



(12)发明专利

(10)授权公告号 CN 105488024 B

(45)授权公告日 2017.10.13

(21)申请号 201510818653.5

G06F 17/30(2006.01)

(22)申请日 2015.11.20

G06N 3/08(2006.01)

(65)同一申请的已公布的文献号

申请公布号 CN 105488024 A

(56)对比文件

CN 104298709 A, 2015.01.21, 说明书第
[0028]-[0047]段.

(43)申请公布日 2016.04.13

审查员 刘渊

(73)专利权人 广州神马移动信息科技有限公司

地址 510627 广东省广州市天河区黄埔大
道西平云路163号广电平云广场B塔12
层自编01单元

(72)发明人 李晨尧 曾洪雷

(74)专利代理机构 北京博雅睿泉专利代理事务

所(特殊普通合伙) 11442

代理人 马佑平 杨国权

(51)Int.Cl.

G06F 17/27(2006.01)

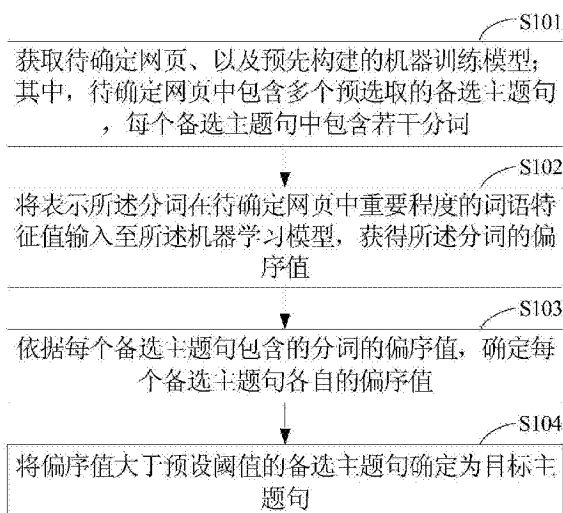
权利要求书3页 说明书10页 附图2页

(54)发明名称

网页主题句的抽取方法及装置

(57)摘要

本申请提供了一种网页主题句的抽取方法实施例,首先获取到待确定的网页,待确定网页中包含多个备选的主题句,备选主题句中包含若干分词,分别确定每个分词的词语特征值,将词语特征值输入至预设的机器学习模型中,从而获得分词的偏序值,进一步依据分词的偏序值,确定出备选主题句的偏序值,最终将偏序值大于预设阈值的备选主题句确定为目标主题句。可见,本实施例利用机器学习模型,来获得备选主题句的偏序值,由于机器学习模型可以反映查询语句与召回的网页之间的关联程度,因此,确定出的偏序值更加准确,从而提高了目标主题句的选取准确率。另外,本申请还提供了网页主题句的抽取装置,用以保证上述方法在实际中的应用及实现。



1. 一种网页主题句的抽取方法,其特征在于,包括:

获取待确定网页、以及预先构建的机器学习模型;其中,所述待确定网页中包含多个预选取的备选主题句,每个所述备选主题句中包含若干分词;

将表示所述分词在待确定网页中重要程度的词语特征值输入至所述机器学习模型,获得所述分词的偏序值;

依据每个所述备选主题句包含的分词的偏序值,确定每个所述备选主题句各自的偏序值;

将偏序值大于预设阈值的备选主题句确定为目标主题句;

其中,所述机器学习模型的构建方法具体包括:

获取若干网页样本的网页标题;其中,每个所述网页样本均对应若干用于召回网页样本自身的查询语句,每个所述查询语句均具有权重值,所述权重值用于表示查询语句与该查询语句召回的网页样本之间的关联程度;

针对每个所述网页标题,依据标注值公式

$$\text{label}(\text{term}_j) = \sum_{m=0}^k (v_m * w_m(\text{term}_j) / (\sum_{n=0}^s w_m(\text{term}_n)))$$
, 计算所述网页标题包含的每个分词的标注值;

其中,所述 term_j 表示网页标题中包含的第 j 个分词, $\text{label}(\text{term}_j)$ 表示分词 term_j 的标注值, m 表示某一个查询语句, k 表示所述查询语句的个数, v_m 表示查询语句 m 的权重值, $w_m(\text{term}_j)$ 表示分词 term_j 在查询语句 m 中的权重值, n 表示网页标题中的某一个分词, s 表示网页标题中分词的个数;

针对每个所述网页标题,确定所述网页标题包含的每个分词用于表示分词在网页样本中的重要程度的词语特征值;

依据预设的机器训练算法,对所述若干网页标题中每个分词的标注值及词语特征值进行训练,获得机器学习模型。。

2. 根据权利要求1所述的网页主题句的抽取方法,其特征在于,所述机器训练算法为GBRank训练算法,所述GBRank训练算法中的偏序比较方式为pairwise偏序比较,所述GBRank训练算法中的损失函数为交叉熵。

3. 根据权利要求1所述的网页主题句的抽取方法,其特征在于,所述依据每个所述备选主题句包含的分词的偏序值,确定每个所述备选主题句各自的偏序值,包括:

针对每个所述备选主题句,将所述备选主题句中分词的偏序值进行加权求和,获得所述备选主题句的偏序值。

4. 根据权利要求1~3任意一项所述的网页主题句的抽取方法,其特征在于,还包括:

建立所述待确定网页与所述目标主题句之间的对应关系;

在预设数据库中,保存所述对应关系。

5. 根据权利要求4所述的网页主题句的抽取方法,其特征在于,还包括:

响应于用户在搜索引擎中输入查询语句的操作,确定所述搜索引擎召回的多个召回网页;其中,所述召回网页包含在所述待确定网页中;

在所述预设数据库中,确定每个所述召回网页的目标主题句,并分别将所述输入的查询语句与每个所述召回网页的目标主题句进行匹配,获得匹配值;

依据所述匹配值的大小排序,在所述搜索引擎中,依次排列各个所述召回网页。

6.一种网页主题句的抽取装置,其特征在于,包括:

网页及模型获取模块,用于获取待确定网页、以及预先构建的机器学习模型;其中,所述待确定网页中包含多个预选取的备选主题句,每个所述备选主题句中包含若干分词;

训练模型构建模块,用于预先构建机器学习模型;

其中,所述训练模型构建模块包括:

网页样本标题获取子模块,用于获取若干网页样本的网页标题;其中,每个所述网页样本均对应若干用于召回网页样本自身的查询语句,每个所述查询语句均具有权重值,所述权重值用于表示某一查询语句与该某一查询语句召回的网页样本之间的关联程度,

标题分词标注值计算子模块,用于针对每个所述网页标题,依据标注值公式

$$\text{label}(\text{term}_j) = \sum_{m=0}^k (v_m * w_m(\text{term}_j) / (\sum_{n=0}^s w_m(\text{term}_n)))$$
, 计算所述网页标题包含的每个分词的标注值,

其中,所述 term_j 表示网页标题中包含的第 j 个分词, $\text{label}(\text{term}_j)$ 表示分词 term_j 的标注值, m 表示某一个查询语句, k 表示所述查询语句的个数, v_m 表示查询语句 m 的权重值, $w_m(\text{term}_j)$ 表示分词 term_j 在查询语句 m 中的权重值, n 表示网页标题中的某一个分词, s 表示网页标题中分词的个数,

标题分词特征值计算子模块,用于针对每个所述网页标题,确定所述网页标题包含的每个分词的词语特征值;其中,所述词语特征值用于表示分词在网页样本中的重要程度,

机器训练子模块,用于依据预设的机器训练算法,对所述若干网页标题中每个分词的标注值及词语特征值进行训练,获得机器学习模型;

分词偏序值确定模块,用于将表示所述分词在待确定网页中重要程度的词语特征值输入至所述机器学习模型,获得所述分词的偏序值;

主题句偏序确定模块,用于依据每个所述备选主题句包含的分词的偏序值,确定每个所述备选主题句各自的偏序值;

目标主题句确定模块,用于将偏序值大于预设阈值的备选主题句确定为目标主题句。

7.根据权利要求6所述的网页主题句的抽取装置,其特征在于,所述机器训练子模块训练使用的机器训练算法为GBRank训练算法,所述GBRank训练算法中的偏序比较方式为pairwise偏序比较,所述GBRank训练算法中的损失函数为交叉熵。

8.根据权利要求6所述的网页主题句的抽取装置,其特征在于,所述主题句偏序确定模块包括:

加权求和确定偏序子模块,用于针对每个所述备选主题句,将所述备选主题句中分词的偏序值进行加权求和,获得所述备选主题句的偏序值。

9.根据权利要求6~8任意一项所述的网页主题句的抽取装置,其特征在于,还包括:

主题句保存模块,用于建立所述待确定网页与所述目标主题句之间的对应关系;在预设数据库中,保存所述对应关系。

10.根据权利要求9所述的网页主题句的抽取装置,其特征在于,还包括:

召回网页确定模块,用于响应于用户在搜索引擎中输入查询语句的操作,确定所述搜索引擎召回的多个召回网页;其中,所述召回网页包含在所述待确定网页中;

网页主题句匹配模块,用于在所述预设数据库中,确定每个所述召回网页的目标主题句,并分别将所述输入的查询语句与每个所述召回网页的目标主题句进行匹配,获得匹配值;

召回网页排序模块,用于依据所述匹配值的大小排序,在所述搜索引擎中,依次排列各个所述召回网页。

网页主题句的抽取方法及装置

技术领域

[0001] 本申请涉及互联网应用技术领域,更具体地,是一种网页主题句的抽取方法及装置。

背景技术

[0002] 随着互联网技术的快速发展,互联网已经成为人们获得信息的重要渠道。具体地,信息查询用户可以在搜索引擎中输入某个查询词语,搜索引擎便召回多个网页,以供用户进行选择查看。需要说明的是,为了方便用户的查看,召回的网页是按照与查询语句的相关度,依次排列在搜索引擎中的。

[0003] 其中,相关度即召回网页的主题句与查询语句的相似度。例如,查询语句为“乙肝的症状”,召回网页1的主题句为“乙肝的症状有哪些”,召回网页2的主题句为“乙肝病毒传播途径”。召回网页1的主题句与查询语句更相似,因此,召回网页1与查询语句更相关,进而便会排列在搜索结果的较前位置。可见,网页主题句会直接影响召回网页的排列顺序,进而影响用户对搜索结果的满意度。

[0004] 目前,网页主题句的抽取方式是人工根据一些任意的网页总结抽取规则,依据该抽取规则来确定某个待确定网页的主题句。然而,该种方式确定的主题句准确率较低。

发明内容

[0005] 有鉴于此,本申请提供了一种网页主题句的抽取方法,用以解决技术中确定的主题句准确率较低的技术问题。另外,本申请还提供了一种网页主题句的抽取装置,用以保证所述方法在实际中的应用及实现。

[0006] 为实现所述目的,本申请提供的技术方案如下:

[0007] 本申请的第一方面提供了一种网页主题句的抽取方法,包括:

[0008] 获取待确定网页、以及预先构建的机器学习模型;其中,所述待确定网页中包含多个预选取的备选主题句,每个所述备选主题句中包含若干分词;

[0009] 将表示所述分词在待确定网页中重要程度的词语特征值输入至所述机器学习模型,获得所述分词的偏序值;

[0010] 依据每个所述备选主题句包含的分词的偏序值,确定每个所述备选主题句各自的偏序值;

[0011] 将偏序值大于预设阈值的备选主题句确定为目标主题句。

[0012] 本申请的第二方面提供了一种网页主题句的抽取装置,包括:

[0013] 网页及模型获取模块,用于获取待确定网页、以及预先构建的机器学习模型;其中,所述待确定网页中包含多个预选取的备选主题句,每个所述备选主题句中包含若干分词;

[0014] 分词偏序值确定模块,用于将表示所述分词在待确定网页中重要程度的词语特征值输入至所述机器学习模型,获得所述分词的偏序值;

[0015] 主题句偏序确定模块,用于依据每个所述备选主题句包含的分词的偏序值,确定每个所述备选主题句各自的偏序值;

[0016] 目标主题句确定模块,用于将偏序值大于预设阈值的备选主题句确定为目标主题句。

[0017] 由以上可知,本申请具有如下优点:

[0018] 本申请提供的网页主题句的抽取方法实施例,首先获取到待确定的网页,待确定网页中包含多个备选的主题句,备选主题句中包含若干分词,分别确定每个分词的词语特征值,将词语特征值输入至预设的机器学习模型中,从而获得分词的偏序值,进一步依据分词的偏序值,确定出备选主题句的偏序值,最终将偏序值大于预设阈值的备选主题句确定为目标主题句。可见,本实施例利用机器学习模型,来获得备选主题句的偏序值,由于机器学习模型可以反映查询语句与召回的网页之间的关联程度,因此,确定出的偏序值更加准确,从而提高了目标主题句的选取准确率。

附图说明

[0019] 为了更清楚地说明本申请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,下面描述中的附图仅仅是本申请的实施例,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,还可以根据提供的附图获得其他的附图。

[0020] 图1为本申请提供的网页主题句的抽取方法实施例1的流程图;

[0021] 图2为本申请提供的机器学习模型的构建方法流程图;

[0022] 图3为本申请提供的网页主题句的抽取装置实施例1的结构框图;

[0023] 图4为本申请提供的网页主题句的抽取装置实施例2的结构框图。

具体实施方式

[0024] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本申请一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0025] 参照图1,其示出了本申请提供的网页主题句的抽取方法实施例1的流程。如图1所示,该方法实施例1可以具体包括步骤S101~步骤S104。

[0026] 步骤S101:获取待确定网页、以及预先构建的机器学习模型;其中,待确定网页中包含多个预选取的备选主题句,每个备选主题句中包含若干分词。

[0027] 在实施前,预先构建有机器学习模型,如有监督的机器学习模型。机器学习模型是利用训练方法构建的模型,训练方法使用的训练样本为查询语句召回的多个网页文件。具体地,用户在搜索引擎中输入某个查询语句,搜索引擎会召回与该查询语句相关的网页文件,若用户发现召回的网页文件中,某个(些)网页文件与该查询语句的关联性较大,则会点击查看该某个(些)网页文件。通过监测用户的点击行为,可以收集到各个网页文件的被点击次数,并进一步将该点击次数换算为权重值。具体地,可以收集到各个网页文件对应的查询语句,以及在这个查询语句下,网页文件的被点击次数,并进一步将该点击次数换算为网

页文件在该查询语句下的权重值。

[0028] 可见,机器学习模型的训练方法使用的训练样本,即多个网页文件,具有各自的权重值,该权重值是通过监测用户对该网页文件的点击行为获得的,因此,权重值可以表示出查询语句与召回的网页文件之间的关联程度,或者说,权重值可以反映用户对召回的网页文件的满意度。

[0029] 需要说明的是,构建机器学习模型的目的在于,利用大规模的点击样本对网页文件中的关键词进行标注,从而使得机器学习程序可以从大规模用户点击样本数据中学习词语在网页中的重要程度的规律,并以模型的形式进行表达。

[0030] 获取到的待确定网页可以是使用网络爬虫工具,从互联网等网络中抓取到的网页文件。待确定网页为至少一个,可选地,待确定网页为多个。针对每个待确定网页均按照本实施例的方式,来确定每个待确定网页各自的目标主题句。

[0031] 需要说明的是,待确定网页中包含大量的语句,在大量的语句中选取一些作为备选主题句。再按照以下步骤,在备选主题句中,确定出最终的目标主题句。这样,可以提高网页文件主题句的确定效率。

[0032] 可选地,选取的备选主题句为最可能包含目标主题句的语句,例如,网页标题(title)、文章标题(article title)、外部链接(anchor)、网站面包屑(breadcrumb)及元标签(meta)等域中最可能包含整个网页内容的关键字,因此,在这些域中提取备选主题句。其中,网页标题可以具体为经过标点切分的网页标题。当然,此些域仅仅是示例说明,本申请可以包含但不局限于此些域,还可以是其他。

[0033] 在获得备选主题句后,进一步,可以使用分词工具,对备选主题句进行切分,从而获得多个分词,该些分词也可以称为备选主题词。例如,备选主题句为“乙肝导致发热”,分词分别为“乙肝”、“导致”及“发热”。

[0034] 步骤S102:将表示所述分词在待确定网页中重要程度的词语特征值输入至所述机器学习模型,获得所述分词的偏序值。

[0035] 其中,首先依据预设的特征值确定算法,确定每个分词的词语特征值;其中,词语特征值用于表示分词在待确定网页中的重要程度。

[0036] 具体地,预先设置多个特征值的确定算法,分别用来确定不同种类的词语特征值。具体地,想要计算分词在哪些方面的词语特征,则针对该词语特征,设置相对应的特征值确定算法。

[0037] 需要说明的是,词语特征可以表示某个词语在待确定网页中的重要程度。可选,词语特征为词频(term frequency,TF)、逆向文件频率(inverse document frequency,IDF)、连续命中率、紧密度(proximity)、网页标题的长度、网页类型等任意一种或多种的组合。

[0038] 其中,词频的确定算法为,统计待确定网页的词语总数、以及分词在待确定网页中出现的次数,将分词的出现次数除以词语总数;逆向文件频率的确定算法为,待确定网页的段落总数除以包含分词的段落数,再将得到的商取对数;连续命中率的确定算法为,设置一个窗口,窗口是以分词为开始且长度为预设长度的分词串,计算该窗口在待确定网页中的命中频率;紧密度的确定算法为,计算分词与分词之间的语法关联程度;网页标题的长度的确定算法为,统计网页标题中文字的个数;网页类型的确定算法为,将网页内容的类型确定为网页类型。

[0039] 当然,以上词语特征仅仅是示例说明,本申请并不局限于此,还可以是其他类型的词语特征。

[0040] 确定分词的词语特征值后,将每个分词的词语特征值输入至机器学习模型,获得分词的偏序值。

[0041] 具体地,机器学习模型的输入为备选主题句中分词的词语特征值,输出为分词的偏序值。具体的,针对备选主题句中的每个分词,将该分词的词语特征值输入至机器学习模型中,机器学习模型可以依据词语特征值,计算该分词的偏序值。这样,便获得备选主题句中每个分词的偏序值。

[0042] 步骤S103:依据每个备选主题句包含的分词的偏序值,确定每个备选主题句各自的偏序值。

[0043] 其中,备选主题句中包括多个分词,每个分词具有各自的偏序值。利用分词的偏序值,可以获得备选主题句的偏序值。具体地,确定方式可以是直接将分词的偏序值相加,或者,可以是将分词的偏序值加权求和,具体加权求和过程可以参见下文。

[0044] 需要说明的是,待确定网页中包含多个备选主题句,按照以上方式,分别确定每个备选主题句的偏序值。

[0045] 步骤S104:将偏序值大于预设阈值的备选主题句确定为目标主题句。

[0046] 其中,预先设置偏序阈值,将偏序值大于该预设偏序阈值的备选主题句,确定为目标主题句。

[0047] 由以上的技术方案可知,本实施例提供的网页主题句的抽取方法,首先获取到待确定的网页,待确定网页中包含多个备选的主题句,备选主题句中包含若干分词,分别确定每个分词的词语特征值,将词语特征值输入至预设的机器学习模型中,从而获得分词的偏序值,进一步依据分词的偏序值,确定出备选主题句的偏序值,最终将偏序值大于预设阈值的备选主题句确定为目标主题句。可见,本实施例利用机器学习模型,来获得备选主题句的偏序值,由于机器学习模型训练过程使用的网页样本可以反映与查询语句之间的关联程度,因此,获得的机器学习模型更加准确,从而也就提高了目标主题句的选取准确率。

[0048] 具体来讲,现有技术中,是人工根据一些任意的网页总结抽取规则,依据该抽取规则来确定某个待确定网页的主题句,确定出的主题句准确率较低。原因主要是:人工总结抽取规则使用的网页是任意选择的网页,生成的抽取规则也只是简单条件的累加,例如什么样的条件下,什么样的词重要,或者不重要,这种简单条件很难准确表示词语在一个网页中的重要程度。

[0049] 然而,本实施例中构建的机器学习模型,使用的网页样本是用户使用查询语句召回的网页,并且网页样本可以反映与查询语句之间的关联程度,其中,关联程度可以具体使用用户对召回网页的投票行为表示。可见,本实施例构建的机器学习模型由于使用的以上特征的网页样本,比现有技术中的抽取规则更加准确,因此,本实施例确定的主题句更加准确。

[0050] 在实施中,为了进一步提高目标主题句的确定准确率,还可以根据待确定网页的外部链接,来筛选目标主题句。具体地,在执行步骤S105之前,获取待确定网页的外部链接,外部链接为一段文字内容,判断外部链接与各个备选主题句之间的相似度,根据相似度大小,对各个备选主题句的偏序值进行惩罚,惩罚因子为大于0小于等于1的数值,且相似度越

大,惩罚因子越大。然后,用惩罚后的偏序值执行步骤S105,即将惩罚后的偏序值大于预设阈值的备选主题句确定为目标主题句。

[0051] 例如,待确定网页的备选主题句有两个,分别为备选主题句1“乙肝导致发热”及备选主题句2“乙肝是什么,发热如何降温”。两个备选主题句的偏序值分别为0.5及0.7。

[0052] 待确定网页的外部链接为“乙肝导致发热”,显而易见,备选主题句1与外部链接的相似度更高,因此,备选主题句1的惩罚因子为1,备选主题句2的惩罚因子为0.5。分别对两个备选主题句的偏序值进行惩罚: 0.5×1 , 0.7×0.5 ,即惩罚后的偏序值分别为0.5及0.35。

[0053] 下面对机器学习模型的构建方法进行详细说明。

[0054] 如图2所示,机器学习模型:构建方法的具体流程可以包括步骤S201~步骤S204。

[0055] 步骤S201:获取若干网页样本的网页标题;其中,每个网页样本均对应若干用于召回网页样本自身的查询语句,每个查询语句均具有权重值,权重值用于表示某一查询语句与该某一查询语句召回的网页样本之间的关联程度。

[0056] 其中,机器学习模型的构建需要训练样本,训练样本即本步骤中的网页样本。这些网页样本是从互联网等网络中抓取到的。

[0057] 需要说明的是,为了提高用户对确定的目标主题句的满意度,在构建机器学习模型时,使用的网页样本需要具有权重值,这些权重值可以反映查询语句与召回网页之间的关联程度,该关联程度是可以反映用户的满意度的。

[0058] 具体地,在搜索引擎中输入某个查询语句,该查询语句召回的多个网页样本。这些网页样本的标题可以显示在搜索引擎中。若用户对某个或某些网页样本的内容满意,则可以对该个或这些网页样本进行投票,即点击该个或这些网页样本的标题。因此,可以监测用户对网页样本的点击次数,来确定网页样本的权重值。其中,权重值可以是点击次数除以某个预设值获得的。

[0059] 以上以查询语句的角度,来说明网页样本的权重值。当然,还可以以网页样本的角度,对其进行说明。

[0060] 具体地,某个查询语句可以召回多个网页样本,例如,查询语句 q_i 召回的网页样本的网页标题为 $\{t_{i1}, t_{i2}, \dots, t_{ik}\}$,这些网页样本的权重值分别为 $\{v_{i1}, v_{i2}, \dots, v_{ik}\}$,即查询语句 q_i 可以表示为 $\{(q_i, ((t_{i1}, v_{i1}), (t_{i2}, v_{i2}), \dots, (t_{ik}, v_{ik})))\}$ 。

[0061] 反过来讲,某个网页样本也可以对应多个查询语句,若在这些查询语句下,用户均点击了该网页样本,则表明该网页样本很好地符合这些查询语句的查询需求。因此,该网页样本的主题句可以使用查询语句中的词语来描述。

[0062] 将上述查询语句的数据关系反转后,获得的网页样本的数据关系为 $\{(t_m, ((q_{m1}, v_{m1}), (q_{m2}, v_{m2}), \dots, (q_{mk}, v_{mk})))\}$ 。其中, t_m 为网页标题 $\{t_{i1}, t_{i2}, \dots, t_{ik}\}$ 中的任意一个, $\{q_{m1}, q_{m2}, \dots, q_{mk}\}$ 为多个查询语句, $\{v_{m1}, v_{m2}, \dots, v_{mk}\}$ 为以上网页样本的权重值 $\{v_{i1}, v_{i2}, \dots, v_{ik}\}$ 。

[0063] 从上述获得的网页样本的数据关系 $\{(t_m, ((q_{m1}, v_{m1}), (q_{m2}, v_{m2}), \dots, (q_{mk}, v_{mk})))\}$ 中可以看出,任意一个网页样本 t_m 均对应多个查询语句,分别为 $\{q_{m1}, q_{m2}, \dots, q_{mk}\}$,在该些查询语句下,该网页样本 t_m 的权重值分别为 $\{v_{m1}, v_{m2}, \dots, v_{mk}\}$ 。

[0064] 步骤S202:针对每个网页标题,依据标注值公式

$label(term_j) = \sum_{m=0}^k (v_m * w_m(term_j) / (\sum_{n=0}^s w_m(term_n)))$, 计算网页标题包含的每个分词的标注值;

[0065] 其中, $term_j$ 表示网页标题中包含的第 j 个分词, $label(term_j)$ 表示分词 $term_j$ 的标注值, m 表示某一个查询语句, k 表示查询语句的个数, v_m 表示查询语句 m 的权重值, $w_m(term_j)$ 表示分词 $term_j$ 在查询语句 m 中的权重值, n 表示网页标题中的某一个分词, s 表示网页标题中分词的个数。

[0066] 具体地, 网页样本为多个, 则网页标题也为多个。网页标题中包含多个分词, 针对每个网页标题, 均依据该标注值计算公式, 计算该网页标题包含的每个分词的标注值。需要说明的是, 标注值公式中的 $w_m(term_j)$ 可以由分词工具计算获得。

[0067] 例如, 某个网页样本的网页标题为“乙肝会导致发热吗”, 在监测到的点击数据中, 该网页样本在查询语句1“乙肝导致发热”及查询语句2“乙肝的症状”中被点击。根据点击数据, 计算两个查询语句的权重值 v_m 分别为1.5及0.5。

[0068] 使用分词工具, 对查询语句1“乙肝导致发热”中的分词计算权重值后可知: “乙肝”的权重值 $w_m(term_j)$ 为97, “导致”的权重值 $w_m(term_j)$ 为73, “发热”的权重值 $w_m(term_j)$ 为85。

[0069] 使用分词工具, 对查询语句2“乙肝的症状”中的分词计算权重值后可知: “乙肝”的权重值 $w_m(term_j)$ 为105, “的”的权重值 $w_m(term_j)$ 为5, “症状”的权重值 $w_m(term_j)$ 为85。

[0070] 基于上述数据, 对网页样本“乙肝会导致发热吗”包含的分词的标注值进行计算。具体地, 该网页样本包含的分词分别为“乙肝”、“会”、“导致”、“发热”。其中,

[0071] 分词“乙肝”的标注值为: $1.5*97/(97+73+85)+0.5*105/(105+5+85)=0.839$; 分词“会”的标注值为: $1.5*0/(97+73+85)+0.5*0/(105+5+85)=0$; 分词“导致”的标注值为: $1.5*73/(97+73+85)+0.5*0/(105+5+85)=0.42$; 分词“发热”的标注值为: $1.5*85/(97+73+85)+0.5*0/(105+5+85)=0.5$ 。

[0072] 经过上述计算, 针对每个网页样本, 可以获得该网页样本的网页标题包含的每个分词对应的标注值。使用数据关系表示网页样本的各个标注值, 可以具体为 $\{(doc_j, ((term_{j1}, l_{j1}), (term_{j2}, l_{j2}), \dots, (term_{jk}, l_{jk})))\}$, 其中, doc_j 表示网页样本, $\{term_{j1}, term_{j2}, \dots, term_{jk}\}$ 表示该网页样本的各个分词, $\{l_{j1}, l_{j2}, \dots, l_{jk}\}$ 表示各个分词的标注值。

[0073] 步骤S203: 针对每个网页标题, 确定网页标题包含的每个分词的词语特征值; 其中, 词语特征值用于表示分词在网页样本中的重要程度。

[0074] 其中, 网页标题的词语特征值的确定方法可以参见上述对网页主题句的抽取方法实施例1中步骤S102的说明, 此处并不赘述。

[0075] 步骤S204: 依据预设的机器训练算法, 对若干网页标题中每个分词的标注值及词语特征值进行训练, 获得机器学习模型。

[0076] 其中, 机器训练算法中包含未知量, 预先设置未知量的初始值, 将分词的词语特征值输入设置初始值的机器训练算法后, 可以获得分词标注值的预测值, 将标注值的预测值与步骤S202计算获得的实际标注值进行偏序比较, 根据比较结果, 对预先设置的初始值不断进行修正, 从而获得未知量的最终值。将未知量的最终值代入机器训练算法中, 便获得了机器学习模型。

[0077] 可选的, 机器训练算法为GBRank训练算法, GBRank训练算法中的偏序比较方式为

pairwise偏序比较,GBRank训练算法中的损失函数为交叉熵。

[0078] 需要说明的是,GBRank的训练过程中包含偏序比较过程,偏序比较使用的是pairwise偏序比较,即将两个分词的词语特征值作为输入,获得两个分词标注值的预测值,将两个分词的标注值的预测值与实际值进行偏序比较。

[0079] 偏序比较使用的公式为 $P_{ij} = P(U_i > U_j) = \frac{1}{1 + e^{-\sigma(s_i - s_j)}}$, 该偏序比较公式经过变换后, 最终为: $\frac{\partial C}{\partial s_i} = \sigma \left(\frac{1}{2}(1 - S_{ij}) - \frac{1}{1 + e^{\sigma(s_i - s_j)}} \right) = \frac{\partial C}{\partial s_j}$; 其中, P_{ij} 表示两个分词中一个分词比另一分词更好的概率; e 为预设固定值; σ 为预设固定值, 通常设置为 0.1, 当然, 还可以设置为其他值, 该值越大表示应该产生的误差也越大; s_i 表示分词 $term_i$ 的标注值的预测值; s_j 表示分词 $term_j$ 的标注值的预测值。

[0080] 损失函数的公式为 $C = -\bar{P}_{ij} \log \bar{P}_{ij} - (1 - \bar{P}_{ij}) \log(1 - \bar{P}_{ij})$; 其中, 令 $\bar{P}_{ij} = \frac{1}{2}(1 + S_{ij})$, 则:

$$[0081] \quad C = -\bar{P}_{ij} \log(1 / (1 + e^{-\sigma(s_i - s_j)})) - (1 - \bar{P}_{ij}) \log(e^{-\sigma(s_i - s_j)} / (1 + e^{-\sigma(s_i - s_j)}))$$

[0082]

$$C = -\bar{P}_{ij} \log(1 / (1 + e^{-\sigma(s_i - s_j)})) - \log(e^{-\sigma(s_i - s_j)} / (1 + e^{-\sigma(s_i - s_j)})) + \bar{P}_{ij} \log(e^{-\sigma(s_i - s_j)} / (1 + e^{-\sigma(s_i - s_j)}))$$

$$[0083] \quad C = \log((1 / (1 + e^{-\sigma(s_i - s_j)}))^{-\bar{P}_{ij}} * (e^{-\sigma(s_i - s_j)} / (1 + e^{-\sigma(s_i - s_j)}))^{\bar{P}_{ij}} / (e^{-\sigma(s_i - s_j)} / (1 + e^{-\sigma(s_i - s_j)})))$$

$$[0084] \quad C = \log((e^{-\sigma(s_i - s_j)})^{\bar{P}_{ij}-1} * (1 + e^{-\sigma(s_i - s_j)}))$$

$$[0085] \quad C = \frac{1}{2}(1 - S_{ij}) \log(e) \sigma(s_i - s_j) + \log(1 + e^{-\sigma(s_i - s_j)})$$

[0086] 由于 σ 和 $\log_e$ 都是常数, 因此, 损失函数可以变换为以下公式:

$C = \frac{1}{2}(1 - S_{ij}) \sigma(s_i - s_j) + \log(1 + e^{-\sigma(s_i - s_j)})$; 其中, S_{ij} 是由两个分词的标注值的大小关系计算获得的, 若分词 $term_i$ 的标注值大于分词 $term_j$, 则 S_{ij} 为 1; 若分词 $term_i$ 的标注值小于分词 $term_j$, 则 S_{ij} 为 -1; 若分词 $term_i$ 的标注值等于分词 $term_j$, 则 S_{ij} 为 0。需要说明的是, 该公式中的其他参数可以参见上述说明, 此处并不赘述。

[0087] 经过以上机器训练过程, 最终计算获得的机器学习模型为: 树形结构的机器学习模型, 可以认为是一组判断条件组成的树形结构, 具体结构如决策树。

[0088] 以上网页主题句的抽取方法实施例 1 中, 步骤 S104 (依据每个备选主题句包含的分词的偏序值, 确定每个备选主题句各自的偏序值) 的一种具体实现方式包括:

[0089] 针对每个备选主题句, 将备选主题句中分词的偏序值进行加权求和, 获得备选主题句的偏序值。

[0090] 其中, 备选主题句中包含多个分词, 每个分词均具有偏序值, 为每个分词设置相对应的权重值, 将分词的权重值与偏序值相乘后进行求和, 从而得到该备选主题句的偏序值。

[0091] 在实施中, 可以将网页的目标主题句保存在数据库中, 以供网页搜索时, 对网页进行排序。具体的保存步骤包括:

[0092] 建立待确定网页与目标主题句之间的对应关系; 在预设数据库中, 保存对应关系。

[0093] 具体地,待确定网页可以是多个,每个待确定网页的目标主题句可以是一个,也可以是多个。对于每个待确定网页,建立该待确定网页与自身的目标主题句的对应关系,因此,每个待确定网页的对应关系可以是一个,也可以是多个。

[0094] 需要说明的是,待确定网页可以用网页的唯一性标识表示,如网址URL (Uniform Resoure Locator,统一资源定位器)。因此,数据库中保存的是多个网址URL,及每个网址URL对应的目标主题句。

[0095] 例如,待确定网页为“新浪体育”,该网页的网址URL为“url=sport.sina.com”,为网页确定的目标主题句有两个,分别为“新浪竞技风暴”及“新浪体育”,因此,预设数据库中保存网址URL“url=sport.sina.com”,及该网址URL对应的目标主题句“新浪竞技风暴”及“新浪体育”。

[0096] 建立的数据库,可以用于对用户搜索时,搜索到的网页进行排序。具体如下步骤A1~步骤A3。

[0097] 步骤A1:响应于用户在搜索引擎中输入查询语句的操作,确定搜索引擎召回的多个召回网页;其中,召回网页包含在待确定网页中。

[0098] 其中,用户想要搜索某方面的信息时,可以在搜索引擎中输入与该信息相关的某个查询语句,搜索引擎可以召回相关的多个网页。为了便于描述,可以将召回的网页称为召回网页。

[0099] 需要说明的是,上述待确定网页可以是多个,召回网页为多个,这些召回网页包含在待确定网页中,以上网页主题确定方法实施例可以确定出待确定网页的目标主题句并保存在预设数据库中,也就是说,预设数据库中也保存有此些召回网页的目标主题句。

[0100] 步骤A2:在预设数据库中,确定每个召回网页的目标主题句,并分别将输入的查询语句与每个召回网页的目标主题句进行匹配,获得匹配值。

[0101] 其中,在预设数据库中,查找每个召回网页的目标主题句。预设数据库中保存的召回网页可以是召回网页的网址URL,因此,查找目标主题句的方式可以是,查找到召回网页的网址URL,再查找该网址URL所对应的目标主题句。

[0102] 将用户输入的查询语句分别与召回网页的目标主题句进行匹配,获得匹配值。需要说明的是,若召回网页有多个目标主题句,则分别将查询语句与目标主题句进行匹配,获得每个目标主题句各自的匹配值,可以将最大的匹配值作为查询语句与该召回网页的匹配值。

[0103] 例如,用户在搜索引擎中输入的查询语句为“新浪体育”,搜索引擎召回的某一个网页的网址URL为“url=sport.sina.com”。在预设数据库中查找到该网址URL对应的目标主题句为两个,分别为“新浪竞技风暴”及“新浪体育”。

[0104] 将查询语句“新浪体育”分别与目标主题句“新浪竞技风暴”及“新浪体育”进行匹配,两个匹配的结果分别为0.8及1,则可以将最大值1作为查询语句“新浪体育”与召回网页“url=sport.sina.com”的匹配值。

[0105] 召回网页为多个,则按照上述方法分别计算每个召回网页与查询语句之间的匹配值。

[0106] 步骤A3:依据匹配值的大小排序,在搜索引擎中,依次排列各个召回网页。

[0107] 其中,召回网页与查询语句的匹配值越大,表示召回网页与查询语句的关联度越

高。按照匹配值的大小排序,来排列召回网页在搜索引擎中的前后位置。这样,与查询语句关联度高的网页可以排列在前,更方便用户查询到关联度高的网页。

[0108] 下面对本申请提供的网页主题句的抽取装置进行介绍,需要说明的是,有关网页主题句的抽取装置的说明可以参见上文有关网页主题句的抽取方法,以下并不赘述。

[0109] 参见图3,其示出了本申请提供的网页主题句的抽取装置实施例1的结构。如图3所示,该网页主题句的抽取装置实施例1可以具体包括:网页及模型获取模块301、分词偏序值确定模块302、主题句偏序确定模块303、以及目标主题句确定模块304;其中:

[0110] 网页及模型获取模块301,用于获取待确定网页、以及预先构建的机器学习模型;其中,待确定网页中包含多个预选取的备选主题句,每个备选主题句中包含若干分词;

[0111] 分词偏序值确定模块302,用于将表示所述分词在待确定网页中重要程度的词语特征值输入至所述机器学习模型,获得所述分词的偏序值;

[0112] 主题句偏序确定模块303,用于依据每个备选主题句包含的分词的偏序值,确定每个备选主题句各自的偏序值;

[0113] 目标主题句确定模块304,用于将偏序值大于预设阈值的备选主题句确定为目标主题句。

[0114] 由以上的技术方案可知,本申请提供的网页主题句的抽取装置中,首先网页及模型获取模块301获取到待确定的网页,待确定网页中包含多个备选的主题句,备选主题句中包含若干分词,分词偏序值确定模块302将词语特征值输入至预设的机器学习模型中,从而获得分词的偏序值,主题句偏序确定模块303可以进一步依据分词的偏序值,确定出备选主题句的偏序值,最终目标主题句确定模块304将偏序值大于预设阈值的备选主题句确定为目标主题句。可见,本实施例利用机器学习模型,来获得备选主题句的偏序值,由于机器学习模型可以反映查询语句与召回的网页之间的关联程度,因此,确定出的偏序值更加准确,从而提高了目标主题句的选取准确率。

[0115] 如图4所示,在实施中,网页主题句的抽取装置还可以包括:训练模型构建模块305,用于预先构建机器学习模型;其中,训练模型构建模块305可以具体包括:网页样本标题获取子模块401、标题分词标注值计算子模块402、标题分词特征值计算子模块403、以及机器训练子模块404;其中:

[0116] 网页样本标题获取子模块401,用于获取若干网页样本的网页标题;其中,每个网页样本均对应若干用于召回网页样本自身的查询语句,每个查询语句均具有权重值,权重值用于表示某一查询语句与该某一查询语句召回的网页样本之间的关联程度;

[0117] 标题分词标注值计算子模块402,用于针对每个网页标题,依据标注值公式

$$\text{label}(\text{term}_j) = \sum_{m=0}^k (v_m * w_m(\text{term}_j) / (\sum_{n=0}^s w_m(\text{term}_n))) , \text{计算网页标题包含的每个分词的标注值};$$

[0118] 其中,term_j表示网页标题中包含的第j个分词,label(term_j)表示分词term_j的标注值,m表示某一个查询语句,k表示查询语句的个数,v_m表示查询语句m的权重值,w_m(term_j)表示分词term_j在查询语句m中的权重值,n表示网页标题中的某一个分词,s表示网页标题中分词的个数;

[0119] 标题分词特征值计算子模块403,用于针对每个网页标题,确定网页标题包含的每个分词的词语特征值;其中,词语特征值用于表示分词在网页样本中的重要程度;

[0120] 机器训练子模块404,用于依据预设的机器训练算法,对若干网页标题中每个分词的标注值及词语特征值进行训练,获得机器学习模型。

[0121] 其中,机器训练子模块404训练使用的机器训练算法为GBRank训练算法,GBRank训练算法中的偏序比较方式为pairwise偏序比较,GBRank训练算法中的损失函数为交叉熵。

[0122] 可选地,上述主题句偏序确定模块303包括:

[0123] 加权求和确定偏序子模块,用于针对每个备选主题句,将备选主题句中分词的偏序值进行加权求和,获得备选主题句的偏序值。

[0124] 在实施中,上述任意一种网页主题句的抽取装置还可以包括:主题句保存模块;其中:

[0125] 主题句保存模块,用于建立待确定网页与目标主题句之间的对应关系;在预设数据库中,保存对应关系。

[0126] 进一步地,网页主题句的抽取装置在包含主题句保存模块的基础上,还可以包括:召回网页确定模块、网页主题句匹配模块、及召回网页排序模块;其中:

[0127] 召回网页确定模块,用于响应于用户在搜索引擎中输入查询语句的操作,确定搜索引擎召回的多个召回网页;其中,召回网页包含在待确定网页中;

[0128] 网页主题句匹配模块,用于在预设数据库中,确定每个召回网页的目标主题句,并分别将输入的查询语句与每个召回网页的目标主题句进行匹配,获得匹配值;

[0129] 召回网页排序模块,用于依据匹配值的大小排序,在搜索引擎中,依次排列各个召回网页。

[0130] 需要说明的是,本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。

[0131] 还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括上述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0132] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下,在其它实施例中实现。因此,本申请将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

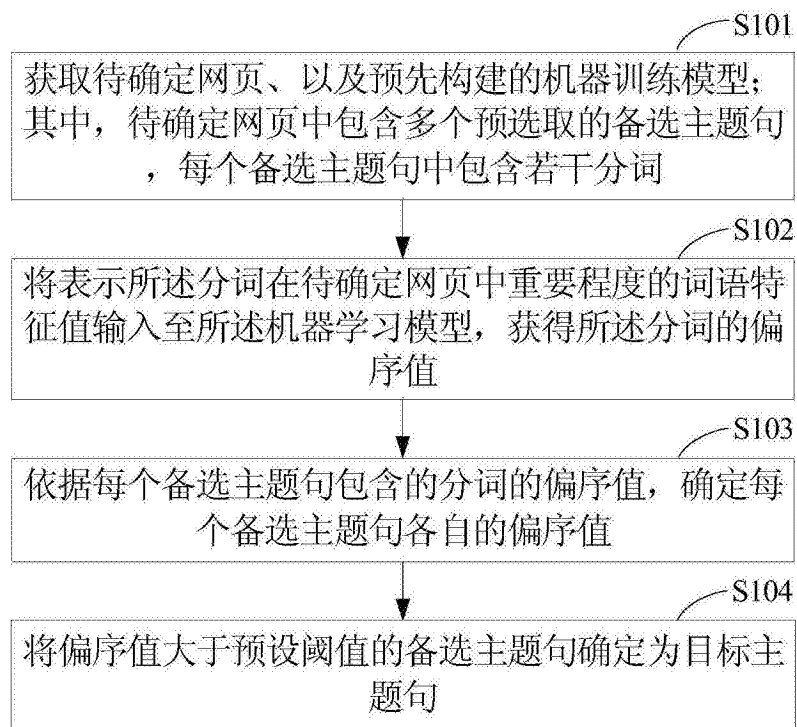


图1

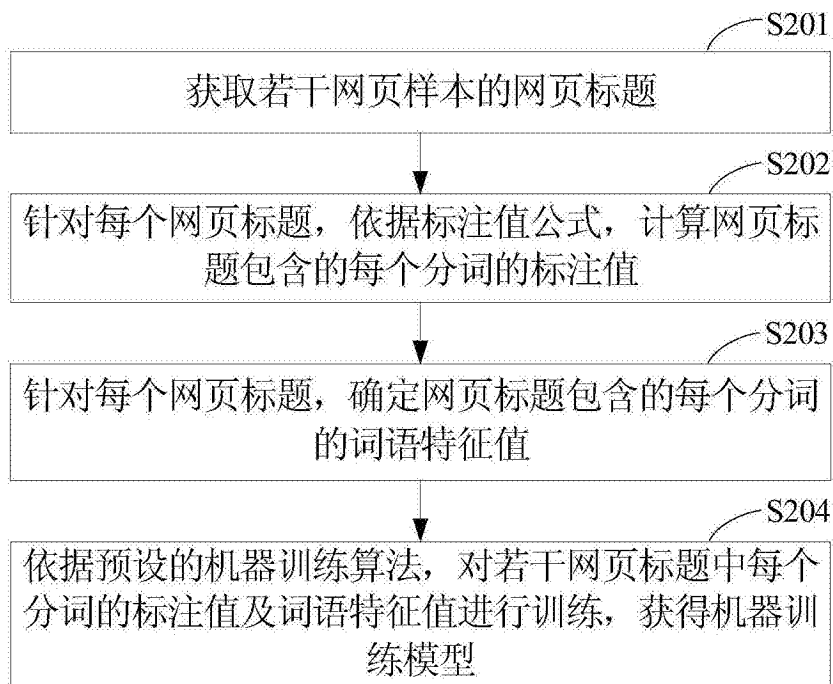


图2



图3

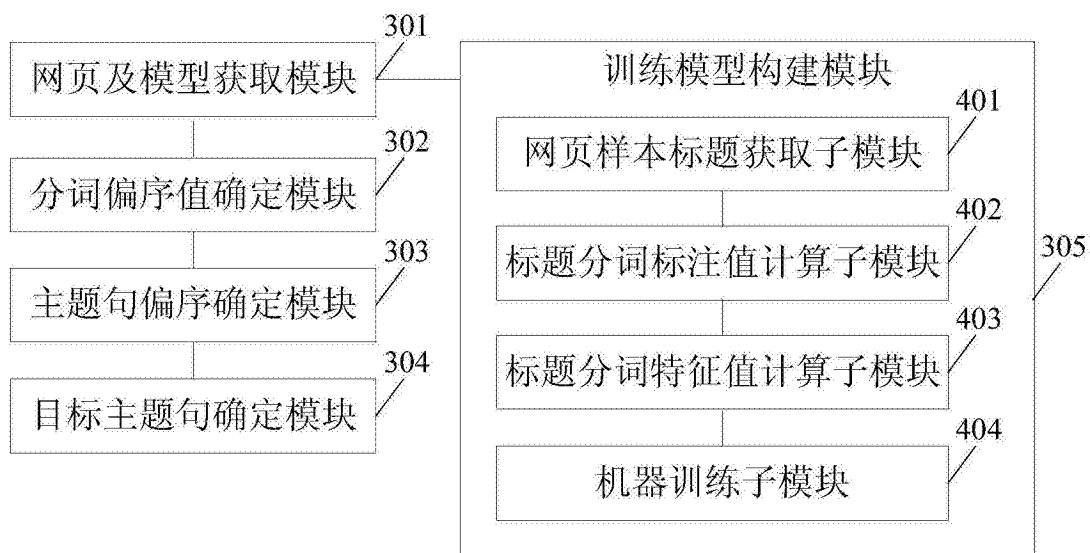


图4