

**内蒙古师范大学计算机科学技术学院**

**毕业设计（论文）开题报告**

**题 目：基于中国图书网经济类书籍的数据分析**

**专 业 ：信息管理与信息系统**

**学 生 ：管贤玉**

**学 号 ：20161102967**

**指导教师 ：苗忠义**

**日 期 ：2019/11/21**

**计算机科学技术学院制**

**说 明**

1. 课题来源及研究的目的和意义

课题来源于想利用学过的知识和现在社会大数据分析的趋势，做一个python的爬虫来抓取一些数据做数据分析。

经济类图书的三大好处文字简洁，案例贴近生活，知识简单易懂。

1）文字简洁，是因为经济类书籍的主要内容基本都写在前言中，且每一处知识都有固定的，配套的专用名词，这些专用名词是在多年的历史长河中流传下来的，是不易更改的，而且深入人们的生活中，类似恩格尔系数，就是食品的消费总额占个人总支出的百分比，GDP，就是国内生产总值，按市场价格计算的一个国家（或地区）所有常驻单位在一定时期内生产活动的最终成果，常被公认为衡量国家经济状况的最佳指标。还有很多耳熟能详的名词。

2）案例贴近生活，比如说“二八原则”，意思是我们只需要花费20%的时间就可以处理或完成80%的事情，也就是我们在生活和学习中需要分清主次，找出主要影响进度的矛盾，并以很高的效率去完成任务。这就是经济学中教授的内容。

另一个经典案例就是啤酒和尿布的故事，沃尔玛超市的员工根据数据分析发现每到周五尿布和啤酒的销量会增加，进一步发现男性消费者群体，进而执行将啤酒摆放在尿布周围的货架上来提高销量这一措施。这也是很实用的案例。

3）知识简单易懂，在浏览相关图书网站时，发现经济类书籍很大一部分介绍知识的都是漫画书，用简单的文字配上生动的图片，更能吸引读者的购买兴趣，观察到有名的书籍有《牛奶可乐经济学》、《和经济学先生约会》等。

这个课题的主要研究目的是在获得网站的书籍数据之后，利用学习过的统计学、管理学、python对这些书籍进行相关频数最高的关键词统计，销量的分析，读者的评论研究，结合现在大众对经济的观点看法，预测经济类书籍的销售趋势，和未来社会经济信息方面的关键词出现时机。

1. 国内外在该方向的研究现状及分析

国内大部分的研究都集中在大数据分析上，相关的有一本书籍《大数据经济学》，百度百科《大数据经济学》是国内第一本关于大数据的经济学著作，指出大数据的本质就是社会经济的离散化解构与重构。《大数据经济学》提出离散主义，指出精神产品大生产与生命时间稀缺性，在此基础上讨论大数据经济学的基本范式，即离散化解构、全息化重构、价值关联。《大数据经济学》探讨了数据理性，论述了理性超越、边界可达与蚂蚁效应，研究了数据成本对企业边界的影响、数据资源的权属。《大数据经济学》最后讨论了大数据平台的构建模式，对人的行为异化等进行了分析，并对未来进行了展望。

也就是利用统计学中的离散知识，例如离散系数，来分析数据是集中还是离散的趋势，重点画的比较广泛，就是关于人类的一切数据，对数据的获取，整合，分析，总结，更复杂的就是针对某一方面的用户画像和数据建模。

国外的大数据分析就更深奥了，2017年4月美国能源部与退伍军人事务部联合发起“百万退伍军人项目（MVP）”，希望借助机器学习技术分析海量数据，以改善退伍军人健康状况。这也是对未来经济发展的一种推进方式，白宫科技政策办公室一直积极与他国展开合作，以预防数字经济监管障碍、促进信息流动和反对数字本地化等。

也就是说现在全球是大数据发展的时代了，且大数据的发展与经济息息相关，基本的出发点就是研究预测新的数据，总结分析旧的数据。

1. 主要研究内容

获取书籍的书名、销量、读者评论、评分、等信息。利用python编写代码获取网页地址，利用x-path获取网页的相关信息，利用线性回归等方法做分析和预测。主要研究在销量前位的书籍的信息和大量出现的经济类关键字，结合现在的经济发展，分析预测未来可能高频出现的词汇，和经济类图书未来的销售趋势。

4. 研究方案

1）首先学习python爬虫的相关知识，了解网络的安全法

2）了解网页的基本结构div，img，href的用法和位置

3）了解x-path和fiddler的使用方法，熟练操作x-path来抓取网页中的有用的信息，观察fiddler获取的网页json报文，并生成相应的CSV、json等文件存到数据库中。

4）回顾学习过的经济学、统计学的知识，用Excel来分析数据，并用Excel制作一个可以使用的用户交互界面。

5）利用一些模型做分析预测。

5．进度安排，预期达到的目标

1）前期学习使用python爬虫，和相关网络安全法律。

2）熟练使用xpath和fiddler获取网页信息

2）中期学习使用Excel的开发者工具，获取5000条以上的数据，并制作出一个可以查询数据的用户交互界面。

3）中后期具体数据具体分析，学习经典的数据模型，对数据进行分析和预测。

4）后期完成论文的编写，制作答辩PPT,整理材料和数据。

6．课题已具备和所需的条件、经费

电脑安装有python、x-path、office2016和谷歌浏览器，可以做简单的数据获取和数据分析，有相应软件的教学课程，书籍等。

7．研究过程中可能遇到的困难和问题，解决的措施

1）在python抓取网站数据的过程中可能会碰到相关网站的封号情况，或者是网站相关路径输入错误造成的数据抓取错误。

解决方案，设计访问网站的username，伪装进站，在代码中加入时间函数，仿照人为浏览网页的情况。网页数据，利用浏览器的开发者选项配合x-path进行数据的精确获取。

2）Excel的使用不当，相关变量分析错误，相关系数计算错误，交互界面的制作不当。

解决方案，利用correl函数分析变量间的相关系数，利用Excel的语言编写宏指令完成交互，显示图表，增删改查。

3）预测方向，方向出问题，回到原始数据中，具体问题具体分析。

8．主要参考文献

[1]薛晨，《国际大数据研究论文的计量分析》-[现代情网]2013年第9期

[2]韩芳芳、范群《我国大数据领域研究论文的计量分析》2013年第8期

[3]李华芳,《思想的盛宴：2007年上半年经济类书籍回顾》-[中国图书评论]第10期

[4]李华芳,《临床经济学的诞生：2007下半年经济类书籍回顾》-[中国图书评论]2008年第10期

[5]马巧红，《经济类书籍借阅率现象透析》

[6]张光芒，《见证了2013年经济轨迹的书籍》-[产权导刊]2014年 第3期

[7]薛永宁，《论“大数据”的应用》-[当代旅游：下旬刊]2019年 第06期

[8]魏凯，《大数据》-[数据通信]2014年 第3期

[9] 王元卓，靳小龙，程学旗,《网络大数据：现状与展望》-[计算机学报]2013年 第6期

[10] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding,《Data Mining with Big Data》2018年

[11] V.Harsha Shastri's, V Sreeprada's,《A Study of Data Mining with Big Data》2016年

[12] Michele Mosca，《Cybersecurity in an Era with Quantum Computers: Will We Be Ready?》-[IEEE security & privacy]2018年 第5期