# Why Build Custom Categorizers Using Boolean Queries Instead of Machine Learning? Robert Wood Johnson Foundation Case Study

Presentation

Joseph A Busch
Taxonomy Strategies, USA
jbusch@taxonomystrategies.com

Vivian Bliss
Taxonomy Strategies, USA
vbliss@taxonomystrategies.com

**Keywords:** automated categorization; Boolean query categorization; auto-classification; text analytics; recall and precision; Robert Wood Johnson Foundation

## Abstract

This abstract provides an update on a project to build a Boolean query categorizer against a set of pre-defined broad categories for the Robert Wood Johnson Foundation (RWJF) a philanthropy dedicated to impacting health and health policy in the United States. Lessons learned building out the categorizer to make it scalable and maintainable are discussed.

## 1. Pre-defined Boolean Queries

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 it 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.

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.

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

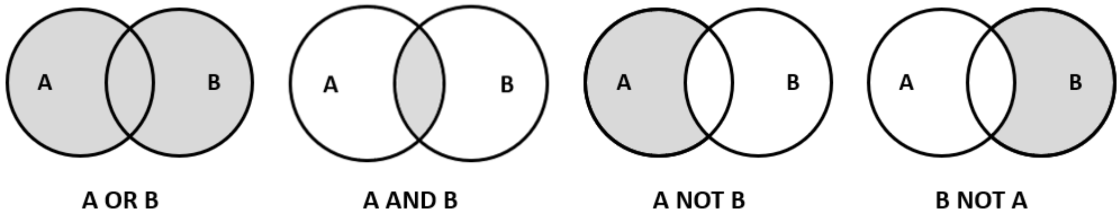


FIG. 1. Boolean query types illustrated using Venn diagrams.

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. 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.

https://doi.org/10.23106/dcmi.952139142

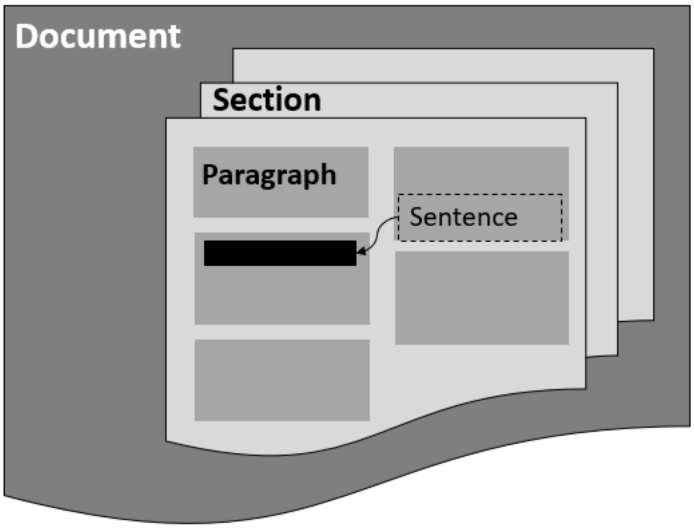


FIG. 2. Proximity searching specifies where query terms are located in documents.

## 2. Case Study

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.

The initial target application for automated categorization is RWJF grant "precis" which are short descriptions of funded projects. Over the last five years, RWJF has made awards ranging from $3,000 to $23 million with time periods ranging from one month to five years. However, most grants are in the $100,000 to $300,000 range, and run from one to three years. (RWJF, 2018) RWJF grants are currently described with metadata including: Program Areas, Types of Support, Grantmaking Interventions, Demographics, Topics and Tags. But the existing descriptive metadata are difficult to use to accurately answer questions about grantmaking trends, thus staff do not use it. Taxonomy Strategies is working on a new metadata scheme and taxonomy to replace the current descriptive metadata. Automated methods will be critical for updating descriptive metadata from the current to the new metadata scheme and values.

In 2017, Taxonomy Strategies 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. (Lexalytics, 2018) This case study was presented in a DCMI Webinar on July 19, 2018. (Busch, 2018)

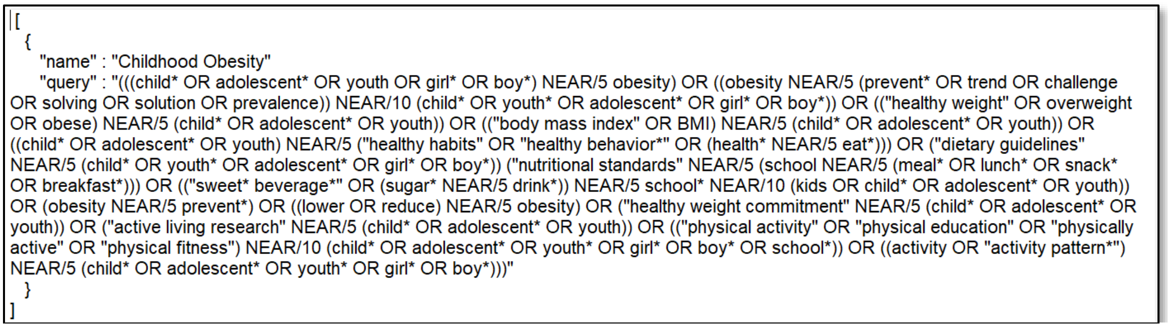
In 2018, Taxonomy Strategies is working with RWJF to: (1) develop requirements for, and suggest how to integrate text analytics and information retrieval software into RWJF staff workflows; (2) develop requirements for, and suggest how to build test collections for refining recall and precision for auto-classification; and (3) develop recommendations for staff roles and processes to support categorization of legacy assets and incoming grantee products.

### 1.1. Breaking down broad topics into simple queries

In the pilot project, Taxonomy Strategies built-up Boolean queries for the four target RWJF Topics. This was done using a text editor as shown in FIG. 3, then the complex query was cut and pasted into the Semantria Web user interface. Semantria validated the queries' syntax and either

https://doi.org/10.23106/dcmi.952139142

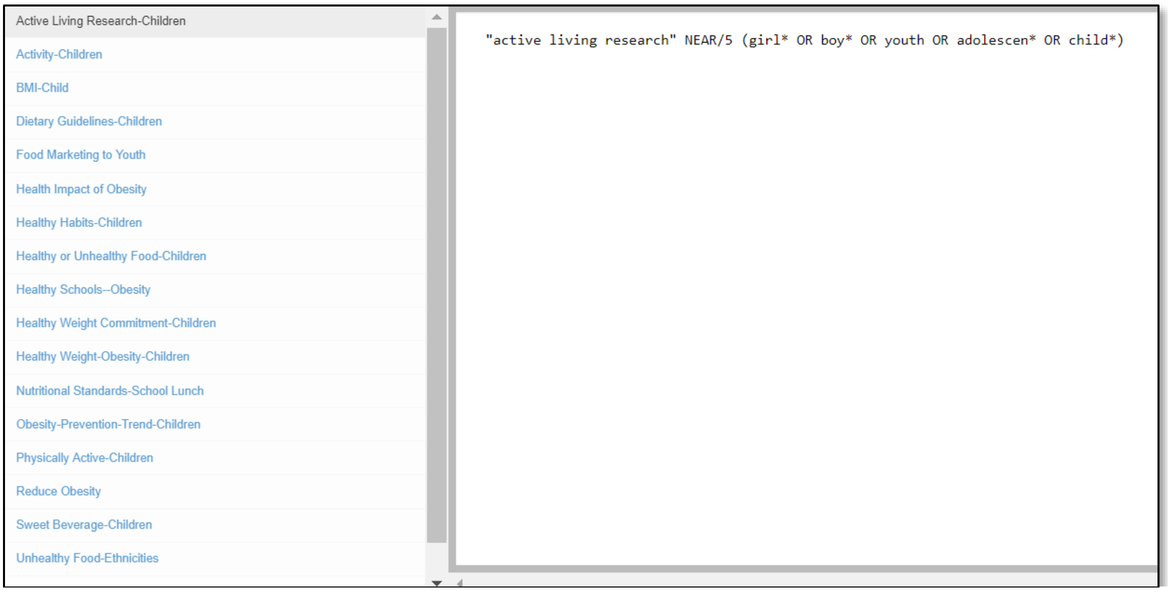
successfully loaded them or returned error messages which needed to be resolved. Eventually each of the four queries was successfully loaded.



```
[
  {
    "name" : "Childhood Obesity"
    "query" : "(((child* OR adolescent* OR youth OR girl* OR boy*) NEAR/5 obesity) OR ((obesity NEAR/5 (prevent* OR trend OR challenge
OR solving OR solution OR prevalence)) NEAR/10 (child* OR youth* OR adolescent* OR girl* OR boy*)) OR (("healthy weight" OR overweight
OR obese) NEAR/5 (child* OR adolescent* OR youth)) OR (("body mass index" OR BMI) NEAR/5 (child* OR adolescent* OR youth)) OR
((child* OR adolescent* OR youth) NEAR/5 ("healthy habits" OR "healthy behavior*" OR (health* NEAR/5 eat*))) OR ("dietary guidelines"
NEAR/5 (child* OR youth* OR adolescent* OR girl* OR boy*)) ("nutritional standards" NEAR/5 (school NEAR/5 (meal* OR lunch* OR snack*
OR breakfast*))) OR (("sweet* beverage*" OR (sugar* NEAR/5 drink*)) NEAR/5 school* NEAR/10 (kids OR child* OR adolescent* OR youth))
OR (obesity NEAR/5 prevent*) OR ((lower OR reduce) NEAR/5 obesity) OR ("healthy weight commitment" NEAR/5 (child* OR adolescent* OR
youth)) OR ("active living research" NEAR/5 (child* OR adolescent* OR youth)) OR (("physical activity" OR "physical education" OR "physically
active" OR "physical fitness") NEAR/10 (child* OR adolescent* OR youth* OR girl* OR boy* OR school*)) OR ((activity OR "activity pattern*")
NEAR/5 (child* OR adolescent* OR youth* OR girl* OR boy*)))"
  }
]
```

FIG. 3. Broad topic Boolean query from 2017 pilot project

In 2018, the process was modified to break up the broad topics into sets of simple queries. The goal was to make the queries more transparent, easier to "read", and easier to maintain as shown in FIG. 4. By "factoring" broad topics in constituent contextual parts, the simple queries could be combined and reused in different contexts. Working with simple contextual queries also facilitated "tuning" to optimize recall and precision.


Active Living Research-Children
Activity-Children
BMI-Child
Dietary Guidelines-Children
Food Marketing to Youth
Health Impact of Obesity
Healthy Habits-Children
Healthy or Unhealthy Food-Children
Healthy Schools--Obesity
Healthy Weight Commitment-Children
Healthy Weight-Obesity-Children
Nutritional Standards-School Lunch
Obesity-Prevention-Trend-Children
Physically Active-Children
Reduce Obesity
Sweet Beverage-Children
Unhealthy Food-Ethnicities

```
"active living research" NEAR/5 (girl* OR boy* OR youth OR adolescen* OR child*)
```

FIG. 4. Broad topic Boolean query broken up into simple queries.

## 1.2. Content collections for query building and testing

Choosing the content collection is a very important step in query building and testing. Busch (1998) suggests a "snowball" method to build up a collection starting with a list of relevant words and phrases to identify a core set of relevant articles from authoritative sources. Then performing a rhetorical analysis of titles, headings, summaries, introductions (at the beginning) and conclusions (at the end) of the content items to build up a list of words and phrases and named entities. Iterating this process a few times and applying some editorial judgement can provide a first draft for a Boolean categorizer.

Alternatively, if a collection of already categorized content items exists, this can be analyzed to generate a first draft for a Boolean categorizer. However, pre-categorized content needs to be carefully assessed to determine if it is relevant and consistently categorized. In the case of RWJF, there was a collection of pre-categorized grant precis, but the quality and completeness of that

https://doi.org/10.23106/dcmi.952139142

categorization was not adequate. Among the anomalies discovered, were formulaic precis and indexing for certain Program Areas especially related to leadership development. The lesson learned is that in some cases, it may be better to build a new set of category examples, than to rely on pre-existing indexing.

## 1.3. Refining recall and then precision

Recall and precision tend to resolve in direct proportion to each other, meaning that generally given an increase in precision there is a comparable decrease in recall, and visa versa. The baseline from which refinements are made is very important. In the 2017 pilot project, the results had 89% precision but only 67% recall, meaning that only 11% of the results were false positives, but 33% of the total collection was not categorized at all. Looking at the trial results for each RWJF Topic shown in FIG. 5 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.

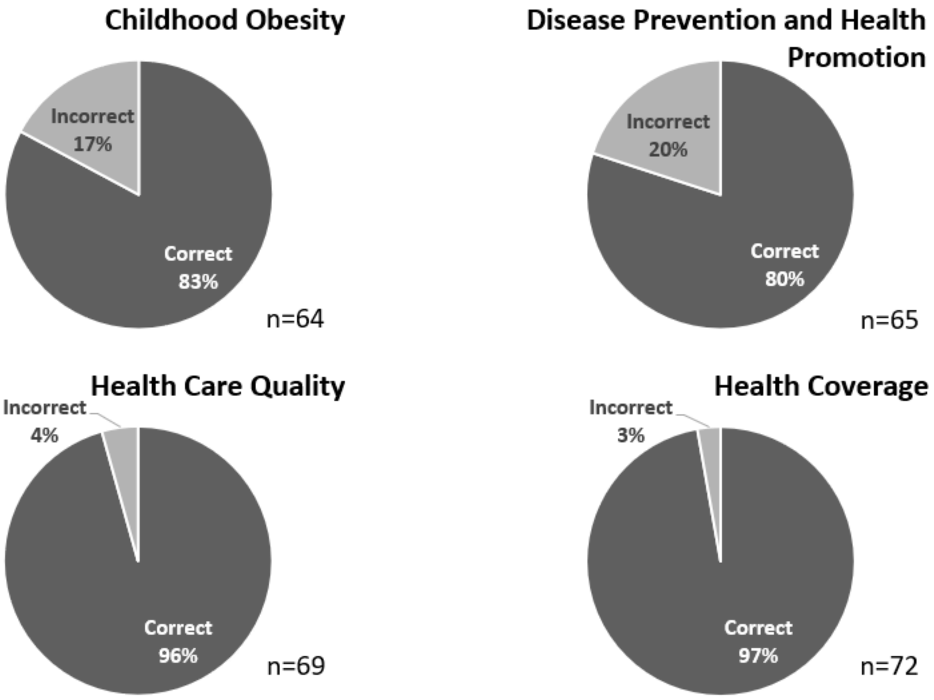


FIG. 5. 2017 pilot project results for each Broad topic.

In 2018, the process of refinement started with optimizing recall as much as possible in a first iteration of Boolean query building, and then optimizing for precision in a second iteration. While the focus of refinement is usually on precision, it is our opinion that optimizing recall is both easier and a better foundation for further refinement. This approach seeks to broaden the scope of the query and eliminate false negatives first to optimize recall, and then in a second iteration focus on the eliminating false positives to optimize precision.

## 1.4. Integrating text analytics into staff workflows

Beyond the development of the Boolean categorizers, developing requirements for integrating automated categorization into RWJF staff workflows raises questions about how these methods will change what people do. From the start, it was a goal to engage the Foundation's program staff directly in the process of categorizing content rather than to provide a fully automated solution to categorizing content. But this has led to some interesting discussions about who should be engaged

https://doi.org/10.23106/dcmi.952139142

in categorization including quality assurance. FIG. 6 shows one of the proposed workflow options for categorizing new grants.

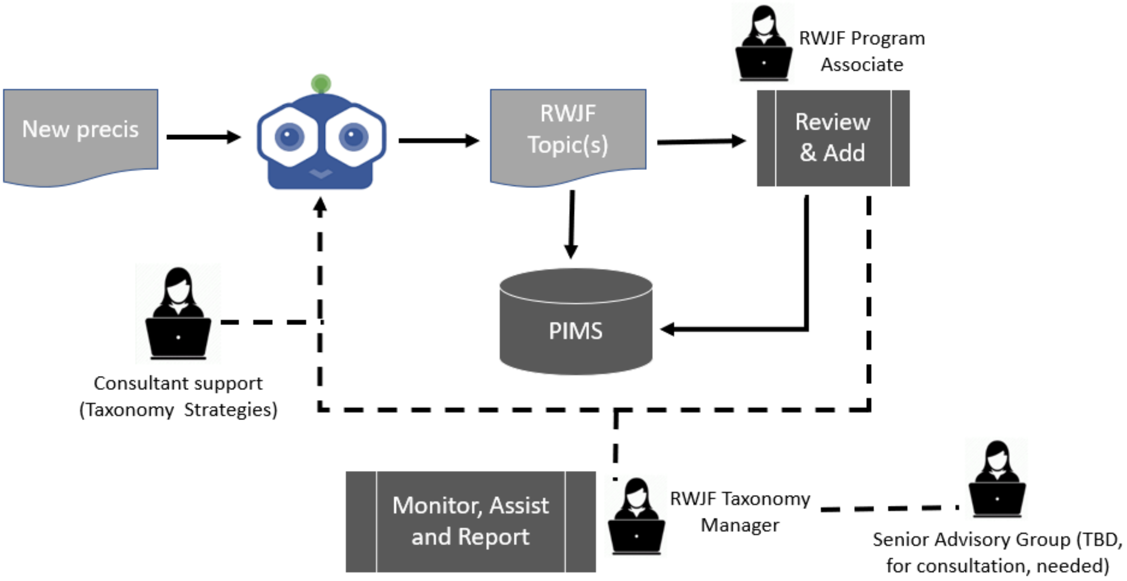

RWJF Program Associate
New precis
RWJF Topic(s)
Review & Add
PIMS
Consultant support (Taxonomy Strategies)
Monitor, Assist and Report
RWJF Taxonomy Manager
Senior Advisory Group (TBD, for consultation, needed)

FIG. 6. One proposed workflow option for categorizing new grants.

Retrospective re-categorization is planned to be a more automated process with a workflow to help users report errors, and a workflow to fix those errors and to inform users when the errors they reported have been fixed.

## 3. Conclusions

Working with RWJF over several years, some helpful lessons have been learned about automated categorization. These are that 1) breaking down broad topics into simple constituent queries facilitates the process of refining recall and precision by making the queries more easily understood and editable; 2) representative test collections are essential for building Boolean categorizers but even when pre-categorized collections exist they should be carefully evaluated for quality and usefulness; 3) it is effective to refine Boolean categorizers by optimizing recall before precision; and 4) automated methods should not replace staff but be a means to engage subject matter experts with content and categorization.

## References

RWJF. (2018). Frequently Asked Questions. Retrieved August 13, 2018. https://www.rwjf.org/en/how-we-work/grants-explorer/faqs.html.

Lexalytics. (2018). Semantria. Retrieved August 13, 2018. https://www.lexalytics.com/semantria.

Busch, Joseph. (2018). GoToWebinar - The Current State of Automated Content Tagging: Dangers and Opportunities. July 19, 2018. Retrieved August 13, 2018. Slides - http://www.taxonomystrategies.com/wp-content/uploads/2018/01/Current%20State%20of%20Automated%20Content%20Tagging-Webinar-20180719.pdf. Script - http://www.taxonomystrategies.com/wp-content/uploads/2018/01/Current%20State%20of%20Automated%20Content%20Tagging-20180719.pdf.

https://doi.org/10.23106/dcmi.952139142