A Semantic Model for Information Retrieval in Documents : an experiment with patient medical records.

Line POULLET, Sylvie CALABRETTO, Jean-Marie PINON.
Laboratoire d'Ingénierie des Systèmes d'Information - INSA de Lyon.
20, avenue Albert Einstein
F-69 621 Villeurbanne Cedex, France

Abstract: Patient medical records contain a great number of information distributed in different kinds of documents: diagnosis, prescription, symptoms observations or radiology analysis, etc. Documents heterogeneity makes specific information retrieval difficult to the medical staff members. This paper shows how a semantic model of documents enables handling information stored in these documents. It enables the definition of a generic semantic structure of a medical record: this structure expresses the implicit content of the each document's element by specifying what kind of information is required. Moreover, it enables to display relevant information depending on the reader.

Keywords: Semantic structuring, meaning representation, structured document, medical record

1. Introduction

Documents are considered to conceptually have three structures :

- a logical structure representing overall organisation of information.
- a layout structure representing presentation of this document.
- and a semantic structure, representing overall organisation of discourse in the document.

Our proposal consists in defining a semantic structure of documents which is bound to logical and layout structures defined by international standards such as ODA (Open Document Architecture) [ISO 8613] or SGML (Standard Generalized Markup Language) [ISO 8879]. This structure expresses the meaning of each document element in a formal way.

This paper shows how the semantics of documents can be expressed in the semantic structure. Document understanding tools are composed of the three model components: the semantic structure, description of semantic elements and a knowledge base for domain description.

The first part of this paper gives an overview of the model. The second part shows how semantic structuring of documents can be efficiently defined using SGML syntax. Using this documents structuring norm, two levels of description may be defined: generic semantic structure (vs. DTD) and specific semantic structure (vs. instanciated document) in order to define an « abstract

interface » of the information stored in documents. The medical patient record is presented as a relevant application for handling semantic structured documents.

2. Model description

The semantic model relies on meaning representation of information units (i.e. the logical units). This meaning representation is distributed in the overall architecture model : the model binds a semantic structure, a logical structure of documents (i.e. a SGML DTD) and a domain model.

The semantic structure contains two levels of description :

1. meaning representation of information units. The Conceptual Graphs formalism is used to represent semantics of document elements ;
2. and document rhetorical organization.

The domain model contains a medical ontology of concepts and relations between concepts, used for guiding the meaning representation of information units.

Generic semantic structure defines the generic organization of documents content for a specific class of patient medical record. Each element of this structure defines the a priori semantic content of the associated logical element(s).

For example, a record for a cardiologist may contain a general description of the patient (civil status, age, ...), a few ECGs with an analysis and some treatments. The generic semantic element *Description of patient* (corresponding to the logical elements : *Civil Status*, *General Information*) specifies the concepts (patient, address, age, sex, ...) and the relations between them (patient is a person who has an address, some medical antecedents, etc.). When instanciating the generic semantic structure (for creating the document), the semantic elements are instanciated: Mr. B. is the patient, he had a coronary thrombosis, etc.

3. Semantic structuring

3.1. Generic semantic structure

A generic semantic structure is defined as a DTD (Document Type Definition) in SGML (Standard Generalized Markup Language) syntax. SGML users have two levels of representation of documents: the generic logical structure, called DTD (Document Type Definition), describes the organisation of a class of documents, in opposite to the specific logical structure which describes an existing document organisation. In the same way, we can define a generic semantic structure as a semantic DTD (see §2) in order to define specific semantic structure of existing patient records. According to semantic representation, we have however define a conceptual semantic structure as a

kind of meta-DTD. It provides the semantic syntax used for defining a semantic DTD for patient records for example.

```
<!DOCTYPE   SemStructure    (
<!ENTITY    Case            « SemanticCase »                                    >
<!ELEMENT   SemStructure    --      (SemElement)+                               >
<!ELEMENT   SemElement      - 0     (Expression |SemElement)+                   >
<!ELEMENT   Expression      - 0     (Verb, QualifiedAttributes+)                >
<!ATTLIST   Expression      id      ID                      #IMPLIED            >
<!ATTLIST   Expression      logEnt  IDREF                   #IMPLIED            >
<!ATTLIST   Expression      Type    #IMPLIED                                    >
<!ELEMENT   Verb            - 0     #PCDATA                                     >
<!ELEMENT   QualifiedAttributes - 0 Concept                                     >
<!ATTLIST   QualifiedAttributes case ENTITY                 #FIXED              >
<!ELEMENT   Concept         - 0     (ConcreteConcept |Expression)               >
<!ELEMENT   ConcreteConcept - 0     #PCDATA                                     >]>
```

Figure 1: The meta semantic structure.

This abstract semantic structure implements the definition of meaning seen in §2. This abstract structure shows how the meaning definition we propose can be implemented in a SGML environment. This grammar defines a semantic structure (SemStructure) as composed of semantic elements (SemElements), i.e. at least (+ occurrence indicator) one semantic element. Semantic elements are Expression, i.e. semantic representation of logical element(s) or are composed of other semantic elements (composite semantic elements).

Note that the element Expression has three attributes :

- the id attribute enabling this element to be referred, for example in a logical element.
- the LogEnt attribute enabling this element to refer to a logical element (the converse link)
- the Type attribute enabling this element to have a rhetorical type.

The links between a logical element and the semantic expression representing it are described by the ID/IDREF mechanism, allowing to refer to a SGML element.

By using this grammar, PatientRecord generic structure can be more precisely described as follows:

```
<!DOCTYPE   Patient Record   (
<!ENTITY    Case.1           « SemanticCase1 »                                    >
<!ENTITY    Case.2           « SemanticCase2 »                                    >
<!ENTITY    Case.3           « SemanticCase3 »                                    >
<!ENTITY    ...                                                                   >
<!ELEMENT   PatientRecord    - 0    (PatientDescription, Observations* &
                                    Prescriptions*)                              >
<!ELEMENT   PatientDescription - 0  (CivilStatus, Antecedents*)                  >
<!ELEMENT   CivilStatus      - 0  ...                                             >
<!ELEMENT   Antecedents      - 0    (Observations+)                               >
<!ELEMENT   Observations     - 0  ...                                             >
<!ELEMENT   Prescriptions    - 0    (Medicine |Tests |Medicine & Tests)           >
<!ELEMENT   Medicine         - 0    (Verb.1, QualifiedAttributes1.1,
                                    QualifiedAttributes1.2, QualifiedAttributes1.3) >
<!ATTLIST   Medicine         id     ID          #IMPLIED                          >
<!ATTLIST   Medicine         logEnt IDREF       #IMPLIED                          >
<!ATTLIST   Medicine         Type = « description »                               >
<!ELEMENT   Verb.1           - 0    #PCDATA                                       >
<!ELEMENT   QualifiedAttributes1.1   - 0    Concept.1                            >
<!ATTLIST   QualifiedAttributes1.1   case=&Case.1   ENTITY   #FIXED               >
<!ELEMENT   QualifiedAttributes1.2   - 0    Concept.2                            >
<!ELEMENT   QualifiedAttributes1.3   - 0    Concept.3                            >
<!ELEMENT   Concept.1        - 0    ConcreteConcept                               >
<!ELEMENT   Concept.2        - 0    ConcreteConcept                               >
<!ELEMENT   Concept.3        - 0    ConcreteConcept                               >
<!ELEMENT   ConcreteConcept  - 0    #PCDATA                                       >
<!ELEMENT   Tests            - 0    #PCDATA                                       >
<!...                                                                             >]>
```

Figure 2: A semantic structure for Patient record - Description of semantic expression Medicine.

In the previous DTD, the only semantic element Medicine has been entirely described.

3.2. Specific semantic structure

Generic semantic elements enable a definition of an *a priori* meaning content by restricting concepts. This mechanism specifies what kind of concrete concepts and what kind of relations are concerned with the semantic entities. The *a priori* semantic content of a kind of document is then clarified. For example, in the semantic structure of documents Patient Record defined in §3.1, the semantic element Medicine can be defined as follows according to SGML syntax.

Generic element Medicine points out that the logical element(s) it is connected with express a relationship « be-prescribed » between a Drug, a Dosage and a Patient. Generic element expresses the a priori meaning content of logical element(s). Specific semantic element Medicine for Mr. B. prescription can be written as follows in his medical record. The specific element represents the meaning of the specific logical element(s) it is connected with.

```
<!DOCTYPE   PatientRecord   (
<!ENTITY    Case.1          « Matter »                                  >
<!ENTITY    Case.2          « State »                                   >
<!ENTITY    Case.3          « Recipient »                               >
<!ENTITY    ...                                                         >
<!ELEMENT   PatientRecord   - 0     (PatientDescription, Observations* &
                            Prescriptions*)                             >
<! ELEMENT  ...                                                         >
<! ELEMENT  Medicine        - 0     (Verb.1, QualifiedAttributes1.1,
                            QualifiedAttributes1.2, QualifiedAttributes1.3)  >
<!ATTLIST   Medicine        id      ID          #IMPLIED                >
<!ATTLIST   Medicine        logEnt  IDREF       #IMPLIED                >
<!ATTLIST   Medicine        Type = « description »                      >
<!ELEMENT   Verb.1          - 0     « be-prescribed »                   >
<!ELEMENT   QualifiedAttributes1.1      - 0     Concept.1               >
<!ATTLIST   QualifiedAttributes1.1      case=&Case.1    ENTITY  #FIXED  >
<!ELEMENT   QualifiedAttributes1.2      - 0     Concept.2               >
<!ELEMENT   QualifiedAttributes1.3      - 0     Concept.3               >
<!ELEMENT   Concept.1       - 0     Drug                                >
<!ELEMENT   Concept.2       - 0     Dosage                              >
<!ELEMENT   Concept.3       - 0     Patient                             >
<!ELEMENT   Drug            - 0     #PCDATA                             >
<!ELEMENT   Dosage          - 0     #PCDATA                             >
<!ELEMENT   Patient         - 0     #PCDATA                             >
<! ...                                                                  >)>
```

Figure 3: Semantic structure with constrained attributes.

```
<!DOCTYPE PatientRecord...
<PatientRecord>
<Medicine id = §§ LogEnt = ### Type = « Description »>
<Verb>Be-prescribed</Verb>
<QualifiedAttribute1.1 case= &Case1><Drug>my_medicine</Drug></QualifiedAttribute1.1>
<QualifiedAttribute1.2 case= &Case2><Dosage>my_dosage</Dosage></QualifiedAttribute1.2>
<QualifiedAttribute1.3 case= &Case2><Patient>Mr. B.</Patient></QualifiedAttribute1.3>
</Medicine>
<       ...
</PatientRecord>
```

Figure 4: Extract of specific semantic structure PatientRecord.

Conclusion

In this paper, we have proposed a formal model to assist users in the retrieval of information stored in documents. This model relies on definition of semantic units, describing the meaning of information units. Documents have three bound structures: layout, logical and semantic structures. This model provides a way of defining a sort of implicit document, with an *a priori* clarified content.

Semantic elements can be used as *semantic structured index* related to the documentary data:

- First, the main index level is the rhetorical type of semantic element. This index defines the kind of required discourse. Prescription of medicine as a description (medical prescription) appreciably differs from a prescription of medicine as an assertion (in the notice of use, for example).
- The second index is the verb which expresses the conceptual relations between the connected concepts: « Be-prescribed » semantically differs from « Be-forbidden » in the same type of semantic element.
- The secondary index are the semantic cases.

This mechanism enables indexing of parts of documents by following the « meaning representation » links between a semantic element and a logical element(s).

We anticipate that such a semantic structure provides tools for ***information retrieval*** in documents because of the formal representation of information. Statistical techniques [Lewis 96] have already been used for semantic retrieval mechanisms. Our approach allows for the extension of definition of indexing and information retrieval, by taking into account formal semantic representations of text. It relies on the integration of two paradigms for representing the same information : structured documents on the one hand and knowledge base on the other hand [Maret 96].

References

[Brachman 85] **BRACHMAN R. & SCHMOLZE J.** An Overview of the KL-ONE Knowledge Representation System. *Cognitive Science.* N° 9, 1985. p. 171-216.

[Dobson 95] **DOBSON, SA. & BURRILL, VA.** Lightweight Databases. In : *proceedings of the 3th International World Wide Web Conference, Darmstadt, avril 1995.* p. 282-288.

[Iso 86 13] International Standard. *Information Processing - Text and Office Systems - Open Document Architecture (ODA) and Interchange Format (ODIF),* 1988.

[Iso 88 79] International Standard. *Information processing - Text and Office Systems - Standard Generalized Markup Language (SGML),* 1986.

[Lewis 96] **LEWIS, D. & SPARK JONES, K.** Natural Language Processing for Information Retrieval. *Communication of the ACM,* 1996, Vol. 39, N° 1, p. 92-101.

[Maret 96] **MARET, P., POULLET, L. & PINON J.-M.** Des modèles pour capitaliser la connaissance au sein des organisations. *ISI (Ingénierie des Systèmes d'Information).* 40 Pages. *To be published.*

[Nanard 88] **NANARD, J. & Al.** Conceptual documents : a mechanism for specifying active views in Hypertext. In: *Proc.ACM Conf. on Document Processing Systems, ACM Press, Santa Fe, 1988.*

[Nanard 93] **NANARD, J. & NANARD, M.** Should Anchors be typed too ? An experiment with MacWeb. In : *proceedings of the Fifth ACM Conference on Hypertext, Seattle, novembre 1993.* p. 51-62.

[Pinon 94] **PINON, J.-M., RICHEZ, M.-A. & FLORY, A.** Support System for Cooperative Edition of Multimedia Documents. In: *Proceedings of OOIS'94, Londres, 19-21 décembre 1994.* Londres: Springer-Verlag, 1994. p. 416-421.