前景与范围文档

2013.12.15

目录

1、业务需求.....2
1.1 背景.....2
1.2 业务机遇.....3
1.3 业务目标与成功标准.....3
1.4 客户与市场需求.....4
1.5 业务风险.....4
2、解决方案的前景.....5
2.1 前景声明.....5
2.2 主要特征.....6
2.3 假设与依赖.....6
3、范围与限制.....6
3.1 第一个版本的范围.....6
3.2 各后续版本的范围.....6
3.3 限制与排除.....7
4、业务背景.....7
4.1 涉众简介.....7
4.2 项目优先级.....8
4.3 操作环境.....8

1、业务需求

1.1 背景

随着互联网的飞速发展，网络上的信息呈爆炸式增长，根据 CNNIC 最新的统计数据显示，截止到 2011 年底，中国网站的数量已经达到 229.6 万。利用 web 数据挖掘技术，可以将海量数据转化为有用的信息，这些信息可以广泛应用于诸如舆情分析、商业营销、市场调研等诸多方面。为了充分利用这些信息，基于互联网的信息采集、大数据分析与分类聚类等技术已经成为人们研究的热点。

选择需要分析的数据时要考虑一些因素，首先是数据是否好获取，由于目前广泛采用的获取网页数据的方法是利用爬虫程序，而这就与被爬去网站所设计的安全性相关，对于禁止合法抓取的网站，就需要开发者进行周全的考虑；其次是对获取数据的分析，此步骤可能会涉及一些较为先进的技术或是已成形的技术，要充分利用复用思想，如如果想对于论坛评论进行分析，就需要一些中文分词算法，才能将大段内容打散并方便存储和分析；当然最终要的是选取数据进行分析时一定要有一定的现实价值意义，因为这才是技术与真实生活良好的结合。

在 Java 高级技术课程中，为本次项目提出了一些要求如下：一是从网络(问答系统、BBS、即时通信工具、微博等)抓取需要的数据；二是将获取到的数据进行除杂等预处中理操作后存入数据库；三是对数据库中的数据进行挖掘、分类、统计、分析…；四是以适当的形式展示数据处理结果。

我们项目组选取获取数据的对象是电子商务网站，选择的原因有已下几点：一是数据源信息量大，即如果要进行大数据分析和处理，数据量是一个关键的指标；二是数据源种类多，即电子商务的业务较为广泛复杂，数据源的种类可以增加分析的复合性，在使分析更为困难的同时，增加了数据源的可选择性以及分析的准确性；三是使用者广泛，即好的电子商务网站通常拥有巨大的用户量，这可以说明数据源较为可靠且具有研究价值，无论是通过研究用户行为从而帮助商家更好的针对用户开展销售策略，还是帮助研究某种新技术对于产品销售的影响，来帮助商家及时改变技术策略等都起到了重要作用。综上所述，研究这样的数据是较好的选择。

通过上述分析，项目选择京东商城作为数据源获取的网站，选择原因有以下几点：一是

京东商城是目前互联网上排名首位的 B2C 网站，以其配送速度、正品保障赢得了消费者的信赖。当需要抓取商品信息，以及制作比价系统（比价系统可以看作针对商品的垂直搜索引擎）时，京东商城都是不可缺少的重要信息源；二是北京京东世纪贸易有限公司副总裁姜海东在讲演中表示“京东最开始基本上通过采销人员经验做事，但现在整个商品品类已过百万，此时发现真的需要依靠数据挖掘来分析内容趋势，所以京东在数据驱动、挖掘方面有非常大的需求”。

京东商城笔记本电脑商品销量评论数据挖掘展示系统（以下简称系统）是利用 java 语言编写的数据挖掘、分析、展示系统，主要通过抓取京东商城的笔记本电脑类型、品牌、介绍、价格、参数、评论（评论内容、评论者 ID、好中差评论数量），并进行分析，将分析结果展示给商家和用户，帮助商家掌握诸如自家所卖的商品哪种较受消费者欢迎，通过用户评论内容判断自家产品与其他商家同类产品相比的竞争优势和不足，帮助商家了解一款新产品的销售趋势对此作出反应等，帮助想买产品的用户了解目前市场上较受欢迎的产品，以及针对自己对某些配件如显卡、CPU 等特征要求较高来选购目前市场上合适的商品品牌等。如果此次项目产生了较好影响，后续版本可能会添加一些基于整个京东网站，与网站方面协商获取更多较为机密数据，来帮助研究整个京东商城的商品销售和用户行为，甚至来比较分析京东商城与其他同类电子商务网站，从而满足京东商城对于利用数据挖掘分析来进行商业决策的需求。目前项目为 1.0 版本，先从分析较为简单的笔记本电脑商品开始。

1.2 业务机遇

利用数据挖掘技术，来达到更积极帮助企业经营决策以及帮助和影响用户行为的目的，初步结合数据挖掘与推荐，目前京东商城对于这方面的系统较为空白，且有较大需求，这对于本系统具有良好的机遇。

1.3 业务目标与成功标准

系统的目标是对京东商城笔记本电脑的数据进行挖掘、分析、展示，要求符合实际需求，且系统设计简单可行。成功标准是系统所分析的结果对于客户有意义。

1.4 客户与市场需求

该系统主要的目标客户为面向京东商城、出售笔记本电脑的京东商城商家、有购买笔记本电脑意向的京东商城用户。

当前已有的趋势分析图，多半是利用较简单的数据来进行图形化的展示，如淘宝网，淘宝网根据其所有数据，对商家给出了与同行相比其商品描述相符程度、服务态度、发货速度的信息，也给出了某种商品价格趋势图和同种商品的款式和型号销量比例图（销量比例可能受库存影响）。这些展示不能满足的需求包括：1.分析较为简单，得出的商家与同行相比的数据经常不被商家认可，即可能只存在一款产品整体拉低了店铺的形象，而整个商家的形象被拉低了，而用户无法如果只根据行业数据判断商家作为选择某种商品的依据，则很有可能错过这个商家提供的较为好的一款产品；2. 数据分析没有明确的针对性，过于宽泛，
遗憾的是，京东商城在利用展示分析数据帮助用户购买这些提高用户体验的方面做得比较不足，而这也说明在这一方面，京东商城有很大的需求。

客户如何使用本产品：不同角色登陆系统，对权限所能访问的功能进行访问和操作。

1.5 业务风险

- 数据获取风险：数据获取困难，导致项目获取不到大量可靠数据，使分析结果不准确或错误，或是会加重爬虫程序的开发成本。这是由于商品介绍页面字段多、布局和样式复杂；页面内容的加载需要用户交互触发，如需要点击选项卡后，页面才会载入评论信息；由于竞争关系，京东商城禁止绝大多数的合法抓取，因而可以预料到京东商城会对爬虫做出比较强的反应。
- 数据分析风险：由于项目组成员对商品的调查不足或不了解，导致分析方向错误
- 与产品开发相关（或无关）风险：
    1) 市场竞争：目前针对京东基本不存在同类产品，所以市场竞争风险发生的几率较低，如果这项风险发生在本系统开发成功之前有可能对本系统的市场竞争力造成较大冲击，可控力为中等。降低风险的方法包括：实时跟进市场动态，获取竞争公司的动态；与京东方面进行沟通交流，对系统功能做推广，多多听取可控用户和商家需求，根据京东要求及时变更需求及进行需求管理；与多个客户建立良好关系，站稳市场；

2) 时间问题：由于系统目前只针对一类商品，虽然不完善，但是开发快速，并且较小范围的数据分析可以在保证系统功能和性能情况下高质量完成。由于时间导致的风险主要体现在无法协定在 deadline 之前完成，可能造成合作终止的风险，风险发生几率极低，对它的控制能力很强。降低风险的方法包括：加强编码组内部协作交流，加快开发进度；在系统即将无法按照既定日期完成时，及时协商；

3) 用户认可问题：由于本系统只针对某一类商品，无法满足其他商品商家、用户对于系统的需求，但可以将系统作为一个类似于原型的系统，通过风险发生几率中等，可控力为中等，如果这项风险发生，系统面临进行较大修改的风险。降低风险的方法包括：通过简易调查，研究用户对于使用系统的感受、制作模拟演示功能模型在实际使用中进行模拟操作与流程跟踪；及时对用户进行宣传培训；

4) 实现问题：实现问题风险主要包括两方面，主要是系统关键技术实现方面，如，风险发生几率中等，可控力较高。降低风险的方法包括：对于系统关键技术的实现，请老师或同学辅导，对不熟悉的关键技术进行系统的学习。

2、解决方案的前景

2.1 前景声明

目标客户：京东商城、笔记本电脑商家、需要购买笔记本电脑的用户。

需求或机会的声明：京东商城需要对其进货库存进行决策并了解笔记本电脑在网站上的销售情况；商家需要了解目前笔记本市场用户最感兴趣的笔记本属性和功能；用户需要针对自己关心的笔记本电脑属性进行笔记本电脑选购和参考目前购买笔记本电脑的潮流趋势

产品名称：京东商城笔记本电脑商品深度 web 数据挖掘展示系统

产品类别：web 应用。

主要竞争产品、当前系统或当前业务过程：数据获取、分析。

新产品的主要竞争优势：可以针对京东商城笔记本电脑产品来帮助京东商城、商家、用户节约成本、提升市场把控和目标性以及较好的购买用户体验。

2.2 主要特征

- 通过爬虫程序与代理服务器技术获取数据；
- 通过 JSOUP、AJAX 页面解析技术、网页去重技术、中文分词技术、正则表达式对获取数据进行解析和处理；
- 将处理后的数据存入数据库；
- 利用挖掘技术对数据库数据进行分析；
- 利用 web 前端技术对分析结果进行展示

2.3 假设与依赖

本系统不涉及本部分内容

3、范围与限制

3.1 第一个版本的范围

- 利用爬虫程序与代理服务器技术获取京东商城某一类笔记本电脑的基本信息，包括：价格、品牌、尺寸（屏幕尺寸和显示比例（宽屏 or 普屏））、售后、便携性（厚度、重量）、外观、其他（蓝牙、无线、3G、麦克风、摄像头、指纹识别、触控 or 非触控）、配置（操作系统、CPU、内存大小、硬盘容量、显存容量、光驱类型）
- 利用 JSOUP、AJAX 页面解析技术、网页去重技术、中文分词技术、正则表达式对获取的网页数据进行处理，得到关键词
- 将关键词数据存放在数据库中
- 构建 SSH 框架
- 对数据库中的可用数据进行挖掘分析，并结合页面展示分析结果
- 根据分析结果对用户与商家给出基本推荐政策

3.2 各后续版本的范围

从数据获取角度：

- 增多商品种类,从单一笔记本商品研究转到笔记本市场多类型研究如(超极本、上网本、

游戏本、平板电脑、便携本等等）

- 增多数据来源网站，从单一获取京东商城笔记本销售数据发展到获取多个典型性电商网站笔记本销售数据
- 与某个电商网站深度合作，获取其部分用户消费信息

从数据分析与展示角度：

- 利用成熟的推荐算法，进行推荐结果展示

3.3 限制与排除

系统由于权限限制无法挖掘用户购买信息和关系信息，无法提供复杂的推荐功能；系统只针对笔记本电脑商家和需要购买笔记本的用户展示分析和推荐结果，对于其他商家和用户不具意义。

4、业务背景

4.1 涉众简介

系统主要面对京东商城、笔记本电脑商家、需要购买笔记本电脑的用户

对于京东商城：

- 可以根据一段时期的不同品牌的销售量分析来决定下一时期应该进多少货存储在仓库中，这是因为仓储管理也需要大量成本，而根据上一时期的销售数据分析可为其下一时期进货量，例如对于销售好的品牌多进货，防止货物短缺，这样做可以为其提前决策，节约成本
- 可以调研出用户目前最关注的笔记本电脑属性

对于笔记本电脑商家：

- 了解目前用户感兴趣的笔记本性能和特点，针对某一消费群体来影响品牌笔记本的设计与研究；

对于需要购买笔记本电脑的用户：

- 可以根据系统对整个京东商城的笔记本市场有一个大体了解，例如了解在一段时间内最受欢迎的电脑品牌或某一款电脑

4.2 项目优先级

- 所有项目必须在能获取到较为全面的数据的前提下执行；
- 面向用户的实际需求，数据分析可靠正确且具有一定意义；

4.3 操作环境

由于目前系统倾向于模型演示系统，因此暂不考虑异地访问
对服务中断的容忍度低，因为要将分析数据转化为图例展示，过慢的展示效果会影响用户体验。