(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 105488209 A
(43)申请公布日 2016.04.13

(21)申请号 201510921247.1

(22)申请日 2015.12.11

(71)申请人 北京奇虎科技有限公司
地址 100088 北京市西城区新街口外大街28号D座112室(德胜园区)
申请人 奇智软件(北京)有限公司

(72)发明人 陈进平

(74)专利代理机构 北京鼎佳达知识产权代理事务所(普通合伙) 11348
代理人 王伟锋 刘铁生

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)

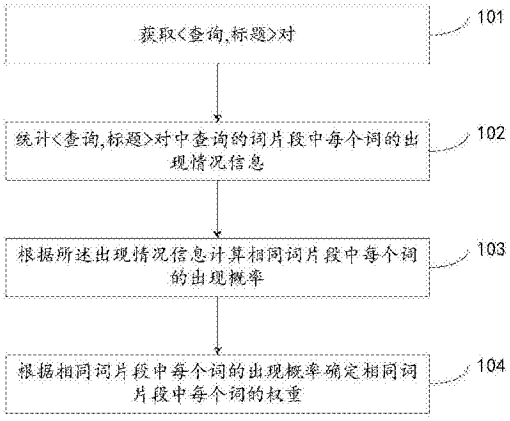
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种词权重的分析方法及装置

(57)摘要

本发明公开了一种词权重的分析方法及装置，涉及互联网技术领域，解决了现有确定term权重的方法无法在互联网搜索引擎环境下准确确定query中term权重的问题。本发明的方法包括：获取<查询，标题>对；统计<查询，标题>对中所述查询的词片段中每个词的出现情况信息；根据所述出现情况信息计算相同词片段中每个词的出现概率；根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。本发明主要用于确定搜索引擎中query的term权重，提高搜索引擎的搜索质量。



CN 105488209 A

1.一种词权重的分析方法，其特征在于，所述方法包括：

获取<查询，标题>对；

统计<查询，标题>对中所述查询的词片段中每个词的出现情况信息；

根据所述出现情况信息计算相同词片段中每个词的出现概率；

根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。

2.根据权利要求1所述的方法，其特征在于，所述获取<查询，标题>对包括：

获取用户点击日志，所述点击日志中包括用户提交的所有查询以及得到的所有标题；

整理所述点击日志，将用户提交的查询与点击所述查询的url得到的标题一一对应，形成<查询，标题>对。

3.根据权利要求1所述的方法，其特征在于，所述统计<查询，标题>对中所述查询的词片段中每个词的出现情况信息包括：

获取<查询，标题>对中所述查询的所有词片段，所述词片段包括所述查询中的每一个词和相邻两个及以上的词组成的词组；

统计所述查询的所有词片段中每个词的出现情况信息。

4.根据权利要求3所述的方法，其特征在于，统计所述查询的所有词片段中每个词的出现情况信息包括：

判断所述查询的词片段中每个词是否在所述查询的<查询，标题>对中对应的标题中出现；

根据判断结果统计所述查询的词片段中每个词的出现情况信息，所述出现情况信息用预设的出现符号以及未出现符号表示。

5.根据权利要求4所述的方法，其特征在于，根据所述出现情况信息计算相同词片段中每个词的出现概率包括：

获取相同词片段所对应的所有标题的总个数；

获取所述相同词片段中每个词在所述对应的所有标题中出现的次数；

用所述次数除以所述对应的所有标题的总个数得到相同词片段中每个词在所述对应的所有标题中的出现概率。

6.根据权利要求5所述的方法，其特征在于，根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重包括：

将相同词片段中每个词在所述对应的所有标题中的出现概率作为所述相同词片段中每个词的权重。

7.一种词权重的分析装置，其特征在于，所述装置包括：

获取单元，用于获取<查询，标题>对；

统计单元，用于统计所述获取单元获取的<查询，标题>对中所述查询的词片段中每个词的出现情况信息；

计算单元，用于根据所述统计单元统计的所述出现情况信息计算相同词片段中每个词的出现概率；

确定单元，用于根据所述计算单元计算的所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。

8.根据权利要求7所述的装置，其特征在于，所述获取单元包括：

获取模块，用于获取用户点击日志，所述点击日志中包括用户提交的所有查询以及得到的所有标题；

整理模块，用于整理所述获取模块获取的所述点击日志，将用户提交的查询与点击所述查询的url得到的标题一一对应，形成<查询，标题>对。

9.根据权利要求7所述的装置，其特征在于，所述统计单元包括：

切分模块，用于获取<查询，标题>对中所述查询的所有词片段，所述词片段包括所述查询中的每一个词和相邻两个及以上的词组成的词组；

统计模块，用于统计所述切分模块获取的所述查询的所有词片段中每个词的出现情况信息。

10.根据权利要求9所述的装置，其特征在于，所述统计单元还用于判断所述查询的词片段中每个词是否在所述查询的<查询，标题>对中对应的标题中出现，以及根据判断结果统计所述查询的词片段中每个词的出现情况信息，所述出现情况信息用预设的出现符号以及未出现符号表示。

11.根据权利要求10所述的装置，其特征在于，所述计算单元包括：

计数模块，用于获取相同词片段所对应的所有标题的总个数；

所述计数模块还用于获取所述相同词片段中每个词在所述对应的所有标题中出现的次数；

计算模块，用于用所述次数除以所述对应的所有标题的总个数得到相同词片段中每个词在所述对应的所有标题中的出现概率。

12.根据权利要求11所述的装置，其特征在于，所述确定单元用于将相同词片段中每个词在所述对应的所有标题中的出现概率作为所述相同词片段中每个词的权重。

一种词权重的分析方法及装置

技术领域

[0001] 本发明涉及互联网技术领域，特别是涉及一种词权重的分析方法及装置。

背景技术

[0002] 随着互联网的发展，互联网中总的存储数据量非常巨大，因此为了使用户能够快速准确的查找到所需要的数据内容，提供互联网搜索服务的厂商就需要对搜索引擎的搜索质量进行优化。其中，权重是搜索引擎给予一个网页的评估值，这个权重可以反映出网页的重要程度，权重越高，说明网页获得更多搜索引擎的信任和认可。而在用户使用搜索引擎的过程中，会在搜索框中提交查询内容，这些查询内容通常称之为query，搜索引擎需要根据query在海量数据中获取有用信息。由于query中具有不同的词语term，其中每个term对于获取有用查询结果而言其重要程度各不相同，因此若要根据query准确获取到目标查询结果就需要参考query中各个term的重要性，也就是需要利用query中term的权重进行目标结果的查询。

[0003] 在现有确定term权重的方法中，通常会利用共同点击、词性以及命名实体来确定term权重，但是这些方法并不是以用户在互联网环境中使用搜索引擎获取内容为基础，从而导致通过上述方法确定的term权重在互联网搜索领域中的参考价值并不高。因此如何在互联网搜索引擎环境下确定term权重成为使用互联网搜索引擎时亟待解决的问题。

发明内容

[0004] 有鉴于此，本发明提出了一种词权重的分析方法及装置，主要目的在于解决现有确定term权重的方法无法在互联网搜索引擎环境下准确确定query中term权重的问题。

[0005] 依据本发明的第一个方面，本发明提供一种词权重的分析方法，包括：

[0006] 获取<查询,标题>对；

[0007] 统计<查询,标题>对中所述查询的词片段中每个词的出现情况信息；

[0008] 根据所述出现情况信息计算相同词片段中每个词的出现概率；

[0009] 根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。

[0010] 进一步的，所述获取<查询,标题>对包括：

[0011] 获取用户点击日志，所述点击日志中包括用户提交的所有查询以及得到的所有标题；

[0012] 整理所述点击日志，将用户提交的查询与点击所述查询的url得到的标题一一对应，形成<查询,标题>对。

[0013] 进一步的，所述统计<查询,标题>对中所述查询的词片段中每个词的出现情况信息包括：

[0014] 获取<查询,标题>对中所述查询的所有词片段，所述词片段包括所述查询中的每一个词和相邻两个及以上的词组成的词组；

[0015] 统计所述查询的所有词片段中每个词的出现情况信息。

[0016] 进一步的，统计所述查询的所有词片段中每个词的出现情况信息包括：

[0017] 判断所述查询的词片段中每个词是否在所述查询的<查询，标题>对中对应的标题中出现；

[0018] 根据判断结果统计所述查询的词片段中每个词的出现情况信息，所述出现情况信息用预设的出现符号以及未出现符号表示。

[0019] 进一步的，根据所述出现情况信息计算相同词片段中每个词的出现概率包括：

[0020] 获取相同词片段所对应的所有标题的总个数；

[0021] 获取所述相同词片段中每个词在所述对应的所有标题中出现的次数；

[0022] 用所述次数除以所述对应的所有标题的总个数得到相同词片段中每个词在所述对应的所有标题中的出现概率。

[0023] 进一步的，根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重包括：

[0024] 将相同词片段中每个词在所述对应的所有标题中的出现概率作为所述相同词片段中每个词的权重。

[0025] 依据本发明的第二个方面，本发明提供一种词权重的分析装置，包括：

[0026] 获取单元，用于获取<查询，标题>对；

[0027] 统计单元，用于统计所述获取单元获取的<查询，标题>对中所述查询的词片段中每个词的出现情况信息；

[0028] 计算单元，用于根据所述出现情况信息计算相同词片段中每个词的出现概率；

[0029] 确定单元，用于根据所述计算单元计算的所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。

[0030] 进一步的，所述获取单元包括：

[0031] 获取模块，用于获取用户点击日志，所述点击日志中包括用户提交的所有查询以及得到的所有标题；

[0032] 整理模块，用于整理所述获取模块获取的所述点击日志，将用户提交的查询与点击所述查询的url得到的标题一一对应，形成<查询，标题>对。

[0033] 进一步的，所述统计单元包括：

[0034] 切分模块，用于获取<查询，标题>对中所述查询的所有词片段，所述词片段包括所述查询中的每一个词和相邻两个及以上的词组成的词组；

[0035] 统计模块，用于统计所述切分模块获取的所述查询的所有词片段中每个词的出现情况信息。

[0036] 进一步的，所述统计单元还用于判断所述查询的词片段中每个词是否在所述查询的<查询，标题>对中对应的标题中出现，以及根据判断结果统计所述查询的词片段中每个词的出现情况信息，所述出现情况信息用预设的出现符号以及未出现符号表示。

[0037] 进一步的，所述计算单元包括：

[0038] 计数模块，用于获取相同词片段所对应的所有标题的总个数；

[0039] 所述计数模块还用于获取所述相同词片段中每个词在所述对应的所有标题中出现的次数；

[0040] 计算模块，用于用所述次数除以所述对应的所有标题的总个数得到相同词片段中

每个词在所述对应的所有标题中的出现概率。

[0041] 进一步的，所述确定单元用于将相同词片段中每个词在所述对应的所有标题中的出现概率作为所述相同词片段中每个词的权重。

[0042] 借由上述技术方案，本发明实施例提供的一种词权重的分析方法及装置，能够在用户大规模使用互联网搜索引擎的过程中获取到<查询，标题>对，并统计查询中的词片段中每个词的出现情况信息，根据每个词的出现情况信息计算相同词片段中每个词的出现概率，根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。而在现有技术中，当确定搜索查询中词的权重时无法基于互联网环境中使用搜索引擎获取内容为基础，从而造成搜索词的词权重确定不准确，进而影响搜索结果的准确性。与现有技术中的这一缺陷相比，本发明能够以用户大规模使用搜索引擎点击形成的日志为基础，在互联网搜索引擎环境下准确确定搜索查询中词的权重，从而有效提高搜索结果的准确性。

[0043] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0044] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0045] 图1示出了本发明实施例提供的一种词权重的分析方法的流程图；

[0046] 图2示出了本发明实施例提供的一种词权重的分析装置的组成框图；

[0047] 图3示出了本发明实施例提供的另一种词权重的分析装置的组成框图；

[0048] 图4示出了本发明实施例提供的另一种词权重的分析装置的组成框图；

[0049] 图5示出了本发明实施例提供的另一种词权重的分析装置的组成框图。

具体实施方式

[0050] 下面将参照附图更加详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0051] 在用户使用搜索引擎时需要提交查询query，查询query中具有不同的词语term，其中每个term对于获取有用查询结果而言其重要程度各不相同，因此若要根据query准确获取到目标查询结果就需要参考query中各个term的重要性，也就是需要利用query中term的权重进行目标结果的查询。在现有确定term权重的方法中，通常会利用共同点击、词性以及命名实体来确定term权重，但是这些方法并不是以用户在互联网环境中使用搜索引擎获取内容为基础，从而导致通过上述方法确定的term权重在互联网搜索领域中的参考价值并不高。

[0052] 为了解决上述问题，本发明实施例提供了一种词权重的分析方法，能够基于互联网搜索引擎环境准确确定用户提交的查询query中各个关键词term的权重，如图1所示，该

方法包括：

[0053] 101、获取<查询,标题>对。

[0054] 在用户使用搜索引擎查询所需要的内容时需要提交包含有关键词term的查询query，搜索引擎根据用户提交的query匹配到一些相关的标题title供用户点击观看，当用户点击相关的title后，本发明实施例就可以将用户提交的query和点击的title进行组合形成<查询,标题>对，也可以记作<query,title>对。

[0055] 102、统计<查询,标题>对中查询的词片段中每个词的出现情况信息。

[0056] 由于搜索引擎在根据用户提交的query在互联网上搜索相应的内容时，需要根据query中每个词term的重要性调整搜索策略，而query中的term出现在query对应的title中的次数越多说明query中该term越重要，因此本发明实施例需要执行步骤102统计大规模的<query,title>对中查询的词片段中每个词的出现情况信息，根据出现情况信息确定词片段中每个词的重要性。

[0057] 103、根据所述出现情况信息计算相同词片段中每个词的出现概率。

[0058] 由于本发明实施例需要统计大规模的<query,title>对，因此所有统计的query中包含有大量的相同词片段，对相同词片段ABC而言，所有包含词片段ABC的query中，各个query对应的title里有部分title包含term-A，部分title里不包含term-A；部分title包含term-B，部分title里不包含term-B；部分title包含term-C，部分title里不包含term-C。也就是说相同词片段中每个词在所有包含所述相同词片段的query所对应的title中的出现概率不相同，因此相同词片段中每个词的重要性也就不一样。由此本发明实施例需要执行步骤103根据相同词片段中每个词在所有包含所述相同词片段的query所对应的title中的出现情况信息计算相同词片段中每个词的出现概率。

[0059] 104、根据相同词片段中每个词的出现概率确定相同词片段中每个词的权重。

[0060] 由于权重是一个相对的概念，针对某个指标而言，该指标的权重是指该指标在整体评价中的相对重要程度。而对本发明实施例而言，某个term的权重就是指该term在其所在的query的词片段中的相对重要程度，同时重要程度越高的term在其词片段所在的query对应的title中出现的概率越高，因此当在步骤103中计算出相同词片段中每个词在所有包含所述相同词片段的query所对应的title中的出现概率之后，就可以根据相同词片段中每个词的出现概率确定相同词片段中每个词的权重，以便搜索引擎根据由大规模统计<查询,标题>对所确定的term权重调整搜索策略，提高搜索结果的准确性。

[0061] 本发明实施例提供的一种词权重的分析方法，能够在用户大规模使用互联网搜索引擎的过程中获取到<查询,标题>对，并统计查询中的词片段中每个词的出现情况信息，根据每个词的出现情况信息计算相同词片段中每个词的出现概率，根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。而在现有技术中，当确定搜索查询中词的权重时无法基于互联网环境中使用搜索引擎获取内容为基础，从而造成搜索词的词权重确定不准确，进而影响搜索结果的准确性。与现有技术中的这一缺陷相比，本发明能够以用户大规模使用搜索引擎点击形成的日志为基础，在互联网搜索引擎环境下准确确定搜索查询中词的权重，从而有效提高搜索结果的准确性。

[0062] 为了更好的对上述图1所示的方法进行理解，作为对上述实施方式的细化和扩展，本发明实施例将针对图1中的步骤进行详细说明。

[0063] 通常用户在使用互联网的过程中会产生大量的点击日志，这些点击日志信息中包括用户在搜索引擎里提交的查询query，所述query点击的统一资源定位符url以及url对应的标题title等数据。由于用户提交的query以及点击所述query的url得到的title通常都具有相互对应的关系，因此通过大规模的统计点击日志信息就可以得到互联网搜索引擎环境下确定搜索关键词term权重的数据基础。由于用户在提交一个query时，有时会点击多个url得到多个相关title，这些title的质量也就是与query的匹配度也会存在高低差异，因此本发明实施例需要对获取的点击日志进行整理，将点击日志中的query与title一一对应，得到<query，title>对。其中，由于用户在提交一个query时，可能点击多个url得到多个对应的title，因此在获得的大规模的<query，title>对中，同一个query也会具有多个<query，title>对。

[0064] 由于用户在搜索引擎里提交query后，搜索引擎需要根据query中每个term(关键词)的相对重要程度也就是权重调整搜索策略，以便获取到准确的搜索结果。而query中每个term的重要程度可以用term在query对应的title中的出现情况来表示，如果在大量query中的某个term在对应的title中出现的次数越多，说明该term越重要。由于各个query中会包含有多种多样的词片段，词片段包括query中的每一个term和相邻两个及以上的term组成的词组，而且各个query中也会包含相同的词片段，就同一个词片段来说，所述同一个词片段中的term在所有包含所述词片段的query对应的title中出现的次数越多，说明在所述词片段中该term越重要。因此，本发明实施例需要统计所有query的词片段中每个term的出现情况信息。为了统计所有query的词片段中每个term的出现情况信息，本发明实施例需要对所有的query进行分词，也就是处理所有的<query，title>对，将各个query进行分词，得到query中的每一个term和相邻两个及以上的term组成的词组也就是上述的词片段，并统计词片段中每个term在其对应的title中的出现情况信息。

[0065] 在统计每个query的所有词片段中每个term的出现情况信息时，可以用预设的出现符号以及未出现符号进行表示。也就是判断query的词片段中每个term是否在所述query的<query，title>对中对应的title中出现，若出现，则用预设出现符号表示，若未出现，则用预设未出现符号表示。例如对于<query:ABCD，title:CDEFG>而言，其query中的一个词片段为ABC，这个词片段ABC中的term-A在title:CDEFG中未出现，则用未出现符号0表示；term-B在title:CDEFG中未出现，则用未出现符号0表示；term-C在title:CDEFG中出现，则用出现符号1表示，因此统计词片段ABC中每个term的出现情况信息就可以用ABC:001表示。

[0066] 当通过上述方式确定<query，title>对中query的词片段中每个term的出现情况信息后，就可以计算相同词片段中每个term的出现概率。具体的在计算相同词片段中每个term的出现概率时，需要获取相同词片段所对应的所有title的总个数。对于同一个词片段而言，就是query中包含所述同一个词片段的所有<query，title>对的总个数，在所有这些<query，title>对的总个数中，部分<query，title>对中的title包含有所述同一个词片段的term，部分<query，title>对中的title不包含有所述同一个词片段的term，因此在获取同一个词片段所对应的所有title的总个数之后，还需要获取同一个词片段中每个term在所述所有title中出现的次数，也就是在所有title中包含某个term的title的个数。用同一个词片段中每个term在所有title中出现的次数除以对应的所有title的总个数得到相同词片段中每个term在对应的所有title中的出现概率。

[0067] 对于同一个词片段而言，其中某个term在其所在query对应的title中的出现频率越高，该term就越重要，因此根据计算得到的相同词片段中每个term的出现概率可以确定相同词片段中每个term的权重。作为一种可选的实施方式，本发明实施例可以将相同词片段中每个term在其对应的所有title中的出现概率作为所述相同词片段中每个term的权重。

[0068] 为了更好的对上述方法进行理解，本发明实施例将以两个<query，title>对为例，对上述过程进行详细说明。这两个<query，title>对分别为<query:ABC，title:CDEF>、<query:ABCDE，title:FGACDHJ>。其中，如果query中的term出现在对应的title中，则用出现符号1表示，如果query中的term未出现在对应的title中，则用未出现符号0表示。

[0069] 在统计<query，title>对中query的词片段中每个term的出现情况时，首先需要对<query，title>对中的query进行分词得到所有词片段，然后统计词片段中每个term的出现情况，也就是以query中的词片段为key，以词片段包含的term在对应的title中出现情况为value进行输出，其处理结果如下：

[0070] 1)在<query:ABC，title:CDEF>对中，

[0071] 包含1个term的:A:0，B:0，C:1

[0072] 包含2个term的:AB:00，BC:01

[0073] 包含3个term的:ABC:001

[0074] 2)在<query:ABCDE，title:FGACDHJ>对中，

[0075] 包含1个term的:A:1，B:0，C:1，D:1，E:0

[0076] 包含2个term的:AB:10，BC:01，CD:11，DE:10

[0077] 包含3个term的:ABC:101，BCD:011，CDE:110

[0078] 包含4个term的:ABCD:1011，BCDE:0110

[0079] 包含5个term的:ABCDE:10110

[0080] 当处理完所有<query，title>对之后，需要根据词片段中每个term的出现情况信息对相同的词片段进行合并，也就是计算相同词片段中每个term的出现概率。以词片段ABC为例，在<query:ABC，title:CDEF>对中，词片段ABC的value值为001；在<query:ABCDE，title:FGACDHJ>对中，词片段ABC的value值为101，其中，term-A在<query:ABC，title:CDEF>对中的title中未出现，而在<query:ABCDE，title:FGACDHJ>对中的title中出现，因此term-A在title中出现的概率为0.5；同理，term-B在title中出现的概率为0，term-C在title中出现的概率为1，因此对于词片段ABC而言，各个term在搜索结果中出现的概率为ABC:0.5、0、1。根据上述统计结果可知，当用户在搜索引擎中提交包含ABC的query时，搜索时需要参考的term的重要性依次为term-C>term-A>term-B。

[0081] 当然，上述只是以两个<query，title>对为例进行的说明，其得到的概率还不具有代表性，只是为了能够清楚说明具体的分析过程。在实际进行分析的过程中，需要按照上述方式大规模的统计<query，title>对才能得到词片段中每个term可靠的出现概率。例如，若统计大量的<query，title>对后得到类似如下数据，词片段ABC:0.7、0.3、0.9，则表示如下含义：在所有包含词片段ABC的query中，点击的title里包含term-A的概率是0.7，包含term-B的概率是0.3，包含term-C的概率是0.9，因此可以认为term-A和term-C的重要性比较高，而term-B的重要性比较低。

[0082] 通过本发明实施例所述的词权重的分析方法，可以大规模的挖掘互联网搜索环境下的词片段以及词片段中包含的term在title里的出现概率，例如如下两个词片段：a)番茄鱼汤：0.75、0.82；b)鱼汤好吗：0.78、0.51。其中，a)中表示所有包含“番茄鱼汤”的query点击的title中，75％包含“番茄”，82％包含“鱼汤”，由于“番茄”有同义词“西红柿”，所以实际上所述title中包含的番茄的概率还要高。在b)中表明所有包含“鱼汤好吗”的query点击的title中，“鱼汤”比“好吗”出现的次数多，“鱼汤”比“好吗”更加重要。

[0083] 本发明实施例利用<query，title>对统计query中term在title里是否出现，并且把出现情况信息通过词片段的value值进行输出，进一步的根据每个词片段的value值统计相同词片段中每个term在title中的出现概率，由此得到词片段中各个term的权重信息，由于这些term的权重信息是基于大规模的互联网搜索环境下的点击日志信息确定的，因此能够有效提高搜索引擎的搜索质量。

[0084] 进一步的，作为对上述图1所示方法的实现，本发明实施例提供了一种词权重的分析装置，如图2所示，该装置包括：获取单元21、统计单元22、计算单元23以及确定单元24，其中，

[0085] 获取单元21，用于获取<查询，标题>对；

[0086] 统计单元22，用于统计获取单元21获取的<查询，标题>对中所述查询的词片段中每个词的出现情况信息；

[0087] 计算单元23，用于根据统计单元22统计的所述出现情况信息计算相同词片段中每个词的出现概率；

[0088] 确定单元24，用于根据计算单元23计算的所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。

[0089] 进一步的，如图3所示，获取单元21包括：

[0090] 获取模块211，用于获取用户点击日志，所述点击日志中包括用户提交的所有查询以及得到的所有标题；

[0091] 整理模块212，用于整理获取模块211获取的所述点击日志，将用户提交的查询与点击所述查询的url得到的标题一一对应，形成<查询，标题>对。

[0092] 进一步的，如图4所示，统计单元22包括：

[0093] 切分模块221，用于获取<查询，标题>对中所述查询的所有词片段，所述词片段包括所述查询中的每一个词和相邻两个及以上的词组成的词组；

[0094] 统计模块222，用于统计切分模块221获取的所述查询的所有词片段中每个词的出现情况信息。

[0095] 进一步的，统计单元22还用于判断所述查询的词片段中每个词是否在所述查询的<查询，标题>对中对应的标题中出现，以及根据判断结果统计所述查询的词片段中每个词的出现情况信息，所述出现情况信息用预设的出现符号以及未出现符号表示。

[0096] 进一步的，如图5所示，计算单元23包括：

[0097] 计数模块231，用于获取相同词片段所对应的所有标题的总个数；

[0098] 计数模块231还用于获取所述相同词片段中每个词在所述对应的所有标题中出现的次数；

[0099] 计算模块232，用于用所述次数除以所述对应的所有标题的总个数得到相同词片

段中每个词在所述对应的所有标题中的出现概率。

[0100] 进一步的，确定单元24用于将相同词片段中每个词在所述对应的所有标题中的出现概率作为所述相同词片段中每个词的权重。

[0101] 本发明实施例提供的一种词权重的分析装置，能够在用户大规模使用互联网搜索引擎的过程中获取到<查询，标题>对，并统计查询中的词片段中每个词的出现情况信息，根据每个词的出现情况信息计算相同词片段中每个词的出现概率，根据所述相同词片段中每个词的出现概率确定所述相同词片段中每个词的权重。而在现有技术中，当确定搜索查询中词的权重时无法基于互联网环境中使用搜索引擎获取内容为基础，从而造成搜索词的词权重确定不准确，进而影响搜索结果的准确性。与现有技术中的这一缺陷相比，本发明能够以用户大规模使用搜索引擎点击形成的日志为基础，在互联网搜索引擎环境下准确确定搜索查询中词的权重，从而有效提高搜索结果的准确性。

[0102] 此外，本发明实施例利用<query，title>对统计query中term在title里是否出现，并且把出现情况信息通过词片段的value值进行输出，进一步的根据每个词片段的value值统计相同词片段中每个term在title中的出现概率，由此得到词片段中各个term的权重信息，由于这些term的权重信息是基于大规模的互联网搜索环境下的点击日志信息确定的，因此能够有效提高搜索引擎的搜索质量。

[0103] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

[0104] 可以理解的是，上述方法及装置中的相关特征可以相互参考。另外，上述实施例中的“第一”、“第二”等是用于区分各实施例，而并不代表各实施例的优劣。

[0105] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0106] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述，构造这类系统所要求的结构是显而易见的。此外，本发明也不针对任何特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本发明的内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0107] 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

[0108] 类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

[0109] 本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单

元或组件组合成一个模块或单元或组件，以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

[0110] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使用。

[0111] 本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本发明实施例的发明名称(如确定网站内链接等级的装置)中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如，计算机程序和计算机程序产品)。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

[0112] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中，这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

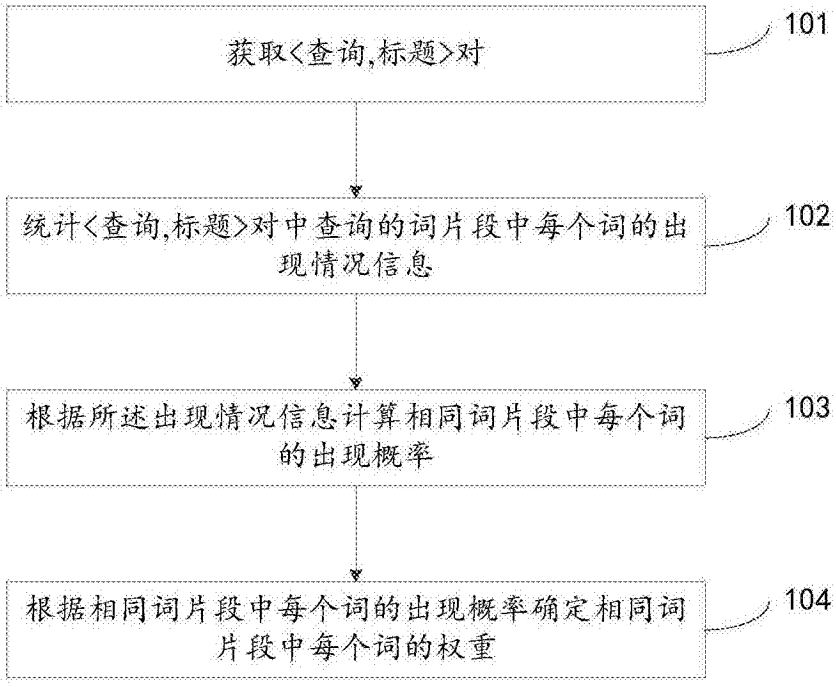

获取<查询,标题>对
101
统计<查询,标题>对中查询的词片段中每个词的出现情况信息
102
根据所述出现情况信息计算相同词片段中每个词的出现概率
103
根据相同词片段中每个词的出现概率确定相同词片段中每个词的权重
104

图1

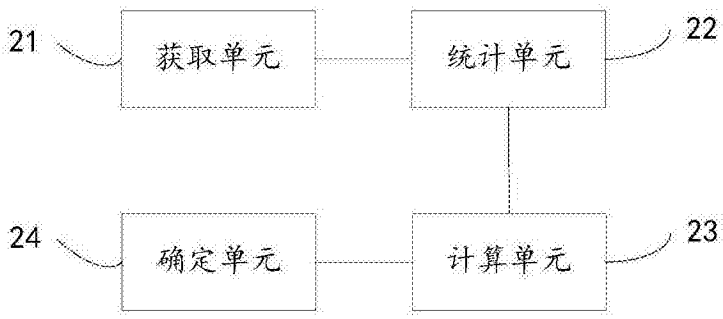

21
获取单元
统计单元
22
24
确定单元
计算单元
23

图2

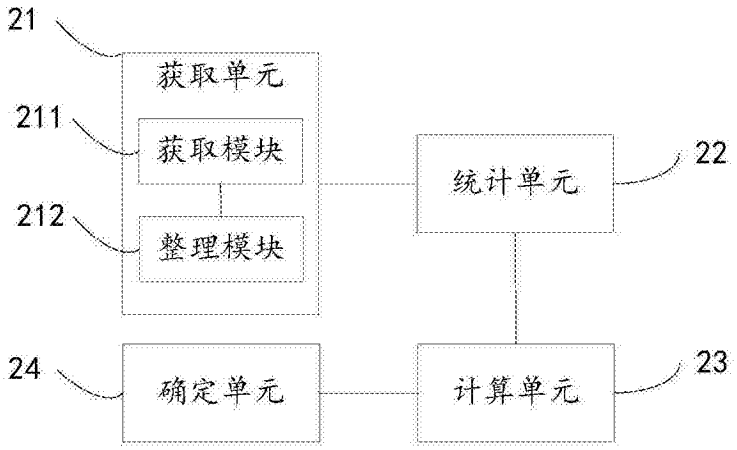

21
获取单元
211
获取模块
212
整理模块
统计单元
22
24
确定单元
计算单元
23

图3

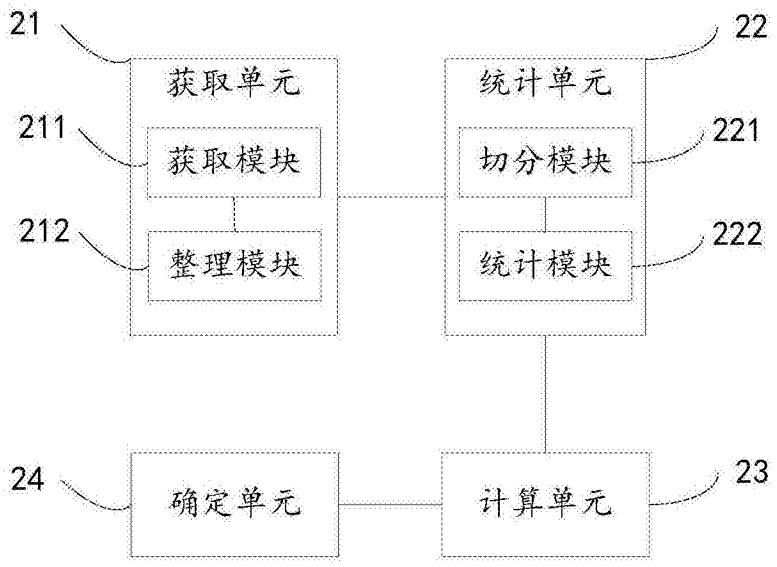

21
获取单元
211
获取模块
212
整理模块
22
统计单元
221
切分模块
222
统计模块
24
确定单元
23
计算单元

图4

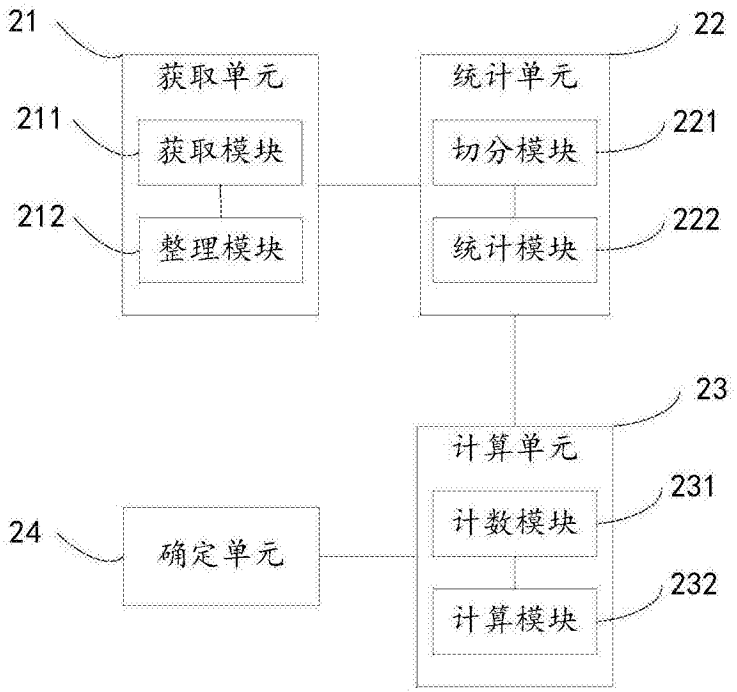

21
获取单元
211
获取模块
212
整理模块
22
统计单元
221
切分模块
222
统计模块
23
计算单元
231
计数模块
232
计算模块
24
确定单元

图5