谢友柏设计科学研究基金

申请书

项目名称：分布式资源环境中产品设计功能知识管理与集成方法研究

申 请 人：胡亚红

依托单位：浙江工业大学

通讯地址：杭州市西湖区留和路 288 号

邮政编码：310023

电子邮件：huyahong@zjut.edu.cn

电　　话：13073638971

申报日期：2021.12.12

1. 基本信息

申请人信息	姓名	胡亚红	性别	女	出生年月	1971.02
	学位	博士	职称	副教授	在站博士后	否
	电话	13073638971	国别	中国		
	通讯地址	杭州市西湖区留和路 288 号浙江工业大学计算机学院				
	电子邮件	huyahong@zjut.edu.cn				
	研究领域	合作设计；分布式系统；大数据处理				
依托单位	单位名称	浙江工业大学				
	联系人	林琼	电话	0571-88320515		
	地址	杭州市拱墅区潮王路 18 号				
	电子邮箱	waglin@zjut.edu.cn				
项目信息	项目名称	分布式资源环境中产品设计功能知识管理与集成方法研究				
	英文名称	Research on Product Functional Knowledge Management and Integration in Distributed Design Resource Environment				
	研究期限	三年				
	申请经费	20 万				
	中文关键词	功能知识表示；知识搜索；聚类算法；分布式设计资源环境				
	英文关键词	functional knowledge representation; knowledge searching; clustering algorithm; distributed design resource environment				

2. 项目摘要

中文摘要	目前，设计知识资源以水平分布结构存在于多个企业之间，设计人员或一个设计团队仅依靠自己掌握的知识来完成产品创新设计变得十分困难。因此，研究分布式资源环境中产品设计知识的管理与高效运用具有重要的实际应用价值。本申请针对产品设计功能知识的管理与集成方法进行研究，主要包括以下四个方面的内容：(1)产品功能知识标准化处理；(2)功能知识单元集成算法研究；(3)分布式环境下功能知识单元的存储管理；(4)功能知识单元簇生成算法研究。通过对产品功能知识单元的有效管理，研究成果能够根据产品功能需求搜索到产品设计需要的功能知识集和功能知识链，从而帮助设计师快速完成产品设计方案，提高产品设计效率和质量。
英文摘要	Nowadays, design knowledge resources exist in multiple enterprises in a horizontal distribution structure. It is very difficult for one designer or a design team to complete product innovation design only by relying on their own knowledge. Therefore, the research on the management and efficient application of product design knowledge in distributed resource environment has important practical value. This research focuses on the management and integration methods of product design functional knowledge. The main contents include the following four aspects: (1) Standardization of product functional knowledge representation; (2) Research on integration algorithms of functional knowledge units; (3) Storage management of functional knowledge units in distributed environment; (4) Research on clustering algorithms of functional knowledge units. Through effective management of product functional knowledge units, the functional knowledge sets and functional knowledge chains which satisfy the product functional requirements can be obtained. The outcomes of this project will enable designers to complete high-quality product designs efficiently.

3. 报告正文

3.1 研究内容及年度进展计划

3.1.1 研究背景与意义

人们在日常生活以及生产实践中常常会自觉或不自觉地进行各种设计。设计为人类有目的活动规划实施结果和路径[1]，成功的设计能够保证活动顺利实施后可以达到人们的预期目标。进行设计需要知识，设计科学的基本定律指出，设计以已有知识为基础，即任何设计问题的解决方案都是由已有知识构成的[2]。在信息爆炸的时代，人类拥有的知识越来越多。充分利用前人留下的宝贵知识资源，能够为高效、高质量地进行设计提供极大的帮助。

随着生活水平的提高，人们对产品的要求已不仅仅是满足物质上的需求，还对产品能够提供的社会功能和精神功能有了更高的要求，因此产品的设计需要更多领域的知识。在这样的环境下，设计以在企业内进行为主逐渐转化为多个企业间的合作设计[3-5]。设计知识资源的结构也由在一个企业中的垂直分布结构转变为在多个企业之间的水平分布结构[1]。于是，设计人员或一个设计团队仅利用自己掌握的知识来完成产品创新设计变得十分困难。设计知识以各种结构化或非结构化的方式被分布式地存储，这给知识的高效使用带来了极大的难度。可以通过一个简单的产品设计来说明这个问题。

进行垃圾分类可以减少环境污染，实现再生资源的利用。但是不少人对于垃圾分类的重要性和如何正确地进行垃圾分类认识不清，所以软件团队决定开发一款帮助公众提高垃圾分类意识和水平的软件。通过用户调查和资料收集，设计团队完成了产品的需求分析，确定了此款软件的主要功能，包括：用户的注册、登录；垃圾分类知识的展示和学习；垃圾的分类识别。每个功能的具体描述见表 1。

表 1 垃圾分类软件功能描述

功能编号	功能名称	功能描述
R1	用户注册	输入：用户名；密码
		输出：注册成功与否标识
		约束：用户名长度在 6-10 位之间； 密码长度在 6-10 位之间，必须同时包含大写字母、小写字母和数字
R2	用户登录	输入：用户名；密码
		输出：登录成功与否标识
		约束：无
R3	垃圾分类知识展示	输入：文字、图片、音频文件、视频文件
		输出：包含文字、图片、音频和视频等内容的页面
		约束：文本字数≤ 300 字 图片大小≤ 3M

		音频文件大小≤10M 视频文件大小≤ 100M
R4	垃圾分类知识的学习	输入：按下键盘 Enter 键或单击鼠标左键
		输出：文件播放
		约束：无
R5	垃圾的分类识别	输入：垃圾的图片
		输出：垃圾的分类
		约束：图片文件大小≤ 3M

为了实现系统的功能，设计师团队对已掌握的知识进行了梳理，发现已具备完成功能 R1、R2 的知识单元 FK1 和 FK2。通过资料的查找，设计师找到了实现R4的知识单元FK4,但是没有找到能实现R3"垃圾分类知识展示"和R5 "垃圾的分类识别" 的知识单元。于是，设计师对 R3 和 R5 进行了功能分解，得到的子功能描述如表 2 所示。

表 2 对 R3 和 R5 进行功能分解的结果

父功能编号	子功能编号	子功能名称	子功能描述
R3	R31	文件上传	输入：文件
			输出：操作成功与否的标识
			约束：无
	R32	信息存入数据库	输入：文本框中文本或文件，数据库中对应的表名
			输出：操作成功与否的标识
			约束：无
	R33	数据库信息搜索	输入：文本中的关键字或文件名关键字，数据库中对应的表名
			输出：满足要求的文字或文件
			约束：无
	R34	排版	输入：文本框或图片的原始位置和大小； 文本框或图片新的位置和大小
			输出：在给定位置或缩放后的文本框或图片
			约束：给定的位置或缩放比例不能超过显示器能够显示的范围
R5	R51	图片上传	输入：图片文件
			输出：操作成功与否的标识
			约束：无
	R52	图片匹配	输入：图片文件
			输出：垃圾类别
			约束：无

对 R3 和 R5 的各子功能进行分析发现，R31“文件上传”和 R51“图片上传”都可以使用知识单元 FK31 完成；R32“信息存入数据库”这个功能可以使用知识单元 FK1 完成；R33“数据库信息搜索”可以使用 FK2。但是 R34 和 R52 没有现有的知识单元与之对应。

于是设计师只能继续对 R34 和 R52 进行功能分解，得到表 3 的结果。

表 3 对 R34 和 R52 进行功能分解的结果

父功能编号	子功能编号	子功能名称	子功能描述
R34	R341	对象移动	输入：对象初始位置的左上角和右下角顶点坐标； 对象目标位置的左上角和右下角顶点坐标
			输出：操作成功与否的标识
			约束：输入数据保证对象大小不发生变化； 移动到新位置的对象能够正常显示
	R342	对象缩放	输入：对象初始的长度和宽度 长度方向的缩放比例 宽度方向的缩放比例
			输出：操作成功与否的标识
			约束：给出的缩放比例保证对象能够正常显示
R52	R521	图片指纹提取	输入：图片文件
			输出：图片对应的指纹
			约束：无
	R522	图片指纹存储	输入：图片指纹，图片数据库中对应的表名
			输出：操作成功与否的标识
			约束：无
	R523	图片匹配	输入：图片指纹，图片数据库中对应的表名
			输出：垃圾类别
			约束：无

再次进行知识搜索，设计师为 R34 和 R52 的所有子功能都找到了合适的知识单元。这时，设计师才拥有了进行“垃圾分类”软件设计需要的所有知识。

从这个例子不难看出，即使是设计这样一个功能简单的系统，设计师也需要花费大量的时间去查找需要的设计知识，更不用说要完成功能复杂的产品设计。由于人工搜索范围和搜索速度的限制，很多情况下，设计师无法找到最适合的知识，从而导致产品开发周期加长、开发成本变高。

因此非常有必要研究知识有效管理方法和智能化的知识搜索算法，以帮助设计人员在进行设计时能够方便地查找到需要的知识，并形成满足用户要求的产品设计方案。

3.1.2 研究目标

本项目的研究目标是在对产品设计的功能知识进行规范化表示和管理的基础上，研究高效的功能知识集搜索算法和功能知识单元链的生成算法，为设计师进行产品集成创新提供支撑。

3.1.3 研究内容

为了能够为产品设计提供所需功能知识，本项目的核心研究内容是功能知识单元的集成算法，包括功能知识单元集的生成算法及功能知识单元链的生成算法。对功能知识进行有效的管理是提高知识单元集成算法效率的关键。因此，本项目还将研究分布式知识资源环境中功能知识的存储和管理方法。知识单元表示的标准化研究可以保证知识单元搜索结果的正确性，将功能知识单元按功能划分为功能簇能够提高功能知识单元的集成效率，因此本项目还将研究知识单元的标准化处理方式和知识簇的生成算法。研究内容的结构图如图 1 所示。

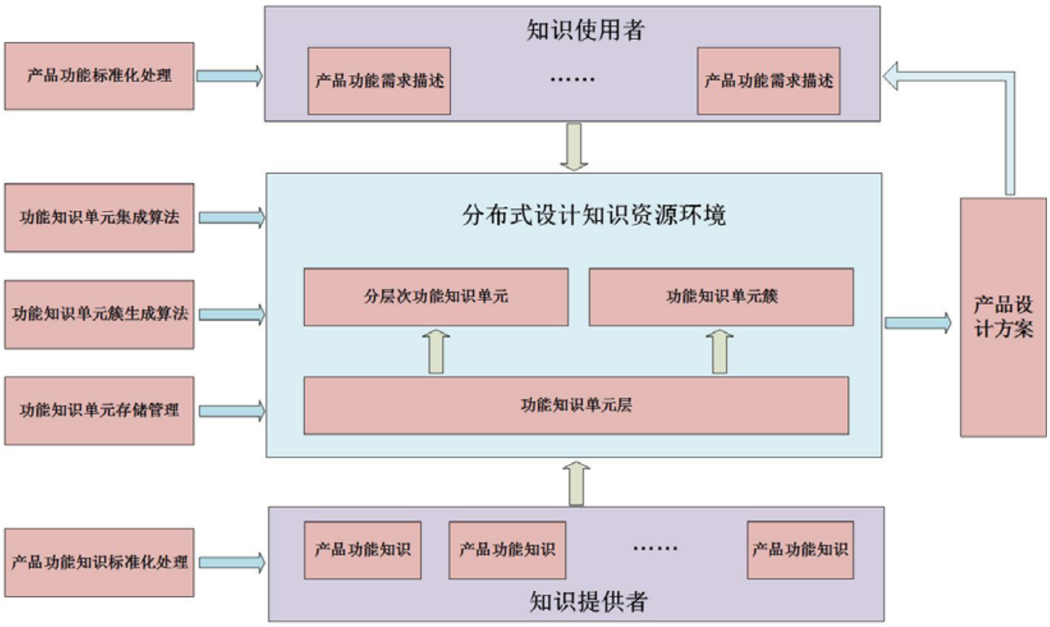


图 1 研究内容结构图

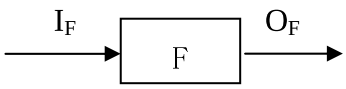
(1)产品功能知识标准化处理

为了实现知识的共享，知识的拥有者需要在分布式设计知识资源环境中发布设计知识，知识的使用者则按一定的规则提交产品的功能需求。为了实现知识发布和知识搜索，产品功能知识需要有合理的表示方法。因为产品的功能需求将被转化为对产品功能的描述，为方便起见，下文不再区分功能需求的表示和功能的表示。已有研究表明，使用功能需求的类型、功能的类型、说明功能输入/输出的关键词和关键词的特征值，能够将产品的功能准确地描述出来[1]。

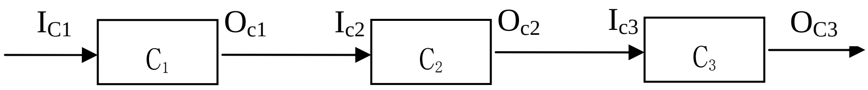
因为表述习惯的差异，不同的知识提供者在描述同样的功能知识时会使用不同的方式。例如相同的质量值，有人用 1Kg，有人用 1000g；同样含义的关键词，有人描述为太阳光，有人表述为日光。如果在进行知识单元搜索时直接根据用户原始的功能表示进行匹配，有时会得出错误的查询结果。因此需要对知识单元表示进行预处理，完成度量单位标准化和关键词表示的标准化。度量单位的标准化将通过特征值单位的判断和特征值数据的转化进行，关键词的标准化则将通过构建同义词库完成。经过预处理得到的标准化的功能知识单元将进入产品设计知识库。对于用户给出的产品需求中包含的约束条件，将研究如何采用特征值进行表示。

(2) 功能知识单元集成算法研究

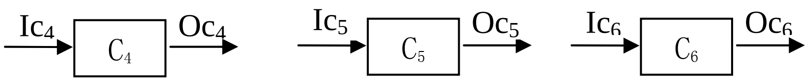
设计师进行产品设计时，如果不能找到满足产品功能需求的知识，就需要对产品功能进行分解，形成多个子功能。如果某个子功能仍较为复杂，可以进一步将其分解为更小的孙功能。图 2 给出了两种不同的功能分解方式，即父功能分解为多个有输入输出约束关系的子功能或分解为多个功能独立的子功能。当然，更多情况是混合使用上述两种方式来完成父功能的分解。



(a) 父功能



(b) 父功能分解为多个有输入输出约束关系的子功能



(c) 父功能分解为多个功能独立的子功能

图 2 父功能分解为多个子功能

图 2 中符号 I 对应功能的输入，O 对应其输出。图 2(a)中，父功能 F 的输入和输出分别为 I_F 和 O_F。当没有功能知识可以完成此功能时，需要将此功能分解为多个子功能。图 2(b)和(c)给出了功能分解的两种情况。图 2(b)中，父功能 F 被分解为三个输入输出相匹配的子功能 C_1、C_2 和 C_3。此时，C_1 的输入 I_{c1} 与 F 的输入匹配，C_3 的输出 O_{c3} 与 F 的输出匹配；并且 O_{c1} 和 I_{c2} 匹配，O_{c2} 和 I_{c3} 匹配。当为子功能 C_1、C_2 和 C_3 都找到设计知识后，设计 F 需要的知识也就具备了，即 $C_1 \rightarrow C_2 \rightarrow C_3$ 构成了实现 F 的功能知识单元链。因为各子功能间存在的约束关系，

在进行父功能设计时，还需要考虑子功能的系统行为和结构之间的匹配。这部分内容的研究非常重要，将在后续的研究工作中进行，本项目暂不涉及各功能知识单元间系统的行为知识和结构知识之间的匹配。

图 2(c)中，父功能 F 被分解为三个功能彼此独立的子功能 C_4、C_5 和 C_6。其中，I_F 与 I_{C4}、I_{C5} 和 I_{C6} 的并集匹配，O_F 与 O_{C4}、O_{C5} 和 O_{C6} 的并集匹配。当子功能 C_4、C_5 和 C_6 设计完成，父功能 F 的设计也就完成了。可以将$\{C_4,C_5,C_6\}$称作实现 F 的子功能知识集。

从图 2 可以看出，进行父功能的分解，需要在分布式的设计知识资源环境中进行功能知识的搜索和/或功能知识单元链的生成。因为知识数量的巨大，依靠人工完成难度过大。因此，本项目将研究功能知识的集成算法（包括知识集生成算法和功能知识单元链生成算法），让计算机辅助设计师完成产品功能的分解，提高新产品设计的效率和质量。

(3) 分布式环境下功能知识单元的存储管理

目前互联网技术得到了广泛应用，大量设计相关知识分布式地存储于设计知识资源环境中。进行产品设计需要用到各种知识，设计师不具备的知识就要到设计资源环境中去寻找。功能知识单元的存储方式会对功能知识集成算法效率产生重要的影响，因此本项目将研究在分布式环境中如何对功能知识单元进行有效的存储和管理。

在存储知识单元信息时，不但需要存储知识单元的输入、输出和功能关键字等，同时还必须将各单元之间的相互关系进行存储。针对不同功能需求的类型，如物质需求、社会需求和精神需求，本项目将研究对应的存储方案。

产品的设计是层次化的设计，一个产品的功能可以分解为若干个子功能。设计师仅需要了解当前产品功能的下一层子功能是否能够满足自己的设计要求，而不需要对子功能的下层功能进行研究。当一个产品设计完成后，得到的产品系统及其子系统（如果原先不存在）可以作为新的功能知识单元为以后的设计提供服务。因此将研究分层次产品功能知识的存储，以缩小功能知识单元集成算法的搜索空间，提高算法效率。

(4) 功能知识单元簇生成算法

根据某种分类原则，将具有同一特征或者应用的知识条和/或较小的知识集放在一起，就形成了知识簇[1]。按照这一思路，可以将具有相同输入或输出的功能知识单元聚合为功能知识单元簇。当进行功能知识单元搜索时，根据目的输入和目的输出查找对应的功能知识单元簇比分别查找各个满足要求的功能知识单元更加高效。本项目将分析现有的聚类算法和产品功能知识表示的特点，对功能知识单元簇的生成算法进行研究。

3.1.4 年度进展计划

2022.1-2022.12

(1) 收集与项目研究内容相关的文献，完成已有研究方法和研究结果的调研；

(2) 设计实现产品功能表示的预处理算法，以获取产品功能的标准化表示；

(3) 分析功能知识单元表示的特点，设计知识库的结构。

2023.1-2023.12

(1) 完成已有的功能知识单元和知识集的层次分析；

(2) 完成已有功能知识单元的知识库分层存储；

(3) 设计实现知识单元簇生成算法。

2024.1-2024.12

(1) 设计实现功能知识单元的集成算法；

(2) 对项目的研究方法和结果进行总结，完成结题报告。

3.2 要解决的核心问题

(1) 功能知识单元集成算法研究

现有的功能知识单元集成算法效率较低、可扩展性差，不适合于功能知识单元数量巨大的应用场景。针对这些问题，本项目将研究基于深度学习和强化学习的功能知识单元集成算法，以快速完成分布式知识资源环境中功能知识集和功能知识链的生成。

(2) 产品功能信息的预处理

分析产品功能规范化表示方法的特点，研究同义词库的构建方式和特征值规范化表示方法，使得具有规范表示的功能知识存入知识库，提高功能知识单元集成算法的性能。

3.3 国内外对该问题研究现状评述及该问题与基金资助指南范围的关联

3.3.1 国内外对该问题研究现状评述

(1) 功能的表示

产品设计的目标是提供具有一定功能的、能够满足用户需求的产品。如何描述功能会对设计过程产生重要的影响，因而很多学者研究了对功能的描述方法[6]。较为经典的模型包括 Pahl 等的 FES（Function-Effect-Solution）模型[7]、Gero 的 FBS（Function-Behavior-Structure）模型[8-10]和 Suh 的公理设计（Axiomatic Design）[11]等。FBS 将设计过程描述为功能、行为和结构之间的转换过程。为了提高产品的竞争力，在 FBS 模型的基础上，文献[12]提出了 PFWSB（Purpose -Function -Working theory-Structure-Behavior，意图-功能-原理-结构-行为）本体模型。该模型强调了以用户为中心的设计思想，在进行产品设计时充分挖掘用户需求（包括

功能性需求和非功能性需求）和产品使用的环境约束，从而保证设计出的产品能够满足用户的要求。基于 PFWSB 模型，该研究分别从意图、功能、原理、行为和结构等层面对设计对象的知识进行了表示。系统建模语言 SysML 支持由软硬件、数据和人共同组成的复杂系统结构的说明、分析、设计及校验[13]。但是它着重于产品系统的分析和优化,不适合新产品从无到有的设计过程。用户需求分为三类，即物质需求、精神需求和社会需求。产品所能提供的功能可以分为四类，即转变、支撑、存储和激励功能[1]。针对以上分析，采用“需求类型+功能类型+目的输入的关键词+目的输出的关键词”的方法进行产品功能的描述，不但为产品功能的表示提供了规范化的方法，同时为在分布式环境中进行功能知识单元搜索提供了极大的方便。

(2) 知识的表示及存储

知识的表示是指用易于计算机处理的方式来描述人脑知识的方法。知识表示是客观事物的机器标识，是一组本体约定和概念模型，是支持推理的表示基础[14]。符号表示方法是传统的知识表示方法,包括逻辑描述[15][16]、霍恩规则逻辑、产生式系统、框架系统[17][18]和语义网络[19]等。这些知识表示方法的缺点是不容易刻画隐性的知识，知识获取需要由专家人工完成。随着深度学习领域研究的发展，用参数化的向量来表示实体和实体间的关系得到越来越多的应用。这种知识表示学习可以在低维空间中高效计算实体和关系的语义联系，从而解决数据稀疏问题，提高知识获取和推理的性能[20]。

知识通常存放于数据库中以方便进行知识查找。知识图谱是一种结构化的知识库，旨在利用图结构建模、识别和推断事物之间的复杂关联关系。知识库根据结构可以分为关系型数据库和图数据库。与关系数据库相比，图数据库能够提供高性能的关系查询、适用多来源数据及进行复杂图的分析，因此得到了广泛的应用。知识图谱在智能制造领域已产生很多的研究成果。Leng等通过分析制造关系的社交交互上下文环境，得到跨企业制造业需求能力匹配的信息[21]。通过利用本体概念自身的语义和概念间的层次结构，文献[22]研究了基于本体语义块相似性匹配的设计知识更新方法。文献[23]对产品设计案例本体进行解析，建立数据库关系模式与本体之间的映射关系，完成了对源数据库的查询。使用知识图谱进行知识的存储有助于进行知识重用。文献[24]结合产品设计过程知识的专业特点，采用四元组形式描述设计过程知识本体，并提出基于双层模块化封装技术的设计过程知识的重用。

(3) 功能知识的集成

充分利用现有产品的功能知识能够降低新产品设计成本，缩短设计时间，

因此已有不少针对知识重用和集成的研究工作。机器人操作系统（ROS，Robot Operating System)提供了一套操控机器人的核心软件，通过对这些软件的重用或扩展，可以快速完成在新平台上软件的开发。但是由于 ROS 中软件包数量巨大，人工去搜索符合功能要求的软件费时费力。文献[25]构建了 ROS 软件包知识图谱 ROSKG（ROS Package Knowledge Graph），用以完成软件包信息的描述、特征提取。ROSKG 能够有效支持对满足一定功能需求的软件包的搜索。该研究通过匹配需求与 ROS 包的描述信息进行，没有涉及到知识集成的问题。

Pahl 等[7]提出将一个产品功能进行分解，通过对分解的各个子功能模块进行输入和输出流的组合来实现产品的目的输出。Liu 等提出在进行新产品设计时，可以先对已有的类似产品进行功能分析和分解。当现有产品的某个功能对应的输出动作和目标产品的输入动作匹配时，可以对此功能设计进行重用，从而设计出满足不同使用环境和不同用户的新产品[26]。该研究中对已有产品的搜索是依靠设计师的经验人工进行专利搜索完成的。

由于设计知识资源数量巨大并且具有水平分布的结构，单个设计师或设计团队进行人工的功能单元分解和组合具有很大的难度，使用智能规划的推理技术，有助于完成设计中的功能知识的自动化集成[27]。智能规划是指从一个初始的状态，通过一系列的动作，得到目标状态。设计状态使用基于流的方式进行表达，功能解的动作采用一阶谓词进行描述[28]。文献[28]还提出了三种启发式的功能知识集成方法，即前向链接集成、后向链接集成和前后向链接综合集成。当给定输入流和目的输出流时，论文提出的知识集成方式可以构造出满足要求的规划解。其中两个流匹配是指它们的类型匹配。文献[29]采用链表存储功能单元，并且定义了功能单元的输入/输出的匹配条件。该文献提出的功能单元链的生成算法可以解决功能单元访问操作冗余的问题和功能单元环的问题。文献[30]更进一步定义了功能单元的匹配度，即输入中被匹配的特征数与输入总特征数的比值。当存在多条功能单元链时，优先选择匹配度大的功能单元链。在获得的功能单元链的基础上，该研究得到了完整的功能单元集的产生过程图，并且对产生的功能结构图进行了评估。

使用有向图能够非常好地表示出各功能单元之间的关系。图模型中，节点表示功能单元,节点之间的边表示功能单元之间的输入/输出匹配关系。这样，查找从目标输入到目标输出的功能单元链的问题，就转化为从给定的起始点到给定的目标点间的路径规划问题。文献[29]和[30]都使用了深度优先搜索算法在图中进行功能单元的查找。当功能单元（对应图中的节点）的数目很大时，图搜索算法会因为图中节点数过多而执行缓慢，甚至无法收敛。通常有多个功能单元的输入与目标输入相同，也会有多个功能单元的输出与目标输出相同，因此进行功能

单元链的搜索与一般的单起始点/目标点的路径规划相比，难度提高了很多。基于深度学习的搜索算法为解决这些问题提供了新的思路。文献[31]采用基于深度学习的融合向量空间模型，完成了生物图像的搜索引擎。文献[32]将增强学习与深度神经网络结合以进行启发式搜索。目前未见将深度学习方法应用于功能单元链生成和功能知识单元集成的相关研究。

参考文献

[1] 谢友柏.设计科学与设计竞争力[M].北京:科学出版社.2018

[2] Xie YB. Four basic laws in design science [J]. Engineering Science. 2014, 2(12): 2-9

[3] 谢友柏.设计科学中关于知识的研究[J]. 中国工程科学. 2013,15(4):14-22

[4] 谢友柏.设计科学的研究和设计竞争力[J]. 中国工程科学. 2014,16(8):4-13

[5] Szykman S, Sriram RD, Regli WC. The role of knowledge in next-generation product development systems[J]. Journal of Computing and Information Science in Engineering. 2001,1:3-11

[6] Erden MS, Komoto H, Beek TJV, et al. A review of function modeling: approaches and applications[J]. Artificial Intelligence for Engineering Design Analysis & Manufacturing, 2008, 22(2):147-169

[7] Pahl G, Beitz W, Feldhusen J, et al. Engineering design: A systematic approach[M]. 3rd Ed. New York: Springer-Verlag, 2007

[8] Gero JS. Design prototypes: a knowledge representation schema for design[J]. AI Magazine. 1990, 11(4):26--36.

[9] Gero JS, Kannengiesser U. The situated function-behaviour-structure framework[J]. Design studies. 2004, 25(4): 373-391.

[10] Galle P. The ontology of Gero's FBS model of designing[J]. Design Studies. 2009, 30(4): 321-339

[11] Nam Pyo Suh. 公理设计——发展与应用[M].谢友柏等译.北京：机械工业出版社.2004

[12] 张执南.产品设计中的知识流理论与方法研究[D].上海交通大学.2011

[13] SysML 精粹系统建模语言概览. https://blog.csdn.net/qq_40921298/article/details/103514267

[14] 陈华钧.知识图谱导论[M].北京:电子工业出版社.2021

[15] 石莲,孙吉贵.描述逻辑综述[J].计算机科学.2006,33(1):193-197

[16] 龚晶,刘检华,赵柏萱等.基于知识的管路布局自动评价技术[J]. 计算机集成制造系统.2014,20(10):2522-2531.

[17] 付炜.基于框架网络结构的专家知识表示方法研究[J].计算机应用. 2002, 22(1):3-6
[18] 邹慧君，王石刚. 基于多层推理机制的机械产品概念设计[J]. 计算机辅助设计与图形学学报.1997,9(6):548-553
[19] 蔡鸿明,何援军,刘胡瑶. 基于分层语义网络的设计资源库建模及实现[J]. 计算机集成制造系统.2005,11(1):73-78.
[20] 刘知远,孙茂松,林衍凯等.知识表示学习研究进展[J].计算机研究与发展. 2016, 53(2): 247-261
[21] Leng J, Jiang P. A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm[J]. Knowledge-Based Systems.2016,100:188-199.
[22] 陈继文,杨红娟,董明晓,等. 基于本体语义块相似匹配的设计知识更新[J]. 机械工程学报，2014，50(7):161-167
[23]徐荣振,高琦,邵祖光,等. 基于本体填充的设计案例获取方法研究[J]. 组合机床与自动化加工技术，2019,(7):125-129
[24] 傅柱,王日芬,丁绪辉. 面向知识重用的设计过程知识语义表示研究[J]. 数据分析与知识发现，2019，3(6):21-29.
[25] Chen L, Mao XJ, Zhang YY, et al. An efficient ROS package searching approach powered by knowledge graph. SEKE 2021
[26] Liu LM, Tan RH, Peng QJ, et al. Function combination of product reconfiguration based on scenario analysis[J]. Computer-Aided Design & Application. 2021, 18(3):532-544
[27] Li X, Zhang ZN, Liu ZL, Xie YB. A novel semi-heuristic planning approach for automated conceptual design synthesis[J]. IMech E, Part C: Journal of Mechanical Engineering Science 2013, 20(2): 100- 125
[28] 李响.面向开放式创新的产品继承设计理论与方法研究[D].上海交通大学.2014
[29] 陈斌.分布式资源环境下设计过程中的功能知识集成理论与方法研究[D].上海交通大学.2019
[30] 刘郡.分布式资源环境下产品设计中功能-结构-行为的知识集成理论与方法[D].上海交通大学.2019
[31] Mishra R, Tripathi SP. Deep learning based search engine for biomedical images using convolutional neural networks[J]. Multimedia Tools and Applications, 2021:1-9.
[32] Zhang W. A learning search algorithm with propagational reinforcement learning[J]. Applied Intelligence, 2021,51:7990-8009

3.3.2 该问题与基金资助指南范围的关联

本项目针对产品设计中功能知识的表示、存储和功能知识单元的集成进行研究，研究内容属于基金资助指南的第三项“知识在设计中流动、集成、竞争、进化的研究”和第四项“面向设计竞争力提升的设计知识高效供给和运用机制研究”。本项目的研究旨在解决功能知识的分类和表示问题，将使用深度学习和强化学习等人工智能算法处理功能知识单元的集成。研究成果有助于基于互联网的知识高效供给与运用。

4.申报条件

申请人本科和硕士就读于西安交通大学计算机系，博士师从谢友柏院士。博士论文针对分布式知识资源的利用进行了研究，对产品设计与产品竞争力的关系、产品的创新设计过程进行了探讨，并完成了异地合作设计环境的设计与实现。在新加坡南洋理工大学和新加坡国立大学从事博士后研究期间，主要从事港口堆场和自动仓储系统的优化调度算法研究。

近年来申请人主要从事分布式系统中大数据的处理和资源分配算法研究，包括分布式数据库的管理、数据特征提取、系统资源分配和作业调度等。在分布式系统的大数据处理研究中，申请人积累了一定的并行处理框架和深度学习算法领域的科研基础。这部分工作有助于本申请中功能知识集成算法的完成。为了能够提高分布式系统的资源利用率、缩短作业的执行时间，需要根据作业的资源需求为其分配合适的系统资源。合理的系统资源的描述方式是进行资源分配的前提。因此，前期工作中对分布式系统中节点资源的表示、用户作业需求和节点匹配算法的探索将会对本申请的研究起到促进作用。

教学方面，申请人十余年来一直教授“实用人机交互”课程，教学效果受到学生好评。该课程强调了以用户为中心的设计理念，着重介绍了获取准确用户需求的重要性及方法，突出了产品运行环境对功能设计的影响以及产品全生命周期设计的思想。这些理念和方法与本项目研究的产品设计完全契合。一方面，申请人可以将该课程的理论应用于本项目研究；另一方面，也可以将本项目的研究成果反哺于教学。本项目中涉及的真实产品设计案例可以提供给学生，帮助他们深入理解设计的要素和精髓，培养他们学以致用的能力和创新精神。

申请人主持参与多项科研项目，相关项目的研究成果能够为本申请的工作起到一定的支撑作用。主要的项目包括：

(1) 国家重点研发计划. “高性能计算环境服务化机制与支撑体系研究（二期）”课题三“支持应用社区的全局资源供给与使用”的子任务“任务的本地提交和全局执行使用”(2018YFB0204000). 2018.5-2021.4. 主持

(2) 863 主题项目. “云制造服务平台关键技术研究”的子课题“面向航天复杂产品的集团企业云制造服务平台开发、系统构建及应用”(SQ2011AA0403829003). 2011.1-2012.8. 参与

(3) 浙江省自然科学基金. “基于无线传感器网络的滑坡预警关键问题研究”(Y1101107). 2010.6~2013.6. 主持

(4) 浙江省自然科学基金. “隐私图像泄漏主动侦查关键问题研究”(LY20F02002). 2020.1-2022.12. 参与

(5) 浙江省自然科学基金. “基于迹零水印及云搜索技术的隐私媒体泄漏追踪模型研究”(LY15F020032).2015.1-2017.12.参与

(6) 浙江省自然科学基金. “基于声誉的 P2P 用户信任模型及信任安全研究”(LY13F020026). 2013.1-2015.12. 参与

(7) 浙江省科技厅重大专项.“基于无线传感器网络的紧急逃生引导系统的产品研制”(2007C13064). 2007.12-2010.12. 参与

(8) 国家自然科学基金重大项目.“支持产品创新的先进制造技术中的若干基础研究”的子课题“支持产品创新设计的知识获取、组织、传递及运用研究”(5999072). 1999.1-2003.7.参与

申请人发表的与本申请相关的主要论文包括：

(1) Zhang Yuchen, **Hu Yahong**, Fang Haoda. Consumer satisfaction oriented residential appliance scheduling algorithms, Systems Science & Control Engineering. 2021, 9:1,663-672.EI检索(20213810926496)

(2) **胡亚红**,潘恩宇,毛家发.基于LSTM和遗传算法的分布式系统资源优化分配算法. 中国高性能计算年会. 2021.10.20-10.23.珠海

(3) **胡亚红**,邱圆圆,毛家发.分布式异构集群中节点优先级调优算法.国防科技大学学报.已录用.EI检索

(4) **胡亚红**,吴寅超,朱正东.节点实时性能自适应的集群资源分配算法. 国防科技大学学报.已录用.EI检索

(5) 朱正东,吴寅超,**胡亚红**,蒋家强. 基于LSTM的集群用户作业执行时间预测模型.计算机工程与科学.已录用

(6) **胡亚红**,吴寅超,朱正东,李小轩.异构集群节点与作业特性感知资源分配算法.计算机工程与应用.网络优先发表

(7) Zheng Jianjing, **Hu Yahong**. Investigation of burst failure of a thick-walled spherical vessel with plastic orthotropy under internal and external pressures. Journal of Applied Mechanics and Technical Physics. 2020,61(6):1033–1041. SCI检索 (000607410800017)

(8) **胡亚红**,王一洲,毛家发.基于PageRank和基准测试的异构集群节点性能评价算法研究. 计算机工程与科学. 2020,42(3):11-16
(9) **胡亚红**,盛夏,毛家发.资源不均衡Spark 环境任务调度优化算法研究.计算机工程与科学. 2020, 42(2):14-20
(10) **Hu Yahong**, Sheng Xia, Mao Jiafa, et al. Spark based real-time proactive image tracking protection model. EURASIP Journal on Information Security. 2019: 3. EI检索(20191806857605)
(11) Mao Jiafa, Wang Weifeng , **Hu Yahong**, Sheng Weiguo. A scene recognition algorithm based on deep residual network. Systems Science and Control Engineering. 2019, 7(1): 243-251. EI检索(20193207273567)
(12) **Hu Yahong**, Zhu Zhengdong, Hsu Wen-jing. Load shuffling algorithms for split-platform AS/RS. Robotics and Computer Integrated Manufacturing. 2010,26(6): 677-685.SCI检索(685YF)
(13) **Hu Yahong**, Zhu Zhengdong, Hsu Wen-jing. AS/RS based yard and yard planning. Journal of Zhejiang University Science A. 2008, 9(8), 1083-1089. SCI检索(332WF)
(14) 朱正东,缪相林,**胡亚红**等. 可用性的语义Web服务通用发现机制. 西安交通大学学报. 2008.42(6): 659-663. EI检索(20082811370946)
(15) 朱正东，武卫国，**胡亚红**等. 一种基于最小覆盖的复杂Web 服务组合方法. 西安交通大学学报. 2008. 42(8):945-949 .EI检索(20083711541276)
(16) 朱正东, **胡亚红**, 伍卫国等. 内容寻址网络的P2P语义Web服务组合系统架构[J]. 西安交通大学学报. 2010,44(2): 6-10. EI检索(20101312808461)
(17) **Hu Yahong**, Huang Shell Ying, Chen Chuanyu. et al. Travel time analysis of a new automated storage and retrieval system. Computers & Operations Research. 2005,32:1515-1544. SCI检索(881ZG)
(18) **胡亚红**,周受钦,姜培林,谢友柏.现代设计中的远程程序调用.中国机械工程, 2001,12(9): 1035~1037. EI检索(2001556802505)
(19) 谢友柏,**胡亚红.** 互联网上进行的合作设计. 北京：机械工业出版社,2000
(20) **胡亚红**,谢友柏.为现代设计创建良好的环境. 机械科学与技术. 1998,6:923~924. EI检索(1999014547875)
(21) **胡亚红**,刘路放,谢友柏,张优云.大型旋转机械摩擦学-动力学设计数据库的设计与实现.润滑与密封.1998,5:28-31. EI检索(1998124500344)
(22) **胡亚红**,刘路放,张优云,谢友柏.摩擦学-动力学设计知识自动获取环境的研究.润滑与密封. EI检索(1998034140491).1997,6:31-34

附上论文“Spark based real-time proactive image tracking protection model”以说明申请人的研究经历。此论文的研究目标是主动发现互联网上被非法盗用的图片，保护图片主人的合法权益。研究中涉及并行处理框架、分布式数据库、数据的存储管理和图片的特征提取、图片搜索等内容。有关分布式系统、数据的存储管理和搜索等内容的研究结果将为本申请的工作提供一定的帮助。

5.依托单位意见

同意该项目申报，如获资助，我单位将为项目开展提供必要的条件。

负责人签名(章)：胡亚红　　依托单位公章

2021 年 12 月 26 日

谢友柏设计科学研究基金申报推荐书

谢友柏设计科学研究基金学术委员会：

我愿意推荐 **胡亚红** 申报名称为“分布式资源环境中产品设计功能知识管理与集成方法研究”的设计科学研究基金项目。

推荐理由：

胡亚红博士期间即开始在分布式知识资源使用领域进行研究工作，并取得了一定的研究成果。近年来主要从事分布式系统中大数据处理和基于深度学习的系统资源分配等优化算法的研究。胡亚红具有严谨的科学态度、做事踏实认真，她主持、参与了多项国家级和省部级项目，均全面完成了任务。因此，胡亚红具备了坚实的理论基础和很强的科研工作能力。针对本项目的申请，胡亚红进行了充分的前期预研工作，已取得了有意义的研究成果。

此外，胡亚红多年来一直教授“实用人机交互”课程，对产品全生命周期的设计过程及以用户为中心的设计理念有深刻的理解。这些都将为本项目的研究工作提供一定的设计理论指导。

综上所述，胡亚红具备承担此项目的能力和对设计科学研究的热情，我强烈推荐胡亚红申报谢友柏设计科学研究基金。如获资助，可保证其具有充足的项目研究时间和研究条件，并就项目执行过程中遇到的问题给予帮助。

推荐人：沈国江（签字）

2021 年 12 月 27 日

推荐人所在单位：浙江工业大学

职称/职务：教授/科研院长

联系方式：0571-85290397

附：体现申请人研究经历的论文

Hu et al. EURASIP Journal on Information Security (2019) 2019:3
https://doi.org/10.1186/s13635-019-0086-2

EURASIP Journal on Information Security

RESEARCH Open Access

Spark-based real-time proactive image tracking protection model

Yahong Hu*, Xia Sheng, Jiafa Mao, Kaihui Wang and Danhong Zhong

Abstract

With rapid development of the Internet, images are spreading more and more quickly and widely. The phenomenon of image illegal usage emerges frequently, and this has marked impacts on people's normal life. Therefore, it is of great importance to protect image security and image owner's rights. At present, most image protection is passive. Most of the time, only when the images had been used illegally and serious adverse consequences had appeared did the image owners discover it. In this paper, a Spark-based real-time proactive image tracking protection model (SRPITP) is proposed to monitor the status of images under protection in real time. Whenever illegal use is found, an alert will be issued to image owners. The model mainly includes image fingerprint extraction module, image crawling module, and image matching module. The experimental results show that in SRPITP, the image matching accuracy rate is above 98.9%, and compared with its stand-alone counterpart, the corresponding time reduction for image extraction and matching are about 58.78% and 61.67%.

Keywords: Fingerprint, Image protection, Spark, Database

1 Introduction

No one would like others to use his belongings unauthorized, especially photos or other images. For example, Mary is a pretty lady and she often shares her photos with friends in Flickr. Accidentally, she found her photo was used as advertisement in an online store. Mary was very angry about the illegal use of her photo, and she wanted to know when this unauthorized use began. In another scenario, Mark is a diligent photographer and he had taken many marvelous photos. One day, he found one of his new photographs was posted in a famous website and the owner of it was someone he never knew. In these circumstances, the rights of Mary and Mark are hurt, and if they can find out the unauthorized use of their photos as early as possible, their loss can be minimized.

Nowadays, information disseminates more rapidly and widely, which makes the security and privacy protection of information really important [1]. The protection of image resources is particularly urgent. Images which need to be protected include pictures, photos, rubbings, and so on. Once images are illegally used in inappropriate situations, the image owners may suffer from great trouble or financial loss [2]. At present, only when illegal use and severe hurt have occurred will the image owners know the fact. Therefore, it is of great necessity to have research on real-time proactive image protection to defend the rights of image owners. There are a vast number of images existing in websites, and many new images appear each day. In this big data environment, traditional stand-alone image processing method can hardly guarantee the image safety in real time [3, 4].

There have been a lot of research results in image processing, while to the best of our knowledge, there is no research on real-time and proactive image tracking protection model till now. In this paper, a Spark-based real-time proactive image tracking protection model (SRPITP) is proposed to find out the illegal use of images and protect the image owners' legitimate rights. This model is deployed in the parallel computing frame Spark to improve the system's real-time performance. The contribution of this paper is to construct a proactive and real-time image tracking protection frame. The frame integrates image crawling, image fingerprint extraction, and image matching.

The remainder of the paper is structured as follows. Section 2 introduces the related research work. Section 3

* Correspondence: huyahong@zjut.edu.cn
School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang, China

describes the structure of SRPITP and explains the related algorithms. Experimental results are shown in Section 4, and Section 5 concludes the paper.

2 Related work

Image protection, authentication, and retrieval are the most relevant topics related with our research. Image protection aims to provide images with high security to keep their confidentiality and integrity. Watermarking and encryption are commonly used methods of image protection. Lots of research has been done in this area [5–8]. Aryal et al. [9] propose a scheme suitable for the hierarchical access control system, where images can be accessed with different access rights. Pareek and Patidar [10] introduce an encryption method for gray-scale medical images based on the features of genetic algorithms. Hu et al. [11] adopt impulsive neural network synchronization technique to intelligent image protection against illegal swiping and abuse. Using the algorithm described in [12], only images with visible watermark or having been cut are allowed to be downloaded. Bhargava et al. [13] propose a method to add invisible watermark to an image, with the user information hidden. It can be seen that existing work of image protection focus on showing image ownership or trying to ensure that people handle the images according to their privileges; however, they cannot find the illegal use of images.

The purpose of image authentication is to solve two kinds of problems, i.e., judging the authenticity of image imaging process and whether the image is tampered after generation [14]. Commonly used technologies for image authentication include digital signature, digital watermarking [15], and perceptual hash. Mao et al. proposed a fingerprint of scene frames based video authentication method which has high accuracy and low storage requirement [16].

Compared with the text-based image retrieval, content-based image retrieval (CBIR) can extract the visual features of images automatically, and the retrieval results are more accurate. The basic steps in CBIR include image feature description, feature extraction, feature compression, and index establishment. Experiments show that SIFT provides the best feature description, but its computing complexity is too high to meet the real-time demand [17, 18]. Haar wavelet decomposition feature has good performance to cope with globally similar images [19]. Support vector machine, association rule [20], and convolutional neural network can also provide very good image retrieval results [21].

The increased amount of images put great pressure to traditional image processing pattern. One way to handle this problem is to upgrade the existing computers; however, this method cannot solve the problem thoroughly and it is also costly. The commonly used method is to apply parallel computing system [22]. Hadoop is one of the standards for big data processing, and it has been adopted by many large enterprises to increase their data processing efficiency. As Hadoop only supports batch processing, it is not suitable to all cases of parallel processing. Therefore, other real-time processing frameworks, such as Storm and Spark, come into being. Spark is a unified analytics engine for large-scale data processing, and it can achieve high performance for both batch and streaming data processing. At the same time, Spark is user-friendly and it can simplify users' programming to a great extent [23].

There are already some systems for massive image processing. An image search engine which copes with matching huge number of high-dimensional features is proposed in [19], and it uses DistFS as the distributed file system. A distributed image retrieval system called DIRS is introduced in [24]. DIRS is also content-based, and the retrieval among massive image data storage is speeded up by utilizing Hadoop. Hadoop framework is presented in [25] with the intention of integrating an image analysis algorithm into the text-based image search engines without degrading the response time. In [26], Hadoop is used to improve the image matching performance. To deal with massive scene images retrieval, [27] puts forward an improved K-means feature clustering-based system and Hadoop is chosen as the parallel computing framework.

The comparison of SRPITP with the above-mentioned systems is listed in Table 1.

3 Spark-based real-time proactive image tracking protection model

The purpose of SRPITP is to protect the ownership and privacy of image owners, and it will send alarm to the image owners as long as the illegal use of their images is found. The model framework is shown in Fig. 1. SRPITP mainly includes the following modules, i.e., image fingerprint extraction, detected image crawling, fingerprint storage, and image matching module.

The detailed image protection process is as follows:

(1) If a user thinks it is necessary to protect his/her images, he/she submits the protection application to SRPITP. The images submitted are checked to see whether they have been in the protected image database already. If not, the application is accepted.
(2) The accepted images are fingerprinted and classified by the fingerprint extraction servers.
(3) The fingerprints of these images and relevant information of the owner are inserted into the protected image database.

Table 1 Comparison of SRPITP with related systems

Literature	System objectives	Image features	Image classification	Matching algorithm	Parallel processing system
[19]	Image retrieval	Harr SIFT	N	Locality sensitive Hashing	DistFS
DIRS [24]	Image retrieval	Content-based visual features	N	Euclidean distance	Hadoop
Online CBIR [25]	Image retrieval	Color and low-level features	N	Distances between the AC's and ACC's	Hadoop
Massive image retrieval [26]	Image retrieval	Color, texture	Y	Euclidean distance	Hadoop
Massive scene image retrieval [27]	Image retrieval	SURF	N	K-means cluster	Hadoop
SRPITP	Finding unauthorized image usage	Element A,D	Y	Hamming distance	Spark

(4) Images in the websites under monitoring are crawled. SRPITP system administrator has the privilege to determine which websites should be monitored.
(5) Image crawling servers receive the images from the websites.
(6) Images obtained from step (5) are tagged and fingerprinted by the fingerprint extraction servers.
(7) The fingerprints and related information of the images obtained from step (6) are saved in the detected image database. The related information includes the names of the uploaders and time of the upload.
(8) Applying the image matching algorithm, the fingerprint matching servers determine whether there exists unauthorized image usage.
(9) If unauthorized image use is found, the fingerprint matching server sends a message to a management server.
(10)The management server sends an alarm to the image owner immediately, and detailed information of the illegal use (such as when and where this usage is found, the uploader of the image) is also sent to the owner. Then, the image owner may take appropriate measures to protect his/her rights.

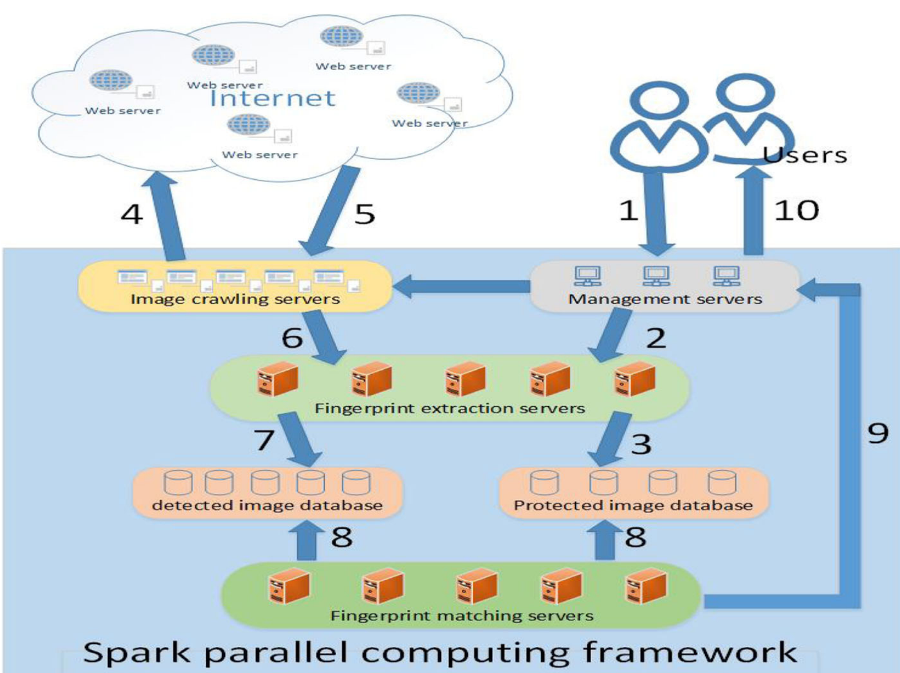


Fig. 1 Structure of the image protection model. This figure gives the components included in the image protection model and also describes the procedure of image protection

With the increased amount of images in websites, using traditional stand-alone computer to accomplish image fingerprint extraction and matching cannot guarantee the real-time performance of our proposed model. Therefore, Spark is applied to enhance the model's throughput. One of the management servers works as the master which monitors the status of the whole system and also responsible for job scheduling. All the other computers act as worker nodes in our system, and they work in parallel to ensure the real-time property of SRPITP.

The details of the main modules are described in the following sub-sections.

3.1 Fingerprint extraction module

Fingerprint extraction algorithm is one of the core algorithms in SRPITP. Compared with other image fingerprint extraction algorithms, the algorithm proposed by Mao et al. [16] has lower calculation complexity and higher accuracy, so it is adopted in our model. For the readers' convenience, the algorithm is described below. The division of an image is shown in Fig. 2.

Algorithm 1: Image fingerprint extraction algorithm[16]

Input: an image

Output: fingerprint of the image

(1) The original color image is transformed to gray image.

(2) The obtained gray image is scaled to the size of 108 × 132.

(3) The image is processed using Gaussian low-pass filtering of 3 × 3 size, with the standard deviation of 0.95.

(4) Image obtained from step (3) is divided into 144 blocks.

(5) Each block is divided into 8 small pieces as shown in Figure 2. Define element A as the average pixel value of pixels in ① to ⑧, then there are altogether 144 such elements. The definition of differential element D is the difference of average pixel value of① and ②, ③ and ④, ⑤ and ⑥ as well as ⑦ and ⑧, then altogether there are 576 D elements. The fingerprint information of an image consists of both element A and D, i.e., the fingerprint of an image is described using 720 elements.

To minimize the storage space usage, quaternion quantization method is applied to represent the image fingerprint data. After the quantization, the fingerprint of each image is 180 words and only 1440 bits are required to store each fingerprint. Section 3.3 gives a detailed description of the image database.

3.2 Image crawling module

In SRPITP, Scrapy is chosen to finish real-time image acquisition from websites determined by the system administrator. Scrapy is a fast, high-level screen capture and web crawler framework which can crawl websites and extract structured data from pages [28].

Scrapy is deployed in the image crawling servers, and these servers execute the crawling at regular intervals. The value of the interval is determined by the system administrator according to the Internet image increase speed. When acquiring images from websites, the image crawling servers adopts the so called incremental crawling policy, i.e., the servers only crawl websites' new added images since last crawl. This policy helps to reduce the number of detected images, so as to reduce the number of fingerprint extraction and image matching, and improve the image protection efficiency greatly.

After being downloaded, the fingerprints of these images will be extracted. Later, the fingerprints and other related information of the images will be inserted into the detected image database for matching. The storage details are shown in Section 3.3.

3.3 Database establishment and Tag classification

As shown in Fig. 1, there are two types of database in SRPITP, i.e., the protected image database and the detected image database. Protected image database contains information of images submitted by the image owners. In order to increase the efficiency of image matching, this database has several data tables holding different types of images, e.g., human figures photography and scenic or animal photos.

After a user submits an image for protection, firstly, the image is classified and a tag is given according to its type, and then, it is fingerprinted. After that, the fingerprint and other necessary information of the image are stored into the corresponding data table in accordance with its type tag. The main fields of each protected image data table include image id, user id, storage address, fingerprint, protection duration, and so on.

Similarly, images obtained from the websites are analyzed, and their type tags are obtained. Their fingerprints together with other related information are stored in the detected image database. The structure of the detected image database is quite similar to that of the protected image database, except that there is only one data table in this database and it has a field to keep the tag values of images.

During matching, there is no need to traverse the whole protected image database to get the result, and only the data table having the same tag with the detected image's tag value needed to be searched. It can shorten the image matching time greatly, and the experimental results are shown in Section 4.

3.4 Image matching module

Efficient and accurate image matching is the basis of SRPITP. Since SRPITP has to handle massive images in real-time, Tag classification method is chosen to help to accomplish image matching faster.

Algorithm 2: Image matching algorithm

Input: detected image database and protected image database

Output: if there exists image illegal usage, returns the information of the images and sends alarm to the image owners

(1) The first record is taken from the detected image database, and its image type and fingerprint *DeFw* are obtained. *DeFw* is transformed back to its corresponding quaternion value *DeF*.

(2) According to the image type, the corresponding protected data table is chosen. (Assume the data table is A, and the number of records in A is *k*).

(3) Let *m*=1.

(4) The word value of the *m*th data in A is taken out and is restored to its corresponding quaternion value *PrF*. The normalized Hamming distance between *DeF* and *PrF* is calculated according to Formula (1).

$$d = \frac{\sum_{i=1}^{L} d_i}{L}, \quad d_i = \begin{cases} 1, DeF_i \neq PrF_i \\ 0, DeF_i = PrF_i \end{cases} \qquad (1)$$

Where i = 1,2, ..., L, L is the fingerprint length.

$$T = \begin{cases} 0, & d = 0 \\ 1, & d \leq Th \\ 2, & d > Th \end{cases} \qquad (2)$$

Where *Th* is the threshold and how to obtain its value is described in Section 4.1.

According to the value of *d* and Formula (2), the indicator *T* can be calculated.

a) When *T* = 0, it means that this image matches with an image in the protected image database and the correspond image record is noted. Go to (8).

b) When *T*= 1, it means that this image is similar to an image in the protected image database to a certain degree. Keep the position and Hamming distance of the corresponding image in the protection data table.

c) When *T*=2, it means that this image is different from all images in the protected image database.

(5) *m*=*m*+1

(6) If *m*<=*k*, go to (4), else go to (7).

(7) a) If there is at least one record satisfying *T*=1, take the record in the protected image database with the smallest Hamming distance, and perform manual review. If it passes the manual check, this record is noted.

b) If there is no record satisfying T=1, the corresponding record is deleted from detected image database.

(8) If there are still images in the detected image database, go to (1), else send the information containing in the noted records to the management server, and the algorithm finishes.

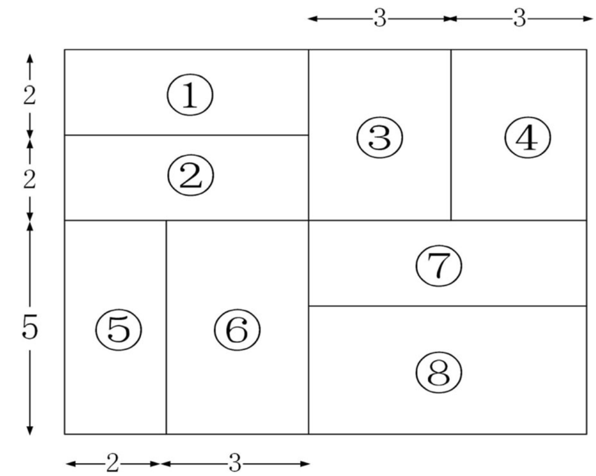


Fig. 2 Division of an image for fingerprint extraction. This figure describes the method to divide an image for the image fingerprint calculation

During the image matching process, image records in the detected database are handled in parallel to improve the system efficiency.

4 Experimental results

4.1 Training image fingerprint matching threshold

The value of threshold *Th* in Formula (2) is very important for the accuracy of the image matching algorithm. If it is too large, the tolerance of SRPITP will increase and miscarriage of justice may occur. On the contrary, if it is too small, leak judgment may happen. To obtain the

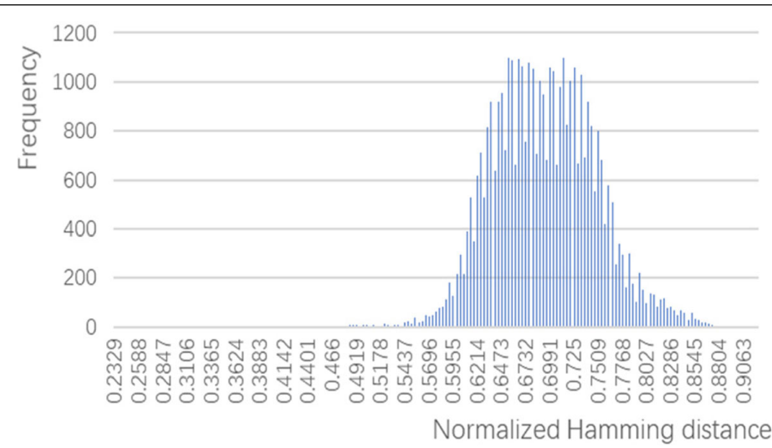


Fig. 3 Normalized Hamming distance distribution between different images. This figure provides the data to determine the value of threshold *Th*. Two hundred different images are chosen, and the normalized Hamming distance between each pair of these images are calculated. This figure shows the Histogram of the calculation results

Table 2 Image matching accuracy rate

Number of protected images	Number of detected images	Accuracy rate
2000	2000	98.90%
2000	4000	99.22%

Table 3 Experiment setup

Number of nodes	A master node and two slave nodes
Node performance	2.5 GHz, 4 cores, 2G memory
Node operating system	Ubuntu16.04
Related software	Spark 2.7, openjdk8, Hadoop 2.7.4

optimal Th, 200 different images were selected to calculate the normalized Hamming distance between each pair of them. Histogram of the calculation results is shown in Fig. 3. The results approximately fit to the Gaussian distribution $N(\mu, \delta^2)$, with the mathematical expectation 0.6915 and the variance 0.0037. An image either belongs to the protected image database or not, then the image matching problem is a kind of two alternative hypothesis testing problem. Define the missing alarm event as not detecting the illegal used image, and the false alarm event as a normal image being treated as illegal usage, it has:

$$P_F = \int_{-\infty}^{Th} \frac{1}{\sqrt{2\pi}\delta} e^{\frac{-(x-\mu)^2}{2\delta^2}} dx \quad (3)$$

where P_F is the false alarm probability, and usually, it is controlled under 1 ppm (parts per million), then the value of Th is calculated as 0.4258.

4.2 Experiment results analysis

Altogether four different types of experiments were conducted, and they are described in detail as follows.

4.2.1 Matching accuracy

This experiment is to test the accuracy of the image matching algorithm in SRPITP. The total amount of data in the protected image database is 2000, and the data volume in the detected image database is 2000 and 4000, respectively. The image matching results are compared with manually checked results. Each experiment was carried out five times, and the average accuracy rate was calculated, as shown in Table 2.

As can be seen from Table 2, the accuracy rate of the image matching is as high as 98.9% and 99.22%, which is quite satisfactory.

4.2.2 Validity of the Tag classification method

As having been shown, Tag classification method is used to distinguish different types of images, and this section demonstrates the feasibility of the method from the experimental point of view. Two sets of databases are used. In one set, the protected image database has the image classification information, and also images under detection have Tag value. For the other set, the protected image database is not classified, and images to be detected have no Tag labels. The total amount of protected images for each experiment is 2000, and the highest

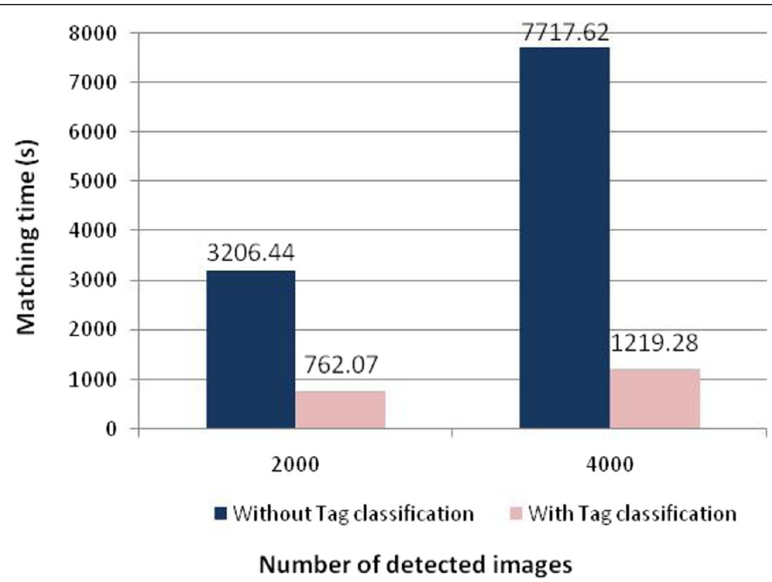


Fig. 4 Image matching time comparison with or without tag classification. This figure provides the execution time comparison of image matching algorithm when using the tag classification method and without using it. The result shows that by using the tag, the image matching time can be reduced a lot

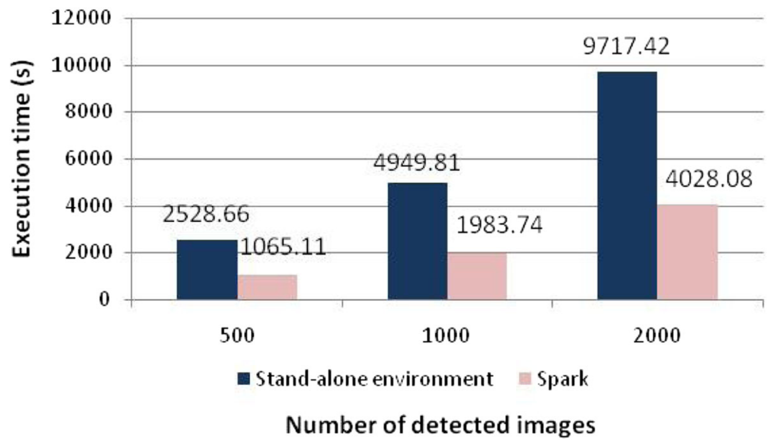


Fig. 5 Comparison of fingerprint extraction time under different operating environment. This figure provides the image fingerprint extraction time using a stand-alone computer and using the Spark-based parallel environment. The result shows that the Spark-based environment proposed by this paper can greatly reduce the extraction time

value of the detected image database reaches 4000. Each experiment was conducted five times, and the average running time was taken. The experimental result is shown in Fig. 4.

It can be seen from Fig. 4 that the time spent for image matching using Tag classification is only 23.77% and 15.80% of the time without Tag classification when the number of detected image data is 2000 and 4000, respectively. The advantage of using image classification is even more obvious when the amount of detected images increases. Thus, Tag classification method has high applicability and feasibility.

4.2.3 Comparison of image fingerprint extraction efficiency

This subsection compares the efficiency of image fingerprint extraction algorithms under stand-alone computer environment and Spark-based environment. The experiment setup is shown in Table 3.

Three sets of experiments were performed with the number of images as 500, 1000, and 2000, respectively. Each experiment was carried out five times, and the average execution time was calculated, as shown in Fig. 5.

It can be seen that with SRPITP, the fingerprint extraction efficiency is improved by 57.88%, 59.92%, and 58.55% when the amount of extracted image data is 500, 1000, and 2000, respectively, and the average improvement is about 58.78%.

4.2.4 Comparison of images matching time

As image matching efficiency is a key indicator of SRPITP, comparative experiments were conducted to show the system performance improvement by applying

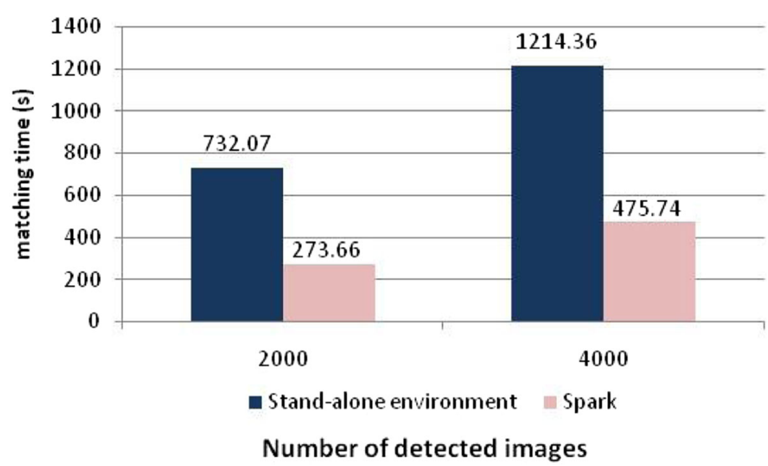


Fig. 6 Comparison of image matching time under different operating environment. This figure provides the image matching time using a stand-alone computer and using the-Spark based parallel environment. The result shows that SRPITP can reduce the matching time to a great extent

Spark. The experiment setup is the same as shown in Table 3. In each experiment, the total data volume of the protected image database was 2000, and the highest value of the detected image database was 4000. Each experiment runs five times, and the average matching time was calculated. The experimental result is shown in Fig. 6.

It is clear that the image matching efficiency of SRPITP is much higher than using the traditional standalone environment. When the numbers of detected images are 2000 and 4000, SRPITP provides 2.67 times and 2.55 times speed compared with those of the standalone ones. In other words, the time reduction is about 62.62% and 60.82% correspondingly, and the average time reduction is 61.72%.

5 Conclusion

In order to solve the problem of image privacy and security protection, a real-time proactive image tracking protection model SRPITP is proposed. Tag classification method and parallel computing framework Spark are adopted to enhance the efficiency of SRPITP.

Future work will be carried out in two aspects. The first is to improve the recognition ability of fingerprint extraction algorithm to handle images with serious attack. The second is to optimize the Spark platform resource allocation algorithm. Using coarse-grained technology to allocate resources dynamically, default Spark only considers CPU resources, and container-level resource adjustment is ignored. More efficient resource allocation algorithms will be proposed to further improve the performance of SRPITP.

Acknowledgements
This work is supported by National Key R&D Program of China (No. 2018YFB0204003 & 2016YFC0701309).

Funding
This work is supported by National Key R&D Program of China (No. 2018YFB0204003 & 2016YFC0701309). They provide help in collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions
HYH provided the overall design of the model proposed in this manuscript. SX integrated each separate program in the Spark environment. MJA provided the image matching algorithm. WKH implemented the matching algorithm, and ZDH conducted the experiments.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 12 September 2018 Accepted: 19 March 2019
Published online: 03 April 2019

References

1. D.G. Feng, Z. Min, L. Hao, Big data security and privacy protection. Chin. J. Comput. Phys. **37**(1), 246–258 (2014)
2. J. Liu, H. Wang, Image protection scheme based on privacy protection in content sharing environment. Comp. Appl. Software **32**(7), 207–211 (2015)
3. X. Lu, X. Han, Research on big data security and privacy protection technology architecture. Inform. Secur. Res. **2**(3), 244–250 (2016)
4. Y. Liu, T. Zhang, X. Jin, et al., Personal privacy protection in big data era. J. Comput. Res. Dev. **52**(1), 229–247 (2015)
5. I.J. Cox, J. Kilian, F.T. Leighton, et al., Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Process. **6**, 1673–1687 (1997)
6. G. Xuan, J. Zhu, J. Chen, et al., Distortionless data hiding based on integer wavelet transform. IEEE Electron. Lett. **38**, 1646–1684 (2002)
7. H.T. Wu, J.L. Dugelay, Y.Q. Shi, Reversible image data hiding with contrast enhancement. IEEE Signal Process Lett. **22**, 81–85 (2015)
8. K. Kurihara, S. Imaizumi, S. Shiota, et al., An encryption-then-compression system for lossless image compression standards. IEICE Trans. Inf. Syst. **E100.D**(1), 52–56 (2017)
9. A. Aryal, S. Imaizumi, T. Horiuchi, et al., Integrated model of image protection techniques. J. Imag. **4**(1), 1–12 (2017)
10. N.K. Pareek, V. Patidar, Medical image protection using genetic algorithm operations. Soft. Comput. **20**, 763–772 (2016)
11. B. Hu, Z. Guan, N. Xiong, et al., Intelligent impulsive synchronization of nonlinear interconnected neural networks for image protection. IEEE Trans. Indus. Inform. **14**(8), 3775–3787 (2018)
12. E. Gomez, P. Cano, L. Gomes, et al., in *Proceedings International Telecommunications Symposium*. Mixed watermarking-fingerprinting approach for integrity verification of audio recording (Nata, Brazil, 2002)
13. N. Bhargava, M.M. Sharma, A.S. Garhwal, et al., in *Proceedings IEEE International Conference on Radar, Communication and Computing*. Digital image authentication system based on digital watermarking (2013), pp. 185–189
14. H.F. Xing, *Research on the Key Technology of Image Authentication* (Hefei University of Technology, Hefei, 2013)
15. C.F. Lee, J.J. Shen, Z.R. Chen, in *Proceedings of 3rd International Conference on Computer and Communication Systems*. A survey of watermarking-based authentication for digital image (2018), pp. 207–211
16. J.F. Mao, G. Xiao, W.G. Sheng, et al., A method for video authenticity based on the fingerprint of scene frame. Neurocomputing. **173**, 2022–2032 (2016)
17. M. Calonder, V. Lepetit, R. Fua, *Proceedings European Conference on Computer Vision, Keypoint Signatures for Fast Learning and Recognition* (2008), pp. 58–71
18. L. Liu, Y. Ma, X. Zhang, et al., Image matching method based on binary SIFT feature description. Comp. Appl. Softw. **33**(12), 152–155 (2016)
19. Y. Zhou, *Million-Order Similar Image Search Engine Based on Distributed Computing* (Zhejiang University, Hangzhou, 2010)
20. G.X. Tan, Z.H. Liu, Learning to rank based approach for image searching. Comp. Sci. **42**(12), 275–278 (2015)
21. L.J. Jin, *Research on Content Based Image Retrieval Technology* (Huazhong University of Science & Technology, Wuhan, 2018)
22. L. Zhang, F. Zhu, H. Zhong, Interactive data preprocessing system based on Spark. Appl. Comput. **25**(11), 84–89 (2016)
23. M. Zaharia, R.S. Xin, P. Wendell, et al., Apache Spark: a unified engine for big data processing. Commun. ACM **59**(11), 56–65 (2016)
24. J. Zhang, X. Liu, J. Luo, et al., in *International Conference on Pervasive Computing & Applications*. DIRS: Distributed image retrieval system based on MapReduce (IEEE, Piscataway, 2010)
25. W. Premchaiswadi, A. Tungkatsathan, S. Intarasema, et al., Improving performance of content-based image retrieval schemes using Hadoop MapReduce. IEEE Int. Conf. High Perform. Comp. Simul. (IEEE, Piscataway, 2013) pp. 615-620
26. Q. Wang, Y.J. Tan, J. Qin, et al., Massive image retrieval based on hadoop distributed platform. J. Nanjing Univ. Sci. Technol. (Natural Science) **41**(4), 442–447 (2017)
27. H.Y. Cui, J.F. Cao, Massive scene image retrieval based on improved distributed K-Means feature clustering. J. Comput. Appl. Softw. **33**(6), 195–200 (2016)
28. Scrapy 1.6 documentation. https://docs.scrapy.org/en/latest/. Accessed 04 Feb 2019