Topic: Lucene

Presenter
Ken Hoang

History

- Doug Cutting
- Lucene (Doug's wife middle name)
- Search software
- Start writing in late 1997 using Java
- 2000 – SourceForge (open source)
- 2001 - Apache adopted Lucene
- 2004 - widely used by web search engines

Why Lucene was written.

- The explosion of the Internet
- Too much information in storage
- Inefficient if crawl hundreds of thousands web pages.

Solution

- Lucene was created
  - more dynamic ways of finding information
  - high performance
  - scalable Information Retrieval(IR) library
  - open-source project implement in Java
  - under Apache Jakarta License

What Lucene is

- A software library
- Not full-featured search application
- Full–featured search applications built on top of Lucene

What Lucene can do

- Allows to add indexing and searching capabilities to applications
- Text indexing
- Make searchable data converted to a textual format
- Fast random access to words

What Lucene can do (cont')

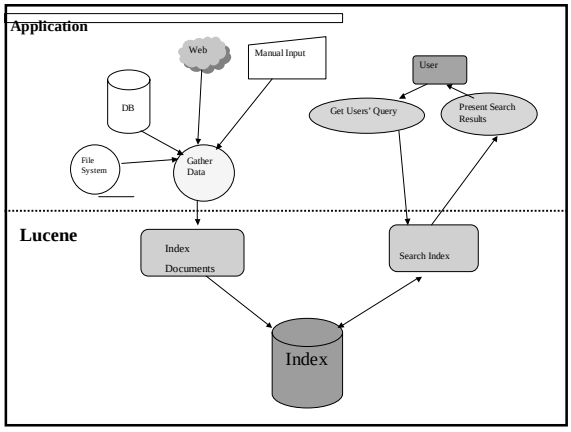
- Search words in
  - Web pages on remote servers
  - Documents stored in local files systems
  - Simple text files
  - MS Word documents
  - HTML or PDF files

Lucene API

- A handful of classes
- Indexing
  - IndexWriter
  - Directory
  - Analyzer
  - Document
  - Field

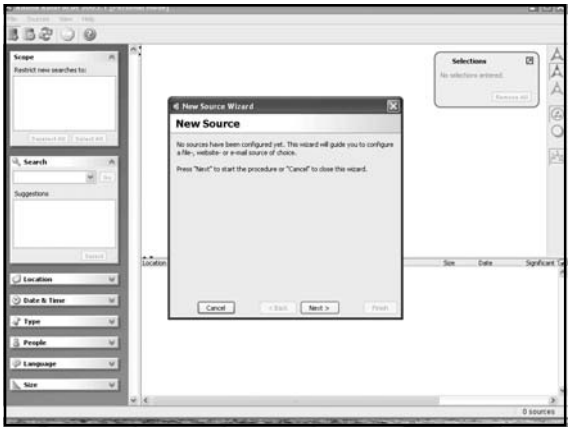
Lucene API (cont')

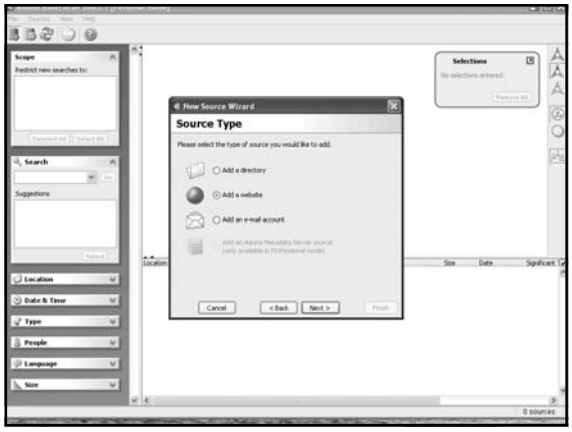
- Analysis
  - converting field text into terms (index representation)
- http://lucene.apache.org/java/docs/api/overview-summary.html
- http://lucene.apache.org/java/docs/

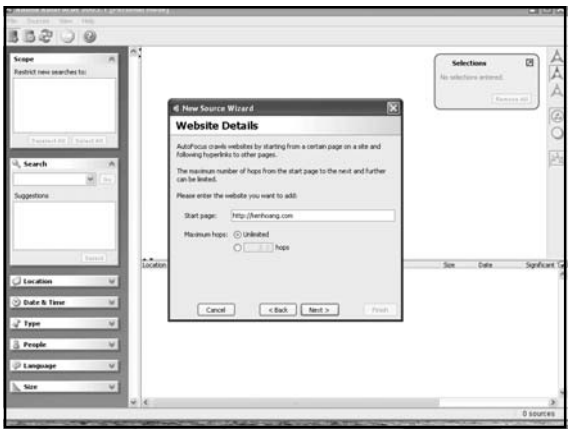


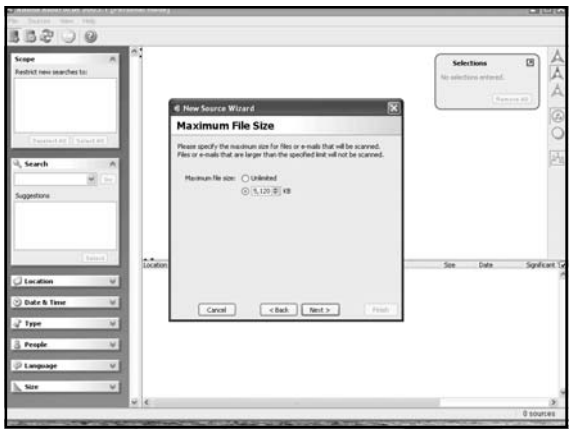
Application

- Autofocus

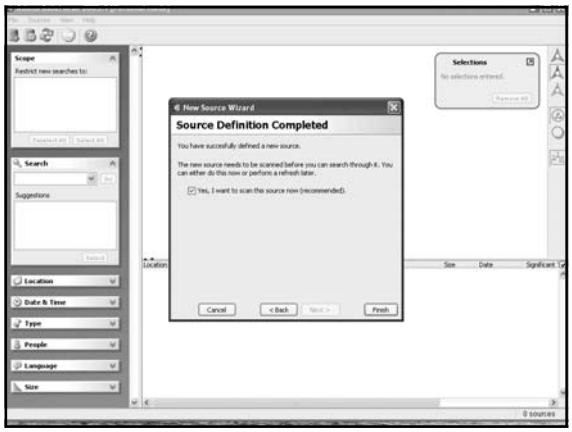

New Source Wizard
New Source
No sources have been configured yet. This wizard will guide you to configure a file-, website- or e-mail source of choice.
Press "Next" to start the procedure or "Cancel" to close this wizard.
Cancel
< Back
Next >
Finish


New Source Wizard
Source Type
Please select the type of source you would like to add.
Add a directory
Add a website
Add an e-mail account
Cancel
< Back
Next >
Finish


New Source Wizard
Website Details
Start page:
Maximum hops:
Unlimited
hops
Cancel
< Back
Next >
Finish

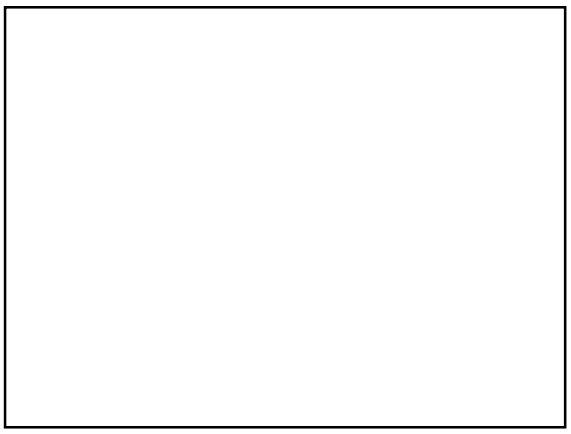

New Source Wizard
Maximum File Size
Cancel
< Back
Next >
Finish

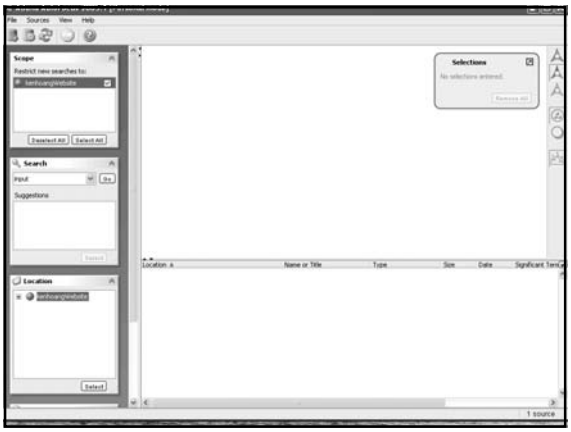

New Source Wizard
Source Name
Name:
Cancel
< Back
Next >
Finish

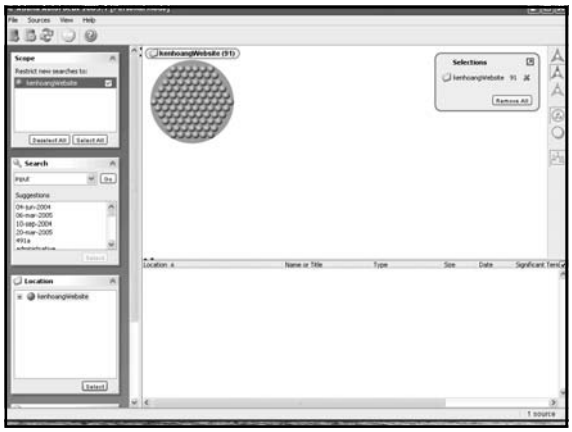

New Source Wizard
Source Definition Completed
You have succesfully defined a new source.
Cancel
< Back
Finish

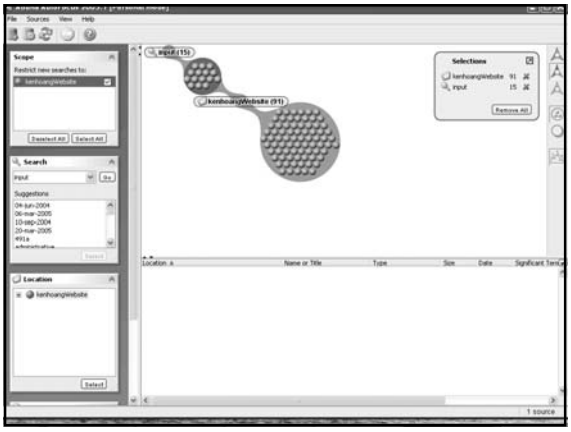

Scanning Progress
Scanning source "kenhoangWebsite"
Stop scanning
Close

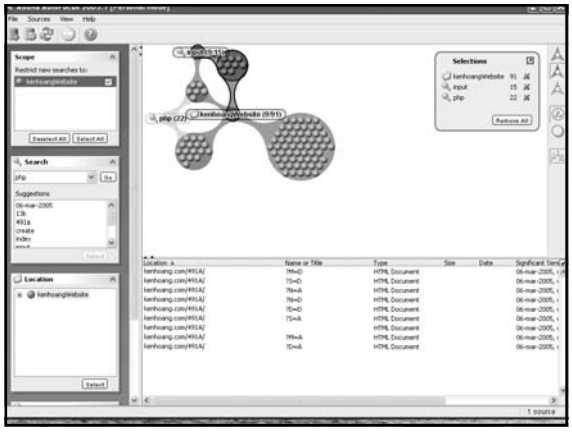

Scanning Progress
Scan completed
Close

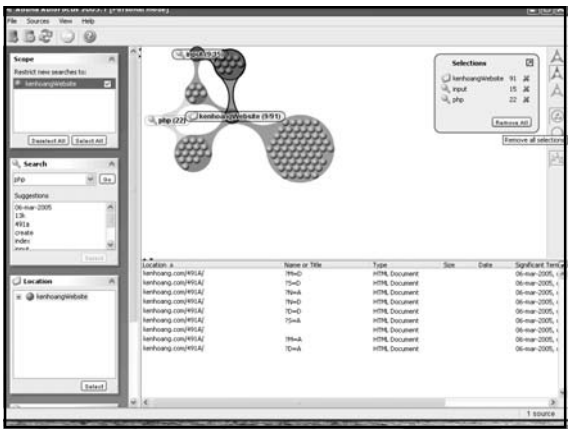


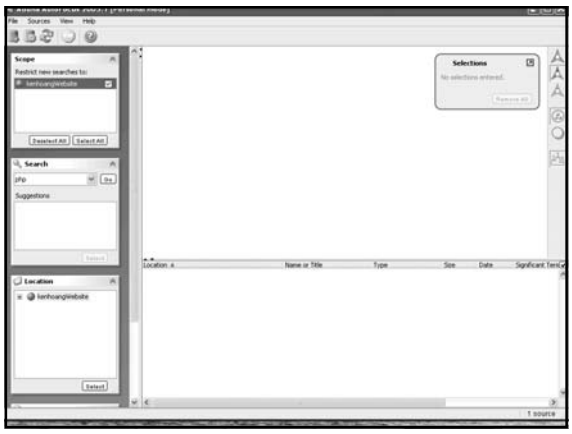












Question?