IBM WebSphere Portal and Lotus Workplace technical symposium
Session Number: B0F2
Session Title: Text Search and Portal Integration
Speaker's e-mail: ayas@il.ibm.com

Aya Soffer, Manager, Search Technologies Dept.
Agenda

- Websphere Portal Search Engine (PSE)
  - Overview and Architecture
  - Main functions
  - Usage Examples and Planning Guidelines
  - Common Components: Lotus Workplace Search
- Demo
- Information Resources
- Q & A
What is the Portal Search Engine? (PSE)

- High level functional overview

- **Administrator**: indexing / collecting content/documents
  - HTTP crawler
  - Indexer component
  - Text analysis functions (taxonomy, categorizer, language tools, summarizer)
  - Simple workflow to control what and how gets indexed

- **End-user**: search
  - web-style search
  - high precision relevance ranking
  - browse through the collection
General information

- Originally developed by IBM Research in Israel
  - Proven technology base with emphasis on search quality
  - Backed by the joint Research and Software group program – Institute for Search and Text Analysis

- Fulltext search technology
  - 100% pure Java implementation
  - Suitable for server as well as client environments
  - Emphasis on highly accurate results - constantly benchmarking and evaluating via official forums such as TREC and INEX
  - Internal interfaces allow for convenient integration in IBM products and solutions
    - Rich set of APIs suitable for simple and complex implementations
    - Easy to customize and extend - Adapt ranking formulas, extend built-in methods, add new document types

- IBM strategic component
Portal Search Engine – where used ....

- Portal Search Engine portlet application:
  - Administer multiple indexes (collections), where each may include multiple sites
  - End-user search portlet for both handling search requests and browsing through the documents in the collection

- Integrated with Portal Document Manager (PDM)
- Integrated with Lotus Workplace 1.1
- Integrated with WebSphere Portal Content Publisher (WPCP)
New key features with Websphere Portal Version 5

- **Taxonomies and categorization**
  - A taxonomy is a hierarchical representation of a set of categories
  - It includes rules per category that are applied to a document through a categorizer
  - Two types of taxonomies available
    - A *pre-defined* taxonomy allowing for simple manipulation (like renaming of categories and definition of new categories)
    - A *rules based* taxonomy which can be built and defined by the user
  - Categorization – process of assigning a document to category(-ies)

- **Summarization**
  - the top ‘3’ key sentences are extracted
  - “the first ‘250’ characters of text” used for CJK and BiDi type languages

- **Document filters**
  - Supports >250 document formats
  - Technology wrapped into the ‘document conversion services’ (DCS) which add support for additional document formats
Conceptual overview – Index build process

Filter

Crawler

Content

Text analysis Components:
• Categorizer
• Summarization
• Document filters

Approval Workflow

Indexer

Collection

1. Metadata injected into original content
2. Approved set of Content “In-basket”
Creating a new document collection is easy

Select the ‘Manage search collections’ portlet

- Create a **new collection**
- Specify a **web site** to collect information/content from
- Click on ‘Start collecting’ icon/text to initiate the index build process

- Processing status and status of the index are shown at the bottom of the portlet, for:
  - the selected site
  - the selected collection (index)
A look at the Manage Collections portlet
Manage Collections Portlet – Options and Status

PSE Manage Indexes

Document Collections

Sites in Collection: c:\indexes\test-rule-based

Site Update Schedule:
2003.07.15 at 11:35 every 1 week

Site Status
Running: No
Number of documents collected: 178
Run time: 0

Collection Status
Last update completed: 2003.07.15 at 04:06
Next update scheduled: 2003.07.22 at 11:35
Number of Active documents: 333
Number of deleted documents: 28

Collection name and location: c:\indexes\test-rule-based
Collection Language: English
Categorizer used: User-Defined
Summarizer used: None

Select ‘Portal Settings’

Browse Documents
End user – Search portlet – detailed view
Administration – advanced options

Portlet for defining a new collection

![Portlet for defining a new collection](image-url)
Administration – advanced options

Portlet for defining a new site
Administration – advanced options

Portlet for defining a schedule for periodic indexing of a site

```
Define Schedule

From: 2003 Jan 01
At: 01 00
Update every: 1 Hour(s)

Create

Scheduled Updates

Delete
```

Cancel
Administration – advanced options

Portlet for defining filters for sites

**PSE Manage Indexes**

*Collection - ci\indexes\test-rule-based  Site: http://w3.haifa.ibm.com/*

Create a rule and click Create to add it to the list on the right.

**Filter Rules**

<table>
<thead>
<tr>
<th>Rule Name:</th>
<th>Apply rule while:</th>
<th>Rule type:</th>
<th>Rule basis:</th>
<th>File Types</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Collecting documents</td>
<td>Exclude</td>
<td>URL text</td>
<td>All</td>
</tr>
</tbody>
</table>

**Filtering Rules**

**While collecting documents:**

- Exclude Filtering Rules
- Include Filtering Rules

**While adding documents to index:**

- Exclude Filtering Rules
- Include Filtering Rules

**Cancel**
Administration – advanced options

Portlet for defining destination categories for the site

![Diagram showing a category tree for defining destination categories.](image)
Administration – advanced options

‘Browse document’ portlet
Administration – advanced options

‘Search’ portlet

PSE Manage Indexes

Search Results

Browse Documents

Manage Indexes

Collection name - c:\indexes\test-rule-based

Search whole document for

For: text search

Search In:

Title
For:

Select document type:

All

Search Under:

All Categories

Hide Advanced Options

Clear Search Results

Results for: text search

100 results found, sorted by relevance

Sort by Date

Hide Details

1 - 10 of 100

Jury - Generic Full-text Java Search Engine | Homepage

Category -root->servlets

Category -root->http

Category -root->haifa

Software and Services : Knowledge Management
http://w3.haifa.ibm.com/dept/software/knowledge.html

Author: harry@il.ibm.com

Description: Software and Services : Knowledge Management

Category -root->http

Category -root->haifa

100%
7/16/03

84%
7/16/03
Administration – advanced options

‘Advanced search’ portlet
Usage example

- Goal: provide a community of users with information about competitors in the market

- How: catalog information such as news articles and related information from external websites

- Additional steps to take:
  - When creating a collection, select “User-defined” from the taxonomy pull-down
  - From the main administration portlet choose “Category tree” in the Manage Collections frame
Category Tree portlet

- Build the taxonomy tree
- then go to ‘Manage Rules’ to define rules for each category
What the rule set looks like ..... 

- a ‘rule’ is essentially a search query one would use to find such specific documents
- you can use ‘+’ and ‘-’ and ‘ “ ‘ and ‘*’ within the rule
Last step – assign categories to each website
Result: search and browse

- **Friendster: A little cash goes a long way?** | CNET News.com
  - Description: Friendster: A little cash goes a long way? | The increasingly popular personal networking site collects a million dollars from venture capitalist Amazon, eBay and Yahoo! in hopes of turning a share of their operation into the Web's next big thing. | September 2, 2003, 9:00 PM PT | Paul Perez

- **Intel puts Tri-Gate transistor on fast track** | CNET News.com
  - Description: Intel puts Tri-Gate transistor on fast track | The chipmaker says the future transistor, designed to let electricity flow more freely inside chips, is moving closer to reality. | June 10, 2003, 9:00 PM PT | Michael Hazel

- **Tablet PC: 21st century notepad?** | CNET News.com
  - Description: Tablet PC: 21st century notepad? | The fast-paced and efficient way to take notes, said a conference attendee in Napa, California, this week at the Computerworld Conference. | August 5, 2003, 6:00 PM PT | Christian Ibarra

- **IT services firms eye foreign labor** | CNET News.com
  - Description: IT services firms eye foreign labor | Companies offering information technology outsourcing see lower-cost foreign labor as a key to growth in 2003, market research firm Gartner Datquest says. | January 30, 2003, 9:00 PM PT | Ed Frauenheim
Planning numbers, performance

- Index size information
  - approximately 40% to 60% of the textual content size of indexed documents/pages

- Indexing throughput
  - crawling/indexing rate between 100 to 200 documents per minute

- Search responsiveness
  - typically a search result page is completed and ready for transmission in less than 0.5 seconds
Additional Information and Resources

**IBM Resources:**


- **Websphere Portal Catalog:**

- **Websphere Portal Developer’s Zone**

- **WebSphere Portal Toolkit** -


- **IBM Lotus Workplace**