(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118673180 B
(45)授权公告日 2024.10.18

(21)申请号 202411163349.7
(22)申请日 2024.08.23
(65)同一申请的已公布的文献号
申请公布号 CN 118673180 A
(43)申请公布日 2024.09.20
(73)专利权人 成都华栖云科技有限公司
地址 610041 四川省成都市高新区天府五街200号4号楼A区9楼
(72)发明人 颜涛 余军 朱愚 黄信云 刘让刚
(74)专利代理机构 成都乐易联创专利代理有限公司 51269
专利代理师 赵何婷
(51)Int.Cl.
G06F 16/78(2019.01)
G06F 16/783(2019.01)
G06N 3/0464(2023.01)
(56)对比文件
CN 118093938 A,2024.05.28
US 2013283301 A1,2013.10.24
审查员 何洋

权利要求书1页 说明书4页 附图3页

(54)发明名称

基于标签检索和多模态向量的视频内容检索方法

(57)摘要

本发明属于视频检索技术领域,公开了一种基于标签检索和多模态向量的视频内容检索方法,该方法包括如下步骤:步骤100:检索前的视频处理;视频处理是指提取视频内容中的标签、语义向量和文字描述;步骤200:根据用户输入的检索式进行检索结果排序并展示;所述排序先采用语义向量排序,然后采用ES得分和多模态向量得分加权平均排序,最后引入标签命中个数的二次排序以获取最终结果,本发明实现了对于视频内容的多维度混合检索,降低单一模型在特殊数据上弊端,同时通过多维度数据的融合检索,极大地提升视频搜索的准确度;此方法能应用到各种视频数据资产管理中,对于视频数据资产的二次使用有很高的经济价值。

CN 118673180 B

1.一种基于标签检索和多模态向量的视频内容检索方法,其特征在于,包括:

检索前的视频处理;视频处理是指提取视频内容中的标签、语义向量和文字描述;

根据用户输入的检索式进行检索结果排序并展示;

其中,所述视频处理是指提取视频内容中的标签、语义向量和文字描述包括如下步骤:

步骤101:采用CV算法对视频内容进行转场切割为数个片段序列;提取每个片段序列中的画面作为片段表征画面;

步骤102:采用AI标签提取所述片段表征画面中的标签;

步骤103:采用CNN卷积网络对所述片段表征画面进行分类,分为正常特征片段表征画面和非正常特征片段表征画面;

对非正常特征片段表征画面进行处理:采用视觉标签模型、画面分割模型、图文问答模型对所述非正常特征片段表征画面的标签、转场分割、文字描述进行综合判定;

综合判定是指再次判定片段表征画面是否存在标签、是否能够进行转场分割以及是否具有文字描述,若无标签和转场分割的片段表征画面确定为无效片段表征画面;反之,则采用图文问答模型提取的文字描述作为文字描述或者以“无具体含义”作为文字描述;

再采用多模态大模型提取语义特征作为语义向量;

步骤104:对正常特征片段表征画面的处理;采用多模态大模型和图文问答模型提取正常特征片段表征画面中的语义向量和文字描述;

根据用户输入的检索式进行检索结果排序是按照如下步骤进行:

步骤201:采用多模态大模型提取用户输入的检索式的语义向量;将检索式的语义向量与视频内容的语义向量进行匹配获取多模态语义向量检索得分,按照多模态语义向量检索得分从高到低进行排序并获取前N个视频内容的数据列表vec_ret;

步骤202:对用户输入的检索式进行实体分词;采用ES搜索引擎对实体分词在数据列表vec_ret中进行搜索并获取ES得分,并利用对数函数f(x)对ES得分进行归一化处理后按照从高到低进行重新排序并获取数据列表tag_ret;

当分词属于人物时且与视频内容的标签对应上,以及当分词对应上人工标签时,给ES得分赋予大于1的权重;

步骤203:将步骤201获取的数据列表vec_ret的多模态语义向量检索得分和步骤202数据列表tag_ret的ES得分平均加权后重新排序获得前M个视频内容的数据列表m_ret,M＜N;

步骤204:根据用户输入的检索式的实体分词对数据列表m_ret中标签命中个数进行统计,以命中个数为优先级进行排序,即有命中个数越多排序越靠前,当命中个数相同时以平均加权得分排序;输出最终排序结果。

2.根据权利要求1所述的基于标签检索和多模态向量的视频内容检索方法,其特征在于,所述多模态语义向量检索得分通过余弦相似度获得。

3.根据权利要求1所述的基于标签检索和多模态向量的视频内容检索方法,其特征在于,步骤202中所述ES得分取多个分词的ES得分平均值。

基于标签检索和多模态向量的视频内容检索方法

技术领域

[0001] 本发明属于视频检索技术领域,具体涉及一种基于标签检索和多模态向量的视频内容检索方法。

背景技术

[0002] 随着互联网技术的不断发展,在网络带宽不断增长的今天,网络视频以其便捷的访问体验、多样化的影片来源以及实时的更新速度吸引了广大的用户,使得网络视频成为了用户网络生活中不可或缺的重要组成部分。随着网络中的存在的各类视频的海量增长,视频用户往往通过视频检索的方式来获取感兴趣的视频内容。

[0003] 目前,视频内容检索技术的发展分为几个不同的阶段,从基于文本的传统视频检索到基于AI标签的跨模态视频检索,再到如今基于大模型的自然语言视频检索;这些技术的进步不仅提高了搜索效率和准确性,而且极大地改善了用户体验。其中,基于多模态特征向量检索方法是通过对视频画面进行多模态的特征分析,利用多模态预训练神经网络模型(CLIP)等多模态大模型提取特征进行向量检索,此类方法可以解决场景语义特征问题,已经广泛应用于图像检索的场景;但是该方法的问题是误检率比较高,业内比较高的模型准确率也仅超过90%。

发明内容

[0004] 本发明的目的在于提供一种基于标签检索和多模态向量的视频内容检索方法,该方法将标签检索和多模态特征向量检索结合实现对视频内容精准检索。

[0005] 为实现上述目的,本发明采用如下技术方案:

[0006] 一种基于标签检索和多模态向量的视频内容检索方法,包括:

[0007] 检索前的视频处理;视频处理是指提取视频内容中的标签、语义向量和文字描述;

[0008] 根据用户输入的检索式进行检索结果排序并展示;

[0009] 其中,所述根据用户输入的检索式进行检索结果排序是按照如下步骤进行:步骤201:采用多模态大模型提取用户输入的检索式的语义向量;将检索式的语义向量与视频内容的语义向量进行匹配获取多模态语义向量检索得分,按照多模态语义向量检索得分从高到低进行排序并获取前N个视频内容的数据列表vec_ret;

[0010] 步骤202:对用户输入的检索式进行实体分词;采用ES搜索引擎对实体分词在数据列表vec_ret中进行搜索并获取ES得分,并利用对数函数f(x)对ES得分进行归一化处理后按照从高到低进行重新排序并获取数据列表tag_ret;

[0011] 步骤203:将步骤201获取的数据列表vec_ret的多模态语义向量检索得分和步骤202数据列表tag_ret的ES得分平均加权后重新排序获得前M(M<N)个视频内容的数据列表m_ret;

[0012] 步骤204:根据用户输入的检索式的实体分词对数据列表m_ret中标签命中个数进行统计,以命中个数为优先级进行排序,即有命中个数越多排序越靠前,当命中个数相同时

以平均加权得分排序;输出最终排序结果。

[0013] 进一步地,所述视频处理是指提取视频内容中的标签、语义向量和文字描述包括如下步骤:

[0014] 步骤101:采用CV算法对视频内容进行转场切割为数个片段序列;提取每个片段序列中的画面作为片段表征画面;

[0015] 步骤102:采用AI标签提取所述片段表征画面中的标签;

[0016] 步骤103:采用CNN卷积网络对所述片段表征画面进行过滤获得正常特征片段表征画面;

[0017] 步骤104:对正常特征片段表征画面的处理;采用多模态大模型和图文问答模型提取正常特征片段表征画面中的语义向量和文字描述。

[0018] 进一步地,对步骤103过滤的非正常特征片段表征画面进行处理:

[0019] 采用视觉标签模型、画面分割模型、图文问答模型对所述非正常特征片段表征画面的标签、转场分割、文字描述进行综合判定;

[0020] 综合判定是指再次判定片段表征画面是否存在标签、是否能够进行转场分割以及是否具有文字描述,若无标签和转场分割的片段表征画面确定为无效片段表征画面确;反之,则采用图文问答模型提取的文字描述作为文字描述或者以“无具体含义”作为文字描述;

[0021] 再采用多模态大模型提取语义特征作为语义向量。

[0022] 进一步地,所述多模态语义向量检索得分通过余弦相似度获得。

[0023] 进一步地,步骤202中所述ES得分取多个分词的ES得分平均值。

[0024] 进一步地,步骤202中当分词属于人物时且与视频内容的标签对应上,以及当分词对应上人工标签时,给ES得分赋予大于1的权重。

[0025] 与现有技术相比,本发明具有如下有益效果:

[0026] (1)在画面语义在进行特征提取前,使用CNN卷积网络对画面进行分类,能快速提取出画面语义描述;然后再次检测被分类为异常画面的界面,避免误分类造成的数据丢失,降低了多模态的误检率;

[0027] (2)因EC检索得分受到时间、检索频率等影响造成了其结果的精准度低,本发明在使用其自评分机制的同时利用数据平均、数值归一化将ES检索得分和多模态余弦相似度得分统一到相同的数据区间,加权平均降低其误差,并且通过人物、人工标签等命中率高的因子加权,提高其精度;

[0028] (3)通过标签命中数量和最终平均加权得分进行二次排序,进一步将命中率高的视频内容排在前面,能够极大地提高视频内容检索结果的准确率。

附图说明

[0029] 图1为本发明的总体流程图。

[0030] 图2为本发明检索前视频处理流程图。

[0031] 图3为本发明检索时处理流程图。

具体实施方式

[0032] 如图1所示,本实施例提供的一种基于标签检索和多模态向量的视频内容检索方法,包括如下步骤:

[0033] 检索前的视频处理;视频处理主要是指提出视频内容中的数据组织形态,数据组织形态为标签、语义向量、文字描述。

[0034] 如图2所示,所述获取视频内容中的数据组织形态具体包括如下步骤:

[0035] 步骤101:采用CV算法对视频内容进行转场切割为数个片段序列;提取每个片段序列中的画面作为片段表征画面;

[0036] 步骤102:采用AI标签提取所述片段表征画面中的标签;

[0037] 步骤103:采用CNN卷积网络对所述片段表征画面进行分类,分为正常特征片段表征画面和非正常特征片段表征画面,其中所述非正常特征片段表征画面是指纯色画面和高度模糊画面的片段表征画面,其余则为正常画面;

[0038] 步骤104:对正常特征片段表征画面的处理;采用多模态大模型和图文问答模型提取正常特征片段表征画面中的语义向量和文字描述;

[0039] 步骤105:对非正常特征片段表征画面的处理,将非正常特征表征画面中有效的画面提取出来,CNN卷积网络过滤掉过多的有效画面,例如纯色背景的有效画面;

[0040] 采用视觉标签模型、画面分割模型、图文问答模型对所述非正常特征片段表征画面的标签、转场分割、文字描述进行综合判定,综合判定是指再次判定片段表征画面是否存在标签、是否能够进行转场分割以及是否具有文字描述,若无标签和转场分割以及无文字描述的片段表征画面确定为无效片段表征画面确;反之,则采用图文问答模型提取的文字描述作为文字描述或者以“无具体含义”作为文字描述;再采用多模态大模型提取语义特征作为语义向量。

[0041] 将提取的视频内容的标签、文字描述和语义向量关联后存储数据库中,本申请通过多种模型实现对视频内容进行处理获取视频内容的标签、文字描述和语义向量,提高后续内容检索的可控度和准确度。

[0042] 检索时的视频内容排序,本实施例通过单排、混排等综合排序,提高检索结果的命中率;如图3所示,具体地包括如下步骤:

[0043] 步骤201:采用多模态大模型提取用户输入的检索式的语义向量;将检索式的语义向量与数据库中视频内容的语义向量进行匹配获取多模态语义向量检索得分,按照多模态语义向量检索得分从高到低进行排序并获取前N个视频内容的数据列表vec_ret;其中,所述多模态语义向量检索得分通过余弦相似度获得。

[0044] 步骤202:对用户输入的检索式进行实体分词;采用ES搜索引擎对分词在数据列表vec_ret中进行搜索并获取ES得分,并利用对数函数f(x)对ES得分进行归一化处理后按照从高到低进行重新排序并获取数据列表tag_ret;

[0045] 因用户输入的检索式分词结果一般都是大于1的数量,即会有两个及两个以上分词,因此针对多个分词的ES得分,本实施例最终的ES得分取多个分词的ES得分平均值。

[0046] 另外,当分词属于人物时且与视频内容的标签对应上,以及当分词对应上人工标签时,本实施例给最终ES得分赋予大于1的权重,提高命中率。人物的指代性更强,其检索的结果精度更高,人工标签属于人为手工标记的标签,人工标签的准确度高于AI标签,因此当

分词属于上述两种情况使,将提高其ES得分。

[0047] 因多模态语义向量检索得分的区间在[0.21,0.35],而ES得分在[3,25],为使ES得分与多模态语义向量检索得分能够在相同区间范围,本申请通过对数函数f(x)对ES得分进行处理,当x在[3,35]区间,使其函数值趋近于[0.21,0.35];当x在[1,3]区间,使其函数值趋近于[0.15,0.21];当x大于35,则为0.35。

[0048] 步骤203:将步骤201获取的数据列表vec_ret的多模态语义向量检索得分和步骤202数据列表tag_ret的ES得分平均加权后重新排序获得前M(M＜N)个视频内容的数据列表m_ret;

[0049] 步骤204:根据用户输入的检索式的实体分词对数据列表m_ret中标签命中个数进行统计,以命中个数为优先级进行排序,即有命中个数越多排序越靠前,当命中个数相同时以平均加权得分排序;输出最终排序结果。

[0050] 本实施例实现了对于视频内容的多维度混合检索,降低单一模型在特殊数据上弊端,同时通过多维度数据的融合检索,极大地提升视频搜索的准确度;此方法能应用到各种视频数据资产管理中,对于视频数据资产的二次使用有很高的经济价值。

[0051] 以上所述仅是本发明优选的实施方式,但本发明的保护范围并不局限于此,任何基于本发明所提供的技术方案和发明构思进行的改造和替换都应涵盖在本发明的保护范围内。

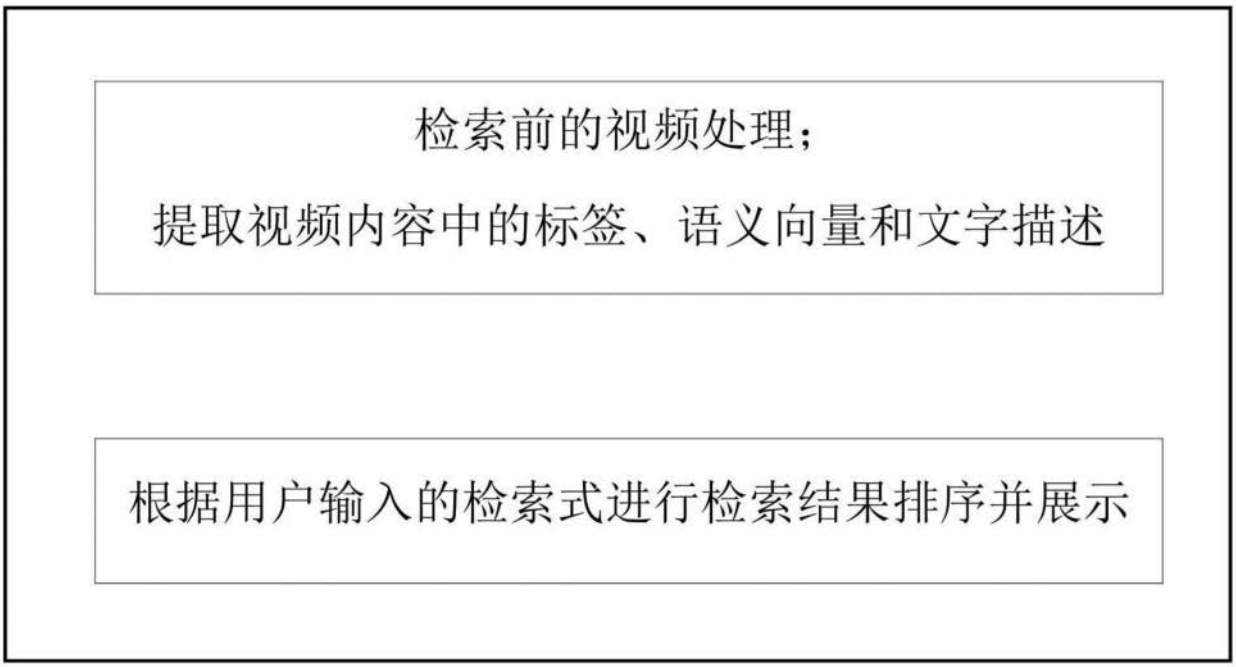

图1

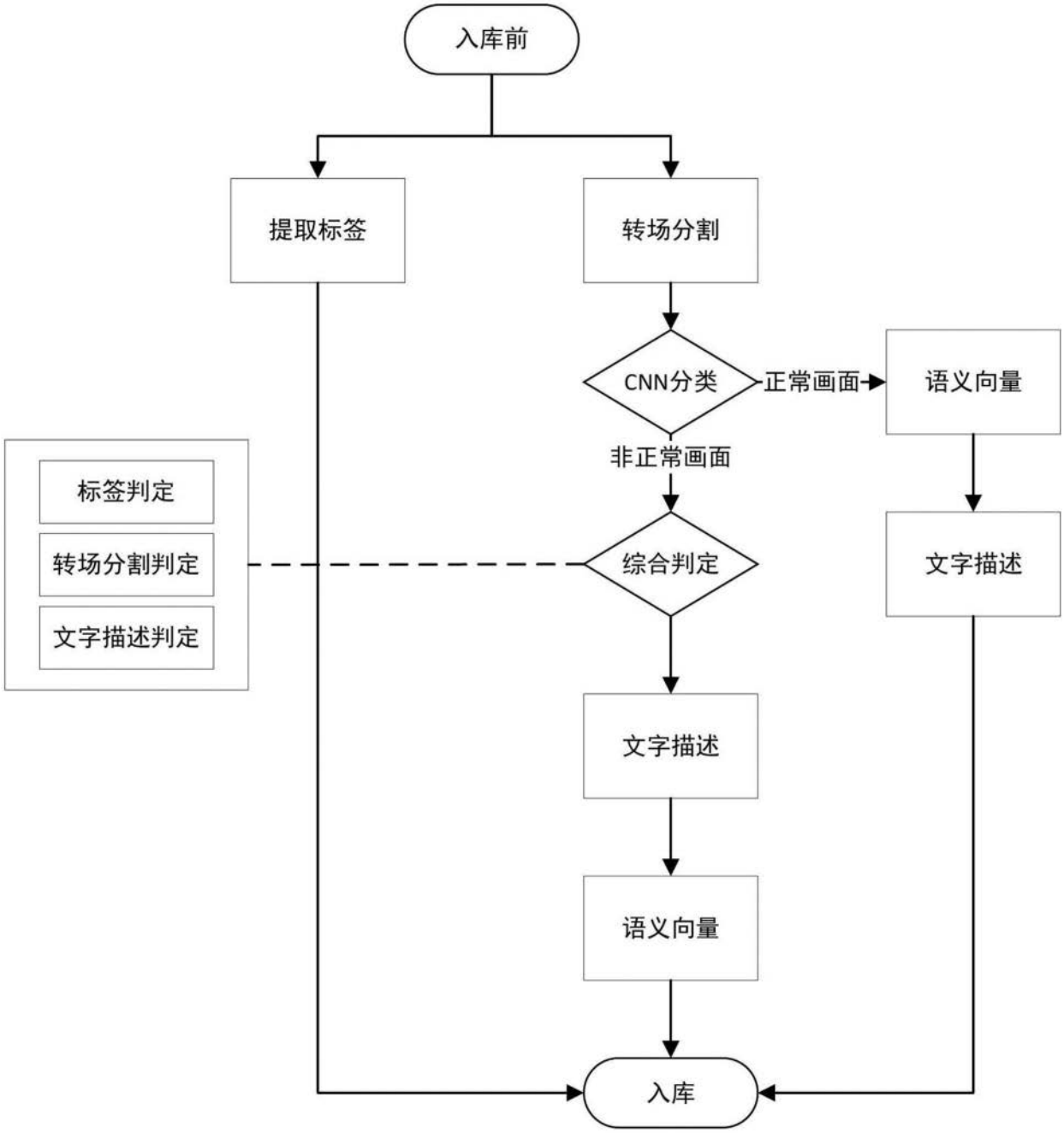
入库前
提取标签
转场分割
CNN分类
正常画面
语义向量
非正常画面
标签判定
转场分割判定
文字描述判定
综合判定
文字描述
文字描述
语义向量
入库

图2

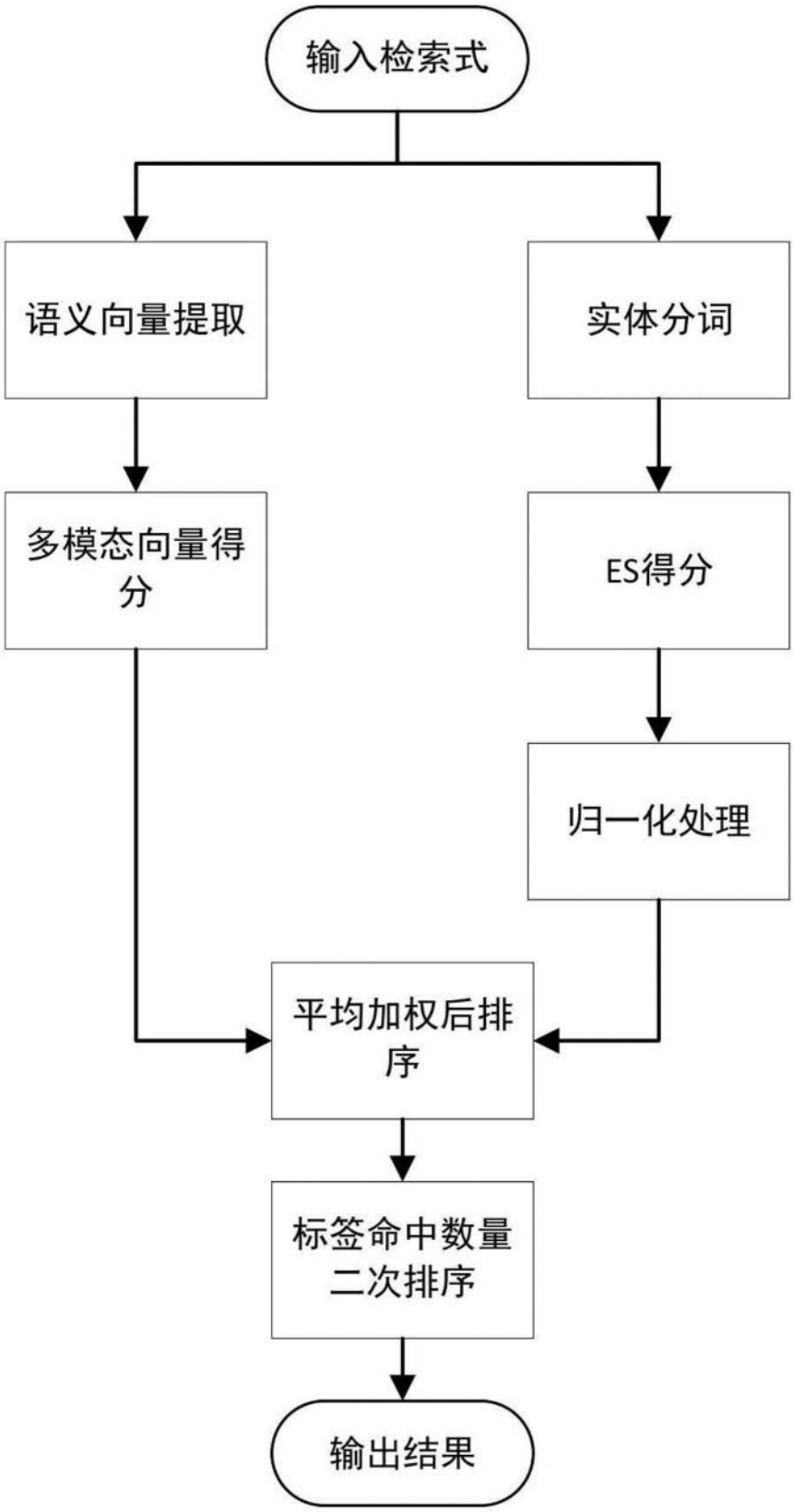
输入检索式
语义向量提取
实体分词
多模态向量得分
ES得分
归一化处理
平均加权后排序
标签命中数量二次排序
输出结果

图3