Expert Systems With Applications 45 (2016) 8–22

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Analyzing natural human language from the point of view of dynamic of a complex network

Guilherme Alberto Wachs-Lopes, Paulo Sergio Rodrigues*

Department of Computer Science, Centro Universitário da FEI, 3972, Castelo Branco Av, São Paulo, Brazil

ARTICLE INFO

Keywords:
Complex networks
Physical measures
Clustering coefficient
Textual information retrieval

ABSTRACT

With increasing amount of information, mainly due to the explosive growth of Internet, the demand for applications of automatic text analysis has also grown. One of the tools that has increased in importance in the understanding of problems related to this area are complex networks. This tool merges graph theory and statistical methods for modeling important problems. In several research fields, complex networks are studied from the various points of view, such as: topology of networks, extraction of physical features and statistics, specific applications, comparison of metrics and study of physical phenomena. Linguistic is one area that has received great attention, particularly due to its close relationship with issues arising from the emergence of large text databases. Thus, many studies have emerged for modeling of complex networks in this area, increasing the demand for efficient algorithms for feature extraction, network dynamic observation and comparison of behavior for different types of languages. Some works for specific languages such as English, Chinese, French, Spanish, Russian and Arabic, have discussed the semantic aspects of these languages. On the other hand, as an important feature of a network we can highlight the computation of average clustering coefficient. This measure has a physical impact on the network topology studies and consequently on the conclusions about the semantics of a language. However its computational time is of $O(n^3)$, making its computing prohibitive for large current databases. This paper presents as main contribution a modeling of two complex networks: the first one, in English, is constructed from a specific medical database; the second, in Portuguese, from a journalistic manually annotated database. Our paper then presents the study of the dynamics of these two networks. We show their small-world behavior and the influence of hubs, suggesting that these databases have a high degree of Modularity, indicating specific contexts of words. Also, a method for efficient clustering coefficient computation is presented, and can be applied to large current databases. Other features such as fraction of reciprocal connections and average connection density are also calculated and discussed for both networks.

1. Introduction

The Theory of complex networks is a set of tools that has gained great attention in many areas of research, whose main objective is to study, from a statistical point of view, large scalable problems with behaviors considered complex. Thus, a considerable range of important problems has been modeled with such tools, including technology areas such as engineering and computer science, but also biological, medical, economic, social, among many others. In particular, the textual human language has received great attention from the point of view of this relatively new area for several reasons. The main reason among them is the possibility of modeling the interrelationship of key elements, such as individual words in an attempt to capture lexical, syntactic and semantic information, and the various levels of human communication through the formalism of a language. Perhaps the most recent work that has generated much debate among researchers is the work of Cong and Liu (2014). In his historical paper, the authors defend the modeling of natural language using complex networks. This work has generated controversy, with criticism and praise from many other researchers. In his work, Cong and Liu suggest possible directions of research, including: "1) relations between the system-level complexity of human language and microscopic linguistic features; 2) expansion of research scope from the global properties to other levels of granularity of linguistic networks; 3) combination of linguistic network analysis with other quantitative studies of language (such as quantitative linguistic)". Also, in this study, the authors present a survey of the three major lines of linguistic research from complex networks approach: "1) characterization of human language as a multi-level system with complex networks; 2) linguistic typological research with the application of linguistic

* Corresponding author. Tel.: +55 11 992463692; fax: +55 11 43532910.
E-mail addresses: gwachs@fei.edu.br (G.A. Wachs-Lopes), psergio@fei.edu.br (P.S. Rodrigues).

http://dx.doi.org/10.1016/j.eswa.2015.09.020
0957-4174/

network and their quantitative measure; and 3) relationships between the system-level complexity of human language (determined by the topology of linguistic networks) and microscopic linguistic (e.g. syntactic) features (as the traditional concern of linguistic)".

The Cong and Liu (2014) work received some contrary arguments, such as Ferrer-i Cancho (2014), Cech (2014), Hudson (2014), Macutek (2014), and favorable arguments such as: Amancio (2014), Chen (2014), Hudson (2014), Kohler (2014), Yu and Xu (2014), Zhao (2014) and Cong and Shuai (2014). The main against argument is related to the fact that complex networks are not able to completely model the human language from a cognitive point of view. On the other hand, favorable opinions argue that many aspects of human language were found, proving the usefulness of networks in several problems modeled as that. Regardless arguments for and against, most researchers have agreed that the Cong and Liu (2014) work laid the modeling of natural language under the light of important scientific discussions, generating important ways and challenges.

Specifically, the paper of Ferrer-i Cancho (2014) presented critics with respect to Cong and Liu (2014) work mainly due to the lack of mention of G.K. Zipf, regarding linguistic aspects of language. However, Ferrer-i Cancho (2014) praised the same work considering the point of view of modeling using complex networks. But Cech (2014) argue that the work of Cong and Liu (2014) exaggerates the power of complex networks in modeling human language.

The authors argument that complex networks is not about human cognition. They claim that human brain networks, as they are small-world, not necessarily implies being efficient in terms of communication between their pairs of vertices. They argue that, if this is true, it must be proved empirically and actual results are mere speculation. Under the same line of thought, Macutek (2014) presents the facts of why complex networks still require a lot of development for modeling of several physical problems and phenomena, mainly including linguistics. The authors point out several advances in the language area and new points of view, however, emphasize that complex networks is not the only tool able to help in that direction.

On the other hand, Amancio (2014) presents arguments supporting the (Cong & Liu, 2014) work demonstrating various applications that can be achieved by complex network models in order to study the semantics of words in texts. Such applications are supported by the extraction of several network features, usually common in this type of modeling. Also, Chen (2014) reinforce the cognitive nature of complex networks for the study of linguistics, as they said "making it possible to integrate findings from different linguistic branches in the same framework and then make closer connections between different linguistic branches". Also, Kohler (2014) presented arguments favorable to the idea that the complex networks when modeling natural languages mutually associating words in texts, allow reveal associated phenomena in different ways. In the work of Yu and Xu (2014), the authors suggest the use of complex networks for the study of natural languages, especially in the case of the interrelationship components in different levels of a natural language.

From the point of view of researchers of the linguist area, Hudson (2014) also shows the vision of the use of complex networks for modeling of natural language. Although Hudson (2014) agrees with the current studies, his paper points directions in which there is demand in the linguistic area, such as the semantic interrelationship of words in different linguistic levels, rather than only between word-word. The author argues that "the most important relation is class-membership. However, there are not only a handful of other special relations but also an open-ended, learnable and hierarchically organized set of relational concepts such as 'subject', 'meaning' and 'base'. One of the challenges for network research is to find out how these different relations interact in producing the overall profile for language's grammar".

Thus, to meet the demands of the language area, both from the point of view of applications, and from the point of view of the study of cognitive functions, so much work in the feature extraction of complex network has gaining in importance, such as the work of Pardo, Antiqueira, Nunes, Oliveira, and Costa (2005), which has studied the ability of complex networks of words in the tasks of automatic summarization of texts. This task is accomplished by calculating metrics such as average degree of the nodes, clustering coefficient, and the study of the network topology from the point of view of degree distribution. In Cech, Macutek, and Liu (2014) is presented a review of complex networks for modeling of natural language from the the point of view of syntactic aspects.

Works that focus on applications of this type of modeling can be easily found in the literature. In Chen (2014) work emphasis is placed on the machine translation applications, text classification, second language learning, child language acquisition and many other linguistic area. And Amancio, Oliveira, and Costa (2014b) explores various measures of similarity between the complex networks texts, in order to capture the semantics of the language in applications such as machine translation, machine translation and quality of authorship recognition.

Many applications, such as summarization of textual and authorship contents are strictly linked to the demand for semantic analysis of networks. As already mentioned, a summarization work was done by Pardo et al. (2005) and authorship by Amancio et al. (2014b). Other applications require finding and characterization of communities, as in the work of Bota and Kovacs (2014), which study the combination of two languages, English and Hungarian, through modeling of complex networking communities. The main feature observed in this study was the degree distribution associated with the number of related hubs for communities formation.

At the same time, the network topology greatly influences the performance of applications based on complex networks, especially from the point of view of its dynamics. For this reason, this particular topic has gained great attention. Then, the investigation of network dynamics is a required topic. In addition, Biemann, Ross, and Weihe (2012) presented a comparison of semantics of natural and artificially generated languages by modeling motifs of words. This work showed that it is possible to distinguish artificially generated languages from natural ones, and Colman and Rodgers (2012) carried out a study to investigate complex network dynamics, observing the connections of preferential attachment. This work shows the importance of studying statistical network for understanding natural phenomena. In Mehri, Darooneh, and Shariati (2012) is addressed complex networks using metrics arising from statistical mechanics, as the Tsallis q-entropy. In this study, it was shown that the morphology of degree distribution can be associated with a q-Gaussian function, and a value of q may be attributed to a specific network allowing applications such as text classification or vulnerability of networks. Under the same idea, Mishkovski, Biey, and Kocarev (2011) propose a network vulnerability study by adding and removing nodes and edges that affect important physical characteristics. The authors formally define vulnerability metrics, and also present vulnerability measures for important network models, such as: brain networks, US power grid networks, collaboration networks, urban transports networks and US power grid networks. Also, Deng, Li, Cai, and Wang (2011) demonstrated the importance of studying the equilibrium of complex networks, particularly under the degree distribution. These works show the importance of this type of study and directly inspired the ideas of our paper.

Regarding the works cited in this section, one can argue that the most recurrent issues in the area of complex networks for natural language modeling through texts are those related to factors such as semantics and cognition of languages; structure and topology of networks; extraction of features for studies of these related factors; comparison of metrics; and applications related to the studied networks. Considering all these issues, several studies have emerged in the interest of studying them for different types of languages. Thus,

Gao, Liang, Shi, and Huang (2014) compares 100 texts in various languages through weighted complex networks modeling the co-occurrences between words. The languages studied were: English, French, Spanish, Russian, Arabic and Chinese. Each one was studied under the light of scale-free and small world topologies. The authors showed that there are signs of greater flexibility in English language regarding the relation to other languages; the Spanish language is more rigid; the French and Spanish language are similar; Russian and Arabic have sparse networks, which means they are languages that have large varieties; and the connections in the Chinese language are very standardized (uniform degree distribution) with fewer words, concluding that the Chinese language is concise. In an earlier work, Li and Zhou (2007) studied, from the point of view of the network structure, the Chinese language, allowing argue important issues as combinatorial behavior of Chinese language and a non-Poisson distribution, suggesting the formation of phono-semantic characters. Likewise, Sheng and Li (2009) showed a weighted complex network to compare the Chinese and English languages which allowed getting important characteristics of the network associated with these languages, including behaviors based on language structure. And Amancio, Nunes, Oliveira, and Costa (2014a) showed various measures of complex networks that capture the text summarization, using for the first time the measure betweeness vulnerability and diversity to analyze written texts in Brazilian Portuguese. The vulnerability measure is also directly linked to physical attributes of the network whose removal can disfigure its structure. The vulnerability measure then is an important factor for the classes of problems that can be modeled by this kind of application.

Regardless the results or arguments for or against, it can be noted that the modeling of complex networks for natural languages has multiple paths with different challenges, having still much work to be done. The modeling of this problem can shed light on many current questions, both in the area of linguistic and in its close relationship with psychology and psychophysic areas.

Our work, proposed in this paper, deals with some important aspects discussed here. Specifically, we present a study on two distinct textual information databases, both modeled as complex networks of co-occurrences of words. As main contribution we present a study of the dynamic behavior of these kinds of networks regarding their physical features and topologies. The first database is in English language, specific to medical context, and the other in Portuguese, specific to journalistic context, therefore, of generic topics. The main feature studied here is related to the influence of hubs in these two types of networks. Regarding this specific feature, the network dynamic are studied under the behavior of adding and removing hubs. Whereas complex networks of words to model Portuguese language are very rare, the study of the related journalistic database is also one of the contributions of this work. As we expected, the main hubs are found stop-words. The observed results show that indeed these hubs carry weak semantic content, as the Text Information Retrieval literature supposes when eliminates such hubs in order to speed up the search process, Baeza-Yates and Ribeiro-Neto (1999). Other two physical features studied here, also from the dynamic point of view, are: Modularity and fraction of reciprocal connection.

The work proposed here improves not only the work of Amancio et al. (2014a), which specifically addresses modeling Portuguese language, but also others addressing various global languages such as in Gao et al. (2014), Li and Zhou (2007), Sheng and Li (2009). In our work, we use mainly a Portuguese database, however, differently from those proposed in Amancio et al. (2014a), Gao et al. (2014), Li and Zhou (2007), Sheng and Li (2009), in our work, as we already said, we also deal with the dynamic aspects of the proposed network.

On the other hand, due to the large volume of information that have emerged, mainly due to the explosive growth of the Internet, there is a need to revise the majority of work from the point of view of large databases. It implies higher computing power and more efficient management algorithms, making the network physical features important research issues, as is the case of clustering coefficient, which is still a challenging task due to its time complexity of brute force $O(n^3)$. The computation of this feature is essential to study the topologies of several networks and consequently their semantic behaviors. However, for the sizes of current databases, this computational time becomes prohibitive. In this sense, as other important contribution, we present an efficient quadratic time method, based on Schank and Wagner (2005) work, for statistical estimation of this specific and important feature, and use it over the databases of our study.

This work is organized as follows. Section 2 presents the main concepts of complex networks and in Section 3 we describe their physical properties and methodology to compute them. In Section 4 the two databases used in our experiments are described. Next, Section 6 describes the experiments and discussed the results. Finally, Section 7 summarizes the results of our experiments and outlines possible future research directions in this area.

2. Complex networks

Complex networks have been used to model a variety of systems. Some studies using complex networks can be found in Erdős and Rényi (1959), Solomonoff and Rapoport (1951), Price (1965), Albert, Jeong, and Barabasi (1999), Newman and Girvan (2004), Newman, Barabasi, and Watts (2006), Allesina and Pascual (2008) and Wachs-Lopes (2011). From social to human language, it has been used as a framework to explore the properties and study the behavior of systems.

The complex networks model is similar to a graph (Newman et al., 2006). Therefore, it can be implemented using the same mathematical background. A graph is an ordered pair $G = (V, E)$ where V is a set of vertices and E is a set of edges, i.e., two-element subsets of V.

Basically, there are two ways to implement a graph. The first one uses an adjacency list. This list contains three columns: the first column indicates the nodes from which the edges leave; the second indicates the nodes to which the edges arrive; and the third indicates the weights of the edges. The second way for implementing a graph is through an adjacency matrix M. This matrix has $|V|$ lines and $|V|$ columns. Each cell $M_{i,j}$ represents the weight of an edge. Each line i corresponds to a node that an edge leaves, and each column j corresponds to a node to which an edge arrives. In both models we can implement direct and indirect (or simple) graphs. However, in the case of a direct graph, the adjacency matrix has only the upper diagonal or the lower diagonal part filled.

When implementing a complex network, we have to keep in mind that its density may be analyzed in order to decide what model to choose. If a dense network is expected, you may consider the adjacency matrix approach. However, if a sparse network is expected, you should use the adjacency list.

3. Network features and methodology of extraction

3.1. Degree distribution (DD)

The degree of a node n is the number of all edges that arrive and leave from it. In directed networks, this property is also divided into two sub-properties: in-degree and out-degree.

The in-degree is the number of edges that arrive at a node and the out-degree is the number of edges that leave it. For weighted networks, this property can be extracted through the sum of incoming/outgoing edges. So, the following property holds:

$$\sum_{i \in V} In(i) = \sum_{i \in V} Out(i), \tag{1}$$

where $In(i)$ is the in-degree and $Out(i)$ is the out-degree of a node i.

In addition to the individual degree of a node, there are two other properties that are normally analyzed for complex networks: the average degree of the network and the degree distribution. The average degree is denoted by $< k >$ and computed as

$$< k >= \frac{|E|}{|V|} \tag{2}$$

This property relates the number of edges with the number of nodes. However, it does not contain information about the dispersion of the degrees.

A property that contains information about dispersion of degrees is the degree distribution (DD). Important studies indicate that scale-free networks follow a degree distribution according to a power-law equation, given by Eq. (3): (Albert et al., 1999; Price, 1965; Réka & Barabási, 2002).

$$P(k) \sim k^{-\gamma} \tag{3}$$

In this work, we compute the degree distribution using two approaches. The first simply counting the number of edges entering and leaving each node. The second is obtained from the summation of the weights for every edge entering or leaving each node.

Section 6.1 shows the results obtained from both methods.

3.2. Weight distribution (WD)

The weight distribution is another physical property used to analyze the topology of a network. It shows how frequent a weight appears in the network and can reveal how much the words belong to the same context. The more specific words co-occur, it means that the language contains concise contexts and more sparse is the network. Thus, the more concentrated is the weight distribution, it means that few words co-occur often. The more the weights are well distributed throughout the network, it means that the language is very general, i.e., linguistically speaking, expressions in such language may be expressed similarly in many different ways. In this paper, we propose the use of the weight distribution to roughly analyze the frequency with which each co-occurrence weight appears on the network.

An interesting way to observe how the behavior of WD property is associated with the network topology is to calculate its entropy. A low entropy means a concentration of information, leading possibly to a network without a clear topology, since in this case the edges have similar weights. However, if the entropy is high, it means a heterogeneous weight distribution. In this case, we cannot infer the network topology without the analysis of other properties (Wachs-Lopes, 2011).

As this property is given by a distribution, it is necessary to define a discretization. At this point, we decided to discretize the weights in 100 values. The justification for this choice is directly related with the application involved, in the case of this work, a complex network of words.

3.3. Averaging clustering coefficient (ACC)

The clustering coefficient indicates how a neighborhood of a node is connected with itself (Newman et al., 2006). The clustering coefficient is a simple ratio of number of edges in a neighborhood and the total number of edges possible to connect all nodes in this neighborhood.

Formally, suppose i is a node in a complex network, L_i the set of nodes that have a connection with i, and $W = \{w_{u,v} | u, v \in L_i\}$ the set of edge weights that connect each node from L_i to another node of L_i. Eq. (4) shows how to obtain the clustering coefficient CC_i of a node i in a directed network.

$$CC_i = \frac{|W|}{|L_i|^2 - |L_i|} \tag{4}$$

This equation only considers the number of connected edges in a neighborhood but not their weights. It means that if an edge has a low weight, it will contribute as one edge instead of its weight.

Considering the limitation of Eq. (4), we propose the use of the weights inside this equation in order to achieve more precision. With this in mind, we change the definition of Eq. (4) to the following:

$$CC_i = \frac{\sum_e W_e}{|L_i|^2 - |L_i|} \tag{5}$$

Besides the individual clustering coefficient of a node, there is another important general property of a complex network: the average clustering coefficient (ACC).

The *ACC* has been studied by many researchers and is used mainly to infer the type of network (Newman et al., 2006; Wachs-Lopes, 2011; Watts & Strogatz, 1998). For instance, Watts and Strogatz (1998) define a complex network as a small world if it contains two properties: a) the mean distance between all nodes (l) is comparable to that of a random network with the same number of nodes and edges, $l/l_{rg} \sim 1$; and b) the ACC is much greater that of a random network, $ACC/ACC_{rg} \gg 1$.

In order to compute the *ACC*, we have to compute the individual clustering coefficients CC_i and take their overall average.

$$ACC = \frac{1}{N} \sum_i CC_i \tag{6}$$

The *ACC* can be infeasible to compute for large and dense complex networks. In this case, its complexity is $O(N^3)$ since we have to compute the individual clustering coefficients. In this work, we propose an approximation of the *ACC* by improving a statistical simplification proposed originally in Schank and Wagner (2005).

Analyzing a complex network of N nodes in terms of probability, a node i has a probability $1/N$ to be randomly chosen. Whereas Eq. (6) is an average over all clustering coefficients, we can compute a clustering coefficient approximation from some random nodes.

This method chooses $Nr \ll N$ nodes randomly and computes the clustering coefficient, according to Eq. (5), considering the original network. After this stage, using Eq. (6), we obtain the average clustering coefficient of the Nr nodes. Note that this is a statistical approximation of the real clustering coefficient. However, this idea is supported by the Central Limit Theory.

$$ACC = \frac{1}{Nr} \sum_{i=0}^{Nr} CC_i \tag{7}$$

Algorithm 1 shows the steps needed to compute the Random-Based average clustering coefficient.

3.4. Modularity

Modularity is a physical property of a network based on a comparison of random networks with the model observed. Therefore, a set of nodes is a cluster if the number of edges between them is greater than that expected if the network was totally random. The quality of a network C for a set of nodes N can be obtained using Eq. (8) (Newman & Girvan, 2004):

$$Q(C) = \sum_{i,j \in N} \left[P_{i,j} - \frac{M_{i,j}}{2m} \right] \delta(c_i, c_j) \tag{8}$$

where $P_{i,j} = \frac{k_i^{out} k_j^{in}}{m^2}$ is the expected probability connection between nodes i and j for a random network with the same number of nodes and edges, k_i^{out} is the number of outgoing edges from i, k_j^{in} is the number of incoming edges to j, and m is the sum of weight matrix.

The δ function is defined as:

$$\delta(c_i, c_j) = \begin{cases} 1 & : \text{ if } i \text{ and } j \text{ are in the same community.} \\ 0 & : \text{ otherwise.} \end{cases}$$

Algorithm 1: Random-based simplification.

Input: (N, E), Nr
Output: Average clustering coefficient
initialization;
Chooses Nr nodes between 1 and N;
$V1 \leftarrow RandomInt(1, size(N), Nr)$;
for *each node* $i \in V1$ **do**
  $V2_i \leftarrow FindNeighbors(N, E, i)$;
  $V3_i \leftarrow 0$;
end
for *each node* $k \in V1$ **do**
  for *each node* $i \in V2_k$ **do**
    for *each node* $j \in V2_k$ **do**
      if $EdgeExists(AD, i, j)$ **then**
        $V3_k \leftarrow V3_k + Weight(N, E, i, j)$;
      end
    end
  end
  $V4_k \leftarrow \frac{V3_k}{size(V2_k)*(size(V2_k)-1)}$;
end
Return $Mean(V4)$

In Newman and Girvan (2004), Eq. (8) was proposed as a measure of clustering quality. Furthermore, many applications were presented and discussed where this measure was successfully used.

In this work, we use the Modularity property to measure the topological organization of the network. The extraction of this property is made from a complex network generated from a supervised database, called Newpaper database (explained in Section 4.2).

The advantage of using a supervised database, as the one proposed in this work, is the possibility of extending the experiments by controlling the amount of topics and, consequently, studying how this metric is affected by this variation. Moreover, a deeper discussion can be made regarding the database features with the this metric.

For this study, we propose the construction of three different complex networks, each one built from the supervised database with a different number of topics: 5, 12 and 21. After all, we will use a clustering algorithm for complex networks proposed in Pons and Latapy (2004) based on random walks for clustering nodes. The focus here is to detect communities, in order to subsequently measure the quality of this clustering.

3.5. Fraction of reciprocal connections (FRC)

Reciprocal connections are pairs of edges connecting two nodes in two contrary directions. Formally, a reciprocal connection exists if $M_{i,j} > 0$ and $M_{j,i} > 0$ for $i \neq j$, where M is the weight matrix. The ratio of the number of reciprocal edges to the number of edges from the network is a measure known as the fraction of reciprocal connections (FRC), and is denoted by ρ.

The construction of complex networks in the proposed work takes into account the order in which the words appear in a document. This means that a word a can connect to a word b, however without the word b connecting to a. A word with this feature clearly presents a syntactic rule for placing it in a sentence. For instance, we can not put an article after a noun to reference it. Thus, the definite article "The" in the phrase "The house is big" can not be placed in another position, as in the phrase "house the is big." This syntactical language stiffness can be obtained by examining the number of reciprocal connections.

Analyzing the syntactic effect on the semantics, one can also speculate that the more the syntax features have well-defined rules, the less ambiguity will be found in the sentences. This will provide a better description of the message that you want to send, making the understanding of the sentences clearer. For example, in the sentence "The peasants are revolting" (extracted from Hart & Parker, 1981) it is not known whether revolting is a verb or an adjective.

In this paper we propose the use of this measure to check how the words of a language change its order. We will measure the fraction of reciprocal connections (FRC) of networks generated from 5, 12 and 21 topics from an supervised database. In addition, we will remove the *hubs* (nodes with the highest degrees) and study the behavior of this metric to verify the importance of hubs in this type of network.

3.6. Average connection density (ACD)

The average connection density is a physical property that relates the number of edges in a network with the number of edges in a totally connected network with the same number of nodes. Therefore, this property varies between 0 and 1, according to the density of the network connections, and is related to the completeness of a graph.

In Sporns (2003), the author comments on studies that infer the topology of complex networks from neurological data through measurements of local densities (in a particular part of the network) and global densities (across the network). So this measure can provide structural information about a network.

Eq. (9) shows how to retrieve the average connection density for directed networks:

$$K_{den} = \frac{|E|}{|N^2| - |N|} \tag{9}$$

where $|E|$ is cardinality of the set of edges and $|N|$ is the cardinality of the set of nodes. According to this equation, when its value is close to zero, there are few edges on the network and it is considered to be sparse. However, when its value is close to one, the network contains many edges, and is considered dense.

All the complex networks generated in this work are built from co-occurrence interactions between words. However, these co-occurrence are related to the type of the database used: specific or generic. In the case of a specific database, what is expected is that both the number of words and topics (clusters) is small, generating a greater number of co-occurrences between words. On the other hand, generic databases tend to contain generic contexts and can represent great diversity of words and topics. Thus, the networks built from generic databases tend to generate a smaller number of co-occurrences between their words. The amount of relationship between the words of the networks can be measured through the average connection density, since this is a measure which is strictly linked to the number of edges and the maximum possible.

It is known that the hubs participate in multiple contexts simultaneously. For a network of words, it means that a hubs can be a generic word, usually a stop-word or a connective as well. As is well known, hubs have a fundamental role in scale-free and small-word networks. Therefore, we propose to study the variation of the average connection density (ACD) according to the variation of the number of hubs.

4. Databases

In order to study some properties of the human language, we created complex networks from two types of databases: a collection of medical texts and a collection of journalistic texts.

Both databases were treated before the construction of the corresponding network. This step is responsible for splitting the databases into a set of documents (D), where each document D_k is a vector of words that preserves the original ordering of the document.

4.1. Scientific database

The scientific texts studied in this work were extracted from the Ohsumed database (Hersh, Buckley, Leone, & Hickam, 1994). This

```
.I 3246
.U
87077345
.S
Br G Wachs 8704; 73(12):1012-4
.M
Example
.T
A simple title.
.P
JOURNAL ARTICLE.
.W
This is a small document.
.A
Author 1; Author 2; Author 3.
```

Fig. 1. An example of a TREC document (Ohsumed).

database contains a collection of abstracts in English from medical papers. In this database, there are 221, 175 distinct words arranged in 54, 710 extracts, where each extract represents a medical paper abstract.

As shown in Fig. 1, each extract is organized with the following labels: abstract(.W), title(.T), authors(.A), keywords(.M), and a few others. Although this database contains all these labels, only the abstract texts were used to build the network.

4.2. Newspaper database

The Folha de São Paulo database[1] contains 649, 490 distinct words arranged in 340, 947 news items in Portuguese. The news in this database was from 1994 and 1995 and is classified into 21 subjects: Agronomy, Brazil, Finance, Jobs, Sports, Teenagers, Real State, Computer Science, World, Television, and few others.

In this work, some experiments take advantage of topic supervision and are carried out with different number of topics. This approach makes possible to analyze the influence of topics on measurements.

5. Complex network construction

In this work we generate two types of complex networks of words. The first one is based on technical texts and the second based on journalistic texts. Both networks are generated by the same method: the co-occurrence of words.

The proposed model represents each word as a node in a network as suggested in Cancho and Solao (2001). When scanning the database, such nodes are connected according to their co-occurrences in the same document. It means that the more often two words appear together in a same document, the greater will be the weight of the edge that connects them.

Also, taking into account the importance of the sequence of words, we must consider modeling the order of the words in a document. With this in mind, the proposed model maps the sequence of words through directed edges. Therefore, if a word i appears before a word j, the created edge goes from i to j.

Another important consideration about our model is that we also take into account the distance between words when computing the co-occurrence weights. This means that the closer two words appear, the greater will be the weight of the edge that connects them, given by Eq. (10).

[1] Project hosted at:http://www.linguateca.pt/

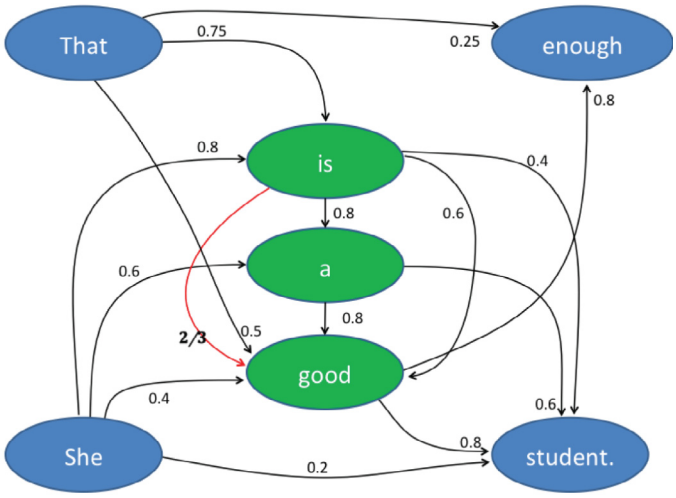


Fig. 2. Network of words obtained from the document in Fig. (1), "This is a small document."

$$w_{i,j} = 1 - \frac{\sum_D Dist(D,i,j)}{\sum_D Size(D)} \tag{10}$$

In Eq. (10), $w_{i,j}$ is the weight between words i and j, D is the set of documents, and $Dist(D, i, j)$ is the distance between words i and j in document D. This function will return 1 if the word i appears next to word j. Finally, $Size(D)$ is the size of document D in number of words.

For a practical example, given all the rules above, consider the following two documents:

$$D_1 = \{\text{'That', 'is', 'good', 'enough.'}\} \tag{11}$$

$$D_2 = \{\text{'She', 'is', 'a', 'good', 'student.'}\} \tag{12}$$

In order to build a complex network from these extracts, we apply Eq. (10) to each co-occurrence of words in the documents. For instance, the co-occurrence of the word "That" with "good" can be measure as $w = 1 - 2/4$, since the size of document is 4 words and the distance between "That" and "good" is 2 words. In the end of the building process we will have the network as shown in Fig. 2. In this specific case, we have to draw attention to the co-occurrence of "is" and "good". These words appear in both texts, resulting in a weight equal to $0.\overline{6}$. We achieve this result by applying Eq. (10). The final result of this process is two complex networks: a technical medical network and a newspaper network.

6. Experiments and results

The experiments in this paper are intended to explore the dynamic behavior of network on the two considered databases: Ohsumed and Newspaper. The first part of experiments evaluates physical properties of the networks that allow to reveal the type of studied network from the point of view of its topology. Thus, we will test: the degree distribution (Section 6.1), the weight distribution (Section 6.2), and the average clustering coefficient (Section 6.3).

In the second part of the experiments will be evaluated the no less important features such as: Modularity (Section 6.4), fraction of reciprocal connections (Section 6.5) and average connection density (Section 6.6). These physical measures will be evaluated only for Newspaper database, since they depend on an annotated database for their evaluation. In addition, this assessment is carried out under a dynamic behavior of the network, as more and more hubs are removed or inserted.

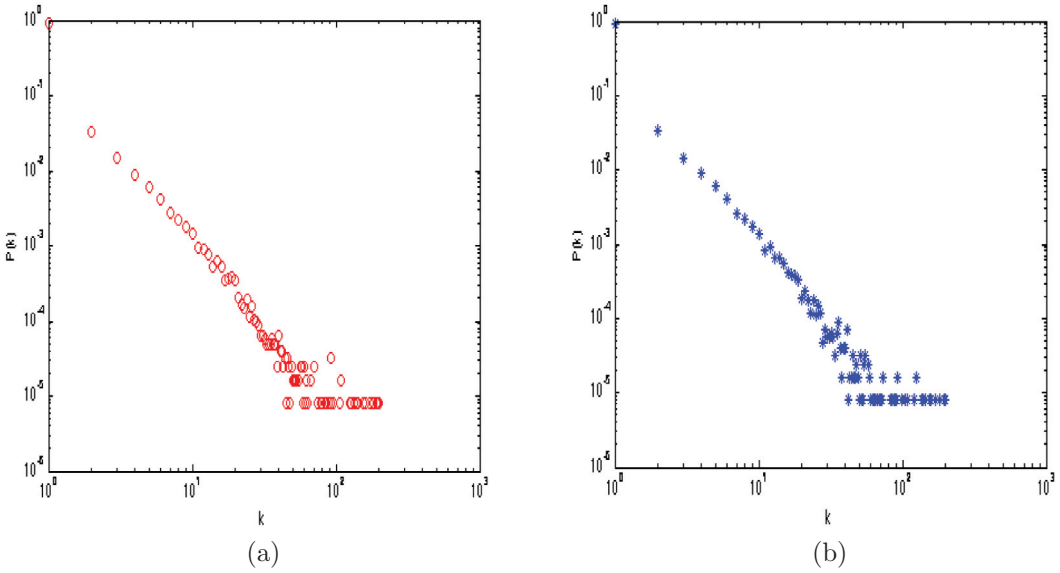


Fig. 3. Network degree distribution for Ohsumed database. (a) Based on incident edges; (b) based on weights of incident edges.

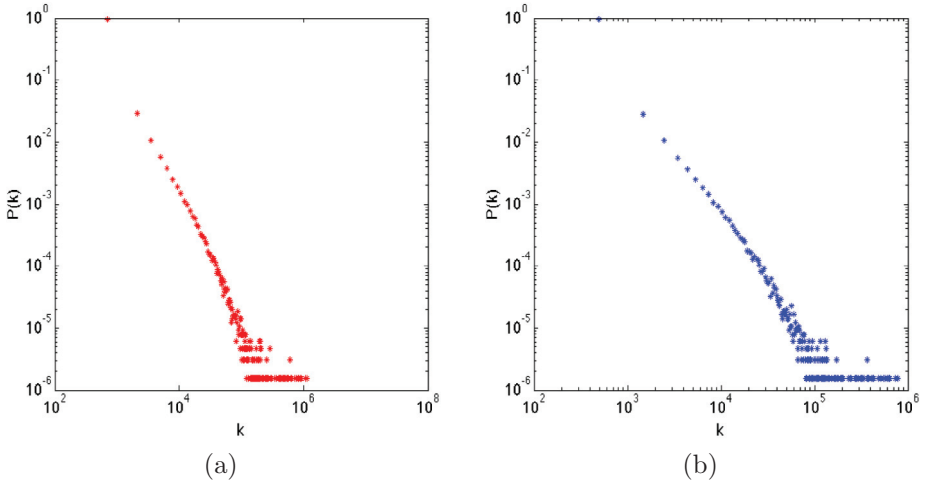


Fig. 4. Network degree distribution for Newspaper database. (a) Based on incident edges; (b) based on weights of incident edges.

6.1. Degree distribution

degree distribution (Section 3.1) is an important topological property that gives an idea of the network connectivity. For each type of context modeled here (medical and journalistic) it was built two types of degree distribution. The former considers only incident edges in each node; and the second considers the sum of incident's weights. Both types produce power-law distributions, as can be seen in Fig. 3a and b for Ohsumed database and in Fig. 4a and b for Newspaper database. These figures indicate that both Ohsumed and Newspaper network have scale-free or small-world behavior.

Considering Eq. (3) for power-law behavior of the four types of networks, the γ values are: $\gamma = 1.077$ for incident edges of Ohsumed (Fig. 3a); $\gamma = 1.1066$ for weighted incident edges of Ohsumed (Fig. 3b); $\gamma = 1.069$ for incident edges of Newspaper (Fig. 4a); $\gamma = 1.1018$ for weighted incident edges of Newspaper (Fig. 4b). These values are consistent with other findings in the literature for other languages (Gao et al., 2014).

Fig. 3a shows the degree distribution for the Newspaper database when only incident edges are considered, and Fig. 3b shows the corresponding distribution when incident edges are weighted. In both cases we can clearly observe the power-law behavior.

Table 1
20 first words of Ohsumed database in decreasing order of node degree.

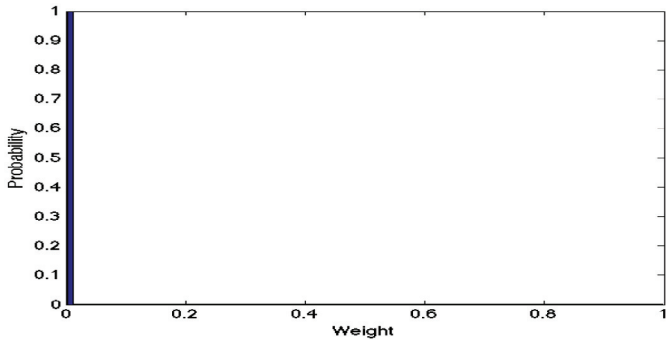
	Word	Degree		Word	Degree
1	of	147452.00	11	by	90021.10
2	the	142448.00	12	for	88009.10
3	and	139092.00	13	that	78924.20
4	in	130601.00	14	from	76556.50
5	to	120512.00	15	or	76393.10
6	a	113976.00	16	on	67856.50
7	with	112310.00	17	is	65898.80
8	was	102333.00	18	as	63063.50
9	were	100101.00	19	patients	62115.10
10	The	99512.00	20	an	62029.60

Fig. 4a shows the degree distribution for the Newspaper database when only incident edges are considered, and Fig. 4b shows the corresponding distribution when incident edges are weighted. In both cases we can clearly observe the power-law behavior.

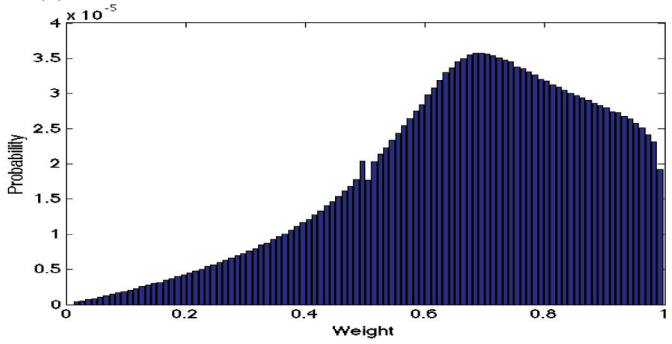
In the case of Ohsumed database, it is also interesting to observe which are the more connected nodes (the so-called hubs) of each type of network. Table 1 shows the rank for the first 20 great hubs in descending order of node degree.

Table 2
10 first words of Newspaper database in decreasing order of node degree for 5, 12 and 21 topics.

	Network of 05 topics		Network of 12 topics		Network of 21 topics	
	Word	Degree	Word	Degree	Word	Degree
1	.	17839.5	.	17937.8	.	17708.8
2	de	15670.2	de	15602.9	de	15546.5
3	e	15153.4	e	14957.2	e	15002.9
4	a	14933.1	a	14755.3	a	14784.2
5	o	14625.9	o	14447.7	o	14445.1
6	que	14110.6	que	13679	que	13984.7
7	do	14095.7	do	13612.8	do	13672.8
8	da	13649.9	da	12920.8	da	13088.9
9	em	12528.1	em	12528.9	em	12642.2
10	para	11837.8	para	11272.2	para	11420.4



(a) Weight distribution for Ohsumed database for all bins



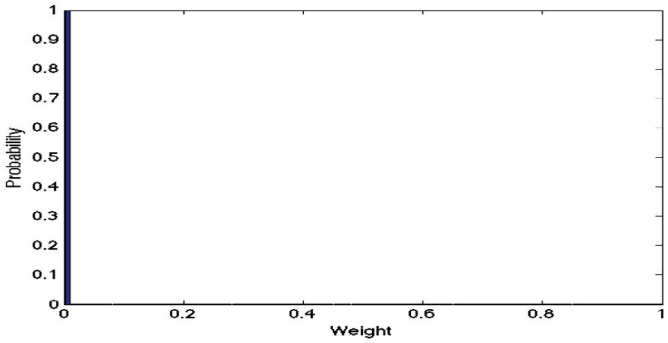
(b) Weight distribution for Ohsumed database for bins with probability greater than zero.

Fig. 5. Weight distribution for Ohsumed database.

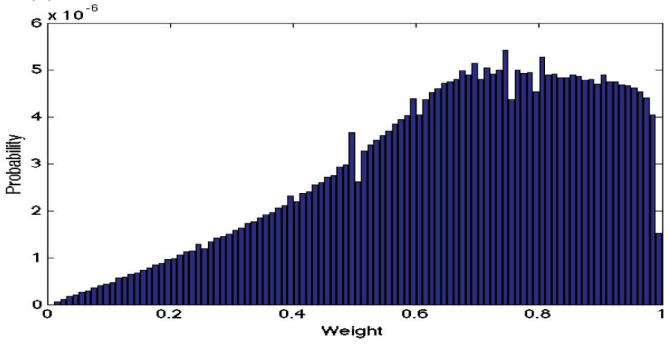
Observing the Table 1, it is easy to conclude that the more connected nodes (hubs) in this database are exactly the stop-words (such as prepositions, connectives, definite and indefinite articles, and so on), as expected. The same can be said for Newspaper database observed in Table 2. However, in the case of Newspaper database Table 2 shows the first 10 great hubs in descending order of node degree for 5, 12 and 21 topics. It is noteworthy that, in the classical literature of textual information retrieval (Baeza-Yates & Ribeiro-Neto, 1999), such words are filtered when building a vector space of documents in order to speed up the search process. So in this paper, among other features, we are interested in the dynamic behavior of networks as we filter hubs in descending order of degree of connectivity.

6.2. Weight distribution

As stated in Section 3.2, the weight distribution can be computed throughout probabilistic density function represented by a histogram of weights. Fig. 5a and b shows two histogram distributions



(a) Weight distribution for Newspaper database for all bins



(b) Weight distribution for Newspaper database for bins with probability greater than zero.

Fig. 6. Weight distribution for Newspaper database.

of weights for a network modeling the Ohsumed database, that has a specific context of the medical field.

The first histogram of Fig. 5a covers all ranges of possible weights between 0 and 1. Note that there is a large concentration of weights (approximately 99% of total edges) equal to zero in the first bins, clearly showing that this is a very sparse network.

In order to observe the curve for histogram bins with probability values greater than zero, the Fig. 5b shows the same histogram of Fig. 5a but now only for bins with probability values greater than zero. It suggests that, for this particular network, a considerable range of words are strongly connected, strengthening the construction of specific contexts, which is expected for this type of base, of specific medical context.

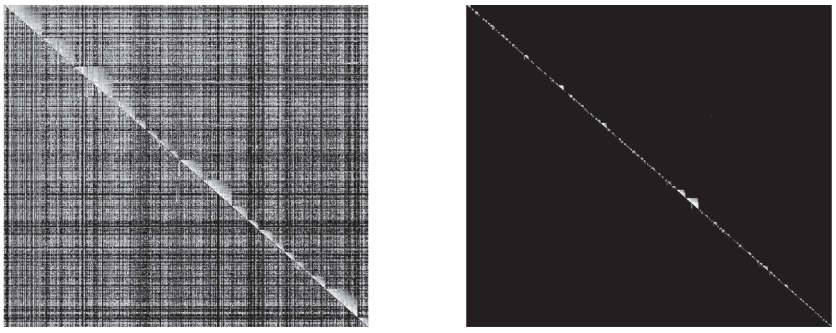
The weight distribution was also observed for the Newspaper database, showing the same behavior observed in Ohsumed. Fig. 6a shows the distribution for all bins and Fig. 6b shows its corresponding distribution for only bins with probability greater than zero.

Observing the weight distributions for Newspaper and Ohsumed databases, we can note that Ohsumed has its highest value around 0.7 and Newspaper database is higher around 0.8 bin. This suggests that the more specific the network, the lower is the vocabulary of words and most frequently is the co-occurrence, which is attenuated by the amount of weights.

6.3. Average clustering coefficient

As stated in Section 3.3, the method to compute the average clustering coefficient used in this paper was based on Schank and Wagner (2005)'s work, which proposes Eq. (4) for non-weighted networks. Thus, since in our paper we work with two weighted networks, we propose Eq. (5).

Once this physical measurement is directly proportional to the number of connections in a neighborhood of a node, it is interesting to compare the performances for both a sparse and a dense networks, showing that this property is similar for both types of



(a) (b)

Fig. 7. Sub-samples screenshot complex networks from Ohsumed database. (a) A dense one; (b) a sparse one.

topologies. Then, the objective of the experiment here is to demonstrate that this important physical measurement can be estimated from a smaller number of nodes (a subset of network), greatly reducing the computational complexity time of $O(n^3)$ order, which could be very time-consuming under the total number of nodes. As described in Section 3.3, this idea is supported by the Central Limit Theorem.

In order to build dense databases to our experiments, and are at the same time small enough to allow the ACC computation with a brute force algorithm only for comparison, we took a small sub-matrix from the overall network. This sub-matrix was built from the first 8870 lines and columns from the total of 221, 175 $\times$ 221, 175 connection matrix of the Ohsumed database. Fig. 7a shows a screenshot of this sub-matrix, where the rows and columns correspond to the sub-sampled words, and each cell has a gray-scale color (in the range where 0 is black and 255 is white) proportional to the connection weight. It may be noted in this case this sub-matrix corresponds to a very dense matrix. Similarly, to construct a sparse corresponding network, we took a sub-matrix of the overall connection matrix. However, this sparse sub-matrix corresponds to the latest 8870 rows and columns of the total connection matrix. Fig. 7b shows a corresponding screenshot for this other sub-matrix, where again the rows and columns correspond to sub-sampled words, and each cell has a gray-scale color proportional to the connection weight. Now, it may be noted that in this case this sub-matrix corresponds to a very sparse matrix. Therefore, both sub-matrices result in two complex networks with a number of nodes small enough to allow the force-brute computation of the original ACC in a feasible time, in order to compare numerically with the corresponding estimated ones.

Fig. 8 shows the ACC when only two nodes are taken randomly from the dense network, repeating this experiment for 100 pairs of nodes randomly taken. In this graph, the central solid line *l* represents the original ACC computed according to Eq. (6) with a brute force algorithm. The horizontal axis stands for 100 different random samples and in the vertical axis are their corresponding approximated AAC. In this particular case, we note that the deviation of the estimated value from the real ACC value, represented by line *l* is large and around 9%, with a maximum of 34% error. However, the average estimated value occurs around the real value. This small number of used nodes corresponds to only 0.1% of the 8870 dense network nodes.

Observing Fig. 9, when the number of involved nodes was increased to 20 (corresponding to 1% of the total nodes from the dense network), the average error falls to 3% and the maximum error was 11% in relation to the average line representing the real value. Similarly, by observing Fig. 10, when the number of nodes involved rises to 500 (corresponding to 25% of nodes of the total nodes from the dense network), the average value falls to 1%, and the maximum value to 3%.

The results shown in Figs. 8–10 for a dense matrix suggest that the method proposed here to calculate the ACC based on only a small

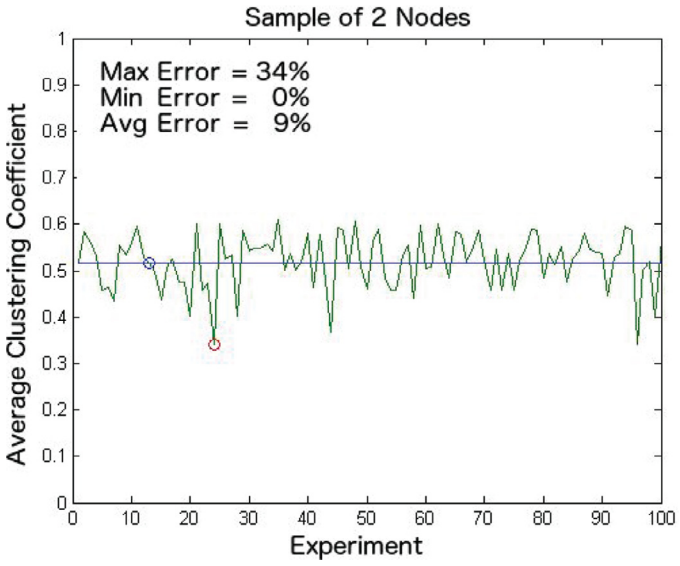


Fig. 8. Av. Clust. Coeff. for a dense matrix with 0.1% nodes from the Ohsumed database.

sample of randomly chosen nodes of a network, fast approaching the calculated value in all nodes for a small portion of the total. This means that it is possible to estimate the ACC for large complex networks where the objective is only to get a rough estimate of its value, for example, to evaluate the type of network studied.

The same procedures may be executed over the corresponding sparse networks of Ohsumed database. Figs. 11–13 show the estimated calculation of the ACC for this type of network. In this case, the real ACC was 0.355, represented by the horizontal solid line in each figure. Similarly, as occurred in the dense matrix, the ACCs were estimated with 0.1%, 1% and 25% of the total nodes. As can be seen, these estimated values were computed with average error of 22%, 7% and 1%, respectively.

Since it is possible to estimate an approximate value for ACC in the case of dense or sparse network, it can be concluded that it is possible to calculate the ACC for any type of network. Thus, we applied the same technique to the entire Ohsumed and Newspaper databases. Results showed that Ohsumed's ACC was around 0.529 and Newspaper's ACC was around 0.517. As expected, Ohsumed's ACC has greater ACC than Newspaper network, since it is a specific context database and words present a more frequent co-occurrence.

Moreover since the ACC of a random graph is E/N^2, where E is the number of graph's edges, we compared these findings with ACC of such graphs using the same number of nodes and edges of Ohsumed and Newspaper databases. The random network's ACC with the same

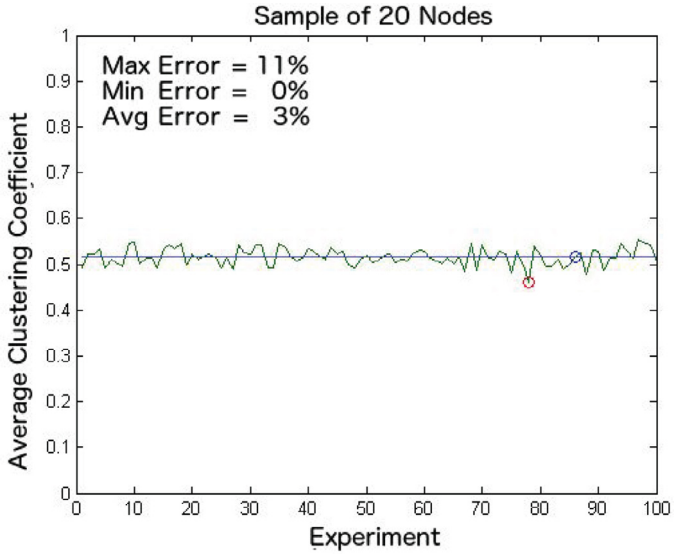


Fig. 9. Av. Clust. Coeff. for a dense matrix with 1% nodes from the Ohsumed database.

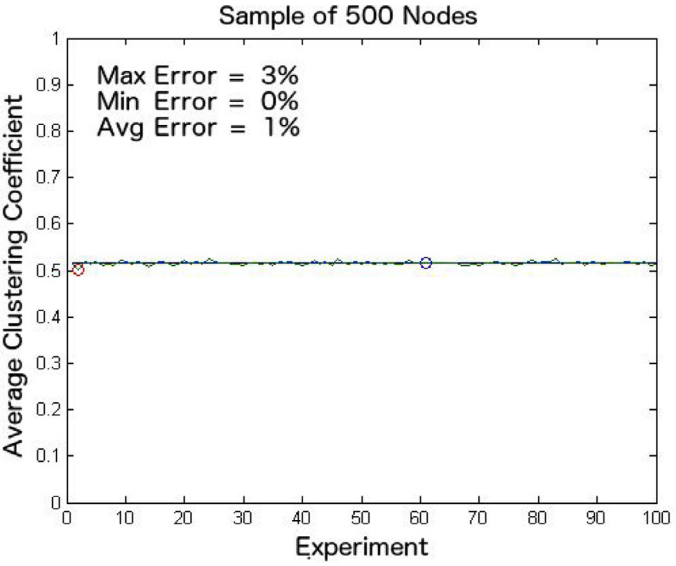


Fig. 10. Av. Clust. Coeff. for a dense matrix with 25% nodes from Ohsumed database.

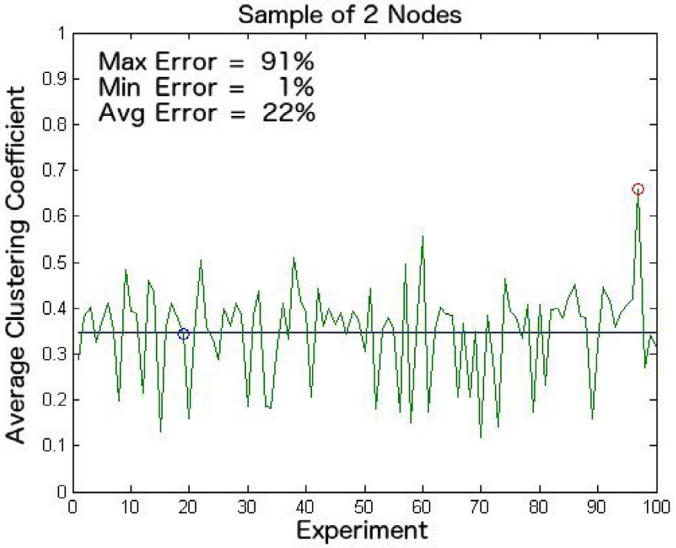


Fig. 11. Av. Clust. Coeff. for a sparse matrix with 0.1% nodes.

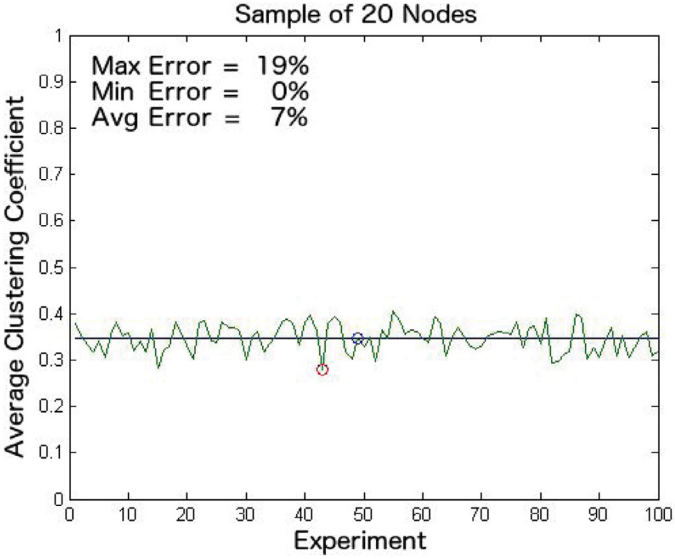


Fig. 12. Av. Clust. Coeff. for a sparse matrix with 1% nodes.

number of nodes and edges of the Ohsumed database is 0.001, and in the case of Newspaper database this value is 0.0001.

Comparing these results, we found that both networks, Ohsumed and Newspaper, have ACCs much greater than their equivalent random networks. According to literature, this strongly suggests that both networks behave as small-world model.

To our knowledge, is the first time that this technique, originated in the work of Schank and Wagner (2005) for nonweighted networks, is applied to weighted networks. Recalling that the ACC computation is essential for estimating the kind of networks in a number of practical applications.

6.4. Modularity

In this experiment, to measure how a network differs from a totally random complex network, we used the algorithm proposed in Newman and Girvan (2004) in order to compute the network Modularity. As discussed in Newman and Girvan (2004), the more modular is a network, the better the clustering (many connections between nodes within the same cluster, and few connections between clusters). For this reason, the Modularity is the physical property that can measure the topological organization of a network with more accuracy.

As discussed earlier, once hubs connect multiple clusters, they strongly affect the Modularity. Therefore, we carried out experiments that measure the influence of this type of node on the Newspaper complex network varying the number of topics as 5, 12 and 21.

Figs. 14–16 show the results of the experiments. The *x*-axis is the entry for the number of nodes in the network and the *y*-axis is the corresponding Modularity, computed according Eq. (8). Note that the values of *x*-axis are in descending order of removing hubs as we move to the right, meaning that the network decreases as more and more hubs are being removed. The nodes are removed according to their degree of connectivity, so nodes with higher degrees (hubs) are removed first. In all figures the maximum Modularity is indicated with a small circle. This point is projected onto the *x*-axis in order to show the size of the network that generated the maximum possible Modularity. The effects of removing hubs over the Modularity can be observed even in a small part of the network. Thus, this experiment was conducted for 3000 words in 5, 12 and 21 topics.

As can be seen clearly in each of the three graphs, as we removed the hubs in descending order of connectivity, Modularity has a strong

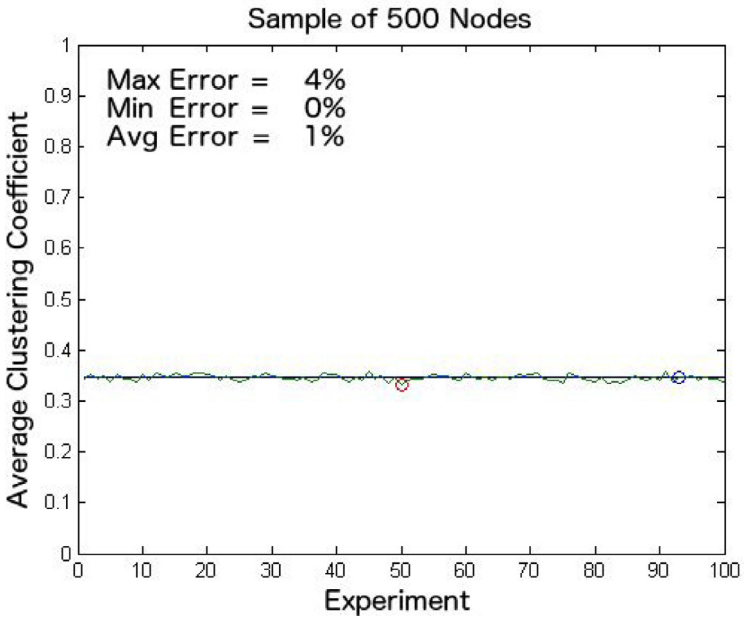


Fig. 13. Av. Clust. Coeff. for a sparse matrix with 25% nodes.

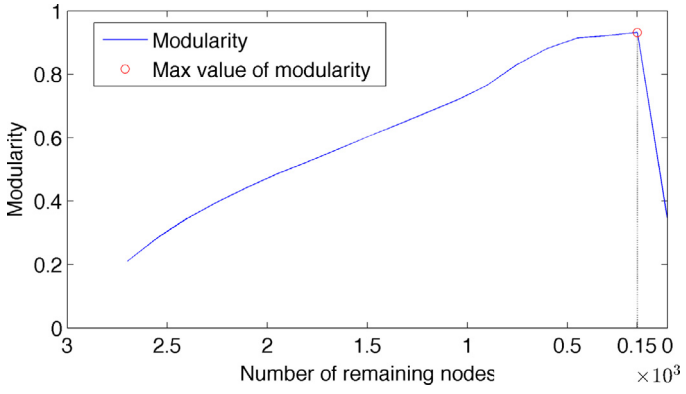


Fig. 14. Modularity from the news complex network with 5 subjects.

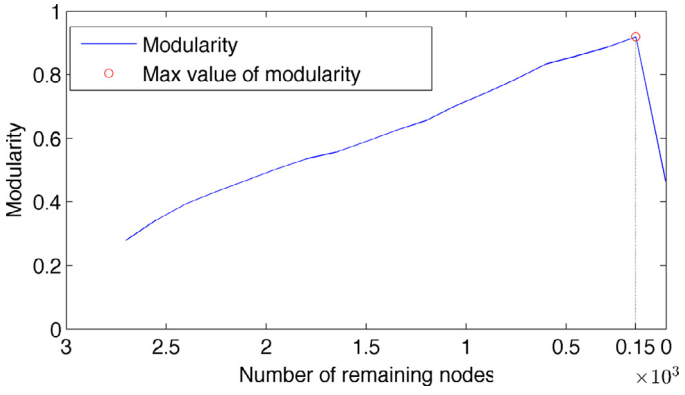


Fig. 16. Modularity from the news complex network with 21 subjects.

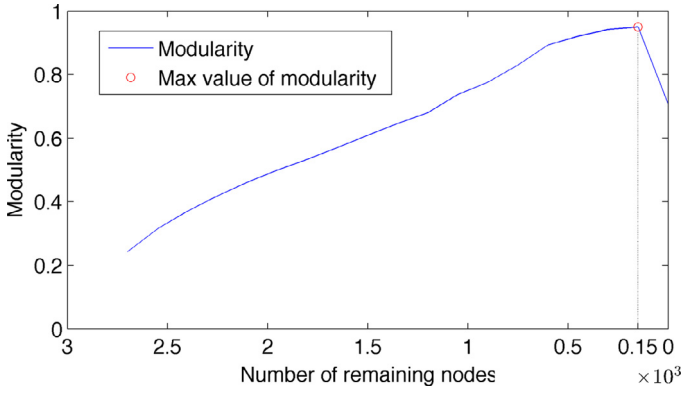


Fig. 15. Modularity from the news complex network with 12 subjects.

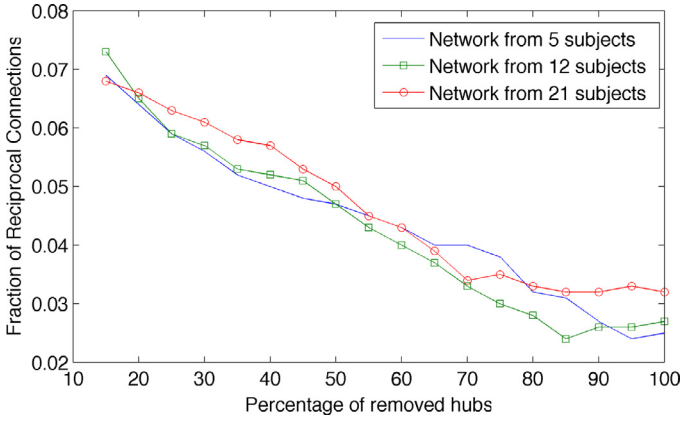


Fig. 17. Fraction of reciprocal connections (FRC) results.

growth reaching a peak (indicated by a small circle) to less than 200 nodes when the Modularity slumps.

This behavior can be explained by the role of hubs. Since hubs are nodes connecting many other nodes within or outside clusters, removing a hub decreases the connection between clusters, increasing the overall Modularity. However, if such removal continue indefinitely, as in the case of the experiments here, at some point we will remove key-nodes, which are those that maintain some clusters together, causing an abrupt decrease of overall Modularity. For the experiments here, the abrupt fall of overall Modularity with few remaining nodes on the network, suggests that it is a network with high level of Modularity and contexts. Networks with this type of

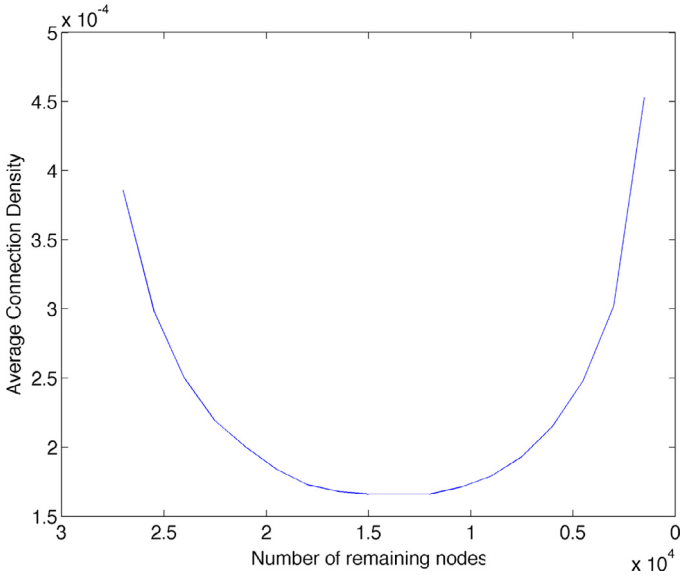


Fig. 18. Average connection density from the Newspaper complex network with 5 topics.

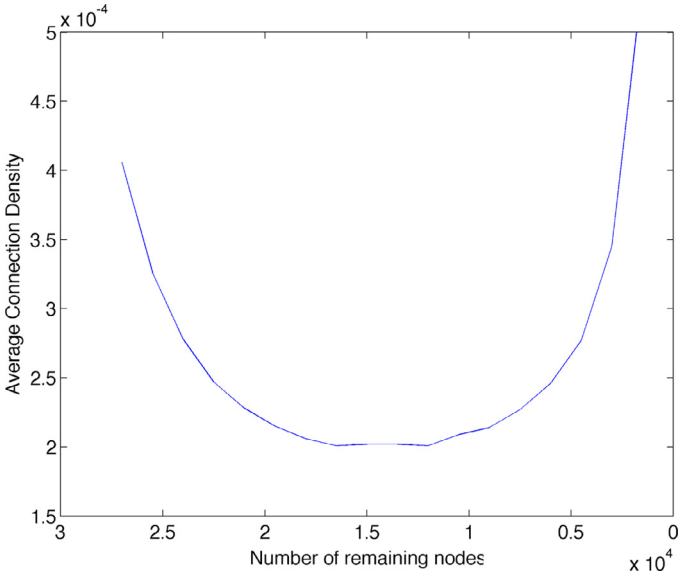


Fig. 19. Average connection density from the Newspaper complex network with 12 topics.

behavior favoring applications that are based on linguistic elements such as machine translation, authorship verification or semantic interpretations of texts, since such applications are highly dependent of linguistic contexts.

6.5. Fraction of reciprocal connections (FRC)

This experiment highlights what happens to this physical property (FRC) as we remove more and more hubs, as in the previous experiment, from a network based on textual information. The graph of Fig. 17 shows the simultaneous behavior for the three different (5, 12 and 21) topics of Newspaper database. In this figure each curve corresponds to different complex network, the x-axis is the percentage of removed nodes, and the y-axis is the corresponding fraction of reciprocal connections (FRC).

For all curves, the FRC clearly decreases as more and more nodes are removed. This due to the frequency of hubs in the database. For a word to be highly connected with many others (a hub), it should appear often in the database. It means that the probability of words having reciprocity is proportionally high, since it can occur in different positions in a textual sentence. It suggest that hubs are nodes having higher reciprocal connections. As we remove such nodes we are decreasing the corresponding FRC.

The seemingly straight fall FRC lines shows that there is a considerable proportion of words with reciprocal connections on the database, regardless the amount of words. Combining this result with the previous experiment for Modularity, which suggest that removing hubs increases Modularity, we can speculate on the degree of flexibility of the order of words in a text. Languages that have a high degree of Modularity and high FRC tend to be less rigid and more flexible in their sentences. A comparison of languages under the light of these features is a future path that must be investigated.

6.6. Average connection density K_{den}

As previously stated, the average connection density ($0 \leq (K_{den}) \leq 1$) is a metric that correlates the number of edges in a network with the maximum possible. This property measures the sparsity of the studied network and a fully connected network with the same number of nodes. This experiment also measures the influence of hubs under this property.

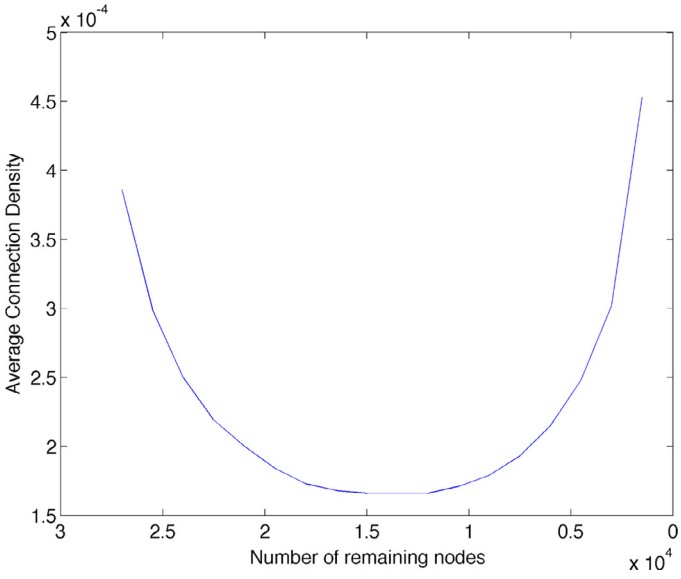


Fig. 20. Average connection density from the Newspaper complex network with 21 topics.

So, as in the previous experiment, we removed the hubs as the K_{den} is computed. Also, this removal is tested in three networks generated using the Newspaper database varying only the number of topics: 5, 12 and 21. Figs. 18–20 show the results.

As can be seen, all curves has the same behavior with a global minimum dividing the graph into a downward part and a growing one. To demonstrate the descending part, consider the graph of Fig. 21. In this graph, K_{den} can be obtained as the following:

$$K_{den} = 2 \cdot \frac{|E|}{|N|^2 - |N|}$$

$$K_{den} = 2 \cdot \frac{33}{20^2 - 20}$$

$$K_{den} = 0.17$$

Take now the highest degree nodes (3 and 9). Removing them from the network (Fig. 22) and calculating K_{den} at this stage, we note

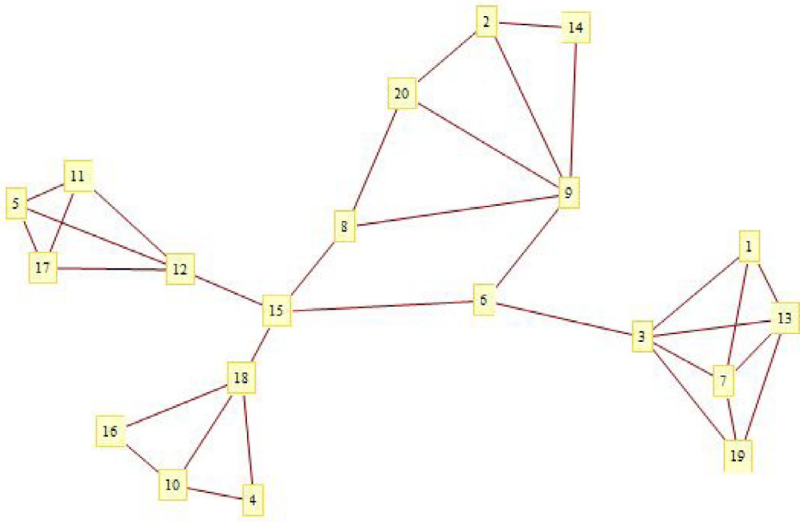


Fig. 21. Illustration for the descending part of the average connection density curve after removing the hubs.

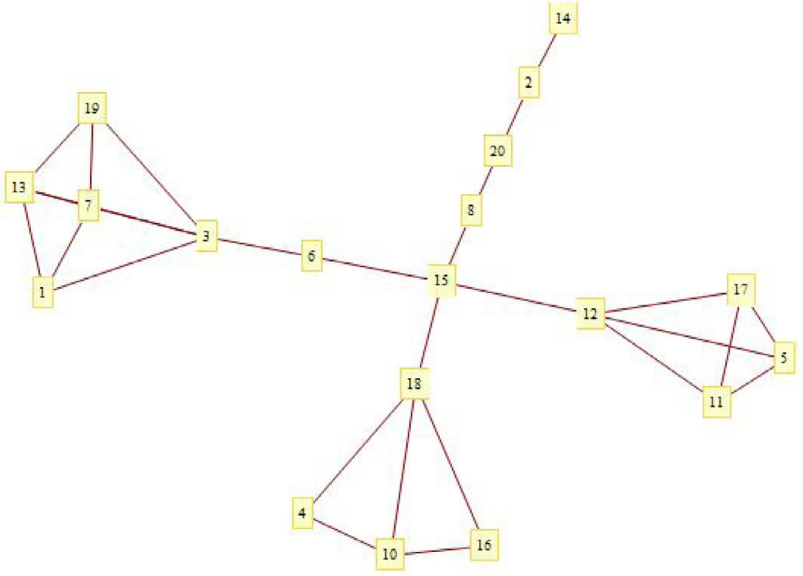


Fig. 22. Illustration for the ascending part of average connection density curve after removing the hubs.

that its value will be less than the previous:

$$K_{den} = 2 \cdot \frac{|E|}{|N|^2 - |N|}$$

$$K_{den} = 2 \cdot \frac{28}{19^2 - 19}$$

$$K_{den} = 0.16$$

This behavior occurs because the reduction rate in the numerator is smaller than the denominator. However, this value tends to increase as the hubs are removed. If we continue to remove them, in the extreme case, we reach only two nodes and one edge in the network, resulting in a $K_{den} = 1.0$. It explains the growing part of the graph.

This measure is independent of the number of topics in the Newspaper database, however it is highly sensitive to the number of hubs, achieving a value of equilibrium (global minimum). It suggests that, for a network with many hubs, the smaller the K_{den} value more sparser is the network. And for a network with small number of words and few hubs the K_{den} is higher.

For networks with high Modularity and number of words and a high value of K_{den}, one can speculate that there are a large number of contexts and that the role of the hubs is essential to connect these contexts. Languages with this feature tend to generate lots of sentences for the same semantic expression, being more flexible.

7. Conclusion

This paper studies some important properties of complex networks applied over textual information. These properties are: degree distribution (DD), weight distribution (WD), average clustering coefficient (AAC), Modularity, Fraction of reciprocal connections (FRC), and average connection density (ACD). We consider two types of textual information database: one from newspaper (called here Newspaper) and other from scientific medical journal (called here Ohsumed), building two distinct complex networks.

Also, we carried out our study from a standpoint of weighted networks, allowing a degree of flexibility in the relationship between the words in a text. This view is slightly different from those

presented so far in the literature. The drawback of this model is that it can generate denser networks which sometimes results in secondary problems, e.g., with physical properties extraction. In the specific case of average clustering coefficient (ACC), we proposed a simplified statistical approximation in order to compute it in a quadratic acceptable computational time. Our proposed statistical approach demonstrates to be feasible when computing this specific physical property since, in a sparse or a dense network of 10, 000 words, the error was up to 5%

The degree distribution (DD) showed that the behavior of these complex networks is similar to that of small-world networks, since they follow the power-law equation with γ close to those achieved in the literature for similar languages. In addition, it revealed that those words known as stop-words are those with the highest degree, the so called hubs. The identification of this class of words, regardless of the type of database (generic or specific), shows a direction for new studies and systems for those who wish to identify automatically and dynamically this type of words.

Another important experiment conducted in this work involved the Modularity. This experiment showed that the hubs are the nodes (stop-words) that link the clusters, which are similar to human language contexts.

Another result that suggests that the studied complex network seems to have the a small-world network behavior is the ACC value. Experiments showed that the ACC extracted from our networks are much greater than the expected ACC from a random network with same number of nodes and edges.

Besides the physical properties studied in this paper, there are many other properties that must be investigate in the future, such as: betweenness, average path length, diameter, degree of connectivity, and node centrality, to name a few. Perhaps one of major challenges to computing these properties is their computational complexity, which is mostly $O(n^3)$.

The studies in the linguistics area based on complex networks lead directly to the viability of important applications for today's problems. In addition to traditional textual information retrieval based on management and mining of large databases, as well as in storage and transmission of information between accessing points, it is also possible to predict the impact of this type of research in modern expert systems applications. Some of these applications have already been mentioned in this paper and have been studied by many researchers. Among them we can find: authorship recognition, machine translation, text summarization throughout semantic capture and community detection, and others.

The main current challenges are related to search semantic content in texts. A process close related to the level of human consciousness, still under study in the field of psychology and neuroscience. It is not possible with current technology developing mathematical and computational models able to interpret the full contents of a text in a highly abstracted level correlated with human consciousness. The task of semantic text matching needs comparisons not only in terms of significance of contents, but also in terms of evaluation of text quality with strong reflections and impacts in nowadays and future applications.

However, the actual findings of linguistics area may, now and in the future, in addition to improving current applications cited here, enable the development of expert systems capable of translating the semantic content of texts to the level of information abstraction. Such translation allows the semantic and rhetoric comparison of two texts, even in different languages and having few possible words in common. This is a direct application of context detection observed in clusters (or communities) of words. These kinds of pattern matching can improve the text comparison in various different languages which lead to improvement of machine translation area and other correlated applications.

Acknowledgment

The authors would like to thank the CNPq (Project 301858/2007-1) and CAPES (Project 094/2007), the Brazilian agencies for Scientific Financing, FAPESP (Sao Paulo Research Foundation), as well as to FEI (Ignacian Educational Foundation) a Brazilian Jesuit Faculty of Science Computing and Engineering, for the support of this work.

References

Albert, R., Jeong, H., & Barabasi, A. L. (1999). The diameter of the world wide web. *Nature, 401*, 130–131.
Allesina, S., & Pascual, M. (2008). Network structure, predator prey modules, and stability in large food webs. *Theoretical Ecology, 1*, 55–64.
Amancio, D. R. (2014). Aperspective on the advancement of natural language processing tasks via topological analysis of complex networks. *Physics of Life Reviews, 11*, 641–643.
Amancio, D. R., Nunes, M. G. V., Oliveira Jr., & Costa, L. F. (2014a). Extractive summarization using complex networks and syntatic dependency. *Physica A, 391*, 1855–1864.
Amancio, D. R., Oliveira Jr. Costa, L. F. (2014b). Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts.
Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Boston, M.A., USA: Addison-Wesley Longman Publishing Co., Inc.Isbn: 020139829X.
Biemann, C., Ross, S., & Weihe, K. (2012). Quantifying semantics using complex networks. In *Proceedings of the COLING 2012: tecnnical papers* (pp. 263–278).
Bota, A., & Kovacs, L. (2014). The community structure of word association graphs. In *Proceedings of the 9th international conference on applied informatics: Vol.1* (pp. pp.113–120). Eger, Hungary, January 29-Febrary 1, 2014.
Cancho, R. F., & Solao, R. V. (2001). The small world of human language, proceedings of the royal society of London. *Series B, Biological Sciences, 268*, 2261–2266.
Cech, R. (2014). On the interpretation of complex network analysis of language. *Physics of Life Reviews, 11*, 624–625.
Cech, R., Macutek, J., & Liu, H. (2014). *Complex networks and their applications, towards a theoretical framework for analyzing complex linguistic networks* (pp. 167–186). Springer-Verlag Berlin Heidelberg.
Chen, X. (2014). Language as a whole – A new framework for linguistic knowledge integration. *Physics of Life Reviews, 11*, 628–629.
Colman, E. R., & Rodgers, G. L. (2012). Kinects of node splitting in evolving complex networks. *Physica A, 391*, 6626–6631.
Cong, J., & Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews, 11*, 598–618.
Cong, T., Shuai, L., & Wu, Y. (2014). Extending network approach to language dynamics and human cognition. *Physics of Life Reviews, 11*, 639–640.
Deng, W., Li, W., Cai, X., & Wang, Q. (2011). The exponential degree distribution in complex networks: Non-equilibrium network theory, numerical simulation and empirical data. *Physica A, 390*, 1481–1485.
Erdős, P., & Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen, 6*, 290–297.
Ferrer-i Cancho, R. (2014). Beyond description. *Physics of Life Reviews, 11*, 621–623.
Gao, Y., Liang, W., Shi, Y., & Huang, Q. (2014). Comparison of directed co-ocurrence networks of six languages. *Physica A, 393*, 579–589.
Hart, J., & Parker, B. (1981). The peasants are revolting. *Wizard of Id Series*. Fawcet Gold Medal.
Hersh, W., Buckley, C., Leone, T. J., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '94* (pp. v192–201). Springer-Verlag, Dublin, Ireland.
Hudson, R. (2014). The view from linguistics. *Physics of Life Reviews, 11*, 619–620.
Kohler, R. (2014). Linguistic complex networks as a young field of quantitative linguistic. *Physics of Life Reviews, 11*, 630–631.
Li, J., & Zhou, J. (2007). Chinese character structure analysis based on complex networks. *Physica A, 380*, 629–638.
Macutek, J. (2014). Complex networks are not (so much) privileged. *Physics of Life Reviews, 11*, 635–636.
Mehri, A., Darooneh, A. H., & Shariati, A. (2012). The complex networks approach for authorship attribution of books. *Physica A, 391*, 2429–2437.
Mishkovski, I., Biey, M., & Kocarev, L. (2011). Vulnerability of complex networks. *Commun Nonlinear Sci Numer Simulat, 16*, 341–349.
Newman, M. E., Barabasi, A.-L., & Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton University Press.
Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *American Physical Society, 69*(2), 026113.
Pardo, T. A. S., Antiqueira, L., Nunes, M. G. V., Oliveira, & Costa, L. F. (2005). Using complex networks for language processing: the case of summary evaluation. In *Proceedings of the international conference on communications, circuits and systems* (pp. 2678–2682).
Pons, P., & Latapy, M. (2004). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications, 10*, 284–293.
Price, D. S. (1965). Network of scientific papers. *Science, 149*, 510–515.
Réka, A., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 1*(74), 47–97.

Schank, T., & Wagner, D. (2005). Approximating clustering-coefficient and transitivity. *Journal of Graph Algorithms and Applications, 9*(2), 265–275.
Sheng, L., & Li, C. (2009). English and Chinese languages as weighted complex networks. *Physica A, 388*, 2561–2570.
Solomonoff, R., & Rapoport, A. (1951). Connectivity of random nets. *Bulletin of Mathematical Biophysics, 13*, 107–117.
Sporns, O. (2003). Graph theory methods for the analysis of neural connectivity patterns. In R. Kötter (Ed.), *Neuroscience Databases* (pp. 171–185). Springer, US.
Wachs-Lopes, G. A. (2011). *Um modelo de redes complexas para análise de informações textuais*. Ignatian Educational Foundation (FEI) Master thesis.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature, 6684*(393), 440–442.
Yu, S., & Xu, C. (2014). Properties of language networks and language systems. *Physics of Life Reviews, 11*, 626–627.
Zhao, Y. (2014). Three lines to view language network. *Physics of Life Reviews, 11*, 637–638.