ISSN 2278-5612

Research system of semantic information in medical videoconference based on conceptual graphs and domain ontologies

Zwidi Afef[1], Yengui Ameni[2], Neji Mahmoud[3]
MIRACL, University of Sfax, Tunisia
[1]Afefsultannezwidi@yahoo.com
[2]Ameni_iseah@yahoo.fr
[3]Mahmoud.neji@fsegs.rnu.tn

ABSTRACT:

The multiplication of the number of AudioVisual Documents (AVD) engendered a problem while searching for information within gigantic databases of which we are incapable to index their contents completely manually. Indeed, several complex difficulties are put by these documents because of the vertiginous increase of the quantity of the multimedia data to be treated and the specification met in the representation and the extraction of their contents in particular semantics of the fact that these documents contain three types of media (text, sound, image). AVDs can be classified in professional broadcasted videos (movies, emissions), sporting videos, video controlling, videoconference… etc. In this paper, we propose a model of representation of the semantic contents of videoconferences' documents in medicine based on the conceptual graphs taking into account the different modalities. This model is based on the concepts' extraction and the semantic relations between them and appeals ontology domain.

Keywords:

Semantic indexation, concept, semantic relations, conceptual graphs, audiovisual document, medical ontology.

Council for Innovative Research
Peer Review Research Publishing System
Journal: International Journal of Management & Information Technology

Vol. 7, No. 2
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 2278-5612

INTRODUCTION

Nowadays, we assist a continuous development of information technology. These new technologies have enabled the rapid development of material production technology and information management. The progress of production tools of information such as video conferencing has enabled the production of a huge amount of information. This rapid increase in the volume of information has created the problem of "how to find information that interests us in this great mass of information?"

To treat this problem, the IRS has been developed in order to select from a volume of information, the pertinent information vis-à-vis an information need. These SRI aim to connect two representations; one is of user needs, and the other is of the documents' content using a correspondence function.

The system we are building consists of three modules, ontologies, a documents' database and an index database. The system's modules are: the analysis module, the indexing and the search module shown in **Figure 1**.

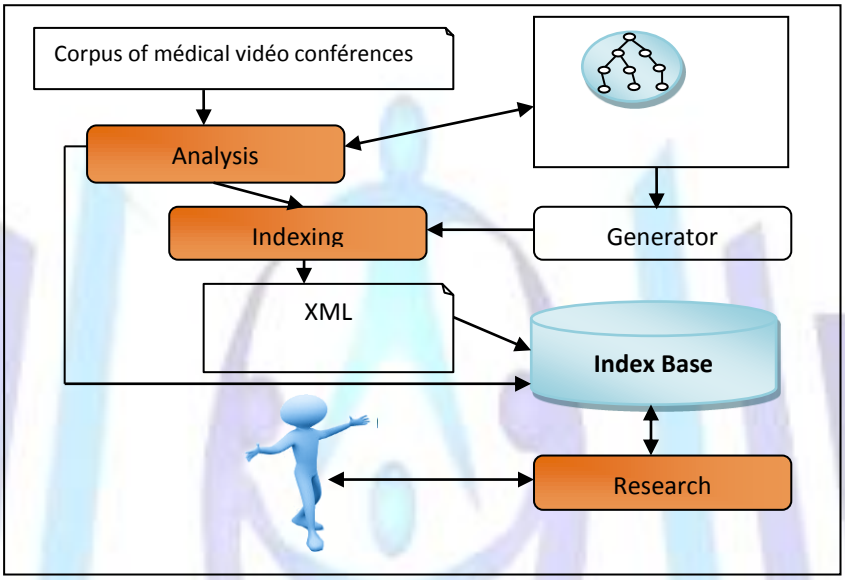

Figure 1: General architecture

The analysis module (also called modeling module) is responsible for navigating and analyzing the documents that are in the documents' database. It is to define a model for the description of media content that are appropriate for the representation of the videoconferences' content. This model must take into account the information's items in different levels of descriptions and the various present media which are adapted to be integrated into an indexing and retrieval system. The analysis's result is an XML file.

The indexing module is in charge of navigating and indexing the documents which are the results of analysis module using the concepts of the ontology. The result of the indexing is stored in a database index. Indexing a document is done by the system administrator who submits the address of the document to be indexed to the indexing module.

In case of a user wanting to search a videoconference, the request will be sent to the search module.

In what follows, we will describe only the analysis and indexing modules of our system.

Ontologies have considerably improved the pertinence of results in the search for audiovisual documents. That is why we opted for the method of indexing using ontology. This improvement is due to the fact that the indexing process takes on consideration the different concepts and relations between these concepts (given by the ontology). Therefore, unlike the methods based on simple and static keywords that interest in whether a word exist or not in a document, this method takes into account the semantics of the terms to search for.

The next part of our paper is devoted to present, firstly, the analysis module (modeling) of videoconferences documents. And, it describes, secondly, the indexing module of videoconferences document completed by the use of ontologies domain.

1. ANALYSIS MODULE

The information of videoconference content can be represented in several levels: physical information consisting of binary data of the content that is not usable by the computer, and description information that can transform physical information into exploitable knowledge by the user, this strengths the interface between man and machine, and then easily exploit the video content. To provide multiple levels of abstraction in the exploitation of the videoconference content, we propose a modeling schema in two levels: structure and semantics.

Modeling the structure of the content describes the organization that may represent the information content. This model is often based on the classical structure of a video document (the sequence, the scene, the plan). The descriptions in this level are automatically calculated using visual descriptors. This level of modeling is free from all semantic description.

Videoconferences are hierarchically structured in scenes, maps and images. This structure reflects the process of creating the videoconference. We are only interested in the visual descriptors to clear this hierarchical organization. It is defined by means of a top-down approach via successive niceties. For a user, browsing a hierarchical structure is certainly easier than navigating a flat structure.

The main interest of this organization is that it can be automatically extracted by releasing the semantic content.

As for the semantic modeling, it is an abstraction that allows you to link low level descriptions of the real world. A modeling schema for describing the meaning of the descriptions located at the structure. We exploit the notion of concepts and conceptual relationships to present occurrences (information items) described in the structure part.

We will use the conceptual graph formalism to present different ideas in a videoconference using concepts and relationships. It is recommended to describe the contents of each of the structural elements to extract excerpts from the videoconference that answer specific requests by navigating conceptual graphs.

The conceptual graph model is a modeling approach having the specificities to be formal, to represent knowledge and to be concrete in the sense that there are efficient tools for manipulating modeled knowledge. It allows modeling the knowledge of a domain using graphs, based on a support. This modeling approach is intentional, and is provided in semantic first-order logic, and assumes a closed world for its reasoning **[4].**

In our modeling approach, we develop a modeling schema of semantic knowledge combining the two types of modeling (hierarchical and semantic) and regardless to the videoconference content, that's to mean that we can apply it on all videoconferences.

1.1. Modeling of visual content

In every facet defining correspondence with an informative content type on all images objects, is associated a model describing the images objects, the relations between them and the operations defined on these descriptions. The specified facets are supplementary in order to go with the interpretations that model and each instance of the general model is a combination of facets to translate the wealth of pertinent characteristics of the images.

Two main different categories of facets: the physical facet which represents the entity perceived by the human eye in its plane and two-dimensional representation, and logical facet collecting the interpretations of the image and all of its most semantic descriptors. The logic facet is subdivided into four facets whose combination provides the symbolic characterization of the image: structural, spatial facets, symbolic and signal.

a) The structural facet

The structural facet represents the decomposition of an image into image objects. Each image object can be decomposed into sub-objects images. The composition relation associated with this facet is the relation "contains" that involves spatial inclusion, the regions corresponding to the component objects are included in the geometric boundaries of the region described by the decomposed object.

The structural facet is represented by a conceptual graph whose nodes are the image objects and the arcs are the instances of the composition relation.

b) *The spatial facet*

The spatial facet describes the geometrical information on relative to spatial objects associated with the images objects as well as the spatial relations between them. This facet allows the characterization of an image's objects by their shapes and their relative positions. A spatial object is defined by giving of a geometric shape (point, segment, polygon) corresponding to its contour. The spatial facet is represented in the classical spaces in order to give the model a greater generality. These spaces are first the Euclidean space combining the notions of scalar product, orthogonality, angles and standards. This space allows operations such as the calculation of the centroid of the area, length, width, height and the polygon encompassing [8].

The spatial sub-facet to specify spatial relations (relative position, direction) of the image objects. This sub-facet is shown by the following graph:

$$[Io] \rightarrow (SpC) \rightarrow [Io2]$$

Considering Io1 and Io2 two images objects representing respectively concepts "Expert" and "chirurgical operation" in the following example: "The videoconference segments show an expert explaining a surgery." The representation with conceptual graph formalism exploiting a spatial description is described by the graph below:

$$[Io1] \rightarrow (front) \rightarrow [Io2]$$

c) The symbolic facet

The symbolic facet is the representation of the semantic content of an image and is defined as a given symbolic objects associated with the images objects as well as relations corresponding to the description of scenes or actions involving those objects. The symbolic facet is trying to take into account the multiple interpretations regarding the semantics conveyed by the image. It is strongly constrained by the application to the extent that the term "sense" is related to the comprehension of the domain of the application as well as an indexing language chosen to express the relations between the elements of knowledge brought to light [8].

d) Facet signal

Facet signal contains information on the visual content of "low level" such as color, texture, spatial positions, etc. This information is represented in the form of numerical descriptors. In many cases, it is possible to use this information to infer a semantic description by the aggregation of a number of these low-level criteria.

Modeling the signal facet is inspired by the work of Mr. Belkhatir **[3].** The signal facet describes the visual content of the document videoconference in terms of visual perception of videoconference information. It allows you to specify the low-level visual features of videoconference. Formally, we denote the elements of this facet by image descriptors (Ids). These descriptors are not necessarily symbolic, but they can be used to infer semantic descriptions.

The signal facet is divided into four sub-facets:

- The sub-facet color to present the colors characteristics in the visual content of the document. This sub-facet is presented in the following graph:

  [Io] → (has color) → [(<col> AND]

  [Io] → (has color) → [(<col>OR]

  With <col> AND and <col> OR representing respectively a combination of 11 Boolean values representing the concepts “colors” as already presented in [8].

- The sub-facet texture is used to describe the texture properties in the visual content. This sub-facet is presented in the following graph:

  [Io] → (has texture) → [<tex>AND]

  [Io] → (has texture) → [<tex>OR]

  With <tex> AND et <tex>OR representing respectively a combination of 11 Boolean values representing the texture concepts as already presented in [8].

- The sub-facet motion to specify the movements of the camera of the images objects and their trajectories. This sub-facet is presented in the following graph:

  [Io] → (has motion) → [<mvt>AND]

  [Io] → (has motion) → [<mvt>OR]

  With <mvt> AND and <mvt> OR representing respectively a combination of 8 Boolean values representing the motion concepts as already presented in [8].

1.2. Modeling of audio content: Extracting terms

The extracting process consists in detecting terms in a documentary context. A documentary context is defined as a textual unit inside of XML document; it may represent a sentence, a paragraph, or a logic element of the logical structure (the Text nodes in the XML documents). First of all, for each term (or one of synonyms of this term), we seek his presence in the Treated document.

Then, we calculate the number of occurrences of each term in the document which is the cumulative of all terms found and their synonyms. The frequency of occurrence of each word in the document is equal to the number of occurrences of each term in the document divided by the total number of terms in the document.

The purpose of this step is to extract all the terms of the document likely to represent concepts in the ontology. These terms correspond to different inputs (or nodes) in the ontology. For this purpose, we use a technique that consists of projecting the ontology on the document. This is done by the ontology browsing using a parser developed for this reason to identify ontology concepts that occur as terms in the document (detailed in the indexing module)

We set three goals for this: extract simple term, compound terms and specific terms.

ISSN 2278-5612

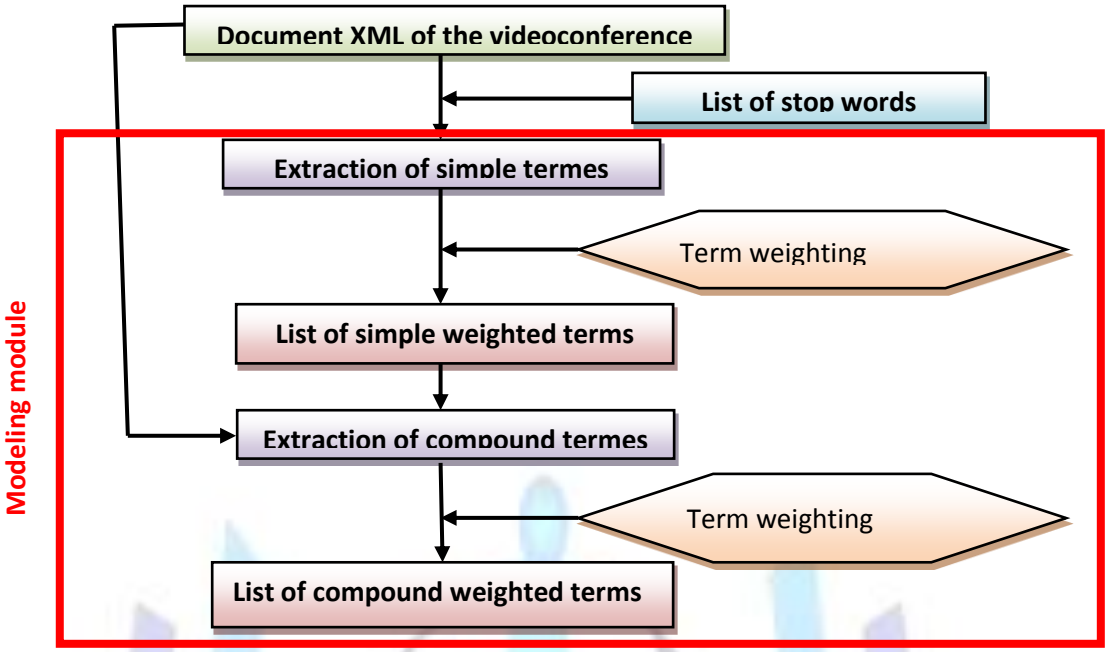

Figure 2: Analysis Module

1.2.1 Extraction of simple terms

a. Extraction of empty words candidate

An empty word candidate is a word likely to be an empty word. In this step, we assign to each word category: empty or full. Stop words (or stop words in English) are words that are common to all the texts in a same language. They have a functional utility. In English, the obvious empty words could be "the", "the", "of", "this", "in", "to", etc…In a monolingual context where all the documents in the corpus are in the same language, empty words are mainly characteristic words of the language such as prepositions, articles, etc.. In this context, empty words are called also grammatical words. So there is no need to index or use in the process of information retrieval. In a text, an empty word is a non-significant unlike a full word.

In contrast to recent work, we keep the pronouns and we consider them as full words as they can refer to simple or compounds words which may be semantically rich.

A term is considered semantically rich if it is either common or specific. In fact, for every non-empty and unweighted term, we check if it is a specific word. A specific word is a low weighted one but rich from its semantic side. It may even have a single occurrence in the videoconference to be indexed for example, the word “mini camera” which appears only once time in the videoconference “laparoscopic surgery in Strasbourg”. This term is considered as a specific term for laparoscopic surgery because it can’t perform this type of surgical operation without introducing this mini camera in the patient digestive system. In this case, despite its low weight, the term a low weighted is considered specific. So it will be added to the list of simple terms.

b. Extraction simply by removing stopwords

To extract the simple terms, we proceed by elimination of empty words. All the words in the corpus consist of two subsets: a subset of empty words and a subset of full words as simple terms. Thus, simple terms are identified by the elimination of empty words from all words that composes the vocabulary of the corpus.

For example, for the following line of the XML document of the videoconference "videosurgery in Starsbourg"

<line> The surgical act ... consists of ... incision of millimeters of few ... **</ line>**

We get:

- Lmv: The, of, few ...

- Lts: surgical, act, incision, millimeters...

c. Weighting simple terms

In this step, we assign a weight to each term that represents its discriminating power and its representative power in the document where it appears.

Indeed, a term doesn't represent adequately the document only if its importance degree in this document is significant. In the literature, we distinguish two types of weighting: local and global.

➢ ***The local weighting***

The local weighting consists of measuring the representative power of a term in a document from the corpus [2]. It uses local information of the term in a given document. This weighting is calculated as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum k \in T_c \; n_{k,j}} \qquad (1)$$

Where:

- $n_{i,j}$ is the number of occurrences of the word i in document j;
- $n_{k,j}$ is the number of occurrences of the term k in the document j. The denominator is the number of occurrences of words in the document in question;
- T_c is the set of terms in the corpus.

➢ ***The overall weighting***

This weighting can assign to a term a measure reflecting its importance in the corpus of documents. A term that appears in the majority of documents is less valuable for distinguishing documents from each other.
In our approach, we are interested in using the local weighting because indexing is done document by document.
Referring to the videoconference "videosurgery in Starsbourg", we find that the term "act" appeared five times in the videoconference. And we find that the total number of simple terms is 189 words, where the weighting of the term "act" is 5/189 = 0.026....

d. Extraction algorithm of simple terms

The algorithm of our approach of extracting simple terms from the documents is as follows:

Extraction algorithm of simple terms of a videoconference

Inputs :
CDV : copy of the XML document of the videoconference // not to change the original document
DMV : document XML of empty words
Lp1 : {who, which, that, those, …}
Lp2 : {latter, former, thereof, …}
Lp3 : {she, he, they, we, …}

Outputs :
Lts: list of simple terms of videoconference
Lm: list of words of videoconference

Variables :
Lmv : list of empty words of the videoconference
m, m1 : words
nbm : number of words non empty in the videoconference
occm : occurrence of a word m
doc : document
ph : sentence
lmphi : list of words of the i^{th} sentence of the videoconference
nph : number of sentences in the videoconference

```
Begin
// Pretreatment: decomposition of the document on words and phrases.
// delete of the empty words from the videoconference
for each sentence ph of doc do
  for each word m in ph do
    if m ∈ Lmv then
       delete m from ph
    end if
  end for
```

```
end for
// Term extraction and weighting
for each lmph_i of doc do
  for each mot m du lmph_i do
    add m to lm
    occm←0
    Nbm←nbm + 1
    J←i
    While (j< = nph) do
      for each word m1 of the sentence lmph_j do
        if m is the first word in lmph_j then
          P←j+1
          While (p<=nph and the first word in lmph_p ∈ LP3) do
            Occm←occm+1
            Delete the first word from lmph_p
            P ←p+1
            end
          Else
         if m is the first word lmph_j then
            if the first word in lmph_{j+1} ∈ LP2 then
              Occm← occm+1
              Delete the first word from lmph_{j+1}
            End if
          else
            if (m1 directly follows m and m1 ∈LP1) then
              Occm← occm+1
              Delete m1 from lmph_j
            else
              if m=m1 then
                 Occm← occm+1
                 Delete m1 de lmph_j
              End if
            End if
          End if
        End if
      End for
    End while
    weight (m)←occm / nbm
    if weight (m)> threshold then
        Add m to lts
     else
      if m is a specific word then
        Add m to lts.
      End if
    End if
  End for
End for
End
```

Algorithm 1: Extraction of simple termes

1.2.2. Extraction of compound terms

a. Extraction of compound terms based on mutual information

To designate a new concept in a field, the principle is to avoid creating a new term which would result in a rapid explosion of the lexicon **[5].** This new term, a compound term, is created from existing lexical data. These compound terms are combinations of two or more words **[7].** With a new concept, there are no new terms, but there are new combinations of words to describe it. These combinations are sequences of words that will be considered as new terms.

In our method, we adapt the approach of F. Harrathi **[5]**. To extract compound terms, we use an iterative and incremental process. It allows discovering new words from existing ones. The process proceeds to extract new terms from an initial list of known terms by using a statistical measure: Adapted Mutual Information (AMI). We start from the list of simple terms. We calculate subsequently the value of the IMA of each pair of words. We do not propose to take into account the frequency of an empty word for the calculation of the IMA. For example, the term « hospital of Strasbourg », the frequency of the empty word « of » will not be taken into account and will be replaced by the value of the frequency of the simple term « hospital ». During the extracting process of compound terms, the term « hospital of » is marked as a « term of construction». This term is deleted at the next iteration. The couple of terms that the value of the IMA is less than a threshold value are accepted as compound terms. The process stops when at iteration no new term is extracted. For a couple of words (m_i,m_j)**,** the adapted mutual information is calculated as follows :

$$\mathrm{CTFij} = \left(1 - \frac{1}{\mathrm{long(ti)}}\right) * \mathrm{TFij} + \frac{1}{\mathrm{long(ti)}} \sum_{\mathrm{K}\in \mathrm{i}} \mathrm{TFk\,j} \qquad (2)$$

The process of extraction of compound terms we use is composed of three steps:

1. **Initialization:** in this step, we initialize the list of composed words by the contents of the list of simple words;
2. **Discovery of new terms:** in this step, we calculate the mutual information between an item from the list of compound terms and a word from corpus;
3. **Adding of new terms:** at this stage, if we find a value of mutual information superior to a given threshold, we add the couple of words to the list of compound terms.

For example, for the following line of XML document of the videoconference "videosurgery in Starsbourg"

<ligne>The surgical act… consists of … incision of few … of millimeters**</ligne>**

And calculating the value of IMA, we get:

$$\mathrm{IMA}\,(m_i,m_j) = \begin{cases} -\log_2\left(\frac{f(m_i,m_j)}{f((m_i)*f(m_j)}\right) & \text{if } m_j \text{ a non} \\ -\log_2\left(\frac{f(m_i,m_j)}{f((m_i)*f(m_i)}\right) & \text{if } m_j \text{ is an empty} \end{cases}$$

- **Lts :** surgical, act, mini, incision, millimeters, …
- **Ltc:** surgical act, mini incision.
- If the number of occurrences of the word "surgical" is equal to the word "act" then these two words are deleted **Lts**. Otherwise, their occurrences as simple terms reduce by the number of occurrences of the compound word "surgical act ".

b. Algorithm of extraction of compound terms

The algorithm of our approach of extracting compound terms from the documents is as follows:

Algorithm of extraction of compound terms
Inputs : Lts : list of simple terms of the videoconference Seuil_IMA : value of threshold of Adapted Mutual Information Seuilf : threshold of frequency
Outputs: Ltc : list of compound terms of the videoconference
Variables : m : word t : term

tc : compound term
nmtc : number of words in the compound term
nbm : number of words in the list of simple words
nmti : number of words in the term ti
CTF : frequency of a compound term
TF : frequency of a simple term

```
Begin
// Discovering of new compound terms
i←1
While  i<=nbm do
     J ←i+1
   // Adding of new terms
     Tc←ti
   While (the value of theIMA (tc, mj) >seuil_IMA) do
       Concat (tc, «_», mj)
       J ←j+1
   End while
   If  (nmtc>= 2)  then

       If  CTF_ij>seuilf then
           Add  tc to ltc
       End if
       If  tc is a specific word then
           Add tc to ltc.
       End if
   End if
   i←j
end while
End
```

Algorithm 2: **Extraction of compound terms**

c. Weighting compound terms

At this stage, we calculate the weigh to reflect the importance of the term in the document. This weight depends on three factors: the frequency of the compound word in this document, the weights of simple terms that compose it and length of the compound term.

Measurement weighting of compound terms that we have proposed in this manuscript called Compound Term Frequency (CTF). It will be calculated by the function 3.

$$\mathrm{CTFij} = (1 - \frac{1}{\mathrm{long(i)}}) * \mathrm{TFij} + \frac{1}{\mathrm{long(i)}} \sum_{\mathrm{K \in i}} \mathrm{TFk} \qquad \textbf{(3)}$$

i : a compound term

j : a document

k : a simple terme

Long (i) : number of words in the compound term i

To test this function, we apply it to a simple term Ti. We conclude in this case that CTij will be equal to FTi.

$$\mathrm{CTFij} = \left(1 - \frac{1}{1}\right) * \mathrm{FTij} + \frac{1}{1} \sum_{\mathrm{K \in i}} \mathrm{FTij} = 0 * \mathrm{TFij} + \frac{1}{1} \mathrm{FTij} = \mathrm{FTij} = \mathrm{FTi}$$

ISSN 2278-5612

2. INDEXING MODULE

Indexing is a step that consists of analyzing the document while organizing of the documentary fund to produce a set of keywords, also called ***"descriptors"*** background, which the system can easily manage and use in the process of further research.

This module allows extracting the concepts of basic documents (see Figure 3). In SRI a document is considered a medium that conveys information. The result of this indexing module is represented as a list of descriptors. These descriptors ; concepts and semantic relations are extracted by exploring ontology of the domain in order to improve its semantic description. They are plotted, then, using the conceptual graph formalism. This representation is called an index document.

The primary criterion in the extraction of descriptors must always be the potential value of a concept as an element in the expression of document content and in its information search.

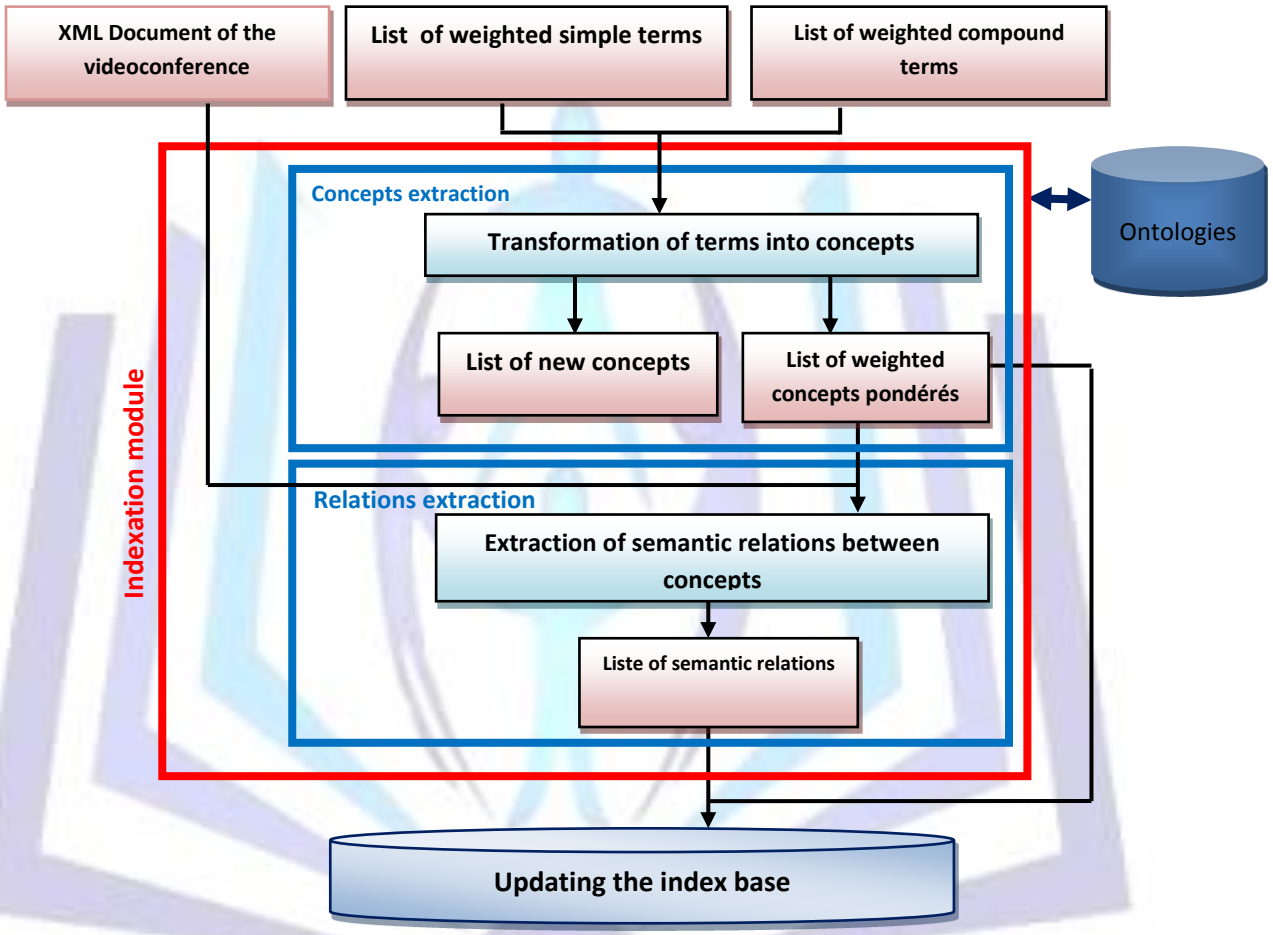

Figure 3: Indexation Module

2.1. Extracting Concepts

During this step, we extract concepts from medical videoconferences. These concepts are denoted in XML documents with simple or compound words. These terms have been taken during the previous steps. To complete the correspondence between the terms and concepts associated with these terms , we use one or more medical ontologies, given an ontology consists of a set of concepts C and a set of relations between these concepts R.

Thus, the structure of an ontology is defined by O = {C, R, $\leq^{c}$ }

Where

- C: all the ontology concepts ;
- R : the set of relations between the concepts of the ontology ;
- $\leq^{c}$: C×C is a partial order on C, it defines the hierarchy of concepts,

– $\leq^{c}(c_1, c_2)$ means that c_1 subsumes c_2 (relationship oriented).

In ontology, each concept is identified by a unique identifier. For each concept, one or more terms have been associated. These terms are called "labels". These are divided into "preferred" label and "alternative" label. Alternative labels are considered synonymous of favorite labels. For example, the concept of "C0001365" in the UMLS has a preferred label and two alternative labels (see Figure 4).

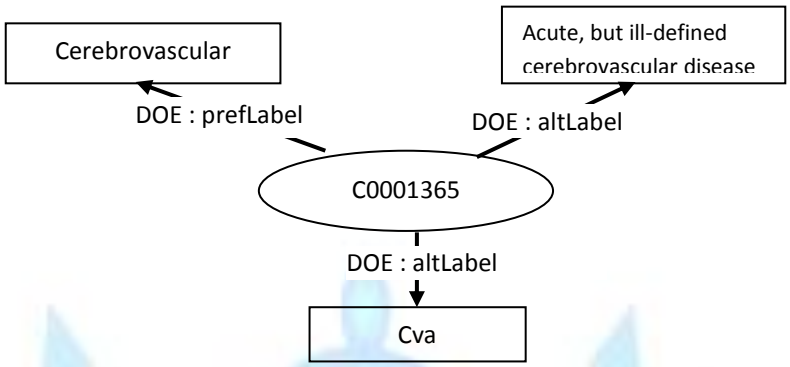

Figure 4: Example of a concept described by DOE

The concept of the figure above is described as follows:

```
<rdf:RDF>
  <owl:Class rdf:ID="C0001365">
      <DOE:prefLabel> Cerebrovascular disease
 </DOE:prefLabel>
      <DOE:altLabel>Acute, but ill-defined cerebrovascular  disease</DOE:altLabel>
      <DOE:altLabel> Cva</DOE:altLabel>
  </owl:Class>
</rdf:RDF>
```

Extract 1: XML Extract of the concept C0001365

In the ontology, a set of terms is used to label the concepts and relations between concepts. This set forms the ontology vocabulary and will be noted: V_O

with $V_O = \{V_{OC}, V_{OR}\}$

Where

V_{OC} : All terms used to denote the ontology concepts;

V_{OR} : All terms used to denote the relations of ontology.

Using the set V_{OC}, we define the operator reference term S_c. This operator is used to determine the / these concept (s) denoted by a term.

An operator reference concept L_c is defined. These two operators are determined by the following equations:

$$\begin{cases} \forall t \in V_{OC}, S_C(t) = \{c \in C \,/\, \text{c is denoted by t}\} \\ \text{and} \\ \forall c \in C, L_C(c) = \{t \in V_{OC} \,/\, \text{t denote c}\} \end{cases} \quad (4)$$

Referring to the same example mentioned before, we get to the concept of "C0001365"

S_c ("Cerebrovascular disease") = {"C0001365"} **and**

S_c ("Acute, but ill-defined cerebrovascular disease") = {"C0001365"} **and**

S_c ("Cva") = {"C0001365"} **and**

L_c ("C0001365") = {"Cerebrovascular disease", "Acute, but ill-defined cerebrovascular disease", "Cva"}

The method which we propose to extract concepts from an XML document of a medical videoconference is to assign to each term of a document the concepts that are related to. To identify concepts related to each term, we use the relation Sc defined beyond.

For semantic indexing based on the concepts, the descriptors that describe the concepts are represented by the terms found in the modeling module. These terms are projected on an external semantic resource to identify the concepts that are associated to them as well as the relationships between them. In our work, we choose three ontologies (ONTOMÉNÉLAS, UMLS, SNOMED-CT). These are judged the most useful among semantic resources for the indexing of audiovisual documents covering the medical field.

In our system, the user chooses the ontology to be used: ONTOMÉNÉLAS for cardiac surgery, SNOMED-CT for clinical terminologies and UMLS in the case of general medicine.

In the case where the ontology does not satisfy the user, this latter can choose another ontology for extracting the associated concepts.

2.1.1. Ambiguity of terms

The problem of ambiguity arises when the terms of the association of words to concepts. There are two types of ambiguity: a linguistic ambiguity and semantic ambiguity.

a. Ambiguous language

This problem is encountered in the case of multilingual documents. It is to find two words belonging to different languages but with the same form in a text. Such ambiguity is not treated in our work because we use monolingual documents.

b. Semantic ambiguity

This case is when we find many concepts denoted by the same term (a term can be the label of several concepts in the ontology). To solve this problem, we seek another concept C 'in relation to the concept C denoted by the ambiguous term t in the ontology. If we find the concept C ', we consider the concept C as a concept denoted by the term t. If not, we use another ontology. In case where we don't not find any concept in relation to the concept C (using the three ontologies), we take all the concepts denoted by the term in question.

2.1.2. Weighting concepts

During this stage, we weight the list of extracted concepts. A concept is represented in a videoconference by one or several words whose frequency of each is already calculated in the extraction of simple terms and compound terms. The Frequency of Concept (CF) is equal to the average of frequencies of the terms that represent it in the document.

$$FCij = \frac{1}{n} * \sum_{i=1}^{n} FTi_j \qquad (5)$$

- FCij : frequency of a concept i in a document j
- n : the number of terms judged as representatives of concept i in a document j
- $\sum_{i=1}^{n} FTi_j$ sum of the frequencies of the terms considered to be representative of the concept i in document j

The weighting of concepts tries to sort them proportionally to their importance in the videoconference document. The most important concept is the one having the highest frequency. We will, then, order the according to their frequencies from the more to less frequent.

ISSN 2278-5612

The algorithm of the method of concepts extraction is as follows:

Algorithme of concepts extraction

Inputs

Lts : list of simple terms

Ltc : list of compound terms

Lco : list of concepts of the ontology

Outputs

Lcp : list of weighted concepts

Variables

Lt : list of terms // ***formed by simple terms and compound terms***

t : a term

c : a concept

DV : document XML of videoconference

Begin

// ***initialisation of list of terms***

Lt ← *Lts* ∪ *Ltc*

// ***identification of concepts associated to non ambiguous terms***

For each term *t* in *DV* **do**

identify the concepts associated to *t* // ***we use Sc***

If *t* is not ambiguous **then**

Add (*DV*, *C*, *weight* (*t*, *DV*)) to *Lcp* // ***c is the identified concept***

End if

End for

// ***identification of associated concepts to ambiguous semantically***

For each term *t* ambiguous in *DV* **do**

Identify the associated concepts with t

Search in *DV* a *C1* denoted by the term *t1* that rise in the same sentence as *t*

If *C1* exist **then**

C ← the set of concepts denoted by *t* that are in relation in the ontology with the concept *C1*

Add (*DV*, *C*, weight (*t*, *DV*)) to *Lcp* // ***c is the identified concept***

else

Add (*DV*, *C*, weight (*t*, *DV*)) to *Lcp* // ***C is the set of concepts associated with t***

End if

End for

End

Algorithm 3: Extraction of concepts

2.2. Extraction of semantic relations between concepts

The importance of taking into account semantic relations lies in the fact that they can considerably improve the efficiency of search for videoconference. Also, indexing using the concepts and the relations between them is much more efficient than using only the concepts **[Harrathi , 09]**.

In order to extract semantic relations between concepts extracted in the previous section, we rely on the used semantic resources, in our case the ontologies. These relations are defined in ontologies by relations types. We admit the hypothesis mentioned in **[6]:** "a relations between two concepts of a document if these two concepts appear in the same sentence, and if the semantic resource defines the semantic relations".

If we take the extract 3 of XML of medical videoconference and using the thesaurus UMLS, we detect the concepts C0334046 and C1302773 denoted by the terms "mild dysplasia" and "squanous low grade intraepithelial lesion".

Applying the hypothesis of Maisonnasse, we find that concepts C0334046 andC1302773 belong to the same sentence, and that these two concepts are connected by the relation "is_finding_of_disease". Using UMLS, we find that this relation is defined as semantics. It is the relations R54390434 of UMLS.

```
<? XML version = “1,0” encoding “iso-8859-1” standalone = “Yes”?;>
  <DOC>
  <ID> 0062782 </ID>
      <Diangnostic> mild dysplasis of squanous epithelium CIN I, LSIL ..  (6278)
      Low grads squanous intraepithelial lesion, coilocyte. </Diagnostic>
      <Description> Atypical cells corresponding to a mild dysplasia. Small air bubbles.
      </Description>
   </Doc>
```

Extract 3: XML Extract of a medical videoconference

Our extraction of semantic relations between concepts algorithm is as follows:

Algorithm of extraction of semantic relations between concepts

Inputs

Lc : list of concepts

Lro : list of semantic relations of the ontology

Lt : list of the terms of the ontology

Outputs

Lrs : list of the semantic relations

Variables

ph : a sentence

Lcph : list of concepts of the sentence *ph*

ci, *cj* : concepts belonging to *Lc*

r : relation

DV : XML document of the videoconference

Begin

For each sentence ph in *DV* **do**

// extract the list of concepts included in the sentence ph Lcph

For each couple of concepts *ci* and *cj* in ph **do**

For each terme *t* in the listof termes *Lt* de *ph* and *t* does not belong to *Lcph* **do**

r ← t

```
        If r belongs to Lro then
          add r to Lrs
        End if
      End for
    End for
  End for
End
```

Algorithm 4: Extraction of semantic relations between concepts

2.3. Updating the index base

With the arrival of a new document, updating the index is performed according to the algorithm 5. The index is updated by linking the document added to the concepts of the ontology and saving the frequency of occurrence of each concept. This frequency will be helpful for sorting documents found during the search module.

Algorithm of updating of the index : Adding of a document

Inputs

D1 : a new document

Outputs

Index : Index up to date

Variables

Ci : concept associated with document

Cj : Concept associated with the index

Ri : Semantic relation

Pi : weight of the concept *Ci*

NbDoc : Number of the documents corresponding to *Ci*

```
Begin
  Change of the state of the document D1 to be added
  For each concept Ci in the document D1 do
    If Ci does not exist in the Index then
      Add Ci
      Connect Ci to the corresponding Cj using the relation Ri
      save the weight Pi of Ci in the Index
      Incrementing of NbDoc corresponding to Ci
    End if
      Incrementing of NbDoc corresponding to Ci
      Save the weight Pi of Ci in the Index
    End if
  End for
End
```

Algorithm 5: Taking into account the adding of new document.

For a deleted document, its status is changed « deleted » before updating the different relative information of the document in the database. Therefore, this document will not be considered by the queries. The corresponding concepts to

this document should be removed if they do not determine other documents. The removal of the document is made by the algorithm 6

Algorithm Updating index: deletion of a document
Inputs *D1* : document to be deleted
Outputs *Index* : updating index
Variables *Ci* : concept associated to the document *Cj* : Concept associated with the index *Ri* : semantic relations *Pi* : weight of concept *Ci* *NbDoc* : Number of documents related to *Ci*

Begin
Change the state of the document D1 to "Deleted" status
/ / the document will be ignored by all queries
For each concept *Ci* of the document *D1***do**
Incrementing of *NbDoc* corresponding to *Ci*
If *NbDoc* = 0 **then**
Delete *Ci*
Delete *Ri* connecting *Ci* to *Cj*
End if
End for
End

Algorithm 6: Taking into account the removal of a new document

During evolution of a document in time, additions, modifications or deletions of information can be made. These events require updating indexes, in order to keep the coherence between the index and the documents of the corpus, because they can cause the contribution of new concepts, concepts removal or modification of frequencies of appearance of concepts.

The addition of new information in a document can cause the following events:

- indexing by other concepts of ontologies;
- indexing by a new concept in the document;
- indexing by new relations between the concepts.

The index is updated by the algorithm 7

Algorithm Updating index: paper editing - adding information
Inputs *D1* : edited document *M1* : modified part of the document
outputs *Index* : Updated index
Variables *Ci* : concept associated with *M1* *Cj* : concept associated with *D1* *Ck* : concept associated with *Index*

Ri : Semantic relation *Pi* : weight of concept *Cj*

Begin
For each concept *Ci* to *M1* **do**
If *Ci* doesn't exist in the list of *Cj* **then**
If *Ci* doesn't exist in *Index* **then**
Add *Ci*
Connect *Ci* to *Cj* or *Ck* by *Ri*
Update *Pi* corresponding to *Ci*
Else
Update *Pi* corresponding to *Ci*
End if
else
increment *Pi*
End if
End for
End

Algorithm 6: Editing a document: adding information

Deleting information in the document may generate the following events:

- removal of the existing relations between concepts presented in the deleted information;
- Deletion of concepts.

The index is updated by the algorithm 8.

Algorithm Updating index: document modification – information deletion
Inputs *D1* : edited document *M1* : Deleted part of the document
Outputs *Index* : Updated index
Variables *Ci* : concept associated with *M1* *Cj* : concept associated with *D1* *Ck* : concept associated with *Index* *Ri* : semantic relation *Pi* : weight of concept *Cj* *NbDoc* : number of documents associated with *Ci*

Begin
For each concept *Ci* to *M1* **do**
Decrement *Pi*
If *Pi* = 0 **then**
Decrement *NbDoc*
If *NbDoc* = 0 **then**
Delete *Ci*
Delete *Ri* connecting *Ci* to *Cj* or *Ck*
End if
End if
End for
End

Algorithm 8: Updating index.

Changing information in an audiovisual sequence means:

- deleting an information;
- adding information.
- This explains the following:
- deleting concepts
- removing relations;
- indexing by another concept of the ontology;
- indexing by a relation between two concepts.

In any case (adding, deleting or updating document corpus), updating the index is carried out only on the concerned document. This reduces indexing time while maintaining the coherence between corpus and the index to find the most relevant documents.

3. EXAMPLE

Our indexing model has been tested on a corpus containing several medical conferences. Among these, we mention videosurgery in Strasbourg.

After analyzing the videoconference, we get the following results:

➢ ***Audio :***

Using FoxTab Video to MP3 software for convert video conferencing "videosurgery in Strasbourg" in an audio file and using the Dragon Naturally Speaking software for transcription of auditory content into text, we get the following:

« The six heart surgery that occured last week by means of a robot which highlighted the fantastic technical progress made in recent years ...»

➢ **Texte :**

We use the approach of Sophie Schupp to analyze the texts contained in the videoconference, we detect the following:

- European Institute Tele Surgery;
- Professor Joel Leroy;
- University Hospital of Strasbourg;
- Professor Jacques Marescaux;
- Pdt IRCAD European Institute Tele Surgery...

➢ **Image :**

We use the software Advanced XVideo Converter for extracting images, we select key frames and we describe them manually.

- Professor Joel Leroy;
- University Hospital of Strasbourg;
- Conference Room contains 160 specialists;
- Operating Room contains surgeons...

➢ **Videoconference :**

By combining information derived from each modality together, eliminating redundant information and correcting mispronunciations, we get:

- European Institute of Tele-Surgery;
- Dr. Jeffrey W. Wilson
- University Hospital of Strasbourg;
- Conference Room contains 160 specialists;
- Operating Room contains Surgeons;
- Professor Joel Lerog;
- « The six heart surgery that occurred last week by means of a robot which highlighted the fantastic technical progress made in recent years ...»

To respectively extract all simple terms, compound terms, we apply the extraction algorithms mentioned in the section « indexing medical content of the videoconferences » in this article.

Frequent simple and compound terms and specific terms of videoconference are implicitly extracted and their union forms the list of terms. This list is used in the extraction of semantic descriptors module. Extraction of concepts and semantic relations will be carried out by the extraction of concepts and semantic relations algorithms mentioned in the same section «indexing medical content of the videoconferences" in this article. The algorithms of extraction of simple and compound terms lead to a correct result (90%), that is to say almost all simple and compound words, whatever the number of words that make up each compound term, are correctly extracted. In the treated corpus, the maximum number of words in a compound term is four for example; we can mention « Institute of Tele surgery of Strasbourg » and « European Institute of Tele surgery ».

We give, in the following, some results and their percentages of test for a selected videoconference from our corpus.

For the extraction of simple terms, they are all extracted and their occurrences are calculated correctly while taking into account the pronouns relating to these terms.

We obtain, by applying the algorithm for extracting compound terms, 40 compounds terms in which there are three terms who are not compound ones. The percentage of successful extraction of compound terms is (40-3) / 40 = 92%. If the pronouns relating thereto are not taken into account, the calculation of frequencies of these terms is 100% correct. The following figure shows the relative XML document of the found compound terms.

compound term.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<racine>
  <exemple>
    <compound term>six heart surgeries</compound term>
    <compound term>last week</compound term>
    <compound term>fantastic technical progress</compound term>
    <compound term>recent years</compound term>
    <compound term>Tele surgery</compound term>
    <compound term>mini incision</compound term>
    <compound term>French technology</compound term>
    <compound term>Institute Research</compound term>
    <compound term>Cancer Strasbourg</compound term>
    <compound term>surgical act</compound term>
    <compound term>intra abdominal colon incision</compound term>
    <compound term>tens millimetres</compound term>
    <compound term>160 specialists</compound term>
    <compound term>laparoscopic surgery</compound term>
    <compound term>conference room</compound term>
    <compound term>160 specialists</compound term>
    <compound term>surgeons eye</compound term>
    <compound term>mini camera</compound term>
    <compound term>patient digestive system</compound term>
    <compound term>surgical act</compound term>
    <compound term>surgical act</compound term>
    <compound term>years 3000 surgeons</compound term>
    <compound term>French technology</compound term>
    <compound term>Professor Leroy Joiil</compound term>
    <compound term>new techniques</compound term>
    <compound term>Institute Tele surgery Strasbourg</compound term>
    <compound term>surgery tables</compound term>
    <compound term>slaughter house</compound term>
    <compound term>colon cancert</compound term>
    <compound term>immediate response</compound term>
    <compound term>video surgery</compound term>
    <compound term>computer surgery</compound term>
    <compound term>third millennium medicine</compound term>
    <compound term>institute Strasbourg</compound term>
    <compound term>laparoscopic surgery</compound term>
    <compound term>conservative surgeons</compound term>
    <compound term>French technology</compound term>
    <compound term>Ina fr</compound term>
    <compound term>University Hospital Strasbourg</compound term>
    <compound term>European Institute Tele surgery</compound term>
    <compound term>operating room</compound term>
    <compound term>Professor Leroy Joiil</compound term>
    <compound term>Doctor Miyake</compound term>
    <compound term>New York</compound term>
    <compound term>Doctor Jeffrey WWilson</compound term>
    <compound term>Professor Jacques Marescaux</compound term>
    <compound term>Pdt IRCAD European institute</compound term>
    <compound term>Tele surgery</compound term>
  </exemple>
</racine>
```

Figure 5: XML Document of the found compound terms

In the indexing module, as examples of the obtained results, the following figure shows the relative XML document of the found concepts

Concepts.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<racine>
  <exemple>
    <concept>Tele surgery</concept>
    <concept>surgical act</concept>
    <concept>160 specialists</concept>
    <concept>laparoscopic surgery</concept>
    <concept>Professor Leroy Joiil</concept>
    <concept>European Institute Tele surgery</concept>
    <concept>surgery</concept>
    <concept>surgeons</concept>
    <concept>act</concept>
    <concept>Doctor Jeffrey WWilson</concept>
    <concept>Doctor Miyake</concept>
    <concept>Professor Jacques Marescaux</concept>
    <concept>intra abdominal colon incision</concept>
    <concept>last week</concept>
    <concept>mini camera</concept>
    <concept>mini incision</concept>
    <concept>operating room</concept>
    <concept>patient digestive system</concept>
    <concept>six heart surgeries</concept>
    <concept>surgeons eye</concept>
    <concept>third millennium medicine</concept>
    <concept>video surgery</concept>
  </exemple>
</racine>
```

Figure 6: XML Document of the found concepts

After the extracting the semantic descriptors, we represent them in a form of a conceptual graph. The concepts are presented within braces « [name_concept] » and relations are shown between parentheses « (name_relation) ». Concepts and semantic relations between them are connected by arrows (→). The figure below illustrates the conceptual graph relative to the videoconference «videosurgery in Strasbourg ».

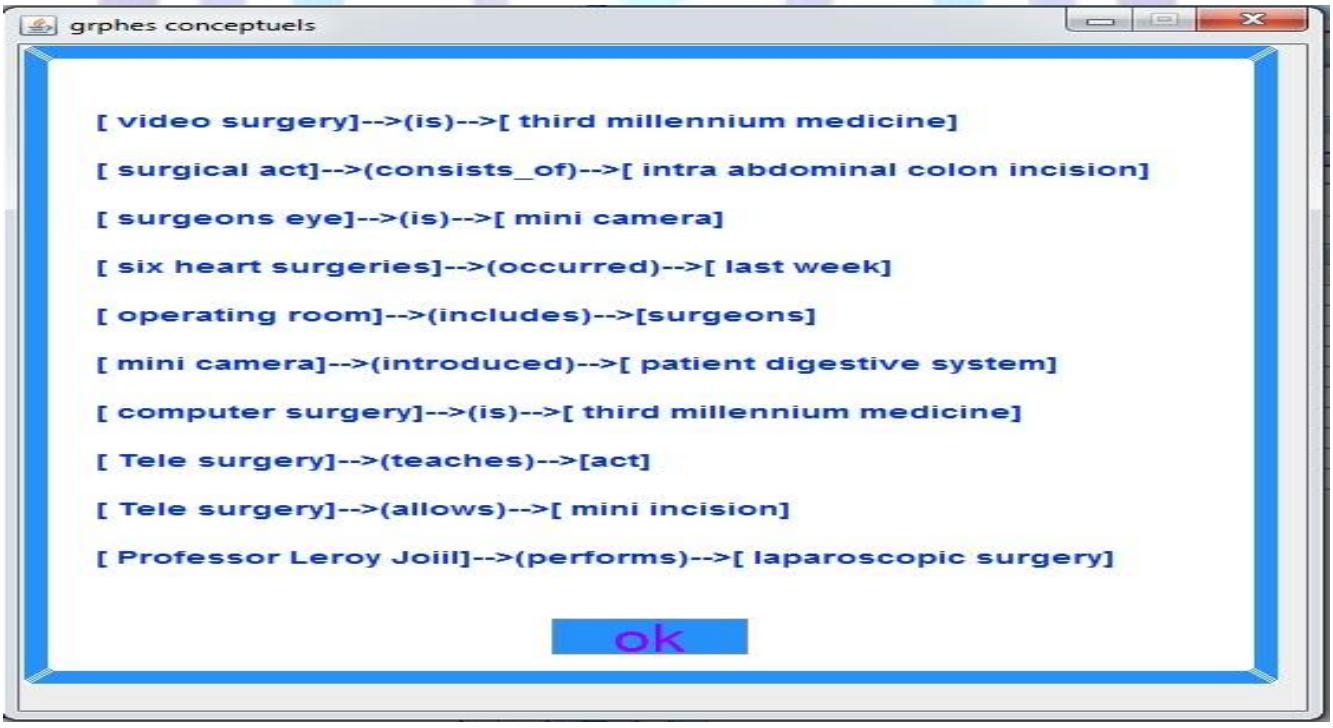

Figure 7: Display interface index of the conceptual graph.

For the research module, we have not yet tested it on our corpus.

4. CONCLUSION

We detailed in this paper two modules of our SRI that is specific to a videoconferences base: analysis and indexing while using one or more ontology (s) of medical field.
The analysis module is specified in two forms: structural and semantic. Its objective is to implement all characteristics of an audiovisual document. In structural modeling, we describe the organization of videoconference that can represent the information content. As for the semantic modeling, the first step, we extract simple terms by eliminating empty words. Then, using the measure of the IMA, we extract the compound terms. The semantically rich and infrequent simple and compound terms are considered as specific terms. The list of simple, compound and specific terms forms the list of terms.
Then we presented the detailed steps that we followed for indexing medical videoconferences. These steps are based on statistical measures and semantic resources such as ontologies and thesaurus. We compare the list of terms to those of existing concepts in resources. The concepts that are found and denoted by the terms will be added to a list of concepts.

Finally, and based on the hypothesis of Maisonnasse, we determine the list of semantic relations between concepts.

REFERENCES

[1] Baea-Yates R., Riberto-Neto B., "Modern Information Retrieval", New-York : ACP Press, Addison-Wesley, 1999.

[2] Baziz M., " Conceptual indexing guided by ontology for information retrieval" , PhD thesis, Institute of Computer Science research of Toulouse, Paul Sabatier University, 2005.

[3] Belkhatir M., " Integration Signal, Symbol for indexing and retrieval of still images ", thesis of University Joseph Fourier, Grenoble I, 2005.

[4] Djamel A., Zighed A., Venturini G., " Knowledge extraction: State and perspectives ", Review of New Information Technologies Edited by Djamel A. Zighed et Gilles Venturini RNTI-E-5 2008.

[5] Harrathi F.," Extraction of concepts and relations between concepts from multilingual documents: Statistical and ontological approach ", thesis of the National Institute of Applied Sciences of Lyon, 2009.

[6] Maisonnasse L., " supports vocabularies for systems-oriented information search accuracy: application to graphs for the desired medical information ". University Joseph Fourier-Grenoble.I. PhD in Computer Science.

[7] Smadja F., "Retrieving collocations from text: Xtract", Journal Computational Linguistics, Volume 19, Number 1, March 1993, pp: 143-177.

[8] Yengui A., Neji M. : "Semantic annotation formalism of video-conferencing documents", International Conference IADIS CELDA 2009 Rome, Italy, 20-22 november 2009