Physica A 390 (2011) 1192–1197

Contents lists available at ScienceDirect

Physica A

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

Semantic networks based on titles of scientific papers

H.B.B. Pereira [a,b,*], I.S. Fadigas [b], V. Senna [a], M.A. Moret [a,c]

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

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

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

ARTICLE INFO

Article history:
Received 28 April 2010
Received in revised form 27 October 2010
Available online 14 December 2010

Keywords:
Semantic networks
Complex networks
Social network analysis

ABSTRACT

In this paper we study the topological structure of semantic networks based on titles of papers published in scientific journals. It discusses its properties and presents some reflections on how the use of social and complex network models can contribute to the diffusion of knowledge. The proposed method presented here is applied to scientific journals where the titles of papers are in English or in Portuguese. We show that the topology of studied semantic networks are small-world and scale-free.

1. Introduction

Social network analysis and complex network theory have been used to research the behavior and structure of several complex systems through networks, e.g. technological networks [1–3], biological networks [4,5], social networks [6–8], organization networks [9], information networks [10–12], and semantic networks [13–15], among others.

Semantic networks are generally contextualized. Several works have been developed to investigate the connection of words from a semantic association perspective and/or the frequency of pairs of words [14–17].

In this paper we propose a discussion on the use of a semantic network based on titles of papers published in scientific journals as a method to analyze the efficiency of diffusion of information. To achieve this, we analyzed scientific journals of different fields (i.e. Interdisciplinary, Agricultural Sciences, Biology, Chemistry, Computer in Education, Engineering, Geography, Health Sciences, Human Sciences, Linguistics, Mathematics Education, Physics and Statistics) and in two languages (English and Portuguese).

We highlight that the analysis of semantic networks of titles of scientific papers is an attempt to understand the association between scientific texts and their titles. This allows us to check the dependence of the titles with respect to (1) the jargon and technical terms of the study field or journal, (2) the research activities taking place in a given period and (3) the remarkable scientists and their models.

This paper is organized as follows. We present in Section 2 the data and the method used to build the proposed semantic networks. In Section 3 we build on the previous one and examine the results obtained through some indices of social and complex networks. Finally, we present the conclusions of this paper in Section 4.

2. Building semantic networks of titles

The semantic networks based on titles of scientific papers we propose are networks which the vertices are words and the edges are connections between words that appear in the same title. To represent a network, we use a graph $G = (V, E)$ that

* Corresponding author at: Programa de Modelagem Computacional, SENAI Cimatec, Av. Orlando Gomes 1845, 41.650-010, Salvador, BA, Brazil.
E-mail address: hbbpereira@gmail.com (H.B.B. Pereira).

0378-4371
doi:10.1016/j.physa.2010.12.001

Table 1
General rules for manual pre-processing of titles of papers.

Rules	Description
R1	Each title consists of one sentence.
R2	Graphic signs, such as period, semicolon, question mark, exclamation point and ellipses are eliminated.
R3	Names should form a single word. For instance, "Bose–Einstein" should be converted to "boseeinstein", or "Albert Einstein" should be converted to "alberteinstein".
R4	Ordinal numbers should be written as follows: "first", "second", etc.
R5	Numbers should be written textually. For instance, "onezero" in place of "10".
R6	Composite words should be considered as only one word. For instance, "Rio de Janeiro" should be converted to "riodejaneiro", or "e-mail" should be converted to "email".
R7	Words incorrectly spelt, should be corrected.
R8	Specialized language should be kept as much as possible.
R9	Words repeated in the same title should be excluded, leaving only one occurrence of the word.
R10	Word strings that are jointly meaningful, are made into a single word (e.g. blackhole, computerscience.).
R11	Titles in another language should be translated into the language of analysis (e.g. an article published in a journal whose main language is Portuguese, with a title in another language should be translated into Portuguese).

is a mathematical structure and consists of two sets: V (finite and not empty) and E (binary relation on V). The elements of V are called vertices and the elements of E are called edges [18]. In our semantic networks, each edge has two vertices associated to it.

As mentioned previously, the data set used is composed of scientific journals published in English (Agricultural and Forest Entomology—**AFE**; Antipode: A Radical Journal of Geography—**ARJG**; Applied Psycholinguistics: Psychological and Linguistic studies Across Languages and Learning—**APPL**; Chemistry and Biology—**CB**; Human Relations: Towards the integration of the Social Sciences—**HR**; **Nature**; Physical Review A—**PRA**; Physical Review B—**PRB**; Physical Review C—**PRC**; Physical Review D—**PRD**; Physical Review E—**PRE**; Physical Review Letter—**PRL**; Probabilistic Engineering Mechanics—**PEM**; **Science**; Sociology of Health and Illness—**SHI**) and Portuguese (Boletim de Educação Matemática—**BOLEMA**; Boletim GEPEM—**GEPEM**; Educação Matemática em Revista—**EMR**; Folhetim de Educação Matemática—**Folhetim**; Revista Brasileira de Informática na Educação—**RBIE**; Revista do Professor de Matemática—**RPM**; **Zetetiké**).

The criteria used to select the scientific journals published in English are an impact factor greater than one; the journals should be available on the Internet; and each journal should represent as well as possible one area of knowledge, including interdisciplinary fields. For the scientific journals published in Portuguese, we chose to concentrate on journals that dealt with Mathematics Education.

The method for constructing the proposed semantic networks basically consists of (1) elimination of words without intrinsic meaning and (2) changing the remaining words to their canonical form, as suggested in Refs. [14,15]. Each title is a network where all vertices (i.e. words) are interconnected, generating cliques (a clique is a subset of vertices in a graph G that are mutually adjacent to one another [18]). Words that appear in more than one title are vertices of connection between the titles. In this way, we construct a semantic network based on papers' titles published in a given journal.

In order to build up the semantic network based on titles, a pre-processing is done that consists of applying general rules defined to minimize possible inconsistencies and to standardize the analysis for the different journals. These rules are shown in Table 1.

After pre-processing, the words go through a set of UNITEX programs [19], to address issues such as ambiguities, the deletion of grammatical words (e.g. articles, personal and possessive pronouns, possessive adjectives, statements, questions, adverbs etc.) and separation of the canonical or inflected forms of words from the rest of the items of grammatical classification generated by the UNITEX programs.

Fig. 1 depicts the resulting network of two titles – T01: "Specialized hepatocyte-like cells regulate Drosophila lipid metabolism" [20] and T02: "Identification and expansion of human colon-cancer-initiating cells" [21] – after applying the method described above. The word "cell" is the point of connection between titles T01 and T02. When the whole network is finally built, these kinds of vertices are likely to become central points.

3. Results and discussion

For the proposed analysis, a set of indices from social network analysis and complex network theory were used to quantify and interpret the network properties. Authors such as [6–8,22–26] present a detailed discussion on several indices to study properties of social and complex networks.

Within this context, we have chosen some indices from complex network theory to characterize topologically the proposed semantic networks. Additionally, we have studied some aspects related to (multi/inter)disciplinarity of the scientific journals as glinted from their titles' semantic networks.

Although studies on social and complex networks are now mature, still there is a lack of standardization in the use and formalization of some concepts related to networks. Therefore, we give a short glossary of terms used in this article.

Number of vertices (N) Total number of vertices or cardinality of set V, i.e. $N = |V|$.
Number of edges (M) Total number of edges or cardinality of set E, i.e. $M = |E|$.

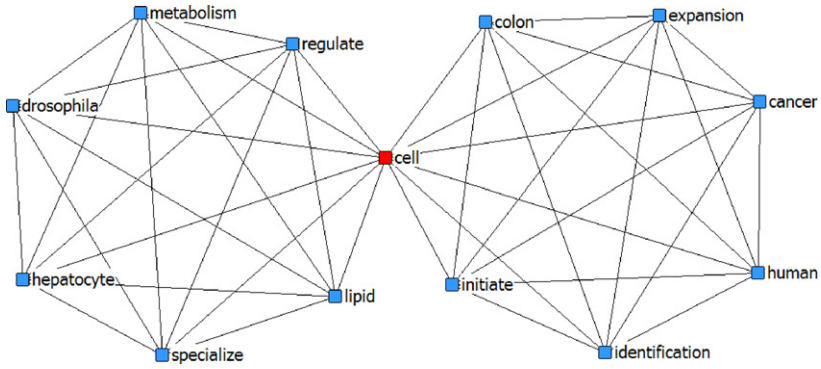

Fig. 1. Excerpt of a semantic network based on two titles of papers published in Nature journal, Volume 445, Issues 7123 and 7125.

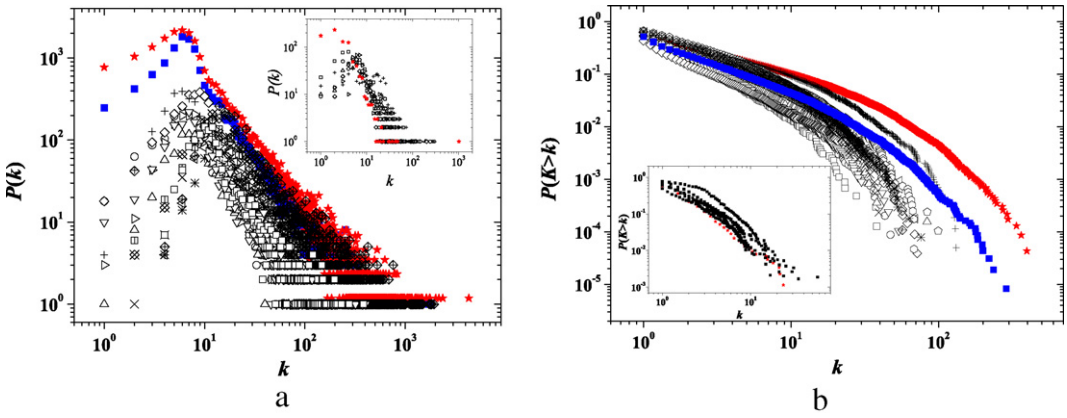

Fig. 2. Degree distributions of the semantic networks based on titles of paper belonging to the journals studied. (a) Degree distributions of scientific journals published in English. The upper inset represents the degree distributions of journals published in Portuguese; (b) Cumulative degree distributions of scientific journals published in English. The lower inset represents the cumulative degree distributions of journals published in Portuguese.

Degree of a vertex (k_i) The degree of a vertex i, denoted by k_i, is the number of edges incident on the vertex i.

Degree distribution ($P(k)$) The degree distribution is the probability distribution of number of connections of all vertices over the whole network.

Average degree ($\langle k \rangle$) The average degree of an undirected network is the average of k_i denoted by $\langle k \rangle = \frac{1}{N}\sum_i^N k_i$.

Average clustering coefficient (C) Mean clustering coefficient of the vertices of the network, $C = \frac{1}{N}\sum_i^N C_i$. The clustering coefficient of a vertex i, denoted by C_i, measures the proportion of existing edges between neighbors of vertex i, E_i and the maximum possible number of edges: $C_i = \frac{2E_i}{k_i(k_i-1)}$; this index shows the extent to which an individual's friends are friends with each other [1].

Average minimal path length (L) Average geodesic distance, considering the shortest path between two 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 number of edges, between the vertices i and j.

Density (Δ) Density of an undirected network, denoted by $\Delta = \frac{M}{N(N-1)/2}$, is the total of existing edges (M), divided by the maximum possible number of edges ($N(N-1)/2$).

From the network topology perspective, all semantic networks based on titles of papers have small-world properties, as they all have a large value of C ($0.68 \leq C \leq 0.80$ for semantic networks from scientific journals in English, and $0.60 \leq C \leq 0.83$ for semantic networks from scientific journals in Portuguese) and small values of L (semantic networks from scientific journals in English present $2.46 \leq L \leq 2.98$ and semantic networks from scientific journals in Portuguese present $2.46 \leq L \leq 3.65$), when compared to a random network with same values of N and $\langle k \rangle$ (typically $0.0046 \leq C \leq 0.024$ for similar values of L).

Degree distributions of all semantic networks based on paper's titles used in this research, as can be observed in Fig. 2, have a similar behavior, i.e., follow power laws according to the probability $P(k) \sim k^{-\gamma}$ (values of γ are shown in Fig. 3). These results seem to vindicate a tendency of the power laws in this kind of semantic networks. Similar results from analysis

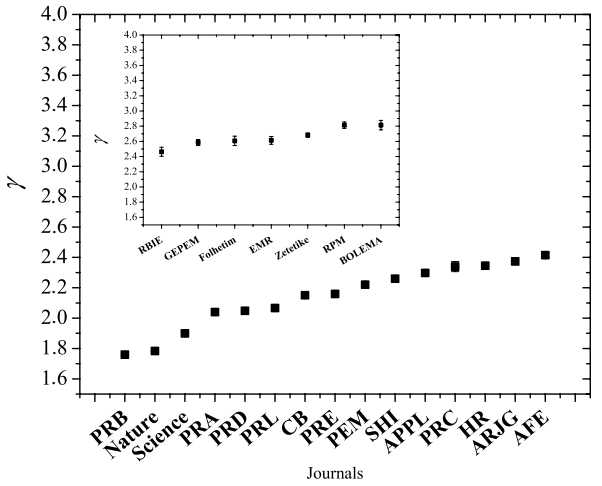

Fig. 3. Values of γ for the degree distributions ($P(k) \sim k^{-\gamma}$) of the scientific journals published in English. The upper inset shows the values of γ for the degree distributions of the scientific journals published in Portuguese.

of texts can be found in [14,15]. In Fig. 2(a), there is a scattering of data (k) in the tail, corresponding to large vertices (i.e. hubs), for which $P(k)$ is low (i.e. words that have a very high frequency of occurrence). This fanning out effect is due to the existence of vertices with high degree that seldom occur. As it is necessary to define a method for choosing the appropriate point to perform the linear fit, we used the cumulative distribution function (Fig. 2(b)) to reduce the noise in the tail.

It is important to note that due to the peculiar character of the semantic networks based on titles of papers, instead of adding vertices to those already existing, cliques are added. Thus, the topology and properties of the network as a whole reflect how the cliques (titles of papers) are associated themselves (not the individual words). We highlight that the connection of cliques depends on the occurrence of at least a vertex (a common word) in two components (i.e. an existent clique and the new one). These networks could be referred to as "networks of cliques".

The growth model that describes these networks, generated through the continued addition of a random number of new vertices imply, for instance, that the degree distribution $P(k)$ can be obtained from the frequency distribution of words and the size distribution of titles. That is, random values of the degree distribution can be approximated as $\sum_{i=1}^{w}[S(t) - 1]$, where w is a random variate from the frequency distribution of words, usually Zipf's law, and $S(t)$ is a variate from the size distribution of titles typically a Poisson Distribution. We believe that such networks have some characteristics that differentiate them from the theoretical models proposed by [1,2,27].

From a dynamics perspective, semantic networks based on paper's titles can be explained through a model of growth similar to film actors [1,28] or co-authorship networks [29].

The study of the topology of semantic networks gave us clues regarding disciplinarity of the journals. Each journal has its own jargon and specific technical terms. Many properties of the studied networks seem to be related to the diversity of vocabulary of the journals. In this paper, we used the density (Δ) and the average minimal path length (L) as starting points to examine some of the journals characteristics.

The density of a semantic network based on titles of papers reflects the quantity of links between words. This indicates a tendency of the titles to link through a large number of words. Fig. 4 depicts the density values of the semantic networks examined.

The small-world phenomenon observed in these semantic networks means that the words are very close in geodesic terms. That is, small L values imply very short paths between titles or words belonging to a title of the scientific journals. Fig. 5 suggests that multidisciplinary journals have high L values in comparison with disciplinary journals.

4. Concluding remarks

We highlight that word network analysis is an attempt of understanding the associations between a scientific text and its title. Thus the main focus is an human language when it is used to make a connection between the central idea of a scientific article, generally expressed in its title, and the complete description of the proposal.

The main purpose of the analysis of the scientific journals as presented here, is to show that social network analysis and complex network theory are convenient tools for studying semantic networks. These tools can contribute to the diffusion of knowledge through a better understanding of the breadth of a journal's coverage and its (multi)disciplinary degree. Within this context, it is important to highlight that our word network analysis is an attempt to understand the association between the contents of a scientific journal and the title of articles it publishes.

In summary, the proposed semantic networks mirror conceptual schemas associated to editorial lines of scientific journals. Semantic networks based on titles of papers can guide authors/researchers in deciding where to submit their results.

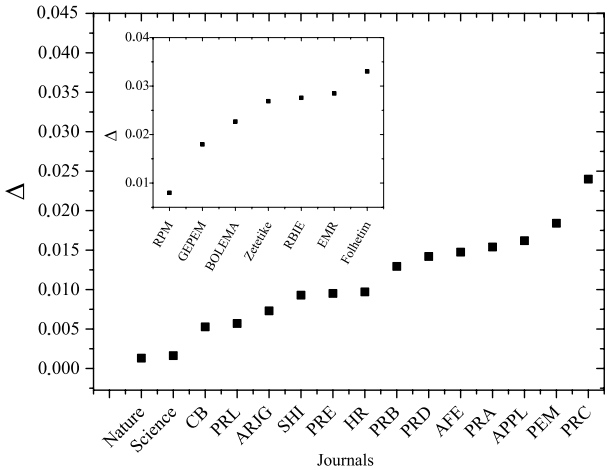

Fig. 4. Density of semantic networks from the scientific journals published in English. The upper inset represents the density of semantic networks from the scientific journals published in Portuguese.

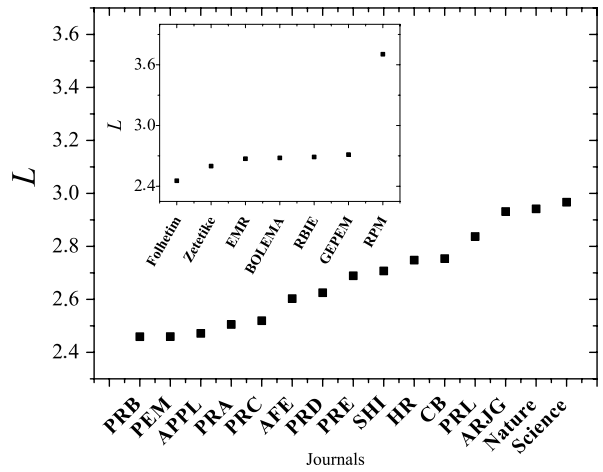

Fig. 5. Values of L for the semantic networks from the scientific journals published in English. The upper inset represents L for the semantic networks from the scientific journals published in Portuguese.

Finally, the results obtained so far suggest a similarity in the behavior of the examined semantic networks, regardless of the official language of the selected scientific journal (here Portuguese and English). The results also point out that the density Δ (Fig. 4) and the average minimal path length L (Fig. 5) are useful measures of information associated to the complexity of the semantic networks.

Acknowledgement

This work received financial support from CNPq (Brazilian federal grant agency).

References

[1] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 6684.
[2] A.-L. Barabási, R. Albert, Science 286 (1999) 5439.
[3] M. Faloutsos, P. Faloutsos, C. Faloutsos, Comput. Commun. Rev. 29 (1999) 251.
[4] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.-L. Barabási, Nature 407 (2000) 651.
[5] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, Y. Åberg, Nature 411 (2001) 907.
[6] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences), Cambridge University Press, Massachusetts, 1994.
[7] J.P. Scott, Social Network Analysis: A Handbook, SAGE Publications, London, 2000.
[8] P.J. Carrington, J. Scott, S. Wasserman, Models and Methods in Social Network Analysis, Cambridge University Press, New York, 2005.
[9] H.B.B. Pereira, M.C. Freitas, R.R. Sampaio, DataGramaZero 8 (2007) 1.
[10] D.J. Price, Science 149 (1965) 510.
[11] M.E.J. Newman, Phys. Rev. E 64 (2001) 016131.

[12] M.E.J. Newman, Phys. Rev. E 64 (2001) 016132.
[13] R.F.I. Cancho, R.V. Solé, Proc. R. Soc. Lond. B Biol. Sci. 268 (2001) 2261.
[14] S.M.G. Caldeira, T.C.P. Lobão, R.F.S. Andrade, A. Neme, J.G.V. Miranda, Eur. Phys. J. B 49 (2006) 523.
[15] G.M. Teixeira, M.S.F. Aguiar, C.F. Carvalho, D.R. Dantas, M.V. Cunha, J.H.M. Morais, H.B.B. Pereira, J.G.V. Miranda, Int. J. Mod. Phys. C 21 (2010) 333.
[16] A. Caramazza, Nature 380 (1996) 485.
[17] D.L. Nelson, C.L. Mcevoy, T.A. Schreiber, The University of South Florida word association, rhyme, and word fragment norms, 1998. http://www.usf.edu/FreeAssociation/.
[18] J. Gross, J. Yellen, Graph Theory and its Applications, CRC Press, Boca Raton, 1999.
[19] S. Paumier, UNITEX 2.0 User Manual, 2008. Eletronic version, www-img.univ-mlv.fr/~unitex/UnitexManual2.0.pdf.
[20] E. Gutierrez, D. Wiggins, B. Fielding, A.P. Gould, Nature 445 (2007) 275.
[21] L. Ricci-Vitiani, D.G. Lombardi, E. Pilozzi, M. Biffoni, M. Todaro, C. Peschle, R. De Maria, Nature 445 (2007) 111.
[22] R. Albert, A.L. Barabasi, Rev. Modern Phys. 74 (2002) 47.
[23] M.E.J. Newman, SIAM Rev. 45 (2003) 167.
[24] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Phys. Rep. 424 (2006) 175.
[25] L.F. Costa, F.A. Rodrigues, G. Travieso, P.R. Villas Boas, Adv. Phys. 56 (2007) 167.
[26] S.H. Strogatz, Nature 410 (2001) 273.
[27] P. Erdös, A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17.
[28] J.J. Collins, C.C. Chow, Nature 393 (1998) 409.
[29] M.E.J. Newman, Proc. Natl. Acad. Sci. 101 (2004) 5200.