信息检索综合课程设计
课程介绍

主讲人：李正华
苏州大学计算机学院
2019年3月12日

主要参考陈文亮老师课件

自我介绍

- 姓名：李正华
- 电子邮件：zhli13@suda.edu.cn
- 个人主页：http://hlt.suda.edu.cn/~zhli
- 课程主页：http://hlt.suda.edu.cn/index.php/Ir-2019-spring
- 研究方向：自然语言处理、人工智能

学习目的

- 学习信息检索基础知识
- 动手构建一些小系统（编程实践）
- 拿到本门课学分

提纲

- 什么是信息检索？
- 为什么要学习信息检索？

提纲

- 什么是信息检索？
- 为什么要学习信息检索？

什么是信息检索

- 现场问答
- 来几个互联网应用例子.....

- 你们来之前

信息检索

SafeSearch on ▼

Web Images Books News Maps More ▾ Search tools

About 9,400,000 results (0.25 seconds)

信息检索- 维基百科，自由的百科全书
zh.wikipedia.org/zh/信息檢索 ▾ 转为简体网页
資訊檢索（英语：Information Retrieval）是指搜尋資訊的科學，如在文件中搜尋資訊、搜尋文件本身、搜尋描述文件的metadata或是在資料庫中進行搜尋，無論是在相關 ...

文本信息检索- 维基百科，自由的百科全书
zh.wikipedia.org/zh/文本信息檢索 ▾ Translate this page
文本信息检索是针对文本的信息檢索技术。在技术社区中，文本信息检索常常被等同于信息检索技术本身。 相对视频、音频检索而言，文本信息检索是发展较快也较 ...
历史介绍 - 模型 - 倒排文档索引技术 - 关键词权重

信息检索- MBA智库百科
wiki.mbalib.com/wiki/信息检索 ▾ Translate this page
信息检索（Information Retrieval）"信息检索"一词出现于20世纪50年代,又称信息存贮与检索、情报检索，是指将信息按一定的方式组织和存储起来，并根据信息用户的 ...

信息检索_互动百科
www.baike.com/wiki/信息检索 ▾ Translate this page
信息检索一词出现于20世纪50年代，又称信息存贮与检索、情报检索，是指将信息按一定的方式组织和存储起来，并根据信息用户的需要找出有关的信息的过程和技术 ...

信息检索- 搜搜百科
baike.soso.com/v88322.htm ▾ Translate this page
信息检索（Information Retrieval）是指信息按一定的方式组织起来，并根据信息用户的需

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

信息检索 百度一下 推荐：用手机随时随地上百度

信息检索_百度百科

信息检索（Information Retrieval）是指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。狭义的信息检索就是信息检索过程的后半部分，...

起源 定义 类型 主要环节 热点 检索原因

baike.baidu.com/ 2013-10-09

excel 中信息检索怎么关闭了_百度知道

9个回答 - 提问时间: 2011年06月26日

最佳答案: 视图--任务窗格 取消了.

zhidao.baidu.com/link?url=FQ23aECvp2JL4vA... 2011-02-25

excel表格里如何取消信息检索?	8个回答	2011-09-16
怎么关闭excel的信息检索?	2个回答	2012-06-28
信息检索	3个回答	2012-01-08

更多知道相关问题>>

相关书籍


信息检索导论

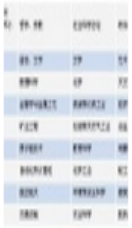

网络信息检索与利用

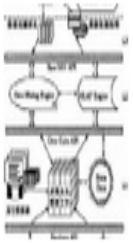

信息检索教程


文献信息检索与利用

相关数据库术语


万方数据库


中文科技期刊数据库


知识发现

信息检索

编辑

信息检索（Information Retrieval）是指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。狭义的信息检索就是信息检索过程的后半部分，即从信息集合中找出所需要的信息的过程，也就是我们常说的信息查寻（Information Search 或Information Seek）。

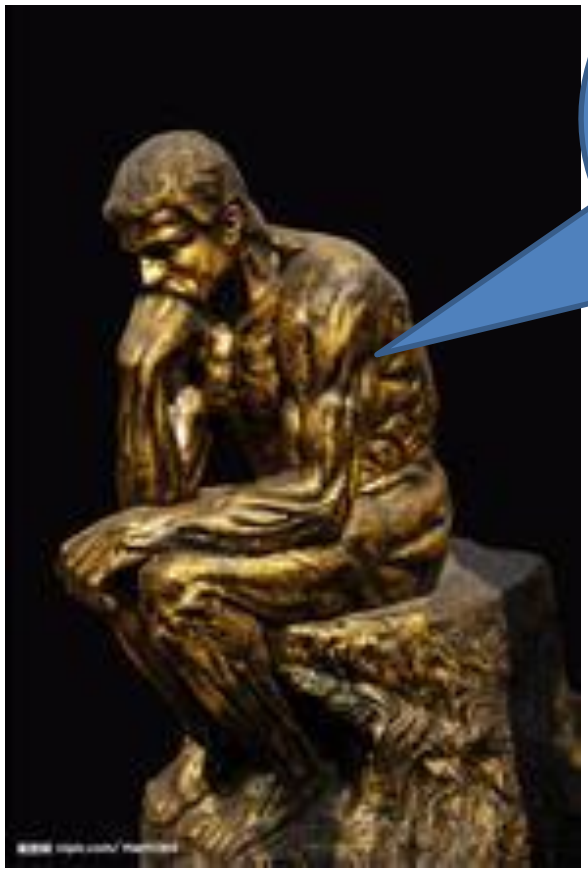
信息检索（Information Retrieval）是指从信息资源的集合中查找所需文献或查找所需文献中包含的信息内容的过程

匹　配　信息检索也是一个匹配过程。　信息检索过程　包括信息处理和检索两个方面

目录

1 起源
2 定义
3 类型
4 主要环节
5 热点
6 检索原因
7 四个要素
8 检索方法
9 同名书籍一
10 同名书籍二
- 图书信息
- 内容简介
- 图书目录
11 同名书籍三
- 图书信息
- 内容简介
- 图书目录
12 同名书籍四
- 图书信息
- 内容简介
- 图书目录
13 同名书籍五
- 图书信息
- 内容简介
- 图书目录

- 接着



信息检索挺有趣的哈！先整件衣服吧，有点冷！

淘宝网

宝贝 店铺

男士西服

全部 搜索

上海：5℃ Powered by 一淘

所有分类 › 该条件下查找 共 30.35万 件宝贝

品牌: 美特斯邦威 邦仕普 七匹狼 杰珂波菲 海澜之家 G2000 雅戈尔 满速 柒牌 杉杉 +多选

花花公子 森马 木吉猫斯 阿玛尼 罗蒙 金利来 圣得西 卡宾 与狼共舞 可可西

风格: 时尚都市 商务绅士 青春流行

上衣尺码: 均码 46 48 50 52 54 160/84(XS) 160/80(XS) 165/88A 165/88B 170/92A +多选

相关分类: 西服套装 西装 西裤 流行男装 拍卖会 男包 服饰配件/... 项链/耳饰/... 淘宝动漫

你是不是想找: 男士休闲西服 | 西服套装男士 | 男士西服外套 | 男士西服上衣 | 男士立领西服 | 西装 | 男士正装 | 休闲西服 | 男士西服相关

所有宝贝 天猫 二手 值得买 1/100

排除关键字 确定 消费者保障 品质承诺 退换货保障 正品保障 旺旺在线 海外商品 货到付款 信用卡 全新

综合 人气 销量 信用 最新 价格 - 所在地 合并卖家

VICUTU威可多 商务西装上装西服套

太平鸟男装 风尚系列 新款正品 西服

2013秋装新款 GXG实体新品 男士休

【反季特卖158包邮】小西服男款修

掌柜热卖

3折秒杀 罗蒙男士西服套装结婚

¥1588.00 免运费

最近成交2471笔 如实描

包邮疯抢 男西服套装 商务休闲

¥360.00 ¥698.00 免运费

最近成交951笔 如实描

- 接着

世纪佳缘 jiayuan.com

搜索会员 我要找：女朋友 年龄：22 至 28 岁 地区：--请选择-- --请选择-- 有照片 搜 索 高级搜索

昨日有9,015会员宣布他们征友成功

中国严肃婚恋交友网站

迄今已成就2,603,051对佳侣

最新注册人数

27,751,458

截止到2010年09月06日24点

马上注册 会员登录

最新会员 | 《我们约会吧》嘉宾 NEW | 《非诚勿扰》 | 真诚榜 | 高端猎婚 | 安全交友 | 最新活动

世纪佳缘创始人-小龙女

创始人-小龙女

“网络第一红娘”，促成的姻缘不计其数，而她本人也成功在世纪佳缘上“秒杀”到一个好老公...

- 中国企业“未来之星”
- 迎向婚恋网站美好未来
- 偶然创业与第一红娘
- 网络红娘与她的理想
- 严肃地做浪漫的事
- 现代红娘，网上佳缘
- 小龙女七夕谈爱情
- 众里寻他千百“度”

什么是信息检索？

- 提问（随机点名系统）

三个应用例子的共同特征

- 给定需求(或者是对象)，从信息库中找出与之最匹配的信息(或对象)
  - Google/百度的例子：需求 “信息检索”
  - 淘宝的例子：对象 “男士西服”
  - 世纪佳缘网的例子：
    - 对象 “女朋友”！

信息检索的一些官方定义

- 给定用户需求，返回满足该需求的信息的一门学科。通常涉及信息的获取、存储、组织和访问。
- 从大规模非结构化数据的集合中找出满足用户信息需求的资料的过程。
  - 非结构化数据通常指文本
  - 什么是结构化数据？（提问）

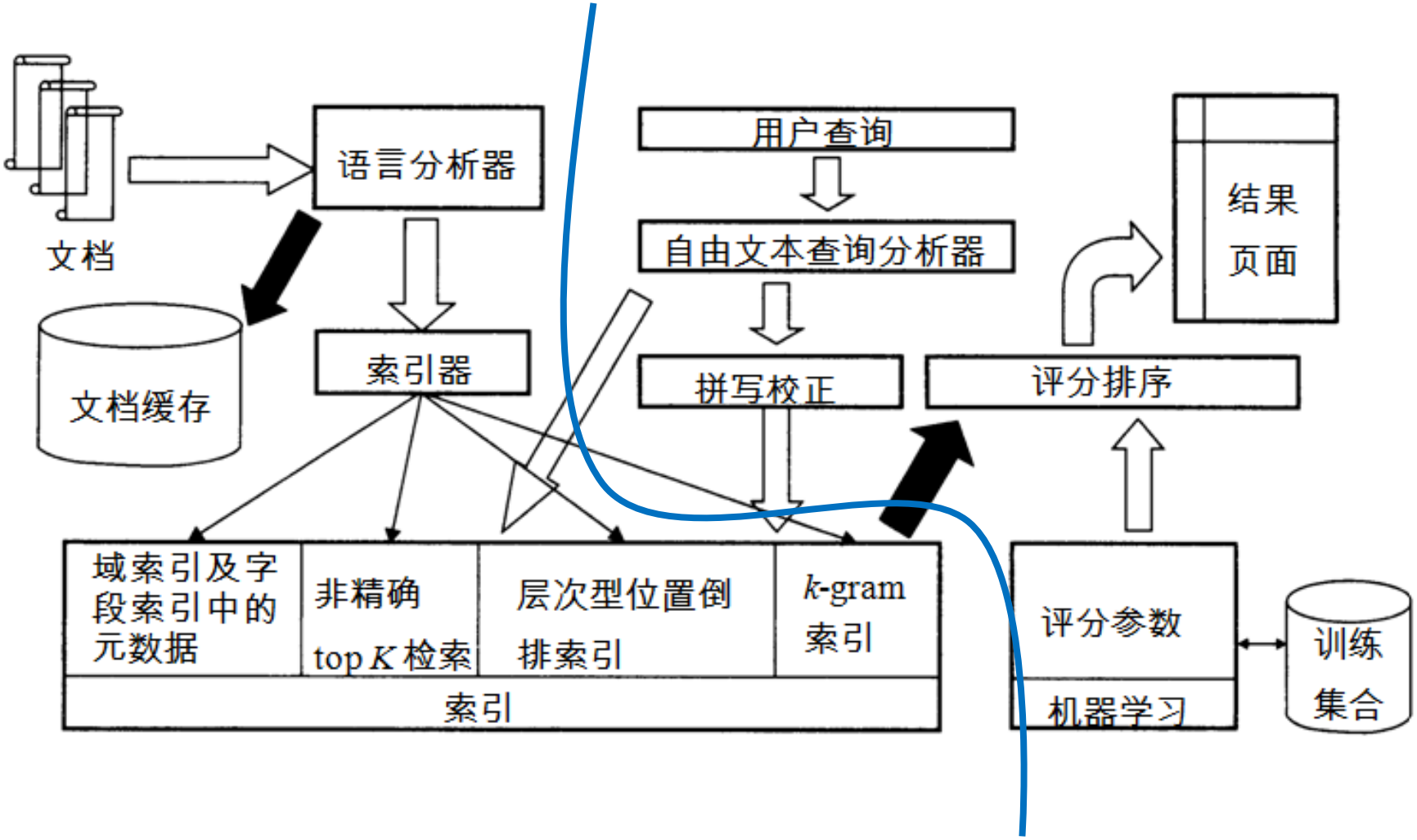
本课程的内容

- 主要关注面向文本数据
- 几部分内容：
  - 爬虫（Crawler），得到网页（wget可以递归爬取静态网页html）
  - 网页正文提取（html文件的处理），得到文档
  - 中文分词
  - 文档存储（倒排；快速查询）
  - 用户query分词
  - 检索：找到相关文档
  - 相关文档简单排序

本课程不涉及内容

- 非文本数据
- Query深度分析和扩展
- 网页链接分析（PageRank）等复杂排序方法
- …

完整的搜索系统示意图


文档
语言分析器
用户查询
自由文本查询分析器
结果
页面
文档缓存
索引器
拼写校正
评分排序
域索引及字
段索引中的
元数据
非精确
top K 检索
层次型位置倒
排索引
k-gram
索引
索引
评分参数
机器学习
训练
集合

信息检索技术的应用

輿情分析

Google 谷歌

Baidu 百度

搜索

amazon.com

Alibaba.com Global trade starts here.™

推荐

IR技术

情报处理

facebook

Tencent 腾讯

挖掘

内容安全

信息检索应用系统

- 搜索系统
  - Web搜索引擎
  - IBM Waston问答系统
  - ……
- 推荐系统
  - 淘宝网
  - 豆瓣网
  - 当当网

从信息规模上分类

- 个人信息检索：个人相关信息的组织、整理、搜索等。桌面搜索(Desktop Search)、个人信息管理(PIM = Personal Information Management)、个人数字记忆(Personal Digital Memory)
- 企业级信息检索：在企业内容文档的组织、管理、搜索等。内容管理(Content Management)
- Web信息检索：在超大规模数据集上的检索。

提纲

- 什么是信息检索？
- **为什么要学习信息检索？**
- 课程情况

直接经济效益-能赚钱啊！

- 世界级牛公司
  - 很多互联网的公司：Google, baidu, … 高市值公司
- 软件工程师
  - 年薪高

市场发展的需求

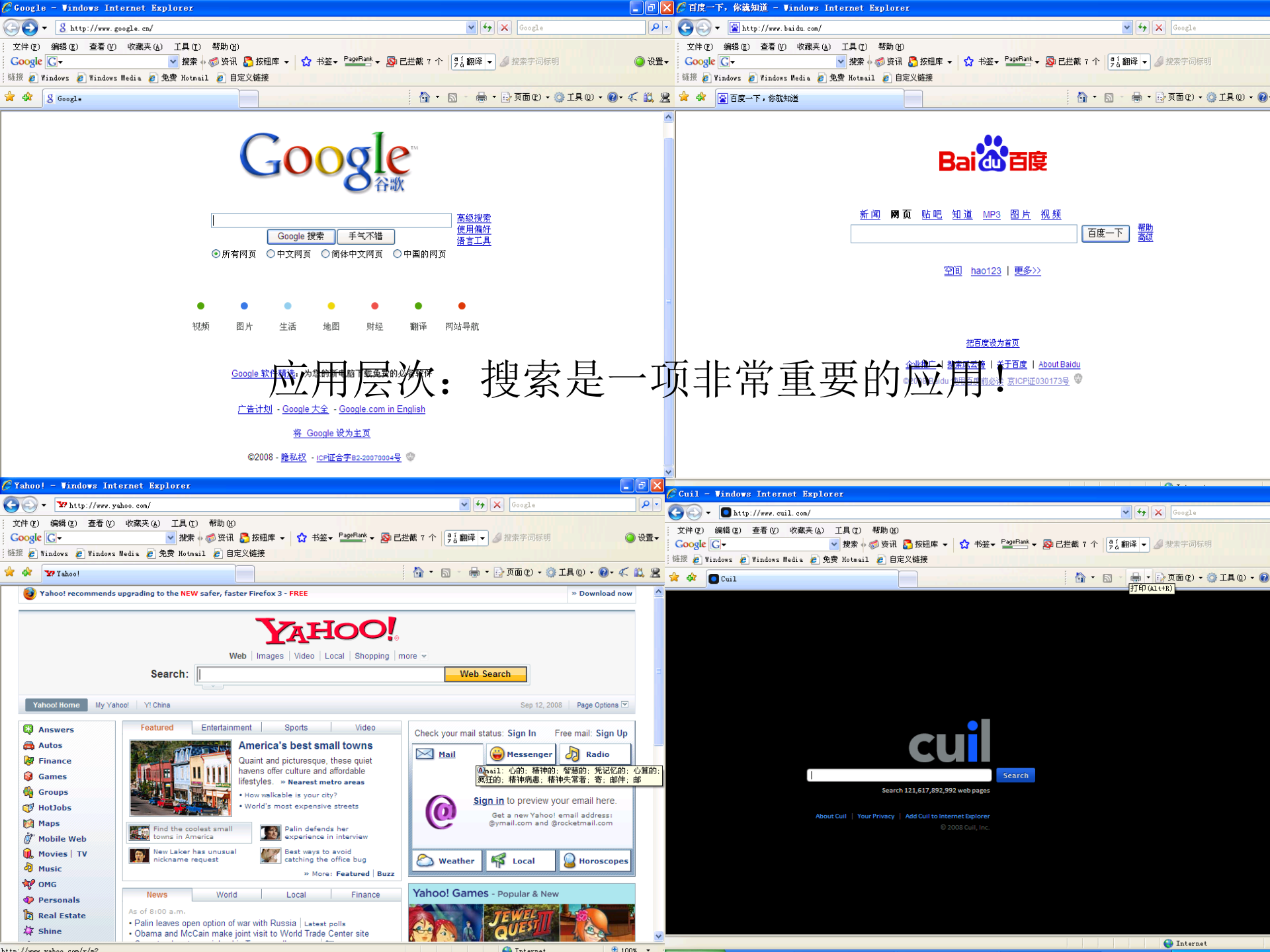
- 用户需要信息检索技术：互联网的信息量太大、噪音太多，寻找所需要的信息非常不容易
- 公司需要信息检索技术：搜索引擎改变了很多传统的生活方式，Yahoo、Google、Baidu，还有一些公司如Microsoft、Sina、Sohu、Tecent、Netease都加入到这个搜索技术的竞争。不只是搜索引擎才需要信息检索技术，电子商务(如亚马逊网站、阿里巴巴)、社交网(微博、Facebook、twitter、校内网)、数字图书馆、大规模数据分析等都需要信息检索技术
  - 人才的竞争：搜索相关人才人数出现缺口，他们非常抢手，待遇如日中天
  - 是不是泡沫：2000年左右出现的网络泡沫和现在的互联网有什么不同，搜索引擎在其中占什么位置？

几个应用需求

- 移动搜索
- 产品搜索
- 专利搜索
- 广告推荐
- 消费行为分析
- 网络评论分析
- SEO营销
- ……

对相关专业学生的基本要求

- 信息检索技术是内容应用特别是互联网内容应用的核心技术，可以说在这些应用中无处不在
- 信息检索将会成为一门计算机专业的基础学科
  - 搜索(狭义的信息检索)的三个层次



应用层次：搜索是一项非常重要的应用！

外媒：没有Google就活不下去的四个网站

http://www.sina.com.cn 2008年09月08日 08:41 IT168.com

导语：NYDailyNews.com周六发表文章，介绍了因谷歌而生的几个网站。它们使用了谷歌的搜索引擎，但风格却大相径庭。

1. elgooG(elgoog.rb-hosting.de/index.cgi)

这是谷歌的一个“真正的镜像”网站，因为一切都是倒过来的，你甚至需要输入字母倒序的搜索关键词。而且，它的“ykcuL gnileeF m'I”，即"I'm Feeling Lucky" 的搜索效果要比普通谷歌的好。

2. Googlefight(www.googlefight.com)

这个网站可以将两个搜索关键词进行“PK”，看哪个的结果数量更多，结果以动画数字的形式显示。一些经典的PK组合包括“God vs. Satan”(上帝对撒旦)、“Luke Skywalker vs. Darth Vader”(天行者对黑武士，即经典影片《星球大战》中的正邪双方代表)以及“pen vs. sword”(笔对剑)。

3. Blackle(www.blackle.com/)

这个网站以几乎全黑的屏幕显示Google。据网站设计者介绍，“Blackle更节能，因为屏幕大部分都是黑色”。

面对节能效果的质疑，该网站也承认其黑屏方式节能有限，不过，该网站仍宣称，“我们认为，每次登入Blackle都是一种提醒，节能要从小处做起”。

4. Goolge Loco(www.googleloco.com)

当你在搜索框中输入关键词时，每个字母都会转换成不同的图标，整个搜索框也会随着图标的变换而改变颜色。不过搜索结果和普通谷歌的一样。(叶西)

中间层次：搜索是极其重要的API

oogle Fight : Make this fight with googleFight Beijing VS Shanghai - Windows Internet Explorer

http://www.googlefight.com/index.php?lang=en_GB&word1=Beijing&word2=Shanghai

Google

google 失败 搜索 资讯 按钮库 书签 PageRank 已拦截 7 个 翻译 搜索字词标明 google

Windows Windows Media 免费 Hotmail 自定义链接

Google Fight : Make this fight with googl...

页面(P) 工具(O)

Googlefight

- The classics
- Funny fights
- Fight of the month
- Last 20 fights

Results on Google :

Beijing	Shanghai
133,000,000 results	121,000,000 results

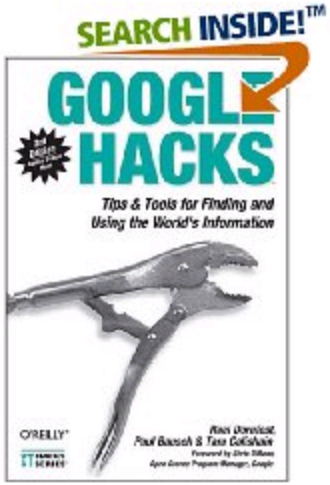
Beijing

Shanghai

Make a fight

Internet 100%

介绍一本有趣的书！



Share your own customer images
Search inside this book

Please tell the publisher:

Google Hacks: Tips & Tools for Finding and Using the World's Information (Hacks) [ILLUSTRATED] (Paperback)

by Rael Dornfest (Author), Paul Bausch (Author), Tara Calishain (Author)

(61 customer reviews)

List Price: ~~$24.99~~

Price: $16.49 & eligible for **FREE Super Saver Shipping** on orders over $25. Details

You Save: $8.50 (34%)

Special Offers Available

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Monday, September 15? Order it in the next 21 hours and 28 minutes, and choose **One-Day Shipping** at checkout. See details

› **41 used & new** available from $3.74

核心层次：搜索是未来操作系统的重要组成部分！

Q&A

- 有什么问题？