Emerging Technologies for Knowledge Management

- *Ramana Rao, CTO & SVP*
- *Inxight Software, Inc*

www.ramanarao.com
www.inxight.com

Extracting Value from Content

	Interact	Deliver	Organize	Analyze	Collect
1st Gen: Search/ Browsing	•Web Pages •Query/Results •Directories	•Web Servers •Query Engine	•Human Authors •Indexing Engines •Human Catalogers	•Stem & Phrases	•Crawlers
2nd Gen: Corporate Portals	•Portal UI	•Portal Server	•Metadata Repository		•Adaptors
3rd Gen: Interaction & Content Enhancement	•Visualization	•Personalization	•Categorization •Clustering	•Summarization •Entity/Concept Extraction •Link Analysis	

Beyond Search & Browse

- Search
  - Precise, but brittle … leaves users searching, not finding …
- Browse
  - Robust, but vague … leaves users wandering & lost, not found …
- Opportunity is to blend Browsing & Search
  - Categorization
  - Information Extraction
  - Information Visualization

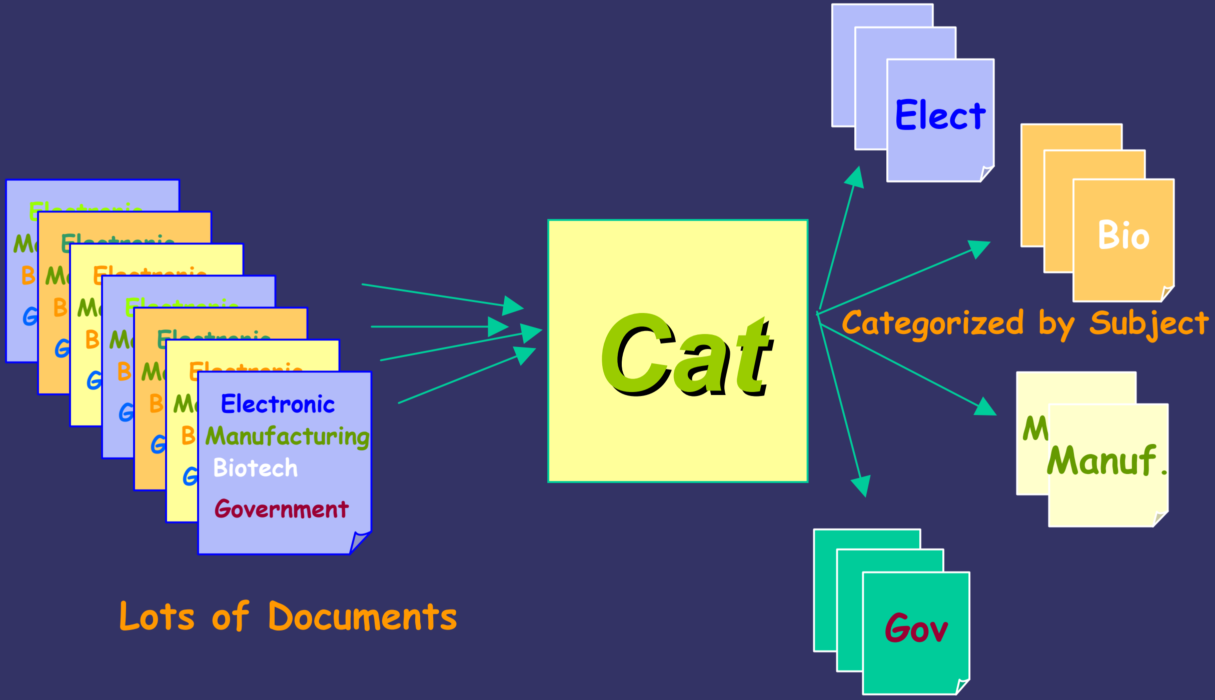
Automatic Categorization

What:

Subject Categorization classifies textual documents into categories based on what they are about.

Why:

Increases Efficiency & Effectiveness of Systems and People that utilize the content



Categorization: Publishers vs. G2000

Electronic Publishers & Aggregators

- Inherent to Product
- Taxonomy-Savvy
- Content Tagging

- **Corpus:** large and dynamic (>1M docs, >5K new docs per day)
- **Accuracy:** mission critical
- **Taxonomy:** have one, and it is likely complex; have pre-existing workflow and understand process of management
- **Training set:** have training data of appropriate quantity and quality

Global 2000 Enterprises

- Reuse of Knowledge Assets
- Taxonomy-Challenged
- Document Access & Routing

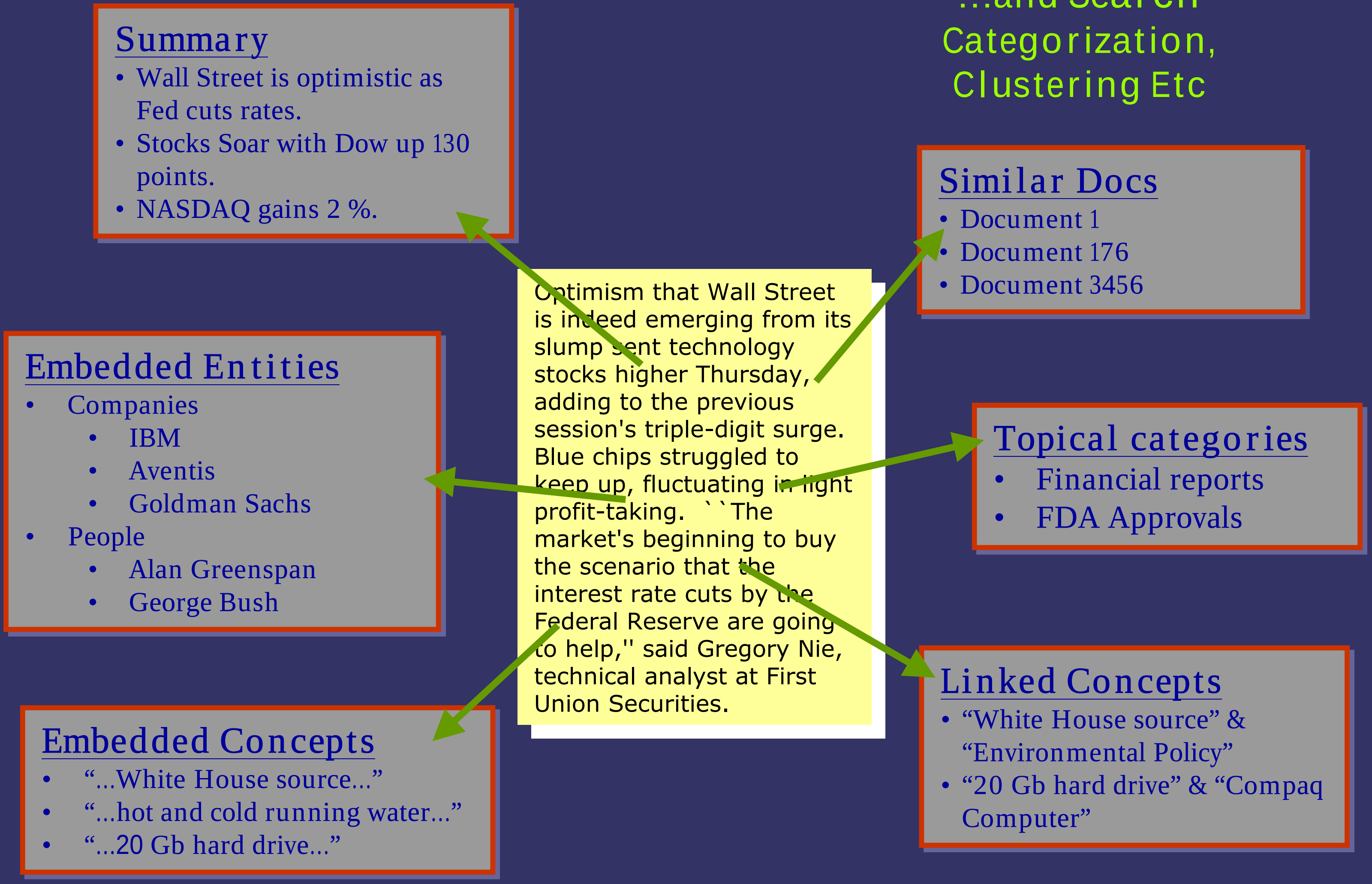
- **Corpus:** moderate to large (>100K docs, possibly >1M docs)
- **Accuracy:** not mission critical
- **Taxonomy:** no or limited pre-existing taxonomy; require extensive taxonomy workflow support
- **Training set:** typically no pre-existing training set

Information Extraction

- Information extraction is about pulling elements out of documents and collections that guides the more intelligent use of content
- Often characterized as metadata that provides context
- Types of metadata include:
  - Noun phrases
  - Named entities (e.g., people, companies, places, products)
  - Key sentences
  - Concepts and topic relationships
  - Similarity between documents, paragraphs and phrases

MetaData from Information Extraction …

…and Search Categorization, Clustering Etc



Optimism that Wall Street is indeed emerging from its slump sent technology stocks higher Thursday, adding to the previous session's triple-digit surge. Blue chips struggled to keep up, fluctuating in light profit-taking. ``The market's beginning to buy the scenario that the interest rate cuts by the Federal Reserve are going to help,'' said Gregory Nie, technical analyst at First Union Securities.

Summary

- Wall Street is optimistic as Fed cuts rates.
- Stocks Soar with Dow up 130 points.
- NASDAQ gains 2 %.

Embedded Entities

- Companies
  - IBM
  - Aventis
  - Goldman Sachs
- People
  - Alan Greenspan
  - George Bush

Embedded Concepts

- "…White House source…"
- "…hot and cold running water…"
- "…20 Gb hard drive…"

Similar Docs

- Document 1
- Document 176
- Document 3456

Topical categories

- Financial reports
- FDA Approvals

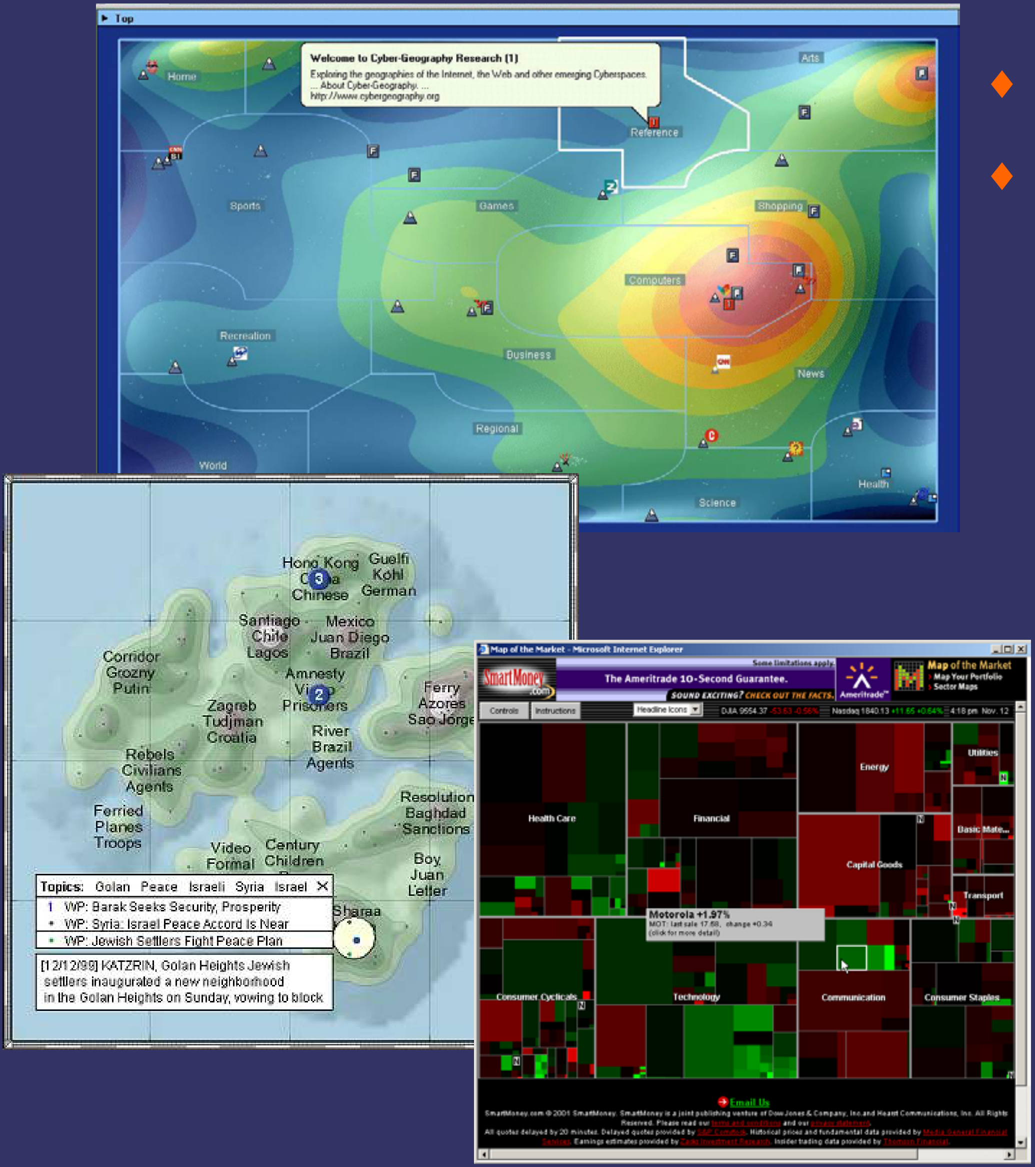
Linked Concepts

- "White House source" & "Environmental Policy"
- "20 Gb hard drive" & "Compaq Computer"

Information Visualization

- The Role of Content Visualization
  - Provides maps of large content spaces … and also the means for getting to specific documents or items
  - Thus support early or organizing processes like orientation, assessing, survey, etc. … as well as tune very focused processes like direct walk navigation
- Nature of the Solution
  - Leverage our visual/spatial skills
  - Like browsing, but shows much more, maps not just pages
  - Can eleminate mechanical overheads of browsing
  - Can integrate with searching more tightly
- Two key types of Content Visualizations
  - Content Terrain Maps
  - Wide Widgets

Content Terrain Maps



Welcome to Cyber-Geography Research [1]

Exploring the geographies of the Internet, the Web and other emerging Cyberspaces.
... About Cyber-Geography ...
http://www.cybergeography.org

Topics: Golan Peace Israeli Syria Israel

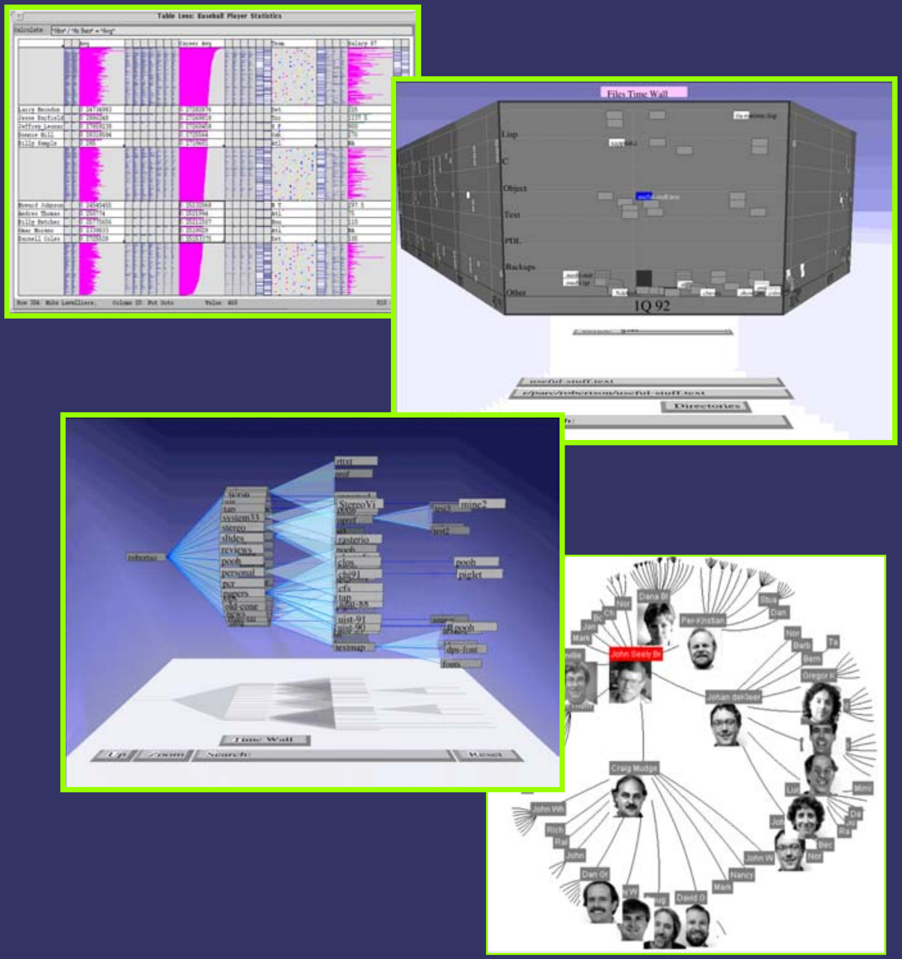
1 WP: Barak Seeks Security, Prosperity
• WP: Syria: Israel Peace Accord Is Near
• WP: Jewish Settlers Fight Peace Plan

[12/12/99] KATZRIN, Golan Heights Jewish settlers inaugurated a new neighborhood in the Golan Heights on Sunday, vowing to block

Motorola +1.97%
MOT: last sale 17.68, change +0.34
(click for more detail)

- Analogy to geographic maps
- Organize on 2-d surface with Terrain created by contours, colors, regions, based on:
  - *Human Design*
  - Categorization
  - Automatic Clustering based on Document Similarity
  - Metadata

Wide Widgets



- high bandwidth widgets for interacting w/ large collections
- arranged on a spine
  - *Hierarchical*
    Cone Tree, Spiral Calendar, Hyperbolic Tree Browser
  - *Temporal*
    Perspective Wall, Time Lens
  - Pages -
    Document Lens, Web Books
  - *Calendars* -
    Spiral Calendar
  - *Tabular* -
    Table Lens, Time Lattice

DEMO

To be continued …

- rao@inxight.com
  - Don’t hesitate to write …
- www.ramanarao.com
  - Papers from talks
  - Newsletter to start in May focused on Intelligent Information Access
- www.inxight.com
  - White papers
  - Visualization demos & free downloads