中文信息检索引擎中的若干技术

吴栋　　滕育平

(南开大学组合数学研究中心 核心数学与组合数学教育部重点实验室, 天津 300071)

摘　要　本文论述了在开发中文信息检索系统中所涉及到的两项关键技术，即中文分词技术和检索技术。对中文分词技术，本文介绍了一种改进的正向最大匹配切分算法，以及为消除歧义引入的校正策略，并在此基础上结合统计方法处理未登录词。针对检索技术，本文综述了几种最常用的检索模型的原理，并对每种模型的优缺点进行了简要分析。最后对给出的分词算法进行了测试，测试表明本文给出的分词算法准确度和效率能够满足实用的要求。

关键词　信息检索　搜索引擎　分词技术　检索技术

1 引言

随着社会的不断进步，特别是在互联网迅猛发展的今天，人们在不断地接触形形色色的信息，同时也要对这些信息进行过滤，从而提取出对自己真正有用的内容。为了达到这个目的，人们开发出了众多的检索引擎，有针对Web进行搜索的Goolge、百度等，也有针对各行业开发的专题检索系统。目前，国内的每个行业、领域都在飞速发展，这中间产生了大量的中文信息资源，为了能够及时准确的获取最新的信息，中文检索引擎是必然的产物。中文检索引擎与西文检索引擎在实现的机制和原理上大致雷同，但由于汉语本身的特点，必须引入对于中文语言的处理技术，而中文分词技术就是其中很关键的部分。

2 中文检索引擎的基本原理

常见的中文检索引擎主要完成两方面的任务：

1．信息的规范化。将搜集来的信息按照一定的方式进行组织管理，使之成为可以高效检索的信息库。

2．信息的检索和表达。以索引好的信息库作为信息基础，利用信息库已被索引的特点，实施快速检索，同时根据用户的需求将检索结果进行输出。

其中，信息的规范化包括分词和索引（以及资料的搜集和整理）、更新（维护）两部分；信息的检索包括搜索、结果输出两部分。整个信息处理和检索过程如图1所示：

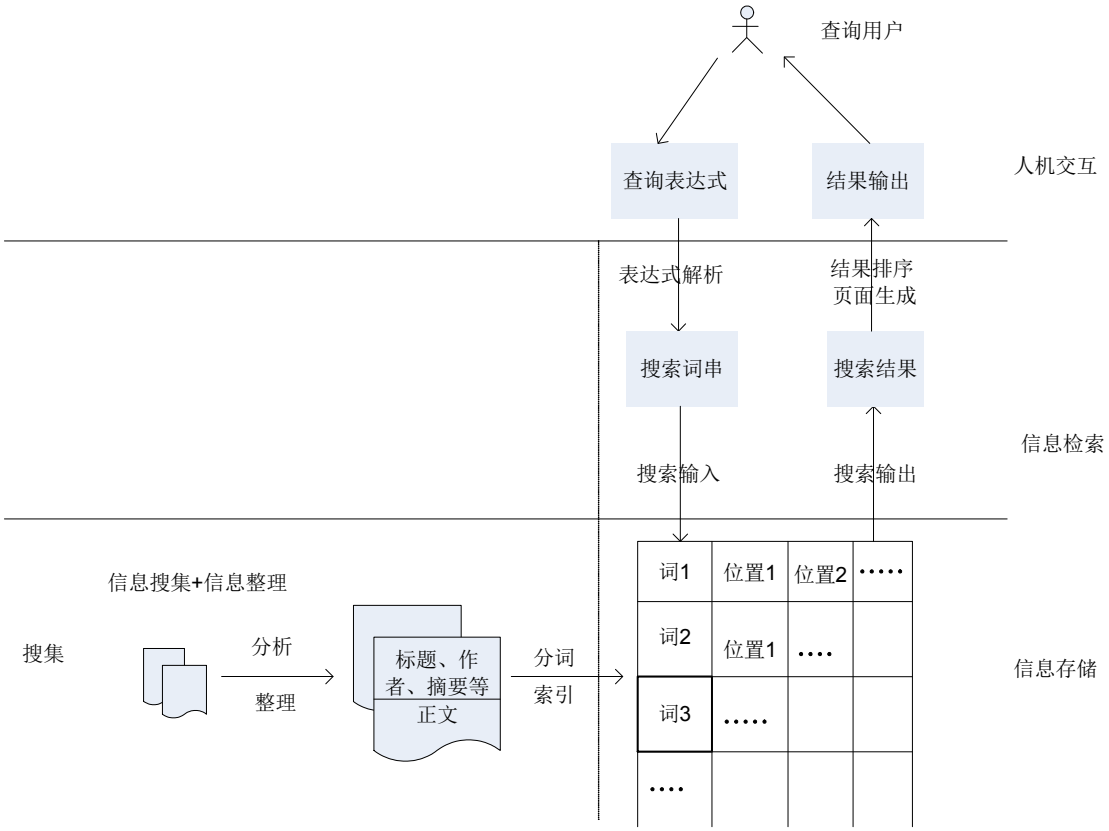


图1 中文信息处理和检索过程

3 中文分词技术

3.1 汉语的特点

词是最小的、能独立活动的、有意义的语言成分。因此，通常的检索引擎都是以每一个独立的词为单位建立索引，在查询时按照检索词出现的位置和频率对文档进行输出。英语文本是小字符集上的已充分分隔开的词串,而汉语文本是大字符集上的连续字串，并且在词与词之间并没有明显的分割标记。故而存在一个对汉语中的词加以识别的问题，即中文检索引擎首先必须对原文进行切分词。如果不切词（按字检索）,可能检索的结果与用户的查询要求会大相径庭，例如当检索德国货币单位"马克"时,就会把"马克思"检索出来,而检索"华人"时会把"中华人民共和国"检索出来。因而进行切词,可以大大提高检索的准确率。

中国的汉字是示意文字，总数有几万个，在由国家标准总局颁布的《信息交换用汉字编码字符集--基本集》（即GB2312-80）中共收录了一级和二级常用汉字共 6763 个，而在 Unicode 编码中更是收录多达 20902 个汉字。据统计，在常用汉语中，90%以上使用的是二字词和三字词，也有使用四字词和五字词。知道这些汉字的特点，对于我们选择合理的切分算法是有益的。

3.2 一般的分词技术

由于书面汉语是字的序列，词与词之间没有间隔标记，使得词的界定往往模糊不清。即使这样，在过去的时间里，人们在汉语的自动分词技术的研究上还是做了很多工作，设计了许多实用、高效的算法。通常的方法主要分为两类[1]：第一类主要基于字典、词库的匹配和词的频度统计，这类方法实用、具体，比较容易实现；第二类方法主要基于句法、语法分析，并结合语义分析，通过对上下文内容所提供信息的分析对词进行定界，这类方法试图让机器具有人类的理解能力，其原理较为晦涩，一般不易实现。

常用的切词算法如下：

1)最大正向匹配法(Maximum Matching Method)

通常简称为 MM 法。其基本思想为：设 D 为词典，MAX 表示 D 中的最大词长，str 为待切分的字串。MM 法是每次从 str 中取长度为 MAX 的子串与 D 中的词进行匹配。若成功，则该子串为词，指针后移 MAX 个汉字后继续匹配，否则子串逐次减一进行匹配。

2)逆向最大匹配法(Reverse Maximum Matcing Method)

通常简称为 RMM 法。RMM 法的基本原理与 MM 法相同，不同的是分词的扫描方向，它是从右至左取子串进行匹配。统计结果表明，单纯使用正向最大匹配的错误率为 1/169，单纯使用逆向最大匹配的错误率为 1/245，RMM 法在切分的准确率上比 MM 法有很大提高。

3)基于词频的统计方法

统计方法一般不依赖于词典，而是将原文中任意前后紧邻的两个字作为一个词进行出现频率的统计，出现的次数越高，成为一个词的可能性也就越大。在频率超过某个预先设定得阈值时，就将其作为一个词进行索引。这种方法能够有效地提取出未登录词。

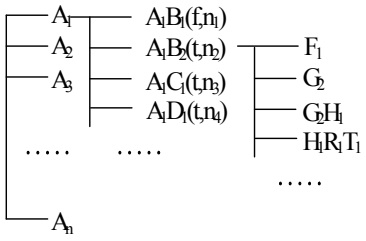
3.3 一种改进的 MM 算法

MM 法和 RMM 法的缺点在于对词典的完全性有很强的依赖性，而且无法很好的解决歧义问题，有人提出了双向匹配法，即针对一个字符串，分别从两个方向进行处理，但这种方法只有检错功能，却不能自动进行校正，给出正确结果。由于一个词在不同的文章中出现的次数通常不一样，因此采用统计方法对词的切分准确度并不太高。

鉴于以上几种方法的优缺点，人们自然想把这几种方法结合起来，扬长避短。这里，介绍一种改进的 MM 算法。

3.3.1 词典存储格式

采用分层存储的形式，一共分为 3 层，形成树型结构，如下所示（每一个字母代表一个字）。



一层存储所有单字。第二层保存所有的双字词和多字词的前两个字（因为，也许会出现 ABC 为词，但 AB 不是词的情况），并对两者做不同标记(t/f)。每一个可成词的单字对应一系列第二层结点，用来存储所有以该字为词首的双字（包括上述两种情况）。并且，在这里，针对每一个双字，需要记录以该双字为词首的所有词的最大长度，实际中，可以保存除去该双字部分的最大长度(记为 n)。第三层存储以某一双字为首的所有词。为了减少存储空间，只存储除去该双字以外的部分（如上图所示）。每一层各结点需按某种次序排列，可使用 hash、二分查找等方法进行查询。采用这种层次的存储结构，可以很快把查询词的工作缩小到一个很小的范围内，有利于分词效率的提高。

3.3.2 匹配方法(MM 方法)

由于词库中的最大词长通常大于所切分出的词长，为了提高切分的效率，不采用逐次减一个字的方法，而是使用正向逐一增长的方法。

假设对一个句子 C_1C_2……进行分词处理，算法描述

如下：

1) 两个字（开始时为 C_1C_2），在词典中查询 C_1C_2 是否存在

2) 不存在，则 C_1 为单字词，一次分词结束，返回 1。

3) 存在，判断 C_1C_2 是否为词，并从词典中获取该词下层节点汉字的最大长度，设为 n

4) 若 n=0，一次分词结束，保存结果。

5) 否则，i=2，转 6)。

6) i=i+1，若 i=n+3，转 8)；否则，转 7)。

7) 再取一个字（此处为 C_i），判断第三层中是否有以 $C_3\cdots\cdots C_i$ 开始的字(不需要恰好匹配，只要匹配开始的 i 个字就可以了)。

8) 若存在，分词结束，返回最近一次能够恰好匹配的 $C_3\cdots\cdots C_j(j<i)$，并与 C_1C_2 组合成词。如果是 C_1C_2，则根据 C_1C_2 的标记判断是双字词还是分为两个单字词。

9) 否则，转 6)。

3.3.3 歧义词处理

汉语中的歧义结构主要有两种：交集型歧义和组合型歧义。据统计，汉语中的交集型歧义字段约占全部歧义字段的 90%。所以，处理好交集歧义字段在很大程度上能保证一定的分词精度。鉴于汉语中多数的词组、短语为偏正结构，中心词在后，而修饰词在前，故而在进行歧义校正时，我们让交集歧义字优先与右边的子段组成词，而其余的字段则尽可能的向左组词。

设 $C_1C_2\cdots\cdots C_n$ 是连续型交叉歧义字段，具体的歧义校正策略如下：

A．主导策略

1) 指针移向 C_n，调用分词算法对以 C_n 为首字的词进行查找。

2) 若句子中 C_n 可以和后面的字构成词（设 $C_n\cdots\cdots C_m$ 为构成的最长词），则对 C_n 进行标记。

3) 移向 C_m，继续对 C_m 进行处理，方法类似于 2)，直到找到没有歧异的词为止。

4) 不妨设 C_m 与其后的字不成词，此时让 C_n 优先与右边的子段组成词，即切分 $C_n\cdots\cdots C_m$ 为一词。

5) 对 C_n 之前的部分做最大正向匹配，歧义处理结束。

B．辅助策略

在汉语中许多字是多义字，由于上下文环境的不同，这些字既可以作为只具语法意义或功能意义的虚词，也可以与其他字组合构成实词，如“的”、“地”、“了”等。统计结果表明，当这些字作为虚词时，通常作为词的尾字出现，而构成实词时，往往出现在词的首位，或中间部位，所以对这些字如果直接采用主导策略，往往会造成切分错误。因此，我们对这些字引入辅助策略。

1) 在使用主导策略第一步时，判断 C_n 是否是上述的多义字

2) 若是，且 C_n 是某个词的词尾字，同时 C_n 无法与其后的字构成词，此时将 C_n 视为虚词，并作为单独一个词进行切分，而对 C_n 之前的部分做最大正向匹配。

3) 否则，继续采用主导策略。

3.3.4 统计方法运用

由于词典的不完全性，许多词可能不会在字典中登录，为了处理句子中的未登录词，我们在原有的算法中嵌入词频统计方法，将某些出现频率较高的连续字段作为一个词切分，我们首先对频度设定一个阈值 f。

设已对 $C_1\cdots\cdots C_n$ 进行切分，由切分算法和歧义处理算法得到 $C_1\cdots\cdots C_i$ 为一个词，$C_j\cdots\cdots C_n$ 为一个词，C_i 与 C_j 之间皆为单字词，即 $C_1\cdots\cdots C_i$ 和 $C_j\cdots\cdots C_n$ 是相邻最近的两个多字词，则将 $C_{i+1}\cdots\cdots C_{j-1}$ 作为一个多字词进行词频统计，在对文章全部切分完毕之后，若 $C_{i+1}\cdots\cdots C_{j-1}$ 的出现次数达到 f 时，则将其看作一个词，否则，将其拆分为单字词。

同时，对于相同或相近专业和领域建立起动态词库，将由统计得到的词不断加入词库中，可以实现对词典的动态维护。

通过将基于词典的处理方法和基于频率的统计方法结合起来，不仅保证了切分速度快、精度高的优点，而且能够结合上下文，最大限度的识别人名、地名、专业术语等未登录词。

4 检索技术

根据查找相关信息的实现方式不同，常见的信息检索引擎有布尔逻辑模型、模糊逻辑模型、向量空间模型和概率检索模型等几类。

4.1 布尔逻辑模型

布尔逻辑模型是最简单的检索模型，也是其他检索模型的基础。

设文本集 $D=(d_1,d_2,d_3,\cdots,d_n)$,$d_i(i=1,2,\cdots,n)$为文本集中某一文档;又设 $T_i=(t_{i1},t_{i2},\cdots,t_{im})$为 di 的标引词集合,则对于形如 $Q=W_1\wedge W_2\wedge\cdots\wedge W_k$ 的检索式,如果 $W_1\in T_i,W_2\in T_i,\cdots,W_k\in T_i$,则 d_i 为查询 Q 的命中文档,否则 d_i 为 Q 的不命中文档;而对于形如 $Q=W_1\vee W_2\vee\cdots\vee W_k$ 的检索式,如果至少存在某个 $W_j\in T_i(j=1,2,\cdots,k)$,则 d_i 为 Q 的命中文档,否则 d_i 为不命中文档。

用户根据所检索关键字在检索结果中的逻辑关系递交查询，查询模块根据布尔逻辑的基本运算法则来给出查询结果。

布尔检索模型原理简单易理解，容易在计算机上实现并且具有检索速度快的优点。但是最终给出的查询结果没

有相关性排序，不能全面反映用户的需求，功能不如其他的检索模型。

4.2 模糊逻辑模型

模糊逻辑模型以模糊数学作为理论基础，设置单个的检索词 w 在文档 d 中的隶属度 u，u∈[0,1]，u 越大代表 w 和文档 d 的相关性越高。用户给出查询要求，查询模块根据模糊逻辑运算给出查询的结果，并能够按照相关度排序。

模糊逻辑模型能够克服布尔逻辑模型检索结果的无序性，但是给查询词设置准确的隶属度有一定困难。

4.3 向量空间模型

向量空间模型[4]将文档映射为一个特征向量 $V(d)=(t_1, \omega_1(d); \cdots; t_n, \omega_n(d))$，其中 $t_i(i=1,2, \cdots,n)$为一列互不雷同的词条项，$\omega_i(d)$为 t_i 在 d 中的权值，一般被定义为 t_i 在 d 中出现频率 $tf_i(d)$ 的函数，即 $\omega_i(d)=\psi(tf_i(d))$。在信息检索中常用的词条权值计算方法为 TF-IDF 函数 $\psi = tf_i(d) \times \log(\frac{N}{n_i})$，其中 N 为所有文档的数目，$n_i$ 为含有词条 t_i 的文档数目。TF-IDF 公式有很多变种，下面是一个常用的 TF-IDF 公式：

$$\omega_i(d) = \frac{tf_i(d)\log(\frac{N}{n_i}+0.1)}{\sqrt{\sum_{i=1}^{n}(tf_i(d))^2 \times \log^2(\frac{N}{n_i}+0.1)}}$$

根据 TF-IDF 公式，文档集中包含某一词条的文档越多，说明它区分文档类别属性的能力越低，其权值越小；另一方面，某一文档中某一词条出现的频率越高，说明它区分文档内容属性的能力越强，其权值越大。

两文档之间的相似度可以用其对应的向量之间的夹角余弦来表示，即文档 d_i，d_j 的相似度可以表示为

$$Sim(d_i,d_j) = \cos\theta = \frac{\sum_{k=1}^{n}\omega_k(d_i)\times\omega_k(d_j)}{\sqrt{(\sum_{k=1}^{n}\omega_k^2(d_i))(\sum_{k=1}^{n}\omega_k^2(d_j))}}$$

进行查询的过程中，先将查询条件 Q 进行向量化，主要依据布尔模型:

当 t_i 在查询条件 Q 中时，将对应的第 i 坐标置为 1，否则置为 0，即

$$q_i = \begin{cases} 1 & t_i \in Q \\ 0 & t_i \notin Q \end{cases}$$

从而文档 d 与查询 Q 的相似度为

$$Sim(Q,d) = \frac{\sum_{i=1}^{n}\omega_i(d)\times q_i}{\sqrt{(\sum_{i=1}^{n}\omega_i^2(d))(\sum_{i=1}^{n}q_i^2)}}$$

根据文档之间的相似度，结合机器学习的一些算法如神经网络算法，K-近邻算法和贝叶斯分类算法等，可以将文档集分类划分为一些小的文档子集。

在查询过程中，可以计算出每个文档与查询的相似度，进而可以根据相似度的大小，将查询的结果进行排序。

向量空间模型可以实现文档的自动分类和对查询结果的相似度排序，能够有效提高检索效率；它的缺点是相似度的计算量大，当有新文档加入时，则必须重新计算词的权值。

4.4 概率检索模型

概率检索模型是在布尔逻辑模型的基础上为解决检索中存在的一些不确定性而引入的。概率检索模型有多种形式，常见的为第二概率检索模型,首先设定标引词的概率值，一般是对检索作业重复若干次,每一次检索用户对检出文档进行相关性判断。再利用这种反馈信息,根据每个词在相关文档集合和无关文档集合的分布情况来计算它们的相关概率,将词的权值设计为:

$$\log\frac{p(1-p)}{p'(1-p')}$$

其中 P,P′ 分别表示某词在相关文档集和无关文档集中出现的概率。某一文档的权值则是它所含的标引词权值之和,于是,文档 d 与用户查询 Q 相关概率可定义为:

$$S(d,Q) = \sum \log\frac{p_w(1-p_w)}{p'_w(1-p'_w)}$$

其中 p_w 和 p_w'分别为 w 在相关文档和无关文档中的概率。上式中右边和式是对所有出现在文档 d 和查询 Q 中的词 w 求和，即 w∈d∩Q.

概率模型有严格的数学理论基础，采用了相关反馈原理克服不确定性推理的缺点，它的缺点是参数估计的难度比较大，文件和查询的表达也比较困难。

以上介绍了几种传统的检索模型，随着检索技术的不断发展，新的检索技术也不断涌现，出现了诸如并行信息检索系统、演绎信息检索系统、基于超文本技术的信息检索系统、分布式检索系统和智能检索系统等[5]。这些新的技术代表了检索技术的发展方向。

5 实验结果

我们对设计的切分算法作了程序上的实现，采用的语料库来自由北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作的 PFR 人民日报标注语料库（版本 1.0）。

本文在以下环境中实现了切分算法：

CPU	内存	操作系统	开发环境
P4　1.5G	256M	Win2000	VC++6.0

切分结果：

文件大小	汉字（个）	用时（秒）	切分准确率
3.55 MB	1839414	483.544	91．57%

统计结果：

统计词数（个）	人　名（个）	地名（个）	其它有意义的词（个）	有效率
1996	622	90	323	51．85%

结果分析：

我们的词典总共收录了 130152 个词，基本上覆盖了常用词汇。切分结果表明采用一个比较完全的词典，再配合以快速的切分算法和适当的校正策略，从而使得无论是切分效率还是切分正确率，都是令人满意的。根据词频统计出的结果中，也有很大一部分是有意义的词汇，说明用统计方法处理未登录词包括人名和地名也是有效的。最后，我们还将经过统计得到的有意义的词加入词典进行再次切分，得到的准确率为 92.34%，比原结果提高了 0.77%，可见在原有的切分算法上再辅助以统计方法，可以有效提高切分的准确度。

6 结束语

要开发高性能的中文检索引擎，快速、可靠的中文分词算法和准确、高效的检索技术是至关重要的，针对不同领域和需求，需要采取不同的策略和方法，本文仅起抛砖引玉的作用。随着科学技术的发展，人们必然需要针对性更强的中文检索引擎，因此，专业化、 深层次的中文检索引擎将是今后的发展方向。

[参考文献]

[1]严威, 赵政. 开发中文搜索引擎汉语处理的关键技术. 计算机工程　Vol .25, No.61, 1999, pp5-6.

[2]姚天顺, 朱靖波等. 自然语言理解（第 2 版）. 北京：清华大学出版社, 2002.

[3]Tom M. Mitchell. 机器学习. 曾华军, 张银奎等译.北京：机械工业出版社，2003.

[4]G.Salton, A.Wong, C.S.Yang. On the specification of term values in automatic indexing. Journal of Documentation, 1973 , 29(4):351～372.

[5]贾同兴. 人工智能与情报检索.北京：北京图书馆出版社, 1997.7.

Some Techniques for Information Search Engines for Chinese

WU Dong　　TENG Yu-ping

(Center for Combinatorics, Laboratory of Pure Mathematics and Combinatorics , Nankai University, Tianjin 300071, P.R. China)

Abstract. Two key techniques in the development of Chinese Information Retrieval System are discussed in this paper, i.e., Chinese word segmentation and search technique. For Chinese word segmentation, the paper presents an improved MM segmentation algorithm, the revise strategy for disambiguation, and the statistic method for unknown words recognition based on the previous methods. For search technique, the paper summarizes the principle of several kinds of search models, and analyzes the advantages and disadvantages of each model simply. At last, the given segmentation algorithm is evaluated, and the results reveal that the veracity and efficiency of the algorithm can satisfy the applied request.

Key words.　information retrieval , search engine, word segmentation, search technique

作者简介 吴栋（1980-），男，上海人，博士研究生，主要研究方向：组合数学；　滕育平　（1980-），男，湖北孝感人，硕士研究生，主要研究方向：组合数学与计算机软件.

第一作者	第二作者				联系地址	邮编	省市	电话	E-mail	稿件名称
吴栋	滕育平				南开大学组合数学研究中心	300071	天津	022-23508038	tengyuping@notionsoft.com	中文信息检索引擎中的若干技术