Models of Dynamic Web Content*

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

Web pages are created, modified and removed at unspecified times by their owners. The frequency and extent of changes to Web pages vary across sites and across pages within sites. These changes influence the quality of the information retrieved from the Web and to some extent the delays experienced by the users. Our study focuses on the analysis of a news Web site with the objective of modeling the dynamic behavior of its content. We identify weekly and daily patterns in the page creation and update processes and we analyze the degree of change of the content of the pages. The study shows that many new pages were created every day, whereas existing pages but front pages, were updated few times and typically to a limited extent.

1. Introduction

The large amount and variety of information available on the Web open many challenging issues. It is necessary to locate useful and up–to–date information and to deliver it to the users with minimal delays. Scheduling strategies in Web crawling are used to improve the quality of the information retrieved from the Web. Content delivery strategies are used to distribute and replicate the information with the objective of improving the performance perceived by the users. All these strategies are influenced by the dynamic nature of the Web. Web pages are created, modified and removed at unspecified times by their owners. The frequency and extent of changes to Web pages vary across sites and across pages within sites. For example, pages that display data such as, headline news, weather reports, stock activity reports, change frequently, whereas pages that display data such as, legislations, historical information, seldom change. To provide users with up–to–date information without overprovisioning, it is then necessary to understand how often new pages are created and how often and to what extent existing pages are modified.

The dynamic nature of the Web has been studied since late nineties (see, e.g., [1, 2, 3, 6, 10, 12]). These studies focused on different aspects of the evolution of Web pages. In [4] a large number of Web pages is analyzed to experimentally derive frequency estimators that can be used to improve the effectiveness of Web crawling and Web caching. Models that capture the characteristics of dynamic Web content are presented in [14]. These models are derived from the analysis of the content of six representative sites. The evolution of several million of Web pages is studied in [7]. The paper analyzes the degree of change of these pages and identifies the attributes that are correlated with their change intensity. The content of the pages is compared by applying a similarity metric based on syntactic document sketches. The study shows that the rate of change varies across top–level domains and past changes to a page are a good predictor of future changes. Moreover, larger pages tend to change more often and to a larger extent than smaller pages. An analysis of the human perception of changes to Web pages is presented in [8]. The study considers the relevance of three categories of changes, namely, content, presentation and structure of Web pages. The results show that the perception of a content change is highly correlated with the perception of an overall change. In [5] changes to Web pages are studied in the framework of supporting maintenance of distributed collections of Web materials. Context based methods are applied to compare Web pages.

This paper analyzes the dynamic behavior of the content of a popular news Web site, i.e., the MSNBC web site [11], with the objective of identifying typical patterns and modeling the page creation and update processes. We derive models of the behavior of the pages by studying page update instances and page change frequency. Moreover, we classify pages according to their characteristics.

* This work was supported in part by the Italian Ministry of Education, Universities and Research (MIUR) in the framework of the FIRB–Perf project.

The changes to pages were characterized by two distinct behaviors, that is, pages were created and updated whenever new events happened and deterministically or quasi–deterministically to keep the site "alive". Changes varied significantly from page to page and some parts of a page changed more frequently than others. We have also observed that, even though most of the pages were created and modified during Eastern Time working hours, the site became quite dynamic, that is, new pages were created and existing pages updated, late in the evenings. Furthermore, the degree of change to a page ranged from minor typographical corrections to replacement of large chunks of its content, whereas the structure of a Web page changed very rarely.

This study has important performance implications in that our findings and models could be used to differentiate caching policies and to improve Web crawling and content distribution strategies by avoiding, for example, retransmissions of unchanged pages.

The paper is organized as follows. Section 2 describes the experimental setup used for the study. Section 3 focuses on the analysis of the Web pages and presents the models that describe their dynamic behavior. Finally, Section 4 concludes the paper by discussing the performance implications of the behavior of the site and by outlining future research directions.

2. Experimental setup

To study the dynamic behavior of the MSNBC news Web site, we actively monitored the site by periodically downloading the Web pages to track changes, that is, to determine if and to what extent pages have changed since last download. At each access we downloaded all HTML files that were on the site, including the new files uploaded to the site since last access.

Our monitoring focused on the HTML files classified in the MSNBC site under five major categories: news, business, sports, entertainment and health. In particular, we downloaded the so–called front pages, that is, the pages containing the headline news of the site, as well as the news pages, that is, the pages containing the actual text of the news. Moreover, among the front pages, we downloaded the main front page that corresponded to the home page of the site.

The choice of the granularity of the downloads was a challenging issue. The granularity had to be fine enough to capture all updates and coarse enough to minimize the time spent to download all files over the network. By monitoring the site at intervals of various lengths, we discovered that the front pages were characterized by more frequent updates than the news pages. As a result of this analysis, we chose to download front pages every 5 minutes, whereas it was enough to download news pages every 15 minutes.

As part of the experimental setup, we extracted the core of the news pages, that is, the actual body of the news. Moreover, we applied some filters based on HTML tags to extract fragments, e.g., top stories section, headline news sections, from the front pages. Indeed, front pages consisted of well identified fragments that seemed interesting to analyze separately. For these purposes, we did some post–processing of all the HTML files downloaded from the site.

Note that HTML files were usually rather small. On the average, the size of a news page was equal to 42Kbytes and did not vary significantly across pages belonging to different news categories and between successive instances of the same page. In the case of front pages, the size of the corresponding HTML files varied across categories from a minimum of 42Kbytes to a maximum of 70Kbytes, corresponding to the entertainment and sports categories, respectively.

The site was monitored for a period of 19 continuous weeks starting mid November 2004 and ending late March 2005. It is worth to point out that the number of front pages on the site was almost constant since they were part of the core structure of the site itself, whereas news pages were added daily to the site. The number of different front pages associated to the five news categories considered in our study was 76 and very few of them were added during our monitoring period. On the contrary, the number of news pages tended to become larger and larger. Hence, to avoid redundant transfers and an uncontrolled increase of the download time, we identified the pages that were "active", that is, still worth monitoring. For this purpose, we assumed that a news page was no longer active, if it was not updated during a five days interval. To identify active pages, we compared successive instances of the core of each page. Indeed, the size of the corresponding HTML file was not a good indicator of actual changes to the file itself due to the presence of contents, such as, advertisements, that typically changed at every access. On the average, at every access, i.e., every 15 minutes, we downloaded from the site 500 active news pages. Moreover, we accessed the site every 5 minutes to download 76 front pages.

3. Data analysis

To characterize the dynamic behavior of the content of the MSNBC Web site, we started with an exploratory analysis of its temporal evolution. More specifically, we focused on the dynamic properties of the creation and update processes of the Web pages. Note that we analyzed separately the news pages and the front pages because of their different scope within the site.

The total number of unique news pages collected during our monitoring interval was equal to 10,902. Among them,

the static pages, that is, pages whose core did not receive any update during their lifetime, was equal to 5,654 and accounted for 51.86% of the global number of unique pages. The remaining 5,248 pages were dynamic, that is, their core was modified at least once. Figure 1 shows the fraction of dynamic pages within each of the five categories of news considered in our study. The complement of each bar displays the fraction of static pages. As can be seen, the fraction of dynamic pages varied across categories. The sports category was characterized by the largest fraction (59.9%) of dynamic news, that is, 2,785 out of 4,649 sports pages were dynamic. The entertainment category was characterized by the largest fraction (65.97%) of static news, that is, 572 out of 867 pages. Note that the fraction of dynamic pages within the business category could have been larger. This is because we could not capture the summary report of stock indices displayed on the pages as it was implemented as a script embedded into the corresponding HTML file.

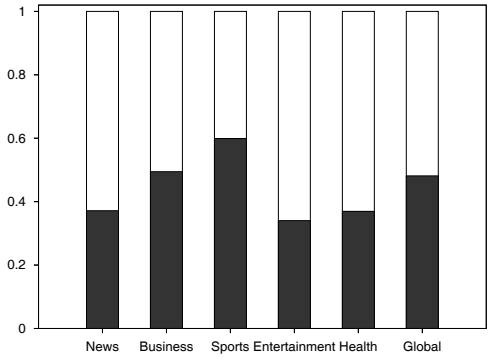

Figure 1. Fraction of dynamic pages within each news category.

Table 1 presents some descriptive statistics of the number of changes to dynamic pages. On the average, each page received 2.898 changes during its lifetime, even though there were pages in the sports category modified as many as 88 times.

Category	mean	std dev.	max	unique pages
News	2.702	3.672	38	1,192
Business	2.253	1.611	14	676
Sports	3.288	5.159	88	2,785
Entertainment	2.904	2.893	22	295
Health	1.766	1.407	13	300
Global	2.898	4.284	88	5,248

Table 1. Descriptive statistics of the number of changes to dynamic pages.

The mean time between two successive changes to a given page was equal to 208 minutes, with a standard deviation equal to 452. Moreover, about 73.5% of the changes occurred within the first three hours of the lifetime of the pages and 90% within the first eight hours. We have also observed that a few of the changes were not correlated to the occurrence of any specific event. This was the case, for example, of editorials that were usually updated in a sort of deterministic way.

The analysis of the temporal similarity of the content of page instances was based on the vector model used in the framework of information retrieval [13]. For this purpose, each page was represented by a vector, whose components are its unique words. Within each page we focused on the core of the corresponding news, that is, its text and HTML tags, thus including the URLs of all its embedded objects.

Let N denote the size of the vocabulary, that is, the global number of distinct words in all instances of a given page. The $j-$th instance of a page was then mapped into this $N-$dimensional space and represented by a vector $d_j = (w_{1,j}, w_{2,j}, ..., w_{N,j})$, where $w_{i,j}$ denotes the weight of word i in the $j-$th instance of the page. In our study, the weights corresponded to the raw frequency of the words, that is, the number of times each word appeared in the page. With this assumption, words that occurred frequently in a page were considered more important than infrequent words.

To assess the similarity between successive instances of a page, we used the cosine coefficient (see [9, 17]). According to this metrics, the similarity between instances d_j and d_k of a page is given by:

$$sim(d_j, d_k) = \frac{d_j \cdot d_k}{|d_j| \times |d_k|} = \frac{\sum_{i=1}^{N} w_{i,j} \times w_{i,k}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,k}^2}}$$

This metrics yields values in the range $[0, 1]$. Values close to 1 denote a high degree of similarity, i.e., the angle between the two vectors is close to 0 degrees.

To determine to what extent successive instances of a page changed, we computed the cosine coefficient between pairs of instances. The analysis has shown that the average cosine coefficient computed for all pages was equal to 0.96125 and corresponded to a 16 degrees angle. This means that successive instances were rather close to each other in terms of the frequency of their words. Moreover, despite previous studies, our results did not show any specific correlation between the degree of change and the size of the page. Indeed, the size of successive instances of a given page typically varied to a limited extent.

A refinement of this analysis was aimed to study to what extent a page changed during its lifetime. For this purpose, we compared the first instance of a page with all its successive instances and computed the corresponding cosine coefficients. Figure 2 shows the distribution of the cosine coefficient computed for all dynamic pages. Since more than 90%

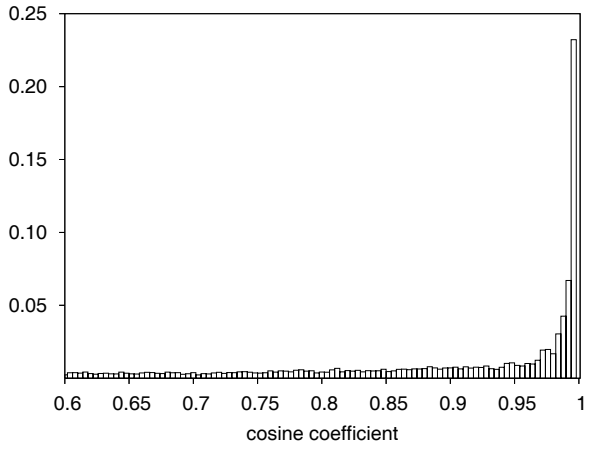

Figure 2. Distribution of the cosine coefficient of similarity for the dynamic pages.

of the values were greater than 0.6, the diagram plots the distribution in the range $[0.6, 1]$. The average cosine coefficient was equal to 0.866 and corresponded to a 30 degrees angle. The distribution shows that differences between the first instance of a page and its successive instances tended to become larger, even though they were not as large as expected. Indeed, there was a large fraction of pages (45.5%) that received one update only. The median of the distribution was equal to 0.943 and the angle was of 19.44 degrees.

To further investigate the dynamic behavior of the content of the site, we studied the evolution of new pages and the changes to existing pages as a function of the time of the day and of the day of the week. Note that for the sake of consistency with the MSNBC site, the times reported in what follows refer to the Eastern Time zone.

Figure 3 shows the weekly behavior of the number of pages created per hour. The diagram refers to all the pages created during the monitored period of 19 weeks. The granularity of the rate is one hour. The black area of the diagram shows the creation rate of static pages. The labels on the x axis are centered at noon of each day. New pages were created at a rate of 3.415 per hour. Among them, static pages were created at a rate of 1.771 per hour. By looking at the diagram, we can easily identify a daily pattern. Rates

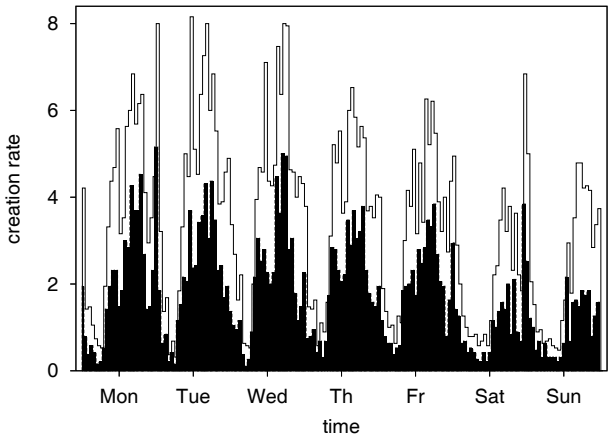

Figure 3. Weekly behavior of the page creation rate per hour.

were higher during working hours, even though there were a few high peaks around midnight. These peaks typically corresponded to sports events. Moreover, as expected, fewer pages were created over week–ends. The rate was equal to 2.272 pages per hour on Saturdays, whereas it was almost double, i.e., 4.386, on Tuesdays.

The daily pattern of the creation rate varied over the five categories of news considered in our study. Figure 4 shows the weekly creation rate of sports pages. Very few pages

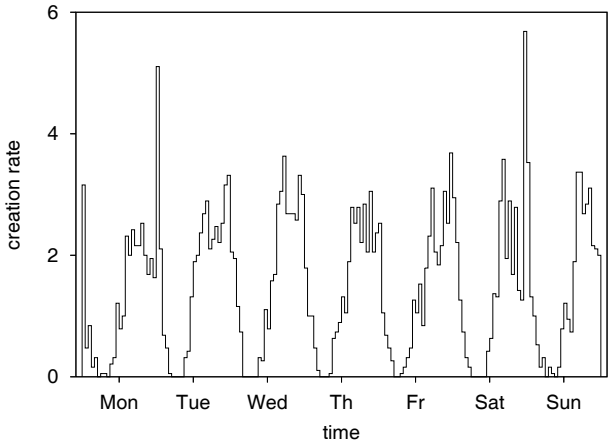

Figure 4. Weekly behavior of the creation rate of sports pages per hour.

were created early in the morning, whereas the rate started to increase steadily around noon. The rate was high until very late at night and exhibited higher peaks on Monday and Saturday nights. This diagram shows that the dynamic behavior of the news pages was highly dominated by the sports pages, which accounted for about 43% of the pages analyzed in our study.

To refine these results and get further insights in the dynamic behavior of the page creation process, we analyzed the hourly rate over different days. Table 2 reports a few statistics of the page creation rate over the various days of the week. The statistics refer to all static and dynamic pages

Day	mean	std dev.	unique pages
Monday	3.680	1.956	1,678
Tuesday	4.386	2.381	2,000
Wednesday	4.129	2.152	1,883
Thursday	3.675	1.793	1,676
Friday	3.379	1.671	1,541
Saturday	2.272	1.575	1,036
Sunday	2.386	1.576	1,088
Global	3.415	2.037	10,902

Table 2. Descriptive statistics of the page creation rate per hour.

created during the monitored period. The standard deviation was always smaller than the corresponding mean, that is, the creation rate across the 19 weeks was characterized by low variability.

To summarize the main statistical measures of the hourly creation rate we used the box–and–whisker diagrams as they are very useful in spotting differences in distributions. Figure 5 shows the diagram corresponding to the pages created on the 19 Wednesdays considered in our analysis. The diagram plots the median and two measures of dispersion, namely, the range and inter–quartile range. The upper and lower boundaries of the boxes correspond to the 25 and 75 percentiles. The solid line within each box represents the median. As can be seen, the median, inter–quartile range and range of the number of pages created between 8:00 and 9:00 were equal to 4, 3.5 and 6, respectively. Moreover, the 75 percentile of the hourly rate never exceeded 10.5. By looking at the height of the boxes, we discovered that the hourly distributions were rather narrow, that is, the number of pages created per hour across different weeks did not vary significantly. Our analysis has shown that this conclusion holds on most days, even though we detected a few large differences. For example, the range of the rate between 11:00 and 12:00 on Tuesdays was as large as 23. On

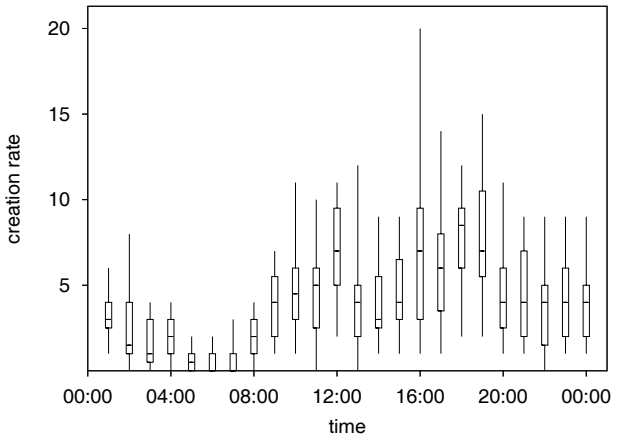

Figure 5. Box–and–whisker diagram of the hourly creation rate for all Wednesdays.

week–end days the distributions were even narrower. The inter–quartile range was always lower than 5.5.

Figure 6 shows the weekly behavior of the number of changes per hour to dynamic pages. The diagram refers

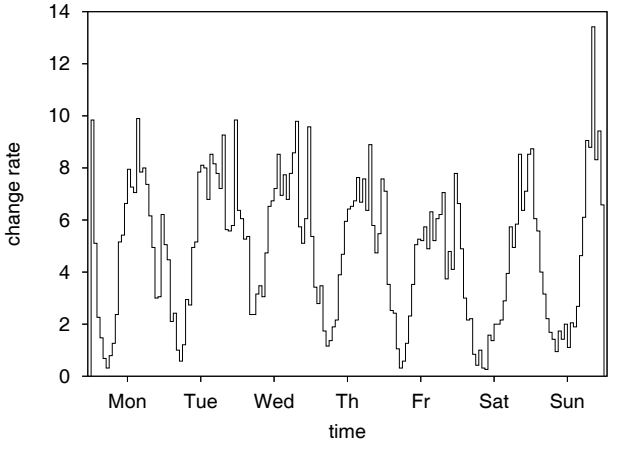

Figure 6. Weekly behavior of the change rate per hour of dynamic pages.

to all changes received by the pages. The labels on the x axis are centered at noon of the corresponding day. The rate of change is equal to 4.848, with a standard deviation of 2.707. The rate is higher on weekdays than on week–end days, even though the highest peak, that is, 13.42 changes

per hour, occurred late on Sunday evening. A refinement of the analysis has shown that the peak was due to sports events that caused a large number of changes to the corresponding pages. On Sundays the rate of change to sports pages was equal to 3.781 per hour, whereas the rate computed for the news pages is equal to 0.621. Note that the overall change rate was equal to 4.71 on Sundays, whereas it was about 22% higher on Wednesdays. This proves once more the dominant effect of the sports pages on the dynamic behavior of the content of the pages.

To obtain analytic representations of the dynamic behavior of the pages, we applied numerical fitting techniques based on the linear least squares method. We modeled the daily and weekly behaviors of the creation and change rates using polynomial functions. Before applying the fitting techniques, we smoothed the measured data to take out local fluctuations. The smoothing was based on either 3–term or 5–term moving average estimators. Moreover, to avoid instability problems in the computation of the coefficients of the polynomials, we scaled the independent variable, that is, the time of the day, in the range $[-0.5, 0.5]$. The reliability of the models was assessed by applying standard criteria, such as, coefficient of determination and F statistic (see [15, 16]). These criteria were also used to choose the optimal degree of the polynomial functions. For the various models, the optimal degree ranged between 5 and 16. In particular, the daily change rates were best described by polynomials of degree 5 or 6, whereas polynomials of degree 15 were used to represent the weekly change rates.

To study the similarities of the rates across different days and weeks, we applied clustering techniques. The rates were described by the coefficients identified by the least squares method. Figure 7 shows the polynomial function that models the change rate of the five weekdays. The function corresponds to the centroid, i.e., the geometric center, of one of the two clusters obtained by analyzing the change rates over the 19 weeks considered in our study. This cluster groups the rates of 15 weeks. The rates of the remaining 4 weeks were grouped in a second cluster as they were characterized by much higher values. Note that all rates were modeled by polynomials of degree 13.

As a final step to fully describe the dynamic behavior of the content of the MSNBC Web site, we analyzed the front pages of each of the five categories of news considered in our study, as well as the main front page, which corresponded to the home page of the site. In particular, we focused on their composition and rate of change. Moreover, due to their peculiarities, we were not interested in studying to what extent they changed whenever they were modified. We were rather interested in analyzing the fraction of fragments that changed between successive instances of a given front page.

As already pointed out, each front page typically consisted of distinct fragments. In particular, within each page, we could identify the so–called top stories section and a variable number of headline news sections containing links to other front pages and to news pages. The number of fragments varied across front pages and during the monitored interval. The sports front pages were characterized by the largest and most variable number of fragments. On the average, a sports front page consisted of 27 fragments, whereas the home page consisted of 16 fragments only. Within each page there were fragments that changed more frequently than others. Actually, some contents could be considered more important than others and consequently some changes more important than others. Our analysis has also shown that the large majority of the changes (92.8%) to the home page involved two fragments at most.

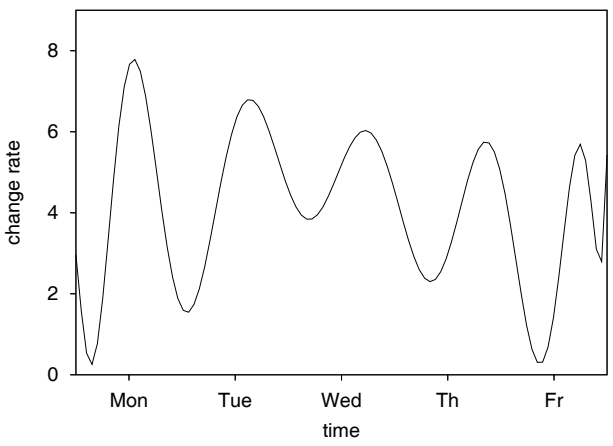

Figure 7. Model of the change rate corresponding to a cluster centroid.

Table 3 presents descriptive statistics of the time, expressed in minutes, between two successive changes to front pages. The home page was very dynamic. On the average, it was updated every 12.2 minutes. Moreover, 90% of its updates occurred within 25 minutes and 95% within 35 minutes.

Figure 8 shows the weekly behavior of the number of changes per hour to the home page. The labels on the x axis are centered at noon of each day. The rate was equal to 4.82 changes per hour. The standard deviation was much smaller than the corresponding mean, i.e., 1.76. It is interesting to point out that the hourly rate was always greater than zero.

As for the case of the news pages, we could recognize a daily pattern. The rate was lower on week–end days and higher on weekdays. The rate was equal to 3.91 changes per hour on Saturdays, whereas it was equal to 5.62 on Wednesdays. The top stories section was the fragment of the home

Category	mean	std dev.	percentiles 90	95
News	32.1	47.77	70	110
Business	26.2	51.58	55	97
Sports	15.3	40.90	25	35
Entertainment	43.7	82.29	92	195
Health	61.1	110.03	145	215
Home Page	12.2	11.83	25	35

Table 3. Descriptive statistics of the time (in minutes) between successive changes to front pages.

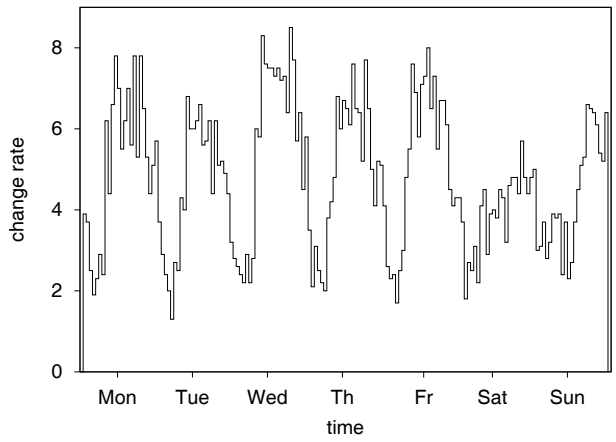

Figure 8. Weekly behavior of the change rate per hour of the MSNBC home page.

page with the highest change rate, that is, 1.38 changes per hour. For this fragment, the change rate was independent of the days, that is, there was no difference between weekdays and week–end days. Similar considerations hold for the sports front page. Their change rate was equal to 3.92 per hour and did not change significantly across days.

4. Conclusions

The highly dynamic nature of Web contents influences the quality of information retrieved from the Web and to some extent the delays experienced by the users. To provide users with up–to–the–minute information with minimal delays, it is important to implement caching and content replication policies that take into account the dynamic behavior of Web contents.

Our study focused on the analysis of the content of the MSNBC news Web site with the objective of identifying models able to describe the evolution of the site. The analysis has shown that many pages were created every day either because of the occurrence of new events and at some specific times of the day. Moreover, front pages changed very often and within them the top stories section was the most dynamic fragment. On the contrary, despite what expected, news pages were not so dynamic. Each news page was updated only few times and whenever it changed, it was typically modified to a rather limited extent. Let us remark that we did not analyzed the types of content, such as, advertisement, that typically changed at every access as a function of pre–defined algorithms or of the cookies set by the site.

We identified daily and weekly patterns in the page creation and update processes. The analytical models obtained by applying fitting techniques were able to capture and reproduce these patterns. Our findings and models have important performance implications in that they could be used to improve scheduling strategies of search engines and to refine data distribution policies.

As a future work, we plan to extent the analysis to further investigate the impact of the dynamic behavior of Web content on the performance perceived by the users. We also plan to use these models in the framework of content replication strategies.

Acknowledgment

The authors wish to thank Clara Parisi for her valuable help in setting up the experimental environment.

References

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

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

[3] L. Cherkasova and M. Karlsson. Dynamics and Evolution of Web Sites: Analysis, Metrics and Design Issues. In *Proc. of the 6th IEEE Symposium on Computers and Communications*, pages 64–71, 2001.

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

[5] Z. Dalal, S. Dash, P. Dave, L. Francisco-Revilla, R. Furuta, U. Karadkar, and F. Shipman. Managing Distributed Collections: Evaluating Web Page Changes, Movement, and Replacement. In *Proc. of the 4th ACM/IEEE Joint Conference on Digital Libraries*, pages 160–168, 2004.

[6] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of Change and other Metrics: A Live Study of the World Wide Web. In *Proc. of the USENIX Symposium on Internet Technologies and Systems*, 1997.

[7] 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.

[8] L. Francisco-Revilla, F. Shipman, R. Furuta, U. Karadkar, and A. Arora. Perception of Content, Structure, and Presentation Changes in Web–based Hypertext. In *Proc. of the 12th ACM Conference on Hypertext and Hypermedia*, pages 205–214, 2001.

[9] D. L. Lee, H. Chuang, and K. Seamons. Document Ranking and the Vector–Space Model. *IEEE Software*, 14(2):67–75, 1997.

[10] M. Mikhailov and C. E. Wills. Exploiting Object Relationships for Deterministic Web Object Management. In *Proc. of the 7th International Workshop on Web Content Caching and Distribution*, 2002.

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

[12] V. N. Padmanabhan and L. Qiu. The Content and Access Dynamics of a Busy Web Site: Findings and Implications. In *Proc. of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 111–123, 2000.

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

[14] 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.

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

[16] R. E. Walpole, R. H. Myers, and S. L. Myers. *Probability and Statistics for Engineers and Scientists*. Prentice Hall, 1998.

[17] J. Zobel and A. Moffat. Exploring the Similarity Space. *ACM SIGIR Forum*, 32(1):18–34, 1998.