吉林大学本科毕业设计（论文）外文文献翻译

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 学 院 | 软件学院 | | | | 专业 | 软件工程 |
| 学生姓名 | 宋治佳 | | | | 学号 | 55130511 |
| 指导教师 | 韩冬冰 | | | | 职称 | 讲师 |
| 合作导师 |  | | 职称 |  | 单位 |  |
| 设计（论文）题目 | | 基于Node的北京租房信息推荐系统的设计与实现 | | | | |
| 外文文献翻译（翻译字数不少于3000字）  **摘要**  聚焦爬虫工具下载与用户指定主题相关的网页。大多数现有的聚焦爬虫工具都是关键字驱动的，并且不考虑与关键字相关联的上下文。这导致检索了大量不知道逻辑是否相关的网页。因此，单是基于关键字的策略对于聚焦爬虫的设计是不够的，因为就用户的需求而言，上下文相关性更重要。本文提出了一种基于上下文驱动的聚焦爬虫（CDFC）的设计，该爬虫仅搜索和下载高度相关的网页，从而减少网络流量。它还采用类别树：一个灵活的用户界面，显示着网络上主题的大类。由于CDFC仅下载相关且可信的文档，与以往的方式相比，该设计显着减少了搜索引擎侧所需的存储空间。  *索引：搜索引擎，爬网程序，超文本文档系统，类别树，软件代理*   1. **介绍**   万维网（WWW）是一个不断扩大的超文本文件的大集合[1]。它代表一个非常大的分布式超文本系统，涉及数十万个单独的站点。它是基于客户端-服务器的架构，其允许用户通过向搜索引擎提供关键字来发起搜索，搜索引擎又从互联网收集并返回所需的网页。由于Web上存在极大量的页面，搜索引擎往往依赖于搜寻器来收集所需的页面。爬虫[2]遵循文档中存在的超链接来下载和存储搜索引擎的网页。  当前的商业搜索引擎维护着大量的网页[3,7]，并且容易找到几千个匹配项用于平均查询。因此，搜索引擎可能相应与用户关键词不相关的数千个网页的列表。 网络搜索引擎尝试覆盖整个网络，并提供关于所有可能的主题的查询[4]。事实上，从用户的角度来看，搜索返回10,000和50,000个匹配没有任何区别，反而可能因为匹配的数量变得太大而不能筛选，导致信息过度使用的问题。  网页的搜索质量可以通过聚焦爬行来改进[5,6,12]，其目的在于仅搜索和检索与特定主题相关的WWW的子集。因此，聚焦的爬行器为信息过度使用的问题提供了潜在的解决方案。现有的焦点爬虫[6,7]采用不同的策略来计算词语在web文档中的频率。如果高频率词与主题关键词匹配，则认为文档是相关的。但是当前抓取工具无法在下载之前分析网页中关键字的上下文。例如，“spider”一词有各种解释。对于web程序员，它是在搜索引擎中使用的软件程序的名称;到一般的计算机用户它表示卡的游戏，但是对外行人，它是昆虫的简单名称。因此，主题相关性不是焦点爬虫的唯一问题，但上下文相关性也应该考虑[10]。如果用户发出一个关键字，则其相关上下文也必须是已知的。  在本文中，提出了一种上下文驱动聚焦爬虫（CDFC）的设计，在灵活的交互式类别树[5]中为用户提供关键字的上下文。基于代理的设计不仅克服了现有聚焦爬虫的复杂耗时计算，而且显着降低了网络流量。根据聚焦爬虫的需求，相关信息应该在最少的时间内由用户收集和检索，所提出的架构减少了文档的搜索时间，并且搜索引擎侧的信息数据库变得更易于管理。   1. **相关工作**   基于相似性的爬虫，其对具有锚文本或URL中的目标关键词的URL进行排序，可能是对聚焦爬行的第一个影响[9]。基本焦点是首先抓取更重要的页面，即查看页面的重要性的各种度量，例如与驾驶查询的相似性，指向该页面的页面的数量（后退链接），页面排名，位置等。页面等级算法[11]通过页面包含的链接的质量成比例地对每个页面的每个链接进行加权来计算页面的分数。因此，如果网页从许多其他页面链接，则网页将具有高页面排名，并且如果这些参考页面也是好的页面，即具有高页面排名得分，则分数将更高。在HITS算法[8]中，权限页面被定义为与特定主题或搜索查询相关的高质量页面，并且中心页面是提供指向其他权限页面的指针的页面。基于这些，网页与被计算为识别网页上下文的权威评分和集线器得分相关联。  另一个聚焦爬虫[7]采用种子关键字，用于从一些标准搜索引擎（如Google）中查找种子URL。种子URL用于在TF.IDF算法的帮助下基于迭代地计算字频率来获取种子页面。该算法用于从种子网页中找出更多数量的关键词以表示主题。之后，计算网页和主题关键词之间的向量相似性，以查看该页面是否与主题相关。  Diligenti等人[6]使用一个通用的搜索引擎来获取网页链接到一个特定的文档，并建立一个页面的上下文图。然后，该图用于训练一组分类器以基于它们到目标的预期链路距离将文档分配给不同类别。实际上，为每个种子文档构建图形和分类器，其中层被构建到指定级别。因此，爬虫获得与目标主题直接或间接相关的主题的知识。  仔细观察可用的聚焦爬虫[5-9,11]可以看出这些爬虫具有以下缺点：   1. .对于每个web文档的字频率的迭代计算的问题使得搜索过程昂贵。 2. .网页的相关性在下载之前是未知的。 3. .在搜索启动之前，网页的相关上下文是未知的。 4. .具有关键字搜索的搜索引擎的用户界面不灵活。   这篇论文提出的方法有效地解决了上述问题。一个使用增强的超文本文档结构与用于在搜索引擎侧提供用户界面的类别树相结合的基于上下文驱动的聚焦爬虫已经被设计出来。  **2.1增强的超文本文件**  WWW上的超文本文件信息系统是很庞大的，以分布式和非线性文本系统的形式组织的。通过HTTP和HTML提供检索和呈现超链接文档的标准方式。XML通过允许网页创建者使用他们自己的一组标记标签提供了更多的灵活性。此功能可用于在超文本文件中进行增强，以适合网络爬行[14]。在PARCAHYD项目[13]和[16,17]中设计的爬虫旨在使用增强的超文本文件的概念来提高爬虫的性能和质量问题。例如，为了管理易失性信息，通过易失性标签[15]来标记文档的可变信息，该易失性标签又从文档中提取出来它们相关的易失性信息。然后将标签及其内容存储在具有与文档相同名称但不同扩展名（.TVI）的文件中。支持.TVI和其他相关增强的超文本文件[14-17]被称为增强超文本文件。  **2.2类别树**  类别树[5]用于搜索引擎中的图形用户界面。它是具有示例关键字的预定义规范主题分类法。要运行特定实例，必须以两种形式提供初始输入。用户必须选择和/或细化分类法中的特定主题节点，并且还可能需要提供附加的示例关键字。然后，用户在相应的主题或类别节点中选择他感兴趣的示例关键字。随后，这些选择被提交给搜索引擎。  **3.上下文聚焦爬行器（CDFC）的设计**  对于本文，已经将所需信息的上下文扩展到超文本文件。其中，在作者创建超文本文件时，明确标记称为“关键字”和“上下文”的标签名称。  在保存文档时，所有关键字标签以及上下文和关键字标签被提取出来并且分别存储在具有相同名称但具有不同扩展名（例如.TOC）的文件中。  类别树也已经适当地作出修改用于所提出的设计，使得上下文还用类别示例来显示。用户选择一个类别节点（比如Internet），然后显示其相关示例。当用户选择一个例子（例如Crawler）时，显示两个相关联的上下文，最后，用户选择上下文（例如一般信息）。实际上，修改的类别树是在图形界面中预先指定的各种类别的集合，其示出了在这些类别下的各种示例及其上下文。用户可以通过在顺序中选择类别上下文来选择任何相关联的上下文。然而如果需要，用户可以插入新的示例，这可以稍后由爬行器链接到上下文。  为了爬取，CDFC使用表1中列出的三个代理，即用户代理，匹配代理和Dbase代理。下面给出关于这些代理及其相关组件的简要讨论：   1. 接受用户从类别树中选择类别节点和相关关键字。它将此信息发送到匹配器代理以从数据库检索相关联的上下文及其链接。 2. 在从匹配器代理获取关联上下文及其链接时，用户代理在类别树中向用户显示上下文列表。 3. 用户代理接受用户选择的上下文并显示其所有相应的链接。 4. 在用户选择链接时，用户代理将此链接发送到Retrieve\_Doc\_Process以从数据库检索文档。如果文档不存在于数据库中，它会将链接传递给搜寻器以下载文档。   **4.性能优势**  所提出的爬行器以类别树的形式呈现灵活且交互式的用户界面，使得指导用户选择合适的关键词及其用于web搜索的上下文。 CDFC仅下载高度相关的文档，其数量非常少，从而减少了用户面临的信息过度问题的问题。此外，减少了网络流量，因为不相关的网页不会被下载。  **5.综述**  所提出的基于上下文驱动的聚焦爬行器（CDFC）的设计是基于扩充的超文本文件，其中关键字的上下文以TOC（上下文表）的形式存储。与类别树耦合的TOC提供关键字的上下文。这种设计不仅避免了用于导出用户关键字的上下文的昂贵的复杂计算，而且显着地减少了网络流量。此外，下载的文档的质量符合用户选择的主题和上下文。 | | | | | | |

附件（双击打开）：