文献检索方法的研究与改进①

陈利东

(宁波职业技术学院 图书馆, 宁波 315800)

摘 要: 通过对当今常用检索方法的研究与分析, 提出一种相对新颖的检索模式与方法, 并对这种新的检索方法进行理论分析与实践验证. 新的检索模式还将把最常用的几项文案处理功能集合在同一平台下, 为用户提供更为便捷、准确、高效的信息服务.

关键词: 数据库; 检索; 文献; 预览

Research and Improvement on Document Retrieval Method

CHEN Li-Dong

(Ningbo Polytechnic Library, Ningbo 315800, China)

Abstract: This paper presents a new retrieval model and method by research and analyzing the commonly used retrieval methods. And it proposes a theoretical analysis and practical verification of the method. The new mode also integrates the most commonly used document processing functions in a platform for providing more convenient, accurate and efficient information services.

Keywords: databases; retrieval; document; preview

1 引言

随着计算机网络的飞速发展, 网络检索功能也在不断地更新, 从简单功能到功能全而完善, 从专业性到普及性, 从固定模式到自由模式. 在多式多样的检索方法面前, 本文将深入研究现今网络检索技术的发展与应用, 探索其检索方法的先进性, 并对目前常用的检索方法进行对比分析, 查找其中差距. 整合现有的网络检索技术与先进的软件查询功能, 提出一个较为新颖的检索理念, 并对这个理念进行初步的实现.

2 常用检索方法的研究与分析

目前, 常用的信息检索模式概括性的可分为: 主题语言检索方式和分类语言检索方式两种[1], 由此衍生出来的具体检索方法就有多种多样, 如: 分类导航检索、统一快速检索、高级组合类检索等等[2-4].

用户在信息检索时, 会根据自身的需求而采用不同的检索方法, 来实现自己的信息需求目标. 这里将用户的需求大致分为以下两种:

① 用户已知需要检索文献信息的具体内容, 如: 文献信息的题目、作者、发表的刊物等. 那么, 在利用检索方法时, 只需在相关的数据库内, 进行精确的检索, 就可以得到需要的文献.

② 用户已知需要查找文献的范围和方向, 但没有指定具体哪篇文献信息. 如: 为了科研项目而查找相关信息; 为了了解某领域的最新进展情况; 为了组织某类专题活动等等.

对于第一种情况本文不作讨论分析, 现针对第二种情况作为研究对象, 并得到相关调查结果为: 用户会在一切可查询信息资源内, 下载浏览大量的与之相关的文献信息. 在检索过程中, 因用户得到最初信息的不完整性, 如: 检索文献数据库所得到的最初结果一般只包含了文献信息的题名、刊名、作者、摘要、关键词等等, 只有信息的大概内容, 如果需要文献的详细内容就必须浏览全文; 各类互联网信息检索的最初结果一般只包含信息名称和与检索词相关的非常简要的网页信息内容、网页地址等, 用户如需详细了解

① 收稿时间:2013-10-05;收到修改稿时间:2013-11-20

信息的详细内容, 须进入该网页. 当今互联网利用率最高的搜索引擎 Google 和百度, 在检索结果内, 为部分信息提供了一个预览功能: 百度利用网页快照功能[5], 而 Google 与百度不同的是直接在当前页面的旁边提供给用户一个预览页面. 但是, 用户如果需要网页的详细内容还是要进入该网站内浏览.

这就存在一个现象: 用户在查找浏览文献信息过程中, 会将大量的宝贵时间浪费在查阅不能为用户目的服务的信息资源上, 因此也就浪费了大量的财力、人力和物力等资源.

3 本文研究的目的与平台开发基础

基于上述研究对象的特性, 确定本文研究的目的: 研发一种更为先进的检索模式与方法, 能为用户节约大量的宝贵时间, 减少大量的信息垃圾, 节约大量的网络资源. 能为用户提供更为便捷、准确、高效的信息服务, 节省大量的人力、财力和物力. 也将使其成为一个全新的文献检索模式.

为了实现新检索模式, 本文将研发一个新的软件平台, 并作为一个控制类插件, 应用到数据库的检索系统中. 此平台研发的基础如下:

① 实验数据库建立. 在研发此平台之前, 首先需建立一个实验性数据库, 此数据库需包含: 结构化数据(包括关系数据库或面向对象数据库中的数据)、非结构化数据(没有固定结构的数据)、半结构化数据(具有隐含结构, 但缺乏固定或严格结构的数据)[6]. 以便于平台验证的需求.

② 数据库内容应尽量包含目前常用的各类数据. 如: 图书、论文、报纸等文本数据; 图片、网页、视频等超文本数据.

③ 新检索模式平台的兼容性. 实验数据库是自建的, 基本不存在兼容性问题, 但是考虑到将来的推广, 还应在不同的数据库检索平台进行测试, 完善其兼容性, 使其有长远发展的可行性.

4 检索方法的改进与创新

本文研究内容就是结合各种先进的检索技术,如: 全文检索技术[7-8]、关键词定位[9]、网络抓取功能[10]等, 整合用户文献操作的一些常用功能, 如: 复制、查询下一个、下载文献等, 并加以改进, 形成一个统一的检索模块, 为用户提供简便、快捷的文献查看方式, 以节省用户的各项资源. 这一模块的研究与应用, 将为传统检索模式带来一个全新的文献查看方式, 使得文献检索功能更完善、更简便、更高效.

现今的数据库基本上都支持全文检索功能, 本方案的实现也将建立在全文检索的基础上. 用户的关键词在全库中进行检索, 并在检索到的每个文献后面添加一个预览功能, 用户可点击预览项得到一个显示框. 框内信息是文献全文内容中关键词所在的部分信息, 并提供以下可操作功能: 1、可调节显示文献内容文字的多少, 系统将设定一个高低限度(默认 100 字); 2、可查看此文献内关键词所在的下一个章节内容; 3、可复制显示的文字内容; 4、可直接下载此文献.

实现 1 功能: 对此文献进行垂直深入检索, 并定位出关键词所在位置(如同 word 里的查找功能), 同时抓取关键词所在位置前后共 100 个文字的内容, 并显示在预览框内(这是默认形式), 用户可根据需要对显示文献内容的多少进行调节, 可在调节框内输入显示文字数量进行重新抓取, 对于文字数量系统将设定上下限.

实现 2 功能: 查看关键词在文献中下一个位置,具体显示方法同功能 1, 如果全部检索完毕将给用户一个提示.

实现 3 功能: 可对预览框内的文字进行复制操作, 与普通复制功能相同.

实现 4 功能: 用户可直接点击下载, 将此文献保存在用户指定位置.

整个预览功能模块将直接嵌入到数据库检索平台内, 在检索结果中, 对每个文献提供预览功能按钮, 以方便用户使用. 预览的内容是: 根据用户检索的关键词在全文中检索其所在的位置, 并抽取关键词所在的前后一定量的文字内容, 提供给用户预览, 用户可以通过预览相应的文字内容来确定这篇文献是否是自己需要的, 如果是就下载使用该文献, 若不是就继续浏览该文献剩余关键词所在的文献内容, 或者退出浏览其它文献.

预览功能的开发与实现可以为用户节约大量的宝贵时间, 并为用户节省大量的人力、财力和物力, 也将是一个全新的文献检索方式.

5 新检索方式的实现与应用[11-12]

基于上述理论分析与检索平台关键模块的构建,

现初建一个数据库与检索平台，来验证新检索方式的可行性与检索结果的正确性.

首先构建一个功能窗口，来实现新检索方式的主要功能模块，包括字数控制、下一个位置、复制、下载. 如图 1:

图 1　功能窗口

第二步，建立一个数据库检索平台，如图 2:

图 2　检索平台

第三步，将新检索方式的功能窗口植入到检索平台中，并在每条文献后面插入“预览”按钮. 如图 3 检索关键词“中国动漫”，得到结果为:

图 3　检索结果

第四步，查看预览结果，并与原文对照，验证其可行性与正确性. 以第一条检索结果“解读中国动漫的发展误区”为例，如图 4 和图 5:

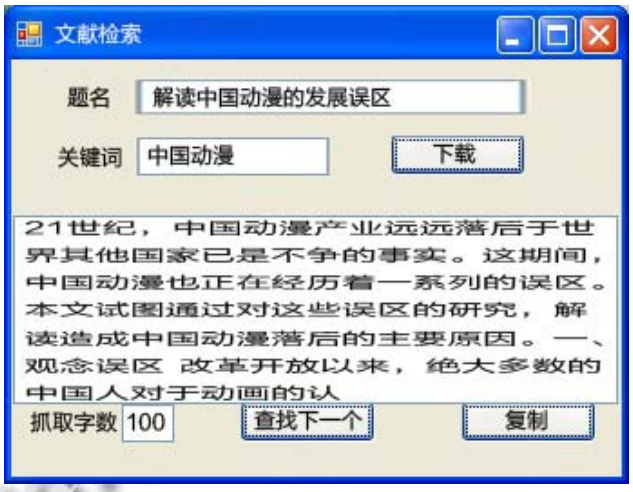

图 4　预览结果

图 5　预览结果

从以上两者的对比中可以得出，此次预览抓取 100 个字是成功的，也验证了文献检索预览功能是成功的.

6　结束语

本文对各种检索技术与文本处理功能进行改进和整合，在此基础上研发并实现了新的检索模式. 目前，该平台已完成初步测试，并取得了较好的效果. 因条件所限，该系统与不同平台兼容性测试还有待于继续完善，预览平台可操作功能过于单一也有待于丰富，这也是该项目今后发展的方向. 本文希望能为读者起到抛砖引玉作用，在未来看到相关的研究进一步深入和发展.

参考文献

1 孙悦民.网络信息系统三种核心检索方式的分析.高效图书馆工作,2009,29(5):54–55,62.

2 宋乐平.中文数据库分类检索能力研究.图书馆学研究,2010,(2):63–66.

3 俞肇元,袁林旺,罗文,胡勇,闾国年.?边界约束的非相交球树实体对象多维统一索引.软件学报, 2012,23(10):2746–2759.

4 李洪梅.数字国书馆异构赏源统一楂索研究.图书馆学刊,2013,(2):49–53.

5 张梁平.搜索引擎,”百度”推荐—谈搜索引擎.现代情报,2004,(2):208–209,212.

6 杜小勇,李曼,王珊.本体学习研究综述.软件学报,2006,17(9):1837–1847.

7 李梅,王庆林.中文全文检索技术的研究及实现.情报学报,2003,22(1):10–17.

8 郑庆华,张炜.超文本全文检索技术的研究与实现.西安交通大学学报,2001,35(4):376–381.

9 王军.词表的自动丰富—从元数据中提取关键词及其定位.中文信息学报,2004,19(6):36–43.

10 徐健;张智雄.基于 Nutch 的 Web 网站定向采集系统.现代图书情报技术,2009,25(4):1–6.

11 甘利人,冯颖,白晨.榜样信息干预下用户检索方法决策的观察学习研究.情报学报,2012,31(7):770–784.

12 郑文晖.Web2.0 在高校图书馆阅读服务中的应用研究.情报资料工作,2012,(3):95–98.