Physics Letters A 379 (2015) 1627–1632

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Word ranking in a single document by Jensen–Shannon divergence

Ali Mehri [a,*], Maryam Jamaati [b], Hassan Mehri [c]

[a] Department of Physics, Faculty of Science, Noshirvani University of Technology, Babol, Iran
[b] Department of Physics, Iran University of Science and Technology, Tehran, Iran
[c] Department of Mathematics, Faculty of Science, Islamic Azad University, Tehran, Iran

ARTICLE INFO

Article history:
Received 1 March 2015
Received in revised form 18 April 2015
Accepted 20 April 2015
Available online 22 April 2015
Communicated by C.R. Doering

Keywords:
Word ranking
Text mining
Jensen–Shannon divergence
Entropy
Information theory

ABSTRACT

Ranking the words in human written texts, according to their relevance to text context, plays a crucial role in many text mining tasks. Highly relevant words concentrate in some limited areas, while the irrelevant ones have nearly random spatial distribution throughout the text. But in the randomly shuffled version of the text, all word types are distributed at random. The difference between spatial distribution of words in the original version of a text and its shuffled version seems a proper criterion for word relevance ranking. In this procedure, spatial distribution of each word type in the document is defined by box counting method. Then we apply Jensen–Shannon divergence to measure the difference between probability distributions of each word in the original text and its shuffled version. This metric properly distinguishes relevant words from irrelevants without requiring any previous knowledge about text structure.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Rapidly growing demands for information retrieval from natural and artificial languages strongly motivate a wealth of scientific research on data mining. Human language is one of the important manifestations of natural languages, and its emergence could be regarded as a significant transition in hominid evolution [1]. It has special importance in human communication, culture, and even intelligence [2]. Human beings employ language as an adaptive equipment to communicate and express their opinions. The brain has a restricted capacity to store lexicons [3]. On the other hand, human needs unlimited concepts to achieve a successful communication. The increasing need for new concepts is resolved by establishing complex syntactic and semantic relations between limited set of stored words and symbols. The grammatical and semantic complexity of human language discriminates between mankind and other species.

A great deal of human knowledge has been included in the written part of language. Several universal features establish complexity of human written texts [4,5]. A written text can be considered as a symbolic sequence. Many natural symbolic sequences such as language, music, genetic codes and neural signals are commonly applied for information conveyance. Due to the increasing interest for automatic information retrieval from textual data, much research has been made on automatic text analysis. Typical text mining tasks include automatic translation, summarization, classification, authorship analysis, keyword extraction, *etc.* One of the key steps in almost all text mining tasks is ranking the words via their relevance according to text context.

Different approaches have been presented for ranking the words in the literature [6]. Luhn in his seminal work computed a relative measure of significance for words by taking advantage of Zipf's law [7]. He ignored the most frequent and the rarest words, and picked the middle ones as the relevant words. Many other scientists have employed distance between two successive word tokens, as statistical information, for words importance ranking [8–12]. In this regard, standard deviation of distance between consecutive occurrences of a term is successfully applied in ranking process [8,10,13]. Carpena and his colleagues found that the important words have greater standard deviation, because the spatial distribution of relevant words is more inhomogeneous than that for common irrelevant terms. In a work of different nature, Mihalcea and co-workers represented the text by a graph; word types are graph nodes connecting the co-occurrences by links. They introduced TextRank metric similar to google's PageRank, by exploiting complex network theory, to classify the semantic level of words in the text [14]. Keywords tend to form clusters just in the ambient regions of the analyzed text. Zhou and Slater proposed the average of cluster indices as an index to detect relevant words [9].

* Corresponding author. Tel.: +98 1132332071; fax: +98 1132334203.
E-mail address: alimehri@nit.ac.ir (A. Mehri).

http://dx.doi.org/10.1016/j.physleta.2015.04.030
0375-9601/© 2015 Elsevier B.V. All rights reserved.

Word ranking problem can be handled by entropic criteria. The first entropic measure has been proposed to quantify information content of word types in structured texts [15]. Herrera and Pury defined probability distribution for words according to their relative frequency in different chapters of text. Mehri and Darooneh introduced some other entropic measures, have in mind relative distances between consecutive occurrences of a word type in the text [12]. They also employed nonextensivity parameter, borrowed from nonextensive statistical mechanics, for word relevance ranking process [11]. In this case, nonextensivity parameter reveals long-range correlation in the spatial distribution of relevant word types. Recently, Yang et al. classified the distances between two successive occurrences of a word type into two groups: intra-cluster (intrinsic) and inter-cluster (extrinsic) [16]. They attributed a probability to intra-cluster and inter-cluster distribution of words, and then applied the entropy difference between the intrinsic and extrinsic modes to tackle this problem.

In this work, we will try to rank the words according to their relevance to text subject, taking advantage of Jensen–Shannon divergence (JSD). We apply box counting method to define a probability distribution for words in the text. The difference between spatial distribution of a word in the original text and its shuffled version is measured by JSD. The basic assumption is that, the relevant words are distributed in a special manner to imply author's purpose. These important words tend to concentrate in certain regions of the text and make clusters, while the irrelevant words commonly spread uniformly throughout the text. Consequently, we expect the spatial distribution of relevant words significantly differs in the original text and its shuffled version. On the contrary, the spatial distribution of irrelevant words, in the both of original and random shuffled texts are alike. In other words, JSD values for relevant words should be greater than its values for irrelevant ones.

The paper is organized as follows. Section 2 is devoted to describe the theoretical background of JSD. Then we will explain text segmentation method to ascribe spatial probability to word types and calculate JSD for them in Section 3. In the fourth section, JSD value is applied to rank the words in some representative documents, according to their relevancy with the notion of text. Then the results are presented, and followed by the related discussions. Finally we end the paper with a conclusion.

2. Entropy and Jensen–Shannon divergence

Entropy, as a key concept in the realm of statistical physics, is applied to extract macroscopic properties of natural and artificial systems from their microscopic details. Clausius introduced the thermodynamical definition of entropy. Afterwards, its first statistical description was represented by Boltzmann and completed by Gibbs:

$$E(P) = \sum_{i=1}^{M} p_i \log(p_i^{-1}), \tag{1}$$

where probability distribution, $P = \{p_1, p_2, \ldots, p_M\}$, contains occurrence probability of all microstates. M denotes the total number of microstates corresponding to a single macrostate of many body system and p_i is the occurrence probability of ith state. From this statistical point of view, entropy can be interpreted as the degree of order/disorder in the many body systems. Later, Hartley and Shannon derived the same form for uncertainty from the axioms of information theory. Intuitively, entropy measures amount of information or uncertainty in a random variable related to a natural process. Zero value for entropy represents only one certain outcome for a random variable. On the other hand, if all outcomes are equally likely, entropy will have its maximal value [17].

The relative entropy or divergence is a measure of the distance between two probability distributions. Various divergences are defined as natural measures of (dis)similarity between distributions. The most important measure of divergence in information theory is Kullback–Leibler divergence (KLD) between two probability distributions P and Q of a random variable [18]. In statistics, it arises as an expected logarithm of the likelihood ratio. The relative entropy, $KLD(P||Q)$, is a measure of the inefficiency of assuming that the distribution is Q, when the true distribution is P [17]:

$$KLD(P||Q) = \sum_{i=1}^{M} p_i \log(\frac{p_i}{q_i}). \tag{2}$$

This functional, which is known in classical information theory as a cross-entropy or a directed divergence, measures uncertainty in relative rather than absolute terms [19]. Relative entropy is always non-negative and is zero if and only if $P = Q$. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality [17].

Jensen–Shannon divergence (JSD) is a symmetrized, finite and smoothed version of Kullback–Leibler divergence [21,22], which is defined as

$$JSD(P||Q) = \frac{1}{2}KLD(P||M) + \frac{1}{2}KLD(Q||M), \tag{3}$$

where $M = (P + Q)/2$. In probability theory and statistics, JSD is applied as a popular method to measure the similarity between two probability distributions. It is also known as information radius (IRad) or total divergence to the average. The JSD is bounded by 1, given that one uses the base 2 logarithm $0 \leq JSD(P||Q) \leq 1$. For Napierian logarithm, which is commonly used in statistical mechanics, the upper bound is $\ln(2)$: $0 \leq JSD(P||Q) \leq \ln(2)$ [20].

3. Words relevance ranking by Jensen–Shannon divergence

Authors arrange the words in the text and spread them in a particular manner to convey their message. As stated before, significant words, which are relevant to text subject, appear in specific parts of text to imply the considered idea. On the contrary, the unimportant terms (*i.e.*, articles, prepositions, conjunctions, *etc.*) are distributed almost homogeneously along the text by grammatical necessities. From physical point of view, it seems that in the meaningful natural and artificial texts, relevant words attract each other and tend to make clusters in ambient regions of the system. While the irrelevant words do not interact between themselves, and thus they appear randomly in the text since they do not feel each other. As a result, they set a nearly random spatial distribution.

Now let us assume a random shuffled version of the text under consideration. In shuffling process all words of original text are distributed randomly to create the shuffled version. It is clear that, all word types are distributed randomly in the shuffled text. Both of the relevant and irrelevant words in this shuffled version have rather homogeneous spatial distribution. Consequently, the spatial distribution of important words in the original text and its shuffled version differ too much. Whereas, they are almost analogous to each other in the case of unimportant word types. This feature can be used in words ranking by their relevance to text context. We will exploit JSD to reveal the discrepancy between spatial distributions of words in the original text and its shuffled version.

First of all, it is necessary to define probability for spatial distribution of words in the text. We apply the so called box counting method, which is formerly used to calculate the dimension in some fractal systems. In this procedure, a text with length L_t is partitioned to N_l boxes with equal size l: $N_l = [L_t/l]$, where $[x]$ returns integer part of x. The spatial probability of a

Watch your thoughts; they become words. Watch your words; they become actions. Watch your actions; they become habits. Watch your habits; they become character. Watch your character; it becomes your destiny.

Fig. 1. A nice quote with length $L_t = 31$ is segmented to $N_4 = [31/4] = 7$ partitions with size $l = 4$. The word *'your'* appears in $n_4(your) = 5$ boxes. Hence the probability of spatial distribution for *'your'* by partitioning with boxes of size $l = 4$ will be proportional to 5/7 ($p_4(your) \propto 5/7$).

word type w is defined as $p_l(w) \propto n_l(w)/N_l$, where $n_l(w)$ denotes the number of boxes with length l which contain the word w. In Fig. 1, we illustrate probability calculation for spatial distribution with a simple example. Therein Fig. 1, a short quote, with length $L_t = 31$, is partitioned to $N_4 = [31/4] = 7$ sections with equal length of 4. The word *'your'* occurs in 5 boxes out of total 7 boxes which cover whole sample text. The spatial probability of this word type, when the text is partitioned to boxes with size $l = 4$, can be defined by using the relative number of boxes which it appears in them: $p_4(your) \propto n_4(your)/N_4 = 5/7$. It is worth noting that $n_l(w)/N_l(w) = 1$ means that the word w is homogeneously distributed in the text partitions with size l. As discussed before, our basic assumption is that highly relevant words should be concentrated in some limited areas. Conversely, a common word should occur homogeneously everywhere in the meaningful texts. But in the random shuffled texts, both of relevant and irrelevant terms are distributed randomly. Let $P(w) = \{p_1(w), p_2(w), \ldots, p_l(w)\}$ and $Q(w) = \{q_1(w), q_2(w), \ldots, q_l(w)\}$ respectively denote the spatial probability distribution sets of word w in the original text and its shuffled version. Here $p_l(w)$ and $q_l(w)$ refer to spatial probability of word w in the original text and its shuffled version when are partitioned with boxes of size l. The spatial distribution of a word with frequency $f(w)$ can be calculated theoretically by using random distribution hypothesis:

$$q_l(w) \propto \begin{cases} \frac{f(w)}{N_l(w)} & \text{if } f(w) < N_l(w) \\ 1 & \text{otherwise} \end{cases}$$

In this way, we can calculate the JSD without a random shuffled version of text, with less computation complexity. P and Q should be normalized by a normalization factor. Now, it is convenient to write Eq. (3) in a new form:

$$JSD(P(w)||Q(w)) = \frac{1}{2}\sum_{l=2}^{L/2} p_l(w)\log\left(\frac{p_l(w)}{m_l(w)}\right) + \frac{1}{2}\sum_{l=2}^{L/2} q_l(w)\log\left(\frac{q_l(w)}{m_l(w)}\right), \quad (4)$$

where $m_l(w) = (p_l(w) + q_l(w))/2$ and $L_t/2$ is maximum size of text partitions. Summation index, l, refers to box size in text partitioning process.

In the next section, we will check if JSD method is able to discriminate between relevant and irrelevant words. In this way, we will try to rank the words in terms of their relation with text subject in two representative texts, using JSD between probability distribution of words in the original text and its shuffled version.

4. Experimental evaluation

For the evaluation of the proposed ranking method, we adopt two representative texts, and make use of JSD to rank their words' information content. The book *The Origin of Species* by Charles Darwin as a prototypical real long text, along with the paper *The Role of Entropy in Word Ranking* [12] by Mehri and Darooneh as a representative short text, are used to evaluate the performance of JSD in word ranking.

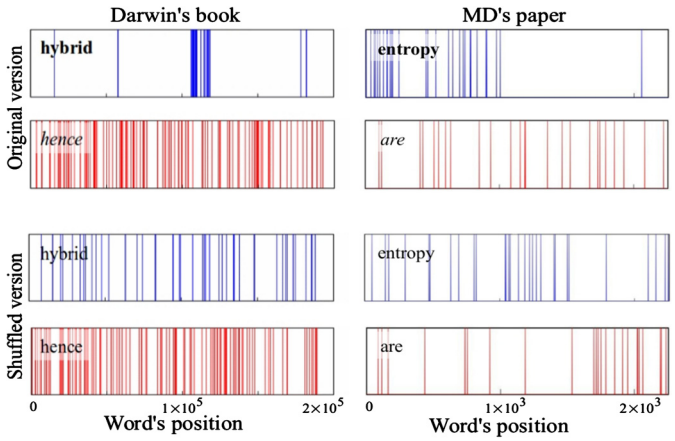

Fig. 2. Absolute positions of a pair of relevant–irrelevant words in Darwin's book (left panel) and MD's paper (right panel) counted from their beginning. The sketched spectra belong to (**hybrid**, *hence*) from Darwin's book, and (**entropy**, *are*) from MD's paper. The spectrum of relevant/irrelevant words are distinguished by blue/red color. The important words, **hybrid** and **entropy**, tend to form clusters in some regions of original texts (blue graphs in upper panels). Whereas the unimportant ones, *hence* and *are*, are spread haphazardly throughout the texts (red graphs in upper panels). Lower panels show spectra of the mentioned words in random shuffled version of texts under consideration. Both of relevant and irrelevant words are distributed almost homogeneously in shuffled documents. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the pre-processing step, we first withdrew all non-alphabetic characters, *i.e.*, numbers and punctuation symbols. Moreover, the table of contents, the glossary and the analytical index were removed from the texts. All words were changed to lowercase and then a simple tokenization method based on whitespaces was applied [23]. In our experiments, the word type is defined as any different string of letters between two whitespaces. Thus words like *organ* and *organs* correspond to different word types. After pre-processing, Darwin's book has $N_t = 191\,525$ words, and its vocabulary includes $N_v = 8535$ word types. The length of Mehri and Darooneh's paper (MD paper) is $N_t = 2894$, and it contains $N_v = 625$ distinct words. Fig. 2 displays the most significant difference between spatial distributions of relevant and irrelevant word types in meaningful texts. We pick a relevant–irrelevant pair of words from each above mentioned corpora, and depict their spectra in original (upper panels) and shuffled (lower panels) version of texts. (**hybrid**, *hence*) with almost 120 occurrences from Darwin's book as a long text, and (**entropy**, *are*) with almost 30 occurrences from MD's paper as a short text, are our choices to illustrate their spectra. Blue and red graphs display spectra for important and unimportant words, respectively. It is clear that, in the original version of human written texts highly relevant words concentrate in some special portions; while common and typically irrelevant words are almost homogeneously distributed along the texts. But in the shuffled version of aforesaid texts both relevant and irrelevant words are spread randomly in the texts. JSD is able to quantify (dis)similarity between spatial probability distributions of a word type in the original text and its shuffled version. Therefore, it can be used to discriminate between relevant and irrelevant words.

We calculate JSD between spatial probability distributions for all word types in the original and shuffled version of two representative texts. Then we sort all word types according to their JSD value in descending order. Highly relevant words are positioned in the beginning of this sorted list. The behavior of words' JSD versus their normalized rank is plotted in Fig. 3. We ignore data points with zero values of JSD due to mathematical difficulties in preparing log–log graph. Blue diamonds and red squares respectively indicate JSD data for Darwin's book and MD's paper words. It can be seen that, JSD-rank data obey Zipf-like law with two

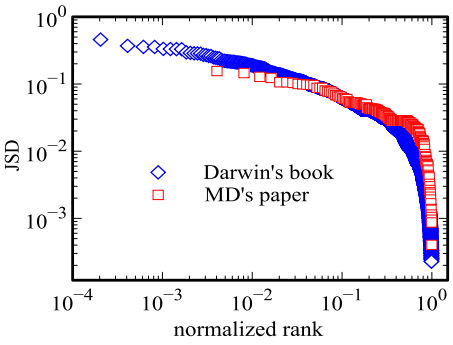

Fig. 3. Jensen–Shannon divergence of words as a function of their normalized relevancy rank, for Darwin's book (blue diamonds) and MD's paper (red squares). JSD-rank data obey Zipf-like law with two different power law exponents; A gentle slope region for relevant words and a steep incline for irrelevant ones. JSD value for the most relevant word is four order of magnitude greater than its value for the most irrelevant word. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Most relevant/irrelevant words extracted from the book *The Origin of Species*, by Charles Darwin, using Jensen–Shannon divergence (JSD). The first column contains 10 words with highest JSD values and the third column includes 10 words with lowest nonzero JSD values. The relevant words (keywords) highlighted by bold fonts. We also report JSD value for each word. It is clear that, the most retrieved keywords closely relate to the subject of the book. The most irrelevant words, which appear in italic fonts are prepositions, conjunctions, articles, *etc.* and there is no apparent connection between these words and the topic of document.

Relevant	JSD	Irrelevant	JSD $\times 10^4$
wax	0.44	*it*	0.62
lamellae	0.36	*simultaneous*	0.61
sterility	0.34	*society*	0.58
cuckoo	0.33	*other*	0.53
pedicellariae	0.33	*in*	0.51
slaves	0.33	*progenitors*	0.46
spheres	0.32	*all*	0.39
neuters	0.30	*few*	0.39
cell	0.30	*not*	0.38
vibracula	0.29	*to*	0.32

Table 2
Ten most relevant words and ten most irrelevant words extracted from the paper *The role of entropy in word ranking* by Mehri and Darooneh (Ref. [12]), using Jensen–Shannon divergence (JSD). The relevant/irrelevant words are determined by bold/italic fonts. The second and the fourth column contain JSD values for all tabulated words.

Relevant	JSD	Irrelevant	JSD $\times 10^2$
entropy	0.16	*words*	0.21
type	0.15	*means*	0.19
precision	0.13	*out*	0.17
information	0.12	*centrality*	0.14
bg	0.11	*an*	0.13
values	0.11	*slope*	0.11
random	0.10	*sd*	0.10
hp	0.10	*matrix*	0.09
occurrences	0.10	*given*	0.08
fmeasure	0.09	*its*	0.04

different power law exponents. The first part of JSD-rank graph, with a small exponent, belongs to important words with significant information content. There are $N_{ret} \simeq 1500$ ($N_{ret} \simeq 85$) words in this part of the graph for Darwin's book (MD's paper), which can be selected as retrieved glossary. The tail of the graph, with greater exponent, belongs to unimportant words with poor information content. It is worth remarking that, there is roughly three order of magnitudes difference between JSD values for most relevant and most irrelevant word types.

Tables 1 and 2 contain our results in word relevancy ranking by JSD. Ten most relevant words and ten most irrelevant words of

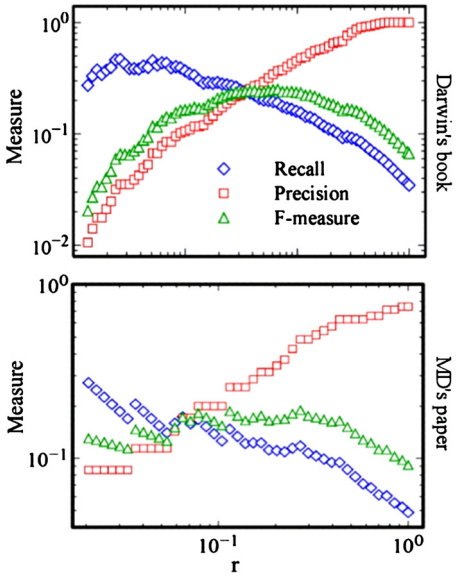

Fig. 4. Recall (blue diamonds), precision (red squares) and F-measure (green triangles) as functions of normalized length of retrieved glossaries, $r = N_{ret}/N_v$, for (upper) Darwin's book and (lower) MD's paper. F-measure reaches its maximum value at $r = 0.07$ and $r = 0.25$ fractions of sorted list for Darwin's book and MD's paper, respectively. Therefore, the best glossaries for these texts contain first $N_{ret} = 0.07N_v$ and $N_{ret} = 0.25N_v$ words from the beginning of the sorted list by using JSD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Darwin's book and MD's paper are listed in Tables 1 and 2, respectively. In each table, the first column includes the relevant words with bold fonts, and the third column displays irrelevant words with italic fonts. The most relevant words in the first column have highest JSD values. The listed irrelevants are those word types with lowest nonzero JSD values. It should be noted that, we ignore many unimportant terms and misprints, with frequency equal to 1 or 2, which get a zero value for JSD. It is easy to see that, the reported JSD values for most relevants are some order of magnitudes greater than its values for most irrelevants.

Recall, precision and F-measure are defined for quantitative evaluation of ranking methods [23].

$$R = \frac{N_{rel \bigcap ret}}{N_{rel}}, \tag{5}$$

$$P = \frac{N_{rel \bigcap ret}}{N_{ret}}, \tag{6}$$

$$F = \frac{2RP}{R+P}, \tag{7}$$

where N_{rel} is total number of relevant words in the prepared benchmark glossary and N_{ret} denotes the number of words within the retrieved glossary of text. $N_{rel \bigcap ret}$ stands for number of words within intersection set of relevant and retrieved lists. We use a prepared glossary with $N_{rel} = 283$ relevant words as a benchmark to evaluate accuracy of retrieved relevant words from Darwin's book [15]. For MD's paper, we make a glossary containing all important words of title, keyword list and abstract of the paper after elimination of their common words. This handy prepared glossary has $N_{rel} = 35$ relevant words. The retrieved glossary of a text can be prepared by picking first N_{ret} important words from the beginning of sorted list of text words according to their relation to text context. The best choice for N_{ret} value maximizes the evaluation criterion such as F-measure. We select different fractions ($r = N_{ret}/N_v$) from beginning of the sorted list of words as retrieved relevant words to obtain recall, precision and F-measure. Fig. 4 illustrates recall, precision and F-measure as a function of

Table 3
Comparison of the results of ranking some selected words of the book *The Origin of Species* written by Charles Darwin, using JSD, ED^2_{nor}, TextRank and C. The selected words (pairs of relevant–irrelevant) with their frequency are listed in the first column. Keywords are highlighted in bold. It is clear that JSD shows acceptable ranking results like previous methods.

Frequency/Word	JSD	ED^2_{nor}	TextRank	C
1749/**species**	**ants**	**organs**	*by*	**species**
1669/*by*	**organs**	**organic**	**species**	**plants**
520/**forms**	**genera**	**inhabitants**	*its*	**forms**
516/*when*	**inhabitants**	**plants**	*when*	**selection**
497/*its*	**birds**	**modification**	**forms**	**structure**
492/**selection**	*why*	**selection**	*than*	**organs**
475/**natural**	**plants**	**birds**	**natural**	**genera**
451/*than*	**forms**	**structure**	**selection**	**natural**
390/**plants**	**system**	**genera**	**plants**	**inhabitants**
384/*thus*	**selection**	**ants**	*thus*	**organic**
306/**nature**	**conditions**	**natural**	*under*	**ants**
305/*under*	**organic**	**forms**	**nature**	**conditions**
295/**conditions**	**modification**	**conditions**	*now*	**birds**
290/*now*	**structure**	**system**	**conditions**	*why*
263/**structure**	**species**	*why*	**structure**	**modification**
260/*do*	**natural**	*together*	*do*	*thus*
211/**genera**	*its*	*present*	*might*	*when*
211/*might*	*seem*	*now*	**birds**	**system**
181/**birds**	*under*	**animal**	*together*	**animal**
180/*present*	*now*	**nature**	**organs**	**nature**
179/**organs**	*together*	*always*	**genera**	*than*
174/*whole*	**animal**	*might*	*present*	*by*
158/*together*	*present*	*do*	*whether*	*under*
157/**organic**	*might*	*when*	*whole*	*now*
156/**modification**	*when*	*whole*	**modification**	*might*
155/*why*	**nature**	*under*	*why*	*do*
138/*whether*	*whether*	*thus*	**organic**	*whole*
137/**inhabitants**	*always*	*whether*	**inhabitants**	*together*
108/**animal**	*do*	*than*	*always*	*present*
108/*always*	*unless*	*seem*	**animal**	*always*
50/**system**	*whole*	*its*	*seem*	*thus*
50/*seem*	*than*	*unless*	**system**	*seem*
30/**ants**	*thus*	*by*	**ants**	*unless*
30/*unless*	*by*	**species**	*unless*	*whether*

retrieved list fraction for Darwin's book (upper panel) and MD's paper (lower panel). For Darwin's book, the maximum value of F-measure, $F_{max} = 0.25$, is located at $r = 0.07$ of list fraction. This means that, the best choice for retrieved glossary of this book is a list which includes first $N_{ret} = 0.07N_v$ of entries in the sorted vocabulary list by JSD. In the case of MD's paper, we have $F_{max} = 0.20$ in $r = 0.25$. Attaining the greater F-measure value in smaller list fraction for Darwin's book shows that the presented method has better ranking results in long documents.

To compare the performance of JSD method with previous methods, we apply them to rank the 34 selected representative words of Darwin's book [11,16]. These adopted words are 17 pairs of relevant–irrelevant words with almost same frequency of occurrence. Each of the pairs are chosen randomly from different frequency ranges. Table 3 represents ranking results by JSD, ED^2_{nor} [16], TextRank [14] and C [10]. The relevants/irrelevants are highlighted in bold/italic fonts. The first column shows the selected words and their frequency. One can see that our new proposal has reliable results in words ranking process.

For a quantitative comparison between JSD and some other important word ranking methods, we calculate recall, precision and F-measure for them by the method which is introduced in Herrera and Pury's paper [15]. This method needs not N_{ret} to calculate recall and precision. The evaluation results are reported in Table 4. It contains recall (R), precision (P) and F-measure (F) for ranking the word of Darwin's book, as a representative long document, and MD's paper, as a representative short document, by JSD, ED^2_{nor}, TextRank and C. JSD has acceptable ranking results in comparison with the other ranking methods.

It can be easily seen that, JSD properly ranks the words according to their relevance to text subject, without any prior information. We do not take any advantage of text structure in ranking procedure. In other words, one may use this method to extract keywords from unstructured documents.

Table 4
Recall, precision and F-measure for extracted indices by JSD, ED^2_{nor}, TextRank and C for Darwin's book as a long document and for MD's paper as a short document. JSD has one of the highest values for all evaluation parameters in both long and short documents.

Metric	Text	R	P	F
JSD	Book	0.20	0.06	0.09
	Paper	0.09	0.08	0.08
ED^2_{nor}	Book	0.09	0.07	0.08
	Paper	0.03	0.23	0.05
TextRank	Book	0.15	0.13	0.14
	Paper	0.17	0.10	0.13
C	Book	0.42	0.04	0.07
	Paper	0.08	0.06	0.07

5. Conclusion

In summary, in this article we introduce Jensen–Shannon divergence (JSD) as a prolific statistical criterion for word relevance ranking in unstructured documents, without any previous information about the text. JSD quantifies the (dis)similarity between two probability distributions. Here we apply JSD to measure the distance between spatial distribution of words in the original version of a text and in its shuffled version. Words are spread randomly in the shuffled texts. While in the meaningful documents, important words make clusters in ambient regions, and unimportant

words distribute almost homogeneously everywhere. Therefore the distance between spatial distribution of the relevant words in a meaningful text and in its shuffled version will be greater than that of the irrelevants. This means that, the highly relevants have greater JSD values.

We checked JSD capability in words' importance ranking process. JSD successfully ranked the words of two representative texts according to their relevance to their subject. The text structure does not play any significant role in this ranking method. Hence, one may use this criterion to automatically detect and rank the relevant words of an unstructured document without any a priori information.

JSD can be applied in principle to term ranking in all types of natural and artificial languages, such as protein chains, DNA sequences, time series, *etc.*

References

[1] J.M. Smith, E. Száthmáry, The Major Transitions in Evolution, Oxford University Press, Oxford, 1997.
[2] M.A. Montemurro, D.H. Zanette, Entropic analysis of the role of words in literary texts, Adv. Complex Syst. 5 (2002) 7–17.
[3] S. Romaine, The evolution of linguistic complexity in pidgin and creole languages, in: J.A. Hawkins, M. Gell-Mann (Eds.), The Evolution of Human Languages, Addison-Wesley, Redwood City, 1992, pp. 213–238.
[4] R. Ferrer i Cancho, R.V. Solé, The small world of human language, Proc. R. Soc. Lond. B, Biol. Sci. 268 (2001) 2261–2265.
[5] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, Proc. R. Soc. Lond. B, Biol. Sci. 268 (2001) 2603–2606.
[6] M.W. Berry, J. Kogan, Text Mining Applications and Theory, Wiley, New York, 2010.
[7] H.P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (1958) 159–165.
[8] M. Ortuño, P. Carpena, P. Bernaola-Galvàn, E. Muñoz, A.M. Somoza, Keyword detection in natural languages and DNA, Europhys. Lett. 57 (2002) 759–764.
[9] H. Zhou, G.W. Slater, A metric to search for relevant words, Physica A 329 (2003) 309–327.
[10] P. Carpena, P. Bernaola-Galvàn, M. Hackenberg, A.V. Coronado, J.L. Oliver, Level statistics of words: finding keywords in literary texts and symbolic sequences, Phys. Rev. E 79 (2009) 035102(R).
[11] A. Mehri, A.H. Darooneh, Keyword extraction by nonextensivity measure, Phys. Rev. E 83 (2011) 056106.
[12] A. Mehri, A.H. Darooneh, The role of entropy in word ranking, Physica A 390 (2011) 3157–3163.
[13] C. Carretero-Campos, P. Bernaola-Galván, A.V. Coronado, P. Carpena, Improving statistical keyword detection in short texts: entropic and clustering approaches, Physica A 392 (2013) 1481–1492.
[14] R. Mihalcea, Random walks on text structures, in: A. Gelbukh (Ed.), CICLing, in: Lect. Notes Comput. Sci., Springer, Heidelberg, 2006, pp. 249–262.
[15] J.P. Herrera, P.A. Pury, Statistical keyword detection in literary corpora, Eur. Phys. J. B 63 (2008) 135–146.
[16] Z. Yang, J. Lei, K. Fan, Y. Lai, Keyword extraction by entropy difference between the intrinsic and extrinsic mode, Physica A 392 (2013) 4523–4531.
[17] T. Cover, J. Thomas, Elements of Information Theory, John Wiley and Sons, New York, 1991.
[18] M. Mezard, A. Montanari, Information, Physics and Computation, Oxford University Press, Oxford, 2009.
[19] G.J. Klir, Uncertainty and Information, John Wiley and Sons, New Jersey, 2006.
[20] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory 37 (1991) 145–151.
[21] D.M. Endres, J.E. Schindelin, A new metric for probability distributions, IEEE Trans. Inf. Theory 49 (2003) 1858–1860.
[22] F. Ôsterreicher, I. Vajda, A new class of metric divergences on probability spaces and its statistical applications, Ann. Inst. Stat. Math. 55 (2003) 639–653.
[23] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, 1999.