DOI: doi.org/10.47363/JAICC/ICADCCS2026/2026(5)1

International Conference on AI, Data Science, Cybersecurity, Cloud Architectures, and Software Engineering

Conference Proceedings April 22, 2026 - Germany

Using Domain Taxonomy to Analyze Tendencies of its Development

Boris Mirkin[1,2]

[1]Professor Emeritus, Computer Science, Birkbeck, University of London, UK

[2]Professor, Computer Science, Higher School of Economics, Moscow, Russian Federation

Abstract

There are several approaches to the computerized analysis of tendencies in research within a domain. One should mention most popular approaches such as intercitation analysis, recognition and analysis of most influential studies, and content analysis. This study follows a different approach involving domain taxonomy (hierarchical classification) as the main source of categories and subjects for further analysis. We illustrate working of our approach over the domain of Data Science. We take relevant parts of the ACM Computing Classification System, in a slightly modified and updated form, to serve as the domain taxonomy, so that its obtained 381 leaf concepts are categories under consideration. We downloaded abstracts of 67,757 research papers published in 59 relevant Springer journals from 2003-2024 as the text sample for our analysis. We developed and validated three mathematical methods forming a backbone of our approach.

These are:
(i) Method for assessment of extent of relevance between categories and texts involves the currently popular idea of using the conditional probability of the last symbol in a string with respect to its previous substring
(ii) Method for computing square topic-to-topic co-relevance matrix with an adequate measure involving the so-called Kulczynski index and
(iii) A special additive-spectral fuzzy clustering method, FADDIS.

These are implemented in our GOT software. We divided the period 2003-2024 into two subperiods and computed a 381,67757 topic-to-text relevance matrix for each of the subperiods. We found seven fuzzy clusters, pertaining to main areas of current interest in data science, at each of the two subperiods. Our clusters show both stability in the research interests across the 22-year period and changes within them. These results demonstrate effectiveness of our approach. The main issue in extending that to other domains is lack of taxonomies.

Joint work with Luis Caneco, Susana Nascimento (both from the New Lisbon University, Portugal) and Yana Gagarina, Dmitry Frolov (both from NRU HSE, RF).

Support from the Basic Research Program of HSE University in Moscow including computational resources of HPC facilities in the university is gratefully acknowledged.