

文本挖掘中的中文分词算法研究及实现

许高建¹, 胡学钢², 王庆人¹

(1. 安徽农业大学 信息与计算机学院, 安徽 合肥 230036;

2. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009)

摘要:文本挖掘是指使用数据挖掘技术, 自动地从文本数据中发现和提取独立于用户信息需求的文档集中的隐含知识。而中文文本数据的获得是依靠中文信息处理技术来进行的, 因而自动分词成为中文信息处理中的基础课题。对于海量信息处理的应用, 分词的速度是极为重要的, 对整个系统的效率有很大的影响。分析了几种常见的分词方法, 设计了一个基于正向最大匹配法的中文自动分词系统。为了提高分词的精度, 对加强歧义消除和词语优化的算法进行了研究处理。

关键词:中文分词; 歧义消除; 最大匹配; 词语优化

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2007)12-0122-03

Research and Realization of Chinese Text Classification Algorithms on Text Mining

XU Gao-jian¹, HU Xue-gang², WANG Qing-ren¹

(1. School of Information & Technology, Anhui Agricultural University, Hefei 230036, China;

2. School of Computer & Information, Hefei University of Technology, Hefei 230009, China)

Abstract: Text mining uses the data mining technique to find and extract the cryptic knowledge automatically from text files, which is self-existent the information users needed. Chinese text data is achieved by Chinese information handling. So text participle is a basic question for discussion on Chinese information handling. The rate of text participle is most important especially in applied in great information handling, and it affects the efficiency of whole system. This paper analyzes some ways in text participle, and designed a Chinese-text-participle-system based on most-matching from left to right. In order to improve the participle precision, the algorithms of eliminating different meanings and words optimization are dealt with.

Key words: Chinese text participle; different meanings eliminating; most matching; word optimization

0 引言

随着 Internet 技术的飞速发展, 互联网上的信息缤纷繁多, 如何让用户快速准确搜索定位到自己所需要的资源, 越来越成为各行业关注的焦点。而中文分词技术的停滞不前是一直阻碍中文搜索质量提高的至关重要的因素。对于中文来说, 词是承载语义的最小单位^[1], 这就好像把英文单词之间的空格都去掉, 人们看到的是一片没有意义的字母。因此, 中文自动分词就成为中文知识管理系统必须解决的问题。单个汉字一般很难单独表达一定的含义, 而中文的词是没有自然分隔符的, 需要采取一定的技术手段将词准确地分离出来, 况且单个的字在句子中有时可以与其前后都

能组成词, 在不同的语言环境下, 它的组词方式也不同。可是对于西文来说, 基本上不用经过分词就可以直接进入检索技术、短语划分、语义分析等更高一层的技术领域, 而对于中文, 只有越过这个技术瓶颈问题, 分词的准确率足够高、分词速度足够快, 中文的信息处理技术才有可能和西文的信息处理技术相媲美。

1 几种中文分词算法

以下为几种中文分词算法^[2]。

1) 基于字符串匹配的分词方法。

这种方法又叫做机械分词方法, 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)。按照扫描方向的不同, 串匹配分词方法可以分为正向匹配和逆向匹配; 按照不同长度优先匹配的情况, 可以分为最大 (最长) 匹配和最小

收稿日期: 2007-02-04

基金项目: 安徽省科技计划项目 (2007ZD-7021010)

作者简介: 许高建 (1974-), 男, 安徽肥东人, 讲师, 研究方向为计算机应用、文本挖掘; 胡学钢, 教授, 研究方向为人工智能、数据挖掘。

(最短)匹配;按照是否与词性标注过程相结合,又可以分为单纯分词方法和分词与标注相结合的一体化方法。常用的几种机械分词方法如下:

(1)正向最大匹配法(由左到右的方向)。

比如:“我是大学生”。由左至右获取字符,并与词库匹配。将“我”作为首字符,与词库匹配,若成功,则继续获取字符,即将“我是”与词库匹配;若失败,则将“我”作为一个词切分开来,并将“是”作为首字符与词库匹配;反之,继续获取字符并与词库匹配。

(2)逆向最大匹配法(由右到左的方向)。设 D 为词库, MAX 为 D 中最大词长, Str 为待切分字符串。由右至左地从 Str 里获取长度为 MAX 的字符串,与词库匹配。若成功,则作为一个词切分开来;反之,减去一个字符继续与词库匹配。

(3)最少切分(使每一句中切出的词数最小)。

首先扫描字符串,在待分析字符串中识别和切分出一些带有明显特征的词,比如:“的”、“了”、“着”等,以这些词作为断点,可将原字符串分为较小的串再进行机械分词,从而减少匹配的误差率。

还可以将上述各种方法相互组合,例如,可以将正向最大匹配方法和逆向最大匹配方法结合起来构成双向匹配法。由于汉语单字成词的特点,正向最小匹配和逆向最小匹配一般很少使用。一般说来,逆向匹配的切分精度略高于正向匹配,遇到的歧义现象也较少。实际使用的分词系统,都是把机械分词作为一种初分手段,还需通过利用各种其它的语言信息来进一步提高切分的准确率。

2)基于理解的分词方法。

这种分词方法是通过人工对句子的语法进行定义,当计算机接收到一个句子,以标点符号为分隔符,首先判断它属于哪种类型的句子,模拟人对句子的理解,达到识别词的效果。这种分词方法需要使用大量的语言知识和信息。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。

3)基于统计的分词方法。

从形式上看,词是稳定的字的组合,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。因此字与字相邻共现的频率或概率能够较好地反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计,计算它们的互现信息。定义两个字的互现信息,计算两个汉字 X、Y 的相邻共现概率。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,便可认为此字组可能构成了一个词。这种方法只需对语料中的字组频度进

行统计,不需要切分词典,因而又叫做无词典分词法或统计取词方法。但这种方法也有一定的局限性,会经常抽出一些出现频度高、但并不是词的常用字组,例如“这一”、“之一”、“有的”、“我的”、“许多的”等,并且对常用词的识别精度差、时空开销大。

2 基于正向最大匹配的分词研究及实现

2.1 正向最大匹配法分词特点

在目前阶段,单纯的理解式的切分方法还不成熟,尚处于研究阶段;而机械匹配式的切分方法已经相当成熟,其中的代表算法——最大匹配法已经被国内很多研究机构所采用^[3]。但其局限也是很明显的:效率和准确性受到词库容量的约束;对于歧义切分无法有效地克服。文中在后面也对歧义消除作了简单的探讨。

2.2 正向最大匹配法分词的实现

分词系统可包含两大模块:分词算法、分词词库,如图 1 所示。当待切分文本输入后,分词算法调用分词词库查询词语,而分词词库则向分词算法反馈该词语是否存在于词库中。其中,若是词库中未出现的字,作为一个词处理。

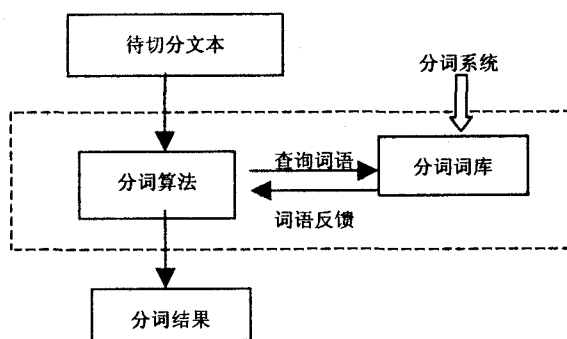


图 1 分词系统结构

2.2.1 算法实现

$k = \text{Len}(\text{Str})$; $n = 1$; $\text{Str1} = \text{Left}(\text{Str}, n)$ // k 记录待切分文本长度, Str1 正向获取字符, n 为计数器

$\text{out1: strsql} = \text{"select * from ciku where ciyu = " \& Str1 \& " ;"}$

$\text{DataEnv.rsciku.Open strsql}$ // 与词库匹配

If $\text{DataEnv.rsciku.EOF}$ Then // 匹配失败

If $\text{Len}(\text{Str1}) = 1$ Then // 该词为单个字

$\text{Strprint} = \text{Strprint} \& \text{Str1}$

继续获取字符并与词库匹配

Else

$\text{Strprint} = \text{Strprint} \& \text{Str2}$ // 该词为多个字

继续获取字符并与词库匹配

End If

Else

$\text{Str2} = \text{Str1}$; $n = n + 1$; $\text{Str1} = \text{Left}(\text{Str}, n)$ // 匹配成功

继续获取字符并与词库匹配

End If

打印 //待切分文本的词语已经切分完

2.2.2 分词结果

从图 2 中可以看出,句子得到了分隔,但是,第二句的切分出现了歧义:应该切分为“大学 生活 丰富”,可是计算机却切分成“大学生 活 丰富”。所以,得再添加一些算法来最大限度地消除歧义。

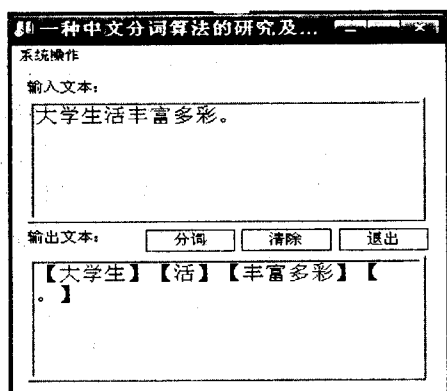


图 2 分词结果显示

3 歧义消除

3.1 歧义的出现

汉语博大精深,十分复杂,在不同语言环境下对句子的理解就可能不同,所以由计算机依据词库切分的词语往往会产生歧义,这就需要插入一些算法最大限度地消除歧义。如上文提到的“大学生生活丰富”,计算机切分结果产生了错误,对此,可以运用“退一字组合法”来消除这类似的歧义。

3.2 歧义消除算法

3.2.1 “退一字组合法”的算法

Strprint 存储已经切分的词语

If Len(Str1)=1 Then //若匹配成功的词为单个字

Str2 = Right(Strprint) //取出最后一个字

Str3 = Str2 & Str1

strsql = "select * from ciku where ciyu = '" & Str3 & "'"

DataEnv.rsciku.Open strsql //与词库匹配

If DataEnv.rsciku.EOF Then //失败

Strprint = Strprint & Str1

Else //成功

k = Len(Strprint):k = k - 1

Strprint = Left(Strprint,k) //Strprint 里去除 Str2 里的字

Strprint = Strprint & Str3

End If

End If

3.2.2 歧义消除结果

尽管“退一字组合法”消除了类似“大学生生活丰富

多彩”这样的简单歧义,但依然存在着问题:对于“他还不了解手的作用”这句话,分词系统在切分过程中,调用了“退一字组合法”程序段却产生了歧义(如图 3 所示)。那么如何消除类似这样的歧义呢?所以,就得使用下文的“词语优化”来消除这类似的歧义。

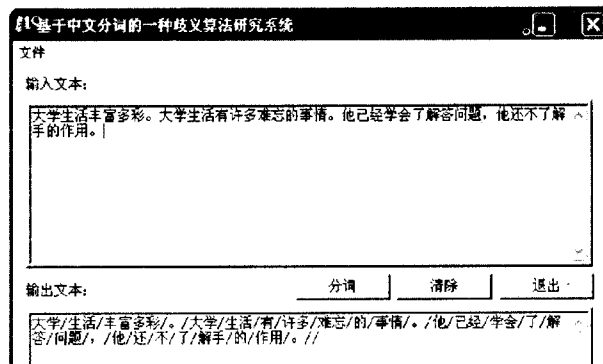


图 3 歧义消除显示

3.3 词语优化

3.3.1 词语优化概念

“词语优化”可以用一句话概括:“词库承认则合成,反之则切分”。所谓“词库承认”就是在分词库中可以查询到的。利用“词语优化”可以有效地处理一些合成词,如“曼联”、“国际米兰”等;当然也有局限性,就是过于依赖分词库。

3.3.2 词语优化算法

//Strprint 存储已经切分的词语

(由左至右从 Strprint 里寻找词语)

Do

Load(Str1) //寻找到第一个词语,存入 Str1 里

Load(Str2) //寻找到第二个词语,存入 Str2 里

Str3 = Str1 & Str2

与词库匹配

If 词库承认 Then

从 Strprint 里去除 Str1、Str2

Strprint = Str3 & Strprint

Else

Strprint1 = Strprint1 & Str1

从 Strprint 里去除 Str1

End If

While(Strprint 里词语查询完成)

3.3.3 优化结果

经过上述算法处理后的结果如图 4 所示,可以看出,计算机这时切分完全正确。

4 结束语

中文分词主要应用于信息检索^[4]、机器翻译等多方面,利用它可以将互联网上的文章进行归类,进而方

(下转第 172 页)

用户的权限,使得一般用户不能随意对数据进行操作,不能对一些保密数据进行查看,系统的启动画面如图 3 所示。

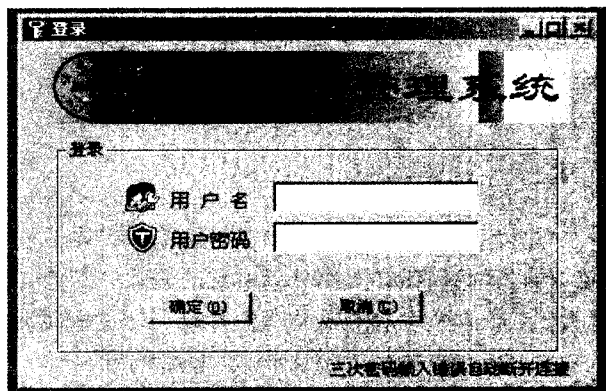


图 3 系统的启动画面

6 系统测试

测试是为了发现程序中的错误而执行程序的过程,好的测试方案是极可能发现迄今为止尚未发现错误的测试方案。测试任何产品都有两种方法:如果已经知道了产品应该具有的功能,可以通过测试来检验是否每个功能都能正常使用;如果知道产品的内部工作过程,可以通过测试来检验产品的内部动作是否按照规格说明书的规定正常进行。前一种方法称黑盒测试,后一种方法称为白盒测试。本软件测试采用黑盒测试的方法,仅对系统功能进行测试。大型软件测试步骤分为模块测试、子系统测试、系统测试、验收测试

和并行测试。本软件直接进行验收测试。系统测试结果显示本系统能够实现系统模块所包含的所有功能,具有界面友好、易于操作、简便实用的特点。

7 结束语

本系统界面友好,用户用鼠标点击即可实现各项功能。整个软件在开发过程中使用模块化程序设计方法,充分考虑了各种实际情况,显示出良好的实用性。该系统能够为教师和学生工作管理人员提供快捷的管理学生成绩的手段,取代了人们长期以来所使用的单纯的人工管理方式,提高了工作效率,也对学生工作管理的科学化、正规化起到了重要的推动作用。

参考文献:

- [1] 曹建军,刘永娟,刘咏梅,等.基于 VB 计算机多媒体技术试题库管理系统开发[J].计算机技术与发展,2006,16(9):154-156.
- [2] 丁建丽,韩清,孙丽.基于 VB 的中小型书店图书管理系统的设计与实现[J].新疆大学学报:自然科学版,2003,20(1):46-49.
- [3] 吴涵.基于 VC++ 的研究生信息管理系统的设计与实现[J].计算机技术与发展,2006,16(12):184-186.
- [4] 刘艳菊,陈伟,耿蕊.基于 VB 的文档管理系统的设计与实现[J].齐齐哈尔大学学报,2004,20(3):72-73.
- [5] 秦乐乐,蒋佳,崔连生.基于 VB 的学生信息管理系统的设计与实现[J].河北工业科技,2006,23(4):206-209.

(上接第 124 页)

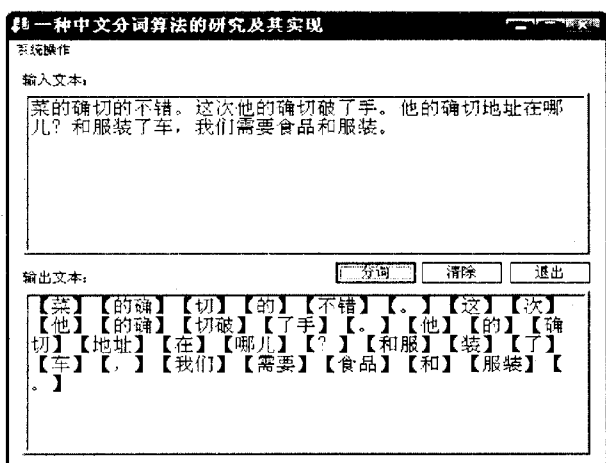


图 4 词语优化结果显示

便人们的学习、工作和查找资料。实际应用的统计分词系统都要使用一部基本的分词词典(常用词词典)进行串匹配分词,同时使用统计方法识别一些新的词,即将串频统计和串匹配结合起来,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^[5]。

关于文中介绍的中文分词理论及算法,依然存在着问题。如:无法处理新词的识别与录入、由 VB 语言编写且数据库为 Access 导致分词速度比较慢(平均 103 字/秒)。这有待于进一步的深入研究和探索。

参考文献:

- [1] 李振星,徐泽平,唐卫清.全二分最大匹配快速分词算法[J].计算机工程与应用,2002,38(11):106-109.
- [2] 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.
- [3] 张滨,晏蒲柳,李文翔,等.基于汉语句模的中文分词算法[J].计算机工程,2004(1):134-135.
- [4] 耿骞,毛瑞.汉语自然语言检索中的词法分析处理[J].情报科学,2004(4):466-499.
- [5] 蒋澄,马范援,蒋思杰.中英文 WWW 搜索引擎的信息处理[J].计算机工程,1999,25(4):372-381.