文章编号:1671 - 6833(2014)05 - 0044 - 05

一种改进的中文分词在主题搜索中的应用

许智宏, 张月梅, 王 一

(河北工业大学 计算机科学与软件学院, 天津 300401)

摘 要: 主题搜索的核心内容是以中文分词为基础的内容匹配,而中文分词的准确性以及对未登记词的识别率问题仍是目前主题搜索的瓶颈.提出了一种改进最大匹配中文分词算法 IMMM,通过词库预处理、未登录词处理和歧义消除等策略,并将主题分类和分词词典的存储相结合,构造了一个主题搜索系统.实验证明,改进后的算法较传统的搜索算法在搜索准确率方面有了较大的改进,系统整体搜索效率有明显提高.

关键词: 最大匹配;主题搜索;词库;中文分词

中图分类号: TP3 **文献标志码:** A doi:10.3969/j.issn.1671 - 6833.2014.05.011

0 引言

随着网络技术的发展、Web2.0 时代的到来以及信息需求的变化,互联网上每天都有海量信息生成、共享和更新,通用搜索无论是软件还是硬件都面临着巨大的挑战.为了解决这一问题,人们提出了基于主题的搜索技术[1-3].它将重点放在特定领域的信息挖掘上,它的出现为人们获取准确的 Web 信息提供了方便.信息检索一般是基于关键词索引技术,以词作为索引必须首先切分出单个的中文词语,这样就到达了查询数据量广的优点,但是难于表达丰富的信息量却成了查询的瓶颈.所以中文分词的准确性、正确性直接影响到查询结果.

中文分词指的是将一个汉字序列切分成一个个单词的过程[4].传统的中文分词算法分为三类:基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法[5].这三类分词方法有着各自的特点和不足.第一类分词法是按照一定的策略将汉字字符串与一个词典中的词进行匹配[6].第二类分词法通过让计算机模拟人对句子的理解,由于汉语语言知识的笼统、复杂,难以将各种语言信息组织成机器可直接读取的形式,所以目前基于理解的分词系统还处于初始阶段.第三类分词法所统计的对象是多元的.最常见的是基于字与字之间的结合频率[7]来决定.

1 中文分词在主题搜索中的应用

1.1 中文分词在主题搜索中的应用

图 1 所示为整个搜索引擎的架构,其中索引建立与查询系统主要是中文分词和文档检索.用户输入关键词或语句,经过中文分词,将分词的结果作为查询的依据,对爬行回来的文档或网页进行检索,最终返回结果.

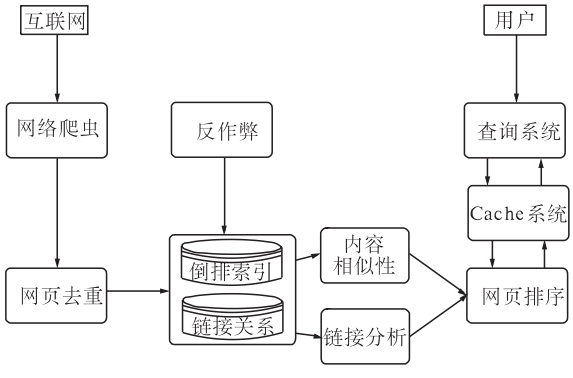

图 1 搜索引擎整体架构

Fig. 1 The overall architecture of search engine

在搜索引擎中使用的分词方法,有其特定的要求[8].①时间开销不能太大;②在准确率和时间开销上找到一个平衡点;③分词的切分粒度应该尽可能使长词优先匹配;④要求识别出来的未登记词要尽可能的准确;⑤搜索引擎的分词使用

收稿日期:2014 - 05 - 27;**修订日期:**2014 - 07 - 17

基金项目:河北省高等学校科学技术研究青年基金项目(20111122)

作者简介:许智宏(1970 -),女,河北张家口人,河北工业大学副教授,博士,主要从事分布式计算研究,E-mail:xuzhihong@scse.hebut.edu.cn.

的词典要符合互联网用户词语习惯;⑥搜索引擎分词在处理网页文本和用户查询时需要使用同样的分词方法.

搜索引擎对分词的特定要求使得一般的中文分词算法并不能很好地提升搜索引擎的效率.综合传统的三类中文分词算法的思想,作者将实现基于字符串匹配的分词方法的改进.

2　基于字符串匹配分词算法机制分析

2.1　正向、逆向最大匹配算法

正向最大匹配算法 FMMM 的基本思路是在计算机中存放一个已知的词表.假设词表中有最长长度为 N 的字符,那么被处理文档的前 N 个字符就被作为匹配字符串 z,令 z 和词表中的词条进行依次对比,如果词表中存在词条与 z 相同,那么匹配成功,z 便作为一个词被切分出去,假如没有匹配成功,就将 z 从末尾处减去一个字符,组成新的 z,再次与词表中的词条进行匹配,如果匹配成功,则切分出去,否则继续减少 z 的字符,重复此过程,直至匹配成功[9].

逆向最大匹配算法 RMMM 基本策略与正向最大匹配算法的思路基本相同,不同的是匹配的方向不同,前者是自左向右,而后者是自右向左.

2.2　最大匹配算法分析

来自统计结果显示利用正向最大匹配的错误率为 1/169,逆向最大匹配方法的错误率为 1/245[10].从结果来看错误率貌似很低,但是就是这样的错误率在现实生活中还是不能被接受的,所以必须对错误率进行降低.由结果可知逆向最大匹配的错误率要小于正向最大匹配,但是由于逆向词表的建立与维护都存在一定的难度[11],所以逆向最大匹配并不能很好地应用于实际操作中.

3　最大匹配算法的改进及应用

3.1　查询系统构建

查询系统是主题搜索的重要组成部分,也是中文分词的重要应用领域,中文分词的改进主要将在其中发挥效应,重新构建的系统共可分为 3 个大模块.

(1)信息树词表的构建、更新与维护.为了实现高效率的中文分词,构建一个高效的词表是基础前提,本算法的实现也是基于词表的良好维护与更新.

(2)最大匹配.应用最大匹配对用户的输入的字符串进行词条匹配,其中对匹配成功与匹配失败的词条进行处理.

(3)未登记词的判断与处理.此功能模块主要对上以模块匹配失败的词条进行记录、存储,并当再次出现时进行频率更新,更新信息树词表.

3.2　词表的构建、更新与维护

为了构建高效的词表,必须考虑 3 个方面:对词典查询速度;词典存储利用率;词典维护的效率.主题搜索的主旨即根据用户定制的主题内容进行搜索,所以在构建词表时将词表的内容分类储存,本算法中词表分为 2 个主要模块,3 个层次:第一个模块是主题分类,第二个模块(未登记词条)是第一个模块中不存在词条.他们的关系如图 2 所示.

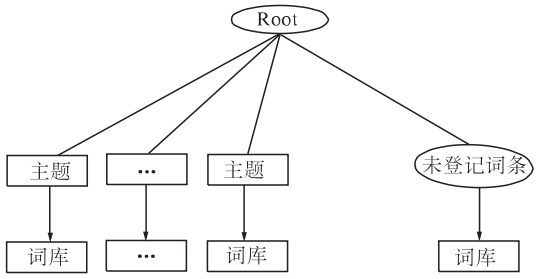

图 2　主题与词库关系图

Fig. 2　Figure between theme and thesaurus

笔者主要研究中文分词在主题搜索中的应用,所以为搜索的每个分类建立了相应的词库,词库的结构如图 3 所示,每个分类下的词条存储都改变以往的存储方式,将该主题(相关主题信息)下能组成词汇的汉字建立索引表,按字典中 26 个英文字母排序排列,每个字组成的词汇按照词汇再次按照索引顺序排序[12].$W(i,1,1)\,W(i,1,2)\cdots W(i,1,k)\cdots W(i,j,1)\,W(i,j,2)\cdots W(i,j,k)$,第一个值 i 是第一个汉字在索引中的顺序,第二个值 j 表示该词在以该汉字开头的词条中长度为 j,第三个值 k 表示在长度为 j 的词条中的位置.

对于未登记词条中词条的排序按主题分类下词库的构建进行,未登记词条词库的结构在每一词条增加频率属性 F,词库表中属性的设置为:$W(i,1,1,F)\,W(i,1,2,F)\cdots W(i,1,k,F)\cdots W(i,j,1,F)\,W(i,j,2,F)\cdots W(i,j,k,F)$,便于在判断时作为参考.例如第一次分词中出现一个未能分词成功的词“工大”,那么未登记词库中增加词条“工大”,并且初始化“工大”的词频 F 等于 1,如果以后的分词过程中再次出现未能分词成功的词中含

有“工大”,则“工大”的词频 F 加1,变为2,由此类推,当某个词的词频到达一定的阈值T时,将其加入到正式词库中.

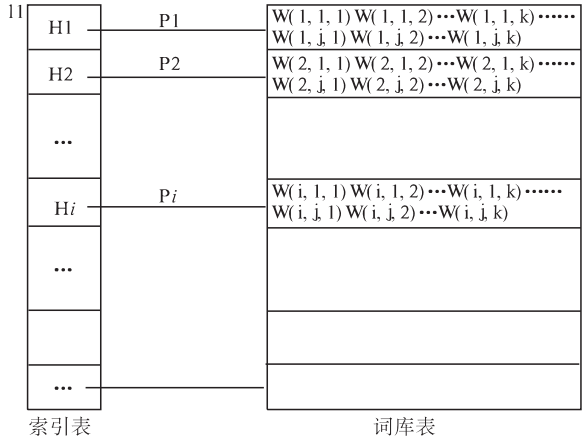

图3 词库结构图

Fig.3 Structure of thesaurus

3.3 改进的最大匹配

由于中文中多数词是由两个汉字以上组成的,所以本算法IMMM将在此基础上进行改进.算法描述:分词时(待分词词长必须大于2).首先根据待分词的词串开头汉字找到以该字开头的所有词组,然后与待分词词串的前两个汉字匹配,如果匹配成功,则按照词长按照由长到短的顺序检索词库,再和待分词词串匹配,如果再次匹配成功,则切分为一个词,然后继续匹配下一个词;如果匹配失败,将待匹配词串从尾部减少一个字重新匹配.如果前两个汉字匹配失败,则可以直接宣布分词失败,将第一个汉字单独进行存储,然后以第二个汉字为开头,再次进行匹配,直至待分词词串分词结束.这样将匹配次数的大大减少,缩小了匹配时间.正确率、效率都相应的提高.

IMMM算法:

```
Int i;
While(还有下一个词)//
若还有下一个词,则继续分词 ……..①
{
Search(当前词中第一个汉字);
//检索出当前词中第一个汉字开头的词组
HeadMatch(当前词前两个汉字)//匹配前两个汉字
If (success){//如果成功
AgSearch()//检索出当前词中前两个汉字开头的词组
For(i=当前词长度;i>0;i--)
//由词长按由长到短的顺序检索词库
{
If Match(当前词)//匹配成功
{
Split(当前词);//切分出一个词
Get(下一个词语);//取下一个词语
Break;
}
Else{
Search(当前词中前两个汉字)
//检索出当前词中前两个汉字开头的词组
当前次长度减1;
}
}
Else{
分词失败
Minus(待分词)//去除第一个汉字
Goto ①
}
}
```

3.4 未登记词的判断与处理

如果分词后出现剩余词条,首次出现的词会被存储在一个中间词库中,并为词库中的每个词条设置一个词频属性,如果下次分词后剩余词条中依然出现中间词库中的词,则将这个词的词频数加1,当词频数等于某一个设定的值时,自动将这个词添加至大词库的未登记词库中,未登记词库中的词条的词频数也是根据搜索匹配结果的次数进行增加,到达一定值后根据所在主题分类将其添加至主题分类下的此表中,不断更新词库系统.

4 IMMM算法验证

笔者根据中文算法的设计实现了基于ictclas4j(java修改版)的分词系统与Nutch搜索相结合的主题搜索.ictclas4j(java修改版)的分词系统的中分词方法改为本文所设计算法IMMM,经过对改进的ictclas4j系统分词效果的测试显示,系统的分词结果与准备好的结果相似度很高,分词的正确率到达了92.3%.

为满足用户的使用习惯,主题搜索系统的词库是基于搜狗《2012细胞词库》构建的,主题分为城市、自然科学、人文科学、社会科学、工程应用科学、娱乐、医学、艺术、生活、运动休闲、农林渔畜、电子游戏等12大类,基本上概括了目前的各个行

业领域,词条数量达 30 多万条. 系统首先通过分词系统 ictclas4j 实现分词处理,然后将分词结果在 Nutch 搜索结果中进行索引检索,返回结果.

4.1　实验结果分析

由于不同的搜索引擎采取不同的爬虫、分词方法,并不能笼统的说那个算法在搜索引擎中性能好,当前百度搜索引擎分词主要采用双向最大匹配算法 BMMM,google 采用的是逆向最大匹配算法 RMMM[13]. 为此我们将主要搜索引擎中的分词算法应用在相同的环境中,得出的结果如表 1 所示.

表 1　部分主题词搜索返回结果

Tab.1　Part returns results of the keyword search

分词算法	主题词数量/个	搜索测试次数	平均返回结果	最差准确率/%	最高准确率/%	平均准确率/%
BMMM	100	200	9 186	85.51	97.30	91.45
Nutch	100	200	7 820	80.54	93.81	87.15
RMMM	100	200	8 768	83.82	96.43	90.10
IMMM	100	200	9 223	85.30	98.12	91.71

实验采用相同的实验环境对不同的算法进行了对比,表 1 中的结果表明,百度采用的 BMMM 算法要比 google 的 RMMM 以及 Nutch 的要好,经过我们改进的最大匹配算法 IMMM 与 BMMM 算法相差并不太多,从结果来看分词的正确率还有 8.29% 的错误,原因可能是词表中并不能真正的包含所有词汇,而且对词表的修复需要系统的长时间使用,才能实现未登记词对词表的更新. 所以经过长时间的词表更新与维护系统的正确率必将得到很好的提高. 整个实验过程顺利并且达到了预期的目的.

5　结论

笔者为主题搜索构建了信息树的词表,使得主题搜索中的搜索效率显著提高,在此基础上实现了最大匹配的准确实现,并在算法的实现过程中对已建立的词表进行不断自我完善.

同时,算法也存在一些不足的地方,中文的结构相比于英文要复杂的多,而主题搜索的词表建立过程中并不能真正准确的实现分类,这使分词的匹配难免出现错误,随着算法的不断运用,词表中词量将不断增多,匹配过程将变得耗时、低效,因此词库在自我完善的同时需要进行维护、更新,而这也将是需要进一步研究的问题.

参考文献:

[1] CHO J. Crawling the web:Discovery and maintenance of large—scale Web data[D]:PhD. Thesis Stanford University,2001.

[2] 王新,刘晓霞. 基于关联规则挖掘的垂直元搜索引擎研究[J]. 计算机工程,2011,37(4):76 - 77,80.

[3] 王旭仁,杨硕,宋蓓. 房地产垂直搜索引擎的设计和实现[C]. 2011 年信息技术、服务科学与工程管理国际学术会议,2011:174 - 177.

[4] 何嘉. 基于遗传算法优化的中文分词研究[D]西安:电子科技大学电子系,2012.

[5] 付年均,彭昌水,王慰. 中文分词技术及其实现[J]. 软件导刊,2011,10(1):18 - 21.

[6] 万建成,杨春花. 书面汉语的全切分分词算法模型[J]. 小型微型计算机系统,2006,24(7):1247 - 1251.

[7] 钱揖丽,郑家恒. 文本切分知识获取及其应用[J]计算机工程与应用,2003(2):63 - 64.

[8] 张敏. Web 文本信息检索的方法研究[D] 北京:清华大学,2003,07.

[9] 黄魏,高兵,刘异,等. 基于词条组合的中文文本分类方法[J]. 科学技术与工程,2010,10(1):85 - 88.

[10] COME D E. Internetworking With TCP/IP Vol Ⅰ: Principles, Protocols, and Architectures 5th Edition [M]. 2005.

[11] 何国斌,赵晶璐. 基于最大匹配的中文分词概率算法研究[J]. 计算机工程,2010,36(5):173 - 175.

[12] 闻玉彪,贾时银,邓世坤,等. 一种改进的最大匹配算法中文分词算法[J]. 计算机技术与发展,2011,21(10):92 - 94.

[13] 周满英. 百度和谷歌的中文分词技术浅析[J]. 中国索引,2011(2):44 - 46.

Application of an Improved Chinese Word Segmentation Technology in Topic Search

XU Zhi-hong, ZHANG Yue-mei, WANG Yi

(Hebei University of Technology, School of Computer of Science and Engineering, Tianjin 300401, China)

Abstract: The topic search's core is the contents of the match which is based on the Chinese word segmentation, but the Chinese word segmentation's accuracy and unregistered word's recognition is still the bottleneck of the topic search. This paper proposed an improved maximum matching of word segmentation algorithm IMMM. In improved algorithm designed thesaurus pretreatment, unknown word processing and disambiguation strategy, combined subject categories and sub-word dictionary storage, finally construct a topic search system. The algorithm results show that the improved algorithm is better than traditional algorithms, and the search accuracy rate has been greatly improved. The system's efficiency is improved.

Key words: maximum matching; topic search; thesaurus; Chinese word segmentation

---

(上接第43页)

[10] KIANI M, GHOVANLOO M. The circuit theory behind coupled-mode magnetic resonance-based wireless power transmission [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2012, 59(8): 1 – 10.

Two Modeling Methods Equivalent Analysis of the Magnetic Coupling Power Transmission System

LI An-xin[1], ZHANG Jiang-fei[1], ZHANG Zu-long[2]

(1. SIPPR Engineering Group Co., Ltd, Zhengzhou 450000, China; 2. Fuyang City Power Supply Corporation of State Grid, Fuyang 236001, China)

Abstract: With the magnetic resonant coupling wireless power transfer system as the research object in this paper, the magnetic resonance system dynamic model is made by using coupling mode theory, from the energy field coupling and decay characteristic angle explains the circuit meaning of the key parameters, and then the system output power and transmission efficiency expression are derived in this paper. Through contrast analysis of the output power and transmission efficiency expression which is derived by equivalent circuit method, it is concluded that the two kinds of modeling method with consistency conclusion as to the same coupling system. Finally, the quantitative relationships between the output power and coupling coefficient and between the transfer efficiency and coupling coefficient are given, which verify the correctness of the theoretical analysis results, and the consistency of two kinds of modeling method in the output power and transmission efficiency is verified indirectly.

Key words: magnetic resonant; wireless power transmission; equivalent circuit; couple mode theory; coupling coefficient