Text Analytics and Auto-Categorization in Semantic Web Applications

Jim Wessely
Advanced Document Sciences
Who Am I?

Jim Wessely

• President, **Advanced Document Sciences** (founded 1995)
• Content management application design since 1990
• Text mining, analysis, and auto-categorization since 1993
• Taxonomy design with auto-categorization since 1998
• Ex-principal consultant for search, taxonomy, and portal architectures, IBM Global Services
• Clients include ConocoPhillips, Turner Broadcasting, National Cancer Institute, Bloomberg Research, US Army, US Marine Corp, IBM Investor Relations, IBM Almaden Research, MetLife, Anadarko Petroleum, State Farm Insurance, DuPont, and others
Agenda

• Why Has Unstructured Data Been Such A Problem?
• Textual Information Access (what are we actually *trying to do*?)
• What Auto-categorization Is and Is Not
• Text Analytics Technologies And Techniques
• Auto-Categorization Technologies
• Taxonomy and Ontology for Auto-categorization
• Challenges of Text with Semantic Web Applications
• Thoughts on Best Practices
Why Has Unstructured Data Been A Problem For Such A Long Time?
Findability: Information Is All Over The Place

THE CHALLENGE

• It’s difficult to enter information into systems
• Then it’s difficult to find the information
• Keyword search engines are imprecise
• Lots and lots of data, poor accessibility
• Too much irrelevant clutter and noise
Aboutness: What Are Any Of These About?
Why Is This Difficult?

Human Factors

- Physical or Cognitive Properties of an Individual or Human Social Behavior which Influence Functioning of Technological Systems
- Limiting Factor = Human Behavior
A Few Definitions

- **Categorization**: Organization of objects into one or more categories

- **Category**: a general concept that marks divisions or coordinations in a conceptual scheme

- **Classification**: Organization of objects into a single, specific class

- **Corpus**: a body, or collection, of document objects

- **Document**: An individual *object* which contains informational content. May contain text, audio, imagery, graphics, video, etc., or any combination thereof

- **Entity**: a term or phrase contained in textual content
A Few More Definitions

• **Entity Extraction**: identification and extraction or tagging of specific entities in documents or a corpus

• **Lexicon**: The vocabulary of a subject or text corpus

• **Relevance**: A *subjective* value indicating the significance of information to an individual user

• **Taxonomy** (*totally unofficial and very loose definition)*: A structure defining relationships between concepts or objects for categorization or classification (commonly hierarchical)

• **Thesaurus**: A list of subject headings or descriptors usually with a cross-reference system for use in the organization of a collection of documents for reference and retrieval
Textual Information Access

OK… What is it that we are actually trying to do?
Text Retrieval (a.k.a. Enterprise Search)

Common enterprise text retrieval functionality:

- Text Indexes
- Document filters
- Metadata filters
- Multiple search techniques
  - Key terms
  - Wild cards
  - Boolean operators (AND, OR, NOT, Less Than, etc.)
  - Proximity searching
  - Thesaurus or lexicon integration (synonyms related terms)
  - Should access multiple repositories and applications
- Federated Search & Metasearch
- Administrator configurable
- Social tagging (this is relatively new)
- Semantic search enhancement
  - Is this really an improvement?
Federated Search
Browsing Textual Content

• Category browsing

Requirements:
  - Identify the categories
  - How does the content get categorized?
    - Auto-categorization (best if properly done)
      - Consistent, scalable, fast, difficult, no perfect technologies or methods
Auto-Categorization *And* Text Retrieval
Autocat, Search, & Metadata = Findability

Advanced Filter Options
- Hide and Cancel Search Filter
- Hide and Cancel Content Type Filters
- Show Associated Collections

Search Filter Query:
- corrosion OR erosion

Content Type:
- Best Practice

Content SubType:
- Any

Topic Browse
- Arctic Technology
  - Construction and Materials (6)
  - Offshore Infrastructure (2)
  - On-Shore Infrastructure (7)

Documents
- Arctic Technology (9 documents)
  - Corrosion Under Insulation (CUI)
    - Corrosion Under Insulation (CUI)...Even for the TSA coated surface corrosion would develop in the areas with damaged coating after some months at 120 C. (The attached report was prepared in 1995 and at that time the photos were not digital, so the posted electronic version of the report has no pictures.)...Laboratory test result indicated that a TSA coating using wire drawn from aluminum alloys designed for galvanic anodes will provide long term cathodic protection.
What Auto-categorization Is And What It Is Not

Why use it and why not to
When to use it and when not to
First - What Is Auto-categorization?

Software that associates content objects with categories in an organizational scheme

• Automatically associates document objects with one or more categories

• Usually using proprietary (and differing) methods

• Requires human intervention to create “rules”

• Most commonly used for textual content
Not To Be Confused With…

- **Entity extraction tools**
  - Software that can identify and extract specific terms and phrases from textual document content
    - From a list you provide
    - Or based on known patterns in text
  - Auto-categorization vendors sometimes include these in their products, but they are supporting players, although sometimes very useful

- **Taxonomy management software**
  - Maintenance of complex term and relationship data structures
  - Autocategorization software usually only supports hierarchical structures
Why Use Auto-categorization?

• Scalability
  – Concisely retrieve information from multiple, large and diverse content repositories

• Findablility and Aboutness
  – Organize information according to what it is about to increase “relevance”

• Simplify and speed up access to knowledge content

• Simplify and speed up entry of knowledge content
Why Use Auto-categorization?

- Improve information access and findability
- Improve search interfaces and results
  - Categorize search results
  - Search within a category and subcategories
- “Push” similar documents out to your users
- Uncover “themes” in your enterprise content
- Consistent categorization
- Find content in multiple topic areas
- Leverage and reuse knowledge across stovepipes
Why *Not* to Use Auto-Categorization

- To totally replace your staff of human indexers (Computers can’t *really* read or understand context)
- You already have rich metadata associated with your content
- You need close to 100% accuracy
- You’re not prepared to devote some resource to building and managing rules
- You need a categorization structure that won’t support autocategorization
- You have a small amount of content
- You have content types that can’t be autocategorized
When Does It Work Best?

- When you can get by with “simple,” easily maintained rules
- When your content is technical, or contains lots of specific terminology
- When your project can support dedicated resources on an ongoing basis to create and maintain complex rules
- With knowledge content that is very descriptive
## Categorizers For Beginners

<table>
<thead>
<tr>
<th>Method</th>
<th>Required Setup &amp; Maintenance</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical / Clustering</td>
<td>Little: But human input (taxonomies, key words, etc.) can improve results</td>
<td>Often disappointing if large scale, but can work well for limited topic range</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Some: Gather representative documents</td>
<td>Often disappointing, but algorithms are always improving</td>
</tr>
<tr>
<td>Linguistic/Lexical</td>
<td>More: Gather and rank full corpus or representative words and concepts</td>
<td>Better, but it’s important to understand how the product uses the terms you choose</td>
</tr>
<tr>
<td>Semantic Rules</td>
<td>Most: Gather words and concepts and understand rule syntax</td>
<td>Best, but rule syntax and operators are key and they vary by product</td>
</tr>
</tbody>
</table>

All of These Are **Content-Based Categorization Methods**
What Is The Typical Process?

• Gather content from your domain, and examine it with the autocategorization tool(s)
  – Gather as much content as you can; this is important!
  – Analyze your content, using both human and machine methods
  – Gather the terms that will be used as categories (usually your taxonomy)
  – Gather term variants
    ▪ Synonyms, Acronyms, Narrower terms, Related terms
      – These terms typically drive the categorization process

• Understand your software’s rule syntax
  – Minimum occurrence, parts of speech, proximity, exclusion, etc.
  – If clustering, understand how you can “influence” the process and results

• Test against large, representative samples of content or the complete content set if available
Case Study - Army Knowledge Online

Before Autocategorization:

A query was made specific by using a phrase and the Boolean “and” expression.

Full text search of AKO content found 2,283 documents, but no indication of how to find the Field Artillery repository area on the site.

In a separate results area, 227 more documents were identified, but still no indication of how to find the Field Artillery document area.
After: Army Field Artillery School

Yahoo!-like categories provide easy access to content in context to the immediate needs of users.

Hierarchical sub-categories provide the granularity that gets users to what they need.

Direct access to frequently referenced material.

Concept lists help users to see why things are in the categories, and help narrow the document list.
Integrated Autocat And Search

Search results show different paths in the hierarchy for context, phrases containing the search term(s), and number of documents in result set.

Related terms display allows user to finely tune searches using terms they hadn’t even thought of!

Context display displays automatically generated summary so that the whole document doesn’t have to be retrieved just to see if it is truly relevant.
Why Are Autocat Rules Needed?

- Rules declare how content is associated with categories in the taxonomic structure
- Rules can be machine generated, human generated, or a combination of the two
- Two primary types of autocategorization engines
  - Mathematical representation of text
  - Semantic and lexical representation
- Humans *must* be in the loop
Autocat Rule Creation

• Statistical Clustering
  – Useful for narrow subject range
  – Useful for visualization

• Machine learning
  – Positive and negative training examples

• Semantic
  – Build a lexicon
  – Purely semantic/lexical
    • associate lexical terms with nodes of the taxonomy
  – Boolean and complex rule building
    • Use Boolean and other operators to make conditional rules
How The Rules Work

• Mathematical representation
  – Training sets / machine learning
    ▪ Bayesian
    ▪ Support Vector Machine
    ▪ Neural Networks
    ▪ Statistical Clustering

• Semantic representation
  – Parts of speech disambiguation
  – Identification of sentence structure
  – Entity identification
Text Analysis
Technologies and Techniques

Various ways to work with all of this pesky unstructured stuff
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.
# Entity Extraction (EE)

Can be the easiest, cheapest and quickest way extract “metadata”

The proposed merger between Mega, Inc. and CNA Systems, Incorporated, has been postponed, Mega CEO Joe Smith said in an analyst call. “CNA’s 1st quarter revenue dropped by 32%, and they lost 23 million dollars,” Smith explained. CNA Systems sources blame weak sales in China. CNA shares (CNAI) fell 47 percent to $9.84 on May 12, the first trading day after the announcement.

<table>
<thead>
<tr>
<th>Company</th>
<th>Mega, Inc., CNA Systems, Incorporated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>May 12</td>
</tr>
<tr>
<td>Person</td>
<td>Joe Smith</td>
</tr>
<tr>
<td>Person Position</td>
<td>Mega CEO</td>
</tr>
<tr>
<td>Currency</td>
<td>23 million dollars, $9.84</td>
</tr>
<tr>
<td>Measurement</td>
<td>32%, 47 percent</td>
</tr>
<tr>
<td>Country</td>
<td>China</td>
</tr>
<tr>
<td>Concept</td>
<td>proposed merger, analyst call, 1st quarter revenue weak sales, first trading day</td>
</tr>
<tr>
<td>Event: M&amp;A</td>
<td>The proposed merger between Mega, Inc. and CNA Systems, Inc. has been postponed</td>
</tr>
</tbody>
</table>
Some Things EE Can Extract

- Address + sub parts
  - City
  - Company
  - Country
  - District
  - Currency
  - Date
  - Facility
  - GeoCoord / MGRS
  - Holiday
  - IIR Format Entities
  - Internet Address
  - Language
  - Month
  - Organization + sub types
  - Percent
  - Person + sub parts
  - Peoples

- Place_Political
  - Place_Region
  - Place_other
  - Phone
  - Position
  - Product
  - Publication
  - SSN
  - State
  - Vehicles
  - Weapons
  - YearGeneral
  - Member of
  - Organization + sub types
  - Company information
  - Management change
  - Merger & Acquisition
  - Product release

- Automatically identify the people, places, companies, etc., from within unstructured data
- Does not carry the development cost typically associated with taxonomies or ontologies
- Promotes information discovery by recognizing people, places, companies, and things based on understanding of language instead of out-of-date lists and training sets
In the Beginning…
Statistical Analysis and Clustering

• Statistical procedures, often using single word entities (not phrases), to identify similarities or commonalities between the terms in a corpus
• Common to use “clustering” algorithms to group statistically similar content together
• Sometimes the algorithms perform clustering inside the clusters to create preliminary hierarchies
• Statistical algorithms: Look at word frequency, co-occurrences of words, word placement and grouping, etc. within documents as ways to identify relationships
• Often assume that position of terms to one another (proximity) implies relational significance
• Another common assumption is that term closest to the cluster centroid describes the category
• Works best over narrow topic ranges
• No way for the computer to identify context in which the terms are used
• Language independent and scales well
• Subject to statistical error
2D Visualization Of Clustered Text
3D Visualization Of Clustered Text

Spatial Paradigm for Information Retrieval and Exploration (SPIRE) – Pacific Northwest Labs
Clustering In Commercial Software

“Clustering dynamically categorizes content into hierarchical themes based on the content itself at query time” - Vivisimo

Clustering shows related themes for users to quickly drill down into or discover new themes.

Structured navigation allows users to drill into metadata such as file type, authors, departments, etc.
Adaptive/Machine Learning

Support Vector Machine
Bayesian Analysis
Neural Networks

• Use mathematical representations of textual content
• Often assumes that the positions of terms (proximity) and co-occurrence to one another implies relational significance
• Can create a mathematical “signature” for user interest profiles
• Easier to create visualizations since the data is already numerical
• Typically language independent
• Often doesn’t scale easily
• Can be very compute intensive
A Bayesian Training Toolset

Image courtesy of Quiver, Inc.
Learning By Example (LBE)
Semantic/Linguistic Analysis

- Parts of speech analysis
- Identifies semantically significant terms
- Single term and phrase extraction
- Creates/uses lexicon that is representative of the corpus
- More difficult to use visualization tools
- Can create a personal lexicon
- Interest matching algorithms can range from simple (term entities) to very sophisticated (semantic profiles)
- Language dependent
Semantic Analysis: Fact Extraction

Category Tree display of Category dimensions including Entities, Fact Categories, and Term Categories. Events can also be included as part of the Category tree dimension (but are not in this example).

Constrained on the “Complaints” Branch of the category tree

Displaying the corresponding Complaint values also as categories. The user can choose to display either the root values or the different levels of the category tree – the same as other dimensions that contain categories.
Categorization of Extracted Facts
Fact Extraction – Graphical Representation
Auto-Categorization Technologies

Making Unstructured Content Go Where You Want It To Go
Early Lexical Rules

Environment
Climate
Arid/Desert
+desert
+dune
+dunes
+effects of mirages
+sand
+sandy
-desert shield
-desert shield/storm
-desert storm
-opfor desert lore
-sand bag
-sand hook
-sand table
Weighted Lexical Rules

Image courtesy of Semio, Inc.
Weighted Linguistic Rules

Image Courtesy of Teragram/SAS® Text Analytics
Boolean Semantic Rules

Image Courtesy of SAP/BusinessObjects (was Inxight)

SemTech 2011 – Text Analytics and Auto-categorization in Semantic Web Applications

© Copyright 2011 Advanced Document Sciences, All Rights Reserved
Testing Your Rules

Image Courtesy of SAP/BusinessObjects (was Inxight)
More Rule Testing

Image Courtesy of SAP/BusinessObjects (was Inxight)
Expanding Rules With Variant Terms

Image Courtesy of SAP/BusinessObjects (was Inxight)
Term Enhancement via Text Semantics

© 2010 Concept Searching – Used with Permission
Concept Term Suggestion

© 2010 Concept Searching – Used with Permission

SemTech 2011 – Text Analytics and Auto-categorization in Semantic Web Applications

© Copyright 2011 Advanced Document Sciences, All Rights Reserved
Concept Term Selection

© 2010 Concept Searching – Used with Permission
Concept Term Feedback

© 2010 Concept Searching – Used with Permission
Steps For Creating Your Rule Set

1. Test
2. Refine & enhance
3. Test
4. Refine & enhance
5. Validate and obtain feedback

• Repeat steps 1 through 5 as necessary
Who is Best to Perform this Work?

Knowledge Engineers and Super Searchers

• Employ deep content analysis

• Ability to communicate with subject matter experts (SMEs)

• Attention to detail despite sometimes mundane work

• Patience for sometimes arcane syntax

• “Word people”
Preparing For Autocategorization

• Select your autocategorization technology very carefully

• Learn your autocategorization software well

• Explore your autocategorization technology and test your skill

• Don’t expect 100% precision and recall

• Be patient. You will need to refine and refine again
Taxonomies and Ontologies

For Auto-categorization and Semantic Web-based Applications
Auto-Categorize to a Defined Structure

- Auto-categorization organizes content to a pre-defined target structure – a taxonomy or metadata schema.
- You will need to create a taxonomy or metadata schema.
- Good luck finding two people that agree on a definition of a “taxonomy”.
- There is more than one way to create a taxonomy.
- There is more than one type of structure for a “taxonomy”.
- As we have seen, there are multiple ways to categorize textual content to a taxonomy.
What kind of Taxonomy - Four Types

- Flat

  - Laurie
  - Mike
  - John
  - Barb
  - Jim
  - Elena
  - Randy

- Hierarchical

  - Transportation
    - Automobiles
      - Truck
      - Sedan/Coupe
      - Sports Car
      - SUV
    - Aircraft
    - Boat
    - Rail
      - Jet
      - Propeller
        - Passenger
        - Military
Four Taxonomy Types (continued)

• Faceted

**Business Function**
- Accounting
- Finance
- Human Resources
- Legal
- IT
- Marketing
- Production
- Sales
- ...

**Location**
- Brazil
- Canada
- Columbia
- Mexico
- Nicaragua
- Panama
- United States
- Venezuela
- ...

**Content Type**
- Abstract
- Compliance
- Contract
- Financial Report
- Incident Report
- Invoice
- Market Plan
- Map
- ...

• Poly-Hierarchical
Which Type Should I Use?

- **Flat**
  - Simple classification
  - Often used for records or document management and file systems
  - Try not to use more than 16 nodes, 20 at most

- **Hierarchical**
  - Enterprise content management
  - Knowledge sharing & management
  - When content is hierarchical in nature
  - When users access information by drilling down

- **Faceted**
  - Records/document/content management
  - When using a database to manage a metadata schema

- **Networked**
  - When content has complex relationships that must be retained
Taxonomies for Your Application

Considerations:

• What type of taxonomy do you need to use?
• How many taxonomies do you need?
• What kind of talent are you willing to pay for either internally or externally?
• What type of taxonomy does your application need?
• How quickly do you need to be up and running?
• Can you start simply and build in the future?
• Are there regulations that dictate what your application (therefore your taxonomy) needs to do?
• What will your total cost of ownership be?
Using Pre-Built Taxonomies

• No pre-built taxonomy will provide anything close to a 100% fit. You **will** need to modify it, or settle for something that isn’t usually a very good fit for most applications.
  - Examine the taxonomy to see how well it suits your application
  - Determine if you can use it without very much modification
  - Identify how difficult it will be to customize the taxonomy
  - Identify what needs to be done to adapt the taxonomy to your software or application
  - Identify the tools required to efficiently modify the taxonomy
  - Estimate the cost to modify it and the level of effort required

• Governance and change control policies and practices are still required

• Taxonomies from a few of the software vendors really only support for that vendor’s software (vendor lock-in?)

• Most pre-built taxonomy vendors only sell hierarchical taxonomies
  - That may be fine if that’s what your application needs to use
Analytics To Aid Term Identification

Structured navigation allows users to drill into metadata such as file type, authors, departments, etc.
Term Expansion & Suggestions
Challenges of Text in Semantic Web Applications
W3C and Text

W3C Recommendations regarding handling of textual content:

W3C Recommendations regarding auto-categorization:
SKOS

“vocabularies define the concepts and relationships (also referred to as “terms”) used to describe and represent an area of concern”

Tom Gruber: http://www.w3.org/standards/semanticweb/ontology

• This is the closest match for any W3C recommendations that have to do with textual information and applications

• Defines a “vocabulary”

• It is essentially just a thesaurus
  – Essentially a listing of terms, their synonyms, and [almost arbitrarily] “related” terms

• NOT a rich semantic representation of anything

• Not particularly useful for rich, full auto-categorization or taxonomy/ontology design
Auto-Categorization and Semantic Web Applications

A Few Solution Architecture Approaches
One Text Analytics Flow for SemWeb

UIMA (IBM) based

Copyright © Mondeca 2011
Mondeca’s Semantic Web Tools

ITM
Ontology Management: Schemas, Thesauri, Terminologies, Knowledge Management

CA-Manager
Content management
Annotate content

Taxonomy Classifier
Publication management
Classify content

Semantic Portal
Content delivery

Terminologies, ontology models
Terminologies, classification taxonomies
Terminologies
SPARQL Query Construction
Mondeca’s Content Classifier Overview

Taxonomy classification engine

- Based on RDF triplestore
- Loads terminology and metadata
- Infer on terminology
  - OWL & SKOS inference
  - Custom rules
- Apply SPARQL classification rules
- Optionnally, simplifies RDF structure

Content classified with additional
\textit{dcterms:subject} and \textit{dc:subject} properties
BBC’s 2010 World Cup Architecture

- Underlying framework “publishes” metadata (in RDF), not content
- RDF is contained in a triplestore, which allows “agile modeling”
- An ontology defines assets: stories, blogs, profiles, images, video and statistics
- Assets are “tagged” with metadata
- Stats are stored separately in XML and transformed into RDF
- Metadata queries (SPARQL) enable dynamic page aggregation

A Few Thoughts on Best Practices

How to Actually Make It Work
Best Practices

• Gain an understanding of your content (if possible)
  – Collect a full corpus of content or representative groups of “good” documents if the full corpus is too large
  – You’ll always be surprised at what you find and learn, especially with large collections

• Take advantage of your subject matter experts

• Take advantage of vendor training when available

• Avoid short cuts when writing rules
  – Beware of using words with different meanings
  – Beware of acronyms
  – Yes, it’s very tedious, but it absolutely requires your attention

• Plan for lots of rule testing, or better yet, use a vendor product that shows auto-categorization results as you create the rules
A Couple Implementation Suggestions

• Plan for technical integration
  – Develop a technical architecture that accounts for each of the “moving parts” of your solution
    ▪ Taxonomy
    ▪ Content analysis tools
    ▪ Categorizer
    ▪ Integration with other applications
      – Search engines/Search UIs
      – Content management systems
      – Content distribution systems

• Always do a PoC or pilot. Always.
Questions?

Jim Wessely
President
Advanced Document Sciences
Steamboat Springs, Colorado

jwessely@adocs.com
970-875-0556

Providing Innovative, Advanced Information Access