文章编号：2096-1472(2018)-09-41-03
DOI:10.19644/j.cnki.issn2096-1472.2018.09.012

科研人员画像系统设计与实现

金冈增，李 娜，郑建兵，高 明
(华东师范大学数据科学与工程学院，上海 200062)

摘 要：用户画像主要用于精准营销、用户征信和个性化推荐等，该技术已广泛应用于电信、电子商务、社交网络等领域。在很多领域，由于数据孤岛的存在，还没有真正实现数据赋能，如在研究生教育评价领域，一方面教育主管部门大多采用抽样、问卷调查等形式对研究生教育进行评价；另一方面，海量的科研数据散落在互联网上，如智立方、谷歌学术、百度学术、DBLP和Web of Science等网站收录了大量科研人员论文信息，学习经历和工作经历散落在各类招聘网站上。为了实现精准的研究生教育质量评价，需要整合各类科研数据，打破数据孤岛。因此，本文以互联网数据为基础，在数据融合的基础上设计并开发了一款科研人员画像系统，辅助科研人员、科研机构、教育主管部门等开展智能决策。

关键词：科研人员画像；数据孤岛；数据清洗；数据融合
中图分类号：TP391 **文献标识码**：A

Design and Implementation of the Scientific Researchers Profiling System

JIN Gangzeng,LI Na,ZHENG Jianbing,GAO Ming
(*School of Data Science and Engineering*,*East China Normal University*,*Shanghai* 200062,*China*)

Abstract: User profiling is mainly used for precise marketing,credit investigation and personalized recommendation.The technology has been widely used in telecommunications,e-commerce,social networking and other fields.In many fields,due to the existence of data islands,data empowerment has not yet been realized.For example,in the field of postgraduate education evaluation,on the one hand,most education authorities use sampling and questionnaires to evaluate graduate education.To achieve accurate graduate education quality assessment,it is necessary to integrate various types of scientific research data to break the island of data.For example,there are a large number of research papers which are distributed in heterogeneous web platforms,such as Zhicub,Google Scholar,Baidu Academic,DBLP,and Web of Science,etc.Their education background and work experience are scattered on various recruitment websites.Therefore,based on the Internet data,this paper designs and implements a researcher profiling system based on data fusion to assist scientific researchers,research institutes,and educational authorities in making intelligent decisions.

Keywords:researcher profiling;data island;data cleaning;data fusion

1 引言(Introduction)

二十多年来，随着信息技术的飞速发展，在社交网络、医疗、教育和社会治理[1]等领域数据量都呈现出爆炸式增长。在当前的“大数据时代”，有效处理海量数据的能力已成为各行各业数据赋能的关键[2]。由于数据孤岛问题的存在，致使难以获取完整而准确的用户信息，数据的价值难以发掘，因此实体匹配已经成为数据集成的关键任务之一[3]。

目前，研究生教育质量评价除了政府主导推动的学位点和学科评价之外，近年来第三方排名评价，以及由研究型大学自主开展的学科国际评价不断涌现[4]。虽然目前评价趋于多元化，但是学位中心作为教育主管部门大多采用抽样、问卷调查的形式对研究生教育进行评价。随着我国研究生教育大众化的推进，这种粗犷式的决策方式越来越无法准确地评估研究生教育质量。另一方面海量的科研数据散落在互联网上，如智立方、谷歌学术、百度学术、DBLP和Web of Science等网站收录了大量科研人员论文和项目资助信息。为了实现精准的研究生教育质量评价，需要整合各类科研数据，打破数据孤岛。

伴随着数据分析和挖掘技术不断发展，“用户画像”的概念逐渐成型[5]。因此，本文以学位中心数据为基础，互联网数据为补充，在数据融合的基础上设计并开发了一款科研人员画像系统，辅助科研人员、科研机构、教育主管部门等开展智能决策。该系统能够对科研人员从基本信息、项目资助情况、论文情况，以及合作者情况进行画像。在此基础上，该系统还能对这些信息进行有效聚合，从而实现面向学校、行政区域和不同学校类别的多样化群体画像。这些画像可以有效地对科研人员、科研机构的研究生教育进行评价，从而可以帮助研究生根据自身需求选择合适的导师，帮助科研机

基金项目：国家重点研发计划(2016YFB1000905)；国家自然科学基金(61402177，61502236，61472321)；上海市农业推广项目(T20170303)．

构进行自查自省，帮助教育主管部门监控研究生教育质量，提升科学决策的能力。

2 系统目标与架构(System goal and architecture)

2.1 系统目标

该系统不仅需要准确地进行多层次画像，而且能够有效地辅助科研人员、科研机构、教育主管部门等开展智能决策。因此该系统必须具备以下几个特征。

(1)数据获取的全面性：为了全面衡量科研人员的科研表现，辅助研究生教育质量评估，该系统需要实时爬取各类科研信息，其中包括发表论文、项目资助、教育经历和工作经历等信息。

(2)数据融合的准确性：来自不同信息孤岛的碎片化数据可能存在数据不一致、数据缺失等数据质量问题，系统需要匹配来自多个数据源的科研人员信息，并运用数据融合技术提升数据质量。

(3)评价体系的合理性：科研人员画像是碎片化科研数据向高精度标签的转化过程，因此该系统需要构建一套合理的、多层次的和全面的标签体系，以便于合理地评价科研人员，以及科研机构的科研表现。

2.2 系统架构

如图1所示，科研人员画像系统由数据获取层、数据匹配与融合、数据分析与挖掘、画像子系统和辅助决策层构成。其中数据获取层进行数据采集、清洗和数据采样；数据匹配与融合层进行姓名消歧从而实现数据融合；数据分析层主要对数据进行深度分析挖掘得出有意义的结论；画像子系统则是对数据进行多层次画像，从而实现辅助决策的作用。

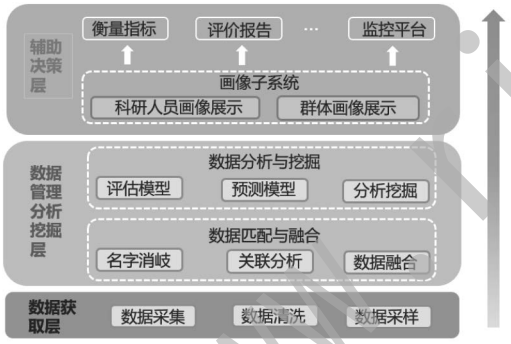

图1 系统逻辑架构设计图

Fig.1 System logic architecture design

2.2.1 数据获取层

数据获取层是整个系统的基础，它主要由数据采集、数据清洗和数据采样三大子模块组成。数据采集模块主要通过爬取智立方、百度百科，以及论文数据库中获取完整的计算机学科数据。为了实现多源数据的实时获取，采用了Python的Scrapy框架实现了分布式数据爬取。其采集的数据主要包括以下三大类：

基本信息：包括科研人员的基本信息、供职单位基本信息，以及其相关科研人员的教育经历等信息。

论文信息：包括其每篇论文的发表年份、发表刊物、主题、引用量，以及合作作者等信息。

项目信息：包括科研人员主持或者参与的各种不同等级的项目，以及获得的资助情况。

初步采集的数据质量差，存在数据缺失、数据与常识不符，以及信息冗余等问题。因此需要进行数据清洗从而保证数据的完整性和真实性等。

2.2.2 数据匹配与融合

随着数据规模迅速扩大，数据匹配面临的挑战越来越多，主要包括数据的海量性、异构性、隐私性、相依性和低质性等挑战。由于数据是多源获取的，数据存在冲突、不一致和歧义等问题。通过用户实体匹配，可以将科研人员在多个数据源上的信息进行融合，以便更全面地了解其整体情况，对用户精准画像、多层次画像提供依据。

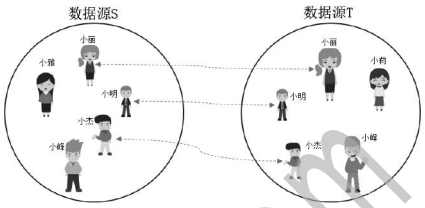

图2 多数据源实体匹配

Fig.2 Entity matching in multiple data sources

多数据源实体匹配主要是关联不同数据源之间的实体，为更全面地构建用户画像打下基础。本系统基于无监督的匹配方法将来自不同数据源的实体进行匹配[6]，并利用数据融合方法消除数据冲突、不一致和歧义等问题，提升数据质量[7,8]。

2.2.3 数据分析与挖掘

依据科研人员画像系统的建设目标，数据分析与挖掘旨在构建科研人员科研评估模型，通过对科研数据的分析、预测和挖掘，实现科研数据向知识的转化。该模块主要挖掘科研人员的研究兴趣，分析研究兴趣的前沿性，以及构建科研人员的影响力模型(如合作网络中的PageRank值[9]、度中心值、亲密度值等关键指标)。

2.2.4 画像子系统

依据定义的一整套标签体系，画像子系统在对数据分析与挖掘的基础上，实现科研人员数据知识化和标签化。特别针对群体画像而言，基于科研人员画像，利用聚类和聚合分析等方法实现群体画像。

其中画像标签体系一级类别共有两类分别为科研人员画像和群体画像，在这两个一级画像类别下又分别包含了基本信息、科研产出、项目资助、学术影响力这四个二级标签。在每类标签体系里详细定义标签信息。

2.2.5 辅助决策层

在画像子系统的基础上，辅助决策层可以直观地利用各项衡量指标对不同科研人员、科研机构进行横向和纵向的分析和比较；辅助研究生教育质量评估；监控异常以辅助管理部门决策。

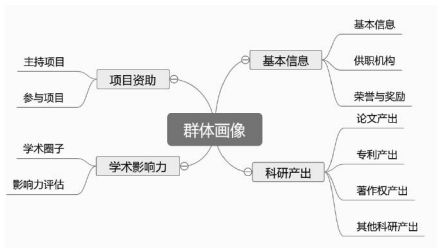

图3 群体画像标签体系

Fig.3 Group profiling label system

3 系统功能(System functions)

3.1 全国画像

全国画像主要描绘了全国具有一定影响力的科研人员的

分布概况，以及各地区一流高校分布情况，如图6所示。图中第二栏展示了全国大学数量、985高校和211高校数量、双一流高校数量和教师总数。正中的热力图展示了全国具有一定影响力的科研人员的分布，由该图可以充分反映出华东，以及华北中的北京、天津等地区是科研密集区，而西北地区则分布十分稀疏。此热力图所反映的信息与其左边的各区域高校分布图互相印证。图中左下角的区域分布图则反映三种类别一流高校在各大区域的数量的区域分布图则反映三种类别一流高校在各大区域的数量，可以看出华东地区在各方面指标上都是领先的。

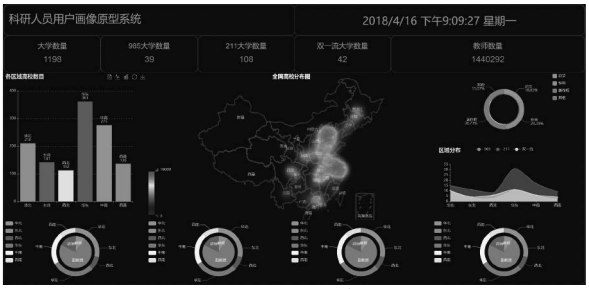

图4 全国画像

Fig.4 National profiling

3.2 区域画像

在全国画像的热力图中点中某一区域则会进入相应区域的画像展示。以北京为例，从第二行来看，北京大学数量、高水平大学数量，以及高层次人次入选数量都遥遥领先其他省市。结合地图，以及各区域高校数目图，可以发现海淀区的大学数量远高于其他地区。将鼠标放于右上角的学校影响力排名图中，则会出现某一学校各指标的数值喝排名情况，这些指标综

图5 区域画像

Fig.5 Regional profiling

合起来可以反映某一高校的科研影响力。

3.3 学校画像

点击区域画像上的坐标点，即可进入学校画像界面。以清华大学为例，第二行显示该校一级学科博士和硕士点个数、长江学者人数、国家千人计划人数和国家杰青人数等。通过展示科研人员影响力反映清华大学某学科在全国的科研实力。然后采用图表分别从清华大学论文发表情况、学术成果产出情况和主持或参与国家自然科学基金情况等。高校论文情况则选取类似的一些高校分从论文作品数、引用量，以及H指数这三个方面进行横向比较。近五年人均论文情况图则从时间维度上观察人均论文发表情况的变化趋势。学术成果产出统计图则是通过饼图来展示人均发表期刊文章数、人均发表会议论文数、人均申请专利数等占比情况，从图中分析可知该校发表期刊文章数占一半以上。近八年主持或者参与国家基金委项目情况图可以在柱形图和折线图之间来回切换，该图也是从时间维度上观察主持或者参与项目的变化情况。最后在整体页面的右上角有一个下拉框，点击下拉框则可以显示出该校所有科研人员的姓名，点击其姓名即可进入具体科研人员画像页面。

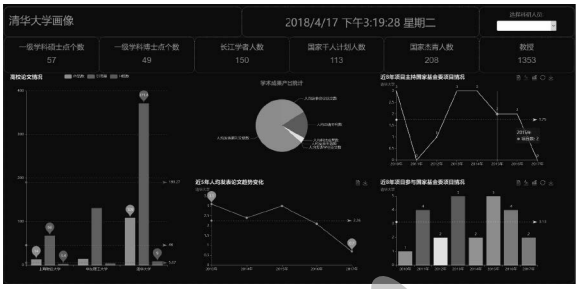

图6 学校画像

Fig.6 School profiling

3.4 科研人员画像

科研人员画像首先展示的是其姓名、供职单位、职称、所获荣誉和奖励的基本情况，从这些信息可以对该科研人员有直观地认识。图中六张图表分别从论文发表情况、项目情况、科研合作情况和学术影响力等多个维度衡量其科研实力。论文情况箱型图则是通过箱型图反映作品数、引用数和H指数这三个指标在整体中所处的位置，从图中所知该教师这些指标处于中位数之上。近五年发表论文情况则反映该科研人员科研产量的变化趋势。获资助论文数箱型图展示了该科研人员获得国家级、省级、市级，以及校级等资助情况所处位置。项目数箱型图则是从参与，以及主持国家级和国际级项目的角度进行分析。中心性排名箱型图则先分别计算了该科研人员相关人数、PageRank值、度中心性、紧密中心性，以及中介中心性等衡量影响力的指标。

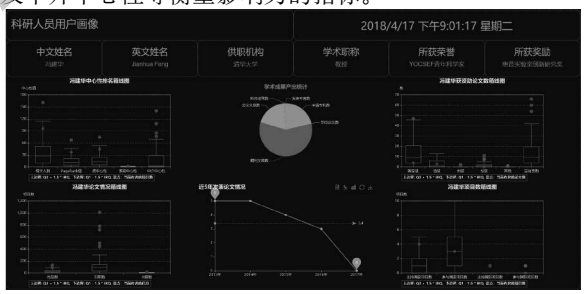

图7 科研人员画像

Fig.7 Researcher profiling

4 结论(Conclusion)

本文首先论述了科研人员画像的背景、意义，以及总结目前国内外研究现状。然后介绍了该系统的系统目标，以及系统逻辑架构。主要以数据获取的全面性、数据融合的准确性，以及评价体系的合理性为目标。系统逻辑架构则是从数据获取层、数据匹配与融合、数据分析与挖掘、画像子系统，以及辅助决策层自下而上构建。最后从全国画像、区域画像、导师画像，以及科研人员画像这四个方面对该系统功能进行了详细的介绍。

参考文献(References)

[1] Wu X,Zhu X,Wu G Q,et al.Data mining with big data[J].IEEE transactions on knowledge and data engineering,2014,26(1):97-107.

[2] Sowmya R,Suneetha K R.Data Mining with Big Data[C].

(下转第28页)