北京交通大学

硕士学位论文

搜索引擎中中文分词与纠错模块的设计与实现

姓名：李晓东

申请学位级别：硕士

专业：软件工程

指导教师：陈旭东

20081201

摘要

随着互联网的蓬勃发展，各种应用服务层出不穷，搜索引擎是其中最流行的一种服务，仅次于电子邮件。伴随着搜索引擎的普及，越来越多的人通过使用搜索引擎，获取日常工作和生活中需要的信息。

搜索引擎涉及多种技术，其中，自然语言处理技术是其中重要的一种，它可以帮助搜索引擎提高查询准确度，并丰富搜索引擎的特色功能。中文分词技术与中文纠错技术是自然语言处理技术的两个分支，可应用在搜索引擎的分析系统与检索系统中，对于提高用户检索效率和检索结果准确度具有十分重要的意义。

论文选题来源于一个提供旅游信息搜索的搜索引擎系统。文中对中文分词技术和中文查询词纠错技术进行了研究，并从软件工程的角度出发，设计与实现了系统的中文分词模块与纠错模块。具体的工作包括：

（1）对一种基于规则的分词算法进行了改进，使用双 hash 词典结构降低了算法正向、反向最大匹配时的匹配次数；使用正反最大匹配切分策略取代逐词切分策略，降低了算法切分字符串的次数，提高了算法的分词效率。

（2）通过相关算法研究提出并实现了一种基于拼音 hash 词典的同音别字词纠错算法，应用于搜索引擎系统中，用于同音别字词的纠错。

（3）改进了双字驱动词典的结构，并结合新的字符串模糊匹配算法对漏字多字查询词进行纠错，增强了系统对漏字多字查询词的纠错能力。

（4）将词典技术与 Ajax 技术相结合，实现了查询词智能提示功能。

（5）采用 N-gram 切分的新词识别算法，用于发现未登录词，实现分词词库的动态更新，提高了中文分词模块的分词准确度。

（6）在漏字多字查询词纠错建议中，选择相似度在一定范围内、词频数高于某一阀值的纠错建议词条，作为相关查询词，模拟了相关查询词推荐功能的实现。

关键词：搜索引擎；中文分词；中文纠错；
分类号：

ABSTRACT

ABSTRACT:

With the vigorous development of the Internet, there have been a wide range of applications; the search engine is one of a wide range of applications, second only to e-mail services. With the popularity of search engines, more and more people through the use of search engines to get the information they need in work and life.

Search engine covers a broad range of technology, of which natural language processing is an important one, it can help search engine to improve the query precision, and enrich the functions of search engine. Chinese word segmentation and error correction technology are two branches of natural language processing technology, and are used in the analysis system and retrieval system; improve the user search for efficiency and accuracy of search results.

Dissertation topics from a travel search engine system. Papers on the Chinese word segmentation Chinese technical and query error correction techniques have been studied. From the perspective of software engineering, design and implement the Chinese word search engine modules with error correction module. Specific works include:

(1) Improve a rule-based segmentation algorithm. use double hash Dictionary to reduce of the number of the forward, reverse biggest match; use the positive and negative biggest matching segmentation strategy to replace the term-by-segmentation strategy , The segmentation algorithm reduces the number of strings to improve the efficiency of the segmentation algorithm.

(2) Through studying the relevant algorithms, implement a pinyin hash dictionary algorithm to corect the words including wrong character which has the same pinyin. Implement the new algorithms to the search engine system for correction.

(3) Improved two-word dictionary-driven structure, combined with a new string matching algorithms blurred leaking word of the multi-word query for correction words, to strengthen the system of multi-word query word leaking word of the error correction capability.

(4) Dictionary technology and Ajax technology have prompted inquiries word intelligent features.

(5) Using N-gram split of the term of the new identification method for the find unknown word, the word thesaurus to achieve dynamic update and improve the Chinese

word segmentation module accuracy.

(6) omitted the word in the multi-word correction proposed query words, the choice of similarity to a certain extent, the frequency is higher than the threshold of a correction term recommendations, as a related query, the relevant inquiry analog functions recommended by the word Realized.

KEYWORDS: Search engine; Chinese word segmentation; Chinese correction;

CLASSNO:

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：李晓东 签字日期：2008 年 12月 05日

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：李晓东　　导师签名：陈小平

签字日期：2008年12月06日　　签字日期：2008年12月05日

致谢

本论文的工作是在我的导师陈旭东老师的悉心指导下完成的，陈旭东老师严谨的治学态度和科学的工作方法给了我极大的帮助和影响。在此衷心感谢三年来陈旭东老师对我的关心和指导。

陈旭东悉心指导我们完成了实验室的科研工作，在学习上和生活上都给予了我很大的关心和帮助，在此向陈旭东老师表示衷心的谢意。

陈旭东老师对于我的科研工作和论文都提出了许多的宝贵意见，在此表示衷心的感谢。

在实验室工作及撰写论文期间，张伟 、袁凯等同学对我论文中的相关研究工作给予了热情帮助，在此向他们表达我的感激之情。

另外也感谢家人，他们的理解和支持使我能够在学校专心完成我的学业。

1 引言

1.1 研究背景与意义

互联网的初期阶段，网站相对稀少，信息查找也较为简单。但伴随互联网爆炸性的发展，有数以亿计的网页向互联网的使用者提供信息，因此单纯依赖人工的搜索效率太低。为满足大众信息检索需求，专业搜索网站应运而生。在国内尽管搜索引擎发展起步相对较晚，但是国内搜索引擎却发展的异常繁荣，成为整个搜索引擎行业中一个重要的组成部分。

在搜索引擎系统发展的过程中，为了满足广大用户的需求，研究者不断地将其它技术应用到搜索引擎系统中，其中自然语言处理技术是其中重要的一项。

自然语言处理技术的应用，可以帮助搜索引擎提高查询精度，丰富搜索引擎的特色功能。以百度为例，百度支持中文分词、查询词纠错、相关查询词提示等功能，这些功能帮助百度更好地支持中文搜索，使其成为搜索引擎市场的后起之秀。

1.2 本文的主要工作

本文从软件工程设计和实现的角度对自然语言处理中的两种技术，中文分词与中文纠错技术进行了研究，主要工作有：

（1）对一种基于规则的分词算法进行了改进，使用双 hash 词典结构降低了算法正向、反向最大匹配时的匹配次数；使用正反最大匹配切分策略取代逐词切分策略，降低了算法切分字符串的次数，提高了算法的分词效率；

（2）通过相关算法研究提出并实现了一种基于拼音 hash 词典的同音别字词纠错算法，应用于搜索引擎系统中，用于同音别字词的纠错。

（3）改进了双字驱动词典的结构，并结合新的字符串模糊匹配算法对漏字多字查询词进行纠错，增强了系统对漏字多字查询词的纠错能力。

（4）将词典技术与 Ajax 技术相结合，实现了查询词智能提示功能。

（5）采用 N-gram 切分的新词识别算法，用于发现未登录词，实现分词词库的动态更新，提高了中文分词模块的分词准确度。

（6）在漏字多字查询词纠错建议中，选择相似度在一定范围内、词频数高于某一阀值的纠错建议词条，作为相关查询词，模拟了相关查询词提示功能的实现。

1.3 本文的内容组织

论文的内容一共由五章组成：

第一章是引言。引言中概述了本文研究背景、意义、内容以及应用背景。分析了自然语言处理技术对搜索引擎系统积极的影响与贡献。

第二章是技术综述。本章概述了搜索引擎的发展及其趋势、技术目标、原理、结构组成以及自然语言处理技术在搜索引擎系统中的作用；分析了中文的特点，介绍了中文分词技术的发展，中文分词技术的难点。介绍了中文纠错技术及其发展状况。

第三章是中文分词模块。本章介绍了中文分词模块在搜索引擎系统中的作用；分析了一种支持词库动态更新的模块设计方式，并采用该方式对中文分词模块进行了设计；分析了一种基于规则的中文分词算法，从词典结构，切分策略两个方面改进了算法；介绍了一种 N-gram 切分的新词识别算法，使用它进行未登录词的发现。

第四章是中文查询词纠错模块。本章介绍了中文查询词模块在搜索引擎系统中的作用，提出并实现了一种基于拼音 hash 词典的纠错算法，用于同音别字词的纠错；提出并实现了一种新的字符串模糊匹配算法，结合多字驱动词典对漏字多字词进行纠错；分析并实现了查询词智能提示功能以及相关查询词推荐功能。

第五章是总结。概述了本文的内容，并提出了下一步的工作。

1.4 本文的应用背景

本文设计并实现的中文分词模块与中文纠错模块，应用于一个主业务是旅游信息搜索的搜索引擎系统中。

2 技术综述

2.1 搜索引擎概述

2.1.1 搜索引擎的发展

搜索引擎是互联网上最受欢迎的应用之一，为用户提供信息检索的服务。

搜索引擎的发展是一个漫长的过程，人们对搜索引擎的认识也在不断的加深，搜索引擎主要有如下的服务方式[2]。

(1)目录式搜索引擎

在互联网的早期阶段，信息检索系统的数据信息通常是人工发现，依靠采集者手动地进行分类。用户可以方便地在这种人工维护，按照某种主题分类体系编制的信息系统去查找目标信息，这种系统就是人们所熟知的目录检索系统。此类系统的代表有早期的雅虎（Yahoo）、Open Directory Project（DMOZ）、LookSmart、About。

目录检索系统多采用网状结构，目录包含多层，层数多为4级，如图2.1所示。用户通过分级的目录的引导，寻找相关信息。例如，如用户想利用网络资源目录查找有关修电脑的信息，在雅虎分类搜索引擎上的检索路径是：所有类目->家庭服务->电脑/数码维修维护->修电脑/笔记本。在互联网的初级阶段，目录式搜索引擎风靡一时，它极大地提升了用户检索信息的效率。

目录式搜索引擎的特点是经过信息管理专业人员、分类专家的人工设计和编制，提高了检索的准确性，适合于查找综合性、概括性的主题概念，或对检索准确度要求较高课题；不足是数据库的规模相对较小，检索到的信息数量有限、更新不及时，不能及时地检索信息 、而且系统是涉及过多的人工维护，导致运营的成本过高。

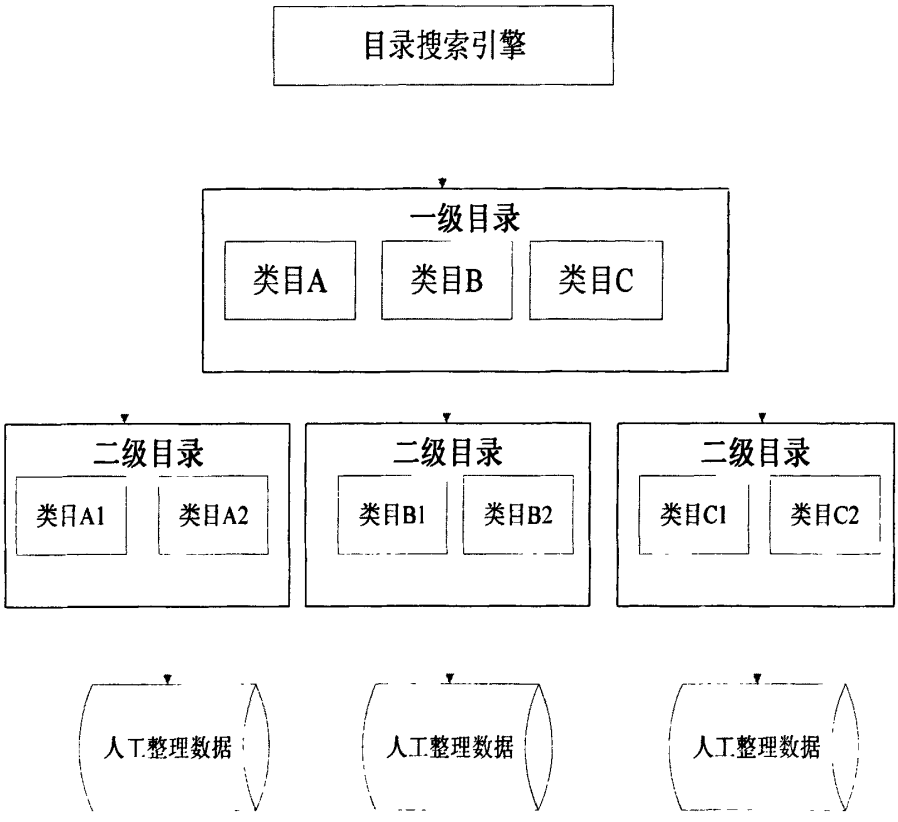


图 2.1 目录式搜索引擎

Figure 2.1 Directory search engine

(2)全文搜索引擎

全文搜索引擎是针对互联网所有网页进行全文检索的搜索引擎。全文搜索引擎系统中包含下载子系统，该子系统按照某种策略从互联网上提取各个网站的信息，这些被提取的信息由索引子系统建立索引，当用户使用全文搜索引擎进行查询时，系统根据用户的输入检索索引数据库，发现与用户查询条件匹配的相关记录，按一定的排列顺序将结果返回给用户。此类系统，国外具代表性的有 Google、Fast/AllTheWeb、AltaVista、Inktomi、Teoma、WiseNut、Lycos 等，国内著名的有百度等第二代商用搜索引擎。其中 Lycos 有其他搜索引擎有不同之处，它是租用其他引擎的数据库，按自定的格式排列搜索结果。

与目录式搜索引擎相比较，全文搜索引擎中人工参与的成分较少，数据采集，索引建立，索引查询都由系统自动完成，所以此类系统的优点是数据信息量大，数据更新及时；缺点是返回的结果信息太多，信息重复情况较为严重，用户需要对返回结果进行繁重的筛选工作。

(3)元搜索引擎

元搜索引擎（如图 2.2 所示）并不包含下载子系统，也就是说该类系统没有属于自己的数据，它是将用户的查询请求转发给不同的搜索引擎，将得到的所有结果集混合后进行重复排除以及重新排序，将处理得到的结果集返回给用户。著名的元搜索引擎有 InfoSpace、Dogpile、Vivisimo 等。

在搜索结果排列方面，有的直接按来源引擎排列搜索结果，如 Dogpile，有的则按自定的规则将结果重新排列组合，如 Vivisimo。

由于互联网的巨大规模，不同搜索引擎都有着不同的抓取策略，且抓取器遍历一次网络的周期都需要数周的时间，因此使用元搜索引擎的优点是利用多个搜索引擎的返回结果可以弥补只用一个搜索引擎时出现信息盲点的问题，但这种系统返回的结果会更多，导致用户需要进行更多的筛选工作，而且用户不能充分使用原搜索引擎的功能。

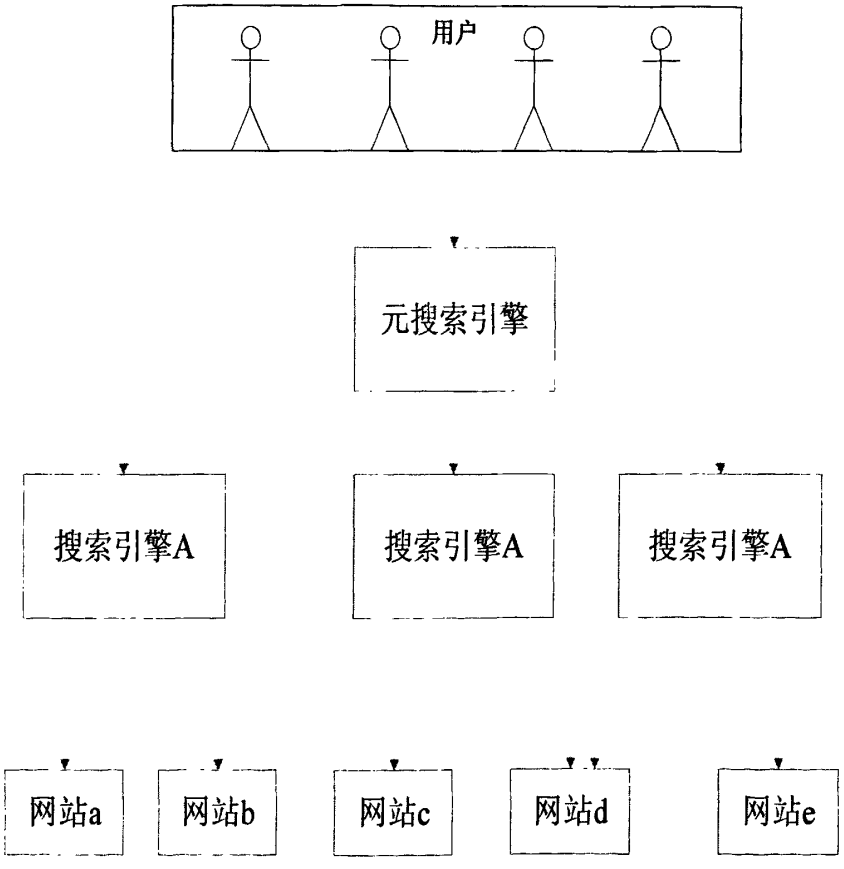


图 2.2 元搜索引擎

Figure 2.2 Meta search engine

(4)移动搜索引擎

移动搜索引擎是指以无线网络为数据传输承载层，将分布在互联网和移动网上的数据信息进行搜集整理，支持用户使用手机、PDA、短信、WAP 等多种特定搜索方式获取所需信息的搜索引擎。移动搜索引擎的搜索内容来自传统互联网资源和移动通信网络资源，是传统搜索引擎在无线领域的延伸。

当前手机用户的数量均远远超过互联网用户数，移动用户的壮大，WAP 市场的成熟和互联网搜索服务的普及为中国移动搜索产业带来了发展机会。尤其是独立 WAP 的兴起,极大地丰富了移动的内容，使得移动搜索服务成为移动互联网用户上网的必要手段。

从 2004 年开始，我国移动搜索业务开始启动，一些增值业务提供商开始进入移动搜索行业，至 2006 年初，众多传统互联网搜索引擎公司如 Google、百度开始正式进入到移动搜索行业。虽然目前移动搜索尚处于起步阶段，但是影响移动搜索的几大因素都呈现良好的增长状态，移动搜索市场将以每年 25-35%的速度增长，而移动搜索用户和移动搜索市场规模将在 2007-2009 年进入高速度增长期，同时随无线接入带宽的加大，移动搜索内容的进一步丰富，终端更加智能化，移动搜索业务将进入一个快速发展阶段。

2.1.2 搜索引擎的技术目标

支持快速，全面，准确，稳定的搜索是搜索引擎最主要的技术目标[2]。

(1)快速查询

从互联网建立开始，人类社会的发展进入信息社会发展的阶段，信息可以说是无处不在的，无论是人们的日常生活，还是工作学习都离不开有价值的信息，因此人们需要一个快速获取信息的方式，这就要求提供信息检索服务的搜索引擎必须满足高速提供服务的要求。

当今公开的搜索引擎的查询速度都在秒这个量级以下，商用搜索引擎的查询速度达到毫秒级，并且能够支持大规模用户的同时访问。

(2)全面查询

全面查询判断的标准是查全率，查全率是指检索出的相关网页数和所有的相关网页数的比率。例如在搜索引擎中查询“北京”，如果世界上包含“北京”这个关键词的网页数为 M，而实际该搜索引擎检索出这 M 条中的 N 条网页，那么查全率为 N/Mx100%，比值越高代表查得越全。

搜索引擎系统是否能查得全，主要取决于网页索引库的大小。索引的网页数

量越多，越有助于提高查全率。

(3)准确查询

准确查询的判断标准是查准率，查准率是检索出的相关文档数与检索出的文档总数的比率。例如在搜索引擎中查询“北京”，在实际检索出的网页数N中，只有M个网页是与查询“北京”相关的，那么查准率为N/M×100%。查全率和查准率的关系，见图2.3。

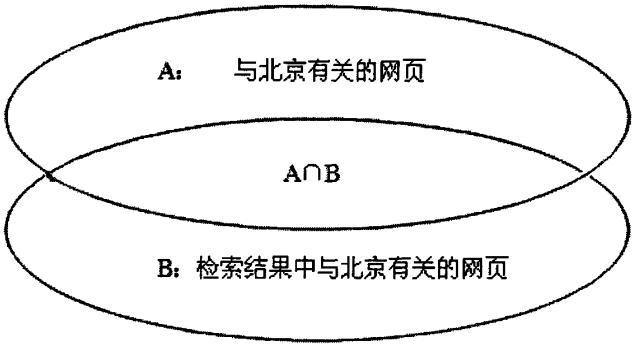


图 2.3 查全率和查准率的关系图

Figure 2.3 The relationship of Recall and precision

$$\text{查全率} = \frac{|A \cap B|}{|A|} \quad ; \quad \text{查准率} = \frac{|A \cap B|}{|B|}$$

其中对集合取||运算的结果表示集合的数量。

对于搜索引擎这种应用，查全率往往是不重要的。衡量的意义也不大，因为没有一个用户会把所有与查询相关的网页都浏览一遍。一般情况下，用户最为关注的仅仅为搜索结果中的前几条。而查准率在很大程度上决定了搜索的质量，在前10条搜索结果(搜索结果首页)中满足用户的查询目的是搜索引擎查准率的主要体现。是否能查得准，主要取决于网页排序。而在中文搜索引擎中，中文分词模块对查准率影响也相当大，分词结果越准确，查准率也相应越高。

(4)稳定查询

稳定是一种基本需求，任何系统包括搜索引擎必须是一个能够长期并稳定地提供服务的系统，因此系统的稳定运行是很重要的需求。在任何情况下可以牺牲检索质量和检索速度，但必须能够提供持续的信息检索服务。

对于搜索引擎来说，查询来自四面八方，查询词也千差万别，同时进行的查询量也非常巨大。稳定地满足这些查询需要，需要在系统的结构上做出权衡，在

文件存储方式、查询系统和索引系统设计等方面都需要考虑稳定性的因素。

2.1.3 搜索引擎的原理与结构

搜索引擎是自动采集数据，分析组织数据，并提供在线检索的一种系统。它的工作原理可以这样描述：采用一定搜索算法在万维网中挖掘指定的相关信息，对采集的数据预处理后分析建立索引，在建立的索引库中按一定的算法进行搜索排序。

根据搜索引擎的工作原理，搜索引擎被分为下载、分析、索引和查询 4 大系统[2]，搜索引擎系统结构[2]如图 2.4 所示。

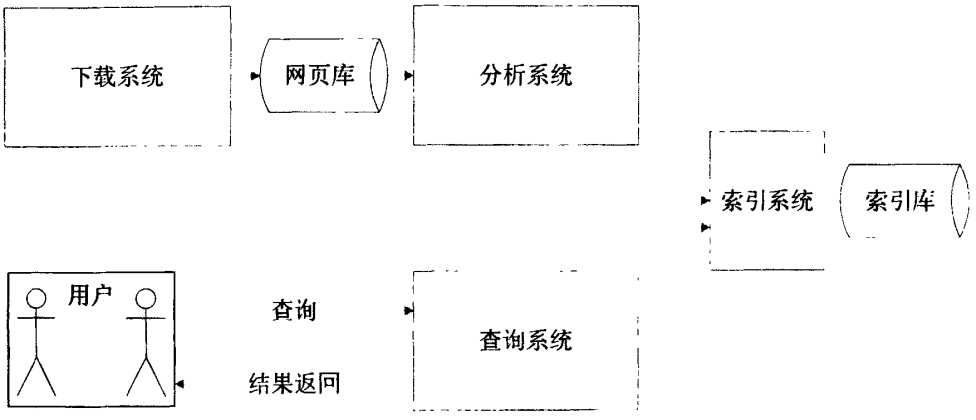


图 2.4 搜索引擎系统结构

Figure 2.4 Search engine system structure

以下 4 大系统的介绍。

(1)下载系统：下载系统负责从万维网上依据某种策略下载各种类型的网页，要求尽可能保持对万维网变化的同步。

(2)分析系统：把输入数据源文档通过一定的处理后存储在全文检索系统中的内部文档。系统并没有规定输入数据源的格式,它只提供了一个通用的结构来接受不同的数据源输入,输入的数据源理论上可以是任意文档, 只要能够设计相应的转换器将数据源构造成文档对象进行文本分词处理, 即可进行索引。全文检索系统目前不但提供了对文本类文档检索的支持,并且加入用户标签功能,使其他格式的数据,通过设置得到检索。

(3)索引系统：索引模块的功能是在加工处理后的数据信息抽取出索引项,用于

表示文档以及生成文档库的索引表。索引项是用来反映文档内容的，如关键词及其权重、短语、单字等等。索引项可以分为单索引项和多索引项两种。单索引项对于英文来讲是英语单词,比较容易处理,因为单词之间有天然的分隔符;对于中文等连续书写的语言，系统通过分词处理。

索引表使用倒排索引表,通过索引项查找相应的文档。索引表能够记录索引项在文档中出现的位置,以便检索模块计算索引项之间的相邻或接近关系。索引算法对索引模块的性能有很大的影响。一个检索系统的有效性在很大程度上取决于索引的质量。

(4)检索系统。当用户输入关键词搜索后，由检索系统程序从网页索引数据库中找到符合该关键词的所有相关网页。利用已经计算好关键词的相关度，进行排序，相关度越高，排名越靠前。最后，将搜索结果的链接地址和页面内容摘要，纠错建议，相关搜索提示等内容组织起来返回给用户。

2.1.4 搜索引擎与自然语言处理技术

自然语言处理技术对于搜索引擎来讲是一种重要的技术，可以帮助搜索引擎提高查询精度，丰富搜索引擎的特色功能。

例如，使用中文分词技术，可以帮助索引子系统建立有效的索引，在线检索系统正确分析用户的查询，返回正确的结果；使用未登录词识别技术帮助搜索引擎的查询词识别能力大幅度提升；使用智能关键词提示功能极大地方便了用户的查询词的输入；使用相关检索词智能推荐技术，在用户一次检索后，会提示相关的检索词，帮助用户查找相关的结果，提高了检索质量；对查询词进行纠错处理，使用户在不能给出正确查询描述信息的情况下，仍然有可能通过搜索引擎的帮助来找到目标信息。

2.2 中文分词技术

2.2.1 中文特点

中文分词是自然语言处理技术中的一个分支，在介绍中文分词技术前，我们需要先了解下中文的特点。

(1)字符数量多

中文常用字符有3000左右，字典中收录的字数要远远超过3000，1990年徐仲舒主编的《汉语大字典》，收字数为54678个，1994年冷玉龙等的《中华字海》，

收字数更是多达 85000 字，这使得汉字在计算机系统的存储与由阿拉伯字母构成的文字有很大的区别，因为汉字的数量远超过一个字节可以表示的最大数字，因此汉字采用双字节，或多字节进行编码。汉字编码中现在主要用到的有三类，包括 GBK、GB2312 和 Big5，。

a：GB2312 又称国标码，由国家标准总局发布，1981 年 5 月 1 日实施，通行于大陆。新加坡等地也使用此编码。它是一个简化字的编码规范，当然也包括其他的符号、字母、日文假名等，共 7445 个图形字符，其中汉字占 6763 个。

b：Big5 又称大五码，主要为香港与台湾使用，即是一个繁体字编码。

c：GBK 是 GB2312 的扩展，是向上兼容的，因此 GB2312 中的汉字的编码与 GBK 中汉字的相同，而且 GBK 中还包含繁体字的编码

(2)词汇量大

汉语中常用的词有几万条，《现代汉语词典》中收录的词约有 6 万个。词由单个或多个字构成，使用得最多的是二字词，其次是单字词，然后是多字词，词条分布情况见表 2.1。虽然汉字词汇量已相当的丰富，但伴随着社会的发展，还是不断地有新词出现。

表 2.1 词条分布情况表

Table 2.1 The sheet of term distribution

词条字数	1	2	3	4	5	6	7
词条数	606	3527	693	622	83	36	3

(3) 一词多意、使用灵活、变化多样

导致理解容易出现歧义的情况。例如，“这台车没锁”这样简单的一个句子就有两种理解方式，这给计算机的中文分析工作带来了非常大的困难。

(4)书写习惯

中文不同于出拉丁字母构成的文字，比如说英语，不论在手工书写还是在计算机系统中，它的词与词之间用空格隔开，这使得计算机处理时可以非常方便地从文档中识别出一个一个的词。而在汉语的情况是，不论手写还是在计算机系统中，它都是以句子为单位，句间用标点符号进行区隔，在句内部，是连续的中文字符，字和词按一定规则排列，它们之间没有分隔。这样，如果要对中文文档进

行基于词的处理，必须先要对其进行词的切分处理，也就是中文分词的工作，才可以正确地识别出每一个词。

(5)发音与字形

中文字符的发音由拉丁字母构成的拼音确定的，但拼音数量有限，所以有非常多的同音字；而中文字符由偏旁、部首和基础字符构成，所以有相当多的形近字。

正是由于中文的上述特点，特别是书写习惯使得计算机必须使用专门算法才可以进行中文分词的工作。

2.2.2 中文分词技术介绍

中文的句子是由连续的汉字序列构成的，将一个中文的汉字序列按着一定的规则切分成有意义的词，就是中文分词。例如，句子“我是警察”在切分后得到结果是“我/是/警察”。

常见的中文分词算法有以下 4 种：

(1)基于词典与规则的分词方法

基于词典与规则的分词方法是先利用机械分词方法对汉字字符串进行粗分，而后利用设定的规则进一步确认切分结果的分词方法。它所使用的机械分词方法，就是按照一定的策略将待分析的汉字串与一个机器词典中的词条进行匹配，匹配成功，则识别出一个词条。按照扫描方向的不同，机械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，分为最大匹配和最小匹配。

常用的几种机械分词方法有：正向最大匹配法;逆向最大匹配法;最少切分。还可以将上述各种方法组合使用，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。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。统计结果表明，单纯使用正向最大匹配的错误率为 1/169，单纯使用逆向最大匹配的错误率为 1/245。

机械分词的精度不能满足实际的需要。实际使用的分词系统，都是以机械分词为基础，通过利用各种其它的语言信息和某种规则来进一步提高切分的准确率。

(2)理解式的分词方法

理解式的分词是通过让计算机模拟人对句子的理解，达到识别词的效果，也称人工智能法。其本质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。

使用这种方法的系统，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，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。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，因此过去基于理解的分词系统还处在试验阶段，近年随着人工智能中的基于心理学的符号处理方法和基于生理学的模拟方法应用到分词方法后,产生了专家系统分词法和神经网络分词法[10].

神经网络具有联想、容错、记忆、自适应、自学习和处理复杂多模式等优点. 不足的是网络连接模型表达复杂、训练过程较长、不能对自身的推理方法进行解释,对未在训练样本中出现过的新的词汇不能给子正确切分.

专家系统具有显式的知识表达形式,知识容易维护,能对推理行为进行解释,可利用深层知识来切分歧义字段;缺点是不能从经验中学习,当知识库庞大时难以维护及在进行多歧义字段切分时耗时较长。

(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的反映成词的可信度。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，会经常抽出一些共现频度高、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。实际应用的统计分词系统都要使用一部基本的分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。

（4）基于字的分词新方法

传统的分词方法，无论是基于规则的还是基于统计的，一般都依赖于一个事先编制的词表。自动分词过程就是通过词表和相关信息来做出词语切分的决策。

与上述方式不同，基于字标注的分词方法实际上是构词方法。即把分词过程视为字在字串中的标注问题。把分词过程视为字的标注问题的一个重要优势在于，它能够平衡地看待词表词和未登录词的识别问题。

在这种分词技术中，文本中的词表词和未登录词都是用统一的字标注过程来实现的。在学习架构上，既可以不必专门强调词表词信息，也不用专门设计特定

的未登录词识别模块。这使得分词系统的设计大大简化。在字标注过程中，所有的字根据预定义的特征进行词位特性的学习，获得一个概率模型。然后，在待分字串上，根据字与字之间的结合紧密程度，得到一个词位的标注结果。最后，根据词位定义直接获得最终的分词结果。在这样一个分词过程中，分词成为字重组的简单过程，其分词结果也是令人满意的。

2.2.3 中文分词技术的发展

在过去的几十年间，中文分词技术不断发展，下面列举3个具有代表性的国内中文分词系统[10]。

1.CDWS 分词系统[10]

国内第一个实用的自动分词系统，它采用的自动分词方法为最大匹配法，辅助以词尾字构词纠错技术。它是汉语自动分词实践的首次尝试，具有很大的启发作用和理论意义。

2.自动分词专家系统[10]

系统将专家系统方法完整地引入到分词技术中。系统使知识库与推理机保持相对独立，知识库包括常识性知识库和启发性知识库，词典使用首字索引数据结构。通过引入专家系统的形式，系统把分词过程表示成为知识的推理过程，即句子“分词树”的生长过程。

3．统计分词系统[10]

系统是一种典型的运用统计方法的纯切词系统，它试图将串频统计和词匹配结合起来。系统出三个部分构成:

①预处理模块，利用显式和隐式的切分标记(标点符号、数字、ASCII 字符以及出现频率高、构词能力差的单字词、数词+单字常用量词模式)将待分析的文本切分成短的汉字串，这大大地减少了需要统计的(无效)字串的数量和高频单字或量词边界串;

②串频统计模块，此模块计算各个已分开的短汉字串中所有长度大于1的子串在局部上下文中出现的次数，并根据串频和串长对每个这样的子串进行加权，再根据词缀集对字串的权值进行提升。如果某个汉字串的权值超过某一阀值，则将此汉字串作为一个新识别的词，将其存入一临时词库中。

③切分模块，首先用临时词库对每个短的汉字串进行切分，使用的是逐词遍历算法，再利用一个小型的常用词词典对汉字短串中未切分的子串进行正向最大匹配分词。对于短汉字串中那些仍未切分的子串，则将所有相邻单字作为一个权值很低的生词。其中每个模块都对待分析的文本进行了一次扫描，因而是三遍扫

描方法。此系统能够利用上下文识别大部分生词，解决一部分切分歧义，但是统计分词方法对常用词识别精度差的固有缺点仍然存在。

2.2.4 中文分词技术的难点

中文自然语言的理解和处理远比西文复杂，主要体现在以下几个方面[11,12]，主要体现在以下几个方面：

（1）分词的规范问题

词的确切概念难以标准化，词的应用领域不同，使得分词规范难以统一，需要达到的分词效果也有很大区别。

（2）歧义切分

对于特定的句子或字符串可能存在多种切分方法，不同的切分方法具有不同的含义，因此会导致歧义。歧义根据不同的情况又分为如下 4 种：

①交集型歧义,在字段 AJB 中，AJ∈W，并且 JB∈W，则 AJB 称为，其中 A，J，B 为字串，w 为词表．例如，“学历史”分为，学历 / 史或者是学 / 历史。

②多义型歧义字段，在字段 AB 中 AB∈W，A∈W，B∈W，则 AB 称为，也称为组合型歧义字段，其中 A，J，B 为字串，w 为词表。例如，a.“门 / 把手 / 坏 / 了”; b.“请 / 把 / 手 / 拿开”。其中“把手”为多义型歧义字段，(1)中“把手”不应该切分，(2)中“把”和“手”都是词，应该切分。

③伪歧义,在真实语言环境下，只有惟一可能的正确切分结果，这样的歧义称为伪歧义。例如：挨 / 批评，爱 / 国家，爱好 / 使，市 / 政府，部门 / 对，国 规定，大会堂 / 会见，等.

④真歧义,在不同的语言环境下,有两种以上可实现的切分结果,这样的歧义称为真歧义.

(3)新词识别

汉字系统是一个开放性系统，可能不断有新词产生，最典型的如人名、地名以

及各类术语，分词系统必须不断更新分词词典。

(4)分词理解的先与后

由于计算机需要靠词的信息来理解文章，因此它只能采用先分词后理解的方法，而分词需要以理解为基础，理解必须先分词。由此产生的逻辑问题决定了不可能有百分之百正确的分词方法。

2.3 中文纠错技术

2.3.1 中文纠错技术介绍

随着计算机技术的发展，方方面面的信息也通过多样的方式被录入到计算机系统中，常用的方式有键盘录入、OCR 识别、电子扫描，但由于人为因素，或录入方式的先天不足，这些方式都不能保证数据信息在录入计算机系统后都还是完全正确的，所以需要纠错技术对错误数据信息进行处理。

纠错技术是自然语言处理技术中的一个分支，是分析发现字符串搭配出错后，通过一定规则对字符串进行纠正的一种技术。而中文纠错技术就是指针对中文信息进行分析、发现错误搭配的汉字字符串并进行纠错处理，早期应用在文本校对系统中，伴随着文本校对系统的发展而不断地趋于成熟。近年来该技术被应用于搜索引擎系统中，用于推测用户输入，在确定查询信息出错时，给出纠错建议。

2.3.2 纠错技术的发展

纠错技术应用在文本校对系统，它的发展体现在文本校对系统的发展过程之中。文本校对系统的研究，始于 20 世纪 60 年代，1960 年 IBMThomasJ．Watson 研究中心在 IBM / 360 和 IBM / 370 用 UNIX 实现了一个 TYPO 英文拼写检查器；1971 年，斯坦福大学的 RalphGorin 在 DEC—10 机上实现了一个英文拼写检查程序 SpellL。

多年来，随着计算机技术的不断发展，新的输入技术不断涌现，如 OCR 识别、语音识别。开展拼写错误校对的研究更加迫切，这方面的研究也在不断取得进展，部分成果已经商品化。

国内在中文文本校对方面的研究开始的较晚，始于 20 世纪 90 年代初期，但发展速度较快。目前有许多科技公司和高等院校或研究机构都投入了一定的人力和财力开展这方面的研究，并取得了一些较好的成果。

目前，国内在文本自动校对方面的研究主要是针对汉语文本开展的。因为中文文本校对主要面向的是含有错误的文本，因此，汉语自然语言理解的研究也就成了计算机中文文本自动校对的基础。由于汉语与英语的不同，在对中文文本进行查错、纠错分析时，必须要基于自然语言的理解技术，通过研究上下文间的依存关系才能实现，因此某些适于英文单词校对的技术和方法对汉语文本并不太适用。

国内在自动文本查错方面主要采用三种方法：a.利用文本上下文的字、词和词性等局部语言特征，包括词性特征、同现特征或相互依存特征，甚至包括字形特征等；b.利用转移概率对相邻词间的接续关系进行分析；c.利用规则或语言学知识，

如语法规则、词搭配规则等。在实际的应用中各种方法并不独立使用，而是混合使用来增强自动文本查错的能力。

(1)基于上下文的局部语言特征的方法，这种方法的一种实现是综合考虑了汉语文本中字、词和词性的局部语言特征以及K距离的语言特征，利用这些上下文特征对目标集中的词进行选择。

(2)基于规则方法，这种方法的一种实现是利用校正文法规则对文稿进行校对，若句子满足校正文法规则，则根据规则把相应字词标记错误，但有限的规则很难覆盖大量难以预料的错误现象，查错能力有限。

(3)基于统计方法，这种方法的一种实现是利用综合近似字集替换，并用统计语言模型评分的方法，其基本思想足以事先整理好字形、字音、字义或输入码相近字的综合近似字集替换待校对句子中的每个汉字，产生许多候选字符串，利用统计语言模型对各候选字符串评分，将评分最高的字符串与待校对文本中的句子进行对照，即可发现错误之所在并提供相对应的正确字。

自动纠错是文本自动校对的一个重要组成部分，它为自动查错时侦测出的错误字符串提供修改建议，辅助用户改正错误。修改建议的有效性是衡量自动纠错性能的主要指标，它有两点要求：

a:提供的修改建议中应该含有正确或合理的建议；

b:正确或合理的修改建议应尽可能排列在所有建议的前面。因此，纠错修改建议的产生算法及排序算法是自动纠错研究的两个核心内容。

目前主要的纠错处理方式有如下3种：

(1)采用模式匹配方法对长词进行纠错处理，但没有充分利用出错字符串的特征，算法计算量大。

(2)替换字表结合主词典，通过加字和换字对侦测出来的错误字符串提供修改建议的纠错算法，但该算法的纠错建议局限于替换字表，没有考虑。上下文启发信息，主要考虑对错字这种错误类型进行纠错，对漏字、多字、易位、多字替换、英文单词拼写等错误类型的纠错能力较弱。

(3)基于似然匹配的纠错建议候选集产生的算法，对漏字、多字、易位、多字替换等错误类型的纠错能力有了较大的提高。

2.4 本章小结

本章阐述搜索引擎的基本原理及其结构，回顾了搜索引擎的发展过程，分析了自然语言处理技术对搜索引擎系统的作用；介绍了中文分词技术与中文纠错技术，及其发展状况，并分析了中两种技术的应用需求与难点。

3 中文分词模块的实现

3.1 中文分词模块介绍

本小节首先介绍模块与系统间的关系,后分析模块的作用。

中文分词模块包含于搜索引擎的子系统分析系统和检索系统中，中文分词模块与搜索引擎的分析子系统的关系如图 3.1 所示。

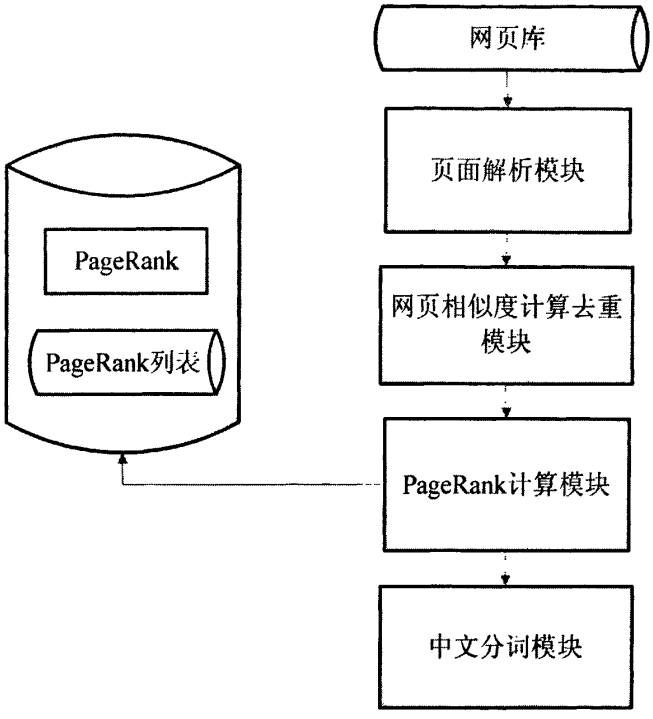


图 3.1 分析系统结构

Figure 3.1 Analysis system structure

中文分词模块与搜索引擎的检索子系统的关系如图 3.2 所示。

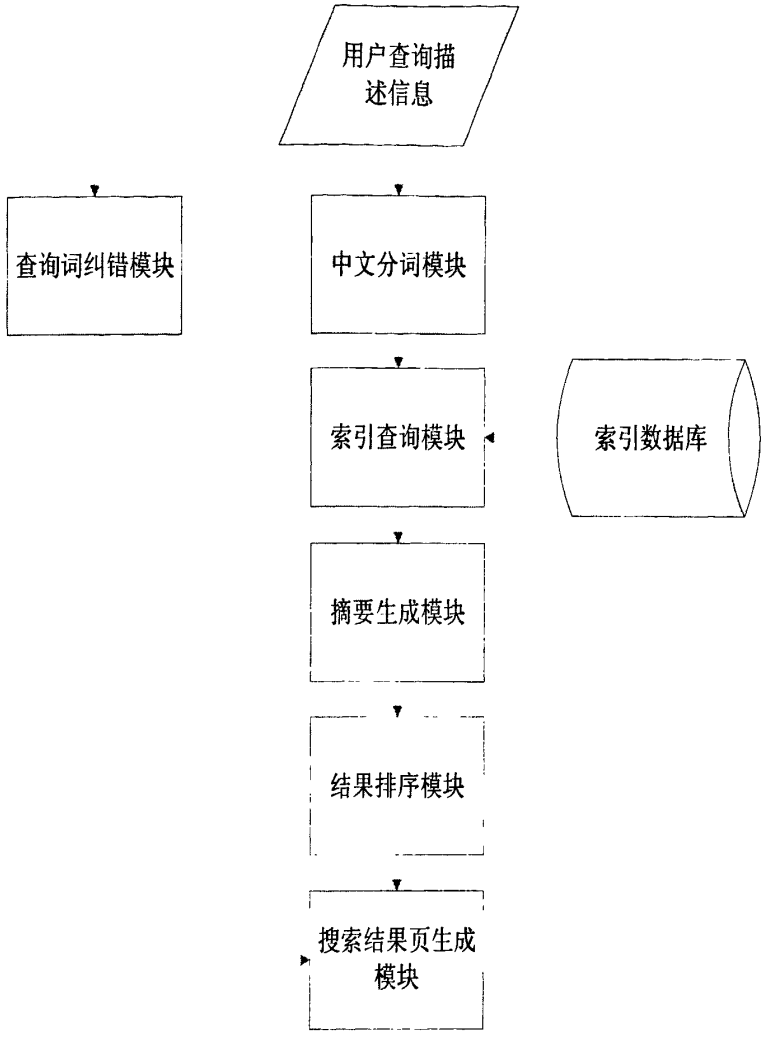


图 3.2 检索系统结构

Figure 3.2 Retrieval system

搜索引擎系统采用倒排表的方式组织索引,如表 3.1 所示。倒排表中存放关键字与包含关键字的文档号，系统通过查找关键字得到包含它的文档列表。

建立倒排表首先要得到文档中的关键字，英文文档中关键字间用空格区隔，计算机很容易得到关键字并建立索引，而中文信息的特点是关键字间没有明显的区隔，因此在分析子系统中，处理通过下载子系统获取的页面时，要先将页面中的信息抽取出来，放入文档中，通过中文分词模块将文档中的连续字符串分割为一个个区分开的关键字，为索引系统使用关键字建立倒排索引做好前期的准备工作。

表 3.1 倒排表结构

Table 3.1 Inverted table structure

关键字字段	文档号字段
……	……
关键字N_i	文档M_1、文档M_2…..
……	……

在检索子系统中，由于对用户的检索信息没有限制要求（例如，限制用户的检索信息中每个关键字必须用特殊符号隔离），所以同样需要中文分词模块将连续的检索信息切分为一组关键字后，利用这些切分出来的关键字去搜索索引数据库，找到包含这些关键字的页面，经过排序等处理后，返回结果。

3.2 模块的设计

在传统语言下切分精度很高的中文分词模块在互联网环境中，与声明的指标有很大差距[6]。传统语言环境相对于互联网的语言环境，语言定义较正规，出现的新词情况比较少，学习和测试的语料学均出自正规文体，字典和处理材料都是封闭的，因而测试得到的精确度较高；在互联网环境下，用来交流的语言，词汇已经发生了变化，因此对分词精确度产生消极的影响，导致分词精度的降低。

在中文分词模块中，传统词典的制定都是由语言领域的专家采集社会生活中用到的词语，按照一定规则编写而成的，词典的变更较小，变更周期较长。而在互联网的环境中，键盘、话筒，扫描仪，鼠标等新的工具的使用在为使用者带来方便的同时，也在影响着人们使用语言的方式。在互联网这个环境下，新生词汇不断地涌现出来，对这于这种情况，用传统词典来理解这些语言就会产生歧义，同样，中文分词模块依赖传统的词典来切分这些词语，也就很得到准确的切分结果。

网络语言与传统语言相比有一些显著的新特性：新词出现和更新速度加快；语言的使用习惯口语化、简写化、不规则化。依靠传统规范的静态词典，在互联网环境下存在一定不足。在互联网环境下，依靠传统规则的静态词典的分词系统，切分动态变化的不规则的网络语言也存在很大的缺陷，有待进一步改善。

由于网络语言的特性，中文分词模块切分词语所依赖的词典如果可以不断地更新，即词库能保证相对于待切分语料的完备性，其切分的精确度就能有很大提高。本文采用了一种支持词库自扩充的设计方式[6]，为中文分词的词库引入未登录词识别的功能。整个模块由未登录词识别和分词两部分组成,如图 3.3 所示，

a.未登录词识别：完成对文本中的未登录词抽取工作、更新词库；

b. 分词：完成字符串的切分工作、去除歧义得到切分结果。

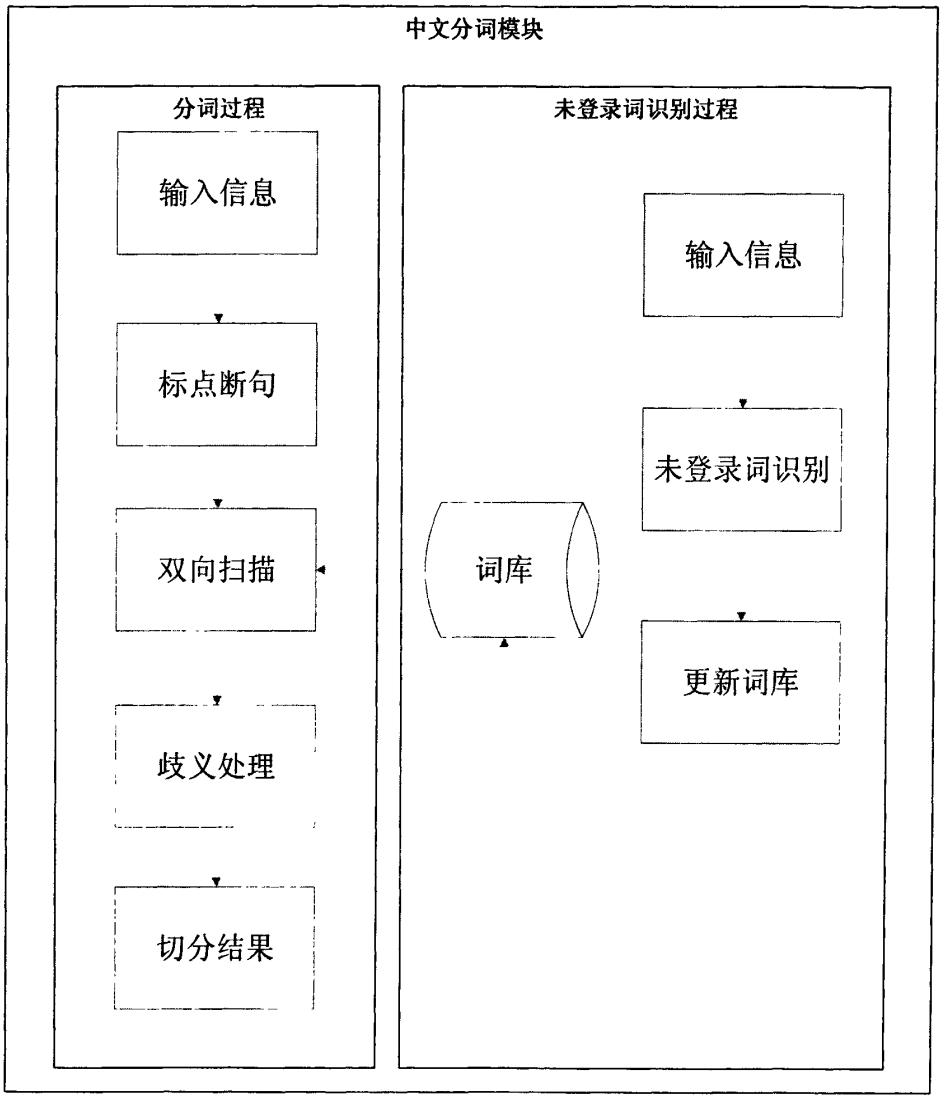


图 3.3 中文分词模块

Figure 3.3 Chinese word segment module

3.3 一种基于规则的中文分词算法的改进

本文改进了一种的基于规则的中文分词算法[1]，作为中文分词模块的分词算法。该算法以“跨度为 1 的前向最大匹配分词算法”为基础，结合词典，词频和词性信息，来判断分词结果的合理性。该分词算法能够在很大程度上消除歧义划分，实验结果表明，该分词算法准确率达 97%以上[1]，但这种算法由于其采用的词典结构形式以及“逐词切分”的切分策略，导致算法的分词速度较慢，本文通过重新设计词典结构，使用新的切分策略来提高算法的分词速度。

3.3.1 算法改进的理论依据

1. 从词典的结构上改进算法。

采用双 hash 词典结构替换原算法的词典结构，降低算法正向、反向最大匹配时匹配的次数。

原算法中一次最大正向匹配的描述是这样的：MAX 为词库中的最大词长，str 为待切分的中文字符串，point 为字符指针(初始时指向 str 的第 1 个字符)，算法首先调用基于字符串匹配的跨度为 1 的前向最大匹配算法，从 str 中 point 所指向的字符开始取长度分别为 2 到 MAX 的子串与词库中的词条进行匹配。若匹配成功，则该子串为词，将得到的词条信息加入到一个临时表中。

由于 MAX 是使用词库的所有词条最大长度，而原算法中，词库里的最大词条长度为 7，所以 MAX=7，利用 2.2.1 小节表 2.1 所记录的词长数据使用全概率公式计算出，汉语词汇的平均词长为 AWordLen=2.3，这使得在原算法进行最大匹配时，平均每一次多 MAX-AWordLen=4.7 次的字符串比较；而使用双 hash 词典的方式，每一次最大匹配是以某一个汉字“A”为首字的最长词条的长度（如，“澳门”的第一个字是“澳”，而以“澳”为第一个字的最长词条长度是 4，该词“是澳大利亚”）作为 MAX 的值，所以改进词典结构后，平均一次的最大匹配次数就是平均词长 AWorkLen=2.3 次，有效地降低了字符串匹配的次数，节省了分词的时间。

2.从切分策率上改进算法。

原算法的切分方式采用的是“调用基于字符串匹配的跨度为 1 的前向最大匹配算法”,实际上就是逐词切分的方式，切分出所有可能的词汇组合。这种切分方式的字符串比较次数过多影响分词的效率，因此本文采用比较次数较少的正反最大匹配的方式进行字符串的切分。

3.3.2 算法的词典设计

原算法中综合考虑词条的词性和词频信息，因此其词库文件的结构如表 3.2 所示。

表 3.2 词库文件结构

Table 3.2 Thesaurus file structure

字　段	字段类型	字段含义
Wordid	char（3）	词条唯一标识
Word	varchar（24）	词条
Mark	char（1）	词条词性（详见表3.3, 3.4）
Freq	int	词条词频
Load	bool	词条是否加入内存

表 3.3 Mask 字段值与词性的对应关系

Table 3.3 The relationship between Mark field and part of speech

Mark 代表的词性	Mark 代表的词性
0 无	8 动词
1 副词	9 动词，副词
2 形容词	10 动词，形容词
3 形容词，副词	11 动词，形容词，副词
4 名词	12 动词，名词
5 名词，副词	13 动词，名词，副词
6 名词，形容词	14 动词，名词，形容词
7 名词，形容，副词	15 动词，名词，形容词，副词

表 3.4 Mark 字段低 4 位与词性的对应关系

Table 3.4 The relationship between low harf part of Mark field and part of speech

0	0	0	0	1	1	1	1
-	-	-	-	动词	名词	形容词	副词

原算法中采用 Hash 词典作为其分词词典的结构，其内存中词库的 C++语言定义如下，hash_map<wstring,CDictItem，ltstr> dict;其中 wstring 是 C++宽字符串；CDictItem 是代表一个词条，包含表 xx 中的所有信息；hstr 是一个函数指针，定义了 CDictItem 进行大小比较时的小于函数， hash_map 会根据这个函数对其中的所有项进行排序、比较等操作。

由于原算法中的 hash 词典通过完整的词条字符串进行 hash 地址计算，确定词条在 hash 词典中的位置，因此算法的一次最大正向匹配过程如下段伪码所示，其中 MAX 是词典中最大词长。

```
for（ i=MAX; i>=2; i-- ）
{
```

```
        if( i 个字符串经 hash 计算存在于 hash 词典中 )
         {
              一次正向最大匹配分词结束;
              break;
         }
    }
```

本文的 3.2.1 小节算法改进的理论依据里分析过这种最大匹配方式的不足是“正向匹配次数过多”，而这是由于原算法中分词词典的结构造成的。

为改进元算中的分词词典结构，本小节对三种主流词典机制[7]进行了研究，它们分别是基于整词二分的分词词典机制、基于 TRIE 索引树的分词词典机制和基于逐词二分的分词词典机制。以下是这三种机制的详细介绍：

（1）基于整词二分的分词词典机制

该机制的词典结构分为词典正文、词索引表、首字散列表等三级。词典正文是以词为单位的有序表，词索引表是指向词典正文中每个词的指针表。通过首字散列表的哈希定位和词索引表很容易确定指定词在词典正文中的可能位置范围，进而在词典正文中通过整词二分进行定位。

（2）基于 TRIE 索引树的分词词典机制

TRIE 索引树是一种以树的多重链表形式表示的键树。基于 TRIE 索引树的分词词典机制由首字散列表和 TRIE 索引树结点两部分组成。TRIE 索引树的优点是在对被切分语句的一次扫描过程中，不需预知待查询词的长度，沿着树链逐字匹配即可；缺点是它的构造和维护比较复杂，而且都是单词树枝，浪费了一定的空间。

（3）基于逐字二分的分词词典机制

这种词典机制是前两种机制的一种改进方案。逐字二分与整词二分的词典结构完全一样，只是查询过程有所区别：逐字二分吸收了 TRIE 索引树的查询优势，即采用的是“逐字匹配”，而不是整词二分的“全词匹配”，这就一定程度地提高了匹配的效率。但由于采用的仍是整词二分的词典结构，使效率的提高受到很大的局限。

通过对 3 种分词词典的分析比较，本文设计了一种双 hash 的分词词典，词典结构如图 3.4 所示。

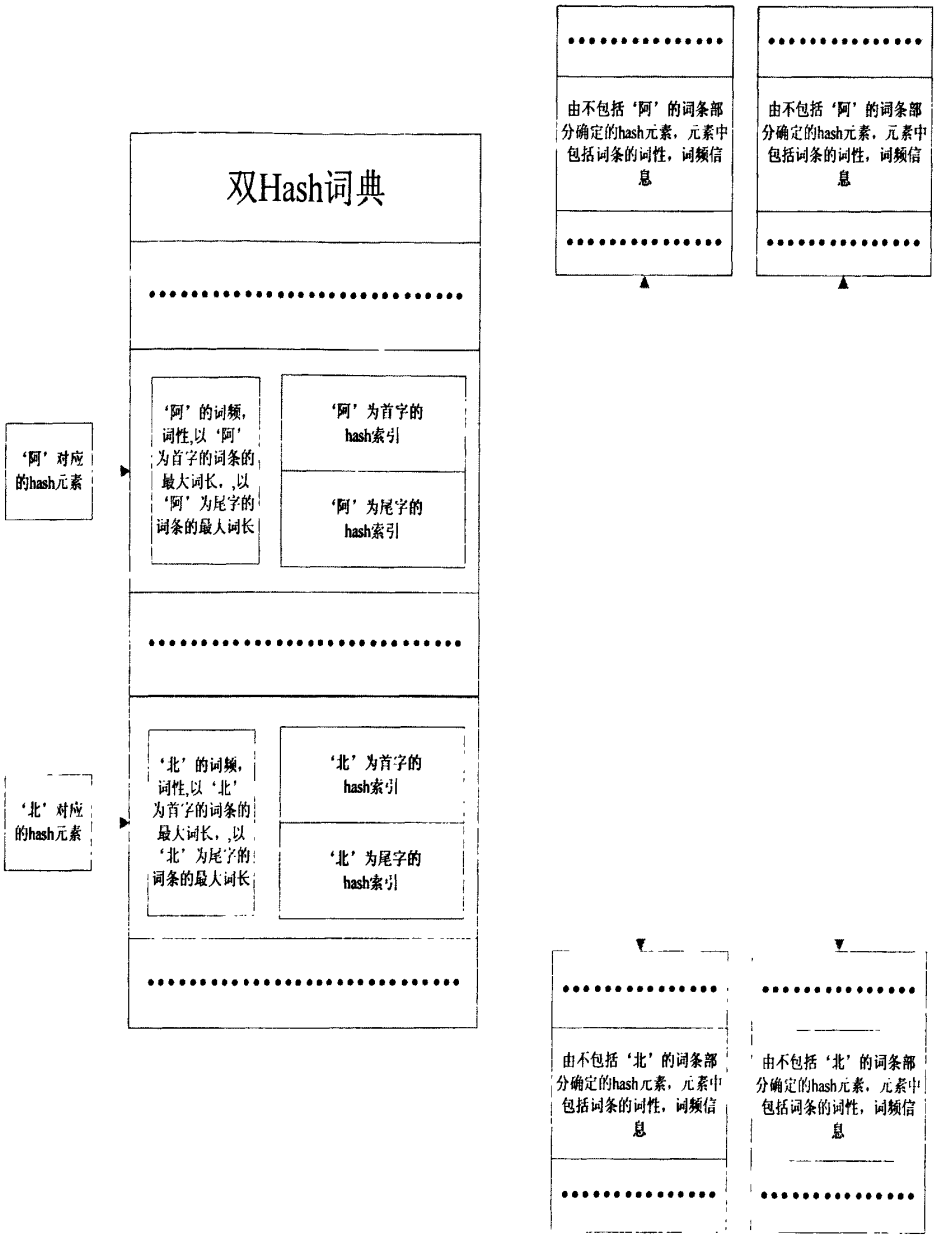


图 3.4 双 hash 词典结构

Figure 3.4 Two-hash dictionary structure

使用这种双 hash 词典结构后，一次正向最大匹配或反向最大匹配的描述如下段伪码所示。

读取一个汉字，从这个汉字对应的 hash 元素中得到这个以这个汉字为首字的词条的最大词长 WORD_MAX；

MAX=WORK_MAX

```
for ( i=MAX; i>=2; i-- )
{
    if( i 个字符串经 hash 计算存在于 hash 词典中 )
    {
        一次正向最大匹配分词结束;
        break;
     }
}
```

在 3.3.1 小节中计算过中文的平均词长为 2.3，假设原算法词典最大词长为 7，则改进词典后，平均一次正向最大匹配节省的比较次数是 4.7。

词典建立算法描述如下：

①读取原始词库文件中一个词条记录。进入步骤②、③。

②取词条的首字，求 Offset 值，Offset=(c1 — 0xb0)*94+(c2 — Oxa1), c1，

c2 为汉字的高低字节，由 Offset 得到首字在第一层 hash 结构里的元素 A，取词条的剩余部分，用 A 中“首字 hash”的“哈希函数”计算得到第二层 hash 结构里的元素 B，将词条的属性信息放入 SUB_ELEM 中，并比较词条长度 LEN 与 ELEMENT 中的首字最大词长 MAX(初值为 0)，如果 LEN>MAX,MAX=LEN。取词条的尾字，求 Offset 值，Offset=(c1 — 0xB0)*94+(c2 — OxA1), c1，

c2 为汉字的高低字节，由 Offse 得到尾字在第一层 hash 结构里的元素 A，取词条的剩余部分，用 A 中“尾字 hash”的“哈希函数”计算得到第二层 hash 结构里的元素 B，将词条的属性信息放入 B 中，并比较词条长度 LEN 与 A 中的尾字最大词长 MAX(初值为 0)，如果 LEN>MAX,MAX=LEN;进入步骤③

③判读原始词库中是否有剩余记录，有进入步骤①,没有进入步骤④。

④词典建立结束。

3.3.3 基于正反最大匹配的切分策略

原算法中使用 “跨度为 1 的强向最大匹配算法”来得到一个字符串所有可能的切分结果，然后在通过算法设定的规则来确定最合理的切分结果，这种切分方式平均一次切分需要进行的切分次数是 LEN，其中 LEN 为字符串长度。

本文中采用正反最大匹配的方式来得到切分结果集，一次正向最大匹配切分是过程是：从字符串头部开始最大匹配成功得到一个词条 Ni 后，向后移动 Mi 个字符，继续最大匹配直至匹配结束，其中 Mi 是词条 Ni 的长度；一次反向最大匹配的切分过程是：从字符串尾部开始最大匹配成功得到一个词条 Qi 后，向前移动

Pi 个字符，继续最大匹配直至匹配结束，其中 Pi 是词条 Qi 的长度。这种方式平均一次切分需要进行的切分次数是 2*（LEN/AVG_DIC），其中 LEN 为字符串长度、AVG_DIC 是词库中的平均词长。

比较两种切分方式可以发现，使用后者切分字符串，可以明显地减少字符串比较的次数，提高分词的速度。

3.3.4 改进后算法的描述

改进后算法的描述如下：

设 D 为词库， str 为待切分的中文字符串，point 为字符指针(初始时指向 str 的第 1 个字符)。首先进行正向最大匹配切分：读取 point 指针指向的汉字 Ci，确定以它为首字的词条最大词长 C_PRE_MAXi，接着调用基于字符串匹配的正向最大匹配算法，从 str 中 point 所指向的字符开始取长度分别为 2 到 C_PRE_MAXi 的子串与 D 中的词条进行匹配。若匹配成功，则该子串为词，将得到的词条信息加入到一个临时表中，并记录该词条的长度 Len 以及子串在 str 中的起始位置 Pos。当完成 C_PRE_MAXi-2+1 词匹配后，指针 point 后移 Len 个中文字符，然后重复上述过程继续匹配，直到 point 指向 str 的最后一个字符，正向最大匹配结束，以同样的方式进行反向最大匹配，将切分的词条加入临时表中，如果词条已经存在，则放弃加入，如果不存在按 Pos 值非递减的顺序加入临时表中。

完成上述过程后得到存放初步分词结果的临时表。然后依据如下中文文法规则，如表 3.5 所示，进行进一步的划分。

①在临时表中，按照从下往上的顺序循环取出一个词条，然后进入②；

②若当前词条的 Pos 减去前一词条的 Len 等于前一词条的 Pos，则保留该词条，然后进入①；若当前词条的 Pos 减去前一词条的 Len 小于前一词条的 Pos，则进入③；

③检查当前词条是否满足中文文法规则，若满足则进入④，否则对该词条进行拆分，然后进入①；

④若当前词条的 Len 不为 1 则进入 5，若 Len 为 1 则检查该词条是否为前一词条的后缀，如果是则删除该词条，不是则保留，然后进入①；

⑤若当前词条的 Pos 等于前一词条的 Pos，则删除其中 Freq 较小的词条，保留 Freq 较大的词条；若 Freq 也相同，则删除 Len 较小的词条，然后进入⑥；

⑥若当前词条的 Pos+Len 大于前一词条的 Pos 且小于前一词条 Pos+Len，则删除该词条然后进入 1，否则进入⑦；

⑦若当前词条的 Freq 大于前一词条的 Freq，则修改前一词条，去掉词条中的

重叠部分；若当前词条的 Freq 小于前一词条的 Freq，则修改当前词条，去掉词条中的重叠部分；若当前词条的 Freq 等于前一词条的 Freq，则保留其中 Len 较大的词条，删除 Len 较小的词条；然后进入⑧；

⑧本次循环结束，进入①。

表 3.5 中文文法规则

Table 4.5 Chinese grammar rules

中文文法最常见的合法组合
1.[形容词]+名词+[副词]+动词+[副词]+[形容词]+[名词]
中文文法中常见的非法组合
1.（前面没有动词）副词+名词
2.动词+形容词（后面没有名词）
3.名词+副词（后面没有动/形容词）
4.形容词+动词

根据上述算法，中文分词的流程如图 3.5 所示。

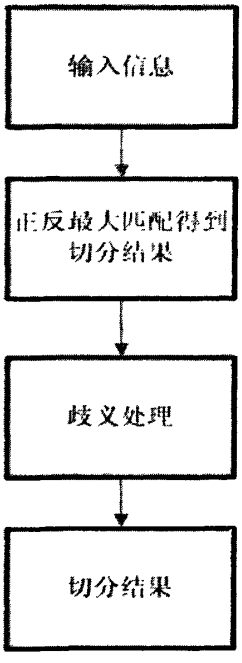


图 3.5 中文分词的流程

Figure 3.5 Chinese word segmentation process

对语句“这是非常情况”的切分过程如表 3.6 所示。

表 3.6 切分示例

Table 3.6 Segmentation example

正反向最大切分语句“这是非常情况”				
Word	Freq	Pos	Len	Mark
这		0	1	0
是非	489	1	2	4
是		1	1	0
非常	7812	2	2	1
常		3	1	0
情况	3134	4	2	4
使用规则处理后的切分结果				
Word	Freq	Pos	Len	Mark
这		0	1	0
是		1	1	0
非常	7812	2	2	1
情况	3134	4	2	4

3.3.5 实验结果

经测试算法改进后速度有明显提升，但分词精度略有下降这是由于采用正反最大匹配切分方式进行字符串切分造成的，因为有部分语句不论正向最大匹配还是反向最大匹配都无法给出正确的切分结果，如“学历史学好”正向最大匹配的结果是“学历/史学/好”、反向最大匹配的结果是“学/历史/学好”。

3.4 未登录词的识别

本文已分析过分词模块中未登录词识别功能的重要性，本小节将详细介绍未登录词的定义，进一步分析未登录词识别的意义，以及如何实现未登录词的识别。

3.4.1 未登录词的定义

未登录词是指包含在机械词典里的词汇，未登录词大致包含两大类：

（1）伴随社会的发展，新出现的通用词或专业术语；

（2）专有名词，如中国人名、外国译名、地名、机构名、等。

前一种未登录词的特点是，在理论上这种未登录词可以人工预先添加到词表中；而后一种未登录词的特点是，不可预期，也就是说无论词表的规模有多么庞

大，也无法包含所有的未登录词。

3.4.2 未登录词识别的意义

互联网上新生词汇产生的较快，数量也较多，而基于词典跟规则的分词算法无法正确切分包含未登录词的语句，且研究发现未登录词(OOV)造成的分词精度失落至少比分词歧义大5倍以上[12]。因此中文分词模块需要动态更新词库，增添新的词条，才可以提高分词的准确性。

3.4.3 未登录词识别的算法

本文的中文模块中采用N—gram切分算法[14]进行未登录词的识别，算法描述如下：

①切词

a.选择训练文本。

b.对训练文本进行预处理，将数据源中出现的标识符替换成空串，将空格、回车符及标点替换为“/”。

c.将预处理文本与停用词典进行模式匹配，文本中含有停用词的地方全部替换成切分标志“/”。

d.N元切分并统计词频。利用停用词、空格、回车换行符、标点等作为分词标志把长的文句拆分为句子片断或子串，利用N—gram算法对拆分后的片段进行切分，记录候选关键词及词频。

②过滤

在数据源够大的情况下，N元切分后的词条，词频越高则成词的可能性越大，而一些低频词往往是错切词，可直接排除，以减少后继各过滤算法的数据量，提高运行速度。设置词频阈值r为2，即排除掉词频为1的词条。对剩下的10% 左右的词条进行以下一系列过滤，包括前停后停词典过滤、相邻词比较、子父串

比较、抽词词典和过滤词典的过滤。

a.基于前停词典和后停词典的规则过滤。利用汉语词汇的构词特点，对切分得到的词条中，首字、尾字不满足构词规范的词条进行过滤，去除这些词条。

b.相邻词比较过滤。词长(设词长为N)相等，并且连续的N－1个字或字符相同，也即只有第一个词条的首(尾)字或字符与第二个词条的尾(首)字或字符不同，其余字或字符要全部相同，满足这两个条件的两词条互称为相邻词。如果两个相邻词的词频相同，则两词均被过滤掉；若其中一个词条的词频高于另一个，则保

留词频高的词条。

c. 子串、父串的比较过滤。子串是由父串中连续若干个字符组成的，表示在父串中的一个子集。子父串比较就是将一个子(父)串与其父(子)串比较过滤。算法中涉及两个参数，一是父串与子串词长间的相对值 P(即父串长度与子串长度之差)，第二个是父串与子串之间词频的相对值 q(即子串词频与父串词频之差)，这两个参数可随自己训练文本的长度与实验要求不同进行调整。对于某一个词条 I 的所有父串来说，当词条 I 的词频与其父串词频之差大于 q 时，若父串与词串 I 的词长之差小于或等于 P，则保留词条 I，而将其父串过滤掉；若词串 I 的词频减去其父串词频的差不大于 q 时，则保留父串，过滤掉词串 I。而对于词条 I 的所有子串来说，当其子串的词频减去词条 I 的词频之差小于或等于 q 时，则将其子串过滤掉，保留词条 I；否则，要再进一步判断词条 I 与其子串的长度之差是否小于或等于 P，若成立，则过滤掉词条 I，并保留其对应的父串。

③筛选

过滤词典是每轮实验数据经切分过滤后，再由人工判别，将过滤算法中未筛掉的错位切分、数字等无意义的词串另作标记，后又将其提取出来加入原有的过滤词典， 因此说本实验的过滤算法是一个不断积累、动态的过程。

3.4.4 实验结果

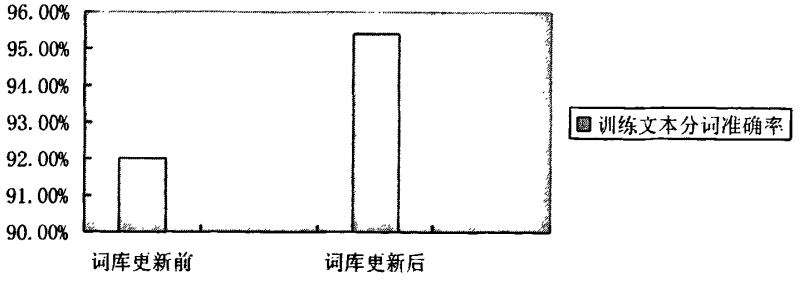


图 3.6 未登录词识别前后的分词准确率比较图

Figure 3.6 Unknown word recognition before and after the word accuracy comparison chart

实验证明引入未登录词识别算法对分词准确率的提高有很大的帮助。图 3.6 为中文分词模块中词库添加未登录词前与添加未登录词后，同一训练文本的分词准确率的对比。

3.5 本章小结

本章分析了中文分词模块在搜索引擎分析子系统与检索子系统中的作用；讨论了未登录词识别的意义，并采用一种支持词库自扩充的设计方式对模块进行设计，保证中文分词模块可以对网络中快速涌现的新词进行正确的切分；从词典设计、切分策略两个方面对一种分词算法进行了改进，提高了该算法的分词效率，并使用改进后的算法作为中文分词模块的分词算法；将一种 N—gram 切分算法应用到中文分词模块中进行未登录词的识别，提高了中文模块的未登录词的识别能力。

4 中文查询词纠错模块的实现

4.1 中文纠错模块的介绍

中文查询词纠错模块包含于搜索引擎的检索子系统中，如图 3.2 所示。在检索系统中，中文纠错模块的作用是：自动分析用户输入的检索词，检测出包含错误的查询词，给出纠错建议。

4.2 中文查询词纠错的需求

中文查询词纠错的需求源自搜索引擎系统中的日志分析。日志分析是搜索引擎中的重要工作，它是优化用户与搜索引擎系统交互的最主要途径。

搜索引擎系统中的日志中包含大量的信息，例如，ip 地址，查询时间，查询词等等，如图 4.1、4.2。研究发现搜索引擎的日志中存在大量包含错误的中文查询信息，这种查询信息降低了搜索引擎的查准率与查全率，因此中文纠错的技术被引入了搜索引擎系统中，来解决由于用户错误输入导致的无效查询问题。

```
127.0.0.1 - - [13/Jul/2008:18:23:11 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 129
127.0.0.1 - - [13/Jul/2008:18:23:11 +0800] "GET /favicon.ico HTTP/1.1" 404 294
127.0.0.1 - - [13/Jul/2008:18:23:27 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 623
127.0.0.1 - - [13/Jul/2008:18:23:51 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 152
127.0.0.1 - - [13/Jul/2008:18:23:52 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 152
127.0.0.1 - - [13/Jul/2008:18:26:03 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 258
127.0.0.1 - - [13/Jul/2008:18:26:26 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 361
127.0.0.1 - - [13/Jul/2008:18:26:27 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 361
127.0.0.1 - - [13/Jul/2008:18:27:23 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 2692
127.0.0.1 - - [13/Jul/2008:18:27:26 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 2692
127.0.0.1 - - [26/Jul/2008:12:29:53 +0800] "GET /Smarty-2.6.19 HTTP/1.1" 301 329
127.0.0.1 - - [26/Jul/2008:12:29:54 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 2692
127.0.0.1 - - [31/Jul/2008:23:31:28 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 2852
127.0.0.1 - - [31/Jul/2008:23:31:34 +0800] "GET /Smarty-2.6.19/ HTTP/1.1" 200 2852
```

图 4.1 系统访问日志文件

Figure 4.1 System access log files

```
218.66.59.84 www.xxxxx.com \xe7\xa6\x8f\xe5\xb7\x9e    [21 p/2008:06:30:05 +0800] "GET / HTTP/1.1" 200 17625 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; MyIE2; .NET CLR 2.0.50727; .NET CLR 1.1.4322)" "-""1530101c98fa90430a6b647149eed926" "bd_soso_image_fp--1221904924--21866598475 "

218.66.59.84 www.xxxxxxx.com \xe7\xa6\x8f\xe5\xb7\x9e    [21 p/2008:06:32:21 +0800] "GET area=com&cmd=uuu&frm=in_ui_train#train HTTP/1.1" 200 41779 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; MyIE2; .NET CLR 2.0.50727; .NET CLR 1.1.4322)" "-" "205d18a0f202acc06d1a91e0001058bc" "bd_soso_image_fp--1221904924--21866598475 "
```

图 4.2 Apache 服务器日志文件

Figure 4.2 Apache server log file

4.3 模块的设计

日志中的错误的中文查询信息有两个特点：

1. 错误信息中不包含错字信息，因为汉字在计算机系统中被包含在相应的字符集中，有着固定的编码，不论是采用音码输入汉字或是采用形码输入汉字，都是在字符集中选择汉字的过程，所以在计算机中不存在错字的情况。

2. 错误信息中最多的是同音别字词错误（如“周洁伦”），也有相当一部分的多字、漏字错误。这是拼音输入法日益流行导致的，因为拼音输入法学习成本低，思维干扰的最少，普及广，所以出现了大量的同音词错误，而多字漏字的情况多是出于不小心的输入造成的。

根据这种错误的特点，本文设计的中文查询词纠错模块使用基于拼音 hash 词典的纠错算法，和多字驱动词典+字符串匹配的纠错算法，利用前者处理包含同音别字的查询信息，利用后者处理包含多字漏字的查询信息。中文纠错模块的结构如图 4.3 所示。

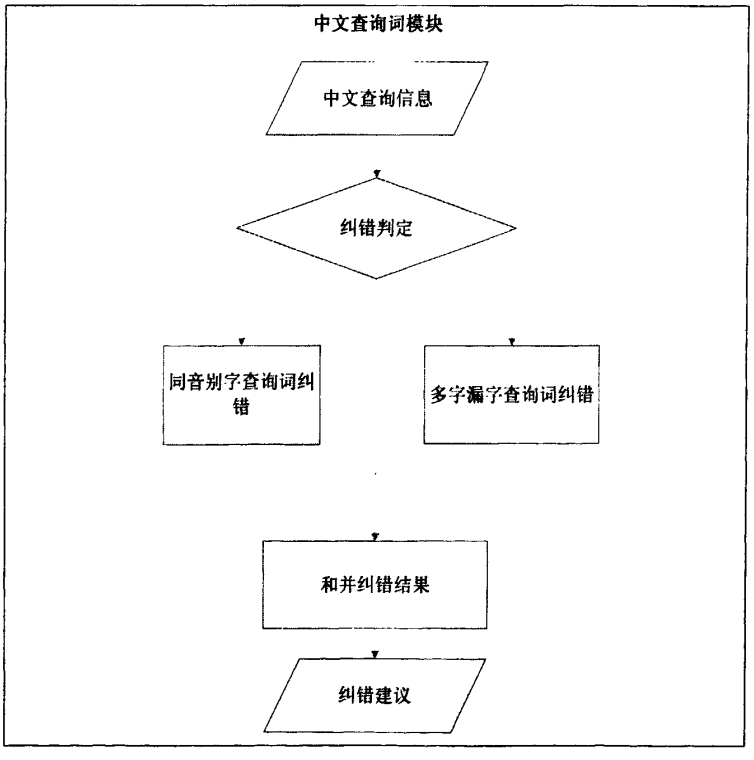


图 4.3 中文查询词纠错模块

Table 4.3 Chinese query term correction module

4.4 同音别字查询词纠错

4.4.1 纠错原理

同音别字查询词的纠错原理，如图 4.4：在判断词条有误后，将错误查询词转化为拼音，由拼音得到纠错词典中收录的同音词，计算它与同音词的相似性，得出纠错建议。

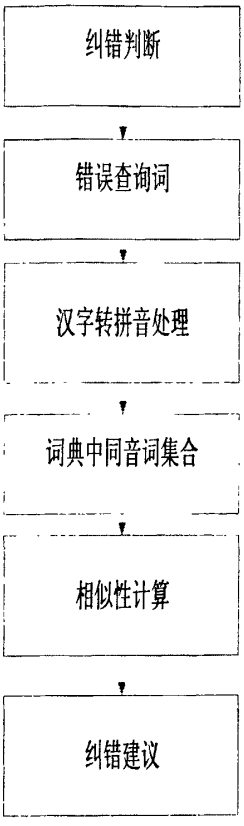


图 4.4 同音别字查询词的纠错原理

Figure 4.4 The Principle of the words with same pinyin and wrong character correction

4.4.2 同音别字词纠错算法

根据上述的纠错原理模型，纠错算法由 4 个部分组成，分别是原始词库建立、纠错词典建立、纠错判定和相似性计算排序。

（1）原始词库建立

原始词库中保存词条和词条对应的词频，其结构如表 4.1 所示。

表 4.1 原始词库文件结构

Table 4.1 Thesaurus of the original document structure

词条	词频
……	……
词条 N1	M1
词条 N1	M2
……	……

原始词库中的词条来源有两处，一是分词模块词库中的高频词汇，二是日志分析得到的检索高频词条和易错词条，这些词条就是系统认定的正确词，也就是经过纠错处理后的纠错建议。

例如，如果“北京交通大学”这个词条存在于原始词库中，那么检索输入“北京郊通大学”或“北京交通大雪”进行查询时，词条会被纠错为“北京交通大学”，但若原始词库中没有词条“北京交通大学”时，输入“北京郊通大学”或“北京交通大雪”进行查询就不会有纠错建议“北京交通大学”，因为系统没有收录这个词，根据原始词库建立的纠错词典中就不会包含这个词，这个词就不会成为纠错建议词条。

（2）创建纠错词典

纠错词典是以拼音为 key 值的链式 hash 结构，如图 4.5 所示，

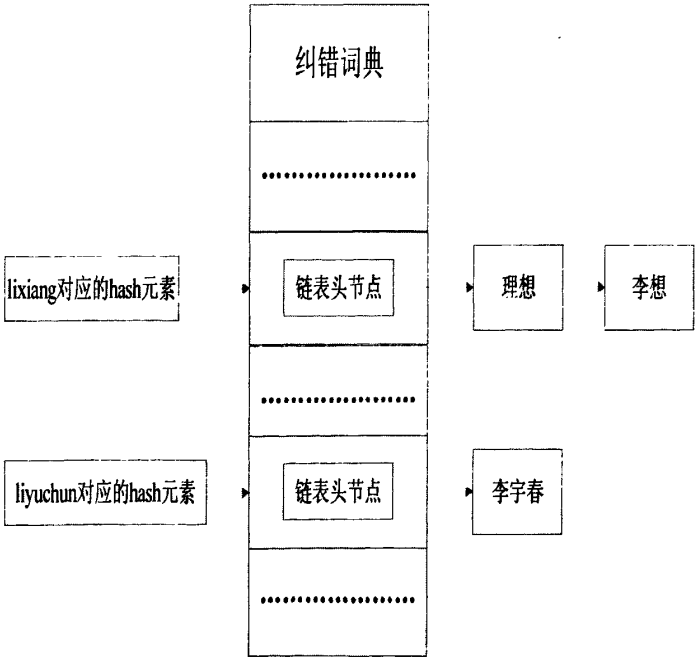


图 4.5 纠错词典结构图

Figure 4.5 Dictionary of error correction

纠错词典的建立过程是：

①从原始词库中读取一个词条，进入步骤 ②。

②将词条的每一个汉字转化为拼音，得到词条的拼音 x ；进入步骤 3。

③以得到 x 为自变量，经由 hash 函数 f（x）得到 x 对应的 hash 元素，将词条加入到相应的 hash 元素的链表中，进入步骤 ④。

④如果关键字源文件还有剩余词条，则进入步骤①，否则进入⑤

⑤词典建立结束。

（3）纠错判定

纠错判定就是根据检索输入来判定是否进行纠错处理。不同于文本校对系统中的查错，搜索引擎中纠错判定较为简单，因为对一个查询词条的纠错是毫秒级的计算，所以当下列条件中的一条或者多条被满足时，系统就会进行纠错处理：

①输入的关键字查询得到的结果集小于某一个阀值；且输入的关键字长度小于某一常数值（例如：20 个字节，10 个汉字）；

②对输入的中文信息分词时发现连续的单字字符串；

③检索输入词条中不包含汉字只包含拼音

④检索输入词条中同时包含拼音与汉字。

（4）相似性计算排序

对字串 Z1Z2…Zn（Zi 为错误字串的一个字）和词 C1C2…Cm（Ci 为纠错建议词中的一个字）；设函数 $same(Z_i, C_i)=\begin{cases}1 & \text{当} Zi=Ci \\ 0 & \text{当} Zi\neq Ci\end{cases}$，$\Sigma Same(Z_i, C_i)/M\times 100\%$ 表示其相似度，M 是纠错建议词的长度。

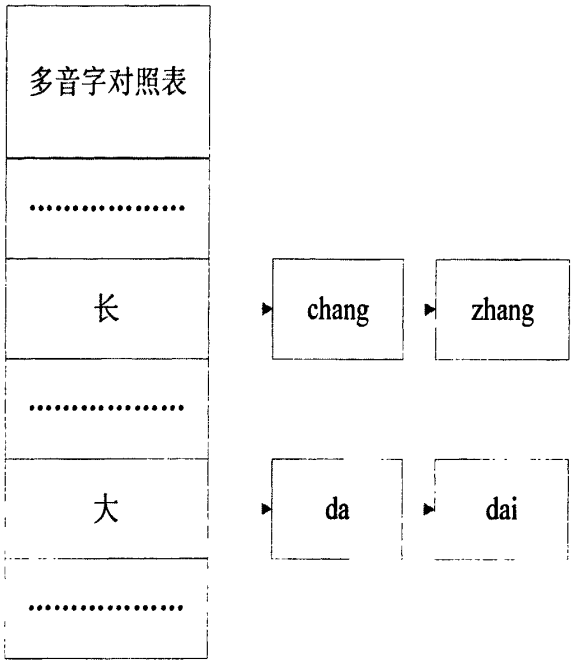
相似性计算排序就是将错误的词条转化为拼音后，由拼音得到一组同音词条，根据上述的函数、相似度计算公式，得到这一组词条与错误词条的相似度，在一组词条的相似度计算结束后，根据相似度对词条排序大到小排序，相似度相同的元素，词频高的排序在前，排序后取前 N（N 为系统设定参数，一般情况下，值取 3）条作为同音词纠错建议。

例如：输入“西按市”去检索信息，“西按市”满足纠错判定条件，将其转化为“xianshi”，由“xianshi”确定的一组其词条为“西安市”、“显示”、“现实”、“县市”，排序取前 N（N=3）项的结果为：西安市、县市、显示。

在上述过程中，将错误的词条转化为拼音时，要考虑多音字的情况。例如，词条“长渡”被转化为“zhangdu”而不是“changdu”时，不就不会得到纠错结果“长度”了。

为了处理多音字的问题，在词条转换为拼音时，需要根据多音字音字对照表 ，

如图 4.6 所示， 判断词条中的汉字是否为多音字，如果汉字在表中存在记录，则为多音字，如果词条包含多音字，则给出词条拼音的所有组合，对组合中每一个词条拼音都进行获取同音词条、相似计算、排序的处理



图表 4.6 多音字对照表

Figure 4.6 Multi-tone character table

根据上述算法，同音别字词的纠错流程如下：

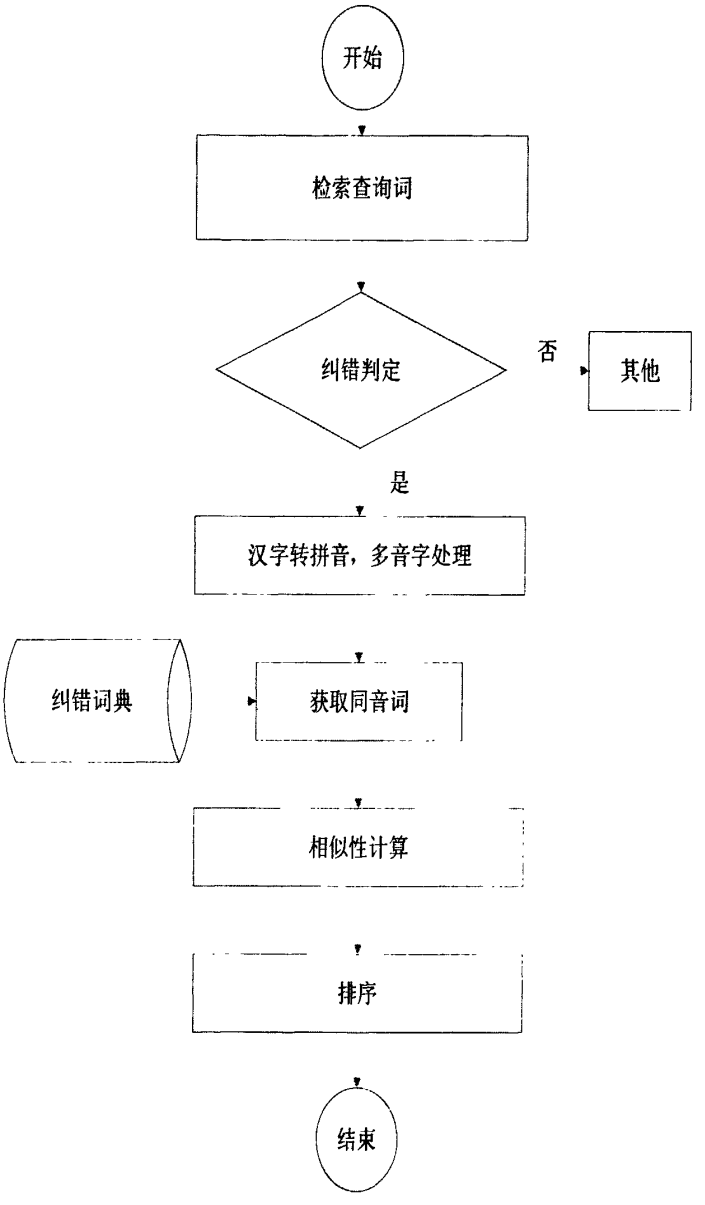


图 4.7 同音别字词纠错流程图

Figure 4.7The flow chart of the words with same pinyin and wrong character

4.5 漏字多字查询词纠错

4.5.1 纠错原理

对漏字，多字的中文词条纠错多采用使用双字驱动双向词典[4,5]，以错误词条

的首字，尾字为其发信息，通过中文字符串模糊匹配方法将发现的相似词条放入集合中，由纠错建议排序程序根据一定的规则对它们进行排序.作为纠错建议，如图 4.8 所示。

本文通过改进字驱动字典的结构，使用新的字符串模糊匹配算法，增强了纠错模块对漏、多字中文的纠错能力。

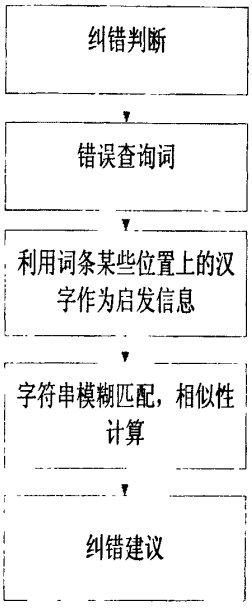


图 4.8 多字漏字词纠错原理图

Figure 4.8 Principle of the words with more or less characters correction

4.5.2 新的字符串模糊匹配算法

本文在设计新的字符串模糊匹配的算法前，对传统的字符串模糊匹配算法，也就是极大似然模糊匹配算法进行了研究[4,5]。极大似然字符串模糊匹配算法描述如下：

对字串 Z1Z2Z3.....Zn (Zi 为错误字串的一个字) 和词 C1C2C3.....Cm (Ci 为纠错建议词中的一个字)；设函数

$$\text{Same}(Z_i,\ C_i) = \begin{cases} 1 & \text{当} Zi = Ci \\ 0 & \text{当} Zi \neq Ci \end{cases}$$

若下列条件成立:

若当 m≥n 时,有:

$\Sigma Same(Z_i , C_i) \geqslant n-1$　　$(1 \leqslant i \leqslant n; Z_1=C_1)$　公式 4.1

或 $\Sigma Same(Zn-i+1 , Cm-i+1) \geqslant n-1$　　$(1 \leqslant i \leqslant n; Zn=Cm)$　公式 4.2

或 $\Sigma Same(Zi , Ci)+\Sigma Same(Zn-i+1 , Cm-i+1) \geqslant n-1$

$(1 \leqslant i \leqslant n; Z1=C1; Zn=Cm)$　公式 4.3

当 $m<n$ 时，有：

$\Sigma Same(Zi , Ci)+Same(Zn-i+1 , Cm-i+1)=m$　公式 4.4

$(2 \leqslant i \leqslant m-1; Z1=C1; Zn=Cm)$　公式 4.4

称字串 Z1Z2Z3.....Zn 与词 C1C2C3.....Cm 似然匹配。

记为 Match (Z1Z2Z3...Zn, C1C2C3...Cm)。

在原算法的基础上

定义相似度

s=similar/M× 100%,

其中，similar=max(ΣSame(Zi , Ci)$\geqslant$n-1 ($1 \leqslant i \leqslant n; Z1=C1$), Σ Same(Zn-i+1 ,Cm-i+1)$\geqslant$n-1 ($1 \leqslant i \leqslant n; Zn=Cm$))

M 是正确字符串的长度。

“北京交通大学”是正确词条，使用该算法对下面四个的错误词条进行相似性计算，结果如表 4.2 所示

表 4.2 纠错结果表

Table 4.2 Correcting the results

错误词条	计算得到的相同汉字个数	相似度
北京交通大	5	83%
京交通大学	5	83%
北北京交通大学的	1	16%
北京的交通大雪	3	33%

由表 4.2 可以发现，使用似然匹配算法对后两个字符串与“北京交通大学”进行相似性在计算时，得出的结果是不理想的。因此本文设计了一种新的字符串模糊匹配算法，描述如下：

字符串 Z：$Z_1Z_2Z_3.....Z_n$ 和 C：$C_1C_2C_3.....C_m$；N，M 分别为其字符串的长度，假设 Z 为错误串，C 为正确字符串同样设函数

$$same(Z_i, C_i) = \begin{cases} 1 & 当Zi=Ci \\ 0 & 当Zi \neq Ci \end{cases}$$

①i，j 分别为 Z，Q 字符串的下标初值设为 0，similar 表示两串相同汉字个数初值设为 0。

②如果 i<N, j<M 进入③，否则进入④ 。

③如果 same(Z_i, C_j)==1，那么 i++；j++；similar++；如果 same(Z_i, C_j)≠1，如果 N>M，i++；如果 N≤M, j++。进入步骤②。

④F=similar/M× 100%, i=N，j=M, similar=0 进入步骤⑤

⑤如果 i≥0, j≥0 进入⑥;否则进入⑦

⑥如果 same(Zi, Cj)==1，那么 i--；j--；similar++；如果 same(Zi, Cj)≠1，如果 N>M，i--；如果 N≤M, j--。进入步骤⑤

⑦ B=similar/M× 100%, 相似度=max（F，B）；分别为 Z，Q 字符串的下标计算的值作为相似度返回。

利用新算法对 Z:“北北京交通大学的”与 C:“北京交通大学”进行匹配的过程如表 4.3 所示，最终相似度的值为 100%

表 4.3“北北京交通大学的”与 “北京交通大学”匹配过程

Table 4.3 Matching process of “beibeijingjiaotongdaxue” and “beijingjiaotongdaxue”

i=0;j=0;	Z_i=‘北’; $C_{j=}$‘北’;	similar=1;
i=1;j=1;	Z_i=‘北’; $C_{j=}$‘京’;	similar=1;
i=2;j=1;	Z_i=‘京’; $C_{j=}$‘京’;	Similar=2;
i=3;j=2;	Z_i=‘交’; $C_{j=}$‘交’;	Similar=3;
i=4;j=3;	Z_i=‘通’; $C_{j=}$‘通’;	Similar=4;
i=5;j=4;	Z_i=‘大’; $C_{j=}$‘大’;	Similar=5;
i=6;j=5;	Z_i=‘学’; $C_{j=}$‘学’;	Similar=6;
得到 F 值，F=100%		
i=7;j=5	Z_i=‘的’; $C_{j=}$‘学’;	similar=0;
i=6;j=5	Z_i=‘学’; $C_{j=}$‘学’;	similar=1
i=5;j=4	Z_i=‘大’; $C_{j=}$‘大’;	similar=2
i=4;j=3	Z_i=‘通’; $C_{j=}$‘通’;	similar=3
i=3;j=2	Z_i=‘交’; $C_{j=}$‘交’;	similar=4
i=2;j=1	Z_i=‘京’; $C_{j=}$‘京’;	similar=5
i=1;j=0	Z_i=‘北’; $C_{j=}$‘北’;	similar=6
得到 B 值，B=100%		
得到相似度，值为 max（F，B）=100%		

下面四个“北京交通大学”的错误词条使用新算法进行相似性计算，结果如表 4.4 所示

表 4.4 新算法计算结果

Table 4.4 The new algorithm results

词条	计算得到的相同汉字个数	相似度
北京交通大	5	83%
京交通大学	5	83%
北北京交通大学的	6	100%
北京的交通大雪	5	83%

4.5.3 多字驱动字典

在包含漏字、多字的中文查询词中，有相当一部分的查询词首字，尾字均不正确，这导致只使用错误词条的首字，尾字为信息的方式将无法对这种查询词进行有效的纠错，因此本文采用多字驱动字典，利用更多地启发信息进行纠错。

多字驱动字典由若干条记录组成，每一条记录中包含一个汉字，不同记录中的包含的汉字不同，每个记录在字典中的位置由汉字的编码经 hash 函数计算后确定，每条记录还包括四个指针，分别指向以记录中汉字为首字，正向第二字，尾字，反向第二字的词条数组。多字驱动词典的结构如图 4.9 所示：

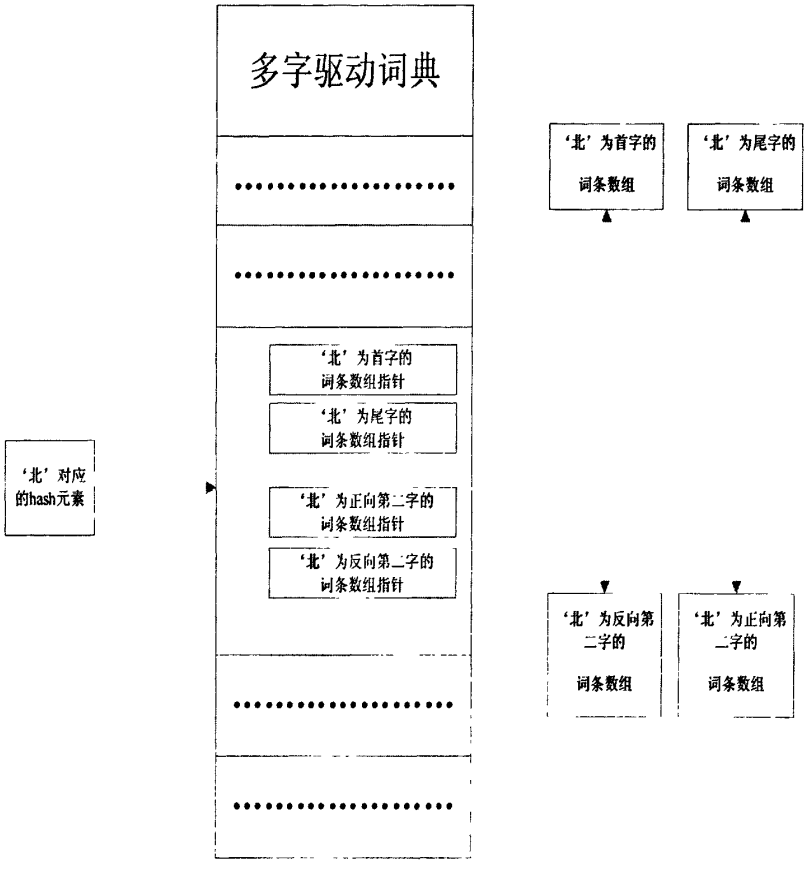


图 4.9 多字驱动词典

Figure 4.9 Multi-drive character dictionary

4.5.4 漏字多字词纠错算法

根据上述的 4.4.1 原理模型，算法由 4 个部分组成，分别是原始词库建立、纠错词典建立、纠错判定和模糊匹配相似性计算排序。

其中原始词库建立，纠错判定与 4.4.2 小节中同音别字词的算法中的原始词库建立，纠错判定相同。纠错词典建立，模糊匹配相似性计算排序两个部分的描述如下：

1．纠错词典建立：

①从原始词库中读取一个词条，进入步骤 ②。

②取得词条的首字，将词条放入以这个字为首字的词条数组里；

取得词条的尾字，将词条放入以这个字为尾字的词条数组里

取得词条的正向第二字，将词条放入以这个字为正向第二字的词条数组里

取得词条的反向第二字，将词条放入以这个字为反向第二字的词条数组里

进入步骤③

③如果关键字源文件还有剩余词条，则进入步骤①，否则进入⑤

④词典建立结束。

2. 模糊匹配相似性计算排序:

首先定义函数 corect(C, STR , n , m),

其中，C 是一个汉字的编码;

STR 是错误字符串;

n 是返回排序后相似的前 n 个词条;

m 是汉字 C 在词条中的位置;

m 可取值为 1、-1、2、-2 分别表示 C 词条首字，尾字，正向第二字，反向第二字。

函数 corect(C, STR , n , m)的功能是：使用 4.5.2 小节中的纠错算法，对以 C 为第 m 个汉字的每一词条与 STR 进行相似度计算，，并返回相似度最高的前 n 个词。

由于采用多字驱动词典，所以对错误词条 S1S2..Sn-2Sn-1Sn 假设为以下的 8 种情况,然后进行处理:

①查询词首字正确，错误信息包好在其他位置，

corect(S1 , S1S2..Sn-2Sn-1Sn , n , 1) 后得到集合 D1。

②查询词尾字正确，错误信息包好在其他位置，

corect(Sn , S1S2..Sn-2Sn-1Sn , n, -1) 后得到集合 D2。

③词条首字信息丢失,

corect(S1 , S1S2..Sn-2Sn-1Sn , n , 2) 后得到集合 D3。

④词条尾字信息丢失，

corect(Sn , S1S2..Sn-2Sn-1Sn , n, -2) 后得到集合 D4。

⑤查询词正向第二字正确，

corect(S2 , S2..Sn-2Sn-1Sn , n, 2) 后得到集合 D5。

⑥查询词反向第二字正确，

corect(Sn-1 , S0S1S2..Sn-2Sn-1 , n, -2) 后得到集合 D6。

⑦查询词首字为无效信息，

corect(S2 , S2..Sn-2Sn-1Sn , n, 1) 后得到集合 D7。

⑧查询词尾字为无效信息，

corect(Sn-1 , S0S1S2..Sn-2Sn-1Sn , n, -1) 后得到集合 D8。例如:

在计算得到 8 个纠错集合后，对所有的词条按相似度对词条排序有大到小排

序，对相似度相同的元素，按词频高的排序在前的方式进行处理，取前 W 个（W 设为 10）作为漏字多字词纠错建议。

根据算法描述，多字漏字词的纠错流程如下。

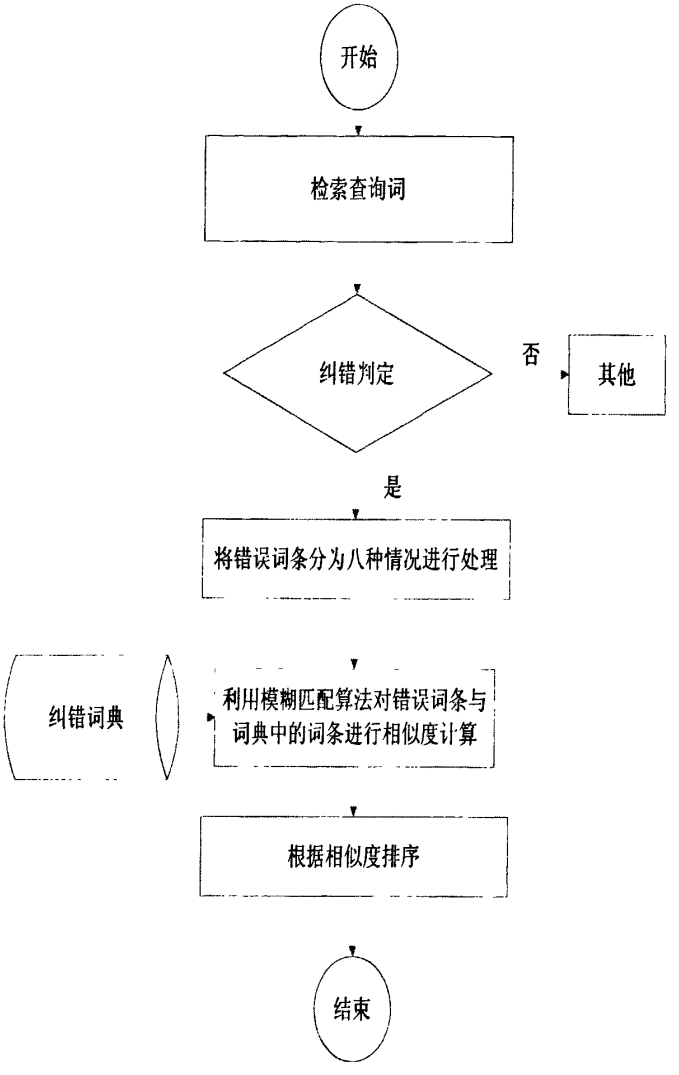


图 4.10 多字漏字词纠错流程

Figure 4.10 The flow chart of the words with more or less characters correction

4.6 纠错功能的扩展应用

4.6.1 中文查询词智能提示

中文查询词智能提示功能定义是：搜索引擎系统在用户输入完整的查询词前，

根据输入的部分汉字字符串 A，匹配词典中存储的查询词，返回包含字符串 A 的查询词候选集合。例如查询“北京交通大学”只输入“北京交通”时，搜索引擎的查询词智能提示如图 4.11 所示。

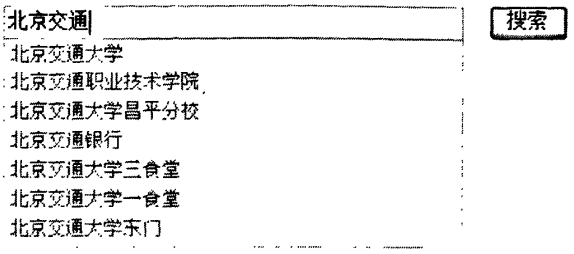


图 4.11 查询词智能提示

Figure 4.11 Intelligent query tips

实现查询词智能提示功能的一项重要技术是 Ajax。Ajax 本身由多个部分组成（如图 4.12 所示）,有 XHTML、CSS 、JavaScript、DOM、XML、XSTL 和 XMLHttpRequest，如图 4.12 所示。AJAX 使用 XHTML 和 CSS 标准化呈现数据，使用 DOM 实现动态显示和交互数据，使用 XML 和 XSTL 进行数据交换与处理，使用 XMLHttpRequest 对象进行异步数据读取，使用 JavaScript 绑定和处理所有数据，其工作原理如图 4.13 所示。

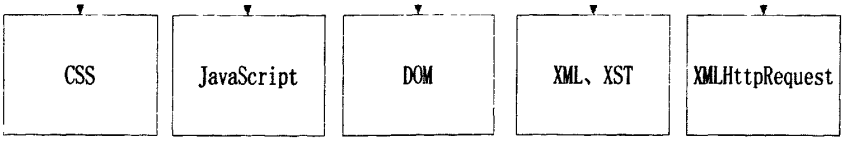


图 4.12 Ajax 技术组成

Figure 4.12 Ajax technical components

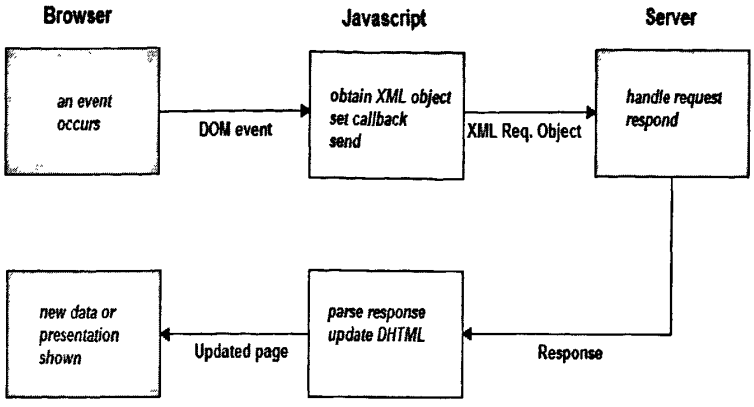


图 4.13 Ajax 技术原理图

Figure Ajax technical schematic

实现查询词智能提示功能的另一项重要技术是查询词提示词典，查询词提示词典以 hash 结构存储着查询词条，结构如图 4.14 所示，用于匹配输入的部分汉字字符串，得到查询词候选集合。

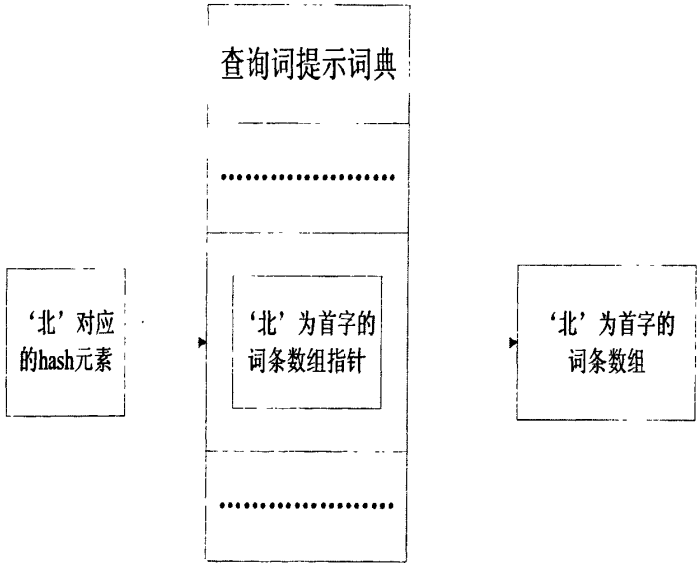


图 4.14 查询词纠错词典

Figure 4.14 Correcting word dictionary

查询词智能提示的过程是:系统的客户端将用户输入的部分查询信息 A 发送给服务器端，服务器端根据 A 的首字 a，查询词典得到所有以 a 开始的词条，从这些词条中进一步匹配，得到包含 A 且词频高于一定阀值的词条集合，并将集合返回给客户端，客户端使用 Ajax 技术完成无刷新的动态加载，实现了查询词的提示。

4.6.2 相关查询词推荐

相关查询词推荐功能就是搜索引擎系统在用户进行一次检索后，返回一组与用户检索相关的查询词。例如输入“大学”进行检索后，返回“大学招生”、“清华大学”、“北京大学”等相关查询词。

本文实现的相关查询词推荐过程是:对纠错处理得到的3个同音词纠错建议词条和10个漏字多字词纠错词条进行排序，取排序结果相似度较低的后10项（前3项作为最终的纠错建议信息）作为相关查询词。

这种方式实现起来较为简易，且能够满足用户对该功能的需求。

4.7 本章小结

本章分析了中文查询词纠错模块在搜索引擎检索子系统的作用，以及中文查询词纠错的需求。提出并实现了一种基于拼音hash词典的纠错算法，用于同音别字词的纠错；实现了一种新的字符串模糊匹配算法，并结合改进了的双字驱动词典对漏字多字词进行纠错；分析并实现了查询词智能提示功能，以及相关查询词推荐功能。

5 总结

在搜索引擎发展的过程中，研究者不断地将其它技术融入搜索引擎系统中，极大地丰富了搜索引擎的功能。本文从软件工程的角度分析了自然语言处理技术在搜索引擎系统中的作用，对自然语言处理技术中的中文分词技术与中文纠错技术进行了重点研究：对一种基于规则的中文分词算法进行了改进，降低了算法切分字符串的次数，提高了算法的分词效率；提出并实现了一种基于拼音 hash 词典的同音别字词纠错算法，用于同音别字词的纠错，提高了系统对同音别字查询词的纠错能力；改进了双字驱动词典的结构，并结合新的字符串模糊匹配算法对漏字多字查询词进行纠错，增强了系统对漏字多字查询词的纠错能力；将词典技术与 Ajax 技术相结合，实现了查询词智能提示功能；采用 N-gram 切分的新词识别算法，实现了分词词库的动态更新，提高了中文分词模块的分词准确度；模拟了相关查询词推荐功能的实现。

进一步的研究工作包括：在改进的中文分词算法中，确定分词结果的文法规则需要进一步的补充与修订，以便提高算法的灵活性；在中文查询词纠错模块中，原始词库需要保持动态更新，才能保证新生的网络词条在检索输入出错时，被系统纠正并返回纠错建议。

展望：自然语言处理技术在搜索引擎系统中的应用是成功的，其他信息检索系统（如文献检索系统、视频检索网站）也可借鉴这种经验，在原系统中加入中文分词、中文纠错等自然语言处理技术，来提高信息检索的质量。

参考文献

[1]傅士光. 基于主题的搜索引擎的研究与实现北京交通大学硕士论文. 2007.10-16
[2]梁斌. 走进搜索引擎. 北京. 电子工业出版社. 2007. 10. 第一版
[3]胡晓青. 网络搜索引擎中文纠错功能实例剖析. 图书情报工作网刊. 2008. 1
[4]张仰森, 曹元大, 徐波. 中文文本自动校错系统中知识库及其构造方法研究. 小型微型计算机系统. 2004. 12
[5]张仰森. 中文校对系统中纠错知识库的构造及纠错建议的产生算法. 中文信息学报. 2000. 第 15 卷, 15 期
[6]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
[7]张培颖, 李村舍. 一种中文分词词典新机制一四字哈希机制. 微型电脑应用. 2006. 第 22 卷第 10 期
[8]谭琼, 史忠植. 分词中的歧义处理. 计算机工程与应用. 2002. 11
[9]隋丽萍, 徐承韬, 李瑞芳. 一个中文全文检索系统的设计与实现. 科技资讯. 2007. 18
[10]祁正华. 基于无词库的中文分词方法的研究. 南京邮电学院硕士学位论文. 2005
[11]冯书晓, 徐新, 杨春梅. 国内中文分词技术研究新进展. 情报检索. 2002. 11
[12]黄吕宁, 赵海. 中文分词十年回顾. 中文信息学报. 第 21 卷, 第 3 期
[13]李群. 文本分词的自动校对. 渤海大学学报(自然科学版). 2006. 第 27 卷, 第 3 期
[14]曹艳,杜慧平,刘竞,侯汉清.基于词表和N—gram算法的新词识别实验.情报科学.2007.第25卷,11期
[15]George W Hart and Anastasios T Bouloutas. Correcting dependent errors in sequences generate by finite2state p rocesses[J]. IEEE Trisections on information theory, July 1993, 39(4) : 124921260.
[16]Joseph J Pollock. Automatic spelling correction in scientific and scho arly text [J]. Communication of the ACM , 1984, (4) : 3582368.
[17]New Generation Operational Suppo~Systems Architecture Overview[S]. TMF GB920，2000，2005.
[18]Richard Soley. The OMG Staff Strategy Group Model Driven Architec. ture Draft 3. 2[S]. 2000，2005.
[19]Enhanced Telecom Operations Map(eTOM)，the Business Process Framework for the Information and Communications Services Industry[S]. TMF GB921 v4. 0，2004，2005.
[20]Shared Information / Data Model-PhaseIII：Concepts，Principles，and Business Entities[S]. TMF GB922 v3. 0，2003，2005
[21]Sun Maosong,Huang Changning.Word Segmentation and Part of Speech Tagging for Unrestricted Chinese Texts: A Tutorial.ICCC'96,Singapore,June 4,1996

作者简历

姓名：李晓东 性别：男

出生日期：1983.07 籍贯：黑龙江省五常市

学历：硕士 毕业院校：北京交通大学

教育经历：

北京交通大学计算机学院	计算机科学与技术	2002/09–2006/07
北京交通大学软件学院	软件工程	2006/09–2009/01

实习经历：

雅普兰科技有限公司	2007/09-2007/12
航天恒星科技有限公司	2008/07-2008/12

完成的工作：

在雅普兰科技有限公司实习期间，参与清华同方总部业务辅助系统的开发工作，完成两个业务模块代码编写与测试的任务；

在航天恒星科技有限公司实习期间，参与风云 3 号卫星地面应用系统的开发工作，完成 5 个数据产品的开发与测试任务。