Analysis and Forecasting of Web Content Dynamics

Maria Carla Calzarossa
Dept. of Electrical, Computer and Biomedical Engineering
Università degli Studi di Pavia
Via Ferrata 5, I–27100 Pavia, Italy
mcc@unipv.it

Daniele Tessera
Dept. of Mathematics and Physics
Università Cattolica del Sacro Cuore
Via Musei 41, I–25121 Brescia, Italy
daniele.tessera@unicatt.it

***Abstract*—Web content changes have a strong impact on search engines and more generally on technologies dealing with content retrieval and management. These technologies have to take account of the temporal patterns of these changes and adjust their crawling policies accordingly. This paper presents a methodological framework – based on time series analysis – for modeling and predicting the dynamics of the content changes. To test this framework, we analyze the content of three major news websites whose change patterns are characterized by large fluctuations and significant differences across days and hours. The classical decomposition of the observed time series into trend, seasonal and irregular components is applied to identify the weekly and daily patterns as well as the remaining fluctuations. The corresponding models are used for predicting the future dynamics of the sites based on their current and historical behavior.**

***Index Terms*—web content dynamics, time series analysis, web crawling, news websites, search engines, forecasting, workload characterization, ARMA models.**

I. Introduction

Web pages appear (and disappear) on the web all the time and their content can vary over time. All these changes are usually driven by the intent and policies applied by webmasters as well as by external events, such as the spontaneous interactions of the users with websites. Search engines and more generally technologies aimed at content discovery, retrieval and management have to cope with this highly dynamic behavior. Therefore, they need to predict how often and to what extent the content of a website changes and adapt their crawling policies accordingly [1].

In this paper we present a methodological framework for modeling and predicting the temporal patterns of web content changes. The approach relies on time series analysis because of its ability to provide a compact description of the data, explain the past behavior of the change patterns and predict future changes. More precisely, starting from the analysis of the properties of the time series representing these patterns, our framework focuses on the decomposition of the time series and on the identification of the models describing the individual components. These models are also the basis for the extrapolations used to predict the future dynamics of the time series.

For testing our approach, we analyze the dynamics of news websites. In fact, the change patterns of their content are quite peculiar since they are often characterized by large fluctuations together with some periodicity and a time dependent behavior. Models – based on trigonometric polynomials and ARMA components – accurately reproduce the dynamics of these patterns. In addition, these models are used for predicting the future changes of the websites.

This paper is organized as follows. Section II reviews some related works. Section III presents the methodological framework devised to model the time series representing the web content dynamics. An application of this approach is described in Section IV. Finally, Section V draws some conclusions and discusses some open research issues.

II. Related work

Web content has been extensively studied by focusing on diverse aspects related to its characteristics and dynamics (see, e.g., [2]–[10]. These works have important implications in the design and implementation of the technologies aimed at content discovery, retrieval and management. In particular, it has been observed by Fetterly et al. [5] that the size of a web page is a strong predictor of both the frequency and degree of change. Moreover, the correlation among changes makes it possible to predict future changes from the past behavior. Shi et al. [10] outline that a large fraction of objects hosted by news and e-commerce sites does not change within the timescale of a week, whereas objects that change within the timescale of a day are characterized by shorter freshness times.

Brewington and Cybenko [3] as well as Cho and Garcia-Molina [4] address the problem of estimating and predicting the change frequency of web content even in presence of an incomplete change history.

Web dynamics has also been investigated in the framework of the increased complexity of websites. The analysis of modern web traffic presented in [11] suggests that web pages are often characterized by an increased size and a larger number of embedded objects. In addition, Butkiewicz et al. [12] observe that news websites are even more complex as their pages consist of a very large number of embedded objects usually hosted by multiple geographically distributed servers.

The extent of page changes is another important aspect investigated in the literature (see, e.g., [13]–[16]). Various similarity measures, such as edit distance, Dice coefficient, cosine coefficient of similarity, are computed for this purpose. More precisely, in [14] these measures are used to identify the pages whose updates involve a very small fraction of their content and adjust the models of change rates accordingly. On

the contrary, Adar et al. [15] study the nature of changes (i.e., changes to content and structure of web pages) in terms of the frequency and amount of change and identify stable and dynamic content within pages.

Machine learning approaches, numerical fitting techniques as well as time series analysis are in general applied to characterize and model the temporal evolution of web content and predict its dynamics (see, e.g., [17]–[22]). In particular, Yang and Leskovec [18] investigate the temporal patterns associated with online textual content by means of time series analysis and derive the shapes characterizing different types of media. This study suggests that press agencies exhibit a rapid rise followed by a relatively slow decay, while bloggers play a key role in the news longevity. An expert predictive framework based on features, such as relatedness to other pages and similarity in the types of changes, is proposed by Radinsky and Bennett [20] for predicting content changes.

Despite other works, we address the problem of modeling and predicting the temporal patterns of web content changes from a methodological perspective. More specifically, starting from the time series representing these patterns, our framework details the steps to be followed for uncovering their properties and identifying the models able to capture and predict their dynamics.

III. Time series modeling

A time series is an ordered set of discrete or continuous observations taken over time at equally spaced time intervals, that is, $\{Y_t\} = \{y_{t_1}, y_{t_2}, ..., y_{t_N}\}$ with $t_1 \leq t_2 \leq ... \leq t_N$ [23].

Time series modeling is not straightforward as it requires various steps dealing with a preliminary analysis of the properties of the time series followed by the detection of its periodicity and the identification of its components. A final step deals with the prediction of the future dynamics of the time series.

A. Preliminary analysis

The preliminary analysis of the data is mainly aimed at understanding the overall properties of the phenomenon under investigation.

In detail, potential outliers – that is, observations that deviate significantly from their neighbors – have to be detected and removed from the data to avoid perturbations in the models. The outlier identification relies on measures, such as median absolute deviation, Z-score.

In addition, the autocorrelation function r_h computed at varying time lags h suggests how similar a sequence is to its previous values. This function is also used to check the randomness of the data and assess the stationarity of the time series.

We recall that r_h is defined as follows:

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

where $\bar{y}$ denotes the mean value of the time series.

These statistical techniques coupled with the visualization of the time series provide an overview of the analyzed phenomenon and of its temporal fluctuations. In particular, it is possible to highlight trends – denoting steadily increasing or decreasing patterns over quite long periods of time – and seasonal effects – denoting periodic patterns that repeat in time, e.g., hourly, daily, weekly, monthly, yearly patterns.

B. Periodicity detection

The detection of the periodicity of the time series over an observation interval T relies on the spectral analysis of the autocorrelation function and more precisely on the computation of the discrete Fourier coefficients f_k associated with the k/T frequencies, that is:

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

The power spectrum density – represented by the square length of each Fourier coefficient – highlights the peaks in the spectrum of the autocorrelation function. Note that the peaks identify the dominant frequencies corresponding to the periods of the repeated temporal patterns in the time series.

C. Time series decomposition

Time series decomposition – that is, the method applied to uncover the complex multi-pattern behaviors summarized by a time series – is the core of time series analysis. This method is particularly useful because it allows a detailed description of the underlying phenomena, thus resulting in accurate forecasting models.

The decomposition consists of detecting the components of the time series and identifying the corresponding models. The model of the overall time series is then obtained by superimposing the models of the individual components.

A classical decomposition is additive, that is, the overall time series is obtained as the sum of three components, namely, trend, seasonal, and irregular. More precisely, the trend represents the long term variations of the time series, whereas the seasonal component represents the periodic contribution related to cyclic effects (e.g., based on the time of the day or on the day of week). Finally, the irregular component takes into account the remaining contributions.

Depending on the characteristics of the time series, different techniques, such as moving average, exponential smoothing, locally weighted polynomial regression, are applied to identify these components [24]. In particular, the Loess regression – a non parametric local weighted regression technique – is applied for smoothing the data and for identifying the trend, seasonal and irregular components of the time series.

The models of the trend and seasonal components are identified using standard numerical fitting techniques. On the contrary, the approach for modeling the irregular component relies on techniques, such as moving average, auto regressive, Box and Jenkins [25]. In particular, depending on the properties of the irregular component with respect to stationarity, Integrated Auto Regressive Moving Average (ARIMA) or Auto Regressive Moving Average (ARMA) models are

derived. The auto regressive and moving average coefficients of these models are obtained using standard numerical fitting techniques. Moreover, statistical diagnostic tests (e.g., Ljung-Box test, Akaike's information criterion) are applied to identify the number of coefficients to be used in the models, that is, their optimal order.

D. Forecasting

The prediction of the future dynamics of the web content relies on the models of the trend, seasonal and irregular components previously identified. More precisely, 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. For the trend and seasonal components, the values are extrapolated from the corresponding models computed at time $t+h$, whereas a moving horizon prediction technique based on the Box-Jenkins approach is applied to forecast the values of the irregular component.

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.

IV. Experimental results

This section presents an application of the proposed framework to model the changes of the content of three news websites. The crawling process devised to periodically download this content is also described.

A. Crawling process

Crawling is a technique typically applied for retrieving and collecting content from the web. As already pointed out, to test the proposed approach we collect the pages of 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]. Because of the peculiarities of these websites – whose content tends to change rapidly and frequently – we download multiple snapshots of each site.

In particular, our crawling process is performed every 15 minutes using a shell script based on the open source `wget` software package [26]. The crawling starts from the front pages of each website and iteratively follows the hyperlinks extracted from these pages. Therefore, each snapshot consists of all web pages that can be directly or indirectly reached from the front pages of the websites. As a consequence, each individual page can be downloaded multiple times, that is, we collect multiple instances of the pages.

Let us remark that the content changes considered in what follows refer to uploads of new page as well as to updates of pages already published on the website.

Therefore, after the crawling process, we parse each page to extract its textual content, that is, the descriptive text of the news story. Moreover, to detect changes to the content of a page, we compute the cosine coefficient of similarity between its consecutive instances [27]. This coefficient identifies whether the content of a page has been modified and provides a measure of the extent of the changes.

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

B. Dataset properties

Table I summarizes the properties of the datasets collected by crawling three news websites. As can be seen, the three

TABLE I
PROPERTIES OF THE DATASETS COLLECTED BY CRAWLING THREE NEWS WEBSITES.

Website	Number of pages	Number of changes	Crawling interval [days]
CNN	8,302	17,804	104
MSNBC	5,436	12,245	84
Reuters	15,157	18,891	63

datasets – that include in total about 29,000 pages addressed by unique URLs – are quite different. For example, slightly more than 50% of these pages are collected from the Reuters website – that appears to be the "richest" in terms of number of unique pages. In addition, thanks to the cosine coefficient of similarity, we identify about 20,000 updates involving mostly the pages of the CNN and MSNBC websites (with 9,502 and 6,809 updates each) and to a much lesser extent the pages of the Reuters website (with 3,734 updates).

Despite these differences, we have observed that the change patterns of the three websites are rather similar. Hence, due to space limitations, in what follows we present the results of the analysis of one website only, namely, the MSNBC website. Similar models have been obtained for the CNN and Reuters time series.

C. Models of the temporal patterns

Figure 1 shows the temporal patterns of the content changes of the MSNBC website taken every hour over a two weeks observation interval. We notice a large variability in these

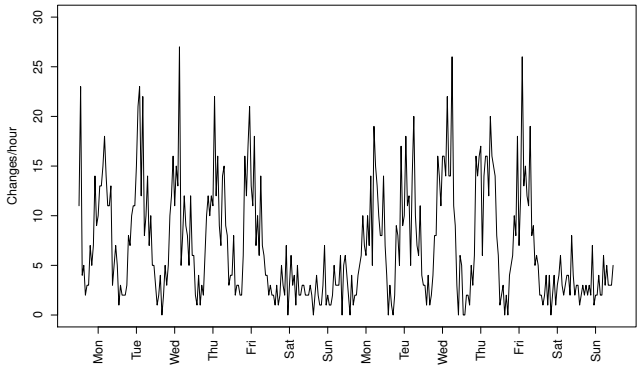

Fig. 1. Temporal patterns of the content changes of the MSNBC website.

patterns with changes not evenly distributed across days and hours and even between the two weeks.

To further investigate the temporal properties of the changes, we compute the autocorrelation function of the time series with

time lags from one hour to one week. The patterns of Figure 2 clearly suggest a periodic behavior of the analyzed changes.

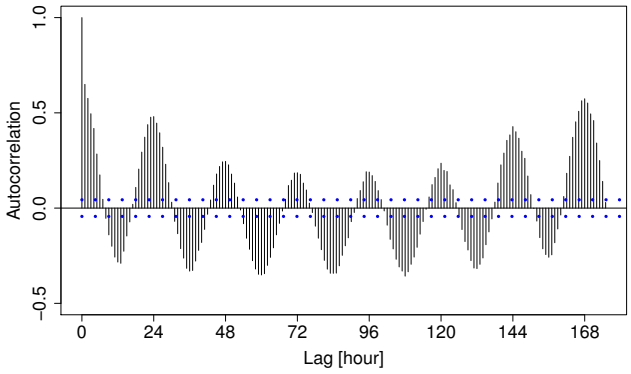

Fig. 2. Autocorrelation function of the time series computed for time lags ranging from one hour up to one week (i.e., 168 hours). The blue dashed lines refer to the 95% confidence bands.

This behavior is also confirmed by the corresponding power spectrum. As shown in Figure 3, the peaks – corresponding to frequencies equal to $10/T$ and $70/T$, T being a ten weeks period — denote weekly and daily patterns. These peaks are

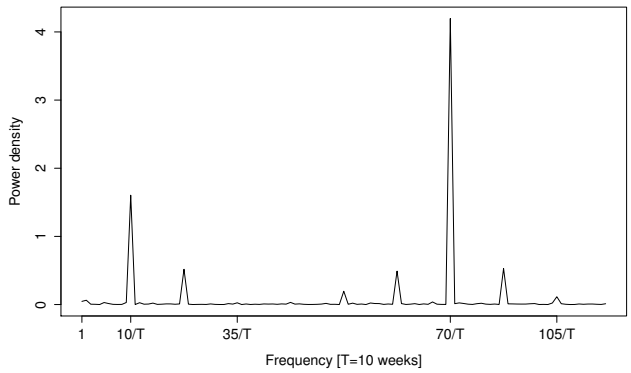

Fig. 3. Power spectrum of the autocorrelation function.

used to identify the periodic components (i.e., trend and daily seasonality) of the time series.

Figure 4 plots the decomposition of the times series shown in Fig. 1. In particular, the estimates of the trend and seasonal components are obtained by applying the Loess method, whereas the irregular component is the remainder of the time series.

To model the behavior of the periodic components we consider trigonometric polynomials whose parameters are identified by applying standard numerical fitting techniques based on the Levenberg Marquardt method. For example, the model of the trend component is a trigonometric polynomial of degree one. In addition, the best fit of the irregular component is represented by an ARMA $(2, 0, 2)$ model.

Figure 5 plots the seasonal component of the time series together with its model, namely, a trigonometric polynomial of degree three described by five parameters.

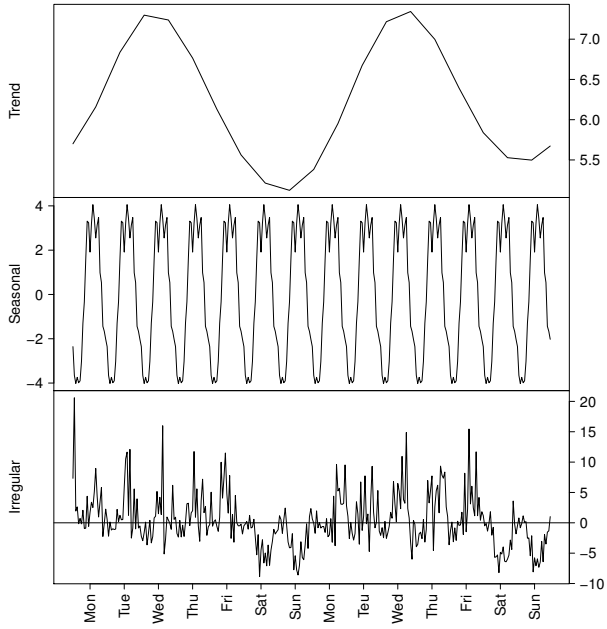

Fig. 4. Decomposition of the time series shown in Fig. 1 into the trend, seasonal and irregular components.

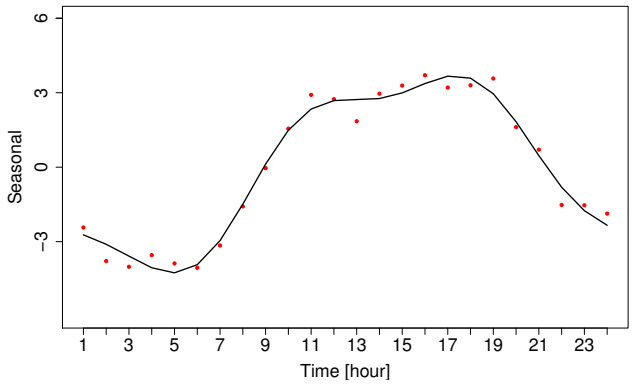

Fig. 5. Seasonal component of the time series (dotted pattern) and corresponding model (solid pattern).

The overall models of the time series are then obtained by superimposing the models derived for the trend, seasonal and irregular components.

D. Change predictions

To predict the future changes based on current and historical data, we extrapolate the trigonometric polynomials that best fit the trend and seasonal components of the time series. On the contrary, the predictions of the irregular components are iteratively derived by means of the Box-Jenkins approach.

Figure 6 shows an example of the predictions over two weeks. Note that to validate our approach we consider data not used for identifying the models of the change patterns. In particular, the diagram plots the empirical data and three predictions corresponding to different time horizons, namely,

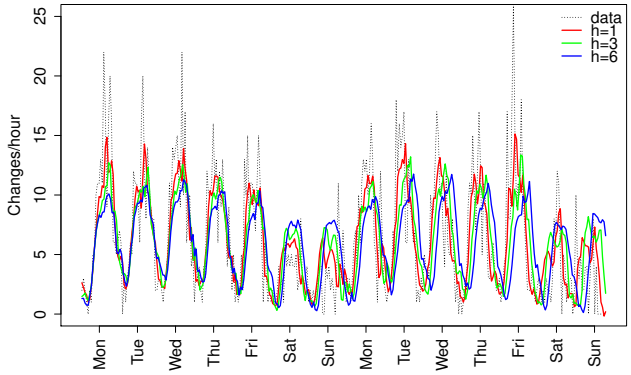

Fig. 6. Predictions of the change patterns for two weeks over three different time horizons, i.e., 1, 3 and 6 hours.

equal to 1, 3 and 6 hours. As can be seen, the curves become smoother as the horizon increases.

In addition, the autocorrelation function of the residuals (i.e., the differences between empirical data and the model) with time lags up to 24 hours (see Fig. 7) suggests that all correlations have been properly taken into account by the model and the few uncorrelated residuals represent the underlying noise.

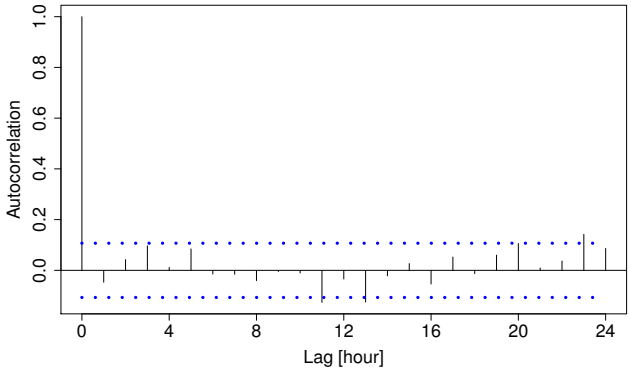

Fig. 7. Autocorrelation function of the residuals. The blue dashed lines refer to the 95% confidence bands.

V. Conclusions

Web content changes regularly over time due to webmaster actions as well as spontaneous interactions of the users with the web. Modeling and predicting the temporal patterns of the content changes are of fundamental importance to efficiently devise technologies aimed at content discovery, retrieval and management. This paper addressed the analysis and forecasting of web content dynamics by presenting a methodological framework based on time series. More precisely, this framework relies on a preliminary analysis of the properties of the time series followed by several steps dealing with the detection of its periodicity, the identification of its constituting components and of the corresponding models as well as the forecasting of its future dynamics.

We tested the proposed approach on datasets collected by crawling three major news websites, namely, CNN, MSNBC and Reuters, because of the peculiarities of their content that tends to change frequently and rapidly over time. The analysis of the time series representing the content changes of these sites has shown that their temporal patterns are characterized by large fluctuations coupled with some periodicity and a time dependent behavior. Hence, each time series was decomposed into three components (i.e., trend, seasonal and irregular components) whose models capture and reproduce the weekly and daily periodicity of the changes and the remaining fluctuations. Moreover, the forecasting approach – aimed at foreseeing the future dynamics of the websites according to their current and historical behavior – provided accurate predictions at time horizons up to three hours ahead without being influenced by the number of values used for the prediction.

As a future work, we plan to extend the proposed approach in different directions, by introducing, for example, a semantic analysis of the content of the pages to assess the actual extent of individual changes. In addition, we plan to consider websites, such as online social networks, whose changes are mainly a consequence of the interactions of the users with the sites.

References

[1] M. Calzarossa, L. Massari, and D. Tessera, "Workload characterization: A survey revisited," *ACM Computing Surveys*, vol. 48, no. 3, pp. 48:1–48:43, 2016.

[2] 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.

[3] B. Brewington and G. Cybenko, "How dynamic is the Web?" *Computer Networks*, vol. 33, no. 1-6, pp. 257–276, 2000.

[4] J. Cho and H. Garcia-Molina, "Estimating frequency of change," *ACM Transactions on Internet Technology*, vol. 3, no. 3, pp. 256–290, 2003.

[5] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A large-scale study of the evolution of Web pages," *Software: Practice & Experience*, vol. 34, no. 2, pp. 213–237, 2004.

[6] 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.

[7] 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.

[8] 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.

[9] V. Padmanabhan and L. Qiu, "The content and access dynamics of a busy Web site: findings and implications," in *Proc. of the Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication - SIGCOMM'00*. ACM, 2000, pp. 111–123.

[10] W. Shi, E. Collins, and V. Karamcheti, "Modeling object characteristics of dynamic Web content," *Journal of Parallel and Distributed Computing*, vol. 63, no. 10, pp. 963 – 980, 2003.

[11] S. Ihm and V. Pai, "Towards understanding modern Web traffic," in *Proc. of the 11th ACM SIGCOMM Conf. on Internet Measurement - IMC'11*. ACM, 2011, pp. 295–312.

[12] M. Butkiewicz, H. Madhyastha, and V. Sekar, "Understanding website complexity: Measurements, metrics, and implications," in *Proc. of the 11th ACM SIGCOMM Conf. on Internet Measurement - IMC'11*. ACM, 2011, pp. 313–328.

[13] A. Ntoulas, J. Cho, and C. Olston, "What's New on the Web?: The Evolution of the Web from a Search Engine Perspective," in *Proc. of the 13th International Conference on World Wide Web (WWW '04)*. ACM, 2004, pp. 1–12.

[14] 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.

[15] 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 Int. Conf. on Web Search and Data Mining - WSDM'09*. ACM, 2009, pp. 282–291.

[16] M. Calzarossa and D. Tessera, "An exploratory analysis of the novelty of a news Web site," in *Proc. Int. Symp. on Performance Evaluation of Computer and Telecommunication Systems - SPECTS 2010*. SCS Press, 2010, pp. 399–404.

[17] Y. Zhang, B. Jansen, and A. Spink, "Time series analysis of a Web search engine transaction log," *Information Processing and Management*, vol. 45, no. 2, pp. 230–245, 2009.

[18] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proc. of the Fourth ACM Int. Conf. on Web Search and Data Mining - WSDM'11*. ACM, 2011, pp. 177–186.

[19] M. Calzarossa and D. Tessera, "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*. IEEE Computer Society Press, 2012, pp. 529–533.

[20] K. Radinsky and P. Bennett, "Predicting Content Change on the Web," in *Proc. of the Sixth ACM Int. Conf. on Web Search and Data Mining - WSDM'13*. ACM, 2013, pp. 415–424.

[21] M. Calzarossa and D. Tessera, "Multivariate analysis of Web content changes," in *Proc. of the 11th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2014)*. IEEE Computer Society Press, 2014, pp. 699–706.

[22] ——, "Modeling and Predicting Temporal Patterns of Web Content Changes," *Journal of Network and Computer Applications*, vol. 56, pp. 115–123, 2015.

[23] J. D. Hamilton, *Time series analysis*. Princeton University Press, 1994.

[24] 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.

[25] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis - Forecasting and Control*, Fourth ed. Wiley, 2008.

[26] Nikšìc, H. et al., *GNU* `wget` *1.18*, 2016, http://www.gnu.org/software/wget/manual/wget.pdf, Last access December 2017.

[27] G. . Salton and M. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.