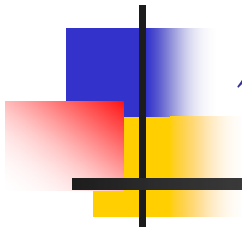
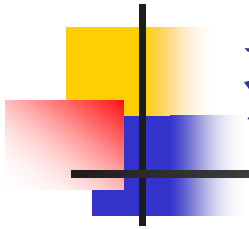


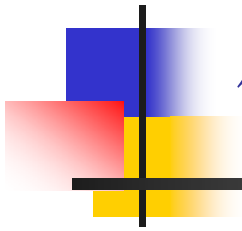
信息检索概述

刘挺
哈工大信息检索研究室
2004.9

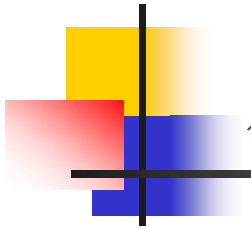


提纲

- 概念
- 体系结构
- 意义
- 历史
- 困难
- 相关领域
- 主要搜索引擎
- 评价
- 信息检索的应用
- 主要研究机构、会议、期刊
- 本课程主要内容

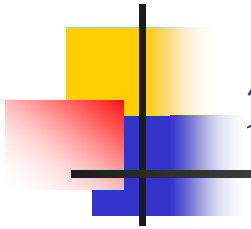


信息检索的概念



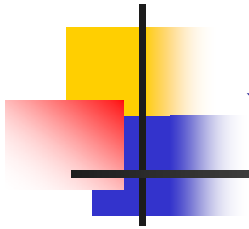
定义

- 信息检索：从非结构化的文档集中找出与用户需求相关的信息
- 和其它相关技术的区别
  - 和数据库的区别
    - 数据库是结构化数据
  - 和情报检索的区别
    - 情报检索介绍如何利用信息检索工具



处理的对象

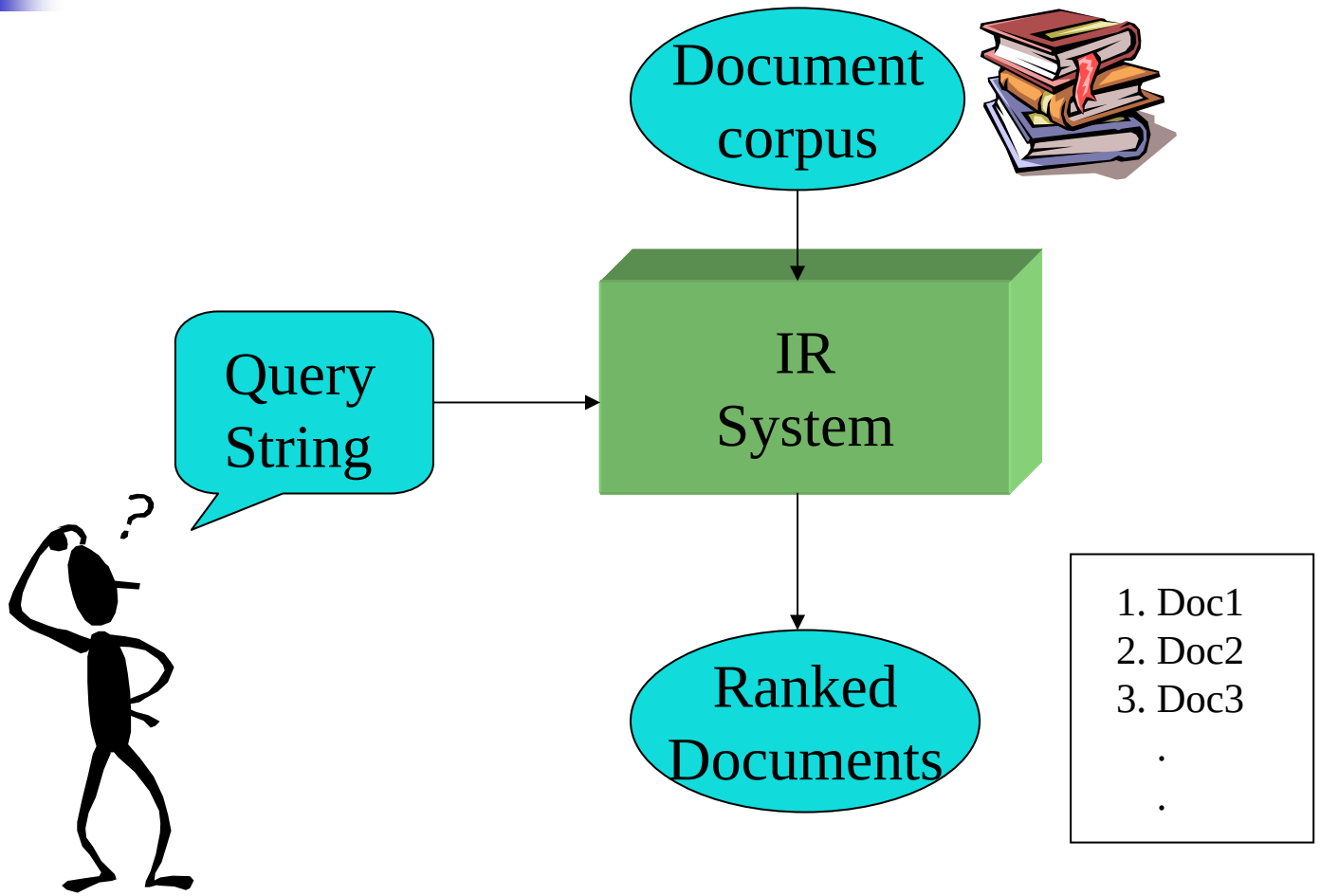
- 非结构化数据
  - 文本数据：新闻、科技论文等
  - 网页：HTML、XML
  - 多媒体数据：图像、视频、图形、音频
- 目前最主要的处理对象是互联网



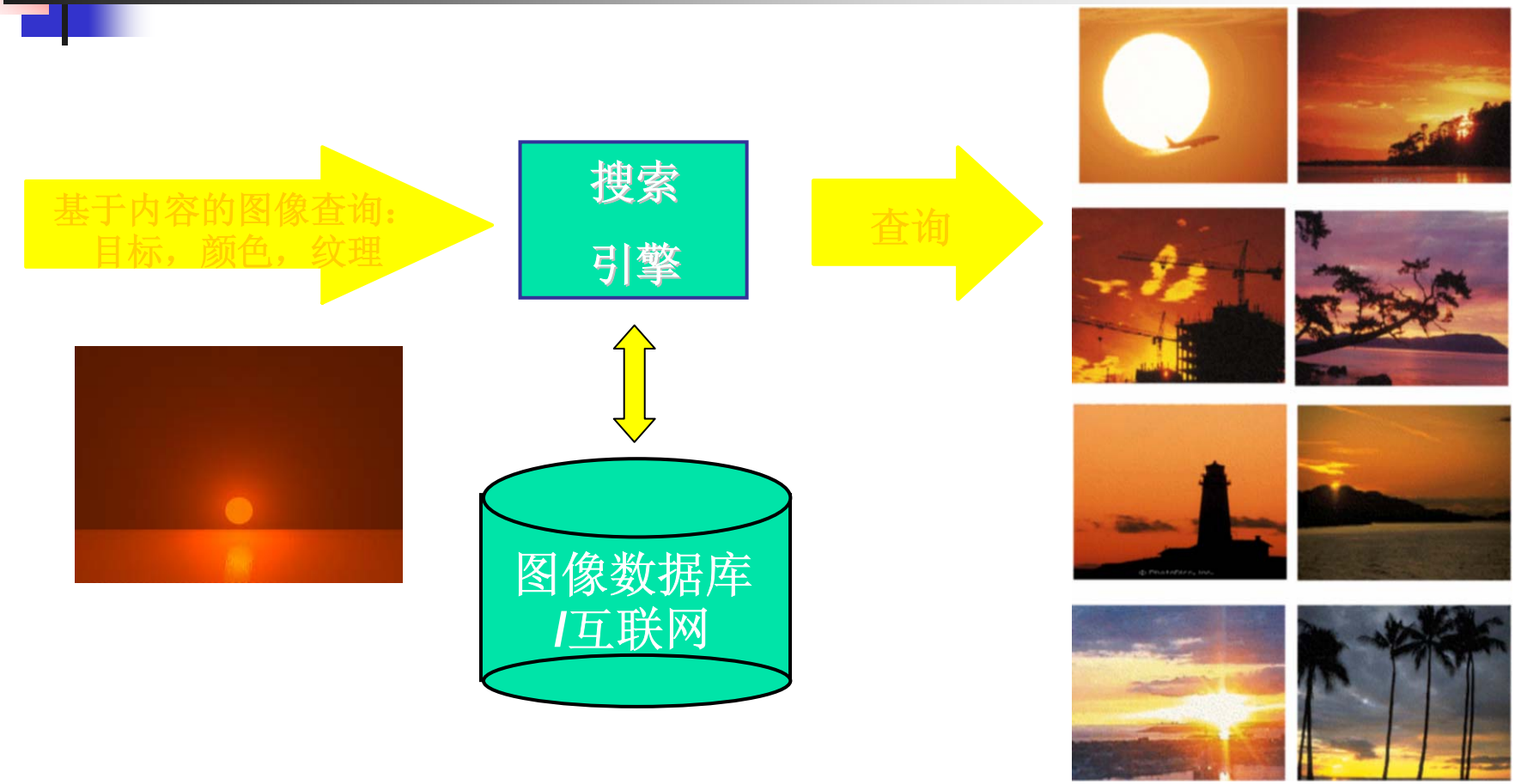
典型的IR任务

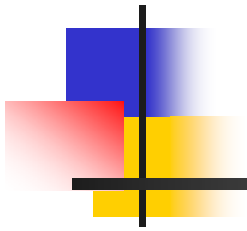
- 给定
  - 自然语言的文档集合
  - 用户的提问(Query)
- 查找
  - 和query相关的经过排序(Rank)的文档子集

IR系统


Document corpus
Query String
IR System
Ranked Documents
1. Doc1
2. Doc2
3. Doc3
.
.

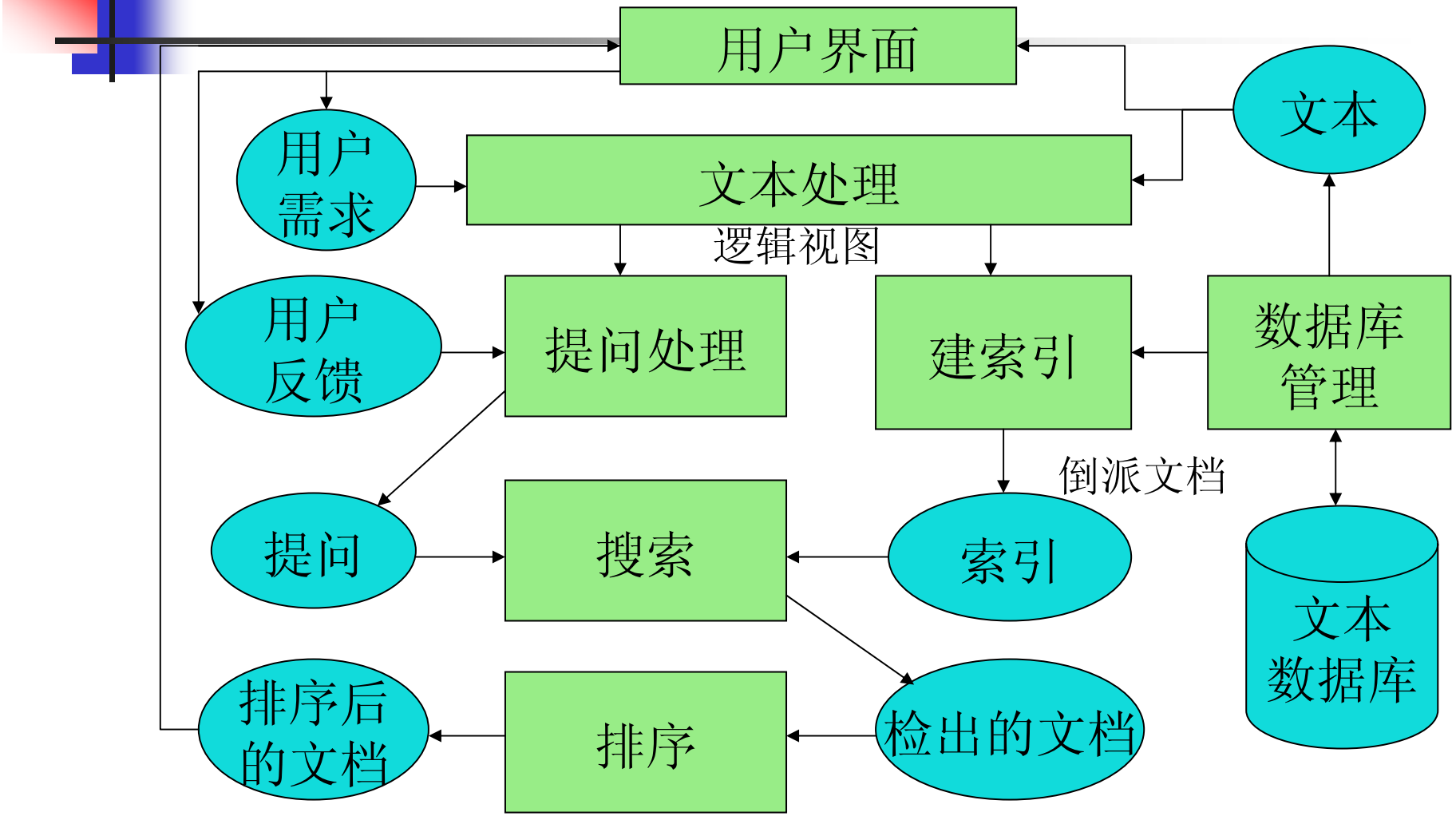
基于内容的图像查询

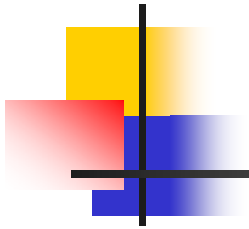




IR系统体系结构

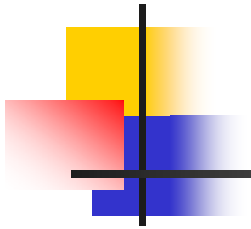
信息检索系统的体系结构





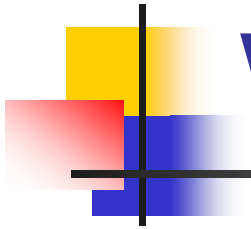
IR系统的组件

- 文本处理形成索引词
  - 删除停用词
  - Stemming（提取词干）
- 建索引
  - 为文档建立倒排索引表
- 搜索
  - 根据倒排索引表检索出与提问相关的文档
- 排序
  - 将检索出的文档根据相关性排序



IR系统的组件

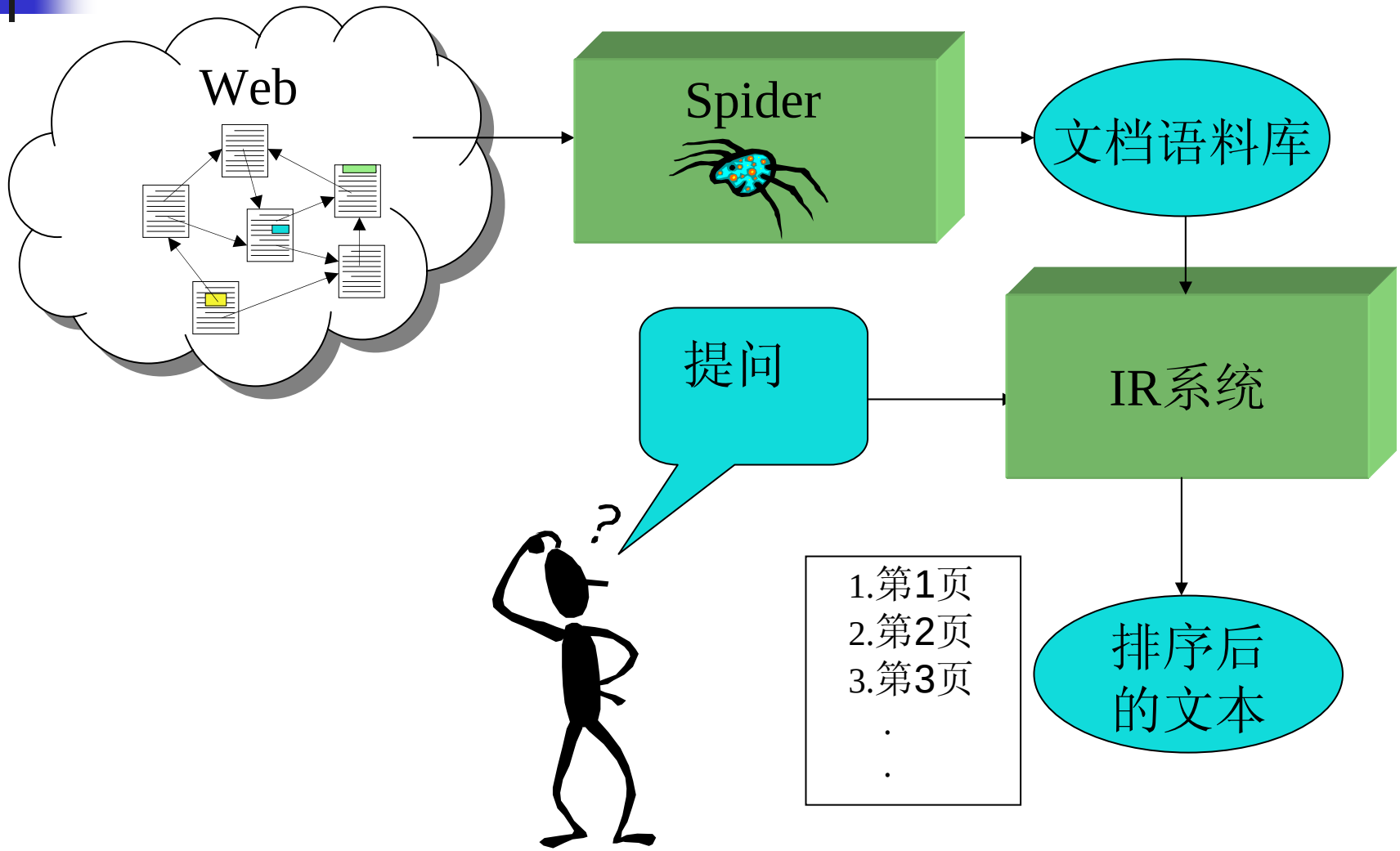
- 用户界面
  - 管理和用户的交互过程，包括：
    - 提问输入和文档输出
    - 相关反馈
    - 结果的可视化
- 提问操作
  - 对提问进行变换，以改进检索结果
    - 根据同义词词典(thesaurus)对提问进行扩展
    - 利用相关反馈对提问进行变换

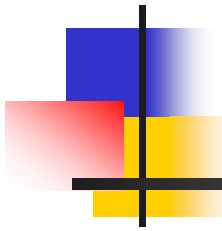


Web搜索

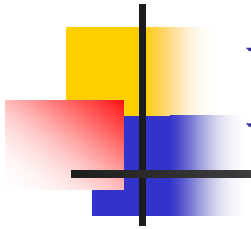
- 将IR技术应用于World Wide Web上的HTML网页
- 和纯文本相比，网页的特点如下：
  - 必须通过在网上“爬行”搜集网页
  - 可以开发结构布局信息
  - 文档的更新是不可控的
  - 可以开发网页之间的链接结构

Web搜索系统



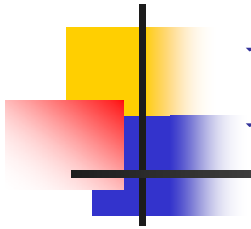


意义



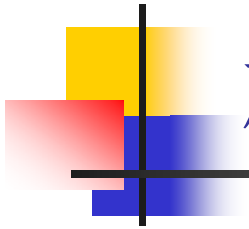
重要性

- 大多数信息都是文本形式的，没有预先定义的格式（例如：邮件、新闻等）
  - 在企业信息化领域，有人统计认为80%的信息是非结构化的
  - 在信息管理向知识管理转变的过程中，文本信息非常关键
- 在非结构化信息中，包括文本信息和多媒体信息，但是文本信息最简洁，最抽象，是人类记载知识的最主要的工具
  - 例如：目前Google的图片检索技术采用的是利用图片周围的文字信息进行的



重要性

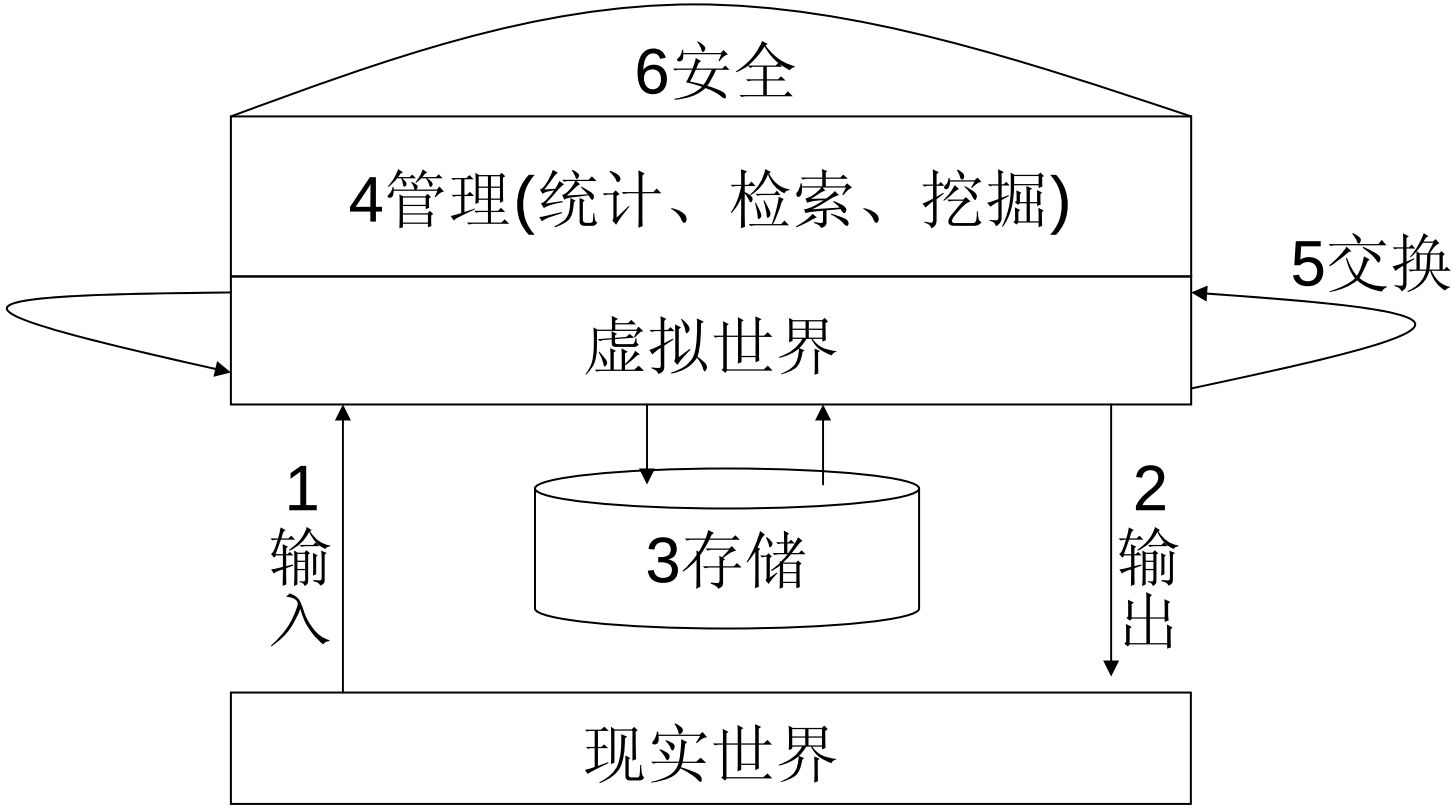
- 传统管理软件需要嵌入IR技术
  - 在SQL数据库中
    - 已采用文本检索技术
    - select * from Employee where Name like '%Lee%'.
  - 在Lotus Notes办公平台上
    - 同样也已采用文本检索技术
- 互联网数据的增长和在线文档（如联机用户手册等）的增长，向IR技术提出迫切需求



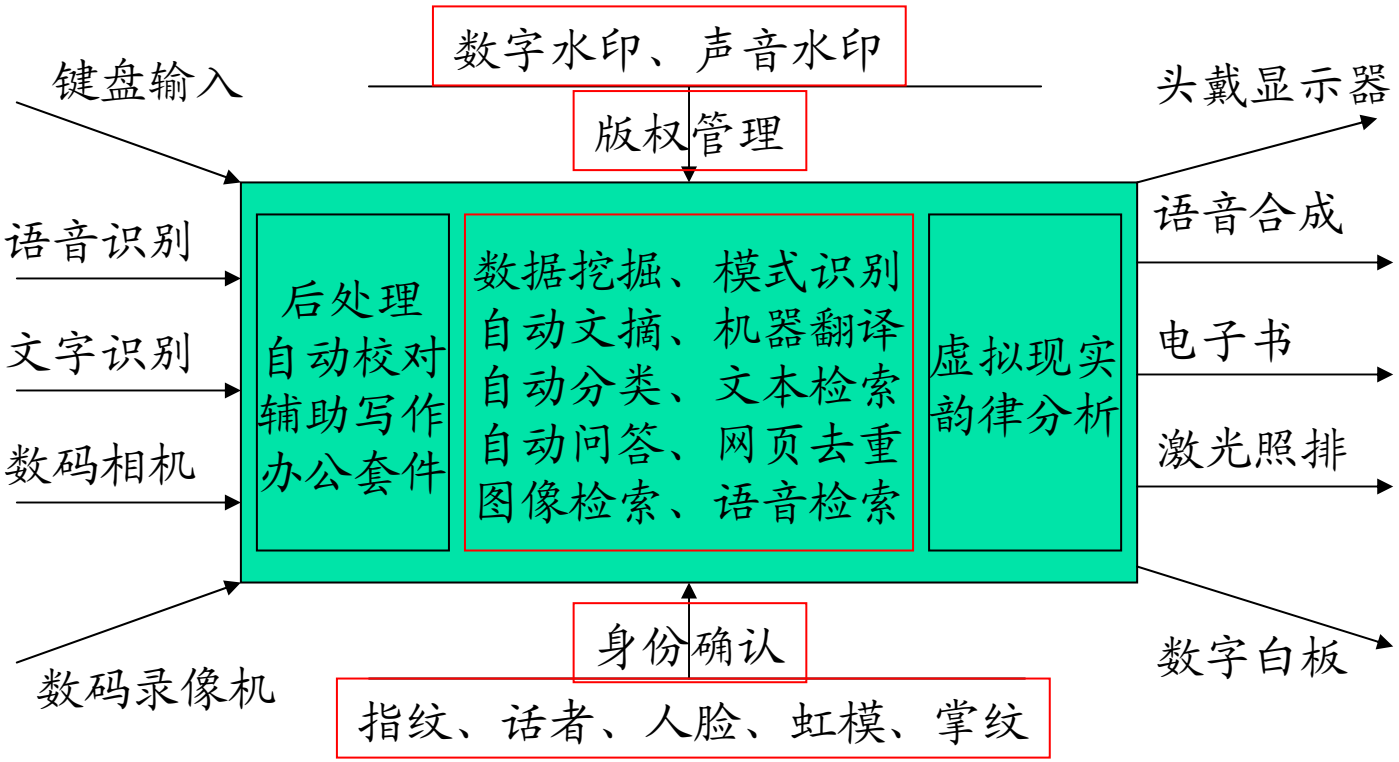
趋势

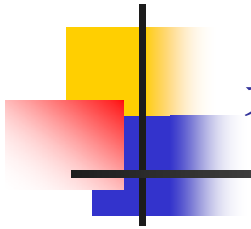
- PC时代
  - Bill Gates:每个人的办公桌上都摆着一台电脑
- 人机交互技术
  - 键盘、鼠标、显示器
  - 中文输入法、激光照排
- 网络时代
  - 网络设备，WWW，互联网浏览器
- 网络安全问题
- 信息的处理和管理问题

虚拟世界和现实世界



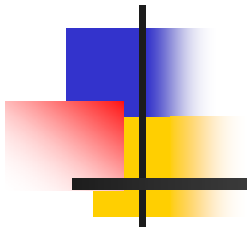
智能计算技术略图


数字水印、声音水印
版权管理
键盘输入
头戴显示器
语音识别
语音合成
文字识别
电子书
数码相机
激光照排
后处理
自动校对
辅助写作
办公套件
数据挖掘、模式识别
自动文摘、机器翻译
自动分类、文本检索
自动问答、网页去重
图像检索、语音检索
虚拟现实
韵律分析
数码录像机
数字白板
身份确认
指纹、话者、人脸、虹模、掌纹

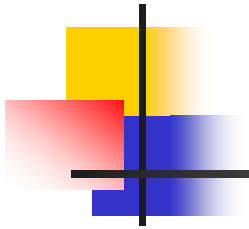


智能计算：从人机交互到内容管理

- 人机交互解决信息录入和呈现的问题
- 在大量信息进入虚拟世界以后，更重要的问题在于如何对这些信息资源进行有效的管理
  - 使用户能够方便快捷地找到想要的信息
  - 使信息保值增值
  - 产生新知
- 竞争不在于拥有多少信息，而在于能够利用多少有价值的信息，因此内容管理至关重要

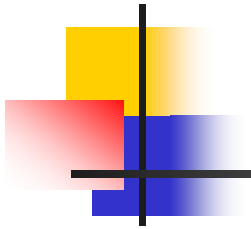


IR的历史



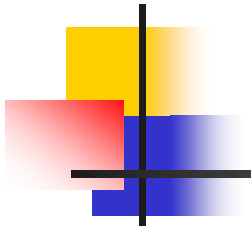
IR的历史

- 1960-70’s:
  - 最初的信息检索系统面向小型的科学文摘数据库、法律和商业文档
  - 检索模型为基本的布尔模型和向量空间模型
  - Cornell University的Prof. Salton和他的学生成为这个领域的先驱



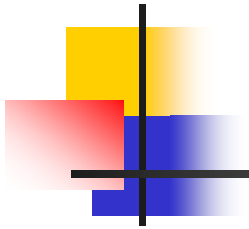
IR历史

- 1980’s:
  - IR技术出现在大型文档数据库中
    - Lexis-Nexis
    - Dialog
    - MEDLINE



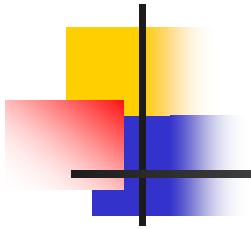
IR历史

- 1990’s:
  - 在互联网上进行对FTP文档进行搜索
    - Archie
    - WAIS
  - 在World Wide Web上进行搜索
    - Lycos
    - Yahoo
    - Altavista



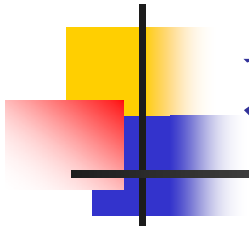
IR 历史

- 1990’s （续）：
  - 有组织地进行评测
    - NIST TREC
  - 智能推荐系统
    - Ringo
    - Amazon（亚马逊网络售书）
    - NetPerceptions
  - 自动文本分类和聚类系统



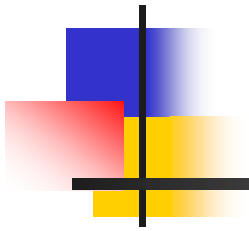
IR历史

- 2000’s
  - 为Web搜索服务的链接分析
    - Google
  - 自动信息抽取
    - Whizbang
    - Fetch
    - Burning Glass
  - 问答系统
    - TREC Q/A track

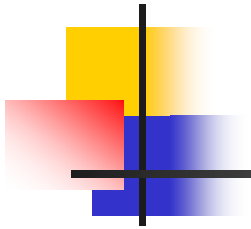


近期的IR

- 2000’s continued:
  - 多媒体IR
    - 图像(Image)
    - 视频(Video)
    - 声音(speech)和音频(Audio)
    - 音乐(music)
  - 跨语言检索Cross-Language IR
    - DARPA Tides项目
  - 自动文摘



IR的困难



国际互联网发展趋势

- 1995-200
  - 1995年11月，50M网页
  - 1997年12月，320M网页
  - 1999年2月，800M网页
  - 2000年，1G网页
  - 2002年，3G网页
- 大量的数据向IR技术提出挑战
  - 以前认为几百兆的数据就是大数据集，现在一个单独的数据库就能够处理10-50G以上的数据

国内互联网

- 中国互联网络信息中心，截至2004年6月30日的统计报告显示
  - 代表互联网发展规模的网民数量半年时间内激增750万，达到8700万人，同上一次调查（2003年12月31日）相比增长9.4%，和去年同期相比增长27.9%。
  - 上网计算机总数3630万，CN域名总数为382216个，WWW站点总数为：约626600个
  - 整个互联网行业呈现发展和复苏，8月3日网易公布的第二季度净收入达2.07亿元人民币，分别较上一季度的1.97亿人民币和去年同期的1.36亿人民币总收入增长5.2%和51.8%。

中国上网人数增长趋势

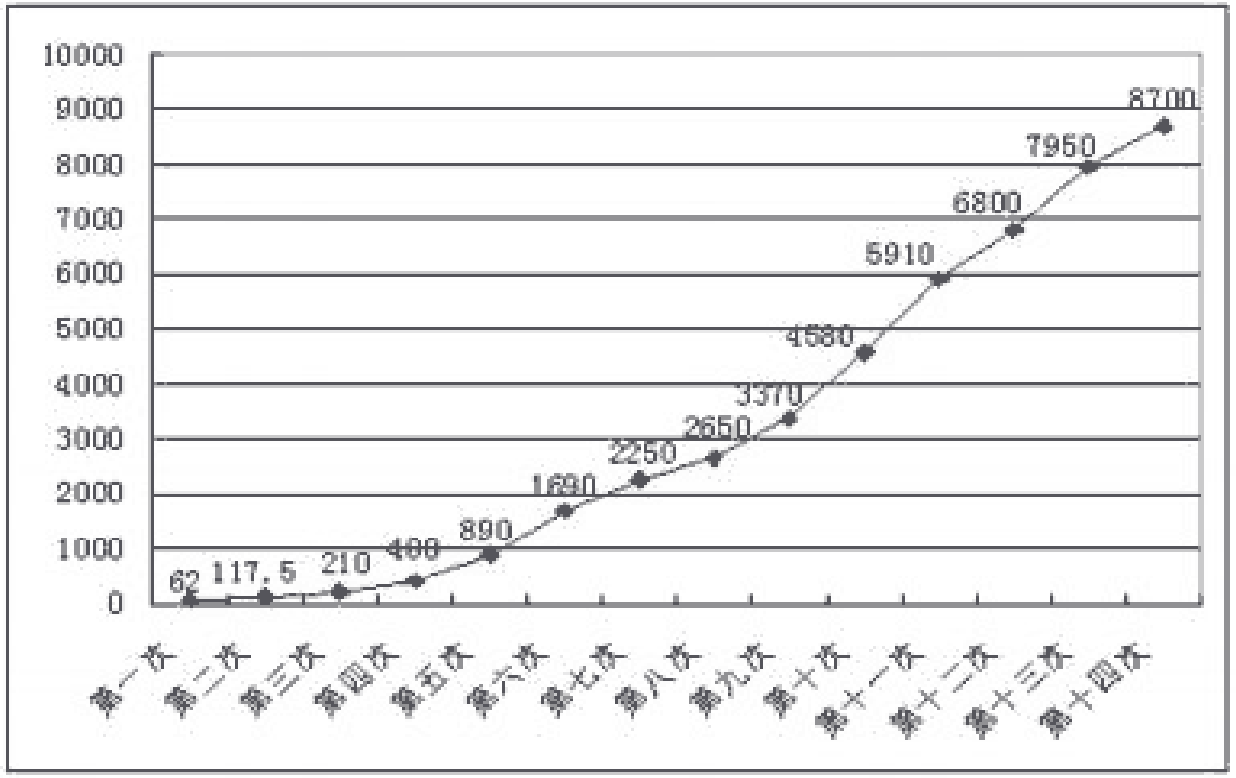


图1-4 历次调查上网用户总数（万人）

中国上网计算机增长趋势

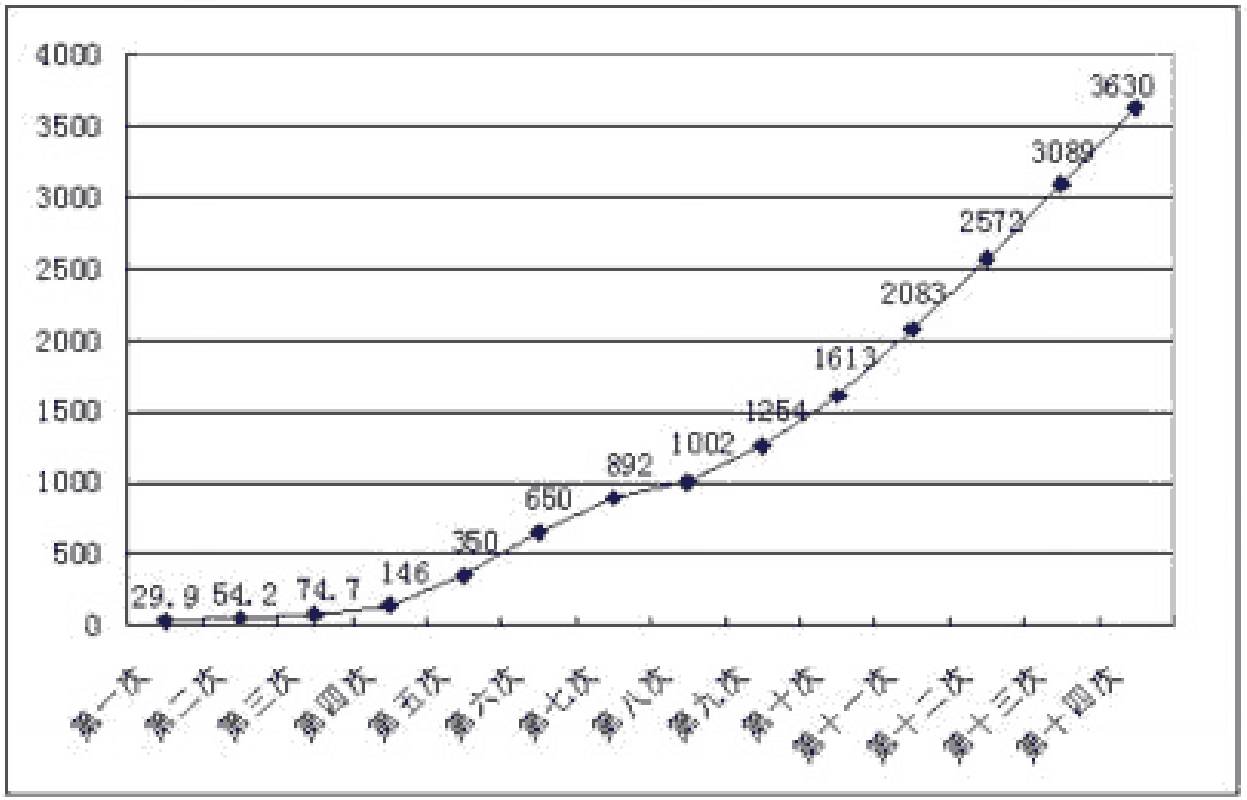
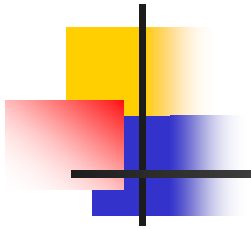
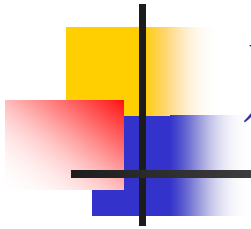


图1-1历次调查上网计算机总数（万台）



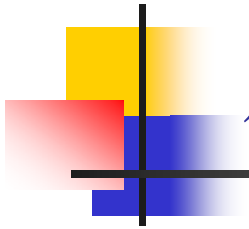
网页数量和字节数

- 来源：TOM科技，2004年3月31日
- 网页数
  - 全国网页总数 311,864,590个
    - 其中：静态网页数 226,725,557个
    - 动态网页数 85,139,033个
    - 静动态网页数比例 2.66:1
  - 平均每个网站的网页数 523.7个
- 网页字节数
  - 全国网页总字节数 6,059,431,526KB
  - 每个网页平均字节数 19.43KB
  - 平均每个网站的网页字节数 10,174.51KB



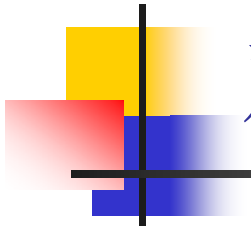
难点所在

- 很难获取非结构化文本的语义信息
  - “select * from Employee where Salary > 100,000”
  - “找出所有关于公司购并的新闻”
  - “找出所有和互联网公司购并相关的新闻”
  - 上述三个问题，一个比一个难
- 检索是在非受限域(unrestricted domains)文档集上进行的
  - 很难对文档的类别事先定义或分类



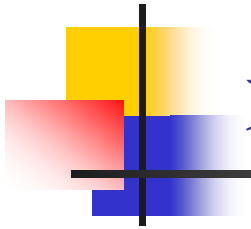
难点所在（续）

- 不同的用户基础
  - 从专家级用户到初级用户
  - 一个系统可能对于专家来说太简单了，而对于初级用户来说又太复杂了
  - 一个系统返回的信息，对于专家来说来泛泛了，但是对于初级用户来说太专业了
- 提问的意图、文档的意图均很难捕获



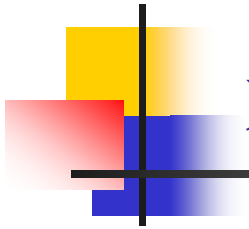
难点所在（续）

- 网页是分布式的和相互连接的
  - 从什么地方开始搜索？这个单一数据库是不同的
  - 信息是如何相互关联的？
- 效率(efficiency)和效果(effectiveness)
  - 在有限的资源内，只能把效率和效果提高到有限的水平
  - 并且，提高效率常常损失效果，反之亦然



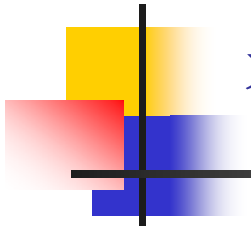
关键词搜索

- 最简单的概念就是关键词在文档中逐字出现
- 稍微严格一点的定义是：提问中的关键词在文档中频繁出现，并且不考虑顺序



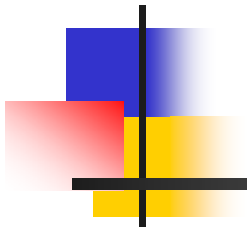
关键词的问题

- 可能找不到同义词
  - “PRC” vs. “China”
  - “电脑”vs. “计算机”
- 可能检索到一些不相关的多义词
  - “bat” (baseball vs. mammal)
  - “Apple” (company vs. fruit)
  - 保安（地名 vs. 保护安全的人员）

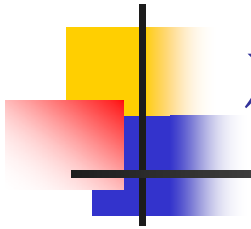


智能信息检索

- 考虑词汇的意义(meaning)
- 考虑词汇的顺序(order)
- 根据直接或间接的反馈适应用户的需求
- 考虑信息来源的权威性(authority)

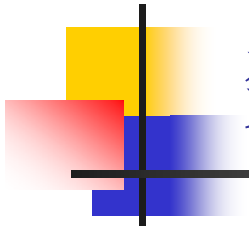


IR相关领域



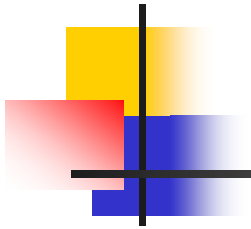
相关领域

- 数据库管理
- 图书和情报科学
- 人工智能
- 自然语言处理
- 机器学习



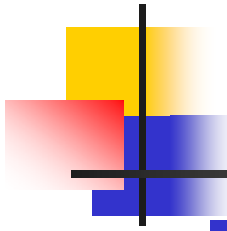
数据库管理

- 专注于研究结构化数据，比如关系表，而不是自由文本
- 专注于处理定义好了的查询式，如SQL
- 查询式和数据的语义都非常清晰
- 近来有向半结构化数据(XML)发展的趋势，和IR越来越接近



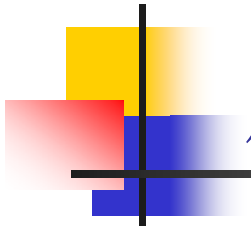
图书馆和情报科学

- 研究信息检索中和人类使用者相关的内容（人机交互、可视化）
- 关心对人类知识的有效分类
- 关心引用(citation)分析和文献计量学(bibliometrics)信息的结构化
- 最近的数字图书馆研究使它和IR距离更近



人工智能

- 研究知识表示、推理和智能行为
- 知识和查询式的形式化:
  - 一阶谓词逻辑
  - 贝叶斯网络
- 最近在Web本体论(Ontology)和智能信息代理(Intelligent Information Agents)使它与IR更接近

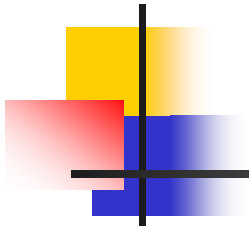


从数据到知识

- 数据(Data)
  - 未经组织的数字、词语、声音、图像等
- 信息(Information)
  - 以有意义的形式加以排列和处理的数据
- 知识(Knowledge)
  - 用于生产的信息（有意义的信息）
  - 信息经过加工处理、应用于生产，才能转变成知识
- 智慧(Wisdom)
  - 应用知识的能力，创新能力

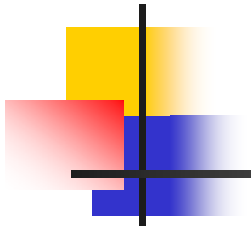
陆汝钤院士——非规范知识处理

- 由于实际应用的迫切需要，计算机科学的研究发生了许多重大的变化。
  - 人们从注重研究对象的形式（form）转向研究对象的内容（content）
  - 从注重研究良构问题（well-structured problem）转向研究病构问题（ill-structured problems）
  - 从注重研究封闭性世界转向研究开放性世界
  - 从研究内涵完整、协调和精确的问题转向研究内涵不完整、不协调和不精确的问题
  - 这些趋势在知识处理的研究中表现为一个过去研究得较少的、十分困难的课题，即非规范知识处理。
- 非规范知识处理的最典型应用领域是因特网上知识的处理
  - 因特网上的知识大部分是非结构或半结构的，它们以各种媒体形式存在，以自然语言为载体，分布在几亿个网页上，每天以百万网页的数量级在增长、消失或改变内容，它充满了各种矛盾的事实、数据和观点，几乎体现了非规范知识的所有特点。
  - 可是，因特网的快速发展与广泛应用要求在开放、动态环境下实现灵活的、可信的、协同的、深层次的知识共享和利用。这个目标的实现在很大程度上依赖于非规范知识处理技术的进步。
  - 目前，任一单一模型的使用效果均有局限，各种模型所得结果的综合是一大问题。这又提出了各种非规范知识的融合问题。正是在这样的背景下，需要系统而深入地开展非规范知识处理的基本理论和核心技术的研究。



自然语言处理

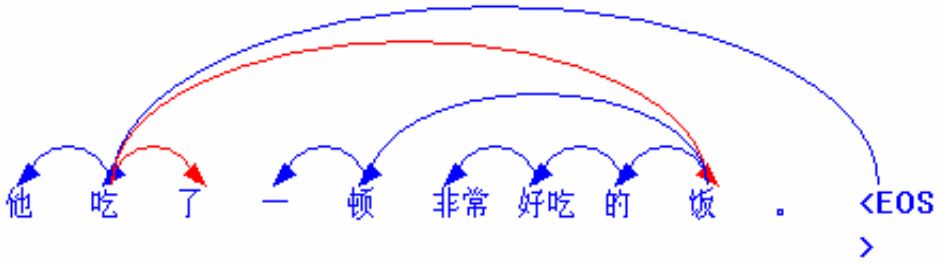
- 研究自然语言文本的句法、语义和语用
- 使检索能够在意义层面而不是仅仅在关键词层面进行

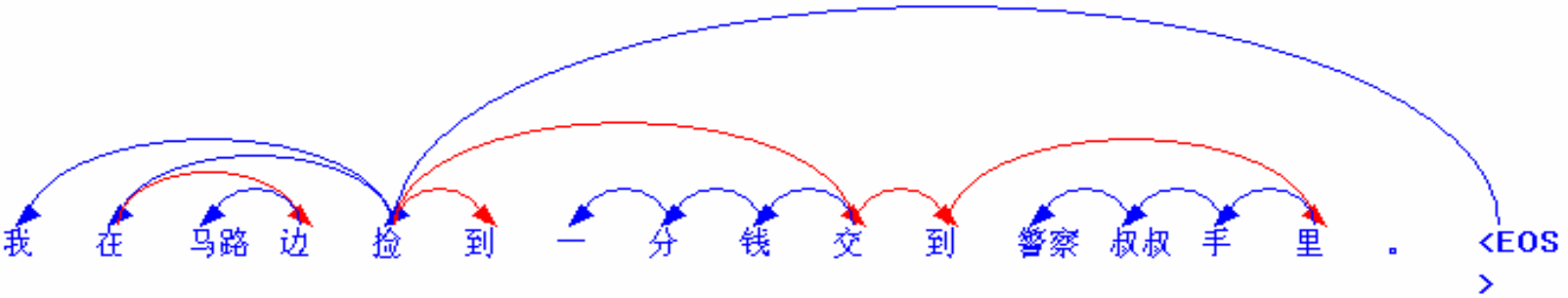


自然语言处理：IR的方向

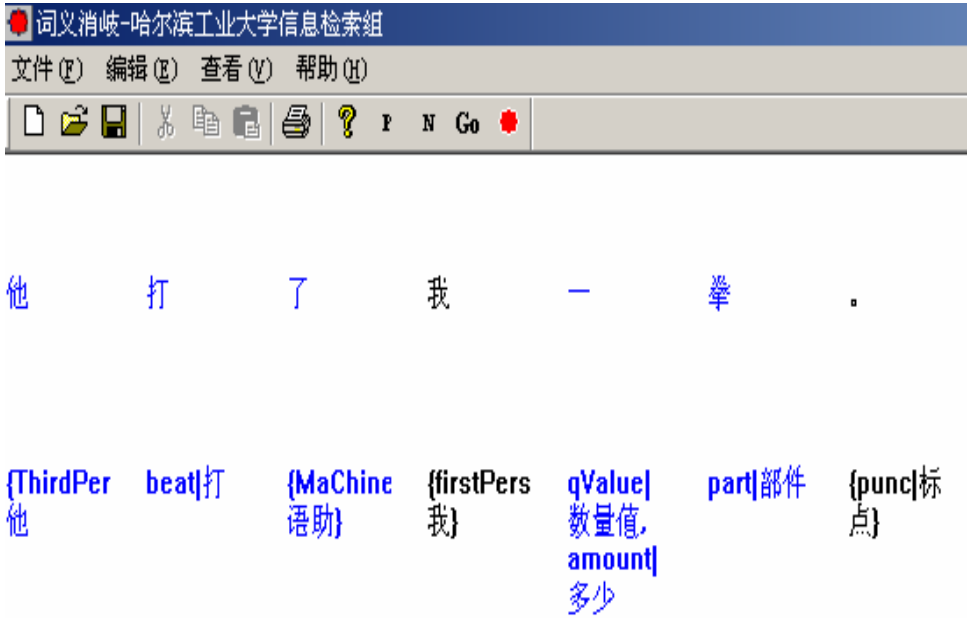
- 根据上下文决定 歧义词的意义：词义消歧(*word sense disambiguation*).
- 识别文本中特殊的信息片断 (*information extraction*).
- 从文本中回答特殊的用自然语言提出的问题

依存文法分析举例

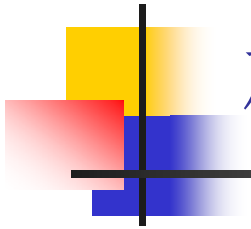




词义消歧

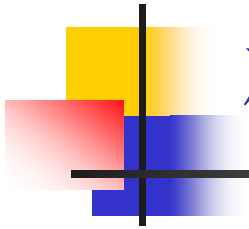


All Not
NounUnit|名量
TakeOutOfWater|捞起
associate|交往
beat|打
build|建造
buy|买
calculate|计算
catch|捉住
compile|编辑
damage|损害
dig|挖掘
draw|画
engage|从事
exercise|锻练
fight|争斗
gather|采集
lift|提升
mix|混合
produce|制造
remove|消除
send|发送
spray|洒下
use|利用
weave|辫编
wrap|包扎
write|写
{LocationIni}
{TimeIni}



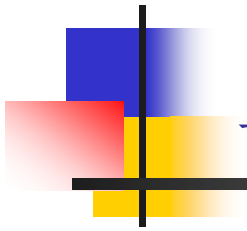
机器学习

- 研究能够通过经验改进自身性能的计算系统
- 有指导的学习(*supervised learning*)
  - 通过从人工标注好的训练样例中学习概念来实现对样本的自动分类
- 无指导的学习(*unsupervised learning*)
  - 事先不经过的人工标注，将样本自动聚为有意义的组



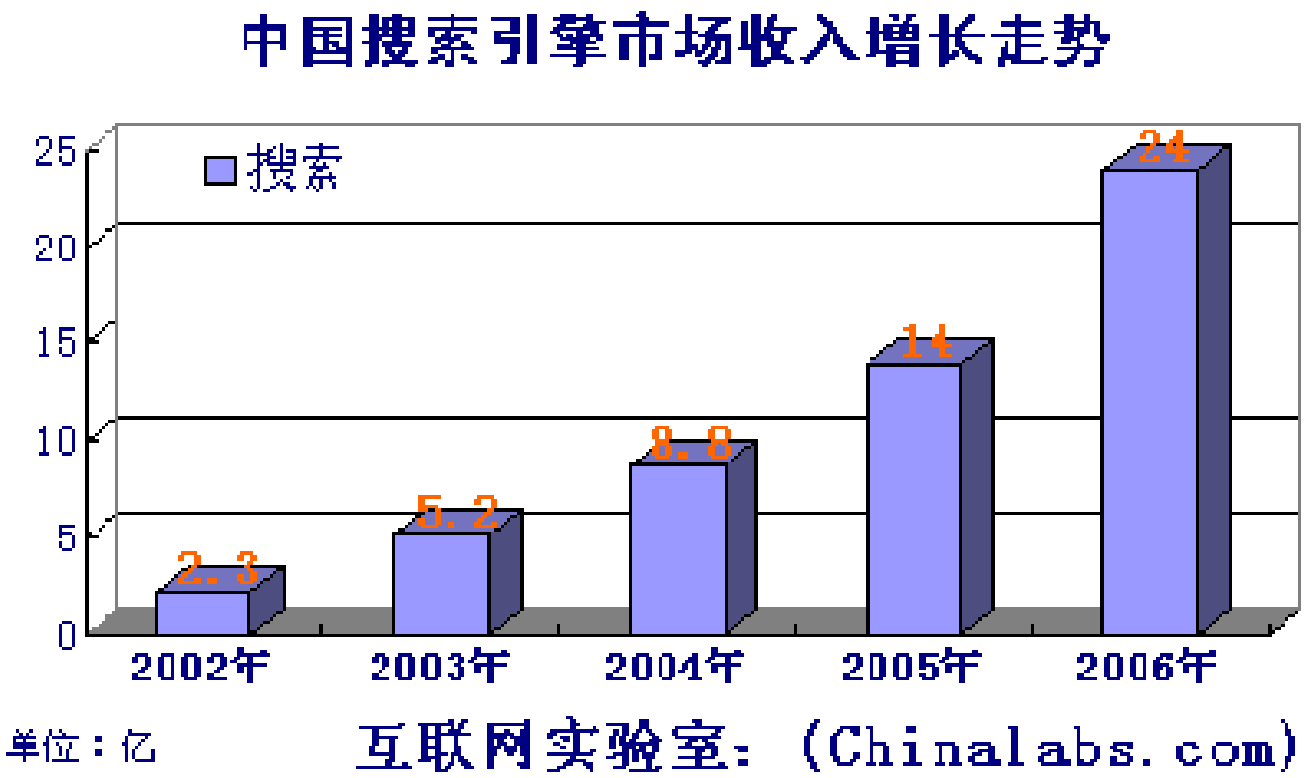
机器学习：IR的方向

- 文本分类
  - 自动层次聚类(Yahoo)
  - 自适应/推送/推荐
  - 垃圾邮件过滤
- 文本聚类
  - 检索结果的自动聚类
  - 自动形成层次体系
- 信息抽取
- 文本挖掘



主要搜索引擎

中国搜索引擎市场


中国搜索引擎市场收入增长走势
搜索
25
20
15
10
5
0
2.3
5.2
8.8
14
24
2002年
2003年
2004年
2005年
2006年
单位：亿
互联网实验室：（Chinalabs.com）

关于搜索引擎的新闻

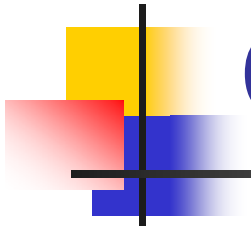
- 2003年底以前，中国搜索引擎市场的格局是：雅虎和Google都提供中文搜索服务，但没有正式进入中国。中国本土的搜索引擎服务商主要是百度、3721、中国搜索(慧聪搜索)。然而，这一切在2004年发生了彻底的变化。
- 2003年11月21日，雅虎中国收购3721公司。3721的搜索服务成为了YHAOO中国的重要组成，YHAOO正式进军中国搜索引擎服务市场。
- 2004年6月15日，Google与其他七家共同投资者一起，收购了有全球最大中文搜索引擎之称的百度的部分股份。Google在上市前终于有了中国搜索的概念。
- 2004年6月21日，雅虎中国除了坚固其门户搜索、3721之外，推出了专门的中文搜索门户网站“一搜(www.yisou.com)”。
- 2004年7月1日，微软公司董事长比尔·盖茨在北京含蓄地表示，要加强MSN搜索开拓中国市场的力度。

全球最大搜索引擎——Google

- 据IDC的预计，全球搜索市场3到5年后将达70亿美元以上，Google在各种搜索引擎中排名第一。
- Google网址：www.google.com
- 技术创业
  - 六年成长史
  - 创始人是两位斯坦福大学学生，而立之年即成为百亿富翁
  - 每个月有数亿人使用
- 走向垄断？
  - 参股百度
  - 左右网民的价值取向
  - 受商业利益驱使，未来很难保证客观公正性

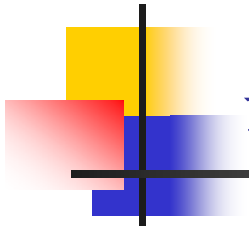
Google


Google搜索: Who is the director of IRLab - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
地址(D) http://www.google.com/search?hl=zh-CN&ie=UTF-8&oe=UTF-8&q=Who+is+the+director+of+IRLab&lr=
高级搜索 使用偏好 语言工具 搜索建议
Who is the director of IRLab
Google搜索
搜索所有网站 搜索所有中文网页 搜索简体中文网页
以下字词太常用，因此未列入搜索范围： Who is the of。 [详情]
所有网站 图像 网上论坛 网页目录
已向英特网搜索 Who is the director of IRLab。
您要不要 只看 中文(简体) 的搜索结果 ？
Information Retrieval Laboratory
... Information Retrieval Lab of HIT was established on March 1, 2001. Prof. Li Sheng is the director of IRLab and Dr. Liu is the vice director. Now IR Lab consists of 4 stuffs(1 professor, 3 associate professor), 6 doctor candidates, 9 master candidates and 5 undergraduates. ...
ir.hit.edu.cn/en/ - 7k - 网页快照 - 类似网页
IF JunGo: Annual Report 1998
... The new leadership with Professor Gustav Andreas Tammann, Astronomisches Institut of the University of Basel, as new President of the Foundation HFSJG, and Professor Erwin Flückiger, Physikalisches Institut of the University of Bern, as new Director of the research ... astro.it/irlab/tirgo/index.html ...
www.ifjungo.ch/flueckiger.htm - 30k - 网页快照 - 类似网页



Google排名

- 输入：信息检索
  - 排名第十（最高时排名第七）
  - 排名1-9的网址中，除了“云网”外，均非技术类网址
- 输入：信息检索研究室
  - 排名第一
- 输入：IRLab
  - 排名第一
- 输入：刘挺
  - 排名第一
- 输入：”Ting Liu”
  - 排名第一

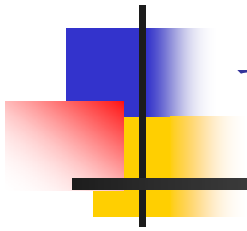


其它主要英文搜索引擎

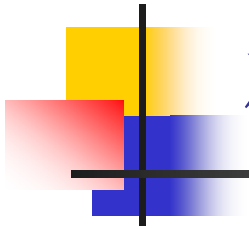
- AOL search.aol.com
- AltaVista www.altavista.com
- AskJeeves www.askjeeves.com
- MSN Search search.msn.com
- LookSmart www.looksmart.com
- Yahoo www.yahoo.com

中文搜索引擎

- 中国网民中至少有79%会经常使用搜索引擎，98%会使用搜索引擎。
- 百度
  - 百度网址：www.baidu.com
  - 北大计算机系学生创办
  - 北京大学李晓明教授继续研究“天网”
  - 天网　pccms.pku.edu.cn
- 中搜
  - http://www.chinasearch.com.cn/
  - 全名“中国搜索”，原名“慧聪”
- 搜狗
  - http://www.sogou.com

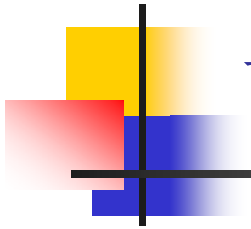


评价



相关性

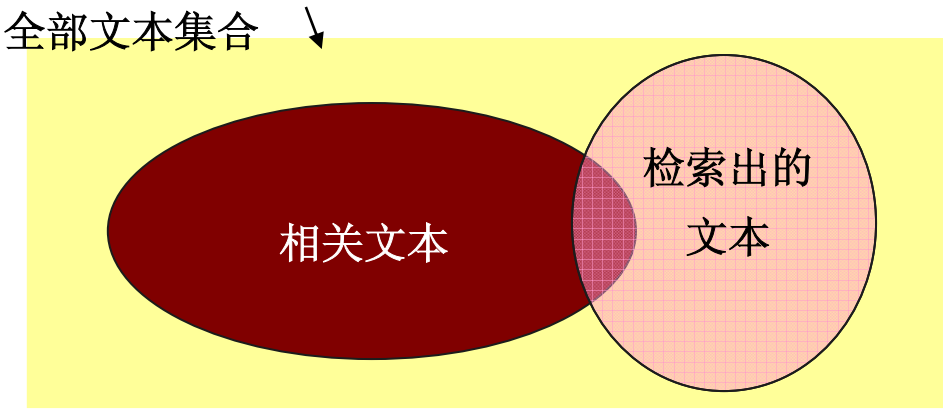
- 相关性是一种主观评价
  - 是不是正确的主题
    - 输入：“和服”；输出：“……咨询和服务……”
    - 由于分词错误，导致检索结果偏离主题
  - 是否满足用户特定的信息需求 (*information need*)
  - 时效性，是不是新的信息
    - 输入：“美国总统是谁”；输出：“克林顿”
    - 信息已经过时
  - 权威性，是否来自可靠的信息源



评价IR系统的困难

- 相关性不是二值评价，而是一个连续的量
- 即使进行二值评价，很多时候也很难
- 从人的立场上看，相关性是：
  - 主观的，依赖于特定用户的判断
  - 和情景相关的，依赖于用户的需求
  - 认知的，依赖于人的认知和行为能力
  - 时变的，随着时间而变化

准确率和召回率

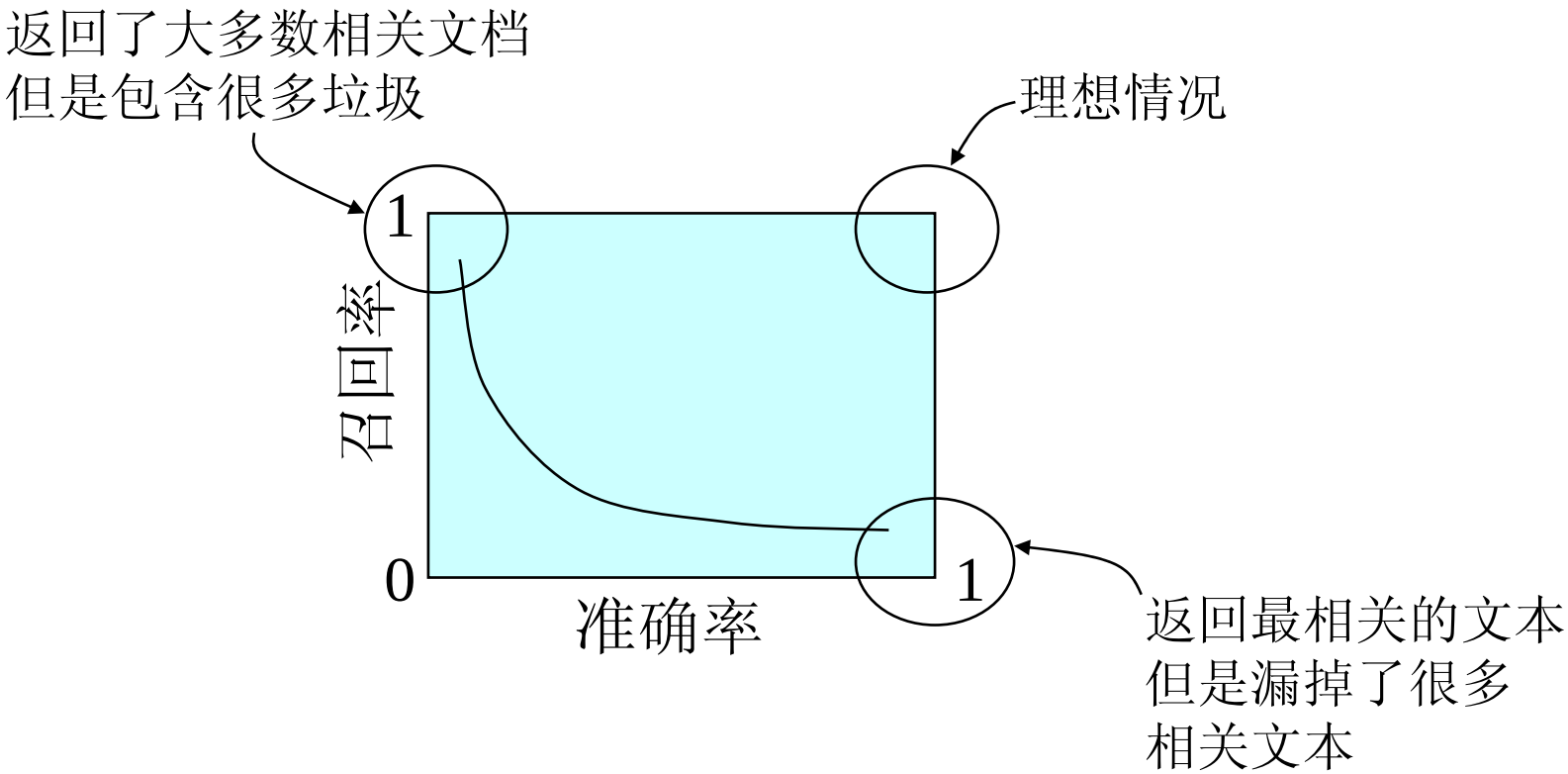


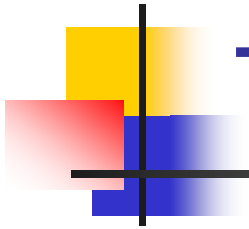
	检出	未检出
不相关	检出且 不相关	未检出且 不相关
相关	检出且 相关	未检出且 相关

召回率=检出的相关文档数/相关文档数

准确率=检出的相关文档数/检出文档数

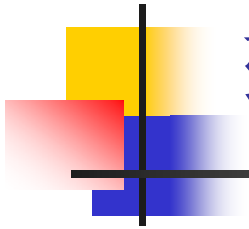
准确率和召回率的关系


返回了大多数相关文档
但是包含很多垃圾
理想情况
1
召回率
0
1
准确率
返回最相关的文本
但是漏掉了很多
相关文本



TREC评测(Benchmark)

- TREC: Text REtrieval Conference (http://trec.nist.gov/)
  - 1992年开始，每年一次
  - 由美国国防部Defense Advanced Research Projects Agency (DARPA)和美国国家标准技术协会National Institute of Standards and Technology (NIST)联合发起
  - 参加者免费获得标准训练和测试数据
  - 参加者在参加比赛时收到最新的测试数据，并在限定时间内作出答案，返给组织者
  - 组织者对各参赛者的结果进行评价
  - 包括检索、过滤、问答等多个主题



输入“和服”

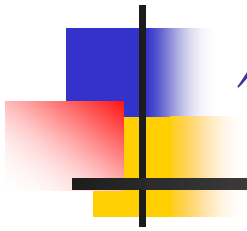
- Google的检索结果
  - 排在前30位的网页绝大多数为日本的“和服”，说明Google进行了有效的分词
- 搜狗
  - 基本正确

百度的检索结果

- [4]青岛东和服装设备
  进入新产品! ENGLISH 地址:青岛市延安一路31号(京剧团对面)·····
- [10]流动人口计划生育管理和服务工作若干规定
  流动人口计划生育管理和服务工作若干规定 第九号部长令·····
- 基本可以，存在错误

中国搜索

- [1]重庆“侦探”商标注册成功 邦德公司获工商认可
  …册范围，将原42类商品和服务商标注册扩大…
- [2]新潮实业:“亚麻”龙头 箭在弦上
  由于所有纺织品和服装配额都将于今年底以前完全取消，近期4元左右的低价纺织股表
- 分词效果不佳！



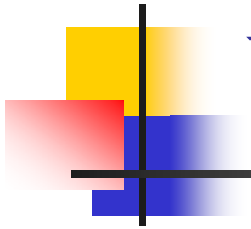
信息检索的应用

数字图书馆

- 自动分类
  - 根据国图分类法，对文本进行自动分类
- 自动标引
  - 自动给出文本的主题词，包括抽词标引和赋词标引两种
- 自动文摘
  - 根据不同比例以及用户的不同需求自动编写文摘
- 定题服务
  - 面向确定主题的情报服务
- 个性化新闻
  - 根据用户的兴趣偏好，
  - 为用户定制新闻

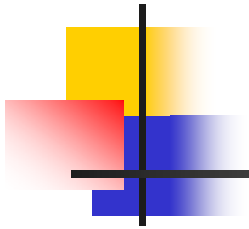
内容安全

- 垃圾邮件过滤
  - 包括广告、黄色和反动邮件的过滤和分析
- 垃圾短信过滤
- 企业商业秘密防泄露
  - 监测从企业内部发出的邮件，封杀包含企业机密的邮件
  - 聊天室和BBS监控
  - 过滤黄色话题或反动言论
- 垃圾短信过滤



商务智能

- 自助呼叫中心
  - 以自动问答的方式，从企业提供的大量技术支持资料中自动获取答案，满足用户的需求
  - 减少呼叫中心的人力服务费用
- 用户投诉信的自动分类和汇总系统
  - 将用户的投诉信自动分发给企业的不同部门去处理
  - 自动发现投诉信中的焦点问题，协助企业决策
- 竞争情报
  - 定制关于互联网上关于竞争对手的各种情报并汇总



电子政务

- 首长办公系统
  - 自动汇总来自各下属部门的文件，并提取重要内容提供给领导阅读
- 政务自动咨询系统
  - 市民通过互联网，以问答的方式咨询政府的政策和办事流程等
- 投诉自动汇总分析系统
  - 将市民的投诉自动分类汇总，以资政府决策
- 行政简报自动编写系统
  - 定期自动编写简报，在政府部门内交流

远程教育

- 自动答疑系统
  - 用户远程提问，系统根据用户的问题收集教材中的相关内容，汇总后提供给用户
- 学生情况调查分析
  - 根据学生的提问情况，自动分析学生的主要问题所在，以便对症下药地改进教学内容

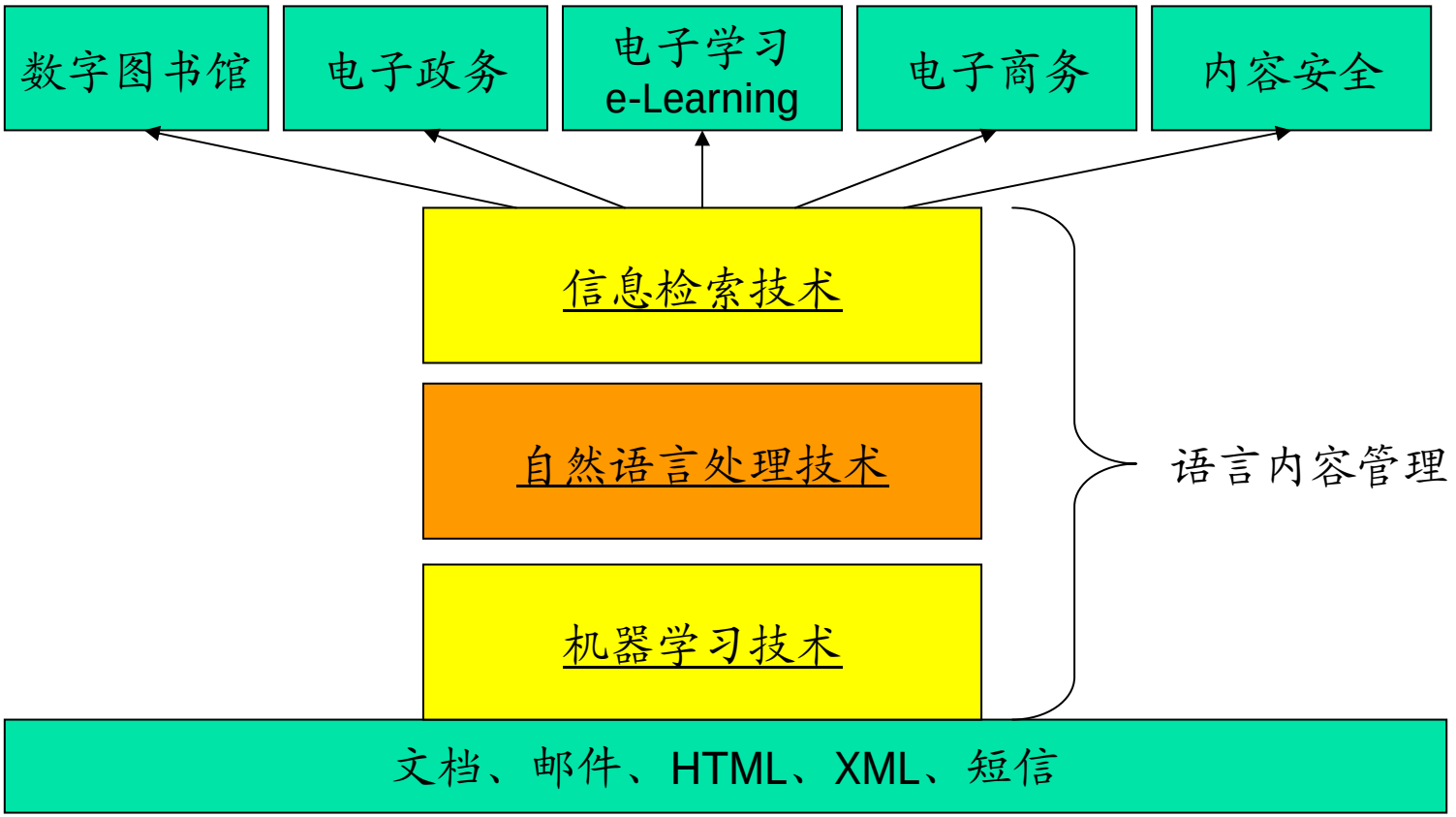
移动计算

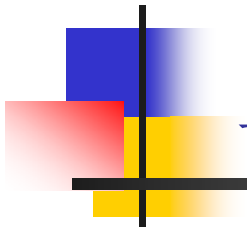
- 短信定制服务
  - 包括新闻、股市资讯等
- 短信汇总服务
  - 电视台或广播电台常常提供在线的短信参与活动，大量短信发送到电视台需要及时地分类汇总，以便主持人作出反应，比如概括出大多数用户最关心的问题等。

军事情报

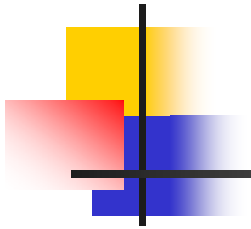
- 国外军事情报的跟踪汇总
  - 重点针对国外互联网进行过滤跟踪，对重要资料进行分析汇总，辅助军事决策
- 国内军事情报的反泄露
  - 发现和拦截泄露军事情报的邮件
- 隐藏于普通文本中的军事情报的过滤技术
  - 文本水印

信息检索应用


数字图书馆
电子政务
电子学习
e-Learning
电子商务
内容安全
信息检索技术
自然语言处理技术
机器学习技术
语言内容管理
文档、邮件、HTML、XML、短信

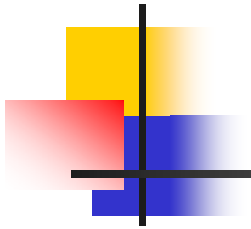


主要研究机构、会议和期刊



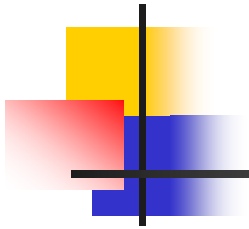
国际研究机构

- 美国CMU语言技术中心LTI
  - Jamie Callan,Yiming Yang
  - http://www.lti.cs.cmu.edu/
- 美国南加州大学ISI
  - Eduard Hovy, Chin-Yew Lin
  - http://www.isi.edu/natural-language/



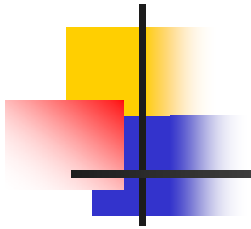
国内研究机构

- 中科院计算所
  - 白硕、程学旗、王斌
  - http://159.226.40.18/
- 复旦大学
  - 吴立德、黄萱菁
  - http://www.cs.fudan.edu.cn/mcwil/irnlp/
- TRS公司（北京信息工程学院）
  - 施水才
  - http://www.trs.com.cn



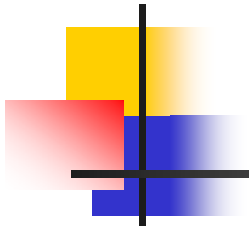
国际会议

1. ACM SIGIR
2. International World Wide Web Conference(WWW)
3. ACM SIGMOD
4. International Conference on Data Engineering (ICDE)
5. ACM Conference on Information and Knowledge Management (CIKM)
6. ACM SIGKDD
7. International conference on Web Intelligence
8. ACM VLDB
9. Europe Conference on Information retrieval(ECIR)
10. International Conference on Machine Learning ICML
11. Asia Pacific Web Conference (APWeb)
12. International Conference on Web Information Systems Engineering
13. International Conference on Artificial Intelligence(IJCAI)
14. TREC(Text REtrieval Conference)



国内会议

- 全国信息检索与内容安全会议
  - 中国中文信息学会信息检索与内容安全专业委员会
  - 上海，2004年11月4-5日
- 全国搜索引擎与网上信息挖掘学术研讨会
  - 中国计算机学会互联网专业委员会
  - 广州华南理工，今年下半年
- 全国计算语言学联合学术会议
  - 中国中文信息学会计算语言学专业委员会
  - 2005年下半年，南京

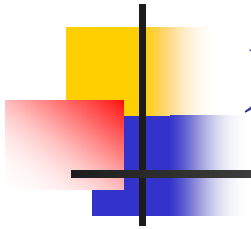


国际期刊

1. ACM Transaction on Information Systems
2. IEEE Transaction on Knowledge and Data Engineering
3. Information Retrieval
4. ACM Transaction on Database system
5. Journal of Intelligent information systems
6. Applied Intelligence
7. Machine Learning
8. Artificial Intelligence
9. Information and Management
10. Information Science
11. Journal of the American Society for Information Science and Technology
12. Information Processing and Management

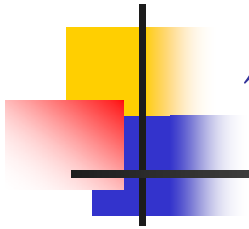
国内期刊

- 中文信息学报
- 情报学报
- 计算机学报
- 软件学报
- 计算机研究与发展
- 自动化学报
- 电子学报
- 高技术通讯



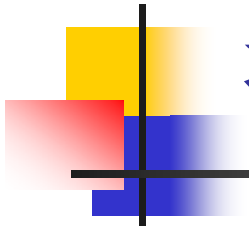
本课的内容

- 信息检索概述
- 信息检索模型
- 搜索引擎
- 信息过滤
- 信息抽取
- 自动文摘
- 问答系统
- 文本分类和聚类
- 数字图书馆



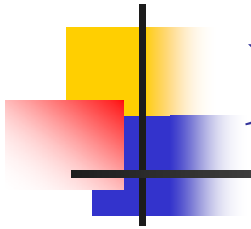
信息检索模型

- 信息检索模型的概述
- 布尔模型
- 向量空间模型(VSM)
- 概率模型
- 模糊集模型
- 扩展的布尔模型
- 潜在语义索引模型(LSI)
- 神经网络模型
- 贝叶斯网络模型
- 信念网络模型



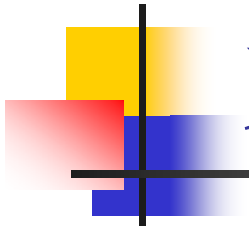
搜索引擎

- 词(Term)处理
  - Stemming技术，词法分析、形态还原，停用词表的构建，语义词典的构建，分词、词性标注和词义消歧等。
- 索引技术
  - 倒排文档(Inverted List)，Signature文件，PAT树和PAT阵列等。
- 提问(Query)处理
  - 提问理解，提问的语义扩展（包括基于局部聚类的提问扩展，基于局部上下文分析的提问扩展，基于相似语义词典的提问扩展，基于统计词典的语义扩展），
- 相关反馈（包括Term权重的重新计算，相关反馈策略的评价等）。
- 搜索引擎技术，集中式体系结构，分布式体系结构，对检索结果的排序(Ranking)问题，Web上的信息采集(Crawler)，元搜索引擎(Meatsearch)等。



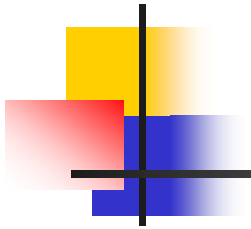
文本过滤

- 过滤系统中的Profile的表示与管理
- 匹配技术
- 介绍各种匹配算法
  - Brute Force算法
  - Knuth-Morris-Pratt算法
  - Boyer-Moore算法
  - Shift-Or算法
  - 多模式匹配算法
  - 模糊匹配算法
  - 多模式模糊匹配算法。
- 过滤系统在信息安全中的应用



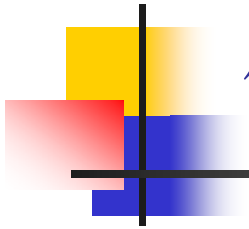
文本分类和聚类

- 特征词抽取
  - TFIDF
  - 信息增益方法
- 文本表示
- 文本相似度计算
- 文本聚类算法
- 文本分类和聚类的应用系统等



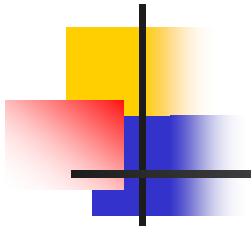
问答式信息检索

- 问题的理解与分类
- 转述(Paraphrasing)
- 答案抽取
- 问答式信息检索的应用



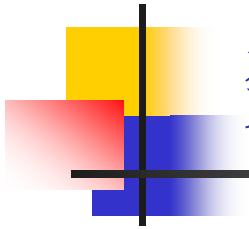
信息抽取

- 命名实体的抽取
- 实体之间关系的抽取
- 事件或观点的抽取等
- 信息抽取在知识发现、商务智能等方面的应用



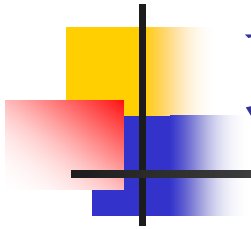
自动文摘

- 机械式文摘
- 基于文本结构的文摘
- 基于理解的文摘
- 与问题相关的文摘
- 多文档文摘



数字图书馆

- 数字图书馆的概念、应用价值，数字图书馆的体系结构、关键技术
- IBM内容管理、TRS内容管理
- 多媒体检索
  - 图像检索、音频检索、语音检索、视频检索等
- 多跨语言检索
  - 提问翻译中的译文选择问题和语义扩展问题，检索结果的翻译问题



联系方式

- 刘挺
  - tliu@ir.hit.edu.cn
  - 86413683-801
- 车万翔
  - car@ir.hit.edu.cn
  - 86413683-806
- 信息检索研究室主页
  - http://ir.hit.edu.cn/