

基于网络爬虫技术对电子商务多肉物种 资源数据的研究

薛华杰*, 张 宁, 傅怡宁, 徐 飞, 王书平, 徐思越, 印丽萍
上海出入境检验检疫局, 上海 200135

摘要:【目的】近年来,通过电子商务平台获取境外珍贵的多肉物种资源已成为一种重要渠道,大量濒危物种通过第三方物流方式非法流入我国。甄别濒危物种并梳理出多肉物种资源重点查验名单,能够为物种资源查验尤其是植物多肉类濒危物种查验提供参考。【方法】以跨境多肉物种资源交易的热门平台“多肉之家”为研究对象,基于网络爬虫技术平台,获取电商多肉植物 926 条种类数据,并对数据进行筛选和归类分析。【结果】“多肉之家”平台上交易的多肉植物共涉及 23 科 878 种,其中包含 18 种 CITES 附录 I 以及 120 种 CITES 附录 II 濒危物种,约占 16%。进一步对濒危植物进行归类分析发现,濒危植物中仙人掌科和大戟科多肉植物种类最多,分别为 66 和 36 种,两者占总计濒危植物种类的 74%。【结论】网络爬虫技术在获取电商类平台的交易植物的种类数据上具有较好的实用性。

关键词: 多肉; 濒危物种; 电商; 网络爬虫; CITES

Analysis of electronic business data for succulents species resource based on web crawler

XUE Huajie*, ZHANG Ning, FU Yining, XU Fei, WANG Shuping, XU Siyue, YIN Liping
Shanghai Entry-Exit Inspection and Quarantine Bureau, Shanghai, 200135, China

Abstract:【Aim】In recent years, obtaining foreign valuable succulent plants through the e-commerce platform has become an important channel and a large number of endangered species illegally were introduced into China through third-party logistics. Screening out endangered species and proposing a key checklist of succulent plants can provide a reference for the identification of species resources, especially for inspection of endangered succulent plants.【Method】Based on the web crawler technology platform, 926 categories of data on succulent plants were obtained from the popular internet trade platform "succulent plant house" in China. The data were then screened and classified according to CITES.【Result】There were 878 species of succulent plants belonging to 23 families from the data obtained from the "succulent plant house", which included 18 endangered species of CITES Appendix I and 120 endangered species of CITES Appendix II. The endangered species accounted for 16% of all the succulent plants imported. Results show that the most endangered plant species belonged to Cactaceae and Euphorbiaceae. The numbers of species in both families were respectively 66 and 36, accounting for 74% of the total endangered plant species.【Conclusion】Results indicate that the web crawler technology platform has good feasibility in obtaining the species data in the e-commerce platform of plant trading.

Key words: succulents; endangered species; electronic business; web crawler; CITES

随着我国经济的高水平发展和人民生活水平的日益提高,国内的资源已无法支撑未来三四亿中产阶层的物质需求和服务追求。因此,从全球获得资源的“跨境消费”已日益成为一种时尚和潮流。2015 年,全国各邮政口岸累计查验进境国际邮包 3600 余万件,共截获禁止进境动植物及其产品

29765 批次,同比增长 74%。作为全国最大的邮政口岸——上海,情况也不容乐观。2015 年,累计查验进境国际邮包 1504 万件,占全国的 1/3,截获禁止邮寄进境物共 4168 批次,同比增长 42%,涵盖活动物、特殊物品、植物繁殖材料等诸多高风险物品,以及龟甲牡丹 *Paeonia suffruticosa* Andr.、星兜 *Astro-*

收稿日期(Received): 2017-07-13 接受日期(Accepted): 2017-10-20

基金项目: 国家质量基础的共性技术研究与应用(2017YFF0210303); 上海检验检疫局科研项目(HK004-2015)

作者简介: 薛华杰,男,高级农艺师。研究方向: 植物检疫

* 通信作者(Author for correspondence), E-mail: xuehj@shciq.gov.cn

phytum asterias Lem.等珍稀濒危物种与腐烂茎线虫 *Ditylenchus destructor* Thorne、贝克剑线虫 *Xiphinema Bakeri* 等检疫性有害生物(国家质量监督检验检疫总局动植物检疫监管司,2016)。

虽然近年来检验检疫系统在物种资源查验,尤其是濒危物种查验方面取得了不错的成绩,但截获的种类偏少,与 CITES 附录 I 和 II 收录的 3 万余种植物相比(宋云等,2015; Convention on International Trade in Endangered Species of Wild Fauna and Flora,2013),仍有巨大差距。口岸一线人员面对海量的物种名单难以有效地开展日常鉴定工作。

本文基于网络爬虫技术平台,通过抓取电商“多肉之家”多肉植物种类信息进行分析,甄别出其中的濒危物种并梳理出多肉物种资源重点查验名单,为物种资源查验,尤其是植物多肉类濒危物种查验提供参考。

1 材料与方法

1.1 网络爬虫技术

网络爬虫(web crawler),又称为网络蜘蛛(web spider)或 Web 信息采集器,是一个自动下载网页的计算机程序或自动化脚本(孙立伟等,2010)。网络爬虫通常从一个称为种子集的 URL 集合开始运行,它将这些 URL 全部放入到一个有序的待爬行队列里,按照一定的顺序从中取出 URL 并下载所指向的页面,分析页面内容,提取新的 URL 并存入

待爬行 URL 队列中,如此重复上面的过程,直到 URL 队列为空或满足某个爬行终止条件,从而遍历 Web。目前已在搜索引擎制作、电商信息分析等领域得到了推广和应用(孙立伟等,2010;王弘巍,2012;杨颂,2010)。

1.2 抓取“多肉之家”物种资源

“多肉之家”为国内最具人气的多肉植物交流论坛之一,不仅在多肉养护种植等方面处于领先地位,而且是跨境多肉物种资源交易的热门平台。

基于“神箭手”(http://www.shenjianshou.cn/)网络爬虫平台,通过对其采集爬虫代码进行编辑,抓取“多肉之家”(www.drzh.net)收录的多肉植物的学名、中文名、科等信息数据。

基于网络爬虫平台,本次抓取共计获得 23 科 926 种物种资源数据。

2 结果与分析

2.1 “多肉之家”物种资源种类构成情况

将抓取的 23 科 926 种数据,经人工甄别筛选,实际有效为 23 科 878 种,见图 1。前 10 名的科为:景天科 Crassulaceae 493 种,百合科 Liliaceae 125 种,仙人掌科 Cactaceae 66 种,番杏科 Aizoaceae 53 种,大戟科 Euphorbiaceae 42 种,龙舌兰科 Agavaceae 19 种,菊科 Asteraceae 17 种,萝藦科 Asclepiadaceae 14 种,马齿苋科 Portulacaceae 12 种,夹竹桃科 Apocynaceae 9 种,牻牛儿苗科 Geraniaceae 5 种,鸭跖草科 Commelinaceae 4 种,胡椒科 Piperaceae 4 种,龙树科 Didiereaceae 3 种,天门冬科 Asparagaceae 2 种,石蒜科 Amaryllidaceae 2 种,唇形科 Lamiaceae 1 种,西番莲科 Passifloraceae 1 种,葡萄科 Vitaceae 1 种,桑科 Moraceae 1 种,葫芦科 Cucurbitaceae 1 种,苦苣苔科 Gesneriaceae 1 种,薯蓣科 Dioscoreaceae 1 种。

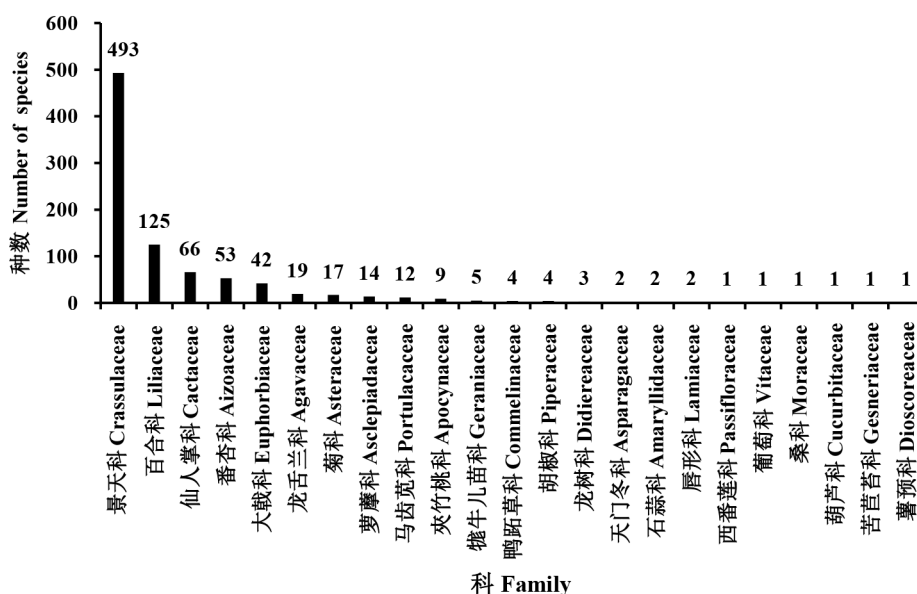


图 1 “多肉之家”物种资源科分布情况

Fig.1 Number of succulent species per family found in www.drzh.net

2.2 “多肉之家”濒危物种与非濒危物种构成情况

“多肉之家”共有 878 种物种资源。其中濒危物种 138 种,CITES 附录 I 濒危物种 18 种,占 2%,CITES 附录 II 濒危物种 120 种,占 14%;其余物种

资源 740 种,占 84%。

“多肉之家”CITES 附录 I 和 CITES 附录 II 濒危物种的详细情况,见表 1。

表 1 “多肉之家”CITES 附录 I 和 CITES 附录 II 濒危物种名单
Table 1 List of endangered species of CITES Appendix I and II in www.drzj.net

科 Family	种或属 Species or genus	CITES	科 Family	种或属 Species or genus	CITES
仙人掌科 Cactaceae	龙舌兰牡丹 <i>Ariocarpus agavoides</i>	I	大戟科 Euphorbiaceae	银手球 <i>Mammillaria gracilis</i>	II
	龟甲牡丹 <i>Ariocarpus fissuratus</i>	I		玉翁 <i>Mammillaria hahniana</i>	II
	连山牡丹 <i>Ariocarpus fissuratus</i> subsp. <i>lloydii</i>	I		克氏丸 <i>Mammillaria hernandezii</i>	II
	黑牡丹 <i>Ariocarpus kotschoubeyanus</i>	I		白鸟 <i>Mammillaria herrerae</i>	II
	龙角牡丹 <i>Ariocarpus scaphiostriis</i>	I		沙堡疣 <i>Mammillaria saboae</i>	II
	三角牡丹 <i>Ariocarpus retusus</i> subsp. <i>trigonus</i>	I		黛丝疣 <i>Mammillaria theresae</i>	II
	星兜 <i>Astrophytum asteria</i>	I		奇仙玉 <i>Matucana madisoniorum</i>	II
	花笼 <i>Aztekium ritter</i>	I		层云 <i>Melocactus amoenus</i>	II
	小极光球 <i>Escobaria minima</i>	I		黄金云 <i>Melocactus broadwayi</i>	II
	帝冠 <i>Obregonia denegrii</i>	I		碧云 <i>Melocactus salvadorensis</i>	II
	精巧丸 <i>Pelecyphora aselliformis</i>	I		大凤龙 <i>Neobuxbaumia polylopha</i>	II
	银牡丹 <i>Pelecyphora strobiliformis</i>	I		狮子王丸 <i>Notocactus pampeanus</i>	II
	多棱玉 <i>Stenocactus multicostatus</i>	I		眩美玉 <i>Notocactus uebelmannianus</i>	II
	姣丽玉 <i>Turbincarpus lophophoroides</i>	I		黄毛掌 <i>Opuntia microdasys</i>	II
	猩猩丸 <i>Turbincarpus valdeianus</i>	I		帝王龙 <i>Ortegocactus macdougallii</i>	II
	尤伯球 <i>Uebelmannia</i> spp.	I		花笠丸 <i>Rebutia neocumingii</i>	II
	美杜莎 <i>Astrophytum caput-medusae</i>	II		习志野 <i>Tephrocactus geometricus</i>	II
	鸾凤玉 <i>Astrophytum myrionostigma</i>	II		大统领 <i>Thelocactus bicolor</i>	II
	姬将军 <i>Austrocylindropuntia subulata</i> f. <i>monst</i>	II		蟹爪兰 <i>Zygocactus truncatus</i>	II
	万重山 <i>Cereus</i> cv. ‘Fairy Castle’	II		皱叶麒麟 <i>Euphorbia decaryi</i>	I
	象牙球 <i>Coryphantha elephantidens</i>	II		峦岳 <i>Euphorbia abyssinica</i>	II
	蜘蛛丸 <i>Discocactus zehntneri</i>	II		红麒麟 <i>Euphorbia aggregata</i>	II
	金星 <i>Dolichothele longimamma</i>	II		铁甲丸 <i>Euphorbia bupleurifolia</i>	II
	金琥 <i>Echinocactus grusonii</i>	II		华烛麒麟 <i>Euphorbia candelabrum</i>	II
	紫太阳 <i>Echinocereus rigidissimus</i> var. <i>rubripinus</i>	II		红彩阁 <i>Euphorbia enopla</i> Boiss. var. <i>enopla</i>	II
	大佛殿 <i>Echinocereus subinermis</i>	II		孔雀丸 <i>Euphorbia flanaganii</i>	II
	千波万波 <i>Echinofossulocactus multicostatus</i>	II		孔雀之舞 <i>Euphorbia flanaganii</i> Cristata	II
	龟甲丸 <i>Echinopsis cinnabarina</i>	II		玉鳞宝 <i>Euphorbia globosa</i>	II
	幻乐 <i>Euphorbia melanostele</i>	II		龙骨冠 <i>Euphorbia grandicornis</i>	II
	巨鹫玉 <i>Ferocactus peninsulae</i>	II		隅田雪 <i>Euphorbia grandidens</i>	II
	日出 <i>Ferocactus recurvus</i>	II		旋风麒麟 <i>Euphorbia groenewaldii</i>	II
	士童 <i>Frailea asterioides</i>	II		柳麒麟 <i>Euphorbia hedyotoides</i>	II
	黑枪丸 <i>Gymnocactus gielsdorfianus</i>	II		魁伟玉 <i>Euphorbia horrida</i>	II
	翠冕冠 <i>Gymnocalycium anisitsii</i>	II		九头龙 <i>Euphorbia inermis</i>	II
	绯花玉 <i>Gymnocalycium baldianum</i>	II		狗奴子麒麟 <i>Euphorbia knuthii</i>	II
	怪龙丸 <i>Gymnocalycium bodenbenderianum</i>	II		帝锦 <i>Euphorbia lactea</i>	II
	勇将丸 <i>Gymnocalycium eurypleurum</i>	II		春峰之辉 <i>Euphorbia lactea</i> f. <i>variegata</i>	II
	绯牡丹 <i>Gymnocalycium mihanovichii</i>	II		白桦麒麟 <i>Euphorbia mammillaris</i> f. <i>variegata</i>	II
	拉根 <i>Gymnocalycium ragonesei</i>	II		贵青玉 <i>Euphorbia meloformis</i>	II
	龙王丸 <i>Hamatocactus setispinus</i>	II		虎刺梅 <i>Euphorbia milii</i>	II
	人字苇 <i>Hatiora epiphylloides</i> subsp. <i>bradei</i>	II		地衣大戟 <i>Euphorbia myrsinites</i>	II
	猿恋苇 <i>Hatiora salicornioides</i>	II		喷炎龙 <i>Euphorbia neohumbertii</i>	II
	光山 <i>Leuchtenbergia principii</i>	II		火巷 <i>Euphorbia nerifolia</i>	II
	杜威丸 <i>Mammillaria duwei</i>	II		布纹球 <i>Euphorbia obesa</i>	II
	金手指 <i>Mammillaria elongata</i>	II		鱼鳞大戟 <i>Euphorbia piscidermis</i>	II
	白玉兔 <i>Mammillaria geminispinia</i>	II		朝驹 <i>Euphorbia pseudocactus</i>	II
	白珠球 <i>Mammillaria geminispinia</i> var. <i>nivea</i>	II		世蟹丸 <i>Euphorbia pulvinata</i>	II

续表 1

科 Family	种或属 Species or genus	CITES	科 Family	种或属 Species or genus	CITES
百合科 Liliaceae	银角珊瑚 <i>Euphorbia stenoclada</i>	II	夹竹桃科 Apocynaceae	银芳锦 <i>Aloe striata</i>	II
	姬麒麟 <i>Euphorbia submammillaris</i>	II		椰子芦荟 <i>Aloe striatula</i>	II
	琉璃晃 <i>Euphorbia susannae</i>	II		千代田锦 <i>Aloe variegata</i>	II
	绿珊瑚 <i>Euphorbia tirucalli</i>	II		惠比须笑 <i>Pachypodium brevicaule</i>	II
	红彩云阁 <i>Euphorbia trigona</i>	II		西巴女王玉栴 <i>Pachypodium densiflorum</i>	II
	土泊大戟 <i>Euphorbia turbiniformis</i>	II	马齿苋科 Portulacaceae	亚阿相界 <i>Pachypodium geayi</i>	II
	贝信麒麟 <i>Euphorbia venenifica</i>	II		非洲霸王树 <i>Pachypodium lamerei</i>	II
	崖大戟 <i>Euphorbia misera</i>	II		光堂 <i>Pachypodium namaquanum</i>	II
	多叶芦荟 <i>Aloe polyphylla</i>	I		象牙宫 <i>Pachypodium rosulatum</i> var. <i>gracilius</i>	II
	绫锦 <i>Aloe aristata</i>	II		白马城 <i>Pachypodium saundersii</i>	II
	洞乳锦 <i>Aloe chabaudii</i>	II		天马空 <i>Pachypodium succulentum</i>	II
	圣诞芦荟 <i>Aloe Christmas Carol</i>	II	龙树科 Didiereaceae	茶笠 <i>Anacampseros baeseckeii</i> Harras	II
	细茎芦荟 <i>Aloe ciliaris</i>	II		银蚕 <i>Anacampseros albissima</i>	II
	塔影锦 <i>Aloe dawei</i>	II		吹雪之松锦 <i>Anacampseros rufescens</i> Sunrise	II
	黑魔殿 <i>Aloe erinacea</i>	II		白花韧锦 <i>Avonia alstonii</i>	II
	青鳄芦荟 <i>Aloe ferox</i>	II		雪嫦娥 <i>Avonia papyracea</i>	II
	羽生锦 <i>Aloe haworthioides</i>	II	龙舌兰科 Agavaceae	红花韧锦 <i>Avonia quinaria</i> subsp. <i>quinaria</i>	II
	帝王锦 <i>Aloe humilis</i>	II		延寿城 <i>Ceraria pygmaea</i>	II
	唐力士 <i>Aloe melanacantha</i>	II		魔针地狱 <i>Alluaudia montagnacii</i>	II
	七宝锦 <i>Aloe mutabilis</i>	II		亚龙木 <i>Alluaudia procera</i>	II
	不夜城 <i>Aloe nobilis</i>	II		马达加斯加树 <i>Didierea madagascariensis</i>	II
	翡翠殿 <i>Aloe squarrosa</i>	II		皇后龙舌兰 <i>Agave victoriae-reginae</i>	II

2.3 “多肉之家”CITES 附录 I 濒危物种构成情况

“多肉之家”CITES 附录 I 濒危物种共计 3 科 18 种。其中仙人掌科 16 种,约占 88%;百合科 1 种,约占 1%;大戟科 1 种,约占 1%。

2.4 “多肉之家”CITES 附录 II 濒危物种构成情况

“多肉之家”CITES 附录 II 濒危物种共计 7 科

120 种。其中马齿苋科 7 种,约占 6%;龙舌兰科 1 种,约占 1%;龙树科 Didiereaceae 3 种,约占 2%;百合科 16 种,约占 13%;仙人掌科 50 种,约占 42%;大戟科 35 种,约占 29%;夹竹桃科 8 种,约占 7%。

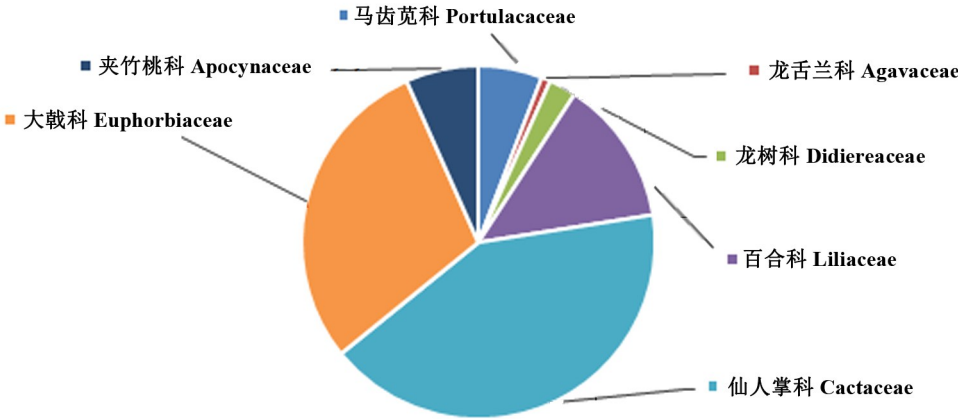


图 2 “多肉之家”CITES 附录 II 濒危物种构成情况

Fig.2 Proportion of endangered species under CITES Appendix II in www.drzj.net

3 讨论

近年来物种资源保护得到质检总局、海关等相关口岸执法部门的高度重视,口岸物种资源查验点

建设不断加强,配套的政策法规也不断完善,但业务量大、人员紧张、技术支撑不足等客观因素限制了物种资源查验工作的开展,而缺乏口岸重点查验

名单,则是导致目前查验工作难以有效推进的主要因素。植物濒危的原因非常复杂,其中自身因素、物理因素、生物因素和人类活动是导致濒危的四大主要原因(盛茂银等,2011)。目前,CITES 附录 I 和 II 收录的 3 万余种植物,并不都是由于国际间非法买卖交易等人类活动导致其濒危。因此,可通过贸易买卖的数据,对 3 万余种植物进行筛查,甄别出易被国际间非法买卖交易的物种,从而建立口岸重点查验名单。“多肉之家”是国内养护交流、交易买卖的多肉植物最具人气的电商平台之一,买卖双方借助该平台进行交易。因此,对该网站的数据进行分析就可以筛选出多肉物种资源重点查验名单(表 1)。

“多肉之家”收录的仙人掌科植物有 66 种,全部是濒危物种,其中龟甲牡丹、星兜等 16 种植物为 CITES 附录 I 濒危物种,帝王龙 *Ortegocactus macdougallii* Alexander、白鸟 *Mammillaria herrerae* Werderm. 等 50 种植物为 CITES 附录 II 濒危物种;“多肉之家”收录的大戟科植物有 42 种,濒危物种约占 86%,其中皱叶麒麟 *Euphorbia decaryi* Guillaumin 为 CITES 附录 I 濒危物种,华烛麒麟 *Euphorbia candelabrum* var. *bilocularis*、柳麒麟 *Euphorbia hedyotoides* N.E. Brown 等 35 种植物为 CITES 附录 II 濒危物种。因此,建议口岸加强对仙人掌科和大戟科的物种资源查验。造型奇特可爱的景天科植物受到多肉爱好者的热捧和追逐,在“多肉之家”电商中被收录了 493 种,约占总数的 56%,但没有一种是濒危物种。不过,景天科植物的种类和数量众多,同时口岸检疫时曾多次截获腐烂茎线虫、贝克剑线虫等检疫性有害生物,因此口岸及其实验室仍需注意对景天科植物的检疫。

当前口岸物种资源的查验主要采取“人—机—犬”三位一体综合查验模式(吴晶等,2016),只能初步判定是否为多肉物种。建议开展现场在线自动识别搜检技术的研究,开发诸如花伴侣、形色、微软识花等物种资源识别 app 软件,从而提高工作效率。此外,物种资源的识别主要是依靠形态学方法,但是大多数物种在口岸并不能被迅速准确识

别,只能按照《中华人民共和国邮政法》等法律法规规定(中华人民共和国中央人民政府,2009),放置一个多月后才能抵达实验室,导致很多样品枯萎、腐烂。另外,截获的不少物种有很多尚处于营养生长阶段,缺少只有在繁殖阶段才会出现的花、果等鉴定到种的特征,上述情况严重影响了形态识别的效果。随着分子生物学,特别是 ITS、psbA-trnH 等 DNA 条形码技术的发展,已达到属甚至种的准确识别(任保青和陈之端,2010)。因此,建议口岸查验实验室加大对 DNA 条形码技术的研发,为物种资源查验提供技术保障。

参考文献

- 盛茂银,沈初泽,陈祥,田兴军,2011. 中国濒危野生植物的资源现状与保护对策. 自然杂志, 33(3): 149–154.
- 任保青,陈之端,2010. 植物 DNA 条形码技术. 植物学报, 45(1): 1–12.
- 宋云,许瑾,赵竹,李明福,陈克,黄炎,陈乃中,2015. 我国进出口贸易中生物物种资源调查分析. 植物检疫, 29(5): 25–29.
- 孙立伟,何国辉,吴礼发,2010. 网络爬虫技术的研究. 电脑知识与技术, 6(15): 4112–4115.
- 王弘巍,2012. 基于亚马逊网站的特定电子商品爬虫设计与实现. 硕士学位论文. 长春: 吉林大学.
- 吴晶,李井干,殷连平,高颖,2016. 中澳新进境邮寄物检疫的比较与分析. 植物检疫, 30(5): 14–17.
- 杨颂,2010. 面向电子商务网站的增量爬虫设计与实现. 硕士学位论文. 长沙: 湖南大学.
- 国家质量监督检验检疫局动植物检疫监管司. 上海局严厉打击虚假申报入境邮包. (2016-05-30) [2017-07-03]. http://dwjyjs.aqsiq.gov.cn/gzdt/201605/t20160530_467478.htm.
- 中华人民共和国中央人民政府,2009. 中华人民共和国邮政法. (2009-04-24) [2017-07-03]. http://www.gov.cn/flfg/2009-04/24/content_1295123.htm.
- Convention on International Trade in Endangered Species of Wild Fauna and Flora, 2013. Appendices I, II and III. (2013-09-13) [2017-07-03]. <https://www.cites.org/eng/app/appendices.php>.

(责任编辑:郭莹)