

小学语文教学语料库的设计与开发

魏顺平, 何克抗

(北京师范大学 现代教育技术研究所, 北京 100875)

摘要: 针对在小学语文扩展阅读资源制作中存在的篇章检索、句子检索、词语检索中的低效、不准确的问题, 本研究通过借鉴前人提出的语料库建设原则和检索功能, 规定了小学语文教学语料库的建库目的、样本大小、样本多样性以及样本加工方式, 并使用 ASP 程序设计语言、SQL SERVER 2000 数据库服务器来采集、加工语料以及开发语料库检索功能, 最终开发出了基于 Web 的小小学语文教学语料库。该语料库收集了文学大类下各个子类的文章以及小学语文阅读材料中强调的趣味性较强的儿歌、顺口溜、谜语等类别的文章, 字数逾 1400 万, 提供篇章检索和句子检索功能, 可用于语文课件制作、教师备课以及学生自主学习。

关键词: 语料库; 小学语文; 资源制作

中图分类号: G434 **文献标识码:** B

一、问题的提出

从 2000 年 9 月开始, 北京师范大学现代教育技术研究所在北京、广东、河北等地的小学进行了一项基于 Web 的、将信息技术与语文教学有效整合的改革试验(下文将其称为“跨越式试验”)。跨越式试验开展了六年, 取得了较大成功。其中, 试验取得成功的因素之一是从小学一年级开始就将识字、阅读、写作结合起来教学, 并为小学生提供大量阅读材料。^[1] 这些阅读材料均由教改试验课题组的人员制作完成。课题组成员在阅读材料开发方面的工作集中在搜寻与当前课文主题密切相关、体裁相近的阅读材料, 对文本进行注音以及给生字找相关的词句等。^[2] 目前, 资源制作人员搜寻文本主要通过各大搜索引擎网站, 如百度、Google 等; 注音通过 MS Office 办公软件中的 Word 自带的注音功能来完成; 给生字找相关的字词句则是通过个人创作、查找电子词典等。然而上述制作资源的方式、方法、工具存在一些不足, 特别是在文本资源检索方面。目前使用的搜索引擎基本上都是通用的, 只能依据所需材料中所包含的文字来查找相应的网页, 而不能依据文章的主题、体裁来搜索“与当前课文主题密切相关、体裁相近的阅读材料”, 更不用说依据作者、标题、出处来查找“名家大作”, 因为搜索引擎服务器没有对网页进行这些属性的标注。因此, 采用搜索引擎查找阅读资源, 其查全率可能较高, 会返回大量的网页, 但准确率很低, 并且从返回的海量结果中去人工找寻所需的资源非常费时, 严重影响了资源查找的效率。

上述在资源制作中存在的问题一定程度上制约着跨越式试验的顺利开展, 是亟需解决的。而要解决

上述问题, 必须建立一个小学语文教学语料库, 该语料库以篇章为单位, 收集适合小学生阅读的各类体裁的文章, “在保证思想性、知识性的基础上, 突出一定的趣味性”^[3] “力求名家大作”^[4], 能反映现当代中国语言发展现状, 要有篇章的各种属性如语体、体裁、标题、作者、主题、出处、字数以及引用课文等多种属性的标注, 支持模糊查询和全文检索, 要能大幅度提高文本素材的查准率, 并具有可扩展性和可编辑性, 最后以 Web 服务的形式发布出来供资源制作人员使用。

笔者通过调研发现, 目前虽有一些专用语料库如武汉大学的汉语现代文学作品语料库、北京师范大学的中学语文教材语料库等可作为小学语文教学语料库的一部分, 但是仍然不能满足跨越式试验突出趣味性的语料要求。并且现有语料库虽然很多, 但是相互之间的共享程度低, 要想获得这些语料库系统的原始数据库进行二次开发几乎不可能, 且直接购买已有的语料库系统价格不菲。还有, 目前的基于 Web 的语料库如国家语委语料库^[5]、北京大学 CCL 语料库^[6]无法提供篇章检索, 满足不了跨越式资源制作的要求。不过, 现有的语料库的建设方法以及基于 Web 的语料库检索功能设计值得借鉴。因此, 笔者将依据自身的需求, 并利用现有研究成果来开发基于 Web 的供小学语文教学用的可提供篇章检索功能的语料库。

二、相关概念

(一) 语料库的定义

语料库通常指为语言研究收集的、用电子形式保存的语言材料, 由自然出现的书面语或口语的样

本汇集而成,用来代表特定的语言或语言变体。经过科学选材和标注、具有适当规模的语料库能够反映和记录语言的实际情况。人们通过语料库观察和把握语言事实,分析和研究语言系统的规律。语料库已经成为语言学理论研究、应用研究和语言工程不可缺少的基础资源。

(二) 语料库的检索功能

关于语料库的检索功能,中外学者多有论述。胡明扬先生提出,语料库的检索系统要保证能够方便地检索到以下各类信息:总字数及字频、总词数及词频、单字和单词的上下文、按词类大类和小类提取总清单、统计分类总数和分类频度、特种词语、语法规则、句类或句型的统计和提取,等等。^[7]杨惠中(2002)认为,检索工具的基本功能包括词表生成、语篇统计、带语境的关键词(KWIC)索引、搭配词统计、词语型式(Pattern)统计、主题词提取(Key Word List)、词丛(Word Cluster)统计、联想词统计及重组、词图(Plot)统计。^[8]

三、小学语文教学语料库的设计

笔者根据所建语料库的使用目的,参照中外学者的观点,确立了语料库建库原则与方法的几个重点。

(一) 语料库的建库目的

笔者所建立的现代汉语语料库是为小学语文教学与研究服务的,属于“专用型”(Specialized)语料库,因此,该语料库不必像“系统型”(Systematic)语料库那样广泛地代表现代汉语的语言事实全貌。跨越式试验中所需要的阅读资源有两大特点,即“突出一定的趣味性”和“力求名家大作”,因此,笔者所选择的语料样本主要来自文学作品特别是儿童文学作品,如儿歌、顺口溜、谜语、诗歌、寓言、童话以及现代文学大家的作品等,而较少涉及经济、政治、哲学等门类的样本。语料库规模大致控制在1000万字左右,因为这是一个基于Web的语料库,同时会有多人使用,如果数据库过大,必然影响检索速度,使用户无法很好地利用语料库。

(二) 语料样本的容量

关于语料库中每个样本应该包含多少个词(字),不同的语料库有不同的做法。一种做法是把原始语料切分为长度在2000—5000词(字)次之内的片段作为样本,另一种做法是采集整个原始文本作为样本。由于本语料库的用途是为小学语文扩展阅读资源提供篇章和例句检索功能,检索的结果要能返回完整文本和完整句子,因此本语料库的语料样本采用原始文本形式,并另建数据表,存储由原始文本分解得来的句子样本。

(三) 语料样本的多样性

根据本语料库的语文教学目的特性,本语料库的语料样本全部属于文学门类,语料样本的多样性体现在文学大门类下的子门类的多样性。语料库具体内容的选择则参照《全日制语文课程标准(实验稿)》^[9]对中小学生课外阅读材料的规定以及《百年百种优秀中国文学图书》丛书。

(四) 语料样本的处理

笔者没有像国家语委或北大中文系那样对所有的语料样本进行语言属性(如词性、语法、语义、语篇、语用等)的人工标注,因此所建立的语料库属于“生语料库”(Raw Corpus)。但是考虑到语料库的不同检索需求,笔者采用两种不同的方法对语料样本进行处理,并分别存放在数据库的不同数据表中。一种处理方法是将每个语料样本当作一条数据记录存放在数据表中,通过检索程序可对关键词进行全文检索,目的是查找扩展阅读所需的篇章。另一种处理方法是将所有语料样本切分成句子,然后将每个句子当作一条数据记录存放在数据表中,通过检索程序可对关键词进行句子检索,目的在于基于字串、词语或短语检索例句。

四、小学语文教学语料库的开发过程

开发过程包括采集语料、加工语料库、开发语料库的检索功能等环节。所采用的技术路线是使用ASP语言写程序,使用SQL SERVER 2000数据库服务器做数据存储服务,以及使用IIS6.0 Web服务器发布语料库检索网站。

(一) 采集语料

1. 语料来源

跨越式试验项目开展已经六年了,试验过程中积累下来的制作完成的阅读资源可以作为语料库的一部分。Internet上有众多的读书网站,笔者可以从下载需要的文章、书籍,这是本语料库资源的主要来源。

2. 语料采集的难点

由于上述来源的资源都没有统一的组织形式,要将这些结构多样的资源转化为结构统一的语料库是采集过程中的难点。再有,所找到的资源都是网页形式的,而语料的全文则是文本形式的,需要对网页中的标记进行清理。

3. 解决方法

由于来自某个网站的资源其格式还是比较统一的,因此可以人工去发现各个网站对文章的各个部分如标题、作者、正文等的标记规则。然而,常常出现的情况是同一个网站中不同页面中对文章的同一部分所用标记也不尽相同,如某网站的网页1中对文

程,需要投入大量的人力。由于个人能力所限,该项工作暂时没有开展,因此整个文本主题表示模块也就停留在设想阶段。

收集文本中的所有术语及术语频次:可通过建立以术语为特征项的向量,并通过最大匹配算法去获得这些术语在文本中出现的频次,从而得到文本的术语向量。

计算文本中的所有概念及概念权重:建立以概念为特征项的向量,利用得到的术语向量,找到与术语相关联的概念并依据术语与概念的距离计算这些概念相对于这个术语的权值,最后将相对于所有术语的权值累加起来得到概念权重。

完成主题标识:将获得的概念向量中的概念按照其权重自高到低排序,选取前 3 个或 4 个概念作为该文本的主题概念,从而完成主题标识。

(三) 语料库的检索功能

前文述及的语料库检索功能太复杂,考虑到服务器的响应速度,不适合通过 Web 来实现。因此,笔者省去了一部分功能,而突出所需的功能,具体如下:

1. 篇章检索功能

检索项包括篇章标题、作者、出处、语体、体裁、字数以及篇章全文。对标题、作者、出处等项的检索提供模糊匹配功能;对语体、体裁等项提供若干选项供用户选择;对语料全文则提供全文检索,可以检索全文中包含的一项或多项内容。

2. 句子检索功能

可进行“关键词居中”(KWIC)检索,关键词为中文(简体)任意字符串。通过设定“跨距”(Span)来限定检索关键词的左右语境范围,跨距以汉字为单位。返回的检索结果以原始语料的句子为最大话语单位,作为检索的上下文语境的最大范围。如果要查看更大范围的上下文语境可以进一步查看该句子所在的篇章。允许对检索结果进行复制或保存。

句子检索功能可实现基于连续字串、短语、复句型的句子检索,方便课件制作人员查找关于某个词语多种应用情景的句子。

五、结束语

本研究设计并开发了基于 Web 的小学语文教学语料库。该语料库以篇章为单位,收集适合中小学生学习阅读的各类体裁的文章,“在保证思想性、知识性的基础上,突出一定的趣味性”“力求名家大作”,属专用语料库,收集的类别涉及文学大类下各个子类如散文、诗歌、小说、戏剧等类别的文章以及跨越式试验中强调的趣味性较强的儿歌、顺口溜、谜语、童话、寓言、故事等类别的文章,字数愈 1400 万,提供

基于标题、作者、出处、体裁、字数以及全文内容的篇章检索。该语料库是国内首个专门为小学语文教学设计的语料库,可帮助小学语文课件制作人员迅速找到符合多种要求的文本,也可用于小学语文教师备课时查找范文、例句时使用,还可作为学生自主学习时的学习工具。然而,该语料库还存在一些不足,如需完成对语料库文章主题的标注,以进一步提高查准率;并需要扩充语料,特别是各个版本小学语文教材的子语料库的建立。

参考文献:

- [1] 何克抗. 儿童思维发展新论和语文教育的深化改革——对皮亚杰“儿童认知发展阶段论”的质疑[J]. 教育研究, 2004, (1): 55—60.
- [2][3][4] 基础教育跨越式发展试验总课题组. 关于跨越式试验资源建设的说明[DB/OL]. <http://www.nsjy.com/images/news/docc/1617204.doc>.
- [5] 国家语委. 国家语委语料库[DB/OL]. <http://219.238.40.213:8080/CpsQrySv.ssf>, 2006-04-03.
- [6] 北京大学汉语语言学研究中心. 北京大学 CCL 语料库[DB/OL]. http://ccl.pku.edu.cn:8080/ccl_corpus/jsearch/index.jsp?dir=xiandai, 2006-04-03.
- [7] 胡明扬. 现代汉语通用语料库的建库原则和设想[J]. 语言文字应用, 1992, (3): 49—56.
- [8] 杨惠中. 语料库语言学导论[M]. 上海: 上海外语教育出版社, 2002.
- [9] 中华人民共和国教育部. 全日制义务教育语文课程标准(实验稿)[M]. 北京: 北京师范大学出版社, 2001.
- [10] 顾益军, 樊孝忠, 于江德, 李良富. 受限领域中文文本主题标引系统研究[J]. 计算机应用, 2004, (1): 6—7.

收稿日期: 2007 年 2 月 3 日

责任编辑: 张静然