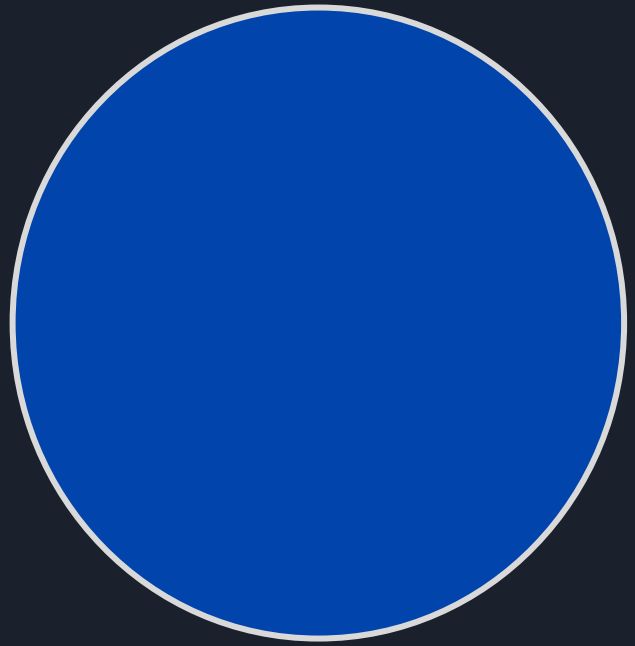
Using Keyword and Indexing Terms to Modify Document Clustering and Labels

Aaron Bohlmann



Objectives

- Modify Corpora
  - Keywords
  - Indexing Terms
  - Impact on clustering performance using PATTIE system
    - Manual clusters as the gold standard
      - Scoping review
    - Autonomous clustering using PATTIE
      - Corpus
      - Search terms
- Expanding search terms with PATTIE
  - Keywords
  - Indexing Terms
  - Effect on cluster label quality

Methods

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j|$$

Test Number	Document Sections Included in Corpus	Keywords and Indexing Terms Status	Search Terms	Results
1	Title, Abstract	None	Machine learning and medication adherence	Purity
2	Title, Abstract	Key words and indexing terms added	Machine learning and medication adherence	Purity
3	Title, Full-Text	None	Machine learning and medication adherence	Purity Label length descriptiveness key topics identified
4	Title, Full-Text	Key words and indexing terms added	Machine learning and medication adherence	Purity
5	Title, Full-Text	None	artificial intelligence, forecast model, machine learning, neural network, prediction modeling, medication adherence, medication nonadherence, medication noncompliance, Medication non adherence, medication persistence, medication compliance, Medication Non Compliance	Purity Label length descriptiveness key topics identified

Results

Test Number	Document Sections Included in Corpus	Keywords and Indexing Terms Status	Search Terms	Results
1	Title, Abstract	None	Machine learning and medication adherence	Purity = 51%
2	Title, Abstract	Key words and indexing terms added	Machine learning and medication adherence	Purity = 58%
3	Title, Full-Text	None	Machine learning and medication adherence	Purity = 70% Average Label length = 3.125 words Descriptiveness = Similar or less descriptive key topics identified = 4
4	Title, Full-Text	Key words and indexing terms added	Machine learning and medication adherence	Purity = 70%
5	Title, Full-Text	None	artificial intelligence, forecast model, machine learning, neural network, prediction modeling, medication adherence, medication nonadherence, medication noncompliance, Medication non adherence, medication persistence, medication compliance, Medication Non Compliance	Purity = 56% Average Label length = 5 words Descriptiveness = Sometimes more descriptive key topics identified = 4

Main Discussion Points /Conclusions

- Highest Purity (70%)
  - Full text corpus
  - With or without keywords / indexing terms
    - Possible explanation
      - Enough text to analyze
      - Key concepts can be determined
        - No need to add indexing terms and keywords to PATTIE
          - Increases system complexity
          - No benefit with full -text corpus
          - Future work: consider weight of keywords/indexing terms
- Expanding search terms using PATTIE
  - Longer labels
  - sometime more descriptive
  - Reduced purity
  - Did not improve
    - Key topic identification using label
  - No need to expand search terms for PATTIE
    - No benefit and increases system complexity
    - Future work: consider splitting clustering and cluster labeling in PATTIE system
      - Preprocessing of keywords / indexing term seemed to cause loss of some semantic meaning

References

1. Stefanowski , Jerzy, and Dawid Weiss. “Extending k-Means with the Description Comes First Approach.” *Matwbn*, Control and Cybernetics, 2007, matwbn.icm.edu.pl/ksiazki/cc/cc36/cc3647.pdf.
2. Manning, Christopher. “Introduction to Information Retrieval.” *Evaluation of Clustering*, 2013, nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html.