Ch 7

基本概念

索引

1. 信息检索中，对照或引导标引信息的排列表
    a. 主题索引：关键词索引、单元词索引、标题词索引、叙词索引
    b. 类号索引
    c. 引文索引（SCI、EI、CSSCI）
2. 将标引的结果（主题、类号）按照一定规律排列的处理技术
3. 任何检索工具都应该由二次文献部分和检索标识组成
4. 索引是一种数据结构，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

索引的存储

1. 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
2. 存储方式：顺序文档、倒排文档、后缀数组、签名文件

顺排文档索引

1. 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
2. 顺序对文档记录检索

倒排文档 倒排索引

1. 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
2. 例如：书最后的单词—页码列表，通过一些关键词，在全书中检索出与之相关的部分
3. 由两部分组成：词汇表、记录表
4. 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
5. 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
6. 使用
    a. 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
    b. 记录表检索：检索出所有找到的单词对应的记录表
    c. 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

特点

1. 快速索引（长query需要更多时间）
2. 灵活性: 不同类型的信息都可以存储在记录表中
3. 如果存储了足够多的信息，则可以支持复杂的检索操作
4. 存储开销较大
5. 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

签名文件

1. 签名文件（signature file）是基于散列（Hash）技术的面向单词的索引结构
2. 一个单词的“签名”是一个位向量
3. 一般将一个文本看作是一块
4. 签名文件索引技术只适用于小规模的文本集合

基本原理

逆波兰表达式 Reverse Polish Notation RPN 后缀表达式

1. 中缀表达式生成的逆波兰表达式是唯一的
2. 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）

文本检索技术

截词

1. 检索者将检索词在自己认为合适的地方截断
2. 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
3. 通常情况下用“*”表示无限截断，用“？”表示有限截断
4. 后截词检索：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
5. 前截词搜索：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
6. 中截词检索：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

加权检索

1. 词加权系统（term weighting system）是最常见的加权检索系统
2. 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）
3. 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
4. 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
5. 词频加权检索：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