

文章编号: 2096-1472(2017)-06-12-02

# 基于XML文档的藏文网页倒排索引的研究与实现

扎西拉旦, 安见才让

(青海民族大学计算机学院, 青海 西宁 810007)

**摘要:** 如今互联网上藏文信息也不断的扩充,藏文搜索引擎作为常用的信息检索的工具和渠道,倒排索引又是搜索引擎的核心技术之一,倒排索引直接影响搜索引擎检索的结果和响应的速度。之所以文章详细介绍了一个自主开发的藏文网页倒排索引系统,它以XML文档的标签内容作为索引对象,定义了文档和文档属性等概念,采用C#语言对文藏文网页正文构建倒排索引的关键技术和实现方法进一步的阐述,实现了基于XML文档的藏文网页倒排索引数据库的底层实现,提供了技术参考。利用这种方法藏文搜索引擎中信息检索的速度和准确率有所提高。

**关键词:** XML; 藏文网页; 倒排索引

**中图分类号:** TP274 **文献标识码:** A

## Research and Implementation of Inverted Index of Tibetan Web Pages Based on XML Documents

ZHAXI Ladan, ANJIAN Cairang

(College of Computer Science, Qinghai Nationalities University, Xining 810007, China)

**Abstract:** As the Tibetan search engine is a commonly used information retrieval tool and channel, and inverted index is one of the core technology of search engines, inverted index directly affects the search results and response speed of the search engine. The paper introduces a self-developed Tibetan web page inverted index system, which uses the tag content of the XML document as the index object, defines the concept of the document and the document attribute, and constructs the inverted index of the text in C# Language. The key technology and the implementation method of the index are further elaborated, and the bottom implementation of the inverted index database based on the XML document is achieved, which provides technical reference for relevant research. Through this method, the efficiency and accuracy of information retrieval in Tibetan search engines have been effectively improved.

**Keywords:** XML; Tibetan web pages; inverted index

### 1 引言(Introduction)

随着互联网的发展和信息大爆炸的今天,互联网上藏文信息也不断的扩充,同时用藏文的网民对藏文信息的获取有着强烈的需求。有了搜索引擎的帮助,使得我们能够快速、便捷的找到所需要的信息<sup>[1,2]</sup>。本文专门构建了7万多条的藏文词典待索引的词库,词库对每网页内容建立倒排索引,设计了基于xml文档的半结构化数据库化的倒排索引写在磁盘上,有利于存储的安全和倒排索引的更新、询速度上得到提升。

### 2 倒排索引基本概念(Inverted index basic concept)

倒排索引一般包括文档、文档集合、文档编号、词项编号、倒排索引、词典、倒排表和倒排文件组成。

文档主要指的是蜘蛛程序爬行获取的网页或者网页内容,而文档这个概念要更宽泛些,以文本形式存在存储的对象,相比网页来说,有更多种形式,比如Word、txt、PDF、HTML、

XML等都属于文档<sup>[3]</sup>。

由于蜘蛛程序在互联网上获取的网页集合或者网页内容过滤后的构成的集合称之为文档集合。

文档编号或者网页编号是蜘蛛程序爬行获取的网页集合内,每个网页赋予一个唯一的内部编号,以此编号作为这个网页的唯一标识,这样快捷的内部处理,每个网页的内部编号即称之为“文档编号”或者网页编号,后文有时会用DocID来便捷地代表文档编号。词项编号:采用与文档编号类似的方法,每一个单词在搜索引擎内部以唯一的编号作为特征代表某个单词,词项编号可以作为识别某个单词的唯一特征。

倒排索引是指单词所出现的网页编号用矩阵的方法来实现的一种具体存储形式,通过倒排索引可以根据某一个单项快速获取包含这个单项的文档列表或者网页集合。倒排索引主要包括“单词词典”和“倒排文件”两个部分组成,单词词

典是指搜索引擎系统的常用的索引单位是单词项,单词词典是由文档集合或者网页集合中出现过的所有单词项构成的字符串集合,单词词典内每条单词索引项记载单词自身的一些信息外指向“倒排列表”的指针。

倒排列表记录了出现过某个单词项的所有藏文网页编号和藏文网页内容摘要及单词项在该藏文网页中出现频率、位置等信息,每条单词项的记录称之为一个倒排项。关键词通过倒排列表匹配,即可获知哪些藏文网页包含某个单词项<sup>[4,5]</sup>。倒排文件指的是所有单词项的倒排列表按顺序地存储到自己制定的磁盘文件目录中,这个文件被称之为倒排文件,倒排文件是存储倒排索引的物理文件。

### 3 藏文网页倒排方法 (Reverberation embodies the XML method)

蜘蛛程序爬行抓取的所有藏文网页的集合称为网页集合,记为:

$$W = \{w_1, w_2, \dots, w_n\},$$

对W中所有藏文网页净化和过滤后的正文进行分词后的所有不重复词条称为词项集合,记为:

$$T = \{t_1, t_2, \dots, t_m\},$$

由W和T形成一个n × m的矩阵来描述藏文网页和单词项之间的关系<sup>[6]</sup>。由于该矩阵是一个高维稀疏阵,开辟的空间较大,为了减小该矩阵的存储空间问题便产生了正排索引和倒排索引。正排索引指的是藏文网页内容进行切分后建立n个正排索引:

$$N_i(i=1, 2, \dots, n) = \{t_j | t_j \in T \wedge t_j \in w_i\},$$

它表示出现在第i藏文网页中的正文切分后的所有不重复词条。切分后的词项集合为:

$$N = \{n_1, n_2, \dots, n_m\},$$

正排索引的基础上把它转换成倒排索引,由N成m个倒排索引。

$$Ik_j(j=1, 2, \dots, m) = \{w_i | w_i \in W \wedge t_j \in N_i\},$$

表示第j个单词项所对应的所有包含该单词项的藏文网页。藏文网页倒排索引集合为:

$$IK = \{Ik_1, Ik_2, \dots, Ik_m\},$$

通常索引的质量是搜索引擎中信息检索系统成功的关键因素之一<sup>[7]</sup>。

### 倒排索引的工作流程(Reversing the indexing workflow)

首先需要爬虫搜集的网页进行正排索引(分词)后再利用藏文词库对正排索引的结果进行倒排索引<sup>[8]</sup>。本系统实验有一千多个藏文网页的正排索引结果进行构建倒排,即工作流程如图1所示。

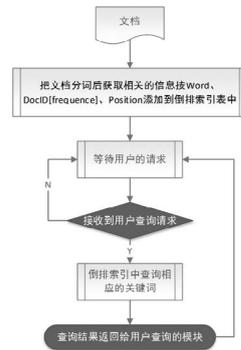


图1 倒排索引的工作流程图

Fig.1 Inverted index of the workflow

### 4 基于XML倒排索引基本思路

首先建立七万多藏文常用单词词典和四千多缩写单词词典后对藏文网页建立倒排索引的,具体模块结构如图2所示。

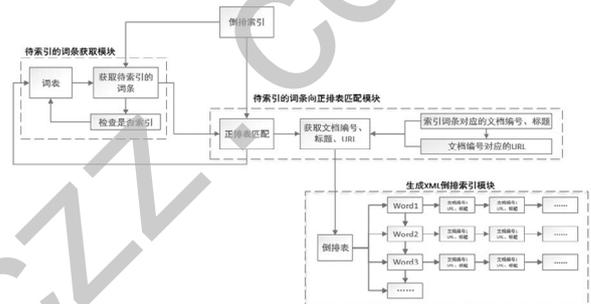


图2 倒排索引的模块结构图

Fig.2 Inverted index of the module structure

图2中包括待索引的词条获取模块、待索引词条向正排匹配模块、生成XML倒排所用模块等三个模块,系统开始将根据检查单词是否索引值来判断后词表中获取单词待索引词条向正排表匹配,词表中从头循环获取每一个不重复的单词到末尾顺序匹配,若有匹配成功,单词项对应的文档编号、网页标题、URL和网页摘要等信息获取,同时利用XML节点结构来构造建立索引做准备工作,XML文档中节点作为一个基本的信存储单元,每个节点有对应的数据存储,是由待索引词条向正排表匹配模块中生成节点和分配,本文将其简记为Word、WordID,其标识为单词和词条编号。把所个二级节点都在父节点<PostingList></PostingList>之间,根据每个单词在网页中出现的频率不同二级节点数量稍有不同,二级节点的节点结构包括如下部分

- ① version.encoding: 文档版本信息与文档编码
- ② <PostingList></PostingList>: 开始与结束
- ③ Word: 词条
- ④ WordID: 词条编号
- ⑤ DocID: 文档编号
- ⑥ Num: 出现的频率
- ⑦ URL: 网址
- ⑧ Title: 网页标题
- ⑨ Concant\_Text: 网页摘要信息

如下是PostingLst.xml文档:

```
<?xml version="1.0" encoding="utf-8"?>
<PostingList>
<Word="语言" WordID="160">
<DocID>PageWithoutTag126</DocID>
<Num>1</Num>
<URL>http://www.bodrigs.com/literature/</
URL>
<Title>青海藏族网青海藏族研究会官网</Title>
<Concant_Text>人类通过语言交流……</Concant_
Text>
……
</PostingList>
```

## 5 结论 (Conclusion)

藏文网页倒排索引直接影响到藏文搜索引擎的检索信息速度快慢及准确率,本文分析了评价藏文搜索引擎系统中倒排索引机制优劣的几个指标,详细介绍了所设计有效的藏文网页倒排索引的方法和过程,它包括单词词表,正排索引匹配模块、XML文档节点结构构造和倒排索引文件生成模块。其中单词词表中有藏文常用的单词项7万多条和4千多藏文缩写词构成,这个基本上覆盖藏文所有的句子,正排表匹配模块中为了提高速率利用二分查找法来完成,生成XML倒排索引模块中正排表匹配模块结果对应的生成节点来完成,词表是用来索引的基本单位集合,索引的最后提出了自己设计思路,设计了一种基于XML文档的藏文网页倒排索引方法。

(上接第16页)

实验结果表明,相同码率下,比例位移法与普通CS方法相比,解码ROI的PSNR更高,从而满足对ROI区域高清晰度的需求。

## 5 结论(Conclusion)

本文通过将位平面位移技术引入CS编码,实现了基于CS的ROI编码方法。通过本文方法,ROI被优先编码、传输和解码重构,有利于高效地进行目标探测和识别;在无线传输带宽较窄或传输意外中断等条件下,也能够保证ROI部分完全或最大程度恢复,尽可能满足面向目标探测识别的图像编码需求。

## 参考文献(References)

- [1] 张河.探测与识别技术[M].北京:北京理工大学出版社,2005.
- [2] JPEG2000 Image Compression Standard[EB/OL].<http://www.jpeg.org/jpeg2000/index.html>.
- [3] D.L.Donoho.Compressed Sensing[J].IEEE Transactions on Information Theory,2006,52(4):1289-1306.
- [4] E.J.Candès S.J.Romberg.Sparsity and incoherence in compressive sampling[J].Inverse Problems,2007,23(3):969-985.
- [5] 杜梅,赵怀慈,赵春阳.兴趣区域优先的多尺度压缩感知渐进

## 参考文献(References)

- [1] HUANG P,LURIE N H,MITRA A S.Searching for experience on the web:an empirical examination of consumer behavior for search and experience goods[J].Journal of Marketing,2009.
- [2] YANG S,GHOSE A.Analyzing the relationship between organic and sponsored search advertising:positive,negative or zero interdependence[J].Marketing Science,2010.
- [3] SENR,KINGR C,SHAWMJ.Buyers'choice of online search strategy and its managerial implications[J].Journal of Management Information Systems,2006.
- [4] 祁坤钰.藏文分词与标注研究[M].兰州:甘肃民族出版社,2015.
- [5] 刘奕群,等.搜索引擎技术基础[M].北京:清华大学出版,2010.
- [6] 安见才让.藏文搜索引擎系统中网页自动摘要的研究[J].微处理机,2010,5:77-80.
- [7] 吴军.数学之美[M].北京:人民邮电出版社,2014.
- [8] 冯志伟.自然语言计算机形式分析的理论与方法[M].北京:中国科学技术大学,2017.

## 作者简介:

北西拉旦(1988-),男,硕士生.研究领域:藏语信息处理工程.

安见才让(1969-),男,博士,教授.研究领域:自然语言信息处理.

编码算法[J].光电子·激光,2015,26(10):2016-2022.

- [6] Z.Wang,et al.Generalized Bitplane-By-Bitplane Shift Method For JPEG2000 ROI Coding.Proc.of the Int.Conf.on Image Processing[C].USA:New York,2002,3:81-84.
- [7] Hsuan-Tsung Wang,S.Ghosh,W.D.Leon-Salas.Compressive sensing recovery from non-ideally quantized measurements. Proc.of the Int.Symp.on Circuits and Systems,China:Beijing,2013:1368-1371.
- [8] S.Mun,J.E.Fowler.Block Compressed Sensing of Images Using Directional Transforms[A].Proc.of the Int.Conf.on Image Processing[C].Egypt:Cairo,2009,12:3021-3024.

## 作者简介:

杜梅(1977-),女,博士,讲师.研究领域:数字图像处理,压缩感知.

曹蔚然(1974-),男,博士,讲师.研究领域:模式识别,数字图像处理.

赵怀慈(1974-),男,博士,研究员.研究领域:复杂系统建模与仿真.