



逸仙时空BBS

搜索引擎设计与实现

邓智平 05372004
信息科学与技术学院
计算机科学与技术

dengzhp@mail2.sysu.edu.cn
2009.5.23 中山大学东校区

大纲

- 1、系统开发背景
- 2、系统设计
- 3、系统实现
- 4、系统演示
- 5、总结展望



回顾BBS

- 计算机网络早期发展而来
- 一般运行在类Unix系统
- 采用纯文本方式表现



逸仙时空

- 建立于1996年
- 校内信息交流平台：
 - Job, EastCampus, Official...
- 中大学子的网上家园：
 - Joke, Girls, Diary, Employee...
- 学术讨论, 第2课堂
 - Linux, ACMICPC...



逸仙时空 (续)

- 约200个讨论区(版面)
- 版面文章 (文本文件)
 - 结构
 - 数目
 - 大小
 - 更新速度
- 精华区文章
 - 结构
 - 大小
 - 更新速度



站友日益增长的需求

- 互联网的高速发展
 - 快速查找信息的需求迫切
- 搜索引擎出现, Google, 百度..
- 新型论坛的兴起
 - 功能丰富(相比bbs 的www界面)
- 新人抱怨或疑问
 - 逸仙时空怎么查找文章?



开发动机和条件

- 常在BBS泡，日久生情
- 把所学知识应用到实际中去
- 让BBS支持全文检索
- 担任过系统维护员
 - 对BBS比较了解一点



搜索引擎的组成

- 下载系统. → 搜集
- 分析系统.
→ 处理
- 索引系统.
- 查询系统. → 服务



设计理念与原则

- 保持简单
 - 选取简单可行的方法
// 搜索引擎很复杂
- 分而治之
 - 降低复杂度
//按版面划分处理集合



系统实现

- Unix-like 系统开发运行环境
- 核心部分 C 语言编写
 - 熟悉
 - 运行效率高
- 一些脚本程序 - Python
 - 开发快速



搜集文章

增量抓取策略

- 1.BBS服务器生成文章列表
- 2.下载文章列表
- 3.利用diff工具得出差异
- 4.去除已删的文章，下载新增文章。



解析HTML

- 观察HTML代码
 - [http://bbs.sysu.edu.cn/bbscon?
board=CS&file=M.1242478980.A](http://bbs.sysu.edu.cn/bbscon?board=CS&file=M.1242478980.A)
- 提取正文
 - <blockquote> **这是正文** </blockquote>
 - 跳过HTML标签<>内的字符，
 - 反转义 < > & ...



过滤ANSI转义序列

- 为终端控制文本格式和其他输出选项
 - 包括彩色显示，光标移动等

Esc[1;31;42m 这是红色背景黄色亮字 Esc[0m
显示为：



- 根据规则过滤



切分词语

- 分词的目的
 - 提高效率
 - 减少使用单字位置信息拼合匹配的次数。
 - 提高搜索准确度
- 针对英文的处理
 - 按空格,标点分开
 - 转化为小写 (case folding)



中文分词

- 基于词典最大匹配
 - 简单
 - 无法处理二义性 (研究生命起源)
- N-Gram切分
 - 照顾了所有的可能
 - 缺点：索引庞大，浪费
- 基于统计语言模型
 - 使分完词后的句子出现概率最大



分词策略

- MMSeg算法 (最大匹配算法的变种)
 - 找到所有从当前位置开始的三个连续词语的块，总长度最大的块为最优解

眼 看 就

眼 看 就要

眼看 就 要

眼看 就要 来

眼看 就要 来了 √



分词策略

仍有多个结果，怎么办？

- 最大平均长度的规则
 - 国际化
- 词语长度变化最小的规则
 - 研究生命起源
- 单字频率和最大的规则
 - 主要是因为



索引

- 为什么要建立索引

如果直接采取暴力法 (BruteForce)

遍历每一篇文档d

对每一个关键字q

计算(d, q)之间的相关性

大部分文档跟查询是无关!!

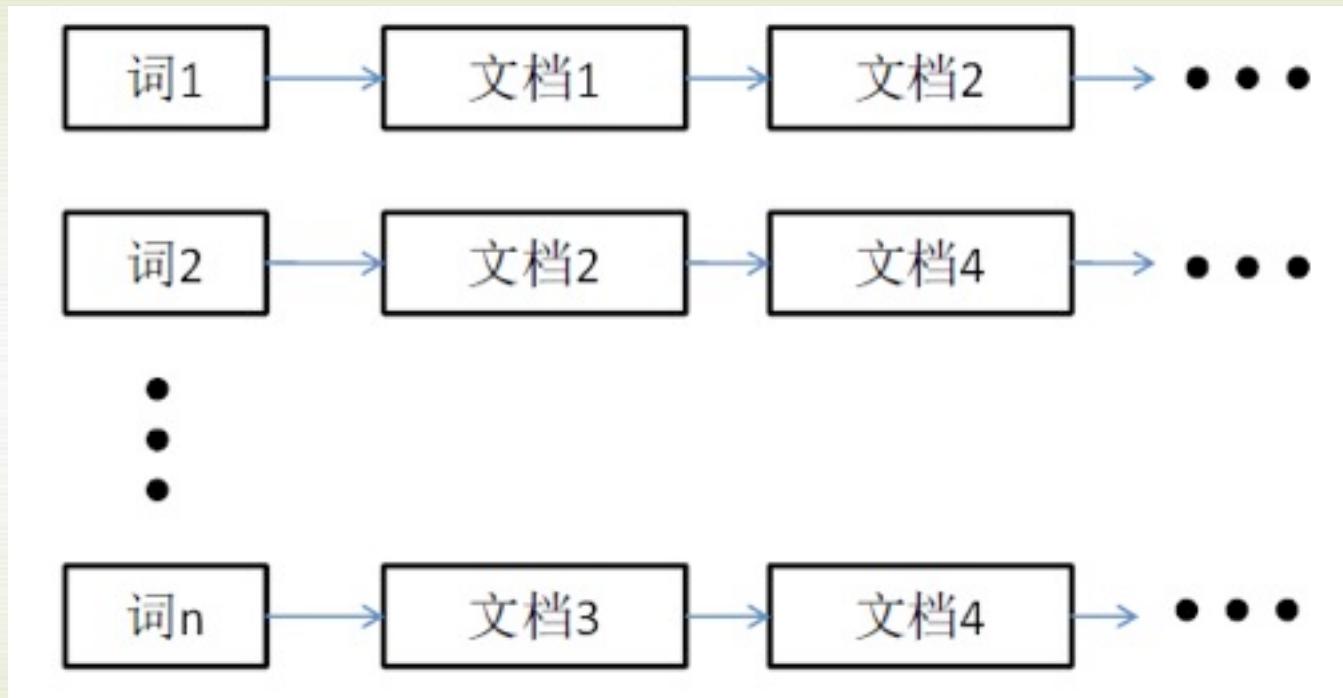


索引(续)

- **Key observation**
 - 相关的文档中必须要含有至少一个检索词
- 记录包含每个检索词的文档列表
- 把"文档-关键词"矩阵转置为"关键词-文档"矩阵



倒排文件



倒排文件(续)

- 效率最高的索引结构
- 每一个关键词(term)对应着一个记录表(postings list)，记录着含有此词语的文档号(DocId)
- 就像书目后面的索引



倒排文件(续)

- 基本的倒排文件组成
 - 关键词词典
 - 对于每个关键词t，记录一个包含它的文档数和一个指向对应的记录表头部的指针。
 - 记录表的集合，
 - 每个记录表保存着对应的文档号列表
 - 关键词在对应文档的出现的频率列表



文档的表示

- 文档的唯一标识符
- 文档路径(或url)? 太占空间了
- 为文档进行编号
 - url进行md5计算，把字符串映射到一个64位的整数
 - 简单地从1开始按处理顺序给文档进行编号



建立索引

```
1 output_file = new file()
2 dict = new hash()
3 while (free memory available)
4 do token = next_token()
5 if token not in dict
6   postinglist = addtodict(dict, token)
7 else postinglist = getpostinglist(dict, token)
8 if full(postinglist)
9   postinglist = doublepostinglist(dict, token)
10 addtopostinglist(postinglist, docid(token))
11 sorted_terms = sortterm(dict)
12 writeblock(sorted_terms, dict, output_file)
```



建立索引(续)

- 算法流程
 - 读取和索引文档直到内存不足时，按词典顺序把记录表写入磁盘，最后扫描多个临时倒排文件，归并成最后的倒排索引文件。
 - 归并的过程
- 使用Berkeley DB来存储索引



索引压缩

- 压缩的好处
 - 减少磁盘空间 **(不是主要动机)**
 - 提高cache的命中率
 - CPU 速度 >> IO速度

在大多数检索系统，压缩记录表要比不压缩要运行得快



编码方案

- 差分编码(gap encoding)
 - 记录表的docID是从小到大有序的,
 - 只保存docID之间的增量
- 记录表里 每个记录里的词频tf较小
 - 定长的数据 浪费
- 变长编码方法
 - 字节对齐
 - 将整数转成二进制, 以7位分段
 - 每段前面加一位标志位,恰好一个字节
 - 首位为1表示为最后1段,0表示还有后续段



索引维护

- 两种策略
 - 重新构造
 - 把新增的文章索引与原有索引合并
- 回顾
 - 版面文章集合较小
 - 更新快
 - 保持简单
- 周期性的产生新索引，待完成后把查询转移到新的索引中进行



向量空间检索模型

- 一个文档被描述成由一系列关键词组成的向量
- 查询也同样被向量空间所表示
- 模型对查询与文档，计算一个相似度

$$\text{Sim}(D, Q) = \frac{D \bullet Q}{|D| * |Q|}$$



关键词权重

- 每一维如何取值?
- 三个基本原则:
 1. 对在较多篇文章出现的词语给予较小的权重
 2. 对在一篇文章出现较多次的词语给予较大的权重
 3. 对于出现较多词语的文章给予较小的权重



相似度计算

- (1) Calculate $w_{q,t}$ for each query term t in q .
- (2) For each document d in the collection,
 - (a) Set $S_d \leftarrow 0$.
 - (b) For each query term t ,
Calculate or read $w_{d,t}$, and
Set $S_d \leftarrow S_d + w_{q,t} \times w_{d,t}$.
 - (c) Calculate or read W_d .
 - (d) Set $S_d \leftarrow S_d / W_d$.
- (3) Identify the r greatest S_d values and return the corresponding documents.



其他

- 摘要
- 快照
- 日志
- 安全隐私



最终系统组成

约3500行C代码

- 下载，解析，分词
- 建立索引，查询

几个Python脚本

- 做一些辅助工作

3个HTML页面

配置文件

- 版面列表
- 字典文件



系统演示

- <http://frgg.3322.org>



系统评价

- 索引
 - 一个版面建立索引用时15秒~2分钟
 - 根据机器配置，负载
 - 已索引 49个版面的文章(总大小5.3G)
索引文件总大小为1007M,约为原数据20%
 - 每小时更新版面文章的索引，达到了实时搜索



系统效果

- BBS用户反馈
Cool, Good~ ...

见EastCampus文章

- <http://bbs.sysu.edu.cn/bbscon?board=EastCampus&file=M.1241974306.A>
- 分析查询日志
- 发布不到20天, 1000 独立IP访问, 5000次查询

最终达到了方便用户的目的!



展望

- 继续完善系统
 - 整合各版面搜索结果
 - 提高搜索准确性
 - 实现在telnet端查询的功能
 - 对文章的元信息建立索引
 - 提供高级个性化的搜索





感谢各位答辩委员！

感谢我的导师！

感谢所有帮助过我的人！

邓智平 05372004

信息科学与技术学院

计算机科学与技术

dengzhp@mail2.sysu.edu.cn

2009.5.23 中山大学东校区