农业搜索引擎中文分词工具对比①

赵 涛, 张太红

(新疆农业大学 计算机与信息工程学院, 乌鲁木齐 830052)

摘 要: 中文分词是中文语言处理技术中最关键的部分, 被作为其中最重要的预处理手段广泛应用. 本文主要分析和比较了 ASPSeek、ICTCLAS、Paoding、MMseg、IK 与 JE 六种分词工具对农业搜索引擎的影响. 结果表明: 在农业搜索引擎的应用效果中, 分词准确性最优的是 ICTCLAS 分词工具, 而 F1 测度最高的是 JE 分词工具.

关键词: 中文分词; 分词工具; 农业搜索引擎

Compare about Chinese Word Segmentation Tools to Agricultural Search Engine

ZHAO Tao, ZHANG Tai-Hong

(College of Computer & Information Engineering, Xinjiang Agricultural University, Urumqi 890052, China)

Abstract: Chinese word segmentation technology is the most critical part of Chinese word processing technology, and it is widely used as one of the most important part in pre-processing methods. This paper mainly analyses and compares the influence of ASPSeek, ICTCLAS, Paoding, MMseg, IK and JE six segmentation tools on agricultural search engine. The results showed that ICTCLAS word segmentation has the most optimization in accuracy, while JE word segmentation has the highest measure of F1 in the application effect of agricultural search engine.

Key words: Chinese word segmentation; word segmentation tools; agricultural search engine

信息化建设的飞速发展, 使得互联网上的信息迅速增长. 为了能够及时准确的获取网页上的信息, 搜索引擎便成为人们快速查找信息和资源的重要手段. 但目前的搜索引擎主要采用基于关键字的查询, 而关键字的简单组合不能明确表述用户的查询意图, 这一问题已成为制约搜索引擎性能提高的瓶颈之一. 由于汉语本身的特点, 必须引入对于中文语言的处理技术, 而中文分词技术就是其中很关键的部分. 目前为止, 还没有完全正确的分词技术, 网络是由无数张网页组成, 其内容无比庞大, 对分词方法的要求就更高. 那么, 这一影响究竟有多大, 中文分词是不是提高搜索引擎性能的关键呢? 这正是本文研究的重点[1].

1 中文分词对搜索引擎的作用

通过近些年的发展, 互联网时刻伴随在我们身边. 网上的信息量也在急剧膨胀, 在这海量的信息中, 各类信息混杂在一起, 要想充分利用这些信息资源就要对它们进行整理, 如果由人纯粹的来做这项工作, 已经是不可能的. 在自然语言处理技术中, 英文是以词为单位的, 词与词之间上靠空格隔开, 而中文是以字为单位, 句子中所有的字连起来才能描述一个意思. 因此, 中文处理技术比西文处理技术相对较难或发展的较晚, 许多西文的处理方法中文不能直接采用, 就是因为中文必需有分词这道工序. 中文分词是其他中文信息处理的基础. 因此, 对于搜索引擎来说, 最重要的并不是找到所有结果, 而是在上百亿的网页中把最相关的结果找到, 并排在最前面, 这也称为相关度排序. 中文分词的准确与否, 常常直接影响到对搜索结果的相关度排序. 分词准确性对搜索引擎来说就十分重要, 但如果分词速度太慢, 即使准确性再高, 对于搜索引擎来说也是不可用的, 因为搜索引擎需要处理数以亿计的网页, 如果分词耗用的时间过长, 会严

① 基金项目:新疆自治区高校科研技术项目(XJEDU2013S13)
收稿时间:2015-07-01;收到修改稿时间:2015-11-25

重影响搜索引擎内容更新的速度. 因此对于搜索引擎来说, 分词的准确性和速度, 二者都需要达到很高的要求[2]. 由此可见, 中文分词的性能对搜索引擎结果的相关性和准确性有相当大的关系.

2 农业搜索引擎简介

随着社会的发展, 人们对信息的要求越来越高, 只是百度、Google、搜狐等综合性搜索引擎不能满足人们各方面的需求, 便出现了垂直搜索引擎, 垂直搜索引擎就是向更加专业化、领域化的方向发展, 随即农业搜索引擎也得到了一系列的发展. 农业搜索引擎属于垂直搜索引擎, 主要为搜索农业信息而开发的检索工具, 专门提供农业信息, 比综合性搜索引擎在解决实际问题时更有效.

2.1 农业搜索引擎的基本原理

农业搜索引擎的基本原理同一般的搜索引擎基本相似, 包括信息的采集、信息的预处理及信息的检索[4]. 唯独不同之处在于农业搜索引擎建立的数据库是跟农业有关的. 信息的采集是通过网络蜘蛛爬虫对互联网上的相关站点进行访问, 然后对抓回的网页进行分析、过滤、和存储, 并对这些信息建立索引. 最后根据用户的要求, 对索引数据库进行访问, 并把检索的结果返回给用户[6].

2.2 农业搜索引擎的发展

2.2.1 国外农业搜索引擎的发展

20 世纪 50 年代到 60 年代, 农业信息化建设开始发展, 80 年代到 90 年代得到了快速发展. 目前农业搜索引擎朝着多元化发展, 如美国农业网络信息中心是由美国国家农业图书馆与一些大学、研究机构及政府机构资源组合而成的, 农业信息服务都是由他们中的每一个成员负责其中的一个部分, 各成员之间也相互提供信息及享受信息. 还有法国的 WEB. AGRISEARCH, 它提供了三种服务: 农业搜索引擎、农业期刊导航和农业站点导航. Agrisurf Search 是由美国一家农业搜索引擎服务的公司从综合搜索引擎中解脱出来的专门提供农业信息的网站, 另外, 此网站还提供农业新闻类与政策类测信息[3]. 国外农业搜索引擎的出现与发展, 为我国农业搜索引擎的发展奠定了良好的基础.

2.2.2 国内农业搜索引擎的发展

目前国内的农业搜索引擎也得到了快速发展, 如“农搜”是全世界数据量最大的汉语农业搜索引擎. 搜农, 是面向农民大户、农业企业、农业科技人员及专业技术协会的农业搜索引擎. 它更加与农业用户的需求相符合. 还有很多如华农在线、中国农业科技信息网农业网站搜索引擎等. 这些网站都为农业信息检索提供了便利条件[5].

3 中文分词

3.1 什么是中文分词

中文分词就是将连续的字序列按照一定的规范重新组合成词序列的过程. 是文本挖掘的基础.

3.2 中文分词的原理

中文分词的基本原理是针对输入文字串(包含中英文数字标点等)进行分词、过滤处理(包括停用词的处理与标点符号的处理), 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[7]. 中文分词的输入输出如图 1 所示.

图 1 中文分词的输入输出图

3.3 一般分词方法

目前采用的分词方法主要有以下几种: 最大匹配法、反向最大匹配法、逐词遍历法、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、邻接约束方法、邻接知识约束方法、专家系统方法、最少分词词频选择方法、神经网络方法等等. 但归纳起来不外乎三类: 一类是基于字符串匹配的分词法, 一般以分词词典为依据. 通过文档中的汉字串和词表中的词逐一匹配来完成词的切分. 一类是理解式切词法, 即利用汉语的语法知识和语义知识以及心理学知识进行分词, 需要建立囫分词数据库、知识库和推理机; 还有一类是基于统计的分词法[8], 它的基础是利用汉字同时出现来组成有意义词的概率, 可以降低统计对语料库的依赖性, 充分利用已有的词法信息, 同时弥补字符串方法的不足.

3.4 中文分词的难点

中文是一门复杂而又灵活的语言, 可以组成很多词汇, 这让计算机理解中文语言便非常困难. 因此,

在实际应用中, 为了切分出更精确的词, 我们还有两大难题需要解决, 就是歧义处理和未登录词的识别.

3.4.1 歧义处理

分词歧义处理之所以是中文分词的困难之一,原因在于歧义分为多种类型. 目前主要分为交集性歧义、组合型歧义和真歧义三种. 交集型歧义字段数量庞大, 处理方法多样; 组合型歧义字段数量较少, 处理起来相对较难; 而真歧义字段数量更为稀少, 且很难处理. 针对不同的歧义类型应采取不同的解决方法. 除了需要依靠上、下文语义信息、增加语义、语用知识等外部条件外, 还存在难以消解的真歧义, 增加了歧义切分的难度[7]. 同时未登录词中也存在着歧义切分的问题, 这也增加了歧义切分的难度. 所以歧义处理是影响分词系统切分精度的重要因素.

3.4.2 未登录词识别

新词, 专业术语称为未登录词. 也就是那些在字典中都没有收录过的词. 未登录词可以分为专名和非专名两大类. 其中专名包括中国人名、外国译名、地名等. 而非专名包括新词、简称、方言词语、文言词语、行业用词等. 无论是专名还是非专名的未登录词都很难处理, 因为其数量庞大, 又没有相应的规范. 而且随着社会生活的变迁, 使未登录词的数量大大增加, 这又为未登录词的识别增加了难度[7]. 因此, 未登录词识别是中文分词的另一大难点.

3.5 中文分词技术的进展

中文分词算法已经被广泛研究, 分词算法多种多样. 目前, 中文分词效果比较好的并且支持 Java 语言的中文分词软件主要包括 ICTCLAS(中科院中文分词软件)、IK、Paoding(庖丁解牛)、MMSEG4J 等中文分词软件, 基于 C++语言的分词方法也有很多, 如 ASPSeek. 在本次设计中, 我分别使用 ASPSeek、ICTCLAS、Paoding、MMSEG4J、IK 以及 JE 分词工具, 并且对它们的分词效果以及农业搜索引擎分词工具的性能予以评测.

3.5.1ASPSeek 分词工具

ASPSeek 是由 Swsoft 公司(2007 年 12 月, SWsoft 更名为 Parallels)使用 C++编写的免费开源互联网搜索引擎, 使用了 STL 库, ASPSeek 单节点可以处理上百万个 Web 页面并提供检索服务, 可以按短语和单词(允许使用通配符)进行布尔搜索. 搜索结果可以限定在特定的时间域的站点、站点空间, 并按照相关性或者时间进行排序.

ASPSeek 支持多语言编码(包括多字节语言如中文). 它为抓取多个站点进行了优化(实现多线程检索, 同步 DNS 查询, 按站点将结果分组, Web 集合等), 同时它也可以用于单个站点的搜索. 其他特性包括支持停词排除和拼写检查, 字符集和语言的预测, 搜索结果的 HTML 模板, 引用和查询词高亮度显示等[18]. 但是由于 ASPSeek 在抓取网页时对抓好的网页进行了自动分词、建倒排索引, 所以 ASPSeek 也可以用作对中文的分词. ASPSeek 是完全基于词典库的分词方法, 并且拥有装载了 25 万词的词典库.

3.5.2 ICTCLAS 分词工具

ICTClAS 分词系统是由中科院计算所的张华平、刘群所开发的一套分词系统, 这是最早的中文开源分词项目之一, 中科院计算机所的 ICTCLAS 分词系统在 2002 年 7 月举行的“973”项目“图像、语音、自然语言理解与知识挖掘”专家组的评测中, 分词正确率高达 97.58%[9];主要功能包括中文分词、词性标注、命名实体识别、新词识别等, 同时支持用户词典[10], 包含的词典是通过统计方法建立的, 对其进行了封装[7]. 该分词系统的主要是思想是先通过 CHMM(层叠形隐马尔可夫模型)进行分词[11], 通过分层,既增加了分词的准确性,又保证了分词的效率. 共分五层, 基本思路: 先进行原子切分,然后在此基础上进行 N-最短路径粗切分,找出前 N 个最符合的切分结果,生成二元分词表,然后生成分词结果,接着进行词性标注并完成主要分词步骤.

3.5.3 庖丁解牛分词工具

庖丁(Paoding)系统是个完全基于 lucene 的中文分词系统. 庖丁解牛分词模块是将输入的字符串中首先识别和切分出带有明显特征的确定词汇, 以这些词汇为间隔点, 把原输入字符串分割成较小的串再进行词典分词. 为了庖丁解牛分词模块采取了最大减小单纯的匹配错误, 匹配方法和最大切分相结合的方式来分词[12]. 另外庖丁解牛分词系统支持纯文本格式, 一行一词, 使用后台线程检测词库的更新, 自动编译更新过的词库到二进制版本并加载, 具有极高效率和高扩展性.

3.5.4 MMseg 分词工具

MMSEG 是用 Chih-Hao Tsai 的 MMseg 算法实现的中文分词器. MMSEG 是一种基于词典的分词算

法, 以正向最大匹配为主, 多种消除歧义的规则为辅. MMSEG 算法主要分为两种: simple 和 complex. simple 算法就是前面提到的最简单的正向最大匹配算法[17]. 为了解决 simple 算法的不足, MMSEG 又提供了另一种选择: complex 算法. 该算法使用了 Chen K.J.和 Liu S. H.于 1992 年提出的一种最大匹配算法的变种. 这种算法的基本思想是: 找到所有从当前位置开始的三个连续词语的块, 总长度最大的块是最优解.

3.5.5 IK 分词工具

IK Analyzer 是一个开源的, 基于 java 语言开发的轻量级的中文分词工具包. 从 2006 年 12 月推出 1.0 版开始, IK 已经推出了 3 个大版本. 最初, 它是以开源项目 Luence 为应用主体的, 结合词典分词和文法分析算法的中文分词组件, 实现了以词典分词为基础的正反向全切分算法, 是 LuceneAnalyzer 接口的实现. 该算法适合与互联网用户的搜索习惯和企业知识库检索, 用户可以用句子中涵盖的中文词汇搜索.

3.5.6 JE 分词工具

JE 分词是一套由 Java 写的分词软件, 提供了很多功能, 比如提供了设定分词粒度的参数, 即可以设定正向最大匹配的字数、提供了 API 增加了词典的动态扩展能力、整理优化了词库、全面支持 lucene3.0 以下的版本[13].

4 实验方法

本文利用 ASPSeek 搜索引擎抓取了新疆兴农网上 10245 张网页, 它的体系结构包含抓取模块、检索模块, 结果显示模块等部分. ASPseek 首先利用抓取的网页, 建立倒排索引, 并将倒排索引存储到特定的数据库中. 在网页抓取的过程中, Index 程序浏览所有的种子站点, 将种子站点的网页存储到临时文件和数据库中. 当抓取程序完毕后, 用户运行相应的命令(index-D)将存储的数据归并到数据库中. 本实验主要使用 ASPSeek、ICTCLAS、Paoding、MMseg、IK 以及 JE 分词工具对这些网页进行了测试.

4.1 实验流程

4.1.1 分词特性比较

本实验首先通过 ASPSeek 搜索引擎抓取 10245 张网页, 由于 ASPSeek 搜索引擎在抓取网页后, 已经对抓好的网页给出了分词结果, 并且对抓好的网页建了倒排索引, 所以在本实验过程中, ASPSeek 都是自行完成的, 不用人工的对文档进行分词和建立倒排索引, 只需对其结果查看. 本实验流程图 2 主要适用于其余五种分词方法. 对这些抓取好的网页进行预处理, 其中预处理包括 html 一些 tag 标记、标点符号的去除等, 然后用各种分词方法结合 lucene 对处理好的文档分词和建立倒排索引.

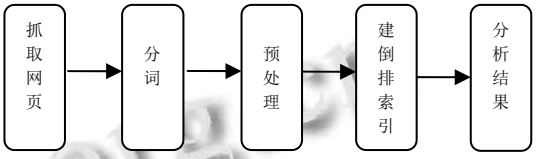

图 2 实验流程图 1

4.1.2 分词工具性能评估

此实验是在上一个实验抓取网页的基础上随机抽取了 30 篇文档, 分别对这 30 篇文档进行人工分词和分词方法分词. 由于 ASPSeek、Paoding、MMseg 三种分词工具都是完全基于词典的方法, 所以本文中对这三种分词工具统一使用 ASPSeek 庞大的 25 万多词汇的词库为标准, 对文档进行分词. 假设人工标定分出的词是正确的, 并且对分出的词去除停用词, 取出分词方法与人工分词分出相同的词, 并且计算每种分词方法的分出词的准确率、召回率和 F_1 测度.

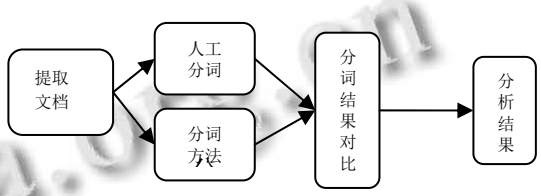

图 3 实验流程图 2

5 实验结果与分析

本实验的 ICTCLAS、Paoding、MMSEG、IK 以及 JE 这五种分词方法都是结合 lucene 在 Eclipse 软件下运行的.

本实验中主要使用了 Lucene 3.0 版本以及 Lucene 2.0 版本, 由于 JE 分词软件只能应用于 Lucene 3.0 以下版本. 其余 4 种分词均和 Lucene 3.0 结合.

实验第一部分根据各种分词工具分出的词以及索引建立时间和索引建立大小进行了比较. 第二部分根据人工和分词工具对文档分词, 并且比较了各种分词工具的性能.

5.1 词汇量、建立索引大小及建立索引时间的比较

表 1 词汇量、建立索引大小及建立索引时间的比较

	ASPSeek	ICTCLAS	Paoding	MMseg	IK	JE
分词总数	93639	90740	130325	126708	97904	249500
索引大小/Mb	38	39.5	53.9	52	57.1	51
索引时间/ms	70000	1054375	1029344	1050047	1116344	63547
字符与数字符号总数	2052	0	0	0	0	0
字符个数	1847	0	0	0	0	0
数字个数	4851	0	307	0	0	0

从表 1 中可知在对相同的文档分词时, 各种分词方法分出的词的数量是不一样的, 其中使用 JE 分词工具分出的词汇最多, 其次是 Paoding 分词, 分词数量最少的是中科院的 ICTCLAS 分词工具.

在使用 ASPSeek 和中科院的 ICTCLAS 建立索引大小差不多的情况下, 即索引的存储空间相当的时候, 它们的索引时间相差 1/3.

在 Paoding、MMseg、IK、JE 四种分词方法的索引存储空间差不多的情况下, JE 分词方法所建索引的速度比其它分词方法节省约 40%.

在 ASPSeek 和 JE 分词工具建立索引时间相当的情况下, JE 分词工具却需要比 ASPSeek 所占存储空间大 25%.

在 ICTCLAS、Paoding、MMseg、IK、分词工具建立索引时间差不多的情况下, 而 ICTCLAS 分词工具却比其它三种分词工具节约 25%的空间.

另外, 只有 ASPSeek 和 Paoding 分词结果中有字符符号和数字符号, 在 ASPSeek 中有 8750 个字符符号和数字符号, 有 1 个停用词; 在 Paoding 分词中有 307 个数字符号. 其它的分词工具都在分词过程调用各自的分词原理已将字符符号和数字符号过滤, 没有显示. 同时在使用各种分词工具得出的结果中仍有一些垃圾词汇, 但人为的干预会导致分词结果的不确定性, 所以在此实验中没有对分词的结果进行处理.

5.2 分词准确率、召回率及 F_1 测度比较

测试文档的召回率、精度以及 F_1 测度分别定义为:

准确率(P)= 识别出来的正确词条数目/文档中分词的总数目

召回率(R)= 识别出来的正确词条数目/人工判别的词条总数

F_1 测度=2RP/(R+P)

表 2 分词准确率、召回率及 F_1 测度比较

	ASPSeek	ICTCLAS	Paoding	MMseg	IK	JE
分词总数	3400	2278	3952	3303	4233	3312
数字字符	22	0	33	0	0	0
与人工分词相同个数	2059	1647	2358	2173	2490	2245
准确率(%)	60.56	72.30	59.67	65.79	58.82	67.78
召回率(%)	67.42	53.93	77.21	71.15	81.53	73.51
F1测度(%)	63.81	61.73	67.32	68.36	68.34	70.53

从表 1 不能直接的判断每种分词方法的准确性. 因此, 实验的最后又对其中的 30 篇文档进行了人工标定分词, 经统计, 人工分出的词汇共有 3054 个. 然后又使用了本文中提到的六种分词方法分别对这 30 篇文档进行了分词, 分词结果如表 2 所示.

从表 2 可以看出, 分词准确率比较高的有中科院的 ICTCLAS 和 JE 分词工具, 准确率最低的是 IK 分词工具, ASPSeek 和 Paoding 分词由于完全基于词典库的分词, 还分出了数字字符. 从分词的召回率可以看出, 比较高的有 IK 和 Paoding 分词方法. F1 测度是一个综合测评的方法, 从结果中可以看出 JE 分词和 MMseg 分词的 F_1 测度较高. 因此, 根据实验得出 ICTCLAS 分词工具的准确率最优,而 JE 分词工具的 F_1 测度最高.

6 结论

在此次实验中, 主要实现了农业搜索引擎中文分词工具的对比. 在农业搜索引擎中加入中文分词算法后, 不仅提高了搜索结果的准确率, 还为农民老百姓等人员带来了方便快捷的服务. 由于不同分词工具分词的原理不同, 基于的词典库不同, 分词的结果及索引建立的时间和存储空间不同, 导致对搜索引擎性能的影响. 本实验中的难点是在实验最后一部分中抓取的网页需要人工标定分词, 专业人士也只能凭借经验和记忆对文档分词, 不可能实现百分之百的分词, 只能降低错误率. 而且基于时间的限制及人员的不足, 不能对所有的文档进行人工标定, 只能随机的对其中一小部分做测试, 这样不仅增加了词识别的难度还增加了工作量并且耗费时间与精力. 望后期可以对这方面有进一步的研究.

参考文献

1 曹桂宏,何丕廉,吴光远,聂颂.中文分词对中文信息检索系统性能的影响.计算机工程与应用,2003.

2 金澎,刘毅.汉语分词对中文搜索引擎检索性能的影响.情报学报,2006,25(1):21–24.

3 章成敏,章成志.国外农业搜索引擎评析.农业网络信息,2004,(11).

4 刘辉林,郭来德,刘兰哲,王光兴.中文农业主题搜索引擎的设计与实现.郑州大学学报,2007,39(2):74–77.

5 彭玉容,杨捧,高媛.农业搜索引擎的发展现状及关键技术研究.安徽农业科学,2010,38(20):10971–10973.

6 杨鸿雁,尚俊平,徐延华,王萌,张宇.农业专业搜索引擎建设探讨.农业图书情报学刊,2005,17(4):83–84.

7 刘件,魏程.中文分词算法研究.微计算机应用,2008,29(8):12–16.

8 刘迁,贾惠波.中文信息处理中自动分词技术的研究与展望.计算机工程与应用,2006.

9 张博,姜建国,万平国.对互联网环境下中文分词系统的一种架构改进.计算机应用研究,2006,(11):176–178.

10 蔡小艳,寇应展,沈巍,郑伟.汉语词法分析系统 ICTCLAS 在 Nutch-0.9 中的应用与实现.军械工程学院学报, 2008, 20(5):63–67.

11 夏天,樊孝忠,刘林.利用 JNI 实现 ICTCLAS 系统的 Java 调用.计算机应用,2004,24:177–182.

12 孙殿哲,魏海平,陈岩.Nutch 中庖丁解牛中文分词的实现与评测.计算机与现代化,2010,6:187–189.

13 蔡小艳,寇应展,沈巍,郑伟. Nutch-0.9 中 JE 中文分词的实现.科学技术与工程,2008,8(17):4881–4884.

14 向晖,郭一平,王亮.基于 Lucene 的中文字典分词模块的设计与实现.信息检索技术,2006,(9).

15 王志嘉,薛质.一种基于 Lucene 的中文分词的设计与测试.信息技术,2010,(12):49–53.

16 费洪晓,康松林,朱小娟,谢文彪.基于词频统计的中文分词的研究.计算机工程与应用,2005.

17 mmseg4j. http://www.oschina.com/project/mmseg4j.

18 ASPSeek 中文网站. http://aspseek.xjau.edu.cn.

19 Foo S, Li H. Chinese word segmentation and its effect on information retrieval. Information Processing and Management, 2004.

20 Liu KY, Zheng JH. Research of automatic chinese word segmentation. Proc. of the First International Conference on Machine Learning and Cybernetics. Beiing. 2002.