

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 Vocab
- ◆ Multiple angles of access
 - Faceted Systems

Content is not “Unstructured”

SIMILAR DOCS

- Document 1
- Document 176
- Document 3456
- ...

CONCEPT LINKER

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

INFORMATION MAP



W lkdfbw sd slkdfjd wkd we
wkwje cwef wkal ckj,wkejnw
eikn wek;jnwekjwnlkjwne
cwac w;eajnwg.

Today, bdwsbou wuwbou
weh bwjwd fooois if hows
when djbwe jsbdbqoo why
what xioajsxb ansiwu qbduw

Then, wdjhw ibwibi wehwb
hue eufu beueuu uwbbk.

SUMMARY

- Wksjd skjsd skf
- Sdkskfjh. Fsds sjha hasdh.
- Dkw dklj cbjhw.

EMBEDDED CONCEPTS

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

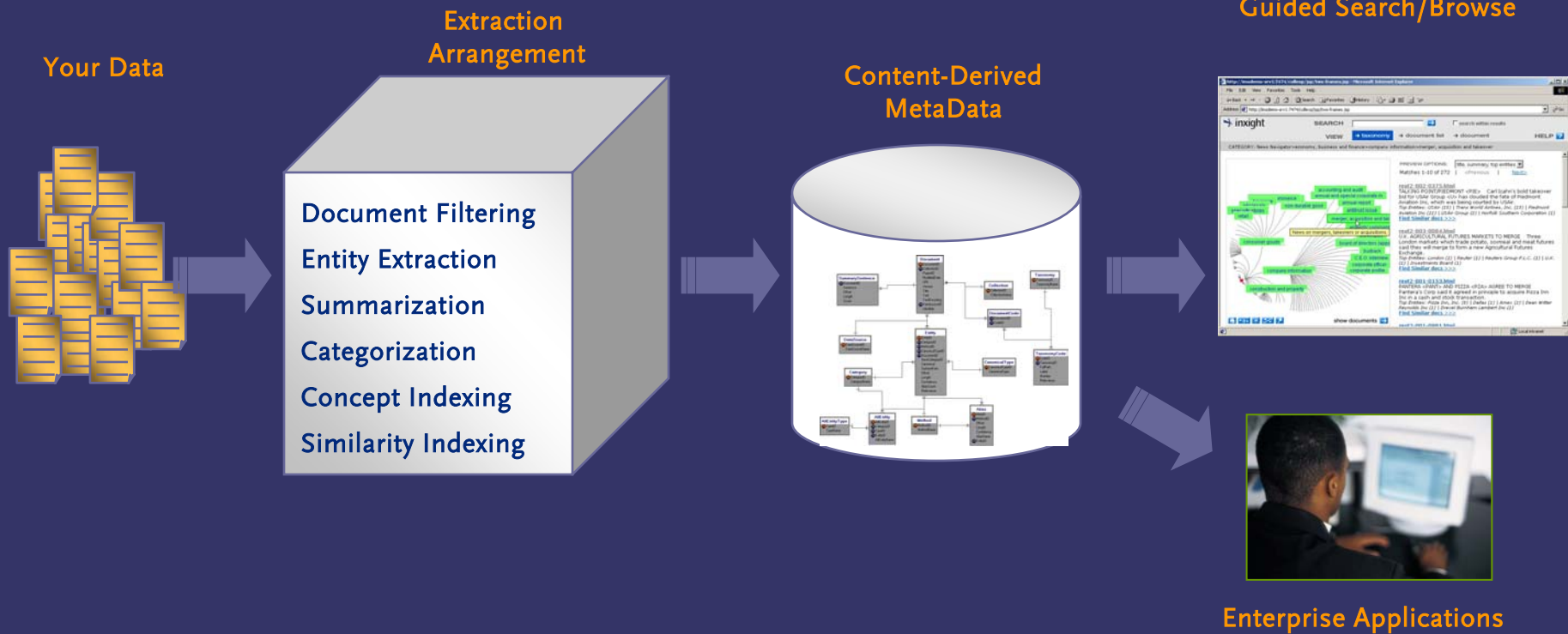
EMBEDDED ENTITIES

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

TOPICAL CATEGORIES

1. Financial results
2. FDA Approvals
3. ...

Organizing & Accessing Content



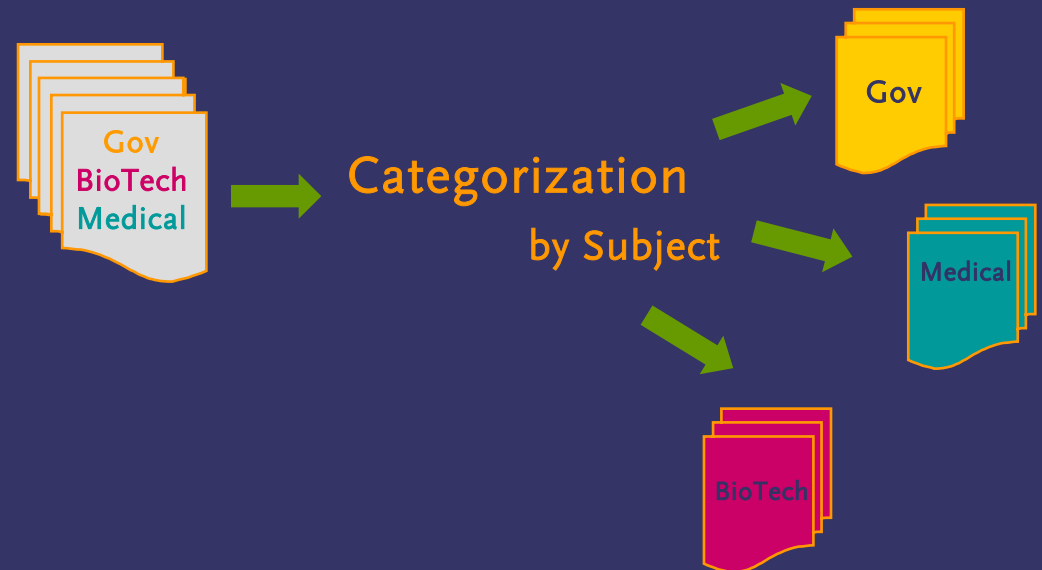
Automatic Categorization

- ◆ 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

The screenshot shows the Inxight search interface. At the top, there is a search bar with the text 'heart' and a 'Search' button. Below the search bar, there are navigation options: 'VIEW' with a dropdown menu set to 'taxonomy', and buttons for 'document list' and 'document'. A 'Search: heart' label is visible on the left, and a 'Display the List of Documents' button is on the right.

On the left side, there are two filter panels:

- concept filter:** A list of concepts with counts: Heart Attack (2), Ischemic Heart Disease (1), Symptom Of Advanced Heart... (1), Previous Heart Attack (1), Symptom Of Heart Failure (1), Congestive Heart Failure (1), Chamber Of The Heart (1), Beating Of The Heart's Ve... (1), and Heart Attack Prevention (1).
- entity filter:** Two sub-sections: 'place_region' with Africa (1), Europe (1), Middle East (1), South America (1), and SW Asia (1); and 'city' with Bombay (1), Campbell (1), Indianapolis, Ind. (2), and Toronto, Canada (1).

On the right side, there are 'PREVIEW OPTIONS' set to 'title, summary, top mentions' and 'Matches 1-6 of 6'. Three document previews are shown:

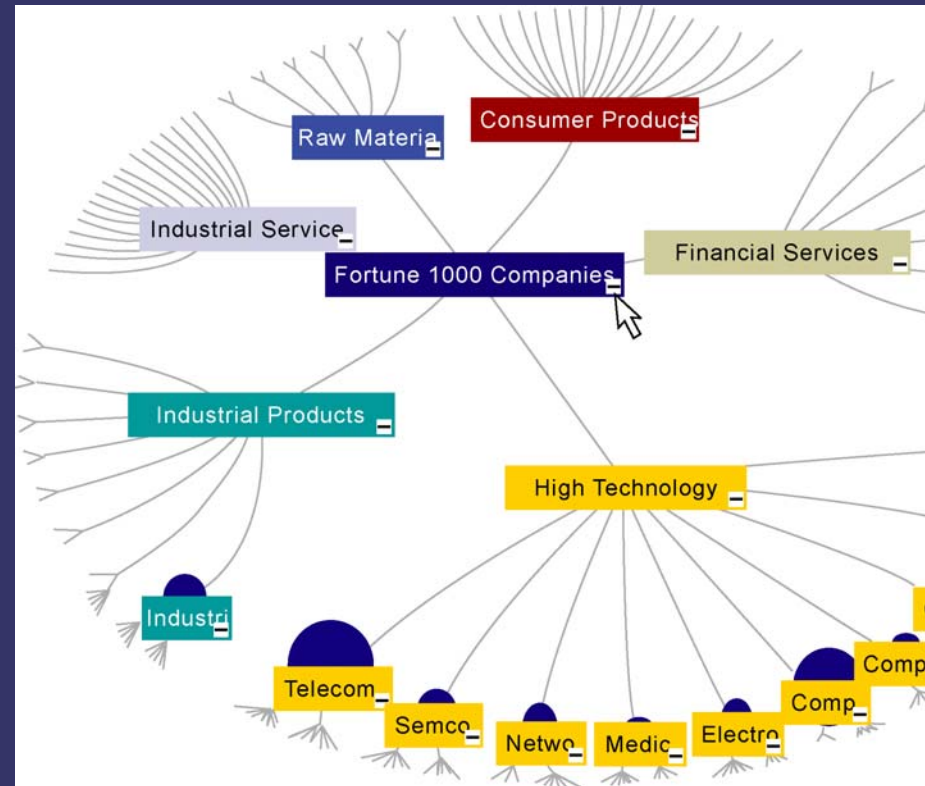
- FDA APPROVES NEW TYPE OF DEFIBRILLATOR.htm**: FDA Talk Papers are prepared by the Press Office to guide FDA personnel in... and accuracy to questions from the public on subjects of current interest. The... coordinates the beating of the left and right ventricles of the heart so that... effectively to pump blood throughout the body. Top 5 Mentions: FDA (3), Guidant Corporation (1), Press Office (1) Find...
- FDA APPROVES NEW INDICATION FOR IMPLANTABLE CARDIOVERTER DEFIB...**: FDA Talk Papers are prepared by the Press Office to guide FDA personnel in... and accuracy to questions from the public on subjects of current interest. FD... indication for Guidant Corporation's implantable cardioverter defibrillators (IC... population to include patients with a history of a heart attack and depressed... Top 5 Mentions: FDA (3), Guidant Corporation (1), Press Office (1) Find...
- FDA APPROVES NEW INDICATION AND LABEL CHANGES FOR THE ARTHRITI...**: FDA Talk Papers are prepared by the Press Office to guide FDA personnel in... and accuracy to questions from the public on subjects of current interest. An... study, however, was that there was a higher cumulative rate of serious card... adverse events (such as heart attacks, angina pectoris, and peripheral vaso... group (1.8%) compared to the naproxen group (0.6%). Top 5 Mentions: FDA (4), Press Office (1) Find Similar docs >>>
- FDA CLEARS NEW PALM TEST FOR SKIN CHOLESTEROL.htm**: FDA Talk Papers are prepared by the Press Office to guide FDA personnel in... and accuracy to questions from the public on subjects of current interest. FD... for marketing was based on a review of the firm's clinical studies that showe... disease or previous heart attack, it could provide 4% to 15% more informati... coronary artery disease beyond that already available with blood cholesterol... Top 5 Mentions: FDA (3), International Medical Innovations Inc. (1), Pres... docs >>>

Information Visualization

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

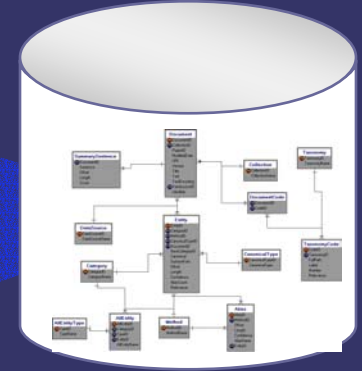
➔ Enables

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



Extracting & Arranging

Content-Based MetaData



SEARCH INDEXING
build index to support querying

CATEGORIZER
Sort documents
by topic against a
known taxonomy

SUMMARIZER
Identify Key
Sentences

FACT EXTRACTION
Extract Roles, Events,
& Relationships

ENTITY EXTRACTION
Extract Proper Nouns

**PHRASE
EXTRACTION**
Identify Concepts

PART OF SPEECH
Identify nouns,
verbs, adverbs,
etc.

STEM
Normalize words
to root forms for
more efficient
indexing

TOKENIZE
Identify
words

LANGUAGE ID
Detect Language

Wlkdfbw sd sklfdjd wdk we
kwkwe cwef wkal ckwj wkejnw
ejkn wek;jnwekjwnlkjwne
cwec w;eajnwg.

Today, bdwsbou wuwbou
weh bwjwd foook if hows
when djbwe jsbdbqoo why
what xioajsxb ansiwu qbduw

Then, wdjhw ibwibi weiwb
hue eufu beueuu uwbbk.

Language Matters

March 2003

<http://www.RamanaRao.com>

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:

Word	Count
bank	2
new	2
zealand	2
flood	2
mailbox	1
free	1
check	1
account	1
breach	1
river	1

Inxight:

Token	Part of Speech	Entity Type	Count
Bank of New Zealand	Proper Noun Group	Company	1
New Zealand	Proper Noun Group	Country	1
flood	verb-present tense		1
flood	noun-singular		1
Mailbox	noun-plural		1
free	adjective		1
checking account	noun phrase		1
bank	noun-plural		1

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

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

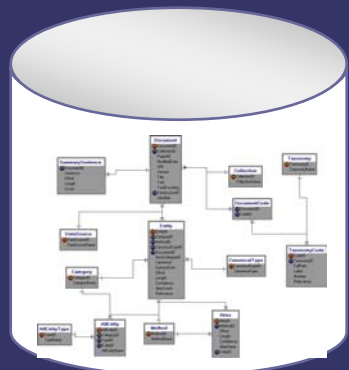
Given a document,
find similar docs

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