

网页信息存储的天网格式

2003 年 4 月

将获取网页信息保存在磁盘中，需要按照规定的格式保存，便于后续处理和提供服务。下面介绍天网格式存储方案。注意这种方案只是顺序保存网页信息，没有索引文件。

原始网页信息的存储格式应当设计为适合长期保存并易于处理，可以作为终端产品提供给用户使用。考虑到终端产品使用的便利性，要求原始网页库的存储格式具备简单性的特点。

由于存储介质都是有寿命的，所以应当考虑当存储介质损坏时数据的可恢复性。例如：磁盘的某个扇区损坏，导致部分数据不能读出，如果剩下的数据仍然可以使用，就能将损失降到最少。对于海量数据来说，在存储和传输的过程中，由于硬件和软件问题导致数据错误是不可避免的。因此，原始网页的存储格式还应当具备容错性的特点。

1. 天网存储格式定义

根据以上考虑，天网存储格式定义如下：

- 1) 一个原始网页库 (RAW_DB) 由若干记录组成，每个记录 (RECORD) 包含一个网页的原始数据，记录的存放是顺序追加的，记录之间没有分隔符；
- 2) 一个记录由头部 (HEAD) 和数据 (DATA) 和空行 (BLANK_LINE) 组成，顺序是：头部 + 空行 + 数据 + 空行；
- 3) 一个头部由若干属性组成，每个属性 (PROPERTY) 是一个非空的行，头部内部不允许出现空行；
- 4) 一个属性包含属性名 (NAME) 和属性值 (VALUE)，并由冒号“:”隔开，顺序是：属性名 + 冒号 + 属性值；
- 5) 头部的第一个属性必须是版本属性，属性名为 version，例如：version: 1.0，该属性表明记录的版本号；
- 6) 头部的最后一个属性必须是数据长度属性，属性名为 length，例如：length: 1800，该属性值必须是数据 (DATA) 的长度 (字节数)，不包括空行的长度；
- 7) 为简化起见，属性名必须是小写的字符串。

注：一个空行 (BLANK_LINE) 仅由一个换行符 (line feed, LF, 即 C 语言中的 “\n”) 组成，因表现为一个空行，所以称为空行。Microsoft Windows 系统和 UNIX 系统在换行机制上有所区别：在 Windows 系统下，一个换行由一个回车符 (carriage return, CR) 和一个换行符组成 (即 C 语言中的 “\r\n”)；而在 UNIX 系统中一个换行仅由一个换行符组成。建议采用 UNIX 系统的换行机制。

2. 当前存储格式版本描述

存储格式允许有多个版本，以满足将来进行扩展的需要。

当前存储格式的版本属性为 1.0。一个记录的存储格式如下 (// 后为注释)：

```
version: 1.0 // 版本号
url: http://www.pku.edu.cn/ // URL
origin: http://www.somewhere.cn/ // 原来的 URL
date: Tue, 15 Apr 2003 08:13:06 GMT // 抓取时间
```

```

ip: 162.105.129.12           // IP 地址
unzip-length: 30233          // 如果数据经过压缩, 则需有此属性
length: 18133                // 数据长度
                               // 空行
XXXXXXXXXX                   // 以下为数据
XXXXXXXXXX
....
XXXXXXXXXX                   // 数据结束
                               // 最后再插入一个空行

```

各属性说明：

version 属性为版本号，以下说明适用版本号为 1.0 的情况。

url 指该网页的 URL，如果因为 HTTP HEAD 中包含 Location 字段而产生网页转向时，该 URL 为最后实际抓取的 URL 地址。该属性是必需的。

origin 指该网页的原始 URL。该属性仅在 HTTP HEAD 中包含 Location 字段而产生网页转向时存在，指向最原始的 URL。

date 属性为该网页的保存时间，保存格式为 RFC822 所制定的格式。该属性是必需的。

ip 属性为该网页所在服务器的 IP 地址。

unzip-length 属性仅在数据经过压缩时存在，记录数据未压缩时的原始长度。

length 属性记录数据长度。

若存在其它未加说明的属性，应用程序可以简单地忽略。

关于数据是否压缩的问题：天网格式并不指定数据是否必须经过压缩。但是压缩的数据必须包含 unzip-length 属性而未压缩的数据不能包含该属性。该属性同时也是解压缩所必需的。如果数据经过压缩，还应附带说明压缩算法，必要时附带压缩函数库及源代码。

3. 数据的可恢复性分析

假设由于数据遭到破坏，只得到其中一个残存的片段。则可按以下步骤找出该残存片段中所有完整的记录：

- 1) 特定字符串“version:”，除非没有一个完整的记录，该字符串肯定能找到。记录该字符串的位置 POS。
- 2) 找到该字符串后，判断其后的数据是否满足存储格式 2) 3) 4) 6) 7) 条件。如果任何一个条件不满足，返回 1，从记录的位置 POS 开始继续查找下一个特定字符串“version”。
- 3) 当满足条件 2) 时，假定这是一个正确的记录，则下一个记录也必定是一个正确的记录。检查该记录满足天网存储格式 2) 3) 4) 5) 6) 7) 条件，如果任何一个条件不满足，说明原先的假定错误，返回 1，从记录的位置 POS 开始，继续查找下一个特定字符串“version”。如果条件都满足，则继续检查下一个记录是否正确。
- 4) 如果连续 3 个记录都是正确的，则认为 1) 所找到的“version”是一个正确的记录的开始，可以依此提取出全部正确的原始网页。

由于原始网页是随机的，而存储格式是严格的，因此经过上述方法得到的记录为错误记录的可能性极小，是完全可以接受的。

4. 其它问题

在实际应用天网存储格式时，应该注意下面两个方面。

- 1) 文件打开模式：文件有两种打开模式：文本（text）模式和二进制（binary）模式。读写原始网页库文件时，应以二进制模式打开。在 Windows 系统中，如果以文本模式打开，可能会产生读数据时“\r\n”被替换为“\n”的现象，导致数据错误。
- 2) FTP 传输：FTP 传输也有两种传输模式：文本（text）模式和二进制（binary）模式。传输原始网页库文件时，应以二进制模式进行传输。如果以文本模式传输，可能会出现“\r\n”被替换为“\n”或“\n”被替换为“\r\n”的现象，导致数据错误。