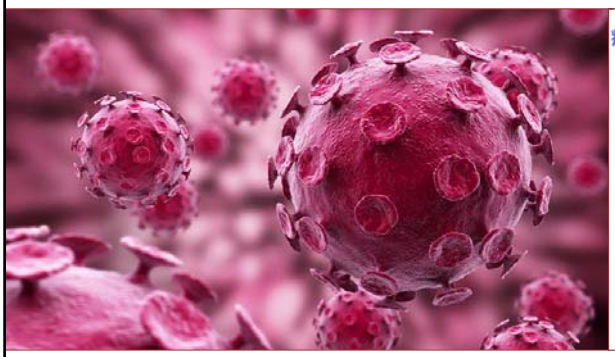
南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

2020的春节：关注



信息检索

2003年SARS与中国互联网发展

2003年5月10日,淘宝网成立

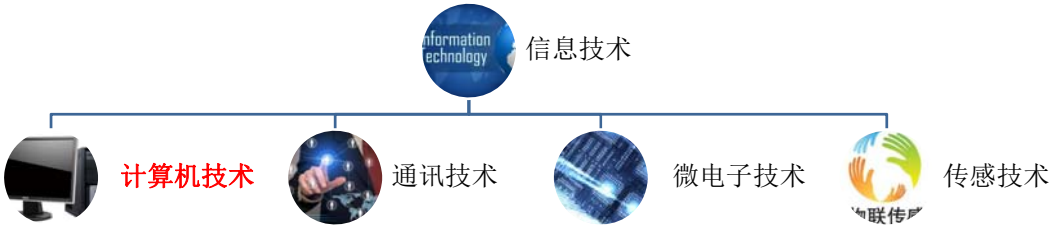
2004年1月,京东多媒体网www.jdlaser.com开通

信息检索

2020热点

信息检索

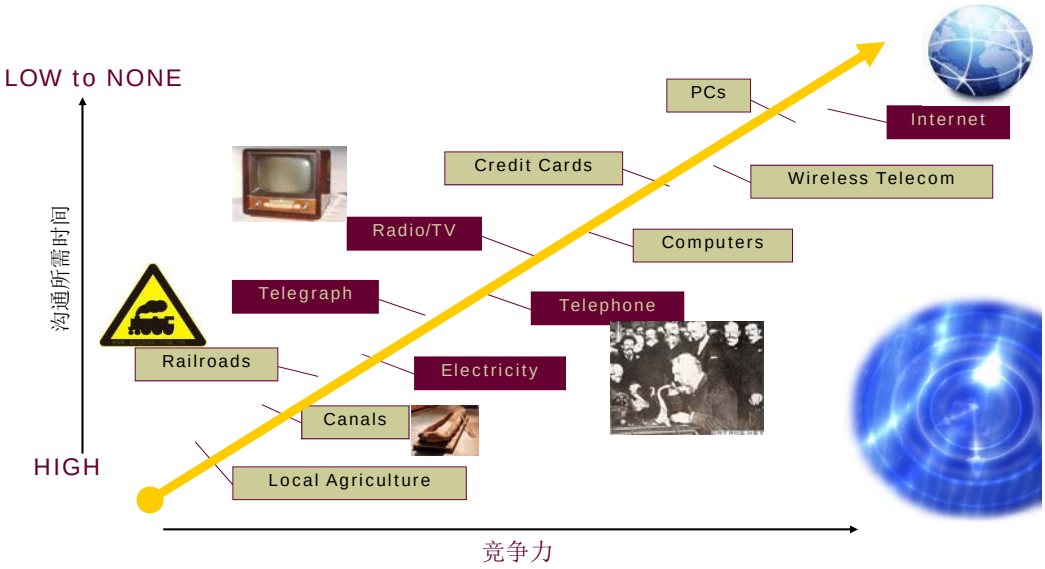
（现代）信息技术的范畴



5

信息检索

信息技术的变迁



6

信息技术的发展节点

阶段	标志	特点
1	语言	猿到人的重要标志
2	文字	信息存储和传递突破时间和空间的限制
3	造纸与印刷术	知识的大规模传递
4	电报电话电视	光（电）速带来传播效率的飞跃
5	计算机及现代通讯	机器开始代替人工

信息检索

关于信息素养

Information Literacy

1979年，全美信息产业协会（AIIL）经过五年的研究后，重新定义信息素质，认为具备信息素质的人即是“**掌握了信息工具利用的知识与技能，并能够应用于解决实际问题的人**”。

具备信息素质的人，能够识别何时需要信息，知道如何查找、评估和有效利用需要的信息来解决实际问题或者做出决策，无论其选择的信息来自于计算机、图书馆、政府机构、电影或者其他任何可能的来源。

—美国图书馆协会（ALA）和美国教育传播与技术协会（AECT）1989年提交给总统委员会的《关于信息素质的报告》。

9

信息检索

美国高等教育信息素质能力标准

Information Literacy Competency Standards for Higher Education

2000年美国大学与研究图书馆协会批准并颁布，2004年美国高等教育协会与独立学院委员会正式通过。
—http://www.ala.org/ala/mgrps/divs/acrl/standards/standards.pdf

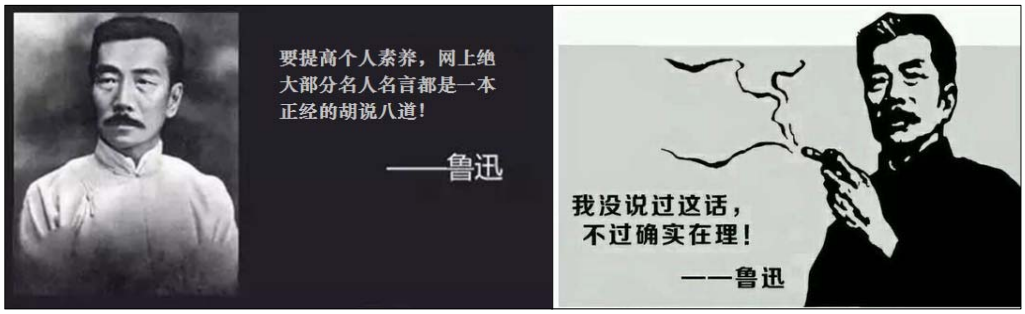
（1）能确定所需信息的性质和范围；
（2）**能有效而又高效地获取所需信息；**
（3）能批判性地评价信息及其来源，并把所选取的信息融入已有知识与价值系统；
（4）能有效地利用信息达到特定目的；
（5）懂得有关信息技术的使用所产生的经济、法律和社会问题，并能在获取和使用信息时遵守公德和法律。

10

不信谣、不传谣

https://www.ifla.org/publications/node/11674

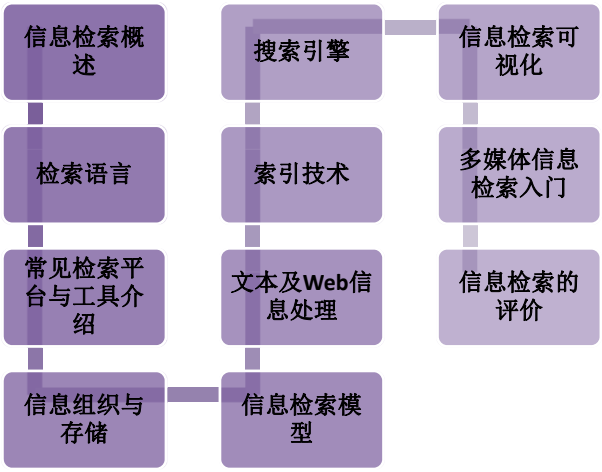
提高信息素养，做合格的信情中人!!!



图片分析

信息管理基本流程

课程框架



教学团队

主讲： 邓三鸿
岳泉

助教： 张艺炜
郭 庆
……

群名称:信息检索教学群
群 号:570666639

教学与考核

教材： 暂无

参考资料： 很多，另附

考核： 平时作业（实践，10%）+综合实践（期末大作业，20%）+闭卷考试（70%）

课程邮箱： njuir@sina.com

建议： 3-5人组成学习小组（数学知识、编程能力、综合分析能力）

每章节至少阅读10篇综述性论文（自己检索？）

从数觉谈起

在一个小的集合里边，增加或者减去一样东西的时候，
尽管未曾直接知道增减，也能够辨认到其中有所变化

信息检索

基本概念-数据

数据是事实或观察的结果，是对客观事物的性质、状态以及相互关系等进行记载的**物理符号**。

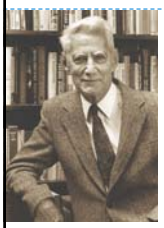
数据可以是离散的，如符号、文字，称为**数字数据**。也可以是连续的值，比如声音、图像，称为**模拟数据**。

数据本身的表现形式不能够完全表达其内容，需要经过解释，才能体现其价值。

数据经过加工后就成为信息。

信息检索

基本概念-信息



1916.4-2001.2

信息论香农（Claude Elwood Shannon）：信息是用来消除随机不确定性的东西。

控制论维纳（Norbert Wiener）：信息就是信息，既不是物质也不是能量。

经济学界认为，信息是反映事物特征的形式，是与物质、能量相并列的客观世界的三大要素之一，是管理和决策的重要依据。

对信息进行提炼、推理后获得的正确的理论是知识。

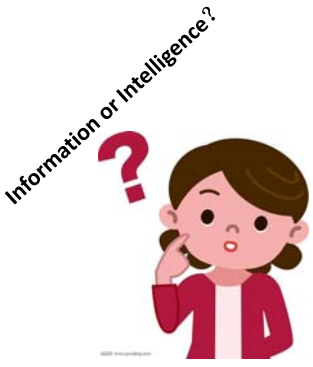
1894.11-1964.3

基本概念-知识

- 知识是信息接收者通过对信息的提炼和推理而获得的正确结论，是人通过对信息对自然界、人类社会以及思维方式与运动规律的认识与掌握，是人的大脑通过思维重新组合的、系统化的信息集合。
- 知识的传输一般遵循如下模式：

传输者的知识 → 数据 → 信息 → 接收者的知识

基本概念-情报



军事情报观对情报的解释是，情报是“以侦察的手段或其它方式获取有关对方的机秘情况”。

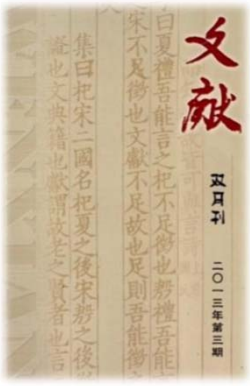
信息情报观对情报的解释是，情报是被人们感受并可交流的信息。

知识情报观对情报的解释是，情报是人们为解决特定问题而被活化了的更为高级，更为实用的知识，是传播中的知识。

大庆油田的典故

- **大庆油田的位置**
  1964年4月20日《人民日报》：大庆精神大庆人—大庆油田确有其事；
  1966年7月《中国画报》：照片—北满，齐齐哈尔和哈尔滨之间， 通过油罐车上土的颜色和厚度证实；
  1966年10月《人民中国》：王进喜事迹介绍，马家窑；
- **大庆油田的规模**
  王进喜的先进事迹介绍，1959年9月
- **大庆油田的加工能力**
  1966年7月《中国画报》：炼油厂反应塔照片—反应塔直径5米，
- 结论：**炼油设备不足，购买日本的轻油裂解设备完全可能。**

基本概念-文献



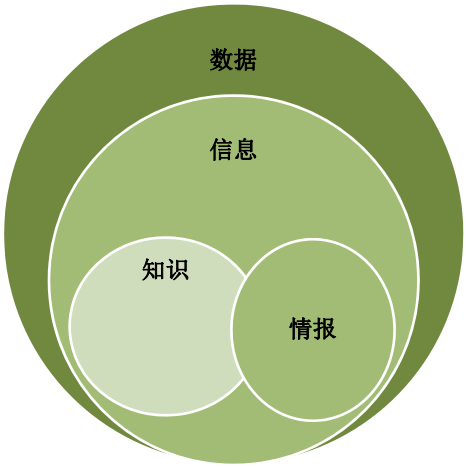
文献是用文字、图形、符号、声频、视频等技术手段记录人类知识的一种载体。

文献是记录、积累、传播和继承知识的最有效手段，是人类社会活动中获取情报的最基本、最主要的来源，也是交流传播情报的最基本手段。

GB/T4894-1985定义：记录知识的一切载体。

基本概念的转换

基本概念的包容



- 数据是事实的数字化、编码化、序列化、结构化；
- 信息是数据在信息媒体上的映射；
- 知识是对信息的加工、吸收、提取和评价的结果；
- 情报是特指的专业信息，是传播中的知识
- 文献是载体

定义：信息检索

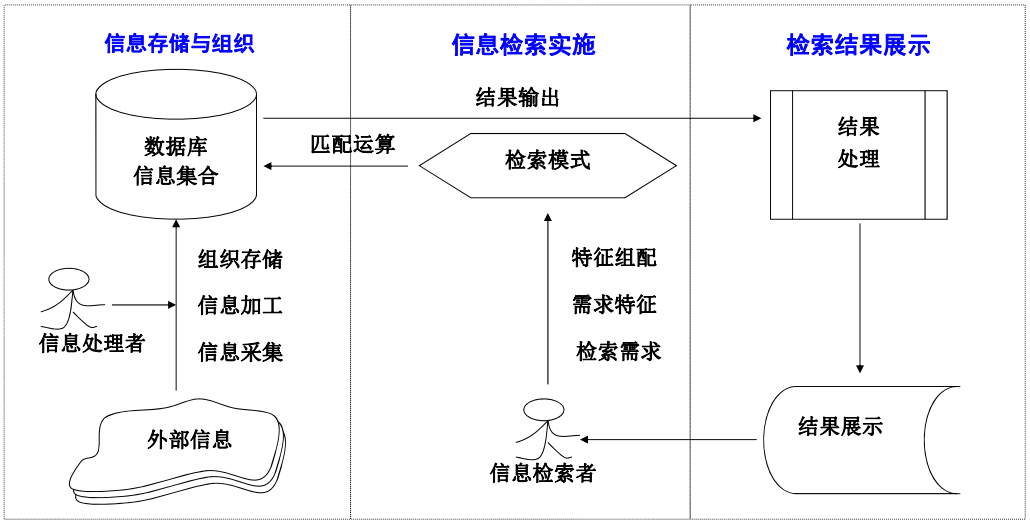
信息检索（Information Retrieval）是用户进行信息查询和获取的主要方式，是查找信息的方法和手段。

狭义的信息检索仅指信息查询（Information Search）。即用户根据需要，采用一定的方法，借助检索工具，从信息集合中找出所需要信息的查找过程。

广义的信息检索是信息按一定的方式进行加工、整理、组织并存储起来，再根据信息用户特定的需要将相关信息准确的查找出来的过程。又称信息的存储与检索。

一般情况下，**信息检索指的就是广义的信息检索**。

信息检索的原理



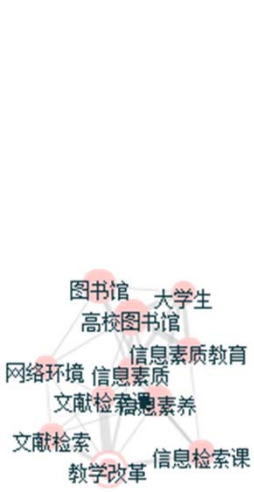
信息检索

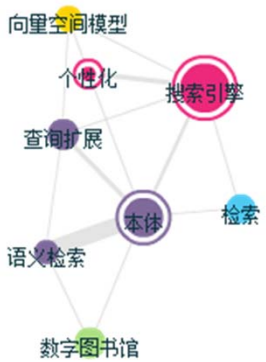
信息检索的意义

- 较全面地掌握有关的必要信息
- 提高信息利用效率，节省时间和费用
- 提高信息素养，加速成才

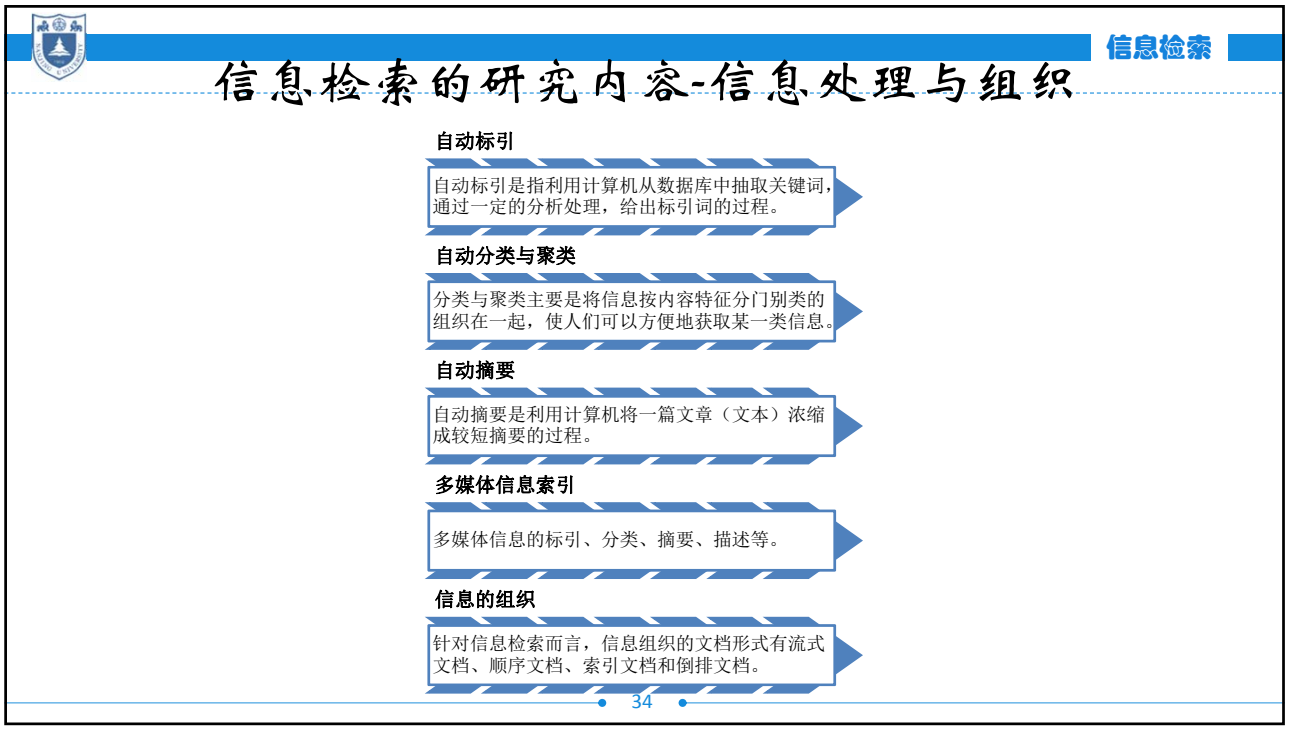
信息检索

信息检索的研究内容





CNKI，2018.3，主题检索



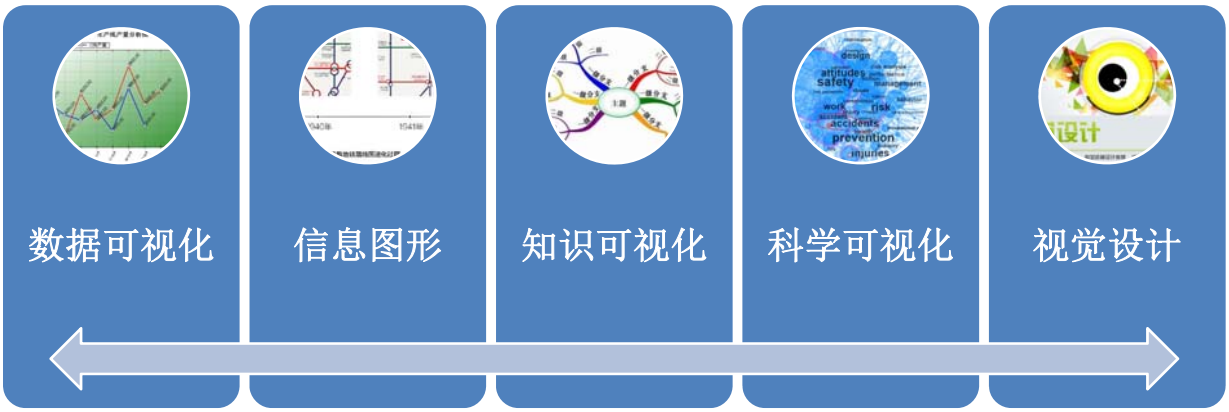
信息检索的研究内容-信息检索技术与方法

信息检索技术与方法是保证检索系统实现**高效**的检索过程、**准确**的检索结果的手段。

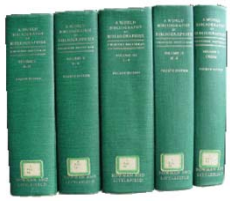
检索技术与方法由检索算法确定，检索算法主要来自于数学理论和方法，采用的数学模型不同，其相应的实现技术与方法也不同。

目前，**常用的检索技术**有：布尔检索、加权检索、全文检索、超文本检索、多媒体检索、智能检索、跨语言跨平台检索等。

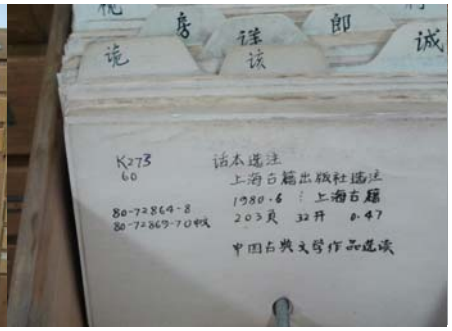
信息检索的研究内容-可视化技术



信息检索的发展-手工



- 书目
- 索引卡片



信息检索的发展-脱机

1948:

C. N. Mooers在其MIT的硕士论文中第一次创造了“Information Retrieval”这个术语。

1960－70年代：

人们开始使用计算机为一些小规模科技和商业文献的摘要建立文本检索系统。

产生了布尔模型(Boolean Model)、向量空间模型(Vector Space Model)和概率检索模型(Probabilistic Model)。

康奈尔大学的Salton领导的研究小组是该领域研究的佼佼者。

伦敦城市大学的Robertson及剑桥大学的SparckJones是概率模型的倡导者。

信息检索的发展-联机

- 1980’s:
  - IR技术出现在大型文档数据库中

- Lexis-Nexis
  - 美国LEXIS-NEXIS公司创始于1973年，其数据库内容很广，其中法规法律方面的数据库是LEXIS-NEXIS的特色信息源，具有非常大的影响力，尤其在法律业界具有很高知名度
- Dialog（1972-）
  - 目前世界上最大的联机检索检索系统之一，包括各学科数据库600多种，可查询研究动态，SCI、EI收录以及专利等情况
- MEDLINE
  - MEDLINE是美国国家医学图书馆的文献数据库

信息检索的发展-Internet

- 1986：现代Internet正式形成（早期为APPAnet）
- 1990’s:
  - 在互联网上进行对FTP文档进行搜索
    - Archie
      - 第一个网络搜索工具：1990年加拿大蒙特利尔McGill大学开发的FTP搜索工具Archie
      - Archie是Internet上用来查找其标题满足特定条件的所有文档的自动搜索服务的工具。
    - WAIS
      - 代表“广域信息服务”(Wide Area Information Service)。Wais作为Internet一项服务， 是唯一由三个商业公司（Apple、Thinking Machines和Dow Jones ）启动的研究计划促成的服务。

http://archie.icm.edu.pl

信息检索的发展-Internet

- 1990's:
  - 第一个网络搜索工具：1990年加拿大蒙特利尔McGill大学开发的FTP搜索工具Archie
  - 第一个WEB搜索引擎：1994年美国CMU开发的Lycos
  - 1995：斯坦福大学博士生开发的Yahoo
  - 1998：斯坦福大学博士生开发的Google，提出PageRank计算公式。
  - 1998：基于语言模型的IR模型提出。

信息检索的发展-前沿

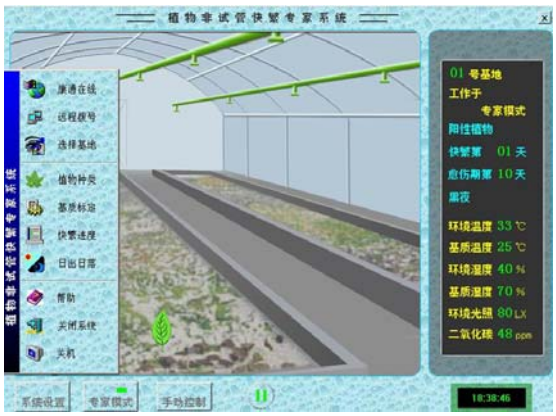
- 2000's :
  - 多媒体IR
    - 图像(Image)
    - 视频(Video)
    - 声音(Speech)和音频(Audio)
    - 音乐(Music)
  - 跨语言检索Cross-Language IR
    - DARPA Tides项目
  - 智能化、个性化IR

信息检索系统的分类-资源

- 书目检索系统
- 全文检索系统
- 多媒体检索系统

信息检索系统的分类-服务功能

- 单纯检索服务系统
- 统计分析系统
- 决策支持系统
- 专家系统



信息检索系统的分类-区域功能

- 单机检索系统
- 联机检索系统
- 网络检索系统

信息检索系统的分类-系统逻辑

- 文献检索
- 数据检索
- 事实检索

化学物质毒性数据库
Chemical Toxicity Database

信息检索系统的评价

- 资源的收录状况
- 数据的质量
- 检索的功能和效率
- 系统的功能
- 检索结果的反馈形式

信息检索的相关学科领域

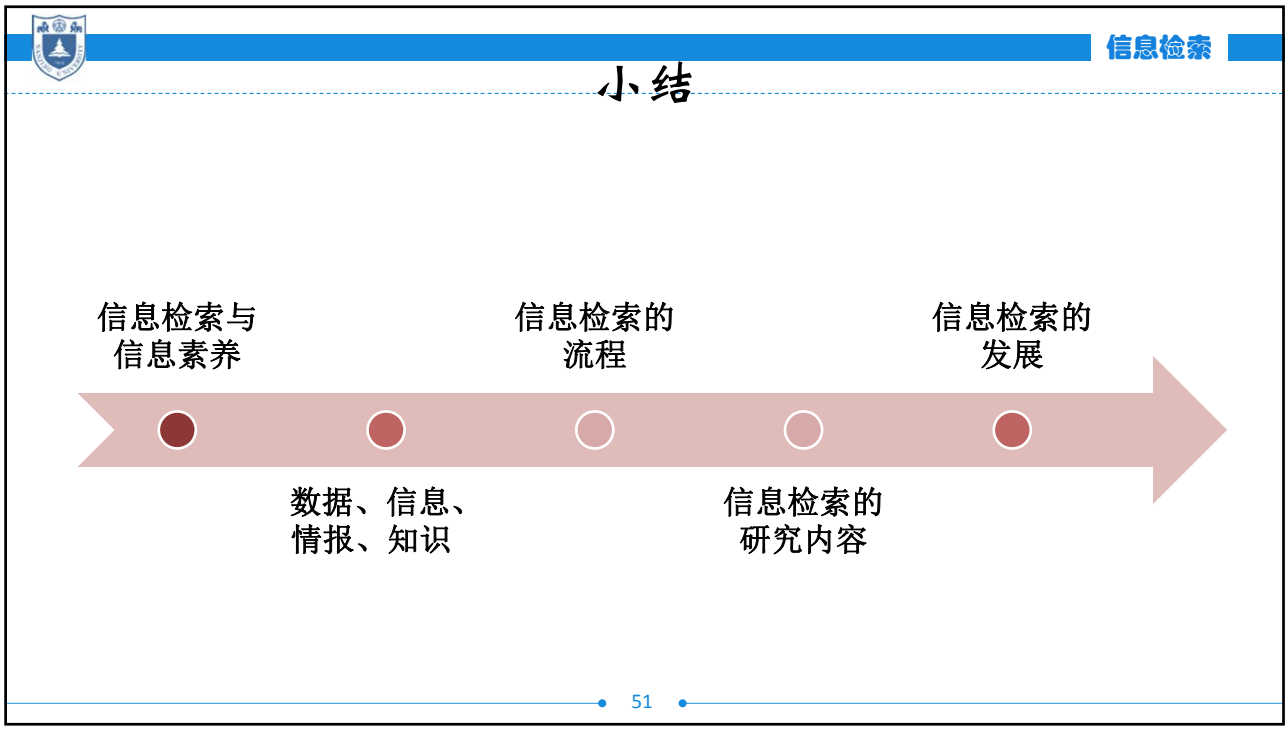
- 数据库管理
- 图书和情报科学
- 人工智能
- 自然语言处理
- 机器学习

信息检索的难点

- 分析技术亟待更新，否则很难有质的突破
- 资源
- 用户
- 系统

信息检索的未来趋势

- **统一的检索界面**
  - ✓ 实现分布式、跨语言、跨平台检索，统一检索界面将成为未来网络信息服务的主流。
- **主动的信息推送服务**
  - ✓ SDI服务/Selective Dissemination of Information
- **多种检索模型融为一体**
  - ✓ 各种模型代表的检索技术融为一体，互相取长补短。
- **可视化技术实用化**
  - ✓ 将在信息检索过程中的各阶段：文档集的表示、检索提问的表达、结果集的检查、用户反馈等过程均使用到可视化技术。
- **检索的智能化**
  - ✓ 信息检索的智能化水平提高、用自然语言检索将成为现实。智能化信息抓取、智能信息处理、智能检索将成为信息信息检索系统的重要组成部分。



南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

02

PART Two

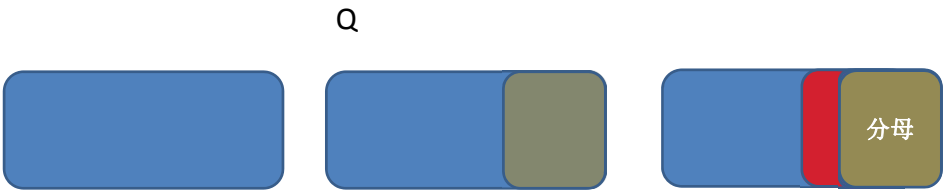
信息检索语言

Information Retrieval Language

预知概念：查全率

- 查全率（Recall ratio）-检出的信息数量与检索系统中相关信息总量之间的比率

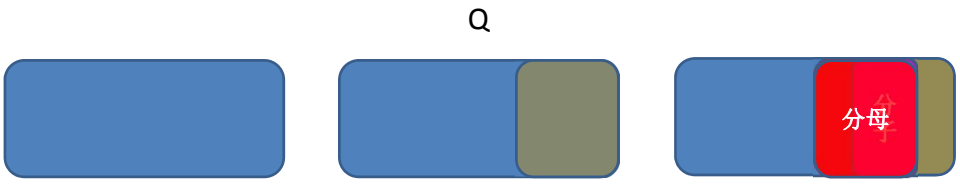
$$R=\frac{\text{检出的相关信息数量}}{\text{系统中的相关信息数量}}\times 100\%$$



预知概念：查准率

- 查准率（Precision ratio）-检出的相关信息数量与检出的信息总量的比率

$$P=\frac{\text{检出的相关信息数量}}{\text{检出的信息总量}}\times 100\%$$



信息检索

2×2表

表12×2表

用户 相关性 判断 / 系统相关性预报	相关文献	非相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

信息检索

漏检率与误检率

表12×2表

用户 相关性 判断 / 系统相关性预报	相关文献	非相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

$$\text{漏检率}(M)=\frac{\text{未检出的相关文献}}{\text{文档中相关文献总量}}\times 100\%=\frac{c}{a+c}\cdot 100\%$$

$$\text{误检率}(N)=\frac{\text{检出的不相关文献量}}{\text{检出的文献总量}}\times 100\%=\frac{b}{a+b}\cdot 100\%$$

R+M=1， P+N=1

局限

查全率的局限性：它是检索出的相关信息量与存储在检索系统中的全部相关信息量之比，但系统中相关信息量究竟有多少一般是不确知的，**只能估计**；另外，查全率或多或少具有“假设”的局限性，这种“假设”是指检索出的相关信息对用户具有同等价值，但实际并非如此，对于用户来说，信息的相关程度在某种意义上比它的数量重要得多。

查准率的局限性：如果检索结果是题录式而非全文式，由于题录的内容简单，用户很难判断检索到的信息是否与课题密切相关，必须找到该题录的全文，才能正确判断出该信息是否符合检索课题的需要；同时，查准率中所讲的相关信息也具有“假设”的局限性。

语言

语言（Language）是采用一套具有**共同处理规则**来进行表达的沟通指令

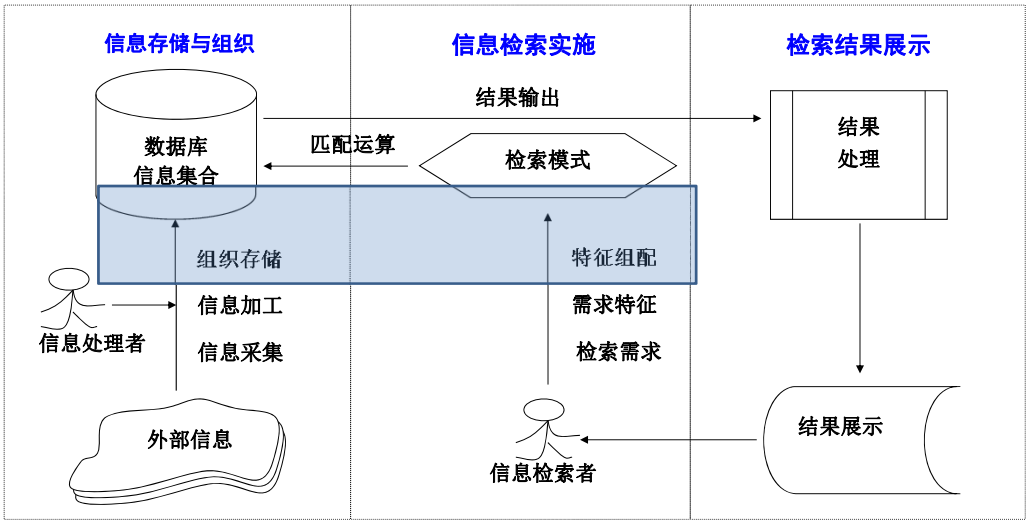
基本概念

检索语言有广义和狭义之分。

广义的检索语言泛指信息检索过程中涉及的人工语言和自然语言。人工语言是根据一定的规则，人为编制而成的检索语言，它有着严格的使用规则，可用于表述文献主要内容，建立信息检索系统。自然语言是人类交流时使用的语言，不受任何限制，未经加工和规范。

狭义的检索语言仅指根据信息检索的需要，按照一定的规则对自然语言进行规范，并专门用于信息标引和用户检索的人工语言。

信息检索语言的功能



IRL的作用

- 标引信息内容特征及某些外表特征，保证不同标引人员表达信息的一致性
- 对内容相同及相关的文献信息加以集中或揭示其相关性
- 使信息的存储集中化、系统化、组织化，便于检索人员按照一定的排列次序进行有序化检索
- 便于将标引用语和检索用语进行相符性比较

例

有三篇文献篇名如下：

文献1：A Model of multimedia information retrieval
文献2：The Information retrieval in chemistry WWW server
文献3：ERIC resources

在对信息存储的过程中，对这三篇文献内容分别进行了分析，并使用检索语言对其进行标引，标引结果为：

文献1：**篇名(title)**：A Model of multimedia information retrieval
主题(subject)：information retrieval, multimedia computer applications

文献2：**篇名(title)**：The Information retrieval in chemistry WWW server
主题(subject)：chemistry, educational materials

文献3：**篇名(title)**：ERIC resources
主题(subject)：educational materials

标引后这三篇文献分别被存储进数据库。

例

在信息检索过程中：

如果用户输入“information retrieval”一词，并将检索范围限定在篇名中，则文献1与文献2符合用户要求，成为检索结果。

如果用户输入“information retrieval”[illegible]中，则只有文献1符合用户要求，成为检索结果。

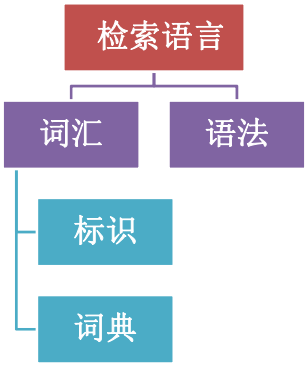
如果用户输入“educational materials”[illegible]中，则文献2和文献3符合用户要求，成为检索结果。

检索式	检索结果	
# 3	43,182	**主题:** (educational materials) *时间跨度=所有年份* *检索语言=自动*
# 2	530,824	**主题:** (information retrieval) *时间跨度=所有年份* *检索语言=自动*
# 1	25,939	**标题:** (information retrieval) *时间跨度=所有年份* *检索语言=自动*

在上述例子中，“information retrieval”、“educational materials”都是检索语言，篇名和主题则是检索语言的标识，检索系统就是通过他们将用户需求与信息内容进行运算匹配，最终找到检索结果的。

由这个例子可以看出，检索语言的主要作用就是对文献的外部特征和内容进行多层次描述，提供多种检索途径，以方便用户从不同角度检索查找。

信息检索语言的组成



词汇：登录在类表、词表中的全部标识，一个标识（分类号、检索词、代码）就是它的语词，而分类表、词表则是它的词典；

语法：如何创造和运用标识（单个或组合）来正确表达信息内容和信息需要。

信息检索

检索语言的分类-描述文献的特征

描述文献外表特征的检索语言

- 题名——题名索引
- 著者——著者索引、团体著者索引
- 文献编号
  - 报告号索引
  - 合同号索引
  - 存取号索引
- 其它——引文索引

描述文献内容特征的检索语言

- 分类语言——体系分类语言、组配分类语言
- 主题语言——标题词语言、关键词语言、单元词语言、叙词语言
- 代码语言——分子式、结构式索引、专利号索引等

信息检索

检索语言的分类-其他

- 按结构或原理
  - 分类语言、主题语言、代码语言和引文语言
- 按信息标识的组合使用方法
  - 先组式语言、后组式语言和散组式语言
- 按语言的规范程度
  - 人工语言和自然语言

信息检索

检索语言的理论基础-概念逻辑

概念（Concept）人类在认识过程中，从感性认识上升到理性认识，把所感知的事物的共同本质特点抽象出来，加以概括，是自我认知意识的一种表达，形成概念式思维惯性。在人类所认知的思维体系中最基本的构筑单位。

概念间的关系

信息检索

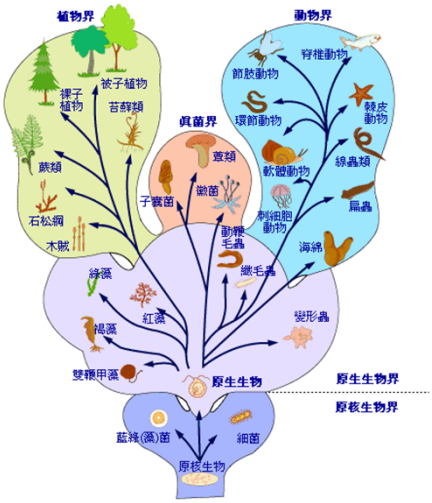
概念逻辑方法

1. **概念的划分与概括（分类）**

建立概念等级体系，用以显示客观世界千差万别的事物之间内在联系。这种结构具有很好的系统性。例如，体系分类法就是用此种逻辑方法的典型。

2. **概念的分析与综合（组配）**

建立概念组配体系，提供从多种途径来进行信息检索的功能而且可以任意选择检索标识的专指度，根据实际需要扩大、缩小改变检索的范围。例如，叙词语言与组配分类法便是应用概念分析与综合的典型。



知识分类

知识分类是对千差万别的事物做系统研究的重要方法，是对各种事物之间的区别和联系从本质上、原理上进行揭示的重要手段，对信息的系统化具有重要的价值，其实质是划分知识单元、组织知识体系，包括学科分类和事物分类。

知识分类应当遵循的两条基本原则是**客观性和发展性**

学科分类是知识分类的主体，事物分类是知识分类的基础

术语学

术语（Terminology）是在特定学科领域用来表示概念的称谓的集合，或者说，是通过语音或文字来表达或限定科学概念的约定性语言符号。

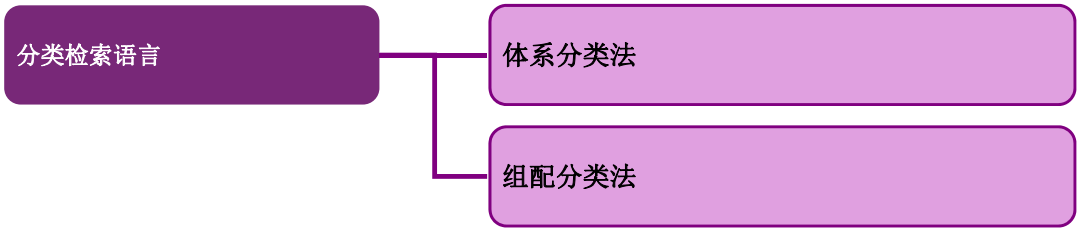
术语是分类表、词表的基本组成要素，检索语言其实就是一个经过精细组织的术语集。

检索语言的创制以术语学的研究成果为基础的。

信息检索

分类检索语言概述

分类检索语言也称分类法，是将许多类目根据一定的原则组织起来，通过标记符号（分类号来代表各级类目和固定其先后次序的分类体系。



信息检索

分类检索语言-体系分类法

体系分类法的结构

微观结构

微观结构指分类法中类目的构成结构。

（1）类目的划分
（2）引用次序
（3）类目的排列
（4）类名的确定
（5）类目之间相互关系的处理

宏观结构

按功能分，体系分类法的宏观结构一般由以下四部分组成：

（1）类目体系
（2）标记系统
（3）说明与注释
（4）类目索引

分类检索语言-体系分类法的特点

体系分类法在实际工作中，主要被用来组织分类排架和统计藏书和建立分类检索系统。

体系分类法的主要特点：

1. 按学科、专业属性构建类目体系，形成按学科、专业集中文献、信息的知识概念系统，从而能够直接地满足用户从学科、专业出发检索课题的需求，可以达到较高的查全率；
2. 采用等级列举式的概念标识系统来揭示概念之间的相互关系，便于用户“鸟瞰全貌”、“触类旁通”、“层层深入”地查找某一专业的信息，用户也无须事先知道事物或概念的确切名称，就可以在一定的类目下通过浏览查到该领域的相关信息；
3. 采用分类号作为主题的标识，不受语种的限制。

分类检索语言-主要体系分类法

国内常见的体系分类法有：

《中国人民大学图书馆图书分类法》，简称《人大法》；
《中国图书馆分类法》，简称《中图法》；
《中国科学院图书馆图书分类法》，简称《科图法》；
《中国档案分类法》

国外常见的体系分类法有：

《杜威十进分类法》（Dewey Decimal Classification），简称DC或DDC
《美国国会图书馆分类法》（Library of Congress Classification），简称LC
《国际十进制分类法》（Universal Decimal Classification），简称UDC.

信息检索

分类检索语言-《中国图书馆分类法》

我国目前广泛使用的分类法是《中国图书馆分类法》。它是由国家图书馆等单位组织全国力量，以学科分类为基础，并结合图书的特性所编制的分类法。它将学科分五大部类，基本序列是：**马克思主义列宁主义毛泽东思想、哲学、社会科学、自然科学、综合性图书**，由5大部类、22个大类、8个总论复分表、4万余条类目组成了一个完善的分类体系。

标记制度采用拉丁字母与阿拉伯数字相结合的混合号码制，用一个字母代表一个大类，以字母的顺序反映大类的序列，在字母后用数字表示大类下类目的划分，数字的设置尽可能代表类的级位，并基本上遵从层累制的原则。

例如：

F—经济（大类）

F2--- 经济计划与管理（二级类）

F25 --物资经济（三级类）

F250 --物资经济理论（四级类）

F251.1 ---物资管理（五级类）....

初版1975年，2010年第五版

信息检索

分类检索语言-《杜威十进分类法》

《杜威十进分类法》由美国的威尔·杜威编制，采用纯阿拉伯数字作为基本标记符号，基本上按照层累制展开，是一部在国际上出现最早、流行最广、影响最大的图书分类法。1876年出版，至1996年出版第21版，四卷本。卷一为编制说明和通用复分表，卷二、卷三为类表，卷四为索引和使用手册。它依据培根的知识分类思想，将图书分为十大类：

- 000 总论　500 自然科学
- 100 哲学　600 技术科学
- 200 宗教　700 美术
- 300 社会科学　800 文学
- 400 语言学　900 史地

（Francis Bacon，1561—1626年）

信息检索

分类检索语言-组配分类法

组配分类表

组配分类表是由编制说明、基本类表、分面类表和分面公式以及通用辅表组成。其建立主要采用了分面分析法（Facet Classification）。

分面分析法是将整个知识领域或某一知识领域按其不同属性分解为若干个不同的分面，每个分面再分解为若干个亚面，每个亚面还可分解为若干个更小的子面，面内列出所属各子目的一种编制分类表的方法。

在组配分类表的编制过程中，需要考虑到分面的引用次序与排列次序、标记符号与标记制度等方面的问题。

信息检索

分面示例

《布利斯书目分类法，BBC》人员通用复分表中人员类型类目，列出的分面有：

the Bliss Bibliographic Classification

按语种划分的群体
按家庭关系分：亲戚、先辈
按婚姻状况分：单身、已婚
按年龄分： 孩子、未成年者
按职业特征分：专职人员、受雇用者
按等级和作用分：政府官员、高级职员

信息检索

分类检索语言-组配分类法的特点

1. 通过简单主题概念的组配，一方面可以简化分类表，缩小类表体积，另一方面能够表达各种复杂主题概念和专深主题概念，并且能够揭示主题因素之间的相互关系；
2. 可以对信息所表达的主题概念进行多方面标引，从而可以实现多途径检索；
3. 可以较为及时地增补新的主题概念，类表修订灵活、方便。

信息检索

主要组配分类法：冒号分类法

阮冈纳赞（S.R.Ranganathan）提出了以分析兼综合原则、分面分析和分面标记为核心的分面分类理论。

《冒号分类法》提出的五个基本范畴的理论。它们依次为：**本体**（Personality）、**物质**（Material）、**动力**（Energy）、**空间**（Space）、**时间**（Time）。通过这五个基本范畴来分析、归纳和组织文献。每个基本范畴都采用特性的指示符表示，即，〔P〕；〔M〕：〔E〕.〔S〕‘〔T〕。在第7版中，又将物质面进一步分解成3个方面：物质〔M〕、物质性质〔MP〕、物质方法〔MM〕。

Ranganathan,Shiyali Ramamrita
1892~1972

信息检索

冒号分类法：图书馆相关

第一层次	第二层次	所属各类
本体	图书馆类型	国家图书馆、大学图书馆、儿童图书馆
物质	图书馆材料	图书、期刊、档案
能量	图书馆活动	分类、编目、流通、馆际互借
空间	空间	中国、江苏等
时间	时间	20世纪80年代等

“2图书馆学”的分面公式为：2〔P〕；〔M〕：〔E〕〔2P〕。
据分面公式，标引《大学图书馆期刊分类工作》，
其类号：P中的类目(大学图书馆：34)，M中的类目(期刊：46)，〔E〕〔2P〕中的类目（分类为51），所以：

2〔P〕；〔M〕：〔E〕〔2P〕
2，34； 46 ： 51

信息检索

对比小结

	体系分类法	分面分类法
优点	便于从学科、专业角度按类检索 清晰显示类目体系层次等级关系 号码简单 编制方法相对简单	准确揭示复杂主题因素 随时生成或增补新的主题概念 多面标引，提供多途径检索 类表便于修订和管理
缺点	不能揭示专深主题 多元检索困难 不利于随时修订	类目直观性差 标记方法复杂 不适合分类排架

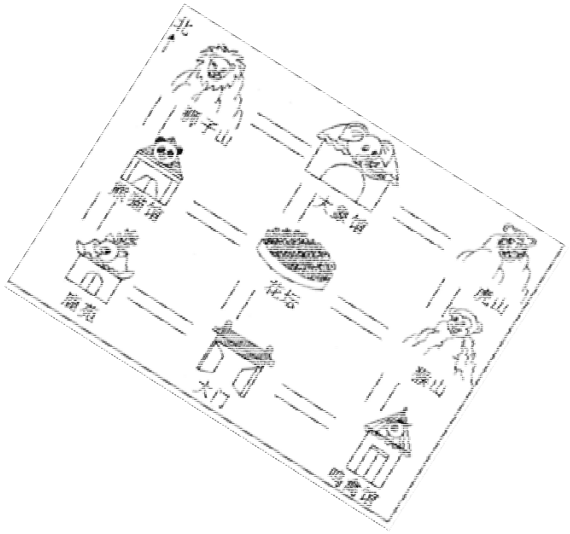
主题检索语言

主题检索语言又称主题法。它采用语词直接作为文献主题标识，按字顺排列主题标识，提供各种检索词语的途径。它从描述事物的特性角度出发，按**文献所论述的事物（即主题）**集中文献，用规范化的名词术语标引和表达文献的主题概念，用参照系统显示事物概念主题词之间的关系。

主题检索语言的主要类型

1	标题词语言
2	单元词语言
3	叙词语言
4	关键词语言

主题检索语言的特点



标题词语言和单元词语言

标题词语言

标题词是从自然语言中选取的、经过规范化处理的、表示**事物概念**的词、词组或短语。标题词按字顺排列，词间语义关系用参照系统显示，并以标题词表的形式体现。

单元词语言

单元词又称元词，是从自然语言中选取，经过规范化处理，表达主题最小的、最基本的、字面上不能再分的名词术语。通过组配来描述文献所论及的事物主题。

单元词语言目前基本发展为叙词语言。

关键词语言

关键词作为信息存储和检索依据的一种检索语言，是直接从原文的标题、摘要或全文中抽选出来，具有实质意义的，未经规范化处理的自然语言词汇。

关键词语言的索引类型：

- 题内关键词索引（KWIC）
- 题外关键词索引（KWOC）
- 词对式关键词索引

叙词语言

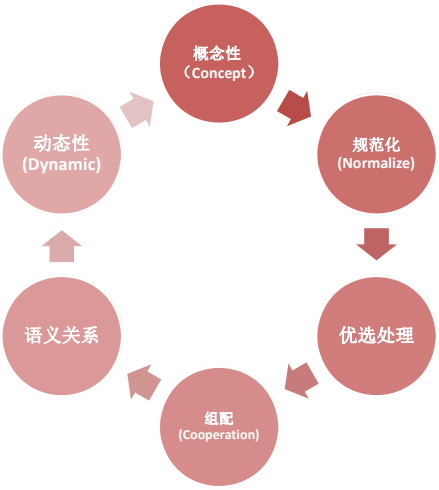
叙词语言也称为主题词，是经过规范化处理的，以基本概念为基础的表达信息内容的词和词组。也叫**受控词**。

叙词语言是以表示单元概念的规范化语词为基础，以**概念组配**为基本原理，对文献主题进行描述的**后组式检索语言**。

叙词语言继承和发展了体系分类语言、组配分类语言、标题词语言、单元词语言、关键词语言等多种检索语言的思想、原理和优点，使其具有多方面的优势，并且已经成为在当今互联网时代下应用最为广泛的人工检索语言之一。

信息检索

叙词的性质



关键词		叙词（主题词）
艾滋病	-------	获得性免疫缺陷综合症
维生素C	-------	抗坏血酸
好奇心	-------	探究行为

信息检索

主要主题词表

《美国国会图书馆主题词表》（Library of Congress Subject Headings，简称LCSH）

《医学主题词表》（Medical Subject Headings）

《汉语主题词表》

《中国分类主题词表》

《社会科学检索词表》

《中国档案主题词表》

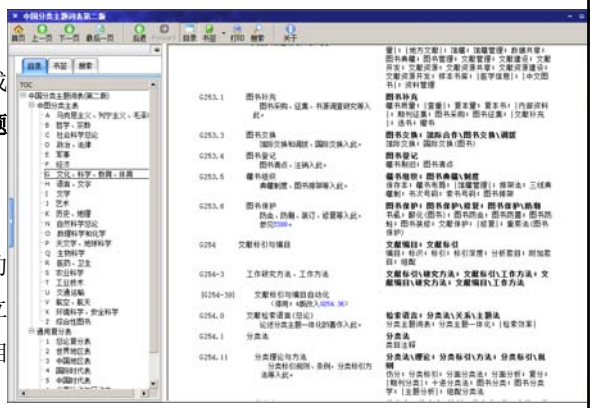
汉语主题词表

该词表是1974年作为“汉字信息处理系统工程”的配套项目开始编制，并在研究和借鉴国外叙词表编制技术的基础上，根据国际标准 ISO2788《单语种叙词表编制规则》，经过近9000人 5年时间的工作，于1980年编成问世。

中国分类主题词表

《中国分类主题词表》是在《中图法》编委会主持下，经全国约40个单位160位专家学者的共同努力，历时8年编制而成的，于1994年6月正式出版。**它是我国第一部分类检索语言和主题检索语言相互兼容对照索引式的一体化词表**。

词表可以使分类标引、主题标引在经过同一主题分析、采用同一标引工具的过程中一次完成，而且能够降低主题标引的难度，提高标引的一致性。同时，由于分类号与主题词之间建立了对应联系，有利于在检索系统中实现分类号与主题词之间的相互转换，从而提高检索效率。



小结：相关分类法

	叙词	单元词	标题词	组配分类法	体系分类法
直观性	好	好	好	差	中
组配	概念组配	字面组配	无	有	无
先组词	部分	无	全部	无	全部
专指性	好	差	好	中	差
检索途径	多	多	少	多	少
检索噪音	小	大	小	中	小
族性检索	中	差	中	好	好
灵活性	好	好	差	好	差
语义关系	有	无	有	无	无
语言类型	先组散组式	后组式	先组式	后组式	先组式

一体化主题检索语言

分类主题一体化检索语言，又称为分类主题一体化词表，是指在一个检索语言系统中，对它们的分类表部分和叙词表部分的术语、参照、标识及索引实施统一的控制，使两者有机地融合为一体，从而能够同时满足分类和主题标引、检索的需要，发挥其最佳的整体效应。

一体化检索语言原理

分类主题一体化检索语言建立在分类检索语言与主题检索语言相通的原理基础之上。

首先，分类检索语言与主题检索语言都是建立在**概念逻辑、知识分类和术语学**的理论基础之上，都应用了概念划分与概括、概念分析与综合的方法。

其次，所采用的表达信息或文献主题概念的标识在本质上相同的，只是表现形式不同而已。

最后，分类检索语言与主题检索语言的处理对象都是**语义单元**，所类集的内容是相同的。

一体化检索语言的功能

分类主题一体化检索语言除了单独具有分类检索语言与主题检索语言的功能外，还具有如下功能：

1. 标引人员可以同时完成分类标引和主题标引，通过标引数据之间的对应转换。
2. 用户既可以从学科、专业出发来进行分类检索，也可以从事物主题出发进行字顺主题检索，提高查全率和查准率。
3. 可以为进行过分类标引而未进行主题标引的书目数据库通过主题词与分类号的转换而提供主题标引，反之亦然。

信息检索

一体化主题检索语言

分类主题一体化检索语言的类型

信息检索

网络信息检索语言

检索语言面临的网络环境

- 信息类型的变化
- 信息数量与质量的变化
- 检索技术的变化
- 信息用户的变化

信息检索

网络信息检索语言-形式变化

网络环境下的分类检索语言

- 分类法的电子化
- 分类体系结构的多维化

信息检索

网络信息检索语言-内容特征

1. 聚类标准的主题化
2. 类目划分的随意性
3. 类目排列非逻辑性
4. 类名的通俗化

在新的网络环境下，分类语言沿着两个方向继续得到发展。

一个方向是积极地调整传统分类法自身，以满足信息资源数量的迅速增长。

另一个方向是抛开传统的分类法，重新建立新的分类体系，即网络分类目录。

自然语言检索

自然语言在信息检索中的应用

自然语言指直接取自文献本身，不经加工和规范的语言，它包含词、词组或句子，没有繁琐规则的约束，不添加任何人工的色彩。

自然语言在信息检索中的应用主要表现为使用关键词的全文检索。

全文检索是指不经过任何标引，而直接通过计算机以自然语言的形式在文本中进行匹配查找。文本中任何字符和字符串均可作为检索入口。

自然语言检索-特点

比较项目 \ 检索语言	情报检索语言	自然语言
检全率	弱	强
检准率	强	弱
标引深度	浅	深
标引一致性	强	弱
标引专指度	低	高
扩检和缩检能力	强	弱
改变检索方向能力	弱	弱
标引成本	高	低
标引速度	慢	快
面向用户能力	差	好
检索成本	低	高
检索人员负担	低	高
词汇更新	慢	快
词表维护	有	无
兼容性	差	好
组织信息资源功能	有	无
适应性	强	弱

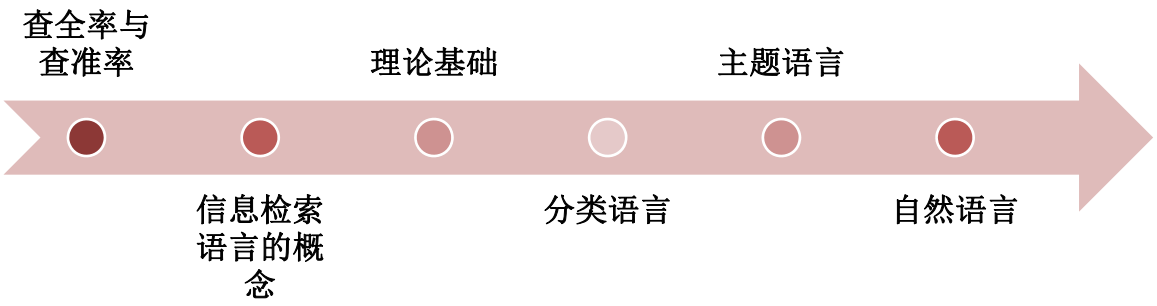
自然语言检索-难点

汉语自动分词问题

词义模糊性、不确定性问题

词间关系的无控制性问题

小结



信息检索

通用复分表例：中国民族表

总论复分表
世界地区表
中国地区表
国际时代表、中国时代表
中国民族表
世界种族与民族表
通用时间、地点复分表

TS959.2 竹、藤、棕、草等加工及制品
草帽、竹篮、藤椅、棕箱等入此。
工艺美术制品入TS93。
参见J528.5。
TS959.3 油漆工艺
一般漆器工艺入此。
建筑油漆工艺入TU767+.3。
TS959.4 纸料工
制盒、裱糊等入此。
TS959.5 制扇、制伞

11/86 七、中国民族表

1. 凡主表中已注明“依中国民族表分”的，均用本表复分。

2. 凡主表中未注明“依中国民族表分”，而需用本表复分时，中国民族号码前需先加中国民族号“2”，并用民族区分标识“ ”。例：中国苗族竹编制品号码为TS959.2“216”。

11	汉族	汉族
12	蒙古族	蒙古族
13	回族	回族
14	藏族	藏族
15	维吾尔族	维吾尔族
16	苗族	苗族

南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

PART Three

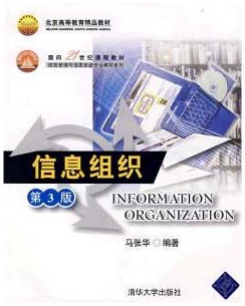
信息组织：标引、描述与存储

Information Organization: Indexing, Description & Storage

信息检索

Information Organization

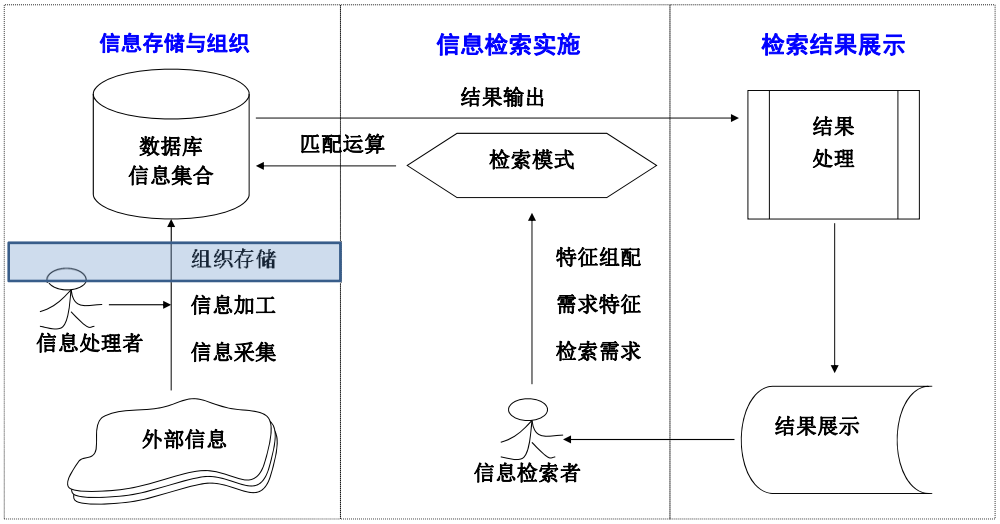
- 马张华：信息组织，亦称为信息资源组织，是根据信息检索的需要，以文本及各种类型的信息资源为对象，通过对其内容特征等的分析、选择、标引、处理，使其成为有序化集合的活动。
- 周宁：信息组织，即信息序化或信息整序，也就是利用一定的科学规则和方法，通过对信息外在特征和内容特征的描述和序化，实现无序信息流向有序信息流的转换，从而保证用户对信息的有效获取和利用及信息的有效流通和组合。



3

信息检索

信息组织在信息检索中的位置



4

信息检索

信息组织的内涵

- **信息组织**是以用户需求为导向，依据信息体自身的属性特征，按照一定的原则、方法和技术，将杂乱无章的信息整理成为有序的信息集合的**活动和过程**。
- **信息组织**的结果是形成各种方便用户利用的有序化的**信息检索系统**，从而达到**信息增值**的目的。
- **信息组织**是信息资源管理的核心和关键性环节，也是信息检索利用的基础。

5

信息检索

信息组织的内容

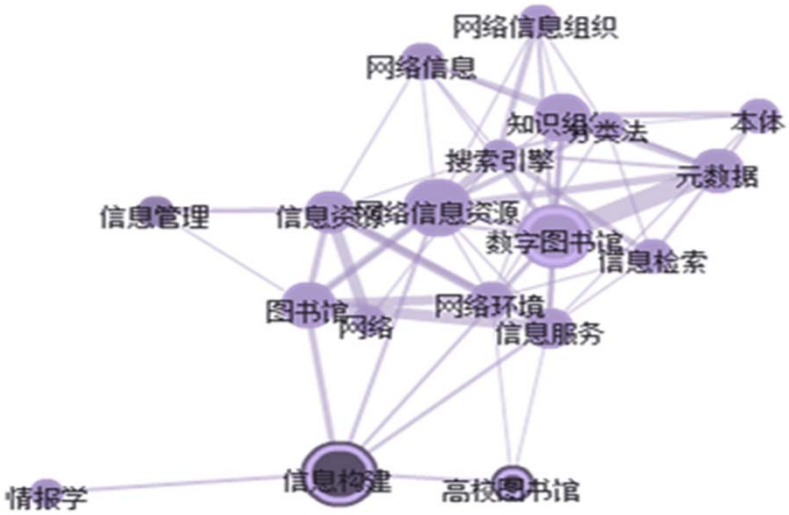
– **信息著录和标引**（信息描述与揭示）
  - **信息著录**实际上是对原始信息的外部属性特征（题名、著者、摘要等）进行描述的过程；
  - **信息标引**是给出信息内容标识（如分类号、主题词等）的揭示过程；
  - 著录、标引的结果是将原始信息制成它的替代记录——二次信息（**元数据**）。

– **信息序化**（信息整序）
  - **信息序化**则是将所有替代记录按照其某种**外部特征**（如著者、题名）和**内容标识**（如分类号、主题词）进行有规律的组织排列，从而构成某种序列（某种目录或索引），各种序列（目录或索引）制作完成并存储以后，就形成了比较完整的检索系统。

6

信息检索

信息组织的研究热点



2018.3 © CNKI

7

信息检索

信息组织发展概要

古代
- 图书编目
- 类书

近代
- 科学分类
- 主题表
- 文摘与索引

现代
- 分类主题一体化
- 自然语言检索

信息检索

标引

标引（Indexing），顾名思义，**标**是标记，**引**是指引，就是通过标记指引人们方便、快捷地找到所需要的信息。

通过对信息的分析，选用确切的检索标识（类号、标题词、叙词、关键词、人名、地名等），用以**反映该文献的内容**的过程。主要指选用检索语言词或自然语言词反映文献主题内容，并以之作为检索标识的过程。

标引是手段，检索是目的，标引为检索服务。


基于LSTM模型的中文图书多标签分类研究*

中图号
TP391

关键词
LSTM
深度学习
……

信息检索

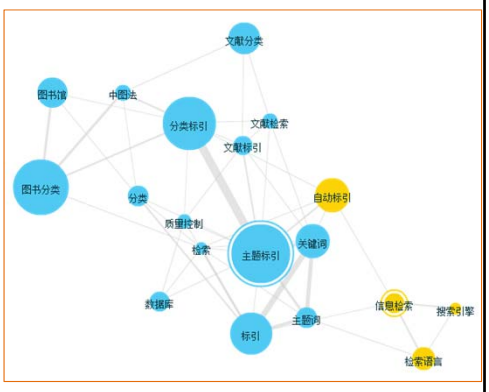
标引的分类

按智能程度

- 人工标引
- 辅助标引
- 自动标引

按标引属性

- 分类标引
- 主题标引
- 代码标引



信息检索

标引的过程

主题分析 → 排重 → 标引 → 审核 → 记录结果

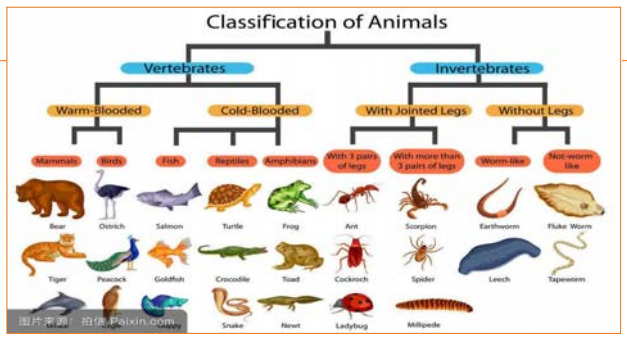
11

信息检索

分类标引

分类标引又称为归类

是依据一定的分类语言，对信息资源的内容特征进行分析、判断、选择，赋予分类标识的过程。



12

信息检索

分类标引要求

- **准确：**一是归类准确，将信息资源归入与其内容相对应的学科和专业；二是归类确切，归入分类体系中最专指、最切合其内容的类目。
- **充分：**指标引者能根据用户使用的需要，充分揭示有检索价值的信息资源的主题。
- **一致：**指对同一主题内容资源的标引结果应保证一致，提高系统的查全率和查准率。
- **适用：**分类标引时应兼顾系统的特点和用户的检索需要，使标引结果适合使用。

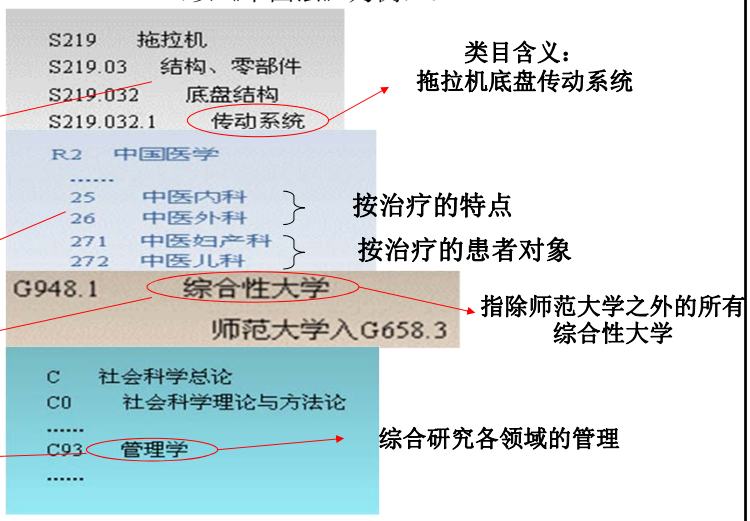
13

信息检索

分类标引方法

- 分类标引的主要方法有：
  - （1）**类目辨析：**
    - 根据上位类和下位类的关系了解类目的含义；
    - 根据同位类的关系了解类目的含义；
    - 根据类目注释了解相关类目的含义和范围；
    - 按照类目体系展开的规律了解类目的含义。

（以《中图法》为例）：

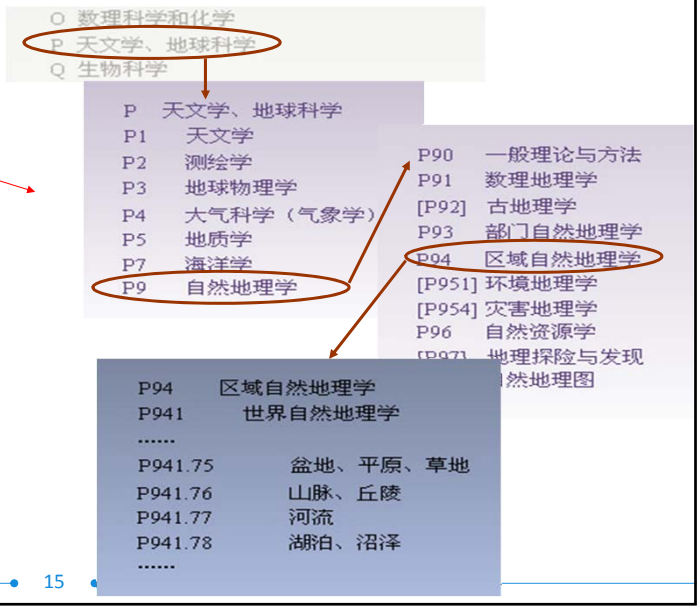


14

信息检索

分类标引方法

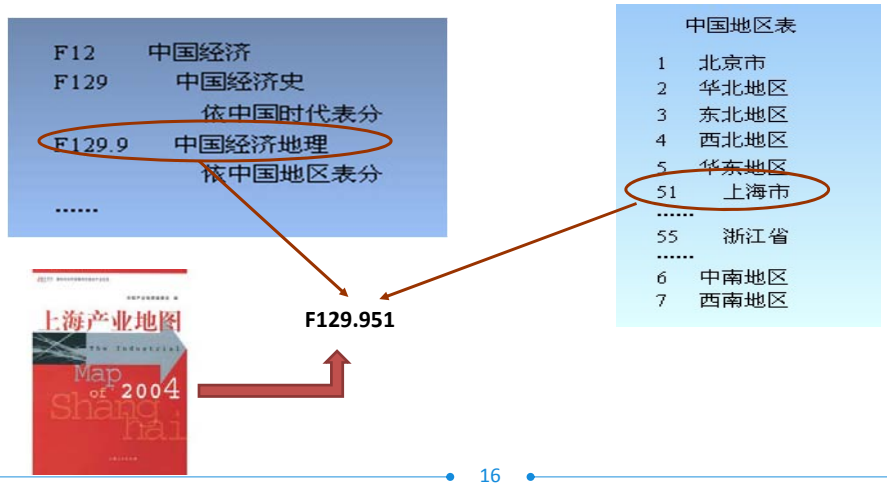
- 分类标引的主要方法有：
  - （2）**号码配置：**
    - 直接获得完整的分类号码：根据信息资源的内容及其特点，采用层层区分的方法，确定其在分类体系中的确切位置，就可得到相应的分类号码。
    - 通过不同号码的组配完成：
      - **使用复分表复分**
      - **仿分**
      - **类间组配**



信息检索

分类标引方法

- 通过不同号码的组配完成：
  - 使用**复分表复分**



信息检索

分类标引方法

- 通过不同号码的组配完成：
  - 仿分：
    - » 仿邻近类目分：指一组**相邻的类目**以相同的分类标准展开时，一般将在前的一个类目详细展开，后面的类目不再展开列举，而是依照已展开的子目细分的形式。 例如，“H32/37 各种常用外国语”下的注释“均可仿H31（英语）分”。
    - » 仿总论性类目分：**专论性类目仿照总论性类目的划分标准细分**。例如，“K21/27 中国各代史”均可仿K20（中国通史）分。
  - 类目仿分的要点：
    - » 按注释规定的范围使用；
    - » 按注释规定的复分次序复分；
    - » 当被仿分的用“/”号连接，且采用借号编号时，部分类目将涉及配号的转换问题。
    - » 注意越级仿分的问题：在标引工作中，根据文献实际论述的主题不需要按类目注释要求依次进行仿分，当跨越规定的某一层次，再继续依其他标准复分或仿分时，须在该复分号或仿分号前加“0”，以保证类目展开的逻辑次序。
    - » 注意复分依据的转换：某些类目凡属各国仿中国分，又涉及时代属性的，应同时将“依中国时代表分”改换成“依国际时代表分”。

17

信息检索

分类标引方法

- 通过不同号码的组配完成：
  - 类间组配

O 数理科学和化学
O1 数学
O3 力学
O4 物理学
O6 化学
O7 晶体学

Z89:O4

Z 综合性图书
Z1 丛书
Z2 百科全书、类书
Z3 辞典
Z4 论文集、全集、选集、杂著
Z5 年鉴、年刊
Z6 期刊、连续性出版物
Z8 图书目录、文摘、索引
Z81 国家总目录
……
Z89 文摘、索引
综合性文摘、索引入此；专科、专题的文摘也入此。
按本本类法体系分，即将各学科的分类号码加于本分类号之后，用组配符号“:”组合。例：化工文摘为Z89:TQ
如愿入有关各类，可在各学科的类号后再加总论复分号-7

18

信息检索

分类标引规则

- 分类标引规则包括：
  - 基本标引规则
  - 一般标引规则
  - *特殊标引规则*

19

信息检索

分类标引规则

- **基本标引规则（基本分类规则）：**
  - （1）信息资源的分类应根据信息资源的性质，按照其各自的特点进行：
    - 科学性质的资源，一般应以其**内容属性**为主要依据，同时兼顾其他特征
    - 文学、艺术形式的资源，通常应根据其特点，则按照其**体裁**、**形式**等标引
    - 特定类型的信息资源，应根据**类目体系的规定**和**使用需要**，同时按**内容**和**形式**归类
  - （2）信息资源的分类必须能体现分类法的**逻辑性、等级性、次第性**
  - （3）信息资源必须归入最切合其内容的类
  - （4）类分的文献必须归入用途最大的类
  - （5）不能单凭题名、篇名的意义归类
  - （6）应注意标引的思想性（揭示内容性质）


20

信息检索

分类标引规则

- 一般标引规则（一般分类规则）：
  - （1）单主题信息资源的分类标引：论述某一特定事物对象
    - 简单单主题信息资源的标引：一般按照主题对象的学科性质归类
    - 方面单主题信息资源的标引：根据论述的方面及各个方面关系归类
  - （2）多主题信息资源的分类标引：论述两个以上事物对象
    - 并列关系主题信息资源的标引：按重点或在前主题归类
    - 从属关系主题信息资源的标引：按大主题归类，也看论述重点
    - 联结关系主题信息资源的标引：按各自特点进行标引
  - （3）丛书、多卷书的分类标引
  - （4）词典、百科全书、年鉴、手册的分类标引
  - （5）目录、索引的分类标引
  - （6）关于对著作的研究、注释的标引
  - （7）特种文献的分类标引
  - （8）非书资料的分类标引
  - （9）网络信息资源的分类标引

21

信息检索

主题标引

- 主题标引是依据一定的主题词表或主题标引规则，赋予信息资源语词标识的过程。
- 具体而言，是在主题分析的基础上，以一定的主题词表或主题标引规则作为依据，将信息资源中具有检索意义的特征转化成相应的主题词，并将其组织成表达信息资源内容特征的标识的过程。

22

信息检索

主题标引方式

- 不同的标引方式，对主题揭示的深度不同。要根据检索系统的设备条件、资源特点、收藏范围、用户需求、标引种类等具体情况选择。常见的主题标引方式：
  - 整体标引：浅标引，概括揭示信息资源的基本主题内容，只揭示其整体性主题（单、多），不涉及从属性主题。适用范围：手工检索，综合性馆藏单位对图书的标引。
  - 全面标引：深标引，充分揭示信息资源所有有检索价值的主题，有利于提高查全率。适用范围：机检系统，专业论文，技术报告。
  - 对口标引：重点标引。只揭示适应其专业需要的主题，进行标引。
  - 综合标引：是一种整体标引。以整套文献为单元（涉及到多个信息单元）进行标引
  - 分析标引：与整体标引或综合标引结合使用。对信息资源中的部分片段或部分构成单元进行标引，利于揭示重要的、特殊的、有检索意义的主题。

23

信息检索

主题标引方法

- 1）主题分析方法
  - 主题类型的分析：
    - 单主题和多主题
    - 单元主题、复合主题和联结主题
    - 主要主题和次要主题
    - 专业主题和非专业主题
    - 显性主题和隐性主题
  - 主题结构的分析（确定主题结构的模式，也就是引用次序）：
    - 显著性引用次序：事物——部件——材料——活动——施动者；
    - 范畴职能引用次序：本体——物质——动力——空间——时间；（印度，阮冈纳赞）

24

信息检索

主题标引方法

- 2）主题概念的转换：
  - 直接转换：指分析出来的主题概念可以直接用主题词表上的一个对应主题词加以表述。
  - 间接转换：指分析出来的主题概念在主题词表中没有现成的主题词直接表达，需将复杂主题概念分解成若干个基本概念，再从主题词表中选取与基本主题概念相对应的主题词。

25

信息检索

例

江泽民同志在中央**计划生育**和**环境保护工作**座谈会上的讲话

- 多主题
- 直接转换：计划生育工作、环境保护工作、座谈会、讲话

中共中央国务院关于**治理**向企业**乱收费、乱罚款和各种摊派**等问题的决定

- 单主题
- 间接转换：治理、三乱

26

信息检索

主题标引方法

- 3）主题标识的确定：
  - 标题的确定：先确定标题的结构形式和级别
    - 单级标题：单词的、词组的、带限定词的；
    - 多级标题（复合标题）：有两个或以上的主题词通过一定的语义逻辑关系组配形成。
  - 机检系统中对词的处理（组配）：
    - 加联号（标识同一主题）
    - 加职号（职能符号，A：动作对象）
  - 标引词的著录：
    - 手工：著录在目录卡片上；将文献号记录在相应标目下，用于主题索引。
    - 机检系统应根据系统要求进行

27

信息检索

主题标引规则

- 包括**选词规则**和**组配规则**：
- 选用标引词的基本规则：
  - 1）正式词标引规则：用正式叙词，保持词形，注意参照系统中带“Y”是入口词；
  - 2）相对专指标引规则：表中刚好有表达所标引主题概念的词，不应选其他词，如：国际公法，一定要先假设词表中有所需最贴切的正式主题词；
  - 3）标引方案优先顺序原则：标引方案优先顺序原则的实施；
  - 4）适度标引规则，标引深度适当。手检检索系统的，分析出2~5个主题概念词即可，机检系统的，分析出4~10个或不限制；原则：各个主题概念应尽可能用足够的叙词完整式表达，需省略时，省必要性最小的词。
  - 5）一致标引规则：相同的主题概念应该用相同的主题词表达，同一标引人员在不同时间或不同的标引人员标引相同的主题概念或概念因素应确保一致。

28

信息检索

标引词组配规则

- 1）概念组配规则：要符合实际逻辑，如：肝外科手术；
- 2）交叉组配优先规则：主题分析先判断是否交叉型，如：畜牧气象学；
- 3）参组叙词相对专指规则：参组的词相对于被标引的主题概念应该是最专指的，不能用其上位或下位词，如：道路运输经营学，应用 公路运输—运输经济
- 4）合理组配规则：用若干主题词组配表达被标引的内容时，要保证组配的合理性，如：淡水养殖鱼类；
- 5）适度组配规则：考察每个参组叙词的有效性，既要防止冗余，又要防止缺词，如：北京工人食道癌的中医疗法、佛山教育；
- 6）明确组配语言规则：组配表达的结果必须明确单一，正确使用组配符号，熟悉组配基本词序；
- 7）分组标引规则：课题是多主题时要防止虚假组配，解决措施是分组，如：钢的抗拉性与铜的传电性；
- 8）不可组配规则：一是词表中有单个专指词时不得组配，如生物物理学，二是专有名词不用分解成普通名词再组配，如：河北工业大学。

29

信息检索

主题标引与分类标引的比较

- 标引对象相同，但揭示信息资源内容的角度不同（内/外）；
- 所使用的标识符不同，主题标引更具直观性，分类标引具有间接性。
- 由于体系结构不同，主题标引具有专指性、灵活性，分类标引具有系统性、稳定性。

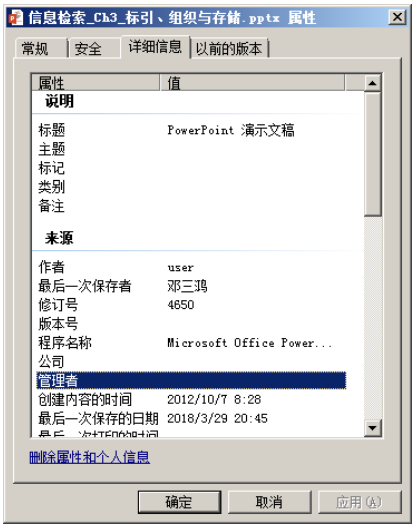
30

信息检索

信息描述

Information Description

- 根据一定的**规则和标准**，对信息资源的**形式特征和内容特征**进行描述并给予记录的过程。
- ***标引**是依据一定的标引规则，在对信息资源内容属性进行分析的基础上，给出其信息属性标识的过程。*

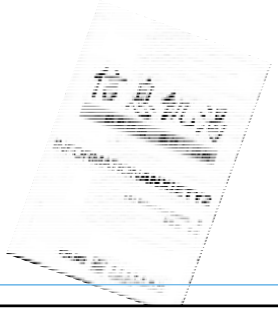


31

信息检索

相关概念-信息构建

- 信息构建（Information Architecture）是当前兴起的信息组织领域的一个热点，IA引入国内后有很多译名：信息构建、信息空间构建、信息建筑学、信息构筑、信息空间构筑、信息构筑体系，港台地区译为资讯架构；
- 美国建筑师Wurman1975年创造的一个新概念。其初衷是看到了信息的收集、组织、整理、表达和提供与建筑师从事的开发设计工作有相似之处，因此把信息（Information）与建筑（Architecture）有机结合，视为一体。
- Wurman在其专著Information Anxiety中将信息构建简要定义为：“**指组织、标识、导航和检索系统的设计，目的是帮助用户查找和管理信息**”。



32

信息检索

文献著录标准

GB/T 3792-1983 -> GB/T 3792-2009

- 题名和责任说明项
- 版本项
- 文献特殊细节项
- 出版发行项
- 载体形态项
- 丛编项
- 附注项
- 标准编号及获得方式项

GB/T 3792《文献著录》共分 9 部分：
第 1 部分：总则；
第 2 部分：普通图书；
第 3 部分：连续性资源；
第 4 部分：非书资料；
第 5 部分：档案；
第 6 部分：测绘制图资料；
第 9 部分：电子资源。

33

信息检索

编目

图书馆编目工作就是对文献资源进行分类、编制目录、建立馆藏目录体系的过程。

包含：印前编目、在版编目、集中编目、联合编目等概念

士兵无编组，军旅难有纲纪；典籍无目录，读者何从求索。是图书固不可以无编目

来新夏（1923-2014）

34

信息检索

印前编目与CIP、联合编目的区别

区别方面	印前编目	CIP	联合编目
编目机构	出版发行领域与图书馆领域联合	新闻出版总署信息中心主导、出版单位辅助	图书馆领域内联合
编目主体	出版发行及图书馆领域按专业虚拟分组的编目员	新闻出版总署信息中心编目员	成员馆编目员
编目对象	电子排版“红样”	图书“清样”、“工作单”	纸本图书
编目时间	印刷前	出版过程中	出版后
编目目的	消除重复编目	书目的控制与审查	减少重复编目
编目用途	修正CIP，更广范围的查询、检索及书目发布。	主要为申报、审校	成员馆及用户查询、检索

35

信息检索

CIP

- 图书在版编目 Cataloguing in Publication

- 依据一定的标准，为在出版过程中的图书编制书目数据。
- 需要依据相关的国家标准《普通图书著录规则》、《文献叙词标引规则》以及《中国图书馆图书分类法》和《汉语主题词表》对图书进行著录、分类标引、主题标引。通常印刷在图书书名页背面或版权页上方。

36

信息检索

CIP组成示例

- 图书在版编目数据标题
- 著录数据
- 检索数据
- 其他注记

37

信息检索

Marc

- **Machine Readable Catalog**
- “机器可读目录”，即以代码形式和特定结构记录在计算机存储载体上的、用计算机识别与阅读的目录。
- MARC格式最早由美国国会图书馆研制，始于20世纪60年代。1973年国际标准化组织将MARC格式作为国际标准正式颁布，即现在所说的USMARC格式，或称LCMARC，现在已改名为MARC21。

38

信息检索

Marc的特点

- 字段数量多
- 著录详尽
- 可检索字段多
- 定长与不定长字段结合，灵活实用
- 保留主要款目及传统编目的特点
- 可扩充、修改，并能在实践中不断发展完善

39

信息检索

CNMARC格式概要

中文普通图书机读目录的信息录入

- **0-- 标识块**
- **1-- 编码信息块**
- **2-- 著录信息块**
- **3-- 附注块**
- **4-- 连接款目块**
- **5-- 相关题名块**
- **6-- 主题责任块**
- **7-- 知识责任块**
- **8-- 国际使用块**
- **9-- 国内使用块**
- **记录头标**

参照标准：《新版中国机读目录格式使用手册》，北京图书馆出版社，2004

40

信息检索

CNMARC示例

MARC格式 | 卡片格式 | 典藏数据 | 流通数据 | 期刊签到 | 期刊装订 | 外借历史 | 丛编数据 | 特借数据 |

字段名称	标识	指	字段内容
头标区			oam2
记录标识	001		012015000000
处理时间	005		20150702121212.0
ISBN	010		@a978-7-5180-0476-8@dCNY30.00
处理数据	100		@a20150702d2014 em y0chiy50 ea
作品语种	101	0	@achi
出版国别	102		@aCN@b110000
编码数据	105		@ay z 000yb
形态特征	106		@ar
题名责任	200	1	@a天使在人间@Atian shi zai ren jian@e赫本传奇@f白皙卉著
出版发行	210		@a北京@c中国纺织出版社@d2014
载体形态	215		@a238页@d23cm
提要文摘	330		@a本书记述了著名影星奥黛丽·赫本的传奇人生，内容包括：被上帝亲
其它题名	517	1	@a赫本传奇@Ahe ben chuan qi
个人主题	600	1	@a赫本@g(Hepburn, Audrey),@f1929-1993@x传记
中图分类	690		@aK837.125.78=536@v5
人名等同	701	0	@a白皙卉@Abai xi hui@4著
记录来源	801	0	@aCN@bWXLS@c20150702
馆藏信息	905		@aSCPCFE@b00651176-7@dK837.125.78=536@e4450@y2014@f2

机读格式显示(MARC)

```
000 01147nam 2200337 450
001 0003652536
005 20111017154900.0
010 __ |a 978-7-305-07263-5 |d CNY62.00
092 __ |b 南大11-02-076 |a CN
099 __ |a CAL 012010218419
100 __ |a 20101018d2010 em y0chiy0121 ea
101 0_ |a chi
102 __ |a CN |b 320000
105 __ |a a z 001yy
106 __ |a r
200 1_ |a Internet实用技术 |A Internet Shi Yong Ji Shu |f 主编邓三鸿, 苏新宁
210 __ |a 南京 |c 南京大学出版社 |d 2010
215 __ |a 540页 |c 图 |d 23cm
225 2_ |a 信息管理研究型课程书系 |A Xin Xi Guan Li Yan Jiu Xing Ke Cheng Shu Xi
320 __ |a 有索引
330 __ |a 本书共分12个章节，主要对Internet实用技术知识作了介绍，内容包括Internet常识及相关概念
410 _0 |1 2001 |a 信息管理研究型课程书系
606 0_ |a 互联网络 |A Hu Lian Wang Luo |x 高等学校 |j 教材
690 __ |a TP393.4-43 |v 4
690 __ |a TP393.4 |v 4
701 _0 |a 邓三鸿 |A Deng San Hong |4 主编
701 _0 |a 苏新宁 |A Su Xin Ning |4 主编
801 _0 |a CN |b ZJU |c 20101018
905 __ |a NUL |d TP393.4/H727
920 __ |a 232010 |z 1
998 __ |a ZJU
```

41

信息检索

Marc的局限

- 多用于图书馆
- 专用软件来处理，不适用于互联网
- 格式复杂，修订麻烦
- 适用完整、静态的信息，不易处理动态信息
- 编制记录需要专业人员，耗时

42

信息检索

元数据

MARC格式 | 卡片格式 | 典藏数据 | 流通数据 | 期刊签到 | 期刊装订 | 外借历史 | 丛编数据 | 特借数据

字段名称	标识	指	字段内容
头标区			oam2
记录标识	001		012015000000
处理时间	005		20150702121212.0
ISBN	010		@a978-7-5180-0476-8@dCNY30.00
处理数据	100		@a20150702d2014 em y0chiy50 ea
作品语种	101	0	@achi
出版国别	102		@aCN@b110000
编码数据	105		@ay z 000yb
形态特征	106		@ar
题名责任	200	1	@a天使在人间@Atian shi zai ren jian@e赫本传奇@f白皙卉著
出版发行	210		@a北京@c中国纺织出版社@d2014
载体形态	215		@a238页@d23cm
提要文摘	330		@a本书记述了著名影星奥黛丽·赫本的传奇人生，内容包括：被上帝亲
其它题名	517	1	@a赫本传奇@Ahe ben chuan qi
个人主题	600	1	@a赫本@g(Hepburn, Audrey),@f1929-1993@x传记
中图分类	690		@aK837.125.78=536@v5
人名等同	701	0	@a白皙卉@Abai xi hui@4著
记录来源	801	0	@aCN@bWXLS@c20150702
馆藏信息	905		@aSCPCFE@b00651176-7@dK837.125.78=536@e4450@y2014@f2

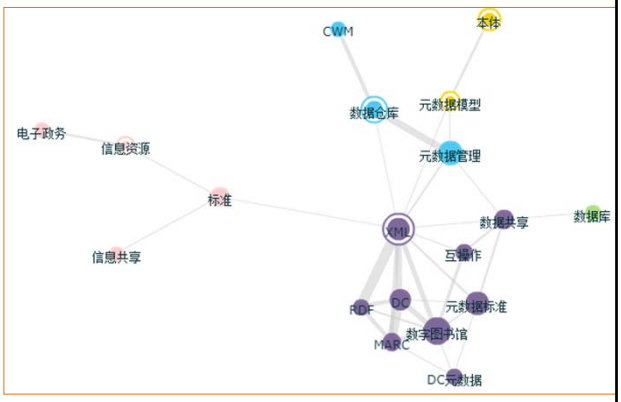
姓 名		性别		出生年月日		照片
籍 贯	省 市（县）		政 治 面 貌		民族	
研 究 生 入 学 年 月		学制		攻读学位	硕士	
所在单位		所学专业		研究方向		
大 学 原 毕 业 学 校		大 学 毕业年月		所学专业		
原工作单位		职务		职称		
学 号		身份证号				
联系电话		手 机 号				
学历	起 止 年 月	学习或工作单位		学生或职务、职称		

信息检索

元数据概念

Metadata

- 关于数据的数据；数据属性描述；
- **data that describes data**
- **data about data**
- **structured data about data**
- **what, where, when, who, how……about data**
- **information about a resource**
- **cataloguing information**
- **data that defines and describes other data （ISO/IEC 11179-1:2015）**



信息检索

元数据的功能

- 支持资源发现 （Resources Discovery）
- 组织数字信息资源 （Digital Resources）
- 支持资源的互操作 （Interoperability）
- 支持数字识别 （Digital Identification）
- 支持存档和保存 （Archiving and Preservation）

45

信息检索

元数据类别

资源类型/应用领域	元数据方案
网络资源	Dublin Core、IAFA Template、CDF、Web Collections
文献资料	MARC（with 856 Field），Dublic Core
人文科学	TEI Header
社会科学数据集	ICPSR SGML Codebook
博物馆与艺术作品	CIMI、CDWA、RLG REACH Element Set、VRA Core
政府信息	GILS
地理空间信息	FGDC/CSDGM
数字图像	MOA2 metadata、CDL metadata、Open Archives Format、VRA Core、NISO/CLIR/RLG Technical Metadata for Images
档案库与资源集合	EAD
技术报告	RFC 1807
连续图像	MPEG-7

46

信息检索

DC元数据:Dublin Core Element Set

http://dublincore.org/about/history/
http://dc.library.sh.cn

- 题名 Title
- 日期 Date
- 创建者 Creator
- 主题Subject
- 出版者Publisher
- 类型Type
- 描述Description
- 其他责任者Contributor
- 格式Format
- 来源Source
- 权限Rights
- 标识符Identifier
- 语种Language
- 关联Relation
- 覆盖范围Coverage

47

信息检索

DC示例

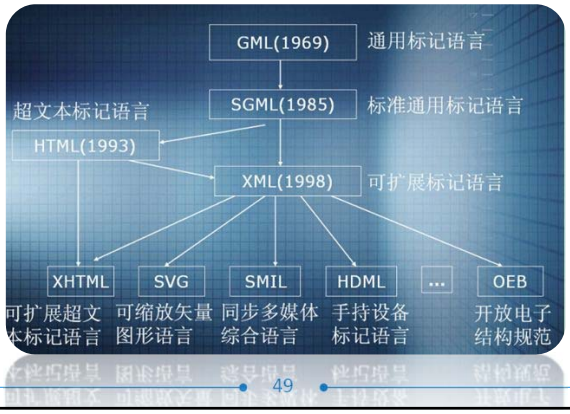
```
<?xml version="1.0" encoding="GB2312" ?>
- <rdf:RDF xmlns:rdf="http://www.w3c.org/1999/02/22-synax-ns#"
    xmlns:dc="http://purl.org/metadata/dublin-core#"
    xmlns:dced="http://purl.org/metadata/dced-v1#">
  - <rdf:Description ID="0310100-A-2">
      <dc:title>神州五号发射成功</dc:title>
      <dc:creator>新华社</dc:creator>
      <dc:subject>中国飞天梦圆</dc:subject>
      <dc:publisher>新华网</dc:publisher>
      <dc:date>2003.10.21</dc:date>
      <dc:language>chi</dc:language>
      <dc:description>神州5号于2003年10月19日在酒泉航天中心发射成功.</dc:description>
      <dc:relation>http://news.xinhuanet.com/audio/2003-10/21/</dc:relation>
      <dc:source>http://www.xinhuanet.com.cn/</dc:source>
      <dc:coverage>神州5号 中国飞天 千年飞天梦</dc:coverage>
      <dc:type>图片</dc:type>
      <dc:format>jpeg</dc:format>
      <dc:identifier>xinsrc_f33e66a1a94d48aeb853bc2f953d20d2.jpg</dc:identifier>
      <dc:right>新华网</dc:right>
      <dc:contributor>新华社</dc:contributor>
      <dced:audience>teacher</dced:audience>
      <dced:level>5</dced:level>
    </rdf:Description>
</rdf:RDF>
```

48

信息检索

标记语言

- Markup Language
- 标记分为两种：一种称为“程序性的标记（Procedard Markup ）”，用来描述文档显示的样式；另一种称为“描述性标记（Descriptive Markup ）”，用来描述文档中的文字的用途。制定“通用标言”的基本思想是把文档的内容与样式分开。



49

信息检索

SGML

- Standard Generalized Markup ·anguage
- SGML以成分和属性规定数据。
- 成分——指明文本的特定单元，如：题名、作者、章节名、段落等。
- 属性——指明成分的特定信息。如注明作者为邓三鸿。
- SGML中对成分的描述由定义符和标识组成。
  - 定义符——用来定义的符号，如：<，>，</,“ 等，可用来结构标识，例如：<author>即是。
- SGML对数据的表述：
- <作者>邓三鸿</作者> *<author>Deng Sanhong</author>*
- <标识>属性</标识>

50

信息检索

SGML特点

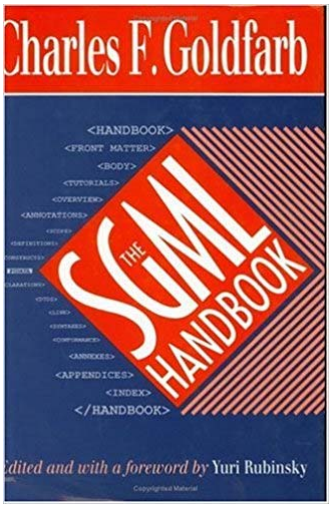
https://www.w3.org/MarkUp/SGML/

优点

- 高稳定性
- 高可携性
- 高完整性

缺点

- 高复杂性
- 费用昂贵



51

信息检索

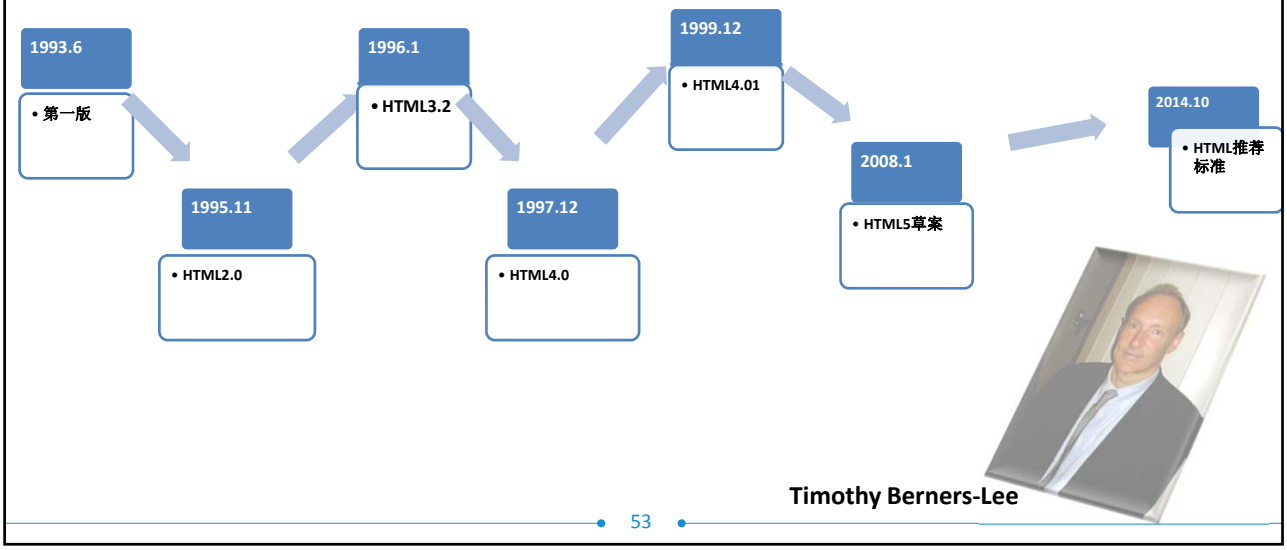
HTML

- Web使用的语言就是HTML
- HTML超文本标记语言（HyperText Markup Language）
- 应用：
  - 出版联网文档，这种文档可包含标题、文字、表格、列表、图像以及声音和影视文件等
  - 通过超文本链接可以检索和阅读联网信息
  - 设计交易单（form）。这是一种用来从读者处收集信息的Web文档，可以与远程服务单位作交易
- HTML使用预先定义的标签（tag）来描述网页中的元素
- HTML**不是编程语言**；纯文本格式

52

信息检索

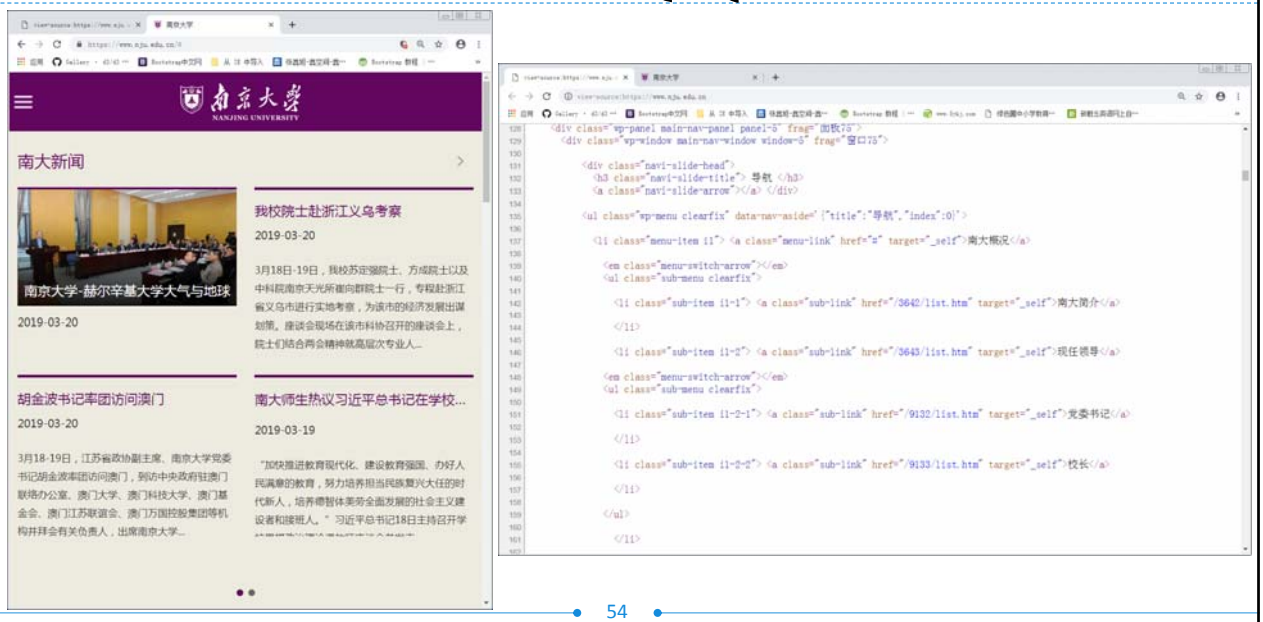
关于HTML



53

信息检索

HTML示例



54

信息检索

XML

- XML 指可扩展标记语言（eXtensible Markup Language）
- XML 是一种标记语言，类似 HTML
- XML 的设计宗旨是传输数据，而非显示数据
- XML 标签没有被预定义。需要自行定义标签。
- XML 被设计为具有自我描述性。
- XML 是 W3C 的推荐标准

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<note>
<to>son</to>
<from>sun</from>
<heading>Reminder</heading>
<body>Don't forget the breakfast!</body>
</note>
```

55

信息检索

XML的作用

- 把数据从HTML中独立
- 简化数据共享
- 简化数据传输
- 简化平台的变更
- 数据用途更广泛
- 创建新的标记语言

56

信息检索

HTML vs XML

比较内容	HTML	XML
可扩展性	不具有扩展性	是源描述语言，可用于定义新的描述语言
侧重点	侧重于如何表现信息	侧重于如何结构化地描述信息
语法要求	不要求标记的嵌套、配对等 不要求标记之间具有一定的顺序	严格要求嵌套、配对和遵循 DTD 的树形结构
可读性及可维护性	难于阅读、维护	结构清晰，便于阅读、维护
数据和显示的关系	内容描述与显示方式整合为一体	内容描述与显示方式相分离
保值性	不具有保值性	具有保值性
比较内容	HTML	XML
编辑及浏览工具	已有大量的编辑、浏览工具	编辑、浏览工具尚不成熟

57

信息检索

DTD

Document Type Definition

- **DTD是一套关于标记符的语法规则**
  - DTD是一种保证XML文档格式正确的有效方法，可以通过DTD文件来看XML文档是否符合规范，元素和标签使用是否正确。
- **DTD规定了语法分析器解释所有细节**
  - DTD文档包含：元素的定义规则，元素间关系的定义规则，元素可使用的属性，可使用的实体或符号规则。
- **可以是XML文档的一部分，但通常是一份单独的或一系列的文档**
  - DTD文件也是一个ASCII的文本文件，后缀名为.dtd
- **使用DTD最大的好处在于DTD文件的共享**
  - 想使用XML进行数据交换的行业或组织可定义自己的DTD

58

信息检索

DTD示例-内嵌

```
<?xml version='1.0' encoding='gb2312'?>
<!DOCTYPE book[                                  <----------根元素的名称
<!ELEMENT book (author,title, description) >     <----------子元素的名称及顺序
<!ELEMENT author (#PCDATA)>                      <----------子元素的数据类型
<!ELEMENT title (#PCDATA)>
<!ELEMENT description (#PCDATA)>
]>                                               <----------结束标签

<book>
   <author>邓三鸿</author>
   <title>信息检索</title>
   <description>南京大学信息管理学院通用教材</ description>
</ book >
```

59

信息检索

DTD示例-外部文件

```
<? xml version='1.0' encoding='gb2312' ?>
<!DOCTYPE book SYSTEM "ex2.dtd">
<book>
   <author>邓三鸿</author>
   <title>信息检索</title>
   <description>南京大学信息管理学院通用教材</description>
</ book >
```

ex2.dtd

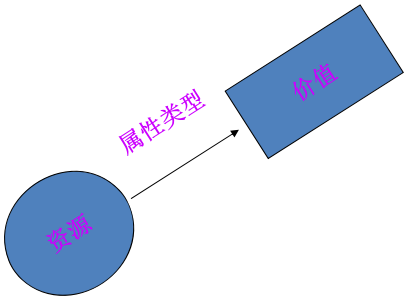
```
<?xml version="1.0" encoding="gb2312"?>
<!ELEMENT book (author,title, description)>
<!ELEMENT author (#PCDATA)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT description (#PCDATA)>
```

60

信息检索

RDF

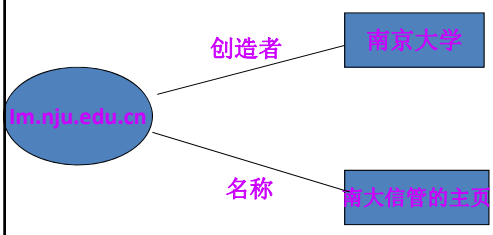
- 资源描述框架（Resources Description Framework，RDF）主要是根据软件内部结构的一般特点，为元数据的发展提供一个有效的模式，支持独立发展和维护元数据集合。通常为一个三元组：成分-属性-值。
- RDF为万维网集团（W3C）支持的描述结构；
- 支持元数据交换使用的规范；
- 以标准的XML形式表示；
- 采用既可以由人或由机器处理的方式；
- 由特定领域团体确定其内容。



61

信息检索

RDF模式示例



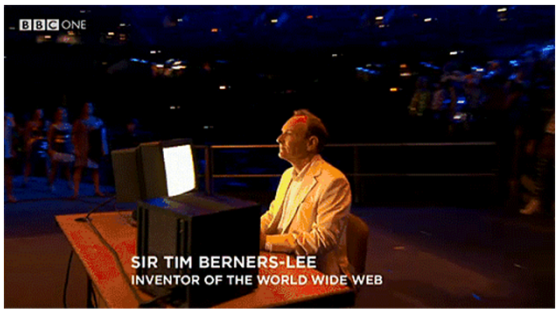

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
  syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.0/">
<rdf:Description about="http://im.nju.edu.cn>
<dc:Creator> 南京大学</dc:Creator>
<dc:Title南大信管的主页</dc:Title>
<rdf:Description>
<rdf:RDF>
```

62

信息检索

语义网与本体

1998年Berners-Lee首次提出语义网这一全新概念，并阐释了语义网的七层体系结构。语义网是指建立一个使用能够表达语义（或机器可处理）的元素来描述信息，以满足智能软件代理对异构、分布信息的有效访问、合理交换、语义处理和准确检索等要求的公开环境。



63

信息检索

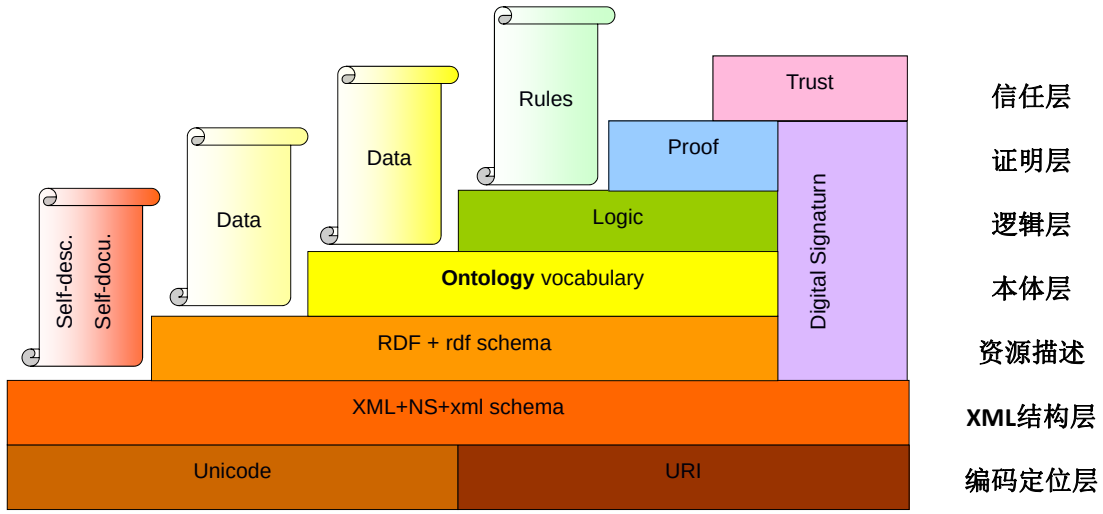
本体

- 本体（Ontology）原本是哲学领域中的概念，是对客观存在的系统解释，描述现实的抽象本质 。在20世纪90年代中期，本体被引入知识工程领域，用于描述知识的内涵，表达知识的语义。
- 一般认为，本体是共享概念模型的形式化规范说明，包含4方面的含义：概念模型、明确性、形式化和共享性 。
  - 本体是一种元数据，它提供丰富原语描述领域的概念模型，澄清领域知识的结构，具有知识表示的能力；
  - 本体可重用，避免了重复的领域知识分析；
  - 本体提供了大量受约束的、明确定义的、机器可处理的统一术语和概念，可以构建完整的“术语表”来定义网络中的数据，使知识共享成为可能；
  - 本体还能够对知识进行推理和验证。

64

信息检索

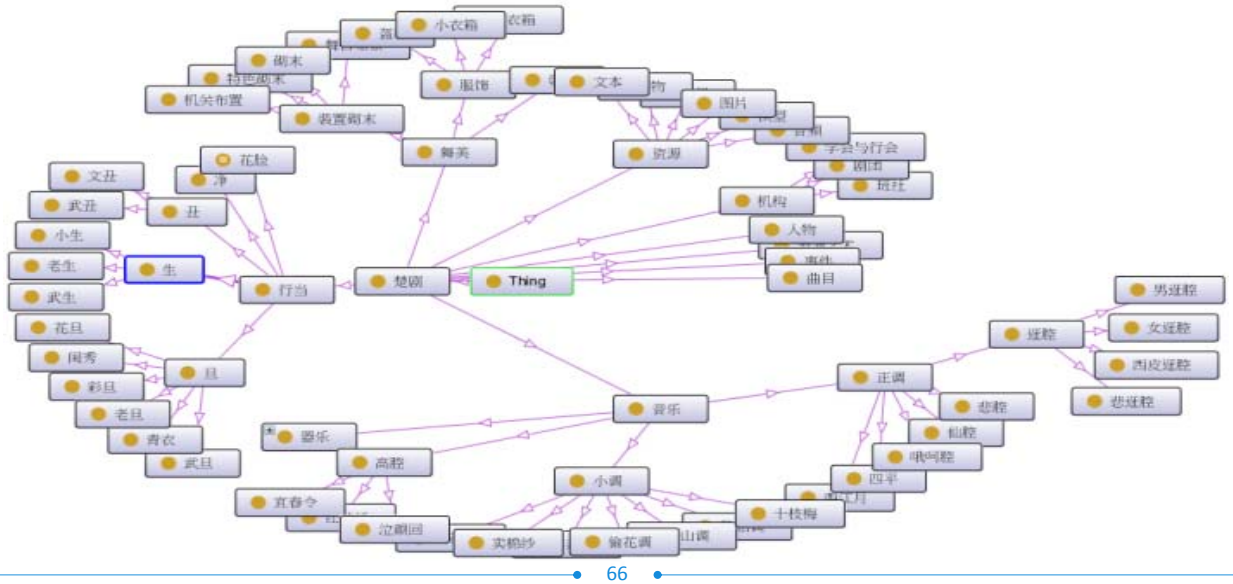
本体的体系结构



65

信息检索

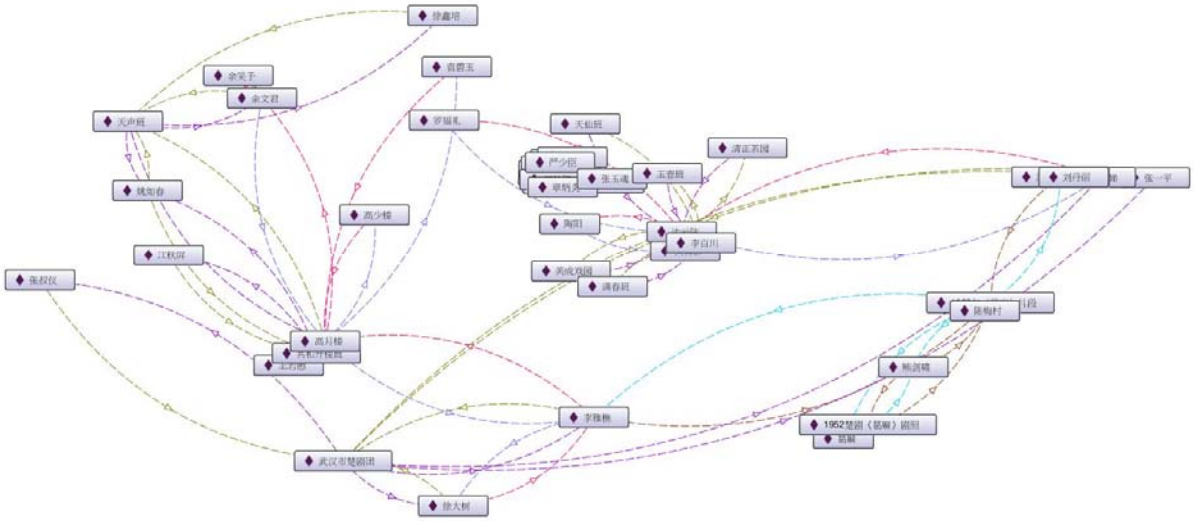
本体示例-概念关系



66

信息检索

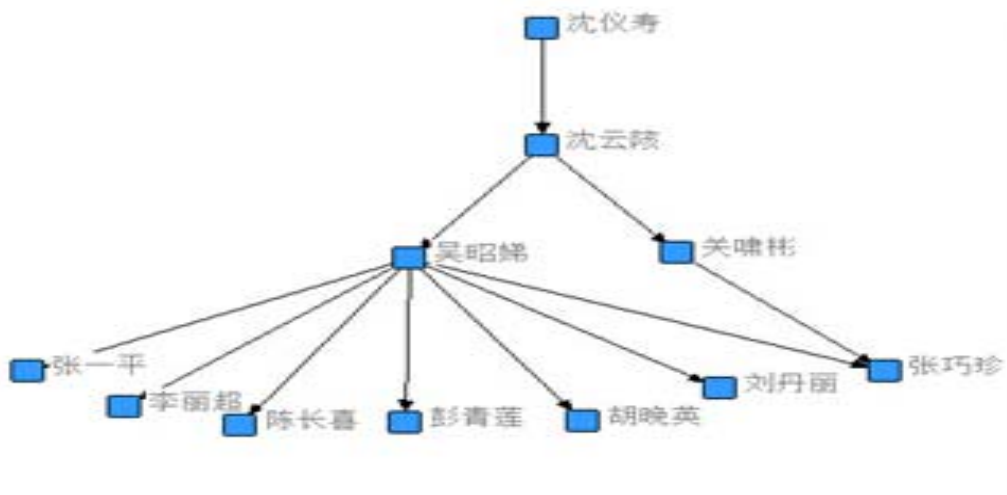
本体示例-可视化



67

信息检索

本体示例-推理



68

信息检索

本体与传统情报检索语言的比较

比较内容	本体	主题语言	分类语言
概念模型	面向对象的认识世界的方法	面向概念的信息表示与检索方法	面向学科的信息表示与检索方法
组成元素	类、属性、实例等	语词及词间关系	类目及类目关系
标识	URI唯一资源标识	语词	类目或类号
形式化程度	较高	较低	较低
概念关系表达	上百种关系	等同、等级、相关三种	包含、并列、交替、相关等关系
层级体系	存在，没有统一标准	有的存在，基本采用学科分类	存在，采用学科分类
适用对象	计算机	人为主，机为辅	人为主，机为辅
应用	语义检索和知识发现	信息内容的主题表示和检索	信息内容的分类表示和检索

69

信息检索

信息存储

- 经过加工整理序化后的信息按照一定的格式和顺序存储在特定的载体中的一种信息活动。目的是为了便于信息管理者和信息用户快速地、准确地**识别、定位和检索**信息。
- 存储的**基本价值**：实现信息跨越时间的传播
- 逻辑存储：文件、数据库、网络.....
- 物理存储（载体）：印刷型、缩微型、声像型、电子型、 .....

70

信息检索

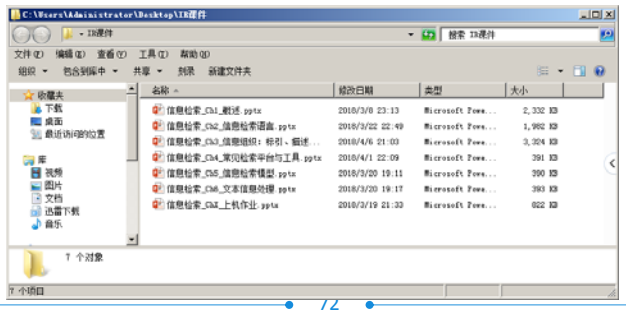
信息存储的要求

71

信息检索

信息的文件存储

- 文件是由创建者所定义、具有文件名的一组相关的信息集合。
- 文件的主要属性：
  - 文件名，文件类型，文件长度，创建者，创建时间，修改时间，文件定位信息 ，文件所包含的信息。
- 为了方便使用、管理系统公共程序和数据以及用户自己的程序和数据而引入文件。
- 为了对外存空间管理和对其上文件的按名访问而引入文件系统。



72

信息检索

文件存储的逻辑结构

- 操作系统感知文件信息的组织形式叫文件的逻辑结构。它包括流式文件（无结构文件）和记录式文件（有结构文件）两种，每种文件信息的逻辑单位分别是字节和记录。
- 流式文件（无结构文件）：
  - 是指对文件内信息不再划分单位，它是依次的一串字节流构成的文件。
- 记录式文件（有结构文件）：
  - 是用户把文件内的信息按逻辑上独立的含义划分信息单位，每个单位称为一个记录。所有记录通常都是描述一个实体集的，有着相同或不同数目的数据项，记录的长度可分为定长和不定长记录两类。

73

信息检索

信息的数据库存储

74

信息检索

文件与数据库：区别

(1)文件系统用文件将数据长期保存在外存上，数据库系统用数据库统一存储数据;

(2)文件系统中的程序和数据有一定的联系，数据库系统中的程序和数据分离;

(3)文件系统用操作系统中的存取方法对数据进行管理，数据库系统用DBMS统一管理和控制数据;

(4)文件系统实现以文件为单位的数据共享，数据库系统实现以记录和字段为单位的数据共享。

75

信息检索

文件与数据库：联系

(1)均为数据组织的管理技术

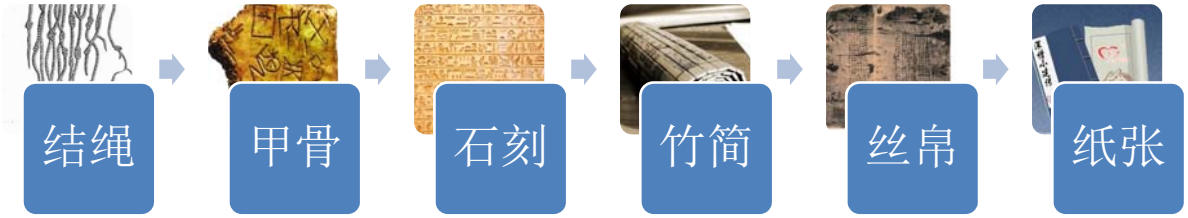
(2)均由数据管理软件管理数据，程序与数据之间用存取方法进行转换

(3)数据库系统是在文件系统的基础上发展而来的

76

信息检索

传统媒介



信息密度低，储藏空间大

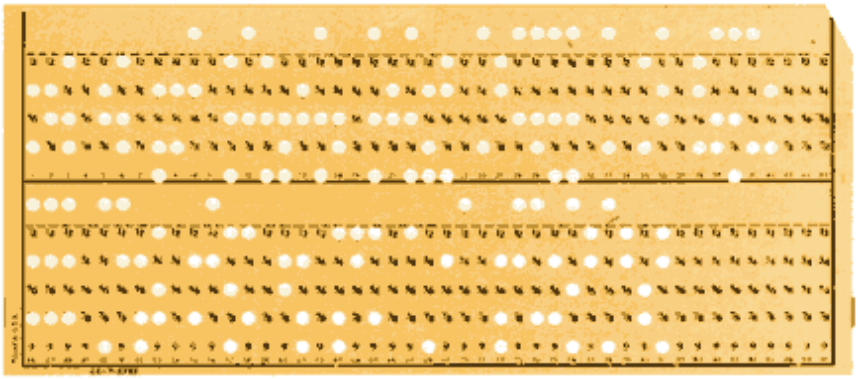
传播复制效率低

管理、应用繁琐

77

信息检索

打孔纸卡



1725年 -- Basile Bouchon
1884年9月23日 -- Herman Hollerith

78

信息检索

穿孔纸带

Alexander Bain（传真机和电传电报机的发明人）在1846年最早使用了穿孔纸带。纸带上每一行代表一个字符。

79

信息检索

胶卷（缩微胶卷）

使用照相技术，将图书报刊等记录有知识信息的一些载体在感光材料上拍摄成缩微影像的复制品，又称缩微品。

记录在缩微胶卷上的缩微影像通过一定的技术手段和方法，可以进行自动化处理、保存、检索、再现和复制还原。

80

信息检索

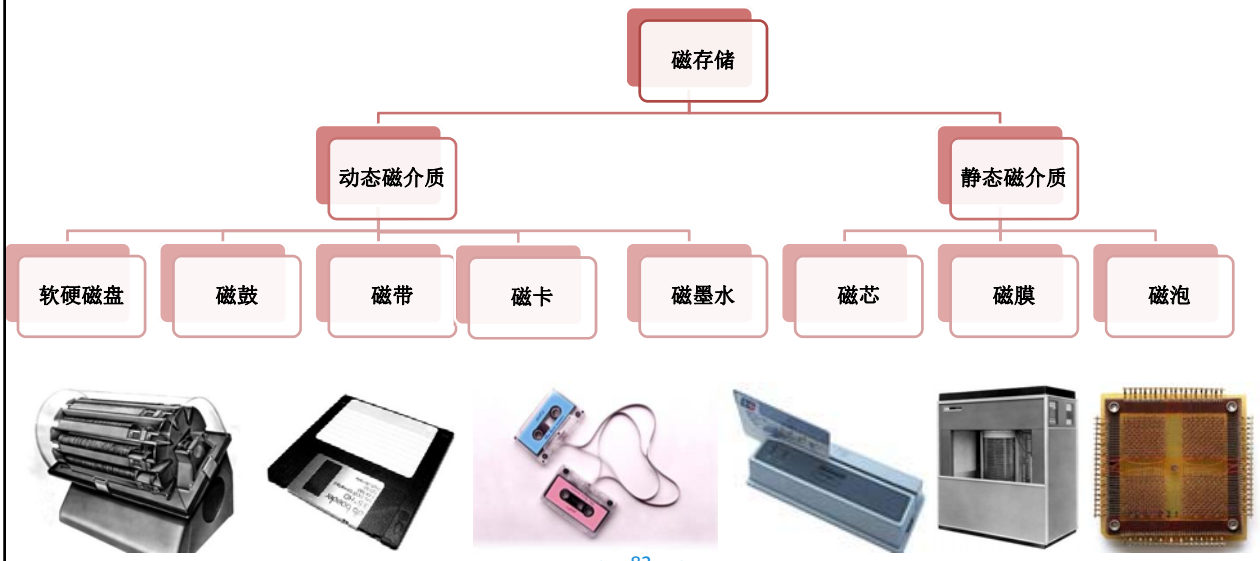
特点

- 优点
  - 缩小保存空间、存储时间长
  - 保护原件、反映原貌
  - 便于文献收集与交流
  - 便于信息的传递
  - 提高办公效率、具备法律凭证作用
- 不足
  - 主要是文字图像小，必须借助于一定的设备方可阅读和查检；
  - 制作、保存和使用条件严格，设备投资费用高

81

信息检索

磁存储



82

信息检索

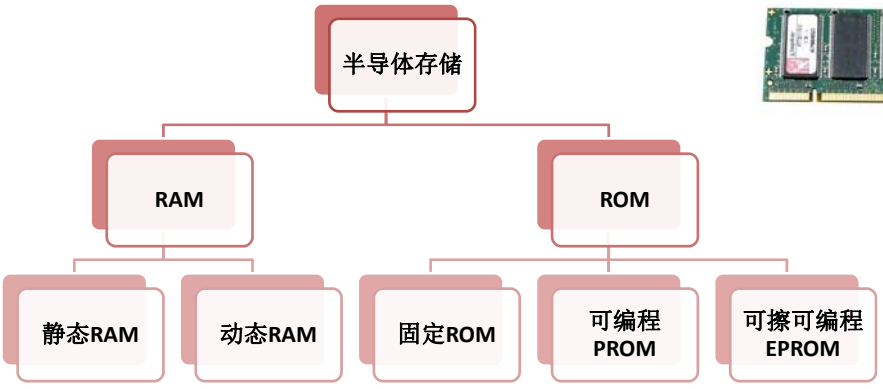
磁存储特点

- 优点：
  - 存储容量大，位价格低；
  - 记录介质可以重复使用；
  - 记录信息可以长期保存而不丢失，甚至可以脱机存档；
  - 非破坏性读出，读出时不需要再生信息。
- 缺点
  - 存取速度较慢
  - 机械结构复杂
  - 对工作环境要求较高。

83

信息检索

半导体存储



速度更快、体积小、可靠性高、易于批量生产
容易静电或者高电压损伤

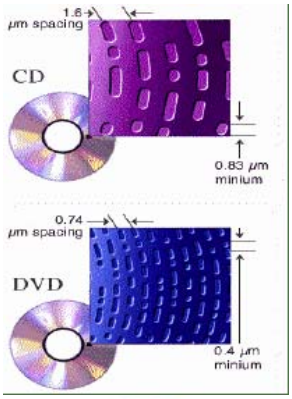
84

信息检索

光存储与光盘

分代	年代	名称	激光类型	存储容量
第1代	1982	CD光盘存储器	红外光	650MB
第2代	1995	DVD光盘存储器	红光	4.7GB
第3代	2006	BD光盘存储器	蓝光	25GB

（注：DVD和BD的容量均为单面单层的容量）

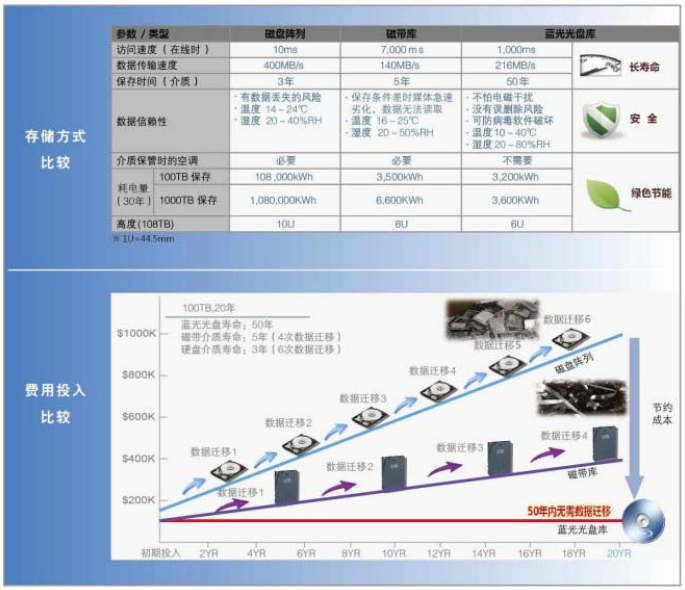


85

信息检索

光存储特点

- 光存储技术具有存储密度高、存储寿命长、非接触式读写和擦除、信息的信噪比高、信息位的价格低等优点
- 优点是在理论上能够永久存储，缺点是这种存储材料极易受摩擦等外部作用而损坏



86

信息检索

复杂存储

- 光盘塔
- 磁盘阵列
- 云存储

信息检索

小结

信息组织 → 标引 → 描述 → 存储

南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

05

PART Five

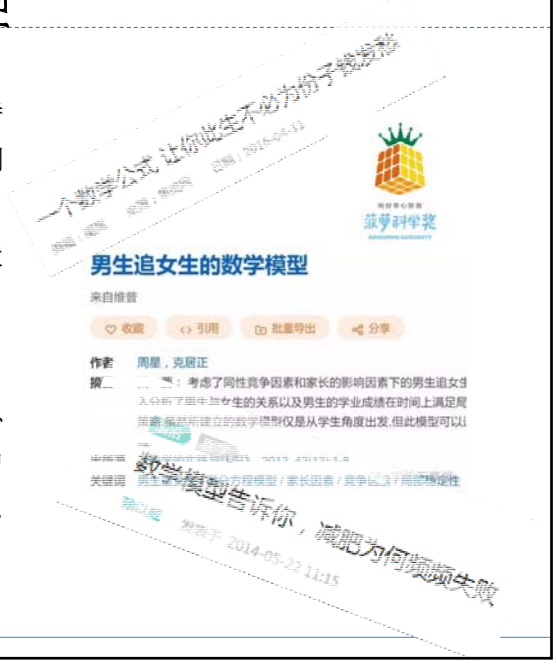
信息检索模型

Information Retrieval Model

信息检索

数学模型

数学模型，是指为了某种特定目的，通过对现实世界的某一特定对象做出一些**必要的简化与假设**，运用适当的数学工具得到的一种数学结构。数学模型具有**保留本质、抑制细节**的功能，它或者能**解释**特定现象的状态和性质，或者能**预测**它的未来状况，或者能提供对处理对象的最优**决策或控制**。



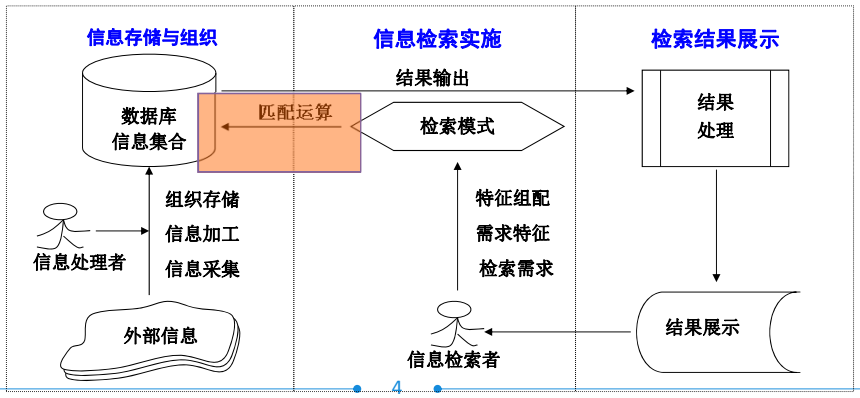
信息检索中的数学模型，就是运用数学的语言和工具，对信息检索系统中的信息及其处理过程加以抽象和编码，表述为某种数学公式，再经过演绎、推断、解释和检验，反过来指导信息检索服务与实践。

3

信息检索

信息检索模型

- **信息检索模型**是指如何对查询和文档进行表示，然后对它们进行相似度计算的框架和方法。
- 本质上是对**相关度**建模。
- 信息检索模型是IR中的核心内容之一。



4

信息检索

IR的数学模型分类表

检索型数学模型（Retrieval）		浏览型数学模型（Browsing）
基于内容的检索模型	基于结构的数学模型（结构化模型）	
集合论模型：布尔模型	非重叠列表（Non-Overlapping Lists）	平面（Flat）
集合论模型：模糊集合模型	邻近节点（Proximal Nodes）	结构导航（Structure Guided）
集合论模型：扩展布尔模型		超文本（Hypertext）
代数模型：向量空间模型		
代数模型：广义向量空间模型		
代数模型：潜在语义索引		
代数模型：神经网络		
概率论模型：（经典）概率模型		
概率论模型：推理网络		
概率论模型：信念网络		

5

信息检索

信息检索的四元组

System=（D，Q，F，R（d_j，q））

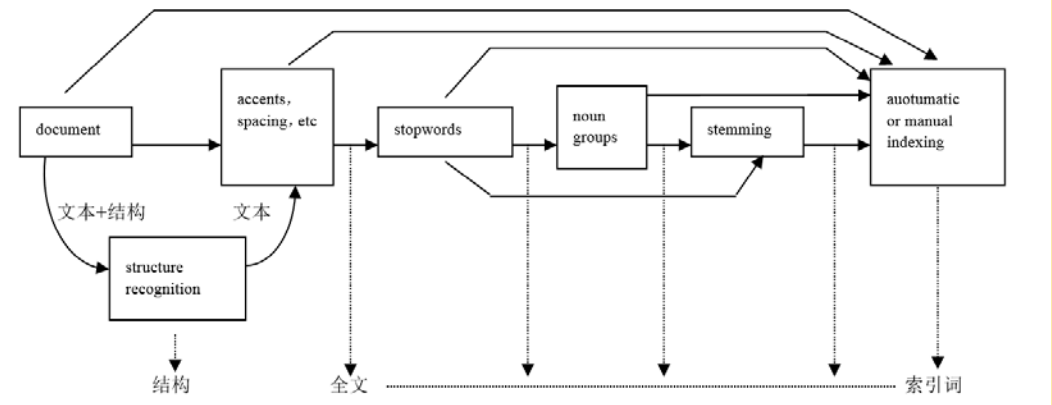
- *D* 信息资源集合
- *Q* 用户信息需求集合
- *F* 信息资源与信息需求的匹配处理框架
- *R*（d_j，*q*）匹配计算函数。

6

信息检索

信息资源集合（D）

- $D=\{d_1, d_2, \ldots\ldots d_n\}$ （$n\geq 0$）



文档逻辑视图：从全文文本到索引词集合

7

信息检索

例：文档的简单表示

◆data, document, information, library, publish（why?）

- 文档d_1=｛*data*, information, library｝
- 文档d_2=｛document, information, library, publish｝
- 文档d_3=｛data, information, publish｝
- 文档d_4=｛data, document, library, publish｝

- 文档d_1=｛1, 0, 1, 1, 0｝/ {data, ~~document~~, information, library, ~~publish~~}
- 文档d_2=｛0, 1, 1, 1, 1｝
- 文档d_3=｛1, 0, 1, 0, 1｝
- 文档d_4=｛1, 1, 0, 1, 1｝

8

信息检索

例：文档的加权表示

◆data, document, information, library, publish

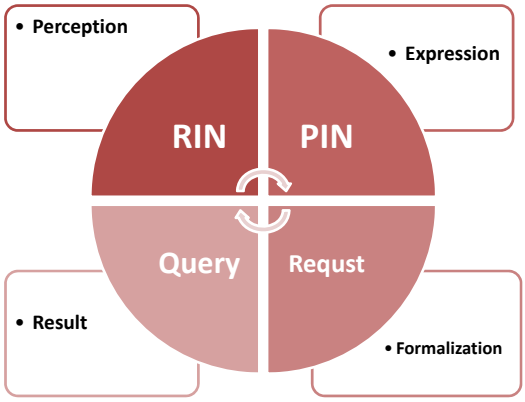
- 文档d_1={data 2, information 1 , library 2}
- 文档d_2={document 1 , information 3, library 2, publish 2}
- 文档d_3={data 3 , information 4, publish 1}
- 文档d_4={data 2, document 1, library 2, publish 1}

- 文档d_1={2,0,1,2,0}
- 文档d_2={0,1,3,2,2}
- 文档d_3={3,0,4,0,1}
- 文档d_4={2,1,0,2,1}

9

信息检索

用户信息需求集合（Q）



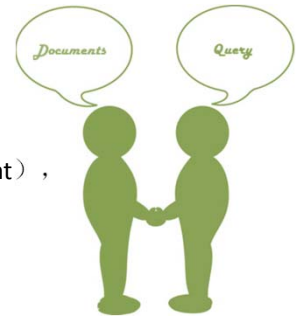
用户信息需求的不同状态

10

信息检索

匹配处理框架（F）

信息检索的根本任务是信息集合（D）与需求集合（Q）之间基于某种相似度规则的匹配处理，匹配处理框架（F）正是寻求在二者之间建立一种沟通与联系机制，提供对文档视图、提问式以及它们之间关系进行模型化处理的框架与规则。



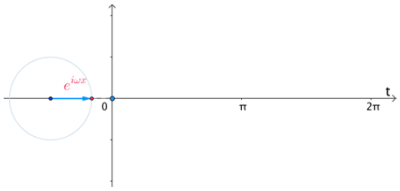
- 对布尔模型而言，匹配规则为**二值相关性判断**（binary relevance judgement）
  匹配运算主要基于集合论的集合基本运算；
- 对向量空间模型而言，匹配规则采用**多值相关性判断**（n-ary relevance judgement），
  匹配处理建立在多维向量空间理论和标准的向量线性代数操作基础之上；
- 而概率模型则依赖**集合论、概率运算和Bayes法则**来完成检索的匹配处理，
  其匹配规则也是多值性的相关性判断。

信息检索

匹配计算函数（R）

匹配函数$R(d_j, q)$用于计算任一文档d_j（$d_j \in D$）与任一提问q（$q \in Q$）形成的文档——提问对(d_j, q)之间的相似度大小。**一般地**，$R(d_j, q)$的函数值为一实数，其取值区间为**[0，1]**。从数学上来讲，匹配函数的选取，要求能够具备以下特点：

- 计算方法简单，计算量小；
- 函数值在取值区间均匀分布；
- 针对某一提问所获取的相关文档集合，能够实现合理的排序输出



信息检索

Classical Models

13

信息检索

预备知识？

14

信息检索

集合论

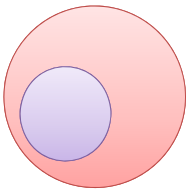
- 定义，简称集（Set）
  - 集合是“确定的一堆东西”，集合里的“东西”则称为元素
  - 由一个或多个确定的元素所构成的整体
- **确定性**
  - 给定一个集合，任给一个元素，该元素或者属于或者不属于该集合，二者必居其一，不允许有模棱两可的情况出现。
- **互异性**
  - 一个集合中，任何两个元素都认为是不相同的，即每个元素只能出现一次。有时需要对同一元素出现多次的情形进行刻画，可以使用多重集，其中的元素允许出现多次。
- **无序性**
  - 一个集合中，每个元素的地位都是相同的，元素之间是无序的。集合上可以定义序关系，定义了序关系后，元素之间就可以按照序关系排序。但就集合本身的特性而言，元素之间没有必然的序。

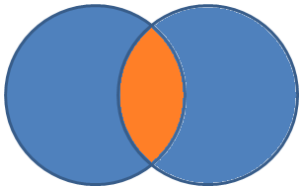
15

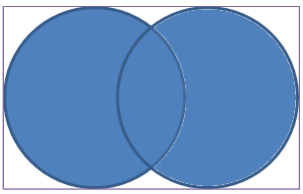
信息检索

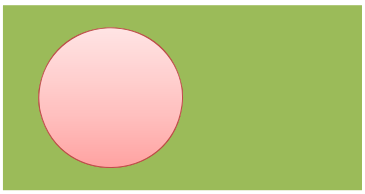
集合分类

- 空集
- 子集
- 交集
- 并集
- 补集









16

布尔模型

Boolean Model

George Boole，1815~1864

理论基础：布尔代数、集合论

- 文档表示
  - 一个文档被表示为关键词的集合
- 查询式表示
  - 查询式（Queries）被表示为关键词的布尔组合，用“与、或、非”连接起来，并用括弧指示优先次序
- 匹配
  - 一个文档当且仅当它能够满足布尔查询式时，才将其检索出来
  - 检索策略基于二值判定标准

布尔模型的基本原理

布尔模型在解释信息检索处理过程时，主要遵守的两条原则：

- 系统索引词集合中的每一个索引词在一篇文档中只有两种状态：出现或不出现。每个索引词的权值$w_{ij} \in \{0,1\}$
- 检索提问式q由三种布尔逻辑运算符“and/∧”、“or/∨”、“not/¬”连接索引词来构成。

根据布尔逻辑的运算规定，提问式q可以被表示成由合取子项（Conjunctive Components）组成的析取范式（Disjunctive Normal Form,简称dnf）形式。

信息检索

布尔运算的dnf形式

根据布尔逻辑的运算规定，提问式q可以被表示成由**合取子项**（Conjunctive Components）组成的**析取范式**（Disjunctive Normal Form,简称dnf）形式。

如：提问式 q = k1 and （k2 or not k3）可写成等价的析取范式形式：

q_{dnf} = (k1 **and** k2 **and** k3) **or** (k1 **and** k2 **and** not k3) **or** (k1 **and** not k2 **and** not k3)

这里q_{dnf}是提问式q的主析取范式。可进一步简化表示 为： q_{dnf}=(1,1,1) or (1,1,0) or (1,0,0)

其中： (1,1,1) or (1,1,0) or (1,0,0)是q_{dnf}的三个合取子项q_{cc}， 他们是一组向量， 由对应的三元组(k1 , k2 , k3)的每一个分量取0或1得到。

19

信息检索

例

- **Q = 病毒 AND （计算机 OR 电脑）AND NOT 医**

D1: …据报道，计算机病毒近日猖獗…

D2: …小王虽然是学医的，但对研究电脑病毒也很感兴趣，最近发明了一种…

D3: …计算机程序发现了爱滋病病毒的传播途径…

D4:…最近我的电脑中病毒了…

请问：列出析取范式，哪些文档会被检索出来？

(病毒 and 计算机 and not 电脑 and not 医) or
(病毒 and not 计算机 and 电脑 and not 医) or
(病毒 and 计算机 and 电脑 and not 医)

20

信息检索

布尔模型的匹配函数F

- 基于以上说明和假设，布尔模型对于任何一篇属于D的文档d_j，定义d_j与用户提问q的匹配函数为
- $sim(d_j,q)=$ 1，如果存在q_{cc}|（$q_{cc}\in Q_{dnf}$）且对于任何k_i，有$g_i(d_j)=g_i(q_{cc})$
- 0，其他
- 在这个式子中，函数g_i定义为$g_i(d_j)=W_{jj}$。现在，假设文档集合D中存在两篇文档d1和d2，其中，d1含有索引词k1和k2，d2含有索引词k1和k3，则他们的文档向量分别为

  d1=（1，1，0） d2=（1，0，1）
- 根据函数匹配$sim(d_j,q)$的定义，我们就不难看出，文档d1与提问式q=k1 AND（K2 OR NOT k3）的匹配函数值为1，即文档d1与提问q是相关的；而文档d2与提问q的匹配函数值为0，表明文档d2与提问q是不相关的。

q=(1,1,0) ^ (1,1,1) ^ (1,0,0)

21

信息检索

优点

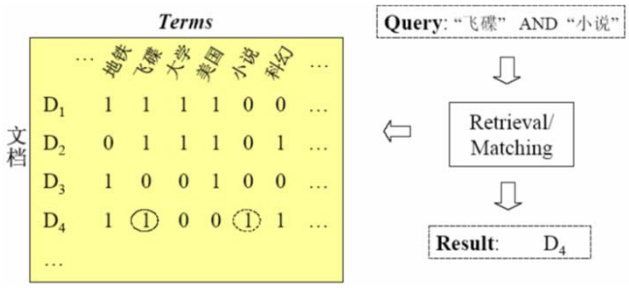
- 到目前为止，布尔模型是最常用的检索模型，因为：
  - 由于查询简单，因此容易理解
  - 通过使用复杂的布尔表达式，可以很方便地控制查询结果
- 相当有效的实现方法
  - 相当于识别包含了一个某个特定term的文档
- 经过某种训练的用户可以容易地写出布尔查询式
- 布尔模型可以通过扩展来包含排序的功能，即“扩展的布尔模型”

22

信息检索

缺点

- 布尔模型被认为是功能最弱的方式，其主要问题在于不支持部分匹配，而完全匹配会导致太多或者太少的结果文档被返回
- 很难控制被检索的文档数量
- 很难对输出进行排序
- 很难进行自动的相关反馈
- 无法体现文档之间的细微差别



信息检索

思考

欲查询2019年斯诺克世锦赛八进四的新闻，如何用布尔模型构建检索式？

(2019 OR 去年) AND

(台球 OR 斯诺克) AND

(世锦赛 OR 世界锦标赛) AND

(8进4 OR 八进四 OR 四分之一决赛)

信息检索

向量空间模型

- 向量空间模型（ Vector Space Model ）是康奈尔大学Salton 1970年代提出并倡导
- 成功应用于SMART（ System for the Manipulation and Retrieval of Text）文本检索系统
- 这一系统理论框架到现在仍然是信息检索技术研究的基础

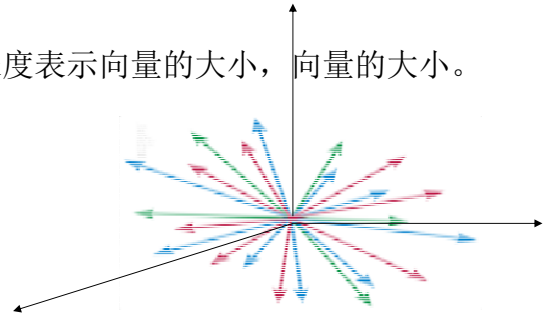
Gerard Salton（1927-1995）

25

信息检索

基本概念

- **Vector**
  - 在数学中，向量（ 欧几里得向量、几何向量、矢量），指具有大小和方向的量。
  - 数学表示：一般印刷用黑体的小写英文字母（a、b、c等）来表示，手写用在a、b、c等字母上加一箭头表示，如$\vec{a}$，也可以用大写字母AB、CD上加一箭头等表示，如，$\overrightarrow{AB}$。
  - 几何表示：用有向线段来表示。有向线段的长度表示向量的大小，向量的大小。
  - 坐标表示（数对），如（2，3，4）
- **Vector Space**
  - 是由一些被称为向量的对象构成的非空集合V

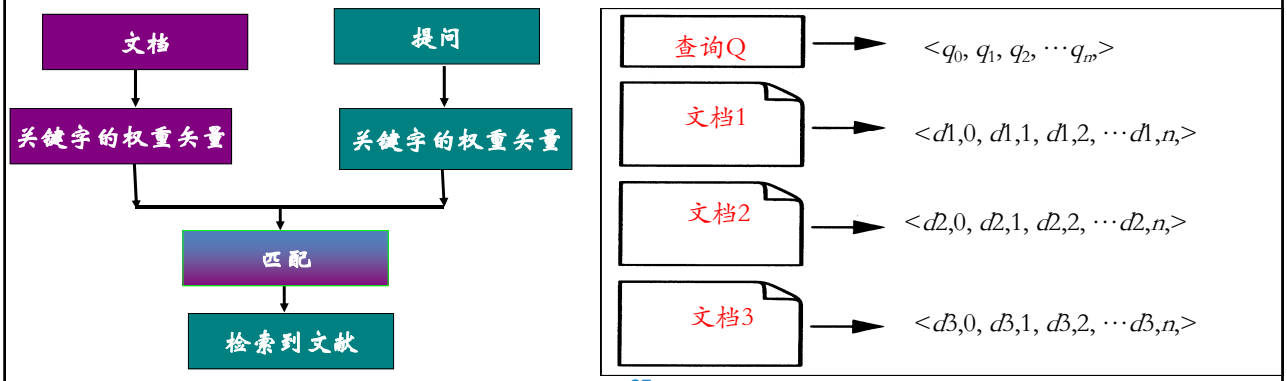


26

信息检索

基本原理

思想：文章的语义通过所使用的词语来表达
方法：每一篇文档用一个向量（特征向量）来表达，查询用一个向量来表达，通过向量来计算相似度。



信息检索

模型描述

- 文档D（**Document**）：泛指文档或文档中的一个片段（文档中的标题/摘要/正文等）。
- 索引项t（**Term**）：指出现在文档中能够代表文档性质的基本语言单位（字、词等），也就是通常所指的检索词，这样一个文档D就可以表示为D（$t_1,t_2,\ldots,t_n$），其中n就代表了检索词的数量。
- 特征项权重W_k（**Term Weight**）：指特征项t_n能够代表文档D能力的大小，体现了特征项在文档中的重要程度。
- 相似度S（**Similarity**）：指两个文档（或文档与查询）内容相关程度的大小。

信息检索

模型的特点

- 基于关键词(一个文本由一个关键词列表组成)
- 根据关键词的出现频率计算相似度
  - 例如：文档的统计特性
- 用户规定一个词项(term)集合，可以给每个词项附加权重
  - 未加权的词项：Q = 〈 database; text; information 〉
  - 加权的词项： Q = 〈 database 0.5; text 0.8; information 0.2 〉
  - 查询式中没有布尔条件
- 根据相似度对输出结果进行排序
- 支持自动的相关反馈
  - 有用的词项被添加到原始的查询式中
  - 例如：Q ⇒〈 database; text; information; *document* 〉

29

信息检索

模型中的问题

- 怎样确定文档中哪些词是重要的词？（索引项）
- 怎样确定一个词在某个文档中或在整个文档集中的重要程度？（权重）
- 怎样确定一个文档和一个查询式之间的相似度？

30

索引项的选择

- 若干独立的词项被选作索引项(*index terms*) or 词表（*vocabulary*）
- 索引项代表了一个**应用中的重要词项**
  - 信息科学图书馆中的索引项应该是哪些呢?

信息科学
文档集

信息
知识
情报
计算机
数据库
….
XML

文档集中的索引项

索引项的选择

- 这些索引项是不相关的 *(或者说是正交的)*，形成一个向量空间 *vector space*

信息科学文档集

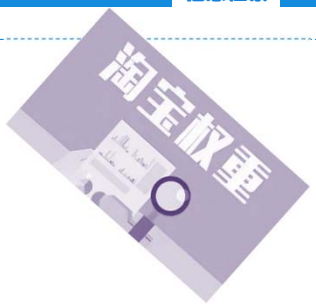
“情报” “科学” “商务”

该文档集中的全部重要词项

- 实际上，这些词项是相互关联的
  - 当你在一个文档中看到“信息”，非常有可能同时看到“科学”
  - 当你在一个文档中看到“信息”，有中等的可能性同时看到“商务”
  - 当你在一个文档中看到“商务”，只有很少的机会同时看到“科学”

信息检索

文档向量的构造



对于任一文档$d_j \in D$，都可将它表示为t维向量形式：

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{ij})$$

其中，向量分量w_{ij}代表第i个索引词k_i在文档d_j中所具有的权重，t为系统中索引词的个数。

在Boolean模型中， w_{ij} ={0，1}，在VSM中，w_{ij}=[0，1]

一篇文档有多个索引词，如何计算每个索引词的权值？

信息检索

词数与词频

- Term Count
  - 指定词项*i*在文档*j*中出现的次数，$n_{i,j}$
- Term Frequency
  - 指定词项在某文档中出现的频率（相对次数）$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$

冠状肺炎	2/1000=0.002
应用	5/1000=0.005
的	35/1000=0.035

信息检索

索引词的权重：tf-idf

- 根据词项在文档（*tf*）和文档集（*idf*）中的频率(frequency)计算词项的权重
- 词项的重要性随着它在**文档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_{ij} = 词项j在文档i中的频率
  - df_j = 词项j的**文档频率**= 包含词项j的文档数量
  - idf_j = 词项j的**逆文档频率**= log (N/df_j)
    - N: 文档集中文档总数
    - 逆文档频率用词项区别文档
  - W_{ij} =词项t_j在文档d_i中的**权重**= $tf_{ij} \times idf_j$

信息检索

*idf*示例

文档总数为1000，出现关键词k_1文档为100篇，出现关键词k_2文档为500篇，出现关键词k_3文档为800篇

N=1000， n_1=100， n_2=500， n_3=800

根据公式： $idf_i = log(N/n_i)$ ，可计算出

$idf_1 = 3 - 2 = 1$

$idf_2 = 3 - 2.7 = 0.3$

$idf_3 = 3 - 2.9 = 0.1$

idf越大，表明**区别（分）文档的能力越强**。

信息检索

文档的词项权重(TF-IDF举例)

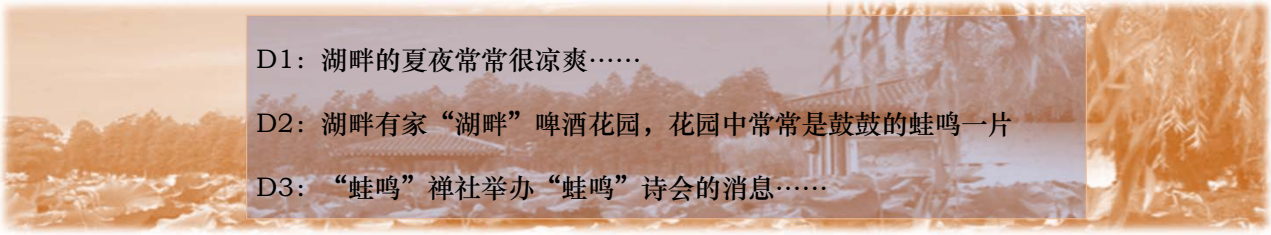
相关文本：“俄罗斯频繁发生恐怖事件，俄罗斯的安全部门加大打击恐怖主义的力度。”

	TF	IDF	TFIDF		TF	IDF	TFIDF
俄罗斯	2	较高	高	安全	1	中等	高
恐怖	2	较高	高	部门	1	较低	低
的	2	非常低	很低	加大	1	较低	低
频繁	1	较低	低	打击	1	中等	高
发生	1	较低	低	主义	1	较低	低
事件	1	较低	低	力度	1	中等	高

37

信息检索

idf计算示例



$N = 3$ $\quad idf_i = \log(\frac{N}{df_i})$

词项	湖畔	夏夜	的	常常	蛙鸣	禅社	诗会
df	2	1	3	2	2	1	1
Idf	0.176	0.477	0	0.176	0.176	0.477	0.477

38

信息检索

查询式的词项权重

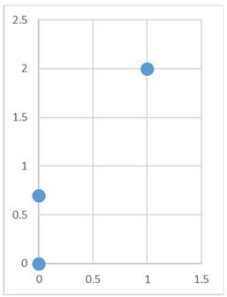
- 如果词项出现在查询式中，则该词项在查询式中的权重为1，否则为0
- 也可以用用户指定查询式中词项的权重
- 一个自然语言查询式可以被看成一个文档
  - 查询式：“有没有周杰伦的歌？”会被转换为:
    <周杰伦, 歌>
  - 查询式：“请帮我找关于俄罗斯和车臣之间的战争以及车臣恐怖主义首脑的资料”会被转换为:
    <俄罗斯 1, 车臣 2, 战争1, 恐怖主义1, 首脑 1>
  - 过滤掉了：“请帮我找”，“和”，“之间的”，“以及”，“的资料”
- 两个文档之间的相似度可以同理计算

39

信息检索

由索引项构成向量空间

- 2个索引项构成一个二维空间（平面），一个文档可能包含0, 1 或2个索引项
  - $d_i = \langle 0, 0 \rangle$ (一个索引项也不包含)
  - $d_j = \langle 0, 0.7 \rangle$ (包含其中一个索引项)
  - $d_k = \langle 1, 2 \rangle$ (包含两个索引项)
- 类似的，3个索引项构成一个三维空间，n个索引项构成n维空间
- 一个文档或查询式可以表示为n个元素的线性组合



40

信息检索

文档集－一般表示

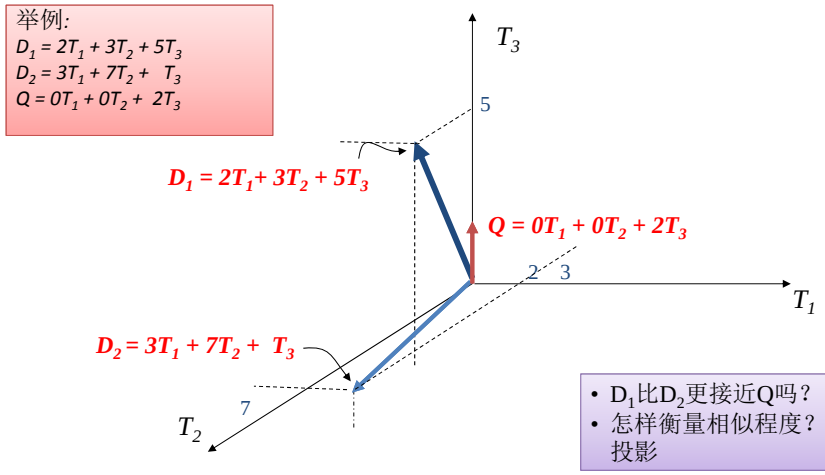
- 向量空间中的N个文档可以用一个矩阵表示
- 矩阵中的一个元素对应于文档中一个词项的权重。“0”意味着该词项在文档中没有意义，或该词项不在文档中出现。

	T_1	T_2	….	T_t
D_1	d_{11}	d_{12}	…	d_{1t}
D_2	d_{21}	d_{22}	…	d_{2t}
:	:	:		:
:	:	:		:
D_n	d_{n1}	d_{n2}	…	d_{nt}

信息检索

图示

举例:
$D_1 = 2T_1 + 3T_2 + 5T_3$
$D_2 = 3T_1 + 7T_2 + T_3$
$Q = 0T_1 + 0T_2 + 2T_3$



- D_1比D_2更接近Q吗？
- 怎样衡量相似程度？夹角还是投影

信息检索

相似度计算

- 相似度是一个函数，它给出两个向量之间的相似程度，查询式和文档都是向量，各类相似度存在于：
  - 两个文档之间（文本分类，聚类）
  - 两个查询式之间（常问问题集）
  - 一个查询式和一个文档之间（检索）

- 人们提出了大量的相似度计算方法，因为最佳的相似度计算方法并不存在。

信息检索

通过计算查询式和文档之间的相似度

- 可以根据预定的重要程度对检索出来的文档进行排序
- 可以通过强制设定某个阈值，控制被检索出来的文档的数量
- 检索结果可以被用于相关反馈中，以便对原始的查询式进行修正。
  (例如：将文档向量和查询式向量进行结合)

信息检索

相似度度量 – 内积(Inner Product)

- 文档D和查询式Q可以通过内积进行计算:

$$\text{sim}(D, Q) = \sum_{k=1}^{t} (d_{ik} \bullet q_k)$$

  - d_{ik}是文档d_i中的词项k的权重，q_k是查询式Q中词项k的权重
- 对于二值向量, 内积是查询式中的词项和文档中的词项相互匹配的数量
- 对于加权向量, 内积是查询式和文档中相互匹配的词项的权重乘积之和

信息检索

内积 – 举例

retrieval database architecture computer text management information

- 二值（Binary）：
  - D = 1, 1, 1, 0, 1, 1, 0
  - Q = 1, 0, 1, 0, 0, 1, 1
  - sim(D, Q) = 3

- 向量的大小 = 词表的大小 = 7
- 0 意味着某个词项没有在文档中出现，或者没有在查询式中出现

- 加权

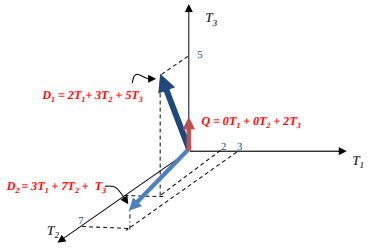
$D_1 = 2T_1 + 3T_2 + 5T_3$

$Q = 0T_1 + 0T_2 + 2T_3$

$D_2 = 3T_1 + 7T_2 + T_3$

$\text{sim}(D_1, Q) = 2*0 + 3*0 + 5*2 = \mathbf{10}$

$\text{sim}(D_2, Q) = 3*0 + 7*0 + 1*2 = 2$



信息检索

内积的特点

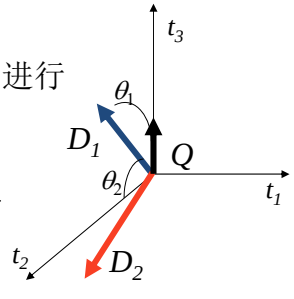
- 内积值没有界限
  - 不象概率值，要在(0, 1)之间
- 对长文档有利
  - 内积用于衡量有多少词项匹配成功，而不计算有多少词项匹配失败
  - 长文档包含大量独立词项，每个词项均多次出现，因此一般而言，和查询式中的词项匹配成功的可能性就会比短文档大。

信息检索

余弦(Cosine)相似度度量

- 余弦相似度计算两个向量的夹角
- 余弦相似度是利用向量长度对内积进行归一化的结果

$$\text{CosSim}(D_i, Q) = \frac{\sum_{k=1}^{t}(d_{ik} \bullet q_k)}{\sqrt{\sum_{k=1}^{t} d_{ik}^2 \bullet \sum_{k=1}^{t} q_k^2}}$$



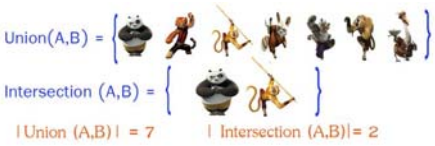
$D_1 = 2T_1 + 3T_2 + 5T_3$ CosSim(D_1, Q) = 5 / √ 38 = **0.81**
$D_2 = 3T_1 + 7T_2 + T_3$ CosSim(D_2, Q) = 1 / √ 59 = 0.13
$Q = 0T_1 + 0T_2 + 2T_3$

用余弦计算，D_1 比 D_2 高6倍；
用内积计算， D_1 比 D_2 高5倍

信息检索

Jaccard Similarity

Jaccard Similarity $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$



Jaccard Similarity J (A,B) = | Intersection (A,B) | / | Union (A,B) |

= 2 / 7

= 0.286

49

信息检索

Jaccard Coefficient

Jaccard Coefficient: $\frac{\sum_{k=1}^{t}(d_{ik} \bullet q_k)}{\sum_{k=1}^{t} d_{ik}^{\,2} + \sum_{k=1}^{t} q_k^{\,2} - \sum_{k=1}^{t}(d_{ik} \bullet q_k)}$

$D_1 = 2T_1 + 3T_2 + 5T_3$ Sim(D_1 , Q) = 10 / (38+4-10) = 10/32 = 0.312
$D_2 = 3T_1 + 7T_2 + T_3$ Sim(D_2 , Q) = 2 / (59+4-2) = 2/61 = 0.033
$Q = 0T_1 + 0T_2 + 2T_3$

- D_1 比 D_2 高9.5倍

50

信息检索

余弦相似性示例

Term / W_{ij} / Doc	…	湖畔	夏夜						
D_1	…	0.176	0.477	0	0.176	0	0	0	…
D_2		0.352	0	0	0.176	0.176	0	0	
D_3		0	0	0					
Q	…	0.176	0.477	0					

Query = "夏夜湖畔的蛙鸣"

Term Wij Doc.	.....	湖畔	夏夜	的	常常	蛙鸣	禅社	诗会	.....
D1	.....	0.176	0.477	0	0.176	0	0	0	.....
D2	.....	0.352	0	0	0.176	0.176	0	0	.....
D3	.....	0	0	0	0	0.352	0.477	0.477	.....
Q	.....	0.176	0.477	0	0	0.176	0	0	.....

$Cos(q,d_1)=0.893 \quad Cos(q,d_2)=0.400 \quad Cos(q,d_3)=0.151$

与查询 q 相似的文档顺序：$d_1 \succ d_2 \succ d_3$

51

信息检索

向量空间模型优点

- 术语权重的算法提高了检索的性能
- 部分匹配的策略使得检索的结果文档集更接近用户的检索需求
- 可以根据结果文档对于查询串的相关度通过Cosine Ranking等公式对结果文档进行排序

52

信息检索

向量空间模型的不足

- 标引词之间被认为是相互独立
- 随着Web页面信息量的增大、Web格式的多样化，这种方法查询的结果往往会与用户真实的需求相差甚远，而且产生的无用信息量会非常大
- 隐含语义索引（LSI）等模型是向量空间模型的延伸

53

信息检索

课堂练习

- 对于下列例子，计算(文档长度以字节数表示，不含标点和空格)，写出计算过程，并判断哪篇文档和查询q更相关。

- **Q：** **"gold silver truck"**
- **D1：** **"Shipment of gold damaged in a fire"**
- **D2：** **"Delivery of silver arrived in a silver truck"**
- **D3：** **"Shipment of gold arrived in a truck"**

54

信息检索

结果

$idf_a = 0$ $idf_{in} = 0$

$idf_{arrived} = 0.176$ $idf_{of} = 0$

$idf_{damaged} = 0.477$ $idf_{silver} = 0.477$

$idf_{delivery} = 0.477$ $idf_{shipment} = 0.176$

$idf_{fire} = 0.477$ $idf_{truck} = 0.176$

$idf_{gold} = 0.176$

docid	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
D_1	0	0	0.477	0	0.477	0.176	0	0	0.176	0	0
D_2	0	0.176	0	0.477	0	0	0	0	0	0.954	0.176
D_3	0	0.176	0	0	0	0.176	0	0	0.176	0	0.176
Q	0	0	0	0	0	0.176	0	0	0	0.477	0.176

55

信息检索

布尔检索示例

飞碟 AND 小说：只能检索出D4，无法显现D1,D2,D3的差异

飞碟 OR 小说：可以检出D1,D2,D4，但无法显现它们的差异

Terms

文档	…	地铁	飞碟	大学	美国	小说	科幻	…
D_1		1	1	1	1	0	0	…
D_2		0	1	1	1	0	1	…
D_3		1	0	0	1	0	0	…
D_4		1	(1)	0	0	(1)	1	…
…								

Query: “飞碟” AND “小说”

⇩

⇦ Retrieval/ Matching

⇩

Result: D_4

56

信息检索

扩展布尔模型

布尔模型和VSM各自有着自己的优点和不足，能否将两者结合起来，克服自身的缺点，发挥相互的长处？

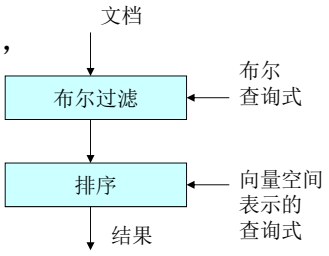
1983年G.Salton及其学生提出一种基于布尔逻辑框架的混合布尔、向量特性的“扩展布尔模型”。

57

信息检索

布尔模型和向量空间模型相结合

- 布尔模型可以和向量空间模型相结合，先做布尔过滤，然后进行排序：
  - 首先进行布尔查询
  - 将全部满足布尔查询的文档汇集成一个文档
  - 用向量空间法对布尔检索结果进行排序



如果忽略布尔关系的话，向量空间查询式和布尔查询式是相同的

58

信息检索

先“布尔”，后“排序”存在的问题

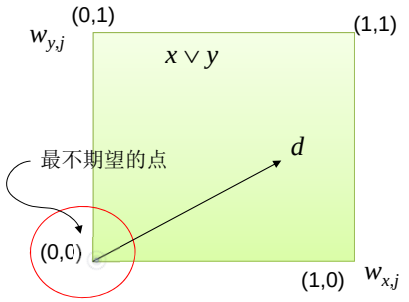
- 如果“与”应用于布尔查询式, 结果集可能太窄, 因而影响了后面的排序过程
- 如果“或”应用于布尔查询式, 就和纯向量空间模型没有多大区别了
- 在第一步，如何最佳地应用布尔模型呢？
- 提出扩展布尔模型

59

信息检索

扩展布尔模型中的“或”关系

- 给定一个或关系的查询式：$x \vee y$
- 假设文档d_i中x和y的权重被归一化在(0,1)区间内:
  - $w_{x,j} = （tf_{x,j} / max_l\ tf_{l,j}）\times（idf_x / max_i\ idf_i）$
  - $\text{sim}(q_{or}, d_j) = [\ (x^2 + y^2)/2\ ]^{0.5}$ where $x = w_{x,j}$ and $y = w_{y,j}$



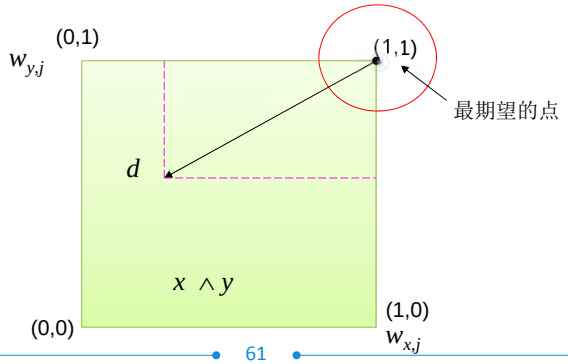
- 一个文档在(1, 1) 处获得最高的权重，此时意味着文档包含了全部两个查询词，并且查询词在文档中的权重也是最高的
- 函数sim()度量了从原点出发的文档向量长度

60

信息检索

扩展布尔模型中的“与”关系

- 给定一个联合的查询式 $x \wedge y$
- $\text{sim}(q_{\text{and}}, d_j) = 1 - \{ [(1-x)^2 + (1-y)^2]/2 \}^{0.5}$
- 函数sim() 表示从**(1,1) 出发到d**的向量长度



61

信息检索

扩展的布尔检索相似度计算示例

飞碟 AND 小说：只能检索出D4，无法显现D1,D2,D3的差异
飞碟 OR 小说：可以检出D1,D2,D4，但无法显现它们的差异

Sim() | Q | 飞碟 or 小说

Query: “飞碟” AND “小说”

Retrieval/ Matching

Result: D_4

Terms

文档	…	地铁	飞碟	大学	美国	小说	科幻	…
D_1		1	1	1	1	0	0	…
D_2		0	1	1	1	0	1	…
D_3		1	0	0	1	0	0	…
D_4		1	1	0	0	1	1	…
…								

62

信息检索

观察

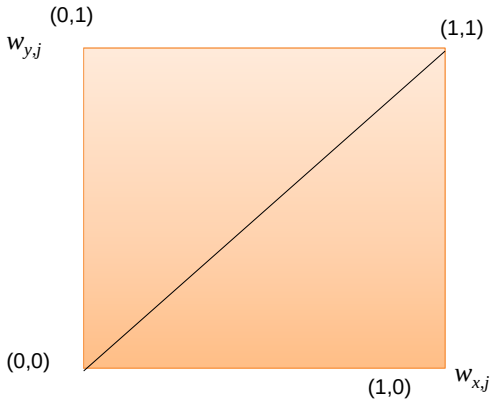
- 如果权值是布尔型的，x出现在文档d_j中，则x在文档d_j中具有权重1，否则为0
- 当d_j 包含x和y时

  $\text{sim}(q_{\text{and}}, d_j) = \text{sim}(q_{\text{or}}, d_j) = 1$
- 当d_j 既不包含x 也不包含y时

  $\text{sim}(q_{\text{and}}, d_j) = \text{sim}(q_{\text{or}}, d_j) = 0$
- 当d_j 包含x 和y二者之一时

  $\text{sim}(q_{\text{and}}, d_j) = 1 - 1/2^{0.5} = 0.293$

  $\text{sim}(q_{\text{or}}, d_j) = 1/2^{0.5} = 0.707$



63

信息检索

观察

- 一个词项的存在将对“或”关系查询式提供0.707的增益值，但对“与”关系查询式仅提供0.293的增益值
  - 一个词项不存在，将给“与”关系的查询式提供0.707的罚分
- 当x 和y 有权值0.5, $\text{sim}(q_{\text{and}}, d) = \text{sim}(q_{\text{or}}, d) = 0.5$
  - 在一个“与”关系查询中，两个词项的权重均为0.5，则相似度为0.5。其中一个权重为1，另一个为0，相似度为0.293。
  - 在“或关系”查询中，情况恰好相反
- 在“与关系”查询中，如果一个词项的权重低于0.5，将给相似度贡献一个较大的罚分

64

信息检索

p-norm 模型

- 扩展布尔模型可以被泛化为m 个查询项:

  $sim(q_{or}, d) = [(x_1^2 + x_2^2 + ... + x_m^2) / m]^{0.5}$

  $sim(q_{and}, d) = 1 - \{ [(1- x_1)^2 + (1- x_2)^2 + ... + (1- x_m)^2] / m \}^{0.5}$

- 它可以被进一步地 泛化为p-norm model:

  $sim(q_{or}, d) = [(x_1^p + x_2^p + ... + x_m^p) / m]^{1/p}$

  $sim(q_{and}, d) = 1 - \{ [(1- x_1)^p + (1- x_2)^p + ... + (1- x_m)^p] / m \}^{1/p}$

- 当$p = 1$时, $sim(q_{or}, d) = sim(q_{and}, d) = (x_1 + x_2 + ... + x_m)/ m$
  - 通过语词-文献权值的和来求合取和析取查询的值，和向量空间中的内积相似
- 当$p = \infty$, $sim(q_{or}, d) = max(x_i)$; $sim(q_{and}, d) = min(x_i)$
  - 模糊逻辑模型(Fuzzy logic model)

65

信息检索

总结

扩展布尔模型的特点：

- 与传统的布尔检索中的倒排文档技术相兼容，支持使用标准布尔逻辑表达的提问式结构；
- 允许在文档和提问式中进行词加权处理；
- 支持按相似度的大小排序输出检索结果；
- 通过调整参数p的值，可灵活选择并得到不同检索结果。

66

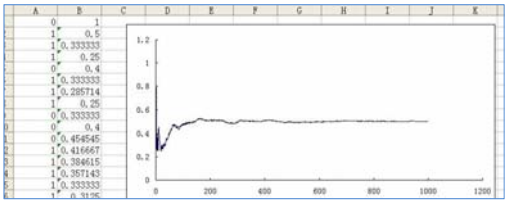
信息检索

预备知识：概率

概率（Probability）亦称“或然率”、“机率”。它反映随机事件出现的可能性大小的量度。（随机事件是指在相同条件下，可能出现也可能不出现的事件。）

- **古典定义：**

$$p(A)=\frac{m}{n},$$



- **统计定义**

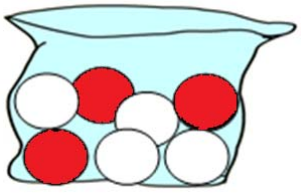
在一定条件下，重复做n次试验，n_A为n次试验中事件A发生的次数，如果随着n逐渐增大，频率n_A/n逐渐稳定在某一数值p附近，则数值p称为事件A在该条件下发生的概率，记做$P(A)=p$。这个定义成为概率的统计定义。

(Jakob Bernoulli,1654－1705)

信息检索

例

袋子中装有10个大小、形状相同的球. 球编号为1－10 （1~6号为红球，其余白球）。把球搅匀，从中任取一球。记 *A*={摸到2号球}, 记 *B*={摸到红球}, *C*={球号大于3的红球}，求事件*A*、*B*、*C* 的概率。



解：$N(S)=10,\ N(A)=1$ $\quad P(A)=\dfrac{N(A)}{N(S)}=\dfrac{1}{10}$

$N(B)=6$ $\quad P(B)=\dfrac{N(B)}{N(S)}=\dfrac{6}{10}$

$N(C)=3$ $\quad P(C)=\dfrac{N(C)}{N(S)}=\dfrac{3}{10}$

信息检索

事件统计的加法原理

设完成一件事有m种方式，第一种方式有n_1种方法，第二种方式有n_2种方法,…,第m种方式有n_m种方法，无论哪种方法都可完成。

则完成这件事共有$n_1 + n_2 + \ldots + n_m$种方法 .

例：从甲地到乙地有三类交通工具可供选择：汽车、火车和飞机。

而汽车有5个班次，火车有3个班次，飞机有2个班次。

则从甲地到乙地共有5+3+2=10种方法.

信息检索

事件统计的乘法原理

若完成一件事有m个步骤，第一个步骤有n_1种方法，第二个步骤有n_2种方法, …，第m个步骤有n_m种方法。则完成这件事共有

$n_1 \times n_2 \times \ldots \times n_m$种方法.

例：若一个人有三顶帽子和两件背心，他可以有多少种打扮？

可有 3×2 种打扮

注意：加法、乘法原理计算概率时非常重要。

信息检索

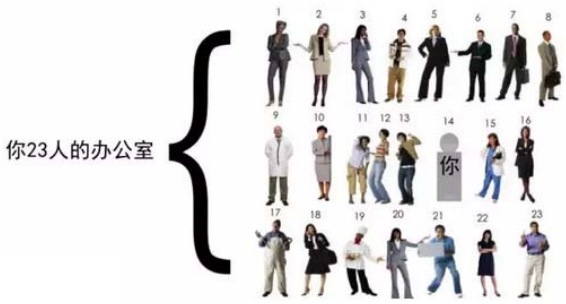
小结

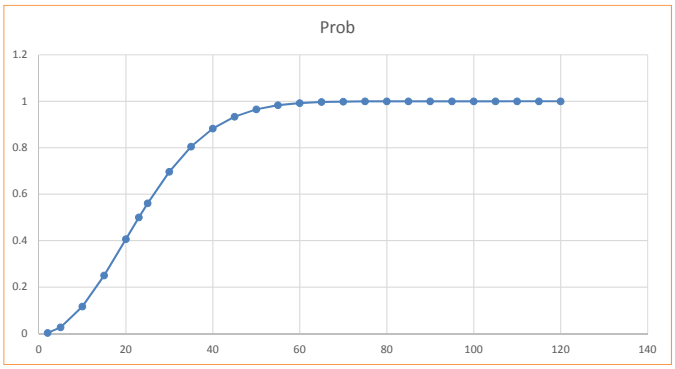
$$0 \leq P(A) \leq 1$$

$$P(A) + P(\bar{A}) = 1$$

信息检索

生日悖论



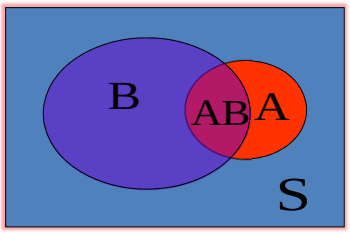


信息检索

条件概率/Conditional probability

设A,B为样本空间S中两个事件,并且P(A)>0. 则称

$P(B|A)=\frac{P(AB)}{P(A)}$ 为事件A发生条件下事件B发生的概率.

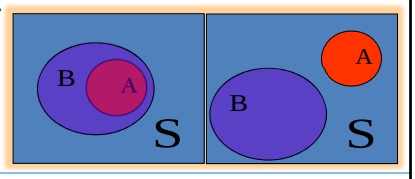


(1) 前提P(A)>0, 否则P(B|A)=0

(2) 求P(B|A)时,样本空间由S缩小至A,在A中确定B发生的可能性.

(3) 一般情况下, P(B)≠P(B|A),两者含义,发生的条件都不相同.

(4) 如果A⊂B,则P(B|A)=1. 如果AB=ϕ, 则P(B|A)=0.



73

信息检索

例1：

掷一颗均匀骰子，*A*={掷出2点}， *B*={掷出偶数点}。求: (1) P(A)； （2）在事件B已发生条件下A发生的概率。

解: S={1, 2, 3, 4, 5, 6}　A={2}　B={2, 4, 6}

(1) 古典概型　$P(A)=\frac{1}{6}$

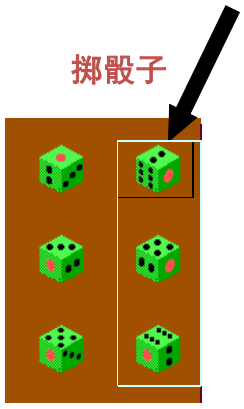
(2) 事件B发生, 意味着$\overline{B}$={1, 3, 5}不会发生.

此时所有可能结果就是$B=S_B$(B发生下的新样本空间)

B中共有3个元素,且等可能性. 其中只有$2\in A$(即AB={2})

因此, $P(A|B)=\frac{N(A|B)}{N(S_B)}=\frac{1}{3}$ (B发生条件下A的概率)

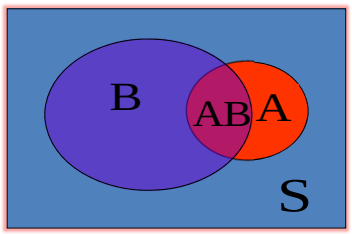
注意： $P(A|B)=\frac{1}{3}=\frac{1/6}{3/6}=\frac{P(AB)}{P(B)}$



74

信息检索

确认一下



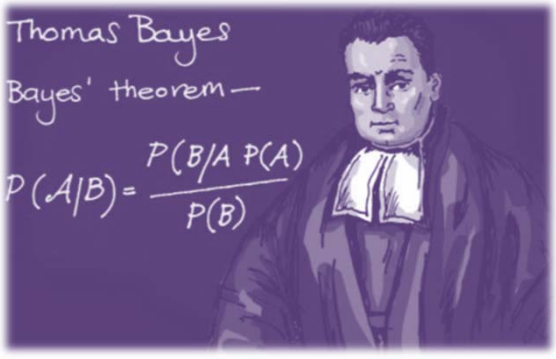
$$P(B|A)=\frac{P(AB)}{P(A)} \quad P(A|B)=\frac{P(AB)}{P(B)}$$

$$P(AB)=P(B|A)*P(A)=P(A|B)*P(B)$$

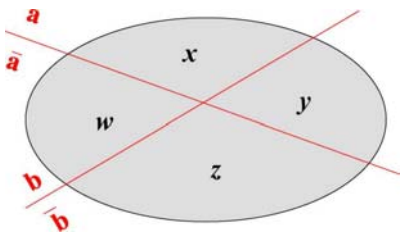
75

信息检索

贝叶斯公式



Thomas Bayes（1702-1761）

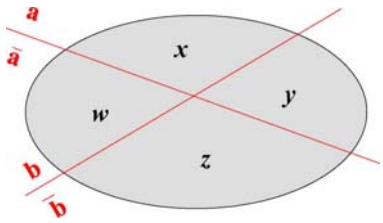


$$P(a)=x+y$$
$$P(b)=x+w$$
$$P(a|b)=\frac{x}{(w+x)}$$
$$P(b|a)=\frac{x}{y+x}$$

76

信息检索

证明



$$P(a \mid b) = \frac{P(b \mid a) \times P(a)}{P(b)}$$

$$P(a \mid b) \times P(b) = P(b \mid a) \times P(a)$$

$$P(a \mid b) \times P(b) = \frac{x}{x + w} \times (x + w) = x$$

$$P(b \mid a) \times P(a) = \frac{x}{x + y} \times (x + y) = x$$

信息检索

例子

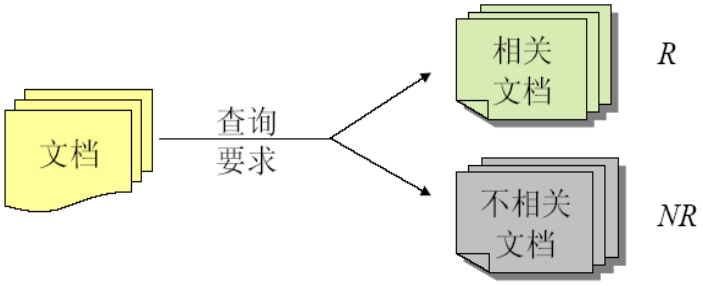
- 2020春冠状肺炎爆发，大量人员需要做检测
  - 1%的人口会感染　P（*well*）=0.99　P（*disease*）=0.01
  - 正常人检测，1%失误　P（*Pos.*|*well*）=0.01　P（*Neg.*|*well*）=0.99
  - 患者检测，1%失误　P（*Neg.*|*disease*）=0.01　P（*Pos.*|*disease* ）=0.99
- 小黄人鲍勃抽检结果为阳性，那他是患者的概率？
  - P（*disease*|*Neg.*）？

$$P(Disease \mid Pos.) = \frac{P(Pos. \mid Disease)P(Disease)}{P(Pos.)}$$

$$= \frac{0.99 * 0.01}{0.0198} = 50\% \ \surd$$

信息检索

概率模型



检索问题即求条件概率问题

If $P(R|d_i,q) > P(NR|d_i,q)$

then d_i是检索结果，否则不是检索结果

79

信息检索

理想状态

文档总数为N

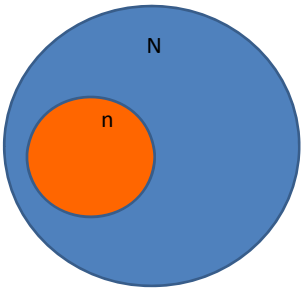
对于检索q而言，相关文档集（R）的文档数为n，（0<=n<=N）

相关文档被检索出来的概率为：

$$P(R) = \frac{n}{N}$$

反之

$$P(\bar{R}) = \frac{N-n}{N}$$



80

信息检索

文档表示

二值（布尔）表达

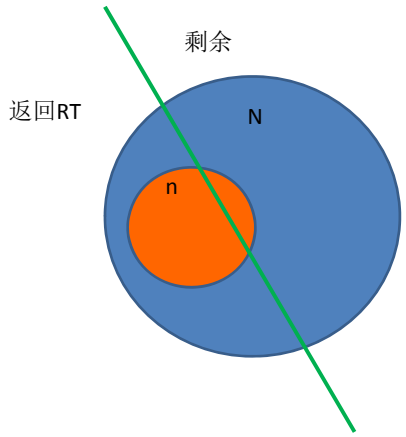
$X=(x_1, x_2, \ldots\ldots, x_n)$

信息检索

检索

对于一次检索q而言，将文档集分成两部分RT和$\overline{RT}$，
我们的目标：

$$P(R|RT) > P(\bar{R}|RT)>=0$$



信息检索

目标值

目标：

$$P(R|RT) > P(\bar{R}|RT)>=0$$

因为：

$$P(R|RT) + P(\bar{R}|RT)=1$$

所以目标：

$$P(R|RT) > 0.5$$

83

信息检索

判别函数

Discrimination function

$$Dis(RT) = \frac{P(R|RT)}{P(\bar{R}|RT)}$$

84

信息检索

判别函数的提升

提高要求：返回结果集中，有效的是无效的3倍以上

$$P(R|RT) > 3 \times P(\bar{R}|RT)$$

那么

$$P(R|RT) > 0.75$$

该值越大，检索结果越好！！！

信息检索

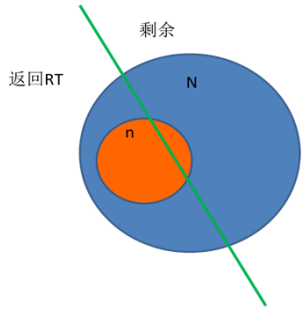
贝叶斯公式代入

$$P(a|b)=\frac{P(b|a)P(a)}{P(b)}$$

$$Dis(RT) = \frac{P(R|RT)}{P(\bar{R}|RT)}$$

$$Dis(RT) = \frac{P(RT|R)\times P(R)}{P(RT)} \times \frac{1}{\frac{P(RT|\bar{R})\times P(\bar{R})}{P(RT)}}$$

$$Dis(RT) = \frac{P(RT|R)\times P(R)}{P(RT|\bar{R})\times P(\bar{R})}$$



信息检索

索引项代入

假设文档是用索引词：t_1,t_2，t_3.....t_n表示，且统计独立

$$P(RT|R) = P(t_1|R \times t_2|R \times t_3|R \times \cdots t_{n-1}|R \times t_n|R)$$

$$P(RT|R) = P(t_1|R) \times P(t_2|R) \times P(t_3|R) \times \cdots P(t_n|R)$$

类似

$$P(RT|\bar{R}) = P(t_1|\bar{R}) \times P(t_2|\bar{R}) \times P(t_3|\bar{R}) \times \cdots P(t_n|\bar{R})$$

信息检索

朴素贝叶斯模型

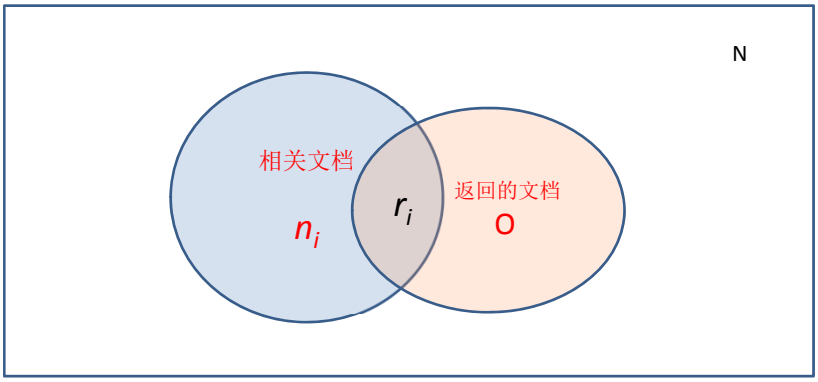
$$Dis(RT) = \frac{P(RT|R)\times P(R)}{P(RT|\bar{R})\times P(\bar{R})}$$

$$Dis(RT) = \frac{P(t_1|R)\times P(t_2|R)\times P(t_3|R)\times\cdots P(t_n|R)}{P(t_1|\bar{R})\times P(t_2|\bar{R})\times P(t_3|\bar{R})\times\cdots P(t_n|\bar{R})} \times \frac{P(R)}{P(\bar{R})}$$

$$Dis(RT) \approx \frac{P(t_1|R)\times P(t_2|R)\times P(t_3|R)\times\cdots P(t_n|R)}{P(t_1|\bar{R})\times P(t_2|\bar{R})\times P(t_3|\bar{R})\times\cdots P(t_n|\bar{R})}$$

信息检索

对于t_i



89

信息检索

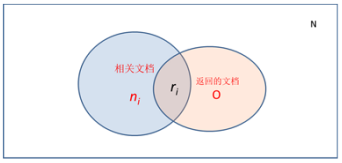
假设

	T1	T2	T3	T4
D1	1	1	1	1
D2	1	0	0	1
D3	0	0	0	1
D4	1	0	0	0
D5	0	1	0	0
D6	0	1	0	1

90

信息检索

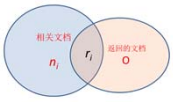
计算t_i相关参数项



Documents	相关	不相关	合计
x_i=1（返回）	r_i	O-r_i	O
x_i=0（未返回）	n_i-r_i	N-R-n_i+r_i	N-O
合计	n_i	N-n_i	N

信息检索

则



$$P(t_i|R)=\frac{r_i}{n_i} \qquad P(t_i|\bar{R})=\frac{O-r_i}{N-n_i}$$

$$\frac{P(t_i|R)}{P(t_i|\bar{R})}=\frac{r_i}{n_i}\times\frac{N-n_i}{O-r_i}=\frac{r_i}{O-r_i}\times\frac{N-n_i}{n_i}$$

信息检索

完善

$$P(t_i|R) = \frac{r_i + 0.5}{n_i + 1} \qquad P(t_i|\bar{R}) = \frac{O - r_i + 0.5}{N - n_i + 1}$$

$$\frac{P(t_i|R)}{P(t_i|\bar{R})} = \frac{r_i}{n_i} \times \frac{N - n_i}{O - r_i} = \frac{r_i + 0.5}{O - r_i + 0.5} \times \frac{N - n_i + 1}{n_i + 1}$$

93

信息检索

例子

查询：”gold silver truck”， 阈值为5

D1： shipment of gold damaged in a fire

D2： delivery of silver arrived in a silver truck

D3： shipment of gold arrived in a truck

D4： rice shipment

D5： truck repair

初次检索，返回D1、D2、D3

94

信息检索

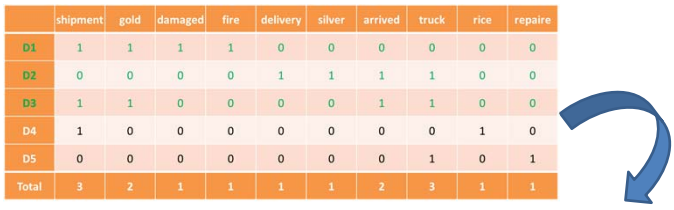
二值矩阵

	shipment	gold	damaged	fire	delivery	silver	arrived	truck	rice	repaire
D1	1	1	1	1	0	0	0	0	0	0
D2	0	0	0	0	1	1	1	1	0	0
D3	1	1	0	0	0	0	1	1	0	0
D4	1	0	0	0	0	0	0	0	1	0
D5	0	0	0	0	0	0	0	1	0	1
Total	3	2	1	1	1	1	2	3	1	1

95

信息检索

统计



变量	Gold	Silver	Truck
N	5	5	5
O	3	3	3
n_i	2	1	3
r_i	2	1	2

96

信息检索

计算

变量	Gold	Silver	Truck
N	5	5	5
O	3	3	3
n_i	2	1	3
r_i	2	1	2

$$P(gold|R)=\frac{r_i+0.5}{n_i+1}=\frac{2+0.5}{2+1}=0.83$$

$$P(gold|\bar{R})=\frac{O-r_i+0.5}{N-n_i+1}=\frac{3-2+0.5}{5-2+1}=0.375$$

$$P(silver|R)=\frac{r_i+0.5}{n_i+1}=\frac{1+0.5}{1+1}=0.75$$

$$P(silver|\bar{R})=\frac{O-r_i+0.5}{N-n_i+1}=\frac{3-1+0.5}{5-1+1}=0.5$$

$$P(truck|R)=\frac{r_i+0.5}{n_i+1}=\frac{2+0.5}{3+1}=0.625$$

$$P(truck|\bar{R})=\frac{O-r_i+0.5}{N-n_i+1}=\frac{3-2+0.5}{5-3+1}=0.5$$

97

信息检索

结果

$$Dis(RT)\approx\frac{P(t_1|R)\times P(t_2|R)\times P(t_3|R)\times\cdots P(t_n|R)}{P(t_1|\bar{R})\times P(t_2|\bar{R})\times P(t_3|\bar{R})\times\cdots P(t_n|\bar{R})}$$

$$Dis(D1)\approx\frac{P(Gold|R)}{P(gold|\bar{R})}=\frac{0.83}{0.375}=2.213$$

$$Dis(D2)=1.875$$

$$Dis(D3)=2.77$$

98

信息检索

概率模型小结

- 优点
  - 文档可以按照他们相关概率递减的顺序来排序。
- 缺点
  - 开始时需要猜想把文档分为相关和不相关的两个集合，一般来说很难
  - 实际上这种模型没有考虑索引术语在文档中的频率（因为所有的权重都是二值的）
  - 假设标引词独立
- 概率模型是否要比向量模型好还存在着争论，但现在向量模型使用的比较广泛。

99

信息检索

三种经典模型小结

	bool	扩展bool	VSM	概率
Q	布尔表达式(析取范式)	布尔表达式(析取范式)	向量	向量
D：ω取值	{0, 1}	[0,1]	[0,1]	{0,1}
F	完全匹配	非完全匹配	非完全匹配	非完全匹配
R	1（匹配） 0（不匹配）	$\sqrt{\frac{w_x^2+w_y^2}{2}}$ $1-\sqrt{\frac{(1-w_x)^2+(1-w_y)^2}{2}}$	$\sum_{k=1}^{n}(w_{ij}\times w_{qj})$	$\frac{P(R_q \mid d_i)}{P(\overline{R}_q \mid d_i)}$
可排序性	较弱	较强	较强	较强
关键词之间			独立	独立

100

信息检索

基于结构的数学模型

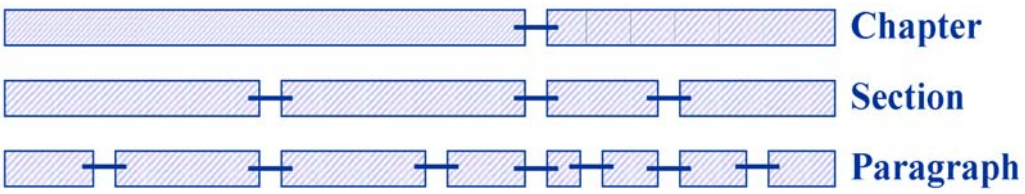
- 非重叠列表 （Non-Overlapping Lists）
- 邻近节点 （Proximal Nodes）

101

信息检索

非重叠列表

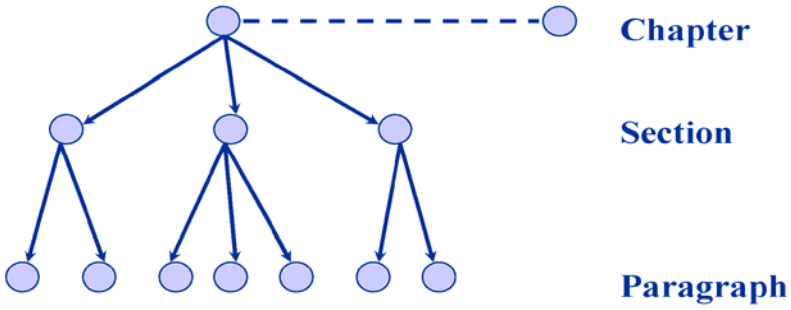
基于非重叠链表的模型是把文档中的整个文本划分为非重叠文本区域，并用链表连接起来；
相同链表中的文本区域没有重叠，而不同链表中的文本区域可能会重叠。



102

邻近节点

文档上定义一个或多个分层索引结构；每个索引结构是一个严格的层次结构。



浏览模型

- 针对浏览（browsing）文献的用户
- 具体分为三种模型
  - 扁平浏览(flat)模型
  - 结构导向(structure guided)模型
  - 超文本(hypertext)模型

扁平浏览模型

平面式浏览是指用户对平面化组织的文档结构进行探寻。对于文档的平面化组织，常见的情况有文档被表示成平面示意图中的一些点（二维），或者被表示为一个线性列表（一维）。用户在这些二维或一维的结构中，通过鼠标、方向键或滚动条等操作来对相关信息进行访问、阅读等。目前，这种浏览模式在信息检索系统的结果处理界面是最为流行的，但检索结果的平面式浏览，仅适用于检索结果数量较少的情形，对各种网络搜索引擎所提供的庞大检索结果集合，这样的浏览方式已成为对用户时间和精力的一个巨大考验。

平面式浏览实现方法简单，并且只能线性地按顺序进行或随机进行，效率较低。

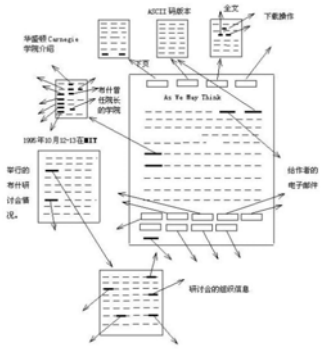
结构导向浏览模型

- 基本思想是把众多文档或信息资源组织到一个树状的类目等级体系中。
- 用户在该结构下，将由上到下，从宽泛到具体，逐步接近所需要的有用信息。
- *例如，对于一部电子图书，就可以根据其目录结构，按照章、节、小节等层次进行有关的浏览与阅读导航。*
- 层次结构式导航由于对信息集合进行了合理的分类，浏览层次与路径清晰，因而效率较高，是一种有效的浏览机制。但是，针对大规模资源集合，如何以自动方式构建其层次组织结构，目前还是一个没有完全解决的问题。

信息检索

超文本浏览模型

- 基本思想是允许以非顺序的方式在计算机屏幕上浏览文本的高层交互式导航结构。
- 由结点和链组成，构成一个有向图。
- 网络空间的迷航与超文本地图。

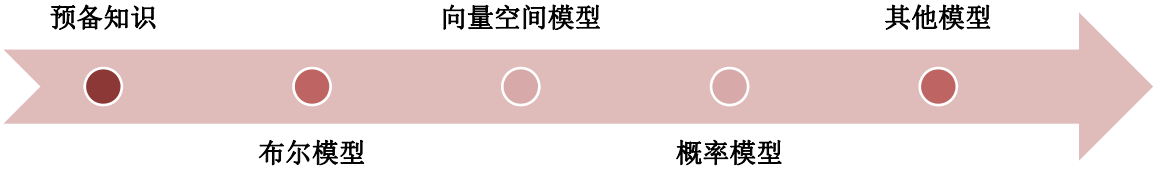


存在问题

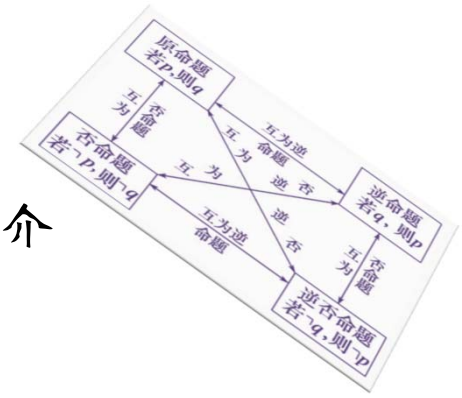
偶然发现：盲目、不可预见
失控：由网络的超链接控制
迷航：在顺连而行中偏离检索目标

信息检索

小结



命题与联结词简介



命题

能表达判断的陈述句称作命题（Proposition）。

命题：具有**确定值**

例：判断下列语句是否为命题:

- 地球外存在智慧生物。
- 1＋1=10。
- 我们去教工食堂吧！ （祈使句）
- 你今年暑假去旅行吗？ （疑问句）
- 克里特岛人说：“克里特岛人都是说谎话者”。 （悖论）

信息检索

真值

命题所表达的判断结果称为**命题的真值**。真值只有“真”和“假”两种，记作True（真）和False（假），分别用符号T和F表示。

由于命题只有两种真值，所以称这种逻辑为二值逻辑。命题的真值是具有客观性质的，而不是由人的主观决定的。

真值是否唯一确定，与是否知道无关。

111

信息检索

示例

再看下面的语句中，哪些语句是命题，如果是命题，指出它的真值：

（1）能整除7的正整数只有1和7本身。
（2）对于每一个正整数n存在一个大于n的素数。
（3）煤是白的。
（4）雪是黑的。
（5）我学英语，或者我学日语。
（6）在宇宙中地球是唯一有生命的球体。
（7）1+101=110
（8）买两张星期六的电影票。
（9）全体立正！
（10）明天是否开会？
（11）天气多好啊！
（12）我正在说谎。

112

信息检索

命题的类型

命题有两种类型：**原子命题**和**复合命题**

- 原子命题：不能分解为更简单的陈述句。
- 复合（分子）命题：由联结词、标点符号和原子命题复合构成的命题。

（1）如果路人甲是犯罪嫌疑人，那么路人甲有犯罪动机。

（2）或者路人甲是犯罪嫌疑人，或者路人甲有犯罪动机。

（3）路人乙的计算机配置合理并且价格低廉。

信息检索

联结词/Logical Connectives

在数理逻辑中，复合命题是由原子命题与逻辑联结词组合而成，命题的连接方式叫做命题联结词或命题运算符。联结词是复合命题中的重要组成部分，为了便于书写和进行推演，必须对联结词作出明确规定并符号化。

本部分仅介绍最常见的5种联结词（亦称真值联结词，逻辑联结词或逻辑运算符）。

否定（Negation）
（一元联结词）

设P为一命题，P的否定是一个新的命题，记作┐P。若P为T，┐P为F；若P为F，┐P为T。联结词“┐”表示命题的否定，称为否定联结词或否定词，读作“非”或“not”。否定联结词有时亦可记作“-”。

P	┐P
T	F
F	T

合取（Conjunction）
（二元联结词）

两个命题P和Q的合取是一个复合命题，记作P∧Q。当且仅当P、Q同时为T时，P∧Q为T，在其他情况下，P∧Q的真值都是F。联结词“∧”称为合取词，读作“和”或“and”。

P	Q	P∧Q
T	T	T
T	F	F
F	T	F
F	F	F

计算示例

① P：1+1=2
Q：地球是行星。
P∧Q：1+1=2与地球是行星。
P∧Q的真值为T。

② P：1+1=2
Q：地球是恒星。
P∧Q：1+1=2与地球是恒星。
P∧Q的真值为F。

③ P：1+1=3
Q：地球是行星。
P∧Q：1+1=3与地球是行星。
P ∧Q的真值为F。

④ P：1+1=3
Q：地球是恒星。
P∧Q：1+1=3与地球是恒星。
P∧Q的真值为F。

析取（Disjunction）
（二元联结词）

两个命题P和Q的析取是一个复合命题，记作P∨Q。当且仅当P、Q同时为F时，P∨Q为F，否则P∨Q的真值为T。联结词“∨”称为**析取词**，读作“或”或“or”。

P	Q	P∨Q
T	T	T
T	F	T
F	T	T
F	F	F

信息检索

条件（Condition）
（二元联结词）

给定两个命题P和Q，其条件命题是一个复合命题，记作P→Q，读作“如果P，那么Q”或“若P则Q”。当且仅当P的真值为T，Q的真值为为F时，P→Q的真值为F，否则P→Q的真值为T。我们称P为前件，Q为后件。

P	Q	P→Q
T	T	T
T	F	F
F	T	T
F	F	T

⊃

信息检索

例

例1 如果某动物为哺乳动物，则它必胎生。

例2 如果我得到这本小说，那么我今夜就读完它。

例3 如果雪是黑的，那么太阳从西方出。

上述三个例子都可用条件命题P→Q表达。

在例1中，P：某动物为哺乳动物，Q：它必胎生。例1表示为P→Q。

在例2中，P：我得到这本小说，Q：我今夜就读完它。P的真值为T，Q的真值为T，P→Q的真值为T。如果P的真值为T，Q的真值为F，P→Q的真值为F。如果P的真值为F，Q的真值为T，P→Q的真值为T。如果P的真值为F，Q的真值为F，P→Q的真值为T 。

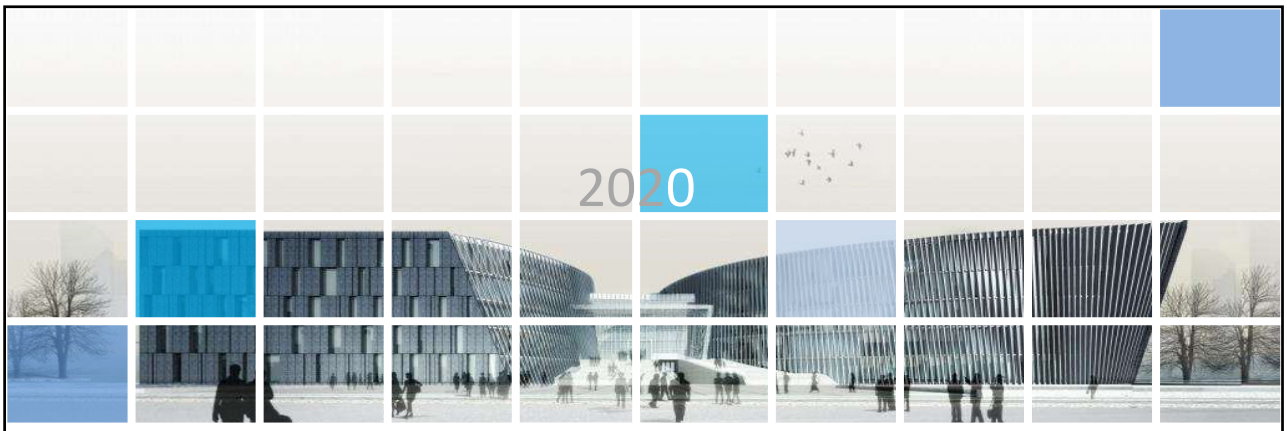
在例3中，P：雪是黑的，Q：太阳从西方出。P的真值为F，Q的真值为F，P→Q的真值为T。

双条件（Double Condition）
（二元联结词）

给定两个命题P和Q，其复合命题P $\rightleftharpoons$ Q称作双条件命题，读作“P当且仅当Q”，当P和Q的**真值相同**时，P $\rightleftharpoons$ Q的真值为T，否则P $\rightleftharpoons$ Q的真值为F。

P	Q	P $\rightleftharpoons$ Q
T	T	T
T	F	F
F	T	F
F	F	T

返回



南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

PART Six

文本信息处理

Text Information Processing

文本处理的内容

采集 → 分词 → 清洗 → 规范 → 标引 → 摘要 → 聚类

Why?

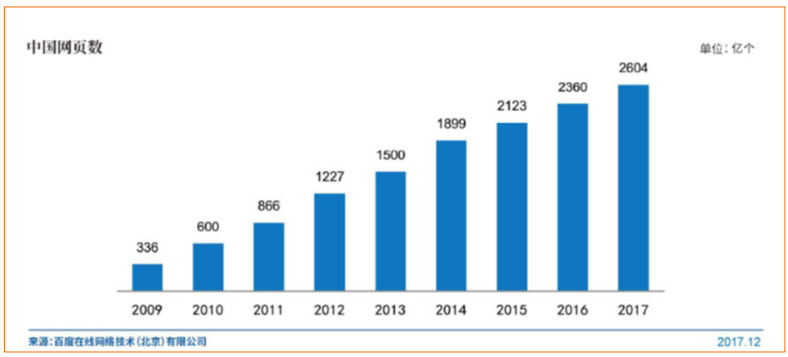
随着互联网的不断发展和日益普及，网上的信息量在爆炸性增长。在 2004 年 4 月，全球 Web 页面的数目已经超过 40 亿，中国的网页数估计也超过了 3 亿。目前人们从网上获得信息的主要工具是浏览器，而通过浏览器得到信息通常有三种方式。第一，直接向浏览器输入一个关心的网址（URL），例如 http://net.pku.edu.cn，浏览器返回所请求的网页，根据该网页内容及其包含的超链文字（anchor text）的引导，获得自己需要的内容；第二，登录到某个知名门户网站，例如 http://www.yahoo.com，根据该网站提供的分类目录和相关链接，逐步"冲浪"浏览，寻找自己感兴趣的东西；第三，登录到某个搜索引擎网站，例如 http://e.pku.edu.cn，输入代表自己所关心信息的关键词或者短语，依据返回的相关信息列表、摘要和超链接引导，试探寻找自己需要的内容。

这三种方式各有特点，各有自己最适合的应用场合。第一种方式的应用是最有针对性的，例如要了解北京大学计算机系网络与分布式系统实验室在做些什么工作，从某个渠道得知该实验室的网址为 http://net.pku.edu.cn，于是直接用它驱动浏览器就是最有效的方式。第二种方式的应用类似于读报，用户不一定有明确的目的，只是想看看网上有什么有意思的消息；当然这其中也可能是关心某种主题，例如体育比赛，家庭生活等等。第三种方式适用于用户大致上知道自己要关心的内容，例如"国有股减持"，但不清楚哪里能够找到相关信息（即不知道哪些 URL 能给出这样的信息）；在这种场合，搜索引擎能够为用户提供一个相关内容的网址及其摘要的列表，由用户一个个试探看是否为自己需要的。现在的搜索引擎技术已经能做到在多数情况下满足用户的这种需要，CNNIC 的信息统计指出，目前搜索引擎已经成为继电子邮件之后人们用得最多的网上信息服务系统。

同时，随着网上信息资源规模的增长，尤其是其内容总体和我们社会的演化发生着越来越密切的联系，研究网上存在的海量信息逐渐成为许多学科关注的一个方向。为此，不少研究人员也有采样搜集特定内容、一定数量网页的需要。

本书以我们设计、实现并维护运行北大"天网"搜索引擎的经验，介绍大规模搜索引擎的工作原理和实现技术。我们要向读者揭示，为什么向搜索引擎输入一个关键词或者短语，就能够在秒钟内得到那么多相关的文档及其摘要，而点击其中的链接就能够被引导到文档的全文，且其中相当一部分可能正是用户需要的。

我们按照上、中、下三篇展开相关的内容。上篇讲搜索引擎的基本工作原理，要解决的是为什么搜索引擎能提供如此信息查找服务的问题，以及它在功能上有什么本质的局限性。这一篇的内容包括网页的搜集过程，网页信息的提取、组织方式和索引结构，查询提交和响应的过程以及结果产生，等



CNNIC统计，2018.2

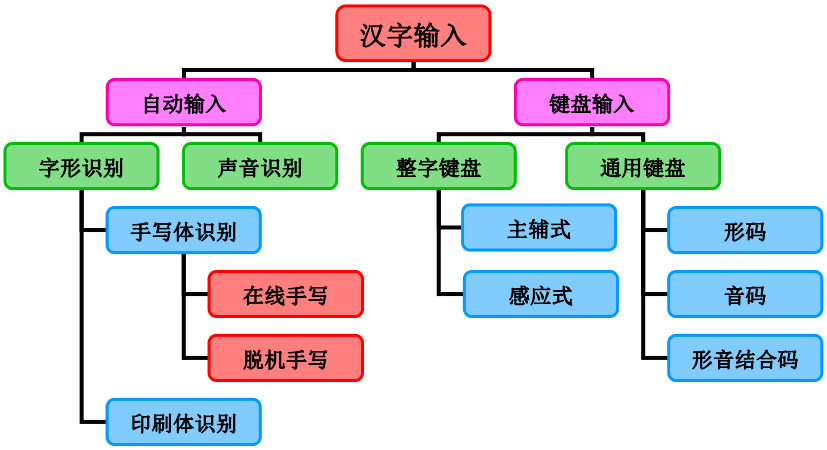
中文信息处理发展史

张华平. 中文信息处理技术发展简史
张普. 共和国的中文信息处理60年

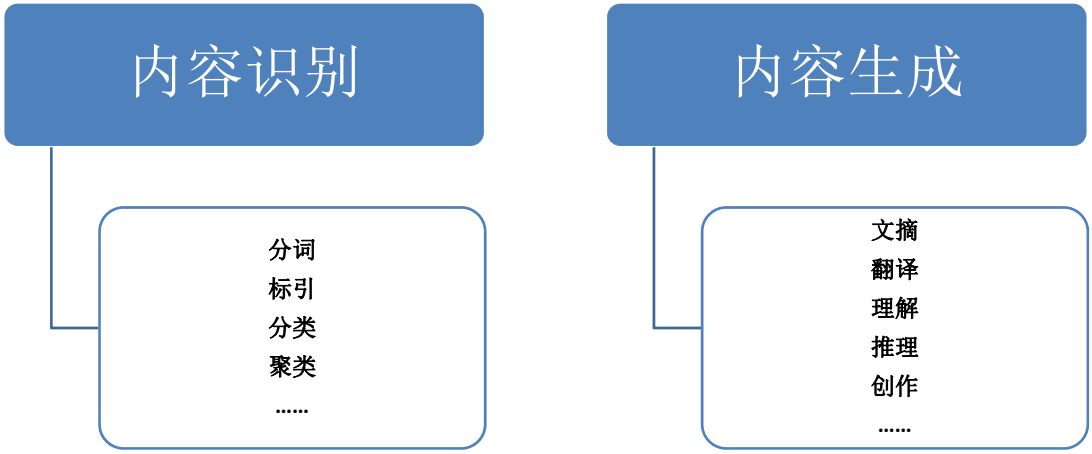
中文信息处理≠中文文本信息处理

- 信息的两个层次
  - 符号层 —— 中文 / 汉语 / 汉字
  - 内容层 —— 符号所承载的意义
- 中文信息处理的两个层次
  - 字符处理（输入、存储、输出等）
  - 内容处理（词语切分，词性标注，结构分析，意义理解，推理，翻译……等等）

符号层的汉字处理技术

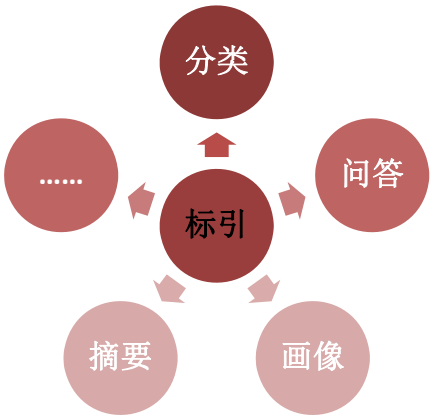


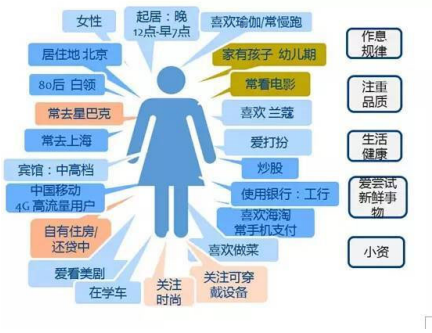
内容层的信息处理



信息检索

标引很重要!





9

信息检索

（人工）标引流程

主题分析 → 排重 → 标引 → 审核 → 记录结果

10

自动标引的定义

文献标引：指对所收集的文献给出标识导引，这些标识包括文献标题、作者名、分类号和主题词等。

文献标引作业流程

- 文献文本分析
- 特征信息（主题词、关键词及其他标识）的提取与描述
- 建立索引或倒排档

自动标引（Automatic Indexing）：自动标引就是用**机器抽取或赋予索引词**，一旦编制好程序和规则，就**不需要人工干预**。

自动标引的意义

- 适应信息资源快速增长的需要
- 效率高、成本低
- 稳定性好、一致性好

什么是词？

最小的能够独立运用的语言单位。

——《信息处理用现代汉语分词规范》

缺乏操作标准。

国家标准 GB/T 13715-92
《信息处理用现代汉语分词规范》

1 主题内容与适用范围
1.1 主题内容
本规范规定了现代汉语的分词原则，以满足信息处理的需要。它对汉语信息处理的规范化，对各种汉语信息处理系统之间的兼容性有重要的作用。
1.2 适用范围
本规范适用于汉语信息处理各领域，其他行业和有关学科可以参考使用。
汉语信息处理各领域可以根据其专门需求，进一步补充和细化本规范的规定。

2 引用标准
GB12200 汉语信息处理词汇

3 术语
以下术语引自 CB 12200。
3.1 汉语信息处理
用计算机对汉语的音、形、义等信息进行的处理。
3.2 词
最小的能独立运用的语言单位。
3.3 词组
由两个或两个以上的词，按一定的语法规则组成，表达一定意义的语言单位。
3.4 分词单位
汉语信息处理使用的、具有确定的语义或语法功能的基本单位。它包括本规范的规则限定的词和词组。
3.5 汉语分词
从信息处理需要出发，按照特定的规范，对汉语按分词单位进行划分的过程。

分词？明句读！

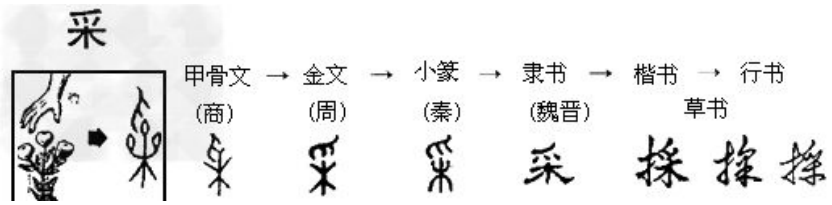
下雨天留客天留我不留

中文（自动）分词就是要由机器在中文文本中词与词之间加上标记

关于汉字和汉语

- 汉语文本是基于单字的，汉语的书面表达方式也是以汉字作为最小单位的，词与词之间没有显性的界限标志，因此分词是汉语文本分析处理中首先要解决的问题
- 汉字
  - 表意
  - 简洁、严谨
  - 无时态、语态、性、格的变化

- 《说文解字》（东汉）：9353字
- 《玉篇》（南朝）收录16,917字
- 《广韵》（宋代）收字26,194字
- 《字汇》（明朝）收录33,197字
- 《康熙字典》（清朝）收录47,043字
- 《汉语大字典》（1992年）5.6万
- 《中华字海》（1994年） 8.6万



分词的意义

•正确的机器自动分词是正确的中文信息处理的基础

◆**文本检索**

- *和服 / 务 / 于三日后裁制完毕，并呈送将军府中。*
- *王府饭店的设施 / 和 / 服务 / 是一流的。*

◆文语转换

- *他们是来 / 查 / 金泰 / 撞人那件事的。（“查”读音为cha）*
- *行侠仗义的 / 查金泰 / 远近闻名。（“查”读音为zha ）*

◆词频统计（汉语中最常用的词是哪个？）

◆句法分析、语义分析、机器翻译、语音合成、自动分类、自动摘要、自动校对……

英语不需要分词？

- 英语中不是完全没有词语切分问题，不能仅凭借空格和标点符号解决切分问题。
  - 缩写词
    N.A.T.O　i.e.　m.p.h　Mr.　AT&T
  - 连写形式以及所有格结尾
    I'm　He'd　don't　Tom's
  - 数字、日期、编号
    128,236　+32.56　-40.23　02/02/94　02-02-94
  - 带连字符的词
    text-to-speech　text-based　e-mail　co-operate
- 英语中的切分通常被叫做**Tokenization**
- 和中文相比，英语切分问题较为容易

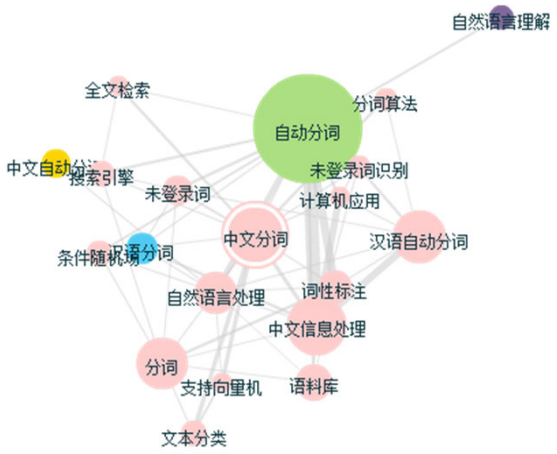
分词的方法

- 基于词典
  - 基于字符串（词）匹配的分词方法：按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行配，若在词典中找到某个字符串，则匹配成功 。可以切分, 否则不予切分。
  - 实现简单，实用性强，但机械分词法的最大的缺点就是词典的完备性不能得到保证。
- 无词典
  - 基于统计（n-gram、HMM）
  - 基于规则
  - 基于字标注

信息检索

分词的难点

- 歧义消解
- 未登录词识别



信息检索

切分歧义类型

- **交集型歧义**
  - 对于汉字串AJB，AJ、JB同时成词
  - 例：结合/成，结/合成　　乒乓球拍卖完了
- **组合型歧义**
  - 对于汉字串AB，A、B、AB同时成词
  - 例：门/把手/坏/了，请/把/手/拿/开
- **混合型歧义**
  - 同时包含交集型歧义和组合型歧义
  - 例：　这样的/人/才能/经受住考验
    这样的/人才/能/经受住考验
    这样的/人/才/能/经受住考验
- 中文文本中，交集型歧义与组合型歧义出现的比例约为1:22。

切分歧义真伪

- **真歧义**
  - 歧义字段在不同的语境中确实有多种切分形式
  - 例：　地面积
    这块/地/面积/还真不小
    地面/积/了厚厚的雪
- **伪歧义**
  - 歧义字段单独拿出来看有歧义，但在所有真实语境中，仅有一种切分形式可接受
  - 例：　挨批评
    挨/批评（√）　挨批/评（×）
- 对于交集型歧义字段，真实文本中伪歧义现象远多于真歧义现象

伪歧义	94%			
真歧义	6%	多种切分形式均匀分布	0.72%	将技术/应用/于/项目　精力/应/用于/学习
		一种切分形式占优	5.28%	解除/了/职务　方程的/解/除了/0还有1

未登录词

- 虽然一般的词典都能覆盖大多数的词语，但有相当一部分的词语不可能穷尽地收入系统词典中，这些词语称为**未登录词**或**新词**
- 类别：
  - 专有名词：中文人名、地名、机构名称、外国译名、时间词
  - 重叠词：“高高兴兴”、“研究研究”
  - 口语：“吃饭饭”、“潘西”
  - 派生词：“一次性用品”
  - 与领域相关的术语：“互联网”

我/看见/周星驰/同/张学友/打招呼

I saw Stephen Chow greeting Jacky Cheung

我/看见/周星驰/同/章/学友/打招呼

I saw Stephen Chow greeting Zhang school friends

正向最大匹配

Forward Maximum Matching method, FMM

1. 设自动分词词典中最长词条所含汉字个数为I；
2. 取被处理材料当前字符串序数中的I个字作为匹配字段，查找分词词典。若词典中有这样的一个I字词，则匹配成功，匹配字段作为一个词被切分出来，转6；
3. 如果词典中找不到这样的一个I字词，则匹配失败；
4. 匹配字段去掉最后一个汉字，I--；
5. 重复2-4，直至切分成功为止；
6. I重新赋初值，转2，直到切分出所有词为止。

特点

- “市场/中国/有/企业/才能/发展/”
- 对交叉歧义和组合歧义没有什么好的解决办法
- 错误切分率为1 / 169
- 往往不单独使用，而是与其它方法配合使用

逆向最大匹配分词

Backward Maximum Matching method, BMM

- 分词过程与FMM方法相同，不过是从句子(或文章)末尾开始处理，每次匹配不成功时去掉的是前面的一个汉字
- “市场/中/国有/企业/才能/发展/
- 实验表明：逆向最大匹配法比最大匹配法更有效，错误切分率为1／245

双向匹配法

Bi-direction Matching method, BM

- 比较FMM法与BMM法的切分结果，从而决定正确的切分
- 可以识别出分词中的交叉歧义
- 算法时间、空间复杂性较高

常用规则

- **颗粒度越大越好**
  - 公安局长:“公安/局长”、“公安局/长”、“公安局长”
- **非词典词越少越好**
  - 技术和服务：“技术/和服/务” 、“技术/和/服务”
- **总体词数越少越好**

- 我们在野生动物园玩
  - FMM：我们/在野/生动/物/园/玩
  - BMM：我们/在/野生动物园/玩

基于理解的分词

- 通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。
- 由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统多处在试验阶段。

基于统计的分词

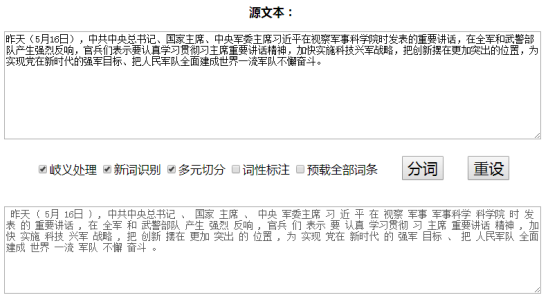
- 基于统计的分词方法：基本原理是根据字符串在语料库中出现的统计频率来决定其是否构成词。
- 无词典分词法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字符串，如“这一”、“之一”以及“提供了”等等。
- 在实际应用的统计分词系统中都要使用一部基本的分词词典（常用词词典）进行串匹配分词，即将字符串的词频统计和字符串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

基于字标注的分词

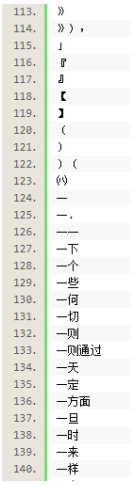
- 把分词过程视为字在字串中的标注问题。由于每个字在构造一个特定的词语时都占据着一个确定的构词位置（即词位），假如规定每个字最多只有四个构词位置：即B（词首），M（词中），E（词尾）和S（单独成词），那么下面句子甲的分词结果就可以直接表示成如乙所示的逐字标注形式：
  - (甲)分词结果：/上海/计划/本/世纪/末/实现/人均/国内/生产/总值/五千美元/。
  - (乙)字标注形式：上/B 海/E 计/B 划/E 本/S 世/B 纪/E 末/S 实/B 现/E 人/B 均/E 国/B 内/E 生/B 产/E 总/B 值/E 五/B 千/M 美/M 元/E 。/S
- 基于字标注的方法通过改进未登录词识别能力，提升了分词系统的总体性能。
- “基于字标注的方法+机器学习”成为中文分词主流技术，算法复杂度较高。

常见分词工具

- Ictclas（NLPIR）
  - http://ictclas.nlpir.org/
- Jieba
  - https://github.com/fxsjy/jieba
- SnowNLP
  - https://github.com/isnowfy/snownlp
- THULAC
  - https://github.com/thunlp/THULAC-Python



停用词



- Stop Words/**禁用词、非用词**
- 在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词。
- **停用词表**是一种特殊的词表，在这个词表中含有冠词、虚词、叹词、连词、介词以及语义泛泛的词等一切在上下文中没有检索意义的词
  - 哈工大停用词词库
  - 四川大学机器学习智能实验室停用词库
  - 百度停用词表
  - ……

（英文）词干化

- Stemming
- 英语单词有时态、单复数等等变化
  - 如beat/beating/beated
- 有成熟工具
  - coreNLP（哈佛）
  - SnowballStemmer（Python）

```
原句 ：jack had been to china there months ago. he likes china very much,and he is falling love with this country
词干化: jack have be to china there month ago . he like china very much , and he be fall love with this country
```


自动标引的类型

从标引工作的**自动化程度**来说，自动标引分为：
- 全自动标引（Automatic Indexing）；
- 半自动标引（机助标引：Automated indexing/Computer Aided Indexing）。

半自动标引基本技术实现是：
（1）文献纪录（题目等著录项目）键入终端后显示在屏幕上；
（2）操作人员移动光标从题目中抽取关键词；
（3）利用人机对话方式输入与标题内容有关的隐含概念词，以保证主题标引的全面性；同时删除计算机程序错误组配的词。
（4）根据词库中的参照系统将关键词转换成标准主题词，进行上位登录。
词库（主题词表）是计算机辅助标引的核心。

从**标引词的来源**去划分，自动标引分为：
- （自动）抽词标引
- （自动）赋词标引

抽词标引

➢ **自动抽词标引/自由词标引**

利用计算机直接从文献题名、文摘或正文中自动抽出能表达文献主题的词作为标引词，并自动生成关键词索引或倒排档。抽词标引的标引词只能来源于文献本身的文内关键词，所以也称为**自由词标引**。

类别：

- **主关键词**标引：要求计算机从抽出的全部关键词中选出少量主要关键词做索引词。
- **全关键词**标引：把除停用词以外的全部关键词抽出，直接做索引词。

优点：无需主题切换，接近自然语言。

缺点：

- 标引用词不规范，影响查全率；
- 同义词检索降低系统的时间效率；
- 难以找出词和词之间的相互关系，很难进一步利用语义信息。

赋词标引

自动赋词标引/受控词标引

让计算机模仿人的赋词标引方法，分析文献的内容，选取与文献主题相符或密切相关的语词符号作为索引词。其标引词是由描述词组成的，这些词不一定来源于文献本身所用的词，而是选自**预先编制的词表**，所以叫**受控词标引**。

优点：

- 规范化用词
- 词表可以反映词的“类－属”关系。

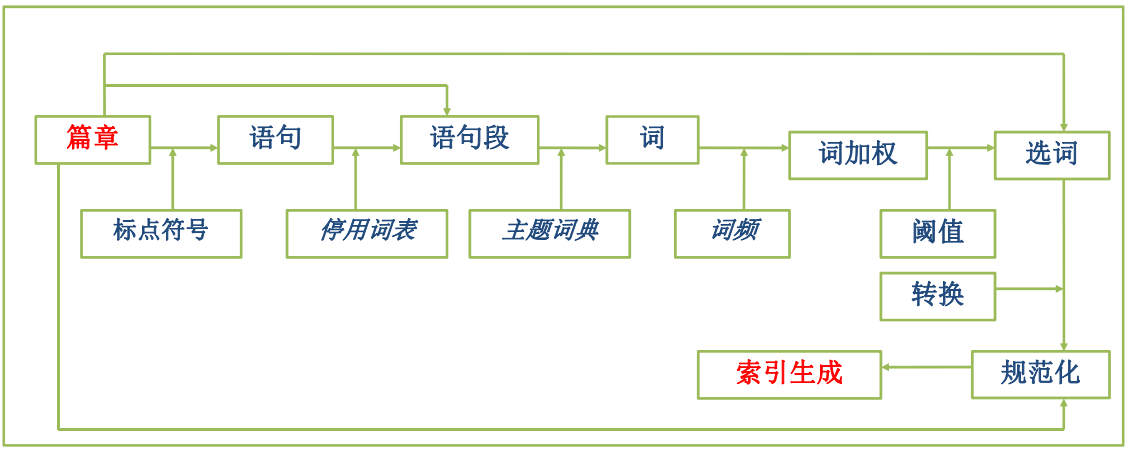
缺点：

- 受控词标引往往有一定的标引误差；
- 词典面临老化的问题；
- 主题词表对用户来说往往是一个负担；

自动赋词标引是在自动抽词标引的基础上发展起来的。
最合理的标引方法：混合标引方法

例

自动标引的流程



标引源

- 全文
  - 数据量大，处理麻烦
- 标题
  - 主要标引源；信息量少，歧义多，标引质量差
- 文摘
  - 主要标引源，大部分情况下够用
- 首尾章节
- 章节的首尾段
- 段落的首尾句

确定关键词

- **统计法**
  - 绝对词频统计法
    - 以词在文章中出现的绝对频次为根本依据确定文章的中心关键词，理论基础是**齐夫定律**。
  - 词频权重法
    - 除考虑词频外，还考虑词的位置、词的词性、词本身的价值、词的长度等因素，对词进行加权，然后根据权值大小确定关键词。
  - 机器学习标引（统计学习）
- 语言法
  - 句法分析法
  - 语义分析法
- 人工智能法

词频统计标引法

- 1958年由美国学者卢恩（Luhn）提出。
- 主要思想：词在文献中的出现频率是该词对该篇文献重要性的有效指标，文献中只有词频介于高频和低频之间的那部分中频词最适合作为标引词。
- 词频统计标引法的理论依据是齐夫第一定律、齐夫第二定律。

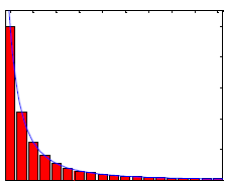
Zipf定律

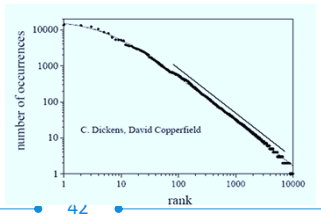
Zipf's Law

词频分布定律

如果把一篇较长文章（>5000）中每个词出现的频率统计起来，按照高频词在前、低频词在后的递减顺序排列，并用自然语言给这些词编上等级序号，即频次最高的词的等级为1，频次次高的等级为2，……，频次最小的词等级为D（或L），若用f表示等级为r的词在文献中出现的**相对频次**，则有：$f_r \cdot r = C$ （C是一个常数，大约等于0.1）.

齐普夫分布曲线：如果用横坐标表示词的等级序号r，纵坐标表示相应的频次fr，我们就可以得到一条双曲线，即齐普夫分布曲线。





$$f = cr^{-\gamma}$$

George Kingsley Zipf（1902–1950）

Luhn标引算法

- 给定m篇文献组成的一个集合，设第k个词在第i篇文献中发生的频率f_{ik}。
- 决定该词在整个文献集上的发生频率：$f_k = \sum f_{ik}$
- 按照f_k的大小将词降序排列，
- 用试错法确定高频词和低频词的阈值。
- 去掉高频词和低频词后，将余下的中频词选作标引
  - Luhn在自动标引中使用的文献，长度在500～5000字之间，为每篇文献选择的标引词数量定在10～24个词之间。近似平均值为16。

世界上第一个自动标引算法

Zipf第二定律

- 低频词定律
- 文章中词频为n的词与词频为1的词数量上有数学关系

$$\frac{I_n}{I_1} = \frac{3}{4n^2 - 1}$$

- A.D.Booth修正为

$$\frac{I_n}{I_1} = \frac{2}{n^2 + n}$$

改进的标引方法

- 存在一个词由高频行为转为低频行为的临界区（critical region），只有处于临界区内的词才最适于描述文献的主题（Goffman）。

$$n = (-1 + \sqrt{1 + 8I_1}) / 2$$

- 以n为临界区的中点，以最高词频处为临界区的上界，取与n到上界之间等级距离相等的另一端为临界区的下界，位于临界区内的词经过筛选即可选为标引词。

标引词加权

词的权值（Weight）一般表示该词的重要程度 。

标引词加权：就是根据标引用词（符号）所代表的内容在文献中的地位和作用的大小（或说与文献的亲疏程度）给予这些词（符号）以相应的数值。

- 词频加权法
  - 绝对频率加权法
  - 相对频率加权法
- 词位置加权法
- 其他加权的方法

相对频率加权法

- 考虑的因素：
  - 词在某个特定文献内的使用频次
  - 词在特定文献集内的使用频次
- 方法：
  - Tf-idf
  - 逆文档频率加权法

逆文档频率加权法

$$W_{ik} = \frac{F_{ik}}{DF_k}$$

$$DF_k = \sum_{i=1}^{n} df_{ik} \qquad df_{ik} = \begin{cases} 1 & F_{ik} \geq 1 \\ 0 & F_{ik} = 0 \end{cases}$$

- F_{ik}为词k在文献i中的出现频率；
- DF_k为词k的**文档频率**。

标引词的权与标引词的文档频率有互逆关系，因此这种标引加权方法叫“逆文档频率加权法”，根据这种加权方法进行的标引叫“逆文档频率加权标引”。

位置加权法

根据词的位置进行加权的方法称为**位置加权法**。

1）标题
2）文摘
3）首尾章节
4）章节的首尾段
5）段落的首尾句

位置	权值
主标题中词汇	2.0
其他标题中词汇	1.8
文摘中的词汇	1.6
首尾章节词汇	1.3
首尾段（句）词汇	1.1
其他位置词汇	1.0

其它加权的方法

1）词性

2）词本身的价值

3）词的长度

4）词的特定位置，如：

- 文献中用括号括起来的部分：ISDN（综合业务数据网）；
- 用破折号引出来的部分，“数据的自动识别输入——条码技术”；
- 用“所谓”等所引出的部分，如“所谓的预置关键词”，其中的实词往往也应当给予特别的加权。

5）词的颜色、字体等（Web）

……

标引结果

- 抽词标引
  - 直接标注标引结果
- 赋词标引
  - 关键词与受控词（主题词、副主题词、特征词）之间存在着一定的关系（如同义词关系、上位关系、下位关系等）。
  - 使用一定的方法，将以上提取的关键词转换为受控词。
    - 使用关键词-受控词对照表
    - 利用词汇相似度

单汉字标引

- 自然语言标引
- 在标引时将概念词拆成单汉字，以单汉字为处理单位，利用汉字索引文件实现自动标引和逻辑检索
- 处理过程：计算机对处理的文本逐一抽字，并去掉无意义的虚字；对剩下的字建立单字索引文件。
  - 搜索引擎就是单汉字标引，且不去虚词
- **优点**
  - 不分词，简单容易
  - 字匹配，查全率高
- **缺点**
  - 索引规模大
  - 速度慢

分类与聚类

Classification and Clustering

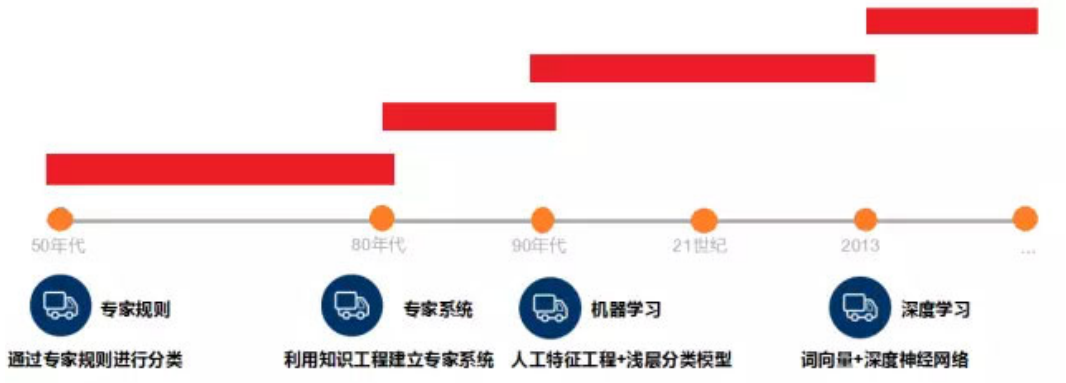
淳于髡

BC 386～BC 310

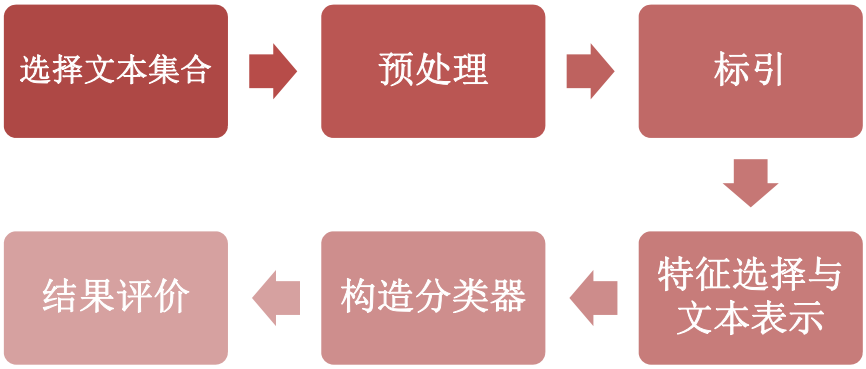
（广义）自动分类概述

- **自动聚类：**从待分类对象中提出特征，然后将提出的全部特征进行比较，再根据一定的原则将具有相同或相近特征的对象定义为一类，并设法使各类中包含的对象大致相等；特点是“先有文档后有类”。（Unsupervised Learning）
- **自动归类（狭义的自动分类）：**指在给定的分类体系下，分析被分类对象的特征，使之与各种类别中对象所具有的共同特征进行比较，然后将对象划归为特征最接近的一类并赋予相应的分类号。特点是“先有类（表）后有文档”。（Supervised Learning）
- **类号的自动转换：**针对多部分类法并存的现状而提出的，有利于分类标准化。

文本分类发展简史



文本分类的步骤



应用

- 垃圾邮件的判定(spam or not spam)
  - 类别{spam，not-spam}
- 新闻出版按照栏目分类
  - 类别{政治，体育，军事，…}
- 词性标注
  - 类别{名词，动词，形容词，…}
- 词义排歧
  - 类别{词义1，词义2，…}
- 计算机论文的领域
  - 类别 ACM system
    - H: information systems
    - H.3: information retrieval and storage

文本分类的模式

- 从类别数目来分
  - –2类（binary）问题，类别体系由两个互补类构成，一篇文本属于或不属于某一类。
  - –多类（multi-class）问题，类别体系由三个或者以上的类别构成，一篇文本可以属于某一个或者多个类别，通常可以通过拆分成多个2类问题来实现，也有直接面对多类问题的分类方法
- 从是否兼类看分
  - –单标签（single label）问题：一个文本只属于一个类
  - –多标签（ multi-label）问题：一个文本可以属于多类，即出现兼类现象

关于分类体系

- 分类体系的构建可以是按照语义，如（政治、经济、体育），也可以按照其他标准（中文的、英文的；）
- 分类体系一般由人工构造，可以是层次结构，如中图分类
- 对于计算机而言，分类体系就是一个目录树，训练样本文本就是最后的叶节点，对计算机而言，只需要训练样本及其对应类别信息，不需要考虑类别标签的意义。

人工方法和自动方法

- **人工方法：**（人工总结规则）
  - **优点**
    - 结果容易理解；如足球and 联赛+体育类
  - **缺点**
    - 费时费力
    - 难以保证一致性和准确性(40%左右的准确率)
  - 专家有时候凭空想象，没有基于真实语料的分布
  - 代表方法：人们曾经通过知识工程的方法建立专家系统(80年代末期)用于分类。
- **自动的方法**(学习从训练语料中学习规则)
  - **优点**
    - 快速
    - 准确率相对高(准确率可达60%或者更高)
    - 来源于真实文本，可信度高
  - **缺点**
    - 结果可能不易理解（比如有时是一个复杂的数学表达式）

规则方法与统计方法

- 规则方法通过得到某些规则来指导分类，这些规则往往人可以理解
- 统计方法通过计算得到一些数据表达式来指导分类
- 二者无本质区别，都是试图得到某种规律性的东西来指导分类，统计方法得到的数学表达式也可以算是一种规则
- 目前统计方法是主流

过程

- 两个步骤
  - 训练（trainning，即从训练样本集中学习分类的规律）
  - 测试（test，根据学习来的规律对新文本进行类别判定）
- 文本表示
  - 无论训练还是测试，都得先分析文本的特征（feature，或标引项term），然后把文本变成这些特征的某种适于处理的形式，通常采用向量表示形式或直接使用某种统计量

特征提取

- 预处理
  - 分词、去停用词、词干化、……
- 文本表示
  - 向量空间模型
- 降维技术
  - 特征选择（Feature Selection）
  - 特征重构（Re-parameterisaton，如LSI，潜在语义索引）

Term的粒度

Character，字：中/国/人/民/银/行
Word，词：中国/人民/银行
Phrase ，短语：中国人民银行
Concept，概念
同义词：开心/高兴/兴奋_
相关词cluster，word cluster ：葛非/顾俊
N-gram，N元组：中国/国人/人民/民银/银行

许多学者认为：（英文分类中）使用优化合并后的Words比较合适

Feature Selection

- 在文本分类问题中遇到的一个主要困难就是高维的特征空间
  - 一份普通的文本在经过文本表示后，如果以字/词为特征，它的特征空间维数将达到几千，甚至几万
  - 大多数学习算法都无法处理如此大的维数
- 在不牺牲分类质量的前提下尽可能降低特征空间的维数
- 特征选取的任务将信息量小，不重要的词汇从特征空间中删除，减少特征项的个数
- 在许多文本分类系统的实现中都引入了特征提取方法

Feature Selection

- 对每类构造k 个最有区别能力的term
- 例如：
  - 计算机领域：
    - *主机、芯片、内存、编译* …
  - *信息管理：*
    - *检索，索引，分类，摘要，* …

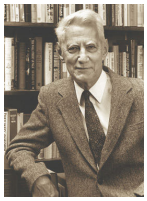
特征选择-DF

- 基于DF（Document Frequency）的选择
  - DF小于某个阈值的去掉，没有代表性
  - DF大于某个阈值的去掉，没有区分度
  - 优点：降低向量计算的复杂度，去掉部分噪声，提高分类的准确率，且简单易行。
  - 缺点：稀少的词具有更多的信息，因此不宜用DF大幅度地删除词

Term的熵（Entropy）

$$Entropy(t) = -\sum_i P(c_i \mid t) \log P(c_i \mid t)$$

- 该值越大，说明分布越均匀，越有可能出现在较多的类别中
- 该值越小，说明分布越倾斜，词可能出现在较少的类别中



Claude Elwood Shannon
1916—2001

特征选择-IG

- 信息增益（Information Gain，IG）
  - 某term为整个分类所能提供的信息量（不考虑某特征的熵和考虑该特征的熵的差值）
    - 1.计算不含任何特征整个文档的熵;
    - 2.计算包含该特征的文档的熵
    - 3.前者-后者
    - 4.选择Top K作为特征
  - 优点：准，选择的特征是对分类有用的特征
  - 缺点：有些信息增益较高的特征出现的频率较低

特征选择- CHI（χ^2）

- 概念：CHI衡量的是特征项t(i)和类C(j)之间的相关联程度。假设t(i)和C(j)之间符合具有一阶自由度的卡方分布，如果特征对于某类的卡方统计值越高，它与该类之间的相关性越大，携带的信息越多，反之则越少。
- 特点：**只统计文档是否出现词，而不管出现了几次。夸大了低频词的作用。**

	c	~c
t	A	B
~t	C	D

文本的总数为N=A+B+C+D

$$\chi^2(t,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

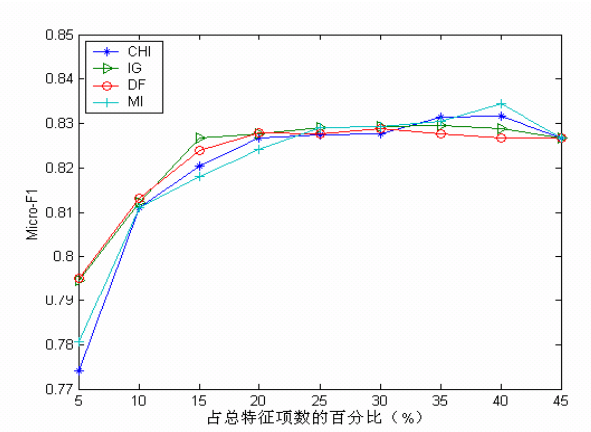
特征选择-MI

- MI（Mutual Information）指互信息，越大，则特征t(i)和类C(j)之间共同出现的程度越大，如果两者无关，那么互信息=0。
- 步骤：两种方法，和CHI一样，最大值方法和平均值法
- 优点：如果某个特征词的频率很低，那么互信息得分就会很大，因此互信息法倾向"低频"的特征词。
- 缺点：相对的词频很高的词，得分就会变低，如果这词携带了很高的信息量，互信息法就会变得低效

$$I(t)=\sum_i P_r(c_i)\log\frac{P_r(t\mid c_i)}{P_r(t)}$$

对比

- 可以看出CHI，IG，DF性能好于MI
- MI最差
- CHI，IG，DF性能相当
- DF具有算法简单，质量高的优点，可以替代CHI，IG



文档表示

	T_1	T_2	....	T_t
D_1	d_{11}	d_{12}	...	d_{1t}
D_2	d_{21}	d_{22}	...	d_{2t}
:	:	:		:
:	:	:		:
D_n	d_{n1}	d_{n2}	...	d_{nt}

指把文本转换为易被计算机理解的形式。即在对文本进行分词等预处理之后把文本表示为计算机可以识别的格式。

文本表示的模型有：

布尔逻辑模型（Boolean Model）

向量空间模型（VSM，Vector Space Model）

潜在语义索引（LSI，Latent Semantic Indexing）

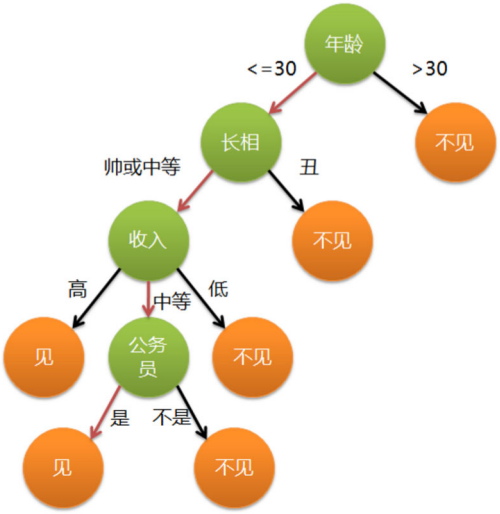
概率模型（Probabilitic Model）

.....

分类方法

- 基于规则
  - ✓ **决策树（Decision Tree）**
  - ✓ 决策规则（Decision Rule Classifiers）
- 基于统计
  - ✓ 回归（Regression）
  - ✓ kNN
  - ✓ SVM
  - ✓ Rocchio
  - ✓ 朴素贝叶斯（Naïve Bayes）
  - ✓ 多重神经网络（Neural Networks）
  - ✓ 基于投票的方法（Voting methods）
  - ✓ Online Linear Classifiers

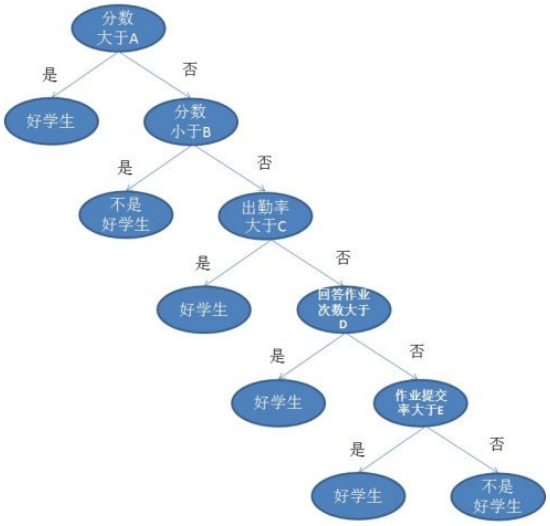
决策树

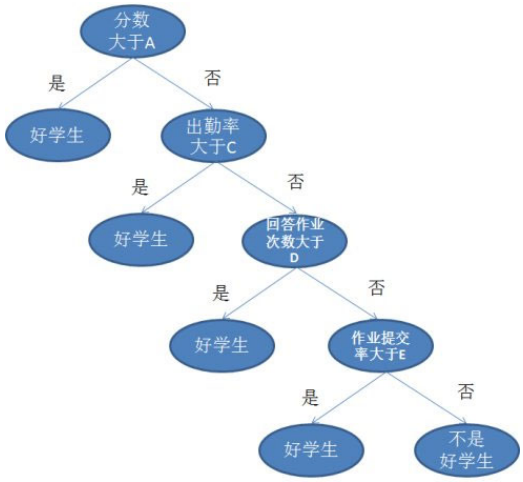


示例

学生编号	分数	出勤率	回答问题次数	作业提交率	分类:是否好学生
1	99	80%	5	90%	是
2	89	100%	6	100%	是
3	69	100%	7	100%	否
4	50	60%	8	70%	否
5	95	70%	9	80%	否
6	98	60%	10	80%	是
7	92	65%	11	100%	是
8	91	80%	12	85%	是
9	85	80%	13	95%	是
10	85	91%	14	98%	是

结果





KNN

- k-Nearest Neighbor
- 给定一个经过分类的训练文档集合，在对新文档（即测试文档或待分类文档）进行分类时，首先从训练文档集合中找出与测试文档**最相关的k篇文档**，然后按照这**k**篇文档所属的类别信息来对该测试文档进行分类处理。

步骤

- 目标：基于训练集X的对y分类
- 在训练集中，寻找和y最相似的训练样本x

$$sim_{MAX}(y) = MAX_{x \in N} sim(x, y)$$

- 得到k个最相似的集合A，A为X的一个子集

$$A = \{x \in N \mid sim(x, y) = sim_{\max}(y)\}$$

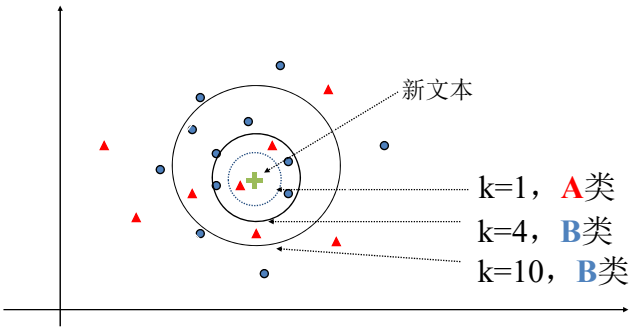
- 设n1，n2分别为集合中属于c1，c2的个数

$$p(c_1 \mid y) = \frac{n1}{n1+n2} \qquad p(c_2 \mid y) = \frac{n2}{n1+n2}$$

- 如果p(c1|y)>p(c2|y)，判为c1，否则判为c2

79

示例



k常取3-15，经常需要测试。

分类的评测

	属于此类	不属于此类
判定属于此类	a	b
判定不属于此类	c	d

准确率(precision) = a/(a + b)
召回率(recall) = a/(a + c)

$$F_1 = \frac{2\ pr}{p\ +\ r}$$

自动聚类技术概述

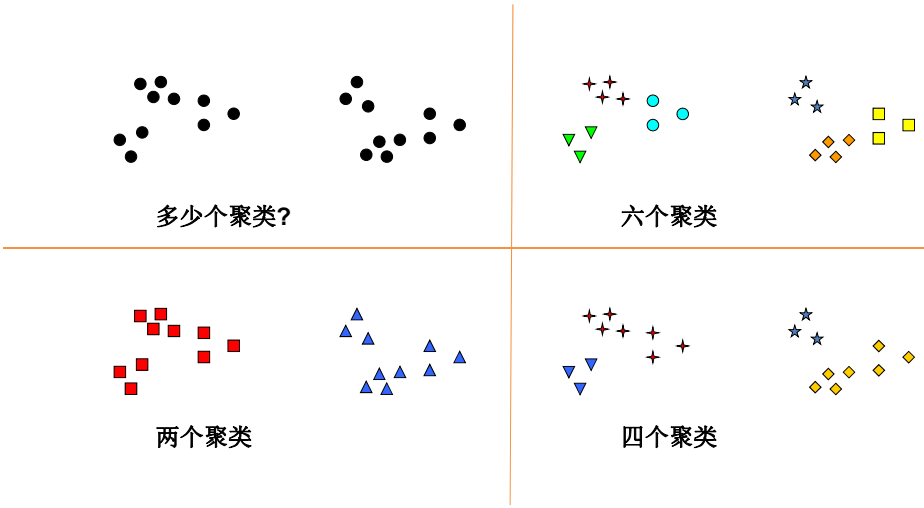
“文本聚类”（text clustering），就是完全根据文本文档的内容相关性来组织文档集合，将整个集合分成若干个类，并使得属于同一类的文档尽量地相似，属于不同类的文档差别明显。

聚类的描述：

- （文档）聚类是将一系列文档按照相似性聚团成子集或者簇(cluster)的过程
- 簇内文档之间应该彼此相似
- 簇间文档之间相似度不大
- 聚类是一种最常见的无监督学习(unsupervised learning)方法。

聚类假设：在考虑文档和信息需求之间的相关性时，同一簇中的文档表现互相类似。

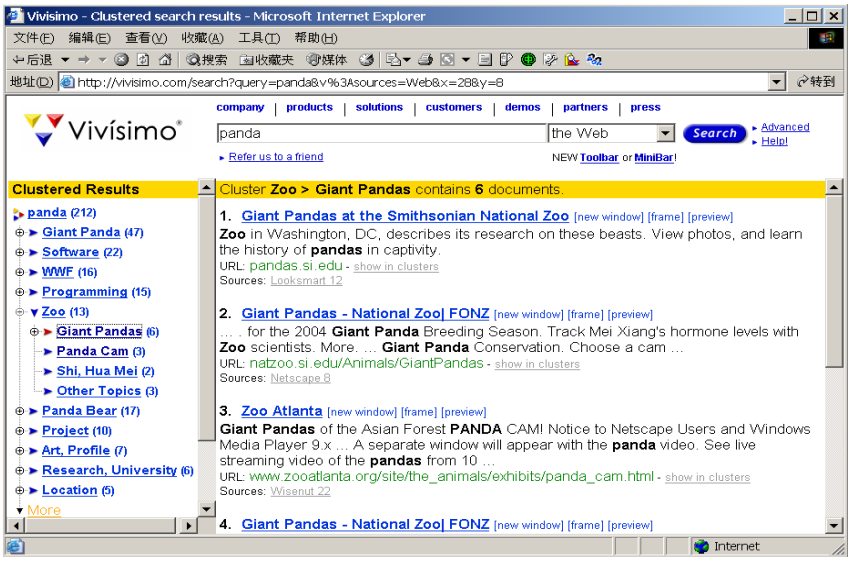
聚类－模糊性



聚类对信息检索的意义

- 提高检索的查准率
  - 相似文档一般对相同查询相关度比较高。
- 提高向量空间模型的检索速度
  - 物理上或逻辑上相似的文档放在一起形成一个类，因此有利于提高检索效率。
- 提供导航
  - 把多个文档聚集在一起，提供用户在聚类层次上浏览文档集，而不需要浏览每篇文章。因此有助于帮助用户快速定位到有用的信息区域。

检索聚类示例



聚类的核心问题

- 文档表示
  - Bag of words, VSM
- 相似性度量
  - 理想：语义上相似
  - 实际：统计上相似
  - 度量方法
    - Cos()
    - 欧式距离
- 算法

聚类算法一览

- 层次算法（Hierarchical algorithms）
  - 自低向上（*bottom-up*）
    - 凝聚（agglomerative）
  - 自顶向下（*partitional, top-down*）
    - 分裂（divisive）
- 平面算法（“flat” algorithms）
  - 划分
    - 随机选择文档集的初始划分
    - 通过迭代精练聚类
  - 密度

层次聚类法

等级聚类又称为分层聚类、等级聚类、系统聚类、谱系聚类，是一种可以利用谱系结构或树状结构图来描绘聚类过程的方法，也是进行聚类分析时应用最多的方法。特别适用于对小样本场合（样本量在100以内比较合适）。

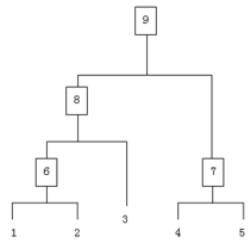
层次聚类可以分为分解法和凝聚法：

- 分解法（Top-down）：在聚类开始时，将所有的文献都看成是一类，然后再根据距离或相似性，不断进行分解，直到每篇文献都自成一类为止。
- 凝聚法（Bottom-up）：聚类开始将每篇文献看成一类，然后再根据距离或者相似性，不断进行合并，直到将所有文献都归结为一类为止

层次聚类-凝聚法

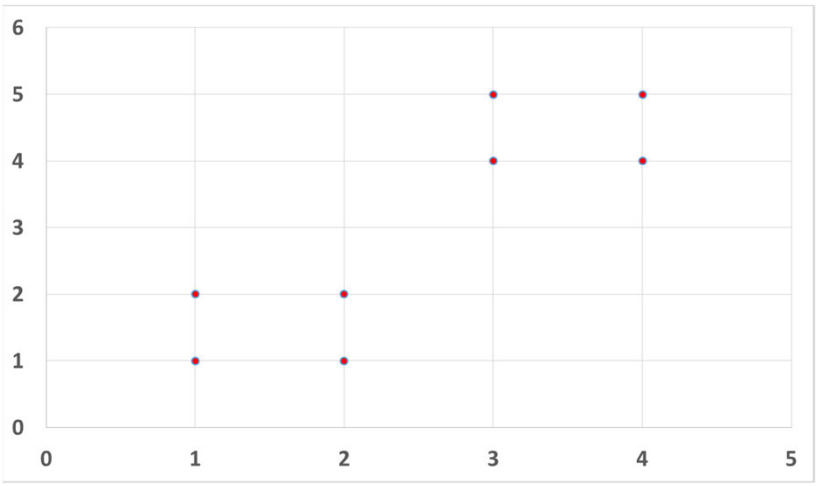
以凝聚法为例，层次聚类的主要步骤有：

- 将每篇文献视为一类，选择度量距离的方法，计算点（文献）与点之间的距离，并将最近的两篇文献聚为一类；
- 选择计算类与类之间距离的方法，计算类与类之间的距离，并将最近的两类进行合并；
- 如果合并后的类数大于1，继续进行类与类之间的合并，直到所有文献合并为一类；
- 绘制等级聚类的谱系图，并根据研究目的、相关的专业理论等选择确定最后的分类结果。



示例

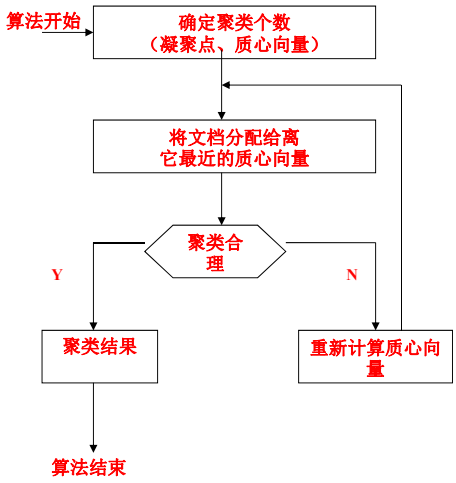
序号	属性1	属性2
1	1	1
2	1	2
3	2	1
4	2	2
5	3	4
6	3	5
7	4	4
8	4	5



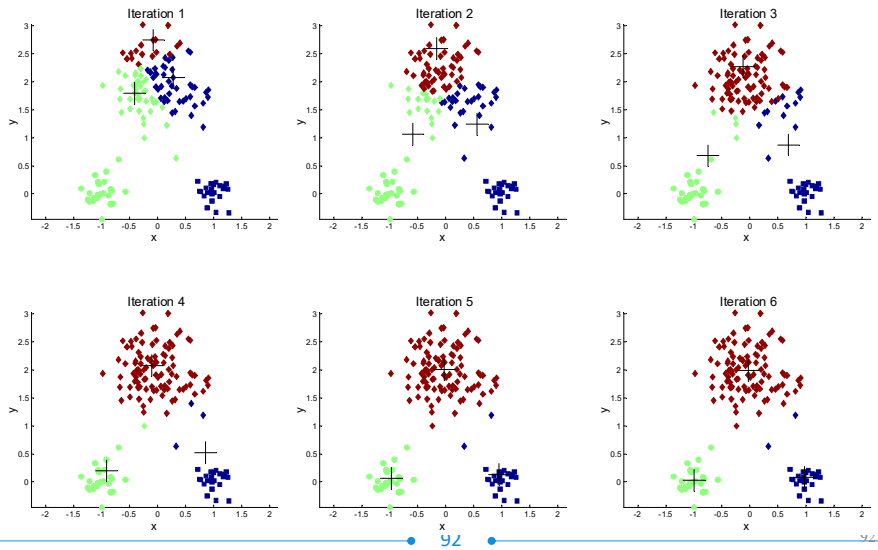
快速聚类法

动态动聚类法又称为： k -均值（k-means）聚类、快速聚类，文本聚类的默认或基准算法。

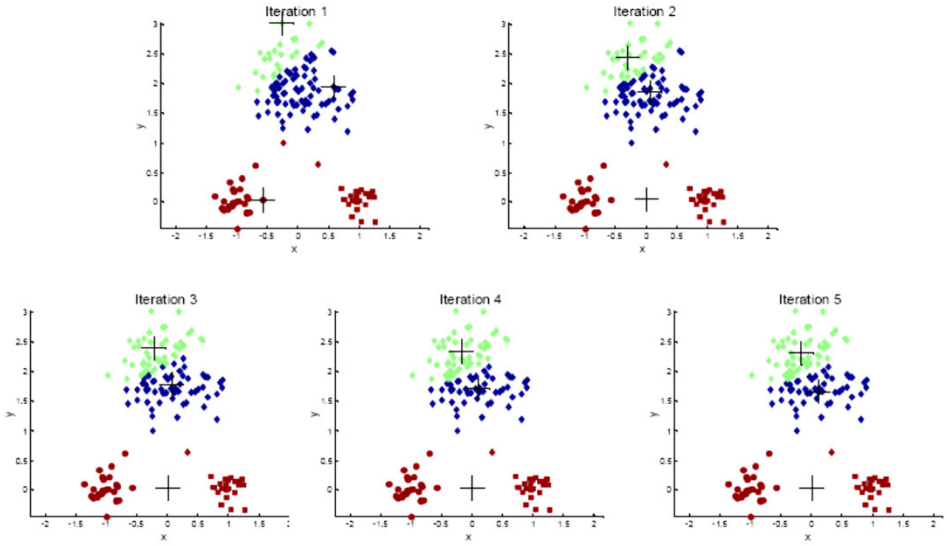
动态聚类的**基本思想**是：先对所要分类的事物作一个初始的分类，然后按照某种最优的原则修改不合理的初始分类，直至分类被认为比较合理时为止，形成最终的聚类结果。



K-means示例



K-means示例2



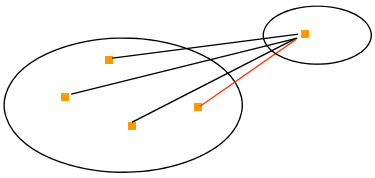
聚类的相似度问题

- 文档相似度: $sim(d_i, d_j)$.
  - 向量夹角余弦
- 聚类间相似度：
  - 聚类中心点（Centroid）：用中心向量表示聚类，聚类间相似度采用向量夹角余弦。
  - 单链（Single Link）：两个聚类间最相似文档的相似度来表示聚类相似度。
  - 全链（Complete Link）：两个聚类间最不相似文档的相似度来表示聚类相似度。
  - 组平均（Group Average）：两个聚类间文档的平均相似度来表示聚类相似度。
  - 其它方法

单链

- 聚类c_i和c_j相似度

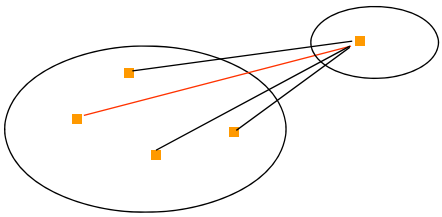
$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$



全链

- 聚类c_i和c_j相似度

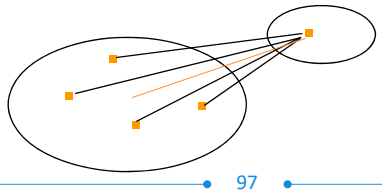
$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$



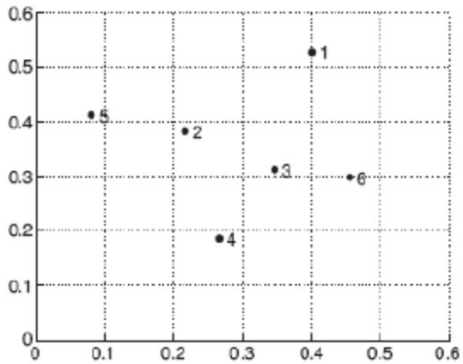
组平均

- 聚类c_i和c_j相似度

$$sim(c_i, c_j) = \frac{1}{MN} \sum_{x \in c_i} \sum_{y \in c_j} sim(x, y)$$

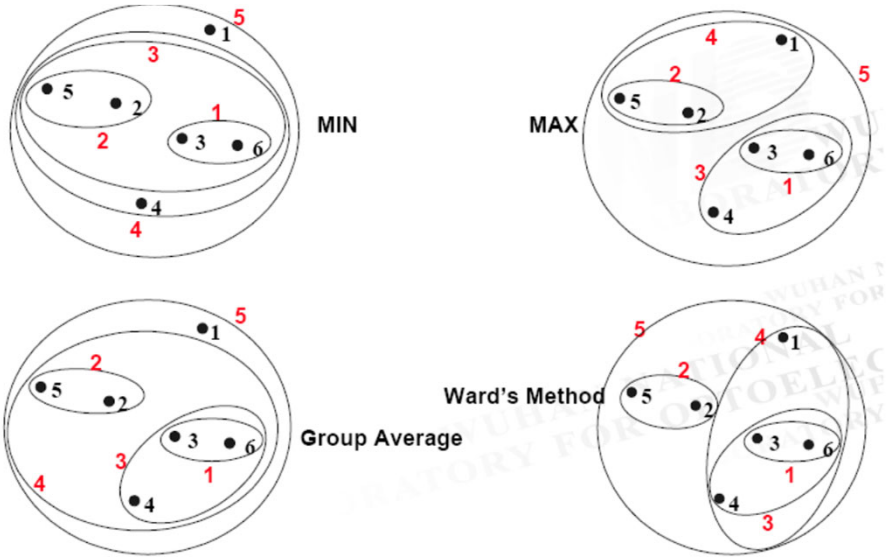


示例



Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

示例（结果）



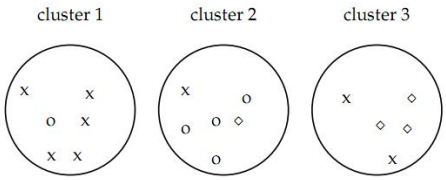
聚类效果的评价

- 文档关联评测及其指标
  - 聚类错误率
  - 聚类全面率
  - 聚类准确率
- 恢复人工分类能力的评测及其指标
  - 重合率
  - 纯度
  - 熵值

聚类的测评-纯度

这里，n_r^i 是属于预定义类i且被分配到第r个聚类的文档个数，n_r为第r个聚类类别中的文档个数。

$$P(S_r) = \frac{1}{n_r}\max(n_r^i) \qquad Purity = \sum_{r=1}^{k}\frac{n_r}{n}P(S_r)$$



purity = (5+ 4 + 3) / 17 = 0.71

小结

南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

07

PART Seven

文本索引和搜索技术

Index & Retrieval Tech.

信息检索

索引技术的内容

概述 → 倒排索引 → 后缀数组 → 签名文件索引 → 检索技术

3

信息检索

为什么要索引？

4

信息检索

索引与标引？

- Index
- 标引
- 索引：IR中，对照或引导标引信息的排列表
  - 主题索引
    - 关键词索引、单元词索引、标题词索引、叙词索引……
  - 类号索引
  - 引文索引（SCI、EI、CSSCI……）

5

信息检索

索引表现形式

- 将标引的结果（主题、类号）按照一定规律排列的处理技术
- 任何检索工具都应该由二次文献部分和检索标识组成

检索工具	二次文献形式	主题表现形式
卡片式	文献卡片	导卡
书本式	文献条目	主题索引
机读顺排文档	条目字段	主题词字段
机读倒排文档	条目区	主题词区

6

信息检索

索引的（存储）价值

- 所谓建立索引，是指将待搜索的信息进行一定的**分析**，并将分析的结果按照**一定的组织方式存储**起来，通常是存储在文件中.
- 存储了分析结果的文件的集合就是所谓的**索引**.
- 准确定义：**索引**（Index）是一种**数据结构**，其将关键词与包含该关键词的文档（或关键词在文档中的位置）建立了一种映射关系，以**加快检索的速度**。

7

信息检索

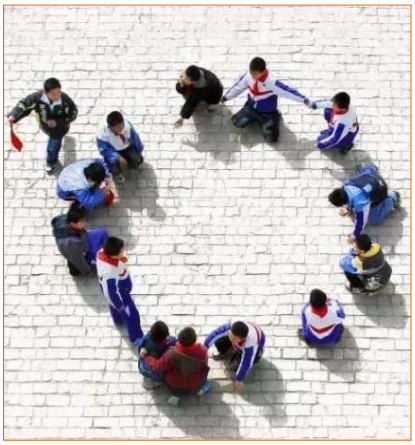
索引的存储方法

8

信息检索

顺排文档索引

- 思想
  - ✓ 将文档中的每一条记录依次去匹配用户的检索提问集合，文档处理完毕后，将各提问的命中结果归并分发给有关用户。
- 定义
  - ✓ 用文档中记录一条一条去匹配提问的，是顺序对文档记录检索的方法
- 关键技术
  - ✓ 采用列表处理方法将提问逻辑式（检索式）变换成等价的提问展开式，按提问展开表的内容对顺排文档的每篇文献进行检索
- 查询方法
  - ✓ ***表展开法***、*逻辑树法等*



9

信息检索

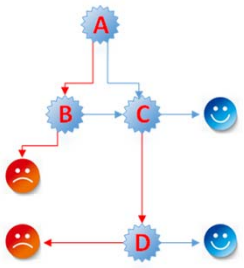
查询的表展开法

1968年菊池敏典提出，其主要思想是采用列表处理方法将逻辑提问式即检索式变换为等价的**提问展开表**，然后按提问展开表的内容对文档中的每一条记录进行检索。

将经典**布尔逻辑检索**的逻辑提问表达式转换为逻辑检索表，每个检索词的检索组配关系要求能够用表进行精确映射，检索的记录是否最终命中检索需求要能准确反映出来。

（A+B）*（C+D）的展开检索基础表

地址	检索词	条件满足指向	条件非满足指向
1	A	3	2
2	B	3	落选
3	C	命中	4
4	D	命中	落选



10

信息检索

特点

- **优点**
  - 凡是可不查阅的文献属性一定不查
  - 凡是可不再查阅的检索词一定不再查
  - 节省机器的查比时间，早期得到广泛应用
- **缺点**
  - 其效率在某种程度上依赖原提问式的书写.
  - 在实际的定额批式检索中，很多提问中往往含有很多相同的检索词，由于每个提问都要与主文档进行匹配处理，因此一批提问中这些相同的检索词和同一文献要重复匹配多次

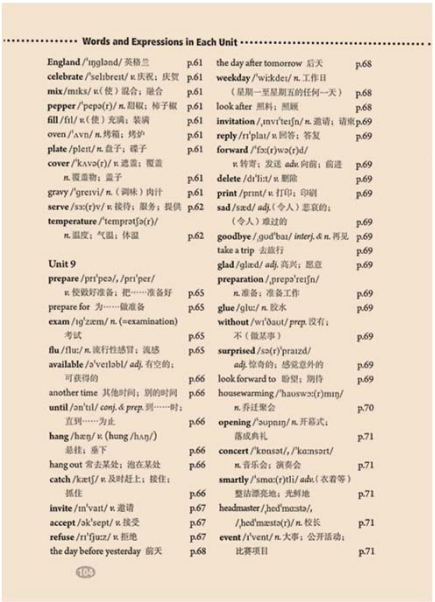
信息检索

倒排文档

(Inverted File)

也称**倒排索引**，索引对象是文档或文档集合中的**单词**等，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**，是对文档或文档集合的一种**最常用的索引机制**。

如:有些书往往在最后提供的索引(单词—页码列表表)就可以看成是一种倒排索引，即通过一些关键词，在全书中检索出与之相关的部分。


Words and Expressions in Each Unit

England /'ɪŋgland/ 英格兰 p.61
celebrate /'selɪbreɪt/ v. 庆祝；庆贺 p.61
mix /mɪks/ v.（使）混合；融合 p.61
pepper /'pepə(r)/ n. 胡椒；柿子椒 p.61
fill /fɪl/ v.（使）充满；装满 p.61
oven /'ʌvn/ n. 烤箱；烤炉 p.61
plate /pleɪt/ n. 盘子；碟子 p.61
cover /'kʌvə(r)/ v. 遮盖；覆盖 n. 覆盖物；盖子 p.61
gravy /'greɪvi/ n.（调味）肉汁 p.61
serve /sɜːv(r)v/ v. 接待；服务；提供 p.62
temperature /'temprətʃə(r)/ n. 温度；气温；体温 p.62

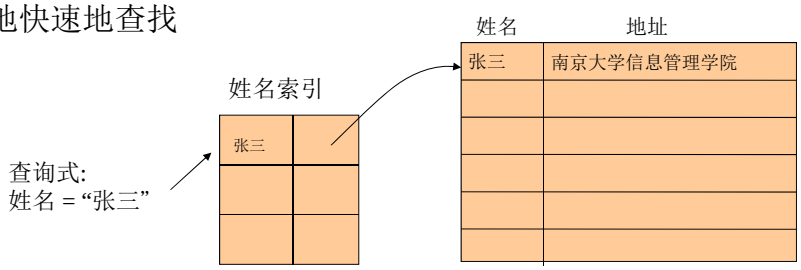
Unit 9

prepare /prɪ'peə/, /prɪ'per/ v. 使做好准备；把……准备好 p.65
prepare for 为……做准备 p.65
exam /ɪg'zæm/ n. (=examination) 考试 p.65
flu /fluː/ n. 流行性感冒；流感 p.65
available /ə'veɪləbl/ adj. 有空的；可获得的 p.66
another time 其他时间；别的时间 p.66
until /ən'tɪl/ conj. & prep. 到……时；直到……为止 p.66
hang /hæŋ/ v.（hung /hʌŋ/）悬挂；垂下 p.66
hang out 常去某处；泡在某处 p.66
catch /kætʃ/ v. 及时赶上；接住；抓住 p.66
invite /ɪn'vaɪt/ v. 邀请 p.67
accept /ək'sept/ v. 接受 p.67
refuse /rɪ'fjuːz/ v. 拒绝 p.67
the day before yesterday 前天 p.68
the day after tomorrow 后天 p.68
weekday /'wiːkdeɪ/ n. 工作日（星期一至星期五的任何一天） p.68
look after 照料；照顾 p.68
invitation /ˌɪnvɪ'teɪʃn/ n. 邀请；请柬 p.69
reply /rɪ'plaɪ/ n. 回答；答复 p.69
forward /'fɔː(r)wə(r)d/ v. 转寄；发送 adv. 向前；前进 p.69
delete /dɪ'liːt/ v. 删除 p.69
print /prɪnt/ v. 打印；印刷 p.69
sad /sæd/ adj.（令人）悲哀的；（令人）难过的 p.69
goodbye /ˌgʊd'baɪ/ interj. & n. 再见 p.69
take a trip 去旅行 p.69
glad /glæd/ adj. 高兴；愿意 p.69
preparation /ˌprepə'reɪʃn/ n. 准备；准备工作 p.69
glue /gluː/ n. 胶水 p.69
without /wɪ'ðaʊt/ prep. 没有；不（做某事） p.69
surprised /sə(r)'praɪzd/ adj. 惊奇的；感觉意外的 p.69
look forward to 盼望；期待 p.69
housewarming /'haʊswɔː(r)mɪŋ/ n. 乔迁聚会 p.70
opening /'əʊpnɪŋ/ n. 开幕式；落成典礼 p.71
concert /'kɒnsət/, /'kɒnsɜːt/ n. 音乐会；演奏会 p.71
smartly /'smɑː(r)tli/ adv.（衣着等）整洁漂亮地；光鲜地 p.71
headmaster /ˌhed'mɑːstə/, /ˌhed'mæstə(r)/ n. 校长 p.71
event /ɪ'vent/ n. 大事；公开活动；比赛项目 p.71

信息检索

在关系数据库上建索引

- 这种想法也被应用于数据库技术中，即对数据库中需要经常进行检索的域建立索引结构，进行快速的查询。
- 索引结构: *hashing, B+-tree*
- 可以索引全部记录，在全部记录上进行搜索
- 精确地快速地查找



13

信息检索

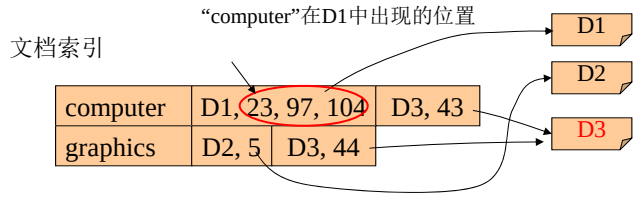
倒排文档的组成

- 倒排文档一般由两部分组成：词汇表（vocabulary）和记录表（posting list）
- 词汇表是文本或文本集合中所包含的所有不同单词（索引项）的集合。
- 对于词汇表中的每一个单词，其在文本中出现的位置或者其出现的文本编号构成一个列表，所有这些列表的集合就称为记录表。

14

信息检索

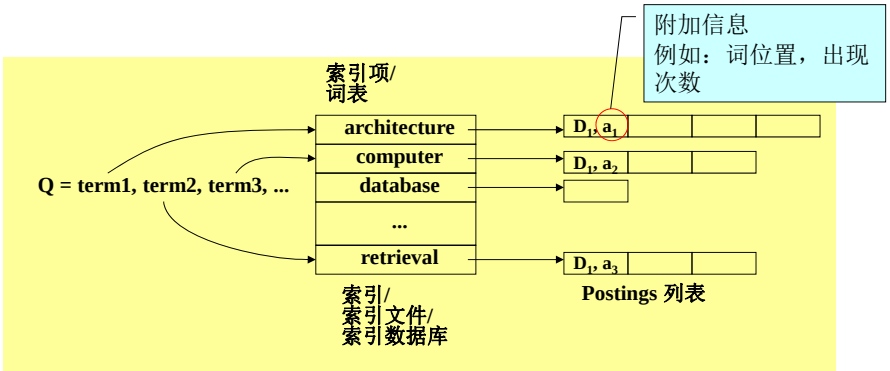
对文档进行索引



- 可以进行部分匹配: '%comput% '
- 可以进行短语搜索:查找包含“computer graphics”的文档

信息检索

一般的倒排索引



- 索引文件可以用任何文件结构来实现
- 索引文件中的词项是文档集合中的词表

信息检索

例子

文本

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
这	是	一本	关于	信息	检索	的	教材	。	介绍	了	检索	的	基本	技术	。	…

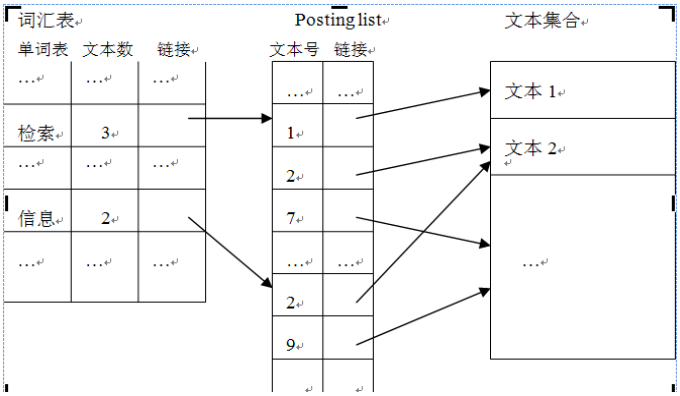
倒排文件

词汇表	记录表
技术	15, …
教材	8, …
检索	6, 12, …
信息	5, …
…	…

信息检索

以文本为记录表

记录表既可以存储文本中单词的编号位置，也可以指向单词首字母的字符位置，还可以是其所在的文本编号，下图是一个以文本为记录表的情况



信息检索

倒排文档的使用

- 词汇表检索
  - 将出现在查询中的单词分离出来，并在词汇表中进行检索；
- 记录表检索
  - 检索出所有找到的单词对应的记录表；
- 记录表操作
  - 对检索出的记录表进行处理，实现短语查询、相邻查询或布尔查询等。

19

信息检索

提问式的逆波兰展开

- 逆波兰
  - 1929年波兰的逻辑学家卢卡西维兹提出将运算符放在运算项后面的逻辑表达式，又称“逆波兰表达式”（Reverse Polish Notation，RPN，或逆波兰记法），也叫后缀表达式（将运算符写在操作数之后）
  - 用于数据结构等多个场合
- 福岛算法
  - 例如：逻辑提问式 A*(B+C)+D （中缀表达式）
    逆波兰表达式： ABC+*D+ （后缀表达式）
    波兰表达式：+*A+BCD （前缀表达式）
- 福岛（Fukujima）算法首先要进行提问式的转换。

Jan Łukasiewicz（1878-1956）

20

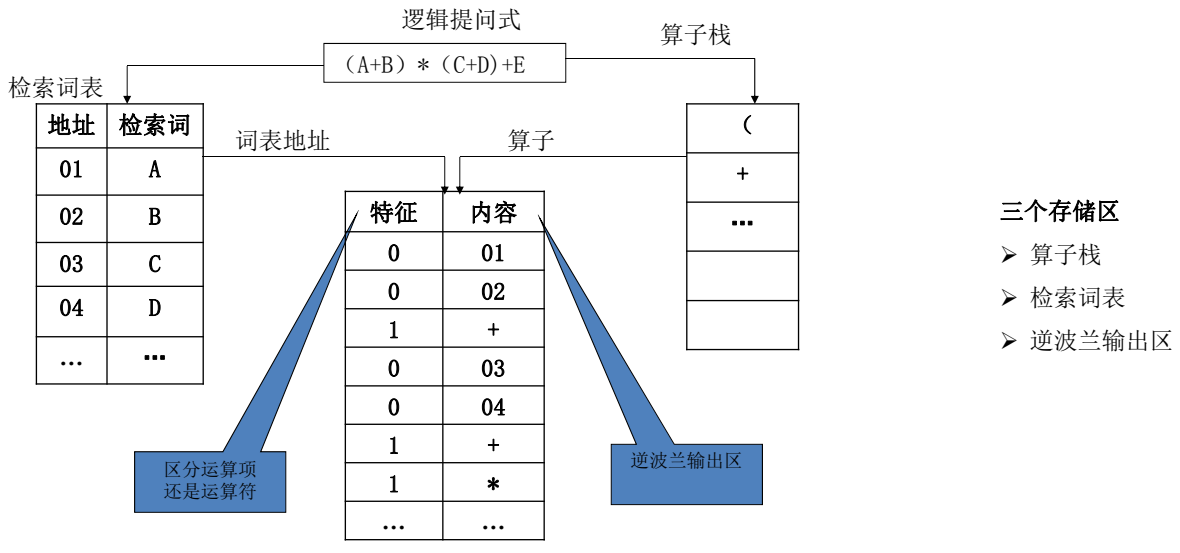
信息检索

逆波兰表达式

- 中缀表达式生成的逆波兰表达式是唯一的
- 注意：
  - a+(b-c)*d 和 (b-c)*d+a 和 a+d*(b-c) 的值是完全一样的。但是，中缀形式不同，产生的逆波兰式必然是不同的。
  - a+(b-c)*d： abc-d*+
  - (b-c)*d+a： bc-d*a+
  - a+d*(b-c)： adbc-*+

信息检索

逆波兰法处理示意图



三个存储区

- 算子栈
- 检索词表
- 逆波兰输出区

信息检索

检索过程示例

- 以提问式（A+B）*（C+D）+E.为例，说明其工作区占用情况，以及由逆波兰表示式转换为一组检索指令的过程。

工作区W

1	A结果文献号码集合 1)	C结果文献号码集合 4)	(A+B)*(C+D)结果文献号码集合 7)	…
2	B结果文献号码集合 2)	D结果文献号码集合 5)	E结果文献号码集合 8)	
3	A和B“或”结果文献号码集合 3)		(A+B)*(C+D)和E“或”结果文献号码集合 9)	
4	C和D“或”结果文献号码集合 6)			
…	…			
n=7	(A+B)*(C+D)+E.结果文献号码集合 10)	转储		

信息检索

倒排索引的特点

- 快速索引（长query需要更多时间）；
- 灵活性: 不同类型的信息都可以存储在记录表中；
- 如果存储了足够多的信息，则可以支持复杂的检索操作；
- 存储开销较大；
- 更新、插入和删除都需要很高的维护开销，倒排索引相对静态的环境（很少插入和更新）中使用比较好。

信息检索

后缀数组

- 在倒排文档中，文本被看作是由单词组成的序列➔限制了倒排文件的应用
  - 情况1：词组查询在使用倒排文件查询时就比较困难，因为不但需要记录每个单词在文档中的位置，还要在合并时判断其是否相邻并构成词组；
  - 情况2：在某些应用中，如基因数据库，不存在词的概念。
- 解决方案：使用**后缀数组（suffix array）**
- 在后缀数组中，可以将文本看作是一个长的字符串，文本中的每个位置都被认为是文本的一个后缀，即一个从当前文本位置到文本末尾的字符串。
- 索引的位置可以是每个字符的位置、单词的位置或者汉字的位置等。
- **后缀数组**就是对文本中的所有后缀按照词典序存放每个后缀对应的起始位置的一个列表

信息检索

后缀数组的构造

原始文本，按字的顺序位置索引

0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	…
这	是	一	本	关	于	信	息	检	索	的	教	材	。	介	绍	了	检	索	的	基	本	技	术	。	…

首先截取每个后缀的前n个字节，作为类似倒排文件中的“词汇表”(此处n=4，2汉字)

0	2	12	16	22	34	44	…
这是…	是一…	信息…	检索…	教材…	检索…	技术…	…

文本中的部分后缀，按位置索引

44	16	34	22	2	12	0	…
技术…	检索…	检索…	教材…	是一…	信息…	这是…	…

相同的部分后缀，按字典序索引

信息检索

后缀数组的使用

- 在使用后缀数组进行检索的时候，将每个查询同样截取前*n*个字节，并于索引中进行查找
  - 如果没有找到，则表明不包含所需查询
  - 如果查找成功，则需要在相应的文本位置上，进行进一步的字符串比较，以确定文本中是否包含查询
- 实例
  - 查找“信息检索”，首先截取4个字节“信息”，并于后缀数组中查找到了位置12，接着在原文中的12号位置，找到“信息检索”字符串，则查找成功；
  - 查找“信息过滤”，则原文中的12号位置不能找到“信息过滤”字符串，则查找失败；
  - 更一般的例子，如果输入的查询为“数据库”，在索引结构中不能找到“数据”，则查找直接失败返回；

44	16	34	22	2	12	0	…
技术…	检索…	检索…	教材…	是一…	信息…	这是…	…

信息检索

后缀数组的分析

- 对于需要大数据量的检索问题，后缀数组并不适用
- 因为构造出的后缀数组需要占用大量的空间，通常是原文本的1.7倍
- 和倒排文档相比，后缀数组里面储存了较多的重复信息
- 同时，由于每个后缀位置的编号不能存储相对位置的变化，所以很难被压缩，需要较多的存储空间；
- 后缀数组的另外一个缺点是不容易对检索结果进行排序，因为计算查询单词的词频比较耗时；
- **实际上,后缀数组的大部分功能可以通过倒排文档来实现!**
  - 例如可以倒排索引文本中的二字串或者三字串，从而提高召回率

签名文件

信息检索

- 签名文件（signature file）是基于散列（Hash）技术的面向单词的索引结构
- 索引占用的空间大约为原始文档集的30~40%
- 在检索时需要顺序比较，适用于小规模的文本
- 在大多数应用中，其性能不如倒排文件

29

关于Hash

信息检索

- **散列/哈希**
  - ✓ 任意长度的数据映射到有限长度的域上
- **要求**
  - ✓ 抗冲突能力
  - ✓ 抗篡改能力
- **应用**
  - ✓ 加密
    - MD5，把一些不同长度的信息转化成杂乱的128/256bits的编码
  - ✓ 信息摘要/查询


```
MD5("version1.2") = "d9671a07d41dc45815aa9e8fbf88148f"

MD5("version2.1") = "51535660790f4243b9cb5dba1a0bc207"
```

30

信息检索

词的Signatures

- 一个单词的“签名”是一个位向量，由F位组成，其中有m位置1;
- 如下图给出了一段文本，以及文本中部分单词的“签名”示例，其中F=12，m=4;

这 是 一本 关于 信息 检索 的 教材 。 介绍 了 检索 的 基本 技术 。…

文本

单词	词“签名”
技术	001 000 110 010
教材	000 010 101 001
检索	000 000 011 110
信息	101 000 100 001
…	…
…	

31

信息检索

词Signature的生成算法

```
输入：词W，参数F和m
输出：词W的F位“签名”S
算法：
(1) 将W转换为ASCII值，然后转换为32位整数I
      例如：free = 66726565 (hex)
(2) 初始化：
      a) S的F位全置0;
      b) srandom(i);                          //初始化随机种子
      c) j = 0;
(3) while (j < m){
      p = random();                           //生成32位随机数
            p = p mod F;                      //映射到0和F-1之间
      if (S[p] == 0) {                        //确保m位置1
            S[p] = 1;
              j++;
            }
      }
(4) 结束，返回S
```

32

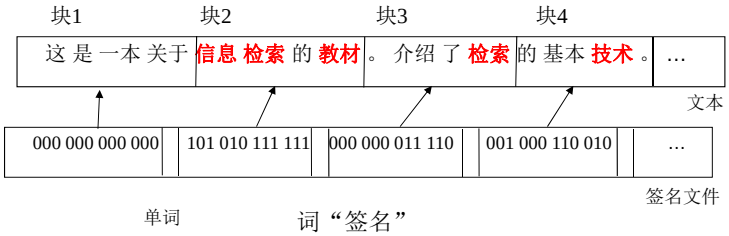
信息检索

签名文件

- 如果仅对每一个单词生成相应的“签名”，则在查询的时候，需要逐个比较，非常浪费时间。
- 因此，通常把文本分成若干块，每个块有多个单词，块的“签名”是通过块中单词的“签名”的按位“**或**”操作得到的
- 因此，所谓的签名文件，就是文本中块“签名”的序列；
- 下图给出了一段文本以及其对应的签名文件

信息检索

一段文本对应的块签名



单词	词“签名”
技术	001 000 110 010
教材	000 010 101 001
检索	000 000 011 110
信息	101 000 100 001
…	…
…	

000 010 101 001
000 000 011 110
101 000 100 001

块2的签名 101 010 111 111

信息检索

签名文件的使用和维护

- 使用签名文件检索单个单词的过程是，首先对这个单词使用相同的算法生成其“签名”S，并与所有文本块的签名S_i依次比较，即计算S & S_i是否与S相等；
- 如果相等，说明S中含有的位，S_i中也有，即S_i文本块可能含有要查找的单词→形成候选文本块
- 对所有候选文本块执行字符串匹配，确定是否真正含有要查找的单词；

块2的签名	101 010 111 111
信息的签名	101 000 100 001
And运算	101 000 100 001

35

信息检索

签名文件的使用和维护

- 使用签名文档时，可能会出现误检的情况，即虽然某个单词不存在于文本块中，但是其“签名”仍然能够与该块的“签名”相应位匹配。
- 造成误检的原因有两方面：
  - 主要原因：相匹配的位来自块中不同单词的“签名”。
    例如，假设查询单词“计算机”的“签名”是101 010 010 000，其与块2的“签名”相匹配，但是显然块2中，不含有“计算机”，这时就造成误检，原因就是“计算机”的“签名”101 000 011 000中，1和3位与“信息”的“签名”相匹配，8和9位与“检索”的签名相匹配；
  - 次要原因：**散列冲突**，即两个不同的单词具有相同的“签名”，然而如果F足够大，这种冲突的可能性很低。
- 与倒排文件的维护相比，签名文件的维护非常容易
  - 添加文本时，只需要将文本分块，然后生成块的“签名”，追加到签名文件末尾即可；
  - 删除文本时，只需要在签名文件中删除相应文本块的“签名”即可

块2的签名	101 010 111 111

36

信息检索

分析

- 为了加快检索速度，需要减少顺序匹配的次数，需要将块做得足够大，一般将一个文本看作是一块；
- 而为了减少误检的发生率，“签名”的位数又要足够多，经验表明，取文本平均长度的30%~40%为宜;
- 签名文件由于组织简单，因此较容易生成，维护费用较小，它是在倒排文件和全文扫描之间做了空间和时间的平衡，适合于**小文本集合和查询频率较低**的系统
- 签名文件的主要缺点：
  - 索引占用的空间和检索的时间复杂性与原始文档集成线性关系。
  - 误检的发生数也基本和文本集合中的文本个数成正比。
  - 即使对于比较短的查询，也需要大量的磁盘访问操作。
  - 签名文件很难对频率和权重信息进行编码，这就很难支持排序操作。
- **签名文件索引技术只适用于小规模的文本集合。**

信息检索

文本搜索技术

- 前面的文本检索技术需要事先建立索引，然后才能快速查找。
- 在某些应用中,这种建立索引的方法并不适用
  - 情况1:在签名文件的候选块确认过程中，就需要在块中查找某一查询是否真正存在；
  - 情况2:在文本过滤技术中，一般文本仅需查询一次，这就没必要建立索引；
  - 情况3:在搜索引擎结果后处理中，需要对搜索结果中包含的查询关键词进行加亮显示，也需要用到文本搜索技术;
- 快速的文本搜索非常必要!

文本搜索技术

信息检索

- 精确的字符串匹配问题可以如下描述:
  - 结定一个长度为m的模式$P(p_1p_2\cdots p_m)$，以及一个长度为n的长文本$T(t_1t_2\cdots t_n)$，其中$n>>m$。在文本T中找出模式P出现的位置。
- 三种常用的精确匹配算法
  - *BF*算法
  - *KMP*算法
  - *BM*算法

39

BF算法

信息检索

- BF算法是**B**rute-**F**orce（蛮力）算法的简称
- 是一种简单、直接、容易实现的字符串匹配算法
- 基本思想：
  - 将模式P和文本T中的m个字符的子串$t_kt_{k+1}\cdots\cdots t_{k+m-1}$进行匹配，$1<k<n$。
  - 若模式和子串匹配，则返回匹配的位置，
  - 若模式和子串不匹配，则从t_{k+1}位置开始新的考察

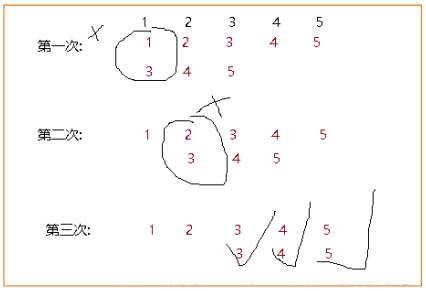
40

信息检索

BF算法

输入： *文本T和模式P*
输出： *文本T中如果包含模式P返回匹配位置，否则返回匹配不成功*
算法：

```
i = 1; j = 1;
while (i ≤ m and j ≤ n) {
    if (pi == tj) {
        i = i + 1; j = j + 1;
    } else {
        j = j - i + 2; i = 1;
    }
}
if (i > m)
    return 匹配位置j;
else
    return "匹配不成功";
```



41

信息检索

BF算法分析

- 在BF算法中，可以更改循环条件为*j≤n-m+1*，使得循环可以更早地终止（只返回首次位置）；
- 因为存在O(n)个文本位置，在最坏的情况下，验证每个位置需要花费的时间为O(m)，所以BF算法的最长时间为O(mn)
- BF算法的平均时间为O(n)，因为对于一个随机文本，一般进行O(l)次比较之后，就可以发现错误的匹配。
- BF算法无需进行任何形式的预处理，**实现简单，**被大多数程序设计语言所采用
  - 如C语言中的字符串查找函数strstr()使用的就是BF算法
- BF算法的**时间复杂性**与其他算法比较还是比较高的，在对效率要求较高的系统中，应该避免大量使用

42

KMP算法

信息检索

- D.E.Knuth、J.H.Morris和V.R.Pratt同时发现的改进的模式匹配算法/克努特—莫里斯—普拉特操作
- 该算法是第一个可以在O(n+m)的时间内完成串模式匹配的算法，**但平均来说**，它的效率并不比BF算法快很多。
- 基本思想
  - 每当匹配过程中出现字符串比较不等时，不像BF算法那样仅将模式向右“滑动”一个位置，而是利用已经得到的“部分匹配”结果将模式向右“滑动”尽可能远的一段距离后，继续进行比较。
  - 可以避免对那些能够推断出不匹配的位置进行徒劳的操作

位置：	1	2	3	4	5	6	7	8	
文本：	a	b	d	a	d	e	f	g	
模式：	a	b	d	f					
		a	b	d	f				← BF 算法仅移动一个字符
				a	b	d	f		← KMP 算法移动三个字符

Donald Ervin Knuth（1938）

KMP算法

信息检索

- KMP算法原则
  - 从**匹配成功**的**子模式**中找出“能够相互匹配的**最长的前缀**和**后缀**”
  - 在使用KMP算法进行模式匹配时，需要根据模式事先构造一个**shift跳转表**，用来决定在某个位置匹配失败时应该移动多少个字符
- Shift表的构造方法
  - 如果当前不匹配的位置为*j*，重复字串的长度为*k*，则跳过的字符个数为*j-k-1*

KMP算法-示例(BF)

信息检索

BBC ABCDAB ABCDABCDABDE
ABCDABD

BBC ABCDAB ABCDABCDABDE
ABCDABD

BBC ABCDAB ABCDABCDABDE
ABCDABD

BBC ABCDAB ABCDABCDABDE
ABCDABD

BBC ABCDAB ABCDABCDABDE
ABCDABD

45

KMP算法-示例(优化)

信息检索

BBC ABCDAB ABCDABCDABDE
ABCDABD

BBC ABCDAB ABCDABCDABDE
ABCDABD

如果当前不匹配的位置为*j*，**重复字串**的长度为*k*，则跳过的字符个数为*j-k-1*

7-2-1=4

46

信息检索

KMP算法的shift表

- 模式P=“abcabcacab”的shift表

匹配失败位置	1	2	3	4	5	6	7	8	9	10
模式字符	a	b	c	a	b	c	a	c	a	b
重复子串长度	0	0	0	0	1	2	3	4	0	1
跳过字符数	1	1	2	3	3	3	3	3	8	8

如果当前不匹配的位置为j,重复字串的长度为k,则跳过的字符个数为j-k-1

- KMP算法的shift表可以在O(m)的时间复杂度下，通过对关键词的分析而获得,m是模式的长度
- KMP算法花费O(m+n)的时间来解决模式匹配问题

信息检索

BM算法

- Boyer和Moore提出
- 是另一个与KMP算法截然不同的却同样拥有线性时间复杂度的算法
- BM算法在实际的模式匹配中跳过了很多无用的字符，不必对无用的字符进行匹配（而在KMP算法中，文本串中的每个字符都是需要进行比较操作的）
- 这种跳跃式的比较方式，使BM算法的效率极高，特别是在大字符集上进行字符串的模式匹配时优势更加明显。
- 在理论和实践上，**BM算法都比KMP算法更有效**
- 基本思想
  - 假设模式的长度为m。先令模式和文本左对齐，然后对模式中最右一个字符p_m与其在文本中相对应的字符t_m进行比较。如果比较的结果是不匹配，那么我们接着考察t_m在模式中出现的最右位置，然后根据这个位置移动模式，使其和t_m对齐

BM算法

信息检索

- 将模式和文本左对齐后，模式中最后一个字符与其在文本中对应字符比较的结果有下面两种可能
  - 第一种情况:t_m根本没有在模式中的任何一个位置出现,那么就可以放心大胆地将模式向后移动m个字符,然后将模式中最后一个字符与它现在所对应的文本中的字符(即t_{2m})进行比较.
  - 第二种情况:如果t_m是模式中的第n个字符,那么就可以将模式向后移动m-n个字符.

- 文本字符G没有在模式中,可以将模式向后移动m个字符

- 文本字符C现现在模式中的第2个字符，可以将模式向后移动4-2个字符

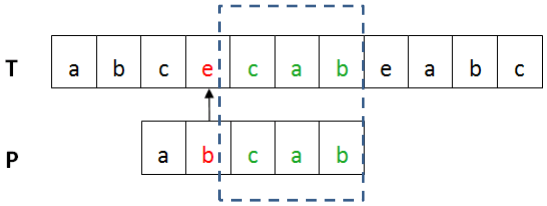
BM算法-好后缀规则

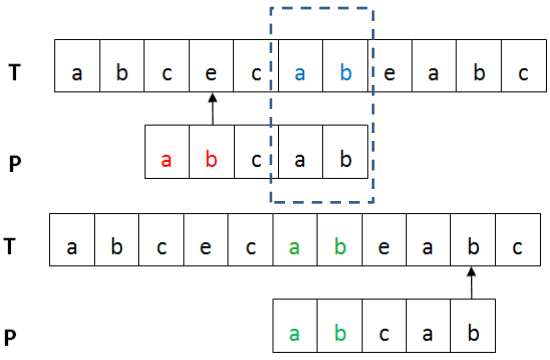
信息检索

若发现某个字符不匹配的同时，已有部分字符匹配（好后缀）成功，则按如下两种情况讨论：

第一种情况，如果在P中位置t处已匹配部分P'在P中的某位置t'也出现，且位置t'的前一个字符与位置t的前一个字符不相同，则将P右移使t'对应t方才的所在的位置。

第二种情况，如果在P中任何位置已匹配部分P'都没有再出现，则找到与P'的后缀P''相同的P的最长前缀x，向右移动P，使x对应方才P''后缀所在的位置。

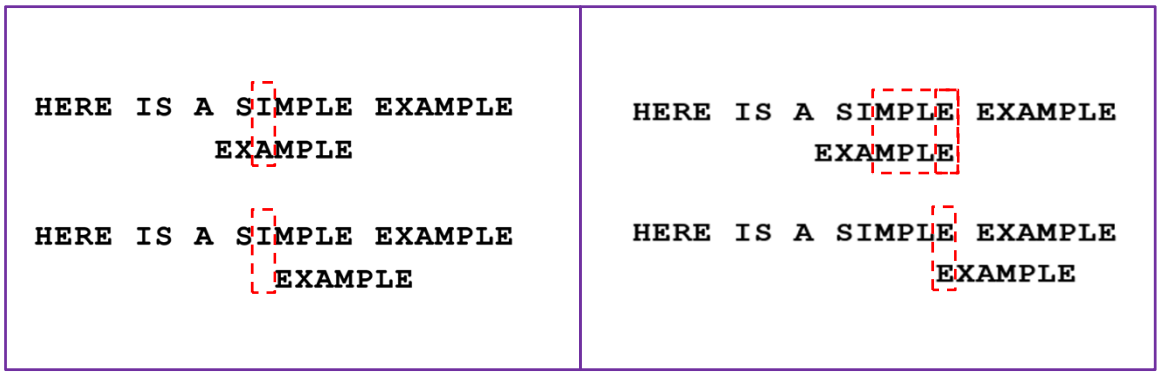




信息检索

BM算法-选择

好后缀规则与坏字符规则中移动步数较大者作为最后移动步数



51

信息检索

模式匹配算法的选择

- 如果模式的长度很小（1到3个字符），可以使用BF算法，因为其实现简单，而且不需要额外构造跳转表
- 如果字母表很大，KMP算法是一个选择，因为模式中含有重复的情况较少
- 除此以外，特别是对于长文本来说，BM算法是最佳选择。

52

信息检索

文本检索技术

53

信息检索

布尔检索

- 布尔逻辑算符及其应用
  - 常用的布尔逻辑算符有三种，分别是逻辑与AND、逻辑或OR、逻辑非NOT，用以表达两个检索词之间的逻辑关系。下面分别简释他们各自的含义与用法。
- （1）逻辑与——“AND”或 *
  - 检索词A与检索词B若用“AND”组配，则提问式可写为“A AND B”或者“A * B”。检索时，数据库中同时含有检索词A和检索词B的文献，才算是命中文献。

54

信息检索

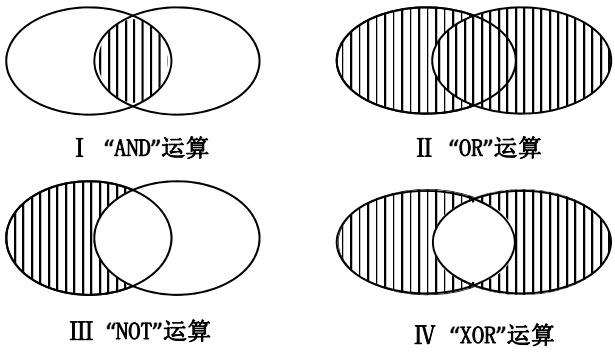
布尔检索

- （2）逻辑或——“OR”或 +
  - 检索词A和检索词B若用“OR”组配，则提问式可写为“A OR B”或者“A + B”。检索时，数据库中的文献凡含有检索词A或者检索词B或者同时含有检索词A和B的，均为命中文献。
- （3）逻辑非——“NOT”或 －
  - 检索词A和检索词B若用“NOT”进行逻辑组配，则可写为“A NOT B”或者“A － B”。对于这个提问式，数据库中凡含有检索词A而不含检索词B的文献，为命中文献。
- （4）逻辑异或 ——”XOR”或⊕（少）
  - 检索词A和检索词B若用异或XOR组配，可写为“A XOR B”或者“A⊕B”。该检索式的检索结果为：含有检索词A的文献命中，含有检索词B的文献命中，但同时含有A和B的文献不命中。

55

信息检索

布尔运算示意图



56

信息检索

注意事项

- （1）布尔检索执行顺序。三种布尔检索运算符之间的优先顺序为**NOT、AND、OR**。有括号时，先执行括号内的逻辑运算。有多层括号时，先执行最内层括号中的运算。
- （2）不同检索工具的布尔逻辑检索有不同的表现形式，在使用的时候，要注意先了解相关的使用规则。

CNKI 例1

要求检索钱伟长在清华大学或上海大学时发表的文章。

检索式：AU =钱伟长 and (AF =清华大学 or AF =上海大学)

CNKI 例2

要求检索钱伟长在清华大学期间发表的题名或摘要中包含“物理”的文章。

检索式：AU =钱伟长 and AF=清华大学 and (TI =物理 or AB = 物理)

57

信息检索

截词检索

- 所谓截词（truncation），是指检索者将检索词在自己认为合适的地方截断；
- 截词检索，则是用截断的检索词的一个局部去数据库中进行检索，凡是能与这个词局部中的所有字符（串）相匹配的文献，即为命中文献。
- 截词符号在不同的信息检索系统中表示不同，但功能是相同的。
- 通常情况下用“*”表示无限截断，用“？”表示有限截断。

58

信息检索

后截词检索

后截断是最常用的截词检索技术。将截词符号置放在一个字符串右方，以表示其右的有限或无限个字符不影响该字符串的检索。从检索性质上讲，后截断是**前方一致**检索。

- *【例】coagula**
- *可检出的词汇有：coagula、coagulable、coagulant、coagulase、coagulate、coagulation、coagulative、coagulator等*
- *【例】mold??*
- *可检出的词汇有：mold、molded、molder等，但却不能检出下述词汇：moldery、molding、moldman、moldwash。*

59

信息检索

前截词检索

与后截断相对，前截断是将截词符号置放在一个字符串左方，以表示其左的有限或无限个字符不影响该字符串的检索。从检索性质上讲，前截断是**后方一致**检索。

- *【例】* meter*
- *可检出的词汇如下：meter、cubic-meter、macro-meter、macrometer、mini-meter、minimeter、square-meter……，但是检不出meterage、metering、meterman等。*

60

信息检索

中截词检索

- 中截断又称为“通用字符法”或“内嵌字截断”或“屏蔽”。这种截断是把截断符置于一个检索词的中间，允许检索词的中间有若干形式的变化。一般地，中截断仅允许有限截断。
- 英语中有些单词的拼写方式有英式、美式之分，有些词则有某个元音位置上出现单复数不同，如：organization↔organisation，man↔men，woman↔women等等。文献数据库中与之相似的例子是大量存在的。若希望不漏检，使用这种词进行检索时就要用中截断的处理方法。
  - 比如，上述词汇在用于检索时可写成：organi? ation，m? n，wom? n。

61

信息检索

限制检索

- 字段检索是限定检索词在数据库记录中出现的字段范围的一种检索方法。
- 在检索系统中，数据库设置、提供的可供检索的字段通常分为主题字段和非主题字段两大类。每个字段都有一个用两个字母表示的字段代码。
- 在CNKI中各字段代码及对应名称说明如下：
  - SU='主题',TI='题名',KY='关键词',AB='摘要',FT='全文',AU='作者',FI='第一责任人',AF='机构',JN='中文刊名'&'英文刊名', RF='引文',YE='年',FU='基金',CLC='中图分类号',SN='ISSN',CN='统一刊号',IB='ISBN' ,CF='被引频次'

62

信息检索

加权检索

- 词加权系统（term weighting system）是最常见的加权检索系统。
- 检索者根据对检索需求的理解选定检索词，同时对提问中的每一个检索词（概念）给定一个数值以表示其重要性程度，即权（weight）。在检索中，先查找这些检索词在数据库记录中是否存在，然后计算存在检索词的记录所包含的检索词的权值总和，通过与预先给定的阈值（threshold）进行比较，权值之和达到或超过阈值的记录视为命中记录，命中记录的输出按照权值总和从大到小排列输出。这种用给检索词加权来表达提问要求的方式，称为词加权提问逻辑。

信息检索

加权检索

【例】以“住房补贴政策”为检索课题，给检索词“住房”、“补贴”和“政策”分别赋予权值4、5、3，阈值T=5。检索时，在关键词文本框中输入“住房/4*补贴/5*政策/3”，单击查询，则依所含关键词的权重检出相应记录，命中文献按权值递减排列如下：

住房，补贴，政策	*权和=12≥5*
住房，补贴	*权和=9≥5*
补贴，政策	*权和=8≥5*
住房，政策	*权和=7≥5*
补贴	*权和=5≥5*

信息检索

词频加权检索

词频加权检索是根据检索词在文档记录中出现的频率来决定该词的权值，而不是由检索者来指定检索词的权值。在这一方面，词频加权就消除了人工干预因素。

TI='转基因$ 2（CNKI）

65

信息检索

小结

66

南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

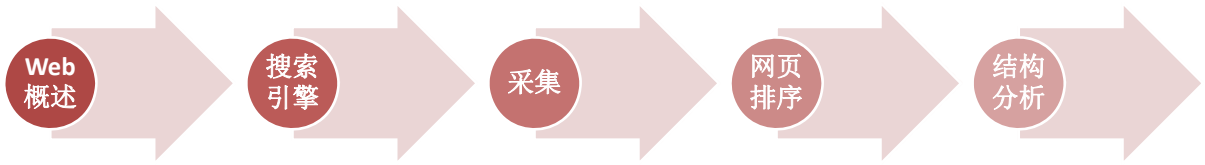
PART Eight

Web信息采集与搜索引擎

Search Engine

信息检索

Web信息处理概要



3

信息检索

Internet与Web

- Internet
  - 因特网，国际互联网
  - 1969秋，ARPAnet，1972年国际联网
  - TCP/IP
- World Wide Web
  - 万维网
  - 1999.12，Tim命名WWW
  - Http

Tim Berners-Lee（1955-）

4

信息检索

暗网与深网

Darknet或Dark Web：需要通过特殊软件、特殊授权、或对电脑做特殊设置才能连上的网络；

Deep Web：互联网上那些不能被标准搜索引擎索引的非表面网络内容。

暗网是深网的一个子集。

5

信息检索

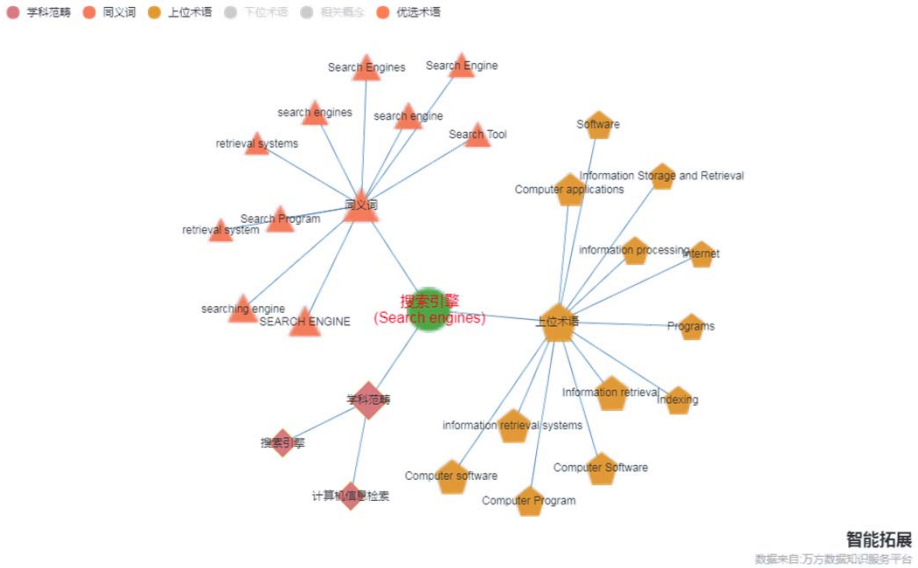
搜索引擎

- Search Engine
- 根据一定的策略、运用特定的计算机程序从互联网上搜集信息，在对信息进行组织和处理后，为用户提供检索服务，将用户检索相关的信息展示给用户的系统。

6

信息检索

SE的相关研究领域



2018.12

信息检索

搜索引擎的发展

- 起源：FTP文件搜索
  - （以Archie为代表，1990-）
- 第一代搜索引擎：分类目录
  - （以雅虎为代表，1995-）
- 第二代搜索引擎：关键词搜索引擎
  - （以Google为代表，1998-）
- 第三代搜索引擎：智能搜索引擎（自然语言）
  - （发展中，如ask.com）

信息检索

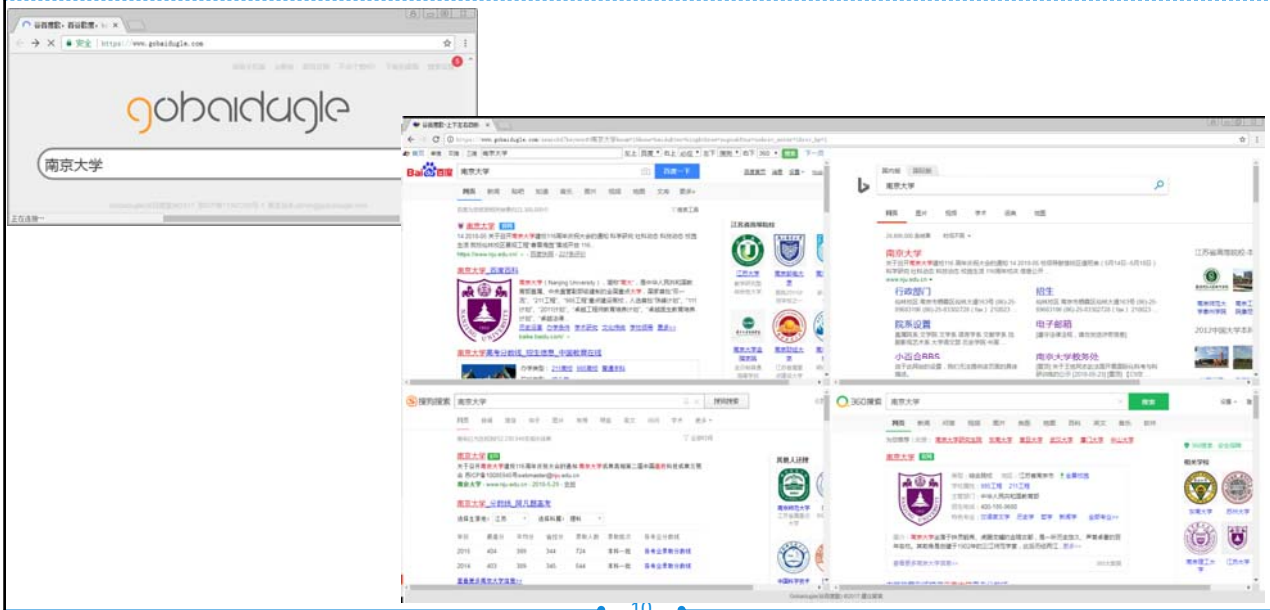
搜索引擎的分类

- 根据检索方式分类
  - 分类目录、关键词搜索引擎、混合搜索引擎
- 根据信息覆盖范围及适用用户群分类
  - 综合搜索引擎、专用搜索引擎（垂直搜索引擎）
- 根据搜索范围分类
  - 独立搜索引擎、集成搜索引擎/元搜索引擎

9

信息检索

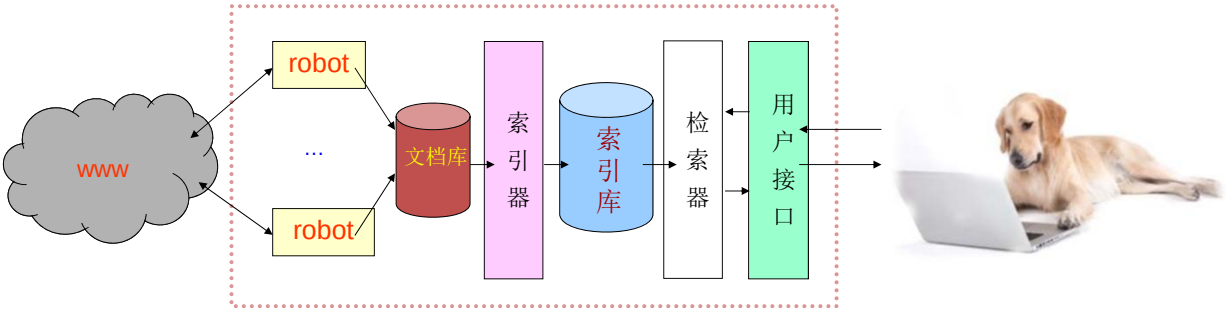
集成/元搜索引擎



10

信息检索

搜索引擎的基本结构



采集 处理 服务

11

信息检索

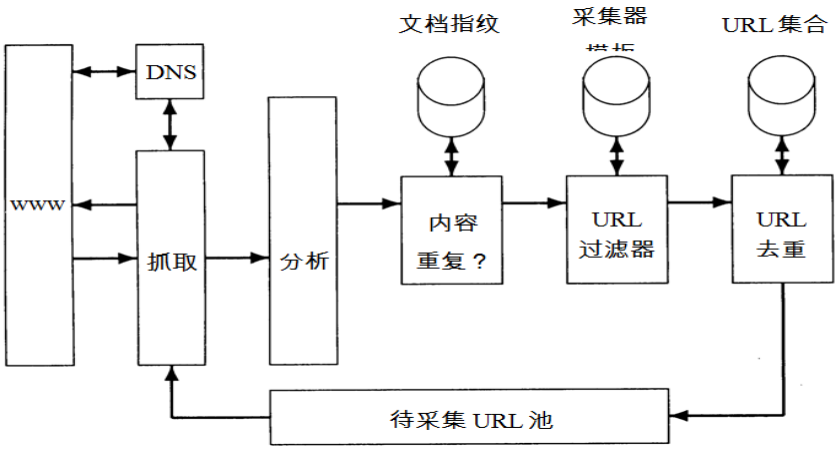
基本的采集过程

- 初始化采集URL种子队列；
- 重复如下过程：
  - *从队列中取出URL*
  - *下载并分析网页*
  - *从网页中抽取更多的URL*
  - *将这些URL放到队列中*
- 基本假设：**Web的连通性很好**

12

信息检索

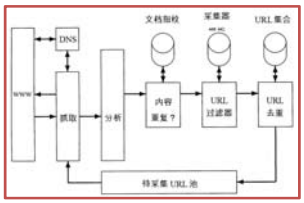
单机采集架构

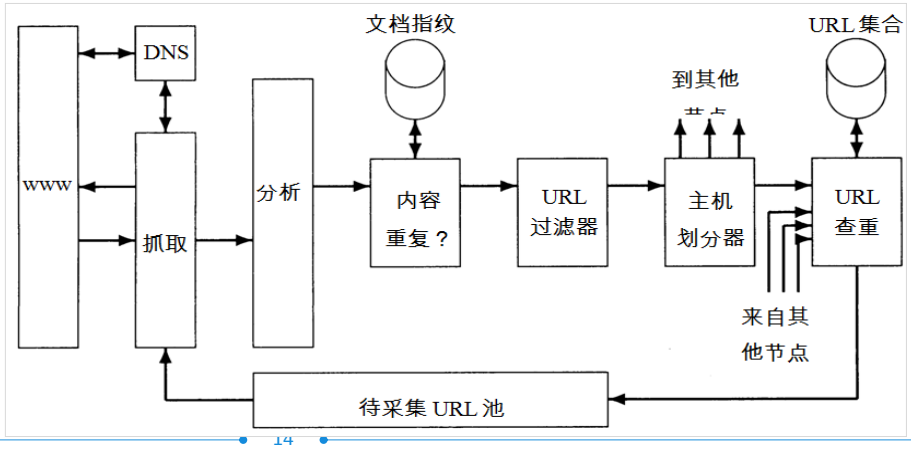


13

信息检索

分布式采集器





14

信息检索

爬行策略


定点策略


定题策略


广度优先

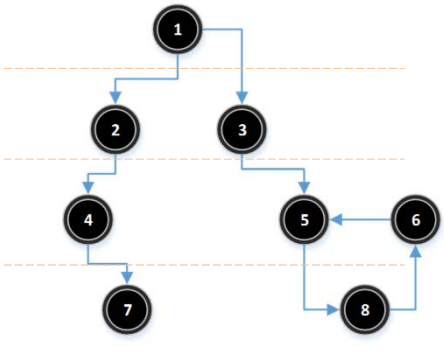

深度优先


大站/要站优先


Partial PageRank/OPIC排序

信息检索

广度优先与深度优先



Breadth-First Search，即广度优先搜索，又称作宽度优先搜索，或横向优先搜索，简称**BFS**

1-2-3-4-5-7-8-6

Depth-First Search，即深度优先搜索，简称**DFS**

1-2-4-7-3-5-8-6

采集时机

- **即时抓取**
  - 用户提交查询的时候即时去网上抓取网页
  - 缺点：系统效益不高（重复抓取网页）
- **预先搜集（直接或间接）**
  - 定期搜集
    - 每次搜集替换上一次的内容
    - 优点：实现简单
    - 缺点：时新性（freshness）不高；重复搜集带来的额外宽带开销
- **增量搜集**
  - 开始时搜集一批网页，以后
    - 只搜集新出现的网页
    - 搜集那些在上次搜集后有过改变的网页
    - 发现自从上次搜索后已经不再存在了的网页，并从网页库中删除
  - 优点：每次搜集的网页量不是很大，可以经常启动搜集过程；时新性比较高
  - 缺点：系统实现比较复杂；不仅搜集过程复杂，而且后续创建索引的过程也很复杂

注意事项

效率

如何利用尽量少的资源（计算机设备、网络带宽、时间）来完成预定的网页搜集量

礼节

网页被搜索引擎索引，从而可能得到更多的访问流量
搜索引擎的“密集”抓取活动阻碍了用户通过浏览器的访问

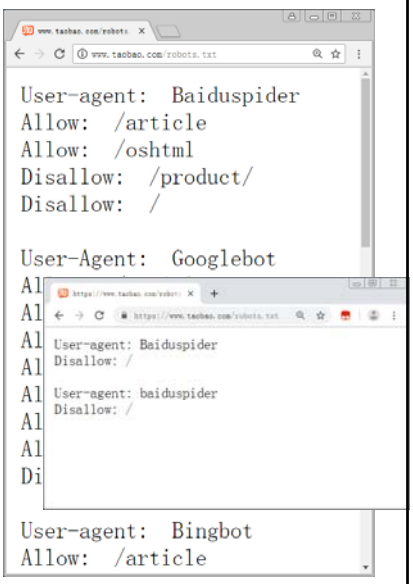
质量

在有限的时间，搜集有限的网页，不要漏掉那些很重要的网页
保证每个网页不被重复抓取

信息检索

Robots协议

- **网站通过该协议告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取。**
  - ✓ User-agent　　*User-agent: Baiduspider*
  - ✓ Disallow　　*Disallow: /cgi-bin/*.htm*
  - ✓ Allow　　*Allow: .gif$*
  - ✓ Sitemap
- **meta Robots标签**
  - ✓ *<meta name="robots" content="all" />*
  - ✓ name可以针对需要进行修改
  - ✓ content的4个指令：index、noindex、follow、nofollow
    - index 指令告诉搜索机器人抓取该页面；
    - follow 指令表示可以沿着该页面上的链接继续抓取下去；



19

信息检索

几个概念

- **锚文本**
  - 把关键词做一个链接，指向别的网页，这种形式的链接（文本）就叫作**锚文本**
- **入链**
  - In-link/inbound link
- **出链**
  - Out-link/outbound link
- **中心节点**
  - Hub
- **权威节点**
  - Authority

20

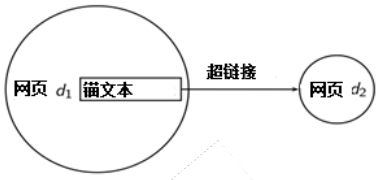
信息检索

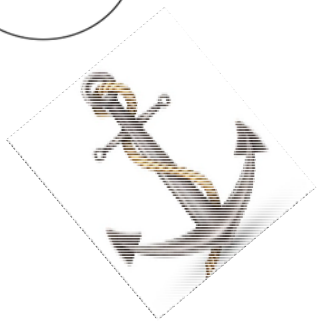
网页排序与链接分析

- 早期搜索引擎主要是比较查询与页面的相关度
  - TF-IDF、SVM、Cosine……
- 链接分析，源于对Web结构中超链接的多维分析。
- 类似于引文分析
  - 论文的价值可以用引用频次来衡量
- 竞价排名？！

信息检索

Web是一个有向图





假设1: 超链接代表了某种质量认可信号

- 超链 $d_1 \rightarrow d_2$ 表示 d_1的作者认可 d_2 的质量和相关性

假设 2: 锚文本描述了文档 d_2 的内容

- 这里的锚文本定义比较宽泛，包括链接周围的文本

- 例子： “You can find cheap cars <a href =http://…>*here* . ”
- 锚文本: “You can find cheap cars here”

链接中心

iSchools

University of Pittsburgh School of Information Sciences

信息检索

锚文本的价值

- 后者往往效果好于前者

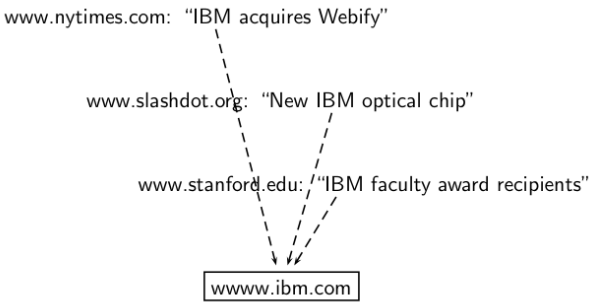
  [d_2中文本] vs. [d_2中文本] + [锚文本 → d_2]

- 例子: 查询 *IBM*
  - IBM 的版权页匹配上
  - 很多作弊网页匹配上
  - IBM的wikipedia页面
  - 可能与IBM 的主页并不匹配!
  - … 也许 IBM 的主页上大部分都是图

- 而按照 [锚文本 → d_2] 来搜索效果会比较好
  - 这种表示下，出现IBM最多的是其主页 www.ibm.com

信息检索

锚文本指向示例



- 锚文本往往比网页本身更能揭示网页的内容
- 在计算过程中，锚文本应该被赋予比文档中文本更高的权重

信息检索

引用分析

- 引用分析：科技文献中的引用分析
- 一个引用的例子: “Miller (2001) has shown that physical activity alters the metabolism of estrogens.”
- 可以把“Miller (2001)” 看成是两片学术文献之间的超链接
- 在科技文献领域使用这些“超链接”的一个应用:
  - 根据他人引用的重合率来度量两篇文献的相似度，这称为共引相似度
  - 在Web上也存在共引相似度：Google中提供的 “find pages like this” 或者 “Similar” 功能

信息检索

引用分析

- 另一个应用: 引用频率可以用度量一篇文档的影响度
  - 最简单的度量指标：每篇文档都看成一个投票单位，引用可以看成是投票，然后计算一篇文档被投票的票数。当然这种方法不太精确。
- 在Web上： **引用频率=入链数**
  - 入链数目大并不一定意味着高质量...
  - ... 主要原因是因为存在大量作弊链接…
- 更好的度量方法： 对不同网页来的引用频率进行加权
  - 一篇文档的投票权重来自于它本身的引用因子
  - 会不会出现循环计算？答案是否定的，实际上可以采用良好的形式化定义

信息检索

PageRank

Larry Page（1973.3-）

PageRank

1998.1申请，2001.9授权，US专利号：6,285,999

Google排名的重要组成部分

主要思想

拥有**越多、越重要入链**的页面越有价值

27

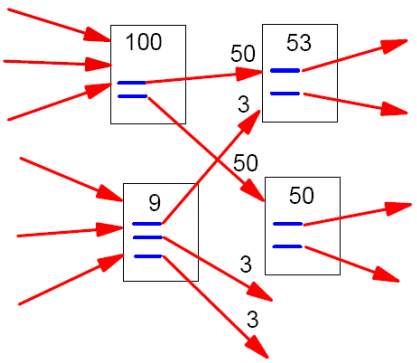
信息检索

原始的PageRank公式

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

$R(u)$和$R(v)$是分别是网页u、v的PageRank值，B_u指的是**指向**网页u的网页集合、N_v是网页v的**出链**数目。

一个网页的PageRank等于所有的指向它的网页的PageRank的分量之和（c为归一化参数）。网页的每条出链上每个分量上承载了相同的PageRank分量。



28

信息检索

PageRank的特点

(1)一个网页如果它的入链越多，那么它也越重要(PageRank越高)；

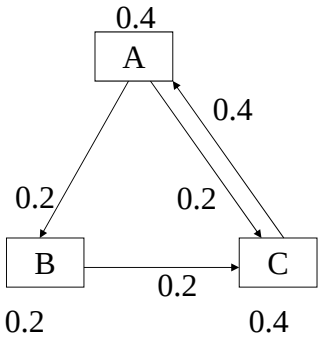
(2)一个网页如果被越重要的网页所指向，那么它也越重要(PageRank越高) 。

类比：(1) 打电话； (2)微博粉丝

29

信息检索

简单计算的例子(c=1)



R(A)=R(C)

R(B)=0.5R(A)

R(C)=R(B)+0.5R(A)

R(A)+R(B)+R(C)=1

解上述方程得：

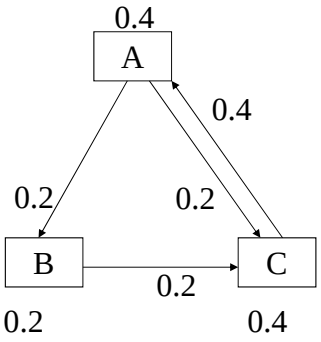
R(A)=R(C)=0.4

R(B)=0.2

30

信息检索

简单计算的例子(c=1)：迭代法求解



$R(A)=R(C)$
$R(B)=0.5R(A)$
$R(C)=R(B)+0.5R(A)$
$R(A)+R(B)+R(C)=1$

迭代次数	R(A)	R(B)	R(C)
0	1/3	1/3	1/3
1	1/3	1/6	1/2
2	1/2	1/6	1/3
3	1/3	1/4	5/12
…	…	…	…
收敛	2/5	1/5	2/5

31

信息检索

转化成矩阵形式

- 令$\boldsymbol{R}$表示所有N个网页的PageRank组成的列向量，令网页间的连接矩阵$\boldsymbol{L}=\{l_{ij}\}$，$P_i$有链接指向$P_j$时，$l_{ij}=1$，否则$l_{ij}=0$。对$\boldsymbol{L}$的每行进行归一化，即用$P_i$的出度$N_i$去除得到矩阵$\boldsymbol{A}=\{a_{ij}\}$，$a_{ij}=l_{ij}/N_i$，则有($\boldsymbol{A}^{\mathrm{T}}$表示$\boldsymbol{A}$的转置矩阵)：

$$\boldsymbol{R}=c\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R} \iff c^{-1}\boldsymbol{R}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}$$

根据线性代数中有关特征向量和特征值的理论，R是矩阵A^{T}的c^{-1}特征值对应的特征向量

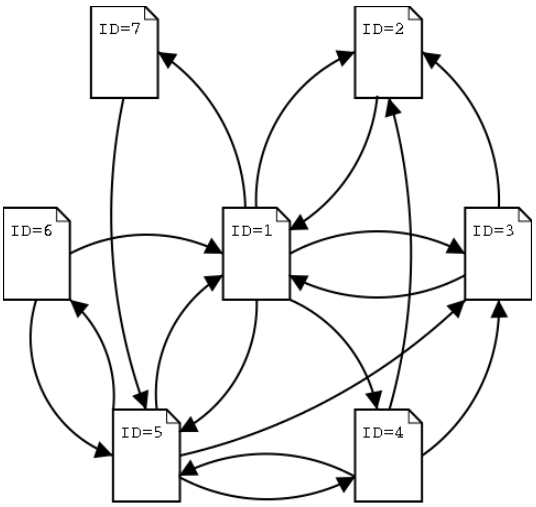
$R(A)=R(C)$
$R(B)=0.5R(A)$
$R(C)=R(B)+0.5R(A)$

$$\begin{bmatrix} R(A) \\ R(B) \\ R(C) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \end{bmatrix} \begin{bmatrix} R(A) \\ R(B) \\ R(C) \end{bmatrix}$$

32

信息检索

一个稍微复杂的例子



Page ID	OutLinks
1	2,3,4,5,7
2	1
3	1,2
4	2,3,5
5	1,3,4,6
6	1,5
7	5

$$A=\begin{pmatrix}0&1&1&1&1&0&1\\1&0&0&0&0&0&0\\1&1&0&0&0&0&0\\0&1&1&0&1&0&0\\1&0&1&1&0&1&0\\1&0&0&0&1&0&0\\0&0&0&0&1&0&0\end{pmatrix}$$

33

信息检索

计算过程

则归一化后A =

$$\begin{pmatrix}0&1/5&1/5&1/5&1/5&0&1/5\\1&0&0&0&0&0&0\\1/2&1/2&0&0&0&0&0\\0&1/3&1/3&0&1/3&0&0\\1/4&0&1/4&1/4&0&1/4&0\\1/2&0&0&0&1/2&0&0\\0&0&0&0&1&0&0\end{pmatrix}$$

$R=cA^{T}R$，令c=1，解得

$$R=\begin{pmatrix}0.69946\\0.38286\\0.32396\\0.24297\\0.41231\\0.10308\\0.13989\end{pmatrix}$$

Normalized =

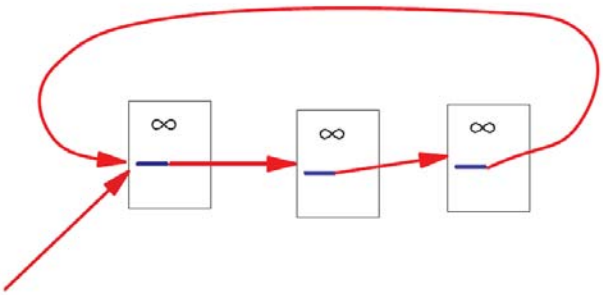
$$\begin{pmatrix}0.303514\\0.166134\\0.140575\\0.105431\\0.178914\\0.044728\\0.060703\end{pmatrix}$$

34

信息检索

原始PageRank的一个不足

A loop:

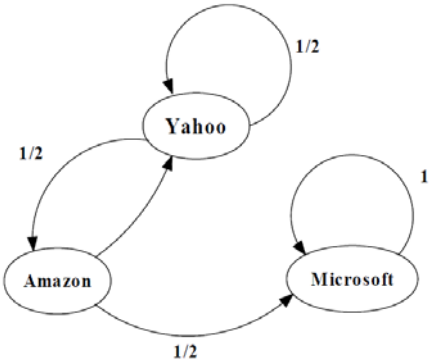


图中存在一个循环通路，每次迭代，该循环通路中的每个节点的PageRank不断增加，但是它们并不指出去，即不将PageRank分配给其他节点！

35

信息检索

一个例子



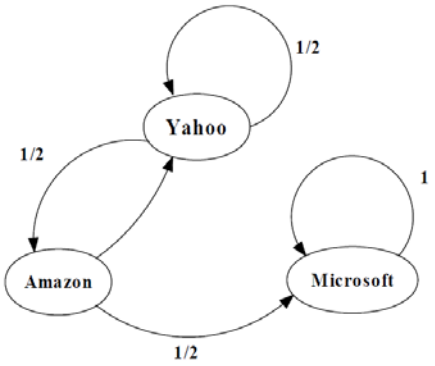
$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/6 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

36

信息检索

一个例子



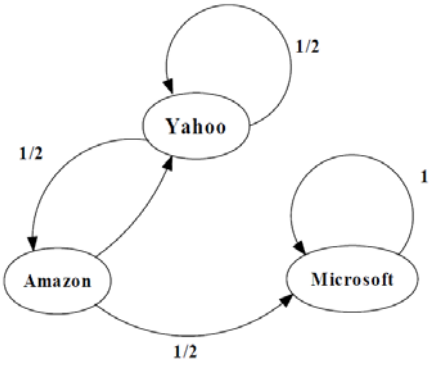
$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/6 \\ 7/12 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/6 \\ 1/2 \end{bmatrix}$$

37

信息检索

一个例子



$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 5/24 \\ 1/8 \\ 2/3 \end{bmatrix} \begin{bmatrix} 1/6 \\ 5/48 \\ 35/48 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

38

信息检索

改进的PageRank公式

随机冲浪或随机游走(Random Walk)模型：到达u的概率由两部分组成：一部分是直接随机选中的概率(1-*d*)或(1-*d*)/*N*，另一部分是从指向它的网页顺着链接浏览的概率，则

$$R(u)=(1-d)+d\sum_{v\in B_u}\frac{R(v)}{N_v} \quad \text{或} \quad R(u)=\frac{(1-d)}{N}+d\sum_{v\in B_u}\frac{R(v)}{N_v}$$

上述两个公式中，后一个公式所有网页PageRank的和为1，前一个公式的PageRank和为$N(1-d)+d$ 。

可以证明，PageRank是收敛的。计算时，PageRank很难通过解析方式求解，通常通过迭代方式求解。*d*通常取0.85

39

信息检索

PageRank面对的Spamming问题

- SEO (Search Engine Optimization)：通过正当或者作弊等手段提高网站的检索排名(包括PageRank)排名。
- 因此，实际中的PageRank实现必须应对这种作弊，实际实现复杂得多。实际中往往有多个因子(比如内容相似度)的融合。

40

信息检索

IBM的HITS算法

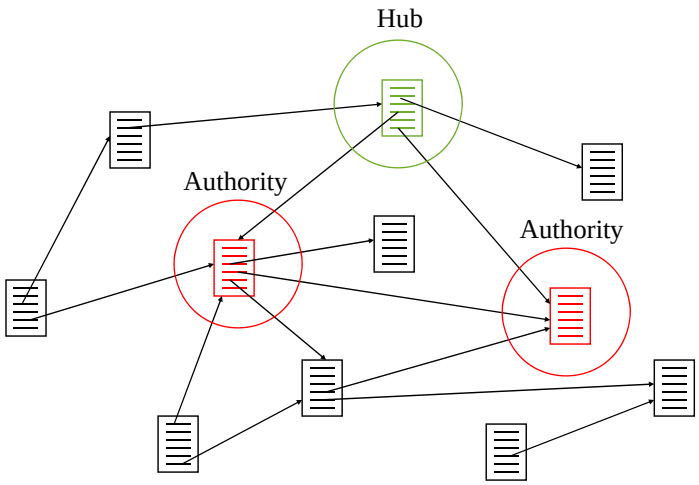
- HITS(Hyperlink-Induced Topic Search)

- 每个网页计算两个值
  - Hub：作为目录型或导航型网页的权重
  - Authority：作为权威型网页的权重

41

信息检索

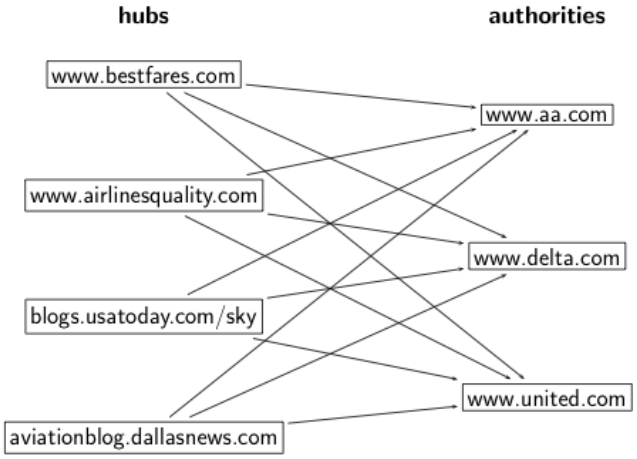
Hub & Authority示意



42

信息检索

例子



43

信息检索

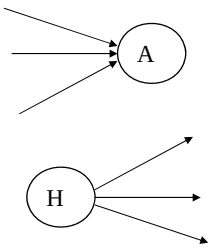
计算方法

$A(p) = \sum H(q_i)$

(其中q_i是所有链接到p的页面)

$H(p) = \sum A(r_i)$

(其中r_i是所有页面p链接到的页面)



(1) 一个网页被越重要的导航型网页指向越多，那么它的Authority越大；

(2) 一个网页指向的高重要度权威型网页越多，那么它的Hub越大。

HITS算法也是收敛的，也可以通过迭代的方式计算。

44

信息检索

HITS的计算过程

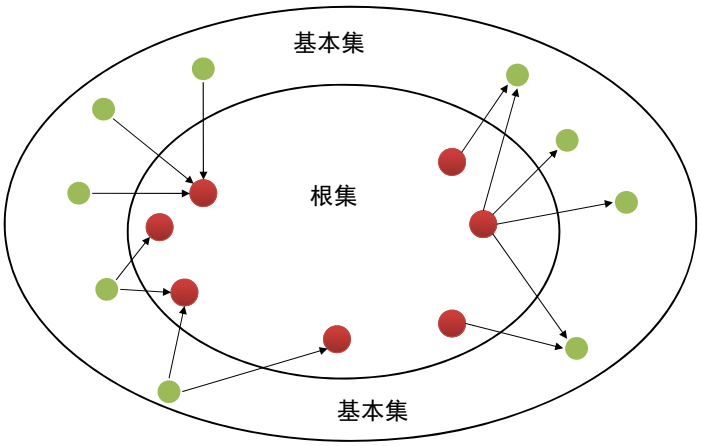
- 首先进行Web搜索；
- 搜索搜索的结果称为根集(root set)；
- 将所有链向种子集合和种子集合链出的网页加入到种子集合；
- 新的更大的集合称为基本集(base set)；
- 最后，在基本集上计算每个网页的hub值和authority值 (该基本集可以看成一个小的Web图)。

45

信息检索

根集和基本集

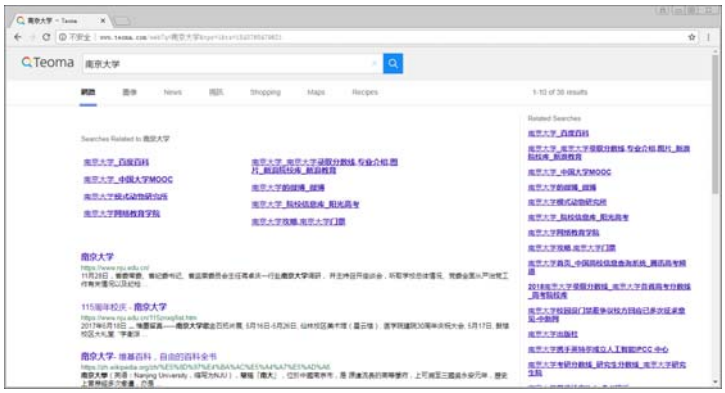
- 根集往往包含200-1000个节点
- 基本集可以达到5000个节点



46

信息检索

HITS缺点



- 计算效率低
  - 实时计算
- 主题漂移
  - 扩展集可能与主题无关
- 易作弊
- 结构不稳定
  - 删改个别少数关系，结果可能变化大

信息检索

PageRank vs. HITS

- 网页的PageRank与查询主题无关，可以事先算好，因此适合于大型搜索引擎的应用。
- HITS算法的计算与查询主题相关，检索之后再进行计算，因此，不适合于大型搜索引擎。

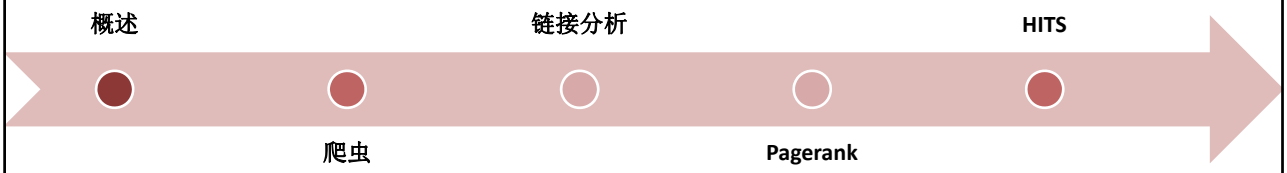
信息检索

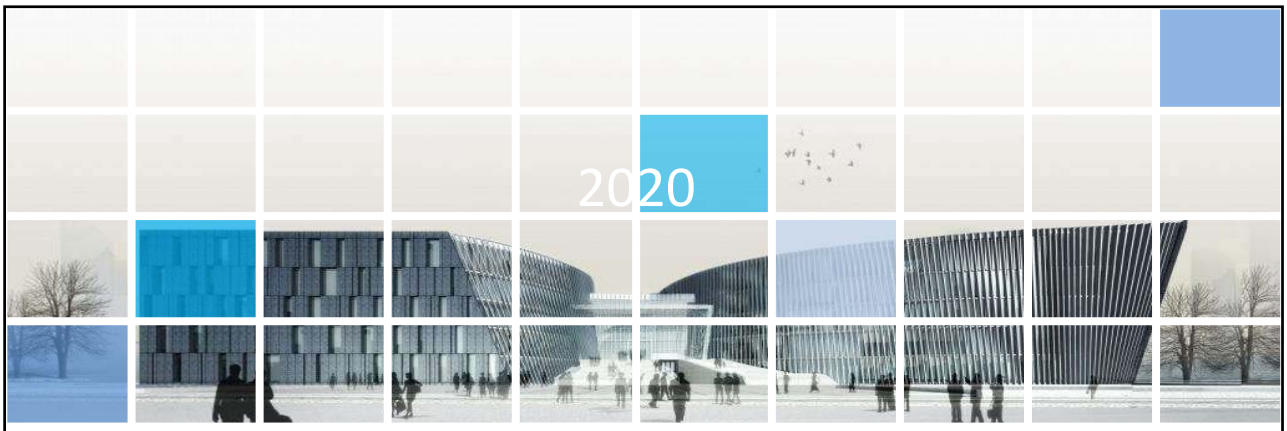
后续

- Web文本结构分析
- 正文提取
- 复杂网络分析
- .....

信息检索

小结





南京大学信息管理学院

信息检索

邓三鸿
njuir@sina.com

PART Ten

信息检索的评价

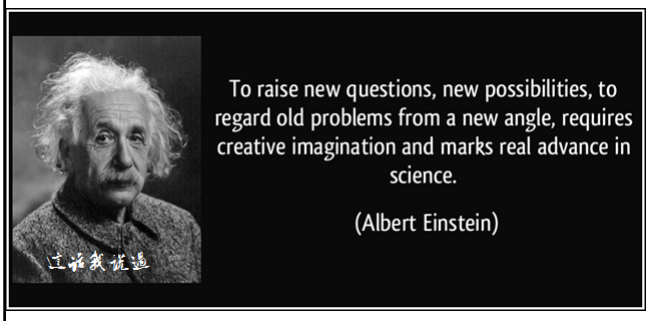
Evaluation of IR

关于评价

评价（Evaluation）：发现和收集关于某种活动的数据，从中判断该项活动的质量及达到预期目标程度的行为。简单地说，评价就是对系统的价值和效率进行测评。

信息检索系统评价：根据给定的指标体系，采用一定的方法和程序，对信息检索系统的功能、特性和运营状况进行**评测**，或对有关假设、预期效益、性能值进行**验证**，以确定系统达到了何种水平、投入成本是否值得、是否可以改进和如何改进，乃至系统是否应生存下去。

IR评价的意义



了解已有检索系统的功能，找出缺陷并改进；

比较各种检索系统的优劣；

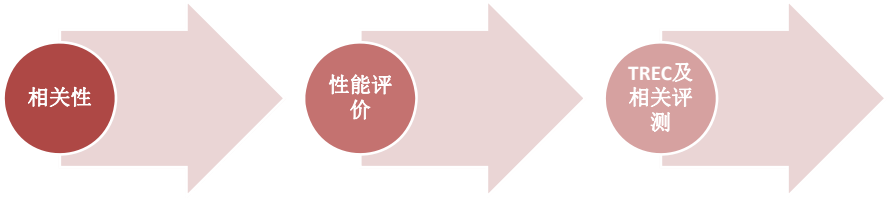
提高效率和效益；

有助于新的检索系统的设计；

丰富信息检索的理论。

信息检索

IR评价的相关内容



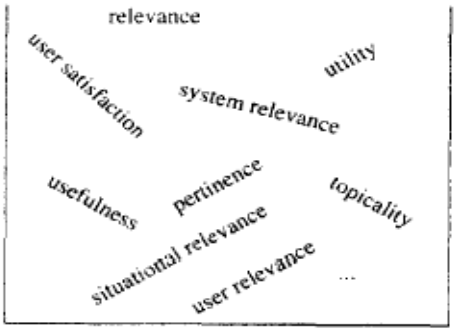
5

信息检索

确定性与相关性

数据检索是“确定性”的
信息检索是“相关性”的

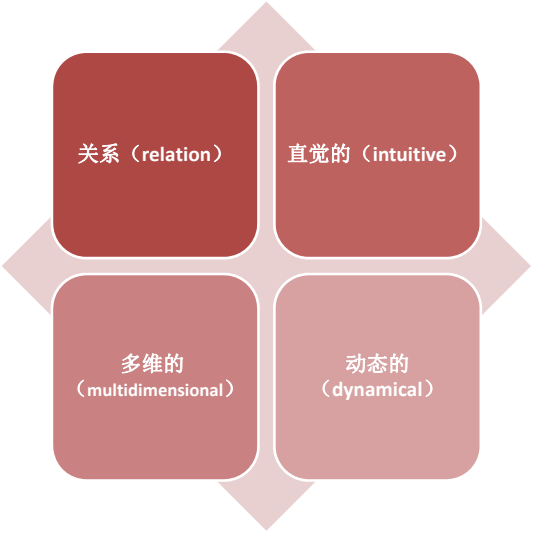
信息检索中的“相关性”主要是指检索系统针对用户的信息需求从文档集合中检出的文档与用户需求之间的一种匹配关系。



6

信息检索

相关性的本质特征



信息检索

米扎罗的相关性问题模型

- **信息源**

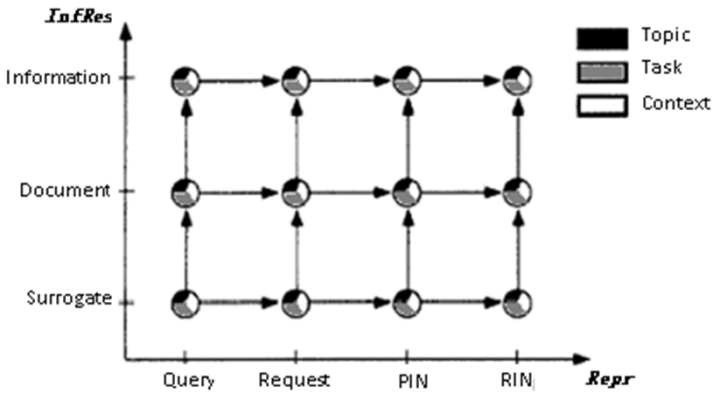
  Surrogate < Document < Information

- **用户信息需求**

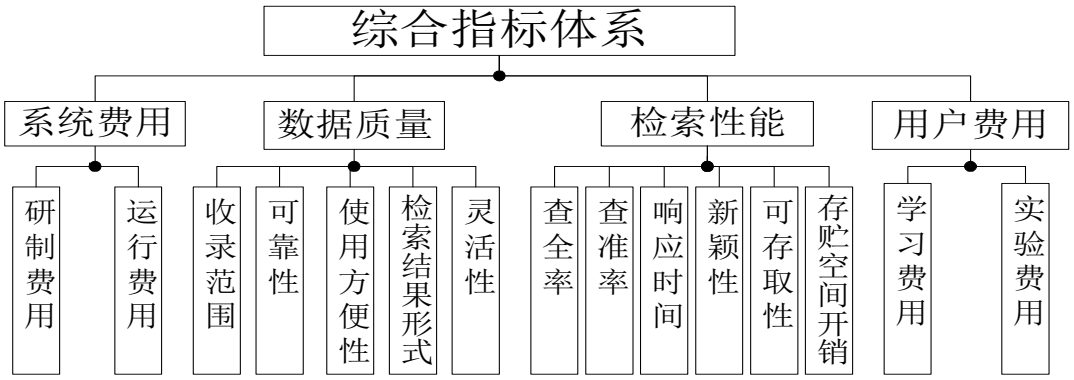
  RIN < PIN < Request < Query

- **时间**
- **组件**

  主题、任务、情境或语境



综合评价体系（参考）



信息检索性能评价及评价指标

- *系统性能指标*
  - 时间效率、空间开销、响应速度……
- **系统角度的相关性判断及评价指标**
  - P、R、F、E……
- 用户角度的相关性判断及评价指标
  - 涵盖率、新颖率……

基本条件

- **一个文档集合*C*。**
  - 系统将从该集合中按照用户查询检出相关文档
- **一组用户查询{q_1, q_2, …, q_n}。**
  - 每个用户查询q_i描述了用户的信息需求
- **对应每个用户查询的标准相关文档集{R_1, R_2…, R_n}。**
  - 该集合可由人工方式构造
- **一组评价指标。**
  - 这些指标反映系统的检索性能。通过比较系统实际检出的结果文档集和标准的相关文档集，对它们的相似性进行量化，得到这些指标值

评价任务示例

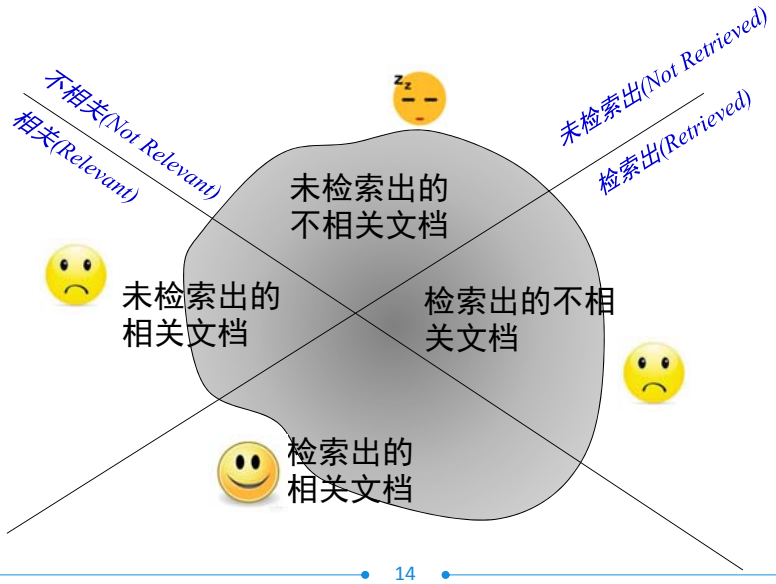
系统&查询	1	2	3	4	…
系统1，查询1	d_3	d_6	d_8	d_{10}	
系统1，查询2	d_1	d_4	d_7	d_{11}	
系统2，查询1	d_6	d_7	d_3	d_9	
系统2，查询2	d_1	d_2	d_4	d_{13}	

缓冲池（Pooling）

对于大规模语料集合，列举每个查询的所有相关文档是不可能的事情，因此，不可能准确地计算召回率！！！

缓冲池（Pooling）方法：对多个检索系统的Top N个结果组成的集合进行标注，标注出的相关文档集合作为整个相关文档集合。这种做法被验证是可行的，在TREC会议等多种测试中被广泛采用。

整个文档集合的划分



信息检索

检索性能评价2*2表

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

15

信息检索

基本评价指标

- 准确率（Precison）
- 召回率（Recall）
- 调和指标F、E
- 平均准确率（AP）

16

查准率

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

查准率/准确率（**Precision ratio**）

——检出的相关信息数量与检出的信息总量的比率

$$P=\frac{\text{检出的相关信息数量}}{\text{检出的信息总量}}=\frac{a}{a+b}$$

查全率

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

查全率/召回率（**Recall ratio**）

——检出的信息数量与检索系统中相关信息总量之间的比率

$$R=\frac{\text{检出的相关信息数量}}{\text{系统中的相关信息数量}}=\frac{a}{a+c}$$

信息检索

Ps:漏检率与误检率

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

$$\text{漏检率（M）}=\frac{\text{未检出的相关文献}}{\text{文档中相关文献总量}}\times 100\%=\frac{c}{a+c}\cdot 100\%$$

$$\text{误检率（N）}=\frac{\text{检出的不相关文献量}}{\text{检出的文献总量}}\times 100\%=\frac{b}{a+b}\cdot 100\%$$

R+M=1， P+N=1

19

信息检索

囊括值

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

- 囊括值（Generality）-系统中相关信息数量与系统的信息总量的比率

$$\text{Generality}=\frac{\text{系统中相关信息数量}}{\text{系统的信息总量}}=\frac{a+c}{a+b+c+d}$$

20

信息检索

非相关检出率

	相关文献	不相关文献	总计
被检出文献	a	b	a+b
未检出文献	c	d	c+d
总计	a+c	b+d	a+b+c+d

- 非相关检出率（Fallout ratio）-检出的非相关信息数量与系统中的非相关信息总量的比率

$$\text{Fallout} = \frac{\text{检出的非相关信息数量}}{\text{系统中的非相关信息数量}} = \frac{b}{b+d}$$

信息检索

P-R例子

- $R_q=\{d_3,d_5,d_9,d_{25},d_{39},d_{44},d_{56},d_{71},d_{89},d_{123}\}$
- 通过某一个检索算法得到的排序结果：

(100%,10%)

1. d_{123} ●
2. d_{84}
3. d_{56} ●
4. d_6
5. d8

(66%,20%)

(50%,30%)

6. d_9 ●
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25} ●

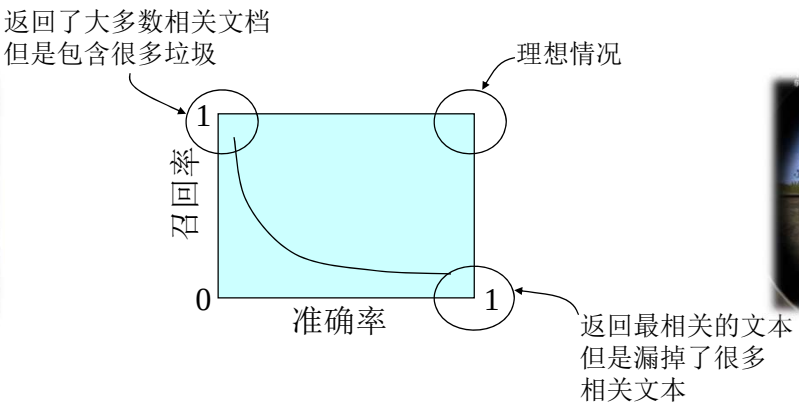
(40%,40%)

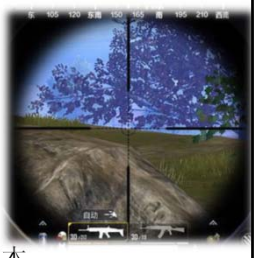
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 ●

(33%,50%)

(Precision, Recall)

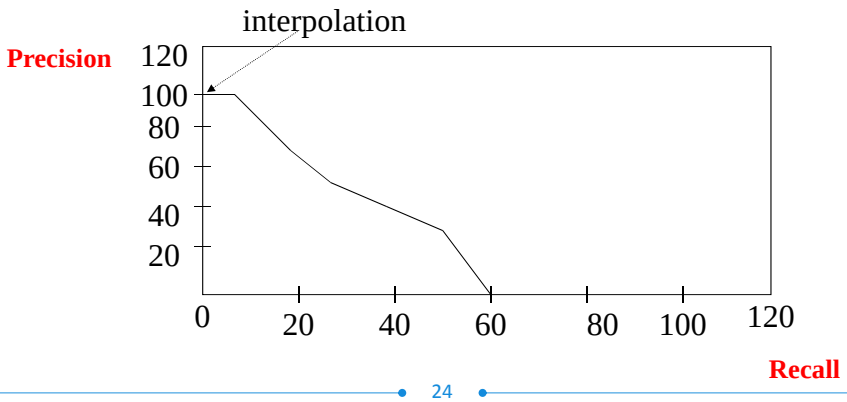
P-R的关系曲线





11点标准P-R曲线

- 11个标准查全率水平所对应的查准率: 0%, 10%, 20%, …, 100%



平均准确率

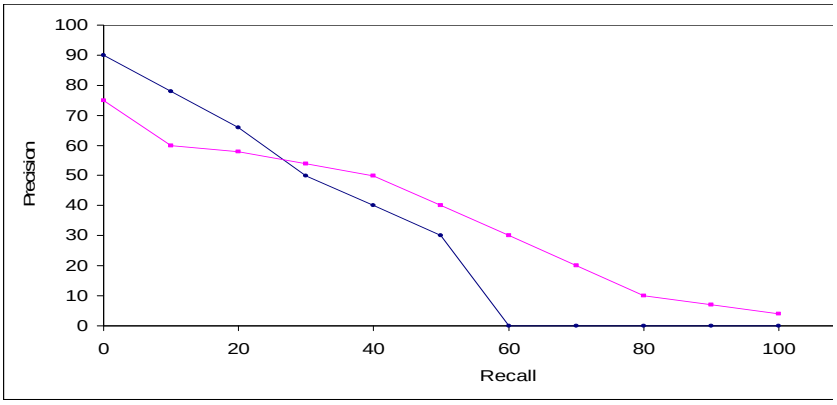
为了评价某一算法对于**所有测试查询**的检索性能，对**每个召回率水平下的准确率**进行平均化处理，公式如下：

$$\overline{P}(r)=\frac{\sum_{i=1}^{N_q}P_i(r)}{N_q}$$

- N_q: ——使用的查询总数
- $P_i(r)$ ——在召回率为r时的第i个查询的准确率

比较示例

- 对多个查询，进行平均，有时该曲线也称为：准确率/召回率的值。
- 如下为两个检索算法在多个查询下的准确率/召回率的值。
  - 第一个检索算法在低召回率率下，其准确率较高。
  - 另一个检索算法在高召回率下，其准确率较高



P-R评价的问题

- 两个指标分别衡量了系统的某个方面，但是为比较带来了难度，究竟哪个系统好？大学最终排名也只有一个指标。

  解决方法：单一指标，将两个指标融成一个指标

- 两个指标都是基于集合进行计算，并没有考虑序的作用

  举例：两个系统，对某个查询，返回的相关文档数目一样都是10，但是第一个系统是前10条结果，后一个系统是最后10条结果。显然，第一个系统优。但是根据上面基于集合的计算，显然两者指标一样。

  解决方法：引入序的作用

- 召回率难以计算

  解决方法：Pooling方法，或者不考虑召回率？

P-R的综合-F值

调和平均值（ Harmonic Mean ）是将准确率和召回率加权平均的评价方法

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2 \times P \times R}{P + R}$$

F∈[0,1]

信息检索

例子

1. d123
2. d84
3. d56 ●
4. d6
5. d8

(33.3%,33.3%)

6. d9
7. d511
8. d129 ●
9. d187
10. d25

(25%,66.6%)

11. d38
12. d48
13. d250
14. d113
15. d3 ●

(20%,100%)

$$F(3)=\frac{2}{\frac{1}{0.33}+\frac{1}{0.33}}=0.33 \quad F(8)=\frac{2}{\frac{1}{0.25}+\frac{1}{0.67}}=0.36 \quad F(15)=\frac{2}{\frac{1}{0.20}+\frac{1}{1}}=0.33$$

29

信息检索

P-R的综合-E指数

$$E=1-\frac{1+b^2}{\frac{b^2}{R}+\frac{1}{P}}=\frac{(b^2+1)\times P\times R}{b^2\times P+R}$$

E∈[0,1]

b为用户指定的参数，可以允许用户调整P和R的相对重要程度

- b=1时，E=1-F。这表示E指数和F指数互补
- b>1时，表示准确率P的重要性大于召回率R
- b<1时，表示召回率R的重要性大于准确率P

30

单值评价指标

MAP

P@10

R准确率

准确率直方图

MAP

- Mean Average Precision, 平均准确率均值
- 单个查询的平均准确率是逐个考察排序中每个新的相关文档，然后对其准确率值进行平均后的平均值;
- 查询集合的平均准确率是每个查询的平均准确率MAP的平均值，MAP的计算公式如下:

$$MAP = \frac{1}{r}\sum_{i=1}^{r}\frac{i}{\text{第}i\text{个相关文档的位置}}$$

•r 为相关文档数

- MAP是反映系统在全部查询上性能的单值指标
- 系统检索出来的相关文档位置越靠前， MAP就可能越高.
- 如果系统没有返回相关文档，则MAP默认为0.

计算MAP举例

$$MAP = \frac{1}{r}\sum_{i=1}^{r}\frac{i}{\text{第}i\text{个相关文档的位置}}$$

- $R_q=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- 通过某一个检索算法得到的排序结果：

1. d_{123} •	6. d_9 •	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} •	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d8	10. d_{25} •	15. d_3 •

AP=(1+0.66+0.5+0.4+0.33)/5=0.578

p@10

p@10——系统对于查询返回的**前10个结果的准确率**.

- 对于搜索引擎系统来讲，由于没有一个搜索引擎系统能够保证搜集到所有的网页，所以召回率很难计算，因而**准确率成为目前的搜索引擎系统主要关心的指标**
- 考虑到用户在查看搜索引擎结果时，往往希望在第一个页面（如10个结果）就找到自己所需的信息，因此P@10能比较真实有效地反映在真实应用环境下所表现的性能

信息检索

p@10计算

- $R_q=\{d_3,d_5,d_9,d_{25},d_{39},d_{44},d_{56},d_{71},d_{89},d_{123}\}$
- 通过某一个检索算法得到的排序结果：

1. d_{123} •	6. d_9 •	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} •	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d8	10. d_{25} •	15. d_3 •

0.4

信息检索

R-Precision

- 单个查询的*R*准确率是指检索出**R篇文档时的准确率**.
- R是当前检索中**相关的文档**总数
- 查询集合中所有查询的R准确率是每个查询的R准确率的平均值.

$$R-\Pr ecision=\frac{\text{前}R\text{篇文档中实际相关文档数}}{R}$$

信息检索

R-Precision计算

$$R-\text{Precision}=\frac{\text{前}R\text{篇文档中实际相关文档数}}{R}$$

- $R_q=\{d_3,d_5,d_9,d_{25},d_{39},d_{44},d_{56},d_{71},d_{89},d_{123}\}$
- 通过某一个检索算法得到的排序结果：

1. d_{123} •	6. d_9 •	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} •	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d8	10. d_{25} •	15. d_3 •

10-precision=4/10=0.4

信息检索

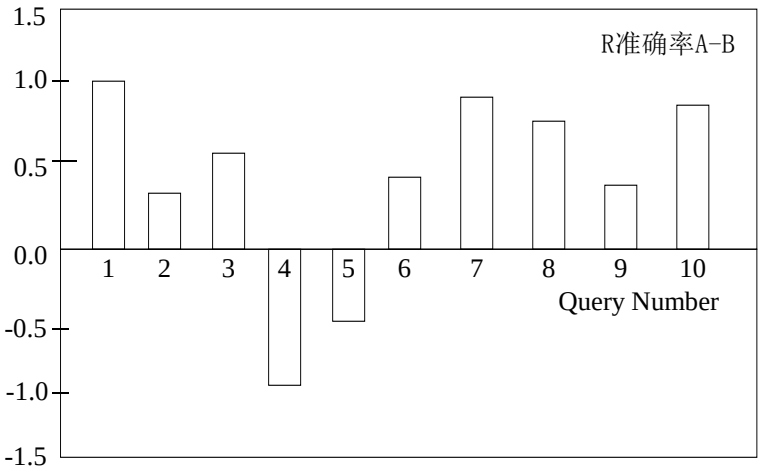
准确率直方图

- 多个查询的**R-Precision**测度
- 用来比较两个算法的检索纪录
- 用$RP_A(i)$和$RP_B(i)$分别表示使用检索算法A和检索算法B检索第i个查询时得到的R准确率,它们之间的差值

$$RP_{A-B}(i)=RP_A(i)-RP_B(i)$$

- $RP_{A-B}=0$:对于第i个查询，两个算法有相同的性能
- $RP_{A-B}>0$:对于第i个查询，算法A有较好的性能
- $RP_{A-B}<0$:对于第i个查询，算法B有较好的性能

准确率直方图：例



单指标评价小结

- 随着信息技术以及互联网的发展，信息检索研究所采用的数据集越来越大，因此构建完整的相关判断越来越难；
- 在相关判断不完整的情况下，采用现有评价方法得出的测试结果会有失公正；
- 对于搜索引擎这样的对高相关性文档进行检索的任务来讲，传统的评价方法也无法很好地对任务评测；
- 特殊指标： ~~*Bpref、N(D)CG、单一相关文档检索的评价*~~

面向用户的评价

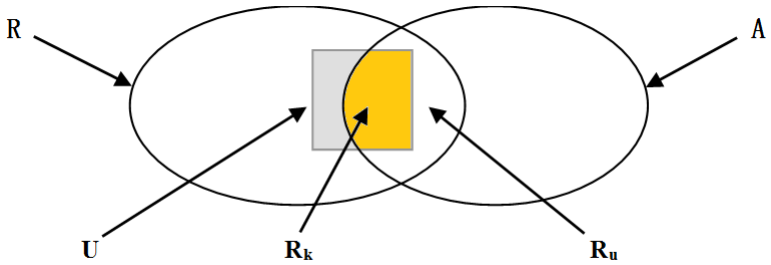
- 面向用户的测度方法/User-Oriented Measures
  - 覆盖率：实际检出的相关文献中用户一致的相关文献所占比例

$$coverage = \frac{|R_k|}{|U|}$$

  - 新颖率：检出的相关文献中用户未知的相关文献所占的比例

$$novelty = \frac{|R_u|}{|R_u| + |R_k|}$$

覆盖率和新颖率



R ——相关文档集

A ——返回文档集

U ——用户的相关文档集

R_k ——返回的、用户已知的文档集

R_u ——返回的，用户未知的文档集

$$coverage = \frac{|R_k|}{|U|} \qquad novelty = \frac{|R_u|}{|R_u| + |R_k|}$$

统一评测

- 同一个算法在不同的数据条件下得到的结果差异很大；
- 没有统一的测试方法和共同的数据集合，几乎不可能比较不同算法；
- 数据采集需花费很大的人力物力，而由政府学术机构或者学术团体组织的开放技术评测，可以为科研提供一种统一的、普遍认可的评价基准和大型测试集，节省了各个研究者重复采集数据而造成的重复付出，对整个领域的科学研究和技术进步起到很大的推动作用；
- 通过技术评测可以提出新的研究问题

国外的评测I

- **The Cranfield Experiments**, by *Cyril W. Cleverdon*

  1957 –1968 （上百篇文档集合）

  http://ir.dcs.gla.ac.uk/resources/test_collections/cran/

- **SMART System**, by *Gerald Salton*

  1964-1988 （数千篇文档集合）

Gerald Salton, 1927–1995

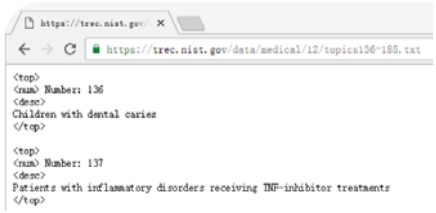
国外的评测II

- **TREC评测**
  - 文本检索会议（Text Retrieval Conference，TREC）是信息检索(IR) 界为进行检索系统和用户评价而举行的活动，它由美国国家标准技术协会(NIST) 和美国高级研究计划局（DARPA） 共同资助，始于1992年。
  - 检索评测中的奥运会！！
- **NTCIR评测**
  - NTCIR(NACSIS Test Collection for IR Systems)始于1998年，是由日本国立信息学研究所（National Institute of Informatics，简称NII）主办的搜索引擎评价型国际会议
- **CLEF评测**
  - CLEF于2000年开始筹办，是欧洲各国共同合作进行的一项长期研究计划，主要想通过评测信息科技技术，促进欧洲语言中的各种单一语言以及多语言信息技术的发展，
  - CLEF的目标只在于跨语言信息检索以及多语言信息检索方面

TREC评测

- 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研究所National Institute of Standards and Technology （NIST）联合发起
  - 参加者免费获得标准训练和开发数据
  - 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

TREC目的



- 促进基于大规模测试文档集的检索研究。
- 为了反映现实系统的主题多样性，必须保证有足够的实验语料集，TREC的文献集合一般在G级左右，包括50~100万篇文献（近几年更大，可达数千万记录，T级存储）；
- 建立一个开放的论坛来交流研究思想，使与会者能交流研究的成果与心得，促进企业学术机构和政府部门之间的交流沟通。
- 通过展示检索方法在解决实际问题中的有效性，来加速实验室技术的商业化产品转换。
- 通过提供大型的语料库、统一的测试程序，有系统地整理评测结果，达到改善文本检索评价和检验方法的目标。

TREC早期任务

- Ad hoc检索任务（传统的批处理检索）
- 类似图书馆里的书籍检索，即书籍库（数据库、文档集合）相对稳定不变，而用户的查询要求是千变万化的
- 主要研究任务包括对大数据库的索引查询、查询的扩展等
- 固定主题检索任务（Information Routing）
- 用户的查询要求相对稳定，而文档集常常发生变化
- 研究的主要任务不是索引，而是对**用户兴趣的建模**，即如何为用户兴趣建立合适的数学模型

信息检索

TREC评测的评价方法

- 概括表统计
- 准确率-召回率平均值
- 文档级别平均值
  - 平均准确率

发布Track → 报名 → 用户测试与提交 → 评估 → 交流

信息检索

参加过TREC的部分单位

Corp.	University	Asian Organization
IBM	MIT	Singapore U. (KRDL)
AT&T	CMU	KAIST
Microsoft	Cambridge U.	Tinghua U. (大陆的清华) TREC11
Sun	Cornell U.	Tsinghua U.(Taiwan) TREC7
Apple	Maryland U.	Taiwan U. TREC8&9&10
Fujitsu	Massachusetts U.	Hongkong Chinese U. TREC9
NEC	New Mexico State U.	Microsoft Research China TREC9&10
XEROX	California Berkeley U.	Fudan U. TREC9&10&11(复旦)
RICOH	Montreal U.	ICT TREC10&11(中科院计算所)
CLRITECH	Johns Hopkins U.	HIT TREC10(哈工大)
NTT	Rutgers U.	北大、软件所、自动化所等
Oracle	Pennsylvania U.	还有更多的大陆队伍逐渐加入……

信息检索

TREC评测的任务（Tracks）

https://trec.nist.gov/tracks.html

- 2018 TREC Tracks
  - CENTRE Track
  - Common Core Track
  - Complex Answer Retrieval Track
  - Incident Streams Track
  - News Track
  - Precision Medicine Track
  - Real-Time Summarization Track

信息检索

NTCIR

http://research.nii.ac.jp/ntcir/index-en.html

➢ NTCIR评测

NTCIR（NACSIS Test Collection for IR Systems）始于1998年，是由日本国立信息学研究所（National Institute of Informatics，简称NII）主办的搜索引擎评价型国际会议

➢ 主要评测任务

✓ 传统的日文、中文、韩文、英文的单语ad hoc任务.

✓ 最重要的任务是跨语言信息检索。若以C、J、K、E分别代表中文、日文、韩文、英文，则有C→CJKE、J→CJKE、K→CJKE、E→CJKE等极为复杂的检索任务。

✓ 另外一个比较重要的任务是中枢语言信息检索，这个任务是模拟在语言资源不足的情况下进行跨语言信息检索。

如要进行C→K的跨语言信息检索，但是没有中韩双语词典，只好借用中英词典以及英韩词典，此时，英语就被视为中枢语言。

CLEF

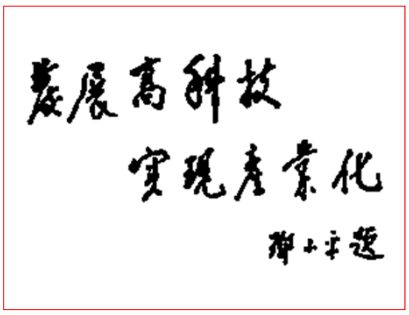
http://clef.isti.cnr.it

http://www.clef-initiative.eu/

- CLEF（Cross-Language Evaluation Forum）评测
  - CLEF（2000-2009）是欧洲各国共同合作进行的一项长期研究计划，主要想通过评测信息科技技术，促进欧洲语言中的各种单一语言以及多语言信息技术的发展。
  - CLEF的目标只在于跨语言信息检索以及多语言信息检索方面
- CLEF的评测任务
  - 跨语言文本检索：包括三个子任务，即单语检索、双语检索以及多语检索。
  - 跨语言专利数据检索：主要是使用专业领域上下文的信息进行单语言以及跨语言的信息检索。
  - 交互式跨语言检索（Interactive Cross-Language Retrieval (iCLEF) ）:尝试模拟实际检索环境下使用者与检索系统的互动情形，以改善信息检索系统的性能。
  - 多语问答:是一种跨语言QA检索评测
  - 图像跨语言检索/跨语言语间检索

国内863评测介绍

- 全名
  - 863计划中文信息处理与智能人机接口技术评测（1991-2005）
- 组织者
  - 国家高技术研究发展计划（863计划）
- 方式
  - 通过网络进行
  - 各单位在自己的环境中运行参评系统
  - 2005年11月召开研讨会
- 2005年度评测内容
  - 机器翻译
  - 信息检索
  - 语音识别



863评测介绍—信息检索评测

- 项目：相关网页检索
- 任务定义：给定主题，返回数据中与该主题相关的网页。
- 数据：CWT100g (中文Web测试集100g)
  - 根据天网搜索引擎截止**2004**年**2**月**1**日发现的中国范围内提供**Web**服务的**1,000,614**个主机，从中采样**17,683**个站点，在**2004**年**6**月搜集获得**5,712,710**个网页（有效网页：**5,594,521**）
  - 包括网页内容和**Web**服务器返回的信息
  - 真实容量为**90GB**。

主题

- **主题**（Topic）模拟了用户需求，由若干字段组成，描述了用户所希望检索的信息。主题和查询的区别在于：主题是对信息需求的陈述，查询则是信息检索系统的实际输入。
- 主题由4个字段组成：
  - **编号**（num）
  - **标题**（title）
  - **描述**（desc）
  - **叙述**（narr）

主题实例

- <top>
- <num>编号：020
- <title> 下载"香奈儿"
- <desc> 描述：mp3格式歌曲“香奈儿”的下载地址
- <narr> 叙述：仅检索具有歌曲“香奈儿”下载地址的网页。有关“香奈儿”的介绍不在检索范围内。提供非mp3格式下载地址的页面不在检索之列。
- </top>

查询的构造

- 自动方式和人工方式
- 自动方式是指在没有任何人为因素的影响下根据主题构造查询的方式
- 除此之外的方式均为人工方式
  - 只允许以人工方式构造查询，不允许在检索过程中加入任何人为因素。
  - 最多返回1000条排序结果

	MAP	R-Precision	P@10
第一名	自动化所0.3175	哈工大0.3672	清华0.6280
第二名	哈工大 0.3107	自动化所0.3607	哈工大0.6240
第三名	清华大学0.2858	清华0.3293	自动化所0.5540

重要会议/小组

SIGIR Special Interest Group on Information Retrieval — http://sigir.org/

Text REtrieval Conference (TREC) — *...to encourage research in information retrieval from large text collections.* — https://trec.nist.gov/

中国计算机学会 CCF — https://www.ccf.org.cn/

CIPS — http://www.cipsc.org.cn/

小结

相关性问题 → 性能评价 → TREC简介