Categorizing the Web
Bootstrapping Personalized Content Management

Daniel P. Lulich, CTO
April 2001
Content Management Architecture

Unstructured

- WEB PAGES
- NEWS FEEDS
- E-MAIL
- DOCUMENTS
- MESSAGES
- ARCHIVES
- TRANSCRIPTS
- CHAT

Structured

- CRM
- CUSTOMER SERVICE
- E-Publishing
- DATA MINING
- CONTENT MANAGEMENT
- CORPORATE PORTALS
- DOC MANAGEMENT
- BUSINESS INTELLIGENCE

Categorization Toolkit

Ontology Definition
Data Collection
Training Testing
What are the Pitfalls?

• The Category Definition Problem
• The Deployment Problem
• Meeting User Expectations
• The Bootstrapping Problem

Some Possible Solutions:

• Adding Personalization to the Architecture
• Bootstrapping with an Archetypal Ontology
• Building this Ontology from the Web
The Category Definition Problem

Archetypal Ontology

- Who is the user?
- What does the user do?
- What does the user know?
- What does the user need to know?

Some classes of users may be able to share an ontology, but in general the ontology depends on the user.
The Deployment Problem

- Categorization toolkits are offered as a solution to the content management problem.

- Toolkits are only a partial solution because:
  - They assume the user is a categorization expert.
    - Knows what to build
    - Has the time to build it
    - Knows when it’s good enough
  - They do not focus on the user’s business problem.
  - Precision & Recall are not a proxy for the user’s cost function.

- The user typically fails to achieve desired results with the toolkit. Then a system integrator is hired as the categorization expert to rescue the project.

- The user’s experience is late delivery and an enormous cost overrun.
Meeting User Expectations

• Users understand their business problems and can articulate what they expect from the categorization solution.

• The real difficulty is meeting expectations of all users while dynamically satisfying the needs of individuals.
  - This appears to require us to read each user’s mind.

• Solution: “Closing the Loop”
  - Model the user’s expectation over time by building personalization into categorization.

• Dynamic Sources for Personalization
  - Collaborative model: Ask the user
  - Learned model: Observe user’s behavior
  - Prototypical model: Assume the user fits a known model
  - Functional model: Understand the user’s role
  - Historical model: Profile of user’s behavior over larger time intervals
  - Derived model: Ask an oracle
The Bootstrapping Problem

• Users don’t know what categories they need. They don’t know where to start.
• Solution: Bootstrap category selection with a large inventory of prefabricated classifiers.
• The Web appears to be a good source for these categories. Why?
  - The web is a large living library.
  - Tens of Thousands of labeled pages in categories.
  - Users are familiar with a variety of “ontologies” such as Yahoo, ODP, Business.com…
  - Others have tried with some success:
    - Mladenic D., Grobelnick M. 1998
    - Dumais S., Chen H. 2000
Content Categorization Architecture

- WEB PAGES
- NEWS FEEDS
- E-MAIL
- DOCUMENTS
- MESSAGES
- ARCHIVES
- TRANSCRIPTS
- CHAT

Categorization Engine

- Ontology Definition
- Data Collection
- Training
- Testing

Ontology Generator

"create the new user model"

Archetypal Ontology

Personalization Engine

- Archetypal Users
- User Feedback
- User Behavior

WWW

CRM
- CUSTOMER SERVICE
- E-Publishing
- DATA MINING
- CONTENT MANAGEMENT
- CORPORATE PORTALS
- DOC MANAGEMENT
- BUSINESS INTELLIGENCE
Very Large Scale Categorization
Built from the Web

Category Space Robot - the OntoBot

- On-the-fly categorization into 2000 categories.
- Hierarchical classifier.
- Trained with 3 million web pages.
- Ontology derived from the ODP.
- Limping versions with 40K categories.
- Astounding emergent behavior.
What is Your Precision and Recall?

- I honestly can’t report formal accuracy. Why?
  - We were unable to label all the testing and training data.
    - Grading 3M web pages into 2000 categories is too expensive.
  - The ODP directory is not a real hierarchy.
  - Many ODP categories are not topical.
  - There is an incredible amount of noise in the data
    - Web pages
    - Hierarchy
    - Categories

- Anecdotally we have found:
  - 85% of the time the answers are right.
  - 10% of the time they are explainable.
  - 5% of the time they make no sense.

- We believe it is not about precision and recall. It is about meeting user expectations. Each user has a different cost function. We have users whose cost functions meet our anecdotal performance.

- Closing the loop allows us to adapt via the user’s cost function.
What We Have Learned

• Existing text categorization theory does not serve us any longer. We are pushing outside the envelope. We are working in noisy spaces very much like nature.

• When you have 1000’s of classifiers and large ontologies things get interesting.
  - Emergent behavior
  - More is better – Critical mass.

• Accuracy numbers distract us from the real problem: solving the user’s business problem.

• The hierarchy is important and can be leveraged for computational performance and accuracy
  - D’Alessio et. al. 2000

• We are becoming expert at working with Web data
  • Graphics, Image only, Flash
  • Frame Pages, Redirects, Links
  • Mixed content pages (like portals)
  • Script-heavy pages
  • Meta Tags