CHAPTER 14

A STUDY OF THE DYNAMIC BEHAVIOR OF A WEB SITE

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

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

Interactive Web services make use of highly dynamic contents. To design efficient mechanisms for the replication and distribution of these contents and to improve the QoS perceived by the users, it is important to understand how often and to what extent contents change. This paper addresses these issues by studying the dynamic behavior of the contents of a popular news site. The contents are analyzed in terms of various metrics to assess their similarity. Groups of news with homogeneous characteristics are also identified.

1. Introduction

The increased pervasiveness of interactive services offered over Internet opens new performance challenges. These services are typically used by a large number of users who share various types of contents that are updated as a consequence of external events or of actions performed by the users themselves. The peaks of load on the servers and the peaks of traffic over the network can cause delays in delivering the requested contents and propagating the updates. These delays have a negative impact on the QoS perceived by the users who might even end up by accessing out of date contents.

To cope with QoS requirements without overprovisioning the systems,

solutions, such as, Content Distribution Networks, and peer-to-peer systems, have been adopted. These solutions allow the large scale replication of shared contents, hence, require efficient data distribution mechanisms.

The analysis of the update process of Web contents is the starting point of any study aimed at assessing the performance and scalability of the proposed solutions. These issues have been addressed from different perspectives.[1,2,3,7,8,15] A combination of empirical data and analytic modeling is used to estimate the change rate of Web pages.[2] The stochastic properties of the dynamic page update patterns and their interactions with the corresponding request patterns are addressed by Challenger et al.[4] Their study shows that the pages of highly dynamic sport sites are characterized by relatively large bursts of updates and periodic behavior. Cho and Garcia–Molina[7] introduce estimators to study the frequency of change to a data item (e.g., a Web page) in the presence of incomplete change history of the item itself. The process of creation and modification of HTML files of the MSNBC news site is addressed by Padmanabhan and Qiu.[14] Their study shows that files tend to change little when they are modified and modification events tend to concentrate on a small number of files. Metrics that consider the introduction of new files and their influence on the requests to the Web site are introduced by Cherkasova and Karlsson to study the dynamics and evolution of Web sites.[6] The impact of new files is also addressed in the framework of media sites.[5] Fetterly et al.[9] analyze the rate and degree of change of Web pages and the factors correlated with change intensity. The study shows that document size is strongly related to both frequency and degree of change. In particular, large documents change more often and more extensively than smaller documents.

This chapter studies the evolution of a news Web site with the objective of analyzing how often and to what extent the contents of Web pages change. We chose the MSNBC site as we consider it a good representative of the most popular news sites. Since we did not have access to any server log, we had to monitor the site. In this respect and many others, our study differs from a previous study on the same site.[14] In particular, we focused on the "core" of the HTML files and we studied the amount of change to each news between two successive downloads. Various metrics are used to assess the similarity of the successive versions of the news.

The chapter is organized as follows. Section 2 presents the experimental set up used for the data collection. Section 3 describes the results of our analysis. Finally, conclusions and future research directions are outlined in Section 4.

2. Data collection

Our study relies on data collected from the MSNBC news Web site[13]. Since we were interested in the actual contents of Web pages, we resorted to active monitoring and we downloaded the HTML files from the site. From these repeated accesses we could detect changes to each file and estimate the change frequency and the extent of each change. Let us remark that by change we mean any modification to the file.

Our monitoring adopted a conservative approach in that at each access we downloaded all the HTML files that were on the site, including the set of files newly introduced into the site since the last access. Note that we only downloaded the HTML files classified on the MSNBC site under five major categories, namely, news, business, sports, entertainment, health. On the average the number of new files downloaded over a 24 hours period is equal to 104.36, with a distribution characterized by low variability. Even though the number of new files can be as large as 156 and as small as 33, their standard deviation is equal to 29.93, that is, much lower than the corresponding mean.

To avoid an uncontrolled increase of the download time, we did some post–processing of the files as to identify the files still worth monitoring. In particular, we decided to filter out from the list of files to be downloaded, the files that were not active, that is, did not receive any update over a period of five consecutive days. Let us remark that the size of the HTML files was not a good indicator of actual updates due to the presence of various types of contents, e.g., advertisement, that typically change at every access. Hence, we did parse the files to identify changes to their "core". Note that at each access we downloaded on the average about 500 files.

News Web sites tend to update their contents periodically whenever something new happens. Hence, the granularity of the monitoring interval, that is, how often to download a particular file, was a crucial issue for the accuracy of our analysis. If a file changes once a hour, it might be unnecessary to download it every minute, whereas it might be insufficient to download it once a day. As a consequence, the granularity of the monitoring has to be fine enough as to capture all updates and coarse enough as to minimize the time spent to download all files over the network.

For this purpose, we monitored the MSNBC Web site over intervals of various lengths. Our analysis has shown that a granularity of 15 minutes is a good tradeoff between the frequency of changes and the download time. The measurements were then collected at regular intervals of 15 minutes

from the mid of November 2004 for a period of approximately 16 weeks.

To remove the fluctuations due to the warm–up effects of the monitoring process, we discarded the first week of measurements. Similarly, we discarded the "tail" of the measurements. The total number of HTML files successfully downloaded was about 5 millions. Among them, the number of unique files was equal to 9,209. Note that each file was downloaded at least 480 times, that is, we collected at least 480 "versions" of the same file.

3. Results

As a preliminary analysis, we focused on the overall characteristics of the HTML files. Table 1 presents some descriptive statistics of the size of the HTML files, with a breakdown into the five news categories considered in our study. As can be seen, files tend to be rather small, their average size is about 42Kbytes, and there is little variation among the categories. The table also shows the number of unique files downloaded for each category.

Table 1. Descriptive statistics of the size, expressed in bytes, of the HTML files.

Category	mean	standard dev.	min	max	unique files
News	40,812	4,321	13,895	83,980	3,393
Business	39,863	3,173	13,849	57,963	1,448
Sports	45,425	4,777	13,894	72,175	2,588
Entertainment	43,557	4,277	28,164	92,018	915
Health	39,636	4,028	14,295	59,942	865
Global	42,524	4,949	13,894	92,018	9,209

Since we were interested in analyzing the core of the news, that is, their actual contents, we did some preprocessing of each HTML file to extract the text of the news and the HTML tags making up its layout. Moreover, as we did not download the objects, e.g., images, videos, embedded in the HTML files, we considered as part of the core of the news the HTML tags and URLs of all embedded objects, but advertisement. Indeed, we have seen that tags and URLs are good indicators of the contents of the corresponding objects. In what follows, our analysis then focuses on the core of the news.

The number of dynamic news, that is, news that received at least one update over their monitoring period, is equal to 4,252, and accounts for about the 46.2% of the total number of unique files downloaded. The remaining 4,957 news can be classified as static, in that they did not receive any update during their monitoring period. Note that the number of dynamic news is not evenly distributed across the five categories. For example,

about 39% belong to the sports category, whereas less than 8% belong to the entertainment category.

Table 2 presents the statistics of the number of changes to the 4,252 dynamic news. As can be seen, this number varies across the categories. On the average, each news received 2.732 changes, even though there are news belonging to the sports category that received as many as 118 changes.

Table 2. Descriptive statistics of the number of changes to the dynamic news.

Category	mean	standard dev.	max	total number of changes
News	2.738	3.871	49	3,431
Business	2.398	2.905	48	1,698
Sports	3.077	4.575	118	5,084
Entertainment	2.650	2.918	22	832
Health	1.761	1.504	13	572
Global	2.732	3.859	118	11,617

The distribution of the intervals between two successive changes to a news is characterized by a mean value equal to 214.18 minutes and a standard deviation about four times larger. This distribution is highly positively skewed. Note that the time stamp associated to each news corresponds to the time measured when the first byte of the corresponding HTML file was received from the site.

Figure 1 shows the distribution of the changes detected on all news as a function of the time elapsed since the news was first introduced into the site. Note that the figure plots the changes occurred within the first 72 hours, that is, 99.6% of the total number of changes. As can be seen, a large fraction of the changes, namely, about 78%, occurs within the first 12 hours, and 97% within the first 48 hours. The distribution of number of changes varies across the five categories of news considered in our analysis. For example, about 80% of changes to the health news occur within the first six hours, and about 98% within the first 12 hours. For the sports news, about 76% of the changes occur within the first 12 hours and 96% within 48 hours.

Similarly, our analysis has shown that the majority of the news, namely, 80%, received all their updates within 12 hours. This percentage goes up to 96% if we consider an interval of 48 hours.

To discover similarities in the characteristics of the news, we applied clustering techniques. Each news was described by three parameters, namely, number of changes, average time between two successive changes,

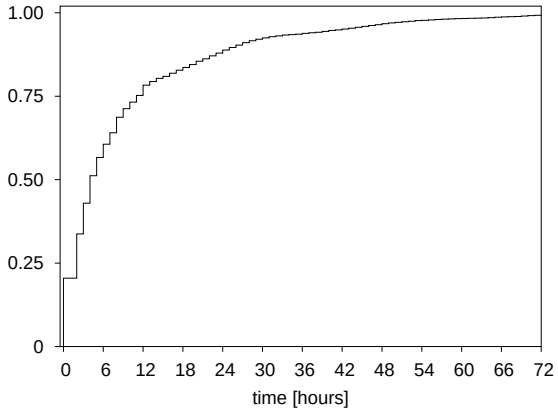

Fig. 1. Distribution of the number of changes as a function of time.

and average size. Clustering yields a partition of the news into four groups. Table 3 presents the centroids, i.e., the geometric centers, of the groups. As can be seen, the news belonging to group 1, that accounts for about 62% of the news, are characterized by a small number of changes and by an interval between changes of approximately 2 hours. Group 4 is the smallest group, only about 6.5% of the news, and contains the news characterized by a number of changes, namely, 11.587, more than 4 times larger than the overall average. The news with largest size are grouped in cluster 2, whereas the news with the largest interval between changes belong to cluster 3.

Table 3. Centroids of the four groups obtained for the dynamic news.

	number of changes	time between changes [min]	size [bytes]	fraction of news
Group 1	1.871	111.929	3,303	62%
Group 2	2.688	154.711	6,872	20%
Group 3	2.253	587.184	3,785	11%
Group 4	11.587	121.889	4,911	7%

The analysis of the composition of each cluster in terms of the categories of the news has shown that about 52% of the news of cluster 4 belong to the sports category, whereas the majority of the news of clusters 2 and 3 belong to the news category. About 65% of the business news and 76% of the health news belong to cluster 1.

To better understand the dynamic behavior of the news, we have analyzed to what extent their contents change. For each news, we first computed the difference between two successive versions in terms of their size, namely, $|s_j - s_{j+1}|$, where s_j and s_{j+1} denote the size of the j–th and of the $j+1$–th version of a news, respectively. This analysis has shown that on the average two successive versions of a news differ by 480.45 bytes, even though this value varies over the categories. For example, for the news category, the average difference is equal to 556.6 bytes. To compare news of different size, we scaled the difference over the maximum size of each news. Figure 2 shows the corresponding distribution computed over all news classified in the news category.

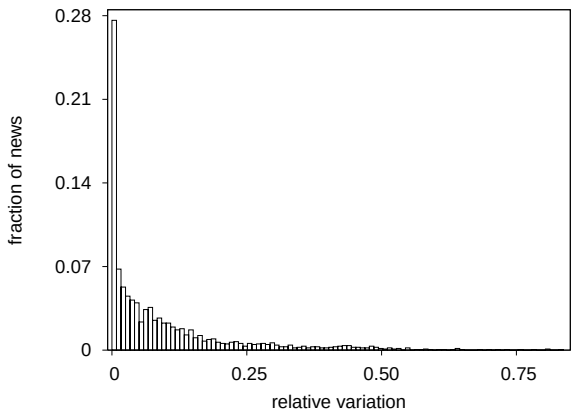

Fig. 2. Distribution of the relative variation of the size of successive versions of all news in the news category.

To quantify the amount of change to a news between two successive versions, we applied the vector model of information retrieval.[12] Each news is represented as a vector, whose components are the unique words that it contains. Let N denote the size of the vocabulary, that is, the global number of distinct words in all versions of a specific news. The $j-$th version of a news is then mapped into an N–dimensional space, namely, it is 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 version of the news. In our analysis, as weight we used the raw term frequency, that is, the number of times each word occurred in the core of the news. Hence, words that occur frequently in a news are more

important than infrequent words. Moreover, despite information retrieval applications, we considered all words including the so–called "stop words".

The similarity measures are based on the various metrics, e.g., cosine, Jaccard and Dice coefficients.[10,11,16] In our analysis we used the cosine metrics. According to this metrics, the similarity between two documents d_j and d_k is given by:

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

The values of this metrics are in the range between 0 and 1. The cosine coefficient is equal to 0 for two versions of a news without any word in common, whereas it is equal to 1 for two identical versions, that is, with the same word frequency.

Figure 3 shows the distribution of the cosine coefficient computed for all the dynamic news considered in our analysis. The values refer to the

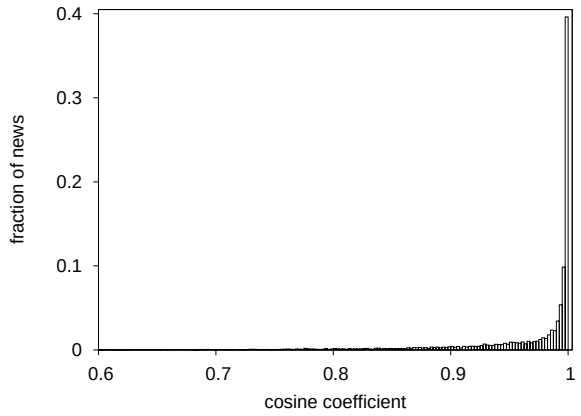

Fig. 3. Distribution of the cosine coefficient of similarity computed between pairs of successive versions of each individual news.

similarity between pairs of successive versions of the news. The figure refers to the coefficients in the range between 0.6 and 1, that accounts for 99.4% of the values. As can be seen, the distribution is skewed towards 1, with an average value equal to 0.96125. This means that in terms of words, the successive versions of the news are rather similar. It is interesting to point out that the minimum of the distribution, that is equal to 0.02916, corresponds

to news belonging to the business category. Similarly, the minimum value of the cosine coefficient computed for the news belonging to the news and health categories are rather small, i.e., 0.03275 and 0.03469, respectively. On the contrary, the corresponding values for the sports and entertainment news are one order of magnitude bigger, namely, 0.31885 and 0.57093, respectively.

Another application of the cosine similarity was aimed at testing to what extent a news changes with respect to its first version, that is, which of the successive versions is closest to the first version. For this purpose, we computed the cosine coefficient between the first version of a news and all its successive versions. The average value of the cosine coefficient is equal to 0.84887, that is, much lower than the value previously computed for successive versions of each news. As expected, the average is even smaller, that is, equal to 0.80038, for the news belonging to the news category. Figure 4 shows the distribution of the cosine coefficient of similarity for the news category. As can be seen, successive versions tend to differ significantly from the first version of the news.

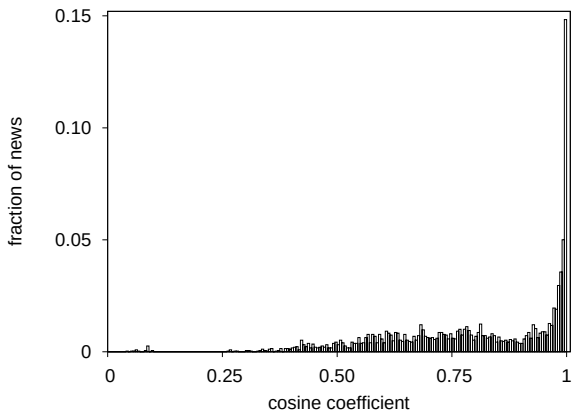

Fig. 4. Distribution of the cosine coefficient of similarity for the news category computed between the first version of a news and all its successive versions.

4. Conclusions

Dynamics of Web contents have to be taken into account when making a decision about caching, content distribution and replication. We studied the

evolution of a popular news Web site with the objective of understanding how often and to what extent its contents change. Despite what expected, we found that the core of most of the news tends to change little and not very often, even though this behavior varies across categories. For example, sports news change more often, whereas news belonging to the news category tend to change to a larger extent. Similarity measures have also shown that successive versions of the same news do not differ significantly, whereas the difference increases when we compare the first version with the following ones. As a future work, we plan to study the performance implications of this behavior and extend the analysis to the whole HTML files.

Acknowledgments

This work was supported by the Italian Ministry of Education, Universities and Research under the FIRB programme. Authors wish to thank Clara Parisi for her valuable help in setting up the experimental environment.

References

1. A. Barili, M. Calzarossa, and D. Tessera. "Characterization of Dynamic Web Contents" LNCS 3280 - Springer, 2004, 648-656 (In Computer and Information Sciences - ISCIS 2004).
2. B. E. Brewington and G. Cybenko. "How Dynamic is the Web?" *Computer Networks*, **33**, (2000), 257-276.
3. B. E. Brewington and G. Cybenko. "Keeping Up with the Changing Web". *IEEE Computer*, **33**(5), (2000),52-58.
4. J. R. Challenger, P. Dantzig, A. Iyengar, M. S. Squillante, and L. Zhang. "Efficiently Serving Dynamic Data at Highly Accessed Web Sites". *IEEE/ACM Transactions on Networking*, **12**(2), (2004), 233-246.
5. L. Cherkasova and M. Gupta. "Analysis of Enterprise Media Server Workloads: Access Patterns, Locality, Content Evolution, and Rates of Change". *IEEE/ACM Transactions on Networking*, **12**(5), 2004,781-794.
6. L. Cherkasova and M. Karlsson. "Dynamics and Evolution of Web Sites: Analysis, Metrics and Design Issues", 2001, 64-71, (In Proc. of the 6th IEEE Symposium on Computers and Communications, 2001).
7. J. Cho and H. Garcia-Molina. "Estimating Frequency of Change". *ACM Transactions on Internet Technology*, **3**(3), 2003,256-290.
8. F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. "Rate of Change and other Metrics: A Live Study of the World Wide Web", 1997, X-Y (In Proc. of the First USENIX Symposium on Internet Technologies and Systems, 1997).
9. D. Fetterly, M. Manasse, M. Najork, and J. Wiener. "A Large-Scale Study

of the Evolution of Web Pages". *Software - Practice and Experience*, **34**(2), (2004), 213-237.

10. D. L. Lee, H. Chuang, and K. Seamons. "Document Ranking and the Vector-Space Model". *IEEE Software*, **14**(2), (1997), 67-75.
11. L. Lee. "Measures of Distributional Similarity", 1999, 25-32, (In Proc. of the 37th Conference on Association for Computational Linguistics, 1999).
12. M. J. McGill. *Introduction to Modern Information Retrieval.* (McGraw-Hill, New York), 1983.
13. "MSNBC Web site". `http://www.msnbc.com`.
14. V. N. Padmanabhan and L. Qiu. "The Content and Access Dynamics of a Busy Web Site: Findings and Implications", 2000, 111-123, (In Proc. of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 2000).
15. W. Shi, E. Collins, and V. Karamcheti. "Modeling Object Characteristics of Dynamic Web Content". *Journal of Parallel and Distributed Computing*, **63**(10), (2003), 963-980.
16. J. Zobel and A. Moffat. "Exploring the Similarity Space". *ACM SIGIR Forum*, **32**(1), (1998), 18-34.