Journal of Network and Computer Applications 56 (2015) 115–123

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Modeling and predicting temporal patterns of web content changes

Maria Carla Calzarossa[a,*], Daniele Tessera[b]

[a] Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, via Ferrata 5 – I-27100 Pavia, Italy
[b] Dipartimento di Matematica e Fisica, Università Cattolica del Sacro Cuore, via Musei 41 – I-25121 Brescia, Italy

ARTICLE INFO

Article history:
Received 24 August 2014
Received in revised form
20 January 2015
Accepted 18 June 2015
Available online 4 July 2015

Keywords:
Web dynamics
Time series analysis
Forecasting
Search engines
Workload characterization
ARMA models

ABSTRACT

The technologies aimed at Web content discovery, retrieval and management face the compelling need of coping with its highly dynamic nature coupled with complex user interactions. This paper analyzes the temporal patterns of the content changes of three major news websites with the objective of modeling and predicting their dynamics. It has been observed that changes are characterized by a time dependent behavior with large fluctuations and significant differences across hours and days. To explain this behavior, we represent the change patterns as time series. The trend and seasonal components of the observed time series capture the weekly and daily periodicity, whereas the irregular components take into account the remaining fluctuations. Models based on trigonometric polynomials and ARMA components accurately reproduce the dynamics of the empirical change patterns and provide extrapolations into the future to be used for forecasting.

1. Introduction

The huge amount of content available on the Web coupled with its large variety and the complex types of user interactions opens up new significant research challenges. To fully exploit the potential of the Web as a global information repository, it is important to devise powerful technologies able to efficiently retrieve Web content and adapt to its highly dynamic nature and usage patterns (Menascé and Almeida, 2001). Among these technologies, search engines play a key role as they represent for the vast majority of the users their primary entry point for exploring the Web.

The content of a website changes whenever new pages are uploaded, existing pages are removed or their content updated. The dynamics of these changes varies from site to site and from time to time. To be of any value to the users, search engines have to cope with this variability and index in a timely manner new Web content as well as changes of existing one. To optimize the crawling activities, that is, minimize the costs for download, storage and management of the content and maximize its freshness, it is therefore important for these tools to be able to predict how often the content of a site changes.

In this paper we present an extensive investigation aimed at modeling and predicting the dynamics of Web content changes. In particular, our study focuses on the change patterns of three major news websites. The choice of these sites is mainly motivated by the peculiarities of their content that is highly time-sensitive and whose relevance tends to rapidly decay with time. Indeed, news websites are expected to deliver in a timely manner the latest stories as well as their latest developments.

The modeling approach adopted in our study relies on time series because of their ability to capture temporal patterns. As a result of their analysis, we obtain models that accurately explain the change dynamics of the three websites. These models are the basis for forecasting, that is, predicting the future dynamics of the websites based on current and historical behavior. Note that, even though different intent may lead to different patterns in the Web content changes, our approach can be applied to any type of website.

The paper is organized as follows. Section 2 presents an overview of the literature on the issues related to Web dynamics. The approach proposed for modeling and forecasting the change dynamics is described in Section 3. The data collection process implemented for crawling the websites under investigation is described in Section 4. The experimental results are presented in Section 5. Finally, Section 6 draws some conclusions by summarizing the major findings.

2. Related work

The dynamics of Web content and its impact in the design of the technologies for content discovery, retrieval and management have been addressed in the literature under different perspectives (Baeza, 2004; Brewington and Cybenko, 2000; Cho and Garcia-Molina, 2003; Fetterly et al., 2004; Ke et al., 2006; Kwon et al.,

[*] Corresponding author.
E-mail addresses: mcc@unipv.it (M.C. Calzarossa), daniele.tessera@unicatt.it (D. Tessera).

http://dx.doi.org/10.1016/j.jnca.2015.06.008

2006; Lim et al., 2001; Padmanabhan and Qiu, 2000). More recently, several studies examined Web dynamics in the framework of the increased complexity of websites. In particular, the analysis of modern Web traffic presented in Ihm and Pai (2011) shows that pages have become quite complex and are often characterized by an increased size and a large number of objects. Moreover, Butkiewicz et al. (2011) observe that news websites are in general more complex than other websites as they load significantly more content usually stored on local as well as remote servers. Furthermore, the work on dynamic Web content generation presented in Ravi et al. (2009) emphasizes that dynamic pages change more frequently than static pages, thus making caching very tricky.

A fine grain characterization of the evolution of Web content is introduced in Adar et al.(2009). 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. Stable and dynamic content within each page are identified. The evolution of a news website is addressed in Calzarossa and Tessera (2008) in terms of rates of page creations and updates. The study shows that these rates are characterized by some well defined patterns. In addition, since most updates involve a very small fraction of the page content and very few are more extensive and spread over the entire page, the corresponding models can be adjusted accordingly. Various methods for predicting content changes on the Web are introduced in Radinsky and Bennett (2013) by considering an expert predictive framework based on features, such as relatedness to other pages and similarity in the types of changes.

The concept of information longevity, that is, the lifetime of the content that appears and disappears from the Web, introduced by Olston and Pandey (2008), is considered as a key aspect for the description of Web evolution and for the development of effective crawling policies. Starting from the observation that there is no correlation between information longevity and change frequency, the paper presents 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 characterization of dynamic content of news and e-commerce sites presented in Shi et al. (2003) highlights that a large fraction of objects does not change within a time scale of a week, whereas objects that change within the timescale of a day are characterized by short freshness times. These findings have important implications on content reusability and caching policies. The analysis of the novelty of the content of a news Web site is addressed in Calzarossa and Tessera (2010) under two different perspectives, namely, an horizontal perspective that focuses on the content of the individual articles and considers how fast and to what extent each article is modified, and a vertical perspective that considers the entire collection of articles posted on the site and takes into account the evolution of the Web site.

Time series analysis has been applied by Yang and Leskovec (2011) to study the temporal patterns associated with online textual content and derive the shapes characterizing the different types of media. For example, press agencies exhibit a rapid rise followed by a relatively slow decay, whereas bloggers influence the news longevity. Another interesting application of time series analysis is presented in Zhang et al. (2009), where the predictive models developed for Web searching are based on the dynamic patterns of the interactions between users and search engines.

A temporal modeling framework that captures the dynamic nature of Web behaviors is introduced by Radinsky et al. (2012) and further explored in Radinsky et al. (2013). The proposed models include the typical characteristics observed in query and URL click behavior of Web searchers, that is, trend, periodicity and surprise disruptions.

Despite other studies, we represent the temporal patterns of Web content changes as time series whose analysis provides models able to explain their dynamics. In addition, this approach allows us to accurately forecast the future dynamics of the websites based on their current and historical behavior. In this respect and some others, this study enhances and complements early results presented in Calzarossa and Tessera (2012) on different data samples.

3. Modeling approach of the temporal patterns

The modeling approach adopted to study the dynamics of the news websites considered in this study relies on the application of time series analysis because of its ability to capture temporal patterns. Let us recall that a time series Box et al. (2008) is a collection of N ordered observations taken sequentially in time at equally spaced time intervals, that is, $\{Y_t\} = \{y_{t_1}, y_{t_2}, \ldots, y_{t_N}\}$ with $t_1 \le t_2 \le \ldots \le t_N$.

Before modeling the time series, it is important to have a preliminary look at the data to understand their overall properties. In particular, it is worth investigating the presence of outliers, that is, observations that appear inconsistent with the neighboring observations of the time series. Measures, such as median absolute deviation, work well for this purpose. To avoid anomalous perturbations in the overall models, outliers are usually removed from the time series. In addition, depending on their nature (e.g., additive, innovation, level shift) it might be necessary to build a specific model for the identified outliers.

Furthermore, the autocorrelation function is another very powerful tool for investigating the properties of the data and highlighting statistical dependencies. The autocorrelation function r_h at lag h is defined as

$$r_h = \frac{\sum_{i=1}^{N-h}(y_{t_i} - \overline{y})(y_{t_{i+h}} - \overline{y})}{\sum_{i=1}^{N}(y_{t_i} - \overline{y})^2}$$

where $\overline{y}$ denotes the mean value of the time series. The function computed at varying time lags allows us to check the randomness of the data. Note that autocorrelations close to zero for any time lag denote a random behavior, whereas one or more values significantly different from zero denote a nonrandom behavior, that is, the presence of some sort of statistical dependence. In addition, to assess the stationarity of the time series, i.e., whether its mean and variance remain constant over time, it is necessary to investigate the time invariance properties of the autocorrelation function. As we will explain in more detail, some methods for time series analysis, e.g., the Box–Jenkins approach, are applicable to stationary time series only.

The results of these analyses together with the plots of the time series provide an overview of the phenomenon and highlight its temporal variations. In particular, the plots (see, e.g., Fig. 1) are important to discover trend cycles indicating a long-term movement in the mean level, that is, a tendency to grow or decrease rather steadily over quite long periods of time, and seasonal factors, that is, patterns that repeat in time with some periodicity, e.g., hourly, daily, weekly, monthly.

The identification of the periodicity of the time series relies on the spectral analysis of the autocorrelation function and in particular on the discrete Fourier coefficients f_k associated with the k/T frequencies, namely

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

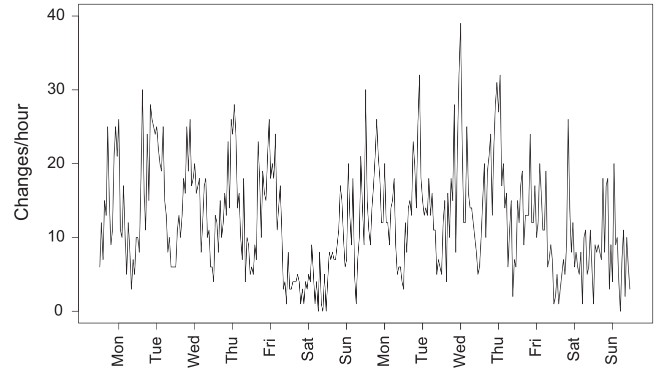

Fig. 1. Dynamics of the content changes of the Reuters website over a two weeks interval. The labels on the *x* axis are centered at 12 noon.

Table 1
Main characteristics of the crawling activities performed on the three news websites.

Website	Unique pages	Updates	Crawling intervals [days]
CNN	8,302	9,502	104
MSNBC	5,436	6,809	84
Reuters	15,157	3,734	63

Note that *i* denotes the complex imaginary unit and *T* refers to the observation interval of the time series.

Diagrams, such as periodograms, that plot $|f_k|^2$ as a function of the frequencies, highlight the peaks, if any, in the spectrum of the autocorrelation function. These peaks denote the dominant frequencies corresponding to the periods of the temporal patterns in the observations.

In case of a non-stationary time series, it is necessary to make it stationary by isolating and removing the deterministic components previously identified, namely, the trend and seasonal components. After adjusting the original time 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 effects and with stable fluctuations.

The decomposition of the time series relies on an additive approach, that is, $Y_t = T_t + S_t + X_t$, where T_t, S_t and X_t denote the trend, seasonal and irregular components, respectively. Various smoothing techniques, such as the locally weighted polynomial regression or Loess method (Cleveland et al., 1990) can be applied for the decomposition of the time series.

The individual components are then considered separately to derive mathematical models able to explain the observed variability. The estimation of the deterministic components relies on standard numerical fitting techniques. Moreover, goodness of fit tests and the nested model selection procedures are applied to choose the models that best fit these components (Trivedi, 2002).

On the contrary, before estimating the irregular component it is necessary to explore its characteristics with respect to statistical dependence and random behavior. In general, whenever the corresponding autocorrelation function is characterized by large values for small lags only, the estimation of the irregular component requires Auto Regressive Moving Average (ARMA) models.

We recall that an ARMA (*p*,*q*) model of a stationary time series X_t at time t_i can be expressed as

$$x_{t_i} = \sum_{k=1}^{p} \phi_k x_{t_{i-k}} + \omega_{t_i} + \sum_{k=1}^{q} \theta_k \omega_{t_{i-k}}$$

where *p* and *q* are the orders of the autoregressive and moving average terms, $\{\omega_t\}$ refers to the white noise, ϕ_k and ϕ_k denote the autoregressive and moving average coefficients, respectively. The coefficients of the ARMA model are derived by applying numerical fitting techniques, whereas the identification of the optimal orders of the model is based on the principle of parsimony. This means that models with fewer parameters are preferable to models with many parameters as long as their descriptive power does not significantly change. More specifically, this identification is based on various model selection criteria (e.g., AIC, BIC, combined with model checking and diagnostic tests (e.g., Ljung–Box tests)) aimed at assessing the statistical independence of the residuals obtained as a result of the numerical fitting (Kyriazidou, 1998).

Once the models have been identified, they can be used for the forecasting, that is, to predict the dynamics of the websites. More specifically, $\hat{Y}_{t+h}$, i.e., the predicted value of the time series at time $t+h$, is obtained as the sum of the values predicted by the models of the trend, seasonal and irregular components, that is, $\hat{Y}_{t+h} = \hat{T}_{t+h} + \hat{S}_{t+h} + \hat{X}_{t+h}$. For the trend and seasonal components these values are extrapolated by the corresponding models computed at time $t+h$. On the contrary, a moving horizon prediction technique based on the Box–Jenkins approach is applied to forecast the values of the irregular component. Hence, for obtaining $\hat{X}_{t+h}$ this approach requires *h* one-step ahead iterations. The forecast accuracy is finally assessed by means of various descriptive measures, such as absolute and relative errors of the predictions with respect to the empirical values.

4. Data collection

Our investigation focuses on three major news websites, namely, the websites owned by the CNN[1] and MSNBC[2] cable news channels and by the Reuters news agency[3]. As the content of news websites tends to change frequently and rapidly, to accurately capture their dynamics, we collect multiple snapshots of each site at a rather fine grain. More specifically, we crawl the sites every 15 min by initially downloading their front pages. We then extract the hyperlinks contained in these pages and we iteratively download the pages addressed by these hyperlinks. Each snapshot of the websites therefore includes pages posted on the sites since the previous download as well as pages already downloaded and still "reachable" from the front pages.

For the crawling of the websites, we develop a shell script that relies on the open source `wget` software package (GNU wget, 2014) to download the pages using the HTTP protocol. On average, at each snapshot we download some 95, 250 and 160 pages for the CNN, MSNBC and Reuters websites, respectively. It is interesting to point out that these numbers vary across sites, nevertheless, for each site they do not vary significantly across snapshots.

The Web pages downloaded from the three websites consist of a template and a large variety of dynamically generated objects, such as images, banners, videos, advertisements and recommendations, often customized according to users preferences, and whose number and temporal patterns vary from page to page. For these reasons, in our study we do not consider the entire Web pages, instead we focus on the descriptive text of the news stories. Hence, after each download, we parse the Web pages to extract their textual content. Note that in what follows with the term Web page we refer to its textual component only.

As already pointed out, the content of a website changes, whenever new pages are posted or existing pages are updated or

[1] http://www.cnn.com
[2] http://www.msnbc.com
[3] http://www.reuters.com

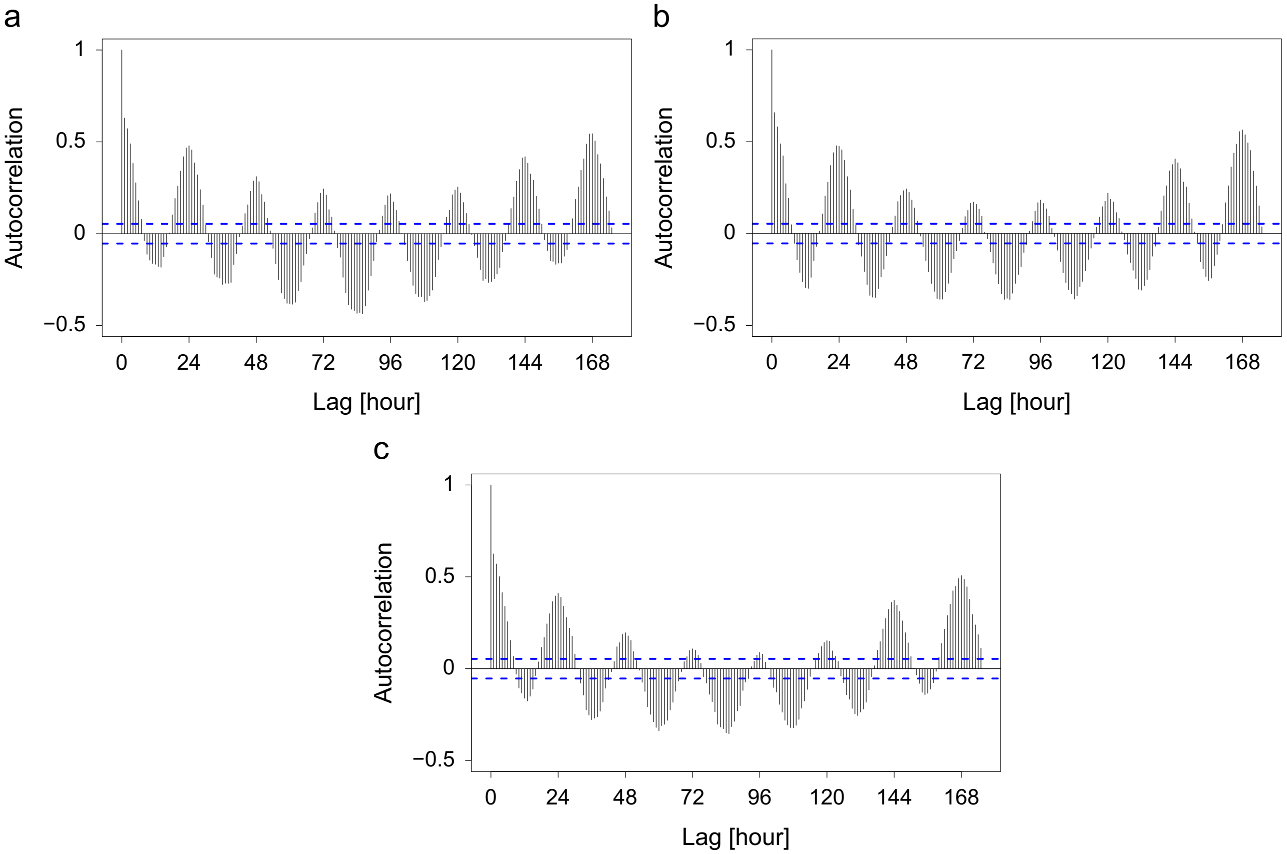

Fig. 2. Autocorrelation functions computed at varying time lags for the content changes of the (a) CNN, (b) MSNBC, and (c) Reuters websites.

deleted. Therefore, as part of some preliminary data processing, we analyze the pages downloaded at each snapshot for detecting possible updates. More specifically, we implement a detection mechanism that relies on the computation of the cosine coefficient of similarity between consecutive instances of a given page (Salton and McGill, 1983). We recall that an update is denoted by a value of the coefficient different than one. The use of the cosine coefficient is motivated by its ability to capture very precisely all updates and avoid marking as updated pages that contain exactly the same set of words but differ by a blank space only.

5. Experimental results

5.1. Overall characteristics of the datasets

The main characteristics of the datasets obtained from the crawling activities are summarized in Table 1. As a result of several weeks of crawling, we download in total some 29,000 unique Web pages, that is, pages identified by a unique URL. More than half of these pages are downloaded from the Reuters website and less than 20% from the MSNBC website. In addition, by computing the cosine coefficients of similarity we identify some 20,000 updates. By comparing the number of unique pages and the number of updates, we can infer differences in the policies adopted by the website administrators with respect to content management. In particular, we notice the tendency of the Reuters website to modify its content by uploading new pages instead of updating the existing ones. In this case, updates involve on average only about one page out of four. On the contrary, pages posted by CNN or MSNBC are updated more frequently, that is, on average more than one update per page.

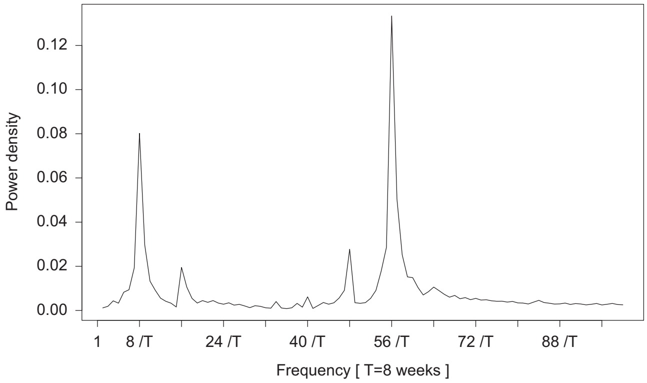

Fig. 3. Power spectrum of the autocorrelation function obtained for the Reuters time series.

To study the dynamics of the three websites we focus on their content changes. More specifically, we analyze changes in terms of uploads of new pages and updates of existing ones. We do not consider page removal because pages are often moved to some long-term archives and seldom disappear from the sites, even though the corresponding hyperlinks are no longer included in the corresponding front pages.

On average, during our crawling intervals, we detect 7.5, 6.1 and 12.5 changes per hour of the content of the CNN, MSNBC and Reuters websites, respectively. Nevertheless, these changes are characterized by a large variability. Moreover, as shown in Fig. 1, they are not evenly distributed across weeks, days and hours. The diagram, that refers to the changes detected over a two weeks interval on the Reuters website, clearly shows daily and weekly periodic patterns and a time dependent behavior with large

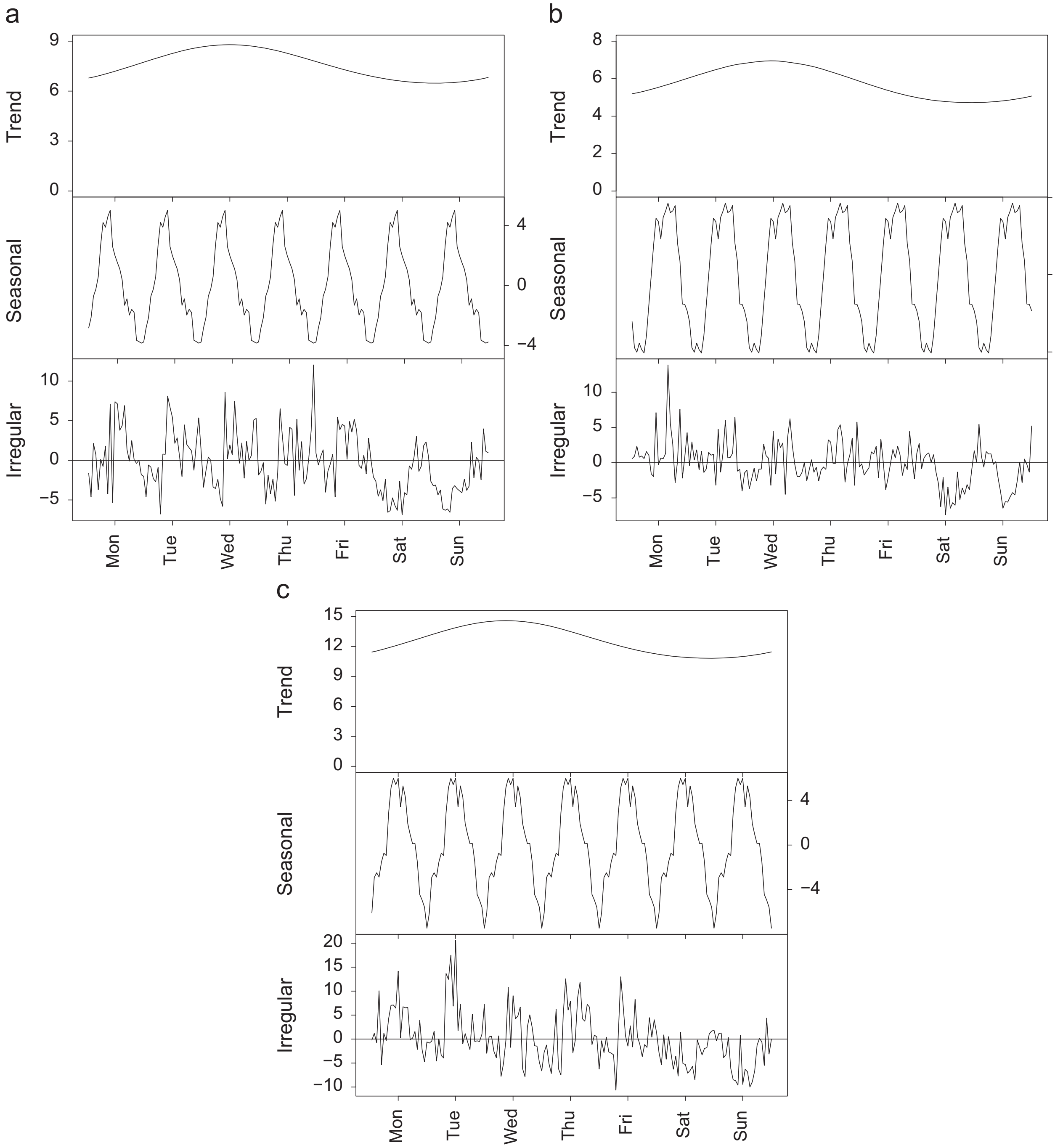

Fig. 4. Decomposition of the time series representing the content changes of the (a) CNN, (b) MSNBC, and (c) Reuters websites into their trend, seasonal and irregular components.

Table 2
Parameters of the trigonometric polynomials of degree one that best fit the trend components of the three time series.

	a_0	a_1	b_1
CNN	7.56	0.91	−0.71
MSNBC	5.76	0.94	−0.60
Reuters	12.54	1.62	−0.97

fluctuations from hour to hour, especially between day and night hours, and from day to day, especially between weekdays and weekend days. Similar patterns are detected for the changes of the CNN and MSNBC websites, although for the MSNBC we notice a substantial variability on weekends. Let us remark that the times considered in our study refer to the time zone of the individual websites.

In addition, before modeling the dynamics of the changes, we analyze the three datasets with the objective of identifying and removing outliers, that is, unusual observations caused by some anomalous conditions or external events, including, among the others, network and crawling failures. In particular, we do not discover any outlier in the Reuters dataset, whereas we identify four outliers in the other datasets. All these outliers correspond to observations whose number of changes is about an order of magnitude greater than the corresponding mean, that is, exceeding 60 and 70 changes per hour. Moreover, as these observations are a sort of ”singleton”, to avoid any perturbation in the modeling phase, we replace each of them with the values obtained via numerical interpolation of the neighboring observations.

5.2. Models of the change patterns

To model the dynamics of the three websites, we partition each dataset in two sections. The training section, namely, the data referring to the first eight weeks of each crawling interval, is used to build the model. The validation section, namely, the data referring to the remaining weeks, is used to evaluate the forecasting accuracy of the model.

As discussed in Section 3, to investigate the temporal properties of Web content changes and in particular their time dependent behavior, we first analyze the autocorrelation functions computed with time lags varying from one hour up to one week, that is, 168 h, (see Fig. 2). Despite some differences in magnitude, the repeated patterns shown in the three diagrams clearly highlight a periodic behavior. In particular, the test of independence performed at 5% significance level fails, as most of the values fall outside the 95% confidence bands delimited by the horizontal dashed lines, thus confirming the non-randomness of these patterns. The periodic behavior is also confirmed by the power spectrum of the autocorrelation functions (see, e.g., Fig. 3) whose peaks correspond to frequencies equal to $8/T$ and $56/T$, T being the period, that is, eight weeks. These peaks denote weekly and daily patterns, respectively.

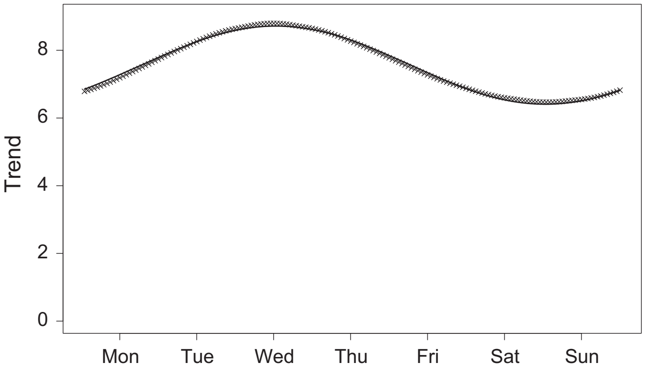

Fig. 5. Estimated trend component (dotted pattern) and corresponding model (solid pattern) obtained for the CNN time series.

To take into account these characteristics, we represent the changes of each site with a periodic time series whose observations Y_t are taken every hour over 24 h intervals. The peaks detected in the autocorrelation spectra allow us to identify the weekly and daily periodic components of the time series, that is, its trend and daily seasonality. More specifically, the application of the Loess method provides estimates of these periodic components. The irregular component is then obtained as the remainder of the time series.

From the decomposition of the three time series (see Fig. 4), we can easily identify their temporal patterns. In particular, the trend components are characterized by a periodic behavior that captures the weekly patterns of the changes, with most changes concentrated during weekdays and fewer changes occurring during weekend days. In general, even though their mean levels differ, we notice a rather similar behavior across sites. On the contrary, by looking at the seasonal components that capture the daily periodic patterns, we can observe some significant differences, that is, changes mostly concentrated in the morning hours for the CNN and Reuters websites, and spread over the entire afternoon for the MSNBC website. Finally, the decompositions include the irregular

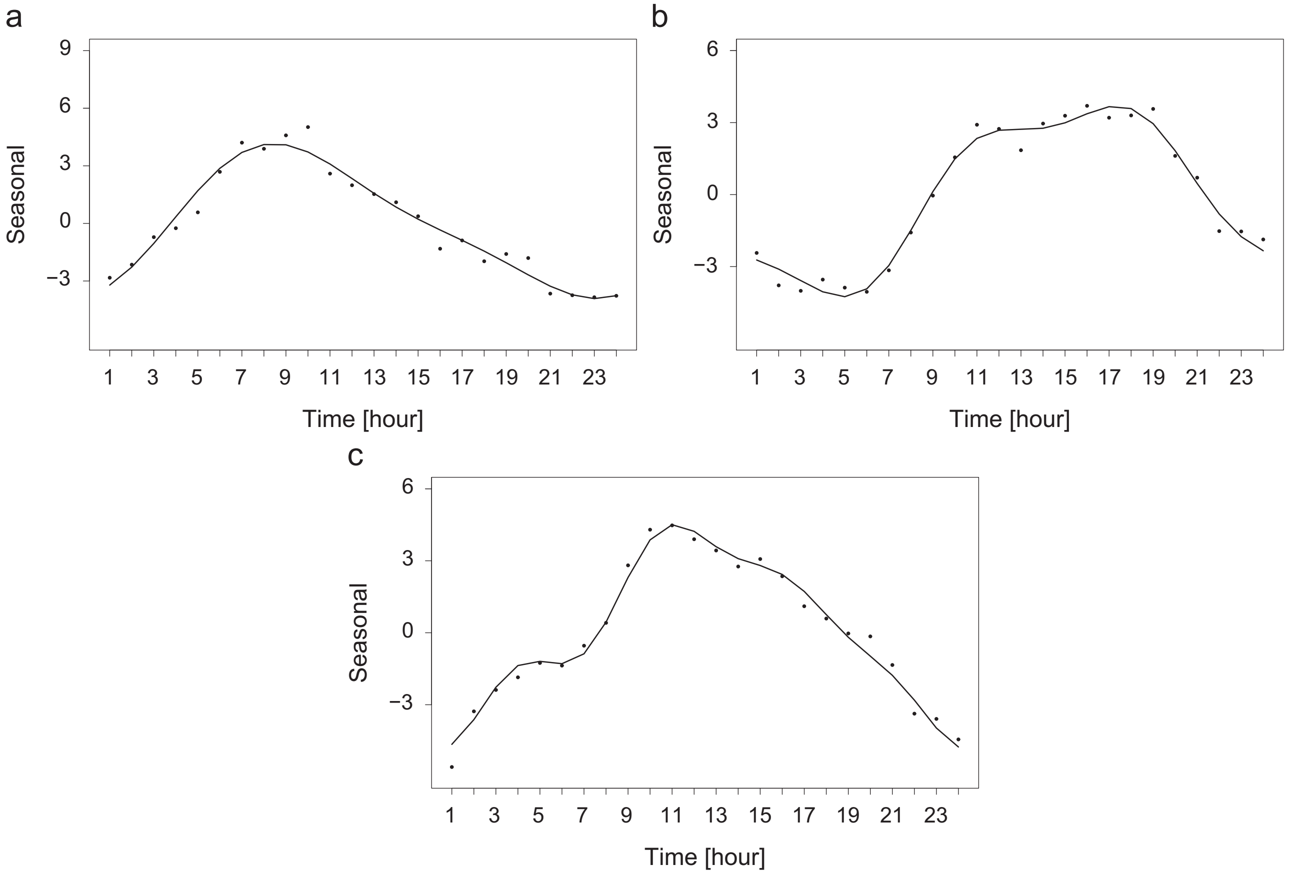

Fig. 6. Estimated seasonal components (dotted patterns) and corresponding models (solid patterns) of the (a) CNN, (b) MSNBC, and (c) Reuters time series.

Table 3
Parameters of the trigonometric polynomial of degree three that best fits the seasonal component of the MSNBC time series.

a_1	b_1	a_2	a_3	b_3
-2.39	-3.08	-0.34	0.57	0.36

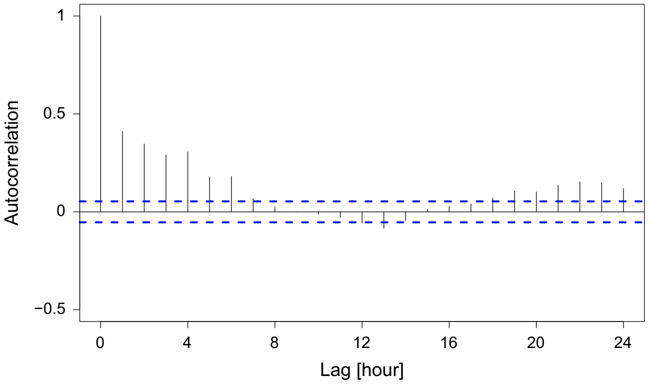

Fig. 7. Autocorrelation function of the irregular component of the MSNBC time series.

Table 4
Parameters of the ARMA models (2, 0, 2) obtained for the irregular components of the three time series.

	ϕ_1	ϕ_2	θ_1	θ_2	σ_ω
CNN	1.4456	-0.5372	-1.2112	0.4309	3.14
MSNBC	-0.0657	0.7322	0.3520	-0.4295	3.09
Reuters	0.1845	0.5879	0.1099	-0.4638	5.04

components that take into account the fluctuations not described by the trend and seasonal components.

Because of the periodic behavior of the deterministic, i.e., trend and seasonal, components of the time series, for their estimation we resort to trigonometric polynomials of the following form:

$$g(t) = a_0 + \sum_{k=1}^{n} \left(a_k \sin\left(2\pi k \frac{t}{T}\right) + b_k \cos\left(2\pi k \frac{t}{T}\right) \right)$$

where n denotes the degree of the polynomial and T the period. Numerical fitting techniques based on the Levenberg Marquardt method allow us to identify the parameters of the polynomials. In addition, to determine the degree of the polynomials that best fit each component and the number of parameters to be considered, we apply goodness of fit tests, i.e., F-test, and model selection criteria, i.e., AIC.

The models of the trend components of the three websites, that is, the periodic slow moving fluctuations around the mean level, are trigonometric polynomials of degree one, whose parameters, summarized in Table 2, reflect the differences previously highlighted. Figure 5 shows the model of the trend of the CNN time series.

A similar approach is applied to model the seasonal components of the three time series (see Fig. 6). Note that the polynomials do not include the parameter a_0 since the seasonal components obtained via the Loess method are characterized by a mean value equal to zero. In particular, the model of the CNN seasonal component is a polynomial of degree two described by three parameters only, namely, $a_1 = 2.87$, $b_1 = -2.39$ and $b_2 = -0.82$. Indeed, according to the AIC model selection criterion, the parameter a_2 is not considered as its value is very small and its 95% confidence interval include the zero.

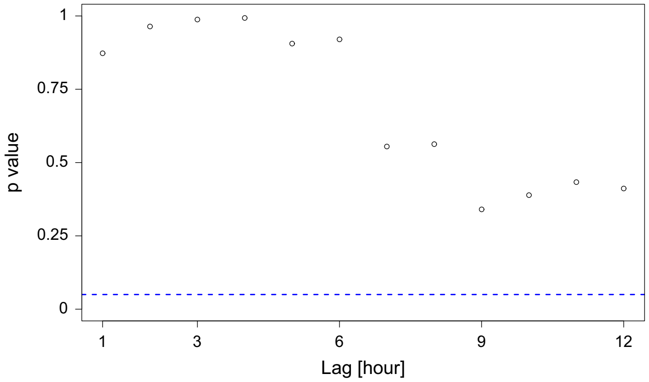

Fig. 8. Ljung-Box test applied to the residuals of the ARMA (2, 0, 2) model of the irregular component of the Reuters time series. The dashed line denotes the 0.05 significance bound.

In the case of the seasonal component of the MSNBC time series, the model is a trigonometric polynomial of degree three, described by five parameters (see Table 3). Similarly, a trigonometric polynomial of degree one described by two parameters, namely, $a_1 = 1.36$ and $b_1 = -5.36$, best fits the seasonal component of the Reuters time series.

The analysis of the irregular components of the time series highlights the presence of large short term autocorrelations (see Fig. 7). Moreover, statistical tests for normality, such as Anderson–Darling, reject the null hypothesis. Therefore, for the estimation of these components we consider ARMA models. In particular, ARMA models (2, 0, 2) best fit the irregular components. The parameters of the models and the standard deviation of the white noise σ_ω are presented in Table 4.

Several tests are applied to assess the goodness of fit of ARMA models. For example, the Ljung–Box test for randomness (see Fig. 8) applied to the residuals of the ARMA models shows the lack of any correlation. All p-values are greater than the 0.05 significance bound.

By superimposing the models derived for the trend, seasonal and irregular components of each time series, we obtain the final models that fully explain the dynamics of the content changes of the three websites. Snapshots over a two weeks interval of the empirical data and the final models are plotted in the diagrams of Fig. 9. As can be seen, the models accurately capture the empirical data, even though they are somehow characterized by less variability. While the mean values of the empirical data and of the models do not differ, the standard deviation obtained for the models is usually smaller. For example, in the case of the MSNBC time series, its value is equal to 3.51, compared to 4.73 of the empirical data. Nevertheless, we can observe that the models are good estimates of the overall change patterns.

5.3. Forecasting

The validation of the forecasting approach described in Section 3 is based on the data not used for model identification, that is, the data referring to one, four and six weeks of the crawling intervals for the Reuters, MSNBC and CNN websites, respectively. As already pointed out, this approach allows us to predict future changes based on their current and historical data. More specifically, from the trigonometric polynomials that best fit the trend and seasonal components of the time series, we extrapolate the corresponding predictions, whereas the predictions of the irregular components are iteratively derived by means of the Box–Jenkins approach.

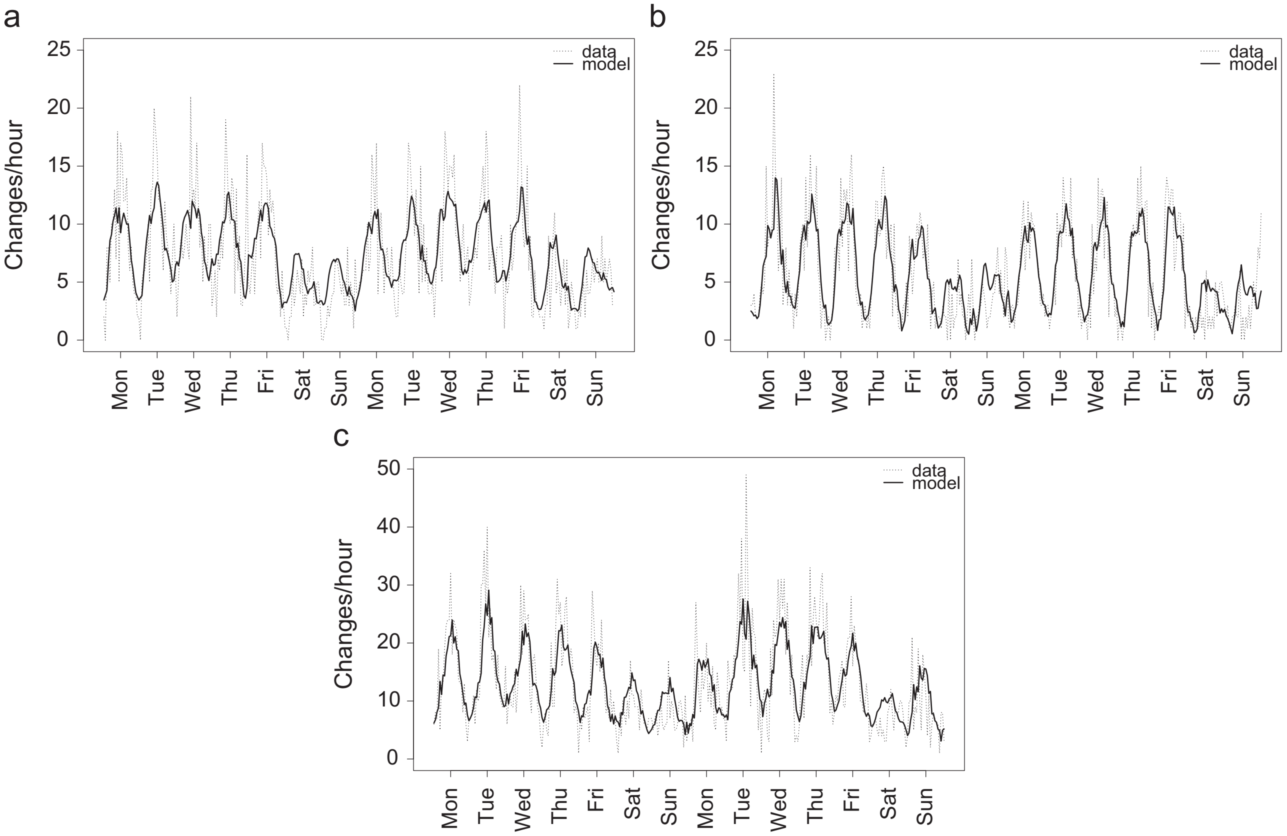

Fig. 9. Temporal patterns of the changes of the (a) CNN, (b) MSNBC, and (c) Reuters websites and corresponding models obtained by time series analysis.

Table 5
Mean absolute errors computed between the empirical data and the forecasts obtained at different prediction horizons.

	Prediction horizon [hours]					
	1	2	3	4	5	6
CNN	2.45	2 .43	2.53	2.76	3.10	3.45
MSNBC	2.48	2.56	2.89	3.26	3.68	4.11
Reuters	4.28	4.49	4.80	5.33	5.81	6.30

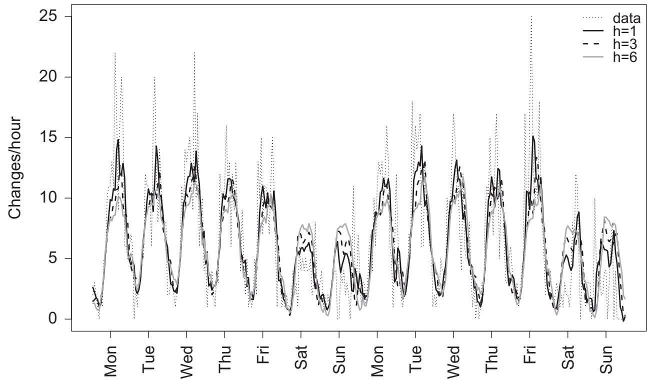

Fig. 10. Forecasts of the change patterns of the MSNBC website over two weeks validation data.

The forecast accuracy is assessed in terms of mean absolute errors computed between the empirical data and the predictions (see Table 5). As can be seen, the error increases as the prediction horizon increases from one hour up to six hours. In particular, the accuracy of the forecasts tends to significantly degrade at horizons larger than three hours. Moreover, it has been noticed that at larger prediction horizons, the forecasts cannot adequately explain the variability of the empirical data. For example, the standard deviation of the MSNBC validation data is equal to 5.30, compared to 3.80 and 3.05 of the values predicted at 1 and 6 h ahead, respectively.

The predictions obtained for two out of the four weeks validation data of the MSNBC website are displayed in Fig. 10. The diagram plots the empirical data and three forecasts, each corresponding to a different horizon, namely, h equal to 1, 3 and 6 h. We clearly notice that the curves become smoother as h increases. We recall that the mean absolute error is equal to 4.11 for $a=6$ horizon, that is, approximately 65% larger than the error corresponding to the forecast with a prediction horizon equal to 1 h.

Finally, it is worth noting that these results show the robustness of the forecast approach whose accuracy does not depend on the number of values being predicted.

6. Conclusions

Web content changes continuously over time. Content types and user interactions influence the change patterns. In this paper, we devise a methodological approach aimed at modeling and predicting the dynamics of the content changes of three major news websites. Let us remark that even though change behavior and page posting may depend on the policies adopted by individual websites, this approach is general enough and applies to any type of website.

Our investigation highlights significant differences in the patterns of the content changes of the three websites considered in our study. In general, the patterns are characterized by large fluctuations coupled with some periodicity and a time dependent behavior. To capture these effects, we represent content changes as periodic time series whose analysis allows us to describe and model their dynamics. The daily and weekly patterns of the changes are captured by the seasonal and trend components of the time series. Numerical fitting techniques identify the parameters of the trigonometric polynomials that best fit these components. Moreover, ARMA models are used to fit the irregular components of the time series. The final models accurately explain the empirical patterns of the content changes of each website. These models represent the basis for forecasting, that is, for predicting future dynamics based on the current and historical behavior. Note that good predictions can have a significant impact on all decision models used by technologies aimed at discovery, retrieval and management of Web content. In our study, the forecasting approach applied to the three validation datasets is rather robust as it provides accurate predictions at horizons up to 3 h ahead without being influenced by the number of values used for the prediction.

References

Adar D, Teevan J, Dumais ST, Elsas JL. The web changes everything: understanding the dynamics of web content. In: Proceedings of the second ACM international conference on web search and data mining—WSDM'09, ACM; 2009. p. 282–91.

Baeza-Yates R, Castillo C, Saint-Jean F. Web dynamics, structure and page quality. In: Levene M, Poulovassilis A, editors. Web Dynamics: Adapting to Change in Content, Size, Topology and Use, Springer; 2004. p. 93–109.

Box GEP, Jenkins GM, Reinsel GC. Time Series Analysis: Forecasting and Control. 4th ed. Hoboken, NJ: Wiley; 2008.

Brewington BE, Cybenko G. How dynamic is the Web? Comput Netw 2000;33(1–6):257–76.

Butkiewicz M, Madhyastha HV, Sekar V. Understanding website complexity: measurements, metrics, and implications. In: Proceedings of the 11th ACM SIGCOMM conference on Internet measurement—IMC'11, ACM; 2011. p. 313–28.

Calzarossa M, Tessera D. Characterization of the evolution a news Web site. J Syst Softw 2008;81(12):2236–344.

Calzarossa M, Tessera D. An exploratory analysis of the novelty of a news Web site. In: Proceedings of the international symposium on performance evaluation of computer and telecommunication systems—SPECTS 2010, SCS Press; 2010. p. 399–404.

Calzarossa M, Tessera D. Time series analysis of the dynamics of news websites. In: Proceedings of the 13th international conference on parallel and distributed computing, applications and technologies—PDCAT'12, IEEE Computer Society Press; 2012. p. 529–33.

Cho J, Garcia-Molina H. Estimating frequency of change. ACM Trans Internet Technol 2003;3(3):256–90.

Cleveland RB, Cleveland WS, McRae JE, Terpenning. I. STL: a seasonal-trend decomposition procedure based on Loess (with discussion). J Off Stat 1990;6:3–73.

Fetterly D, Manasse M, Najork M, Wiener J. A large-scale study of the evolution of Web pages. Softw: Pract Exp 2004;34(2):213–37.

Free Software Foundation. GNU `wget Manual` (access July 2014). ⟨http://www.gnu.org/software/wget/manual/wget.pdf⟩.

Ihm S, Pai VS. Towards understanding modern Web traffic. In: Proceedings of the 11th ACM SIGCOMM conference on Internet measurement—IMC'11, ACM; 2011. p. 295–312.

Ke Y, Deng L, Ng W, Lee DL. Web dynamics and their ramifications for the development of web search engines. Comput Netw 2006;50(10):1430–47.

Kwon S, Lee S, Kim S. Effective criteria for web page changes. In: Zhou X, Li J, Shen H, Kitsuregawa M, Zhang Y, editors. Frontiers of WWW Research and Development—APWeb 2006, Lecture Notes in Computer Science, vol. 3841, Springer; 2006. p. 837–42.

Kyriazidou E. Testing for serial correlation in multivariate regression models. J Econom 1998;86(2):193–220.

Lim L, Wang M, Padmanabhan S, Vitter J, Agarwal R. Characterizing web document change. In: Wang X, Yu G, Lu H, editors. Advances in web-age information management, Lecture Notes in Computer Science, vol. 2118, Springer; 2001. p. 133–44.

Menascé DA, Almeida VAF. Capacity Planning for Web Services: Metrics, Models, and Methods. Upper Saddle River, NJ, Prentice Hall, 2001.

Olston C, Pandey S. Recrawl scheduling based on information longevity. In: Proceedings of the 17th international conference on world wide web—WWW'08, ACM; 2008. p. 437–46.

Padmanabhan VN, Qiu L. The content and access dynamics of a busy Web site: findings and implications. In: Proceedings of the conference on applications, technologies, architectures, and protocols for computer communication—SIGCOMM'00, ACM; 2000. p. 111–23.

Radinsky K, Bennett PN. Predicting content change on the web. In: Proceedings of the sixth ACM international conference on web search and data mining—WSDM'13, ACM; 2013. p. 415–24.

Radinsky K, Svore K, Dumais S, Teevan J, Bocharov A, Horvitz E. Modeling and predicting behavioral dynamics on the Web. In: Proceedings of the 21st international conference on world wide web—WWW'12, ACM; 2012. p. 599–608.

Radinsky K, Svore KM, Dumais ST, Shokouhi M, Teevan J, Bocharov A, Horvitz E. Behavioral dynamics on the web: learning, modeling, and prediction. ACM Trans Inform Syst 2013;31(3):16:1–37.

Ravi J, Yu Z, Shi W. A survey on dynamic Web content generation and delivery techniques. J Netw Comput Appl 2009;32(5):943–60.

Salton G, McGill MJ. Introduction to Modern Information Retrieval. New York: McGraw-Hill; 1983.

Shi W, Collins E, Karamcheti V. Modeling object characteristics of dynamic Web content. J Parallel Distrib Comput 2003;63(10):963–80.

Trivedi KS. Probability and Statistics with Reliability. Queuing and Computer Science Applications. 2nd ed. New York: Wiley; 2002.

Yang J, Leskovec J. Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on web search and data mining—WSDM'11, ACM; 2011. p. 177–86.

Zhang Y, Jansen BJ, Spink A. Time series analysis of a Web search engine transaction log. Inf Process Manag 2009;45(2):230–45.