# Characterization of the Evolution of a News Web Site

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

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

**Abstract**

The Web has become a ubiquitous tool for distributing knowledge and information and for conducting businesses. To exploit the huge potential of the Web as a global information repository, it is necessary to understand its dynamics. These issues are particularly important for news Web sites as they are expected to provide fresh information on current world events to a potentially large user population. This paper presents an experimental study aimed at characterizing and modeling the evolution of a news Web site. We focused on the MSNBC Web site as it is a good representative of its category in terms of structure, news coverage and popularity. Specifically, we analyzed how often and to what extent the content of this site changed and we identified models describing its dynamics. The study has shown that the rate of page creations and updates was characterized by some well defined patterns that varied as a function of time of day and day of week. On the contrary, the content of individual pages changed to a different extent. Most updates involved a very small fraction of their content, whereas very few were more extensive and spread over the whole page. By taking into accounts all these aspects, we derived analytical models able to accurately capture and reproduce the evolution of the news Web site.

# 1 Introduction

As the Web is getting richer and richer of all kinds of information, the major challenge faced by the users is the ability to locate the most relevant and up-to-date information. In this framework, tools, such as, search engines and alerting/publishing services, play a key role. To meet user requirements and expectations, these tools have to address the challenging issues arising from the highly evolving Web, where pages are created, modified and eventually deleted by their owners.

Page creations produce new information, whereas page updates modify existing information. To keep the information fresh, all these changes need to be captured in a timely manner. Similarly, when a page disappears, it is important to detect as soon as possible its deletion.

News Web sites face the very dynamic setting of world events where they are expected to provide up-to-date information about current events to a potentially large user population. Thus, new pages are often created and the content of existing pages frequently modified. On the contrary pages are seldom deleted: they are usually stored in some historical news archives.

This paper analyzes the evolution of news Web sites from a methodological perspective with the objective of identifying models able to describe their dynamics. In particular, to assess how often and to what extent the content of news Web sites changed, we focused on the MSNBC Web site [18], since we believe it is a good representative of this category of sites in terms of structure, news coverage and freshness and popularity. For this site, we studied the creations of new pages and updates of the content of existing pages. We did not address the persistence of Web pages in that, even though pages were no longer "on–line", that is, reachable by links available on the site, their bookmarks usually worked.

The analysis of the measurements collected by monitoring the MSNBC news Web site over 19 continuous weeks has shown that page creations and updates were both characterized by some well defined daily and weekly patterns that persisted over the entire monitoring period. On the contrary, not all updates were equally extensive. Some updates involved the whole page, whereas others were rather clustered and limited to a very small fraction of its content.

As a main contribution of our work, we then identified models that captured and described the evolution of the MSNBC Web site. Our investigation has also shown that these models could accurately describe the behavior of news web sites other than the MSNBC's.

Let us remark that analytical models describing the evolution of a Web site are very useful to predict the future behavior of the site from its past behavior and to customize, for example, the crawling strategies of search engines and the polling policies of services, such as, RSS feeds.

The rest of the paper is organized as follows. Section 2 presents the related work on Web dynamics. Section 3 introduces the experimental environment of our study and presents some preliminary findings. The models characterizing the evolution of the MSNBC Web site are described in Section 4. The outcomes of the analysis of the content updates are discussed in Section 5. Finally, Section 6 draws some conclusions and outlines future developments.

## 2 Related work

Web dynamics has been addressed in the literature under two different dimensions, namely, growth, that is, page creation, and update, that is, page content change. An overview of these topics is presented in [16], where the focus was on the four factors that characterize Web dynamics, i.e., size, pages, link structures, and user interests, and on their influence on the design and development of search engines.

Large-scale studies dealing with the characterization of the evolution of Web sites have been presented in several papers (see e.g., [1, 2, 5, 6, 9, 12, 21]). In particular, Fetterly et al. [12] analyzed several million of pages with the aim of measuring the rate and the degree of changes to Web pages. The statistical observations of the measurements showed that page size was a strong predictor of both frequency and degree of change. Similarly, even though the average degree of change varied widely across top-level domains, changes were rather correlated, that is, past changes to a page were a good predictor of future changes.

Brewington and Cybenko [5, 6] focused on the frequency and nature of modifications of a large sample of Web pages with the aim of estimating revisit rates of search engines. Exponential distributions were used to model the times between individual page changes. These distributions were the basis of the models of the change rates and of the definition of a metric that measured the currency of a page.

Several estimators of the change frequency of Web pages in presence of an incomplete change history have been developed by Cho and Garcia-Molina [9]. The estimators assumed page changes modeled by Poisson processes. Nevertheless, the authors showed the effectiveness of the estimators even in the presence of pages not following the Poisson model.

Web dynamics has also been studied under a different perspective by considering to what extent the content of a Web page changed whenever it was updated. Ntoulas et al. [19] focused on the evolution of content and link structure by analyzing weekly snapshots of some 150 Web sites for one year. The study showed that there was a high turnover of Web pages and hyperlinks. Moreover, the rate of content change for a given page was likely to remain consistent over time. In [7] the cosine coefficient of similarity was applied to evaluate the degree of change

of Web pages; the analysis has shown that most pages changed to a limited extent independently of their size. Gabrilovich et al. [13] presented a methodology for filtering news stories based on the novelty of the information. The proposed algorithms analyzed how information evolved over time within individual articles and across articles and quantified the importance and relevance of each update. An analysis of Web document changes aimed at improving the methods to keep Web indices up-to-date is presented in [17].

Our study differs from previous studies in that we derived analytical models able to describe the evolution of a Web site without making statistical assumptions on the behavior of the creation and update processes in that they might influence the characteristics and the accuracy of the resulting models. Moreover, the analysis of the content of the pages and of the amount of new information introduced by each update allowed us to fine tune our models. Hence, the resulting models did only take into into account the updates that could be classified as "most extensive".

## 3 Preliminary analysis

The basis of our study is represented by measurements collected by monitoring the MSNBC news Web site [18] for 19 continuous weeks, that is, 133 days, starting mid November 2004. Despite most of other studies, we collected snapshots of the site at a very fine grain by downloading every 15 minutes the pages belonging to the six major news categories available on the site, that is, business, entertainment, health, news, sports, and technology and science.

To investigate the dynamics of the site, we analyzed both the creations of new Web pages and the updates of the content of existing pages. During our monitoring interval, 14,165 new pages were uploaded to the site, that is, on average 106.5 pages per day. Of these pages, 42.74% belonged to the sports category, 25.24% to the news, 10.74% to the business, 9.11% to the technology and science category, and the remaining pages were almost evenly distributed between the entertainment and health categories. Figure 1 shows the average number of pages created per day, broken down according to the six categories analyzed in our study. As can be seen, the creation process was characterized by a large variability across days. There were many more uploads of new pages on weekdays than on week-end days. In particular, about twice as many pages were uploaded on Tuesdays (i.e., 136) than on Saturdays (i.e., 67). This is mainly due to the lack of most business activities on Saturdays.

A snapshot of the temporal evolution of the number of pages created over six weeks starting Monday, January 10, 2005, is shown in Figure 2. Even though there is a clear weekly pattern, the number of pages uploaded per day of the week did

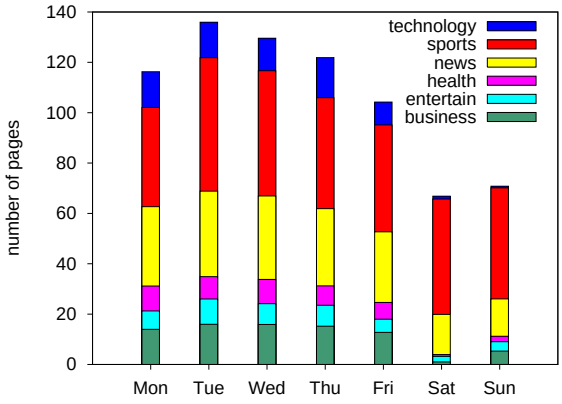


Figure 1: Average number of pages created per day.

not vary significantly across weeks and across categories. Their standard deviations were typically an order of magnitude lower than the averages and the corresponding confidence intervals were very small.

The analysis of the behavior of the creations across the six categories of pages has shown, as expected, that sports pages represented the large majority of the new pages uploaded to the site over the week-ends (i.e., 68.6% on Saturdays and 62.3% on Sundays), whereas almost all business and technology pages were uploaded during weekdays; for example, on average some 13 technology pages per day were uploaded on weekdays and one page over week-end days.

Page updates were another key aspect considered to study the evolution of the Web site. Specifically, our analysis focused on the actual content of the pages, thus ignoring all kinds of banners and navigation bars appearing in every page. The HTML files downloaded at each snapshot of the site were parsed to extract the content of each page and assess whether it was modified since the previous download.

As a result of this analysis, during our monitoring interval, we recorded 21,918 updates, that is, 164.8 updates per day and 1.55 updates per page. These updates involved about 70 different Web pages every day. Figure 3 plots the distribution of the number of updates. Since the distribution was characterized by a very long tail (one page received as many as 271 updates), we restricted the diagram to the range $[0, 20]$, that accounted for the 81.4% of the updates. As shown by the figure, the updates were not evenly distributed across pages. For the large majority of the pages, once a page was created, it was likely that it did go through either few updates or no updates at all. In particular, 7,421 pages, that is, 52.39%, did not

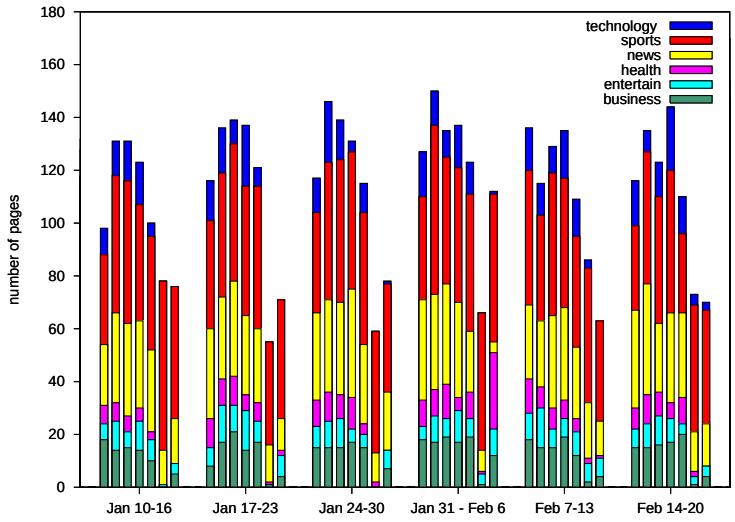


Figure 2: Temporal evolution of the number of pages created per day.

receive any update, and 3,042, that is, 21.48%, were modified only once. The average number of updates of the 6,744 pages modified at least once, was equal to 3.25. It is interesting to outline that 4,082 updates (i.e., 18.6% of the updates) were distributed across 99 pages only, with an average of 41 updates per page.

We then refined the analysis and studied the overall behavior of the site in terms of changes, i.e., creations of new pages and updates of the content of existing pages. Figure 4 shows a snapshot of the daily evolution of the site taken over the first eight weeks of our monitoring interval. The diagram plots for each day, starting Monday November 15, 2004, the number of changes occurred to the site. By looking at the figure, we can easily identify weekly patterns. Many more changes occurred to the site during weekdays than during week-end days, and, in particular, on Saturdays. The site was also characterized by very light activities during holidays. This phenomenon could be easily recognized during the Thanksgiving week (that started on November 22, 2004). The number of changes to the site on Thanksgiving day was equal to 120, compared to an average of 294 changes observed during the entire monitoring interval on Thursdays.

The analysis of the breakdown of the changes into creations and updates has shown that the new pages uploaded daily to the site accounted for the 39.36% of the total changes. This fraction was typically higher on weekdays (i.e., 40.61%) than on week-end days (i.e., 36.24%).

To further investigate the variability of the evolution of the site across the 19

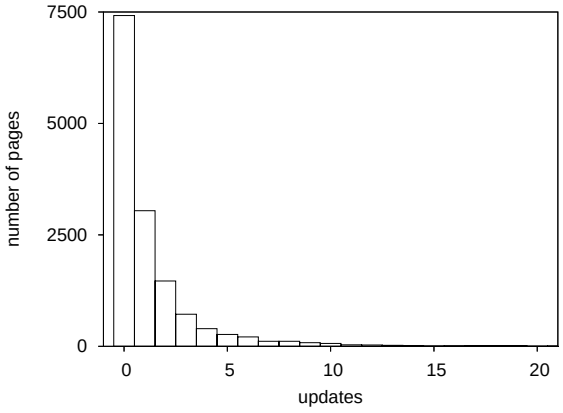


Figure 3: Distribution of the number of updates of the pages.

weeks, we represented the changes, that is, creations and updates, by means of a box-and-whisker plot. The diagram shown in Figure 5 plots the median and two measures of dispersion, namely, the range and inter–quartile range of the number of changes for each day of the week. The upper and lower boundaries of each box correspond to the 25th and 75th percentiles of the distribution of the changes, each computed over 19 days. The solid line within each box represents the median. We can see once more that changes to the site varied significantly as a function of the day of the week. Tuesdays, with a range and an inter–quartile range equal to 139 and 47, respectively, were characterized by the lowest variability. The variability was very high on Fridays, with the range equal to 308 and the median equal to 270.

The evolution of the site was also studied by analyzing the rate of changes, that is, the number of changes to the site per hour, over a 24 hours period. The change rate measured for one day, that is, Tuesday December 22, 2004, is shown in Figure 6 (dotted pattern). Despite the very high peak at midnight, the change rate was characterized by a typical diurnal pattern with more changes during day time than during night time. Specifically, the midnight peak was followed by a steady decrease of the change rate until 6am. The pattern was then characterized by a continuous increase with some fluctuations in the afternoon. Let us remark that the times plotted in the diagram refer to the Eastern time zone, as reported by the MSNBC Web site.

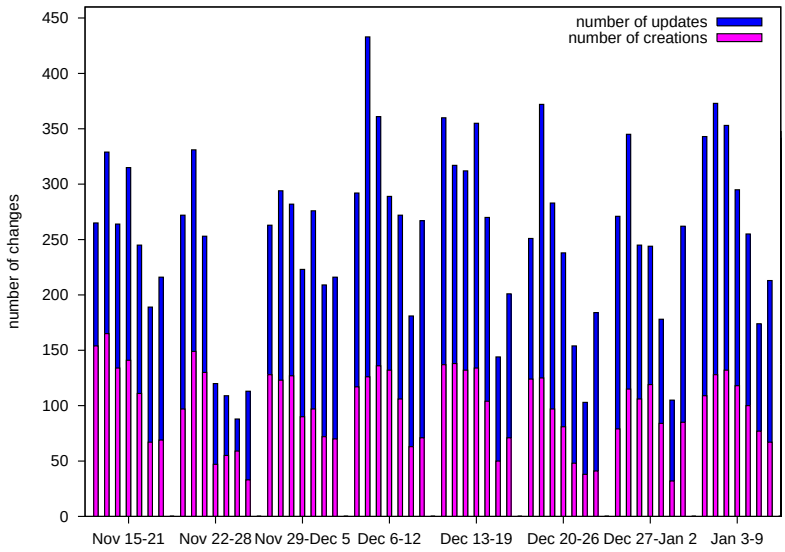


Figure 4: Snapshot of the daily evolution of the site over eight weeks.

# 4 Models of the change rate

In this section, we will present the models that characterize the dynamics of the MSNBC news Web site and its daily patterns. These models will be obtained by applying numerical fitting techniques to the hourly change rates measured on the site for each day. The use of clustering techniques will then allow us to discover similarities in the models of the change rates of different days.

Note that the models will specifically refer to the MSNBC Web site. Nevertheless, the methodological approaches proposed in what follows are very general and can be applied for modeling the dynamics of other Web sites.

To study the evolution of the MSNBC Web site, we applied numerical fitting techniques to the change rates. Indeed, as already pointed out, we did not make any statistical assumption on the underlying processes. Moreover, despite what found in other contexts, the Kolmogorov-Smirnov test, applied to assess whether the experimental data came from a Poisson model, failed.

As we were studying rate functions, we searched our models among the exponential polynomials of the following form:

$$\hat{y}(\tilde{t}) = \exp\left(\sum_{j=0}^{n} c_j \times \tilde{t}^j\right)$$

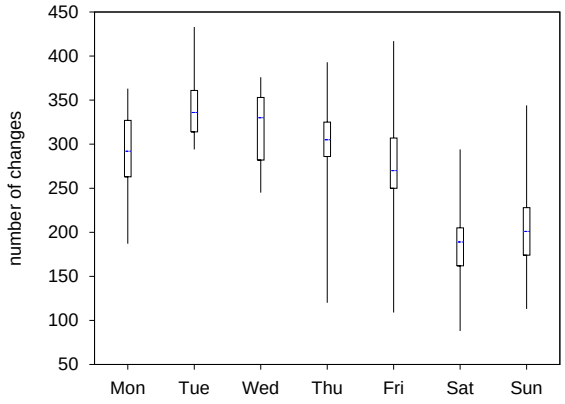


Figure 5: Box-and-whisker diagram of the changes to the site over the monitoring interval of 19 weeks.

where $n$ denotes the degree of the polynomial, $c_j,\ j = 0, \ldots, n$, are the unknown parameters to be identified by the fitting techniques, and $\tilde{t}$ represents the scaled time. Indeed, to prevent instability problems in the identification of the parameters, we scaled the time $t$ that described the hourly rate from the range $[0, 23]$ to the range $[-0.5, 0.5]$, that is:

$$\tilde{t}_i = \frac{t_i - \frac{(t_{\max} - t_{\min})}{2}}{t_{\max} - t_{\min}}, \quad i = 0, \ldots, 23.$$

The identification of the parameters that make a function best fit the experimental data relied on the application of the least squares based on the Levenberg-Marquardt method [4]. For the rate of each day, we generated exponential polynomials of degree $n$ ranging from 2 to 11.

To assess the goodness of fit between the measured data and the fitted model, the parameters of the functions identified for each day were analyzed for statistical significance by applying various tests [3, 22]. We computed the residuals, that is, the difference between the experimental data $y(\tilde{t}_i)$ and the corresponding fitted value $\hat{y}(\tilde{t}_i)$ predicted by the least squares, and we studied their signs and the corresponding run length. Moreover, we computed the adjusted coefficient of determination $R^2$, that is, the ratio of the variation explained by the model to the total variation. The tests of the goodness of fit were significant for all the fitted models of degree three and higher.

The solid curve of Figure 6 is an example of the model that fits the change rate measured on December 22, 2004. The curve represents an exponential polynomial

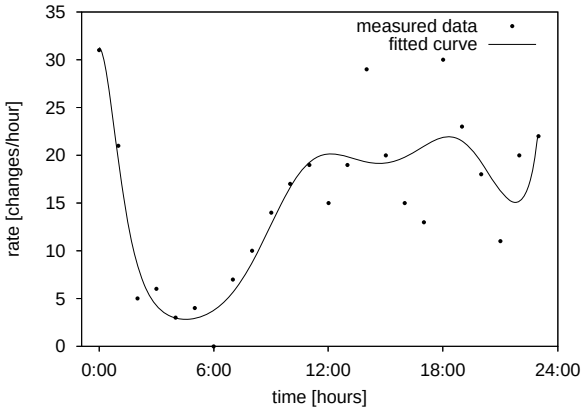


Figure 6: Hourly change rate of December 22, 2004.

of degree seven, whose parameters are presented in Table 1. All the statistical tests showed that the measured data was well described by the fitted curve. Specifically, the value of the adjusted coefficient of determination for this curve was equal to 0.783, that is, the fitted curve explained 78.3% of the variation of the measured data. Moreover, the residuals were randomly distributed above and below zero and did not exhibit any particular pattern. The corresponding number of runs was equal to 13.

| $c_0$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ |
|---|---|---|---|---|---|---|---|
| 2.817 | 4.298 | -14.875 | -40.536 | 73.660 | 281.049 | -91.373 | -768.307 |

Table 1: Parameters of the exponential polynomial function of degree seven that fits the change rate measured on December 22, 2004.

Since exponential polynomial functions could become ill conditioned as their degree increased, we addressed the problem of identifying the simplest model that fitted the data adequately. Specifically, we compared pairs of models to assess whether a more complex model, that is, a model with an additional parameter, outweighed the cost of its complexity. We performed an $F$ test [11] based on the following equation:

$$F_n = \frac{(SSE(n) - SSE(n+1))/1}{SSE(n+1)/(M-n-2)} \tag{1}$$

where $SSE(n)$ and $SSE(n+1)$ are the sums of squares of the residuals computed

for exponential polynomial functions of degree $n$ and $n+1$, respectively, and $M$ denotes the number of data points (i.e., 24) used by the fitting.

To test the null hypothesis $H_0$ that the exponential polynomial function of degree $n+1$ did not fit the data significantly better than the function of degree $n$, we compared the ratio $F_n$ to the tabulated value $F(\alpha, \nu_1, \nu_2)$, with $\alpha = 0.05$, $\nu_1 = 1$ and $\nu_2 = M - n - 2$. We accepted the null hypothesis $H_0$ at the 5% level of significance, if the computed ratio was lower than the corresponding tabulated value, otherwise we rejected the null hypothesis and accepted the model of degree $n+1$.

Table 2 presents the results of the $F$ test applied to the exponential polynomials fitting the data points shown in Fig. 6. The ratios $F_n$, computed according to Eq. (1), and the corresponding tabular values computed at the 5% level of significance, refer to functions of degree $n$ ranging from 3 to 10. Let us remark that the function of degree two failed the goodness of fit tests. As can be seen, the null

| $n$ | $F_n$ | $F_{\alpha=0.05}$ | $\nu_1, \nu_2$ |
|---|---|---|---|
| 3 | 15.3885 | 4.3807 | 1, 19 |
| 4 | 0.1940 | 4.4139 | 1, 18 |
| 5 | 0.1714 | 4.4513 | 1, 17 |
| 6 | 4.9554 | 4.4940 | 1, 16 |
| 7 | 0.0878 | 4.5431 | 1, 15 |
| 8 | 0.1508 | 4.6001 | 1, 14 |
| 9 | 0.9012 | 4.6672 | 1, 13 |
| 10 | 6.0250 | 4.7472 | 1, 12 |

Table 2: $F$ test performed on the exponential polynomial functions fitting the change rate measured on December 22, 2004.

hypothesis was accepted for the functions of degree 4, 5, 7, 8 and 9, whereas it was rejected for the functions of degree 3, 6 and 10. Therefore, the function of degree seven, that is, a model described by eight parameters, could represent a good compromise between simplicity and accuracy.

An extensive application of the $F$ test to all measured data has shown that exponential polynomials of degree up to seven adequately fitted the large majority of the change rates analyzed in our study. Hence, without loss of generality, to describe the change rates of all days, we used exponential polynomials of degree seven, although the rates of some days could be adequately modeled with fewer parameters.

To discover similarities in the models of the change rates of different days, we applied clustering techniques [15] to a sample of the exponential polynomial

functions of degree seven. Note that this approach is particularly suitable for the analysis of the dynamics of Web sites in that it detects the variability of the models of different days and helps in assessing their representativeness. Specifically, we focused on the 56 models of the first eight weeks of our monitoring interval. Each of these functions, described by the eight parameters identified by the fitting techniques, was represented as a point in an eight-dimensional space. The $k$-means clustering algorithm classified the functions in four groups. Two groups (cluster 1 and cluster 2) contained approximately one third (i.e., 17) and two thirds (i.e., 36) of the models, respectively, and the two remaining groups (cluster 3 and cluster 4) contained only three models in total, two in one group and one in the other group. It is interesting to point out that these three models were set apart by clustering as outliers because their parameters were very different from the parameters of the functions that fitted the remaining change rates of the site.

As representatives of the two main clusters, we used their centroids, that is, their geometric centers. Figure 7 shows the exponential polynomial functions whose parameters correspond to the geometric centers of these clusters. By look-

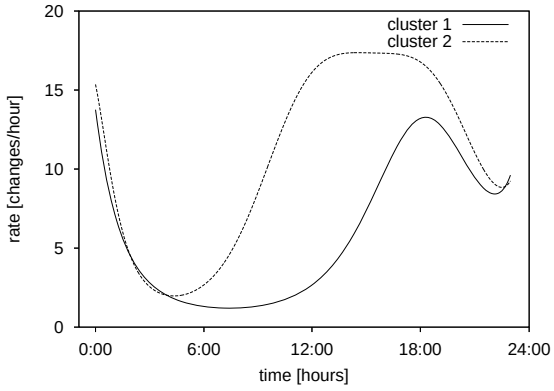


Figure 7: Exponential polynomial functions corresponding to the geometric centers of the two main clusters.

ing at the figure, we can notice that the two functions were quite different. In particular, the function corresponding to cluster 1 was characterized by a very low rate during the morning, followed by a slowly increasing rate around noon and a peak around 6pm. On the contrary, the function corresponding to cluster 2 was characterized by a much higher rate throughout the 24 hours.

By analyzing the composition of each group, we discovered that more than 88% of the components of cluster 1 were the models of change rates measured on Saturdays and Sundays, whereas all the components of cluster 2, but one, re-

ferred to the models measured on weekdays. We could then conclude that cluster 1 and cluster 2 summarized the evolution of the Web site over week-end days and weekdays, respectively.

The spread of the models within each of the two main clusters is shown in Figure 8. The envelopes represent the upper and lower bounds computed from the models of each cluster. We can notice that the envelopes follow rather closely the pattern of the functions corresponding to the cluster centroids, with some exceptions at the beginning and at the end of the day.

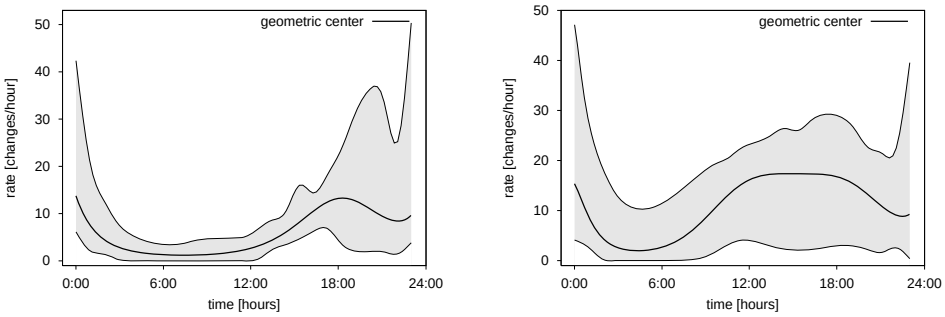


Figure 8: Envelopes of cluster 1 (a) and cluster 2 (b).

To assess the representativeness and the persistence of the daily patterns identified by clustering, we analyzed the exponential polynomial functions of degree seven that described the change rates of the remaining 77 days not considered in the sample used for clustering. In particular, we evaluated the similarity of these models with the groups previously identified, by testing whether they could be assigned to any of these groups. For consistency with the $k$-means clustering algorithm, the similarity criterion used for this evaluation was the Euclidean distance computed in an eight-dimensional space. The procedure applied for this assessment can be summarized as follows:

1. computation of the Euclidean distance $d_{ij}$ between model $i,\ i = 1, \ldots, 77$ and the centroid of cluster $j,\ j = 1, 2, 3, 4$;
2. computation of $\tilde{d}_{ij} = d_{ij} - D_j^{\max}$, $D_j^{\max}$ being the maximum distance of the models of cluster $j$ from the corresponding centroid;
3. assignment of model $i$ to cluster $l$, provided that $\tilde{d}_{il} \leq 0$ and $\tilde{d}_{il} = \min_j \tilde{d}_{ij}$.

An example of one of the functions assigned to cluster 2 according to this procedure and the envelope of the cluster itself is shown in Figure 9. We can notice that the

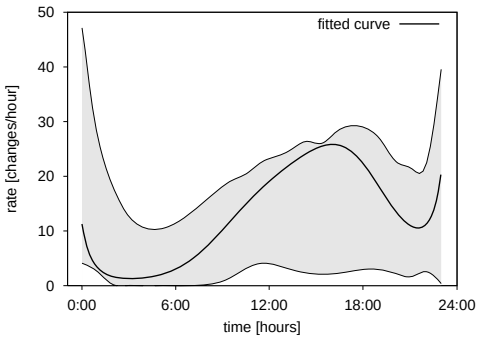


Figure 9: Envelope of cluster 2 and fitted curve assigned to this cluster.

curve accurately follows the pattern of the corresponding envelope.

An extensive application of the assignment procedure outlined above has shown that all models but three, could be assigned to one of the clusters. In particular, 17 models (i.e., 22.08%) were assigned to cluster 1 and the remaining 57 (i.e., 74.03%) to cluster 2, whereas none was assigned to any of the two outliers clusters. Hence, all these results confirmed that the two main clusters could precisely capture the evolution of the MSNBC Web site during the time period of 19 weeks.

To assess the representativeness of the models specifically derived for the MSNBC news Web site, we then refined our investigation and studied the evolution of other news Web sites. In particular, we analyzed the measurements collected on three news Web sites, namely, CBSNews [8], CNN [10], Reuters [20] Web sites, for a period of four weeks starting mid January 2008. Even though the behavior of these sites was slightly different from the behavior of the MSNBC Web site, in terms, for example, of fractions of pages created and updated per day and extent of each update, their evolution did not differ significantly. Their change rates could be accurately modeled by exponential polynomial functions of degree seven. Figure 10 shows an example of the model that describe the change rate of the Reuters Web site on January 31, 2008.

The application of the assignment procedure previously described to the models of the change rates of these three news Web sites has shown that most of them could be ascribed to one of the two main clusters identified from the analysis of the MSNBC site, with a clear subdivision between weekdays and week-end days. Thus, these results confirmed that the applicability of our methodological approach and more generally the representativeness of the models obtained for the MSNBC

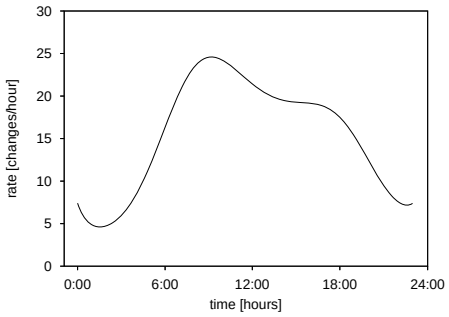


Figure 10: Model of the change rate of the Reuters news Web site on January 31, 2008.

Web site. Dynamics of news Web sites was rather stable and did not change significantly from site to site and even over time.

# 5 Content updates

The study of the evolution of the site was complemented by considering another dimension, that is, the extent of the updates made to individual pages. This section presents the methodological approach and the metrics proposed to quantify the extent of the modifications involved in each update and the impact of these metrics on the models of the change rates.

As already pointed out, we focused on the content extracted from the HTML files downloaded during our monitoring interval. We considered the 6,744 pages that received at least one update after they have been uploaded to the site and we studied to what extent each update modified the content of the page itself. Note that in what follows, the terms page and content are used interchangeably.

Not all updates were equally extensive: some involved minor typographical modifications, whereas others involved some major rewriting of the pages. As a first observation, it is interesting to outline that the successive updates to each of these 6,744 pages did not result in a significant variation of their size. Most of the updates (i.e., 66.1%) resulted in a page expansion, whereas 26.8% of the updates resulted in a page shrinkage, and the remaining 7.1% did not produce any change to the page size. On average successive updates increased or decreased the size of each page by approximately the same amount of bytes, that is, some 578 bytes.

Let us remark that the mean size of a page when it was first uploaded to the site, was equal to 3,531 bytes. Moreover, by comparing the initial size of the pages with the size of their last update, we noticed that the size of 5,107 pages (i.e., 71.73%)

increased approximately of one third, that is, 1,145 bytes, whereas the size of 1,322 pages (i.e., 19.6%) decreased of about 18%, namely, 637 bytes. For the remaining 315 pages the size did not change. The large majority of these pages received one update only.

To quantify the extent of modifications involved in each update as to assess, for example, whether the updated page was worth the cost of a download, we derived a more detailed quantitative description of the updates. In particular, we compared the successive updates of each page by computing their corresponding character-level edit distance [14]. Given two strings of characters, $X = X_1X_2 \ldots X_p$ and $Y = Y_1Y_2 \ldots Y_q$, of length $p$ and $q$, respectively, each representing the characters of a page, the edit distance $ed(X, Y)$ is defined as the minimum number of operations (i.e., insertions, deletions, substitutions) that transform $X$ into $Y$, that is:

$$ed(X, Y) = \min\{\gamma(S)|\ S \text{ is an edit transformation of } X \text{ into } Y\ \}$$

$\gamma(S)$ being the total cost of the operations (assumed in our case equal to one for every operation) required by the transformation $S$. Moreover, as the size of pages varied, to make the edit distance computed for each page and across pages comparable, we normalized $ed(X, Y)$ according to the longest transformation $S$.

The average normalized edit distance computed for the 21,918 updates detected in our monitoring interval was equal to 0.2134. This value corresponded to 972 edit operations, involving one character each. By looking at the corresponding distribution (see Figure 11), we can notice that it was highly skewed towards zero, that is, a good number of updates involved very few edit operations. In particular, the normalized edit distance of 5,128 updates, i.e., 23.4% of the updates, was smaller than 0.01, namely, these updates involved at most 16 operations. In addition, the normalized edit distance of some 75% of these updates was smaller than 0.005. Furthermore, it is interesting to point out that the normalized edit distance of about 15% of the pages (i.e., 1,056) was always smaller than 0.01. This means that these pages were always modified to a very limited extent.

Apart from these updates that could be considered as minor corrections rather than actual modifications of the pages, we have detected some major updates that involved as many as 53,232 operations. Let us remark that some pages were as large as 84,375 bytes, even though the size of most pages was much smaller and typically characterized by little variability.

As a further step towards a more detailed characterization of the evolution of the site, we used the outcomes of the analysis of the normalized edit distance to adapt the models of the daily change rates (see Sect. 4) as to take out the effects of the updates that could be considered as less extensive. Indeed, the reduction of the number of updates to be considered by services, such as, news aggregators and

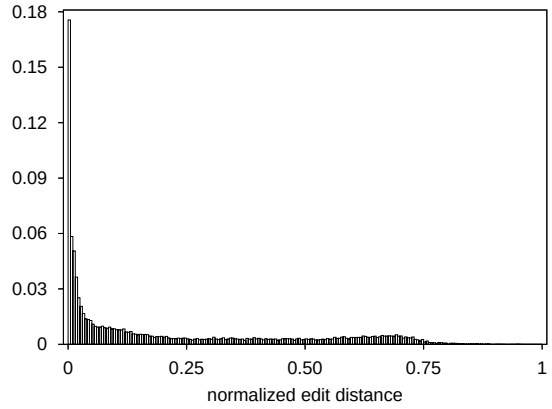


Figure 11: Distribution of the normalized edit distance of the page updates.

RSS feeds, could represent an important gain.

For this purpose, we applied two correction factors to the functions previously identified: a factor $a$ that took into account the weight of updates and a factor $b$ that took into account the weight of the "most extensive" updates. Indeed, the change rates included both creations of new pages and updates of existing pages, nevertheless, this adaptation process was not affected by page creations.

The factor $a$ was then obtained as the fraction of the updates over the overall number of changes to the site. The factor $b$ was obtained as the fraction of updates whose normalized edit distance was larger than a predefined threshold $\beta$ to be chosen according to the objective of the study. The resulting function $\hat{y}^*(\tilde{t})$ is given by:

$$\hat{y}^*(\tilde{t}) = (1 - a) \times \hat{y}(\tilde{t}) + a \times b \times \hat{y}(\tilde{t})$$

Figure 12 shows two examples of the measured rates and the corresponding adjusted models $\hat{y}^*(\tilde{t})$ both filtered to take out the updates identified as "not extensive". In these examples, the threshold $\beta$ was set to 0.01. The factors $a$ and $b$ were equal to 0.6064 and 0.766, respectively. These two factors were computed over the entire monitoring interval of 19 weeks. Figure 12 (a) plots the change rate of November 30, 2004, whereas Figure 12 (b) the change rate of December 22, 2004. As can be seen, both models adapted quite well to the corresponding filtered experimental data. Let us remark that the unfiltered data and the original model of December 22, 2004 are shown in Fig. 6.

The metric used to assess the accuracy of the proposed approach was the mean squared error computed for the adjusted model $\hat{y}^*(\tilde{t})$ with respect to the filtered experimental data. The values of this metric for the models shown in Fig. 12 were

equal to 15.34 and 18.96, respectively. These values were even smaller than the corresponding values computed by the fitting techniques for the model $\hat{y}(\tilde{t})$ with respect to the actual measured data (equal to 15.61 and 23, respectively).

From the application of this approach to all the change rates measured and modeled during our monitoring interval, we could observe that that the accuracy of most of the adjusted models was very good. Moreover, by testing various values of the threshold $\beta$, we could conclude that the models obtained with $\beta = 0.01$ were very precise, in that, despite the filtering of the less extensive updates, the patterns and dynamics of the changes to the site were accurately captured.

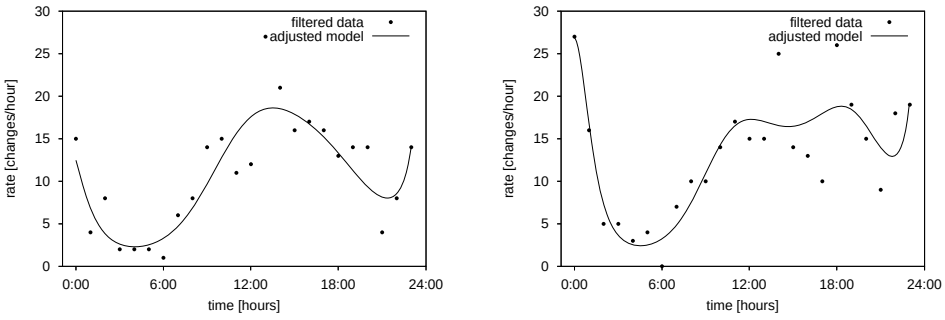


Figure 12: Filtered data and adjusted model of the change rates of two days.

# 6 Conclusions

The content of Web sites evolves continuously as a result of uploads of new pages and updates of existing pages. We studied the evolution of the MSNBC news Web site by considering how often and to what extent its content changed. Our preliminary findings have shown that, as expected from a news Web site, the behavior was different on weekdays and on week-end days. In particular, the site was more dynamic on weekdays than on week-end days. In addition, we noticed that on week-end days there was very little activity in the earlier hours of the day and most of the changes occurred late in the afternoon and in the evening. These patterns were rather stable and did not vary significantly over a monitoring interval of several months. Finally, a good fraction of the pages uploaded to the site either never changed or changed to a rather limited extent.

We then proposed a methodology that allows the identification of models able to reproduce the dynamics of the site in terms of page creations and page updates and to take into account how extensive the updates are, that is, whether they are worth the cost of a download. The proposed methodology can be easily automated in that it basically requires the application of fitting techniques for the identification of the parameters of the analytical models describing the evolution of the Web site and the use to clustering techniques to classify these models.

The models of the change rates could be used to control the crawling frequency of search engines and to develop delivery strategies and infrastructures that maximize the freshness of the information without penalizing too much the resources. A reduction of the number of downloads requested to keep a given level of freshness is an important goal for many of the new emerging services.

Even though the results of our study referred specifically to the characterization of the MSNBC news Web site, we have noticed that these models could accurately describe the dynamics of other news Web sites. Moreover, we believe that our findings are rather general and the methodological approach proposed in this paper can be applied to model the evolution of any type of Web content.

As a future work, we plan to focus the analysis on the novelty of the information available on Web sites used for social networking. These sites are characterized by the presence of multiple authors posting and sharing information. Hence, it is important to study the persistence of the information and to understand the role of the underlying infrastructure.

## Acknowledgments

The authors would like to thank the anonymous reviewers for the valuable comments that have helped in improving the paper. A special thank goes to Clara Parisi for her continuous support in the various phases of the analysis.

## References

[1] R. Baeza-Yates, C. Castillo, and E.N. Efthimiadis. Characterization of National Web Domains. *ACM Transactions of Internet Technology*, 7(2):Article No. 9, 2007.

[2] R. Baeza-Yates and B. Poblete. Dynamics of the Chilean Web Structure. *Computer Networks*, 50(10):1464–1473, 2006.

[3] D.M. Bates and D.G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, 1998.

[4] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.

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

[6] B.E. Brewington and G. Cybenko. Keeping Up with the Changing Web. *IEEE Computer*, 33(5):52–58, 2000.

[7] M. Calzarossa and D. Tessera. Models of Dynamic Web Contents. In *Methodologies, Techniques and Tools for Performance Evaluation of Complex Systems*, pages 26–33. IEEE Computer Society Press, 2005.

[8] CBSNews Web site. http://www.cbsnews.com.

[9] J. Cho and H. Garcia-Molina. Estimating Frequency of Change. *ACM Transactions on Internet Technology*, 3(3):256–290, 2003.

[10] CNN Web site. http://www.cnn.com.

[11] N.R. Draper and H. Smith. *Applied Regression Analysis - Third Edition*. Wiley, 1998.

[12] 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.

[13] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing Personalized Newsfeeds Via Analysis of Information Novelty. In *Proc. of the 13th ACM International Conference on World Wide Web WWW'04*, pages 482–490, 2004.

[14] D.M. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1999.

[15] J.A. Hartigan. *Clustering Algorithms*. Wiley, 1975.

[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*, 50:1430–1447, 2006.

[17] L. Lim, M. Wang, S. Padmanabhan, J.S. Vitter, and R. Agarwal. Efficient Update of Indexes for Dynamically Changing Web Documents. *World Wide Web*, 10(1):37–69, 2007.

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

[19] 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 International Conference on World Wide Web WWW'04*, pages 1–12, 2004.

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

[21] K.M. Risvik and R. Michelsen. Search Engines and Web Dynamics. *Computer Networks*, 39(3):289–302, 2002.

[22] K.S. Trivedi. *Probability and Statistics with Reliability, Queueing and Computer Science Applications – Second Edition*. Wiley, 2002.