网页搜索爬虫时效性系统

叶顺平
yeshunping@gmail.com
分享时间：2013/12/30

大纲

一．爬虫时效性系统的目标

二．时效性系统的整体架构

三．时效性系统主要模块介绍

3.1 Rss/sitemap 系统介绍

3.2 泛爬系统与时效性的关系

3.3 种子调度系统

3.4 种子的挖掘

3.5 种子的更新机制

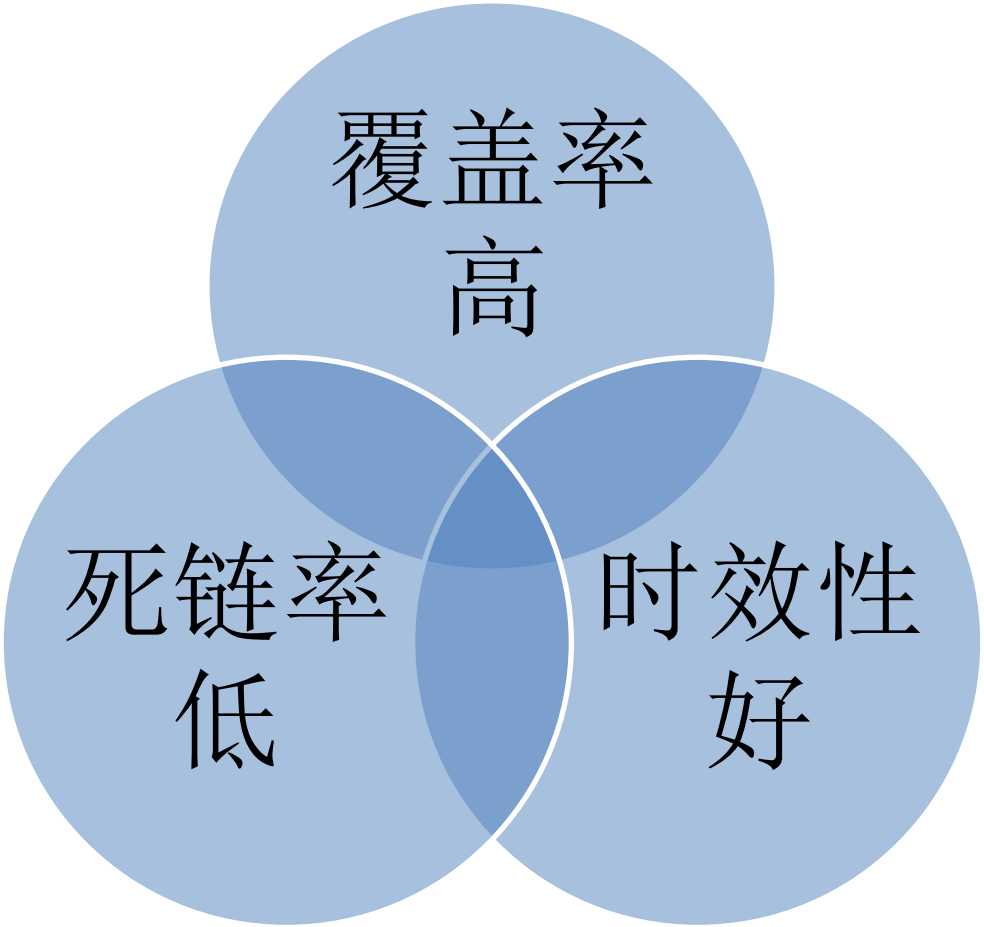
3.6 抓取系统与javascript解析

3.7 外部合作数据的引入

四、爬虫时效性数据的后续处理

五、有待改进的问题

网页爬虫的几个主要目标

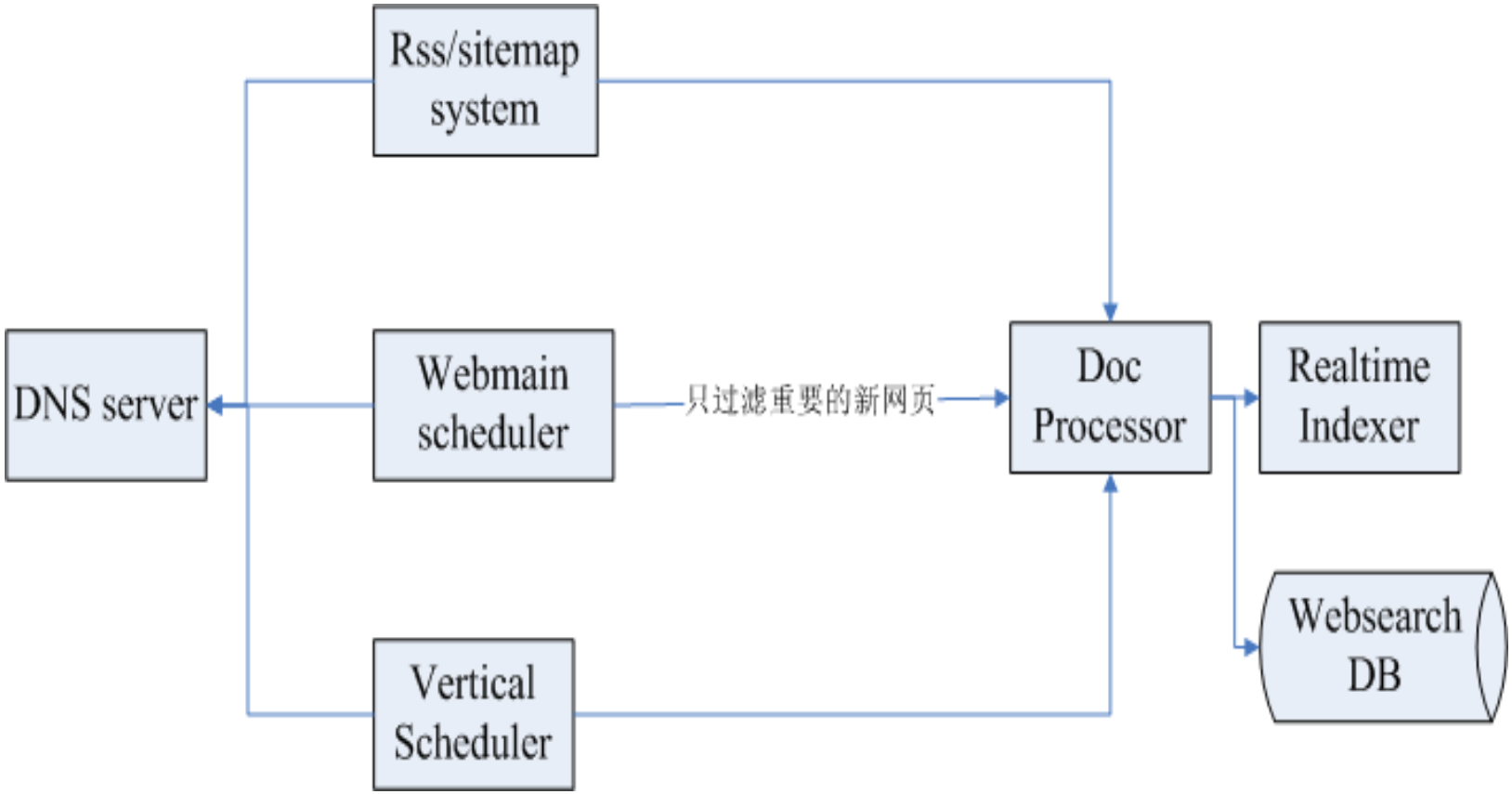


爬虫时效性系统的目标

新网页收录全

新网页收录快

时效性系统的整体架构


Rss/sitemap system
DNS server
Webmain scheduler
只过滤重要的新网页
Doc Processor
Realtime Indexer
Websearch DB
Vertical Scheduler

时效性问题的基本解决方案

天级以上

- 解决系统：网页大爬虫
- 链接规模：万亿级别
- Scan周期：一天左右一轮
- 抓取规模：几亿到几十亿/天

天级以内

- 解决系统：时效性系统
- Hub规模：10-100万级别
- Scan周期：几秒钟到天级
- 抓取规模：百万到千万/天

时效性系统主要模块介绍

- Rss/sitemap 系统介绍
- 泛爬系统与时效性的关系
- 种子调度系统
- 种子的挖掘
- 种子的更新机制
- 抓取系统与javascript解析
- 外部合作数据的引入

Rss/sitemap 系统介绍

rss

- http://en.wikipedia.org/wiki/RSS
- http://www.rssboard.org/rss-specification

- 豆瓣最新影评rss
- http://www.douban.com/feed/review/movie

rss_douban_movie.xml

Rss Cont

- <item>
- <title>xxx</title>
- <link>http://movie.douban.com/review/6456744/</link>
- <description>xxx</description>
- <content:encoded>xxx</content:encoded>
- <dc:creator>半辈子</dc:creator>
- <pubDate>Wed, 11 Dec 2013 01:27:21 GMT</pubDate>
- <guid isPermaLink="true">xxx</guid>
- </item>

Sitemap protocols

- w3c:
- http://www.sitemaps.org/

- Google:
- https://support.google.com/webmasters/answer/156184
- https://support.google.com/webmasters/answer/71453

- Baidu:
- http://www.baidu.com/search/sitemap_help.html

Where is sitemap---robots.txt

- Robots.txt of douban
- http://www.douban.com/robots.txt

- User-agent: *
- Disallow: /subject_search
- Disallow: /amazon_search
- …
- Sitemap: http://www.douban.com/sitemap_index.xml
- Sitemap: http://www.douban.com/sitemap_updated_index.xml
- ……

douban_robots.txt

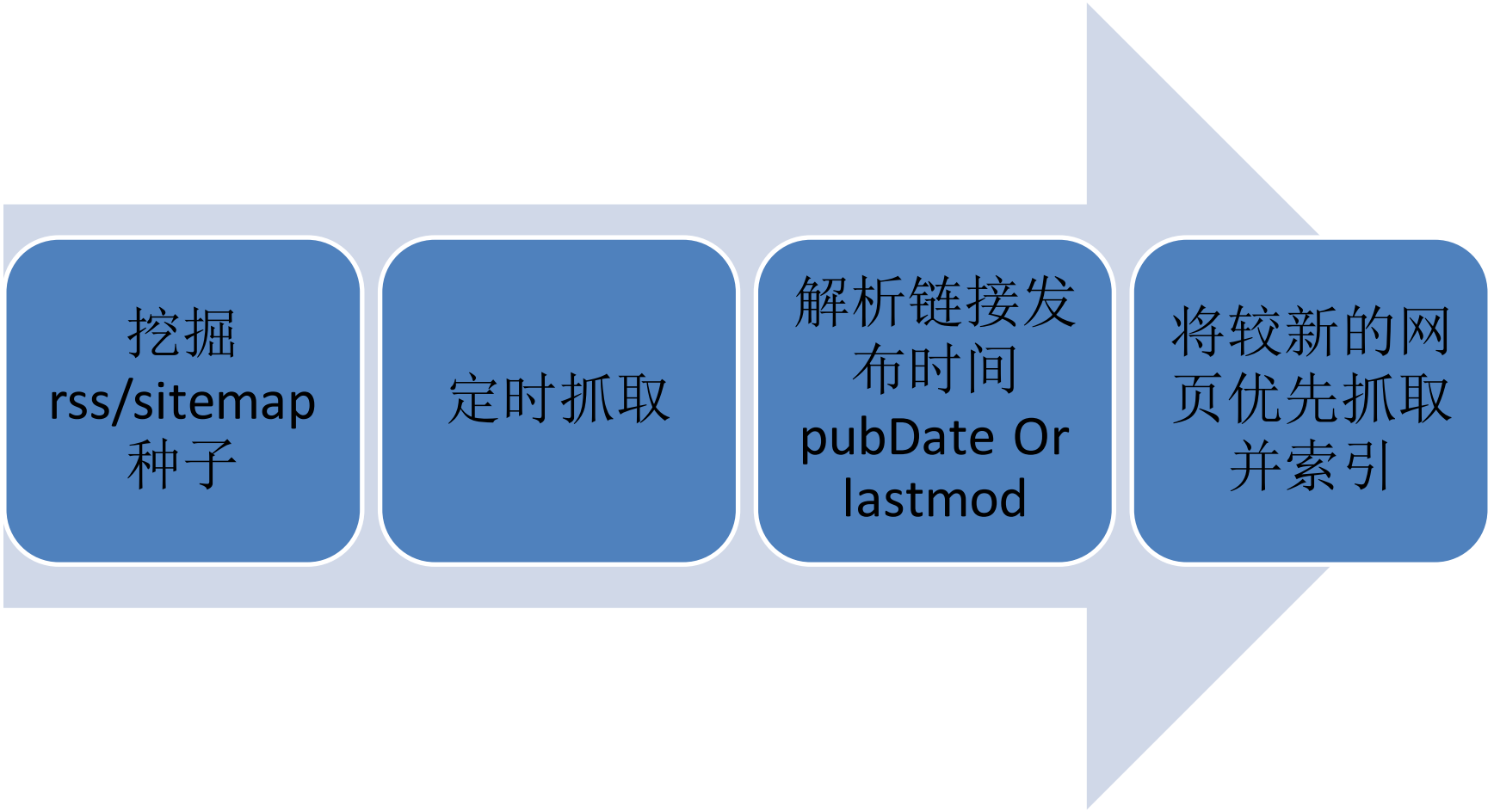
Sitemap Cont

- http://ent.people.com.cn/news_sitemap.xml

- <urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
- <url>
- <loc>
- http://ent.people.com.cn//n/2013/1229/c1012-23970224.html
- </loc>
- <lastmod>2013-12-29</lastmod>
- <changefreq>daily</changefreq>
- <priority>0.8</priority>
- </url>

ent.people.com.cn_news_sitemap.xml

时效性系统如何利用Rss/sitemap



Rss/sitemap 一些统计值

- rss种子数：　　几十万
- sitemap种子：　　几千万
- 日均发现新链接数：　　几千万
- 有sitemap的时效性网站：　　~10%
- 这些网站sitemap有新链接比例：　~25%

泛爬系统与时效性的关系

1，时效性系统收录全，有助于提高覆盖率

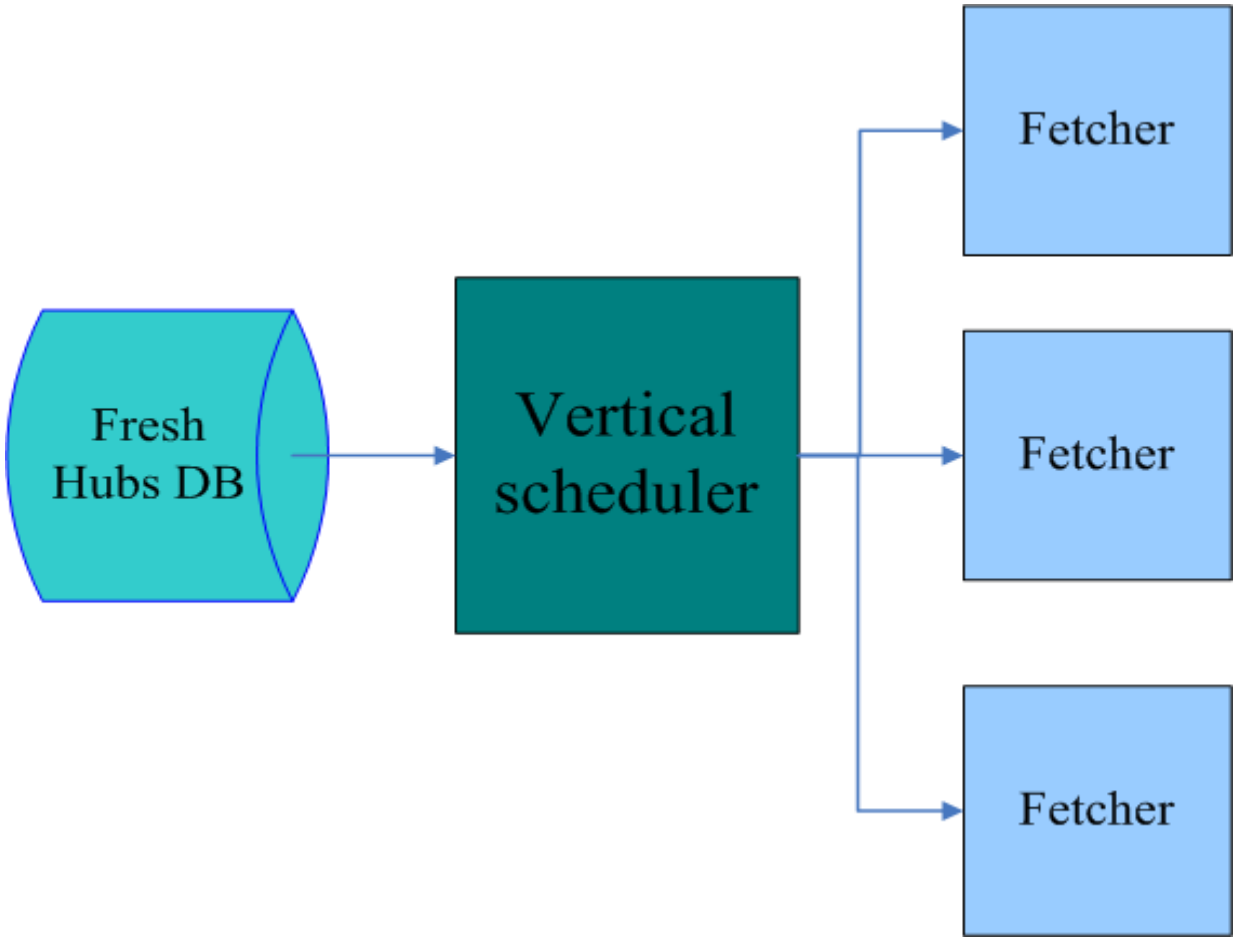
由于泛爬系统link规模庞大，周期较长，对于转瞬即逝的新链接，如果无法及时抓取，可能就比较难发现部分链接。(抓取时已经指向其它网页了)

2，泛爬系统设计良好，也有助于提高时效性网页的高覆盖率

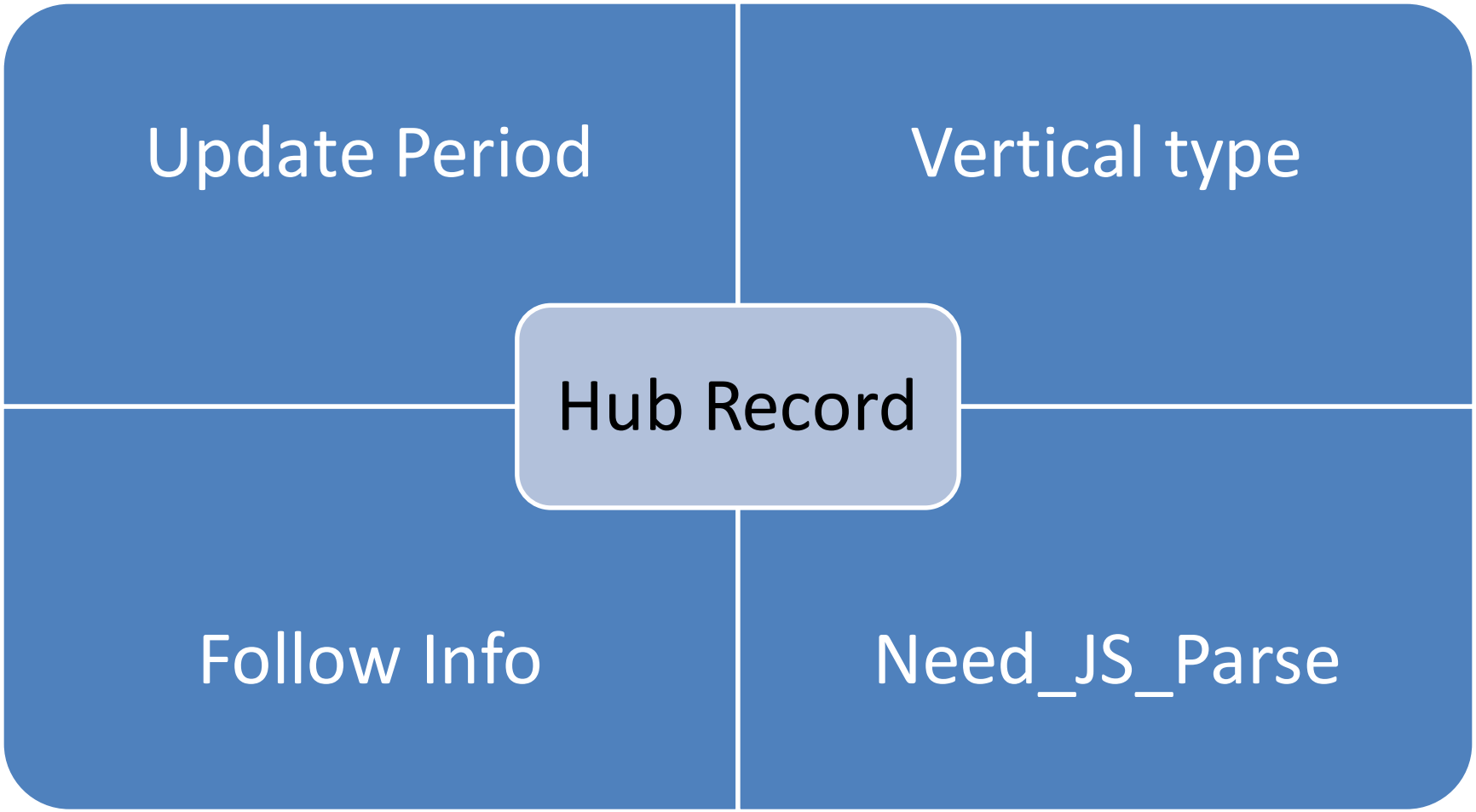
a)泛爬需要尽可能缩短调度周期，

b)泛爬抓取新数据，需要优先抓取新出现的、高质量的网页

种子调度系统



Hub DB Record fields



种子的挖掘

- 站点地图
- 导航条
- 面包屑
- 基于页面结构特征
- 基于页面变更规律
- ……

种子的更新机制

1，记录每个种子的抓取历史，follow的链接信息

2，定期根据种子的外链更新特征，重新计算种子的更新周期

更新周期

1，秒级

2，分钟级别

3，小时级别

4，天级别

抓取系统与javascript解析

部分种子页面，下载下来的html没有外链信息，需要使用浏览器才有。(浏览器下载Html后执行相关js脚本，进行了后续的网络下载)

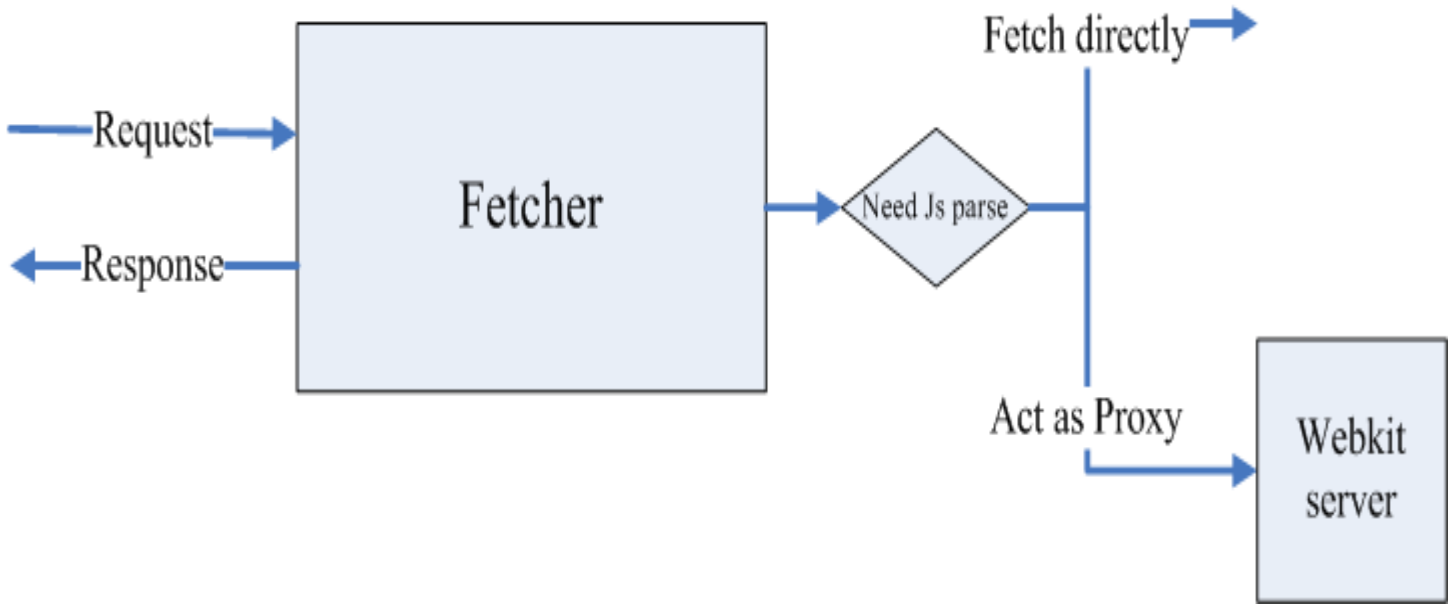
比如：

http://roll.news.qq.com/

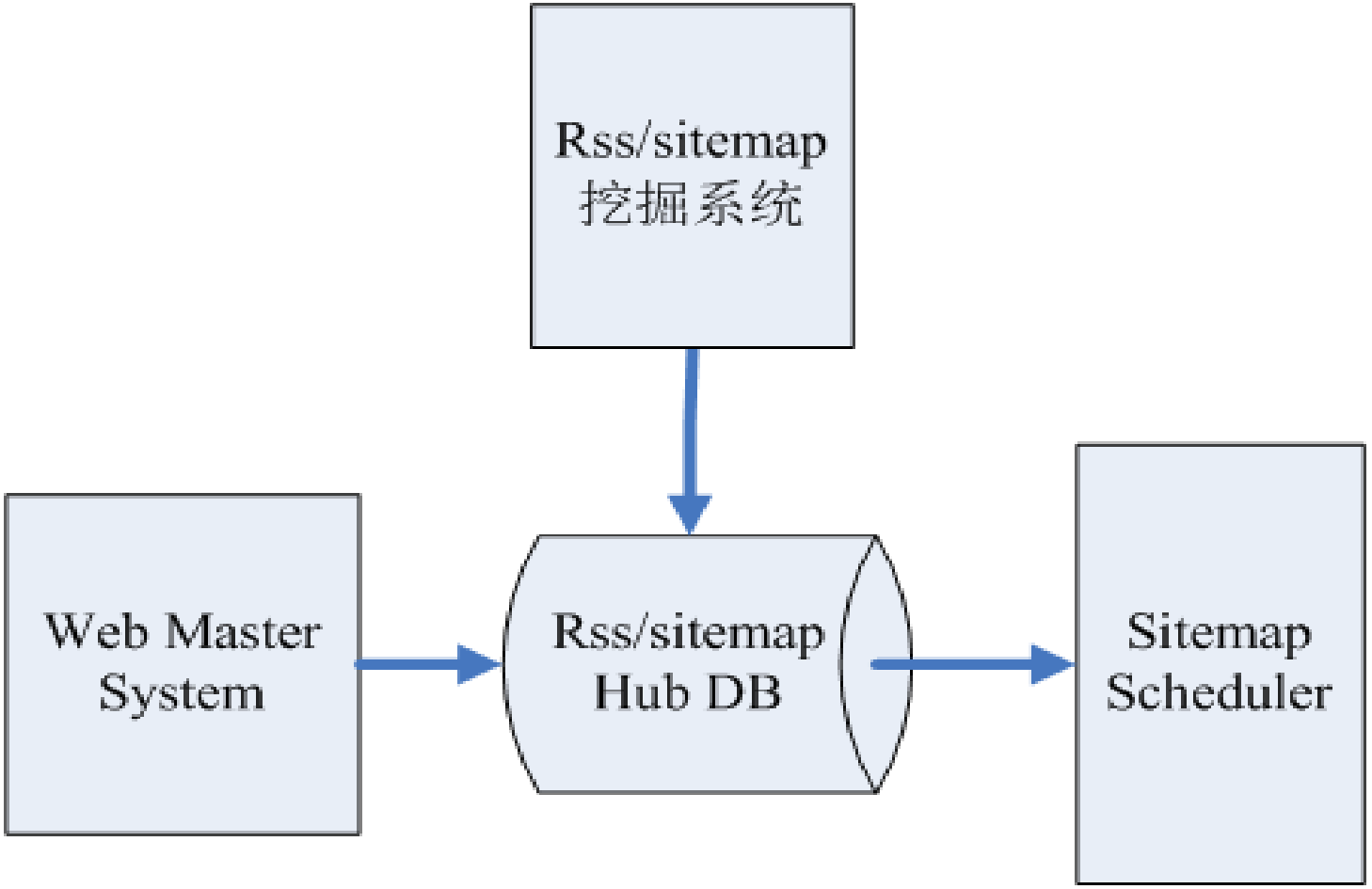
http://roll.news.sina.com.cn/s/channel.php#col=89&spec=&type=&ch=&k=&offset_page=0&offset_num=0&num=60&asc=&page=1

解决方案	使用开源项目	抓取性能
• 使用浏览器进行抓取，搭建一个基于浏览器抓取的抓取集群。	• Qtwebkit	• 几十个页面/单台每秒 • 几百万/单机每天

Different Fetchers With The Same RPC



外部合作数据的引入


Rss/sitemap
挖掘系统
Web Master
System
Rss/sitemap
Hub DB
Sitemap
Scheduler

爬虫时效性数据的后续处理

解析

- Publish time解析与过滤
- Doc type识别与过滤(比如新闻可能指向广告网页等)
- 。 。 。

建库

- 建库周期短
- 线上检索生效快

权重

- Pr周期长
- 实时计算外链信息等
- 新闻，论坛等有更多的页面属性(评论数，转发数，回帖数，查看数，分享数等)

需要改进的其他问题

- 1，js页面的抓取性能
- 2，种子挖掘的覆盖率
- 3，新种子的发现时间
- 4，种子调度周期的改善
- 5，时效性关键词作弊网页的识别过滤
- 6，时效性网页转载较多，权威性与原创性的识别
- ……

Q & A

Thanks