In the Proceedings of the 3rd IEEE International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2002), Lecture Notes in Computer Science (LNCS), volume 2412, H. Yin et al. (Eds), pages 173-178, Manchester, UK, August 2002.

Study of the Regularity of the Users' Internet Accesses

Nicolas Durand and Luigi Lancieri

France Telecom R&D Caen
{nicola.durand, luigi.lancieri}@francetelecom.com

Abstract. The aim of this study is to investigate relationship between past users' behavior (described by access patterns) and future one. The two main ideas are first to explore the possible users' characterization that can be extracted from access pattern. This allows to measure and to have a better understanding of users' behavior. This knowledge allows us to build new services as building interest communities based on a comparative approach and clustering. The second idea is to see if these characterizations can be useful to forecast future access. This could be useful to prefetch web data in proxy-cache. We show that there are some partial mathematical models binding the users' behavior to the repetition of queries.

1 Introduction

Discovery of frequent patterns [1, 2] in the past accesses of the web users is a very interesting subject of study if we wish to model their behavior. One of the applications of these studies is to prefetch web information [3]. In our context, the prefetching consists in loading in a local proxy-cache [4] some distant web information having a high probability to be accessed by the users in the future. The aim is to improve the access speed by decreasing the latent period (response time to a query). Thus, the performances of such methods depend on finding regularity or rules on users behavior. The model of user's behavior is also interesting for other purposes including social study [5]. A good model can be used to simulate humans interactions and to have a better understanding of human access to knowledge. A model can also be used to optimize existent services as clustering users in groups of interest having a comparable behavior. So, we made an analysis of users accesses to Web through an operational proxy-cache during a long period of time, in order to determine the level of time regularity and behavioral correlation. Contrary to Web server, proxy-cache allows having a better reflect of users behavior since its trace covers larger information requests (potentially the full Web).

The organization of this document is as follows. In Section 2, we discuss the context of our experiment and the data. Then in Section 3, we describe our method. In Section 4, we detail a batch analysis of the results. In Section 5, we discuss a temporal analysis. In Section 6, we compare the behavior of the users. In Section 7, we discuss previous works on the analysis of the users behavior. We conclude in Section 8.

2 Context of the experimentation

We used the log files of two operational proxies-caches of France Telecom R&D (at Caen), over a period of 17 months (from January 1-st, 1999 to May 31, 2000) concerning 331 users. The total number of queries is 1510358 corresponding to 392853 different objects (a given object can be asked several time).

The log files was purified, we only kept the queries corresponding to an object of mime type: "text". This approach answers the following logic. From a cognitive point of view, the textual pages are strongly linked to the explicit step of the users [5]. In terms of probability, the inclusive objects are primilary textual type accesses. In semantic terms, text is also more easily definable than for example images or video sequences. The remaining data corresponds on average to: 1186 different URLs per user, and approximately 8.6 queries per user and per day.

3 Description of the regularities analysis method

At first, we studied the level of regularity on different sequences of access without worrying about the chronology (see Section 4). Then we made a study to determine the level of time regularity of the users' accesses (see Section 5). The global redundancy is the ratio between the number of queries and the number of unique URLs over the total period of consultation. We computed the percentage of common queries between two different sequences of queries that we noted $CQTS$ (Common Queries in Temporal Sequences). These two variables are calculated per user and on average on all the users. All user queries are sorted chronologically and split into sequences of equal length. This corresponds to a percentage of the total number of user queries and is recorded as T. Figure 1 illustrates the process. For the first sequence noted S_1, we compute the number of common URLs with the following sequence noted Q_2, this corresponds to a space of Δ=0. We repeat this process until all the space of sequence is covered: let (S_i, Q_j) be the pairs of studied sequences, and n the total number of sequences, then $i \in [1, n/2]$ and $j \in [i+1, n/2+i]$. In this way, we have the same number of computed values for each of the chosen sequences S_i.

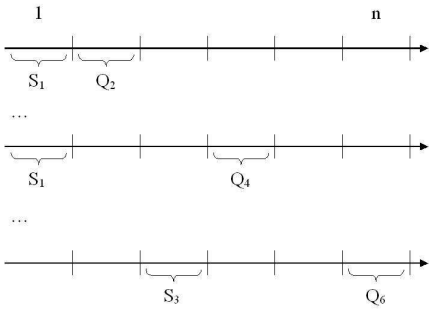

Fig. 1. Method of analysis

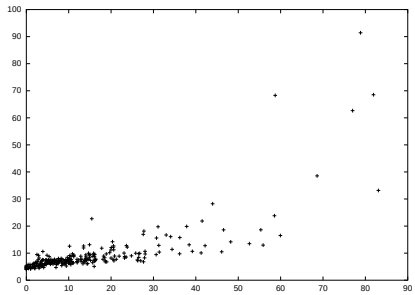

Fig. 2. Global redundancy according to $CQTS$

4 Batch analysis

For every user, we calculated the global redundancy of these accesses and the average of $CQTS$ (with T=5%). We obtained the curve illustrated by the Figure 2, where we have in abscissa, $CQTS$, and in ordinate, the global redundancy. We notice that $CQTS$ and the global redundancy vary in the same direction according to a geometrical law with an increasing dispersal according to the redundancy. In the extreme cases (all URLs identical or all URLs different), the level of $CQTS$ or global redundancy is identical, and in the intermediate cases, the global redundancy grows much slower than $CQTS$.

The redundancy of the queries can be interpreted as illustrating the behavior of the user [5]. For example a monolithic behavior (concentrated accesses on few web sites) will have a strong global redundancy while a more scattered behavior (distributed accesses on a large number of URLs) will have a low redundancy. We observe here the highlighting of the coherence of users' accesses. Indeed, if the accesses were done at random, $CQTS$ would statistically be very close to the global redundancy.

5 Temporal analysis

In this section, we studied the value of users' $CQTS$ according to the space of time between the sequences of queries (see Figure 1). We made vary the size of the studied sequences (T=1, 2, 3, 5, 10%). The maximal value of $CQTS$ is realized when the space is minimised. The behavior of this curve (except the maximum value of $CQTS$) is independent of the users and shows a strong temporal coherence. The maximal value is not very different according to the size of the studied sequences. It is between 23,6% and 25,3%. We notice that the level of $CQTS$ according to the time separating both compared sequences, follows a sub exponential law. Indeed, after a logarithmic transformation (log X-axis, log A-axis), the increase of the redundancy according to the increase of the time separating sequences is constant (see Figure 3). The slopes of the curves of the Figure 3 have a coefficient of regression of 0.99, that shows an excellent correlation.

The sub exponential law binding these two parameters can be expressed in the following way. Let Δ be the space between the studied sequences, $cqmax$ be the maximal value of $CQTS$, T be the size of studied sequences (expressed in percentage of all the queries), cq be the $CQTS$ value, and k be a constant.

We have the following relation: $log(cq) = -kTlog(\Delta) + log(cqmax)$,
where $-kT$ is the slope of the straight line. Thus, we have:

$$cq = cqmax.\Delta^{-kT}$$

So, it means that we can determine the value of $CQTS$ knowing the size of the studied sequences and the value of the space of time. We can verify with the Figure 4, that the slope of the straight lines of the Figure 3 evolves linearly according to the size T of the sequences. It brings out that the global behavior of the function cq (derived of the affine function of the Figure 3) is independent of the user whose the behavior intervenes only with the value of $cqmax$.

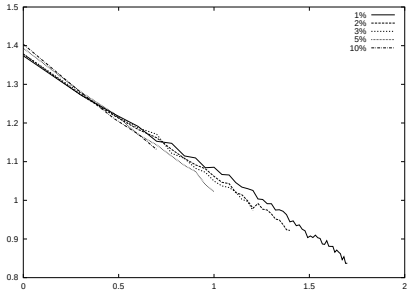

Fig. 3. Logarithm transformation of the temporal analysis

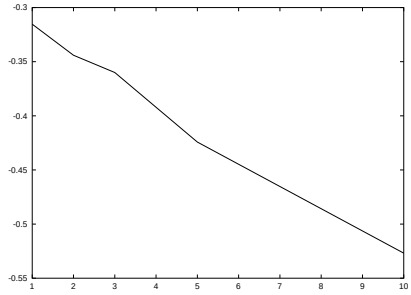

Fig. 4. Evolution of the slope according to T

6 Comparative analysis of users

We are interested here in the relations between the behavior of the users, with their maximal $CQTS$ value, and the average of $CQTS$ values. With the Figure 5, we studied the evolution of the maximum $CQTS$ value, $cqmax$, according to the rank of the maximal $CQTS$ value of the user, for T=5%. We notice, by disregarding the extremes, that there is a relatively proportional evolution, what confirms the conclusions of the Section 5.

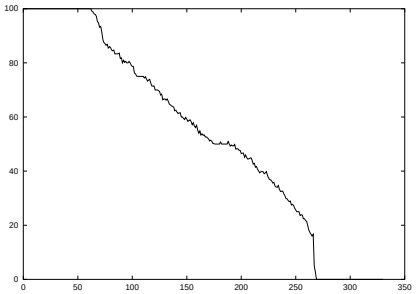

Fig. 5. Maximal $CQTS$ value according to the user rank

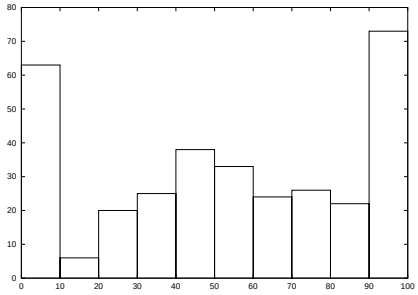

Fig. 6. Distribution of the maximal $CQTS$ values among users

The distribution of the values of the maximum $CQTS$ is represented in the Figure 6. We have in ordinate the number of users, and in abscissa the maximum $CQTS$ value. We notice by disregarding the extremes, that the distribution is almost homogeneous. For a $CQTS$ included between 90 and 100%, we have 73 users (22%). This high number explains itself because of a user having a low number of queries and a minimum of regularity, has a $CQTS$ very high for a low space of time (Δ).

Then, we sorted out the users according to their $CQTS$ value (with T=5%). The Figure 7 indicates in abscissa the rank of bigger $CQTS$, and in ordinate the $CQTS$ value. Remember that the more $CQTS$ is high, the more the future

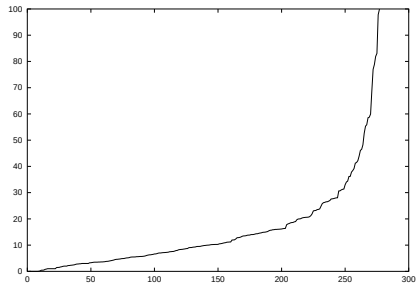

Fig. 7. $CQTS$ according to the users rank

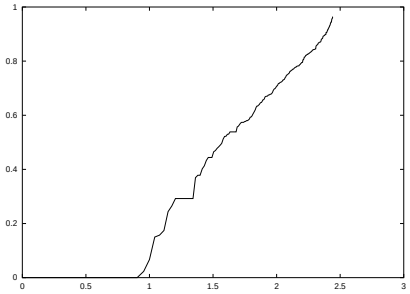

Fig. 8. Hyperbolic tangent transformation of the Y-axis of the logarithm transformation of the Figure 7

accesses will resemble the past ones. We notice clearly that there is a small group of user having a relatively high $CQTS$ (10% of the users with more than 30%). There are very few users with a very high $CQTS$ (2% of the users with more than 70%). Just like the Section 5, we applied some mathematical transformations to try to discover a underlying law to this evolution. We computed the logarithm transformation on the ordinate of the curve of the Figure 7, and we applied the hyperbolic tangent transformation (see Figure 8). The curve of the Figure 8 has a coefficient of linear regression of 0.98 that corresponds to a good correlation. We notice that the increase of the $CQTS$ according to the increase of rank among users classified by increasing $CQTS$, is a constant (in the space of the transformation logarithm / hyperbolic tangent logarithm). It means that if the user A is 3 times as regular as the user B, we can estimate the $CQTS$ of B given the $CQTS$ of A.

7 Related works

To study the regularity of the accesses, the notion of self similarity is a very interesting tool for the modeling. The self-similarity (heavy tail, Zipf's law) expresses that a series of data shows the same characteristic in different scales (fractals). In the case of a chronological series, the self-similarity implies the dependence between the short term and the long term (spatial and temporal locality [6]). Leland and al [7], and Paxson and al [8], showed respectively in a LAN context and a WAN context, that the forecast of traffic was much more reliable by taking into account a stream distribution of self-similar data rather than a distribution of Poisson. While in the case of Poissonian or Markovian streams, the level of variation or disparity of the queries tends to decrease in the time, the real traffic reveals a stability on various scales of time. Studies [9] tend to show that web tracks do not follow strictly Zipf's law, but which remain nevertheless self-similar. Other works on the variability of the users' behavior of Web were realized by taking into account the semantic aspect underlying the queries [10, 5].

8 Conclusion

We studied the regularity of the users' accesses of proxies thanks to the $CQTS$ computation. Except the statistics information on these accesses, we can highlight some natural conclusions on the users' behavior and thus to forecast it better. The users, at least a great majority, are not regular in their accesses. Indeed, the maximal rate of $CQTS$ is low at 25.3%. In spite of this lack of regularity, we brought out the important coherence of the accesses. This internal logic suggests underlying laws, at least by the rational component of the user's behavior. We are far from having discovered a general model of the behavior of the user (assuming that it is possible). However, we showed that there are partial models binding the users' behavior to the repetition of queries. We also showed that the temporal coherence of the accesses was a factor independent from users. We notice the strong diminution of the $CQTS$ value in the time. That means that if we mine the access patterns by using log files at the given moment, then the obtained patterns can really be used only over a short period.

In future works, we will go more to depth: at the level of the keywords of the consulted documents. The $CQTS$ value at the level of the consulted keywords should be higher (a keyword can be common to several URLs). Furthermore, this will allow us to get the centres of interests of the users, which vary less quickly in the time and allow a higher level of description.

References

[1] R. Agrawal and R. Srikant. Mining Sequential Patterns : Generalizations and Performance Improvements. In *EDBT'96*, Avignon, France, March 1996.

[2] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of Frequent Episodes in Event Sequences. Technical Report C-1997-15, Helsinki, February 1997.

[3] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos. Effective Prediction of Web-User Accesses: A Data Mining Approach. In *WebKDD'01 Workshop*, 2001.

[4] Q. Yang, H. H. Zhang, and T. Li. Mining Web Logs for Prediction Models in WWW Caching and Prefetching. In *ACM SIGKDD'01*, San Francisco, 2001.

[5] L. Lancieri. *Memory and Forgetfulness: Two Complementary Mechanisms to Characterize the Various Actors of the Internet in their Interactions.* PhD thesis, University of Caen, France, December 2000.

[6] V. Almeida, A. Bestravos, M. Crovella, and A. de Oliveira. Characterizing Reference Locality in the WWW. In *PDIS'96*, Miami Beach, December 1996.

[7] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the Self-Similar Nature of Ethernet Traffic. In *ACM SIGCOMM93*, pages 183–193, San Francisco, 1993.

[8] V Paxson. Fast Approximation of Self-Similar Network Traffic. Technical Report LBL-36750, University of California, Berkeley, April 1995.

[9] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. On the Implications of Zipf's law for Web Caching. In *the 3rd Int. WWW Caching Workshop*, June 1998.

[10] S. Legouix, J. P. Foucault, and L. Lancieri. A Method for Studying the Variability of Users' Thematic Profile. In *WebNet2000 AACE*, San Antonio, 2000.