文章编号:1001 - 9081(2010)07 - 1941 - 03

面向 GIS 基于专有名词优先的中文分词方法

罗　浩[1],魏祖宽[1],金在弘[2]
(1. 电子科技大学 计算机科学与工程学院 成都 610054; 2. 永同大学校 计算机工学科,韩国 永同郡 370701)
(lhao77@gmail.com)

摘　要:提出了一种面向地理信息系统领域的基于专有名词优先的中文分词方法:利用专业词典、通用词典和同义词词典相结合的词典机制,优先切分专有名词,对粗分结果利用 Trigram 模型进行消歧而获取最终结果。实验证明,该分词算法对专业文献的分词处理具有较好速度和准确性。

关键词:中文分词;专业词典;Trigram 模型;同义词词典
中图分类号: TP391　　**文献标志码**:A

Chinese word segmentation for GIS based on priority special name

LUO Hao[1], WEI Zu-kuan[1], KIM Jae-hong[2]
(1. *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan* 610054, *China*;
2. *Department of Computer and Information Engineering, YoungDong University, Yongdong-Gun Chungbuk* 370701, *Korea*)

Abstract: A Chinese word segmentation algorithm for Geographic Information System (GIS) based on priority special name was designed: use dictionary mechanism which combines synonyms dictionary, general dictionary and special dictionary, cut the sentences by special name firstly, and get the segmentation result of disambiguating with Trigram mode lastly. The experimental results show that the segmentation algorithm has good speed and accuracy in segmentation processing of professional literature.

Key words: Chinese word segmentation; special dictionary; Trigram mode; synonyms dictionary

0　引言

当前,搜索引擎被广泛使用,是互联网领域中仅次于电子邮件的应用。搜索引擎主要分为三部分:信息采集、信息索引和信息检索。中文分词技术作为信息索引和信息检索的关键技术,首先要能够准确快速地分词,此外与搜索引擎其他模块之间良好的兼容也至关重要[1]。所谓中文分词,即把中文汉字序列切分成有意义的词。目前各种分词算法非常多[2],大致可以分为以下三类:基于字符串匹配的分词、基于理解的分词和基于统计的分词[3]。基于字符串匹配的分词,又称为机械式分词或基于字典的分词,是按照一定策略将待分的汉字语句与一个充分大的词典中的词条进行匹配。若在词典中找到该字符串,则匹配成功。机械式分词的三要素是:分词词典、文本扫描顺序和匹配原则,其主要优点是简单、易于实现;其主要缺点是容易产生歧义切分,新词识别能力弱。基于理解的分词又称基于人工智能的分词,是指通过句法和语义的分析,利用句法信息和语义信息来处理歧义现象。它主要包括三个部分:分词子系统、句法语义子系统和总控部分。这种分词方法模拟了人对句子的理解过程,它需要使用大量的语言知识和信息。目前基于理解的分词系统尚不成熟。基于统计的分词方法,又叫无词典分词法,根据字符串在语料库中出现的统计频率来决定其是否构成词。其主要优点是能有效地实现上下文识别生词、自动消除歧义;其主要缺点是时间空间开销大,训练时间长,对长词识别能力差,不能从根本上消歧。

在一个成熟可靠的中文分词系统中,只使用某一种分词方法是不可行的,需要充分结合各种方法来解决实际问题。本文在对现有中文分词技术研究的基础上,结合地理信息系统(Geographic Information System,GIS)领域的应用背景,提出了一种改进的中文分词算法并应用在 GIS 领域。它可以对 GIS 领域的中文文档进行高效准确的分词,并通过实验进行了验证。

1　分词系统总体框架

本分词系统由词典组织和分词算法两部分组成。系统结构如图 1 所示。

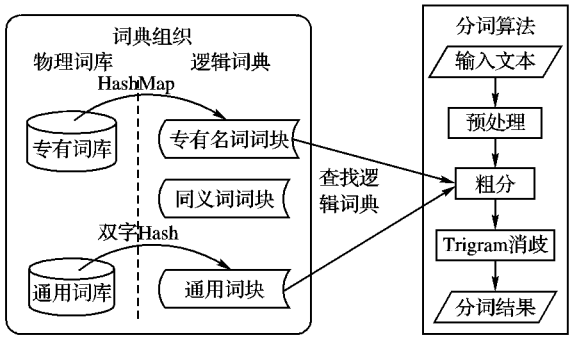

图 1　分词系统总体结构

系统采用基于字典的最大逆向匹配法对输入语句进行粗分,将粗分结果利用 Trigram 模型进行消歧,获取最大概率的切分结果。

收稿日期:2010 - 01 - 20;**修回日期**:2010 - 03 - 05。

作者简介:罗浩(1984 -),男,湖北武穴人,硕士研究生,主要研究方向:搜索引擎技术、空间数据库; 魏祖宽(1968 -),男,四川成都人,副教授,博士,主要研究方向:空间数据库、3G 的应用、数据库技术; 金在弘(1960 -),男,韩国首尔人,副教授,博士,主要研究方向:地理信息系统、数据库技术。

词典组织是中文分词系统的重要组成部分,因为系统在分词过程中会频繁查询分词词典,所以提高对分词词典的查询速度能有效提升分词系统的整体性能。另外,在搜索引擎这个大规模、开放的语言环境中,需要周期性地对词典进行新词登录、词条删除等维护工作,这就要求分词词典能够灵活、快速地更新。本文通过专有词典和通用词典相结合,并对词典的内存数据结构进行设计,很好地完成了这两个目标。

分词算法是中文分词系统的核心。本文在通过匹配专有词典和通用词典获取粗分结果的基础上,通过 Trigram 模型进行消歧,使用 Viterbi 算法求解出最大概率的分词结果,作为最终分词结果。

2 词典组织

目前,大多数开源分词系统都是使用一个通用词典。这种方式有两个缺点:首先,通用词典很大,一般词条数在十万以上,这样会影响分词速度;此外,一些在通用词典中不存在的专有名词无法切分出来,这必然会影响分词准确率。使用专有词典,从待分的字符串中识别出专有名词,以其作为断点就能将原字符串切分成更小的字符串,有利于提升分词效果。

词典组织的两个要素是词条来源和存储结构。本文结合实际应用情况,提出了以下方案。

2.1 词条来源

从分词效率和分词准确性两方面综合考虑,对于词条来源的问题,我们采用了专有词典与通用词典相结合的方式。其中专有词典由三部分组成:人名部分[4],一些业内专家的名字被收录其中;机构名称部分,主要包括一些业内机构名称;专业术语部分,包括当前已知晓专业术语。通用词库主要是中国科学院语言研究所的 ICTCLAS 分词系统中的词库,再添加上从新的训练语料中发现的词,最后过滤出存在于专有名词词库中的词条[5]。与通用词典相比,专有词典的词条数目较小,这样在匹配时与专有词典进行优先匹配能有效提高效率。在专业文献中,专业词汇出现率较高,通过优先匹配专有词典能获取较为准确的粗分结果。

针对通用词典,考虑到长词匹配不成功对分词效率影响较大的问题,例如,词典中有这样一个词条:“爱拼才会赢”。现在有一个句子“爱拼才会胜”,在进行最大匹配时,只有匹配到“胜”才能发现匹配失败,然后继续做四字匹配。本分词系统将通用词典一分为二,把四字以上的词条提出,构成长词词典,四字以下的作为短词词典,匹配时先匹配长词词典,匹配不成功再匹配短词词典。

2.2 存储结构

针对存储结构,我们为通用词典和专有词典选择了不同的方案。考虑到匹配通用词典时效率更加重要的情况,双字词语和三字词语在词库中比例高达 85% 以上的实际,通用词典的存储结构采用双字哈希数组的方式[6]。

词典结构如图 2 所示,分为三部分:首字 Hash 索引、次字 Hash 索引和剩余字串组。

首字 Hash 索引通过使用 Hashtable 实现,其每个单元包括三项内容:关键字为词的第一个汉字;成词否为 bool 型,标示单个首字是否构成词;次字 Hash 索引指针为指向以首字起始的所有词语的第二个汉字的索引。

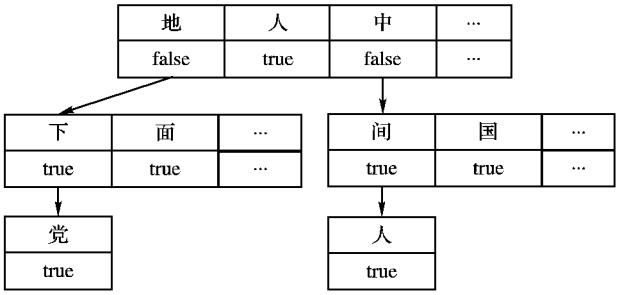

图 2 双字哈希词典结构

次字 Hash 索引与首字 Hash 索引类似,区别在于其关键字是词的第二个汉字(仅对能跟第一个字构成“词或词的前缀”);其第三项的指针是剩余字串组指针,指向以起始双字串起始的所有词语的剩余字串有序数组。

剩余字串组是以起始双字串开始的所有词语剩余字串的有序数组,每个单元包括两项内容:剩余字串,即除去词的起始两字后的剩余部分;成词否,用来标示从首字到对应位置的子串是否也构成一个词,该项必然为 true,目的是与前面两项保持一致,方便处理。

以上描述的双字哈希数组的存储结构很适合逆向最大匹配算法。匹配时,首先从字符串 S 的第 i 位开始查找,匹配 $S[i]S[i+1]$,匹配成功则继续匹配后续字符串,否则转到 S 的第 $i+1$ 位,进行下一次匹配。例如,匹配 S = “以数字地形模型为基础”,假设最大词长为 4,则匹配情形如下:

1)匹配“型为基础”,在首字 Hash 索引中定位到“型”字开始的索引项;

2)由于“型”字成词否选项为假,转到“型”的次字 Hash 索引继续匹配;

3)结果,“型为”并不成词,重新匹配“为基础”;

4)在匹配“为”字的次字 Hash 索引中的索引项时,发现“为基”并不成词,重新匹配“基础”;

5)“基础”能够成词,继续匹配“形模型为”;

6)最终切分结果为“以/数字/地形/模型/为/基础”。

2.3 同义词词库

在匹配专有词典时,多词同义的情况十分普遍。例如,用户在搜索关键词“中科院”时,关键词“中国科学院”的结果也被整合进来更加符合用户意图。为解决这一问题,我们设计了同义词配置模块,其关键在于同义词文件的物理结构,其物理结构如图 3 所示。

行号	同义词1,	同义词2,	…	, 同义词n
1.	阳光	日光	…	太阳光线
2.	诺依曼	冯诺依曼	…	冯·诺依曼
3.	洋人	外国人	…	老外
4.	电脑	计算机	…	PC机
⋮	⋮	⋮		⋮

图 3 同义词文件物理结构

分词系统首先从同义词文件[7]中逐行读取数据,然后保存到一个 HashMap 中,以词条为关键字,行号为值。这样在生成倒排索引时,只要两个词条的值相同,就把它们的词频和 URL 等信息整合在一起。另外,为了保证分词的一致性,专有词典和通用词典中不能出现相同的词条,每次更新词典时

要做全面检查。

3　分词算法

分词算法详细步骤如下所示。

1)编码转换,把输入文本的编码统一转换成 UTF-16 编码,并存放到缓冲区中,供步骤 2)使用。原因是 UTF-16 是定长格式,转换后方便统一处理。

2)统一化,把全角的数字和英文字母统一转换为半角。

3)数字提取,由于数字和汉字不可能组合成词,先把数字提取出来,可以进一步提高分词准确性和效率。

4)英文提取,提取出输入语句中所有英文,原因同上。

5)无效字符处理,一些符号如“%”、“^”等不用处理,可删除后把结果存放到新的缓冲区。

6)读入专有词典,将专有词典读入内存,保存为相应数据结构。

7)匹配专有词典,使用逆向最大匹配算法先后与读入内存的专有词典进行匹配,结果保存到缓冲区。

8)匹配通用词典,使用逆向最大匹配算法与通用词典进行匹配,获得粗分结果。

9)使用 Trigram 模型进行消歧,所谓 Trigram 模型是 N-gram的一种,其中 $N = 3$。公式如下:

$$P(w_1 w_2 \cdots w_n) = P(w_1)\ P(w_2 \mid w_1) P(w_3 \mid w_1 w_2) \cdots P(w_n \mid w_{n-2}\ w_{n-1})$$

通过计算,取使得概率最大的分词方式。比如:假设“催眠曲子” 通过切分词典有两种可能”催眠曲/子”和“催眠/曲子”。需要分别计算 p(催眠曲) * p(子) 和 p(催眠)p(曲子),找出概率最大的一种切分方法。可使用 Viterbi 算法进行求解[8]。所谓 Viterbi 算法的目的是:给定观察序列 O 以及模型 λ,如何选择一个对应的状态序列 S ,使得 S 能够最为合理地解释观察序列 O。Viterbi 算法[9]通过动态规划的方法找到概率最大的序列 O。

我们通过对通用语料库和专业语料库[10]进行比较,发现有相当多的通用语料库中高频词条,如、“今天上午”、“于今年”、“去年底”、“想起”等,在专业领域语料库变为中低频,而专业领域语料库中一些高频的词条,如“地理学报”、“相关系数”等,在通用语料库中的词频较低。对语料库进行数据统计,给予专业语料库和通用语料库不同权重,最终得出各词条概率系数,能有效解决这一问题。

4　实验分析

将该分词系统与中国科学院的 ICTCLAS 分词系统和 C++ 实现的逆向最大匹配算法进行了比较[11]。在这些页面中随机选择了 5 段文字,首先通过人工分词得到正确分词结果,然后分别使用 3 种分词方法对这 5 段文字进行分词。所谓分词准确度是指正确切分的词数占正确结果总词数的百分比。我们从国内权威的 GIS 网站 GIS 帝国上随机挑选了 80 篇文章进行分词测试,分词测试结果如表 1 所示。

以上实验数据表明,本文的分词算法采用了基于专有词汇优先的粗分以及使用 Trigram 模型消歧,比单纯的逆向最大匹配更加复杂,所以分词速度上会略有不足。但对 GIS 领域的文献而言,本算法提高了 GIS 专有词汇的识别率,因而准确性高于前两种分词方法。而且随着分词词典的进一步完善,效果会有更大提升。总而言之,本算法的分词速度和分词准确度均较佳,能满足专业领域搜索引擎的应用需要。

表 1　分词测试结果比较

算法	总耗时/s	平均速度/kBps	准确度/%
逆向最大匹配	21.23	2835.65	81.72
ICTCLAS	947.34	63.54	90.32
专业领域分词	157.22	382.93	93.89

5　结语

本文提出的分词方法的分词速度和准确度能有效满足面向专业应用领域信息处理的要求,但不足之处在于它不能自动识别新词。可以通过以下方法解决:根据业界最新的语料库,通过统计学方法并加上人工审核来有效控制词库。本方法没有实现真正的歧义识别,使用中文的语法和词性分析能进行真正的歧义识别,但会耗费大量时间和系统资源。这两个问题有待进一步研究。

参考文献:

[1] JIANG WENBIN, HUANG WENBIN, LIU QUN, *et al*. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging[EB/OL]. [2009-10-05]. http://www.cis.upenn.edu/~lhuang3/cascaded.pdf.

[2] 马玉春, 宋涛瀚. Web 中中文文本分词技术研究[J]. 计算机应用, 2004, 24(4): 134-136.

[3] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[C]// Readings in Speech Recognition. San Francisco: Morgan Kaufmann Publishers, 1990: 257-286.

[4] 张峰, 樊孝忠, 许云. 基于统计的中文姓名识别方法研究[J]. 计算机工程与应用, 2004, 40(10): 53-57.

[5] QIN YING, ZHANG SUXIANG, WANG XIAOJIE. Combining multi-knowledge for Chinese word segmentation disambiguation[C]// Proceedings of the 6th International Conference on Intelligent Systems Design and Applications. Washington, DC: IEEE Computer Society, 2006, 551-556.

[6] 张锋, 樊孝忠. 基于最大熵模型的交集型切分歧义消解[J]. 北京理工大学学报, 2005, 25(7): 590-594.

[7] HUANG C, ZHAO H. Chinese word segmentation: A decade review[J]. Journal of Chinese Information Processing, 2007, 21(3): 8-18.

[8] ZHANG RUIQIANG, KIKUI G, SUMITA E. Subword-based tagging by conditional random fields for Chinese word segmentation[C]// Proceedings of the Human Language Technology Conference of the NAACL. Morristown, NJ: Association for Computational Linguistics, 2006: 193-196.

[9] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3): 8-18.

[10] 赵海, 揭春雨. 基于有效子串标注的中文分词[J]. 中文信息学报, 2007, 21(5): 8-13.

[11] 顾铮, 顾平. 信息抽取技术在中医研究中的应用[J]. 医学信息学, 2007(20): 27-29.