第34卷 第21期 计 算 机 工 程 2008年11月
Vol.34 No.21 Computer Engineering November 2008
·软件技术与数据库· 文章编号：1000—3428(2008)21—0045—03 文献标识码：A 中图分类号：TP311.13

维、哈、柯全文搜索引擎检索器的关键技术

吐尔地·托合提，维尼拉·木沙江，艾斯卡尔·艾木都拉
(新疆大学信息科学与工程学院，乌鲁木齐 830046)

摘 要：研究维、哈、柯全文搜索引擎检索器的关键问题，提出有效的解决方法，包括在用户计算机没有安装本地输入法和字库的情况下输入维、哈、柯文检索词并正常显示搜索结果，针对具有高拼写错误率的维、哈、柯文检索词进行检错、纠错处理，返回给用户正确而全面的搜索结果等。实验结果表明，该方法为用户提供方便的同时明显提高了维、哈、柯文搜索引擎的查全率和查准率。

检索词：在线处理；检错；纠错；词根切分；同化处理

Key Techniques of Uyghur, Kazak, Kyrgyz Full-text Search Engine Retrieval Server

TURDI Tohti, WINIRA Musajan, ASKAR Hamdulla
(Information Science and Technology Institute, Xinjiang University, Urumqi 830046)

【Abstract】This paper studies the key problems of Uyghur, Kazak, Kyrgyz full-text search engine retrieval server and proposes an effective solution, including inputting Uyghur, Kazak, Kyrgyz keywords and shows normal search results without installing the local input method in users' computers, detecting and correcting Uyghur, Kazak, Kyrgyz keywords with high spelling mistake rate, returning the correct and comprehensive search results to users, etc. Experimental results indicat that the solution provides convenience for users and obviously improves the precision and recall of Uyghur, Kazak, Kyrgyz search engine.

【Key words】online processing; error detection; error correction; root segmentation; assimilation processing

1 概述

随着Internet和WWW的发展，Internet上的资源越来越丰富，基于Internet的各类信息检索服务随之诞生并获得了迅速的发展。中国有非常优秀的中文、英文搜索引擎，如Google, Baidu，但这些搜索引擎没有解决少数民族语言文字特征方面的关键问题(民文在线输入及显示、标准字符编码、网页布局及书写方向、检索词语处理等)，完全不能满足广大少数民族网络用户的信息检索需求。针对维吾尔文、哈萨克文、柯尔克孜文(以下简称维、哈、柯文)等少数民族语言的搜索引擎的研究至今还处在空白阶段，没有一个比较成熟的搜索引擎。为了能够在网上快速检索以本民族语言发布的信息而开发一个多文种搜索引擎是新疆少数民族面临的一个亟待解决的问题。图1是本文设计的维、哈、柯文搜索引擎的体系结构[1-2]。

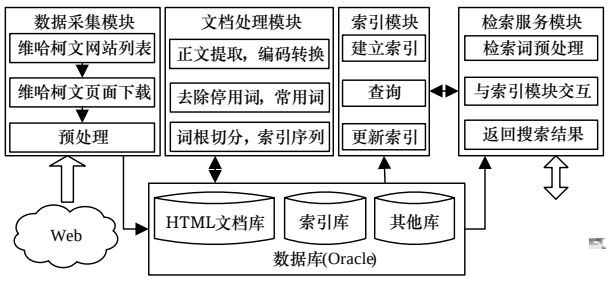

图1 维、哈、柯文搜索引擎系统架构

检索器(retrieval server)是本搜索引擎的用户接口，如果该模块中的一系列关键问题没有得到彻底的解决，就不能满足少数民族网络用户的信息检索需求，还会严重影响该搜索引擎的查准率、查全率等主要性能。针对以上实际情况，本文研究了实现民文搜索引擎检索器的相关技术，其中包括维、哈、柯文在线处理[3]、检索词的检错与纠错[4]、检索词词根切分[5]、同化(弱化)处理等。

2 维、哈、柯文的特点及检索器的技术难点

维、哈、柯文字母与中英文不同的是：(1)书写方向相反，汉字和西文的书写方向是从左到右，而维、哈、柯文的书写是从右向左。(2)维吾尔文由32个字母组成，哈文由33个字母组成，柯文由30个字母组成，各有120多个字符形式。维、哈、柯文的一个字母根据在单词中位置的不同会有4种变形。即首写、中写、尾写形和独立形式。在文字输入时，要根据字母在文字中的位置确定使用何种形式。字母不等宽且某字母的4种形式也不等宽。维、哈、柯语中的每一个词由这些不等宽的120多个字母前后相连而成，书写出错率比较高，词缀直接连接在词根上表示不同的含义。本文的检索器需要解决的技术难点主要有：(1)可直接在Web搜索页中输入维、哈、柯文检索词(keywords)并正常显示维、哈、柯文搜索结果，用户不需要安装民文输入法和字库。(2)在用户输入的检索词出现拼写错误的情况下，确保查准率。(3)根据检索词词根，搜索同词根范围的全部内容，保证查全率。

基金项目：新疆维吾尔自治区高技术研究与发展计划基金资助项目(200612115)；新疆维吾尔自治区高校科研计划基金资助重点项目(XJEDU2006113)

作者简介：吐尔地·托合提(1975－)，男，讲师、硕士研究生，主研方向：自然语言处理，信息检索；维尼拉·木沙江，教授；艾斯卡尔·艾木都拉，教授、博士生导师

收稿日期：2007-12-20 **E-mail：**turdy@xju.edu.cn

3 维、哈、柯文在线处理技术

为了实现在线输入与显示，根据维、哈、柯文输入法，创建了将键盘上的字母转换成对应的维、哈、柯文字符的映射表。这样，无须在本地安装任何民文输入法就可以输入民文。由于键盘上的字母包括大写、小写共有 52 个字符，而维、哈、柯文分别有 32 个、33 个和 30 个字母，因此在映射表中使用了 26 个小写字母和部分大写字母。比如，用 26 个小写字母和大写字母 D, F, G, H, J, K 作为维吾尔文的输入字符，对于其他大写字母，系统不做任何处理，即当用户输入其他大写字母时，系统不显示任何维吾尔文字母。还使用了 Microsoft 的字体处理工具 WEFT(Web Embedding Font Tool)、UTF-8-ASCII 兼容的多字节 Unicode 编码法和某些脚本语言代码创建了维、哈、柯文 Unicode 索引表 Keymap，利用 JavaScript 提供的 CaptureEvent 函数在客户端上挂一个“浏览器级”钩子函数来处理键盘和鼠标输入信息。每当按下和释放键时，钩子函数都会被使用，一旦读入虚拟键码，它立即调用 Window.Event 函数。Window.Event 函数通过调用 Window.Event.Keycode 函数将虚拟键码映射到维、哈、柯文 Unicode 索引表，得到对应的 Unicode 编码后将它放入到自动选型模块进行自动选型，最后输出模块里的 CharAt(),addchar() 等函数，将字符串发送到网页指定的位置上。具体方法如下：

(1)使用 WEFT 创建“嵌入字体”并把它放在服务器上。

(2)在 HTML 文件中插入以下代码：

```
<head>
…
<script src=”uyghur.js”></script>
…
</head>
```

(3)在网页中的文本框代码中插入以下代码：

```
Onkeypress=” addchar(this,event);”
```

例如：

```
<textarea name=”search” rows=”8” cols=”80” wrap=”soft” onkeypress=”return addchar(this,event);”></textarea>
```

由此可以实现网页中维、哈、柯文 Unicode 字符的输入和显示。

4 检索词的检错与纠错技术

维、哈、柯文中的一个词是由 120 多个字母前后相连而成，字符集中相似字母较多，拼写出错率比较高(计算机输入或手写出错率可达 50%以上)，词缀与词根相连表示不同的含义。为了能够快速、正确地搜索本民族语言的、与用户输入的检索词有关的内容，开始搜索之前需要对用户输入的检索词进行检错及纠错处理。

(1)检索词检错。维、哈、柯文无论是印刷体还是手写体都具有草书特点，拼写出错率较高。如，用户要搜索与检索词“学校”有关的内容，则输入维文检索词مەكتەپ(学校)是正确的，但是输入了بەكتەپ，مەكيەپ，ەكتەپ等是错误的(维语中没有这些词)。因此，要快速、正确地搜索相关内容的网页，在用户输入检索词并开始搜索之前对检索词进行检错是很有必要的。在研究中笔者收集了大量的词语，分别创建了维吾尔语、哈萨克语和柯尔克孜语词语库作为检错依据。先从词语库中查找每一个检索词，如找到，则是正确的，否则，是错误的。

(2)检索词纠错(候选词选择)。当用户输入的检索词出现拼写错误时，搜索引擎检索器应该给用户返回拼写错误提示或者返回若干个候选检索词让用户再次确定，保证检索词与搜索结果的正确性。在研究中发现，文本中 60%以上的错误是由“单个错误的拼写”(single-error misspelling)引起的。“单个错误”就是指下列错误中的某一个：1)插入，把كېلىش错误地打成كېپلىش；2)脱落，把كېلىش错误地打成كلىش；3)替代，把كېلىش错误地打成كىلىش；4)换位，把كېلىش错误地打成كېلڭىش。

为了提高纠错效率，笔者统计并收集了拼写出错率在 95%以上的约 6 000 个单词的错误拼写和正确拼写，并建立了对照词语库。算法流程是：先从对照词语库中查找拼写错误的检索词，如找到，则直接得到正确的检索词，否则，用最小编辑距离(minimum edit distance)算法和长度优先算法找到与用户输入的检索词最接近的若干个(本算法中限定为 5 个)候选词返回给用户。2 个符号串之间的最小编辑距离是指把一个符号串转换为另一个符号串时所需的最少编辑操作(插入、删除和替换)次数。例如，كىلىشكىلىش之间的编辑距离是 1(一个字符替换操作)，根据最小编辑距离与字符长度匹配，返回给用户匹配概率最大的 5 个候选词，让用户再次确定。

这 2 部分都嵌入到一个类库文件 Spelling.dll 中，每个部分都适当地提供了实现其功能的接口。在内核技术方面，把词语库里的所有词放到微软 .Net Framework2.0 以后的版本所提供的 Dictionary 类中，由此极大地提高了检错纠错速度。

5 检索词词根切分

词根切分算法流程如图 2 所示。

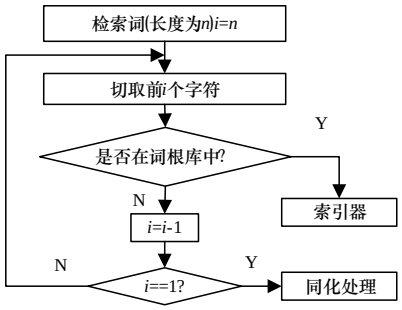

图 2 词根切分算法流程

维、哈和柯语都是黏性语言，在这一类语言中，词(word)是最小的独立运用的语言单位。词由词根和附加成分构成。如：用户输入的维文检索词جۇڭگونىڭ نوپۇسى(中国的人口)中，جۇڭگونىڭ=جۇڭگو+نىڭ、نوپۇسى=نوپۇس+ى。搜索引擎应该能够搜索包含词根جۇڭگو(中国)和نوپۇس(人口)的所有网页。因此，设计高准确率的切分算法对检索词进行词根切分，能够给用户提供比较全面的搜索结果，这也是本搜索引擎检索器的关键技术之一。在设计中，分别创建了维、哈和柯文词根库作为主要的切分依据。

6 同化(弱化)处理

同化处理算法流程如图 3 所示。

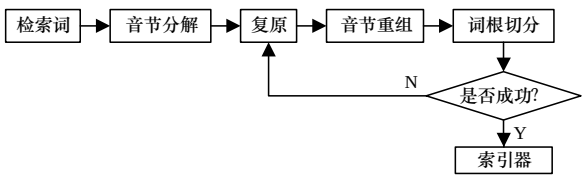

图 3 同化处理算法流程

如果检索词没有拼写错误或者有错误但已纠正，通过词根切分算法还是找不到词根，那就表明该词中产生了同化现

象。同化是指有的维、哈、柯文词的最后一个音节的有些字母(元音或辅音)可能被替换成另外一些字母。如，维文词根连接某些词缀时其词根最后一个音节的ئا ,ئە 被替换成ئى或ئې，这种现象称为元音同化。解决同化问题是维、哈、柯文搜索引擎检索器的又一技术重点和难点。在本算法中首先进行音节分解，然后从最后一个音节开始检查，如音节中有元音字ئى或ئې，则分别替换成ئا 和ئە ，这一过程称为复原。对复原后的音节再进行重组，然后进行词根切分。

7 网页相关内容提取、格式处理、搜索结果返回

检索器的工作流程如图 4 所示。

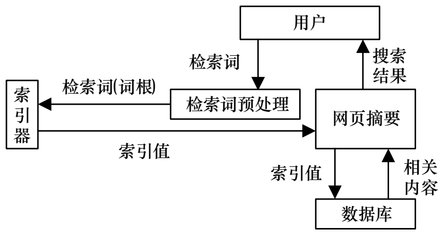

图 4 维、哈、柯文搜索引擎检索器工作流程

用户提交检索词、经过检索词预处理模块并提取检索词词根之后，检索器激活通信服务程序与索引器 Socket 连接，将检索词词根发送给索引器，并从索引器得到与检索词相关的索引值，包括该检索词出现的网页个数、前 10 个网页在索引库中的 ID 号、每一个网页中检索词出现的次数、位置、长度等信息，然后交给本模块处理。本模块根据索引信息从索引库提取每个网页 ID 号对应的网页标题(title)、URL 和网页内容(content)。再根据检索词在该网页中的出现次数、出现位置、长度等信息提取前 5 个检索词和每个检索词前后部分内容构成长度为 600 个字符的网页摘要，再加上网页标题、URL 等内容生成搜索结果页面返回给用户。

8 结束语

本文对维吾尔文、哈萨克文、柯尔克孜文在线处理、检索词的检错与纠错处理、检索词词根切分、同化(弱化)处理等多文种搜索引擎检索器的关键技术做了深入的研究，并解决了维、哈、柯文在线输入及显示问题，以及对搜索引擎的查准率和查全率至关重要的检索词预处理问题，完全排除了维、哈、柯文信息检索中的语言障碍。该搜索引擎已经以 www.tapkak.com 为域名开始提供维、哈、柯文信息检索服务，得到了广大维吾尔族、哈萨克族、柯尔克孜族网络用户的认可，成为新疆地区唯一的少数民族语种的全文搜索引擎。

参考文献

[1] 李晓明, 闫宏飞, 王继民. 搜索引擎——原理、技术与系统[M]. 北京: 科学出版社, 2005: 20-27.
[2] 徐宝文, 张卫丰. 搜索引擎与信息获取技术[M]. 北京: 清华大学出版社, 2003: 120-121.
[3] 维尼拉·木沙江, 艾尔肯·伊米尔. 维文 Unicode 在线处理技术与实现[J]. 新疆大学学报, 2004, 21(3): 332-334.
[4] 米吉提·阿不力米提. 在多文种环境下的维吾尔语文字校对系统的开发研究[J]. 系统工程理论与实践, 2003, 23(5): 117-124.
[5] 古丽拉·阿东别克, 米吉提·阿布力米提. 维吾尔语词切分方法初探[J]. 中文信息学报, 2004,18(6): 61- 65.

~~~~~~

（上接第 39 页）

(2)在同样标注 1 000 条样本的条件下，分别使用 PRank 算法与 Active PRank 算法进行标注，使用相同的测试集进行对比实验。实验结果如图 3 所示。从图 3 可以看出，在同样标注 1 000 条样本的条件下，Active PRank 算法的 *MAP* 值和 *NDCG* 值都高于 PRank 算法。因此，使用 Active PRank 算法在同等标注量的条件下，无论是排序的整体效果，还是排序结果中顶部序列的准确性，都优于 PRank 算法。

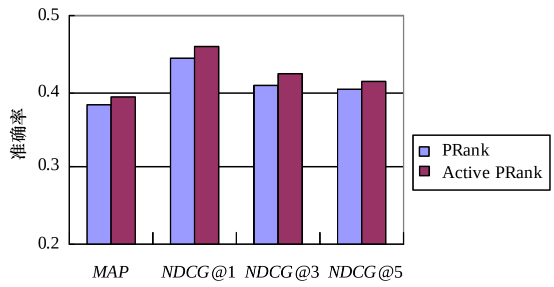

图 3 Active PRank 算法与 PRank 算法的准确率比较

(3)当 *MAP* 值达到并稳定在 0.35 以上时，使用 Active PRank 算法只需要标注约 300 条样本；而在同等条件下，使用 PRank 算法需要标注约 700 条样本。可见，使用 Active PRank 算法可在保证排序模型性能的前提下，减少样本的标注量。

6 结束语

本文介绍了排序学习和主动学习的思想，并将主动学习的思想引入排序学习中，提出 Active PRank 算法。实验结果证明，该算法可以减少样本标注量，降低标注代价，提高排序结果的正确率。

参考文献

[1] Crammer K, Singer Y. Pranking with Ranking[C]//Proceedings of the Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: [s.n.], 2001.
[2] Herbrich R, Graepel T, Obermayer K. Large Margin Rank Boundaries for Ordinal Regression[M]//Smola A, Bartlett P, Scholkopf B. Advances in Large Margin Classifiers. [S. l.]: MIT Press, 2000.
[3] Seung Sebastian, Opper M, Sompolinsky H. Query by Committee[C]//Proceedings of the 15th Annual ACM Workshop on Computational Learning Theory. California, USA: Morgan Kaufmann Publishers Inc., 1992.
[4] Yates B, Ribeiro R A, Neto B. Modern Information Retrieval[M]. Boston, USA: Addison-Wesley Longman Publishing Inc., 1999.
[5] Jarvelin K, Kekalainen J. Cumulated Gain-based Evaluation of IR Techniques[J]. ACM Transactions on Information Systems, 2002, 20(4): 422-446.
~~~~~~