inxight

An Evaluation of Modern Categorization Systems

Ian Hersey
VP, Linguistic Products
www.inxight.com

Why Categorize?

- **Let users find information more easily**
  - Categories provide a “map” into a collection
  - Can be used as a search/delivery filter
  - Can help provide a single interface for disparate content sources
  - Can combine with other metatags for powerful content exploration experience:

    “I’m looking for a Gartner <doc_source=“Gartner”> market research report <doc_genre=“market research”> on knowledge management <doc_topic=“KM”>.”

inxight

Why is Categorization Difficult?

- **Humans are expensive, inconsistent and slow**
- **Machines are cheap, consistent and fast, but dumb**

…but that's not all…

- **The problem is *language***
  - – Ambiguity
  - – Complexity

inxight

Approaches to Categorization

- **Three basic approaches: manual, unsupervised and supervised**
  - – Manual techniques use human-built rules and/or keywords
  - – Unsupervised techniques uses statistical processing to separate documents into clusters based on common terms
  - – Supervised techniques use a training set + statistical processing

inxight

Approaches to Categorization

- **Manual techniques**
  - – Lets users exert a high degree of control, but…
  - – Difficult to scale
  - – Non-adaptive
  - – Labor intensive
- **Unsupervised techniques**
  - – Can be a good starting point, but…
  - – Difficult to scale
  - – Non-adaptive

Approaches to Categorization

- **Supervised techniques**
  - Require a training set
  - Training set should be well-coded
  - Requires minimum number of examples per code
- **Desirable features**
  - User feedback can be incorporated into the system; system "learns" as new documents are categorized without retraining the entire system
  - System scales to large taxonomies, can apply multiple categories
  - System provides taxonomy creation and management tools

inxight