ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.5,No.10, April 2009, pp.2605-2607

E-mail: xsjl@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

中文分词算法概述

龙树全, 赵正文, 唐华

(西南石油大学 计算机科学学院,四川 成都 610500)

摘要:当前搜索引擎技术被广泛地应用,这使得全文检索技术和中文分词技术的研究逐渐深入。中文分词是中文信息的关键技术之一,其质量高低直接影响中文信息处理效率。文章致力于研究中文分词算法,对多种中文分词算法、自动分词系统的理论模型进行了详细的阐述和讨论,为中文分词的进一步发展提供基础和方向。

关键词:中文分词;全文检索;算法;搜索引擎;歧义切分

中图分类号:TP391.1 **文献标识码**:A **文章编号**:1009-3044(2009)10-2605-03

Overview on Chinese Segmentation Algorithm

LONG Shu-quan, ZHAO Zheng-wen, TANG Hua

(Department of Computer Science and Technology, Southwest Petroleum University, Chengdu 610500, China)

Abstract: Currently, the search engine technology has been widely used,which brings in-depth researches to full-text search technology and Chinese segmentations; Chinese Segmentation is one of the key technologies of Chinese information, it directly affects the quality of Chinese information processing efficiency.This article dedicated to Research on Chinese Segmentation Algorithm, described in detail and discuss to some kinds of Chinese Segmentation Algorithms, Theoretical model of Auto-Segmentation system. Provide foundation and direction for the further development of Chinese segmentations.

Key words: chinese segmentations; full-text search; algorithm; search engine; ambiguous word segmentation

1 引言

自然语言处理是人工智能的一个重要分支。中文分词是中文自然语言处理的一项基础性工作,也是中文信息处理的一个重要问题。随着搜索引擎技术的广泛应用,全文检索技术和中文分词技术也逐步受到广泛的研究和应用,然而到目前为止,还没有完全成熟实用的中文分词系统面世,这成为严重制约中文信息处理发展的瓶颈之一。本文致力于研究中文分词算法,通过分词算法对分词的质量做出客观的判断和评估,从而为中文分词的进一步发展提供基础和方向。

2 中文分词技术综述

2.1 全文检索技术

所谓全文检索是指计算机索引程序通过扫描文章中的每一个词,对每一个词建立一个索引,指明该词在文章中出现的次数和位置,当用户查询时,检索程序就根据事先建立的索引进行查找,并将查找的结果反馈给用户的检索方式。在中文文档中根据是否采用分词技术,索引项可以是字、词或词组,由此可分为基于字的全文索引和基于词的全文索引。

基于字的全文索引是指对于文章中的每一个字都建立索引,检索时将词分解为字的组合。对于各种不同的语言而言,字有不同的含义,比如英文中字与词实际上是合一的,而中文中字和词有很大分别。此方法查全率较高,但查准率较低。有时会出现令人啼笑皆非的检索结果,如检索货币单位"马克"时,会把"马克思"检索出来。

基于词的全文索引是指对文章中的词,即语义单位建立索引,检索时按词检索,并且可以处理同义项等。英文等西方文字由于按照空白切分词,因此实现上与按字处理类似,添加同义处理也很容易。中文文字则需要切分字词,以达到按词索引的目的。对中文文档进行切词,提高分词的准确性,抽取关键词作为索引项,实现按词索引可以大大提高检索的准确率。

2.2 中文分词技术

中文分词与英文分词有很大的不同,对英文而言,一个单词就是一个词,而汉语是以字为基本的书写单位,词语之间没有明显的区分标记,需要人为切分。中文分词系统是利用计算机对中文文本进行词语自动识别的系统,对其研究已经取得了很多成果,出现了众多的算法。根据其特点,可以将现有的分词算法分为四大类:基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法和基于语义的分词方法等。

3 中文分词方法

中文分词方法的基本原理是针对输入文字串进行分词、过滤处理,输出中文单词、英文单词和数字串等一系列分割好的字符串。中文分词模块的输入输出如图 1 所示。

图 1 中文分词原理图

3.1 基于字符串匹配的分词方法

这种方法又叫作机械分词方法、基于字典的分词方法,它是按照一定的策略将待分析的汉字串与一个"充分大的"机器词典中的词条进行匹配。若在词典中找到某个字符串,则匹配成功(识别出一个词)。该方法有三个要素,即分词词典、文本扫描顺序和匹配原则。文本的扫描顺序有正向扫描、逆向扫描和双向扫描。匹配原则主要有最大匹配、最小匹配、逐词匹配和最佳匹配。

收稿日期:2009-02-11

1) 最大匹配法(MM)。基本思想是:假设自动分词词典中的最长词条所含汉字的个数为 i,则取被处理材料当前字符串序列中的前 i 个字符作为匹配字段,查找分词词典,若词典中有这样一个 i 字词,则匹配成功,匹配字段作为一个词被切分出来;若词典中找不到这样的一个 i 字词,则匹配失败,匹配字段去掉最后一个汉字,剩下的字符作为新的匹配字段,再进行匹配,如此进行下去,直到匹配成功为止。统计结果表明,该方法的错误率为 1/169。

2) 逆向最大匹配法(RMM)。该方法的分词过程与 MM 法相同,不同的是从句子(或文章)末尾开始处理,每次匹配不成功时去掉的是前面的一个汉字。统计结果表明,该方法的错误率为 1/245。

3) 逐词遍历法。把词典中的词按照由长到短递减的顺序逐字搜索整个待处理的材料,一直到把全部的词切分出来为止。不论分词词典多大,被处理的材料多么小,都得把这个分词词典匹配一遍。

4) 设立切分标志法。切分标志有自然和非自然之分。自然切分标志是指文章中出现的非文字符号,如标点符号等;非自然标志是利用词缀和不构成词的词(包括单音词、复音节词以及象声词等)。设立切分标志法首先收集众多的切分标志,分词时先找出切分标志,把句子切分为一些较短的字段,再用 MM、RMM 或其它的方法进行细加工。这种方法并非真正意义上的分词方法,只是自动分词的一种前处理方式而已,它要额外消耗时间扫描切分标志,增加存储空间存放那些非自然切分标志。

5) 最佳匹配法(OM)。此法分为正向的最佳匹配法和逆向的最佳匹配法,其出发点是:在词典中按词频的大小顺序排列词条,以求缩短对分词词典的检索时间,达到最佳效果,从而降低分词的时间复杂度,加快分词速度。实质上,这种方法也不是一种纯粹意义上的分词方法,它只是一种对分词词典的组织方式。OM 法的分词词典每条词的前面必须有指明长度的数据项,所以其空间复杂度有所增加,对提高分词精度没有影响,分词处理的时间复杂度有所降低。

由上面的算法,不难看出基于字符串匹配的分词方法的优缺点:

优点:简单,易于实现。

缺点:1)匹配速度慢;2)存在交集型和组合型歧义切分问题;3)词本身没有一个标准的定义,没有统一标准的词集;4)不同词典产生的歧义也不同;5)缺乏自学习的智能性。

3.2 基于理解的分词方法

该方法又称基于人工智能的分词方法,其基本思想就是在分词的同时进行句法、语义分析,利用句法信息和语义信息来处理歧义现象。它通常包括三个部分:分词子系统、句法语义子系统和总控部分。在总控部分的协调下,分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断,即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。目前基于理解的分词方法主要有专家系统分词法和神经网络分词法等。由于汉语语言知识的笼统、复杂性,难以将各种语言信息组织成机器可直接读取的形式,因此目前基于理解的分词系统还处在试验阶段。

1) 专家系统分词法。从专家系统角度把分词的知识(包括常识性分词知识与消除歧义切分的启发性知识即歧义切分规则)从实现分词过程的推理机中独立出来,使知识库的维护与推理机的实现互不干扰,从而使知识库易于维护和管理。它还具有发现交集歧义字段和多义组合歧义字段的能力和一定的自学习功能。

2) 神经网络分词法。该方法是模拟人脑并行,分布处理和建立数值计算模型工作的。它将分词知识所分散隐式的方法存入神经网络内部,通过自学习和训练修改内部权值,以达到正确的分词结果,最后给出神经网络自动分词结果。

3) 神经网络专家系统集成式分词法。该方法首先启动神经网络进行分词,当神经网络对新出现的词不能给出准确切分时,激活专家系统进行分析判断,依据知识库进行推理,得出初步分析,并启动学习机制对神经网络进行训练。该方法可以较充分发挥神经网络与专家系统二者优势,进一步提高分词效率。

3.3 基于统计的分词方法

该方法的主要思想:词是稳定的组合,因此在上下文中,相邻的字同时出现的次数越多,就越有可能构成一个词。因此字与字相邻出现的概率或频率能较好反映成词的可信度。可以对训练文本中相邻出现的各个字的组合的频度进行统计,计算它们之间的互现信息。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时,便可以认为此字组可能构成了一个词。该方法又称为无字典分词。

该方法所应用的主要的统计模型有:N 元文法模型、隐 Markov 模型和最大熵模型等。在实际应用中一般是将其与基于词典的分词方法结合起来,既发挥匹配分词切分速度快、效率高的特点,又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

3.4 基于语义的分词方法

语义分词法引入了语义分析,对自然语言自身的语言信息进行更多的处理,如扩充转移网络法、知识分词语义分析法、邻接约束法、综合匹配法、后缀分词法、特征词库法、矩阵约束法、语法分析法等。

1) 扩充转移网络法。该方法以有限状态机概念为基础。有限状态机只能识别正则语言,对有限状态机作的第一次扩充使其具有递归能力,形成递归转移网络(RTN)。在 RTN 中,弧线上的标志不仅可以是终极符(语言中的单词)或非终极符(词类),还可以调用另外的子网络名字分非终极符(如字或字串的成词条件)。这样,计算机在运行某个子网络时,就可以调用另外的子网络,还可以递归调用。词法扩充转移网络的使用,使分词处理和语言理解的句法处理阶段交互成为可能,并且有效地解决了汉语分词的歧义。

2) 矩阵约束法。其基本思想是:先建立一个语法约束矩阵和一个语义约束矩阵,其中元素分别表明具有某词性的词和具有另一词性的词相邻是否符合语法规则,属于某语义类的词和属于另一词义类的词相邻是否符合逻辑,机器在切分时以之约束分词结果。

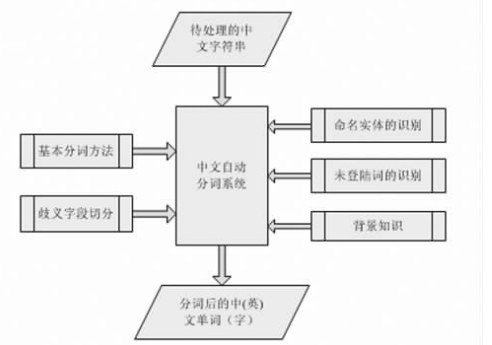

图 2 中文自动分词系统的框架

4 中文分词算法中的难点

4.1 歧义问题

歧义切分字段处理一个汉语句子是以连续字串的形式书写的。

由于可能存在歧义,分词并不是一个简单的从输入串中发现合法词的过程。一个句子经常对应几个合法词序列,因此,汉语分词中的一个重要问题就是在所有这些可能的序列中选出一个正确的结果。歧义切分是自动分词中不可避免的现象,是自动分词中一个比较棘手的问题。对歧义切分字段的处理能力,严重影响到汉语自动分词系统的精度。实践表明,只用机械匹配进行分词,其精度不可能高,虽然有时也能满足一些标准不高的需要,但不能满足中文信息处理高标准的要求。

4.2 未登录词识别问题

未登录词辨别未登录词包括中外人名、中国地名、机构组织名、事件名、货币名、缩略语、派生词、各种专业术语以及在不断发展和约定俗成的一些新词语。是种类繁多,形态组合各异,规模宏大的一个领域。对这些词语的自动辨识,是一件非常困难的事。

5 自动分词的评价准则

自动分词系统的最主要的工作是进行分词。对于分词而言,不仅要求所研制的软件在分词的正确率和速度方面满足一定的要求,而且要象开发大型传统软件那样,在各个阶段不断地进行评价,其目的主要是检查它的准确性和实用性,分词的评价主要有以下几个方面:

5.1 分词正确率

书面汉语的文本可以看成是字符序列,分词的正确率直接影响更高一级的处理。现有的分词系统切分错误主要集中在歧义字段和专有名词(如人名、地名、机构名和未登录词等)。为了获得分词系统切分正确率,应该进行整体测试,歧义测试和专业词测试。因此,自动分词系统的切分正确率的基本公式为:

$$S = \sum_{i=1}^{3} \beta_i S_i$$

其中,S_1,S_2,S_3。分别为总体测试、歧义测试和专业词测试的正确率;
B_i(i=1,2,3)为三种测试加的权值。

5.2 切分速度

切分速度是指单位时间内所处理的汉字个数。在分词正确率基本满足要求的情况下,切分速度是另一个很重要的指标,特别对于算法不单一,使用辅助手段,诸如联想,基于规则,神经网络,专家系统等方法更应注意这一点。通常中文信息处理的文本数量是相当大的,因此必须考虑方法是否能使系统总开销合理。在人机交互方式下处理歧义问题的策略和人机接口的设计,有时会严重地影响切分速度,这也是应考虑的因素。

5.3 功能完备性

自动分词方法除了完成分词功能外,还应具备词库增删、修改、查询和批处理等功能。

5.4 易扩充性和可维护性

这是提供数据存储和计算功能扩充要求的软件属性,包括词库的存储结构,输入/输出形式的变化等方面的扩展和完善。这项指标与系统清晰性、模块性、简单性、结构性、完备性以及自描述性等软件质量准则有直接的联系,对于研究实验性质的软件是非常重要的,因为这类软件需要不断提高与改进,使之适应中文信息处理的各种应用。

5.5 可移植性

可移植性是指方法能从一个计算机系统或环境转移到另一个系统或环境的容易程度。一个好的分词方法不应该只能在一个环境下运行,而应该稍作修改便可在另一种环境下运行,使它更便于推广。

6 结论

由于中文的独特性,目前还没有完美的中文分词算法。中文分词算法的进一步完善应该在已经取得的成绩的基础上,综合运用多种方法,并引入新的模型和方法,通过不断探索,使中文分词算法越来越完善。

参考文献:

[1] 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.
[2] 曹桂宏,何丕廉,吴光远,等.中文分词对中文信息检索系统性能的影响[J].计算机工程与应用,2003(19):78-79.
[3] 刘开瑛.中文文本自动分词和标注[M].北京:北京商务印书馆,2000.
[4] Chien Lee-Feng.PA T-tree-based adaptive keyphrase extraction for intelligentChinese information retrieval [J].Information Pro-cessing andManagement,1999(35):501-521.

龙树全(1982-),男,四川阆中人,硕士,主要研究方向:计算机软件与理论,.net 分布式应用程序。

赵正文(1969-),男,博士,教授,主要研究方向:数据库系统实现技术,数据仓库,数据挖掘;
唐华(1983-),男,硕士,主要研究方向:计算机软件与理论。