



文本复制检测报告单(全文对照)

№:ADBD2018R_2017110615244720181017200038719806713784

检测时间: 2018-10-17 20:00:38

检测文献: 80636129435913428_田肇阳_近似新闻合并及正负面评价

作者: 田肇阳

检测范围: 中国学术期刊网络出版总库

中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库

中国重要会议论文全文数据库

中国重要报纸全文数据库

中国专利全文数据库

互联网资源(包含贴吧等论坛资源)

英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)

港澳台学术文献库

优先出版文献库

互联网文档资源

图书资源

CNKI大成编客-原创作品库

个人比对库

时间范围: 1900-01-01至2018-10-17

检测结果

总文字复制比: 9.7%

跨语言检测结果: 0%

去除引用文献复制比: 9.7%

去除本人已发表文献复制比: 9.7%

单篇最大文字复制比: 3%

重复字数: [1066]

总字数: [10984]

单篇最大重复字数: [333]

总段落数: [1]

前部重合字数: [38]

疑似段落最大重合字数: [1066]

疑似段落数: [1]

后部重合字数: [1028]

疑似段落最小重合字数: [1066]

指标: ☐ 疑似剽窃观点 ☒ 疑似剽窃文字表述 ☐ 疑似自我剽窃

☐ 一稿多投 ☐ 过度引用 ☐ 疑似整体剽窃 ☐ 重复发表

表格: 0

脚注与尾注: 0

(注释: ■ 无问题部分 ■ 文字复制比部分 ■ 引用部分)

1. 80636129435913428_田肇阳_近似新闻合并及正负面评价

总字数: 10984

相似文献列表 文字复制比: 9.7% (1066) 疑似剽窃观点 (0)

1	基于语义分析的网络信息采集算法研究与应用 赵佳鹤(导师: 王秀坤) - 《大连理工大学硕士论文》 - 2006-12-05	3.0% (333) 是否引证: 否
2	互联网短文本信息分类关键技术研究 - 豆丁网 - 《互联网文档资源 (http://www.docin.com/p-108213576.html)	2.8% (313) 是否引证: 否

3	<u>搜索引擎中中文WEB文本自动分类研究</u> 刘宏伟(导师: 孟小华) - 《暨南大学硕士论文》 - 2007-04-01	2.8% (303) 是否引证: 否
4	中文分词技术 - 深之JohnChen的专栏 - 博客频道 - CSDN.NET - 《网络 (http://blog.csdn.net/byxdaz/article/details/5815677) 》 - 2013	2.6% (286) 是否引证: 否
5	中文分词原理 - xiaomin1991222的专栏 - 博客频道 - CSDN.NET - 《网络 (http://blog.csdn.net/xiaomin1991222/article/details/50981853) 》 - 2017	2.6% (286) 是否引证: 否
6	中文分词技术(中文分词原理) - u010384318的专栏 - 博客频道 - CSDN.NET - 《网络 (http://blog.csdn.net/wbgxx333/article/details/11178693) 》 - 2013	2.6% (286) 是否引证: 否
7	文本相似性检测----中文分词技术 - Johline的博客 - CSDN博客 - 《网络 (http://blog.csdn.net/johline/article/details/59109811) 》 - 2017	2.6% (281) 是否引证: 否
8	互联网舆情信息获取与分析研究 - 豆丁网 - 《互联网文档资源 (http://www.docin.com/p-669365673.html) 》 - 2017	2.5% (280) 是否引证: 否
9	《跨语言文本相似性检测》第一周一前期调研 - Johline的博客 - CSDN博客 - 《网络 (http://blog.csdn.net/johline/article/details/59111894) 》 - 2017	2.4% (269) 是否引证: 否
10	互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 (http://www.docin.com/p-669364954.html) 》 - 2017	2.4% (263) 是否引证: 否
11	互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 (http://www.docin.com/p-122549484.html) 》 - 2013	2.4% (263) 是否引证: 否
12	<u>中文分词技术及其应用初探</u> 余战秋 - 《电脑知识与技术》 - 2004-11-27	2.3% (253) 是否引证: 否
13	<u>民航发动机健康管理数据库设计与故障诊断</u> 张煜(导师: 李书明;黄燕晓) - 《中国民航大学硕士论文》 - 2016-06-30	2.3% (251) 是否引证: 否
14	关于现代中文分词技术的综述 - 《互联网文档资源 (http://wenku.baidu.com/view/bf11ce4be518964bcf847cff.html) 》 - 2017	2.3% (251) 是否引证: 否
15	信息检索论文 - 《互联网文档资源 (http://wenku.baidu.com/view/708f660390c69ec3d5bb753c.html) 》 - 2017	2.2% (242) 是否引证: 否
16	<u>英文自动问答系统中数值型问句的理解研究</u> 赵龙(导师: 刘亚清;黄威) - 《大连海事大学硕士论文》 - 2016-05-01	2.1% (234) 是否引证: 否
17	<u>社交网络舆情传播监督管理系统的设计与实现</u> 郑罡(导师: 阚忠良;姚德明) - 《黑龙江大学硕士论文》 - 2015-10-20	1.9% (212) 是否引证: 否
18	<u>基于用户兴趣度的网络信息过滤模型研究</u> 王翠平(导师: 刘培玉) - 《山东师范大学硕士论文》 - 2007-04-27	1.9% (212) 是否引证: 否
19	中文分词算法 - llandrj的博客 - CSDN博客 - 《网络 (http://blog.csdn.net/llandrj/article/details/49412141) 》 - 2017	1.9% (212) 是否引证: 否
20	[PDF][精品文]基于用户兴趣度的网络信息过滤模型研究 - 豆丁网 - 《互联网文档资源 (http://www.docin.com/p-475602698.html) 》 - 2017	1.9% (212) 是否引证: 否
21	[PDF][精品文]基于用户兴趣度的网络信息过滤模型研究 - 豆丁网 - 《互联网文档资源 (http://www.docin.com/p-475602698.html) 》 - 2016	1.9% (212) 是否引证: 否
22	<u>基于条件随机场和空间推理的地理编码方法</u> 周海(导师: 李宏伟) - 《解放军信息工程大学硕士论文》 - 2015-04-20	1.8% (196) 是否引证: 否
23	<u>论四头双导程蜗杆车削挂轮的选配</u> 梁宗斌; - 《现代商贸工业》 - 2017-08-05	1.5% (165) 是否引证: 否
24	<u>基于Hadoop平台分布式SVM分类研究</u> 蔡鑫怡; - 《电脑迷》 - 2018-06-21	1.4% (155) 是否引证: 否
25	<u>基于经济普查大数据的上海“三新”经济发展态势研究</u>	1.4% (154)

	杭敬;苑立波;张志远; - 《统计科学与实践》 - 2016-11-25	是否引证: 否
26	<u>吕苏语口语标注语料的自动分词方法研究</u> 于重重;操镭;尹蔚彬;张泽宇;郑雅; - 《计算机应用研究》 - 2016-07-15 1	1.3% (147) 是否引证: 否
27	<u>媒体情绪能够影响投资者情绪吗——基于新兴市场门槛效应的研究</u> 黄宏斌;刘树海;赵富强; - 《山西财经大学学报》 - 2017-10-30 1	0.9% (103) 是否引证: 否
28	<u>汉语篇章连贯性自动分析方法研究</u> 王小虎(导师: 钟茂生) - 《华东交通大学硕士论文》 - 2015-06-30	0.8% (92) 是否引证: 否
29	<u>面向篇章的省略恢复及其在机械设计中的应用</u> 万棋顺(导师: 赵克) - 《西安电子科技大学硕士论文》 - 2008-01-01	0.7% (75) 是否引证: 否
30	<u>熔融金属红外热像测温精度的研究</u> 高悦(导师: 马翠红;李北丹) - 《华北理工大学硕士论文》 - 2016-12-05	0.6% (66) 是否引证: 否
31	<u>金泽大厦建筑工程项目绿色施工管理研究</u> 胡通文(导师: 蔡为民;刘美秀) - 《天津工业大学硕士论文》 - 2018-01-26	0.5% (60) 是否引证: 否
32	<u>习近平:各级党政机关和领导干部要学会通过网络走群众路线</u> - 《共产党员》 - 2016-05-03	0.3% (38) 是否引证: 否
33	<u>对网络直播乱象说“不”</u> 曹振国; - 《求学》 - 2017-02-15	0.3% (37) 是否引证: 否

	原文内容	相似内容来源
1	<p>此处有 38 字相似</p> <p>中, 互联网正在成为一个日趋发展的平台和日渐重要的媒介。互联网不是有百利而无一害的, 在具有丰富信息的同时具有许多不良信息。网络空间是亿万民众共同的精神家园。网络空间天朗气清、生态良好, 符合人民利益。因此, 建立一个负面信息过滤系统迫在眉睫。简要说来, 本文完成了以下几个任务: 文章获取, 即利用URLLIB模块从百度爬</p>	<p>习近平:各级党政机关和领导干部要学会通过网络走群众路线 - 《共产党员》-2016-05-03 (是否引证: 否)</p> <p>1. 和政府工作提的还是对领导干部个人提的,不论是和风细雨的还是忠言逆耳的,我们不仅要欢迎,而且要认真研究和吸取。习近平强调,网络空间是亿万民众共同的精神家园。网络空间天朗气清、生态良好,符合人民利益。网络空间乌烟瘴气、生态恶化,不符合人民利益。我们要本着对社会负责、对人民负责的态度,依法加强网络空间治理,加强网络内容建</p> <p>对网络直播乱象说“不” 曹振国;-《求学》-2017-02-15 (是否引证: 否)</p> <p>1. 中产生的低俗文化,已给互联网环境造成巨大的伤害与损失。2016年4月,习近平总书记在网络安全和信息化工作座谈会上强调:“网络空间是亿万民众共同的精神家园。网络空间天朗气清、生态良好,符合人民的利益。”国家网信办出台规定对网络环境进行净化,正当其时。●观点三棱镜◆“网红”不该沦为低俗的代名词 人民网评:当一夜</p>
	<p>此处有 35 字相似</p> <p>图2-3所示: 图2: 程序运行截图 图3: 程序运行结果3. JIEBA分词4.1 JIEBA模块介绍 对于中文来说, 词是最小的能够独立活动的有意义的语言成分。汉语是以字位单位, 不像西方语言, 词与词之间没有空格等的标志指示词的边界。分词问题为中文文本处理的基础性工作, 对于本文的中文信息处理起到关键</p>	<p>中文分词技术 - 深之JohnChen的专栏 - 博客频道 - CSDN.NET - 《网络 (http://blog.csdn.net/byxdaz/article/details/5815677) 》 (是否引证: 否)</p> <p>1. (0) 收藏 举报 算法lucene文档搜索引擎中文分词库classification 中文分词技术 一、为什么要进行中文分词? 词是最小的能够独立活动的有意义的语言成分, 英文单词之间是以空格作为自然分界符的, 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, 因此, 中文词语分析是中文</p>

2		<p>中文分词原理 - xiaomin1991222的专栏 - 博客频道 - CSDN 《网络 (http://blog.csdn.net/xiaomin1991222/article/details/509 (是否引证: 否)</p> <p>1. 一、为什么要进行中文分词? 词是最小的能够独立活的语言成分, 英文单词之间是以空格作为自然分界符的, 字为基本的书写单位, 词语之间没有明显的区分标记, 因此, 词语分析是中文</p> <p>中文分词技术(中文分词原理) - u010384318的专栏 - 博客 - CSDN.NET - 《网络 (http://blog.csdn.net/wbgxx333/article/details/11178693 (是否引证: 否)</p> <p>1. 一、为什么要进行中文分词? 词是最小的能够独立活有意义的语言成分, 英文单词之间是以空作为自然分界符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, 因此, 中文词语分析是中文信</p> <p>文本相似性检测----中文分词技术 - Johline的博客 - CSDN 博客 - 《网络 (http://blog.csdn.net/johline/article/details/59109811) 》 (是否引证: 否)</p> <p>1. , 唯独词没有一个形式上的分界符, 虽然英文也同样存在短语的划分问题, 不过在词这一层上, 中文比之英文要复杂的多、困难的多。 词是最小的能够独立活动的有意义的语言成分, 英文单词之间是以空作为自然分界符的, 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, 因此, 中文词语分析是中文信</p>
3	<p>此处有 31 字相似</p> <p>的, 可以分成一下两种形式。结婚 / 的 / 和 / 尚未 / 结婚 / 的 结婚 / 的 / 和尚 / 未 / 结婚 / 的 未登录词存在两种情况, 一是已有的词表中没有收录的词, 二是已有的训练语料中未曾出现过的词。据学界普遍认为对于大规模真实文本来说, 未登录词对于分词的精度的影响远超歧义切分。一些网络新词, 自造词通常都属于这些词。</p>	<p>汉语篇章连贯性自动分析方法研究 王小虎-《华东交通大学硕士论文》-2015-06-30 (是否引证: 否)</p> <p>1. 过程中可以结合上下文的语义分析, 甚至韵律分析去分析它。最后就是未登录词问题, 未登录词又称生词, 可以有两种解释: 一、已有的词表中 没有收录的词, 二、已有的训练预料中未曾出现过的词 (集外词)。由于目前的汉语分词大多数采用基于大规模训练语料的统计方法, 因此通常将集外词与未登录词等一起 来。</p>
4	<p>此处有 33 字相似</p> <p>/ 未 / 结婚 / 的 未登录词存在两种情况, 一是已有的词表中没有收录的词, 二是已有的训练语料中未曾出现过的词。据学界普遍认为对于大规模真实文本来说, 未登录词对于分词的精度的影响远超歧义切分。一些网络新词, 自造词通常都属于这些词。汉语分词方法主要有以下几种: ——基于字典、词库匹配的分词方法 (基于规则)</p>	<p>汉语篇章连贯性自动分析方法研究 王小虎-《华东交通大学硕士论文》-2015-06-30 (是否引证: 否)</p> <p>1. 究领域名称, 如三聚氰胺、苏丹红、禽流感等等; 四、其他专用名词, 如新出现的产品名, 电影、书籍等文艺作品的名称。 对于大规模篇章来说, 未登录词对于分词精度的影响比歧义切分还大。对于一个分词系统来说, 如果不能实时获取训练预料, 一旦出现新词, 那么切分结果可能会出错。自从汉语自动分词这一概</p>
	<p>此处有 277 字相似</p> <p>学界普遍认为对于大规模真实文本来说, 未登录词对于分词的精度的影响远超歧义切分。一些网络新词, 自造词通常都属于这些词。汉语分词方法主要有以下几种: ——基于字典、词库匹配的分词方法 (基于规则) ——基于词频度统计的分词方法 (基于统计) ——</p>	<p>英文自动问答系统中数值型问句的理解研究 赵龙-《大连海事大学硕士论文》-2016-05-01 (是否引证: 否)</p> <p>1. 算法大体可分为H类; 基于字符串匹配的分词, 基于理解的分词, 基于统计的分词。(1)基于字符串匹配的分词也叫机械分词, 是指将待分的字符串和一个充分大的机器词典中的词条进行匹配。匹配又可细分为正向匹配</p>

基于知识理解的分词方法。基于字符串匹配分词，机械分词算法。将待分的字符串与一个充分大的机器词典中的词条进行匹配。分为正向匹配和逆向匹配；最大长度匹配和最小长度匹配；单纯分词和分词与标注过程相结合的一体化方法。所以常用的有：正向最大匹配，逆向最大匹配，最少切分法。实际应用中，将机械分词作为初分手段，利用语言信息提高切分准确率。优先识别具有明显特征的字，以这些字为断点，将原字符串分为较小字符串再机械匹配，以减少匹配错误率，或将分词与词类标注结合。相邻的字同时出现的次数越多，越有可能构成一个词语，对语料中的字组频度进行统计，基于词的频度统

和逆向匹配、最大长-5- 第1章绪论 度匹配和最小长度匹配、单纯分词和分词与标注过程相结合的一体化方法。实际

2. 机器词典中的词条进行匹配。匹配又可细分为正向匹配和逆向匹配、最大长-5- 第1章绪论 度匹配和最小长度匹配、单纯分词和分词与标注过程相结合的一体化方法。实际应用中，利用机械分词进行初次划分，并利用语言信息提高切分准确率。划分过程结合词类标注，优先识别具有明

3. 匹配、最大长-5- 第1章绪论 度匹配和最小长度匹配、单纯分词和分词与标注过程相结合的一体化方法。实际应用中，利用机械分词进行初次划分，并利用语言信息提高切分准确率。划分过程结合词类标注，优先识别具有明显特征的字并W这些字为分隔点，将原字符串分为较小字符串再机械匹配，从而降低匹配错误率。(2)基于理解分词，是指分词过程中利用分词系统模拟人对句子的理解来进行句法语义分析。

基于语义分析的网络信息采集算法研究与应用 赵佳鹤-《大连理工大学硕士论文》-2006-12-05（是否引证：否）

1. 但计算机如何也能理解其处理过程就称为分词算法。现有的分词算法可分为三大类[2v]:基于字典、词库匹配的分词方法、基于词的频度统计的分词方法和基于知识理解的分词方法。(1)基于字符串匹配的分词方法

2. 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(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，常用的方法如下: ①正向最大匹配法(Maxi

3. 一种方法是改进扫描方式，称为特征扫描或标志切分，优先在待分析字符串中识别和切分出一些带有明显特征的字，以这些字作为断点，可将原字符串分为较小的串再来进机械分词，从而减少匹配的错误率。另一种方法是将分词和词类标注结合起来，利用丰富的词类信息对分词决策提供帮

面向篇章的省略恢复及其在机械设计中的应用 万棋顺-《西安电子科技大学硕士论文》-2008-01-01（是否引证：否）

1. 义性，在词法分析阶段很难能够一次性的正确识别。根据国内外的发展状况，综合了大量参考文献，已有的汉语分词方法[26]主要的有以下三大类：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。本系统的词法分析系统是一个基于知识的分词系统，根据领域词库把以汉字为单位的输入段落转换成以独立

搜索引擎中中文WEB文本自动分类研究 刘宏伟-《暨南大学硕士论文》-2007-04-01（是否引证：否）

1. 两种方法正是目前中文分词领域最为常用的。3.2中文分词算法 自动分词的基本方法有:基于字符串匹配的分词方法和基于统计的分词方法。3.2.1基于字符串匹配的分词算法 这种方法又称为机械分词方法，它是按照一

定的策略将待分析的汉字串与一个

2. 为常用的。3.2中文分词算法 自动分词的基本方法有:
基于字符串匹配的分词方法和基于统计的分词方法。

3.2.1基于字符串匹配的分词算法这种方法又称为机械分词方法,它是按照一定的策略将待分析的汉字串与一个充分大的词典中的词条进行匹配,若

基于条件随机场和空间推理的地理编码方法 周海-《解放军信息工程大学硕士论文》-2015-04-20 (是否引证: 否)

1. 配的时间复杂度为字符串本身长度。对于长度为n的字符串,其时间复杂度为O(n),而最大匹配平均复杂度为O(n²)。(4)将机械分词作为初分手段,利用语言信息提高切分准确率。优先识别具有明显特征的词,然后以这些词为断点将原字符串分割为较小字符串再进行机械分词,从而减少匹配错误率;先用匹配法分词,发现歧义,向前看两词语,对此三个词运用启发式的消歧规则,根据规则(最长匹配,词语长度,语素,概率等规则)

互联网舆情信息获取与分析研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-669365673.html>) 》 - (是否引证: 否)

1. e 3- 1 Participle Technology 互联网舆情信息获取与分析研究 23 自动分词算法可分为三大类,基于字典、词库匹配的分词方法,基于词频度统计的分词方法和基于知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词,如,最大匹配法、最小分词方法等。这类方

2. 典中找到某个字符串,则匹配成功。识别出一个词,根据扫描方向的不同分为正向匹配和逆向匹配。根据不同长度优先匹配的情况,分为最大,最长,匹配和最小,最短,匹配。根据与词性标注过程是否相结合,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。常用的方法如下,最大正向匹配法 (Maximum Matching

互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-669364954.html>) 》 - (是否引证: 否)

1. 图如图2所示。图2 分词技术 Figure 2 Participle Technology 自动分词算法可分为三大类,基于词典、词库匹配的分词方法,基于词频统计的分词方法和基于知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词,是一种应用最广泛的机械分词算法。常用的

2. 典中找到某个字符串,则匹配成功。识别出一个词,根据扫描方向的不同分为正向匹配和逆向匹配。根据不同长度优先匹配的情况,分为最大,最长,匹配和最小,最短,匹配。根据与词性标注过程是否相结合,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。常用的方法如下,最大正向匹配法(Maximum Matching

互联网短文本文信息分类关键技术研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-108213576.html>) 》 - (是否引证: 否)

1. 3- 1 分词技术 Figure 3- 1 Participle Technologies 自动分词算法可分为三大类,基于字典、词库匹配的分词方法,基于词频统计的分词方法和基于知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词,如,最大匹配法、最小分词方法等。基于该

2. 典中找到某个字符串,则匹配成功。识别出一个词,根据扫描方向的不同分为正向匹配和逆向匹配。根据不同长度优先匹配的情况,分为最大、最长、匹配和最小、最短、匹配。根据与词性标注过程是否相结合,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。常用的方法如下,最大正向匹配法 (Maximum Matching

互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-122549484.html>) 》 - (是否引证: 否)

1. 图如图2所示。图2 分词技术 Figure 2 Participle Technology 自动分词算法可分为三大类,基于词典、词库匹配的分词方法,基于词频统计的分词方法和基于知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词,是一种应用最广泛的机械分词算法。常用的

2. 典中找到某个字符串,则匹配成功。识别出一个词,根据扫描方向的不同分为正向匹配和逆向匹配。根据不同长度优先匹配的情况,分为最大、最长、匹配和最小、最短、匹配。根据与词性标注过程是否相结合,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。常用的方法如下,最大正向匹配法(Maximum Matching

《跨语言文本相似性检测》第一周一前期调研 - Johline的博客 - CSDN博客 - 《网络 (<http://blog.csdn.net/johline/article/details/59111894>) 》 (是否引证: 否)

1. bcdefg,k设为2, 那得到的词语就是ab,bc,cd,de,ef,fg。中文分词技术 现有的分词算法可分为三大类: 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。字符串匹配的分词方法 这是种常用的分词法,百度就是用此类分词。字符串匹配的分词方法, 又分为3种分词方法。(1) 正向最

吕苏语口语标注语料的自动分词方法研究 于重重;操镭;尹蔚彬;张泽宇;郑雅;-《计算机应用研究》-2016-07-15 1 (是否引证: 否)

1. 典、词库匹配的分词方法旨在将待分词的字符串,与容量够大、内容够丰富的计算机词库中的词条进行匹配。按照匹配原则的不同,基于词典的方法分为:正向匹配法和逆向匹配法;最小长度匹配法和最大长度匹配法;单纯分词和分词与标注相结合的综合性方法。目前,较为典型的分词匹配算法是正向最大匹配法、逆向最大匹配法以及最少切分法[8,9]。在大多数实际应

中文分词技术及其应用初探 余战秋-《电脑知识与技术》-2004-11-27 (是否引证: 否)

1. 大匹配法组合)我们还可以稍做改进,其一是改进扫描

方式,称为特征扫描或标志切分,即首先在待分析字符串中识别和切分出一些带有**明显特征**的词,以**这些词作为断点**,将原字符串分为较小的串,然后再进行机械分词,从而**减少匹配的误差率**。另一种方法是将分词和词类标注结合起来,利用丰富的词类信息对分词决策提供帮助,同时在标注过程中又反过来对分词结果进行检验、

中文分词技术 - 深之JohnChen的专栏 - 博客频道 -

CSDN.NET - 《网络

(<http://blog.csdn.net/byxdaz/article/details/5815677>) 》

(是否引证: 否)

1. 分 (包括向前、向后、以及前后相结合)、最少切分、全切分等等。二、中文分词技术的分类 我们讨论的分词算法可分为三大类: **基于字典、词库匹配的分词方法; 基于词频度统计的分词方法和基于**知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词, 如: 最大匹配法、最小分词方法等。这类方

中文分词原理 - xiaomin1991222的专栏 - 博客频道 - CSDN.NET - 《网络

(<http://blog.csdn.net/xiaomin1991222/article/details/509>)

(是否引证: 否)

1. 分 (包括向前、向后、以及前后相结合)、最少切分、等。二、中文分词技术的分类 我们讨论的分词算法可分为三大类: **基于字典、词库匹配的分词方法; 基于词频度统计的分词方法和基于**知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词, 如: 最大匹配法、最小分词方法

民航发动机健康管理数据库设计与故障诊断 张煜-《中国民航大学硕士论文》-2016-06-30 (是否引证: 否)

1. 正意义上分隔符, 虽然也存在同样的分隔问题, 不过在词汇分隔上, 英文要比中文简单许多。分词算法可分为三大类: **基于字典、词库匹配的分词方法; 基于词频度统计的分词方法和基于**知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词, 如: 最大匹配法、最小分词

中文分词技术(中文分词原理) - u010384318的专栏 - 博客频道 - CSDN.NET - 《网络

(<http://blog.csdn.net/wbgxx333/article/details/11178693>)

(是否引证: 否)

1. 分 (包括向前、向后、以及前后相结合)、最少切分、分等等。二、中文分词技术的分类 我们讨论的分词算法可分为三大类: **基于字典、词库匹配的分词方法; 基于词频度统计的分词方法和基于**知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词, 如: 最大匹配法、最小分词方法等。这类方

文本相似性检测----中文分词技术 - Johline的博客 - CSDN.NET - 《网络

(<http://blog.csdn.net/johline/article/details/59109811>) 》

(是否引证: 否)

1. 分 (包括向前、向后、以及前后相结合)、最少切分、

	<p>全切分等等。二、中文分词技术的分类 我们讨论的分词算法可分为三大类：基于字典、词库匹配的分词方法；基于词频度统计的分词方法和基于知识理解的分词方法。第一类方法应用词典匹配、汉语词法或其它汉语语言知识进行分词，如：最大匹配法、最小分词方法等。这类方</p> <p>关于现代中文分词技术的综述 - 《互联网文档资源》 (http://wenku.baidu.com/view/bf11ce4be518964bcf847c) (是否引证：否)</p> <p>1. 中文文本中词与词则无明显的界限，这就影响了关键词配。三、中文分词技术的分类 我们讨论的分词算法可分为基于字典、词库匹配的分词方法；基于词频度统计的分词知识理解的分词方法。第一类方法应用词典匹配、汉语词汉语语言知识进行分词，如：最大匹配法、最小分词方法</p> <p>信息检索论文 - 《互联网文档资源》 (http://wenku.baidu.com/view/708f660390c69ec3d5bb7) (是否引证：否)</p> <p>1. 信息来进一步提高切分的准确率。一种方法是改进扫描特征扫描或标志切分，优先在待分析字符串中识别和切分显特征的词，以这些词作为断点，可将原字符串分为较小械分词，从而减少匹配的错误率。另一种方法是将分词和起来，利用丰富的词类信息对分词决策提供帮助，并且在反过</p> <p>汉语篇章连贯性自动分析方法研究 王小虎-《华东交通大学硕士论文》-2015-06-30（是否引证：否）</p> <p>1. 机结合能相 互补充对方的缺点，使得系统具有较高的精度和效率。目前使用得最多的基于规则的分词方法是最大匹配法，该方法可以分为正向最大匹配法和逆向最大匹配法。所谓正向最大匹配就是首先设定一个最大词长 MAXL，并按照 从左到右的顺序在词串中选取一个长度为 MAXL 的字串。</p>
6	<p>此处有 31 字相似</p> <p>，对语料中的字组频度进行统计，基于词的频度统计的分词方法是一种全切分方法。JIEBA是基于统计的分词方法，JIEBA分词采用了动态规划查找最大概率路径，找出基于词频的最大切分组合。[8] 该方法主要基于句法、语法分析，并结合语义分析，通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分</p> <p>媒体情绪能够影响投资者情绪吗——基于新兴市场门槛效应的研究 黄宏斌;刘树海;赵富强;-《山西财经大学学报》-2017-10-30 1（是否引证：否）</p> <p>1. 中文分词工具采用了Jieba分词工具。该工具基于前缀词典实现词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图。采用了动态规划查找最大概率路径,找出基于词频的最大切分组合。对于未登录词,采用了基于汉字成词能力的HMM模型,使用了Viterbi算法。对数据集进行分词后,一共产生了287 632</p> <p>基于Hadoop平台分布式SVM分类研究 蔡鑫怡;-《电脑迷》-2018-06-21（是否引证：否）</p> <p>1. .4jieba分词主要原理jieba分词目前最好用中文分词之一,它基于Trie树结构实现高效的词图扫描,生成句子中汉字。采用了动态规划查找最大概率路径,找出基于词频的最大切分组合有可能成词情况所构成的有向无环图(DAG)。对于未登录词,采用了基于汉字成词能力的HMM模型,使用了Viterbi算法。</p> <p>基于经济普查大数据的上海“三新”经济发展态势研</p>

	<p>究 杭敬;苑立波;张志远;-《统计科学与实践》-2016-11-25 (是否引证: 否)</p> <p>1. 进行分词。1.分词算法该组件基于前缀词典实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图(DAG);采用了动态规划查找最大概率路径,找出基于词频的最大切分组合;对于未登录词,采用了基于汉字成词能力的HMM模型,使用了Viterbi算法。2.分词模式应对不同的分词需要,该组件提供了</p> <p>吕苏语口语标注语料的自动分词方法研究 于重重;操镭;尹蔚彬;张泽宇;郑雅;-《计算机应用研究》-2016-07-15 1 (是否引证: 否)</p> <p>1. 型的分词方法和基于词感知机算法的分词方法等[10~12]。目前主流的分词方法中,结巴(jieba)分词方法[13,14]采用了动态规划查找最大概率路径,找出基于词频的最大切分组合,对于未登录词,采用了基于汉字成词能力的隐马尔可夫模型,使用了维特比算法。Jieba分词是国内程序员用Python开发的一</p>
7	<p>此处有 227 字相似</p> <p>JIEBA是基于统计的分词方法, JIEBA分词采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合。[8]该方法主要基于句法、语法分析, 并结合语义分析, 通过对上下文内容所提供信息的分析对词进行定界, 它通常包括三个部分: 分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力, 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。因此目前基于知识的分词系统还处在试验阶段。[8] JIEBA中文分词模块支持三种分词模式: A. 精确模式, 试图将句子最精确地切开, 适合文本分析; B. 全模式, 把句子中所有的可以成词</p> <p>搜索引擎中中文WEB文本自动分类研究 刘宏伟-《暨南大学硕士论文》-2007-04-01 (是否引证: 否)</p> <p>1. 的语言模型叫作二元模型2—Grain。3.2.3基于知识理解的分词算法 该方法主要基于句法、语法分析, 并结合语义分析, 通过对上下文内容 所提供信息的分析对词进行定界, 它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力, 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, 因此目前基于知识的分词 系统还处在试验阶段。3.3传统分词方法改进 对传统分词方法的改进, 要以提高系统整体性能为前提。在这个前提下应</p> <p>中文分词技术 - 深之JohnChen的专栏 - 博客频道 - CSDN.NET - 《网络 (http://blog.csdn.net/byxdaz/article/details/5815677) 》 (是否引证: 否)</p> <p>1. 策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。4). 基于知识理解的分词方法。 该方法主要基于句法、语法分析, 并结合语义分析, 通过对上下文内容所提供信息的分析对词进行定界, 它通常包括三个部分: 分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力, 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。因此目前基于知识的分词系统还处在试验阶段。5). 一种新的分词方法 并行分词方法: 这种分词方法借助于一个含有分词词库的管道进行, 比较匹配过程是分步进行的, 每一</p> <p>中文分词原理 - xiaomin1991222的专栏 - 博客频道 - CSD</p>

《网络

(<http://blog.csdn.net/xiaomin1991222/article/details/509>

(是否引证: 否)

1. 策算法决定最优的切分结果。它的优点在于可以发现所歧义并且容易将新词提取出来。 4). 基于知识理解的分词方法主要基于句法、语法分析,并结合语义分析,通过对所提供信息的分析对词进行定界,它通常包括三个部分:系统、句法语义子系统、总控部分。在总控部分的协调下,可以获得有关词、句子等的句法和语义信息来对分词歧义。这类方法试图让机器具有人类的理解能力,需要使用大量和信息。由于汉语语言知识的笼统、复杂性,难以将各种织成机器可直接读取的形式。因此目前基于知识的分词系统阶段。 5). 一种新的分词方法 并行分词方法: 这种分词于一个含有分词词库的管道进行,比较匹配过程是步进行

互联网舆情信息获取与分析研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-669365673.html>) 》 -

(是否引证: 否)

1. 策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。 ,4,基于知识理解的分词方法。 该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分,分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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。因此目前基于知识的分词系统还处在试验阶段。 ,5,并行分词方法 这种分词方法借助于一个含有分词词库的管道进行,比较匹配过程是分步进行的,每一步可以对进入管道中

[PDF][精品文]基于用户兴趣度的网络信息过滤模型研究 - 豆丁网 - 《互联网文档资源

(<http://www.docin.com/p-475602698.html>) 》 - (是否

引证: 否)

1. 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。 ,基于知识理解的分词方法该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分,分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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,因此目前基于知识的分词系统还处在试验阶段,对于任何一个成熟的分词系统来说,单独依靠某一种算法都不可能实现,需要综合不同的算法。 ,特征选择方法训练集中包含了大量的词

[PDF][精品文]基于用户兴趣度的网络信息过滤模型研究

- 豆丁网 - 《互联网文档资源

(<http://www.docin.com/p-475602698.html>) 》- (是否引证: 否)

1. 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。基于知识理解的分词方法该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分,分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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,因此目前基于知识的分词系统还处在试验阶段,对于任何一个成熟的分词系统来说,单独依靠某一种算法都不可能实现,需要综合不同的算法。特征选择方法训练集中包含了大量的词

互联网短文本信息分类关键技术研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-108213576.html>) 》- (是否引证: 否)

1. 策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。4,基于知识理解的分词方法。该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分,分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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。因此目前基于知识的分词系统还处在试验阶段。5,并行分词方法 这种分词方法借助于一个含有分词词库的管道进行,比较匹配过程是分步进行的,每一步可以对进入管道中的词

基于语义分析的网络信息采集算法研究与应用 赵佳鹤-《大连理工大学硕士论文》-2006-12-05 (是否引证: 否)

1. 自动消除歧义的优点。(3)基于知识理解的分词方法该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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,因此目前基于知识的分词系统还处在试验阶段。对于任何一个成熟的分词系统来说,单独依靠某一种算法都不可能实现,需要综合不同的算法。

基于用户兴趣度的网络信息过滤模型研究 王翠平-《山

东师范大学硕士论文》-2007-04-27 (是否引证: 否)

1. 的优点。 2.3.3 基于知识理解的分词方法 该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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,因此目前基于知识的分词系统还处在试验阶段。对于任何一个成熟的分词系统来说,单独依靠某一种算法都不可能实现,需要综合不同的算法。

社交网络舆情传播监督管理系统的设计与实现 郑罡-《黑龙江大学硕士论文》-2015-10-20 (是否引证: 否)

1. 模型和决策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。
4、基于知识理解的分词方法该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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。因此目前基于知识的分词系统还处在试验阶段。
5、一种新的分词方法——并行分词方法这种分词方法借助于一个含有分词词库的管道进行,比较匹配过程是分步进行的,每一步可以对

民航发动机健康管理数据库设计与故障诊断 张煜-《中国民航大学硕士论文》-2016-06-30 (是否引证: 否)

1. 分结果。它的优点 在于可以发现所有的切分歧义并且容易将新词提取出来。
4、基于知识理解的分词方法。该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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。因此目前基于知识的分词系统还处在试验阶段。而由于已有的民航发动机故障案例信息的文本描述只是并非长篇大段,且考虑到应用难度选取机械分词的方法对故障案例进行分

中文分词技术(中文分词原理) - u010384318的专栏 - 博客 - CSDN.NET - 《网络
(<http://blog.csdn.net/wbgxx333/article/details/11178693>)
(是否引证: 否)

1. 策算法决定最优的切分结果。它的优点在于可以发现所

切分歧义并且容易将新词提取出来。4). 基于知识理解的方法。该方法主要基于句法、语法分析,并结合语义分析通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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,因此目前基于知识的分词系统还处在试验阶段。5). 一种新的分词方法 并行分词方法:这种分词方法借助于一个含有分词词库的管道进行,比较匹配过程是分步进行的,每一

文本相似性检测----中文分词技术 - Johline的博客 - CSDN 博客 - 《网络
(<http://blog.csdn.net/johline/article/details/59109811>)》
(是否引证: 否)

1. 策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。4). 基于知识理解的分词方法。该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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,因此目前基于知识的分词系统还处在试验阶段。5). 一种新的分词方法 并行分词方法:这种分词方法借助于一个含有分词词库的管道进行,比较匹配过程是分步进行的,每一

中文分词技术及其应用初探 余战秋-《电脑知识与技术》-2004-11-27 (是否引证: 否)

1. 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。2.3基于知识理解的分词方法该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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,因此目前基于知识的分词系统还处在试验阶段。对于任何一个成熟的分词系统来说,单独依靠某一种算法都不可能实现,需要综合不同的算法。据了解,我国海量科技的分词算法就采用

中文分词算法 - llandrj的博客 - CSDN博客 - 《网络
(<http://blog.csdn.net/llandrj/article/details/49412141>)》
(是否引证: 否)

1. 统计语言模型和决策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。基于知识理解的分词方法 基于句法、语法分析,并

结合语义分析，通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分：分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力，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。因此目前基于知识的分词系统还处在试验阶段。参考 <http://www.cnblogs.com/flish/archive/2011/08/08/2131031>。

《跨语言文本相似性检测》第一周一前期调研 - Johline的博客 - CSDN博客 - 《网络

(<http://blog.csdn.net/johline/article/details/59111894>) 》

(是否引证：否)

1. 模型和决策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。基于知识理解的分词方法 该方法主要基于句法、语法分析，并结合语义分析，通过对上下文内容所提供信息的分析对词进行定界，它通常包括三个部分：分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力，需要使用大量的语言知识和信息。由于汉语语言知识的统、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。因此目前基于知识的分词系统还处在试验阶段。一种新的分词方法：并行分词方法 这种分词方法借助于一个含有分词词库的管道进行，比较匹配过程是分步进行的，每一步可以

信息检索论文 - 《互联网文档资源

(<http://wenku.baidu.com/view/708f660390c69ec3d5bb7>)

(是否引证：否)

1. 决策算法决定最优的切分结果。它的优点在于可以发现歧义并且容易将新词提取出来。（三）基于知识理解的分法主要基于句法、语法分析，并结合语义分析，通过对上供信息的分析对词进行定界，它通常包括三个部分：分词语义子系统、总控部分。在总控部分的协调下，分词子系关键词、句子等的句法和语义信息来对分词歧义进行判断。让机器具有人类的理解能力，需要使用大量的语言知识和语语言知识的笼统、复杂性，难以将各种语言信息组织成取的形式。因此目前基于知识的分词系统还处在试验阶段种新的分词方法 并行分词方法：这种分词方法借助于一个库的管道进行，比较匹配过程是分步进行的，每一

关于现代中文分词技术的综述 - 《互联网文档资源

(<http://wenku.baidu.com/view/bf11ce4be518964bcf847c>)

(是否引证：否)

1. 决策算法决定最优的切分结果。它的优点在于可以发现歧义并且容易将新词提取出来。3.4基于知识理解的分词主要基于句法、语法分析，并结合语义分析，通过对上下供信息的分析对词进行定界，它通常包括三个部分：分词法语义子系统、总控部分。在总控部分的协调下，分词子得有关词、句子等的句法和语义信息来对分词歧义进行判

法试图让机器具有人类的理解能力,需要使用大量的语言信息。由于汉语语言知识的笼统、复杂性,难以将各种语言机器可直接读取的形式。因此目前基于知识的分词系统还

段。3.5一种新的分词方法 并行分词方法:这种分词方法含有分词词库的管道进行,比较匹配过程是分步进行的,每

互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-669364954.html>)》 - (是否引证: 否)

1. 策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。 ,4,基于知识理解的分词方法。该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分,分词子系统、句法语义子系统、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 上海交通大学工学硕士学位论文 互联网媒体信息热点主动发现系统的详细设计 22法和语义信息来对分词歧义进行判断。这类方法试

2. 子等的句 上海交通大学工学硕士学位论文 互联网媒体信息热点主动发现系统的详细设计 22法和语义信息来对分词歧义进行判断。这类方法试图让机器具有人类的理解能力,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。因此目前基于知识的分词系统还处在试验阶段。 ,5,并行分词方法 这种分词方法借助于一个含有分词词库的管道进行,比较匹配过程是分步进行的,每一步可以对进入管道中的词

互联网媒体信息热点主动发现关键技术研究 - 豆丁网 - 《互联网文档资源 (<http://www.docin.com/p-122549484.html>)》 - (是否引证: 否)

1. 策算法决定最优的切分结果。它的优点在于可以发现所有的切分歧义并且容易将新词提取出来。 ,4,基于知识理解的分词方法。该方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分,分词子系统、句法语义子系统、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 上海交通大学工学硕士学位论文 互联网媒体信息热点主动发现系统的详细设计 22法和语义信息来对分词歧义进行判断。这类方法试

2. 子等的句 上海交通大学工学硕士学位论文 互联网媒体信息热点主动发现系统的详细设计 22法和语义信息来对分词歧义进行判断。这类方法试图让机器具有人类的理解能力,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。因此目前基于知识的分词系统还处在试验阶段。 ,5,并行分词方法 这种分词方法借助于一个含有分词词库的管道进行,比较匹配过程是分步进行的,每一步可以对进入管道中的词

基于条件随机场和空间推理的地理编码方法 周海-《解放军信息工程大学硕士论文》-2015-04-20 (是否引证: 否)

1. 中文分词思想:在分词的同时进行句法、语义分析,利

		<p>用句法信息和语义信息来处理歧义现象。该系统通常包括三个部分:分第15页词子系统、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。这种分词方法需要大量语言知识和信息,由于汉语语言知识笼统和复杂,很难将各种语言信息组织成机器可直接读取的形式。目前,基于</p> <p>2. 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。这种分词方法需要大量语言知识和信息,由于汉语语言知识笼统和复杂,很难将各种语言信息组织成机器可直接读取的形式。目前,基于理解的方法的研究有专家系统和神经网络的算法[25][62][67],主要利用语法、句法以及语义分析及上下文信息进行分词和标注,</p> <p>英文自动问答系统中数值型问句的理解研究 赵龙-《大连海事大学硕士论文》-2016-05-01 (是否引证: 否)</p> <p>1. 匹配,从而降低匹配错误率。(2)基于理解分词,是指分词过程中利用分词系统模拟人对句子的理解来进行句法语义分析。系统包括总控部分、分词子系统、句法语义系统。在总控部分控制下,分词子系统通过句法语义系统获得有关词、句子等的句法和语义信息并对分词产生的歧义进巧判断。(3)基于统计分词,是指通过相邻的字同时出现的次数进行判断和统计,出现的次数越多越有可能构成一个词语。这种方法不</p>
8	<p>此处有 67 字相似</p> <p>难以将各种语言信息组织成机器可直接读取的形式。因此目前基于知识的分词系统还处在试验阶段。[8] JIEBA 中文分词模块支持三种分词模式: A. 精确模式,试图将句子最精确地切开,适合文本分析; B. 全模式,把句子中所有的可以成词的词语都扫描出来,速度非常快,但是不能解决歧义; C. 搜索引擎模式,在精确模式的基础上,对长词再次切分,提高召回率,适合用于搜索引擎</p>	<p>基于Hadoop平台分布式SVM分类研究 蔡鑫怡;-《电脑迷》-2018-06-21 (是否引证: 否)</p> <p>1. 有可能成词情况所构成的有向无环图(DAG)。对于未登录词,采用了基于汉字成词能力的HMM模型,使用了Viterbi算法。支持三种模式:精准模式,将句子最精准地切开;全模式,把句子中所有可以成词的词语都扫描出来;搜索引擎模式,在精准模式的基础上,对长词再次切分,提高召回率。此外jieba还支持繁体分词,支持自定义词典,MIT授权协</p>
9	<p>此处有 31 字相似</p> <p>切开, 适合文本分析; B. 全模式, 把句子中所有的可以成词的词语都扫描出来, 速度非常快, 但是不能解决歧义; C. 搜索引擎模式, 在精确模式的基础上, 对长词再次切分, 提高召回率, 适合用于搜索引擎。全模式是这样的: 我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学。而精确模式则是这样的:</p>	<p>基于Hadoop平台分布式SVM分类研究 蔡鑫怡;-《电脑迷》-2018-06-21 (是否引证: 否)</p> <p>1. 型,使用了Viterbi算法。支持三种模式:精准模式,将句子最精准地切开;全模式,把句子中所有可以成词的词语都扫描出来;搜索引擎模式,在精准模式的基础上,对长词再次切分,提高召回率。此外jieba还支持繁体分词,支持自定义词典,MIT授权协议。1.5词/文本向量化(Word2Vec):word2vec</p>
	<p>此处有 95 字相似</p> <p>分词; 添加自定义词典, 即开发者可以指定自己自定义的词典, 以便包含 JIEBA 词库里没有的词; 关键词提取; 词性标注; 并行分词等。JIEBA分词的算法策略是基于前缀词典实现高效的词图扫描, 生成句子中汉字所有可能成词情况所构成的有向无环图。此后采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合。JIEBA还可以添加自定义词典。目前, 网络迅速发展, 出现了许多词典中没有但实际上十分常用的词语,</p>	<p>基于经济普查大数据的上海“三新”经济发展态势研究 杭敬;苑立波;张志远;-《统计科学与实践》-2016-11-25 (是否引证: 否)</p> <p>1. 使用Python3.5软件下的jieba 0.38分词组件对经济普查调查单位的“主要业务活动(或主要产品)”文本集合进行分词。1.分词算法该组件基于前缀词典实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图(DAG);采用了动态规划查找最大概率路径,找出基于词频的最大切分组合;对于未登录词,采用了基于汉</p>

10	<p>在新闻报道中也频频出现（</p>	<p>字成词能力的HMM模型,使用了Viterbi算法。2.分词模式应对不同的分词需要,该组件提供了</p> <p>媒体情绪能够影响投资者情绪吗——基于新兴市场门槛效应的研究 黄宏斌;刘树海;赵富强;-《山西财经大学学报》-2017-10-30 1 (是否引证: 否)</p> <p>1. 感词典构建。考虑构建情感词典的需要,我们首先对数据集的所有新闻进行了分词,中文分词工具采用了Jieba分词工具。该工具基于前缀词典实现词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图。采用了动态规划查找最大概率路径,找出基于词频的最大切分组合。对于未登录词,采用了基于汉字成词能力的HMM模型,使用了Viterbi算法。对数据集进行分词后,一共产生了287 632</p> <p>基于Hadoop平台分布式SVM分类研究 蔡鑫怡;-《电脑迷》-2018-06-21 (是否引证: 否)</p> <p>1. .4jieba分词主要原理jieba分词目前最好用中文分词之一,它基于Trie树结构实现高效的词图扫描,生成句子中汉字。采用了动态规划查找最大概率路径,找出基于词频的最大切分组合有可能成词情况所构成的有向无环图(DAG)。对于未登录词,采用了基于汉字成词能力的HMM模型,使用了Viterbi算法。</p> <p>吕苏语口语标注语料的自动分词方法研究 于重重;操镭;尹蔚彬;张泽宇;郑雅;-《计算机应用研究》-2016-07-15 1 (是否引证: 否)</p> <p>1. 型的分词方法和基于词感知机算法的分词方法等[10~12]。目前主流的分词方法中,结巴(jieba)分词方法[13,14]采用了动态规划查找最大概率路径,找出基于词频的最大切分组合,对于未登录词,采用了基于汉字成词能力的隐马尔可夫模型,使用了维特比算法。Jieba分词是国内程序员用Python开发的一</p>
11	<p>此处有 34 字相似</p> <p>上十分常用的词语,在新闻报道中也频频出现(如:“厉害了”,“惊”),这些词语的使用意与原意并不相同,此时需要建立新词典。虽然JIEBA有新词识别能力,但是自行添加新词可以保证更高的正确率,用户也可以通过建立新词典来防止歧义。JIEBA还可以去停用词。去停用词的意思是有一个文件存放要改的文章,一个文件存放</p>	<p>基于经济普查大数据的上海“三新”经济发展态势研究 杭敬;苑立波;张志远;-《统计科学与实践》-2016-11-25 (是否引证: 否)</p> <p>1. 句子最精确地切开,适合本文的文图1“三新”经济的传统认定方法与文本挖掘认定方法比较本分析。3.自定义“三新”经济特征词典虽然jieba有新词识别能力,但是自行添加新词可以保证更高的正确率。由于本文的“三新”经济特征词典大部分为新登录词,本文使用“添加自定义词典”的功能指定该词典,以便能够对主要业务活动进行精</p>
	<p>此处有 167 字相似</p> <p>落了,从未感觉到时间过得如此之快。蓦然回首,有太多难以忘怀的时刻。在英才计划研究报告结题论文即将完成之际,本文作者要特别感谢牛建伟教授和李青峰老师的的热情关怀和悉心指导。在本文作者研究和论文撰写的过程中,教授和老师都倾注了大量的心血和汗水,无论是在论文的选题、构思和资料的收集方面,还是在论文的研究方法以及成文定稿方面,本文作者都得到了教授和老师悉心细致的教诲和无私的帮助,特别是牛教</p>	<p>论四头双导程蜗杆车削挂轮的选配 梁宗斌;-《现代商贸工业》-2017-08-05 (是否引证: 否)</p> <p>1. 我非常珍惜这一具有挑战性的工作,愿做一个永不服输的探索者。在本论文完成之际,向所有帮助过我的老师、工友们表示衷心的感谢!特别要感谢指导老师罗国荣老师的热情关怀和悉心指导。在我撰写论文的过程中,罗国荣老师倾注了大量的心血和汗水。从开题报告的修改、论文的架构拟定到最终定稿,他给予了殷切的指导,提出了许多宝贵的意见。无论是在论文的选题、构思和资料的</p>

12	授广博的学识、深厚的学术素养、严谨的治学精神和一丝不苟的工作作风, 以及李老师解答问题的耐心、对于本文作者的爱心, 都使本文作者终生受益, 本文作者在此表示真诚地感激和诚挚的谢意。 在论文的	收 2. 罗国荣老师倾注了大量的心血和汗水。从开题报告的修改、论文的架构拟定到最终定稿,他给予了殷切的指导,提出了许多宝贵的意见。无论是在论文的选题、构思和资料的收集方面,还是在论文的研究方法以及成文定稿方面,都得到了罗国荣老师悉心细致的教诲和无私的帮助,特别是他严谨的治学精神和一丝不苟的工作作风使我受益匪浅,在此表示真诚地感谢和深深的谢意。由于理论知识水平比较有限,论文中的有些方式、方法的阐述难免有疏漏和不足的地
		熔融金属红外热像测温精度的研究 高悦-《华北理工大学硕士学位论文》-2016-12-05 (是否引证: 否) 1. 学术上的精心指导和生活上的关怀表示敬意和感谢。同时我要感谢李北丹导师的热情关怀和悉心指导。在我撰写论文的过程中, 无论是在论文的选题、构思和资料的收集方面, 还是在论文的研究方法以及成文定稿方面, 我都得到了李老师悉心细致的教诲和无私的帮助在此表示深深的谢意。 此外, 我要感谢家人以及同学的支持与关怀, 感谢他们对我的生活无微不至的 关怀和照顾, 在我困难
		金泽大厦建筑工程项目绿色施工管理研究 胡通文-《天津工业大学硕士学位论文》-2018-01-26 (是否引证: 否) 1. . 51 天津工业大学硕士学位论文52 mm致谢 论文完成之际, 首先我要以最诚挚的谢意, 感谢我的导师, 感谢他的热心关 怀和悉心指导。在我撰写论文的过程中, 教授老师倾注了大量的心血和汗水。其 次, 感谢老师们为我传道授业解惑的, 在学习过程中, 我不仅学到了专业知识, 也从你们的身上学到了治学和人生态度。老师敏锐的学
		互联网短文本信息分类关键技术研究 - 豆丁网 - 《互联网文档资源 (http://www.docin.com/p-108213576.html) 》 - (是否引证: 否) 1. 中遇到的困难,他都耐心地予以提示或解答。在本文写作期间,李翔老师给了我很多帮助和指点,使我能够顺利地完成本文的写作。李翔老师拥有渊博的知识、敏锐的洞察力、严谨的工作态度和踏实的作风,是我终生学习的榜样。在此向李翔老师致以最诚挚的谢意。 还要对林祥老师表示感谢。他带领我参与到实验室的项目中,在项目的开发



跨语言检测结果: 0%

原文内容	相似内容来源
指 标	
疑似剽窃文字表述	
1.	
基于字符串匹配分词, 机械分词算法。将待分的字符串与一个充分大的机器词典中的词条进行匹配。分为正向匹配和逆向匹配; 最大长度匹配和最小长度匹配; 单纯分词和分词与标注过程相结合的一体化方法。所以常用的有: 正向最大匹配, 逆向最大匹配, 最少切分法。实际应用中, 将机械分词作为初分手段, 利用语言信息提高切分准确率。优先识别具有明显特征的词, 以这些词为断点, 将原字符串分为较小字符串再机械匹配, 以减	

少匹配错误率,

2. 方法主要基于句法、语法分析,并结合语义分析,通过对上下文内容所提供信息的分析对词进行定界,它通常包括三个部分:分词子系统、句法语义子系统、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。这类方法试图让机器具有人类的理解能力,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。
3. 支持三种分词模式: A. 精确模式,试图将句子最精确地切开,适合文本分析; B. 全模式,把句子中所有的可以成词的词语都扫描出来,
4. 分词等。 JIEBA分词的算法策略是基于前缀词典实现高效的词图扫描,生成句子中汉字所有可能成词情况所构成的有向无环图。此后采用了动态规划查找最大概率路径,找出基于词频的最大切分组合。
5. 特别感谢牛建伟教授和李青峰老师的的热情关怀和悉心指导。在本文作者研究和论文撰写的过程中,教授和老师都倾注了大量的心血和汗水,无论是在论文的选题、构思和资料的收集方面,还是在论文的研究方法以及成文定稿方面,本文作者都得到了教授和老师悉心细致的教诲和无私的帮助,特别是牛教授广博的学识、深厚的学术素养、严谨的治学精神和一丝不苟的工作作风,

说明: 1.仅可用于检测期刊编辑部来稿,不得用于其他用途。
2.总文字复制比:被检测文章总重合字数在总字数中所占的比例。
3.去除引用文献复制比:去除系统识别为引用的文献后,计算出来的重合字数在总字数中所占的比例。
4.去除本人已发表文献复制比:去除作者本人已发表文献后,计算出来的重合字数在总字数中所占的比例。
5.指标是由系统根据《学术期刊论文不端行为的界定标准》自动生成的。
6.红色文字表示文字复制部分;绿色文字表示引用部分。
7.本报告单仅对您所选择比对资源范围内检测结果负责。

8.Email: amlc@cnki.net  <http://e.weibo.com/u/3194559873>  http://t.qq.com/CNKI_kycx
<http://check.cnki.net/>