(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110750995 B
(45)授权公告日 2023.06.02

(21)申请号 201911037419.3

(22)申请日 2019.10.29

(65)同一申请的已公布的文献号
申请公布号 CN 110750995 A

(43)申请公布日 2020.02.04

(73)专利权人 上海德拓信息技术股份有限公司
地址 200233 上海市徐汇区虹漕路448号1幢9楼

(72)发明人 袁赛杰 谢赟 韩欣 杨锐

(74)专利代理机构 上海湾谷知识产权代理事务所(普通合伙) 31289
专利代理师 张恒

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 16/36(2019.01)
G06F 18/214(2023.01)
G06N 3/04(2023.01)

(56)对比文件
CN 106874378 A,2017.06.20
CN 110148043 A,2019.08.20
WO 2016145480 A1,2016.09.22
US 2018366013 A1,2018.12.20
王仁武;袁毅;袁旭萍.基于深度学习与图数据库构建中文商业知识图谱的探索研究.图书与情报.2016,(01),全文.
朱木易洁;鲍秉坤;徐常胜.知识图谱发展与构建的研究进展.南京信息工程大学学报(自然科学版).2017,(06),全文.

审查员 张玲

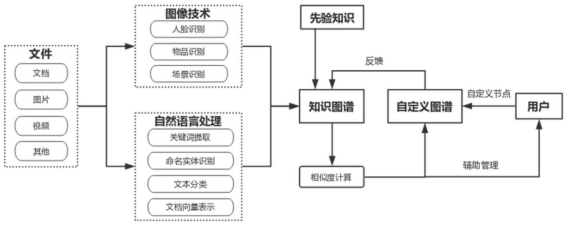
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种基于自定义图谱的文件管理方法

(57)摘要

本发明公开了一种基于自定义图谱的文件管理方法,包括:步骤S1,向图数据库中导入先验知识;步骤S2,针对现有所有文件进行处理,将处理结果结合先验知识形成包含各个文件属性和拓展属性的知识图谱,并保存在图数据库中;步骤S3,用户输入或选定各个节点词,并反馈给知识图谱,根据节点词计算相似度,取符合各节点词的相似度的文件与对应节点词建立连接,构建自定义图谱;步骤S4,基于自定义图谱和图数据库,用户进行管理和搜索文件。本发明结合用户的一定反馈信息,将文件管理分级组织,辅助用户检索出更加符合查询需求的内容。



CN 110750995 B

1.一种基于自定义图谱的文件管理方法,其特征在于,包括:

步骤S1,向图数据库中导入先验知识;

步骤S2,针对现有所有文件进行处理,将处理结果结合先验知识形成包含各个文件属性和拓展属性的知识图谱,并保存在图数据库中;

步骤S3,用户输入或选定各个节点词,并反馈给知识图谱,根据节点词计算相似度,取符合各节点词的相似度的文件与对应节点词建立连接,构建自定义图谱;

步骤S4,基于自定义图谱和图数据库,用户进行管理和搜索文件;

所述的先验知识指的是从网络上爬取的中国县级以上行政区划的上下级关系表;

所述的步骤S2包括:

针对文档型文件,通过命名实体识别、关键词提取、文本分类和文档向量表示进行处理;

针对图片型文件,通过关键词提取、人脸识别、物品检测和场景识别进行处理;

针对视频型文件,通过关键词提取和截取部分帧进行人脸识别,进行处理;

针对去除文档型、图片型和视频型以外的其他类文件,通过关键词提取进行处理;

将处理结果与先验知识进行融合,得到包含各个文件属性和拓展属性的知识图谱,存入图数据库;

所述的命名实体识别指:采用词性分析工具获取句子中人名、地名和机构名;

所述的关键词提取包括:

对文档型文件进行的内容关键词的提取;以及

对所有类型文件分别进行的标题关键词的提取;

所述的文本分类指:采用文本卷积神经网络方法训练得到的分类器,在两份训练集上分别得到通用文本分类模型和针对教育行业的教育类文本分类模型,利用通用文本分类模型和教育类文本分类模型进行分类;

所述的文档向量表示指:将文档转化为向量表示;

所述的物品检测指:识别图片中包含的物品;

所述的场景识别指:识别图片中包含的拍摄场景;

所述的人脸识别指:识别图片中的人脸,构成人物库;

在图数据库中导入省市区名称字典,再获取句子中地名;

在图数据库中导入高等教育学校及企业的字典,再获取句子中机构名,在获取机构名时通过短语检测,把类似机构名的短语进行拼接获取;

所述步骤S3中包括:

用户输入或选定节点词,并反馈给知识图谱,查找符合该节点词的相似度的文件,为自定义图谱中对应于该节点词的节点添加实体类型Dir和实体关系类型u_define;存在如下情况:

当节点词为人名/地名/机构名,则将与之关联的文件实体与Dir实体建立u_define连接;

当节点词为文本分类中出现的类别名,则将判别属于该类别的文件与Dir实体建立u_define连接;

当节点词为任一场景名、物品名或者是用户为人物库标注的名称,则将相关图片/视频

与Dir实体建立u_define连接；

当节点词为标题关键词，首先将标题中含该标题关键词的文件与Dir实体建立u_define连接，针对文档型文件，将其中有关联的文档型文件成为一簇，与其他文档进行相似度计算，计算文档表示向量之间的欧式距离来表示文档间的相似度，若相似度大于预设值，则将文档与Dir实体建立u_define连接；针对非文档型文件的标题内容计算相似度，通过计算相同词汇的占比得到标题与标题之间的相似度，若相似度大于预设值，则将文档与Dir实体建立u_define连接；

当用户输入的节点词在知识图谱中不存在，由用户来手动进行部分文件的关联，之后再自动将其他文件与Dir实体关联的文件计算相似度，若相似度大于预设值，则建立其他连接；

重复上述步骤，将对应各节点词的节点进行连接组织得到一个网状结构的文件目录，定义为自定义图谱。

2.根据权利要求1所述的基于自定义图谱的文件管理方法，其特征在于，通用类文本分类包含体育，财经，房产，家居，教育，科技，时尚，时政，游戏，娱乐，彩票，股票，社会，星座；教育类文本分类包含仪器设备，党政，基建，外事，教学，科研，行政，财会。

3.根据权利要求1所述的基于自定义图谱的文件管理方法，其特征在于，所述的人脸识别过程中，利用开源的insightface模型得到人脸识别的检测结果，构成人脸库；

所述的物品检测过程中，利用YOLO v3目标检测模型检测图片中的特定的物体；

所述的场景识别过程中，利用深度卷积网络得到的分类模型识别图像的拍摄场景。

一种基于自定义图谱的文件管理方法

技术领域

[0001] 本发明涉及文件管理方法,尤其涉及基于自定义图谱的文件管理方法。

背景技术

[0002] 计算机的文件系统是一套实现了数据的存储、分级组织、访问和获取等操作的抽象数据类型。随着网络的迅猛发展,每个人收集的资料都会相当庞大,随着自身事业和生活的变化,文件的区分界限也越来越模糊并显的有些杂乱,与之带来的困境就是找文件变得越来越困难。对于网页的搜索有搜索引擎的支撑,但对个人的文件搜索仅凭词汇匹配查找结果是不够的。随之而来的是各大网盘产品,可以将文件放在云端,并提供群组,共享圈等服务,但也意味文件将越来越难规范分级组织。

[0003] 针对这一问题,需要设计一个半自动的能辅助查找的文件管理方法。

发明内容

[0004] 本发明的目的在于提供一种基于自定义图谱的文件管理方法,结合用户的一定反馈信息,将文件管理分级组织,辅助用户检索出更加符合查询需求的内容。

[0005] 实现上述目的的技术方案是:

[0006] 一种基于自定义图谱的文件管理方法,包括:

[0007] 步骤S1,向图数据库中导入先验知识;

[0008] 步骤S2,针对现有所有文件进行处理,将处理结果结合先验知识形成包含各个文件属性和拓展属性的知识图谱,并保存在图数据库中;

[0009] 步骤S3,用户输入或选定各个节点词,并反馈给知识图谱,根据节点词计算相似度,取符合各节点词的相似度的文件与对应节点词建立连接,构建自定义图谱;

[0010] 步骤S4,基于自定义图谱和图数据库,用户进行管理和搜索文件。

[0011] 优选的,所述的先验知识指的是从网络上爬取的中国县级以上行政区划的上下级关系表。

[0012] 优选的,所述的步骤S2包括:

[0013] 针对文档型文件,通过命名实体识别、关键词提取、文本分类和文档向量表示进行处理;

[0014] 针对图片型文件,通过关键词提取、人脸识别、物品检测和场景识别进行处理;

[0015] 针对视频型文件,通过关键词提取和截取部分帧进行人脸识别,进行处理;

[0016] 针对去除文档型、图片型和视频型以外的其他类文件,通过关键词提取进行处理;

[0017] 将处理结果与先验知识进行融合,得到包含各个文件属性和拓展属性的知识图谱,存入图数据库。

[0018] 优选的,所述的命名实体识别指:采用词性分析工具获取句子中人名、地名和机构名;

[0019] 所述的关键词提取包括:

[0020] 对文档型文件进行的内容关键词的提取;以及

[0021] 对所有类型文件分别进行的标题关键词的提取;

[0022] 所述的文本分类指:采用文本卷积神经网络方法训练得到的分类器,在两份训练集上分别得到通用文本分类模型和针对教育行业的教育类文本分类模型,利用通用文本分类模型和教育类文本分类模型进行分类;

[0023] 所述的文档向量表示指:将文档转化为向量表示;

[0024] 所述的物品检测指:识别图片中包含的物品;

[0025] 所述的场景识别指:识别图片中包含的拍摄场景;

[0026] 所述的人脸识别指:识别图片中的人脸,构成人物库。

[0027] 优选的,在图数据库中导入省市区名称字典,再获取句子中地名;

[0028] 在图数据库中导入高等教育学校及企业的字典,再获取句子中机构名,在获取机构名时通过短语检测,把类似机构名的短语进行拼接获取。

[0029] 优选的,所述步骤S3中包括:

[0030] 用户输入或选定节点词,并反馈给知识图谱,查找符合该节点词的相似度的文件,为自定义图谱中对应于该节点词的节点添加实体类型Dir和实体关系类型u_define;存在如下情况:

[0031] 当节点词为人名/地名/机构名,则将与之关联的文件实体与Dir实体建立u_define连接;

[0032] 当节点词为文本分类中出现的类别名,则将判别属于该类别的文件与Dir实体建立u_define连接;

[0033] 当节点词为任一场景名、物品名或者是用户为人物库标注的名称,则将相关图片/视频与Dir实体建立u_define连接;

[0034] 当节点词为标题关键词,首先将标题中含该标题关键词的文件与Dir实体建立u_define连接,针对文档型文件,将其中有关联的文档型文件成为一簇,与其他文档进行相似度计算,计算文档表示向量之间的欧式距离来表示文档间的相似度,若相似度大于预设值,则将文档与Dir实体建立u_define连接;针对非文档型文件的标题内容计算相似度,通过计算相同词汇的占比得到标题与标题之间的相似度,若相似度大于预设值,则将文档与Dir实体建立u_define连接;

[0035] 当节点词在知识图谱中不存在,由用户来手动进行部分文件的关联,之后再自动将其他文件与Dir实体关联的文件计算相似度,若相似度大于预设值,则建立其他连接;

[0036] 重复上述步骤,将对应各节点词的节点进行连接组织得到一个网状结构的文件目录,定义为自定义图谱。

[0037] 优选的,通用类文本分类包含体育,财经,房产,家居,教育,科技,时尚,时政,游戏,娱乐,彩票,股票,社会,星座;教育类文本分类包含仪器设备,党政,基建,外事,教学,科研,行政,财会。

[0038] 优选的,所述的人脸识别过程中,利用开源的insightface(人脸识别模型)得到人脸识别的检测结果,构成人脸库;

[0039] 所述的物品检测过程中,利用YOLO v3(目标检测模型)检测图片中的特定的物体;

[0040] 所述的场景识别过程中,利用深度卷积网络得到的分类模型识别图像的拍摄场

景。

[0041] 本发明的有益效果是:本发明基于自定义图谱,分别对文档,图片,视频采用不同的处理机制,结合各大领域的深度学习应用,为文件展示更多不同层面的表达信息,辅助用户检索出更加符合查询需求的内容。适用于个人或多人的文件管理机制。在一定程度上减少管理成本,且适应各个用户的管理方式。对文件而言,同样增加了描述维度,基于本发明的搜索将更符合用户需求,更易反馈用户所期望的结果。

附图说明

[0042] 图1是本发明的基于自定义图谱的文件管理方法的流程示意图;

[0043] 图2是本发明中对于文档型文件的处理流程示意图;

[0044] 图3是本发明对于非文档型文件的处理流程示意图;

[0045] 图4是本发明根据自定义图谱中的节点匹配文件的处理流程示意图;

[0046] 图5是本发明中相似度计算的示意图。

具体实施方式

[0047] 下面将结合附图对本发明作进一步说明。

[0048] 请参阅图1,本发明的基于自定义图谱的文件管理方法,结合图像、自然语言处理领域的多种技术手段,并结合用户根据自身需求定义的分级组织策略来得到的有助于用户高效管理文件及搜索文件的一种文件管理方法。包括下列步骤:

[0049] 步骤S1,向图数据库中导入先验知识;先验知识指的是从网络上爬取的中国县级以上行政区划的上下级关系表。引入此先验知识的目的是为了匹配更准确,比如当用户自定义节点为“江苏”,不能只看包含“江苏”的文件,需要去关注江苏以下,比如“南通”,“南京”等城市的相关文件。为后续命名实体识别出的地名做辅助判断。

[0050] 步骤S2,针对现有所有文件进行处理,将处理结果结合先验知识形成包含各个文件属性(包括名称、大小、格式、文件路径、下载量、收藏量等)和拓展属性(针对文档型文件包含识别出的人名、地名、机构名、关键词及类别等;针对图片型和视频型文件包含识别出的人像、物品及场景)的知识图谱,并保存在图数据库中。即通过各大抽取技术和分类技术得到节点信息,并将其保存在图数据库中。文件包含的类型有文档,图片,视频,其他。不同类型的文件处理流程是不同的。如图2、图3所示,具体包括:

[0051] S21,针对文档型文件,通过命名实体识别、关键词提取、文本分类和文档向量表示进行处理。如下:

[0052] 1)命名实体识别:命名实体主要采用词性分析工具获取句子中人名、地名、机构名等实体。由于实体识别的识别准确率有限,为了降低该问题的影响力,一方面引入了外部词典,对于地名而言,在获取之前向图数据库中导入了省市区名称字典;对于机构名而言,在获取之前向图数据库中导入了高等教育学校及企业的字典;另一方面做了短语检测,把较大可能为机构名的短语进行了拼接获取。

[0053] 2)关键词提取:关键词提取分两个部分,一是标题关键词,二是内容关键词,做这个划分的目的是标题关键词要比内容关键词更为重要些。文档型文件需要区分,其他都是标题中的关键词。所以在搜索结果匹配上的展现,应当把标题中匹配到的结果相对展示在

靠前位置。关键词是能够表示文档主题的词语或者短语，且关键词大多为名词，一篇文档中的关键词极有可能频繁出现，但文档中频繁出现的词也并不少；此处采用的处理方法是，先进行分词技术，对文本进行拆分，然后统计词频，去除停用词，在剩下的词中若文本出现在标题中，则作为标题关键词，若文本出现在文档的首个段落中，则作为内容关键词。

[0054] 进一步地了解，提取关键词是在命名实体识别之后将剩下的词通过词频来计算重要性，取topN，若关键词出现在标题中，则作为标题关键词，剩下的则为内容关键词；对于非文档型的文件采用分词方法，保留名词的方式得到标题关键词。

[0055] 3)文本分类：此处的文本分类都是采用文本卷积神经网络方法训练得到的分类器，在两份训练集上分别得到了通用文本分类模型和针对教育行业的文本分类模型；两份训练集分别源自于由清华大学的自然语言处理实验室开源的中文文本分类数据集THUCNews，其中包含新闻界常见的14个类别，比如娱乐，财经，星座等类别；另一份数据来源是人工收集，人工获取各大院校的教育官方网站上各个栏目的文章及报导。两个分类器的准确率分别可达98.7％和94.6％。考虑到分类器一定会给出一个概率最大的类别，但有可能最大概率的类别也是判断错误的，所以在这边加一个阈值判断的操作，若概率最大和概率次大的差值小于0.5，则不输出任何类别，从而保障准确度。通用类文本分类包含14个类别，分别是体育，财经，房产，家居，教育，科技，时尚，时政，游戏，娱乐，彩票，股票，社会，星座；教育类文本分类包含8个类别，分别是仪器设备，党政，基建，外事，教学，科研，行政，财会。

[0056] 4)向量表示：将文档转化为向量表示的目的是为了后续的相似度计算。自google在2018年10月底公布BERT在11项自然语言处理任务中的卓越表现后，各大技术应用上的准确率都有提升。此处采用bert(预训练语言模型)，输入一篇文档，可以得到一个768维向量，该向量在一定程度上就表达了该文档的主题内容。

[0057] S22，针对图片型文件，通过关键词提取、人脸识别、物品检测和场景识别进行处理。处理图片的流程相对较多，首先他同样有文件名，和其他类型文件的处理是一致的；其次是需要物品检测和场景识别的技术得到图片中包含的物品和拍摄场景判断；最后走人脸识别，判断是否存在人物，若存在，则需要去构建人物库，人物库的构建是通过得到人脸表示向量和计算相似度得到一个人物的多张图片信息，此处该技术计算过程不是重点，不做具体阐述，由此可以得到人物编号，且支持用户对人物进行标注，所以在人物库中有些人物是由名称的，有些人物没有，取决于用户有没有标注。人脸识别过程中，利用开源的insightface(人脸识别模型)得到人脸识别的检测结果，构成人脸库。物品检测过程中，利用YOLO v3(目标检测模型)检测图片中的特定的物体。场景识别过程中，利用深度卷积网络得到的分类模型识别图像的拍摄场景。

[0058] S23，针对视频型文件，通过关键词提取和截取部分帧进行人脸识别，进行处理。可以把视频看作是一帧一帧的图片，所以实则是和图片一致的处理流程，但考虑到计算量的问题，所以这边视频只是按周期的取帧，然后通过人脸识别的技术结果进行汇总得到视频中出现过的人物。

[0059] S24，针对去除文档型、图片型和视频型以外的其他类文件，通过关键词提取进行处理。因为能获取的内容只有文件名，即只需要对文件名做命名实体识别；由于文件名不会很长，所以不适合选用词频来提取关键词，此处采用分词后保留名词词性词语作为关键词。

[0060] S25,将处理结果与先验知识进行融合,得到包含各个文件属性和拓展属性的知识图谱,存入图数据库。知识图谱中包含多种实体类型和实体关系类型,见表1。

[0061]

类型	名称	含义
实体	File	文件
	Person	人
	Place1	地名的县级划分
	Place2	
	Place3	
	Organization	机构
	Tkey	标题关键词
	Ckey	内容关键词
	Comlabel	通用场景分类标签
	Edulabel	教育类分类标签
	Fvector	文档特征向量
	Pscene	图像场景识别
	Pobject	图像目标检测
	Pface	图像人脸检测
关系	p_has	文件的图像识别结果
	belongs_to	属于(区域划分)
	title	文件含标题关键词
	content	文件含内容关键词
	t_has	文件的文本分类结果
	contains	文件包含人/地/机构

[0062]

[0063] 表1

[0064] 步骤S3,用户输入或选定各个节点词,并反馈给知识图谱,根据节点词计算相似度,取符合各节点词的相似度的文件与对应节点词建立连接,构建自定义图谱。用户根据自身的管理习惯构建自定义图谱。图数据库根据节点词进行检索匹配,会出现两种情况,一是未匹配到任何实体,二是匹配到实体。如图4所示,具体包括:

[0065] S31,用户输入或选定节点词,即:用户可以通过两种方式确定自定义图谱中的节点,一种是对所有文件的一个大致了解后给出中心词汇(自行输入),另一种是根据对上述多种技术分析结果的聚合统计选择某一关键词作为节点词(提供高频关键词,地名,人名,

机构名等内容进行选择)。并反馈给知识图谱,查找符合该节点词的相似度的文件,在图数据库中,为自定义图谱中对应于该节点词的节点添加实体类型Dir和实体关系类型u_define。存在如下情况:

[0066] 当节点词为人名/地名/机构名,则将与之关联的文件实体与Dir实体建立u_define连接;

[0067] 当节点词为文本分类中出现的类别名,则将判别属于该类别的文件与Dir实体建立u_define连接;

[0068] 当节点词为任一场景名、物品名或者是用户为人物库标注的名称,则将相关图片/视频与Dir实体建立u_define连接;

[0069] 当节点词为标题关键词,首先将标题中含该标题关键词的文件与Dir实体建立u_define连接,针对文档型文件,将其中有关联的文档型文件成为一簇,与其他文档进行相似度计算,计算文档表示向量之间的欧式距离来表示文档间的相似度(如图5),若相似度大于预设值(比如0.7),则将文档与Dir实体建立u_define连接;针对非文档型文件的标题内容计算相似度,通过计算相同词汇的占比得到标题与标题之间的相似度(如图5),若相似度大于预设值,则将文档与Dir实体建立u_define连接。例如:通过计算相同词汇的占比得到标题与标题之间的相似度,比如[“机器学习”,“教程”,“手册”]和[“机器学习”,“深度学习”]之间的相似度为(1*2)/(3+2),同样定义阈值保留部分文件。

[0070] 当节点词在知识图谱中不存在(即用户输入的是新词),此处不采用模糊匹配的方式去图数据库中查找较为匹配的节点,这边是期望与用户进行交互,由用户来手动进行部分文件的关联,之后系统再根据用户添加的文件通过计算相似度来将其他相关文件与Dir实体进行关联,这么做的目的是为了提高分类的准确性,如果单单从模糊匹配的角度把文件与用户定义的节点关联起来,可以会发生较大的错误,导致后面计算相似度依然把关联关系错误连接。

[0071] S32,重复上述步骤,将对应各节点词的节点进行连接组织得到一个网状结构的文件目录,定义为自定义图谱。从而得到自身设计的图谱,理清文件的分级组织。

[0072] 步骤S4,基于自定义图谱和图数据库,用户可以多维度的管理和搜索文件,以获取最切合自身需求的查询结果。

[0073] 通过上述的操作和计算过程,用户可自行组织文件管理的分级策略,从而从自身关注视角对文件进行编排归纳,在此自定义图谱的基础上,做相关的文件展示和维度也有了多样性,不同用户可拥有不同的管理界面和搜索倾向性,从而个性化的探索用户的搜索需求,返回用户期望中的搜索结果。

[0074] 以上实施例仅供说明本发明之用,而非对本发明的限制,有关技术领域的技术人员,在不脱离本发明的精神和范围的情况下,还可以作出各种变换或变型,因此所有等同的技术方案也应该属于本发明的范畴,应由各权利要求所限定。

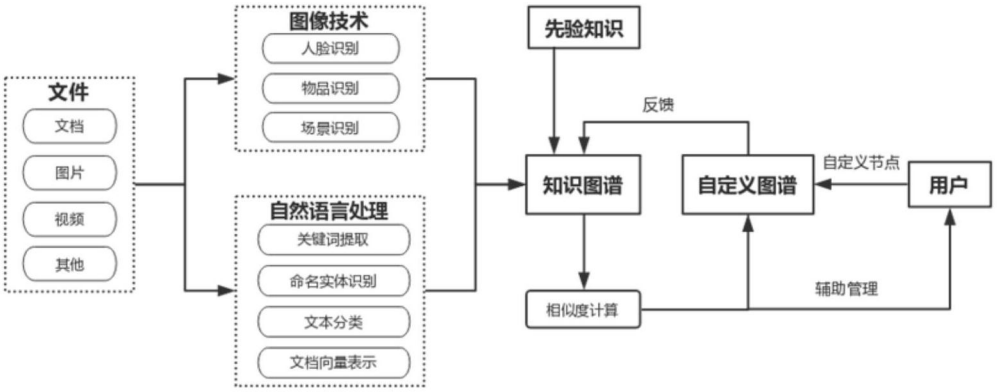

文件
文档
图片
视频
其他
图像技术
人脸识别
物品识别
场景识别
自然语言处理
关键词提取
命名实体识别
文本分类
文档向量表示
先验知识
反馈
知识图谱
自定义图谱
自定义节点
用户
相似度计算
辅助管理

图1

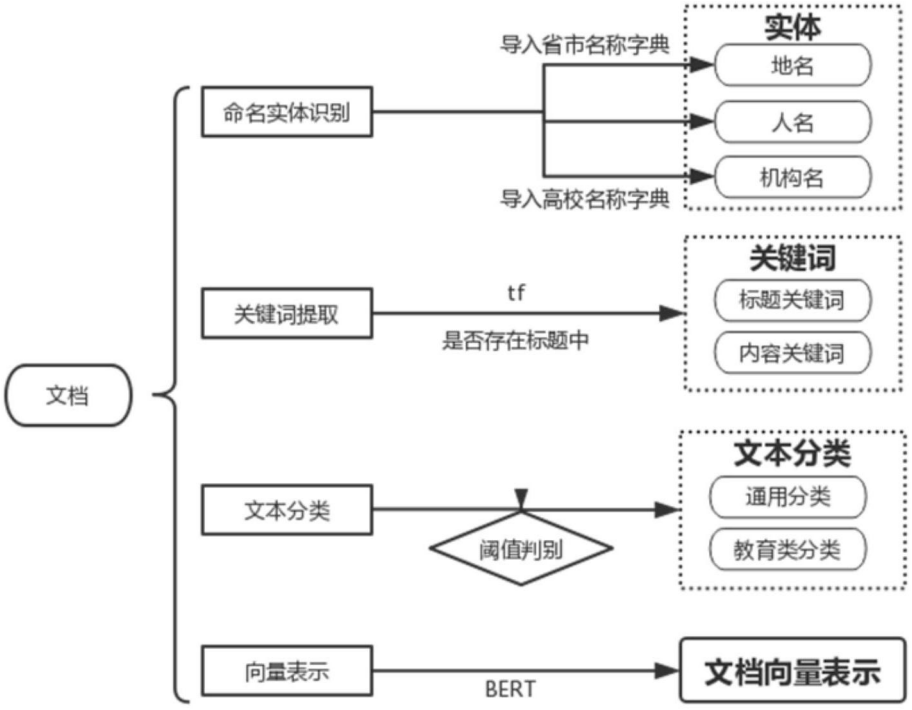

文档
命名实体识别
导入省市名称字典
导入高校名称字典
实体
地名
人名
机构名
关键词提取
tf
是否存在标题中
关键词
标题关键词
内容关键词
文本分类
阈值判别
文本分类
通用分类
教育类分类
向量表示
BERT
文档向量表示

图2

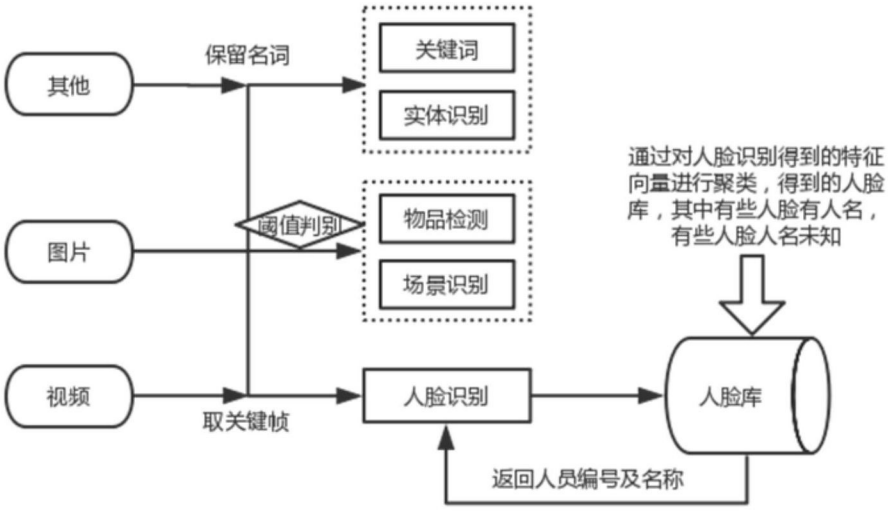

其他
保留名词
关键词
实体识别
图片
阈值判别
物品检测
场景识别
视频
取关键帧
人脸识别
人脸库
通过对人脸识别得到的特征向量进行聚类，得到的人脸库，其中有些人脸有人名，有些人脸人名未知
返回人员编号及名称

图3

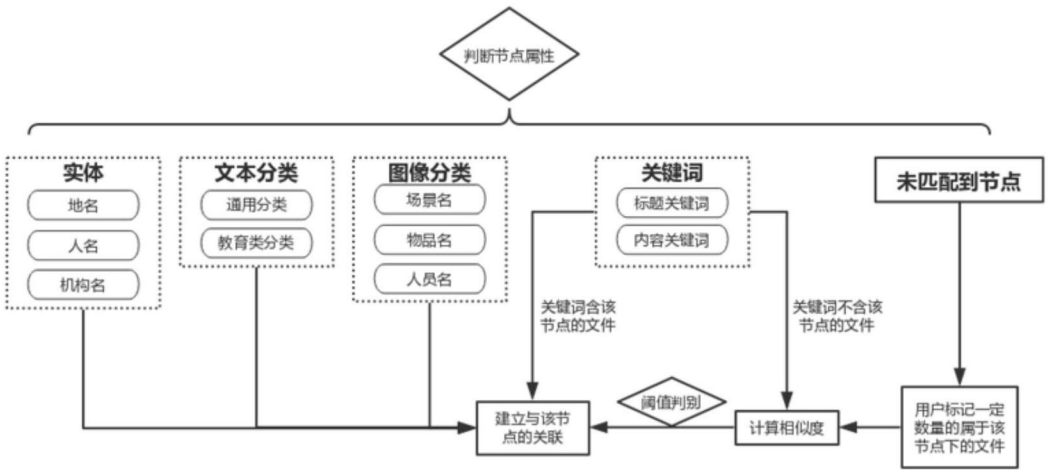

判断节点属性
实体
地名
人名
机构名
文本分类
通用分类
教育类分类
图像分类
场景名
物品名
人员名
关键词
标题关键词
内容关键词
未匹配到节点
关键词含该节点的文件
关键词不含该节点的文件
阈值判别
建立与该节点的关联
计算相似度
用户标记一定数量的属于该节点下的文件

图4

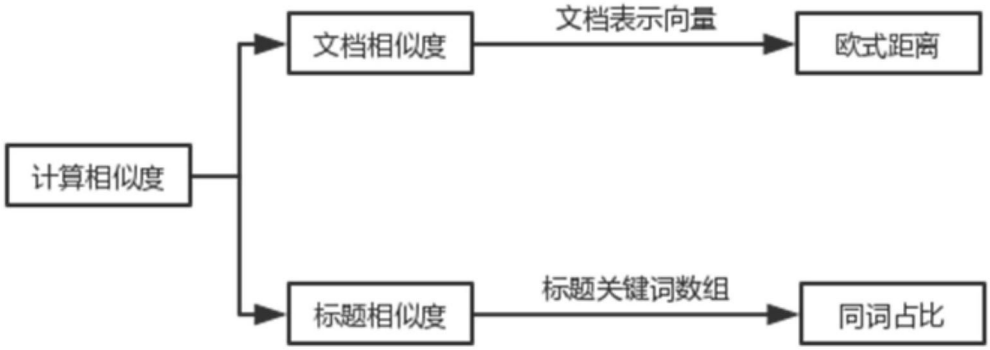

计算相似度
文档相似度
文档表示向量
欧式距离
标题相似度
标题关键词数组
同词占比

图5