2017-10-27

# 高效搜索技巧

郭延锐

合肥锐云智能科技有限公司

# 修订

| 版本 | 日期 | 作者 | 修订内容 |
| --- | --- | --- | --- |
| V1.0 | 20171027 | 郭延锐 | 第一次提交 |

Github 项目主页获取最新版本：

https://github.com/henry-yanrui/Sharpen-Cloud-Intelligent-Technology.git

# 一、信息检索定义

信息检索是一种从信息资源中获取需要的资源一种方法

检索方法分为两种：

1. 基于文本的检索；

2. 其它基于内容的检索（比如图片搜索、音频搜索）。信息检索是一种科学的信息搜索 的方法，可以用来检索文档、检索文档的内容、音频、视频等。

# 二、信息检索在个人学习中的应用

养成信息检索的思维方法，在科研和工作生活的过程中查找专业权威的资料指导学习研究，提高自学的能力。主要有在如下几个地方的应用：

## 2.1 数据库文献检索

国内有知网、维普、万方等数据库，国外有 Wiley InterScience、SCI 、IEEE 等数据库。一般的数据库搜索引擎提供了一个搜索对话框的入口，用户输入关键词进行查找，这是最基本的查找方法，针对于不同的数据库网站各自有自己高级搜索功能，可以更加细致全面的查找数据库的资料。

## 2.2 搜索引擎检索

目前市场上用的比较多的搜索引擎主要为百度、google、bing 等搜索引擎，一般搜索引擎有搜索入口，用户输入关键词进行查找，与数据库查找一样，搜索引擎也有搜索语法，可以按照文件种类搜索、信息内容的日期搜索、来源网站搜索等。与数据库搜索不同的是，多数搜索引擎可以搜索图片（按照文件种类搜索），用来查找相关内容的图片，用户也可以上传图片用来查找相似图片。

## 2.3 第三方网站检索

主要有两点原因：

a. 某些网站禁止搜索引擎爬取资料，因此只能在网站内部搜索。

b. 由于搜索引擎搜索的资料时相似内容过多，信息筛选有困难，在特定网站检索可以大大缩小查找范围，查找结果更加准确。

# 三、信息检索基本概念

## 3.1 基本概念

关键词（keyword）

关键词揭露了作者论证中核心的结构，描述了一篇文献中涉及到的核心部分，出现在文献的标题、摘要及文章中。对于一个问题来说，关键词描述了一个问题的本质，由关键词及其它修饰词组成一个问题。关键词一般是现成的词组。如“信息采样”、“压缩感知”“稀疏表示”、“观测矩阵”等。

单元词（uniterm）

单元词属于不可分割的词汇，可以独立存在的最小概念单元，也是在中文分词中能够被分类的最小单元，单元词也即单词，比如“信息采样”词组可以分为“信息”与“采样”两个单词。

## 3.2 中文分词方法

在书面表达的过程中是以字为最小组成单位，在计算领域的自然语言理解中是以词（单词、词组）为最小语言成分，为独立的活动的单元。中文分词就是将汉字串找到分割边界，便于计算机处理汉字串。理解中文分词的方法有助于建立高效的检索式。

目前中文分词主要有四种方法，简述如下：

1. 基于字符串匹配的分词方法

此方法也称机械分词方法、基于字典的分词方法，顾名思义，它就是将带分析的汉字串与计算数据库中机器词典进行匹配。

2. 基于语义的分词方法

传统的基于关键词字符匹配的信息检索中，参与匹配的只有外在的表现形式,而非它们所表达的全部概念，用户很难简单地用关键词或关键词串来真实地表达真正需要检索的内容。语义分析方法把信息检索从关键词匹配的层面提高到概念(语义)的层面,从概念意义上来认知和处理检索用户的请求。

3. 基于理解的分词方法

此方法也称人工智能分词方法，基本思想是在分词的同时进行句法、语义分析，利用句法、语义分析的来处理歧义现象。

4. 基于统计的分词方法

此方法也称无字典分词，词是字的组合，单词和词组中的字与字同时存在的概率较高，在实际使用的时候统计出相邻词之间的频度，频度越高，是单词或者词组的可能性

则越大。分词结合上下文识别生词、自动消除歧义的优点。

并非所有的检索引擎都采用了最先进的人工智能等分词方法，Lucene 是一个开源的搜索引擎框架，是一个高效的基于 Java 的全文检索库，能够为搜索引擎的爱好者提供一套可以构建搜索引擎的应用程序接口（API），在很多网站内部使用。但是 Lucene 的排序算法主要是根据信息检索的向量空间模型来计算的，文档和查询条件之间越接近，则该文档的权值就越高。为了达到最优化的搜索，**这个时候需要我们手动的将需要查找的汉字串有效的分割开，提高检索成功率。**

图 3.1 Lucene 站内搜索引擎

# 四、数据库信息检索

## 4.1 数据库信息检索的意义

官方数据库中的资料是人类智慧的结晶，科研的过程中查找数据库的意义不言而喻，在个人日常学习及工程实践方面也具有广泛的应用。在生活中可以查找疾病的成因；在学习中可以查找相关问题的详细研究；在工程实践中可以进行项目可行性研究及调研新技术新方案。

## 4.2 数据库中内容

以万方数据库分类为例，主要分为以下几种文献：

**期刊论文资源**

期刊论文是全文资源。收录自 1998 年以来国内出版的各类期刊 6 千余种，其中核心期刊 2500 余种，论文总数量达 1 千余万篇，每年约增加 200 万篇，每周两次更新。

**学位论文资源**

学位论文是全文资源。收录自 1980 年以来我国自然科学领域各高等院校、研究生院以及研究所的硕士、博士以及博士后论文共计 136 万余篇。其中 211 高校论文收录量占总量的 70%以上，论文总量达 110 余万篇，每年增加约 20 万篇。

**会议论文资源**

会议论文是全文资源。收录了由中国科技信息研究所提供的，1985 年至今世界主要学会和协会主办的会议论文，以一级以上学会和协会主办的高质量会议论文为主。每年涉及近 3000 个重要的学术会议，总计 97 万余篇，每年增加约 18 万篇，每月更新。

**专利资源**

专利是全文资源。收录了国内外的发明、实用新型及外观设计等专利 2400 余万项，其中中国专利 331 万余项，外国专利 2073 万余项。内容涉及自然科学各个学科领域，每年增加约 25 万条，每两周更新一次。

**成果资源**

成果是题录资源。主要收录了国内的科技成果及国家级科技计划项目。总计约 50 余 万项，内容涉及自然科学的各个学科领域，每月更新。

**法规资源**

法规是全文资源。收录自 1949 年建国以来全国各种法律法规 28 万余条。内容不但包括国家法律法规、行政法规、地方法规，还包括国际条约及惯例、司法解释、案例分析等。

**标准资源**

标准是题录资源。综合了由国家技术监督局、建设部情报所、建材研究院等单位提供的相关行业的各类标准题录。包括中国行业标准、中国国家标准、国际标准化组织标准、国际电工委员会标准、美国国家标准学会标准、美国材料试验协会标准、美国电气及电子工程师学会标准、美国保险商实验室标准、美国机械工程师协会标准、英国标准化学会标准、德国标准化学会标准、法国标准化学会标准、日本工业标准调查会标准等 26 万多条记录，每月更新。

**企业信息**

企业信息是题录资源。始建于 1988 年，是国内最早商业化运作的企业信息库，收录了国内外各行业近 20 万家主要生产企业及大中型商贸公司的详细信息及科技研发信息。每月更新。

**西文期刊论文**

西文期刊论文是全文资源。收录了 1995 年以来世界各国出版的 12634 种重要学术期刊，部分文献有少量回溯。每年增加论文约百万余篇，每月更新。

**西文会议论文**

西文会议论文是全文资源。收录了 1985 年以来世界各主要学协会、出版机构出版的学术会议论文，部分文献有少量回溯。每年增加论文约 20 余万篇，每月更新。

**科技动态**

收录国内外科研立项动态、科技成果动态、重要科技期刊征文动态等科技动态信息，每天更新。

### 4.3 数据库专业检索


图 4.1 中国知网专业检索

# 五、搜索引擎检索

## 5.1 google 搜索语法

参考下面的百度

## 5.2 baidu 搜索语法

高级搜索提供了一个可视化的搜索对话框，用户不必记忆搜索语法，可以便捷的使用。下面介绍的是在普通对话框中使用的搜索语法。

### 5.2.1 包含关键词 +

默认搜索语法为包含全部关键词，所有包含在汉字串中的标点符号（前后均无空格）以及空格均被认为是连词符号，在搜索时候被忽略。

需要注意的是以下结果的区别：

**锐云 +智能 +科技**

搜索结果包含锐云（锐云、锐、云）和智能（智能、智、能）和科技（科技、科、技）


## 锐云 智能 科技

同上，词与词之间空格分割会自动添加“+”。

## 锐云 +智能科技

所查询结果含有锐云（锐云、锐、云）和智能科技（智能科技、智能、科技）


## 锐云智能科技

所查询结果含有锐云智能科技（锐、云、智能、智能科技、科技、云智能、云智能科等）


## 5.2.2 排除关键词 –

### cpa 资格证

返回大量推广信息，用户难以分辨。


### cpa 资格证 -推广

不含有任何广告信息

### 5.2.3 整体关键词 “”

在中文全角的双引号之间的内容不可拆分，在检索时作为一个整体使用。

**“锐云智能科技”**


### 5.2.4 或关键词 |

对关键字做或逻辑操作，可与其它语法并列使用。

**锐云 | 智能 | 科技**

查询结果含有锐云或者智能或者科技，并列关系


**(锐云 | 智能 | 科技) 有限公司**

所查询结果含有锐云或者智能或者科技和有限公司，注意这里的“()”为西文半角括号用来组合逻辑使用。


## 5.2.5 文档格式 filetype

支持 pdf、doc、xls、ppt、rtf、all（这里的 all 值得是包含 pdf、doc、xls、ppt、rtf 格式，不支持新版本 office 的格式，比如 docx）

**ADI 系统方案精选 filetype:pdf**

检索结果第二个为 pdf 文档，点击可下载


## 5.2.6 标题 intitle

在标题中含有关键词

**intitle:震惊!**

返回的结果中标题中含有待查询关键词，如下图所示。


## 5.2.7 URL inurl

在 URL 中含有关键词

**inurl:震惊!**

返回的结果中是 url 含有关键词，需要注意的是，url 都是对关键词编码之后的结果，对 url 解码可看到 url 中含有的关键词


## 5.2.8 站内 site

在站内含有关键词

**"搜索语法" insite:zhihu.com -百度**

这里使用“-百度”主要是避免百度自己的产品优先排列对搜索结果影响


# 六、 总结

## 6.1 避免使用非陈述句或句子

在 3.2 节说到中文分词的方法，虽然说现在有各种先进的人工智能的分词方法，但是用户将一个问题提取出关键词再检索，还是会大大提高准确度。


图 6.1 疑问句式搜索


图 6.2 关键词搜索

## 6.2 外文资料优先检索 google、bing 等搜索引擎

百度是全球最大的中文搜索引擎，对于外文的支持并不是很友好，例如，检索 c primer plus 这本书，百度与 bing 的结果如下：


图 6.3 c primer plus filetype:pdf

在百度中检索文档，优先推荐百度文库的文档，文档种类繁多，有时候需要的结果并不在前几个搜索结果中。


图 6.4 c primer plus filetype:pdf –百度文库

避免百度文库的干扰，将百度文库关键词排除，但是没有任何关键词相关的资料

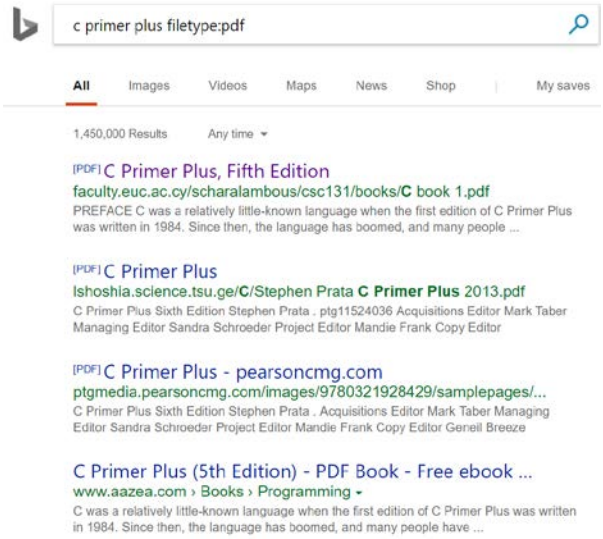


图 6.5 在 bing 中检索 c primer plus filetype:pdf

在 bing 搜索引擎中检索，第一个结果即为需要查询的结果，效率大大提高，这并非个例，绝大多数的外文资料更容易在国际搜索引擎上检索成功。

## 6.3 使用间接检索

搜索引擎的内容较多，信息冗余严重，提取有时候会存在一些困难，这个时候需要

使用站内检索的功能，一种方法是使用 " site" 语法，另一种方法是搜索出目标网站，直接进入网站内部搜索。比如需要了解常用三极管信号和价格，直接在搜索对话框检索获取到非常多的冗余信息，如图 6.6 所示。这时可以先检索 "电子元件采购平台 -推广 -广告" 获取到目前市场上比较正规的电子元件采购平台，从平台中检索信息。如图 6.7 及 6.8 所示。


图 6.6 "NPN 三极管采购"


图 6.7 "电子元件采购平台 -推广 -广告"

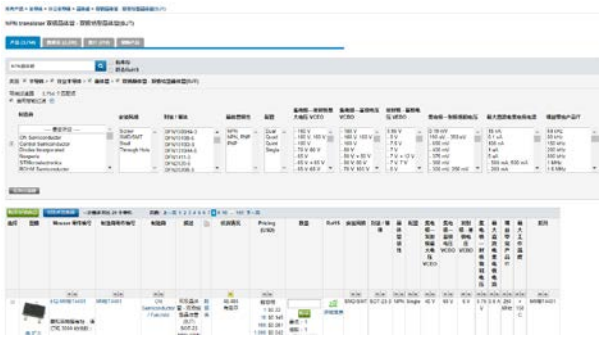

图 6.8 mouser 电子采购平台

## 6.4 建立私有数据库

对于搜索引擎来说，准确的关键词可以获取到准确的结果，对于经常使用搜索引擎的同学来说需要建立一个属于自己的关键词库。另外在平时的学习过程中记录下一些搜索引擎的技巧，必定事半功倍。

# 附录 常用信息检索网站

1. google scholar
2. 百度学术
3. google patents
4. Wikipedia
5. Github
6. 中西文科技文献服务平台

# 参考文献

[1] Uberti H Z, Scruggs T E, Mastropieri M A. Keywords Make the Difference![J]. Teaching Exceptional Children, 2003.
[2] 龙树全, 赵正文, 唐华. 中文分词算法概述[J]. 电脑知识与技术, 2009, 5(10):2605-2607.
[3] 张启宇, 朱玲, 张雅萍. 中文分词算法研究综述[J]. 情报探索, 2008(11):53-56.
[4] 王兆宇, 乐嘉锦, WangZhaoyu,等. 基于 Lucene 的个性化站内搜索引擎的研究[J]. 计算机应用与软件, 2011, 28(12):188-190.