SMART CLUSTERING

A New Approach to Prioritizing Toxicological Literature for Systematic Review

Michelle Cawley, Jessica Wignall, and Arun Varghese | ICF

Dream big.
Then call ICF.

Contact
Michelle Cawley
Arun Varghese

michelle.cawley@icf.com
arun.varghese@icf.com

ICF

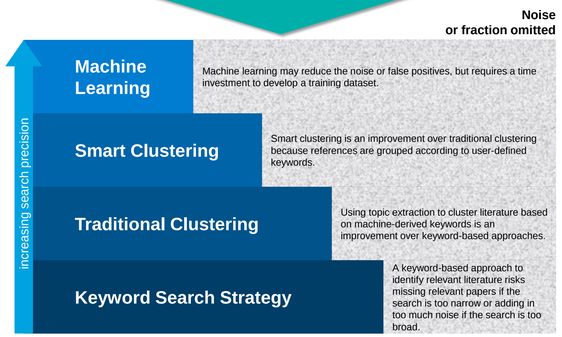
Get Smart
About Prioritizing References

The problem: Broad comprehensive literature searches conducted for systematic reviews may result in tens of thousands of results. Distinguishing relevant literature from background noise is time and labor intensive. Methods to reduce level of effort may sacrifice recall (i.e., the proportion of relevant references identified).

A new solution: With smart clustering a user-defined set of keywords is used to control how the literature is clustered. Our smart clustering algorithm automatically weights a set of topic words and each document in the unclassified corpus is then ranked according to relevance to the topic.

Why is it better?: Both traditional and smart clustering segregate an unclassified corpus of abstracts into a user-specified number of groups. However, with traditional clustering algorithms users have no control over where clusters center which may necessitate trial-and-error iterations until the topic of interest is identifiable in a focused set of clusters. Unlike clustering, machine learning requires an investment of time to create a training dataset. In some instances smart clustering may be a better option.

We asked: How does smart clustering perform in terms of recall compared to traditional topic clustering and machine learning? If clusters are created around user-defined keywords instead of machine-derived keywords can we reduce level of effort without sacrificing recall (i.e., missing relevant references)?



The Power of ToxSyn

What is ToxSyn?

ToxSyn is a toxicological semantic similarity discovery engine. It helps users discover synonyms, antonyms, hyponyms and hypernyms. It also provides quantitative scores of term similarity.

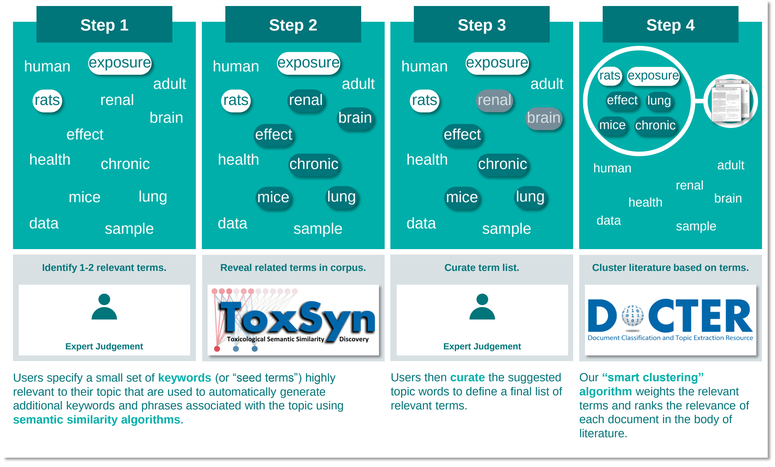
How does ToxSyn work?

ToxSyn assumes that words with similar neighbors or contexts are similar in meaning. Its underlying algorithm uses neural networks to predict the word vector that best defines each word in a given vocabulary. Measures of distance between these word vectors serve as metrics of term similarity.

What data does ToxSyn need?

ToxSyn learns by reading as much text as you make available. ToxSyn's default libraries include millions of pre-processed abstracts from toxicology, risk assessment, exposure science and other topics. Users can also choose to read in their own textual data.

Smart Clustering



Users specify a small set of **keywords** (or "seed terms") highly relevant to their topic that are used to automatically generate additional keywords and phrases associated with the topic using **semantic similarity algorithms**.

Users then **curate** the suggested topic words to define a final list of relevant terms.

Our **"smart clustering" algorithm** weights the relevant terms and ranks the relevance of each document in the body of literature.

Methods

- **Input Dataset:** We used a curated set of 8,225 references that had been previously classified for a chemical risk assessment. Studies were classified for relevance to genotoxicity and hazard identification. Genotoxicity and Hazard Identification studies represented approximately 10% and 4% of the corpus respectively.
- **Process**: We used three methods to identify papers from this corpus, including smart clustering, topic extraction, and machine learning. Each approach was repeated for hazard identification and genotoxicity.

Step 1 Convert corpus to numerical matrix using TF/IDF. Each word in a given document is assigned a score that represents its importance in distinguishing the document.

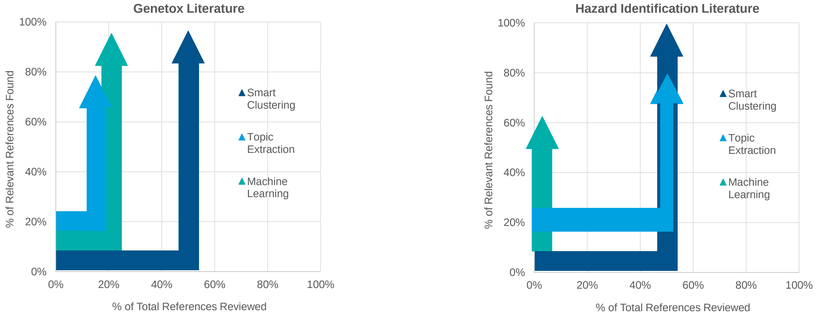
Step 2 Create a list of keywords to describe the topic. ToxSyn semantics similarity algorithm can be used to reveal related keywords. User may curate and weight keywords generated by ToxSyn.

Step 3 Create a topic score for each document based on weighted sum of the TF/IDF scores of each topic keywords from Steps 1 and 2.

Step 4 User selects an appropriate cutoff to achieve desired level of recall based on available resources.

Approach	Input Requirements	How it Works	Algorithms
Smart Clustering	Low	Each cluster is defined by a set of **user-defined keywords** that most closely define the group.	Boyer Moore Horspool Knuth Morris Plath
Topic Extraction (Traditional Clustering)	None	Each cluster is defined by a set of **machine-selected keywords** that most closely define the group.	k-Means Non-Negative Matrix Factorization (NMF) Latent Dirichlet Allocation (LDA)
Machine Learning	High	Using a **manually-developed training dataset** machine learning predicts relevance of all unclassified references.	Naive Bayes k-Nearest Neighbors Linear Support Vector Machines

Your Recall, Your Choice



- Smart clustering, like traditional clustering, has very low input requirements. We saw improved recall over topic extraction (traditional clustering) using smart clustering approach.
- We did not curate the keywords provided by ToxSyn for this simulation. We expect that results would be improved with an expert-curated keyword list.
- Machine learning may outperform clustering but comes with the tradeoff of requiring a training dataset.
- Smart clustering may be appropriate in some situations where a topic of interest does not cluster neatly using traditional clustering algorithms or when a training dataset is not available.
- We conclude that our smart clustering algorithm could offer significant time and cost savings to toxicologists engaged in systematic literature reviews.