Journal of Informetrics 10 (2016) 439–453

Contents lists available at ScienceDirect

Journal of Informetrics

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

Modeling and visualization of media in Arabic

Zeev Volkovich[a,b,*], Oleg Granichin[c], Oleg Redkin[c], Olga Bernikova[c]

[a] Department of Software Engineering, ORT Braude College, Karmiel 21982, Israel
[b] Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, USA
[c] Research Laboratory for Analysis and Modeling of Social Processes, Saint Petersburg State University, 7-9, Universitetskaya Nab., St. Petersburg 199034, Russia

ARTICLE INFO

Article history:
Received 18 December 2015
Accepted 29 February 2016
Available online 23 March 2016

Keywords:
Media quantization
Visualization
Arabic text segmentation

ABSTRACT

In this paper, a novel method for analyzing media in Arabic using new quantitative characteristics is proposed. A sequence of newspaper daily issues is represented as histograms of occurrences of informative terms. The histograms closeness is evaluated via a rank correlation coefficient by treating the terms as ordinal data consistent with their frequencies. A new characteristic is introduced to quantify the relationship of an issue with numerous earlier ones. A newspaper is imaged as a time series of this characteristic values affected by the current social situation. The change points of this process may indicate fluctuations in the social behavior of the corresponding society as is evident from changes in the linguistic content. Moreover, the similarity measure created by means of this characteristic makes it possible to accurately derive the groups of homogeneous issues without any additional information. The methodology is evaluated on sequential issues of an Egyptian newspaper, "Al-Ahraam", and a Lebanese newspaper, "Al-Akhbaar". The results exhibit the high ability of the proposed approach to expose changes in the linguistic content and to connect them with changes in the structure of society and the relationships in it. The method can be suitably extended to every alphabetic language media.

1. Introduction

The spread of digitized resources to a high level of the information society allows automatically retrieving data from observations using modern technologies. In this regard, new approaches are required for processing documents on the basis of a scrupulous analysis of graphics and morphology; in particular, analyzing peculiarities of the entire linguistic system, for example the Arabic mass media, as a whole.

Any mass media may be considered as a system operating according to its own underlying rules. Thus, the system behavior and properties depend on the internal features and stimulus as well as on surrounding factors. One of the possible illustrations is a relationship between the language of the mass media and the society. Like any other kind of media, newspaper editorial processes may be considered as a feedback system, which reflect changes in the natural and social environment in some ways through personal (author's) or corporative (editorial policy) perceptions and evaluations of reality. In other words, a modern newspaper can be imagined as a mirror reflecting changes in the life of the society. The language of the press itself has many peculiarities, which differ from one printed edition to another and depend on the articles' categories and on the publishing period.

* Corresponding author at: Department of Software Engineering, ORT Braude College, Karmiel, 21982, Israel. Tel.: +972 504041359; fax: +972 49901852.
E-mail address: vlvolkov@braude.ac.il (Z. Volkovich).

http://dx.doi.org/10.1016/j.joi.2016.02.008
1751-1577/

This paper aims at emerging methods for the mathematical modeling and visualization of media in Arabic, and possibly in other languages, using new quantitative attributes. It strives to expose one kind of discrete markers, which are essential for identifying points signaling changes in the linguistic content connected to fluctuations in the political and economic life of a society. We consider a sequence of daily newspaper issues and represent them as histograms of informative terms chosen according to a central criterion. The shape closeness of these histograms is evaluated via a rank correlation coefficient by treating the terms as ordinal data consistent with their frequencies. Aiming to identify changes in the issues' style, we introduce a new characteristic representing the mean rank correlation of a current issue with numerous earlier ones. Change points of this feature can indicate, according to our perception, fluctuations in social behavior causing changes in the linguistic content. Further, a similarity measure created through the introduced average rank correlation makes it possible to derive linguistically homogeneous groups of issues in an accurate manner and without any additional information.

The proposed methodology is evaluated on editorial texts published in the "Al-Ahraam" ("The Pyramids") newspaper for the periods from 1.1.2010 to 31.1.2010, 1.1.2011 to 30.9.2011, and 1.1.2014 to 30.6.2014[1]; and the "Al-Akhbaar" ("The News") newspaper for the period 1.1.2010 to 31.12.2010[2]. These periods include several significant events in the political and economic life of the appropriate societies (particularly, the 'Arab Spring' (Arabic: ربيع عربي, 'ar-rabī 'al-arabī), as well as changes in the official ideology.

The newspaper "Al-Ahraam" was founded in 1875 in Alexandria. It is the second oldest newspaper in Egypt and the most famous daily, not only in the country, but in the Arab world as a whole. It covers a wide array of problems ranging from politics and economy to sport and family issues and is, apparently, the most dominant newspaper in the Arab world. The other source is the Lebanese independent daily newspaper "Al-Akhbaar" ("The News"), which was founded in 1938 and is recognized as one of the most popular in Lebanon. Although these newspapers are published in different countries and are very dissimilar by their inherent nature, and the first one is an official newspaper while the second one is independent, they simultaneously report on the start of the Arab Spring. Al-Ahraam (published in Egypt) reflects the events at a higher resolution because this newspaper is more influenced by those events and their dramatic developments happen mainly in Egypt.

2. Material and mathematical preliminary

Our aim is to demonstrate the ways in which important social events can be recognized using the proposed methodology. As was mentioned earlier, to this end we consider sequential issues of two well-known daily newspapers "Al-Ahraam" (Arabic:الأهرام ; "The Pyramids") and "Al-Akhbaar" (Arabic: "الاخبار The News") published during the stated periods of particular interest.

2.1. Vector Space Model

The first attempts of formalization and automatic parsing of Arabic texts date to the beginning of the 1960s. However, implementation of such models was restricted both by difficulties of pure linguistic problems on one hand and backwardness of software on the other. Moreover, methods whose effectiveness in English or similar languages was proved, appeared to be ineffective when applied to Arabic (Beesley, 1989). Only starting in the 1980s (Koskenniemi, 1983), the development of Arabic linguistics brought methods of mathematical analysis, and technical advances made it possible to transform attitudes in various fields related to Arabic text processing such as computer translation, text segmentation, automatic conjugation and lemmatization, and optical character recognition. Nevertheless, in spite of all these positive developments, the accuracy of the software designed for Arabic appears to be insufficient.

In this paper, we use a relatively simple yet robust *N*-gram based version of the common Vector Space Model (*VSM*). The general *VSM* representation ignores grammar and the order of terms but preserves the variety of terms. Each document is characterized through a terms frequency table against the vocabulary containing all the words (or "terms") in all documents in the corpus. The tables are deemed as vectors in a linear space having dimensionality identical to the vocabulary size. We now consider the three most acceptable methods for vocabulary construction.

2.1.1. Bag of Words (BoW) model

In this model, a document is represented as the distribution of its words. To reduce the space dimensionality, the stop-words are commonly removed because it is doubtful that these frequently arising words provide functional mining. Note that Arabic linguistic material has many peculiarities (Redkin & Bernikova, 2011) and is very challenging for researchers due to the richness of its vocabulary, cursiveness of its script and diversity of written variants of words, well-developed system of verbal conjugation and declension of nouns, variety and complexity of the paradigms of verbal and noun forms, regional lexical and morphological variants, and large number of prefixes, suffixes, and particles and their written forms. In addition to all that, unlike other languages, Arabic has a system of broken plurals, which brings more difficulties. The picture becomes even more complex when taking into consideration local Arabic dialects and Modern Standard Arabic, which differ from one

[1] http://ahram.org.eg/.
[2] http://al-akhbar.com.