Physica A 495 (2018) 49–58

Contents lists available at ScienceDirect

Physica A

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

On the role of words in the network structure of texts: Application to authorship attribution

Camilo Akimushkin [a], Diego R. Amancio [b,*], Osvaldo N. Oliveira Jr. [a]

[a] São Carlos Institute of Physics, University of São Paulo, Avenida Trabalhador São-carlense 400, São Carlos, São Paulo, Brazil
[b] Institute of Mathematics and Computer Science, University of São Paulo, Avenida Trabalhador São-carlense 400, São Carlos, São Paulo, Brazil

HIGHLIGHTS

- The authorship is addressed by combining network metrics and vocabulary usage.
- Strong evidence of the influence of authors on the network properties of specific words have been found.
- A generalized similarity measure based on word ranks is proposed.
- Multi-Dimensional Scaling allows the simultaneous implementation of dissimilarity matrices.
- The results are robust among different collections composed of various authors.

ARTICLE INFO

Article history:
Received 15 May 2017
Received in revised form 18 October 2017
Available online 19 December 2017

Keywords:
Complex networks
Word semantics
Authorship attribution
Similarity measures
Burstiness
Intermittency

ABSTRACT

Well-established automatic analyses of texts mainly consider frequencies of linguistic units, e.g. letters, words, and bigrams. In a recent, alternative approach, medium and large-scale text structures were used in opposition to the belief that text structure is dominated by the language features. In this paper, we introduce a generalized similarity measure to compare texts which accounts for both the network structure of texts and the role of individual words in the networks. The similarity measure is used for authorship attribution of three collections of books, each composed of 8 authors and 10 books per author. High accuracy rates were obtained with typical values between 90% and 98.75%, much higher than with the traditional term frequency-inverse document frequency (tf-idf) approach for the same collections. These accuracies are also higher than those obtained solely with the topology of networks. We conclude that the different properties of specific words on the macroscopic scale structure of a whole text are as relevant as their frequency of appearance; conversely, considering the identity of nodes brings further knowledge about a piece of text represented as a network.

1. Introduction

The huge volume of written text produced everyday makes it imperative to use automatic tools to retrieve relevant information, e.g. with text summarization, polarity analysis, citation analysis, and document classification [1–6]. An essential step in many of these tasks is to compare pieces of texts, as in classification of texts into categories [3] and in search engines where typically a list of texts relevant to a given query is retrieved. A special case is the pairwise comparison,

* Corresponding author.
E-mail addresses: diego.raphael@gmail.com, diego@icmc.usp.br (D.R. Amancio).

https://doi.org/10.1016/j.physa.2017.12.054
0378-4371/

where one searches for similarities between pairs of texts, which is actually a typical subtask in the authorship attribution process [7]. Automatic authorship attribution has been made with varied strategies [8], from the use of first-order statistics of linguistic elements to the processing of text represented as networks [9,10]. For example, the frequency of characters [11,12], phonemes [13], and morphemes [14,15] has been explored, with texts normally modeled as lists of individual words, i.e. word order is disregarded. The archetype of such models is the so-called bag-of-words (BoW) model [16], where the text is represented as the set of its constitutive words by counting the number of appearances for each word. Word frequencies, which follow Zipf's law [17,18], can then be used straightforwardly as attributes in a machine learning scheme [19] or to further build specific similarity measures.

Variations of the BoW model have been developed to address possible biases, e.g. the tendency of larger texts of being more likely to be considered similar to any other. These variations include the use of the term frequency-inverse document frequency (tf-idf) statistic [20,3], where lower relevance is assigned to words frequent in the document as well as in the whole collection. The model has also been modified to incorporate other kinds of data, such as in the bag-of-features model used for image analysis [21]. Another important modification is to consider n-grams, i.e., groups of n adjacent words [22,23], in an attempt to take syntactic information into account, since the BoW model disregards word ordering. In other types of work, the syntactic roles of the words in sentences are used for authorship attribution [24,25]. It must be noted, nevertheless, that all of these approaches are based on the counting of features, even if some consider small-scale structural relationships.

An alternative perspective has been developed in recent years from the discovery that language features may be best described by complex network models [26–33]. The structure of a text, for instance, can be mapped onto a co-occurrence network [9], which is characterized by power-law distributions [18,34], and core–periphery structures [35]. Even though the general features of these complex networks remain analogous for texts in the same language, the network representation can also be used for classification tasks, particularly for authorship attribution [9,36,37].

While frequency-based methods overlook all structural relationships among words farther than in the same sentence, several methods based on co-occurrence networks ignore the identity of the words (i.e. which actual word corresponds to a given node), thus characterizing the texts only on the basis of the network topology. In this study, we reconcile both viewpoints to show that, from a network perspective, words can play relevant roles in the structure of a text besides their frequencies.

2. Related works

The authorship attribution problem has been extensively studied along the last few years. Several features have been found informative in the context of identifying the subtleties in authors' style. Among the most used attributes are those relying on the following resource types [8]:

- *Lexical* and *character* features: word length, word n-grams, vocabulary richness, character n-grams, sentence lengths, etc.
- *Syntactical* features: phrase structure, part-of-speech, etc.
- *Semantical* features: number of synonyms and semantical dependencies, etc.

More recently, additional statistical methods have been proposed, including those based on complex networks. Complex networks have been used to model and study many linguistic phenomena, such as complexity [38–41], semantics [2] and citations [42]. In the authorship attribution context, approaches were used to examine the global properties of word adjacency networks, without taking into account the labels of nodes after building the underlying networks. In [43], the authors use the labels of words to analyze local network features of word adjacency networks, where the selected words as features are the most frequent ones. Because some network features depend on network size, only texts of same length were considered in the analysis. In [41], the authors focused on the analysis of transition matrices between function words. Therefore, the label of the nodes was also used for comparison purposes. As we shall show, our method also uses a local approach based on node labels. However, differently from previous approaches, the ranking of words (according to a given network measurement) is used as a feature selector. In addition, the absolute values of network measurements are not used as features: our method relies on the relative importance of words given by word rankings. This is an advantage of our method because word rankings are less sensitive to text length than traditional network measurements [44]. The absolute values of network measurements are only used to rank the importance of words. Our methodology also differs from previous ones since we used the multidimensional scaling (MDS) technique to optimize the performance. To our knowledge, such projection method has not been used to network text analysis for stylometry purposes. Furthermore, the ranking strategy allowed us to use only a limited number of measurements. This is in contrast to the work in [43], where a much larger number of topological measurements was used to yield optimized results. The method we propose is explained in detail in the next section.

3. Methods

The methodology used here to address the authorship attribution task consists of four steps: (i) construct a co-occurrence network for each text; (ii) obtain various dissimilarity matrices for the collection using the proposed similarity metrics (see below); (iii) join the various dissimilarity matrices with multi-dimensional scaling [45]; and (iv) analyze the resulting data with standard supervised learning algorithms [19]. These steps are described in detail below. The model was applied to three

collections of 80 literary texts. Each collection contains 10 texts per author for 8 authors from the 19th century, with 22 of the 24 authors being native English writers (details of the collections are included in the Supporting Information).

3.1. Network construction and characterization

Texts are pre-processed for constructing the networks, with stopwords, such as articles and prepositions, being removed, and lemmatization being applied to reduce different forms to a common base form. Lemmatization is assisted by a part-of-speech tagger based on entropy maximization [46], in order to solve ambiguities in mapping words to their lemmatized form. It must be noted that throughout the text we refer to "words", even though a more precise term is "lemmas" owing to the preprocessing done, i.e. the observed multiplicity of word roles is not caused by word polysemy.

From the resulting pre-processed text, a co-occurrence (or word adjacency) network is built, where each distinct word is a node and two nodes are connected if the words appear consecutively in the text. The link is directed according to the natural reading order. For instance, the title of this paper generates the network: role → word → network → structure → text → application → author → attribution. Each link has a default weight equal to one, which is increased by one unit each time the pair of words appears again in the text.

Networks were characterized in this study by four well-known node-local metrics:

1. Degree (k_i): this metric corresponds to the number of links attached to a node. As a consequence of the construction rules imposed by co-occurrence networks, there is strong correlation between this metric and the word frequency.
2. Average shortest path length (l_i): this is the typical distance between two nodes of the network, given by:

$$l_i = N^{-1} \Sigma_j d_{ij}, \tag{1}$$

   where d_{ij} is the shortest path length between nodes i and j, and N is the number of nodes. This metric is useful to identify keywords in written texts, irrespectively of the word frequency [47]. Low values of l are not only associated to the frequent words, but also to the words appearing close to other relevant words in the text.
3. Betweenness centrality (B_i): the betweenness is the fraction of all shortest paths that pass through the node, i.e.

$$B_i = \sum_{i \neq j \neq k} \frac{n_{jk}^{(i)}}{n_{jk}}, \tag{2}$$

   where $n_{jk}^{(i)}$ is the number of shortest paths from j to k passing through i and n_{jk} is the total number of shortest paths from j to k. In text analysis, the betweenness can be interpreted as a measure to quantify the ability of a word to appear in restricted or wider contexts [9].
4. Intermittency (I_i): the intermittency is a measure that quantifies the spatial distribution of a given word along a text. To define this measure, consider the text as a sequence of tokens. This sequence generates, for each word i, a time series $T^{(i)} = \{t_1^{(i)}, t_2^{(i)}, \ldots, t_{f_i}^{(i)}\}$, where $t_j^{(i)}$ corresponds to the position of the jth occurrence of the word i. The interval recurrence (τ) for word i is defined as the spatial difference between two occurrences, i.e. $\tau_j^{(i)} = t_j^{(i)} - t_{j-1}^{(i)}$. The set of all values of $\tau_j^{(i)}$, i.e. $\mathcal{T}^{(i)} = \{\tau_1^{(i)}, \tau_2^{(i)}, \ldots\}$ is used to quantify the regularity of the appearance of i along the sequence of tokens. More specifically, this regularity is computed using the intermittency defined as:

$$I_i = \sigma_{\mathcal{T}} / \langle \mathcal{T} \rangle = \left[\frac{\langle \mathcal{T}^2 \rangle}{\langle \mathcal{T} \rangle^2} - 1 \right]^{1/2}, \tag{3}$$

   where $\sigma_{\mathcal{T}}$ and $\langle \mathcal{T} \rangle$ are the standard deviation and average of $\mathcal{T}$, respectively. In text networks, the intermittency measures the words most related to the subject being addressed [47].

We have decided only to use a small set of network measurements to show, as a proof principle, that a simple set of topological features is able to detect stylistic subtleties in texts. The clustering coefficient was not used because, in previous works, we found that both clustering coefficient and betweenness measure similar linguistic properties [9].

The time complexity of the proposed method is most affected by the time complexity of computing the most costly network measurements, namely the betweenness and the shortest path lengths. The time complexity of computing shortest paths between all pairs of nodes is $\Theta(N^3)$. In an unweighted and sparse graph, the computation of shortest paths can be optimized with the Brandes' algorithm [48]. In this case, the computation takes $\mathcal{O}(N \times E)$ steps, where E is the total number of edges.

3.2. Similarity metrics

The novelty introduced in this work is to compare the words representing the most relevant nodes in the network topology, in contrast to previous approaches where only the statistics of topological metrics were taken into account [9]. We consider as the most relevant the nodes possessing the highest degree and betweenness. As for the other metrics, namely average shortest paths and intermittency, we chose the nodes with lowest values. We tested the largest shortest paths but

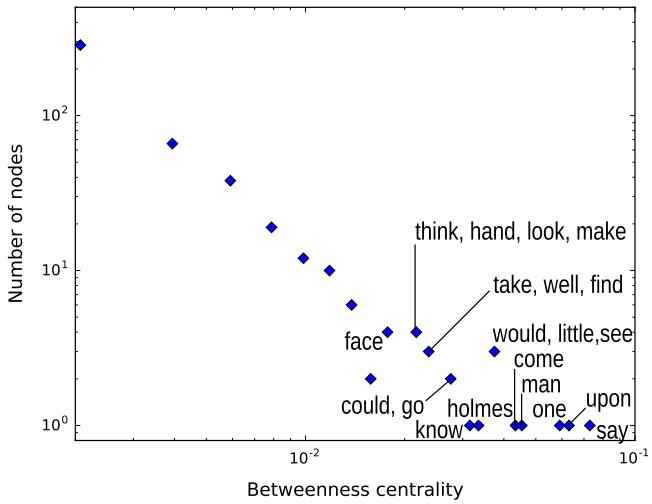


Fig. 1. Betweenness centrality distribution for "The Memoirs of Sherlock Holmes". The top 20 nodes are labeled by the corresponding words.

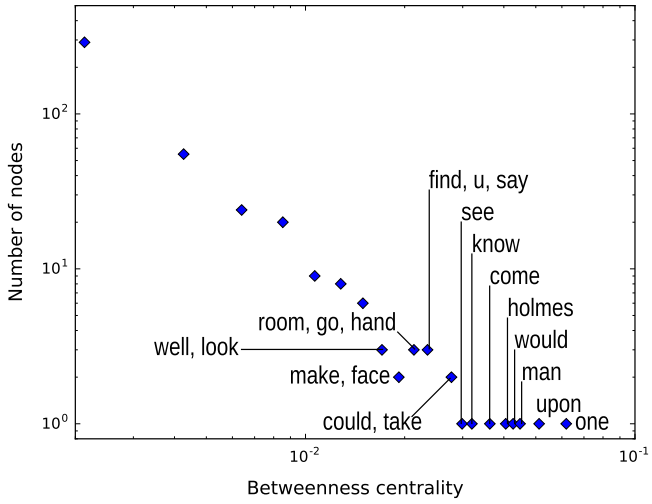


Fig. 2. Betweenness centrality distribution for "The Return of Sherlock Holmes". The top 20 nodes are labeled with the corresponding words.

the results were not as good. Intermittency obeys a power law with positive exponent. We hypothesize that two pieces of text will be similar if there is significant overlap in the words (nodes) considered most relevant in both texts.

Figs. 1 and 2 show the distributions of betweenness centralities for two books from Arthur Conan Doyle: The Memoirs of Sherlock Holmes and The Return of Sherlock Holmes, which could be expected to be similar since these books were written by the same author in the same series of novels. In both figures, the highest centralities belong to the same words (18 out of 20) which also occupy almost the same relative positions. We shall therefore test the hypothesis that not only the frequency of usage but also the long-scale topological metrics of the organization of words may be reliable signatures of authorship.

To quantify this we introduce a similarity measure between pairs of texts as follows: for each network metric considered we give a rank R to each word w from a subset V of top words with unique properties. In our approach the importance of a word depends on the particular network metric considered. As mentioned above, we select the words with the highest connectivity and betweenness centrality, and with the smallest values of shortest path and intermittency. For these two latter metrics the smallest values were chosen because their distributions present power laws with positive coefficients. We choose sets of 100 top words since in subsidiary experiments we observed that the interval between 50 and 150 words gave the best results. A ranking is assigned to each word, starting with the maximum value (100 in this case) for the word with the most extreme value (e.g. "say" for the betweenness centrality of "The Memoirs of Sherlock Holmes"), and decreasing in one unit for each consecutive word until reaching the last of the top words which receives a ranking value of one. With these

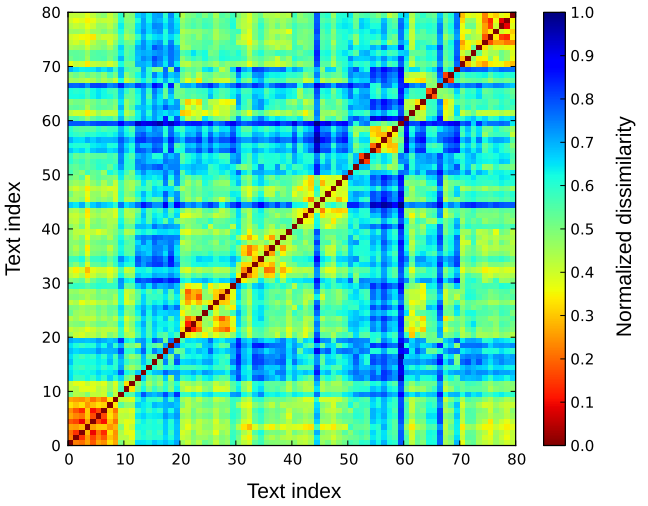


Fig. 3. Betweenness centrality dissimilarity matrix for the second collection, each decade corresponds to texts from the same author.

rankings, the similarity between two texts A and B for a given network metric is given by

$$A \cdot B = \sum_{w \in V_A \cap V_B} R_A(w) R_B(w), \tag{4}$$

that is, if a word is present in the top words subsets of both texts, the product of its rankings adds to their similarity.

This similarity metric is guaranteed to be high only if the same words occupy similar positions in the distributions such as in Figs. 1 and 2, with higher influence from the highest-ranked words. Eq. (4) implies that the norm or similarity of a text with itself is always the same, that is, $A \cdot A = B \cdot B = n(n+1)(2n+1)/6$, where n is the size of V. We therefore normalize all similarities for this value to be one and the minimum value to be zero, and define the dissimilarity D_{AB} between two texts as being one minus this normalized value. It is worth noting that other similarity metrics could be used to compare texts, but the dot product adopted here appears to be the most straightforward, as it is done in bag-of-words methods [3].

With all the values D_{AB} we produce a dissimilarity matrix for each metric. The dissimilarity matrix for the betweenness centrality of one of the collections is shown in Fig. 3, where the indices 0 to 9 correspond to texts from the first author, texts 10 to 19 to the second author and so on. Note that, in general, texts from the same author appear to be closer among themselves compared to texts from different authors even if they are relatively separated (e.g. texts 10–19 and 50–59).

3.3. Combining dissimilarity matrices

One strength of the approach is the ability of observing different aspects of the network structure simultaneously. Each metric yields a different dissimilarity matrix; hence, we can observe the similarity between texts at different scales. We now combine information from the distinct metrics in order to have useful data for the classification algorithms.

In this study we employed two strategies for the input into the classification algorithms. In the first, we simply used the whole of the dissimilarity matrices for the different metrics, i.e. with dissimilarities as attributes. In the second strategy, we reduced the dimensionality of the dissimilarity metrics with Multi-dimensional scaling (MDS) [49], with the aim of capturing the highest similarities while eliminating possible unnecessary information that may harm the classification task. MDS was conceived to map dissimilarities into positions in a space so that the distances between these positions reproduce as well as possible the original input dissimilarities. The space obtained is usually intended to have a small dimensionality and the algorithm is largely used for visualization purposes. The positions obtained when applying the algorithm to map one of the dissimilarity matrices to a two-dimensional space is presented in Fig. 4. Rather than using dissimilarity matrices, here we obtained a more clear visualization of similarities between same-authors texts. It must be noted that Fig. 4 is an approximate representation of the original data: not only does the distance among points differ from the dissimilarity matrix but the inclusion of more dimensions could also reveal differences from the apparent dissimilarities of the bidimensional case. Nevertheless, it is clear that already for this case same-author texts are in general closer to each other.

We use MDS to map the four dissimilarity matrices of each collection into four subspaces and then join these subspaces into a space of bigger dimension: if we write the positions in each subspace as a matrix $M \times N_i$ where M is the number of points (80 texts per collection in our case) and N_i is the dimensionality of the subspace, then the positions in the composed space are given by a $M \times (N_1 + N_2 + N_3 + N_4)$ matrix where each row is composed joining head to tail the corresponding rows of the positions on the subspaces.

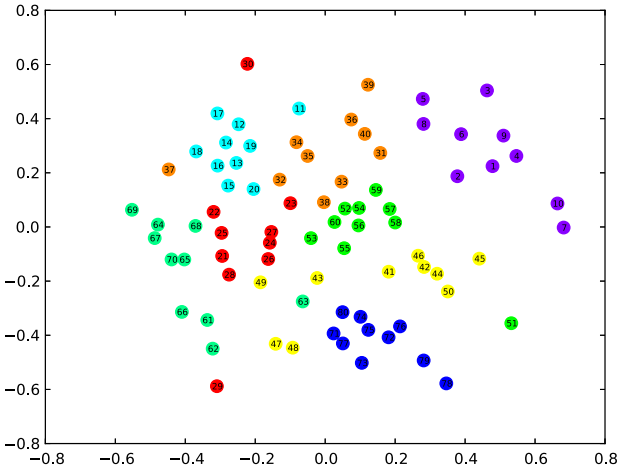


Fig. 4. Bi-dimensional MDS mapping of the betweenness centrality dissimilarities for the third collection. Numbers correspond to texts indices, colors correspond to authors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Instead of a bi-dimensional mapping such as that of Fig. 4, the dimensionalities N_i are calculated based on the stress or cost function (the difference between the actual and the obtained dissimilarities). As stress is a monotonically decreasing function of the number of dimensions we set a threshold of 10% of the value for one dimension (also known as the elbow method) which was usually found to be reached at $N_i = 6$.

3.4. Data analysis

The final positions on the composed space are the attributes for the data analysis algorithms. Analysis is done with supervised learning algorithms from the main types currently in use: tree-based J48; K-Nearest Neighbors (KNN); Naive Bayesian (NB); and Radial Basis Function Network (RBFN). For evaluation purposes, 10-fold and one-leave-out cross-validation [50] were applied.

For KNN a small value ($k = 3$) was chosen in order to consider only the local neighborhood; $k = 1$ was discarded as this choice reduces to using the simpler Nearest Neighbor algorithm and $k = 2$, along with all even numbers, is not recommended due to the possibility of ties. For all other parameters the default values were used, as suggested in [19].

For comparison purposes, the authorship attribution was also performed using the standard tf-idf model. Since tf-idf returns a dissimilarity matrix, we used MDS in this single matrix to apply the same classification algorithms in both approaches.

4. Results and discussion

For the three collections used, the success score with Zero Rule majority classifier is $1/8 = 12.5\%$. The results are outstanding as shown in Tables 1 (10-fold cross-validation) and 2 (one-leave-out), especially when MDS was used. It seems therefore that reducing the dimensionality actually amounted to an efficient feature selection, probably eliminating data that brought noise to the analysis. With MDS, typical accuracy rates were above 90% and the maximum value was 98.75% obtained with KNN for the third collection which corresponds to only one text (out of 80) not correctly classified. These scores greatly surpassed the values obtained by applying the tf-idf method, for which the mean scores among collections were 36.67% for J48, 66.25% for KNN, 63.75% for NB, and 65% for RBFN, as shown in Fig. 5. These scores demonstrate the added value of using the network structure over relying only on the frequency of appearance of features.

A random forest algorithm with tf-idf was tested showing an improvement of the scores; however, the values reached ranged between 60 and 70%. Significantly, the higher scores for the approach introduced here are maintained when changing the classification algorithm (KNN, NB, and RBFN), which indicates the robustness of the proposed metrics. Also worth noting is that the present approach outperforms a previous one where the topology of networks was used without considering node labels [37], for which the accuracy rates for the second collection studied here were 63.75% with J48, 88.75% with KNN, 81.25% with NB, and 83.75% with RBFN. In addition, the present approach is less demanding, both computationally and conceptually, than the previous one.

Taken together, the results indicate that, apart from the frequency of appearance and syntactical relations, certain words are essential to the structure of a text as a whole. The procedure to identify such words using complex networks has been successful, since utilization of these words is author-dependent. The co-occurrence network procedure allows one to observe the features of a word at different scales in the text. For instance, words with low intermittency, i.e. whose appearance in the

Table 1
Accuracy rates (in percentage) in identifying the authors in the three collections, using several machine learning algorithms. The validation was performed using the 10-fold cross-validation strategy. Results are shown with the input comprising the whole dissimilarity matrices (without MDS) and applying MDS on the matrices. For the three last algorithms MDS improved accuracy in all cases.

	J48	KNN	NB	RBFN
Without MDS				
Collection 1	73.75	85.00	85.00	62.50
Collection 2	73.75	83.75	82.50	78.75
Collection 3	75.00	92.50	80.00	71.25
Using MDS				
Collection 1	72.50	87.50	92.50	90.00
Collection 2	63.75	97.50	93.75	96.25
Collection 3	73.75	98.75	92.50	95.00

Table 2
Accuracy rates (in percentage) in identifying the authors in the three collections, using several machine learning algorithms. The validation was performed using the leave-one-out strategy. Results are shown with the input comprising the whole dissimilarity matrices (without MDS) and applying MDS on the matrices. For the three last algorithms MDS improved accuracy in all cases.

	J48	KNN	NB	RBFN
Without MDS				
Collection 1	72.50	85.00	83.75	56.25
Collection 2	68.75	81.25	81.25	78.75
Collection 3	68.75	92.50	77.50	75.50
Using MDS				
Collection 1	73.75	85.00	91.25	91.25
Collection 2	57.50	95.00	92.50	92.50
Collection 3	68.75	98.75	91.25	93.75

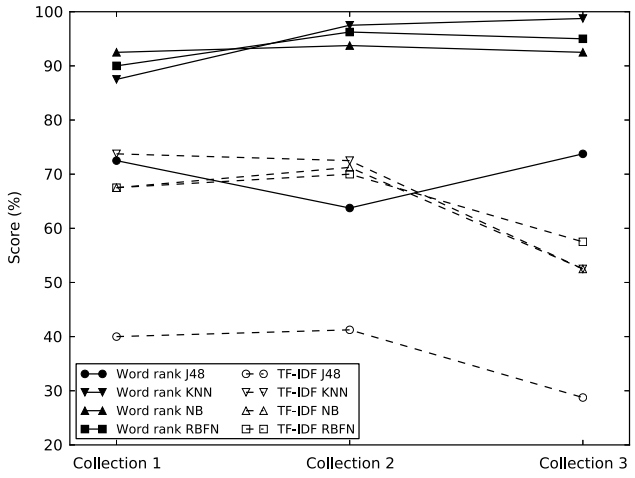
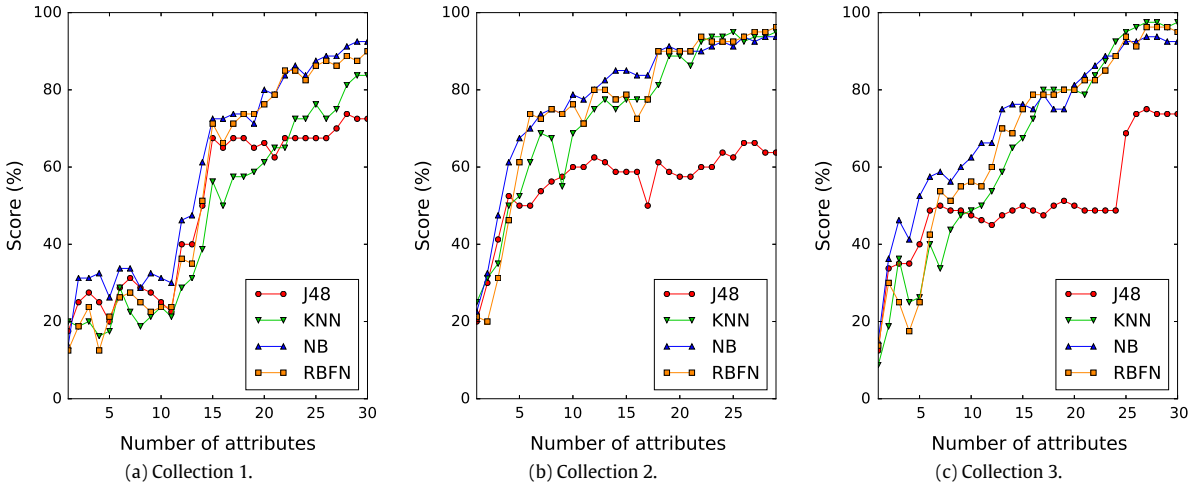


Fig. 5. Scores obtained with the method introduced here, also using MDS, and with tf-idf for the three collections of texts.

text is highly periodic, had a high relevance for the authorship attribution, even though a word can have a low intermittency and may be absent in most paragraphs.

Ours is a unified framework for multivariate analysis of texts. In contrast to other multivariate approaches, the generalized similarity measure (4) allows one to easily introduce new features to the scheme using some of the many node-local network metrics in existence. Care must be taken, however, because not all metrics are useful and some can even lower the performance. For example, we tested the clustering coefficient (characterized by a bell-shaped distribution) and eigenvalue centrality without success. Even though the computation of the similarity measures between documents resembles that of tf-idf (i.e. cosine similarity) there are significant differences, particularly those related to the much smaller sizes of the vectors, with no repeated values. A question to be further studied is the optimal ranking procedure: we chose a ranking *a la* Zipf because of the presence of power law distributions, but other rankings could be possible. While both tf-idf and our



(a) Collection 1. (b) Collection 2. (c) Collection 3.

Fig. 6. Success scores calculated with subsets comprising the attributes with the highest variances. After applying MDS a total of 30 attributes are used for the first and the third collection, while for the second collection this number was 29.

Table 3
The five attributes with the highest variances of each collection. Columns represent the ordering of attributes by variance. Attributes follow the nomenclature: D for degree (connectivity); I for intermittency; SP for shortest path; and BC for betweenness centrality. The numbers stand for the index of the corresponding dimension after applying MDS.

	1	2	3	4	5
Collection 1	SP5	I3	BC3	SP9	SP8
Collection 2	SP7	SP5	D2	D4	SP3
Collection 3	SP7	D2	SP1	SP5	BC2

method account for the heterogeneity of sizes of texts, our ranking procedure has two principal advantages: computation is faster and, most importantly, the ranking does not have to be repeated every time the collection is modified, which is especially advantageous with big collections.

In order to illustrate the relative importance of the attributes, Fig. 6 shows the success scores calculated with a feature selection based on variance threshold, i.e. from each collection subsets are constructed using the attributes with the highest variances. Using few attributes gives small scores (left ends of the figures), which means that the present approach is inherently multivariate. We also present the five attributes with the highest variances in Table 3, most of which are related to the shortest path. For instance, the seventh dimension of the MDS mapping of the shortest path dissimilarity matrix is the attribute with the highest variance for both the second and the third collections.

5. Conclusions

We have introduced an approach by which the representation of text with complex networks is enhanced by considering the words corresponding to the nodes. This is done with a similarity metric to compare two pieces of text where the presence of the most relevant words, according to network metrics, is taken into account. When the data obtained with the similarity metrics were used as input into machine learning algorithms, a high accuracy was achieved which reached 98.75% for one of the book collections. Significantly, the accuracy was considerably higher than for traditional methods based on tf-idf, being also higher than other network approaches that did not consider the label of the nodes. Also relevant is that the performance was improved with dimensionality reduction with MDS, which is advantageous owing to the lower computational cost.

With regard to the limitations, one should emphasize that the present approach is not useful for very short texts (such as a summary of an article). Moreover, even if a person has a characteristic writing fingerprint owing to their particular way to learn a language [51], the traits that define such a fingerprint are probably complex and not bounded to one single measure. A possible way to address this limitation would be to extend the method to employ other metrics and multi-node distributions. As for applications rather than authorship attribution, the approach proposed could be used for part-of-speech analysis of network distributions and resolution of word polysemy.

Acknowledgments

This work was supported by CNPq (Brazil) and FAPESP (grants 2014/20830-0, 2013/14262-7 and 2016/19069-9).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2017.12.054.

References

[1] W. Liang, Spectra of english evolving word co-occurrence networks, Physica A 468 (2017) 802–808.
[2] T.C. Silva, D.R. Amancio, Word sense disambiguation via high order of learning in complex networks, Europhys. Lett. 98 (5) (2012) 58001.
[3] C.D. Manning, P. Raghavan, H. Schütze, et al., Introduction to Information Retrieval, Vol. 1, Cambridge University Press, Cambridge, 2008.
[4] M.Z. Asghar, A. Khan, S. Ahmad, M. Qasim, I.A. Khan, Lexicon-enhanced sentiment analysis framework using rule-based classification scheme, PLoS One 12 (2) (2017) e0171649.
[5] D.R. Amancio, O.N. Oliveira Jr., L.F. Costa, Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index, J. Inform. 6 (3) (2012) 427–434.
[6] M.P. Viana, D.R. Amancio, L.F. Costa, On time-varying collaboration networks, J. Inform. 7 (2) (2013) 371–378.
[7] P. Juola, Authorship attribution, Found. Trends Inf. Retr. 1 (3) (2006) 233–334.
[8] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60 (3) (2009) 538–556.
[9] D.R. Amancio, E.G. Altmann, O.N. Oliveira Jr., L.F. Costa, Comparing intermittency and network measurements of words and their dependence on authorship, New J. Phys. 13 (12) (2011) 123024.
[10] D.R. Amancio, Authorship recognition via fluctuation analysis of network topology and word intermittency, J. Stat. Mech. Theory Exp. 2015 (3) (2015) P03005.
[11] F. Peng, D. Schuurmans, S. Wang, V. Ešelj, Language independent authorship attribution using character level language models, in: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 2003, pp. 267–274.
[12] H.J. Escalante, T. Solorio, M. Montes-y-Gómez, Local histograms of character n-grams for authorship attribution, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, 2011, pp. 288–298.
[13] C. Forstall, W. Scheirer, Features from frequency: Authorship and stylistic analysis using repetitive sound, J. Chicago Colloq. Digit. Humanit. Comput. Sci. 1 (2010).
[14] O.V. Kukushkina, A. Polikarpov, D.V. Khmelev, Using literal and grammatical statistics for authorship attribution, Probl. Inf. Transm. 37 (2) (2001) 172–184.
[15] C.E. Chaski, Who's at the keyboard? authorship attribution in digital evidence investigations, Int. J. Digit. Evidence 4 (1) (2005) 1–13.
[16] Z. Harris, Distributional structure, Word 10 (2–3) (1954) 146–162.
[17] G.K. Zipf, The Psycho-Biology of Language, Houghton, Mifflin, 1935.
[18] R. Ferrer-i-Cancho, R.V. Solé, Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited, J. Quant. Linguist. 8 (3) (2001) 165–173.
[19] D.R. Amancio, C.H. Comin, D. Casanova, G. Travieso, O.M. Bruno, F.A. Rodrigues, L.F. Costa, A systematic comparison of supervised classifiers, PLoS One 9 (4) (2014) e94137.
[20] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1) (1972) 11–21.
[21] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, Int. J. Comput. Vis. 73 (2) (2007) 213–238.
[22] V. Kešelj, F. Peng, N. Cercone, C. Thomas, N-gram-based author profiles for authorship attribution, in: Proceedings of the conference pacific association for computational linguistics, PACLING, Vol. 3, 2003, pp. 255–264.
[23] R. Clement, D. Sharp, Ngram and bayesian classification of documents for topic and authorship, Lit. Linguist. Comput. 18 (4) (2003) 423–447.
[24] H. Baayen, H. Van Halteren, F. Tweedie, Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Lit. Linguist. Comput. 11 (3) (1996) 121–132.
[25] J. Rygl, K. Zemková, V. Kovár, Authorship verification based on syntax features, in: Proceedings of Sixth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN, 2012, pp. 111–119.
[26] J. Cong, H. Liu, Approaching human language with complex networks, Phys. Life Rev. 11 (4) (2014) 598–618.
[27] Z.-K. Gao, S.-S. Zhang, W.-D. Dang, S. Li, Q. Cai, Multilayer network from multivariate time series for characterizing nonlinear flow behavior, Int. J. Bifurcation Chaos 27 (04) (2017) 1750059.
[28] Z.-K. Gao, Y.-X. Yang, P.-C. Fang, Y. Zou, C.-Y. Xia, M. Du, Multiscale complex network for analyzing experimental multivariate time series, Europhys. Lett. 109 (3) (2015) 30005.
[29] G. Carlsson, F. Mémoli, A. Ribeiro, S. Segarra, Axiomatic construction of hierarchical clustering in asymmetric networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 5219–5223.
[30] Z.-K. Gao, W.-D. Dang, Y.-X. Yang, Q. Cai, Multiplex multivariate recurrence network from multi-channel signals for revealing oil-water spatial flow behavior, Chaos 27 (3) (2017) 035809.
[31] Z.-K. Gao, M. Small, J. Kurths, Complex network analysis of time series, Europhys. Lett. 116 (5) (2016) 50001.
[32] S. Segarra, A. Ribeiro, Stability and continuity of centrality measures in weighted graphs, IEEE Trans. Signal Process. 64 (3) (2016) 543–555.
[33] B. Luzar, Z. Levnajic, J. Povh, M. Perc, Community structure and the evolution of interdisciplinarity in slovenia's scientific collaboration network, PLoS One 9 (4) (2014) e94429.
[34] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, Proc. R. Soc. B 268 (1485) (2001) 2603–2606.
[35] M. Choudhury, D. Chatterjee, A. Mukherjee, Global topology of word co-occurrence networks: Beyond the two-regime power-law, in: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 2010, pp. 162–170.
[36] A. Mehri, A.H. Darooneh, A. Shariati, The complex networks approach for authorship attribution of books, Physica A 391 (7) (2012) 2429–2437.
[37] C. Akimushkin, D.R. Amancio, O.N. Oliveira Jr., Text authorship identified using the dynamics of word co-occurrence networks, PLoS One 12 (1) (2017) e0170527.
[38] R.V. Solé, B. Corominas-Murtra, S. Valverde, L. Steels, Language networks: Their structure, function, and evolution, Complexity 15 (6) (2010) 20–26.
[39] D.R. Amancio, O.N. Oliveira Jr., L.F. Costa, Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts, Physica A 391 (18) (2012) 4406–4419.
[40] S. Segarra, M. Eisen, A. Ribeiro, Authorship attribution using function words adjacency networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 5563–5567.
[41] S. Segarra, M. Eisen, A. Ribeiro, Authorship attribution through function word adjacency networks, IEEE Trans. Signal Process. 63 (20) (2015) 5464–5478.

[42] D.R. Amancio, M.G.V. Nunes, O.N. Oliveira Jr., L.F. Costa, Using complex networks concepts to assess approaches for citations in scientific papers, Scientometrics 91 (3) (2012) 827–842.
[43] H.F. de Arruda, L.F. Costa, D.R. Amancio, Using complex networks for text classification: Discriminating informative and imaginative documents, Europhys. Lett. 113 (2) (2016) 28007.
[44] D.R. Amancio, Probing the topological properties of complex networks modeling short written texts, PLoS One 10 (2) (2015) e0118394.
[45] I. Borg, P. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer, 2005.
[46] B.B. Greene, G.M. Rubin, Automatic Grammatical Tagging of English, Department of Linguistics, Brown University, Providence, Rhode Island, 1971.
[47] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, A.M. Somoza, Keyword detection in natural languages and dna, Europhys. Lett. 57 (5) (2002) 759.
[48] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2) (2001) 163–177.
[49] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1) (1964) 1–27.
[50] Z.-K. Gao, Q. Cai, Y.-X. Yang, N. Dong, S.-S. Zhang, Visibility graph from adaptive optimal kernel time-frequency representation for classification of epileptiform eeg, Int. J. Neural Syst. 27 (04) (2017) 1750005.
[51] H. Van Halteren, H. Baayen, F. Tweedie, M. Haverkort, A. Neijt, New machine learning methods demonstrate the existence of a human stylome, J. Quant. Linguist. 12 (1) (2005) 65–77.