(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107491447 B
(45)授权公告日 2021.01.22

(21)申请号 201610408229.8
(22)申请日 2016.06.12
(65)同一申请的已公布的文献号
申请公布号 CN 107491447 A
(43)申请公布日 2017.12.19
(73)专利权人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦
(72)发明人 成幸毅 林荣逸 吕钦 李磊
(74)专利代理机构 北京鸿德海业知识产权代理有限公司 11412
代理人 袁媛
(51)Int.Cl.
G06F 16/242(2019.01)
G06F 16/2453(2019.01)
(56)对比文件
CN 104933183 A,2015.09.23
CN 104615767 A,2015.05.13
CN 101131706 A,2008.02.27
US 2012233140 A1,2012.09.13
审查员 田志方

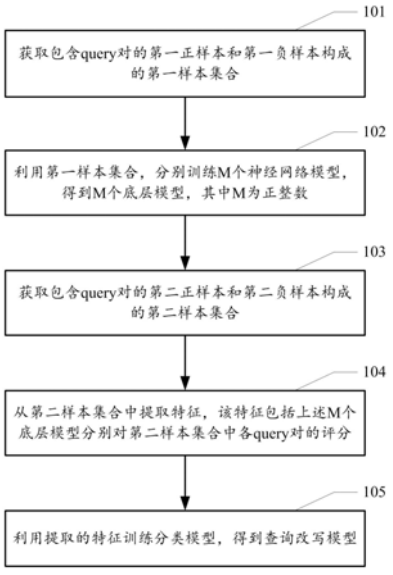
权利要求书3页 说明书11页 附图5页

(54)发明名称

建立查询改写判别模型、查询改写判别的方法和对应装置

(57)摘要

本发明提供了一种建立查询改写判别模型、查询改写判别的方法和对应装置，其中建立查询改写判别模型的方法包括：利用包含query对的第一正样本和第一负样本构成的第一样本集合，分别训练M个神经网络模型，得到M个底层模型，所述M为正整数；从包含query对的第二正样本和第二负样本构成的第二样本集合中提取特征，所述特征包括所述M个底层模型分别对所述第二样本集合中各query对的评分；利用提取的特征训练分类模型，得到查询改写判别模型。本发明利用了前沿的机器学习技术，以学习文本表达的潜在关联，从而实现查询改写的准确判别。



CN 107491447 B

1.一种建立查询改写判别模型的方法,其特征在于,该方法包括:

利用包含query对的第一正样本和第一负样本构成的第一样本集合,分别训练M个神经网络模型,得到M个底层模型,所述M为正整数;

从包含query对的第二正样本和第二负样本构成的第二样本集合中提取特征,所述特征包括所述M个底层模型分别对所述第二样本集合中各query对的评分;

利用提取的特征训练分类模型,得到查询改写判别模型。

2.根据权利要求1所述的方法,其特征在于,所述第一样本集合采用如下方式获取:

从搜索日志中获取被点击url的相似度大于或等于第一阈值的两个query构成的query对作为第一正样本,和/或,从由已有改写规则得到的改写词表中,获取人工选择出的原query的优质改写query,由该原query和优质改写query构成的query对作为第一正样本;

从搜索日志中获取被点击url的相似度小于或等于第二阈值的两个query构成的query对作为第一负样本;

其中所述第一阈值高于所述第二阈值。

3.根据权利要求2所述的方法,其特征在于,所述第二样本集合采用如下方式获取:

从搜索日志中获取被点击url的相似度大于或等于第三阈值并且小于或等于第四阈值的两个query构成的query对,所述第三阈值大于所述第二阈值,所述第四阈值小于所述第一阈值;

依据人工对所述query对进行的标注结果,将人工标注为表述相同含义的query对作为第二正样本,将人工标注为表述不同含义的query对作为第二负样本。

4.根据权利要求2或3所述的方法,其特征在于,对正样本进行以下过滤中的至少一种:

若query对中两个query对应的搜索结果中排在前q个的共同url个数小于预设的个数阈值,则过滤掉该query对,q为预设的正整数;

若query对中两个query分别去掉停用词后得到相同的表述,则过滤掉该query对;

若query对中两个query包含不同的数字内容,则过滤掉该query对;

若query对中两个query对应的url总点击次数小于预设的点击次数阈值,则过滤掉该query对;

若query对中的一个query为另一个query的纠错表述,则过滤掉该query对。

5.根据权利要求2或3所述的方法,其特征在于,对负样本进行以下过滤中的至少一种:

若query对中的各query均不是具有预设需求的query,则过滤掉该query对;

若一个query存在于多个query对,则保留其中m个query对,其他过滤掉,所述m为预设的正整数。

6.根据权利要求1所述的方法,其特征在于,所述神经网络模型包括以下至少一种:

基于多层感知机的神经网络BOW_NN、卷积神经网络CNN、双向递归神经网络BiRNN。

7.根据权利要求1所述的方法,其特征在于,所述特征还包括以下中的一种或任意组合:

统计特征、距离特征、位置特征、词语重要性特征、语义特征以及同义词改写特征。

8.根据权利要求1所述的方法,其特征在于,所述利用提取的特征训练分类模型,得到查询改写判别模型包括:

利用提取的特征分别训练N个分类模型,得到N个高阶模型,所述N为大于1的正整数;

对所述N个高阶模型进行选择和集成，得到查询改写判别模型。

9.根据权利要求8所述的方法，其特征在于，所述分类模型包括以下至少一种：

梯度递归决策树GBDT、支持向量机SVM、逻辑回归LR、随机森林RF、多层感知器MLP。

10.根据权利要求8所述的方法，其特征在于，对所述N个高阶模型进行选择和集成，得到查询改写判别模型包括：

利用测试集对所述N个高阶模型的结果进行测试评分，所述测试集包含已确定改写评分的query对；

依据测试评分选择其中P个高阶模型，所述P小于或等于所述N；

对所述P个高阶模型进行加权处理，得到查询改写判别模型。

11.一种判别查询改写的方法，其特征在于，该方法包括：

从待判别query对中提取特征，所述特征包括M个底层模型对该query对的评分，所述M为正整数；

将提取的特征输入查询改写判别模型，得到所述查询改写判别模型的判别结果；

其中所述M个底层模型和所述查询改写判别模型是采用如权利要求1至10任一权项所述方法得到的。

12.一种建立查询改写判别模型的装置，其特征在于，该装置包括：

第一样本获取单元，用于获取包含query对的第一正样本和第一负样本构成的第一样本集合；

第二样本获取单元，用于获取包含query对的第二正样本和第二负样本构成的第二样本集合；

第一训练单元，用于利用所述第一样本集合，分别训练M个神经网络模型，得到M个底层模型，所述M为正整数；

特征提取单元，用于从所述第二样本集合中提取特征，所述特征包括所述M个底层模型分别对所述第二样本集合中各query对的评分；

第二训练单元，用于利用所述特征提取单元提取的特征训练分类模型，得到查询改写判别模型。

13.根据权利要求12所述的装置，其特征在于，所述第一样本获取单元，具体用于采用如下方式获取所述第一样本集合：

从搜索日志中获取被点击url的相似度大于或等于第一阈值的两个query构成的query对作为第一正样本，和/或，从由已有改写规则得到的改写词表中，获取人工选择出的原query的优质改写query，由该原query和优质改写query构成的query对作为第一正样本；

从搜索日志中获取被点击url的相似度小于或等于第二阈值的两个query构成的query对作为第一负样本；

其中所述第一阈值高于所述第二阈值。

14.根据权利要求13所述的装置，其特征在于，所述第二样本获取单元，具体用于采用如下方式获取所述第二样本集合：

从搜索日志中获取被点击url的相似度大于或等于第三阈值并且小于或等于第四阈值的两个query构成的query对，所述第三阈值大于所述第二阈值，所述第四阈值小于所述第一阈值；

依据人工对所述query对进行的标注结果，将人工标注为表述相同含义的query对作为第二正样本，将人工标注为表述不同含义的query对作为第二负样本。

15.根据权利要求13或14所述的装置，其特征在于，所述第一样本获取单元和所述第二样本获取单元，还用于对正样本进行以下过滤中的至少一种：

若query对中两个query对应的搜索结果中排在前q个的共同url个数小于预设的个数阈值，则过滤掉该query对，q为预设的正整数；

若query对中两个query分别去掉停用词后得到相同的表述，则过滤掉该query对；

若query对中两个query包含不同的数字内容，则过滤掉该query对；

若query对中两个query对应的url总点击次数小于预设的点击次数阈值，则过滤掉该query对；

若query对中的一个query为另一个query的纠错表述，则过滤掉该query对。

16.根据权利要求13或14所述的装置，其特征在于，所述第一样本获取单元和所述第二样本获取单元，还用于对负样本进行以下过滤中的至少一种：

若query对中的各query均不是具有预设需求的query，则过滤掉该query对；

若一个query存在于多个query对，则保留其中m个query对，其他过滤掉，所述m为预设的正整数。

17.根据权利要求12所述的装置，其特征在于，所述神经网络模型包括以下至少一种：

基于多层感知机的神经网络BOW_NN、卷积神经网络CNN、双向递归神经网络BiRNN。

18.根据权利要求12所述的装置，其特征在于，所述特征还包括以下中的一种或任意组合：

统计特征、距离特征、位置特征、词语重要性特征、语义特征以及同义词改写特征。

19.根据权利要求12所述的装置，其特征在于，所述第二训练单元，具体用于：利用提取的特征分别训练N个分类模型，得到N个高阶模型，所述N为大于1的正整数；对所述N个高阶模型进行选择和集成，得到查询改写判别模型。

20.根据权利要求19所述的装置，其特征在于，所述分类模型包括以下至少一种：

梯度递归决策树GBDT、支持向量机SVM、逻辑回归LR、随机森林RF、多层感知器MLP。

21.根据权利要求19所述的装置，其特征在于，所述第二训练单元在对所述N个高阶模型进行选择和集成，得到查询改写判别模型时，具体执行：

利用测试集对所述N个高阶模型的结果进行测试评分，所述测试集包含已确定改写评分的query对；

依据测试评分选择其中P个高阶模型，所述P小于或等于所述N；

对所述P个高阶模型进行加权处理，得到查询改写判别模型。

22.一种判别查询改写的装置，其特征在于，该装置包括：

特征提取单元，用于从待判别query对中提取特征，所述特征包括M个底层模型对该query对的评分，所述M为正整数；

判别单元，用于将所述特征提取单元提取的特征输入查询改写判别模型，得到所述查询改写判别模型的判别结果；

其中所述M个底层模型和所述查询改写判别模型是采用如权利要求12至21任一权项所述装置得到的。

建立查询改写判别模型、查询改写判别的方法和对应装置

【技术领域】

[0001] 本发明涉及计算机应用技术领域，特别涉及一种建立查询改写判别模型、查询改写判别的方法和对应装置。

【背景技术】

[0002] 在搜索引擎中为了改善搜索结果，引入了查询改写这一技术。通过将用户输入的query进行改写，使得搜索结果能够召回改写后的query对应的搜索结果，从而使得用户需求的表达更加准确。

[0003] 在现有的查询改写技术中，主要是基于一些人工制定的规则，例如片段改写规则、调序改写规则、链式改写规则、省略改写规则，等等。然而，中文自然语言博大精深，字里行间体现了我国数千年的文化底蕴和先人智慧，基于人工制定的规则进行查询改写时，往往达不到较高的准确度要求。例如，在基于片段改写规则时，将“老干妈”改写为“老干娘”；在基于调序改写规则时，将“北京南到深圳”改写为“南京到深圳北”；在基于链式改写规则时，将“湖北汽车票”改写为“湖北车票”，再进而改写为“湖北火车票”；在基于省略改写规则时，将“美股的行情”改写为“美的行情”……显然这些查询改写的准确度是比较差的。因此急需一种判别一个query是否可以用于另一query的查询改写的方式。

【发明内容】

[0004] 有鉴于此，本发明提供了一种建立查询改写判别模型、查询改写判别的方法和对应装置，以便于准确判别一个query是否可以用于另一query的查询改写。

[0005] 具体技术方案如下：

[0006] 本发明提供了一种建立查询改写判别模型的方法，该方法包括：

[0007] 利用包含query对的第一正样本和第一负样本构成的第一样本集合，分别训练M个神经网络模型，得到M个底层模型，所述M为正整数；

[0008] 从包含query对的第二正样本和第二负样本构成的第二样本集合中提取特征，所述特征包括所述M个底层模型分别对所述第二样本集合中各query对的评分；

[0009] 利用提取的特征训练分类模型，得到查询改写判别模型。

[0010] 根据本发明一优选实施方式，所述第一样本集合采用如下方式获取：

[0011] 从搜索日志中获取被点击url的相似度大于或等于第一阈值的两个query构成的query对作为第一正样本，和/或，利用已有改写规则确定出原query的优质改写query，由该原query和优质改写query构成的query对作为第一正样本；

[0012] 从搜索日志中获取被点击url的相似度小于或等于第二阈值的两个query构成的query对作为第一负样本；

[0013] 其中所述第一阈值高于所述第二阈值。

[0014] 根据本发明一优选实施方式，所述第二样本集合采用如下方式获取：

[0015] 从搜索日志中获取被点击url的相似度大于或等于第三阈值并且小于或等于第四

阈值的两个query构成的query对，所述第三阈值大于所述第二阈值，所述第四阈值小于所述第一阈值；

[0016] 依据人工对所述query对进行的标注结果，将人工标注为表述相同含义的query对作为第二正样本，将人工标注为表述不同含义的query对作为第二负样本。

[0017] 根据本发明一优选实施方式，对正样本进行以下过滤中的至少一种：

[0018] 若query对中两个query对应的搜索结果中排在前q个的共同url个数小于预设的个数阈值，则过滤掉该query对，q为预设的正整数；

[0019] 若query对中两个query分别去掉停用词后得到相同的表述，则过滤掉该query对；

[0020] 若query对中两个query包含不同的数字内容，则过滤掉该query对；

[0021] 若query对中两个query对应的url总点击次数小于预设的点击次数阈值，则过滤掉该query对；

[0022] 若query对中的一个query为另一个query的纠错表述，则过滤掉该query对。

[0023] 根据本发明一优选实施方式，对负样本进行以下过滤中的至少一种：

[0024] 若query对中的各query均不是具有预设需求的query，则过滤掉该query对；

[0025] 若一个query存在于多个query对，则保留其中m个query对，其他过滤掉，所述m为预设的正整数。

[0026] 根据本发明一优选实施方式，所述神经网络模型包括以下至少一种：

[0027] 基于多层感知机的神经网络BOW_NN、卷积神经网络CNN、双向递归神经网络BiRNN。

[0028] 根据本发明一优选实施方式，所述特征还包括以下中的一种或任意组合：

[0029] 统计特征、距离特征、位置特征、词语重要性特征、语义特征以及同义词改写特征。

[0030] 根据本发明一优选实施方式，所述利用提取的特征训练分类模型，得到查询改写判别模型包括：

[0031] 利用提取的特征分别训练N个分类模型，得到N个高阶模型，所述N为大于1的正整数；

[0032] 对所述N个高阶模型进行选择和集成，得到查询改写判别模型。

[0033] 根据本发明一优选实施方式，所述分类模型包括以下至少一种：

[0034] 梯度递归决策树GBDT、支持向量机SVM、逻辑回归LR、随机森林RF、多层感知器MLP。

[0035] 根据本发明一优选实施方式，对所述N个高阶模型进行选择和集成，得到查询改写判别模型包括：

[0036] 利用测试集对所述N个高阶模型的结果进行测试评分，所述测试集包含已确定改写评分的query对；

[0037] 依据测试评分选择其中P个高阶模型，所述P小于或等于所述N；

[0038] 对所述P个高阶模型进行加权处理，得到查询改写判别模型。

[0039] 本发明还提供了一种判别查询改写的方法，该方法包括：

[0040] 从待判别query对中提取特征，所述特征包括M个底层模型对该query对的评分，所述M为正整数；

[0041] 将提取的特征输入查询改写判别模型，得到所述查询改写判别模型的判别结果；

[0042] 其中所述M个底层模型和所述查询改写判别模型是采用上述方法得到的。

[0043] 本发明进一步提供了一种建立查询改写判别模型的装置，该装置包括：

[0044] 第一样本获取单元，用于获取包含query对的第一正样本和第一负样本构成的第一样本集合；

[0045] 第二样本获取单元，用于获取包含query对的第二正样本和第二负样本构成的第二样本集合；

[0046] 第一训练单元，用于利用所述第一样本集合，分别训练M个神经网络模型，得到M个底层模型，所述M为正整数；

[0047] 特征提取单元，用于从所述第二样本集合中提取特征，所述特征包括所述M个底层模型分别对所述第二样本集合中各query对的评分；

[0048] 第二训练单元，用于利用所述特征提取单元提取的特征训练分类模型，得到查询改写判别模型。

[0049] 根据本发明一优选实施方式，所述第一样本获取单元，具体用于采用如下方式获取所述第一样本集合：

[0050] 从搜索日志中获取被点击url的相似度大于或等于第一阈值的两个query构成的query对作为第一正样本，和/或，利用已有改写规则确定出原query的优质改写query，由该原query和优质改写query构成的query对作为第一正样本；

[0051] 从搜索日志中获取被点击url的相似度小于或等于第二阈值的两个query构成的query对作为第一负样本；

[0052] 其中所述第一阈值高于所述第二阈值。

[0053] 根据本发明一优选实施方式，所述第二样本获取单元，具体用于采用如下方式获取所述第二样本集合：

[0054] 从搜索日志中获取被点击url的相似度大于或等于第三阈值并且小于或等于第四阈值的两个query构成的query对，所述第三阈值大于所述第二阈值，所述第四阈值小于所述第一阈值；

[0055] 依据人工对所述query对进行的标注结果，将人工标注为表述相同含义的query对作为第二正样本，将人工标注为表述不同含义的query对作为第二负样本。

[0056] 根据本发明一优选实施方式，所述第一样本获取单元和所述第二样本获取单元，还用于对正样本进行以下过滤中的至少一种：

[0057] 若query对中两个query对应的搜索结果中排在前q个的共同url个数小于预设的个数阈值，则过滤掉该query对，q为预设的正整数；

[0058] 若query对中两个query分别去掉停用词后得到相同的表述，则过滤掉该query对；

[0059] 若query对中两个query包含不同的数字内容，则过滤掉该query对；

[0060] 若query对中两个query对应的url总点击次数小于预设的点击次数阈值，则过滤掉该query对；

[0061] 若query对中的一个query为另一个query的纠错表述，则过滤掉该query对。

[0062] 根据本发明一优选实施方式，所述第一样本获取单元和所述第二样本获取单元，还用于对负样本进行以下过滤中的至少一种：

[0063] 若query对中的各query均不是具有预设需求的query，则过滤掉该query对；

[0064] 若一个query存在于多个query对，则保留其中m个query对，其他过滤掉，所述m为预设的正整数。

[0065] 根据本发明一优选实施方式，所述神经网络模型包括以下至少一种：

[0066] 基于多层感知机的神经网络BOW_NN、卷积神经网络CNN、双向递归神经网络BiRNN。

[0067] 根据本发明一优选实施方式，所述特征还包括以下中的一种或任意组合：

[0068] 统计特征、距离特征、位置特征、词语重要性特征、语义特征以及同义词改写特征。

[0069] 根据本发明一优选实施方式，所述第二训练单元，具体用于：利用提取的特征分别训练N个分类模型，得到N个高阶模型，所述N为大于1的正整数；对所述N个高阶模型进行选择和集成，得到查询改写判别模型。

[0070] 根据本发明一优选实施方式，所述分类模型包括以下至少一种：

[0071] 梯度递归决策树GBDT、支持向量机SVM、逻辑回归LR、随机森林RF、多层感知器MLP。

[0072] 根据本发明一优选实施方式，所述第二训练单元在对所述N个高阶模型进行选择和集成，得到查询改写判别模型时，具体执行：

[0073] 利用测试集对所述N个高阶模型的结果进行测试评分，所述测试集包含已确定改写评分的query对；

[0074] 依据测试评分选择其中P个高阶模型，所述P小于或等于所述N；

[0075] 对所述P个高阶模型进行加权处理，得到查询改写判别模型。

[0076] 本发明还提供了一种判别查询改写的装置，该装置包括：

[0077] 特征提取单元，用于从待判别query对中提取特征，所述特征包括M个底层模型对该query对的评分，所述M为正整数；

[0078] 判别单元，用于将所述特征提取单元提取的特征输入查询改写判别模型，得到所述查询改写判别模型的判别结果；

[0079] 其中所述M个底层模型和所述查询改写判别模型是采用上述建立查询改写判别模型的装置得到的。

[0080] 由以上技术方案可以看出，本发明将自学习得到的底层模型对query对的评分作为特征，并用以训练分类模型，从而得到查询改写判别模型，这种方式利用了前沿的机器学习技术，以学习文本表达的潜在关联，从而实现查询改写的准确判别。

【附图说明】

[0081] 图1为本发明实施例提供的建立查询改写判别模型的方法流程图；

[0082] 图2为本发明实施例提供的一个建立查询改写判别模型的实例图；

[0083] 图3为本发明实施例提供的查询改写判别的方法流程图；

[0084] 图4为本发明实施例提供的一个查询改写判别的实例图；

[0085] 图5为本发明实施例提供的建立查询改写判别模型的装置结构图；

[0086] 图6为本发明实施例提供的判别查询改写的装置结构图。

【具体实施方式】

[0087] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0088] 在本发明实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。在本发明实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”

也旨在包括多数形式,除非上下文清楚地表示其他含义。

[0089] 应当理解,本文中使用的术语“和/或”仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0090] 取决于语境,如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”或“响应于检测”。类似地,取决于语境,短语“如果确定”或“如果检测(陈述的条件或事件)”可以被解释成为“当确定时”或“响应于确定”或“当检测(陈述的条件或事件)时”或“响应于检测(陈述的条件或事件)”。

[0091] 本发明颠覆性地采用有监督的机器学习技术,通过大规模数据和算法去发现语义表达的规律。下面通过实施例对该方法进行详述。

[0092] 图1为本发明实施例提供的方法流程图,如图1中所示,该方法可以包括以下步骤:

[0093] 在101中,获取包含query对的第一正样本和第一负样本构成的第一样本集合。

[0094] 在此处所采用“第一”的限定方式,主要是为了与后续出现的“第二”所限定的样本数据进行区分,并没有任何语义上的限制,后续出现的“第二”也是如此。

[0095] 本步骤中获取的第一样本集合是基于大数据的样本集合,对于正样本而言,可以采用但不限于以下两种:

[0096] 第一种:利用已有改写规则确定出原query的优质改写query,由该原query和优质改写query构成的query对作为第一正样本。

[0097] 正如背景技术中所述,目前query改写主要基于人工制定的改写规则,这些改写规则中有一些改写query是非常优质的,那么在本发明实施例中,可以从由已有改写规则得到的改写词表中,选择出query对,该query对包含的是原query和优质改写query。

[0098] 第二种:从搜索日志中获取被点击url的相似度大于或等于第一阈值的两个query构成的query对作为第一正样本。如果两个query的搜索结果相似,那么一般可以认为它们的语义(意图)也是相似的。特别是对于中高频query,经过了用户的点击校验,搜索结果的相关性一般较强。采用这种方法,可以用较小的成本获取大量意图相似的query对,且能够涵盖大部分领域,作为底层模型的正样本。

[0099] 其中两个query对应的被点击url的相似度可以采用多种方式衡量,在此列举一种方式:

[0100] 假设query对中的两个query分别为:queryLeft和queryRight,两个query对应的共同被点击url构成的集合为overlapUrls,如果

[0101] min(overlapClickRatioLeft,overlapClickRatioLeft)>0.3并且

[0102] max(overlapClickRatioLeft,overlapClickRatioLeft)>0.6,则认为queryLeft和queryRight存在比较多的共同点击,即对应的被点击url的相似度满足作为正样本的要求。需要说明的是,上述0.3和0.6为一种优选的阈值选择,但并不限于这些数值。

[0103] 其中,$$\text{overlapClickRatioLeft}=\frac{\sum_{u\in\text{overlapUrls}}\text{clickLeft}(u)}{2+\sum_{u\in\text{leftUrls}}\text{clickLeft}(u)}$$

[0104] $$\text{overlapClickRatioRight}=\frac{\sum_{u\in\text{overlapUrls}}\text{clickRight}(u)}{2+\sum_{u\in\text{rightUrls}}\text{clickRight}(u)}$$

[0105] leftUrls为queryLeft对应的被点击url构成的集合，rightUrls为queryRight对应的被点击url构成的集合，clickLeft(u)为queryLeft对应的u的被点击数量，clickRight(u)为queryRight对应的u的被点击数量。

[0106] 对于采用上述方式得到的正样本，可以采用以下方式中的至少一种进行过滤处理：

[0107] 第一种过滤：若query对中两个query对应的搜索结果中排在前q个的共同url个数小于预设的个数阈值，则过滤掉该query对。例如，如果两个query对应的搜索结果中排在前10个的url中共同url的个数小于3个，则说明该query对中的两个query在语义上并没有那么相似，可以从正样本集合中过滤掉该query对。

[0108] 第二种过滤：若query对中的两个query分别去掉停用词后得到相同的表述，则过滤掉该query对。

[0109] 第三种过滤：若query对中两个query包含不同的数字内容，则过滤掉该query对。例如“跑男第三季”和“奔跑吧兄弟第四季”，里面包含相冲突的数字内容，意图差别较大，因此不适合作为查询改写的正样本。

[0110] 第四种过滤：若query对中两个query对应的url总点击次数小于预设的点击次数阈值，则过滤掉该query对。这种情况很大程度上说明这种query的表述并不合适，才使得搜索结果没有命中用户的需求，因此这部分query对并不适合作为正样本。

[0111] 第五种过滤：若query对中的一个query为另一个query的纠错表述，则过滤掉该query对。例如“连单杠”和“练单杠”，后者是对前者中错别字的纠正，那么这种就不适合作为查询改写的正样本。

[0112] 对于负样本而言，可以从搜索日志中获取被点击url的相似度小于或等于第二阈值的两个query构成的query对，作为第一负样本。第二阈值小于第一阈值。同样采用这种方法，可以用较小的成本获取大量意图相似的query对，且能够涵盖大部分领域，作为底层模型的负样本。

[0113] 其中两个query对应的被点击url的相似度可以采用多种方式衡量，在此列举一种方式：

[0114] 假设overlapQ为query对对应的被点击url中，排在前N个的url的交集，若某query对满足如下条件，则将其作为负样本：

[0115] $1 \leqslant overlap15 \leqslant 3$ 且 $0 \leqslant overlap10 \leqslant 2$ 且 $overlap5 = 0$ 且 $clickLeft \geqslant 2$ 且 $clickRight \geqslant 5$。其中，clickLeft为queryLeft对应的所有url的被点击次数，clickRight为queryRight对应的所有url的被点击次数。

[0116] 对于负样本，可以执行以下过滤方式中的至少一种：

[0117] 第一种过滤：若query对中的各query均不是具有预设需求的query，则过滤掉该query对。例如，假设查询改写是针对的结构化搜索，则若query对中的各query均不具有结构化搜索需求，则过滤掉该query对。

[0118] 第二种过滤：若一个query存在于多个query对，则保留其中m个query对，其他过滤掉，m为预设的正整数。例如，每个query最多存在于5个query对中，即每个queryLeft最多保留5个queryRight，可以从中随机选取5个，其他过滤掉。

[0119] 采用上述方式获取到的样本是非常大规模的，可以得到上亿级别。

[0120] 在102中，利用第一样本集合，分别训练M个神经网络模型，得到M个底层模型，其中M为正整数。

[0121] 神经网络模型能够针对样本自动进行特征的学习，最终得到的底层模型能够对任意输入的query对进行评分，该评分可以看做是该query对中的queryRight作为queryLeft的查询改写的评分。

[0122] 在本发明实施例中，神经网络模型可以采用诸如基于多层感知机的神经网络(BOW_NN)、卷积神经网络(CNN)、双向递归神经网络(BiRNN)。由于神经网络模型的实现机制以及对于文本的学习过程为较为成熟的技术，在此不再赘述。其中，Char-BiRNN是一种优选的双向递归神经网络，其优点是无需对输入进行分字，其学习效果明显优于其他神经网络。

[0123] 另外，单个神经网络模型过于垄断，风险较大，在本发明可以采用多个结构差异的神经网络模型，即上述M可以是2以上的值，经过训练分别得到多个底层模型。

[0124] 但若仅仅使用底层模型来进行查询改写的判别，则准确度仍然不高，经过测试发现，底层模型对查询改写的判别准确度通常在70％左右。这主要是由于样本分布以及大部分样本特征过于明显所造成的，对于一些边界的样本并未严格区分，造成了精度不足。为了克服这一问题，继续执行以下步骤建立更准确和高阶的模型。

[0125] 在103中，获取包含query对的第二正样本和第二负样本构成的第二样本集合。

[0126] 第二样本集合主要是选取边界数据，使得模型能够区分更加细微的意图差别，例如“糖尿病治疗”和“糖尿病病因”应该判定为不相似，但其依据上述第一负样本的获取策略可能并不能获得，因为两者在被点击url上具有一定相似性，但相似度并没有那么低(未低于第二阈值)。因此需要挖掘一些比较边界的样本。可以首先从搜索日志中获取被点击url的相似度大于或等于第三阈值并且小于或等于第四阈值的两个query构成的query对，所述第三阈值大于所述第二阈值，所述第四阈值小于所述第一阈值；然后将该部分query对提交给人工进行标注，将人工标注为表述相同含义的query对作为第二正样本，将人工标注为表述不同含义的query对作为第二负样本。

[0127] 由于对于边界上的样本的标准精度要求会更高，因此在此可以采用多个标注人员，例如由三个熟悉查询改写的工程师分别进行标注，然后去异求同。这种方式得到的第二样本集合的数据大概在上万级别。

[0128] 需要说明的是上述103与101、102的执行顺序并不加以限制，也可以与步骤101同时执行，也可以在步骤101之后执行，图1中所示顺序仅为其中一种实现顺序。

[0129] 在104中，从第二样本集合中提取特征，该特征包括上述M个底层模型分别对第二样本集合中各query对的评分。

[0130] 将上述第二样本集合分别输入上述M个底层模型后，就可以分别得到各底层模型对各query对的评分，该评分可以作为训练最终查询改写判别模型的特征。这一特征实际上是将多个底层模型从大规模训练样本上学习到的知识迁移到人工标注的边界样本上。

[0131] 除了上述特征之外，还可以包含一些其他特征，从而训练得到高阶模型。例如以下特征中的一种或任意组合：

[0132] 1)统计特征。例如统计query中词语term的个数或占比，term可以采用n-gram的形式；统计是否数字占比。

[0133] 2)距离特征。例如确认两个query之间的jaccard距离或者编辑距离等。其中

jaccard距离为query对中两个query共现的term数量与query对所包含term的总数量。

[0134] 3)位置特征。例如确认两个query中共同的term在两个query中的位置方差均值。

[0135] 4)词语重要性特征。例如query对中term的tf-idf特征。

[0136] 5)语义特征。例如term的词性、句子成分等等。

[0137] 6)同义词改写特征。例如确认query对中属于同义词的term。

[0138] 在105中,利用提取的特征训练分类模型,得到查询改写判别模型。

[0139] 本步骤中训练的分类模型可以是一个,即训练的该一个分类模型即得到查询改写判别模型。

[0140] 作为一种优选的实施方式,本步骤中可以训练多个分类模型,即利用提取的特征分别训练N个分类模型,得到N个高阶模型,N为大于1的正整数;然后再对N个高阶模型进行选择和集成,得到查询改写判别模型,这种方式得到的查询改写判别模型实际上是一个集成模型。

[0141] 本步骤中涉及的分类模型可以采用GBDT(梯度递归决策树)、SVM(支持向量机)、LR(逻辑回归)、RF(随机森林)、MLP(多层感知器)等等。上述的N个分类模型可以是不同类型的分类模型,也可以是相同类型的分类模型但采用不同的模型参数。

[0142] 例如,可以利用从第二样本集合中提取的特征,训练N个GBDT,这N个GBDT分别采用不同的模型参数(例如深度、决策树数量、学习力等参数),这样就可以得到N个高阶模型。可以直接将这N个模型进行集成,但由于这N个模型中不一定所有模型都能够达到预期的判别准确率,因此可以从这N个模型中选取出能够达到预期判别准确率的模型进行集成。

[0143] 在此可以利用测试集对这N个高阶模型的结果进行测试评分,其中测试集中包含了一些已确定改写评分的query对,然后将这些query对分别输入N个高阶模型,得到各高阶模型分别对各query对的评分,然后将得到的评分与测试集中各query对的改写评分进行比较,得到这N个高阶模型的结果的测试评分,例如可以采用AUC体现测试评分。然后依据测试评分可以从中选择出P个高阶模型,例如选择测试评分大于预设测试评分阈值的高阶模型,P小于或等于N。

[0144] 选择出P个高阶模型后,可以采用加权的方式对这几个高阶模型进行集成,得到查询改写判别模型。即可以为这几个高阶模型分配各自的权值,这些权值用于在利用查询改写判别模型判别一个query是否是另一个query的查询改写时,可以将各个高阶模型对该query对的评分进行加权处理后得到的最终高评分作为查询改写判别模型的评分,据此评分来产生判别结果。

[0145] 举一个具体的实施例:

[0146] 如图2所示,采用上述101中所示的方式得到大数据样本,采用上述103所示的方式得到边界样本。利用大数据样本分别训练BOW_NN、CNN、BiRNN三个模型。然后将边界样本输入训练得到的三个模型,分别得到对边界样本中各query对的评分,将这三个模型的评分作为特征,连同其他从边界样本中提取出的诸如统计特征、距离特征、位置特征、词语重要性特征、语义特征、同义词改写特征等,一起用于训练N个GBDT模型,然后从中选出P个BGDT模型进行集成后,得到最终的查询改写判别模型。

[0147] 完成查询改写判别模型的建立后,若采用该模型进行查询改写判别的过程可以如图3所示,包括以下步骤:

[0148] 在301中，从待判别query对中提取特征，该特征包括上述各底层模型对该query对的评分。

[0149] 假设要判别query对中的一个query是否为另一个query的查询改写，则可以将该query对输入上述实施例中训练得到的M个底层模型，会得到各底层模型对该query对的评分。将这M个评分作为特征，再进一步结合从该query对中提取的统计特征、距离特征、位置特征、词语重要性特征、语义特征以及同义词改写特征等(训练查询改写判别模型时采用了哪些特征，在此就从待判别query对中提取哪些特征)。

[0150] 在302中，将提取的特征输入查询改写判别模型，得到查询改写判别模型的判别结果。

[0151] 若查询改写判别模型是由多个高阶模型集成得到的，那么本步骤实际就是将提取的特征分别输入各高阶模型，得到各高阶模型对该待判别query对的评分，然后依据各高阶模型的权值，对这些评分进行加权处理，例如加权求和或加权求平均，依据最终得到的评分来判别待判别query对中的一个query是否为另一个query的查询改写。

[0152] 以图2所示查询改写判别模型举一个实施例：

[0153] 如图4所示，待判别query对输入BOW_NN、CNN和BiRNN三个底层模型，得到三个输出评分。从待判别query对中提取统计特征、距离特征、位置特征、词语重要性特征、语义特征、同义词改写特征等特征，连同上述三个评分一起作为特征输入查询改写判别模型，该查询改写判别模型是由P个GBDT模型集成而成的，由这P个GBDT模型输出的评分进行加权处理，最终得到查询改写判别模型的判别结果。

[0154] 以上是对本发明所提供方法进行的描述，下面对本发明提供的装置进行详细描述。

[0155] 图5为本发明实施例提供的建立查询改写判别模型的装置结构图，如图5所示，该装置可以包括：第一样本获取单元01、第二样本获取单元02、第一训练单元03、特征提取单元04和第二训练单元05，各组成单元的主要功能如下：

[0156] 第一样本获取单元01负责获取包含query对的第一正样本和第一负样本构成的第一样本集合。

[0157] 具体地，第一样本获取单元01可以采用如下方式获取第一样本集合：

[0158] 从搜索日志中获取被点击url的相似度大于或等于第一阈值的两个query构成的query对作为第一正样本，和/或，利用已有改写规则确定出原query的优质改写query，由该原query和优质改写query构成的query对作为第一正样本。

[0159] 从搜索日志中获取被点击url的相似度小于或等于第二阈值的两个query构成的query对作为第一负样本；其中第一阈值高于第二阈值。

[0160] 第二样本获取单元02负责获取包含query对的第二正样本和第二负样本构成的第二样本集合。

[0161] 具体地，第二样本获取单元02可以采用如下方式获取第二样本集合：

[0162] 首先，从搜索日志中获取被点击url的相似度大于或等于第三阈值并且小于或等于第四阈值的两个query构成的query对，第三阈值大于第二阈值，第四阈值小于第一阈值。然后，依据人工对query对进行的标注结果，将人工标注为表述相同含义的query对作为第二正样本，将人工标注为表述不同含义的query对作为第二负样本。

[0163] 对于采用上述方式获得的正样本，第一样本获取单元01和第二样本获取单元02可以对正样本进行以下过滤中的至少一种：

[0164] 第一种过滤：若query对中两个query对应的搜索结果中排在前q个的共同url个数小于预设的个数阈值，则过滤掉该query对，q为预设的正整数。

[0165] 第二种过滤：若query对中两个query分别去掉停用词后得到相同的表述，则过滤掉该query对。

[0166] 第三种过滤：若query对中两个query包含不同的数字内容，则过滤掉该query对。

[0167] 第四种过滤：若query对中两个query对应的url总点击次数小于预设的点击次数阈值，则过滤掉该query对。

[0168] 第五种过滤：若query对中的一个query为另一个query的纠错表述，则过滤掉该query对。

[0169] 对于负样本而言，第一样本获取单元01和第二样本获取单元02可以进行以下过滤中的至少一种：

[0170] 第一种过滤：若query对中的各query均不是具有预设需求的query，则过滤掉该query对。

[0171] 第二种过滤：若一个query存在于多个query对，则保留其中m个query对，其他过滤掉，m为预设的正整数。

[0172] 第一训练单元03负责利用第一样本集合，分别训练M个神经网络模型，得到M个底层模型，M为正整数。其中，神经网络模型可以包括但不限于：BOW_NN、CNN、BiRNN等。神经网络模型能够针对样本自动进行特征的学习，最终得到的底层模型能够对任意输入的query对进行评分，该评分可以看做是该query对中的queryRight作为queryLeft的查询改写的评分。

[0173] 特征提取单元04负责从第二样本集合中提取特征，其中特征包括M个底层模型分别对第二样本集合中各query对的评分，还包括统计特征、距离特征、位置特征、词语重要性特征、语义特征以及同义词改写特征等中的一种或任意组合。

[0174] 第二训练单元05负责利用特征提取单元04提取的特征训练分类模型，得到查询改写判别模型。本步骤中训练的分类模型可以是一个，即训练的该一个分类模型即得到查询改写判别模型。作为一种优选的实施方式，第二训练单元05可以利用提取的特征分别训练N个分类模型，得到N个高阶模型，N为大于1的正整数；对N个高阶模型进行选择和集成，得到查询改写判别模型。

[0175] 其中，分类模型可以采用GBDT、SVM、LR、RF、MLP等中的一种或任意组合，采用的多个分类模型可以是不同类型的分类模型，也可以是相同类型的分类模型，但采用不同的模型参数。

[0176] 第二训练单元05在对N个高阶模型进行选择和集成，得到查询改写判别模型时，可以直接利用这N个高阶模型进行集成，得到查询改写判别模型。也可以利用测试集对N个高阶模型的结果进行测试评分，测试集包含已确定改写评分的query对；依据测试评分选择其中P个高阶模型，P小于或等于N；对P个高阶模型进行加权处理，得到查询改写判别模型。

[0177] 图6为本发明实施例提供的判别查询改写的装置结构图，如图6所示，该装置包括：特征提取单元11和判别单元12，各组成单元的主要功能如下：

[0178] 特征提取单元11负责从待判别query对中提取特征，特征包括M个底层模型对该query对的评分，M为正整数。该底层模型为上述实施例训练得到的，将这M个评分作为特征，再进一步结合从该query对中提取的统计特征、距离特征、位置特征、词语重要性特征、语义特征以及同义词改写特征等。特征提取单元11提取的这部分特征与图5所示实施例中特征提取单元04提取的特征一致。

[0179] 判别单元12负责将特征提取单元11提取的特征输入查询改写判别模型，得到查询改写判别模型的判别结果。若查询改写判别模型是由多个高阶模型集成得到的，那么判别单元12实际就是将提取的特征分别输入各高阶模型，得到各高阶模型对该待判别query对的评分，然后依据各高阶模型的权值，对这些评分进行加权处理，例如加权求和或加权求平均，依据最终得到的评分来判别待判别query对中的一个query是否为另一个query的查询改写。

[0180] 本发明实施例提供的上述方法和装置，可以用于准确判别一个query是否可以用于另一个query的查询改写，其可以用于线下的改写词库的建立和优化，也可以用于线上进行query改写的判别和选择，还可以用于其他多种应用场景，本发明在此不再一一穷举。

[0181] 在本发明所提供的几个实施例中，应该理解到，所揭露的装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，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。

[0182] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0183] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

[0184] 上述以软件功能单元的形式实现的集成的单元，可以存储在一个计算机可读取存储介质中。上述软件功能单元存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)或处理器(processor)执行本发明各个实施例所述方法的部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0185] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

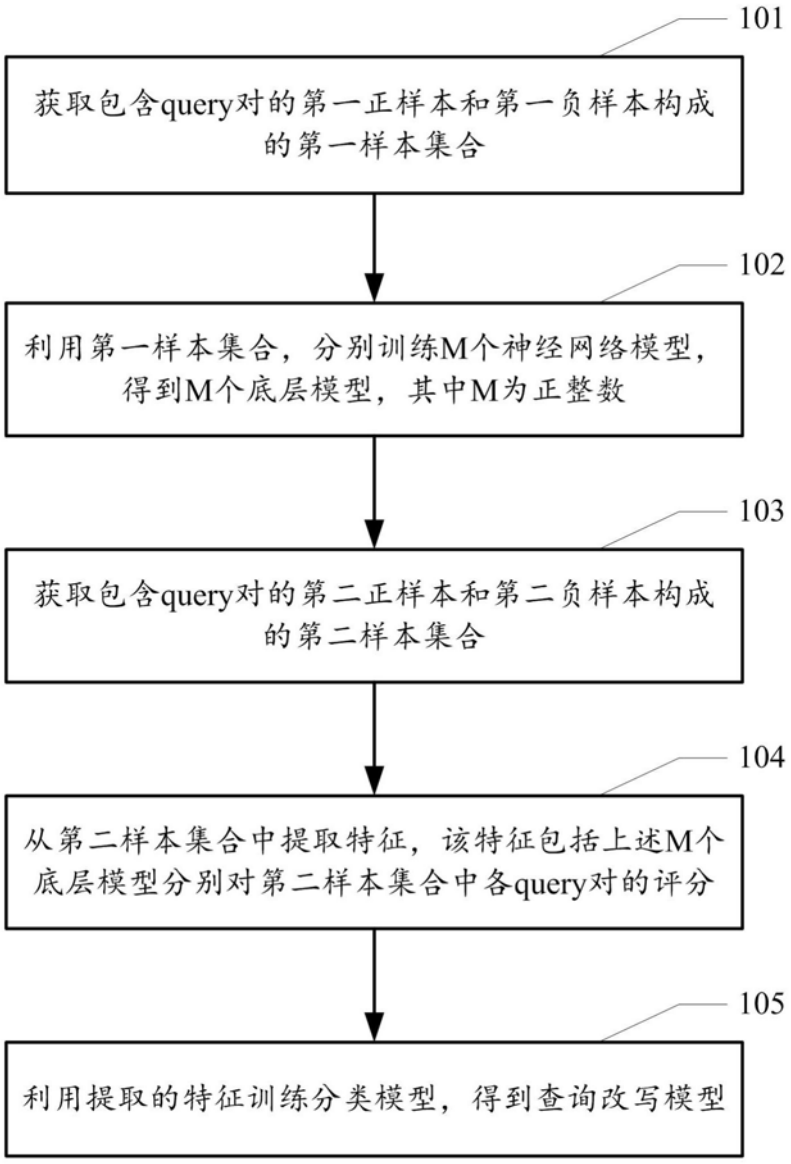

101
获取包含query对的第一正样本和第一负样本构成的第一样本集合
102
利用第一样本集合，分别训练M个神经网络模型，得到M个底层模型，其中M为正整数
103
获取包含query对的第二正样本和第二负样本构成的第二样本集合
104
从第二样本集合中提取特征，该特征包括上述M个底层模型分别对第二样本集合中各query对的评分
105
利用提取的特征训练分类模型，得到查询改写模型

图1

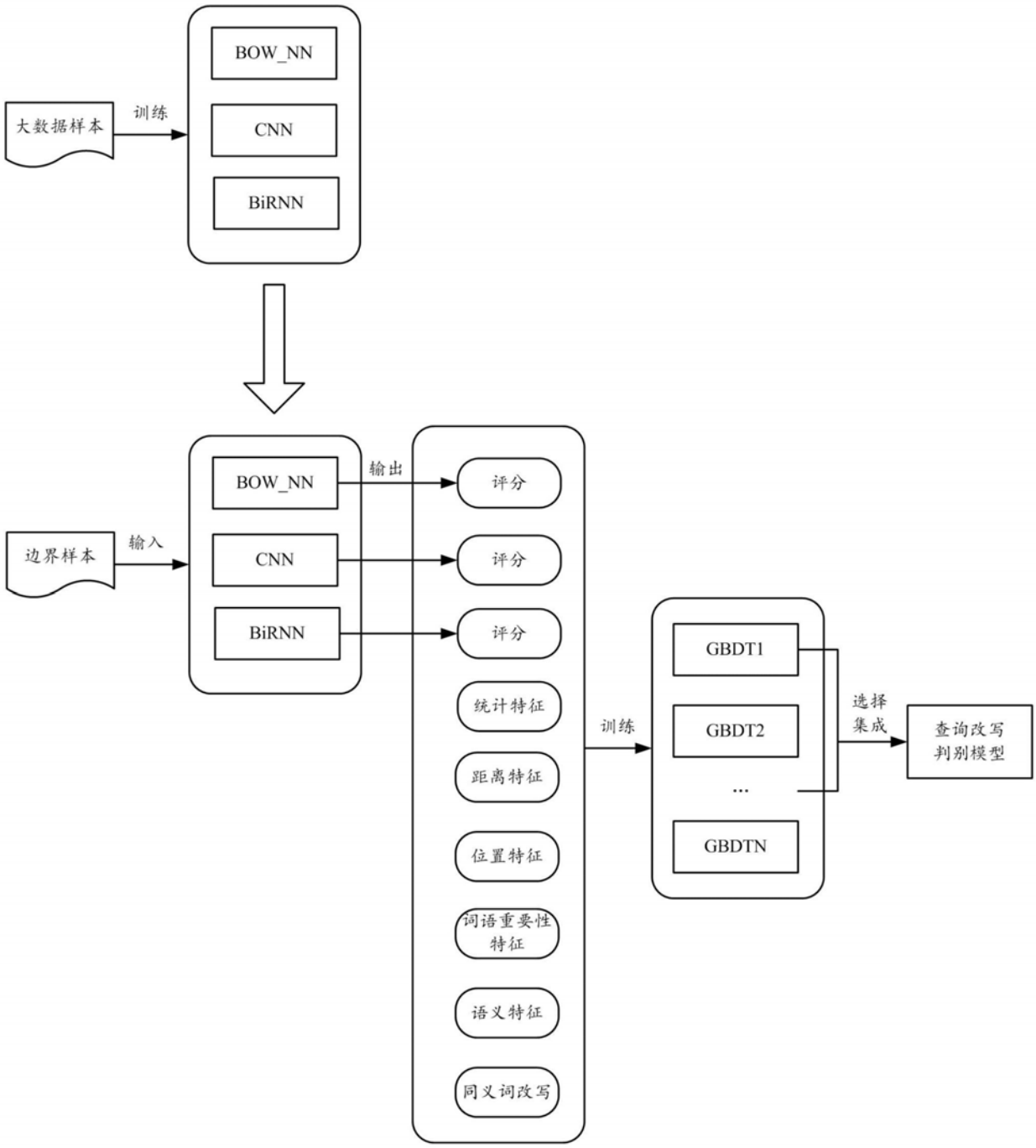

大数据样本
训练
BOW_NN
CNN
BiRNN
边界样本
输入
BOW_NN
CNN
BiRNN
输出
评分
评分
评分
统计特征
距离特征
位置特征
词语重要性特征
语义特征
同义词改写
训练
GBDT1
GBDT2
...
GBDTN
选择集成
查询改写判别模型

图2

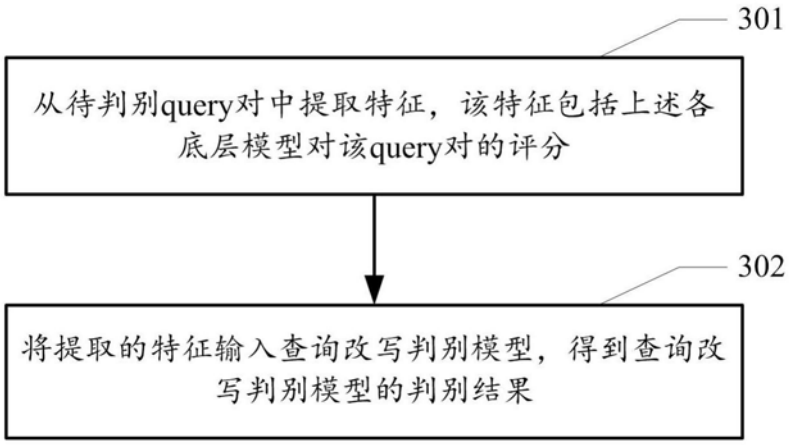

301
从待判别query对中提取特征，该特征包括上述各底层模型对该query对的评分
302
将提取的特征输入查询改写判别模型，得到查询改写判别模型的判别结果

图3

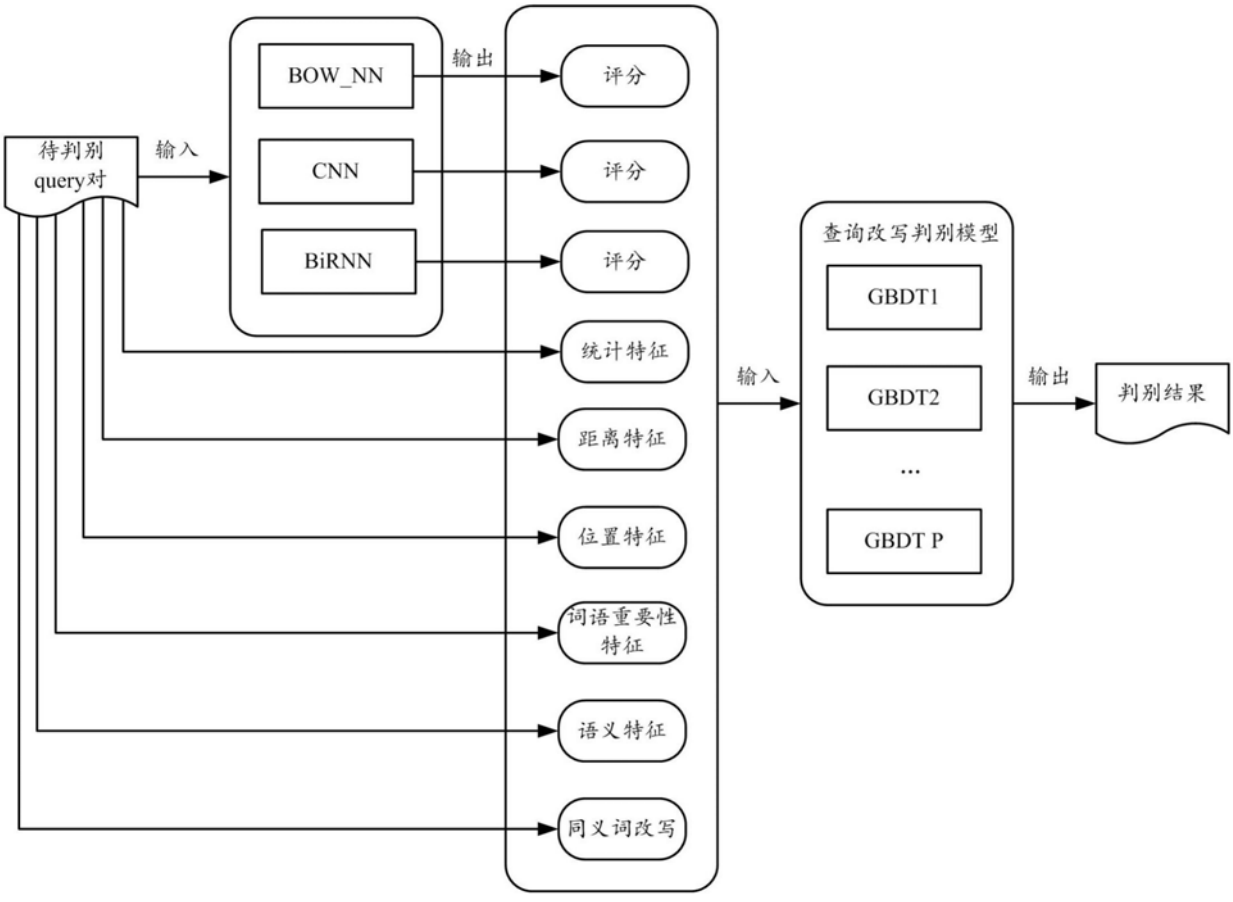

待判别query对
输入
BOW_NN
CNN
BiRNN
输出
评分
评分
评分
统计特征
距离特征
位置特征
词语重要性特征
语义特征
同义词改写
输入
查询改写判别模型
GBDT1
GBDT2
...
GBDT P
输出
判别结果

图4

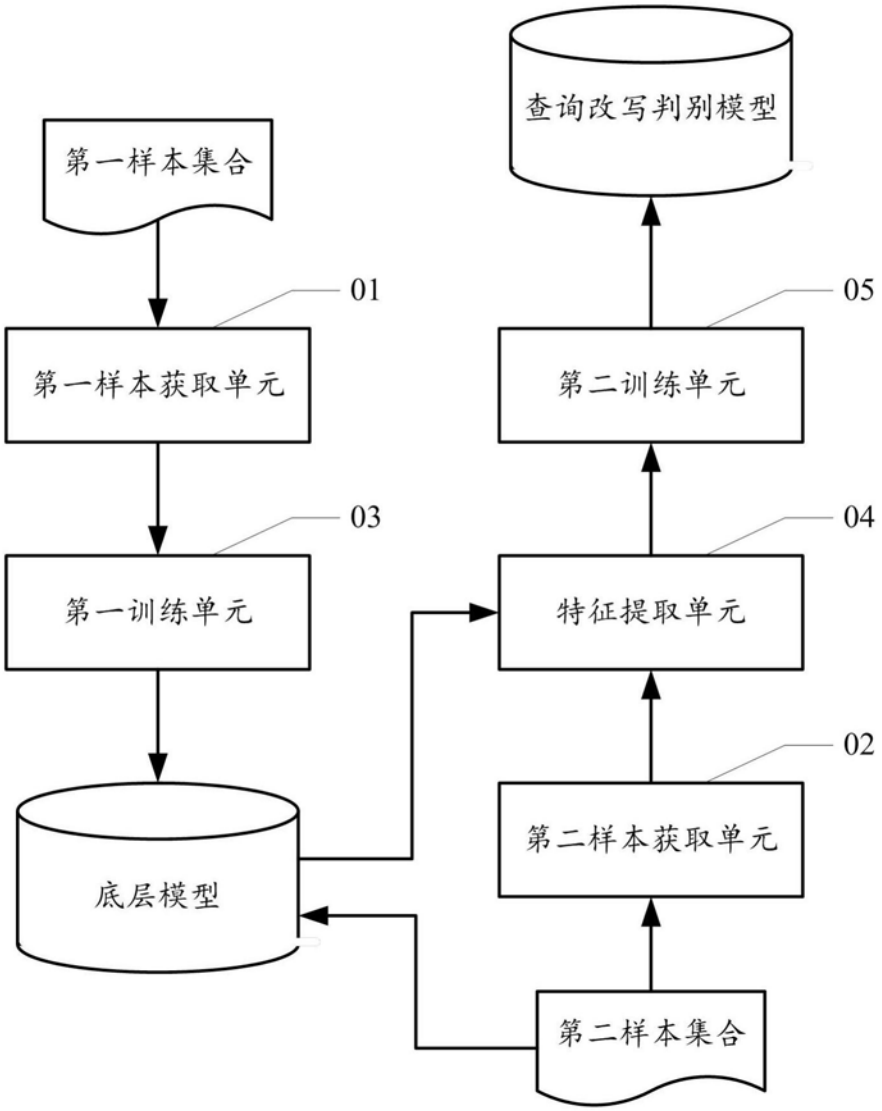

查询改写判别模型
第一样本集合
01
05
第一样本获取单元
第二训练单元
03
04
第一训练单元
特征提取单元
02
底层模型
第二样本获取单元
第二样本集合

图5

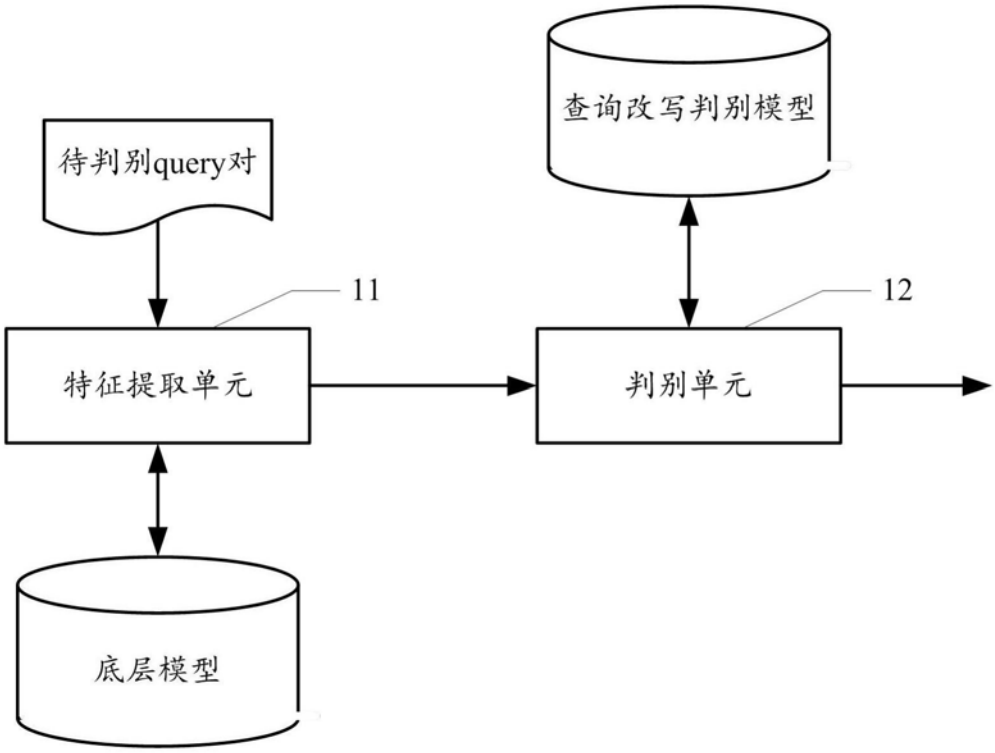


图6