Simulating Time Saved for Manual Screening of Literature for IRIS Assessments Using Text Analytics

Michelle Cawley, Arun Varghese, Heidi Hubbard, Cara Henning | ICF

It all starts with the literature and…

Risk assessments developed for EPA's Integrated Risk Information System (IRIS) begin with a broad, comprehensive literature search to identify all relevant literature. IRIS literature searches often result in thousands and at times tens of thousands of results that must be screened for relevance.

reviewing literature takes a lot of time, but…

Distinguishing relevant literature from background noise is time and labor intensive and holds up the process of developing assessments. Literature searches for IRIS chemicals by design have low search precision to ensure high recall—most results are not relevant.

text analytics can save time without sacrificing recall.

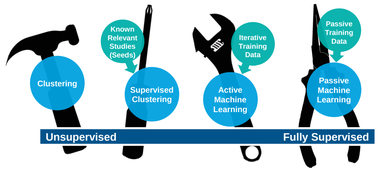
Text analytics methods have been used successfully to reduce the level of effort for manual screening. Here we simulate the use of supervised clustering with an ensemble approach coupled with machine learning to reduce time spent manually screening literature for multiple IRIS assessments.

Text Analytics Approaches Available in DoCTER

Text analytics methods available in DoCTER range from unsupervised algorithms (e.g., clustering) to fully supervised methods (e.g., machine learning).

Supervised machine learning algorithms perform well in terms of retrieval metrics such as recall and precision, but require a sizeable training dataset that is expensive to develop.

Supervised clustering algorithms have comparable retrieval efficiency and transparency to supervised algorithms, but require minimal training data.



Simulating Time Saved with Text Analytics Compared to Manual Screening

Benzo(a)pyrene | *n*-Butanol | *t*-Butanol | Trimethylbenzenes | Hexavalent Chromium | Arsenic

- We ran simulations on the six IRIS chemicals shown above using screening data available in EPA's Health and Environmental Research Online (HERO) Database.
- For each chemical, we used DoCTER to evaluate how well supervised clustering with a 9-model ensemble approach would have identified hazard ID literature (i.e., toxicology and epidemiology studies) using 100 randomly selected seeds (50 positive and 50 negative) (Method 1).
- For four chemicals (BaP, *n*BA, *t*BA, TMB) we also investigated whether additional time could be saved without a significant effect on recall by using a two-stage approach of supervised clustering followed by machine learning (Method 2 & Method 3).
- Time saved was calculated as the number of studies that were eliminated by DoCTER x 1 minute/study x 2 screeners/study.

Visualizing Supervised Clustering with a 9-Model Ensemble Approach

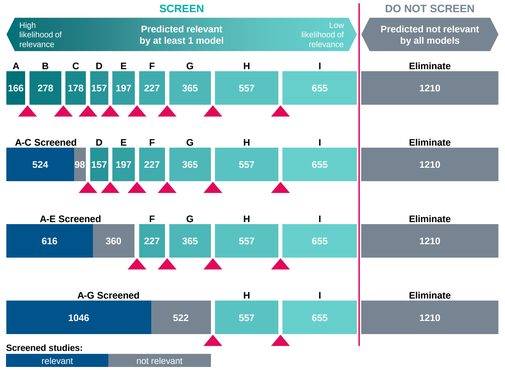
Using an ensemble approach, studies are predicted for relevance using multiple models. For this research we ran supervised clustering using nine models based on three cluster sizes (10, 20, and 30) and three algorithms:

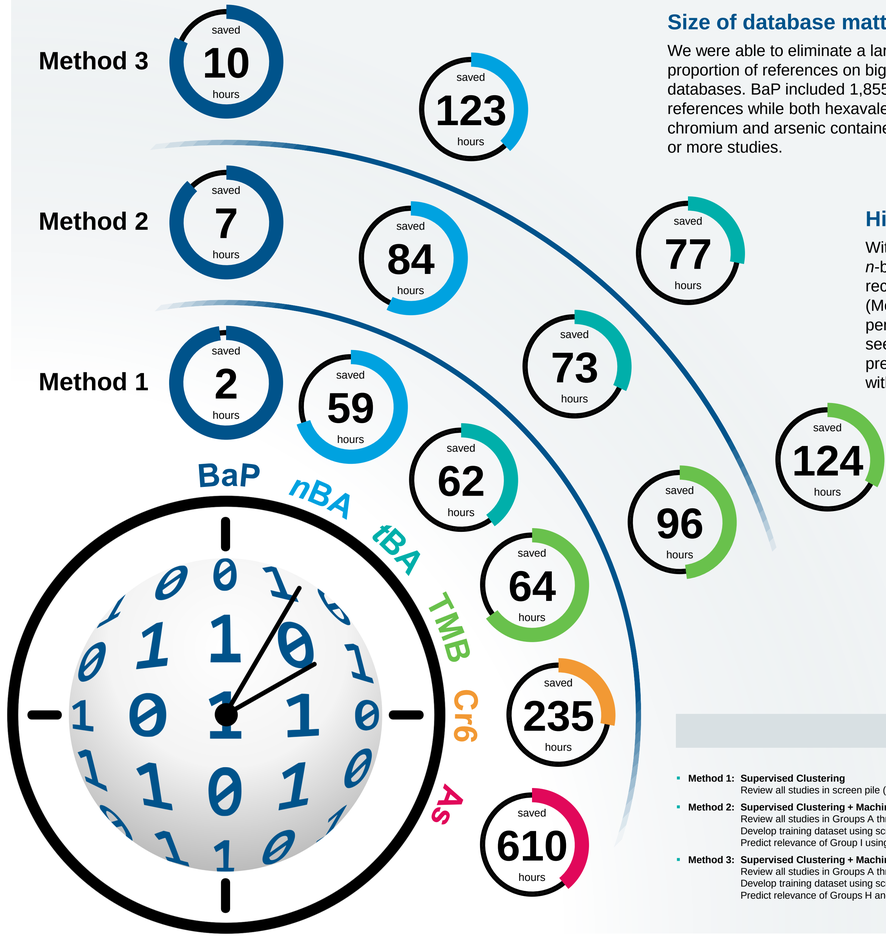
- Latent Dirichlet Allocation
- K-Means
- Non-Negative Matrix Factorization

After all models are run the database is split into two piles: Screen and Do Not Screen. Studies in the screen pile are further prioritized based on likelihood of relevance such that:

- Group A contains studies "voted" likely to be relevant by all nine models.
- Group I contains studies "voted' likely to be relevant by only one model.

Options at each ▲ checkpoint:
1. Keep screening
2. Stop screening
3. Modify Approach (e.g., switch to machine learning)





Size of database matters

We were able to eliminate a larger proportion of references on bigger databases. BaP included 1,855 references while both hexavalent chromium and arsenic contained 10,000 or more studies.

High recall is achievable

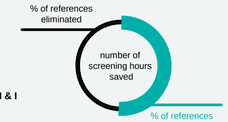
With the exception of Method 3 on *n*-butanol we reached over 90 percent recall of relevant hazard ID papers. (Method 3 on *n*-butanol reached 75 percent recall.) Using a chemical-agnostic seed set is an alternative approach for low precision search results like those found with *n*-butanol.

Interested in the details?

See Varghese, A; Cawley, M; and Hong, T. (2017). *Supervised clustering for automated document classification and prioritization: A case study using toxicological abstracts* in the journal Environment Systems and Decisions for a detailed discussion of supervised clustering using an ensemble approach.

Key

- **Method 1: Supervised Clustering**
  Review all studies in screen pile (i.e., Groups A through I).
- **Method 2: Supervised Clustering + Machine Learning on Group I**
  Review all studies in Groups A through H.
  Develop training dataset using screening results.
  Predict relevance of Group I using machine learning.
- **Method 3: Supervised Clustering + Machine Learning on Groups H & I**
  Review all studies in Groups A through G.
  Develop training dataset using screening results.
  Predict relevance of Groups H and I using machine learning.

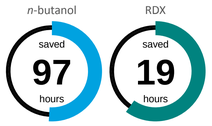


Predicted versus Actual Recall

5%

In each simulation, predicted recall and actual recall were closely aligned—they did not vary more than 5 percent, demonstrating that DoCTER predicted recall is reliable and unbiased.

Testing "Ready Made" Seed Sets: A Chemical-Agnostic Approach



- For search results with very few relevant results, finding enough seeds can be a problem.
- We simulated how well a chemical-agnostic seed set developed using known hazard ID studies for benzo(a)pyrene, Cr6, arsenic, and perchloroethylene was able to predict relevant references for *n*-butanol and RDX – both low precision datasets.
- Chemical-agnostic seed sets require a multistage approach of supervised clustering and machine learning. Using this two-stage process on *n*-butanol and RDX we achieved a 93 to 95% recall and eliminated 49 to 40% of studies from manual screening, saving an estimated 97 to 19 hours respectively.

What's Next for DoCTER?

- **DoCTER Online**, coming in 2018, will contain all DoCTER functions including Active Machine Learning.
- **DoCTER-Ex** is DoCTER's prototype for automated extraction.
- **New AI technology** being added to DoCTER to enhance performance.

DoCTER
Document Classification and Topic Extraction Resource

Contact: Michelle Cawley | michelle.cawley@icf.com or Arun Varghese | arun.varghese@icf.com