Journal of Informetrics 10 (2016) 124–131

Contents lists available at ScienceDirect

Journal of Informetrics

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

Hidden revolution of human priorities: An analysis of biographical data from Wikipedia

Ilia Reznik, Vladimir Shatalov*

National Technical University of Ukraine "Kiev Polytechnic Institute", Slavutych Branch, 6 Heroiv Dnepra St., 07101 Slavutych, Kiev Region, Ukraine

ARTICLE INFO

Article history:
Received 9 September 2015
Received in revised form 3 December 2015
Accepted 3 December 2015

Keywords:
Data mining
Knowledge extraction
Biographical data
Lifespan
History
Wikipedia

ABSTRACT

An innovative study of Wikipedia biographical pages is presented. It is shown that the dates of some historical cataclysms may be reproduced from peculiarities of lifespan changes over time. Time dependence of number of biographical pages related to a year has a broken linear trend in logarithmic scale. It shows a sudden change of the slope from 0.0006 to 0.008 per year near 1700 AC. Presumably, this reflects the emergence of new ways of information dissemination associated with printing of books and newspapers. Cultural or historical significance of a person is measured using a number of proper Wikipedia references. We divided human activity into nine categories using keyword search. They cover over 97% of the extracted data. Time dependencies of shares of each category reveal evolution of priorities or interests of mankind. Finally, categories were merged in just two classes. We call them *Personal* and *Public*, introducing a new index of human priorities as a ratio of *Personal* to *Public*. Being quite constant during almost the entire history, the index shows a sharp jump at the end of the 20th century mainly due to growth of *Sport* and *Art* groups over all others. We consider this as a kind of revolution.

1. Introduction

A well-rounded look at the early history and development of Wikipedia was given by Lih (2009). Wikipedia contains near 35 million articles in 288 different languages in total. More than 12,000 new pages are created every day. We believe Wikipedia implicitly reflects human interests and lifestyle, clearly demonstrating various dimensions of human activity during different periods of history.

Nowadays Wikipedia itself becomes an object of research by means of data mining. A comprehensive survey by Piatetsky-Shapiro (2007) describes the evolution of the data mining and knowledge discovery field over the last 10 years. An overview of mining subjective data on the Web and of recent advances in the area has been presented by Tsytsarau and Palpanas (2012). Moreover, the latter authors discuss several methods of data extraction and try to sketch the future research directions in the field. Besides, some results of Wikipedia mining are reviewed by Nakayama et al. (2010). Furthermore, Ye et al. (2009) have explored how to generate series of summaries based on Wikipedia articles and developed a method to combine wiki concepts and non-textual features. Alfonseca et al. (2013) have extracted a collection of large structured data sets of timely anchored attributes from the revision history of the English Wikipedia. A supervised learning approach for automatic key phrase extraction has been proposed by Abulaish and Anwar (2012). Meanwhile, Milne and Witten (2013) have introduced a

* Corresponding author. Tel.: +380 502114531.
E-mail address: vladishat@gmail.com (V. Shatalov).

http://dx.doi.org/10.1016/j.joi.2015.12.002
1751-1577/

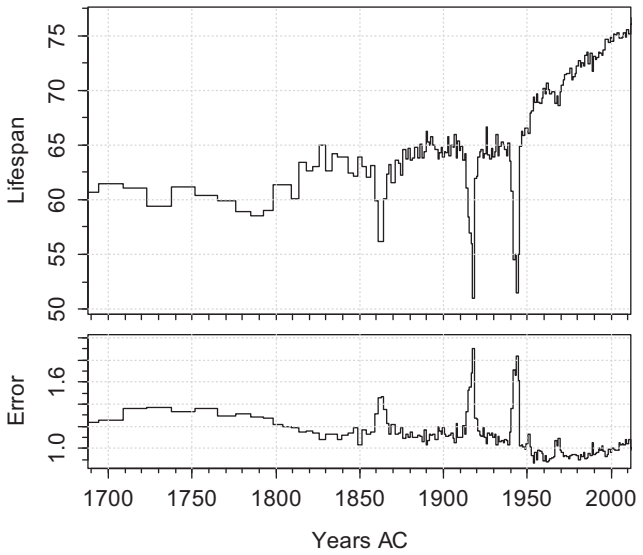

Fig. 1. Lifespan (top) and standard error of the mean value (bottom) in years.

Wikipedia miner toolkit as an open-source software system that allows researchers and developers to integrate Wikipedia's rich semantics into their own applications. Finally, Viseur (2014) has proved the high reliability of the Belgian biographical data extracted from Wikipedia; and the list of works may be continued.

To sum up, Wikipedia data are discussed in detail in a wide variety of works. Meanwhile, only a few of them contain numerical results and forecasts. For example, some quantitative outputs may be found in Bhagavatula, Noraset, and Downey (2013), where the electricity consumption table taken from Wikipedia was considered and correlation between electricity consumption and some other properties of countries (e.g. CO_2 emissions, GDP, etc.) had been estimated.

In general, we consider Wikipedia a unique source for the analysis of human culture. The aim of our present paper is to employ biographical data stored in Wikipedia to get quantitative measures of different areas of human activity, which usually are described just in qualitative terms. Thus, using data mining of Wikipedia articles we would like to draw attention to certain particular historical events and their sociological estimations.

In the next section the details of data extraction and estimations of page significance are presented and successful detection of catastrophic events is demonstrated. Then we propose a set of categories for classifying biographical pages. Finally, a novel index is introduced as the *Personal* to *Public* Ratio and its time dependence is studied in detail. We conclude with the summary and discussions.

2. Imprints of historical events

As the first step, we downloaded the XML dump of English edition of Wikipedia articles for February, 2015. A special parser helped in extracting all the pertinent Web pages from this dump. Then, the biographical pages were separated from the regular ones by the fact that the birth and/or death date is present in an information block ("Info-Box") on a page. This still does not guarantee that such pages are devoted solely to human beings: the information about some famous animals, for example, might also contain birth and death dates. However, we have made sure that the contribution of such cases is negligibly small.

Dates in Wikipedia are presented in multiple ways, namely, sometimes as incomplete data sets and sometimes in the form of several reasonable guesses. Thus, their automatic extraction in all the cases of interest is hardly possible. Nevertheless, a corpus containing 632,092 biographical pages could be successfully extracted. Both birth and death year are known for 207,533 records. For the rest of them, either birth or death year could be located, but without any information about age. The large amount of extracted data enabled us to perform statistical analysis, even for some relatively short historical periods. This way we were able to obtain a table with the following columns: title of the page (usually the name of a person), birth year, death year, age, number of links to the relevant page from other Wikipedia pages, and list of Wikipedia categories in the way they are presented at the bottom of every page.

To sum up, the extracted data set spans a rather long period from 5000 BC to nowadays. Below we present a couple of examples of its immediate usage. The first one is a list of the top-10 most cited persons for several centuries, presented in Table 1. More details about our definitions of *Public* and *Personal* classes in the last column will be given later on.

Next, we checked if events of global scale might somehow be tracked using the extracted biographical data. To achieve this, the time dependence of lifespan was investigated. As the density of data significantly varies in the course of time, time periods containing at least 500 pages were selected to average out the ages. Fig. 1 (top) shows a histogram for the last two