A methodological approach for time series analysis and forecasting of web dynamics

Maria Carla Calzarossa[1], Marco L. Della Vedova[2], Luisa Massari[1], Giuseppe Nebbione[1], and Daniele Tessera[2]

[1] Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Pavia, Italy, {mcc, luisa.massari}@unipv.it, giuseppe.nebbione01@ateneopv.it
[2] Dipartimento di Matematica e Fisica, Università Cattolica del Sacro Cuore, Brescia, Italy, {marco.dellavedova, daniele.tessera}@unicatt.it

Abstract. The web is a complex information ecosystem that provides a large variety of content changing over time as a consequence of the combined effects of management policies, user interactions and external events. These highly dynamic scenarios challenge technologies dealing with discovery, management and retrieval of web content. In this paper, we address the problem of modeling and predicting web dynamics in the framework of time series analysis and forecasting. We present a general methodological approach that allows the identification of the patterns describing the behavior of the time series, the formulation of suitable models and the use of these models for predicting the future behavior. Moreover, to improve the forecasts, we propose a method for detecting and modeling the spiky patterns that might be present in a time series. To test our methodological approach, we analyze the temporal patterns of page uploads of the Reuters news agency website over one year. We discover that the upload process is characterized by a diurnal behavior and by a much larger number of uploads during weekdays with respect to weekend days. Moreover, we identify several sudden spikes and a daily periodicity. The overall model of the upload process – obtained as a superposition of the models of its individual components – accurately fits the data, including most of the spikes.

Keywords: web dynamics, temporal patterns, time series analysis, forecasting, performance modeling, search engines, ARMA models.

1 Introduction

The web is a huge repository of information that provides users with an enhanced experience by combining many different content forms, e.g., text, audio, images, video, animations. This complex information ecosystem is regularly updated to keep the content fresh and attract at the same time the interest of the users. New pages are uploaded, existing pages are updated and eventually removed.

All these changes are often the result of combined effects that involve the management policies of the websites, the behavior of the users as well as external events. For example, news websites are generally updated to report the

latest news stories and their developments as well as to keep the websites "alive". The changes of social media websites are mainly driven by the activities and interactions of their users who post and share content, add comments and likes. Corporate websites are periodically updated to advertise and promote the companies and their business and improve customer perception and search engine rankings.

These highly dynamic scenarios challenge all technologies aimed at discovery, retrieval and management of web content and in particular search engines. In fact, to avoid wasting resources, reduce bandwidth usage and server load and keep web pages fresh, these technologies need to adjust their crawling policies according to the dynamics of the websites [4, 15]. Hence, it is necessary to derive accurate predictions of the frequency and extent of website changes.

The problem of predicting the future behavior of a phenomenon based on its past behavior can be addressed under different perspectives [18]. In this paper we investigate this problem in the framework of time series analysis – a popular method used for modeling and making forecasts of temporal data. More precisely, we present a general methodological approach for studying the dynamics of any phenomenon that can be described by a time series. In fact, even though time series analysis and forecasting techniques are well defined, their application requires particular care. Our approach tries to overcome this issue by addressing time series analysis as a sequence of steps dealing with the characterization of the overall statistical properties of the time series, the identification of the underlying patterns describing its behavior, the formulation of suitable models and finally the use of these models for predicting the future behavior. Moreover, we include in the framework a novel approach for accurately detecting and modeling the spiky patterns that might be present in a time series.

As an application of the proposed approach, we investigate the dynamics of the Reuters news agency website[3]. Nevertheless, we outline that this approach is general enough and can be easily applied to study and predict the dynamics of various types of web services and applications (e.g., content delivery, video streaming, mobile apps and embedded ads) as well as of the traffic they generate.

In this paper we focus on the analysis of the time series representing the patterns of page uploads. In fact, for news websites these patterns are usually characterized by a time-dependent behavior with well defined periodicity and large fluctuations. Hence, to predict future uploads from past uploads, it is critical to identify models that accurately explain these behaviors.

We summarize our contributions as follows:

- definition of a methodological framework for time series analysis and forecasting,
- identification and modeling of spiky/bursty patterns, and
- application of the proposed approach to study the dynamics of the Reuters news agency website.

The layout of this paper is the following: Section 2 reviews the state of the art, while Section 3 presents the methodological approach proposed for time series

[3] http://www.reuters.com

analysis and forecasting. The dataset considered in the study and the results of the analysis and prediction of the dynamics of the Reuters website are addressed in Sections 4 and 5, respectively. Section 6 summarizes the paper and outlines possible research directions.

2 Related work

The problem of estimating and predicting web dynamics has been studied under different – although complementary – angles. Some works specifically focused on the changes of individual web pages (see, e.g., [1, 13, 17, 20, 22]), while others studied the overall evolution of websites (see, e.g., [3, 5–10]). These works have important implications on content reuse and caching and more generally on information retrieval technologies.

In the framework of page changes, the extensive analysis presented by Fetterly et al. [13] suggests that changes of web pages are somehow correlated, thus future changes can be easily predicted from past changes. Similarly, Shi et al. [22] outline that within news and e-commerce websites, objects are characterized by different freshness times with most objects that do not change within the timescale of a week and fewer objects that change within the timescale of a day.

Lim et al. [17] analyze and quantify consecutive changes of individual pages by means of two measures, namely, distance and clusteredness measures. Their study shows that in general changes are small and rather clustered. A similar approach has been applied in [5] to assess the extent of page changes and adjust the models of change rates of the websites accordingly. Measures, such as edit distance, cosine coefficient of similarity, are used for this purpose.

Content change prediction is addressed by Radinsky and Bennet [20] through an expert predictive framework that takes into account various features, such as degree and relationships among changes and similarity in the types of changes. A temporal modeling framework that captures the dynamic nature of Web behaviors is presented in [21]. The proposed models include the typical characteristics observed in query and URL click behavior of Web searchers, that is, trend, periodicity and surprise disruptions.

The temporal patterns of the content changes of three major news webites have been studied in [8]. The patterns of each website are represented as periodic time series whose models explain their dynamics and are the basis for the forecasting.

Yang and Leskovec [23] investigate the temporal patterns associated with online textual content by formulating a time series clustering problem that allows the identification of the shapes characterizing different types of media.

The problem of predicting the time between changes of web pages under blind sampling is addressed by Li et al. [16]. A stochastic modeling framework where updates and sampling follow independent point processes is proposed.

An interesting survey on different approaches applied for quantifying changes and predicting their frequency and dynamics is provided by Oita and Senellart [19].

In this work, we address the problem of modeling and predicting the dynamics of web content changes by devising a systematic methodological framework based on time series analysis. This framework is general and can be easily applied for investigating the characteristics of any temporal data and make forecasts.

3 Methodological framework

A time series is a sequence of discrete or continuous observations collected at equally spaced time intervals, i.e., $\{Y_t\} = \{y_{t_1}, y_{t_2}, ..., y_{t_N}\}$ with $t_1 \leq t_2 \leq ... \leq t_N$ [14]. As already pointed out, although the techniques for time series analysis and forecasting are well defined, for a proper application of these techniques it is necessary to define a systematic methodological approach.

The workflow of Figure 1 summarizes the methodological framework proposed in this paper. Starting from the background knowledge of the phenomenon being investigated and from the raw data transformed into a time series, it is necessary to gain some preliminary insights into the behavior of this time series through an exploratory data analysis (EDA) of its temporal patterns. In particular, from the statistical properties of the time series it is possible to explain the inherent structure that has to be modeled. These models are then used to make forecasts. The details of each of these steps are presented in what follows.

3.1 Exploratory data analysis

The exploratory analysis of the data is an important step for understanding the overall behavior and the statistical properties of the time series under investigation. Visualization and statistical techniques work for this purpose.

More precisely, the exploratory analysis includes the computation of descriptive statistics, such as mean, percentiles, autocorrelations. In particular, the autocorrelation function at varying time lags is particularly useful in the analysis of a time series since it suggests how similar a sequence is to its previous values. Moreover, autocorrelations allow for checking the randomness of the data and assessing the stationarity of the time series.

In addition, visualization techniques are applied to obtain an overview of the temporal patterns of the time series (see Fig. 2). Their visual inspection highlights recognizable patterns, such as trend, seasonal or cyclic. The trend denotes steadily increasing or decreasing patterns over quite long periods of time. The seasonality denotes a behavior that repeats in time on a regular basis over a fixed period, e.g., each month, each year. On the contrary, a cyclic pattern denotes a behavior that repeats over a variable period.

Time series visualization is also very useful for recognizing sudden rises followed by falls in the data. The nature of these spiky or bursty patterns depends on the intrinsic characteristics of the phenomenon described by the time

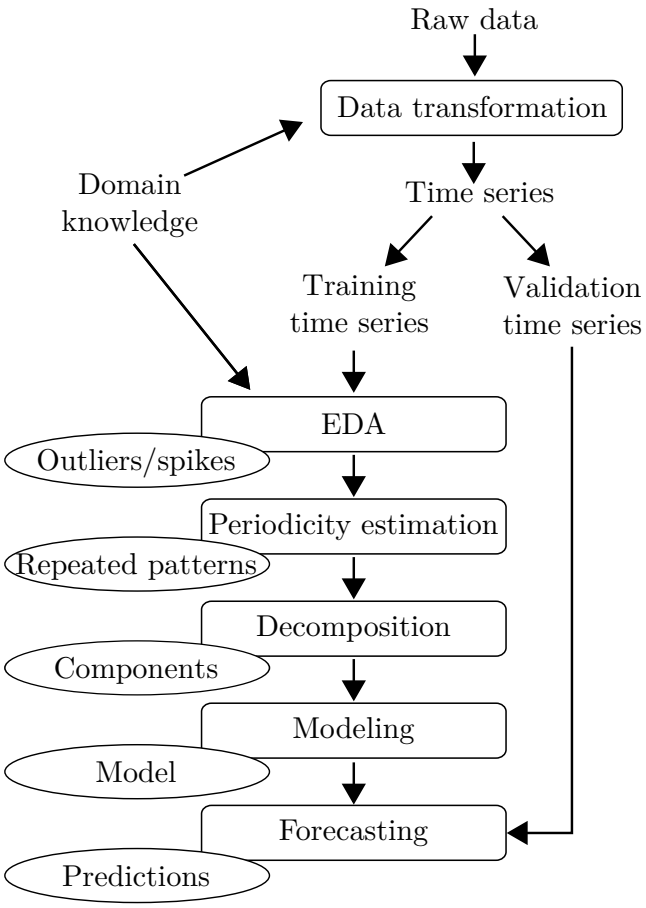


Fig. 1. Methodological framework for time series analysis and forecasting.

series. They might represent typical behaviors or anomalous behaviors, thus corresponding to potential outliers. We recall that outliers are defined as the observations in the series that are significantly different from the rest of the observations.

Statistical measures, such as median absolute deviation, Z-score, are applied for the identification of outliers and more generally of spiky patterns.

All these patterns must be treated with particular caution since they might affect the time series analysis and have negative effects on its models. In general, once the patterns have been identified, a good practice is to remove the corresponding observations from the data and replace them with observations obtained by interpolation over neighbor observations. Nevertheless, as we will discuss in Sect. 3.3, the presence of typical spiky patterns has to be properly included in the final model of the time series.

3.2 Periodicity estimation

The detection of periodic behaviors in a time series is another important step toward time series modeling. Hence, the periods – usually not known a priori – have to be accurately estimated.

Spectral analysis is a popular method used for this purpose. This is because this method characterizes the frequency representation of a signal. Peaks in the frequency domain will correspond to periods in the time domain. Thus, by analyzing peaks and finding the dominant frequencies, it is possible to estimate the periods of the repeated temporal patterns.

More precisely, the spectral analysis applied to the autocorrelation function of the time series relies on the computation of the discrete Fourier coefficients f_k associated with the k/T frequencies, that is:

$$f_k = \sum_{j=0}^{N-1} y_{t_j} e^{-i2\pi \frac{j}{N} k} \ , \ k = 0, 1, 2, ..., N-1.$$

The power spectrum density – represented by the absolute value of each Fourier coefficient – highlights the peaks in the spectrum of the autocorrelation function.

3.3 Decomposition

Time series decomposition is primarily applied to better understand its properties, exploring its behavior over time and improve forecasts. In general, a time series exhibits a huge variety of patterns whose classification is at the basis of the decomposition. In fact, the components have to correspond to the underlying pattern categories.

A classical decomposition approach of a time series relies on an additive model that includes deterministic parts, e.g., the trend and seasonal components, and stochastic parts, e.g., the irregular component corresponding to the random noise. Hence, the time series Y_t is given by $Y_t = T_t + S_t + \epsilon_t$, where T_t, S_t and ϵ_t denote the trend, seasonal and irregular components, respectively.

Depending on the characteristics of the time series, smoothing techniques, such as moving average, exponential smoothing, locally weighted polynomial regression, Loess regression, are applied for identifying these components [11].

The estimation of the deterministic components is obtained by fitting appropriate models to the data, while the estimation of the stochastic component – depending on its statistical dependence and random behavior – relies on techniques, such as moving average, auto regressive, Holt-Winters, Box and Jenkins [2].

Another important step proposed in this methodology to improve the forecasts is aimed at including in the final model of the time series the contribution of the spiky patterns identified by the exploratory data analysis (see Sect. 3.1). For this purpose, it is necessary to detect and model the temporal behavior of these patterns. In particular, classifiers (e.g., decision trees, logistic regression) applied to some short term historical data of the time series allow for predicting spikes. By fitting these models, we estimate the probability associated with a

future observation being a spike. Moreover, these patterns – depending on their behavior – are described by simple models, such as split, tailing, fronting. The time series final model is then adjusted by adding the contribution of the model chosen to represent the patterns.

3.4 Forecasting

The final step of the methodological framework deals with making forecasts using the models previously identified. This step is rather straightforward. In fact, the predicted value of the time series $\hat{Y}_{t+h}$ at time $t+h$ is obtained by superimposing the values predicted by these models. In detail, for the deterministic components, the new values are extrapolated from the corresponding models computed at time $t+h$. On the contrary, approaches, such as the Box-Jenkins approach, are applied to compute the forecasts of the stochastic component, while the forecasts of the spiky patterns rely on classification techniques applied to short term historical data.

The evaluation of the performance of the forecasts at varying time lags h is based on standard measures of accuracy (e.g., mean error, mean absolute deviation, mean absolute percent error, mean squared error and its square root).

4 Dataset

To test our methodological approach, we analyzed the temporal patterns of the uploads of new pages on the Reuters news agency website over one year. In what follows we describe the dataset considered in this study and its main characteristics.

4.1 Description

The dataset relies on a publicly available unofficial Reuters dataset[4] that stores information about the archival time of the web pages together with their title – referred to as news title in what follows – and the corresponding URL. From this huge dataset – that spans several years from 2007 until 2016 – we extracted the data of 50 weeks since January 4, 2015 that refers to 893,905 pages.

Before applying our methodology, we applied some preliminary transformations to this raw data (see Fig. 1). In particular, since we were interested in modeling and predicting the dynamics of the upload process of new pages rather than their archival process – which is usually of little interest for search engines and similar technologies – it was necessary to adjust the timestamps associated with the pages. For this purpose, we crawled the Reuters website – using the URLs stored in the dataset – to discover the actual publish time of the web pages. In detail, to avoid overloading the website, we applied this process to a sample of 13,546 pages, that is, about 1.5% of the pages. For each of them, we

[4] https://github.com/philipperemy/Reuters-full-data-set

extracted the `og:article:published_time` metadata tag[5] used to specify when the page was first published. We discovered that a page is archived on average 19.65 hours after its upload. As expected, the archival process is rather deterministic: the corresponding standard deviation is only 0.47. Hence, by subtracting this average from the archival time, we obtained an accurate estimation of the publish time – used in what follows to describe the upload process.

Another step of the data transformation process deals with approximately 30,000 news titles including the keyword "UPDATE". A manual inspection of a sample of the corresponding pages has shown that these pages were updated once or multiple times after their first upload. Hence, not to mix the upload and update processes, we discarded these observations. The resulting dataset consists of the data of 864,304 pages.

4.2 Characteristics

The exploratory analysis of the data is aimed at gaining some preliminary insights into the time series describing the behavior of the page upload process and into the content of the news titles. We first characterized the dynamics of the website in terms of number of page uploads per day. Figure 2 shows the temporal patterns of this time series over the 50 weeks analyzed in this study. We notice large fluctuations, where the number of uploads per day ranges from 206 up to 4,822 and it is much lower during weekends with respect to weekdays. On average about 3,300 pages are uploaded during a weekday, whereas only about 355 during weekend days (see Table 1 for the details).

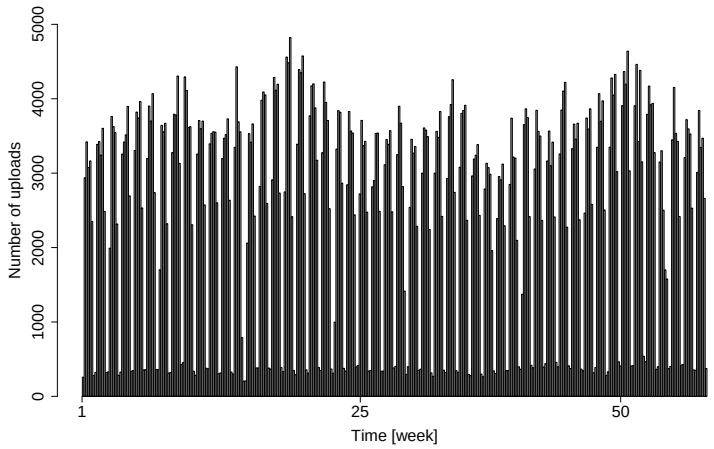


Fig. 2. Temporal patterns of the time series representing the number of uploads per day over 50 weeks.

[5] http://ogp.me/

	Mean	St. dev.	Min	Max
Weekdays	3,315.15	681.9	790	4,822
Weekend days	355.16	53.6	206	540
Overall	2,469.44	1,458	206	4,822

Table 1. Basic statistics of the number of uploads per day broken down for weekdays and weekend days.

The analysis of the number of uploads at a finer granularity, i.e. per hour, confirms these findings, namely, big differences between weekdays and weekend days (see Fig. 3). This was expected since the Reuters website is mainly focused

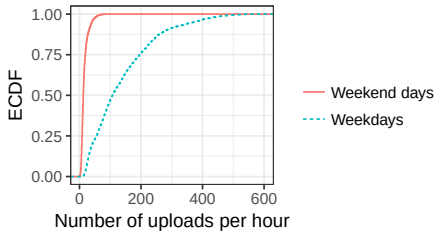


Fig. 3. Cumulative distribution function of the number of page uploads per hour for weekdays and for weekend days.

on business and financial news.

An overview of the temporal patterns of the time series over three weeks is shown in Figure 4. The figure clearly suggests diurnal patterns characterized by sudden spikes with as many as 500 uploads in an hour. We outline that the values of approximately 10% of the observations of the overall time series exceed 250, while only 1% exceed 450.

This spiky behavior is also highlighted in the boxplot of Figure 5 showing the number of uploads per hour for each day of the week across all weeks. In general, Tuesdays are characterized by the largest variability. Moreover, the website is more active during the mid days of the week.

This characterization, together with the business-oriented focus of the news published on the Reuters website, has suggested that the dynamics of the website is mainly relevant during weekdays. Hence, the time series analysis addresses the page upload dynamics over weekdays only. We analyze the data of 828,788 pages – accounting for approximately 96% of the data.

Another interesting aspect considered in the exploratory data analysis deals with news titles. Although this analysis is not strictly related to the dynamics of the website, it provides some insights in the content of the news. In particular, we analyzed these titles in terms of the words they consist of, i.e., the single units of textual information (tokens).

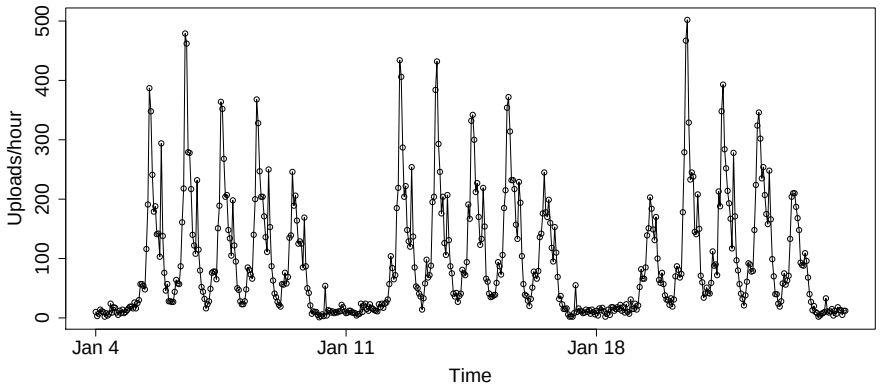


Fig. 4. Temporal patterns of the number of uploads per hour over three weeks.

The number of words per title does not significantly vary: 62% of the titles consist of a number of words between 8 and 13. On average a title includes 10.91 words with a standard deviation of 4.38 words.

To avoid redundancy, inflectional and related forms of a word, we applied Natural Language Processing (NLP) techniques, such as tokenization, stop words removal and stemming [12]. After this process, we obtained 124,914 unique stems (out of 319,165 unique word). The distribution of the popularity of the top 30 stems is shown in Figure 6. These stems account for 950,736 occurrences, that correspond to 13.4% of the total number of occurrences. In particular, the most popular stem, i.e., *announc* occurs 86,104 times. As expected, most of the stems are related to the financial domain.

Additionally we performed topic modeling in order to extract topics from news titles. For this purpose we applied a graphical probabilistic model, namely, Latent Dirichlet Allocation (LDA) to the titles of the pages uploaded over three weeks – starting May 4, 2015. We labeled each title with the most relevant topic identified by the LDA. For example, it is interesting to point out that by considering three topics, news titles are subdivided into three sets including 32%, 38% and 30% of the pages. The temporal patterns of the number of page uploads per hour subdivided according to these topics is shown in Figure 7. As can be seen – even though the page published during weekend days mainly refer to one topic – in general the website is characterized by a mix of pages covering different topics that does not depend on the time of the day and the day of the week.

5 Results

In this section, we present the results of the analysis of the time series referring to 250 weekdays, namely, a "training" time series consisting of 5,400 observations

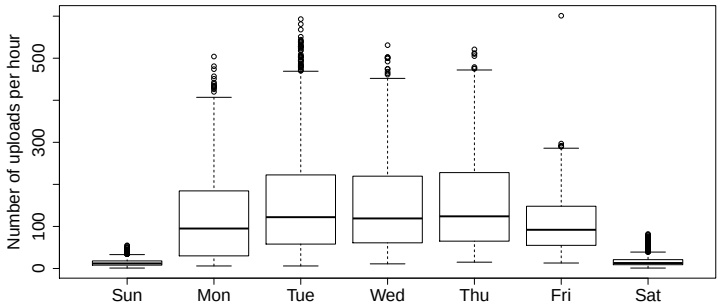


Fig. 5. Boxplot of the number of page uploads per hour for each day of the week.

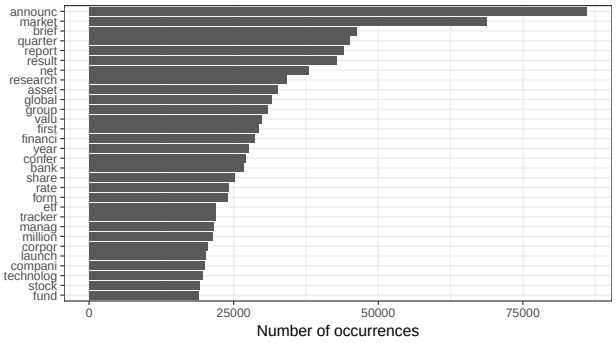


Fig. 6. Popularity of the top 30 stems appearing in the news titles.

– referring to the uploads per hour over 225 consecutive days – and a "validation" time series consisting of 600 observations – referring to the uploads of the remaining 25 days.

As previously discussed, the upload patterns exhibit large variability across days, hours and even weeks (see Fig. 4). To further investigate the properties of these temporal patterns, we analyze in the lag plots of Figure 8 the overall behavior of time series to assess whether there is any autocorrelation structure. The observations tend to group around the diagonal for small time lags, thus exhibiting a positive autocorrelation. On the contrary, for larger time lags the observations are more scattered.

The patterns of the autocorrelation function with time lags from one to 120 hours (i.e., five days) – summarized in Figure 9(a) – clearly suggest a periodic behavior of the uploads. All values fall outside the 95% confidence bands highlighted in the diagram by dashed lines. Similarly, the power spectrum of the autocorrelation function shown in Figure 9(b) confirms that the time series ex-

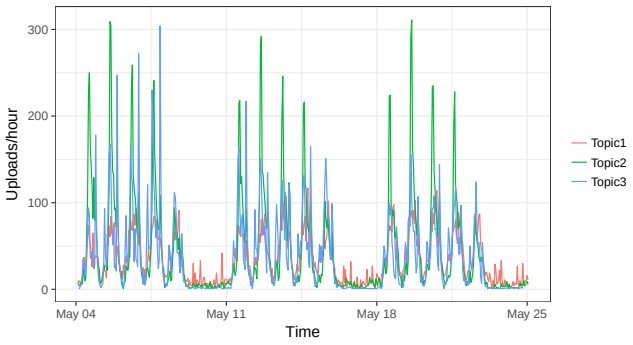


Fig. 7. Temporal patterns of the number of uploads per hour subdivided according to the topics identified by LDA.

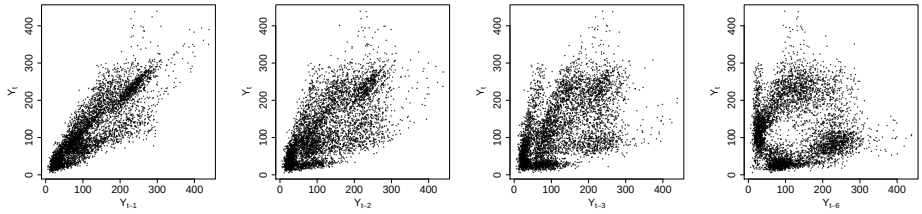


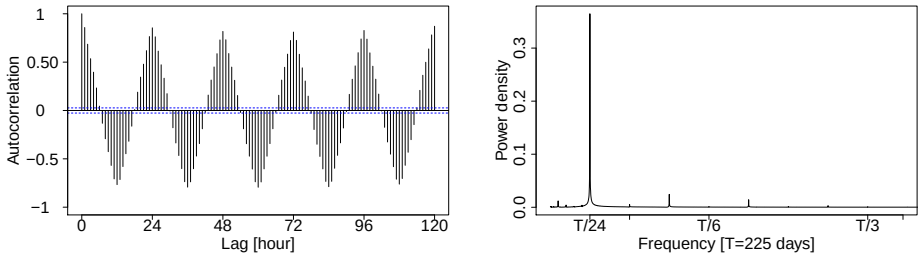
Fig. 8. Lag plots of the time series at varying time lags, i.e., 1, 2, 3 and 6.

hibits a certain periodicity. More precisely, the peak at frequency $24/T$ indicates the presence of a daily periodicity. This finding will be used for the identification of the deterministic components of the time series.

As outlined in Sect. 3.3, the decomposition of the time series into deterministic, i.e., trend and seasonal, and stochastic, i.e., irregular, parts relies on an additive approach. In detail, we applied the Loess method to estimate the trend and seasonal components, while the irregular component corresponds to the remainder of the time series.

An example of the decomposition of the time series representing the upload patterns over five days is shown in Figure 10. Note that we applied the decomposition to the "adjusted" time series where the spikes previously identified have been replaced with observations obtained by interpolation over neighbor observations.

Because of the characteristics of the deterministic components, we selected their models in the family of trigonometric polynomials and we applied least square techniques to fit the models to the data. In details, the model identified for the trend component is a trigonometric polynomial of degree four with eight parameters including the intercept. The seasonal component is modeled



(a) autocorrelation function (b) power spectrum

Fig. 9. Autocorrelation function of the time series computed for time lags ranging from one to 120 hours (a) and corresponding power spectrum (b).

by a trigonometric polynomial of degree one with two parameters. On the contrary, the best fit of the irregular component is represented by an ARMA model $(1, 2) \times (1, 0)_{24}$.

Since the final model has to include the contribution of the spiky patterns previously identified, we applied a logistic regression to predict whether the observation y_t corresponds to a spike. In particular, the model takes into account the time t together with y_{t-1} and the difference between y_{t-1} and y_{t-2}.

An example of the overall model of the time series over ten days is shown in Figure 11. We notice that the model accurately fits the data even though – because of their peculiarities – some of the spikes have not been precisely captured. The root mean squared error computed over the entire "training" time series is equal to 40.1.

The final model is used for making forecasts of the future dynamics of the website. For this purpose, we used the "validation" time series consisting of the observations over 25 days. More precisely, we extrapolate the trigonometric polynomials that best fit the trend and seasonal components of the time series, while the predictions of the irregular component rely on the Box-Jenkins approach. The logistic regression model previously identified has been used to predict the spiky patterns. Figure 12 shows an example of the predictions over ten days with a time horizon h equal to one hour.

We outline that our methodological approach has several advantages with respect to other approaches (e.g., Holt-Winters, Seasonal ARIMA, Recurrent Neural Networks). In details, the periodicity estimation and the identification of the underlying patterns (e.g., spiky patterns) are very useful for understanding and exploring the properties of the time series and improving forecasts. For example, sophisticated methods, such as RNNs, do not provide any insights in the temporal patterns. In addition, their application is usually computationally intensive and requires large training datasets. Similarly, SARIMA models do not break down the contributions of the underlying patterns of the time series.

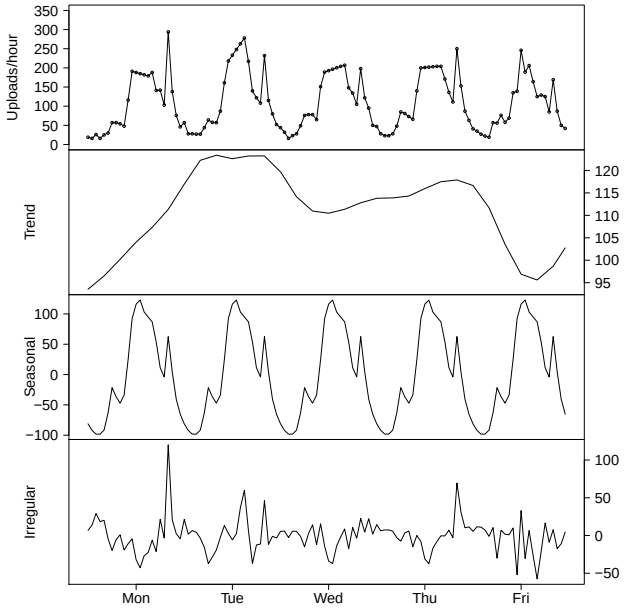


Fig. 10. Temporal patterns of the time series over five days and decomposition into trend, seasonal and irregular components. The labels on the x axis are centered at 12 noon.

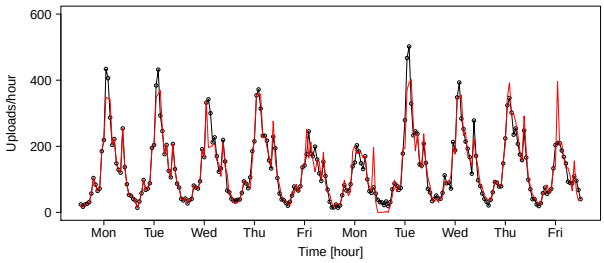


Fig. 11. Overall model (red curve) of the upload patterns (represented by circles) over ten days. The labels on the x axis are centered at 12 noon.

On the contrary, even though Holt-Winters models take into account trend and seasonal components, they describe these components in terms of a sequence of coefficients and smoothing equations.

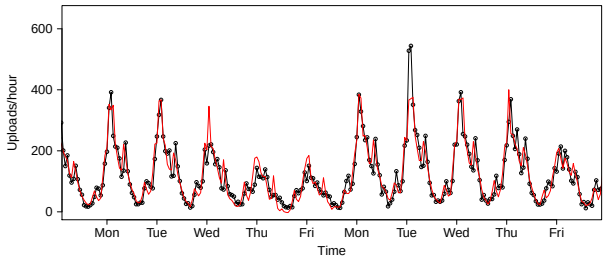


Fig. 12. Predictions of the upload patterns over ten days with a time horizon of one hour. The red curve refers to the predictions, the circles to data. The labels on the x axis are centered at 12 noon.

6 Conclusions

The web is a large information ecosystem where content changes over time as a consequence of combined effects involving the management policies of the websites, the behavior of the users and external events. These dynamics challenge technologies aimed at content management and retrieval.

Time series analysis is a valid and well defined method to model and predict the behavior of temporal data. Nevertheless, its application requires particular care. In this paper we proposed a general methodological framework for time series analysis and forecasting that specifically addresses the estimation of the its periodicity, the detection and modeling of the spikes and the decomposition of the time series into its underlying patterns.

The methodology has been applied to investigate and predict the dynamics of the Reuters news agency website. The page upload process of this website is characterized by a diurnal pattern and a much larger number of uploads during weekdays with respect to weekend days. Moreover, this process exhibits several sudden spikes. From the analysis of the content of the news titles we observed that the pages published on the website cover different topics that do not depend on the time of the day and on day of the week.

The individual components of the time series have been independently modeled and these models have then been used for making forecasts.

We outline that the proposed methodological approach – although tested in this paper in the framework of web dynamics – is general enough and can be applied to model and predict the behavior of any phenomenon represented as a time series.

As a future work, we plan to investigate the dynamics of the access patterns of web robots and identify differences and similarities between the patterns of good and malicious robots. Another possible research direction is in the area of topic modeling to classify pages and assess the relationships between web dynamics and the topics being addressed.

References

1. Adar, E., Teevan, J., Dumais, S.T., Elsas, J.: The web changes everything: Understanding the dynamics of web content. In: Proc. of the 2nd ACM Int. Conf. on Web Search and Data Mining - WSDM'09. pp. 282–291. ACM (2009)
2. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. Wiley, 5th edn. (2015)
3. Brewington, B., Cybenko, G.: How dynamic is the Web? Computer Networks **33**(1-6), 257–276 (2000)
4. Calzarossa, M., Massari, L., Tessera, D.: Workload characterization: A survey revisited. ACM Computing Surveys **48**(3), 48:1–48:43 (2016)
5. Calzarossa, M., Tessera, D.: Characterization of the evolution a news Web site. Journal of Systems and Software **81**(12), 2236–2344 (2008)
6. Calzarossa, M., Tessera, D.: Time series analysis of the dynamics of news websites. In: Proc. of the 13th Int. Conf. on Parallel and Distributed Computing, Applications and Technologies - PDCAT'12. pp. 529–533. IEEE Computer Society Press (2012)
7. Calzarossa, M., Tessera, D.: Multivariate analysis of Web content changes. In: Proc. of the 11th ACS/IEEE Int. Conf. on Computer Systems and Applications (AICCSA 2014). pp. 699–706. IEEE Computer Society Press (2014)
8. Calzarossa, M., Tessera, D.: Modeling and Predicting Temporal Patterns of Web Content Changes. Journal of Network and Computer Applications **56**, 115–123 (2015)
9. Calzarossa, M., Tessera, D.: Analysis and Forecasting of Web Content Dynamics. In: Proc. of the 32nd Int. Conf. on Advanced Information Networking and Applications Workshops. pp. 12–17. IEEE Computer Society (2018)
10. Cho, J., Garcia-Molina, H.: Estimating frequency of change. ACM Transactions on Internet Technology **3**(3), 256–290 (2003)
11. Cleveland, R., Cleveland, W., McRae, J., Terpenning, I.: STL: A Seasonal-Trend Decomposition procedure based on Loess (with discussion). Journal of Official Statistics **6**, 3–73 (1990)
12. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research **12**, 2493–2537 (2011)
13. Fetterly, D., Manasse, M., Najork, M., Wiener, J.: A large-scale study of the evolution of Web pages. Software: Practice & Experience **34**(2), 213–237 (2004)
14. Hamilton, J.D.: Time series analysis. Princeton University Press (1994)
15. Ke, Y., Deng, L., Ng, W., Lee, D.L.: Web dynamics and their ramifications for the development of web search engines. Computer Networks **50**(10), 1430–1447 (2006)
16. Li, X., Cline, D.B.H., Loguinov, D.: Temporal update dynamics under blind sampling. IEEE/ACM Transactions on Networking **25**(1), 363–376 (2017)
17. Lim, L., Wang, M., Padmanabhan, S., Vitter, J., Agarwal, R.: Characterizing Web Document Change. In: Wang, X., Yu, G., Lu, H. (eds.) Advances in Web-Age Information Management, Lecture Notes in Computer Science, vol. 2118, pp. 133–144. Springer (2001)
18. Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting - Methods and Applications. Wiley, 3rd edn. (1998)
19. Oita, M., Senellart, P.: Deriving Dynamics of Web Pages: A Survey. In: Proc. of the 1st Int. Temporal Workshop on Web Archiving - in conjunction with WWW 2011. pp. 25–32 (2011)

20. Radinsky, K., Bennett, P.: Predicting Content Change on the Web. In: Proc. of the 6th ACM Int. Conf. on Web Search and Data Mining - WSDM'13. pp. 415–424. ACM (2013)
21. Radinsky, K., Svore, K., Dumais, S., Shokouhi, M., Teevan, J., Bocharov, A., Horvitz, E.: Behavioral Dynamics on the Web: Learning, Modeling, and Prediction. ACM Transactions on Information Systems **31**(3), 16:1–16:37 (2013)
22. Shi, W., Collins, E., Karamcheti, V.: Modeling object characteristics of dynamic Web content. Journal of Parallel and Distributed Computing **63**(10), 963 – 980 (2003)
23. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proc. of the 4th ACM Int. Conf. on Web Search and Data Mining - WSDM'11. pp. 177–186. ACM (2011)