AUTO-CLASSIFICATION & TAXONOMY: NATURAL ALLIES

DEC 4, 2012
AIIM - ECM BREAKFAST & SEMINAR

Stephanie Lemieux
A LITTLE ABOUT ME…

Founder & principal at Dovecot Studio

• Taxonomy, information architecture, search, content management

Current & past work includes:

MLIS from McGill University
TODAY’S TOPICS

Auto-classification 101
Integrated taxonomy & auto-indexing tools
SharePoint
Selection advice
TAGGING IS INCONVENIENT

It’s time-consuming
It’s subjective
There’s often too much content to tackle

“You get to ask people for 3 fields. Make ‘em count.”

-- Instruction during a recent digital asset management project
SO WHY DON'T WE GET MACHINES TO HELP?
AUTO-CLASSIFICATION
(Over-simplified)

Document preparation
- Split into language blocks (paragraphs, headings, etc.), formatting, layout, etc.

Parsing
- Entity extraction
- NLP: parts of speech, phrases
- terms, variants, etc.

Weighting
- Frequency
- Location in text, phrase
- Proximity
- Combination
- Format of text

Classification
- If threshold reached
- Can influence search results

This is where rules vs. stats comes into play...
STATISTICAL

Often use Bayes theorem: measures “degrees of belief”
(or in this case, “degrees of aboutness”)
  • use word frequency & location to identify salient concepts

Document-training sets
  • match word/concept patterns to categories
  • often need sets of 50+ documents
  • poor document choice can cause pollution/noise
EXAMPLE CASE: ASTRAL
Android, le système d'exploitation pour mobiles du géant de l'internet Google, prend à Apple une part de plus en plus grande du marché des téléphones et des tablettes informatiques, grâce à une innovation rapide et un large éventail d'engins moins coûteux.

"La variété est une des forces" d'Android, estime Stephen Baker, un analyste du cabinet de recherche NPD Group. "Avoir beaucoup de gens fabriquant beaucoup de choses, couvrant toute une gamme de prix, avec de multiples marques (...) fait une grande différence", souligne-t-il.
APPROACH

Auto-classifier & entity extractor

“Skill cartridges” -- training sets

• Some are off the shelf (people names, locations, companies, etc.)
• Some are custom – built against subject taxonomy

Provide 30 “best of show” articles for each taxonomy term

Test for false positives & missed documents

Machine-aided indexing – content creator triggers process if desired
Maison > Décoration > Peinture et faux finis

Faux fini : un mur « hallucinant »

Travaux de peinture : les bons outils

Peindre du papier peint?

Grossesse, peinture et teinture à cheveux

Bienvenue à vous toutes, futures mamans, qui avez pris la peine de venir me poser vos questions. Je suis extrêmement touchée de votre confiance.

Vicky veut savoir s'il y a des contre-indications à se teindre les cheveux pendant la grossesse et quels sont les dangers potentiels liés à la peinture. Voici sa question :

« Bonjour, j'aimerais savoir s'il y a un risque associé à la coloration (mèches, teinture) des cheveux lors de la grossesse? Mon médecin me dit qu'il ne devrait pas y avoir de problèmes, mais j'ai lu des opinions contraires et je ne sais plus à qui me fier. J'en suis à ma 16e semaine et j'ai pour l'instant décidé de ne pas en faire, mais je souhaiterais connaître votre opinion sur le sujet. Aussi, pourriez-vous m'informer quant aux dangers potentiels liés à la peinture? Merci beaucoup! »
DIFFICULTIES

• Because tagging drives content publication, some staff do not trust the auto-classifier, prefer to manually tag

• Temis does not take into account synonyms from the taxonomy – not a true taxonomy tool

• Changing the taxonomy means re-training the tool – big investment

• Rules of taxonomy vs. auto-indexing
  • 5+ documents = warrant for a taxonomy term
  • Need 30 to train the categorizer…
RULES-BASED

Boolean (and, or, not) rules to find positive or negative evidence of a match

More control over behaviour

\[\text{but with power comes great responsibility, and work!}\]

Success depends on quality of rules

- E.g. Match category “Apple” if phrase found close to “iPhone” or “Steve Jobs”, but not “fruit” or “farm”

Weighting (how discriminatory a term is for a category) + frequency used to adjust scoring
EXAMPLE CASE: FOUNDATION CENTER

Organization that tracks philanthropic giving

• Database of grantmakers (+100,000) and grants given in the US

• Classify grants from data in 990 forms (tax statements)

• Challenge:
  • Grant descriptions are very short – often a few words only
### 990 Finder

**Notice:** The IRS has announced processing delays of the 990 for filing years 2007-2009. Learn more...

Your query: *(Organization Name: Ford, State: TN)*

1987 documents matched. 100 documents displayed.

### Search Again

<table>
<thead>
<tr>
<th>ORGANIZATION NAME</th>
<th>STATE</th>
<th>YEAR</th>
<th>TOTAL AMOUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A J Ford Ministry Inc.</td>
<td>TX</td>
<td>2006</td>
<td>5000</td>
</tr>
<tr>
<td>Aj Ford Ministries Inc.</td>
<td>TX</td>
<td>2003</td>
<td>9700</td>
</tr>
<tr>
<td>Aj Ford Ministry Inc.</td>
<td>TX</td>
<td>2004</td>
<td>9700</td>
</tr>
<tr>
<td>Aj Ford Ministry Inc.</td>
<td>TX</td>
<td>2005</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust</td>
<td>NH</td>
<td>2003</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust U/W</td>
<td>NH</td>
<td>2006</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust U/W</td>
<td>NH</td>
<td>2004</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust U/W</td>
<td>NH</td>
<td>2005</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust Uw</td>
<td>NH</td>
<td>2007</td>
<td>9700</td>
</tr>
<tr>
<td>Alice P Ford Trust</td>
<td>NH</td>
<td>2010</td>
<td>9700</td>
</tr>
<tr>
<td>Alice P Ford Trust Uw</td>
<td>NH</td>
<td>2011</td>
<td>9700</td>
</tr>
<tr>
<td>Alice P Ford Trust Uw</td>
<td>NH</td>
<td>2009</td>
<td>9700</td>
</tr>
</tbody>
</table>

### Grants and Contributions Paid During the Year

<table>
<thead>
<tr>
<th>RECIPIENT NAME AND ADDRESS</th>
<th>RECIPIENT RELATIONSHIP AND PURPOSE OF GRANT</th>
<th>RECIPIENT STATUS</th>
<th>AMOUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>APRICAN WILDLIFE FOUNDATION</td>
<td>NONE GENERAL OPERATIONS</td>
<td>PUBLIC CHARITY</td>
<td>50</td>
</tr>
<tr>
<td>ALCIOHPS OF THE TREASURE COAST, INC.</td>
<td>NONE GENERAL OPERATIONS</td>
<td>PUBLIC CHARITY</td>
<td>10,000</td>
</tr>
<tr>
<td>AMERICAN ASSOCIATION OF MUSEUMS</td>
<td>NONE GENERAL OPERATIONS &amp; PROGRAM</td>
<td>PUBLIC CHARITY</td>
<td>48,000</td>
</tr>
<tr>
<td>ANIMAL HUMANE SOCIETY</td>
<td>NONE GENERAL OPERATIONS</td>
<td>PUBLIC CHARITY</td>
<td>1,000</td>
</tr>
<tr>
<td>ASSOCIATES OF THE JFB LIBRARY</td>
<td>NONE PROGRAM</td>
<td>PUBLIC CHARITY</td>
<td>1,342</td>
</tr>
</tbody>
</table>

### Example Case: Foundation Center

**JAMES FORD BELL FOUNDATION**

**FORM 990-PF**

**Statement**

Your query: *(Organization Name: Ford, State: TN)*

1987 documents matched. 100 documents displayed.

### Search Again

<table>
<thead>
<tr>
<th>ORGANIZATION NAME</th>
<th>STATE</th>
<th>YEAR</th>
<th>TOTAL AMOUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A J Ford Ministry Inc.</td>
<td>TX</td>
<td>2006</td>
<td>5000</td>
</tr>
<tr>
<td>Aj Ford Ministries Inc.</td>
<td>TX</td>
<td>2003</td>
<td>9700</td>
</tr>
<tr>
<td>Aj Ford Ministry Inc.</td>
<td>TX</td>
<td>2004</td>
<td>9700</td>
</tr>
<tr>
<td>Aj Ford Ministry Inc.</td>
<td>TX</td>
<td>2005</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust</td>
<td>NH</td>
<td>2003</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust U/W</td>
<td>NH</td>
<td>2006</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust U/W</td>
<td>NH</td>
<td>2004</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust U/W</td>
<td>NH</td>
<td>2005</td>
<td>9700</td>
</tr>
<tr>
<td>Alice Ford Trust Uw</td>
<td>NH</td>
<td>2007</td>
<td>9700</td>
</tr>
<tr>
<td>Alice P Ford Trust</td>
<td>NH</td>
<td>2010</td>
<td>9700</td>
</tr>
<tr>
<td>Alice P Ford Trust Uw</td>
<td>NH</td>
<td>2011</td>
<td>9700</td>
</tr>
<tr>
<td>Alice P Ford Trust Uw</td>
<td>NH</td>
<td>2009</td>
<td>9700</td>
</tr>
</tbody>
</table>

### Grants and Contributions Paid During the Year

<table>
<thead>
<tr>
<th>RECIPIENT NAME AND ADDRESS</th>
<th>RECIPIENT RELATIONSHIP AND PURPOSE OF GRANT</th>
<th>RECIPIENT STATUS</th>
<th>AMOUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>APRICAN WILDLIFE FOUNDATION</td>
<td>NONE GENERAL OPERATIONS</td>
<td>PUBLIC CHARITY</td>
<td>50</td>
</tr>
<tr>
<td>ALCIOHPS OF THE TREASURE COAST, INC.</td>
<td>NONE GENERAL OPERATIONS</td>
<td>PUBLIC CHARITY</td>
<td>10,000</td>
</tr>
<tr>
<td>AMERICAN ASSOCIATION OF MUSEUMS</td>
<td>NONE GENERAL OPERATIONS &amp; PROGRAM</td>
<td>PUBLIC CHARITY</td>
<td>48,000</td>
</tr>
<tr>
<td>ANIMAL HUMANE SOCIETY</td>
<td>NONE GENERAL OPERATIONS</td>
<td>PUBLIC CHARITY</td>
<td>1,000</td>
</tr>
<tr>
<td>ASSOCIATES OF THE JFB LIBRARY</td>
<td>NONE PROGRAM</td>
<td>PUBLIC CHARITY</td>
<td>1,342</td>
</tr>
</tbody>
</table>
APPROACH

Home-grown system

- Rules based
- Crawl grant text for only certain grantmakers who had 990s in a specific format
- Takes many rules for just one category:
  - E.g. Day care
    - Not near dog, doggie, pet, cat, etc.
    - Not near adult
  - E.g. Nursery (synonym)
    - Not near plant, tree, etc.
- Rules-based gives them the control they need to deal with the lack of text
- Moving to new system (taxonomy tool) in 2013
SOME KEY FEATURES

Stemming, canonical form
Co-referencing or proximity (preceding & forward)
Zone-based relevancy (title/body, intro, metadata, etc.)
Entity extraction (including events, dates, people, etc.)
Weighting
Automatic vs. machine-aided options
## PROS & CONS

<table>
<thead>
<tr>
<th>Statistical</th>
<th>Rules-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Work involved in building good training sets</td>
<td>• Work involved in building exhaustive rules (mitigated by taxonomy tools)</td>
</tr>
<tr>
<td>• If there’s a problem, can be difficult to diagnose &amp; rectify/retrain</td>
<td>• If there’s a problem, go back to the rule set &amp; tweak</td>
</tr>
<tr>
<td>• Machine-learning can augment accuracy or lead to pollution (accuracy can wax &amp; wane)</td>
<td>• System doesn’t evolve without new rules, but high degree of control (accuracy mostly increases)</td>
</tr>
</tbody>
</table>

Interesting read: Automatic Indexing ROI - A Case Study Comparison of Rule Based and Statistical Approaches
Auto-classification starts with a rich taxonomy, thesaurus or ontology

- Provides values and their “evidence” - associated terms & phrases (NTs, RTs)
- Rules or patterns created against the evidence

Combo taxonomy management & auto-classification in one tool

Some auto-classifiers come with pre-built taxonomies & rules
POPULAR COMBO TOOLS

Semaphore by Smartlogic
  • www.smartlogic.com

M.A.Istro by Data Harmony
  • www.dataharmony.com

SAS Enterprise Content Categorization (formerly Teragram)
  • http://www.sas.com/text-analytics/index.html

Concept Searching
  • http://www.conceptsearching.com/

Most taxonomy/indexing combos are rule-based
HOW THEY WORK

Components:

- Taxonomy/ontology/thesaurus manager
- Rule base (usually auto-created from taxonomy)
  - Category-based
  - Entity or fact-based
  - Candidate (e.g. “Report written on”)
- Rule builder/editor
- Text analysis/NLP engine
- Concept extractor
- Reporting/statistics
- Language packs
- Editorial review workbench (for machine-aided vs. automatic classification)
M.A.I.STRO: DATA HARMONY

Thesaurus manager (no custom relationship types)

Rules-based (auto-creation from taxonomy)

Developed for both automatic & assisted indexing

- Includes statistical comparisons between human & machine (identify noise & misses)

Includes concept extractor

SharePoint connector
RULE BUILDER VIEW: M.A.ISTRO
Methane Lakes Raise Hopes of Life on Titan

In the dark, chilly tropics of Saturn's moon Titan, the Cassini probe has spotted what appears to be liquid methane lakes. Titan has a rocky, icy surface coated by a thick atmosphere of nitrogen and methane. It's a place of mountains and several possible cryovolcanoes, which are thought to erupt methane rather than water. It's the only extraterrestrial object to have ever shown clear evidence of stable bodies of surface liquid water, which is only been spotted in the planet's polar regions. In 2005, exo-meteorologists saw evidence of lakes and seas on Titan, rather like the water cycle on Earth — where methane evaporates and then rains down as a liquid. In a recent paper, scientists from the University of Arizona in Tucson, told Nature.

However, the discovery of the liquid methane lake in the desert-like tropical regions of the planet remains to be confirmed and be transported to the cooler poles. "Lakes at the poles are easy to explain, but lakes in the tropics are the most intriguing," said the lead author.

The lake that Grift and his colleagues found is 37 miles long, 25 miles wide and at least appears to be a black mark on near-infrared wavelengths (which can penetrate Titan's thick atmosphere) with a diameter of 12 miles. It may be "shallow enough to see through from Earth, with knee-to-ankle-level depths".

So where is the methane coming from? One potential source could be the methane that is being released from the moon's equatorial regions as they evaporate from above. That's the idea, it adds to the number of places on Titan where there might be liquid water. Of course it's entirely plausible that it might not be a lake at all. A possible organic compound on Titan could be widespread, but if that's the case, it adds to the number of places on Titan where there might be methane down there. If so, there could be methane life there.

A NASA mission to Titan to look at its complex chemistry in more detail, named the Titan Mole, is due to land in 2030 and will be able to detect methane below the surface. It's still in development and will be a great opportunity to find out if there's life on Titan. So far, only 17 percent of Titan's surface has been covered in maps, which they believe could be important in detecting liquid methane below the surface. In recent months in a sea of methane on Titan that would be the perfect place to look for liquid methane. That's the idea, it adds to the number of places on Titan where there might be methane down there.
SEMAPHORE : SMARTLOGIC

Full enterprise-class ontology manager
Rules-based (auto-create from taxonomy)
Lots of features, including text mining/entity extraction
Starting on big data fact extraction
20 types of business rules
SharePoint connector
Search enhancement module (faceted search, etc.)
## Rule Base Editor Example: SMARTLOGIC

### Rule Editor

<table>
<thead>
<tr>
<th>Rule Type</th>
<th>Name</th>
<th>Weight</th>
<th>Data</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>meta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>content</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>category</td>
<td>Apple</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>link</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>combine</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>text</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>max</td>
<td></td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>phrase</td>
<td></td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>any</td>
<td></td>
<td>60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Text Rule**

The basic rule is a **text** rule, used to capture a single or multiple words. It can be used as a standalone rule or as the building block for more complex rules, for example, **phrase** and **sentence** rules.

**Weight**

The weight of the rule determines how much it will influence the classification.
BARCLAYS CAPITAL does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report.

Investors should consider this report as only a single factor in making their investment decision. Please use ANALYST (R) CERTIFICATION (R) AND IMPORTANT DISCLOSURES BEGINNING ON PAGE 12.

1. Apple Posts Record Quarterly Results Turning Focus from Steve Jobs Leave of Absence.

On 1/16, Apple reported fiscal 2Q11 EPS of $0.64, significantly beating our recently raised estimate of $0.53 and the recently raised consensus of $0.59. Earnings upside was primarily driven by much better than expected revenue and margins, while a lower than expected tax rate and higher other income played a minor role. Revenue of $19.74 billion (+$7.1B vs. our recently raised estimate by about $2.5 billion and consensus by $2.3 billion due to better than expected sales of iPods, iPhones, Macs and iPads.) Gross margin in the quarter of 38.5% was well above our estimate of 37% and 250bps above Apple’s original guidance due to a favorable component cost environment as well as leverage on higher revenue and lower other product costs (w/o freight, warranty, support). Operating margin in the quarter was 29.3%—ahead of our 27.3%.

Cash from operations in the quarter was $9.77 billion with strong free cash flow of $8.55 billion or 57.1% of GAAP.

Apple posted a very strong CAG.

2. In our opinion.

With regard to products, Apple beat our unit estimates across the board posting 5.10 million iPhones (up 86% yoy and +15% qoq) vs. our 5.3 million estimate, iPad sales of over 4.1 million vs. our 4.4 million estimate, iMac sales of 7.3 million vs. our 6.5 million estimate, and Mac units of 10.5 million (+17% qoq), or our 17.55 million estimate. This very strong quarter report left no holes to punch in the fundamental story. We believe the above consensus revenue and EPS guidance and new products to come bring potential for further upside. We continue to believe the company is in very capable hands with COO Tim Cook and the rest of the team. Our price target for Apple is $450 based on $1 FY12 EPS estimate of $27.00 excluding interest income ($0.31) but adding back forward 12-month cash ($85). Please see our more detailed earnings note “Quarters Ahead: Even By Apple’s Standards” from 1/16/11 for more information.


In conclusion, we view today’s news as quite positive as it suggests that Jobs has likely completed his medical leave of absence and is eventually going back to work. Jobs’ continued presence at Apple will be key to the company’s long-term success.

Windows 10 • 5:55 AM • 4/13/2016
## Classification Review Example: SMARTLOGIC

**Classification Review Tool**

### Browse Analysis: POST Notes by:
- Documents
- Terms

### Terms for Document

<table>
<thead>
<tr>
<th>Rulebase Class</th>
<th>Term Name</th>
<th>IPSV v5</th>
<th>IPSV v4</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPSV</td>
<td>Abuse</td>
<td>0.51</td>
<td>0.0</td>
</tr>
<tr>
<td>IPSV</td>
<td>Professional people</td>
<td>0.48</td>
<td>0.0</td>
</tr>
<tr>
<td>IPSV</td>
<td>Parents</td>
<td>0.76</td>
<td>0.0</td>
</tr>
<tr>
<td>IPSV</td>
<td>Families</td>
<td>0.79</td>
<td>0.0</td>
</tr>
<tr>
<td>IPSV</td>
<td>Education and skills</td>
<td>0.53</td>
<td>0.0</td>
</tr>
<tr>
<td>IPSV</td>
<td>Mental health</td>
<td>0.66</td>
<td>0.9</td>
</tr>
<tr>
<td>IPSV</td>
<td>Primary schools</td>
<td>0.0</td>
<td>0.09</td>
</tr>
<tr>
<td>IPSV</td>
<td>School welfare, behaviour and atten</td>
<td>0.0</td>
<td>0.83</td>
</tr>
<tr>
<td>IPSV</td>
<td>Learning disabilities</td>
<td>0.7</td>
<td>0.85</td>
</tr>
<tr>
<td>IPSV</td>
<td>Hyperactivity</td>
<td>0.82</td>
<td>0.6</td>
</tr>
<tr>
<td>IPSV</td>
<td>Behaviour disorders</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>IPSV</td>
<td>Hyperactivity</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>IPSV</td>
<td>Behaviour support services</td>
<td>0.66</td>
<td>0.77</td>
</tr>
<tr>
<td>IPSV</td>
<td>Young people</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>IPSV</td>
<td>Children</td>
<td>0.0</td>
<td>0.73</td>
</tr>
<tr>
<td>IPSV</td>
<td>Lone parents</td>
<td>0.0</td>
<td>0.72</td>
</tr>
<tr>
<td>IPSV</td>
<td>Disabled children</td>
<td>0.0</td>
<td>0.69</td>
</tr>
<tr>
<td>IPSV</td>
<td>Emotional problems</td>
<td>0.5</td>
<td>0.63</td>
</tr>
<tr>
<td>IPSV</td>
<td>Mental disorders and illness</td>
<td>0.0</td>
<td>0.68</td>
</tr>
<tr>
<td>IPSV</td>
<td>Psychiatry</td>
<td>0.62</td>
<td>0.67</td>
</tr>
<tr>
<td>IPSV</td>
<td>Care</td>
<td>0.5</td>
<td>0.67</td>
</tr>
<tr>
<td>IPSV</td>
<td>Children in care</td>
<td>0.0</td>
<td>0.66</td>
</tr>
</tbody>
</table>

### Model Statistics (run IPSV v4)
- Avg number of docs per term: 1.35
- Standard deviation: 1.19
- Max number of docs per term: 10
- Number of scoring terms: 330
- Most popular terms: Population and migration

### Corpus Statistics (run IPSV v4)
- Avg no of terms per tagged doc: 9.55
- Standard deviation: 8.21
- Max number of terms per doc: 50
- Number of tagged documents: 56
- Number of all documents: 59
- Most termed documents: Treating Problem Behaviour in Children.pdf
SAS ENTERPRISE CONTENT CLASSIFICATION

Formerly Teragram
Full enterprise-class ontology manager
Rules-based
Includes concept/entity extractor
Very granular rules control
Pre-built taxonomy kits
APIs to SharePoint, Endeca, Fast, Documentum
Support for 27 languages
RULE-BASE EDITOR
EXAMPLE : SAS

You can define Predicate rules, using interrelated Concepts, to locate Facts.
SHAREPOINT

Automatic application of metadata & content types
Move documents to different libraries
Assisted or automatic indexing
Within SharePoint or in other Office tool
CONCEPT SEARCHING EXAMPLE
HOW TO EVALUATE?

Clear use cases & goals
Scorecard approach
Provide sample taxonomy & data for demos!
  • Don’t settle for IPSV demos that have been tweaked over years
Cost (esp. prof services) & set-up time
User-interface
Integration with other systems (CMS, search, etc.)

Note: some tools are stronger in taxonomy management than others; consider using connectors between a taxo tool and an auto-classifier
KEEP IN MIND...

+/- 60% accuracy out of the box

Success depends on:

- Richness of taxonomy
- Quality of documents
- Time spent building complex rules
THANK YOU!

Stephanie Lemieux
Dovecot Studio
stephanie@dovecotstudio.com
514-804-7833