Time series analysis of the dynamics of news websites

Maria Carla Calzarossa
Dipartimento di Ingegneria Industriale e Informazione
Università di Pavia
via Ferrata 1 – I-27100 Pavia, Italy
mcc@unipv.it

Daniele Tessera
Dipartimento di Matematica e Fisica
Università Cattolica del Sacro Cuore
via Musei 41 – I-25121 Brescia, Italy
daniele.tessera@unicatt.it

***Abstract*—The content of news websites changes frequently and rapidly and its relevance tends to decay with time. To be of any value to the users, tools, such as, search engines, have to cope with these evolving websites and detect in a timely manner their changes. In this paper we apply time series analysis to study the properties and the temporal patterns of the change rates of the content of three news websites. Our investigation shows that changes are characterized by large fluctuations with periodic patterns and time dependent behavior. The time series describing the change rate is decomposed into trend, seasonal and irregular components and models of each component are then identified. The trend and seasonal components describe the daily and weekly patterns of the change rates. Trigonometric polynomials best fit these deterministic components, whereas the class of ARMA models represents the irregular component. The resulting models can be used to describe the dynamics of the changes and predict future change rates.**

***Keywords*-Web dynamics; time series analysis; search engines; news websites**

I. Introduction

The explosive growth and variety of content available on the Web coupled with the complex types of user interactions open up new significant challenges. Powerful technologies able to efficiently retrieve Web content and adapt to its highly dynamic nature and usage patterns need to be devised. Among these technologies, search engines play a key role as they represent the primary entry point of many users into the Web.

The content of a website can change under three different circumstances: upload of new documents, update of the content of existing documents, removal of existing documents. Of course, the frequency of changes varies from site to site; while several sites are rather static, that is, their content seldom changes, other sites, such as, news websites, are very dynamic, that is, their content changes frequently and rapidly. Hence, search engines need to refresh and index in a timely manner the content of these websites to be of any value to the users. Nevertheless, the policies implemented by search engines are usually driven by some conflicting requirements: they have to provide a good coverage of the sites, maximize the freshness of their content and at the same time minimize the costs for downloads, storage and management. It is then important for these tools to be able to predict how often and to what extent the content of the site changes as to adjust their crawling activities accordingly.

In this paper we study the temporal patterns of the changes of the content of news websites with the objective of identifying models able to capture and predict their evolution. Changes are represented as time series, that is, ordered sequences of observations, whose analysis summarizes and models their properties and temporal patterns. The models of the temporal behavior of the change rates will be very useful for refining the crawling policies of search engines as to present users with the newest available content.

The choice of news websites is motivated by their time-sensitive content whose relevance tends to decrease with time. Hence, their evolution is particularly interesting from search engines perspective. Moreover, we remark that even though the models refer to news websites, the applicability of the methodological approach proposed in this paper encompasses any type of website characterized by a dynamic behavior.

The paper is organized as follows. After an overview of the literature on Web dynamics given in Section II, the methodological approach followed for the analysis of the evolution of the news websites is introduced in Sect. III. The experimental results are described in Section IV. Finally, Section V summarizes the major findings and outlines future research directions.

II. Related work

The design of search engines has to take into account the characteristics of websites and in particular their dynamics. In the literature, Web dynamics have been studied under some different perspectives, see, e.g., [1], [2], [3], [4], [5], [6], [7], [8]. For example, in [3], authors focus on growth and update dynamics, whereas a survey of the major research challenges on Web dynamics in the framework of four dimensions, namely, size, pages, link structures, and user interests, is presented in [5].

Other aspects related to how much a Web document changes and whether changes are clustered, are investigated in [1] where authors define two measures that quantify these changes and discuss the implications of this approach for Web information system maintenance. A fine grain

characterization of the evolution of Web content is presented in [9]. More specifically, the analysis focuses on the nature of changes, that is, changes to content and structure of Web pages, by considering both frequency and amount of change as a function of the page types. The outcome of this study deals with the identification of stable and dynamic content within each page.

The concept of longevity of the information found on the Web, that is, the lifetime of the content that appears and disappears from the Web, is introduced by Olston and Pandey in [10]. Authors consider this characteristic as a key aspect for the description of Web evolution and for the development of effective crawling policies. In particular, starting from the observation that there is no correlation between information longevity and change frequency, they propose a generative model for dynamic Web content, where pages are seen as a set of independent fragments each characterized by its own temporal behavior and change profile.

The amount and type of changes to Web page content and the corresponding user visit behavior are studied in [11] to discover relationships and associations between Web dynamics and user accesses. In this context, authors define metrics and methods to characterize these patterns and identify their relationships. The paper shows that different visit patterns resonate with different kinds of change, for example, with the rate of change of interesting content.

Our main contribution is in the framework of predicting Web dynamics and deals with the application of time series analysis techniques to the identification of models able to represent the temporal behavior of the change rates of news websites.

III. Methodological approach

The methodological approach adopted to study the dynamics of news websites relies on time series analysis. The choice of time series is motivated by their ability to capture the temporal patterns that characterize the change rates of the sites. Let us recall that a time series is a collection of ordered observations of a random variable X taken sequentially in time at equally spaced time intervals, that is, $\{X\} = \{x_{t_1}, x_{t_2}, ..., x_{t_n}, ...\}$ with $t_1 \leq t_2 \leq ... \leq t_n \leq ...$ [12].

Our methodology consists of several steps aimed at identifying and understanding the temporal patterns of the change rates and predicting their future behavior based on past behavior.

The first step focuses on a descriptive analysis with the objective of discovering the properties of the observations, for example, whether there are unusual observations, such as peaks that can be considered as outliers and removed from the data. Moreover, the statistical dependence between pairs of observations needs to be investigated. In this respect, the autocorrelation function is a very powerful tool to assess the presence of time dependent observations and their degree of dependence. As a further analysis, plots of the observations as a function of time provide a general view of the phenomenon and highlight its time variations, that is, whether there is any trend, indicating a long-term change in the mean level, that is, a tendency to grow or decrease rather steadily over quite long periods of time, or seasonality, showing a pattern that repeats in time with some periodicity, e.g., hourly, daily, weekly, monthly.

The main goal of the following explanatory step is to isolate and remove the components previously identified, namely, trend and seasonal components, such that after adjusting the original series with respect to these components, the remaining variability, i.e., the random noise or irregular component, is a stationary process without any deterministic and predictable trend or seasonal effect and with stable fluctuations.

Various approaches, such as, smoothing techniques based on the Loess approach, that is, locally weighted polynomial regression, can be adopted for the decomposition of the time series [13]. Each component will then be estimated separately to derive a mathematical model able to explain the observed variability. Numerical fitting techniques are well suited for estimating the deterministic components, whereas the estimation of the irregular component relies on the class of the Auto Regressive Moving Average (ARMA) models. We recall that an $ARMA(p, q)$ model of a time series z_t at time t can be expressed as:

$$z_t = \sum_{i=1}^{p} \phi_i z_{t-i} + \omega_t + \sum_{i=1}^{q} \theta_i \omega_{t-i}$$

where p and q denote the orders of the autoregressive and moving average terms, ω_t refers to the white Gaussian noise of the time series, ϕ_i and θ_i are the p autoregressive parameters and the q weights of the white noise of the model, respectively.

Note that the principle of parsimony should drive the identification of the order of the autoregressive and moving average terms of the ARMA model. Of course, models with a smaller number of parameters are preferable to bigger models as long as both have similar predictive power.

The final step of our methodological approach deals with forecasting, that is, based on the identified model we can predict future observations of the change rate.

IV. Experimental results

In this section, we present the experimental results obtained by applying the methodology introduced in Sect. III. In particular, we focus on the temporal behavior of three major news websites, namely, CNN [14], MSNBC [15] and Reuters [16] websites. To capture their dynamics, we monitored these sites for several weeks by taking snapshots every 15 minutes and downloading at each snapshot all documents published on the sites. The main characteristics

of our crawling activities are summarized in Table I. Of

	Number of documents	Crawling interval
CNN	8,302	104 days
MSNBC	5,436	84 days
Reuters	11,570	91 days

Table I
MAIN CHARACTERISTICS OF THE CRAWLING ACTIVITIES ON THE THREE NEWS WEBSITES.

these documents, some were modified once or more times after they were published on the sites, whereas the content of some others never changed. For example, we have detected some 9,482 updates to about half of the documents uploaded on the CNN website, that is, 2.5 updates per document, whereas the content of only one fourth of the documents of the Reuters website was updated.

Let us remark that due to space limitations, in what follows, we mainly present the results obtained for the CNN news website.

A snapshot over a two weeks period of the behavior of the change rate, that is, the number of changes per hour, of the content of the CNN website is shown in Figure 1. Note that in our analysis changes refer to uploads of new

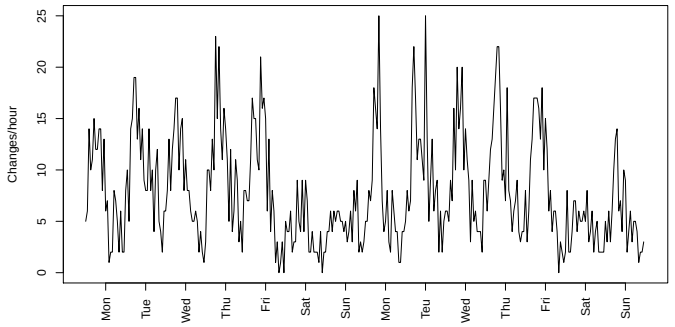

Figure 1. Change rate over a two weeks period of the content of the CNN website as a function of time.

documents and updates of existing documents. We do not consider document removal as documents are often moved to some sort of off-line archives and seldom disappear from the site.

As we can see from the figure, changes are characterized by large fluctuations from hour to hour, especially with respect to day and night hours, and from day to day, especially with respect to weekdays and week-end days. Moreover, we can observe both daily and weekly periodic patterns as well as a time dependent behavior. These observations are supported by the corresponding autocorrelation function computed with one hour time lags (see Fig. 2). The horizontal dashed lines plotted in the figure correspond to the reference values of the independence test at 5% significance

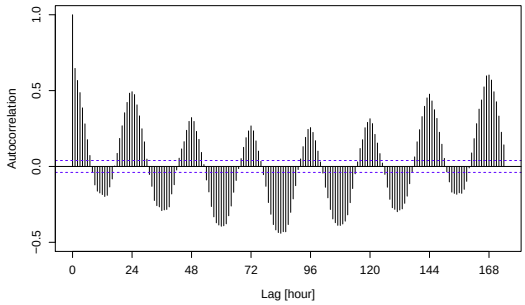

Figure 2. Autocorrelation function at one hour time lag of the change rate of the content of the CNN website.

level. The test fails as most of the values fall outside these boundaries. This means that autocorrelation effects have to be included in the following analysis.

These results and the analysis of the temporal behavior of the change rates of the three news websites lead us to represent the rates with time series whose observations are taken every hour over 24 hours intervals. Moreover, because of time dependence, we apply the standard decomposition of the time series into a seasonal, a trend and an irregular component, using an additive approach. In particular, by applying the Loess approach, we estimate the seasonal and trend components with moving windows of a day and a week, respectively. Figure 3 shows the decomposition obtained for the snapshot of the time series plotted in Fig. 1 and referring to the CNN change rate.

We notice that the trend component exhibits a periodic behavior that captures the weekly patterns of the changes of the content of the website, with much fewer changes during week-end days than during weekdays, whereas the seasonal component captures their daily patterns, whose hourly rate varies as a function on the time of the day

The figure includes the plot of the irregular component, i.e., the remainder of the time series that takes into account the fluctuations not described by the deterministic components.

A snapshot of the change rate over a two weeks period of the MSNBC news website together with the decomposition of the corresponding time series is plotted in Figure 4. We can easily identify daily and weekly patterns whose temporal behavior is rather similar to what detected for the CNN website.

The descriptive step is followed by the explanatory step, dealing with the estimation in the frequency domain of the components previously identified. In details, to model the deterministic components of the time series we use

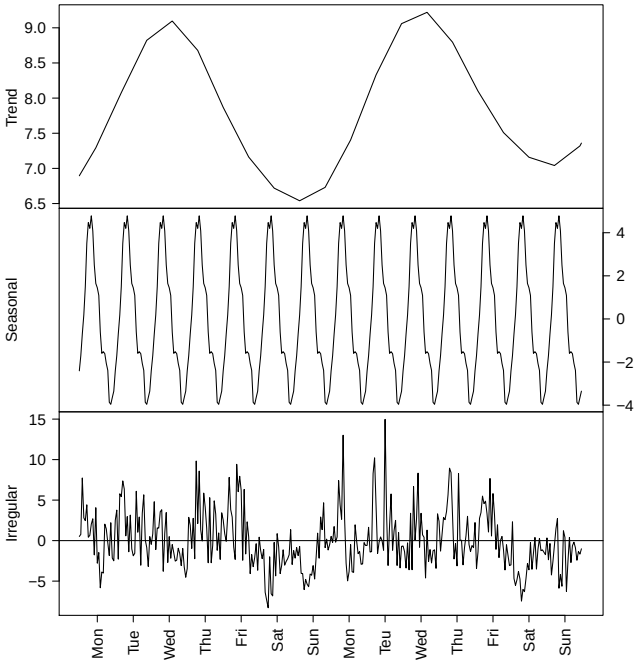

Figure 3. Decomposition of the time series shown in Fig. 1 into seasonal, trend and irregular components.

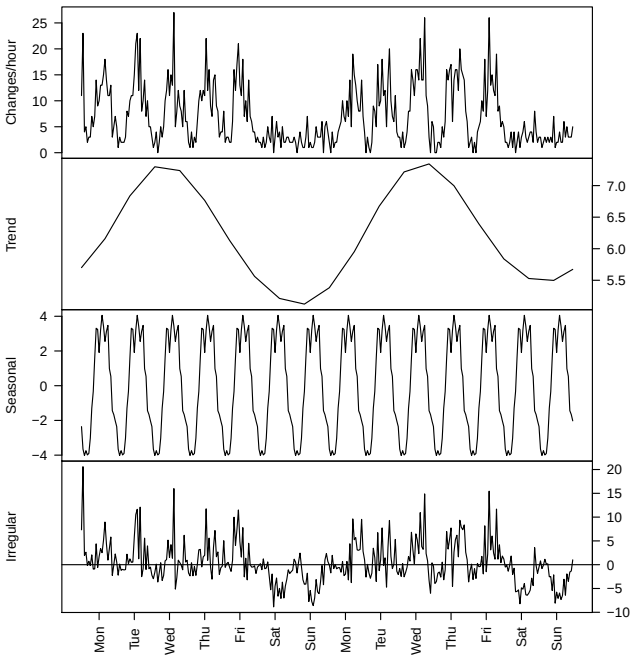

Figure 4. Change rate over a two weeks period of the content of the MSNBC website and decomposition of the corresponding time series.

trigonometric polynomials of the form

$$a_0 + \sum_{i=1}^{n} \left(a_i sin\left(2\pi i \frac{t}{T}\right) + b_n cos\left(2\pi i \frac{t}{T}\right)\right)$$

where n denotes their degree and T the period. Spectral analysis and numerical fitting techniques identify a polynomial of degree four as the best fit of the seasonal component of the CNN change rate. Table II summarizes the corresponding parameters. Note that the intercept a_0 is equal to zero be-

a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4
2.99	-2.31	-0.24	-0.86	0.05	0.02	0.35	0.21

Table II
PARAMETERS OF THE TRIGONOMETRIC POLYNOMIAL FITTING THE SEASONAL COMPONENT OF THE CNN TIME SERIES.

cause the mean of the seasonal component is zero. Moreover, we remark that we ignore two of the parameters of the polynomial, namely, a_3 and b_3, because their values are definitely too small and they are not considered significant for the overall model accuracy by standard model selection tests. Hence, the polynomial model plotted in Fig. 5 does only include six parameters.

A trigonometric polynomial of degree three, whose parameters are listed in Table III, best fits the seasonal com-

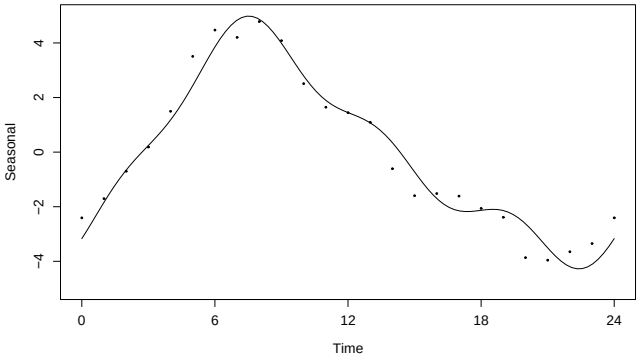

Figure 5. Seasonal component (dotted pattern) and corresponding model (solid curve) obtained for the CNN website.

ponent of the Reuters change rate. By applying the model selection tests, we discover that parameters a_2 and b_2 can be discarded, thus, we represent the model with four parameters only. In the case of the seasonal component of the MSNBC

a_1	b_1	a_2	b_2	a_3	b_3
0.30	-3.84	-0.09	-0.09	0.52	-0.12

Table III
PARAMETERS OF THE TRIGONOMETRIC POLYNOMIAL FITTING THE SEASONAL COMPONENT OF THE REUTERS TIME SERIES.

change rate, the model is a trigonometric polynomial of

degree two, described by three parameters because parameter b_2 can be ignored.

Similarly, we use trigonometric polynomials to model the trend components. On the contrary, the properties of the irregular component analyzed in terms of partial autocorrelations lead us to consider an ARMA model. Autocorrelations are also used to identify the orders p and q of the model. By fitting the irregular component of the time series with ARMA models of increasing orders, we find that the best fit corresponds to a seasonal ARMA model that takes into account the daily autocorrelated patterns, that is, an $ARMA(1,1) \times (1,1)_{24}$. Figure 6 plots, over a two weeks period, the final model obtained for change rate of the CNN website against the original time series.

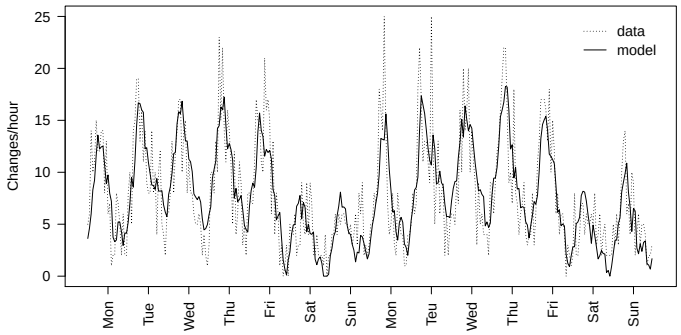

Figure 6. Model of the change rate of the CNN website.

For the irregular components derived for the change rates of the content of the MSNBC and Reuters websites, we obtain an $ARMA(1,1)\times(1,1)_{24}$ and an $ARIMA(0,1,1)\times(0,0,2)_{24}$, respectively. Note that the use of an Auto Regressive Integrated Moving Average model for the irregular component of the Reuters change rate is motivated by the moderate level of non stationarity of this component.

Once we have obtained these models, we can use them to predict future observations of the change rates.

V. Conclusions

The content available on the Web is rapidly evolving and this evolution needs to be detected in a timely manner. Our study has investigated the properties and the behavior of the changes of the content of three major news websites as a function of time. The decomposition of the time series used to describe the change rates has shown some periodic patterns characterizing these websites. In particular, the seasonal and trend components of the time series capture daily and weekly patterns, where the change rate varies continously from hour to hour and from day to day, with significant differences between day and night hours and between weekdays and week-end days. These time dependences are taken into account by the class of the ARMA models used to fit the irregular components of the time series. The final models accurately fit the empirical patterns of the change rates and can be used as good predictors of the future behavior of the changes.

As future work, we plan to investigate to what extent the content of a website changes whenever it is modified as to assess the novelty of consecutive snapshots and drive the crawling policies of search engines.

References

[1] L. Lim, M. Wang, S. Padmanabhan, J. Vitter, and R. Agarwal, "Characterizing web document change," in *Advances in Web-Age Information Management*, ser. Lecture Notes in Computer Science, X. Wang, G. Yu, and H. Lu, Eds. Springer, 2001, vol. 2118, pp. 133–144.

[2] W. Koehler, "Web page change and persistence – A four-year longitudinal study," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 2, pp. 162–171, 2002.

[3] K. Risvik and R. Michelsen, "Search engines and web dynamics," *Computer Networks*, vol. 39, no. 3, pp. 289 – 302, 2002.

[4] R. Baeza-Yates, C. Castillo, and F. Saint-Jean, "Web dynamics, structure and page quality," in *Web dynamics: Adapting to change in content, size, topology and use*, M. Levene and A. Poulovassilis, Eds. Springer, 2004, pp. 93–109.

[5] Y. Ke, L. Deng, W. Ng, and D.-L. Lee, "Web dynamics and their ramifications for the development of web search engines," *Computer Networks*, vol. 50, no. 10, pp. 1430–1447, 2006.

[6] S. Kwon, S. Lee, and S. Kim, "Effective criteria for web page changes," in *Frontiers of WWW Research and Development - APWeb 2006*, ser. Lecture Notes in Computer Science, X. Zhou, J. Li, H. Shen, M. Kitsuregawa, and Y. Zhang, Eds. Springer, 2006, vol. 3841, pp. 837–842.

[7] M. Calzarossa and D. Tessera, "Characterization of the evolution a news Web site," *Journal of Systems and Software*, vol. 81, no. 12, pp. 2236–2344, 2008.

[8] ——, "An exploratory analysis of the novelty of a news Web site," in *Proc. International Symposium on Performance Evaluation of Computer and Telecommunication Systems – SPECTS 2010*. SCS Press, 2010, pp. 399–404.

[9] E. Adar, J. Teevan, S. T. Dumais, and J. Elsas, "The web changes everything: Understanding the dynamics of web content," in *Proc. of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*. ACM, 2009, pp. 282–291.

[10] C. Olston and S. Pandey, "Recrawl scheduling based on information longevity," in *Proc. of the 17th International Conference on World Wide Web - WWW '08*. ACM, 2008, pp. 437–446.

[11] E. Adar, J. Teevan, and S. Dumais, “Resonance on the Web: Web dynamics and revisitation patterns,” in *Proc. of the 27th International Conference on Human factors in computing systems - CHI ’09*. ACM, 2009, pp. 1381–1390.

[12] G. E. P. Box and G. M. Jenkins, *Time Series Analysis, Forecasting, and Control*. Holden-Day, 1976.

[13] R. Cleveland, W. Cleveland, J. McRae, and I. Terpenning, “STL: A Seasonal-Trend Decomposition procedure based on Loess (with discussion),” *Journal of Official Statistics*, vol. 6, pp. 3–73, 1990.

[14] CNN International Edition Web site, http://edition.cnn.com.

[15] MSNBC Web site, http://www.msnbc.com.

[16] Reuters Web site, http://www.reuters.com.