Scientific Formulas Extraction
Oriented Towards Web Summarizing File Cards

SHAHNAZ BEHNAMI
LaLICC (Language, Logic, Informatics, Cognition, Communication), CNRS, UMR 8139
Paris-Sorbonne University
96, Boulevard Raspail, 75006
PARIS, FRANCE

Abstract: - This paper is devoted to the results of the Filimage system using Internet oriented technologies. Our solution is applied to textual documents and the main objective is to provide efficiency the e-document extraction containing mathematical objects. This work deals with the problem of the scientific formulas and the process designed to enable a semantic filtering. The significant interest of the adapted strategy is to bring out the multimedia contents use to perform an information retrieval. We introduce a description of the Contextual exploration method which is carried out for the text semantic filtering. Based on the access to user-relevant information, the system lies on the Natural Language Processing (NLP) to identify the visual components (formulas, tables, diagrams, images...) and the textual extracts at the same time which be used as their comments by A.M.A (Automatic Matching and Associating) operation. The paper gives an overview of the development and shows how is implemented the system. In order to illustrate the results of this process, we expose them by a visualization example containing automatic extracted scientific formulas. At last, we present how the extracted information are stored (scientific formulas, textual comments, title…) for a subsequent using of the results.

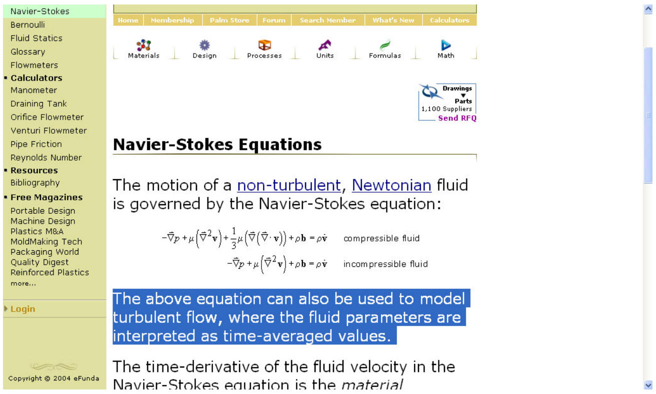
Key-Words: - A.M.A operation, Contextual exploration, Filimage system, scientific formulas extraction, semantic filtering.

1 Introduction

The increasing number of user's need when they browse web pages and carry out the research into the relevant information sources, makes emerge the potential useful of the visual components. Indeed, the systematic resort to the images and the contents of the formulas is a growing web practice. One of the most significant interaction is taking place between the "scientific formulas" and textual expressions in the technical "e-documents". It means that the automatic textual filtering is one of the challenge to be succeed, including an exhaustive and synthetic view of the conveyed information. Various researches are available concerning the multimedia contents exploration [5]. The information retrieval notion depends on the treated document's feature. Henceforth, the extraction of the relevant textual annotations is inevitable and acquires the highest strategic interest to be associated with the different components of the downloaded web pages. A computational linguistics tool based on the semantic analysis have to bring one's mind to bear on quick users access to the contents. This work focus on the contribution of the different visual components (formulas…) as a cognitive aid making easier to understand the text. Otherwise, the comprehension get better thanks to perception and memorisation through the adaptable multimedia interfaces using internet technology.

2 Problematic

To answer the question: "how to make the conditions of access better, while dealing with the users' requests?" [13], the solution must suggest an ability to, view, transmit, use again and store all the formulas accompanied with the quantitative parameters inserted into the tables and the graphics... Fig. 1 shows an equation followed by it relevant explanation contents. The sentence underlined in blue is the relevant content to be extracted with it previous equation (Fig.1).



Navier-Stokes Equations

The motion of a non-turbulent, Newtonian fluid is governed by the Navier-Stokes equation:

$$-\vec{\nabla}p + \mu\left(\vec{\nabla}^2\mathbf{v}\right) + \frac{1}{3}\mu\left(\vec{\nabla}\left(\vec{\nabla}\cdot\mathbf{v}\right)\right) + \rho\mathbf{b} = \rho\dot{\mathbf{v}} \qquad \text{compressible fluid}$$

$$-\vec{\nabla}p + \mu\left(\vec{\nabla}^2\mathbf{v}\right) + \rho\mathbf{b} = \rho\dot{\mathbf{v}} \qquad \text{incompressible fluid}$$

The above equation can also be used to model turbulent flow, where the fluid parameters are interpreted as time-averaged values.

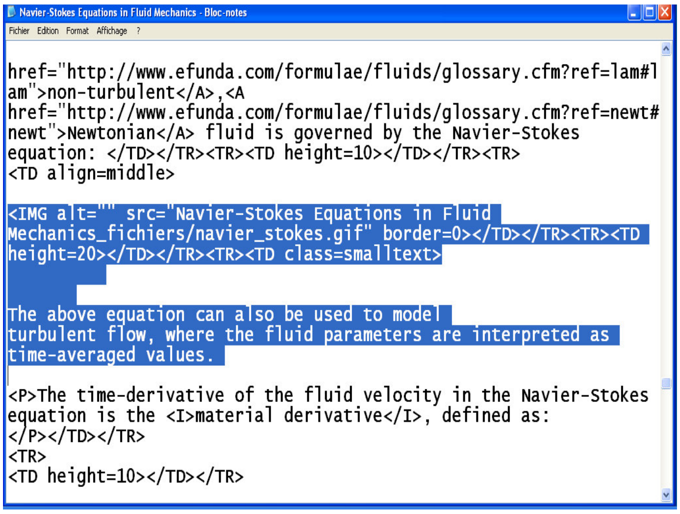
The time-derivative of the fluid velocity in the Navier-Stokes equation is the *material*

Fig. 1. This web site contains a relevant textual segment referring to "Navier-Stockes Equation". www.efunda.com/formulae/fluids/navier_stokes.cfm

We will see in the following sections that the automatic extraction of the scientific formulas brings up some problems inherent to their identification. Some systems in information retrieval have been proposed [6], [11], numerous works rely on numerical treatment, pattern cognition, Markov network processes… to extract the information and propose diverse solutions [1], [9], [10], [12], [14]. However, a multimedia extraction system that retrieves the multiple components using the Automatic Matching and Associating (A.M.A) operation is a relevant issue to be realized. The tool gap is the key issue for both visual and textual objects filtering and Filimage system makes one's contribution to fill this user's need [3]. The proposed system exploits the relevant information available in both visual and textual context. The aim of this work is to identify through an adapted strategy, the textual expressions with regard to scientific formulas in the web pages. By this way, it contributes to make progress in the linguistic engineering. To this effect, we have developed a system to address design usability with a view to automate the treatment of the different objects stored in the databases. The next section describes the problem solution.

3 Problem Solution

The association between the extracted components brings out the most relevant information according to the user's point of view. The adopted strategy relies on the document's structure particularly about the design to obtain the optimal filtering of mathematical objects. The interest points of this treatment resides in a filtering-orientated towards one and/or the other one components (formulas, textual expressions) taking accounts of the databases variety (textual and visual). The design and the realization of this system use the web technologies to log on the downloaded pages [4]. An adopted principle relies on the encoding of each object's feature (textual and visual), providing means of analysis. The first issue to be solve is the extraction's task. The automatic treatment proceed via the source code tags of the structured documents (HTML, XML) through the technique object-oriented. The system has the ability to filter the web sites by using the appropriates objects addresses (URL). Fig. 2, is the example of a HTML code source to extract the relevant objects according to the mark-up tag. The extracted object mark-up tag for instance is: <img…src="…_ fichiers/Navier-Stockes.gif">... The extraction requires a separated and adapted process. Indeed, the proposed solution consists of the filtering through the realization of specific modules. This decision overcomes the complexity of each object to provide an enhanced result. Therefore, the linguistics decision-making are necessary while the different operations are in progress.



Navier-Stokes Equations in Fluid Mechanics - Bloc-notes
Fichier Edition Format Affichage ?

```
href="http://www.efunda.com/formulae/fluids/glossary.cfm?ref=lam#l
am">non-turbulent</A>,<A
href="http://www.efunda.com/formulae/fluids/glossary.cfm?ref=newt#
newt">Newtonian</A> fluid is governed by the Navier-Stokes
equation: </TD></TR><TR><TD height=10></TD></TR><TR>
<TD align=middle>

<IMG alt="" src="Navier-Stokes Equations in Fluid
Mechanics_fichiers/navier_stokes.gif" border=0></TD></TR><TR><TD
height=20></TD></TR><TR><TD class=smalltext>


The above equation can also be used to model
turbulent flow, where the fluid parameters are interpreted as
time-averaged values.

<P>The time-derivative of the fluid velocity in the Navier-Stokes
equation is the <I>material derivative</I>, defined as:
</P></TD></TR>
<TR>
<TD height=10></TD></TR>
```

Fig. 2. This source code is corresponding to the "Navier-Stockes Equation and its textual comment "The above equation…be used to…".

On one side, Filimage system is undertaking treatments in order to etablish a connection with text, this allows to achieve a semantic filtering of the textual comments. On the other side, it identifies the mathematic formulas... and associates them automatically to the previously extracted textual comments. The combination results are stored and displayed to access.

4 Semantic contents Filtering

Developed according to the semantic filtering and based on Contextual exploration method (CE), the obtained result performances increase considerably [8]. The principle of CE requires linguistic knowledge and allow to find thanks to indices. The use of heuristic allow the acquisition of linguistic resources which are the indices, the exploration rules and the databases. The first step of the indices classification consists of their systematic collect and will constitutes the basis of the linguistic analysis. It means that the text semantic filtering consists of determining a number of rules to be used by the extraction system. The CE declarative rules allow to find the most relevant indicator with the associated indices, co-presents in the same context [7]. Otherwise, a set of rules helps filtering of the textual contents according to the selected task, for instance: extracting the information depending on user's profiles (automatic text summarizing, technological intelligence...). The relevant information are textual comments, thematic segments, titles, summing-ups and conclusions. In fact, when a textual segment is located as being related to a formula, it is extracted then is combined with the corresponding formula as its textual comment.

The main features of this method is illustrated by a context example. As shown in below Fig. 3, the identification of markers in a context is based on CE semantic filtering method [8]:

This ***diagram***…reveals.…

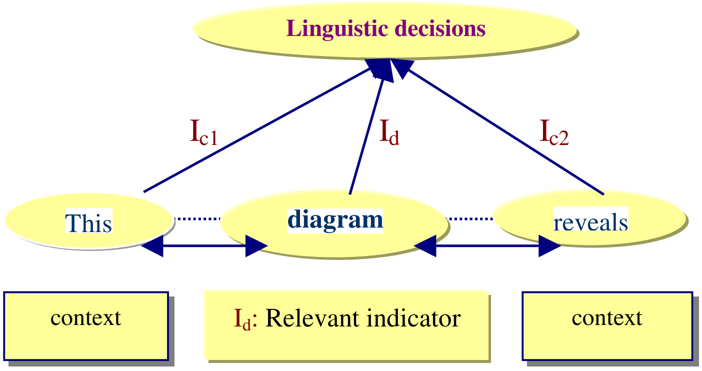
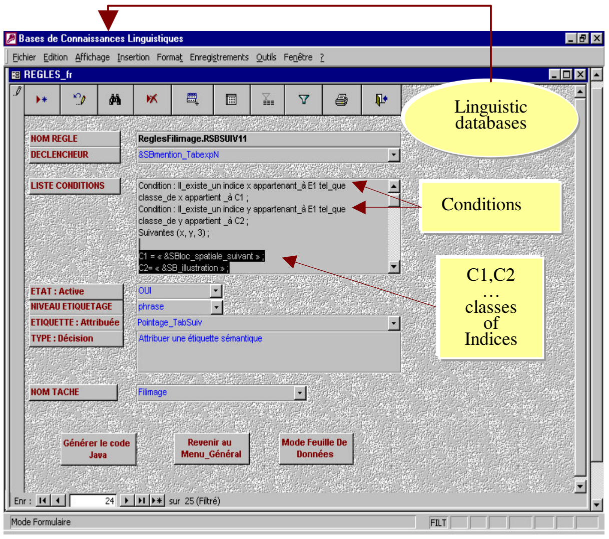


Fig. 3. Linguistic decision-making expressed in the exploration rule.

Here is one released rule example in the Filimage system database. Each CE rule must verified the relevant indicator, indices classes and the listing of the conditions to tag the sentence to be extracted.



NOM REGLE: ReglesFilimage.RSBSUIV11

DECLENCHEUR: &SBmention_TabexpN

LISTE CONDITIONS:
Condition : Il_existe_un indice x appartenant_à E1 tel_que classe_de x appartient _à C1 ;
Condition : Il_existe_un indice y appartenant_à E1 tel_que classe_de y appartient _à C2 ;
Suivantes (x, y, 3) ;

C1 = « &SBloc_spatiale_suivant » ;
C2= « &SB_illustration » ;

ETAT : Active: OUI

NIVEAU ETIQUETAGE: phrase

ETIQUETTE : Attribuée: Pointage_TabSuiv

TYPE : Décision: Attribuer une étiquette sémantique

NOM TACHE: Filimage

Enr : 24 sur 25 (Filtré)

Fig. 4. This CE rule example is a screen copy of the linguistic knowledge basis.

We describe in the following, the different steps of realization. The operational architecture is presented on the Fig. 5.

5 Development and implementation

The system implementation find it realization in Filimage. It is based on the sequential processing. The different phases of this optimization destined to the formulas semantic filtering are illustrated. The analysis levels with the successive treatments are included in the filtering process. Firstly, an elementary analysis is launched so as to produce what can be call the pre-processing. At the output of this operation, the HTML mark-up tags are automatically deleted by the source code reading. The aim of this phase is to send back a text suitable for the semantic analysis. Secondly, the formulas are identified and extracted. Subsequently, two different components that had been separated during the treatment, are again combined into one new structured e-document. Nevertheles, an appropriate proceed is designed for the extraction task in mind to integrate the scientific formulas. This operation is called A.M.A (Automatic Matching and Associating) and is required to display the final result [2]. A new web page is presented to the user in the browser offering an overall visualization of the extracted document. We should point out that a relevant association requires the necessary conditions to achieve a coherent result display. The operational's aim system is to provide an adapted result for each components which are integrated according to a coherent spatial organization of the extracted objects following the initial document composition. The below diagram shows the successive treatments.

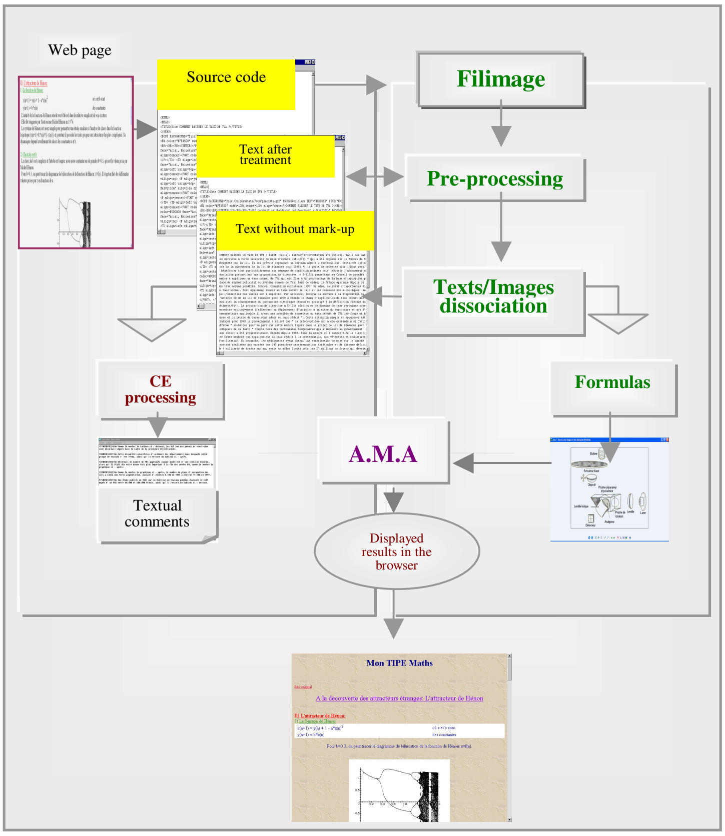


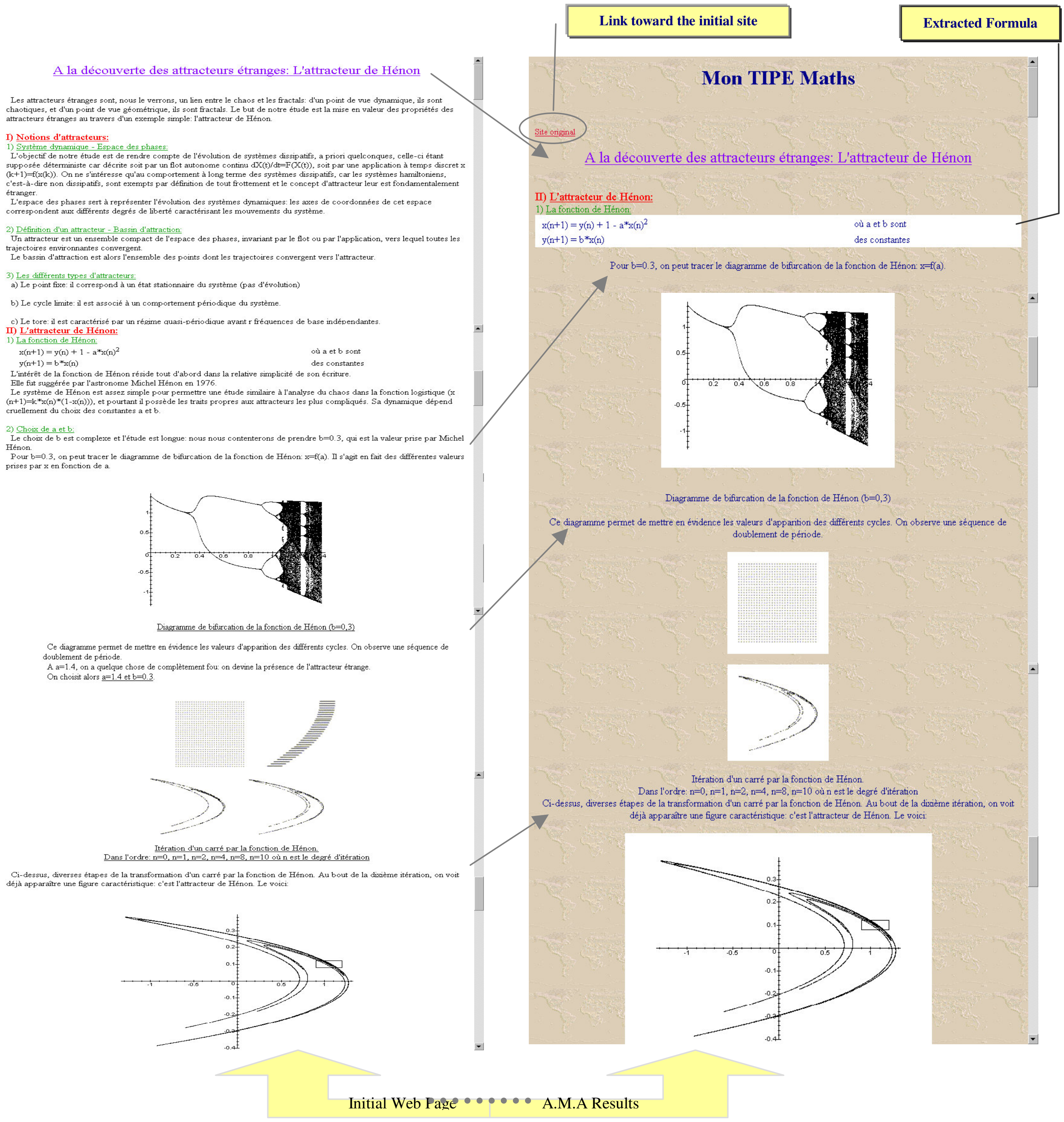
Fig. 5. Architecture overview.

In the following section, a representative example of a mathematic formula is showed, including its legends and its textual comments.

5.1 Example

We expose through the result presentation of a French web page containing a scientific formulas (mathematics). It illustrates a relevant exemple of Filimage.

This screen copy (on the left) shows the initial web page. The result of the A.M.A (Automatic Matching and Associating) operation is displayed (on the right) and it comes from the automatic treatment completed by Filimage. The relevant textual segments are identified, are extracted and then are associated with their corresponding formulas (tables, graphics…).



A la découverte des attracteurs étranges: L'attracteur de Hénon

Les attracteurs étranges sont, nous le verrons, un lien entre le chaos et les fractals: d'un point de vue dynamique, ils sont chaotiques, et d'un point de vue géométrique, ils sont fractals. Le but de notre étude est la mise en valeur des propriétés des attracteurs étranges au travers d'un exemple simple: l'attracteur de Hénon.

I) Notions d'attracteurs:

1) Système dynamique - Espace des phases:

L'objectif de notre étude est de rendre compte de l'évolution de systèmes dissipatifs, a priori quelconques, celle-ci étant supposée déterministe car décrite soit par un flot autonome continu dX(t)/dt=F(X(t)), soit par une application à temps discret x(k+1)=f(x(k)). On ne s'intéresse qu'au comportement à long terme des systèmes dissipatifs, car les systèmes hamiltoniens, c'est-à-dire non dissipatifs, sont exempts par définition de tout frottement et le concept d'attracteur leur est fondamentalement étranger.

L'espace des phases sert à représenter l'évolution des systèmes dynamiques: les axes de coordonnées de cet espace correspondent aux différents degrés de liberté caractérisant les mouvements du système.

2) Définition d'un attracteur - Bassin d'attraction:

Un attracteur est un ensemble compact de l'espace des phases, invariant par le flot ou par l'application, vers lequel toutes les trajectoires environnantes convergent.

Le bassin d'attraction est alors l'ensemble des points dont les trajectoires convergent vers l'attracteur.

3) Les différents types d'attracteurs:

a) Le point fixe: il correspond à un état stationnaire du système (pas d'évolution)

b) Le cycle limite: il est associé à un comportement périodique du système.

c) Le tore: il est caractérisé par un régime quasi-périodique ayant r fréquences de base indépendantes.

II) L'attracteur de Hénon:

1) La fonction de Hénon:

$x(n+1) = y(n) + 1 - a*x(n)^2$ où a et b sont
$y(n+1) = b*x(n)$ des constantes

L'intérêt de la fonction de Hénon réside tout d'abord dans la relative simplicité de son écriture.

Elle fut suggérée par l'astronome Michel Hénon en 1976.

Le système de Hénon est assez simple pour permettre une étude similaire à l'analyse du chaos dans la fonction logistique (x(n+1)=k*x(n)*(1-x(n))), et pourtant il possède les traits propres aux attracteurs les plus compliqués. Sa dynamique dépend cruellement du choix des constantes a et b.

2) Choix de a et b:

Le choix de b est complexe et l'étude est longue: nous nous contenterons de prendre b=0.3, qui est la valeur prise par Michel Hénon.

Pour b=0.3, on peut tracer le diagramme de bifurcation de la fonction de Hénon: x=f(a). Il s'agit en fait des différentes valeurs prises par x en fonction de a.

Diagramme de bifurcation de la fonction de Hénon (b=0,3)

Ce diagramme permet de mettre en évidence les valeurs d'apparition des différents cycles. On observe une séquence de doublement de période.

A a=1.4, on a quelque chose de complètement fou: on devine la présence de l'attracteur étrange.

On choisit alors a=1.4 et b=0.3.

Itération d'un carré par la fonction de Hénon.
Dans l'ordre: n=0, n=1, n=2, n=4, n=8, n=10 où n est le degré d'itération

Ci-dessus, diverses étapes de la transformation d'un carré par la fonction de Hénon. Au bout de la dixième itération, on voit déjà apparaître une figure caractéristique: c'est l'attracteur de Hénon. Le voici:

Mon TIPE Maths

Site original

A la découverte des attracteurs étranges: L'attracteur de Hénon

II) L'attracteur de Hénon:

1) La fonction de Hénon:

$x(n+1) = y(n) + 1 - a*x(n)^2$ où a et b sont
$y(n+1) = b*x(n)$ des constantes

Pour b=0.3, on peut tracer le diagramme de bifurcation de la fonction de Hénon: x=f(a).

Diagramme de bifurcation de la fonction de Hénon (b=0,3)

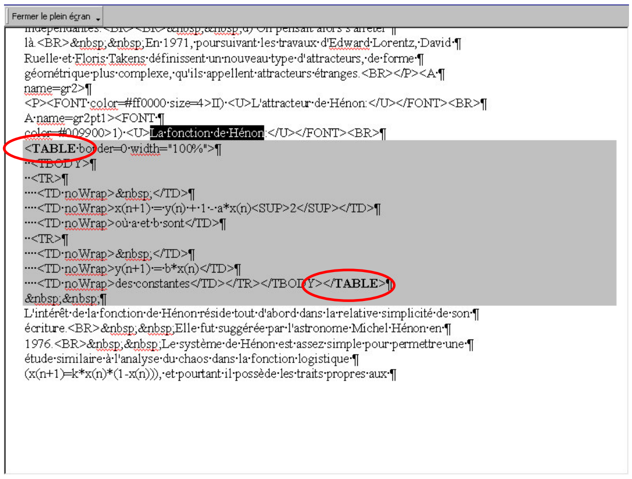
Ce diagramme permet de mettre en évidence les valeurs d'apparition des différents cycles. On observe une séquence de doublement de période.

Itération d'un carré par la fonction de Hénon.
Dans l'ordre: n=0, n=1, n=2, n=4, n=8, n=10 où n est le degré d'itération
Ci-dessus, diverses étapes de la transformation d'un carré par la fonction de Hénon. Au bout de la dixième itération, on voit déjà apparaître une figure caractéristique: c'est l'attracteur de Hénon. Le voici:

Fig. 6. In this example a whole relevant components (on the right) are extracted (title, textual comments, formulas, captions, diagrams…).

This Web page is coming from: www-ensimag.imag.fr/eleves/Guillaume.Molleda/tipe2.htm

The screen copy of the source code is corresponding to the previous example and shows how the specific HTML mark-up tags allow to identify this mathematical formula. It is encoded as a table object: <TABLE>,</TABLE>. We must notice that the object is automatically extracted whatever the visual entity's type inserted into a web page.



```
indépendantes.<BR><BR>&nbsp;&nbsp;d)·On·pensait·alors·s'arrêter·¶
là.<BR>&nbsp;&nbsp;En·1971,·poursuivant·les·travaux·d'Edward·Lorentz,·David·¶
Ruelle·et·Floris·Takens·définissent·un·nouveau·type·d'attracteurs,·de·forme·¶
géométrique·plus·complexe,·qu'ils·appellent·attracteurs·étranges.<BR></P><A·¶
name=gr2>¶
<P><FONT·color=#ff0000·size=4>II)·<U>L'attracteur·de·Hénon:</U></FONT><BR>¶
A·name=gr2pt1><FONT·¶
color=#009900>1)·<U>La·fonction·de·Hénon:</U></FONT><BR>¶
<TABLE·border=0·width="100%">¶
··<TBODY>¶
··<TR>¶
····<TD·noWrap>&nbsp;</TD>¶
····<TD·noWrap>x(n+1)·=·y(n)·+·1·-·a*x(n)<SUP>2</SUP></TD>¶
····<TD·noWrap>où·a·et·b·sont</TD>¶
··<TR>¶
····<TD·noWrap>&nbsp;</TD>¶
····<TD·noWrap>y(n+1)·=·b*x(n)</TD>¶
····<TD·noWrap>des·constantes</TD></TR></TBODY></TABLE>¶
&nbsp;&nbsp;¶
L'intérêt·de·la·fonction·de·Hénon·réside·tout·d'abord·dans·la·relative·simplicité·de·son·¶
écriture.<BR>&nbsp;&nbsp;Elle·fut·suggérée·par·l'astronome·Michel·Hénon·en·¶
1976.<BR>&nbsp;&nbsp;Le·système·de·Hénon·est·assez·simple·pour·permettre·une·¶
étude·similaire·à·l'analyse·du·chaos·dans·la·fonction·logistique·¶
(x(n+1)=k*x(n)*(1-x(n))),·et·pourtant·il·possède·les·traits·propres·aux·¶
```

Fig. 7. Source code of the web page presented in Fig. 4.

5.2 Subsequent results using

The creation of file cards can be destined to the page web summarizing offering a new use to the extracted and stored components. It means that all the relevant extracted objects (formulas, frames…) and contents (titles, subtitles, textual comments, captions, sources) can be exploited once more. Indeed, the results are oriented towards one and/or the other one according to the objects required. Thus, the adapted solution can be destined to the indexation of the visual objects.

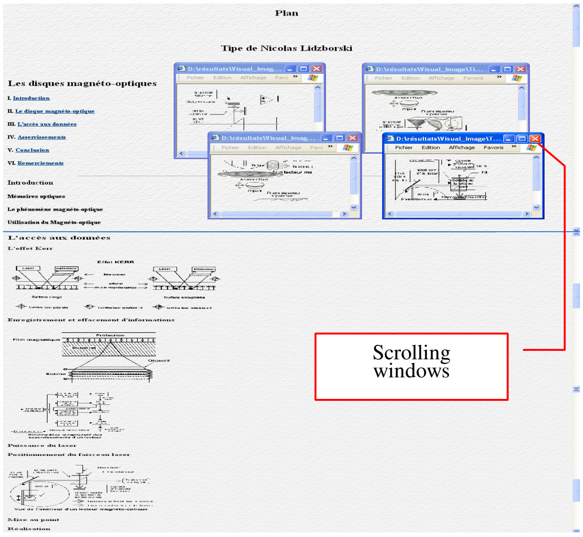


Fig. 8. Screen copy of the windows results.

The two display options offer the possibilty for a subsequent results access by scrolling windows. The extracted visual objects are preceded by the title and the subtitles of the page in order to keep the hierarchic structure of the initial document. We can see in the following Fig. 8. these optional functionalities (visualization of the windows simultaneously or not).

6 Conclusion

This new synthesis system is a high specification level developed in order to automate the design specifically applied to the mathematical formulas. Filimage system contributes to the systematic exploration of the e-document's structure (formulas, title, caption, frame…) in an optimal way and offers the ability to automate the extraction task. As a result, the system is built for re-using, manipulating the formula, the special symbols and the relevant contents. Often, the web pages contain the mathematics formulas and their so particular textual expressions. Therefore, the obtained results coming from tested document confirm the efficiency of the computational method not only for text semantic filtering but also for the multimedia objects extraction. The A.M.A (Automatic Matching and Associating) operation destined to establish the relation between the extracted components and allows a visualization of the results. This process reduces the initial document's volume and provides great flexibility producing an automatic optimized displaying. The mains interest of our strategy are presented. The innovation of this system is to provide an extraction solution to multimedia contents. The extracted visual and textual objects is operating as a semantic entity enable to participate in setting up the organizational more personalized e-learning process. Furthermore, the establish relation between the different objects is to be especially adapted to the human perception and understanding teaching. In this way, the developed strategy is particularly significant on-line and takes a leading part in interactive training. Besides, Filimage get the ability to evolve towards a multilingual filtering thanks to the possibility of integrating the languages through the incremental linguistic databases. One of the further research directions is to pursue the experiments improving an online user's interface.

References:

[1]Y. Bai, D. Qi, Q. Pu, N. Mastorakis, A data mining algorithm based on genetic algorithm, *The World Scientific and Engineering Academic Society (WSEAS),* Taiwan, 2004.

[2] Sh. Behnami, Filimage: Automatic Matching and Associating (A.M.A) for Extraction of E-document Components, *The World Scientific and Engineering Academic Society (WSEAS),* Crete, Greece, 2004.

[3] Sh. Behnami, Filimage System: Web's Images and Texts Automatic Extraction, *The World Scientific and Engineering Academic Society (WSEAS),* Izmir, Turkey, 2004.

[4] Sh. Behnami, *Filtrage sémantique des commentaires textuels associés aux images des documents électroniques*, Thèse de doctorat, Université Paris-Sorbonne, 2003.

[5] A. E. Cawkell, An introduction to Image Processing and Picture Management, *Journal of Document and text Management*, 1(1), 1993.

[6] T.C. Craven, Abstracts produced using computer assistance, *JASIS*, 518:745-756, 2000.

[7] J.P. Desclés, Ingénierie linguistique: enjeux, domains and methods, *conference à l'université Klément d'Okrid*, Sofia, Bulgarie, 2003.

[8] J.P. Desclés, Systèmes d'exploration contextuelle, Co-texte et calcul du sens, (Claude Guimier), *Presses universitaire de Caen*, pp. 215-232, 1997.

[9] R. Fidel, The image retrieval task: implications for the design and evaluation of image databases, *The New Review Hypermedia and Multimedia*, vol. 3, 181-199, Taylor Graham, London, UK, 1997.

[10] M. Pinto, F.-W. Lancaster, Abstracts and abstracting in knowledge discovery, *Library Trends*, 48 (1) 234-248, University of Illinois, 1999.

[11] F. Piroi, B. Buchberger, An environment for bulding mathematical knowledge libraries, *In A. Aspert & Al, editors, MKM conference, Springer Verlag*, 2004.

[12] E. Svenonius, Access to nonbook materials: The limits of subject indexing for visual and aural languages, *ASIS*, 458: 600-606, 1994.

[13] M., Shephred & Al. The role of user profiles for News Filtering, Journal of the American Society for Information Science and Technology, *JASIS*, 522: 149-160, 2001.

[14] H. Wang, O. Lin, Q. Pu, N. Mastorakis, A distributed adaptative learning environment, *The World Scientific and Engineering Academic Society (WSEAS),* Salzburg, Austria, 2004.