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/

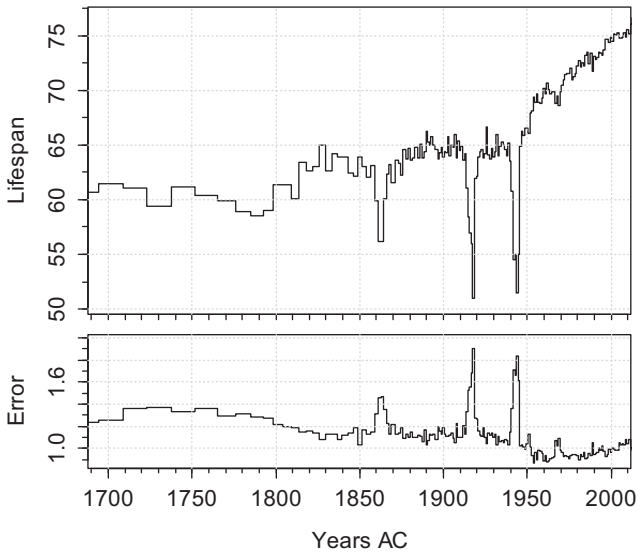

Fig. 1. Lifespan (top) and standard error of the mean value (bottom) in years.

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

Table 1
Most cited people by century.

1st-century	Cited	Class	16th-century	Cited	Class
Jesus	8098	Public	William Shakespeare	10,032	Personal
Pliny the Elder	3170	Personal	Henry VIII of England	5502	Public
Plutarch	2744	Personal	Elizabeth I of England	4678	Public
Saint Peter	2588	Public	Martin Luther	3307	Public
John the Baptist	2561	Public	Charles V, Holy Rom E	3178	Public
Ovid	2135	Personal	Philip II of Spain	2785	Public
Tacitus	1990	Personal	Mary I of England	2212	Public
Nero	1782	Public	Michelangelo	1985	Personal
Josephus	1451	Personal	Mary, Queen of Scots	1961	Public
Trajan	1399	Public	Henry IV of France	1900	Public
18th-century	**Cited**	**Class**	**19th-century**	**Cited**	**Class**
George Washington	7985	Public	Abraham Lincoln	8535	Public
Napoleon	5502	Public	Queen Victoria	7169	Public
Thomas Jefferson	5047	Public	Theodore Roosevelt	5689	Public
Carl Linnaeus	3392	Personal	Charles Dickens	4868	Personal
Benjamin Franklin	3249	Public	Charles Darwin	4429	Personal
George III of the UK	3104	Public	Woodrow Wilson	4228	Public
Immanuel Kant	2498	Personal	Andrew Jackson	3443	Public
John Adams	2281	Public	Mark Twain	3090	Personal
Voltaire	2275	Personal	Edgar Allan Poe	3087	Personal
James Madison	2209	Public	Oscar Wilde	3060	Personal
20th-century	**Cited**	**Class**	**21st-century**	**Cited**	**Class**
George W. Bush	18,349	Public	Roger Federer	6803	Personal
Bill Clinton	12,700	Public	Kanye West	6354	Personal
Ronald Reagan	11,350	Public	Rihanna	6133	Personal
Michael Jackson	9580	Personal	Eminem	5754	Personal
Bob Dylan	9144	Personal	Britney Spears	5709	Personal
Madonna (entertainer)	8856	Personal	Lady Gaga	5586	Personal
Elizabeth II	8821	Public	Snoop Dogg	5495	Personal
Elvis Presley	8603	Personal	Rafael Nadal	5472	Personal
Robert Christgau	8308	Personal	Serena Williams	5324	Personal
John F. Kennedy	7925	Public	Lil Wayne	4635	Personal

centuries, where the time axis reflects the death year. Of course, these results may differ from the official lifespan statistics for the latter contains child mortality as well. Meanwhile, the people referenced in Wikipedia were rarely dying in childhood: they needed to get quite famous in order to justify their appearance in Wikipedia.

The lifespan plot demonstrates a slow rising trend with short-term fluctuations. Remarkably, the pronounced dips in the curve are statistically significant. For example, the dip at 1865 (Fig. 1, top) corresponds to a seven years decrease in lifespan, while the standard error of the mean value is less than one and haft years (Fig. 1, bottom). Their ratio, 7:1.5, far exceeds the 1.96 value required for the regular 95% confidence of a statistical conclusion. Not to excessively overload this graph, we skip here the error bars showing the standard confidence intervals.

The dips in the lifespan around 1915 and 1945 are obviously caused by the World War I, the 1918 flu pandemic, and the World War II. The less intense dip at 1865 possibly corresponds to the American Civil War (but also to the Austrian-Prussian and the Prussian-French wars around the same time); the weak dip at 1969 (it is also statistically significant)—to the Vietnam War. We guess the less pronounced features in Fig. 1 might also be related to wars: at 1810—the Napoleonic Wars, at 1795—the French Revolution and so on.

The top and bottom parts of Fig. 1 demonstrate strong anti-correlation. Dips of lifespan in wartime are obvious while peaks in corresponding errors are not because sampling sizes are the same. To study this, we calculated the detailed distributions of the lifespan for the two war periods and the two relatively peaceful ones preceding them (Fig. 2). These frequency histograms actually represent mortality rates for the selected time periods. As we can immediately see, the wartime histograms appear to be significantly distorted and about two times wider than “peaceful” ones. Broadening of distribution leads to an increase in the standard deviation reflecting the behavior of statistical error of the mean value mentioned above.

Further on, we investigated the time dynamics of the number of pages per year. To this end we assigned:

(a) the half-sum of birth and death dates (if both are known) or
(b) the year of birth plus 30 or
(c) the year of death minus 30 (if just one of the dates is known)

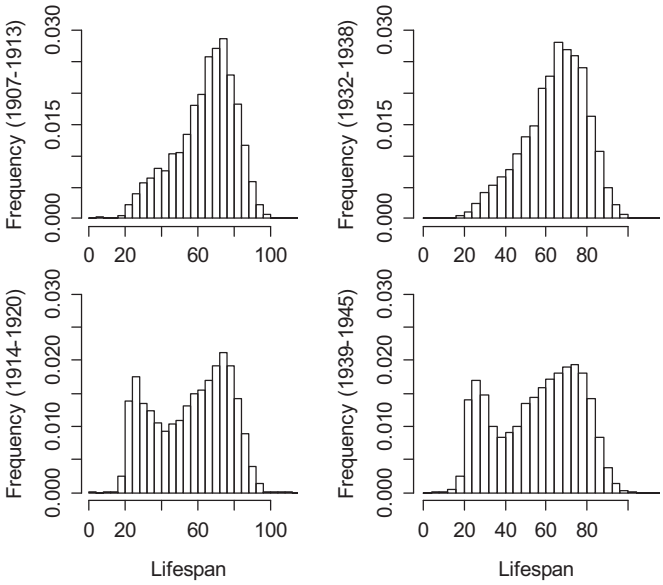

Fig. 2. Lifespan distributions at World Wars (bottom) and preceding "peaceful" periods (top).

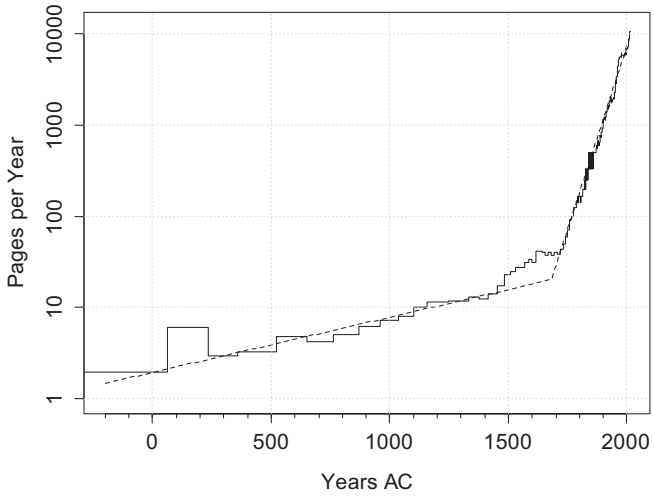

Fig. 3. Number of biographical pages per year (solid line) and its trends (dashes lines).

as the value of "the year of activity" to every biographical page. Of course, the number 30 is quite arbitrary. However, any reasonable value cannot noticeably influence the final results grouped by centuries or even decades. This way we could assign the "year of activity" to all extracted pages.

Fig. 3 presents a histogram of the number of biographical pages versus the year of activity. The shape of the resulting distribution is quite surprising. It has a broken linear trend that shows a sudden change of the slope from 0.0006 to 0.008 per year near 1700 AC. We tend to associate such a behavior with the advent of the newspaper era. Indeed, the newspapers seem to be the primary sources of information (see "The list of the oldest newspapers" in Wikipedia). In addition, we consider the hump after 1500 AC to correspond to the advent of the book printing era that took place in the 15th century.

After the database of biographical pages was created, we scanned the same Wikipedia dump again, but now to calculate a number of pages of all kinds which contain references to each of the biographical pages under study. In such a way, the resulting value ought to represent the statistical weight of each individual person in the total knowledge base, which resembles to some extent the citation index used to measure the importance of scientific works. If no article refers to the person of our study, then he/she probably did not contribute significantly to the worldwide processes of his/her time. Vice versa, any person with high citation index probably contributed a lot—even to the current human culture. Time dependence of average number of citations is presented in Fig. 4.

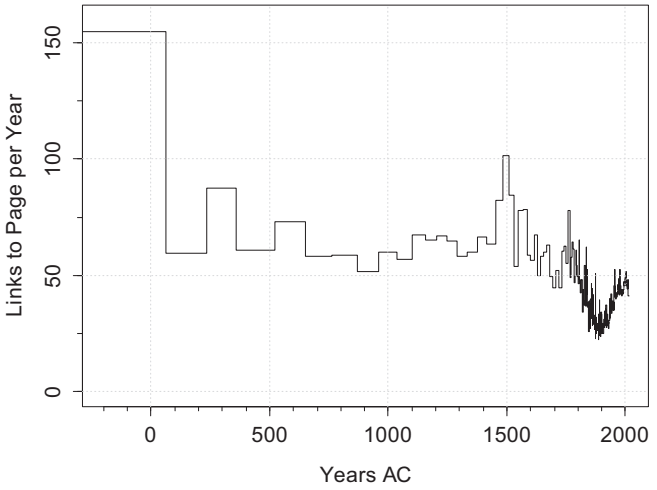

Fig. 4. Time dependence of average number of links to a page.

Remarkably, the obvious prevalence of citing the ancient history is followed by almost constant values of "page significance" in the Middle Ages (Fig. 4). Then we can observe the surge of citations to the Renaissance as well as the subsequent decrease near the 18–19th century. Next, we believe the increase in the 20th century is a result of the 'outburst' of the social activity of people, and of the steadily rising interest in contemporary history. Another reason of the cross-citing increase may be the tighter than ever overlap between different areas of human activity.

3. Evolution of priorities

We now categorize areas of personal activities (and, accordingly, of human culture) in just a few broad and easily understandable terms. We choose the categories *Art*, *Business*, *Medicine*, *Military*, *Merchant*, *Politics*, *Science*, *Sport*, and *Religion*. Unfortunately, Wikipedia descriptive categories accompanying each article are not immediately suitable for such a classification because of the really huge amount of them (229,229 in our case) and due to their excessive specificity as well. Thus, we face a classification problem. It consists in mapping of the enormous number of Wikipedia categories into the categories mentioned above. To solve the problem, we selected several dozens of typical keywords uniquely matching each of our broader categories. For example, keywords "poet", "actor", "music", "writer", "singer", etc., could be associated with *Art*; "artillery", "battalion", "infantry", "soldier" etc.—with *Military*; "church", "mosque", "rabbi", "monk", "mullah" etc.—with *Religion*, and so on.

Our classification was done in the following way. We perform keyword search in the list of categories, shown at the bottom of every personal Wikipedia page. As soon as an occurrence of any of our keywords is found, the page is categorized accordingly. For instance, a page is categorized as *Military* if "artillery" is found. If "mullah" is also found in the same page, the page is also categorized as *Religion*. If nothing is found, we perform the same search in the article itself. Searching categories first speeds up the whole process and leads to lesser number of false positives. Absence of our keywords in Wikipedia categories is typical for short articles only. Large enough article may contain any word. However, detecting a keyword in such a case does not necessarily lead to correct classification. In most cases search within Wikipedia categories only is enough.

A person may appear in several categories in which his/her activity is mentioned in Wikipedia. Therefore, a biographical page may refer to several of our categories. As always, such a "bag-of-words" approach may lead to false conclusions. This could happen, e.g., when the words, surrounding a keyword, change the meaning of it. The simplest example is appearance of a keyword with a negation. Also, not everybody might agree with the classification of "martial art" as a kind of *Art*. But we had to neglect such cases.

To estimate the errors related to the above situations, we selected several random sets of about a hundred pages and manually verified the automatic classification in question. We found that only 2.5% of the total number of cases were not classified (but this could be corrected, if we would extend the keywords set) and about 5% of pages were classified incorrectly. Such errors might be considered representing an error level low enough to allow performing the study to be presented below. We believe that the results of a more sophisticated categorization approach would not change our conclusions significantly.

Fig. 5 shows the resulting shares of the categories mentioned above. For example, 17% of pages have *Art* category only, 3% of pages have *Science* category only, and 1% has no other category than *Art* and *Science*. The shares falling below 1% of pages are not shown. They cover about 17% of pages in total. Only 2.5% of the biographical pages remain non-categorized and will not be counted further on. Thus, our present choice of categories seems reasonable.

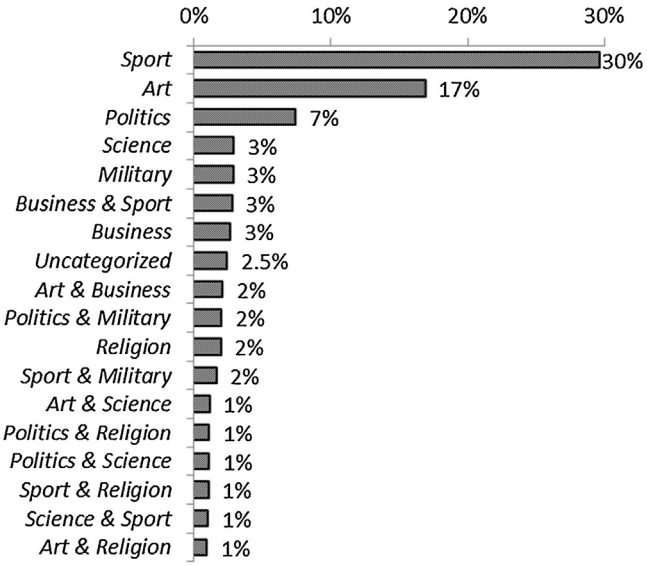

Fig. 5. Shares of categories and combinations of categories (only exceeding 1% are shown).

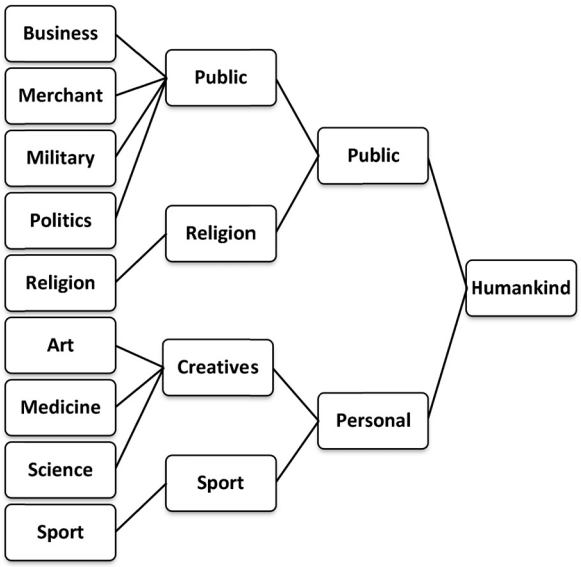

Fig. 6. Tree structure of the biographical pages classification.

In order to simplify our study of changes over time in human priorities, we aggregated our nine categories into four broader groups:

1. Public as the aggregation of Business, Merchant, Military and Politics;
2. *Creatives* as the aggregation of the *Art*, *Medicine* and *Science*;
3. *Religion* as the category *Religion* itself;
4. *Sport* as the category *Sport* itself.

These groups can be easily divided into two almost non overlapping classes. We label them as *Public* and *Personal*. Specifically, we define a person as *Public* if he/she became famous due to social interactions with other people. This way politicians, businessmen, merchants, military men as well as clergymen are considered *Public*. However, people of arts, medicine, sciences or sports could come into history due to their personal achievements. Therefore, we assign them to the *Personal* class. Our resulting classification is illustrated by the tree structure shown in Fig. 6.

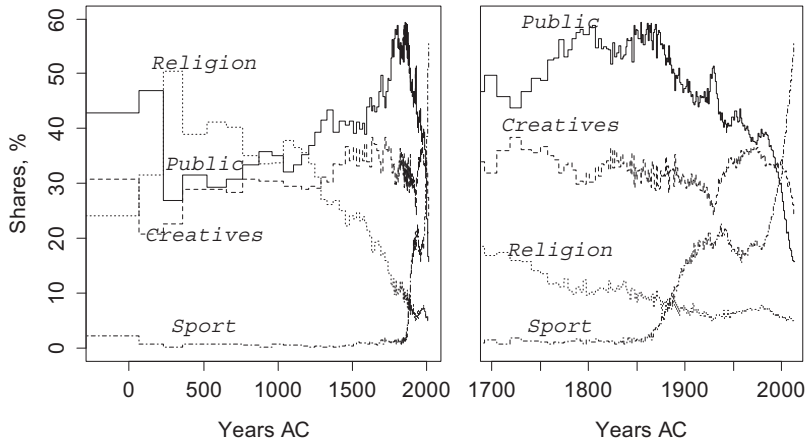

Fig. 7. Shares of main groups of categories from 200 BC to 2000 AC (left) and detailed results for 1700–2000 AC (right).

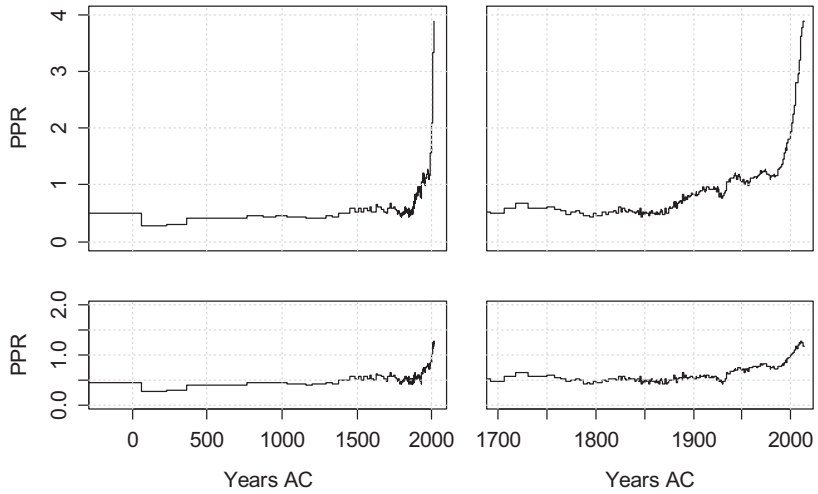

Fig. 8. *Personal* to *Public* Ratio (PPR) in different time periods. Bottom: without *Sport.*

Next, we calculated the time dependence of the shares of each of the four groups mentioned above. Results are shown in Fig. 7. They reflect the evolution of mankind's priorities. We conclude that the share of *Creatives* remains relatively constant over time except for the last decades, as it becomes suppressed by *Sport*. The share of *Religion* dominates from 300 to 1200 AC and since then steadily decreases until nowadays. *Politics* competes with *Religion* until 1200. Then it dominates all the others until the end of the 19th century. Afterwards it goes down to become dominated by *Sport.* Interestingly, the latter, starting from 1880s, increases rapidly to 20% in 1920. In the 1940s this growth is dominated by *Creatives*, and, after a period of stagnation, *Sport* exhibits an avalanche-like growth from 1990 until nowadays dominating strongly all the others groups. The rapid growth of the *Sport* share since 1880 is mostly caused by the emergence of the European football in its modern form (rules stated in 1888). In fact, 52% of *Sport* pages are related to football. However, the actual reason for *Sport* growth at 1990 is not quite clear. Anyhow, now the *Sport* group occupies about one third of the Wikipedia biographical pages.

To demonstrate the changes in priorities of people in a more cogent manner, we plotted the ratio of the number of pages in the *Personal* class and the number of pages in the *Public* class (Fig. 8, top). Remarkably, the *Personal* to *Public* Ratio demonstrates a surprisingly steady behavior in all time periods until the end of the 20th century, when it starts increasing about eight times due to the growth in the groups of *Sport* and *Art*.

If we now exclude the *Sport* group (see Fig. 8, bottom), the growth factor reduces to 2.5. This is also large enough value. The major contribution to this growth comes from the *Art* group, which is nowadays formed mostly by movie actors and pop stars. The changes in the categories to which the top people belong (cf. Table 1) is in accordance with our above conclusion, indicating the drastic restructuring of the conventional human activities. This might be seen as a kind of hidden revolution in the human perception of the world: *Personal* strongly dominates *Public* in the last decades. A possible reason for this could be the widespread use of the Internet, allowing people to communicate in a much more personalized way. Other mass media (in particular television) are not likely to be relevant because they have much longer history than the Internet.

4. Conclusions

Wikipedia, being a knowledge database, is a promising object for any kind of data mining. Our work illustrates a result of such an approach. It allows us to explore changes in people's interest in historical persons over time.

Wikipedia mining is capable of quite accurately detecting the dates of historical cataclysms based upon changes in lifespan and in the distribution of the mortality rate. This indicates the reliability of the biographical data extracted from Wikipedia.

The 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 that reflects the emergence of new primary sources of information spreading, such as book and newspaper printing.

The number of links to a page (citations) is proposed as a measure of page significance. Time dependence of average number of citations is presented.

The classification suggested here splits the conventional human activity areas into nine main categories, and covers about 97% of the Wikipedia biographical pages available. The time dependence of the grouped categories clearly reflects the inevitable evolution of human priorities.

We have introduced a new index of human priorities, namely the *Personal* to *Public* Ratio. The time dependence of this ratio exhibits a kind of hidden revolution in human priorities. Being almost a constant over centuries it increases eight times in the last decades due to the *Sport* and *Art* group growth.

Hence, Wikipedia biographical pages can be used as a unique self-consistent source of information for studies in historical sociology. Wikipedia data mining may reveal changes over time in the human perception of the world, and may also serve as an independent reliable quantitative method of investigation of historical events.

Acknowledgements

We dedicate the article to our teacher K. B. Tolpygo, the great physicist and outstanding person, whose birth centenary is celebrated in 2016. He instilled in us a taste and interest in the exploration of unusual patterns in the nature and society and thus stimulated this work. Evgeni B. Starikov and Alex Leytes's useful advices and invaluable assistance are greatly appreciated. We would like to express our sincere gratitude to editor-in-chief Ludo Waltman for attentive critical reading and comments.

References

Abulaish, M., & Anwar, T. (2012). A supervised learning approach for automatic keyphrase extraction. *International Journal of Innovative Computing, Information and Control*, *8*(11), 7579–7601. ISSN 1349-4198.

Alfonseca, E., Garrido, G., Delort, J.-Y., & Peñas, A. (2013). WHAD: Wikipedia historical attributes data. *Language Resources and Evaluation*, *47*(4) http://dx.doi.org/10.1007/s10579-013-9232-5

Bhagavatula, C. S., Noraset, T., & Downey, D. (2013). Methods for exploring and mining tables on Wikipedia. In *Proceedings of the ACM SIGKDD interactive data exploration and analytics (IDEA). ACM, 2013.* ⟨http://users.eecs.northwestern.edu/∼csb939/docs/p19-bhagavatula.pdf⟩.

Lih, A. (2009). *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia* (1st ed.). New York: Hyperion (ISBN 978-1-4013-0371-6) 1-4013-0371-4.OCLC232977686.

Milne, D., & Witten, I. H. (2013). An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, *194*(1), 222–239.

Nakayama, K., Pei, M., Erdmann, M., Ito, M., Shirakawa, M., Hara, T., et al. (2010). Wikipedia mining. In *Wikipedia as a corpus for knowledge extraction.* , http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.178.3046.

Piatetsky-Shapiro, G. (2007). Data mining and knowledge discovery 1996 to 2005: Overcoming the hype and moving from "university" to "business" and "analytics". *Data Mining and Knowledge Discovery*, *15*(1), 99–105.

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*(3), 478–514. http://dx.doi.org/10.1007/s10618-011-0238-6

Viseur, R. (2014). Reliability of user-generated data: The case of biographical data in Wikipedia. In *OpenSym '14* , http://dx.doi.org/10.1145/2641580.2641581 ACM 978-1-4503-3016-9/14/08.

Ye, S., Chua, Tat-Seng, & Lu, J. (2009). Summarizing definition from Wikipedia. In *Proceedings of the 47th annual meeting of the association for computational linguistics and the fourth international joint conference on natural language processing of the AFNLP* http://dx.doi.org/10.3115/1687878.1687908. Source: DBLP Conference: ACL 2009