An exploratory analysis of the novelty of a news Web site

Maria Carla Calzarossa

Dipartimento di Informatica e Sistemistica
Università di Pavia
I-27100 Pavia, Italy
mcc@unipv.it

Daniele Tessera

Dipartimento di Matematica e Fisica
Università Cattolica del Sacro Cuore
I-25121 Brescia, Italy
daniele.tessera@unicatt.it

Abstract

The growing amount of information published on the Web, combined with its dynamic nature, opens many challenging issues dealing with management and retrieval of the information and provisioning of the underlying infrastructures. Search engines have to meet two conflicting requirements: minimize the number of downloads and provide up-to-date information. In this paper, we present the results of an exploratory analysis aimed at investigating the novelty of the content of a news Web site. We analyzed the Web site from an horizontal perspective by focusing on the content of the individual articles and from a vertical perspective by focusing on the entire collection of articles published on the site. These two perspectives allowed us to study how fast and to what extent articles were modified and to model the evolution of the Web site.

I. Introduction

The Web has become the primary platform for disseminating and sharing on a very large scale personal and professional information and for conducting many kinds of businesses. This explosive growth opens new challenging issues dealing with management and retrieval of the information and with provisioning of computing power, storage and bandwidth for the underlying infrastructures. The highly dynamic nature of many Web sites, where new pages are often uploaded and existing pages are modified and eventually disappear, makes all these issues even more challenging.

Search has clearly emerged as a key enabling technology to facilitate users navigate in the huge amount of information available on the Web. Search engines are expected to crawl, index, group and cache the information as to present users with updated and relevant Web pages. Similarly, RSS feeds are expected to distribute in a timely manner information captured from very many different Web sources. The crawling frequency and its targets are driven by some conflicting tradeoffs. As the information on the Web is rather fluid, its novelty varies and is not always worth the cost of a download. For example, news Web sites regularly upload new stories and generate multiple versions of the same story. This phenomenon is even more relevant in social media Web sites where users collaboratively create, evaluate and share information. On the other hand, the actual importance and relevance of some Web content, such as, bids, stocks, news, mainly rely on its freshness, that is, new information has to be captured and distributed as soon as posted. In this framework, it is important to assess the originality of a Web page or of an entire site and how often and to what extent the content changes and needs to be refreshed.

The literature is rich of papers dealing with the analysis of the dynamic behavior of Web sites. Some of these papers (see, e.g., [1], [2], [3]) studied the evolution of Web sites in terms of frequency and nature of changes. Other papers (see, e.g., [4], [5], [6]) focused on the degree of change of Web pages by considering various aspects, such as, size, content and link structure.

Because of their characteristics and popularity, news Web sites received some special attention. Some studies analyzed the mechanisms for news discovery and ranking, whereas few papers specifically tackled their evolution. Gabrilovich et al. [7] addressed the problem of filtering news stories according to some measures of information novelty by analyzing the inter- and intra-document dynamics and considering how information evolves over time from article to article and within individual articles. Kutz and Herring [8] focused on the content of the headlines stories displayed on three news Web sites with the objective of classifying and understanding the types of changes taking

place. A methodological approach for the characterization of a news Web site is presented in [9], where the evolution of the site is described by means of analytical models that capture and reproduce its dynamics in terms of rates of page creations and updates and extent of content changes.

This paper presents an exploratory analysis of the content of a news Web site with the aim of assessing its novelty and more specifically how fast and to what extent content evolves over time.

Our study is motivated by the compelling need to select and extract novel information from the growing amount of news published on the Web and reduce the number of validations required to maintain a given level of freshness. Indeed, the characteristics of news Web sites make it difficult to assess what is actually fresh. New articles are frequently posted and older articles eventually dropped and migrated to some sort of off-line archives. Moreover, the novelty of a story tends to fade with time and thus the attention users pay to it.

The paper is organized as follows. Section II describes the data sample used in the analysis. The evolution over time of the content of individual articles is presented in Section III, whereas the novelty and evolution of the entire collection of articles published on the Web site are discussed in Section IV. Finally, some conclusions are drawn in Section V.

II. Data sample

The basis of our study consists of the Web pages of the international edition of the CNN Web site [10]. We took multiple snapshots of the site by crawling it for a period of approximately 15 weeks, from April 2008. The snapshots were taken at a rather fine grain, that is, every 15 minutes. Indeed, as we will explain in more details later on, we noticed that this interval could capture the behavior of the site with a good accuracy and a reasonable usage of bandwidth and server resources.

At each snapshot we retrieved all Web pages listed on the site as "published in the past 24 hours", that is, on the average 94 pages per snapshot. Specifically, we downloaded HTML files for some 6.3Mbytes per snapshot.

During our monitoring interval, a total of new 8,508 articles were published on the site. Hence, for each of these articles, we downloaded the corresponding HTML files until the article disappeared from the list of the past 24 hours articles.

To study the properties of the site in terms of novelty, variety and evolution, we focused our analysis on the actual content of each article, thus ignoring markups, banners, navigation bars, advertisements and all sorts of add-ons included in Web pages for various purposes. Hence, the preliminary step of our analysis deals with the parsing of the HTML files to strip all these components and extract their core text. By comparing the texts of the consecutive downloads of each article, we could assess whether these texts were ever modified during the monitoring interval. We then classified the articles in two groups: static articles, whose text never changed, and dynamic articles, whose text was updated once or more during the monitoring interval.

Table I presents some statistics of the files downloaded from the Web site. The average size of an HTML file is

	HTML file	Core text	Number of articles
Static articles	69,610	2,988	4,564
Dynamic articles	70,559	3,201	3,944
Overall	70,050	3,087	8,508

TABLE I. Size (in bytes) of the articles downloaded from the Web site.

approximately 68Kbytes, whereas the text extracted from each file after its parsing represents only a very small fraction, that is, about 3Kbytes. In general, the size of dynamic articles tends to be slightly larger than the size of static articles. Nevertheless, these values do not vary significantly across articles. The standard deviation of the size of HTML files is an order of magnitude smaller that the corresponding average. Moreover, even though some core texts are rather large, i.e., 20Kbytes or larger (see Figure 1), the size of the texts of the majority of the articles does not exceed 5Kbytes.

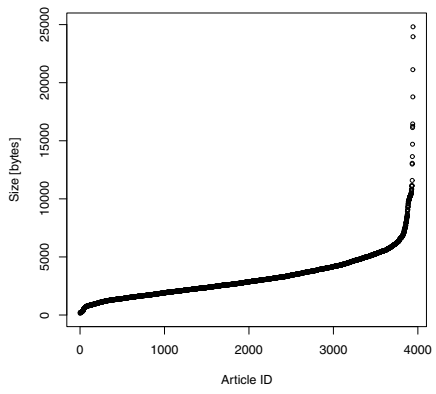


Fig. 1. Size of the core text of the articles.

The preliminary analysis of the data sample has shown that the static articles account for the 53.6% of the articles published on the site, whereas the remaining articles, namely, 3,944, were dynamic and received on average 2.6 updates each. Figure 2 plots the behavior of the articles with respect to their number of updates. Note that we

used a log scale on the y-axis. While the majority of dynamic articles (51.7%) was modified only once, 12 articles received more than 25 updates each. The total number of updates of these 12 articles is equal to 713 and accounts for about 6.9% of the updates. In particular, an article describing the 2008 China earthquake was modified as many as 90 times while it was listed among the past 24 hours articles of the Web site.

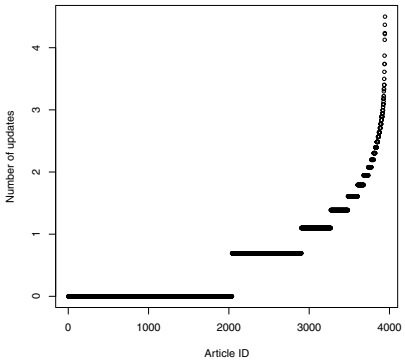


Fig. 2. Number of updates of the dynamic articles.

It is interesting to outline that the updates of the multimedia contents (e.g., images, videos) that are part of the HTML files are characterized by a slightly different behavior. More precisely, we have detected 4,191 updates that affect only the multimedia components. Moreover, the updates of 1,084 articles did not modify their core text, they only involved their multimedia content. Nevertheless, as these types of contents are often seen as add-ons aimed at enhancing the experience of the users, we decided to ignore them and focus our analysis on the core of the articles. Hence, from now on with the term article we refer to its actual content, that is, its core text.

III. Temporal evolution of the articles

A preliminary characterization of the temporal evolution of the Web site was addressed from an horizontal perspective, that is, we investigated how the information evolves within each article by analyzing its behavior in terms of updates and more specifically to what extent an article changed whenever it was modified. Hence, in this analysis we considered the different versions of the 3,944 dynamic articles published on the site.

The average inter-update times of the articles, that is, the times between two consecutive updates, are plotted in Figure 3. For about one fourth of the articles the inter-

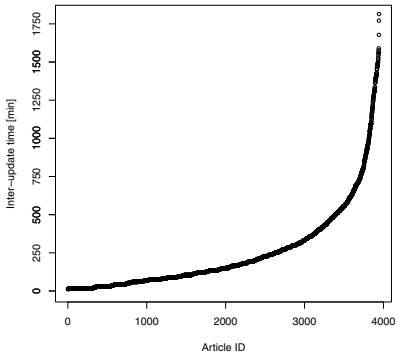


Fig. 3. Inter-update times of the dynamic articles.

update time did not exceed 67 minutes and for half of the articles was within 149 minutes, even though we have detected an article characterized by an inter-update time as large as 1,814 minutes. More specifically, the mean inter-update time of the articles is equal to 248 minutes, that is, approximately four hours, whereas the corresponding standard deviation is slightly larger, namely, equal to 278.7.

It is worth to point out that the first update of an article occurred on average some 219 minutes after the article was first posted on the Web site. Moreover, by analyzing the articles modified once, we could observe that their mean inter-update time is slightly longer, namely, 274 minutes with a standard deviation equal to 339. Indeed, the vast majority of the articles characterized by very large inter-update times received one update only. In general, as we will discuss in more details later on, these articles were modified to a non negligible extent, whereas for most of the articles modified within an hour of their publication, the changes affected their text to a rather limited extent.

By analyzing the variation of the size of articles after an update, we noticed that the majority of the updates, namely, 67.84%, added some extra text to the previous version of the article. On the average the size increased by some 444 bytes. On the contrary, whenever an update removed text, the article was shortened by approximately 539 bytes. For about half of the dynamic articles, all their updates added text (some 413 bytes each), thus reflecting a typical behavior of news Web sites where articles are posted as soon as something new happens, e.g., breaking news, and eventually updated to include the latest developments of the story. Finally, it is worth noting that for very few articles, namely, 46, the updates did not modify their size.

As a further refinement of the analysis, we quantified to what extent a change modified the content of an article. This analysis is based on the application of the vector space

model used in the framework of information filtering and retrieval [11]. Each article was represented by a vector, whose components are the unique words used in the article itself.

As we were interested in analyzing to what extent information evolves within an article, we mapped every article and each of its updates into an $n-$dimensional space, n being the size of the vocabulary, that is, the total number of distinct words used by both. For our analysis, the average size of the vocabulary was equal to 342 words. More than 55% of these words were common to both texts. Let us recall than an article consists on average of some 255 words.

An article i was represented by a vector $d_i = (w_{i,1}, w_{i,2}, ..., w_{i,n})$, where $w_{i,k}$ denotes the weight of word k, $(k = 1, 2, ..., n)$. As weights, we simply used the term frequency, that is, how often word k occurs in article i. The similarity was computed by means of the cosine coefficient (see [12]). According to this metrics, the similarity between article i and its update j, each represented by the corresponding vector d_i and d_j, is given by:

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

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, whereas values close to 0 denote little or no similarity at all as the vectors are orthogonal.

The mean value of the cosine coefficient of similarity computed between an article and each of its updates is equal 0.9225. The details of the behavior of these coefficients are shown in Figure 4. From the figure we notice that for about three fourth of the articles, that is, 3,000 articles, the value of the cosine coefficient was larger than 0.9, whereas for only some 100 articles it did not exceed 0.5. This means that most of the updates did introduce minor changes in the articles and major changes involved only few articles. Moreover, we did not discover any correlation between the cosine coefficient of similarity and the inter-update time. The correlation coefficient computed between these two parameters was equal to -0.1797. Figure 5 shows the scatter diagrams of the cosine coefficients of similarity as a function of the inter-update times for the 2,038 articles with one update (Fig. 5(a)) and for the remaining 1,906 articles updated more than once (Fig. 5(b)). These diagrams reveal some interesting peculiarities of the articles. The concentration of articles in the top left corner of both diagrams denotes the presence of many articles with very short inter-update times and very high degree of similarity. On the contrary, most of the articles characterized by a low degree of similarity have been updated once and sometimes even some hours after their publication on the Web site. Let us remark that the average cosine coefficients of similarity are equal to 0.9338 and 0.9111 for the articles updated once and for the articles with multiple updates, respectively.

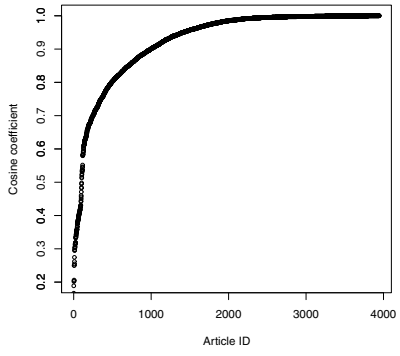


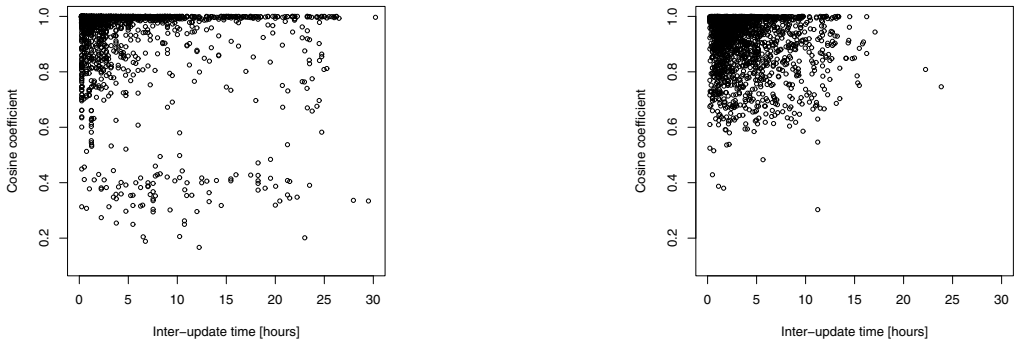
Fig. 4. Cosine coefficient of similarity computed between an article and all its updates.

To study in details the behavior of the articles with multiple updates, we computed the cosine coefficient of similarity for the pairs of consecutive updates of every article. As expected, the mean value of the coefficient, equal to 0.9483, is larger than the value computed in the previous analysis, that is, the information tends to evolve to a smaller degree.

Another complementary measure used to assess the evolution of the information within dynamic articles is the character-level edit distance. This metrics computes the minimum number of operations, i.e., insertions, deletions, substitutions, required to transform one text into another [13]. For the pairs of updates of every article, we computed the edit distance normalized with respect to the longest transformation. Such a normalization allows a fair comparison of the values obtained for each article and across articles. The average value of the normalized edit distance is equal to 0.2, that corresponds to 618 edit operations of one character each to transform one text into the other, thus confirming that most of the consecutive updates modified the articles to a limited extent.

IV. Temporal evolution of the snapshots

The variety and novelty of the information available on the Web site were investigated more specifically by analyzing the site from a vertical perspective. We focused our attention on the collection of articles downloaded at a given snapshot and studied the evolution and the similarity



(a) (b)

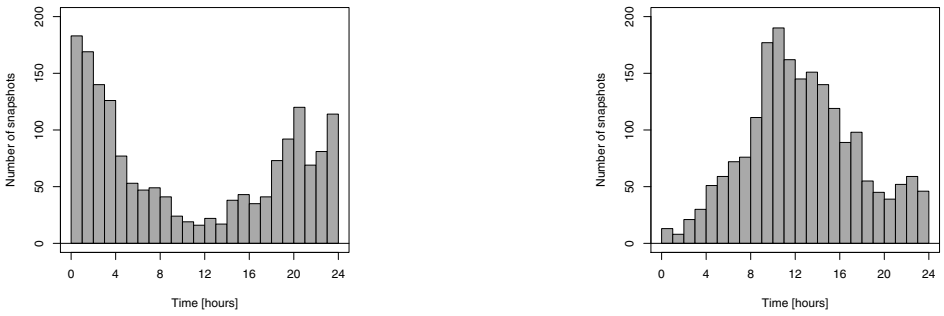
Fig. 5. Scatter plots of the inter-update times versus the cosine coefficients of similarity for the articles updated once (a) and more than once (b).

between consecutive snapshots.

As already pointed out, we collected some 10,000 snapshots by crawling the Web site every 15 minutes. At each snapshot, we downloaded on average 94 articles, in general, fewer over week-end days than over weekdays.

To study the evolution of the Web site, we compared pairs of consecutive snapshots and assessed their differences in terms of uploads of new articles, updates of existing articles and removals of the articles from the list of the last 24 hours articles. A preliminary characterization of the snapshots has shown that these activities are not evenly distributed across snapshots: uploads are distributed across some 5,102 of the snapshots, updates across 5,912 and removals across 4,668. Moreover, in 1,689 snapshots (about 17%) no activities have been detected, that is, the content of the Web site did not change for an interval of at least 15 minutes. About a half of these snapshots were clustered in 329 sequences whose average length is equal to 2.6. Most of these sequences are concentrated over night, namely, about 50% between midnight and 4:30am and another 25% after 7:30pm, with the highest peak between midnight and 1am (see Fig. 6 (a)). On the contrary, the site was very dynamic during the day and especially between 10am and 5pm, where we have detected a strong concentration of activities. For example, more than one third of the snapshots taken in this time interval were characterized by all types of activities, that is, consecutive snapshots were different in terms of uploads, updates and removals of articles (see Fig. 6 (b)).

As a further development of this analysis, we tested whether the dynamic behavior of the site followed any specific pattern and whether it was characterized by any type of dependencies. In particular, we focused our attention to the time interval between 10am and 5pm where the largest number of changes took place. We explored the use of finite state, discrete time, homogeneous Markov chains as models for representing the sequences of activities that describe the evolution of the Web site. The states of the Markov chain were represented by tuples $(n_{upl}, , n_{upd}, n_{del})$, where the subscripts upl, upd and del refer to uploads, updates and removals, respectively. Each component of the tuple can be either 0 or 1, denoting the presence or the absence of the corresponding activity, respectively. Indeed, we were not interested in the multiplicity of these activities, we were rather interested in assessing the occurrence of any of the three types of activities that make the Web site change. For example, the state $(1, 1, 0)$ denotes that the snapshot of the Web site changed with respect to the previous snapshot because of uploads and updates, whereas no articles were removed. The chain was represented by a total of eight different states. For each state, we computed the transition probabilities p_{ij} of moving from state i to state j, $i, j = 1, 2, ..., 8$. We then applied the standard tests, at the 0.05 significance level, to assess the time dependence, the homogeneity and the order of the corresponding Markov chain [14]. The test for time dependence has rejected the hypothesis of a process represented by an independent multimodal distribution. The other tests have shown that a first order model characterized by homogeneous transitions is a suitable representation of the evolution of the entire Web site. This means that the transitions of the site through the various states did only depend on the current state. The limiting state probabilities computed from the transition probability matrix have indicated that some states are more probable than others. The most probable state, with a limiting state probability equal to 0.3431, is the state $(1, 1, 1)$, that includes to all types of changes with respect



(a) (b)

Fig. 6. Distribution over the 24 hours of the snapshots with no activity (a) and of the snapshots with all types of activity (b).

to the previous snapshot. The less probable state is the state $(0, 0, 1)$ with limiting state probability equal to 0.0327. These Markovian models could be then exploited by search engines for tuning their crawling strategies.

V. Conclusions

News web sites are characterized by a peculiar behavior where articles are posted as soon as something new happens and eventually modified to refine the text or to reflect the latest developments of the stories. Moreover, the relevance of news articles tends to decrease as time elapses. Hence, it is important for crawlers to detect in a timely manner newly uploaded articles and distinguish them from older articles. Similarly, it is important to recognize updates that are worth the cost of a download.

The exploratory analysis of the CNN news Web site presented in this paper has shown that the updates of the articles follow different patterns. Some articles never changed during our monitoring interval, whereas others changed to a rather limited extent and major changes involved only few articles. The analysis of the entire collection of articles published on the site has shown that the site was very dynamic during the day, where we have detected a strong concentration of uploads, updates and removals of articles. The sequences of activities describing the evolution of the Web site were then modeled by a first order Markov chain characterized by homogeneous transitions.

As a future work, we plan to cluster similar articles in groups and to study the evolution and the persistence of these groups. Moreover, we will analyze to what extent the recommendation mechanisms used within social networks influence the relevance of the articles and help in filtering the vast streams of news articles.

References

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

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

[3] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A Large-Scale Study of the Evolution of Web Pages," *Software: Practice & Experience*, vol. 34, no. 2, pp. 213–237, 2004.

[4] 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*, 2004, pp. 1–12.

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

[6] R. Baeza-Yates and B. Poblete, "Dynamics of the Chilean Web Structure," *Computer Networks*, vol. 50, no. 10, pp. 1464–1473, 2006.

[7] 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*, 2004, pp. 482–490.

[8] D. Kutz and S. Herring, "Micro-Longitudinal Analysis of Web News Updates," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences HICSS'05*, vol. 4. IEEE Press, 2005, p. 102a.

[9] M. Calzarossa and D. Tessera, "Characterization of the evolution of a news Web site," *Journal of Systems and Software*, vol. 81, no. 12, pp. 2336–2344, 2008.

[10] CNN International Web site, http://edition.cnn.com.

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

[12] D. L. Lee, H. Chuang, and K. Seamons, "Document Ranking and the Vector–Space Model," *IEEE Software*, vol. 14, no. 2, pp. 67–75, 1997.

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

[14] T. W. Anderson and L. Goodman, "Statistical Inference about Markov Chains," *Annals of Mathematical Statistics*, vol. 28, pp. 89–110, 1957.