Modeling Hypermedia with Conceptual Graphs

Ammar KHEIRBEK, Yves CHIARAMELLA
LGI-IMAG BP 53
38041, Grenoble cedex 9 - FRANCE
Tel: (33) 76 51 46 00 ext 3356, Fax: (33) 76 44 66 75
Ammar.Kheirbek@imag.fr

Abstract
Hypermedia is devised to provide the final user several pre-defined paths (links) to access information that is split into several units of storage (nodes). Creation and organization of a graph of nodes/links are critical: manual processing of multimedia documents is expensive and impractical, hard link models are not flexible, and updating methods of nodes/links are not efficient. An automated assistant process for the construction of the hypermedia network is then needed. This automated process allows essentially to retrieve relevant links between documents. Some studies ([1], [2]) were developed to provide automatic assistance techniques for the processing and the creation of links. But, these works don't use any knowledge representation. We believe that an elaborated model of knowledge representation is indispensable to represent semantic and structural information of documents. From this kind of associated information to documents, the automatic assistance will be intelligent and provide effective solutions to hypermedia problems. Another source of problems is the notion of browsing: users use links to browse from a node to another until they match their needs. The final user then needs to know where he is in the network, to know where to go and how to go ([3]). Solutions were developed to solve these problems like the use of historic, annotations or assistant semantic information. An interesting approach is to integrate IR functions into hypermedia. This integration of browsing and IR queries will give efficient techniques to classical hypermedia systems, and will give also a very powerful interface to classical IR systems that suffer in general from the quality of their interfaces. At any moment the user can formulate queries (of IR type) or continue his browsing using hypermedia links. In this paper we show how these two approaches may be combined in a more efficient way within the framework of a global model based on conceptual graphs (CGs) [7] theory for structures and knowledge. A final problem with hypermedia is modeling. With the increasing number of hypermedia systems, we feel always the need of a general formalization of such systems. More research was focused on this point ([4], [5]), but we are always far from a general universal formalism because researchers haven't decided yet which are the common organizing principles and fundamental concepts of such systems.
In [6] we have presented a first attempt for modeling of hypermedia system based on CGs [7]. In this paper we present a more elaborated model that integrates hypermedia navigation with IR querying and an elaborated knwoledge representation. We believe that IR and hypermedia are two complementary domains that must be combined to provide the user a more efficient and effective environment for accesing and retrieving information.

Keywords: Hypermedia, Link construction, Structured documents, Knowledge Representation, Information Retrieval, Conceptual Graphs.

1. Hypermedia model

1.1. Introduction
The hypermedia model is classically organized in two levels: the hyperindex and the hyperbase. Our approach about this architecture is to consider that both components may be viewed as different kinds of knowledge that have to be made explicit (ie. accessible to the user) either while browsing or when he is querying the system. As will be seen later, this integrated view of a hypermedia system is very powerful in the sense that it helps for a proper integration of browsing/querying approaches, while it also allows the user to issue much more classical queries that combine content and structure as parts of the needed information. Following our first approach presented in [6] we suggest to use CGs as an unifying formalism for both knowledge and structures. For more details about the theory of CGs see [7] and [6].

1.2. The model

We define a hypermedia model as a 4-tuple: $\mathcal{H}(\mathcal{T}, \mathcal{M}, \mathcal{R}, \mathcal{I})$ where:

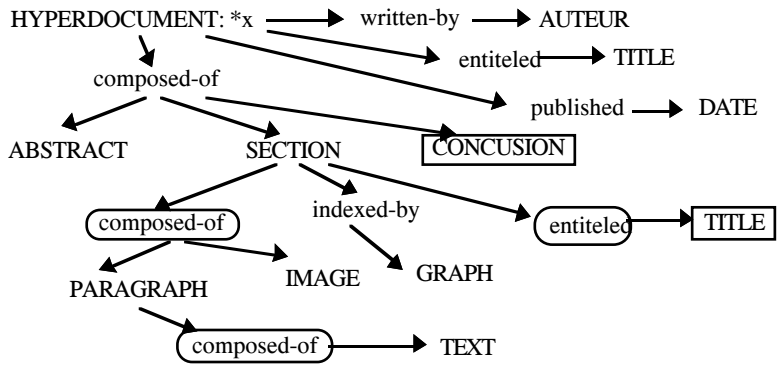
a. $\mathcal{T}=\mathcal{T}_S$

This lattice is based on generic/specific order relation where the symbol $\perp$ stands for the absurd (minimal) type. $\mathcal{T}_a$ is the sub-lattice that defines all the application domain knowledge, ie. contains concept types used for the indexing of documents (eg. TUMOR, CANCER, LUNG in medical applications). $\mathcal{T}_S$ is the sub-lattice where all object types that will be presented in the hyperbase will be defined. Object types can correspond to atomic objects (eg. TEXT, IMAGE, SOUND, ...) or complex (ie. structured) objects (eg.. ABSTRACT, SECTION, CONCLUSION, PARAGRAPH, HYPERDOCUMENT, ...). Note that the CG theory allows the definition of new object types or domain types by combining existing types (see example below). In such cases the new type will be placed below the variable concept of its definition.

Example

In this example, the new type ARTICLE, derived from the concept type HYPERDOCUMENT will be placed below HYPERDOCUMENT in $\mathcal{T}_S$ (ARTICLE ≤ HYPERDOCUMENT).

Type ARTICLE (x) **is**



b. $\mathcal{M}$ is the set of pre-defined CGs that are used as primitive knowledge to build every CG needed to represent the hypermedia system. This construction is based on the four canonical CG operators: copy, restrict, join and simplify [7]. CGs belonging to $\mathcal{M}$ are canonical, minimal (ie. they cannot be derived from any other graph of $\mathcal{M}$) and have a correct meaning. For the representation of every needed knowledge, $\mathcal{M}$ is partitioned into two distinct sub-sets $\mathcal{M}_S$ and $\mathcal{M}_a$, where $\mathcal{M}_a$ contains the graphs needed to generate domain knowledge, and $\mathcal{M}_S$ contains those needed to represent system knowledge (see figure 1 for some examples of such graphs).

c. $\mathcal{R}$: the lattice of conceptual relation types which is also a part of hyperindex definition. Here also, we have two distinct lattices $\mathcal{R} = \mathcal{R}_S$

hyperbase (browsing relations), and $\mathcal{R}_a$ holds for conceptual relation types that express domain knowledge (non browsing relations or links). In $\mathcal{R}_S$ we distinguish three different kinds of relations:

c.1. Hyperbase relation: they correspond to links between hyperbase objects, and can be classified into three sub-classes:

c.1.1. External links: such as *same-author*, *same-date*, *same-editor*: these symmetric links are calculated from attributes associated to hyperbase objects, as author, editor or date (see section 2).

c.1.2. Structural links: they describe the logical structure of documents such as *composed-of* or *part-of*. Other links are considered as structural like *illustrated-by*, *referenced-by*, *commented-by*.

c.1.3. Semantic links: such as *same-subject*. These links are calculated from internal attributes (indexes) associated to hyperbase objects.

c.2. Hyperindex relations: these are the generic/specific relations presented above and that define the type lattices $\mathcal{T}$, $\mathcal{R}$.

c.3. Hyperbase / hyperindex relation: this relation defines entry points from the hyperindex to the hyperbase and from the hyperbase to the hyperindex. They link hyperdocuments in the hyperbase to minimal canonical graphs ($\mathcal{M}$).

d. $\mathfrak{I}$ is the set of data objects (ie. strings, numbers, images, sounds, …) that constitute the atomic objects of the hyperbase.

From the sets $\mathfrak{T}$, $\mathfrak{M}$, $\mathfrak{R}$ and $\mathfrak{I}$ we can construct every conceptual graph that is needed to design the hyperindex, the hyperbase and the relation between the two components.
In Figure 1 there are examples of each set in this model.

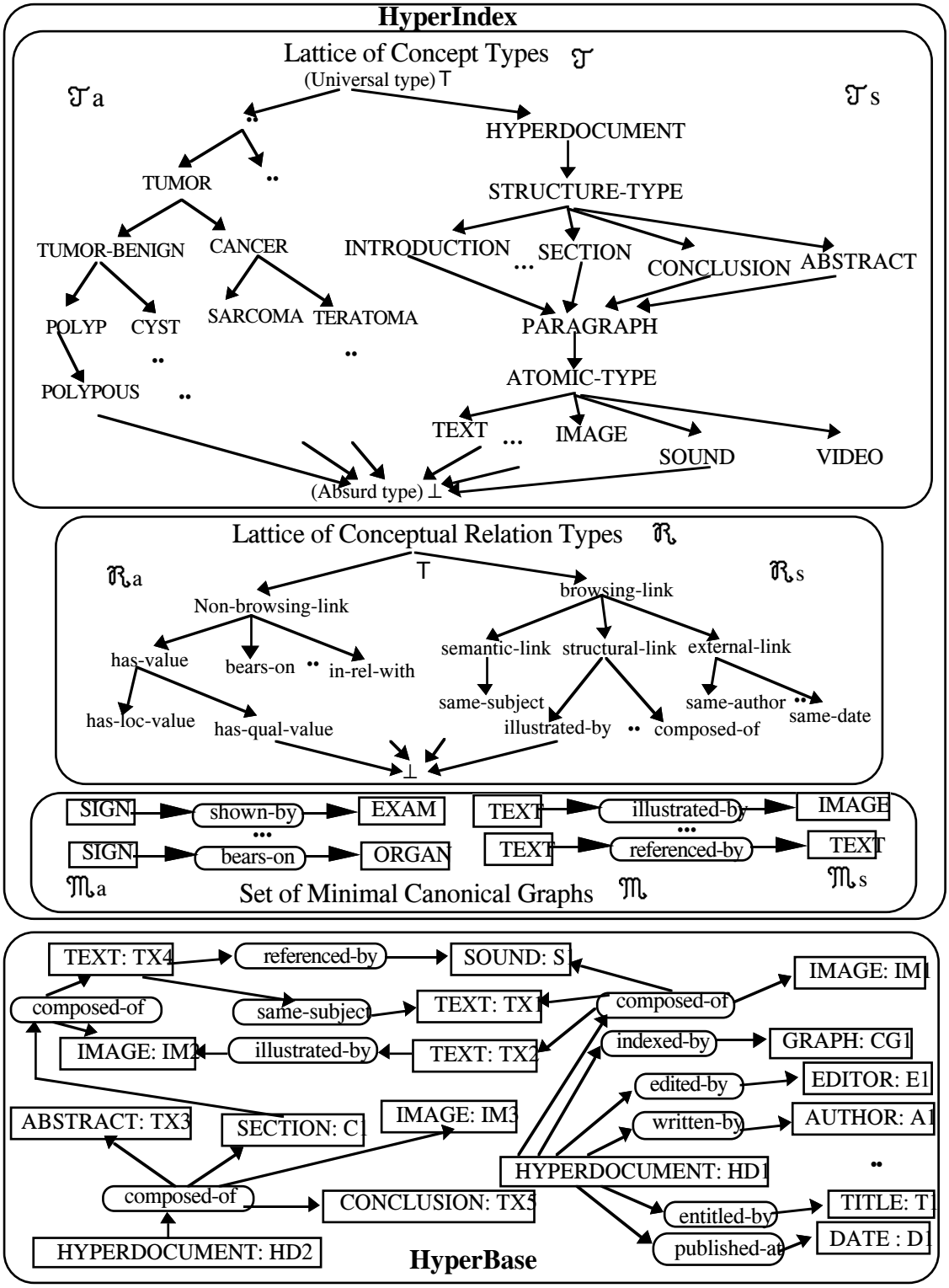


Figure 1: two-level hypermedia architecture

2. The IR model

In this section we extend our hypermedia model towards an IR model. In general, an IR model must contain a model for documents, a model for queries and a matching function (figure 2).

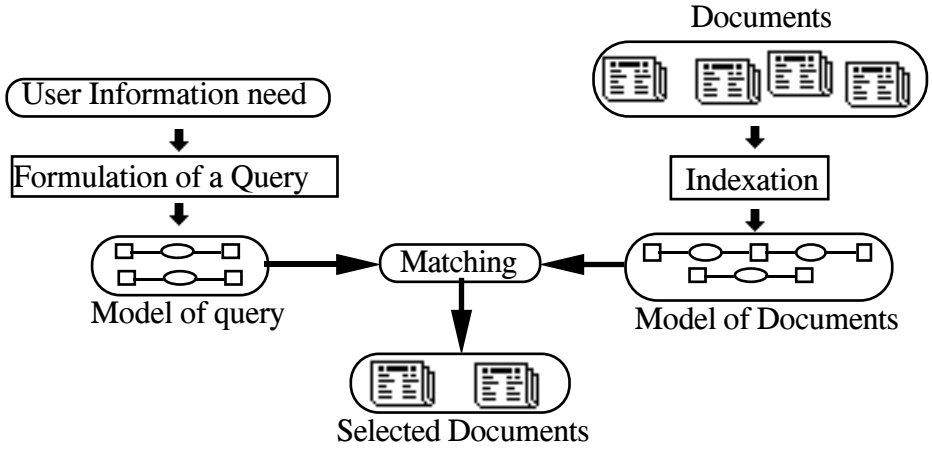


Figure 2: An IR model

In respect to our hypermedia model we propose the following:

2.1. Model of documents: the general knowledge about concept types ($\mathfrak{T}$), relation types ($\mathfrak{R}$) and minimal canonical CGs (m) presented before allows the representation of every kind of document, and the description of their semantic content (ie. their indexing). At this point we may notice that CGs allows the definition of classical attributes that are commonly attached to hypermedia nodes and documents. Figure 3 shows some examples of minimal canonical graphs that are used for expressing an attribute (AUTHOR), structural hierarchy (composed-of), indexing relation (indexed-by) and cross-references between documents (illustrated-by).

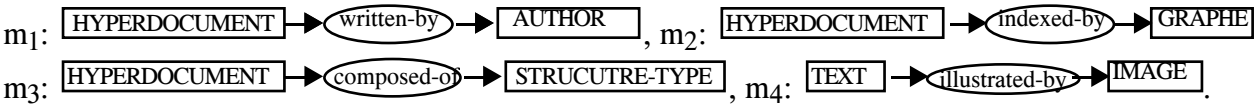


Figure 3: minimal canonical graphs for generating document representation by CGs

2.2. Model of queries: any query CG will be generated using the same model components presented above for the description of documents in the hyperbase. At this point we notice that queries may combine domain knowledge and structural knowledge (eg. find documents that talk about topic t and illustrated by images).

2.3. Matching function: is based on the logical model for IR introduced by *Van Rijsbergen* [8], and which is based on the principle: "given a query Q and a document D, D is relevant to Q if D logically implies Q". When queries and documents are represented using CGs, the CG theory tells that this logical implication is simply expressed by $\Phi(D) \supset \Phi(Q)$, where $\Phi(D)$ and $\Phi(Q)$ are the first-order predicates that are associated to the CGs D and Q. The CG theory also tells that this implication holds whenever there exist a so-called projection of Q into D or said, in other words, $D \leq Q$ [9], [7].
New extensions of the CG theory [10] include the notion of fuzzy CG which correspond to certain knowledge. This aspect is of extreme importance for the design of IR matching function that commonly takes into account the notion of uncertain matching [8]. This aspect, which also part of our model cannot be developed in this short paper.

Example
Figure 4 shows a representation of a document and a query in CGs. Note that the query can be projected on the document, so the document is considered relevant to the query. The projection of Q in D is illustrated by gray boxes.

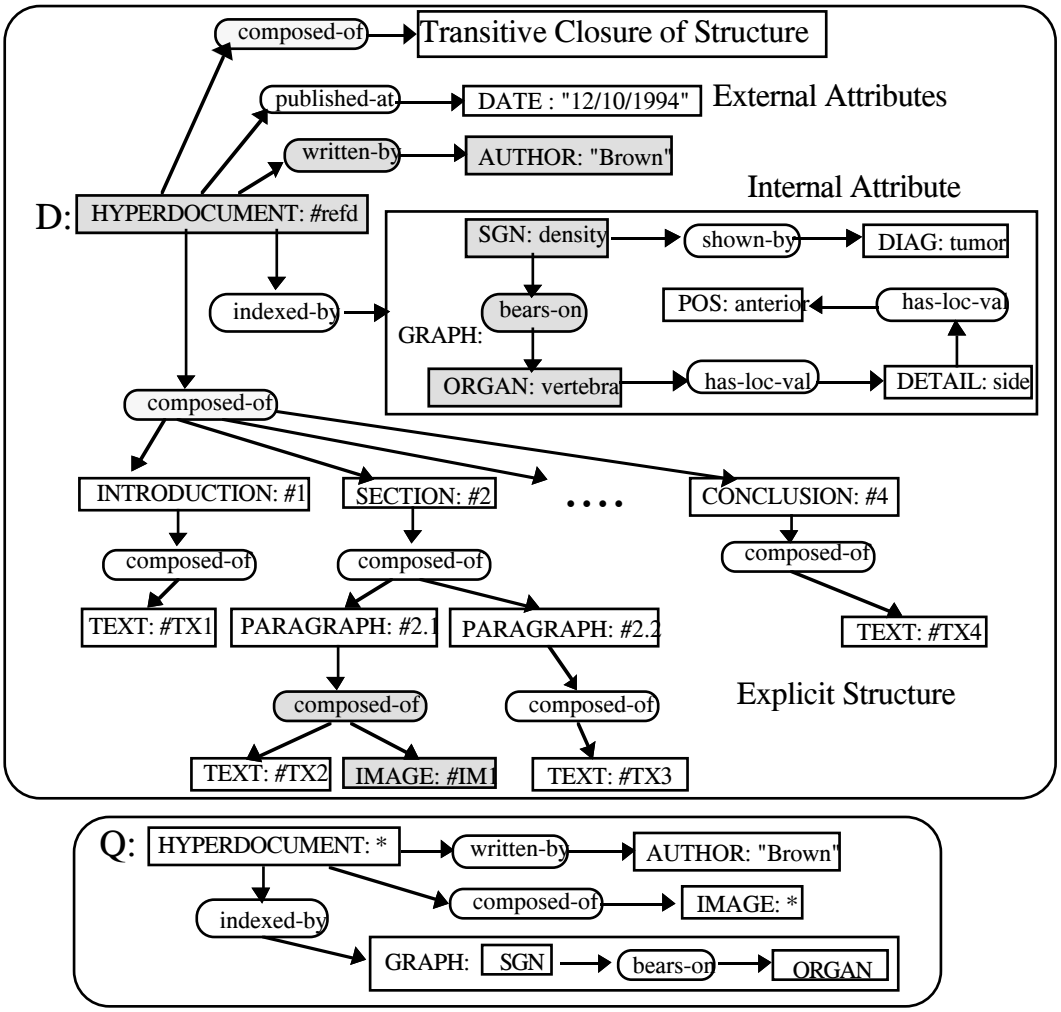


Figure 4: Example of a document, a query and the projection operation

3. Applying the model

Applying this model to a particular application implies several steps. The first step consists in defining the hyperindex components (sets $\mathfrak{T}$, $\mathfrak{R}$ and $\mathfrak{M}$). At this stage we have to notice that many basic types (for concepts and relations) and minimal CGs could be provided as a basic common knowledge for large classes of applications. Also it has to be noticed that the design of the hyperindex can be done in an incremental way by combination of available types and minimal CGs when needed. A next step is the building of the hyperbase itself, which uses the various types of concepts and relations of the hyperindex. Instances of nodes are created and linked either by logical structure links or by any other kind of links relevant for the application.

The model includes a complete strategy for indexing structured data that allow the retrieval of the most relevant components of complex objects, given a particular query ; this important feature of the model cannot be developed in this short paper. More intuitively the matching function used for retrieving information can be used also to for comparing documents (or components of documents), and constitutes a good basis for an automated or an assisted process for setting these links.

RIME is an implementation of our model combining hypermedia and IR. The application concerns medical date combining textual, structured medical reports and related X-ray images. The prototype is built on an OODBMS (O2 from O2 Technology), and its hypermedia interface is being developed using Mosaic of NCSA.

In our application of our conceptual view of structures, transformation of SGML structures to CGs structures has been developed in the RIME project. This tool allows an automatic integration of SGML documents in the hyperbase with the advantage of having explicit structure knowledge about these documents.

4. Conclusion

In this paper we have presented an approach to hypermedia systems that integrates an elaborated view of hypermedia systems (especially a high-level knowledge representation), and an equally elaborated view of IR systems (based on logical IR models). This integration has been made feasible by considering all aspects hybrid system as particular classes of knowledge (meanly domain and structured knowledge). This integration allows the design of systems that offer to the users the ability to browse and search in the same environment, and to take full advantage of the natural complementarity of these two means of satisfying information needs in large and complex information system. Important aspects such as indexing of complex structure and uncertain matching of CGs have been only mentioned in this paper, though they are well advanced in the development in the RIME project.

6. References

[1] G. Salton, C. Buckley. *On the Automatic Generation of Content Links in Hypertext.* Department of Computer Science, Cornell University, Ithaca, NY 14853-7501, rap. 89-993, April 1989.

[2] J. Allen, G. Salton. *The Identification of Text Relations Using Automatic Hypertext Linking.* Workshop on Intelligent Hypertext, CIKM'93, 5 Nov. 1993.

[3] T.P. Van Der Weide, P.D. Bruza. *Two level Hypermedia: An improved architecture for Hypertext.* Proc. of the Database and Expert System Applications, DEXA'90, Springer Verlage, Vienne, Autriche, Septembre 1990.

[4] P. Garg. *Abstraction mechanisms in Hypertext.* Com. of the ACM, Vol. 31, N° 7, July 1988.

[5] G. Richard, A. Rizk. *Quelques idées pour une modélisation des systemes hypertextes.* T.S.I. Technique et Science Informatique, Vol. 9, N°6, 1990.

[6] A. Kheirbek. *Two-level Hypermedia Model Based on Conceptual Graph Theory.* Workshop on Intelligent Hypertext, CIKM'93, 5 Nov. 1993.

[7] J. F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine.* Addison-Wesley publishing company, 1984.

[8] C.J. Rijsbergen. *A New Theoretical Framework for Information Retrieval.* Proc. of the ACM Conference on Research and Develoment in Information Retrieval, Pisa, September 1986, p. 194-200.

[9] J. P. Chevalet. *Un modèle logique de Recherche d'Informations appliqué au formalisme des Graphes Conceptuels, le prototype ELEN et son expérimentation sur un corpus de composants logiciels.* Thèse de l'Université J. Fourier, Grenoble, 15 mai1992.

[10] V. Wuwongse, M. Manzano. *Fuzzy conceptual graphs*. Proc. of ICCS'93, Quebec City, Canada, August 1993. Lecture Notes in Artificial Intekkigence, 699.