。

|  |  |  |  |
| --- | --- | --- | --- |
| 起始时间 | 完成时间 | 计划工作内容 | 备注 |
| 2015.04.01 | 2015.04.03 | 分类器参数设置界面实现 | — |
| 2015.04.04 | 2015.04.06 | 分类效果计算界面实现 | — |
| 2015.04.07 | 2015.04.10 | 日志查询统计界面实现 | — |
| 2015.04.11 | 2015.04.13 | 文件导出界面实现 |  |
| 2015.04.13 | 2015.05.20 | 系统测试 | — |
| 2015.05.20 | 2015.06.20 | 撰写、修改论文，参加毕业答辩 | — |

表2-27 后期工作进行及时间安排**附件：本科毕业设计(论文)中期检查意见表**

|  |  |
| --- | --- |
| 基地导师意见（需写具体内容） | |
| 签 字：  年 月 日 | |
| 校内导师意见（需写具体内容） | |
| 见邮件截图  签 字：  年 月 日 | |
| 中期检查小组意见 | |
| 结论：◎ 通过  ◎ 警告  ◎ 不通过  具体意见： | 评委签字：      年 月 日 |

附件：导师意见邮件截图

