CEU CENTRAL EUROPEAN UNIVERSITY

AUTOMATIC IDENTIFICATION OF RELEVANT CONCEPTS IN SCIENTIFIC PUBLICATIONS

A TALK BY

ALESSIO CARDILLO

Ecole Polytechnique Federale de Lausanne (EPFL),
Lausanne, Switzerland

THURSDAY, APRIL 27, **2:00 P.M.** | NÁDOR U. 15., **ROOM 106**

ABSTRACT | Recently, scientists have devoted many efforts to study the organization and evolution of science by leveraging the textual information contained in scientific documents such as: keywords and terms extracted from title/abstract. However, only few studies focus on the analysis of the core of a document, i.e., its body. The access to the whole text of documents allows to study, instead, the organization of scientific knowledge using networks of similarity between articles based on their whole content.

In this talk, I will use the concepts extracted from the documents/articles available within the ScienceWISE platform to build the network of similarity between them. However, such network possesses a remarkably high link density (36%). As a consequence, attempts of associating groups of documents (communities) to a given topic are of limited success. The reason is that not all the concepts are equally informative and thus equally useful to discriminate the articles. The presence of “generic concepts” gives rise to spurious similarities responsible for a large amount of connections in the system.

To get rid of such concepts, I will introduce a method to gauge their relevance according to an information-theoretic approach. The significance of a concept c is encoded by the distance between its maximum entropy, $S_{\max}$, and the observed one, S_c. After removing concepts having an entropy within a certain distance from the maximum, I rebuild the similarity network and analyze its community structure (topics). The consequences of pruning concepts are twofold: the number of links decreases, as well as the noise present in the strength of similarities between articles. Hence, the filtered network displays a more well-defined community structure, where each community contains articles related to a specific topic. Finally, the method can be applied to any kind of documents, and works also in a coarse-grained mode since it is able to identify the relevant concepts for a certain collection of documents, allowing the study of a documents corpus at different scales.

BIO | Alessio Cardillo is currently postdoc research fellow at the Ecole Polytechnique Federale de Lausanne (EPFL) in Switzerland. After obtaining his MSc in Physics at University of Catania (Italy), he moved to Zaragoza (Spain) for his PhD. His research interests focus on the analysis of the structure of networked systems, like: urban mobility and street patterns, scientific collaborations, collections of documents and multiplex networks. He is also interested in the emergence of collective behaviours like cooperation or synchronization by means of coevolutionary dynamics.

ORGANIZED BY THE CENTER FOR NETWORK SCIENCE