信息检索导论--
lucene

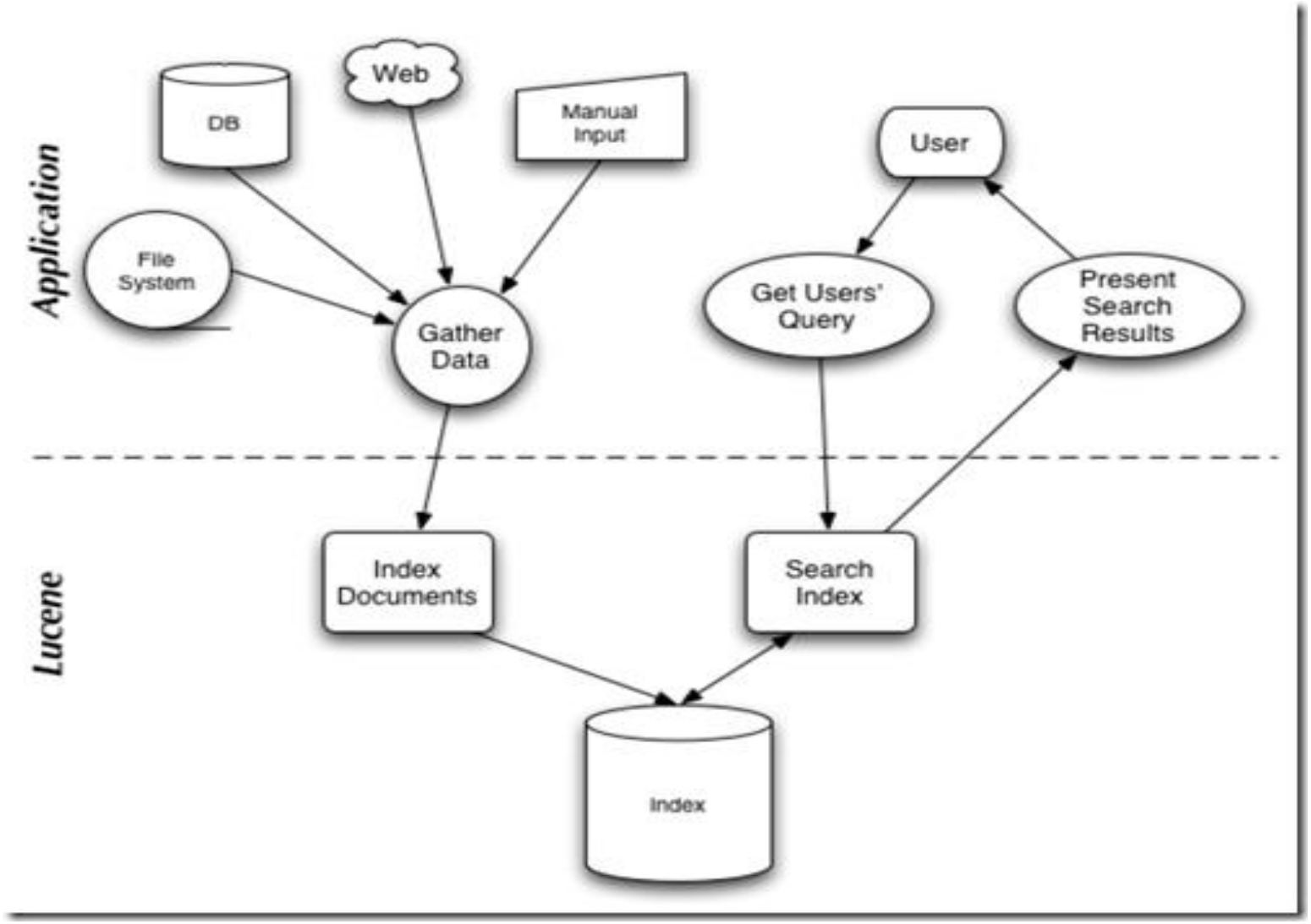
授课人：李永可
邮箱：lyk@xjau.edu.cn

欢迎进入lucene世界

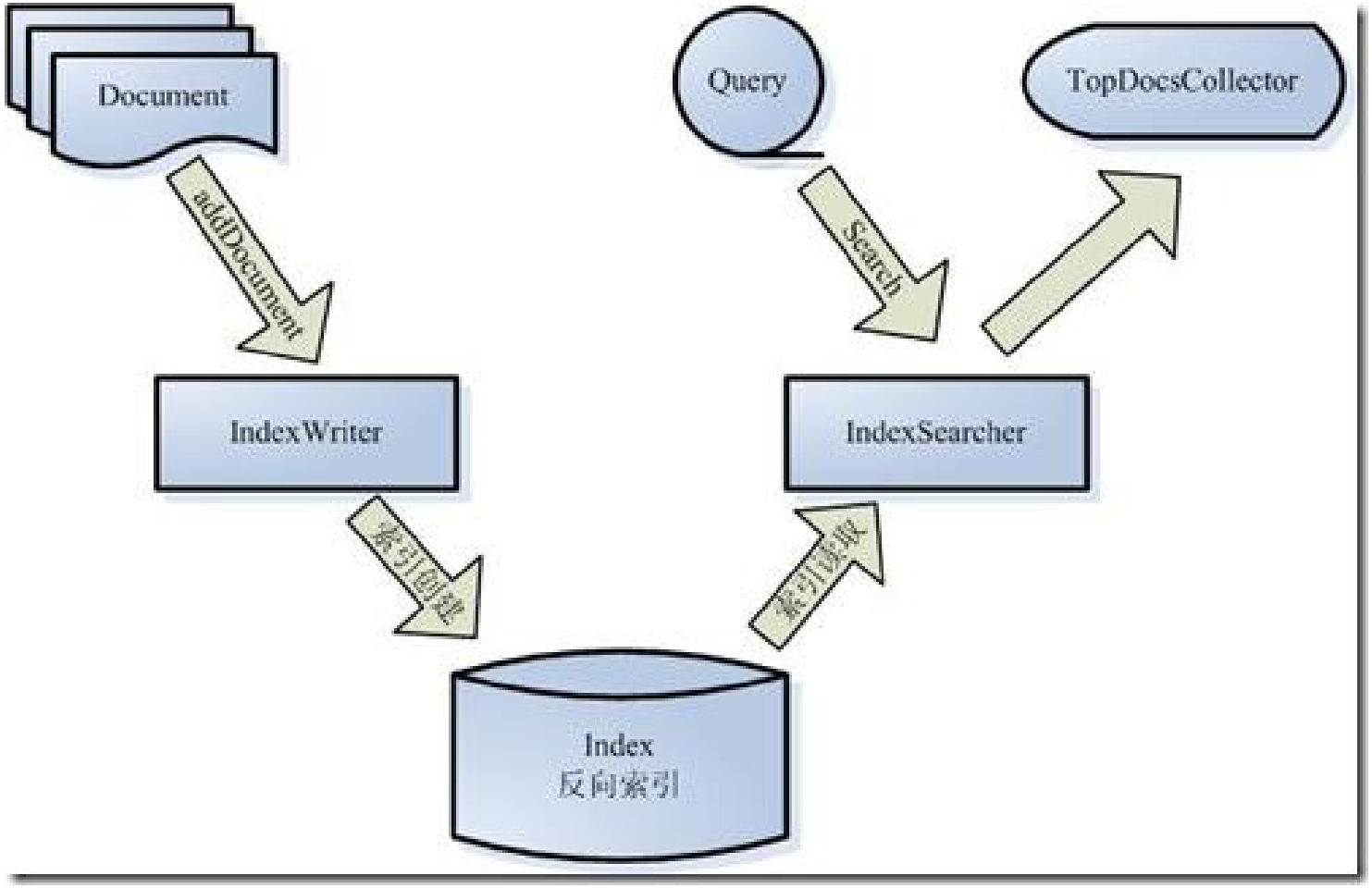
lucene

- 一个高效的，可扩展的，全文检索库。
- 全部用Java实现，无须配置。
- 仅支持纯文本文件的索引(Indexing)和搜索(Search)。
- 不负责由其他格式的文件抽取纯文本文件，或从网络中抓取文件的过程。

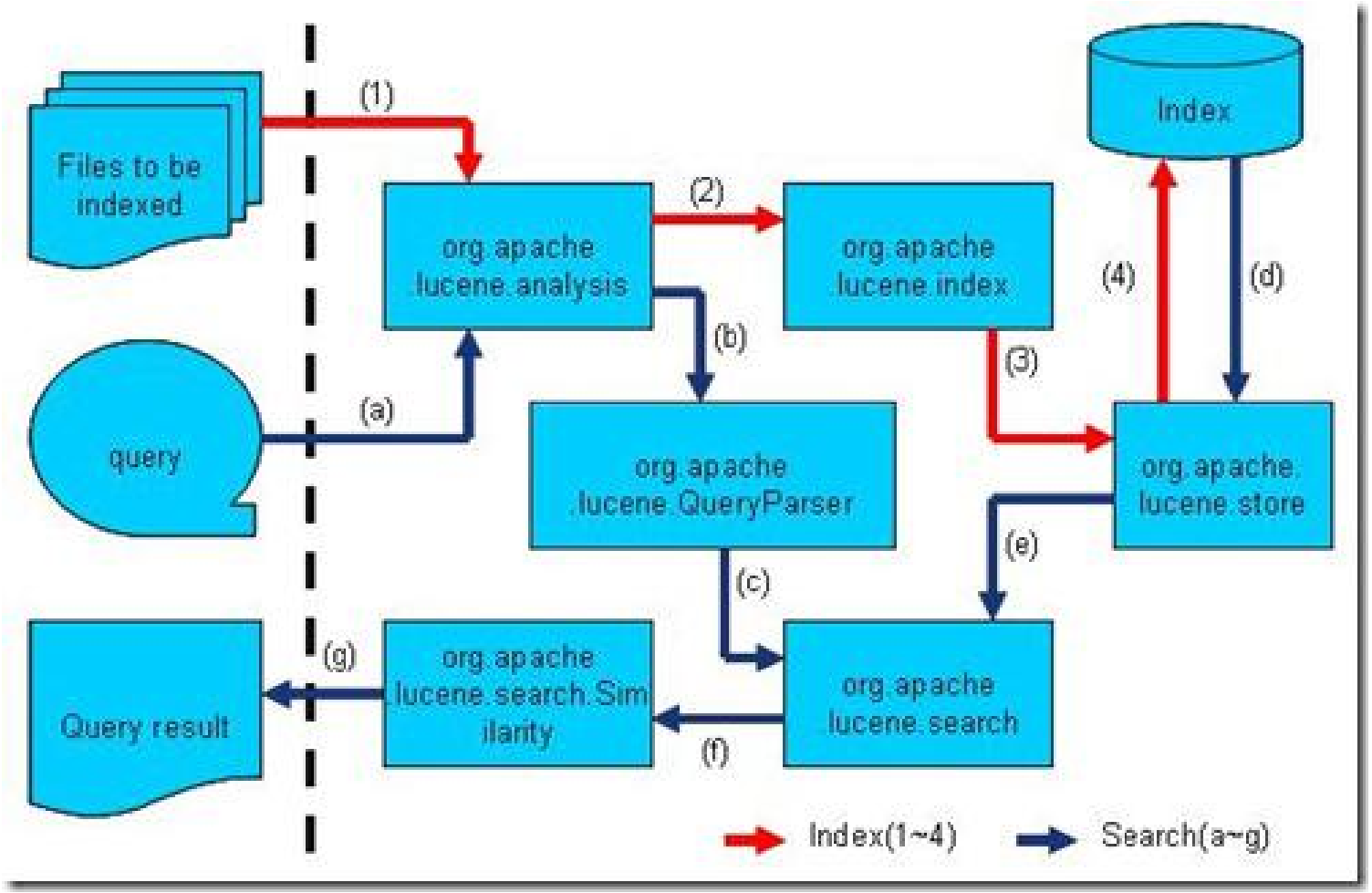
Lucene架构和过程



Lucene各个组件


Document
addDocument
IndexWriter
索引创建
Index
反向索引
索引读取
IndexSearcher
Search
Query
TopDocsCollector

lucene包结构


Files to be indexed
(1)
org.apache.lucene.analysis
(2)
org.apache.lucene.index
Index
(4)
(d)
query
(a)
(b)
org.apache.lucene.QueryParser
(3)
org.apache.lucene.store
(e)
(c)
Query result
(g)
org.apache.lucene.search.Similarity
org.apache.lucene.search
(f)
Index(1~4)
Search(a~g)

- Lucene的analysis模块主要负责词法分析及语言处理而形成Term。
- Lucene的index模块主要负责索引的创建，里面有IndexWriter。
- Lucene的store模块主要负责索引的读写。
- Lucene的QueryParser主要负责语法分析。
- Lucene的search模块主要负责对索引的搜索。
- Lucene的similarity模块主要负责对相关性打分的实现。

IndexWriter

- **Lucene3.1以前版本**

  IndexWriter(Directory d, Analyzer a, boolean create, IndexWriter.MaxFieldLength mfl)

- **Lucene3.1以后版本**
- **IndexWriter**(Directory d,IndexWriterConfig conf)

Directory

- 索引文件存放路径
- FSDirectory fdir=FSDirectory.open(new File(filepath));

Analyzer

- 索引索引分词器
- Analyzer ik=new IKAnalyzer();

create

- 若create为true则将增量更新索引，如果为false则将删除原来的索引建立新索引。

IndexWriter.MaxFieldLength

- Limited限制索引长度
- Unlimited不限制索引长度

IndexWriterConfig

- **IndexWriterConfig**(Version matchVersion, Analyzer analyzer)
  IndexWriterConfig conf=new IndexWriterConfig(Version.lucene_35,ikanalyzer)

IndexWriter初始化

- FSDirectory fdir=FSDirectory.open(new File(filepath));
- Analzer ik=new IKAnalyzer();
- IndexWriter writer=new IndexWriter(fdir,ik,true, IndexWriter.MaxFieldLength.unlimited);

文档Document

- 文档是我们建索引的基本单位，不同的文档是保存在不同的段中的，一个段可以包含多篇文档。
- 新添加的文档是单独保存在一个新生成的段中，随着段的合并，不同的文档合并到同一个段中。
- Document doc=new Document();

域Field

- 一篇文档包含不同类型的信息，可以分开索引，比如标题，时间，正文，作者等，都可以保存在不同的域里。
- 不同域的索引方式可以不同
- **Field**(String name, String value, Field.Store store, Field.Index index)
- Field f1=new Field
  （’title’,title,Field.Store.YES,field.Index.ANALYZED）

创建索引完整过程

- FSDirectory fdir=new FSDirectory(new File(path));
- Analyzer ik=new IKAnalyzer();
- **IndexWriterConfig conf=new IndexWriterConfig(Version.lucene_35,ik);**
- **IndexWriter writer=new IndexWriter(fdir,conf);**
- **Document doc=new Document();**

创建索引完整过程

- Field f1=new Field(“title”,title,Field.Store.YES,Field.Index.ANALYZED);
- Field f2=new Field(“content”,content,Field.Store.YES,Field.Index.ANALYZED);
- doc.add(f1);
- doc.add(f2);
- writer.addDocument(doc);
- writer.close();