Scientific Concept Evolution Tracker

Nic Bolton
University of Toronto
Toronto, Canada
nic@cs.toronto.edu

Abstract

With the sheer volume of research being published, the history and context of how scientific ideas evolve are often difficult to visualize through the noise. Terminology in science can be dynamic—the semantic meaning of terms such as “neural networks”, “entropy”, or “plasma” shift significantly over decades as new research subfields emerge. Traditional information retrieval systems and static vector databases index semantic meanings as fixed points in high-dimensional space, which flattens the temporal dimension and hides the evolutionary history of these concepts. This report introduces the Scientific Concept Evolution Tracker (SCET), a comprehensive system designed to ingest, index, and analyze large-scale scientific corpora to quantify this semantic drift. SCET is built for scale with PostgreSQL for storing metadata and Milvus for embeddings. We introduce a methodology that combines unsupervised clustering (K-Means) with temporal segmentation (Decision Tree Regression) to automatically identify distinct “eras” of a concept’s life cycle. We demonstrate the system’s capabilities through case studies, such as the divergence of “Transformer” from electrical engineering to natural language processing, and provide a quantitative analysis of system performance on a dataset sourced from arXiv.

1 Introduction

Science is a cumulative endeavor, yet the language of science is fluid. A core challenge in bibliometrics and the “Science of Science” is understanding how scientific consensus and terminology evolve. For a researcher entering a new field, understanding the historical context of a term is as critical as understanding its current definition. For example, a query for “Attention” in 2005 would yield results dominated by cognitive psychology and neurobiology. The same query in the late 2010s and early 2020s is overwhelmingly dominated by field of machine learning. This is the phenomenon that we are interested in, that is, being able to quantify how a concept’s relevance or association with specific fields change over time.

The introduction and combination of Large Language Models (LLMs) and vector databases have revolutionized semantic search. By representing text as dense vectors in a high-dimensional space, we can capture semantic similarity beyond simple keyword matching. However, most vector search implementations treat the document corpus as a static snapshot. They are designed to answer the question, “What is semantically similar to this query now?” rather than “How has the meaning of this query changed over time?”.

This project addresses the following question. How can we design a scalable vector database system that can quantify the semantic evolution of scientific concepts and identify pivotal publications that drive these shifts?

SCET was developed to answer this question. It consists of a pipeline that:

(1) Ingests and indexes scientific abstracts using a hybrid embedding strategy
(2) Retrieves context-aware results using a weighted hybrid search
(3) Clusters results into sub-concepts associated with a given query
(4) Produces a set of time periods that represent “stable” eras of a sub-concepts association/relevancy
(5) Identifies papers that are “pivotal” to the shift into these eras

The remainder of this paper is organized as follows. We first review the background and related work in NLP and bibliometrics. We then discuss the details of the methodology and system architecture, followed by a presentation of the experimental results and case studies. We conclude by discussing limitations and future work.

2 Background and Related Work

The problem to solve (tracking the evolution of scientific concepts) sits at the intersection of Natural Language Processing (NLP), Information Retrieval (IR), and the “Science of Science”. This section reviews the historical progression of these fields and the specific technologies that enable SCET.

2.1 The Evolution of Information Retrieval

The field of Information Retrieval has evolved through several distinct paradigms. An overview of the history is listed below. It is important to note that although these systems perform well within their own goals, they all lack the ability to capture semantic and contextual meaning within terms.

2.1.1 Boolean Logic. The earliest IR systems relied on semantics used within set theory, where queries were built with operators such as AND, OR, NOT. These systems could make retrievals at a high level of precision by using Inverted Indices. These work as a key-value store, that is, each term is a key that is associated with a list of pages that contain the term. This results in a very fast and precise system [3].

2.1.2 TF-IDF. Term Frequency Inverse Document Frequency (TF-IDF) introduced the concept of weighting the importance of words. This system involves calculating a score for a term, derived by balancing how frequently a term appears in a specific document (TF) against how rarely the term appears across the entire collection of documents (IDF). Consequently, these scores often favour the terms that best characterize the topics involved in a collection of documents. [1].

2.1.3 Vector Space Model. The Vector Space Model (VSM) generalized the idea of weighting terms by representing documents as vectors in a multi-dimensional space, where each dimension corresponds to a distinct term in the corpus. In this model, the relevance of a document to a query is measured by the similarity between their respective vectors. The similarity measurement is often done using Cosine Similarity

$$\cos(\boldsymbol{q}, \boldsymbol{d}) = \frac{\boldsymbol{q} \cdot \boldsymbol{d}}{\|\boldsymbol{q}\|\|\boldsymbol{d}\|} \tag{1}$$

for TF-IDF weights of the query $\boldsymbol{q}$ and the document $\boldsymbol{d}$.

This allows for ranked retrieval results based on the angle between vectors, rather than a binary inclusion/exclusion. However, traditional VSMs treat terms as orthogonal dimensions, meaning they cannot capture semantic relationships between synonyms (e.g., "car" and "automobile") without explicit expansion [1].

2.1.4 BM25. Okapi Best Matching 25 (BM25) represents a significant evolution in probabilistic information retrieval. While sharing similarities with TF-IDF, BM25 introduces two critical improvements: term saturation and document length normalization. Unlike TF-IDF, where the score increases linearly with term frequency, BM25 applies a saturation function so that the value of a term diminishes as it is repeated (thus, preventing long repetitions of keywords from dominating results). Additionally, it penalizes long documents (which naturally contain more terms) to prevent them from unfairly dominating search results. BM25 remains a strong baseline for lexical search tasks today.[2].

Acknowledgments

Acknowledgements go here. Delete enclosing begin/end markers if there are no acknowledgements.

References

[1] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press, USA.

[2] Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. 0–.

[3] Gerard Salton, Edward A. Fox, and Harry Wu. 1983. Extended Boolean information retrieval. *Commun. ACM* 26, 11 (Nov. 1983), 1022–1036. doi:10.1145/182.358466