21 July 2025 · CC-BY 4.0

Qeios

Peer Review

Review of: "Quantifying Hot Topic Dynamics in Scientific Literature: An Information-Theoretical Approach"

Chico Q. Camargo[1,2]

1. Department of Computer Science, University of Exeter, Exeter, United Kingdom; 2. The Alan Turing Institute, London, United Kingdom

This paper proposes a framework for quantifying "concept drift" in scientific literature by modeling scientific terms as nodes in a co-occurrence network and tracking their changes over time. By leveraging changes in node embeddings and neighborhood structures, the authors aim to capture semantic evolution and identify emerging, merging, or declining concepts. The work is motivated by the need to systematically measure conceptual shifts in fast-moving fields and is illustrated using examples from biomedical literature.

But as it is, the work is currently very hard to understand and disentangle.

Here are our recommendations to improve this work:

____

1. Distance/Metric

The NVI metric the author is using is actually the Rajski distance [1][2], defined in 1961, which satisfies the properties of a metric (triangle inequality, non-negativity, indiscernibility, and symmetry). For this metric, d(X,Y) = 0 does not imply X=Y though, but rather, it implies that X can be completely determined from Y. Knowing X eliminates all uncertainty about Y.

[1] - https://www.sciencedirect.com/science/article/pii/S0019995861800557

[2] - https://en.wikipedia.org/wiki/Mutual_information#Variations

2. Collocations

doi.org/10.32388/2887W4

The approach is vulnerable to collocation effects. For example, "artificial intelligence" as a bigram might collapse distances between "artificial" and "intelligence" if these terms are rare individually but frequent together.

A hierarchical decomposition (trigrams, bigrams, unigrams) could help mitigate this issue. Successive filtering of higher-order n-grams before lower-order terms would improve robustness.

3. Link prediction validation

Regarding predicting [new] links within a network of concepts: see the https://github.com/iarai/science4cast competition. This might be worth mentioning. Incorporating such benchmarks would strengthen claims about predictive capability and link emergence. And doing a bit of a literature review on approaches like that might give the reader a better idea of how much of the work here is novel and how much is similar to other past approaches.

4. Mathematical clarity and rigour

Variables such as (k,m) should be clearly defined where they first appear, rather than assuming familiarity. This way, the reader doesn't have to go around the rest of the manuscript trying to understand or confirm what (k,m) are.

And in Eq. (2), x and c seem to be referring to the same thing? I would use a single letter for both then. This is a bit unclear since the formula for P(x_k, t) does not involve x.

The time notation could be simplified: using "t" or "t_0 to t_0 + t" is clearer than "t_0 + t," which is potentially confusing.

Finally, terms like "stable conceptual hubs" need precise operational definitions. For example, do they refer to high-degree nodes with consistent neighborhood overlap? Including explicit examples (e.g., domain-specific jargon or methodological anchors) would enhance reader understanding.

5. Overall clarity

The Minimum Spanning Tree (MST) appeared a bit out of nowhere and wasn't explained in much detail. Can you expand that part? Why is the MST able to identify "the most influential conceptual transitions, filter out minor fluctuations, and highlight the dominant patterns of conceptual change"?

Figure 1 is not clear at all. It's pretty, but in practice, all the reader sees is a bunch of red and blue bubbles that seem to be connected to each other. I suggest either removing it or making a much simpler version, with maybe 3 years of data, and possibly fewer words. Otherwise, it just looks messy.

For instance, you mention "the temporal convergence of concepts toward 'cybersecurity'", but I took some time to understand what you were trying to represent. But if you made all links grey, except for the links to cybersecurity, that would maybe become clearer. Or maybe plot blue and red separately? Essentially, we need fewer lines on the figure.

The way it is, for every word mentioned, I had to go on the plot and spend some time trying to disentangle the image. This could be a lot easier if the words were in alphabetical order, if red and blue were separate plots... anything to reduce the complexity of the figure.

Same goes for all other figures in the same format.

6. Broader framing

The paper could benefit from situating its approach within related literature on semantic change detection (e.g., historical word embeddings, dynamic topic models), and also on approaches like the link prediction competition mentioned above. A comparison or acknowledgment of differences would strengthen theoretical grounding.

7. Minor points

Figures could be annotated more clearly to highlight key shifts or illustrate different hub types.

____

Reviewed by: **Chico Q. Camargo** and **Owen Saunders**

Declarations

Potential competing interests: No potential competing interests to declare.