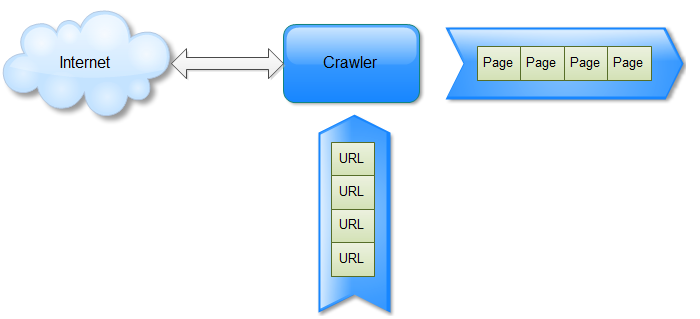
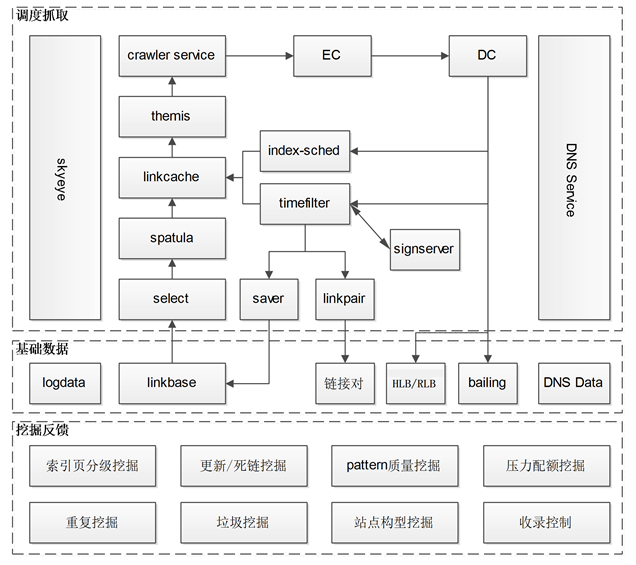
## 背景：

Spider简言之就是一个实现了HTTP协议的客户端：接收一系列URL，向网站发起抓取，输出一系列网页



衡量的指标：快、准、全、新；对应的是时效性、准确性、覆盖率、一致性

## Spider架构：



各模块的功能：

总体来看，spider架构如上图所示：

线上调度抓取

* 两个全局基础服务

[DNS](http://wiki.babel.baidu.com/twiki/bin/edit/Ps/Spider/SpiArchTopic/DNS_Service?topicparent=Ps/Spider/SpiArchTopic.SpiderArchOverview12_20;nowysiwyg=0)：站点IP解析及robots数据服务

skyeye：日志收集汇总系统。目标是得到链接在环上各模块调度的足迹

* select：链接选取模块 根据调度规则从Linkbase里选取出需要抓取的链接（包含对已抓链接的CHK和新链接的GET)，发往CS([CrawlerSystem](http://wiki.babel.baidu.com/twiki/bin/edit/Ps/Spider/SpiArchTopic/CrawlerSystem?topicparent=Ps/Spider/SpiArchTopic.SpiderArchOverview12;nowysiwyg=0" \o "CrawlerSystem (this topic does not yet exist; you can create it)))。
* spatula：流量抹平模块
* [linkcache](http://wiki.babel.baidu.com/twiki/bin/view/Ps/Spider/SpiArchTopic/SE_STAR_UnifiedSpiderArchitecture)：实时链接调度系统
* [themis](http://spark.baidu.com/spaces/themis/spaceviews/frameShow#t=2522)：统一压力控制系统
* [crawler service](http://wiki.babel.baidu.com/twiki/bin/edit/Ps/Spider/SpiArchTopic/Spider-cs-agile?topicparent=Ps/Spider/SpiArchTopic.SpiderArchOverview12_20;nowysiwyg=0)：抓取服务平台
* EC：单页面计算（信息提取，页面价值，页面类型，变化判断）
* DC：链接规则替换、删除 负责把抓取结果包拆分成网页包和链接包（又细分为状态包和提链包）。分别发往网页库和链接库 DC还负责按照挖掘或人工的规则对链接进行处理（替换、删除、降权）。
* index-sched：翻页索引页调度策略
* saver：linkbase更新模块(合并抓取状态包和提链包)
* linkpair：产出最近调度产生的链接对(A follow B)
* timefilter：时效性链接过滤模块，用于降低时效性链接抓取延时
* signserver：link库全库已抓、删除链接签名数据库。用于实时判断链接是否为新链接

基础数据

* linkbase：链接库
* bailing：网页库
* HLB：超链历史库，用于站点构型挖掘
* [RLB](http://wiki.babel.baidu.com/twiki/bin/view/Ps/Spider/SpiArchTopic/RLB)：反链库，用于链接关系挖掘与反链计算
* logdata：各模块日志。目前CS日志用于挖掘比较多
* 链接对：最近两个月调度链接follow关系。用于索引页分级挖掘。
* DNS数据：DNS dump数据，包含IP和robots。

挖掘反馈

* 索引页分级挖掘：index-check挖掘链接对数据，计算出链接贡献度和强度谱，用于索引页分级和hub页挖掘。影响覆盖率/时效性CHK调度
* 更新/死链挖掘：通过抓取状态挖掘死链接和死链规律，推送到检索端进行屏蔽。以prefix或pattern形式反馈到link库，根据更新率模型影响更新率调度。
* pattern质量挖掘：effect统计pattern中已抓部分质量，反馈到Link库。影响覆盖率/时效性GET调度。linkuniq通过pattern指向关系挖掘，产出低质量索引页pattern，影响覆盖率CHK调度。
* 压力配额挖掘：confcenter综合各种数据，产出IP、站点可承受的压力，以及select在各站点上的选取配额。数据分别应用于themis和select。
* 重复挖掘：dedup-mine通过网页库挖掘重复规则，更新到DC在线生效。同时挖掘出待替换、删除的链接，通过dedup-rm回灌link库、网页库删除。
* 垃圾挖掘：通过rubbish-mine、anchor-mine等挖掘出垃圾页面，通过arbitor综合判断后进行降权或删除。通过xbesom挖掘内容死链报错句子，反馈至EC进行在线内容死链识别。
* 站点构型挖掘：通过数据挖掘的方法理解站点结构，产出一些基础方法和数据。例如：链接pattern提取，页面集合聚类，指向关系研究，稳定内链挖掘等。
* 收录控制挖掘：挖掘计算站点上线，控制linkbase的收录规模，在linkbase一二层之间换入换出，防止出现收录异常暴增情况。

外部反馈

* 站点反馈：通过站长平台的sitemap、ping等进行数据推送。
* 用户反馈：主要是sobar。
* 建库反馈：网页是否通过garbage筛选，是否在线上。
* 前端反馈：点击、展现等数据。

## 常见问题及对应模块：

    未收录问题主要可拆解为如下方向：

    1. 链接发现，链接没有被发现，链接库中没有该链接

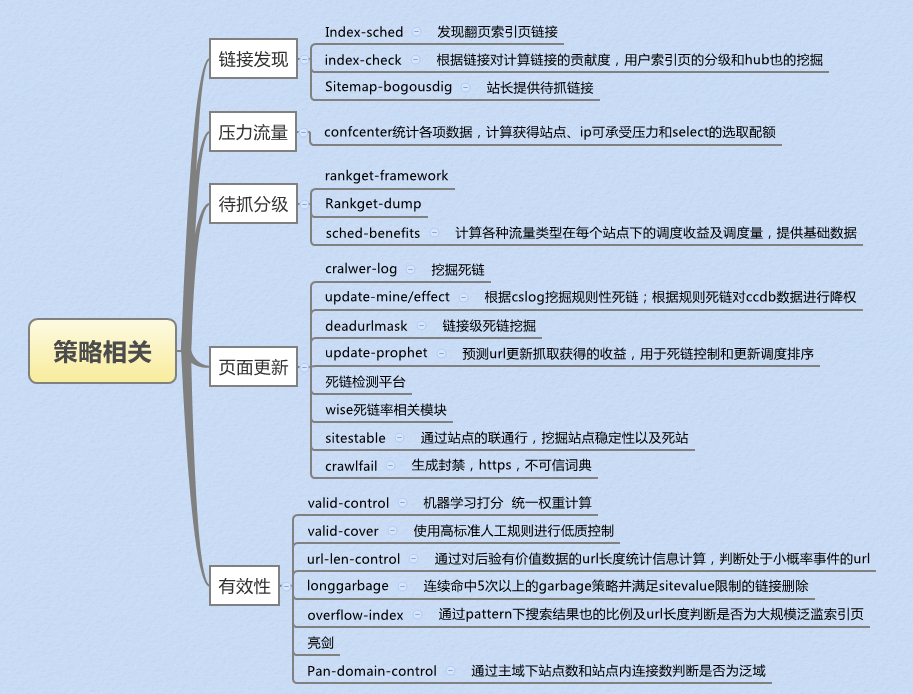
    2. 链接调度，链接已经被发现，但是没有被抓取

    3. 页面更新，死链，或者页面未及时更新，导致被删除或者降权

    4. 有效性，被各种有效性策略删除或者降权

    5. 价值，被低质和重复策略降权

    6. PAGE，主要是EC属性计算错误，DC误删除等问题



## Spider 3.0

一个统一的linkbase管理模块，取代select、saver、timefilter、signserver等一系列模块，简化抓取环架构；

链接质量打分、更新概率预测、新链接贡献度计算等策略，作为插件在linkbase管理模块中生效；

调度判断由天级批量计算，改进到单个链接入库时实时触发计算，与一定频率的全库扫描计算相结合。

