(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117648504 A
(43)申请公布日 2024.03.05

(21)申请号 202210973356.8

(22)申请日 2022.08.15

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 李烜

(74)专利代理机构 华进联合专利商标代理有限公司 44224
专利代理师 董慧

(51)Int.Cl.
G06F 16/9538(2019.01)
G06F 16/903(2019.01)
G06F 18/22(2023.01)
G06F 18/23(2023.01)

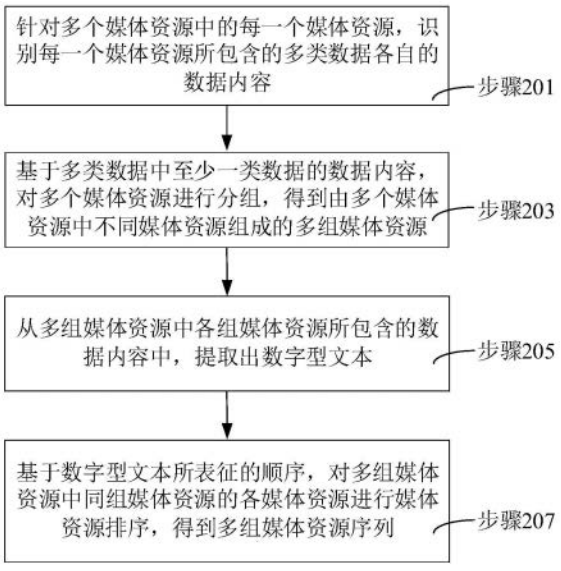
权利要求书3页 说明书17页 附图11页

(54)发明名称

媒体资源序列的生成方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及一种媒体资源序列的生成方法、装置、计算机设备、存储介质和计算机程序产品。方法包括:针对多个媒体资源中的每一个媒体资源,识别每一个媒体资源所包含的多类数据各自的数据内容;基于多类数据中至少一类数据的数据内容,对多个媒体资源进行分组,得到由多个媒体资源中不同媒体资源组成的多组媒体资源;从多组媒体资源中各组媒体资源所包含的数据内容中,提取出数字型文本;基于数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。本申请可以提高针对媒体资源的排序准确性。



CN 117648504 A

1.一种媒体资源序列的生成方法,其特征在于,所述方法包括:

针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容;

基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源;

从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本;

基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

2.根据权利要求1所述的方法,其特征在于,所述媒体资源所包含的多类数据包括内容标签数据以及内容标题数据;

所述针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容包括:

提取所述多个媒体资源中的每一个媒体资源的查询关键词;

基于所述查询关键词,查找包含所述查询关键词的媒体资源所包含的内容标签数据以及内容标题数据。

3.根据权利要求1所述的方法,其特征在于,所述媒体资源所包含的多类数据包括封面图像特征数据以及封面文本数据;

所述针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容包括:

提取所述多个媒体资源中的每一个媒体资源的封面图像;

对所述封面图像进行特征提取处理,得到封面图像特征数据,并对所述封面图像进行光学字符识别处理,得到封面文本数据。

4.根据权利要求1所述的方法,其特征在于,所述基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源包括:

获取各类数据内容对应的分组优先级;

基于所述分组优先级,依次通过各类数据内容对未分组的媒体资源进行分组,得到多组媒体资源。

5.根据权利要求1所述的方法,其特征在于,所述媒体资源所包含的多类数据包括内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据;

所述基于所述分组优先级,依次通过各类数据内容对未分组的媒体资源进行分组,得到多组媒体资源包括:

基于所述内容标签数据,对所述媒体资源进行分组,得到内容标签组媒体资源以及第一待分组媒体资源;

基于所述封面图像特征数据,对所述第一待分组媒体资源进行分组,得到封面图特征组媒体资源以及第二待分组媒体资源;

基于所述内容标题数据,对所述第二待分组媒体资源进行分组,得到内容标题组媒体资源以及第三待分组媒体资源;

基于所述封面文本数据,对所述第三待分组媒体资源进行分组,得到封面文本组媒体

资源。

6.根据权利要求5所述的方法,其特征在于,所述基于所述内容标签数据,对所述媒体资源进行分组,得到内容标签组媒体资源以及第一待分组媒体资源包括:

识别所述内容标签数据中的资源名称标签,并确定各资源名称标签中的媒体资源数;

基于媒体资源数大于或等于二的所述资源名称标签得到内容标签组媒体资源,基于媒体资源数小于二的所述资源名称标签得到第一待分组媒体资源。

7.根据权利要求5所述的方法,其特征在于,所述基于所述封面图像特征数据,对所述第一待分组媒体资源进行分组,得到封面图特征组媒体资源以及第二待分组媒体资源包括:

基于所述封面图像特征数据,得到所述第一待分组媒体资源中各媒体资源之间的封面图像欧氏距离矩阵;

基于所述封面图像欧式距离矩阵对所述第一待分组媒体资源进行聚合处理,得到封面图像特征聚簇;

确定所述封面图像特征聚簇对应的轮廓系数,基于所述轮廓系数对所述封面图像特征聚簇进行切分,得到封面图像特征类簇;

将媒体资源数大于或等于二的所述封面图像特征类簇作为封面图特征组媒体资源,将媒体资源数小于二的所述封面图像特征类簇作为第二待分组媒体资源。

8.根据权利要求5所述的方法,其特征在于,所述基于所述内容标题数据,对所述第二待分组媒体资源进行分组,得到内容标题组媒体资源以及第三待分组媒体资源包括:

确定所述第二待分组媒体资源中各媒体资源之间的标题文本相似度;

根据标题文本相似度大于或等于预设标题文本相似度阈值的媒体资源得到内容标题组媒体资源,根据文本相似度小于预设标题文本相似度阈值的媒体资源得到第三待分组媒体资源。

9.根据权利要求1所述的方法,其特征在于,所述方法从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本包括:

获取所述媒体资源所包含的所述数据内容中的文本型数据;

通过正则表达式从所述文本型数据中提取数字型文本。

10.根据权利要求1至9中任意一项所述的方法,其特征在于,所述基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列包括:

获取所述数字型文本对应的文本位置,并将所述数字型文本转化为数值型数据;

基于所述文本位置以及所述数值型数据,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

11.根据权利要求1至9中任意一项所述的方法,其特征在于,所述针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容之前,还包括:

获取资源排序请求;

确定所述资源排序请求所指定的目标内容生产者;

在预设资源数据库中查找所述目标内容生产者生产的多个媒体资源;

所述针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容包括:

针对所述内容生产者生产的多个资源数据中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容。

12.根据权利要求11中任意一项所述的方法,其特征在于,所述权利要求1-9所述的方法通过媒体资源序列生成模型实现,所述在预设资源数据库中查找所述目标内容生产者生产的多个媒体资源之后,还包括:

将所述目标内容生产者生产的多个媒体资源输入所述媒体资源序列生成模型,得到所述媒体资源序列生成模型输出的多组媒体资源序列;

获取目标对象对应的媒体资源浏览信息;

当确定所述多组媒体资源序列中包含所述媒体资源浏览信息对应的目标媒体资源时,得到所述媒体资源浏览信息中所述目标媒体资源的浏览完成度;

当基于所述浏览完成度确定满足资源推荐条件时,确定所述多组媒体资源序列中所述目标媒体资源下一序列的待推荐媒体资源;推送所述待推荐媒体资源至所述目标对象。

13.一种媒体资源序列的生成装置,其特征在于,所述装置包括:

数据内容识别模块,用于针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容;

媒体资源分组模块,用于基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源;

文本提取模块,用于从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本;

序列生成模块,用于基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

14.一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,其特征在于,所述处理器执行所述计算机程序时实现权利要求1至12中任一项所述的方法的步骤。

15.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至12中任一项所述的方法的步骤。

媒体资源序列的生成方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及计算机技术领域,特别是涉及一种媒体资源序列的生成方法、装置、计算机设备、存储介质和计算机程序产品。

背景技术

[0002] 随着计算机技术与网络技术的发展,出现了各种类型媒体网站,用户可以根据实际的需要在媒体网站浏览视频、音乐或者文章等媒体资源。而对于由内容生产者(content producer,CP)提供的一系列媒体资源,有时需要根据需要将同一序列下的媒体内容进行归纳并排序,从而有效地引导用户按照序列顺序来浏览媒体资源。

[0003] 目前,对于媒体资源序列的归纳与排序,一般依赖于内容生产者自身对媒体资源的标注内容,然而由于标注内容在创建时的标准完全取决于内容生产者的理解,因此内容形式较为复杂,内容标准不统一,容易导致媒体资源归纳排序不准确。

发明内容

[0004] 基于此,有必要针对上述技术问题,提供一种能够提高系列媒体资源排序准确率的媒体资源序列的生成方法、装置、计算机设备、计算机可读存储介质和计算机程序产品。

[0005] 第一方面,本申请提供了一种媒体资源序列的生成方法。所述方法包括:

[0006] 针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容;

[0007] 基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源;

[0008] 从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本;

[0009] 基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0010] 第二方面,本申请还提供了一种媒体资源序列的装置。所述装置包括:

[0011] 数据内容识别模块,用于针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容;

[0012] 媒体资源分组模块,用于基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源;

[0013] 文本提取模块,用于从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本;

[0014] 序列生成模块,用于基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0015] 第三方面,本申请还提供了一种计算机设备。所述计算机设备包括存储器和处理

器,所述存储器存储有计算机程序,所述处理器执行所述计算机程序时实现以下步骤:

[0016] 针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容;

[0017] 基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源;

[0018] 从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本;

[0019] 基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0020] 第四方面,本申请还提供了一种计算机可读存储介质。所述计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现以下步骤:

[0021] 针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容;

[0022] 基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源;

[0023] 从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本;

[0024] 基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0025] 第五方面,本申请还提供了一种计算机程序产品。所述计算机程序产品,包括计算机程序,该计算机程序被处理器执行时实现以下步骤:

[0026] 针对多个媒体资源中的每一个媒体资源,识别所述每一个媒体资源所包含的多类数据各自的数据内容;

[0027] 基于所述多类数据中至少一类数据的所述数据内容,对所述多个媒体资源进行分组,得到由所述多个媒体资源中不同媒体资源组成的多组媒体资源;

[0028] 从所述多组媒体资源中各组媒体资源所包含的所述数据内容中,提取出数字型文本;

[0029] 基于所述数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0030] 上述媒体资源序列的生成方法、装置、计算机设备、存储介质和计算机程序产品,通过先针对多个媒体资源中的每一个媒体资源,识别每一个媒体资源所包含的多类数据各自的数据内容,而后即可基于数据内容来对媒体资源进行分组及排序;而后基于多类数据中至少一类数据的数据内容,对多个媒体资源进行分组,得到由多个媒体资源中不同媒体资源组成的多组媒体资源;通过媒体资源内的各类数据内容,可以有效地对媒体资源进行分组,从而为后续的排序处理提供数据基础。继而从多组媒体资源中各组媒体资源所包含的数据内容中,提取出数字型文本;由于数字型文本一般可以表征内容顺序,可以在对媒体资源进行分组后,再提取出其对应的数字型文本来进行后续的分组,最终则是基于数字型文本所表征的顺序,对所述多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列,通过媒体资源的分组以及排序,可以有效地对杂乱的媒体资源

进行排序,保证媒体资源归纳排序准确性。

附图说明

[0031] 图1为一个实施例中媒体资源序列的生成方法的应用环境图;

[0032] 图2为一个实施例中媒体资源序列的生成方法的流程示意图;

[0033] 图3为一个实施例中依照份子优先级来对视频数据进行分组步骤的流程示意图;

[0034] 图4为一个实施例中聚类簇中树状结构的示意图;

[0035] 图5为一个实施例中基于轮廓系数进行切分的示意图;

[0036] 图6为一个实施例中基于轮廓系数进行切分所得到的同组媒体资源示意图;

[0037] 图7为一个实施例中基于正则表达式来抽取数字型文本的示意图;

[0038] 图8为一个实施例中视频资源序列的生成方法的流程示意图;

[0039] 图9为一个实施例中视频组的聚类识别过程与排序过程的示意图;

[0040] 图10为一个实施例中数据内容提取过程的示意图;

[0041] 图11为一个实施例中视频排序过程的示意图;

[0042] 图12为一个实施例中序列视频的示意图;

[0043] 图13为一个实施例中短视频应用发现页界面的示意图;

[0044] 图14为一个实施例中短视频发现页分发场景处理流程的示意图;

[0045] 图15为一个实施例中媒体资源序列的生成装置的结构框图;

[0046] 图16为一个实施例中计算机设备的内部结构图。

具体实施方式

[0047] 为了使本申请的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本申请进行进一步详细说明。应当理解,此处描述的具体实施例仅仅用以解释本申请,并不用于限定本申请。

[0048] 本申请涉及人工智能((Artificial Intelligence,AI)领域,是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。本申请具体涉及了人工智能中的自然语言处理(Nature Language processing,NLP)与机器学习(Machine Learning,ML)技术。

[0049] 其中,自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此,这一领域的研究将涉及自然语言,即人们日常使用的语言,所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括

文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。机器学习是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0050] 在本文中,需要理解的是,所涉及的术语:

[0051] 内容生产者(content producer,CP):是指提供内容服务的个人或单位。

[0052] 视频序列:指视频平台的内容生产者在进行创作时,由于主观流量效果考量或平台时长限制,将一个主题或一个系列的内容,分为多集或多部来生产。这些同一个作者生产的同系列内容组成的内容单元称为视频系列,按一定的先后顺序,共同组成一个视频序列。

[0053] Vid:单个视频的唯一id,用来指代独一无二的一个视频

[0054] 聚类:将样本的集合分成由类似的样本组成的多个簇的过程被称为聚类。簇中的对象要求尽量与同一个簇中的对象相似,与非同一个簇中的对象相异。

[0055] 光学字符识别(Optical Character Recognition,OCR):是一种识别图片等内容中的文字的技术。

[0056] IP内容:指和某些版权或者有独立主题制作有关的剧集剪辑内容。

[0057] 压字:封面图上面的字幕或字符型文本,可被OCR识别。

[0058] 本申请实施例提供的媒体资源序列的生成方法,可以应用于如图1所示的应用环境中。其中,终端102通过网络与服务器104进行通信。数据存储系统可以存储服务器104需要处理的数据。数据存储系统可以集成在服务器104上,也可以放在云上或其他服务器上。当用户希望对多媒体平台上某些多媒体内容进行分组并排序时,可以通过本申请的媒体资源序列的生成方法来实现分组以及序列生成的处理,首先用户可以通过终端102向服务器104发送序列生成请求,并指定需要处理的媒体资源。服务器104接收序列生成请求,而后查找到相对应的媒体资源后,针对多个媒体资源中的每一个媒体资源,识别每一个媒体资源所包含的多类数据各自的数据内容;基于多类数据中至少一类数据的数据内容,对多个媒体资源进行分组,得到由多个媒体资源中不同媒体资源组成的多组媒体资源;从多组媒体资源中各组媒体资源所包含的数据内容中,提取出数字型文本;基于数字型文本所表征的顺序,对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。其中,终端102可以但不限于是各种台式计算机、笔记本电脑、智能手机、平板电脑、物联网设备和便携式可穿戴设备,物联网设备可为智能音箱、智能电视、智能空调、智能车载设备等。便携式可穿戴设备可为智能手表、智能手环、头戴设备等。服务器104可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

[0059] 在一个实施例中,如图2所示,提供了一种媒体资源序列的生成方法,以该方法应用于图1中的服务器104为例进行说明,包括以下步骤:

[0060] 步骤201,针对多个媒体资源中的每一个媒体资源,识别每一个媒体资源所包含的多类数据各自的数据内容。

[0061] 其中,媒体资源是指由内容提供者发布在内容发布平台上的内容数据,如视频、音乐或文章等数据。多个媒体数据可以是同一种数据也可以是不同种的数据,如媒体资源都

为视频,或者媒体资源中一部分视频一部分文章等。媒体资源可以由同一个内容提供者发布。媒体资源所包含的多类数据则是指媒体中所包含不同类内容数据,比如对于一个视频,其包含有封面图、视频标题以及视频内容等数据,而基于封面图又可以提取文本或者图像特征等。针对媒体资源,将这些数据提取出来,即可得到数据内容。

[0062] 具体地,本申请具体适用于针对指定的媒体资源来分组,并对分组后的媒体资源进行排序。当终端102方的当用户需要对某些媒体资源进行分组排序时,可以通过发送相应请求至服务器104,来请求服务器104对请求中指定的多个媒体资源进行分组以及排序。服务器104在接收到请求后,先在资源数据库中查找到请求指定的多个媒体资源,而后针对每一个媒体资源,都查找其相关的多类数据,得到每个数据所对应的不同类数据内容。数据内容的识别查找方法具体可以为直接查找相关数据库,或者直接从媒体资源提取得到。在其中一个实施例中,本申请用于对视频数据进行处理,此时可以通过分布式数据库查询各个视频资源对应的内容标题或者内容标签等数据,或者直接从媒体资源中提取,同时对于存在封面图的视频资源,可以直接提取封面图的图像特征向量,或者直接通过光学字符识别从封面图识别出文本内容作为媒体资源的数据内容。

[0063] 步骤203,基于多类数据中至少一类数据的数据内容,对多个媒体资源进行分组,得到由多个媒体资源中不同媒体资源组成的多组媒体资源。

[0064] 其中,多类数据中至少一类数据的数据内容,对媒体资源进行分组具体是指可以预先制定不同类数据内容的优先级,而后基于优先级来依次通过不同类的数据内容对媒体资源分组,先基于第一优先级的数据内容来对媒体资源中的资源数据分组,而后若是存在部分无法分组的资源数据。则基于第二优先级的数据内容来对这些无法分组的资源数据分组,直到所有数据内容用完,或者媒体资源全都分组完成。当分组完成后,得到即为不同类型的多组媒体资源。

[0065] 具体地,当服务器104识别出媒体资源所包含的多类数据各自的数据内容后,为了有效地进行后续的排序处理,可以先对多个媒体资源进行分组,将同一系列的媒体资源分到同一个组内,才可以有效地进行排序。而分组的依据具体可以为媒体资源所包含的多类数据各自的数据内容,依照数据优先级,依次通过各类数据内容来对媒体资源进行分组的判定,如果判定媒体资源可分组,则构建出相应的一组媒体资源,而剩余不可分组的媒体资源则可基于优先级中下一类数据内容来分组,通过多次循环,可以实现对各个媒体资源的分组,得到多组媒体资源。在其中一个实施例中,本申请用于对视频数据进行处理,用于生成视频序列,此时识别出的数据内容具体包括了内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据这四类数据。优先级从高到低依次为内容标签数据、封面图像特征数据、内容标题数据和封面文本数据。而分组过程中,则首先可以通过内容标签数据来进行分组,必有将具有相同内容标签的两个视频分为一组,而不存在相同内容标签的待分组视频进入下一轮次的分组,即通过封面图像特征数据来识别出封面图相似的待分组视频中的多组视频,之后再依次通过内容标题数据和封面文本数据来对上一轮次分组后无法分组的视频资源进行分组。当所有的数据内容用完后,得到即为分组完成的多组视频资源。

[0066] 步骤205,从多组媒体资源中各组媒体资源所包含的数据内容中,提取出数字型文本。

[0067] 其中,数字型文本是指包含数字类内容的文本数据,其包括各种的数字文本内容,

具体包括了1、2、3等阿拉伯数字的数据，或者one、two、three等英文数字的数据，或者一、二、三等中文数字的数据，此外对于中文数字，还可以零、壹、贰、叁等大写的中文数字，或者甲、乙、丙、丁等天干型数字，或者子、丑、寅、卯等地支型数字。

[0068] 具体地，由于数据内容中的数字型文本往往可以表征媒体资源的顺序数据，如对于包含多个不同剧集的连续剧资源组，其包含的数据内容有内容标题数据。具体包含有第一季第1集、第一季第2集、第二季第1集等内容标题数据。因此，可以提取出其中包含的数字型文本数据来对同一组内的不同媒体资源进行排序处理。在其中一个实施例中，提取数字型文本的过程具体可以通过正则表达式来实现，而在另一个实施例中，则可以通过机器学习模型来进行提取。

[0069] 步骤207，基于数字型文本所表征的顺序，对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序，得到多组媒体资源序列。

[0070] 其中，媒体资源序列是指对分组得到的各组媒体资源进行排序整理后的资源组数据，通过数字型文本可以对一个组别内的多个不同媒体资源进行排序，具体可以依据数字型文本从小到大或者从大到小的顺序，对同组媒体资源进行媒体资源的排序处理，得到一个完成的媒体资源序列。

[0071] 具体地，在进行排序时，媒体资源中提取出的数字大小一般表征了媒体资源在组内的序列顺序，因此可以基于这些数字型文本来实现对同一组内媒体资源数据的排序处理。在排序时，由于从数据内容中提取出的数字型文本可能包含多个。因此，可以对不同位置的数字型文本授予不同的优先级，比如优先标题中的数字型文本，先根据标题内的数字型文本进行排序，无法排序时，再基于封面图提取出的文本来进行排序。在其中一个实施例中，在排序处理时，为了进行统一性的比较，可以先将数字型文本统一转化为数值数据，再基于数值数据的顺序来对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序，得到多组媒体资源序列。在其中一个实施例中，本申请用于对电视剧类的视频数据进行处理，用于生成包含多个剧集序列的视频序列，此时，可以从指定的视频资源所包含的视频标题类数据中，比如从“第X集”中提取出包含数字型文本X，而后再针对同组的多个不同视频，比较X的大小，而后从小到大的顺序，生成各个剧集对应的剧集序列。而后当用户浏览视频时，如果用户观看完视频序列中在前的视频，则可向其推荐序列中当前观看完的视频之后的一个视频，保证视频的推荐效果。

[0072] 上述媒体资源序列的生成方法中，通过先针对多个媒体资源中的每一个媒体资源，识别每一个媒体资源所包含的多类数据各自的数据内容，而后即可基于数据内容来对媒体资源进行分组及排序；而后基于多类数据中至少一类数据的数据内容，对多个媒体资源进行分组，得到由多个媒体资源中不同媒体资源组成的多组媒体资源；通过媒体资源内的各类数据内容，可以有效地对媒体资源进行分组，从而为后续的排序处理提供数据基础。继而从多组媒体资源中各组媒体资源所包含的数据内容中，提取出数字型文本；由于数字型文本一般可以表征内容顺序，可以在对媒体资源进行分组后，再提取出其对应的数字型文本来进行后续的分组，最终则是基于数字型文本所表征的顺序，对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序，得到多组媒体资源序列，通过媒体资源的分组以及排序，可以有效地对杂乱的媒体资源进行排序，保证媒体资源归纳排序准确性。

[0073] 在一个实施例中，媒体资源所包含的多类数据包括内容标签数据以及内容标题数

据,步骤201包括:提取多个媒体资源中的每一个媒体资源的查询关键词;基于查询关键词,查找包含查询关键词的媒体资源所包含的内容标签数据以及内容标题数据。

[0074] 其中,查询关键词是指用于在数据库中查找媒体资源相关数据所用的检索键值。如对于视频,可以用Vid来作为单个视频的唯一标识,指代独一无二的视频,而在数据库中检索时,则可将Vid作为查询关键词查找,得到Vid所对应视频的内容标签数据以及内容标题数据。

[0075] 具体地,本申请中用于实现分类的媒体资源内数据具体包括内容标签数据以及内容标题数据。而内容标签数据和内容标题数据这两个数据具体可以从大数据分布式仓库中提取。在将媒体资源存储至资源库时,可以同时确定媒体资源所对应的内容标签数据和内容标题数据,而后将这两者与媒体资源的查询关键词进行关联存储,而后需要查找这些数据时,直接通过开发结构化查询语言进行汇总计算后,拉取进行本地化存储即可得到。在其中一个实施例中,本申请用于对电视剧类的视频数据进行处理,用于生成包含多个剧集序列的视频序列。此时,在将电视剧中的视频数据存储视频库时,针对每一个视频,都可以将其视频标题以及视频标签都与视频Vid关联存储至大数据分布式仓库。而后,在需要对视频进行整理时,可以预先选定好需要排序的视频,而后提取这些视频的Vid,再基于Vid从大数据分布式仓库中查找到相应的视频标题数据以及视频标签数据以便进行后续的排序。本实施例中,通过查询关键词来查找内容数据中的内容标签数据以及内容标题数据,可以有效保证内容数据提取的效率。

[0076] 在一个实施例中,媒体资源所包含的多类数据包括封面图像特征数据以及封面文本数据,步骤201包括:提取多个媒体资源中的每一个媒体资源的封面图像;对封面图像进行特征提取处理,得到封面图像特征数据,并对封面图像进行光学字符识别处理,得到封面文本数据。

[0077] 其中,封面图像是指用于展示媒体资源相关内容所用的封面,用户可以自行根据视频内容来选择合适的封面图。如对于视频,可以将视频中的某个画面作为封面图,也可以根据视频内容另外设计封面图。而对于文章,则可以直接将文章标题作为封面图等。特征提取处理具体是指提取封面图像上的图像特征,得到该封面图像对应的嵌入特征向量。而光学字符识别处理是指电子设备检查纸上打印的字符,通过检测暗、亮的模式确定其形状,然后用字符识别方法将形状翻译成计算机文字的过程;即,针对印刷体字符,采用光学的方式将纸质文档中的文字转换成为黑白点阵的图像文件,并通过识别软件将图像中的文字转换成文本格式,供文字处理软件进一步编辑加工的技术。而本申请中,则是直接对封面图像进行光学字符识别,来得到封面图像上所包含的文本内容。

[0078] 具体地,本申请中用于实现分类的媒体资源内数据还包括封面图像特征数据以及封面文本数据,而这两者都是来自于封面图,因此首先需要得到媒体资源所对应的封面图像。而封面图像同样可以从大数据分布式仓库中提取。而对于数据的存储过程,如对于在将电视剧中的视频数据存储视频库的过程,针对每一个视频,都可以通过爬虫或者对应的数据接口,获取到视频的封面图像,而后将封面图像与视频Vid关联存储至大数据分布式仓库。在其中一个实施例中,本申请用于对电视剧类的视频数据进行处理,用于生成包含多个剧集序列的视频序列。此时,在将电视剧中的视频数据存储视频库时,针对每一个视频,都可以将封面图像与视频Vid关联存储至大数据分布式仓库。而后,在需要对视频进行整理

时,可以预先选定好需要排序的视频,而后提取这些视频的Vid,再基于Vid从大数据分布式仓库中查找到相应的封面图像。而后,通过预先训练好的机器学习模型,从得到的封面图像中提取出封面图像特征数据。同时,通过光学字符识别来对封面图像进行处理,得到封面文本数据。本实施例中,通过查询封面图像来确定内容数据中的封面图像特征数据以及封面文本数据,可以有效保证内容数据提取的效率与准确性。

[0079] 在其中一个实施例中,步骤206包括:获取各类数据内容对应的分组优先级;基于分组优先级,依次通过各类数据内容对未分组的媒体资源进行分组,得到多组媒体资源。

[0080] 其中,分组优先级指的是不同类型的数据内容所对应的分组顺序是不同的,而在分组时,也是先基于优先级高的数据内容来对媒体资源进行分组,而后在依次通过优先级更低的其他数据内容来对无法分组的部分媒体资源进行再次分组。

[0081] 具体地,可以依赖不同的数据内容,通过不同的数据聚合方法来实现对媒体资源的分组处理,由于媒体资源的种类杂乱且繁多,而分组信息也根据资源类型的不同隐藏在不同类的数据内容中,但是一般同一序列下的媒体资源,其所对应的数据内容是相同的,如对于一部电视剧视频,其对应的数据内容可能是封面图类似,也可能是视频标题类似,或者视频标签中都包含有电视剧的剧名标签等。因此,可以根据实际的分组需要,预先选择好应用于分组所用的数据内容,并据此为不同类型的数据内容设置好分组优先级。在分组时,优先通过分组优先级高的数据内容来进行筛选分组,识别出可基于高分组优先级高的数据内容分组的媒体资源并分组。而后再通过其他类型的分组优先级更低的数据内容来依次对无法分组的资源数据再次分组,直到分组完成。本实施例中,依次通过不同优先级的数据内容对未分组的媒体资源进行分组,可以有效地通过多轮分组尽可能地将同类型的媒体资源分配到同一组别中,从而有效提高针对媒体资源进行排序处理的准确性。

[0082] 在其中一个实施例中,媒体资源所包含的多类数据包括内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据;

[0083] 基于分组优先级,依次通过各类数据内容对未分组的媒体资源进行分组,得到多组媒体资源包括:基于内容标签数据,对媒体资源进行分组,得到内容标签组媒体资源以及第一待分组媒体资源;基于封面图像特征数据,对第一待分组媒体资源进行分组,得到封面图特征组媒体资源以及第二待分组媒体资源;基于内容标题数据,对第二待分组媒体资源进行分组,得到内容标题组媒体资源以及第三待分组媒体资源;基于封面文本数据,对第三待分组媒体资源进行分组,得到封面文本组媒体资源。

[0084] 具体地,本申请中媒体资源所包含的多类数据包括内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据这四类数据,而其对应的分组优先级为内容标签数据为第一优先级,封面图像特征数据为第二优先级,内容标题数据为第三优先级,封面文本数据为第四优先级。因此,在分组时,其对应的分组顺序也是先通过内容标签数据对全量的媒体资源进行分组,得出可以通过内容标签数据分组的若干组内容标签组媒体资源以及无法依赖内容标签数据分组的第一待分组媒体资源。而后则是依赖封面图像特征数据,来对第一待分组媒体资源进行分组,同样得到多组封面图特征组媒体资源,以及无法通过封面图像特征数据分组第二待分组媒体资源。同理,可以根据内容标题数据,对第二待分组媒体资源进行分组,得到内容标题组媒体资源以及第三待分组媒体资源,最终则是依赖封面文本数据,对第三待分组媒体资源进行分组,得到封面文本组媒体资源,此处得到的封面文本

组媒体资源包含有可以通过封面文本数据分组的媒体资源。此外,对于无法通过内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据分组,即找不到类似数据的媒体数据,可以在封面文本组媒体资源中一个资源单列一组,同时直接放弃对其进行排序处理,因为组内只含有单个媒体资源,无需排序。在一个实施例中,申请用于对电视剧类的视频数据进行处理,用于生成包含多个剧集序列的视频序列,此时,可以根据预先设置,将分组优先级从高到低设置为内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据这四类数据,并如图3所示,可以依照这四类数据依次对由内容生产者发布的电视剧视频资源进行分组处理,得到多个待排序的电视剧剧集组,而后则可进一步地对这些电视剧剧集组进行一一排序,得到有序的电视剧剧集组。在另一个实施例中,本申请的方案还可用于对剪辑类视频进行分组,此时,可以调整各类数据对应的优先级,比如将内容标题数据作为第一优先级,而后依次设置各类数据的优先级以实现有效地分组。本实施例中,通过设置内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据这四类数据的优先组来依次对媒体资源进行有效地分组,从而保证针对媒体数据分组的有效性。

[0085] 在其中一个实施例中,基于内容标签数据,对媒体资源进行分组,得到内容标签组媒体资源以及第一待分组媒体资源包括:识别内容标签数据中的资源名称标签,并确定各资源名称标签中的媒体资源数;基于媒体资源数大于或等于二的资源名称标签得到内容标签组媒体资源,基于媒体资源数小于二的资源名称标签得到第一待分组媒体资源。

[0086] 其中,资源名称标签指的是内容标签数据中的一类标签,用于表征内容的名称,如对于电视剧媒体资源,其包含有剧名标签,通过剧名标签可以将属于同一个电视剧的不同视频划分到同组内。资源名称标签中的媒体资源数即表示含有资源名称标签的资源一共有多少个,如果媒体资源数大于或等于2,说明一个资源名称标签下包含有多个媒体资源,可以对其分组,若是媒体资源数等于1,说明该资源名称标签仅含有1个媒体资源,无法对单个资源分组,或者媒体资源中根本不含有资源名称标签,也说明无法通过内容标签数据,需要借助其他数据来实现对媒体资源的分组。

[0087] 具体地,对于媒体资源中的内容标签,其包含有一个特殊的标签,即资源名称标签。该标签能否反映媒体资源的实际名称。因此,在对媒体资源进行分组时,可以先查找媒体资源对应的内容标签数据,而后从其中提取出资源名称标签。并在提取完成之后,针对每一个资源名称标签下的媒体资源数,只有一个资源名称标签中包含有大于或等于二的媒体资源数,即可根据该资源名称标签构建一个内容标签组,而后将对应的媒体资源置于内容标签组中。否则将其作为第一待分组媒体资源以待后续过程的分配处理。在一个实施例中,申请用于对电视剧类的视频数据进行处理,用于生成包含多个剧集序列的视频序列。此时,针对多个待处理的视频,可以提取出每个视频所对应的剧名标签,并优先使用剧名标签作为分组依据,并将某个剧名标签下的所有内容划入同一内容单元,如“XX传”、“XX传说”等剧名标签。如果某个剧名标签下,存在大于等于2个内容,视为剧名标签可用。没有被剧名标签选入序列的视频数据会进行下一步的分组处理。本实施例中,通过资源名称标签来实现针对媒体资源的分组,可以有效保证媒体资源分组过程的准确性。

[0088] 在其中一个实施例中,基于封面图像特征数据,对第一待分组媒体资源进行分组,得到封面图特征组媒体资源以及第二待分组媒体资源包括:基于封面图像特征数据,得到第一待分组媒体资源中各媒体资源之间的封面图像欧氏距离矩阵;基于封面图像欧式距离

矩阵对第一待分组媒体资源进行聚合处理,得到封面图像特征聚簇;确定封面图像特征聚簇对应的轮廓系数,基于轮廓系数对封面图像特征聚簇进行切分,得到封面图像特征类簇;将媒体资源数大于或等于二的封面图像特征类簇作为封面图特征组媒体资源,将媒体资源数小于二的封面图像特征类簇作为第二待分组媒体资源。

[0089] 其中,欧氏距离即欧几里得度量,指在m维空间中两个点之间的真实距离,或者向量的自然长度。而在本申请中,媒体资源之间的欧氏距离矩阵指的是两个媒体资源所对应封面图像特征数据之间的向量距离。对于封面图像,每个媒体资源的封面都会对应唯一的N维嵌入特征向量,在N维向量空间中,可以计算得到任意两个向量之间的欧式距离矩阵。轮廓系数为衡量距离效果的指标,取值范围为[-1,1],越接近于1说明聚类效果越好,越接近-1说明聚类效果越差。公式如下:若a为某个点s与所有簇内点的距离的均值;b为某个点与所有簇外点的距离均值。则样本s的轮廓系数为(b-a)/max(a,b)。样本整体的轮廓系数为所有样本的均值。

[0090] 具体地,对于封面图像特征聚类的过程,可以先计算出第一待分组媒体资源中各媒体资源之间的封面图像欧氏距离矩阵,两个n维向量a(x_{11},12,13,・・・,$_{1n}$)与b(x_{21},22,23,・・・,x_{2n})间的欧式距离矩阵具体表示为:

[0091] $$d_{12}=\sqrt{\sum_{k=1}^{n}\left(x_{1k}-x_{2k}\right)^{2}}$$

[0092] 在得到欧式距离矩阵后,继而可以通过距离矩阵来实现第一待分组媒体资源内各个媒体资源的聚类处理,利用距离矩阵可进行凝聚的层次聚类,即将样本集中的所有的样本点都当做一个独立的类簇,假如共有M个媒体资源样本,根据所得的距离矩阵,距离最小的两个类簇c1和c2合并类簇c1和c2为一个类簇。在合并之后的M-1个簇的样本中,再次将其中最接近的两个簇合并。如果簇中有多个点的话,采用均值的方式,求两个簇中两两点的所有距离的均值作为两个簇的均值。最终能够生成簇内的树状结构,该结构可以参考图4所示。在生成树状结构后,由于目标是寻找封面图像除了压字之外完全一致的内容,因此在一个具体的实施例中,经过数据与业务实践,在(0,1]之间按照0.05的间隔(([0.05,0.1,0.15,0.2,0.25,0.3,0.35,0.4,0.45,0.5,0.55,0.6,0.65,0.7,0.75,0.8,0.85,0.9,0.95,1.]),循环在不同距离下的类簇切分值,分别计算不同场景下的轮廓系数,取轮廓系数最高的切分点。切分后的点会分别归类到不同的类簇中,而如果某个类簇下存在两个及以上的内容,即记为封面图类簇可用。基于轮廓系数的切分过程示意图如图5所示,其中横轴为切分距离,纵轴为轮廓系数。而基于封面图像特征数据所得到的同一组媒体资源具体参照图6所示。本实施例中,通过基于欧氏距离的资源聚合以及基于轮廓系数的聚簇切分,可以有效地将封面图像类似的不同媒体资源分到一起,保证媒体资源分组的准确性。

[0093] 在其中一个实施例中,基于内容标题数据,对第二待分组媒体资源进行分组,得到内容标题组媒体资源以及第三待分组媒体资源包括:确定第二待分组媒体资源中各媒体资源之间的标题文本相似度;根据标题文本相似度大于或等于预设标题文本相似度阈值的媒体资源得到内容标题组媒体资源,根据文本相似度小于预设标题文本相似度阈值的媒体资源得到第三待分组媒体资源。

[0094] 其中,标题文本相似度用于描述不同标题文本之间的向量距离,对于两个标题文本,其文本相似度=两个文本去重字符子集的重复文本字数/较短的文本字数。最终的距离为1-相似度。

[0095] 具体地,对于内容标题数据,可以通过标题文本相似度来计算两个标题文本之间的距离,而后基于标题文本的距离来将标题文本类似的部分标题数据聚合到一起,从而得到有效的内容标题组媒体资源。首先,先获取各个媒体资源所对应的标题文本,而后基于上述公式计算出文本相似度以及文本距离,在得到文本距离后,如上述针对特征距离的计算过程,可以先基于文本距离来对第二待分组媒体资源中的媒体资源进行聚类,再通过轮廓系数对封面图像特征聚簇进行切分,得到各个内容标题组。同理,对于封面文本数据也可以类似内容标题数据一样处理。先确定第三待分组媒体资源中各媒体资源之间的封面文本相似度,再基于封面文本相似度来聚类以及分组,得到最终的封面文本组媒体资源,以及无法分组部分媒体资源。在其中一个实施例中,针对视频进行标题文本相似度的处理,所得到的类似标题文本包括:title1_1='心机丈夫吞弹自尽,却给妻子一个大号惊喜!高能复仇《XXXX》#电影AABB大赛';title1_2='高能复仇爽片,处心积虑丈夫饮弹自尽,却给妻子留下致命陷阱!#电影AABB大赛';title1_3='全程高能的复仇爽片:心机丈夫吞弹自尽,却给妻子一个大号惊喜!#电影AABB大赛'。本实施例中,通过标题文本相似度来对不同类型的媒体资源进行分组,可以有效聚合标题文本类似的部分媒体资源,保证资源分组的有效性。

[0096] 在其中一个实施例中,方法从多组媒体资源中各组媒体资源所包含的数据内容中,提取出数字型文本包括:获取媒体资源所包含的数据内容中的文本型数据;通过正则表达式从文本型数据中提取数字型文本。

[0097] 其中,正则表达式是又称规则表达式,是一种文本模式,包括普通字符(例如,a到z之间的字母)和特殊字符(称为"元字符"),是计算机科学的一个概念。正则表达式使用单个字符串来描述、匹配一系列匹配某个句法规则的字符串,通常被用来检索、替换那些符合某个模式(规则)的文本。

[0098] 具体地,针对媒体资源所对应的各种数据内容,可以抽取出其对应的文本型数据,例如对于内容标签数据,可以直接获取标签文本作为文本型数据,而对于内容标题数据以及封面文本数据则可以直接将这些数据作为文本内容。由于媒体资源的数据内容中数字型文本一般都会以特殊的形式进行展示,所以可以通过构建正则表达式来从中抽取相应的数字型文本。例如,对于电视剧,其内容标题中包含有集数的信息,如第十五集,或者第13话等。通过正则表达式,可以抽取出其中和数字相关的关键文本表示。如对于第十五集可以抽取出数字型文本为十五。基于正则表达式来抽取数字型文本的示意图具体可以参照图7所示。本实施例中,通过正则表达式来进行内容提取,可以有效地从文本型数据中提取出数字型表示,从而保证媒体资源排序的准确性。

[0099] 在其中一个实施例中,步骤207包括:获取数字型文本对应的文本位置,并将数字型文本转化为数值型数据;基于文本位置以及数值型数据,对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0100] 其中,文本位置具体是指该数字型文本在媒体资源中的位置,例如位于媒体标签中,位于标题文本中,或者是来自于封面图文本中等,不同位置的数字型文本所表达的意义不同,通过识别数字型文本对应的文本位置,可以有效地提高媒体资源排序的准确性。而数

值型数据是指统一化的数值数据,例如可以将汉字的数字“十五”转化为数值型数据15,通过将不同类型的数字型文本转化为数值型数据,继而可以实现对这些数据的排序处理,而不转化的话处理过程较为复杂,需要针对不同类型的数字数据采用不同的排序方法。

[0101] 具体地,由于文本中可能会同时出现多组数字标识(如“《第109回:XXX诈病赚YY,夺走三代江山,一统天下》标题中,同时出现了“109”、“三”、“一”等多处数值相关表述),因此可以采用贪婪的匹配从左到右的第一个出现的数值,作为第一排序。匹配从右到左的第一个数值型,作为第二排序。仅当第一排序为空或相同时,才考虑第二排序,既能兼容上述例子,也能解决如下的标题问题:“女二号比女一号出彩的热播韩剧第五集”,同理在抽取文本内容时,同样拥有优先级,以内容标题为先,内容标题无顺序的(即上述抽取方式抽取不到关键词),再抽取封面图像文本数据等其他类型的文本数据,从而保证对媒体资源排序的准确性。本实施例中,通过数值型数据的转换以及文本位置的识别,能够在文本中出现多处数字内容时也保证排序的准确性。

[0102] 在其中一个实施例中,步骤201之前,还包括:获取资源排序请求;确定资源排序请求所指定的目标内容生产者;在预设资源数据库中查找目标内容生产者生产的多个媒体资源;步骤201包括:针对内容生产者生产的多个资源数据中的每一个媒体资源,识别每一个媒体资源所包含的多类数据各自的数据内容。

[0103] 其中,资源排序请求用于请求服务器104来对指定的资源数据进行排序,包含有用于指定排序目标的目标内容生产者信息。内容生产者是指生成资源数据,并将资源数据发布在服务器104内的用户,比如对于视频网站,用户可以将自制的视频上传到视频网站上以供其他用户浏览,这个上传视频的用户就是内容生产者。

[0104] 具体地,本申请的媒体资源序列的生成方法可以用于对各个内容生产者所生成的媒体资源进行分组以及排序,保证这些媒体资源分组排序的准确性,同时减少这些内容生产者的工作量。首先,当终端102需要对某个内容生产者所生产的资源数据进行排序时,工作人员可以在终端102上生成该内容生产者对应的资源排序请求,而后服务器104在接收到资源排序请求后,通过解析资源排序请求,可以确定工作人员所指定的目标内容生产者,而后则可基于确定的目标内容生产者,在媒体资源数据库中,查找到由该内容生产者所生成的所有资源数据,并将这些资源数据作为媒体资源排序的目标数据。在其中一个实施例中,资源排序请求中除了指定目标内容生产者之外,还可以指定时段信息,而后基于时段信息在媒体资源数据库中,查找到时段信息对应的媒体资源,而后将这些时段信息对应的媒体资源作为分组排序的对象。在另一个实施例中,资源排序请求处理指定目标内容生产者之外,还可以携带内容标签信息,而服务器104则可以根据内容标签信息筛选出媒体资源数据库中,该内容标签信息对应的部分媒体资源,而后在基于目标内容生产者对这部分媒体资源再进行一次筛选,得到的即为最终需要分组排序的媒体资源。本实施例中,先通过资源排序请求来指定需要排序的目标内容生产者生产的多个媒体资源,而后再对这些媒体资源进行排序,可以有效保证排序对象的合理性,从而提高媒体资源分组排序的准确性。

[0105] 在其中一个实施例中,上述的媒体资源序列的生成方法可以通过媒体资源序列生成模型实现,在预设资源数据库中查找目标内容生产者生产的多个媒体资源之后,还包括:将目标内容生产者生产的多个媒体资源输入媒体资源序列生成模型,得到媒体资源序列生成模型输出的多组媒体资源序列;获取目标对象对应的媒体资源浏览信息;当确定多组媒

体资源序列中包含媒体资源浏览信息对应的目标媒体资源时,得到媒体资源浏览信息中目标媒体资源的浏览完成度;当基于浏览完成度确定满足资源推荐条件时,确定多组媒体资源序列中目标媒体资源下一序列的待推荐媒体资源;推送待推荐媒体资源至目标对象。

[0106] 其中,可以将本申请的媒体资源序列的生成方法打包成一个媒体资源序列生成模型,通过将将目标内容生产者生产的多个媒体资源输入媒体资源序列生成模型,即可得到媒体资源序列生成模型输出的多组媒体资源序列。媒体资源浏览信息是指目标对象对媒体内容的浏览数据,比如用户在观看视频时,即可生成该视频对应的视频浏览信息。浏览完成度是指目标对象对媒体资源的观看程度,包括观看的时长等信息,资源推荐条件具体是指观看时长有没有达到推荐阈值等等。

[0107] 具体地,本申请的媒体资源序列的生成方法打包成一个媒体资源序列生成模型,而后工作人员只需要指定需要分组排序的目标内容生产者,服务器在通过目标内容生产者确定需要排序的媒体资源后,即可将这些媒体资源输入到模型中,由模型对这些媒体资源进行分组排序,得到多组媒体资源序列。在生成媒体资源序列后,还可基于生成的媒体资源序列完成资源推荐的功能。目标对象在浏览媒体资源时,服务器104可以得到相应的媒体资源浏览信息,只要目标对象浏览的目标媒体资源在多组媒体资源序列中,即可完成推荐。而后即可识别目标媒体资源的浏览完成度,如果浏览完成度达到了推荐阈值,满足资源推荐条件,则可判断目标对象希望进一步地浏览系列媒体资源内容,此时可以确定多组媒体资源序列中目标媒体资源下一序列的待推荐媒体资源。而后将这个待推荐媒体资源推荐给目标对象,目标对象则可基于反馈的待推荐媒体资源来进行后续的浏览,保证浏览过程的连续性。本实施例中,通过媒体资源的推荐,可以有效保证目标独享浏览过程的连续性。

[0108] 本申请还提供一种应用场景,该应用场景应用上述的媒体资源序列的生成方法。具体地,该发音评测方法在该应用场景的应用如下:

[0109] 当视频平台在对平台中各个用户所发布的视频内容进行整理时,为了对其中的序列视频进行有效整理,可以借助本申请的媒体资源序列的生成方法来实现对用户所发布视频的分组与排序。首先,平台可以先向搭载有本申请媒体资源序列的生成方法的服务器指定需要整理的视频,主要可以提供视频的Vid来指定待处理的目标视频,而后服务器在确定待处理的视频资源后。整体处理的框架可以参照图8所示,首先根据多种聚合方法,对视频进行聚合分组,得到不同组的媒体资源。其次根据聚合得到的视频组,识别视频组内视频的排序。最终还可以对于视频组内的排序进行校验,对于识别不到排序的内容单元,进行丢弃,不进入最终的视频内容池中,该过程中视频组的聚类识别过程与排序过程可以参照图9所示。其中,视频的分组过程可以依照多种数据内容来视频是否可分进行判断,并依次进行分组。而提取数据内容的过程可以参照图10所示,当平台指定需要为某个IP相关的视频进行整理排序时,服务器可以基于视频的Vid从数据仓库中提取到这些视频相关的内容标题数据与内容标签数据,同时通过爬虫爬取到这些视频相关的封面图像,而后通过图像特征处理,提取出封面图像所对应的封面图像特征数据,并对封面图像进行光学字符识别处理,得到封面文本数据。而后对这些数据进行数据结构化处理,从而得到可以用于分组的数据内容。基于内容标签数据,对视频进行分组,得到内容标签组视频以及第一待分组视频;基于封面图像特征数据,对第一待分组视频进行分组,得到封面图特征组视频以及第二待分组视频;基于内容标题数据,对第二待分组视频进行分组,得到内容标题组视频以及第三待

分组视频;基于封面文本数据,对第三待分组视频进行分组,得到封面文本组视频,而后将无法分组的视频从合集中放弃掉。对于内容标签数据的分组过程,可以先识别内容标签数据中的剧名标签,并确定各剧名标签中的视频数量;基于视频数量大于或等于二的剧名标签得到内容标签组视频,基于视频数量小于二的剧名标签得到第一待分组视频。而对于封面图像特征,则可以先基于封面图像特征数据,得到第一待分组视频中各视频之间的封面图像欧氏距离矩阵;基于封面图像欧式距离矩阵对第一待分组视频进行聚合处理,得到封面图像特征聚簇;确定封面图像特征聚簇对应的轮廓系数,基于轮廓系数对封面图像特征聚簇进行切分,得到封面图像特征类簇;将视频数大于或等于二的封面图像特征类簇作为封面图特征组视频,将视频数小于二的封面图像特征类簇作为第二待分组视频,而后的标题文本与封面文本处理过程类似,都是先确定各视频之间的文本相似度,而后基于文本相似度进行聚合并基于轮廓系数进行拆分,得到各个标题文本组视频与封面文本组视频,而无法聚和的部分视频则被舍弃。而在分组完成之后,则是排序的过程可以参照图11,包括关键词匹配提取与关键词转化排序两个过程,具体可以将数字型文本作为关键词,可以先获取媒体资源所包含的数据内容中的文本型数据;而后通过正则表达式从文本型数据中提取数字型文本,而排序过程则是先获取数字型文本对应的文本位置,并将数字型文本转化为数值型数据;基于文本位置以及数值型数据,对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。通过设置不同位置下文本的优先级,可以优先地将内容标题等位置下的数字作为排序的依据得到可靠的视频。而得到视频的分组与排序结果后,即可基于这些结果来实现对视频的推荐。如图12所示,图12具体为序列内容的形态,标题或封面图中一般会包含上、中、下或者一、二、三、四等能表示顺序的统一主题。生成的内容单元和内容顺序,将被推荐系统加载,用于在播放完成第N集时,自动续播下一集的视频内容。此外,本申请还可以应用于短视频发现页分发场景的情况,如图13所示,用户点击短视频应用下的发现条目观看视频,在分发过程中,当完播率大于某个阈值,且播放时长大于一定阈值时,会触发短剧跟随,在下一刷的推荐系统精排队列中出现(或强插)当前短视频内容所属序列中下一个视频内容的vid。具体地处理过程可以参照图14所示。研究发现,通过对短剧跟随内容池的内容特征的丰富,提升了序列内容的覆盖度和标注准确度。同时在下游使得推荐系统对于内容的选择更加多元,提升了推荐系统的推荐效率同时,也通过短剧跟随的页面展示形式,提升了用户的关注转化以及对内容生产者的品牌认知。扩充了50万拥有序列标记的视频,占推荐分发高时效内容池的17.54%、抽样评估100个IP序列(覆盖内容量660条),准确序列98个,准确度98%。而针对不同的内容池,进行了线上的A/B实验,实验结果显示:核心指标对比大盘持平有略微提升,曝光人均播放量提升0.15%,后台订阅率,人均后台订阅次数提升约5%。而实验组和对照组相比,在短视频发现页影视综剧名标签的人均曝光播放数据,实验组有约0.7%的提升。

[0110] 应该理解的是,虽然如上的各实施例所涉及的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,如上的各实施例所涉及的流程图中的至少一部分步骤可以包括多个步骤或者多个阶段,这些步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤中的步骤或

者阶段的至少一部分轮流或者交替地执行。

[0111] 基于同样的发明构思,本申请实施例还提供了一种用于实现上述所涉及的媒体资源序列的生成方法的媒体资源序列的生成装置。该装置所提供的解决问题的实现方案与上述方法中所记载的实现方案相似,故下面所提供的一个或多个媒体资源序列的生成装置实施例中的具体限定可以参见上文中对于媒体资源序列的生成方法的限定,在此不再赘述。

[0112] 在一个实施例中,如图15所示,提供了一种媒体资源序列的生成装置,包括:

[0113] 数据内容识别模块1502,用于针对多个媒体资源中的每一个媒体资源,识别每一个媒体资源所包含的多类数据各自的数据内容。

[0114] 媒体资源分组模块1504,用于基于多类数据中至少一类数据的数据内容,对多个媒体资源进行分组,得到由多个媒体资源中不同媒体资源组成的多组媒体资源。

[0115] 文本提取模块1506,用于从多组媒体资源中各组媒体资源所包含的数据内容中,提取出数字型文本。

[0116] 序列生成模块1508,用于基于数字型文本所表征的顺序,对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0117] 在其中一个实施例中,媒体资源所包含的多类数据包括内容标签数据以及内容标题数据;数据内容识别模块1502具体用于:提取多个媒体资源中的每一个媒体资源的查询关键词;基于查询关键词,查找包含查询关键词的媒体资源所包含的内容标签数据以及内容标题数据。

[0118] 在其中一个实施例中,媒体资源所包含的多类数据包括封面图像特征数据以及封面文本数据;数据内容识别模块1502具体用于:提取多个媒体资源中的每一个媒体资源的封面图像;对封面图像进行特征提取处理,得到封面图像特征数据,并对封面图像进行光学字符识别处理,得到封面文本数据。

[0119] 在其中一个实施例中,媒体资源分组模块1504具体用于:获取各类数据内容对应的分组优先级;基于分组优先级,依次通过各类数据内容对未分组的媒体资源进行分组,得到多组媒体资源。

[0120] 在其中一个实施例中,媒体资源所包含的多类数据包括内容标签数据、封面图像特征数据、内容标题数据以及封面文本数据;媒体资源分组模块1504具体用于:基于内容标签数据,对媒体资源进行分组,得到内容标签组媒体资源以及第一待分组媒体资源;基于封面图像特征数据,对第一待分组媒体资源进行分组,得到封面图特征组媒体资源以及第二待分组媒体资源;基于内容标题数据,对第二待分组媒体资源进行分组,得到内容标题组媒体资源以及第三待分组媒体资源;基于封面文本数据,对第三待分组媒体资源进行分组,得到封面文本组媒体资源。

[0121] 在其中一个实施例中,媒体资源分组模块1504具体用于:识别内容标签数据中的资源名称标签,并确定各资源名称标签中的媒体资源数;基于媒体资源数大于或等于二的资源名称标签得到内容标签组媒体资源,基于媒体资源数小于二的资源名称标签得到第一待分组媒体资源。

[0122] 在其中一个实施例中,媒体资源分组模块1504具体用于:基于封面图像特征数据,得到第一待分组媒体资源中各媒体资源之间的封面图像欧氏距离矩阵;基于封面图像欧式距离矩阵对第一待分组媒体资源进行聚合处理,得到封面图像特征聚簇;确定封面图像特

征聚簇对应的轮廓系数,基于轮廓系数对封面图像特征聚簇进行切分,得到封面图像特征类簇;将媒体资源数大于或等于二的封面图像特征类簇作为封面图特征组媒体资源,将媒体资源数小于二的封面图像特征类簇作为第二待分组媒体资源。

[0123] 在其中一个实施例中,媒体资源分组模块1504具体用于:确定第二待分组媒体资源中各媒体资源之间的标题文本相似度;根据标题文本相似度大于或等于预设标题文本相似度阈值的媒体资源得到内容标题组媒体资源,根据文本相似度小于预设标题文本相似度阈值的媒体资源得到第三待分组媒体资源。

[0124] 在其中一个实施例中,文本提取模块1506具体用于:获取媒体资源所包含的数据内容中的文本型数据;通过正则表达式从文本型数据中提取数字型文本。

[0125] 在其中一个实施例中,序列生成模块1508具体用于:获取数字型文本对应的文本位置,并将数字型文本转化为数值型数据;基于文本位置以及数值型数据,对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序,得到多组媒体资源序列。

[0126] 在其中一个实施例中,还包括媒体资源识别模块,用于:获取资源排序请求;确定资源排序请求所指定的目标内容生产者;在预设资源数据库中查找目标内容生产者生产的多个资源数据,数据内容识别模块1502具体用于:针对内容生产者生产的多个资源数据中的每一个媒体资源,识别每一个媒体资源所包含的多类数据各自的数据内容。

[0127] 在其中一个实施例中,上述的媒体资源序列生成方法通过媒体资源序列生成模型实现,装置还用于:将目标内容生产者生产的多个资源数据输入媒体资源序列生成模型,得到媒体资源序列生成模型输出的多组媒体资源序列。装置还包括媒体资源推荐模块,用于:获取目标对象对应的媒体资源浏览信息;当确定多组媒体资源序列中包含媒体资源浏览信息对应的目标媒体资源时,得到媒体资源浏览信息中目标媒体资源的浏览完成度;当基于浏览完成度确定满足资源推荐条件时,确定多组媒体资源序列中目标媒体资源下一序列的待推荐媒体资源;推送待推荐媒体资源至目标对象。

[0128] 上述媒体资源序列的生成装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行以上各个模块对应的操作。

[0129] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是服务器,其内部结构图可以如图16所示。该计算机设备包括处理器、存储器、输入/输出接口(Input/Output,简称I/O)和通信接口。其中,处理器、存储器和输入/输出接口通过系统总线连接,通信接口通过输入/输出接口连接到系统总线。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质和内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储媒体资源序列的生成相关的数据。该计算机设备的输入/输出接口用于处理器与外部设备之间交换信息。该计算机设备的通信接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种媒体资源序列的生成方法。

[0130] 本领域技术人员可以理解,图16中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设

备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0131] 在一个实施例中,还提供了一种计算机设备,包括存储器和处理器,存储器中存储有计算机程序,该处理器执行计算机程序时实现上述各方法实施例中的步骤。

[0132] 在一个实施例中,提供了一种计算机可读存储介质,存储有计算机程序,该计算机程序被处理器执行时实现上述各方法实施例中的步骤。

[0133] 在一个实施例中,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各方法实施例中的步骤。

[0134] 需要说明的是,本申请所涉及的用户信息(包括但不限于用户设备信息、用户个人信息等)和数据(包括但不限于用于分析的数据、存储的数据、展示的数据等),均为经用户授权或者经过各方充分授权的信息和数据,且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

[0135] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一非易失性计算机可读取存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、数据库或其它介质的任何引用,均可包括非易失性和易失性存储器中的至少一种。非易失性存储器可包括只读存储器(Read-Only Memory,ROM)、磁带、软盘、闪存、光存储器、高密度嵌入式非易失性存储器、阻变存储器(ReRAM)、磁变存储器(Magnetoresistive Random Access Memory,MRAM)、铁电存储器(Ferroelectric Random Access Memory,FRAM)、相变存储器(Phase Change Memory,PCM)、石墨烯存储器等。易失性存储器可包括随机存取存储器(Random Access Memory,RAM)或外部高速缓冲存储器等。作为说明而非局限,RAM可以是多种形式,比如静态随机存取存储器(Static Random Access Memory,SRAM)或动态随机存取存储器(Dynamic Random Access Memory,DRAM)等。本申请所提供的各实施例中所涉及的数据库可包括关系型数据库和非关系型数据库中至少一种。非关系型数据库可包括基于区块链的分布式数据库等,不限于此。本申请所提供的各实施例中所涉及的处理器可为通用处理器、中央处理器、图形处理器、数字信号处理器、可编程逻辑器、基于量子计算的数据处理逻辑器等,不限于此。

[0136] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0137] 以上所述实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对本申请专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范围。因此,本申请的保护范围应以所附权利要求为准。

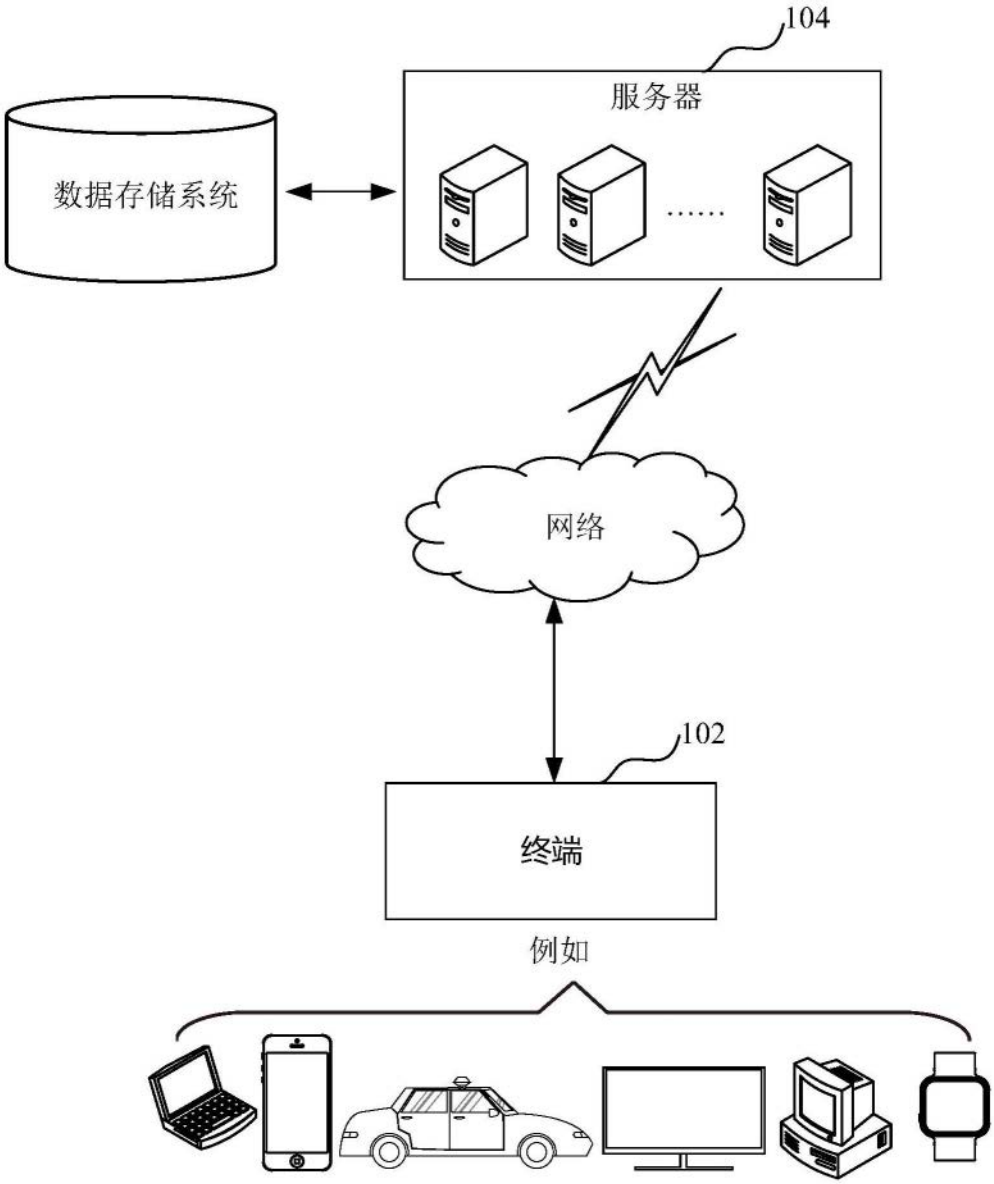


图1

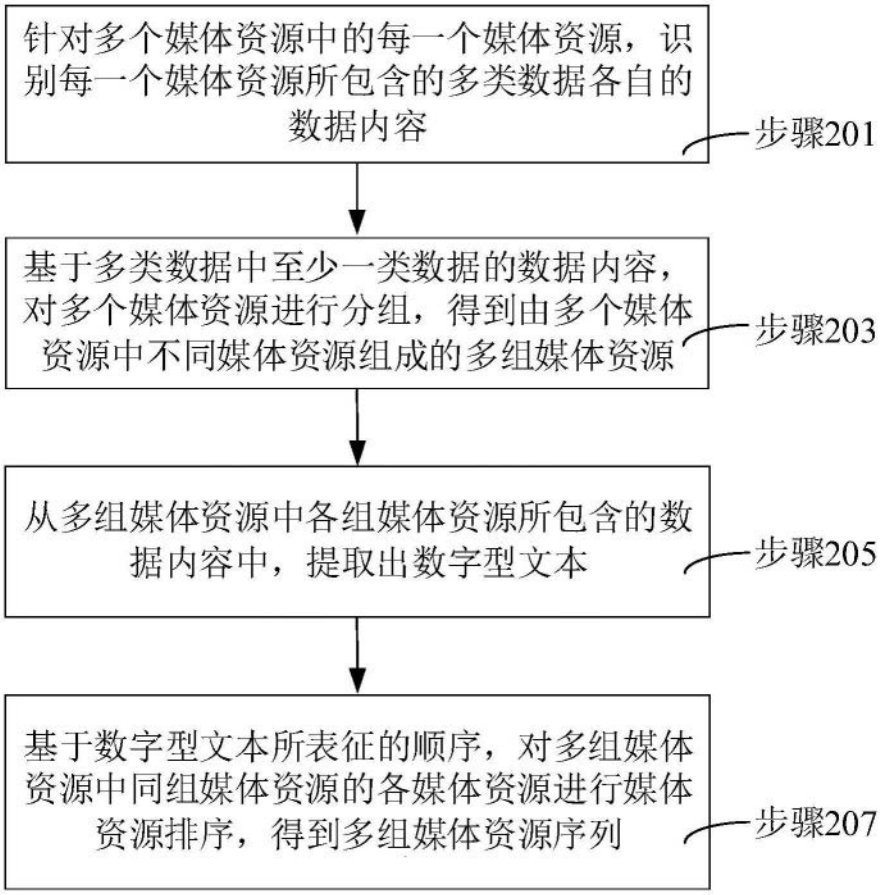

针对多个媒体资源中的每一个媒体资源，识别每一个媒体资源所包含的多类数据各自的数据内容
步骤201
基于多类数据中至少一类数据的数据内容，对多个媒体资源进行分组，得到由多个媒体资源中不同媒体资源组成的多组媒体资源
步骤203
从多组媒体资源中各组媒体资源所包含的数据内容中，提取出数字型文本
步骤205
基于数字型文本所表征的顺序，对多组媒体资源中同组媒体资源的各媒体资源进行媒体资源排序，得到多组媒体资源序列
步骤207

图2

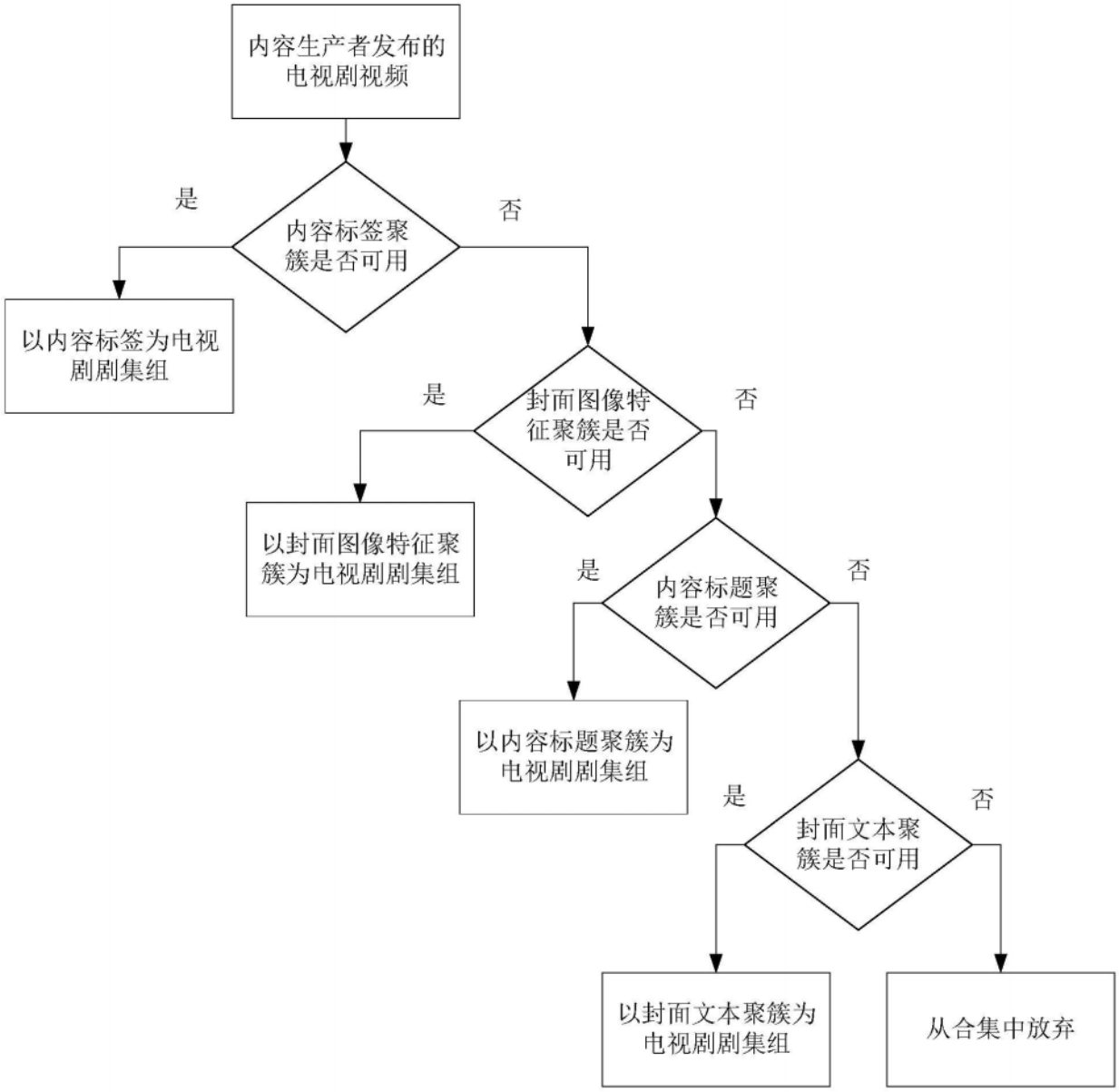

内容生产者发布的电视剧视频
内容标签聚簇是否可用
是
否
以内容标签为电视剧剧集组
封面图像特征聚簇是否可用
是
否
以封面图像特征聚簇为电视剧剧集组
内容标题聚簇是否可用
是
否
以内容标题聚簇为电视剧剧集组
封面文本聚簇是否可用
是
否
以封面文本聚簇为电视剧剧集组
从合集中放弃

图3

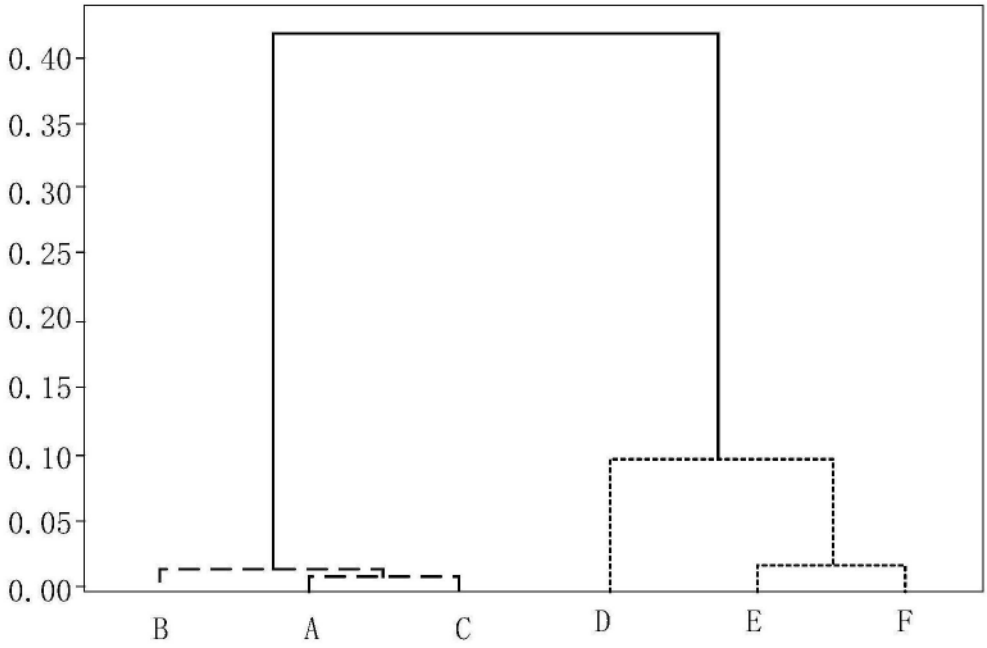

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
B
A
C
D
E
F

图4

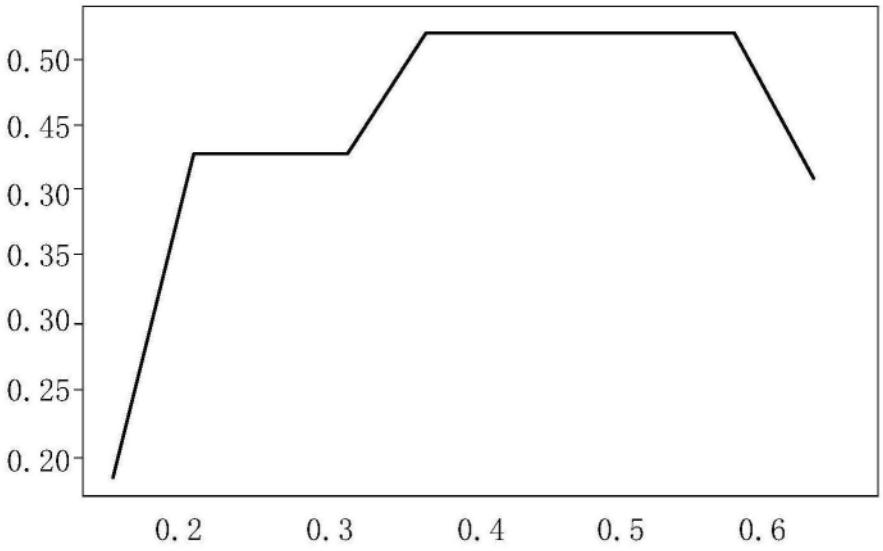

0.50
0.45
0.30
0.35
0.30
0.25
0.20
0.2
0.3
0.4
0.5
0.6

图5

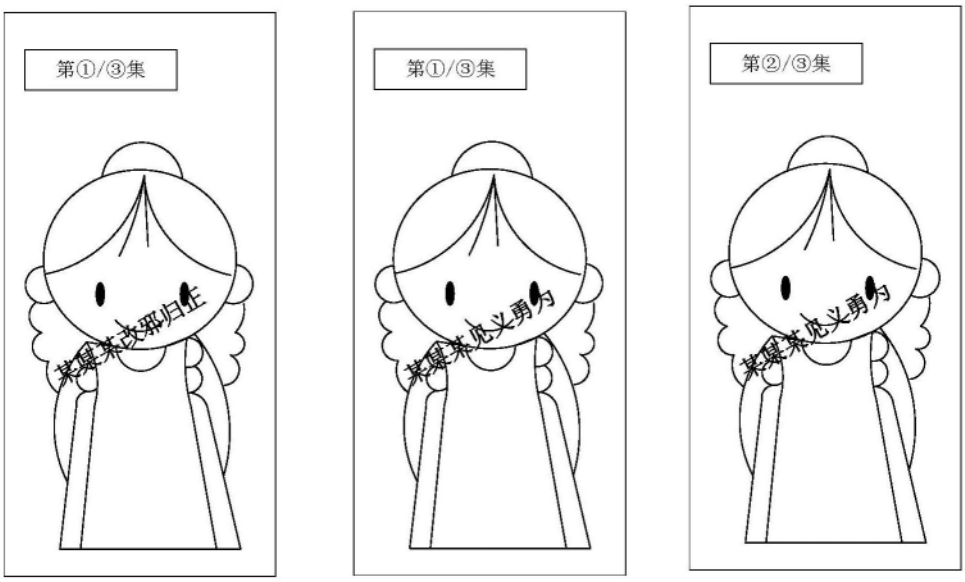

第①/③集
某键某改邪归正
第①/③集
某键某见义勇为
第②/③集
某键某见义勇为

图6

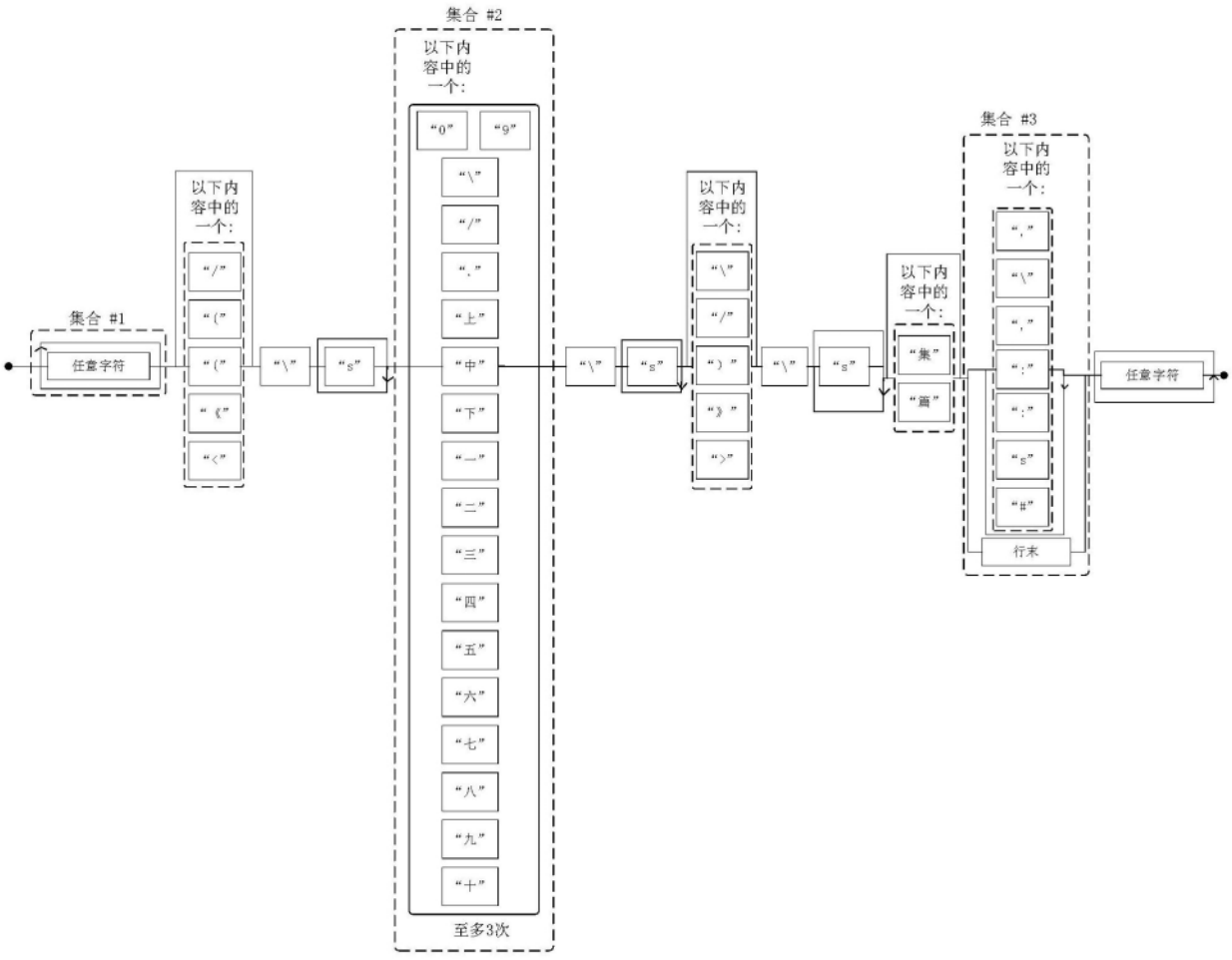

集合 #1
任意字符
以下内容中的一个:
“/”
“(”
“(”
“《”
“<”
“\”
“s”
集合 #2
以下内容中的一个:
“0”
“9”
“\”
“/”
“、”
“上”
“中”
“下”
“一”
“二”
“三”
“四”
“五”
“六”
“七”
“八”
“九”
“十”
至多3次
“\”
“s”
以下内容中的一个:
“\”
“/”
“)”
“》”
“>”
“\”
“s”
以下内容中的一个:
“集”
“置”
集合 #3
以下内容中的一个:
“,”
“\”
“,”
“:”
“:”
“s”
“#”
行末
任意字符

图7

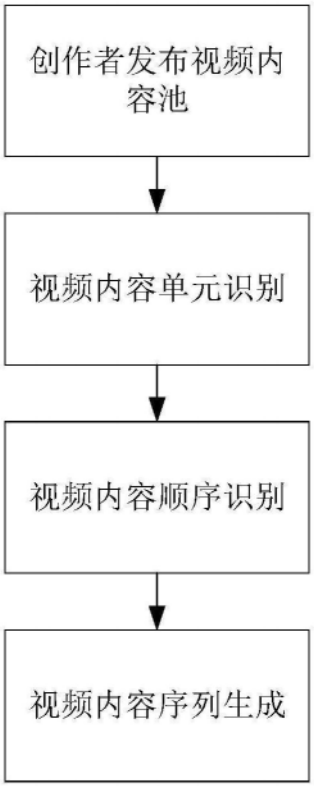


图8

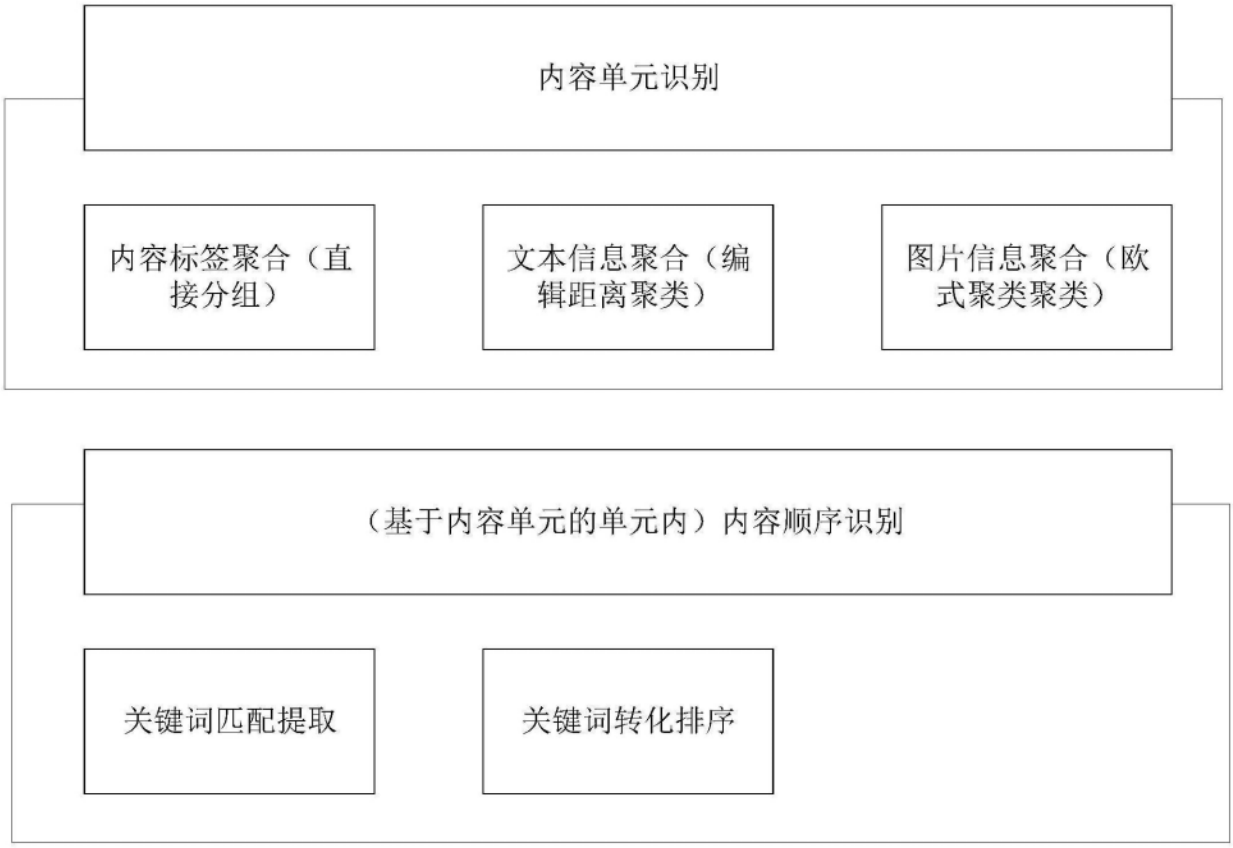

内容单元识别
内容标签聚合（直接分组）
文本信息聚合（编辑距离聚类）
图片信息聚合（欧式聚类聚类）
（基于内容单元的单元内）内容顺序识别
关键词匹配提取
关键词转化排序

图9

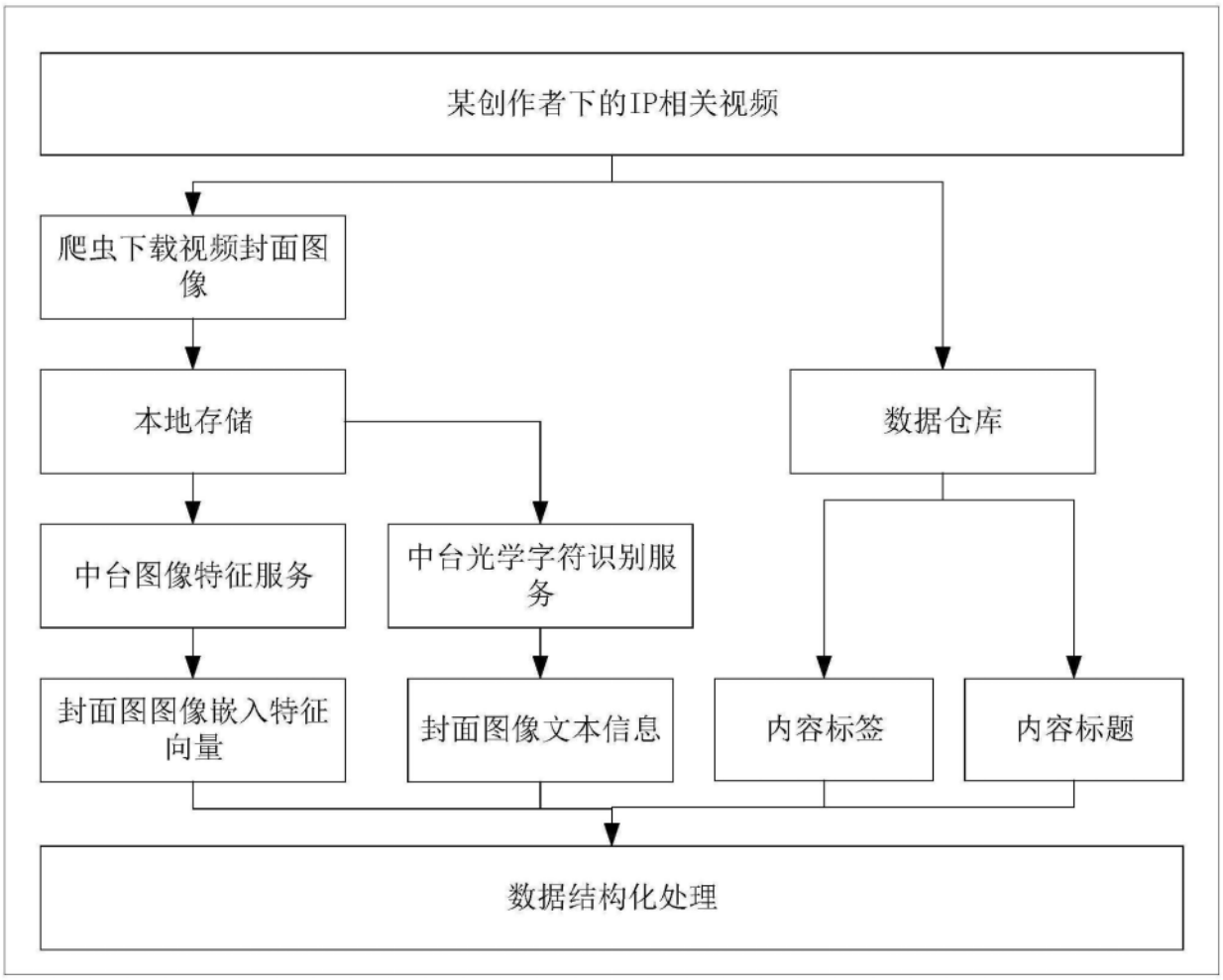

某创作者下的IP相关视频
爬虫下载视频封面图像
本地存储
数据仓库
中台图像特征服务
中台光学字符识别服务
封面图图像嵌入特征向量
封面图像文本信息
内容标签
内容标题
数据结构化处理

图10

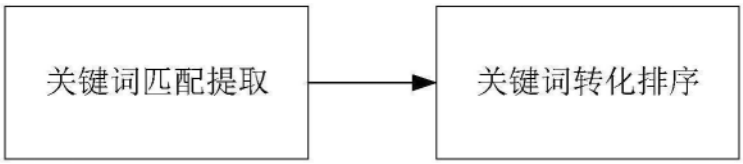

关键词匹配提取
关键词转化排序

图11


订阅 精选 发现 热点 电视剧 记录 二
《片名》第1/3集：一句简介
@某某某
+订阅
XX
评论
分享
英雄不问出处，评论不看字数
首页
圈儿
VIP会员
消息
个人中心
《片名》第2/3集：一句简介
@某某某
+订阅
XX
评论
分享
评论有你，心动不已
《片名》第3/3集：一句简介
@某某某
+订阅
XX
评论
分享
笑点泪点齐飞，评论等你发挥

图12


订阅 精选 发现 热点 电视剧 记录
内容内容内容内容内容内容内容内容内容
@某某某
某某某
看正片
+订阅
XX
评论
分享
笑点泪点齐飞，评论等你发挥

图13

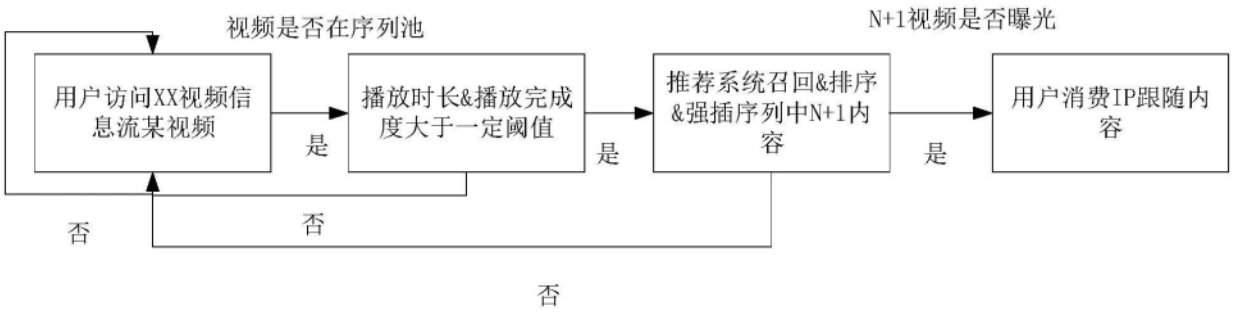

视频是否在序列池
用户访问XX视频信息流某视频
播放时长&播放完成度大于一定阈值
是
推荐系统召回&排序&强插序列中N+1内容
N+1视频是否曝光
是
用户消费IP跟随内容
否
否
否

图14

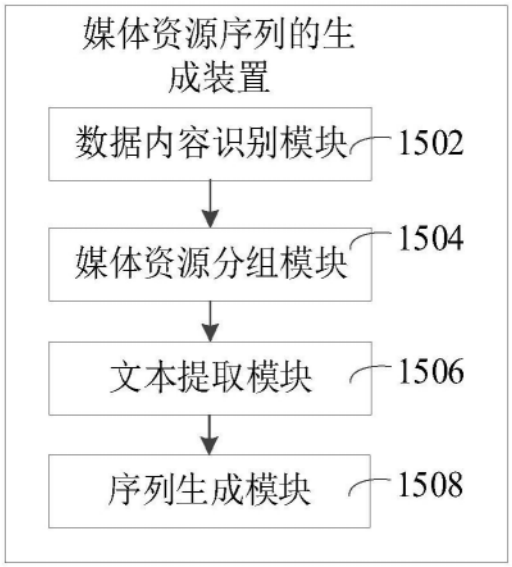

媒体资源序列的生成装置
数据内容识别模块
1502
媒体资源分组模块
1504
文本提取模块
1506
序列生成模块
1508

图15

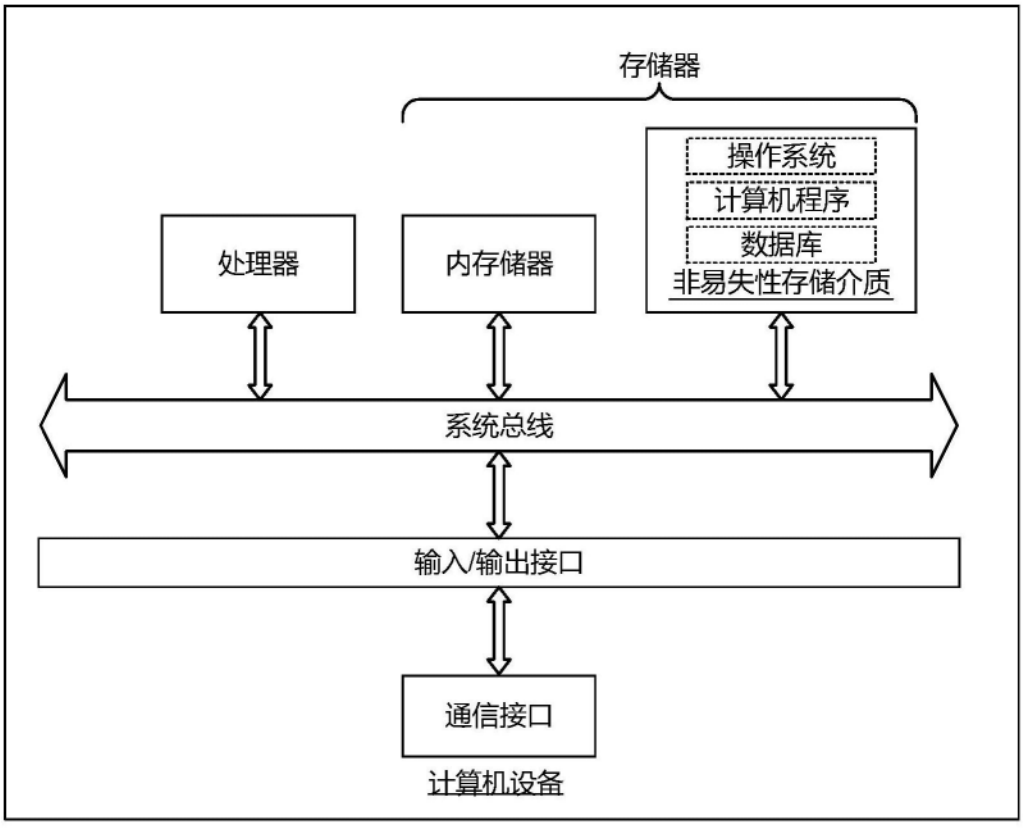

存储器
操作系统
计算机程序
数据库
非易失性存储介质
处理器
内存储器
系统总线
输入/输出接口
通信接口
计算机设备

图16