# 搜索引擎详细设计说明书

## 摘要

信息的飞速增长时搜索引擎成为人们查找信息的首选工具目前在中文搜索引擎领域国内的搜索引擎已经和国外的搜索引擎效果上相差不远之所以能形成这样的局面有一个重要的原因就在于中文和英文两种语言自身的书写方式不同。这其中对于计算机设计的技术就是中文分词技术。

本设计主要是实现一个中文分词器通过中文分词实现对中文句子的分析将其拆分成词语再将其应用在搜索引擎中使该搜索引擎实现对中文的搜索。系统中的分词算法采用基于分词词典的机械分词方法，这种方法是按照正向最大匹配的策略，将要分析中文句子与字典中的词条进行匹配而把中文句子拆成词

通过使用该分词器可以自动被中文句子准确的拆分成词，拆分速度快。运用在搜索引擎的分析器中可以对中文信息进行分析实现了对中文句子的搜索，提高了搜索引擎对中文搜索的功能和效率。并且本系统具有良好可扩展性和可移植性，可以移植到大部分使用lucene分析器的搜索引擎中

Abstract

The rapid growth of information search engines become the preferred tool for finding information Chinese search engine has been the effect on the search engine to be close Is able to develop this kind of situation there is one important reason for this is that both Chinese and English language to write their own different ways including the computer technology is the Chinese word segmentation technology

This design is the realization of a Chinese-term components Through the Chinese word segmentation and analysis of the Chinese sentence will be split into its term And its application in search engines so that the realization of the Chinese search engine search System of the algorithm using the word dictionary based on the mechanical-term approach which is in accordance with the largest positive match will be the strategy of the Chinese phrases and terms in the dictionary to match then split into the Chinese word sentence

Through the use of the term of the sub-components of the Chinese sentences can be automatically split into precise words split fast Use the search engine analyzer information to the Chinese and the Chinese phrase search improved search engines to the Chinese search function and efficiency And this system has good scalability and portability can be transferred to the use of most of the Lucene search engine analyzer

## 引言

本课题主要研究中文分词技术在搜索引擎中的应用与实现通过使用中文分词技术实现一个中文分词器把中文的汉字序列切分成有意义的词再将其应用到搜索引擎中以实现搜索引擎对中文搜索的支持提高搜索速度和效率

引擎发展至今已经有十几年的历史，而国内开始研究搜索引擎是在上世纪末本世纪初。目前在中文搜索引擎领域国内的搜索引擎和国外的搜索引擎效果上相差不远。之所以能形成这样的局面，有一个重要的原因就在于中文和英文两种语言自身的书写方式不同。由于中文是由联系文字组成，缺乏有效的间隔。虽有句段分隔，但在进行机器语言学习文本羽翼理解分析过程中都需要以词组为最小单位。因策实现中文分词相对西方语言来讲更加的复杂和困难。这其中对于计算机设计的技术就是中文分词技术。国内外的研究院校如北大清华中科院北京语言学院东北大学IBM研究院微软中国研究院等都一直关注并研究中文分词技术

研究这个中文分词组件的主要目的是服务于搜索引擎。通过使用词典将要分析的中文语句与词典中的词条进行匹配来提高粉刺速度通过使用正想最大匹配策略提高粉刺的精确度将该粉刺器应用到使用Lucene的搜索引擎中，不但可以实现对中文的搜索提高搜索功能，还可以对中文分词提高搜索精度，应用到其他方面可以使得计算机快速准确处理中文信息。

### 第一章 中文分词概述

随着搜索引擎的发展分词技术也运用而生在早起搜索引擎一直是用于对英文信息的搜索，先进搜索引擎已经运用在各种方面，也要实现对中文的搜索而对中文搜索就是使用中文分词技术

#### 1.1 中文分词简介

在英文的行文中单词之间是以空格作为自然分界符的，而中文只是字句和段可以通过明显的分界符来简单划界，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。但是在词这一层上中文与之英文要复杂的多苦难的多。这就要求在对中文网页进行分析之前先要将网页中的句子切个成一个个的词级序列。这就是中文分词。

#### 1.2 中文分词技术的应用

中文分词技术是中文信息处理的基础，在一切中文信息处理系统中