



(12) 发明专利

(10) 授权公告号 CN 101499098 B

(45) 授权公告日 2012.07.11

(21) 申请号 200910118150.1

(22) 申请日 2009.03.04

(73) 专利权人 阿里巴巴集团控股有限公司

地址 英属开曼群岛大开曼岛

(72) 发明人 陈华

(74) 专利代理机构 北京同达信恒知识产权代理
有限公司 11291

代理人 魏杉

(51) Int. Cl.

G06F 17/30 (2006.01)

(56) 对比文件

US 2008/0256051 A1, 2008.10.16,

US 6658423 B1, 2003.12.02,

CN 101154224 A, 2008.04.02, 全文 .

审查员 欧阳琦

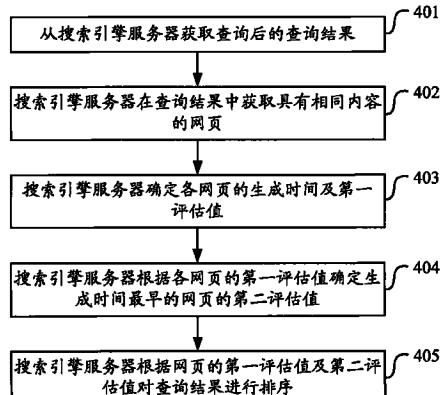
权利要求书 2 页 说明书 9 页 附图 5 页

(54) 发明名称

一种网页评估值的确定及运用的方法、系统

(57) 摘要

本申请公开了一种网页评估值的确定及运用的方法、系统，包括：从搜索引擎服务器获取具有相同或接近相同的内容的网页；确定各网页的生成时间及第一评估值；根据各网页的第一评估值确定生成时间最早的网页的第二评估值。进一步的，搜索引擎服务器根据网页的第二评估值对查询结果进行排序。使用本申请能够为搜索引擎的结果排序增加一个新的重要的排序参数，大幅度提高内容类查询词的搜索结果效果，能够使得用户在网页搜索中找内容类网页的查询结果满意度大幅度提高。



1. 一种利用计算机对网页评估值进行确定的方法,其特征在于,包括如下步骤:

从搜索引擎服务器系统获取具有相同或接近相同的内容的网页;

搜索引擎服务器系统确定所述各网页的生成时间及第一评估值,所述第一评估值为根据包括外链在内的数据而形成的评估值;

搜索引擎服务器系统根据所述各网页的第一评估值确定生成时间最早的网页的第二评估值,以根据各网页的所述第一评估值和所述第二评估值进行排序,所述第二评估值为与所述生成时间最早的网页具有相同或接近相同的内容的网页的第一评估值之和与第一加权系数的乘积加上所述生成时间最早的网页的第一评估值与第二加权系数的乘积。

2. 如权利要求1所述的方法,其特征在于,所述具有相同或接近相同的内容的网页,包括数字指纹相同的网页。

3. 如权利要求2所述的方法,其特征在于,所述获取具有相同或接近相同的内容的网页,包括:

获取各网页中非第一段和非最后一段的中间内容最长的段落或段落非第一句和非最后一句的最长句子,并生成数字指纹;

根据数字指纹确定各网页内容是否相同后获取具有相同或接近相同的内容的网页。

4. 如权利要求1所述的方法,其特征在于,所述确定所述各网页的生成时间,包括下列方式之一或者其组合:

根据网页统一资源定位符URL包含的时间确定;

根据内容类网页中的时间确定;

根据抓取网页的时间确定;

根据最早将网页收入索引的时间确定。

5. 如权利要求1所述的方法,其特征在于,所述第二评估值大于所述第一评估值。

6. 如权利要求5所述的方法,其特征在于,所述第一加权系数与所述第二加权系数的取值相同或不同。

7. 一种根据权利要求1至6任一项所述的网页评估值对搜索查询结果进行排序的方法,其特征在于,包括如下步骤:

从搜索引擎服务器系统获取查询后的查询结果;

搜索引擎服务器系统根据各网页的第一评估值及生成时间最早的网页的第二评估值对查询结果排序。

8. 如权利要求7所述的方法,其特征在于,进一步包括:

搜索引擎服务器在查询结果中显示每个网页的转载次数。

9. 一种搜索引擎服务器系统,其特征在于,包括:

爬虫系统,用于获取具有相同或接近相同的内容的网页;

索引系统,用于确定各网页的生成时间及各网页的第一评估值,并根据各网页的第一评估值确定生成时间最早的网页的第二评估值,以根据各网页的所述第一评估值和所述第二评估值进行排序,所述第一评估值为根据包括其他网页指向在内的数据而形成的评估值确定,所述第二评估值为与所述生成时间最早的网页具有相同或接近相同的内容的网页的第一评估值之和与第一加权系数的乘积加上所述生成时间最早的网页的第一评估值与第二加权系数的乘积。

10. 如权利要求 9 所述的搜索引擎服务器系统，其特征在于，所述索引系统进一步用于根据网页的数字指纹确定各网页是否具有相同或接近相同的内容。

11. 如权利要求 9 所述的搜索引擎服务器系统，其特征在于，所述索引系统包括：

数字指纹生成单元，用于获取各网页中非第一段和非最后一段的中间内容最长的段落或段落非第一句和非最后一句的最长句子，并生成数字指纹；

比较单元，用于根据数字指纹确定各网页内容是否相同；

获取单元，用于根据数字指纹确定各网页内容是否相同后，获取具有相同或接近相同的内容的网页。

12. 如权利要求 9 所述的搜索引擎服务器系统，其特征在于，所述索引系统进一步用于根据下列方式之一或者其组合确定网页生成时间：

网页统一资源定位符 URL 包含的时间；

内容类网页中的时间；

抓取网页的时间；

最早将网页收入索引的时间。

13. 如权利要求 9 所述的搜索引擎服务器系统，其特征在于，所述索引系统还用于根据各网页的第一评估值及第二评估值对查询结果排序。

14. 如权利要求 13 所述的搜索引擎服务器系统，其特征在于，所述索引系统进一步用于在查询结果中显示每个网页的转载次数。

一种网页评估值的确定及运用的方法、系统

技术领域

[0001] 本申请涉及信息处理技术,特别涉及一种利用计算机对网页评估值进行确定及运用的方法、系统。

背景技术

[0002] 搜索引擎从互联网上抓取网页,在用户查询时,会找到满足用户关键字的所有网页,然后按照相关度排序,以便排在前面的搜索结果更符合用户的需求。由于相关度是一个非常复杂、基于很多参数计算出来的结果,因而也就存在着利用各种各样的算法和参数来计算相关度的技术方案,并且,一般来说各个搜索引擎厂商用的参数和算法也都各不相同。

[0003] 例如,Google 在 1997 年提出了 Page rank 这种提高相关度算法的参数以及计算这个参数的算法。大概可以这样理解 Page rank,重要网页链接出去的目标网页,会获得重要的权值,被越多重要网页指向的网页,page rank 越高,也就越重要。

[0004] 现有搜索引擎在处理内容类查询词的排序时,普遍解决的不够好。重点体现在如下两点:

[0005] 1、使用外链计算 Page rank 以判断重要网页的方式基本不起作用,导致排在搜索结果前面的结果,很大程度上并不是用户最想看到的结果。

[0006] 2、现有的搜索引擎通常用排重技术处理内容相同或接近相同的网页。例如,在抓取网页时并不储存某些内容重复的网页,或在收到用户搜索请求后不显示某些内容重复的网页或将该等内容排在搜索结果的后面。如果没有合适的链接数据,搜索引擎有可能根据 Page rank 算法将原创网页忽略或排后,而把转载的网页排在前面。因此,现有的搜索引擎并没有考虑到内容相同的不同网页对结果排序的影响。

发明内容

[0007] 本申请提供一种利用计算机对网页评估值进行确定及运用的方法、系统,用以提高对查询结果的反馈准确性。

[0008] 本申请实施例中提供了一种利用计算机对网页评估值进行确定方法,包括如下步骤:

[0009] 从搜索引擎服务器系统获取具有相同或接近相同的内容的网页;

[0010] 搜索引擎服务器系统确定所述各网页的生成时间及第一评估值,所述第一评估值为根据包括外链在内的数据而形成的评估值;

[0011] 搜索引擎服务器系统根据所述各网页的第一评估值确定生成时间最早的网页的第二评估值,以根据各网页的所述第一评估值和所述第二评估值进行排序,所述第二评估值为与所述生成时间最早的网页具有相同或接近相同的内容的网页的第一评估值之和与第一加权系数的乘积加上所述生成时间最早的网页的第一评估值与第二加权系数的乘积。

[0012] 较佳地,所述具有相同或接近相同的内容的网页,包括数字指纹相同的网页。

[0013] 较佳地,所述获取具有相同或接近相同的内容的网页,包括:

- [0014] 获取各网页中非第一段和非最后一段的中间内容最长的段落或段落非第一句和非最后一句的最长句子，并生成数字指纹；
- [0015] 根据数字指纹确定各网页内容是否相同后获取具有相同或接近相同的内容的网页。
- [0016] 较佳地，所述确定所述各网页的生成时间，包括下列方式之一或者其组合：
- [0017] 根据网页统一资源定位符 URL 包含的时间确定；
- [0018] 根据内容类网页中的时间确定；
- [0019] 根据抓取网页的时间确定；
- [0020] 根据最早将网页收入索引的时间确定。
- [0021] 较佳地，所述第二评估值大于所述第一评估值。
- [0022] 较佳地，所述第一加权系数与所述第二加权系数的取值相同或不同。
- [0023] 本申请还提供了一种根据网页评估值对搜索查询结果进行排序的方法，包括如下步骤：
- [0024] 从搜索引擎服务器系统获取查询后的查询结果；
- [0025] 搜索引擎服务器系统根据各网页的第一评估值及生成时间最早的网页的第二评估值对查询结果排序。
- [0026] 较佳地，进一步包括：
- [0027] 搜索引擎服务器在查询结果中显示每个网页的转载次数。
- [0028] 本申请提供了一种搜索引擎服务器系统，包括：
- [0029] 爬虫系统，用于获取具有相同或接近相同的内容的网页；
- [0030] 索引系统，用于确定各网页的生成时间及各网页的第一评估值，并根据各网页的第一评估值确定生成时间最早的网页的第二评估值，以根据各网页的所述第一评估值和所述第二评估值进行排序，所述第一评估值为根据包括其他网页指向在内的数据而形成的评估值确定，所述第二评估值为与所述生成时间最早的网页具有相同或接近相同的内容的网页的第一评估值之和与第一加权系数的乘积加上所述生成时间最早的网页的第一评估值与第二加权系数的乘积。
- [0031] 较佳地，所述索引系统进一步用于根据网页的数字指纹确定各网页是否具有相同或接近相同的内容。
- [0032] 较佳地，所述索引系统包括：
- [0033] 数字指纹生成单元，用于获取各网页中非第一段和非最后一段的中间内容最长的段落或段落非第一句和非最后一句的最长句子，并生成数字指纹；
- [0034] 比较单元，用于根据数字指纹确定各网页内容是否相同；
- [0035] 获取单元，用于根据数字指纹确定各网页内容是否相同后，获取具有相同或接近相同的内容的网页。
- [0036] 较佳地，所述索引系统进一步用于根据下列方式之一或者其组合确定网页生成时间：
- [0037] 网页统一资源定位符 URL 包含的时间；
- [0038] 内容类网页中的时间；
- [0039] 抓取网页的时间；

- [0040] 最早将网页收入索引的时间。
- [0041] 较佳地，所述索引系统还用于根据各网页的第一评估值及第二评估值对查询结果排序。
- [0042] 较佳地，所述索引系统进一步用于在查询结果中显示每个网页的转载次数。
- [0043] 本申请有益效果如下：
- [0044] 在本申请实施中，首先获取具有相同或接近相同的内容的网页；然后确定各网页的生成时间及评估值；最后再根据各网页的评估值确定生成时间最早的网页的评估值。
- [0045] 由于在方案中通过对生成时间这一参数考虑到了网页是否为原创，从而确定了与生成时间为依据的、判断网页实际评估值的方案，因此克服了在使用外链计算Page rank以判断重要网页的方式时，导致排在搜索结果前面的结果并不能代表其评估值的问题。
- [0046] 进一步的，还充分利用了内容相同的不同网页之间的评估值之间的关系，并将其用于改进搜索结果的排序，因此提高了查询结果反馈的准确性。

附图说明

- [0047] 图 1 为本申请实施例中 Copy Rank 在搜索引擎结果中的效果示意图；
- [0048] 图 2 为本申请实施例中网页评估值的确定方法实施流程示意图；
- [0049] 图 3 为本申请实施例中转载网页与原创网页 Copy Rank 关系示意图；
- [0050] 图 4 为本申请实施例中根据网页评估值对查询结果进行排序的方法实施流程示意图；
- [0051] 图 5 为本申请实施例中搜索引擎服务器系统结构示意图；
- [0052] 图 6 为本申请实施例中搜索引擎服务器系统运用环境结构示意图；
- [0053] 图 7 为本申请实施例中利用计算机对网页搜索查询结果进行排序的方法实施流程示意图；
- [0054] 图 8 为本申请实施例中搜索引擎服务器系统结构示意图。

具体实施方式

- [0055] 下面结合附图对本申请具体实施方式进行说明。
- [0056] 发明人在发明过程中注意到：
- [0057] 1、内容类网页往往外链很少，因此使用外链计算 Page rank 以判断重要网页的方式基本不起作用，从而导致排在搜索结果前面的结果，很大程度上并不是用户最想看到的结果。
- [0058] 2、对于内容不同的不同网页，搜索引擎都把他们当做干扰搜索结果的负面因素，要么被搜索引擎直接扔掉，要么将 page rank 降的很低。但其实这些内容相同的不同网页，对于改进搜索结果排序具有非常重要的作用。
- [0059] 鉴于此，本申请提出了为搜索引擎的结果排序增加一个新的、重要的排序参数，大幅度提高内容类查询词的搜索结果效果的技术方案。使得网页搜索中找文章的查询结果满意度大幅度提高。下面先对网页评估值的确定实施方式进行说明，在对将该网页评估值运用于返回查询结果以提高搜索准确性的实施方式进行说明。
- [0060] 实施中，借用 Google 对某个网页的重要性评估的评估值 Page Rank 的概念，将本

申请中网页的评估值称为 Copy rank, 其代表了一种用于改进搜索引擎相关度排序的参数和产生这个参数的算法, 适用于优化内容类查询的搜索结果排序。它利用互联网上文章的转载次数, 计算原创网页的 Copy rank, 并对转载网页进行聚合。搜索引擎在计算相关度时, 综合 page rank、关键词匹配程度等传统计算相关度的参数和 Copy rank, 一起计算出一个新的相关度值。在搜索引擎显示结果时, 也显示转载数目, 以帮助用户最快判断互联网上符合此查询词的最佳结果。

[0061] 图 1 为 Copy Rank 在搜索引擎结果中的效果示意图, 如图所示搜索结果, 版本 (转载次数) 越多的文章, 越有可能是用户想要看到的文章。

[0062] Copy Rank 的确定主要包括三个因素, 一是判断网页内容是否基本相同; 二是判断网页的真实发布时间; 三是判断谁是原创网页, 下面进行说明。

[0063] 图 2 为网页评估值的确定方法实施流程示意图, 如图所示, 在进行评估值确定时可以包括如下步骤:

[0064] 步骤 201、从搜索引擎服务器系统获取具有相同或接近相同的内容的网页;

[0065] 步骤 202、搜索引擎服务器系统确定各网页的生成时间及第一评估值;

[0066] 步骤 203、搜索引擎服务器系统根据相同或接近相同的内容的各网页的第一评估值确定生成时间最早的网页的第二评估值。

[0067] 实施中, 在步骤 201 中, 具有相同或接近相同的内容的网页包括数字指纹相同的网页。

[0068] 则获取具有相同或接近相同的内容的网页, 可以包括:

[0069] 从搜索引擎服务器获取各网页中非第一段和非最后一段的中间内容最长的段落或段落非第一句和非最后一句的最长句子, 并生成 MD5;

[0070] 根据数字指纹确定各网页内容是否相同后获取具有相同或接近相同的内容的网页。

[0071] MD5 是 message-digest algorithm 5(信息 - 摘要算法) 的缩写, 被广泛用于加密和解密技术上, 它可以说是文件的“数字指纹”。任何一个文件, 无论是可执行程序、图像文件、临时文件或者其他任何类型的文件, 也不管它体积多大, 都有且只有一个独一无二的 MD5 信息值, 并且如果这个文件被修改过, 它的 MD5 值也将随之改变。因此, 实施中可以通过 MD5 来确定各网页内容是否具有相同或接近相同的内容, 即, 通过对比同一文件的 MD5 值, 来校验这个文件是否被“篡改”过。MD5 的作用在于: 当下载了文件后, 如果想知道下载的这个文件和网站的原始文件是否一样, 就可以给下载的文件做个 MD5 校验。如果得到的 MD5 值和网站公布的相同, 可确认所下载的文件是完整的。如有不同, 说明你下载的文件是不完整的: 要么就是在网络下载的过程中出现错误, 要么就是此文件已被修改。一般正规的站点, 都会提供文件 md5 校验码。

[0072] 判断网页内容是否相同, 具体采用的办法可以是在所有文章类网页中寻找非第一段和非最后一段的中间最长段落, 生成 MD5 作为网页指纹, 作为判断相同的依据。对于只有两个以内段落的文章, 取段落非第一句和非最后一句的最长句子, 生成 MD5 作为网页指纹, 作为判断相同的依据。如果两个网页的网页指纹一样, 则说明两个网页的整篇内容是相同的。

[0073] 具体实施中, 寻找非第一段和非最后一段的中间最长段落, 以及取段落非第一句

和非最后一句的最长句子生成 MD5 作为网页指纹,是因为发明人在发明过程中注意到:通常第一段和最后一段、第一句和最后一句被改动的频率很高,并不能代表文章的真实内容,因此选用非第一段和非最后一段、非第一句和非最后一句来生成 MD5。

[0074] 实施中,可以通过 MD5 来判断两个文件之间是否相同,本领域技术人员易知,当在执行步骤 201 获取具有相同或接近相同的内容的网页时,并不仅限于采用通过 MD5 判断内容一致的方式,其他能够比较出两个网页内容是否一致的技术手段均可采用,其最终目的在于当存在内容相同的不同网页时,不会把他们当做干扰搜索结果的负面因素来直接扔掉,并将其用于改进搜索结果排序。

[0075] 在步骤 202 中,在确定网页的生成时间时可以包括:

[0076] 根据网页统一资源定位符 URL 包含的时间,和 / 或,根据文章类网页中的时间确定网页生成时间。

[0077] 实施中,判断网页的真实发布时间,可以采用计算机程序抽取的方式获得。由于目前大部分网站的网页都是动态生成的,因而网页服务器返回的 Last-modified(最后修改时间) 字段已经没有什么意义,因此可以从网页正文等处抽取时间。抽取时间可以按以下算法:

[0078] 首先判断 URL 中是否含有时间,例如下面的一个例子中的 URL(Uniform Resource Locator,统一资源定位符) 中便含有时间:

[0079] <http://news.sina.com.cn/w/2009-01-15/184017052431.shtml>;

[0080] 然后通过程序便有可能把 2009-01-15 抽取出来。实施中,具体的抽取手段可以包括:A、列举常用的时间格式,并建立时间格式维表用以存储常用的时间格式;B、按照分割符对 URL 进行切分;C、将切分后的每一部份在时间格式维表中进行查询,若与该维表中的时间格式相匹配,则说明该 URL 中含有时间,便可以提取该时间。

[0081] 如果 URL 中没有时间,则从文章正文中获取。文章正文中的时间格式有很多种,实施中只要根据实际情况将计算机程序考虑周全,便可以尽快找到时间。如下面的例子中文章正文中便含有时间:

[0082] 2009 年 01 月 15 日 18:40 中国网

[0083] 2009 年 12 月 27 日 23:35

[0084] 通过程序便可以很容易的把 2009 年 12 月 27 日 23:35 抽取出来。

[0085] 实施中,在具体的实现手段上,可以通过分析网页中各种时间、日期格式的代码,用正则表达式匹配等任意程序方式来进行获取。如果程序不能确定生成时间,则取当前抓取的时间作为生成时间。实施中不论如何实现计算生成时间,其目的在于将获取的生成时间用于识别各个转载网页的原创版本。

[0086] 实施中可以在抓取网页、建立网页索引时就判断生成时间,并将生成时间储存在网页索引的一个字段(FIELD)里。

[0087] 实施中,当存在无法从文章或 URL 中抽取生成时间的情况时,可以使用抓取网页的时间作为生成时间,也可以把最早收入索引的时间假定为文章生成时间。

[0088] 在通过上述方式确定出内容相同的网页以及其生成时间后,便可以判断出谁是原创网页,即,在所有相同网页中,找到真实发布时间最早的网页,即为原创网页。

[0089] 下面对步骤 202 中的评估值进行说明。

[0090] 首先对 Page Rank 进行说明,以便更深入的理解本申请中所定义的 Copy Rank, Page Rank 是 Google 对某个网页的重要性评估的评估值,是 Page Rank,而不是“ Site Rank(网站评估值)”,不是对整个网站的评估值。如果一个网站首页的 Page Rank 是 5,那只是说首页那个页面的 Page Rank 是 5,而不是说整个网站是 5。Google 的 Page Rank 不针对网站而言,只针对页面,一个个的页面。

[0091] 某个页面的 Page Rank 值,主要来自于指向这个页面的所有链接所代表的那些页面。所谓“所有链接”包括两部分:本网站之外的外部链接和本网站内的其他页面的内部链接。也就是说,任何一个页面的 Page Rank 值,是由外部链接和内部链接共同作用而产生的。而不只是由外部链接或只由内部链接单方面作用而产生。假设一个网站的首页因为有两个 Page Rank 为 5 的外部链接指过来,加上还有更多的内部链接指向首页,才使网站首页的 Page Rank 为 5。

[0092] 同样道理,在本申请实施中,当在步骤 203 根据相同或接近相同的内容的各网页的评估值确定生成时间最早的网页的评估值时,便可以确定第二评估值为与生成时间最早的网页具有相同或接近相同的内容的网页的第一评估值之和与第一加权系数的乘积加上所述生成时间最早的网页的第一评估值与第二加权系数的乘积。

[0093] 也就是说,Copy Rank 使得原创网页获得了所有转载网页的权重,即,Copy Rank 可按如下公式计算:

[0094] 原创网页的 Copy Rank = Σ 每个转载网页的 Page Rank*w1+ 原创网页的 Page Rank*w2;其中 W1 和 W2 为加权系数,W1 和 W2 的值可在实施中按照需要自行设定,并且 W1 和 W2 的取值可以相同也可以不同。

[0095] 需要说明的是,本申请实施例中用以说明评估值的是 Page Rank,但是,实际上根据包括其他网页指向在内的数据而形成的评估值都可以用来实现。

[0096] 另外,实施中可以在抓取网页下来后生成 Copy Rank,也可以定期更新所有网页的 CopyRank。

[0097] 实施中,在确定原创网页时还可以根据历史数据或经验建立一个网站黑名单和/或白名单,属于白名单上网站的网页假定为原创网页,而属于黑名单上网站的网页假定为非原创网页。

[0098] 图 3 为转载网页与原创网页 Copy Rank 关系示意图,如图所示,将外链网页给所有转载网页的评估值权重,全部给了原创网页,相当于从外部看,这些网页外链的评估值都给了原创网页。

[0099] 图 4 为根据网页评估值对查询结果进行排序的方法实施流程示意图,如图所示,在将网页评估值运用于返回查询结果以提高搜索准确性的实施过程中可以包括如下步骤:

- [0100] 步骤 401、从搜索引擎服务器系统获取查询后的查询结果;
- [0101] 步骤 402、搜索引擎服务器系统在查询结果中获取具有相同或接近相同的内容的网页;
- [0102] 步骤 403、搜索引擎服务器系统确定各网页的生成时间及第一评估值;
- [0103] 步骤 404、搜索引擎服务器系统根据各网页的第一评估值确定生成时间最早的网页的第二评估值;

- [0104] 步骤 405、根据各网页的第一评估值及第二评估值对查询结果排序。
- [0105] 搜索引擎服务器系统在步骤 405 的实施中便可以根据评估值对查询到的网页排序,比如按评估值大小排序后依次返回并显示给查询的用户。
- [0106] 进一步的,搜索引擎服务器系统还可以在查询结果中显示每个网页的转载次数。
- [0107] 基于同一申请构思,本申请实施例中还提供了一种搜索引擎服务器系统,由于系统解决问题的原理与网页评估值的确定方法、根据网页评估值返回查询结果的方法相似,因此系统的实施可以参见方法的实施,重复之处不再赘述。
- [0108] 图 5 为搜索引擎服务器系统结构示意图,如图所示,搜索引擎服务器系统中可以包括:
- [0109] 爬虫系统 501,用于获取具有相同或接近相同的内容的网页;
- [0110] 索引系统 502,用于确定各网页的生成时间及各网页的第一评估值,并根据各网页的第一评估值确定生成时间最早的网页的第二评估值。
- [0111] 实施中,索引系统可以进一步用于根据网页的 MD5 确定各网页是否具有相同或接近相同的内容。
- [0112] 网页获取模块中可以包括:
- [0113] MD5 生成单元,用于获取各网页中非第一段和非最后一段的中间内容最长的段落或段落非第一句和非最后一句的最长句子,并生成 MD5;
- [0114] 比较单元,用于根据 MD5 确定各网页内容是否相同;
- [0115] 获取单元,用于根据 MD5 确定各网页内容是否相同后,获取具有相同或接近相同的内容的网页。
- [0116] 实施中,索引系统可以进一步用于根据网页 URL 包含的时间,和 / 或,根据内容类网页中的时间确定网页生成时间。
- [0117] 实施中,索引系统还可以进一步用于根据各网页的第一评估值确定生成时间最早的网页的第二评估值时,确定所述第二评估值为与所述生成时间最早的网页具有相同或接近相同的内容的网页的第一评估值之和与第一加权系数的乘积加上所述生成时间最早的网页的第一评估值与第二加权系数的乘积。
- [0118] 实施中,索引系统还可以进一步用于根据包括外链在内的数据而形成的评估值确定各网页的第一评估值。
- [0119] 索引系统还可以用于根据根据各网页的第一评估值及第二评估值对查询结果排序。
- [0120] 实施中,索引系统还可以进一步用于在查询结果中显示每个网页的转载次数。
- [0121] 图 6 为搜索引擎服务器系统运用环境结构示意图,如图所示,网络中包括有根据网页评估值对查询结果进行排序的索引系统 601、网页 602(代表产生网页的各种实体,具体的网页可以表现为服务器等,实施例中用网页来指代这类实体仅是为了描述方便,同时,这类实体可以有很多,图中仅以一个示意)、用户端 603(图中仅用一个示意)、爬虫系统 604、查询系统 605。
- [0122] 由图也可见,索引系统 601 与爬虫系统 604 也构成了搜索引擎服务器系统,需要说明的是,图中各功能实体的连接方式有通过网络连接,也有以直线表示的直接连接,但是,该图仅为示意图,实际实施中,可以根据实际需要进行网络架构,比如:爬虫系统与索引系

统通过因特网连接,而非局域网连接等,只要各实体之间能实现数据交互的连接方式均可实施本申请。

[0123] 实施中,网页 602 提供各种网页内容,爬虫系统 604 可以在网络中采集各种网页信息,并将网页信息储存在一个或多个服务器上。本申请中的索引系统 601 根据采集到的网页信息建立索引,以便快速处理查询请求。索引系统 601 还可以确定网页的第一评估值和第二评估值,并根据上述第一评估值、第二评估值进行网页的排序。所述排序可以在爬虫系统采集到网页信息之后立刻进行,也可以在收到用户端的查询请求之后再进行,本申请对此并不做限定。

[0124] 当用户端 603 通过网络到查询系统 605 中进行查询时,查询系统 605 便可以根据排序装置 601 的排序结果,将用户端 603 所需的信息返回,使得用户获得的查询结果排序准确,能够真实反映查询结果之间的关系。

[0125] 由上述实施例可以看出,本申请在实施时使用了内容被转载的次数和基于转载计算出来的 Copy Rank 值,Copy rank 是能够改进搜索引擎相关度排序的参数,适用于优化内容类查询的搜索结果排序。能够利用互联网上文章的转载次数计算原创网页的 Copy rank,并对转载网页进行聚合,因此在搜索引擎计算相关度时,便可以综合如 page rank 等根据包括外链在内的数据而形成的评估值、关键词匹配程度等传统计算相关度的参数和 Copy rank 一起计算出一个新的相关度值,在搜索引擎显示结果时,也显示转载数目,以帮助用户最快判断互联网上符合此查询词的最佳结果,因而能够提高搜索引擎返回结果的准确性。本领域技术人员易知,搜索引擎包括网页搜索引擎、图片搜索引擎、软件搜索引擎等,本申请的技术方案能够提高搜索引擎结果的准确性,包括对搜索结果排序顺序的影响(使得转载次数更高的结果排在前面),也包括对搜索结果界面的影响(在结果页面上显示转载的次数,在结果页面上优先展示原创内容)等。

[0126] 本申请实施例中为搜索引擎的结果排序增加一个新的重要的排序参数,大幅度提高内容类查询词的搜索结果效果,能够使得用户在网页搜索中找文章的查询结果满意度大幅度提高。

[0127] 为了描述的方便,以上所述装置的各部分以功能分为各种模块或单元分别描述。当然,在实施本发明时可以把各模块或单元的功能在同一个或多个软件或硬件中实现。

[0128] 图 7 为利用计算机对网页搜索查询结果进行排序的方法实施流程示意图,如图所示,当进行排序时可以包括如下步骤:

[0129] 步骤 701、从搜索引擎服务器系统获取具有相同或接近相同的内容的网页;

[0130] 步骤 702、搜索引擎服务器系统确定所述各网页的生成时间;

[0131] 步骤 703、搜索引擎服务器系统根据所述各网页的生成时间的先后顺序进行排序。

[0132] 进一步的,还可以包括:搜索引擎服务器系统根据所述各网页的生成时间以及外链数据进行排序。

[0133] 图 8 为搜索引擎服务器系统结构示意图,如图所示,包括:

[0134] 爬虫系统 801,用于从搜索引擎服务器系统获取具有相同或接近相同的内容的网页;

[0135] 搜索引擎服务器 802,用于搜索引擎服务器系统确定所述各网页的生成时间,并根据所述各网页的生成时间的先后顺序进行排序。

[0136] 搜索引擎服务器 802 还可以进一步用于根据所述各网页的生成时间以及外链数据进行排序。

[0137] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品形式。

[0138] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0139] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0140] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0141] 尽管已描述了本申请的优选实施例,但本领域内的技术人员一旦得知了基本创造性概念,则可对这些实施例作出另外的变更和修改。所以,所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

[0142] 显然,本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样,倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内,则本申请也意图包含这些改动和变型在内。

山水美 山水风华的QQ空间 明朝那些事

1347年——1352年做和尚主要工作是撞钟 1352年——1368年造反(这个猛) 1368年——1398年主要工作是做皇帝 一切的事情都从1328年的那个夜晚开始,农民朱五四的妻子陈氏生下了一个男婴,大家都知道了,这个男婴就是后来的朱元璋。大凡皇帝出世,...

user.qzone.qq.com/736002462/blog/1193997150 4K 2007-12-12 - 百度快照

少年朱元璋 当年明月--维普资讯网

一切的事情都从1328年的那个夜晚开始,农民朱五四的妻子陈氏生下了一个男婴,大家都知道了,这个男婴就是后来的朱元璋。大凡皇帝出世,后来的史书上都会有一些类似的怪象记载。比如刮风啊,下暴雨啊,冒香气啊,天上星星闪啊,到处放红光啊,反正就...

www.cqvip.com/qk/88654X/200705/24418371.html 29K 2008-11-20 - 百度快照

99书城 99读书人

一切事情都从1328年的那个夜晚开始,农民朱五四的妻子陈氏生下了一个男婴,大家都知道了,这个男婴就是后来的朱元璋。大凡皇帝出世,后来的史书上都会有一些类似的怪象记载。比如刮风啊,下暴雨啊,冒香气啊,天上星星闪啊,到处放红光啊,反正...

www.99shucheng-99dushuren.com/ 35K 2009-2-26 - 百度快照

明朝那些事儿-当年明月-电子书

《明朝那些事儿》这篇文主要讲述的是从1344年到1644年这三百年间关于明朝的一些事情。作者:当年明月中国友谊出版公司 ...第一章一切的事情都从1328年的那个夜晚开始,农民朱五四的妻子陈氏生下了一个男婴,大家都知道了,这个男婴就是后来的朱...

www.du8.com/books/nov821.html 18K 2008-12-8 - 百度快照

明朝那些事儿(第壹部)-洪武大帝/当年明月-图书-卓越亚马逊

三百年明朝那些事儿。从我们的第一位主人公写起,要写三百多年,希望我能写完!说起来,我也写了不少东西了,本来只是...一切的事情都从1328年的那个夜晚开始,农民朱五四的妻子陈氏生下了一个男婴,大家都知道了,这个男婴就是后来的朱元璋。...

图 1

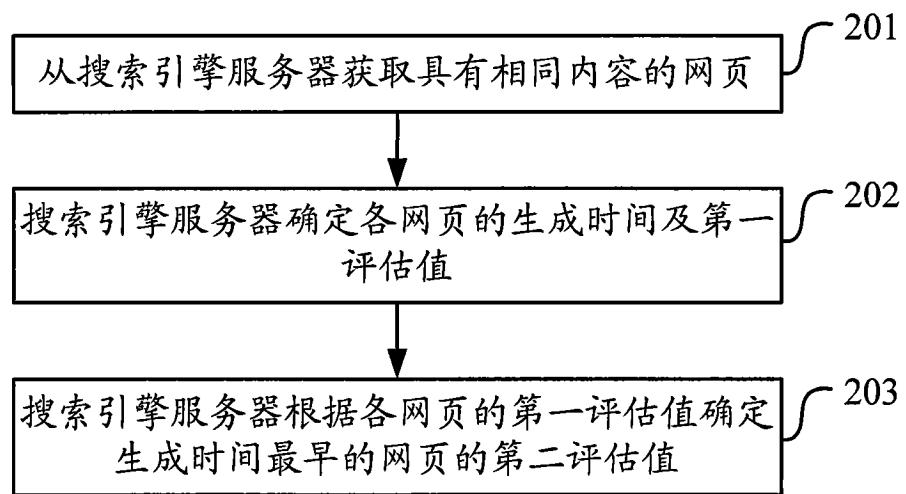


图 2

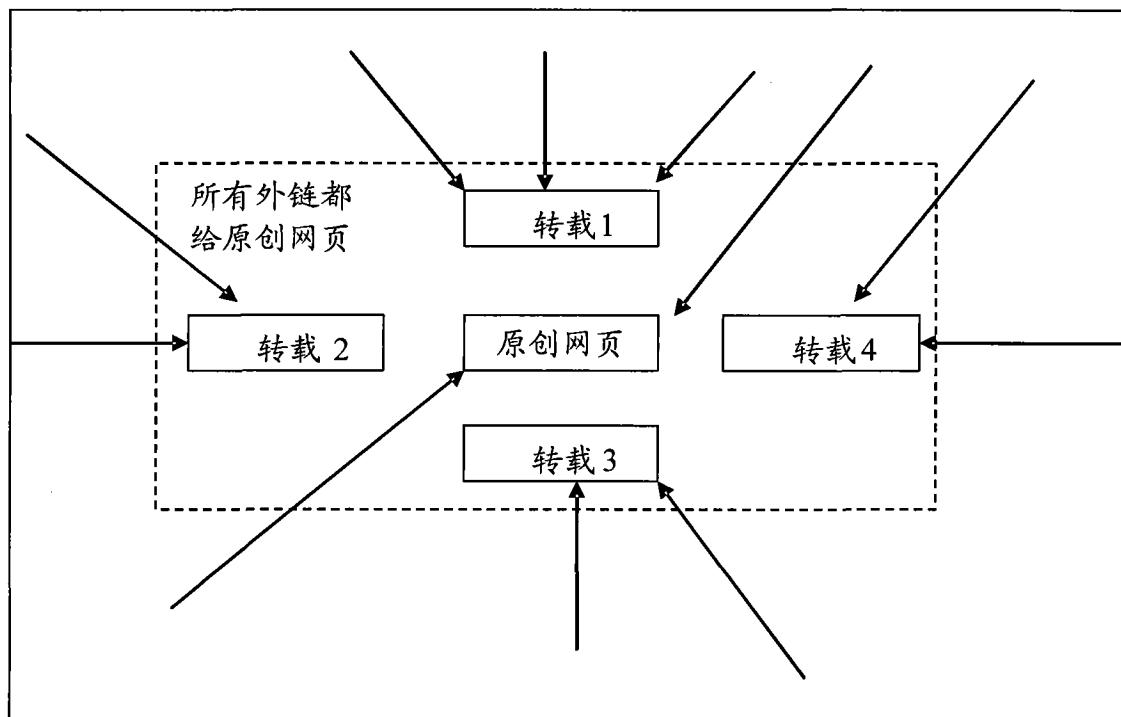


图 3

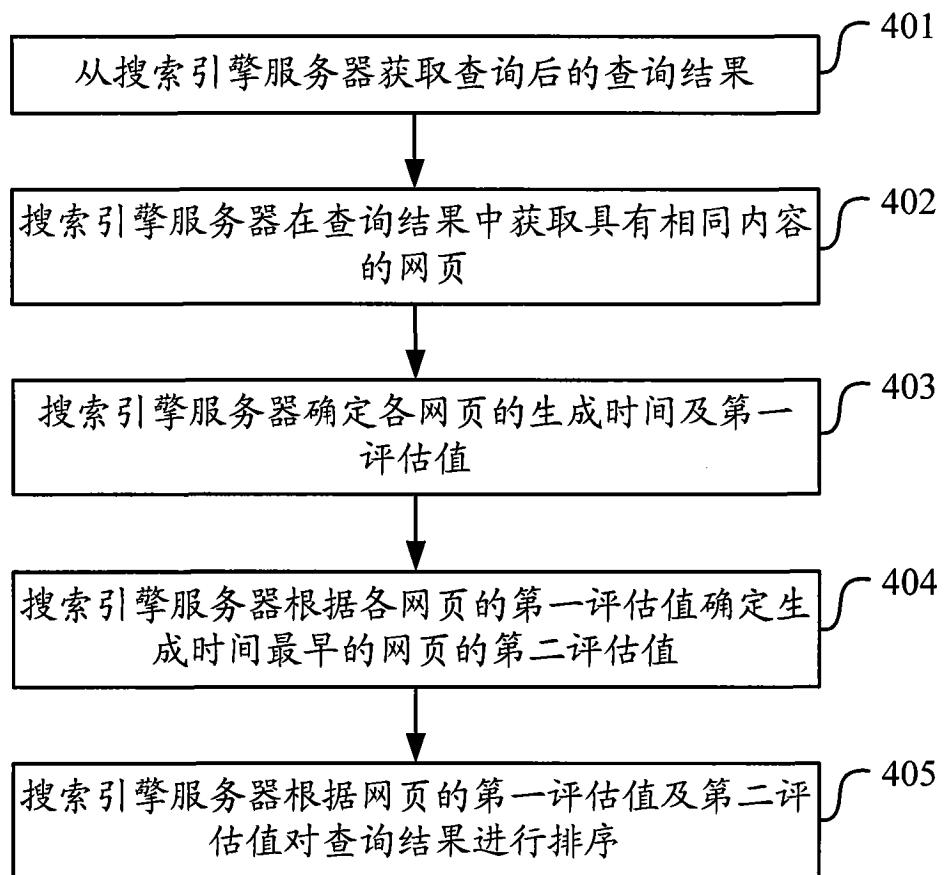


图 4

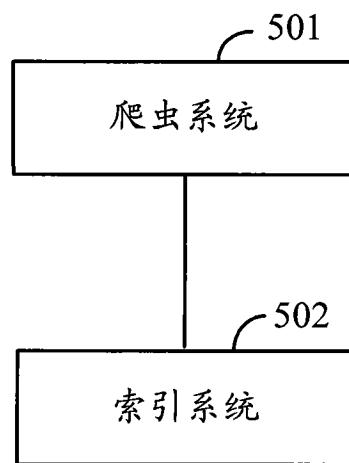


图 5

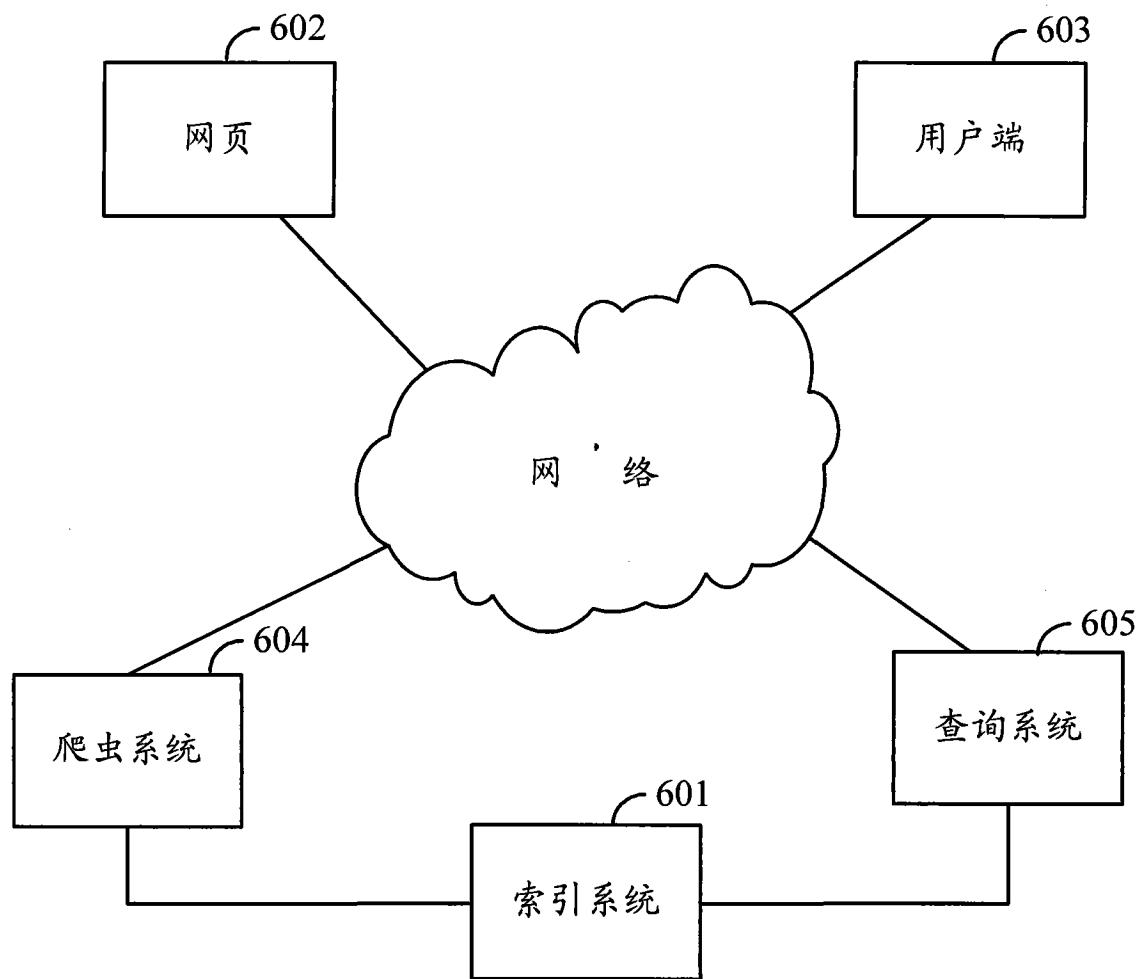


图 6

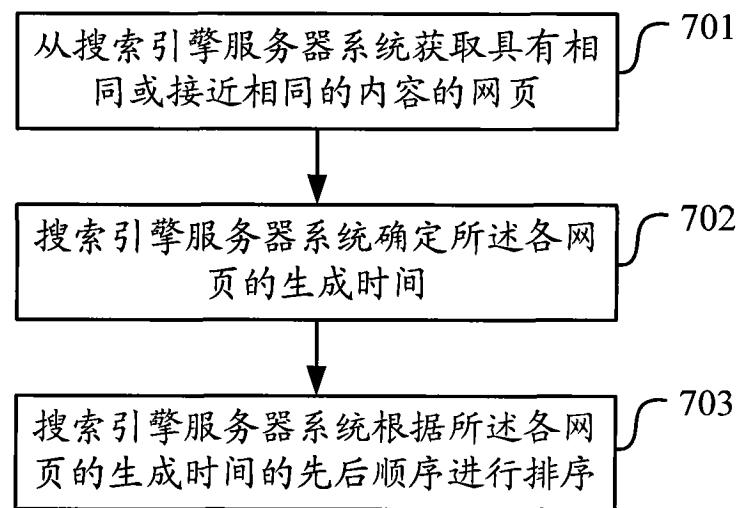


图 7

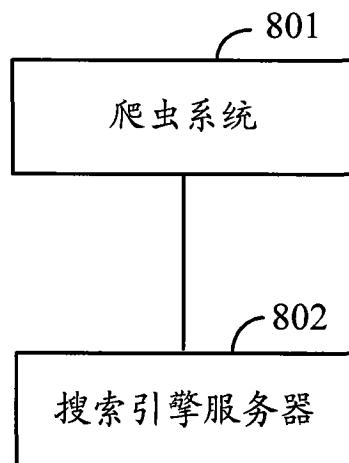


图 8