文章编号:1671-8860(2009)08-0943-05　　文献标志码:A

一种中文自然语言表达交通信息的跨阶分词算法

陆　锋[1]　刘焕焕[1,2]　陈传彬[1,3]

(1　中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室,北京市朝阳区大屯路甲 11 号,100101)
(2　中国矿业大学(北京)资源与安全工程学院,北京市海淀区学院路丁 11 号,100083)
(3　福州大学福建省空间信息工程研究中心,福州市工业路 523 号,350002)

摘　要:在分析中文分词算法和交通信息自然语言表达特点基础上,提出了一种自然语言表达交通信息的跨阶匹配分词算法,以适应动态出行信息服务对数字形式结构化实时交通信息的迫切需求。该算法充分考虑了交通信息自然语言描述词库记录长度特点,通过设置对应的中文分词阶数,将传统中文分词的字符串指针 1 阶跨越方法改进为依词库性质变化的多阶跨越方法,对可能成词的中文字符串进行整体处理,极大地提高了自然语言表达交通信息的实时分词与理解效率。通过与改进 MM(maximum matching)算法的实验比较,本方法在理解成功率和容错性相同的情况下,效率比 MM 分词算法提高了 10 倍以上。

关键词:交通信息;中文自然语言处理;分词;跨阶法
中图法分类号:P208

目前,实时交通信息采集和发布技术在各大城市日趋成熟,对提高城市交通管理和公众出行效率起到了很大的作用。然而,来源于浮动车和感应线圈采集的交通流信息,难以覆盖整个路网,也难以获取突发性点状交通信息;来源于短信息、电话、监控摄像头的交通信息目前只能通过人工处理方式加以利用,或直接通过语音广播形式发送,并且发送量十分有限,无法服务于日益普及的动态导航过程。因此,迫切需要开展自然语言表达交通信息的实时分词技术研究,进而与路网空间信息进行实时融合,通过各种通讯协议发布,服务于广大的交通出行者。

目前的中文分词算法主要有以下三种[1-5]:① 机械分词法。又称词典式切分法,包括最大匹配法(MM 算法)、部件词典法、词频统计法、设立标志法、并行分词法、词库划分和联想匹配法等。② 语义分词法。切分时引入语义分析,如邻接约束法、综合匹配法、后缀分词法、特征词库法、约束矩阵法、语法分析法等。③ 人工智能法。模拟人脑思维功能进行分词,如神经网络分词法和专家系统分词法。目前,各种流行的分词算法有着不同的技术特点和适应性。机械分词法较为简洁,易于实现,尤其是最大匹配法及其改进算法,在工程上得到了广泛的应用[6-8]。但机械分词法难以处理未登录词,无法有效克服歧义切分。语义分词法切分精度较高,但在引入了语义分析的同时也增大了时空开销。人工智能法研究还处于初步阶段,效率不高,不易实现。

由于目前还不能达到真正意义上的自然语言理解,对于特定的应用领域,可以根据自然语言描述的规则,建立受限语法规则,获取输入自然语言中的关键字词信息[9,10]。因此,对于中文自然语言表达交通信息的自动化理解,应当根据交通信息自然语言表达的特点,开展分词方法研究。从应用特点上看,交通信息具有很强的实时性,只关心能否快速、正确地处理地点、方向、数值偏移量、事件等类型的关键词汇,忽略无关词汇。自然语言理解过程对分词算法的时间效率要求较高。从信息特征上看,交通信息描述较为简短,采用规范的陈述句句型,关键词汇具有长词优先的特点,存在较少的理解歧义。此外,由于理解目的在于和空间信息匹配,所有涉及的地址、方位和事件词汇已经在

收稿日期:2009-05-20。
项目来源:国家 863 计划资助项目(2006AA12Z209,2007AA12Z241);国家自然科学基金资助项目(40871184);中国科学院知识创新工程重点方向性资助项目(KZCX2-YW-308)。

词汇库中存在,不需要考虑未登录词的处理。因此,比较适合采用机械分词法中的匹配方法。

然而,传统的MM算法在匹配分词过程中通常采用逐渐减一字方式处理,处理效率低下。而词库中的最大词长通常大于所切分出的词长,即使采用正向逐一递增循环方式处理,虽然在一定程度上提高了传统MM方法中文分词的效率,但是依然采取逐字匹配方法,没有考虑词库记录长度的特点,对可能成词的中文字符串进行特别处理。也有文献基于长词优先原则,提出了首字扩词分词法,根据查询自然语言字符串首字符对词汇库的记录进行筛选,并在缩小匹配范围时根据筛选后的子词库记录长度决定匹配长度[11]。该算法避免了MM算法在匹配过程中固定不变的字符加减方式,更适合字符串长度完全随机的匹配分词过程。而交通信息词库记录字符串长度具有很强的规律性,应当利用这一特征尽早判断可能成词的中文字符串,尽可能避免无谓匹配过程。因此,本文提出一种采用跨阶匹配的中文自然语言表达交通信息分词算法。

1 跨阶分词算法原理

以中文自然语言表述的实时交通信息通常表示形式为"<地址> + {方向}+{偏移量} + <事件>"。交通信息描述词库包括地址词库、方向词库和事件词库。地址词库存储交通信息发生地址;方向词库存储描述交通信息时采用的方向信息,如"由东向西"、"从南到北"等;而事件词库则存储交通信息所对应的具体事件,如"车行缓慢"、"追尾"、"两车刮蹭"等。各词库中记录长度分布具有一定的规律。以笔者开发的北京市交通信息处理与融合系统为例,其中交通信息涉及的地址库记录长度分布如表1所示。

从表1可以看出,自然语言描述的交通信息发生地址长度具有一定的统计规律,地址库中99.3%的记录长度大于2,尤以记录长度为4或5居多。方向信息和事件信息也具有类似的统计规律。因此,采用经典或改进的MM算法进行分词时,指针在整个句子中逐字移动,并进行累加词库匹配,没有利用交通信息词库记录长度分布的统计规律,对可能成词的中文字符串进行整体处理,无疑是一种比较低效的方法。因此,本文基于交通信息自然语言描述词库记录长度特点,设置对应的中文分词阶数,提出跨阶匹配分词思想,将传统中文分词的字符串指针一阶跨越方法改进为依词库性质变化的多阶跨越,以提高分词效率。

表1 交通信息描述词库地址库记录长度分布

Tab.1 Length Distribution of Address Records in Real-time Traffic Information

记录长度	记录数	比例/%
2	29	0.71
3	797	19.48
4	1 480	36.18
5	1 218	29.77
6	427	10.44
≥7	140	3.42
合计	4 091	100

对中文自然语言描述的实时交通信息进行分词处理时,首先将整个句子从左侧开始,与地址库进行匹配,然后将分词后的字符串再分别与地址库(接受可能出现的二重地址描述)、方向库及事件库进行匹配。与地址库进行匹配时,根据地址库记录长度分布特点,设置初始阶数为3,根据匹配结果设置指针的前移或后移;与方向库进行匹配时,按照方向库中记录的最大长度设置阶数为4,成词则成功切分,若不成词,指针前移,再次与方向库进行匹配;与事件库进行匹配时,根据事件库记录长度分布特点,设置初始阶数为2,根据匹配结果设置指针的前移。

2 数据结构与算法流程

2.1 数据结构

本文采用了一种多层模式的数据结构。其具体描述如下:最大的层数为词库中最大词的单字数。其中,字母表示一个字,数字表示成词标志,0代表当前字串不能单独成词,只是某一词汇的一部分,1则代表可以成词。第一层存储单字,第二层存储以第一层的字串为前缀的双字或者双字词,第三层存储以第二层的字串为前缀的三字或者三字词,依此类推。采用多层树型结构,能不断

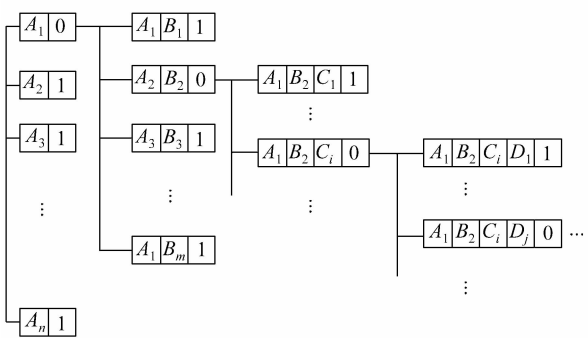

图1 词库存储数据结构

Fig.1 Data Structure of word Library

限定匹配范围,提高分词算法的效率。

2.2　算法流程

本文提出的跨阶中文匹配分词算法思想如下。

目标:对一个以中文自然语言描述的交通信息语句 $C_1C_2\cdots C_n$ 进行分词处理。

其定义符号如下:A 地址库;O 方向库;E 事件库;currentWord 为当前与词库比较的词汇;F 为地址词汇切分标记,初始为 false。

1) 读入 C_k,判断 F 值,若是 true,则判断 C_k 是哪个词库的前缀字符,若是 A 的前缀则转入步骤 2),若是 O 的前缀转入步骤 3),若是 E 的前缀转入步骤 4),若是 A 与 O 或 E 的前缀,转入步骤 5)。若 C_k 不是 3 个词库的前缀,将 C_k 从整个句子中切分出去,读入下一个可能成词的字符;若 F 为 false,只判断 C_k 是否是 A 的前缀,如是则转入步骤 2),否则将 C_k 从整个句子中切分出去。

2) 如 C_k 是 A 的前缀字符,读入 $C_{k+1}C_{k+2}$,判断 $C_kC_{k+1}C_{k+2}$ 是否也是 A 的前缀,如是则判断是否可成词,不成词继续循环读入下一个字符,再进行判断,如可成词则将 $C_kC_{k+1}C_{k+2}$ 切分出来作为候选地址,然后判断 $C_kC_{k+1}C_{k+2}$ 是否同时也是 A 中另外某词的前缀,若是,再循环往下读字符并进行判断,否则将 $C_kC_{k+1}C_{k+2}$ 作为成功切分地址,F 设为 true。如 $C_kC_{k+1}C_{k+2}$ 不是 A 的前缀,指针后退 2 位,以 C_{k+1} 作为起始字符转入步骤 1)。

3) 如 C_k 是 O 的前缀字符,读入 $C_{k+1}C_{k+2}C_{k+3}$,判断 $C_kC_{k+1}C_{k+2}C_{k+3}$ 是否可成词,如是则成功切分;不成词则指针前移,判断 $C_kC_{k+1}C_{k+2}$ 是否成词。依此类推,若成词则进行成功切分,如始终不成词,则以 C_{k+1} 为开始字符转入步骤 1)。

4) 如 C_k 是 E 的前缀字符,则读入 C_{k+1},判断 C_kC_{k+1} 是否是成词前缀,如是则判断是否成词,成词则成功切分,否则循环继续读入下一个字符,再进行判断。如 C_kC_{k+1} 不是成词前缀,以 C_{k+1} 作为起始字符转入步骤 1)。

5) 读入 C_{k+1},判断 C_kC_{k+1} 是否是 O 或 E 的成词前缀,若是则转入步骤 3)或步骤 4),若不是则判断 C_kC_{k+1} 是否是 A 的成词前缀,如是则转入步骤 2),否则以 C_{k+1} 作为起始字符转入步骤 1)。

其算法流程如图 2 所示。

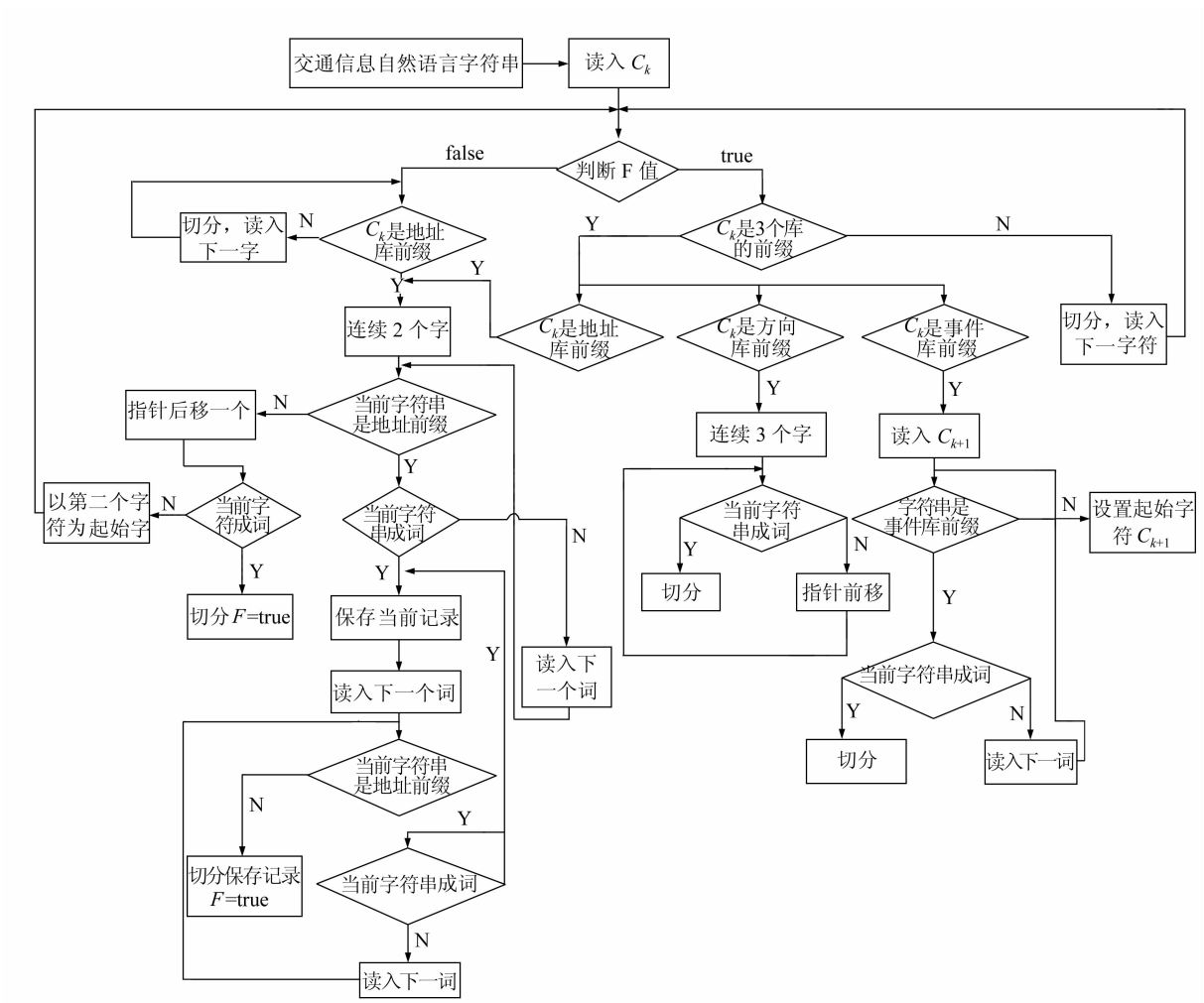

图 2　自然语言交通信息跨阶分词算法流程

Fig. 2　Flow Chart of Cross-step Word Segmentation Algorithm for Traffic Information in Natural Chinese

以实时交通信息“健翔桥由南向北行驶缓慢”为例,来介绍跨阶中文分词方法的运行过程:首先读入“健”字,写入currentWord,直接与地址库进行比较,判断currentWord是地址库中词的前缀。继续读入“翔桥”,currentWord修改为“健翔桥”,能够单独成词,此外地址词库中也不存在其他以“健翔桥”为前缀的词,则成功切分出地址为“健翔桥”。然后将currentWord置空,将地址切分标志F设为true;读入“由”字,写入currentWord,判断其为方向库的前缀,则读入“南向北”,修改currentWord为“由南向北”,判断成词,则成功切分出方向信息。再将currentWord置空,读入“行”字,写入currentWord,判断其为事件库中词的前缀,继续读入“驶”字,currentWord修改为“行驶”,“行驶”也是成词前缀,不能单独成词,继续读入“缓”和“慢”字符,currentWord置为“行驶缓慢”,可以单独成词,则成功切分出对应交通事件。

值得注意的是,交通信息中还可能存在数字型偏移量,偏移量数值难以预先在词库中逐一枚举。针对这一问题,本文采用数值型偏移量与字符串匹配分开处理策略,先以跨阶法处理输入的自然语言表达的交通信息,再对无法匹配的剩余字符串进一步处理,从而一次性提取出数字信息,以此达到偏移量切分目的。

3 效率实验

在上述算法研发的基础上,笔者采用Java技术实现了实时交通信息的跨阶中文分词算法。实时交通信息来源于北京交通广播电台,空间信息采用基于路幅段模型构建的路网数据集,符合地理导航数据库国际标准GDF4.0中1层定义[12]。采用Oracle 10g数据库管理系统完成所有数据的管理工作。选择北京市五环内城市路网作为示范区域,并随机收集了2007-11-09:08:00～18:00这一时段内北京交通广播电台发布的实时交通信息,共计400条,如图3所示。

实时路况信息

ID /	路况信息
1	西长安街双向都有交通管制措施建议出城的绕行莲花池东路，进城的方向可以绕行阜石路
2	三元西桥到三元桥由西向东行驶缓慢
3	西二环北段南北双向都是车多流量大
4	北沙滩桥由北向南主辅路行驶缓慢
5	机场高速的收费站由北向南的方向行驶缓慢
6	东便门桥由南向北方向行驶缓慢
7	航天桥到劲松桥由北向南的方向车多
8	马甸桥出京方向车多
9	南三环草桥由东向西最内侧的车道施工完毕，交通正在恢复当中
10	东四十条桥由北向南的路行驶缓慢
11	官园桥以西由东向西的方向行驶缓慢
12	德胜门桥由东往西的方向车多
13	赵公口桥由东向西的方向中间车道和外侧车道各有一起有事故，造成后车比较拥堵
14	东便门桥到建国门桥南向北方向车辆行驶缓慢
15	天宁寺桥南向北方向车辆行驶缓慢

图3 自然语言描述的实时交通信息

Fig. 3 Real-time Traffic Information Represented in Natural Chinese

实时交通信息是对交通状况的即时反映,而且具有很强的时效性。针对实时出行和智能导航,实时交通信息的处理效率至关重要。对400条交通信息分别用跨阶分词算法和改进的MM分词算法进行中文分词,跨阶分词算法和改进的MM分词算法理解成功率均为98%,容错性也完全相同,但改进的MM分词算法耗时为2 951 ms,而跨阶分词算法耗时仅为229 ms,跨阶分词算法比改进MM中文分词的效率提高了10倍以上,体现出很好的效率优势。

本文所提出的自然语言表达交通信息的跨阶匹配分词算法,充分考虑了交通信息自然语言描述词库记录长度特点,通过设置对应的中文分词阶数,将传统中文分词的字符串指针一阶跨越方法改进为依词库性质变化的多阶跨越,对可能成词的中文字符串进行成词处理。在理解成功率和容错性相同的情况下,该算法极大地提高了自然语言表达交通信息的实时分词与理解效率,效率比经典的MM分词算法提高10倍以上。后续研究中将进一步提高算法的容错性,使其可处理以长句或组合多句表达的交通信息,以尽可能提高自然语言表达的实时交通信息的自动化、智能化处理水平。

参 考 文 献

[1] 文庭孝,邱均平,侯经川. 汉语自动分词研究展望[J]. 现代图书情报技术,2004(7):6-9

[2] 张春霞,郝天永. 汉语自动分词的研究现状与困难[J]. 系统仿真学报,2005,17(1):138-143

[3] 邱均平,文庭孝,周黎明. 汉语自动分词与内容分析

法研究[J]. 情报学报,2005,24(3): 309-317
[4] 黄昌宁,赵海. 中文分词十年回顾[J]. 中文信息学报,2007(9): 8-19
[5] Fu G,Kit C,Webster J J. Chinese Word Segmentation as Morpheme-based Lexical Chunking[J]. Information Sciences,2008,178: 2 282-2 296
[6] 乐小虬,杨崇俊,于文洋. 基于空间语义角色的自然语言空间概念提取[J]. 武汉大学学报·信息科学版,2005,30(12):1 100-1 103
[7] 胡斌,汤伟,刘晓明. 基于自然语言理解的文本标图系统设计与实现[J]. 解放军理工大学学报(自然科学版),2005,6(2):132-136
[8] 马林兵,龚健雅. 空间信息自然语言查询接口的研究与应用[J]. 武汉大学学报·信息科学版,2003,28(3):301-305
[9] 龙毅,张翎,胡雷地,等. 移动 GIS 中语音与自然语言的应用模式探讨[J]. 测绘科学技术学报,2008,25(1): 8-12
[10] 徐爱萍,边馥苓. GIS 中文查询系统的词典设计与分词研究[J]. 武汉大学学报·信息科学版,2006,31(4):348-351
[11] 吴静,蔡砥,王铮. 地理信息系统中自然语言查询的分词处理与应用[J]. 地球信息科学[J],2005,7(3):67-71
[12] 蒋捷,韩刚,陈军. 地理导航数据库[M]. 北京:科学出版社, 2003

第一作者简介:陆锋,博士,研究员,博士生导师,研究方向包括空间数据库管理技术、交通 GIS 理论与技术、LBS 与导航数据库技术、城市发展与城市 GIS 技术等。
E-mail:luf@lreis.ac.cn

A Cross-step Word Segmentation Algorithm for Understanding Traffic Information Represented in Natural Chinese Language

LU Feng[1]　*LIU Huanhuan*[1,2]　*CHEN Chuanbin*[1,3]

(1　LREIS, Institute of Geographic Sciences and Natural Resources Research, CAS, A11 Datun Road, Beijing 100101, China)
(2　College of Resources and Safety Engineering, China University of Mining and Technology, D11 Xueyuan Road, Beijing 100083, China)
(3　Spatial Information Research Center, Fuzhou University, 523 Gongye Road, Fuzhou 350002, China)

Abstract: A novel cross-step word segmentation algorithm is proposed to process real-time traffic information represented in natural Chinese in this paper, to meet the urgent need of real-time traveling information service, for dynamic traffic information. Considering the record length distribution of the word libraries depicting real-time traffic information, this algorithm sets corresponding steps of word segmentation for address, direction and event libraries, and improves the one step running of the string pointer in classical Chinese word segmentation to flexible multiple steps running, so as to aggregate possible Chinese words efficiently. A case study shows that the proposed algorithm runs 10 times faster than an improved MM algorithm, whilst keeping similar accuracy and robustness. The authors argued that the presented algorithm is greatly helpful to the automatic and intelligent processing of the real-time traffic information, and facilitate the development of travel information services.

Key words: traffic information; natural Chinese processing; word segmentation; cross-step algorithm

About the first author: LU Feng, researcher, Ph.D, Ph.D supervisor, Research interests include spatial database technologies, GIS for transportation and LBS applications, urban development, etc.
E-mail: luf@lreis.ac.cn