(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 103678597 A
(43)申请公布日 2014.03.26

(21)申请号 201310684066.2

(22)申请日 2013.12.13

(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 侯小虎

(74)专利代理机构 北京智汇东方知识产权代理事务所(普通合伙) 11391
代理人 康正德 薛峰

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

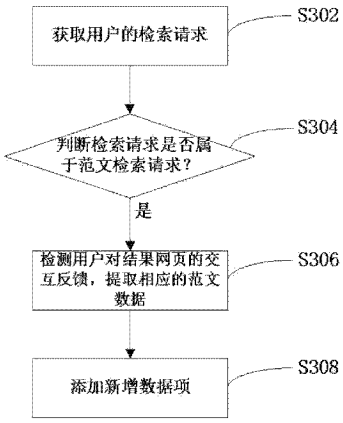
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种范文网页数据库的优化方法和装置

(57)摘要

本发明提供了一种范文网页数据库的优化方法,包括:获取用户的检索请求;根据所述检索请求中携带的关键词,判断所述检索请求是否属于范文检索请求;如果是,检测用户在检索结果页中对结果网页的交互反馈,根据所述交互反馈提取对应的结果网页中的范文数据;将所述关键词、所述结果网页中的范文数据及结果网页URL作为新增数据项添加到范文网页数据库中。根据本发明所述方法,提供了一种及时补充范文网页数据库的机制,不断扩充数据库内容以满足更多用户的需求,提升召回率。另外,本发明还提供了一种相应的优化装置。



CN 103678597 A

1. 一种范文网页数据库的优化方法，包括：

获取用户的检索请求；

根据所述检索请求中携带的关键词，判断所述检索请求是否属于范文检索请求；

如果是，检测用户在检索结果页中对结果网页的交互反馈，根据所述交互反馈提取对应的结果网页中的范文数据；

将所述关键词、所述结果网页中的范文数据及结果网页 URL 作为新增数据项添加到范文网页数据库中。

2. 根据权利要求 1 所述的优化方法，其中，添加步骤包括

判断结果网页 URL 是否包含在范文网页数据库的现有范文网页数据项中；

如果是，则不添加所述新增数据项；

如果否，则添加所述新增数据项。

3. 根据权利要求 1 或 2 所述的优化方法，其中新增数据项中的关键词对应于范文网页数据项的范文类型；其中添加步骤包括

在范文网页数据库中，确定与新增数据项具有相同范文类型的现有范文网页数据项的项数；

如所述项数小于预定数量，则添加所述新增数据项；

如所述项数大于或等于预定数量，则不添加所述新增数据项。

4. 根据权利要求 1-3 任一项所述的优化方法，其中被提取的结果网页是问答社区网页，包括提出问题的主楼块和回答问题的次楼块；其中提取范文数据的步骤包括

将所述关键词与网页主楼块的文字内容进行匹配；

如匹配，判断次楼块的文字内容的字数是否大于预定阈值；

如果是，确定字数大于预定阈值的次楼块为待提取次楼块；且

提取该结果网页的范文数据；其中所述范文数据包括：待提取次楼块的文字内容的标题，待提取次楼块的文字内容的正文，待提取次楼块的文字内容的字数。

5. 根据权利要求 1-4 任一项所述的优化方法，确定待提取次楼块的步骤还包括：

根据所述关键词确定元关键词；

将所述元关键词与字数大于预定阈值的次楼块的文字内容进行匹配；

如匹配，确定匹配的次楼块为待提取次楼块。

6. 根据权利要求 1-5 任一项所述的优化方法，所述网页中待提取次楼块为多个，则所述网页对应的范文网页数据项包括与待提取次楼块数量相应的多个范文数据。

7. 根据权利要求 1-6 任一项所述的优化方法，其中被提取的结果网页是文字网站网页，其网页包括正文标题和正文内容；其中提取范文数据的步骤包括：

将所述关键词与正文标题进行匹配；

如匹配，提取该结果网页的范文数据；其中所述范文数据包括：正文标题，正文内容，和正文内容的字数。

8. 根据权利要求 1-7 任一项所述的优化方法，其中被提取的结果网页是文库资源网站网页，其网页包括范文文档的 URL 资源链接和描述对应范文文档的文字内容；其中提取范文数据的步骤包括：

将所述关键词与描述对应范文文档的文字内容进行匹配；

如匹配，经由所述 URL 资源链接下载所述范文文档；

提取该网页的范文数据；其中所述范文数据包括：描述范文文档的文字内容，和所述范文文档。

9. 一种范文网页数据库的优化装置，包括：

请求获取单元，适于获取用户的检索请求；

请求判断单元，适于根据所述检索请求中携带的关键词，判断所述检索请求是否属于范文检索请求；

范文数据单元，适于检测用户在检索结果页中对结果网页的交互反馈，并根据所述交互反馈提取对应的结果网页中的范文数据；

数据添加单元，适于将所述关键词、所述结果网页中的范文数据及结果网页 URL 作为新增数据项添加到范文网页数据库中。

10. 根据权利要求 9 所述的优化装置，其中，数据添加单元还适于

判断结果网页 URL 是否包含在范文网页数据库的现有范文网页数据项中；

如果是，则不添加所述新增数据项；

如果否，则添加所述新增数据项。

一种范文网页数据库的优化方法和装置

技术领域

[0001] 本发明涉及互联网搜索领域，特别是涉及一种用于范文搜索的范文网页数据库的优化方法和装置。

背景技术

[0002] 范文搜索是网页搜索中一种很重要的需求，被搜索的范文类型众多，包括但不限于各类公文、文秘书信、工作计划、总结报告、心得体会、演讲致辞、作文作业、各种论文等等。在例如学生非放假期间、年底工作总结期间的高峰期，每天可以占到总网页搜索量的1%左右。实际中，大多数范文需求对于字数都有固定要求，因而很多用户都会在进行范文搜索时输入字数，如“读后感400字”、“奖学金申请书800字”等。即使在搜索时没有明确将字数输入，也会存在有对于该范文字数的潜在要求；例如，论文类的范文一般不低于8000字；入党申请书类的范文一般要求有3000～5000字，等等。

[0003] 对于范文搜索，目前存在的问题主要有两个：一是当前检索机制只能通过标题、网页内容匹配来命中字数的需求，对于没有相关字数的网页排序不公平；由于找不到相应的字数的字段，也使得召回率不足；二是在目前检索结果的标题摘要下，用户只能通过相应字段飘红来判定是否是自己想要的信息，对于很多有欺骗性质的页面、字数是否满足要求等都是没有预期的。

[0004] 图1示出了当前范文搜索的搜索结果页示意图，用户输入的范文搜索请求是“以家为题的作文350字”；在搜索结果页中，除了第一条结果的标题摘要直接命中350字外，其它结果都不知道有多少字数，只能将“350字”这个关键词丢弃来进行排序，对于一些潜在与350字非常接近的结果来说就显得非常不公平；用户也不知道什么结果是好结果，只能逐个点击查看，效率比较低下。

发明内容

[0005] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的用于范文搜索的范文网页数据库的优化方法和相应的装置。

[0006] 依据本发明的一个方面，提供了一种范文网页数据库的优化方法，包括：

[0007] 获取用户的检索请求；

[0008] 根据所述检索请求中携带的关键词，判断所述检索请求是否属于范文检索请求；

[0009] 如果是，检测用户在检索结果页中对结果网页的交互反馈，根据所述交互反馈提取对应的结果网页中的范文数据；

[0010] 将所述关键词、所述结果网页中的范文数据及结果网页URL作为新增数据项添加到范文网页数据库中。

[0011] 可选地，添加步骤包括：判断结果网页URL是否包含在范文网页数据库的现有范文网页数据项中；如果是，则不添加所述新增数据项；如果否，则添加所述新增数据项。

[0012] 可选地，新增数据项中的关键词对应于范文网页数据项的范文类型；其中添加步

骤包括：在范文网页数据库中，确定与新增数据项具有相同范文类型的现有范文网页数据项的项数；如所述项数小于预定数量，则添加所述新增数据项；如所述项数大于或等于预定数量，则不添加所述新增数据项。

[0013] 可选地，被提取的结果网页是问答社区网页，包括提出问题的主楼块和回答问题的次楼块；其中提取范文数据的步骤包括：将所述关键词与网页主楼块的文字内容进行匹配；如匹配，判断次楼块的文字内容的字数是否大于预定阈值；如果是，确定字数大于预定阈值的次楼块为待提取次楼块；且提取该结果网页的范文数据；其中所述范文数据包括：待提取次楼块的文字内容的标题，待提取次楼块的文字内容的正文，待提取次楼块的文字内容的字数。

[0014] 可选地，确定待提取次楼块的步骤还包括：根据所述关键词确定元关键词；将所述元关键词与字数大于预定阈值的次楼块的文字内容进行匹配；如匹配，确定匹配的次楼块为待提取次楼块。

[0015] 可选地，所述网页中待提取次楼块为多个，则所述网页对应的范文网页数据项包括与待提取次楼块数量相应的多个范文数据。

[0016] 可选地，被提取的结果网页是文字网站网页，其网页包括正文标题和正文内容；其中提取范文数据的步骤包括：将所述关键词与正文标题进行匹配；如匹配，提取该结果网页的范文数据；其中所述范文数据包括：正文标题，正文内容，和正文内容的字数。

[0017] 可选地，被提取的结果网页是文库资源网站网页，其网页包括范文文档的 URL 资源链接和描述对应范文文档的文字内容；其中提取范文数据的步骤包括：将所述关键词与描述对应范文文档的文字内容进行匹配；如匹配，经由所述 URL 资源链接下载所述范文文档；提取该网页的范文数据；其中所述范文数据包括：描述范文文档的文字内容，和所述范文文档。

[0018] 根据本发明的另一方面，提供了一种范文网页数据库的优化装置，包括：

[0019] 请求获取单元，适于获取用户的检索请求；

[0020] 请求判断单元，适于根据所述检索请求中携带的关键词，判断所述检索请求是否属于范文检索请求；

[0021] 范文数据单元，适于检测用户在检索结果页中对结果网页的交互反馈，并根据所述交互反馈提取对应的结果网页中的范文数据；

[0022] 数据添加单元，适于将所述关键词、所述结果网页中的范文数据及结果网页 URL 作为新增数据项添加到范文网页数据库中。

[0023] 可选地，数据添加单元还适于：判断结果网页 URL 是否包含在范文网页数据库的现有范文网页数据项中；如果是，则不添加所述新增数据项；如果否，则添加所述新增数据项。

[0024] 可选地，新增数据项中的关键词对应于范文网页数据项的范文类型；其中数据添加单元还适于：在范文网页数据库中，确定与新增数据项具有相同范文类型的现有范文网页数据项的项数；如所述项数小于预定数量，则添加所述新增数据项；如所述项数大于或等于预定数量，则不添加所述新增数据项。

[0025] 可选地，被提取的结果网页是问答社区网页，包括提出问题的主楼块和回答问题的次楼块；其中范文数据单元还包括：匹配单元，适于将所述关键词与网页主楼块的文字

内容进行匹配；次楼块确定单元，适于如匹配，判断次楼块的文字内容的字数是否大于预定阈值；如果是，确定字数大于预定阈值的次楼块为待提取次楼块；以及提取单元，适于提取该结果网页的范文数据；其中所述范文数据包括：待提取次楼块的文字内容的标题，待提取次楼块的文字内容的正文，待提取次楼块的文字内容的字数。

[0026] 可选地，次楼块确定单元还适于：根据所述关键词确定元关键词；将所述元关键词与字数大于预定阈值的次楼块的文字内容进行匹配；如匹配，确定匹配的次楼块为待提取次楼块。

[0027] 可选地，所述网页中待提取次楼块为多个，则所述网页对应的范文网页数据项包括与待提取次楼块数量相应的多个范文数据。

[0028] 可选地，被提取的结果网页是文字网站网页，其网页包括正文标题和正文内容；其中范文数据单元还包括：匹配单元，适于将所述关键词与正文标题进行匹配；提取单元，适于如匹配，提取该结果网页的范文数据；其中所述范文数据包括：正文标题，正文内容，和正文内容的字数。

[0029] 可选地，被提取的结果网页是文库资源网站网页，其网页包括范文文档的 URL 资源链接和描述对应范文文档的文字内容；其中范文数据单元还包括：匹配单元，适于将所述关键词与描述对应范文文档的文字内容进行匹配；下载单元，适于如匹配，经由所述 URL 资源链接下载所述范文文档；提取单元，适于提取该网页的范文数据；其中所述范文数据包括：描述范文文档的文字内容，和所述范文文档。

[0030] 本发明所述范文网页数据库建立在搜索引擎服务器侧，包括一定数量的范文网页数据项，所述范文网页数据项包括了范文类型、范文网页的范文数据和范文网页对应的 URL，所述范文数据典型地包括范文的标题、正文和字数。用户发出范文搜索请求时，在网页爬虫抓取的基础网页库进行常规搜索的同时，还在范文网页数据库中进行搜索。由于范文网页数据库中包含了各种范文网页的范文标题、正文以及字数，使得真正与用户所要求的范文字数相同、相近的范文网页出现在搜索结果页中，并且能在搜索结果排序时能排在前面，进一步还能在搜索结果页中将范文字数显示给用户，由此提升搜索质量和用户体验。

[0031] 根据本发明的范文网页数据库的优化方法和装置，对于在范文网页数据库中未检索到的、且能满足用户范文需求的网页，通过用户的点击不断反馈补充到范文网页库中，不断扩充范文网页数据库中范文网页数据项的数量以提升召回率，满足更多用户的需求。并且，由于新增数据项是根据用户对结果网页的点击反馈进行的，所以新增数据项与用户需求相关度非常高，从而能够提高下一步经由范文网页数据库检索的质量。

[0032] 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

[0033] 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：

[0034] 图 1 是现有技术的范文搜索的搜索结果页示意图；

[0035] 图 2 是本发明所述范文网页数据库的数据结构示意图；

[0036] 图 3 是根据本发明一个实施例的范文网页数据库的优化方法的流程图；

[0037] 图 4 是本发明所述范文网页数据库的一范文网页数据项的示意图；

[0038] 图 5 是根据本发明另一实施例的范文网页数据库的优化装置的框图。

具体实施方式

[0039] 下面将参照附图更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。

[0040] 本发明所述范文网页数据库，包括多个范文网页数据项，每一范文网页数据项对应一范文网页，具体包括该范文网页的范文类型、范文网页的范文数据和范文网页对应的 URL。典型地，范文网页的范文数据包括范文标题、范文正文和范文字数。图 2 示意性地示出了本发明所述范文网页数据库的数据结构示意图。

[0041] 范文网页数据库建立在搜索引擎的服务器端。首先，通过网络爬虫对互联网上范文资源网站的范文网页进行抓取；网络爬虫是一个技术成熟的、能自动提取互联网上网页的程序，它根据既定的规则为搜索引擎从互联网上下载网页，是搜索引擎的重要组成。所有被网络爬虫抓取的网页将会被存贮在服务器侧；同时可进行一定的分析、过滤，建立索引，生成供用户检索使用的基础检索库（或索引数据库）；网络爬虫可以在全网络范围内进行范文网页的抓取，也可以在指定的多个范文资源网站的范围内进行网页抓取，所述指定的范文资源网站可以由搜索服务提供商和 / 或用户不断进行添加和更新。然后，针对存储在服务器侧的、已被抓取到的范文网页，根据与所要提取的范文类型相对应的关键词，提取该类型范文网页的范文数据。具体地，首先需要将关键词与范文网页的内容进行匹配；如匹配，则从范文网页的内容中提取范文数据。最后，基于已提取的范文网页的范文数据，建立范文网页数据库。

[0042] 当用户在浏览器客户端发起检索请求时，搜索引擎的服务器获取用户的检索请求，例如“入党申请书 3000 字”，此请求表示用户希望搜索到字数在 3000 字左右的入党申请书范文。服务器接收到检索请求后，会对检索请求的内容进行分析，获取其中的检索项，对于上述检索请求，其检索项为“入党申请书”和“3000 字”；根据检索项“入党申请书”，判断其为范文检索请求。

[0043] 当判断用户的检索请求是范文检索请求时，进一步根据检索请求携带的关键词，在范文数据库中进行检索。检索请求携带的关键词，对应于前述检索项，包括主题关键词和字数关键词。例如，检索请求“入党申请书 3000 字”，其主题关键词为“入党申请书”，字数关键词为“3000”。在范文数据库的检索过程中，可先将主题关键词与范文类型进行匹配，然后根据二者匹配情况，进一步将主题关键词在相应的范文标题和 / 或范文正文中进行匹配。示例性地，对于检索请求“入党申请书 3000 字”，在范文类型和 / 或范文标题和 / 或范文正文中对主题关键词“入党申请书”进行匹配，若匹配成功，即判断主题关键词检索命中。在主题关键词检索命中之后，在命中的范文网页数据项中，进一步将字数关键词与范文网页数据项的范文字数进行匹配，并建立二者的匹配度。示例性地，对于检索请求“入党申请

书 3000 字”，若一范文网页数据项的范文字数为 3000 字，则匹配度为 100%；另一数据项的范文字数为 5000 字，则匹配度为 60%。优选地，若字数匹配度低于某一阈值，例如匹配度低于 50%，可在检索结果中舍弃此范文网页数据项对应的范文网页。

[0044] 在检索步骤之后，提取检索到的与所述关键词匹配的范文网页数据项的关键项信息，并在检索结果页中显示所述关键项信息。其中，被提取的关键项信息包括范文标题和范文字数，范文标题可以让用户初步了解范文的具体主题，范文字数则能够让用户直观了解该范文字数是否满足自己的需求。优选地，被提取的关键项信息还可包括部分范文正文，以使得用户能够在不点击打开范文网页的情况下就知晓范文的部分内容。在搜索结果页中，根据范文字数与字数关键词的匹配度由高到低对在范文网页数据库中检索到的范文网页进行排序。

[0045] 另一方面，在范文网页数据库中进行检索的同时，还根据请求内容的关键词，在基础网页库进行检索。如上所述，基础网页库，即搜索引擎的索引数据库，包括由网络爬虫下载存储到服务器侧的网页；在基础网页库的检索是互联网搜索领域的一项成熟技术，在此不再赘述。最后，在检索结果页中合并在范文网页数据库检索到的范文网页和在基础网页库检索到的范文网页。具体地，在检索结果页中，在范文网页数据库检索到的范文网页排在基础网页库检索到的范文网页的前面。

[0046] 综上所述，由于范文网页数据库中包含了各种范文网页的范文标题、正文以及字数，使得真正与用户所要求的范文字数相同、相近的范文网页出现在搜索结果页中，并且能在搜索结果排序时能排在前面，进一步还能在搜索结果页中将范文字数显示给用户，由此提升搜索质量和用户体验。但是，如果范文网页数据库中某种范文类型的范文网页数据项的数量不足，就很有可能不能满足用户的范文检索需求，导致范文网页数据库的上述优点不能发挥，最后在检索结果页中呈现给用户的还是从基础网页库中得出的检索结果。

[0047] 在本实施例中，提供了一种范文网页数据的优化方法，在搜索引擎的服务器侧执行。图 3 示出了所述优化方法的流程图，包括步骤 S302-S308，其中：

[0048] 步骤 S302：获取用户的检索请求；

[0049] 步骤 S304：根据所述检索请求中携带的关键词，判断所述检索请求是否属于范文检索请求；

[0050] 步骤 S306：如果是，检测用户在检索结果页中对结果网页的交互反馈，根据所述交互反馈提取对应的结果网页中的范文数据；

[0051] 步骤 S308：将所述关键词、所述结果网页中的范文数据及结果网页 URL 作为新增数据项添加到范文网页数据库中。

[0052] 本实施例方法从步骤 S302 开始，搜索引擎的服务器获取用户的检索请求，例如“入党申请书 3000 字”，此请求表示用户希望搜索到字数在 3000 字左右的入党申请书范文等等。

[0053] 服务器接收到检索请求后，会对检索请求的内容进行分析，获取其中的检索项，即此时进入步骤 S304。对于检索请求“入党申请书 3000 字”，其检索项为“入党申请书”和“3000 字”。根据检索项“入党申请书”，步骤 S304 判断其为范文检索请求。关于对检索请求进行分析获取检索项，有多种成熟的技术能够实现，在此不再赘述。

[0054] 当判断用户的检索请求是范文检索请求时，进一步根据检索请求携带的关键词

(对应于前述检索项,例如主题关键词“入党申请书”和字数关键词“3000”),同时在范文数据库和基础网页库中进行检索,并在检索结果页中合并在范文网页数据库检索到的范文网页和在基础网页库检索到的范文网页,此时进入到步骤S306,检测用户在检索结果页中对结果网页的交互反馈,根据所述交互反馈提取对应的结果网页中的范文数据。具体地,在步骤S306中,服务器检测用户在浏览器侧对搜索结果页中提供的结果网页(包括在范文网页数据库检索到的范文网页,和在基础网页库检索到的范文网页)的选择,即检测来自浏览器侧的用户对某结果网页的URL链接的点击动作的反馈,服务器判断该结果网页是与用户范文需求较为相关的一范文网页,并根据此反馈对该结果网页进行解析、提取其中的范文数据。

[0055] 可选地,所述结果网页是问答社区网站的范文网页;问答社区网站以一个主楼块提出问题,多个次楼块回答问题的形式,直接、快速地满足了用户搜索答案的需求,几乎能解决日常生活中的所有问题,由此也形成了一个巨大的内容资源。目前,国内已有很多较有影响的问答社区网站,例如百度知道,360问答,搜搜问问,天涯问答等等。下面以检索请求“入党申请书3000字”为例详细描述步骤S306如何提取问答社区网站的范文网页的范文数据。

[0056] 首先,判断主楼块的文字内容与主题关键词“入党申请书”是否匹配;其中,问答社区网页的主楼块和各个次楼块的文字内容,已经由网络爬虫提取出来。当主楼块的文字内容与主题关键词“入党申请书”匹配时,判断次楼块的文字内容的字数是否大于预定阈值。根据一篇入党申请书通常所要求的最低字数,设定该预定阈值,例如2000字,低于预定阈值2000字的次楼块将被舍弃。因为在问答社区网页中,很多次楼块的内容极有可能不是对主楼块所提问题的答案,例如次楼块内容为“不知道”、“我也很想知道”等等;而字数大于2000字的次楼块,就极有可能是一篇真正的入党申请书的范文。

[0057] 当然,对于不同的范文类型,应该对范文字数设置不同的预定阈值。例如,对于类型为“请假条”的范文,其预定阈值可以设置得相对较低,例如10字。

[0058] 优选地,如次楼块的文字内容大于了预定阈值2000字,可基于关键词进一步对次楼块进行筛选,判断字数大于预定阈值的次楼块的内容与元关键词是否匹配。这里的元关键词,或者为关键词本身,或者从关键词中提取得来。对于本示例中的关键词“入党申请书”,确定其元关键词为“申请”和“入党”。在问答社区网页中,次楼块通常可由任何网络用户进行添加,因此字数大于预定阈值的次楼块的内容也有可能与主楼块所提问题无关,例如次楼块为网络用户恶意粘贴的广告等。通过将次楼块内容与元关键词进行匹配,可以进一步确定次楼块的内容与入党申请书的相关性。另一方面,次楼块中一篇关于入党申请书的范文也有可能未出现“入党申请书”的完整字段,而元关键词“申请”和“入党”保证了不会将其遗漏。若次楼块与元关键词匹配,确定该次楼块为待提取次楼块,即确定该次楼块的内容包含主题为“入党申请书”的范文。

[0059] 最后,从所述待提取次楼块的“入党申请书”范文中提取范文数据,包括范文标题,范文正文,以及范文字数。从次楼块的文字内容中提取上述范文数据的具体实现,并非本发明的发明点所在,其实现细节在此不再赘述。

[0060] 发明人注意到,对于问答社区网站的一个范文网页,其多个次楼块都有可能被确定为待提取次楼块,即多个次楼块都能满足字数要求和元关键词匹配,故该范文网页对应

的范文网页数据项将包括多条范文数据，如图 4 所示，其中每条范文数据与每个待提取次楼块的内容对应。

[0061] 通过本实施例，准确提取了问答社区网站中所包含的范文数据，最大限度地去除了无效内容或恶意广告内容。

[0062] 可选地，所述结果网页是文字网站的范文网页；文字网站的网页以文字为主，在网页的主要区域内以诸如文稿的形式介绍该网页的主要内容，例如新闻网站、博客网站等等。典型地，文字网站的网页都包括正文标题和正文内容，通过网络爬虫能够获取这些信息。下面仍将以主题关键词“入党申请书”为例来介绍针对文字网站的范文网页的范文数据提取方法。

[0063] 根据本实施例的范文数据提取方法，对于文字网站，首先将关键词“入党申请书”与正文标题进行匹配；如匹配，确定该网页为类型为“入党申请书”的范文网页，则进一步在正文内容中提取正文标题，正文内容，和正文内容的字数，将其作为该网页的范文数据。

[0064] 可选地，所述结果网页是文库资源网站的范文网页，文库资源网站能够为用户提供各种文章、论文的下载服务，例如万方数据网等等。典型地，文库资源网站的网页包括范文文档的 URL 资源链接和描述该范文文档的文字内容。

[0065] 根据本实施例的范文数据提取方法，对于文库资源网站，首先将关键词“入党申请书”与文库资源网页中描述范文文档的文字内容进行匹配；如匹配，确定该文库资源网页为类型为“入党申请书”的范文网页，则经由所述 URL 资源链接下载所述范文文档；进一步地，提取描述范文文档的文字内容和所下载的范文文档作为该网页的范文数据。

[0066] 在步骤 S306 提取好范文数据后，进入步骤 S308，将所述关键词、所述结果网页中的范文数据及结果网页 URL 作为新增的范文网页数据项添加到范文网页数据库中。其中，所述关键词对应于范文网页数据项的范文类型。

[0067] 可选地，在向范文网页数据库添加新增数据项之前，判断结果网页 URL 是否包含在范文网页数据库的现有范文网页数据项中；如果否，则添加所述新增数据项；如果是，则不添加所述新增数据项。由此，避免了在范文网页数据库中重复添加相同范文网页的数据项。

[0068] 可选地，在向范文网页数据库添加新增数据项之前，先在范文网页数据库中确定与新增数据项具有相同范文类型的现有范文网页数据项的项数；如所述项数小于预定数量，则添加所述新增数据项；如所述项数大于或等于预定数量，则不添加所述新增数据项。例如，对于范文类型“入党申请书”，所述预定数量设置为 10 万；若范文网页数据库中范文类型为“入党申请书”的已有范文网页数据项为 8 万条，说明该类型的范文网页数量不足，则继续添加所述新增数据项。

[0069] 本发明另一实施例提供了一种范文网页数据库的优化装置，设置在搜索引擎的服务器侧；图 5 示出了根据本实施例的范文网页数据库的优化装置的框图，包括单元 502-508。

[0070] 当用户向搜索引擎发出检索请求时，请求获取单元 502 获取用户的检索请求；然后请求判断单元 504 根据请求内容的关键词，判断所述检索请求是否是范文检索请求。

[0071] 当判断用户的检索请求是范文检索请求时，进一步根据检索请求携带的关键词，同时在范文数据库和基础网页库中进行检索，并在检索结果页中合并在范文网页数据库检

索到的范文网页和在基础网页库检索到的范文网页。此时，范文数据单元 506 检测用户在检索结果页中对结果网页的交互反馈，并根据所述交互反馈提取对应的结果网页中的范文数据。具体地，服务器检测用户在浏览器侧对搜索结果页中提供的结果网页的选择，即检测来自浏览器侧的用户对某结果网页的 URL 链接的点击动作的反馈，服务器判断该结果网页是与用户范文需求较为相关的一范文网页，并根据此反馈对该结果网页进行解析、提取其中的范文数据。

[0072] 可选地，范文数据单元 506 包括：匹配单元，适于将关键词与问答社区网页主楼块的文字内容进行匹配；次楼块确定单元，适于在关键词与主楼块匹配时，判断次楼块的文字内容的字数是否大于预定阈值，并且如果次楼块文字大于预定阈值，则确定字数大于预定阈值的次楼块为待提取次楼块；以及提取单元，适于提取该网页的范文数据；其中所述范文数据包括待提取次楼块的文字内容的标题，待提取次楼块的文字内容的正文，待提取次楼块的文字内容的字数。优选地，在确定次楼块的内容字数大于预定阈值后，次楼块确定单元进一步将字数大于预定阈值的次楼块的内容与元关键词（关键词本身，或根据所述关键词确定）进行匹配；如匹配，确定匹配的次楼块为待提取次楼块。

[0073] 可选地，范文数据单元 506 包括匹配单元，适于将所述关键词与文字网页的正文标题进行匹配；以及提取单元，适于在关键词与正文标题匹配时，提取该网页的范文数据；其中所述范文数据包括：正文标题，正文内容，和正文内容的字数。

[0074] 可选地，范文数据单元 506 包括匹配单元，适于将所述关键词与文库资源网页中描述对应范文文档的文字内容进行匹配；下载单元，适于在关键词与描述文字匹配时，经由所述 URL 资源链接下载范文文档；和提取单元，适于提取该网页的范文数据；其中所述范文数据包括：描述范文文档的文字内容，和所述范文文档。

[0075] 在范文数据提取完成之后，数据添加单元 508 将所述关键词、所述结果网页中的范文数据及结果网页 URL 作为新增范文网页数据项添加到范文网页数据库中，其中所述关键词对应于范文网页数据项的范文类型。

[0076] 可选地，数据添加单元在向范文网页数据库添加新增数据项之前，判断结果网页 URL 是否包含在范文网页数据库的现有范文网页数据项中；如果是，则不添加所述新增数据项；如果否，则添加所述新增数据项。

[0077] 可选地，数据添加单元在向范文网页数据库添加新增数据项之前，先在范文网页数据库中确定与新增数据项具有相同范文类型的现有范文网页数据项的项数；如所述项数小于预定数量，则添加所述新增数据项；如所述项数大于或等于预定数量，则不添加所述新增数据项。

[0078] 根据本实施例所述的范文数据库优化方法和装置，提供了一种及时补充范文网页数据库的机制，尤其是在范文网页数据库中无结果网页或结果网页很少的情况下，通过用户对基础检索库中得到的结果网页的互动反馈，及时将该结果网页的范文数据补充到范文数据库中，从而不断扩充数据库内容以满足更多用户的需求，提升召回率。

[0079] 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。

[0080] 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。

[0081] 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。

[0082] 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。

[0083] 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。

[0084] 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范文网页数据优化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。

[0085] 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。

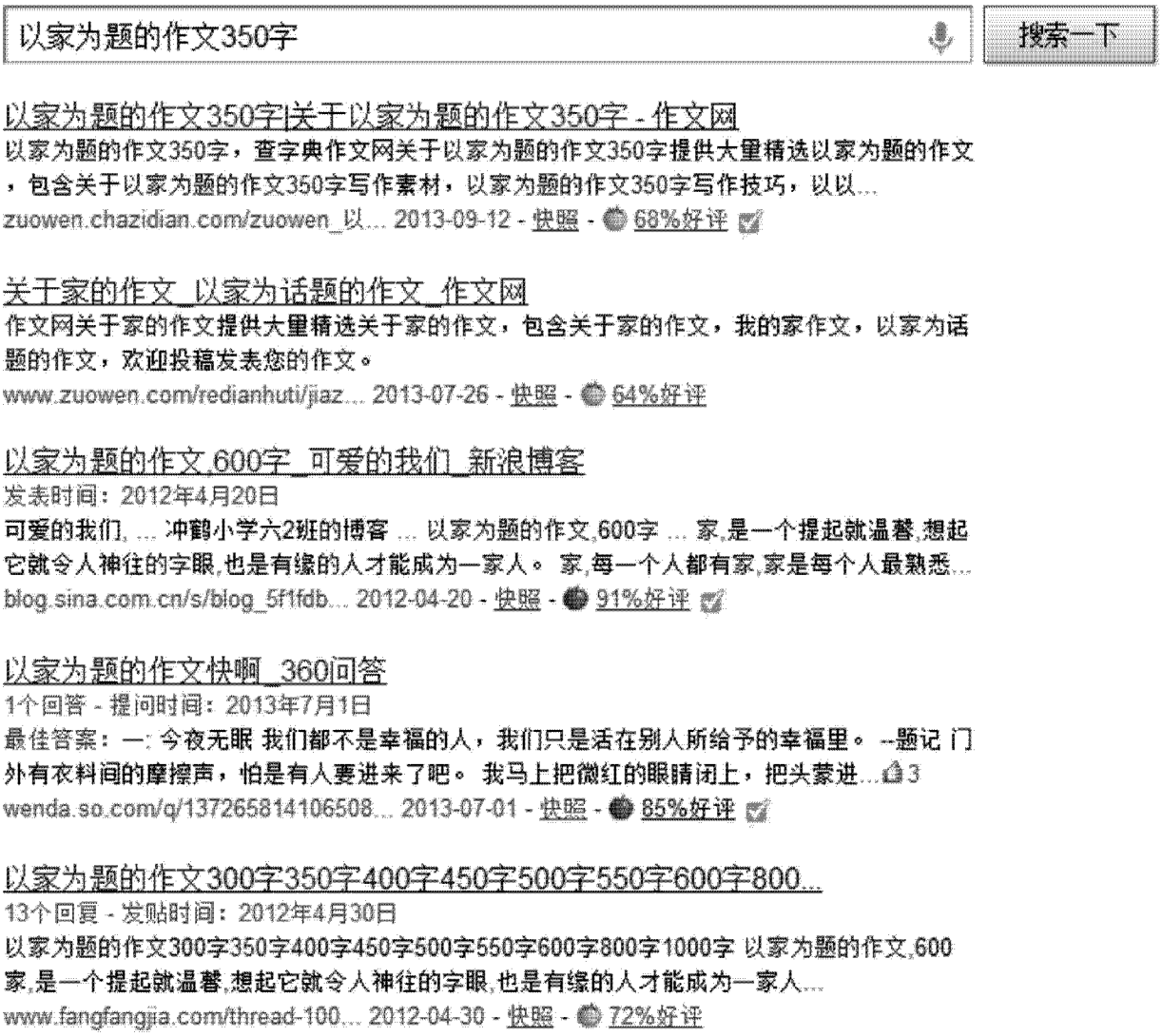


图 1

文类型	范文数据			范文网页 URL
	范文标题	范文正文	范文字数	
奖学金申请书	××	××××××	500	×××
工作总结	××	××××××	1000	×××
…	…	…	…	…
…	…	…	…	…

范文网页数据项 1

范文网页数据项 2

图 2

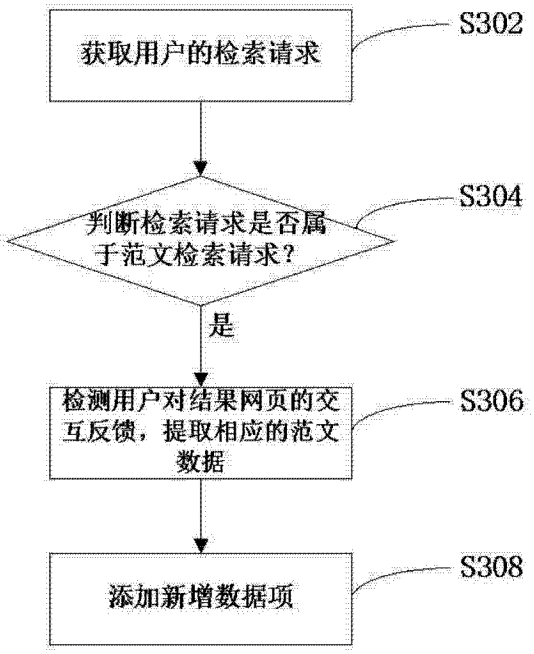


图 3

入党申请书	AA	AAAA	2000	URL	范文网页数据项
	BB	BBBB	3000		
	CC	CCCC	4000		

图 4

图 5