E-textbook 全文检索①

陈 辉, 戚佳慧, 吴 敏

(中国科学技术大学 现代教育技术中心, 合肥 230026)

摘 要: 基于电子教材的特殊应用的需求, 在传统的 web 页面全文检索技术基础上, 设计了电子教材的全文检索系统. 它包含教材文档处理模块、索引服务模块和检索服务模块. 根据电子教材的结构需求, 定义了索引文件数据结构、文本文件数据结构、索引条目数据结构及结果排序的分数模型. 通过系统的实现, 为电子教材学习者提供了快速准确的检索服务, 提高学习者学习效率.

关键词: 全文检索; web 信息抽取; 电子教材

E-textbook Full-text Retrieval

CHEN Hui, QI Jia-Hui, WU Min

(Center of Modern Educational Technology, University of Science and technology of China, Hefei 230026, China)

Abstract: Based on web pages full-text search technology. Design a full-text retrieval system to meet the special demand of e-textbook. It contains teaching materials document processing module, the indexing service module and the retrieval service module. According to the special feature of e-textbooks, we define the data structure of the index file, data structure of the text file, data structures of index entries and the score model for ranking. The retrieval system provides a fast and accurate search results to the learner.

Key words: full-text retrieval; web information extraction; e-textbook

电子书随着计算机软硬件技术的发展正迅速发展, 市面上已经出现多种不同形式的电子书应用. 电子书最简单的形式就是将纸质文本进行数字化, 使其能在终端设备上查看, 更为复杂的电子书形式是集合了文本、图片、音频和视频等丰富的用户体验, 供用户娱乐、学习使用. 电子教材属于电子书的一种特殊形式, 它的目的是用于课程教学[1]. 电子教材(e-textbook)是满足一定课程标准一种特殊形式的电子书[2], 与传统纸质教材不同, 电子教材充分利用计算机技术的优势来提高教材的可用性, 从而提高学习者的学习效率[3].

全文检索是电子教材的主要附加功能之一[4], 全文检索减短了学习者获取特定知识的时间, 提高了学习者的学习效率, 文章介绍学习型电子书的全文检索系统的设计和实现. 电子教材与普通文本不同, 内容包含文本、图片、视频等多种媒体表现形式, 同时为了提高教材的教学性, 电子教材包含更多的用户交互, 所以电子教材的格式也不是单纯的单文档形式, 而是依据特殊格式制作的多文档的内容集成. 全文检索技术在因特网、电子图书馆及各种资料库中有着广泛的应用, 它能快速准备的找到用户所关心的文档, 并将结果呈现给用户. 电子教材用户同样存在对电子教材的全文检索的需要, 电子教材设计实现的一种主流方式就是基于 HTML5 进行设计, 对文本内容的检索类似于传统页面的检索, 但又有着特殊之处. 电子教材的全文检索类似于传统因特网、电子图书馆的全文检索系统, 都需要根据页面的内容进行索引生成、检索结果程序等. 但传统的全文搜索一般只能定位到文档级别, 而对于电子教材需要定位每个电子教材逻辑页面的内容级别中, 需要明确能区分出每个电子教材逻辑页面内包含检索关键字相关内容的段落句子信息,

① 基金项目:安徽省教育研究项目(10JDGC015)

收稿时间:2014-03-16;收到修改稿时间:2014-04-21

并能根据相应的信息定位到电子教材的各个逻辑页面中，所以不能直接适用传统 web 站点的全文检索方式进行电子教材的全文检索系统的设计.

1　电子教材基本情况

目前有多个公司推出了电子教材，但还没有统一的内容格式标准，但在电子教材结构上都有课本、单元、小节、文本区的顺序关系，如 kno 和 smartCourse. 通用电子教材的内容组织结构为:

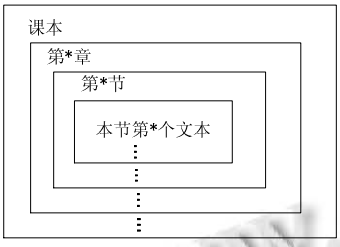

图 1　系统总体框图

学习型电子(教材)书的教学设计和软件设计研究与实现是安徽省教育研究重点项目，在学习型电子书中的教材内容由 HTML 页面组成，每一个 HTML 页面负责一个小节的内容表现，基于 HTML 进行电子教材的设计也是目前主流的技术之一.

学习型电子教材全文检索系统的目的是为了提高电子教材的可用性，全文检索需要能够根据用户输入的搜索关键字，快速获得本书内与关键词相关的内容，不仅要求结果上需要呈现出检索结果所在的具体课本章节位置，还要能够根据用户的选择定位到小节内的具体文本中，并高亮相关的文本结果. 一个好的检索结果，应该能够根据用户的行为对检索结果进行排序，争取将用户期望的结果在最前面现实出来，同时应该尽可能根据用户的输入提供给用户所需的结果，满足用户在无法准确输入关键词的情况下，获得所需要的结果. 通过检索系统，学习者可以快速获取需要的学习知识点，从而提高学习效率，增强了电子教材对学习者学习效率提高的能力.

2　检索系统总体设计

电子教材全文搜索系统分为以下几个部分:

(1)教材文档处理模块[5,6]，教材编写者根据电子教材规范上传教材. 文档处理模块遍历每个单元小节文件，提取出文件内文本内容，记录文本位置信息，对文本进行分词计算、去除停词、标准化单词等操作，最后生成教材的内容索引文件.

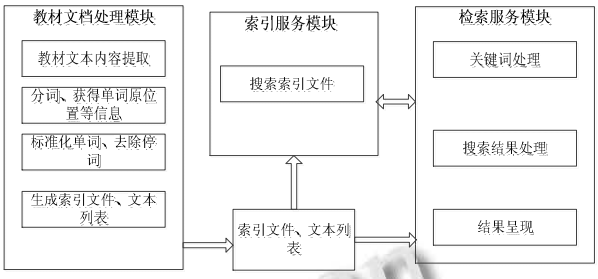

图 2　系统总体设计

(2)索引服务模块，根据检索服务模块提供的处理后关键词在索引文件中查找，并返回初步结果列表.

(3)检索服务模块，对用户输入关键词进行预处理并提交给索引服务模块查询，根据返回的初步索引结果列表，对结果进行同一逻辑句合并处理，相关度计算处理，得到最终索引结果列表. 根据最终索引结果列表，查找对应的文本列表，抽取对应关键字‘句子’描述信息，排序并返回给用户.

3　模块功能实现

3.1　教材文档处理模块

该模块负责对电子教材进行处理，每当教材制作者上传一本教材，模块就对教材进行处理，得到索引文件和文本列表文件，这两个文件为搜索系统的核心部分.

(1)web 页面的信息抽取，遍历每本教材所有的页面，识别页面中的标签内容，提取个页面内的文本内容块，将图片、音频、视频等信息去除，对文本进行分词存储，同时记录文本所在教材的位置信息并标识，将各个文本信息存入文本列表中，文本信息的数据结构如下:

```
文本数据结构
class text{
   String textId        //文本唯一标识
   String unitId;       //单元标识
   String cardId;         //小结标识
   String uuid;         //教材文本标识
   String[] words;    //文本内容
}
ArrayList<text> textlist; //文本列表
```

(2)生成一个页面文章副本，对副本文本中的每个英文单词进行标准化操作[7]，支持用户模糊查询，如单词 ‘cells’ 来自词根 ‘cell’，标准化后 ‘cells’ 在处理时按照 ‘cell’ 处理. 使用事先定义好的无效关键词进行过滤[8]，形成单词索引库，其各个索引词数据结构如下:

索引条目数据结构

```
class wordIndex{
   String index;        //关键词
   String textId;       //所在文本唯一标识
   String wordId;       //文本中单词偏移值
}
```

(3)对生成的基本索引库进行去重处理，生成索引文件[9]. 为了加快查找速度，索引文件采取两级索引模式，第一重根据单词首字母进行索引，第二重为单词内容匹配，索引文件数据结构为一个链表形式，如图所示:

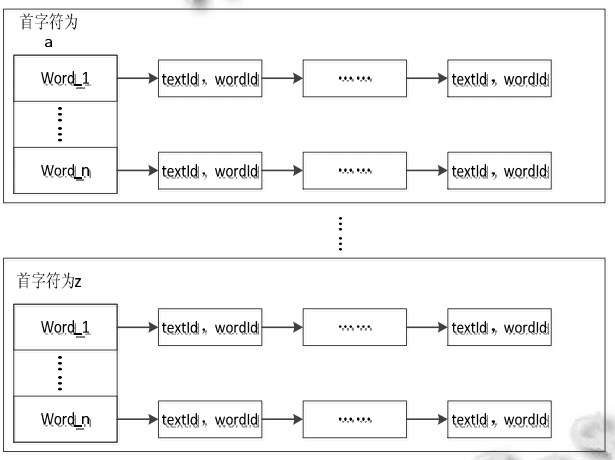

图 3 索引数据结构

生成的索引文件是整个系统的核心，为快速检索服务提供基础. 目前基于电子教材特殊应用，索引文件为一次生成，但只需将生成索引文件过程改为动态执行，就可以满足动态添加页面生成索引的需求.

3.2 索引服务模块

根据检索服务模块提交的关键词查找索引文件，并返回索引结果供检索服务模块处理. 检索服务模块将用户输入内容进行处理后，提交给索引服务模块一个字符串数组，字符串数组中包含需要检索的单词列表. 索引服务模块对每个有效关键字分别进行检索，这样做的目的是尽可能的能够给用户返回足够的有效信息.

在电子教材中，每本教材页面数量相对较少，索引文件占用的内存也相对较少，所以可以将用户输入的有限个数的有效关键字进行处理，以保证能够尽可能多的返回内容，保证用户能在结果中查找到自己想要的结果. 当然，没有根据用户关键字内容和行为对结果进行排序的尽可能多的结果对用户来说也是无效的，所以在检索服务模块中需要对结果进行处理.

3.3 检索服务模块

检索服务模块负责接收用户输入，处理索引服务模块返回的初步索引结果，后查找文本列表将最终结果返回给用户.

(1)关键词处理，根据用户的输入，对关键进行分词处理，记录用户输入单词的顺利及位置，对单词进行标准化操作[7]，并除去无效关键词等. 如用户输入 <The Porter Stemming>，经过预处理后提交给索引服务模块的内容为[‘poter’， ‘stem’]，同时记录获得关键词数位 2.

(2)搜索结果处理，索引服务模块根据提交的关键字数组，依次查询每个单词的结果，获得结果数据列表，结果数据列表的数据结构为:

```
ArrayList<result > resultList;
class result{
   ArrayList<wordIndex> result;
}
```

resultList 返回的结果仅是各个没有处理的关键词结果，它的长度由处理后有效关键字个数决定.

对于电子教材全文检索与普通检索和简单字符匹配不同，检索系统系统需要根据关键字个数、关键字顺序对初步索引结果进行处理. 检索系统需要给用户呈现的不是内容所在的文本，而需要精确到文本内‘句子’级别，‘句子’为足够体现文本内容的单词组，设计时定义为 20 个单词. 而‘句子’内容又由多个单词组成，这些单词不一定完全包含在用户的关键字中，并且顺序也不一定完全一样，这就需要判断 resultList 中各个列表的结果是否属于一个‘句子’中. 判断属于一个句子的关键字组，只保存第一次存储标记为该句子的关键字索引信息.

句子命中关键字个数越多，可以认为句子为期望的结果可能性越大，同时考虑用户期望结果应和用户行为有关，用户当前行为在电子书教材的表现为用户

当前正在阅读的页面信息. 定义积分模型[10], 用于对结果进行排序, 与积分相关的内容有: ‘句子’命中关键词个数、用户当前所在页面与各个结果所在页面关系. 分数计算公式为:

$$score(doc)=\sum_{i=1}^{n} score(item_i)$$

图 4 描述了检索服务系统对索引服务系统返回的初步索引结果的处理过程. 经过处理后, 最终得到一个带有分数的索引结果, 与用户期望越接近的结果分数也越高.

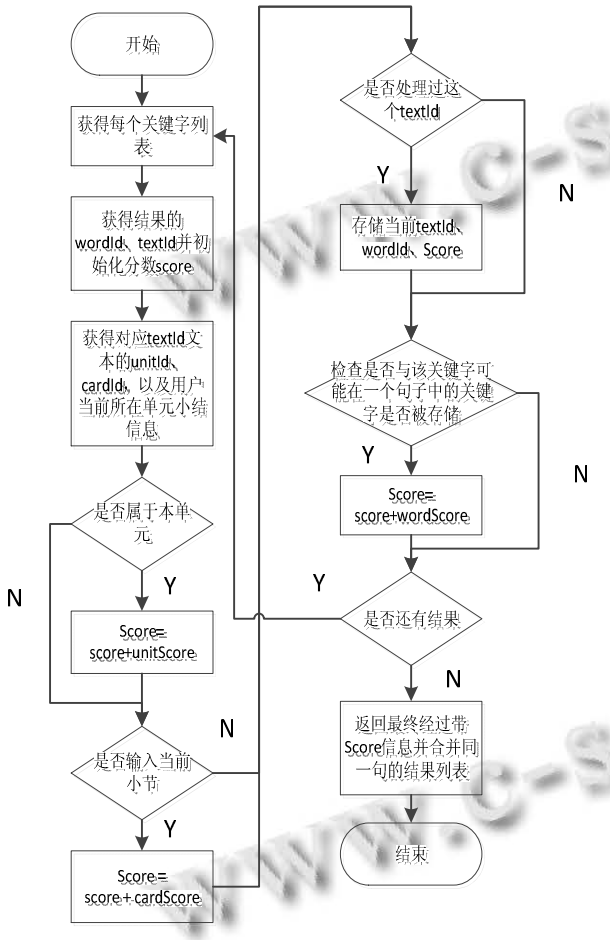

图 4　初步处理结果流程图

(3)按照分数高低值将所查询到的结果进行排序, 得到一个排序列表, 这个列表中的元素由 wordIndex 对象和其对应分数组成. 根据 wordIndex 中的信息, 可以获得该组结果对应的文本位置, 以及关键词在文本中的偏移量, 偏移量加上前后单词信息, 就形成了呈现给用户的句子信息. 同时, 由于记录了关键词个数以及各个关键词, 对于呈现出来的信息还可以对相应关键词进行高亮, 从而使用户更好的获得期望的结果.

3.4 实现效果展示

学习型电子教材的全文检索系统检索效果图如图 5 和图 6, 其中图 5 展示了学习者输入特定的检索关键字后检索系统返回的结果列表, 返回结果列表中不仅包含了关键字的章节位置信息, 还提供了包含该关键字的课文句子信息, 这样用户就可以根据这些信息快速的定位到自己需要查找的信息内容, 实现了检索结果信息的精确定位. 在检索结果中不仅包含了学习者输入的关键字, 还包含与学习者输入关键字相关的用户可能感兴趣的关键字内容, 这样更有利于用户获得自身不明确关键字信息的内容.

当学习者点击检索结果列表项时, 系统就会跳转到具体的内容页中, 内容页中包含关键字的部分就会高亮来提醒用户位置, 加快学习者信息定位, 如图 6.

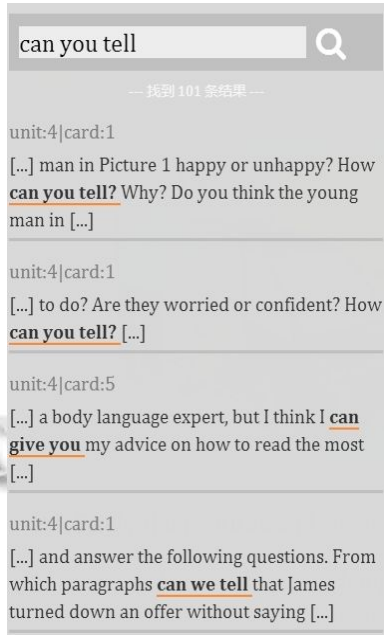

图 5　搜索结果列表

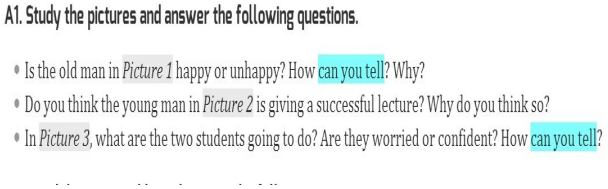

图 6　搜索结果高亮

4　结语

本文在常见全文检索技术的研究分析的基础之上, 结合电子教材的特殊应用, 设计了基于电子教材特殊

应用的全文检索系统. 在设计过程, 充分考虑了电子书的应用环境、使用流程、用户行为等特征, 系统在电子教材应用中性能较佳, 能够快速准确的返回用户期望内容并提供准备的定位信息, 从而提高学习者获取知识的效率, 增强了电子教材提高学习者学习效率的能力.

目前, 我们的设计局限于单一语种的电子教材内容, 后期还要增加对中文、中英文混合文本的支持, 这些支持并不会影响系统的整体结构, 只需要在分词计算上进行优化. 另外, 目前分数模型较为简单, 虽然能够较好的排序结果, 但没有考虑用户已搜索结果的影响等其他因素.

参考文献

1 Gong CH, Chen C, Zhang JB, Huang RH. The development process of e-Textbook for K-12 schools in South Korea and its inspiration to China. ICECE2011, 2011 Int. Conf. on Electrical and Control Engineering, Piscataway: IEEE Conference Publications. 2011. 6879–6883.

2 Davidson AL, Carliner S. Characteristics of effective e-textbooks: Lessons from the literature. IPCC2013, 2013 IEEE International Professional Communication Conference, Piscataway: IEEE Conference Publications. 2013. 479–80

3 Lai JY, Ulhas KR. Understanding acceptance of dedicated e-textbook applications for learning involving Taiwanese university students. Electronic Library, 2012, 3(30): 321–338.

4 Cristy T. Developing a plug-in tool to make OneNote an E-textbook. 2012 2nd Workshop on Developing Tools as Plug-ins, Piscataway. IEEE Conference Publications. 2012. 84–85.

5 冯进,丁博,史殿习,张曙熹,许凯.XML 解析技术研究.计算机工程与科学,2009,2(31):120–124.

6 张维刚,徐永东,雷小强,何辉.Web 全文检索中间件的设计与应用.计算机应用,2011,31(8):2261–2264.

7 Karen S, Jones PW. Reading in Information Retrieval. Morgan Kaufmann Publishers, 1997: 589.

8 付仲恺,秦华.在未分类英文文档集中挖掘相关词的方法.计算机工程与应用,2009,45(5):151–163.

9 邓攀,刘功申.一种高效的倒排序索引存储结构.计算机工程与应用,2008,44(31): 49–152.

10 涂新辉,何婷婷,罗景.一种全文检索系统的设计与实现.计算机工程,2005,17(31):55–57.