

Conceptual Exploration of Topic Maps

Benedicte Desclefs – Le Grand
PhD Student
Laboratoire d Informatique de Paris 6
8 rue du Capitaine Scott 75015 Paris,
France
tel : 33 1 44 27 75 12
fax : 33 1 44 27 74 95
email : Benedicte.Le-Grand@lip6.fr
web : <http://www.lip6.fr/rp/~blegrand>

Michel Soto
Associate Professor
Laboratoire d Informatique de Paris 6
8 rue du Capitaine Scott 75015 Paris,
France
tel : 33 1 44 27 88 30
fax : 33 1 44 27 74 95
email : Michel.Soto@lip6.fr
web: <http://www.lip6.fr/rp/~ms>

KEYWORDS

XML, Topic Maps Visualization, 3D, Galois classification algorithm

BIOGRAPHIES

Benedicte Desclefs – Le Grand was born in 1975. She received her engineer diploma from the Institut National des Telecommunications in 1997 and is currently a PhD student at LIP6 (Laboratoire d'Informatique de Paris 6) in the Network Department.

Her research deals with Virtual Reality and its use for complex systems visualization. She has been working on XML for several years and she is particularly interested in topic maps; Benedicte was a speaker at ACM CIKM'99 (Conference on Information and Knowledge Management) in Kansas City, at Markup Technologies'99 in Philadelphia and at XML Europe'2000 in Paris. She is a founding member of TopicMaps.Org.

Michel Soto is an associate professor at the LIP6 Laboratory (Laboratoire d'Informatique de Paris 6), University Pierre et Marie Curie, Paris, France. His main research interests are currently virtual worlds interoperability and complex systems visualization.

Introduction

Topic maps provide a bridge between knowledge representation and information management. Topics and associations build a semantic network above information resources, which allows users to navigate at a higher level of abstraction.

As stated in the ISO 13250 standard [ISO13250], “a topic map defines a multidimensional topic space”. A topic has one or more names (basename, dispname, sortname) within a scope; it can also have occurrences and may play a role in zero or more associations. Associations, association roles and occurrence roles have a type which is a topic, etc. A topic map is actually a high-dimensional knowledge base.

Moreover, topic maps may contain lots of topics and associations, and it might be very difficult for a user to know where to start within the topic map to explore it.

In this article, we describe a new means of interpreting and inferring information from topic maps. This analysis is based on Galois classification algorithm; it allows to find new properties about topics and associations and helps users in their information retrieval process.

This paper is organized as follows: in the first part, we describe users' needs and current navigation systems' weaknesses. In the second part, we explain our approach's basic principles. Finally our first results are discussed.

1. What are users' needs ?

Topic maps can be used in two different ways, depending on whether the user has precise needs or not.

As far as the precise search is concerned, topic maps query languages are being implemented and seem to be appropriate means of finding an answer to a specific request. The topic map only needs to be filtered so as to display only relevant information. But this only concerns a limited number of scenarios.

If there is no specific request from the user, it might be very difficult to know where to start and how to navigate within the topic map. In this case, the global topic map needs to be considered, as there is no information about the user's interests. The problem is that topic maps can be very large and complex. A few user interfaces have been developed so far. A very interesting and intuitive one is available on Infoloom Inc. web site [Infoloom] and allows us to navigate through GCA conferences' proceedings, as shown on figure 1. Usually, a list of topics – or associations - is displayed, and users choose a current topic – or association. When a topic is selected, everything about this topic is displayed – its attributes, the associations it is involved in, the topics it is associated with.

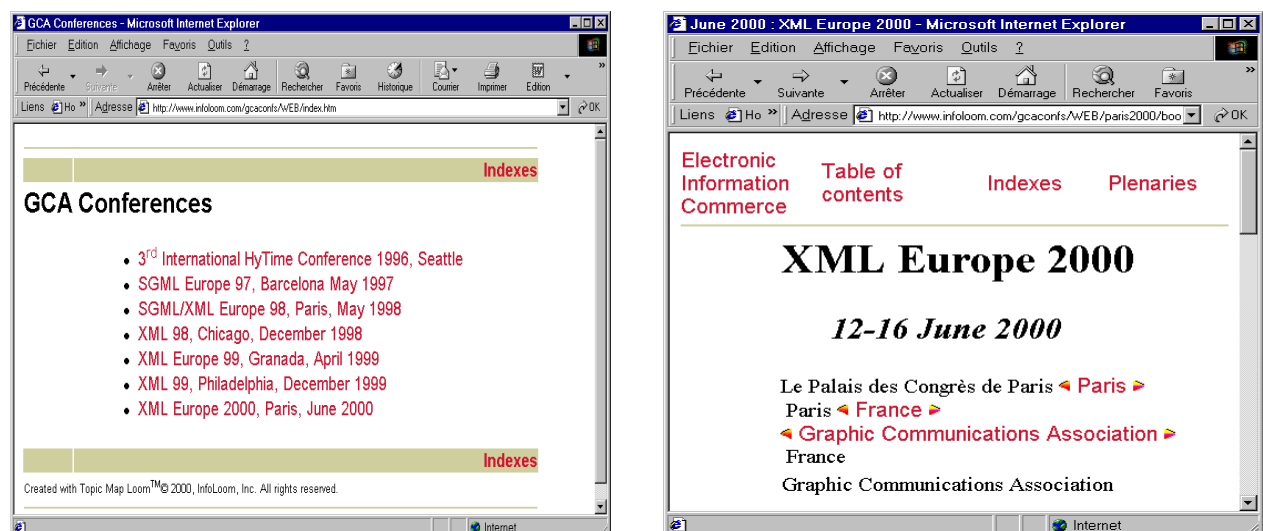


figure 1. Navigation through GCA conferences proceedings

This navigation is intuitive and provides a lot of information. However, the problem lies in the choice of the starting point – the first selected topic or association.

At LIP6, we developed a topic map visualization tool [LeGrand2000] based on virtual reality techniques and especially the use of 3D and interaction (see figure 2). However, the same problem as stated before appears when the topic map is very large.

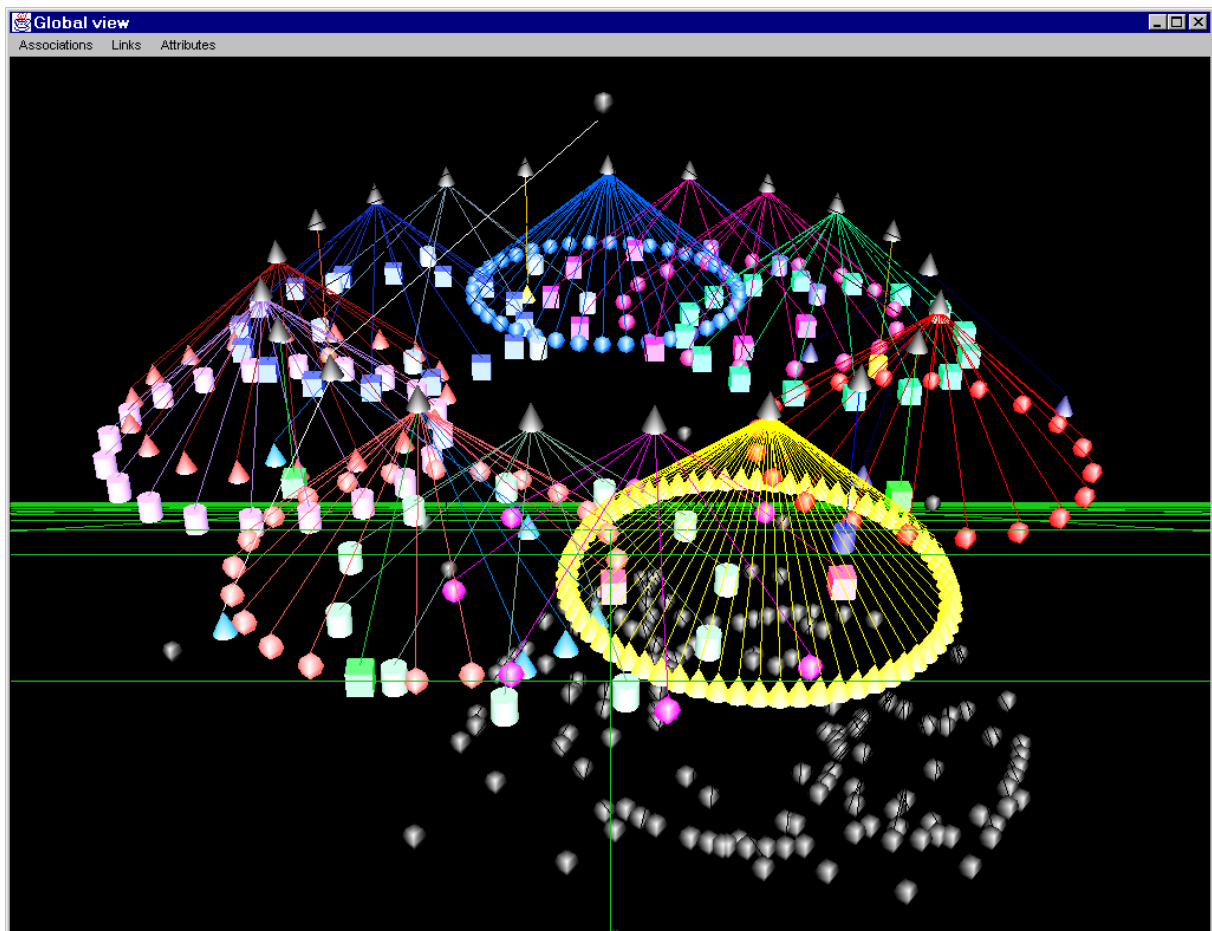


Figure 2. Topic map 3D interactive visualization

We now aim at helping users understand the global meaning of the topic map. We want to provide navigation hints to help them decide where to start and where to go within the topic map. Our approach's basic concepts are explained in the second part of this paper.

2. Conceptual analysis of topic maps

In the first part, we showed that it could be very difficult to navigate within a topic map - which may be a very large and complex structure - when there is no precise research goal. We suggest that some information can be inferred from the initial topic map, allowing users to refine their scope of interest.

The *scope* attribute is intended to provide topics and associations' context and validity. However, scope is not always correctly defined and sometimes even not used at all. We investigated another way of inferring "meta-information" about the topic map, through the use of Galois classification algorithm and Galois lattices.

2.1. Introduction to Galois lattices

The notion of Galois lattice for a relationship between two sets is the basis of a set of conceptual classification methods. This notion was introduced by Barbut and Monjardet in

[Bar70]. Galois lattices consist in grouping objects into classes that materialize concepts of the domain under study. Individual objects are discriminated according to the properties they have in common.

Let us first introduce Galois lattices basic concepts.

Let two finite sets E and E' (E consists of a set of objects and E' is the set of these objects' properties), and a binary relation $R \subseteq E \times E'$ between these two sets. Figure 3 shows an example of binary relation between two sets. According to Wille's terminology [Wil92], the triple (E, E', R) is a formal context which corresponds to a unique Galois lattice. It represents natural regroupings of E and E' elements.

Let $P(E)$ a partition of E and $P(E')$ a partition of E' . Each element of the lattice is a couple, also called concept, noted (X, X') . A concept is composed of two sets $X \in P(E)$ and $X' \in P(E')$ which satisfy the two following properties :

$$X' = f(X) \text{ where } f(X) = \{ x' \in E' \mid \forall x \in X, xRx' \}$$

$$X = f'(X') \text{ where } f'(X') = \{ x \in E \mid \forall x' \in X', xRx' \}$$

A partial order on concepts is defined as follows :

Let $C1=(X1, X'1)$ and $C2=(X2, X'2)$,

$$C1 < C2 \Leftrightarrow X'1 \subseteq X'2 \Leftrightarrow X2 \subseteq X1$$

This partial order is used to draw a graph called a Hasse diagram, as shown on figure 3. There is an edge between two concepts $C1$ and $C2$ if $C1 < C2$ and there is no other element $C3$ in the lattice such as $C1 < C3 < C2$. In a Hasse diagram, the edge direction is upwards. This graph can be interpreted as a representation of the generalisation / specialization relationship between couples, where $C1 < C2$ means that $C1$ is more general than $C2$ (and $C1$ is above $C2$ in the diagram).

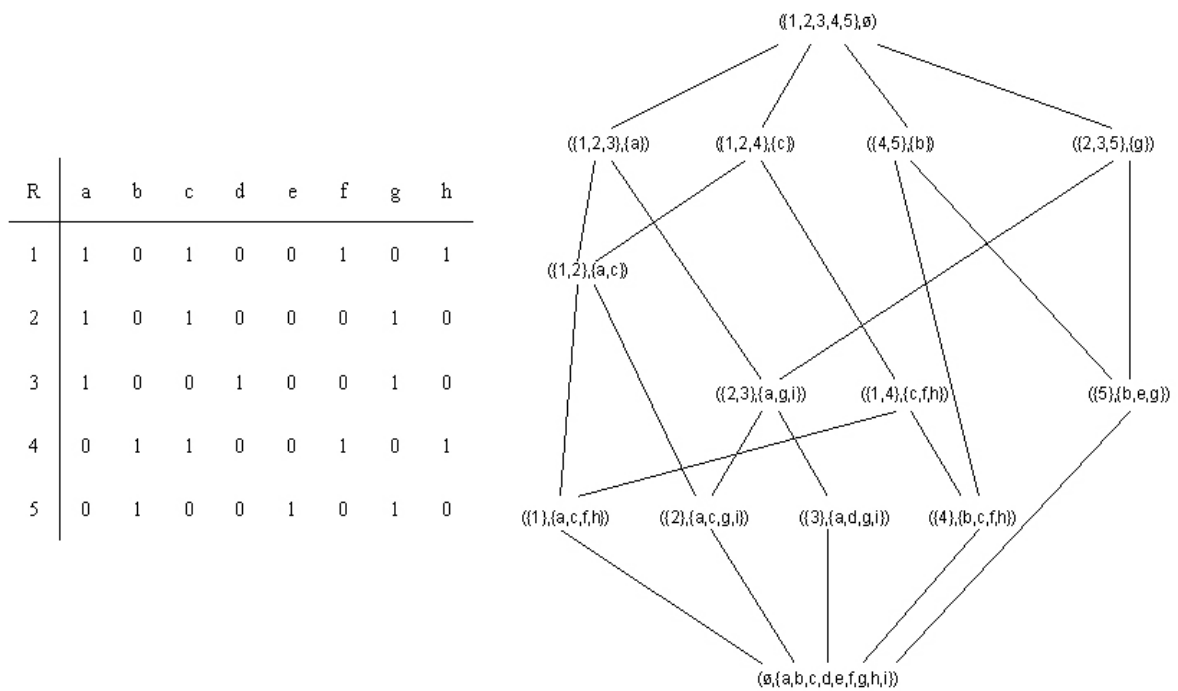


figure 3. Binary relationship and associated Galois lattice representation (Hasse diagram)

The concept lattice shows the commonalities between the classes of the context. The first part of a concept is the set of classes – or objects - and the second part reveals their common properties.

2.2. Generation of Galois algorithm input

We generated Galois lattices from different topic maps, written in XML. The first step was to create the input of the Galois algorithm, which consists of several pairs. Each pair contains one object and its associated properties. Topic maps are multidimensional knowledge bases. Topics have many characteristics, such as names, types, roles in associations, roles in occurrences, all of which depend on context – called scope. Association role and occurrence role are topic themselves. Associations have types... All these attributes should appear as objects or properties. We will explain the different possibilities in the following of this article.

Once the input is created, the Galois lattice is incrementally built. The algorithm we implemented is an adaptation of the one proposed in [Godin98].

The output of the classification is a lattice of concepts, i.e. interconnected sets of objects with their common properties. In next section, we will explain how this lattice can be interpreted and how we use these results to help users in their navigation through a topic map.

We followed two distinct scenarios – intuitive and recursive - to generate the input of the algorithm.

Intuitive scenario

We first had to decide what would be objects in our topic map. Both topics and associations are fundamental. The object symbolizing a given topic is its *id* attribute. For an association, the corresponding object is its *type* attribute.

Then we need to define these objects' properties. Topics and associations are both XML elements. We chose to characterize these elements with all their attributes, their children's attributes and more for associations, which we will explain later.

Consider the following XML fragment :

```
<topic id="testarossa" types="ferrari" >
  <topname scope="vehicle" >
    <basename>Ferrari Testarossa</basename>
    <dispname>Testarossa Ferrari</dispname>
    <sortname>FerrariTestarossa</sortname>
  </topname>
  <occurs>
    <type="or-webSite" href="http://www.ferrari.com"/>
  </occurs>
</topic>
<assoc type= "at-modelCountry">
  <assocrl role="or-model" href="testarossa"/>
  <assocrl role="or-country" href="italy" />
</assoc>
```

The corresponding input of the Galois algorithm is :

```
(testarossa, [ferrari, vehicle, or-webSite, http://www.testarossa.nl])
(at-modelCountry, [or-model, testarossa, or-country, italy, carrera,
germany, iso13250-AssociationType])
```

A topic object is characterized by all its attributes and its children's attributes. All *testarossa*'s properties are contained in the corresponding topic element or in its children. There is no need to study other parts of the topic map to find the properties of a topic object.

The properties of an association object reflect all topics linked by this association and the roles they play. The whole topic map needs to be parsed to find all topics connected by an association. It's not the same as topics, as some of the association's properties are not contained in the *assoc* element itself.

Recursive scenario

In the recursive scenario, all properties defined in the first scenario become objects themselves (even if they are neither topics nor associations). With this input, more concepts are generated by the classification algorithm. An object *http://www.ferrari.com* is created, although there is no corresponding topic or association.

The second modification is that properties of associations are added to the topics involved in these associations: we learn that *testarossa* is involved in the *at-modelCountry* and *at-modelOwner* associations, which we did not know in the previous scenario.

Associations' properties are unchanged.

```
(testarossa, [ferrari, vehicle, or-webSite, http://www.ferrari.com, at-
modelCountry, at-modelOwner])
(http://www.ferrari.com, [testarossa, spider])
```

```
(at-modelCountry, [or-model, testarossa, or-country, italy, carrera, germany, isol3250-AssociationType])
```

2.3. Lattice representation, navigation and simplification

Once the input generation scenario is chosen, the lattice is computed and visualized as a 3D interactive Hasse diagram. This graph can be translated, rotated in any direction so that the user can put interesting parts in the foreground. Two windows display objects and properties contained in the concept selected in the 3D representation, as shown on figure 4. This lattice was obtained from a topic map about The Clash available on Techquila Web page [Techquila].

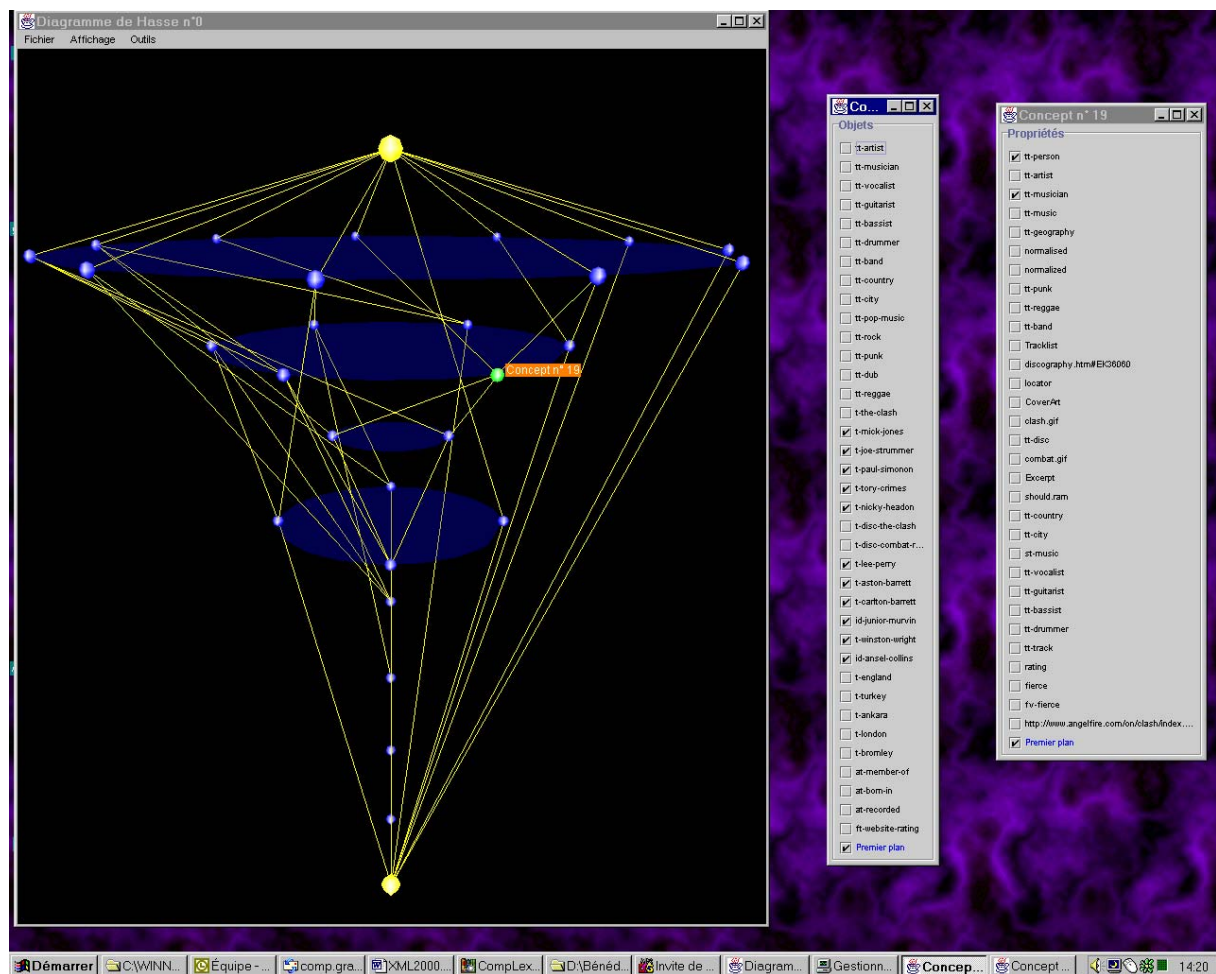


figure 4. Galois lattice 3D representation and navigation

When the mouse's pointer is on a node, this concept's parents and children are highlighted. This prototype was created with the Java3D API [Java3D].

As explained before, objects and properties are generated automatically, therefore some of them may not be relevant to the user. Therefore it is possible to select relevant objects and properties from the complete and automatically-generated list. Once this selection is done, the lattice is computed again with the new input.

This lattice allows us to study the main concepts in the topic map and the relations between them – some of them have nothing in common, others are connected. On figure 4, a given node is more specific than its parents and more general than its children.

However, these lattices need to be interpreted in more detail, so as to provide the information we need to help users navigate through topic maps.

3. Results interpretation

3.1. Associations analysis

In order to interpret the lattice, it is easier to consider topics and associations separately. First we focused on associations and considered lattices with associations objects only. We got very interesting results. By studying resulting lattices, we were able to distinguish different kinds of associations. Some of them have many commonalities with other associations – they are called *regular* - whereas others are *singular* (they belong to very few concepts of the lattice, which means they have very little in common with other associations).

Analysing associations allows us to deduce information about the topic map. If all associations have many commonalities, the topic map is very specialized. On the other hand, if there are many singular associations, the topic map is quite general as it deals with different subjects. Our conceptual analysis shows the *morphology* of topic maps.

After generating the lattice, we exploit the results by characterizing associations according to their regularity / singularity. This computation provides the user with a classification of associations. If the user is interested in a singular association, it is very easy to find the corresponding parts of the topic map. If only regular associations are relevant, singular ones can be pruned from the lattice, and consequently from the topic map.

An object is singular if there is a low similarity between this object and the all others. Let us explain how the similarity between two objects is computed. This similarity measure is a function which assigns a value of matching coefficient to a pair of vectors. Each vector included a set of attributes that characterize an object. Example of such attributes are the number of concepts this object appears in, the degree of commonality it has with other objects, the number of objects there is in each concept it belongs to, and finally its number of occurrences in the topic map. We have not decided yet which attributes will be used in vectors.

The matching coefficient techniques we chose is the Cosine measure method described in [Yu2000]. All attributes have different weights representing their importance. Let V1 and V2 two vectors symbolizing the objects of two Galois concepts.

$$V1 = \{(V1.attr1, V1.w1), (V1.attr2, V1.w2), \dots (V1.attrN, V1.wN)\}$$
$$V2 = \{(V2.attr1, V2.w1), (V2.attr2, V2.w2), \dots (V2.attrN, V2.wN)\} ,$$

where V1.attr1 is an attribute and V1.w1 its weight.

The similarity of two objects is defined as :

$$S_{1,2} = \frac{\sum(V_{1,wi} * V_{2,wi})}{\sqrt{\sum(V_{1,wi})^2} \cdot \sqrt{\sum(V_{2,wi})^2}}$$

Our similarity measures will be used to cluster objects and determine relevance of associations to users needs.

Lattice and topic map pruning

Once associations are categorized, it is possible to decide which classes are interesting or not. Irrelevant associations may be deleted from the topic maps, as well as topics which are involved only in these associations. This will significantly simplify the structure and will prevent users from losing time with irrelevant information.

Associations aggregation

An aggregation of topics and associations is essential if the topic map is very large. There is far too much information to display everything in a global view. It is not only impossible, because the screen would be cluttered, but also irrelevant. If the user does not know what he is looking for, he is not interested in details at the beginning. He needs to narrow his scope of research before. We can compare this with a geographic map ; a map of the world cannot be precise. If one is interested in details, one must precise his interest, for example choose a specific country. We need to provide different scales in topic maps visualization.

One of the problems in topic maps visualization is the very high number of dimensions. For example, there can be many different associations types and differentiating them on the screen is impossible. There are not enough different visual cues - icons, colours, shapes, textures – available. One way of solving this problem is to display similar associations the same way which reduces the number of different association types. The topic map becomes more general. Two associations are said to be similar if the measure of similarity between them is superior to a threshold. The degree of generality of the topic map can be adapted by adjusting this threshold.

3.2. Topics analysis

The similarity between topic objects can be measured the same way as the similarity between associations and similar conclusions can be drawn.

However, additional analyses can be done if we consider the lattice consisting of topic objects only. We find out that topics can be grouped in a concept whereas they are not associated with each other – by an association - in the topic map. On the other hand, some topics have no properties in common although they are semantically related by topic maps associations. Finally, some topics are grouped both by Galois concepts and topic maps associations. In this case, associations between these topics may be redundant. We are currently studying this issue to decide if these associations should be deleted form the topic map.

Conclusion

In this article, we presented our work about topic maps conceptual analysis. This technique is based on Galois lattices and provides a classification of topics and associations. The interpretation of results and the measure of similarity between concepts allows us to categorize topics and associations according to their singularity or regularity. Users may choose what classes they are interested in. Therefore it is possible to prune the topic map by deleting irrelevant topics and associations. Remaining objects can be aggregated if they are considered as similar. In the future, we will carry on implementing and testing this theory.

Bibliographie

[Bar70] Barbut M., Monjardet B., *Ordre et classification, Algebre et combinatoire*, Tome 2, Hachette, 1970

[Godin98] Godin, R., Chau T., *Comparaison d'algorithmes de construction de hierarchies de classes*, Rapport de recherche, Universite du Quebec, Montreal, 1998

[Infoloom] <http://www.infoloom.com/gcaconfs/WEB/index.htm>

[ISO13250] International Organization for Standardization, ISO/IEC 13250, *Information Technology-SGML Applications-Topic Maps*, Geneva: ISO

[Java3D] *Java3D API Specification* Version 1.1.2, June 1999-06-24, Sun Microsystems

[LeGrand2000] Le Grand B., Soto M., *Information Management - Topic Maps Visualization*, XML Europe 2000, Paris, France, June 2000

[Techquila] <http://www.techquila.com/tmsamples/index.htm>

[Wil92] Wille R., *Concept lattices and conceptual knowledge systems*, Computers and Mathematics Applications, 23, n 6-9, p; 493-515, 1992

[Yu2000] Yu H., Ghorbani A., Bhavsar V., Marsh S., *Keyphrase-Based Information Sharing in the ACORN Multi-agent Architecture*, Mobile Agents for Telecommunications Applications 2000, Paris, France, September 2000