Outline

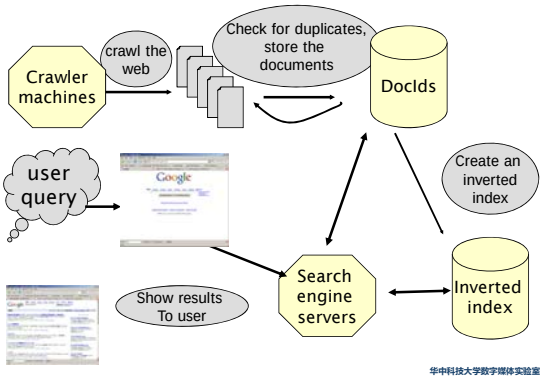
http://media.hust.edu.cn

- 搜索引擎体系结构
- 小型搜索引擎 - 目标与功能
- 采集工具Larbin简介
- 数据分析与预处理工具
- 全文检索工具Lucene简介
- 小型搜索引擎的设计与实现
- 实用化问题
- 参考文献

华中科技大学数字媒体实验室

搜索引擎体系结构

http://media.hust.edu.cn



华中科技大学数字媒体实验室

典型的全文搜索引擎

http://media.hust.edu.cn

- 采集器
- 分析器
- 索引器
- 检索器
- 人机接口
- 索引数据库

华中科技大学数字媒体实验室

Outline

http://media.hust.edu.cn

- 搜索引擎体系结构
- 小型搜索引擎 - 目标与功能
- 采集工具Larbin简介
- 数据分析与预处理工具
- 全文检索工具Lucene简介
- 小型搜索引擎的设计与实现
- 实用化问题
- 参考文献

华中科技大学数字媒体实验室

搜索引擎体系结构

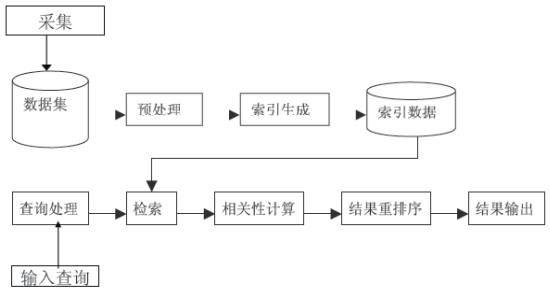
http://media.hust.edu.cn

- 目标
  - 理解信息检索技术的原理
  - 实际搭建一个可运行的实验系统
- 要实现的功能
  - 采集
  - 格式转换：HTML2TXT,PDF2TXT, PS2TXT…
  - 词法处理：分词、Stemming
  - 创建索引
  - 增加媒体搜索模块
  - 全文检索

华中科技大学数字媒体实验室

简单搜索引擎流程

http://media.hust.edu.cn



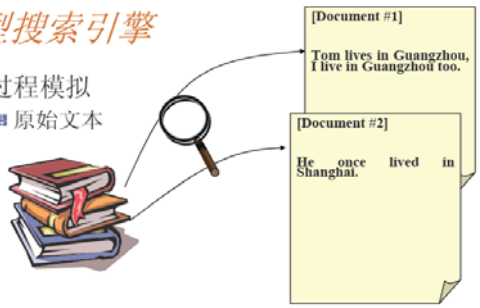
华中科技大学数字媒体实验室

简单搜索引擎流程

http://media.hust.edu.cn

小型搜索引擎

- 过程模拟
  - 原始文本



华中科技大学数字媒体实验室

简单搜索引擎流程

http://media.hust.edu.cn

- 预处理之后的文本

[Document #1']
[tom]
[live]
[in]
[guangzhou]
[.]
[i]
[live]
[in]
[guangzhou]
[too]
[.]

[Document #2']
[he]
[once]
[live]
[in]
[shanghai]
[.]

华中科技大学数字媒体实验室

简单搜索引擎流程

http://media.hust.edu.cn

- 倒排索引

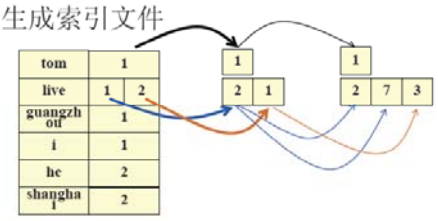
[tom] 1
[live] 1,2
[guangzhou] 1
[i] 1
[he] 2
[shanghai] 2

华中科技大学数字媒体实验室

简单搜索引擎流程

http://media.hust.edu.cn

- 生成索引文件

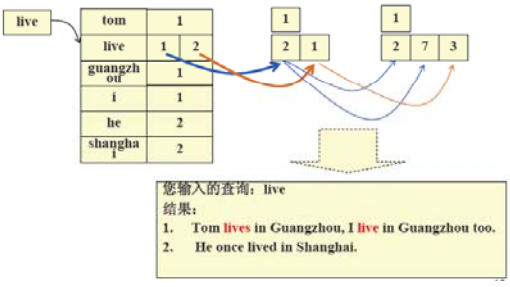


华中科技大学数字媒体实验室

简单搜索引擎流程

http://media.hust.edu.cn

- 检索



华中科技大学数字媒体实验室

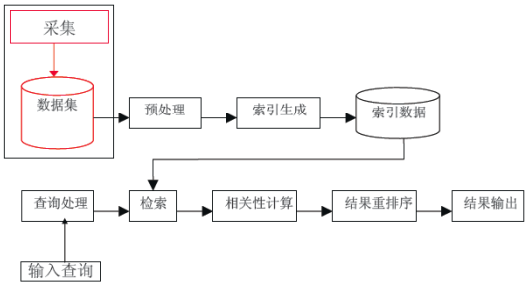
Outline

http://media.hust.edu.cn

- 搜索引擎体系结构
- 小型搜索引擎 - 目标与功能
- 采集工具Larbin简介
- 数据分析与预处理工具
- 全文检索工具Lucene简介
- 小型搜索引擎的设计与实现
- 实用化问题
- 参考文献

采集工具

http://media.hust.edu.cn



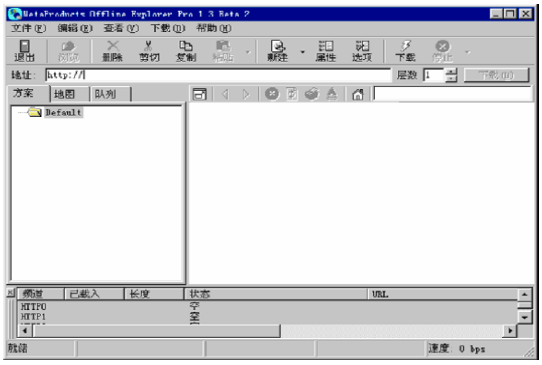
采集工具

http://media.hust.edu.cn

- Offline Explorer(离线浏览软件)
  - 小型网页采集器
  - 性质：商业软件
  - 作者：MetaproductsInc.
  - 平台：Windows

采集工具

http://media.hust.edu.cn



采集工具

http://media.hust.edu.cn

- wget
  - 命令行形式的小型采集器
  - 性质：GNU GPL
  - 作者：HrvojeNiksic
  - 平台：Unix/Linux/Windows
  - 功能：支持FTP/HTTP下载，支持断点续传

采集工具

http://media.hust.edu.cn

- wget
  - 用法
    - wget[option] [URL]
  - 举例：
    - wget-r -l4 -t0 http://www.hust.edu.cn/
    - wget-t 0 -w 31 -c -B ftp://user@password:ftp.mtgroup.ict.ac.cn/smt-i filelist.txt-o down.log &

采集工具

http://media.hust.edu.cn

- Larbin简介
  - 全功能的网页爬虫
  - 性质：GNU GPL
  - 作者：SébastienAilleret
  - 网址：http://larbin.sourceforge.net/index-eng.html
  - 版本：V2.6.3
  - 平台：Linux/FreeBSD
  - 后续支持：Nutch

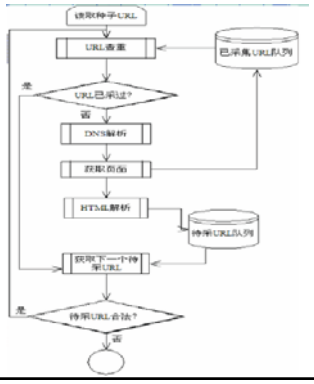
采集工具

http://media.hust.edu.cn

- Larbin简介
  - 功能特点
    - 高速，低资源占用
    - 从种子站点开始全网采集
    - 支持远程输入站点采集
    - 限定域/站点/文件类型采集
    - 现场保存和续采
    - 网页去重
    - 状态检测与显示

Larbin采集流程

http://media.hust.edu.cn



Larbin简介

http://media.hust.edu.cn

- 运行
  - ./larbin[-c larbin_conf] [-scratch]
- 结果存放
  - save/dxxxx/fxxxx⇔index
- 运行状态查看
  - http://server_run_larbin:8081/

Larbin简介

http://media.hust.edu.cn

- 配置与部署
  - 配置文件
  - 外部环境变量设置：Larbin.conf
  - 内部变量设置：
    options.h
    types.h

Larbin配置与部署

http://media.hust.edu.cn

[Larbin.conf]

- From,UserAgent: robot协议
- httpPort,inputPort: 管理与交互
- pagesConnexions,dnsConnexions: 并发采集占用的网络资源
- depthInSite: 每个站点的采集深度
- noExternalLinks: 限定站点
- waitDuration: 采集间隔
- Proxy: 代理
- StartUrl: 种子Url
- limitToDomain: 限定域，如cn, uk等
- forbiddenExtensions: 排除文件类型

Larbin配置与部署

http://media.hust.edu.cn

[options.h]

- #define DEFAULT_OUTPUT // do nothing...
- #define SIMPLE_SAVE // save in files named save/dxxxxxx/fyyyyyy
- #define MIRROR_SAVE // save in files (respect sites hierarchy)
- #define STATS_OUTPUT // do some stats on pages
- #define SPECIFICSEARCH //Set up a specific search
- #define FOLLOW_LINKS //follow links in pages
- #define LINKS_INFO //associate to each page the list of its sons
- #define NO_DUP //suppress duplicate pages
- #define MAXBANDWIDTH 200000

Larbin配置与部署

http://media.hust.edu.cn

[types.h]

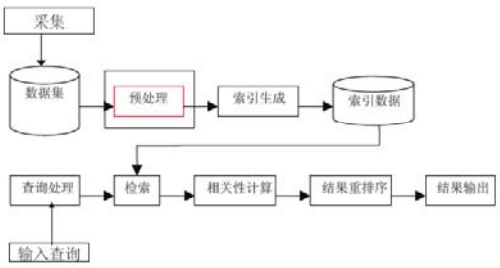
- #define hashSize 64000000 //max number of urls that can be fetched
- #define timeoutPage 30 // default time out
- #define maxPageSize 100000
- #define saveDir "save/" // if we save files,where files are stored
- #define indexFile "index.html" //for MIRROR_SAVE

Outline

http://media.hust.edu.cn

数据预处理

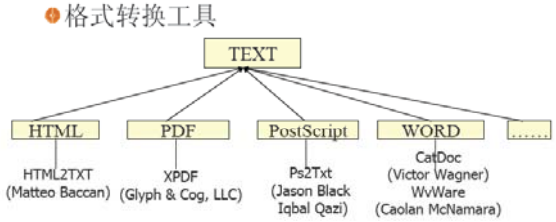
http://media.hust.edu.cn



数据分析与预处理工具

http://media.hust.edu.cn

格式转换工具



数据分析与预处理工具

http://media.hust.edu.cn

格式转换工具

名称	功能	接口
HTML2TXT	去处 HTML 页面中的标签，把转义字符替换为普通字符，得到纯文本文件	void dfConvert(char * Buffer)
XPDF	把 PDF 文件转换为文本文件	int pdf2txt(char* pdffile, char* txtfile)
Ps2Txt	把 PostScript 文件（包括由 DVI 格式转换而来的 PostScript 文件）转换为文本文件	void dviparse(FILE* source) void psparse(FILE* source)
CatDoc	把 Office 文件（Doc, Excel, Powerpoint）转换为文本文件（支持 Office97）	int doc2txt(char* docfile) int xls2csv(char* xlsfile) int ppt2txt(char* pptfile)

数据分析与预处理工具

http://media.hust.edu.cn

- 中文分词工具 - ICTCLAS（ Chinese Lexical Analysis System ）
  - 中文分词、命名实体识别工具
  - 性质：自然语言处理开放资源许可证
  - 作者：张华平，刘群
  - 网址：http://www.nlp.org.cn/project/project.php?proj_id=6
  - 平台：Windows/Linux
  - SIGHAN第一届国际中文分词比赛Bakeoff多项第一名

数据分析与预处理工具

http://media.hust.edu.cn

- 中文分词工具 - ICTCLAS
  - 功能
    - 中文分词
    - 词性标注
    - 中文命名实体识别：人名、地名、机构名

数据分析与预处理工具

http://media.hust.edu.cn

- 中文分词工具 - ICTCLAS
  - API
    - 初始化：
      - boolICTCLAS_Init()
    - 运行：
      - ICTCLAS_ParagraphProcess(char*sParagraph,char*sResult)
    - 退出：
      - boolICTCLAS_Exit()
    - 配置文件
      - Configure.xml

数据分析与预处理工具

http://media.hust.edu.cn

- 词根还原库 - Snowball
  - 用于IR的词根还原库
  - 性质：开源
  - 作者：Martin Porter
  - 网址：http://snowball.tartarus.org/index.php
  - 平台：Windows/Linux

数据分析与预处理工具

http://media.hust.edu.cn

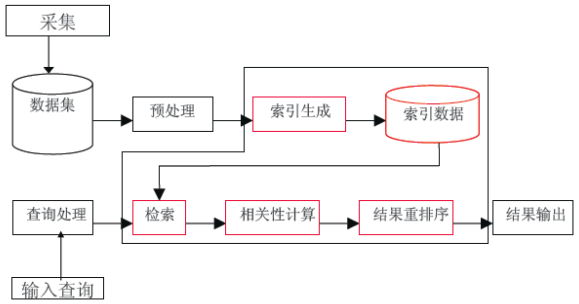
- 词根还原库 - Snowball
  - API
    - 初始化：
      structsb_stemmer* sb_stemmer_new(constchar * algorithm, const char * charenc)
    - 运行
      const sb_symbol* sb_stemmer_stem(structsb_stemmer* stemmer, const sb_symbol* word, intsize)
    - 退出
      void sb_stemmer_delete(structsb_stemmer* stemmer)

Outline

http://media.hust.edu.cn

全文搜索

http://media.hust.edu.cn



华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

- Lucene简介
  - 完整、高效、易用、可扩展的IR库
  - 性质：Apache License
  - 作者：Doug Cutting
  - 网址：http://lucene.apache.org/
  - 版本：LuceneJava 2.2
  - 平台：跨平台
  - 支持：Apache Jakarta项目

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

- Lucene特点
  - 基于Java实现
  - 可扩展、高性能索引工具：
  - 动态内存<=1M
  - 增量式索引
  - 索引数据大小≈(20%~30%)×源数据大小

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

- Lucene功能
  - 结果排序 - 最好结果优先
  - 强大的查询表达式处理功能 - 短语、通配符、模糊查询等
  - 字段检索（如作者、标题、内容等）
  - 指定日期范围检索
  - 分字段排序
  - 支持多索引检索与结果合并
  - 支持更新与检索并发执行

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

- Lucene的其他语言版本
  - CLucene - Lucene implementation in C++
  - dotLucene - Lucene implementation in .NET
  - Lucene4c - Lucene implementation in C **closed**
  - LuceneKit - Lucene implementation in Objective-C (Cocoa/GNUstep support)
  - Lupy - Lucene implementation in Python
  - NLucene - another Lucene implementation in .NET (out of date)
  - Zend Search - Lucene implementation in the Zend Framework for PHP
  - Plucene - Lucene implementation in Perl
  - KinoSearch - a new Lucene implementation in Perl
  - PyLucene - GCJ-compiled version of Java Lucene integrated with Python via SWIG
  - MUTIS - Lucene implementation in Delphi
  - Ferret - Lucene implementation in Ruby

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

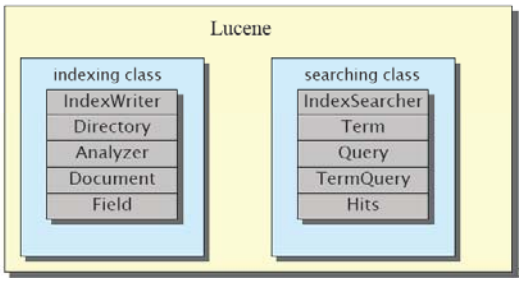
- Clucene简介
  - Lucene的C++实现
  - 作者：Ben van Klinken
  - 协议：LGPL, Apache License 2.0
  - 网址：http://sourceforge.net/projects/clucene/
  - 平台：Windows/Linux/FreeBSD/...
  - 版本：0.9.20(clucene-core-stable)
  - 与Lucene接口基本一致
  - 速度快于Lucene

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

Lucene结构



全文检索工具

http://media.hust.edu.cn

- Lucene的核心接口
  - 索引器
    - IndexWriter：创建与更新索引数据
    - Directory：获取Lucene索引的位置
    - Analyzer：从文本中提取要建索引的符号
    - Document：字段(Field)的集合，可视为虚拟的文档（网页、Email或文本）
    - Field：每个Document由一个或多个Field组成，每个Field对应于可进行标引和检索的一部分数据，即每个数据单元可以用(Field, Value)表示。

全文检索工具

http://media.hust.edu.cn

简单示例—索引

[indexing implementation]

```
void IndexFiles(char* source_data_path, char* index_storage_path)
{
    StandardAnalyzer analyzer;
    IndexWriter* writer = new IndexWriter(path, &analyzer);
    indexDocs(writer, index_storage_path);
    writer->optimize();
    writer->close();
    delete writer;
}// IndexFiles
```


全文检索工具

http://media.hust.edu.cn

简单示例—索引

[indexing implementation]

```
void indexDocs(IndexWriter* writer, char* directory)
{
  DIR* dir = opendir(directory);
  while (dir != NULL)
  {
    if (dir->isdirectory())
    {
        indexDocs(writer,dir->GetFullPath());
    }else{
        Document* doc = new Document();
        doctxt = GetDocText(doc);
        doc->add( *Field::UnIndexed(_T("path"), dir->GetFullPath()) );
        doc->add( *Field::Text(_T("contents"), doctxt) );
        writer->addDocument(doc);
    }//if
    dir->readdir();
  }//while
  closedir(dir);
}// indexDocs
```


全文检索工具

http://media.hust.edu.cn

Lucene核心接口

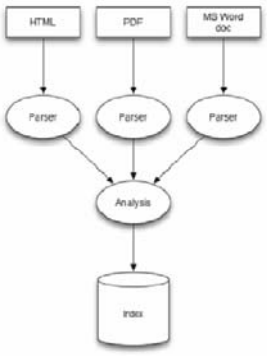
Field的类型

Field method/type	Analyzed	Indexed	Stored	Example usage
Field.Keyword(String, String) Field.Keyword(String, Date)		✔	✔	Telephone and Social Security numbers, URLs, personal names Dates
Field.UnIndexed(String, String)			✔	Document type (PDF, HTML, and so on), if not used as a search criteria
Field.UnStored(String, String)	✔	✔		Document titles and content
Field.Text(String, String)	✔	✔	✔	Document titles and content
Field.Text(String, Reader)	✔	✔		Document titles and content

全文检索工具

http://media.hust.edu.cn

- 中文检索问题
  - 标引过程



全文检索工具

http://media.hust.edu.cn

- 中文检索问题
  - 标引核心接口
    - Analyzer ⇒ChineseAnalyzer
      - TokenStream* tokenStream(constTCHAR* fieldName, Reader* reader)
    - Tokenizer
    - Filter

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

ChineseAnalyzer的实现

```
TokenStream* ChineseAnalyzer::tokenStream(const TCHAR* fieldName, Reader* reader)
{
        reader->read(pBuf,iTxtLen);
        char* seg;

        try
        {
                ICTCLAS_ParagraphProcess(buf,seg);
        }
        catch (...)
        {
                cout<<"segment error!\n";
        }

        this->tmpReader=_CLNEW StringReader(seg);

        return _CLNEW WhitespaceTokenizer(this->tmpReader);
}
```

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

- Lucene的核心接口
  - 检索器
    - IndexSearcher：打开并在索引数据中查找
    - Term：用于检索的基本单位，形式为(Field，Value)
    - Query：检索表达式类型
    - TermQuery：最基本的Query类型，返回在指定字段包含指定关键词的文档
    - Hits：检索结果集

华中科技大学数字媒体实验室

全文检索工具

http://media.hust.edu.cn

简单示例—检索

[Searching implementation]

```
void SearchFiles(const char* path, char* userquery)
{
  IndexReader* reader = new IndexReader(path);
  StandardAnalyzer analyzer;
  Query* q = QueryParser::parse(userquery,_T("contents"),&analyzer);
  Hits* h = s->search(q);
  for (int i=0; i<h->length(); i++)
  {
    Document* doc = h->doc(i);
    printf("Result%d: %s\n", i, doc->get("path"));
  }//for
  delete reader;
}// SearchFiles
```

华中科技大学数字媒体实验室

Outline

http://media.hust.edu.cn

华中科技大学数字媒体实验室

小型搜索引擎的设计与实现

http://media.hust.edu.cn

- 运行采集器
- 构造索引器
- 构造检索服务器
- 构造用户交互接口

华中科技大学数字媒体实验室

小型搜索引擎的设计与实现

http://media.hust.edu.cn

- 运行采集器
  - 典型设置

```
[larbin.conf]
From luoweihua@ict.ac.cn
UserAgent larbin_2.6.3
#noExternalLinks
waitDuration 30
depthInSite 5
startUrl http://www.sina.com.cn/
```


小型搜索引擎的设计与实现

http://media.hust.edu.cn

- 运行采集器
  - 典型设置

```
[options.h]
#define SIMPLE_SAVE
#define FOLLOW_LINKS
#define NO_DUP
#define MAXBANDWIDTH 200000
[types.h]
#define saveDir "save/"
```


小型搜索引擎的设计与实现

http://media.hust.edu.cn

- 运行采集器
  - ./configure
  - gmake
  - ./larbin

小型搜索引擎的设计与实现

http://media.hust.edu.cn

- 构造索引器

```
❶Preprocess_Webpages(Web_Page_Path, New_Data_Path, DocId);
❷Build_Index_By_Lucene_IndexWriter(New_Data_Path);
```


小型搜索引擎的设计与实现

http://media.hust.edu.cn

- 数据预处理

```
void Preprocess_Webpages(string Web_Page_Path, string New_Data_Path, unit& DocId)
{
  DIR* dir = opendir(Web_Page_Path);
  dirent* de = readdir(dir);
  while (de != NULL)
  {
    if (de.isdirectory())
    {
        Preprocess_Webpages(de.getfullpath(), New_Data_Path, DocId);
    }else{
        Preprocess_EachPage(de.getfullpath(),New_Data_Path, DocId);
    }//if
    de = readdir(dir);
  }//while
  closedir(dir);
}// Preprocess_Webpages
```


小型搜索引擎的设计与实现

http://media.hust.edu.cn

- 数据预处理

```
Preprocess_EachPage(string Page_Path,string New_Data_Path, unit& DocId)
{
  int Type = GetFileType(Page_Path);
  switch(Type)
  {
    case HTML: Htm2Txt(GetFileText(Page_Path), NewText);
               break;
    case PDF: Pdf2Txt(GetFileText(Page_Path), NewText);
               break;
    case DOC: Doc2Txt(GetFileText(Page_Path), NewText);
               break;
    default: break;
  }//switch
  string SegText;
  ICTCLAS_Paragraph(NewText, SegText);
  SnowBall(SegText);
  StoreResult(SegText, DocId++);
}// Preprocess_Webpages
```

小型搜索引擎的设计与实现

- 构造检索服务器

```
void SearchServer()
{
    InitServer();
    StartSocket(server_socket);
    while (true)
    {
        socket client_socket = accept(server_socket);
        beginthread(Search_Daemon, client_socket);
    }//while
    CloseServer();
}//SearchServer
```

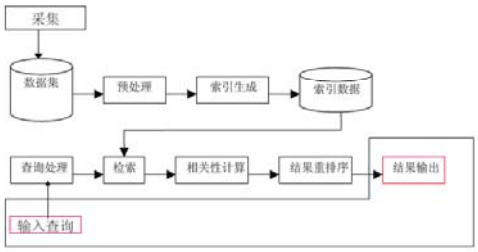

小型搜索引擎的设计与实现

- 检索

```
void Search_Daemon(socket client_socket)
{
        GetUserQueryBySocket(buffer);
        DecodeUserQuery(buffer, UserQuery);
        Search_Result_By_Lucene(UserQuery, AllHits);
        EncapsulateResult(AllHits, Result_String);
        ReturnResult(Result_String, client_socket);
}// Search_Daemon
```


小型搜索引擎的设计与实现

用户交互接口



构造用户接口

- 用户接口
  - Web浏览器界面
  - 用户查询请求处理程序

构造用户接口

- Web浏览器界面
  - 输入查询表达式

构造用户接口

- Web浏览器界面
  - 结果输出

构造用户接口

http://media.hust.edu.cn

- 用户查询请求处理程序
  - CGI或脚本语言（ASP,PHP,JSP,etc）
  - 功能
    - 获取用户查询式：把用户通过Form输入的查询语句封装发送给检索服务器
    - 显示结果：从检索服务器获取结果，缓存并分页呈现给用户

华中科技大学数字媒体实验室

构造用户接口

http://media.hust.edu.cn

❶获取用户查询

```
$Query_and = $_Get["query_and"];
$Query_or = $_Get["query_or"];
$Query_not = $_Get["query_not"];
$Total_Query = GenerateTotalQuery($Query_and,$Query_or,$Query_not);
$ClientSock = new ClientSocket($Address, $Port);
$ClientSock->SendToServer($Total_Query);
```

华中科技大学数字媒体实验室

构造用户接口

http://media.hust.edu.cn

❷显示结果

```
$Result_String = $ClientSock->receive();
$AllResults = DecodeResult($Result_String);
Foreach ($AllResults as $EachResult)
{
    Echo "DocId : " . $EachResult->docid . "<br>";
    Echo "DocPath:" . $EachResult->docpath . "<br>";
}//Foreach
```

华中科技大学数字媒体实验室

Outline

http://media.hust.edu.cn

华中科技大学数字媒体实验室

实用化问题

http://media.hust.edu.cn

- 系统架构
- 数据更新策略
- 数据去重
- 格式识别与转换
- 内码识别与转换
- 有针对性的优化策略
- 结果排序与优化
- 负载均衡
- 缓存策略
- ……

华中科技大学数字媒体实验室

Outline

http://media.hust.edu.cn

华中科技大学数字媒体实验室

参考文献

http://media.hust.edu.cn

- Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. http://www-db.stanford.edu/~backrub/google.html
- Lan Huang. A Survey on Web Information Retrieval Technologies. http://citeseer.ist.psu.edu/336617.html
- cneie.com. 中文搜索引擎技术揭密：系统架构. http://www.cneie.cn/news/20064/20064291616555.html
- 卢亮. Larbin 一种高效的搜索引擎爬虫工具. http://www.wespoke.com/archives/0008/9.php
- Martin Porter. snowball manual. http://snowball.tartarus.org/compiler/snowman.html
- Doug Cutting, Otis Gospodnetic , etc. Overview of Apache Lucene. http://lucene.apache.org/java/docs/
- ERIK HATCHER,OTIS GOSPODNETIC. Lucene in Action. Manning Publications Co. Greenwich CT.
- 李晓明，闫宏飞，王继民. 搜索引擎－原理、技术与系统. 北京:科学出版社,2005

Project Requirement

http://media.hust.edu.cn

- 请利用本课程介绍的工具或自选工具，设计一个小型的搜索引擎实验系统
- 具体要求：
  - 以同一导师或同一课题组分组，每组4-6人，其中选一个人作为组长
  - 本项目占总分的50%
  - 要求提交如下内容
    - 设计文档和测试报告
    - 程序代码和演示PPT
  - 方案汇报：2008年12月9日，汇报时间：15分钟
  - 完成时间：2009年1月5日以前

Project 1：基于搜索引擎架构的视频特征提取

http://media.hust.edu.cn

- 功能要求
  - 视频文件的搜索与分类
  - 在线提取视频特征（以某一种特征为例）
  - 按照mpeg-7对特征进行描述
- 性能要求
  - 如何提高特征提取的速度
  - 如何处理“断点续传”的问题
- 人员需要
  - 周玲（组长）、张强、余华飞、张宝印、李松

Project 2：基于哼唱的MP3搜索引擎

http://media.hust.edu.cn

- 功能要求
  - MP3文件的搜索
  - MP3文件按照元数据(参考iTune)进行分类组织
  - 离线提取音频特征（以某一种特征为例）
  - 实现基于哼唱的搜索
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
  - 马尧（组长）、王淞、邓珂、张勇、刘吉鹏、马红欣、戈峰

Project 3：基于颜色特征的图像搜索引擎

http://media.hust.edu.cn

- 功能要求
  - 图像文件的搜索
  - 图像文件的分类（文件类型、分辨率、创作时间等）组织
  - 离线提取颜色特征（以某一种颜色特征为例）
  - 实现颜色特征的图像搜索
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
  - 戴振（组长）、李振环、文耀光、陈勇飞、周渊、殷实

Project 4：基于内容的电视节目预告搜索引擎

http://media.hust.edu.cn

- 功能要求
  - 实现对国内所有电视台节目预告的搜索
  - 按照内容分类组织节目预告，建立统一的节目预告表
  - 实现基于节目内容的预告搜索，并提供提醒服务
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
  - 张富强（组长）、陈萌、胡春龙、刘浩、刘锐

Project 5：高校跳蚤市场商品搜索引擎

http://media.hust.edu.cn

- 功能要求
  - 实现对省内高校跳蚤市场商品的搜索
  - 同一商品按价格、质量等信息排序
  - 实现对商品的图片浏览功能
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
  - 廖红虹（组长）、李晓振、向杰、孙自龙、黎单、徐姗、王洁

华中科技大学数字媒体实验室

Project 6：基于内容的图书搜索引擎（暂定）

http://media.hust.edu.cn

- 功能要求
  - 实现对中文图书和书评的搜索
  - 按照元数据对图书进行分类(书名、作者、出版社、出版日期等)组织
  - 实现基于内容的图书搜索和书评搜索
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
  - 高一鸣（组长）、王健、吕文松、方红霞、何珂程、周苗苗

华中科技大学数字媒体实验室

Project 7：基于内容的电影搜索引擎

http://media.hust.edu.cn

- 功能要求
  - 实现对武汉搜索影院电影预告和相关电影影评的搜索
  - 按照内容分类组织电影预告，建立统一的电影预告表
  - 实现基于电影内容的预告和影评搜索，并提供提醒服务
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
  - 黄进东（组长）、谢莹、朱传聪、胡吉柱、肖骁、张深毅

华中科技大学数字媒体实验室

Project 8：基于内容的视频点播搜索引擎

http://media.hust.edu.cn

- 功能要求
  - 实现对国内著名视频点播网站的搜索
  - 按照内容分类（文件格式、片名、长度、导演、男女主角等等）组织各大网站的视频节目
  - 实时监测测用户的网络状况，实现基于不同的网络状况提供不同的文件类型
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
  - 闻乃松（组长）、王珊珊、程格平、江兴才

华中科技大学数字媒体实验室

Project 9：基于内容的课件搜索引擎

http://media.hust.edu.cn

- 功能要求
  - 实现对国内高校上网课件的搜索
  - 课件类型包括：PPT、视频
  - 按照课程内容实现对课件的分类组织（名称、主讲教师、开课单位等）
- 为该搜索引擎设计一种以上的运营模式
- 人员需要
- XX

华中科技大学数字媒体实验室