Distinguishing between Positive and Negative Opinions with Complex Network Features

Diego R. Amancio, Renato Fabbri, Osvaldo N. Oliveira Jr.,
Maria G. V. Nunes and Luciano da F. Costa
University of São Paulo, São Carlos, São Paulo, Brazil
diego.amancio@usp.br, renato.fabbri@gmail.com, chu@ifsc.usp.br,
gracan@icmc.usp.br, ldfcosta@gmail.com

Abstract

Topological and dynamic features of complex networks have proven to be suitable for capturing text characteristics in recent years, with various applications in natural language processing. In this article we show that texts with positive and negative opinions can be distinguished from each other when represented as complex networks. The distinction was possible by obtaining several metrics of the networks, including the in-degree, out-degree, shortest paths, clustering coefficient, betweenness and global efficiency. For visualization, the obtained multidimensional dataset was projected into a 2-dimensional space with the canonical variable analysis. The distinction was quantified using machine learning algorithms, which allowed an recall of 70% in the automatic discrimination for the negative opinions, even without attempts to optimize the pattern recognition process.

1 Introduction

The use of statistical methods is well established for a number of natural language processing tasks (Manning and Schuetze, 2007), in some cases combined with a deep linguistic treatment in hybrid approaches. Representing text as graphs (Antiqueira et al., 2007), in particular, has become popular with the advent of complex networks (CN) (Newman, 2003; Albert and Barabasi, 2002), especially after it was shown that large pieces of text generate scale-free networks (Ferrer i Cancho and Sole, 2001; Barabasi, 2009). This scale-free nature of such networks is probably the main reason why complex networks concepts are capable of capturing features of text, even in the absence of any linguistic treatment. Significantly, the scale-free property has also allowed CN to be applied in diverse fields (Costa et al., 2008), from neuroscience (Sporns, 2002) to physics (Gfeller, 2007), from linguistics (Dorogovtsev and Mendes, 2001) to computer science (Moura et al., 2003), to mention a few areas. Other frequently observed unifying principles that natural networks exhibit are short paths between any two nodes and high clustering coefficients (i.e. the so-called small-world property), correlations in node degrees, and a large number of cycles or specific motifs.

The topology and the dynamics of CN can be exploited in natural language processing, which has led to several contributions in the literature. For instance, metrics of CN have been used to assess the quality of written essays by high school students (Antiqueira et al., 2007). Furthermore, degrees, shortest paths and other metrics of CN were used to produce strategies for automatic summarization (Antiqueira et al., 2009), whose results are among the best for methods that only employ statistics. The quality of machine translation systems can be examined using local mappings of local measures (Amancio et al., 2008). Other related applications include lexical resources analysis (Sigman and Cecchi, 2002), human-induced words association (Costa, 2004), language evolution (Dorogovtsev and Mendes, 2002), and authorship recognition (Antiqueira et al., 2006).

In this paper, we model texts as complex networks with each word being represented by a node and co-occurrences of words defining the edges (see next section). Unlike traditional methods of text mining and sentiment detection of reviews (Tang et al., 2009; Pennebaker et al., 2003), the method described here only takes into account the relationships between concepts, regardless of the semantics related to each word. Specifically, we analyze the topology of the networks in order to distinguish between texts with positive and negative opinions. Using a corpus of 290 pieces of

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 83–87,
Uppsala, Sweden, 16 July 2010. ©2010 Association for Computational Linguistics

Before pre-processing	*After pre-processing*
The projection of the network data into two dimensions is crucial for big networks	projection network data two dimension be crucial big network

Table 1: *Adjacency list obtained from the sentence "The projection of the network data into two dimensions is crucial for big networks".*

text with half of positive opinions, we show that the network features allows one to achieve a reasonable distinction.

2 Methodology

2.1 Representing texts as complex networks

Texts are modeled as complex networks here by considering each word (concept) as a node and establishing links by co-occurrence of words, disregarding the punctuation. In selecting the nodes, the stopwords were removed and the remaining words were lemmatized to combine words with the same canonical form but different inflections into a single node. Additionally, the texts were labeled using the MXPost part-of-speech Tagger based on the Ratnaparki's model (Ratnaparki, 1996), which helps to resolve problems of ambiguity. This is useful because the words with the same canonical form and same meaning are grouped into a single node, while words that have the same canonical form but distinct meanings generate distinct nodes. This pre-processing is done by accessing a computational lexicon, where each word has an associated rule for the generation of the canonical form. For illustrative means, Table 1 shows the pre-processed form of the sentence "The projection of the network data into two dimensions is crucial for big networks" and Figure 1 shows the network obtained for the same sentence.

Several CN metrics have been used to analyze textual characteristics, the most common of which are out-degree (k_{out}), in-degree (k_{in}), cluster coefficient (C) and shortest paths (l). Here we also use the betweenness (ϱ) and the global efficiency (η). The out-degree corresponds to the number of edges emanating from a given node, where the weight of each link between any two nodes may also be considered, being referred to as out-strength. Analogously, the node's in-degree is defined as the number of edges arriving at a given node. The network's k_{out} and k_{in} are evaluated by calculating the average among all the nodes, note that such global measures k_{out} and k_{in} are always equal. Regarding the adjacency matrix to represent the network, for a given node i, its k_{out} and k_{in} are calculated by eqs 1 and 2, where N represents the number of distinct words in the pre-processed text:

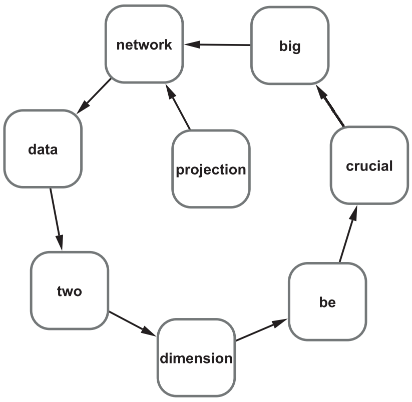

Figure 1: *Network obtained from the sentence "The projection of the network data into two dimensions is crucial for big networks".*

$$k_{out}(i) = \sum_{j=1}^{N} W_{ji} \tag{1}$$

$$k_{in}(i) = \sum_{j=1}^{N} W_{ij} \tag{2}$$

The cluster coefficient (C) is defined as follows. Let S be the set formed by nodes receiving edges of a given node i, and N_c is the cardinality of this set. If the nodes of this set form a completely connected set, then there are N_c(N_c-1) edges in this sub graph. However, if there are only B edges, then the coefficient is given by eq. (3):

$$C(i) = \frac{B}{N_c(N_c - 1)} \tag{3}$$

If N_c is less than 1, then C is defined as zero. Note that this measure quantifies how the nodes connected to a specific node are linked to each other, with its value varying between zero and one.

The shortest paths are calculated from all pairs of nodes within the network. Let d_{ij} be the minimum distance between any two words i and j in the network. The shortest path length l of a node i is given in equation 4.

$$l(i) = \frac{1}{N-1} \sum_{j \neq i} d_{ij} \quad (4)$$

Another measure often used in network analysis is the global efficiency (η), which is defined in equation 5, and may be interpreted as the speed with which information is exchanged between any two nodes, since a short distance d_{ij} contributes more significantly than a long distance. Note that the formula below prevents divergence; therefore, it is especially useful for networks with two or more components. The inverse of η, named *harmonic mean of geodesic distances*, has also been used to characterize complex networks.

$$\eta = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \quad (5)$$

While l and η use the length of shortest paths, the betweenness uses the number of shortest paths. Formally, the betweenness centrality for a given vertex v is given in equation 6, where the numerator represents the number of shortest paths passing through the vertices i, v and j and the denominator represents the number of shortest paths passing through the vertices i and j. In other words, if there are many shortest paths passing through a given node, this node will receive a high betweenness centrality.

$$\varrho(v) = \sum_{i} \sum_{j} \frac{\sigma(i,v,j)}{\sigma(i,j)} \quad (6)$$

2.2 Corpus

The corpus used in the experiments was obtained from the Brazilian newspaper Folha de São Paulo[1], from which we selected 290 articles over a 10-year period from a special section where a positive opinion is confronted with a negative opinion about a given topic. For this study, we selected the 145 longest texts with positive opinion and the 145 longest text with negative opinions[2], in order to have meaningful statistical data for the CN analysis.

2.3 Machine Learning Methods

In order to discriminate the topological features from distinct networks we first applied a technique for reducing the dimension of the dataset, the canonical variable analysis (McLachlan, 2004). The projection of network data into a lower dimension is crucial for visualization, in addition to avoids the so-called "curse of dimensionality" (Bishop, 2006). To calculate the axes points for projecting the data, a criterion must be established with which the distances between data points are defined. Let S be the overall dispersion of the measurements, as shown in equation 7, where ζ is the number of instances (ζ = 290), $\overrightarrow{x_c}$ is the set of metrics for a particular instance and $\langle \overrightarrow{x} \rangle$ is the average of all $\overrightarrow{x_c}$.

$$S = \sum_{c=1}^{\zeta} \left(\overrightarrow{x_c} - \langle \overrightarrow{x} \rangle \right) \left(\overrightarrow{x_c} - \langle \overrightarrow{x} \rangle \right)^T \quad (7)$$

Considering that two classes (C_1 = positive opinions and C_2 = negative opinions) are used, the scatter matrix S_i is obtained for each class C_i, according to equation 8, where $\langle \overrightarrow{x} \rangle_i$ is the analogous of $\langle \overrightarrow{x} \rangle$ when only the instances belonging to class C_i is taken into account.

$$S_i = \sum_{c \in C_i} \left(\overrightarrow{x_c} - \langle \overrightarrow{x} \rangle_i \right) \left(\overrightarrow{x_c} - \langle \overrightarrow{x} \rangle_i \right)^T \quad (8)$$

The intraclass matrix, i.e. the matrix that gives the dispersion inside C_1 and C_2, is defined as in equation 9. Additionally, we define the interclass matrix, i.e. the matrix that provides the dispersion between C_1 and C_2, as shown in equation 10.

$$S_{intra} = S_1 + S_2 \quad (9)$$

$$S_{inter} = S - S_{intra} \quad (10)$$

The principal axes for the projection are then obtained by computing the eigenvector associated with the largest eigenvalues of the matrix Λ (McLachlan, 2004) defined in equation 11. Since the data were projected in a two-dimensional space, the two principal axes were selected, corresponding to the two largest eigenvalues.

$$\Lambda = S_{intra}^{-1} S_{inter} \quad (11)$$

Finally, to quantify the efficiency of separation with the projection using canonical variable analysis, we implemented three machine learning algorithms (decision tree, using the C4.5 algorithm (Quinlan, 1993); rules of decision, using the

[1] http://www.folha.com.br

[2] The average size of the selected corpus is 600 words.

RIP algorithm (Cohen, 1995), and Naive Bayes algorithm (John and Langley, 1995)) and evaluated the accuracy rate using the 10-fold-cross-validation (Kohavi, 1995).

3 Results and Discussion

The metrics out-degree (k_{out}), in-degree (k_{in}), shortest paths (l), cluster coefficient (C), betweenness (ϱ) and global efficiency (η) were computed for each of the 145 texts for positive and negative opinions, as described in the Methodology. The mean values and the standard deviations of these metrics were used as attributes for each text. This generated a dataset described in 10 attributes, since the average k_{in} is equal to the average k_{out} and the standard deviation of η is not defined (in other words, it is always zero). Figure 2 shows the projection of the dataset obtained with canonical variable analysis, illustrating that texts with different opinions can be distinguished to a certain extent. That is to say, the topological features of networks representing positive opinion tend to differ from those of texts with negative opinion.

The efficiency of this methodology for characterizing different opinions can be quantified using machine learning algorithms to process the data from the projection. The results are illustrated in Table 2. Again, the distinction between classes is reasonably good, since the accuracy rate reached 62%. Indeed, this rate seems to be a good result, since the baseline method[3] tested showed an accuracy rate of 53%. One also should highlight the coverage found for the class of negative reviews by using the C4.5 algorithm, for which a value of 82% (result not shown in the Table 2) was obtained. This means that if an opinion is negative, the probability of being classified as negative is only 18%. Thus, our method seems especially useful when a negative view should be classified correctly.

Method	*Correctly classified*
C4.5	58%
Rip	60%
Naive Bayes	62%

Table 2: *Percentage of correctly classified instances.*

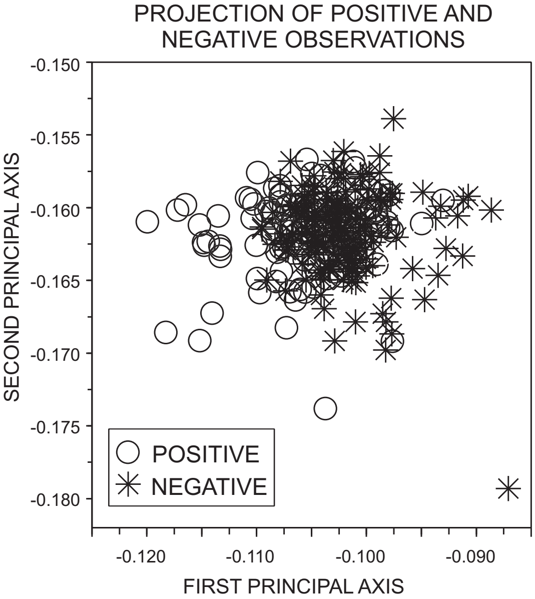

Figure 2: *Projection obtained by using the method of canonical variables. A reasonable distinction could be achieved between positive and negative opinions.*

4 Conclusion and Further Work

The topological features of complex networks generated with texts appear to be efficient in distinguishing between attitudes, as indicated here where texts conveying positive opinions could be distinguished from those of negative opinions. The metrics of the CN combined with a projection technique allowed a reasonable separation of the two types of text, and this was confirmed with machine learning algorithms. An 62% accuracy was achieved (the baseline reached 53%), even though there was no attempt to optimize the metrics or the methods of analysis. These promising results are motivation to evaluate other types of subtleties in texts, including emotional states, which is presently being performed in our group.

Acknowledgements: Luciano da F. Costa is grateful to FAPESP (05/00587-5) and CNPq (301303/06-1 and 573583/2008-0) for the financial support. Diego R. Amancio is grateful to FAPESP sponsorship (proc. 09/02941-1) and Renato Fabbri is grateful to CAPES sponsorship. We also thank Dr. Oto Araujo Vale very much for supplying the corpus.

[3] The baseline method used as attributes the frequency of each word in each text. Then, the algorithm C4.5 was run with the same parameters used for the methodology based on complex networks.

References

C. D. Manning and H. Schuetze. 1999. Foundations of Statistical Natural Language Processing. *The MIT Press*, First Edition.

L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr. and L. da F. Costa. 2007. Strong correlations between text quality and complex networks features. *Physica A*, 373:811–820.

M. E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Review*, 45:167–256.

R. Z. Albert and A.L. Barabasi. 2002. Statistical Mechanics of Complex Networks. *Rev. Modern Phys.*, 74:47–97.

R. Ferrer i Cancho and R. V. Sole. 2001. The small world of human language. *Proceedings of the Royal Society of London B*, 268:2261.

A.L. Barabasi. 2009. Scale-Free Networks: a decade and beyond. *Science*, 24 325 5939 412–413.

L. F. da Costa, O. N. Oliveira Jr., G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, L. E. C. da Rocha. 2008. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. *arXiv* 0711.3199.

O. Sporns. 2002. Network analysis, complexity, and brain function. *Complexity*, 8(1):56–60.

D. Gfeller, P. LosRios, A. Caflisch and F. Rao. 2007. Complex network analysis of free-energy landscapes. *Proceedings of the National Academy of Science USA*, 104 (6):1817–1822

S. N. Dorogovtsev and J. F. F.Mendes. 2001. Language as an evolving word web. *Proceedings of the Royal Society of London B*, 268:2603.

A. P. S. de Moura, Y. C. Lai and A. E. Motter. 2003. Signatures of small-world and scale-free properties in large computer programs. *Physical Review E*, 68(1):017102.

L. Antiqueira, O. N. Oliveira Jr., L. da F. Costa and M. G. V. Nunes. 2009. A Complex Network Approach to Text Summarization. *Information Sciences*, 179:(5) 584–599.

M. Sigman and G.A. Cecchi. 2002. Global Organization of the Wordnet Lexicon. *Proceedings of the National Academy of Sciences*, 99:1742–1747.

L. F. Costa. 2004. What's in a name ? *International Journal of Modern Physics C*, 15:371–379.

S. N. Dorogovtsev and J. F. F. Mendes. 2002. Evolution of networks. *Advances in Physics*, 51:1079–1187.

L. Antiqueira, T. A. S. Pardo, M. G. V. Nunes, O. N. Oliveira Jr. and L. F. Costa. 2006. Some issues on complex networks for author characterization. *Proceeedings of the Workshop in Information and Human Language Technology*.

H. Tang, S. Tan and X. Cheng. 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36:7 10760–10773.

J. W. Pennebaker, M. R. Mehl and K. G. Niederhoffer. 2003. Psychological aspects of natural language. use: our words, our selves. *Annual review of psychology*, 54 547-77.

D. R. Amancio, L. Antiqueira, T. A. S. Pardo, L. F. Costa, O. N. Oliveira Jr. and M. G. V. Nunes. 2008. Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, 19(4):583-598.

A. Ratnaparki. 1996. A Maximum Entropy Part-Of-Speech Tagger. *Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania.*

G. J. McLachlan. 2004. Discriminant Analysis and Statistical Pattern Recognition. *Wiley.*

C. M. Bishop. 2006. Pattern Recognition and Machine Learning. *Springer-Verlag New York.*

R. Quinlan. 1993. C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers.*

W. W. Cohen. 1995. Fast Effective Rule Induction. *12 International converence on Machine Learning*, 115–223.

G. H. John and P. Langley. 1995. Estimating Continuous Distribution in Bayesian Classifiers. *11 Conference on Uncertainty in Artificial Intelligence*, 338–345.

R. Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2*, 12:1137-1143.