Physica A 392 (2013) 1481–1492

Contents lists available at SciVerse ScienceDirect

Physica A

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

Improving statistical keyword detection in short texts: Entropic and clustering approaches

C. Carretero-Campos, P. Bernaola-Galván, A.V. Coronado, P. Carpena *

Departamento de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga. 29071, Málaga, Spain

ARTICLE INFO

Article history:
Received 13 March 2012
Received in revised form 20 September 2012
Available online 5 December 2012

Keywords:
Keyword detection
Linguistic
Statistical analysis
Entropy

ABSTRACT

In the last years, two successful approaches have been introduced to tackle the problem of statistical keyword detection in a text without the use of external information: (i) The entropic approach, where Shannon's entropy of information is used to quantify the information content of the sequence of occurrences of each word in the text; and (ii) The clustering approach, which links the heterogeneity of the spatial distribution of a word in the text (clustering) with its relevance. In this paper, first we present some modifications to both techniques which improve their results. Then, we propose new metrics to evaluate the performance of keyword detectors based specifically on the needs of a typical user, and we employ them to find out which approach performs better. Although both approaches work well in long texts, we obtain in general that measures based on word-clustering perform at least as well as the entropic measure, which needs a convenient partition of the text to be applied, such as chapters of a book. In the latter approach we also show that the partition of the text chosen affects strongly its results. Finally, we focus on short texts, a case of high practical importance, such as short reports, web pages, scientific articles, etc. We show that the performance of word-clustering measures is also good in generic short texts since these measures are able to discriminate better the degree of relevance of low frequency words than the entropic approach.

1. Introduction

Statistical keyword extraction and ranking is a key problem in information science. The goal is to develop a statistical method to automatically detect and rank the relevant words of a text without any a priori information. To achieve this objective, several approaches have already been suggested. The first one was proposed by Luhn [1] and was based on an analysis of the frequency of occurrence of words in the text: the words with very high (common) or very low (rare) frequency were excluded and the remainder were considered as keywords. A frequency analysis approach works properly with a collection of documents taken as a reference for comparison [2–5], but it is not sufficient for a single text. There are words which are frequent and relevant, or rare and relevant, and it is clear that a randomization of the text preserves the frequency but destroys the information (in this case none of the words is relevant independently of its frequency).

In recent years two different strategies have been considered to tackle the problem of keyword detection in a text without using an external corpus: the clustering and the entropic approaches.

The clustering approach. In 2002, in a seminal work by Ortuño et al. [6], a crucial relationship was shown between the significance of a word and its spacial distribution. Relevant words, those which are more closely related to the main topics of the text, have a very inhomogeneous distribution. They are usually concentrated in certain regions, present large frequency fluctuations and tend to form clusters. Using a physical language, the different appearances of a relevant word present a high-degree of self-attraction, therefore giving rise to regions with high frequency (clusters) and regions where the word is

* Corresponding author.
E-mail address: pcarpena@ctima.uma.es (P. Carpena).

0378-4371/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2012.11.052

more rarefied. The origin of this self-attraction is related to the structure of the information: an important concept is used more often in regions of the text where it is discussed, and appears more scarcely when a different concept is transmitted or analyzed. In contrast, common words, as articles, prepositions, etc, are placed randomly everywhere in the text and have a quite homogeneous distribution. In a physical language, the different appearances of a non-relevant word do not interact between themselves. These physical concepts of self-attraction or absence of interaction come from the analysis of energy levels of quantum disordered systems, where the first tools similar to the ones we present in Section 2 were used to analyze if the energy levels present attraction, repulsion or absence of interaction (see [7] and references therein). Using that connection between the clustering of a word and its relevance, Ortuño et al. [6] defined an effective method for automatic keyword detection which is based on the statistical analysis of the distributions of distances between successive occurrences of a word. Afterwards, following the same hypothesis, Zhou and Slater [8] proposed another measure that detects an increase of the clustering and is not affected by a single unusual word location in the text. More recently, Carpena et al. [9] significantly improved the method defined in Ref. [6]. They introduced a new approach that combines both the information provided by the clustering of the word and by its frequency.

The entropic approach. Another fundamentally different keyword-detection technique was proposed by Herrera and Pury [10]. They used Shannon's entropy of information to define a new method based on the information content of the sequence of occurrences of each word in the text. A partition of the text is needed to calculate the entropy of any word. They considered 'The Origin of the Species' by Charles Darwin as a corpus sample and its chapters as the natural partition, and showed that this entropic method also provides accurate results and has a performance as good as or better than the clustering approaches proposed in Refs. [6,8]. Other recent entropic approaches confirm these results [11]. In [10] a version of the glossary of the book prepared by hand is employed to identify the relevant words to the text, and the comparison is done by adapting to the problem of keyword detection concepts commonly used in the information retrieval context: precision and recall.

Both approaches have proven to be successful in long texts (book-type), and the measures of word relevance employed by them are described in Section 2. However, some important questions remain open: (i) *The possibility of improvements in both techniques*. We present some improvements for both techniques which are described in Sections 2 *and* 3. (ii) *The use of appropriate metrics to quantify the performance of keyword detectors*. As we wrote above, there are some previous attempts to go beyond qualitative results and quantify the performance of keyword detectors, mainly through the adaptation of the concepts of precision and recall [10]. However, we discuss in Section 4 that such metrics are not appropriated in the keyword detection problem, since they rely on the completeness of the results and imply a large number of putative keywords, but the typical keyword detector user only looks for a reduced number of them. Thus, in Section 4 we propose two metrics which are more convenient for this purpose. Once the metrics have been defined, we use them to compare the performance of the entropic approach [10] and the most recent version of the clustering approach [9]. For that quantitative comparison we use the book 'The Origin of the Species' by Charles Darwin as our entry text due to the availability of a convenient hand-prepared glossary [10] which can be used as a benchmark for the results. (iii) *The possible weaknesses of both approaches*. In particular, the entropic approach requires a partition of the entry text to be applied, and we systematically study the dependence of the results on the chosen partition in Section 5. (iv) *The performance of keyword detectors in short texts.* The problem of keyword detection in short texts is especially important from the practical point of view (scientific articles, web pages, etc.). In this important case, and due to the statistical nature of the measures used in both methods, worse results are expected because of the small size of the sample. Thus, it is important to quantify if both methods work properly when faced with this problem. In Section 6 we study the performance of both methods when applied to short texts in two ways. First, we use as our sample text the shortest chapter of 'The Origin of the Species' (about 3% of the total length of the book), for which we can use the available glossary as a benchmark and quantify conveniently the results. Secondly, we present qualitative results of both methods when applied to several Wikipedia entries with lengths in the range of 500–3000 words, which we take as our generic short texts. Finally, in Section 7 we present our conclusions.

2. The measures of word relevance

As we discussed in the Introduction, we focus on the entropic measure defined by Herrera and Pury [10], E_{nor}, and the clustering measure defined by Carpena et al. [9], C. Here we review both methods and introduce some improvements in C, leading to the new measures C_0 and C_1.

2.1. The entropic measure E_{nor}

The measure E_{nor} uses Shannon's entropy for its definition and for that it needs a previous partition of the text. Suppose we have a text with length N (i.e., composed of N words) and we divide it into P parts. For every word type w, a probability measure over the partition $\{p_i(w)\}$ can be defined as follows:

$$p_i(w) = \frac{f_i(w)}{\sum_{j=1}^{P} f_j(w)} \quad (i = 1, \ldots, P), \tag{1}$$

where $f_i(w)$ is the relative frequency of occurrence of the word type w in the i-th part.