

硕士研究生学位论文

|  |  |
| --- | --- |
| 题目： | “海淘”市场热销产品 |
|  | 识别方法设计与实现 |

|  |  |
| --- | --- |
| 姓 名： | 董丹丹 |
| 学 号： | 1301210603 |
| 院 系： | 软件与微电子学院 |
| 专 业： | 软件工程 |
| 研究方向： | 电子商务与物流 |
| 导师姓名： | 李杰 教授 |

二〇一五 年 七 月

版 权 声 明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者授权，不得对本论文进行复制、抄录、拍照、修改或以任何方式传播等行为。否则，引起有碍作者著作权益之问题，将可能承担法律责任。

摘要

近几年来，中国的电子商务行业飞速发展。其中，“跨境电子商务”中的进口部分俗称为“海淘”发展迅速，被称为电子商务的又一片蓝海。从2010年海淘交易额仅120亿元，到2013年的767亿元，2014年预计超过1500亿，市场前景广阔。如何获得精准的“海淘”市场热销产品信息对于电商企业扩大“海淘”市场份额具有重大意义。

美国电子商务企业E集团创建了中文“海淘”页面与其英文网站进行链接，以期进一步提高其中国 “海淘”市场份额。如何精准的识别“海淘热销产品”信息并用于中文页面展示，对进一步扩大E集团中国“海淘”市场份额有重大的商业意义；同时对其他电子商务企业运营工作将具备重要的参考意义。

本文针对电商企业掌握的数据资源进行热销商品信息的识别，包括两个方面。第一、基于内部数据仓库TERADATA：依据“海淘”业务流程进行数据挖掘，完成热销产品识别方法的分析与设计；第二、基于外部“什么值得买”海淘论坛：利用爬虫采集信息，采取文本挖掘方法，完成热销产品识别方法的分析与设计。

具体研究内容包括：第一、基于内部数据仓库TERADATA进行数据挖掘，建立“基于海淘流程的历史购买模型”获得历史购买热销产品信息；建立“基于搜索关键词的供求缺口模型”获得潜在需求的热销产品信息。第二、对上述两个模型完成自动化设置，包括利用SAS完成代码自动更新运行设置，利用KETTLE, VBA等进行结果自动化制作以及展示设置。第三、基于企业外部论坛，采取爬虫框架SCRAPY从“什么值得买”海淘论坛中采集四年的帖子文本数据，利用PYTHON进行文本挖掘，获得热销产品信息。

综合两套方法获得的信息，完成对中国消费者“海淘”热销产品信息的识别，为电子商务企业E集团提供运营决策支持，较大程度提高了其市场交易额，扩大了其中国“海淘”市场份额。同时，本文的研究对于电子商务企业识别市场需求工作具有一定的参考作用。

关键词：海淘，TERADATA，供需模型，SCRAPY，结巴分词，文本挖掘

Design and implementation of method for

discovering “Hai Tao” popular items

Dandan Dong ( Software Engineering )

Directed by Professor Jie Li

ABSTRACT

In recent years, China's e-commerce industry has a great development. The "Import of cross-border e-commerce", which is commonly known as "Haitao", because of its rapid development, it is honored as another piece of blue ocean of e-commerce. Haitao had increased from only 12 billion yuan to 76.7 billion yuan between 2010 to 2013, and was expected to more than 150 billion in 2014.Haitao has a great market future. How to precisely discover the hot items for haitao market is definitely important to expanding the market cap of e-commerce company.

In order to further expand market cap of E group(the US e-commerce business company)in China haitao market, E group created the "Haitao" page in Chinese to link with US web site. So discovering the hot items to display on the haitao page is so crucial to improve business performance and expand the market cap of E group. Besides, it is also with great reference to other e-commerce company’s operation.

This paper is to discover the hot items information based on the e-commerce company resources, including two aspects. Firstly, based on the enterprise internal TERADATA data warehouse, to finalize the analysis and design of the method of discovering the hot items of haitao based on process of Haitao" through data mining. Secondly, based on SMZDM haitao BBS to finalize the analysis and design of the method of discovering the hot items of haitao based on SCRAPY web information from SMZDM through text mining.

Detail research of this paper including: Firstly, Based on TERADATA DW to build “The historical purchase model based on haitao process” to find out the hot items and "the gap of supply and demand model based on keywords" to find out the popular products for potential demand. Secondly, do the automation of the two models with SAS and KETTLE VBA. Thirdly, for the external haitao BBS, use SCRAPY to collect four years (2011-2014) text BBS information, do the text mining by PYTHON to discover the hot items information for haitao.

Combine the information of the two sets of method to finalize the method of discovering of hot items for haitao, which is to support E group operation decision. This method has greatly improved the haitao market cap and gross domestic buying of Chinese consumers. Meantime, this paper has crucial reference to other e-commerce group operation works including discovering hot items.

KEY WORDS**:** “Hai Tao”, TERADATA, The Supply And Demand Model, SCRAPY, Stammer Participles, Text mining.

目 录

[第一章 绪论 1](#_Toc420520410)

[1.1 研究背景与意义 1](#_Toc420520411)

[1.2 研究目标与内容 5](#_Toc420520412)

[1.3 研究方法 8](#_Toc420520413)

[1.4 论文主要贡献 9](#_Toc420520414)

[1.5 论文框架结构 9](#_Toc420520415)

[第二章 研究现状 11](#_Toc420520416)

[2.1 跨境电商研究现状 11](#_Toc420520417)

[2.2 爬虫技术应用现状 11](#_Toc420520418)

[2.3 文本挖掘研究现状 12](#_Toc420520419)

[2.4 电商市场需求研究现状 14](#_Toc420520420)

[第三章 主要技术介绍 17](#_Toc420520421)

[3.1 数据仓库TERADATA 17](#_Toc420520422)

[3.2 自动化过程 20](#_Toc420520423)

[3.3 网络爬虫原理 21](#_Toc420520424)

[3.3.1 工作原理及关键技术 21](#_Toc420520425)

[3.3.2 网页搜索策略 22](#_Toc420520426)

[3.3.3 爬虫框架SCRAPY 22](#_Toc420520427)

[3.4 文本挖掘 23](#_Toc420520428)

[3.4.1 文本预处理 24](#_Toc420520429)

[3.4.2 分词模型HMM隐马尔可夫 25](#_Toc420520430)

[3.4.3 分词算法VITERBI 27](#_Toc420520431)

[3.5 数据可视化 28](#_Toc420520432)

[第四章 企业内部信息识别方案设计 29](#_Toc420520433)

[4.1 基于海淘流程的历史购买模型 30](#_Toc420520434)

[4.1.1 “海淘”流程介绍 30](#_Toc420520435)

[4.1.2 甄别中国消费者购买信息 31](#_Toc420520436)

[4.1.3 数据结构及模型设计 32](#_Toc420520437)

[4.1.4 模型代码设计 33](#_Toc420520438)

[4.2 基于搜索关键词的供求缺口模型 34](#_Toc420520439)

[4.2.1 模型背景介绍 35](#_Toc420520440)

[4.2.2 供需缺口模型 35](#_Toc420520441)

[4.2.3 模型代码设计 39](#_Toc420520442)

[4.3 代码及报告自动化 39](#_Toc420520443)

[4.3.1 基于SAS的代码自动更新 40](#_Toc420520444)

[4.3.2 基于KETTLE的自动化数据处理 43](#_Toc420520445)

[4.3.3 基于VBA的自动化数据处理 46](#_Toc420520446)

[4.3.4 基于Tableau的可视化 48](#_Toc420520447)

[4.4 结果展示 48](#_Toc420520448)

[4.5 本章小结 51](#_Toc420520449)

[第五章 企业外部信息识别方案设计 53](#_Toc420520450)

[5.1 数据抓取 53](#_Toc420520451)

[5.1.1 数据抓取准备 53](#_Toc420520452)

[5.1.2 SCRAPY爬虫的工作流程 54](#_Toc420520453)

[5.1.3 爬虫功能及模块设计 55](#_Toc420520454)

[5.1.4 数据采集问题解决 57](#_Toc420520455)

[5.2 数据结构设计 58](#_Toc420520456)

[5.3 文本挖掘 65](#_Toc420520457)

[5.3.1 文本预处理 65](#_Toc420520458)

[5.3.2 文本挖掘分析 67](#_Toc420520459)

[5.4 结果展示 69](#_Toc420520460)

[5.5 本章小结 71](#_Toc420520461)

[第六章 总结与展望 73](#_Toc420520462)

[6.1 本文主要工作 73](#_Toc420520463)

[6.2 本文创新点 73](#_Toc420520464)

[6.3 不足与展望 74](#_Toc420520465)

[参考文献 75](#_Toc420520466)

[附录 79](#_Toc420520467)

[A 重要代码 79](#_Toc420520468)

[B “海淘”热销产品信息结果展示 82](#_Toc420520469)

[C 文本挖据&数据可视化工具 83](#_Toc420520470)

[致谢 85](#_Toc420520471)

[北京大学学位论文原创性声明和使用授权说明 86](#_Toc420520472)

图表目录

[图1.1 中国网购市场交易规模季度走势 1](#_Toc420520628)

[图1.2 海淘流程 2](#_Toc420520629)

[图1.3 中文“海淘”背景介绍 3](#_Toc420520630)

[图1.4 项目整体背景介绍 4](#_Toc420520631)

[图1.5 热销产品识别总体方法 6](#_Toc420520632)

[图1.6 内部信息识别流程 7](#_Toc420520633)

[图1.7 外部信息识别流程 7](#_Toc420520634)

[图1.8 研究方法 8](#_Toc420520635)

[图1.9 本文结构 10](#_Toc420520636)

[图2.1 爬虫技术总结 12](#_Toc420520637)

[图2.2 需求预测方法 15](#_Toc420520638)

[图2.3 电商市场需求研究总结 16](#_Toc420520639)

[图3.1 TERADATA 数据仓库并行处理机制 18](#_Toc420520640)

[图3.2 TERADATA 数据仓库检索机制 19](#_Toc420520641)

[图3.3 SCRAPY基本工作流程 23](#_Toc420520642)

[图3.4 知识发现流程 24](#_Toc420520643)

[图3.5 文本挖掘流程 24](#_Toc420520644)

[图3.6 文本特征提取过程 25](#_Toc420520645)

[图3.7 隐马尔可夫模型 26](#_Toc420520646)

[图3.8 HMM模型分词例子 27](#_Toc420520647)

[图3.9 VITERBI模型预测例子 27](#_Toc420520648)

[图4.1 本章总览 29](#_Toc420520649)

[图4.2 基于TERADATA建模识别热销产品信息 29](#_Toc420520650)

[图4.3 历史购买模型设计过程 30](#_Toc420520651)

[图4.4 “海淘”整体流程 30](#_Toc420520652)

[图4.5 基于“海淘”流程甄别过程 31](#_Toc420520653)

[图4.6 代码构建逻辑图 34](#_Toc420520654)

[图4.7 历史购买模型设计过程 35](#_Toc420520655)

[图4.8 基于关键词搜索的市场空间模型 36](#_Toc420520656)

[图4.9 市场不饱和状态分析 36](#_Toc420520657)

[图4.10 供给增加成交增加 37](#_Toc420520658)

[图4.11 供给增加成交增加续 38](#_Toc420520659)

[图4.12 供求缺口模型整体流程介绍 38](#_Toc420520660)

[图4.13 供求缺口模型逻辑图 39](#_Toc420520661)

[图4.14 权限设置 40](#_Toc420520662)

[图4.15 通过Putty登陆SAS 42](#_Toc420520663)

[图4.16 运行schedule.txt 42](#_Toc420520664)

[图4.17 SPOON自动化工程界面展示 43](#_Toc420520665)

[图4.18 SPOON Automation时间设置 44](#_Toc420520666)

[图4.19 加载数据库参数 44](#_Toc420520667)

[图4.20 从数据仓库中加载数据 45](#_Toc420520668)

[图4.21 加载数据模块处理 45](#_Toc420520669)

[图4.22 获取系统时间 46](#_Toc420520670)

[图4.23 数据透视表连接数据仓库 48](#_Toc420520671)

[图4.24 热销产品类别信息 49](#_Toc420520672)

[图4.25 热销产品品牌信息 49](#_Toc420520673)

[图4.26 热销产品详细信息 50](#_Toc420520674)

[图4.27 Tablueau热销类别对比数据图 50](#_Toc420520675)

[图4.28 中文页面信息展示图 51](#_Toc420520676)

[图5.1 论坛获取信息整体流程 53](#_Toc420520677)

[图5.2 SCRAPY工作流程 54](#_Toc420520678)

[图5.3 Item模块设置 56](#_Toc420520679)

[图5.4 Setting模块设置 56](#_Toc420520680)

[图5.5 数据采集结果 57](#_Toc420520681)

[图5.6 XML数据存储结果 57](#_Toc420520682)

[图5.7 数据库表结构 59](#_Toc420520683)

[图5.8 数据清洗 66](#_Toc420520684)

[图5.9 品牌关键词词频统计 67](#_Toc420520685)

[图5.10 类别关键词词频统 67](#_Toc420520686)

[图5.11 商城关键词词频统计 68](#_Toc420520687)

[图5.12 类别词频统计结果展示图 69](#_Toc420520688)

[图5.13 热门类别信息 70](#_Toc420520689)

[图5.14 服装类热门分类信息 70](#_Toc420520690)

[图5.15 服装类热门品牌信息 70](#_Toc420520691)

[图5.16 商城市场占有率 71](#_Toc420520692)

[表2-1 数据挖掘分类 13](#_Toc420520693)

[表2-2 数据挖掘基本方法 14](#_Toc420520694)

[表3-1 数据库定义语言 Data Definition Language(DDL) 19](#_Toc420520695)

[表3-2 数据库操作语言Data Manipulation Language (DML) 19](#_Toc420520696)

[表3-3 数据库控制语言Data Definition Language(DDL) 19](#_Toc420520697)

[表4-1 购买信息分类数据结构 32](#_Toc420520698)

[表4-2 脚本命令含义 43](#_Toc420520699)

[表4-3 模型效果展示 48](#_Toc420520700)

[表5-1 回帖评论索引表 59](#_Toc420520701)

[表5-2 站点索引表 60](#_Toc420520702)

[表5-3 版面索引表 60](#_Toc420520703)

[表5-4 品类关键字索引表 60](#_Toc420520704)

[表5-5 品牌关键字索引表 60](#_Toc420520705)

[表5-6 商城关键字索引表 61](#_Toc420520706)

[表5-7 情感关键字索引表 61](#_Toc420520707)

[表5-8 金额关键字索引表 61](#_Toc420520708)

[表5-9 主题商城关键字统计表 61](#_Toc420520709)

[表5-10 主题情感关键字统计表 62](#_Toc420520710)

[表5-11 主题金额关键字统计表 62](#_Toc420520711)

[表5-12 部分商城词库 62](#_Toc420520712)

[表5-13 部分品牌 词库 63](#_Toc420520713)

[表5-14 部分类别词库 63](#_Toc420520714)

[表5-15 评价词词库 64](#_Toc420520715)

[表5-16 折扣信息词库 64](#_Toc420520716)

[表5-17 货币信息词库 65](#_Toc420520717)

[表5-18 项目管道任务 66](#_Toc420520718)

# 第一章 绪论

## 1.1 研究背景与意义

近几年来，中国互联网的高速发展，基数巨大的互联网用户推动中国电子商务行业的迅速发展，中国IT研究中心发布的《2014年Q1中国移动网购市场调研报告》中指出，中国2014年Q1 “网购”总体市场规模达6479亿元，环比增长近12%;“网购”用户数量已超过3.1亿[1][2][3]，如下图1.1显示。



图1.1 中国网购市场交易规模季度走势

其中，跨境电子商务作为电子商务中的黑马迅猛发展，中国报告大厅数据显示：仅2014上半年我国跨境电商交易额达3 万亿元，其中进口（海淘）增长速度远大于出口贸易。中国电子商务研究中心的数据显示，中国2013 年电子商务市场交易规模达10.2 万亿元，其中跨境电商的进出口部分交易额约为3.1 万亿元。跨境电商的进口部分海外代购市场规模达767 亿元，同比增长58.8%，预计2014 年这一数字将达到1549 亿元。中国已然成为美国跨境“网购族”的第二大目的国，覆盖率高达39%[4]。

跨境电子商务的进口部分，俗称为“海淘”。图1.2描述了其整个过程，首先需要注册电商网站，第三方支付软件账号。通过电子商务相关网站进行购物，通过电子订购单发出购物请求并进行在线付款。通常通过多币种信用卡或者第三方支付如PayPal，支付宝等，款到后发货。之后由海外购物网站卖家通过国际快递直邮到国内，或由转运公司代为收货物后再转寄回国。通过海关入境，期间会涉及海关申报以及关税环节，最后在国内通过快递服务邮寄到消费者手中。

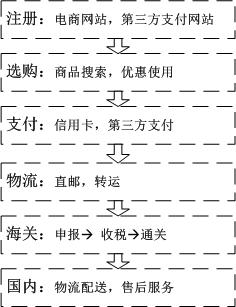


图1.2 海淘流程

在上述电子商务迅速发展的宏观背景下，E集团在中国发展“海淘”具备优势，包括各大电商企业大力发展海淘业务、国内政策支持、E集团自身的强大电商背景。

首先：宏观背景下国内大型电商企业纷纷介入：莫岱青高级分析师提到“国内各大电商巨头大力发展跨境电商业务”。其中早期有天猫，1号店，“亚马逊中国”，“淘宝全球购”等国内电子商务巨头大力发展海淘，此后，苏宁易购，“聚美优品”，“唯品会”等国内知名电商企业全部都在抢占“海淘”市场，都希望建立自己的物流体系以及跨境品牌渠道[3]。

其次：国内政策支持。莫岱青指出：由于国内大型主流跨境电商企业加入“海淘”市场，在政策、品牌、物流等诸多方面都会比转运、代购更有优势：（1）、政策方面：基于国家优惠政策支持，各大电商大多采用“海外直邮+保税进口”作为商品入境渠道，并且享受出口退税；(2)、品牌方面：“海外购”、天猫、京东、苏宁易购、1 号店、“聚美优品”、“唯品会”等电子商务企业在国内具备较高的知名度，地域辐射范围较广，品牌信誉度高，使得“海淘”业务具备了较高的信任度；（3）、物流速度方面：由于国家政策支持以及企业一体化物流供应链的建设，直邮服务整个购买周期由半个月可压缩到3 天，甚至更短。相关学者预测15 年将是跨境电子商务飞速发展非常重要的一年，主流电子商务企业纷纷抢占跨境份额。在当前人民币升值宏观背景下，“海淘”这项具备潜力的跨境电子商务势必持续走高，同时各大电商企业将面临巨大的竞争压力[3]。

再次，自身具备电商优势：曾经退出中国境内市场的电商巨头E公司在中国的“海淘”市场中占据重要市场份额。E集团的出口业务中，中国地区的“海淘”市场规模以及业务增长率远远高于其他国家，基于上述较为有利的国际宏观背景，中国良好的政策支持以及国内广阔的“海淘市场环境”，结合该企业自身的情况，E公司投入大量资金大力发展中国的“海淘”市场。E公司采取各种战略，包括与中国境内的成型电商网站合作，结成战略伙伴关系来扩大中国市场交易额等方案。

但是，E公司在发展中国“海淘”市场时遇到了各种实际问题，如下图1.3所示E公司在中国的业务发展受限存在以下几方面的原因：

首先，与中国电商企业合作，主控性差；合作过程中，由于中国合作伙伴更多的关注自身的利益，加之跨境电商中，中国消费者较多的购买衣服等存在尺码，颜色等问题的产品，退货问题严重；由于E公司在合作中不占有主动权，无法左右其经营策略，同时给了中国合作伙伴较高的优惠政策，却没能带来良好的市场表现，导致E公司在中国的市场扩展计划以损失千万美金而告终。

其次，中国消费者自主“海淘”存在较多障碍；包括：语言问题，交流时效性，风险控制等。英语是“海淘”必备的技能，对英语的不熟悉导致中国消费者无法直接通过E公司的美国网站进行购买；同时与国外卖家的交流工具主要是电子邮件，缺乏实时性。此外对于海淘的各种环节，不甚熟悉。如果借助第三方网站或代购又存在着较大的信用风险，这些因素都极大程度的降低了消费者的购买热情。

最后，跨境贸易中存在各种问题；包括：卖家发货周期长，物流时间长，物流方案选择，海关申报，关税等一系列问题，“海淘”服务属于一种各个环节都独立的状态。只有消费者对于各个环节需要都熟悉才能较好的驾驭“海淘”这种购物方式。同时，中国消费者比较注重“价格”之中除了质量之外的重要品质，E公司在中国境内投放的广告有限，导致中国消费者对于E公司已经慢慢的淡忘，对“海淘”中国消费者多数想到亚马逊，而并非E公司，这种市场相关的品牌选出也是影响市场份额的重要因素。如图1.3所示，将自主“海淘”模式存在的问题进行了汇总分析。

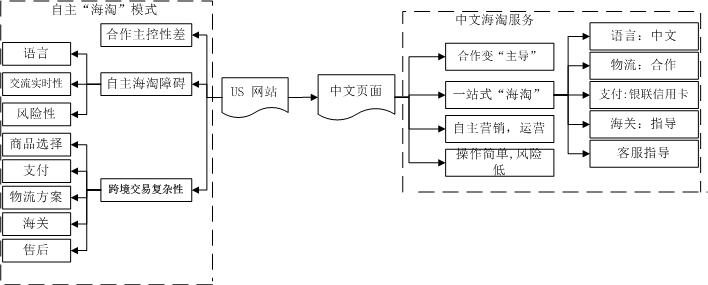


图1.3 中文“海淘”背景介绍

基于上述各种背景，E集团的“中国市场扩展”部大胆的提出中文“海淘”这个市场扩展计划，意图打造一个自主设计运营的一体化的“海淘”服务，在整个“海淘”环节中与中国的企业进行合作，提供整套解决方案。为解决语言问题，自建中文“海淘”页面与其英文网站进行连接，并提供“海淘”整套流程中文指导。针对海淘的各个环节与中国合作伙伴进行合作：(1)流量提供网站：海淘论坛，海淘返利网，海淘优惠券投放等合作伙伴合作。作为其流量的来源，并提供较多的佳品展示机会，优惠券信息等，用来吸引消费者前来购买；(2)银行：由于“海淘”交易中需要用美元等外币作为支付币种，所以E公司除了自身的第三方支付方式之外，与中国境内银行合作，使得消费者可以用带有银联标示的信用卡直接进行支付，为支付环节带来了便利；(3)指定优秀卖家，保证商品质量以及发货速度：对于海淘页面上展示的产品均是市场部通过对历史数据的分析筛选出的优秀大卖家，并与之进行沟通。可以充分保证跨境交易中中国卖家的发货速度和商品的质量，免去消费者与卖家沟通的麻烦与时滞。(4)多种转运方案：为消费者准备了快速直邮、经济型转运服务两种选择 ，提供便捷的物流服务；(5)海关申报，关税指导：对于各种类别的物品不同的海关申报指导；(6)依托强大的E集团客服，为中国消费者提供“海淘”各个环节指导以及问题解决。

因此，E公司为了进一步扩展其中国市场交易额，提高中国“海淘”市场份额，自建中文“海淘”页面与其主体英文网站相链接，并提供一站式海淘服务，如图1.4。解决了语言障碍，降低了消费者在E公司网站上进行“海淘”的门槛。如何获得精准的“海淘”市场需求，识别出“海淘”热销产品用于E公司中文页面的展示，可以较大程度的提高E公司的中国地区市场交易规模。为此，如何获得精准的“海淘”热销产品成为本本文研究的目标。

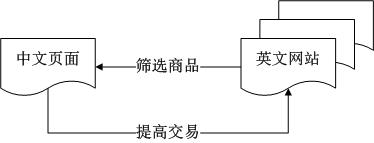


图1.4 项目整体背景介绍

中文页面可以极大程度上的解决语言，商品筛选，支付，物流，关税等海淘流程中存在的瓶颈问题。如上图所示，中文页面通过对英文网站上繁多的商品进行筛选，识别出热销产品进行展示，其交易将正向作用于英文网站的交易额，对于提高E集团的中国市场交易额，进一步提高“海淘”市场份额具有重要的作用。因此，如何精准的识别“海淘”市场热销商品成为本文研究的主要内容。

此外，“海淘”这种跨境电商市场火速发展，如何把握消费者购买需求和市场走向是电商企业十分关注的。“海淘”相关的论坛中消费者通过发帖回帖交流购物心得，宝贝分享等，此种帖子之中驻留了大量的信息。通过对帖子内容进行挖掘分析可以获得消费者在“海淘”领域喜爱的产品，讨论较多的品牌信息等；通过上述分析可以找出“海淘”热销产品信息，将上述信息用于“海淘”市场扩展计划。SMZDM（什么值得买）“海淘”论坛较为典型，帖子内容中存在大量的海淘热销产品信息，将作为本章的分析对象，用于获取“海淘”热销商品信息。

近年来，电商企业广泛的对用户在其网站上的评论进行分析挖掘，获得提升服务以及产品质量的重要依据。同时，评论中挖掘的信息用于辅助市场营销。上述评论信息大多限于企业内部数据仓库中存储的评论信息，企业外的论坛中，更多消费者进行经验交流，产品评价，信息推荐等活动。因此，外部“海淘”论坛中的评论信息对于企业的挖掘“海淘”市场需求具备重要的意义。此外，什么值得买论坛在海淘论坛中较为活跃，属于优质论坛。用户数量众多，帖子内容丰富，较多的评价了海淘中的商品信息。因此对于上述论坛内容进行分析将极大可能获得“海淘”热销商品的相关信息。同时，该论坛的数据结构简单，页面涉及较为简单，易于被爬虫抓取，并进行解析。

技术层面，较为成熟。爬虫技术SCRAPY成熟，可以节省用户的开发时间，只需要进行各项参数设置便可以轻松的采集论坛帖子内容进行分析。利用文本挖掘方法可以对上述信息进行进一步的挖掘分析。企业内部的数据仓库TERADATA内信息可以对外部文本挖掘提供各项字典信息，用于文本挖掘的词频统计环节。包括：商城信息，产品品牌信息，类别信息等等各类数据字典。基于企业内部信息优势以及外部论坛的重要作用，综合考虑论坛质量，最终选定什么值得买论坛作为分析对象，采用SCRAPY框架采集数据，结巴分词进行中文分词，PYTHON进行文本预处理以及文本挖掘。

综上所述，针对上述研究背景本文基于商业背景，依据企业的现实需求，进行识别海淘热销商品信息的分析与设计并最终付诸实践，分别包括对内部数据仓库的数据挖掘以及对外部“海淘”论坛的文本挖掘。其研究成果直接涉及当今发展最迅速的商业领域-电子商务。不仅具有重要的商业意义，同时有助于电子商务企业获得精准的“海淘”市场需求，对于进一步提高市场份额具备重大作用，并具备对电子商务行业的运营决策提供指导作用。

## 1.2 研究目标与内容

基于上述项目背景，本文将识别中国消费者“海淘”市场热销产品方法的设计与实现作文本文的研究目标。包括基于内部数据仓库TEDADATA的信息识别方案以及外部“什么值得买”海淘论坛的信息识别方案的设计与实现。

第一：基于企业内部TERADATA（数据仓库）贮存的海量信息分析；首先，基于已购买信息：本文将基于E公司内部数据仓库TERADATA沉淀的大量数据中获得中国消费者在E公司网站的全部购买信息，建立基于“海淘”流程的挖掘规则，最终得出中国消费者在E公司网站“海淘”的热销商品信息。其次，基于关键词搜索信息：TERADATA中记录了用户的每一次关键词搜索的信息，搜索关键词可展示潜在购买需求。因此对于消费者搜索较多，但网站自身提供较少的商品，当关键词的搜索量与提供量之间存在较大差距时，则认为存在潜在市场需求。针对本情况，作者提出了结合经济学模型的供求关系理论，采取寻找供需缺口的方式从企业内部数据库获得潜在相关需求信息。

第二：基于“什么值得买”海淘论坛帖子内容，对消费者反馈的信息进行分析；国外商品从品质，种类，价格等方面较国内商品具备较大的优势。因此分享海淘心得，产品信息，优惠信息的网站热度升温。国内目前针对 “海淘”进行交流的平台主要有：什么值得买（SMZDM）,REBATESME, 海淘贝，LETSEBUY，海淘一线，55海淘网，洋码头等。海淘论坛上用户的发帖回帖内容，隐藏了大量信息，反映了用户对商品的购买情况，关注程度，对商品的满意程度，购买愿望，购买趋势，以及对于“海淘”服务流程的态度等。针对以上信息本文将基于文本识别方法，探索文本挖掘技术在“海淘”市场需求分析上的应用。“海淘”相关交流平台上的文本数据，进行处理、分析，以获得辅助市场决策的数据挖掘，最终获得中国“海淘”市场热销产品信息。基于本文的研究背景与研究目标，主要研究内容可以用下图1.5进行概括：

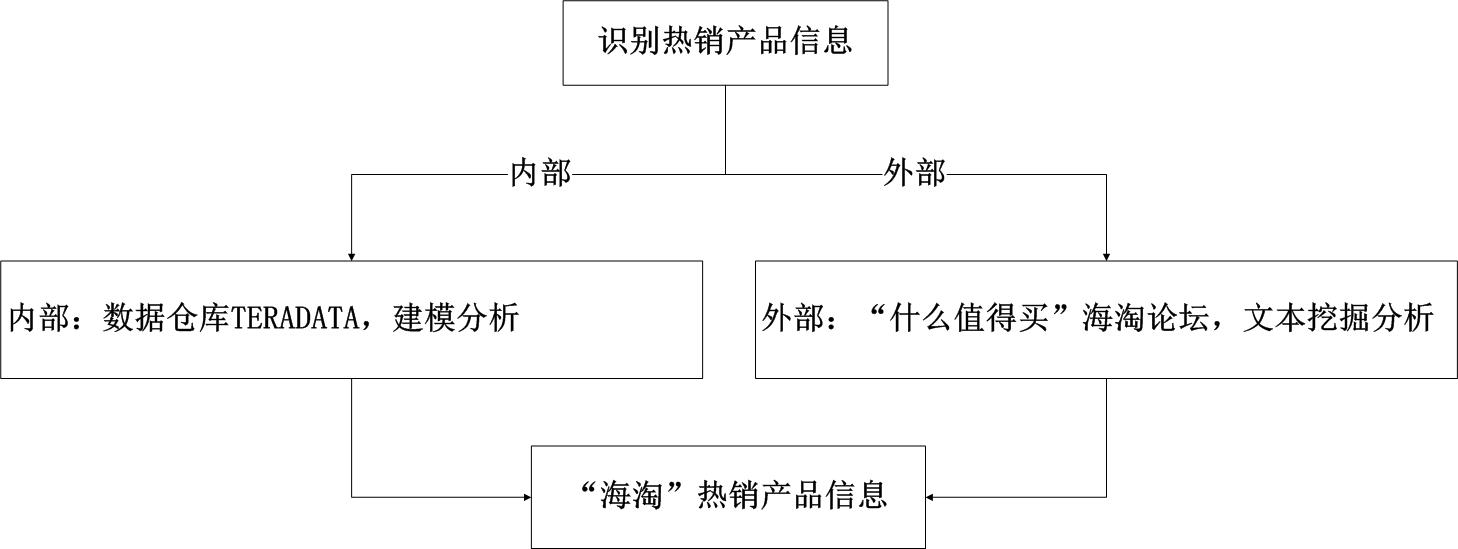


图1.5 热销产品识别总体方法

首先，基于企业内部数据，本文将研究并实现自动化的一套获取热销产品的流程。第一：信息获取模型设计；(1)基于企业内部数据仓库TERADATA存储的海量消费者购买信息，依据“海淘”整套流程为建模依据，获得历史热销产品信息；(2)基于消费者搜索信息，本文将设计针对关键词的“供需缺口模型”；对上述两个模型完成分析设计，如图1.6所示。

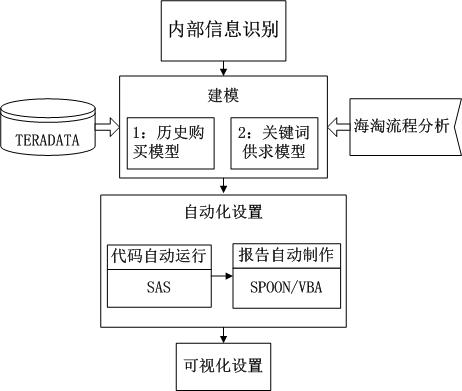


图1.6 内部信息识别流程

第二：代码以及报告自动化体系；模型设计完成后，应实际应用意义企业需要重复的进行模型更新以及报告的制作工作。为了提高效率与准确率，本文研究并实现了对于报告的整体自动化设置，包括模型代码自动化运行以及报告自动化制作发送。本文采取软件的结合使用（TERADATA+SAS+SPOON/VBA），编程实现完成数据的自动获取，报告的自动制作，并在指定的时间将相关信息发给指定的邮件列表。

第三：报告可视化；对于获得模型结果数据本文将采取可视化技术，将其通过Pivot Table，Excel图表展示，以及采取Tableau进行专业化的结果展示，发布到互联网，可以供企业通过互联网平台在线实时访问。

其次，基于“海淘”相关论坛内容，本文以SMZDM(什么值得买)为例进行数据获取。

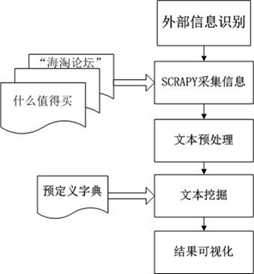


图1.7 外部信息识别流程

本文将搭建基于“网络爬虫SCRPAY+文本挖掘（PYTHON）+可视化（Excel/ PivotTable）”的一套挖掘信息的方法，完成其设计与实现。研究内容主要包括以下几点：(1)采用SCRAPY爬虫技术从SMZDM（什么值得买）获得文本数据；(2)对文本数据进行文本清洗、去噪声、去重、分词等预处理；(3)建立相应的文本数据库，利用已有字典进行词频统计；(4)基于文本数据库和统计方法进行“海淘”市场热销产品分析；(5)完成“海淘”热销产品信息的可视化过程。

## 1.3 研究方法

本文是基于E公司实际项目为基础，基于理论与实践相结合的研究方式。具体的研究方法包括：(1)方法设计与工程实践相结合。(2)定性分析与定量分析结合。(3)文献阅读与工程实践研究相结合。基于“海淘”流程各个环节进行数据挖掘分析，以期获得“海淘”热销产品的方法体系的设计与实现。基于“什么值得买”这一“海淘”论坛的研究方法采取理论与工程实践相结合，对整体的方案，操作流程，数据结构进行设计，对整套方法设计完成后，付诸工程实践，定量与定性分析，并在实践中不断的修正，最终获得整套可行且效果较好的方案。最后将两套方法体系所得到的结果进行融合分析，得出热销产品信息，基于上述方法体系本文的研究方法可以用图1.8表示。

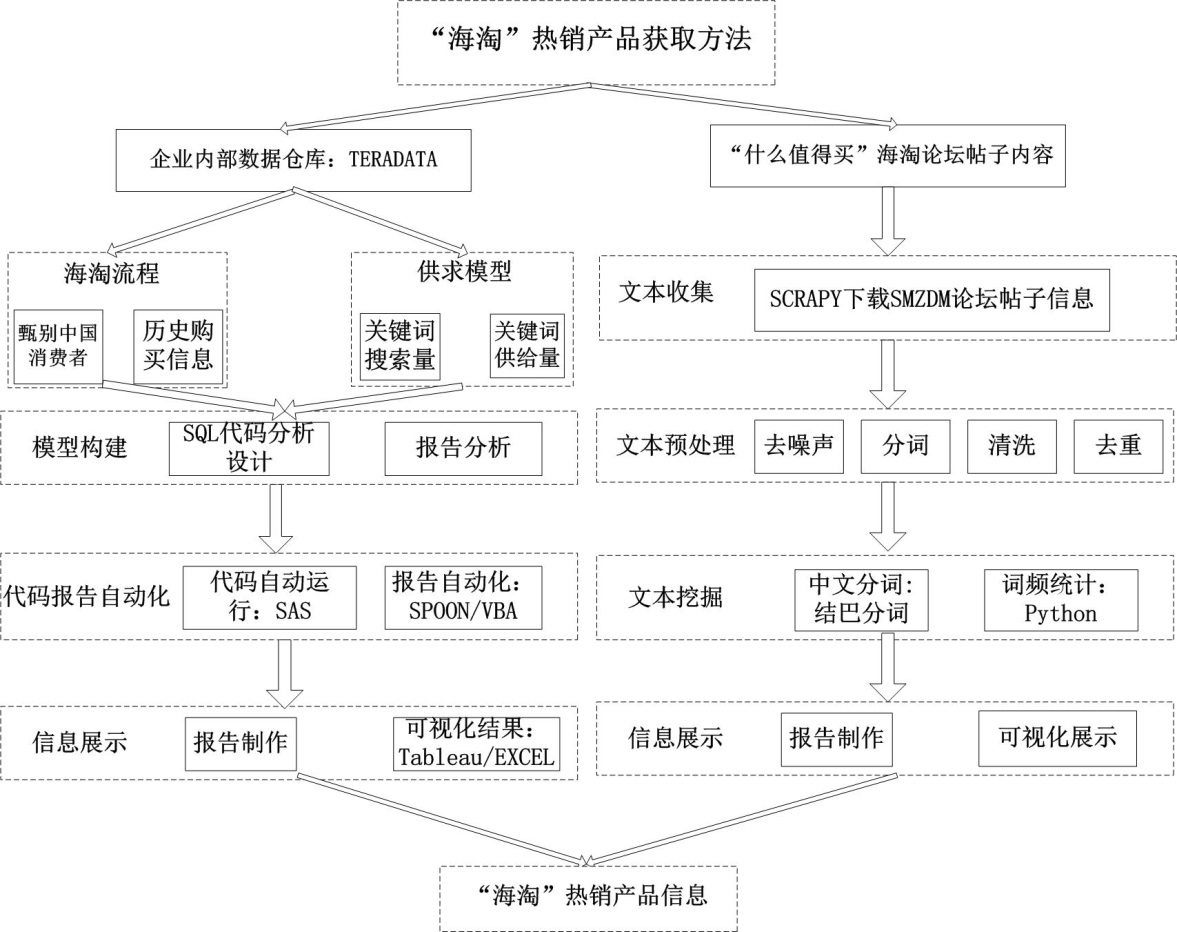


图1.8 研究方法

本文采取一个目标，两条并行途径的方法完成对于海淘热销产品的识别。内部采用数据挖掘方法，外部采用文本挖掘方法。图1.8中左侧部分详细介绍了数据挖掘的建模思想，模型构建，代码及报告自动化，信息展示四个环节。右侧部分详细的讲述了外部论坛文本挖掘的流程包括文本采集，文本预处理，文本挖掘，信息展示的过程，本图完成了对本文工作的高度总结。

## 1.4 论文主要贡献

本文完成了对于“海淘”热销产品信息识别的整套方案分析与设计；包括：基于内部数据仓库两类模型分析建设，基于外部论坛文本挖掘方案的设计与实施；最后综合两类结果获得了“海淘”热销产品信息。本文的研究对于E集团提高“海淘”市场占有率具备重大意义，对于其他电子商务企业具备重要的借鉴意义。

本文完成了整套获取“海淘”热销产品的方法，包括理论知识、方法体系以及具体的项目实践过程。内部方案的设计：建立“基于海淘流程的历史购买数据分析模型”以及“基于关键词搜索的供需缺口模型”，实现模型代码更新运行以及报告自动化的过程；外部方案设计：实现了SCRAPY抓取“什么值得买”论坛文本信息，采用PYTHON对外部论坛获得的文本数据进行挖掘，制作成报告并应用到企业的实际商业运营之中

## 1.5 论文框架结构

根据企业的实际业务需求与具体的实践过程，本文将围绕一个主题，两套方法共同完成企业所需的“海淘”热销产品信息的获取。采取业务梳理，分析思路，方法设计，工程实践相结合的方法进行本文的撰写，具体内容如下图1.9所示。基于上述框架本文的篇章组织结构如下：

第一章、绪论：简要介绍了“海淘”热销商品识别方法的背景信息、本文的研究意义、研究方法、研究过程，及本文的主要研究工作和论文结构简介。

第二章、研究现状：介绍了国内外在电子商务领域研究情况，爬虫应用情况，文本挖掘的研究现状，以及目前对电子商务市场需求的研究现状。采取的主要方法体系，相关技术发展状况。

第三章、相关技术介绍：详细介绍了项目中采用的各项技术包括数据仓库，自动化技术，爬虫技术，文本挖掘技术，可视化技术。

第四章、企业内部“海淘”热销产品信息识别方案设计：介绍了基于企业内部数据仓库TERADATA，建立“历史购买模型”以及“潜在需求模型”。介绍了模型背景、模型设计逻辑、代码设计。综合两类模型分析得出“海淘”热销产品信息。

第五章、企业外部“海淘”热销产品信息识别方案设计：介绍了基于企业外部“什么值得买”“海淘”论坛，采取爬虫、文本挖掘等方法识别出“海淘”热销商品信息的方案设计与实现。

第六章、总结与展望：总结贡献、创新点与不足之处并展望了未来的研究方向。



图1.9 本文结构

# 第二章 研究现状

## 2.1 跨境电商研究现状

跨境电商作为电子商务的重要领域被学者较多的研究。目前，国内学者对于跨境电商的研究主要集中在物流领域，信用，跨境电商发展模式，策略，营销，经营发展战略，金融，数据挖掘，电商对对外贸易影响，人才培养，出口退税，海关监管，法律，品牌管理，针对某一领域的跨境电商构建等方面对跨境电商进行研究。

典型代表为：上海社会科学院的穆承刚硕士论文研究“我国小额跨境电子商务模式研究”，对跨境电商进行了模式特征分析与模式归类，并对比了四类模式-自营商，平台式，综合服务商，企业应用式。并预测了我国跨境电商的发展趋势-综合模式是未来发展趋势[6]。浙江大学单丹丹硕士论文利用仿真软件Swarm模拟复杂的系统环境，进行电子商务信用研究，文中建立了动态博弈，静态博弈两种模型，得出来静态博弈时交易主体更倾向于选择失信，动态博弈交易中主体信用值不断提升，主体更倾向于守信。安徽大学杨璐硕士论文从理论角度对电商在国际贸易中的应用及发展现状进行研究[8]。

对于“海淘”这种新兴迅速发展的跨境电商的研究较少，这种新型的电子商务模式包括完整的链条，多个环节。包括国外电商网站购物，优惠信息使用，第三方支付，跨境物流（转运/直邮），海关申报，关税等环节。国内对于该领域的研究较少，且主要集中于理论研究层面，技术研究较少。对于热销产品信息这种营销方面的研究较为缺乏。

## 2.2 爬虫技术应用现状

爬虫技术主要用于搜索引擎，近年来被广泛的应用于信息采集，是文本挖掘领域必备的技术。目前研究方向包括与算法的结合应用，面向主题的搜索爬虫分析设计。

爬虫技术在电子商务领域的应用主要是基于电子商务网站信息采集，典型代表为湖南大学杨颂硕士论文 [9]本文中针对电子商务“比价网”的采集电子商务网站增量信息的增量爬虫系统“Heritrix”进行了分析与设计。北京交通大学王静硕士论文中采用MHRW算法进行抽样，SCRAPY进行数据采集，从在线社会网络视角分析买家，商品，卖家在“淘宝分享平台”上的关系。讨论三者构成的网络特征，得出了买家分享行为相似性对于买家以后的分享行为的影响。

爬虫技术在论坛数据分析领域，主要用于论坛数据采集。基于采集到的数据进行研究主要包括用户分类，获取热点话题，热点新闻等。典型代表：哈尔滨工业大学胡亚楠硕士论文“社交网络数据获取技术与是实现”[10]设计并实现了针对Twitter，Facebook，“人人网”等论坛的爬虫，采集Deep Web数据进行聚类分析，得到社交网络上的热点新闻信息以及热点话题信息，并采用FLASH+SILVER LIGHT等技术进行可视化展示。北京邮电大学陈希的硕士论文 [11]采用基于HttpClient爬虫技术采集“人人网”的用户数据并存入数据库之中，利用社交网络理论以及客户细分理论进行研究，采取决策树，聚类等数据挖掘方法进行客户细分，为识别细分市场提供差异化服务奠定基础。

此外，综合搜索引擎如百度，谷歌等无法满足专业领域工作人员的搜索准确率需求，主题搜索引擎面向主题的爬虫诞生。近年来相关研究较多，包括面向垂直搜索引擎的爬虫，面向特定行业的主题爬虫，面向特定领域的垂直搜索引擎爬虫。典型代表为：武汉理工大学戴支荣硕士论文 [12]讲述了数据网页爬虫，页面解析技术，并设计了针对气象工作人员的主题搜索引擎，记性个性化设置推荐与预览等功能。复旦大学王烨硕士论文“基于广度优先的主题爬虫的设计与实现”[13]解决了广度优先算法构建主题搜索引擎的各种问题，包括搜索方案，关联度计算，URL消重，WEB页面存储等，提出了对搜索规模和准确率进行了均衡处理的搜索引擎方案。

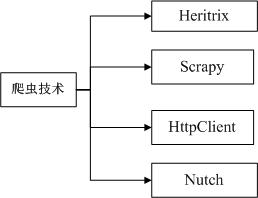


图2.1 爬虫技术总结

总之，对于爬虫的研究集中在搜索引擎层面以及电子商务，社交网络等研究的信息采集层面。目前较为成熟的爬虫方法包括：JAVA开发的基于Hadoop分布式爬虫Nutch，基于JAVA开发的Heritrix，基于PYTHON开发的SCRAPY框架，Beautiful Soup，mechanize，selenium，cola等爬虫框架。目前最为常用的为Nutch，Heritrix，SCRAPY框架。

## 2.3 文本挖掘研究现状

文本挖掘是数据挖掘的分支，常常被认为是知识发现TDT的重要环节。在电子商务领域，文本挖掘主要研究方向为个性化推荐，推荐算法，信用评价，评论情感分析，评价挖掘等领域的研究；

此外，文本挖掘主要应用领域包括：搜索引擎优化，潜在语义分析，Web挖掘，SNS热点话题，分类系统实现，互联网新闻研究，与CRM系统的结合使用，电子商务评论因素研究，网络舆情热点分析，量化投资领域，预测分析，企业竞争情报分析，物流领域应用，社交网络领域，分类、聚类系统的设计与实现；金融财经领域的预测分析；医学领域，生物领域中文本挖掘的应用也较多。

国内的研究典型例子包括：清华大学何叶青在其硕士论文“网上交易文字评论的卖家因素分析—基于淘宝网实证研究”[14]该文章采用文本挖掘方法将用户评论进行了数字量化处理，建立模型进行实证研究，得出了不同类型的买家对评论采取的态度以及做法，最终得出对买家，与卖家的政策建议。此文代表文本挖掘在量化分析领域的应用；哈尔滨工业大学徐东亮硕士论文 [15]中主要对比分析了聚类方法K-Means，K-melodies算法，对于论坛帖子内容相似度进行计算，得出聚类结果，进而得出话题热度评分结果。此文代表文本挖掘领域不同算法在实际项目中的应用效果分析；中国科学技术大学代其峰在其硕士论文 [16]中提出了SPS（相似产品单元算法）对电子商务现有类目体系构建算法进行优化，并建立了基于SPS算法的个性化推荐系统，解决类目过粗，产品分类不精问题，同时解决了长尾问题，存在的稀疏性问题，最终提高了推荐的精确度。

总体而言，文本挖掘作为数据挖掘的重要分支，需要进行文本分词，预处理，特征提取等操作，文本挖掘的方法与数据挖掘较为相似。数据挖掘的主要分类与聚类的方法可以用下面两个表格进行总结。

表2-1 数据挖掘分类

|  |  |
| --- | --- |
| 关联规则 | 对数据之间的相互关联进行挖掘 |
| 时序规则 | 数据之间的因果分析 |
| 分类分析 | 依据已有分类，采取数据挖掘方法建立分类模型，依据规则对数据进行合理的划分为不同集合 |
| 聚类分析 | 输入的数据无 类型标记，按照规则将数据划分为不同集合 |
| 预测分析 | 采取回归方法，利用历史数据预测未来数据 |
| 数据总结 | 对数据进行汇总分析 |
| 模式发现 | 序列模式发现，依赖关系，异常和趋势发现等 |

表2-2 数据挖掘基本方法

|  |  |
| --- | --- |
| 统计方法 | (1)回归分析（多元回归，自回归等），(2)判别分析（贝叶斯，非参数），(3)聚类分析（平面划分聚类，层次聚类，密度聚类，系统聚类，动态聚类），(4)“探索”分析（主成分分析，相关性分析） |
| 神经网络方法 | (1)前向算法（BP算法），(2)“自组织”神经网络（“自组(3)织”特征映像，竞争学习） |
| 机器学习方法 | (1)“归纳 学习”法（规则归纳，决策树），(2)基于范例的学习方法，(3)遗传算法 |
| 数据仓库方法 | (1)多维分析，(2)联机分析，(3)属性归纳方法，(4)数据分析，(5)建模 |

文本挖掘理论研究领域主要对各种算法进行改进，对文本挖掘的各个环节进行优化处理，将文本挖掘应用在电子商务，金融，财经，医药，生物，互联网等各个领域之中，解决信息爆炸问题。

## 2.4 电商市场需求研究现状

目前对于市场需求方面的研究主要包含以下几个方面：物流方面预测，基于互联网搜索量对某种商品进行预测，大部分研究针对个体的个性化推荐以及算法的研究，以及针对供应链的研究，针对销售额的预测。

物流方面预测的方法主要包括：定性分析，回归分析，时间序列分析，趋势外推方法。基于互联网搜索量的预测中，北京邮电大学刘佳霖[17]以百度指数与“移动飞信”用户数量为基础，采取时间序列分析模型，引入了百度指数与移动用户数。探索了电影票房这种商品的互联网搜索量与市场需求之间的关系。

针对互联网搜索量与市场需求的研究目前国外学者研究较多。Hyunyong Choi, Hal Varian 2009年采用Google Trends关于汽车的搜索记录，包括福特汽车，雪佛兰，日本丰田等为预测对象，建立模型发现上述商品销售量与其关键词在互联网上的搜索量具有重要的正相关性[18]。Simeon Vosen, Torsten Schmidt在2011年研究互联网搜索记录与美国私人消费之间的关系，发现当引入互联网搜索量与CCI，MCSI三个变量进入需求模型。结果显示，新变量是模型的准确程度有所提升，互联网搜索量进入模型后的模型的准确率提升程度最大[19]。Nicholas Lincoln在2011年研究iphone，iPad的销售量与internet搜索量（google trends上的搜索量数据为数据来源），广告投入的关系，得出结论Iphone的销售与其互联网搜索量有较大关系，与广告投入无关。IPad销售则与因素都无关[20]。Gauri Kulkarni, P.K.Kannan,Wendy Moe在2012年研究了互联网搜索量对于电影票房的预测模型。对电影发布当周的票房以及后续票房利用模型进行预测，结果显示：Internet搜索量的引入可以较大程度提高预测精度[21]。

针对个体的个性化推荐方法的研究，推荐策略推荐算法的选择与推荐系统的构建是目前国内学者研究的主要方向。目前推荐算法的类型包括基于规则，效用，知识，内容的推荐以及协同过滤推荐[23]。其中,复旦大学吴海峰在其硕士论文中提出采用层次分析法完成多Agent的物流服务推荐系统设计与实施[22]。对于供应链的研究逐渐从供应链SC向需求链DC转变，之所以做出上述转变主要原因是：SCM对中心要就较高，由于SC的计划复杂性以及协调成本高SC重构困难，所以目前偏重于SC向DC转变[24]。

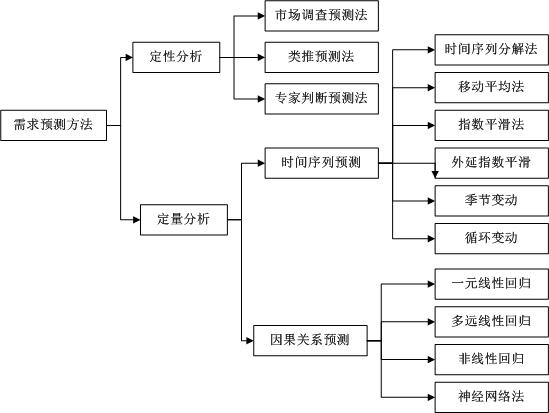


图2.2 需求预测方法

目前，对于销量的预测的研究主要方法包括：定性分析与定量分析。具体方法由哈尔滨工业大学樊营营在其硕士论文中进行了总结[25]。上图中可见，目前需求预测的方法中“统计方法”较多，神经网络发作为机器学习的方法正逐渐被学者使用。实验数据显示，神经网络法以其灵活性以及超强的处理大数据的能力正在被更多的采用。

国外研究学者对于预测的研究主要采取ARIMA，Box-Jenkins，Winter指数平滑等，采用销售数据本身建模对未来进行数据预测,ARIMA对于短期预测很有效果[26][27]。对于神经网络的研究表明：神经网络比统计方法更能体现季节性特征以及相关趋势，能够更好的进行销售预测[28][29]。B.Formby在论文中提出Stable Seasonal Pattern稳定性波动模式在销售预测中的实际应用[30]，上海交通大学者将上述方法与SVM结合使用对销售额进行了预测。

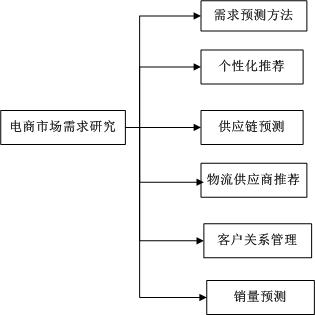


图2.3 电商市场需求研究总结

Sebastian Klapdor，Eva M.Anderl(2014)等人在Finding the wright words[33]中总结了在电子商务付费搜索领域中关键词的宽度以及关键词描述的准确性对交易的影响。Qihui Lu，Nan Liu(2013)[31]采取斯塔克伯格模型，纳什均衡模型分析对比了不同的定价策略对市场需求以及自身利润的影响。Kaiquan Xu, Stephen S.Y. Liao（2012）研究了社交网络用户对于企业以及产品的评价对于企业的声誉以及商业表现的影响，进一步提出了企业利用社交网络作为维护自身形象，开展营销工作的方法及建议 [32] 。

综上可知目前学术界对于电商市场需求的研究主要针对算法的研究与完善，不同算法组合实践等，对于实际市场需求的产品信息研究较少。具体应用如上图所示，主要集中在需求方法的分析设计，包括统计方法，数据挖掘类方法；个性化推荐，主要针对个性化算法的创新以及应用领域的创新为主；供应链以及物流方面的推荐与预测，客户关系管理，销量预测等方面的研究主要集中于方法的选择领域，通过对比各种方法的效率，最终决定方法的选择。本文总结上述思想，进行电商市场需求分析研究，完成对电商企业实际业务逻辑的梳理，基于实际业务流程的完成数据挖掘模型与方法的设计，并最终获得运用在实际业务中的电子商务热销产品信息。

# 第三章 主要技术介绍

## 3.1 数据仓库TERADATA

TERADATA介绍：TERADATA公司成立于1976年，进军OLAP即联机分析处理这一市场；1984年，TEARDATA退出来MPP（基于海量并行处理技术）的商用系统，MPP基于Data Base Computer的硬件平台，TERADATA OS操作系统；1992年被AT&T公司收购，后与NCR合并，1996年NCR推出TERADATA的关系型数据库管理系统RDBMS，此RDBMS基于UNIX平台，此后1998年，该公司再度推出基于Windows NT平台的RDBMS[34]；

“作为关系型数据库管理系统的TERADATA，可以把企业内部的交易信息经过处理存储到关系型数据库之中；TERADATA采取结构化查询语言SQL进行数据获取，其复杂的内部结构适用于处理复杂的数据仓库查询，数据处理相关的应用；同时灵活多变的SQL语言适用于创建复杂的数据挖掘模型，简单易懂。TERADATA凭借其良好的扩展性，处理能力从GB级别提升到了100TB以上[35]”。从功能上分，数据仓库TERADATA的框架可以分为以下四部分：

第一：数据装载功能；实现将企业的交易数据转换为数据仓库内的关系型数据的过程，涉及到数据提取Extraction，数据转换Transformation，数据加载Loading过程，即ETL；对操作数据以及“源数据”经过抽取，过滤，清洗，转换等操作后加载到TERADATA数据仓库中。此种转换包括逻辑部分转换以及物物理部分转换：逻辑部分转换需要在数据仓库中建立相关逻辑模型，逻辑建模的目标为解决企业业务中的问题，TERADATA的出发点为解决企业的业务问题；物理部分转换指基于逻辑模型而形成的物理模型；后通过数据抽取，转换，加载(ETL)过程，将数据加载到物理表之中。

第二：数据管理功能；数据管理是TERADATA数据仓库的核心功能，数据仓库需要具有优秀并行处理性能的RDBMS关系数据库管理系统、以及扩展能力强的服务器。NCR公司恰好具备了上述条件，它拥有TERADATA数据库管理系统作为其引擎，以及专门配套数据仓库设计的并行处理服务器，其特点是灵活的扩展能力，可以依据要求增加机柜变成处理能力更强的存储系统。

第三：信息访问功能；TERADATA数据仓库为用户提供界面化，可交互处理的前端工具，方便用户完成对数据仓库中的数据访问，编程和灵活处理使用。用户可以使用SQL语言在信息采集界面进行数据获取逻辑的编辑，数据挖掘，数据预处理等相关的工作。同时，TERADATA数据仓库支持企业级的业务级别，全球的分析师可以在不同的地方灵活高效的听过客户界面完成对数据仓库的访问，进行数据挖掘，数据分析。

第四：数据仓库支持基础功能；TERADATA数据仓库将前三个功能有机的结合，构成了完整的解决方案。包括加载元数据服务，data mining 数据仓库建模，Data Warehouse与网络管理，数据字典，专业技术咨询服务等工作[34][35]，如图3.1所示。

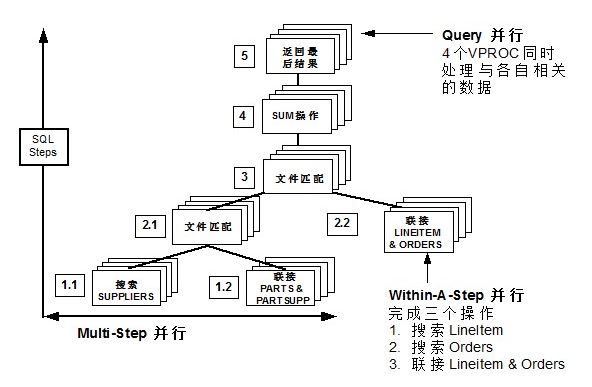


图3.1 TERADATA 数据仓库并行处理机制

TERADATA具备强大的并行处理能力，称为多维并行处理机制，这项功能使得TERADATA具备了强大的数据管理能力，TERADATA将巨大的数据量拆分成中国的小任务，以此可以处理TB级的数据。包括“查询并行（Query），步内并行（Within a step），多步并行(Multi step)三大特点” [36]。查询并行：基于HASH数据分配机制完成，每个Access Module Processer都是一个虚拟处理器VPROC，VPROC之间独立的负责各自的任务，如完成表的选择，内外全链接，聚集，索引检索，投影，排序等相关操作都是有VPROC独立的负责； 步内（within a step）并行：对于复杂的Query，首先TERADATA会对其进行优化分解，得到一下小步骤，每一个小步骤都会被发给一个独立的VPROC处理。对于复杂的关系运算，一个VPROC将启动多个进程同时处理，成为within a step处理。多步并行，“步内并行”中的步骤同时进行称为多步并行。

TERADATA具备分布式的数据检索机制如下图3.2所示：

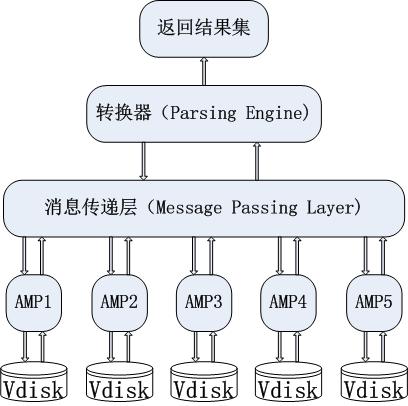


图3.2 TERADATA 数据仓库检索机制

TERADATA SQL：SQL（结构化查询语言）是一种非常接近英文的查询语言，非过程的编程语言。目前已然成为所有的关系数据库的查询语言。TERADATA支持标准的ANSI SQL，并在此基础上作了部分扩展，这和大部分商用数据库是一样的。SQL是一种基于集合概念进行操作的编程语言，以集合的方式输出结果。

表3-1 数据库定义语言 Data Definition Language(DDL)

|  |  |
| --- | --- |
| SQL语句 | 功能 |
| CREATE | (1)定义新的表、索引、视图、宏、触发器和存储过程 |
| DROP | (2)删除表、索引、视图、宏、触发器和存储过程 |
| ATLTER | (3)表结构与保护机制的调整 |

表3-2 数据库操作语言Data Manipulation Language (DML)

|  |  |
| --- | --- |
| SQL | 语句功能 |
| SELECT | 执行关系查询操作(1)选择、(2)投影、(3)连接、(4)合并、(5)交集） |
| INSERT | 向数据库的表中插入一条新记录 |
| UPDATE | 修改数据库的表中记录的值 |
| DELETE | 删除数据库的表中指定的记录 |

表3-3 数据库控制语言Data Definition Language(DDL)

|  |  |
| --- | --- |
| SQL | 语句功能 |
| GRANT | 给用户授予某种权限 |
| REVOKE | 删除用户某种权限 |
| GIVE | 转移用户所有权关系 |

资料来源：TERADATA SQL基础教程

此外TERADATA具备联机帮助功能。TERADATA的帮助系统主要由三条命令组成，分别是(1)HELP，(2)SHOW，(3)EXPLAIN。其中，HELP命令可以提供有关数据库中各种目标的信息；此外，SHOW命令用来显示这些目标的结构，包括创建目标的DDL语句；最后，EXPLAIN命令以英文方式显示系统处理SQL交易请求的执行过程；SQL采用与英语相似的语法结构，从目标出发编写代码。常用的语法包括以下几个方面：

(1)逻辑与条件表达,BETWEEN…AND,IN,LIKE,CASE,AND/OR,NOT；

(2)子查询，EXISTS，IN；

(3)属性与函数，CHARACTERS函数，TRIM函数，FORMAT格式；

(4)表连接，左、右、全连接，自连接，交叉连接，子查询与表连接；

(5)分组与聚合，GROUP BY,WHERE,ORDER BY ,HIVING；

(6)总结与小计，WITH BY, WITH,DISTINCT；

(7)OLAP 相关函数，移动平均，移动查分，累计，排队汇总等；

## 3.2 自动化过程

电子商务企业的商务部门，数据分析部门日常工作中设计到数目惊人的商业报告，这些商业报告发送的准时性与准确性要求较高，且大多数报告制作过程繁琐，如果人工制作，则浪费了了巨大的人力物力成本。故在实际工作过程中企业会有各种Automation Report（报告自动化）的需求以及实现方式。E公司内部的大部分报告都需要进行自动化设置，便于完成从数据仓库内表的更新到数据仓库数据的抽取，加载，转换，以及报告的制作全部过程的完成军采用软件系统组合的方式完成自动化运行的过程。E公司主要采取三种Automation（自动化）的方法，包括SAS Automation，Spoon Automation，VBA Automation三种方式的组合使用。

SAS Automation；

SAS（Statistics Analysis System）是美国北卡罗来纳大学的两名研究生编制，1976年成立SAS研究所。该软件早期致力于统计分析，目前SAS是决策支持领域的Information System大型集成信息系统。SAS具备强大的数据统计分析功能，可以高速高效的处理海量的数据。“SAS主要完成的任务包括：数据访问(Data Access)、数据管理(Data Management)，数据呈现(Data Presentation)，数据分析(Data Analysis)” [37]。本文主要应用SAS的自动化驱动功能，采取SAS与TERADATA数据仓库相连。SAS Automation整个过程需要的软件包括:WinSCP, Putty, SAS Account，A folder in SAS。

简要步骤如下：(1)启动PUTTY进行相关设置，登录SAS。(2)在SAS file中为每个任务建立独立文件夹，建立同名的SQL文件，KSH文件；(3)对KSH进行一系列配置，SAS运行命后可以利用KSH文件驱动与TERADATA数据仓库链接，并在指定的时间，按照指定的频率完成指定SQL代码的运行，完成SQL代码对于数据仓库内数据完成的一系列操作。(4)最后将SQL运行的成功与失败的结果通过电子邮件发送给指定的邮件接收者。

SAS自动运行代码这一功能为企业带来了很多便利，节省了企业相关工作人员的时间，提高了工作效率，降低了工作中出错，遗漏的情况发送频率。SAS做好了数据仓库中相关表的更新操作，为下一步数据抽取，进行模型分析，报告制作打下了坚实的基础。

KETTLE Automation；

KETTLE是“KETTLE E.T.L. Environment”的简写，它被设计用来帮用户实现ETL需要：抽取，转换，装入和加载数据：中文名字 “水壶”代表程序员MATT“希望把各种数据放到一个壶里，然后以一种指定的格式流出。”Spoon是KETTLE的一个图形用户界面的框架工具，它允许用户运行或者转换任务。自动化数据抽取转换，装入加载数据的工具。本文中将采取Spoon与数据仓库相连，在指定的自动的完成相关操作，自动的从数据仓库抽取数据，对数据进行操作，导入到EXCEL中，并对EXCEL中的数据进行处理，最终将已做好的报告发送给指定的用户。Spoon将SAS更新的数据抽取，加载，后转入到指定的EXCEL中，同时对EXCEL中的数据进行处理，制作出符合要求的商业报告，并发送。

VBA Automation：

微软自带可编程软件VBA为公司的自动化报告工作贡献了很多力量。可通过建立VBA工程，进行编程操作，实现自动化报告制作过程中的数据获取、转换、加载工作，并将获得的数据加载到指定的EXCEL中，之后完成对数据的自动处理，并在指定的时间开启操作，并将最终的结果发送给指定的用户。VBA Automation可以完成的工作与Spoon一致，实际工作中会将两者同时待命，当主题的Spoon发生故障时，会启动备份的VBA进行替代工作，保障自动化报告制作发送工作的顺利进行。

## 3.3 网络爬虫原理

### 3.3.1 工作原理及关键技术

网络爬虫以自动采集网页内容为目标，从初始的采集初始URL网页开始，递归的获取网页内容，最终搜集整个网站的信息，直到满足事先设定的条件为止。网络爬虫的工作原理如下：第一步，根据网页搜索算法，不断的从当前页面上获得新的URL放入队列。第二步，按照一定的搜索策略从队列中选择出下一步需要抓取的网页URL，重复上述过程，直到满足一定的条件为止 [38]。

### 3.3.2 网页搜索策略

爬虫工具的网页搜索策略：深度优先搜索，广度优先搜索，最佳优先三种策略。深度优先容易使得爬虫Trapped,目前较为流行的是广度优先搜索策略与最佳优先搜索策略。(1)广度优先：在获取URL时，先将得到的URL全部放入队列，完成当前层次的搜索后，才继续进行下一层次的搜索。此种算法涉及URL与主题相关性的判断，对相关度低的网页进行过滤，剪枝。缺点在于：可能导致数据冗余，大量下载无关网页并过滤，降低算法的效率。(2)深度优先：对于满足条件的URL将抓取到深层次，没有超链接或者满足搜索条件才会跳出，回到上一层次继续抓取URL。缺点：先来先服务，理论上可以获得完整的信息，但是往往存在网页嵌套太多，爬虫将陷入问题，故此策略并不常用。(3)最佳优先：按照某种网页分析算法对与目标网页主题相关的URL进行相似度，相关性的预判。如果相关则将该URL放入目标队列中，否则将该URL跳过。本策略执行过程中，得到一个URL后可采取深度优先或广度优先算法获得下一个URL。缺点：最佳优先算法是一种局部最优的策略，因此抓取路径上已经被筛选掉的网页，即使与当前主题相关的URL也会被忽略。

### 3.3.3 爬虫框架SCRAPY

SCRAPY是一个爬虫框架，基于PYTHON为了“爬取”网站数据，为提取结构性数据而编写的应用框架。可以应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。SCRAPY用途广泛，可以用于数据挖掘、监测和自动化测试等相关领域。SCRAPY 使用了 Twisted异步网络库来处理网络通讯。该框架最初是为完成页面抓取 ( 网络抓取 )所设计的， 也可以应用在获取API所返回的数据，或者通用的网络爬虫。整体架构如下图3.3所示，SCRAPY主要包括了以下组件：

(1)引擎（SCRAPY Engine）：处理整个爬虫系统数据流并触发事务。

(2)调度器(Scheduler)：接受引擎发的请求，将其压入队列之中，并在引擎再次请求的时候返回。

(3)下载器(Downloader)：下载指定URL的网页内容，将内容返回给蜘蛛。

(4)蜘蛛(Spiders)：制订特定域名或网页的解析规则。

(5)项目管道(Item Pipeline)：主要任务是清洗、验证和存储蜘蛛从网页中抽取的项目，数据。当页面被蜘蛛解析后，将被发送到项目管道，并经过几个特定的次序处理数据。

(6)下载器中间件(Downloader Middlewares)：位于SCRAPY引擎和下载器之间的钩子框架，主要用于处理SCRAPY引擎与下载器之间的请求request及响应response速度与方向。

(7)蜘蛛中间件(Spider Middlewares)：介于SCRAPY引擎和蜘蛛之间的钩子框架，主要工作是处理蜘蛛的响应输入和请求输出方向。

(8)调度中间件(Scheduler Middlewares)：介于SCRAPY引擎与调度器之间的中间件，控制从SCRAPY引擎发送到调度的请求和响应的方向。



图3.3 SCRAPY基本工作流程

使用SCRAPY可以很方便的完成网上数据的采集工作，提供了很强的功能使得scraping很易用,高效率。如：支持从HTML、XML源中选择并提取数据的功能；支持多种导出数据的方式,包括XML,CSV以及JSON三种格式。

Interactive Scraping Shell Console（交互式爬虫控制台）,对编写与调试爬虫均非常有用; Web Management Console（WEB控制台），监视和控制爬虫; Telnet Console(Telnet 控制台),在SCRAPY进程中可以不受限制地访问PYTHON Console控制台。用于调试爬虫工程;此外，一些内嵌的工具,用于记录分析Logging，统计数据搜集(Collecting Stats)，并发送电子邮件通知给指定的人(Sending Email Notifications )[38]。

## 3.4 文本挖掘

文本挖掘是指从大量文本数据中抽取出事先未知的、可以理解的、最终可用的知识的整个过程。同时运用这些知识可以更好地组织信息，供将来参考 [39]。文本挖掘是一个交叉学科领域。文本挖掘是从数据挖掘发展而来，因此其定义与数据挖掘定义类似，是数据挖掘的扩展，如果数据挖掘的研究对象是文本这种类型的数据时，便称作文本挖掘。与传统的数据挖掘相比，文本挖掘的特殊之处主要表现在：

处理对象不同，数据挖掘以数据库中的高度结构化数据为处理对象，用表结构关系进行知识发现；文本挖掘则以非结构化的文档为对象，或者是半结构化，良结构化的数据（图书馆数据库中的文档结构，包含标题、作者、出版日期这种结构化数据与摘要，正文这种非结构化数据）。文本数据如果具有较好结构的也可以用关系数据库进行存储实现 [39]。通常将文本挖掘通常作为知识发现的一个基本步骤，文本知识发现主要由以下步骤组成：

(1)文本采集：使用爬虫将网页信息抓取过来，建立文本仓库。(2)文本预处理：对获得的网页文档进行数据清洗，去噪声，去重等操作；将文本数据转化为文本挖掘工具可以进行操作的形式；(3)文本挖掘：在完成文本预处理后，可以利用统计学方法、机器学习方法、模式识别等方法提取面向特定应用目标的知识或模式；文本挖掘包含的技术：文本分类，文本聚类，文本关联分析，文本结构分析，文本摘要，分布分析，趋势预测等技术；(4)模式评估：利用已经定义好的评估指标对采集的知识或模式进行评价；如果符合文本挖掘的结果要求，就存储该模式以备用户后续使用；否则返回前面环节进行重新调整和改进，进行新的发现过程；(5)结果展示：将挖掘出的知识或模式以可视图形界面化形式展示给读者[40][41][42]。



图3.4 知识发现流程

文本挖掘的步骤包括：文本数据采集操作、文本挖掘预处理技术、分类、聚类、信息提取、信息提取的概率模型、预处理应用、可视化方法、链接分析、文本挖掘应用等[43][44]，如下图所示：

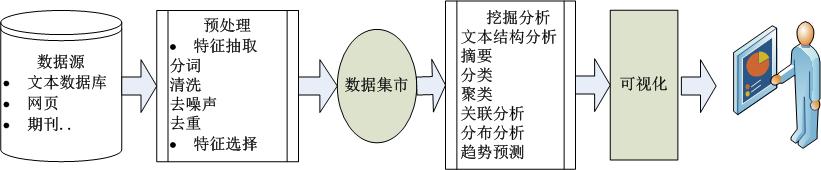


图3.5 文本挖掘流程

本文中主要完成文本挖掘的数据预处理，挖掘分析，可视化的过程。

### 3.4.1 文本预处理

文本预处理环节的包括：特征抽取与特征选择。

“特征”指从文本中分理出的可以用来表示文本的基本语言单位。通常文本特征抽取需要分中英文分别进行，具体步骤如下图所示。“英文文档集”需要进行消除停词，词干抽取工作获得特征词集合。“中文文档集”需要进行消除停词，词语切分等处理，最终获得特征词集合。

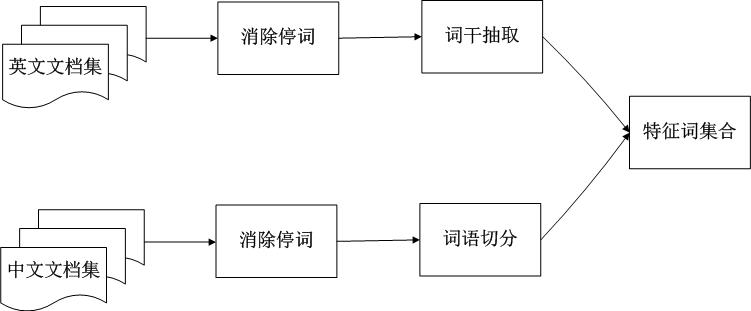


图3.6 文本特征提取过程

分词环节，本文采用基于PYTHON的结巴分词模块。结巴分词：是国内的程序员基于PYTHON语言实现的中文分词工具。其工作原理为基于词典进行分词，致力于打造最好的PYTHON分词组件。主要的功能包括：(1)分词：目前结巴分词具备全模式，精确模式，搜索引擎模式三种分模式。其中全模式速度最快，但是存在歧义，将所有词语都拆分出来。精确模式适合文本分词，可以对句子进行精确的分词处理。本文主要采取精确模式进行分词。“搜索引擎模式在精确模式之上，对较长的词再进行切分，提高了召回率，比较适用于搜索引擎的分词处理 [45]” 。(2)自定义词典：用户可以增加自定义词典，补充结巴分词的词库。(3)关键词提取：返回TF/IDF权重最大的关键词，可以指定返回数量。(4)词性标注：结巴分词采用ICTCLAS兼容的词性标记法，对分词操作得到的每个词语的词性进行标注。

结巴分词模块分词的工作原理：

(1)基于Trie树形结构高效地实现“词图扫描”；(2)基于包含句子中全部汉字所有可能成词的情况，生成所构成的有向无环图（DAG)； (3)基于动态规划原理查找最大概率路径, 获得基于词频的最大切分词组；(4)基于汉字“成词能力”的“HMM模型”，利用VITERBI算法完成对于未登录词的处理.

### 3.4.2 分词模型HMM隐马尔可夫

隐马尔可夫模型是一种统计模型，用于描述状态间的转移，依赖于前面N个状态的过程。在HMM(隐马尔科夫)模型中，假设时刻t的观察值 只依赖于t时刻所处的状态，这中模式确保了序列中的所有观察值可以保持互相独立。如图3.7所示。

三个基本问题:

(1)给定一个模型，如何高效快速的计算出某一个观察序列发生的概率；

(2)给定一个模型和一个观察序列，如何识别出导致这一观察序列发生概率最大的引导状态序列；

(3)给定一个模型以及一个观察序列，如对模型的参数进行适当调整使得这一序列的概率发生的最大；

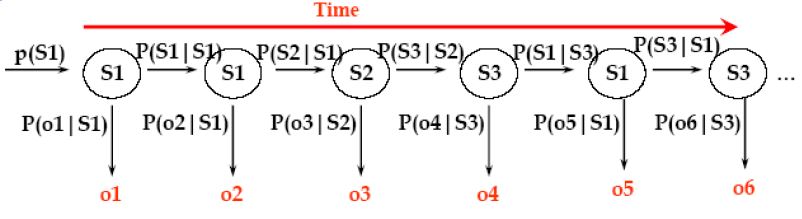


图3.7 隐马尔可夫模型

三个独立性假设:

(1)t时刻的状态只依赖于t-1时刻的状态；

(2)t时刻所生成的观察值又只是依赖于t时刻所处的的状态值；

(3)状态与具体时间无关；

HMM模型包含五类元素，两个状态集，三个概率矩阵。

S:隐含状态，

O:可观测状态，

π：初始状态概率向量，

A：转移状态概率矩阵，

B：符号发出的概率矩阵；

N：隐含状态的数目，

M：可观测状态数目。

隐马尔可夫模型框架应用相应算法解决以下三个问题：

(1)评估问题：forward前向算法，用于解决语音识别问题；

(2)解码问题：VITERBI算法，用于解决汉语分词问题；

(3)学习问题：Baum-Welch算法（向前-向后算法）；[46]

用公式表示HMM模型为：，可以用简要表示。下图3.8为通过HMM模型进行句子分词的例子。隐马尔科夫模型主要实现对于词性的标注。

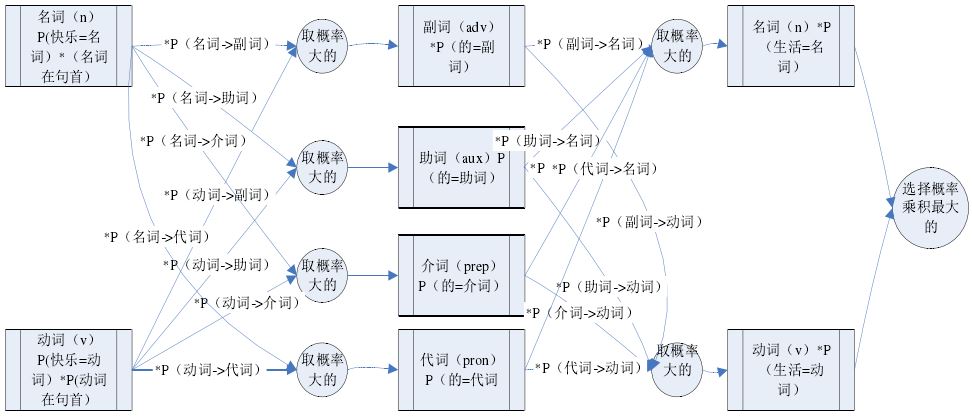


图3.8 HMM模型分词例子

### 3.4.3 分词算法VITERBI

VITERBI算法是一种基于动态规划的搜索算法，用于解决HMM模型标注词性过程中寻找最优路径。不仅可以找到较好的一条状态转移路径，而且可得出该路径对应的输出(output)概率。该算法将根据既定现象进行本质挖掘，寻找最优路径。例如基于HMM模型的词性标注过程对于每个句子，在若干词性组合的过程中，根据统计分析，计算每个组合出现的概率，确定概率较大的组合结果，最终获得每个词的词性标注。

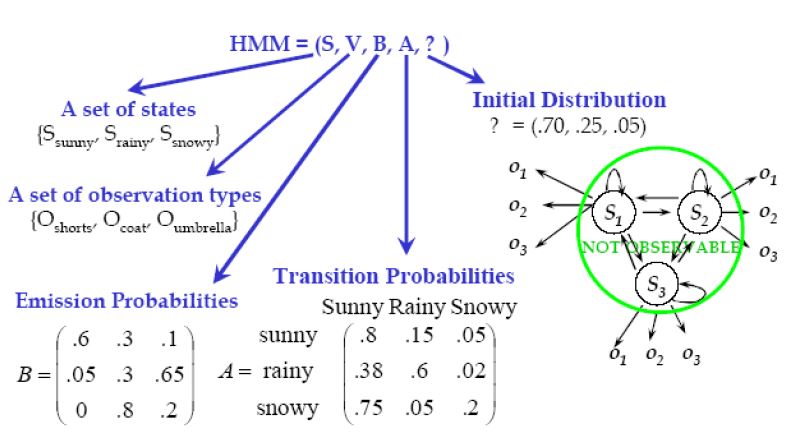


图3.9 VITERBI模型预测例子

维特比 VITERBI算法通过字典进行词性标注的过程如下：首先通过对字典的统计分析，获得各种词性出现在句首的比例，两种词性在句子中相邻的次数，每个词不同词性出现次数。利用上述知识，对论坛中帖子内容以及回帖内容进词性标注[47]。

## 3.5 数据可视化

数据可视化伴随着统计学的诞生而产生，采取图形图像，记录量化信息的思想。从15世纪便产生，随着计算机技术的高速发展数据可视化技术伴随着商务智能技术得到了极大的发展。目前数据可视化与数据分析，大数据分析相结合，作为数据分析的主要手段。由数据可视化，进一步发展出信息可视化，知识可视化，思维可视化，可视化分析等多个领域 [49]。本文涉及的数据可视化技术包括Tableau，Excel Pivot table。

Tablueau 被称为数据领域的梵高，作为桌面系统中最简单成熟功能强大的商业智能工具软件，不要求强迫用户编写自定义代码，新的控制台可以进行自定义配置，与各种数据库进行无缝连接，提供强大的分析能力。Tablueau控制台非常灵活且具高度的动态性。Tablueau将数据运算与美观的图表完美地结合起来，操作简单，可实现大量数据拖放到数字“画布”上的效果。同时图表是一种动态的图表，展示的结果会随着其所连接的数据的变化而变化。其产品具备轻松整合，交互性，免费的特点。同时Tableau Server允许企业将数据可视化结果展示到互联网上，不受时间空间的限制。利用Tablueau Reade用户可以对用户采用Desktop创建的交互式数据可视化工作部文件，使用按过滤、排序以及调查得到的数据结果进行交流。

Tablueau作为世界五百强企业的首选，Tablueau Desktop，Tablueau Server，Tablueau Reader，Tableau Public可以为企业提供整套的商务智能可视化服务。Tableau主要面对非技术人员，用户可以轻松的对已有的数据进行可视化处理、可交互的即时展示与分析。数据可视化是Tableau的核心功能，主要包括以下两方面：

(1)独创的VizQL数据库：Tableau是来自斯坦福的数据科学家们为了实现数据完美的展示以及后期处理展示等作用，把关系型数据库Database的语法与可以完成驱动图像展示的编程脚本进行结合，开发出了VizQL，全新的可视化关系型数据库，可以多种角度且高效率，更加直观的展示多维数据仓库的逻辑关系以及结果。

(2)用户体验上对易用性的完美呈现：Tableau具备易用且美观的使用界面，能够处理规模巨大多维的数据时，从不同角度实时的设置下看到数据所表现出来的规则。

# 第四章 企业内部信息识别方案设计

本章针对电子商务企业构建“海淘”中文页面与英文页面进行链接，识别“海淘”热销商品的项目背景，完成方法的分析与设计。基于企业内部数据仓库TERADATA，建立两类模型：第一：基于“海淘”流程的购买模型；第二：基于搜索关键词的供求模型。第一类模型将识别历史购买中的热销产品，第二类模型将识别潜在购买需求中的热销产品。综合两类模型得到“海淘”热销产品信息，用于中文“海淘”页面展示，将扩展中国市场份额并指导运营工作，详细内容如下图4.1所示。

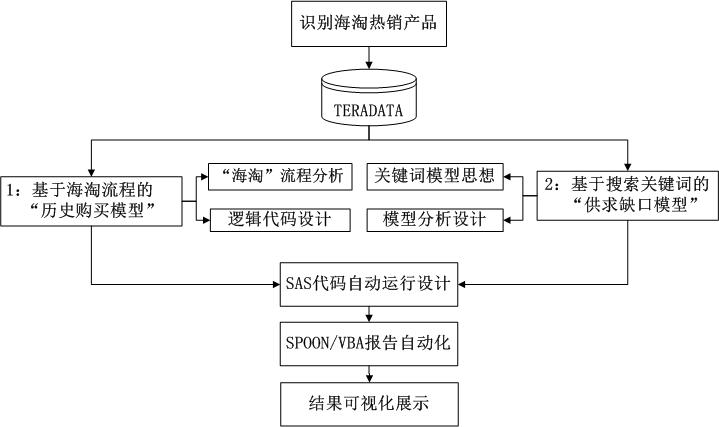


图4.1 本章总览

首先，为了完成热销商品识别的目标，针对“海淘”具体流程，本文将基于TERADATA数据仓库的大数据分析，为“海淘”项目提供了完整的数据支持。通过对企业数据仓库内存储的海量数据进行分析，获取中国消费者历史消费数据，以及搜索信息。针对历史购买数据，依据“海淘”流程建立历史购买模型，分析热销产品信息。基于搜索关键词代表的潜在购买需求，建立基于关键词的供求缺口模型，获得潜在热销产品信息，如图4.2所示。两个模型将分别在4.1节与4.2节中进行讲述。

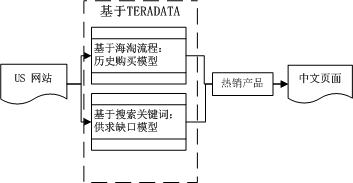


图4.2 基于TERADATA建模识别热销产品信息

## 4.1 基于海淘流程的历史购买模型

本节将对第一个模型的设计思路，数据结构，模型代码设计进行详细的分析介绍。首先，介绍“海淘”具体流程；其次，甄别全部中国消费者消费信息；再次，获取信息的数据结构分析，设计模型结构；最后：分析两类模型的代码设计思路。

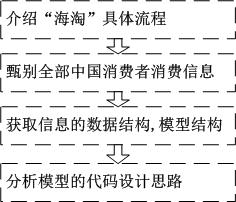


图4.3 历史购买模型设计过程

### 4.1.1 “海淘”流程介绍

“海淘”的整体流程包括以下几个部分：国外电商网站注册🡪收货地址填写（选择直邮服务的中国境内地址/转运服务的公司仓库地址）🡪优惠信息领取🡪商品选择🡪支付🡪物流方案选择🡪海关申报、清关🡪国内运输🡪售后服务等，具体如下图4.4所示。

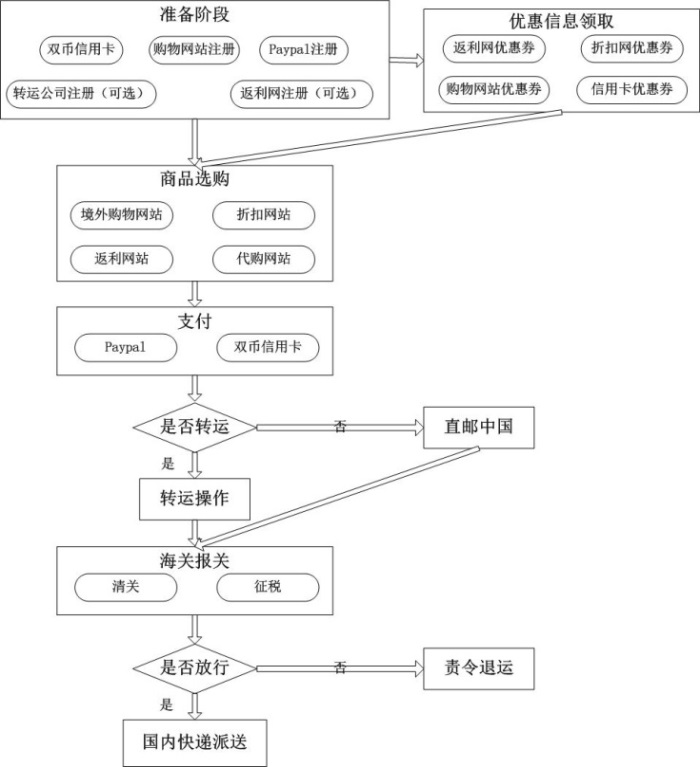


图4.4 “海淘”整体流程

“海淘”的各个流程在E公司的数据仓库中均有记录，可以通过对各个流程的数据分析来获得全部中国消费者的消费记录。例如，对物流环节进行鉴别，包括以下两种情况(1)消费者选择的直邮地址为中国地区地址；(2)消费者寄往美国的地址为中美转运公司的美国仓库地址；则可以判定该消费者为中国消费者，即使其个人信息中存在错误(将消费者所属国家被标注为US等)也可以判定出该消费者为中国消费者。在对历史数据进行分析时需要将上述两种情况鉴别出来，并加入到中国消费者的数据中。基于上述思路依据跨境交易的全部流程对消费者信息进行甄别，获得中国消费者信息库，以及中国消费者的购买信息，从中找出热销产品信息；从而获得：基于“海淘”整体流程分析的热销产品获取模型。

### 4.1.2 甄别中国消费者购买信息

分析全部中国消费者“海淘”的全部购买记录，从中得出“热销产品”信息。共需完成四个步骤：(1)获得全部中国消费者信息；(2)获得中国消费者全部购买记录；(3)对全部购买信息进行数据挖掘分析，分类汇总；(4)对于“海淘”页面展示商品进行筛选，如图4.5所示。

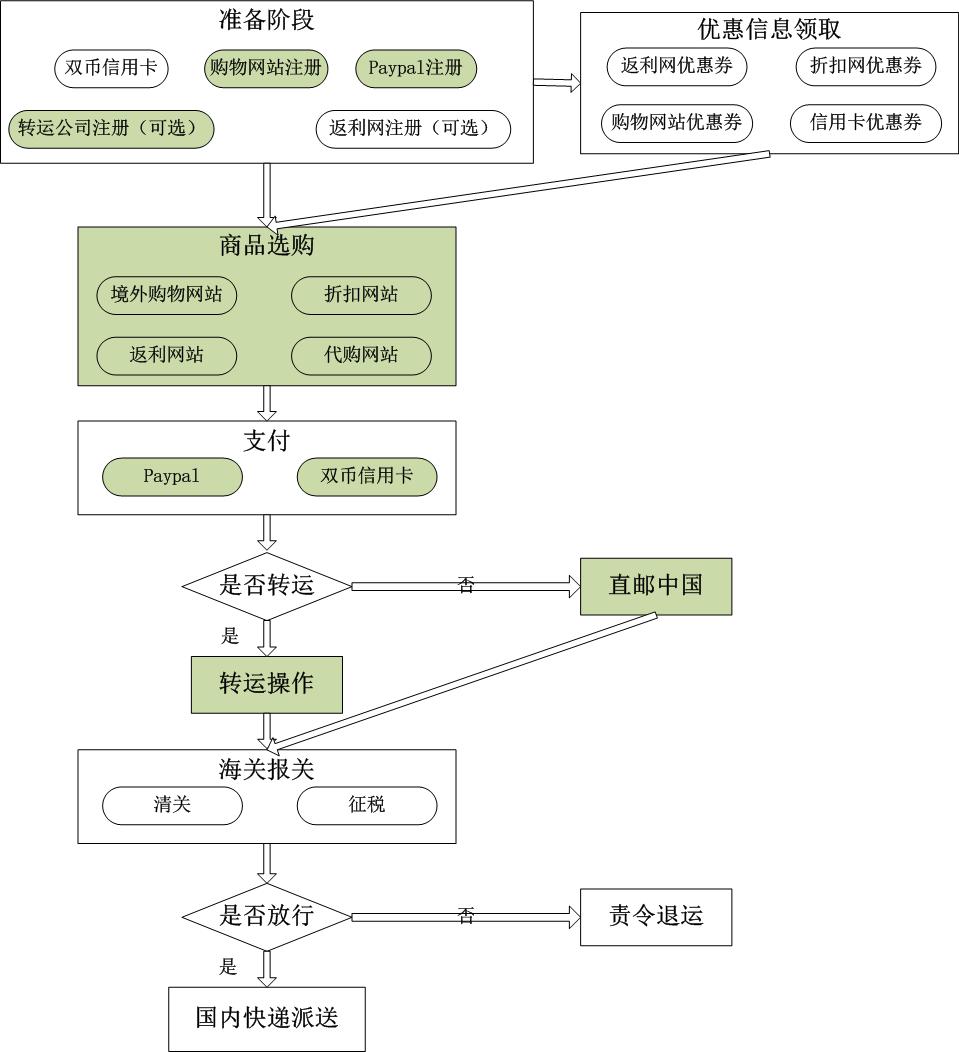


图4.5 基于“海淘”流程甄别过程

为了保证数据完整性需要对全部数据进行统计，需要将消费者在“海淘”过程中存在的错误全部纠正。比如，注册时并未将国家选为中国，这将严重的影响数据的完整性，据统计此种错误出现的概率达到30%-40%。本模型将从“海淘”的各个环节对消费者所犯错误进行甄别，以期获得全部中国消费者的购买信息。

基于上述标记颜色的环节进行中国消费者购买甄别，主要在以下环节进行：(1)注册信息， E公司网站的注册信息以及PayPal两者有一个的国家选择中国（包括大陆，香港，台湾，澳门在内）则将消费者的购买定义为是中国消费者购买；(2)购买环节：如果消费者购买时的IP地址多次显示为中国的IP则认为是中国消费者购买；(3)支付环节：如果支付是采用的是中国境内的信用卡通过PayPal进行支付，则认为是中国消费者购买；(4)物流环节：如果购买时消费者选择直邮中国，则认为是中国消费者购买；如果消费者选择将商品寄送到中美转运公司的美国仓库地址与联系方式，则认为是中国消费者的购买；

### 4.1.3 数据结构及模型设计

对基于上一小节方法获得的全部中国消费者的购买信息进行采集。数据结构设计如下：包括全部交易信息：(1)买家个人信息：买家ID，用户名，注册时间，E-MAIL, Phone Number 注册国家信息，注册站点信息，邮寄地址信息，第一次成功交易时间等；(2)卖家个人信息：卖家ID，卖家店铺名称，卖家信息用评价分数，卖家的商品总数,卖家销售的产品相关信息等；(3)详细交易信息：交易编码(Transation\_id)、商品编码（Item\_id）、交易时间、交易金额、交易数量、交易站点；(4)购买产品信息：产品ID，所属最高级别的分类（Catergory），初级分类(Meta)，一、二、三级分类(CAT\_LV1, CAT\_LV2, CAT\_LV3)，品牌，产品标题(Auction\_title)，详细信息（尺寸，颜色，内存大小等），商品单价（Price）,商品所在地点（Item\_location）。

表4-1 购买信息分类数据结构

|  |  |
| --- | --- |
| 信息分类 | 详细信息 |
| 买家个人信息 | 买家ID，用户名，注册时间，国家信息，E-MAIL,PHONE、邮寄地址等； |
| 卖家个人信息 | 卖家ID，卖家店铺名称，卖家信用评价分数，卖家的商品总数等 |
| 详细交易信息 | 交易ID，交易网站（US,UK,AU..），交易时间，交易金额，交易数量； |
| 购买产品信息 | 产品ID，产品所属国家，Category，Meta，LV1\_CAT、LV2\_CAT,LV3\_CAT(分类)，Brand，产品标题，详细信息，交易金额，交易数量 |

基于上述基础信息对消费者购买信息进行多角度，全方位的查看，包括：

(1)基于产品类别Category，Meta，LV1\_CAT，LV2\_CAT，BRAND；

(2)基于产品所在网站，查看消费者在美国、英国、日本、澳大利亚站购买记录；

(3)基于时间维度：消费者在各个季节、月份、重要节日等分别倾向于购买什么；

(4)基于海淘流程维度：转运热销商品，直邮热销商品，“银联卡”热销商品等。

对全部购买信息进行数据挖掘分析；通过对购买商品的时间维度：精确到月与周，星期，时间点；品类信息：各个维度的类别，品牌，品名。基于上述维度，获取中国消费者购买的TOP产品信息。本章采用结构化查询语言SQL于数据仓库TERADATA，基于“海淘”这样跨境交易购买全部流程按照分析逻辑进行数据挖掘建模，最终获得所需热销产品信息。

最后，“海淘”中文页面运营中，对于“海淘”页面展示商品进行筛选；此处考虑如下：中文页面上线之前与中文页面上线之后有所不同；上线之前需要对北美站点全部中国消费者购买信息进行获取分析即可；上线之后，除了完成上述步骤之外，还需要对中文页面已经展示的信息进行分析，对正在展示的商品取得的销售成果进行判断，检测该商品是否继续用做“海淘”热销产品。

中文页面产品筛选主要思路如下：对该商品的点击量，转化率，以及带来的销售额，销售数量，新卖家数量，等维度进行检测，如果不达标则考虑后续过程中不再将该商品加入到其中了。同时中文页面启动以后，E公司与中国合作伙伴合作，包括通过对各个搜索引擎投放关键词广告的LINKTECH，SEM（百度广告投放），以及返利网Rebatesme，SMZDM，社交网络SNS等途径进行热销产品推销，对于上述关键词带来的销售量，转化率进行检测，转化率较低的关键词便不再进行关键词投放。同时带来交易量较少的商品选择取消其展示的资格。

### 4.1.4 模型代码设计

本节基于TERADATA，采用的语言为SQL，详述了代码逻辑。首先：基于“海淘”流程获得全部中国消费者信息：基于甄别中国消费者的方法进行代码设计。如下图4.6种左半部分关于中国消费者信息的甄别环节。

建立逻辑，基于数据仓库中已有的表，建立符合要求的表，用于获得全部中国消费者信息；基于上述步骤中识别出的全部中国消费者信息，获得全部中国消费者三年内的购买记录（本文选择2011.12-2014.12）。将需要的数据按照标准，比如一年的购买金额超过一定的金额等条件进行数据筛选。筛选出总体的热销产品信息，类别，品牌，标题等信息。针对不同的需要对数据进行抽取，建立独立的表用于分析，例如下图中包含的转运热销产品，直邮热销产品，“银联卡”热销产品等类别。

从下图4.6中可以看出，基于“海淘”流程的代码设计思路分为几步；首先，采取“海淘”流程为判断依据，甄别中国消费者信息；其次，采集全部中国消费者的购买记录；再次：针对中国消费者全部购买记录进行基于时间，品类，品牌，物流，支付等各个维度的挖掘分析；最后：将上述信息进行关键词转换，获得热销产品的关键词信息。本步骤获得的热销产品信息以及热销产品关键词将被最终将被用作“海淘”中文页面进行展示。

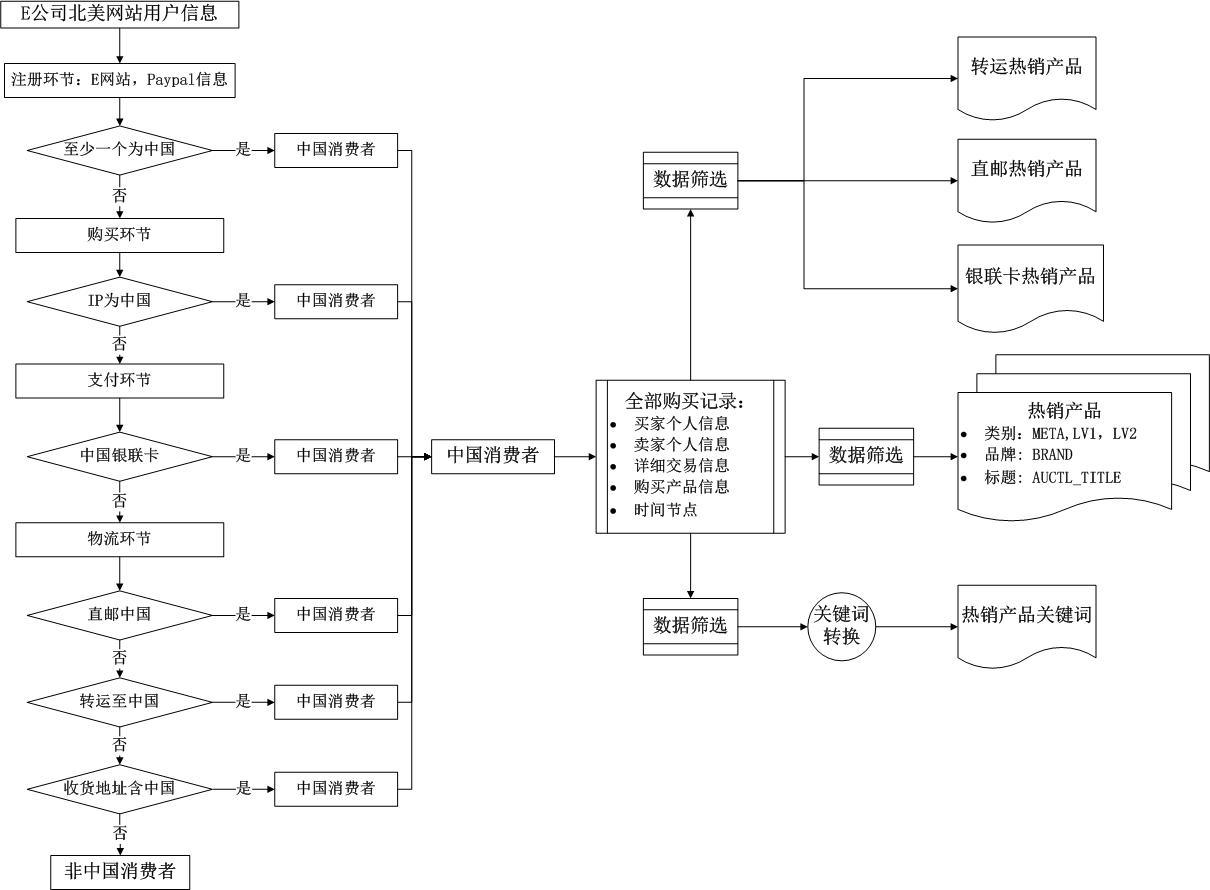


图4.6 代码构建逻辑图

## 4.2 基于搜索关键词的供求缺口模型

本节将对“基于搜索关键词的供求缺口模型”的设计背景，设计思路，数据结构，模型代码设计进行详细的分析介绍。首先，介绍项目背景；其次，介绍经济学领域供求模型，基于搜索记录的“供需缺口”模型；再次，模型代码设计。具体内容如下图4.7所示。

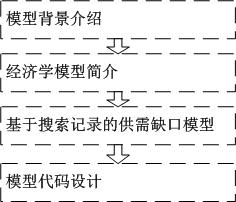


图4.7 历史购买模型设计过程

### 4.2.1 模型背景介绍

消费者在E公司网站对服务器发出的每一次请求都会被数据仓库TERADATA记录下来。通常，消费者的搜索记录与最终的购买产品之间存在一定的关联。某些关键词如iPhone，MacBook，iPad，Nike等较多中国消费者喜爱的产品都会成为搜索关键词中被大量搜索的，同时交易量也是排名靠前的。

但是，当某些关键词被消费者大量的搜索，而网站中相应产品的供应量较少时，则意味着潜在的市场需求。例如：近年来中国大陆大部分城市被雾霾困扰，空气净化器（Air Cleaner）以及其知名品牌，Blue air，IQair，Honeyware等的搜索量较大，但是由于国际运费较高，卖家销售地理范围将中国剔除掉，以及转运费用较高等原因导致很多消费者最终无法进行购买。这种搜索量大，但是最终交易较少的将是潜在的市场需求，此种产品放在“海淘”中文页面销售到中国，将给E公司带来更多交易。

此外，E公司内部数据开发部门开发出来基于Hadoop的关键词归类工具L，该工具会自动识别关键词，将其进行分类处理，比如上文提到的对Iphone的多种搜索方法，调动工具L可以进行自动识别。调用此工具，将买家的搜索记录进行分类处理，卖家的提供记录进行处理，两者进行匹配。 本节将调用工具L并基于供给与需求之间的较大差距的关键词建立“供需”缺口模型，找出潜在热销产品信息。

### 4.2.2 供需缺口模型

经济学的供需模型是基础的模型之一。其基本思想为某种商品的供给与需求都是与价格相关的，价格提高，供给量增加，需求量减少；反之，价格降低，供给减少，需求增加。商品价格的不合理导致该商品的供给量与需求量之间存在着差距。

本文中将借助上述思想，提出买家角度：对关键词搜索量而定义的某商品需求量，以及站在卖家角度：对某种关键词的某种商品展示数量，代表该商品的供给量。如果消费者对某个关键词的搜索量较大，但是E公司的网站上相应的在卖品较少，则说明该商品具备较大的市场空间。可通过指导卖家增加相应产品的供应量来增加市场交易额，若将本商品列在中文页面上，则预期会带来较多的交易。

基于上述思想建立下图4.8，X代表各个被检测的关键词，Y代表相应的关键词搜索量以及卖家在卖的产品关键词数量。当代表某种商品的某个关键词其搜索量巨大，但是该种商品被提供的数量较少，则意味着该商品的市场潜力较大。如果中文页面中提供相应的产品，并进行良好的营销与“海淘”服务则可能获得较好的市场表现。

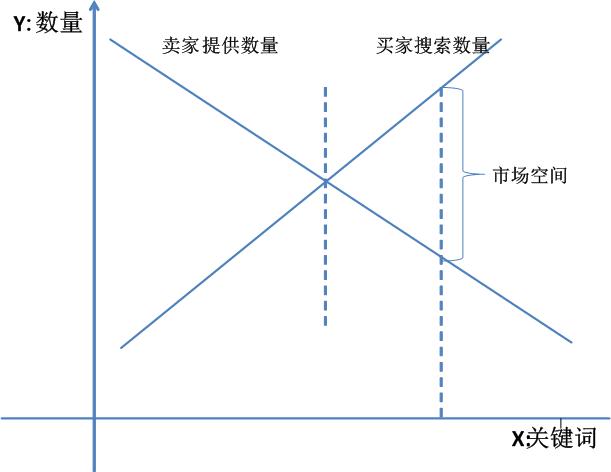


图4.8 基于关键词搜索的市场空间模型

基于经济学供需模型将上述思想转化为基于关键词搜索与提供差距的商品供需缺口模型。

X轴：关键词数量差=买家搜索量—∑权重\*卖家提供的商品量；

Y轴：代表卖家的供给量或者买家的需求量；用于如下图4.9。

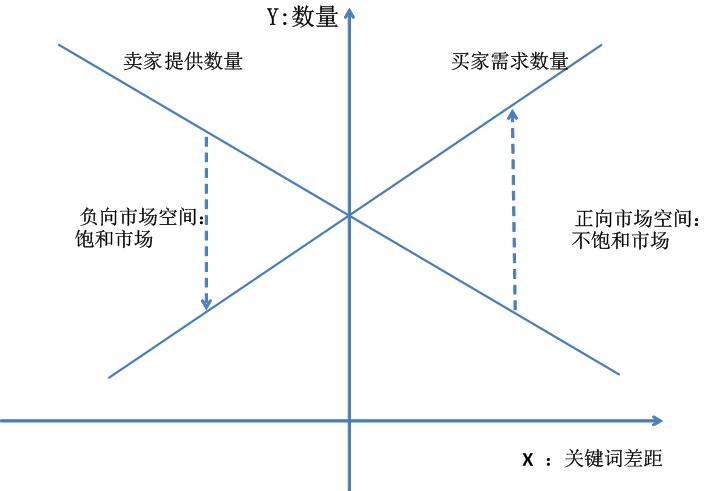


图4.9 市场不饱和状态分析

图4.9种展示出，当某种商品需求量大于供给量较多时，代表该商品存在一种正向的不饱和市场空间，为具备潜力的热销产品；反之，当某商品的需求量较少，供给量充足，此时则代表饱和市场（此处饱和市场仅为关键词这个角度来看，并不代表该商品为非热销产品）。此时从关键词搜索角度，无法体现该商品具备较大市场空间。需要从成交量，成交额，点击率等方面进行进一步的度量。因此将坐标轴右侧的关键词代表的商品作为目标产品，当两者差距达到一定的程度后，将该产品作为候选产品。

当某种商品的供给量与需求量之间存在正向的不匹配的情况出现，则代表市场空间。“搜索记录”通常被转化为关键词，对关键词实现精确的匹配，并与数量充足、质量可信的商品供给对于电商网站的交易额起着重要作用。当E公司对关键词进行精确的解读，并调动卖家提供了更多优质的此类产品，会带来如下结果：供给增加，市场差距缩小，成交量会增加。此效应是中文“海淘”页面希望带来的结果。当供给增加后，市场差距将缩小，供求匹配的机会增加，此时交易量将增加。具体过程如图4.10所示。

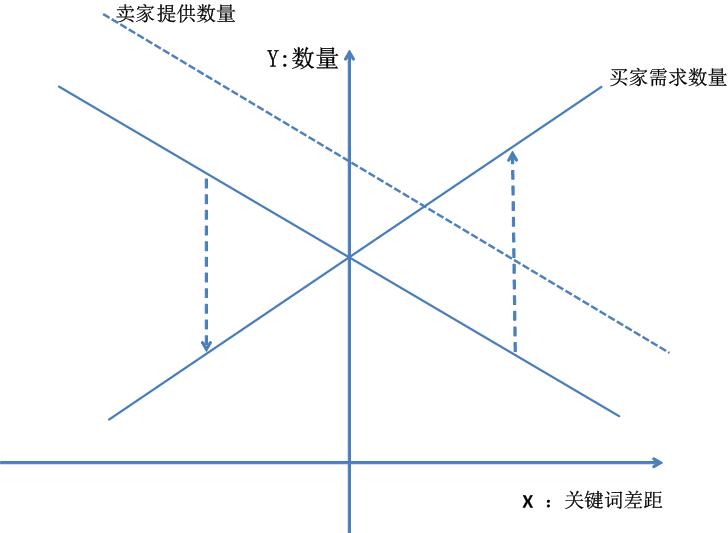


图4.10 供给增加成交增加

如下图4.11所示，当商品的供给与需求同步提高后，其交易数量以及金额将可以预见的会增加。如图所示，预期带来Y1🡪Y2的变化过程。E公司有专门的分类工具L负责对关键词进行数据挖掘，实现各种关键词的分类，完成了用户输入的关键词列表与数据仓库中的关键词的匹配。例如：IPHONE，iphone， APPLE CELL PHONES,apple phone等一系列的关键词可以精确的匹配到iphone，本文将直接利用上述工具L进行应用，不在进行此类技术的分析设计。

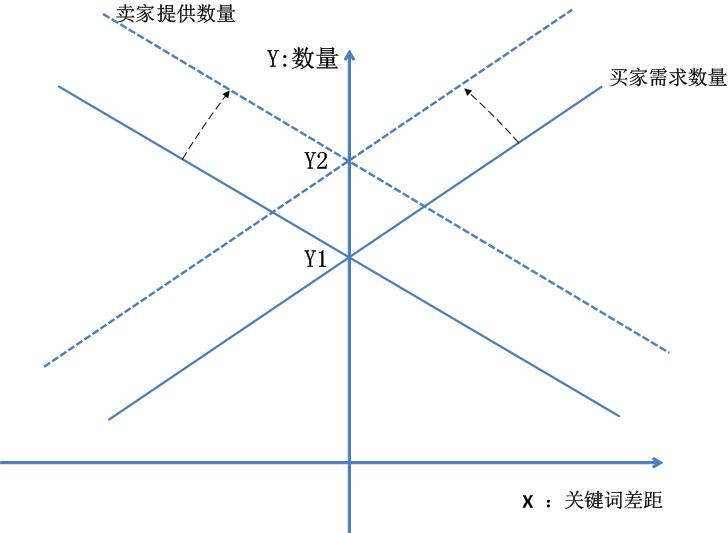


图4.11 供给增加成交增加续

基于上述对模型设计思想的分析基础之上，本模型的实现过程如下图4.12所示：(1)调用已有“将搜索记录转会为关键词”的工具L，对中国消费者的搜索记录进行分类汇总，同时对US站点的卖家提供的商品关键词分类汇总，计算出每一个关键词的供求差距，将信息存储在数据仓库表中。(2)锁定美国站点，从数据库中获取差距较大的关键词列表。对关键词信息进行归类，获得该关键词所属的Category，Meta，LV1\_CAT，LV2\_CAT，BRAND等信息。(3)筛选取出适于“海淘”的相关类别的关键词作为候选关键词。最终，将结合本模型与上一步模型获得的信息进行综合的考虑，获得中国消费者“海淘”热销产品信息。

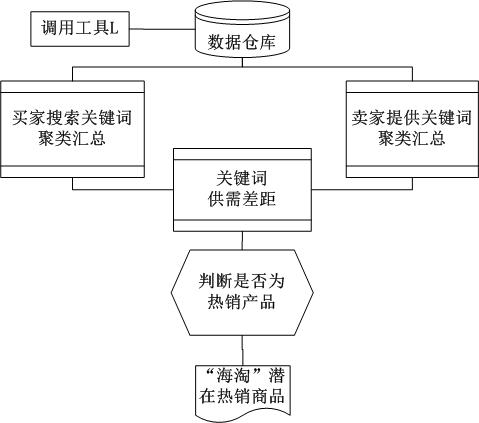


图4.12 供求缺口模型整体流程介绍

### 4.2.3 模型代码设计

基于搜索记录获取供需差距的模型主要涉及E公司内部数据部门开发的基于Hadoop的工具L，将相近的关键词进行聚类。例如不同的iphone搜索方法均会被归结为iphone，本节模型的设计将直接利用上述工具。如图4.13所示完成热销商品识别以及分类。

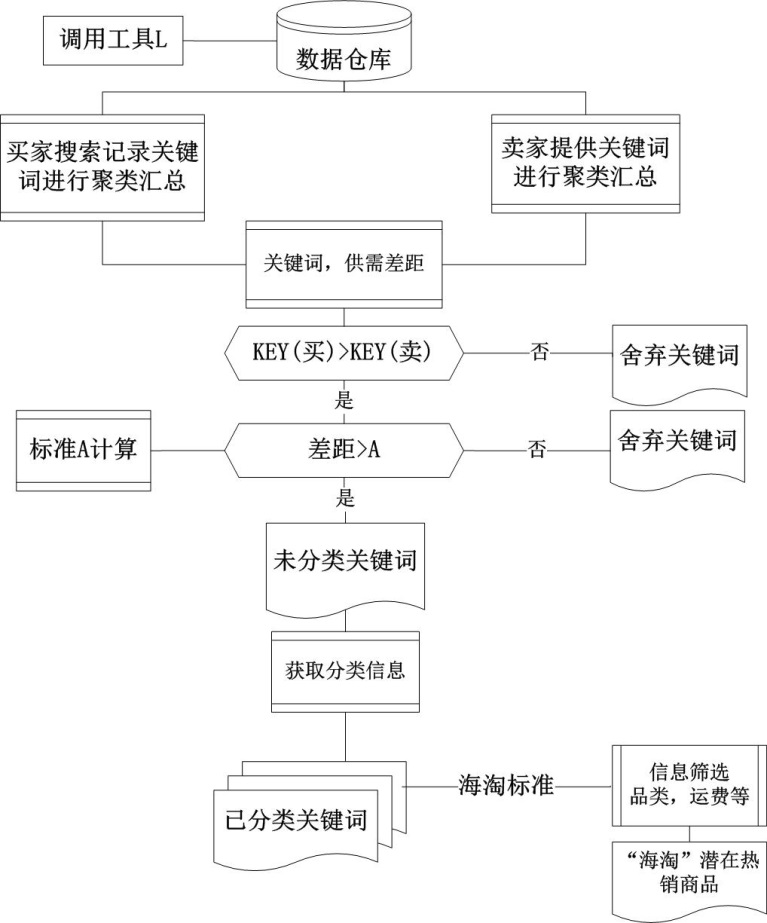


图4.13 供求缺口模型逻辑图

本模型主要完成对于数据仓库中买家与卖家提供的搜索记录进行对比分析，对两者的差距采用模型进行度量，得出搜索较多，提供较少的商品关键词。调用数据仓库中已有信息，获取上述关键词所属的类别信息，包括CAT,META,CAT\_LV1，CAT\_LV2，CAT\_LV3，BRAND，AUCTION\_TITLE等信息。依据符合“海淘”标准的筛选条件，对获得的关键词信息进行筛选，比如去掉“古董”等类别的商品关键词。对于标准“A”的计算，本文采用各大类商品META即第一分类级别的分类中，关键词“买家搜索/卖家提供” 的四分位点针对每个META类别进行计算与筛选。

## 4.3 代码及报告自动化

建立模型之后，为了提高效率以及准确率以及降低人力成本，本节将针对上述模型完成基于SAS的代码自动化更新运行以及报告制作发送自动化过程。

### 4.3.1 基于SAS的代码自动更新

对于设计成功的代码，根据业务需要，需要重复的完成对数据表的更新操作，如果人工完成将浪费很多时间且容易遗忘，导致任务疏漏。此外，TERADATA在众多任务同时运行的时候将出现速度降低等问题。故采取SAS Automation功能，让SAS自动运行脚本，完成在数据仓库中按找定义好的逻辑进行建表，更新基于模型建立的表中数据等操作。Software:

(1) WinSCP，Putty;(2)A SAS account;

(3)A folder in /sas/cac like: /sas/cac/dandanDong

具体的步骤如下，第一步：申请SAS Account以及SAS 文件夹， 安装Putty，WinSCP。第二步：在SAS文件夹中建立与自动化报告同名文件夹，如GEOX代表“市场扩展”组需要运行的代码存放的文件夹，设置权限，创造者自己为OWNER，包括：可读（R）、可写（W）、可执行（E）三种权限，其他用户有用读的权利，无修改权利等。

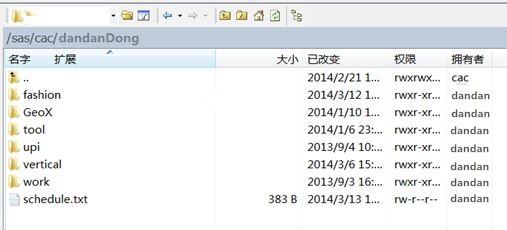


图4.14 权限设置

第三步：在Geox中建立Geox.sql，Geox.ksh；Geox.sql文件用于存放需要自动运行的SQL代码，设置权限，自己可读可写可执行，Geox.ksh(WinSCP 编辑的)文件，并设置权限；KSH文件是脚本语言可以识别的文件 ；

第四步：将SQL代码存放到上述Geox.sql文件中，比如本段完成对表ams\_creative\_Geox\_v的更新操作。

CREATE VOLATILE TABLE ams\_creative\_Geox\_v AS---------1

(………

)WITH DATA PRIMARY INDEX(cal\_dt,ams\_Geox\_id)

ON COMMIT PRESERVE ROWS;

DELETE FROM P\_CAC\_T.ams\_creative\_Geox

WHERE cal\_dt IN

(SEL cal\_dt FROM P\_CAC\_T.ams\_creative\_Geox WHERE cal\_dt >= CURRENT\_DATE - 30)AND cal\_dt >=

(SEL MIN(cal\_dt) FROM ams\_creative\_Geox\_v);

INS INTO P\_CAC\_T.ams\_creative\_Geox

SEL \* FROM ams\_creative\_Geox\_v a

WHERE NOT EXISTS

(SELECT \* FROM P\_CAC\_T.ams\_creative\_Geox z

WHERE z.cal\_dt=a.cal\_dt

AND z.ams\_Geox\_id=a.ams\_Geox\_id);

此段SQL代码代表对表的一些更行操作。由于E公司的数据仓库存在两天的时滞，且每天晚上九点完成更新，故自建逻辑基于数据仓库的表记性更新时在每天早上7点以后则可以获得两天前的完整数据（美国时间为准）。KSH文件是脚本语言可以识别和驱动的程序，KSH用于自动化的完成：(1)数据仓库的连接；(2)用户名密码输入；(3)自动运行指定的SQL代码；(4)自动阐述日志文件。

#!/bin/ksh

jobName='Geox'

jobFolder=/export/home/cac/dandanDong/Geox

loginFile=/export/home/cac/dandanDong/work/logins

loginString=`grep -i mozart $loginFile`

sqlFile="$jobFolder/Geox.sql"

kshFile="$jobFolder/Geox.ksh"

toEmail="lqiang@e.com,daDong@e.com,zhoudong@e.com"

jobDesc="This is a job for \"$Geox - Mozart\".\n\nSQL File: $sqlFile\n\n$Memo\n\n"

jobDesc=$jobDesc"Back End Server: "`hostname`"\n"

jobDesc=$jobDesc"CronJoblog: /export/home/cac/dandanDong/work/schedules/CronJobLog/cronjob.log"

详细的KSH 代码付于附录之中。

第五步：指定每个任务的运行时间，对dandanDong文件夹中的schedule.txt进行设置，完成对每个任务的运行时间分、时、日、月、周的设定，格式如下：

\* \* \* \* \* file path

m h d m w (分 时 日 月 周 )，之间存在空格键。例如：

00 12 14 \* \* /sas/cac/dandanDong/upi/upi.ksh 代表每月14日，12点运行；

00 08 \* \* \* /sas/cac/dandanDong/ GeoX/GeoX.ksh 代表每天08点运行；

00 02 \* \* 2 /sas/cac/dandanDong/vertical/weekly\_vertical.ksh代表每周二02点运行；

第六步：登陆Putty，加载SAS，登陆SAS；

Open your putty

Host Name: phximdsas02.phx.E.com

Port: 22

Saved Sessions: \*\*\* (eg:sas/hadoop) 🡨 代表进入SAS或者HADOOP

Login SAS Server

Login as : dadong

Password: welcome123

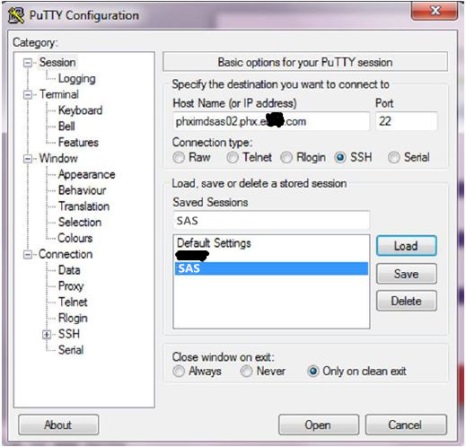


图4.15 通过Putty登陆SAS

登陆SAS之后可以直接调用执行schedule.txt完成其中的全部任务，或者可以直接完成对某一项任务时间的设定。

crontab /sas/cac/dandanDong/schedule.txt 执行文件里所有命令；

crontab 00 12 13 \* \* \* \* /sas/cac/dandanDong/upi/upi.ksh 执行指定任务。

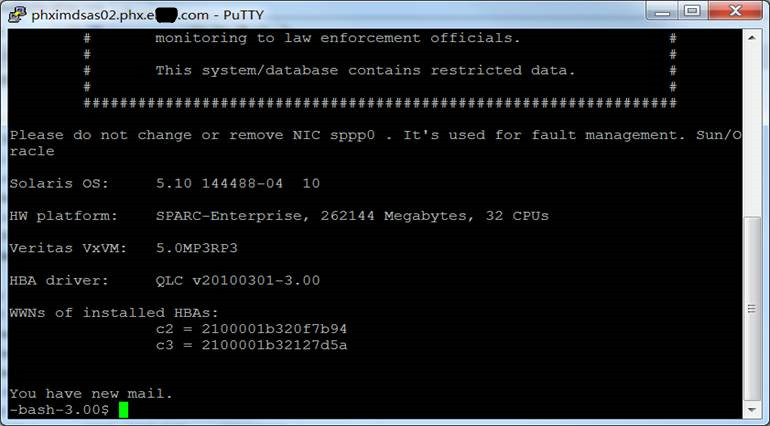


图4.16 运行schedule.txt

如图4.16展示了基于SAS平台进行自动化设置中调用任务时间列表的过程。表4-2中列举了上述过程中可以调用的一部分命令列表。

表4-2 脚本命令含义

|  |  |
| --- | --- |
| 命令 | 含义 |
| crontab –e | 进入编辑模式 |
| crontab –l | list所有的命令 |
| Ctrl + c | 退出当前页面 |
| Shift + zz | 保存并退出 |
| Shift + :q! | 强制退出 |
| Shift + wq | 保存退出 |

第七步：等待结果Email，KSH中包含Email列表，SAS会发送数据更新结果给指定的list；如果Email结果显示为：GeoX.SQL SUCCESS！代表数据更新正确完成，可以进行下一步操作；否则检查Log日志查询错误原因，完成对SQL的修改，修改schedule.txt中的代码执行时间，重新执行上述运行步骤并等待结果。

### 4.3.2 基于KETTLE的自动化数据处理

在上一步完成SAS对任务表的更新之后，需要利用SPOON，VBA自动化的完成数据抽取，转化，加载，并最终以报告的形式发送给指定的邮件列别(Email List)。此处需要注意SPOON，VBA只能完成对数据的读取操作，无法完成数据更新操作，所以本步骤必须在SAS数据更新操作执行完毕之后进行，工程实现的最终结果如下图4.17所示。

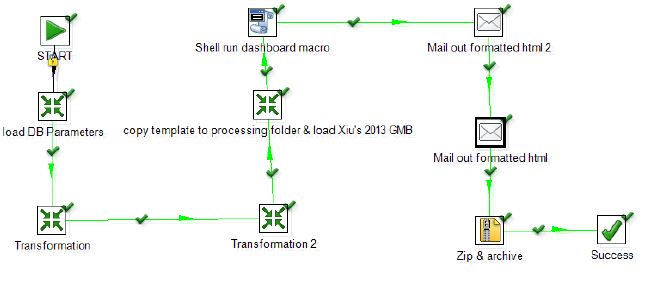


图4.17 SPOON自动化工程界面展示

SPOON自动化ETL设置：准备工作，安装JDK,安装JDBC,运行SPOON.bat成功安装SPOON并配置其运行环境之后，需要建立SPOON工程，具体步骤如下：(1)设定启动时间,连接数据仓库，输入用户名，密码；(2)完成对数据仓库中已更新表的数据抽取；(3)导入数据到指定Excel template的指定位置；(4)驱动dashboard将EXCEL中制作表格，图表等结果进行截图操作；(5)发出带有附件（Report结果）的邮件给指定的邮件列表；(6)压缩结果保存在指定的位置；以每日市场交易金额日报的自动化为例进行展示：包括对每日的结果建立表格和图形比较的形式展示。

在SPOON中建立工程CN-DAILY GMB TRACKING-EMAIL.kjb 具体步骤如下：

第一步：设定启动时间：如图4.18所示，模块START，双击START,设置重复频率，启动时间；



图4.18 SPOON Automation时间设置

第二步：连接数据仓库，完成数据库仓库的载入，连接；本工程中所有箭头都代表完成顺序，比如本例中，先加载相关参数，后将上述参数对应的输入到系统变量中。如图4.19所示：

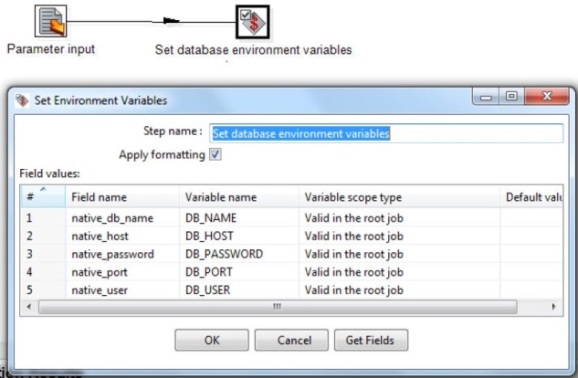


图4.19 加载数据库参数

第三步：完成对数据仓库中已更新表的数据抽取，存储到指定位置，并对其进行结果处理。将完成数据读取的SQL代码添加到Extract data中，运行时将完成对数据仓库表中数据的读取，如图4.20所示；

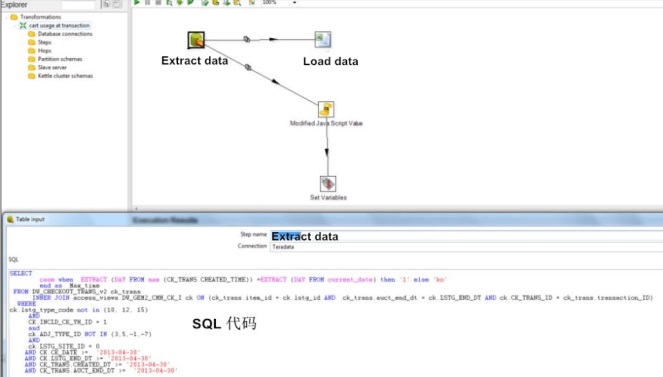


图4.20 从数据仓库中加载数据

load data的file&sheet中需要加载dashboard 以及input template.xls。Content中将数据结果中的字段GMB,BI，CAT,LV3，BRAND等需要用于制作报告的字段信息提取出来存放到指定excel文件input template.xls🡪output sheet🡪C5单元格开始进行数据写入。加载的已经编辑好的dashboard（内含有VBA程序）用于对input template中的数据进行数据可视化结果中截取指定的表格范围进行拍照保存成JPEG格式的图片，作为结果的附件，如图4.21所示。

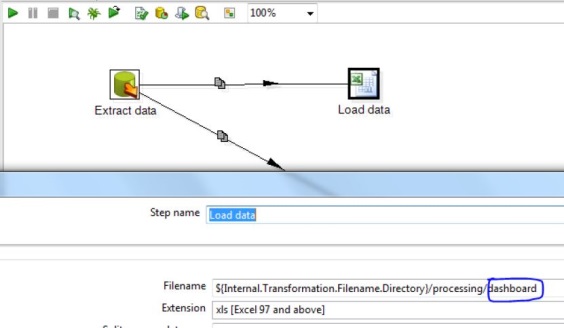


图4.21 加载数据模块处理

第四步：邮件结果编辑；首先：邮件标题设置，邮件标题中涉及到系统时间，故需要完成对系统时间的获取，get week data中完成对系统时间的获取，email subjet中用于完成邮件格式的编辑，最后将邮件标题存到系统变量中，如图4.22所示。

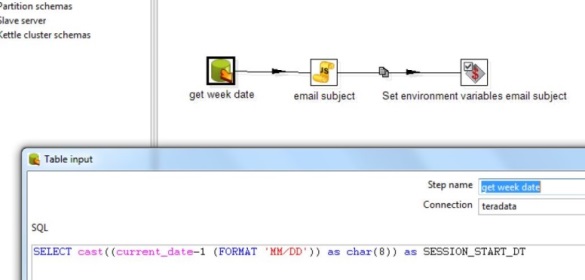


图4.22 获取系统时间

第五步：完成对邮件页面以及邮件发送列表的编辑；上述工程中的mail out formatted html将完成邮件页面的结果编辑，input template中的结果作为邮件附件，同时将代表结果信息汇总的图表的照片进行截取并将结果照片添加到邮件正文，按照要求发送给指定的邮件地址列表。

第六步：压缩数据结果并存储到指定位置。结果文件将会被压缩放入到指定的文件夹中 …\archive\dataXXX.zip，在此处可以查询到所有的数据结果压缩文件。将上述工程应用到对于热销产品自动获取的最终结果即可，但是热销产品的获取的结果需要对EXCEL模板input template完成如下处理：将数据加载到EXCEL 中，将input template中制作出固定格式的数据透视表Pivot table，包含筛选，排序等操作，对输入的数据信息进行汇总处理。

### 4.3.3 基于VBA的自动化数据处理

VBA实现Automation的方式比较复杂，因为VBA没有成型的框架，所以需要手动的编写每一步的程序。VBA作为SPOON中存在错误时的备份操作。

编写的程序由需求决定，具体步骤如下：

步骤一：设定运行时间

步骤二：连接数据仓库

步骤三：确定数据库中的数据已经更新

步骤四：导入数据到EXCEL中、

步骤五：在EXCEL中制作表格，图表以及相关的结果。

步骤六：发出带有附件（Report结果）的邮件

步骤七：压缩结果并删除之前的数据。

VBA代码需要完成对每一步骤每一个模块的分别编写。将需要执行的SQL代码直接添加到VBA代码中，注意每行不可好过255个英文字符。

Private Sub Workbook\_Open()

'Application.AskToUpdateLinks = True

'Stop

Dim wb As Workbook

'FileN = "dashboard " & Format(Date, "yyyy-mm-dd") & ".xls"

FileN = "dashboard.xls"

FilePath = ThisWorkbook.Path & "\processing\"

Workbooks.Open Filename:=(FilePath & FileN), UpdateLinks:=True

Set wb = Workbooks(FileN)

'ThisWorkbook.UpdateLink Name:=FilePath & FileN, Type:=xlExcelLinks

'Workbooks(FileN).Close savechanges:=False

Call createTables(wb)

Call createCharts(wb)

Workbooks(FileN).Close savechanges:=False

Workbooks(1).Close savechanges:=False

'Application.EnableEvents = True

'Application.DisplayAlerts = False

End Sub

Sub createCharts(wb As Workbook)；

此外，当需要从数据仓库中将数据添加到EXCEL中制作Pivot Table，且需要重复操作的可以采取如下方法：对于制作好的Pivot Table进行数据仓库连接操作。数据🡪属性🡪定义，如图4.23所示。连接字符串：数据库，用户名，密码的设置，用于数据库连接；命令文本：输入读取数据的SQL代码（仅限读取功能，无更新功能）。数据库连接代码如下：

"Provider=MSDASQL.1;Persist Security Info=True;Extended

Properties="DSN=mozart;UID=dadong;Upassword=Apple123;DATABASE=access\_views; AUTHENTICATION=TD2;" 。此操作引入了一段VBA操作，通过对格式固定的Pivot table 采取以上操作，可以简化工作流程，只需待数据表更新完成后，直接刷新数据透视表（Pivot table）既可以完成对数据加载，对数据透视表的更新操作。本操作是半自动的完成数据报告的更新。

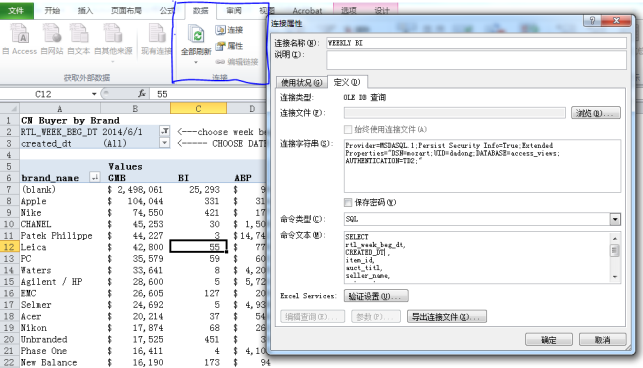


图4.23 数据透视表连接数据仓库

### 4.3.4 基于Tableau的可视化

Tableau可以数据完成精美的数据可视化，相当于高级版的Pivot table并通过Tableau Server发布到Interent，登陆E公司的内网便可以在全球各地对报告进行访问。并可以根据自己的需求，选择对数据进行筛选观察。可以通过Tableau制作热销产品中的重点产品表现报告，以图表的格式。作为对热销产品筛选的途径。

具体步骤：(1)Tableau与数据仓库相连接，从指定表中读取数据；(2)对数据进行分类，measures标尺（数字型）和Dimensions维度；(3)对数据进行X,Y轴摆放，形成对应关系；(4)数据操作：进行公式编辑，增加filter，对图表中增加Marks等；(5)可以进行图展示，也可展示具体数字的表格；(6)可以将多个相关联的报告放在统一的仪表板中，进行数据的动态相连以及互动。完成了为E公司全部获取流量的网站的统计并进行图表展示。可以通过Tableau制作热销产品中的重点产品商业表现报告，以图表的格式，作为对热销产品信息展示的途径。

## 4.4 结果展示

采用上述模型对数据修正后得出的结论如下表4-3所示：

表4-3 模型效果展示

|  |  |  |  |
| --- | --- | --- | --- |
| 适用数据 | 买家数量 | GMB | BI |
| 修正前 | a | b | c |
| 修正后 | 1.36a | 1.5b | 1.44C |

表格中数据的数据来源为：E公司数据仓库内基于14年上半年历史数据。从上述对比中可以看出，经过采用上述模型对中国消费者信息进行甄别后，可以将中国消费者真正消费的商品更加全面的搜集，展示，分析，提高的比例高达50%。基于历史购买记录获得的热销产品信息，一个星期之内中国消费者购买产品中按照类别分类，按照类别与时间进行分类，各个类别按照总销售额进行分类获得结果如下图4.24所示：

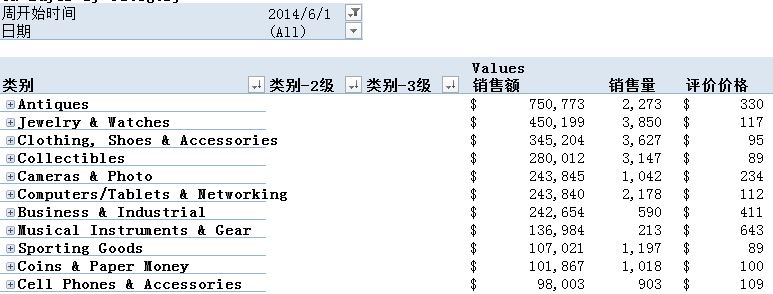


图4.24 热销产品类别信息

此数据为一个星期的数据，此外，按照品牌，具体标题等分类与之结果形式相似。

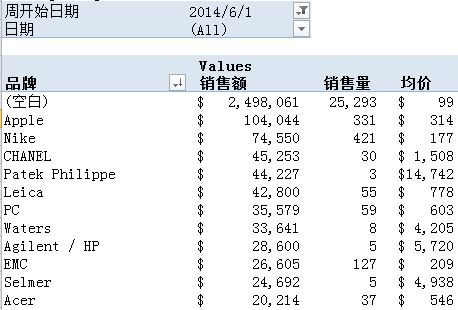


图4.25 热销产品品牌信息

上图4.25对于热销产品的品牌进行展示，可见苹果代表的电子系，NIKE代表的运动系，CHANEL代表的奢侈品系产品是海淘热门产品。上述表格中将产品细化到具体的标题维度，综合上述表格中信息可以获得一个星期内热销产品的具体信息，包括类别，类别的多级分类(META，LV2，LV3)，品牌信息，此外可以细化到具体每个类别，每个品牌的那一款商品是热销的。本表格中较多的补充信息存放在附录B之中。其中，AUCT\_TITL代表商品的标题，ITEM\_ID代表商品编码，CAT\_LVL3\_NAME，代表商品三级分类，GMB，代表商品销售额。具体结果如下图4.26所示。

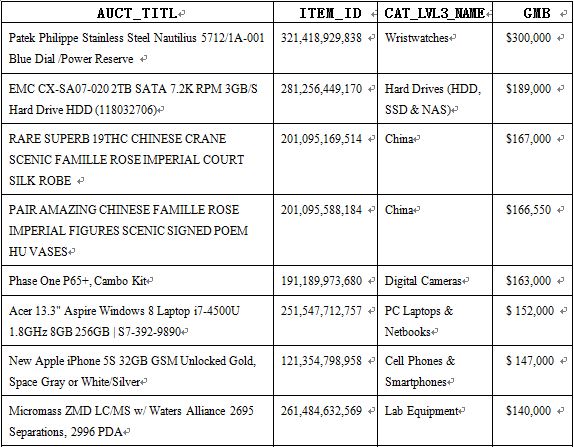


图4.26 热销产品详细信息

此外，可以通过Tableau制作热销产品中的重点产品表现报告，以图形式，作为对热销产品筛选的途径。并通过tableau server发布到internet上，如图4.27所示，便于全球范围内在线访问。

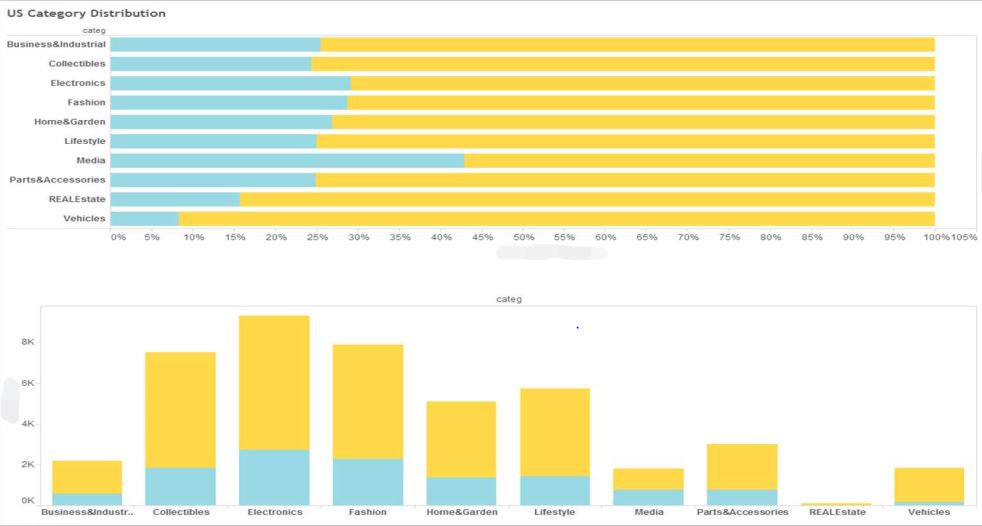


图4.27 Tablueau热销类别对比数据图

上图展示了热销品类CAT在13年以及14年的销售数据对比。可见，热销商品类别在14年较13年销售额存在较大的进步。较为突出的是电子产品，销售额进步较大。最终将采取上述方法获得的热销产品中，符合“海淘”标准的进行筛选。综合两类模型分析得到的结果，以内部信息为主，外部信息为辅，进行热销产品筛选。最终将选定的热销产品用于中文页面展示，并进行相应的营销活动。图4.28表示的中文页面展示的商品信息来源于英文网站的具体商品，本网页展示的商品信息会随着时间以及活动安排进行变换。筛选商品基于上述模型以及海淘这项跨境交易的具体特征以及要求进行。

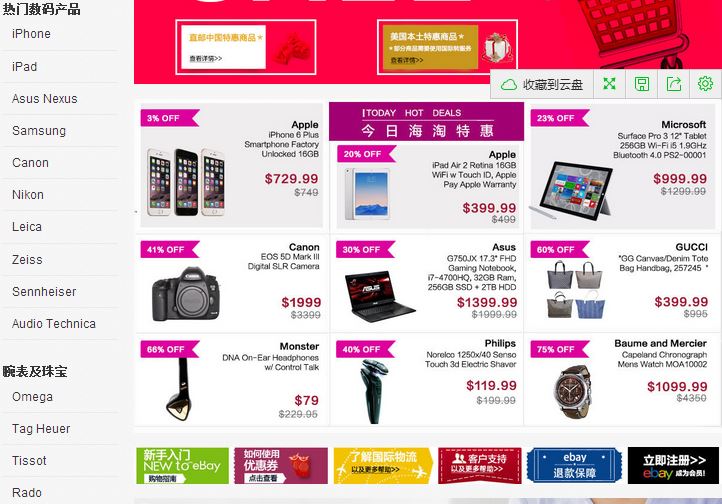


图4.28 中文页面信息展示图

## 4.5 本章小结

本章基于企业内部数据仓库TEARDATA储存的海量数据进行分析，完成“海淘”热销产品识别，主要建立了两类模型：

第一类模型：基于历史购买信息按照“海淘”流程进行数据挖掘得出中国消费者购买产品的热销产品信息。本模型将中国消费者消费的信息识别准确率提高了很大的比例，获得了更加准确的中国消费者购买信息。最终获得热门商品的类别，品牌，标题等具体信息。

第二类模型：基于TERADATA存储的海量信息寻找供求之间差距较大的商品关键词。找出了中国消费者对于每一类别中的商品搜索较多的关键词，用于进行热门产品筛选的辅助支持。

完成逻辑代码设计后，依据实际需求，设计并实现了自动化程序：包括基于SAS自动化代码更新运行；基于SPOON，VBA进行自动化的数据抽取，转换，加载，自动化报告的制作发送。此外，采用Tableau进行报告可视化设计，制作可以通过Internet全球访问的联动仪表板dashboard。最后，综合分析上述两个模型，获得“海淘”热销产品信息，用于“海淘”中国市场扩展商业活动。

# 第五章 企业外部信息识别方案设计

本章将完成以“什么值得买(SMZDM)”“海淘”论坛为研究对象，识别“海淘”热销产品信息的整套方法的分析与设计。如图5.1所示：将采用爬虫以及文本挖掘技术对SMZDM论坛帖子内容进行分析，最终获得热销产品信息。包括：(1)采用SCRAPY爬虫框架进行论文帖子内容抓取，文本预处理，将数据存储在XML文件中；(2)采用“结巴分词”进行分词处理；(3)利用PYTHON进行文本挖掘，基于字典的词频统计；(4)将得到的结果进行汇总分析，制作商业报告；(5)最终获得“海淘”热销产品相关信息。

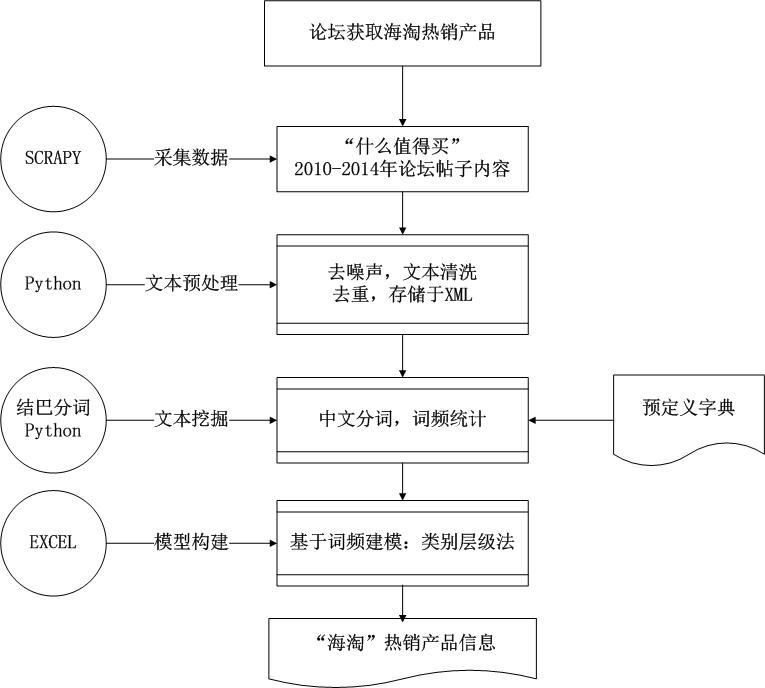


图5.1 论坛获取信息整体流程

## 5.1 数据抓取

### 5.1.1 数据抓取准备

本章主要采取网络爬虫SCRAPY针对“海淘”热门论坛，什么值得买（SMZDM）论坛进行数据抓取。SMZDM（什么值得买）是一个中立的、致力于帮助广大网友买到更有性价比的商品的指南网站,每天推送国内与海外的优惠商品、促销活动、消费资讯。“什么值得买”的内容主要包括(1)“优惠精选”、(2)“海淘专区”、(3)“发现频道”、(4)“晒物广场”、(5)“经验盒子”几个频道，其中“晒物广场”是“海淘”这种“网购”产品的经验分享平台。 “晒物广场”用于展示真实的产品和购物信息 ，其展示的商品中大部分是“海淘”相关的晒物，涵盖电脑数码类产品、服饰鞋包类、运动户外等各大品类，已成为网友“海淘”购买前参考的重要“数据库”；“经验盒子：用于分享“海淘”的各种购物经验。

本章将抓取“晒物广场”与“经验盒子”中的帖子内容，基于上述内容可以获得消费者通过“海淘”购买的商品信息，用于后续分析。经过对“什么值得买”网页分析，得出本论坛的页面逻辑简单。针对每个帖子，只需要将其全部“下一页”的信息抓取即可获得全部帖子信息。因此，本文采用深度优先策略进行网页抓取，将网页中存在噪声的部分，如广告，推荐等进行标签清洗将其删除，帖子的主体内容以及回帖内容作为分析的对象。

### 5.1.2 SCRAPY爬虫的工作流程

第三章第三节已经详述过爬虫框架SCRAPY的基本工作过程。此处将直接进入工作流程介绍：

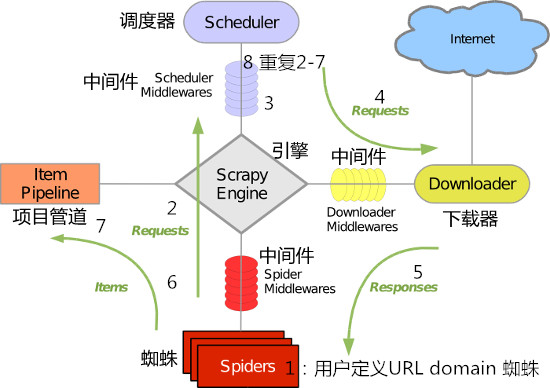


图5.2 SCRAPY工作流程

引擎：SCRAPY Engine，调度器：Scheduler，下载器：Downloader，蜘蛛：Spiders，项目管道：Item Pipeline，下载器中间件：Downloader Middlewares，蜘蛛中间件：Spider Middlewares，调度中间件：Scheduler Middlewares。本章采用深度优先策略DFS (Depth -First- Search，3.3.2节抓取策略中有详细介绍)进行数据抓取工作，工作步骤如下：

(1)用户定义目标URL（网址），Domain（域名），同名蜘蛛；

(2)引擎打开域名，找到蜘蛛，向蜘蛛请求URL并发送给调度器；调度器将URL压入队列中；引擎再次向调度器请求，调度器返回一个URL给引擎；

(3)引擎将URL传送给下载器（下载器中间件确定请求方向），下载器下载；

(4)下载器将返回的网页（Response）传送给蜘蛛进行网页处理（蜘蛛中间件将确定请求方向）；

(5)蜘蛛收到Response 后返回Scraped items并产生新的请求给引擎；

(6)引擎传送Scraped items给项目管道（进行详细分析，过滤以及存储操作），将新的请求传给调度器；

(7)重复步骤(2)直到调度器没有请求产生为止，引擎将关闭域名，抓取数据结束；

### 5.1.3 爬虫功能及模块设计

本节将讲述建立爬虫的过程，以爬取SMZDM什么值得买网站为例,具体步骤包括：

(1)创建一个SCRAPY项目：smzdmDmoz；

(2)定义提取的Item；

(3)编写爬虫 spider 并提取 Item；

(4)编写 Item Pipelines 来存储提取到的Item(即数据)；

(5)编写中间件；

首先：创建一个SCRAPY项目：smzdmDmoz;命令：SCRAPY startproject smzdmDmoz此时系统将会产生:

smzdmDmoz/

SCRAPY.cfg 🡨项目的配置文件

smzdmDmoz/

items.py 🡨项目的item文件

pipelines.py 🡨项目的项目管道文件

settings.py 🡨项目的设置文件

spiders/ 🡨存放蜘蛛代码的文件

其次，对蜘蛛（spiders）模块进行设置；

包括：domain\_name(smzdm.com)、URL(http://www.smzdm.com/)，其中提取下一步链接的规则（深度优先策略）以及从网页中采集数据的规则在parse方法中进行实现；此方法是回调函数（本函数用于初始的request获得response返回后调用），此处需要对页面进行解析（由于SMZDM页面简单，只需要提取下一页，抓取完全部帖子信息即可），返回(1)需要继续抓取本帖子的下一页(2)需要对数据的两种结果进行存储； 从页面中采集数据环节：首先需要清楚网页结构；其次：明确需要被抓取信息的网页位置。需要抓取的信息包括论坛帖子发送的用户以及对帖子进行回复的用户的全部个人信息，以及讨论回复的信息进行具体的记录，此外需要对回复的用户针对每个讨论回复的顺序进行排号。最后利用XPath描述采集信息的路径，并用HtmlXPathSelector完成将信息传输到指定的数据存储模块。

第三，对Item以及Pipeline进行定义；

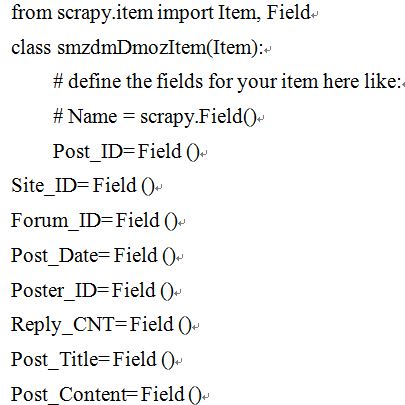


图5.3 Item模块设置

蜘蛛分析出的数据需要保存到Pipeline中必须在Item中进行定义，以便数据后期在Pipeline中进行详细分析，过滤，以及存储操作。本模块需要完成对字段的定义，此处需要参考下一节的数据结构。Pipeline接收Item数据后需要进行以下处理：清除无用的HTML，检查前面步骤数据的合法性，去重，存储数据到本地。本项目将数据存储在XML文件中。

第四：完成对中间件的设置；“中间件”用于完成两个模块中数据流传递，设置需要三个步骤：根据功能需要，选择中间件，在已有工程目录中增加中间件，并在配置文件中增加中间件参数。

最后：setting设置；



图5.4 Setting模块设置

此模块中定义了蜘蛛的全部参数，并设置参数将item，pipeline加载进来。

### 5.1.4 数据采集问题解决

(1)数据存储问题：本工程设计数据量较大，结合处理时间等因素的考虑，选取XML格式存储数据（这些数据已经进行了初步的标签清洗，但是还需要进一步处理才可以进行文本挖掘）。初步得到的数据如下：



图5.5 数据采集结果

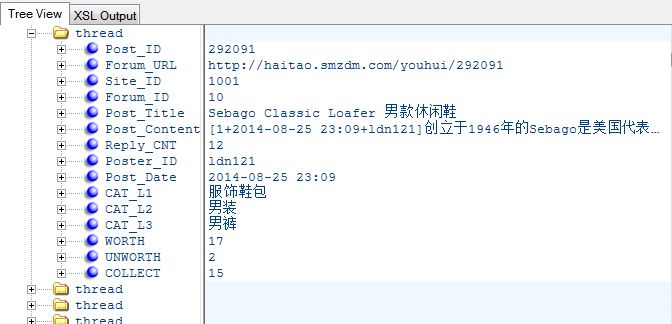


图5.6 XML数据存储结果

(2)下一页处理：

SMZDM的“晒物广场”里页面分别由帖子，回帖组成。例如，帖子是：http://shaiwu.smzdm.com/p/161345因为本帖子回帖较多，故存在下一页。例如：

http://shaiwu.smzdm.com/p/161345/p2#comments，初始页面增加了/p2#comments再下一页http://shaiwu.smzdm.com/p/161345/p3#comments，p2变为了p3。同样的规则适合所有的帖子页面。此处需要蜘蛛按照这个规则进行组合，得到“下一页”链接的URL，发送给Engine继续进行数据抓取。

(3)抓取策略：

本章采取深度优先(DFS)方式对SMZDM论坛进行数据抓取，针对每一个帖子将全部数据抓取出来，再进行下一个帖子的抓取。时间段为2010-2014年这五年期间的全部数据。

(4)访问频率限制：

SMZDM网站属于一个中型网站，其服务器的健壮性不如大型网站，该网站设置了对于同一个IP地址请求频率的限制，经过测试其限制频率为4-5S左右，因此为了防止被封IP，设置了访问时间延时为5S。

(5)访问安全性设计：

对中型网站进行数据抓取过程中可能产生a、硬件问题（计算机当机），b服务器访问失败。针对上述问题本工程进行了如下设计：a:对中间件“Scheduler Middlewares”，“Downloader Middlewares”进行设置，增加定时器，定时对已经抓取的数据进行备份操作；b:根据下载器接受服务器反馈的Http状态码进行判断，如果为“200”代表服务器响应成功，否则代表相应失败，此时Scheduler Middlewares“中间件”会对此时的参数（调度器中的请求队列中的请求，本请求的参数，数据抓取的环境）进行记录，待网页恢复相应后继续抓取。

## 5.2 数据结构设计

经过爬虫采集的论坛内帖子内容，经过文本预处理，文本挖掘等处理步骤，之后采用已有的字典进行词频统计等相关操作后的结果存入到数据仓库TERADATA中，数据存储的结构如下图5.7所示。数据结构包括三类：帖子索引表，关键字索引表，关键字统计表；数据抓取结果将按照下表进行分类，粗体标记的为表的主键。(1)帖子索引表：抓取文本后进行文本数据归类，数据结构梳理之用。(2)关键字索引表：对关键字进行编码，便于管理。(3)关键字统计表：记录各类关键字词频统计的结果。TERADATA提取出的品类，品牌等词库类表用做词频统计依据的数据字典。在图5.7中，中间表为从论坛中采集帖子内容获得的数据，索引表主要内容来自于数据仓库TERADATA中存储的数据，依据所以表对于帖子内容进行文本挖掘，得出上边热销产品信息以及主题关键词统计结果。

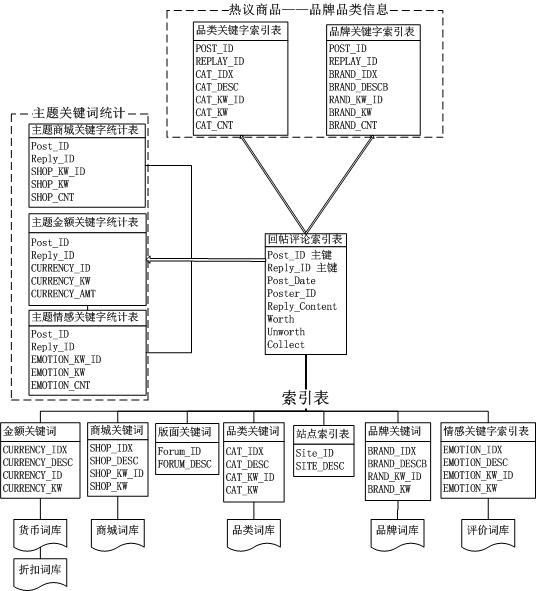


图5.7 数据库表结构

表5-1至表5-3展示爬虫从论坛中采集数据的结果，表5-4至表5-8用于展示从数据仓库TERADATA中得出的各种关键词的索引表。表5-9至表5-11用于存储依据字典进行词频统计的结果。

表5-1 回帖评论索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **Post\_ID** | Num | 帖子唯一编号 |
| **Reply\_ID** | Num | 回帖编号，1对应楼主，2对应沙发，3对应板凳等 |
| Post\_Date | Date | 回帖日期 |
| **Poster\_ID** | VarChar | 回帖人ID |
| Reply\_Content | VarChar | 每个回帖的完整内容 |

表5-2 站点索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **Site\_ID** | Num | 站点编号，如1对应ETAO，2对应LETSEBUY，3对应SMZDM等 |
| SITE\_DESC | VarChar | 站点文字说明，如ETAO，LETSEBUY，SMZDM等 |

表5-3 版面索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **Forum\_ID** | Num | 版面编号或分类编号，如1对应“境外网购攻略--新手指南”版面，2对应“海淘经验汇报--香港区”版面等 |
| FORUM\_DESC | VarChar | 版面文字说明，如境外网购攻略--新手指南，海淘经验汇报--香港区等 |

关键字索引表

表5-4 品类关键字索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **CAT\_IDX** | Num | 总品类唯一编号，如1：电脑数码，2：母婴等 |
| CAT\_DESC | VarChar | 总品类说明，如母婴，包包 |
| **CAT\_KW\_ID** | Num | 品类关键字编号，如1：奶粉，2：钱包等 |
| CAT\_KW | VarChar | 品类关键字，如奶粉，钱包 |

表5-5 品牌关键字索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **BRAND\_IDX** | Num | 总品牌唯一编号，如1：爱她美，2：喜宝等 |
| BRAND\_DESC | VarChar | 品牌说明，如爱她美，喜宝 |
| **BRAND\_KW\_ID** | Num | 品牌关键字编号，如1：爱她美，2：Aptamil 等 |
| BRAND\_KW | VarChar | 品牌关键字，如爱她美，Aptamil 等 |

表5-6 商城关键字索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **SHOP\_IDX** | Num | 总商城唯一编号，如1：Amazon，2：eBay等 |
| SHOP\_DESC | VarChar | 商城说明，如Amazon，eBay等 |
| **SHOP\_KW\_ID** | Num | 商城关键字编号，如1：美亚，2：日亚等 |
| SHOP\_KW | VarChar | 商城关键字，如美亚，日亚等 |

表5-7 情感关键字索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **EMOTION\_IDX** | Num | 感情色彩唯一编号，如1：负面，2：中性，3：正面 |
| EMOTION\_DESC | VarChar | 感情色彩说明，如负面，中性，正面 |
| **EMOTION\_KW\_ID** | Num | 感情关键字编号，1：投诉，2：杯具，3：恶心 |
| EMOTION\_KW | VarChar | 感情关键字，如投诉，杯具，恶心等 |

表5-8 金额关键字索引表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **CURRENCY\_IDX** | Num | 货币唯一编号，如1：美金，2：欧元，3：日元 |
| CURRENCY\_DESC | VarChar | 货币说明号，如美金，欧元，日元等 |
| **CURRENCY\_ID** | Num | 货币关键字编号，如1：美刀，2美金，3：$ |
| CURRENCY\_KW | VarChar | 货币关键字，如美刀，美金，$等 |

关键词统计表：

表5-9 主题商城关键字统计表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **Post\_ID** | Num | 帖子唯一编号 |
| **Reply\_ID** | Num | 回帖编号，1对应楼主，2对应沙发，3对应板凳等 |
| **SHOP\_KW\_ID** | Num | 帖子涉及的商城（销售网站）关键字编号，如1对应美亚，2对应ebay等。多个关键字分多条记录 |
| SHOP\_KW | VarChar | 帖子涉及的商城关键字，如美亚，ebay。多个关键字分多条记录 |
| SHOP\_CNT | Num | 该商城关键字出现的次数 |

表5-10 主题情感关键字统计表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **Post\_ID** | Num | 帖子唯一编号 |
| **Reply\_ID** | Num | 回帖编号，1对应楼主，2对应沙发，3对应板凳等 |
| EMOTION\_KW\_ID | Num | 帖子涉及的感情关键字编号，如1对应投诉，2对应推荐等。多个关键字分多条记录 |
| EMOTION\_KW | VarChar | 帖子涉及的感情关键字，如投诉，推荐。多个关键字分多条记录 |
| EMOTION\_CNT | Num | 感情关键字出现的次数 |

表5-11 主题金额关键字统计表

|  |  |  |
| --- | --- | --- |
| 字段名 | 字段类型 | 字段说明 |
| **Post\_ID** | Num | 帖子唯一编号 |
| **Reply\_ID** | Num | 回帖编号，1对应楼主，2对应沙发，3对应板凳等 |
| CURRENCY\_ID | Num | 帖子涉及的货币关键字编号，如1对应美刀，2对应美金，3对应RMB等。多个关键字分多条记录 |
| CURRENCY\_KW | VarChar | 帖子涉及的货币关键字，如美刀、欧、日元、RMB等 |
| CURRENCY\_AMT | Num | 该货币的总金额 |

由于本项目的针对性较强，对于爬虫项目提供了主题词库。包括商城词库，品牌词库，类别词库，情感词库。上述词库来自于数据仓库TERADATA 以及国内主流电子商务网站分类列表。此外，折扣信息表以及感情色彩信息表来自于主流“海淘”论坛专业术语。

表5-12 部分商城词库

|  |  |  |
| --- | --- | --- |
| 商城词库 | 文字编码 | 数字编码 |
| EBAY | SMSM0000 | 10000000 |
| “亿贝” | SMSM0001 | 10000001 |
| 易趣 | SMSM0002 | 10000002 |
| PAYPAL | SMSM0003 | 10000003 |
| 贝宝 | SMSM0004 | 10000004 |
| 亚马逊 | SMSM0005 | 10000005 |
| 美亚 | SMSM0006 | 10000006 |
| 美国亚马逊 | SMSM0007 | 10000007 |
| 日亚 | SMSM0008 | 10000008 |
| 日本亚马逊 | SMSM0009 | 10000009 |
| AMAZON | SMSM0014 | 10000014 |
| AMZ | SMSM0015 | 10000015 |
| VITACOST | SMSM0016 | 10000016 |
| VT | SMSM0017 | 10000017 |
| DRUGSTORE | SMSM0018 | 10000018 |
| DR | SMSM0019 | 10000019 |
| IHERB | SMSM0020 | 10000020 |

表5-13 部分品牌 词库

|  |  |  |
| --- | --- | --- |
| 品牌 | 文字编码 | 数字编码 |
| ASICS （爱世克斯） | BCBC0000 | 01000000 |
| Armani Exchange （A|X） | BCBC0001 | 01000001 |
| Adidas （阿迪达斯） | BCBC0002 | 01000002 |
| Aptamil （爱他美） | BCBC0003 | 01000003 |
| ASOS (US)  （ASOS） | BCBC0004 | 01000004 |
| ArcTeryx （始祖鸟） | BCBC0005 | 01000005 |
| AVENT （新安怡） | BCBC0006 | 01000006 |
| Apple （苹果） | BCBC0007 | 01000007 |
| Aubrey Organics奥泊丽有机 | BCBC0008 | 01000008 |
| Abbott （雅培） | BCBC0009 | 01000009 |

表5-14 部分类别词库

|  |  |  |  |
| --- | --- | --- | --- |
| LV1\_CAT | LV2\_CAT | 文字编码 | 数字编码 |
| -孕妇保健- | 维生素 | BMMH0000 | 02010000 |
| 01 | DHA | BMMH0001 | 02010001 |
|  | 钙 | BMMH0002 | 02010002 |
|  | 镁 | BMMH0003 | 02010003 |
| -奶粉辅食- | 一段奶粉 | BMMF0000 | 02020000 |
| 02 | 二段奶粉 | BMMF0001 | 02020001 |
|  | 三段奶粉 | BMMF0002 | 02020002 |
|  | 孕妇奶粉 | BMMF0005 | 02020005 |

表5-15 评价词词库

|  |  |  |
| --- | --- | --- |
| 评价词 | 文字编码 | 数字编码 |
| 喜欢 | CWCW0000 | 10000000 |
| 很好 | CWCW0001 | 10000001 |
| 好 | CWCW0002 | 10000002 |
| 支持 | CWCW0018 | 10000018 |
| 特别 | CWCW0019 | 10000019 |
| 快 | CWCW0020 | 10000020 |
| 慢 | CWCW0021 | 10000021 |

表5-16 折扣信息词库

|  |  |  |
| --- | --- | --- |
| 折扣信息 | 文字编码 | 数字编码 |
| % off | DIDI0000 | 11000000 |
| Discount | DIDI0001 | 11000001 |
| Reduction | DIDI0002 | 11000002 |
| Rebate | DIDI0003 | 11000003 |
| \*折 | DIDI0010 | 11000010 |
| 折扣价 | DIDI0011 | 11000011 |
| 半价 | DIDI0012 | 11000012 |
| 减价 | DIDI0013 | 11000013 |
| 降价 | DIDI0014 | 11000014 |
| 甩卖 | DIDI0015 | 11000015 |
| 酬宾 | DIDI0016 | 11000016 |
| 优惠 | DIDI0017 | 11000017 |
| 买一送一 | DIDI0018 | 11000018 |

表5-17 货币信息词库

|  |  |  |
| --- | --- | --- |
| 货币 | 文字编码 | 数字编码 |
| 美元 | CMCM0000 | 12000000 |
| 美金 | CMCM0001 | 12000001 |
| 美刀 | CMCM0002 | 12000002 |
| 刀 | CMCM0003 | 12000003 |
| $ | CMCM0004 | 12000004 |
| dollar | CMCM0005 | 12000005 |

## 5.3 文本挖掘

### 5.3.1 文本预处理

文本预处理包括“分词”，“消停词”，“去噪声”，“清洗”，“去重”操作[50]。

第一、分词；英文分词通常以空格作为标志，切割词，删除数字、分割符，转换为小写，去除停用词如介词、助动词等，同根替换。中文分词需要充分的考虑以下因素：分词的准确率，功能完备性，分词速度，工具的可扩充性，分词软件的易维护性，可移植性等因素。综合考虑各种因素之后，选择PYTHON的结巴分词模块进行文本分词。本文采用PYTHON的“结巴分词”模块对抓取的数据结果进行分词处理。PYTHON模块基于HMM“隐马尔可夫模型”VITERBI算法进行文本分词操作。结巴分词包括三种分词模式，对“全模式”“精确模式”“搜索引擎模式”进行对比，文本例子：我来到北京清华大学。

全模式：我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学

精确模式：我/ 来到/ 北京/ 清华大学

搜索引擎模式 ：我/ 来到/ 北京/ 清华/ 华大/ 大学/ 清华大学

考虑到分词的粒度与词义理解程度，“精确模式”更匹配本文对于分词的需求，故本文采用精确模式。

第二、消停词；文本中含有无意义的“啊 呀”等词语或者标点符号，自定义停用词字典，内容包括语气词，表情词，等。采用本文采用结巴分词模块调用停用词字典进行消除停用词处理。

第三、文本去噪声、清洗、去重；网页中存在的如广告、版权、导航栏、友情链接等部分属于与主题无关的内容，称之为网页噪声，多为模板噪声，可分类为全剧噪声与局部噪声[51]。局部噪声：广告，导航栏，交互插件，友情连接等，图片，表情。全剧噪声：需要分析很多网页才能确定的噪声，多由转载造成，包括网页转载，冗余对象，镜像站点等。



图5.8 数据清洗

文本清洗源自于数据清洗，指为提高数据质量而进行数据修复，删除不合规范的数据，一致性检查，有效性检查，缺失处理等操作，文本清洗则指将与主题无关文本进行去除操作，获得符合文本挖掘的数据[52]。本章中需要进行的文本去噪声，清洗，去重等操作都在项目管道（item pipeline）中进行。

表5-18 项目管道任务

|  |  |
| --- | --- |
| 名称 | 操作 |
| 文本清洗 | 对获得的HTML网页的数据进行清洗 |
| 文本验证 | 对解析的数据进行验证，检查是否包含需要的字段 |
| 文本去重 | 检查是否具备重复数据 |
| 文本存储 | 将解析到的数据进行存储（本文：XML文件中） |

此外，本章中的文本清洗包括两部分：(1)、抓取数据过程中：SMZDM网页上存在广告宣传栏，友情链接，网页导航栏等网页上的一些内容并不是本项目需要的信息，在解析网页时将其删除，即文本清洗操作。

(2)、获得的数据进行清洗：包括无关数据的一致性，有效性，缺失值处理。例如：此外对最终导入到EXCEL中的数据进行清洗，需要将上述错位的数据进行修正，null改为数字0，并将文字信息进行调整。

### 5.3.2 文本挖掘分析

(1)文本词频统计：

词频统计的基本原理：词是组合稳定的字，字组合出现概率越高或频率越高则成词的可信度越高。目前比较成熟的算法包括：互信息原理，N-Gram统计模型，t测试方法，利用字典统计法。前三者的思路为利用统计方法计算单字组合成词的概率与独立的概率进行比较，如果成词概率较大则组成词。利用字典进行统计的方法，利用字典对整个文档进行扫描，针对每一个词都需要对文档进行遍扫描，计算每个词出现的频率[53]。

本章采用PYTHON对分词后的文本进行词频统计，以空格为标志，此处采用E公司根据公司内部数据仓库提取的：品牌词库，类别词库，情感词库，商城词库等相应的词库字典（词库具体信息参考上一节），参照这些字典对分词结果进行词频统计，得到结果如下：从上述结果中可以看出，基于论坛帖子内容，依据数据字典进行词频统计的结果中，每一个帖子内的词频统计结果如图5.9至图5.11所示。上述信息并非汇总的结果，需要进一步的整体结果需要采用数据透视表，对上述数据进行汇总分析。



图5.9 品牌关键词词频统计



图5.10 类别关键词词频统



图5.11 商城关键词词频统计

对于每个帖子内容进行商城关键词词频统计。上述结果仅为针对每个帖子编号进行统计的结果，如果总体的数据，需要采用数据透视表对上述结果进行汇总分析。

此外，涉及到货币关键词，情感关键词等关键词的词频统计表，此处不再赘述。

词频统计代码如下：

#coding:utf-8

word\_lst = []

word\_dict = {}

with open("中文.txt","r") as f1 ,open("词次数.txt",'w') as f2:

for line in f1:

word\_lst.append(line.split('/'))

for item in word\_lst:

for item2 in item:

if item2.strip() not in "，！。“”" :

if item2 not in word\_dict:

word\_dict[item2] = 1

else :

word\_dict[item2] += 1

for key in word\_dict:

print key,word\_dict[key]

f2.write(key+' '+str(word\_dict[key]))

(2)词频统计汇总

经过词频统计之后，需要对相似词进行合并，如“美元”，“美刀”，“$”,“刀”，“美金”等可以归属于“美元”这一种货币，亚马逊，美国亚马逊，美亚，日亚，日本亚马逊，德亚，德国亚马逊，英国亚马逊，AMAZON，英亚，AMZ全部都属于“亚马逊”。此处采用数据透视表进行分组，将上述数据进行分组汇总处理。

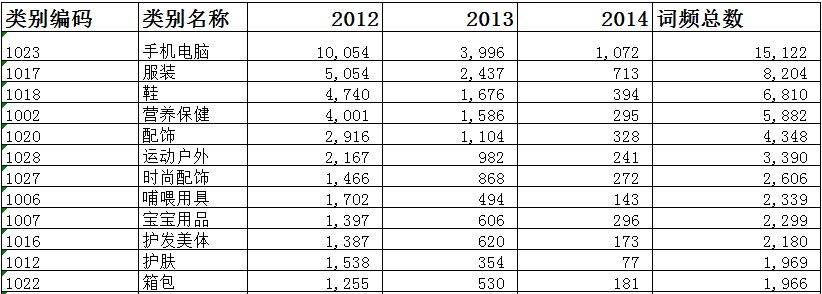


图5.12 类别词频统计结果展示图

(3)文本分析

对经过数据预处理的数据采用统计方法进行词频统计，采用E公司内部数据仓库采集的根据统计词频数对上述类别关键词进行汇总处理。本文主要采取EXCEL中的数据透视表进行数据分析处理，处理结果如图5.12所示。

其中，“手机电脑，服装，鞋，营养保健，配饰，运动户外，时尚配饰，哺喂用具，宝宝用品，护发美体”作为帖子中被讨论频率最高的前十大类别。通过对数据仓库内的表进行连接查询，可以得出具体的产品信息包括如下逻辑：

CAT->META->LV2,3->BRAND->AUCTION\_TITLE。

## 5.4 结果展示

对于热销商品，需要其类别CAT，META，LV2，LV3，具体品牌，相关链接标题等详细的信息，对经过统计分析得出最受热议的商品类别；对排名前三的类别信息进行挖掘，得出相应类别下一层分类LV3的信息，以及被讨论最多的品牌信息。将上述信息存入到TERADATA数据库中，进行挖掘分析，可以得出排名TOP的商品信息，追溯到相关的帖子中可以找出具体是哪一款产品比较热销。对于上述分析数据结果进行图形化表示，得到如下:热销产品中排名前三的商品类别为第一类“手机电脑”，第二类“服装”，第三类“鞋”，上述商品符合“海淘”的特征，满足作为“海淘”热销产品的条件。

本章选择排名第二的信息进行展开分析，得出服装类销售的热门分类以及服装类销售的热门品牌。可完成CAT->META->LV2,3->BRAND->AUCTION\_TITLE的信息采集。

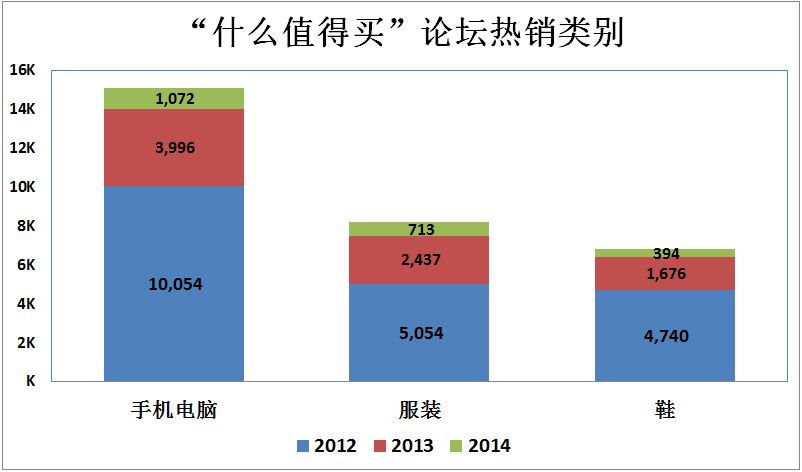


图5.13 热门类别信息

针对热门产品类别中的服装类进行进一步的信息挖掘,获取第三级别的分类信息：

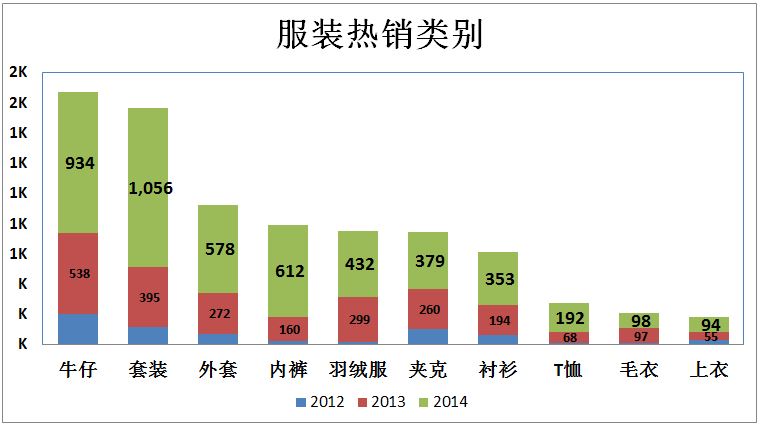


图5.14 服装类热门分类信息

针对服装类产品，进行进一步的信息挖掘，获取热门品牌信息：



图5.15 服装类热门品牌信息

最后，采用词频统计结果，对市场占有率进行分析，可见帖子中讨论的商城中亚马逊占了75%的份额。由此可以推出，消费者进行海淘的商品中，大部分来自于亚马逊外国网站。

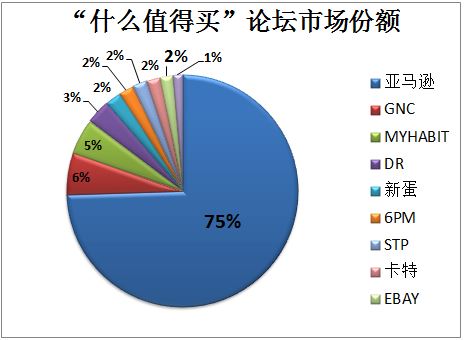


图5.16 商城市场占有率

将上述热销产品包括CAT->META->LV2,3->BRAND->AUCTION\_TITLE的详细信息制作成商业报告的形式，用于支持企业的“海淘”市场商业决策。本节得到的数据将与企业内部数据仓库获得的数据进行合并使用，最终获得海淘热销产品信息。

## 5.5 本章小结

本章针对电子商务企业的实际需求，完成了基于企业外部“什么值得买”论坛帖子内容进行挖掘热销产品信息的方法分析与设计。利用爬虫SCRAPY从“什么值得买（SMZDM）”海淘论坛中采集论坛帖子内容，利用PYTHON进行文本挖掘分析，最终采集“海淘”热销产品详细信息，可用于指导电子商务企业的实际运营工作。

首先，SCRAPY采集论坛数据。分析SMZDM网页结构，论坛内帖子规则，设计被采集的信息数据结构，采用SCRAPY框架进行数据爬取，包括对网页的筛选以及标签清洗。

其次，采用PYTHON进行文本预处理，包括去噪声，文本清洗，去重操作以及PYTHON“结巴分词”模块进行文本分词；再次，采用PYTHON依据E公司提供的类别，品牌，商城，情感等词典进行词频统计。

最后，针对词频统计结果，按照数据结构存入数据仓库TERADATA相应的表中，将获得的数据进行数据分析，获得热销商品类别->初始分类->二三级分类->品牌->标题信息，采用图的形式对上述数据进行可视化处理。得到最终的数据分析结果，用于E公司内部数据仓库进行数据分析时辅助分析。综合企业内外部获得的信息，得到热销产品详细信息，用于开展“海淘”中文页面商业活动 (本文得出的小部分具体数据于附录中进行展示) 。

# 第六章 总结与展望

## 6.1 本文主要工作

在电子商务飞速发展的时代，“海淘”作为电商的又一片蓝海深受各大电子商务企业的重视。基于实际项目：E公司自建“海淘”中文页面与英文网站进行连接，期望提高中国“海淘”市场交易额，识别热销产品信息用于“海淘”中文页面商业活动成为本文研究的内容。本文立足于电子商务企业实际运营中的需求，为获得“海淘”热销产品信息建立了一套信息识别方案，并完成方案设计与实施。本文采用了数据挖掘，文本挖掘方法识别信息，从企业内部与外部同时着手：

基于内部数据：

(1)建立两个模型a、基于“海淘”流程的历史购买数据分析模型，b、基于关键词搜索的供需缺口理论模型，详细讲述了两类模型的思想；(2)基于企业内部数据仓库TERADATA进行数据挖掘分析，完成模型设计，代码设计；(3)进行自动化设置，包括基于SAS的代码自动更新运行，基于SPOON，VBA的商务报告自动制作与发送，并完成基于TABLEAU的数据可视化过程。

基于外部“什么值得买”海淘论坛数据：

(1)选定“什么值得买”SMZDM作为分析对象，利用PYTHON的SCRAPY爬虫框架，采取DFS深度优先策略，按照预先定义好的数据结构进行论坛帖子信息采集。(2)利用PYTHON行预处理操作，利用结巴分词模块（采用HMM模型VITERBI算法）采用精确模式进行文本分词，进行去噪，去重，数据清洗等文本预处理。(3)利用PYTHON按照自定义的数据字典进行词频统计，将获得的信息存入数据仓库TERADATA的表中。(4)对最终结果进行分析汇总，获得热销产品的描述信息，并制作成可视化的图表展示结果。

最后，综合企业内外部数据进行挖掘分析后得到的可视化结果可用于辅助市场营销决策分析。本文是一篇理论应用于实际的典型例子，将数据挖掘、文本挖掘技术应用于电子商务企业的实际工作之中，为电子商务企业识别市场需求的热销产品信息提供指导。

## 6.2 本文创新点

在方案设计与实现的过程中，本文存在以下创新点：

第一：将经济学中“供需理论”的思想应用于对搜索关键词的分析建模中；根据供需理论，提出买家搜索量较大，卖家提供较少的关键词所代表的商品存在较大的市场空间。如果所属的类别适合于“海淘”这种跨境贸易，可以被作为“海淘”热销产品进行中文页面展示。

第二：采用企业内部数据仓库挖掘出的词库进行文本挖掘的词库字典，获得了更加精准的分析方向，提高了针对“海淘论坛”进行文本挖掘的效率与准确率。

第三：本文完成了从逻辑分析，数据获取到商务报告制作的全部自动化设计过程，提高企业的工作效率与准确率，对于企业的实际工作中报告制作自动化提供指导。

第四：本文将跨境电商的分支“海淘”的整体流程，各个环节都进行了详细的讲述，目前对该领域的研究论文较少。

## 6.3 不足与展望

本文实现了理论与实践相结合的，在目前最热门的电子商务市场需求领域进行研究，采取数据挖掘，文本挖掘相结合的方式进行分析，最终获得“海淘”热销商品信息。本文的研究对于电子商务企业的获取市场需求具备重要的意义；对跨境电商运营具备重要的指导作用；对电商的广泛领域具备重要的参考作用。由于研究时间以及条件的限制论文存在以下不足，需要继续研究：

第一：基于企业内部数据建立的模型仅仅针对“海淘”这一市场领域，存在较大的发展空间；第二：企业外部“海淘”论坛数据采集范围有限：仅仅采集了一个论坛的帖子，数据仅为四年数据；第三：仅仅对热销商品信息进行分析，未对“海淘”服务提升提出建议；第四：依赖于E公司内部数据仓库提供了主题词，没有建立自己提取主题词的机制；未实现对相关文本数据的实时爬取；第五：外部信息目前处于辅助内部信息分析的地位，作用有限。

基于上述不足未来可以继续研究的方向包括：

第一、进一步优化基于企业内部数据提出的基于“海淘”流程的历史数据获取模型，以及关键词搜索模型，进一步扩大研究范围，为电商的其他领域提供指导。

第二、拓展文本采集对象范围，优化爬虫算法，获得更多论坛的完整数据；

第三、优化文本挖掘方法，分析服务中存在的问题，为服务提升提供指导；

第四、建立自动聚类学习提取主题词的机制，防止对于主题词库的依赖；

第五、建立一个基于云端的爬虫程序，对相关文本数据进行实时爬取，并建立GUI，便于用户使用和发现近期的海淘热点；

第六、建立并完善企业内外部数据的应用互补机制，提高热销产品获取的准确率。

# 参考文献

[1] 商务部报告：http://www.199it.com/archives/261953.html

[2] 中国报告大厅：http://www.chinabgao.com/k/dianshang/13056.html

[3] 莫岱青. 2014 年中国网络零售行业十大事件盘点[期刊]. 国际商报, 2014, A02:1-10

[4] 中国电子商务研究中心：2014年“海淘”十大热门产品.http://www.100ec.cn/detail--6223203.html

[5] CNNIC：2014年第34次中国互联网络发展状况报告.http://www.199it.com/archives/257541.html

[6] 穆成刚. 我国小额跨境电子商务模式研究[硕士学位论文]. 上海: 上海社会科学院,2014.

[7] 单丹丹. 基于Agent-Swarm仿真的跨境电子商务交易信用优化优化研究[硕士学位论文]. 杭州: 浙江大学, 2014.

[8] 杨璐. 电子商务对我国对外贸易的影响及对策研究[硕士学位论文]. 合肥: 安徽大学, 2014.

[9] 杨颂. 面向电子商务网站的增量爬虫设计与实现[硕士学位论文]. 长沙: 湖南大学,2010.

[10] 胡亚楠. 社交网络数据获取技术与实现[硕士学位论文]. 哈尔滨: 哈尔滨工业大学,2011.

[11] 陈希. 基于R语言数据挖掘的社交网络客户细分研究[硕士学位论文]. 北京: 北京邮电大学, 2011.

[12] 戴支荣. 基于Lucene的面向主题信息搜索系统的关键技术分析及应用[硕士学位论文]. 武汉: 武汉理工大学, 2011.

[13] 王烨. 基于广度优先的主题爬虫的设计与实现[硕士学位论文]. 上海: 复旦大学, 2011.

[14] 何叶青. 网上交易文字评论的买家因素分析——基于淘宝网实证研究[硕士学位论文]. 北京: 清华大学, 2013.

[15] 徐东亮. 基于聚类分析的网络论坛舆情信息挖掘技术研究[硕士学位论文]. 哈尔滨: 哈尔滨大学, 2010.

[16] 代其峰. 基于电子商务的个性化推荐系统研究[硕士学位论文]. 合肥:中国科学技术大学,2011

[17] 刘佳霖.引入互联网搜索量的市场需求预测模型研究[硕士学位论文].北京:北京邮电大学,2013

[18]: Hyunyoung Choi, Hal Varian. Predicting the Present with Google Trends[R]. Technical Report, Google INC,2009.

[19]: Simeon Vosen, Torsten Schmidt. forecasting of Private Consumption: Survey-based Indicator vs Google Trends [J].journal of forecasting.2011:1002-1013

[20]Nicholas Lincoln. The Relationship between Internet Marketing Search Volume and Product Sales

[D], Ohio State: The Ohio State University.2011

[21] Gauri Kulkarni, P.K.Kannan, Wendy Moe. Using Online Search Data to Forecast New Product Sales [J].Decision Support System.52.2012:604-611.

[22]吴海峰.基于多Agent的物流服务提供商推荐系统设计[硕士学位论文].上海.复旦大学.2011.

[23] Lawrence R. D., Almasi G. S., Kotlyar V., et al. Personalization of Supermarket Product Recommendations[J]. Data Mining and Knowledge Discovery, 2011, 5(1/2), 11-32.

[24] Roger D.H. Warburton, J.P.E. Hodgson, E.H. Nielsen. Exact solutions to the supply chain equations for arbitrary, time-dependent demands Original Research [J]. International Journal of Production Economics,2014: 195-205

[25]樊营营.电子商务环境下需求链产品需求预测模型研究[硕士学位论文].哈尔滨.哈尔滨工业大学.2008.

[26]Prins J. Time Series Models. In: NIST/SEMATECH e-Handbook of Statistical Methods[M]. (2003), available at http://www.itl.nist.gov/div898/handbook.

[27] Sanjita Jaipuria, S.S. Mahapatra. An improved demand forecasting method to reduce bullwhip effect in supply chains [J]. Journal of Expert Systems with Applications,2014:2395-2408.

[28] Mehdi Khashei, Ali Zeinal Hamadani, Mehdi Bijari. A novel hybrid classification model of artificial neural networks and multiple linear regression models [J]. Expert Systems with Applications, 2012:2606-2620

[29] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao. Data mining techniques and applications – A decade review from 2000 to 2011[J]. Expert Systems with Applications,2009: 11303-11311

[30] Fomby T B. Stable Seasonal Patter (SSP) Model [D]. Texas.Southern: Methodist University, 2010.

[31] Qihui Lu, Nan Liu. Pricing games of mixed conventional and e-commerce distribution channels [J]. Computers & Industrial Engineering.2013: 122-132.

[32] Kaiquan Xu, Xitong Guo, Jiexun Li, Raymond Y.K. Lau, Stephen S.Y. Liao. Discovering target groups in social networking sites: An effective method for maximizing joint influential power [J]. Electronic Commerce Research and Applications, 2012: 318-334

[33] Sebastian Klapdor, Eva M. Anderl, Florian von Wangenheim, Jan H. Schumann. Finding the Right Words: The Influence of Keyword Characteristics on Performance of Paid Search Campaigns [J]. Expert Systems with Applications,2014: 285-301

[34] 刘宝.基于TERADATA数据仓库的负载均衡查询系统的设计与实现[硕士学位论文]. 上海：上海交通大学, 2013.

[35] 齐红霞.中国移动一级经营分析系统的ETL设计与优化[硕士学位论文]. 北京：北京交通大学, 2011.

[36] 王夕萌. TERADATA数据仓库在邮政金融客户管理系统中的设计与实现[硕士学位论文]. 北京：北京交通大学, 2013.

[37] 戚俊伟. 基于SAS的客户营销相应模型设计[硕士学位论文]. 上海：华东理工大学, 2013.

[38] 王静. 基于SCRAPY的电子商务网络测量与网络特征分析[硕士学位论文]. 北京：北京交通大学, 2011.

[39] 陈晓云.文本挖掘若干关键技术研究[硕士学位论文]. 上海.复旦大学, 2005.

[40] Srivastava and Sahami. Text Mining: Classification, Clustering, and Applications. Available at http://www.doc88.com/p-294947828469.html

[41] 沙有闯. 基于Web文本挖掘的网络口碑监测系统研究[硕士学位论文]. 合肥: 安徽大学, 2010.

[42] 黄文江. 中文文本聚类算法分析与研究[硕士学位论文]. 上海: 上海交通大学, 2010.

[43] 菲尔德曼. 文本挖掘 ISBN: 9787115205353 [M]. 北京: 人民邮电出版社, 2009. 1-145

[44] 康东. 中文文本挖掘理论与应用[硕士学位论文]. 苏州: 苏州大学, 2014.

[45] 程志强. 基于“新浪微博”主题的用户影响力研究[硕士学位论文]. 沈阳: 东北大学, 2013.

[46] 郑树锐. 基于HMM模型的顺式调控模块识别方法的研究[硕士学位论文]. 西安: 西安电子科技大学, 2012.

[47] 胡馨云. 基于属性的商品评论情感挖掘研究[硕士学位论文]. 武汉: 华中科技大学, 2013.

[48] 万接喜. 多语言文本聚类研究[硕士学位论文]. 南京: 南京大学, 2013.

[49] 曾悠. 大数据时代背景下的数据可视化概念研究[硕士学位论文]. 浙江: 浙江大学, 2014.

[50] 杨建武，北京大学计算机科学技术研究所-文本挖掘技术讲义，北京大学2014年春季.http://webkdd.org/course/

[51] 陈玉兰. 基于标签特征向量的网页去噪声研究及其应用[硕士学位论文]. 上海: 华东师范大学, 2010.

[52] 卢明泰. Web数据挖掘及其在社交网络领域的应用[硕士学位论文]. 成都: 电子科技大学, 2012.

[53] 炎士涛. 基于词频统计的文本分类模型研究[硕士学位论文]. 上海: 上海师范大学, 2007.

# 附录

## A 重要代码

A.1 数据结果处理Dashboard的VBA代码

Private Sub Workbook\_BeforeClose(Cancel As Boolean)

Call closeExcel

End Sub

Private Sub Workbook\_Open()

'Application.AskToUpdateLinks = True

'Stop

Dim wb As Workbook

'FileN = "dashboard " & Format(Date, "yyyy-mm-dd") & ".xls"

FileN = "dashboard.xls"

FilePath = ThisWorkbook.Path & "\processing\"

Workbooks.Open Filename:=(FilePath & FileN), UpdateLinks:=True

Set wb = Workbooks(FileN)

'ThisWorkbook.UpdateLink Name:=FilePath & FileN, Type:=xlExcelLinks

'Workbooks(FileN).Close savechanges:=False

Call createTables(wb)

Call createCharts(wb)

Workbooks(FileN).Close savechanges:=False

Workbooks(1).Close savechanges:=False

'Application.EnableEvents = True

'Application.DisplayAlerts = False

End Sub

Sub createCharts(wb As Workbook)

Dim cht As ChartObject

Dim strPath As String

strPath = ThisWorkbook.Path & "\processing\chart"

wb.Activate

For Each cht In wb.Sheets(1).ChartObjects

' added next line to activate - fix export issue in excel 2010

' http://social.msdn.microsoft.com/Forums/eu/isvvba/thread/736eeb5f-7f21-4006-87ee-7748f3c83175

cht.Activate

i = i + 1

fName = strPath & i & ".JPG"

cht.Chart.Export Filename:=fName, FilterName:="JPG"

Next

End Sub

Sub createTables(wb As Workbook)

On Error Resume Next

Dim cht As Excel.ChartObject

Dim tbl1 As Excel.Range

Dim tbl2 As Excel.Range

Dim tbl3 As Excel.Range

Dim strPath As String

strPath = ThisWorkbook.Path & "\processing\"

Set tbl1 = wb.Sheets(1).Range("table1")

Set tbl2 = wb.Sheets(1).Range("table2")

Set tbl3 = wb.Sheets(1).Range("table3")

wb.Activate

'don't use Format:=xlPicture

http://www.pcreview.co.uk/forums/copypicture-method-range-class-failed-using-rangename-t3952149.html

tbl1.Activate

tbl1.CopyPicture xlScreen, xlPicture

Set cht = wb.Sheets(1).ChartObjects.Add(0, 0, tbl1.Width, tbl1.Height)

cht.Activate

cht.Chart.Paste

cht.Chart.Export strPath & "table1.JPG", FilterName:="JPG"

cht.Delete

tbl2.Activate

tbl2.CopyPicture xlScreen, xlPicture

Set cht = wb.Sheets(1).ChartObjects.Add(0, 0, tbl2.Width, tbl2.Height)

cht.Activate

cht.Chart.Paste

cht.Chart.Export strPath & "table2.JPG", FilterName:="JPG"

cht.Delete

tbl3.Activate

tbl3.CopyPicture xlScreen, xlBitmap

Set cht = wb.Sheets(1).ChartObjects.Add(0, 0, tbl3.Width, tbl3.Height)

cht.Activate

cht.Chart.Paste

cht.Chart.Export strPath & "table3.JPG", FilterName:="JPG"

cht.Delete

End Sub

Sub CopyRangeToGIF()

' save a range from Excel as a picture

Dim rng As Excel.Range

Dim cht As Excel.ChartObject

Const strPath As String = "C:\Michael\ETL\dandan's automation\processing\"

Application.ScreenUpdating = False

Set rng = Range("C6").CurrentRegion

rng.CopyPicture xlScreen, xlPicture

Set cht = ActiveSheet.ChartObjects.Add(0, 0, rng.Width + 10, rng.Height + 10)

cht.Chart.Paste

cht.Chart.Export strPath & "myfile.gif"

cht.Delete

ExitProc:

Application.ScreenUpdating = True

Set cht = Nothing

Set rng = Nothing

End Sub

A.2 KSH文件代码：

#!/bin/ksh

jobName='Geox'

jobFolder=/export/home/cac/dandanDong/Geox

loginFile=/export/home/cac/dandanDong/work/logins

loginString=`grep -i mozart $loginFile`

sqlFile="$jobFolder/Geox.sql"

kshFile="$jobFolder/Geox.ksh"

toEmail="lqiang@e.com,daDong@e.com,zhoudong@e.com"

jobDesc="This is a job for \"$Geox - Mozart\".\n\nSQL File: $sqlFile\n\n$Memo\n\n"

jobDesc=$jobDesc"Back End Server: "`hostname`"\n"

jobDesc=$jobDesc"CronJoblog: /export/home/cac/dandanDong/work/schedules/CronJobLog/cronjob.log"

recoveryCmd="\nBack End Recovery Command:\n"

recoveryCmd=$recoveryCmd"nohup $kshFile>>/export/home/dandanDong/work/schedules/CronJobLog/cronjob.log 2>&1 &\n"

logFolder=$jobFolder/LOG

if [ ! -d $logFolder ]

then

mkdir $logFolder

echo "mkdir $logFolder"

fi

logFile=$logFolder/$jobName.`date +"%y%m%d%H%M%S"`'.log'

echo "log file: "$logFile

set +e

bteq <<EOF >$logFile

.SET ERROROUT STDOUT

.set session transaction btet

.logon $loginString

select 'SESS', session;

.maxerror 1

.run file=$sqlFile;

EOF

rcode=$?

set -e

echo $jobDesc>>$logFolder/mailcontent

echo "Script log: "$logFile"\n">>$logFolder/mailcontent

if [[ $rcode -ne 0 ]]

then

subject='Script Failed! BTEQ '$jobName

tail -100 $logFile grep -i fail>>$logFolder/mailcontent

echo $recoveryCmd>>$logFolder/mailcontent

mail -s "$subject" "$toEmail" <$logFolder/mailcontent

rm $logFolder/mailcontent

echo $jobName" fail. Send mail to $toEmail ("`date`")"

exit 1

else [[ $rcode -eq 0 ]]

subject='Script Success! BTEQ '$jobName

mailx -s "$subject" "$toEmail" <$logFolder/mailcontent

rm $logFolder/mailcontent

echo $jobName" succeed. Send mail to $toEmail ("`date`")"

fi

exit 0

## “海淘”热销产品信息结果展示

附表B.1：一周内热销产品具体信息

|  |  |  |  |
| --- | --- | --- | --- |
| AUCT\_TITL | ITEM\_ID | CAT\_LVL3\_NAME | GMB |
| Patek Philippe Stainless Steel Nautilius 5712/1A-001 Blue Dial /Power Reserve | 321,418,929,838 | Wristwatches | $300,000 |
| EMC CX-SA07-020 2TB SATA 7.2K RPM 3GB/S Hard Drive HDD (118032706) | 281,256,449,170 | Hard Drives (HDD, SSD & NAS) | $189,000 |
| RARE SUPERB 19THC CHINESE CRANE SCENIC FAMILLE ROSE IMPERIAL COURT SILK ROBE | 201,095,169,514 | China | $167,000 |
| PAIR AMAZING CHINESE FAMILLE ROSE IMPERIAL FIGURES SCENIC SIGNED POEM HU VASES | 201,095,588,184 | China | $166,550 |
| Phase One P65+, Cambo Kit | 191,189,973,680 | Digital Cameras | $163,000 |
| Acer 13.3" Aspire Windows 8 Laptop i7-4500U 1.8GHz 8GB 256GB | S7-392-9890 | 251,547,712,757 | PC Laptops & Netbooks | $ 152,000 |
| New Apple iPhone 5S 32GB GSM Unlocked Gold, Space Gray or White/Silver | 121,354,798,958 | Cell Phones & Smartphones | $ 147,000 |
| Micromass ZMD LC/MS w/ Waters Alliance 2695 Separations, 2996 PDA | 261,484,632,569 | Lab Equipment | $140,000 |
| Netural antique red coral necklace 210 gram . Very good color | 141,310,871,176 | jewelry | $135,000 |
| 2014 New Hurom Slow Juicer Extractor HH-SBF11 Fruit Vegetable Citrus | 141,291,946,679 | Small Kitchen Appliances | $134,890 |
| AUTH CHANEL CROCODILE PINK DOUBLE FLAP CC LOGO SHOULDER BAG | 171,346,768,364 | Handbags & Purses | $130,000 |

## 文本挖据&数据可视化工具

1. Tablueau官网

http://www.tableau.com/

1. 结巴分词示例

http://jiebademo.ap01.aws.af.cm/extract

1. PYTHON官网

https://www.PYTHON.org/

[4]文本挖掘书籍Text Mining Classification, Clustering, and Applications

http://www.doc88.com/p-294947828469.html

# 致谢

时光荏苒，岁月如梭，北大的研究生学习接近尾声。蓦然回首，二十余载的求学时光即将画上圆满的句号。

首先：衷心地感谢我的导师李杰教授。本文是在我的导师李杰教授的悉心指导、谆谆教诲、严格把关的情况下完成的。在我论文的撰写过程中李杰教授数次对我的论文进行指导、答疑解惑、使我论文的各个方面不断完善。李杰教授在学术方面循循善诱、鼓励创新、尊重他人思想的教育方法使我的论文研究得以深入进行并顺利完成。李杰教授严谨求实的治学态度、精益求精的工作作风、正直诚恳、与人为善的待人风范对我影响极其深远，将指引我今后的工作与生活。李杰教授在指导论文的同时教会了我很多做事做人的准则，这些是更高层面的知识，使我在技术与写作能力提高的同时，领会到了更多人生的哲理与处世准则，将受用一生。同时，感谢我实习公司的项目组长朱一超经理以及高级分析师陈泓，他们为我提供了项目研究条件，使我在项目实践中得到了技术与思想的层面的锻炼与提高。同时言传身教，使我对职业规划与人生有了更深的感悟。

正是以上老师的培养与教诲，使我分析问题的思路不断拓展、解决问题的方法更加科学可行。同时我的技术水平、理论知识和综合素质得以不断提升。

其次：感谢北京大学软件与微电子学院的段莉华老师，孙圣力老师，刘京老师在论文开题答辩时给予我论文选题以及思路的指导、并给出建设性的修改意见，在此表示深深的谢意。

再次，感谢北大软件学院的梅佳师兄和陈杭师兄，实习与论文撰写过程中他们给予我很多关怀与帮助；感谢同门师兄弟们对我论文完成情况的关怀帮助，让我感受到了同学间真挚的爱。感谢父母对我无私的爱，无尽的关怀，他们的关爱对我论文的完成有很大的帮助。

最后，衷心地感谢参与本文评审的各位专家教授学者，感谢您们能在百忙之中抽出宝贵时间审阅本文。

# 北京大学学位论文原创性声明和使用授权说明

**原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

**学位论文使用授权说明**

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
* 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
* 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日