Why Build Custom Categorizers Using Boolean Queries Instead of Machine Learning? Robert Wood Johnson Foundation Case Study
Joseph Busch and Vivian Bliss

Agenda

❖ Pre-defined Boolean Queries
❖ Case Study
❖ Lessons Learned
Boolean queries

- Basic operators
  - AND (conjunctive)
  - OR (disjunctive)
  - NOT (negation)

- Venn diagrams

  ![Venn Diagrams](image)

  - A OR B
  - A AND B
  - A NOT B
  - B NOT A

Proximity operators

- Proximity search (specified distance). Hint: Proximity operators and syntax are not standardized.
  - NEAR
  - NOT NEAR
  - FOLLOWED BY
  - NOT FOLLOWED BY
  - SENTENCE
  - FAR

![Proximity Operators Diagram](image)
Query syntax

- Bounded phrase
  - Usually quotation marks, e.g.
    "health insurance"
- Truncation (right, left, internal)
  - Usually an asterisk, e.g.
    child*
    "pre-existing condition**"
- Nested statements
  - Parentheses (that must match up)
    ("health insurance" AND (children* OR "pre-existing condition**"))

How to create a Boolean query (1)

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 item.
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.
How to create a Boolean query (2)

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 1-7.

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.

---

**Agenda**

- Pre-defined Boolean Queries
- Case Study
- Lessons Learned
Applying the RWJF taxonomy to grants

- **Grant awards**
  - $3,000 to $23 million
  - Mostly $100,000-$300,000

- **Grants period of performance**
  - 1 month to 5 years

- **Grants description**
  - Metadata including Program Area, Type of Support, Grantmaking Intervention, Demographics, Topics, and Tags
  - But metadata is difficult to use to understand answer grantmaking trends
New RWJF taxonomy objectives

- Better coding
- Better data
- Better insights
- Better usage
- Better integration

 Automated methods will be critical for updating descriptive metadata from the Current Taxonomy to the new metadata scheme and values (the Better Taxonomy)

2017 pilot project

Childhood Obesity  Disease Prevention and Health Promotion  Health Care Quality  Health Coverage
2018 project

- Operational text classifier using less complex Boolean queries
- Requirements for building test collections to refine recall and precision for auto-classification
- Methodology for refining recall and precision in stages
- Requirements for integrating text analytics and information retrieval software into RWJF staff workflow

Breaking down broad topics into simple queries

Broad topic Boolean query from 2017 pilot project

```

```

Broad topic Boolean query broken up into simple queries

```

```

```
Content collections for query building and testing

- **Option 1:** Build-up a test collection using the “snowball” method.
  - Use relevant words & phrases to identify a core set of relevant articles from authoritative sources
  - Perform rhetorical analysis to build up lists of words, phrases and named entities
  - Iterate with editorial judgement

- **Option 2:** If available, use an existing categorized collection.
  - Carefully assess existing categorized content to determine if it is relevant and consistently categorized, and representative of categorization application target content.
  - Potential problems:
    - Categorization is incorrect or over-tagged (more than 3 Topics)
      - E.g., RWJF.org topics are associated with grants based on a mapping from more detailed PIMS Topics, not directly from subject matter expert indexing.
    - Content is formulaic rather than distinct
      - E.g., RWJF leadership development grants often have a lot of boiler plate paragraphs in their executive summary.
    - Content is not representative of target collection
      - E.g., Only most recent content, not example from across the whole collection.

- **Option 3:** Re-index an existing representative collection.

Option 1: Building and Option 3: Re-indexing could take less time than compensating for problems with Option 2: Using an existing categorized collection.

- **Option 1:** Build-up a test collection using the “snowball” method.
- **Option 2:** If available, use an existing categorized collection.
- **Option 3:** Re-index an existing representative collection.
Overall 2017 pilot project results

Refinement is a tradeoff between recall and precision in categorization

Categorized to Pre-defined Category

Correct Pre-defined Category

2017 pilot project results for each Broad Topic

Childhood Obesity

Disease Prevention and Health Promotion

Health Care Quality

Health Coverage

n=64

n=65

n=69

n=72
Boolean query categorizer refinement process

- In first iteration, optimize recall as much as possible
- In second iteration, optimize for precision (and recall as necessary)

Optimize recall and then refine precision (and recall as necessary)
Integrating text analytics into staff workflows

Agenda

- Pre-defined Boolean Queries
- Case Study
- Lessons Learned
Lessons learned about building Boolean categorizers

- 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.
- Representative test collections are essential for building Boolean categorizers, but even when pre-categorized collections exist they may not be the best or most cost-effective option.
- It is effective to refine Boolean categorizers by optimizing recall before refining precision (and recall as necessary).
- Automated methods should not replace staff but be a means to engage subject matter experts (operational as well as senior level staff) with content and categorization.

References

Resources on recall and precision


Questions

Joseph Busch, jbusch@taxonomystrategies.com joseph@semanticstaffing.com (m) 415-377-7912

Vivian Bliss, vbliss@taxonomystrategies.com (m) 425-417-7628