doi:10.3969/j.issn.1672-4623.2017.11.005

众包图像数据采集与聚类分析方法探讨

余晓敏[1]，陈尔刚[2]，季　鹏[2]，郭　涛[2]，秦　昆[2]
（1. 湖北省基础地理信息中心，湖北 武汉 430074；2. 武汉大学 遥感信息工程学院，湖北 武汉 430079）

摘　要：众包图像是由大众经过一定方法获取后通过互联网向公众或相关机构提供的一种开放式图像数据。利用网络爬虫工具在互联网上爬取了一定数量的众包图像，并分别探讨了单张图像聚类方法和多张图像聚类方法，以期为众包技术如何服务于智慧小城镇规划管理提供技术参考。利用 K-means 聚类方法对单张众包图像进行聚类，并探讨了分别利用 Python 语言和 Java 语言编程实现图像聚类的方法；利用层次聚类方法对多张众包图像进行聚类。

关键词：众包图像；智慧小城镇；数据采集；图像聚类；K-means 聚类；层次聚类
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2017）11-0016-02

众包地理数据，也称众源地理数据，是由大量非专业人员志愿获取并通过互联网向大众或相关机构提供的一种开放地理空间数据[1-3]。由大众使用的智能手机或普通相机拍摄并与大众共享的图像数据是其中的一种重要类型，可称为众包图像数据。众包图像数据主要来源于一些互联网网站，如百度图片（http://image.baidu.com/）、Google 图片（http://images.google.com）、Instagram（http://www.instagram.com/）和 Flickr（http://www.flickr.com）等。对众包图像数据进行聚类分析，可有效探测其中的聚类模式，并进行场景分析，对于众包图像在小城镇规划与管理中的应用具有重要作用。

图像聚类是图像数据挖掘的重要工具[4-6]。常用的图像聚类方法包括：K-means 聚类算法、层次聚类算法、SOM 自组织神经网络聚类算法和 FCM 模糊聚类算法等。虽然学者们提出了大量的图像聚类方法，但图像聚类仍面临一些挑战，如聚类算法的鲁棒性问题、高效聚类问题、聚类类别数自动确定问题、特征降维问题和相似或相异度计算问题等[7]。

本文对众包图像的采集与聚类方法进行了研究。首先利用网络爬虫工具爬取众源图像，再利用 K-means 聚类、层次聚类等方法对图像进行聚类分析，探索众源图像中的聚类模式。

1　基于 Python 网络爬虫的众源图像采集

利用 Python 语言编写了一个网络爬虫工具。在爬取网络数据时，首先导入 Python 语言库，包括 os、urllib、urllib2 等；再根据函数输出设置将要处理的网页地址以及图像爬取后存储的路径。例如，进行具体实验时，在“百度图片”中搜索“建筑”关键词，将得到的网址作为实验对象存储在网页的地址变量中，并选择一个文件夹（如 D:\ImageDownload）作为下载图像的存储路径；然后爬虫程序向地址变量发送网络访问请求并读取图片。由于图像数据可能具有多种格式，程序读取网络图片时，考虑了主流数据格式并使这些数据格式的图像都可被正确访问与下载；程序下载时，会实时显示下载进度。

2　单张图像聚类分析

本文选取 K-means 聚类算法对不同场景下的单张图像进行实验，实验过程中分别使用 Python 和 Java 两种语言进行编程。

2.1　基于 Python 的单张图像聚类

利用 Python 语言编程实现单张图像的聚类。图像聚类时，使用 Python 语言模块中的函数，所调用的语言库包含计算机视觉相关的语言模块以及显示图像的语言模块。Scipy.cluster 库是专门用于聚类的模块，使用该模块可快速实现图像数据聚类。首先定义图像数据像素点的 RGB 值，再定义需要处理的图像数据文件的位置路径以及图像名称、迭代次数、聚类类数等信息，最后调用相关函数利用 K-means 聚类算法对图像进行聚类。本文分别选择建筑物、道路等多种场景的众包图像进行了单张图像的聚类实验。使用 Python 语言对其中以建筑物为主体的单张图像进行 K-means 聚类。在分析图像前需判断图像最大聚类的数目，将 K 值从2起逐渐枚举到最大值，找到类簇指标折线的拐点，以确定 K 值，本文选择的类簇指标为 K 个类簇的平均质心距离的加权平均值。分别选择 K=10 和 K=5 进行聚类，

收稿日期：2017-09-06。
项目来源：国家科技支撑计划课题资助项目（2015BAJ05B01）。

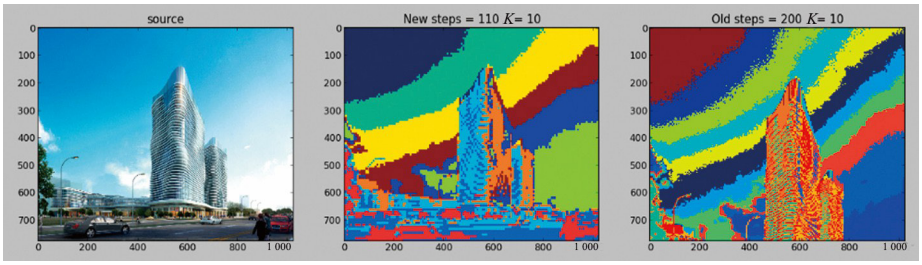


图 1　单张图像聚类（K=10）

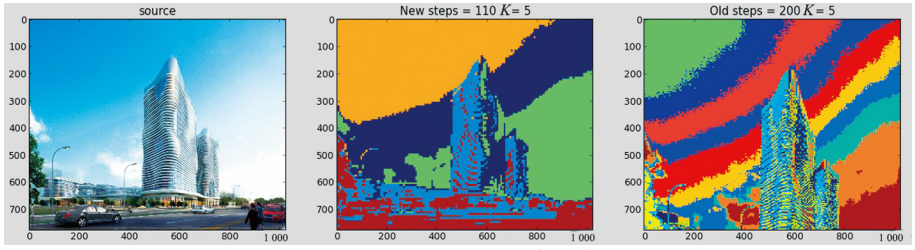


图 2　单张图像聚类（K=5）

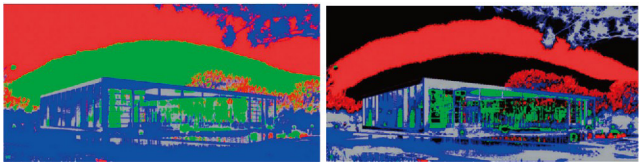
聚类结果存在一定的区别。从图 1、图 2 可以看出，采用 K=5 的聚类结果较好。K 值的选取需根据图像的具体特点确定，这样才能得到较为合适的结果。

2.2　基于 Java 的单张图像聚类

本文也通过 Java 语言实现了 K−means 聚类算法对单张图像的聚类分析。其算法思路为 [8]：①随机选择 K 个中心点；②用 RGB 值作为特征值，计算所有点到 K 个中心点的特征距离，选择距离最近的中心点为其所在的簇；③重新计算 K 个簇的中心；④重复步骤②和③，直至簇类不再发生变化或达到最大迭代值为止；⑤输出结果。

利用 Java 语言编程实现 K−means 聚类算法对单张图像的聚类，对如图 3 所示的图像进行聚类实验。分别采用 K=4 和 K=7 进行图像聚类，K 值确定方法与 §2.1 类似，迭代次数均为 10 次，聚类结果如图 4 所示。

图 3　聚类分析原图



a 图像聚类（K=4）　　b 图像聚类（K=7）

图 4　图像聚类结果

K 值的选择对聚类结果产生了影响，这是由原图像本身的特点所决定的。原图像中相同种类的物体由于光影的原因显示了不同的颜色，这种情况下应选择较小的 K 值将同类物体归并成一类；而 K 值较大时则将原本属于同一类的物体分成了不同类别。

使用 Java 语言进行 K−means 聚类的代码量比使用 Python 多很多，这是因为 Python 语言在实现 K−means 聚类的过程中调用了一些自身语言库中的函数。

3　多张图像的层次聚类分析

本文在进行图像聚类时，针对多张图像选取了层次聚类算法进行实验。使用 Python 语言选择层次聚类算法对多张图像进行聚类。层次聚类算法主要用于多张图像的聚类，其算法思路为基于距离生成一棵和相似度相关的树，首先将每一个样本划为单独的一类，计算两个类间的距离或相似度；然后在所有类中找到距离最近的两个，将它们归到同一个类别中；重复上面的工作直至结束 [9]。

本文使用 Python 的函数库进行层次聚类分析，引用 Numpy 库中的相关函数，还需引用与聚类相关的模块。对建筑物、道路、植被等场景的多张图像层次聚类的实验结果如图 5 所示。

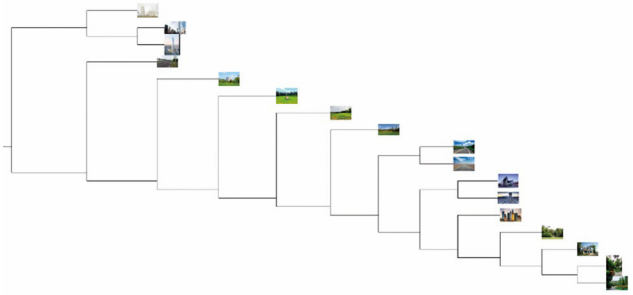


图 5　多张图像的层次聚类结果

图 5 中共使用了 17 张图像进行层次聚类，其中同类景物聚类结果为同层次的有 10 张，不同类图像均归为不同类，根据场景的不同图像逐层进行了归类，没有出现不同景物被归为同一类的情况。从统计结果可以看出，在聚类簇状图里，以 3 种不同场景（建筑物、道路、植被）为主体的图像基本上都被正确地分类到了所属分支，且场景越相似分支距离越近。（下转第 20 页）

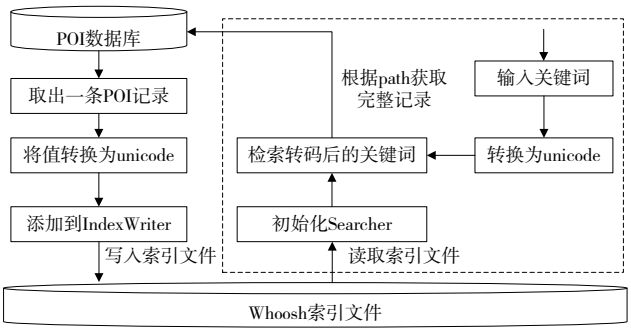


图 2　利用 Whoosh 进行 POI 全文检索流程图

表 1　不同检索方式的实验结果对比

检索关键词	检索方式			
	“like %keyword%”方式		Whoosh 全文检索方式	
	返回记录数 / 条	耗时 /s	返回记录数 / 条	耗时 /s
“测绘”	17	23	17	0.003
“市政府”	506	20	472	0.068
“交通大学”	129	23	8 018	0.094
“西北工业大学”	840	24	840	0.228

由表 1 可以看出，Whoosh 全文检索方式的查询效率比数据库模糊查询方式的效率高出 3 个数量级；另外从检索到的记录数量上来看，Whoosh 全文检索的方式一般要比“like %keyword%”方式返回的记录多，这主要是因为采用 Whoosh 进行全文检索时，先对关键词进行了分词处理。在百万级别上，Whoosh 全文检索方式将文本的检索耗时减少到 ms 级，空间数据检索基本做到了实时响应，提高了发现和利用效率。

实验还发现，Whoosh 全文检索方式的耗时随着字符串长度的增加而增加，主要是因为在全文检索时先对关键词进行了分词处理，增加了检索的复杂度；同时导致两种检索方式检索出的记录数不一致。

4　结　语

本文采用 Whoosh 全文检索引擎对 POI 数据进行了空间数据全文检索实验，与关系数据库模糊检索方式相比，其极大地提高了数据的检索效率。影响全文检索效率的主要因素包括是否对搜索结果进行排序、是否为多域检索等。Whoosh 全文检索引擎的模块化设计和高效率的优势，使得它可将空间检索应用到包括元数据检索在内的空间数据组织和管理领域。本文实验仅针对文档中的词，而没有考虑文档中可能出现的错别字、近义词以及汉语语义的丰富性等因素，要提高全文检索的准确性还需进行错别字纠正、语义分析等复杂的自然语言处理，因此未来的研究方向是实现空间语义级别的检索。此外，全文检索过程中创建的检索文件大小会随数据量（文档的大小和数目）的增加而增加，因此还应研究大数据量条件下全文检索索引的存储方式。

参考文献

[1] 管建和，甘剑峰．基于 Lucene 全文检索引擎的应用研究与实现 [J]. 计算机工程与设计 ,2007,28(2):489-491
[2] 李永春，丁华福．Lucene 的全文检索的研究与应用 [J]. 计算机技术与发展 ,2010,20(2):12-15
[3] 张林曼，吴升．地理编码系统中地址匹配引擎的设计与实现 [J]. 测绘信息与工程 ,2008,33(6):12-14
[4] 方志，夏立新，刘启强．中外全文检索研究的现状及趋势 [J]. 图书情报知识 ,2006(5):71-75
[5] 王富强，王青山，张立朝，等．基于 Lucene 的数据库全文信息检索 [J]. 测绘科学 ,2008,33(3):184-186
[6] 杨柳．空间数据全文检索方法研究 [J]. 测绘工程 ,2012,21(6):8-12
[7] Matt Chaput. Whoosh Documentation: Release 2.7.4 [EB/OL]. (2016-06-19)[2016-07-12].http://media.readthedocs.org/pdf/whoosh/latest/whoosh.pdf

第一作者简介：周海，硕士研究生，主要从事地理编码、空间数据挖掘、GIS 应用开发等工作。

（上接第 17 页）

4　结　语

本文探讨了来自于互联网的众包图像数据的采集与图像聚类方法。对于图像采集部分，使用 Python 语言编写了爬虫工具在网页上爬取合适的图像数据，以便后续的聚类分析。在进行单张众包图像聚类时，分别使用 Python 语言和 Java 语言编程实现 K-means 聚类算法的聚类分析，并在此基础上尝试了不同场景下的图像聚类效果。在进行多张众包图像聚类时，使用 Python 语言编程实现层次聚类算法的聚类分析，后期还需在多场景分类、聚类效果的提升方面进一步研究。

参考文献

[1] Heipke C. Crowd Sourcing Geospatial Data[J]. ISPRS Journal of Photogrammetry and Remote Sensing,2010,65(6):550-557
[2] 单杰，秦昆，黄长青，等．众源地理数据处理与分析方法探讨 [J]. 武汉大学学报（信息科学版）,2014,39(4):390-396
[3] 单杰，贾涛，黄长青，等．众源地理数据分析与应用 [M]. 北京：科学出版社 ,2017
[4] 赵春晖，李雪源，崔颖．混合编码方式的图像聚类算法 [J]. 通信学报 ,2017,38(2):1-9
[5] 艾凌云．基于蚁群算法和粗糙集方法的图像聚类分析研究 [J]. 西北大学学报（自然科学版）,2011,41(5):808-812
[6] 郭庆锐，许建龙，孙树森，等．基于颜色重心和 K-means 的彩色图像聚类分割算法 [J]. 浙江理工大学学报 ,2010,27(4):580-584
[7] 王骏，王士同，邓赵红．聚类分析研究中的若干问题 [J]. 控制与决策 ,2012,27(3):321-328
[8] Likas A, Vlassis N, Verbeek J J. The Global K-means Clustering Algorithm[J]. Pattern Recognition,2003,36(2):451-461
[9] Rokach, Lior, Oded Maimon. Data Mining and Knowledge Discovery Handbook[M]. US: Springer,2005:321-352

第一作者简介：余晓敏，博士，主要从事卫星遥感影像处理和应用。