网络信息内容获取

网络信息内容获取模型

1. 网络信息内容获取模型：
    - **信息检索**：信息需求者主动在网上搜索需要的信息；
    - **信息推荐**：网络信息服务系统从网上固定信息源或提供商获取信息，并通过固定的频道向用户发送；
    - **信息交互**：双向的信息交流；
    - **信息浏览**：相当于传统情况下的阅读、观看、倾听等行为。

搜索引擎技术

1. 中文搜索引擎的关键技术：（1）网络内容分析；（2）网络索引；（3）查询解析；（4）相关性计算。
2. 网上采集算法：自动化下载网页的计算机程序或自动化脚本，是搜索引擎的重要组成部分。
    - 通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层网络爬虫。
    - 获取原理：初始化URL集合 -> 信息获取 -> 信息解析 -> 信息判重。
    - 抓举策略：深度优先遍历、宽度优先遍历、反向链接数、Partial PageRank、OPIC、大站优先。
3. **排级算法**：PageRank与HITS。
4. PageRank：
    - **核心思想**：在互联网上，如果一个页面被很多其他页面所链接，说明它受到普遍的承认和信赖，那么它的排名就高。
    - **算法内容**：

    > 设页面集合为T，页面$t \in T$在转移过程中的PageRank计算公式为：
    >
    > $$\Pr(t) = (1-d) + d\left(\sum_{p \in T, p \neq t}(\frac{\Pr(p)}{|p|})\right)$$
    >
    > 其中$|p|$表示页面p的出度；d为影响因子，可取$d = 0.85$。

    - 优点：（1）直接高效；（2）主题集中。
    - 缺点：
        - 完全忽略页面内容，干扰挖掘结果；
        - 结果范围窄；
        - 影响因子与网页获取数量缺乏科学性。
5. HITS（Hyperlink-Induced Topic Search）：
    - 基本定义：

        - **Authority页面（权威页面）**：指与某个领域或某个话题相关的高质量网页；
        - **Hub页面（枢纽页面）**：指包含了很多指向Authority页面链接的页面。
        - **枢纽值**：所有导出链接指向页面的权威值之和。
        - **权威值**：所有导入链接所在页面的枢纽值之和。
    - 核心思想：相互增强关系。
        - 基本假设：（1）好的Authority页面会被很多好的Hub页面指向；（2）好的Hub页面会指向很多好的Authority页面。
    - 算法内容：构建根集合 -> 扩展集合Base -> 计算扩展集Base中所有页面的Hub指与Authority值 -> 排序并输出结果。
    - 优点：（1）知识范围扩大；（2）搜索时部分地考虑了页面内容，结果更具科学性。
    - 缺点：
        - **计算效率差，实时性差**；
        - **“主题漂移”**；
        - **易被作弊者操纵结果**：作弊者可以建立一个很好的Hub页面，再将这个页面链接指向作弊页面，可提高作弊页面的Authority值；
        - **结构不稳定**：在扩展集Base内，如增删个别页面或改变少数连接关系，HITS算法的排名结果就有极大改变。
6. 搜索引擎优化（Search Engine Optimization）：
    - 具有良好素养和道德观念的SEO，力图通过优化网站结构、提高页面质量等方法提高页面排名；
    - 寻找“捷径”提高页面的排名，往往是垃圾信息的制造者。

数据挖掘技术

1. 数据挖掘：通过从数据库（包括互联网上的信息内容）中抽取隐含的、未知的、具有潜在使用价值的信息的过程。
2. Web挖掘技术：（1）内容挖掘；（2）结构挖掘；（3）使用挖掘。

信息推荐技术

1. 信息推荐：
    - 组成要素：**推荐候选人、用户、推荐方法。**
    - 形式化定义：

> 设C是所有用户的集合，S是所有可以推荐给用户的商品对象的集合。效用函数$u(\cdot)$用以计算对象s对用户c的推荐度（如提供商的可靠性和商品的可取性），即：
>
> $$u: C \times S \rightarrow R$$
>
> R是一定范围内的全序的非负实数，信息推荐要研究的问题就是，对给定用户c，找到推荐度R最大的那些对象s^*，即：
>
> $$\forall c \in C,\ s^* = \argmax_{s \in S} u(c, s)$$

2. 信息推荐算法：（1）**基于内容推荐**；（2）**协同过滤推荐**；（3）**组合推荐。**
3. 基于内容推荐：根据**用户已选择的对象**，推荐其他类似属性的对象。

> **对象内容特征**（s）：以对象的文字描述为主；
>
> **用户资料模型**（c）：取决于机器学习算法。
>
> 结合对象内容特征和用户资料模型，最终的效用函数定义如下：

> $$u(c,s)=\text{score}(\text{CotentBasedProfile}(c),\text{Content}(s))$$
>
> Score的计算有不同方法，例如可以使用向量夹角余弦的距离计算方法：
>
> $$u(c,s)=\cos(\tilde{w}_c,\tilde{w}_s)=\sum_{i=1}^{k}w_{i,c}w_{i,s}/\left(\sqrt{\sum_{i=1}^{k}w_{i,c}^2}\sqrt{\sum_{i=1}^{k}w_{i,s}^2}\right)$$

4. 协同过滤推荐；推荐**相似用户**所选择的对象。
    - 基本思路：
        1. 找到与当前用户c相似的其他用户c'；
        2. 计算对象s对于用户的效用值$u(c',s)$；
        3. 利用效用值对所有对象进行（加权）排序，找到最适合c的对象s^*。
    - 分类：（1）启发式方法；（2）基于模型的方法。
5. 组合推荐：
    - **后**融合组合推荐：融合两种及以上推荐方法各自产生的推荐结果，判断哪一结果更优。**属于结果层次上的融合。**
    - **中**融合组合推荐：以一种推荐方法为框架，融合另一种推荐方法。
    - **前**融合组合推荐：直接融合各种推荐方法。

信息还原技术

1. 信息还原技术：（1）电脑还原技术；（2）网页还原技术；（3）多媒体信息还原技术。
2. 电脑还原技术：
    - 软件还原：本地还原、远程还原。
    - 硬件还原：主板集成型、独立网卡型。
3. 网页还原技术：数据包捕获技术、协议还原技术、网络内容还原技术。
4. 多媒体信息还原技术：（1）基于解码器；（2）基于封装；（3）基于远程线程注入。