文章编号：2095-2163(2019)03-0276-04 中图分类号：TP311 文献标志码：A

基于房地产大数据的自动估价系统研究

董睿琳[1]，董 楠[2]
（1 哈尔滨第六中学，哈尔滨 150001；2 哈尔滨工业大学 软件工程股份有限公司，哈尔滨 150001）

摘 要：大数据颠覆了人们对吃、穿、行的思考方式与习惯。而在“住”的方面，房地产一直以来都和金融业有着千丝万缕的联系，房地产大数据对于金融业来说有着至关重要的意义。依托于房地产大数据的自动股价平台可以为银行等金融机构带来决策性的意义，降低自身持有抵押品的风险。本项目是在物联网、大数据、下一代互联网的背景下提出的房地产评估系统。大规模发展 IPv6 下一代互联网，将会给互联网核心技术及大数据带来历史性发展机遇。当前房地产行业面临转型，要通过科技智慧化手段实现管理增效、技术增收，而物联网能够给地产行业转型升级提供有力支撑。

关键词：房地产大数据；金融业；网络爬虫；自动估价

Research on automatic evaluation system based on real estate big data

DONG Ruilin[1], DONG Nan[2]
(1 Harbin No.6 High School, Harbin 150001, China; 2 Software Engineering Co. Ltd., Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Big data has subverted the way of thinking and habits about eating, wearing and walking. Real estate has always been inextricably linked with the financial industry, real estate big data for the financial industry has a vital significance. Automated stock price platform based on real estate big data can bring decision-making significance to banks and other financial institutions and reduce the risk of holding collateral. This project is a real estate evaluation system under the background of Internet of Things, big data and next generation Internet. Large-scale development of IPv6 next generation Internet will bring historic development opportunities for core technology of Internet and big data. At present, the real estate industry is facing a transformation. It is necessary to achieve management efficiency and technology income by means of scientific and technological wisdom. The Internet of Things can provide strong support for the transformation and upgrading of the real estate industry.

【Key words】 real estate big data; finance; Internet worm; automatic valuation

0 引 言

随着网络信息技术的不断进步，大数据时代已悄然来临，大数据也在各行各业中陆续得到广泛的应用，而且正在逐渐改变着人们的社会生活[1]。

IPv6 下一代互联网的大规模发展，将会给互联网核心技术及大数据带来历史性发展机遇。物联网、大数据与房地产密不可分，未来数据资产在房地产中的价值体现也越来越重要。当前房地产行业面临转型，要通过科技智慧化手段实现管理增效、技术增收，而物联网能够给地产行业转型升级提供有力支撑。

房地产业因其运转周期长、融资量大离不开金融业的支持，而金融业则将房地产业视为一种安全性和收益性都很高的优良资产和黄金业务。房贷业务几乎成为大部分银行信贷板块中的主推项目。由于国内社会信用制度尚不规范，整个社会的商业信用体系也有待完善，导致银行在很大程度上将可能面对一定的商业风险。因此银行需要对押品进行估价，实时掌握押品的价值。数据是前瞻性的，收集历史数据，目的是为了预知未来[2]，为可能到来的金融风险做准备。

1 房地产大数据现状及存在问题

1.1 国内现状

作为国内知名的房产经纪公司，链家很早之前就已开始了大数据探索尝试，在大数据的构建、应用上已取得了初步成就。

禧泰房地产数据有限公司是国内最早设立的专业房地产大数据公司，早在 2005 就开始从事房地产数据的收集、整理和研究应用。该公司于 2017 年度提供房产自动估价服务 8 000 万笔、服务房产交易用户超过 1 亿人次（以上数字来源自禧泰官网）。自动估价系统已经逐渐替代传统的房地产评估公司的人工估价业务。

基金项目：赛尔网络下一代互联网技术创新项目（NGII20160901）。

作者简介：董睿琳（2001-），女，高中生，主要研究方向：数据挖掘、数据分析；董 楠（1995-），女，学士，助会，主要研究方向：财务软件分析。

收稿日期：2019-01-18

1.2　国外现状

CoreLogic 公司是全世界最大的房地产数据分析服务商。该公司将政府公开信息、客户特供和第三方数据构建成复杂而又庞大的大数据库，就美国而言，覆盖了 99.8%以上人口，超过 1.47 亿人的财产记录，搜罗了超过 930 万人的按揭贷款申请，超过美国 99%县、市及特殊税收管辖权的纳税记录，超过 7.95 亿次房地产交易历史数据，占据租赁市场约 70%的 23 万活跃的租户/业主记录，每年可提供超过 2 500 万的信用报告，甚至包括空间地理与国家防汛数据(以上数据源自链家研究院)。

1.3　中国房地产大数据存在的问题

目前，中国对房地产大数据的管理是匮乏的，没有专门的机构来整理记录房地产大数据，房地产大数据仍处于杂乱无章的状态，具有真实性低、规范性差等特点，这给房地产大数据的应用造成了巨大的困难。网络中的挂牌和出售数据需经过去重、清洗后才能在日常实际生活中投付使用。本项目在清洗与去重中采用了时间与空间相结合的方法，并在数据中根据不同属性采用了取极值的操作，去重率在 80%以上。

2　房地产大数据将改变传统房地产业

随着下一代互联网和物联网的发展，以及房地产大数据的日趋完备，人们居住的房子将会被赋予更多的网络属性，智慧生活，智能家居就目前而言已是触手可及。可以这样说，未来优秀的企业都是大数据公司，都将基于大数据生态环境让自己更高效、更智慧地参与现代市场竞争。所以在现如今的金融业、房地产行业中，必须拥有卓越的大数据体系和平台整合能力，而不是如同当下仅仅关注的只是销售排名的数字。

如果说传统工业代表着过往，互联网科技代表着现在，那么以大数据为代表的智慧科技则代表着未来。金融业与大数据的深度融合是大势所趋。

目前，传统房地产估价行业中大多数公司依然采用传统的人工方式进行评估，这种方式不仅费时、费力，而且在操作上也不具备公开透明性。通过基于房地产大数据的自动评估系统能够实时批量地对房产进行评估，能够给银行减少时间成本，同时还可降低金钱成本。

3　通过网络爬虫抓取地产大数据

本次项目研究中，获取数据的方法主要为网络爬取。网络爬虫，又被称为网页蜘蛛、网络机器人，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。另外，一些不常使用的名字还有蚂蚁、自动索引、模拟程序或者蠕虫[3]。通用爬虫的设计架构如图 1 所示。

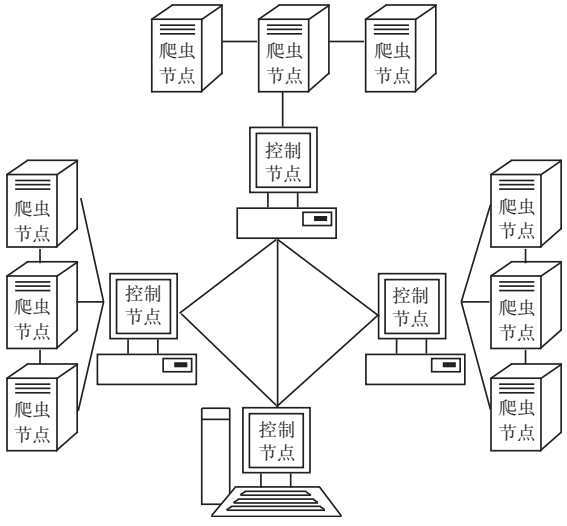

图 1　通用爬虫架构图

Fig. 1　General reptilian architecture diagram

在各种以数据作为设计运行基础的实验中，数据重要性是不言而喻的。不仅需要数据作为各种模型的基本计算和训练依据，产生更多、更准确的特征来构建和模拟构建效用相当的仿真模型，通过这些模型对新产生的数据进行预估和处理，从而提高模型的利用效果。随着网络的迅速发展，互联网成为大量信息的载体，如何有效地找到自己需要的信息，并加以提取和利用即成为一个巨大的挑战。

定向抓取相关网页资源的聚焦爬虫可以帮助研究者解决这一问题。聚焦爬虫是一个自动下载网页的程序，可根据既定的抓取目标，有选择地访问互联网上的网页与相关的链接，获取所需要的信息。与通用爬虫(general purpose web crawler)不同，聚焦爬虫并不追求大范围的覆盖，而将目标定为抓取与某一特定主题内容相关的网页，为面向主题的用户查询准备数据资源。

4　房地产大数据的清洗加工

分布式的数据抓取系统，散布在不同位置的数据中心，若干台抓取服务器，若干套爬虫程序，构成了一个分布式的抓取系统，用于存储各个阶段的历史数据。借助于成熟的分布式系统基础架构 Hadoop 开发分布式程序，充分利用集群的威力进行高速运算和存储。基于 IPv6 的部署实现，进一步推动 IPv6 在研发

实践中的普及应用。最终结合本次研发需求制定一套适合该项目的自动估价模型,相对传统估价模式对高成本的估价方式,自动估价技术的运营成本明显降低。Hadoop 架构图如图 2 所示。

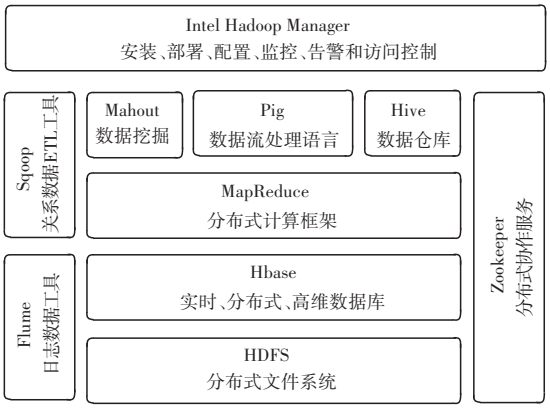

图 2 Hadoop 架构图

Fig. 2 Architecture diagram of Hadoop

研究中,建立了房地产数据仓库,将海量的原始数据存放于数据仓库中,通过自动化的脚本流程自动整理与清洗数据。并依托于清洗后的结果进行统计分析,将分析后的结果以准实时的方式存放于应用数据库中。

数据仓库可以存储各个阶段的历史数据,为房地产价格的分析起到事半功倍的作用。

爬虫抓取到的数据经过格式化处理后送至数据仓库的增量层,然后经过清洗去重处理后送入到全量层。在全量层对其进行统计,再将统计后的结果传送到应用服务层。整个过程的执行周期为一天。做到数据的准实时。整个处理流程如图 3 所示。

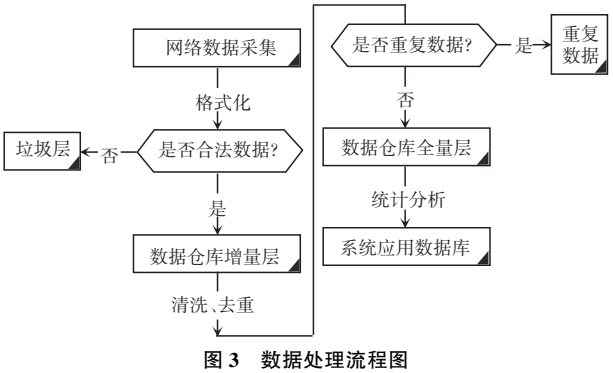

图 3 数据处理流程图

Fig. 3 Data processing flow chart

5 自动估价系统的实现

在数据日渐开放、并已全面进入大数据时代的背景下,审时度势的实践者可以利用自动估价技术对中国的房地产估价和经纪行业带来变革。事实证明,AVM[4]并没有使估价机构丢失了原有的业务而受到威胁,反而使其可通过利用 AVM 技术为客户提供更加丰富全面的估价服务,同时也保证了自己的收益,创造了良好的客户关系。

大数据是房地产估价方法的基础。房地产估价方法包括比较法、收益法、成本法、假设开发法。例如,比较法中交易实例的搜集、房地产状况调整,收益法中的资本化率的确定,全部需要大数据[5]。本项目实质为一个垂直搜索模型,通过输入房屋的具体地址信息,评估房子的价格,展示房子的属性。基于此,这里给出了研发系统的首页设计效果见图 4。继而,关于房地产小区详情页和小区其它信息页的界面效果则分别如图 5 和图 6 所示。

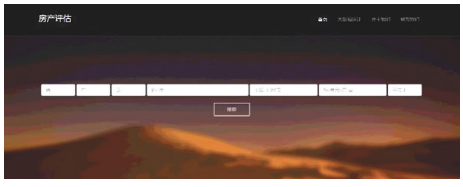

图 4 系统主页

Fig. 4 System home page

图 5 小区详情页

Fig. 5 Community details page

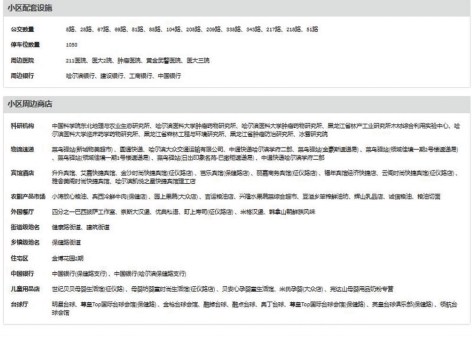

图 6 小区其它信息页

Fig. 6 Other information pages of community

基于房地产大数据的自动估价可以为房地产实现更为精确的自动估价,可以解决各级信贷审批人员缺乏便捷全面的房地产综合全景信息工具的问题。自动估价在提供房地产自动估价、人工估价和价格走势等多维度分析的同时,还可呈现相关的楼盘综合信息、市场动态和周边设施配置,可以有效地提高信贷审批业务的工作效率。 (下转第 281 页)