基于卷积神经网络的近代报纸广告图片聚类方法*

王海洋 / 清华大学统计学研究中心和工业工程系
邓　柯 / 清华大学统计学研究中心和工业工程系
陈　静 / 南京大学艺术学院

摘　要：近代报纸广告作为重要的历史研究文献资料，日益受到人文学者的重视。用图像识别技术对海量广告图像进行内容解析是图像研究和广告研究的新方向。常规路线通过图像识别技术对单幅广告图片进行内容解析，错误率较高，并且需要大量人力进行校对，成本很高。此研究跳出对单幅广告图像内容解析的微观视角，从整体考虑海量图像的内容解析问题。利用深度学习将广告图像转化为特征向量，计算图像的相似度，此研究建立了从广告图像中识别相似模式的技术工具。应用相关方法对《中国商业广告档案库》中收录的广告图像进行处理，将图像解析工作的效率提高了近六倍。

关键词：近代报纸广告　内容分析　图像解析　卷积神经网络　图片聚类

引　言

近代报纸广告作为重要的历史研究文献资料，日益受到来自历史、艺术、经济史、文学、广告学、传播学等领域的研究者们重视，对当代人文研究具有重要意义。广告刊登的版面、频次、尺幅大小等信息能够为研究相应产品、企业和行业的商业历史提供大量信息；对广告图像内容，如广告主题、图像特征、叙事方

*本文为国家社会科学基金重大项目“基于大数据技术的古典文学经典文本分析与研究”（18ZDA238），国家自然科学基金（11771242），中央高校基本科研业务费专项资金（项目编号011514370105）和北京智源人工智能研究院研究员专项基金（BAAI2019ZD0103）资助阶段性成果。

式、艺术风格等，进行深入分析，能够反映近代社会经济、文化、社会生活的诸多维度，更是人文学者们十分关心的重要问题。近年来，有关近代图像，尤其是广告图像的研究越来越受到学界重视。比如，陈平原教授以晚清画报为例对传统中国的“左图右史”与西学东渐之“图像叙事”结盟，进而汇入到以“启蒙”为标识的现代化进程的研究；[①]梅嘉乐（Barbara Mittler）教授从妇女杂志、广告切入对以上海为代表的近现代中国城市视觉的研究；[②]安克强（Christian Henriot）教授对摄影技术和照片塑造近代中国形象的研究；[③]葛凯（Karl Gerth）教授对近现代中国的消费文化和民族国家的研究；黄克武教授对于近代医药与身体观与广告图像的关系研究；[④]白露（Tani Barlow）教授所开展的关于近代广告中的“摩登女性”形象的研究，将广告、中国现代商品社会的出现和本土社会学（借助中国启蒙社会思想来解释日常生活、阐释精英化社会关系与社会实践等的思想或著作）放置于特定的历史语境中进行考察，以包括Cutex（蔻丹）、BAT（英美烟草公司）、双美人牌化妆品、安迪生电灯泡、卜内门化肥等商品的一系列广告为例，说明商品在现代中国社会是如何与社会进化思想联系起来的；本土社会学是如何与性感的摩登女性的图像建立关系的；广告又是如何将这种对现代性的另类的幻化图像自然化的。[⑤]这些学者的研究都为近代广告研究提供了具有开拓意义的视角和途径。

内容分析是近代报纸广告研究的核心问题之一。现代报纸广告研究中，图像是非常重要的内容要素。一方面是因为对于广告史，从早期的“告白”逐渐细化出商品广告、商业服务广告、娱乐广告、文教类广告和社会广告等，涵盖日常生活各个方面。形式也从早期的文字为主发展为更富有叙事性的插图为主，图像成为广告宣传意图和营销策略的重要体现。另一方面则是因为以广告为代表的现代商业艺术形式发展出了特殊的艺术风格：卡通、手绘、线条为主的绘图形式体现了现代性视觉文化的重要转变，以视觉化的方式展示了现代公民所享受的新商品世界，一个理想化的世界：在这里现代商品、药品、社会活动（开车、跳舞、清洁、缝制甚至施肥）都是标准化的、“现代的”。

---

①陈平原：《左图右史与西学东渐：晚清画报研究》，北京：生活·读书·新知三联书店，2018年。

②B. Mittler, “Defy(N)ing Modernity: Women in Shanghai's Early News-Media (1872-1915),” *Jindai Zhongguo funü shi yanjiu* (*Research on the History of Women in Modern China*), November 2013, p.217.

③C. Henriot, “Visualising China, 1845-1965: Moving and Still Images in Historical Narratives,” *Chineses Studies*, Leiden: Brill, 2012.

④黄克武：《从申报医药广告看民初上海的医疗文化与社会生活，1912—1926》，《中研院近代史研究所集刊》1988年第17期。

⑤Tani E. Barlow, “Buying In: Advertising and the Sexy Modern Girl Icon in Shanghai in the 1920s and 1930s,” *The Modern Girl Around the World*, ed. The Modern Girl Around the World Group, 2008, pp. 288-316.

尽管广告图像的重要性日益凸显，但研究广告的方式却缺少行之有效的图像筛选方法。广告图像的特殊性在于：为了达到有效的说服营销，同一个版式的广告图像会反复出现。而在传统的广告研究中，学者主要靠人力去收集、浏览和筛选图像类型，对同一版式的出现频率进行统计，从中看出广告图像所体现的商家营销策略和受众接受度。由于目前报纸类档案库和数据库主要依据报纸期刊类的连续出版物的出版序号存档，因此广告是作为版面内容的一部分出现的，并不是独立的对象。尽管某些档案/数据库提供了相关的搜索功能，但搜索结果也多基于广告条目的文献信息，如日期、期号、主题，偶见广告商品名或广告语。就图像内容而言，从大规模处理海量数据的角度来对图像进行计算分析的思路，尚不多见。目前图像识别技术的发展，使其在多领域、多场景中的施用日益成熟。运用前沿图像识别技术对大量广告图像的内容进行解析，从中自动提取出人文学者感兴趣的各种要素，具有技术可行性，是非常有吸引力的解决方案。据我们所知，到目前为止还没有看到这方面的研究工作公开发表。

本文力图探索运用基于卷积神经网络的深度学习技术对海量近代广告图片进行自动化内容分析的技术路径。我们首先尝试了实施图像处理和分析的常规技术路径：运用图像识别工具对大量广告图像逐一进行内容解析，从图像中提取出文本、图形、构图、风格等关键特征后，再结合人文研究的不同视角进行定量分析和研究。但由于近代报纸广告图像的特殊属性，运用上述基于单幅图像识别的技术路线进行内容解析时错误率较高，仍需要投入大量人力对各个图片的解析结果逐一进行校对，周期长，成本高。这些困难成为制约近代报纸广告大规模、系统性研究的主要瓶颈。对此，我们认为可以跳出对单幅广告图像进行内容解析的微观视角，从整体上考虑对一大批广告图片同时进行内容解析。该路线主要包含四个步骤：首先，对广告图像进行预处理，统一其格式；其次，运用深度学习技术将广告图像转化为“特征向量”；第三，利用得到的特征向量对不同广告图像之间的相似度进行度量，通过相似度生成大量广告图片之间的“关联图谱”；最后，依据得到的关联图谱将多幅内容高度相似的广告图像聚合为一个“图片类”，对归属于同一个图片类的多幅广告图片实施同步内容解析，并通过“整合分析”将来自各单幅图像解析的结果相互印证后输出最终的图像解析结果。

与对单幅图像逐一进行识别的常规技术路线相比，这个新的技术路线具有明显的技术优势：首先，能够很好利用不同广告图片之间可能具有的相似特征来实现信号增强，从而大幅提高图像解析算法的精确度和稳健性；其次，能够通过整合分析自动识别单幅图像解析中的潜在错误，为下一阶段的人工校对提供指引；

第三，这种技术框架和数据组织形式的建立可以为其他更具针对性的人工智能技术开发和应用创造有利的条件，推动相关研究的智能化水平快速提高。这些技术优势极大地提高了近代广告图片内容分析的智能化程度，大幅减轻对人工校对的依赖。

将这一技术路线应用于海量近代报纸广告的实际数据分析，我们发现大量的广告图像在内容和形式上具有高度的一致性，本质上对应于一则商业广告，仅以略有差异的形式在不同日期的一份甚至多份报纸上重复出现。运用上述基于相似模式识别和匹配的技术路线，我们对海量广告图像实施了有效的聚类分析，将版式相同或仅有细微差别、大量重复出现的“广告图像”压缩成为一个“唯一图像”，以不同的“唯一图像”而不是“广告图像”作为基本的研究对象展开后续研究。从人文研究的角度来看，这种新技术路线的应用具有如下几方面的积极意义：首先，将“广告图像”和“唯一图像”的概念区别开来，使得人文研究的基本思路和逻辑更为清晰；其次，将大量高度类似的“广告图像”聚合成一则“广告”，大幅降低了数据管理、处理和分析的难度，使得人文学者能够更容易摆脱各种数据噪声的干扰，聚焦于问题的本质；第三，类似的研究思路不仅仅局限于近代报纸广告的研究，对其他具有类似特征的研究对象，如手绘风格的漫画、线描为主的绘画作品等，亦具有借鉴意义。

上述技术路线的一个核心前提要素在于开发有效的广告图片聚类工具。本文聚焦于对这个核心技术要素的构建和阐释，并通过模拟实验和实证研究印证其可行性和有效性。受限于篇幅，不再对技术路线的另一核心要素“整合分析”做详细介绍。我们将在后续的研究论文中对相关内容做更为深入的介绍。本文的结构如下：在第一节，我们将简要介绍本研究所涉及的近代报纸广告图像数据的基本情况；第二节详细说明我们设计开发的基于深度学习的广告图像聚类方法；第三节通过实验测评该方法的性能，并和其他方法进行比较；第四节展示运用新方法实际处理海量近代报纸广告图像数据的结果；最后，对本文进行总结和讨论。

一、近代报纸广告图像数据简介

本研究所涉及近代报纸广告图像数据主要来自于《中国商业广告档案库》[①]收录的广告图像。该档案库是一个学术驱动型的在线档案馆，收录了19世纪末至20世纪中叶在中国主要报纸上刊登的国际品牌广告的高质量数字图片，从五个

①The Ephemera Project, “Chinese Commercial Advertisement Archive (CCAA),” Rice University, Luce Foundation and Nanjing University, https://ccaa.nju.edu.cn/html/index.html.

最主要的因条约而被迫开放口岸的港口城市中，选取了最重要的几种商业报纸，包括上海的《申报》、天津的《大公报》、沈阳的《盛京时报》、汉口的《汉口中西报》和广州的《越华报》。档案库中的每张广告图片都附有可用于研究的基于都柏林数据规范的元数据，其中除文献信息如日期、期号、版面等外，还包括描述性数据，如动物、植物、人物性别、人物年龄、商品品牌和商品类别等。其内容涵盖该报纸上刊登的大多数外国公司所生产和销售的所有商品广告。表1总结了档案库中数据资料的基本信息。图1给出档案库中典型广告图片的示意图。从中可以看出，相关数据资料在时间、空间、行业、商品上具有广泛的分布，在广告样式上具有多样性。这确保了这些数据资料具有广泛的代表性。本研究旨在开发对不同广告图像的相似模式进行自动识别和匹配的技术工具，并对广告图片实施精准、高效的聚类分析，将对应于同一则广告的不同广告图片聚成一类。

表1 《中国商业广告档案库》收录近代报纸广告数据基本情况汇总

报纸	发行城市	时间跨度	广告图像数量
《申报》	上海	1873—1940	约 60,000
《大公报》	天津	1906—1934	67,544
《盛京时报》	沈阳	1907—1938	49,381
《汉口中西报》	汉口	1908—1937	15,222
《越华报》	广州	1930—1938	3,488

图1 CCAA收录的广告示意图

二、研究方法

为了将版式相同或仅有细微差别、大量重复出现的“广告图像”压缩成为一个“唯一图像”，我们将基于卷积神经网络的深度学习技术与迁移学习技术相结合，开发出被称为RNGI的图片聚类方法。图2展示了该方法的主要步骤和实施流程。算法以一组待聚类的图片为输入，通过“预处理”“特征提取”“构建关联图谱”和“后处理”四个步骤，最终得到输入图片的聚类结果作为输出。下面，我们将逐一介绍算法中各步骤的技术细节。

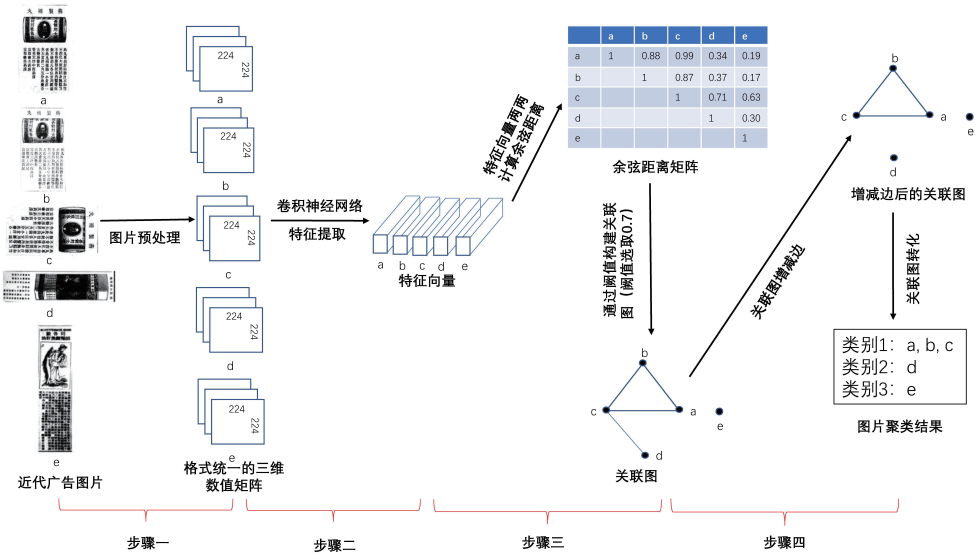

图2　基于深度学习的民国广告图片聚类算法流程示意图

首先，是预处理环节。这一环节的主要目的是将不同尺幅大小、存储格式、摆放角度的原始图片统一到一个规范的框架，以便于后续步骤的实施。具体包含如下几方面的操作。最开始是异常图片识别和删除。由于数据管理和数据存储方面的某些原因，在大型图片档案库中不时会包含一些异常图片，比如并无任何内容的“空图”、发生格式损坏无法读取的“坏图”等。将这些异常图片导入后续的数据分析既无意义，又可能会对算法产生干扰。因而，应在开始阶段识别并删除。其次，是图片大小及存储格式转化。为了卷积神经网络能够方便地对图片进行处理，我们将所有输入图片统一转化为解析度为224×224像素的RGB图片。存储格式为RGB的图片的数学本质是一个三维矩阵，其中前两个维度代表图片

像素点阵的行和列，第三个维度代表每个图片像素在红（R）、绿（G）、蓝（B）三个颜色通道的色彩值。在本文中，所有经过预处理的输入图片最终均转化为224×224×3的三维矩阵。转化方式为先将像素矩阵的行列像素分别平均划分为224份（不是224的倍数，则最后一份加上余数）；然后对224×224的像素小格，求每个像素的平均值，此时得到行列像素均为224的图片；再将图片通过颜色分解，转化为彩色图片，得到224×224×3的三维矩阵。最后，是图片角度旋转。此步骤主要是用于修正待聚类图片中没有摆正的图片。可旋转的度数包括90度、180度、270度，由计算机算法自动识别并完成旋转。

第二步是特征提取环节，主要通过基于卷积神经网络的深度学习和迁移学习技术将每个经过预处理的图片转化为一个固定维度的特征向量，实现图像特征提取。[①]此处，我们使用经由图片数据库ImageNet[②]预训练过的VGG16[③]卷积神经网络作为图片特征提取的主要工具。VGG16是为图像识别及分类任务而搭建的一个深度神经网络，是近年来非常流行的一个图像处理深度学习构架。VGG16共由16层网络构成，图3给出了其网络结构的基本构架。该网络以224×224像素的RGB图片为输入，通过由卷积层（convolution）和池化层（max pooling）构成的13层网络不断进行卷积操作，将维度为224×224×3的原始输入图片逐步转化为一个维度为7×7×512的特征张量，然后再通过3个全连接层进一步将图片的特征信息转化为一个维度为1000的特征向量，并最终通过一个softmax层实现图像分类功能，整个网络共包含约1.3亿个自由参数。当全部网络参数被确定以后，任何一个尺寸为224×224像素的RGB图片通过多层网络的层层迭代计算，将输出一个在图像类别空间上的概率分布，从而实现图像识别分类。

然而，要让这种规模庞大的深度神经网络真正发挥作用，需要使用海量的带有分类标记的图像数据进行参数训练，而且参与训练的图像最好与需要进行处理的图像具有较好的相似性。因而，在理想情况下，为了训练VGG网络对近代报纸

①Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," In Advances in Neural Information Processing Systems, 2012, pp. 1097-1105; C. Szegedy et al., "Going Deeper with Convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594; K. He et al., "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

②J. Deng et al., "ImageNet: A Large-scale Hierarchical Image Database," 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

③Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," In International Conference on Learning Representations (ICLR), 2015.

广告图片进行识别，我们应该收集大量的近代报纸广告图片，对它们进行分类，并给属于同一类的每张图片打上一个相同的类别标签。比如，我们可以通过人工归类的方法把对应于同一则广告的不同广告图片归为一类，并打上该类的标记。但是，由于近代报纸广告图片的特殊性，这种做法并不现实。首先，《中国商业广告档案库》中所收录的广告图片数量达到67,000份之多，依靠人工归类打标在短时间内难以做到。其次，尽管档案库中的67,000份广告图片数目已经比较大了，但是对于训练VGG16这样的大规模网络仍旧远远不够。这意味着通过自建训练数据集来训练VGG16网络是不现实的，我们必须采用其他方法解决网络的训练问题。“迁移学习”（transfer learning）是解决这类问题的一个有效手段。

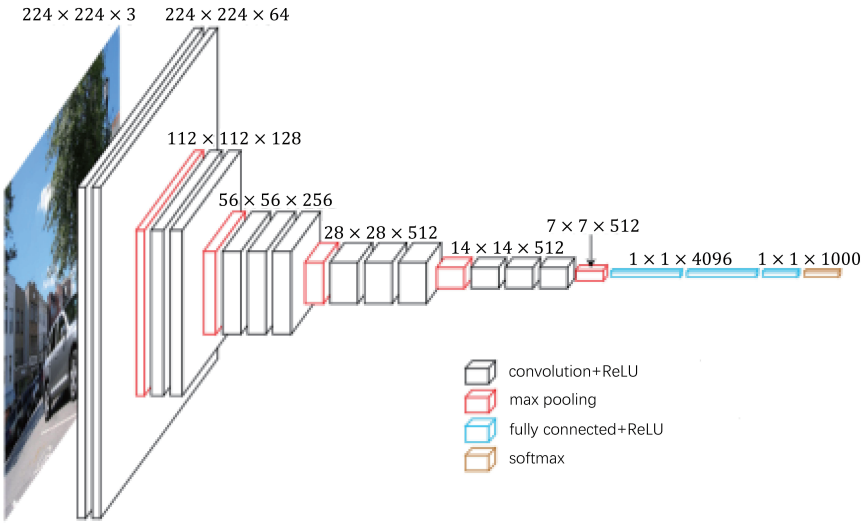

图3　VGG16网络构架示意图[①]

所谓“迁移学习”就是将其他领域或者问题里的训练数据或是训练信息迁移到当前领域或问题中来使用，以解决当前领域或问题中训练信息不足或训练难以实施的问题。为了解决本研究中训练数据难以建立且样本量偏小的问题，我们利用迁移学习的思路，采用图片数据库ImageNet中的图片代替近代报纸广告图片对网络参数进行训练。图片数据库ImageNet包含1,400多万幅图片，涵盖各种动植物、自然实体等2万多个类别，其中超过100万张图片有明确的类别标注，涵盖超过1,000种图像类别，是对VGG16网络进行参数训练的理想数

①Image source: https://heuritech.wordpress.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/.

据集。在这里，我们直接采用经由ImageNet数据训练出来的VGG16网络参数来搭建面向近代报纸广告图片的深度神经网络。这种直接依据ImageNet训练结果进行迁移学习的策略在许多应用中被证明是普遍有效的。考虑到近代报纸广告图片与ImageNet图像数据库中以各种动植物和自然实体彩色照片为主的图片有较大差异，为了防止根据VGG网络提取的特征过度拟合ImageNet中的图像分类，因而远远偏离近代报纸广告图像的特征，我们选择将VGG16网络中最后3个全连接层全部舍弃，而选取网络中13个卷积层中的最后一层输出的7×7×512特征张量作为广告图片的特征。将这个特征张量拉伸为一个向量，我们最终得到了一个长度为7×7×512=25088的数值型向量作为输出图片的特征向量。

第三步是构建关联图谱环节，目标是构建能够反映图片之间相似程度的“关联图谱”。我们会根据图片的特征向量计算每两个待聚类图片的余弦距离，并通过设定阈值，将距离信息转化为关联图谱。特征向量包含了图片的许多信息，一般而言，两张待聚类图片的特征向量距离越小，图片越会被聚成同一类。基于这一性质，我们通过设定特征向量距离的阈值来衡量两个图片是否属于同一类。以每个待聚类图片作为关联图的一个顶点，判断两个待聚类图片的特征向量之间的距离是否小于或等于预设阈值（在近代广告去重问题中使用阈值为0.7）。若两个待聚类图片的特征向量之间的距离小于或等于预设阈值，则将这两个待聚类图片对应的顶点用边相连。因此在特征向量足够准确的理想情况下，同一类图片所对应的顶点之间，都会有边相连。可以通过DFS（Depth First Search，深度优先搜索）或者BFS（Breath First Search，广度优先搜索）算法找到关联图中的各个连通分支，每个连通分支所包含的顶点，代表同一类图片。

第四步是后处理环节，会对第三步得到的每个连通分支进行增边和减边的处理。首先进行减边处理。对于每个连通分支的每条边，判断该条边的两个顶点所连接的顶点数是否小于该连通图的总顶点数，若该条边的两个顶点所连接的顶点数之和小于该连通分支总顶点数，则删除该条边。减边处理能有效地提高图片聚类的精度，防止出现某一个图片由于其特征向量的提取不够精确，导致仅仅与大类图片中的一张图片相似，却被归为该大类的情况。在减边处理完成之后，可以对每个减边处理后的连通图进行增边处理。若两个顶点同属一个连通分支，但是没有相连的边，则在所述两个顶点之间增加一条边。增边处理的目的是将每个连通分支补全为完全图，以表达图片间完整的相似关系。得到后处理的关联图谱后，会将其转化为图片的聚类结果。上述关联图谱中每个连通分支代表一个聚类后的图片类别，每一个连通分支中的点，代表该类别所包含的图片。

三、实验验证

在本节中，我们将通过实验来测评新方法的性能，并和其他方法进行比较。我们共设计实施了两个实验。在第一个实验中，一个仅包含100张近代广告图片的小型数据集被用作验证集来进行方法的测评和比较，其中30张对应于同一条广告的图片应被归为一类，其余70张分别对应于70条不同广告的图片应各自单独成类。进而，在该实验中样本总数N=100，目标类别数K=1+70=71。在第二个实验中，我们选取了一个包含1,000张近代广告图片的中型数据集来进行测评，其中含有3个各分别包含363、348、282个相似图片的图片大类和7个单独为一类的图片小类。换句话说，在该实验中有N=1000及K=3+7=10。这两组测评数据的特性与在实际工作中所遇到的图像数据集的特性较为接近。

在两个实验中，我们共选取了三种能够实现图片聚类的方法进行比较。不同于本研究使用的RNGI方法，其他两个方法的归类过程是先计算每两个图片的数值距离，再在数值距离上设定阈值。如果距离小于某一阈值，则两张图片为同一类。这两个方法中，数值距离选取为欧式距离和hash距离。[①]

我们通过如下两个指标来衡量一个广告图片聚类算法的性能：

错误聚类率=错误聚类图片数/总图片数

未聚类率=未识别为对应类图片数/总图片数

其中，“错误聚类图片数”是指对于A、B两个图片大类，B类的图片被错误识别为A类的图片总数。“未识别为对应类图片数”是指本应归入A类，但未被找到的总数（可能被归入其他类，或者被单独识别为一类）。错误聚类率是衡量算法聚类准确率的重要指标，错误聚类率越小，每个类别中识别错误的图片越少，聚类准确率越高。未聚类率是衡量算法聚类覆盖率的重要指标，未聚类率越小，未被识别进入对应类别的图片数量越少，算法聚类覆盖率越高。用下面的例子来进一步说明两个指标。假设共有5张图片，图片A、B、C、D为第一类，E单独为第二类。而算法识别结果中A、B、C识别为第一类，D为第二类，E为第三类。则算法识别结果中没有出现某一类中两张图片不属于同一类的情况，错误识别率为0/5=0%。另一方面，图片D未被准确识别成第一类，因此未聚类率为1/5=20%。

①X. Lv, Z. J. Wang, “Compressed Binary Image Hashes Based on Semisupervised Spectral Embedding,” IEEE Transactions on Information Forensics and Security, 2013, vol. 8, no. 11, pp. 1838-1849.

表2　近代报纸广告图片聚类方法的测评和比较

	实验一（N=100）		实验二（N=1000）	
聚类算法	错误聚类率	未聚类率	错误聚类率	未聚类率
欧式距离法	9%	17%	30.4%	64.5%
Hash 距离法	15%	4%	53.6%	37.4%
RNGI 算法	0%	0%	0%	2.7%

表2总结了在两组不同的实验条件下，三种不同图像聚类方法的功效。从中可以看出，本文提出的RNGI方法在两个实验中均达到了最高的精度，且比其他方法的结果有大幅度提升。这说明本文设计的方法能够精准高效地实现目标图片聚类。

四、实际应用

我们将本研究的RNGI图片聚类算法应用于所有67,000张近代广告图片进行大规模图片聚类。结果共发现了11,841类相似图片。所有类别包含图片数的大小分布图和主要类别的图片示例如图4：

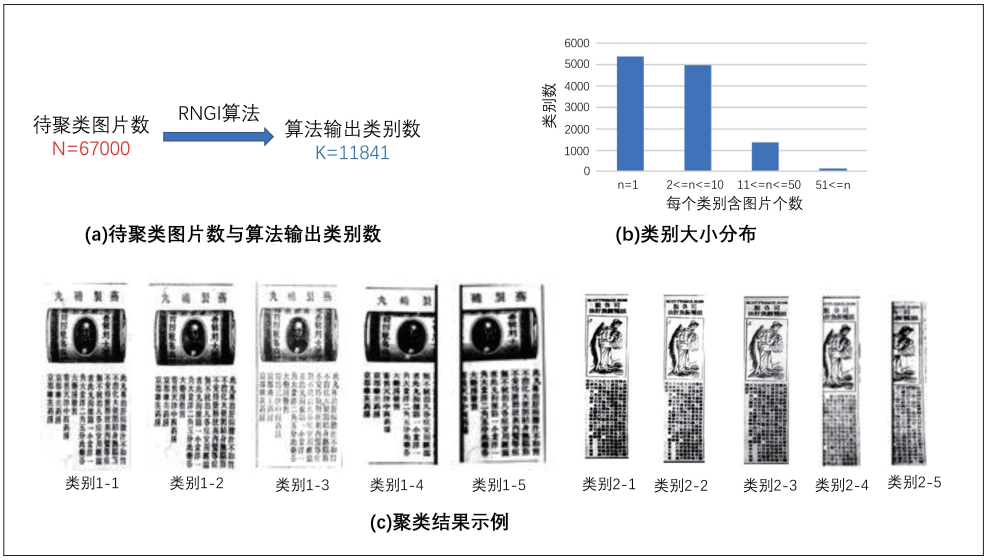

图4　近代报纸广告图片聚类结果

总　结

本文针对海量近代报纸广告分析的问题，跳出单个广告图像内容解析的微观视角，提出了用近代报纸广告图片聚类结果的每一类别进行整合分析，代替传统人文学者逐一分析广告的思路。本文同时提出了RNGI算法，通过深度学习技术将广告图像转化为特征向量，实现计算机自动高效地聚类海量图片。这种新的研究思路和算法极大地提高了近代广告分析的工程效率，该方法能够推广到更广泛的图像分析领域。

Clustering Methods of Modern Newspaper Advertisements via Convolutional Neural Network

Wang Haifeng, Deng Ke, Chen Jing

Abstract: As important historical documents, commercial advertisements in pre-modern newspapers have received more and more attentions from humanists. Content analysis of mass advertisement images in pre-modern newspaper has becoming an active research direction in contemporary humanities. Most existing algorithms process the advertisement images one by one separately. However, due to the special characteristics of advertisement images in pre-modern newspapers, directly applying image recognition to implement content analysis for a single advertisement image often suffers from noticeable errors, and it typically needs a lot of manpower to verify the results, leading to high financial and operational costs. In this article, we propose to abandon the conventional strategy to implement content analysis for individual advertisement images and consider the problem in a global view. We develop in this study a highly efficient tool by measuring the similarity of different advertisement images based on their feature vectors obtained from deep learning. Applying the proposed strategy to process the advertisement images archived in the *China Commercial Advertisement Archive*, we improved the research efficiency by nearly 6 times.
Keywords: Modern Newspaper Advertisements; Content Analysis; Image Processing; Convolutional Neural Network; Clustering Analysis

（编辑：赵薇）