[Slide 1] Current State of Automated Content Tagging

Joseph Busch

[Slide 2] AI, IR and Automated Tagging

About a year ago I went to a conference where there was a lot of discussion about Artificial Intelligence – AI. Several vendors gave presentations, but there was not a lot of detail about what they considered AI and how this differed from Information Retrieval or IR. IR is the use of technology to process content to identify patterns based on a combination of Natural Language Processing (NLP) and statistics.

[Slide 3] The natural world is filled with patterns. Some patterns may be inherent in the nature of things, consider a snowflake. **[Slide 4]** Other patterns may be overlaid on the natural world to make sense of it in human terms, consider the constellations of the zodiac. **[Slide 5]** Yet others may be overlaid on the human-made world to make sense of large amounts of information, **[Slide 6]** consider file naming and filing systems.

[Slide 7] People create guidelines and procedures to try to make naming and description complete and consistent. **[Slide 8]** But studies have shown that people are inherently inconsistent.[i] Even the same person on different days or at different times of day make different choices when they categorize and describe content. It's interesting to note that IR methods perform somewhat more consistently than people. But the mistakes they make are often more extreme, but they are also more consistent and more easily discovered than the inconsistencies made by human indexers.

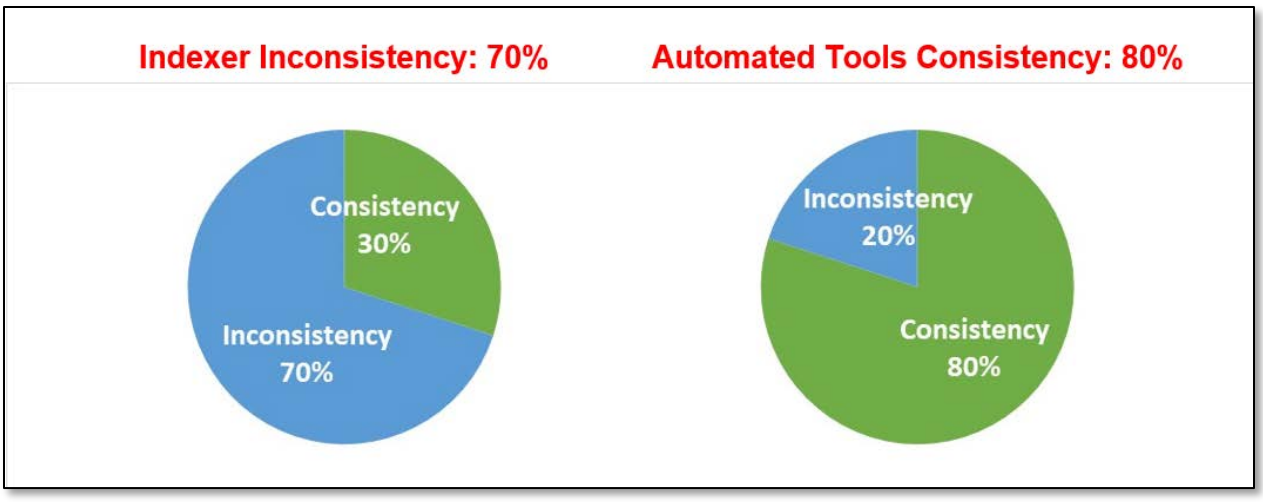

Figure 1-Human indexers vs. automated methods.

[Slide 9] A better scenario is to use automation to process content and then to let human experts review and "approve or improve" the results. This leads to more consistency, and the improvements made by humans provide feedback that can be used to improve the algorithms over time.

So, what's all this talk about AI really about? I think that much of what's called AI is really automation, not autonomous machine intelligence – although applications may sometimes appear to be smart. There are some important trends that are encouraging this type of large scale automation.

[Slide 10] The first trend is Cloud Computing. Standing up data processing intensive applications used to require appropriately specifying and acquiring a powerful computer and lots of data storage capacity, as well as installing and configuring complicated software and aggregating large content collections. All of that infrastructure had to be maintained. Today you can subscribe to cloud services that provides whatever capacity is needed, and it can be extended.

[Slide 11] The second trend is readily available IR software as a service. IBM Watson is a well-known example of so-called AI software as a service. IBM offers a bunch of solutions and tools including IR tools as the IBM Cloud (formerly branded as Blue Mix). Other vendors large and small offer SAAS that includes IR services such as GE Predix, Lexalytics, Data Harmony, Expert System, Mondeca, PoolParty and Smart Logic.

[Slide 12] The third trend is the Internet of Things or IOT. More and more commercial and consumer machines and devices are gathering and sharing data. This may be a big industrial machine such as a GE turbine in a power plant or a Rolls-Royce jet engine, or a Samsung smart refrigerator or an Amazon Echo, or an Apple iPhone or Dell laptop computer.

[Slide 13] With all this data, processing power, and ready access to IR tools what can we do with content to automatically extract useful metadata? Here's a quick summary of some of the common IR methods.

- Entity extraction is a method to identify named entities such as people, organizations, places, events, themes, etc.
- Sentiment analysis is a method to generate a relative positive or negative measure of the tone or intent of a content item.
- Keyword extraction is a method to identify and rank meaningful keywords and phrases in content.
- Summarization is a method to identify and extract key sentences from content that summarize its content.
- Predefined Boolean query is a method to generate rules to identify predefined topics that are relevant descriptors for a content item.
- Training set is a collection of discrete examples that are used to identify predefined topics fully or semi-automatically that are relevant descriptors for a content item.
- Statistical categorizer is a method to automatically identify words and phrases that are closely associated with pre-defined topics.

AI is simply intelligence exhibited by machines. Today, most AI is based on the processing of large collections of content that are being created by people and their interaction with each other (Social Networks) and machines (Applications), as well as machines doing the functions they have been built to perform (IOT).

AI uses IR to process collections of content and data. The difference is that the patterns in the results are used as rules by machines rather than people. These rules are called machine learning. This works well with some applications where patterns are less ambiguous or nuanced by cultural bias, such as machine telemetry or image recognition, but not so well with language. It all depends on the nature and quality of the examples used to train the algorithms. In fact, this is no different than training IR statistical categorizers. Having discrete examples is critical to accuracy. As far as I can tell AI is the automation of IR.

[Slide 14] Pre-defined Boolean Queries

[Slide 15] In machine learning, all you need to provide is lots of content. The system figures out what it's about. But the problem with machine learning is that is opaque, it's difficult to understand why an item is considered relevant. Categories are generic, may be irrelevant, can be biased, and are difficult to change or tune.

But what if you want to categorize a collection against a set of pre-defined categories? One way to do this is to develop a set of Boolean queries that scope the context for each category. This is much more transparent than machine learning, and it provides relevant categories. But it requires a lot of work to set up, and specialized skills.

[Slide 16] A Boolean query is a type of search that combines keywords or phrases with AND, OR, and NOT operators.

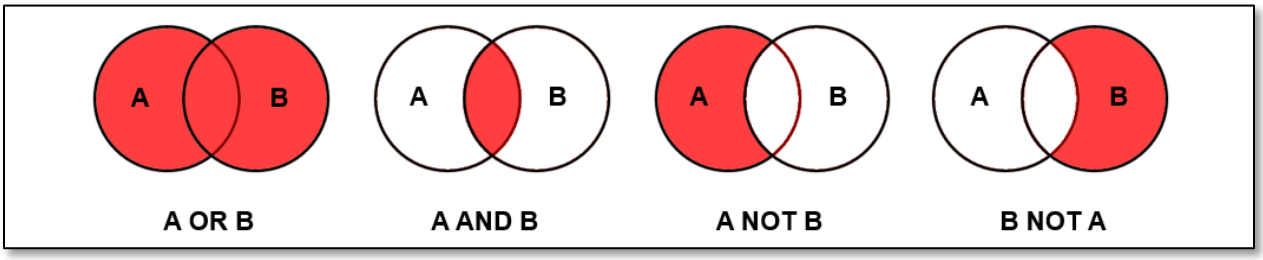

Figure 2-Boolean query types illustrated using Venn diagrams.

[Slide 17] Boolean queries are often used with proximity search. Proximity searching is a way to search for two or more words that occur within a certain number of words from each other, or within a section of a document. Unfortunately, Proximity operators and syntax are not standardized.

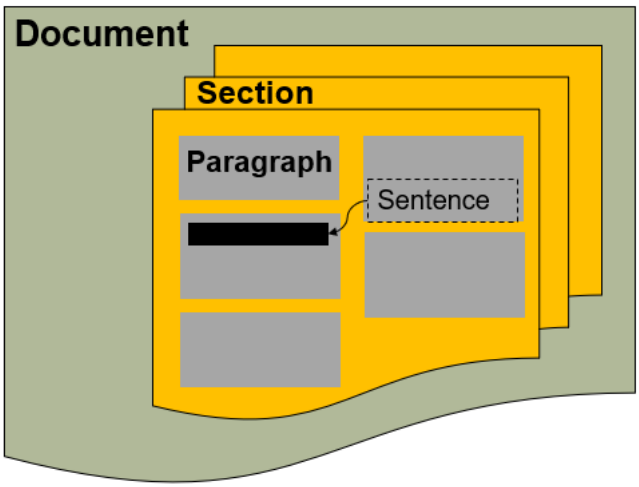

Figure 3-Proximity searching specifies where query terms are located in documents.

[Slide 18] The query syntax for Boolean queries also includes bounded phrases usually with quotations; right, left, and internal truncation; and nested statements with parentheses that match up.

[Slide 19] Here's a method that can be used to develop a Boolean query for a pre-defined topic.

1) Brainstorm a list of 10 relevant words and phrases.
2) Use that list to identify 10 relevant items (articles, videos, websites, etc.) E.g., do a Google search, search Google Scholar, search the NYT (or any other newspaper that you subscribe to), search Library of Congress Chronicling America (1789-1963), etc.
3) Review 10 relevant items and write down the words and phrases that provide a context for the theme/topic/concept. Titles, headings, summaries, introductions (at the beginning) and conclusions (at the end) are good areas to focus on without having to read the whole thing.
4) Note any named entities (people, organizations, events, laws, etc.) that are closely associated with the theme/topic/concept. E.g., for gun violence Gabrielle Giffords, Michael Bloomberg, Doctors Against Gun Violence, March for our Lives, etc.

[Slide 20]

5) Consolidate the terms. Identify duplicates, synonyms, as well as any concepts that you want to combine even if they are not synonyms. Re-label the term as needed to reflect the concept/category. Also consider and note any other relationships between terms. Prioritize the terms. Rank from 1-N, most relevant to least relevant. Hint: Rank each term by higher, medium, lower relevance, then sort and rank from 1-N.
6) Write a query for each term. Note that regular plurals (-s, -es, -ies) are usually (but not always) included automatically, but you always need to specify irregular plurals, e.g., "mice".
7) Qualify the scope for each term. Does the term require any qualification of the scope, e.g., by population, setting, geography, etc.? Validate that the term is disjunctive, distinct, and requires no further qualification.
8) Combine the terms into a single nested query with an OR operator.

[Slide 21] Case Study

[Slide 22] The Robert Wood Johnson Foundation (RWJF) is the largest philanthropy dedicated solely to health in the United States. Taxonomy Strategies has been working with RWJF to develop an enterprise metadata framework and taxonomy to support needs across areas including program management, research and evaluation, communications, finance, etc. We have also been working with RWJF on methods to apply automation to support taxonomy development and implementation within their various information management applications.

Last year we developed a pilot categorizer for 4 pre-defined Topics that describe some of the focus areas for RWJF programs and grantmaking:

- Childhood Obesity,
- Disease Prevention and Health Promotion,
- Health Care Quality, and
- Health Coverage

using Lexalytics Semantria. The goal was to determine the feasibility of building pre-defined Boolean categorizers for RWJF Topics. How difficult would it be to scope the context for each category? How accurate and comprehensive would the categorizer be?

[Slide 23] We identified a test collection of content items with known correct Topics—400 short form items which were summaries published on the RWJF.org website. **[Slide 24]** We built-up Boolean queries for the four target RWJF Topics using the method described in the previous section. This was done using a text editor, then we cut and pasted them into the Semantria Web user interface. Semantria validated the queries' syntax and either successfully loaded them or returned error messages which needed to be resolved. Eventually each of the four queries was successfully loaded. Short form items were then categorized using the Semantria for Excel plug in. In this case, results were returned and evaluated in Excel.

[Slide 25] Overall, the results from both test collections and user interfaces had 89% precision but only 67% recall. Meaning that only 11% of the results were false positives, but only 67% of the total collection was categorized. Generally, recall and precision measures work in opposition to each other. Meaning that decreasing precision would increase recall, and in this instance a fully optimized implementation would have approximately 78% precision and 78% recall. Our target was 80%.

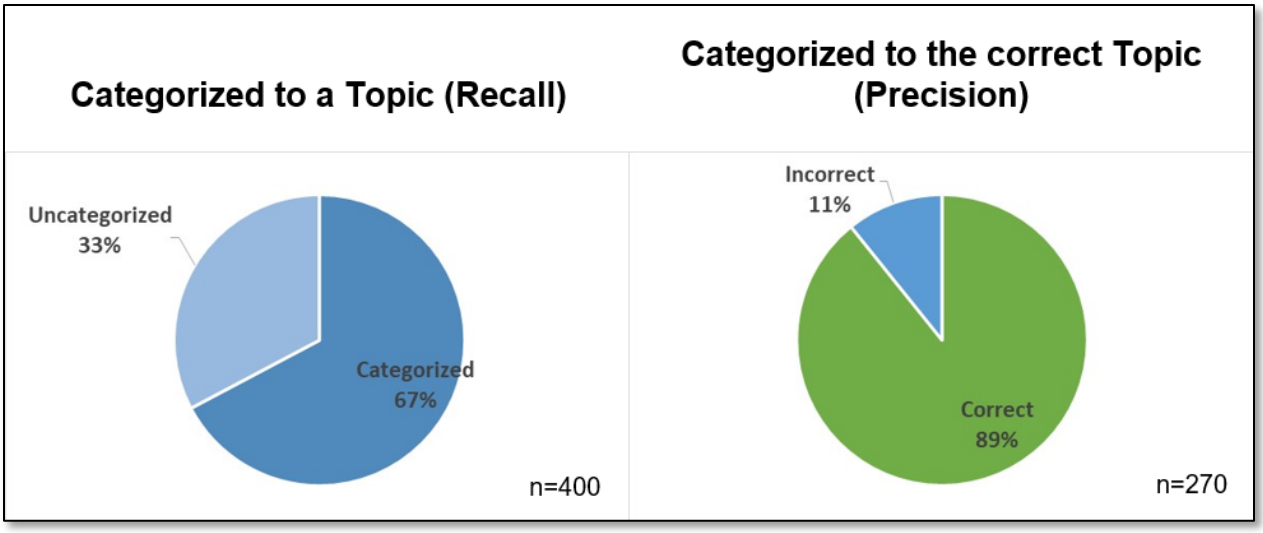

Figure 4-Overall trial results.

[Slide 26] Looking at the trial results for each RWJF Topic showed that the most precise results were for Health Care Quality and Health Care Coverage, and the least precise results were for Childhood Obesity and Disease Prevention and Health Promotion. But overall, the results were impressive given that the Topics are broad and potentially ambiguous. Based on these results it was decided to build and operationalize a categorization service for all of the RWJF Topics. I'll be talking about the results of this next project at the DCMI-18 meeting in Porto in September.

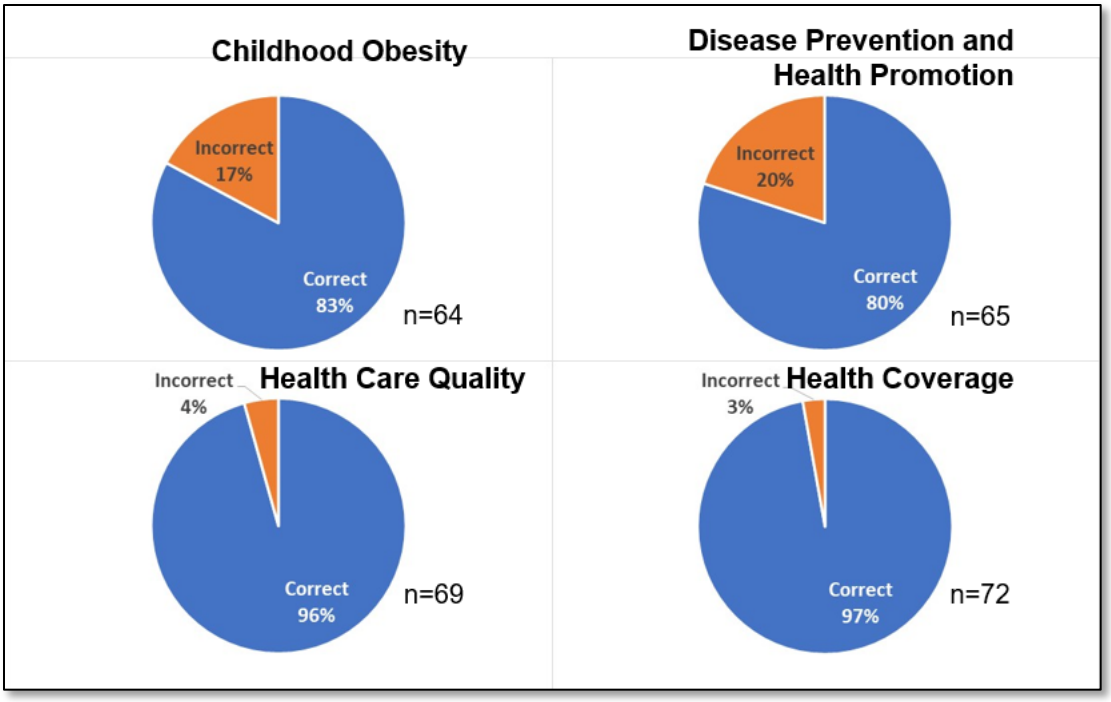

Figure 5-Trial results for each Topic.

[Slide 27] Summary

Much of what is being called AI is based on natural language processing (NLP) and IR methods. There is renewed interest in using IR methods to tag content so that it can be aggregated, analyzed and more effectively used by organizations. Cloud computing, software as a service (SAAS), and the Internet of Things (IOT) have removed some of the barriers to automated categorization. **[Slide 28]** While the tools and applications exist and are affordable, good implementation skills are hard to find. Training and expertise such as that in the Dublin Core community are needed.

[Slide 29] More Information

Performance Comparison of 10 Linguistic APIs for Entity Recognition. https://www.programmableweb.com/news/performance-comparison-10-linguistic-apis-entity-recognition/elsewhere-web/2016/11/03. Last checked 7/4/2018.

Top 27 Free Software for Text Analysis, Text Mining, Text Analytics. http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/. Last checked 7/4/2018.

Is there any free tool available for text classification? https://www.quora.com/Is-there-any-free-tool-available-for-text-classification. Last checked 7/4/2018.

Satnam Alag. Collective Intelligence in Action. https://www.manning.com/books/collective-intelligence-in-action. Last checked 7/4/2018.

Haralambos Marmanis and Dmitry Babenko. Algorithms of Intelligent Web. https://www.manning.com/books/algorithms-of-the-intelligent-web. Last checked 7/4/2018.

[Slide 30]

---

[i] "The importance of the indexing task in IR has led to much interest in studies of *inter-indexer consistency*---i.e., of the extent to which agreement exists among different indexers on the sets of index terms to be assigned to individual documents. These studies have consistently concluded that recorded levels of consistency vary markedly, and that high levels of consistency are rarely achieved.[1] Similar levels of inconsistency are routinely observed in the manual execution of related cognitive tasks,[2] such as the selection of terms to use as names for concepts or objects,[3] the formulation of queries for searching document databases,[4] and the estimation of relevance of the documents retrieved in such a search.[5] The insertion of links in hypertext documents may be viewed as being analogous to the assignment of index terms to such documents:[6] the present paper summarizes and discusses the main results of a study[7] that drew this analogy in seeking to determine the extent to which different people produced similar link structures for the same hypertext documents." Furner, J., Ellis, D., and Willet, P. "Inter-linker consistency in the manual construction of hypertext documents." *ACM Computing Surveys* (December 1999)

[1] Leonard, L.E. *Inter-indexer consistency studies, 1954-1975: a review of the literature and summary of study results*. Graduate School of Library Science, University of Illinois, Urbana-Champaign, IL, 1977.
[2] Saracevic, T. "Individual differences in organizing, searching and retrieving information." In: *Proceedings of the 54th ASIS Annual Meeting*, 82-86, 1991.
[3] Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. "The Vocabulary problem in human-system communication." *Communications of the ACM*, 30, 964-971, 1987.
[4] Bates, M.J. "Subject access in an online catalog." *Journal of the American Society for Information Science* 37, 357-376, 1986.
[5] Lesk, M.ER and Salton, G. "Relevance assessments and retrieval system evaluation." *Information Storage and Retrieval* 4, 343-359, 1969
[6] Liebscher, P. "Hypertext and indexing." In: Challenges in Indexing Electronic Text and Images, Learned Information, Medford, NJ, 82-86, 1994.
[7] Furner, J., Ellis, D. and Willett, P. "The Representation and comparison of hypertext structures using graphs." In: Information Retrieval and Hypertext, Kluwer, Norwell, MA, 75-96, 1996.