An Informational Model for Cooperative Information Gathering

Hassan TOUT
University of Toulouse 1, Place Anatole France
31042 Toulouse Cedex
33 5 61 63 35 60, tout@univ-tlse1.fr

ABSTRACT
A Cooperative Information Gathering system may be abstracted as three communicating models: an informational model (IM), an organizational model (OM) and a task model (TM). This paper focuses on the informational model describing the Universe of Discourse (UoD) and the users' subjects of interest. It presents a new way of looking at ontology as a super-topic. Then, ontology in our IM is not used only to help a CIG system to deal with the heterogeneities of information sources, but also to help it to capture the user's subjects of interest.

Keywords
Ontology, Topic, context, Cooperative Information Gathering.

1. INTRODUCTION
Cooperative Information Gathering (CIG) is a paradigm which considers information retrieval as a distributed problem solving [14] carried out by intelligent and cooperating Agents. The complexity of this area is mainly due to three factors:

- *Information Sources nature*: the information sources involved in a CIG system and accessed by agents are inherently heterogeneous, distributed, autonomous and dynamic.
- *Tasks complexity*: a CIG task is not a simple database query but rather a problem that must be decomposed into inter-dependant tasks, each of which must be, in turn, intelligently allocated to appropriate agents thanks to middle agents (e.g Matchmakers, Brokers [10]). Tasks complexity is also due to the necessity of involving users during the process. Users define and refine their questions, reformulate them and validate the final results. To cooperate with users, the system must model users' preferences including their topics of interest.
- *Organizational complexity*: the diverse resources (information sources and middle agents) involved in a CIG system may belong to different organizations, each one imposing its own rules and communication protocols. To adequately exploit resources, organizations and their features (locality, capacity, quality of service, protocols) must be described to be retrieved and exploited at run time.

One possible way to deal with this complexity, -and to ease CIG system's understandability, design and implementation-, is to use a well-known software engineering principle: the separation of concerns [6]. Separation of concerns aims at decomposing a system in communicating sub-systems, each one corresponding to a relevant abstraction. Each abstraction requires a model to be structured and described. Following this practice, a CIG system may be thought as three communicating models: an organizational model, an informational model, and a task model. They are described below.

The Organizational model (OM) has two roles. First it structures resources in class of agents sharing the same organization structure. A class is called a *role* when it comprises agents having the same capabilities, and *organizational unit* for agents belonging to a same organization structure. This requires in particular the definition of a Capabilities Description Language (see for example LARKS [Sycara, 99]). Second, the organizational model attributes, to each resource, authorization to perform a task or a sub-task. Roles and organizational units are abstraction that can be used to define a CIG task without referring explicitly to individual agents but rather to the quality they must have. In real time, a CIG system assigns sub-tasks to individual agents, possibly selected by middle agents. Therefore, change of individuals has no consequence for the task definition.

The Informational Model (*IM*) describes the structure of two types of information. It first describes the semantic structure of the Universe of Discourse (UoD) i.e. the structure of the domain(s) being covered by the CIG system. This abstract and common model avoids to describe the contents of the different Information Sources which moreover are numerous, not known a priori, and evolutives. Second, the IM describes each *user's subjects of interest* i.e. the information items he/she needs in varied contexts. A User subject of interest is a view of the UoD. *A context* is a computer, an environmental or a user situation which influences the way a subject of interest is explored at run time. Hence, the response to a query may have variable forms (structure, volume, …) depending on user's preferences and profiles, his present intentions and situations (hurried…). For example, additional information –not explicitly asked in the query- may be provided to a user if the system has detected this need through its interaction with the user. At the opposite, to avoid information overload, the system may exploit and reason on meta-information [9] and provides synthetic information instead of too numerous information. Exceptional situations (unavailability of Information Sources, inefficient network, …), also, may occur and require user guidance. In other words, a context provides the user with dynamic views on the UoD.

A Task model describes the structure of each task: decomposition in sub-tasks, inter-dependence between them, and their coordination. This model refers to both the organizational model, which defines and organizes the set of potential performer agents, and the information model, which determines the objects to be retrieved and processed.

Objectives of the paper

This paper discusses the need of ontologies in an informational model of a CIG system. It follows how ontologies help a CIG system to deal with semantic and structural heterogeneities of information sources in the UoD. It also follows how to build

ontologies using topics and how relations between attributes of various topics and ontologies can help the CIG system to capture users subjects of interests.

Organization of the paper

The organization of the rest of this paper is as follows. Section 2 presents the concept of *topic* to capture users subjects of interest. Topics are the pivotal concept since it makes up Ontology and allows the connection with the task and the organizational models. Section 3 defines the *ontology concept*, justifies the necessity of using multiple ontologies as a common model in the context of CIG and presents ontology as a super-topic.

2. Topics to capture users subjects of interest

2.1 Topic's definition

In [5], *a topic* is defined as a general term related to the signification of a set of expressions. In our context, we define a topic as a set of information or attributes belonging to a subject. Then, we can define *a user topic of interest* as a set of topics that interest the user.

To represent a hierarchy of subjects, a topic can include others. A topic that doesn't include others is called *an elementary topic* and contains only closely related attributes. A user interested by an elementary topic T is supposed to be interested by all the attributes composing T.

Example of elementary topic: in the context of Vacation organization, "departure time", "arrival time", "departure date" and "arrival date" belong to a same elementary topic " Flight Schedule ". Indeed, when a user is interested by a flight's departure time, he is inevitably interested by the flight's departure date, arrival date, and arrival time.

A topic can belong to various ontologies, and an ontology may be considered as a super Topic (as detailed in section 3). A topic can be represented differently in various ontologies.

2.2 Users' topics of interest

A User topic of interest is a set of topics that interest the user in a given context. To identify topics that interest a given user we must take into consideration the user's question, the user's preferences, the environment and the computing features. Section 4 presents how current situations can help to determine a user's topic of interest.

The information relevant for a user in a given context are not restricted to a topic but may belong to several related topics. The set of topics belonging to a user topic of interest are linked by a relation called Relation Inter-Topics (RT). A RT link connects 2 attributes A and B from two different elementary topics T1 and T2 according to the following syntax: RT(T1.A:T2.B) which may be interpreted as "when a user U is interested by the topic T1 (called departure topic of RT), it may be interested by the topic T2 (called arrival topic of RT) on condition that the value of the attribute A be imputed to the attribute B. To be valid, arrival and departure attributes (B and A in the example) must have the same structure.

For example, the following fact "a user who wants to go to a town W can be interested by the weather in W", can be expressed by a RT link that links attribute "town-name" in the topic *means of transport* to the attribute "town-name" in the topic *weather*.

A CIG system should navigate through the RT links to provide interesting information to the users according to the context.

3. Ontology to deal with heterogeneity

The autonomy of information sources is one of the main issues of the CIG field. "Autonomy is the absence of a common control" [2] over the various information sources. It exists at different levels including:

1. the Service level. Each Information source decides by itself what services to offer, how to offer these services (how cooperation is established), and to whom.
2. the Design level: Information sources are built independently with different query languages, data structures and semantics.

To deal with Service Autonomy, a CIG system must record information to select adequately the relevant IS at run time. An Agent Capabilities Language is required to express this information in the organizational model.

Design autonomy induces semantic and structural heterogeneities of information sources and query languages.

To deal with the problem of *query language heterogeneity* recent works in the CIG field (like TSIMMIS [4]) exploit *translators*. A translator is associated with an information source and converts received queries into requests understood and executable by the information source.

Semantic heterogeneity occurs when the same information is represented by different expressions in various sources (synonyms), or when an expression is used in various sources to represent two different information (homonyms). To deal with semantic and structural heterogeneity problems, the construction of *a common model* to the global system is indispensable. Thus, relations between terms in the common model and terms in different information sources help the system to identify semantic and structural relations between terms in various information sources.

Given the dynamic and the number of information sources, the common model should describe the UoD (domains treated by the system) rather than the structure of the information sources themselves. Doing so, we don't have to update the common model when information sources structures are updated. Only relations between terms in the common model and terms in the different information sources have to be updated.

In this article, we use the ontology concept to construct the common model. Here, we consider ontology as a *"specification of conceptualization*" [7] and conceptualization as "*a set of concepts, relationships, objects, and constraints that define a semantic model of a subject of interest*" [8].

In our context, where the modeled universe may be huge and open, we also need to manage several domains, and consequently we need ontology mechanisms enabling semantic interoperability between information belonging to different domains. To achieve this goal, two approaches are possible:

1. The use of a common global ontology to several domains (e.g. SIMS [1], IM [11], InfoSleuth [3]). Using a global

ontology could facilitate the task of integration and semantic interoperation across different sources. In the other hand, it is very difficult to create and manage such an ontology [12] and also to use it because of the huge number of concepts and relations between concepts it contains.

2. The use of multiple ontologies (e.g. OBSERVER [12]): in such an approach, several ontologies are used, each one being specific to a domain. This solution needs also to represent the semantic relations that can exist between concepts from different ontologies. This solution is more realistic than the use of a global one since it is more easier to build (modularity, reusability) and to manage. Besides, it is easier to use since we can focus on the main ontology, corresponding for example to the user's subject of interest, and then navigate to the other ontologies if needed.

In our informational model we have chosen the second solution that obviously has more advantages.

3.1 Ontologies representation formalism

Many works in this area have chosen logical languages to represent ontologies and to specify relationships between ontological terms. As an example we quote the description logic used in OBSERVER [12] to construct ontologies.

Other works have chosen graphical representations to represent both semantic and structural relations between ontological terms in a simple manner, easy to be understood by the user and to be updated by an expert.

In this article we propose a graphical representation model inspired from the ONION system model [13]. In ONION an ontology is represented by a directed labeled graph where nodes represent ontological terms and an arc between two nodes has as label either a verb in natural language or a pre-defined semantic relationships.

The four semantic relationships defined in [13] are: "InstanceOf" (arcs with the label I in figure 1), "SubclassOf" (arcs with the label S in figure 1), "AttributeOf" (arcs with the label A in figure 1), and "Semantic Implication" (arcs with the label SI in figure 1). The first three relationships can relate nodes within a same ontology, when "Semantic Implication" relates nodes from different ontologies (domains). Term1 is related to Term2 by this relationships means that Term2 is a subset of Term1.

In addition to these relations, we have proposed the Inter-Topics Relation (RT) already described in section 2 and which enables to navigate through topics inside ontologies.

In a given domain, the topics and their links constitute a sub-graph of the corresponding ontology. Example: *means of transport* is a topic of the ontology "vacation organization". Thus, an ontology is a super-topic that include topics.

Finally, our multiple ontologies model can be viewed as an agreement on terms signification and terms structure between the user and all information sources in the system.

4. REFERENCES

[1] Y. Arens et al., "Query reformulation for dynamic information integration". Journal of Intelligent Information Systems, 6(2-3):99-130, 1996.

[2] Paolo Atzeni et al., "Database cooperation: classification and middleware tools". *Journal of Database Management, 2000*.

[3] R. Bayardo et al., "Infosleuth: Agent-Based semantic integration of information in open and dynamic environments", In Proceedings of the 1997 ACM International Conference on the Management of Data (SIGMOD), Tucson, Arizona, May 1997.

[4] S. Bergamaschi. Cooperative Information Agent, "Extraction of Information Highly Heterogeneous Source of Textual Data", First International Workshop, CIA'97. February 1997, Kiel, Allemagne.

[5] S. Cazalens "Formalisation en logique non standard de certaines méthodes de raisonnement pour fournir des reponses cooperatives, dans des systèmes de bases de données et de connaissances", PhD thesis Rapport, Université Paul Sabatier, Toulouse, 1992.

[6] C. Ghezzi, M. Jazayeri, D. Mandrioli. Fundamentals of Software Engineering. Prentice-Hall International Editions, 1991.

[7] T.R. Gruber, "A translation approach to portable ontologies", *Knowledge Acquisition*, 5(2) : 199-220, 1993.

[8] P.D. Karp et al., " XOL : An XML-Based Ontology Exchange Language", February 2000. http://www.ai.sri.com/cgi-bin/pubs/list_document_object.pl?doc_uri=/pubs/technotes/aic-tn-1999:559/.

[9] V. Kashyap et al., "Semantic Heterogeneity in Global Information Systems: the Role of Metadata, Context and Ontologies, in Cooperative Information Systems Trends and Directions, ACADEMIC PRESS, 1998.

[10] Matthias Klusch, introduction of the book "Intelligent Information Agents", Springer-Verlag Berlin Heidelberg 1999.

[11] A.Y. Levy et al., "Data model and query evaluation in global information systems", Journal of Intelligent Information Systems, 5(2):121-143, 1995.

[12] E. Mena et al., "Observer : An Approach for Query processing in Global Information Systems based on Interoperation across Pre-existing Ontologies", *Distributed and Parallel Databases Journal*,1999.

[13] P. Mitra et al. "A Graph-Oriented Model for Articulation of Ontology Interdependencies", VII Conference on Extending Database technology, Konstanz-Germany, 2000.

[14] T. Oates, M. V. N. Prasad, V. Lesser, "Cooperative Information Gathering: A Distributed Problem Solving Approach", UMass Computer Science Technical Report 94-66-version 2.