Information Mining
a Component of EIP V8.1

Stefan Orban
Data Management Development
IBM Lab Böblingen, Germany
Email: stefan_orban@de.ibm.com
EIP Framework

- Solutions
  - Visual and Non-Visual Beans
    - C++, Java
- EIP Services
- Thick Client Visual and non-Visual Beans
- Thin Client Enabler (non-Visual Beans)
- Connector
- Federation
- Workflow
- Information Mining
- DB2
- Browser

- Content Manager
- Databases
- Business Intelligence

IBM Germany 09/05/2002
Text Mining
Finding Information in a Sea of Content
What is Text Mining?

- Information extraction
  - Analysis of textual data to extract information for a particular application

- Knowledge discovery
  - Search, organize, relate, ... documents
How does text mining work?

Treat text as data

- Prose does not have an explicit structure
  - Any structure is implicit
  - *Text Mining uses non-structural features*

- Documents may have some structure
  - Title, abstract, author, ...
  - `<XML>` may be deep or shallow
  - *Text Mining makes limited use of document structure*

- Metadata is highly structured
  - *Text Mining can generate new metadata*
**Text Mining technology**

Automate tasks previously done manually

- Recognize significant items in text: names, technical terms, abbreviations, etc.
- Identify the language of a document
- Group related documents based on their content, without requiring predefined classes
- Extract key information from a document to create a summary
- Assign documents to one or more user-defined categories

Diagram:

- Summarization
- Clustering
- Language Identification
- Feature Extraction
- Categorization
Use EIP Information Mining to ...

- set up an *Information Warehouse* that keeps relevant information about sets of relevant documents
- create the contents of this warehouse automatically using the text analysis functions
  - Language Identification, Summarization, Information Extraction, Categorization, Clustering
- organize your information with different taxonomies created with the Information Structuring Tool
- create applications using Java Beans or Service API to...
  - import documents, apply mining functions, and store the results in the Meta Data Store
  - retrieve information from the Meta Data Store
  - automatically filter information and route it to the appropriate people
Information Structuring Tool Sample
EIP Information Mining

list the summaries of all documents in category X that contain "yz.."

meta data

apply text mining functions

working set of documents

crawl the Web

import from backend

Catalog 1

- business
- government
- science
- arts

Catalog N

- football
- tennis
- baseball
- boxing
| **documentname:** | c:\demodocs\arts\design\design3.html |
| **document:** | 1 / 1 |
| **categories:** | arts/design |
| **language:** | en |
| **summary:** | Moore College of Art and Design – educating women since 1848, Moore is a 4-year degree granting institution, both the professional and fine arts areas. |
| **features:** | Personalize Yahoo,Yahoo,Glasgow Festival of Architecture, Web Sites AltaVista News Stories Net Events Found 8,Find Booksmerchants,Site Matches,Interior Design Institute,Web |
| **title:** | Moore College of Art and Design |
| **author:** | Glasgow Festival of Architecture |
| **content:** | Arts: Design Arts: Graphic Design: Magazines |
| | • Australian Graphic Design Cafe – resource for graphic and web designers, includes articles, tips, messageboard, and database. |
| | • Design Mind – insight, commentary, advice and so... |
Conclusion

Information mining functions add ability to...

- ...collect documents from the Web
- ...create Meta Data catalogs of document sets
- ...automatically assign documents in a catalog to categories
- ...search for documents in categories
- ...automatically summarize documents
Text Mining Sample

Categorization of Newspaper Articles at MediaProfiler, Barcelona, in the context of EIP & Content Manager
Text Mining Sample
The process at MediaProfiler before the change

Newspapers

Scanner

ORACLE DB
- images
- text (OCR)
- categories

Get page image & Cut article & OCR

Manually assign Categories and store in ORACLE DB

Sports
- Soccer
- Scating

Banking
- Stocks
- Credits
- Accounts

Medicine
- Viruses

Use Oracle Context to search for category & key words

Customer A

Customer B

Customer C

IBM Germany 09/05/2002
Objectives of solution

- Automate categorization process
- Use category assignment as preprocessing step
  - allow later manual intervention
- Supply flexible solution
  - handle category changes easily
- Store images, text & meta data of articles in CM
  - meta data will consist of e.g.
    - Date/Time
    - Title of article
    - Originating newspaper
    - category
Text Mining Sample (cont)

The Information Mining solution

Newspapers

Scan / OCR -> Store in CM -> Run search on date and pass result to categorizer -> Add assigned categories to related CM items

CM DB
- date/Time
- title of article
- orig newspaper
- images / text
- categories & conf. level (filled later)

Run parametric search by "date", "category" AND full text search like "IBM" AND "DB2" AND "Oracle"

take probes of returned results, change / send to customer

Customer A
Customer B
Customer C

Soccer
Scating
Stocks
Credits
Accounts
Viruses

IBM Germany    09/05/2002
Information Mining a Component of EIP