A Computational Model of Information Retrieval with UMLS

Jean-Jacques Robert, M.S., Michel Joubert, Ph.D., Laurent Nal, M.S., Marius Fieschi, M.D., Ph.D.
CERTIM. Faculté de Médecine.
Boulevard Pierre Dramard. F-13326 Marseille Cedex 15. France

A high level representation of data would clarify the complex collection of medical concepts, terms and relationships derived from standard classifications that the Unified Medical Language System contains. A conceptual model is described which represents the data structure. A second objective of this conceptual model is to provide users with the capability to build queries to information databases as easily as possible on the basis of this data structure. The methods used to build this model are semantic networks and conceptual graphs. The object-oriented computational model which implements this conceptual model is detailed. It reuses part of the generic C++ classes of the National Institutes of Health library. New classes are added to this library to implement the needed functionalities.

1. INTRODUCTION

The Unified Medical Language System (UMLS) is a major project of the U.S. National Library of Medicine [8]. Three main components constitute the UMLS data structure: the so-called Metathesaurus (Meta-1) [13], the Semantic Network [9] and the Information Sources Map [10]. The UMLS is a complex collection of medical concepts, terms and relationships derived from standard classifications [5]. To optimally benefit from the rich resources the UMLS provides, researchers and developers must understand its structure and data relationships. Four successive versions have been delivered to date. Among them, the third version of the Metathesaurus [16] presented a simpler schema than the previous versions. Nevertheless, as noticed by authors, a high level representation of data would clarify the Metathesaurus and Semantic Network contents and potential applications [15].

The fixed objective is to help users in formulating queries to information databases as naturally as possible. To achieve this goal, it is necessary to give them the ability to operate the closeness link between their view of concepts they use and the representation of these concepts in computer systems. The principle we adopt is to build a logical and conceptual interface between users' applications and information databases. In a previous work, we proposed to build a conceptual model for information retrieval with UMLS [6]. This model, based on conceptual graphs [12] and semantic networks [14], provides the ability to express declarative knowledge relevant to medical concepts and the relationships which link them to each other. The exploitation of the conceptual model results in a query translated by a graph which is further matched to data graphs derived from information databases. This model, we will present briefly in section 2, is conceived to assist end-users to query large information databases. It has been used experimentally to query patients records [7] and is currently being experimented with documents databases. The object-oriented computational model which implements this conceptual model is presented in section 3.

2. THE CONCEPTUAL MODEL

In the following we will only be concerned with the Metathesaurus and the Semantic Network of UMLS. The core concepts which have been isolated in Meta-1 are connected to generic types of concepts in the Semantic Network. These types are interconnected by semantic relationships. The data structure of Meta-1 is based on hierarchies and associations. The association relationship links a given term to related terms and to a preferred term. The hierarchy structures the preferred terms into more generic terms and more specific ones. This hierarchical relationship divides Meta-1 into several so-called microthesauri, according to a local specificity of concepts. The presence of these microthesauri translates the various contexts in which a medical concept can be viewed.

A first objective of the conceptual model is to represent this data structure. We organise the above elements according to the diagram of figure 1 conceived for semantic networks exploitation [11]. When applied to the UMLS components, this structure shows the three following levels:

- the core concepts identified in Meta-1 constitute the lowest level,
- the upper level is the ontology of types of concepts which constitutes the basis of the UMLS Semantic Network,
- the intermediary level is constituted by views which represent microthesauri in Meta-1.

0195-4210/94/$5.00 © 1994 AMIA, Inc.

A given concept may have various instances in different views. For instance, an aneurysm can be either viewed as an acquired abnormality -thus, connected to this related type of concept- or as a congenital abnormality -thus, connected to this latter type. Congenital abnormality and acquired abnormality are two distinct types in the UMLS Semantic Network. Semantic relationships, which are not represented in figure 1, are set between types of concepts to translate the UMLS Semantic Network totally.

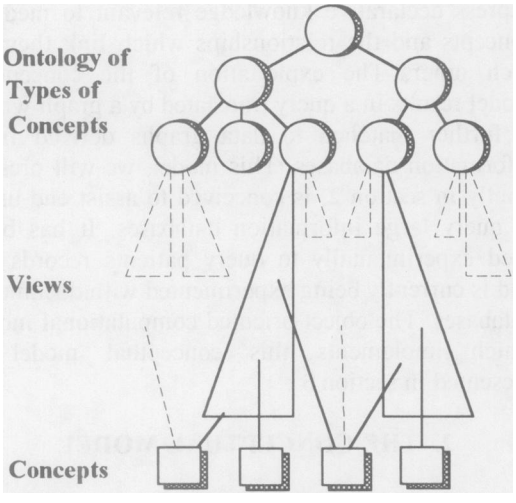

Figure 1: Organisation of concepts in views connected to the ontology of types

A second objective of the conceptual model is to provide users with the capability to build conceptual graphs which involve instances of concepts in views interconnected by instances of semantic relationships inherited from the Semantic Network. These graphs are built at the intermediate level of figure 1. Since the semantic relationships set between types are inherited neither by the concepts nor by their instances automatically, the Semantic Network must be considered as a frame knowledge which describe valid associations and not as an expert knowledge base which should describe relevant associations. For instance, a coronary aneurysm cannot be diagnosed by every diagnostic procedure, but, when viewed as a cardiovascular disease, it can be diagnosed by some (but, not every) cardiovascular diagnostic procedure, such as an angiography. Valid conceptual graphs are built under the control of the Semantic Network. The end-users determine the medical relevance of these graphs. The figure 2 shows the building of an elementary conceptual graph. If the type A (e.g., Diagnostic Procedure) is linked to the type B (e.g., Disease or Syndrome) by the means of the semantic relationship R_1 (e.g., diagnoses), then instances of concepts a (e.g., Angiography) and b (e.g., Coronary Arteriosclerosis) in their respective views V_1 (e.g., Cardiovascular Diagnosis) and V_2 (e.g., Cardiovascular Diseases) may be linked by an instance of the above relationship R_1. Valid operations on conceptual graphs - such as join, restriction and projection - allow the users to build complex graphs which translate full sentences such as "treatment by angioplasty of coronary diseases diagnosed by angiography" [7,12]. This is illustrated in figure 2 by a join of the above graph with a second elementary graph on the same b linked in this latter by R_2 (e.g., treated by) with c (e.g., Angioplasty), issued from C (Therapeutic Procedure), viewed in V_3 as a Cardiovascular Therapeutic Procedure.

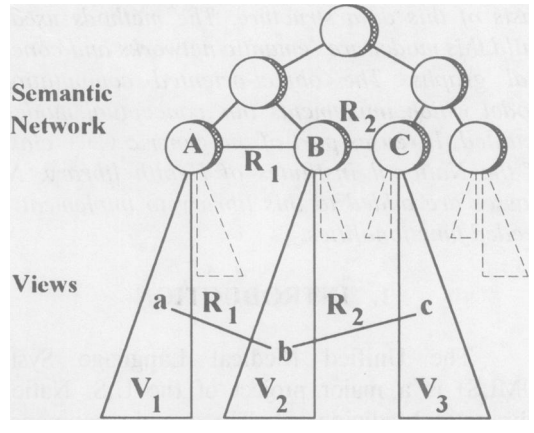

Figure 2: Building a conceptual graph from the Semantic Network

A typical exploitation of the model when building a conceptual graph is as follows:

- select a concept
- select a view on this concept: the related type is then selected automatically,
- select a relationship involving this type: the destination type is selected automatically,
- select a view connected to this latter type,
- select an instance of a concept in this view.

This sequence of operations builds an elementary conceptual graph which links two nodes, the two selected instances of concepts, by the means of an edge, the instance of the selected semantic relationship. Iterations of such a sequence allow the building of more complex connected graphs (each node is linked to another node, at least) which involve several instances of concepts linked by various relationships.

Records in an information database are patients records or bibliographical references, for instance. Fields are identified in information

records which contain codes or key-words which index the records. A conceptual graph represents a user's focus or interest in an information database, i.e. a query to this database which translates a full sentence as exemplified above. Such a graph expresses a generic query which is satisfied by coded or indexed data in an information database which

- either matches the nodes exactly,
- or is the code of a descendant of a node in the relevant view.

For instance, if a patient record contains the coded data which expresses that (s)he has been treated by a transluminal angioplasty for a coronary arteriosclerosis diagnosed by a coronary angiography, the previous query is satisfied, since transluminal angioplasty, coronary arteriosclerosis and coronary angiography are descendants of angioplasty, coronary diseases and angiography in their respective views.

The conceptual model unifies the representation of the Semantic Network and the conceptual graphs by the means of (weakly) connected graphs [4]. In this way, conceptual graphs appear as specializations of the graphs at the semantic level, since they have the same structure and the former graphs are obtained from the latter ones in replacing types nodes by concepts instances nodes.

3. THE COMPUTATIONAL MODEL

The objects involved in the conceptual model are complex. They are concepts, instances of concepts, types of concepts, semantic relationships, graphs, and so on. Powerful tools are needed to design and implement them. Object-orientation allows capabilities to design computer programs [2] which store and treat complex data [1]. Object-oriented programming languages provide the facility to define hierarchies of classes of objects in which the subclasses inherit the properties of their ancestor classes automatically. Objects have the properties of the classes of which they are instances. So, the classes constitute a model for the instantiated objects. The current success of object-programming languages leads developers to create powerful libraries of generic classes of objects. The National Institutes of Health (NIH) Library of Classes [3], implemented in C++ language, provides a large set of hierarchically structured classes. This library contains generic classes of objects such as "list", "ordered list", "set", "dictionary", and so on, which can be reused with benefit. The NIH library of classes is partly represented in figure 3.

A graph is designed as a set of 3-uples: two nodes and the directed relationship which links them. In our case, we needed to implement classes such as:

- "concept", instantiated by concepts or types of concepts,
- "relationship", instantiated by semantic relationships both in conceptual graphs and in semantic networks,
- "tree", instantiated both by the ontology of types and by concepts instances structured in views,
- "graph unit", instantiated by graphs nodes designed as 3-uples,
- "graph", instantiated both by semantic networks and by a conceptual graphs,

and their related procedures among which are the valid operations on graphs. So, we specialized the hierarchy of classes of the NIH library in such way that these new classes benefit from the already defined properties. For instance, since a procedure is defined in the class "set" which verifies that an instantiated set does not contain duplicated elements, to make the class "graph" a subclass of the class "set" entails that the former class inherits this property. Figure 3 presents the hierarchy of classes we implemented. It shows a part of the NIH library we reused and how we implemented new classes as subclasses in it. We can remark that the classes "concept" and "relationship" are new classes directly attached as subclasses of the root class "object", when "node", "tree" and "graph" are attached to classes in the hierarchy and thus benefit from the inheritance.

Object - Root of the NIH library of classes
  Concept - Unique representation of a concept
  Relationship - Semantic relationship
  Collection - Abstract class for collections
    ArrayOb - Array of objects references
      GraphUnit - 3-uple graph unit
    Set - Collection of non duplicated objects
      Dictionary - Set of associations
        IdentDict - Dictionary
      Graph - Set of graph units
      Tree - Tree of objects references
  Iterator - Collection iterator
  Link - Abstract class of links for linked lists
    LinkOb - Link with an object reference
    LookUpKey - Dictionaries associations
      Assoc - Association of references

Figure 3: Part of the computational hierarchy of classes. Bold names refer to the classes added into the NIH library

The first time a concept is referenced, an object which instanciates it is created in memory. The selected view of this concept is also instantiated by the means of tree-structured references to the related instantiated concepts. When a relationship links two instantiated concepts, it is also instantiated. A graph unit is created which references the two concepts and the relationship instances. According to this structure, when a graph is created, it is instantiated as a set of instances of units. In this way, no object is duplicated in memory. Duplication is avoided by the use of references to objects identified internally by the object-oriented mechanism. Such a computational instanciation process is summarised by figure 4 which details instances of the concerned objects with an example. It shows the instanciation of the graph which translates the previous query "treatment by angioplasty of coronary diseases diagnosed by angiography".

Users queries are instantiated by the means of graphs after a navigation in UMLS components as described previously. When an information database is queried, data graphs issued from individual records are also instantiated according to the data model the Semantic Network provides. The query graph is successively compared to the individual data graphs by the means of a matching process which implements an operation on graphs named "projection" [12]. This process verifies that the query, or a specialization of it, as described above, is a subset of the individual data graph.

The software we developed in C++ implements the above computational model. The UMLS data (the Semantic Network and Meta-1 components), excluding the chemical data, have been imported from the relational format delivered by the NLM [16] into the Oracle relational database system which allows easy communications with C and C++ programs.

4. DISCUSSION

Efficiency in information retrieval requires all the most integrated system as the information is complex in form and type. The main problem the users encounter is to express queries when the semantics of the domain does not explicitly appear in the information database. This problem is increased by the diversity of the methods to

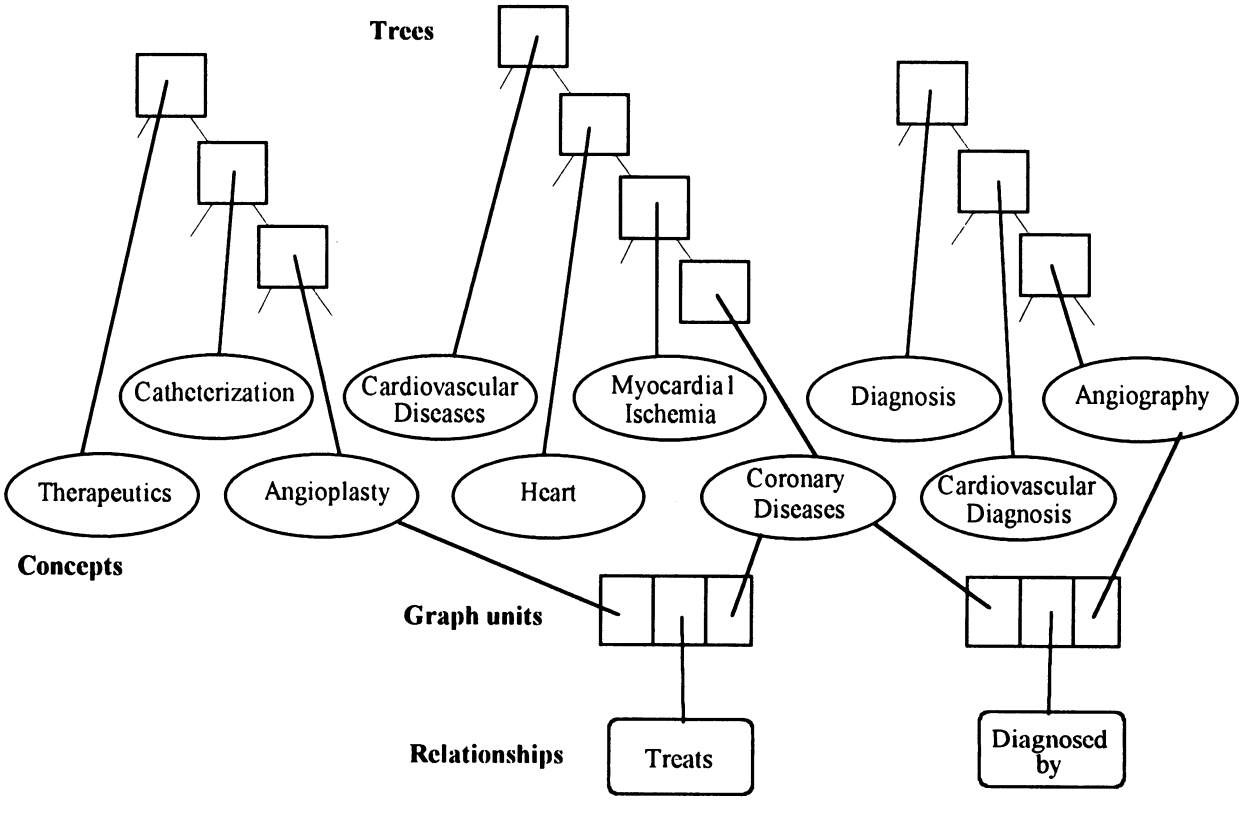

Figure 4: An example of instantiated graph

implement these systems where representation and storage are commonly proprietary even if they offer a high level of query language. We did have an approach, based on the current results of the UMLS project, which mixes the usage of both key-words coming from a standardized vocabulary (Meta-1) and core concepts isolated from it. The Semantic Network of UMLS structures, by the means of semantic relationships, the types of concepts to which the core concepts of Meta-1 are connected. The fixed goal is to help users in formulating queries to information servers as easy as possible. Thus, the semantic approach is significant as well because, using knowledge of the domain, it is able to represent part of the meaning though associations between the concerned concepts.

After we have represented the conceptual model by the means of conceptual graphs and operations on these graphs, it was necessary to select powerful tools to implement it. Object-orientation is a suitable approach to design such systems and permitted us to define the generic classes of the computational model to implement. The C++ language and the library of classes developed by the NIH are efficient supports for implementation. This kind of implementation is general enough to be done with other object-oriented languages and databases. Promising experiments have been done with patients records databases. Complementary experiments with documents databases are underway currently and should conclude to the efficiency of this conceptual approach and the suitability of the developed software. We hope that our object-oriented conceptual and computational modelling of UMLS components will be helpful for designers concerned with information science, and especially for other UMLS users and developers.

ACKNOWLEDGEMENTS

This work has been partially founded by the French Ministry of Education. We thank the U.S. National Library of Medicine which graciously furnished us the UMLS data.

REFERENCES

[1] R.G.G. Cattell. Object Data Management: object-oriented and extended relational database systems. Addisson Wesley; 1991.

[2] P. Coad, E. Yourdon. Object-Oriented Design. Prentice-Hall, 1991.

[3] K. Gorlen, S. Orlow, P. Plexico. Data Abstraction and Object-Oriented Programming in C++. John Wiley & Sons, 1991.

[4] F. Harary. Graph Theory. Addison Wesley, 1971.

[5] B.L. Humphreys, D.A.B. Lindberg. Building the Unified Medical Language System. Proc. 13rd SCAMC. L.C. Kingland editor. IEEE Computer Society Press, 1989: 475-480.

[6] M. Joubert, M. Fieschi, J-J. Robert. A Conceptual Model for Information Retrieval with UMLS. Proc. 17th SCAMC. C. Safran editor. McGraw-Hill, 1993: 715-719.

[7] M. Joubert, M. Fieschi, J-J. Robert, A.G. Tafazzoli. Users Conceptual Views on Medical Information Databases. Int. J. Biomed. Comput. In press.

[8] D.A.B. Lindberg, B.L. Humphreys, A.T. McRay. The Unified Medical Language System. Methods of Information in Medicine 32, 1993: 281-291.

[9] A.T. McRay. The UMLS Semantic Network. Proc. 13rd SCAMC. L.C. Kingland editor. IEEE Computer Society Press, 1989: 503-507.

[10] D.R. Masys, B.L. Humphreys. Structure and Function of the UMLS Information Sources Map. Proc. MEDINFO 92. K.C. Lun, P. Degoulet, T.E. Piemme, O. Rienhoff editors. North-Holland, 1992: 1528-1521.

[11] L. Shastri. Why Semantic Networks? In: Principles of Semantic Networks: Exploration in the representation of knowledge. J.F. Sowa editor. Morgan Kaufmann, 1991: 109-136.

[12] J.F. Sowa. Conceptual Structures: information processing in mind and machine. Addison Wesley, 1984.

[13] M.S. Tuttle, D.D. Sheretz, M.S. Erlbaum, N.E. Olson, S.J. Nelson. Implemeting Meta-1: the first version of the UMLS metathesaurus. Proc. 13rd SCAMC. L.C. Kingland editor. IEEE Computer Society Press, 1989: 483-487.

[14] W.A. Woods. Understanding Subsumption and Taxonomy: a framework for progress. In: Principles of Semantic Networks: exploration in the representation of knowledge. J.F. Sowa editor. Morgan Kaufmann, 1991: 45-94.

[15] Y. Yang, C.G. Chute. A Schematic Analysis of the Unified Medical Language System. Proc. 16th SCAMC. M.E. Frisse editor. McGraw-Hill, 1992: 204-208.

[16] UMLS Knowledge Sources, 4th Experimental Edition. National Library of Medicine, 1993.