

Intelligent Access to Data and Knowledge Bases via User's Topics of Interest *

Sylvie CAZALENS Robert DEMOLOMBE
ONERA/CERT

2, Av. Edouard BELIN

31055 TOULOUSE, France

{cazalens, demolomb}@tls-cs.cert.fr

tel: (...) +33 61-55-70-62

fax: (...) +33 61-55-71-72

June 17, 1997

Abstract

Retrieving relevant information in Data and Knowledge Bases containing a large number of different types of information is a non trivial problem. That is the reason why, in areas like Decision Support Systems, or Computer Aided Design, users need the help of Intelligent Systems.

Such Intelligent Systems must be able to provide the users with useful information which is not explicitly requested. For this purpose, we consider a method where the user's needs are represented in terms of their topics of interest. This kind of user representation is integrated in an appropriate Modal logic, that allows the system to reason about the user's representation, the user's query and the DKB content, in order to determine the useful information.

The main contribution of the paper is to present a first attempt to the formalization of the concepts of topic, interesting topic and interesting piece of information. Several criteria are investigated for these definitions, in particular to define a partial order on interesting pieces of information.

*This work has been partially supported by the EEC, in the context of the Basic Research Action, called MEDLAR.

1 Introduction.

The aim of an Intelligent System when accessing large Data and Knowledge Bases (DKB) is to help users to retrieve **useful** information. In many cases, a user's query characterizes only part of the useful information for two possible reasons: first, in Data and Knowledge Bases containing hundreds of relations, there are so many different types of information that the user may ignore the existence of some of them. Second, it is generally too boring for users to entirely describe all the useful information. In these cases, the system must be able to reason about the description of a given situation defined by: 1) a query or a dialogue, 2) a user representation, and 3) a DKB content, in order to determine the useful information.

To define such reasoning methods we have to make precise what we mean by useful information. Roughly speaking, useful information is information that is needed to reach a given goal, or information that helps the user to select or make more precise a goal, or information that is logically related to the query. We briefly sketch three kinds of reasoning methods associated to these three definitions of useful information.

- **The plans and goals method** [1, 2, 7, 8]: plans are sequences of actions representing users' typical behaviours; they are also called scripts or frames in the literature [?, ?]. The achievement of a plan enables the user to reach a given goal. Generally, the actions of the plan have preconditions which have to be satisfied for the action to be performed. Some preconditions are to know given pieces of information, which are the useful information. For example, in order to take a train, one has to buy a ticket and go to the platform. Hence, he has to know the ticket price, the platform number and the departure time. So, by reasoning from the plans and the user's goal, the system can derive information to be provided to the user.
- **The topics of interest method** [11, 5, 9, 10]: When the system fails to recognize what the user's goal is, it may be the case that, in fact, the user has no precise goal in mind. In such a situation, if the system can determine what the user's topics of interest are, it can provide him with information related to these topics. This information may help the user to refine his goal definition. For example, if a user wants to know the price of a return ticket to New-York, then the system can infer that expenses are one of his topics of interest, and it can also provide the user with the possible special fares.

- **The logical links method:** in that case, useful information may be the information that explains why a direct answer holds. For example if the direct answer is “there is no flight to Milano this morning”, the explanation may be: “because there is a strike” or “because there is fog”. Useful information may also be the information one has to make sure by himself, when querying an incomplete DKB, in order to get a direct answer. An example of such conditional answer could be: “there is a flight to Milano today if there is no fog”. In general, useful information, in the context of a query q , is of the form: $q \leftarrow p$ and p , for explanations, or: $q \leftarrow p$, for conditional answers.

These three methods are complementary, and it would be very convenient for an intelligent system to have a formal description of each of them in the same formalism. There already are several proposals to formalize the Plans and Goals method in modal logic (Epistemic and Dynamic modal logic), [7, 15, 13, 8]. However, as far as we know, there is no formalization of the Topics of Interest method in the same modal framework. The objective of this paper is to contribute to define such a formalization. We do not consider here any implementation, or efficiency aspects.

In the following, we first recall what the topics of interest approach consists in, and we show where the problems of formalization are. Then we show how the topics, the topics of interest, and the interesting pieces of information are defined in this approach.

2 The Topics of Interest Method

Several works [11, 5, 9, 10] have already underlined how much it can be useful to take into account the user’s topics of interest, when answering cooperatively to a user’s question. The main ideas are summerized in the following:

The user’s topics of interest are one of the components of the user’s model, together with his beliefs, goals, and may be other information. Roughly speaking, the topics are used to represent information about the meaning of words and sentences used in the description of the world. Of course, they capture only part of the meaning. In a Data and Knowledge Base context they mainly inform about the meaning of the predicates and formulas. For example in Figure 1. , the predicates *departure-time* and *arrival-time* belong to the topic **timetable**, while *frequency* and *validity* belong to the topic **condition of validity**.

Figure 1: Links between predicates and topics.

They can be used in a very simple way : if a user asks a question about the *departure-time* of a given flight, then he may also be interested by the *arrival-time* of the same flight. Note also that the topics can be organized into a hierarchy. For example (see Figure 2), the topics **timetable** and **condition of validity** belong to the more general topic **time**. Within this hierarchy, the links are of type “is-a”.

Figure 2: A hierarchy of topics.

It is worth noting that the link between a sentence and a topic is independent of the truth value of the sentence. For example, the fact that “the predicate *Departure-time* is related to the topic **Timetable**”, is independent of the truth value of the sentence “the *Departure-time* of flight AF001 is 1 pm”. The first sentence is about the meaning of the predicate; it is

linguistic information, while the second one describes the real world. The hierarchy of topics is also independent of any truth value. This is the basic reason why the concept of topic is not easy to formalize in standard logic. Indeed in standard logic two formulas having the same truth value in every world state are equivalent, and they can be substituted one each other, even if their meanings are completely different. However, as it is said in the introduction we want to adopt a common theoretical framework with the Plans and Goals method, to keep the complementarity of the two approaches. For this reason we have tried to integrate the concepts of topics and interesting topics into an Epistemic Logic.

3 Formalization of the topics

The Epistemic Logic semantics is defined in terms of possible worlds. Possible worlds represent different states of the world compatible with the user's beliefs, or in general with an agent's beliefs. It is a non trivial issue to represent topics, and topics of interest in this framework. Indeed, the links between topics and formulas, and the topic hierarchy are not part of the application domain description. They are linguistic information and then they are independant of what is true in a given world. Hence, we have to adapt standard Modal Logic to represent this information.

Another issue is to define the structure of the sets of formulas related to a given topic. These sets of formulas are called topic extensions.

3.1 Topic Extensions

In the simplest approach, we can define no structure at all on a given topic extension. In that case, topic extensions should be explicitly defined by sets of formulas. That would be extremely heavy, and even impossible if the set is infinite.

A more realistic approach is to define a structure on a topic extension. We have defined two kinds of structures where a set of primitive propositions (not necessarily atomic formulas) is explicitly given for each topic. In addition, general rules are defined to characterize, for each topic, how the overall topic extension is defined from the set of primitive propositions. We have considered two kinds of general rules:

- The first one is: **a formula F belongs to topic T iff each subformula of F belongs to T , or is a primitive proposition.** Topic extensions defined by this rule have a structure similar to that of

“production fields” defined by Siegel and Bossu in [4, 3]. It has also been adopted by Inoue in [14]. The constraints imposed by this rule are quite strong. For example, if in a formula like $A \vee B$, formula A belongs to T , but formula B does not belong to T , then $A \vee B$ does not belong to T . The consequence is that if we want to retrieve all the formulas related to the topic T , we do not get $A \vee B$.

- The second one is: **a formula F belongs to topic T iff at least one of its subformulas belongs to T** . According to this definition, the formula $A \vee B$ of the previous example belongs to T . In our mind, the justification for this rule can be found in the analogy between topics and key-words in the area of Information Retrieval. Indeed, in Information Retrieval, if a key-word is associated to a chapter of some book, or to a section of a paper, then, it is associated to the overall book, or to the overall paper. Here we consider that subformulas play a similar role as chapters or sections in a document.

3.2 Topic Hierarchy

Topic extensions can also be structured via a structure on topics themselves. If a topic $T1$ has a more general meaning than $T2$, then $T1$ extension must contain $T2$ extension. This corresponds to the general idea of structuring object types via an “is-a” relation. In the following, we adopt the notation: $T2 \prec T1$ to represent the fact that $T1$ is more general than $T2$.

In this section, we have presented how topics can be formally represented; in the next section, we shall see how topics of interest are integrated in the formalism of Epistemic Logic.

4 Topics of Interest and Interesting Formulas.

In a given situation, a user is interested by information related to some particular topics. These topics are called the user’s topics of interest. There are different possible methods to determine what the interesting topics are: for example, in [9], the interesting topics are defined as the topics related to some predicate in a user’s query; we can also imagine that the system directly asks the user what his topics of interest are. However it is not the purpose of this paper to present methods for acquiring this information,

and we assume that the system knows what the current topics of interest are.

Our idea in formalizing topics of interest and interesting formulas, was to take inspiration from the formalization of the Logic of General Awareness by Fagin and Halpern in [12]. Indeed the concepts of interest and awareness share an important common feature which is their independence with respect to the truth value of the sentences. The fact that one is aware of sentence p does not depend of the truth value of p , and, in the same way, the fact that one is interested by p does not depend on the truth value of p . In the following, we first recall what the Logic of the General Awareness is, and then present our formalization. We assume that the reader is familiar with basic notions of Modal Logic (as explained for example in [6]).

4.1 The Logic of General Awareness

This logic was defined to get rid of the omniscience problem: it is not realistic to accept that a user, or an agent in general, is able to derive **all** the logical consequences of an explicit set of formulas. To avoid this problem Halpern and Fagin introduced a distinction between explicit beliefs (represented by the modality B), and implicit beliefs (represented by the modality L). Explicit beliefs are the subset of implicit beliefs the agent is **aware of**.

A Krikpe model for General Awareness is a tuple:

$$M = (S, \pi, \mathcal{A}, \mathcal{B})$$

where S is a set of states, $\pi(s,p)$ is a truth assignment function for each state s and each primitive proposition p , \mathcal{B} is a relation on $S \times S$. \mathcal{B} is serial, transitive and euclidean and represents the accessibility relation for the agent's beliefs. \mathcal{A} is a function associating to each state s an arbitrary set of formulas $\mathcal{A}(s)$, representing the formulas the agent is aware of at state s . The truth relation is defined by:

$$M,s \models true,$$

$$M,s \models p \text{ where } p \text{ is a primitive proposition, iff } \pi(s,p) = true$$

$$M,s \models \neg\varphi \text{ iff } M,s \not\models \varphi$$

$$M,s \models \alpha \wedge \beta \text{ iff } M,s \models \alpha \text{ and } M,s \models \beta$$

$$M,s \models A\varphi \text{ iff } \varphi \in \mathcal{A}(s)$$

$$M,s \models L\varphi \text{ iff } M,t \models \varphi \text{ for all } t \text{ such that } (s,t) \text{ belongs to } \mathcal{B}$$

$$M,s \models B\varphi \text{ iff } \varphi \text{ belongs to } \mathcal{A}(s) \text{ and } M,s \models L\varphi.$$

In [12], Fagin and Halpern consider different possible structures for the sets of formulas $\mathcal{A}(s)$.

4.2 Formalization of the Interesting Formulas

The formalization of the interesting formulas is based on similar ideas. A Kripke model capturing the concept of Interest is a tuple:

$$M = (S, \pi, \mathcal{T}, \mathcal{B})$$

where S is a set of states, $\pi(s,p)$ is a truth assignment function for each state and each primitive proposition p , \mathcal{B} is a relation on $S \times S$. \mathcal{B} is serial, transitive and euclidean. It represents the accessibility relation for the user's beliefs. \mathcal{T} is a function associating to each state s , a set of topics $\mathcal{T}(s)$, representing **the current topics of interest at state s** . We have considered that the set of topics a user is interested in may depend on what he has in mind, that is, it may depend on his belief state. As for $\mathcal{A}(s)$, the set of current topics of interest can be any set of topics, or a structured one. We have adopted the very intuitive structure defined by the rule: **if a topic T is a current topic of interest, then all the topics which are more specific than T are also current topics of interest**. That means that the interest is inherited in the hierarchy of topics.

The truth relation for the logical connectives and the the beliefs is defined as usual. We abandon the differentiation between explicit and implicit belief which is no more relevant here.

$$M, s \models true,$$

$$M, s \models p \text{ where } p \text{ is a primitive proposition, iff } \pi(s,p) = true$$

$$M, s \models \neg\varphi \text{ iff } M, s \not\models \varphi$$

$$M, s \models \alpha \wedge \beta \text{ iff } M, s \models \alpha \text{ and } M, s \models \beta$$

$$M, s \models B\varphi \text{ iff } M, t \models \varphi \text{ for all } t \text{ such that } (s,t) \text{ belongs to } \mathcal{B}.$$

We introduce the notation $I\varphi$ to represent that φ is an interesting formula. There are several possible variants in the formal definition of $I\varphi$, i.e. to define:

$$M, s \models I\varphi$$

All of the following definitions of $I\varphi$ at state s refer to $\mathcal{T}(s)$. We assume that $\mathcal{T}(s)$ gives a complete description of the user's state of mind with regard to his topics of interest at state s .

1. First definition : “ φ is an interesting formula” iff at least one of the current topics of interest contains φ :

$$M, s \models I\varphi \text{ iff } \exists T (T \in \mathcal{T}(s) \text{ and } \varphi \in T)$$

According to this definition, any formula in the extension of a topic of interest is an interesting formula. Of course this definition may

lead to a too large set of interesting formulas. If that is the case, to remove this drawback, we can define a partial order on the interesting formulas, as presented in the following section, in order to select the most interesting formulas

2. Another alternative definition is to make explicit the topic with regard to which the formula is interesting. That leads to slightly change the syntax, and incorporate the name of the topic in the operator I, using the notation $I_T\varphi$. In that case, the definition is:

$$M, s \models I_T\varphi \text{ iff } \varphi \in T \text{ and } T \in \mathcal{T}(s)$$

Because of the structure of $\mathcal{T}(s)$, we can deduce the following rule, which reflects a kind of inheritance:

$$M, s \models I_T\varphi \text{ if } \varphi \in T \text{ and } T \prec T' \text{ and } M, s \models I_{T'}\varphi.$$

3. A more constraining definition would be to impose that interesting formulas belong to all the current topics of interest extensions. We can easily see that this definition would not be sensible because in many cases the intersection of all the current topics is an empty set.

Comments : in the above formalization, the current topics of interest in two different possible worlds are independently defined, even if one world is accessible from another one by the accessibility relation representing the beliefs. In reality, it may happen that topics of interest in a mental state depend on the topics of interest in the previous mental state. We are perfectly aware of this over simplification, and this issue will need further investigations.

Another simplification comes from the fact this definition of $I\varphi$ intuitively says that φ is interesting if φ is related to an interesting topic. However, in practice, formulas like: $p \wedge p$, or: $p \wedge \neg p$, are interesting for no user. Hence, we will need to refine the definition of $I\varphi$ by using logical criteria, for example, removing $I\varphi$ if φ is a tautology or a contradiction, or if φ is not minimal in some sense or by using pragmatic criteria, for example, removing formulas which are not in some normal form which is more easy to understand for users.

4.3 Ordering Interesting Formulas

The set of interesting formulas generated by the topics of interest method may be too large, and it can be useful to have criteria to order formulas in function of their interest. This allows to provide the users with the

most interesting formulas in a first step. In further steps, if the user wants to acquire more information, he can interactively request the formulas of lower interest.

Here we consider several criteria to order formulas.

4.3.1 Syntactic Criteria

We call syntactic criteria, criteria which are defined in function of structural properties, independently of an application domain. The following criteria seem to be natural.

- If formula $F1$ can be fully decomposed into subformulas which are primitive propositions belonging to a topic of interest $T1$, and if that property does not hold for formula $F2$, then $F1$ is said to be more interesting than $F2$. The intuition is that formula $F1$ is preferred to formula $F2$ because it is “fully” relevant to $T1$.
- Let $T1$ and $T2$ be two topics of interest, such that $T1 \prec T2$, i.e. $T1$ is more specific than $T2$. Let $F1$ be a formula in the extension of $T1$, and $F2$ a formula in the extension of $T2$. If $F2$ does not belong to the extension of $T1$, then $F1$ is more interesting than $F2$. The intuition is that the more specific is the interesting topic to which a formula belongs, the more interesting is this formula.
- Let $F1$ and $F2$ be two formulas. Let $n1$ (respectively $n2$) be the number of topics of interest to which $F1$ (resp. $F2$) belongs. Then formula $F1$ is more interesting than formula $F2$ if $n1 > n2$. The intuition is that the more important is the number of topics of interest to which the formula belongs, the more interesting is this formula.

4.3.2 Semantic Criteria

Ordering criteria can also be based on the application domain. For example in the user representation, we can define a partial order on the topics which reflects the user’s preferences.

A simple ordering criterion can be defined as follows : formula $F1$ is more interesting than formula $F2$, if $F1$ belongs to a topic $T1$, and $F2$ belongs to a topic $T2$, and $T1$ is greater than $T2$, according to the given partial order. Though this definition seems to be intuitive, it is in fact confusing because formulas can belong to several topics. For example, it may be the case that $F1$ also belongs to $T'1$, $F2$ also belongs to $T'2$, and

T'1 is lower than T'2. In such a case it is no longer clear which of the two formulas should be preferred.

In fact, the partial order on the topics can be viewed as a sort of preference, as defined by Shoham in [16]. The problem of combining preferences has still been extensively investigated in other areas, and it is well known that it is not a trivial issue to define intuitive composition rules between preferences. That will also need further investigations.

5 Conclusion.

We have recalled at the beginning three possible methods which can be implemented in an intelligent system to help users to retrieve information. Then we focused on the method based on the topics of interest. The main contribution of this paper is to propose a first attempt to formalize the notions of topic, topic of interest and interesting formula, in a variant of the Epistemic logic, which takes inspiration from the logic of General Awareness [12].

Issues requiring more investigations have been explicitly mentioned in the paper. Another important issue will be to define a Proof Theory associated to the Model Theory. The Model Theory we presented, gives a precise definition for the new concepts, but a Proof Theory is needed for computational objectives.

We think the concept of topic can have many applications for Intelligent Information Retrieval in its broad sense, because it gives a more abstract description of stored information than predicates do, and because terms used for topic definitions come from natural language and are easier to understand by any user. Possible applications are, for example, homogeneous description of distributed databases which have been designed independently. Another application is to provide a uniform description of multimedia database contents. Indeed, topics can be used as well to describe document contents or pictures, and they can be viewed as a sophisticated extension of the concepts of keywords and thesaurus, combined with automated reasoning techniques.

References

- [1] J. F. Allen. Speech acts. In S. C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, volume 2, pages 1062–1065. Interscience, 1987.

- [2] J. F. Allen and C. R. Perrault. Analysing intention in utterance. *Artificial Intelligence*, 15(3):143–178, 1980.
- [3] J.M. Boi, E. Innocente, A. Rauzy, and P. Siegel. Production fields : A new approach to deduction problems and two algorithms for propositional calculus. *The Journal of Artificial Intelligence*, To appear, 1991.
- [4] G. Bossu and P. Siegel. Saturation, Non-monotonic Reasoning, and the Closed World Assumption. *Artificial Intelligence*, 25, 1985.
- [5] S. Cazalens. A Cooperative Answering Real Toy Example. In *Milestone I Deliverable, ESPRIT 3125 (MEDLAR)*, 1990.
- [6] B. F. Chellas. *Modal Logic: An introduction*. Cambridge University Press, 1988.
- [7] P. R. Cohen and H. J. Levesque. Persistence, Intention, and Commitment. In A. L. Lansky M. P. Georgeff, editor, *Reasoning about actions and plans*, pages 297–340, Timberline, USA, 1986.
- [8] P.R. Cohen, J. Morgan, and M.Pollack. *Intentions in Communication*. The MIT Press, 1990.
- [9] F. Cuppens and R. Demolombe. Cooperative Answering: a methodology to provide intelligent access to Databases. In *Proc. of 2d Int. Conf. on Expert Database Systems*, Tysons Corner, Virginia, 1988.
- [10] F. Cuppens and R. Demolombe. How to recognize interesting topics to provide cooperative answering. *Information Systems*, 14(2), 1989.
- [11] R. Demolombe. Cooperative Access to Data and Knowledge Bases (Tutorial). In *Proc. of 17th Int. Conf. on Very Large Data Base Conference*, Barcelona, 1991.
- [12] R. Fagin and J. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.
- [13] M.P. Georgeff and A.N. Rao. Modeling Rational Agents within a BDI Architecture. In *Proc. 2nd Int. Conf. on Knowledge Representation and Reasoning*. Morgan Kaufmann, 1991.
- [14] K. Inoue. Consequence-Finding Based on Oredered Linear Resolution. In *Proc. of International Joint Conference on Artificial Intelligence*, Sydney, 1991.

- [15] C. Seguin. Une étude logique de la coopération dans les systèmes multi-agents. Rapport de dea, Université de Toulouse (IRIT), 1988.
- [16] Y. Shoham. *Reasoning about change*. The MIT Press, 1988.