Ch 8

基本概念
  深网
  搜索引擎
  锚文本
  入链
  出链
基本原理
  搜索引擎的分类
  基本采集过程
  爬行策略
  HITS Hyperlink-Induced Topic Search
基本流程
  PageRank
    特点

基本概念

深网

1. 暗网：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
2. 深网：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
3. 暗网是深网的一个子集

搜索引擎

1. 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

锚文本

1. 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
2. 锚文本往往比网页本身更能揭示网页的内容
3. 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

入链

1. In-link/inbound link

出链

1. Out-link/outbound link

基本原理

搜索引擎的分类

1. 检索方式：分类目录、关键词搜索引擎、混合搜索引擎
2. 信息覆盖范围及适用用户群：综合搜索引擎、专用搜索引擎（垂直搜索引擎）
3. 搜索范围：独立搜索引擎、集成搜索引擎/元搜索引擎

基本采集过程

1. 初始化采集URL种子队列
2. 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
3. 基本假设：Web的连通性很好

爬行策略

1. 定点策略
2. 定题策略
3. 广度优先：Breadth-First Search，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

4. 深度优先：Depth-First Search，深度优先搜索，简称DFS
5. 大站/要站优先
6. Partial PageRank/OPIC排序

HITS Hyperlink-Induced Topic Search

1. 每个网页计算两个值
    a. Hub：作为目录型或导航型网页的权重
    b. Authority：作为权威型网页的权重
2. 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
3. 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
4. 缺点
    a. 计算效率低，实时计算
    b. 主题漂移，扩展集可能与主题无关
    c. 易作弊
    d. 结构不稳定：删改个别少数关系，结果可能变化大
5. 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
6. 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

基本流程

PageRank

1. 拥有越多、越重要入链的页面越有价值
2. 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)
3. 网页的每条出链上每个分 量上承载了相同的PageRank分量

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）