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/$ – see front matter
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. $M = |E|$.