Auto-Categorization for Records and Information Management

The Good, the Bad, and the Ugly

Ben Licciardi
Taxonomy and Information Management Analyst
Agenda

• Setting the Scene – Classification, Taxonomy, and Human Indexing
• The Dilemma – So Much Content, So Little Time...
• To the Rescue – Auto-Categorization Rides into Town
• Character Study – What Makes Auto-Categorization Tick?
• Riding into the Sunset – Is Auto-Categorization the Hero for You?
• Meet the Cast – A Few Auto-Categorization Tools
• Around the Campfire – Auto-Categorization and RM Roundtable
What is classification?

Classification is the process of organizing and describing objects by placing them into categories.
What is taxonomy?

Often, the categories we use to classify objects are pulled from a controlled vocabulary or taxonomy.

**Taxonomy**

Fruit
- Apples
- Bananas
- Blueberries
- etc...

Tableware
- Bowls
- Cups
- Dinnerware Sets
- Flatware
What is taxonomy?

• Broadly speaking, a taxonomy is any system of organization that divides and orders concepts into groups or categories.

• In the corporate setting, a taxonomy usually refers to a controlled vocabulary designed to help a business or organization classify and organize information assets in order to solve a problem or accomplish a goal.
A taxonomy in action...
Setting the Scene – Classification, Taxonomy, and Human Indexing
How do objects get categorized?

One option is *human* indexing, where a person...

- Examines and interprets an object,
- Finds appropriate terms in the taxonomy to describe the object
- Applies those terms to the object in the form of metadata
Recipe Example

Chicken Cacciatore (Hunter's Wife's Chicken)

(Spollo alla Cacciatore)

SERVES 4

America knows this Italian favorite as chicken cacciatore (hunter's-style), but it's really alla cacciatore, named in honor of the hunter's wife—who, all over northern Italy, might traditionally cook the dish on the eve of the hunt as fuel for the chase.

4 tbsp. extra-virgin olive oil
2 medium onions, peeled and chopped
2 cloves garlic, peeled and minced
1 3-lb. chicken, cut into 8 pieces
1 cup dry white wine
1 28-oz. can peeled whole san marzano tomatoes, chopped, juice reserved
1 bay leaf
1 tsp. minced fresh rosemary leaves
1/4 cup minced flat-leaf parsley
Salt and freshly ground black pepper
1 cup strong chicken stock

1. Heat oil in a large pan over medium-high heat. Add onions and cook, stirring occasionally, until soft, about 10 minutes. Add garlic and cook, continuing to stir, for about 2 minutes more. Push onions to

Metadata

Course: Main Course
Cuisine: Italian
Difficulty: Easy
Main Ingredient: Chicken, Tomato
Technique: Sauté
Source: Saveur
Benefits of human indexing

• We’re really good at understanding and interpreting objects
• We’re great at understanding context and how that influences classification
• We’re accurate (when well-trained)

Challenges of human indexing

• We’re slow
• We cost money
• We get bored and lazy
• We’re not always consistent in the way we describe objects (we’re all different)
Human indexing is great, but in some settings it can be problematic...

**TOO. MUCH. CONTENT.**
Tagging takes time and money. Sometimes it’s simply not feasible to manually index a large collection.

**Low Levels of Interest**
The people responsible for tagging may not take the job all that seriously.

**Consistency**
Different people tag in different ways. Even the same person might tag in different ways over time.
Auto-Categorization is designed to overcome some of these challenges.

*Takes the job of indexing away from humans and hands it over to computer software*

- Computers can easily move through large amounts of data
- Computers never get bored
- Computers are consistent – they only do what you tell them to do
How does auto-categorization work?

As with human indexing, the first thing auto-categorization tools attempt to do is understand the object in question.
Auto-categorization tools determine the “aboutness” of an object based on:

- Where the object lives in the directory structure
  - Who created the object
- The text contained within the object (text analytics)
- Any other metadata already associated with the object
Based on this interpretive analysis, auto-categorization tools then apply taxonomy terms to objects automatically.

They do this in a couple of ways:

- Rules Logic
- Machine-Learning
Rules Logic

One of the most basic techniques used in auto-categorization is if/then rule logic.

If a recipe exists in the folder “Chicken Recipes”...

Then apply the tag “Chicken” to the Main Ingredient metadata field.
Rules Logic

These rules can be very simple (as in the last example) or a sophisticated combination of weighted variables.

Contains the string “General Motors” in the Title field. **Add 40 points.**

Contains the string “General Motors” in the body text more than 3 times. **Add 20 points.**

Contains the the string “GM” in the body text more than three times. **Add 10 points.**

Contains the string “genetically modified” in the Title field. **Subtract 40 points.**

Etc...

**If** score is greater than or equal to 50...

**Then** apply the term “General Motors” to the Subject field.
Rules Logic

In a way, this is reminiscent of database searching, where the system creates a sophisticated search string and applies tags based on the results of the inquiry.

### Logical Operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR</td>
<td>Retrieves all records that contain at least one of the search terms.</td>
<td>A or B</td>
</tr>
<tr>
<td>AND</td>
<td>Retrieves all records that contain all of the search terms.</td>
<td>A and B</td>
</tr>
<tr>
<td>NOT</td>
<td>Eliminates records that include a search term or group of search terms.</td>
<td></td>
</tr>
</tbody>
</table>

Caution: Use the NOT operator carefully. You may untag records accidentally.

### Proximity Operators

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Example</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>(W)</td>
<td>S SOLAR(W) ENERGY &lt;br&gt; S PAGO(W) PAGO &lt;br&gt; S RETINAL SCAN</td>
<td>Requests that terms be adjacent to each other and in the order specified. Can also be used to retrieve identical terms. Empty paired parentheses are interpreted to be the (W) operator.</td>
</tr>
<tr>
<td>(nW)</td>
<td>S SOLAR(nW) ENERGY</td>
<td>Requests that terms be within n words of each other and in the order specified.</td>
</tr>
<tr>
<td>(N)</td>
<td>S AIR(N) POLLUT? &lt;br&gt; S JOHNSON(N) JOHNSON</td>
<td>Requests that terms be adjacent but in any order. Can also be used to retrieve identical terms.</td>
</tr>
</tbody>
</table>
Machine Learning

Another technique auto-categorization tools use is machine learning, where the system uses sophisticated algorithms to tag objects based on examples.

Character Study – What Makes Auto-Categorization Tick?

User provides the tool with multiple examples of documents that should be tagged with “Chicken.”

The tool analyses these documents and makes notes of their characteristics.

When new documents that resemble the examples enter the library, the tool applies the term “Chicken.”
Is auto-categorization the answer to my information management problems?

Maybe. But keep in mind, there’s no such thing as a silver bullet in the world of categorization...
All forms of categorization – manual and automated – require time and work, and no method is accurate all the time.

• Auto-categorization tools require human intervention to set-up, fine-tune, and maintain.
• If you don’t invest time and energy into the software, metadata quality suffers.
• In fact, some claim auto-categorization is less accurate compared to human indexing.
Auto-categorization could be a great option for you if...

- You have more documents than you can manually index
  - Speed of indexing is a priority
- You have the time, knowledge, and money to invest in auto-categorization software
- You have a strong familiarity with the content in your collection
  - Your collection is made up of text documents
Human indexing could be a great option for you if...

- You have enough people and enthusiasm to manually tag your collection
  - You need a high level of metadata accuracy
  - You have images and videos in your collection
There are a number of auto-categorization tools out there. Here are a few:

- Autonomy

- Data Harmony MAIstro

- Concept Searching

- OpenText

- SmartLogic
Here’s a quick look at one of those tools, Concept Searching:

Document Type Taxonomy

Weighted “Clues”
Round-Table Discussion

*Implications for Records Management*

- Metadata accuracy and compliance
  - Print and digital files
  - Scanned files and OCR
  - Archiving and disposal
  - Other issues?
Ben Licciardi

ben.licciardi@ppc.com

(703) 748-7263