Autonomous filter engine based on knowledge acquisition from the Web

Luigi Lancieri, Pierre Agostini, Nicolas Saillard, Samuel Legouix
Centre National d'Etudes des Télécommunications France Telecom

Abstract

This paper presents a filter tool that is easy to built and to use. It allows sorting and rating various textual documents according to their thematic content.

The method is based on a semantic comparison between documents and a thematic profile. This comparison involves a knowledge database autonomously built from the Web with an adaptive learning process. The autonomy (minimum human intervention) is an important characteristic of this tool. This approach allows to easily building a thematic dedicated knowledge database in any language.

The possible use of this tool can be to automatically send Internet information (Web pages, news,..) to specific users. It is also possible to consider this method as a way to avoid unsuited content from the Web.

1 Introduction

Filtering data is since a long time a very important matter. With the growth of Internet this question is more than ever up to date. For example, everyone knows that finding information on Internet is not always very easy, because of the large amount of data that the network contains. More generally, our concern is not only to find what is interesting to download but also to avoid what is unsuited.

Search engines, for example, are since a long time very helpful to find information in the web. Push technologies also allow having filter and retrieval capabilities [10]. Recent researches tend to improve these devices but some problems remain.

To avoid the download of unsuited information from the Internet is also a big concern, for example, regarding child protection on Internet connectivity. In this field also, several interesting studies have been done but the ratio cost on efficiency is not very good. Our system can be used as a helpful component that can contribute to reach this goal.

One of the different ways recently explored to address the question of filtering and rating information is to apply to the web the studies done in the field of linguistic and artificial intelligence.

The work presented in this paper is situated in this context. We would like to show that it is possible to build an easy going and low cost tool to compare information. As we will see later different methods or systems to achieve the same goal already exist. These systems are sometimes effective in the general case but they are not very adaptive and are often expensive.

Our scope is to build a specialized knowledge database that can easily be adapted to any subject in any language. Our method allows, thanks to the richness of Internet, to do it automatically with satisfactory results.

2 Distributed multimedia document model

We give here a brief overview of the model used; the complete version [1] is available for more details. This model was initially designed 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 a semantic locality. This means that words that are close (in the text) are statistically supposed to have a close meaning.

The general model describes, in addition, relations between context of the words and multimedia objects. Here, we only describe the words related model.

2.1 The learning process

In this model we consider that a word is a neuron that is made active if sufficient of its high valued inputs are activated. The goal is to create links between words (semantic network) 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.

Progressively, the neural net will become richer by analyzing a large amount of pages and by updating the weights of the links. This is the learning process.

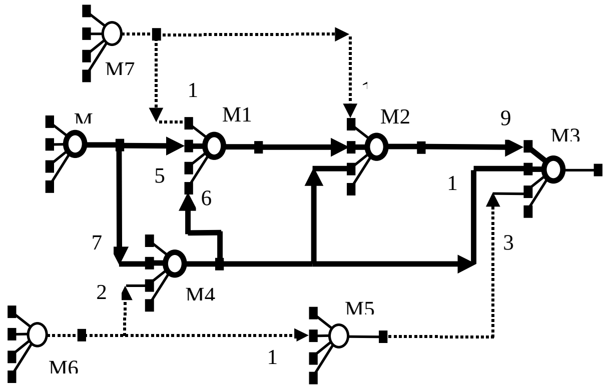

Fig1: Words' connection in the neural net

2.2 Mathematics modeling

Let's define the C matrix as an *n x n* (maximum number of neurons) dimension of integer numbers. It formalizes the inter-neuron connectivity 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 it).

The C matrix can be considered as a representation of *n* vectors in an *n* dimensional vectorial space. The *n* elements of a vector to represent are the weights corresponding to the n inputs of a neuron (words).

Car	100
Motor	70
....	.....
Wheels	50
Road	30
......	....
.....	.....
Security	10
.....	....

Fig 2: The vector for the word "Car"

This view can be compared to the logic approach in semantic linguistic [11] where each word or concept is described by its elementary component (car = vehicle + motor + wheels +......+..). As we will see the coefficients (vector components) are defined automatically by a learning process.
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).

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

Although it is not the main concern of this paper, it is interesting to briefly say that the model also allows to extract the profile of a user [1]. This can be done by computing the Eagan vectors or the iterated power of the Ci matrix (subset of C matrix for the user i).

One interesting point regarding this model is that it is possible characterize all elements of information (words, user's profile, web pages,...) in the same algebraic space. So, it is possible to make a comparison between these elements. The proxy cache is an important device and we will briefly see in the next section that it allows tracking the users profile.

2.3 Analysis of pages

In this 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.

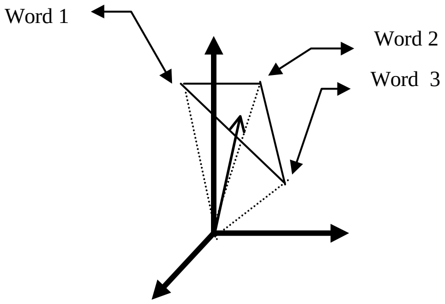

Fig 3: Example of barycentre determination

Now we consider the occurrences value of all significant words contained in a specific page. If we compute the barycentre of these words we obtain a value corresponding to a semantic evaluation of a page. The barycentre is obtained by computing the resulting vector from the words' vectors (in C matrix) pondered by the words' occurrences in the page. So, it becomes possible to evaluate and compare semantic distances between a page and a specific theme.

This method was already used in combination with a proxy cache content to improve a search engine called ISB (Interactive shared bookmark) [2]. The use of a cache is very helpful because its content (the most downloaded pages) is consistent with the thematic profile of users. In ISB, we show that it was possible to avoid themes unrelated with user's profile. (e.g. avoid electrical network or road network if the user is only interested on computer network). We also showed how it was possible to automatically organize an architecture of caches [3] using a thematic organization (users are connected to caches depending on their profile proximity) instead of a geographical one. To group users according to thematic criteria allows to increase the caches' performances (hit rate).

3 Architecture

This section describes the global and the functional architecture of our filtering system (filter). Filter is associated with a cache and a semantic engine in a local area network. The following diagram shows a global view of this architecture. We see that LAN's users get connected to the Internet through a proxy cache and almost all the pages that the users download go in the cache. In our case the filter engine will act in the LAN as a virtual user.

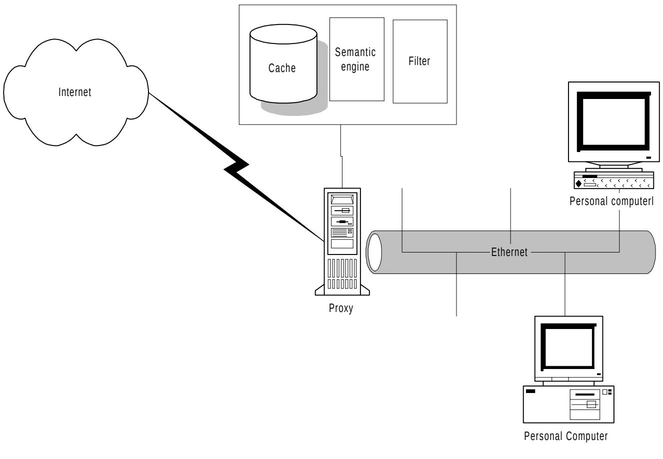

Fig 4: **General architecture**

The following diagram describes the modules of the autonomous filter.

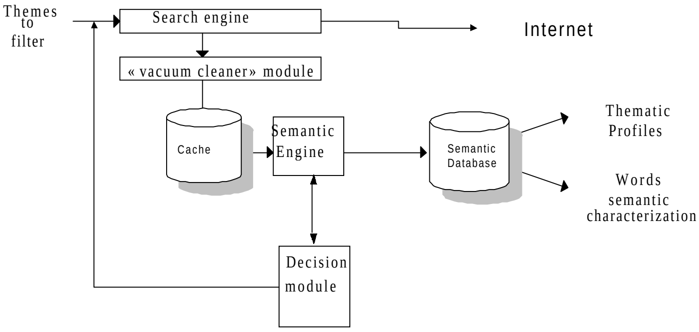

Fig 5: **Functional architecture**

We use an access module made up with a regular search engine service (Hot bot, altavista, ...) connected with a "vacuum cleaner" module that allows to download pages the addresses of which are turned back by the search engine (including a level of sub directory). The access module is feed by a user who gives 3 or 4 keys words (boot strap) basically related with the theme that will be used to built the knowledge database.

The access module gets connected to the Web and automatically feed the proxy cache with thousands of pages that become little by little more or less consistent with the chosen theme. The decision module will help to sort all these pages and feed the semantic engine with the most related pages.

These pages will be analyzed according to the process described in chapter 2 and the resulting inter words relations will fill the knowledge database.

Since search engine have multi lingual capabilities, it is possible to fill the knowledge database with any subjects (depend on the boot strap key words) in any language with no extra difficulties.

Before going ahead, we need to answers 2 questions:

- What level of consistency with the chosen theme can we expect to obtain with such a system?
- Can we increase this database consistency by an autonomous process?

The following section will answer these two questions. We present a comparison between a human rating of a set of Web pages and our system one. We also show that an automatic feed back system (decision module) allows converging automatically to a best consistency.

4 Methodology

The purpose of this method is to compare a human ranking ability to the automatic system. The first step was to “manually” find on the web a set of 60 HTML documents that more or less matches a specific subject (we chose the subject “ car” - automobile in French). The documents has a comparable length (2 “A4” format pages), and only 1 document contains one time the word “automobile”. So, this word only appears one time in a set of about 60 000 words. This set of pages was ranked according the human feeling level of relevancy with the subject “car”. Each person from a group of six was asked to analyze the set of 60 documents without knowing the results of the others (this took about 3 hours per tester).

The following table shows an average evaluation of the thematic content of the set of 60 pages.

Subject	Ratio %
Automobile (car)	32,3577586
Société	13,7887931
Histoire	7,56465517
Economie	6,87068966
Industrie	6,62931034
Media (public relation)	5,80172414
Futur	5,12931034
Culture	4,63362069
Transport	4,15086207

Fig 6: Evaluation of tested pages thematic content

We postulate that the human evaluation is the reference and all results provided by the system should be compared with the human estimation. However, as we will see next, we observed that the human subjective feeling sometimes leads to large differences in the evaluations of the testers.

We compute the system evaluation of the 60 tested pages in several steps related to a knowledge database’s level of learning. At each step we compare the vector barycentre of each page to test with the thematic vector of the theme “car” extracted from the knowledge database. Let’s remember that this vector is made up with the weights of all words linked to the word “automobile”.
Here is the different steps used to evaluate the efficiency of the filter system.

1)The knowledge database is only filled with the set of the 60 pages to test.

2) The previous step to which we add 200 pages with no relation with the theme “car” (we choose arbitrarily subjects related with the nature: animal, flowers, water, ..).

3) Step one to which we add a set of 100 and 500 pages highly related to the subject “car”.

4) Step one to which we add a mixing of 100 pages related to the theme and 100 pages with no relation at all.

5 Human subjectivity

It is well known that human perception and feeling can be very different from one person to another. This can also be observed in ranking pages because of a subjective interpretation. Each user has a personal feeling on assessment matter, the difference of sensitivity at peripheral elements such as images or even the personal knowledge of the tested field may modify the judgment. This can partly explain one part of the differences obtained between the human and the system estimation.

The following table gives a view of the user’s subjectivity regarding the ranking of the set of documents. We compute the correlation coefficient (0 = no correlation, 1= total correlation) between the users’ evaluation and between the users and the system one.

The 6 firsts items (LL to DV) identify the human testers. The item “syst” identify the filter system. The items “Aver” and “best” are respectively the average combination of all the 6 human users and the best combination.

LL	PA	NS	FM	SL	DV	Syst	Aver	Best	
1,00	0,75	0,76	0,74	0,59	0,79	0,43	0,90	0,81	LL
	1,00	0,75	0,72	0,44	0,70	0,46	0,87	0,80	PA
		1,00	0,72	0,53	0,75	0,55	0,91	0,95	NS
			1,00	0,49	0,71	0,53	0,86	0,90	FM
				1,00	0,44	0,35	0,64	0,55	SL
					1,00	0,37	0,87	0,79	DV
						1,00	0,53	0,58	Syst
							1,00	0,96	Aver
								1,00	Best

Fig 7: Coef. of correlation between users and system

This table shows that the heterogeneity between the users is on the same level that between users and the filtering system. For example the correlation between user SL and PA is equal to 0,44 whereas the correlation between the system and the users vary from 0,35 to 0,55.

These following two curves represent the evaluation of two testers that were chosen according to the representative divergence of their ranking. (FM, vs NS coef 0,72)

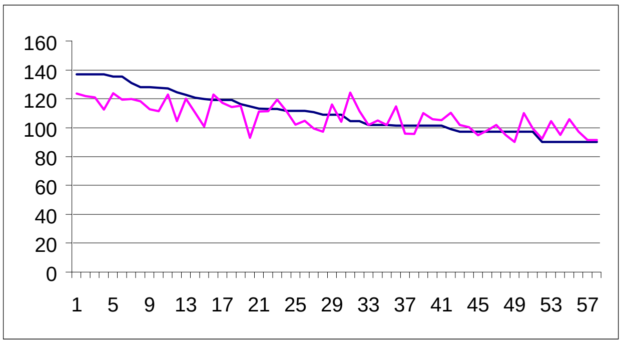

Fig 8: Divergences between Humans evaluations

We clearly see that there is not a high divergence on the pages that are completely relevant with cars or not relevant at all. But the divergences are in the middle of the scale where there is a high level of subjectivity.
The following plot shows the human (best combination) compared to the system evaluation. (coef 0,58)

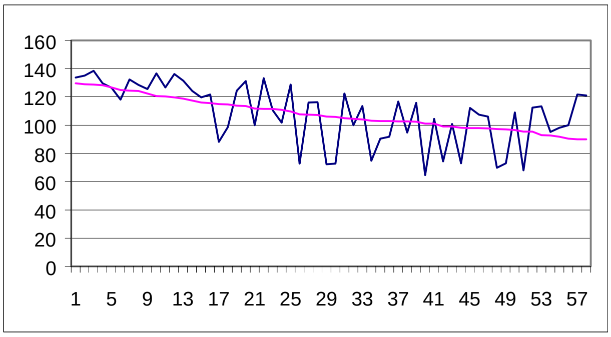

Fig 9: Human vs system evaluation

We see that the global trend of the system estimation follows the human one. Of course, all values do not match exactly but it is not possible to have errors between relevant and non-relevant pages (extremes parts of the plot). Let's remember that the word chosen as representative of the central theme (car) does not practically appear in the tested pages. We think that the level of the learning process achieved by the system also probably causes the differences between the human and the system estimation. We develop this point now.

6 Learning level

As many systems based on artificial intelligence the performances of neural network architectures highly depend on the learning level. The metaphoric comparison with human mind clearly shows that someone that never hear about cars can not rank pages that contain or not cars stuff.

For our system, learning means downloading HTML pages that contain more or less information on cars. This step was automatically done with the search engine. The following table shows according to the 4 steps previously described, the rise of the correlation coefficient between the system and the human evaluation. As we fed more the knowledge database (100 and 500 pages) we came closer to the human evaluation.

Set of feed	Coef. Correlation
60 tested pages (TP)	0,38
TP+ 200 random	0,43
TP + 100 “car”	0,47
TP+500 “car”	0,58

Fig 10: Impact of the learning process

We put in the following plot two curves that represent two learning levels in order to highlight the improvement. (TP vs TP+500)

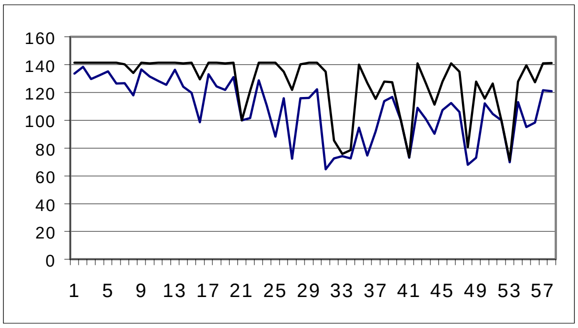

Fig 11: Impact of the learning process

We see that the more the system learns the more it matches with the human evaluation. In order to build an optimal semantic database, we have to select the pages to feed the system. The more this set of page will be consistent with the theme to filter, the more the results will be good.

However, the gain in knowledge acquisition is not linearly linked to quantity of pages used to fill the database. The best growth is obtained at the beginning of the learning process and progressively alleviates itself. This is also consistent with the well-known phenomena of knowledge acquisition of young children. Some researchers in this field say that the humans learn more in their 5 first years than in all the rest of their life.

7 Module of decision

In order to obtain an autonomous filtering tool, we have to take the decision if a specific page contribute or not to the convergence of the knowledge database. Without such a control process, the convergence would take a long time.

We have thousands of pages which are turned back by the access module (search engine and vacuum cleaner). The question is: what page contributes to build an optimal semantic database. We know, by experience, that the first pages returned by a search engine have a good level of consistency. So, even if we can not compare these pages with the semantic database (it is empty at the beginning), we can accept them and start to fill it. Progressively key words will appear in the database and will be used to compare the following pages.

A soon as the key needed word (e.g "car") appears in the database, we extract its representative vector. We also select the less no null weighted word in the "car" vector and we extract from the database its corespondent vector (see § 2.2). We compute the distance between these 2 vectors in order to have a reference. We postulate that this reference represents the distance between a related and a non-related theme (with "car").
To take the decision of acceptance of new pages in the database we compute the distance between this page and the vector "car" if this distance is up to half the reference we accept the page. Regularly, we compute a new reference that becomes more and more reliable.

Actually, it is very a very approximate method especially at the beginning, but it allows to converge rather rapidly.

8 Performances

This is a very important question. In the experimental version, the overall computation delay (analyze of pages, distance measure and take of decision) is less than 2 second for each textual document on a Pentium 200 Bi processor.

As, the constitution of the knowledge database can be done in batch mode the computation delay is not a real problem, but it can take a long time (more than 15 h of continuous computation for 500 pages).

Once the database has achieve a good level of learning, we can use it both real time or batch mode to filter information. Of course, the real time mode is the most difficult to manage and we have to take carefully into account the delay in the applications.

A typical example of this kind of application is the online filter when a user downloads a specific page. In this case the filter can be use to accept or to refuse the download of the page. In this last case the filter send a warning html page to the user instead of the requested page.

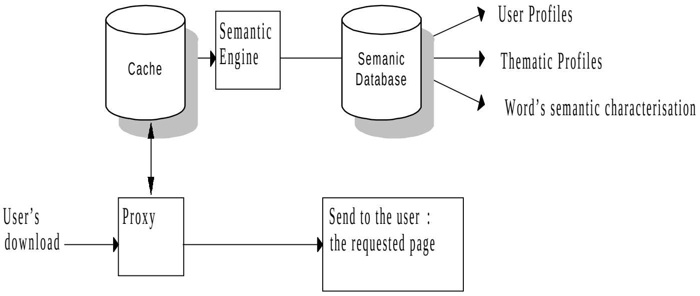

Fig 12: On line filtering

In order to have a better view of this question, it is interesting to compare this delay to the one needed to have a regular Internet access. In France, the average delay [4] (only during working hours) is around 5 sec with a maximum around 20 sec. Of course, this delay depends on several criteria including the geographical localization of the server. Nevertheless, we think that the page retrieval process and the rating process can be done in parallel. Considering that, in regular Web browsers, the text is downloaded before the images and that the analysis can be launched at the same time as the images' download.

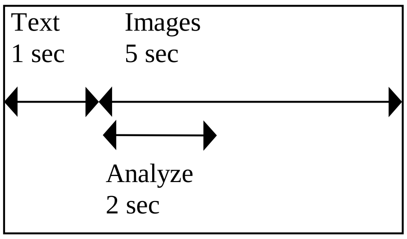

Fig 13: Parallel processing

In this case, the only drawback is that user can not have the progressive display of the page but he has to wait a bit more (3 sec in the previous figure). However, the delay to display the global page is not increased.

We see that the filter has enough time to take the decision: send or not the requested page to the user. In the case concerning already requested pages, the analysis delay is reduced because it was done previously.

We can also consider that the power of the machine increases rapidly. Actually, we estimate that it is possible to reduce the latency time to less than 1 second per page by optimizing the development.

9 Related Works

The use of the 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 build 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 the 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 network [14] [15] are also used by several authors to characterize and explore information.

Filtering Internet content is a problem since everybody can access the WWW and especially children. This problem is critical for corporation, schools and Internet Access Providers but also for parents who want to prevent their children from accessing sexual materials.

Thus, different methods exist to block access to certain types of document and which have been implemented in commercial products [8,9]. These products set legal problems and all the methods used can not guarantee to block all non-desired documents.

Filtering algorithm should detect web documents, which concerns given themes. Web is in constant evolution that is why filtering must also be dynamic. We expect to set up the themes and the quality of filtering for all these themes. Within different environment, theme to ban would differ as the security level.

Filtering content can either be done by a proxy cache or by a navigator with one of these methods:

- Using an URL ban list is the simplest method but it does not provide good results because of the difficulty to maintain such a list.
- So, we would prefer the document rating systems. PICS [7] define HTML tags which associate a rate for each theme to a document. This system can only work if the persons who maintain documents accept to rate their documents.
- The last system consists in using an engine that permits rating as documents are accessed. This slows down access time and forbids human rating but gives entire control on the filter mechanism.

None of the systems described in this category are, today completely reliable.
Rating content is done by using one or several of these three methods:

- Human rating is a safe solution but can not be realistic due to the number of documents on the web. We should prefer an auto-rating system.
- To use a dictionary is the most developed solution in order to rate a document. The presence of certain words in a text is a good criterion to block access to this text. This system is too static and can not provide a refined evaluation.
- Intelligent systems based on artificial intelligence heuristic [9] can provide rate closest to human evaluation and are able to learn the filtering criteria by itself.

10 Conclusion

This paper has presented a method to build an autonomous filter engine that allows to automatically classify information from the web. One of the main contribution of this method is the ratio efficiency on cost with the use of regular tools or services (e.g search engine). It also allows building a specialized knowledge database that can easily be adapted to any subject in any language. Our method allows, thanks to the richness of Internet, to do it automatically with satisfactory results.
This kind of system can be used with profit in batch mode, for example, to sort information and directly to send it to interested users. The real time mode can also be interesting especially when security considerations are important. In this case, users may experience low extra delay.

11 References

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

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

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

[4] http://www.serveurs-nationaux.jussieu.fr/cache/mrtg/
Statistics on French national cache “ Renater ”. We consider delay from root cache to a distant server.

[5] T.Krauskopf, J.Miller, P.Resnick, and G.W.Treese, “ Label Syntax and Communication Protocols ”, World Wide Web Consortium, May 5 1996, (http ://w3.org/PICS/labels.html)

[6] J.Miller, P.Resnick, and D.Singer, “ Rating Services and Rating Systems (and their Machine Readable Descriptions), “ World Wide Web Consortium, May 1996, (http ://w3.org/PICS/services.html)

[7] (PICS) “ Platform for Internet Content Selection ” (http ://www.w3.org/PICS/)

[8] SafeSurf (http ://www.safesurf.com)

[9] valuWeb http://calvin.ptloma.edu/~spectre/evaluweb/)

[10] http://www.backweb.com/html/knowledge.html

[11] Semantics; Geoffrey Leech; a Pelican Original

[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 Edition Eyrolles.

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