Mapping emergent knowledge in Wikipedia through complex networks

Keywords: complex networks; cultural analytics; historical networks; network analysis; knowledge discovery in databases.

Extended Abstract

Just as complex systems exhibit emergent behavior based on how their elements are interconnected, the way in which pieces of information are related in databases can give rise to emergent knowledge that is hardly imaginable from the individual components. In this context, the framework of complex networks can provide the appropriate mathematical tools to unveil and make meaningful the hidden knowledge buried in databases. In particular, we focus on the Wikipedia corpus and demonstrate how the network of internal links between articles yields new, previously undiscovered knowledge.

Wikipedia is one of the most outstanding collective creations in which millions of anonymous editors work independently to build the most significant source of knowledge that humanity has ever seen. Interestingly, in addition to the explicit knowledge presented in its articles, a vast amount of implicit learning emerges from the dense network of internal links connecting people, ideas, and works. This extensive network (~160M connections in the English version) can be represented as a directed graph and has been used in several studies, ranging from computing semantic relatedness [1] to natural language processing [2].

In this work, we propose a mathematical approach to transform the directed network of internal links in Wikipedia into a weighted, undirected network that reveals and quantifies subtle connections among individuals, places, oeuvres, and ideas. The proposed methodology is based on computing the structural relationship between each pair of nodes using the normalized Google distance [3]. Thus, we obtain a relatedness matrix of N x N ($N \sim 7$ M), where each element represents the strength of the structural relationship between the corresponding Wikipedia articles. Once we have this matrix, we can map and analyze how specific elements interact with one another. This methodology allows us to approach several interdisciplinary problems that were previously out of reach. For example, we can quantitatively study the interactions between people, places, or even ideas; we can map and analyze the interactions among cultural entities; and we can also explore these interactions for a given period (such as the Renaissance [4] or Enlightenment [5]) or follow their evolution over several centuries.

As a proof of concept for the proposed methodology, we will examine in this talk two cases: a) on the one hand, we will study the networks of relevant writers over the last three centuries. We will see how unknown patterns emerge and allow us to predict future behaviours; b) the second case will focus on the interdisciplinary cultural network that connected art, science, and philosophy in Europe during the seventeenth century (see Figure 1). The results align with well-established historical knowledge of the period and, more importantly, provide new insights into how elements in these networks interact with each other. In particular, we found that nodes within a given cluster, related respectively to art, science, or philosophy, interact with nodes in the same cluster following a core-periphery behaviour. In contrast, inter-cluster interactions across disciplines follow a power law distribution.

References

[1] "Computing semantic relatedness using Wikipedia-based explicit semantic analysis". In: Int Joint Conf Artificial Intell 7 (2007) 1606–1611.

[2] "Wikipedia-based semantic interpretation for natural language processing". In: J Artificial Intell Res 34 (2009) 443–498.

[3] "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links". In: Association for the Advancement of Artificial (2008) 25-30.

[4] "Mapping the networked context of Copernicus, Michelangelo, and della Mirandola in Wikipedia". In: Advances in Complex Systems, 25 (2022) 2240010.

[5] "Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks". In: Humanities and Social Sciences Communications 12 (2025) 462.

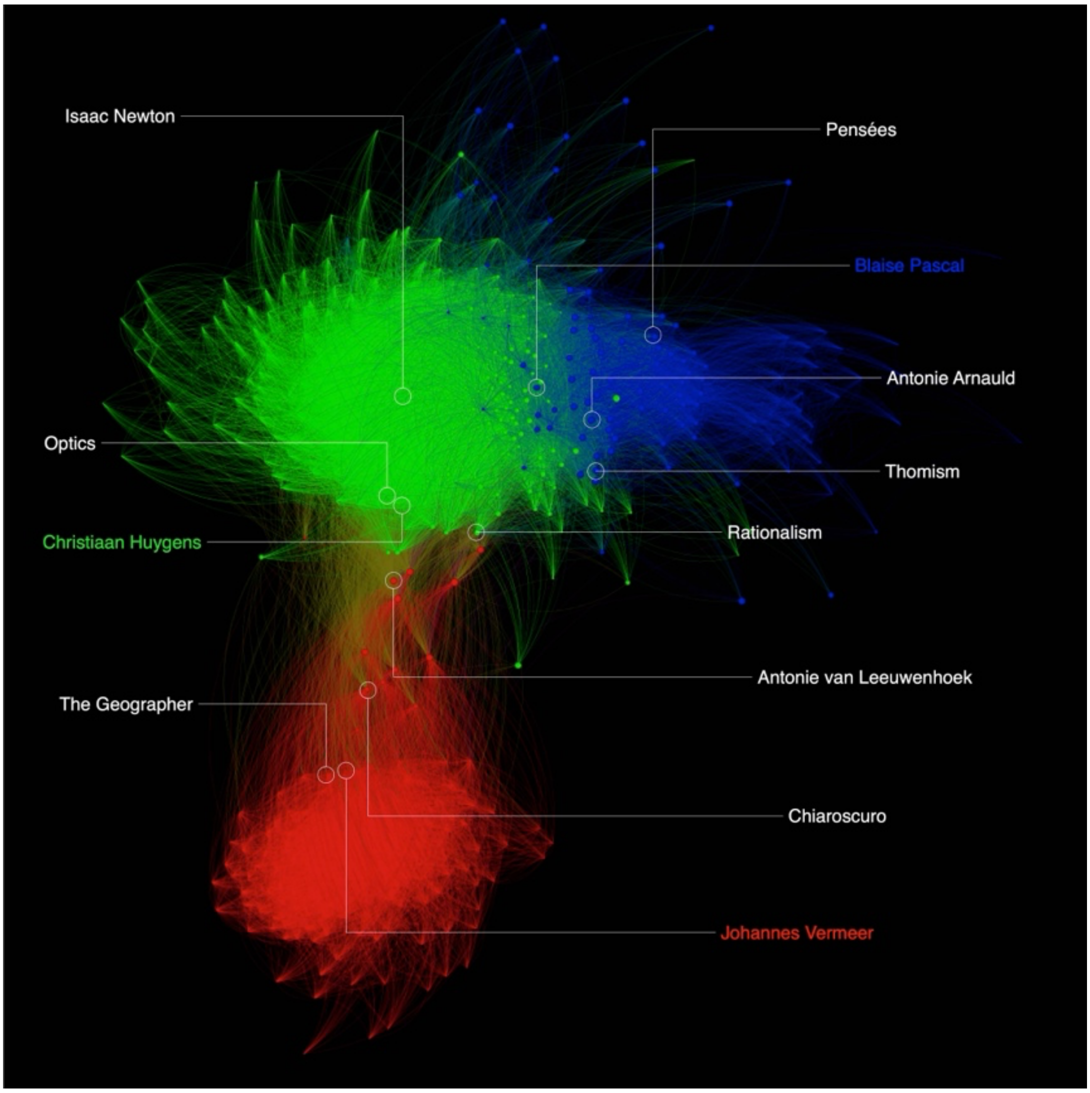

Figure 1. Cultural network based on Johannes Vermeer, Christiaan Huygens, and Blaise Pascal. Each point represents a Wikipedia article whereas each line represents the relationship between them calculated using the normalized Google distance. Note the strong interaction between science (green) and philosophy (blue) and the relative isolation of art (red).