面向信息检索的藏文文本索引策略研究

万福成 何向真 夏建华 杜玉祥
（西北民族大学中国民族信息技术研究院国家民委教育部重点实验室，兰州，730030）
E-mail:wanfucheng@126.com

摘要：互联网文本数量持续爆炸式增长，用户通过互联网查找信息变得更加困难，响应时间得不到满足。本文针对藏文文本身的语言学特点,探讨一种面向信息搜索的藏文文本索引建立策略，建立一种高效的藏文文本索引，以提高藏文信息检索速度。

关键词：信息检索;藏文文本; 索引技术

中图分类号：TP391　　　**文献标识码**：A

Research of Tibetan text index strategy for information retrieval

Wan Fucheng He Xiangzhen Xia Jianhua Du Yuxiang
（Key lab of China’s National Languages Information Technology, Northwest University for Nationalities,730030, Lanzhou Gansu, ）
E-mail:wanfucheng@126.com

Abstract: The quantities of Internet texts grow fast, users search information through the Internet become more and more difficult, answering time of internet is not enough. This paper is to discuss a kind of information search strategy for Tibetan text index, in order to build strategy and establish a highly efficient Tibetan text index, improve the retrieval speed of Tibetan information.

Keywords: information retrieval; Tibetan text; Index technology

0　引言[1]

用户在互联网中搜索信息时，面对是浩瀚的信息海洋，希望可以在尽可能短的时间没找到对自己有用的信息，据统计，用户等待打开一个互联网页面的时间是 7 秒，超过 7 秒，用户在很大程度上会放弃当前页面操作，从而转向其他互联网页面。因此，在最短的时间内对用户的请求进行响应，是每一个搜索引擎都要面对的任务，那么，如何提高检索速度，索引技术就成为了提高搜索引擎检索速度的关键。藏文搜索引擎技术也是如此，藏文文本索引技术是藏文搜索技术的关键。

本文在借鉴汉语索引技术的基础上，结合藏语本身的特点，对藏文文本进行分词、动词归一化处理,虚词归类处理，然后，根据藏文文本词权重建立倒排索引，通过藏语汉语句对齐、分词后的语料进行实验，对实验效果进行分析，索引速度和检索速度得到明显的提升。

1　相关研究

索引技术是搜索引擎的核心技术，人们一直没有停止在索引技术方面的研究工作。文献[1]针对藏语紧缩词识别的问题，提出了一种基于条件随机场的紧缩词识别方法，实验结果表明，基于条件随机场的紧缩词识别方法快速、有效，而且可以方便地与分词模块相结合，显著地提高了藏语分词的效果。文献[2]提出了基于合并因子的多种格式文件索引技术，有

---

[1] [基金项目]西北民族大学科研创新项目（ycx13015） [作者简介] 万福成（1985-)，男，黑龙江人，博士研究生，研究方向中文信息处理.何向真，男，博士研究生，研究方向中文信息处理.夏建华，男，博士研究生，研究方向智能信息处理

效解决了索引大数据量文件速度慢，甚至造成内存溢出等问题。文献[3]提出一种文本索引词项相对权重计算方法，有效地提高索引词对文本内容识别的准确性。文献[4] 构建一个基于哈希索引内存表的模型来提升应用系统查询数据速度。文献[5]在T-树的基础上设计一种新的索引结构，在处理区间查询操作时其效率有明显的提高，也能够很好地解决数据插入、删除操作所造成的数据溢出问题。文献[6]对索引词的自动提取和检索模型技术进行了论述。

2　藏文文本索引模型建立

汉字以字为基本单位，藏文以音节为基本单位，在目前，对于汉语来说，有很多搜索引擎，例如人们广泛使用的谷歌、百度、搜狐等等，但是对于藏文来说，还没有成型的藏文搜索引擎，正是基于这样的现状，本文在汉语搜索引擎的架构下，利用汉语搜索引擎技术，结合藏文语法本身的特点，来具体探讨高效的藏文文本索引策略。

2.1 藏文分词文本准备

藏语是一种拼音文字，有 30 个辅音字母和 4 个元音字母，由这些字母组成音节，由音节构成词。音节之间用音节点“.”作为分隔符，例如“ང་རང་དགེ་རྒན་ཡིན།”（我是老师），句子以下垂符“།”结尾。经过互联网爬虫程序抓取到的网页大量存到文本文件中，对文本文件去除头信息，将内容重新存到文本文件中，藏文分词同中文分词在自然语言处理领域具有相同的地位，因为藏文同中文一样，并没有像英语那样以空格来切分词语，因此，汉语和藏语的处理，首先要进行分词，分词也对句法分析、机器分析、信息抽取等领域有着重大的意义，对抓取的藏文文本文件进行分词，可以以更大粒度建立向量空间，分此后的文本，如：“/དོན་/གྲུབ་/།/རྡོ་རྗེ ”，汉语文本同样进行分词处理。

2.2 词向量空间模型

向量空间模型在文本索引中应用较为广泛，本文以词为单位，建立向量空间模型。向量空间模型，基本思想是：文本空间由 n 个文档组成，每个文档由数量不等的词组成，这样向量空间模型就可以表示成 n 个索引文档 ***Document***（d_1，d_2，⋯，d_n），m 个索引词 ***Term***（t_1，t_2，⋯，t_m）构成，这样就可以形成 $m \times n$ 阶的矩阵，如下式所示

$$A=\left(w_{ij}\right)_{m\times n}=\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ & \cdots & & \\ w_{m1} & w_{m2} & w_{m3} & w_{m4} \end{pmatrix}$$

其中，w_{ij} 为第 i 个文档的第 j 个词的权重，这样每一个行向量为一个索引文本向量。

2.3 倒排索引策略

向量空间由关键词构成，通过每一个关键词建立倒排索引，将指向同一个关键词的不同文本做映射，倒排索引如图 1 所示。

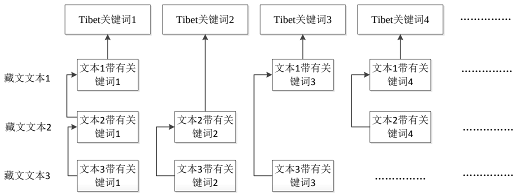

图 1　倒排索引示意图

通过 Hashtable 这种数据结构，将关键词和文本映射数据存储，这样建立起以藏文文本关键字为 key 的倒排索引，hashtable 的 value 值就是每个包含此关键词的文本链表，在这里，对藏文关键词的选取根中文有所不同，如果仅从句子的长度上来看，同样一句话，以概率来

计算，藏语的表示要明显长于汉语，如果将所有的词，包括实词，虚词都进行关键字的索引，势必严重影响索引的时间，进而影响到搜索时间，因此，要将藏语进行词类划分，将藏文实词逐一进行索引，将虚词进行归类后再进行索引，具体的分类和索引方案在第 2 节中有所提及。

2.4 索引词权重计算方法

倒排索引结构是通过计算每一个索引词在相应文本中的权重计算排序之后得到的，计算方法采用的是 tf-idf，tf 是文档频率，是每一个关键词在对应的文本中出现的频率，idf 是逆文档频率，即如果这个关键词在整个文本空间内出现的次数越少，代表这个文档代表的信息量就会越多，就会更容易匹配，从而计算得分会高。当然，这里同样也要考虑到实词和虚词的不同，实词可以按照上面的方法来进行计算，虚词，需要进行归类以后再进行计算，在实际计算中，为了更好的进行比较，将文档频率做平方根计算，分数的计算范围缩小，将逆文档频率做对数计算，这样计算之后进行排序，一次链接到相应的关键词链表，为了真正能够达到快速的效果，采用词项列表常驻内存的方式，这样，可以提高访问速度，倒排索引以词项在系统中的标识(Word ID)为索引项的键值，再输入 Word ID 后，可以得到一个文档标识(Doc ID)的列表，即对每一个文档对应的出现记录信息列表。图 2 给出了此项列表和倒排索引的结构，同一词项在倒排索引项中的文档标识是按照逐次递增的顺序进行排列的，通过常驻内存的词项列表来对具体的文本文件进行关联，通过指针的方式来查找具体的文本，指针指向的是文本文件的地址，而不是实际的具体文件，查找具体的文件，需要等到呈现具体的搜索列表项时才进行真正的物理文件的读取，这样做，也是为了达到减少索引时间和搜索时间的目的。

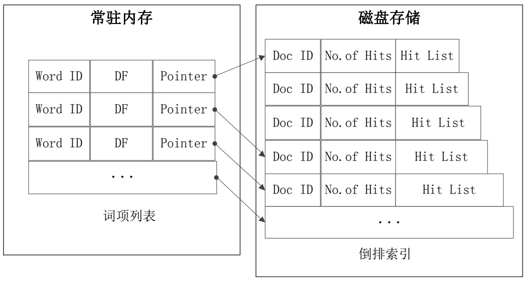

图 2 词项列表和倒排索引的结构

3 索引流程设计与实现

对藏文文本建立索引，最简单的方式是文本不经过处理，直接在每一个藏文字节的基础上进行索引，这样做的好处是方便、快速，无论什么样的藏文文本都可以按照这种方式建立索引，粒度可以是单个字节，也可以使双字节，三字节等等，然而，这样做，粒度小时，建立索引的时间会比较多，粒度大时，切分出来的音节不一定会是一个关键词，因此，本文在藏文分词的基础上，对每个文本进行分析，索引建立的流程如图 3 所示。

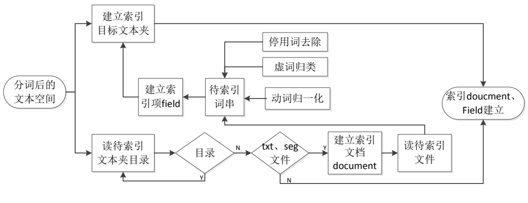

图 3 文本索引建立流程图

在索引过程中，对索引词串处理的过程中，汉语和藏语存在很明显的区别。

首先从动词上来看，汉语缺少动词的形态变化，不区分过去时，现在时和将来时，在

藏语上动词要对时态进行区分，有些动词的过去时和将来时跟动词的原形相同，有些动词过去时和将来时都是要进行区分的。因此，在藏语处理中，将藏语动词做归一化处理，将过去时，将来时的动词都替换成原型进行归一化处理，这样可以降低索引数量，提高索引速度。

藏文虚词和汉语中的虚词在句子中占的比重有所不同。例如对同一个句子，“中国政府和澳大利亚政府今天签署了一项协议，澳方将向中国提供1.5亿澳元优惠财政政策。”。藏语对应翻译的句子为“ཀྲུང་གོ་དང་ ཨོ་སི་ཁྲི་ཏུ་ལི་ཡའི་སྲིད་གཞུང་གིས་དེ་རིང་ཨོ་ཕྱོགས་ཀྱིས་ཀྲུང་གོ་ལ་ཕན་ལྡན་ ནོར་སྲིད་ཨོ་སྒོར་དུང་ཕྱུར་ 1.5ཀྱི་དངུལ་བུན་བ དང་རྒྱུའི་གྲོས་ཆོད་ཐོ་འགོད་བྱས།”（该句来源于宾大中文树库）。汉语的句子中虚词有“和”、“了”、“将”、“向”4 个，而在藏语中除了这 4 个以外，还有 3 个代表领属关系的格助词“འི་”，有 2 个代表施事的格助词“གིས”，有 1 个代表对象的格助词“ལ་”，因此，藏语在虚词的使用上更频繁，需要对虚词进行特殊的归类处理，可分为领属类、目的类、施事类等等。这样有效地进行分类以后，建立索引模型。

此外，藏文跟中文的字串一样，有一些停用词也要进行删除处理，这样藏文经过和中文的不同的语处理，就可以用于建立属于藏文自身的索引模型。

3.1 生成索引目录

对藏文的关键词建立索引以后，生成索引文件，索引文件存在于索引目录之下，索引目录存放生成的文本索引，同时这个目录也是信息搜索查找的目录，搜索信息首先搜索信息的索引，然后根据索引信息查找信息所在的文件，路径以及内容。生成索引目录，为建立文本索引做准备。索引目录通常为本地的磁盘文件目录，这样建立起来的索引，可以长时间的保存起来，方便用户搜索，这样做是考虑到搜索文件会遇到频繁的读文件操作，建立索引又会频繁的写文件，因此，定期对索引目录进行备份，也是必须要进行的工作。

3.2 遍历待索引文件目录

分此后的文本空间，文本和文件夹共存，因为需要对所有文本建立索引，那么就需要以深度优先算法遍历每一个文件，为每一个文件建立索引，形成每一个文本的索引 document。索引 document 的建立，可以用来接收索引项存放的数据。

3.3 读取分词项建立索引 field

读取每一个文本，由于分词后的藏文文本是以“/དོན་/སྒྲུབ/།/རྡོ་རྗེ་ ”这样的形式存在，由于存在特殊的字符“/”，在处理字符串流时，将“/”替换为空格，通过 whitespaceAnalyzer（空格分词法）来将字符串中的词分割，将分割出的藏文关键词，循环建立索引，流程如图 4 所示。

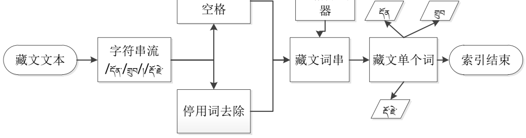

图 4 索引 field 建立流程图

文本经过图 3 一系列的处理过程，就可以在整个文本空间中为每一个文本建立索引 document，为每一个 document 中的词建立索引项 field。将 document 和 field 存入到索引目录中，为检索服务做准备，检索项和检索文档，都是检索内容的重要组成。

3.4 检索 query 的分词

检索流程中还需要对用户的查询语句 query 做分词处理，同时也将特殊字符替换，去除停用词，通过布尔查询，将这些关键词做联合查询处理，在索引目录文件夹内搜索，得到含有这些关键词的文本，这些文本都是经过计算，按照得分高低排列。

4　测试结果分析

本文为了测试索引和检索效果，本文采用了用于机器翻译评测的藏汉双语对齐语料库，一共101629行对齐经过分词之后的藏语和汉语文本进行实验，文本数量为3384个，汉语语料库总大小为8.79G，藏语生语料为9.69G，处理后的藏语语料为9.17G。索引和搜索工具采用开源lucene软件，通过对汉语、藏语未处理、藏语经过处理的文本分别进行索引，查看索引时间，以这个词“དཀར་ཆག”为例，进行搜索，查看索引和搜索时间。索引时间的计算均为5次实验时间的平均值，效果如表3所示。

表1 藏语特征索引、搜索对比表

	藏语					
	未分词	分词	未归一化	归一化	虚词分类	虚词未分类
索引(ms)	7806	6173	6374	6289	7126	6923
搜索(ms)	2713	1016	1326	1135	1926	1824

由表1可以看出，在同样的藏文预料上分别进行分词、动词归一化处理、虚词归类处理，对比没有进行处理的语料进行索引和搜索，分词对藏文索引的影响比较大，能够显著提高索引和搜索速度，动词归类和虚词归一化对索引和搜索时间也有所改进，但是效果不明显，原因可能是，根据藏文动词词典来看，藏文动词归一化处理后有1100多个藏文动词，跟海量的索引信息相比还是比较小的，藏文虚词只是针对格助词的用法进行了归类，并没有进行细致的区分，因此，提升的效果不是很明显。

表2 藏语、汉语索引和检索时间对比表

	汉语	藏语	
		未处理	处理后
索引(ms)	7205	7301	6257
搜索(ms)	2035	1849	1032

由表2可以看出，在藏汉平行语料库中，经过处理之后建立索引的时间要小于没有处理建立索引的时间，随着语料库的规模增大，索引建立的时间也要小于同等规模的没有处理的语料库，同时，信息检索的时间也比较少，因此，将文本进行预处理，经过分词，特殊字符替换，文本分析，停用词去除，之后建立索引，可以明显提升索引时间和检索时间，此外，将待搜索的句子进行预处理也是提升检索速度的一个很重要的方面。

5　结语和下一步的工作

目前，没有成型的藏文搜索引擎，本文基于倒排索引和文本权重建立索引策略，建立搜索引擎，文本在经过分词、动词归一化和虚词归类处理的基础上建立索引策略，通过实验进行对比，索引时间和检索时间都有所减少。下一步的工作是对比不同索引和搜索方法进行横向对比，得出相应的结论。

参考文献：

[1] 李亚超,加羊吉，宗成庆,于洪志.基于条件随机场的藏语自动分词方法研究与实现,中文信息学报[J],2012,7(2).

[2] 孙广路,易成岐,郎非,基于合并因子的多种格式文件索引技术,哈尔滨理工大学学报[J],2012,17(2).

[3] 蓝海洋,周杰韩,张和明,文本索引词项相对权重计算方法与应用,计算机工程与应用[J],2003,39(5).

[4] 杨燕明,鲁志军,陈煜,孙权,一种基于哈希索引的内存表模型,计算机应用与软件[J],2012,29(1).

[5] 王平,朱敏,姜雪,一种优化的T-tree索引算法,计算机应用与软件[J],2011,28(2).

[6] 励子闰，余青松，陈胜东．基于全文检索引擎的信息检索技术的应用研究[J]．计算机与数字工程，2008，36(9).