垂直搜索引擎中分词技术的算法研究

邹 嵩,赵诗阳,周新志
(四川大学 电子信息学院,四川 成都 610064)

摘 要:为了提高垂直搜索引擎的检索效率,介绍了垂直搜索引擎中的分词技术。文中主要通过研究最大长度匹配算法,提出了改进后的最大长度匹配算法以提高检索效率。改进后的算法显示,分词效果得到了一定的提升,体现了相对于普通最大长度匹配算法的优势。且通过将改进后的方法与普通最大长度匹配算法相比较可知,改进后的算法提高了搜索的正确率,提升了检索的效率,是一种对最大长度匹配算法的有效改进,由此也体现了搜索引擎中算法合理设计可以提升搜索性能。

关键词:垂直搜索引擎;分词技术;最大长度匹配算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2012)02-0131-03

Word Segmentation Algorithm in Vertical Search Engine

ZOU Song,ZHAO Shi-yang,ZHOU Xin-zhi
(College of Electronics and Information Engineering,Sichuan University,Chengdu 610064,China)

Abstract:In order to improve vertical search engine's search efficiency,describe the word segmentation in vertical search engine. It studies the characteristics of the maximum length matching algorithms,propose the improving of the maximum length matching algorithm to improve search efficiency. The improved algorithm shows that,it results to a certain improvement to the segmentation,reflected the advantage compared to the ordinary maximum length matching algorithms. And compared to the maximum length matching algorithms,the improved algorithms can improve the search accuracy rate,enhance the efficiency of retrieval,and it has a effective improvement to the maximum length matching algorithms,and it also reflects that the rational design of search engine algorithms can improve search performance.

Key words:vertical search engine;word segmentation;maximum length matching algorithms

0 引 言

目前,传统的搜索引擎,在针对各行业相关专业文档搜索查询方面已有所局限,在一些专业领域,通用搜索引擎的检索必然会带来"搜索噪音"、"信息过载"、"信息冗余"等一系列负面结果,不能真正实现高效与便捷。为了解决以上一系列问题,采用具有专业领域检索优势的垂直检索思想,并在此基础上进行优化,提高检索过程的准确度和查全度来完善整个系统的核心部分,最终实现专业行业领域的文档信息检索与管理显得十分必要[1]。而分词技术是任何一个垂直搜索引擎的"灵魂",能否将检索请求中的关键字通过正确的理解从句子中正确切分出来并进行匹配查找,是一个搜索引擎能否得到正确结果的关键过程。分词的准确与否会直接影响到检索结果的正确率。因此,文中将主要研究垂直搜索引擎中分词技术的算法的改进[2]。

收稿日期:2011-07-04;**修回日期**:2011-10-15
作者简介:邹 嵩(1985-),男,四川泸州人,硕士研究生,研究方向为电子与通信工程。

1 搜索引擎中的分词技术

现行常用的分词技术有单汉字分词、双汉字分词和最大长度匹配算法[3]。

1.1 单汉字分词

按照单汉字分词法切分,将文章中的每一个汉字都输入到索引库中,设:$S=C_1C_2C_3C_4\cdots C_{n-1}C_n$,$n$ 为需要进行切分的字符串,其中 $C_i(i=1,2,3,\cdots,n)$ 为单个的汉字字符,按照单汉字切分,结果为:$S=\{C_1\},\{C_2\},\{C_3\},\cdots,\{C_{n-2}\},\{C_{n-1}\},\{C_n\}$。如将"中国人民银行"分解为"中"、"国"、"人"、"民"、"银"、"行",用户查询的语句根据单汉字匹配原则来进行检索。这一方法的优点是灵活,可以有很高的查全率。但是它的问题也显而易见,那就是它需要设计复杂的单汉字匹配算法,并且由于中文中有很多存在歧义的地方,所以很可能会返回一些跟用户输入无关的结果,例如:查"国人"同时也会返回"中国人民银行"[4]。

1.2 双汉字分词

又称为二分法,就是将每两个字当作一个词语进行切分。$S=C_1C_2C_3C_4\cdots C_{n-1}C_n$,为需要进行切分的字

符串，$C_i(i=1,2,3,\cdots,n)$ 为单个的汉字字符，按照双汉字切分，结果 $S=\{C_1C_2\},\{C_2C_3\},\{C_3C_4\},\cdots,\{C_{n-2}C_{n-1}\},\{C_{n-1}C_n\}$。如将“山洪预警系统设计”切分为“山洪”、“洪预”、“预警”、“警系”、“系统”、“统设”、“设计”。同样，在查询“山洪预警”一词时，也采用相同的拆分方法，变为“山洪”、“预警”两个词，把它们按“and”的关系构成查询条件，定位到相应位置。双汉字分词的最大优点是简单的算法实现，且免去了词库维护工作，缺点是其冗余大、搜索效率不高。

1.3 最大长度匹配算法

目前比较常用的分词算法是最大匹配法（又称 MM 法）。MM 法是一种应用广泛的分词方法，在其分词过程中，它只使用一个分词词库[5]。MM 法的基本思想如下：设词典中最长的词为 K 个字，则每次均从句子开始位置起取一个长度为 K 的字串，依次将它与词库中的词匹配，若词库中确有这样一个字长为 K 的词，则成功匹配，就把这个字串作为一个词从句子切分出去。接着找到句子剩余部分的起始位置，依上法同样取另一个字长为 K 的字串，重复以上方法，直到把句子切分完为止。当词库中找不到一个匹配当前字串的词条，则应删掉该字串尾部一个字，另生成一个 $K-1$ 字长的字串，再到词库中查询。如成功匹配，把该字长为 $K-1$ 的字串作为一个词，从句子中切分出去；如匹配失败，重复以上步骤，把该 $K-1$ 字长的字串尾部再删去一个字，生成一个 $K-2$ 的字串去词库中匹配，直到成功匹配[6]。

假设一个需要进行切分的字符串：$S=C_1C_2C_3C_4\cdots C_{n-1}C_n$，其中 $C_i(i=1,2,3,\cdots,n)$ 代表单个汉字字符，字典中最长的词字长为 i，以 i 字长进行第一次切分，如成功匹配，结果如下：

$$S=S_1+S_2,S_1=\{C_1\cdots C_i\},S_2=\{C_{i+1}\cdots C_n\}$$

如匹配失败，则以 $i-1$ 字长进行再一次匹配，如匹配成功，则其结果如下：

$$S=S_1+S_2,S_1=\{C_1\cdots C_{i-1}\},S_2=\{C_i\cdots C_n\}$$

以此类推，将 S_1 切分出来后，再同样递归处理 S_2。

最大匹配算法的特点是“长词优先”，与普通匹配相比，可以很好地提高检索速率[7]。与正向最大长度匹配算法的取词与步骤相同，只是从字符串的最后一个自开始向前取词匹配，这种匹配算法叫做逆向最大长度匹配算法，既 RMM。RMM 匹配方式现阶段也越来越受到人们在设计分词方式时的青睐，因其在一些特殊检索中的分词可以达到更为精确的目的[8]。

2 最大长度匹配算法的改进

在普通最大长度匹配算法中，无论是正向还是逆向，都是在开始的局部范围内的最大匹配，即每个句子的前 M 个字符或者后 M 个字符进行匹配。在短句当中，这样的匹配方式还可以很好地解决问题，但是如果待匹配的句子长度远远大于最大匹配词的长度，那么这种匹配方式的弊端就显而易见了，只是局部的匹配不能完全囊括整个句子，使得在句子中间的词汇没有匹配成功，从而造成检索中的漏词和检索不完全[9]。

针对单独的正向或是逆向最大匹配算法存在的问题，在垂直搜索系统中，充分利用专业类环境，在库中建立专业类词库，首先根据词库中的专有名词的最大长度来确定 MAX_Length 的值，解决了匹配算法中盲目选择最大长度的问题，并通过正向与逆向最大匹配算法结合的方式来构成正反双向匹配算法，让正反最大匹配算法可以在匹配过程中弥补各自的不足，从很大程度上提高检索的精确度。在这里，在提出双向匹配算法的同时，针对正向和逆向的匹配过程中的分词歧义和匹配不全问题，对最大长度的匹配算法提出改进，提出“窗口”匹配的思想来进一步提升分词匹配效果[10]。

2.1 正向最大匹配算法的改进

对于正向最大匹配算法的改进：

Step1：假设一个待匹配的句子 S 的长度为 n，即 $S=C_1C_2C_3C_4\cdots C_{n-1}C_n$，其中 $C_i(i=1,2,3,\cdots,n)$ 为单个的汉字字符。如果 n 的值小于等于 1，则分词匹配结束，直接跳到 Step7；当 n 大于 1 时，进入 Step2；

Step2：从库中查找最大长度匹配词的值 $M=$MAX_Length，“窗口”的起始位置为待匹配句子的第一个字，记为 $j=0$；

Step3：当 $n-j$ 大于等于 M 且 M 大于等于 1 时，则从 j 的位置开始向后截取长度为 M 的子串进行匹配，若匹配不成功则进入 Step4，若匹配成功则进入 Step5；当 $n-j$ 小于 M 且 M 大于等于 1 时，进入 Step6；当 M 小于 1 时，进入 Step7；

Step4：$j=j+1$，进入 Step3；

Step5：将匹配成功的分词取出，$j=j+M$，进入 Step3；

Step6：M 值减 1，$j=0$，进入 Step3；

Step7：匹配结束。

可以得到改进的正向最大匹配算法的流程图如图 1 所示。

2.2 逆向最大匹配算法的改进

对于逆向最大匹配算法的改进：

原始的 RMM 算法中，其匹配方式和正向匹配方式基本相同，只是在匹配行进的方向上互逆，这一点不同已经可以在匹配过程中形成较大的结果差异。而在改进的匹配算法中，由于窗口的加入，逆向匹配方式的一个难点就是窗口起点的定位和挪动方向的控制，这

也是与正向匹配方式相比较下的重要差别之一[11]。

逆向最大匹配算法描述如下:

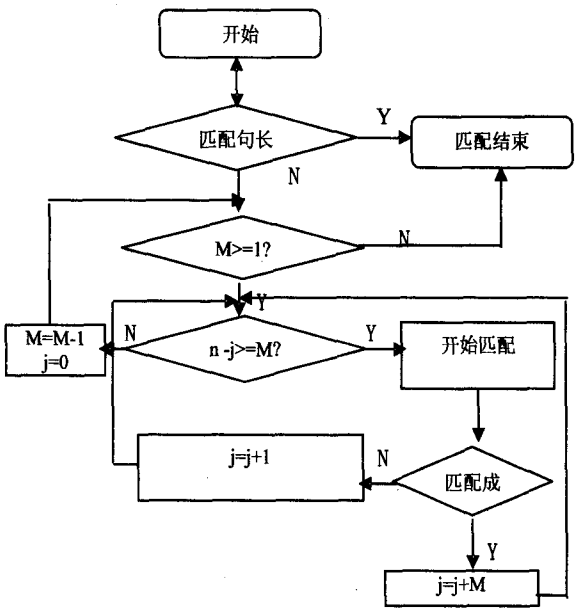

图 1　正向最大匹配算法的改进流程图

Step1:假设一个待匹配的句子 S 的长度为 n,即 $S = C_1C_2C_3C_4\cdots C_{n-1}C_n$,其中 $C_i(i=1,2,3,\cdots,n)$ 为单个的汉字字符。如果 n 的值小于等于 1,则分词匹配结束,直接跳到 Step7;当 n 大于 1 时,进入 Step2;

Step2:从库中查找最大长度匹配词的值 M = MAX_Length,"窗口"的起始位置为待匹配句子的最后一个字,记为 $n-p$,$p=0$;

Step3:若 $n-p$ 小于 0,则匹配结束,直接进入 Step7;当 p 大于等于 M 且 M 大于等于 1 时,则从 $n-p$ 的位置开始向前截取长度为 M 的子串进行匹配,若匹配不成功则进入 Step4,若匹配成功则进入 Step5;当 p 小于 M 且 M 大于等于 1 时,进入 Step6;当 M 小于 1 时,进入 Step7;

Step4:$p=p+1$,进入 Step3;

Step5:将匹配成功的分词取出,$p=p+M$,进入 Step3;

Step6:M 值减 1,$p=0$,进入 Step3;

Step7:匹配结束。

逆向最大匹配算法的改进流程图如图 2 所示。

3　最大长度匹配算法的改进效果

经过上一节,文中已详细分析了正向与逆向的最大长度匹配算法的匹配特点,以及改进的最大长度匹配算法的改进后的优势,下面以几个句子的分词结果为具体例子来说明改进后的最大长度匹配在分词中的改进效果。

句子 1:"发展中国家人民"。

句子 2:"大学生活动点"。

当使用 MM 算法进行分词的时候,句子 1 可以得出结果"发展/中国/家人/民",而句子 2 可得出结果"大学生/活动/点"。当使用 RMM 算法时,句子 1 可以得出结果"发展中/国家/人民",而句子 2 可以得出结果"大学/生活/动点"。在这种情况下,显然人脑可以清楚地辨别出,句子 1 用 RMM 算法做出的分词结果正确,句子 2 用 MM 算法得出结果正确。

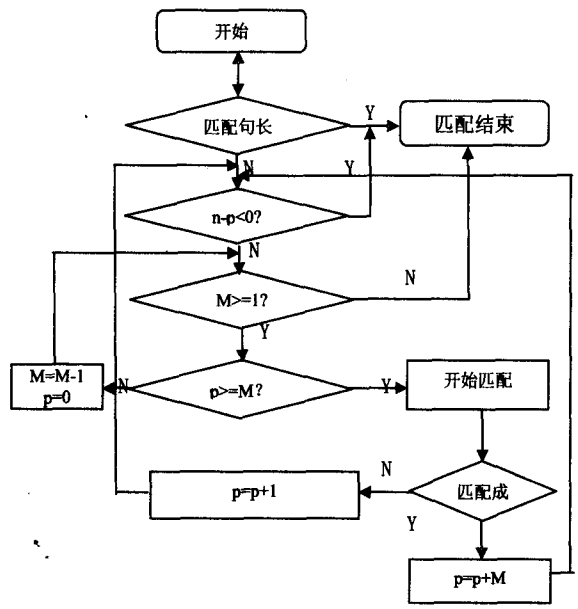

图 2　逆向最大匹配算法的改进流程图

在这里即可以得知,当单独使用 MM 算法或者 RMM 算法进行分词的时候,若这两类句子同时在分词工作中出现,则必然会有一类不能做出正确的分词结果。而此时若用正向和逆向结合的最大长度匹配算法来分词,则可以在对照专业词库的同时找出哪一种匹配方式最为合适,从而得出正确的分词结果,即在匹配中判断出句子 1 用 RMM 算法和句子 2 用 MM 算法。

而对于"窗口"式匹配思想的特点和优势,这里在第三个典型的句子来举例说明:

句子 3:"重点加强天然气运输工程管理"。

根据上面的句子特点,用普通的正向或者逆向最大长度匹配算法进行分词以后都只可以得到一种结果为"重点/加强/天然气/运输/工程/管理"。而在气田地面工程设计中,"天然气运输工程"本身作为一个专业词汇,是石油行业研究的重点。若将这一个词切分成"天然气/运输/工程"来分别匹配,即不能达到满足用户期望的检索结果的目的。显然在这个时候,即使用普通最大长度匹配算法,无论是正向还是逆向都不能做出最好的分词结果。那么采用"窗口"匹配方式,当 M_Length 值 M 取值为 7,且窗口位置为 $j=4$ 时,因库中含有该专有名词并做相应正确匹配。可正确切分

(下转第 137 页)

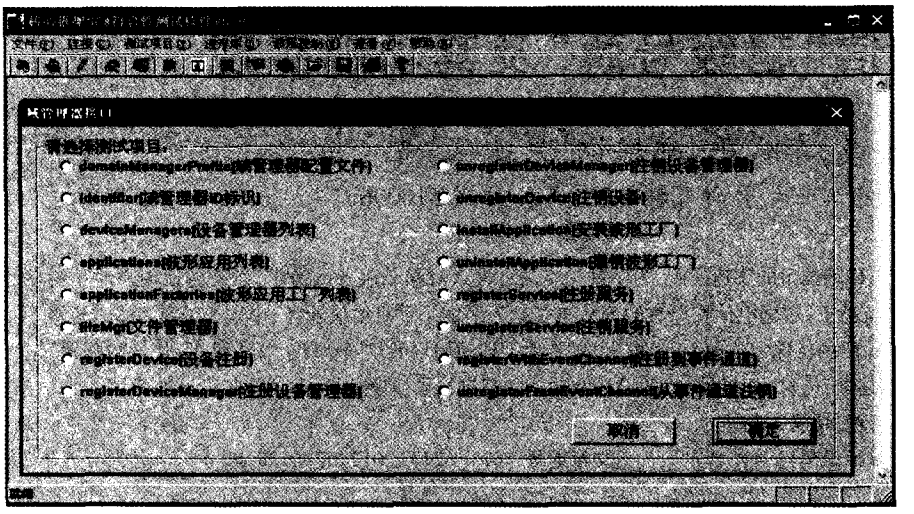

图 5　域管理器接口测试界面

Joint Program Office. Software Communication Architecture Specification[S]. USA:JTRS Joint Program Office,2006.

[2] Object Management Group. The common object request broker:architecture and specification[S]. USA:Object Management Group,2000.

[3] 赵秋明,林志堂,杨莹莹,等. 基于 SCA 的小型化核心框架研究[J]. 计算机测量与控制,2010,18(5):1145-1147.

[4] Schmidt D C,Kuhns F. An Overview of the Real-time CORBA Specification[J]. IEEE Computer,2000,33(6):56-63.

[5] Henning M,Vinoski S. 基于 C++ CORBA 高级编程[M]. 徐金梧,徐　科,吕志民,译. 北京:清华大学出版社,2000.

[6] 代　霞,黄劲松. 基于 CORBA 综合网络配置管理的设计与实现[J]. 计算机技术与发展,2008,18(2):91-93.

[7] 李　振,曹谢东,刘世齐. 基于 CORBA 的油气田异构信息系统多源集成[J]. 计算机技术与发展,2006,16(6):60-62.

[8] 李雅萍,杨尚森,李　阳. CORBA 技术在 SCA 系统中的应用[J]. 计算机工程与设计,2008,29(16):4200-4203.

[9] 李　方,张　虹. GIOP 协议和 CORBA 的性能优化[J]. 微计算机信息,2006,22(7):7-10.

[10] 祁明龙,阚文第,杨　俊. CORBA 与 DCOM 桥接设计与实现[J]. 计算机技术与发展,2008,18(5):105-107.

[11] 洪锡军,刘献科,张　激. 基于 SCA 的无线通信技术研究[J]. 计算机工程,2005,31(8):120-122.

[12] 朱其亮,郑　斌. CORBA 原理及应用[M]. 北京:北京邮电大学出版社,2001.

(上接第 133 页)

出“天然气运输工程”,得出分词结果“重点/加强/天然气运输工程/管理”。在这里可以看出,改进的最大长度匹配算法可以遵从“长词优先”的要求,并针对石油行业的专业要求进行切分,从而得出正确的分词结果[12]。

4　结束语

文中重点针对分词过程中的最大长度匹配算法进行研究分析,讨论 MM 和 RMM 算法的特点与弊端,提出改进的“窗口”式最大长度匹配算法,并详细介绍算法过程,并且与普通最大长度匹配算法相比较并阐明其优势,通过比较可知,改进后的算法提高了搜索的效率。

参考文献:

[1] 张翠英,亢临生. 三字歧义链自动分词方法[J]. 情报学报,1998,17(3):203-207.

[2] 黄春毅. 一种自适应搜索引擎的构建研究[J]. 情报检索,2006(2):163-164.

[3] 吴右昕. 网络搜索引擎的发展趋势分析[J]. 当代传播,2007(3):73-74.

[4] 郭祥昊,钟义信,杨　丽. 基于两字词簇的汉语快速自动分词算法[J]. 情报学报,1998,17(5):352-357.

[5] 袁占亭,张爱民,张秋余. 基于概念的 web 信息检索[J]. 计算机工程与应用,2003,39(36):173-181.

[6] 赵诗阳. DGP 系统中基于库的垂直检索技术的优化[D]. 成都:四川大学,2011.

[7] Camarilla M J,Etzioni O. A search engine for natural language applications[C]//Proc of the 14th International Conference on World Wide Web. New York:ACM,2005.

[8] Bemers-Lee T,Hendler J,Lassila O. The Semantic Web[J]. Scientific American,2001,284(5):34-43.

[9] Vailaya A,Figueiredo A T,Jain A K,et al. Image classification for content-based indexing[J]. IEEE Transactions on Image Processing,2001,10(1):117-130.

[10] 向　晖. 基于 Lucene 的中文字典分词模块的设计与实现[J]. 现代图书情报技术,2006(8):46-50.

[11] 郭　辉,苏中义,王　文,等. 一种改进的 MM 分词算法[J]. 微型电脑应用,2002,18(1):13-15.

[12] 赵曾贻,陈天娥,朱　兰. 一种基于语词的分词方法[J]. 苏州大学学报(自然科学),2002,18(3):44-48.