Internet User Behavior: Compared Study of the Access Traces and Application to the Discovery of Communities

Luigi Lancieri, *Member, IEEE*, and Nicolas Durand

***Abstract*—With an ever-increasing emphasis on human activity (idea exchange, shopping, gaming, etc.) being mediated through the data network, the understanding of Internet users' behavior has become a rising challenge. Research dealing with the analysis and modeling of Internet user behavior can be roughly split in to two main approaches. The first is based on sociocognitive observation of users' practices in a standardized context. The second approach focuses on the analysis of productions and the traces of users' activity. This paper relates to the latter approach and presents a comparative analysis of Internet navigation traces (URLs versus keywords) to characterize individual or group-of-users' behavior when accessing the Web. The proposed models are based on the study of accesses redundancy seen as global static parameters and from the angle of time evolution. We also study the use of these models, in particular, to categorize a population of users in communities of interests. This study enables us to draw some conclusions on the compared performances of the two kinds of trace exploitation, as raw information, as well as the self-similar properties of the models.**

***Index Terms*—Data mining, Internet, redundancy, self-similarity, user modeling.**

I. Introduction

THE INTERNET adds a new dimension to the interactions between people through information access. All of us acknowledge that Internet provides easier worldwide access to huge quantities of data, but it is also a medium offering new capacities in more direct human exchanges (forums, chat, e-mail, etc.). Using computer technology, many Internet devices also generate traces that memorize all kinds of activity from that of the internal device itself (e.g., processor, memory load, etc.) to that of the external interactions with the devices (from other devices or human users). Initially, this tracing capacity was allowed in order to control the devices in case of failure, but more recently, this capacity has been studied in order to have a better knowledge of the user and to offer him/her enhanced services. One of the best known examples is that of e-commerce sites such as "Amazon.com," which makes recommendations to the user by making a cross analysis of what other users buy (e.g., if most users interested in cars also ask for sports books, the fact that you ask for a sports book means that you will probably also be interested in car books).

Such an inference requires that the knowledge of user activity (who, when, what) be recorded in devices (e.g., web-server traces) in order to discover models as reliable as possible. Over the last few years, a lot of studies have concentrated on the problem of behavior modeling from web navigation traces (see Sections II and VIII). Apart from the identification and the validation of the model itself, some of the questions that remain to be studied, are as follows.

1) The choice of the type of raw traces to be exploited.
2) Among the different kind of available traces, which ones should be privileged?
3) Is the granularity information of these traces important to consider?
4) Does it provide useful information or, on the contrary, too many details, i.e., only adding noise?

Access to web pages such as information having a minimum of structures (books, movie, etc.), or more generally, the access to knowledge, results from a complex cognitive behavior. We think that the study of this behavior and its relation to the access to structured data (here, web pages) can be very useful in modeling user activity. In this context, we have compared two aspects of access to web information. The first one is direct and relates to a global document materialized by its URL [1]. The second is more indirect and concerns themes more granular and is represented by the most frequent keywords associated with a given URL. From this point of view, we may wonder if there is any relation between these two modes of accessing information. We thus carried out an analysis of user access by using an operational proxy cache (that memorizes both the user activity and the consulted content) over a long period, in order to determine the access regularity level. One other question we have tried to answer is whether user global behavior can be anticipated by analyzing a limited-time-period behavior. In other words, is there a temporal consistency in user accesses? Our paper will provide an answer to these questions.

The organization of the paper is as follows. Since traces are important material for us, we first give some background on the possible origin and use of traces. In Section III, we present the context of our experimentation and we describe our method of study of the access regularities. Then, in Section IV, we describe the principle of self-similarity, which we consider as a key descriptor of the user behavior. In Section V, we present our results obtained with a batch approach of the access

Manuscript received February 10, 2003; revised March 23, 2004. This paper was supported by France Telecom R&D. This paper was recommended by Associate Editor J. Miller.

L. Lancieri is with France Telecom R&D, Caen 14000, France (e-mail: luigi.lancieri@francetelecom.com).

N. Durand is with the University of Caen, Bd Maréchal Juin, Caen 14000, France (e-mail: ndurand@info.unicaen.fr).

Digital Object Identifier 10.1109/TSMCA.2006.859095

regularity, and a temporal analysis. Section VI gives a comparative study of user behaviors. In Section VII, we cluster users in communities of interest. In each case, the results obtained with URL-based traces and those obtained with the keywords are compared in parallel. Section VIII is dedicated to the state of the art, and we conclude in Section IX.

II. Origins and Use of Traces

Devices or user-activity traces can be obtained at almost every level of the network. We may split up these sources of traces into three categories. The question is always the same: to identify mainly "who" (device or user) made "what" transaction and "when." The differences between the three categories deal with the precision of the information obtained for the "what" question and with the facility to obtain data.

The first category involves low-level network devices including hubs, routers, or switches, for example. These traces allow physicals properties of networks to be measured, such as the number of information packets [transmission control protocol (TCP) segments, IP datagrams, etc.] over a period or ranking users according to the quantity of their downloads, etc. Numerous studies based on these kinds of traces have been carried out in order to study the statistical behavior of the networks. Since these devices did not directly manage application-level data, they deliver little information on semantic aspects of user behavior. It is possible to see that a server has a certain number of visits each day, but it is difficult to say that one user is interested in cars and another in sports, for example. The second category relates to higher level information sources, such as a web server, or an end-user PC. The information available at this level, including the consulted content itself (e.g., keywords from documents downloadable from the server), allows more details on the semantic aspect of the user behavior to be collected. For example, web-server traces allow the content or the topics mostly appreciated by the users to be known. The main drawback of this approach is that the results of the analysis correspond only to restricted sources (only the content of the topics stored in one web server or the limited activity of one user). It is possible to make such a study on several users or servers, but it is not easy to manage and can be intrusive for users (often a sensor/spyware on the user PC is needed). The last category involves mediator devices, such as proxy caches that enable all web servers of the Internet (i.e., all possible topics) to be potentially targeted by a large audience (i.e., several hundred people of a local area network). Since these devices also store the contents downloaded by the users, it becomes possible to obtain more accurate details on user behavior with reasonable statistical accuracy. This solution still has the drawback of not catching the entire consulted document (only the cacheable part). However, this part remains about 70% of the whole document, which corresponds to a high quantity of data. Our study is managed within the context of this category.

Finally, and from a more general point of view, it also important to keep in mind that, unfortunately, even if laws exist, traces can lead to some misuses regarding individual privacy (see [2] for a general discussion on the ethics and legal aspects of trace uses).

TABLE I
Two Records of Acces.Log File From a SQUID Proxy Cache

953706308.914 AA.BB.CC.DD P_HIT/200 1886
GET **http://www.server1.com/page1.htm** - DEFAULT_PARENT/
139.100.0.37 **text/html**
0/3276/3276 1071359/238/227409
4426/6920/7168 11594/1059765/-2 -2/-2/-2 -2/-2/-2

953706309.563 EE.FF.GG.HH TCP_MISS/404 350
GET **http://www.server1.com/image.gif** - DEFAULT_PARENT/
139.100.0.37 **image/gif**
0/33752/33752 774494/255/26354
35004/7256/7492 42496/731998/758200 -2/-2/-2 -2/-2/-2

III. Context of the Experimentation

To feed our study, we used the log files and the contents of two operational proxy caches located at France Telecom R&D (Caen, France). The collected data cover a period of 17 months, concerning 331 users. The number of users is not very high, but the long time span covered by the data gives a reasonable statistical consistency. The total number of queries is 1 510 358, corresponding to 392 853 different objects (a given object can correspond to several queries). The log files were purified to keep only the queries corresponding to a textual object. We justify this approach from a cognitive point of view, since the textual pages are strongly related to the intentionality (explicit step) of the users (the multimedia objects included are rather, in terms of probability, a consequence of a primary access, which is generally textual) [3]. Moreover, the text is easier to characterize in semantic terms than images or video. Lastly, the purpose of this restriction is also to reduce the set of queries, and consequently, to facilitate the exploitation of the data. The textual objects correspond to 4563 queries per user, and 1186 different URLs consulted per user, i.e., approximately 8.6 queries per user and per day and 2.2 different objects consulted per user and per day (84 750 different words). After the filtering stage, the objects corresponding to these queries were inserted in a PostgreSQL database using JAVA/Java Database Connectivity (JDBC) Application Program Interface (API), which facilitates processing of raw data. Before being refined, the raw traces had the format described in Table I.

The interpretations of all the information contained in these logs are beyond the scope of this paper (see [4] for more details). What we need to understand is that a record is added on the log file for each user's request. For example, if a user downloads a web page with five images, the proxy cache will adds six records in this log file. We use only four fields (in bold) in each record. The first one is a time stamp used for the temporal study, and the second (e.g., AA.BB.CC.DD) is the IP address of the computer (in our study, one computer relates to one user, but it is not always the case). The two others fields are the URLs of the downloaded content and, finally, the Multipurpose Internet Mail Extension (MIME) type of the content. We first sort all the lines based on the MIME type in order to select only textual content (text, html, pdf, ps, etc.). Other lines, related to nontextual content (e.g., line 2 in Table I) and URLs linked to the major search engine or portal (based on a list), are deleted. An automatic method of identifying little informative and nonautonomous content (little text and a lot of links such as a search engine response, portal, etc.) is to compute the density of the

TABLE II
REFINED TRACES

Time stamp; IP address; URL; word-1, word-2, word-3,..., word-10.

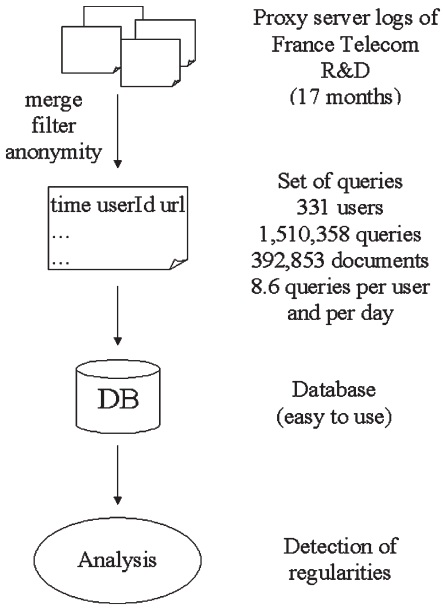

Fig. 1. Context of the experimentation.

TABLE III
PSEUDOCODE OF THE LOGIC FOR COLLECTING DATA

```
For each user access recorded in the proxy log file.
   If the access correspond to textual data
      Then retrieve the data from the origin Web server
         If the data are available
         Then
         Begin
            Extract 10 most clean keywords
            Save in the database: Kwords, User ID, URL, Time
         End if
      End if
   End if
End for
```

web page [3], that is, the ratio between the numbers of the links in the page towards external pages divided by the number of words on the page. Then, we get the content corresponding to the URLs from which we extract the most frequent keywords. In this study, we focused only on the ten most frequent keywords, but the impact of the number of keywords has also been studied [5]. The conclusion of this latter study is that ten words give sufficient precision, and this precision is degraded by adding too many words. After carrying out all these operations, we obtain refined traces (see Table II).

Fig. 1 illustrates the general processing of the data, and Table III details the data collecting.

As we said, the two types of symbolic series that we studied are the consequence, on different levels, of an intentional action of knowledge acquisition by users. In Fig. 2, the Su series first consist of the consulted URLs, the second series Sw, the included significant corresponding words.

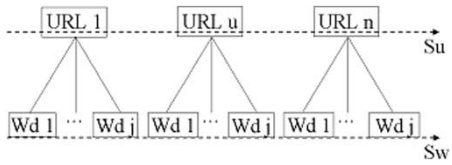

Fig. 2. Two symbolic data series of which we studied the evolution.

In the first step, we studied the level of regularity on different access sequences, ignoring the chronology (see Section V-A). In this step, the question is mainly how the users returned to the same web sites (URLs) or to the same themes (keywords). Then, we carried out a study to determine the level of regularity of these series in time, in order to study the temporal consistency (see Section V-B.). Whereas these two sections try to characterize the accesses, Sections VI and VII are dedicated to the characterization of the users. All the studies are based on the analysis of redundancies. Indeed, the redundancy of the queries can be interpreted as illustrating user behavior. For example, a monolithic behavior (accesses concentrated on few sites) will have a strong global redundancy, whereas a more dispersed behavior (a lot of different URLs) will have a weak redundancy. The global redundancy (gR) is the complement ratio between the number of unique URLs (or keywords) and the number of total queries (or keywords) over the total period of consultation

$$\mathrm{gR} = 100 - 100\left(\frac{\#\text{Unique items}}{\#\text{total item}}\right). \tag{1}$$

We see that gR takes the value 0 if all elements of a set are different (no redundancy). To the contrary, a high redundancy implies that all elements are similar. The partial redundancy (pR) is measured between two sets of queries and evaluates the ratio of items (contained in the requests) from one set present in the second set. This metric can be very useful in evaluating user-activity forecast methods (see Section VIII)

$$\mathrm{pR}(Q1, Q2) = 100\left(\frac{\#\text{ items } Q1 \in Q2}{\#\text{ items } Q1}\right). \tag{2}$$

The whole set of queries is organized chronologically and cut out in sequences of equal size, noted T (which corresponds to a percentage of the total number of queries). For example, let us assume that for user U1, we have a set of queries divided into four sequences $Q1$–$Q4$ with the following requests:

$$Q1 = A, B, B, B$$
$$Q2 = C, C, A, B$$
$$Q3 = A, C, B, B$$
$$Q4 = A, B, B, D.$$

The global redundancy (gR) would be equal to $100 - (100 \times 4/16) = 75\%$, whereas the partial redundancy (pR) between the sequences $Q2$ and $Q4$ would be equal to $100 \times (2/4) = 50\%$.

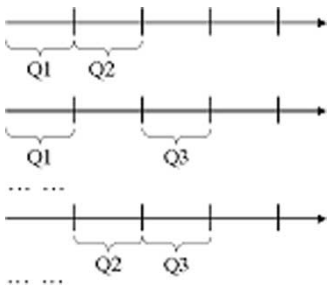

Fig. 3. Process of analysis.

TABLE IV
PSEUDOCODE OF THE ANALYSIS PROCESS

```
For each user
For Sequence size T from 1 % to 20 % of |S| (i.e. the number of
items in the user's stream S)
    Number of sequences N=|S|/T
    For sequence position I from 0 to N/2
      For sequence position J from I+1 to I+N/2
        Compute pR (QI, QJ)
      End For J
    End For I
End For T
End For Each user
```

We illustrate the process of temporal analysis (pR) in Fig. 3. For the first sequence, noted $Q1$, we compute pR, with the following sequence noted $Q2$; this corresponds to a spacing of $\Delta = 0$. We reiterate this process until all the queries are covered (sequence Qi compared to $Q(i + N/2)$, where N is the total number of sequences per user). Fig. 3 and Table IV describe this process.

In order to illustrate the relationships between gR an pR, we can also say that gR offers a macroscopic and general view of access redundancy, whereas pR, as a microscope, enables us to see a lot of details over time, but without an overview of the context. Therefore, gR and pR are two complementary measures that can be computed per user and on average for all the users.

IV. Self-Similarity: A Consistent Model for User Profile and Interactions

As we will see in the state of the art, statistical tools or models are often used when the quantity of data increases, even if semantic description is the main goal of the study. For example, in document indexing, average or standard-deviation values are often integrated into more complex metrics. The interest of taking statistical value into account is to reduce the descriptive space of a phenomenon. Definitely, too high a reduction can be damaging and causes a high loss of information. For example, the average can be useful when the variations are not too high, but can be useless when the process varies, for example, in an exponential fashion. One of the challenges in data analysis is to identify adequate description value, consistent with the goal of the measure. This difficult task can be made easier by first discovering an underlying model. In our previous example, the knowledge that our process varied in an exponential fashion is very helpful because we can describe its behavior by the single value of the parameter of an exponential law. This parameter, as a single number, is a reduced description, but much more accurate than the average value. Our interest in self-similarity comes from the fact that it models a lot of natural behavior that can be approached by a single value (parameter of the law).

The subexponential distributions are the most known of the self-similar laws (e.g., Pareto, log normal, Weibull). Several characteristics differentiate these distributions from the more traditional ones (e.g., Gauss, exponential, Poisson). For example, its representation in log–log coordinate corresponds to a straight line (characterized by its slope $-a$). Self-similarity expresses that a set of data exhibits the same characteristic on different scales. In the case of time series, this characteristic implies the dependence between the short and the long term (temporal consistency). Indeed, whatever the scale of time, the characteristic (e.g., variability) is the same. These kinds of distributions also have an infinite variance if $a \leq 2$, and an infinite average if $a \leq 1$. When "a" decreases, an important part of the distribution is concentrated in its tail, which justifies its name of "heavy tailed." Thus, a random variable following a subexponential law can reach very high values with a high probability.

Since the beginning of the 20th century, several authors (Mendelbrot, Zipf, Hurst, Whittle, etc.), have studied and applied self-similar properties to different scientific fields. A very easy way to describe self-similarity deals with Zipf's generalized law, which is a reasonable description of web traces [6] (see Section VIII for application). Zipf, in 1929, applied a simplified version of this law $(a = 1)$ to the relations between words in a given text, and thereafter, it was also applied in a sociological context [7]. The origin most usually evoked to justify this law is "the human choice" or "the intelligent choice" (e.g., choice of the words in a text). Many studies put forward this law to describe the popularity of the documents on the Web, or to predict the level of access to a document according to its rank of popularity. Recently, some authors (e.g., in [8] and [9]) have put forward the hypothesis that it is largely influenced by the type of computer file systems (relationship with memory) and user behavior. Self-similarity was used to realize simulators of web caches [10], information research [11], or storage systems [12].

V. Results

We present in this section the main results of our study, beginning with the characterization of the accesses, and then its impact on the user and community behavior characterization.

A. Batch Analysis

For each user, we computed the gR of their accesses and the average of the pR (with $T = 5\%$, i.e., 17 months of activity for one user is split into 20 sequences). The results are presented in Figs. 4 and 5, where each point represents one user with his gR in ordinate and his average pR in abscissa.

We remark that gR and pR move in the same direction according to a geometrical law, with an increasing dispersion

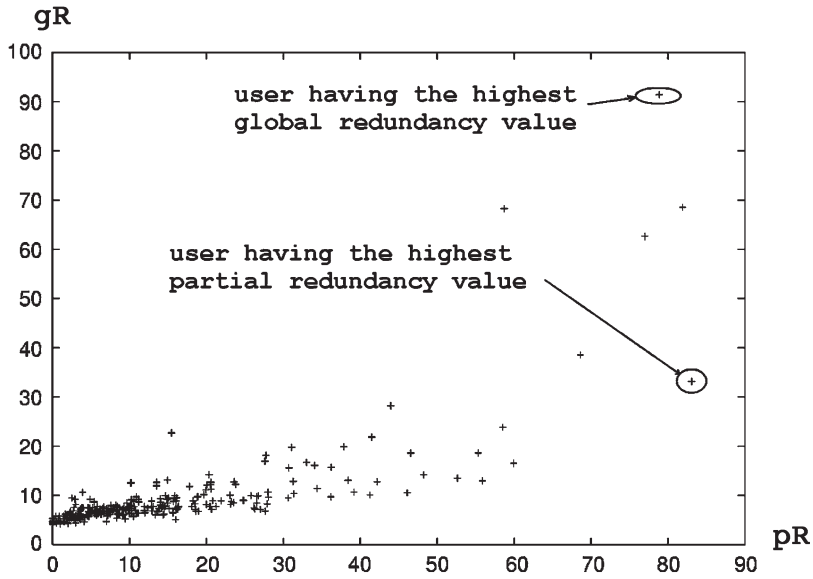

Fig. 4. Global redundancy according to partial redundancy (URLs).

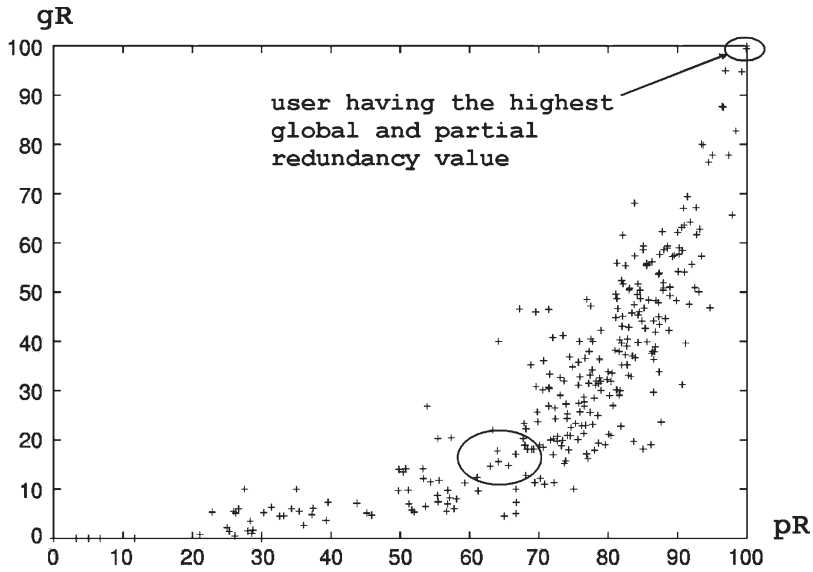

Fig. 5. Global redundancy according to partial redundancy (keywords).

according to the redundancy (more evident in the URLs than in keywords). In the extreme cases (all identical URLs/keywords or all different URLs/keywords), the level of global or partial redundancy tends to be identical [point (0,0) and (100,100) in the graphs]. This is logical because, if, for example, all the (macroscopic) (gR) accesses are made randomly (gR tends to approach 0), the probability that pR (microscopic view) would also reveal random behavior is high. The same logic can be applied to the other extreme behaviors. We also see that gR increases less rapidly than pR, for low values of pR. If we consider that these graphs show the relationship between both the level of macroscopic (gR) and microscopic (pR) consistency, it seems that the access consistency is much more rapidly perceptible at microscopic level than at a macroscopic one. For example, the users in the circle (in Fig. 5) seem to have almost random access considering the low value of gR, whereas the same users have high access consistency considering a high value of pR.

Let us also note that gR of the keywords is well distributed over the whole spectrum of the values of pR, whereas for URLs, it is more concentrated on the low values of pR. This can be explained by the fact that the number of possible keywords is small, considering the number of possible URLs (i.e., several different Web pages can use exactly the same words differently organized). The number of keywords rises quickly, but it is rapidly restricted by the limits of the language. The number of URLs is potentially much higher (combination of the keywords). The two sets increase very differently according to the time of consultation (see Fig. 6). Taking into account the respective size of the two sets, it is much more probable to find redundancies for the words than for the URLs.

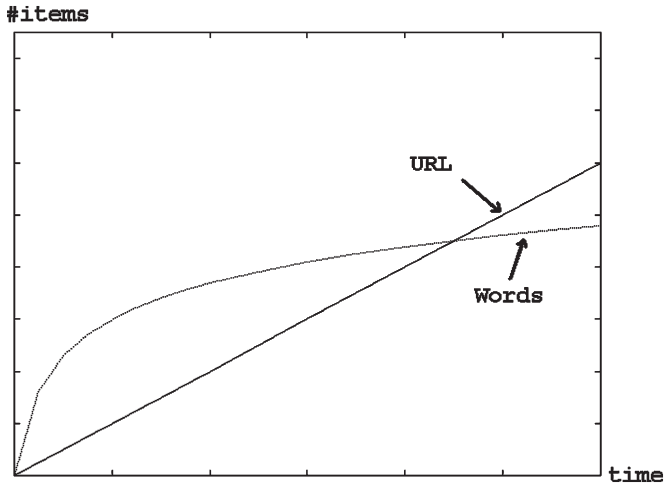

Fig. 6. Evolution of the number of consulted URLs and keywords according to the time.

However, even if the distribution of data is different in keywords than in URLs, the dependence between pR and gR seems to follow the same law in both cases. Since the distribution is more homogeneous in the keywords, we can try to identify an underlying law. We obtain the following relation between pR and gR by analyzing the data of Fig. 5

$$\mathrm{gR} = 100\left(\frac{\mathrm{pR}}{100}\right)^{\alpha}. \tag{3}$$

The value of alpha seems to be a constant near 4 and provides a good correlation with our data (0.89 for the keywords, i.e., Fig. 5, and 0.83 for the URLs, i.e., Fig. 4).

Figs. 7 and 8 show the relation between words and URLs, respectively, in terms of gR and pR. Each point corresponds to a user, and represents his average measures. The association with a subjacent law is not clear (a geometric law like (3), with alpha equal to 0.35, gives a correlation coefficient equal to 0.66). A good correlation would imply that we can accurately express (i.e., that there is a strong link) the redundancy of URLs from those of the keywords and vice versa. This is not really the case here, but it is an approach. We also remark that the relation between keywords and URLs is clearer (coherence) in the case of strong gR or weak pR.

This will be clearer in the next section but, with the figures represented by graphs in Figs. 4 and 5, and 7 and 8, we have begun to observe the accesses consistency. Indeed, if the accesses were made randomly, the partial redundancy would be statistically very close to the global one (equal distribution). From a cognitive point of view, we can make a link between this consistency of the accesses and the causality on one hand, and human intention on the other hand.

B. Temporal Analysis

In this section, we studied the value of the user pR according to the space between the sequences of queries (see Fig. 3). The results are presented in Figs. 9 and 10, respectively, for URLs and keywords. We have, in abscissa, the spacing between two sequences (Δ), and in ordinate, the average of pR of all the users. Whereas in the last section, we analyzed the global impact of pR, here, we varied the size of the sequences to be studied ($T = 1\%, 2\%, 3\%, 5\%, 10\%, 20\%$) in order to better take into account temporal consistency of accesses. For example, let us explain how we obtain the point "A" in Fig. 9. We split the flow of requests (consulted URLs) into 100 sequences ($T = 1\%$). Then, we compute, for all the users, the $\mathrm{pR}(q1, q2)$, where $q1$ and $q2$ are distant from 30 sequences ($\Delta = 30$). We compute the average of all these pRs in order to obtain the point "A."

We remark that the maximum value for pR is obtained when the spacing is smallest ($\Delta = 0$, i.e., two consecutive sequences), marking a strong temporal consistency. That means that the closer the two sequences are in the time, the more the accesses are redundant. This maximum value is not very different according to the size from the studied sequences. It varies between 23.6% and 25.3% (the same value for keywords is about twice as high).

We also notice that the level of pR according to time, separating the two compared sequences, follows a subexponential law. Indeed, after logarithmic transformation, the increase in the redundancy according to the increase in the space of time separating the sequences is constant.

We applied a logarithmic transformation of the curves in Figs. 9 and 10, and obtained Figs. 11 and 12. The lines of Figs. 11 and 12 have a coefficient of regression equal to 0.99, which shows an excellent correlation.

The subexponential law binding these two parameters is expressed in the following way. Let T be the size of the studied sequences (expressed as a percentage of the total queries), pR be the partial redundancy, Δ be the spacing between the two studied sequences and pRmax be the maximum value of the pR, and k be a constant. We have the following relation, where $(-kT)$ is the slope of the straight line

$$\log(\mathrm{pR}) = -kT\log(\Delta) + \log(\mathrm{pRmax}).$$

After some transformations, we obtain

$$\mathrm{pR} = \mathrm{pRmax}(\Delta)^{-kT}. \tag{4}$$

If we consider that the curves of Figs. 11 and 12 are expressed by the relation $y = ax + b$, where a is the slope and b the ordinate at the origin, we can make the following remark when $T\%$ rises. For URLs, a varies and b is quasi-fixed, whereas the opposite occurs for keywords.

We can check with Fig. 13 that the slope of Fig. 11 moves linearly according to the sequence size T. Previous observations mean that we can determine the value of pR, knowing the size of the studied sequences and the value of the spacing. This means that the behavior of the function pR is independent of the user whose behavior intervenes only with pRmax value. Thus, the temporal consistency linked to $\mathrm{pR} = f(\Delta, T)$ is user independent because it is directly and mainly linked to the size and spacing between the sequences. For each user, the model ensures with a high probability that for a specific delta, their pR will fall along the plot (see Fig. 11) corresponding to the

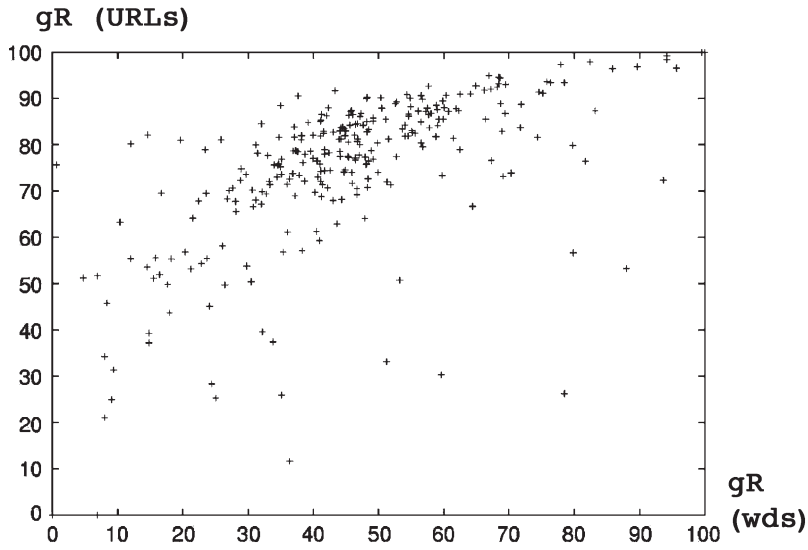

Fig. 7. Average gR of URLs according to average gR of keywords.

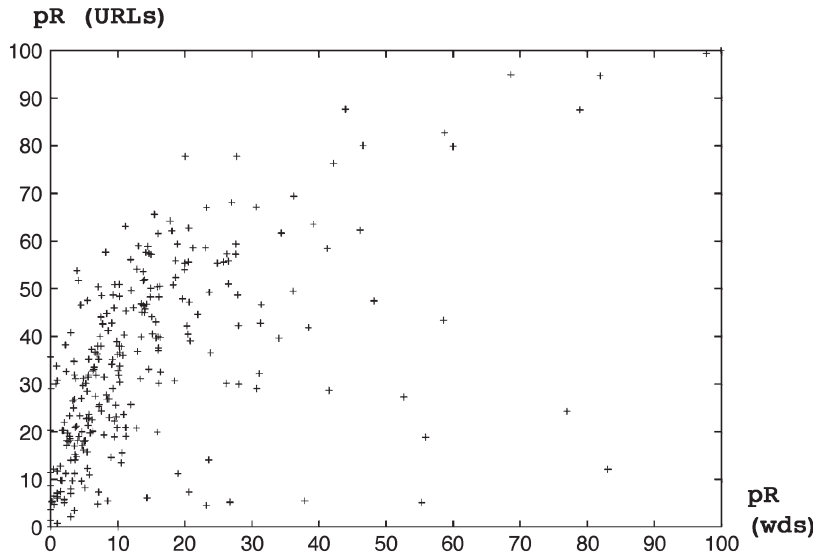

Fig. 8. Average pR of URLs according to average pR of keywords.

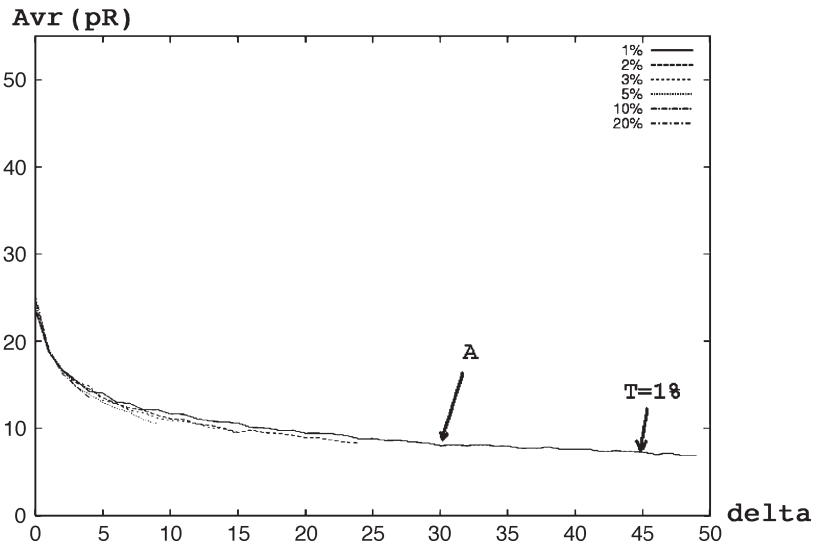

Fig. 9. Average of the partial redundancies according to the space between the sequences (URLs).

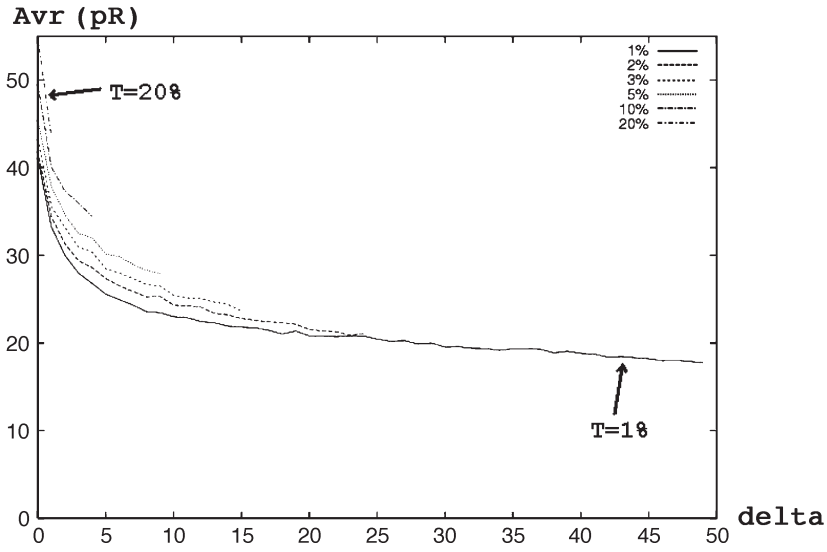

Fig. 10. Average of the partial redundancies according to the space between the sequences (keywords).

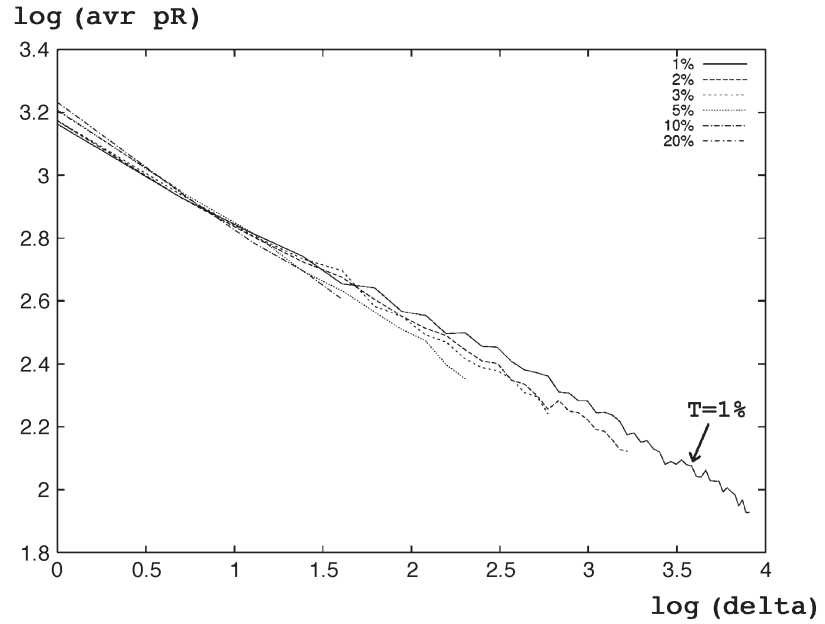

Fig. 11. Logarithmic transformation on the curves of Fig. 9 (URLs).

value of T (for which both T and Δ are user independent). Of course, this is only true in terms of probability, since we have used the average pR. The general law we can highlight is that, whoever the user is, the temporal consistency falls rapidly in a subexponential fashion when two sequences move away from each other.

VI. Comparative Analysis of User Behaviors

The distribution of the values of the maximum pR among the users is represented in Figs. 14 and 15 (numbers of users in ordinate versus class of pRmax values in abscissa, $T = 5\%$). Let us remark that this parameter is more homogeneously distributed among URLs than among keywords. For a pRmax value between 80% and 100%, we have with the URLs, 90 users (27% of the users), and with the keywords, 190 users (58% of the users). We remark clearly that there is a small group of users having a partial redundancy relatively high. The population, having a maximal pR under 50%, represents 45% of the total users with URLs and only 15% with keywords.

Following what we have seen in Figs. 4 and 5, this is another clue that implies that URLs are less discriminate parameter than words. Indeed, a low partial redundancy is not informative (equivalent to random, equiprobability, or inconsistency).

VII. The Community of Interests

In order to validate the potential of the two different kinds of traces, we clustered the users, based on URLs and on keywords. For us, a community of interest is a cluster of users

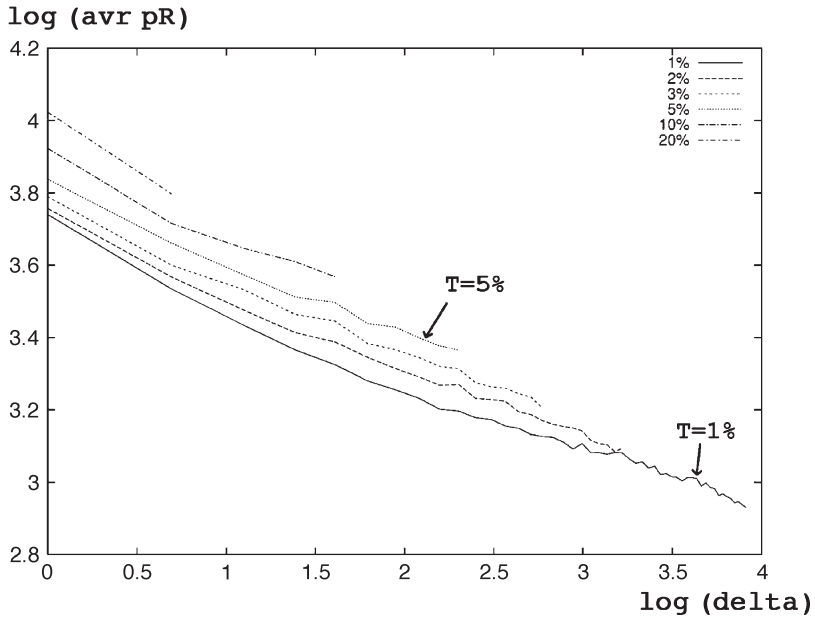

Fig. 12. Logarithmic transformation on the curves of Fig. 10 (keywords).

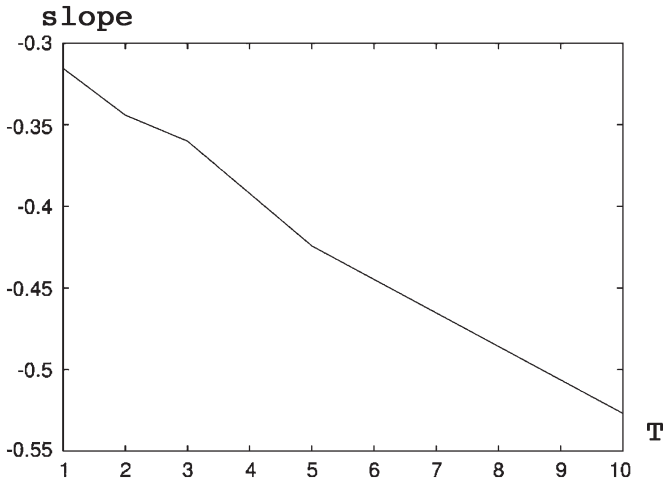

Fig. 13. Evolution of the slope according to T (URLs).

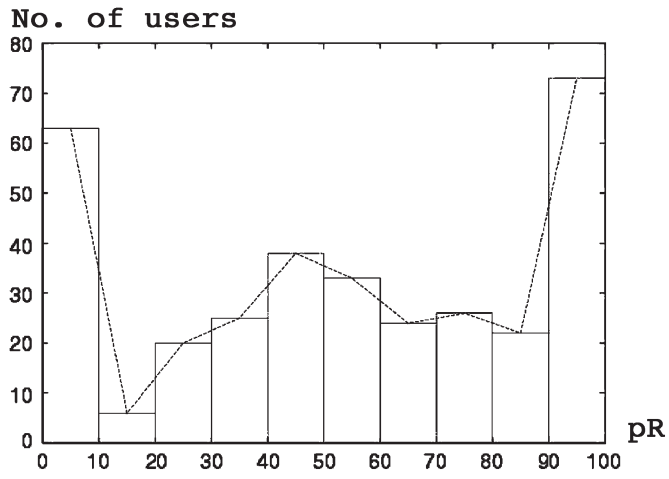

Fig. 14. Distribution of the maximal partial redundancy among users (with URLs).

automatically detected according to their shared interests. We used two kinds of categorization algorithms: an algorithm of "hard clustering" producing a partition, and an algorithm of "soft clustering" discovering overlapping clusters. The first method allows a user to be only in one cluster, whereas in the second one, the user can be in several clusters.

The first algorithm is a method of hierarchical agglomerative clustering (HAC) [13]. From a matrix of distance between users, this method successively creates several clustering (partitions of users) by agglomerating the most similar groups. A measure of quality allows the best clusters to be selected. This measure favors groups having a good intracluster similarity and a good intercluster dissimilarity. Given that we have qualitative data (URLs and keywords), the methods of computation of distance between two users are limited. To evaluate the similarity, it is possible to use several coefficients: matching, Jaccard, Dice, or cosine [14]. We chose the cosine coefficient, which gives the best results, to compute the matrix of the distances

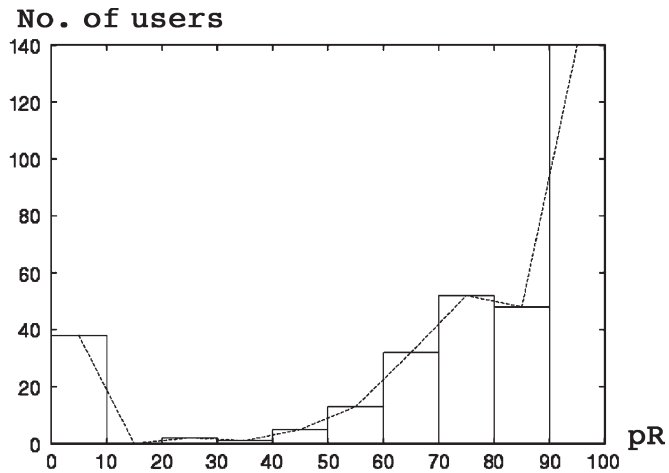

Fig. 15. Distribution of the maximal partial redundancy among users (with keywords).

TABLE V
AGGLOMERATIVE CLUSTERING OF THE USERS (HAC)

	URLs	keywords
No. of clusters	280	91
Avr no. of users per clusters	1.14	3.5
Max no. of users in a cluster	8	32

TABLE VI
DISCOVERY OF OVERLAPPING CLUSTERS (ECCLAT)

	URLs	Keywords
minfr	35%	35%
No. of clusters	3	25
Max. no. of shared items	1	16

between users. The cosine coefficient between two users $u1$ and $u2$ is defined by

$$D_{\cos ine}(u1, u2) = 1 - \frac{|u1 \cap u2|}{\sqrt{|u1| \times |u2|}}.$$

$u1$ is the set of items (URLs or keywords) corresponding to user 1. Thus, $|u1|$ is the number of items.

We have noticed that the clusters, obtained by using the URLs, are not really useful (see Table V). We have 280 clusters for 319 users (almost one user per cluster!). This comes from the fact that while being based on the URLs, there is a very weak similarity between the users. On the other hand, by using the keywords, we have fewer clusters and a better distribution of the users. With the URLs, we have 261 clusters containing only one user and only two clusters with more than five users (representing 15 users). With the keywords, we have only 39 clusters with only one user and 15 clusters with more than five users (representing 192 users).

In the second experiment, a user can be present in several clusters, making it very useful for certain applications (possibility to have several points of view). Let us take for example a user who is interested in "fishing" and in "cars," then, it is completely possible and comprehensible that he is present in a cluster corresponding to "fishing," and also in another cluster corresponding to "cars." This shows his different centers of interest.

To discover such clusters, we used Extraction of Clusters from Concepts LATtice (ECCLAT), which we developed and presented in [15]. It extracts the interesting clusters from the frequently closed item-sets lattice. A closed item set checks an important property for clustering: It gathers a maximal set of items (URLs or keywords) shared by a maximal number of users. In other words, this allows the capture of the maximum amount of similarity. A closed item set associated with the corresponding set of users is seen as a cluster. We selected the most meaningful clusters by using an evaluation measure and a greedy algorithm. The method has two parameters: minfr (minimum frequency), corresponding to the minimum number of users in a cluster, and M, corresponding to the minimal number of different users between two selected clusters. The parameter M allows the control of the overlapping of users. Here, we set M to the minimum (i.e., 1), because we do not wish a pseudopartition of the users, but a set of overlapping clusters. The results are represented in Table VI.

We remark that for a threshold of 35% (111 users), which is enough, we find 25 clusters with the keywords, whereas three with the URLs. Moreover, we observe that the greatest number of shared items for a cluster is 1 for the URLs, and 16 for the keywords. This shows that there is a good correlation between the similarity of users behavior by using the consulted keywords. Finally, and as supposed, the use of keywords for identifying and characterizing communities is more efficient than the use of URLs.

VIII. RELATED WORKS

Most of the studies on user behavior are based on the traffic measure and statistics from traces. Abdulla *et al.* [16] studied the shared user behavior by identifying the invariants (file types, etc.) in a collection of ten traces representing traffic seen by different proxy caches, each representing a community (education, business. . .). Balachandran *et al.* [17] analyzed user behavior and network performance in a public area wireless network, in order to characterize wireless users in terms of a parameterized model for use with analytic and simulation studies. They analyzed user behavior from a statistical point of view (user session duration, data rates, application popularity, user mobility). Other studies on the variability of the web user behavior were made specifically by taking into account the semantic aspect of the queries [18]–[20].

Concerning the use of user behavior characterization, it is interesting to consider the research work links to the predictability of the user activity (recommendation systems, prefetching, etc). The prefetching systems [21]–[23], for example, aim at identifying the consulted URL in advance, so it can be downloaded by an artificial agent and be locally (rapidly) available for user consultation. As intuition suggests, it could be very difficult to forecast user behavior since human activity is mainly nondeterministic. However, even if it is true, there has been a lot of academic and industrial work in this field. These studies are interesting for us because a reasonable efficiency requires an accurate characterization and adequate models of human behavior.

An approach consists in determining a subset of the links of the page in the course of consultation, classified according to a degree of growing predictive estimation. In [21], and [24]–[26], the decision making must be very fast and seldom makes it possible to use complex forecasts methods. Some studies [21], [27] show that if all the links were prefetched, the probability of forecast would reach 69% (45% for the ten best links and 20% for the best 5). An analysis of URLs [28], [29] or of keywords [26], [27], [30] contained in the past consulted documents is often used as a base for future activity. This profile of future consultation can be treated by algorithms based on clustering [31], learning [32], [33], decision trees [34], or hidden Markov chains [35], [36] in order to obtain rules of decision to identify future activity.

Whereas the preceding methods aim at selecting a reduced subset of known resources to evaluate those that will be consulted soon, others rather aim to discover resources potentially useful [30], [37]. These forecasting techniques correspond to the needs for recommendation systems that are assimilated to information filtering systems because the ideas and the methods are very close [38]. There are two types of filtering: content-based filtering and collaborative filtering. Content-based filtering identifies and provides relevant information to users on the basis of the similarity between the information and the profiles [37], [39], [40]. Collaborative filtering finds relevant users who have similar profiles, and provides the possibility of sharing documents they like [41], [42].

From the model point of view, the notion of self-similarity has also been studied in different fields. Leland *et al.* [43] and Paxson [44] showed, respectively, with LAN and WAN traces, that the forecast of traffic was much more reliable by taking into account a self-similar distribution data stream rather than a Poisson one. Whereas in the case of Poissonian or Markovian streams, the level of variation or disparity of the queries tends to decrease in time, the real traffic reveals stability on various scales of time. Crovella and Bestravos [9] underlined the self-similarity in the level of activity on the Web, for example in file transmission time or in the level of the server inactivity, generally distributed in a subexponential way. Some authors also linked part of this latency time to what they call "Think time" (the time the user thinks before taking action). For Pitkow and Recker [45], the subexponential nature of the "the think time" distribution takes its roots in the cognitive functioning of the human being.

IX. Conclusion

We studied the regularity of the user access using a method of analysis of the global and partial redundancy of URLs and of keywords. Apart from the statistical information on these accesses, we can emphasize some conclusions to better characterize user behavior. One of the interests of this paper is to study the phenomenon of user characterization and behavior by a compared approach of several of its parameters (macroscopic/symbolic: URLs and microscopic/granularity: keywords). The study shows that the dependence between the two parameters is well correlated.

The second conclusion we can give is that, even if the relationship between URLs and keywords distribution is not clear, keywords allow a more powerful description than URLs, even if keywords require more processing capacity. This is clear, considering the clustering study. We can highlight laws describing the temporal coherence of the access and the interesting model capacity of self-similarity. This internal logic suggests laws driven, at least, by the rational side of the user behavior.

We are far from having discovered a general model of user behavior (supposing that it is possible). However, we showed that there are partial models binding, for example, user behavior to redundancies. We also showed that the temporal access coherence is independent of the users. In future work, we will perform other experiments with data containing more users, and we will focus on the model based on keywords to investigate predictability capacity. One of the perspectives of our work will be to investigate the feasibility and the performance of the forecast of user activity [26], [38]. As we have seen in the state-of-the-art section, there are a lot of studies in this field, which represent a big challenge from a scientific point of view, as well as from a commercial one.

Acknowledgment

The authors wish to thank their colleagues for their help and constructive comments for this study. They would also like to thank S. Lenoir and V. Murphy for their helpful recommendations on the English used in this paper.

References

[1] N. Durand and L. Lancieri, "Study of the regularity of the users' Internet accesses," in *Proc. Intelligent Data Engineering and Automated Learning (IDEAL)*, Manchester, U.K., Aug. 2002, pp. 173–178.

[2] L. Lancieri, "Reusing implicit cooperation, a novel approach to knowledge management," *TripleC Int. J.*, vol. 2, no. 1, pp. 28–46, 2004.

[3] ——, "Memory and forgetfulness: Two complementary mechanisms to characterize the various actors of the Internet in their interactions," Ph.D. thesis, Dept. Comput. Sci., Univ. Caen, Caen, France, 2000.

[4] *SQUID Proxy Cache Log Files*. [Online]. Available: http://www.tenon.com/support/webten/papers/squidlog.shtml

[5] L. Lancieri and N. Durand, "Evaluating the impact of the user' profiles dimension on its characterization effectiveness: Method based on the evaluation of user' communities' organization quality," in *Proc. IEEE Int. Symp. Computational Intelligence Measurement Systems and Applications (CIMSA)*, Lugano, Switzerland, Jul. 2003, pp. 130–134.

[6] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "On the implications of Zipf's law for Web caching," in *Proc. 3rd Int. WWW Caching Workshop*, Manchester, U.K., Jun. 1998, pp. 1–14.

[7] B. Mandelbrot, *The Fractal Objects*. Flamarion edition, 1984.

[8] V. Almeida, A. Bestravos, M. Crovella, and A. de Oliveira, "Characterizing reference locality in the WWW," in *Proc. Parallel and Distributed Information Systems (PDIS)*, Miami Beach, FL, Dec. 1996, pp. 92–103.

[9] M. Crovella and A. Bestravos, "Self-similarity in World Wide Web traffic: Evidence and possible causes," *IEEE ACM Trans. Netw.*, vol. 5, no. 6, pp. 835–846, Dec. 1997.

[10] N. Saillard, "Simulation for cache mesh design," in *Proc. TERENA-NORDUnet Networking Conf.*, San Diego, CA, 1999, pp. 1–6.

[11] C. Gomez, B. Selman, and N. Crato, "Heavy tailed probability distributions in combinatorial search," in *Principles and Practice of Constraint Programming*, G. Smolka, Ed. Berlin: Springer-Verlag, 1997, pp. 121–135.

[12] I. Noros, "A storage model with self-similarity input," *Queueing Syst.*, vol. 16, no. 3/4, pp. 387–396, Sep. 1994.

[13] P. Ronkainen, "Attribute similarity and event sequence similarity in data mining," Univ. Helsinki, Helsinki, Finland, Tech. Rep. C-1998-42, Oct. 1998.

[14] W. B. Frakes and R. Baeza-Yates, *Information Retrieval, Data Structure and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[15] N. Durand and B. Crémilleux, "ECCLAT: A new approach of clusters discovery in categorical data," in *Proc. 22nd Annu. Int. Conf. Knowledge Based Systems and Applied Artificial Intelligence (ES)*, Cambridge, U.K., Dec. 2002, pp. 177–190.

[16] G. Abdulla, A. Edwards, A. Fox, and M. Abrams, "Shared user behavior on the World Wide Web," in *Proc. WebNet*, Toronto, Canada, 1997, pp. 1–7.

[17] A. Balachandran, G. M. Voelker, P. Bahl, and P. Venkat Rangan, "Characterizing user behavior and network performance in a public wireless LAN," in *Proc. Int. Conf. Measurement and Modeling Computer Systems (SIGMETRICS)*, Marina del Rey, CA, Jun. 2002, pp. 195–205.

[18] L. Lancieri, "Description of Internet user behavior," in *Proc. IEEE Int. Joint Conf. Neural Network (IJCNN)*, Washington, DC, 1999, pp. 2514–2519.

[19] ——, "The concept of informational ecology or interest of the information re-use in the company," in *Proc. 3rd Int. Conf. Enterprise Information Systems*, Setubal, Portugal, 2001, pp. 188–193.

[20] S. Legouix, J. P. Foucault, and L. Lancieri, "A method for studying the variability of users' thematic profile," in *Proc. WebNet*, San Antonio, TX, 2000, pp. 906–907.
[21] L. Fan, P. Cao, W. Lin, and Q. Jacobson, "Web prefetching between low-bandwidth clients and proxies: Potential and performance," in *Proc. Joint Int. Conf. Measurement and Modeling Computer Systems (SIGMETRICS)*, Atlanta, GA, May 1999, pp. 178–187. [Online]. Available: http://www.cs.wisc.edu/~cao/publications.html
[22] R. P. Klemm, "WebCompanion: A friendly client-side Web prefetching agent," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 4, pp. 577–594, Jul./Aug. 1999.
[23] T. Palpanas and A. Mendelzon, "Web prefetching using partial match prediction," in *Proc. Web Caching Workshop*, San Diego, CA, Mar. 1999, pp. 1–21.
[24] V. N. Padmanabhan and J. C. Mogul, "Using predictive prefetching to improve World Wide Web latency," *Comput. Commun. Rev.*, vol. 26, no. 3, pp. 22–36, Jul. 1996.
[25] J. Griffioen and R. Appleton, "Reducing file system latency using a predictive approach," in *Proc. USENIX Tech. Conf.*, Boston, MA, 1994, pp. 197–207.
[26] L. Lancieri and N. Durand, "Activity forecast of the Internet users based on the collective intelligence," in *Proc. IASTED Int. Conf. Artificial Intelligence and Application (AIA)*. Innsbruck, Austria, Feb. 2004, pp. 770–775.
[27] B. D. Davison, "Predicting Web actions from HTML content," in *Proc. 13th ACM Conf. Hypertext and Hypermedia (HT)*, College Park, MD, Jun. 2002, pp. 159–168.
[28] Y. Aumann, O. Etzioni, R. Feldman, M. Perkowitz, and T. Shmiel, "Predicting event sequences: Data mining for prefetching Web-pages," in *Proc. Int. Conf. Knowledge Discovery Databases (KDD)*, New York, Aug. 27-31, 1998, pp. 1–11.
[29] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos, "Effective prediction of Web-user accesses: A data mining approach," in *Proc. WebKDD Workshop (WebKDD)*, San Francisco, CA, 2001, pp. 169–176.
[30] P. K. Chan, "A non-invasive learning approach to building Web user profiles," in *Proc. WebKDD*, San Diego, CA, 1999, pp. 7–12.
[31] S. H. Kim, J. Y. Kim, and J. W. Hong, "A statistical, batch, proxy-side Web prefetching scheme for efficient Internet bandwidth usage," in *Proc. Network+Interop Engineers Conf.*, Las Vegas, NV, May 2000, pp. 1–6.
[32] Q. Yang, H. H. Zhang, and T. Li, "Mining Web logs for prediction models in WWW caching and prefetching," in *Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, Aug. 2001, pp. 473–478.
[33] T. I. Ibrahim and C.-Z. Xu, "Neural Net based pre-fetching to tolerate WWW latency," in *Proc. 20th Int. Conf. Distributed Computing Systems (ICDCS)*, Taipei, Taiwan, Apr. 2000, pp. 636–643.
[34] T. S. Loon and V. Bharghavan, "Alleviating the latency and bandwidth problems in WWW browsing," in *Proc. USENIX Symp. Internet Technologies and Systems (USITS)*, Monterey, CA, Dec. 1997, pp. 1–12.
[35] R. R. Sarukkai, "Link prediction and path analysis using Markov chains," in *Proc. 9th Int. World Wide Web Conf.*, Amsterdam, The Netherlands, May 2000, pp. 377–386.
[36] D. Duchamp, "Prefetching hyperlinks," in *Proc. 2nd USENIX Symp. Internet Technologies and Systems (USITS)*, Boulder, CO, Oct. 1999, pp. 127–138.
[37] D. S. W. Ngu and X. Wu, "SiteHelper: A localized agent that helps incremental exploration of the World Wide Web," in *Proc. 6th Int. World Wide Web Conf.*, Santa Clara, CA, 1997, pp. 691–700.
[38] N. Durand, L. Lancieri, and B. Cremilleux, "Recommendation system based on the discovery of meaningful categorical cluster," in *Proc. Int. Conf. Knowledge-Based Intelligent Information & Engineering Systems (KES)*, Oxford, U.K., Sep. 2003, pp. 857–863.
[39] M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill &Webert: Identifying interesting Web sites," in *Proc. 13th Nat. Conf. Artificial Intelligence*, Portland, OR, 1996, pp. 54–61.
[40] H. Lieberman, "Letizia: An agent that assists Web browsing," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Montreal, QC, Canada, Aug. 1995, pp. 924–929.
[41] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "GroupLens: Applying collaborative filtering to usenet news," *Commun. ACM*, vol. 40, no. 3, pp. 77–87, Mar. 1997.
[42] A. Moukas, "Amalthaea: Information discovery and filtering using a multi-agent evolving ecosystem," *Int. J. Appl. Artif. Intell.*, vol. 11, no. 5, pp. 437–457, 1997.
[43] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of Ethernet traffic," in *Proc. ACM Special Interest Group Data Communication (SIGCOMM)*. San Francisco, CA, 1993, pp. 183–193.
[44] V. Paxson, "Fast approximation of self-similar network traffic," Univ. California, Berkeley, Tech. Rep. LBL-36750, Apr. 1995.
[45] J. E. Pitkow and M. M. Recker, "A simple yet robust caching algorithm based on dynamic access patterns," in *Proc. 2nd World Wide Web Conf.*, Chicago, IL, 1994, pp. 1–8.

Luigi Lancieri (M'99) received the Eng. degree in automatic control engineering from the University of Lille, Lille, France, the Ph.D. degree in computer science, in 2000 and the H.D.R degree (accreditation to supervise research) from the University of Caen, Caen, France, in 2004.

Since 1993, he has been with the France Telecom research Labs in Caen, where he currently has his main activity. He also has been teaching as an Assistant Professor at the University of Caen. His research interest covers Web mining and the study of human factors in data networks such as collective intelligence.

Nicolas Durand received the M.S. and Ph.D. degrees in computer science from the University of Caen, Caen, France.

He is currently teaching as an Assistant Professor in Computer Science and Data Network Engineering at the University of Caen. His main research interest covers clustering algorithms and data mining.