(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116340552 B
(45)授权公告日 2024.07.02

(21)申请号 202310016743.7
(22)申请日 2023.01.06
(65)同一申请的已公布的文献号
申请公布号 CN 116340552 A
(43)申请公布日 2023.06.27
(73)专利权人 北京达佳互联信息技术有限公司
地址 100085 北京市海淀区上地西路6号1幢1层101D1-7
(72)发明人 王英俊
(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
专利代理师 申健
(51)Int.Cl.
G06F 16/48(2019.01)
G06F 18/22(2023.01)
G06F 18/214(2023.01)
G06N 3/045(2023.01)
G06N 3/08(2023.01)
(56)对比文件
CN 115374278 A,2022.11.22
CN 111274494 A,2020.06.12
审查员 陈茜

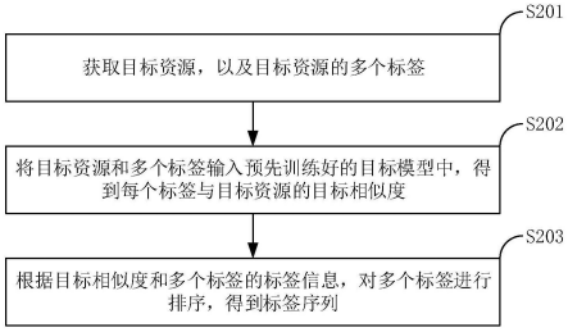
权利要求书2页 说明书15页 附图7页

(54)发明名称

一种标签排序方法、装置、设备及存储介质

(57)摘要

本公开关于一种标签排序方法、装置、设备及存储介质,涉及互联网技术领域,可以实现对多媒体资源的多个标签进行准确排序。该方法包括:获取目标资源,以及目标资源的多个标签;将目标资源和多个标签输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度;目标模型是根据样本数据训练初始模型得到的;样本数据包括:样本资源,以及样本资源的多个样本标签;根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。



CN 116340552 B

1.一种标签排序方法,其特征在于,包括:

获取样本数据;所述样本数据包括:样本资源,以及所述样本资源的多个样本标签;每个所述样本标签包括:多个字符;

根据所述样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型;所述模型整体损失是根据损失集合得到的;所述损失集合包括:第一损失和第二损失;所述第一损失用于表征所述样本资源的样本资源特征与所述样本标签的样本标签特征的差异;所述第二损失用于表征所述字符的预测值与所述字符的真实值的差异;所述初始模型包括第一特征提取模型、第二特征提取模型;所述根据所述样本数据训练初始模型,包括:将所述样本资源输入所述第一特征提取模型,得到所述样本资源的样本资源特征;将每个所述样本标签输入所述第二特征提取模型,得到每个所述样本标签的样本标签特征;当所述第二特征提取模型为多层神经网络模型时,所述样本标签特征包括所述多层神经网络模型中的多个隐含层数据;每个所述样本标签包括:多个字符;所述第二特征提取模型包括:BERT模型;对输入的每个所述样本标签的至少一个所述字符进行遮蔽,得到每个所述样本标签对应的遮蔽训练数据;所述遮蔽训练数据包括:每个所述样本标签中,与至少一个所述字符一一对应的至少一个遮蔽位置;将所述遮蔽训练数据输入所述BERT模型,得到所述遮蔽位置的字符的预测值;

获取目标资源,以及所述目标资源的多个标签;

将所述目标资源和多个所述标签输入所述目标模型中,得到每个所述标签与所述目标资源的目标相似度;

根据所述目标相似度和多个所述标签的标签信息,对多个所述标签进行排序,得到标签序列。

2.根据权利要求1所述的标签排序方法,其特征在于,所述初始模型还包括:判别模型;所述损失集合,还包括:所述判别模型的对抗损失;所述标签排序方法,还包括:

将所述样本数据输入所述判别模型,得到所述样本数据的数据类型的预测值;所述数据类型用于表征所述样本数据为所述样本资源或者所述样本标签;所述对抗损失用于表征所述数据类型的预测值与所述数据类型的真实值的差异。

3.根据权利要求1所述的标签排序方法,其特征在于,所述根据所述样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型,包括:

将训练数据输入所述初始模型,确定软目标标签,以及所述模型整体损失;所述训练数据包括所述样本数据;所述软目标标签包括:多个所述样本标签中,对应的所述样本标签特征与所述样本资源特征的相似度大于预设阈值的标签;

当所述模型整体损失不满足所述收敛条件时,更新所述初始模型的模型参数,并将所述软目标标签和所述样本资源作为所述训练数据,输入更新后的初始模型中,直至所述更新后的初始模型对应的模型整体损失满足所述收敛条件。

4.根据权利要求1所述的标签排序方法,其特征在于,所述标签信息包括:所述标签的维度信息和/或属性信息;所述维度信息用于表征所述标签的类型特征;所述属性信息用于表征所述标签的数据来源、出现位置、相互关系中的至少一种;

所述根据所述目标相似度和多个所述标签的标签信息,对多个所述标签进行排序,得到标签序列,包括:

根据所述目标相似度,对多个所述标签进行排序,得到初始序列;

根据所述标签信息和预设优先级策略,更新所述初始序列,得到所述标签序列。

5.一种标签排序装置,其特征在于,包括:获取单元和第一处理单元、第二处理单元;

所述第二处理单元,被配置为执行获取样本数据;所述样本数据包括:样本资源,以及所述样本资源的多个样本标签;每个所述样本标签包括:多个字符;

所述第二处理单元,还被配置为执行根据所述样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型;所述模型整体损失是根据损失集合得到的;所述损失集合包括:第一损失和第二损失;所述第一损失用于表征所述样本资源的样本资源特征与所述样本标签的样本标签特征的差异;所述第二损失用于表征所述字符的预测值与所述字符的真实值的差异;所述初始模型包括第一特征提取模型、第二特征提取模型;所述第二处理单元,具体被配置为执行:将所述样本资源输入所述第一特征提取模型,得到所述样本资源的样本资源特征;将每个所述样本标签输入所述第二特征提取模型,得到每个所述样本标签的样本标签特征;当所述第二特征提取模型为多层神经网络模型时,所述样本标签特征包括所述多层神经网络模型中的多个隐含层数据;每个所述样本标签包括:多个字符;所述第二特征提取模型包括:BERT模型;对输入的每个所述样本标签的至少一个所述字符进行遮蔽,得到每个所述样本标签对应的遮蔽训练数据;所述遮蔽训练数据包括:每个所述样本标签中,与至少一个所述字符一一对应的至少一个遮蔽位置;将所述遮蔽训练数据输入所述BERT模型,得到所述遮蔽位置的字符的预测值;

所述获取单元,被配置为执行获取目标资源,以及所述目标资源的多个标签;

所述第一处理单元,被配置为执行将所述目标资源和多个所述标签输入所述目标模型中,得到每个所述标签与所述目标资源的目标相似度;

所述第一处理单元,还被配置为执行根据所述目标相似度和多个所述标签的标签信息,对多个所述标签进行排序,得到标签序列。

6.一种电子设备,其特征在于,所述电子设备包括:

处理器;

用于存储所述处理器可执行指令的存储器;

其中,所述处理器被配置为执行所述指令,以实现如权利要求1-4中任一项所述的标签排序方法。

7.一种计算机可读存储介质,所述计算机可读存储介质上存储有指令,其特征在于,当所述计算机可读存储介质中的指令由电子设备的处理器执行时,使得所述电子设备能够执行如权利要求1-4中任一项所述的标签排序方法。

一种标签排序方法、装置、设备及存储介质

技术领域

[0001] 本公开涉及互联网技术领域,尤其涉及一种标签排序方法、装置、设备及存储介质。

背景技术

[0002] 随着网络技术飞速发展,各种多媒体资源已经成为大数据的主体。标签可以精确提炼多媒体资源的内容,利用标签可以实现为订阅标签的用户精准推送多媒体资源。

[0003] 但是,多媒体资源通常对应多个标签,例如,打篮球的视频,对应的标签可以包括:运动、男生、操场、篮球等。由于不同的标签与多媒体资源的关联程度不同,当关联程度越高时,多媒体资源更容易引起订阅该标签的用户的兴趣。因此,对多媒体资源的多个标签进行排序,是多媒体资源的推广过程中可以改进的地方。

发明内容

[0004] 本公开提供一种标签排序方法、装置、设备及存储介质,可以实现对多媒体资源的多个标签进行准确排序。

[0005] 本公开实施例的技术方案如下:

[0006] 根据本公开实施例的第一方面,提供一种标签排序方法,包括:获取目标资源,以及目标资源的多个标签;将目标资源和多个标签输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度;目标模型是根据样本数据训练初始模型得到的;样本数据包括:样本资源,以及与样本资源对应的多个样本标签;根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。

[0007] 可选的,初始模型包括第一特征提取模型、第二特征提取模型;标签排序方法,还包括:获取样本数据;将样本资源输入第一特征提取模型,得到样本资源的样本资源特征;将每个样本标签输入第二特征提取模型,得到每个样本标签的样本标签特征;当第二特征提取模型为多层神经网络模型时,样本标签特征包括多层神经网络模型中的多个隐含层数据。

[0008] 可选的,每个样本标签包括:多个字符;第二特征提取模型包括:BERT模型;标签排序方法,还包括:对输入的每个样本标签的至少一个字符进行遮蔽,得到每个样本标签对应的遮蔽训练数据;遮蔽训练数据包括:每个样本标签中,与至少一个字符一一对应的至少一个遮蔽位置;将遮蔽训练数据输入BERT模型,得到遮蔽位置的字符的预测值。

[0009] 可选的,标签排序方法,还包括:获取第一损失和第二损失;第一损失用于表征样本资源特征与样本标签特征的差异;第二对比损失用于表征字符的预测值与字符的真实值的差异;根据模型整体损失训练初始模型,直至满足收敛条件,得到目标模型;模型整体损失是根据损失集合得到的;损失集合包括:第一损失和第二损失。

[0010] 可选的,初始模型还包括:判别模型;损失集合,还包括:判别模型的对抗损失;标签排序方法,还包括:将样本数据输入判别模型,得到样本数据的数据类型的预测值;数据

类型用于表征样本数据为样本资源或者样本标签;对抗损失用于表征数据类型的预测值与数据类型的真实值的差异。

[0011] 可选的,根据样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型的方法,包括:将训练数据输入初始模型,确定软目标标签,以及模型整体损失;训练数据包括样本数据;软目标标签包括:多个样本标签中,对应的样本标签特征与样本资源特征的相关度大于预设阈值的标签;当模型整体损失不满足收敛条件时,更新初始模型的模型参数,并将软目标标签和样本资源作为训练数据,输入更新后的初始模型中,直至更新后的初始模型对应的模型整体损失满足收敛条件。

[0012] 可选的,标签信息包括:标签的维度信息和/或属性信息;维度信息用于表征标签的类型特征;属性信息用于表征标签的数据来源、出现位置、相互关系中的至少一种;根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列的方法,包括:根据目标相似度,对多个标签进行排序,得到初始序列;根据标签信息和预设优先级策略,更新初始序列,得到标签序列。

[0013] 根据本公开实施例的第二方面,提供一种标签排序装置,包括:获取单元和第一处理单元;获取单元,被配置为执行获取目标资源,以及目标资源的多个标签;第一处理单元,被配置为执行将目标资源和多个标签输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度;目标模型是根据样本数据训练初始模型得到的;样本数据包括:样本资源,以及与样本资源对应的多个样本标签;第一处理单元,还被配置为执行根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。

[0014] 可选的,初始模型包括第一特征提取模型、第二特征提取模型;该标签排序装置,还包括:第二处理单元;第二处理单元,被配置为执行获取样本数据;将样本资源输入第一特征提取模型,得到样本资源的样本资源特征;将每个样本标签输入第二特征提取模型,得到每个样本标签的样本标签特征;当第二特征提取模型为多层神经网络模型时,样本标签特征包括多层神经网络模型中的多个隐含层数据。

[0015] 可选的,每个样本标签包括:多个字符;第二特征提取模型包括:BERT模型;第二处理单元,还被配置为执行对输入的每个样本标签的至少一个字符进行遮蔽,得到每个样本标签对应的遮蔽训练数据;遮蔽训练数据包括:每个样本标签中,与至少一个字符一一对应的至少一个遮蔽位置;将遮蔽训练数据输入BERT模型,得到遮蔽位置的字符的预测值。

[0016] 可选的,第二处理单元,还被配置为执行获取第一损失和第二损失;第一损失用于表征样本资源特征与样本标签特征的差异;第二对比损失用于表征字符的预测值与字符的真实值的差异;根据样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型;模型整体损失是根据损失集合得到的;损失集合包括:第一损失和第二损失。

[0017] 可选的,初始模型还包括:判别模型;损失集合,还包括:判别模型的对抗损失;第二处理单元,还被配置为执行将样本数据输入判别模型,得到样本数据的数据类型的预测值;数据类型用于表征样本数据为样本资源或者样本标签;对抗损失用于表征数据类型的预测值与数据类型的真实值的差异。

[0018] 可选的,第二处理单元,具体被配置为执行将训练数据输入初始模型,确定软目标标签,以及模型整体损失;训练数据包括样本数据;软目标标签包括:多个样本标签中,对应的样本标签特征与样本资源特征的相关度大于预设阈值的标签;当模型整体损失不满足收

敛条件时,更新初始模型的模型参数,并将软目标标签和样本资源作为训练数据,输入更新后的初始模型中,直至更新后的初始模型对应的模型整体损失满足收敛条件。

[0019] 可选的,标签信息包括:标签的维度信息和/或属性信息;维度信息用于表征标签的类型特征;属性信息用于表征标签的数据来源、出现位置、相互关系中的至少一种;第一处理单元,具体被配置为执行根据目标相似度,对多个标签进行排序,得到初始序列;根据标签信息和预设优先级策略,更新初始序列,得到标签序列。

[0020] 根据本公开实施例的第三方面,提供一种电子设备,可以包括:处理器和用于存储处理器可执行指令的存储器;其中,处理器被配置为执行指令,以实现上述第一方面中任一种可选的标签排序方法。

[0021] 根据本公开实施例的第四方面,提供一种计算机可读存储介质,计算机可读存储介质上存储有指令,当计算机可读存储介质中的指令由电子设备的处理器执行时,使得电子设备能够执行上述第一方面中任一种可选的标签排序方法。

[0022] 根据本公开实施例的第五方面,提供一种计算机程序产品,包含指令,当其在电子设备的处理器上运行时,使得电子设备执行上述第一方面中任一种可选的标签排序方法。

[0023] 本公开提供的技术方案至少带来以下有益效果:

[0024] 基于上述任一方面,本公开中,在获取到目标资源,以及目标资源的多个标签之后,可以将目标资源和多个标签,输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度。其中,目标模型是根据样本数据训练初始模型得到的。样本数据包括:样本资源,以及与样本资源的多个样本标签。然后,可以根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。由于本公开融合了考虑标签与资源的相似度,以及标签自身的标签信息,因此,可以实现对目标资源的多个标签的合理排序,进而使得目标资源的推广可以更精准地契合用户的兴趣标签,提高推广效率。

附图说明

[0025] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理,并不构成对本公开的不当限定。

[0026] 图1为本公开实施例提供的一种标签排序系统示意图;

[0027] 图2示出了本公开实施例提供的一种标签排序方法的流程示意图一;

[0028] 图3示出了本公开实施例提供的一种标签排序方法的流程示意图二;

[0029] 图4示出了本公开实施例提供的一种训练初始模型的示意图;

[0030] 图5示出了本公开实施例提供的一种标签排序方法的流程示意图三;

[0031] 图6示出了本公开实施例提供的一种标签排序方法的流程示意图四;

[0032] 图7示出了本公开实施例提供的一种标签排序方法的流程示意图五;

[0033] 图8示出了本公开实施例提供的一种标签排序方法的流程示意图六;

[0034] 图9示出了本公开实施例提供的一种标签排序方法的流程示意图七;

[0035] 图10示出了本公开实施例提供的一种标签参数的示例图;

[0036] 图11示出了本公开实施例提供的一种标签排序装置的结构示意图一;

[0037] 图12示出了本公开实施例提供的一种标签排序装置的结构示意图二。

具体实施方式

[0038] 为了使本领域普通人员更好地理解本公开的技术方案,下面将结合附图,对本公开实施例中的技术方案进行清楚、完整地描述。

[0039] 需要说明的是,本公开的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本公开的实施例能够以除了在这里图示或描述的那些以外的顺序实施。以下示例性实施例中所描述的实施方式并不代表与本公开相一致的所有实施方式。相反,它们仅是与如所附权利要求书中所详述的、本公开的一些方面相一致的装置和方法的例子。

[0040] 还应当理解的是,术语“包括”指示所描述特征、整体、步骤、操作、元素和/或组件的存在,但并不排除一个或多个其他特征、整体、步骤、操作、元素和/或组件的存在或添加。

[0041] 本公开所涉及的数据可以为经用户授权或者经过各方充分授权的数据。

[0042] 以下对本公开涉及到的现有技术进行解释。

[0043] 一、对比文本-图像预训练模型(contrastive language-image pre-training, CLIP)

[0044] 在目前图像分类领域中,训练的模型通常会遇到以下问题:模型需要用到大量的格式化标注数据,这些标注数据获取通常成本高昂;模型在当前数据集的效果比较好,但是可能模型的泛化能力较差,同时迁移到新的训练任务也比较困难等。与此同时,互联网上面已经存在了大量的图像文本对(在网页中,开发者一般都会为图片添加一段文字备注),实际上这些素材可以作为已经标注好的数据集,利用这些数据集进行训练,既能解决获取标注数据成本高昂的问题,同时也因为互联网上的数据量比较大和数据本身差异较大,更容易获得泛化能力较强的模型。

[0045] CLIP模型就是基于上述概念,使用图像文本对,分别将文本和图像进行编码,之后进行训练,得到的用于匹配图像和文本的预训练神经网络模型。其目标是将图像与文本的相似度提高,核心流程比较简洁。

[0046] 二、BERT(bidirectional encoder representations from transformers)模型

[0047] 1、BERT模型是一种双向编码器,旨在通过在左右上下文中共有的条件计算来预先训练来自无标号文本的深度双向表示。因此,经过预先训练的BERT模型只需一个额外的输出层就可以进行微调,从而为各种自然语言处理任务生成最新模型。

[0048] 2、MLM(mask language model)和NSP(next sentence prediction)

[0049] 为了能够更好训练BERT网络,在BERT的训练过程中引入两个任务,MLM和NSP。对于MLM任务来说,其做法是随机掩盖掉输入序列中的随机标识Token(即用“[MASK]”替换掉原有的Token),然后在BERT的输出结果中取对应掩盖位置上的向量进行真实值预测。

[0050] 一些如问答、自然语言推断等任务需要理解两个句子之间的关系,而MLM任务倾向于抽取Token层次的表征,因此不能直接获取句子层次的表征。为了使模型能够有能力理解句子间的关系,BERT使用了NSP任务来预训练,简单来说就是预测两个句子是否连在一起。具体的做法是:对于每一个训练样例,在语料库中挑选出句子A和句子B来组成,50%的时候句子B就是句子A的下一句(标注为Is Next),剩下50%的时候句子B是语料库中的随机句子(标注为Not Next)。接下来把训练样例输入到BERT模型中,进行二分类的预测。

[0051] 三、独热编码(one-hot)

[0052] One-Hot编码,又称为一位有效编码,主要是用来解决类别型数据的离散值问题,采用N位状态寄存器来对N个状态进行编码,每个状态都由他独立的寄存器位,并且在任意时候只有一位有效。

[0053] One-Hot编码是分类变量作为二进制向量的表示。这首先要求将分类值映射到整数值。然后,每个整数值被表示为二进制向量,除了整数的索引之外,它都是零值,它被标记为1。

[0054] 使用one-hot编码,将离散特征的取值扩展到了欧式空间,离散特征的某个取值就对应欧式空间的某个点。将离散特征通过one-hot编码映射到欧式空间,是因为,在回归,分类,聚类等机器学习算法中,特征之间距离的计算或相似度的计算是非常重要的,而我们常用的距离或相似度的计算都是在欧式空间的相似度计算,计算余弦相似性,基于的就是欧式空间将离散型特征使用one-hot编码,确实会让特征之间的距离计算更加合理。

[0055] 独热编码解决了分类器不好处理属性数据的问题,在一定程度上也起到了扩充特征的作用。它的值只有0和1,不同的类型存储在垂直的空间。

[0056] 四、交叉熵(cross entropy)

[0057] 交叉熵主要用于度量两个概率分布间的差异性信息。语言模型的性能通常用交叉熵和复杂度(perplexity)来衡量。交叉熵的意义是用该模型对文本识别的难度,或者从压缩的角度来看,每个词平均要用几个位来编码。复杂度的意义是用该模型表示这一文本平均的分支数,其倒数可视为每个词的平均概率。平滑是指对没观察到的N元组合赋予一个概率值,以保证词序列总能通过语言模型得到一个概率值。

[0058] 五、软目标(soft targets)更新

[0059] 为了保证目标模型的模型参数在每次迭代中都会更新,相当于更新间隔为1。软目标更新依赖一个动量更新的编码器(encoder),将当前模型参数输入目标算法,得到更新后的模型参数。

[0060] 如背景技术中所描述,多媒体资源通常对应多个标签,例如,打篮球的视频,对应的标签可以包括:运动、男生、操场、篮球等。由于不同的标签与多媒体资源的关联程度不同,当关联程度越高时,多媒体资源更容易引起订阅该标签的用户的兴趣。因此,对多媒体资源的多个标签进行排序,是多媒体资源的推广过程中可以改进的地方。

[0061] 基于此,本公开实施例提供一种标签排序方法,在获取到目标资源,以及目标资源的多个标签之后,可以将目标资源和多个标签,输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度。其中,目标模型是根据样本数据训练初始模型得到的。样本数据包括:样本资源,以及与样本资源的多个样本标签。然后,可以根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。由于本公开融合了考虑标签与资源的相似度,以及标签自身的标签信息,因此,可以实现对目标资源的多个标签的合理排序,进而使得目标资源的推广可以更精准地契合用户的兴趣标签,提高推广效率。

[0062] 以下结合附图对本公开实施例提供的标签排序方法进行示例性说明:

[0063] 图1为本公开实施例提供的一种标签排序系统示意图,如图1所示,该标签排序系统中可以包括:平台服务器110和标签排序设备120。

[0064] 其中,平台服务器110和标签排序设备120之间可以通过有线网络或者无线网络连

接。

[0065] 平台服务器110可以是一些应用平台的数据服务器。例如,应用平台可以是短视频应用服务平台、新闻服务平台、直播服务平台、购物服务平台、外卖服务平台、共享服务平台、功能性网站等。其中,短视频应用服务平台提供的多媒体资源可以为一些短视频作品,新闻服务平台提供的多媒体资源可以为一些新闻信息,直播服务平台提供的多媒体资源可以为直播作品等,其余不再一一赘述。本公开对多媒体资源服务平台的具体类型并不作限制。

[0066] 一些实施例中,平台服务器110可以是单独的一个服务器,或者,也可以是由多个服务器构成的服务器集群。部分实施方式中,服务器集群还可以是分布式集群。本公开对平台服务器110的具体实现方式也不作限制。

[0067] 标签排序设备120可以是为资源的标签进行排序的设备,标签排序设备120可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云计算服务的云服务器。

[0068] 在一些实施例中,标签排序设备120还可以包含有数据库或与数据库连接,数据库中存储有资源的资源数据和标识数据,其中,资源可以包括视频、图片、广告、新闻等可以在互联网上传播的内容。

[0069] 具体的,标签排序设备120对目标资源的多个标签进行排序后,可以将标签序列发送给平台服务器110。平台服务器110根据标签序列,向对应的用户推广目标资源。

[0070] 应理解的,图1所示的标签排序系统仅为示例,在实际应用中,可以由终端或者服务器独立执行本公开实施例的标签排序方法,也可以由终端和服务器配合执行本公开实施例的标签排序方法,本公开实施例对具体的标签排序系统不作限定。

[0071] 需要说明的是,上述平台服务器110和标签排序设备120均可以称为电子设备。

[0072] 本公开实施例提供的标签排序方法可以应用于前述图1所示的应用场景中的标签排序设备120。

[0073] 如图2所示,该标签排序方法可以包括:

[0074] S201、标签排序设备获取目标资源,以及目标资源的多个标签。

[0075] 可选的,目标资源可以包括:图像资源、文字资源、音频资源、视频资源等,例如短视频平台推送的视频、购物平台推送的商品等。

[0076] 可选的,标签可以是由多个字符组成的文本。

[0077] 在一种可以实现的方式中,平台服务器可以向标签排序设备发送待排序的目标资源和多个标签,相应的,标签排序设备可以对目标资源的多个标签进行排序。

[0078] S202、标签排序设备将目标资源和多个标签输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度。

[0079] 其中,目标模型是根据样本数据训练初始模型得到的。样本数据包括:样本资源,以及样本资源的多个样本标签。

[0080] 可选的,目标模型和初始模型可以为CLIP模型、或者基于CLIP模型的改进模型。

[0081] 可选的,目标模型和初始模型中均可以包括:判别模型、第一特征提取模型、第二特征提取模型。其中,判别模型用于识别输入数据的数据类型,第一特征提取模型用于提取资源特征,第二特征提取模型用于提取标签特征。

[0082] 可选的,第一特征提取模型可以为多模的嵌入模型。例如,河图(HeTu)模型。当目标资源为视频资源时,第一特征提取模型可以从视频封面、视频帧识别、语音识别等多方面提取资源数据。

[0083] 可选的,第二特征提取模型可以包括:BERT模型、注意力模型。

[0084] 在一种可以实现的方式中,标签排序设备可以将输入目标模型的数据先输入判别模型,划分出目标资源和标签。然后,分别将目标资源输入第一特征提取模型,得到目标资源特征,将标签输入第二特征提取模型,得到目标标签特征。后续,匹配目标资源特征与目标标签特征,得到每个标签与目标资源的目标相似度。

[0085] S203、标签排序设备根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。

[0086] 可选的,标签信息可以包括:标签的维度信息和/或属性信息。维度信息用于表征标签的类型特征。属性信息用于表征标签的数据来源、出现位置、相互关系中的至少一种。

[0087] 可选的,粒度特征包括:标签的数据来源、出现位置、相互关系中的至少一种。

[0088] 可选的,标签的相互关系包括:相似关系、同义关系、上下位关系等。

[0089] 示例性的,相似关系例如美妆与彩妆,同义关系例如日光与阳光,上下位关系例如爵士舞与古风爵士。

[0090] 可选的,数据来源包括:通过模型、检索、以及切词等得到的。出现位置表示标签出现在文本中的位置,以及文本前后特殊字符,例如标签首次出现的顺序,在文本中是否出现书名号等特殊形式等。相互关系表示文本结构中标签的相互关系,例如游戏英雄属于游戏等。

[0091] 上述实施例提供的技术方案至少带来以下有益效果:由S201-S203可知,在获取到目标资源,以及目标资源的多个标签之后,可以将目标资源和多个标签,输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度。其中,目标模型是根据样本数据训练初始模型得到的。样本数据包括:样本资源,以及与样本资源的多个样本标签。然后,可以根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。由于本公开融合了考虑标签与资源的相似度,以及标签自身的标签信息,因此,可以实现对目标资源的多个标签的合理排序,进而使得目标资源的推广可以更精准地契合用户的兴趣标签,提高推广效率。

[0092] 在一种可选的实施例中,当初始模型包括第一特征提取模型、第二特征提取模型时,在图2示出的方法实施例的基础上,本实施例提供一种可能实现方式。如图3所示,该标签排序方法,还包括:

[0093] S301、标签排序设备获取样本数据。

[0094] 在一种可以实现的方式中,结合现有技术,标签排序设备可以获取到网络中大量的图像文本对或者视频文本对,将图像或者视频作为样本资源,描述资源的文本作为样本标签,得到样本数据。

[0095] S302、标签排序设备将样本资源输入第一特征提取模型,得到样本资源的样本资源特征。

[0096] 在一种可以实现的方式中,标签排序设备可以通过第一特征提取模型可以对样本资源执行光学字符识别(optical character recognition,OCR)、自动语音识别

(automatic speech recognition,ASR)等操作,得到样本资源特征。

[0097] S303、标签排序设备将每个样本标签输入第二特征提取模型,得到每个样本标签的样本标签特征。

[0098] 可选的,当第二特征提取模型为多层神经网络模型时,样本标签特征包括多层神经网络模型中的多个隐含层数据。

[0099] 在一种可以实现的方式中,标签排序设备可以针对标签进行从输入层到隐含层的转换并提取对应的隐含层数据。

[0100] 可选的,多层神经网络模型包括多个隐含层,本公开实施例以12个隐含层为例进行说明,隐含层的层数不作限定。

[0101] 在一种可以实现的方式中,当第二特征提取模型包括:BERT模型和注意力模型时,标签排序设备可以将样本标签输入BERT模型,得到表征向量,然后对表征向量进行注意力学习,得到样本标签特征。

[0102] 示例性的,如图4所示,标签排序设备获取到的样本数据包括:短视频01与标签(tag)001、短视频02与标签002。

[0103] 短视频01通过HeTu模型进行特征提取,得到样本资源特征,即河图嵌入值(HeTu embedding)A。标签001通过BERT模型,得到表征向量:[CLS]天、上、飞……[SEP],然后对表征向量进行注意力学习,得到样本标签特征,即标签嵌入值(Tag embedding)a。

[0104] 短视频02通过HeTu模型进行特征提取,得到样本资源特征,即河图嵌入值B。标签002通过BERT模型,得到表征向量:[CLS]星、辰、大……[SEP],然后对表征向量进行注意力学习,得到样本标签特,即标签嵌入值b。

[0105] 在一种可以实现的方式中,标签排序设备可以将样本资源特征与样本标签特征配对得到样本对。

[0106] 在一种可以实现的方式中,标签排序设备可以与样本资源特征对应的样本标签特征确定为正样本对,并将样本数据中,除正样本对以外的其他样本对确定为负样本对。

[0107] 结合上述示例,标签排序设备可以将样本资源特征(河图嵌入值A、河图嵌入值B),分别与样本标签特征(标签嵌入值a、标签嵌入值b)分别配对,得到样本对,包括:样本对A-a、样本对A-b、样本对B-a、样本对B-b。

[0108] 其中,样本对A-a与短视频01对应,样本数据B-b与短视频02对应。因此,标签排序设备可以将样本对A-a、样本对B-b确定为正样本对,样本对A-b、样本对B-a确定为负样本对。

[0109] 上述实施例提供的技术方案至少带来以下有益效果:由S301-S303可知,标签排序设备可以在获取到样本数据之后,将样本资源输入第一特征提取模型,得到样本资源特征,将每个样本标签输入第二特征提取模型,得到样本标签特征。本公开实施例提供一种对输入的数据进行特征提取的方法,以使得后续可以提取后的特征进行匹配,确定资源与标签的相似度。

[0110] 同时,当第二特征提取模型为多层神经网络模型时,样本标签特征包括多层神经网络模型中的多个隐含层数据。因此,本公开实施例可以保留隐含层数据,以使得对标签提取出来的样本标签特征更丰富,进一步提高了模型的泛化能力。

[0111] 在一种可选的实施例中,当每个样本标签包括多个字符、第二特征提取模型包括:

BERT模型时,在图3示出的方法实施例的基础上,本实施例提供一种可能实现方式。如图5所示,该标签排序方法,还包括:

[0112] S501、标签排序设备对输入的每个样本标签的至少一个字符进行遮蔽,得到每个样本标签对应的遮蔽训练数据。

[0113] 其中,遮蔽训练数据包括:每个样本标签中,与至少一个字符一一对应的至少一个遮蔽位置。

[0114] S502、标签排序设备将遮蔽训练数据输入BERT模型,得到遮蔽位置的字符的预测值。

[0115] 具体的,为了能够更好训练BERT网络,在BERT的训练过程中引入两个任务,MLM和NSP。对于MLM任务来说,其做法是随机遮蔽掉至少一个字符对应的随机标识Token(即用“[M]”替换掉原有的Token),然后在BERT的输出结果中取对应遮蔽位置上的向量进行真实值预测。

[0116] 示例性的,结合图4,标签排序设备可以对“天”、“星”进行遮蔽,然后输入注意力模型进行学习。

[0117] 上述实施例提供的技术方案至少带来以下有益效果:由S501-S502可知,标签排序设备对输入的每个样本标签的至少一个字符进行遮蔽,得到每个样本标签对应的遮蔽训练数据,然后,标签排序设备可以将遮蔽训练数据输入BERT模型,得到遮蔽位置的字符的预测值。本公开提供了一种对模型进行预训练的方法,可以通过遮蔽样本标签的字符,训练模型得到完整标签,实现模型的深度学习,进一步提高了模型的准确性。

[0118] 在一种可选的实施例中,在图5示出的方法实施例的基础上,本实施例提供一种可能实现方式。结合图3,如图6所示,该标签排序方法,还包括:

[0119] S601、标签排序设备确定第一损失和第二损失。

[0120] 其中,第一损失用于表征样本资源特征与样本标签特征的差异。

[0121] 在一种可以实现的方式中,可以通过损失函数得到第一损失。可选的,损失函数可以为公式(1)。

[0122] $$L_{con} = -log\frac{exp(Sim(q_i,k_j)/\tau)}{\sum_{i=1}^{B} exp(Sim(q_i,k_j)/\tau)} \quad (1)$$

[0123] 其中,L_{con}为第一损失,q_i为第i个样本资源特征,k_j为第j个样本标签特征,τ为可训练的温度系数,B为样本数据的个数。

[0124] 可选的,公式(1)中的向量内积可以通过公式(2)得到。

[0125] $$Sim(z_i,Z_j)=l_2(z_i)\cdot l_2(z_j) \quad (2)$$

[0126] 其中,l_2范数定义为向量所有元素的平方和的开平方,z_i、z_j为任意元素。

[0127] 在一种可以实现的方式中,考虑到实际应用中的模型损失(loss),第一损失的损失值可以通过交叉熵的形式得到。此时,第一损失如公式(3)所示。

[0128] $$L_{con}=L_C(S(q,k),y(q,k)) \quad (3)$$

[0129] 其中,L_C表示交叉熵公式,y(q,k)表示样本数据的独热编码(one hot)。当样本数据为正样本数据时,独热编码为1,当样本数据为负样本数据时,独热编码为0。

[0130] S(q,k)为样本数据的归一化函数指数,满足公式(4)。

[0131] $$S(q,k)=\frac{exp(Sim(q_i,k_j)/\tau)}{\sum_{i=1}^{B}exp(Sim(q_i,k_j)/\tau)} \quad (4)$$

[0132] 在一种可以实现的方式中,第二对比损失用于表征字符的预测值与字符的真实值的差异,如公式(5)。

[0133] $$L_{ce}=-log\frac{exp(q^m)}{\sum_{n=0}^{c}exp(q^n)} \quad (5)$$

[0134] 其中,L_{ce}为第二损失。

[0135] S602、标签排序设备根据样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型。

[0136] 其中,模型整体损失是根据损失集合得到的,损失集合包括:第一损失和第二损失。

[0137] 可选的,收敛条件包括:当样本数据为正样本对时,模型整体损失小于预设上限值,对应的样本资源特征与样本标签特征的差异值小,即相似度高。反之,当样本数据为负样本对时,模型整体损失大于预设下限值,对应的样本资源特征与样本标签特征的差异值大,即相似度低。

[0138] 当模型整体损失不满足收敛条件时,表示模型未达到训练目标,此时,需要调整模型参数。

[0139] 在一种可以实现的方式中,可以按照公式(6)调整模型参数。

[0140] $$f_{target}=m*f_{target}+(1-m)*f_{online} \quad (6)$$

[0141] 其中,f_{target}为更新后的目标(target)参数,f_{online}为当前模型的在线(online)参数,m为调整系数。

[0142] 可以得到更新后的样本数据的独热编码如公式(7)所示。

[0143] $$\hat{y}(q,k)=\alpha S(q,k_m)+(1-\alpha)y(q,k) \quad (7)$$

[0144] 其中,k_m表示将样本标签输入动量更新的编码器后,得到的样本标签特征。

[0145] 然后,将公式(7)得到的独热编码代入公式(3),得到调整后的第一损失,然后进一步确定调整后的模型整体损失。

[0146] 当模型整体损失满足收敛条件时,表示模型达到训练目标,可以将训练后的初始模型确定为目标模型。

[0147] 上述实施例提供的技术方案至少带来以下有益效果:由S601-S602可知,标签排序设备确定第一损失和第二损失。然后,标签排序设备可以根据样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型。本公开提供了一种训练初始模型得到目标模型的具体实现方法,以使得目标模型可以达到训练目标,可以提高通过目标模型得到标签与资源的相似度的精准度。

[0148] 在一种可选的实施例中,当初始模型还包括判别模型时,损失集合,还包括判别模型的对抗损失。在图6示出的方法实施例的基础上,本实施例提供一种可能实现方式。结合图6,如图7所示,该标签排序方法,还包括:

[0149] S701、标签排序设备将样本数据输入判别模型,得到样本数据的数据类型的预测值。

[0150] 其中,数据类型用于表征样本数据为样本资源或者样本标签。对抗损失用于表征

数据类型的预测值与数据类型的真实值的差异。

[0151] 具体的,为了进一步减少资源和标签的差异,可以通过标签排序设备识别样本数据的数据类型。

[0152] 此时,需要引入判别模型的对抗损失,如公式(8)所示。

[0153] $L_{adv}=-(\log(\sigma(q_i))+\log(1-\sigma(k_i)))$ (8)

[0154] 其中,σ为数据判别模型的识别概率,L_{adv}为对抗损失。

[0155] 此时,模型整体损失如公式(9)所示。

[0156] $$\mathrm{Loss}=\min_{\theta}\max_{\delta}(L_{con}+L_{ce}-L_{adv}) \quad (9)$$

[0157] 上述实施例提供的技术方案至少带来以下有益效果:由S701可知,标签排序设备可以将样本数据输入判别模型,得到样本数据的数据类型的预测值。相较于人工标识数据类型,本公开可以进一步降低数据处理成本。

[0158] 在一种可选的实施例中,在图6或图7示出的方法实施例的基础上,本实施例提供一种可能实现方式。结合图6,如图8所示,S602中,标签排序设备根据样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型的方法,包括:

[0159] S801、标签排序设备将训练数据输入初始模型,确定软目标标签,以及模型整体损失。

[0160] 其中,训练数据包括样本数据。

[0161] 具体的,在模型训练过程中,需要进行多次迭代更新模型参数。每次迭代时,模型可以得到样本资源特征与样本标签特征的相似度的预测值。当相似度小于或者等于预设阈值时们可以认为该样本标签与样本资源的相关程度较低,对应的参考价值较低,此时,可以筛除该样本标签。因此,可以将多个样本标签中,对应的所述样本标签特征与样本资源特征的相似度大于预设阈值的标签确定为软目标标签,用于下一次对模型进行训练。

[0162] S802、当模型整体损失不满足收敛条件时,标签排序设备更新初始模型的模型参数,并将软目标标签和样本资源作为训练数据,输入更新后的初始模型中,直至更新后的初始模型对应的模型整体损失满足所述收敛条件。

[0163] 上述实施例提供的技术方案至少带来以下有益效果:由S801-S802可知,标签排序设备可以将训练数据输入初始模型,确定样本标签对应的软目标标签,以及模型整体损失。当模型整体损失不满足收敛条件时,标签排序设备可以更新初始模型的模型参数,并将软目标标签和样本资源作为训练数据,输入更新后的初始模型中,直至更新后的初始模型对应的模型整体损失满足所述收敛条件。本公开提供了一种训练初始模型得到目标模型的具体实现方法,可以筛除迭代过程中参考价值较低的样本标签特征,避免了算力资源的浪费,提高了模型训练效率。

[0164] 在一种可选的实施例中,当标签信息包括标签的维度信息和/或属性信息时,在图2示出的方法实施例的基础上,本实施例提供一种可能实现方式。结合图2,如图9所示,S203中,标签排序设备根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列的方法,包括:

[0165] S901、标签排序设备根据目标相似度,对多个标签进行排序,得到初始序列。

[0166] 在一种可以实现的方式中,标签排序设备可以将多个标签,按照对应的相似度从

大到小进行排序。这样一来,标签在标签序列越靠前,意味着与目标资源的资源内容的关联程度越高。

[0167] S902、标签排序设备根据标签信息和预设优先级策略,更新初始序列,得到标签序列。

[0168] 在一种可以实现的方式中,为了能够将标签粒度越细(即标签描述越详细)的标签靠前展示,标签排序设备可以根据标签的维度信息或者属性信息中的一种信息,对初始序列进行一次调整,得到标签序列。

[0169] 在另一种可以实现的方式中,为了进一步细化标签粒度,标签排序设备可以依次基于维度信息和属性信息对初始序列进行两次调整。本公开实施例对基于维度信息和属性信息的调整顺序不作限定。

[0170] 本公开实施例以先考虑标签的维度信息为例进行说明。

[0171] 标签排序设备可以先考虑标签的维度信息,例如主题标签、元素标签、知识产权(intellectual property,IP)标签、风格标签,形式标签等,基于类型特征的预先设定的重要程度对初始序列进行第一次调整。

[0172] 然后,在第一次调整之后,发现类型特征的标签还有不同的属性信息,例如,同样是人脸IP,有的根据人脸模型检测检索,有的来源图文校验人脸IP。因此,还可以根据粒度特征对第一次调整后的初始序列进行第二次调整。

[0173] 示例性的,如图10所示,标签的维度信息的优先级顺序可以包括:第一维度、第二维度、第三维度和第四维度。其中,第一维度包括:兴趣点(point of interest,POI)等。第二维度包括:主题标签、形式标签、类目、商品标签等。第三维度包括:明星IP、网红IP、名人IP等。第四维度包括:影视综IP、菜品IP、音乐IP、动漫IP、游戏IP等。

[0174] 一个美食视频的标签可以包括:主题标签为“美食分享”、形式标签为“烹饪教学”、类目为“美食分享”、网红IP为“网红A”、菜品IP为“烤乳猪”。

[0175] 可以理解的,在实际实施时,本公开实施例所述的标签排序装置可以包含有用于实现前述对应标签排序方法的一个或多个硬件结构和/或软件模块,这些执行硬件结构和/或软件模块可以构成一个电子设备。本领域技术人员应该很容易意识到,结合本文中所公开的实施例描述的各示例的算法步骤,本公开能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本公开的范围。

[0176] 基于这样的理解,本公开实施例还对应提供一种标签排序装置。图11示出了本公开实施例提供的标签排序装置的结构示意图一。如图11所示,该标签排序装置包括:获取单元1101和第一处理单元1102。

[0177] 获取单元1101,被配置为执行获取目标资源,以及目标资源的多个标签。

[0178] 第一处理单元1102,被配置为执行将目标资源和多个标签输入预先训练好的目标模型中,得到每个标签与目标资源的目标相似度;目标模型是根据样本数据训练初始模型得到的;样本数据包括:样本资源,以及与样本资源对应的多个样本标签。

[0179] 第一处理单元1102,还被配置为执行根据目标相似度和多个标签的标签信息,对多个标签进行排序,得到标签序列。

[0180] 可选的,初始模型包括第一特征提取模型、第二特征提取模型;该标签排序装置,还包括:第二处理单元1103。

[0181] 第二处理单元1103,被配置为执行获取样本数据;将样本资源输入第一特征提取模型,得到样本资源的样本资源特征;将每个样本标签输入第二特征提取模型,得到每个样本标签的样本标签特征;当第二特征提取模型为多层神经网络模型时,样本标签特征包括多层神经网络模型中的多个隐含层数据。

[0182] 可选的,每个样本标签包括:多个字符;第二特征提取模型包括:BERT模型;第二处理单元1103,还被配置为执行对输入的每个样本标签的至少一个字符进行遮蔽,得到每个样本标签对应的遮蔽训练数据;遮蔽训练数据包括:每个样本标签中,与至少一个字符一一对应的至少一个遮蔽位置;将遮蔽训练数据输入BERT模型,得到遮蔽位置的字符的预测值。

[0183] 可选的,第二处理单元1103,还被配置为执行获取第一损失和第二损失;第一损失用于表征样本资源特征与样本标签特征的差异;第二对比损失用于表征字符的预测值与字符的真实值的差异;根据样本数据训练初始模型,直至模型整体损失满足收敛条件,得到目标模型;模型整体损失是根据损失集合得到的;损失集合包括:第一损失和第二损失。

[0184] 可选的,初始模型还包括:判别模型;损失集合,还包括:判别模型的对抗损失;第二处理单元1103,还被配置为执行将样本数据输入判别模型,得到样本数据的数据类型的预测值;数据类型用于表征样本数据为样本资源或者样本标签;对抗损失用于表征数据类型的预测值与数据类型的真实值的差异。

[0185] 可选的,第二处理单元1103,具体被配置为执行将训练数据输入初始模型,确定软目标标签,以及模型整体损失;训练数据包括样本数据;软目标标签包括:多个样本标签特征中,对应的样本标签特征与样本资源特征的相关度大于预设阈值的标签;当模型整体损失不满足收敛条件时,更新初始模型的模型参数,并将软目标标签和样本资源作为训练数据,输入更新后的初始模型中,直至更新后的初始模型对应的模型整体损失满足收敛条件。

[0186] 可选的,标签信息包括:标签的维度信息和/或属性信息;维度信息用于表征标签的类型特征;属性信息用于表征标签的数据来源、出现位置、相互关系中的至少一种;第一处理单元1102,具体被配置为执行根据目标相似度,对多个标签进行排序,得到初始序列;根据标签信息和预设优先级策略,更新初始序列,得到标签序列。

[0187] 如上所述,本公开实施例可以根据上述方法示例对存储设备进行功能模块的划分。其中,上述集成的模块既可以采用硬件的形式实现,也可以采用软件功能模块的形式实现。另外,还需要说明的是,本公开实施例中对模块的划分是示意性的,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式。例如,可以对应各个功能划分各个功能模块,也可以将两个或两个以上的功能集成在一个处理模块中。

[0188] 关于上述实施例中的标签排序装置,其中各个模块执行操作的具体方式、以及具备的有益效果,均已经在前述方法实施例中进行了详细描述,此处不再赘述。

[0189] 本公开实施例还提供一种标签排序装置。图12示出了本公开实施例提供的标签排序装置的结构示意图二。该标签排序装置可以包括至少一个处理器221,通信总线222,存储器223以及至少一个通信接口224。

[0190] 处理器221可以是一个中央处理器(central processing units,CPU),微处理单元,ASIC,或一个或多个用于控制本公开方案程序执行的集成电路。

[0191] 通信总线222可包括一通路,在上述组件之间传送信息。

[0192] 通信接口224,使用任何收发器一类的装置,用于与其他设备或通信网络通信,如电子设备、以太网,无线接入网(radio access network,RAN),无线局域网(wireless local area networks,WLAN)等。

[0193] 存储器223可以是只读存储器(read-only memory,ROM)或可存储静态信息和指令的其他类型的静态存储设备,随机存取存储器(random access memory,RAM)或者可存储信息和指令的其他类型的动态存储设备,也可以是电可擦可编程只读存储器(electrically erasable programmable read-only memory,EEPROM)、只读光盘(compact disc read-only memory,CD-ROM)或其他光盘存储、光碟存储(包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等)、磁盘存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质,但不限于此。存储器可以是独立存在,通过总线与处理单元相连接。存储器也可以和处理单元集成在一起。

[0194] 其中,存储器223用于存储执行本公开方案的应用程序代码,并由处理器221来控制执行。处理器221用于执行存储器223中存储的应用程序代码,从而实现本公开方法中的功能。

[0195] 在具体实现中,作为一种实施例,处理器221可以包括一个或多个中央处理器,例如图12中的中央处理器0和中央处理器1。

[0196] 在具体实现中,作为一种实施例,标签排序装置可以包括多个处理器,例如图12中的处理器221和处理器225。这些处理器中的每一个可以是一个单核(single-CPU)处理器,也可以是一个多核(multi-CPU)处理器。这里的处理器可以指一个或多个设备、电路、和/或用于处理数据(例如计算机程序指令)的处理核。

[0197] 在具体实现中,作为一种实施例,标签排序装置还可以包括输入设备226和输出设备227。输入设备226和输出设备227通信,可以以多种方式接受用户的输入。例如,输入设备226可以是鼠标、键盘、触摸屏设备或传感设备等。输出设备227和处理器221通信,可以以多种方式来显示信息。例如,输出设备227可以是液晶显示器(liquid crystal display,LCD),发光二级管(light emitting diode,LED)显示设备等。

[0198] 本领域技术人员可以理解,图12中示出的结构并不构成对标签排序装置的限定,可以包括比图示更多或更少的组件,或者组合某些组件,或者采用不同的组件布置。

[0199] 本公开还提供了一种包括指令的计算机可读存储介质,所述计算机可读存储介质上存储有指令,当所述计算机可读存储介质中的指令由标签排序装置的处理器执行时,使得标签排序装置能够执行上述所示实施例提供的标签排序方法。例如,计算机可读存储介质可以为包括指令的存储器223,上述指令可由标签排序装置的处理器221执行以完成上述方法。

[0200] 可选的,计算机可读存储介质可以是非临时性计算机可读存储介质,例如,所述非临时性计算机可读存储介质可以是ROM、RAM、CD-ROM、磁带、软盘和光数据存储设备等。

[0201] 本公开还提供了一种包含指令的计算机程序产品,当其在电子设备上运行时,使得电子设备执行上述所示实施例提供的标签排序方法。

[0202] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本公开的其

它实施方案。本公开旨在涵盖本公开的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本公开的真正范围和精神由下面的权利要求指出。

[0203] 应当理解的是，本公开并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求来限制。

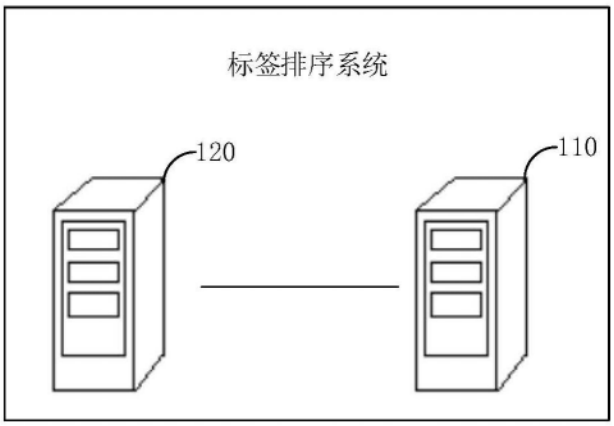

标签排序系统
120
110

图1

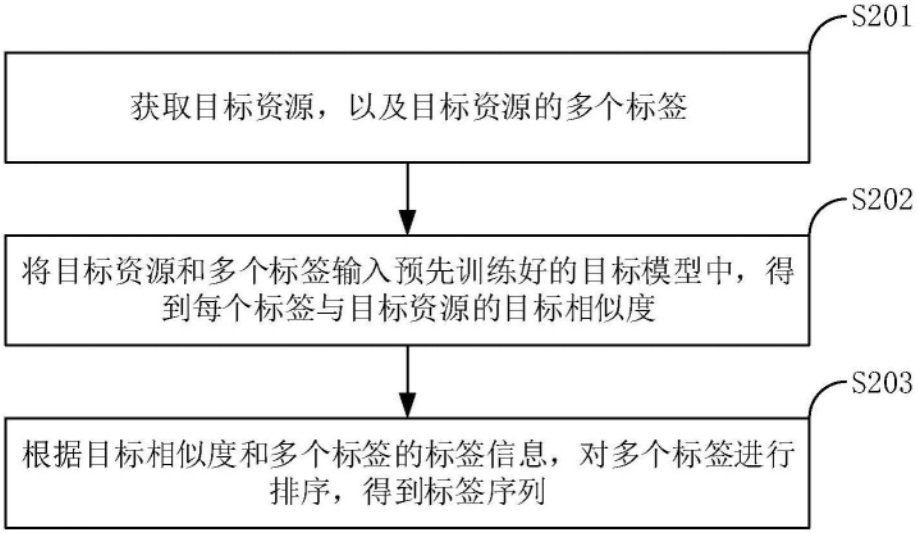

S201
获取目标资源，以及目标资源的多个标签
S202
将目标资源和多个标签输入预先训练好的目标模型中，得到每个标签与目标资源的目标相似度
S203
根据目标相似度和多个标签的标签信息，对多个标签进行排序，得到标签序列

图2

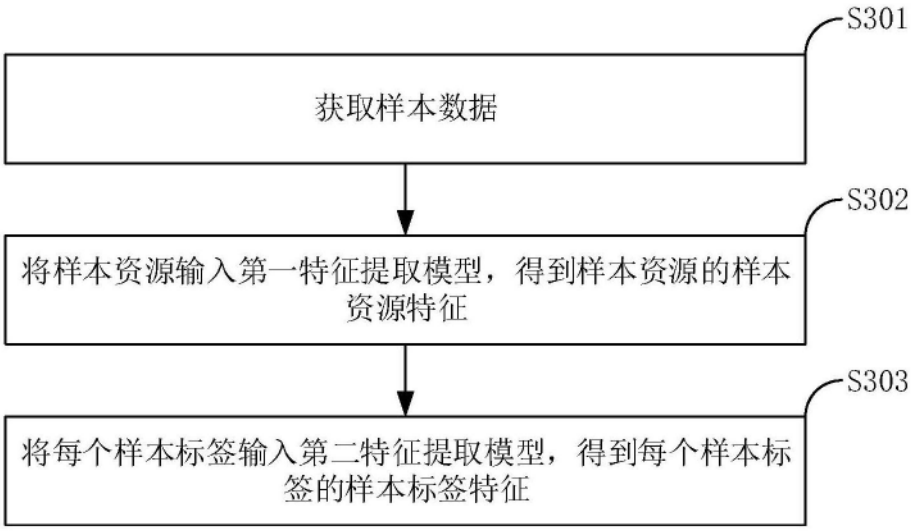

S301
获取样本数据
S302
将样本资源输入第一特征提取模型，得到样本资源的样本资源特征
S303
将每个样本标签输入第二特征提取模型，得到每个样本标签的样本标签特征

图3

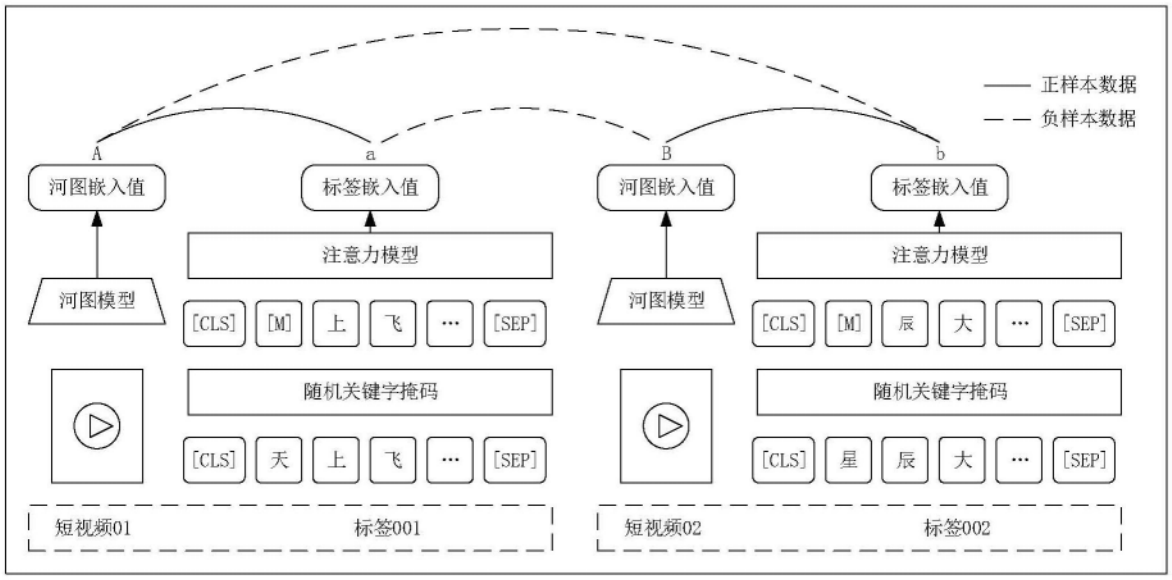

正样本数据
负样本数据
A
a
B
b
河图嵌入值
标签嵌入值
河图嵌入值
标签嵌入值
注意力模型
注意力模型
河图模型
河图模型
[CLS] [M] 上 飞 … [SEP]
[CLS] [M] 辰 大 … [SEP]
随机关键字掩码
随机关键字掩码
[CLS] 天 上 飞 … [SEP]
[CLS] 星 辰 大 … [SEP]
短视频01
标签001
短视频02
标签002

图4

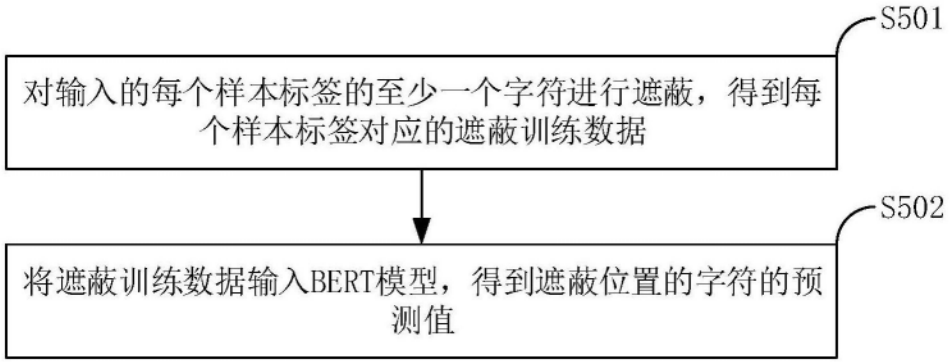

S501
对输入的每个样本标签的至少一个字符进行遮蔽，得到每个样本标签对应的遮蔽训练数据
S502
将遮蔽训练数据输入BERT模型，得到遮蔽位置的字符的预测值

图5

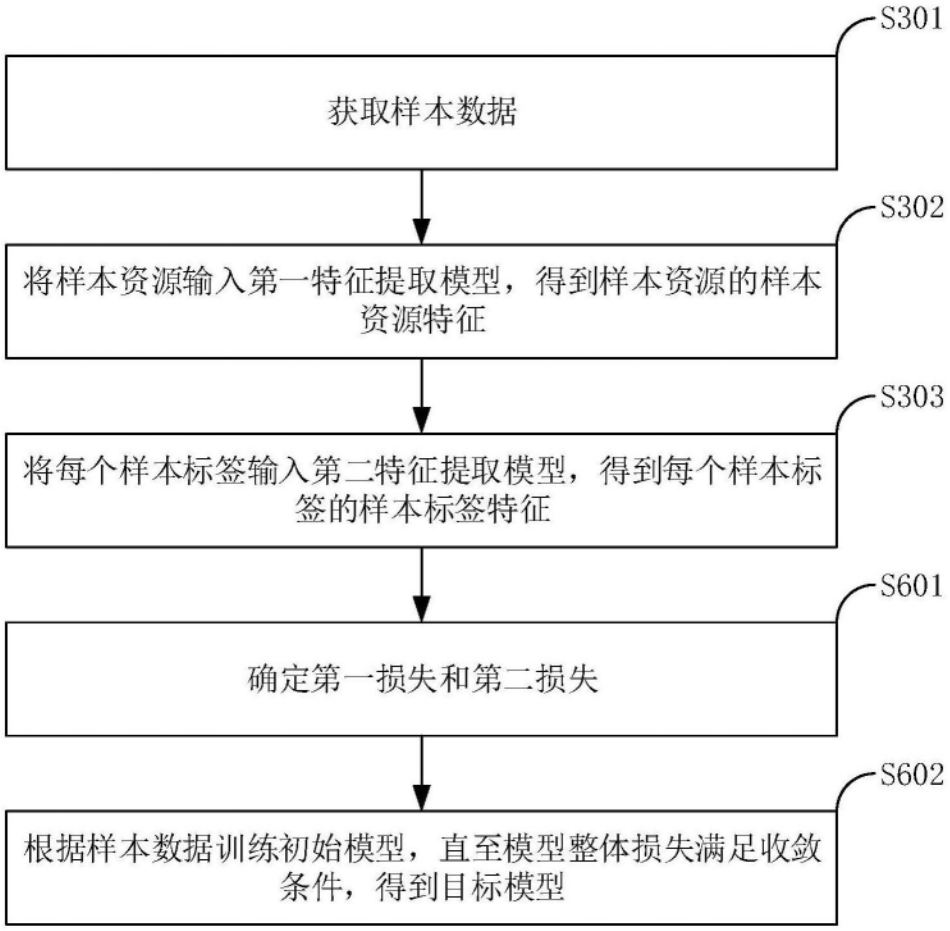

S301
获取样本数据
S302
将样本资源输入第一特征提取模型，得到样本资源的样本资源特征
S303
将每个样本标签输入第二特征提取模型，得到每个样本标签的样本标签特征
S601
确定第一损失和第二损失
S602
根据样本数据训练初始模型，直至模型整体损失满足收敛条件，得到目标模型

图6

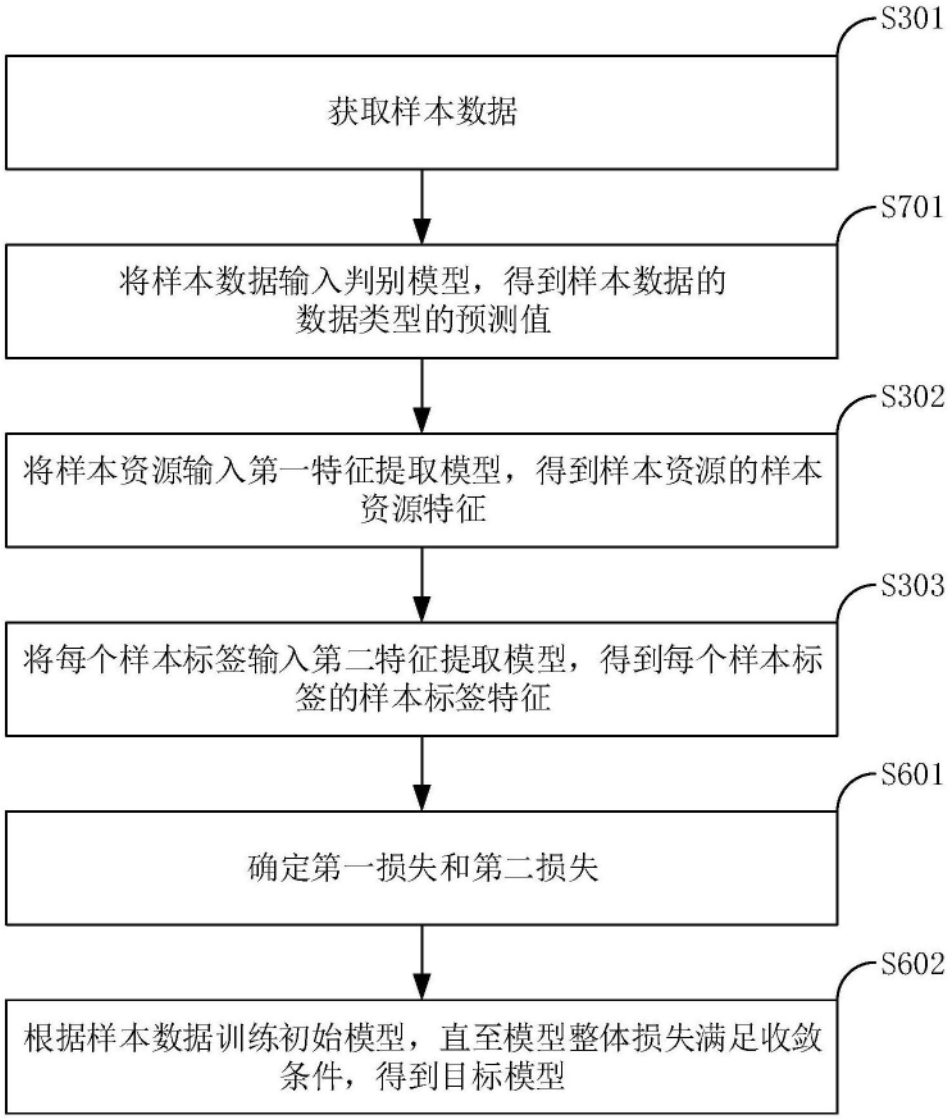

S301
获取样本数据
S701
将样本数据输入判别模型，得到样本数据的数据类型的预测值
S302
将样本资源输入第一特征提取模型，得到样本资源的样本资源特征
S303
将每个样本标签输入第二特征提取模型，得到每个样本标签的样本标签特征
S601
确定第一损失和第二损失
S602
根据样本数据训练初始模型，直至模型整体损失满足收敛条件，得到目标模型

图7

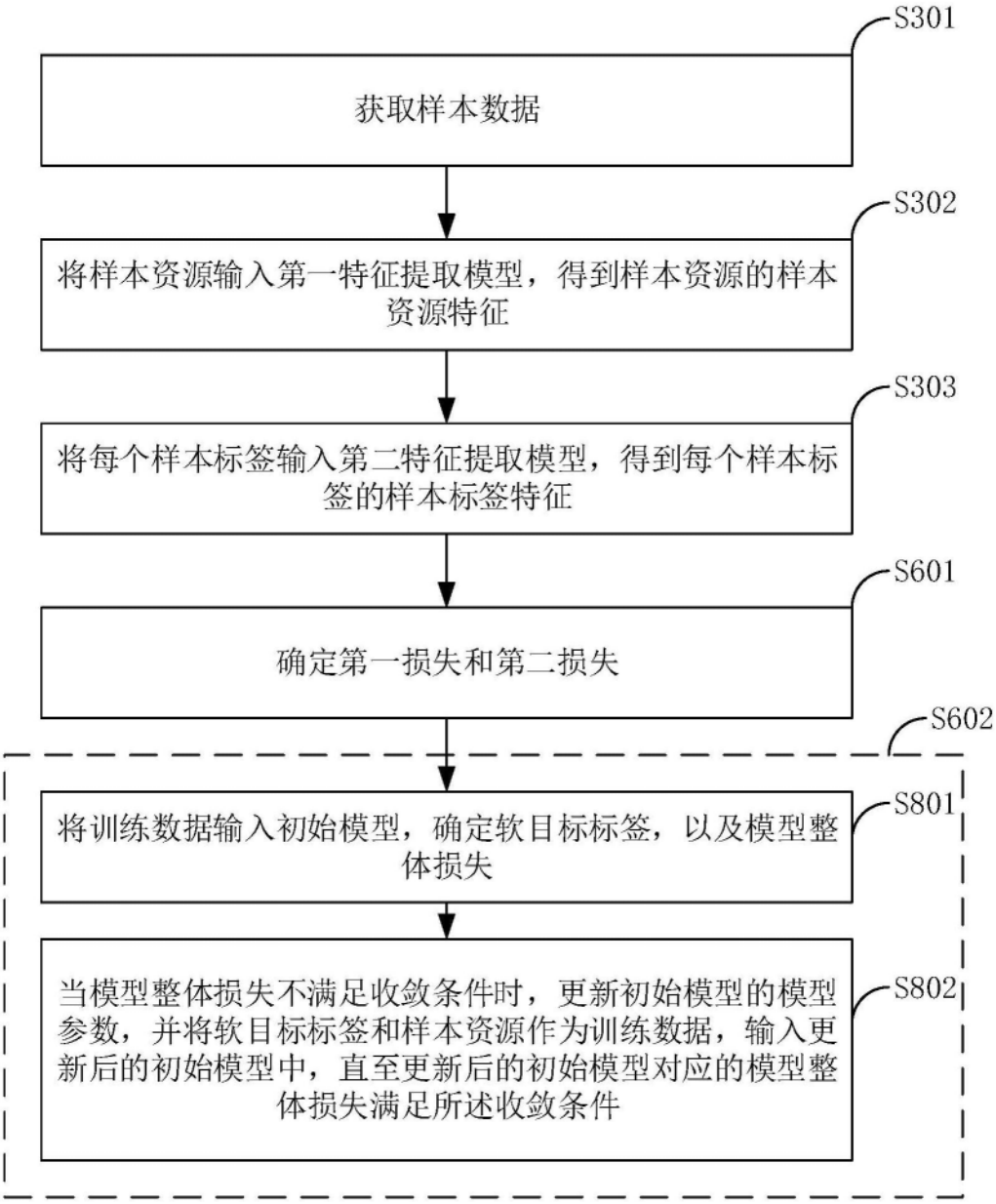

S301
获取样本数据
S302
将样本资源输入第一特征提取模型，得到样本资源的样本资源特征
S303
将每个样本标签输入第二特征提取模型，得到每个样本标签的样本标签特征
S601
确定第一损失和第二损失
S602
S801
将训练数据输入初始模型，确定软目标标签，以及模型整体损失
S802
当模型整体损失不满足收敛条件时，更新初始模型的模型参数，并将软目标标签和样本资源作为训练数据，输入更新后的初始模型中，直至更新后的初始模型对应的模型整体损失满足所述收敛条件

图8

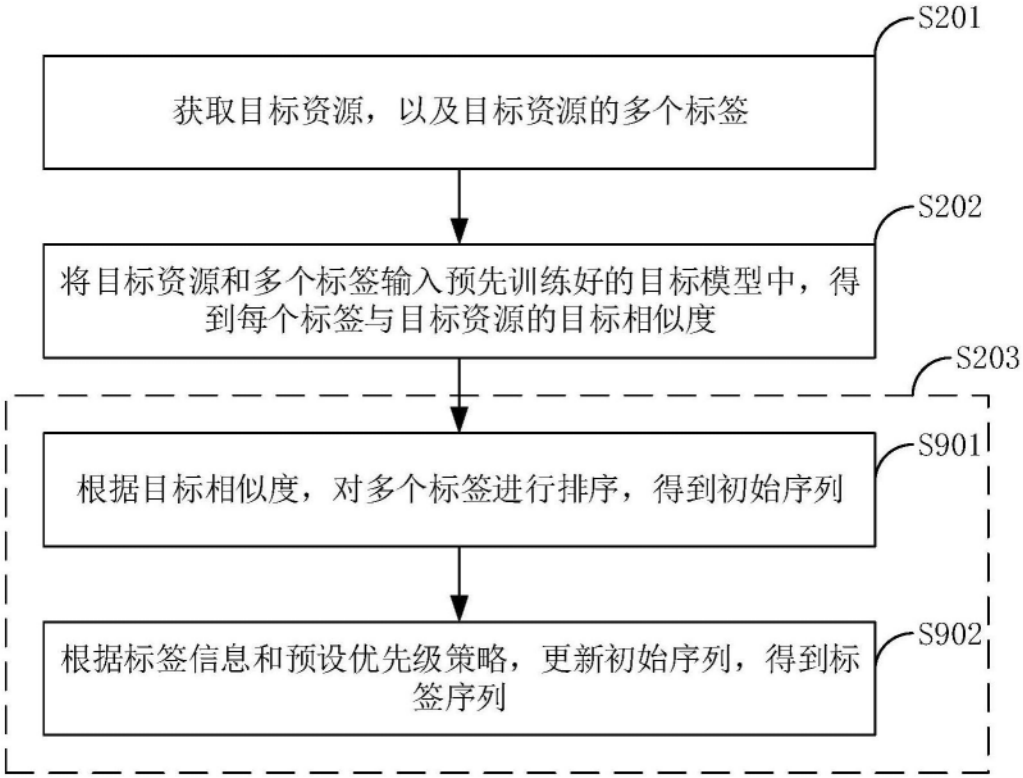

S201
获取目标资源，以及目标资源的多个标签
S202
将目标资源和多个标签输入预先训练好的目标模型中，得到每个标签与目标资源的目标相似度
S203
S901
根据目标相似度，对多个标签进行排序，得到初始序列
S902
根据标签信息和预设优先级策略，更新初始序列，得到标签序列

图9

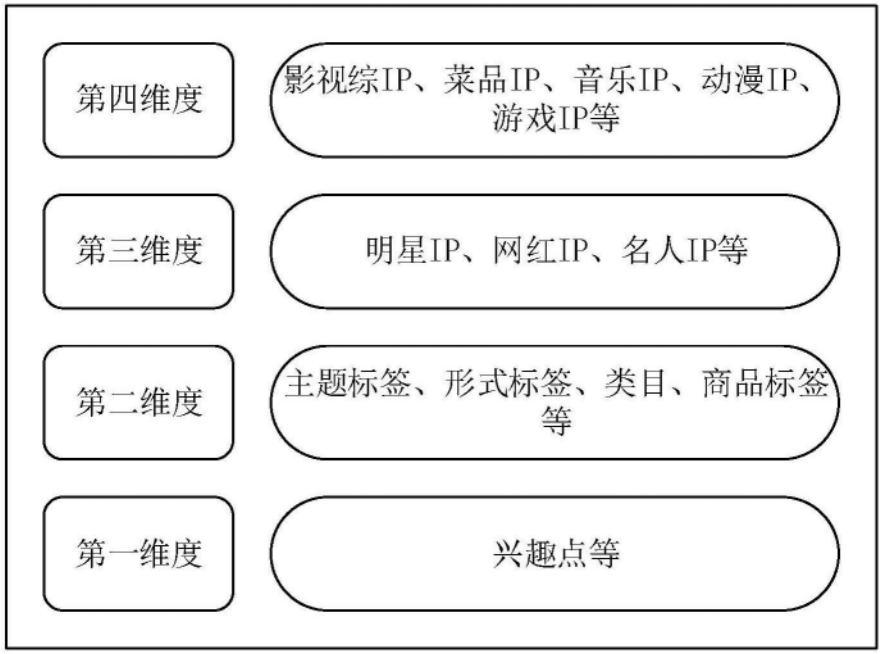

第四维度
影视综IP、菜品IP、音乐IP、动漫IP、游戏IP等
第三维度
明星IP、网红IP、名人IP等
第二维度
主题标签、形式标签、类目、商品标签等
第一维度
兴趣点等

图10

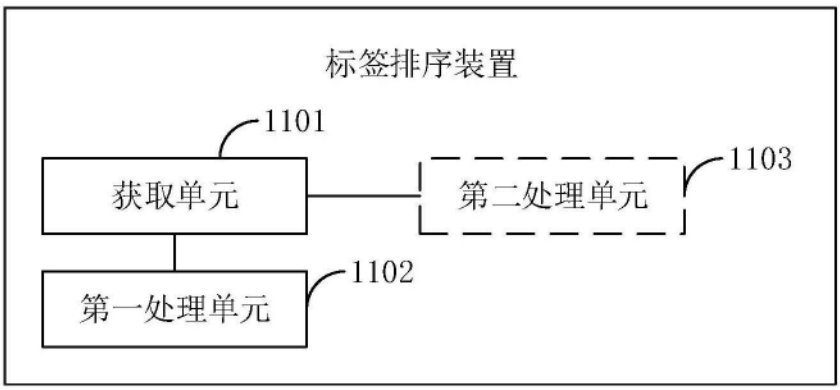

标签排序装置
1101
获取单元
第二处理单元
1103
第一处理单元
1102

图11

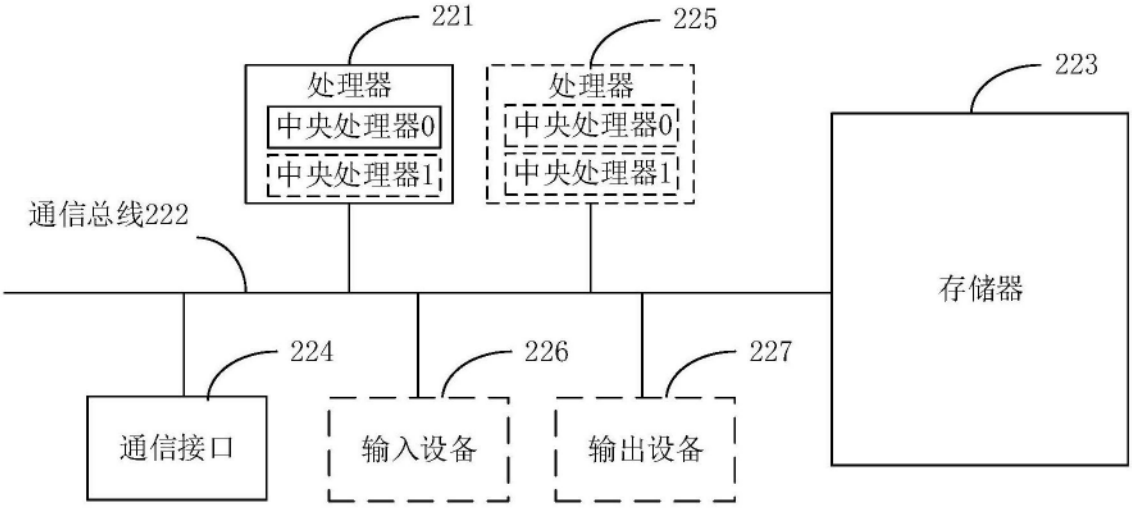

221
225
223
处理器
中央处理器0
中央处理器1
处理器
中央处理器0
中央处理器1
通信总线222
存储器
224
226
227
通信接口
输入设备
输出设备

图12