(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113392196 A
(43)申请公布日 2021.09.14

(21)申请号 202110622823.8

(22)申请日 2021.06.04

(71)申请人 北京师范大学
地址 100875 北京市海淀区新街口外大街19号

(72)发明人 余胜泉 陈鹏鹤 刘杰飞 徐琪 陈玲 卢宇

(74)专利代理机构 北京京万通知识产权代理有限公司 11440
代理人 刘浩 许天易

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/953(2019.01)
G06K 9/62(2006.01)

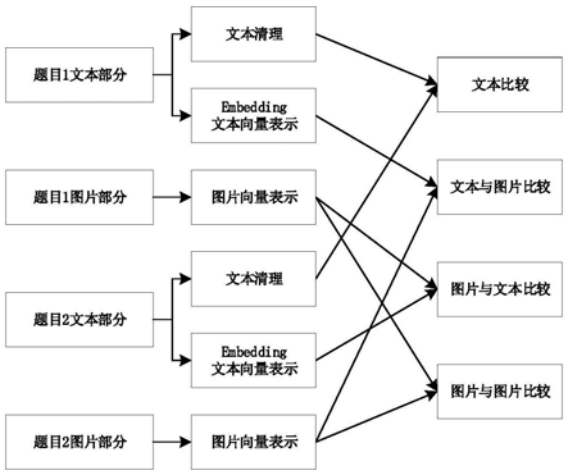
权利要求书2页 说明书7页 附图2页

(54)发明名称

一种基于多模态交叉比较的题目检索方法和系统

(57)摘要

本发明提供了一种基于多模态交叉比较的题目检索系统和方法，系统包括：题目数据解析模块、题目相似度计算模块和结果输出模块；其中，题目数据解析模块，用于接收用户输入的题目信息，并进行预处理及结构化整理；题目相似度计算模块，用于交叉计算用户输入的题目与题库中的题目的文本表示和图片表示的相似度，并加权计算其综合相似度；结果输出模块，用于将综合相似度大于预先设置的科目阈值的题库中的题目及答案等相关信息返回用户。通过本发明的系统，可以使得各科题目在题目库中的检索结果更加准确。



CN 113392196 A

1.一种基于多模态交叉比较的题目检索系统，其特征在于，所述系统包括：题目数据解析模块、题目相似度计算模块和结构输出模块；其中，

题目数据解析模块，用于接收用户输入的题目信息，并进行预处理；

题目相似度计算模块，用于计算用户输入的题目与题库中的题目的相似度；

结果输出模块，用于将所述相似度大于预先设置的科目阈值的所述题库中的题目返回用户。

2.根据权利要求1所述的题目检索系统，其特征在于，在所述题目相似度计算模块中，包括：

a.将题目标题清理后文本和题目内容清理后文本拼接后作为文本表示，将题目图片文本识别内容和题目图片作为图片表示；

b.将用户输入题目的文本表示与图片表示用T1、P1表示，题库中题目的文本表示与图片表示用T2、P2表示，计算T1与T2、T1与P2、P1与T2、P1与P2的相似度，分别用S1、S2、S3、S4表示；

c.计算综合相似度s。

3.根据权利要求2所述的题目检索系统，其特征在于，所述相似度S1采用Jaccard方法进行计算；通过余弦相似度计算相似度S2、S3和S4；优选的，将题目图片文本识别通过BERT模型转换为向量表示，将题目图片通过LeNet卷积网络模型转换为向量，然后将这两个向量拼接作为图片表示的向量化表示。

4.根据权利要求2所述的题目检索系统，其特征在于，综合相似度s的计算公式为：

$$s = \frac{W1S1 + W2S2 + W3S3 + W4S4}{4}$$

W为科目权重。

5.根据权利要求4所述的题目检索系统，其特征在于，科目权重为：

科目	权重（w1、w2、w3、w4）	科目	权重（w1、w2、w3、w4）
语文	5.5，2，2，0.5	物理	5，1，1，3
历史	5，2，2，1	数学	4，2，2，2
地理	5，2，2，1	化学	5，1，1，3
政治	4，2，2，2	生物	4，2，2，2

6.根据权利要求1所述的题目检索系统，其特征在于，在结果输出模块中，若综合相似度大于用户输入的题目所对应的科目阈值，则将题库中问题作为候选问题。

7.根据权利要求6所述的题目检索系统，其特征在于，所述科目阈值为：

科目	阈值	科目	阈值
语文	0.8	物理	0.7
历史	0.8	数学	0.5
地理	0.7	化学	0.5

政治	0.6	生物	0.5

8.一种基于多模态交叉比较的题目检索方法,包括:

步骤1、接收用户输入的题目信息;

步骤2、计算接收的题目与题库中题目的相似度;

步骤3、对于相似度大于科目阈值的题库中题目作为候选问题返回给用户。

9.根据权利要求1所述的题目检索方法,其特征在于,所述步骤2包括:

a.获取题目信息的文本表示和图片表示;

b.将题目的文本表示与图片表示用T1、P1表示,将题库中题目的文本表示与图片表示用T2、P2表示,然后交叉比较四部分内容的相似度S1、S2、S3、S4;其中,S1采用Jaccard方法计算;计算S2、S3、S4时采用余弦相似度计算方法;

c.计算综合相似度s。

10.根据权利要求9所述的题目检索方法,其特征在于,在所述步骤3中,将综合相似度大于用户输入题目所对应的科目阈值的题库中的题目作为候选问题。

一种基于多模态交叉比较的题目检索方法和系统

技术领域

[0001] 本发明涉及计算机技术领域，具体涉及一种基于多模态交叉比较的题目检索方法和系统。

背景技术

[0002] 近年来，随着互联网和人工智能技术的发展，题目问答系统得到的很大的发展，为个性化教学提供了很大的帮助。如何根据用户的问题在题库中快速并准确地检索到与用户输入问题相同或相似的问题然后给出答案变得越来越重要。

[0003] 当前，题目检索系统的实现方式一般是通过比较题目间的文本相似度来实现的，用户将用于描述题目信息的文本传递给题目检索系统，题目检索系统通过比较用户输入题目文本与题库题目文本之间的文本相似度，然后选取相似度最大的题目作为检索结果返回给用户。若用户输入的题目信息为图片的形式，则通过比较图片的相似度来比较题目的相似度。

[0004] 目前文本相似度计算方法主要分为两类。分别是基于字符的文本相似度计算方法及基于向量空间的文本相似度计算方法。

[0005] 基于字符的文本相似度计算方法如传统的编辑距离、汉明距离、Jaccard、LCS等方法，通过直接比较两个文本字符串间相同字符及其序列关系来评估其文本相似度。

[0006] 基于向量空间的文本相似度计算方法如TF-IDF、BM25、以及将文本进行向量表示后通过余弦相似度来计算以及通过神经网络来直接比较文本间的相似情况。

[0007] 随着多媒体的发展，用户在描述题目信息的时候，除了通过纯文本的方式来描述，当前更多的情况是通过文本并结合图片共同描述题目的信息。

[0008] 当前市面上主流的题目检索服务及系统只支持面向文本题目或者图片题目的检索方式，比如“搜题易”(地址为:https://www.xuesai.cn/souti/)小猿搜题等，如图1和图2所示。

[0009] 现有技术主要有以下几种问题，下面说明并描述其问题所在。

[0010] (1)仅输入文本内容，对于数学、物理或者语文中看图说话等形式的题目，用户经常会出现不知如何表述，或者表述不清的状况。

[0011] (2)仅输入图片内容，虽然能返回正确答案，但答案可能并不能够满足用户的需求，比如图2所示，用户对于答案中向量加法的三角形法则可能不熟悉，这样即使看到答案，可能也不清楚如何做题。因此，如果能让用户输入其需求，则能更好地为用户辅文。

[0012] (3)现在也有问答系统让用户同时输入题目文本部分与题目图片部分，但其通过图片文本识别技术将题目图片部分识别为文本表示，然后将题目文本部分与识别出的图片文本内容直接拼接后比较题目间的文本相似度。由于图片文本识别技术的缺陷，会出现题目内容识别错误的情况，比如同一个题目由于光线、角度等不一样，最后识别出的结果也不一样，这种情况会导致原来相同的题目由于图片文本识别的错误而判定为不相似。

发明内容

[0013] 针对现有技术的不足，本发明提供一种基于多模态交叉比较的题目检索方法和系统，通过(1)题目数据解析模块将用户的输入题目进行结构化整理，(2)通过题目相似度计算模块逐一计算用户输入题目与候选题目之间的相似度，并将结果的信息返回给用户。

[0014] 为实现上述目的，本发明通过以下技术方案予以实现。

[0015] 根据本发明的一方面，提出一种基于多模态交叉比较的题目检索系统，包括：题目数据解析模块、题目相似度计算模块和结构输出模块；其中，

[0016] 题目数据解析模块，用于接收用户输入的题目信息，并进行预处理；

[0017] 题目相似度计算模块，用于计算用户输入的题目与题库中的题目的相似度；

[0018] 结果输出模块，用于将所述相似度大于预先设置的科目阈值的所述题库中的题目返回用户。

[0019] 进一步地，在所述题目相似度计算模块中，包括：

[0020] a.将题目标题清理后文本和题目内容清理后文本拼接后作为文本表示，将题目图片文本识别内容作为图片表示；

[0021] b.将用户输入题目的文本表示与图片表示用T1、P1表示，题库中题目的文本表示与图片表示用T2、P2表示，计算T1与T2、T1与P2、P1与T2、P1与P2的相似度，分别用S1、S2、S3、S4表示；

[0022] c.计算综合相似度s。

[0023] 进一步地，所述相似度S1采用Jaccard方法进行计算；通过余弦相似度计算相似度S2、S3和S4；优选的，将题目图片文本识别通过BERT模型转换为向量表示，将题目图片通过LeNet卷积网络模型转换为向量，然后将这两个向量拼接作为图片表示的向量化表示。

[0024] 进一步地，综合相似度s的计算公式为：

[0025] $$s = \frac{W1S1 + W2S2 + W3S3 + W4S4}{4}$$

[0026] W为科目权重。

[0027] 进一步地，科目权重为：

[0028]

科目	权重(w1、w2、w3、w4)	科目	权重(w1、w2、w3、w4)
语文	5.5,2,2,0.5	物理	5,1,1,3
历史	5,2,2,1	数学	4,2,2,2
地理	5,2,2,1	化学	5,1,1,3
政治	4,2,2,2	生物	4,2,2,2

[0029] 进一步地，在结果输出模块中，若综合相似度大于用户输入的题目所对应的科目阈值，则将题库中问题作为候选问题。

[0030] 进一步地，所述科目阈值为：

[0031]

科目	阈值	科目	阈值
语文	0.8	物理	0.7
历史	0.8	数学	0.5
地理	0.7	化学	0.5

政治	0.6	生物	0.5

[0032] 根据本发明的另一方面，提出一种基于多模态交叉比较的题目检索方法，包括：

[0033] 步骤1、接收用户输入的题目信息；

[0034] 步骤2、计算接收的题目与题库中题目的相似度；

[0035] 步骤3、对于相似度大于科目阈值的题库中题目作为候选问题返回给用户。

[0036] 进一步地，所述步骤2包括：

[0037] a.获取题目信息的文本表示和图片表示；

[0038] b.将题目的文本表示与图片表示用T1、P1表示，将题库中题目的文本表示与图片表示用T2、P2表示，然后交叉比较四部分内容的相似度S1、S2、S3、S4；其中，S1采用Jaccard方法计算；计算S2、S3、S4时采用余弦相似度计算方法；

[0039] c.计算综合相似度s。

[0040] 进一步地，在所述步骤3中，将综合相似度大于用户输入题目所对应的科目阈值的题库中的题目作为候选问题。

[0041] 本发明与现有技术相比的有益效果为：

[0042] (1)本发明对于题目检索，采用了基于多模态的交叉比较方法，相比于直接通过题目文本相似度比较及比较题目图片的相似度，极大提高了题目比较的准确度。

[0043] (2)本发明对于交叉比较的四部分相似度进行题目综合相似度计算时，根据科目的不同，基于经验设置了不同的权重参数组合，针对不同科目的题目计算，提高了题目比较的准确度。

[0044] (3)本发明对于题目相似度的判定采用题目比较的综合相似度结果与预先设置的不同科目对应的阈值进行比较，提高了题目比较的准确度。

附图说明

[0045] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0046] 图1是现有问答系统的界面示意图；

[0047] 图2是现有问答系统的另一种界面示意图；

[0048] 图3是根据本发明一个实施例的题目检索系统的结构示意图；

[0049] 图4是根据本发明一个实施例的题目数据解析的流程示意图；

[0050] 图5是根据本发明一个实施例的交叉比较的示意图。

具体实施方式

[0051] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清查、完整的描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0052] 本发明提出一种基于多模态交叉比较的题目检索系统，能够应用在题目问答系统

上，接收用户输入的题目，在题库中进行检索，并将相同及相似问题或答案返回给用户，进而帮助用户完成学习任务。如图3所示，系统包括如下模块：

[0053] (1)题目数据解析模块：接收用户输入的题目信息，包括题目所属学科信息、问题标题、问题说明、及问题图片、问题类型信息。题目数据解析模块可以设置在前端网页或者移动端APP。

[0054] 题目数据解析模块在收到这些题目数据后，先将题目数据分为题目文本数据与题目图片数据，并分别对其进行预处理。题目数据解析模块如图4所示。

[0055] 题目文本数据的预处理包括文本清理与分词。其中，文本清理的工作包括统一编码处理、去除非法字符、去除emoji符号、颜文字符号、HTML标签等符号、去除无效字符。

[0056] 其中无效字符是经过对题库中题目数据进行分析统计后整理的与题目内容无关的字符，主要包括如下：

[0057] 谢、在线求解、请、老师们、老师、帮忙一下、看一下、求助、不会、求解、各位、帮忙、辛苦、您好、麻烦、解一下、帮我讲一下、发一下、这个、希望可以帮我解答、题目怎么做、问一下这个怎么回答、帮我、检查、检查一下、不太会、不太懂、不太明白、不懂、不明白、不清晰、讲解、帮我看看、拜托、解答、希望老师、帮忙看下、教我一下、看不懂题、怎么写、怎么解、怎么求、怎么列、怎么做、详细、对不对、对吗、不理解、读不懂题、帮我分析、问一下、给我讲讲、帮我讲讲、不怎么会、不是特别明白、不怎么会、没有思路、没思路、讲一下、不知道、指点一下、求指教、希望详细、模、卷、中考、考试、期末、通知、版本。

[0058] 题目图片数据为问题图片，需要对其进行图片文本识别，通过现有的图片识别服务将图片中的题目信息识别为文本信息，然后将这些数据进行结构化整理。

[0059] 题目文本和题目图片数据预处理后组合成题目数据，题目数据的结构如下所示，其中题目数据中存储的图片信息为图片地址。

[0060] {

[0061] "question_id":题目数据唯一标识符，

[0062] "question_title_original":题目标题原文本信息，

[0063] "question_content_original":题目内容原文本信息，

[0064] "question_type":题目类型，

[0065] "question_create_time":题目创建时间，

[0066] "question_subject":学科信息，

[0067] "question_pic":题目图片信息，

[0068] "question_pic_content":题目图片文本识别内容，

[0069] "question_title_clean":题目标题清理后文本，

[0070] "question_content_clean":题目内容清理后文本

[0071] }

[0072] 例子：

[0073] {

[0074] "question_id":00063274-1008-4525-9b7d-9f297f72bde1，

[0075] "question_title_original":请老师帮忙看一下这个关于《陋室铭》题目怎么做?，

[0076] "question_content_original":分析题目中的材料,请问对于“山不在高,有仙则名。”作者想表达的思想是什么?,

[0077] "question_type":问答题,

[0078] "question_create_time":2019-06-27 20:58:28,

[0079] "question_subject":语文,

[0080] "question_pic":https://cs.101.com/v0.1/download/actions/direct?dentryId＝9e563a62-a96a-49e2-a238-a810b550b001&serviceName＝fep,

[0081] "question_pic_content":《陋室铭》:山不在高,有仙则名。水不在深,有龙则灵。斯是陋室,惟吾德馨。苔痕上阶绿,草色入帘青。谈笑有鸿儒,往来无白丁。

[0082] "question_title_clean":关于《陋室铭》,

[0083] "question_content_clean":分析题目中的材料,请问对于“山不在高,有仙则名。”作者想表达的思想是什么

[0084] }

[0085] (2)题目相似度计算模块,用于将题目数据与题库中的题目数据逐一比较,将相同及相似题目信息返回给用户。如图5所示,具体的比较方法如下:

[0086] a.将题目标题清理后文本和题目内容清理后文本拼接后作为文本表示,将题目图片文本识别内容与题目图片自身共同作为图片表示。

[0087] b.将题目1的文本表示与图片表示用T1、P1表示,题目2的文本表示与图片表示用T2、P2表示,然后交叉比较四部分内容的相似度,即(T1,T2)、(T1,P2)、(P1,T2)、(P1,P2)的相似度,分别用S1、S2、S3、S4表示。

[0088] 在计算相似度S1也就是比较题目文本表示间的相似度时,采用Jaccard方法进行计算,先将待比较文本表示进行分词,然后去除停用词及标点符号,通过比较文本所包含的词的交集及并集,计算其相似度,具体计算公式如下所示。

[0089] $$S1 = J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

[0090] 其中J(A,B)∈[0,1]。

[0091] 在计算相似度S2与S3,也就是文本表示与图片表示的相似度时,先将文本表示通过BERT模型进行向量化表示,然后将图片表示也进行向量化表示,具体为:将题目图片文本识别同样通过BERT模型转换为向量表示,将题目图片通过LeNet卷积网络模型转换为向量,然后将这两个向量拼接作为图片表示的向量化表示,然后通过余弦相似度计算向量间的相似度。其中余弦相似度计算公式如下所示。

[0092] $$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

[0093] 在计算相似度S4,也就是比较图片表示的相似度时,将题目图片同时进行向量化表示,然后计算其余弦相似度。

[0094] 当获取到交叉比较的相似度结果后,根据预先统计分析后的先验经验,对不同的相似度比较值赋予其不同的权重Wi,则两个待比较题目的综合相似度表示为:

[0095] $$s = \frac{W1S1 + W2S2 + W3S3 + W4S4}{4}$$

[0096] 对于各个学科，设置相应的权重如下(科目权重可以根据经验设置，也可以通过神经网络的训练获得)：

[0097] 表1权重设定情况表

[0098]

科目	权重(w1、w2、w3、w4)	科目	权重(w1、w2、w3、w4)
语文	5.5,2,2,0.5	物理	5,1,1,3
历史	5,2,2,1	数学	4,2,2,2
地理	5,2,2,1	化学	5,1,1,3
政治	4,2,2,2	生物	4,2,2,2

[0099] 这样做的好处是可以提高比较的准确度，相较于原来的方法，交叉比较的方法将题目的文本内容和图片的文本内容分开比较，在进行比较的时候采用了更细化的方式分析题目内容。

[0100] 如下面的例子举例：

[0101] 题目1：

[0102] 题目内容：分析题目中的材料，请问对于“山不在高，有仙则名。”作者想表达的思想是什么?

[0103] 题目图片文本内容：《陋室铭》：山不在高，有仙则名。水不在深，有龙则灵。斯是陋室，惟吾德馨。苔痕上阶绿，草色入帘青。谈笑有鸿儒，往来无白丁。

[0104] 题目2：

[0105] 题目内容：请问“苔痕上阶绿，草色入帘青。”是什么意思?

[0106] 题目图片文本内容：《陋室铭》：山不在高，有仙则名。水不在深，有龙则灵。斯是陋室，惟吾德馨。苔痕上阶绿，草色入帘青。

[0107] 如果直接将题目文本内容与题目图片内容进行拼接然后使用Jaccard计算题目1的题目内容与题目2的题目内容相似度为0.68。

[0108] 而采用交叉比较的方式进行比较后，题目1的文本内容与题目2的文本内容的相似度为0.05，题目1的图片内容与题目2的文本内容的相似度为0.14，题目1的文本内容与题目2的图片内容的相似度为0.19，题目1的图片内容与题目2的图片内容的相似度为0.95，根据预先设置的语文学科的权重信息(5.5,2,2,0.5)，综合计算题目1与题目2的相似度为(5.5*0.05+2*0.14+2*0.19+0.5*0.95)/4＝0.3，其结果说明两者的相似度小，或者说不相似。而实际上，通过人工也可以判断出这两道题目是不相似的，因此，交叉比较的结果0.3比之前的方法的结果0.68更能反应题目的相似度。

[0109] (3)结果输出模块，用于将计算得到的综合相似度与预先设置的科目阈值进行比较，其中针对不同的学科，设置了不同的比较阈值，如表2所示。若综合相似度大于用户输入的题目所对应的科目阈值，则认为题库中问题与用户输入的问题相似，将其作为候选问题。阈值同样可以基于数据集和人工智能网络训练获得。比较题库中的全部题目后，对候选问题根据综合相似度大小进行排序，选取前五个题目返回给用户；如果不足五个，则将全部满足条件的候选题目返回给用户；如果不存在，则提示用户没有查找的相似问题，将用户输入的问题作为新问题进行发布。

[0110] 表2阈值设定情况表

[0111]

科目	阈值	科目	阈值

语文	0.8	物理	0.7
历史	0.8	数学	0.5
地理	0.7	化学	0.5
政治	0.6	生物	0.5

[0112] 根据本发明的另一方面，提出一种基于多模态交叉比较的题目检索方法，包括：

[0113] 步骤1、接收用户输入的题目信息；

[0114] 步骤2、计算接收的题目与题库中题目的相似度；

[0115] 步骤3、对于相似度大于科目阈值的题库中题目作为候选问题返回给用户。

[0116] 在步骤1中，接收题目数据后，先将题目数据分为题目文本数据与题目图片数据，并分别对其进行预处理。具体过程见题目解析数据模块中描述。

[0117] 在步骤2中，a.获取用户输入题目信息的文本表示和图片表示；b.将题目的文本表示与图片表示用T1、P1表示，将题库中题目的文本表示与图片表示用T2、P2表示，然后交叉比较四部分内容的相似度S1、S2、S3、S4；其中，S1采用Jaccard方法计算；计算S2与S3时，先将文本表示通过BERT模型进行向量化表示，然后将题目图片也进行向量化表示，具体为将题目图片文本识别同样通过BERT模型转换为向量表示，将题目图片通过LeNet卷积网络模型转换为向量，然后将这两个向量拼接作为图片表示的向量化表示，然后通过余弦相似度计算向量间的相似度；计算相似度S4时，计算题目图片和题库中题目图片的向量化表示的余弦相似度；c.计算综合相似度s。具体的计算方法如上文所述。

[0118] 在步骤3中，将计算得到的相似度与预先设置的科目阈值进行比较，其中针对不同的学科，设置了不同的比较阈值，如表2所示。若相似度大于用户输入的题目所对应的科目阈值，则认为题库中问题与用户输入的问题相似，将其作为候选问题。阈值同样可以基于数据集和人工智能网络训练获得。比较题库中的全部题目后，对候选问题根据相似度大小进行排序，选取前五个题目返回给用户；如果不足五个，则将全部满足条件的候选题目返回给用户；如果不存在，则提示用户没有查找的相似问题，将用户输入的问题作为新问题进行发布。

[0119] 以上实施例仅用于说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或替换，并不使相应的技术方案的本质脱离本发明各实施例技术方案的精神和范围。


搜题易
请输入您的问题
搜题
热门考试：大学试题、大学本科、成考、自考、考研、高中、初中、更多考试

图1


拍照搜题
海量题库 秒出解析
已知三角形OBC中,点A是BC中
点,D是OB上的点,且$OD=2DB$
,DC和OA交于点E,设
$\overrightarrow{OA}=\vec{a},\overrightarrow{OB}=\vec{b}$
(1)用$\vec{a},\vec{b}$表示向量$\overrightarrow{OC},\overrightarrow{DC}$;
(2)若$\overrightarrow{OE}=\lambda\overrightarrow{OA}$求实数$\lambda$的值.
答案解析
(1)利用向量共线的条件,用向量加法的三角形法
则,可以用向量$\vec{a},\vec{b}$表示向量$\overrightarrow{OC},\overrightarrow{DC}$
(2)利用向量共线及向量相等的条件结合向量加法
的三角形法则,可求λ的值
本题主要考查向量加法的三角形法则,向量共线\
向量相等的条件,关键是要熟悉向量的各个知识点

图2


题目数据解析模块
题目相似度计算模块
检索结果

图3

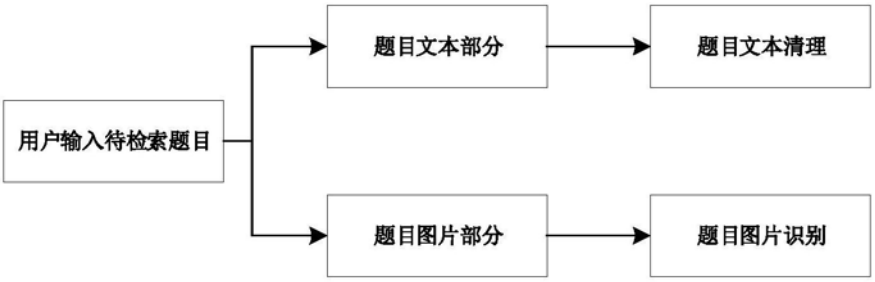

用户输入待检索题目
题目文本部分
题目文本清理
题目图片部分
题目图片识别

图4

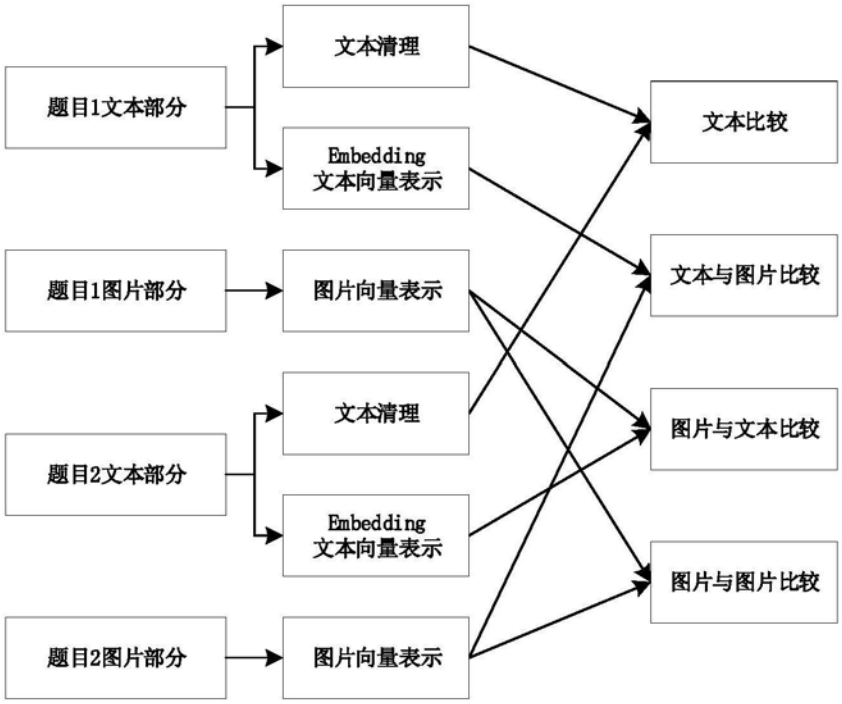

题目1文本部分
文本清理
Embedding
文本向量表示
题目1图片部分
图片向量表示
题目2文本部分
文本清理
Embedding
文本向量表示
题目2图片部分
图片向量表示
文本比较
文本与图片比较
图片与文本比较
图片与图片比较

图5