Physica A 441 (2016) 81–84

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Density: A measure of the diversity of concepts addressed in semantic networks

H.B.B. Pereira [a,b], I.S. Fadigas [c], R.L.S. Monteiro [a,b], A.J.A. Cordeiro [a,d], M.A. Moret [a,b,*]

[a] Programa de Modelagem Computacional, SENAI Cimatec, Av. Orlando Gomes 1845, 41.650-010, Salvador, BA, Brazil

[b] Universidade do Estado da Bahia, Rua Silveira Martins, 2555, 41.150-000, Salvador, BA, Brazil

[c] Departamento de Ciências Exatas, Universidade Estadual de Feira de Santana, Campus Universitário, Módulo 5, 44031-460, Feira de Santana, BA, Brazil

[d] Centro Universitário Estácio da Bahia, Campus Gilberto Gil, 41770-130, Salvador, BA, Brazil

HIGHLIGHTS

- We propose a novel methodology to grow up cliques networks.
- We propose density to measure the diversity of concepts addressed in semantic networks.
- We measure the Euclidean distance between different scientific journals.
- The density of clique networks may be of interest for the definition of new bibliometric parameters.

ARTICLE INFO

Article history:
Received 6 February 2015
Received in revised form 15 July 2015
Available online 24 August 2015

Keywords:
Complex networks
Self-organized criticality
Euclidean distances

ABSTRACT

In this paper, we studied density effects in semantic networks constructed from a database of titles of papers published in scientific journals as a parameter to indicate the diversity of concepts in a journal. The proposed method essentially consists of fixing the number of titles for all of the studied scientific journals and analyzing the behavior of the density variation curves with regard to the inclusion of cliques (that is, complete networks associated with the titles). We observed that density behaves as a critically self-organized object when titles (cliques) are included in the network.

In recent decades, there has been intrinsic evidence that many physical, economic, and biological phenomena as well as other complex systems show a critical phenomenon characterized by a behavior that follows a power law. In this context, self-organized criticality (SOC) [1,2] was proposed to analyze such complex systems. One advantage of SOC is that it has an enormous intuitive appeal, which explains why it has been widely discussed, especially regarding natural phenomena and complex systems. The SOC methodology has been applied to analyze avalanches [1], hydrophobicity [3], the evolution of species [4,5], cellular behaviors [6–8], proteins [9–14], and epidemics [15–23].

When considering complex networks, certain properties allow us to explain or predict the behaviors of this type of complex system. Regarding that there is no detailed specification of the initial conditions, semantic networks based on titles of scientific papers present an evolution phenomenon of a spatial structure with scale-invariant and self-similarity

* Corresponding author at: Universidade do Estado da Bahia, Rua Silveira Martins, 2555, 41.150-000, Salvador, BA, Brazil.
E-mail addresses: hbbpereira@gmail.com (H.B.B. Pereira), isfadigas@gmail.com (I.S. Fadigas), roberto@souzamonteiro.com (R.L.S. Monteiro), antonio.cordeiro@gmail.com (A.J.A. Cordeiro), mamoret@gmail.com (M.A. Moret).

http://dx.doi.org/10.1016/j.physa.2015.08.024
0378-4371/

properties, as proposed in Ref. [2]. Besides, new titles are randomly selected to build the semantic networks. In this paper, we apply SOC to analyze density effects in complex semantic networks constructed based on titles from journals.

Density is a physical property of matter that is defined as the ratio between the mass of an object and its volume. In this study, the object in question is a structure composed of vertices and edges, known as a network. Specifically, we studied the density of semantic networks based on titles from journals as a parameter to indicate the diversity of concepts addressed by a scientific journal.

Semantic networks are networks whose vertices are words and whose edges consist of connections between the words that appear in the same unit of meaning, that is, in a sentence, a paragraph, or a title of the analyzed discourse. They are, therefore, systems for the representation of knowledge. Recently, certain studies about complex semantic networks based on linear discourses were proposed [24–26].

The proposed semantic networks are represented as graphs $G = (V, E)$, which are mathematical structures and consist of two sets: V (finite and not empty) and E (binary relation in V). The elements of V are called vertices, and the elements of E are called edges [27]. In the analyzed semantic networks, there are no multiple edges (more than one edge between the same pair of vertices) or loops (each edge has two vertices associated with it).

In this study, we used the following properties to characterize the semantic networks based on titles from journals: n is the total number of vertices or the cardinality of the set V (that is, $n = |V|$); $m$is the number of edges or the cardinality of the set E (that is, $m = |E|$). For each semantic network proposed, the maximum possible number of connections between the words (that is, vertices selected after pre-treatment of the titles of the papers) is determined by $n(n-1)$. The density Δ is the ratio between the number of connections present in the studied semantic network and the maximum number of possible connections:

$$\Delta = \frac{2m}{n(n-1)}. \tag{1}$$

We used scientific journals from different areas (Interdisciplinary, Agrarian Sciences, Biology, Chemistry, Engineering, Geography, Health Sciences, Human Sciences, Linguistics, Physics, and Statistics) with papers published in English. The data set is the same as used for Ref. [25]: AFE—Agricultural and Forest Entomology; ARJG—Antipode: A Radical Journal of Geography; APPL—Applied Psycholinguistics; CB—Chemistry and Biology; HR—Human Relations; Nature; PRA—Physical Review A; PRB—Physical Review B; PRC—Physical Review C; PRD—Physical Review D; PRE—Physical Review E; PRL—Physical Review Letter; PEM—Probabilistic Engineering Mechanics; Science; and SHI—Sociology of Health and Illness.

To construct a semantic network based on the titles of papers published in scientific journals, we used the method proposed by Refs. [26,28]. Essentially, after preserving only words with intrinsic meaning and eliminating grammatical words (for example, articles, pronouns, prepositions, connectors, abbreviations, and interjections), we applied the general rules for pre-processing the titles of papers proposed by Ref. [25]. Each title is then represented by a network where all of the vertices (that is, words) are interconnected, generating a clique (one clique is a maximal subset of vertices in one graph G that are mutually adjacent to each other [27]). Words that appear in more than one title are connection vertices between the titles.

This type of semantic network is called a "clique network" and has different properties from other types of networks because of its construction process, which is based on juxtaposition and/or overlapping (that is, the process of juxtaposition means connecting two cliques by only one common vertex, and the overlapping process means connecting two cliques by two or more shared vertices, as suggested by Ref. [29]). We recall that the building of semantic networks based on titles of scientific papers is based on the addition of titles (or more specifically cliques).

In previous studies (e.g. Ref. [25]), we considered all of the titles from each journal ($370 \leq$ quantity of titles $\leq 35{,}163$). Thus, our analysis considered 300 titles from each journal.

In addition to what was mentioned previously, we also used the following: k_i is the degree of vertex i, that is, the number of edges incident on vertex i; $P(k)$ is the probability distribution of the number of connections of all of the vertices over the entire network (that is, the distribution of degrees); $\langle k \rangle$ is the average of k_i defined by $\langle k \rangle = \frac{1}{n}\sum_i^n k_i$; C represents the average clustering coefficient of the network vertices, $C = \frac{1}{n}\sum_i^n C_i$, where the clustering coefficient of one vertex i, defined as C_i, measures the proportion of edges existing between neighbors of the vertex i, m_i and the maximum possible number of edges in this neighborhood: $C_i = \frac{2m_i}{k_i(k_i-1)}$ [30]; and L is the minimum average path length or the average geodesic distance between all of the vertices of a connected network: $L = \frac{1}{n(n-1)}\sum_{i\neq j} d_{ij}$, where d_{ij} is the geodesic distance, in terms of the number of edges, between the vertices i and j.

Fig. 1 shows the variation of the density when new titles are introduced individually into the network, up to 300 titles. It is worth emphasizing that the data shown are from the journal *Nature*. The observed shape of the curve suggests to us a behavior similar to the one observed in objects that show self-organized criticality. The analysis of the evolution of the density of this type of semantic network (that is, clique networks) indicates that when there are a small number of titles, an elevated density is obtained; however, after adding a large number of titles to the network, a small density is obtained. That said, when there is no connection or the connection occurs taking into account few vertices, the network density (Δ) greatly decreases. This suggests that new words (new vertices) were added and the semantic networks self-organize. The relationship between density and the quantity of titles (cliques) obeys a power law with Pearson's correlation coefficient $R = -0.998$, $F_{value} = 72920.403$ and $Prob > F \rightarrow 0$. The power law in the network density analogously behaves as a