J Search Engine 原理

00648333 陈志杰(Joyan)　　00648332 揭忠(lezhengyi)

December 22, 2007

1　分组情况

陈志杰（00648333），揭忠（00648332）

2　工作流程

2.1　抓取部分

文件名 crawl

用法 crawl dir-name

说明 递归访问dir-name下的文件，在dir-name下建立.tianwang.raw原始文件。

备注 天网文件格式：

```
    version:1.0
    url:foo.txt
    date:Tue, 15 Apr 2003 08:13:06 GMT
    length:12345

    XXXXXXXXXXX
    XXXXXXXXXXXXX
    ...
    XXXXXX

    version:1.0
    ...
```

2.2　索引部分

文件名 docoff

用法 docoff xxx.raw

说明 将raw中每个原始文件的url的MD5与其正文在raw文件中的偏移关联，存入xxx.didx

备注 *.didx文件格式：

```
1    DocID offset
2    DocID offset
3    ......
```

文件名 rawseg

用法 rawseg xxx.raw xxx.didx

说明 根据xxx.didx对 xxx.raw 进行切词处理，存入 xxx.raw.seg

备注 *.raw.seg文件格式：

```
1    DocID term1 term2 ...
2    DocID term1 term2 ...
3    ......
```

文件名 sort

用法 sort xxx.didx > xxx.didx.sort

说明 排序

文件名 preinvert

用法 preinvert xxx.raw.seg

说明 将xxx.raw.seg转化为准倒排文件xxx.piidx

备注 *.piidx文件格式：

```
1    term(lock 8char) DocID(lock 8char HEX)
2    term DocID
3    ......
```

文件名 sort

用法 sort xxx.piidx >xxx.piidx.sort

说明 排序

文件名 invert

用法 invert xxx.piidx.sort

说明 将准倒排文件转化为正式的倒排文件,存入xxx.iidx

备注 *.iidx文件格式：

```
1    term DocID DocID docid ...
2    term DocID DocID ...
3    ......
```

2.3 查询部分

文件名 query

用法 query keyword

说明 将keyword切词，分别求出其docid集合，合并后将docid按照词频排序,输出到res.txt文件中

备注 res.txt文件格式：

```
1    docid offset1 offset2 offset3 ...
2    docid offset1 ...
3    ......
```

2.4 用户部分

文件名 query.htm

用法 用浏览器打开

说明 将res.txt翻译成html格式并显示出来。

备注 条目格式：

文件名
摘要摘要摘要摘要摘要摘要摘要摘要摘要摘要
摘要摘要摘要摘要摘要摘（关键词高亮）
文件路径