# Distributed Multimedia Document Modeling

Luigi Lancieri
*Centre national d'études des télécommunications France Telecom*
*luigi.lancieri@cnet.francetelecom.fr*

## Abstract

*First, this paper describes a model to represent a set of heterogeneous data widely distributed. Secondly, we propose a method to use this model based on a specific learning process. This approach allows to build up dynamically a « semantic database » consistent with a user's group profile. To reach this goal, the database is built through using data contained in Intranets caches.*
*With approximate criteria of search it becomes possible to retrieve document or multimedia components which best fit with the request. It is also possible to « semantically» compare documents or evaluate the « semantic distance » between a document and a particular theme. The model also provides a method for self-extracting themes from the « semantic database ».*
*The first step is to define a semantic architecture based on neural nets. The second step provides an algebraic model of this architecture represented by a single matrix.*
*Doing so it is possible to easily evaluate relations (e.g. semantic links) in multimedia documents, by computing some mathematical properties of this matrix (e.g. Euclidean distances).*

## 1. Introduction

We consider the case of a multiple site entity connected to Internet through caches. This situation is common, it corresponds to company's Intranets or university's campus. Later in this document we explain more accurately what are the main functionalities of caches and why it is so interesting to use them. What is important to know at present is that the caches contain all Internet's data (HTML pages, pictures, sound files...) downloaded by users of a considered site. The model described in this paper integrates all information linked with HTML's pages : Address of the page, words, multimedia component (e.g pictures). One of the interests of this model is to simplify intuitive query analysis. This property allows an increase of performances and uses flexibility. For example it becomes easy to formulate the following kind of query : Looking for a piece of classic music, looking for information in English by formulating the request in French or filtering information (Web Site) close to a particular theme.

First we give some details about caches and a global architecture of the model. After that, we present a mathematics modeling approach and associated properties. Finally we give some perspectives and other possible applications of such model.

## 2. Why using caches

In Intranet, architecture caches associated with proxies are used as intermediary between a group of users and Internet. So all the user's accesses go by the proxy and all the information downloaded is copied in the cache. When a user wants to see a page already downloaded by another user this page can be provided directly by the cache. So users do not have to wait after long distance pages retrievals. The cache is periodically refreshed to provide up to date information to users. 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. So that caches are more and more used in Intranets.

What is important in our concern is that the contain of the cache reflects the interests, the habits, the needs, in one word the profile of a user's group [5]. As described in the rest of this document, cache data's will help us to build a « semantic database » consistent with the profile of users.

## 3. General presentation

To reach this goal, we aim to build up a model of HTML pages as multimedia document referenced with an URL (addresses).

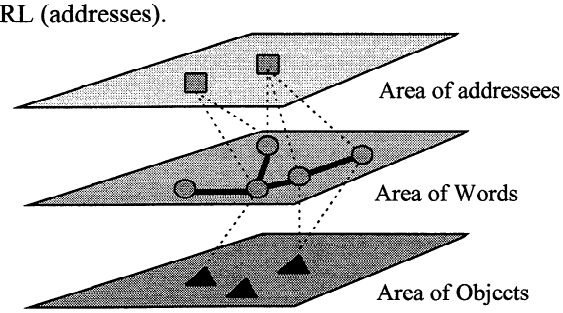


**Figure 1. General model**

As it is suggested by the above picture, pages' addresses, significant words (see later) and multimedia objects (sound, image, video) are classified in three areas.

As we will see later, this architecture's « intelligence » is located in the area of words. It is possible to move from the area of words to both area of addresses or area of objects. This allows to reach (just as if the relation is rough) from a keyword to both the pages' addresses (web site or cached version entire page) or a multimedia object. We will see more precisely later that it is also possible to get others various characteristics of these elements.

We note that area of addresses and area of objects elements are not inter-connected but, if we just make a three areas cross section, we show an inter-connected heterogeneous network. So, it is possible to consider the whole set of elements (addresses, words, objects) as two orthogonal planes with an inter-connected network of elements. The first one is the same as the area of words as presented in the above picture. The associated neural net, looks like Hopfield's network (associative memory). The second one is orthogonal and contains heterogeneous inter-connected elements. The associated neural net is near Kohonen topological map. The modeling of these two areas will be discussed later in this document.

## 4. To model HTML pages

As said before HTML pages can be considered as multimedia documents because they contain, except the main text itself, various objects as sound, image or video. Modeling HTML's pages involves to organize these kinds of elements. Before, let's formulate some definitions.

The word « *word* » (*m or n*) is intended as a significant word, which does not include article, preposition, verb, etc. If it appears ten times in a page (A) it has an « *occurrence ratio* » equal to ten.

It is necessary to identify a rule as precisely as possible which allows to determine if a word can be considered as significant. To reach this goal it is possible to use either a manual or an automatic approach.

The manual approach requires to build up a database with major knew non significant words. This method is not optimum.

The automatic method consists in the following intuitive principle : in a document, the less significant words appear more often as the most significant one. We can observe, for example, that most of the phrases begin with articles. So, it appears logical to postulate that a statistical study on apparition frequency ratio ($\overline{X}(m)$) of words in a document allows to separate significant and most of non significant words. Let's consider *Ls* (see tuning value) as a border frequency limit that separates significant and non significant words and S the set of significant words.

$$S = \left\{ m \in A \middle| \overline{X}(m) < Ls \right\}$$

Words in « *high relation* » ( R ) are words associated with the same themes or closed to them (e.g. bread and baker). The model will be able to extract words in high relation with a given word. This subject is discussed later with the computation of « distance » between words and the set up of various relations levels.

The « *page density* » of a hyper-text document is the ratio between the amount of hyper-links and words in the page. We consider that in conjunction with its density a page contains more or less themes. We also consider that in a same phrase, the words are in relation with the same unique theme. A null density means that all the information is contained in the page. At the opposite, a high density means that the major part of the information is contained outside the page (addresses linked on HTML's pages).

An « *heterogeneous page* » means that it develops a high amount of themes. We consider that most of pages have a low level of heterogeneity. This is important for the convergence of the neural net's learning process. (See later)

The estimation of « *heterogeneity level* » (TH ) is conform to the last hypothesis. Let's consider OC, the average occurrence of words in a page. Let's consider E the set of words that has an occurrence higher than OC, or more generally higher than an occurrence limit L.

$$E = \left\{ m \in A \middle| OC(m) > L \right\}$$

We define the heterogeneity level as the amount of higher relation (F) in E by the amount of words in E ratio :

$$F = \left\{ (m,n) \in E^2 \middle| R \right\} \qquad TH = \frac{\#F}{\#E}$$

An other way to formulate this is the following : Themes which are the most present in the page are they or not highly different ? This computation will be significant only if the neural net is over a certain learning level.

### 4.1 Structure of words

In our model we consider that a word is a neuron. The neuron is made active if sufficient of its high valued inputs are activated.

#### 4.1.1 The learning process

The goal is to create links between words if it does not exist or to update the weight of the link if it exists. We consider three cases : Words M1 and M2 are in the same

phrase, the same paragraph or only in the same page. We increase the link between the two words respectively by 3, 2 or 1.

Two words M1 and M2 with the same occurence involves two symmetric links. If the occurrence value of M1, for example, is higher than M2, we make only one connection from M2 to M1. Progressively, the neural net will become richer by analyzing pages and updating links weight. This is the learning process. Let's take an example.

We are looking for words that are in high relation with the word M. Activating itself the neuron M passes the signal on neuron M1, M2, M3 and M4. In this model, making a neuron active is done if the average sum of active inputs weight is higher or equal than the average of all input weights (decisional function).

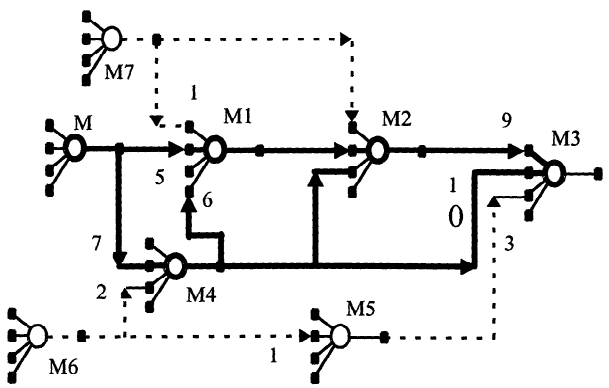


**Figure 2. Level of activation in the neural net**

As shown by this picture, we can segment neuron activation in several levels of influence or relations. M1 and M4 have the higher influence level on M whereas M2 and M3 have a secondary influence level because they are connected to M through M1 and M4.

### 4.2 Structure of objects and addresses

Each object is involved in two representations. The first one describes its physical characteristics as: Its nature (identified by its file type : sound, image, video). Its size. Its location : local or distant (URL).

The second one is a logical or associative characteristic. It consists in pointing, more or less, « strongly » at this object some words that will describe it. In the following picture, the word M2 describes better the object O1 than the word M1. We can also say that the words M1 and M2 contribute together to precise the context link of the object O1.

As in the structure of the words, weights allocated during the update to the link object to words will be 3, 2 or 1 according to the level of proximity between them, in the same phrase, paragraph or page (see paragraph 4.1.1).

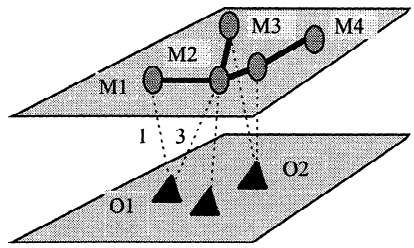


**Figure 3. Links between words and objects.**

In his basic sense, the modeling of the structure of objects does not directly correspond to classical neural net because, in this case, the learning process is not direct. It comes through the structure of words. When a weight is applied to one word to object's link, it is final. On the other hand, the structure of words, which is in constant learning, makes possible indirect link. between words and objects. In the figure 3, for example, the word M4 is not connected to an object but it has a link with M3 that is connected to O2.

The logic of the structure of addresses is the same as the structure of the objects. The only differences are located in physical characteristics since it is composed with URL or other information (density, heterogeneity rate). Here also, the purpose is first to describe one page with its characteristics including most significant keywords and second to reach a most significant page according to keywords.

## 5. Mathematics modeling

Referring to the general presentation, a multimedia document can be modeled by 2 orthogonal planes where 2 structures are embedded. We call it « words structure » referencing to the area of word and « transversal structure » referencing to the cross section of the three areas : words, addresses and objects. Let's define a C matrix that represents inter-neuron connection and their input weight.

### 5.1 Principle of the C matrix

The C matrix is an $n \times n$ (maximum number of neurons) dimensions of integer numbers. It formalizes the connectivity inter-neuron as a graph with oriented and weighted arcs.

We consider that each of the $n$ originator neurons is potentially connected to the other n destination neurons (including itself). In the reality the maximum number of destination neurons is equal to $n-1$, but considering the

real case, which is a sub set of a general case, adds complexity and does not bring special interest.

The neurons have $n$ possible inputs. Each of these inputs will receive the output of one originator neuron and will be affected by a weight. We impose that the output of the $i$ index of originator neuron $(1 \le i \le n)$ will be connected to the same index input of a destination neuron (1). This does not cause any problem because there is as many inputs to a neuron as neurons in the net.

If $\vec{O}$ is a vector of integers [0,1]where a single no null element validates the originator neuron index for which we want to know the connected destination neuron and $\vec{D}$ the result vector that gives these destinations connected neurons, we have the relation :

$$[\vec{D}] = [C][\vec{O}]$$

According to § 4.1.1, the C matrix is not symmetric, so to calculate its transpose make sense. If a single no null element of an O' vector represents the destination neuron index for which we want to know the source neurons, D' is the result vector that gives these source neurons. According to (1) D' represents also the connected input indexes. We have the relation :

$$[\vec{D'}] = [C]^T[\vec{O'}]$$

Let's take, as example, a 4 neurons net with inputs weights equal to 1.

$$\begin{pmatrix}1\\0\\1\\0\end{pmatrix} = \begin{pmatrix}0100\\0011\\1100\\1000\end{pmatrix}\begin{pmatrix}0\\1\\0\\0\end{pmatrix} \qquad \begin{pmatrix}0\\0\\1\\1\end{pmatrix} = \begin{pmatrix}0011\\1010\\0100\\0100\end{pmatrix}\begin{pmatrix}0\\1\\0\\0\end{pmatrix}$$

$$[\vec{D'}] = [C]^T[\vec{O'}] \qquad [\vec{D}] = [C][\vec{O}]$$

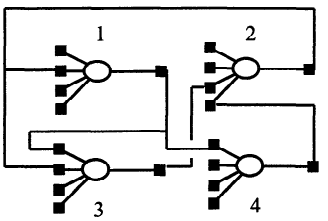


**Figure 4. Example of connectivity**

The calculation shows that the originator neuron number 2 is connected to the destination neurons number 1 and 3. We also see that the sources of neuron 2 are neuron 3 and 4. This last remark involves according to (1) that the source of neuron 2 is connected on the input 2 of destination neuron 3 and on the input 2 of destination neuron 4. We can check this computation with the matrix associated picture.

## 5.2 Building the C matrix

Initially the C matrix is null because it represents all the neurons without connections. The column vector Oj which contains a single $j$ line equal to 1 which represents the destination neuron index to connect, is used to update the matrix C. This one is built progressively with new connections adding to the precedent state of the C matrix the Oj and D vectors product .

$$[C_t] = \sum_j [C_{t-1}] + [O_j]\,[D]^T$$

Let's take an example. We are on the initial situation and we want to update C matrix for the following connections : neuron 1 to input 1 of neurons 2 and 3, neuron 2 to input 2 of neuron 1, neuron 3 to input 3 of neuron 1 and finally the neuron 4 which is not connected. To simplify the presentation we consider that the input weights are equal to 1. So we have the relation :

$$\begin{pmatrix}0110\\1000\\1000\\0000\end{pmatrix} = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}(0110) + \begin{pmatrix}0\\1\\0\\0\end{pmatrix}(1000) + \begin{pmatrix}0\\0\\1\\0\end{pmatrix}(1000) + \begin{pmatrix}0\\0\\0\\0\end{pmatrix}(0000)$$

Let's take another example. We have a C matrix including the input weights computed according to the same method as in the previous example. The C matrix models a 4 elements neural net (since each neuron have 4 inputs). We are looking for the various inputs and the associated weights corresponding to the various destination neurons connected to the originator neuron 2.

$$\begin{pmatrix}2\\3\\0\\1\end{pmatrix} = \begin{pmatrix}0200\\5316\\1001\\2109\end{pmatrix}\begin{pmatrix}0\\1\\0\\0\end{pmatrix}$$

The result shows that the neuron 2 is connected to the input 2 of the destination neurons 1, 2 and 4 and that these inputs associated weights are respectively 2, 3 or 1.

## 5.3 The C matrix properties

The C matrix can be considered as a representation of $n$ vectors in a $n$ dimensional vectorial space. The $n$ elements of a vector to represent 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 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).

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

In the following paragraph, we will show how to identify words as themes. So, we will have necessary means to determine the distance between words and themes.

## 5.4 Identifying themes

We consider that themes are words to which it is the most often made reference, directly or not. If we make reference to the neural model, we can associate these words to the « winners » neurons on which other neurons more often focus.

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 arcs unit weighted. The goal is to determine the most significant tops (i.e. highest influence).

Perron and Frobenius [1] have used the graph theory to solve this kind of problems. They have shown that the « iterated powers » of the matrix associated to the graph (the same that 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. In the second example of § 5.2, the computation gives the following results.

$$p^1(1) = 2 \; ; \; p^2(1) = 15 \; ; \; p^3(1) = 2 \; ; \; p^4(1) = 12$$

We can check in the matrix that the node 2 an 4 are really the most significant. Increasing the value of *k* allows to determine best indirect node influence and possibly to confirm the best direct one (k=1). In order to determine the overall best 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 converge to a proper vector of the C matrix. These tops are called the leader tops. We can decide that the *n%* of the best filled words will be themes. We will be sure, according to the Perron-Frobenius theorem that this words will be the most significant on the net.

As we said before, we can now automatically get out themes from a document, but it is interesting to consider a complementary method : Initialing a neural net. This method consists in « manually » pointing out words as theme. This have several interests. For example, in pointing out several equivalent words in various language, we can have language gateways which allow to retrieve information in one language queering in other one. In an other hand, associating manual and automatic method allows a better, and fastest focus (convergence) of a learning process.

## 5.5 Analyze of pages

We remember that in the model each word identified by its coordinates in the vector spaces can be expressed as a linear combination of other words. So, it is possible to compute the barycentre of a set of words. With the barycentre defined as the center of proportional distances between *n* points, we postulate that this computation can be estimated as a « semantic combination » of the initial set of words.

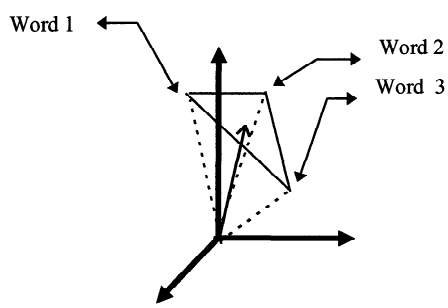


**Figure 5. Example of barycentre determination**

Now we consider for a specific page each significant word associated with its occurrence value. If we compute the barycentre of these words pondered by its occurrence we obtain a value corresponding to a semantic evaluation of a page.

It is important to see that doing so we combine a set of the word's representative value (the page) with a model consistent with the user's profile. So, it becomes possible to evaluate and compare semantic distances between a page and a specific theme.

## 5.6 Valid graph model

The goal here is to present at a *t* instant the transformation of the C matrix into a G matrix which

represents a static graph. This transformation only consists in taking under consideration only the activated neurons as top of a graph and eliminate the other ones.

Remember that Cij is the weight of an input i of a destination neuron i on which is connected a originator neuron j. As formulated before a neuron is activated if its solicited inputs have an average weight over the average weight of all the inputs. So the transformation algorithm from C to G is the following.

$$\textbf{Gij=1 if Cij} \geq \textbf{k}\frac{1}{n}\sum_{j} Cij \quad ; \quad \textbf{Gij=0 if not}$$

K is a selectivity factor between 0 and n. A null k factor means that all the neurons are considered as top of a valid graph.A n k value means thatno neurons isget as top of the graph, so, G becomes null. In first approximation it seems interesting to put a K value to 1. The G matrix will give us an image of the activated neurons consistent with the figure 2. The question is to know which neurons are activated and if those neurons are directly or not connected to the initial activated neuron. Considering this, the nearest are the most significant. This approach is comparable to the computing of a distance between the words seen before.

### 5.7 Transversal structure

We model here the structure of objects addresses. This model looks like the topological maps ( auto-organizer neural net ) proposed by Kohonen in 1982, whose work is directly inspired from a biological model. To be more precise, it consists, in our case, in a double Kohonen net with the same competitive layer, the area of words. The two input layers correspond to the area of objects and addresses. The goal is here to shows relations such as :what are the most significant objects or addresses in relation with a word or a group of words. The transversal model can be represented by 2 matrix (O and A) which make a link between the words vector Vm and the object vector Vo or the addresses vector Va.

$$[Va]= [A]\,[Vm] \quad ; \quad [Vo]= [O]\,[Vm]$$

The A and O elements correspond to the weight of the connection words, to addresses or words to objects.

## 6. Tuning elements

Various tuning elements allow to optimize the model.

- The occurrence limit §4
- Selectivity value §5.6
- The n% of accepted winners § 5.4

The influence of these parameters on the model stability and on the convergence of the learning process is variable.

## 7. Conclusion and perspectives

This proposed modeling method allows to formulate easily the dynamic of the learning process particularly with neural nets. This allows to use known result from classic algebra or graph theory, for example Euclidean distance or Perron-Frobenius theorem. This kind of model could probably be done without including neural nets but with possible more complexity. The application made here to the multimedia modeling and Internet caches, can be transposed to other comparable problems where uncertainty factors are involved. We can imagine, for example : operational search, transport, optimizing production management and so on.

## 8. References


[1] La théorie des Graphes (Graph theory)
Claude Bérgé Dunod

[2] Des réseaux de neurones (Neural's Nets)
Eric Davalo et Patrick Naim Eyrolles

[3] Précis de recherche opérationnelle
(Operational research handbook)
Robert Faure Dunod

[4] Neural Network
Usenet Group Frequently asked questions

[5] Interactive Shared Bookmark
Luigi Lancieri WebNet 97 Toronto
Association for the Advancement of Computing in Education (AACE)