

**基于新闻的高级中文搜索引擎**

**Advanced Chinese Search Engine based on news**

目录

[第一章 绪论 4](#_Toc471216172)

[1.1 课题背景 4](#_Toc471216173)

[1.2 国内外研究现状分析 4](#_Toc471216174)

[第二章 主要算法详细介绍 5](#_Toc471216175)

[2.1 搜索引擎主要工作原理 5](#_Toc471216176)

[2.2 分词 5](#_Toc471216177)

[2.2.1 最大概率分词算法介绍 6](#_Toc471216178)

[2.3词性标注 7](#_Toc471216179)

[2.3.1 VOLSUNGA 算法介绍 7](#_Toc471216180)

[2.4检索 8](#_Toc471216181)

[2.4.1 倒排检索介绍 8](#_Toc471216182)

[2.5排序 9](#_Toc471216183)

[2.5.1 TF-IDF介绍 9](#_Toc471216184)

[2.6高级检索功能 9](#_Toc471216185)

[2.6.1 包含全部关键词的检索方式 10](#_Toc471216186)

[2.6.2 包含完整关键词的检索方式 10](#_Toc471216187)

[2.6.3 包含任意一个关键词检索 10](#_Toc471216188)

[2.6.4 不包含关键词检索 10](#_Toc471216189)

[2.6.5 站内搜索 10](#_Toc471216190)

[2.6.6 新闻类别检索 10](#_Toc471216191)

[第三章 系统结果展示及分析 10](#_Toc471216192)

[3.1 初级检索 10](#_Toc471216193)

[3.2高级检索 11](#_Toc471216194)

[3.2.1 按照新闻类别检索 11](#_Toc471216195)

[3.2.2 检索特定网站 12](#_Toc471216196)

[3.2.3 包含任一关键词检索 12](#_Toc471216197)

[3.2.4 包含完整关键词检索 13](#_Toc471216198)

[3.2.5 排除关键词检索 13](#_Toc471216199)

[第四章 系统开发平台与工具说明 14](#_Toc471216200)

[第五章 总结 14](#_Toc471216201)

[5.1 学习体会 14](#_Toc471216202)

[5.2 对课程意见及建议 14](#_Toc471216203)

[5.3 打分与分工 15](#_Toc471216204)

[参考文献 15](#_Toc471216205)

**摘要**

随着大数据时代来临,互联网信息量海量倍增, 浏览网络新闻也已经成为大家了解社会动态的一个重要渠道。搜索引擎为用户提供了一个找海量新闻信息的快速入口，然而，随着信息搜索技术的不断发展，用户对信息搜索的要求也在不断提高。通用搜索技术已经满足不了人们更加个性化的查询请求。因此，提供更高级的搜索功能将有助于为用户提供感兴趣的新闻服务，增强用户的体验性，增加用户的粘性。

本文旨在设计并实现一个基于新闻的高级中文搜索引擎。我们通过从网易新闻、腾讯新闻、新浪新闻、凤凰新闻、今日头条、搜狐新闻6个网站获得8个分类的新闻，分类包括军事、体育、科技、娱乐、社会、国际、国内、数码。然后使用最大概率分词对新闻进行分词，用VOLSUNGA进行词性标注，并为新闻建立倒排索引，最后使用TF-IDF对检索结果进行排序，为用户提供个性化的新闻搜索服务。

**关键词：新闻 高级中文搜索引擎 最大概率分词 倒排索引 TF-IDF VOLSUNGA**

Abstract

With the advent of the big data era, a huge amount of information on the Internet has redoubled, browsing network news has become an important channel for everyone to get the social dynamics. Search engine provides a quick entrance to help users find a large amount of news information, however, with the continuous development of information technology, the users' requirements of information search are also rising. Universal search technology has been unable to meet people's personalized query request. Therefore, providing a more advanced search function will help to provide users with services of finding news, to enhance the users’ experience, to increase users stickiness.

The purpose of this paper is to design and implement an Advanced Chinese search engine based on news. We get eight categories of news data including military, sports, science and technology, entertainment, social, international, domestic and digital from six websites，Netease News, Tencent News, Sina News, Phoenix News, today's headlines and Sohu News，and then use the maximum probability segmentation for the participle about news, and tag the part of speech by VOLSUNGA, and establish inverted index for news, and finally use the TF-IDF to sort the search results, to provide users with personalized news search service.

**Keyword**：news; advanced Chinese search engine; Maximum probability segmentation; Inverted index; TF-IDF; VOLSUNGA

# 第一章 绪论

## 1.1 课题背景

随着网络在世界范围内的不断发展和扩张，网络信息量也跟着大范围的增长，互联网已经成为一个包含有巨大信息量的空间。根据 2014 年中国互联网发展报告指出，中国网民规模截止 2013 底达到 6.18 亿，全年共计新增网民 5358 万人。通过通用搜索引擎得来的内容不够全面，搜索出来的内容，有许多是用户所不想要的内容，这样就造成了用户在使用搜索引擎时会有一个不好的体验，不能够达到用户对搜索结果的满意度。

包括新闻搜索在内的垂直搜索引擎主要专注于特定的网页内容,因此也叫做特色或主题搜索引擎。通常的垂直搜索引擎包括购物，汽车行业，法律信息，媒体信息，医药信息，学术型内容和旅游等。垂直搜索引擎定位于特定的用户搜索内容，它能满足那些需要查找精确和特定信息的用户的要求。垂直搜索引擎在所查找出的结果集相对通用搜索引擎要少，能够更加贴切用户的要求，同时能够对搜索的网页信息进行及时更新，所以说它的搜索结果更具体化和专业。垂直搜索引擎是现阶段搜索引擎的一个重要研究领域，在学术和工业界是一个热门的研究领域。

垂直搜索引擎的三个特点：（1）垂直搜索引擎对查询的信息有更高的精确度，这得益于它有限的搜索范围（2）充分利用领域知识，包括分类法和本体（3）针对用户的特定需求，给用户提供有价值的信息服务。 对比通用搜索引擎，垂直搜索引擎与其的异同点包括以下几方面：（1）不同的服务目的：垂直搜索引擎专注于某一个专门的知识领域，能够给用户提供定制的的搜索体验。（2）搜索方式不同：通用搜索引擎是对整个互联网内的内容进行整个爬取，而垂直搜

索引擎则按照预先设定的网络爬虫，爬取特定领域内的 web 页面，丢弃那些符合要求的页面，所以说这将节约大量的网络资源。

## 1.2 国内外研究现状分析

第一个搜索引擎系统Archie是由Montreal University的学生Alan Emtage、Peter Deutsch和Bill Heelan在1990年开发的。Archie索引互联网上的FTP网站的文件，严格上它还不算是真正的搜索引擎。随着搜索引擎技术的发展，出现了许多著名的搜索引擎：Alta Vista、Yahoo、Google等。随着搜索引擎的不断发展，搜索引擎也多元化发展，不止提供基本的文字搜索功能，例如Google搜索引擎还提供了天气预报、股价、地图、新闻等特殊功能。

垂直搜索专注于某一领域的信息，为用户提供更专、更精、更深的信息搜索服务。国内外也出现了许多垂直搜索引擎系统。 Medical Matrix是美国医学信息学会主办的、1994 年由堪萨斯大学创建的、目前最重要的医学专业的垂直搜索引擎。它是一个可免费进入的临床医学数据库，提供了关键词搜索和分类目录搜索服务。LIBClient-IRISWeb系统是由 North Carolina 大学计算机科学系和法学院联合开发的，提供了对网络上的法律信息进行了全文索引，法律领域的相关人员可以从搜索中获取法律专业的信息。

赛迪 IT 罗盘是国内最具代表性的垂直引擎，它是国内第一个中文 IT 垂直搜引擎，是真正意义上的首个中文领域的垂直搜索引擎。赛迪 IT 罗盘融入了个性化的设计，用户可以参与网站的评价，用户可以个性化定制搜索。Deepdo垂直搜索引擎是面向工作职位的搜索引擎，从 51job、chinahr 等招聘网站上收集数据，并为用户提供工作搜索的垂直搜索引擎。

在国外,D Freitag提出Machine learning for information extraction in

Information domains技术,将机器学习的理论融入信息提取技术中。GuPta等提出DOM一based content extraction of HTML documents技术,利用DOM树对HTML网页分析,进而得到有效信息。Christos Makris等提出Category ranking for personalized search算法,在排序算法中加入用户反馈信息与用户偏好信息。

# 第二章 主要算法详细介绍

## 2.1 搜索引擎主要工作原理

搜索引擎是一个信息检索系统，从互联网中采集各种网络资源信息，将采集来的信息进行组织和处理后存储在数据库中，然后为用户提供搜索功能，帮助用户从海量的互联网信息中快速地找到感兴趣的资源。

搜索引擎的工作原理大致可以分为：

1.搜集信息：搜索引擎的信息搜集基本都是自动的。搜索引擎利用称为[网络蜘蛛](https://zh.wikipedia.org/wiki/%E7%BD%91%E7%BB%9C%E8%9C%98%E8%9B%9B)的自动搜索机器人程序来连上每一个网页上的超链接。机器人程序根据网页链到其中的超链接，就像日常生活中所说的“一传十，十传百……”一样，从少数几个网页开始，连到数据库上所有到其他网页的链接。

2.整理信息：搜索引擎整理信息的过程称为“创建索引”。搜索引擎不仅要保存搜集起来的信息，还要将它们按照一定的规则进行编排。这样，搜索引擎不用重新翻查它所有保存的信息而迅速找到所要的资料。

3.接受查询：用户向搜索引擎发出查询，搜索引擎接受查询并向用户返回资料。搜索引擎每时每刻都要接到来自大量用户的几乎是同时发出的查询，它按照每个用户的要求检查自己的索引，在极短时间内找到用户需要的资料，并返回给用户.

其中，整理信息及接受查询的过程，大量应用了[文本信息检索](https://zh.wikipedia.org/wiki/%E6%96%87%E6%9C%AC%E4%BF%A1%E6%81%AF%E6%A3%80%E7%B4%A2)技术。通常信息检索技术主要包括两个部分：文档匹配和文档排序。

## 2.2 分词

最大概率分词是一种最基本的统计方法分词，它的基本思想是：一个待分割的字符串有多种分词结果，将其中概率最大的那个作为该字符串的分词结果。

### 2.2.1 最大概率分词算法介绍

根据贝叶斯定理和独立性假设：

注： 为句子分词结果的概率，*)*

最优路径中的第个词的累计概率等于它的左相邻词的累积概率乘以自身的概率。

具体步骤如下：(例子：结合成分子时)

1. 对“结合成分子时”，从左到右进行一遍扫描，得到全部候选词：“结”，“结合”，“合”，“合成”，“成”，“成分”，“分”，“分子”，“子”，“时”；

2. 对每个候选词，记录下它的概率值，并将累计概率赋初值为0；

|  |  |
| --- | --- |
| 结 | 0.0037 |
| 结合 | 0.0353 |
| 合 | 0.0049 |
| 合成 | 0.0006 |
| 成 | 0.0423 |
| 成分 | 0.0023 |
| 分 | 0.0312 |
| 分子 | 0.0038 |
| 子 | 0.001 |
| 时 | 0.1043 |

3. 顺次计算各个候选词的累计概率值，同时记录每个候选词的最佳左邻词：

P(结)=P(结)

P(结合)=P(结合)

P(合)=P(结)P(合)

P(合成)=P(结)P(合成)

P(成)=P(合)P(成)

P(成)=P(结合)P(成)

P(成分)=P(合)P(成分)

P(成分)=P(结合)P(成分)

P(分)=P(成)P(分)

P(分)=P(合成)P(分)

P(分子)=P(成)P(分子)

P(分子)=P(合成)P(分子)

P(子)=P(分)P(子)

P(子)=P(成分)P(子)

P(时)=P(子)P(时)

P(时)=P(分子)P(时)

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 候选词 | 累计概率 | 最佳左邻 |
| 0 | 结 | 0.0037 | 无 |
| 0 | 结合 | 0.0353 | 无 |
| 1 | 合 | 0.00001813 | 结 |
| 1 | 合成 | 0.00000222 | 结 |
| 2 | 成 | 0.00149319 | 结合 |
| 2 | 成分 | 0.00008119 | 结合 |
| 3 | 分 | 0.0000465875 | 成 |
| 3 | 分子 | 0.00000567412 | 成 |
| 4 | 子 | 0.00000008119 | 成分 |
| 5 | 时 | 0.000000591811 | 分子 |

4.如果当前词是字符串S的尾词，且累计概率P(wn)最大，则wn就是S的终点词；

5.从开始，按照从右到左的顺序，因此将每个词的最佳左邻词输出，即为S的分词结果。

“时 → 分子 → 成 → 结合”

即：

结合 / 成 / 分子 / 时

## 2.3词性标注

判定给定句子中每个词的语法范畴，确定其词性并加以标注的过程。

### 2.3.1 VOLSUNGA 算法介绍

VOLSUNGA是一种基于统计的词性标注算法。从左往右，对于当前考虑的词，只保留通过该词的每个词类的最佳路径，然后继续将这些路径与下个词的所有词类标记进行匹配，分别找出通往这个词的每个标记的最佳路径，后面的词依旧重复。公式如下：

一个句子由N个词组成，匹配词典，为每个词标上所有可能的词性，令是由个词组成的词串，是词串W对应的标注串，其中是的词性。

其中,

例子如下：

|  |  |  |  |
| --- | --- | --- | --- |
| W = 一 把 青菜 | | | |
|  | 一 | l | 1.0 |
|  | 把 | l | 0.3 |
|  |  | v | 0.6 |
|  | 青菜 | n | 1.0 |

V.S

比较和，发现，则继续

即:

一（l）把（l）青菜（n）

## 2.4检索

文档匹配主要是通过建立索引来实现的，索引方法包括： 简单的词-文档索引(term-document indexing)和倒排索引(inverted indexing)

### 2.4.1 倒排检索介绍

倒排索引源于实际应用中需要根据属性的值来查找记录。这种索引表中的每一项都包括一个属性值和具有该属性值的各记录的地址。由于不是由记录来确定属性值，而是由属性值来确定记录的位置，因而称为倒排索引。倒排索引对象是文档中的单词，用来存储这些单词在一个文档或者一组文档中的存储位置。

例子如下：

|  |  |
| --- | --- |
| 文档序号 | 分词后的句子 |
| 0 | 我0 喜欢1 电脑2 |
| 1 | 我0 说1 学习2 好3 |
| 2 | 我0 讨厌1 学习2 电脑3 |

**↓**

|  |  |
| --- | --- |
| 倒排索引 | |
| 我 | 0-0、1-0、2-0 |
| 喜欢 | 0-1 |
| 电脑 | 0-2 2-3 |
| 说 | 1-1 |
| 学习 | 1-2 2-2 |
| 好 | 1-3 |
| 讨厌 | 2-1 |

## 2.5排序

对匹配的文档设定一个分值，对分值排序，依据排序的先后显示检索结果

### 2.5.1 TF-IDF介绍

TF-IDF(term frequency–inverse document frequency)是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。

一个词i的TF-IDF权重值为:

⚫词频TF表示文档i中词汇j出现的频率，计算公式如下：

其中，表示词j在文档i中出现的次数，分母则是文档j中所有字词出现的次数之和。

⚫逆向文档频率IDF是一个词普遍重要性的度量，由下面的式子计算：

其中，N为所有文档的总数，是指包含词的文档的总数

## 2.6高级检索功能

高级检索（Advanced Search） 也称命令检索。功能包括字段检索，布尔逻辑检索等，利用这些功能搜过引擎可以检索某一Web站点被搜索引擎收录的页面数量，站点被链接的次数，比较各站点的声望高低。另外高级检索最多被应用到一些数据库中，实现精确查找数据的功能。

现有的搜索引擎中，大多都具备了高级检索的功能，这些功能使搜索引擎能够更加有效识别出用户搜查意图，使用户获得更多有效信息，过滤无用信息，并能够有效提高检索效率。

本系统的高级检索功能以及检索格式是参考百度的高级相关检索功能并实现的。

### 2.6.1 包含全部关键词的检索方式

这是最常用的检索方式，对检索语句进行分词获得关键词，检索出正文和标题包含了全部关键词的新闻。

### 2.6.2 包含完整关键词的检索方式

与包含全部关键词不同，包含完整关键词的检索方式检索的内容必须包含了完整搜索语句，关键词之间的顺序是按照检索语句严格相同。因此与包含关键词搜索的差别是，除了需要考虑关键词出现在哪些文章列表，还需要考虑在文章中出现的位置。

### 2.6.3 包含任意一个关键词检索

根据检索语句获得的关键词组，检索出包含任意一个或多个关键词的新闻。

### 2.6.4 不包含关键词检索

将原搜索所得的结果，筛选出包含了指定关键词的新闻，并从检索结果列表中去除。

### 2.6.5 站内搜索

仅在指定的新闻网站进行检索，提高检索效率。

### 2.6.6 新闻类别检索

仅在指定的新闻类别进行检索，提高检索效率和信息有效率。

# 第三章 系统结果展示及分析

## 3.1 初级检索

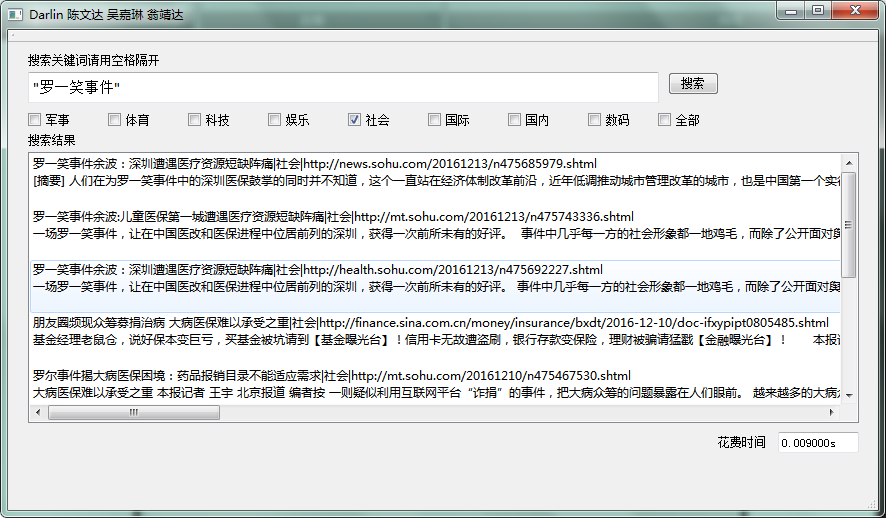
文档搜索主要是通过建立倒排索引(inverted indexing)来实现的，按空格切分用户输入的查询语句（query）得到查询关键词（term），通过倒排索引检索出包含全部关键词的文档，返回相关文档信息。



## 3.2高级检索

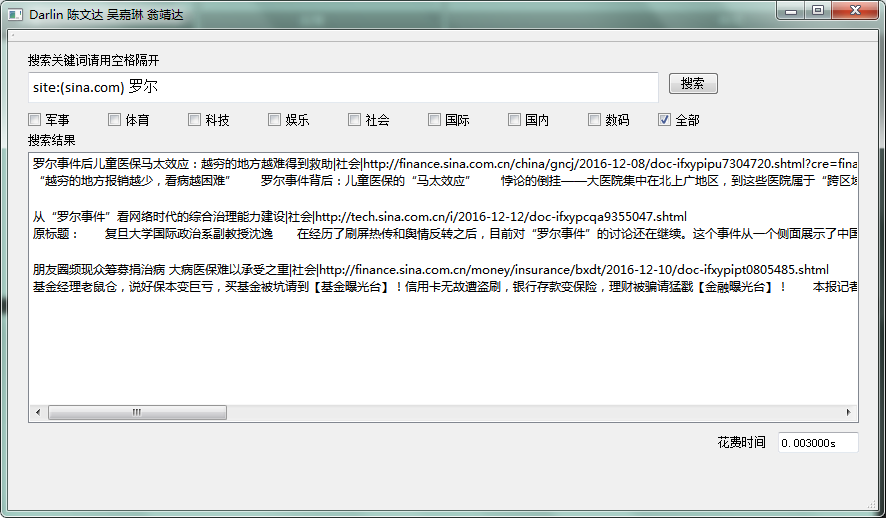
### 3.2.1 按照新闻类别检索

爬取新闻的同时，我们也获取了新闻在出处网站归属的类别信息，可以通过复选检索特定类别的新闻，包括军事, 体育, 科技, 娱乐, 社会, 国际, 国内, 数码8个类别。



### 3.2.2 检索特定网站

系统提供特定网站检索功能，检索格式为：site:(网站地址) query



### 3.2.3 包含任一关键词检索

系统提供包含任一关键词检索功能，检索格式为：(query)



### 3.2.4 包含完整关键词检索

系统提供包含完整关键词检索功能，检索格式为：”query”



### 3.2.5 排除关键词检索

系统提供包含排除关键词检索功能，检索格式为：query -(query)



# 第四章 系统开发平台与工具说明

注意：

1. 程序源码见项目文件。
2. 本程序基于python2.7 与 pyqt4库开发，并提供同步的exe版本。
3. 运行方式

（1）：在命令行中输入 python frame.py

（2）：直接双击frame.exe文件

# 第五章 总结

## 5.1 学习体会

这学期学习了自然语言处理这门课，首先先总结一下学习的内容，主要有分词、词性标注、句法分析、语义分析、信息检索、拼写检查、机器翻译。每上一个新的知识点，老师总会以具体例子一步一步解释该知识点的思路，课后让我们自己打代码或使用开源包去体验和吸收该知识点的内容，觉得这种方式很适合这门难度比较高的课程，让我们学生更容易吸收和记住知识点的内容，比单纯的书本理论教学要好很多。

上完这门课后，我们对于自然语言处理最大的感受是，分词是自然语言处理里面最基础且又最重要的一个环节，可是由于中文表达的多义性和歧义性，中文分词的准确率一直未能同英文分词一般高，分词的结果却又直接影响自然语言处理其他进阶处理的准确率，可以说要让机器充分理解中文自然语言是件较艰巨的任务。另外，我们还发现了几乎每个知识点都有基于规则和基于统计这两大种方法。基于规则的方法，是根据自定义的语法规则来进行语言的处理，而基于统计的方法则不需要太多的规则，只要采取一定的算法策略便可实现。在如今这个大数据时代，如果仅仅依赖基于规则的方法，规则的制定想必数量庞大，相比之下，基于统计的方法比基于规则的方法能节约更多的时间、人力和物力。虽然仅仅上了短短的十几节课，我们已经可以感受到自然语言处理这门课的广度之深，要想学好这门课，并不是这一个学期就可以掌握的，需要我们投入很多的精力，不仅要学好每个知识点的多个算法的思路，还要会通过实践具体实现，这样才能体会每个算法的优缺点，才能较全面地掌握知识点。

## 5.2 对课程意见及建议

由于课程时间紧凑，未能很好地学习到自动问答系统这个知识点。希望以后的课程可以充分展开介绍该知识点。

## 5.3 打分与分工

|  |  |  |  |
| --- | --- | --- | --- |
|  | 翁靖达 | 陈文达 | 吴嘉琳 |
| 翁靖达 |  | 96 | 90 |
| 陈文达 | 96 |  | 90 |
| 吴嘉琳 | 95 | 95 |  |

陈文达：数据库处理、搜索功能及高级搜索功能实现、文档编写

吴嘉琳：最大概率分词、文档编写

翁靖达：倒排索引目录构建、排名、数据爬取、框体实现

参考文献

[1] 邓丹君;基于Lucene的垂直搜索引擎关键技术研究[D];武汉理工大学；2011

[2] 林乐;特定网站新闻检索系统的设计与实现[D];华南理工大学;2013