Physica A 531 (2019) 121728

Contents lists available at ScienceDirect

Physica A

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

Correlation analysis of short text based on network model

Dongyang Yan [a], Keping Li [a,*], Jingjing Ye [b]

[a] State Key Laboratory of Rail Traffic Control and Safety, Beijing 100044, China
[b] School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China

HIGHLIGHTS

- We propose a network model for short text analysis using the long-range correlation.
- The conditions that the words have long-range correlation in the network model are given.
- This method can get the inner-correlation of words in short texts.
- We find the general laws of word correlation in the network model of short texts.

ARTICLE INFO

Article history:
Received 3 January 2019
Received in revised form 9 May 2019
Available online 13 June 2019

Keywords:
Long-range correlation
Network model
Short text
Fluctuation analysis

ABSTRACT

Correlation of words in the text is of great importance in text analysis like text retrieval, keywords extraction, and text clustering. For short text, because of the limited information of text content, it is difficult to catch the correlation well among words. In this paper, we propose an algorithm based on the complex network to calculate the correlation of words in short texts. A new variable *Edge-degree* is proposed and used in studying the network model of texts. By using fluctuation analysis, we give the condition that *Edge-degree* correlation between words exists beyond nearest neighbors. Further analysis shows that numerical results of the fluctuation function of *Edge-degree* act a power law distribution and that the scaling exponent diverges at a long distance under the finite size effect and varies in different texts. The fluctuation function separates the words in a text into different clusters, and this property is used to measure inner-correlation of different words. Hub nodes act a significant influence on the long-range *Edge-degree* correlation through changing the linear trend of the fluctuation function in a log–log plot.

1. Introduction

The development of online information spreading platforms, with Twitter, Facebook and WeChat as examples, is showing a boosting trend, among which short text has become an especially important way to exchange information. Text correlation calculation is a high-profile task in different fields like keyword extraction, text classification and so on. Correlation is a feature to evaluate the degree of closeness of two words (or texts). Generally, the technologies proposed so far to calculate correlation include frequency-based methods, vector-based methods, and ontology-based methods.

A typical frequency-based approach is TF-IDF (term frequency-inverse document frequency). The importance of a word increases proportionately as it appears in a document (term frequency, TF), and decreases inversely as it appears in the corpus (inverse document frequency, IDF). TF-IDF is widely used in text analysis [1,2].

* Corresponding author.
E-mail address: kpli@bjtu.edu.cn (K. Li).

https://doi.org/10.1016/j.physa.2019.121728
0378-4371/

Vector-based methods first transfer documents (words) into vectors based on the idea of vector space model (VSM), then the correlation is represented by the distance of vector, like Euclidean distance or cosine power (called cosine similarity) in vector space. One of the prominent features of this method is the way to map texts into vectors. Bag of words (BoW) [3] represents documents as vectors; Word embedding [4,5] maps words into vectors; Latent semantic indexing (LSI) [6] uses the singular value decomposition (SVD) technique to transform the word frequency matrix into a singular matrix, and then small singular values are eliminated; PLSA [7] is proposed with a better statistical foundation and more proper generative model of data.

Ontology-based methods try to calculate the correlation of words relying on background knowledge (called ontology). The ontology is the large corpora that people have summed up over the last decades like WordNet [8], Wikipedia [9], etc. Correlation of two words can be measured by the existing forms such as the distance of words occurring in the ontology.

These three kinds of methods can get ideal performance for texts with large corpus, but in terms of short texts processing, with limited contextual information, the accuracy is considerably lost in the analysis. The analysis on short texts is becoming a challenging issue because many documents include only a few words and they can be noisy and ambiguous. In this context, discovering relevant information from short texts gets a lot of attention in many areas, especially in the areas of classification [10,11], clustering [12], sentiment analysis [13], and topic modeling [14,15]. To cope with the problem of a small content of words, two strategies are preferred: one strategy is to introduce pseudo-document [14,16]; another strategy is to use the semantic relation of existing sources [11]. However, with the introduction of pseudo-document, the noise will also be unavoidably introduced; and with the bias of the reference sources, the semantic relation strategy may neglect the unique feature of a short text.

Recent years have witnessed the rapid development of network science. Researchers have found that, in a complex system, the emergent characteristic in holistic properties are highly relevant to the interactions between elements on a microscopic level. For example, the properties of a network, such as clustering and fractality, are related to the nearest degree correlation [17–20]. Meanwhile, degree correlation beyond nearest neighbors is an essential factor in describing the non-local properties of a network [21,22]. Network science has also been employed in text analysis. Many ways of applications to text processing have been studied, and with particular relevance to the topological properties of network, text classification [23], text quality [24], keywords extraction [25], authorship [26,27], authors' style recognition [28,29], and text similarity [30] show good performance in practice. Moreover, the proper sampling of short fragments of a text can perform a similar discriminability to the full text in statistical analyses with network tools [10].

The correlation analysis has already been used in text analysis. Zheng et al. [31] have come up with a short text classification method based on correlation analysis. Ning et al. [32] have used semantic correlation of HowNet to do short text classification. The n-gram [33] and skip-gram [34] model, which are based on the short-range correlation, are representative and widely used in text processing. Also, researchers have applied the long-range correlation to detect keywords of texts [35]. In this paper, we propose an algorithm to calculate the correlation of words in short texts. By using the long-range correlation existing in texts [36] and network [37,38], this method focuses on extracting correlation of words effectively without background knowledge and extra training. Our work mainly captures the correlation laws of short texts, which is reflected in the following three aspects: Firstly, we construct the network model of short text and propose variable *Edge-degree*. Secondly, we use fluctuation analysis to calculate the *Edge-degree* correlation of two words and prove that long-range correlation exists in *Edge-degree* correlation of network model constructed from a short text. Finally, an algorithm is proposed to measure the correlation of words.

The rest of this paper is structured as follows: In Section 2 we will introduce the proposed method based on the network model and variable *Edge-degree*. Also, the correlation of *Edge-degree* beyond the nearest neighbor is proved; in Section 3, we present experimental results; finally, in Section 4, we conclude this paper.

2. Our proposed method

In this section, we describe in detail our proposed correlation calculating method. Long-range correlation of variable *Edge-degree* is firstly proved with the tool of fluctuation analysis (see [38–40]). Then this long-range correlation is used to get correlated words.

2.1. Construction of the network model

The network model is a tool to study language from a systematic perspective. The basic form of network is $N = (V, E)$, where N represents the network, V is the set of nodes and E is the set of edges. Set V contains all linguistic units (in English texts, the linguistic unit can be word or letter) in the source text and set E contains all edges which represent the pairwise relation between linguistic units. In this paper, we choose the co-occurrence [41] network to construct the network model for texts, and nodes are connected when they are adjacent (occur together in at most two consecutive words). Take the following sentences (1) to (4) for example, the network is constructed as Fig. 1.

(1) Mary and Lucy are good friends
(2) Mary likes playing basketball and Lucy likes playing football
(3) Mary get hurt once playing games

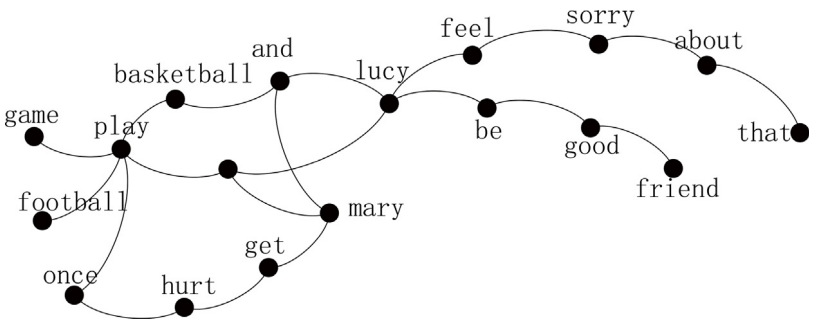

Fig. 1. Word co-occurrence complex network, the same words are converted to the single node and two nodes are linked by an edge if they are adjacent in at least one sentence in the source text.

(4) Lucy feels sorry about that

The weight of an edge represents the distance between this edge's both ends. In this paper, we set the weight of every edge to be 1. Meanwhile, the edge is undirected, which means if two nodes are connected by an edge, they are accessible to each other.

2.2. Definition of Edge-degree correlation

The degree k_i of node i is the number of edges attached to node i. If degree correlation is used in correlation analysis of text network, the characteristics presented in the network structure will also be reflected in the text network. Nevertheless, for text network, taking text features into consideration gives the network some more specific information. For example, two double nodes (assume as $a1, a2$ and $b1, b2$) with k and k' in text network have the same $P(k, k')$, where $P(k, k')$ is the probability that a node $a1$ with k is connected to node $a2$ with k'. But in the source text, the frequency of co-occurrence of words corresponding to $a1, a2$ and $b1, b2$ may be different. Therefore, *Edge-degree* is suggested in this paper, which is the product of the degree of one node in a text network and its corresponding frequency of co-occurrence with another node in the source text. The *Edge-degree* is represented as $kf = k * f$, where k denotes degree and f denotes the frequency of co-occurrence. For nodes $a1, a2$, kf is directive, we use the subscript $a \rightarrow b$ to denote the direction, then $kf_{a1\rightarrow a2} = k_{a1} * f$ and $kf_{a2\rightarrow a1} = k_{a2} * f$, where k_a denotes degree of a. In the following discussion, for convenience, "$a1 \rightarrow a2$" is abbreviated as "$a1, a2$".

Like degree correlation, analysis on the network of the short texts shows that *Edge-degree* correlation also exists beyond the nearest neighbors. To prove it, firstly, we introduce the fluctuation analysis.

Fluctuation analysis is a method to determine whether the records have a long-range correlation (see [39,42]). Consider a record $x_i (i = 1, 2, \ldots, N)$ with the index i corresponding to the recording time. People are interested in the correlation function $C(s)$ of value x_i and x_{i+s} with time scale s. If there exists long-range correlation, $C(s)$ declines as a power law as $C(s) \sim s^{-\gamma}$ with $0 < \gamma < 1$. Usually, the calculation of $C(s)$ is not easy, so people use fluctuation function $F(s)$ instead, and $F(s) \sim s^{\alpha}$, the exponent α of which is related to γ. For *Edge-degree*, fluctuation analysis can be used to determine whether there is a long-range correlation at large distance in short text network. Instead of being treated like records, the distance between two nodes is based on the perspective of the topological relation. For example, nodes $a1$ and $a2$ in network model has shortest paths, second shortest paths ... longest paths and so on, which indicates that the scale s in network model naturally occurs without dividing long paths into shorter segments. To calculate the $C(s)$ of *Edge-degree*, *Edge-degree* is processed as follows:

1. Find all paths that start from a node x_s and end at the node x_e. Paths are not supposed to exist loop.
2. For every path found in step 1, calculate the weighted average of *Edge-degree*:

$$kf_d = \frac{1}{\sum_{i=1}^{d} f_{i,i-1}} \sum_{j=1}^{d} f_{j,j-1} k_j \tag{1}$$

   Where kf_d is the weighted average of *Edge-degree* of paths with length d. d plays the role of length of the path. $f_{i,i-1}$ is the frequency that pairs of nodes in site i and $i-1$ along the path co-occur (neighbor) in one sentence of the source text. k_j is the degree of the node in site j along the path. For the start node x_s, the site of which is 0.
3. Repeat step 1 and step 2 from length $d = 1$ to length $d = max$, where "max" denotes the maximum length of paths start from s and end at e. A nd get all kf_d ($d = 1$ to max).

Each path with length d is deemed to be a segment with scale $s = d$. The correlation function of the start node x_s and the end node x_e is represented as

$$C_{s,e}(d) = \langle (kf_d^{i,i-1} - \overline{kf})(kf_d^{j,j-1} - \overline{kf}) \rangle \tag{2}$$

Where $\langle * \rangle$ denotes the average of $*$, $\overline{kf}$ denotes the average of kf_d at all d (approximate to the average value of $f_{i,j}k_i$, where i and j denote two nodes in the network). $kf_d^{i,i-1} = f_{i,i-1}k_i$, i and j denote node x_i and x_{i-1} along a certain path with length d. If there is long-range *Edge-degree* correlation against long path length d, $C_{s,e}(d)$ is expected to be $C_{s,e}(d) \sim d^{-\lambda}$ and generally the exponent is $0 < \lambda < 1$. Indirectly, the fluctuation function is used to calculate λ. The Fluctuation function is suggested as follow

$$F(d) = \langle (kf_d - \overline{kf}_d)^2 \rangle^{\frac{1}{2}} \sim d^{\alpha} \tag{3}$$

Where $\overline{kf}_d$ is the average value of kf_d with length d, and the relation of exponent α and λ is $\alpha = -\lambda/2$. Next, the mathematical proof is given.

2.3. Mathematical proof

As mentioned above, *Edge-degree* kf_d combines network with the source text, which makes it possible to consider more details of the source text in the network model. For kf_d at a certain length d of all paths from certain x_s to x_e, their variance is

$$\sigma^2(kf_d) = \langle (kf_d - \overline{kf}_d)^2 \rangle \tag{4}$$

Considering that in the network of short text, the difference of each *Edge-degree* is not too significant, $\overline{kf}_d$ of a path can be approximately equal to $\overline{kf}$. Then

$$\sigma^2(kf_d) \approx \langle (\frac{1}{\sum_{i=1}^{d} f_{i,i-1}} \sum_{j=1}^{d} f_{j,j-1} - \overline{kf})^2 \rangle \tag{5}$$

With $f_{i,i-1} \geq 1$, it can be figured out that $d \leq \sum_{i=1}^{d} f_{i,i-1} \leq \mu d$ and $\mu \geq 1$. The value of $\sigma^2(kf_d)$ can be limited to a range $[leq, req]$, where $leq = \langle (\frac{1}{\mu d} \sum_{i=1}^{d} f_{i,i-1}k_i - \overline{kf})^2 \rangle$ and $req = \langle (\frac{1}{d} \sum_{i=1}^{d} f_{i,i-1}k_i - \overline{kf})^2 \rangle$. With Eq. (2), the upper bound is

$$\begin{aligned}
req &\approx \langle (\tfrac{1}{d} \textstyle\sum_{i=1}^{d} f_{i,i-1}k_i - \overline{kf})^2 \rangle \\
&= \tfrac{1}{d^2} \langle \textstyle\sum_{i=1}^{d} (kf_d^{i,i-1} - \overline{kf})^2 \rangle + \tfrac{1}{d^2} \langle \textstyle\sum_{i \neq j}^{i,j \leq d} (kf_d^{i,i-1} - \overline{kf})(kf_d^{j,j-1} - \overline{kf}) \rangle \\
&= \tfrac{1}{d^2} \langle \textstyle\sum_{i=1}^{d} (kf_d^{i,i-1} - \overline{kf})^2 \rangle + \tfrac{1}{d^2} \textstyle\sum_{i \neq j}^{i,j \leq d} C_{s,e}(|j - i|) \\
&= \tfrac{1}{d^2} \langle \textstyle\sum_{i=1}^{d} (kf_d^{i,i-1} - \overline{kf})^2 \rangle + \tfrac{1}{d^2} \textstyle\sum_{k=1}^{d-1} 2(d-k) C_{s,e}(k)
\end{aligned}$$

For paths that from x_s to x_e, if the length d is small, normally the value of path is small, and $\overline{kf}_d \neq \overline{kf}$. When d is large enough (for example, $d > 10$), the assumption of $\overline{kf}_d \approx \overline{kf}$ is appropriate. Then

$$req \approx \frac{1}{d} \langle (kf_d^{i,i-1} - \overline{kf})^2 \rangle + \frac{1}{d^2} \langle \sum_{k=1}^{d-1} 2(d-k) C_{s,e}(k) \rangle$$

If *Edge-degree* is long-range correlated, $C_{s,e}(d) \sim d^{-\lambda}$ and $0 < \lambda < 1$. The second term on the right-hand side of the above equation will dominate if d is large [38,39]. Then $req \approx \frac{1}{d^2} \sum_{k=1}^{d-1} 2(d-k) C_{s,e}(k) \sim d^{-\gamma}$. And similarly, $leq \sim d^{-\gamma}$.

Above all, fluctuation function is suggested as Eq. (3), and

$$F(d) \sim d^{-\gamma/2} \tag{6}$$

For Eq. (6), $\alpha = -\gamma/2$. When $-1/2 < \alpha < 0$, *Edge-degree* shows positive long-range correlation, while $\alpha < -1/2$, *Edge-degree* shows negative long-range correlation [43].

2.4. The measurement of words correlation

As the long-range *Edge-degree* correlation is confirmed between two nodes in the network model of short texts, there is still one question: How to measure quantitatively the difference of long-range correlation of double nodes; or whether the further processing of *Edge-degree* correlation can reflect the difference of nodes in the network? In fact, the differentiation of different words to extract the required information is of important value in the correlation analysis. In order to get the quantitative measure of the difference of different nodes, one formula is suggested as follow

$$Q(i,j) = \frac{1}{2} [\sum_{d=k}^{K} log(F_{i,j}(d)) + \sum_{d=k}^{K} log(F_{j,i}(d))] \tag{7}$$

Where $Q(i,j)$ denotes the quantitative measure of long-range *Edge-degree* correlation of node i and j. It is the sum of the logarithm of $F_{i,j}(d)$ from $d = k$ to $d = K$. If $i = j$ or $Q(i,j) = -\infty$ ($F_{i,j}(d) = 0$ exists with $d \in [k, K]$), $Q(i,j)$ is set to be "0". Eq. (7) is based on the assumption that the trend (reflected by the exponent α of $F(d)$) $F_{i,j}(d)$ is the same between different doubles of nodes i, j of one network, so $Q(i,j)$ can reflect their difference in space. From the definition of $F(d)$, it is easy to find that correlation between two words is not symmetrical: $F_{i,j}(d) \neq F_{j,i}(d)$. To make $Q(i,j)$ symmetrical, $F_{i,j}(d)$ and $F_{j,i}(d)$ are calculated respectively and averaged. In actual calculations, there is no need to add $F_{i,j}(d)$ from $d = 1$ to $d = max$ because part of data, if chosen properly, contains the correct trend of $F_{i,j}(d)$. Besides, the finite size effect of the network leads to continuous "0" results of $F_{i,j}(d)$ when d is near 1 and max. Therefore, $Q(i,j)$ is calculated with the value of d from k to K ($k > 0$ and $K < max$). To facilitate calculation, we magnify the difference among $Q(i,j)$. Then equation (7) is replaced by

$$Q(i,j) = \frac{1}{2}[\prod_{d=k}^{K} F_{i,j}(d) + \prod_{d=k}^{K} F_{j,i}(d)] \tag{8}$$

Although $Q(i,j)$ is the quantitative measure of *Edge-degree* correlation, there is no evidence showing that a large (small) value of $Q(i,j)$ denotes high correlation of node i and j. Nevertheless, this quantitative measure can still mirror internal rules between words in short texts. $Q_{i,j}$ is further processed to be normalized to [0,1] by the following formula

$$Q'(i,j) = \begin{cases} \frac{Q(i,j) - Q_{min} + adj}{Q_{max} + Q_{min} + adj} & Q(i,j) \neq 0 \\ 0 & Q(i,j) = 0 \end{cases} \tag{9}$$

Where Q_{max} denotes the maximum value among $Q(i,j)$. And Q_{min} denotes the minimum value among $Q(i,j)$ except $Q(i,j) = 0$. adj is a very small constant for differentiating $Q'(i,j) = 0$ and $Q'(i,j) \neq 0$ while narrowing the gap between "0" and Q_{min} among $Q'(i,j)$.

One word (i) corresponds to many $Q'(i,j)$ with j changes among other words. $Q'(i)$ is used to represent all $Q'(i,j)$ that i corresponds to, i.e. $Q'(i) = Q'(i,j_1), Q'(i,j_2), \ldots, Q'(i,j_n), Q'(i,i)$. $S = j_1, j_2, \ldots, j_n, i$ is the corpus of corresponding short text. By the statistical characteristics of $Q(i)$ like the mean and variance of its elements, the word i is mapped to space coordinates and is uniquely represented

$$Node(i) = Vector(sta_1, sta_2, \ldots, sta_m) \tag{10}$$

Where $Node(i)$ denotes the coordinate in m-dimensional space and $sta_*(* \in 1, 2, \ldots, m)$ is the statistical characteristic of $Q'(i)$.

After i is mapped into $Node(i)$, the correlation between words can be calculated through the distance of two corresponding nodes.

3. Experimental results

3.1. Variation of scaling exponent in and between texts

As mentioned in Section 2, if $F(d) = \sigma(kf_d) \sim d^{-\alpha}$ and $\alpha < -1/2$, *Edge-degree* shows negative long-range correlation, while $-1/2 < \alpha < 0$, *Edge-degree* shows positive long-range correlation. Experimental analyses show $\alpha < -1/2$, indicating that *Edge-degree* is of negative long-range correlation. Experimental texts are chosen from *The Economist Expresso*, which is the condensed version of *The Economist* and the chosen texts contain 80–100 words. In such texts, most of the words occur only once.

Most of the results show vibration when d is less than 10. In such condition of $d < 10$, there are normally very limited paths being found as is explained in Section 2, because of the finite size effect of the network model. Actually this case happens when d is near max as well, and the assumption that $\overline{kf}_d \approx \overline{kf}$ is inappropriate. Fig. 2 shows the results of *Edge-degree* fluctuation function between certain word and other words. The network is constructed from *Fine time again: Europe v Google* in *The Economist Expresso* and contains 88 nodes. To analyze *Edge-degree* correlation, it is necessary to get $F(d)$ between every two words: $F_{*\to **}$, where $*$ and $**$ denote arbitrary word but $* \neq **$. For certain word β, $F_{\beta\to *}(d)$ denotes all $F(d)$ between β and other words. There exists similarity in $F_{\beta\to *}(d)$: every curve of $F_{\beta\to *}(d)$ follows approximately the same trend. As Fig. 2(a) showing $F_{apps\to fine}(d)$ and $F_{apps\to accuse}(d)$, though the trend of two curves is a little different, in Fig. 2(b), the overall condition of $F_{apps\to *}(d)$ performs a steady trend.

Further analysis shows that the steady trend exists in $F_{*\to **}(d)$. But when different texts are considered, as Fig. 3 shows, the trend of the overall condition of $F_{*\to **}(d)$ is different. To catch the difference of two texts, we choose one word from each text ("apps" from text 1: *Fine time again: Europe v Google*; "market" from text 2: *Thinking about tomorrow: Toyota*) and use the trend of $F_{\beta\to *}(d)$ to approximately represent $F_{*\to **}(d)$.

Because of the existence of the above situations, the trend (represented by α) of $F(d)$ of a text is assumed to be the same.

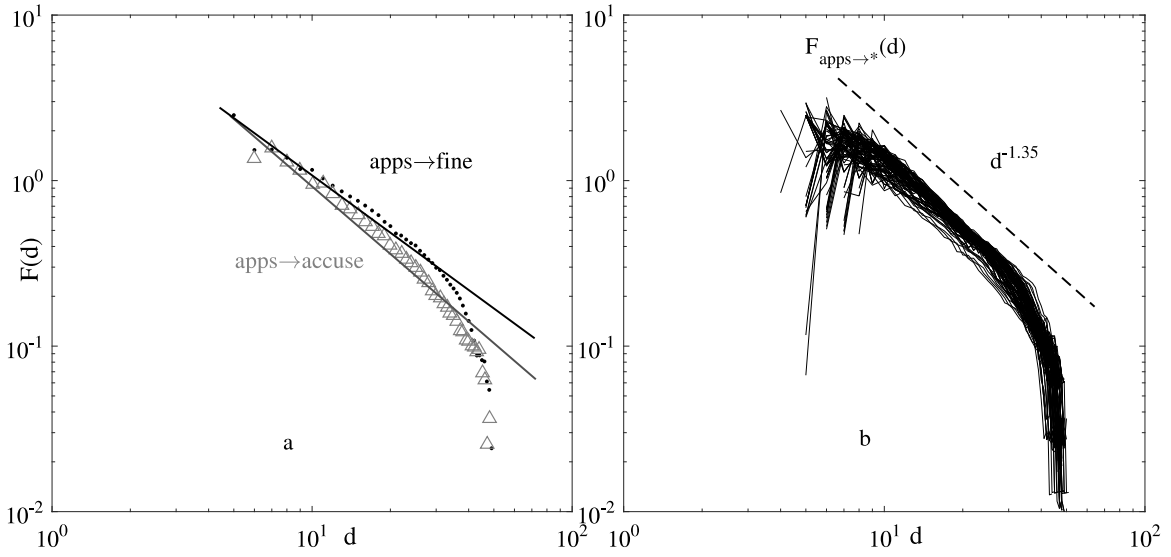

Fig. 2. *Edge-degree* fluctuation function against the length d of paths. The network is constructed from *Fine time again: Europe v Google* in *The Economist Expresso* and contains 88 nodes. Picture (a) shows the results of $F_{apps→fine}(d)$ (solid point) and $F_{apps→accuse}(d)$ (triangle). Picture (b) shows the overall condition of $F_{apps→*}$, $*$ denotes any words except "apps", vibration exists when $d < 10$. (b) The trends of curves are almost the same; but (a) little difference exists. The difference occurs in the value of $F(d)$ against d.

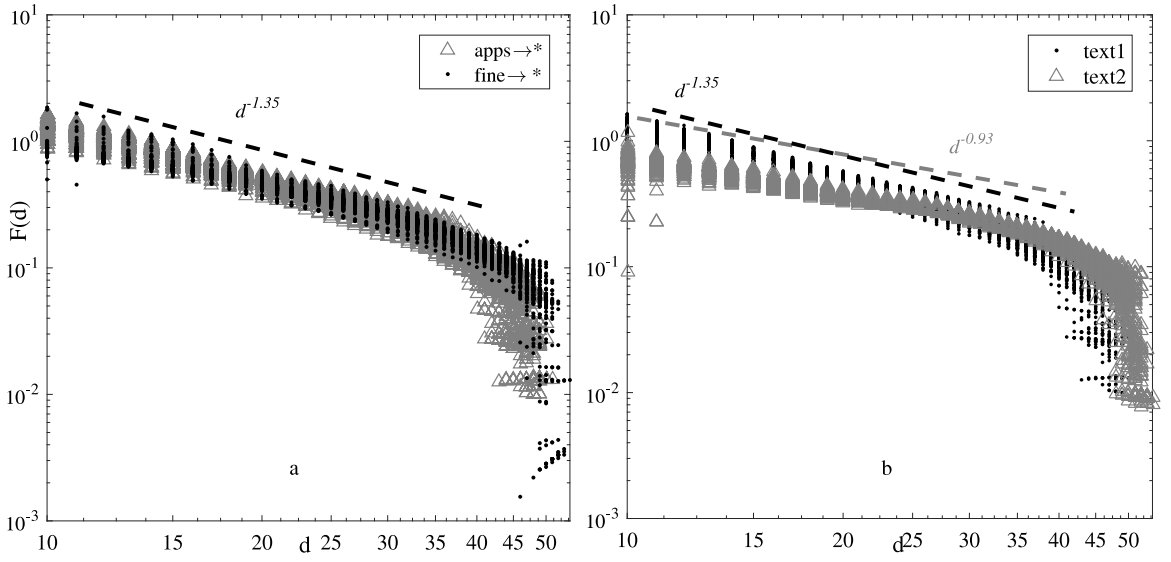

Fig. 3. $F_{\beta→*}(d)$ between words in one text and $F_{*→**}(d)$ between texts. Text 1: *Fine time again: Europe v Google*; text 2: *Thinking about tomorrow: Toyota*. (a) The trend of $F_{apps→*}(d)$ (triangle) and $F_{fine→*}(d)$ (solid point). (b) The trend of $F_{*→**}(d)$ between text 1 (solid point) and text 2 (triangle). Results show that the trend of the overall condition of $F_{*→**}(d)$ is different between texts while the trend of the overall condition of $F_{\beta→*}(d)$ is almost the same in certain text.

3.2. Inner-correlation of words

Eq. (8) in Section 2 is suggested to be the quantitative measurement of *Edge-degree* correlation while the value of $Q(i, j)$ is not used to accurately represent the degree of relation between i and j. Experimental results also illustrate this point. In Table 1, we list words that rank in the top 5 in terms of $Q(i, j)$. Meanwhile, in Table 2, words that rank in the bottom 5 in terms of $Q(i, j)$ are listed. Text 1: *Fine time again: Europe v Google*; Text 2: *Thinking about tomorrow: Toyota*. k and K in Eq. (8) are set to be 15 and 30 respectively and $adj = (Q_{max} - Q_{min})/99$. "a", "after", "out", "of" in Table 1 and "it", "in", "the" in Table 2 are extracted to be top arranged according to $Q(i, j)$. Instead, words that make up the main content tend to be distributed in the middle. This phenomenon occurs among other words. Fig. 4 shows the arranged position

Table 1
Words in the top 5 in terms of $Q(i, j)$.

TEXT	Text 1	Text 1	Text 1	Text 2	Text 2	Text 2
WORD	apps	fine	market	market	Toyota	carmaker
	firm	firm	a	after	after	come
	part	global	firm	will	motor	after
	a	revenue	revenue	already	out	heavier
	bigger	European	European	of	reveal	investor
	rule	offend	maker	investor	disguise	sale

Table 2
Words in the bottom 5 in terms of $Q(i, j)$.

TEXT	Text 1	Text 1	Text 1	Text 2	Text 2	Text 2
WORD	apps	fine	market	market	Toyota	carmaker
	mobile	than	will	it	in	it
	it	important	Google	ahead	it	the
	the	more	than	future	ahead	future
	Google	but	size	the	would	ahead
	than	search	the	firm'	electric	in

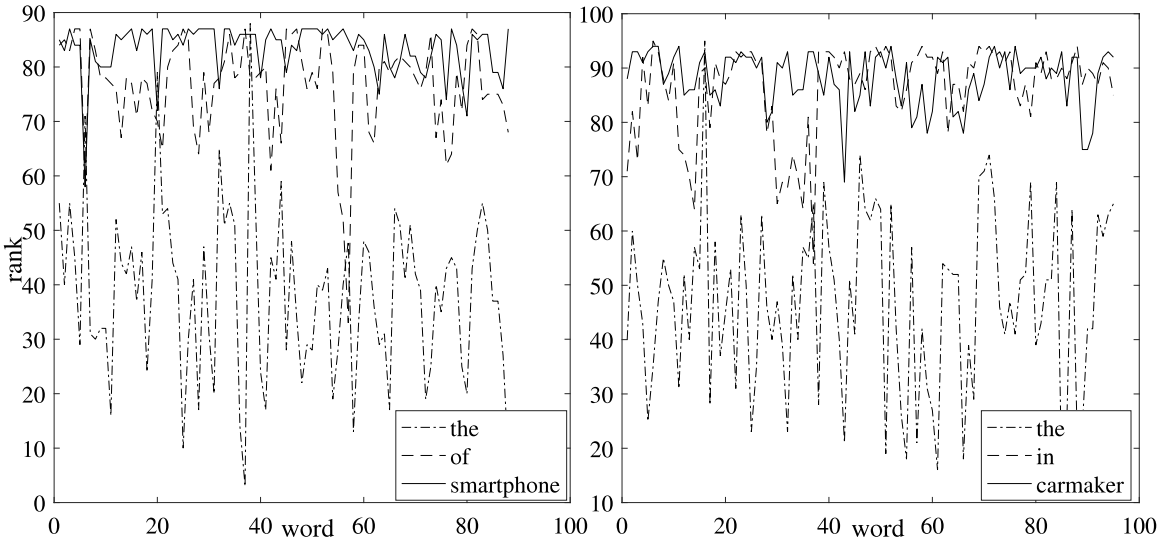

Fig. 4. The arranged position of words according to $Q(i, j)$. (a) text 1: *Fine time again: Europe v Google*. "the" and "of" tend to be arranged in the bottom position (70th–90th) while "smartphone" is arranged in the middle position (30th–60th); (b) text 2: *Thinking about tomorrow: Toyota*. "the", "in" are arranged in the bottom position and "carmaker" is in the middle position.

of words according to $Q(i, j)$ with other words, for example, in Table 1, "rule" is arranged in the 5th place according to $Q(apps, j)$ with word "apps".

To get the whole condition of the distribution of words, mean of $Q'(i)$ of word i is calculated to show its tendency, and variance is calculated to show its fluctuation around the tendency. Fig. 5 shows the distribution of words under the condition that X axis denotes the mean of $Q'(i)$, and Y axis denotes the variance of $Q'(i)$. The results show a striking feature: the mean and variance of $Q'(i)$ show the power law growth. This may indicate that $Q(i, j)$ of different words has intrinsic regularity in short texts.

Specifically, this power law distribution distinguishes the characteristics of $Q'(i)$ of each word, and these characteristics reflect the status of the corresponding word in the text. Fig. 6 shows the results of k-means clustering for words. In the bottom left corner of Fig. 6, μ and σ^2 of $Q'(i)$ are both small. In such situation, words distributed over this part have nearly the same $Q'(i, j)$ with other words. This characteristic is compatible with functional words like 'the', 'in', 'would', 'ahead', 'with', 'that', etc. In the top right corner of Fig. 6, μ and σ^2 are both large. These two large value means that words distributed over this area have μ which are dominated by some relatively large $Q'(i, j)$. It indicates that this kind of word has special relationships with certain words, which may be the words with special existence in text (not the keywords). In the middle of Fig. 6, many nodes are concentrated, and this section contains most of the information in short text.

The above analysis shows that words mapped in space are clustered partly according to their usage in text, and their distance can be a measure for their relevancy to some extent. The further analysis focuses on the distance between words,

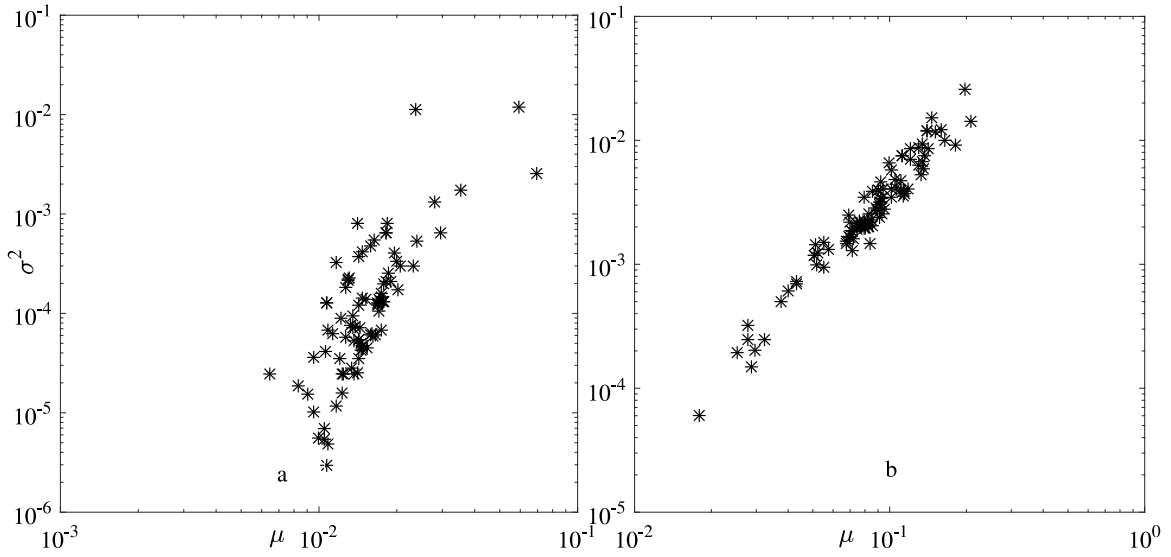

Fig. 5. The distribution of words in (a) text 1: *Fine time again: Europe v Google*; (b) text 2: *Thinking about tomorrow: Toyota*. *X* axis denotes the mean (μ) of $Q'(i)$ of certain word with other words, *Y* axis denotes the variance (σ^2) of $Q'(i)$ of certain word with other words. It presents a linear distribution in the double log coordinate system; this distribution form means that the mean and variance of $Q'(i)$ show the power law growth.

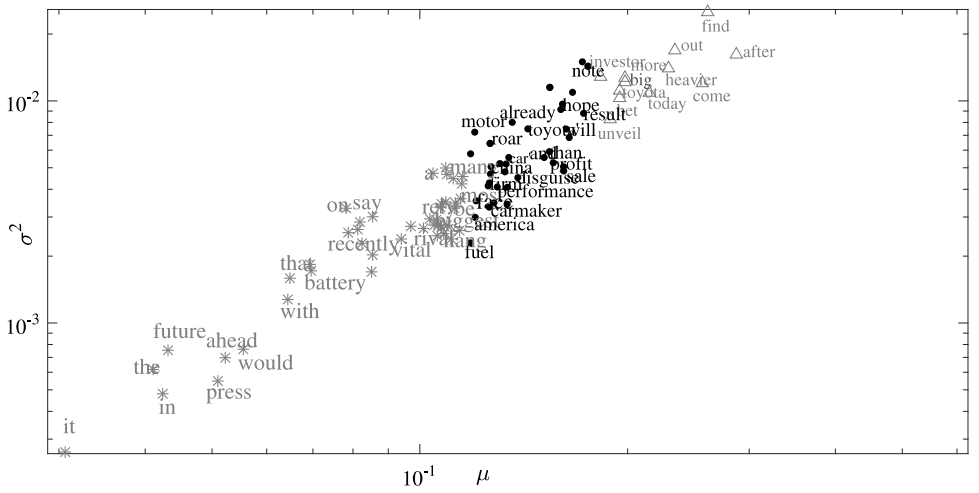

Fig. 6. The distribution of words in the text: *Thinking about tomorrow: Toyota*. Words are clustered into 3 parts by k-means clustering. Words that carry the main content of text, like "motor", "carmaker", "market", "Toyota", etc. tend to be in the middle position of the graph; Functional words like 'the', 'in', 'would', 'ahead', 'with', 'that' lies in the left bottom position.

Table 3
Top 5 related words.

TEXT	Text 1	Text 1	Text 1	Text 2	Text 2	Text 2
WORD	apps	fine	market	market	Toyota	carmaker
	bigger	when	internet	rival	fuel	exhaust
	revenue	important	European	vital	power	first
	almost	penalty	remedy	rely	hydrogen	reveal
	abuse	amount	have	behind	America	hope
	consider	focus	case	biggest	electric	not

Eq. (10) with 2-dimensional vector, where $sta_1 = log\mu$ and $sta_2 = log\sigma^2$, is used to map words into space. Some related words are shown in Table 3. The results are appropriate for this text.

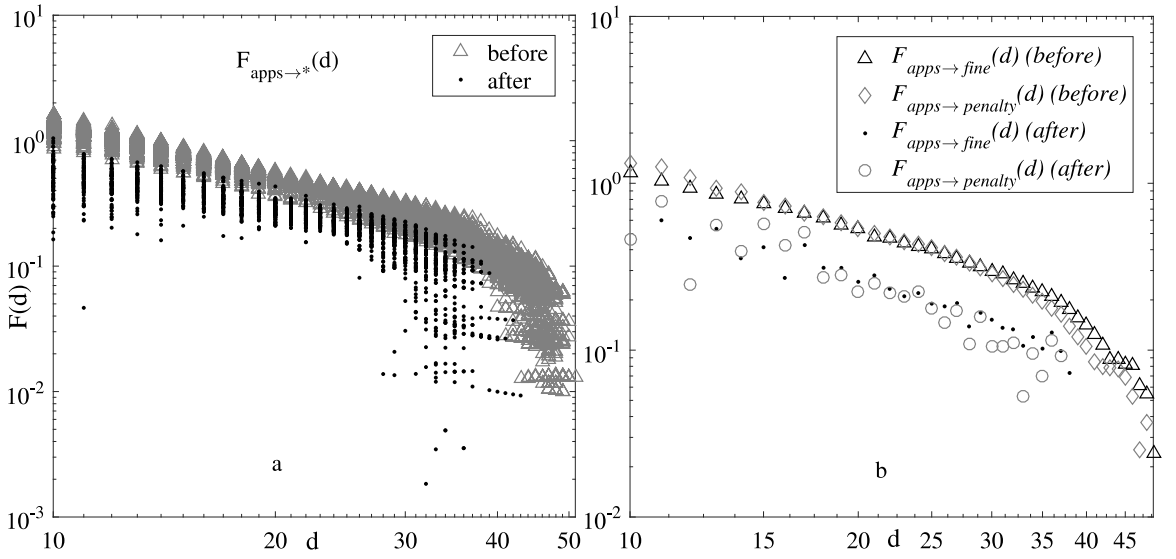

Fig. 7. The variation of $F(d)$ before and after deleting the hub node with the highest degree in the network model. The network is constructed from *Fine time again: Europe v Google*. (a) Overall condition of $F_{apps\to *}(d)$. After removing the hub node with the highest degree, all points in the picture move down, and the distribution of points become sparse. (b) $F_{apps\to fine}(d)$ and $F_{apps\to penalty}(d)$ show that $F(d)$ in log–log plot loses the linear feature.

It is worth to note that, we do not distinguish features that are specific to words, like word formation, word classes, etc. The rank of related words may show a big difference when different situations are considered, meanwhile, when the range of words to be analyzed is narrowed, the results will be more in line with actual needs.

3.3. Effect of the hub nodes

What is the media of long-range Edge-degree correlation? Further analysis shows the very important role of "the" and prepositions like "to", "of", etc. In this section, one hub node is deleted and then $F(d)$ is recalculated. By comparing the difference with the results before deleting hub nodes, the hub nodes are found to be a key factor.

The hub nodes of many texts' network model are prepositions or articles like "of", "to", "the", etc. When hub nodes are deleted, we can find both the connections between most of the words will disappear and the maximum path length and quantity of paths between two words will become smaller. $F(d)$ loses the linear feature in the log–log plot when recalculated, which indicating that long-range *Edge-degree* disappears.

Fig. 7 shows the variation of $F(d)$ after deleting the hub node with the highest degree. Fig. 7(a) shows the overall condition of $F_{apps\to *}(d)$ with some noticeable changes: the points move down and become sparse in the picture after deleting the hub node with the highest degree. Fig. 7(b) further illustrates the changes of $F_{apps\to fine}$ and $F_{apps\to penalty}(d)$: the positional relation of two curves also changes.

4. Conclusion

In this paper, a method to calculate the correlation of words in English short text based on the network model is proposed. First, we introduce *Edge-degree* to consider some information in source text while calculating the correlation of words based on the network model. Then we prove that there exist *Edge-degree* correlation beyond the nearest neighbor of nodes in the network. Finally, we propose a quantitative measure for the existence of difference of double nodes. And we find that the statistical property of these quantitative measures can be used to distinguish different words and map them into space, and then use the distance to be an indicator of correlation of two words. Experimental results show that the proposed method can effectively express the correlation between different words.

The proposed method can be applied in many fields, especially in the analysis of short texts. In natural language processing, correlation calculating is of great importance because the higher accurate measurement of correlation in texts will improve many techniques like keyword extraction, text classification, text summarization.

The regularity of words' distribution after mapped into space shows the inner correlation of words in texts. This regularity may exist in larger texts. Also, some more targeted statistical tools may get the more general form of information distribution in texts.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 71621001) and the Research Foundation of State Key Laboratory of Railway Traffic Control and Safety, China, Beijing Jiaotong University (Grant No. RCS2019ZT001).

References

[1] B. Trstenjak, S. Mikac, D. Donko, KNN with TF-IDF based framework for text categorization, in: Proceedings of the 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 23-26 Oct 2013, vol. 69, Univ Zadar, Zadar, Croatia, 2014, pp. 1356–1364.
[2] A.Z. Guo, T. Yang, Research and improvement of feature words weight based on TFIDF algorithm, in: Proceedings of IEEE Information Technology, Networking, Electronic and Automation Control Conference, 20-22 May 2016, Chongqing, China, 2016, pp. 415–419.
[3] Y. Zhang, R. Jin, Z.H. Zhou, Understanding bag-of-words model: A statistical framework, Int. J. Mach. Learn. Cyb. 1 (2010) 43–52, http://dx.doi.org/10.1007/s13042-010-0001-0.
[4] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.
[5] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, ACM, New York, USA, 2008, pp. 160–167.
[6] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. Tec. 41 (1990) 391–407, http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
[7] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (2001) 177–196, http://dx.doi.org/10.1023/A:1007617005950.
[8] A. Hotho, S. Staab, G. Stumme, Ontologies improve text document clustering, in: Proceedings of the 3rd IEEE International Conference on Data Mining, 19-22 Nov 2003, Melbourne, FL, 2003, pp. 541–544.
[9] X.H. Hu, X.D. Zhang, C.M. Lu, E.K. Park, X.H. Zhou, Exploiting Wikipedia as external knowledge for document clustering, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Jun 28-Jul 01 2009, Paris, France, 2009, pp. 389–396.
[10] D.R. Amancio, Probing the topological properties of complex networks modeling short written texts, PLoS One 10 (2) (2015) e0118394, http://dx.doi.org/10.1371/journal.pone.0118394.
[11] P. Wang, B. Xu, J.M. Xu, G.H. Tian, C.L. Liu, H.W. Hao, Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification, Neurocomputing 174 (2016) 806–814, http://dx.doi.org/10.1016/j.neucom.2015.09.096.
[12] J.M. Xu, B. Xu, P. Wang, S.C. Zheng, G.H. Tian, J. Zhao, B. Xu, Self-taught convolutional neural networks for short text clustering, Neural Netw. 88 (2017) 22–31, http://dx.doi.org/10.1016/j.neunet.2016.12.008.
[13] M.H. Arif, J.X. Li, M. Iqbal, K.X. Liu, Sentiment analysis and spam detection in short informal text using learning classifier systems, Soft Comput. 22 (21) (2018) 7281–7291, http://dx.doi.org/10.1007/s00500-017-2729-x.
[14] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, H. Xiong, Topic modeling of short texts: A pseudo-document view, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 13-17 Aug 2016, 2016, pp. 2105–2114.
[15] T. Shi, K. Kang, J. Choo, C.K. Reddy, Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations, in: Proceedings of the 2018 World Wide Web Conference, 23 April 2018, Lyon, France, 2018, pp. 1105–1114.
[16] Z.Y. Wang, H.X. Wang, Understanding short texts, in: The Association for Computational Linguistics (ACL) (Tutorial), 2016, pp. 1–18.
[17] S.M. Angeles, M. Boguna, Tuning clustering in random networks with arbitrary degree distributions, Phys. Rev. E 72 (3) (2005) 036133, http://dx.doi.org/10.1103/physreve.72.036133.
[18] J. Mneche, A. Valleriani, R. Lipowsky, Asymptotic properties of degree-correlated scale-free networks, Phys. Rev. E 81 (4) (2010) 046103, http://dx.doi.org/10.1103/PhysRevE.81.046103.
[19] R. van der Hofstad, A.J.E.M. Janssen, J.S.H. Leeuwaarden, C. Stegehuis, Local clustering in scale-free networks with hidden variables, Phys. Rev. E 95 (2) (2017) 022307, http://dx.doi.org/10.1103/PhysRevE.95.022307.
[20] Z.W. Wei, B.H. Wang, Emergence of fractal scaling in complex networks, Phys. Rev. E 94 (3) (2016) 032309, http://dx.doi.org/10.1103/PhysRevE.94.032309.
[21] S.H. Yook, F. Radicchi, H. Meyer-Ortmanns, Self-similar scale-free networks and disassortativity, Phys. Rev. E 72 (4) (2005) 045105, http://dx.doi.org/10.1103/PhysRevE.72.045105.
[22] C.M. Song, S. Havlin, H.A. Makse, Origins of fractality in the growth of complex networks, Nat. Phys. 2 (4) (2006) 275–281, http://dx.doi.org/10.1038/nphys266.
[23] H.F. de Arruda, L.D. Costa, D.R. Amancio, Using complex networks for text classification: Discriminating informative and imaginative documents, Europhys. Lett. 113 (2) (2016) 28007, http://dx.doi.org/10.1209/0295-5075/113/28007.
[24] L. Antiqueira, M.G.V. Nunes, O.N. Oliveira, L.D. Costa, Strong correlations between text quality and complex networks features, Physica A 373 (2007) 811–820, http://dx.doi.org/10.1016/j.physa.2006.06.002.
[25] M. Garg, M. Kumar, Identifying influential segments from word co-occurrence networks using AHP, Cogn. Syst. Res. 47 (2018) 28–41, http://dx.doi.org/10.1016/j.cogsys.2017.07.003.
[26] H.F. De Arruda, V.Q. Marinho, T.S. Lima, D.R. Amancio, L.D. Costa, An image analysis approach to text analytics based on complex networks, Physica A 510 (2018) 110–120, http://dx.doi.org/10.1016/j.physa.2018.06.110.
[27] 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, http://dx.doi.org/10.1371/journal.pone.0170527.
[28] D.R. Amancio, Comparing the topological properties of real and artificially generated scientific manuscripts, Scientometrics 105 (3) (2015) 1763–1779, http://dx.doi.org/10.1007/s11192-015-1637-z.
[29] D.R. Amancio, F.N. Silva, L.D. Costa, Concentric network symmetry grasps authors' styles in word adjacency networks, Europhys. Lett. 110 (6) (2015) 68001, http://dx.doi.org/10.1209/0295-5075/110/68001.
[30] H.F. De Arruda, V.Q. Marinho, T.S. Lima, D.R. Amancio, L.D. Costa, An image analysis approach to text analytics based on complex networks, Physica A 510 (2018) 110–120, http://dx.doi.org/10.1016/j.physa.2018.06.110.
[31] C.Y. Zheng, T. Usagawa, Short Chinese text classification based on correlation analysis, in: Proceedings of the 11th International Conference on Information and Communication Technology and System, ICTS, 31 Oct 2017, Surabayb, Indonesia, 2017, pp. 265–268.
[32] Y.H. Ning, L. Zhang, Y.R. Ju, W.J. Wang, S.Q. Li, Using semantic correlation of HowNet for short text classification, in: Proceedings of the International Conference on Advances in Materials Science and Information Technologies in Industry, AMSITI, 11-12 Jan 2014, Xian, China, 2014, pp. 1931–1934 https://doi.org/10.1007/s11192-015-1637-z.
[33] C.Y. Suen, N-gram statistics for natural-language understanding and text processing, IEEE T. Pattern Anal. 1 (2) (1979) 164–172, http://dx.doi.org/10.1109/TPAMI.1979.4766902.

[34] P.F. Liu, X.P. Qiu, X.J. Huang, Learning context-sensitive word embeddings with neural tensor skip-gram model, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI, 25-31 Jul 2015, Buenos Aires, Argentina, 2015, pp. 1284–1290.
[35] M. Jamaati, A. Mehri, Text mining by Tsallis entropy, Physica A 490 (2018) 1368–1376, http://dx.doi.org/10.1016/j.physa.2017.09.020.
[36] E.G. Altmann, G. Eduardo, M.D. Esposti, On the origin of long-range correlations in texts, Proc. Natl. Acad. Sci. USA 109 (29) (2012) 11582–11587, http://dx.doi.org/10.1073/pnas.1117723109.
[37] Y. Fujiki, T. Takaguchi, K. Yakubo, A general formulation of long-range degree correlations in complex networks, Phys. Rev. E 97 (2018) 062308, http://dx.doi.org/10.1103/PhysRevE.97.062308.
[38] D. Rybski, H.D. Rozenfeld, J.P. Kropp, Quantifying long-range correlations in complex networks beyond nearest neighbors, Europhys. Lett. 90 (2) (2010) 28002, http://dx.doi.org/10.1209/0295-5075/90/28002.
[39] J.W. Kantelhardt, E. Koscielny-Bunde, H.H.A. Rego, S. Havlin, A. Bunde, Detecting long-range correlations with detrended fluctuation analysis, Physica A 295 (3–4) (2001) 441–454, http://dx.doi.org/10.1016/S0378-4371(01)00144-3.
[40] D. Rybski, A. Bunde, On the detection of trends in long-term correlated records, Physica A 388 (8) (2009) 1687–1695, http://dx.doi.org/10.1016/j.physa.2008.12.026.
[41] J. Cong, H.T. Liu, Approaching human language with complex networks, Phys. Life Rev. 11 (2014) 598–618, http://dx.doi.org/10.1016/j.plrev.2014.04.004.
[42] D. Koutsoyiannis, Nonstationarity versus scaling in hydrology, J. Hydrol. 324 (1–4) (2006) 239–254, http://dx.doi.org/10.1016/j.jhydrol.2005.09.022.
[43] Q.D.Y. Ma, R.P. Bartsch, P. Bernaola-Galvan, M. Yoneyama, P.C. lvanov, Effect of extreme data loss on long-range correlated and anti-correlated signals quantified bydetrended fluctuation analysis, Phys. Rev. E 81 (3) (2010) 031101, http://dx.doi.org/10.1103/PhysRevE.81.031101.