

Keyword detection in natural languages and DNA

M. ORTUÑO¹, P. CARPENA², P. BERNAOLA-GALVÁN²,
E. MUÑOZ³ and A. M. SOMOZA¹

¹ *Departamento de Física, Universidad de Murcia - Murcia, Spain*

² *Departamento de Física Aplicada II, ETSI de Telecomunicación
Universidad de Málaga - Málaga, Spain*

³ *Facultad de Documentación, Universidad de Murcia - Murcia, Spain*

(received 20 July 2001; accepted in final form 30 November 2001)

PACS. 89.20.-a – Interdisciplinary applications of physics.

PACS. 89.70.+c – Information science.

Abstract. – We show that words in a text present long-range frequency fluctuations due to a strong self-attraction, that is directly related to the relevance of the term to the text considered. The standard deviation of the distance between successive occurrences of a word is an excellent parameter to quantify this self-attraction, and provides us with an effective tool for automatic keyword extraction. DNA sequences also present the same features: “words”, for example codons in the coding part of the sequences, attract between themselves.

A key problem in documentation science is how to extract the relevant words of a text from their statistical properties. Present techniques are mainly based on an analysis of frequency occurrences of words in the text, as first proposed by Luhn [1]. He also noticed that it is convenient to exclude as keywords both common and rare words, and proposed that the ability of words to discriminate content can be represented by a function of the rank peaked at intermediate values [1]. For a collection of documents, modern-term weighting schemes normally use a factor proportional to the frequency of occurrence of a term in a document and another factor of the form $\log(N/n)$, where n/N is the proportion of documents containing the term [2]. Following a different approach, the probabilistic model of information retrieval related the significance of a term to its frequency fluctuations between documents [3–5]. The main problems of this approach are the dependence of these fluctuations on the distribution of document lengths, and the need of a collection of documents instead of a single one to perform the analysis.

The frequency-based (or similar) methods above referred to detect keywords may fail when applied to texts for which no *a priori* information is available, either because words relevant to a text cannot be relevant at all for a different text, or because words relevant to several texts can appear with very different frequencies. Here, we show that the spatial information of a word, *i.e.*, the way in which it is distributed along the text (independently of its relative frequency), is very important to quantify the relevance of the word to the text considered.

To study the spatial distribution of a word, we propose to calculate the nearest-neighbor spacing distribution $p(x)$ of a word in the following way: we number successively all the words

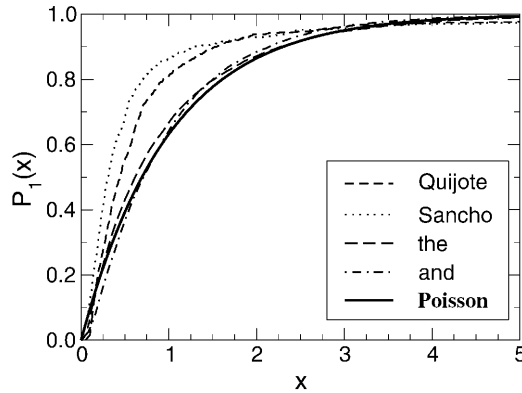


Fig. 1 – Accumulated distribution function of several words of *The Quijote* as a function of the normalized separation: “Quijote” (dashed), “Sancho” (dotted), “the” (long dashed) and “and” (dot-dashed). The thick continuous line corresponds to Poisson distribution.

of a given text and extract the positions corresponding to a given word. Then, for each word, we concentrate on the statistical properties of the set of distances $\{x_i\}$ between successive occurrences. We define $p(x)$ as the relative frequency of occurrence of a given separation x , and $P_1(x) = \sum_{x'=1}^x p(x')$ as its integrated distribution function. Both functions contain the same information, but the latter is more useful in terms of numerical precision. This approach is the same as the one usually adopted in the study of the level statistics of the spectrum of quantum disordered systems, according to the Random Matrix Theory [6]. In this latter case, by studying the properties of $p(x)$ (or $P_1(x)$), where now x is the energy spacing between consecutive levels, one can characterize different localization regimes [7].

To eliminate the dependence on frequency for different words, it is convenient to normalize separations for each word, *i.e.*, to measure them in units of their corresponding mean value, \bar{x} . From now on we will always use normalized variables, $s = x/\bar{x}$. If the words were distributed at random, the integrated distribution P_1 for each word in the continuum limit would be a Poisson distribution:

$$P_1(s) = 1 - \exp[-s]. \quad (1)$$

If a word repels itself, its distribution P_1 will be smaller than Poisson’s for $s < 1$. On the contrary, if a word attracts itself, it will tend to form clusters and P_1 will be larger than Poisson’s at short distances.

In fig. 1 we represent P_1 as a function of s for four different words of *The Quijote* [8]: “Quijote”, “Sancho”, “the” and “and”. The first two words are supposed to be very relevant to the text considered, while the last two are not. The thick continuous line corresponds to Poisson distribution. Figure 1 shows that, while non-specific words run very close to the Poisson curve (*i.e.*, they are distributed at random along the text), content-bearing terms go above this curve for distances smaller than the average separation. These results are independent of the frequency of the words and fairly typical for all the texts studied. It is easy to find an explanation for this behavior: words relevant to a text will normally appear in a very specific context, concentrated in a region of the text, presenting large frequency fluctuations. For example, for a general-physics textbook the word “magnetic” is a relevant keyword. It will appear very often in the context of electromagnetism and relatively seldom elsewhere. Relevant terms will usually contribute to determine the subject of an area of text, and within this area they will very likely appear again. This clustering is more important than

TABLE I – Words with larger σ in *The Bible*.

Word	Frequency	σ	Word	Frequency	σ
jesus	983	24.18	david	1064	8.86
christ	571	18.42	king	2542	8.15
paul	162	11.56	pharisees	87	8.06
disciples	244	10.88	jeremiah	148	8.00
peter	164	10.17	gospel	104	7.91
joab	145	10.03	solomon	305	7.67
faith	247	9.34	mordecai	60	7.45
saul	420	9.17	esther	57	7.43
absalom	108	9.12	joshua	217	7.42
john	137	9.03	elisha	58	7.39

what the previous example could suggest, being observed even for words appearing along the whole text, like “Quijote” and “Sancho”.

The attraction between words was previously reported on the context of speech recognition [9], but the crucial relationship between the significance of a word and the strength of its self-attraction was not established.

Once we have shown that $P_1(s)$ is useful to characterize the relevance of a word to a text, it is convenient to provide a simpler method to automatic keyword extraction. Note that the computation of all the $P_1(s)$ functions associated to all the words of a text can be very time consuming computationally speaking. So it is in practice better to use the standard deviation $\sigma = \sqrt{s^2 - \bar{s}^2}$, which is the simplest possible variable to characterize a normalized distribution and its fluctuations. For a Poisson distribution $\sigma = 1$, while if there is attraction $\sigma > 1$ and in the case of repulsion $\sigma < 1$. σ is very easy to calculate and robust against fluctuations.

We have calculated the standard deviation σ for all words of several texts, including novels, scientific books and articles. The trend is always the same: large σ values generally correspond to terms relevant to the text considered. As an example, in table I we show the 20 words with highest σ in *The Bible* [8]. It is clear that the words in this table are fairly representative of the book. More results for this and other texts can be found in [10]. All of them have been obtained directly, without resort to stop-lists, thesaurus or corpus information, as required by most techniques.

In order to explore in more detail our association of large σ values with high degrees of relevance, we represent in fig. 2 every word of a book on chaos [11] as a point in a two-dimensional σ -frequency space. The open circles correspond to words in the index (selected by the authors and assumed by us to be the relevant terms of the book), and the solid dots to the rest of the words. It is clear that open circles tend to lie to the right of the figure. Moreover, many of the terms with large sigma values not included in the index are anyway good keyword candidates [10], because they are words too frequent to be included in the index, but nevertheless are typical of the topic of the book. The previous tendency can be appreciated in a more quantitative way in the inset of the figure, where we plot the probability density function of σ for all terms of the book (\square) and for the index terms only (\bullet). Our parameter, σ , behaves, on average, similarly to Luhn’s resolving power [1] as a function of frequency, but it also discriminates the degree of relevance among terms of similar frequencies. Thus, we propose to extract keywords automatically from a text by selecting those with highest σ and with a frequency of appearance higher than a given cutoff. The method is very easy

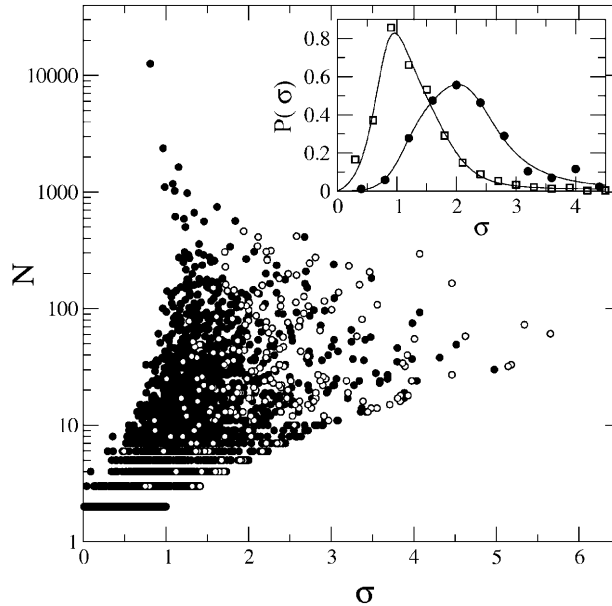


Fig. 2 – Relationship between the frequency of occurrence and the standard deviation σ for all the terms in a technical book [11]. The open circles correspond to words in the index, and the solid ones to the rest of the words. Inset: probability density function of σ for all terms (\square) and for the index terms (\bullet).

to implement and requires little computational resources. The algorithm can be extended to take into account compound terms and can also be used in conjunction with other methods. As σ is a normalized variable, it does not depend on the document length and can be used in collections with a wide distribution of sizes, as in web pages.

The previous ideas can also be applied to DNA strings, in order to test if there exist similarities between natural languages and “DNA language”. This problem has been previously addressed with approaches mainly based on Zipf’s law and Information Theory [12]. In our case, we interpret DNA strings as sequences built from an alphabet of four symbols (A, T, C, G), corresponding to the different nucleotides. In DNA the words more clearly identified, and commonly accepted, are the codons: groups of three nucleotides each of which coding for an aminoacid. The codons are placed consecutively one after the other, without spaces, forming the coding regions, which are separated one from another by non-coding DNA (the proportion of coding regions in a given genome ranges from values larger than 90% in certain prokaryotes to values as small as 3% in the human genome). Outside coding regions probably also exist some kinds of “words” with regulatory functions, promoters, etc. A serious problem in recognizing these “words” is that there is no prior knowledge about their length.

In order to search for a possible “word structure” we divide the DNA sequence into non-overlapping boxes of fixed length n (with $n = 1, 2, 3, \dots$) and associate each box with a word. For DNA sequences of different organisms, we calculate σ for all possible words of length n (4^n), from $n = 1$ to 6 (shifting n times the initial position to take into account all possible phases). The averaged results for four organisms are shown in fig. 3. We have normalized σ to the “discrete” Poisson distribution, so that $\sigma = 1$ corresponds to a purely random sequence. The first feature of this plot is the peak in $n = 3$ for sequences with a large concentration of

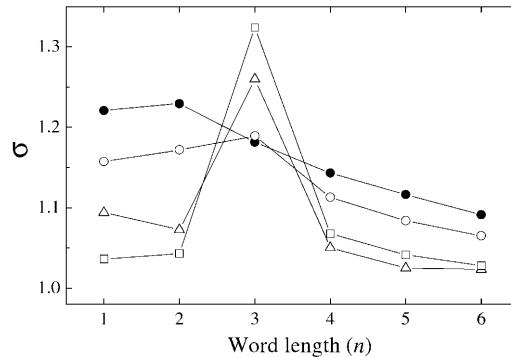


Fig. 3 – σ as a function of the word length for different organisms: *E. coli* (□), *S. cerevisiae* (△), *C. elegans* (○) and *H. sapiens* (●). The dotted line for $\sigma = 1$ corresponds to the random case. We have used the complete genomes of *E. coli* [13], *S. cerevisiae* [14] and *C. elegans* [15], and a contiguous sequence of 22 millions of base pairs from chromosome 22 for *H. sapiens* [16].

coding regions. The height of the peak decreases as the proportion of the non-coding part increases. This means that our method detects the “typical length” of the code. When only coding DNA is considered, the peak at length 3 is always observed, no matter the organism studied (not shown). On the other hand, when the majority of the sequence is non-coding (as in *C. elegans* [15] and, even more, in *H. sapiens* [16]), the results show a bigger attraction for words of length 1 and 2. Our results for $n = 1$ are consistent with the long-range correlations in non-coding DNA previously reported by many authors [17] when considering mononucleotides. The reason is that long-range correlations in mononucleotides imply that there is clustering of these mononucleotides, *i.e.*, regions of abundance of one type of nucleotide followed by regions poor in this type of nucleotide, and this is precisely the condition that produces a high σ value.

Figure 3 also suggests that, if the non-coding DNA stores any information, single nucleotides or dinucleotides are playing an important role. Among all the dinucleotides in human sequences, the CG pairs present the largest value of σ , 1.43. This is due to the high clustering of these pairs, in the so-called “CpG” islands. Other statistical tools have not been so successful, as far as we know, in picking up well-known special biological features in such a straight and simple way.

To conclude, the tools of level statistics are useful for the analysis of texts, and they prove that self-attraction of words is linked to their relevance to the text considered. In particular, σ provides us with an efficient method for keyword extraction, complementary to the existing ones. For DNA, we have shown that the search of subsequences that self-attract leads, as in natural languages, to the extraction of “keywords” within the sequence.

We are grateful to J. L. OLIVER and F. GONZÁLEZ (Universidad de Granada) for useful discussions. We acknowledge financial support of the Spanish DGEISIC for the project numbers PB96-1118, PB96-1120, BIO99-0651-CO2-01 and 1FD97-1358.

REFERENCES

- [1] LUHN H. P., *IBM J. Res. Devel.*, **2** (1958) 159.
- [2] SALTON G. and MCGILL M. J., *Introduction to Modern Information Retrieval* (McGraw-Hill, New York) 1983; SPARK-JONES K., *J. Document.*, **28** (1972) 111; ROBERTSON S. E. *et al.*, *NIST SP*, **500-225** (1995) 109.
- [3] BOOKSTEIN A. and SWANSON D. R., *J. Am. Soc. Inf. Sci.*, **25** (1974) 312; BOOKSTEIN A. and SWANSON D. R., *J. Am. Soc. Inf. Sci.*, **26** (1975) 45.
- [4] HARTER S. P., *J. Am. Soc. Inf. Sci.*, **26** (1975) 197; 280.
- [5] BERGER A. and LAFFERTY J., *Proc. ACM SIGIR'99*, **222** (1999); PONTE J. and CROFT W. B., *Proc. ACM SIGIR'98*, **275** (1998); FUHR N. and BUCKLEY C., *ACM Trans. Inform. Syst.*, **9** (1991) 2.
- [6] BRODY T. A. *et al.*, *Rev. Mod. Phys.*, **53** (1981) 385; MEHTA M. L., *Random Matrices* (Academic Press, San Diego) 1991.
- [7] CUEVAS E., ORTUÑO M., RUIZ J., LOUIS E. and VERGÉS J. A., *J. Phys. Condens. Matter*, **10** (1998) 295.
- [8] Project Gutenberg: <http://www.promo.net/pg/>.
- [9] BEEFERMAN D., BERGER A. and LAFFERTY J., *Proc. ACL/EACL'97*, **373** (1997); NIESLER T. R. and WOODLAND P. C., *Proc. ICASSP-97*, **2** (1997) 795.
- [10] See the web page: <http://bohr.fcu.um.es/words>.
- [11] LERNER I. V., KEATING J. P. and KHMELNITSKII D. E. (Editors), *Supersymmetry and Trace Formulae*, *NATO ASI Ser. B*, Vol. **370** (Kluwer, New York) 1999.
- [12] MANTEGNA R. N. *et al.*, *Phys. Rev. Lett.*, **73** (1994) 3169; MANTEGNA R. N. *et al.*, *Phys. Rev. E*, **52** (1995) 2939; HAVLIN S. *et al.*, *Fractals*, **3** (1995) 269; STANLEY H. E. *et al.*, *Physica A*, **273** (1999) 1.
- [13] BLATTNER F. R. *et al.*, *Science*, **277** (1997) 1453.
- [14] Chromosome sequences obtained from <ftp://ncbi.nlm.nih.gov/genbank>.
- [15] The *C. elegans* Sequencing Consortium, *Science*, **282** (1998) 2012.
- [16] DUNHAM I. *et al.*, *Nature*, **402** (1999) 489.
- [17] PENG C. K. *et al.*, *Nature*, **356** (1992) 168; VOSS R. F., *Phys. Rev. Lett.*, **68** (1992) 3805; LI W. and KANEKO K., *Europhys. Lett.*, **17** (1992) 655.