A Model for Graded Levels of Generalizations in Intensional Query Answering

Farah Benamara
IRIT, 118 route de Narbonne
31062 Toulouse, France.
benamara@irit.fr

Abstract

We present in this paper a model for graded levels of generalizations, within a cooperative question-answering framework. We describe how intensional answers descriptions can be generated when the set of extensional answers set, for a given natural language question, is very large. We develop a variable-depth intensional calculus that allows for the generation of intensional responses at the best level of abstraction.

Introduction

Knowledge discovery aims at the exploration and the analysis of large quantities of data in order to discover meaningful patterns and rules that better organize the data. Several methods for the discovery of various kinds of knowledge, including association, characterization and classification, have been proposed in the context of relational database systems (Park, Chen, & Yu 1995) (Hah, Cai, & Cercone 1993) (Piatetsky-Shapiro 1991). These techniques have been applied in the areas of decision support and market strategies. In this paper, we show how these techniques can be used to provide intensional answers, using logical inferences, within an intelligent question-answering framework.

Data generalization, statistics summarization and generalized rule extraction are essential techniques for intelligent query answering. These techniques, generally called intensional query answering (IQA), tend to abstract over possibly long enumerations of extensional answers in order to provide responses, in general at the highest level of abstraction. In addition, the intensional character of these answers may give hints to users about the structure of the knowledge base and of the domain knowledge and can help to clear up misconceptions. There are many ways to provide intensional answers, let us cite, for example :

1. *introduction of higher level concepts* (Han *et al.* 1996) in the answer (Yoon & Park 1999) is a simple substitution of some low level data in the answer by corresponding super concepts, at an appropriate level w.r.t. the user model, based on the query intent analysis. For example, instead of responding ”Tom is a junior student born in Vancouver on July 1977”, the following InR can be given ”*Tom is an* **undergraduate** *student, born in* **Canada** *in* **1977**”.
2. *data reorganization* aims at e.g. sorting the extensional answer set according to specific criteria generally inferred from question elements. An InR for the question ”give me hotels by the sea side” can be ”*list of hotels* **sorted according to the their distance w.r.t the sea**”.
3. *generalization/exception* (Motro 1994) can be realized by mechanisms that infer generalizers from concepts or rules that better summarize the given answer set. A term, judged as particularly relevant can be chosen as generalizer, even if its extension is larger than the answer set. Then, exceptions, provided they are limited, can be listed. The generalization can be done using statistical information (e.g. aggregation (Shum & Muntz 1988)) or by choosing higher concepts in an ontology (Yoon & Park 1999). For example, a possible InR for: ”which country can I visit without any visa formalities” is : *you can visit* **all the countries of the EEC except the UK and Norway**.

Most of the previous studies on IQA focused on generating intensional responses (InR) at a single level of abstraction using integrity constraints and/or rules of the database without any access to the extensional answer set (see (Motro 1994) for a general overview).

Our approach has substancial differences with these ones: the set of potential generalizers is directly derived via an intensional calculus from the set of direct responses to the question. Recently, (Yoon & Park 1999) used a similar approach by applying data mining techniques for IQA at multiple levels of abstraction in a relational database environment. The originality of our work lies around two major points : (1) the use of the question focus (extracted while parsing the NL question) paired with a rich ontological description of the domain. The goal is to select and rank the set of relevant concepts to be generalized, and (2) the definition of a *variable-depth intensional calculus* paired with a cooperative know how component that determines, via a supervisor, the best level of abstraction for the response, using a conceptual metrics.

In the following sections, we present an algorithm for the discovery of InR at different levels of abstractions using a domain ontology and the question focus. We concentrate in this paper on data reorganization and on the production of generalization/exception responses. Results are integrated and evaluated within the WEBCOOP project (Benamara & Saint-Dizier 2004), an intelligent, cooperative question-answering system.

The Discovery of InR : the framework

Our general framework is WEBCOOP, a logic based question answering system that integrates knowledge representation and advanced reasoning procedures to generate intelligent or cooperative responses to NL queries on the web. A general overview of the system is given in (Benamara & Saint-Dizier 2004). In WEBCOOP, responses provided to users are built in web style by integrating natural language generation (NLG) techniques with hypertexts in order to produce dynamic responses. NL responses are produced from semantic forms constructed from reasoning processes. During these processes, the system has to decide, via cooperative rules, what is relevant and then to organize it in a way that allows for the realization of coherent and informative responses. In WEBCOOP, responses are structured in two parts : (1) the production of explanations that report user misconceptions and then (2) the production of flexible solutions that reflect the cooperative know how of the system. This component is the most original. It is based on intensional description techniques and on intelligent relaxation procedures going beyond classical generalization methods elaborated in artificial intelligence. This component also includes additional dedicated cooperative rules that make a thorough use of the domain ontology and of general knowledge. This paper deals with the intensional component.

Knowledge Representation in WEBCOOP

WEBCOOP has two main sources of information: (1) general knowledge and domain knowledge represented by means of a deductive knowledge base, that includes facts, rules and integrity constraints and (2) a large set of indexed texts, where indexes are logical formulae. Our system being a direct QA system, it does not have any user model.

The first source includes basic knowledge (e.g. countries, distances between towns), and ontological knowledge. Ontologies are organized around concepts where each concept node is associated with its specific lexicalizations and properties. For example, the concept `hotel` has the specific properties `night-rate` and `nb-of-rooms`. Values associated with scalar properties allow for sorts, useful for category (2) above. The aim is to sort concepts according to specified dimensions.

We assume that the most relevant documents w.r.t the user question are found using standard information retrieval techniques and that the relevant paragraphs that respond to the question keywords are correctly extracted from those documents. Then, our knowledge extractor transforms each relevant paragraphs into the following logical representation: $text(F, http)$ where F is a first-order formula that represents some knowledge extracted from a relevant paragraph with address $http$ (Benamara & Saint-Dizier 2003). For example, the following text fragment :

....three star hotels in Cannes....
is represented by:
$text(hotel(X) \wedge localization(X, in(cannnes)) \wedge city(cannes) \wedge category(X, 3star), \ www.cote.azur.fr)$.

Query Representation and Evaluation

The parse of a query allows to identify: the type of the query (yes/no, boolean or entity), the question focus and the semantic representation of the query in first-order logic (conjunctive formula). For example, the question:
Q1: *what are the means of transportation to go to Geneva airport*
has the following logical representation: $(entity, Y : meansoftransportation, \ route(X) \wedge to(X, Z) \wedge bymeansof(X, Y) \wedge meansoftransportation(Y) \wedge airportof(Z, geneva))$.

Evaluating a query is realized in two steps. First, we have to check if the question is consistent with the knowledge base. If neither a misconception nor a false presupposition is detected then the extensional answer set that corresponds to the question can be found in two different ways: (1) from the deductive knowledge base, in that case, responses are variable instances or (2) from the indexed text base, and in that case, responses are formulae which subsumed with the query formula. Roughly, unification proceeds as follows. Let Q (conjunction of terms q_i) be the question formula and F (conjunction of f_j) be a formula indexing a text. F is a response to Q iff for all q_i there is an f_j such that:

1. q_i unifies with f_j or
2. q_i subsumes, via the ontology, f_j (e.g. means-of-transportation(Y) subsumes tramway(Y)), or
3. q_i rewrites, via rules of the knowledge base, into a conjunction of f_j, e.g.: $airportof(Z, geneva)$ rewrites into: $airport(Z) \wedge localization(Z, in(geneva))$.

An Algorithm for the construction of InR

Given the set of extensional answers to a question, which is, in most cases, formulas, as explained above, content determination of an InR is defined as follows: (1) search in the answer set or in a related node in the ontology for the adequate element to be generalized, then (2) find the best level of abstraction for the answer, possibly including a list of exceptions. An intensional supervisor (cf. next section) manages the different operations, including non-determinism (*variable depth intensionality*) using a conceptual metrics.

The discovery of InR begins by searching in the answer set a relevant generalized element. It is important to note, that our algorithm identifies and eliminates those predicates in the response that are not relevant for the abstraction task, e.g. predicates that have the same instanciations in the response logical formula, and predicates that does not have any entry in the domain ontology such as hotel names.

The algorithm that searches for the best element to generalize is as follows:

1. check if the list of extensional answers includes a term that can be generalized or sorted using the question focus.

For example, the question Q1 above can have the following InR : *all public transportation and taxis go to Geneva airport* where the extensional answers, e.g. buses, trolleys and trains, are generalized using the type of the question focus *means_of_transportation*.

2. else, search in the question for a property associated with the focus that has variable values in the answer set. For example, the list of answers to the question :
$hotel(X) \wedge localization(X, [atborderof(Y), in(monaco)]) \wedge sea(Y) \wedge city(monaco)$
(*what are the hotels at the border of the sea in Monaco*) cannot be generalized using the focus $hotel(X)$ which corresponds to hotel names and is therefore eliminated, as described above. The property $localization$ of the concept hotel is selected because all extensional answers include the distance of the hotel to the sea, (distance which is an instance of the semantics of atborderof(Y)). A possible InR is then the list of hotels in Monaco sorted according to their increasing distance to the sea.

3. else, search in the ontology for a property related to the focus on which a generalization can be made. For example, if we ask for *hotels in Cannes with a swimming pool* represented by the formula
$hotel(X) \wedge localization(X, in(cannes)) \wedge equipment(X, Y) \wedge swimmingpool(Y)$,
no possible generalizer can be found neither on the focus X, nor on its properties, localization and equipment. So a possible InR is : ”3 stars and 4 stars hotels in Cannes have a swimming pool” where the generalization is realized on the property ‘hotel category’ associated with the concept hotel.

4. else, no InR can be generated. The answer list is simply enumerated without any intensional treatment.

Given the element in the response on which intensional calculus operates, the next stage is to find the best level of abstraction. Considering the set of instances of this element in the answer set, generalization proceeds by grouping instances by generalizable subsets w.r.t. the ontology. This procedure may be repeated until a single generalizer is found, if it exists, possibly with a few exceptions. The following examples illustrate our approach.

Example 1. Suppose the following fragment of a transportation ontology such as described in the following figure.

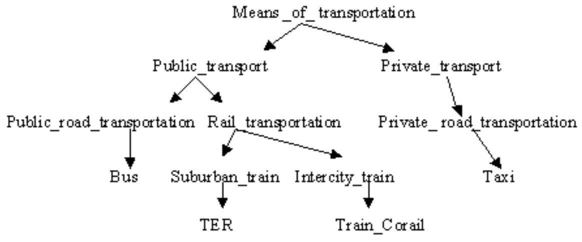


Figure 1: A fragment of the transportation ontology

Suppose again, that the question $Q1$: *what are the means of transportation to go to Geneva airport* has the following extensional answers set: trains, buses and taxis. According to the ontology described in figure 1, this question has the following generalization levels:

1. InR$_a$: all intercity and all suburban trains of Geneva, taxis and buses go to the airport,

2. InR$_b$: all rail transportations, buses and taxis of Geneva go the airport,

3. InR$_c$: most public transportations and taxis of Geneva go to the airport,

4. InR$_d$: most means of transportation of Geneva go to the airport.

InR$_a$ to InR$_c$ are possible generalizations, InR$_d$ is correct, but not very informative because of its proximity to the query[1].

Example 2. Suppose someone asks for *3 star county cottages in southern France between 1th and 10th August* :
$Q2 = (entity, X : country\_cottage,\ stay(W) \wedge accommodation(W, X) \wedge$
$county\_cottage(X) \wedge localization(X, in(Z)) \wedge southfrance(Z) \wedge category(X, 3star) \wedge$
$period(W, P) \wedge between(temp,\ P,\ 1august,\ 10august))$.

Suppose, again, that the inference engine retrieves more than 40 available 3 star country cottages located in different regions in south France, among which :
- county cottages in Foix,
- county cottages in Carcassonne,
- county cottages Pradis in Cannes,
- etc.

This question can be generalized using the property $localization$ associated to the focus $country\_cottage$ (step 2 of the algorithm). According to the following ontology fragment :

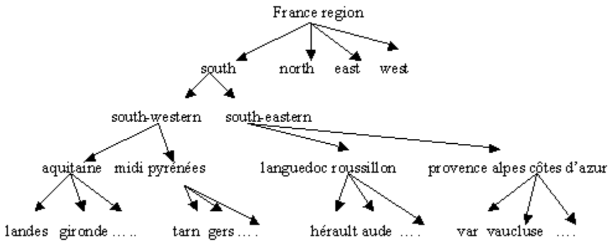


Figure 2: A fragment of a region ontology

and to the corresponding extensional answers set, the question $Q2$ can have the following generalization levels :

1. InR$_e$: 3 star available country cottages in : Aquitaine except for the Landes department, in Midi Pyrénées except for the Tarn department, in Languedoc Roussillon and in Provence Alpes Côte d’Azur regions.

[1] The query itself is often the only InR of the retrieved set of values of which the user is aware (Motro 1994)

2. InR$_f$: 3 star available country cottages in south eastern regions and in south western regions except for the Landes and the Tarn countries,
3. InR$_g$: 3 star available country cottages in south France regions except for the Landes and the Tarn countries.

These examples show that our algorithm derives graded levels of generalizations. It is then necessary to select responses at the best level of abstraction. (Cholvy & Demolombe 1986) performs a syntactic check on the response set that selects only those IA that are not logically subsumed by any others (e.g. IA$_c$ above). They also used another criterion which aims at limiting the set of interesting IA only to answers which share the same vocabulary defined by the user in his question in term of concepts, properties and constants. These techniques are simple and not adequate for our purpose since, first, the choice of the best level of abstraction is automatically performed and, second, we want to give users the possibility of choosing the type of intensional answer which is the most appropriate for them. This is, in our sense, more cooperative, provided the system is not too verbose.

A Supervisor for intensional calculus

In our case, a supervisor manages both the abstraction level and the display of the elaborated InR. In WEBCOOP, we have a *variable depth intensional calculus* which allows the user to choose the degree of intensionality of responses in terms of the abstraction level of generalizers in the ontology. This choice is based on a conceptual metrics $M(C, C')$ that determines the ontological proximity between two concepts C and C' (Budanitsky & Hirst 2001). Considering the form of our ontology, roughly, our similarity metrics considers the sum of the distance in the ontology between the two concept C and C' in terms of the number of arcs traversed combined with the inverse proportion of shared properties, w.r.t the total number of properties on these nodes. This metrics is defined as follows :

$$M(C, C') = NbArc(C, C') + \frac{Card(prop(C) \cup prop(C'))}{Card(prop(C) \cap prop(C'))}$$

The variable depth intensional calculus

Let $ResponseSet = \{InR_1, ..., InR_n\}$ the set of possible generalization levels for a given question and let $generalizer(InR_i)$ a function that returns the list of generalizers for the response InR_i with : $InR_i = all\ Gen_1\ and...and\ all\ Gen_j$. The variable depth intensional calculus is implemented by the predicate : $var\_depth(ResponseSet,\ Choice)$ where the best level of abstraction that corresponds to the variable $Choice$ is determined using the following rule :

$var\_depth(ResponseSet,\ Choice) \longrightarrow$
$list\_metrics(ResponseSet,\ MSet)$,
$max(MSet, Choice)$,

where, $list\_metrics(ResponseSet,\ MSet)$ computes for each $InR_i \in ResponseSet$, the value of $M(Gen_a, Gen_b)$ with $Gen_a, Gen_b \in generalizer(InR_i)$ $(1 \leq a \leq j, 1 \leq b \leq j$ and $a \neq b)$ which corresponds to the conceptual distance between the generalizers Gen_k. Then, depending on the variable $MSet$, the choice of the best level of abstraction is made as follows. If this metrics shows an important distance between concepts, then it is more appropriate to stay at the current level of abstraction. Otherwise, it is best to replace two similar concepts by their immediate mother concept in the ontology, and recursively for the other $Gen_{k'}$.

If we go back to the example 1 in the last section, the supervisor computes, for the response InR$_b$, the metrics M(bus, taxi), M(train, taxi) and M(train, bus). After the computation of the metrics associated to each response level, this strategy allows to choose the InR$_b$ as the best summary. The same strategy is used in example 2, where, for example, for InR$_f$ the metrics M(south_eastern, south_western) is computed. Finally, InR$_f$ is chosen to be the best level of abstraction.

The response display strategy

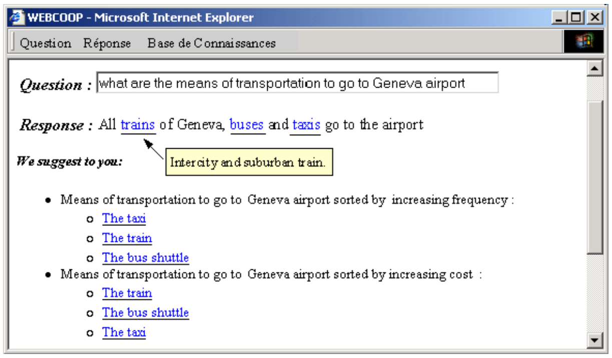
The organization of the response display is as follows. The retrieved intensional answers are structured in two parts. First, the generation of an InR that corresponds to generalization and/or exception of the extensional answers set and then the generation of InRs that corresponds to (1) a sorted list (if responses can be sorted) of the retrieved extensional answers according to different criteria, identified as sortable or (2) another kind of generalization. This strategy avoids the problem of having to guess the user's intent. In example 1 of the previous section, the extensional answers set for the question $Q1$ can be sorted according to two different criteria: the frequency and the cost. In example 2, the list of available 3 star country cottages can be generalized according to the kind of leisure activity practiced in the country cottage like fishing or riding. This criteria is one of the properties of the concept country cottage in the domain ontology.

This second part is best viewed as the expression of the cooperative know-how of the system. For the moment, the maximum number of InR given in the second part is fixed to three which seems to be a manageable number of possibilities. If the number of ordering criteria, for a given concept or property in the ontology, is greater than three then the system arbitrarily chooses three of them. This strategy is still under evaluation.

InR are displayed in our system WEBCOOP using an hyperlink presentation that allows the user to view either the intensional definition of the generalized classes or their extensional enumeration. Because the user is familiar with the concepts he or she is using, we try to keep in the natural language responses, as much as possible, the same syntactic structure and lexicalizations as in the question. The following figures illustrate how intensional answers are displayed in WEBCOOP, for both the two examples in the last section, as explained above.

Conclusion and Perspectives

We proposed in this paper a method for discovering graded levels of generalizations in a cooperative question answer-


WEBCOOP - Microsoft Internet Explorer
Question Réponse Base de Connaissances

Question : what are the means of transportation to go to Geneva airport

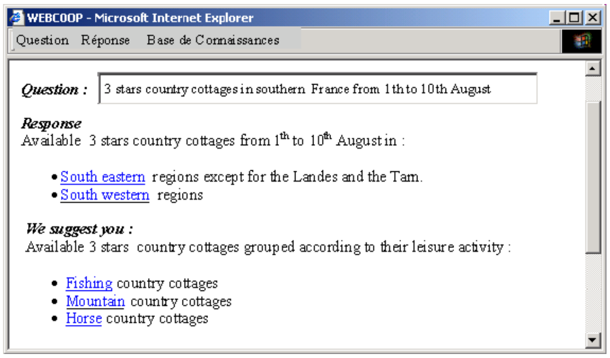
Response : All trains of Geneva, buses and taxis go to the airport

Intercity and suburban train.

We suggest to you:

- Means of transportation to go to Geneva airport sorted by increasing frequency :
  - The taxi
  - The train
  - The bus shuttle
- Means of transportation to go to Geneva airport sorted by increasing cost :
  - The train
  - The bus shuttle
  - The taxi

Figure 3: Intensional answers in WEBCOOP: example 1


WEBCOOP - Microsoft Internet Explorer
Question Réponse Base de Connaissances

Question : 3 stars country cottages in southern France from 1th to 10th August

Response
Available 3 stars country cottages from 1th to 10th August in :

- South eastern regions except for the Landes and the Tarn.
- South western regions

We suggest you :
Available 3 stars country cottages grouped according to their leisure activity :

- Fishing country cottages
- Mountain country cottages
- Horse country cottages

Figure 4: Intensional answers in WEBCOOP: example 2

ing framework when the set of extensional answers set, for a given natural language question, is very large. The originality of our approach mainly lies in the use of the question focus and a rich ontological description of the domain, in order to select a set of relevant concepts to be generalized. We develop a *variable depth intensional calculus* which allows for the generation of intensional answers at the best level of abstraction. A supervisor guides this process using a conceptual metric and manages the response display by structuring InRs in two parts, allowing for a mixed and graded generation of InRs based on different criteria.

On the basis of (Motro 1994), IQA can be evaluated according to three main features : (1) intensional only (pure) versus intensional/extensional (mixed) ; (2) independence from the database instances versus dependence and (3) completeness of the characterization of the extensional answers. Our approach is mixed, dependent and complete since our algorithm computes all non redundant InRs.

For the moment, we are evaluating the linguistic and the cognitive adequacy of the generated intensional answers using an experimental psychology method since an evaluation in TREC style (Voorhees 2002) is not adequate for our purpose. We have developed several experimental protocols and interpretation results are ongoing. In the future, we plan to :

- integrate integrity constraints in the intensional calculus. In this case, the InR will be pure.
- investigate other metrics in order to enhance the system capabilities for choosing the best level of abstraction,
- introduce a user model for the selection of the best response.

References

Benamara, F., and Saint-Dizier, P. 2003. Knowledge Extraction from the Web: an Experiment and an Analysis of its Portability. *revue Vivek, volume 15, numro 1* 3–15.

Benamara, F., and Saint-Dizier, P. 2004. Advanced Relaxation for Cooperative Question Answering. *New Directions in Question Answering, Chapter 21* A paratre.

Budanitsky, A., and Hirst, G. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. *In Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics. Pittsburgh, PA.*

Cholvy, L., and Demolombe, R. 1986. Querying a Rule Base. *Expert Database Conf* 477–485.

Hah, J.; Cai, Y.; and Cercone, N. 1993. Data-driven discovery of quantitative rules in relational databases. 29–40.

Han, J. W.; Huang, Y.; Cercone, N.; and Fu, Y. J. 1996. Intelligent Query Answering by Knowledge Discovery Techniques. *IEEETrans. On Knowledge And Data Engineering* 8:373–390.

Motro, A. 1994. Intensional Answers to Database Queries. *IEEE Transactions on Knowledge and Data Engineering, volume 6 number 3.* 444–454.

Park, J.; Chen, M.; and Yu, P. 1995. An effective hash based algorithm for mining association rules. *In Proc. ACM SIGMOD Intl. Conf. Management of Data, May.*

Piatetsky-Shapiro, G. 1991. *knowledge discovery in database.* AAAI, MIT press.

Shum, C., and Muntz, R. 1988. An Information-Theoretic Study on Aggregate Responses. *In Proceedings of the 14th VLDB Conference* 479–490.

Voorhees, E. M. 2002. Overview of trec 2002. *In NIST Special Publication 500-251: The Eleventh Text Retrieval Conference.*

Yoon, S.-C., and Park, E. K. 1999. An Approach to Intensional Query Answering at Multiple Abstraction Levels Using Data Mining Approaches. *Proceedings of the 32th Hawaii conference on System Sciences.*