

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 sub-fields 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

Papers do not have any fixed sections names. Three levels of heading are supported in ACM papers.

A test reference of Bush’s ‘As We May Think’ [1], to indicate how references are set up.

There is no need to save this file when editing as Overleaf auto-saves the document. Use the green ‘Recompile’ button in the upper menu bar to show or refresh the PDF. The output PDF is not saved into the project but can be downloaded, for reading outside overleaf, using a button at the right-hand end of the Overleaf menu-bar. To download the project files, use the home button in the menu bar to exit the project. Download options are available in your project listing.

Depending on your type of Overleaf account you will be able to share access with fellow authors—see Overleaf’s documentation for details. Shared access also allows for leaving comments in the project.

1.1 Do Not Remove Boilerplate Code

The TEX document includes code to generate sections such as the the copyright block. Do not remove these. Authors of accepted papers will receive sections of code to replace these and customise the final paper accordingly.

Pay attention to the code comments about author information and add/remove authors as necessary; there must be at least one author. It is desirable that all/most authors have an ORCID ID as this is replacing the need for explicit emails that my change as researcher move around. If an ORCID is supplied that author’s name is made the anchor text of a web link to their ORCID page.

Macro codes are offered (e.g. `authornote`) which may be used as well as indicating the corresponding author(s) for communication during publication.

1.2 General points on ACM papers

See the file ‘sample-sigconf.pdf’ provided with this template set. It explains the basics of a number of structural features.

Newer users of \LaTeX should take care to understand \LaTeX ’s special characters that need explicit escaping. Also be aware that typographic quotes and en-dash/em-dash hyphens are not used literally in TEX documents but are indicated with macros.

Overleaf has extensive documentation covering how to indicate typographic elements in \LaTeX code.

1.2.1 A Third-level Heading. Note that level-three headings are inset into the beginning in the first paragraph of that section, regardless of line breaks in the source document.

Sections at all three levels are auto-numbered, so these do not need to be numbered in your text.

Acknowledgments

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

References

- [1] The Atlantic 1945. As We May Think. *The Atlantic Monthly* 176, 1 (1945), 101–108 pages.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CSC2508: Advanced Data Systems, Fall 2025, University of Toronto

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>