**Search Engine 数据结构摘要**

**原始网页存储**（原始）：

网页的保存需要按照一定的格式，以便以后数据的批量处理。这里介绍一种存储数据格式，该格式从北大天网的存储格式简化而来：

网页库由若干记录组成，每个记录包含一条网页数据信息，记录的存放为顺序添加；

一条记录由数据头、数据、空行组成，顺序为：头部 + 空行 + 数据 + 空行；

头部由若干属性组成，有：版本号，日期，IP 地址，数据长度，按照属性名和属性值的方式排列，中间加冒号，每个属性占用一行；

数据即为网页数据**（网页所有内容，包括html标签）**。

需要说明的是，添加**数据收集日期**的原因，由于许多网站的内容都是动态变化的，比如一些大型门户网站的首页内容，这就意味着如果不是当天爬取的网页数据，很可能发生数据过期的问题，所以需要添加日期信息（这里的日期指的是爬虫执行的日期）加以识别。

**原始网页库中的一条网页记录**

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx // 之前的记录

version:1.0 // 记录头部

url:http://ast.nlsde.buaa.edu.cn/

date:Mon Apr 05 14:22:53 CST 2010

IP:218.241.236.72

length:3981

<!DOCTYPE …… // 记录数据部分

<html> …… </html>

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx // 之后的记录

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

保存下来的网页信息具有较好的信息存储格式，但是还是有一个**缺点（如何改进？）**，就是**不能按照网页 URL 直接定位到所指向的网页**。所以，在第一个流程中，需要先建立**网页库的索引**，如此通过索引，我们可以很方便的**从原始网页库中获得某个 URL 对应的页面信息**。之后，我们处理网页数据，对于一个网页，首先需要提取其网页正文信息（**此处应该指的是网页所有内容**），其次对正文信息进行分词，之后再根据分词的情况建立索引和倒排索引，这样，网页的预处理也全部完成。

**网页库的索引**

采用“**网页库名—偏移**”的信息对来定位库中的某条网页记录。

表的内容如下：**url、content、offset、raws**。

URL 是某条记录对应的 URL，因为索引数据库建立之后，我们是通过 URL 来确定需要的网页的；

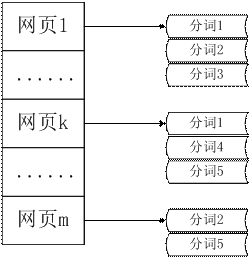
raws 和 offset 分别表示网页库名和偏移值，这两个属性唯一确定了某条记录。

content 是网页**内容的摘要（这是MD5值，不是通常说的摘要，**存储该摘要可以验证之后获取的网页信息是否与原始网页一致**）**，网页的数据量一般较大，把网页的全部内容放入数据库中显得不是很实际，所以我们将网页内容的 MD5 摘要放入到 content 属性中，该属性相当于一个校验码，在实际运用中，当我们根据 URL 获得某个网页信息是，可以将获得的网页做 MD5 摘要然后与 content 中的值做一个匹配，如果一样则网页获取成功，如果不一样，则说明网页获取出现问题。

对网页进行分词…..

有了分词的结果，我们就可以获得一个**正向的索引**，即某个网页以及其对应的分词结果。

**正向索引**



建立正向索引的目的就是通过翻转的操作建立**倒排索引**。

在 dySE 中，正向索引和倒排索引都是采用 **HashMap** 来存储

**倒排索引**



倒排索引中的网页ID映射到**网页库索引**中的某条记录（某个url）

在本文的开头，我们建立了**网页库的索引**，用于通过 URL （网页库的索引中的url）可以直接定位到原始网页库中该 URL 对应的数据的位置。