Automatic Annotation of Images, Pictures or Videos Comments for Text Mining Guided by No Textual Data

Florence Le Priol

LaLIC, Paris-Sorbonne University
28 rue Serpente, 75006 Paris, France
Florence.Le_Priol@paris-sorbonne.fr

Abstract

The Text mining guided by No Textual data (TNT) is not intended to extract the information contained in the images, aiming the information included in the text that describes these images. In other words, it aims to present to the reader the information about the images next to them, regardless of its real position in the document.
Reading, while focusing on no textual data (images, pictures, videos, sounds, etc.), their caption and some comments identified in the text is faster than reading the entire text and gives a good summary of the document.
We present in this article, a tool for automatic text mining guided by no textual data (fixed or moving images, photographs, video, sound ...).
This tool, based on contextual exploration method and EXCOM platform, can automatically annotate the information concerning no textual data, it can make the link between annotated text and corresponding no textual data and propose navigations among these textual and no textual information.

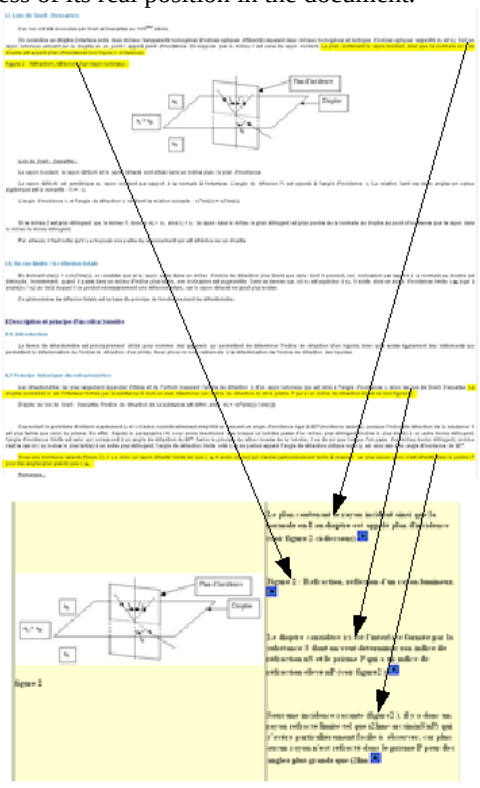
Introduction

The text mining has developed in relation on the type of data, such as tabular data, text, images… However, the objects taking into account are, in reality, not independent.
Let us consider the example of the medical records of a person (Boussaid & al. 05). You can find tabular data such as laboratory tests, textual data as exams and clinical records, X-rays, ultrasound, electrocardiogram curves ...
Searching these complex structures supposes to treat one type of media at the same time, independently of the others. Another example comes from web mining where html documents are treated as a corpus of texts and the other data such as images or sound are ignored, whereas they may also contain some information.
The new challenge of text mining is to take into consideration all types of data simultaneously.
In the present article, we propose an automatic tool for text mining to treat both textual and no textual data (images, pictures, videos, sounds ...). This tool based on the contextual exploration method (Desclés 97) is developed in EXCOM platform (Djioua & al. 06).

System description

Image mining versus text mining guided by images

The image mining includes all the techniques that has an interest for the content of the image. It is similar to image analysis, a subset of signal processing and recognition of objects.
The image mining, therefore, is to identify the forms in a digital image, and by extension in a digital video stream and provide a quantitative description of the image.
The text mining guided by the images is not intended to extract the information contained in the images, it aims to take into account the information contained in the text that describes the images. In other words, it aims to present to the reader the information about the images next to them, regardless of its real position in the document.

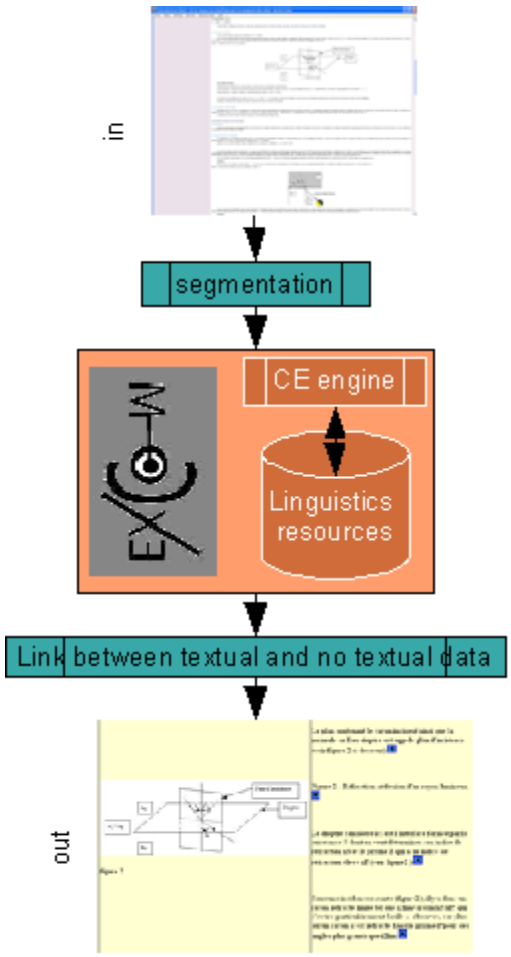


In the previous example, the information is located just

above the image, then several paragraphs below. Once identified, this information can be presented in terms of the image.
The text mining guided by the images, identifies the information describing these images.

Process

The process that we propose is not going to conduct a text mining guided only by the images but by all no textual data, ie images, pictures, video, sound ...
The annotation of information about no textual data is automatic with the platform EXCOM and the contextual exploration (CE) method.
Before running EXCOM, the text is segmented in paragraphs and sentences and after the links between textual and no textual data are built.
The entire process is automatic and implemented with perl, xslt, php and mysql.



Automatic Annotation

EXCOM (Djioua & al. 06) implements contextual exploration.
The main hypothesis of contextual exploration is that the text contains, linguistics indicators independent of knowledge domain. These indicators may however not be enough to assign an annotation to the segment. In this case, clues are sought in the context of the indicator.
The linguistics resources are therefore composed by the annotation, markers (indicators and clues) and rules.
Based on the linguistics resources (markers, rules and annotations) established for the resolution of the text mining guided by the no textual data, the platform EXCOM annote automatically segment with no textual data (annotation : 'notextualdata'), captions (annotation : caption) and all segments that make textual reference or describe no textual data (annotation : 'directcomment').
The text of captions is used in a second phase of the treatment to complete the text mining (annotation : 'indirectcomment').

For example, the rule 'TNT_01' annote a segment with no textual data.

```
<rule name="TNT_01" task="textuelnontextuel"
type="EC">
   <conditions>
      <indicator space_research="sentence"
      type="annotation" value="img"/>
   </conditions>
   <actions>
      <annotation  space="sentence"
      type="notextualdata"/>
   </actions>
</rule>
```

The other example, the rule 'TNT_08' annote a segment commenting on an image. This rule is triggered, for English text, by an indicator of class 'elementnontextuel_en' which includes words like 'graph', 'picture', 'map', 'scene'... It is supplemented by the clue placed on the right of the indicator belonging to the class 'montre_en' (show, indicate...), 'preter_en' (lending...) ou 'extraire_en' (extract...) and the clue placed on the left belonging to the class 'demonstratif'_en (this, these...).

```
<rule name="TNT_08" task="textuelnontextuel"
type="EC">
   <conditions>
      <indicator space_research="sentence" type="list"
      value="elementnontextuel_en"/>
      <clue context="right" space_research="same"
      type="list" value="montrer_en|preter_en|
      extraire_en"/>
      <clue context="left" space_research="same"
      type="list" value="demonstratif_en"/>
   </conditions>
   <action>
      <annotation space="sentence"
      type="directcomment"/>
   </action>
</rule>
```

These rules allow annotating text like :

carefully preserved to help retard slides. If you need more detail, we have a version twice this size.

This photo from False Mt. Troy on Douglas Island shows the urban avalanche paths on the Gastineau Channel side of Mt. Juneau.

The following extract of annotated file corresponds to the text above :

...
<para id="12">
<sentence id="25"
annotation="lalic.excom.textualnotextual.notextualdata"
rule="TNT_01"><indicator cat="img"
src="alaska_en_fichiers/20070214MtJuneauPaths.jpg"
rules="TNT_01" rule="TNT_01"/>
</sentence>
</para>
<para id="13">
<sentence id="26"
annotation="lalic.excom.textualnotextual.directcomment"
clues=" This,shows" position_clues="1,49"
rule="TNT_08"><before_clue/> This <after_clue/>
<indicator rules="TNT_08,TNT_05a" rule="TNT_08">
photo </indicator> <before_clue> from False Mt. Troy on Douglas Island </before_clue> shows<after_clue> the urban avalanche paths on the Gastineau Channel side of Mt. Juneau</after_clue>
</sentence>
</para>

Using captions to complete the text mining

In some domains like biology, captions consist of several sentences. It is interesting to be able to use the text of the caption to complete the text mining. Using the key terms of the caption as markers we can activate the same type of rule as described above but with annotation 'indirectcomment' like in the rule 'TNT_12'.

<rule name="TNT_12" task="textuelnontextuel"
type="EC">
<conditions>
<indicator space_research="sentence" type="list"
value="caption_terms"/>
</conditions>
<action>
<annotation space="sentence"
type="indirectcomment"/>
</action>
</rule>

For example, in this text, we can associate a sentence with the caption with the term 'T. weissflogii'.

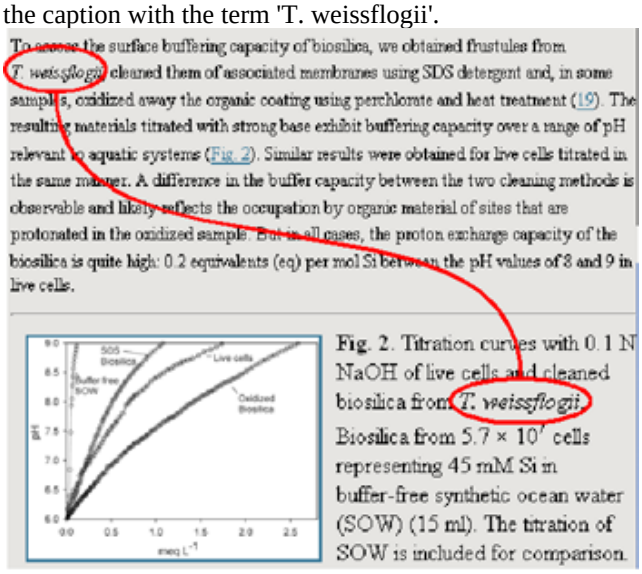

To assess the surface buffering capacity of biosilica, we obtained frustules from *T. weissflogii*, cleaned them of associated membranes using SDS detergent and, in some samples, oxidized away the organic coating using perchlorate and heat treatment (19). The resulting materials titrated with strong base exhibit buffering capacity over a range of pH relevant to aquatic systems (Fig. 2). Similar results were obtained for live cells titrated in the same manner. A difference in the buffer capacity between the two cleaning methods is observable and likely reflects the occupation by organic material of sites that are protonated in the oxidized sample. But in all cases, the proton exchange capacity of the biosilica is quite high: 0.2 equivalents (eq) per mol Si between the pH values of 8 and 9 in live cells.

Fig. 2. Titration curves with 0.1 N NaOH of live cells and cleaned biosilica from *T. weissflogii*. Biosilica from 5.7×10^7 cells representing 45 mM Si in buffer-free synthetic ocean water (SOW) (15 ml). The titration of SOW is included for comparison.

The following extract of annotated file corresponds to the text above :

<para id="7" >
<sentence id="49"
annotation="lalic.excom.textuelnontextuel.indirectcommen
t" rule="TNT_12"> To assess the surface buffering capacity of biosilica, we obtained frustules from <indicator rules="TNT_12" rule="TNT_12"> T. Weissflogii </indicator>, cleaned them of associated membranes using SDS detergent and, in some samples, oxidized away the organic coating using perchlorate and heat treatment (19).
</sentence>
<sentence id="50"
annotation="lalic.excom.textuelnontextuel.directcomment"
rule="TNT_03"> The resulting materials titrated with strong base exhibit buffering capacity over a range of pH relevant to aquatic systems (<indicator cat="reference" rules="TNT_03" rule="TNT_03">Fig. 2</indicator>).
</sentence>

Link between textual and no textual data

It is not enough to annotate text segments as a caption or a direct or indirect comment, it is also important to link the commentary with the proper no textual object.
Building links between annotated textual data and no textual data is carried out automatically according to three situations.

The first situation occurs when the annotated sentence with

reference is tied to a no textual object and a comment resumed this reference. In this case, we have situations more or less complex. The simplest situation is to have the same term in the caption and the sentence like in the following example (the reference to the caption is 'FIG. 1.', The reference of the comment is 'Fig. 1'). More complex situations are when references are not exactly the same, for example the caption says 'Figure 1', the comment 'fig.1a' or comment refers to several figures, 'figures 1 to 3' ... In these cases, the process will link caption and comment ('Figure 1' must be linked with 'Fig. 1a' ; comment with 'figures 1 to 3' must be linked with figure 1 and figure 2 and figure 3).

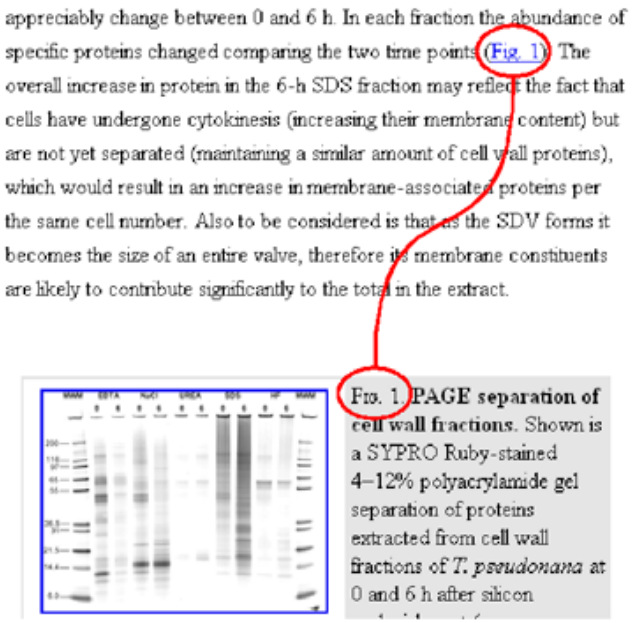
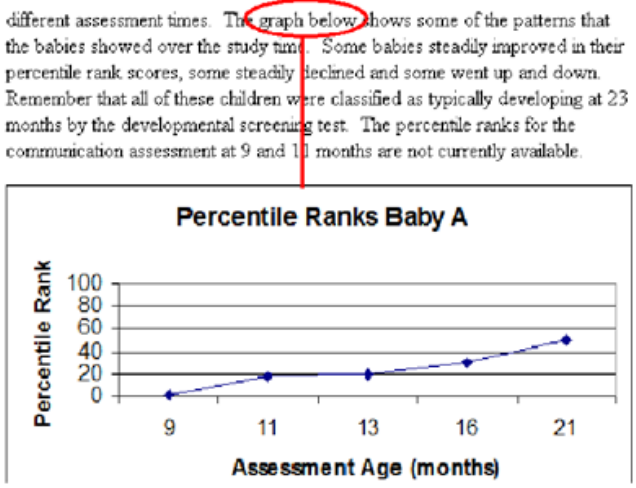

appreciably change between 0 and 6 h. In each fraction the abundance of specific proteins changed comparing the two time points (Fig. 1). The overall increase in protein in the 6-h SDS fraction may reflect the fact that cells have undergone cytokinesis (increasing their membrane content) but are not yet separated (maintaining a similar amount of cell wall proteins), which would result in an increase in membrane-associated proteins per the same cell number. Also to be considered is that as the SDV forms it becomes the size of an entire valve, therefore its membrane constituents are likely to contribute significantly to the total in the extract.

FIG. 1. **PAGE separation of cell wall fractions.** Shown is a SYPRO Ruby-stained 4–12% polyacrylamide gel separation of proteins extracted from cell wall fractions of *T. pseudonana* at 0 and 6 h after silicon

The second situation is such that the comment refers to the no textual object with markers; for example, a sentence specifying the no textual object location such as 'above', 'below' ... In this case, the comment is related to the no textual object with the significance of the marker.



We also have markers that do not indicate the position of the no textual data. In this case we propose some hypothesis.
For example, a sentence annotated 'directcomment' by markers as 'this figure' (cette figure) is linked to the previous no textual object.

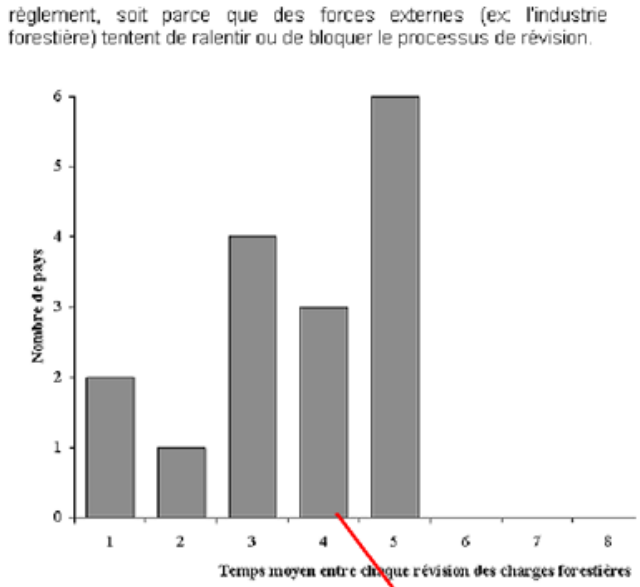

règlement, soit parce que des forces externes (ex: l'industrie forestière) tentent de ralentir ou de bloquer le processus de révision.

Figure 6 Nombre de pays réexaminant ou révisant leurs taxes forestières à des intervalles de temps différents

La figure 6 montre à quelle cadence les pays ont révisé leurs taxes forestières au cours des années récentes. Cette figure concerne principalement la révision des redevances sur la production de bois ronds. Dans certains pays, si l'on tient aussi compte d'autres taxes

The third situation occurs when the sentences are annotated with 'indirectcomment'. This sentences are linked to the caption. The caption is linked to the no textual data with reference (first situation) or with marker (second situation). In this case, the comment is attached to the no textual data by the caption.

Functionalities

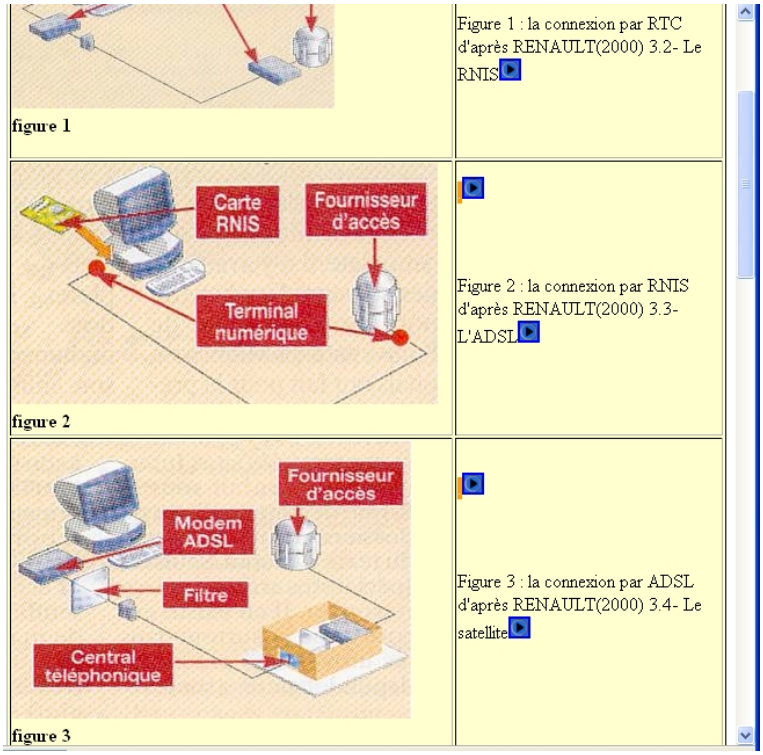
This process allows to build a navigation tool where text and no textual objects are directly linked : no textual data on the left, textual data on the right. The results are indexed and stored for easy use.

Navigation text by text

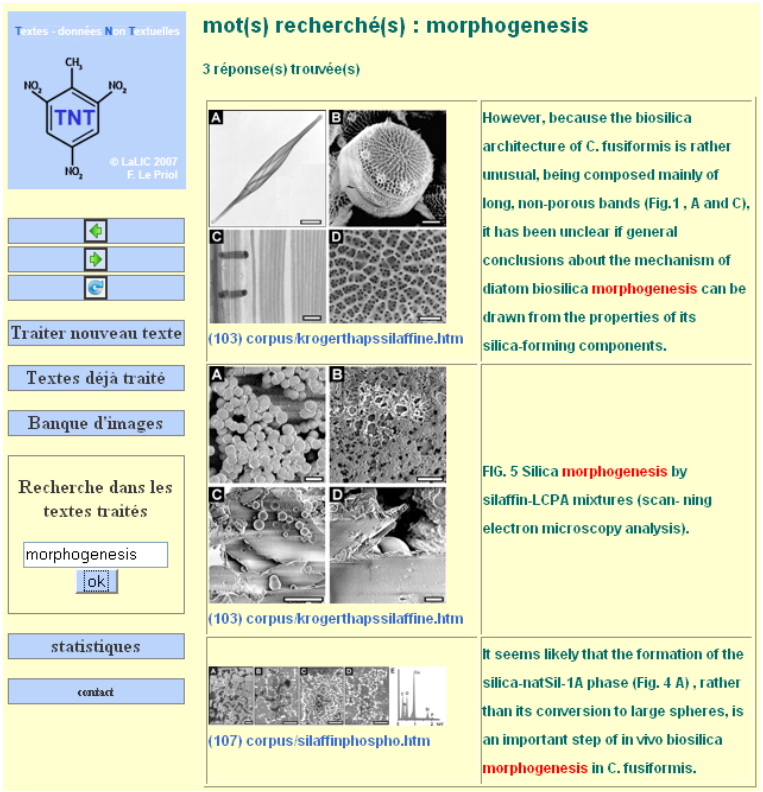
It is possible to choose a navigation text by text by selecting the text to be viewed from the list of treaty texts. In this case, the no textual data belong only to the selected text.
As an example, we show at the end of the article the result obtained for a text in French on “cours de pharmacie sur internet”.

Navigation multi-text

For a multi-text navigation, we should use the intern search engine to view all the images in relation to the search term. This mode allows to have the entire informations about a



Navigation text by text



Navigation multi-text

subject and to discover eventually the polysemy of the searched term.
For example (at the end on this paper), research on "morphogenesis" gives three answers from two texts.

Navigation by no textual data
We usually access the information by a text or by keywords. It is possible to access information while surfing on the basis of no textual objects. One can just click on the image to get comments as a result.

Back to full-text
No matter what kind of navigation we choose, text by text, multi-text or by no textual data, we can always return to full-text to see the context of the no textual data or the context of the comment and to have a complete reading of the document.

Future work

If the text mining guided by no textual data is possible in many knowledge domain, this is particularly interesting in biology where arguments are based on no textual data. Reading an article can be simply faster by looking at the no textual data and captions. This is important while a researcher has to get through an important number of documents. That is why few tools were already proposed for this discipline.
FigSearch is a prototype text-mining and classification system for figures from any corpus of full-text biological papers (Liu & al. 04). The BioText Search Engine (Hearst & al. 07) has ability to search and browse figures and their captions.
However while these tools treat only the captions, the TNT-EXCOM we have presented in this paper takes into account the captions and the comments in the text (identified by reference markers or the terms of the caption). Thus, the information given with the no textual object seems more comprehensive than using existing tools.
The future work is to develop this tool specifically for biology, working with an expert of the domain.

Conclusion

Text mining guided by no textual data is an original and interesting method to access information.
Reading, focusing on no textual data (images, pictures, videos, sounds, etc.) and their caption, as some existing tools in biology do, is fast and gives a good summary of the document. However, this reading can become even better if we could add extracted comments from the text. We have shown that the contextual exploration method is appropriated to annote comments of no textual data and that this tool has a particular interest in biology even it is possible to use it in many different knowledge domains.

References

BEHNAMI S., 2002, Filtrage sémantique des commentaires d'images pertinentes à partir de documents numériques, Thèse de doctorat, Univ. Paris-Sorbonne

BOUDRY C., 2005, Traitement et analyse d'images, URFIST de Paris/Ecole des Chartes, http://www.ext.upmc.fr/urfist/image_numerique/home_image.htm

BOUSSAID O., GANCARSKI P., MASSEGLIA F. & TROUSSE B. (dir.), 2005, Fouille de Données Complexes, revue des nouvelles technologies de l'information (RNTI), Edition Cépaduès

DESCLES J-P., 1997, Système d'exploration contextuelle, in Co-texte et calcul du sens, 215-232. Calif : eds Guimier, Presses Univ. Caen

DESCLES J-P., JOUIS C., OH H-G., REPPERT D., 1991, Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte., in Knowledge modeling and expertise transfert, Eds D. Hérin-Aime, R. Dieng, J-P. Regourd, J-P. Angoujard, 371-400. Calif : IOS Press

DJIOUA B., GARCIA FLORES J., BLAIS A. DESCLES J-P., GUIBERT G. JACKIEWICZ A., LE PRIOL F., NAIT-BAHA L., SAUZAY B., 2006, EXCOM: an automatic annotation engine for semantic information, FLAIRS 2006, Floride, 11-13 Mai

HEARST MA., DIVOLI A., GUTURU H., KSIKES A., NAKOV P., WOOLDRIDGE MA. YE J., 2007, BioText Search Engine : beyond abstract search, Bioinformatics, vol. 23, no 16, pp. 2196-2197

LE PRIOL F., 2000, Extraction et capitalisation automatiques de connaissances à partir de documents textuels. Seek-Java : identification et interprétation de relations entre concepts, Thèse de doctorat, Univ. Paris-Sorbonne

LIU F., JENSSEN TK., NYGAARD V., SACK J., HOVIQ E. , 2004, FigSearch : a figure legend indexing and classification system, Bioinformatics, vol. 20, no 16, pp. 2880-2882

MOURAD G., 2001, Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE, Thèse de doctorat, Univ. Paris-Sorbonne