Multivariate analysis of Web content changes

Maria Carla Calzarossa
Dipartimento di Ingegneria Industriale e Informazione
Università di Pavia
via Ferrata 5 – I-27100 Pavia, Italy
mcc@unipv.it

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

***Abstract*—News websites are expected to deliver in a timely manner the latest stories as well as their latest developments. Thereby, tools, such as, search engines, need to cope with these rapid and frequent content changes by adjusting their crawling activities accordingly. In this paper we explore and model the properties and temporal behavior of the content changes of three major news websites. The dynamics of the changes is characterized by large fluctuations and significant differences from day to day and from hour to hour. However, a certain degree of similarity in the overall patterns of each website exists. In particular, the application of multivariate analysis techniques allows us to identify groups of days with similar change patterns, thus allowing for the customization of the crawling policies adopted by search engines.**

I. Introduction

Web content is characterized by an increased dynamic nature and complex usage patterns that make its management and retrieval rather challenging [5]. The content of a website changes whenever something new is posted or existing content is updated or removed. The frequency of these changes varies from site to site and from time to time. In the case of news and e-commerce websites, it has been shown [21] that a large fraction of objects does not change within a timescale of a week, whereas objects that change within the timescale of a day are characterized by short freshness times. To meet users expectations, search engines have to cope with this variability, index Web content in a timely manner and minimize, at the same time, the costs for its download, storage and management. It is then important for these tools to take into account the temporal behavior of the content changes in their crawling strategies.

In this paper we investigate the properties and the temporal patterns of the content changes of three news websites. In particular, multivariate analysis techniques allow us to model the behavior of the hourly changes across days and identify groups of days with similar patterns. Note that even though our study focuses on news websites, the approach proposed here is general enough and encompasses the behavior and temporal evolution of any type of Web content.

The paper is organized as follows. Section II presents an overview of the literature on the issues related to Web dynamics. The multivariate analysis techniques applied to model the change patterns are briefly described in Sect. III. Section IV summarizes the main characteristics of our datasets, whereas the experimental results are presented in Sections V. Finally, some conclusions and future research directions are discussed in Section VI.

II. Related work

Web dynamics is a challenging problem with important implications in many application domains. In the literature this topic has been addressed under different perspectives. Several papers (see, e.g., [10], [17], [19]) mainly focus on aspects, such as, degree of Web document changes and presence of clustered changes, whereas others (see, e.g., [4], [6], [8], [18]) study the rate of change of Web content. A comprehensive survey of the major research issues in Web dynamics is presented by Ke et al. [16] in the framework of four dimensions, namely, size, pages, link structures, and user interests.

To study Web dynamics, Brewington and Cybenko [4] consider properties, such as, frequency and nature of modifications of Web pages. Further characteristics, e.g., link structure over time, rate of creation of new pages and new distinct content as well as rate of change of the content itself, are explored in [18] by analyzing weekly snapshots collected on some 150 websites for one year. Adar et al. present in [1] a fine grain characterization of the evolution of Web content that focuses on the nature of changes, i.e., changes to content and structure of Web pages. Stable and dynamic content within each page are then identified.

Rate and degree of changes are the basic features selected in [10] for investigating the evolution of Web pages. This large scale experimental study shows that whenever Web pages change, they usually change only in their markup or in trivial ways. In addition, document size is a strong predictor of both frequency and degree of change. In particular, large documents tend to change more often and more extensively than smaller ones. Similarly, some well defined patterns are identified in the rates of page creations and updates studied in [6], where most updates involve a small fraction of the page content, and few are more extensive. In [7] the evolution of Web content is addressed under a different perspective, that is, by considering its novelty, namely, how fast and to what extent a single page and the entire collection of pages posted on websites are modified.

Other important aspects investigated in the framework of Web dynamics refer to the relationships and associations between dynamics and user accesses. Castillo et al. [9] have recently analyzed the life cycle of news articles posted on line to identify classes of news stories and evaluate the user behavior. The amount and type of changes to Web page content and the corresponding user visit behavior, considered in [2], show that different visit patterns resonate with different kinds of change, for example, with the rate of change of interesting content.

Our work complements and enhances the studies on Web dynamics previously described. The main contribution of this paper is two-fold. We characterize and model the properties of Web content changes over time and apply multivariate analysis techniques for classifying the change patterns of the various days.

III. Multivariate analysis techniques

The methodological approach adopted for the analysis of Web content changes relies on the application of various multivariate analysis techniques. These techniques allow us to investigate and discover the main characteristics of the content changes and model their temporal properties. More specifically, as a preliminary step of our exploratory analysis, we examine the change patterns for detecting outliers, that is, unusual and anomalous observations that might be present in the collected data. Statistical techniques applied in combination with visualization techniques work well for this purpose. We point out that whenever outliers are detected, it is usually advisable to remove them to avoid any perturbation in the following analysis.

Moreover, to highlight similarities in the number of changes across hours and days, multivariate analysis techniques, such as, clustering [14], are applied in conjunction with Principal Component Analysis (PCA) [15]. In particular, PCA is applied to reduce the dimensionality of the data, while retaining as much as possible their variation, whereas clustering techniques are used to identify groups of homogeneous observations. Let us recall that PCA linearly transforms correlated parameters into a set of uncorrelated parameters, i.e., the principal components (PCs). These components, ranked in decreasing order of importance, explain and summarize the variability in the original data.

After the selection of the number of principal components to be retained, clustering is applied to the data represented in the m-dimensional subspace defined by the first m PCs. Clustering allows us to build groups with homogeneous characteristics from heterogeneous data such that the variance within groups is minimized and the variance among groups is maximized. The groups identified by clustering are such that observations within each group are sufficiently similar to be treated identically for the purpose of any further analysis. Hence, the centroids, i.e., the geometric centers of the individual groups, can be used as the representatives of the change patterns of the websites.

IV. Data characteristics

Our datasets refer to three major news websites, namely, the websites owned by the CNN[1] and MSNBC[2] cable news channels and by the Reuters[3] news agency. We crawled each site every 15 minutes for several weeks by initially downloading their front pages. We then extract the hyperlinks contained in these pages and iteratively recrawl the sites by following these hyperlinks and downloading the corresponding pages. A snapshot of a website then includes pages posted since the previous download as well as pages already downloaded and whose hyperlinks are still present in the front pages.

For our crawling activities we implement a shell script based on the open source `wget` software package [11] to download the pages using the HTTP protocol. The number of pages downloaded from each site does not vary significantly across snapshots, whereas it varies across sites.

Let us remark that each Web page consists of a template and a large variety of dynamic content, such as, images, banners, videos, advertisements, often customized according to users preferences, whose number and change patterns vary from page to page and from time to time. In our study, instead of focusing on the dynamics of the entire page, we analyze what we consider the most important part of a news story, that is, its textual content. Hence, after each download, we parse the Web pages to extract their content. Note that in what follows with the term Web page we refer to its textual content only.

After each download, we analyze the pages for detecting possible updates. More specifically, among the various metrics used to determine the similarity between documents, our detection mechanisms rely on the cosine coefficient of similarity [20] and on the edit distance [13]. Both metrics are computed between consecutive instances of a given page.

Table I summarizes the results of our crawling activities. Even though the three websites offer the same type of content, that is, news stories, the corresponding datasets differ significantly in terms of both number of Web pages

	unique pages	updates	crawling interval
CNN	8,302	9,502	104 days
MSNBC	5,436	6,809	84 days
Reuters	15,157	3,734	63 days

Table I
Characteristics of the datasets of the three news websites.

[1] http://www.cnn.com

[2] http://www.msnbc.com

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

and number of updates. Moreover, it is interesting to point out that even the percentage of pages involved by updates varies from site to site. For example, updates involve more than half pages downloaded from the CNN and MSNBC websites, and only 15% of the pages of the Reuters website, thus showing the different policies adopted by individual websites with respect to their content management.

V. Experimental results

To study the dynamics of the websites we focus on the number of changes per hour of their content. Let us recall that changes refer to uploads of new Web pages or updates of existing ones. We do not consider page removal as pages seldom disappear from the sites even though their hyperlinks may disappear from the front pages.

Figure 1 shows the temporal patterns of the changes detected for the MSNBC website over a two weeks interval. We observe significant fluctuations and diurnal patterns with most changes concentrated during the day and much fewer over night. On average, during the entire crawling interval we detect 6.1 changes per hour. Moreover, we identify similar characteristics for the change patterns of the other two websites, even though their average number of changes is slightly larger, namely, 7.1 and 12.5 changes per hour for the CNN and Reuters websites, respectively.

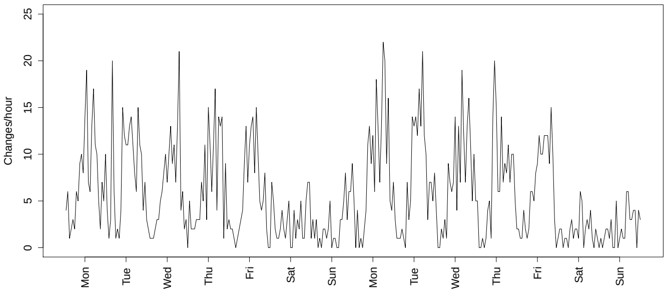
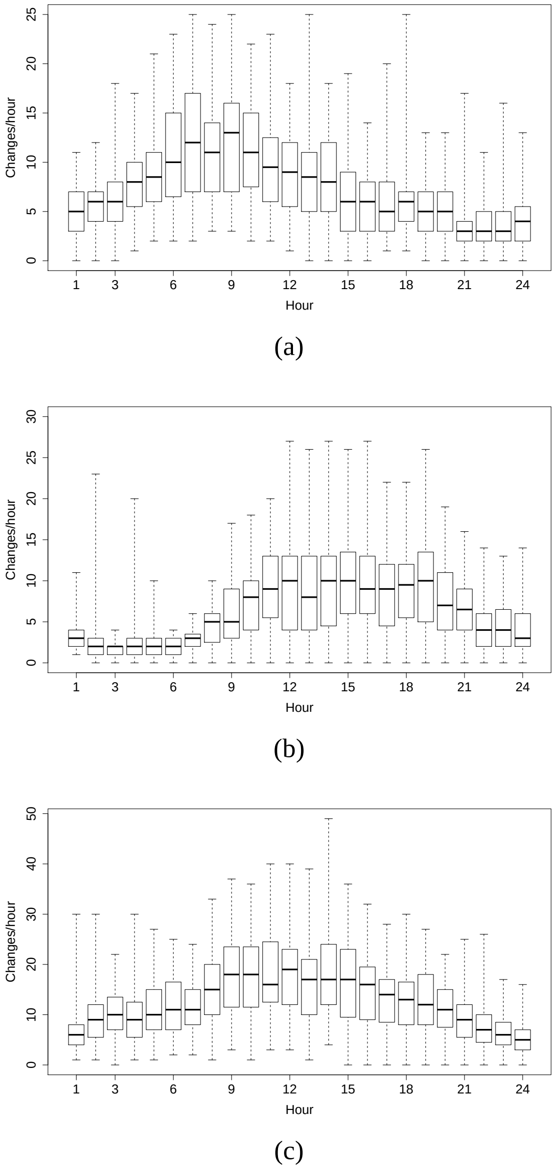


Figure 1. Changes per hour of the MSNBC website over a two weeks interval. The labels on the x axis are centered at 12 noon.

To explore the properties and model the temporal patterns characterizing the change dynamics, we analyze the data collected on each website under two different perspectives, namely, individual hours and daily patterns. In particular, to investigate the variability of the number of changes in each hour, we pool the data of the corresponding hours. On the contrary, to discover similarities across days, we describe each day of our crawling intervals in terms of the number of changes in each of the 24 hours.

The box plots of Figure 2 summarize the distributions of the number of changes in each hour computed over the entire crawling intervals. The diagrams display the medians, denoted by the solid lines drawn in each box, and two measures of dispersion, namely, the ranges and inter–quartile ranges of the distributions. In particular, the upper and lower boundaries of each box correspond to the first and third quartiles. The figure confirms the clear diurnal patterns characterizing the number of changes. In addition, even though each crawling interval spans several weeks, the variability within the individual hours is in general rather small. Nevertheless, we observe some interesting differences among the sites. In the CNN website, most changes occur at 9am or earlier, as denoted by the corresponding peaks. On the contrary, the peaks are slightly shifted towards later hours in the case of the MSNBC and Reuters websites, thus



(a)

(b)

(c)

Figure 2. Box plots of the distribution of the number of changes in the various hours for CNN (a), MSNBC (b) and Reuters (c) websites.

highlighting significant differences in how the content of each website is uploaded and managed.

A deeper inspection of the box plots shows possible outliers in the data. In particular, we notice some ranges much bigger than the corresponding inter–quartile ranges. This is the case, for example, of the hourly changes detected for the MSNBC website at 2am and 4am, whose ranges are about an order of magnitude bigger than the corresponding inter–quartile ranges. We will further explore these anomalous behaviors when analyzing the daily patterns.

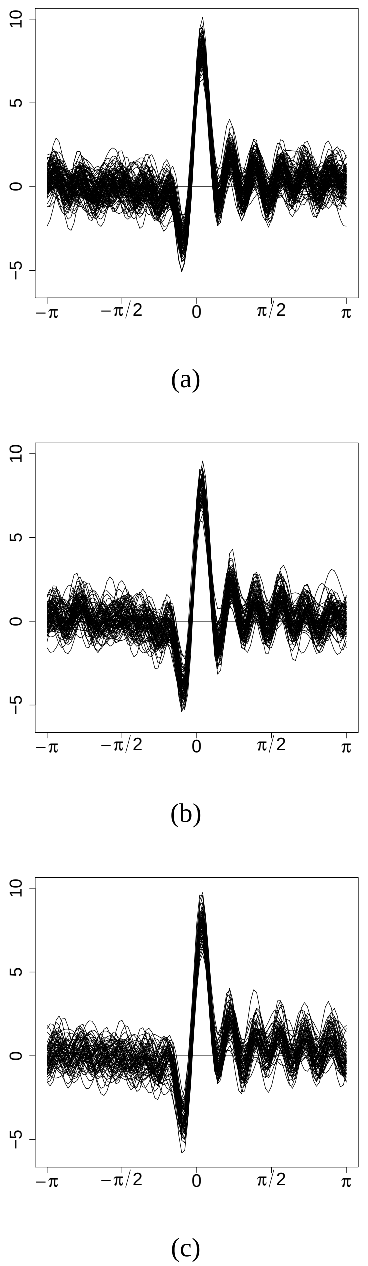
To discover similarities in the change patterns across days, we apply clustering techniques to the observations, i.e., days, represented as points in the multidimensional space of their parameters. As already pointed out, a day is described by 24 parameters, each referring to the number of changes detected in a given hour, that is, from 1am until midnight.

To simplify the description of the datasets and the corresponding models, as an intermediate step towards clustering, we apply the Principal Component Analysis. Moreover, to further investigate the presence of outliers, we resort to statistical and visualization techniques. In detail, to simultaneously display multivariate data, we plot the Andrews curves of the observations in the space of the principal components [3]. Each curve is defined for t in the range $[-\pi, \pi]$ by the following function:

$$\begin{aligned} f(t) \quad = \quad & z_1/\sqrt{2} + \; z_2 \, \sin(t) + z_3 \, \cos(t) + z_4 \, \sin(2t) + \\ & + z_5 \, \cos(2t) + ... + z_{23} \, \cos(11t) + z_{24} \, \sin(12t) \end{aligned}$$

where $z_1, z_2, ..., z_{24}$ denote the principal components of the individual days. Figure 3 plots the Andrews curves that summarize the behavior of the various days. We emphasize that in these types of diagrams, outliers usually appear as single Andrews curves that look different from the rest. In our case, the large datasets make the diagrams rather difficult to interpret. Nevertheless, a deeper inspection shows two potential outliers in the MSNBC dataset. Hence, we complement this conjecture with some quantitative measures of the dissimilarities across days by computing the corresponding Mahalanobis distances. The maximum distance is equal to 66.42, that is, about three times bigger than the median of the corresponding distribution. In addition, its 98th percentile is equal to 62.03, whereas its 97th is equal to 46.29. Thereby, these results confirm the presence of the two outliers visually identified. We emphasize that these outliers are also responsible of the anomalous behaviors shown in the box plot (see Fig. 2 (b)).

After the removal of the two outliers, we apply the PCA to the new datasets. Table II presents the eigenvalues obtained from the correlation matrix and the cumulative percentages of the variance explained by the first ten PCs of the CNN, MSNBC and Reuters datasets. The first component summarizes the largest part of the variance of each of the three datasets, i.e., about one third of their overall variance. The second component, that is orthogonal to the first one, summarizes the largest remaining variance, that is, 13.2%, 7.98% and 8.49%, respectively, and so on.



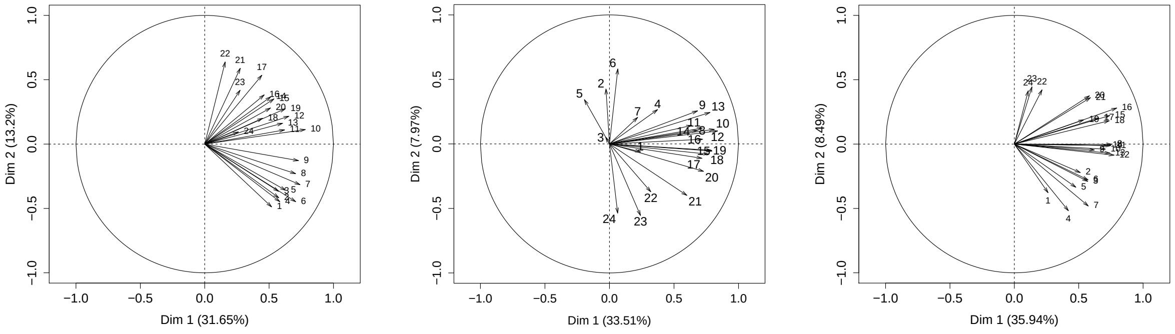
(a)

(b)

(c)

Figure 3. Andrews curves of the CNN, MSNBC and Reuters observations plotted in the space of their principal components.

To characterize and evaluate the relationships between the principal components and the original parameters describing the hourly changes of each day, we compute the corresponding coefficients of correlation, i.e., the loadings of the parameters on the principal components. These coefficients provide estimates of the information shared between



(a)

(b)

(c)

Figure 4. Correlations of the 24 parameters in the space of principal components 1 and 2 for the CNN (a), MSNBC (b) and Reuters (c) datasets.

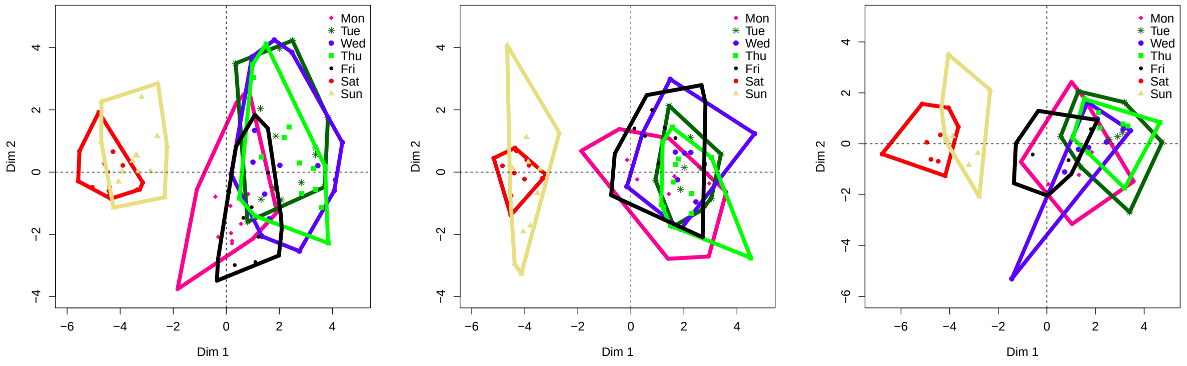
	CNN		MSNBC		Reuters	
PC	Eigen value	Perc. cum. var.	Eigen value	Perc. cum. var.	Eigen value	Perc. cum. var.
1	7.596	31.65	8.043	33.51	8.624	35.94
2	3.167	44.85	1.914	41.49	2.037	44.43
3	1.878	52.67	1.684	48.50	1.585	51.03
4	1.123	57.35	1.352	54.13	1.362	56.70
5	0.994	61.49	1.238	59.29	1.164	61.55
6	0.902	65.25	1.084	63.80	1.099	66.13
7	0.878	68.91	0.964	67.82	1.023	70.39
8	0.811	72.29	0.931	71.70	0.934	74.28
9	0.777	75.53	0.820	75.12	0.775	77.51
10	0.694	78.42	0.751	78.25	0.692	80.40

Table II
EIGENVALUES OF THE CORRELATION MATRIX AND CUMULATIVE PERCENTAGES OF THE EXPLAINED VARIANCE OF THE FIRST TEN PRINCIPAL COMPONENTS (PC).

the parameters and the PCs. Figure 4 plots, in the circle of correlations, the 24 parameters on the two dimensions corresponding to PCs 1 and 2. The parameters, labeled in the figure from 1 to 24 according to the hour they refer to, are represented as points whose coordinates correspond to their loadings on the two PCs. Thus, the positions of the points in the circle denote the relative importance of the parameters for the PCs. By looking at the individual circles, we can see several differences. For the CNN and Reuters datasets, all loadings on the first PC are positive. In addition, for the CNN dataset, almost all loadings are about equal. This means that each parameter is evenly represented in the linear combination. Thereby, the first principal component can be interpreted as an overall dimension that does not clearly differentiate the number of changes in the various hours. On the contrary, for the MSNBC dataset, we observe two negative loadings as well as some points very close to the center of the circle, thus indicating the limited contributions of corresponding parameters in the description of PCs 1 and 2. By analyzing in more detail the correlations associated with the second principal component, we notice that this component describes for every dataset contrasts among the parameters. These contrasts are characterized by a clear pattern for the CNN dataset (see Fig. 4 (a)), where the points lie in different sides of the horizontal axis, namely, 15 above and the remaining nine below the axis. In the case of the MSNBC dataset (Fig. 4 (b)), PC 2 is characterized by some large negative loadings corresponding to parameters 22, 23 and 24. Finally, for the Reuters dataset, we observe (Fig. 4 (c)) that about one third of the correlations are very small, hence, the impact of the corresponding parameters on this PC is negligible, whereas larger correlations are associated with parameters describing the changes in the late evening hours.

The projections of the observations in the space of their first two principal components are plotted in Figure 5. Let us remark that these two dimensions represent slightly less than 45% of the total variance of the CNN and Reuters datasets, and some 42% for the MSNBC dataset. The convex hulls superimposed on each diagram identify the areas covered by the various days of the week. As expected, weekdays and weekend days lie in very different areas of the diagrams. Their separation is mainly explained by the first principal component. Moreover, weekend days overlap to a different extent. An analogous remark applies to weekdays, even though, in general, we notice a larger degree of dispersion mainly explained by PC 2.

Clustering techniques are then applied in the space of the principal components. Since the objective is to use a number of PCs much smaller than the number of parameters without any significant information loss, we base their selection on the corresponding eigenvalues (see Table II). More specifically, we retain the components whose eigenvalues are larger than the average, that is, larger than one. According to this criterion, instead of using 24 parameters, we use five, six and seven PCs, explaining about 61%, 64% and 70% of the variance of the CNN, MSNBC and Reuters datasets, respectively. Note that for the CNN dataset we include the fifth PC since its eigenvalue is almost equal to one, namely,



(a) (b) (c)

Figure 5. Scatter plots of the observations in the space of principal components 1 and 2 for CNN (a), MSNBC (b) and Reuters (c) datasets with superimposition of the convex hulls for the various days of the week.

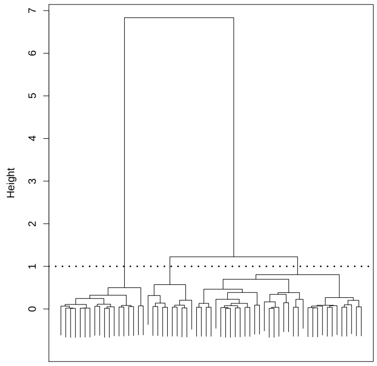


Figure 6. Dendrogram obtained for the Reuters dataset.

0.994.

Among the various clustering techniques, we apply a hierarchical algorithm that computes the similarity between pairs of observations according to the Euclidean distance. Moreover, the dissimilarity between clusters is based on the Ward method that relies on the analysis of variance, that is, the total sum of squared deviations from the mean of a cluster [12].

A graphical representation of the results of the clustering applied to the observations of the Reuters dataset is shown in Figure 6. The dendrogram plotted in the figure displays the observations, whose labels have been omitted for legibility, and the sequence of clusters. The heights of the tree represent the distances between the clusters and are proportional to the within-clusters variance. By pruning the tree at a given level, we obtain the branches that describe the clusters. The horizontal dotted line drawn in the figure at height one corresponds to the level of our pruning. As a result of this cut, we obtain three subtrees each representing a cluster. These clusters group 18, 14 and 31 observations each.

The same approach applied to the CNN and MSNBC datasets yields subdivisions in three clusters. Table III summarizes the composition of the various clusters. The table reports the number of observations of each cluster and their breakdown across days of the week. Clustering confirms the separation between weekdays and weekend days, also highlighted by the convex hulls of Fig. 5. Moreover, the observation referring to weekend days are very homogeneous and belong to a single cluster which contains very few, if any, other observations. On the contrary, most observations referring to weekdays belong to two clusters.

	CNN			MSNBC			Reuters		
	Cluster			Cluster			Cluster		
	1	2	3	1	2	3	1	2	3
Monday	2	11	2	1	6	5	0	5	4
Tuesday	0	10	4	0	4	6	0	0	9
Wednesday	0	11	4	0	7	5	0	2	7
Thursday	0	11	4	0	4	8	0	0	9
Friday	0	12	3	0	11	1	0	7	2
Saturday	15	0	0	12	0	0	9	0	0
Sunday	15	0	0	12	0	0	9	0	0
# of observations	32	55	17	25	32	25	18	14	31

Table III
COMPOSITION OF THE CLUSTERS OBTAINED FOR THE THREE DATASETS.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
cl 1	-4.111	0.302	0.194	0.127	-0.177	0.033	-0.058
cl 2	0.020	-1.301	-0.207	-0.252	0.789	-0.190	0.208
cl 3	2.378	0.413	-0.019	0.040	-0.253	0.066	-0.060

Table IV
CENTROIDS OF THE THREE CLUSTERS OBTAINED FOR THE REUTERS DATASET.

The centroids of the three clusters obtained for the Reuters dataset are presented in Table IV. As can be seen, the clusters are very well defined. PC 1 clearly contrasts cluster 1 with cluster 3, whereas PC 2 contrasts cluster 2 with clusters 1 and 3. Similar remarks apply to the remaining PCs. In particular, we notice that in clusters 2 and 3, five PCs out of seven have opposite signs. Let us recall that the positive big loadings on the first PC denote its strong direct correlation with the original parameters, whereas PC 2 is strongly correlated with the parameters describing the changes in late evening hours. Thus, the observations of cluster 1 are characterized by few hourly changes, whereas in cluster 3 we find observations with many hourly changes mostly concentrated in late morning hours. Cluster 2 is a small cluster whose observations are characterized by few changes in the late evening hours.

The clusters obtained for the CNN and MSNBC datasets exhibit similar behaviors. In detail, for both websites PC 1 contrasts cluster 1 with clusters 2 and 3, whereas PC 2 contrasts cluster 2 with cluster 3. Again, the observations of these two clusters mainly differ in terms of number of changes in the early morning hours and in the late hours of the day. Moreover, for the MSNBC dataset we identify a cluster that groups the observations with the largest number of changes in the afternoon hours.

It is worth to point out that the classification process applied in the space of the principal components is very robust. Despite of the small number of PCs used to describe each observation, the groups identified by this approach do not differ from those obtained using all the 24 original parameters.

VI. Conclusions

Web content changes continuously over time. To explore the characteristics and model the temporal properties of change patterns of three major news websites, we applied a methodological approach based on various multivariate analysis techniques. Our study has shown some interesting findings. First of all, to our surprise the content of these websites in general does not change that often. In addition, the distinct patterns detected for the three websites are mainly due to their different content management policies. Nevertheless, even though the number of changes varies significantly from day to day, some similarities and well defined patterns exist across days. To identify and model these similarities we applied clustering techniques in conjunction with Principal Component Analysis. As a result, we identified groups of days classified according to their change patterns. As expected, weekdays and weekend days were set apart in different groups. Similarly, weekdays are clustered according to their change patterns. In particular, days with the largest number of changes in the late morning hours are clustered together.

The daily patterns associated with each cluster could then be exploited for tuning and validation of the crawling strategies employed by search engines.

Finally, it is worth noting that the combined application of clustering and PCA makes the approach adopted for the classification of the change patterns rather fast and efficient, thus allowing for the analysis of multiple datasets collected on different websites.

As a future work, we plan to study the dynamics of websites deployed in different application domains, e.g., institutional websites, personal websites, blogs, social networks, and introduce additional metrics to explain their complexity.

References

[1] E. Adar, J. Teevan, S. T. Dumais, and J.L. 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*, pages 282–291. ACM, 2009.

[2] E. Adar, J. Teevan, and S.T. Dumais. Resonance on the Web: Web dynamics and revisitation patterns. In *Proc. of the 27th Int. Conf. on Human factors in computing systems - CHI'09*, pages 1381–1390. ACM, 2009.

[3] D. F. Andrews. Plots of High-Dimensional Data. *Biometrics*, 28(1):pp. 125–136, 1972.

[4] B.E. Brewington and G. Cybenko. How dynamic is the Web? *Computer Networks*, 33(1-6):257–276, 2000.

[5] M. Butkiewicz, H.V. Madhyastha, and V. Sekar. Characterizing Web Page Complexity and Its Impact. *IEEE/ACM Transactions on Networking*, 22(3):943–956, 2014.

[6] M. Calzarossa and D. Tessera. Characterization of the evolution a news Web site. *Journal of Systems and Software*, 81(12):2236–2344, 2008.

[7] 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*, pages 399–404. SCS Press, 2010.

[8] M. Calzarossa and D. Tessera. Time series analysis of the dynamics of news websites. In *Proc. of the Thirteenth Int. Conf. on Parallel and Distributed Computing, Applications and Technologies - PDCAT'12*, pages 529–533. IEEE Computer Society Press, 2012.

[9] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the Life Cycle of Online News Stories Using Social Media Reactions. In *Proc. of the 17th ACM Conf. on Computer Supported Cooperative Work & Social Computing - CSCW'14*, pages 211–223. ACM, 2014.

[10] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of Web pages. *Software: Practice & Experience*, 34(2):213–237, 2004.

[11] Free Software Foundation. GNU `wget` Manual. http://www.gnu.org/software/wget/manual/wget.pdf.

[12] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. SIAM, 2007.

[13] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[14] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Survey*, 31(3):264–323, 1999.

[15] I. T. Jolliffe. *Principal Component Analysis - Second Edition*. Springer, 2002.

[16] Y. Ke, L. Deng, W. Ng, and D.-L. Lee. Web dynamics and their ramifications for the development of web search engines. *Computer Networks*, 50(10):1430–1447, 2006.

[17] L. Lim, M. Wang, S. Padmanabhan, J. Vitter, and R. Agarwal. Characterizing web document change. In X. Wang, Ge Yu, and Hongjun Lu, editors, *Advances in Web-Age Information Management*, volume 2118 of *Lecture Notes in Computer Science*, pages 133–144. Springer, 2001.

[18] 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 ACM Int. Conf. on World Wide Web - WWW'04*, pages 1–12, 2004.

[19] K. Radinsky and P.N. 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.

[20] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[21] W. Shi, E. Collins, and V. Karamcheti. Modeling object characteristics of dynamic Web content. *Journal of Parallel and Distributed Computing*, 63(10):963 – 980, 2003.