



苏州大学

# 中文信息检索

苏州大学计算机科学与技术学院

2017年12月24日 12时10分

苏州大学：中文信息处理



# 内容

- ❖ 中文信息检索基础
- ❖ 中文信息检索架构
- ❖ 中文信息检索模型
- ❖ 中文信息检索系统
- ❖ Web信息检索概述
- ❖ 搜索引擎
- ❖ 搜索引擎分类



# 什么是信息检索?

## ❖ Information Retrieval

### ❖ 广义的信息检索

利用感知来获取信息

❖ 看

❖ 听

❖ 摸

❖ 闻

人生活中重要环节

知识的来源





# 什么是信息检索？(续)

## ❖ 狭义的信息检索

✎ 特指计算机信息检索

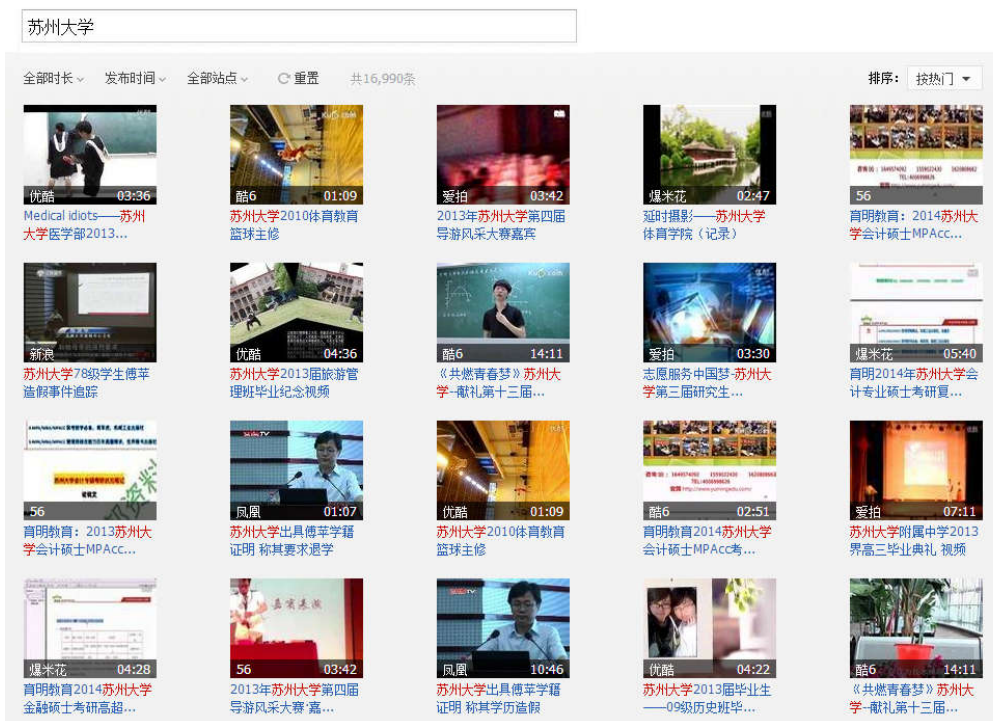
❖ 文本

❖ 图像

❖ 音频

❖ 视频

❖ .....





# 文本信息检索

❖ 对象:

❖ 起源:

❖ 研究, 中找! Inform

❖ 文档

☞ 网页

☞ 邮件

☞ 专业



❑ [发明] 一种中文事件的抽取方法及系统 - 201210182651.8 审中

申请人: 苏州大学 - 申请日: 2012-06-05 - 主分类号: G06F17/27(2006.01)I

发明人: 李培峰; 朱巧明; 周国栋; 朱晓旭...

摘要: 本发明提供一种中文事件抽取方法和系统, 该方法包括: 将待抽取事件的本文依次进行分句、分词、实体识别、句法和依存关系分析; 根据词的内部结构, 将符合抽取条件的词标记为候选触发词; 根据概率、词性和词内部结构将符合过滤条件的触发词过滤掉; 利用最大熵识...

☞ 阅读 - 下载 - 法律状态 - 信息查询 - 同类专利



❑ [发明] 一种微博信息的压缩编码和解码的方法及装置 - 201110298118.3 有权

申请人: 苏州大学 - 申请日: 2011-09-29 - 主分类号: G06F17/22(2006.01)I

发明人: 李培峰; 朱巧明; 刁红军; 朱晓旭; 张玉华...

摘要: 本发明实施例公开了微博信息的压缩编码和解码的方法及装置, 本发明实施例通过设置中文词典、中文符号表和英文字典对使用UCS-2编码的微博进行压缩再编码, 其中压缩编码的方法概括为: 识别UCS-2编码文本中各UCS-2编码的字符类型, 根据识别的字符...

☞ 阅读 - 下载 - 法律状态 - 信息查询 - 同类专利



❑ [发明] 事件信息融合方法和系统 - 201110269307.8 审中

申请人: 苏州大学 - 申请日: 2011-09-13 - 主分类号: G06F17/30(2006.01)I

发明人: 李培峰; 朱巧明; 周国栋; 王红玲; 朱晓旭...

摘要: 本发明公开了一种事件信息融合方法和系统, 用于对事件信息进行抽取、补全、事件聚类 and 融合, 形成事件信息完整度高的完备事件。本发明实施例方法包括: 生成包括多个事件的初选事件集合; 比较初选事件集中的事件与事件抽取模式的相似度, 形成候选事件集合; 甄别...

☞ 阅读 - 下载 - 法律状态 - 信息查询 - 同类专利



❑ [发明] 文本信息抽取方法和系统 - 201110273322.X 有权

申请人: 苏州大学 - 申请日: 2011-09-15 - 主分类号: G06F17/30(2006.01)I

发明人: 李培峰; 朱巧明; 孔芳; 周国栋; 钱龙华...

摘要: 本发明实施例公开了一种文本信息抽取方法, 实现从文本中抽取某种现象或某个事件产生的原因信息, 其方法根据原因种子对从互联网中采集的语句进行分析, 生成原因句抽取模式, 并利用依存关系和依存路径表示原因句的抽取模式, 再基于该抽取模式来抽取原因信息, ...

☞ 阅读 - 下载 - 法律状态 - 信息查询 - 同类专利

作  
ollection)



# 中文信息检索

- ❖ Chinese Information Retrieval
- ❖ 研究从中文文档中找出满足用户提出的信息需求的技术。
- ❖ 特点：
  - ☞ 词语切分问题
  - ☞ 简繁并存问题
  - ☞ 内码转换问题

南京市长江大桥    南京市长 江大桥

武汉市长江大桥    武汉市长 江大桥





# 文档结构

- ❖ 结构化：内容按照结构组织
  - ✧ 数据表文件
- ❖ 非结构化：无格式
  - ✧ 自然文本
- ❖ 半结构化：部分有结构，部分没有结构
  - ✧ 网页，邮件等

From: <pfli@zhhz.org>

Subject: xxxx

Date: Wed, 15 Sep 2004 07:24:01 +0800

在相同文档中经常出现的概念也作为提问概念被激活，这样就可以自然地、灵活地进行概念扩展，让用户和系统相互作用。



# 和数据库检索的差异

## ❖ 数据库检索

- ❧ 结构化文档
- ❧ 半结构化文档
- ❧ 检索结果精确

## ❖ 文本信息检索

- ❧ 无结构文档
- ❧ 半结构文档
- ❧ 检索结果不精确





# 信息检索评测

## ❖ TREC

✧ Text REtrieval Conference

✧ 文本检索会议

✧ 文本检索领域人气最旺、最权威的评测会议

## ❖ ACM SIGIR

✧ ACM Special Interest Group on Information Retrieval

✧ 信息检索顶级会议



# 信息检索的发展

---

- ❖ 手工检索
- ❖ 脱机检索
- ❖ 联机检索
- ❖ 网络检索
  - ☞ 互联网



# 作用

## ❖ 获取知识的捷径

- ❧ 美国普林斯顿大学学生 约翰·菲利普
- ❧ 图书馆公开资料，四个月画出原子弹设计图
- ❧ 威力相当广岛原子弹3/4，造价两千美元

## ❖ 科学研究的向导

- ❧ 阿波罗飞船燃料箱甲醇会引起钛应力腐蚀
- ❧ 付出数百万美元
- ❧ 早在十多年前，有人研究出来，只需在甲醇中加入2%的水
- ❧ 检索时间10多分钟

## ❖ 情报获取手段

- ❧ 美国“棱镜”计划
- ❧ 中情局90%以上情报来自网络检索



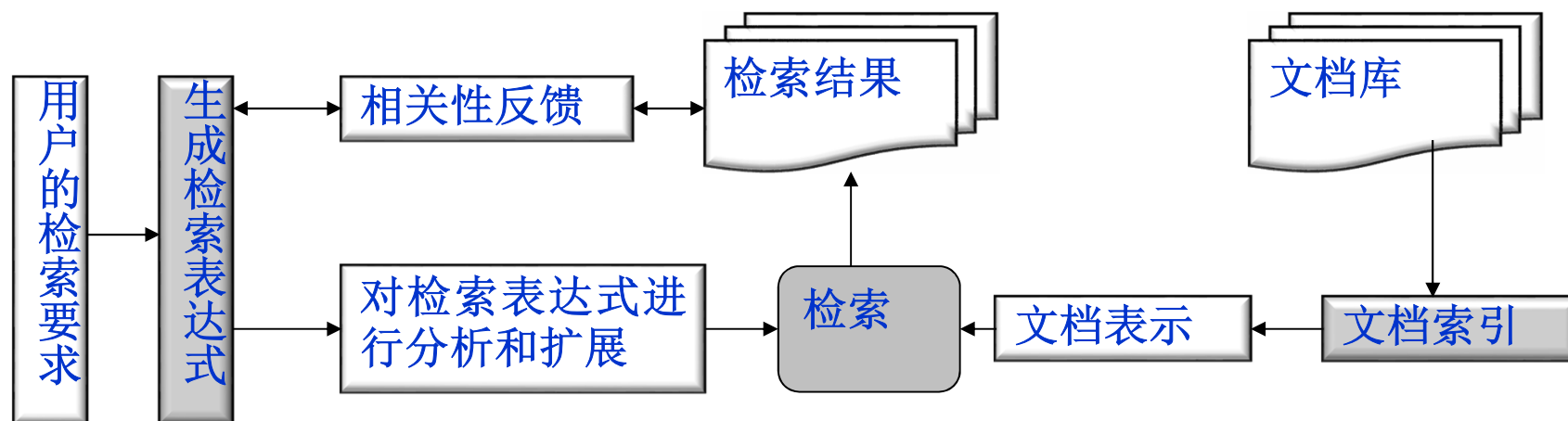
# 内容

---

- ❖ 中文信息检索基础
- ❖ 中文信息检索架构
- ❖ 中文信息检索模型
- ❖ 中文信息检索系统
- ❖ Web信息检索概述
- ❖ 搜索引擎
- ❖ 搜索引擎分类



# 中文信息检索架构



- ❖ 文档索引：整理文档并为这些文档建立索引
- ❖ 文档表示：文档的组织形式
- ❖ 用户检索表达和扩展：生成一个查询表达式，并进行分析和扩展
- ❖ 匹配和检索：从文档库中找出和用户需求相关文档
- ❖ 相关性反馈：把检索结果按相关性反馈给用户



# 性能评价

	真正相关文档RR+NR	真正不相关文档	
系统判定相关 RR+RN (检索出)	RR	RN	→ 准确率
系统判定不相关 (未检索出)	NR	NN	
	↓ 召回率		

❖ 准确率 (Precision) :  $P = RR / (RR + RN)$

❖ 召回率 (Recall):  $R = RR / (RR + NR)$





# 内容

- ❖ 中文信息检索基础
- ❖ 中文信息检索架构
- ❖ 中文信息检索模型
- ❖ 中文信息检索系统
- ❖ Web信息检索概述
- ❖ 搜索引擎
- ❖ 搜索引擎分类



# 检索模型

## ❖ 统计模型

☞ 应用统计手段从被检索文档中查询与用户需求匹配程度最好的文档

## ❖ 语义模型

☞ 对需求实现一定程度语法和语义分析并重新生成查询

## ❖ 统计和语义结合模型



# 统计模型

- ❖ 布尔模型 (Boolean Model)
- ❖ 扩展布尔模型 (Extended Boolean Model)
- ❖ 向量空间模型 (Vector Space Model)
- ❖ 概率模型 (Probabilistic Model)
  - ⌘ 二元独立模型(Binary Independency Model)
  - ⌘ 双泊松模型(Two Poisson Model)
  - ⌘ 推理网络模型(Inference Network Model)
  - ⌘ 信度网络模型(Belief Network Model)
  - ⌘ 贝叶斯网络模型 (Bayesian Network Model)



# 语义模型

- ❖ 知识库
- ❖ 自然语言处理
  - ✧ 语义相似度
  - ✧ 实体指代
  - ✧ .....
- ❖ 潜在语义索引模型
- ❖ 神经网络
- ❖ .....



# 检索表达式

## ❖ 由检索词和逻辑运算符组成

❧ 用检索系统规定的各种算符将检索词之间的逻辑关系、位置关系等连接起来，构成的计算机可以识别和执行的检索命令式

## ❖ 检索表达式分为：

❧ 逻辑表达式

❧ 截词检索表达式

❧ 位置检索表达式

苏州 AND 名小吃

苏州

“苏州大学”



## 项和权值

### ❖ 检索表达式和文档分成项

☞ 项（Term）：文档和检索条件中的字词或短语

❖ 停用词：不具有区分度的普通字、词和短语

❖ 如“我”、“的”等

☞ 索引项：文档中的项

☞ 检索项：检索条件中的项

### ❖ 权值（Weight）

☞ 表示相关性

☞ 一般值为0和1，0表示不相关，1则表示相关





## 项和权值例子

	计算机	电视机	胖娃娃
文档1	1	0	0
文档2	1	1	1
文档3	1	0	0
文档4	0	1	0

项

权值



# 布尔模型

- ❖ 最简单的信息检索模型
- ❖ 能处理逻辑表达式

	计算机	电视机	胖娃娃
文档1	1	0	0
文档2	1	1	1
文档3	1	0	0
文档4	0	1	0

检索1: 计算机 AND 胖娃娃

检索2: 计算机 OR 胖娃娃



# 扩展布尔模型

## ❖ 思想:

- ✧ 计算检索文档中的索引项与用户查询表达式的相似度
- ✧ 按照优先次序排列查询结果

## ❖ 扩展布尔模型常用方法:

- ✧ MMM模型
- ✧ Paice模型
- ✧ P-Norm模型



## P-Norm 模型

### ❖ AND (Term1 AND Term2)

$$\text{sim}(q_{\text{and}}, d_j) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

文档 \ 查询 相似度	计算机 AND 胖娃娃	计算机 OR 胖娃娃
文档1	0.293	0.707
文档2	1	1
文档3	0.293	0.707
文档4	0	0

y表示Term<sub>2</sub>在文档d<sub>j</sub>中的权重 ∈ (0,1)



# 向量空间模型

- ❖ VSM (Vector Space Model)
- ❖ 哈佛大学的Gerard Salton提出
  - ✧ 现代信息检索的奠基人
  - ✧ 开发著名的SMART系统
- ❖ 基本思想
  - ✧ 将文档D和查询Q都用向量表示
  - ✧ 计算文档向量与查询向量相似度
  - ✧ 根据相似度值对检索结果排序





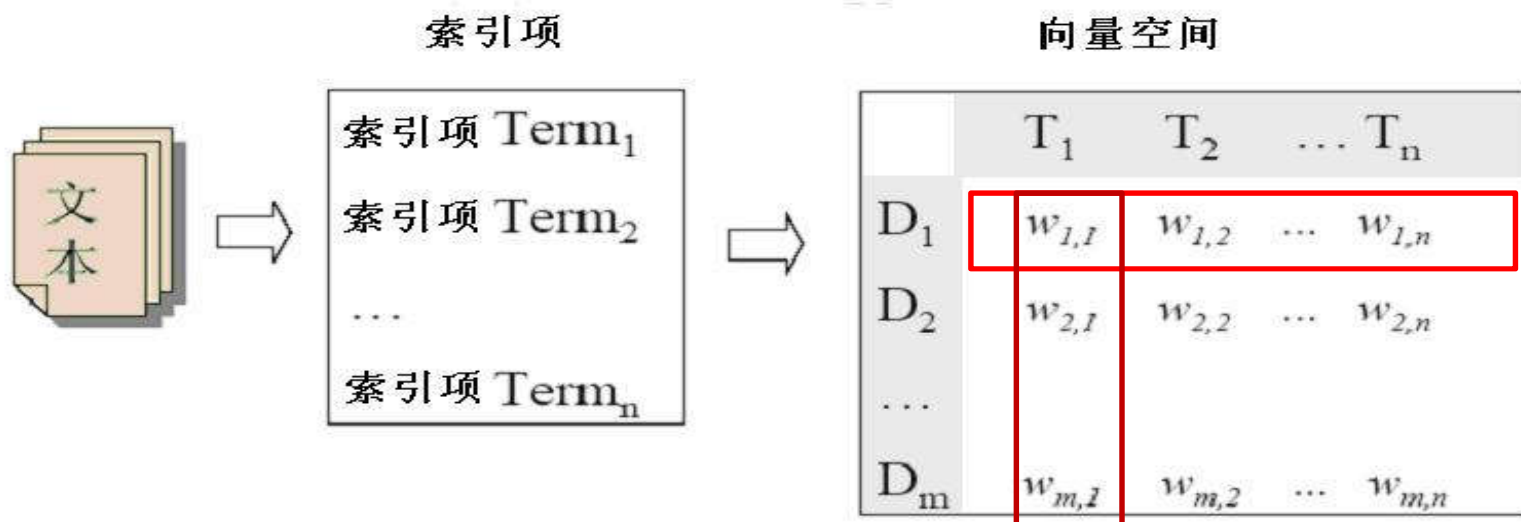
# 向量空间表示

## ❖ 文档空间

每个文档用一个索引项的集合构成的向量表示

## ❖ 项空间

一个索引项在文档集合中的各个文档中权值的集合的向量

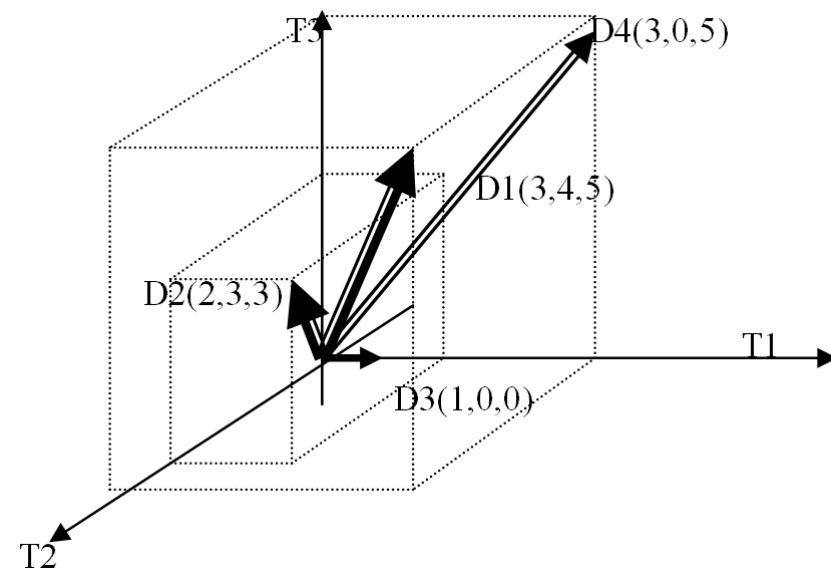






# 例子

	女人(T1)	时装 (T2)	胖娃娃(T3)
文档1 (D1)	3	4	5
文档2 (D2)	2	3	3
文档3 (D3)	1	0	0
文档4 (D4)	3	0	5





# 项选择和权值

## ❖ 项选择

- ✧ 选择最有代表性的项
- ✧ 去掉停用词
- ✧ 选择方法：互信息、CHI等

## ❖ 权值

- ✧ 每个索引项都有一个权值
- ✧ 权值是索引项对文档的重要程度
- ✧ 简单确定方法：是否出现（0未出现；1出现）



# TF\*IDF

- ❖ 项频度\*逆向文档频度加权法

- ❖ 项频度TF (Term Frequency)

- ✧  $\text{term}_i$  在文档  $d_j$  中的出现次数，记做  $\text{tf}_{i,j}$

- ✧  $\text{tf}_{i,j}$  越高， $\text{term}_i$  对于文档  $d_j$  就越重要

- ❖ 文档频度DF (Document Frequency)

- ✧ 含有  $\text{term}_i$  的文档数量，记做  $\text{df}_i$

- ✧  $\text{df}_i$  越高， $\text{term}_i$  在衡量文档之间相似性方面作用越低

- ❖ 逆向文档频度IDF (Inverse DF)

- ✧ 衡量这个项在整个文档的分布情况  $\text{idf}_i = \log \left( \frac{N}{\text{df}_i} \right)$

- ✧ IDF和DF形成反比 (N文档总数)



# 相似度计算

- ❖ 检索条件也表示成一个向量
- ❖ 相似度计算
  - ✧ 计算检索向量和文档向量的相似度
- ❖ 向量间相似程度计算方法：
  - ✧ 内积法（Inner Product）
  - ✧ Dice法（Dice Coefficient）
  - ✧ Jaccard法（Jaccard Coefficient）
  - ✧ 余弦法(**Cosine Coefficient**)



## 余弦法

- ❖ 检索向量  $\vec{q} = (q_1, q_2, \dots, q_n)$
- ❖ 文档向量  $\vec{d} = (d_1, d_2, \dots, d_n)$

$$\text{Cos}(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^n q_i \times d_i}{\sqrt{\sum_{i=1}^n q_i^2} \times \sqrt{\sum_{i=1}^n d_i^2}}$$



## 例子

### ❖ 检索条件

$$\hookrightarrow q=(1, 0, 2)$$

### ❖ 文档

$$\hookrightarrow d_1=(1, 4, 5) \quad d_2=(4, 9, 2) \quad d_3=(0, 0, 10)$$

$$\cos(\vec{q}, \vec{d}_1) = \cos(\theta_1) = \frac{1*1 + 4*0 + 5*2}{\sqrt{1^2 + 2^2} * \sqrt{1^2 + 4^2 + 5^2}} = \frac{11}{\sqrt{5} * \sqrt{42}} = 0.759$$

$$\cos(\vec{q}, \vec{d}_2) = \cos(\theta_2) = \frac{4*1 + 9*0 + 2*2}{\sqrt{1^2 + 2^2} * \sqrt{4^2 + 9^2 + 2^2}} = \frac{8}{\sqrt{5} * \sqrt{101}} = 0.356$$

$$\cos(\vec{q}, \vec{d}_3) = \cos(\theta_3) = \frac{0*1 + 0*0 + 10*2}{\sqrt{1^2 + 2^2} * \sqrt{10^2}} = \frac{20}{\sqrt{5} * \sqrt{100}} = 0.894$$





# 概率模型的缺点

## ❖ 存在问题

❧ 无法处理同义词

❖ 计算机 电脑

❧ 无法处理表达形式的不一致

❖ 苏州大学 苏大

❧ 无法处理以概念为核心的检索

❖ 苹果：水果

❖ 苹果：计算机厂商

❖ 苹果：范冰冰梁家辉主演的电影

❖ .....  
.....



# 语义模型

## ❖ 结合自然语言处理技术

### ☞ 词义消歧

#### ❖ 人名消歧义

### ☞ 简称识别

#### ❖ 中国 中华人民共和国

### ☞ 指代消解


#### ❖ 朱巧明 朱教授 老朱

### ☞ .....

## ❖ 信息检索的发展方向





# DuckDuckGo


×🔍▼更多 ▼


## Apple


The apple is the pomaceous fruit of the apple tree, species *Malus domestica* in the rose family (Rosaceae).

 [更多在Wikipedia](#)


 [Apple Inc.](#)

 [Apples](#)

 [Sequenced genomes](#)



**Apple: Computers, Phones, MP3 Players - Best Buy**  
Shop online for **Apple** products at Best Buy, including Mac computers,...  
[bestbuy.com/site/Brands/Apple/pcmcat128500050005.c?...](#) Sponsored link

 **Apple**  
Shop the **Apple** Online Store (1-800-MY-APPLE), featuring MAC, iPod, iPhone, iPad, iTunes, service, and support.  
[apple.com](#) [More from apple.com ▶](#)

2017年12月24日 12时10分

33

外州八子：中文信息处理



# KNGINE

[Home](#) [Mobile](#) [Technology](#) [Business/Developer](#) [Company](#) [Press Kit](#) [Jobs](#) [Blog](#)



Beta  
KNGINE

CHICAGO

Know

Chicago

Chicago (2002 Movie)

Chicago (band)

Chicago (1998 Movie)

Chicago



book subject, comic book location, fictional setting, film location, governmental jurisdiction, citytown, dated location, location, place with neighborhoods, statistical region, olympic bidding city, organization scope, sports team location, travel destination, business location, employer, hud county place, hud foreclosure area, administrative division, newspaper circulation area

Population: 2.851 Million (2009)



# 内容

- ❖ 中文信息检索基础
- ❖ 中文信息检索架构
- ❖ 中文信息检索模型
- ❖ 中文信息检索系统
- ❖ Web信息检索概述
- ❖ 搜索引擎
- ❖ 搜索引擎分类



# 论文检索

## ❖ C N K I 中国知网

### 🔗 中国期刊全文数据库

- ❖ 目前世界上最大中国期刊全文数据库，累积全文文献约**3200**多万篇
- ❖ 国内公开出版的约**9100**多种综合期刊与专业特色期刊的全文

### 🔗 中国博士硕士学位论文全文数据库

- ❖ 博硕士学位论文全文文献**87.5**万多篇

### 🔗 中国会议论文全文数据库

### 🔗 中国重要报纸全文数据库



# 论文检索

中国期刊全文数据库

逻辑 检索项 检索词 词频 扩展

从 1999 到 2012 更新全部数据 范围全部期刊 匹配模糊 排序时间 每页20 中英扩展

共有记录3955条

序号	篇名	作者	刊名	年/期
1	数字图书馆信息检索过程中情绪研究的测量方法	付亚楠	郑州铁路职业技术学院学报	2012/04
2	国外协同信息检索行为研究述评	吴丹	中国图书馆学报	2012/06
3	基于共现分析的语义信息检索研究	邱均平	中国图书馆学报	2012/06
4	网络信息检索中数据挖掘的应用	杨丽	信息系统工程	2012/11
5	用户协同信息检索行为与系统评价研究——以任务类型和协同能力为视角	邱瑾	现代图书情报技术	2012/09
6	基于书目检索信息的图书推荐系统	刘佳佳	图书情报工作	2012/15
7	粗糙本体支持的信息语义检索	樊皓	计算机工程与设计	2012/12
8	基于词语-概念相关度的关键词语义信息检索方法	吕义	辽宁工业大学学报(自然科学版)	2012/05
9	信息检索课的信息素养教育研究	吴权喜	湖北广播电视大学学报	2012/12
10	偏振旋转的量子私有信息检索方案	易运晖	电子与信息学报	2012/10

❖ [http://library.suda.edu.cn/app\\_cust/database/detail.jsp?key=100](http://library.suda.edu.cn/app_cust/database/detail.jsp?key=100)





# 图书检索

JALIS 江苏省高等教育文献保障系统

超星数字图书馆南京大学图书馆镜像站点

首页 | JALIS主页 | 学校主页 | 下载注册机 | 下载浏览器 | 使用说明 | 超星主页

超星数字图书馆->军事图书馆->世界军事->各种武装力量(各军、兵种)

首页 上一页 下一页 尾页 页次: 1/2页 共28条记录 转到: 1 GO

**信息检索**

书名    
 全部    
 查询 高级检索

**公告栏**

北京时代超星公司与江苏省高等教育文献保障系统管理中心合作建设《超星数字图书馆》镜像站点,向全省高校教师同学提供校园网的阅读、下载服务,2007年初,JALIS《超星数字图书馆》收藏53万种中文电子图书,江苏省内高校承诺遵守相关规定,均可以使用。

**环球军事力量概览**

阅读 下载   
 作者: 汤奇 SS号: 11510886 索书号: E15 出版日期: 2005年12月第1版 页数: 481

**抢占先机 现代军事纵横谈**

阅读 下载   
 作者: 陶宇著 SS号: 11601378 索书号: E15 出版日期: 2005年12月第1版 页数: 130

**特种部队制敌绝技 (下册) (第2版)**

阅读 下载   
 作者: 吴巍 SS号: 11613355 索书号: E156/2: 2 出版日期: 2005年10月第2版 页数: 374

**世界特种部队大观 (上册) (第2版)**

阅读 下载   
 作者: 肖达喜 SS号: 11613364 索书号: E156/1: 2 出版日期: 2005年10月第2版 页数: 180

**世界特种部队大观 (下册) (第2版)**

阅读 下载   
 作者: 肖达喜 SS号: 11613366 索书号: E156/1: 2 出版日期: 2005年10月第2版 页数: 560

军事、经济  
呆等

p?lib=0E10





# 专利搜索

## ❖ SooPAT

**SooPAT**

☐全部专利 ☐发明和实用新型 ☐发明 ☐实用新型 ☐外观 ☐发明授权

显示: ☒ 搜索式 ☐ 两栏式 ☐ 多图式 ☐ 表格式 ☐ 只搜外观

**搜索结果统计**

申请人  
发明人  
申请日  
公开日  
分类号  
外观分类  
更多

**全部**  
有专利权  
无专利权  
审查中

**[发明] 蚕丝香烟滤嘴及其制备方法 - 200710039386.7 [有效]**  
申请人: 苏州大学 - 申请日: 2007-04-12 - 主分类号: A24D3/06(2006.01)I  
地址: 215006江苏省苏州市沧浪区十梓街1号苏州大学...  
摘要: 本发明公开了一种以天然蛋白质纤维蚕丝作为主要原料制备而成的香烟滤嘴,它以蚕丝或去蛹后的蚕茧拉松均匀后形成的纤维网状蚕丝为滤嘴的滤芯材料,并采用在纤维表面涂层、高温高湿汽蒸进行定型处理等方法,赋予蚕丝香烟滤嘴具有一定的硬度、弹性和吸湿能力。试...  
[阅读](#) - [下载](#) - [法律状态](#) - [信息查询](#) - [同类专利](#)

**[发明] 一种细胞培养支架材料及其制备方法 - 200610041017.7 [有效]**  
申请人: 苏州大学 - 申请日: 2006-07-13 - 主分类号: A61L27/34(2006.01)I  
地址: 215006江苏省苏州市沧浪区十梓街1号苏州大学...  
摘要: 本发明公开了一种细胞培养支架材料及其制备方法。它以蚕丝为主要原料,经脱胶、溶解、纯化、干燥形成丝素蛋白,与胶原蛋白或明胶天然蛋白质分别溶解于同种溶剂后,依次采用高压静电纺丝制得以丝素蛋白为基材、胶原蛋白、明胶等为复合层的三维网状丝素复合纳米...  
[阅读](#) - [下载](#) - [法律状态](#) - [信息查询](#) - [同类专利](#)



# 内容

---

- ❖ 中文信息检索基础
- ❖ 中文信息检索架构
- ❖ 中文信息检索模型
- ❖ 中文信息检索系统
- ❖ **Web信息检索概述**
- ❖ 搜索引擎
- ❖ 搜索引擎分类



# 和文本检索的区别

## ❖ Web搜索

棱镜

网页

图片

地图

新闻

更多 ▾

搜索工具

找到约 11,400,000 条结果 (用时 0.36 秒)

## ❖ Web搜索

棱镜的新闻搜索结果



英美沆瀣一气共施“棱镜”计划

中國藝術報 - 10 小时前

英美沆瀣一气共施“棱镜”计划—英国大量参与美国的监控计划并获取海量“个人敏感信息”。

## ❖ 文本信息

美国前总统小布什力挺棱镜计划自曝由他发起

人民网温州视窗 - 8 小时前

“棱镜”再曝惊人内幕欧盟震惊遭美国窃听

环球网 - 1 天前

## ❖ 面向海量

棱镜门\_百度百科

[baike.baidu.com/view/10688863.htm](http://baike.baidu.com/view/10688863.htm) ▾

据美国中情局前职员爱德华·斯诺登爆料：“棱镜”窃听计划，始于2007年的小布什时期，美国情报机构一直在九家美国互联网公司中进行数据挖掘工作，从音视频、图片、 ...

棱镜\_百度百科

[baike.baidu.com/view/47349.htm](http://baike.baidu.com/view/47349.htm) ▾

一种由两两相交但彼此均不平行的平面围成的透明物体，用以分光或使光束发生色散...

2017年12月24日 12时10分



## 面临挑战

### ❖ Google索引网页

- ✧ 1998年：2600万
- ✧ 2000年：10亿
- ✧ 2008年：10000亿
- ✧ 2013年：????

### ❖ Googlebot网页爬虫

- ✧ 每天会走过大约200亿个网页
- ✧ 追踪300亿个左右独立URL链接
- ✧ 搜索请求：大约1000亿次/月





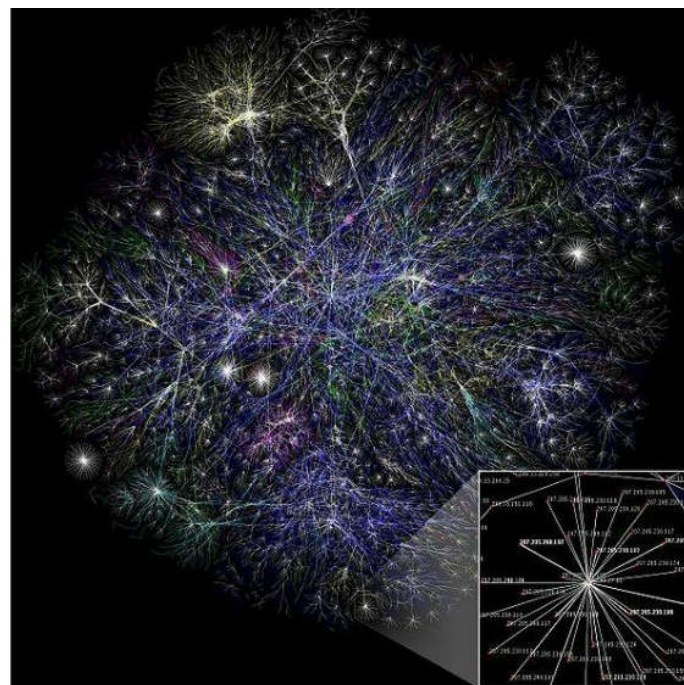


# Web信息检索

❖ 以互联网为文档库进行检索

❖ Web文档特点:

- ❧ 数据分布
- ❧ 数据量巨大
- ❧ 数据的动态性
- ❧ 结构性差且数据冗余
- ❧ 数据质量不高
- ❧ 数据的异构
- ❧ .....





# Web信息检索要点

- ❖ 考虑计算时间
  - ⌘ 关键词匹配为主
- ❖ 考虑性能
  - ⌘ 查询预处理和扩展（分词、同义词扩展等）
  - ⌘ 页面排序方法（海量页面）
- ❖ 可以利用
  - ⌘ 链接
  - ⌘ HTML标记



# 内容

- ❖ 中文信息检索基础
- ❖ 中文信息检索架构
- ❖ 中文信息检索模型
- ❖ 中文信息检索系统
- ❖ Web信息检索概述
- ❖ 搜索引擎
- ❖ 搜索引擎分类



# 什么是搜索引擎

## ❖ 搜索引擎（Search Engine）

- ❧ 以Web页面（或超链接）为检索文档
- ❧ 核心是信息检索技术
- ❧ 一个系统工程：

❖ Web页面的抓取、分类、索引、存储、更新等





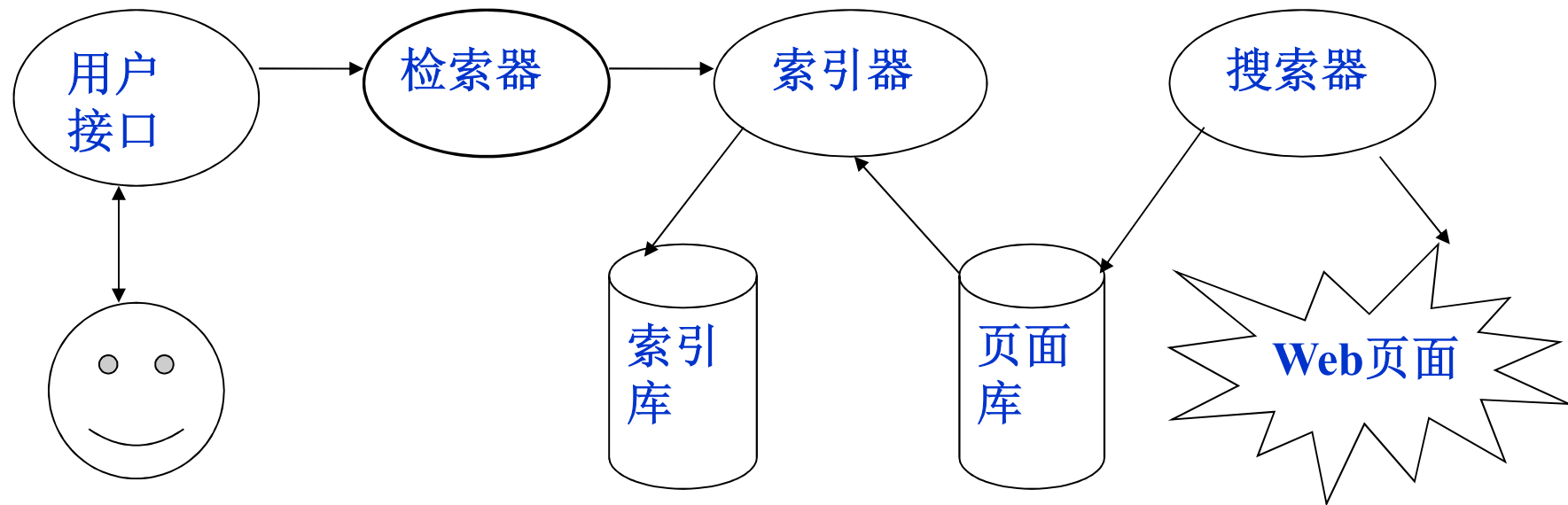


## 发展过程

- ❖ 1990年初:万维网还未出现, Archie、Gopher等搜索工具
- ❖ 1994年4月:Yahoo目录诞生
- ❖ 1994年7月: Lycos-基于robot的数据发掘技术, 并支持搜索结果相关性排序
- ❖ 1995年:第一个中文搜索引擎蕃薯藤 (繁体)
- ❖ 1995年: 元搜索引擎(Meta Search Engine)
- ❖ 1995年12月: AltaVista大量创新功能, 第一个支持自然语言搜索
- ❖ 1998年10月:Google诞生
- ❖ 2000年1月:百度



# 搜索引擎结构





# 搜索器

## ❖ Spider或Crawler

- ❧ 在Internet遍历网址，发现和搜集网页信息
- ❧ 常常是一个机器人（Robot）程序不停运行
- ❧ 要尽可能多、尽可能快地搜集新网页
- ❧ 定期更新已经搜集过的旧网页，以避免死链接和无效链接





# Robots 协议

❖ 爬虫 禁止所有搜索引擎访问网站的任何部分

❖ 网站 **User-agent: \***  
**Disallow: /**



❖ 奇虎 禁止某个搜索引擎的访问

**User-agent: BadBot**  
**Disallow: /**

取

范

保用户个

网站信息





# 索引器

## ❖ 功能

- ❧ 预处理（提取文字；分词；去停止词；消噪）
- ❧ 理解搜索器所搜索的信息
- ❧ 从中抽取出索引项，用于表示网页以及生成页面库的索引表



# 倒排索引

## ❖ 倒排索引

- ❧ 实际应用中需要根据项的值来查找网页。
- ❧ 索引表中的每一项包括项本身和具有该属性值的各网页的地址
- ❧ 倒排：不是由记录来确定属性值，而是由属性值来确定记录位置

项	网页	网页	...	网页
TERM <sub>1</sub>	Doc <sub>i</sub>	Doc <sub>j</sub>	...	Doc <sub>m</sub>
TERM <sub>2</sub>	Doc <sub>i</sub>	Doc <sub>k</sub>	...	Doc <sub>n</sub>
...	...	...	...	...
TERM <sub>s</sub>	Doc <sub>j</sub>	Doc <sub>m</sub>	...	Doc <sub>p</sub>



和信息检索相同

## 分词

## 扩展

## ❖ 查询纠错

## 哪个搜索引擎最最最烂


哪个搜索引擎最烂 百度知道

4个回答 - 提问时间: 2007年07月26日

最佳答案 这个问题,首先除非是专业开发搜索引擎,并知道某个品牌搜索引擎的技术数据的才能从某个侧面说别人烂。所以请不要随便说某个搜索引擎烂了什么的。有...

zhidao.baidu.com/question/316046...html 2013-1-24 - 百度快照

苏洲大学

 您要找的是不

苏州大学高考招

## 最烂的搜索引擎--google (谷歌) 冰冻四尺\_百度空间

好好的**百度搜索引擎**,而且百度就非常的方便快捷.但这几天经常是一用百度搜索就...恰恰也证明了自己的**最烂**。**google**(谷歌)到底是哪些烂人来设置的,不在软件的...

hi.baidu.com/llzdbk/item/74f493dbc51... 2012-9-14 - 百度快照

大家来评评哪个数据库的搜索引擎是最烂的 - 论文投稿 - 小木虫 - ...

你来评一下哪个搜索引擎是最烂的,记住是最烂的,不是最好的。我选美国化学会的,以前的版本还好,新的版本简直就是不让你搜索。...

emuch.net/html/201107/33544...html 2013-6-16 - 百度快照

院校信息: [专业介绍](#) [招生章程](#) [招生计划](#)

阔千



# 检索器

## ❖ 功能

- ✧ 根据用户的查询在索引库中快速检出网页
- ✧ 进行网页与查询的相似度评价
- ✧ 对将要输出的结果进行排序
- ✧ 实现用户相关性反馈机制





# 页面排序

- ❖ 基于相关度排序
  - ⌘ 查询与网页内容相似性
- ❖ 基于重要性排序
  - ⌘ 基于链接分析计算
- ❖ 综合排序

Google

网球

网页

图片

地图

新闻

更多 ▾

搜索工具

找到约 63,600,000 条结果 (用时 0.28 秒)

相关搜索: [网球吧](#) [网球规则](#)



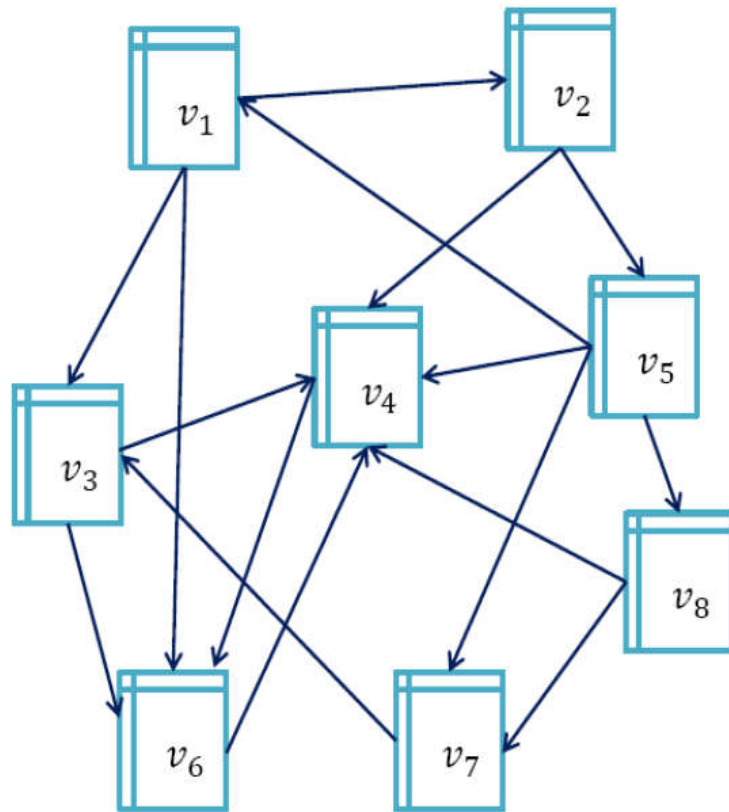
# PageRank

## ❖ Google专有的算法

- ✧ 衡量特定网页对于搜索引擎索引中的其他网页的重要程度
- ✧ 页面的得票数由所有链向它的页面的重要性决定
- ✧ 页面的PageRank是由所有链向它的页面（“链入页面”）的重要性经过递归算法得到
- ✧ 有较多链入的页面会有较高的等级
- ✧ 页面没有任何链入页面，那么它没有等级



# 例子



Adjacent Matrix

$$M = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$



## 例子 (续)

Transition Probability Matrix  $D = \{d_{ij}\}$

ID	PR	Inlink	Outlink
1	0.0250	$v_5$	$v_2, v_3, v_6$
2	0.0259	$v_1$	$v_4, v_5$
3	0.0562	$v_1, v_7$	$v_4, v_6$
4	0.4068	$v_2, v_3, v_5, v_6, v_8$	$v_6$
5	0.0298	$v_2$	$v_1, v_4, v_7, v_8$
6	0.3955	$v_1, v_3, v_4$	$v_4$
7	0.0357	$v_5, v_8$	$v_3$
8	0.0251	$v_5$	$v_4, v_7$

0  
0  
0  
0  
/4  
0  
0  
0

$\pi(v_i)$

$P^T \pi$

$v_j \in \text{inlink}[v_i]$



# HITS

## ❖ Hub网页

- ❧ 提供指向Authority(good source of contents) 网页链接集合的网页
- ❧ 本身可能并不重要，没有几个网页指向它
- ❧ 提供了指向某个主题最为重要的站点链接集合
  - ❖ 比如：课程目录

## ❖ 思想：

- ❧ 好的Hub网页指向许多好的Authority网页
- ❧ 好的Authority网页有许多好的Hub网页指向自己
- ❧ Hub与Authority网页之间的相互加强关系



# 百度

新闻 网页 贴吧 知道 MP3 图片 视频 地图 图书

**Baidu 百度**  **关键字竞标** [设置](#) [高级搜索](#)

把百度设为主页 百度一下，找到相关网页约100,000条

**卓越亚马逊:《图书》特价免运费!**  
卓越亚马逊:《图书》超低价,全球著名的中文网上书店,免费送货,货到付款,百万图书任你选,低至5折,畅销书惊喜折上折,正品保证,假一罚二,15日可退换!  
[www.Amazon.cn](http://www.amazon.cn) 2010-03 - 推广

**买图书到当当,正版图书2折起,免运费!**  
买图书到当当网,免运费!当当网,全球领先的综合性中文网上书店和购物中心!80万种正版图书音像2折起,文学,小说,经管,教材等上百品类任你选!免运费,700个城市支持货到付款!  
[www.DangDang.com](http://www.DangDang.com) 2010-03 - 推广

**中国国家图书馆-中国国家数字图书馆**  
原称北京图书馆。履行搜集、加工、存储、研究、利用和传播知识信息的职责,地上书库19层,地下书库3层,总共可容纳2000万册藏书。主持编制了《中国国家书目》、《民国...  
[www.nlc.gov.cn/](http://www.nlc.gov.cn/) 2010-3-12 - 百度快照

**当当图书 - 全球最大的中文网上书店**  
您现在看到的是新版图书首页,对于这个页面您有什么想法,都欢迎反馈给我们。我们会认真考虑您的意见和建议,把图书首页做得更好,谢谢! 具体建议(不超过500字) ...  
[book.dangdang.com/](http://book.dangdang.com/) 2010-3-11 - 百度快照

**法律图书馆\_法律法规数据库\_法律论文\_法律图书\_中国法律门户网站**  
法律图书馆\_中国法律门户网站,提供法律法规数据库、法学论文、裁判文书、律师黄页、法治动态、司法考试资料、法律图书、法律书刊、法律书摘、著者介绍、出版社介绍等资料...  
[www.law-lib.com/](http://www.law-lib.com/) 2010-3-12 - 百度快照

**超星数字图书馆 | 每天不到1元钱,把图书馆搬回家**  
致力于纸张图文资料数字化技术及相关应用与推广,为国内外图书馆、档案馆和出版社数字化提供了重要的整体解决方案。  
[www.superstar.com/](http://www.superstar.com/) 2010-3-10 - 百度快照

**图书防盗仪系统专业生产厂家**  
龙口伍洋电子有限公司提供图书防盗仪系统  
图书防盗仪生产与销售热线:0535-8515637  
[www.sd-wuyang.com](http://www.sd-wuyang.com)

**顶尖选股机构 黑马短线股强**  
专业权威分析师提供机械走势实时分析,机械,暴涨牛股黑马强力出击,机械。  
[www.180333.com](http://www.180333.com)

**今日最新股票行情 投资理财**  
股票黑马,股票行情分析,天天冲击涨停股票盈利100%!  
[www.gupiaodafa.info](http://www.gupiaodafa.info)

**北京世纪超讯专业的图书防盗**  
北京世纪超讯科技长期致力于研发,生产图书防盗仪,引进国际先进技术,为您提供业界顶尖  
[www.csstech.com.cn](http://www.csstech.com.cn)

**图书防盗设备厂家免费安装**  
专业从事超市防盗,商品防盗的专业制造商,生产的超市防盗系统具有抗干扰力强,性能稳定  
[www.htcdz.com/cp/class](http://www.htcdz.com/cp/class)

**图书防盗 龙口图新**  
我公司专业研发生产优质图书防盗系统  
图书防盗 免费保修三年 终生服务  
[www.lktuxin.com](http://www.lktuxin.com)





# 内容

- ❖ 中文信息检索基础
- ❖ 中文信息检索架构
- ❖ 中文信息检索模型
- ❖ 中文信息检索系统
- ❖ Web信息检索概述
- ❖ 搜索引擎
- ❖ 搜索引擎分类



# 目录搜索引擎

## ❖ Yahoc

人工

人工

信息  
检索



[Yellow Pages](#) - [People Search](#) - [Maps](#) - [Classifieds](#) - [News](#) - [Stock Quotes](#) - [Sports Scores](#)

- [Arts and Humanities](#)  
[Architecture](#), [Photography](#), [Literature](#)...
- [Business and Economy \[Xtra!\]](#)  
[Companies](#), [Investing](#), [Employment](#)...
- [Computers and Internet \[Xtra!\]](#)  
[Internet](#), [WWW](#), [Software](#), [Multimedia](#)...
- [Education](#)  
[Universities](#), [K-12](#), [College Entrance](#)...
- [Entertainment \[Xtra!\]](#)  
[Cool Links](#), [Movies](#), [Music](#), [Humor](#)...
- [Government](#)  
[Military](#), [Politics \[Xtra!\]](#), [Law](#), [Taxes](#)...
- [Health \[Xtra!\]](#)  
[Medicine](#), [Drugs](#), [Diseases](#), [Fitness](#)...
- [News and Media \[Xtra!\]](#)  
[Current Events](#), [Magazines](#), [TV](#), [Newspapers](#)...
- [Recreation and Sports \[Xtra!\]](#)  
[Sports](#), [Games](#), [Travel](#), [Autos](#), [Outdoors](#)...
- [Reference](#)  
[Libraries](#), [Dictionaries](#), [Phone Numbers](#)...
- [Regional](#)  
[Countries](#), [Regions](#), [U.S. States](#)...
- [Science](#)  
[CS](#), [Biology](#), [Astronomy](#), [Engineering](#)...
- [Social Science](#)  
[Anthropology](#), [Sociology](#), [Economics](#)...
- [Society and Culture](#)  
[People](#), [Environment](#), [Religion](#)...

类别  
服务和直接

泡泡网 PCPOP.COM





# 全文搜索引擎

- ❖ Google
- ❖ 百度
- ❖ 机器人自动抓取网页
- ❖ 用户输入查询条件



[新闻](#) [网页](#) [贴吧](#) [知道](#) [音乐](#) [图片](#) [视频](#) [地图](#)

百度一下

[百科](#) [文库](#) [hao123](#) | [更多>>](#)



Google 搜索

手气不错

将 Google 设置为主页

Google.com.hk 使用下列语言: [中文\(繁體\)](#) [English](#)



# 元搜索引擎

- ❖ 没有自
- ❖ 将用户
- ❖ 将返回
- ❖ 理后，
- ❖ 代表：

Web  
中文

网页 资讯 图片 网站 导航

网页 搜魅 中文元搜索

羽绒服 清洗 搜魅

搜魅 > 网页 > 羽绒服 清洗 搜魅聚合 百度 谷歌 搜狗 雅虎 必应 中搜 搜搜

怎样清洗羽绒服\_百度知道

2006年1月4日 ... 我常用洗涤剂洗羽绒服，清洗后把衣服水份滤一下，就搭在衣架上，结果衣服 ... 经验一

<http://zhidao.baidu.com/question/2552023> - 44k

羽绒服的清洗保养方法\_雅虎知识堂

2006年12月26日 ... 如果羽绒服被弄脏后清洗不得当，就会导致保暖性下降，也失去了原先的蓬松感。 ...

<http://ks.cn.yahoo.com/question/1306122613209.html> - 51k

五妙招教你自己清洗羽绒服\_中国经济网——国家经济门户

2009年1月25日 ... 羽绒服内的禽类羽绒为蛋白质纤维，若使用肥皂或普通洗衣粉清洗（更不能使用加酶洗衣粉）

[http://boaoforum.ce.cn/life/right/fwxx/200901/25/t20090125\\_18057653.shtml](http://boaoforum.ce.cn/life/right/fwxx/200901/25/t20090125_18057653.shtml) - 41k

小丽搭配5妙招教你自己清洗羽绒服-搜狐女人

2009年1月15日 ... 羽绒服内的禽类羽绒为蛋白质纤维，若使用肥皂或普通洗衣粉清洗（更不能使用加酶洗衣粉）

<http://women.sohu.com/20090115/n261746996.shtml> - 126k

支招：如何清洗和收藏羽绒服\_清洗|收藏\_服饰频道\_名品网

2009年2月23日 ... 当脱去了厚重的羽绒服后，随之而来的洗涤也就变得十分重要。它既是一个力气活，也是个

<http://dress.mplife.com/help/xiaobianguide/090221/24805326101.shtml> - 29k

清洗羽绒服的妙招，无需干洗就可以的哦~! 阿里巴巴hongmei888666的 ...

2007年8月21日 ... 清洗羽绒服的妙招，无需干洗就可以的哦~!，各位姐妹们好啊，又到了冬季，羽绒服算的

<http://blog.china.alibaba.com/blog/hongmei888666/article/b0-i1541019.html> - 73k

洗涤羽绒服妙法

洗涤羽绒服一忌碱性物，二忌用洗衣机搅动或用手揉搓，三忌拧绞，四忌明火烘烤。 ...

<http://www.hzfx.com.cn/info/html/200472992159.htm> - 11k

序等处。



# 垂直搜索

## ❖ 专注于特定搜索领域和需求

**Baidu 团购** 苏州 | 美食 百度一下

全部团购分类 ▾ 首页 餐饮美食 休闲娱乐 电影 旅游住宿 生活服务 丽人 商品

频道: 全部 **餐饮美食** 休闲娱乐 旅游住宿 生活服务 丽人 商品

分类: **全部** 火锅 447 烧烤 220 西餐 222 海鲜 137 地方菜 1288 烤鱼 42 麻辣香锅 4 日韩料理 181 快餐 54  
蛋糕 561 其他 693 东南亚菜 4 咖啡 131 下午茶 38

特色: **全部** 自助 549 双人套餐 433 多人聚餐 1767

区域: **全部** 吴中区 1529 沧浪区 1011 平江区 726 金阊区 719 虎丘区 325 相城区 305

热门区域: 观前街 462 十全街 494 南门 357 石路 346 凤凰街 580 南浩街 304 邻里中心 330 左岸商业街 383

价格: **全部** 10元以下 463 10元-50元 795 50元-100元 1303 100元-200元 686 200元以上 775



# 中文搜索引擎排名

搜索引擎网站Alexa排名

综合统计 | Alexa排名 | 百度权重 | PR值



百度 www.baidu.com

Alexa周排名: 6 百度权重: 9 PR: 9 反链数: 490546

网站简介: 百度, 全球最大的中文搜索引擎、最大的中文网站。2000年1月创立于北京中关村。

1

得分:4962

[查看榜单](#)



谷歌中国 www.google.com.hk

Alexa周排名: 23 百度权重: 7 PR: 8 反链数: 32752

网站简介: Google是全球最大的并且最受欢迎的搜索引擎, 主要的搜索服务有: 网页搜索, 图片搜索, 视频搜索, 地图搜索, 新闻搜索, 购物搜索, 博客搜索, 论坛搜索, 学术搜索, 财经搜索。2010年3月23日凌晨, Google公司总部发表声明...

2

得分:3586



搜搜 www.soso.com

Alexa周排名: 52 百度权重: 9 PR: 7 反链数: 22811

网站简介: 搜搜是腾讯旗下的搜索网站, 是腾讯主要的业务单元之一。网站于2006年3月正式发布并开始运营。搜搜目前已成为中国网民首选的三大搜索引擎之一, 主要为网民提供实用便捷的搜索服务, 同时承担腾讯全部搜索业务, 是腾讯整...

3

得分:4278



搜狗 www.sogou.com

Alexa周排名: 108 百度权重: 9 PR: 7 反链数: 35762

网站简介: 搜狗是搜狐公司的旗下子公司, 于2004年8月3日推出, 目的是增强搜狐网的搜索技能, 主要经营搜狐公司的搜索业务。在搜索业务的同时, 也推出搜狗输入法、免费邮箱、企业邮箱等业务。2010年8月9日搜狗与阿里巴巴宣布将分...

4

得分:4269



谷歌台湾 www.google.com.tw

Alexa周排名: 134 百度权重: 6 PR: 8 反链数: 1962

网站简介: 谷歌台湾 (www.google.com.tw) 提供网络、新闻、图片等搜索服务。特色包括PageRank网页排行榜、网页存档、搜索结果翻译及类似网页查询。

5

得分:3227

❖ 2013-07-01

2017年12月24日 12时10分



苏州大学

谢谢!

