(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 103246669 B
(45)授权公告日 2018.04.27

(21)申请号 201210027535.9

(22)申请日 2012.02.08

(65)同一申请的已公布的文献号
申请公布号 CN 103246669 A

(43)申请公布日 2013.08.14

(73)专利权人 深圳市世纪光速信息技术有限公司
地址 518057 广东省深圳市南山区粤海街道科技中一路腾讯大厦16层

(72)发明人 杨巍 张立明

(74)专利代理机构 北京华沛德权律师事务所 11302
代理人 刘杰

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

(56)对比文件
CN 1731396 A,2006.02.08,
CN 101636737 A,2010.01.27,
CN 101887426 A,2010.11.17,
魏超 等.网页质量评价体系的研究.《中文信息学报》.2011,第25卷(第5期),第3－8页.

审查员 刘申

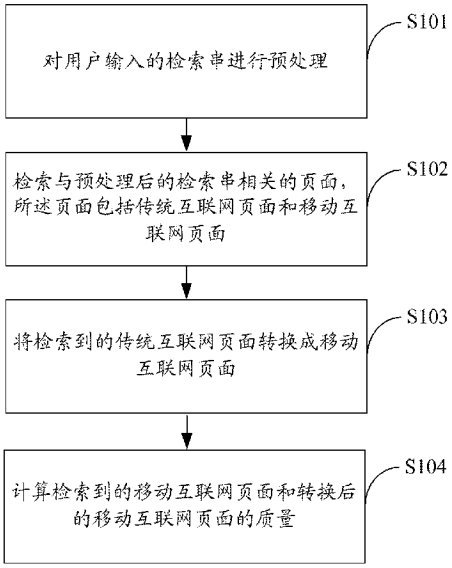
权利要求书1页 说明书6页 附图2页

(54)发明名称

一种移动终端网页质量计算的方法以及装置

(57)摘要

本发明适用于移动终端领域,提供了一种移动终端网页质量计算的方法及装置,所述方法包括下述步骤:对用户输入的检索串进行预处理;检索与预处理后的检索串相关的页面,所述页面包括传统互联网页面和移动互联网页面;将检索到的传统互联网页面转换成移动互联网页面;计算检索到的移动互联网页面和转换后的移动互联网页面的质量。本发明能够有效解决移动终端传统互联网页面与移动互联网页面的页面质量计算以及混合排序的问题。



CN 103246669 B

1.一种移动终端网页质量计算的方法，其特征在于，所述方法包括以下步骤：

对用户输入的检索串进行预处理；

检索与预处理后的检索串相关的页面，所述页面包括传统互联网页面和移动互联网页面；

将检索到的传统互联网页面转换成移动互联网页面；

根据传统互联网页面转换成移动互联网的难易程度计算转换质量；

计算检索到的移动互联网页面和转换后的移动互联网页面的质量，包括：将获得的转换质量作为页面质量的一个影响因子，计算转换后的移动互联网页面的质量。

2.如权利要求1所述的方法，其特征在于，所述转换包括以下至少一个：页面的标题识别、页面的主体内容识别、页面的核心正文识别、与页面正文相关的多媒体信息的识别以及与页面正文相关标签的替换和删除。

3.如权利要求1所述的方法，其特征在于，所述转换质量的计算包括以下至少一个：页面排版质量的计算、页面主体内容质量的计算、页面交互性质量的计算、页面主体内容包含资源质量的计算。

4.如权利要求1所述的方法，其特征在于，所述页面质量的计算包括以下至少一个：检索串与页面文本相关性的计算、页面链接关系的计算、页面作弊情况的计算、页面敏感度的计算、页面排版质量的计算以及页面转换质量的计算。

5.一种移动终端网页质量计算装置，其特征在于，所述装置包括：

预处理单元，用于对用户输入的检索串进行预处理；

检索单元，用于检索与预处理后的检索串相关的页面，所述页面包括传统互联网页面和移动互联网页面；

转换单元，用于将检索到的传统互联网页面转换成移动互联网页面；

转换质量计算单元，用于根据传统互联网页面转换成移动互联网的难易程度计算转换质量；

页面质量计算单元，用于计算检索到的移动互联网页面和转换后的移动互联网页面的质量，包括：将获得的转换质量作为页面质量的一个影响因子，计算转换后的移动互联网页面的质量。

6.如权利要求5所述的装置，其特征在于，所述转换包括以下至少一个：页面的标题识别、页面的主体内容识别、页面的核心正文识别、与页面正文相关的多媒体信息的识别以及与页面正文相关标签的替换和删除。

7.如权利要求5所述的装置，其特征在于，所述转换质量的计算包括以下至少一个：页面排版质量的计算、页面主体内容质量的计算、页面交互性质量的计算、页面主体内容包含资源质量的计算。

8.如权利要求5所述的装置，其特征在于，所述页面质量的计算包括以下至少一个：检索串与页面文本相关性的计算、页面链接关系的计算、页面作弊情况的计算、页面敏感度的计算、页面排版质量的计算以及页面转换质量的计算。

9.一种移动终端，其特征在于，所述移动终端包含权利要求5至8任一项所述的移动终端网页质量计算装置。

一种移动终端网页质量计算的方法以及装置

技术领域

[0001] 本发明属于移动终端领域，尤其涉及一种移动终端网页质量计算的方法以及装置。

背景技术

[0002] 随着移动互联网的发展，越来越多的适合移动终端访问的网站出现在互联网中。在移动终端的信息搜索中，如何将搜索到的传统的互联网页面与移动互联网页面进行混合排序，是移动终端亟需解决的问题。

[0003] 在传统的互联网页面中，PageRank算法是目前比较常见的衡量页面质量的方法之一，它是根据页面的外部链接和内部链接的数量和质量来衡量页面的价值。

[0004] 在移动互联网页面中，对于页面作弊程度上的打分是常见的衡量页面质量的方法之一。该方法根据页面的作弊程度，对每一个页面进行打分，从而衡量出不同页面的质量。

[0005] 虽然PageRank算法可以从页面内外部链接的角度衡量页面的质量，但是由于传统互联网页面的数量与移动互联网页面的数量差距非常悬殊，而且两种页面间的交集也较少，通过PageRank算法计算得到的传统互联网页面的质量远远高于移动互联网页面的质量，无法达到公平衡量两种页面质量的目的。

[0006] 另外，根据页面作弊程度上的打分，虽然可以比较好的区分页面是否作弊。但是，对于两个都没有作弊的页面，则无法区分出两个页面的质量。

发明内容

[0007] 本发明实施例提供一种移动终端网页质量计算的方法，旨在解决现有技术无法有效衡量移动终端页面的质量的问题。

[0008] 本发明实施例是这样实现的，一种移动终端网页质量计算的方法，所述方法包括以下步骤：

[0009] 对用户输入的检索串进行预处理；

[0010] 检索与预处理后的检索串相关的页面，所述页面包括传统互联网页面和移动互联网页面；

[0011] 将检索到的传统互联网页面转换成移动互联网页面；

[0012] 计算检索到的移动互联网页面和转换后的移动互联网页面的质量。

[0013] 本发明实施例提供一种移动终端网页质量计算装置，所述装置包括：

[0014] 预处理单元，用于对用户输入的检索串进行预处理；

[0015] 检索单元，用于检索与预处理后的检索串相关的页面，所述页面包括传统互联网页面和移动互联网页面；

[0016] 转换单元，用于将检索到的传统互联网页面转换成移动互联网页面；

[0017] 页面质量计算单元，用于计算检索到的移动互联网页面和转换后的移动互联网页面的质量。

[0018] 本发明实施例提供一种移动终端，所述移动终端包括所述移动终端网页质量计算装置。

[0019] 从上述技术方案可以看出，本发明实施例将检索到的传统互联网页面转换成移动互联网页面，由于传统互联网页面转换成移动互联网页面之后，与其他的移动互联网页面具有类似的属性特征，从而可以统一、公平的计算移动互联网页面和传统互联网页面的质量。

附图说明

[0020] 图1是本发明实施例一提供的移动终端网页质量计算方法的实现流程图；

[0021] 图2是本发明实施例一提供的传统互联网页面转换的示例图；

[0022] 图3是本发明实施例二提供的移动终端网页质量计算装置的组成结构图。

具体实施方式

[0023] 为了使本发明的技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0024] 本发明实施例将检索到的传统互联网页面转换成移动互联网页面，由于传统互联网页面转换成移动互联网页面之后，与其他的移动互联网页面具有类似的属性特征，从而达到统一、公平的计算移动互联网页面和传统互联网页面质量的目的。

[0025] 为了说明本发明所述的技术方案，下面通过具体实施例来进行说明。

[0026] 实施例一：

[0027] 图1示出了本发明实施例一提供的移动终端网页质量计算方法的实现流程，该方法过程详述如下：

[0028] 在步骤S101中，对用户输入的检索串进行预处理。

[0029] 在本实施例中，所述预处理包括但不局限于检索串切分、同义词替换以及检索串扩展等。

[0030] 在步骤S102中，检索与预处理后的检索串相关的页面。

[0031] 在本实施例中，移动终端在互联网上检索与预处理后的检索串相关的页面，所述页面包括传统互联网页面(遵循http协议的网页)和移动互联网页面(遵循wap协议的网页)。

[0032] 在步骤S103中，将检索到的传统互联网页面转换成移动互联网页面。

[0033] 在本实施例中，对检索到的传统互联网页面进行转换，转换之后的页面适合移动终端的访问。其中，所述转换包括但不局限于以下至少一个：页面的标题识别、页面的主体内容识别、页面的核心正文识别、与页面正文相关的多媒体信息的识别以及与页面正文相关标签的替换和删除。

[0034] 举例说明传统互联网页面的转换过程，如图2所示：

[0035] 1)传统互联网页面的url：http://news.qq.com/a/20101120/000780.htm；

[0036] 2)下载与该URL对应的html源码以及相关的附件信息css/frame/is等；

[0037] 3)对该页面进行解析，查找与<h1>标签对应的文本-“下岗女工取款3千被给3万元

3次还款被银行赶出”，该文本与<title>标签中的文本相似度很高，因此，将该文本作为网页的标题，也作为网页核心正文的开始；

[0038] 4)分析网页核心正文的结尾：由于在网页的某部分存在关键字“相关阅读”，并存在其他新闻的链接。因此，判断出该部分为网页中的相关链接，作为核心正文的结尾部分。核心正文的开始和结尾中间的部分作为该html网页的核心正文，其他部分作为该html页面中的噪声部分；

[0039] 5)根据分析到的网页核心正文开始和结尾位置，将核心正文中出现的图片(“http://img1.gtimg.com/news/pics/hv1/251/29/792/51507446.jpg”和“http://img1.gtimg.com/news/pics/hv1/250/29/792/51507445.jpg”)和视频/flash作为该html页面中正文相关图片和视频/flash；

[0040] 6)对核心正文中的图片进行下载，判断图片的大小以及长宽信息，对大图片进行缩略处理，同时，存储对应的缩略图的地址；

[0041] 7)对核心正文中的视频/flash信息进行关键图像帧提取，对提取出的关键图像帧进行缩略处理，同时，存储对应的缩略图的地址；

[0042] 8)根据html页面、html页面核心正文、正文相关的图片、视频/flash的缩略图地址，对该html页面进行排版处理。排版的过程就是html标签的替换或者删除的过程。根据不同的移动设备，对标签进行不同的处理。如移动设备只支持wap1.0页面，则删除掉html中的标签，只保留<p>，<a>等wap1.0页面支持的标题，同时将核心正文图片/视频/flash地址替换成缩略图地址。如果移动设备支持wap2.0页面，则同样只需将wap2.0页面中不支持的标签进行替换和删除，同时将正文图片/视频/flash地址替换成缩略图地址。对于html页面中的噪声部分，可以根据具体需求不展示或者折叠、隐藏展示等；

[0043] 9)将转换后的页面发送给移动终端设备。

[0044] 进一步的是，本实施例还包括根据传统互联网页面转换成移动互联网的难易程度计算转换质量。其中，所述转换质量的计算包括但不局限于以下至少一个：页面排版质量的计算、页面主体内容质量的计算、页面交互性质量的计算、页面主体内容包含资源质量的计算。

[0045] 具体说明如下：

[0046] 页面排版质量的计算：从转换后的页面排版是否适合移动终端访问来计算质量因子。例如：table节点在移动终端访问时排版效果很差，如果原始页面的主体是由table表格组成，则其转换质量较差。

[0047] 页面主体内容质量的计算：从页面主体内容是否适合移动终端访问而计算质量因子。例如：页面主体内容是视频播放窗口的，由于一般的移动终端不支持视频播放窗口，因此此类页面的转换质量较差。

[0048] 页面交互性质量的计算：从页面主体内容是否需要和用户进行交互而计算质量因子。例如：页面主体是一个JS(Javascript)控制的登录框，由于一般的移动终端不支持JS脚本，因此此类页面的转换质量较差。

[0049] 页面主体内容包含资源质量的计算：从页面主体内容包含资源的质量来计算质量因子。例如：在一个包含下载资源的页面中，其下载资源(如ipa格式、sis格式等)是适合移动终端下载的，则其转换质量较好；下载资源不适合移动终端下载的，则其转换质量较差。

[0050] 针对上述影响转换质量的因子，采用机器学习算法、分支定界方法或者阈值分支法等，建立样本集进行训练，根据训练得到的决策模型来确定各影响转换质量的因子在页面中所占的权重，并根据确定的权重，计算整个页面的转换质量。其中，整个页面转换质量的计算方式包括但不局限于以下几种：1)取各影响转换质量的因子在页面中所占权重的平均值；2)直接将各影响转换质量的因子所占权重值相乘。例如：通过决策模型得到在某一页面(每个页面含有的影响因子的个数不同，所占的权重也会不同)中页面排版质量因子所占权重为0.3、页面主体内容质量因子所占权重为0.3、页面交互性质量因子所占权重为0.2、页面主体内容包含资源质量因子所占权重为0.2，则整个页面的转换质量可以取各影响转换质量的因子在页面中所占权重的平均值即(0.3+0.3+0.2+0.2)/4＝0.25；也直接将各影响转换质量的因子所占权重值相乘得出整个页面的转换质量0.3*0.3*0.2*0.2＝0.0036。

[0051] 本实施例得到的转换质量可以用于后续移动终端页面质量的计算中。

[0052] 在步骤S104中，计算检索到的移动互联网页面和转换后的移动互联网页面的质量。

[0053] 在本实施例中，由于传统互联网页面转换成移动互联网页面之后，与其他的移动互联网页面具有类似的属性特征，因此所有的页面可以采用统一的规则来计算页面质量，并可以根据页面质量计算的结果排序输出各页面。

[0054] 由于目前移动终端设备具有屏幕小、网络带宽小、资源传输慢、多媒体资源展示不丰富、可交互性差等特点。因此，选择从以下至少一个方面：检索串与页面文本相关性、页面链接关系、页面作弊情况、页面的敏感度以及页面排版质量等对页面质量进行计算。

[0055] 具体说明如下：

[0056] 检索串与页面文本相关性：计算预处理后的检索串与页面标题、页面正文的文本相关性，相关性越高，则页面质量越高。

[0057] 页面链接关系：采用PageRank算法，根据页面的外部链接和内部链接的数量和质量来衡量页面的质量。

[0058] 页面作弊情况：从页面是否包含作弊信息，来判断页面质量，作弊信息越多，页面质量越低。所述作弊信息包括：隐藏/堆砌关键词、夹杂锚文本、页面标题和页面正文不相符等。

[0059] 页面的敏感度：从页面的色情程度、反政治程度等角度出发，判断页面质量。

[0060] 页面排版质量：从页面资源(图片/视频播放窗口的连通性)、页面的排版效果(标题/正文是否突出，是否包含悬浮框等)等角度，判断页面的质量。

[0061] 针对上述影响页面质量的因子，采用机器学习算法、分支定界方法或者阈值分支法等，建立样本集进行训练，根据训练得到的决策模型来确定各影响页面质量的因子在页面中所占的权重，并根据确定的权重，计算整个页面的质量。

[0062] 在本实施例中，对传统互联网页面转换后的页面，在计算其页面质量时，还需要考虑其页面转换质量，即将页面转换质量作为页面质量的一个影响因子，预先为其设置固定的权重(例如：0.1)。

[0063] 本发明实施例将检索到的传统互联网页面转换成移动互联网页面，由于传统互联网页面转换成移动互联网页面之后，与其他的移动互联网页面具有类似的属性特征，从而达到统一、公平的计算移动互联网页面和传统互联网页面质量的目的。另外，在移动终端页

面质量的计算中，考虑了传统互联网页面转换成移动互联网的难易程度，使得页面质量的计算更具公平性，传统互联网页面与移动互联网页面的排序更具合理性，从而极大的提高了移动终端用户对搜索结果的满意度。

[0064] 实施例二：

[0065] 图3示出了本发明实施例二提供的移动终端网页质量计算装置的组成结构，为了便于说明，仅示出了与本发明实施例相关的部分。

[0066] 该移动终端网页质量计算装置可以是运行于移动终端(例如手机、掌上电脑、个人数字助理等)内的软件单元、硬件单元或者软硬件相结合的单元，也可以作为独立的挂件集成到移动终端中或者运行于移动终端的应用系统中。

[0067] 该移动终端网页质量计算装置包括预处理单元31、检索单元32、转换单元33和页面质量计算单元34。其中，各单元的具体功能如下：

[0068] 预处理单元31，用于对用户输入的检索串进行预处理；

[0069] 检索单元32，用于检索与预处理后的检索串相关的页面，所述页面包括传统互联网页面和移动互联网页面；

[0070] 转换单元33，用于将检索到的传统互联网页面转换成移动互联网页面；

[0071] 页面质量计算单元34，用于计算检索到的移动互联网页面和转换后的移动互联网页面的质量。

[0072] 进一步的，所述装置还包括：

[0073] 转换质量计算单元35，用于根据传统互联网页面转换成移动互联网的难易程度计算转换质量。

[0074] 在本实施例中，所述转换包括但不限于以下至少一个：页面的标题识别、页面的主体内容识别、页面的核心正文识别、与页面正文相关的多媒体信息的识别以及与页面正文相关标签的替换和删除；所述转换质量的计算包括但不限于以下至少一个：页面排版质量的计算、页面主体内容质量的计算、页面交互性质量的计算、页面主体内容包含资源质量的计算；所述页面质量的计算包括但不限于以下至少一个：检索串与页面文本相关性的计算、页面链接关系的计算、页面作弊情况的计算、页面敏感度的计算以及页面排版质量的计算。

[0075] 本实施例提供的移动终端网页质量计算装置可以使用在前述对应的移动终端网页质量计算方法，详情参见上述移动终端网页质量计算方法实施例一的相关描述，在此不再赘述。

[0076] 本领域技术人员可以理解为上述装置所包括的各个单元只是按照功能逻辑进行划分的，但并不局限于上述的划分，只要能够实现相应的功能即可；另外，各功能单元的具体名称也只是为了便于相互区分，并不用于限制本发明的保护范围。

[0077] 综上所述，本发明实施例将检索到的传统互联网页面转换成移动互联网页面，由于传统互联网页面转换成移动互联网页面之后，与其他的移动互联网页面具有类似的属性特征，从而达到统一、公平的计算移动互联网页面和传统互联网页面质量的目的。而且，在移动终端页面质量的计算过程中，充分考虑到了移动终端设备屏幕小、网络带宽小、资源传输慢、多媒体资源展示不丰富、可交互性差的特点，选择符合移动终端设备特点的页面质量影响因子(如检索串与页面文本相关性、页面链接关系、页面作弊情况、页面的敏感度以及页面排版质量等)对页面质量进行计算。同时，在移动终端页面质量的计算中，考虑了传统

互联网页面转换成移动互联网的难易程度，使得页面质量的计算更具公平性，传统互联网页面与移动互联网页面的排序更具合理性，从而极大的提高了移动终端用户对搜索结果的满意度。

[0078] 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作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

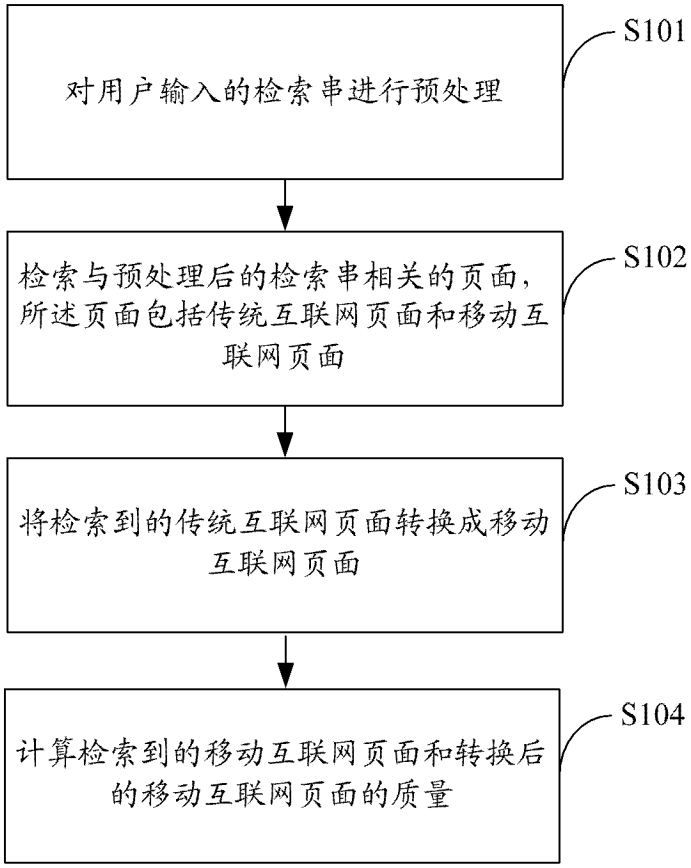

对用户输入的检索串进行预处理
S101
检索与预处理后的检索串相关的页面，所述页面包括传统互联网页面和移动互联网页面
S102
将检索到的传统互联网页面转换成移动互联网页面
S103
计算检索到的移动互联网页面和转换后的移动互联网页面的质量
S104

图1


腾讯新闻
下岗女工取款3千被给3万元 3次还款被银行赶出
搜索新体验
SOSO 搜搜
每日推荐

图2

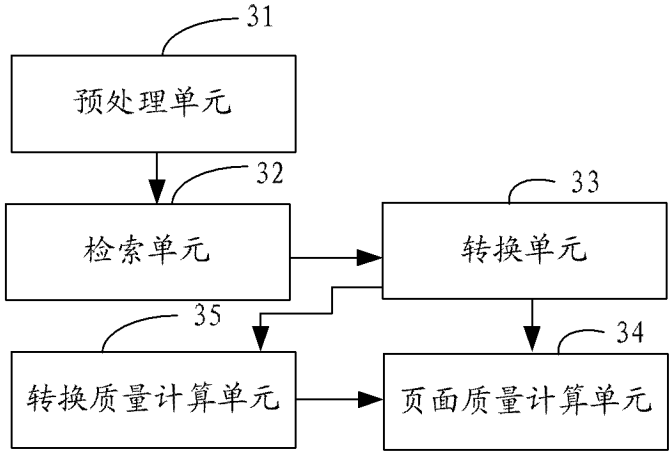


图3