Increasing Content Velocity

- Ramana Rao, CTO & Founder
- Inxight Software, Inc

www.ramanarao.com
www.inxight.com
Agenda

♦ The Problem of Content Underutilization
  • “Content Velocity”
♦ Traditional Library Science Ideas
♦ New Technologies for Organization & Access
  • Categorization
  • Information Extraction
  • Information Visualization
♦ Demonstration
♦ Examples
Content is Underutilized!

- Perhaps, most underutilized asset in large organizations
- Must enable people to find and understand available content in support of their work
- XML, portals, content management don’t address
- Neither search nor browse sufficient alone
- Content Velocity like Inventory Velocity

Enterprises appreciate that Search is Not Enough

According to an IDC survey...

- 62% consider document subject categorization to be important or very important
- 44% consider taxonomies to be important or very important
- 57% consider text mining (i.e. information extraction) to be important or very important
Traditional Library Science Ideas

♦ The Library Experience doesn’t just happen
  - Information Organization – Selection, Indexing & Cataloging
  - Information Access – Guidance, Education
♦ Multiple types of languages
  - Classification Schemes & Controlled Vocabs
♦ Multiple angles of access
  - Faceted Systems
Content is not “Unstructured”

**Summary**
- Wksjd skjsd skf
- Sdkskfjh. Fsds sjha hasdh.
- Dkw dklj cbjhwi.

**Embedded Concepts**
1. “...White House source...”
2. “...hot and cold running water...”
3. “...20 Gb hard drive...”
4. ...

**Embedded Entities**
1. Companies
   1. IBM
   2. Aventis
   3. Goldman Sachs
2. People
   1. Alan Greenspan
   2. ...

**Topical Categories**
1. Financial results
2. FDA Approvals
3. ...

**Concept Linker**
- “White House source” & “Environmental Policy”
- “20 Gb hard drive” & “Compaq Computer”
- ...

**Similar Docs**
- Document 1
- Document 176
- Document 3456
- ...

**Information Map**

Organizing & Accessing Content

Your Data

Extraction
Arrangement

Document Filtering
Entity Extraction
Summarization
Categorization
Concept Indexing
Similarity Indexing

Content-Derived
MetaData

Guided Search/Browse

Enterprise Applications

27 March 2003

Automatic Categorization

Classifies textual documents into categories usually based on what they are about

Enables
- document routing
- end user browsing of categories
- optimizing search & contextualizing results
Information Extraction

♦ Extracts nuggets from documents and collections
  • entities and concepts – people, places, things
  • Facts – relationships & events

Enables
  • tagging content w/ metadata
  • previewing documents
  • monitoring important events
  • finding specific answers
  • optimizing search & categorization

Information Visualization

- Interactive Visual representations that leverage human perceptual/spatial skills.

Enables
- getting an overview of collection
- navigating to specific documents
- interpreting in context
- narrowing search
- assimilating results more quickly
Extracting & Arranging

**Categorizer**
Sort documents by topic against a known taxonomy

**Summarizer**
Identify Key Sentences

**Search Indexing**
Build index to support querying

**Fact Extraction**
Extract Roles, Events, & Relationships

**Entity Extraction**
Extract Proper Nouns

**Phrase Extraction**
Identify Concepts

**Part of Speech**
Identify nouns, verbs, adverbs, etc.

**Tokenize**
Identify words

**Language ID**
Detect Language

**STEM**
Normalize words to root forms for more efficient indexing

**Language Matters**
March 2003

Linguistic Analysis: An Example

1. “Bank of New Zealand floods mailboxes with free checking account offers.”
2. “Banks of New Zealand river breached in flood.”

Others:

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>bank</td>
<td>2</td>
</tr>
<tr>
<td>new</td>
<td>2</td>
</tr>
<tr>
<td>zealand</td>
<td>2</td>
</tr>
<tr>
<td>flood</td>
<td>2</td>
</tr>
<tr>
<td>mailbox</td>
<td>1</td>
</tr>
<tr>
<td>free</td>
<td>1</td>
</tr>
<tr>
<td>check</td>
<td>1</td>
</tr>
<tr>
<td>account</td>
<td>1</td>
</tr>
<tr>
<td>breach</td>
<td>1</td>
</tr>
<tr>
<td>river</td>
<td>1</td>
</tr>
</tbody>
</table>

Inxight:

<table>
<thead>
<tr>
<th>Token</th>
<th>Part of Speech</th>
<th>Entity Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank of New Zealand</td>
<td>Proper Noun Group</td>
<td>Company</td>
<td>1</td>
</tr>
<tr>
<td>New Zealand</td>
<td>Proper Noun Group</td>
<td>Country</td>
<td>1</td>
</tr>
<tr>
<td>flood</td>
<td>verb-present tense</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>flood</td>
<td>noun-singular</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Mailbox</td>
<td>noun-plural</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>free</td>
<td>adjective</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>checking</td>
<td>noun phrase</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>account</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>bank</td>
<td>noun-plural</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Chinese (2), Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hungarian, Italian, Japanese, Korean, Norwegian (2), Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Turkish

27 March 2003

Access & Applications

**Mining**
Analyze and Explore Statistics of Collection

**Routing**
Generate New or Optimize Content Flow

**Guided Access**
Enhanced Search and Navigation using MetaData

**Search**
Present documents that match query

**Clustering**
Organize document or result sets

**Similarity Search**
Given a document, find similar docs

**Concept Linker**
Analyze results using links between embedded concepts

**Visualization**
Present Collections
DEMO
Applications

♦ R&D knowledge sharing
  • Accelerate product time-to-market by sharing internal research reports and data to avoid duplicate research, facilitate connections, accelerate project startup and execution, and buffer against personnel turnover.

♦ Customer and channel management
  • Improve customer and channel satisfaction by filtering inquiries based on content and context and accurately routing them for timely, relevant response

♦ Information aggregation and syndication
  • High precision, high volume document classification to facilitate accurate, fast routing and distributing of large volumes of information
  • Manage information retrieval and routing according in information-intensive environments such as market research, R&D, legal

♦ Government/Intelligence
  • Direct intelligence attention at most important issues analyzing large amounts of multilingual text for phrases, license plate numbers, suspect names, affiliations, connections, relationships, events, etc..

♦ Competitive Intelligence
  • Populate enterprise specific CI databases and collections
To be continued ...

- rao@inxight.com
  - Don’t hesitate to write ...
- www.ramanarao.com
  - Papers from talks
  - Information Flow newsletter
- www.inxight.com
  - White papers
  - Demos & free downloads