# 信息检索概述

* 信息检索：“信息存储于检索”，将信息按照一定的方式组织和存储起来，并能根据用户的需要找出其中相关信息的过程。
* 早期分为文献检索、事实检索、数据检索

新的三分方法：文本检索、数值检索、音频与视频检索

* 信息检索的基本原理：对信息资源集合与信息需求集合的匹配与选择。
* 信息检索发展时期：  
  手工检索：1.以印刷文献为主要检索对象；2.以各类文摘、题录和目录性工具书为可利用的主要检索工具；3.以图书馆的参考咨询部门为开展信息检索服务的中心机构。  
  计算机化检索：1.以各类机读数据库为检索对象；2.各类情报所、联机服务中心作为新兴的信息服务部门而存在；3.信息检索用户逐渐由专业检索人员向个人终端用户转移和扩散，信息检索的社会普及化程度日益提高。  
  网络化检索：（问题）1.超文本/超媒体技术应用；2.多媒体信息检索；3.自然语言理解；4.海量信息资源的组织与检索；5.检索可视化；6.知识检索与语义检索。
* 主要研究问题：1.信息检索理论；2.信息检索工具/系统；3.信息资源及其收集、加工；4.检索技术与方法；5.用户研究与检索策略；6.其他密切相关的自动化处理技术。
* 相关学科：计算机科学、数学、系统科学、语言学与计算语言学、认知心理学。

# 信息检索的数学模型

* 信息检索的数学模型：运用数学的语言和工具，对信息检索系统中的关键要素及重要操作等进行抽象和编码而建立起来的某种数学表达模式。
* System=（D，Q，F，R（dj，q））
* 用户信息需求：潜在的真实需求、意识到的或感知到的需求、表达出的需求、提问。
* 布尔模型的基本原理：1.系统索引词集合（K）中的每一个索引词在一篇文档中只有两种状态：出现或者不出现；2. 用户提问式由3中布尔逻辑运算符“and”、“or”、“not”连接索引词来构成。
* 布尔模型的优缺点：简单、容易理解、简洁的形式化vs.精确匹配策略问题、布尔逻辑表达用户需求的能力问题。
* 向量空间模型的基本原理：1.文档向量的构造；2.提问向量的构造；3.匹配函数的选择及相似度阈值的确定。
* 向量空间模型的技术特征：1.采用部分匹配策略，使得在算法层面上基于多值相关性的判断处理得以实现；2.采用基于统计学方法的词加权处理模式，使检索效果得到显著改善；3.采用对检索结果排序输出的策略，使对检索结果数量的控制与调整具有相当大的弹性与自由度。
* 经典概率模型的基本思想：给定一个用户提问，则检索系统中存在着一个与该提问相关的理想命中结果集合R。如果能已知集合R的主要特征及其描述，则用户的检索要求便不难实现。但问题是，在用户提出检索要求时，并不知道这个理想结果集合的特性。为此，需要在检索伊始就对R的特性进行某种猜测。根据初始的猜测，系统将检索到一个初步的命中结果集合。在此基础上，用户可以对初始检索结果集合中文档相关与否进行判断，或者，由系统对检索结果文档的相关性进行自动判别。根据这些反馈信息，系统便可以在后续的检索处理中不断做出优化与改进，从而在多处交互操作之后使检索结果逐步接近该提问的理想命中结果集合R。在这里，重要的问题是如何进行初始的猜测以及如何通过相关反馈与交互来不断调整、改善检索性能。
* 经典概率模型对信息检索的理论贡献：吸收了相关反馈原理，并在理论上采用了一种更严密的决策方式。

# 信息检索系统及其构成

* 信息检索系统类型：书本式、卡片式、穿孔卡片、缩微式、计算机化、网络；文献检索、数据库管理、自动问答、管理信息、决策支持。
* 信息检索系统的基本结构：概念结构、物理结构（硬件部分、软件部分、信息资源集合）、逻辑结构（功能模块及其相互关系）  
  信息存储：信息选择与采集、信息标引处理、创建/更新数据库及索引文档；信息查询：用户（检索）接口、提问处理与检索匹配；知识组织与管理工具。
* 印刷型信息资源：便于阅读和流传vs.存储密度低、体积庞大、难以实现自动输入和自动检索。图书、期刊、报纸、会议录、学位论文、科技报告、专利文献、技术标准、政府出版物。
* 电子型信息资源：系统依赖性、复杂性、共享性、不安全性。
* 信息资源的选择与采集方式：人工->人机结合->计算机自动控制。
* 标引：对信息资源的各种检索特征进行分析并使之显性化，以便为存储和检索这两个环节提供某种连接的一种重要的信息加工操作。  
  人工标引、自动标引——全自动标引vs.半自动标引、自动抽词标引vs.自动赋词标引（词频统计标引法、句法模式标引法、语义分析标引法）。
* 自动标引流程：待标引机读文本->语句分析（抽词/切词处理）->候选标引词->词频统计->加权计算->阈值确定->标引词确定（->概念转换->规范词语选取->）->索引文档生成。
* 自动标引中的词语加权方案：  
  绝对词频法：方法简单，容易实现，有一定的实用性vs.仅使用中频词作为标引词，简单排除全部高频词和低频词，不仅可能会降低查全率和查准率，而且确定高频词和低频词的截止界也是很困难的。  
  逆文档频率法：计算方法比较简单，加权效果也比较好。  
  词区分值法
* 中文自动标引的词语切分方案：  
  词典切分法：传统vs.词典的构造困难、维护代价大。  
  切分标记法：字典+构词模式vs.仍需要较多的汉语语言知识和相关专业知识。  
  单汉字法：标引深度最大，检索的查全率能得到最充分的保证vs.对于文本中隐含的概念主题难以揭示和处理，在查准率、检索速度、用户智力负担等方面存在较大缺陷。  
  N-gram法。
* 提问处理与检索匹配作业流程：1.接收用户提问；2.提问校验，包括对用户输入的提问式进行语法、格式、用词等方面的检查；3.提问加工，对经过校验的合法、有效的用户提问式（源提问式）进行解释性或编译性的变换与加工，生成便于机器处理的目标提问式；4.检索匹配，即将目标提问式与数据库索引文档的记录进行匹配，把满足匹配要求的数据库记录输出到结果文档中，或直接显示在检索屏幕上。
* 知识组织与管理工具作用：表达、组织、控制。
* 知识组织与管理工具类型：集合或线性结构工具（规范文档、停用词表、（中文）切分词典、代码表）；树型结构工具（（网络）分类表、范畴表、词族表、类目体系）；网（图）状结构工具（大型分类表、叙词表、语义网络、本体）。
* 知识组织与管理工具应用方式：全控制方式、先控制方式、后控制方式、无表应用方式。

# 文本信息检索技术与方法

* 齐普夫定律：在一个给定的文本文档集合中，如果将所有单词按照其出现的频率递减排列，并用自然数依次给单词赋予等级序号1、2、3、…，那么，单词频率与其等级序号的乘积为一个常数。  
  齐普夫定律词表编制、自动标引、倒排文档组织等方面都具有比较重要的理论指导价值。
* 文本预处理：文本词汇分析、停用词去除、词干提取（英）、名词识别。
* 布尔检索提问式的变换处理方法：  
  逆波兰变换法：符合常规的执行顺序；检索速度较快。  
  准波兰变换法：所需内存工作区个数最少的后缀表达式；检索速度较快。  
  吸取范式变换法：对内存空间的占用量达到最少vs.访问倒排文档的次数增加->增加检索次数，对结果集合的逻辑运算带来额外的负担；检索速度较快。
* 截词检索类型：后截断、前截断、中截断；有限截断、无限截断。
* （nW）：两个检索词在检索结果中出现时，相互之间的距离不超过n个单词。  
  （nN）：两个检索词在检索结果中出现时，相互之间的距离不超过n个单词，且两个词出现的先后顺序也可以变化。
* 位置检索的实现：倒排索引文档的支持、对检索提问式进行必要的变换处理。  
  对倒排索引文档的要求：全面、详细地提供每个检索词在数据库中每一层出现时的位置及其他信息（大小写标识、记录号、字段代码、段落编号、句编号、词位置编号）。
* 聚类检索的基本算法思想：1.建立检索系统数据库中文档的向量标识，并进行聚类处理，形成聚类文档；2.接受用户的检索请求，并将检索请求标识或转换为提问向量；3.确定相似度计算函数，并计算提问向量与聚类文档中各文档类向量之间的相似度；4.根据系统中预定义的相似度阈值，选择与提问向量最相似（或者相似度值大于阈值）的类向量作为下一步进行检索的文档范围；5.在选中的类向量所包含的文档集合中，逐一计算提问向量与各文档向量的相似度，并将相似度值大于指定阈值的文档作为命中结果排序输出。

# 多媒体信息检索技术与方法

* 多媒体：利用计算机及相关信息技术对多重媒体进行一体化处理并实现交互应用的技术，它具有集成性、交互性和独立性。
* 语音检索技术的研究类型：利用大词汇语音识别技术进行检索、基于子词单元进行检索、基于关键词确认进行检索、基于说话人辨认进行语音分割和索引。
* 图像信息基于内容检索的类型：示例、描绘、属性和特征说明、浏览。
* 视频信息基于内容检索的类型：基于关键帧的检索、基于运动特征的检索、基于视频语义特征的检索。

# Web与网络搜索引擎

* Web信息资源的特性：海量规模、分散无序、动态更新、种类/形式多种多样、非结构化或半结构化、语义冗余和质量控制缺乏、需求与使用方式个性化。
* 搜索引擎的发展：前身（FTP类检索工具）、早期的搜索引擎、全文搜索引擎、元搜索引擎、垂直搜索引擎、搜索代理系统。
* 搜索引擎的类型：关键词检索型vs.目录浏览型、综合型vs.垂直型、独立vs.元、学术型vs.商业型。
* 独立搜索引擎基本结构：数据采集、数据分析与标引、数据检索、信息挖掘。
* Robot工作原理：1.精选一批高质量的URLs作为初始搜索地址，并将这些地址放入袋搜索的地址列表中；2.取出待搜索地址列表中的第一项，依据并遵循HTTP协议，向有关的WWW服务器发出请求，等待并抓取到相应的HTML文件，下载到本地后进行预处理，并在该HTML文件中抽取指向其他HTML文件的超链接；3.对抽取出的超链接地址逐个进行处理，按照是否已被搜索过分别放入近期已搜索的Web站点列表和尚未搜索的地址列表中去；4.判断程序是否结束，如果没有结束，返回2继续执行。  
  注意事项：1.遍历策略的选择；2.Robot Exclusion协议的遵守；3.网页采集效率与保持WWW服务器正常工作状态的平衡机制；4.镜像站点的识别；5.动态网页的困扰；6.网页更新问题。
* PageRank算法基本思想：1.如果一个页面被多次引用，那么这个页面很可能是重要的；2.如果一个页面尽管没有被多次引用，但却被一个重要的页面引用，那么这个页面很可能是重要的；3.一个页面的重要性被均分，并传递到它所引用的页面。
* PageRank与HITS的异同：利用网页之间的相互链接关系进行递归计算，并赋予它们一定的重要性；query-independent vs. query-dependent、预先进行vs.实时进行、基于网页间的单向链接vs.基于网页间的双向链接、稳定性好vs.敏感。

# 信息检索评价研究

* 信息检索评价研究历史：  
  20c50s初-20c60s中：研究工作主要集中在对各种检索语言和标引方法的比较和评测上，初步探索、提出并形成了一套检索性能评价的指标体系，并为后来的相关研究工作奠定了方法论的基础。  
  20c60s中-20c70s中：不仅评价对象有所转移和扩展，而且在评价试验设计水平上也有了明显提高，评价指标较以往更为合理、完善，并且新增加了新颖率、费用-效益等评价指标，在评价方法上也更趋于多样化。  
  20c70s中-90s初  
  20c90s初-今（TREC时代）
* “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。  
  相关性判断标准：二值、多值、多维。
* 查全率：衡量系统在实施某一检索作业时间检出相关文献能力的一种测度指标，R=a/(a+b)。  
  查准率：衡量系统在实施某一检索作业时检索准确度的一个测度指标，P=a/(a+c)。
* 查全率和查准率的替代性计算指标：平均查全率和平均查准率、调和数F、E测度指标。
* 搜索引擎的性能评价指标：数据库规模与内容、索引方法、检索功能、检索结果处理、分类功能、用户界面、汉字处理、其他。
* TREC试验数据集合：测试文档、检索问题、检索问题的正确答案。
* TREC的检索评测项目：Cross-Language、Filtering、Interactive、High Accuracy Retrieval from Document（HARD）、Query、Questions Answering（QA）、Very Large Corpus、Web、Genomics、Robust Retrieval、Novelty、Spoken Document Retrieval Video。

# 文本聚类

* 文本聚类：根据文本文档的内容相关性来组织文档集合，将整个集合聚集成若干个类，并使得属于同一类别的文档尽量相似，属于不同类别的文档差别明显。
* 文本等级聚类算法的基本思想：1.计算文档的距离系数矩阵；2.合并两个最相似的文档类；3.更新相似矩阵；4.重复2与3，直到最后所有的文档类合并成一个类。
* 文本动态聚类方法的基本思想：1.确定聚类个数k，从文档集合中选择最初的k个文档作为凝聚点，每个凝聚点文档自成一类；2.按照距离最近原则，将剩余（n-k）个文档逐个并入最近凝聚点所代表的类，每并入一篇文档，立即重新计算该类的重心，并用此重心替代原来的凝聚点；3.以最后形成的每个凝聚点代表一类，将全部n篇文档重新聚类，逐个并入最近的凝聚点所属的类，与步骤2相同，每并入一个文档后，就重新计算重心，并以此重心代替原凝聚点，文档集合被重新聚类后，如果与原来的聚类结果不同，就重复步骤3，否则，聚类处理即告完成。  
  注意：初始聚类中心的选取、聚类参数k的选取与调整。
* 评价方法与指标：聚类错误率、聚类全面率、聚类准确率；重合率、纯度、熵值。