基于FAQ的智能答疑系统中分词模块的设计

程节华
(安徽科技学院 计算机系,安徽 蚌埠 233100)

摘 要:在基于FAQ的智能答疑系统中,分词处理是基础和关键。分词质量的好坏直接影响智能答疑系统的准确性。针对实际应用领域的需要,本系统分词模块采取普通词典和专业词典混合的词典设计方案,分词算法采用正向最大匹配的分词算法。为了提高词典的查找速度,分词词典采用Hash表和二维数组的数据结构,根据汉字的内码利用Hash方法,求得在内存的地址,然后计算其索引项的二维数组的下标,对于词典的搜索采用二分查找法。实验结果表明:该分词系统提高了智能答疑系统的效率和准确率。

关键词:自然语言处理;智能答疑;分词
中图分类号:G434 **文献标识码**:A **文章编号**:1673-629X(2008)07-0181-03

Design of Words Module in Intelligent Q/A System Based on FAQ

CHENG Jie-hua
(Department of Computer, Anhui College of Science and Technology, Bengbu 233100, China)

Abstract: In the intelligent Q/A system based on FAQ, words processing is the basis and key point. Words quality directly influences the intelligence Q/A system accuracy. Aiming at special practical field, a design scheme for words module dictionary which combined with general and special ones is presented. Further more; an algorithm based on forward maximum match is also put into use. In order to enhance the dictionary the search speed, words dictionary uses Hash tables and two-dimension array as data structure, uses the Hash method according to the Chinese encoding to obtain the memory address, and then, calculates the index two-dimensional array subscript, uses the binary searching method for the dictionary search. The experimental result indicates that this participle enhances the intelligent Q/A system efficiency and is accurate rate.

Key words: natural language processing; intelligent Q/A; words segmentation

0 引 言

随着网络通信技术的发展和网络应用的普及,现代远程教育已被越来越多的机构和学习者所接受。远程教育是在师生分离的环境下进行的,课程答疑作为现代远程教育学生支持的一个重要环节,可以及时解答学生的疑难问题,帮助学生内化课程的概念,消除学生的学习障碍。

目前远程教育机构提供的网络答疑大多以Internet为基础,非实时的答疑有E-mail、BBS等,实时的答疑有语音在线、视频会议等。这些答疑方式虽然实现简单,但无形中增加了教师的工作量,特别是对某些问题的重复回答,无法实现远程教育课程答疑的及时性、实时性的特点。根据这一实际情况,出现了一批基于WEB的智能答疑系统[1~3]的研究。针对高校网络课程,设计和实现了一个基于FAQ的智能答疑系统,它允许用户以自然语言方式提出问题,并获得相应的解答。系统主要功能模块为分词处理、FAQ查询以及FAQ维护等,文中重点介绍其中的分词处理模块。

汉语自动分词是中文自然语言处理中最基本和主要的步骤,分词的质量直接影响自然语言处理的结果。汉语分词已广泛应用于词频统计、新词识别、计算机辅助词典编撰和词语搭配研究等众多领域。另外在汉语文献处理自动化中如:自动标引、自动摘录、自动分类、信息检索等汉语分词也大有用武之地。对于汉语分词来说存在的主要问题有:

(1)歧义切分字段的识别与处理。

(2)未登录词的识别。

(3)多义词的词性标注。

在汉语自动分词与自动标引的研究与实践上进行了大量的研究,找到了许多解决汉语分词的方法,归纳起来有:最大匹配法、逆向最大匹配法、逐词遍历

收稿日期:2007-10-28
基金项目:安徽省自然科学研究项目(KJ2007B159)
作者简介:程节华(1970-),男,安徽怀宁人,讲师,硕士,研究方向为自然语言处理。

法、设立切分标志法、最佳匹配法、有穷多层次列举法、二次扫描法、高频优先分词法、基于期望的分词法、联想——回溯法、双向扫描法、邻接约束法、扩充转移网络分词法、语境相关法、全自动词典切词法、基于规则的分词法、多遍扫描联想法、部件词典法、链接表法、最少分词词频选择法、专家系统分词法、基于神经网络的分词方法等 22 种。这些方法又大体上可分为两类：一类是基于规则的，大多数中文分词方法都属此类；一类是基于语料库的，如神经网络分词法部分的属于此类。基于规则的分词算法的计算模型均是概率论中的马尔可夫过程又称元语法、隐马尔可夫过程和通信中的信道噪声模型。但无论是马尔可夫过程还是信道噪声模型，最后都归结为计算词频的统计信息，串频和互信息是词频的另一种表现形式。但遗憾的是自然语言远不是一个经过事先精心规划的系统，难以用一套完整的规则去准确地预测正式汉语文本中所出现的各种变异。考虑到实际需要在本分词系统中采用正向最大匹配法，分词词典采用专业词典和普通词典相结合的词典设计方案。

1 分词字典的设计

分词词典是汉语自动分词系统的基本组成部分。自动分词系统所需各类信息(知识)都要从分词词典中获取，分词词典查询速度直接影响分词系统的速度。而现实应用(如因特网上的中文文本检索、汉字与汉语语音识别系统的后处理以及中文文语转换系统的前处理等)均对分词速度提出了迫切要求，因此建立高效快速的分词词典机制势在必行。

文中对用户输入问题进行分词时，所用到的字典是普通分词字典和专业分词字典的混合。其基本设计思想是首先判断用户输入的关键词在专业词典中是否已经存在，若存在则显示该词字典中已有并返回，否则向专业字典中添加该词。接着判断普通字典中是否存在，若不存在则向普通词典中添加该词并对普通词典重新进行排序，若存在则返回。其详细的设计流程图如图 1 所示。

2 分词词典的组织

国内自 20 世纪 80 年代中后期就开展了中文电子词表的研制，现有词表有采用 B+ 树(或其变种)作为词表索引数据结构的，也有利用现成的关系数据库技术的。还有一些系统由于词汇量不是很大，也有用纯文本的。中文字的准确数目目前尚未完全弄清，一种说法是中文字大概有 10 万个左右，常用汉字则只有几千个。图形字符代码表 GB5007－85 共收录汉字 6763 个(一级汉字 3755 个，二级汉字 3008 个)，目前汉字代码体系由输入码、交换码、内码、区位码等构成，很多中文平台都采用内码来处理汉字信息。内码、区位码与交换码之间存在一一映射关系：

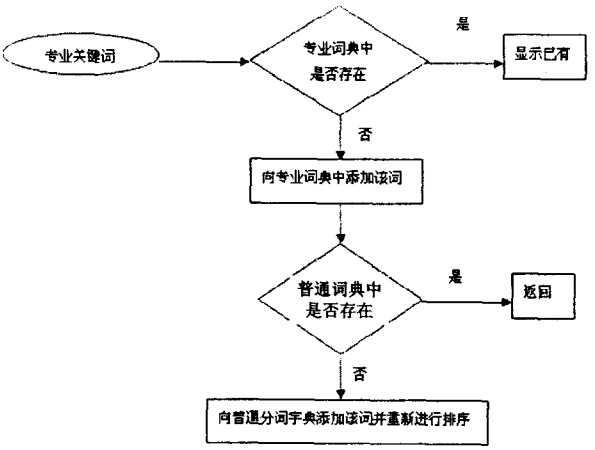

图 1 分词词典的设计

设汉字 i 的内码为 ICCi，区位码为 RCCi，交换码为 ECCi，则

$$\text{HighByte}(\text{ICC}i) = \text{Q}(\text{RCC}i) + 0\text{xa0}$$

$$\text{LowByte}(\text{ICC}i) = \text{W}(\text{RCC}i) + 0\text{xa0}$$

$$\text{HighByte}(\text{ICC}i) = \text{Q}(\text{ECC}i) + 0\text{x80}$$

$$\text{LowByte}(\text{ICC}i) = \text{W}(\text{ECC}i) + 0\text{x80}$$

考虑到中文字的编码体系和中文词的上述特点，分词词典采用如下数据结构[4]，其在主存中的形式如图 2 所示。

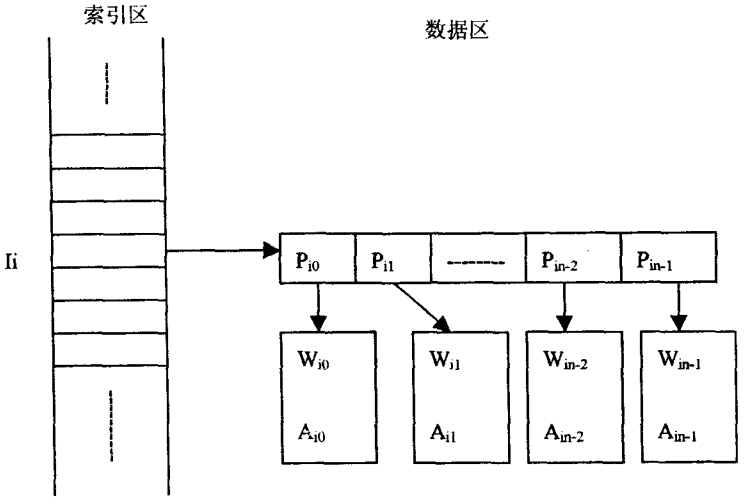

图 2 中文词典在内存中组织方式

其中，Pi：指向所有首字为第 i 个汉字 CCi 的词条的指针；Pik：指向首字为 CCi 的第 k 个词条的指针；Wik：首字为 CCi 的第 k 个词(词条按内码顺序从小到大排列)，不包括首字；Aik 的属性(包括价值、歧义、位性、词性等)；Ii：字 CCi 的索引项，占 5 个字节，其结构

形式如下:

CC*i*	*n*	flag

其中,*n* 为首字为 CC*i* 的词的数目,flag 为是否单独成词标志。在磁盘上进行存储时,每个汉字占一行(共 6763 个汉字),所有以该字为首字的词中间用分割符 S 分开,第 *i* 行的存储形式如下所示:

CC_i	*n*	flag	S	W_{i0} A_{i0}	.....	S	W_{in-1} A_{in-1}

根据上述词典的组织,主要词典搜索算法如下:

设待查的输入项以汉字内码方式表示,首字为 CC*i*,其内码为 ICC*i*,则

$$Q(RCCi) = HighByte(ICCi) - 0xa0$$
$$W(RCCi) = LowByte(ICCi) - 0xa0 \quad (1)$$

利用 Hash 方法求其在内存中的地址 A*i*:

$$Ai = f(Q(RCCi), W(RCCi)) \quad (2)$$

在文中的词表管理系统中,利用 Q(RCC*i*) 和 W(RCC*i*) 计算二维数组的下标:

这种 Hash 方法实质上是一种一一映射,首字不同,地址亦不同,避免了模式冲突问题:利用 CC*i* 可经过式(1)、式(2)的运算而直接得到首字索引项 *Ii*,这一过程不进行任何匹配。找到索引项后,如待查项为单字,还要看 CC*i* 是否能够独立成词,否则根据待查项的第 2 个字进行二分查找,最后返回查找结果。查找算法描述如下:

算法 1 FindWord

(1) 根据 CC*i* 计算两维数组下标:

下标 1 = Q(RCC*i*) - 16 //汉字从第 16 区开始,

下标 2 = W(RCC*i*) - 1 //汉字的位从 1 开始,

根据下标找到 *Ii*,取词数 *n*;

(2) 如果待查项为单字:

如果 CC*i* 能够独立成词,返回 TRUE,否则返回 FALSE;

否则根据 *n* 进行二分查找,直到查找结束。

3　分词模块的设计

对于汉语自动分词现有的分词算法有:字符串匹配算法(如:正向最大匹配法,逆向最大匹配法、双向最大匹配法、逐词遍历匹配法、切割标志法),基于规则的分词方法(如文献[5]提出一套机械切分与语义校正的汉语自动分词方法;文献[6]提出了通过利用邻接约束知识解决分词中歧义问题),基于语料库统计的方法(如:文献[7]提出可以用穷尽句子的所有可能的切分方法,然后通过统一算法来得到可能正确结果;文献[8]是通过有限全切分,然后得到有向图,通过计算最佳路径来实现自动分词;文献[9]提出基于支持向量机(SVM)和 k-NN 向结合一种分类方法来解决交集型伪歧义字段;文献[10]提出了基于最长次长匹配的汉语自动分词算法;文献[11]提出了一种利用句内相邻字之间的互信息及 t-测试差解决汉语自动分词中交集型歧义切分字段方法。

本系统采用简单的词典分词法——正向最大匹配算法[12]。其算法思想为:设 D 为词典,Max 为最大词长,S 为待分词字串.每次分词时从字符串 S 的前端截取 Max 个字符形成子串 W,然后在 D 中查找是否存在 W,若找到,则 W 为词,进行存储;否则,从字串 W 的后端减掉一个字符继续查找,直到找到或剩下最后一个字符,再次截取时起始位置为后移一个词的长度,并重复上述过程,直至 S 为空。图 3 是其设计流程图。

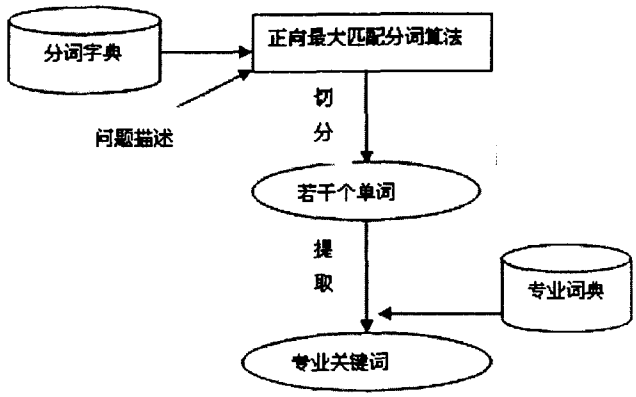

图 3　分词模块的设计

为了方便系统中其他模块引用该分词算法,在设计时把它单独抽象成一个业务逻辑过程,并用一个类进行了封装,这样既可以方便其他模块的调用又方便以后对该算法的不断完善及扩展。

现将 Segment 类中主要方法列举如下,并对其作一些简单介绍:

GetCharType(stirng);判断字符类型

InitWordDics();加载词列表

LoadWords(string ,System. Collection. ArrayList);加载文本词组到 ArrayList

OutArrayList(System. Collections. ArrayList);输出 ArrayList 中的数据

SegmentText(string);分词函数,输入参数为要分词的文本,输出为分词后的文本

SortDic();对加载的词典进行排序

4　结束语

目前基于自然语言形式的智能答疑系统研究还比较薄弱,距离真正实用尚有一定的距离。文中对智能答疑系统中的分词模块做了研究,进一步的研究可以

(下转第 186 页)

在灾害应急中是起着决定性作用的。面向城市地质防灾应急指挥部门,整合地质灾害及相关信息的综合需求和资源为基于 GIS 的城市基础地质与防灾应急系统的数据基础,实现地质灾害与勘察等相关信息资源共享,为地质灾害应急的智能化奠定基础。

3.2 城市地质灾害风险分析、评价与预测模型研究

利用 GIS 空间分析与建模技术建立城市地质灾害风险分析、评价与预测模型的技术路线:

①利用 3S 技术进行城市基础地质调查、地质灾害监测,建立 GIS 平台的地学信息空间数据库和自然灾害风险评估决策模型;

②建立工程体 - 地质体整体稳定性分析评价模型和建立合理、规范的地质灾害空间预测评价指标体系,并通过实时、重复 GPS 监测和航片、RS 解译获得地质灾害发生的时序和空间分布规律,建立不同周期下地质灾害的风险等级,实现对区域、局域层次地质灾害的动态评价,为防灾应急提供支持;

③通过影响城市地质灾害演化和宏观因素的研究,建立估计灾害发生可能性的预测模型,并通过对风险区内建筑物等的数量特征和易损性的研究,对城市地质灾害进行风险评价以及分析地质灾害比较突出的基础地质环境,据此进行城市地质灾害成灾环境和成灾模式研究,分析影响城市地质灾害的主导因素和敏感因素,建立针对性的城市地质灾害应急方案和最优调控方法。

4 结束语

文中的创新点主要表现在:

(1)UML 建模技术在地质领域的应用:采用 UML 语言建立地质领域的概念模型,进一步分析类的属性、行为及相互关系,建立系统的逻辑模型,包括静态结构模型、元数据模型、动态行为模型和 GIS 表现模型等,为地质防灾应急系统开发提供快速高效的模型基础;

(2)基于空间统计与空间机理模型的反演、模拟与预报:通过空间统计与空间机理模型的反演、模拟与预报,对城市地质灾害进行风险评价和分析地质灾害成灾环境和成灾模式,得出影响城市地质灾害的主导因素和敏感因素,帮助决策者进行应急指挥与决策支持。

文中的后期阶段工作为地质信息模型和地质灾害风险分析、评价与预测模型的详细设计与具体实现。

参考文献:

[1] 张宗祜. 环境地质与地质灾害[J]. 第四纪研究,2005,25(1):1-5.

[2] 国土资源部. 地质环境与地质灾害研究[EB/OL]. 2004. http://www.cigem.gov.cn/ReadNews.asp? NewsID = 1186.

[3] 杨起明,廖化荣,黄显艺. 基于 GIS 的地质灾害信息系统的研究[J]. 西部探矿工程,2006,18(6):283-285.

[4] Frankel D S. Model Driven Architecture: Applying MDA To Enterprise Computing[M]. America: Wiley Publishing,2003.

[5] 中科永联高级技术培训中心[EB/OL]. 2007. http://www.itisedu.com/phrase/200603051312555.html.

[6] 李 琦,郭玲玲. 面向数字城市的空间应用服务互操作模型研究[J]. 地理与地理信息科学,2003,19(3):14-17.

[7] 包世泰. 基于 GIS 的地质勘察信息模型研究及其应用[D]. 广州:中国科学院广州地球化学研究所,2004.

(上接第 183 页)

通过扩充和优化专业词典,以及采用更有效的分词算法来提高对用户问句的分词效果。

参考文献:

[1] 杨鸿雁, 唐 棣. 基于 Web 的网络答疑系统的设计与实现[J]. 沈阳师范学院学报:自然科学版,2001(3):22-26.

[2] 陈小茵. 基于自然语言的自动答疑系统设计[J]. 南京广播电视大学学报, 2005(4):85-87.

[3] 侯丽敏,朱 一,周 舫,等. 基于网络的智能答疑系统的研究[J]. 微机发展,2005,15(8):120-123.

[4] 陈桂林,王永成,韩客松,等. 一种高效的中文电子词表数据结构[J]. 计算机研究与发展,2000,37(1):109-115.

[5] 姚天顺,张桂平,吴映明. 基于规则的汉语自动分词系统[J]. 中文信息学报, 1990,4(1):37-43.

[6] 王锡江,王启祥,陈家骏. 基于邻接知识的汉语自动分词系统[J]. 计算机研究与发展,1992,29(11):54-58.

[7] 王 伟,钟义信,孙 建,等. 一种基于 EM 非监督训练的自组织分词歧义解决方案[J]. 中文信息学报,2001,15(2):38-44.

[8] 沈达阳,孙茂松,黄昌宁. 汉语分词系统中的信息集成和最佳路径搜索方法[J]. 中文信息学报,1997,11(2):34-47.

[9] 李 蓉,刘少辉,叶世伟,等. 基于 SVM 和 K_NN 结合的汉语交集型歧义切分方法[J]. 中文信息学报,2001,15(6):13-18.

[10] 黄德根,朱和合,王昆仑,等. 基于最长次长匹配的汉语自动分词[J]. 大连理工大学学报,1999,39(6):121-125.

[11] 孙茂松,黄昌宁,邹嘉彦,等. 利用汉语的二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展,1997,34(5):14-21.

[12] 王 坚. 专业搜索引擎的实现与研究——中文分词算法[J]. 电子科学技术评论,2005(3):77-79.