封面单独打印！

（居中，注意不要缩进）

厦門大學

（小二号宋体，加组）

本 科 毕 业 论 文 （设 计）

（主修专业）

（二号黑体）

（三号宋体，加粗）

面向非结构化企业指标信息的智能处理和可视分析

Indicators of the Unstructured Enterprise Information for Intelligence Processing and Visualization

（三号 Times New Roman 加粗）

阅读请开批注模式

姓　　名：赵四

学　　号：20420132201111

学　　院：化学化工学院

专　　业：化学

年　　级：2013 级

校内指导教师：张三　教授

校外指导教师：（姓名）（职务）

方块全部换成上面一样的虚线

数字用四号 Times New Roman

四号宋体，按照学校发的 Word 版封面格式

四号宋体，姓名与职称之间空两格。下面的**校外指导教师即使没有也原样保留**

所有空格与标点都跟随文字字体和大小。英文用 TNR 或 Arial，中文用宋体或黑体，均取决于当处字体要求。

二〇XX 年六月一日

（四号宋体）

厦门大学本科学位论文诚信承诺书

电子版封面背后空白页删除！

本人呈交的学位论文是在导师指导下独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合相关法律规范及《厦门大学本科毕业论文（设计）规范》。

该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明）。

本人承诺辅修专业毕业论文（设计）（如有）的内容与主修专业不存在相同与相近情况。

必须手签

学生声明（签名）：

年　　月　　日

封面之后、正文之前的页码用大写罗马数字表示。

背面空白页要标有页码，用大写罗马数字表示。

三号宋体，加粗，中间空两格，2 倍行距

知 悉 书

日期中文宋体，数字 TNR。

本人_____年____月开始，在厦门大学化学化工学院____老师的课题组参与了 “_____”（注：此处请填写论文题目）等课题的研究，本人知悉这期间在本课题组所接触的数据和工艺等的知识产权均属于厦门大学所有，受相关法律法规的保护。因此，在即将毕业之际，本人特此保证：

字体要求与前后文一致，务必确认。

该段落号为上方文档格式编号，复制导致段落号消失请到开始-段落里找，不要自己手动输入。

（1）将本阶段获得的实验数据用于任何目的（如：发表学术论文、会议论文、申请专利、产业化等）之前，需经_____老师的书面许可。

Times New Roman

（2）不将与实验相关的秘密以任何形式泄露给他人。

（3）学位论文、实验数据提供给他人阅读、复制之前，需经_____老师的书面许可。

系别:_____

学号和日期的数字用 Times New Roman

专业:_____

学号:_____

必须手签

学生（签字）:

年 月 日

页码用大写罗马数字表示。

背面空白页要标有页码，用大写罗马数字表示。

（小三号黑体，中间空两格，
1.5 倍行距，段前段后 0.5 行）

致　谢

（首行缩进 2 个字符，小四号宋体）

值此论文完成之际，谨向所有关心和支持我的人们致以诚挚的谢意！

首先，我要衷心地感谢我的导师 XXX 教授。……

页码用大写罗马数字表示。

背面空白页要标有页码，用大写罗马数字表示。

（小三号黑体，中间空两格，1.5 倍行距，段前段后 0.5 行）

摘　要

随着信息的发展，出现了越来越多的非结构化信息。并且非结构化信息在政府和企业等的决策中扮演着重要的角色。如何将非结构化数据有效的管理起来，能够进行数据和知识挖掘，提取当中的隐含信息，提供一种形象的可视分析，为政府和企业决策提供支持成为当今亟待解决的主要问题。

（小四号宋体，1.5 倍行距，正文不要设置段前段后距）

本文以北京市科委的指数统计文档为研究对象，主要任务是针对以北京市科委的指数统计文档为代表的非结构化信息的抽取和企业指标信息的可视分析。主要工作包括三个方面：第一，设计了一套以北京市科委的指数统计文档编写规范为标准的确实可行的信息抽取算法；第二，针对抽取出来的指标信息，借助于 Dundas 可视化工具进行可视分析；第三，完成了一个满足客户需求的企业信息库管理系统。

论文从项目背景出发，介绍了系统开发的背景和研究价值。然后，详细介绍了企业指标信息智能处理的可行性和算法设计，以及企业指标信息可视分析的原理及其实现。再次，论文详细阐述了系统的需求，具体介绍了企业信息库管理系统的设计及其实现，最后论文针对企业信息库管理系统进行了分析和评价，并指明了下一步的改进计划。

（小四号黑体）　　（小四号宋体）

关键词：非结构化信息；信息可视化；可视分析

中文分号“；”分开，最后一个关键词不打标点符号

页码用大写罗马数字表示。

背面空白页要标有页码，用大写罗马数字表示。

（小三号 Times New Roman 加粗）

Abstract

With the development of information, there has been an increasing number of unstructured information. And it plays an important role in decision of government and enterprise, etc. How to manage the unstructured information efficiently, mine the data and knowledge, extract the implicit information, provide a visual image analysis, and then support the government and enterprise's decision have become the main issues to be settled urgently.

（小四号 Times New Roman）

In this question for discussion, we mainly have a research in indicator of enterprise documents from the Beijing Science and Technology Commission and try to obtain the indicators of the unstructured information, and then provide a visual image analysis. It includes three aspects: First, to design a set of practical information extraction algorithm; second, through the use of the Dundas Chart toolbox, providing visual analysis; third, completed Enterprise Information Management System which meet customers requirement.

The beginning of the dissertation introduced the background of the project, introduced the background of the system and research value. Second, detailing information extraction algorithms and principles of Information Visualization. Third, the dissertation elaborated the system's requirement, specifically introduced the system design and implementation. Finally, some possible improvements and future works were presented.

（小四号 Times New Roman）

Key words: Unstructured Information; Information Visualization; Visual Analysis; Education Database

（小四号 Times New Roman 加粗）

第二行，悬挂缩进与第一行对齐，缩进精度可到小数点后两位。
该处缩进为 4.10 字符。

英文分号“;”加一个空格，分开，最后关键词不打标点符号

页码用大写罗马数字表示。

分页使用分页符，分节同样。

背面空白页要标有页码，用大写罗马数字表示。

（四号黑体，编号与文字标题之间空一格，中文字体是黑体，英文和数字字体采用 Arial）

一级标题后不加点！

（小三号黑体，中间空两格）

（目录中的虚线要统一格式，均匀、不加粗，各标题的页码和页码**前导虚线**统一为五号 Time New Roman）

目　录

前导虚线必须与页码一致

1 绪论 1

1.1 引言 1

1.2 研究内容和方法 2

1.3 论文组织结构 2

2 系统相关技术概述 4

2.1 非结构化信息处理 4

2.1.1 非结构化信息管理概述 4

2.1.2 信息抽取技术 5

2.2 信息可视化 7

2.3 其它系统技术介绍 8

2.3.1 ASP.NET 简介 8

2.3.2 ASP.NET AJAX 简介 9

2.3.3 ASP.NET Ajax Control Toolkit 组件 11

2.3.4 Dundas Chart 工具箱简介 12

2.4 本章小结 14

3 非结构化信息处理和可视分析 16

3.1 企业指标信息统计分析设计方案 16

3.2 企业指标信息的智能处理 17

3.2.1 企业指标信息文档的结构分析 17

3.2.2 指标信息的提取算法设计 24

3.2.3 指标值的提取算法设计 33

3.3 信息可视化的设计方案 34

3.3.1 信息可视分析过程模型 34

3.3.2 基于 Dundas 的信息可视分析设计 36

3.4 本章小结 39

4 企业信息库管理系统的实现 39

4.1 系统概述及功能 39

4.1.1 开发背景与系统目标 39

4.1.2 系统功能和模块划分 40

4.2 系统的框架设计 42

4.3 指数统计模块的实现 43

4.3.1 统计分析模块的实现 43

（中文小四号黑体，英文和数字 Arial，编号与文字标题之间空一格）

（小四号宋体，编号与文字标题之间空一格，中文字体是宋体，英文和数字字体采用 Times New Roman。行距使用 1.25 倍。）

行距 1.25 倍

页码用大写罗马数字表示。

4.3.2 问卷管理模块的实现 45
4.4 文档资源库模块的实现 46
4.5 系统维护模块的实现 48
4.5.1 用户管理子模块的实现 48
4.5.2 角色管理子模块的实现 49
4.5.3 文档类型定义子模块的实现 49
4.5.4 数据库备份&还原的实现 49
4.7 本章小结 49
5 系统测试及运行结果 50
5.1 系统测试 50
5.2 运行结果 50
5.2.1 统计分析模块的运行结果 50
5.2.2 问卷管理模块的运行结果 52
5.2.3 文档资源库模块的运行结果 54
5.2.4 用户管理子模块的运行结果 55
5.2.5 角色管理子模块的运行结果 56
5.2.6 文档类型定义子模块的运行结果 57
5.2.7 数据库备份&还原的运行结果 58
5.2.8 改善用户体验的工作 59
5.3 本章小结 60
6 结论 61
6.1 论文总结 61
6.2 工作展望 62
参考文献 63
附录 64

双面打印，背面要标有页码，用大写罗马数字表示。

（小三号 Times New Roman 加粗）

（目录中的虚线要统一格式五号 Time New Roman，均匀、不加粗；页码统一为五号 Time New Roman）

（四号 Times New Roman 加粗，编号与英文标题之间空一格）

Content

1 Introduction 1

1.1 Introduction 1

1.2 The structure of this dissertation 2

2 Literature review 4

（小四号 Times New Roman 加粗，编号与英文标题之间空一格）

2.1 Unstructured information management 4

2.1.1 Introduction of unstructured information 4

2.1.2 Information extraction 5

2.2 Information visualization 7

2.3 Other related technologies introduce 8

2.3.1 Introduction of ASP.NET 8

2.3.2 Introduction of ASP.NET AJAX 9

（小四号 Times New Roman，编号与英文标题之间空一格）

2.3.3 ASP.NET Ajax Control Toolkit 11

2.3.4 Dundas Chart Toolkit 12

2.4 Summary 14

3 Unstructured information management and visulization 16

3.1 The design philosophy of enterprise indicators 16

3.2 The design philosophy of enterprise indicators extraction 17

3.2.1 The statistics document’s structure analysis 17

3.2.2 The statistics information extraction algorithm 24

3.2.3 The value of statistic extraction algorithm 33

3.3 The design philosophy of information visualization 35

3.3.1 Information visualization model 35

3.3.2 Information visualization base on Dundas Chart Toolkit 36

3.4 Summary 38

4 Implementation of enterprise infromation management 39

4.1 System profiler and function 39

4.1.1 Development background and overall objective 39

4.1.2 Functional requirements and module division 40

4.2 System architecture 42

4.3 Indicators of statistics module design 43

4.3.1 Statistical Analysis module design 43

页码用大写罗马数字表示。

4.3.2 Questionnaire management module design 45
4.4 Document management module design 46
4.5 System maintenance module design 48
4.5.1 User management sub-module design 48
4.5.2 Role management sub-module design 49
4.5.3 Document attribute management sub-module design 49
4.5.4 Database backup and restore 49
4.7 Summary 49
5 System testing and the running results 50
5.1 System testion 50
5.2 Running results 50
5.2.1 Statistical Analysis module running results 50
5.2.2 Questionnaire management module running results 52
5.2.3 Document management module running results 54
5.2.4 User management module running results 55
5.2.5 Role management module running results 56
5.2.6 Document attribute management module running results 57
5.2.7 Database backup and restore running results 58
5.2.8 Improve the system-experience 59
5.3 Summary 60
6 Conclusions and future works 61
6.1 Conclusions of the dissertation 61
6.2 Future works 62
References 63
Appendix 64

双面打印，背面要标有页码，用大写罗马数字表示。

（正文从另右页开始，每一章应另起页，并从奇数页开始；奇数页页眉为当前章名，小五号宋体。）

一级标题：
（中文小三号黑体，编号和英文 Arial 字体，编号顶格，**与文字标题之间空一格，段前段后 0.5 后，共占 2 行。**）

1 绪论

二级标题:
（中文四号黑体，编号和英文 Arial，编号顶格，与文字标题之间空一格）
1.5 倍行距，段前段后 0 行。

1.1 引言

（小四号宋体，1.5 倍行间距，两端对齐）

随着计算机技术的发展，使海量信息得以存在并迅猛发展。尤其是信息技术的日益普及其应用以后，随着各个行业的信息系统的规模的日益扩大，信息系统在长年累月的运转过程中，积累了庞大的数据资源。然而决策者却很难利用这些数据资源，为企业和政府的决策提供确实有效的帮助。这是因为一方面，在这庞大的数据资源中，非结构化信息占据了主要部分。Gartner的一项调查显示，在今天的社会中，有80% 以上的商业行为依赖于非结构化信息；我们所存储的数据中，85 ℃以上是非结构化信息；每过三个月，我们周围的非结构化信息就会增加一倍[1]。这些数据充分说明，我们周围信息的形态是以非结构化信息为绝对主体的，也可以说我们接触到的信息中绝大部分是非结构化信息。因此对非结构化信息进行管理，能够进行数据和知识挖掘，提取当中的隐含信息，对决策进行支持成为当今亟待解决的主要问题[2]。另一方面，随着信息技术的发展，信息结构越来越复杂，信息更新越来越快，信息规模越来越大，给人们获取信息、理解信息、掌握信息带来了沉重的负担，常常导致“认知过载”、“视而不见”[3][4]。

北京市科学技术委员会在企业指标信息统计分析工作上就存在这两方面的问题，文献[5]介绍了这方面的工作。每年北京市科委都要对北京市企业进行企业指标信息的调查，在长年累月的积累过程中，北京市科委积累了大量的企业指标调查表、项目立项、执行、验收等文档。这些调查表以word形式保存起来，并且调查指标的方式也呈现多样化，存在选择、填空、表格、问答以及这些题目的复合等形式。而且企业指标的调查涵盖范围也很广泛，从企业性质及登记情况到企业财务及信息化投入状况，再到人力状况及信息化支撑状况，到企业信息化基础设施建设状况、企业信息化应用情况，甚至涉及到企业对信息化工程的满意程度的调查。面对海量的非结构化企业指标信息，北京市科委每年都要投入大量的人力、物力、精力，将企业指标信息从word文档中手工提取出来，形成计算机可以识别的结构化的表格信息，再对企业指标信息进行统计分析。即使是这样，仍然存在许多问题：第一，手工抽取企业信息调查表耗时较长，工作强度大。第二，手工抽取数据信息容易出现错误，准确性不能得到有效保证，而且一旦出错，就

除角度（°）外，数字与单位之间必须空格。

%不是单位。
详见 GB 3100-93

正文英文和数字除特别要求字体全部统一 times new roman。

顺序编码制参考文献的标注方式：数字加方括号在右上角，置于句号之前。

（正文之后的页码用阿拉伯数字连续编码，小五号 Times New Roman。）

（偶数页页眉为论文题目，小五号宋体。）

有可能导致整个统计分析结果的错误，进行核对非常困难。第三，即使是将企业指标信息全部准确转成计算机可以识别的表格数据以后，由于数据的多样性，缺少形象的对企业指标信息的统计分析工具。

针对北京市科委的企业指标信息统计分析问题，我的毕业设计结合北京市科委的业务需求，开发了企业信息库管理系统。这个项目来源于国家科技支撑计划项目课题“面向服务的智能化制造技术及示范应用”（课题编号2006BAF01A17）。该项目主要是为了解决北京市科委的指标信息统计分析过程中，存在指数统计困难和文档管理困难两个问题，以业务为主线，主要包括科委文档的管理、企业指标信息的智能处理、企业指标信息的可视分析三个方面的内容。通过为科委中存在的大量信息文档实体构建基础信息模型，来方便用户的日常管理和提高文档的利用率。通过构建应用数据模型，将企业指标信息文档中的非结构化信息智能抽取出来，并存储于数据库当中，将非结构化信息结构化，用成熟的结构化数据管理理论来管理非结构化数据。通过对指标信息的查询，构建信息可视分析模型，使用户可以对知识进行挖掘，提供形象的可视分析，提高北京市科委的企业指标信息的统计分析效率。本项目完成后将会在北京市科委投入使用。

括号字体全中文，正文全宋体，标题全黑体。

二级标题与上文之间空一行。

三级标题：
（中文小四号黑体，编号和英文用 Arial 字体，
编号与文字标题之间空一格）
1.5 倍行距，段前段后 0 行。
与上文之间无空行

1.2 研究内容和方法

1.2.1 研究内容

1.3 论文组织结构

本论文共分为六章，论文首先分析了政府和企业在信息化过程中遇到的两个问题：非结构化信息管理和“认知过载”。并结合北京市科委的企业指标统计分析问题，介绍了毕业设计项目的背景和研究价值，引出了论文所做的主要工作内容。紧接着论文简单概述了毕业设计项目中所用到的各项技术，并针对北京市科委的业务要求提出了信息抽取和基于 Dundas Chart 信息可视化的解决方案。然后论文就项目中的两个技术难点——非结构化信息处理和信息可视分析，详细阐述了信息抽取技术的算法设计和信息可视分析技术的模型设计。在系统实现方面，论文详细介绍了企业信息库管理系统所使用的技术要点：基于 Asp.net 的三层结构（USL-BLL-DAL）的框架设计；在用户体验方面，采用了 Asp.net Ajax 改善

用户的体验。

论文具体安排如下：

第 1 章 简单介绍了企业和政府在信息化过程中遇到的非结构化信息管理困难和“认知过载”问题。针对北京市科委的指标统计分析问题，提出了毕业设计的背景、目标和研究价值。

第 2 章 概述系统中所使用的各项技术及各项技术的国内外发展现状。

第 3 章 详细介绍了针对北京市科委企业指标信息文档的信息抽取技术的算法设计和信息可视分析的模型设计。

第 4 章 介绍了企业信息库管理系统的实现。详细阐述了系统的背景和总体目标，基于表示层（USL）-业务逻辑层（BLL）-数据访问层（DAL）的三层结构的框架设计和功能模块介绍及其实现。

第 5 章 介绍了企业信息库管理系统的系统测试和运行结果。

第 6 章 最后论文总结了毕业设计所做的工作，并且指明了下一步的改进计划。主要是在信息抽取算法的改进，以及在用户体验方面的改进计划。

2 文献综述

每一章均从奇数页开始。如前一章在奇数页结束，那么下一页偶数页留空白。

2.1 非结构化信息处理

2.1.1 非结构化信息管理概述

在引言中，我们提到过“在当今的社会中，我们周围信息的形态是以非结构化信息为绝对主体的， 也可以说我们接触到的信息中绝大部分是非结构化信息。”，那么什么是非结构化信息？非结构化信息具有什么特点？如何管理非结构化信息？

信息可以分为三类：结构化信息，非结构化信息和半结构化信息。

1. 结构化信息——经过严格标引后的数据，一般以二维表的形式存在。如数据库中的表、各种票据信息等等。

结构化信息又分为以下三种：

（1）一维结构化信息。

三级标题以下不允许使用四级标题，直接用（1）（2）（3）等表达分节，详见左侧所示。

一维结构化信息可以进一步分为以下两类：

① 第一类一维结构化信息。

② 第二类一维结构化信息。

（2）二维结构化信息。

（3）三维结构化信息。

2. 非结构化信息——没有经过人为处理的不规整的信息。这些信息更加符合人类交流的方式。如新闻报道、科技文献、散文等等。

3. 半结构化信息——介于结构化信息和非结构化信息之间的。有一定格式约束，这不同于非结构化信息，但局部上，又按人类自然语法组织信息，与结构化信息又有所区别，例如电报报文，通知、公告、指数统计表等等。

非结构化信息具有如下特点：第一，其格式非常多样；第二，标准是多样性的，不像我们结构化的数据一目了然；第三，在技术上非结构化信息比结构化信息更难标准化和理解。所以存储、检索、发布以及利用需要更加智能化的计算机技术。

基于非结构化信息的特点，将非结构化信息结构化，转化为结构化信息进行管理是一个可行的管理方案，而构建的面向用户的企业非结构化信息管理系统必

须具备以下特征[5]：

1. 必须对非结构化信息资源的获取、转换、分析、管理、应用全过程进行分析，提供基于标准工作过程的支持环境。

2. 必须提供标准的对外接口、信息描述方法和定制规范降低定制分析机组件和信息应用组件的复杂性。

3. 必须提供灵活的信息描述资源模式简化信息结构化信息资源库的构建。

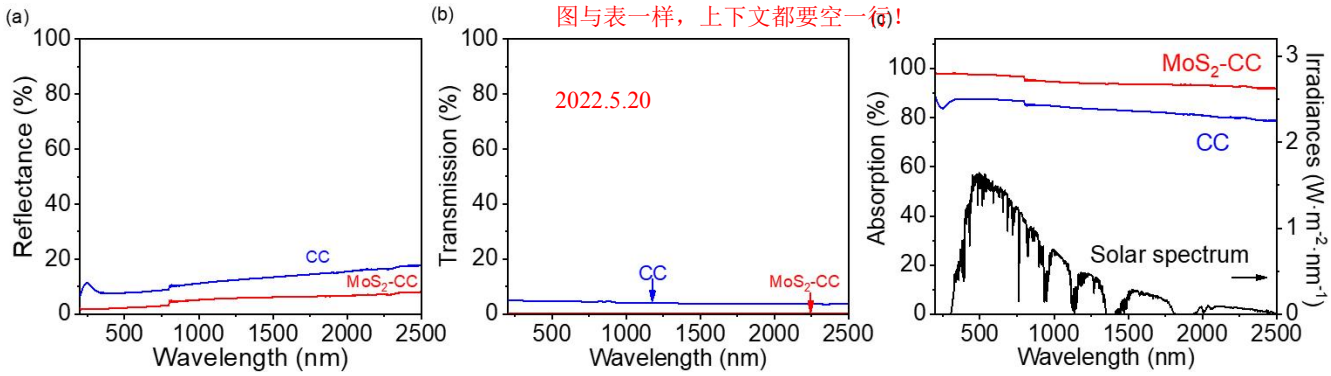
4. 采用自然资源技术以支持高质量的“拉式”信息服务和知识抽取。

5. 提供对外的标准的接口以支持非结构化信息资源管理系统与企业其他应用系统的集成。

6. 提供界面友好的工具方便用户系统管理和应用。

7. 其本身应具有易于扩充、动态发展的能力。

图2-1为基于UIMA（Unstructured Information Management Architecture）的非



图与表一样，上下文都要空一行！

结构化信息管理的架构图，具有一定的指导意义：

图2-1 (a-b)MoS_2-CC复合材料反射透射率；(c)复合材料光吸收率与光功率分布[6]

（图与图标题、图序号为一个整体，不得拆开排版为两页。当空白不够排版该图整体时，可将其后文字部分提前，将图移至次页最前面。建议自己的图表要增加英文图，表题）

（图号、图名居于图下方正中，五号宋体加粗，图号编号须与表格、公式编序方式 统一。）

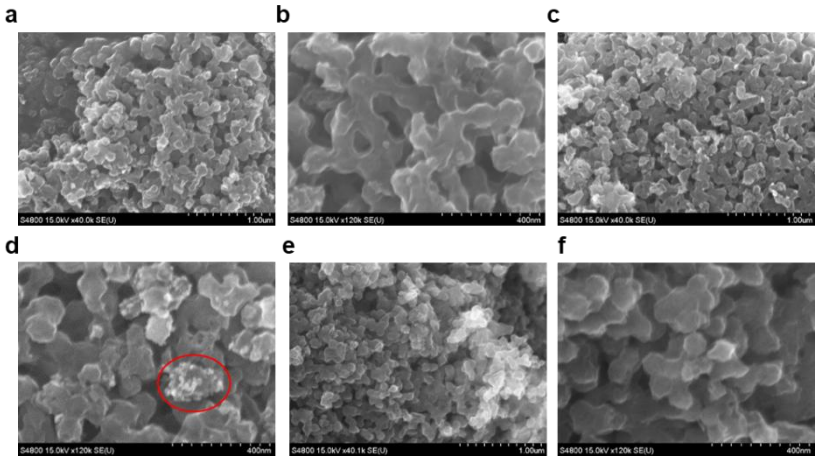
或者 图2-1 MoS_2-CC复合材料吸光能力：(a-b)反射透射率；(c)光吸收率与光功率分布[6]

在把列名映射到 Dundas 里面的图例，而行名则映射为 Dundas 里的轴标签。完成了数据表的映射以后，剩下的就是图表自身形态的改变了。

为了实现 Dundas 形态的改变，我们对 Dundas 的属性进行了分类和总结，如表 2-1 所示：

表 2-1 Dundas 的部分属性表

属性	描述
图表类型 (Chart Type)	条柱型图表(Bar and Column Charts)：条形图、柱状图； 线型图表(Line Charts)：折线图、曲线图、阶梯图； 点图表(Point Charts)：点图、泡泡图； 饼图(Pie Charts)：饼图、圈图； 分区图(Area Charts)：折线分区图、曲线分区图；
条柱宽度 (Point Width)	针对条柱型图表，条柱的宽度。取值从(0,1)。
条柱风格	针对条柱型图表，有默认、砖型、圆形、棱型、明暗变化



数值标签 (Value Label)	是否显示数值标签。
3D 显示	是否 3D 显示。
簇状显示	是否簇状显示。
图例 (Legend)	字体属性；字号属性；显示位置：图表的左边、右边、上面、下面。
标签 (Axis)	字体属性；字号属性。
标题 (Title)	字体属性；字号属性。

选择算子决定了哪些染色体进入下一代。本算法中采用“轮盘赌”的选择方式，它按照染色体的适应值大小来确定该染色体的被选择概率。如果染色体的适应值越大，其被选中的概率越大。个体 r_i 被选中的概率 p(ri)定义如下：

$$p(r_i)=Fitness(c_i)/\sum_{j=1}^{pSize} Fitness(c_j) \qquad \text{(公式 2-1)}$$

$$O_2 + H^+ + e^- \rightarrow {}^*OOH + H^+ + e^- \rightarrow {}^*O + H_2O + H^+ + e^- \rightarrow {}^*OH + H^+ + e^- \rightarrow H_2O \qquad \text{(公式 2-2)}$$

确定了每个染色体的被选择概率后，系统生成一个在[0，1]区间的随机数组，然后与对应染色体的被选择概率比较，如果随机数大于染色体的被选择概率则该染色体被选择，反之被淘汰。

算法 2-1 直线拟合算法

Begin

(1) 对 V_l 中的每个元素 aa，重复执行以下的步骤：

mx=mx+aa.X；

my=my+aa.Y；

mxx=mxx+aa.X*aa.X；

mxy=mxy+aa.X*aa.Y；

(2) If mx*mx-mxx*n=0

拟合失败

公式去掉首行缩进，居于正中间，括号与数字均用 times new roman，所有字大小全用小四，数字之间用横杠相连，即：(公式 X-X)。

物理数学以及简单化学反应公式字体使用 TNR，小四。有机反应公式直接用图。

当公式过长时，公式编号另起一行，置于右下。

Else

k=(my*mx-mxy*n)/(mx*mx-mxx*n)；

b=(my-mx*k)/n；

类公式短语，符号与括号字体跟随英文字体，不要中文。

End

定义 2-1 如果存在一条从 V_i 到 V_j 的路，称 V_i 是 V_j 的前驱节点，而对于(V_i，V_j)∈E，称 V_i 是 V_j 的立即前驱节点，记为 V_i∈iPred(V_j)，称 V_j 是 V_i 的立即后继节点，记为 V_j∈iSucc(V_i)。

定义一个公共容器类型的代码如下：

```
class Container : public Object{
public:
        virtual Object* get();                    //删除并返回当前元素
        virtual void put(Object*);                //在当前元素之前插入
        virtual Object*& operator[] (size_t);     //下标
        //…
};
```

2.4 本章小结

本章详细介绍了针对北京市科委企业指标信息文档的信息抽取技术的算法设计和信息可视分析的模型设计。

首先，我们参考了 UIMA 的非结构化信息的管理体系结构，并结合北京市科委的实际业务要求，提出了自己的非结构化企业指标信息的管理模型。并指出了在这个指标模型当中的两个技术难点：非结构化信息的提取和信息可视分析的实现。然后就存在的两个技术难点展开了详细的分析和设计。

其次，我们详细阐述了信息抽取算法的思想。首先，我们了详细分析了企业指标统计表中存在的规律和模型，抽象出企业指数统计表中存在的五条规则，并提出了用信息抽取技术中的知识工程法进行信息抽取的可行性。为了更好地进行指标信息的提取，我们给出了三条建议。其次，在总结的规则的基础上，我们阐述了企业指标信息分析的流程图。结合科委的业务情况，将企业指标信息的分析分成指标提取和指标值提取两个方面。 紧接着，结合企业指数信息表的规则，我们给出了指标信息提取的体系结构图，包括分块、题目树的构建、题目的分割、题目的细化、基础模型的解析五个步骤。并详细阐述了各个步骤的算法思想。最

后，我们阐述了指标值抽取的算法思想。

最后，我们结合 Card 模型，提出了企业信息管理系统信息可视分析的模型。然后，我们分析 Dundas 工具箱的元素和属性，并详细阐述了该模型在 Dundas 中的实现。

6 结论

6.1 论文总结

论文分析了北京市科委的指标统计分析业务中遇到的两种问题：非结构化信息管理困难和“认知过载”问题，并详细介绍了针对这两种问题国内外系统的解决方案。结合北京市科委的业务需求，提出了建立企业信息库管理系统建设方案。

论文详细介绍了将非结构化信息结构化，利用成熟的结构化信息的管理方案来解决非结构化信息管理的问题，借鉴 UIMA 的体系结构，结合业务需求，提出了信息库管理系统体系结构。并详细阐述了系统中存在的两大技术难点的解决方案——指标信息智能抽取的算法设计和信息可视分析的模型设计。首先就指标信息抽取的算法设计，我们详细分析了北京市科委指标文档存在的规则，提出了利用知识工程法来抽取指标信息的体系结构，详细阐述了指标抽取和指标值抽取的算法设计。其次，在信息可视化方面，我们构建了利用 Dundas 工具箱的可视分析的模型。

本文还介绍了企业信息库管理系统的实现，详细阐述了系统的需求和系统目标，基于三层结构（USL-BLL-DAL）的架构和功能模块设计，同时介绍了系统的主要功能模块和相关流程图。最后介绍了 Ajax 在本系统的运用和改善用户体验所做的工作。论文的主要内容如下：

1. 简单介绍了企业和政府在信息化过程中遇到的非结构化信息管理困难和“认知过载”问题。针对北京市科委的指标统计分析问题，提出了企业信息库管理系统的解决方案。

2. 详细阐述了指标信息智能抽取和信息可视分析的算法设计思想。在指标信息抽取的算法设计思想上，我们详细分析了北京市科委指标文档存在的规则，提出了利用知识工程法来抽取指标信息的体系结构，详细阐述了指标抽取和指标值抽取的算法设计；在信息可视化方面，我们提出了利用 Dundas 工具箱的可视分析的模型。

3. 介绍了企业信息库管理系统的实现。详细阐述了系统的背景和总体目标，基于表示层（USL）-业务逻辑层（BLL）-数据访问层（DAL）的三层结构的框架设计和功能模块介绍及其实现。

4. 最后总结了论文所做的所有工作，并且指明了下一步的改进工作。

6.2 工作展望

当然，企业信息库管理系统中还存在着许多的不足，我们将对它进行进一步的完善和改进：

1. 企业信息库管理系统的界面比较简单，因此在新一轮的迭代开发测试中，对原有的界面设计进行进一步的改进。

2. 非结构化指标信息的抽取算法还有待提高。我们将对非结构化指标信息的抽取算法进行进一步改进，使它能够更加合理的解决其他企业中的类似问题，进一步提高算法的通用性。

3. 在信息可视分析中缺乏交互效果，而且信息可视分析的配置过程比较繁琐，我们将在在下一轮的改进这个配置过程并利用的 Dundas 的支持 Ajax 的新特性增加仪表盘的交互能力。

4. 系统的扩展性需要改进。虽然采用了三层架构可以方便地扩展，但由于快速开发，在代码中直接使用了 SQL 语句进行操作，降低系统的性能，存储过程可以改进这个缺陷。

5. 系统的维护工作，以及文档的完善。

6. 代码的优化。

序号后 Tab 制表符分隔，中文参考文献与英文参考文献内部各行抬头对齐一致。

（小三号黑体，居中）

参考文献

[1] 杨福生，高上凯．生物医学信息处理[M]．北京:高等教育出版社, 1988.

[2] 陈川波．基于半结构化文本信息抽取的简历识别系统[D]．北京:北京邮电大学, 2008.

[3] 张德政，张萍萍．非结构化信息管理[J]．微计算机信息, 2006, 22(3):218-239.

[4] 谢希德．创造学习的新思路[N]．人民日报, 1998-12-25(10).

[5] Foster I, Kesselman C. The Grid: Blueprint for a New Computing Infrastructure [M]. Morgan Kaufmann Publishers, 1998.

[6] Foster I, Kesselman C, Nick J, et al. Grid services for distributed systems integration [J]. IEEE Computer, 2002, 35(6):135-160.

[7] Xiang S D, Scholzen A, Minigo G, et al. Pathogen recognition and development of particulate vaccines: does size matter? [J]. Methods, 2006, 40(1):1-9.

（中文，五号宋体）

（英文和数字，五号 Times New Roman）

调整缩进字符使参考文献正文两端对齐，该处为 1.46 字符

期刊使用全称，不允许简写。

（作者 3 人以下全部列出，3 人以上可只列出前 3 人，后加“, 等.”，外文用“, et al.”。中英文每条文献中的逗号句号统一用英文半角逗号句号加空格取代；页码前的冒号用英文冒号，后不加空格。）

期刊标注必须为（以下只是展示用，详细格式上文所示已有）：
期刊名（全称）， 年份， 卷（期）:页码（xx-xx）.

没有卷可以直接写期，没有明确页码，可以写文章编号。

（小三号黑体，居中，中间空两格）

附　录

论文附录依次用大写字母“附录 A、附录 B、附录 C……”表示，附录内的分级序号可采用“附 A1、附 A1.1、附 A1.1.1”等表示，图、表、公式均依此类推为“图 A1、表 A1、式 A1”等

图、表、公式，格式字体要求同正文一致。

附录 A 计算原始数据

附 A1 Cu_{13} 团簇的林德曼指数-温度曲线数据

温度/K	Lindemann index
200	0.0318254459682325
250	0.0365486017060895
300	0.041359874196529
350	0.147201945312047
370	0.225872393504144
400	0.259485985818768
450	0.282297155382904
500	0.31054533005517
600	0.32352013915776
800	0.330791219871914
1000	0.332383087278731

题目等非具体内容的一般性信息，统一用宋体，数字与英文用 times new roman，字体 11 号，下表均同。

任 务 书

题目			
姓名		学号	
指导教师		职称	
类型	□论文 □设计	写作语言	□中文 □其他_____
来源	□纵向项目 □横向项目 □其它	选题方式	□导师指定 □自选
论文写作要求	（对论文写作目标、思路、方法、进度安排等提出要求）		
支持条件	（可提供必要的资料和数据支持）		
文献阅读要求	（可列出推荐参考文献）		
签名	指导教师签名： 年 月 日		

（括号内填表说明内容保留，以下略同）

（该表由教师填写，可电脑输入打印，填写的字体建议为中文五号宋体、英文和数字五号 Times New Roman，1.3 倍行距，首行缩进 2 个字符，以下略同）

注：不足部分可加页。

（必须由教师用黑色钢［水］笔手签，**日期请填写 1 月份前或开学前的时间**）

（该表由学生根据指导教师提供的任务书填写，可电脑输入打印，该表一面打印不完可双面打印。）

开 题 报 告

题目			
姓名		学号	
研究目标			
研究思路			

（表内各空栏地方可电脑输入打印，建议字体为中文五号宋体、英文和数字五号 Times New Roman，1.3 倍行距，首行缩进 2 个字符，以下略同）

<table>
<tr><td>研究方法</td><td colspan="2"></td></tr>
<tr><td rowspan="6">具体进度安排</td><td>起讫时间</td><td>计划完成内容
（一般可分为资料文献搜索、拟定方案（提纲）、试验或初稿、定稿等阶段）</td></tr>
<tr><td>年 月 日- 年 月 日</td><td></td></tr>
<tr><td>年 月 日- 年 月 日</td><td></td></tr>
<tr><td>年 月 日- 年 月 日</td><td></td></tr>
<tr><td>年 月 日- 年 月 日</td><td></td></tr>
<tr><td>年 月 日- 年 月 日</td><td></td></tr>
<tr><td rowspan="2">签名</td><td colspan="2">学生签名： 年 月 日</td></tr>
<tr><td colspan="2">指导教师签名： 年 月 日</td></tr>
</table>

（必须由学生用黑色钢［水］笔手签，**日期填写建议到 2 月-3 月初，比任务书填写的时间要晚**）

注：不足部分可加页。

（必须由教师用黑色钢［水］笔手签，**日期填写建议到 2 月-3 月初，比任务书填写的时间要晚**）

（该表由教师填写，可电脑输入打印，该表一面打印不完可双面打印。**签名部分必须由教师用黑色钢［水］笔手签，日期请按论文工作各阶段时间填写。**）

教师指导记录

题目			
姓名		学号	
第一阶段指导	（建议字体为中文五号宋体、英文和数字五号 Times New Roman，1.3 倍行距，首行缩进 2 个字符，以下略同） 指导教师签名：　　年　　月　　日		
第二阶段指导	指导教师签名：　　年　　月　　日		
第三阶段指导	指导教师签名：　　年　　月　　日		
第四阶段指导	指导教师签名：　　年　　月　　日		
第五阶段指导	指导教师签名：　　年　　月　　日		

注：不足部分可加页。

（该表由教师填写，可电脑输入打印，该表一面打印不完可双面打印。**签名部分必须由教师用黑色钢［水］笔手签，**）

指导教师评语

题目			
姓名		学号	
（从论文的选题、基本写作规范、文献综述、原创性、工作量、所取得的成果、学生的知识和能力，以及完成论文的态度等方面进行总体评价） （括号内填表说明的内容保留，以下略同） （建议字体为中文五号宋体、英文和数字五号 Times New Roman，1.3 倍行距，首行缩进 2 个字符）			
拟评成绩： 是否具备答辩资格：□具备　　□不具备 指导教师签名：　　年　　月　　日			

注：不足部分可加页。

（前五行由学生电脑输入打印，后面部分由答辩秘书负责填写，答辩记录由学生答辩后整理完提交给答辩秘书）

答 辩 记 录

题目						
学院				专业		
学生姓名				学号		
指导教师				职称		
答辩日期	_____年_____月_____日			答辩地点		
答辩小组	姓名					
	职称					
答辩记录	（包括论文陈述、答辩小组成员提出的主要问题及学生答辩的简要情况）					

（建议字体为中文五号宋体、英文和数字五号 Times New Roman，1.3 倍行距，首行缩进 2 个字符）

答辩记录	（包括论文陈述、答辩小组成员提出的主要问题及学生答辩的简要情况）
答辩小组意见	是否通过答辩：□是　　□否 答辩成绩： 组长签名： 成员签名： 秘书签名：　　年　　月　　日

论文总评成绩	总评成绩评定规则： 参照答辩成绩。 总评成绩： 院长（系主任）签名：　　年　　月　　日