26 July 2025 · CC-BY 4.0

Qeios

Peer Review

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

Jean-Charles Lamirel[1]

[1] University of Strasbourg, Strasbourg, France

This paper presents a method for detecting dynamic hot topics in datasets of research papers (scientific literature). The method supposes that the topics are known in advance or that they are produced by an external topic modeling method. The method is based on a co-occurrence graph in which frequencies are substituted by the normalized variation of information (a well-known information theoretic distance), and the dependencies that are examined are the ones between topic (core words) and the most associated word on a time-slice basis.

The approach is tested with 6 different datasets, but only the results obtained with the first one are partially discussed.

The paper has some interest because the topic tracking problem is really difficult and it is a hot research topic, but it also carries some important problems.

Among the said problems:

The topics must be known, and the approach is limited to finding relationships between core topic words and some related words, which strongly limits the topic discovery of the method in a time-slice based approach. Topics could appear or disappear without any dependencies on the existing ones, especially in research.

The use of word frequencies in the process seriously limits the capabilities of the method to deal with collections of increasing size (results will depend on the size of the time-slices), and the authors are

doi.org/10.32388/GCCRPH

trying to adapt the chosen distance to that problem without completely convincing of their approach. Some equations, like equation (5), are proposed without any proof.

Obtained results would need the validation of experts and quantitative validation measures as well. Moreover, only the results obtained in the first dataset are superficially discussed.

The bibliography is outdated (see, for example, ref. down, and more recent ones are also available).

Some topic modeling methods, like NRM or CFMf (see reference down or more recent ones on CFMf that use clustering jointly with a powerful cluster labeling process), exploit IDF weighting that decreases the dependence of the distance on the collection size. I advise the authors to look at the work described in the references down that do not use information theoretic distances and produce reliable topic tracking results without any latent hypothesis (like the Dirichlet distribution) and/or hyperparameters.

References :

1) Jean-Charles Lamirel

A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research

Scientometrics, Volume 93, Issue 1, pp 151-166.

2) Jean-Charles Lamirel, Francis Lareau, Christophe Malaterre

CFMf topic-model: comparison with LDA and Top2Vec

Scientometrics, Volume 129, Issue 10, pp 6387 – 6405.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/GCCRPH