WEB 应用

1. 请说明信息检索 Information Retrieval 时用户查询的主要形式有哪些？

答：（1）关键词查询：用户可以使用一组（至少一个）关键词（或者 Terms）表达他所需要的信息，目的是查找包含其中一些查询词（至少一个）或者全部查询词的文档。

（2）布尔查询：用户可以使用布尔操作符组成复杂的查询，也就是查询里包含查询词和布尔操作符。

（3）短语查询：查询式包含一些词的一个短语或句子。

（4）邻近查询：查询可以有词语和短语的组合，邻近查询查找那些包含查询词，且允许查询词之间相互有其他词的间隔文档。

（5）全文搜索：查询是一个完整文档，一般用户希望找到相似文档。

（6）自然语言查询：用户通过自然语言来表带自己的想法，然后由系统查询结果。

2.信息检索模型有哪些，简要描述其主要思想？解释向量空间模型中的 TF 及 TF-IDF 的含义。

答：问题 1：

（1）布尔模型 Boolean model：采取用户查询与文档精确匹配的想法，查询和搜索都是基于布尔代数的理论之上。文档和查询在该模型都被表示成一组词。

（2）向量空间模型 Vector space model：文档被表示成一个权值向量，其中的每一个权值都通过词频率表，或者词逆向文档频率表，或者他们的变异版本计算得到的。

（3）统计语言模型 Statistical language model：以概率及统计学为基础的一种模型。每个文档估计一个语言模型，然后基于语言模型根据查询的似然排序。

问题 2：

TF：词频率表，一种权值的取值表或方式。在这种方式中，文档 dj 中 ti 的权值就是在 dj 中 ti 出现的次数，被定义为 fij。

TF-IDF：词逆向文档频率表，一种权值的取值表或方式。将词 ti 的逆向文档频率定义为 idfi，词逆向文档频率权值为：wij=tfij*idfi。

3.简述信息检索算法的几种评估方法 Evalution Measures。

答：①平均查准率：有时我们需要一个简单的查准率去比较相同的查询下，使用不同的检索算法的效果。

②查准率－查全率曲线：根据排序中每个文档的查全率和查准率，绘制曲线：x 轴是查全率，y 轴是查准率。在同一张图表中，画不同的算法，对于相同的查询、相同文档数据集的 PR 曲线进行比较。或者利用多次查询以绘制曲线进行评估。

③排名查准率：选择一些排名位置上的文档，计算查准率。

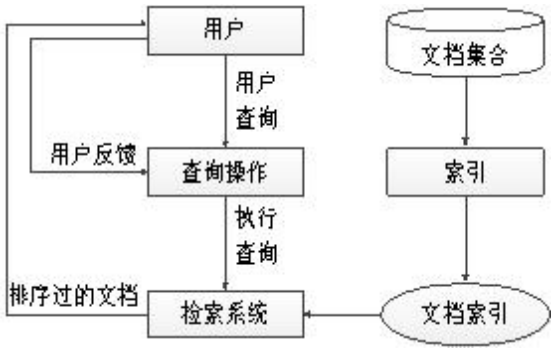
④F－score：也就是计算查全率和查准率的调和平均数。

⑤利用查全率和查准率的平衡点进行评估。

4.简要叙述 IR 系统的基本架构和工作原理。

答：信息检索（Information Retrieval，IR）是搜索的根基，其目的是帮助用户从大规模的文本文档中找到所需信息的研究领域。在用户给出一个能够描述信息

需求的查询后，信息检索系统就会从这些文档中找出和该请求相关的文档集，这也正是搜索引擎的工作原理。
IR 系统的基本架构如图：



用户通过查询操作模块发送一个查询到检索系统。检索模块使用文档索引找到包含这些查询词的文档，并且计算这些文档的相关度分数，然后根据分数给这些文档排序。进过排序的文档返回给用户。同时，文档数据集为了有效的检索已经建立了索引。查询操作需要对查询问题做预处理之后再将查询发送给检索系统，例如需要把自然语言问题转化成可执行的查询问题。索引器模块是为了更有效的查询而建立的。在搜索引擎以及大多数的 IR 系统中会使用到倒排索引。这种索引简单而且有效。检索系统会为每个索引文件计算与查询的相关度分数。文档会根据它们的相关度分数排序来进行反馈。

5.简述关联反馈的目的。简述对 Rocchio 方法中对查询 q 的更新公式的含义。

答：目的：加强检索的性能

更新含义：在 q 的更新公式中简单的加上相关文档的词，减去仅仅出现在非相关文档的词与不具有区分度的词。

1.文本和网页需要做哪些预处理。

答：①传统的文本文档(不含有 HTML 标签)，预处理步骤包括：无用词移出，词干提取，处理数字、连接词、标点以及字母大小写。

②对于网页，在传统的文本文档预处理基础上需要加上：HTML 标签移除，鉴定主要内容块，辨别不同字段，辨别锚文本。

2.简要解释倒排索引 Inverted Index 和描述索引建立的大致过程。

答：①简要解释倒排索引：如果在检索之前，已经有一个数据结构存储了包含每个检索词的对应文档集合，那么检索效率将会急剧提高。因此倒排索引应运而生。倒排索引是现代信息检索系统的核心部分，其组织方式和存储结构对信息检索系统的性能有很大的影响。倒排索引主要由词典和倒排链两个部分组成。词典记录了需要被检索的所有词条项和对应倒排链指针。对于一个查询词条项，查找其是否出现在词典中，如果找到就可以直接获取到倒排链指针，也就是就直接快速获取到了文档集合。为了满足快速索引构建和词项查找需求，词典本身通常是利用 Hash 表或者二叉搜索树形结构实现；出现在词典中的每个词项都对应一个倒排链，最简单的倒排链存储了包含该此项的所有文档 ID 列表。

②索引的建立：在这里我们以使用 Tire 数据结构为例进行建立索引，算法顺序的扫描每个文档中每个词，并且查找与 Tire 中的现有词相同的词。如果发现有相同的词，那么文档的 ID 和其他的信息(例如：词在文档中的偏移量)将被加到词的倒排列表中；如果没有发现相同的词汇，那么将生成一个新的叶子结点，用来代表词。若直至主存已满整个索引不能全部存在主存中，我们把主存中现有的部分索引存到硬盘中，然后在主存中建立余下文档的部分索引。

网络作弊（Spamming）-使用人为的手段，让一些网页高于其应有的排名
作弊技术是指不增加一个网页的信息价值，而通过误导搜索引擎提高网页排名的手段
搜索引擎优化-SEO
网络作弊分为内容作弊（content）和链接作弊(LINK)

内容作弊又称为词组作弊（term）-词组作弊可以出现在任何文本域中：标题 title 元标记 meta-tags 正文 body 锚文本 anchor text 网址 url
两种词作弊技术：重复一些重要词； 大量添加其他不相关词

链接作弊：
链出链接作弊 Out-link
链入链接作弊 in-link-作弊器通常使用的一些技术：创建蜜罐；在网络目录中添加链接；在用户生成内容中添加链接；交换链接；自发添加

3.什么是社会网络分析。Link analysis
答：社会网络分析是对社会网络的关系结构及其属性加以分析的一套规范和方法。
社会网络是一门研究社会中社会实体(组织中的人、或者叫参与者)以及他们之间的活动与关系的学问。而实际上 WEB 就是一个虚拟的社会，在这个虚拟的社会关系中，每张网页可以被看作是一个参与者，而每个超链接这可以被看作是一个关系。许多社会网络研究所得出的结果可以被延伸和利用到 web 范畴中。

4.何谓中心性 centrality？中心性度量的三种主要方式及度量标准分别是什么？何谓权威性 prestige？度量权威性的三种方法是什么？如何度量的？
答：①中心性：用来度量结点在网络中的重要性。对于单个结点或由多个结点组成的群体都可以定义中心性。对于社会网络参与者的著名程度进行度量的标准。
②.1 度中心性：中心参与者是拥有与其他参与者的链接或者链接数目最多，最活跃的参与者。
②.2 接近中心性：主要基于接近度或距离，即它到其他所有参与者的距离要足够短。
②.3 中介中心性：中介性用来度量 i 对于其他节点对的控制能力。也就是其处在非常多结点的交互路径上。
③权威性：被定义为大量链接指向的参与者，也就是说计算一个参与者的权威，只看指向该参与者的链接(链入链接)

④.1 度权威 ④.2 邻近权威 proximity ④.3 等级权威 rank
⑤.1 考察其入度数量
⑤.2 考虑每一个能到达其的另一参与者，也就是到目标参与者的有向路径
⑤.3 在前基础上考虑某些拥有选择权和投票权的参与者的重要性

5.什么是同引分析 co-citation？什么是引文耦合 bibliographic coupling？可能的应用有哪些？

答：①同引分析是被用来度量两篇论文（出版物）之间的相似性。如果论文 i 与论文 j 都被同一论文引用，就说两个论文之间有某种程度上的关系。

②采用相近原则，研究问题方向与同引分析是镜像关系，其将引用同一篇其他文章的两篇论文联系起来。

③应用于学术出版界领域，通过研究引用以期找出作者与它们著作之间关系的计量研究领域，分析的结果有时可以被用来衡量学术作品的权威性与科学性。

PageRank 算法依赖于 Web 的自然特性，它利用 Web 的庞大链接结构来作为单个网页质量的参考。本质上，PageRank 算法将网页 x 指向网页 y 的链接当作是一种投票行为，由网页 x 投给网页 y。然而，PageRank 算法并不只是仅仅考虑网页的得票数，也就是指向该网页的链接数。它也会分析那些投票的网站。那些重要网站投出的选票使得接收这些选票的网页更加重要。这刚好就是社会网络中所提到的**等级权威**（**Rank Prestige**）概念（参见 7.1.2 节）。

6.PAGERANK 算法的基本原理是什么？试用实例说明算法的基本思想，并分析其优点及缺陷。指出可能的改进有哪些？

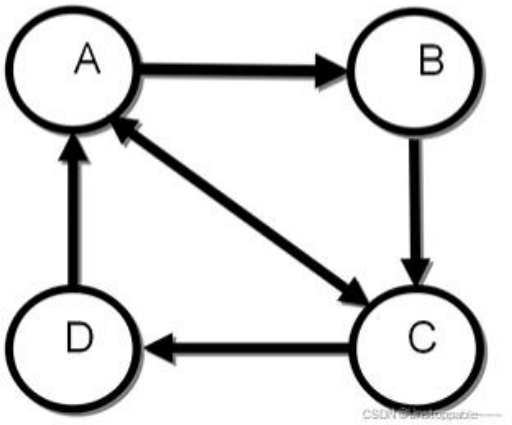
答：①PageRank 算法的基本想法是在有向图上定义一个随机游走模型，即一阶马尔可夫链，描述随机游走者沿着有向图随机访问各个结点的行为。在一定条件下，极限情况访问每个结点的概率收敛到平稳分布，这时各个结点的平稳概率值就是其 PageRank 值，表示结点的重要度。PageRank 是递归定义的，PageRank 的计算可以通过迭代算法进行。

②公式定义

$$PR(a)_{i+1}=\sum_{i=0}^{n}\frac{PR(Ti)_i}{L(Ti)}$$

PR（a）表示当前节点 a 的 PR 值
PR（Ti）表示其他各个节点（能够指向 a）的 PR 值
L（Ti）表示其他各个节点（能够指向 a）的出链数
i 代表当前时刻或迭代次数

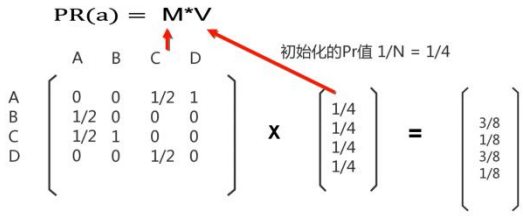
接下来以下图为例进行计算演示：



将四个节点的初始 PR 都设置为 1/4
根据每一个节点（a）的入链节点(Ti)的 PR 值及出链数和自身(a)的 PR 值
不断进行迭代，直到 PR 值不再发生变化
以 A 为例：
A 有两个入链节点 C（出链数为 1，PR=1/4）和 D（出链数为 2，PR=1/4）由计算公式得到：
i=1 时刻的 PR（A） = (1/4)/1 + (1/4)/2 = 3/8

PR 值 / 循环次数 i	PR（A）	PR（B）	PR（C）	PR（D）
i = 0 PR值初始化 = 1/N	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
i=1	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
排名	1	2	1	2

矩阵化计算
借助邻接矩阵（转移矩阵）的表示方式，我们可以简化上述计算，将四个节点的 PR 值转化为 V 向量，并于转移矩阵相乘，可以得到新一轮的 PR 值向量



由此可以得到每一步 PR 值迭代的结果为：MV， MMV， MMM*V 最终会收敛为 M‘ * V

③优点：防止作弊；从全局出发的度量以及其非查询相关的特性。
缺点：不能分辨网页在广泛意义上是权威的还是仅仅在特定的查询话题上是权威的；其没有考虑时间。

④针对传统 PageRank 算法迭代过程复杂、时效性不强、执行速度慢等缺点，我们可以进行优化迭代过程、增加时间因子影响函数、并行化三点改进方向入手进行改进。

重点

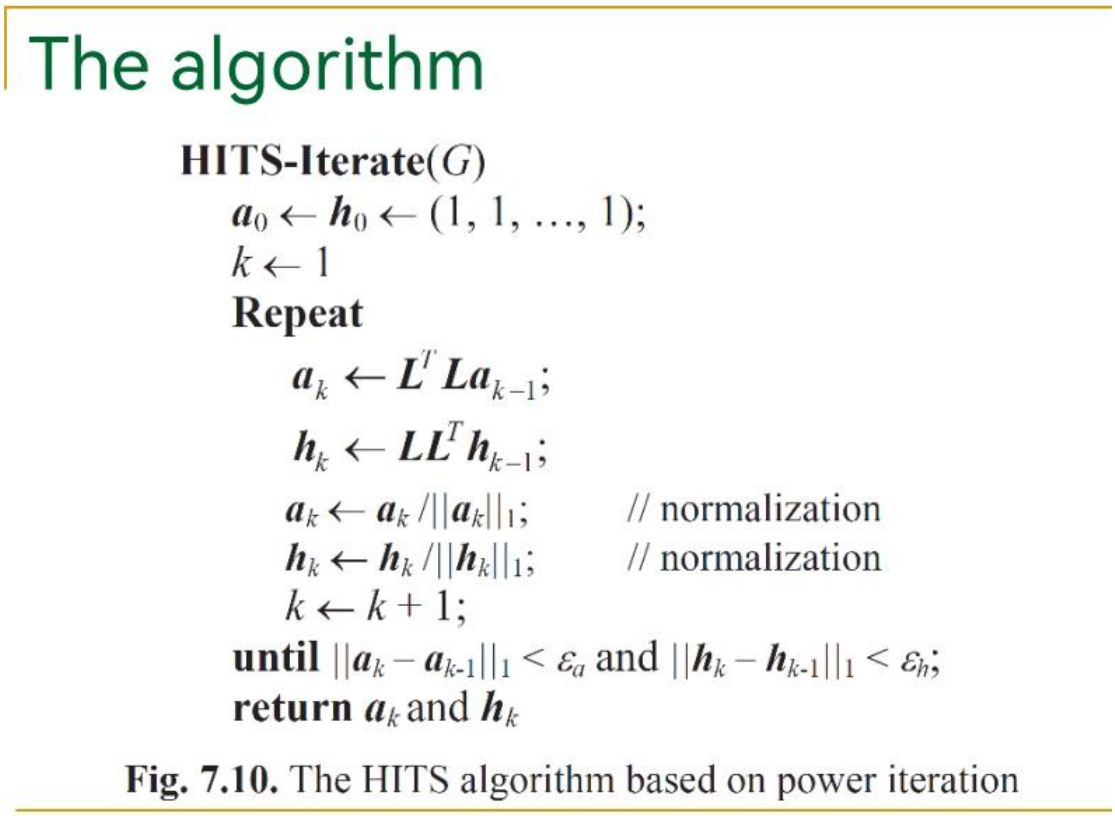


Fig. 7.10. The HITS algorithm based on power iteration

Normalization-数据归一化 方便后续数据处理

1、HITS 算法的基本思想是什么，以矩阵方式写出节点中心性和权威性的迭代方程？与 PAGERANK 算法相比它们的主要异同是什么？HITS 算法的优缺点是什么？

答：①基本思想：HITS 是查询相关，当有查询请求时，其首先展开一个由搜索引擎返回的相关网页列表，给出了权威等级和中心等级两个评级，一个优秀的中心页必然会指向很多优秀的权威页，一个优秀的权威页必然会被很多优秀的中心页指向，二者是一种相互促进的关系。

中心性、权威性迭代方程：

用 a 表示所有权威值的列向量， a=(a(1), a(2),…,a(m)) T

用 h 表示所有中心分值的列向量， h=(h(1), h(2),…,h(m)) T,于是有

a=L T h

h=L a

计算权威值和中心值与计算 PageRank 算法的重要性类似，需要迭代计算：

$$a_k = L^T L a_{k-1}$$

$$h_k = L L^T a_{k-1}$$

其中， a 0= h 0=(1, 1, … ,1)

为了让 a 和 h 中的数值不至于太大，每次迭代之后，都对它们进行归一化处理，使得：

$$\sum_{i=1}^{m} a(i) = 1$$

$$\sum_{i=1}^{m} h(i) = 1$$

注意，选择不同 a 0 和 h 0，最终可能会收敛到不同的向量。

②主要异同：HITS 算法与 PageRank 算法最大的区别是，PageRank 算法是与查询无关的全局算法，PageRank 算法是基于以下两个假设的。数量假设：在 Web 图模型中，如果一个页面节点接收到的其他网页指向的入链数量越多，那么这个页面越重要。质量假设： 指向页面 A 的入链质量不同，质量高的页面会通过链接向其他页面传递更多的权重。所以越是质量高的页面指向页面 A， 则页面 A 越重要。而 HITS 算法与用户输入的查询词是密切相关的，HITS 算法接收到用户查询之后，将查询词提交给搜索引擎，返回的搜索结果中， 提取排名靠前的网页，得到一组与用户查询高度相关的初始网页集合，这个集合被称为根集。 HIST 算法对根集中的网页进行扩充，扩充的原则：凡是与根集内的网页有直接链接指向关系的网页都被扩充进来。

③优点：迭代次数少，收敛速度快，简单且效率高

缺点：主题漂移，易被作弊者操作结果

2、何谓 WEB 爬虫？其主要目的和作用是什么？

答：①：也称蜘蛛或机器人，是能够自动下载网页的程序。

②：爬虫可以将多个站点的信息收集起来，并通过在线（网页被下载后）的或离线的（网页被存储后）的方式，集中进行进一步的分析和挖掘。

3、爬虫的主要分类是什么？请简要说明之。

答:

①通用网络爬虫：门户站点搜索引擎、大型 Web 服务提供商采集数据 爬行范围和数量巨大、爬行页面顺序要求低、并行工作方式，爬取互联网上的所有数

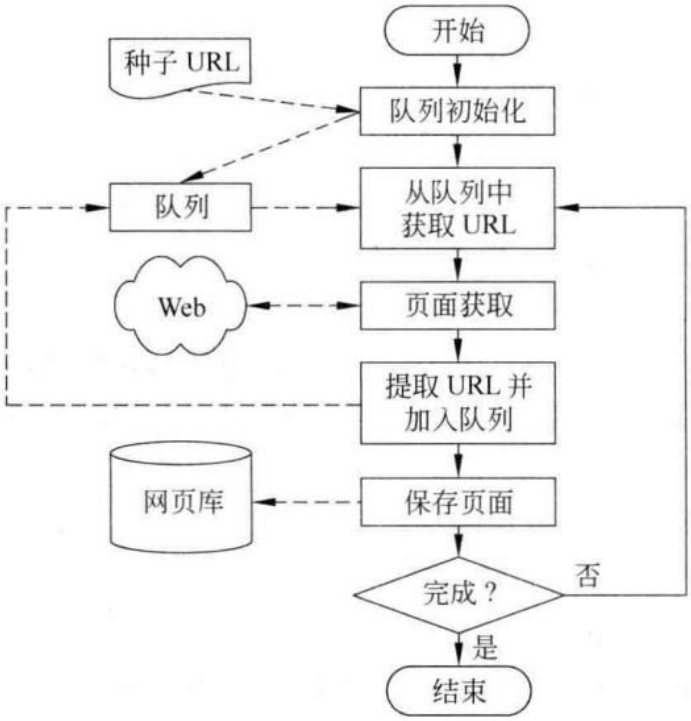
②主题网络爬虫：只爬行特定的数据，节省了硬件和网络资源，页面更新快

③限定网络爬虫：当我们并不打算爬取整个网页，而只是打算爬取某些特定类别的网页能其够爬取用户感兴趣的某一类网页

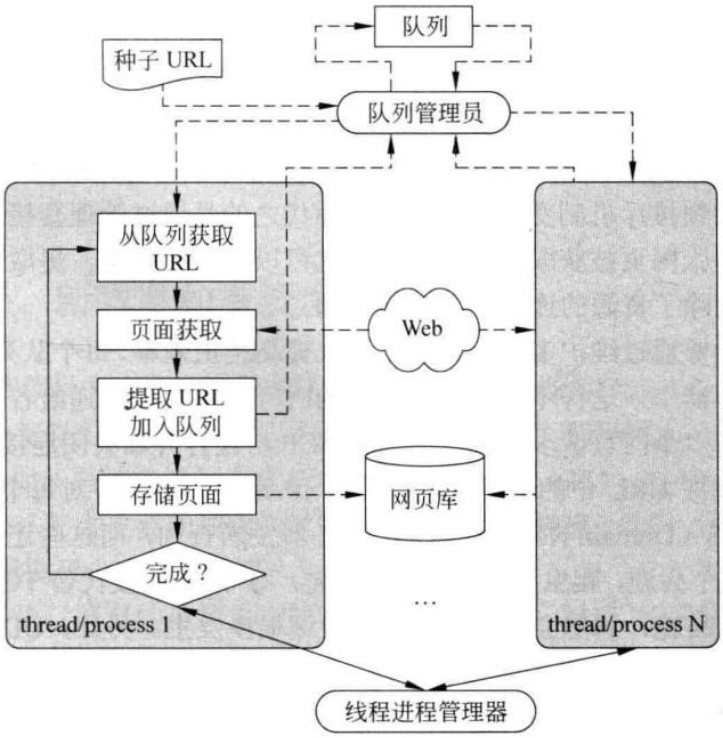
Web Crawling

1、试分析简单序列爬虫、并发爬虫、Simple sequence crawler, concurrent crawler 通用爬虫 universal 及主题爬虫的联系与区别。并分别给出其基本算法框架。

答：①简单序列爬虫，每次只获取一张网页，并不考虑充分利用它的资源。

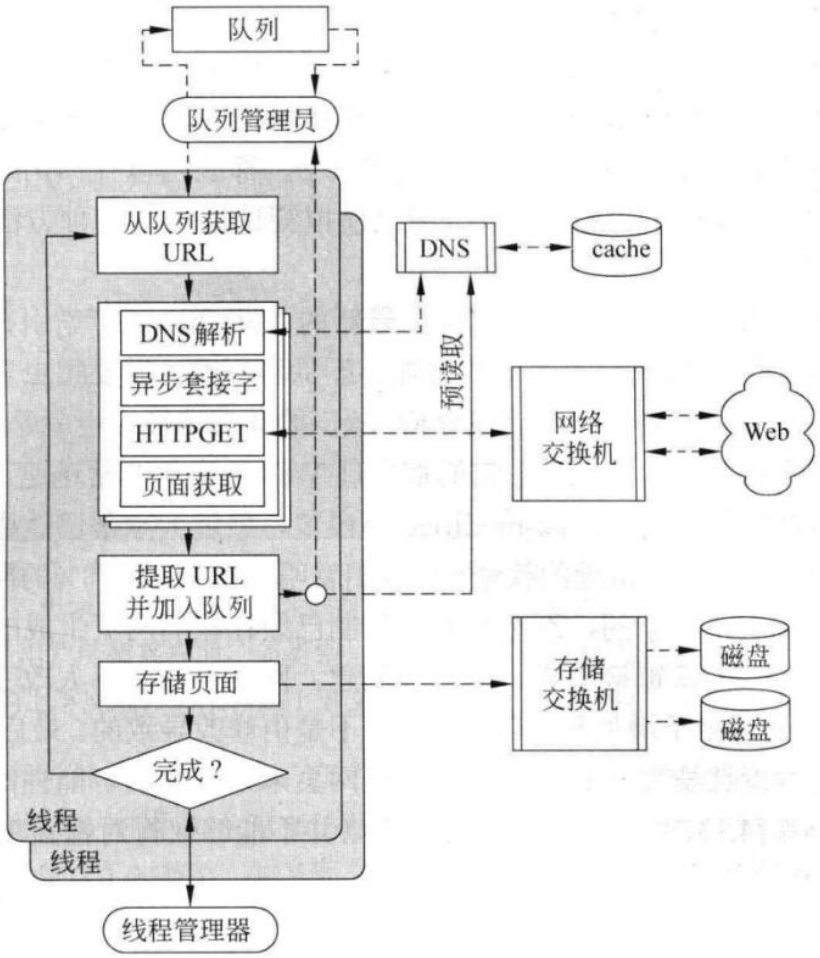


②并发爬虫，采用多线程并发操作执行程序可以大大降低运行时间，提高效率。但处理空队列时要比序列爬虫更复杂。

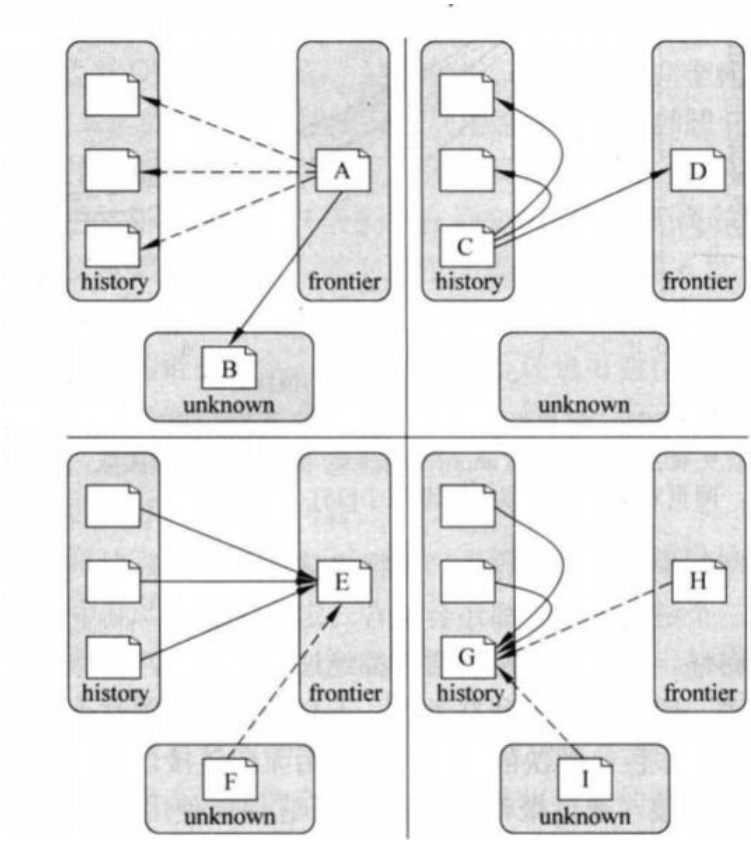


③通用网络爬虫，门户站点搜索引擎、大型 Web 服务提供商采集数据，并行工作方式，爬取互联网上的所有数。与并发爬虫的区别主要集中在：性能(需

要可以美妙处理成千上万网页抓取工作规模)和策略(能够尽量覆盖尽可能多的重要网页，且尽可能维持最新状态)上。



④主题网络爬虫：只爬行特定的数据，节省了硬件和网络资源，页面更新快。



2、网页爬取数据的主要预处理包括哪些步骤？何谓网页解析 parsing？主要包括哪些内容？

答：①预处理：网页获取并下载，网页解析，删除无用词并提取词干，链接提取和规范化等。

②.1 网页解析：爬虫对网页内容进行解析，且可以进一步将解析出的内容用来支持调用爬虫的主程序和保持爬虫不断运行。

②.2 解析内容：可以是只从超链接提取 URL，也可以分析 HTML 代码，以及一些开放或私有的格式。

3、对爬虫爬取的网页主要的组织方式有哪些形式？各有何利弊？

答：①线性结构 这是网站最简单的一种结构，它是以某种顺序组织的，可以是时间顺序，也可以是逻辑甚至是字母顺序。通过这些顺序呈线性地链接。如一般的索引就采用线性结构。线性结构是组织网页的基本结构，复杂的结构也可以看成是由线性结构组成的。

②二维表结构 这种结构允许用户横向、纵向地浏览信息。它就好象一个二维表，如看课表一样。

③等级结构 等级结构由一条等级主线构成索引，每一个等级点又由一条线性结构构成。如网站导航等就是这种结构。在构造等级之前，你必须完全彻底的理解你的网站内容，避免线性组织不严的错误，不方便浏览者。

④网状结构 这是最复杂的组织结构，它完全没有限制，网页组织自由链接。这种结构允许访问者从一个信息栏目跳到另一个栏目去，其目的就是充分利用网络资源和充分享受超级链

4、衡量爬虫质量的标准有哪些常用的方式？

答: ①需要在有限的资源内获取最想要的网页。这就需要考虑网页的重要性与查准率、查全率，更为准确地抓取想要的网页。

②所抓取网页的时效性。因为对于下载到本地的网页，可能源网页已经发生更新，则为确保所抓取网页的有效性，需要尽可能地保证网页的时新性。

③在上面两者的基础上，力求使得抓取的网页更加广。

5、何谓爬虫道德？如何解决爬虫冲突？Crawler ethics and conflicts

答：①爬虫道德问题主要聚焦于按授权情况，网络爬虫可以分为「合法爬虫和恶意爬虫」，前者以符合 Robots 协议规范的行为爬取网页，或爬取网络公开接口、授权接口进行爬取。后者恶意爬虫则是通过分析并自行构造参数对非公开接口进行数据爬取或提交，获取对方本不愿意被大量获取的数据。

②遵守相关协议，且注意：考量被爬企业对于数据信息的保护措施，避免破解或规避被爬企业为保护数据而采取的加密算法、技术保护措施；考量被爬企业设定的获取数据的措施（如需要实名认证、账号密码登录、内部权限），避免伪造实名认证或窃取账号密码、内部权限的形式获取被爬企业的数据；考量被爬企业向一般用户提供数据的方式和业务模式，避免将大量获取（包括利用技术措施模拟正常用户大量获取数据的形式）的数据应用于与被爬企业相竞争的业务；考量抓取的数据信息的属性，避免抓取身份认证信息。

6、结构化数据抽取 structured data extraction 的主要方法有什么？

答：主要采取三种方法：

①手工方法；②包装器归纳；③自动抽取。

	手工方法	包装器归纳	自动抽取
优点	1. 对于任何一个网页都是通用的，简单快捷； 2. 能抽取到用户感兴趣的数据。	1. 需要人工标注数据集； 2. 能够抽取到用户感兴趣的数据； 3. 可以运用到大规模网站的信息抽取	1. 无监督的方法，无需人工进行数据的标准； 2. 可以运用到大规模网站的信息抽取。
缺点	1. 需要对网页数据进行标注，耗费大量的人力； 2. 维护成本高； 3. 无法处理大量站点的情况。	1. 可维护性比较差； 2. 需要投入大量的人力去做标注。	1. 需要相似的网页作为输入； 2. 抽取的内容可能达不到预测，会抽取一些无关信息。

7、包装器归纳方法的主要原理是什么？

答：包装器定义：包装器是一个能够将数据从 HTML 网页中抽取出来，并且将他们还原为结构化的数据的软件程序。

包装器归纳是基于有监督学习的，他从标注好的训练样例集合中学习数据抽取规则，用于从其他相同标记或相同网页模板抽取目标数据。

8、何谓列表页？详情页？

答：①列表页:每个页面都包含几个列表对象

②详情页:网页侧重于一个对象

9、 在结构化数据抽取时，什么是地标？主要的通配符有哪些？

答： ①每一个地标都是一个连续的标志序列，并且被用于定位一个目标项的开头或结尾。

②所学课本中：***_Numeric_***、***_AlphaNum_***、***_Alphabetic_***、***_Capitalized_***、***_AllCaps_***、***_Punctuation_***、***_HtmlTag_*** 为常见通配符。

10、何谓提纯？ 主要的提纯方法有哪些？具体如何实现提纯？

答: ①提纯:通过添加更多的终结符来特化一个析取项

②方法与实现:

地标提纯:通过在一个地标 **li** 的开头或结尾连接一个终结符加长 **li.**

拓扑提纯:通过添加 **1**-终结符增加地标数量,也就是说 **t** 和与之匹配的通配符.

1、在结构化数据抽取时，主动学习的基本原理是什么？什么是协同测试？

答：1 学习是一种帮助自动识别提供信息的未标注样例的方法。

在包装器学习中，该方法按照以下方式进行工作：

1)从未标注的样例集合中随机选取一个较小的未标注样例的子集；

2)手工标注仔集中的样例，并另原先的样例集合为之前的补集；

3)基于标注集学习一个包装器计做 w；

4)将 w 应用于样例集找到一个提供信息样例的集合；

5)该集合等于空集则终止，否则继续到第二步骤。

2 协同测试利用通常可以用多种方式抽取同一个数据项这一事实,使系统可以学习不同的规则，如：向前或后项规则来定位同一个数据项。

2、什么是包装器维护？主要包括什么内容

答：1 对于在包装器生成之后，因需应用于训练样例含有类似的数据并且格式相同的其他网页，而带来的新问题进行维护与处理应对。

2 主要包含包装器验证问题和包装器修复问题。

3、什么是完美析取规则？

答：在序列覆盖中，算法 LearnRule() 中的输出项。特征为：包含尽可能多的正数据项，并且不涵盖所有负数据项。

4、什么是串的编辑距离？什么是树的编辑距离？

答：1 字符串的编辑距离是一种使用最为广泛的字符串匹配和比较技术。

2 两棵树之间的编辑距离是指将一棵树变成另一棵树所需要的最小操作及对应的代价。

例 1：我们想计算下列两个字符串的编辑距离并找到对齐情况：

s_1:　　X G Y X Y X Y X

s_2:　　X Y X Y X Y T X ■

图 9.17 给出了编辑距离矩阵。最终的编辑距离值是 2，也就是右下角单元中的值。图 9.17 还展示了回溯路径。注意一条对角线意味着匹配或者变化，一条垂直线意味着插入，而一条水平线则意味着删除。于是，我们的两个字符串的最终对齐情况为：

s_1:　X G Y X Y X Y _ X

s_2:　X _ Y X Y X Y T X

	s_1	X	G	Y	X	Y	X	Y	X
s_2	0	1	2	3	4	5	6	7	8
X	1	0	1	2	3	4	5	6	7
Y	2	1	1	1	2	3	4	5	6
X	3	2	2	2	1	2	3	4	5
Y	4	3	3	2	2	1	2	3	4
X	5	4	4	3	2	2	1	2	3
Y	6	5	5	4	3	2	2	1	2
T	7	6	6	5	4	3	3	2	2
X	8	7	7	6	5	4	3	3	2

图 9.17　编辑距离矩阵和回溯路径

这个算法的时间复杂度为 $O(|s_1||s_2|)$（以填充矩阵），空间复杂度也是 $O(|s_1||s_2|)$。回溯耗时为 $O(|s_1|+|s_2|)$。

归一化编辑距离（**Normalized Edit Distance**, $ND(s_1, s_2)$）定义为编辑距离除以两个字符串的平均长度：

$$ND(s_1,s_2)=\frac{d(s_1,s_2)}{(|s_1|+|s_2|)/2} \qquad (1)$$

另一个常用的除数是 $max(|s_1|, |s_2|)$。

最后，在数据抽取中，可能不希望“变换一个字符”（这意味着正则表达式中的一个析取式）。一个大的距离可能被用于避免它。我们将在 9.11.2 节中再来讨论这个问题。

5、试用课堂上讲的矩阵算法求解下面两个串的编辑距离 string edit distance。S1=XGYXYYX，S2=XYXYXTX

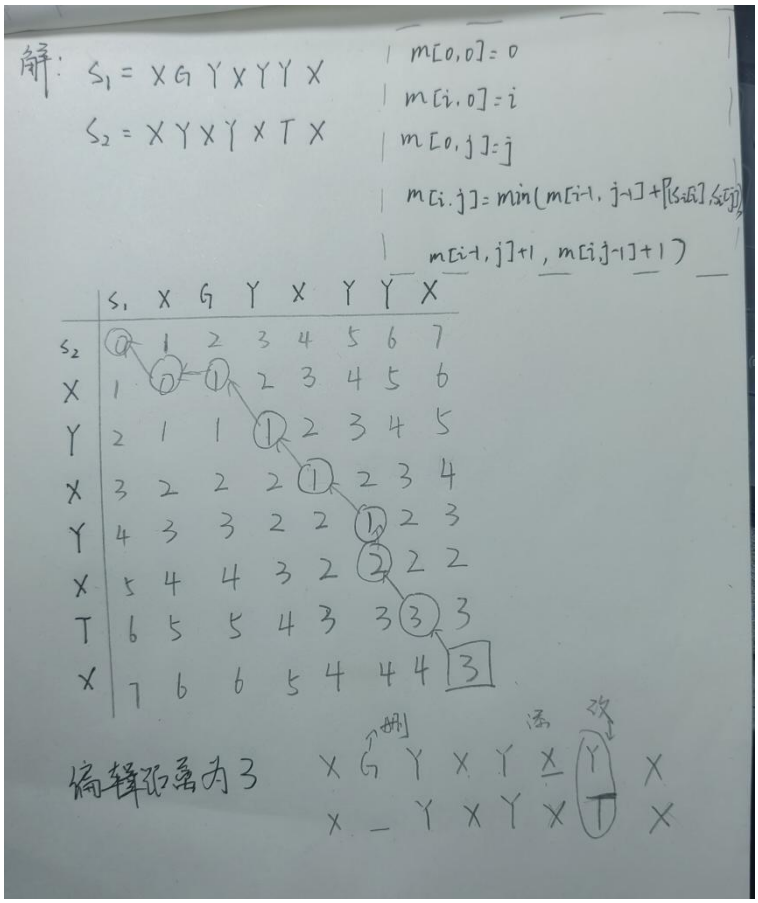
假设给我们两个字符串 s_1 和 s_2，下列递推关系定义了 s_1 和 s_2 的编辑距离 $d(s_1, s_2)$：

$$d(\varepsilon, \varepsilon) = 0 \qquad //\varepsilon 表示一个空字符串$$

$$d(s, \varepsilon) = d(\varepsilon, s) = |s| \qquad //|s|是字符串 s 的长度$$

$$d(s_{1-}+c_1, s_{2-}+c_2) = \min(d(s_{1-}, s_{2-}) + p(c_1, c_2), d(s_{1-}+c_1, s_{2-}) + 1, d(s_{1-}, s_{2-}+c_2) + 1)$$

这里 c_1 和 c_2 分别是 s_1（$= s_{1-}+c_1$）和 s_2（$= s_{2-}+c_2$）的最后一个字符，并且如果 $c_1 = c_2$，则 $p(c_1, c_2) = 0$；否则 $p(c_1, c_2) = 1$。



6、STM 算法的主要原理是什么？试给出算法描述。

答：1 STM 树的匹配定义为:

对于映射 M，任意的结点对(i , j) ，(i , j) ∈ M (i,j 是非根结点)，(parent(i),parent(j))∈M 成立，则结点对(i , j) 匹配.

最大匹配是指匹配到的结点对数目最多的匹配.

2 算法流程：

① BEGIN

② IF 树 A 和 B 的根结点的标记不同

③ RETURN 0;

④ ELSE

⑤ 令 m 为树 A 第 1 层子树的数目;
令 n 为树 B 第 1 层子树的数目;
Initialize $M[i,0]:=0, i=0,\cdots,m$;
$M[0,j]:=0, j=0,\cdots,n$;

⑥ FOR $i=1:m$

⑦ FOR $j=1:n$

⑧ $M[i,j]=\max(M[i,j-1], M[i-1,j], M[i-1,j-1]+W[i,j])$;
其中 $W[i,j]=SimpleTreeMatching(A_i,B_j)$;

⑨ ENDFOR;

⑩ ENDFOR;

⑪ RETURN $(M[m,n]+1)$;

⑫ END

2. 何谓多重对齐？给给出中星算法描述，并用中星算法求解 S={ABCD，XBCD，XABC，YBCD}。

答：1－多重对齐：由于一张网页通常包含两个以上的数据记录，所以需要匹配两个以上的字符串和树。因此，产生一个对所有字符串或树的全局的对齐是至关重要的，而这项任务与操作被称作多重对齐。

2－算法描述：

第一步：选择 Sc 为到其他所有的剩余的字符串编辑距离之和最小的字符串，把它当作一个代表。称作中心字符串 Sc

第二步：定义 M，用于存放对齐后的字符串（此时 M 中只存放了 Sc，

中会出现的情况)除 s'和 c*'，剩下字符串要依据 c*到 c*'的对齐操作进行相同的操作保证同步对齐；

}

CenterStar(S)

1. choose the center star s_c using Equation (3);
2. initialize the multiple sequence alignment M that contains only s_c;
4. **for** each s in S-$\{s_c\}$ **do**
5. let c^* be the aligned version of s_c in M;
6. let s' and $c^{*\prime}$ be the optimally aligned strings of s and c^*;
7. add aligned strings s' and $c^{*\prime}$ into the multiple alignment M;
8. add spaces to each string in M, except, s' and $c^{*\prime}$, at locations where new spaces are added to c^*
9. **endfor**
10. **return** multiple string alignment M

3－解：

$\Sigma d_{ABCD} = 1+2+2 = 5$

$\Sigma d_{XBCD} = 1+2+2 = 5$ ∴选取最小.

$\Sigma d_{XABC} = 2+2+4 = 8$ 弃 ABCD 作为 S_C

$\Sigma d_{YBCD} = 2+2+4 = 8$

循环1: $C^* = S_C$, $S = XBCD$ C^*: A B C D S': X B C D	变化过程: M ABCD ABCD XBCD
2: C^* , $S = XABC$ C^*: − A B C D S': X A B C −	−ABCD −XBCD XABC_
3: C^* , $S = YBCD$ C^*: − A B C D S': Y − B C D	−ABCD −XBCD XABC_ Y_BCD

Information integration 信息集成

1、什么是模式匹配 schema matching？主要的匹配形式有哪些？

答：①模式匹配是指对于两个或更多数据库的模式，当需要在这些模式之间产生映射，就需要把具有相同语义的元素(或属性)映射到一起，这样便可以把多个模式整合为一张全局的统一的模式。

②根据输入信息不同，匹配分为以下类型：

②.1 模式层 Schema-level 的匹配

②.2 域和实例层的匹配 Domain and instance-level

②.3 模式、域和实例的综合匹配

2、模式匹配时的数据预处理包括哪些方面？

答：①预处理 1-分词 Tokenization(断词)：会对模式元素或者属性值这样的项进行分词处理

②预处理 2-扩展 Expansion：把词汇的缩写形式扩展为它们的原型

③预处理 3-移除停用词和词干提取 Stopword removal and stemming

④预处理 4-单词的标准化 Standardization of words：将一个词的不同拼写形式将转化为唯一的标准形式

3、什么是简单领域？Simple domain 什么是复合领域？Composite domain

答：①简单域是指这个值域中的实例值都是单一成分，也就是非合成的，且简单域可以是任何类型的.

②一个 k 元的复合域 d 是一个有序的 k 元组，其中第 i 部分是 d 的第 i 个子域的值，我们用 di 来表示，每个 di 都是一个简单域。一个复合域通常可以通过它包含多种形式分隔符的实例值体现出来。111

4、构建全局搜索界面时 Global Query Interface ，需要满足哪些条件？分别是什么？

答：①结构恰当 Structural appropriateness：全局界面中的属性应也满足源个体界面中一些属性关系的限制，这些限制指引合并算法生成全局界面，每个聚类对应全局界面的一个属性。

②词汇恰当 Lexical(词汇) appropriateness：界面合并之后，集成界面的属性需要被赋予标签，而元素的标签必须精心选择，使之可以准确表达每个元素的含义。

③实例恰当 Instance appropriateness：需要确定全局属性的值域类型，而每个全局界面属性的值域应该涵盖所有源查询界面所对应的属性值，一般我们使用兼容性规则确定其类型。

A unified query interface:

Conciseness 简洁

Completeness 完整性

User-friendliness 用户友好性

5、什么是观点挖掘？Opinion mining

答：观点挖掘其目标是让计算机在语义理解的基础上，从文本中获取有价值的评价信息和观点。广义的观点挖掘也称情感分类、情感分析或文本意见挖掘等。观点由主题（Topic）、发出者（Holder）、断言（Claim）和情感（Sentiment）组成，观点挖掘就是要通过计算机系统，找到文本中的这四种要素，并分析出它们之间的相互联系。

观点挖掘是多学科的综合性研究，涉及文本挖掘、信息抽取、信息检索、机器学习、自然语言处理、概率论、统计分析、本体、可视化技术等方面。按照处理文本的粒度不同，可以分为词语、句子和篇章三个级别的研究；按照处理文本的类别不同，可以分为基于产品评论和基于新闻评论的观点挖掘。

1、文档层次、句子层次，以及特称层次的观点挖掘的主要任务分别是什么？

答：①文档层次：把含有观点的文档整体作为基本信息单元，进行正面、负面或其他情感的分类；

②句子层次：首先对句子进行主观性分类，确定其是主观句还是客观句，接着对得到的主观句进行正面或负面观点的分类；

③特称层次：首先进行方面或特征的抽取，抽取过程中一定要明确方面具体属于哪个实体，再基于不同方面的观点进行正面、负面或中立的分类。

2、什么是比较关系挖掘？比较性句子的比较类型有哪些？比较关系挖掘的基本任务是什么？

答：①比较关系挖掘：利用一个对象和其他相似对象进行比较的形式进行观点挖掘；

②比较性句子的比较类型：可以被分为四种主要类型，前三种类型是等级比较，最后一种是非等级比较：

②．1 不相等的等级比较

②．2 相等的比较

②．3 最高级比较

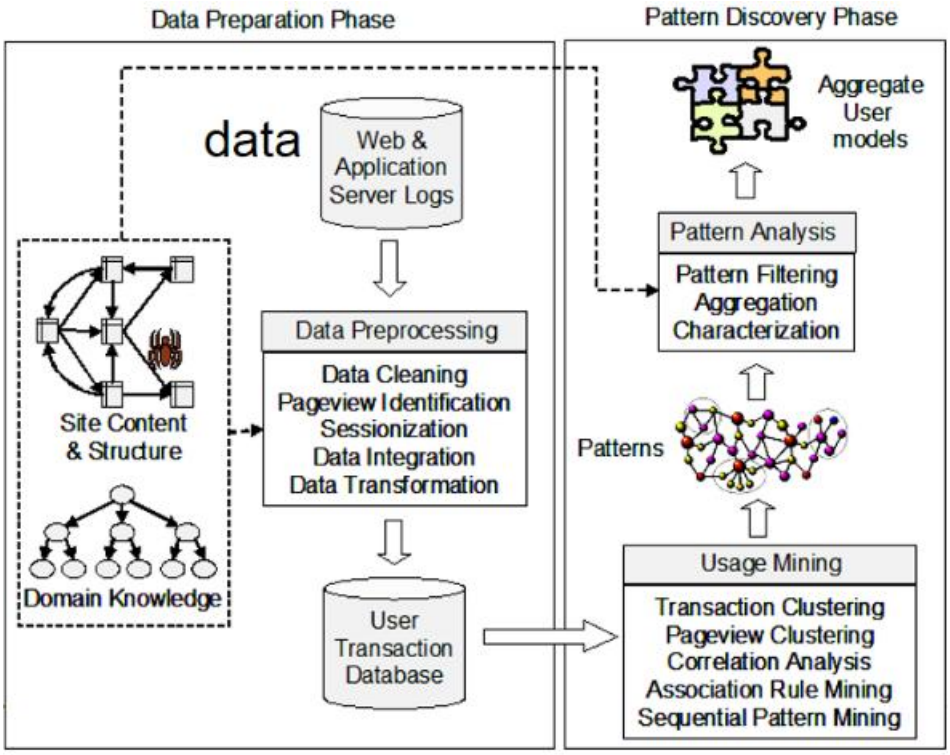
②．4 非等级比较用于比较多个实体方面之间的关系，但并不对他们进行分级；

③基本任务：首先进行等级比较性语句的识别，在进行偏好实体识别，确定一个比较性句子是否带有观点，最后可根据比较实体的共有方面进行排序来比较多个实体，从而给出比较性的观点。

1、什么是 WEB 用法挖掘？其基本流程框架是什么？用法挖掘的主要数据源包括哪些？

答：①定义：WEB 使用挖掘是指自动发现和分析（来自于收集的点击流、用户事务数据和因用户与一个或多个网站交互 WEB 资源所产生或收集其他数据构成的）模式，去捕捉、建模并分析用户与网站交互的行为模式和整体情况。

②基本流程框架：



WEB 使用挖掘可分为三个相互依赖的阶段：数据收集及预处理、模式发现和模式分析。

在预处理阶段，点击流数据被整理并分割成一组用户事务集合，用来表示每个用户对站点的不同访问。其他知识来源也可能被用于数据的预处理或用户事务数据的补充，这些知识来源包括网站的内容或结构，以及来自网站本体的语义领域知识（例如产品目录或概念层次）。

在模式发现阶段，统计学、数据库以及机器学习的方法被用来发现反映用户特定行为的隐藏模式以及对 WEB 资源、会话和用户的简要统计。

在最后阶段，已发现的模式和统计信息将被进一步处理、过滤，进而可能得到聚合的用户模型以投入应用。

③主要数据来源是服务器日志文件，包括 WEB 服务器访问日志和应用服务日志。还有一些其他数据来源包括网站文件和元数据、操作数据库、应用程序模板和领域知识。

Data in Web Usage Mining:

Web server logs file (WEB server visit logs, application server logs)

Site contents

Data about the visitors, gathered from external channels

Further application data

2、Web 使用记录数据预处理的关键步骤是什么？

答：①数据的融合和清理：数据融合是指将来自多个 Web 和应用程序服务器的日志文件进行合并。数据清理通常根据站点不同而不同，涉及的工作有删除对分析无关紧要的嵌入式对象的引用，包括样式文件、图形以及声音文件。

②页面访问识别：页面访问的识别主要依赖于网站的页内结构、页面内容以及基础站点领域知识。

③用户识别：Web 使用记录的分析不需要用户识别的知识。 使用用户活动记录的形式来表示同一个用户的日志活动序列。

④会话识别：会话识别是将每个用户的用户活动记录分成一个一个会话的过程，每个会话代表了一次对站点的访问。

⑤路径完善：客户端或代理端的缓存功能经常会导致对那些被缓存的页面和对象的访问引用的丢失。而缓存而丢失的记录可以通过路径完善探索式的补全，路径完善依靠服务器日志上的站点结构和引用信息完成。

⑥数据整合：为了向模式发现提供最有效的框架，来自多渠道的大量数据必须与预处理过的点击流数据进行整合。

3、Web 使用记录挖掘的主要应用是哪些？什么是个性化推荐系统？什么是协同过滤？

答：①主要应用：

Web 使用挖掘已经作为实现更加个性化、用户友好化以及商业价值最优化的 Web 服务的一个必不可少的工具出现。如;推荐系统和协同过滤、查询日志挖掘。

②个性化推荐：基于内容的推荐根据项目与某个用户过去喜欢的项目之间的相似程度来预测该项目对该用户的效用，从而进行项目推荐。在这种方法中，每个项目通常由一套特征表示。通过调查问卷，或者也可以隐式地从用户交易行为中学得用户资料。通过对用户资料与用同一套特征表示的候选项目的比较获得推荐结果，前 k 个最匹配或最相似的项目将被推荐给用户。

③协同过滤：协同过滤最主要的特点在于它是根据其他志趣相投的用户过去打分或购买过的项目来预测项目对某个用户的效用。该方法通常只利用用户 - 产品交互数据而忽略用户和产品的本身属性。

下面我们将讨论 1 种流行的 CF 算法，即 k 近邻，基于关联规则的预测和矩阵分解。本质上，这种算法直接寻找相似的用户或项目（称作邻居）并用它们来预测目标用户的喜好。包括基于用户和基于项目两种算法。特别地，该算法用 k 近邻分类器来预测用户评分或购买倾向，具体是通过计算目标用户资料（可能是一组项目评分或者一组已访问或购买的项目）与其他用户资料的相关度来寻找数据库中与目标用户具有相似特征或爱好的用户。当 k 近邻被找到以后，预测结果便可以通过对这些邻居中的值进行某种整合而得到。