An algebra for structured documents in the context of the object-oriented approach
Sylviane Schwer, Haider Hamza, André Flory

To cite this version:
Sylviane Schwer, Haider Hamza, André Flory. An algebra for structured documents in the context of the object-oriented approach. 19 pages. 1999. <hal-00265817>

HAL Id: hal-00265817
https://hal.archives-ouvertes.fr/hal-00265817
Submitted on 20 Mar 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
An algebra for structured documents in the context of the object-oriented approach

Haider Hamza
INSA, LSI, 20, Avenue Albert Einstein 69621 Villeurbanne France
e-mail : hamza@lisflury.insa-lyon.fr

Sylvie R. Schwert
Université Paris 13, LIPN, av. J. B. Clément, 93430 Villetaneuse, France
e-mail: schwert@lipn.univ-paris13.fr

André Flury
INSA, LSI, 20, Avenue Albert Einstein 69621 Villeurbanne France
e-mail: flury@lisflury.insa-lyon.fr

Summary

The aim of this paper consists in defining an algebra allowing the request of structured documents in the context of object oriented approach. The operators of this algebra are defined in conformity with the concepts of the object model. It takes into consideration not only the embedding of structures and links between them, but also the informational aspect of the document as well as its structural aspect.

The operands of the proposed algebra are only one type: they are subdatabases. A subdatabase is made of a collection of the databases objects, grouped in classes and interconnected through links. This algebra responds to two objectives. Firstly, it represents a basic nucleus of a declarative query system; it contains the whole of the elementary operations which will be used in the resolution of the request. Secondly, it provides facilities to users to formulate their requests and manipulate the documentary database.

Key Words : Structured documents, Object-oriented, Documentary database, Query, Algebra

1. Introduction

Since the emergence of the relational model, many researches have addressed the problem of the development of documentary databases using the technology of database management relational systems (RDDBMS). For the development of documentary applications, the interest of this approach lies in the possibility of taking advantage of all the functionality of RDDBMS. Nevertheless the capacities of abstraction and the expressiveness of the relational model are insufficient for handling structured documents. Indeed, the representation of the fragments of the documents by several tables results in a heavy modeling and decreases considerably the performance of the system.

Researches aiming at a new generation of DBMS has been carried out to extend the models of data representation and to increase the power of manipulation languages. Indeed, new needs have emerged requiring references to the stored information in terms of the perception of the user and not simply according to their model inside the database. The object oriented DBMSs gives nowadays a quite rich model capable of representing complex data. A document can be modeled by a complex object the components of which may be formed by other objects belonging to different classes. The classes are organized in a hierarchy of inheritance in which the objects of different classes respect the relations of inclusion between them. The technology of OODBMS is well adapted to the management of structured documents such as arborescences of objects with associated methods [Amgh89, Bens89, Mous89, Ham89].

The retrieval of informations is one of the most important functionality of a management system of documentary databases. By system of retrieval of information, we mean the whole mechanisms which allows the user to select documentary information. One of the essential objectives of a retrieval system of informations is to make easy the restitution of a portion of information from a documentary database, in response to a user's request. Different strategies of information retrieval were developed, and most of them were built around an algebra. In fact, G. ting [Guo89] has proposed an algebra made of a set of operators which permits to take into account the different informational aspects of a document. But this algebra is not well adapted to an object-oriented system, because it doesn't consider the inter-objects relationships. In the case of the algebra for the OODBMS, several works were carried out [Ahla93, Clue90, Ham95, Liu93, Shau90, Shau90, Sub95]. However, these algebras are not really adapted to structured documents, since they only take into consideration the informational aspect of the data, and not at all their structural aspect. This is the case, for instance, of the algebra of Show and Zdonik, which is an extension of the algebra of the denormalized models that takes into account some object-oriented concepts. The EXCCESS algebra [van91], is a many-sort algebra i.e. the operators are defined by types of operators. Such an approach in a documentary retrieval context puts in question the uniformity of the operators and the reusing of the results of the requests for a new query. The algebra of Liu [Liu93] is based on a clear distinction between links of association and links of aggregation. We think that in fact this distinction is not relevant: two concepts can have different visions of the same reality. While one can model a relation as a link of composition, most models among the reviewed aboved treat it as a link of association.

The objective of our work consists in defining an algebra that allows the interrogation of documents in the context of an object-oriented approach. The basis of manipulation is the embedding of structures and the links between structures [Schw97]. This algebra permits to take into account the informational aspect as well as the structural aspect of documentary databases.

This paper is divided into 5 sections. The first one being this introduction, section 2 is devoted to the basic concepts of the proposed algebra. Section 3 is dedicated to the description of the algebraic operators and section 4 to examples of the algebraic operations. The last section consists in our conclusion and perspectives.

2. Formalization of a documentary database

In this part we define the basis of our algebra. Starting from a clear definition of what is a documentary database, we defined what is a subdatabase, the single type of our algebra and their operands which fit to the object-oriented ontology.

2.1. Documentary database

As usual in database theory, a documentary database is defined y a schema and an instance of this schema.

Definition 1: Documentary database schema

The documentary database schema can be defined by a finite acyclic connect graph. Formally, a schema $\Sigma$ of documentary database is defined by the triplet $(C, p, \lambda)$ where:

- $C = \{c|c\text{ is a set of classes, where each class } c\text{ is characterized by some properties } p_0, p_1, \ldots, p_n\}$,
- $p : C \rightarrow 2^C$ is an hierarchy function which associates a set of classes to a given class. We note by $p^{-1}$ the inverse relation of the hierarchy function, $p^*$ the transitive hierarchy function of order $n$, $p^*$ the transitive hierarchy function and $p^*$ the strict transitive hierarchy function which are defined as follows:
  - $p^0(c) = c$,
  - $p^n(c) = p^n(c) \cup p^{n-1}(c))$,
  - $p^*(c) = \bigcup_{\alpha \in p^0} p^\alpha$,
  - $p^*(c) = \bigcup_{\alpha \in p^0} p^\alpha$,
- $\lambda : C \rightarrow 2^C$ is a Roll function which associates a set of rolls to a pair of classes. The schema is endowed by the following properties:
  1) $\text{Root}(\Sigma) = \{C \subseteq C \text{ where } C \text{ is a connected component}\}$ (due to the fact that the graph is acyclic and finite)
  2) $(\forall \theta \in C) \bigwedge (\forall p \in p^0(c))$ (due to acyclicity)
  3) $(\forall \theta \in \text{Root}(\Sigma), \bigwedge (p^*(c) \wedge \emptyset)$ (this is the definition of the Root).
  4) $(\forall \theta \in \Sigma \wedge \text{Root}(\Sigma), \bigwedge (p^*(c) \wedge p^*(c) \wedge \emptyset)$ (this is due to the connection property).

Example:

Let $\Sigma = (C, p, \lambda)$ be the schema of the documentary database represented by fig. 1. The graph which corresponds to this schema is represented by fig. 2.
The documentary database schema can also be defined by a finite acyclic connected graph. We denoted it a $\Sigma$-graph

Fig. 1. Example of a schema of a documentary database

$$C=\{\text{Document (title, type, date-creation), Chapter (title, date-creation), Reference (reference), Section (title, date-creation), Content (text, type), Image (title, bitmap), Author (nom, name, address)}\}$$

$$p(\text{Document}) = \{\text{Author, Content, Chapter, Reference}\}$$
$$p(\text{Chapter}) = \{\text{Content, Section, Picture}\}$$
$$p(\text{Section}) = \{\text{Content, Picture}\}$$
$$p(\text{Reference}) = \{\text{Content, Picture}\}$$
$$p(\text{Author}) = \{\text{main-author, co-authors}\}$$
$$\lambda(\text{Document, Content}) = \{\text{introduction, conclusion}\}$$
$$\lambda(\text{Section, content}) = \{\text{body}\}$$
$$\lambda(\text{Section, Picture}) = \{\text{pictures}\}$$
$$\lambda(\text{Reference, Publication}) = \{\text{publication}\}$$
$$\lambda(\text{Chapter, Picture}) = \{\text{pictures}\}$$
$$\lambda(\text{Publication, Author}) = \{\text{authors}\}$$

The documentary database schema can also be defined by a finite acyclic connected graph. We denoted it a $\Sigma$-graph

A documentary database is defined by a schema and an instance of this schema.

**Definition 2 : Documentary database**

A documentary database is defined by $\{\Sigma, O, A\}$ where $\Sigma = (C, p, \lambda)$ is a schema of the database, $O$ is a set of objects and $A$ a set of inter-objects links $A \subseteq O \times O$. If we note by $\tau$ the total function $\tau : O \rightarrow C$ which returns for every object its class, then $\{\Sigma, O, A\}$ must verify the following constraint:

- $\forall (a_i, a_j) \in A, \tau(a_i) \in \text{pr}(\tau(a_j))$

Fig. 2. Representation of the schema of the documentary database under the form of a connected acyclic graph

2.2. Documentary subdatabase

A user who wants to formulate queries on a database, is rarely interested in the whole entities in the database. In most cases, he wants to formulate queries on a portion of this database. For example, a user would like to formulate queries only on the Author class and Document class of our illustration.

Fig. 3. Example of a partial view of an instance of a documentary database
We defined a subdatabase of a database D as a database which is merged into the database D.

**Definition 3 : Subschema**

A subschema \((\psi, \rho_{\psi}, \lambda_{\psi})\) of a schema \((\Sigma, \rho_{\Sigma}, \lambda_{\Sigma})\) is included in the \(\Sigma\) graph, such that:
- \(\psi \subseteq \Sigma\)
- \(\forall e \in \psi \subseteq \rho_{\psi}(e) \subseteq \rho_{\Sigma}(e)\)
- \(\forall e_1, e_2 \in \psi \subseteq \lambda_{\psi}(e_1, e_2) \subseteq \lambda_{\Sigma}(e_1, e_2)\)

Remarks:
- a schema is a subschema of itself.
- a class of a schema is a subschema of the schema.
- \((\psi, \rho_{\psi}, \lambda_{\psi}) = (\psi, \rho_{\psi}, \lambda_{\psi})\) is a subschema of any schema.

**Example:**

Fig. 4 represents the graph of the subschema \(\psi = (\psi, \rho_{\psi}, \lambda_{\psi})\) which is defined as:
\[
\psi = \{ \text{Document (title, type, creation-date), Chapter (title, creation-date), Reference (reference)} \}
\]
\[
\rho_{\psi}(\text{Document}) = (\text{Chapter, Reference})
\]
\[
\rho_{\psi}(\text{Chapter}) = \emptyset
\]
\[
\rho_{\psi}(\text{Reference}) = \emptyset
\]

**Definition 4 : An instance of a subschema**

An instance of a subschema \((\psi, \rho_{\psi}, \lambda_{\psi})\) is a set of objects and a set of links between objects which respectively correspond to the classes in the subschema and to the links between classes. Formally, an instance is defined by the pair:
\[
(\Omega, A)
\]
where \(\Omega\) is a set of objects and \(A\) a set of inter-objects links \(A \subseteq \Omega \times \Omega\) with the following constraint:
- \(\forall (x, y) \in \Omega \times \Omega \text{ where } (r(x, y)) \in \rho_{\psi}(x)\)

**Example:**

Fig. 5 gives a graphical representations of an instance of the subschema that is represented in fig. 4.

**Definition 5 : Subdatabase**

A subdatabase is defined by a subschema and by an instance of this subschema. Formally, a subdatabase is defined by the triple:
\[
(\psi, \Omega, A)
\]

**Example**
8. Distance between two subschemas

Given two subschemas \( \psi_1 = (C_{\psi_1}, \rho_{\psi_1}, \lambda_{\psi_1}) \) and \( \psi_2 = (C_{\psi_2}, \rho_{\psi_2}, \lambda_{\psi_2}) \), the distance between \( \psi_1 \) and \( \psi_2 \), denoted by \( D(\psi_1, \psi_2) \), is defined as:

\[
D(\psi_1, \psi_2) = \text{Min} \left( \text{Len}(\psi_1) \cap \text{Len}(\psi_2) \right)
\]

Example:

The following is a distance between subschemas \( \psi_1 \) and \( \psi_2 \) shown in figure 6:

\[
D(\psi_1, \psi_2) = \text{Min} (\text{Len}(\text{Document}), \text{Len}(\text{Document.Author.Publication})) = 2
\]

9. The upper limit of two subschemas

Given two subschemas \( \psi_1 = (C_{\psi_1}, \rho_{\psi_1}, \lambda_{\psi_1}) \) and \( \psi_2 = (C_{\psi_2}, \rho_{\psi_2}, \lambda_{\psi_2}) \), the upper limit \( \psi_1 \vee \psi_2 \) is the least schema containing \( \psi_1 \) and \( \psi_2 \). The upper limit can be defined as the union of \( \psi_1 \) and \( \psi_2 \) and all the paths between \( \psi_1 \) and \( \psi_2 \) that have their lengths equal to \( D(\psi_1, \psi_2) \). Let us consider the set \( S \) of all the paths between \( \psi_1 \) and \( \psi_2 \) that have their lengths equal to \( D(\psi_1, \psi_2) \). The set \( S \) is defined as:

\[
S = \left\{ C_{\psi_1}, \rho_{\psi_1}, \lambda_{\psi_1}, C_{\psi_2}, \rho_{\psi_2}, \lambda_{\psi_2} \right\} \cap \left( \text{Len}(\psi_1) \cup \text{Len}(\psi_2) \right)
\]

Formally, the upper limit of two subschemas \( \psi_1 = (C_{\psi_1}, \rho_{\psi_1}, \lambda_{\psi_1}) \) and \( \psi_2 = (C_{\psi_2}, \rho_{\psi_2}, \lambda_{\psi_2}) \), is defined as:

\[
\psi_3 = \psi_1 \vee \psi_2 = (C_{\psi_1}, \rho_{\psi_1}, \lambda_{\psi_1}) \text{ where:}
\]

\[
- C_{\psi_3} = C_{\psi_1} \cup C_{\psi_2}
\]

\[
- \forall C \in C_{\psi_1}, \rho_{\psi_1}(C) = \rho_{\psi_2}(C) \cup \rho_{\psi_1}(C)
\]

\[
- \forall C_1, C_2 \in C_{\psi_3}, \lambda_{\psi_3}(C_1 \cup C_2) = \lambda_{\psi_1}(C_1) \cup \lambda_{\psi_2}(C_2)
\]

Example:

An example of an upper limit of two subschemas is shown in fig. 8.
The algebraic operators have as operands a subdatabase and produce a new subdatabase. In this algebra, the closure property is maintained, because the results produced by the queries are structured in the same manner as the operands. Also, this algebra allows to produce a new class of objects that enriches the existing database.

The algebraic operators will be formally defined in the appendix. The examples used to explain these operators will make use of the database shown in fig. 1 and 3. We extend algebraic operators to subdatabases.

3.1. Select (σ)

The **Select** is a unary operator, which operates on a subdatabase to produce a new subdatabase in which objects and inter-objects links satisfy a specified predicate. The resultant subdatabase has the same schema as the operand subdatabase, and the instances of the result is a subset of the instances of the operands which satisfy the predicate. The select operation is denoted \( σ(P) \) where \( X \) is an operand subdatabase and \( P \) is a predicate. The predicate \( P \) is a logical expression that is evaluated to true or false. This logical expression is composed by terms interconnected by logic operators (and, or, not). Each term can be, either a condition on an attribute value, or a condition on an inter-object link. The condition on the attribute values have the following form:

\[
\begin{align*}
\text{a} & \theta \text{constant, or a b} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and Boolean} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and string} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\text{a} & \theta \text{a b constant and} \\
\end{align*}
\]

where \( a \) and \( b \) represent attributes.

- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
- \( a \theta a \) represents attributes,
3.4. Union, Intersection and Difference (\( \cup, \cap, \setminus \))

The union operator is a binary operator which constructs a new subdatabase by combining two subdatabases. The subschema of the result subdatabase is defined as the upper limit of the subschemas of the operands subdatabases. The instance of the result subdatabase is constituted by the union of the set of instances of operands which are extended to the subschema of the result subdatabase. The union operation is denoted \( X \cup Y \).

The intersection operator, as the union operator, is a binary operator which constructs a new subdatabase by combining two subdatabases. The subschema of the result subdatabase is defined as the lower limit of the subschemas of the operands subdatabases. The instance of the result subdatabase is constituted by the intersection of the set of instances of the subdatabases operands. The intersection operation is denoted \( X \cap Y \).

The difference operator, as the union and intersection operators, is a binary operator which constructs a new subdatabase by subtracting from a subdatabase \( X \) of an other subdatabase \( Y \). The subschema of the result subdatabase is defined as the subschema of \( X \). The instance of the result subdatabase is constituted by the instance of \( X \) to which is subtracted the instances of the objects of \( Y \). The difference operation is denoted \( X \setminus Y \).

Example:

Examples of union, intersection and difference operations are shown in fig. 14.

3.5. Join

Unlike the relational data model, in object-oriented models many relationships between objects can be represented within the objects themselves. Hence the join operations are used less frequently in the object algebra than in the relational algebra. However, the existing objects in a database may not explicitly reflect all relationships required by the queries. An explicit join is still needed to handle the cases when the relationship being queried upon is not defined within the object classes. Such relationship is called value-based relationship in contrast with those specified explicitly in the object model. For example, suppose we have a layer document class and a technical document class. A query «Find the layer document which has the same key words as the technical document», requires a value-based join between the classes layer document and technical document through the attribute keywords.

The join condition links the properties of only one class of \( X \) and the properties of only one class in \( Y \). The subschema of the result subdatabase is defined as the concatenation of the subschemas of the subdatabases operands through a new class. This new class is defined as a set of tuples. The tuple is made of two attributes, the first attribute is a set of references to the class \( c_1 \) of \( X \), the second attribute is a set of references to a class \( c_2 \) of \( Y \). The graphical structure of class \( c \) is shown in fig. 15. The instance of the result subdatabase is constituted of a portion of \( Y \) that satisfies the join condition, a portion of \( Y \) that satisfies the join condition, and the union of the instances of...
operands which are extended to the subschema of the result subdatabase. The Join operation is denoted by $X \bowtie [P] Y$ where $X$ and $Y$ are the subdatabases operands and $P$ represents the join predicate.

**Example:**

Let $X$ and $Y$ be the subdatabases represented by the fig. 16. An example of join operation is:

$$X \bowtie \{ \text{Chapter-date-creation} = \text{Part-date-creation} \} Y$$

if, we suppose that the values of the attributes $\text{date-creation}$ of the $\text{reference-guide}$ class and of the $\text{user-guide}$ class are:

- $c_i.\text{creation-date} = "26-03-62"$,
- $c_i.\text{creation-date} = "26-03-62"$,
- $c_i.\text{creation-date} = "05-07-62"$,
- $c_i.\text{creation-date} = "05-07-62"$,
- $p_i.\text{creation-date} = "26-03-62"$,
- $p_i.\text{creation-date} = "26-03-62"$,
- $p_i.\text{creation-date} = "03-07-62"$,
- $p_i.\text{creation-date} = "03-07-62"$.

The result of the join operation is shown in fig. 16. The subschema of the result subdatabase is composed of the subschema of $X$ and of $Y$, and of a new $\text{Chapter-Part}$ class which refer to $\text{Chapter}$ class and $\text{Part}$ class. The $\text{Chapter-Part}$ class is composed of two objects $f_i$ and $f_j$. The object $f_i$ is linked to the objects $\{c_i, p_i\}$ of the $\text{Chapter}$ class and it is also linked to the objects $\{p_i, p_j\}$ because the two sets of objects have the same value of the $\text{date-creation}$ attribute and the join condition is fulfilled. For the same reason, the object $f_j$ is linked to the objects $\{p_i, p_j\}$. However, the object $c_i$ of $\text{Chapter}$ class is not present in the result, because it doesn't fulfill the join condition. Consequently, the objects $\{x, x \}$ of the $\text{Selection}$ class are also omitted from the result.

### 3.6 Grouping ($\gamma$)

The grouping is an unary operator. This operator is also used in NST-Algebra [Gu09]. It is used to transform an association relationship to an aggregation relationship. The grouping operation is defined as $\gamma(X[c_i, c_2, a])$ where $X$ is an operand subdatabase and $\{c_i, c_2, a\}$ is the association relationship to be transformed to an aggregation relationship.

**Example:**

An example of a grouping operation is shown in fig. 17.

### 3.7 Split ($\alpha$)

The split operation is the inverse of the grouping operation. Indeed, the split operation is used to transform an aggregation relationship to an association relationship. The split operation is defined as $\alpha(X[c_i, c_2, a])$ where $X$ is an operand subdatabase and $\{c_i, c_2, a\}$ is the aggregation relationship to be transformed to an association relationship.

**Example:**

An example of the split operation is shown in fig. 18.
3.8 Distribution (δ)

The distribution operator is used to restructure a subdatabase. This restructuration is made by distributing a property of class \( c_2 \) to a class \( c_1 \). The distribution is valid, if there exists a link between \( c_1 \) and \( c_2 \). The distribution operation is defined as \( \delta(X)[c_1, c_2, p] \) where \( X \) is a subdatabase operand, \( p \) is the property of \( c_2 \) to be distributed to class \( c_1 \).

Example:
An example of the distribution operation is shown in figure 19.

4. Examples of expressions of algebraic queries

A storage and a query system of structured documents using the OODHMS technology was proposed in [Ch94]. This system was developed in the Verso project of INRIA. It is based on an extension of the model and query language of O2 OODHMS [Mes90]. In order to request documents by their structure and content, O2SQL was extended to:
- take into account new functions of modeling like the ordered n-tuples and unions of types
- include predicates of textual retrieval like contains.

We will present examples of queries to compare our algebraic approach with the O2SQL query language.

Example 1:
«Find the documents containing the word Computer in the title and which are constituted of two chapters.»
The request is written as follows in the O2SQL query language:

\[
\text{Select } s \\
\text{From } y \text{ in Document} \\
\text{Where } s \text{.title contains } \text{computer } \text{ and count } (s \text{.chapter}) = 2
\]

The algebraic expressions of the request are:
\[
R_1 := \pi [\text{Document, Chapter}] (\pi \text{(Document.title \ contains \ computer \ s) \ where \ (Document} \ \text{.Chapter} = 2)) \\
R_2 := \pi [R_1] (\text{Document}) \\
R_3 := \pi \text{[Document, Chapter]} (\text{Document.title \ contains \ computer \ s) \ where \ (Document} \ \text{.Chapter} = 2)
\]

5. Conclusion
We have presented in this paper an algebra for structured documents in an object-oriented approach. We have defined the concept of subdatabase on which is based the definition of algebraic operators. Indeed, the operands and the query result are subdatabases, as a consequence, the closing condition and the operators orthogonality conditions are respected because the operands and the results of the query operators are structured in the same way. The operators can be regrouped into two categories. A first category that deals with information retrieval operators (select, project, associate, union, intersection, difference and join operators). A second category that deals with information restructuring operators (grouping, split, distribution operators).

Appendix
The formal definition of the algebraic operators are given below:

- Select
Let us consider a subdatabase \( X = (\psi_X, O_X, A_X) \) (where \( \psi_X = (C_X, \rho_X, \lambda_X) \)). The select operation is defined as:

\[
Y = \sigma (\theta (X)) (\psi_Y, O_Y, A_Y) \text{ where } \psi_Y = (C_Y, \rho_Y, \lambda_Y) \text{ such as:} \\
- C_Y = C_X, \\
- \forall c \in C_Y, \rho_Y(c) = \rho_X(c), \\
- \forall c_1, c_2 \in C_Y, \lambda_Y(c_1, c_2) = \lambda_X(c_1, c_2), \\
- O_Y = \theta (\psi_Y \circ \theta (\lambda_Y), P(\theta) = \text{true}).
\]

- Project
Let us consider a subdatabase \( X = (\psi_X, O_X, A_X) \) (where \( \psi_X = (C_X, \rho_X, \lambda_X) \)). The project operation is defined as:

\[
T = \pi (\theta (X)) (\psi_T, O_T, A_T) \text{ where } \psi_T = (C_T, \rho_T, \lambda_T) \text{ such as:} \\
- C_T = \{x_1, x_2, \ldots, x_n\}, \\
- \forall c_1, c_2 \in C_T, \lambda_T(c_1, c_2) = \lambda_X(c_1, c_2), \\
- O_T = \theta (\psi_T \circ \theta (\lambda_T), P(\theta) = \text{true}).
\]

- Associate
Let us consider a subdatabase $X = (\psi_X, O_X, A_X)$ ($\psi_X = (C_X, p_X, A_X)$) and a database $D = (\Xi, O, A)$. The association operation is defined as:

$Z = X \circ \{ \psi_X, O_X, Y \circ \{ \psi_Y, O_Y, A_Y \} \}$

where $Z = (C_Z, p_Z, A_Z)$ such as:

- $C_Z = C_X \cup C_Y$
- $p_Z = p_X \cup p_Y$
- $A_Z = A_X \cup A_Y$

- $X = X \circ \{ \psi_X, O_X, A_X \}$
- $Y = Y \circ \{ \psi_Y, O_Y, A_Y \}$

1. **Union**

Let us consider a subdatabase $X = (\psi_X, O_X, A_X)$ ($\psi_X = (C_X, p_X, A_X)$) and $Y = (\psi_Y, O_Y, A_Y)$ ($\psi_Y = (C_Y, p_Y, A_Y)$) where $Z = (C_Z, p_Z, A_Z)$ such as:

- $Z = X \cup Y$
- $\psi_Z = \psi_X \cup \psi_Y$
- $A_Z = A_X \cup A_Y$

2. **Intersection**

Let us consider a subdatabase $X = (\psi_X, O_X, A_X)$ ($\psi_X = (C_X, p_X, A_X)$) and $Y = (\psi_Y, O_Y, A_Y)$ ($\psi_Y = (C_Y, p_Y, A_Y)$) where $Z = (C_Z, p_Z, A_Z)$ such as:

- $Z = X \cap Y$
- $\psi_Z = \psi_X \cap \psi_Y$
- $A_Z = A_X \cap A_Y$

3. **Difference**

Let us consider a subdatabase $X = (\psi_X, O_X, A_X)$ ($\psi_X = (C_X, p_X, A_X)$) and $Y = (\psi_Y, O_Y, A_Y)$ ($\psi_Y = (C_Y, p_Y, A_Y)$) where $Z = (C_Z, p_Z, A_Z)$ such as:

- $Z = X \setminus Y$
- $\psi_Z = \psi_X \setminus \psi_Y$
- $A_Z = A_X \setminus A_Y$

4. **Join**

Let us consider a subdatabase $X = (\psi_X, O_X, A_X)$ ($\psi_X = (C_X, p_X, A_X)$) and $Y = (\psi_Y, O_Y, A_Y)$ ($\psi_Y = (C_Y, p_Y, A_Y)$) where $Z = (C_Z, p_Z, A_Z)$ such as:

$Z = X \circ \{ \psi_X, O_X, A_X \} \circ \{ \psi_Y, O_Y, A_Y \}$

where $Z = (C_Z, p_Z, A_Z)$ such as:

- $C_Z = C_X \cup C_Y$
- $p_Z = p_X \circ p_Y$
- $A_Z = A_X \cup A_Y$

References


