Inversed N-gram viewer: Searching the space of word temporal profiles

Vincent Buntinx, Frédéric Kaplan

DHLAB, École Polytechnique Fédérale de Lausanne, Switzerland

Studies based on the visualization and analysis of temporal profiles of words using a so-called “n-gram” approach have been popular in recent years [1, 2]. However, most of the studies so far discuss the case of remarkable words which are mainly found due to the researcher’s intuitions for finding “interesting” curves using the n-gram viewer. In this paper, we investigate how we could inverse the problem and automatically explore the space of temporal curves in search for words. For instance, we could be interested in asking the system to retrieve all curves “similar” to a given one. This would entail the definition of a method to describe temporal profiles and classify them according to a predefined distance.

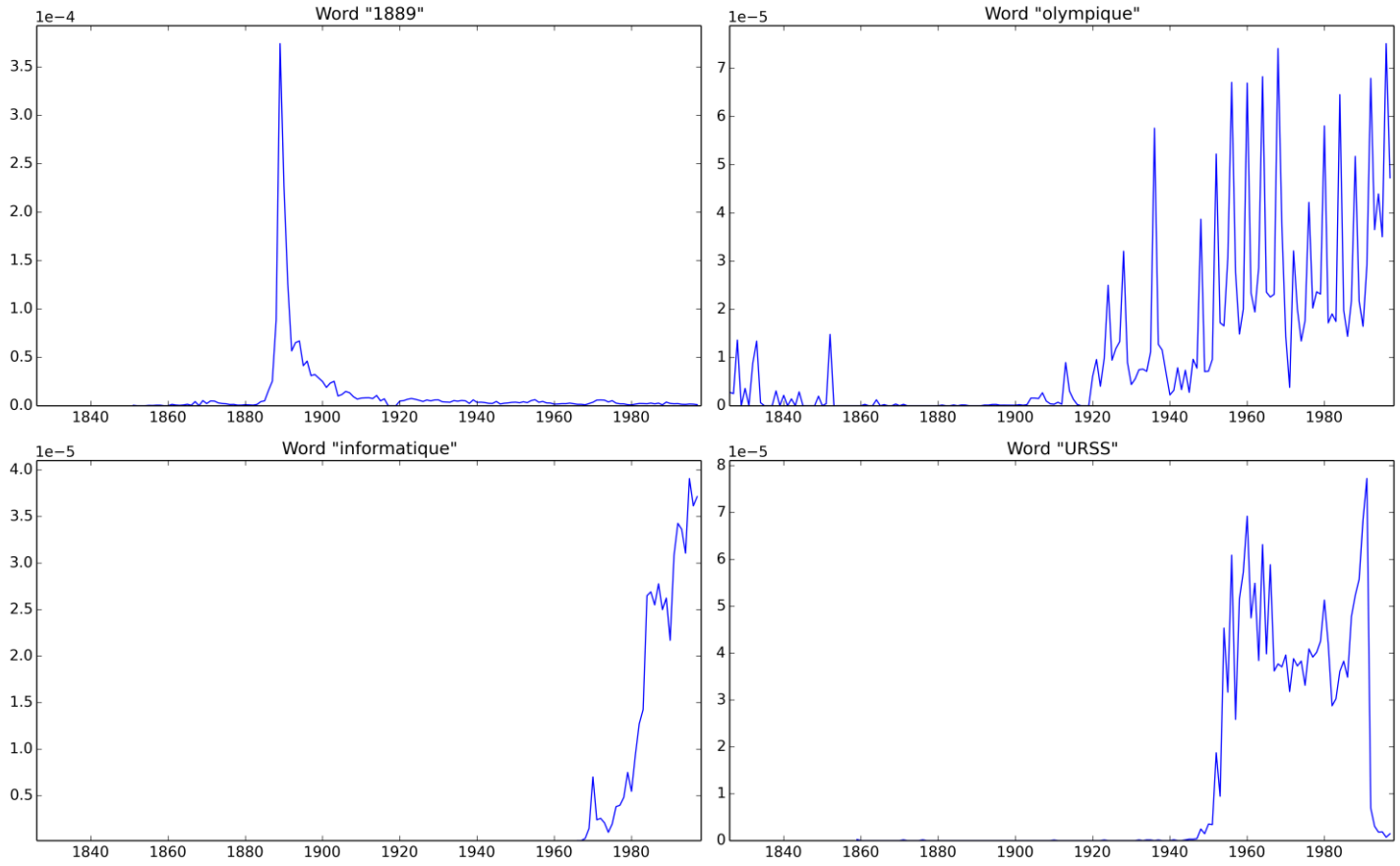

FIG 1:
4 examples of curves: (1) The word “1889” is very popular in the year 1889 but ceases to be as interesting quickly after, (2) The term “olympique” is very popular every 4 years, (3) The term “informatique” keeps getting more popular over time (4) The term “URSS” is very popular only between 1950 and 1990.

The study presented in this paper uses a database of 4 million articles covering a period of 200 years and is composed of digitized facsimiles of "Le Journal de Genève" and "La Gazette de Lausanne". Each article has been OCRed and indexed in an SQL database. For each indexed word, the yearly temporal profile showing the relative number of word occurrences in the corpus has been pre-computed. We designed an n-gram viewer that allows for the simultaneous comparison of up to 4 words. In order to compare these curves with one another and therefore be able to retrieve “similar” curves automatically in the database, we need to find a simple way of describing them. The problem could be approached in a very general manner (how to describe an unknown curve by approximating it on a given decomposition base) or could take into account the fact that we are dealing with temporal profiles that could be described by a given limited number of archetypical curves. The second option implies having some a priori knowledge about the kind of curves one might encounter but has the advantage of allowing for a very compact description of general curves using a family of possible profiles. We choose this second option in this paper.

The first step of our curve analyzer is to associate a given curve to a pre-existing curve family. This can be done in a hierarchical manner. At this stage of our research, we decompose curves into four basic families: “Dirac” curves (with a single peak), periodic curves (with regular peaks of popularity), monotonic linear curves (either increasing or decreasing) and “square” curves (associated with a predefined period). It is clear that these 4 families do not cover the entire spectrum of possible curves but they do provide a possible starting point for investigating the space of “remarkable” curves. To determine if a given curve can be reasonably approximated by one of the 4 families, we carry out a sequence of tests, starting with the periodicity of the curve. To evaluate whether a curve is periodic we simply compute its Fourier transform and automatically check for hidden periodicity. If the curve is detected to be periodic we extract the period and try to fit a “comb” function as defined by the following formula:

$$f(x;a,b,m,t) = \begin{cases} 0, & if\ x < m \\ a, & if\ x \geq m\ \ and\ \ (x-m)\ modulo\ t = 0 \\ b, & if\ x \geq m\ \ and\ \ (x-m)\ modulo\ t \neq 0 \end{cases}$$

The fitting is done using Particle Swarm Optimization method (PSO) [3] and the quality of the fitting with the theoretical curve is measured using the classical least squares error (minimizes the sum of squared residuals).

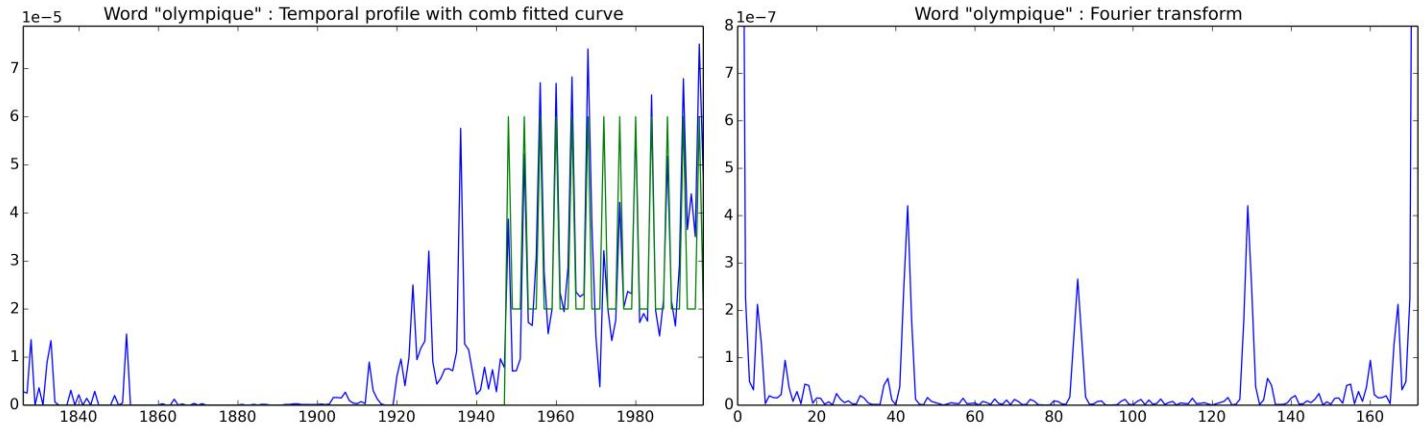

FIG 2:
Temporal profile (blue) of the word “olympique” with fitted curve (green) and Fourier transform (right).

Using the model, the extracted period becomes a way of comparing periodic curves with one another. Fig 3 shows different curves sorted by periodicities.

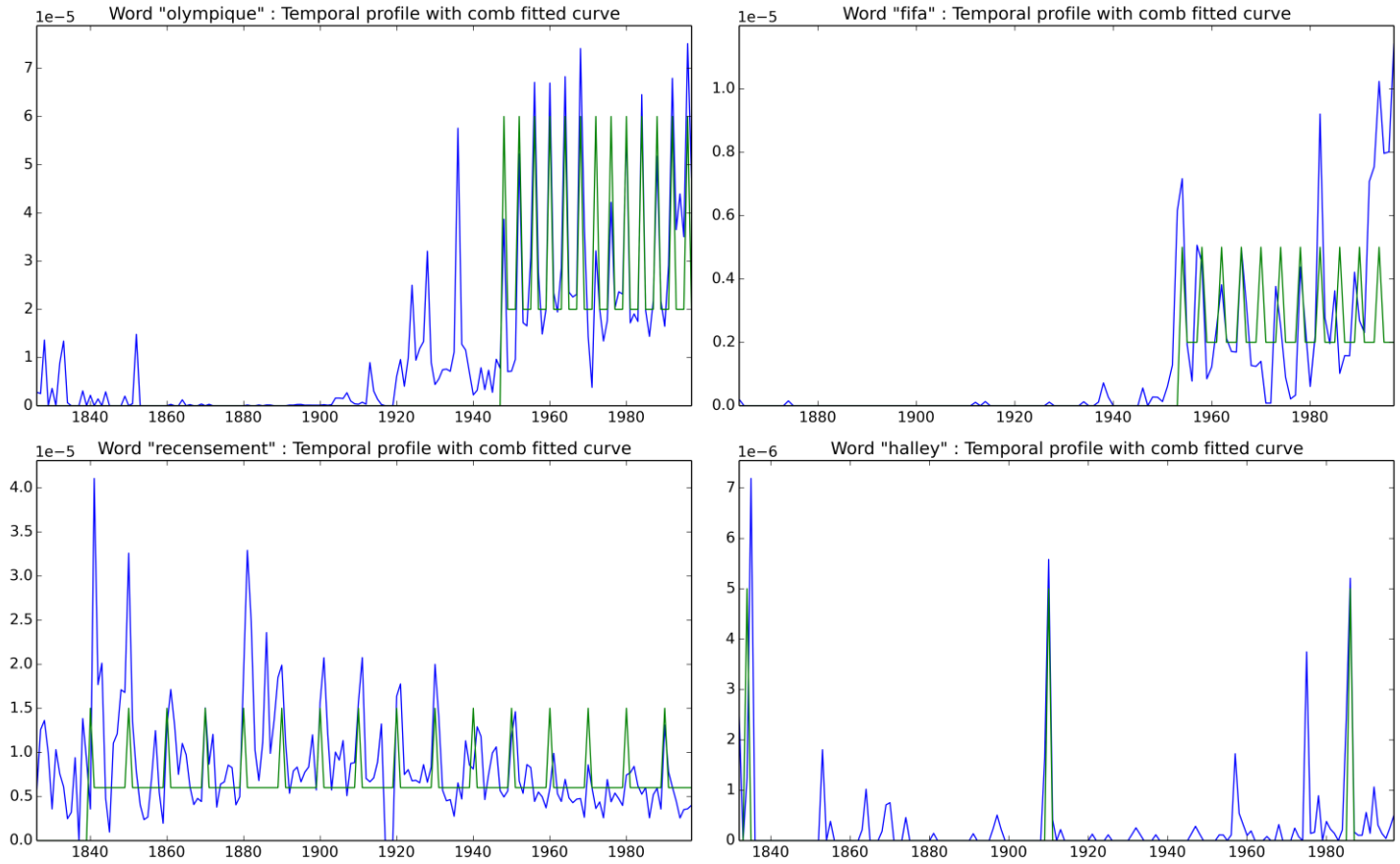

FIG 3:
Temporal profile (blue) of the words "olympique", "fifa", "recensement" and "halley" with fitted curve (green)

For each curve, the sum of the squared residuals between the actual curve and the fitted curve is calculated and this represents the error of the comparison between them. This measure is used to optimize the fitting and allows for determining the category the actual curve belongs to when the optimization is done for all predefined categories of curves.

For "peaks" we try to fit the curve with a classical model used in the analysis of "fads" (e.g. for instance the use of this model in the context of the analysis of Internet Memes [4]) by using the Fréchet curve [5] as defined by the following formula:

$$f(x;a,s,m) = \begin{cases} \frac{a}{s}\left(\frac{x-m}{s}\right)^{-a-1} e^{-\left(\frac{x-m}{s}\right)^{-a}}, & if\ x > m\ \ and\ \ s > 0 \\ 0, & else \end{cases}$$

Using this model, fitted curves can be described using the position and value of the peak, allowing optimization with only one degree of freedom. These three parameters can be used to compare curves with one another as shown in Fig 4.

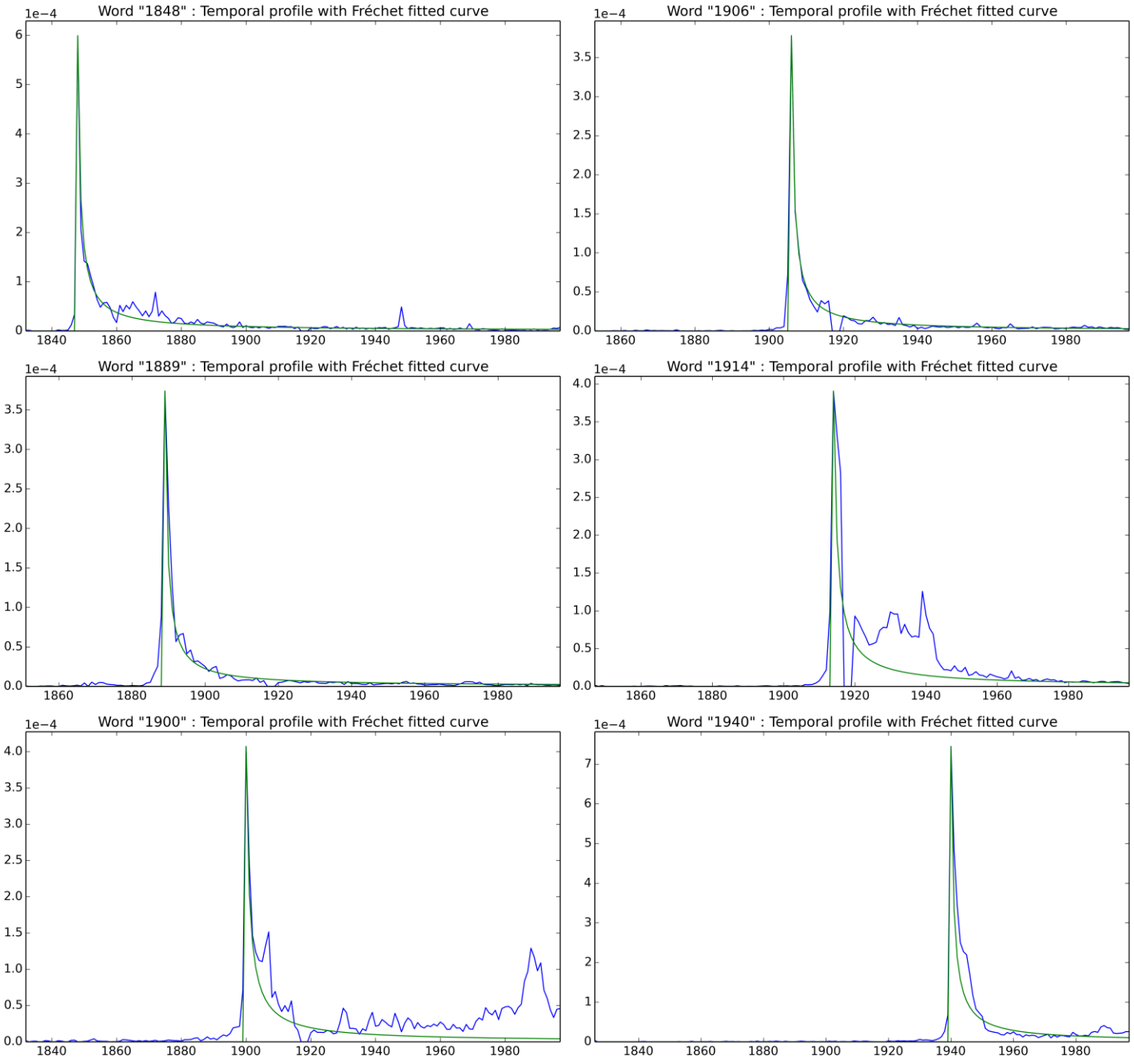

FIG 4:
The year 1889 is much like 1906 using the fitted Fréchet curves. The years 1848 and 1940, because of their historical significance, have rather different parameters despite the fact that they belong to the same family. For the years 1900 and 1914, the error of fitting is higher, meaning that they might belong to a different category of curves.

The same approach can be conducted with linear monotonic curves and “squared” curves.

We believe that inversing the problem of n-gram visualization by enabling automatic search in a space of curves could profoundly transform research in this area, going beyond the intuitive search for remarkable curves. It is likely that many different levels of information, combining semantics and grammatical constraints with historical contexts, are implicitly coded in n-gram temporal profiles.

Understanding how to classify and study these curves is important for harnessing the power of this set of statistical tools. The solution based on families of simple archetypical curves briefly described in this article is certainly not the only way of approaching this question but constitutes an initial attempt to demonstrate the potential of this overall research goal.

References

1. J.-B. Michel et al. *Quantitative Analysis of Culture Using Millions of Digitized Books*. Science 331, 17, 2011. DOI: 10.1126/science.1199644.

2. J.-P. Delahaye and N. Gauvrit. *Culturomics: Le Numérique Et La Culture*. O. Jacob, Paris (France), 2013.

3. I. C. Trelea, *The particle swarm optimization algorithm: convergence analysis and parameter selection*, Information Processing Letters, Volume 85, Issue 6, Pages 317-325, 2003. ISSN 0020-0190, http://dx.doi.org/10.1016/S0020-0190(02)00447-7.

4. C. Bauckhage , K. Kersting and F. Hadiji. *Mathematical Models of Fads Explain the Temporal Dynamics of Internet Memes*. International AAAI Conference on Weblogs and Social Media, 2013.

5. M. I. Fraga Alves and C. Neves. *Extreme Value Distributions*. International Encyclopedia of Statistical Science (pages 493-496), Lovric, Miodrag (Ed.), Springer-Verlag, 2010. ISBN: 978-3-642-04897-5; DOI: 10.1007/978-3-642-04898-2.