Description of Internet User Behavior

Luigi Lancieri
Luigi.lancieri@cnet.francetelecom.fr
Centre National d'Etudes des Télécommunications France Telecom

Abstract

This document presents a way to characterize the dynamic behavior of the Internet user. We use an information structure based on associative memory to store data related to the user's thematic and behavioral profile. To do so, we "follow" the user's activity on the Web using a regular network device called "proxy cache". The behavioral profile is related to the way that the user acquires knowledge when he surfs on the Web. This can help us to automatically distinguish, for example, experienced users from beginners one. Users may regularly need different and superficial information or on the contrary information that is more thorough and precise. This is why we want to study the regularity of the knowledge acquisition. By using mathematical tools like Fourier or Wavelet Transform, we characterize and highlight symmetrical properties in the user's behavior. We also show that it is possible to give visual description of these symmetries in particular by using "Chryzode" as representation.

1 Introduction

The growing success of Internet gives great facilities regarding acquisition of information. The huge quantity of data available on the network corresponds to a large variety of themes and forms (Text, images, video,..). This information can be created, stored, obtained or modified by all kinds of people worldwide. However, Internet is a victim of its own success. There is such a large quantity of data that it is some times very difficult to find what we are looking for.

The pertinent information need not only to be consistent with a specific theme but it is also related with the user "behavioral profile". For example, the need of a young student is not the same as an experienced university researcher, even if they have the same interest. Indeed, it is interesting to notice that the young student and the experienced researcher "surf" the Web differently. By studying the user's Web accesses, it is possible to obtain a knowledge relating to his behavior and his field of interest. We will see that we can get this knowledge by a self oriented and low constraining learning process and build a dynamic information structure comparable to Hopfield associative memory. This structure will contain information extracted from the user's Internet accesses and will be semantically in correlation with his interest profile. Moreover, the knowledge acquisition of the user's behavior is extracted by analyzing how the information structure evolves over the time.

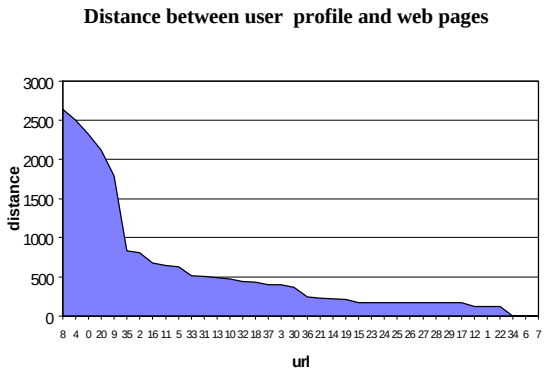
2 Implicitly oriented learning process-IOLP

Before introducing IOLP principle we need to give details on a regular Internet device called "proxy cache". This device works as a memory and it will help us to build a "user oriented" information structure.

In Intranet architecture, the proxies caches are used as an intermediary between a group of users and Internet. So all the users' accesses go towards the proxy and all the information downloaded is copied in the cache memory. When a user wants to see a page already downloaded by another user, this page can be provided directly by the cache. So the users are able to retrieve long distance page rapidly. The cache is also periodically self refreshed to provide the users with up to date information. The other benefits of caches are that they allow limited traffic on the net. This is interesting for all Internet users. Proxy-caches also have security capabilities. For all theses reasons, caches are more and more used in Intranets. What distinguish network cache from static computer memory is that its content is implicitly "user driven". The cache "follows" the user activity on the web, keeps in memory the information often downloaded and forget after a while the information which does not correspond to the user's main interests.

What is important in our concern is that the content of the cache reflects the interests, the habits, the needs, i.e the profile of the users. This is shown by the following curve. As we will describe later in this document, we can compute a "metric" that reflects the "semantic" proximity between components of information.

The following graphic shows the distance between the profile of one user and all the web pages (URL[18]) that this user downloaded over a period of time. We can see that the majority of the pages downloaded matches with his profile (low distance).



As the content of the a cache is " user oriented ", the information structure, built on the basis of an "implicitly supervised learning process ", will provide the characteristics of the users. Moreover, the learning support provided by the cache is free since this device is used in the Local Area Network anyway.

The information structure will be built by making a textual analysis of all Web pages that are in the cache. This can be done separately for each user or for a group. The analysis of the cache needs to be done regularly (e.g each day) because its content partially changes. This drives us to introduce the notion of "short term" and "long term" profile of a user.

3 Mathematics model

This model was initially designed [19] to allow the representation of multimedia web component (embedded image, sound, video) in a multidimensional algebraic space. The basic principle is to correlate the document's component (including words) spatial locality to its semantic locality. This means that words that are close (in the text) are statistically supposed to have a close meaning.

Considering that each word is a neuron, we increase the links between 2 neurons differently according to the fact that they are on the same page, in the same paragraph or sentence. Progressively, the neural net will become richer (thanks to the learning process) by analyzing all the pages that were downloaded by the users. We use a *n x n* (maximum number of neurons) matrix (C) of integer numbers that formalizes the connectivity between neurons as a graph with oriented and weighted arcs.

The C matrix can be considered as a representation of *n* vectors in a *n* dimensional vectorial space where each word can be represented as a linear combination of other words. The *n* elements of a vector to be represented are the weights corresponding to the n inputs of a neuron. In these conditions, an estimated value of the semantic link between two words can be formulated as the Euclidean distance between the two vectors associated with these words. So, if we want to know the « distance » between words 1 and 3, we have, referring to C, the elements of the two representatives vectors : C11, C12,...,C1n and C31,...., C3n (line 1 and 3 of C).

$$d=\sqrt{\sum_j (C1j - C3j)^2}$$

We consider that themes are words to which reference is the most often made, directly or not. If we make reference to the neural model, we can associate these words to the « winner » neurons on which other neurons are more often focused. Let's consider the link between neuron *i* and neuron *j* and *p* the weight of this link. To justify what follows, we admit that this link in a neural net at a *t* moment is equal to *p* links between 2 tops *i* and *j* of one graph. So we have a graph with *n* tops inter-connected by units weighted arc. The goal is to determine the most significant tops (i.e. highest influence). Perron and Frobenius [20] have used the graph theory to solve this kind of problem. They have shown that the « iterated powers » of the matrix associated with the graph (the same as C) associates to each top a value which corresponds to the « influence » of this top in the graph. These values can be computed rapidly with analytical numeric methods.

$$p^i(k) = \sum_j Cij.\, p^j(k-1)$$

This formula computes the order *k « iterated power »* of the node (neuron) *Nj*. The order 1 « iterated power » corresponds to the sum of the matrix C lines. Increasing the value of *k* allows us to determine the best indirect node influence. In order to determine the best overall influence and classify the nodes, it is interesting to compute the « relative influence » of a specific *i* node by the following ratio.

$$IRi = \frac{p^i(k)}{\sum_i p^i(k)}$$

Perron and Frobenius have shown that for all the nodes, this ratio converges to the Eagan vector of the C matrix. These tops are called the leader tops. We can decide that the *n%* of the best valued words will be themes. We will be sure, according to the Perron-Frobenius theorem that these words will be the most significant on the net. The

iterated power is a long term representation. Basically, it is updated every day and then normalized. We may compute the short term variation of the profile as the difference in consecutive long term profile.

This model was first experimented with a search engine called ISB (Interactive Shared Bookmark) [17]. ISB have two main capabilities. First, it is possible to find information related to key words even if the keyword is not in the document. The second, allows us to avoid the "noise" due to distantly related topics using the same words. This is possible because the information structure is "user interest driven". We also showed how it was possible to automatically organize a cache architecture [21] using a thematic organization (users are connected to caches according to their profile proximity instead of a geographical one). Grouping users on thematic criteria allows to increase the cache performance. This experiment also showed us that we can identify communities of interests by measuring the semantic distances between the users' profile.

The use of algebraic model to characterize and to retrieve information is not new [13] and several methods have been used. For example, the latent semantic indexing (LSI) [12] was developed at the Belcore labs for information search purposes. The principle is to built a semantic correspondence between documents and words. LSI uses a singular value decomposition. Each document is represented in an algebraic space by a vector where components are occurrences of most frequent words. The data mining [16] is also a technique which is more and more used with a lot of commercial products. Likewise, neural networks [14] [15] are also used by several authors to characterize and explore information. The main differences with our approach is that we automatically build a user driven information base.

4 Description of the user's behavior

Now, we would like to characterize the dynamic knowledge acquisition of the user. The first order iterated power (foip) is built by adding, for each neuron, the weights of all incoming ones. So, the foip represents the global influence (or importance) of each neuron. The relative position of each neuron to the others in the foip is not due to chance. Each new word is recorded by order of appearance in the pages, downloaded from the Web by a specific user. So, the relative position of the words in the foip is strongly related to the specific dynamic of the user's consultation. For example, if the important words (most weighted) are sparse in the foip, we can conclude that the user does not have a precise strategy in his consultation. Each access corresponds to completely different themes. In this case, the user has several superficial and weakly related fields of interests. On the contrary, if several important words are close, we can conclude that the user has high and more focused field of interests.

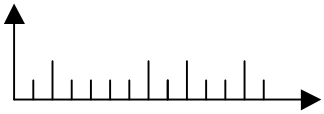


Fig 1 heterogeneous interests

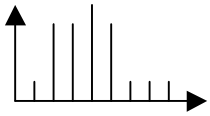


Fig 2 homogeneous f interests

It is also interesting to study the regularity of occurrences between words that have the same level of importance. This will give information on the consistency of the user's dynamic behavior (related to the knowledge acquisition). For example, over the time, a user may have successively different field of deeps interest. This behavior is characterized by the regularity of the acquisition of different knowledge. This could be the case of a journalist who searches information over the Web in order to deliver each day a paper on different topic. This behavior contrasts with the one of an employee in charge of summarizing, for example, the stock market activity and who consults mainly the same pages each day.

In order to obtain information related to the user's behavior, it is interesting to split up the fiop into waves of different periods. Each wave will materialize the simultaneous importance of groups of words and the simultaneous low importance of others. The goal is to study the regularity over different periods of the knowledge acquisition. To do so, we use the properties of the Fourier Transform. This computation is usually done to study temporal signal. The basic principle is to convert a magnitude vs a temporal representation in a magnitude vs a frequency one. For a continuous signal the formula is the following:

$$x(f) = \int_{-\infty}^{+\infty} x(t) e^{-2jpft} dt$$

The result, according to the Moivre principle is composed of real part and an imaginary one that correspond respectively to a cosine and a sine decomposition of the initial signal. Once the transform in the frequency space is done, we can achieve some statistical computations to characterize and compare the users' behavior. In addition to the average and the standard deviation, we can compute the covariance between the user signal and the random one.

Some authors [9][24] highlight the self similarities and the "symmetries" in the browsing activities of Web

surfers. This was done by analyzing the caches' log files. This shows specific access sequences to servers. As we said before, we highlight the same kind of properties in the foip that we consider reflecting the user's knowledge acquisition. Doing so, we expect to build a more representative behavioral profile. Further on, we will see two graphical representation that allow us to compare similarities between the users behavior: the Wavelet Transform and "Chryzode" representation.[23].

The wavelet transform (WT) has born in the 1980s as an alternative to the Furrier transform (FT) for the non stationary signals. Whereas regular FT converts a time series from an amplitude-time domain to an amplitude-frequency one, the WT shows the signal in the frequency-time domain. The working mode consists in scanning the time series with a delta time window in which we compute a FT. When the window is fix with a square wave signal this technique is called the Windowed FT. In the WT, the size of the window changes and it contains a specific signal called the mother wavelet. The great advantage of the WT is to have an adaptive resolution: good time and poor resolution at high frequencies and poor time and good frequency resolution at low frequencies.

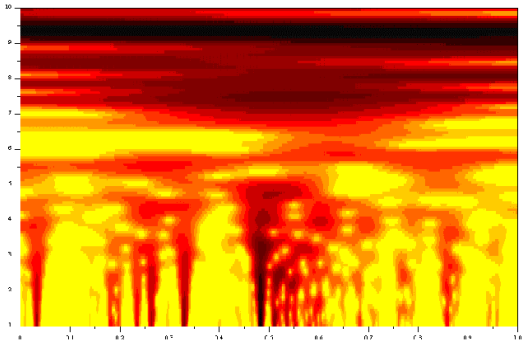


Fig 3 : User WT representation

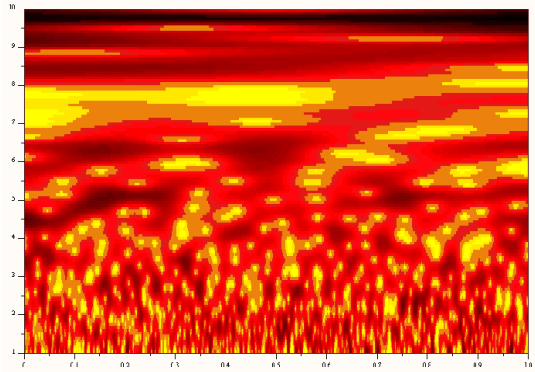


Fig 4 : Pseudo Random WT representation

The fig 3 shows the WT representation of a user foip whereas the fig 4, a random signal. The horizontal axis represents the neuron order and the vertical axis, its corespondent frequency. We interpret the frequency as a regularity between the neuron's relations. Shapes in this figure allow us to isolate the main stable relations on the net.

The Chryzodes (from chryzos and zooïde: writing in a circle) are interesting representations to highlight harmonies and similarities in the dynamic systems. In short, we may say that the chryzodes are graphic patterns of phenomena connected to arithmetical congruencies. To build such a representation, we take the real part (R) and the imaginary part (IM) of the signal's Fourrier Transform. We divide IM and R by its magnitude and we plot for each frequency the corresponding parametric curve. We see that each point (x=R and y=IM) turns around a circle of unit diameter. The distance on the circle between each point (phase evolution) is not constant because it depends on the user's behavior embedded in the original foip. So, we obtained our figures by linking all these points by a line.

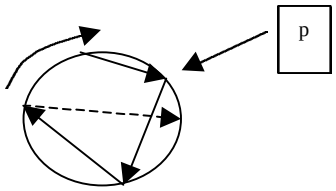


Fig 5: Principle of Chryzode

Once again the following pictures correspond to a user's profile (Fig 6) and the next to a random one (fig 7).

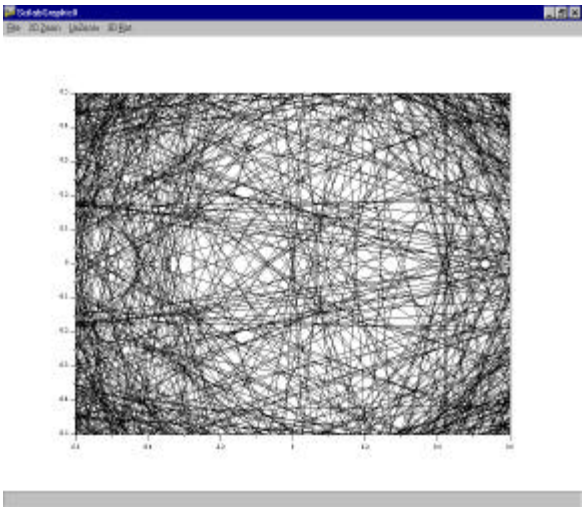


Fig 6: Chryzode User representation (Zoom 50%)

We made a zoom on the center with 50 % of the graphic being represented. Of course, all theses figures are built on a comparative basis (number of point, magnitude,..).

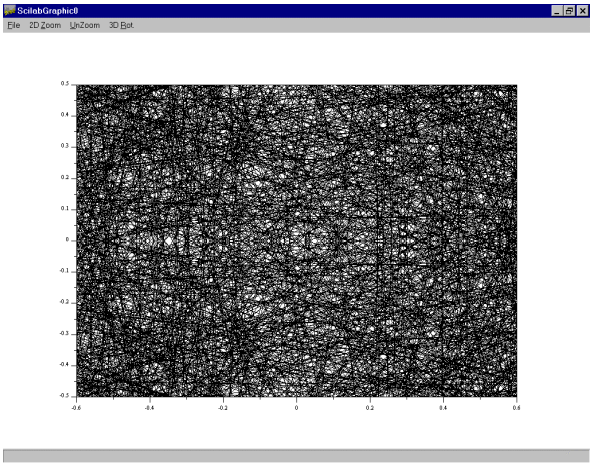


Fig 7:Chryzode Pseudo random view (zoom 50%)

Self similarity or fractal concept [6] is not new but it was popularized in the 1960s by the mathematician Benoit Mandelbrot. He formalized the notion of fractional dimension. Whatever the scale you consider a fractal representation, you see the same structure. In addition to their advantages for modelization, Fractals can also be used to solve practical problems. Mandelbrot used it to measure the coast line of Britany, also more recently, for example, to investigate the forest fire pattern evolution [7] or even to characterize virtual environment [8]. Studies were also done in the domain of knowledge characterization. The Zipf-Mandelbrot law [6] was used to express the self similarity contained in a lexicographic tree. Each word in the tree receives a weight according to its frequency of use. The words are then ranked from the highest weight to the lowest one. The relationship between the frequency (f) of use and the rank (r) in a ranked tree (with k and D constant) follows an hyperbolic law [9] [10] that can be approximated by:

$$r = k + f^{-D} \rightarrow r \sim \frac{1}{f}$$

So, a self similar set of data can be described by a line in a log/log representation. This is what we can observe in the following plot in which we ranked the foip from the highest weighted neurons to the lowest one.

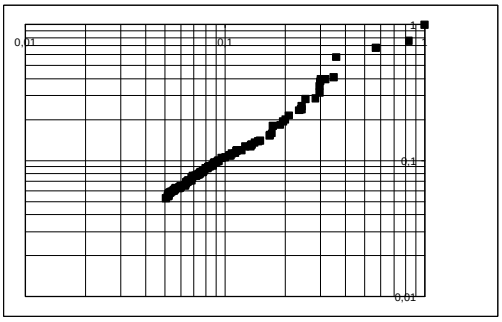


Fig 8: Self similarity in user's foip.

5 Conceptual issues

We saw that it was possible to build an information structure consistent with the users' network knowledge. When we describe documents, servers, group of users or any items of information, we do it from the users' point of view. So, the same items characterized with two users' information structure will appear different. This can be correlated with the subjective human perception and comprehension of his environment. Actually, it is well known that level of knowledge, culture and present concern modify the perception of people. This was demonstrated by M. Minsky inspired by an experiment of J. Piaget[11] .

The role of both distances and barycentre is very important. They are used to build and compare the representation of items of information. A document is represented as the combination of its words (vector) in the space of a user. A server that contains several documents can be described, in the same way, by a combination of its documents (which are described by a combination of its words). In the same way, a network can be described as a combination of the servers that it contains (which become described by,…,its pages,…its words). We can compare this view of composition and interaction in the network to a social metaphor where people become grouped according to their personal or professional affinities.

All these "items of information" (document, server, network,..) are described from the user's point of view. Once again this description is of course partial for 2 reasons. The first one is that the "Network " experience of the user, imbedded in the information structure that characterizes him, is limited (which depends on how long he surfs the web). This is consistent with the knowledge acquisition to the young children described by J. Piaget and M. Minsky.

The second reason is that the user's perception of his network environment is also more or less limited. For example, if he has a total access to a server (e.g in Intranets) he could compute a full characterization of this server. If the server is not completely accessible or if the user is not interested in its entire content, it will only be possible to build a partial view of the server. This is also consistent with what we can usually observe. Our knowledge of an individual, for example, does not only depend on our general experience (what we can understand of the general human behavior) but it also depends on how much time and interest we have in acquiring knowledge about peoples and what they want to let us know about themselves.

6 Conclusion

Moreover to extract a thematic profile, we present a way to express the behavioral profile of Web users. We have seen that this profile contains some symmetrical or self similar properties that deal with the user's knowledge acquisition. Such knowledge is made easily available with proxy caches that allow us to transparently monitor user's activities on the Web. Adding the thematical and behavioral aspect to the user's profile gives us the opportunity to built new attractive services, such as, the possibility to send best suited information to people based on their global profile.

7 References

[1] National Laboratory for Applied Network Research (NLANR); http://www.nlanr.net.

[2] Joël de Rosnay ; L'homme symbiotique ; Editions du Seuil.

[3] François Bourdon ; Systèmes d'information ouverts : Sémantique Interactionnelle des connaissances et systèmes multi-agents ; Rapport d'habilitation à diriger des recherches

[4] Jacques Roubaud, Maurice Bernard ; Quel avenir pour la mémoire; Edition Galimard

[5] The World-Wide Web as a Super-Brain: from metaphor to model , Francis Heylighen & Johan Cybernetics and Systems '96

[6] Les objets Fractals; B. Mandelbrot; Edition Flamarion

[7] Structure and scale of forest fires Patterns; Pedro Pereira-Goncalves (http://fct.unl.pt/~pedro/tese/papers/terra3/index.htm)

[8] Fractal complexity of a cybercity; Centre for advanced spatial analysis University College of London; (http://www.geog.ucl.uk/casa/naru/vcgis98/alpha.html)

[9] Characterizing reference locality in the WWW; V. Almeida, A. Bestravos, M. Crovella, A. de Oliveira

[10] On the implication of Zipf 's law for Web caching; L. Breslau, P. Cao, L.Fan, G.Phillips, S. Shenker

[11] Sciences et avenir; Dossier special Inteligence; December 1998

[12] http://superbook.bellcore.com/~std/lsi.html

[13] Using linear algebra for intelligent information retrieval; M.W. Berry, S.T. Dumais.

[14] Data exploring using self organizing maps; S. Kaski; in Mathematics Computing and Management in Engineering.

[15] Les réseaux de neurones definitions et principes

[16] Introduction au datamining; M. Jambu; Edition Eyrolles.

[17] Interactive Shared Bookmark; Luigi Lancieri In proceedings of WebNet 97 -Association for the Advancement of Computing in Education (AACE)

[18] T. Berners-Lee,, L. Masinter, M. McCahill ; Uniform Resource Locators ; RFC 1738 ; 1994

[19] Distributed Multimedia Document Modeling; Luigi Lancieri ; In proceedings of IEEE Joint Neural Network Conference 1998

[20] La théorie des Graphes (Graph theory) Claude Bérgé; Edition Dunod

[21] Automated organization of caches architecture; Luigi Lancieri WebNet 98 -Association for the Advancement of Computing in Education (AACE)

[23] http://www.chryzode.org or http://www.etca.fr/Users/Pierre%20Germain/CHRYZODES/english/page1.htm

[24] Linking cache performance to user behavior; I. Marshall, C. RoadKnight.; In proceedings of Cache International Workshop 1998.