Available online at www.sciencedirect.com
ScienceDirect
Physica A 385 (2007) 683–688
PHYSICA A
www.elsevier.com/locate/physa

The shape of the past in the World Wide Web: Scale-free patterns and dynamics

Roger Jovani[a,*], Miguel A. Fortuna[b]

[a]Department of Conservation Biology, Estación Biológica de Doñana, CSIC, Avda. Mª Luisa s/n, E-41013 Sevilla, Spain
[b]Integrative Ecology Group, Estación Biológica de Doñana, CSIC, Avda. Mª Luisa s/n, E-41013 Sevilla, Spain

Received 1 June 2007
Available online 24 July 2007

Abstract

Human societies accumulate a great deal of information about past events. People make reference to things that happened in time in different ways and record them in multiple media. We have studied the current use of this information by analysing the frequency of occurrence of numbers associated with years in the World Wide Web (WWW). We found a consistent scale-free reduction in the number of web pages referencing events occurred in increasingly older years. This was found for the entire WWW and separately for web pages written in 12 different languages. From year 2005 to 2006 the increase on the number of web pages associated to each year also decayed as a power-law from recent to old years. Such general pattern reveals that time elapsed to present is the best predictor of the interest or the amount of information on a particular year in the WWW. Moreover, the power-law increase from one year to the next shows that the scale-free shape of past in the WWW is dynamically maintained.

Keywords: Complexity; Culture; Human societies; Language; Memory; History; World Wide Web

1. Introduction

Information about past events travels within societies through time thanks to oral tradition and different storing media such as books, newspapers, tapes, archives, and more recently, multimedia digital files. Most of this huge amount of information about the past is not used by people for communicative purposes, but a small part is referred during communication either because of their link to current interests or their own relevance. The use of information about the past is a central piece of human culture (e.g. Ref. [1]), but it is difficult to study because of the different channels used for communicative purposes, because this pool of information is changing permanently, and because many people need to be studied to achieve a representative sample of information use in societies. In this way, many studies have approached to the loss of memory of individuals [2], but analyses of empirical data on the way societies use information about the past has been rarely done.

*Corresponding author. Tel.: +34 954 23 23 40; fax: +34 954 62 11 25.
E-mail address: jovani@ebd.csic.es (R. Jovani).

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.07.023

We only know of a previous study where references to past events were studied from information published in newspapers. In the present study we have tried to overcome these challenges, and approach to a broader source of communicative channels, by using the World Wide Web (WWW).

The WWW is a decentralised all-purpose communicative media that has become a place where everything may be found. The WWW is not just a place to store information for a long time, but rather a place to communicate things. It is like a virtual showcase where people display information that they would like to share with others. The WWW combines, in a single format, different informative channels, such as books, archives, newspapers, journals, catalogues, government documents, blogs, advertisements, and even chats in on-line forums. Moreover, thanks to powerful search engines, all this information is easily accessible, allowing retrieving at the same time information of any kind communicated in many different formats. In addition, the WWW continuously receives new documents and corrections to old ones, and old web pages disappear, ensuring that the WWW is permanently updated. All of this, in addition to being the largest and most easily accessible dataset, makes the WWW an ideal database to study in real time the use of information by many people.

Numbers are used for many purposes in the WWW [3,4]. Dorogovtsev et al. [3] showed that the frequency of occurrence of numbers in the WWW decayed as a power-law (i.e., a scale-free pattern) from small to large natural numbers (see Fig. 1a,b). They also found that the occurrence of numbers used to indicate calendar years was higher than expected by this general power-law decay, reaching a maximum in the current year number [3]. Dorogovtsev et al. [3] suggested the study of the frequency of numbers in the WWW as a promising way of addressing cultural, psychological, and social human phenomena. Here, we depart from these exciting initial results and suggestions to further explore the use of information about the past in the WWW. Specifically, our aim was to explore how general are power-law decays in the use of information about

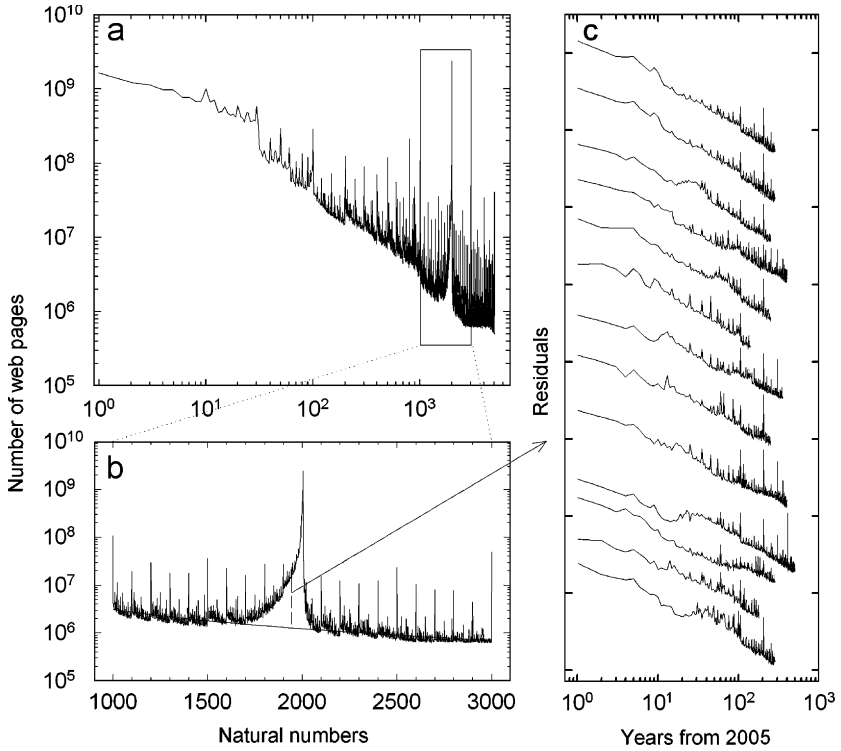

Fig. 1. (a) Frequency of occurrence of natural numbers in web pages written in English in the WWW. (b) Power-law fitting of the frequency of natural numbers from 1000 to 1300 and 2700 to 3000, and an example of residual calculation. (c) Log–log plot of the range of positive residuals for each language or the entire WWW from 2004 to back (e.g. 1 = 2004, 10 = 1995). Data has been shifted in the vertical direction for clarity, and ordered from top to down as in Table 1 according to their fit to a power-law. However, fitted power laws are not shown for clarity.

the past in the more widely used languages in the WWW. Moreover, taking advantage of the continuous growth of the WWW, we studied whether and how this scale-free pattern is dynamically maintained.

2. Materials and methods

In the second week of May 2005 we quantified the number of web pages containing Arabic numerals from 1 to 5000 (hereafter, frequency of numbers) supplied by the most popular search engine, Google (http://www.google.com). We did that separately for the different languages it allows, and also for the entire WWW (web pages written in any language). We first fitted for each language separately a power-law to a set of numbers within the intervals 1000–1300 and 2700–3000 (Fig. 1b). Then, we calculated the difference (i.e., the residuals) between the theoretical occurrence according to the fitted power-law and their real values (see Fig. 1b). In other words, we calculated the surplus of web pages containing numbers associated with years once the frequency of use for other purposes was eliminated. After that, we selected the interval of residuals from 2005 to older year numbers until the first no-positive residual was found. Some available languages did not offer enough sample size for reliable analyses and were not used in the study. We excluded Chinese and Japanese for the likely effect of having an own calendar. Italian was also excluded because the residuals could not be calculated due to an abnormal behaviour in the frequency of occurrence of numbers below 1300.

The number of web pages is constantly increasing. For web pages in the entire WWW (written in any language) we retrieved one year later (in the second week of May 2006) the number of web pages for each number associated with years, and we obtained the increase on the number of web pages for each number.

3. Results

References to recent years were much more frequent than to older ones in all the languages and for the entire WWW (Fig. 1c). The decreasing trend showed a close fit to a power-law with slopes between -0.8 and -1.3 (see Table 1). This scale-free pattern is characterised by a fast decrease in the frequency of occurrence of the recent years and by the presence of a long tail as time goes back, displaying a lineal relationship when both variables are plotted in a log–log axis (see Fig. 1c).

When we retrieved the number of web pages in the WWW one year later we found an increase of ca. one order of magnitude in the number of web pages (Fig. 2a). Even so, the pattern remained the same, references to progressively older calendar years decaying as a power-law ($r_s = 0.9998$) with a $slope = -1.1$ (Fig. 2b), similar to the $slope = -1.2$ found in May 2005. Thus, plots for retrievals from 2005 and 2006 displayed almost parallel lines (Fig. 2a). Note that this means that the increase in the number of pages from one year to the next

Table 1
Languages analysed and their power-law fitting values to the residuals in the number of web pages containing numbers associated with years appearing in Fig. 1c

	R^2	Slope	Std. Err.
WWW	0.960	−1.285	0.016
English	0.957	−1.344	0.017
Portuguese	0.944	−1.264	0.019
French	0.939	−1.102	0.014
Spanish	0.934	−1.276	0.021
Korean	0.925	−1.207	0.029
Danish	0.918	−0.965	0.015
Polish	0.908	−1.053	0.021
Swedish	0.908	−0.978	0.016
German	0.901	−0.792	0.016
Dutch	0.902	−1.030	0.015
Czech	0.897	−1.013	0.026
Russian	0.882	−1.194	0.026

All fits were statistically significant at $\alpha = 0.05$.

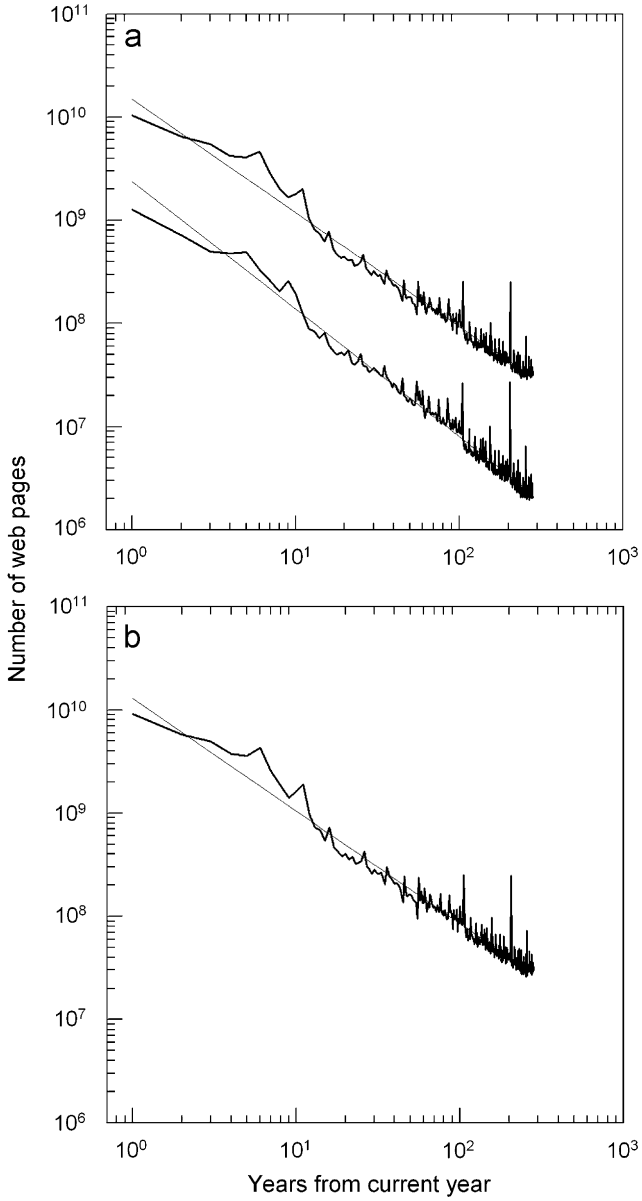

Fig. 2. (a) Comparison of the number of web pages containing each number for data retrieved in 2005 (down) and 2006 (up). The x-axis is the log of (current year (2005 or 2006) less the number retrieved), i.e., 0 is $10^0 = 1$, and thus corresponds to the number 2004 for data retrieved in 2005 and the number 2005 for data retrieved in 2006. (b) Difference on the number of web pages between data retrieved in 2006 *vs*. 2005. Power-law fits are also shown (straight lighter lines).

was very heterogeneous because of the log–log nature of the axis. This increase was around 10^{10} for recent years and $10^{7.5}$ for old ones, that is, the increase was 2.5 orders of magnitude higher for recent than for old years (Fig. 2b).

4. Discussion

Both in 2005 and 2006, and for web pages written in different languages and for the entire WWW, the use of past information decayed as a power-law, meaning that recent years were much more referred than older ones. This could reflect a rapid loss of interest by people for events occurring in the progressively more distant past,

or it could be the indirect outcome of people mainly writing and talking about things that currently concern them (mainly from the near past and present). Under the later scenario, references to past events would then occur because their link with present events or interests, links with more distant events being progressively less probable.

Power-law patterns could originate through different mechanisms, and could be transient stages of the system [5–7]. The fact that we have found in two consecutive years the same pattern found by Dorogovtev et al. for the entire WWW, and for web pages written in different languages, jointly with the associated scale-free increase found here strongly suggest that the power-law decay on information use is not a transient stage. Thus, in this case, the scale-free pattern results from an underlying scale-free process that dynamically maintain the multiplicative relationship between consecutive years. However, whether it is a loss of interest, a perception of loss of relevance of old for current issues, a real loss of information, or a mixture of them, remain open questions.

We have found that there is some amount of noise around the power-law pattern, that is, data does not perfectly match to a straight line in the log–log plots (Figs. 1c, 2a). This merits further study because these local deviations from the power-law could be indicating some surplus/deficit of interest for certain years compared with the adjacent ones. However, it is also much interesting that this noise does not break down the power-law behaviour of past use in the WWW, achieving very high fits to a power-law (Table 1). This means that the decrease of use of information of increasingly older events is constant through different scales of time. This is not trivial at all if we realise how many factors (historical events such as technological advances in storage methods) could potentially have modified this constant rate, introducing inflection points in the straight line. This suggests that what really determines how the information about events that occurred in a given year is reported in the WWW is not the relevance of these events *per se* (however, it may be measured) but their closeness to the present.

We want to emphasise that our results do not confront, but rather complement, historical, sociological and psychological approaches to the way human societies look to the past. What we have found is that the most important fact explaining how often the information about events that occurred in a particular year is reported in the present, is the time elapsed until the present. This means that the details of what happened in each year are not so relevant to understanding the temporal behaviour of the use of information about the past. This decoupling between the details of a system and its high level collective behaviour is a pervasive phenomenon in complex biological, physical, and social systems [7]. Thus, concepts coming from complex systems research and the quantitative tools of statistical physics [7] should be taken into account when approaching to this kind of emerging properties of human culture.

The scope of our results depends on whether the use of information in the WWW can be translated to the use of information in entire societies. The WWW is not accessible to everyone, and the number of people that have enough skills to display a web page is still low. Moreover, the way people communicate in the WWW may be different than other communicative channels. Interestingly, however, the results found here are partially supported by the use of year numbers in some newspapers [8]. However, they found a truncated power-law with an inflection point at 50 years, maybe as a consequence of the lower relevance of ancient past for the scope of newspapers [8]. Further studies will be necessary to clarify the relevance of our results as a detailed description of the present use of historical information. Surely, information contained in the WWW will be a much greater source of information for future generations than the huge amounts of latent information stored in libraries and archives. Thus, although the WWW may have some biases that need further study, the information displayed in it is an important study subject in itself. In any case, further monitoring of this continually updated pool of information will provide important insights into the dynamics in the way societies use information about the past, and thus which information will be really available and used in the future.

Acknowledgements

We thank J. Bascompte, C.J. Melián, A. Hampe, and many others for helpful discussions. This work was funded by the Spanish Ministry of Science and Technology (Fellowship BES-2004-6682 to M.A.F.).

References

[1] J. Jedlicki, Historical memory as a source of conflicts in Eastern Europe, Communists Post-Communist Stud. 32 (1999) 223–232.
[2] J.T. Wixted, The psychology and neuroscience of forgetting, Annu. Rev. Psychol. 55 (2004) 235–269.
[3] S.N. Dorogovtsev, J.F.F. Mendes, J.G. Oliveira, Frequency of occurrence of numbers in the World Wide Web. Physica A 360 (2006) 548–556.
[4] G. Levin, M. Wattenberg, J. Feinberg, D. Becker, D. Elashoff, S. Wynecoop, ⟨http://www.turbulence.org/Works/nums/⟩ 2002.
[5] J.T. Wixted, E.B. Ebbesen, Genuine power curves in forgetting: a quantitative analysis of individual subject forgetting functions, Mem. Cognit. 25 (1997) 731–739.
[6] S. Sikström, Forgetting curves: implications for connectionist models, Cogn. Psychol. 45 (2002) 95–152.
[7] R.V. Solé, J. Bascompte, Self-Organization in Complex Ecosystems, Princeton University Press, Princeton, 2006.
[8] T. Pollmann, R.H. Baayen, Computing historical consciousness. A quantitative inquiry into the presence of the past in newspaper texts, Comput. Humanit. 35 (2001) 237–253.