ACTIVITY FORECAST OF THE INTERNET USERS BASED ON THE COLLECTIVE INTELLIGENCE

Luigi Lancieri
luigi.lancieri@francetelecom.com
France Telecom R&D
42 rue des coutures 14000 Caen
France

Nicolas Durand
ndurand@info.unicaen.fr
University of Caen, Greyc, UMR 6072
14032 Caen Cedex
France

Abstract

The aim of this paper is to compare two modes of activity forecast of the Internet users for the consultation of Web sites. The two modes exploit the phenomena of collective intelligence by using the combined last activity of a group of users as an indicator of the future activity. The first mode, called passive mode, consists in preserving the most popular objects consulted in the past by certain members of the group by hoping that they will be consulted in the future. The second, called active mode, aims at anticipating the download of the objects which correspond to the foreseeable evolution of the group's centres of interest. We present several metric allowing evaluating the effectiveness of the forecast as well as the results of a comparative study relating to 8 days of activity of 2000 users.

Keywords

User profile, community, forecasting

1 Introduction

Many aspects of the human activity progressed thanks to the forecast techniques. Actually, human being cognitive operation naturally integrates processes of forecast which make it possible to anticipate recurring situations while identifying and then exploiting precursory indices. Such techniques can be useful in decision-making aided processes as heuristic to reduce the space of the possible future states of a system.

Every one knows that all is not foreseeable, so far, but also, that all is not useful to forecast. Actually, it is necessary to evaluate the "energy" necessary to obtain a forecast, the success rate but also the impact of this forecast on the effectiveness of the decision. It is obvious, for example, that certain decisions can be satisfied with an average effectiveness of the forecast, whereas in other case the consequences of a bad forecast can be more critical. For instance, the errors of weather forecasting are less critical for the individual than for the air traffic. This shows clearly that the performance of the forecast not correlated from the context of use of the predicted data does not have much significance. In addition, the human activity being essentially non deterministic, it can appear hazardous to try to foresee it, even if the target and the range of the forecasts are restricted.

In the context of these remarks, the aim of this paper is to compare two approaches of Internet users' web pages consultations forecast. To be more precise, our objective is not to foresee in advance all the pages which will be consulted but rather to identify a subset of pages (even weak) with the maximum of guarantees that they will be useful. Our validation process of the quality of the forecast is to verify on real consultations that a more or less large proportion of the predicted pages were actually consulted. The two strategies of forecasts (passive and activate modes) that we compare, are based on the implicit phenomena of collective intelligence. The idea is to use the joint activity of some users to make profiting to the others. The passive approach which is only based on the synergy of the individuals, is distinguished from the active approach where artificial agents are used to optimize the effects induced by the passive approach. We will show that the passive method that allows obtaining a probability to reuse a previously downloaded document near than 30 % (average hit rate of a regular proxy-cache) can be optimized through the active method. In Sections 2 and 3, we describe the approaches and the validation method. Next, we present some results. Before concluding, we present related works.

2 Methodology

2.1 Active approach

The active approach that we propose is based on the detection of new trends in the variation of variables characterizing a user or a group. These detected trends make it possible to build a future profile of the user interest made up of a series of keywords. These keywords will allow, under certain conditions, to find in advance web objects thanks to a search engine that gives URLs that will be downloaded. These objects will be stored in the users' proxy-cache that plays an important role in terms of collective intelligence catalyser. The user profile is a vector of valued keywords established on the basis of analysis of the textual documents consulted over a given period. Significant keywords obtained after a basic lexical treatment (lemmatisation, etc.) are analysed in order to retain only the words corresponding to the dominant topic of interests. The vector of these keywords associated with a standardized frequency constitutes the profile. It can be individual or collective over a short or long term. In accordance with the conclusions drawn from a previous study [1], the size of these profiles is fixed at 10 items, knowing that a more significant size does not bring precision and weighs down the treatment. Taking into account these constraints, the frequencies evolutions of each profile variable (i.e. a keyword) can be described in a chronological reference mark. In the Figure 1, we have time for X-coordinate and the level of weighting in Y-coordinate. We represented here the evolution of the frequency of appearance of the words “car” and “tree” over 4 days.

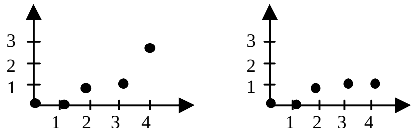

Figure 1: Evolution of the variable “car” vs. “tree”.

We see that the topic "car" is dominating compared to the topic "tree" for the last consultations. The method of trend detection is as follows. For a given interval of time and for each variable, we compute the increase of the weights. Only the variables observed as trend emerging upper than a certain threshold are preserved. For example, the threshold can be the average of the increases.

In the Figure 2, we present the forecasting method. For each user, we calculate the predictive profiles corresponding to the emergent keywords compared with the previous period (period i, compared with period i-1). The most frequent keywords of this profile are used to supply a search engine which provides corresponding URLs which are finally downloaded to supply the proxy-cache for the following period (period i+1). Moreover, we download only the URLs proposed by the search engine for at least 2 users, in order to take into account of the collective intelligence.

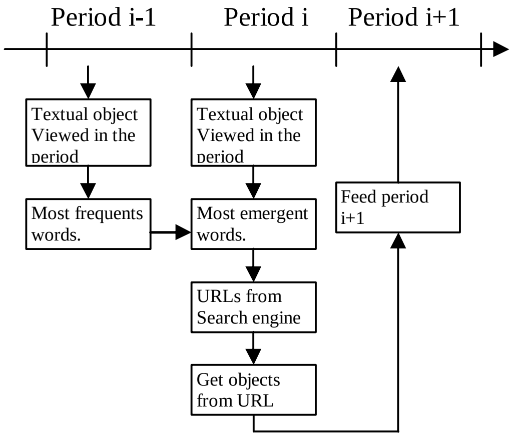

Figure 2: Forecasting method

The objective is to download in advance the predicted URLs. Consequently, the objects are present in the proxy-cache before the user requests them. The fact of checking that the user asks indeed for the predicted URLs is a mode of validation of the optimum of the method possibilities. Indeed the user does not know all the predicted URLs. It is extremely probable that many of not consulted URLs could be consulted if they were known. The method of forecast was implemented on a Squid proxy-cache and was tested in parallel with a not equipped proxy corresponding to the passive method. On the average the hit ratio of a regular proxy cache is about 30 %. This means that a document previously consulted by a group of users has about 30% of probability to be requested on the future (see for example [2] [3]).

2.2 Passive Method

The passive approach is the reference in our study. Its principle comprises a very low algorithmic complexity (shared memory stack of a regular proxy-cache) and makes it possible to evaluate the range of more sophisticated techniques. Indeed in this approach the prediction of the objects which should be consulted in the future is obtained purely by a synergy effect of the users’ actions. In other words, we consider that if an object has a shared interest in the past, it is likely to be consulted in the future. The passive mechanism of prediction thus consists in "simply" placing at the disposal of the users through the proxy, the most popular objects (consulted more than 2 times) over one former period of observation. We will note that this approach aims at replacing or combining "artificial" algorithmic complexity by the natural complexity of the human activity. This is the basic concept of the collective intelligence.

3 Validation Method

To evaluate our approach, we simultaneously replay traces of activity on a passive and active process (see

Figure 3). One of the interests of this comparative approach is to take into account similarly the requests on pages which would have disappeared from web servers.

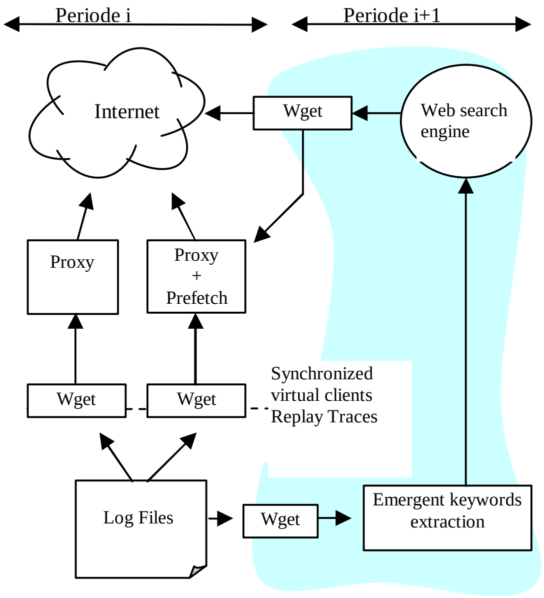

Figure 3: Structures of test for the evaluation of the forecasting method

We initially have a file of traces gathering the activity of 2000 users over 8 days. The computation of the trends is carried out on the text/HTML documents which represent 791,069 requests (227,325 documents). The experiment proceeds in the following way. The initial file is replay on the two processes simultaneously until reaching 7PM each day (time stamp in the file of trace). The replay is then stopped (simulation of the night) during time necessary to make the forecast computations and to find the corresponding objects. As explained previously the emergent words are calculated for each user by comparing during the 4 last hours the keywords (3-7PM) with the 4 previous hours (11AM-3PM). For each user, 3 emergent words are given to a search engine which provides the corresponding URLs. The choice of these parameters will be detailed further. The URLs corresponding to at least 2 users are downloaded and stored in the cache. After this operation, the replay of the file of trace restarts until the tag 7PM of the following day, etc. If a predicted URL is effectively consulted, the proxy-cache log will indicate a hit. As the intuition suggests it, variables as duration period for the estimate of the emergent words or the number of words provided to the search engine affect the effectiveness of the forecast. This is studied in the following section.

4 Results

The following figure indicates the forecasting accuracy according to the duration of the period in days of activity (since 1/8 of days = 0.125 to 3 days). We remark that the best results are obtained when the period is rather short and about 1/4 of days. The fall of performance for a size of lower period is due to a too significant reduction of the number of objects (too short period) what makes the forecast too vague. Thus, theoretically, with a number of infinite objects, the size of the period of the ideal forecast tends towards 0, but from a practical point of view, a too short period is difficult to manage. In addition the quantity of network activity is less penalizing when it is carried out in period of under activity. For all these reasons, we initially chose to carry out these calculations the night, over one reference period, 4 hours all the 24 hours (easy to test in an experimental context).

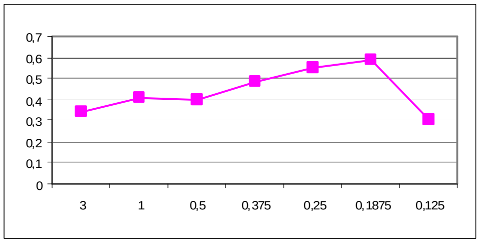

Figure 4: Percentage of forecast according to the period in days.

At first sight, the percentage of URLs correctly predicted is very weak (less than 1 %). In fact, it should not be forgotten that it is here about single URLs. In reality, some URLs are consulted many times and others very little.

Nb word	% max of good forecast
2	0.53
3	0.7
5	0.41

Figure 5: Impact of number of keywords on the forecast.

Another problem to be solved is to evaluate the number of emergent keywords to provide to the search engine to obtain an effective forecast. A second experiment (see Figure 5), shows that 3 is a good choice.

In what follows, value "A" represents the number of request generated over period i to obtain a quantity "B" of URL predicted over the period i+1. According to the precision of the used forecast method, we need a more or less significant number of requests (A) to obtain a given number of URLs (B). The variable "C" measures the number of consultation, carried out by the users on predicted URLs (B). Let us remember that these consultations are made in blind mode since users do not know that the URLs (B) are in the cache. We thus can define the ratio C/B as a first measure of the forecast efficiency. The figure 6 represents this ratio. The doted line represents the active approach whereas the plain line represents the passive one.

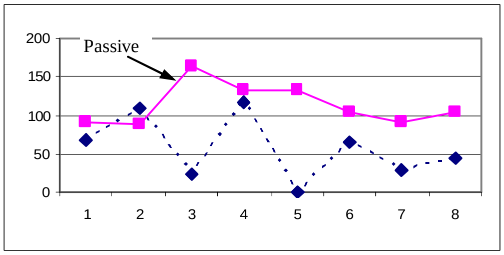

Figure 6: Evolution of the ratio C/B (effectiveness of the forecast)

We remark that the passive approach is better quite all of the time. However, as we said it previously, a given quantity of predicted URL can be generated by a number more or less significant of requests (several requests corresponding to the same URL). By taking into account this parameter, we can compute the forecast output with the ratio B/A (number of predicted URL / number of requests to obtain the predicted URLs). As shows in the Figure 7, the passive method is more effective but it has an output 3 times less significant than the active method. This means that the passive method requires 3 times more requests to obtain a comparable effectiveness with the active method.

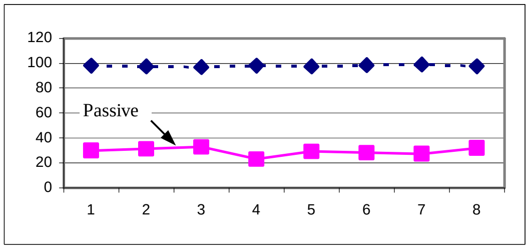

Figure 7: Evolution of the ratio B/A
(Output of the forecast)

Finally, we compute the C/A ratio between the number of requests on the URLs predicted (C) and the number of requests to obtain these URLs (A) (see Figure 8). We obtain a more precise vision of the effectiveness of each method. Indeed this metric take into account the effectiveness of the forecast by balancing it with the "energy" needed to produce the predicted URLs (a number of requests). We note that in 6 times out of 8, the active method is more effective.

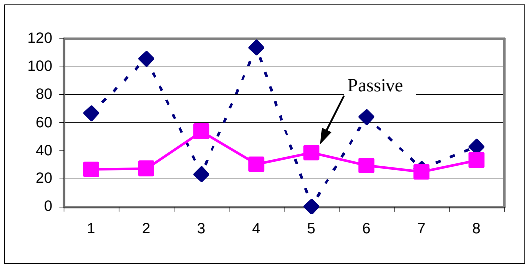

Figure 8: Evolution of the C/A ratio (balanced effectiveness)

5 State of the art and discussion

Many works were carried out on the forecast of the web documents consultation. The practical applications of these works consist, for example, in proposing to the users links on potentially interesting objects or even to preload and store locally the corresponding objects. The studies on the prefetching, the pre-filling or the recommendations systems are included in these categories. Generally speaking, these techniques are sometimes difficult to compare in terms of effectiveness since what is measured can be very different. For example, the forecast of the consultation from a web server viewpoint cannot be compared with the forecast of the consultation from a user point of view. In the first case a simple statistical study on the access to the limited contents of the server can be used as an effective predictive value. In the second case, the user can potentially get the entire Internet what makes the forecast much more difficult.

We can distinguish several architectures being able to imply a more or less strong interaction between the end user, the final server and the mediation systems (e.g. proxy-cache). For instance, some solutions run only on the user's terminal. Other techniques make it possible the origin web server to collaborate with the intranet proxy-cache in order to transmit the more popular contents to them [4] [5]. The autonomous techniques give for the moment limited results in term of forecast but are easy to deploy. Indeed the installation on only one computer or only one site makes it possible to obtain all the possible benefit of the system. This approach has the disadvantage of generating overload traffic which contributes to pollute the Internet already quite charged. Collaborative techniques are very effective but needs to be deployed on all Internet web servers (potential accessible space). This very heavy constraint makes these solutions difficult to exploit. For example [6][3] modified HTTP protocol to be able to communicate the predictions between the web server and the browser.

Several forecast methods can be used in these different architectural contexts (more or less autonomous). The main difficulty is here to identify the type and the quantity of data on which must be carried out the predictive analysis. The second problem is to identify the predictive model which will make the forecast. An example of basic predictive model is that there is a strong probability that a user consults in the immediate future the documents corresponding to the links contained in the consulted document. In this case the predictive data of the analysis are the pages in consultation and the analysis simply consists in parsing the document to extract the links (e.g. images included, links to other pages, etc) which will be downloaded as soon as possible. This model can be derived in a more or less complex way. Such a method can be used on a proxy-cache (WWW Collector [7] or

CacheFlow [8]) or directly on the final user's navigator (NetSonic Pro [9] and WebEarly 3 [10]). In the basic case, there is no additional band-width consumed because the accesses correspond just to an anticipation of the automatic browser requests. An approach a little more sophisticated 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 [6][12][4][13], the decision-making must be very fast and seldom makes it possible to use complex forecasts methods. Some studies [7] [2] shows that if all the links were pre-fetched, the hit rate would reach 69 % (45 % for the 10 best links and 20 % for the best 5) .

Other approaches make a more elaborate selection of the links to pre-fetch. The Klemm's [14] webCompagnon evaluates the cost of pre-loading of each link and pre-fetch only those which have a significant latency (high round trip time). This system is used on the user terminal and the author obtains a reduction of the latency time of 50 % and an increase in the band-width consumed of 150 %. Another method used by Cachflow [8] on its proxy-cache (Adaptative Refresh) consists in refreshing the most required documents automatically. In other cases, the documents detected as potentially interesting for the user, are preloaded for the vacant periods [15] [16]. This last more flexible approach in term of times of analysis makes it possible to use more complexes and resources consuming forecasts methods. An analysis of the URLs ([17] [18]) or of the keywords [19] [2] contained in the past consulted documents is often used as bases of future activity. This profile of future consultation can be treated by algorithms based on clustering [16], learning [30] [20], decision trees [21], or hidden Markov chains [22] [23] in order to obtain decision rules to identify the future activity.

Whereas the preceding methods aim at selecting a reduced subset of known resources to evaluate those which will be consulted soon, others aim rather to discover resources potentially useful [19] [26]. These techniques of forecast correspond to the needs for the systems of recommendation that are assimilated to information filtering systems because the ideas and the methods are very close [24]. 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 [25][26][27]. Collaborative filtering finds relevant users who have similar profiles, and provides the documents they like to each other [28] [29].

6 Conclusion

Our forecasting method deals mainly with the identification of potential interesting web objects. Our results should be seen in the context where the potential target is that of the whole web knowing that human activity is highly non deterministic. Further work will consist in proposing the predicted object to user in order to evaluate the level of consultation whereas this work tested the consultation in blind mode.

7 References

[1] L. Lancieri, N. Durand; Evaluating the Impact of the user profile dimension on its characterization effectiveness: Method based on the evaluation of user community organizations quality, IEEE International Symposium on Computational Intelligence for Measurement systems and applications. (CIMSA2003)

[2] B. D. Davison , Predicting Web Actions from HTML Content: Proceedings of the The Thirteenth ACM Conference on Hypertext and Hypermedia (HT'02)

[3] A. Bestavros. Using Speculation to Reduce Server Load and Service Time on The WWW. In Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM95), Baltimore, November 1995. Also at : http://www.cs.bu.edu/techreports/

[4] E. P. Markatos and C. E. Chronaki. A Top-10 Approach to Prefetching the Web. In INET'98, Geneva, July 1998.

[5] J. Gwetzman and M. Seltzer. The Case for Geographical Pushing-caching. In Proceedings of the HotOs Conference, 1994.

[6]V. N. Padmanabhan and J. C. Mogul. Using Predictive Prefetching to Improve World Wide Web Latency. Computer Communication Review, 26(3) :22-36, July 1996
.

[7] http://shika.aist-nara.ac.jp/products/wcol/

[8] http://www.cacheflow.com/.

[9]http://www.web3000.com/products/NetSonicpro/

[10]http://www.goto.fr/ACH/achpreswe.htm.

[12] J. Gri-oen and R. Appleton. Reducing File System Latency using a Predictive Approach. In USENIX'94 Technical Conference, Boston, 1994.

[13] L. Fan, P. Cao, W. Lin, and Q. Jacobson. Web Prefetching between Low-Bandwidth Clients and Proxies :Potential and Performance. In Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS99), Atlanta, may 1999. Also at: http://www.cs.wisc.edu/~cao/publications.html

[14] R. P. Klemm. WebCompanion: A friendly client-side Web prefetching agent. IEEE Transactions on Knowledge and Data Engineering, July/August 1999.

[15] T. Palpanas and A. Mendelzon. Web Prefetching Using Partial Match Prediction. In Web Caching Workshop, San Diego, CA, March 1999.

[16] S. H. Kim, J. Y. Kim, and J. W. Hong. A Statistical, Batch, Proxy-Side Web Prefetching Scheme for E-cient Internet Bandwidth Usage. In Proceedings of the 2000 Network+Interop Engineers Conference, Las Vegas, May 2000.

[17] Y. Aumann, O. Etzioni, R. Feldman, M. Perkowitz, T. Shmiel. Predicting Event Sequences: Data Mining for Prefetching Web-pages. In Proceedings of the International Conference on Knowledge Discovery in Databases (KDD'98), 1998.

[18] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos. Effective Prediction of Web-user Accesses : A Data Mining Approach. In Proceedings of the WebKDD Workshop (WebKDD'01), San Francisco, 2001.

[19] P. K. Chan. A non-invasive learning approach to building Web user profiles. In Proceedings of WebKDD'99

[20] T. I. Ibrahim and C.-Z. Xu. Neural Net Based Pre-fetching to Tolerate WWW latency. In Proceedings of the 20th International Conference on Distributed Computing Systems (ICDCS2000), Apr. 2000.

[21] T. S. Loon and V. Bharghavan. Alleviating the Latency and Bandwidth Problems in WWW Browsing. In Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS'97), December 1997

[22] R. R. Sarukkai. Link prediction and path analysis using Markov chains. In Proceedings of the Ninth International World Wide Web Conference, Amsterdam, May 2000.

[23] D. Duchamp. Prefetching hyperlinks. In Proceedings of the Second USENIX Symposium on Internet Technologies and Systems (USITS '99), Boulder, CO, Oct. 1999.

[24] Recommendation System Based on the Discovery of Meaningful Categorical Cluster, N. Durand, L. Lancieri, and B. Cremilleux, International Conference on Knowledge-Based Intelligent Information & Engineering Systems, 2003 University of Oxford, UK (KES 2003)

[25] M Pazzani, J. Muramatsu, and D. Billsus. Syskill &Webert: Identifying interesting web sites. In Proceedings of the 13th National Conference on Artificial Intelligence, pages 54–61, Portland, Oregon, 1996.

[26] D.S.W. Ngu and X. Wu. SiteHelper : A Localized Agent that Helps Incremental Exploration of the World Wide Web. In the 6th international World Wide Web Conference, pages 691–700, Santa Clara, CA, 1997.

[27] H. Lieberman. Letizia: An Agent that Assists Web Browsing. In the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI’95), pages 924–929, Montral, Quebec, Canada,, August 1995.

[28] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, and J. Riedl.GroupLens: Applying Collaborative Filtering to Usenet News. Communication of the ACM, 40(3):77–87, March 1997.

[29]A. Moukas. Amalthaea: Information Discovery and Filtering Using a Multi-Agent Evolving Ecosystem. International Journal of Applied Artificial Intelligence,11(5):437–457, 1997.

[30] Q. Yang, H. H. Zhang, and T. Li. Mining Web Logs for Prediction Models in WWW Caching and Prefetching. In the seventh ACM SIGKDD International Conference on knowledge Discovery and Data Mining KDD'01, San Francisco, California, USA, August 2001.