PUC

Series: Monographs in Computer Science
and Computer Applications
Nº 11/75

SEMANTIC ASPECTS OF THE RELATIONAL MODEL
- A RESEARCH OUTLINE -

by

A. L. Furtado

Departamento de Informática

Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente 209 — ZC 20
Rio de Janeiro — Brasil

Series: Monographs in Computer Science
and Computer Applications
Nº 11/75

SEMANTIC ASPECTS OF THE RELATIONAL MODEL
- A RESEARCH OUTLINE - *

by

A.L. Furtado

Series Editor: Larry Kerschberg

December, 1975

* This work was partially supported by the Brazilian Government Agency FINEP under contract Nº 244/CT, and developed by the Information Systems Group, Departamento de Informática, PUC/RJ.

ABSTRACT

The relational data base model has been criticized on the grounds that more emphasis should be given to "meaning" and that relational languages are not what non-mathematical users would need. On the other hand, the unrestricted use of natural language seems too expensive, at the present stage, to be employed with commercial data bases.

A line of research is indicated, involving the determination of a severely restricted subset of natural language, duly adapted to the relational model, to be used as a means of communication with data bases.

KEYWORDS

Data bases, relational models, natural languages, case grammars.

RESUMO

O modelo relacional de banco de dados tem sido criticado com base em que deveria ser dada maior ênfase ao "significado" e em que as linguagens relacionais não são o que os usuários não-matemáticos necessitariam. Por outro lado, o uso irrestrito da linguagem natural parece muito caro, no estágio atual, para ser empregado em bancos de dados comerciais.

Indica-se uma linha de pesquisa que envolve a determinação de um subconjunto de linguagem natural bastante restrito, devidamente adaptado ao modelo relacional, a ser empregado como um meio de comunicação com bancos de dados.

PALAVRAS-CHAVE

Bancos de dados, modelos relacionais, linguagens naturais, gramáticas de casos.

CONTENTS

DISCUSSION

The relational model views data bases as consisting of domains D_1 , D_2 , ..., D_m and relations, defined as subsets of Cartesian products of such domains [1, 2, 3, 4].

Thus an n-ary relation R (D_{i_1}, D_{i_2}, ..., D_{i_n}) is simply defined by

$$R \subseteq D_{i_1} \times D_{i_2} \times \ldots \times D_{i_n}$$

Apart from this algebraic characterization, R has no further meaning. As a consequence, the manipulation of data bases is also conceived in purely mathematical terms, using either a relational algebra or a relational calculus as basic languages [5, 6].

As one would expect, such languages tend to be elegant and precise, but they do not appear to be "natural" for non-mathematical users. A sound contribution is now being offered by researchers comming from the area of computational linguistics [7] , where very important work is in progress, dealing with natural language understanding [8, 9, 10].

It is unfortunate that the complexities and ambiguities, besides problems of size, condemn the unrestricted usage of natural language when a reasonable operation of commercial data bases is considered.

We conjecture that a worthwhile research effort would be to determine what features of the recent studies on natural language could be safely incorporated, and what adaptations should be made to them.

In this informal report, we shall use the following example in order to illustrate the discussion.

Consider the domains:

S# - suppliers
P - parts
J - projects

T - means of transportation
L - local agents
A - geographical areas

and the relations;

S (S#, P, J)
R (L, S#, A)
A (P, T)

Assume that the above relations "mean", respectively:

S - x supplies y to z - where $x \in S\#$, $y \in P$, $z \in J$
R - x represents y in z - where $x \in L$, $y \in S\#$, $z \in A$
A - x is transported by z - where $x \in P$, $y \in T$

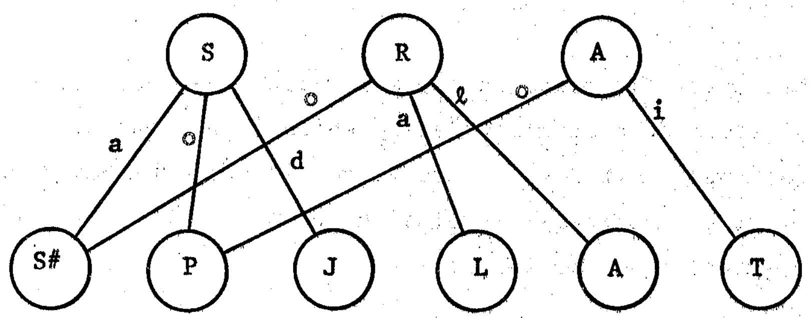
Of course there may be other (occasionally many) alternative phrasings with the same meaning. In particular:

S - y is supplied by x to z
S - z obtains its supply of y from x

where y and then z are taken as the subject of the sentence. Considerations of this sort have led to regarding syntactical categories (subject, object, etc.) as "surface" concepts, which can be usefully replaced by the concept of case* . Following in part [8], we can thus represent the three relations, using the cases:

a - agentive
o - objective
d - dative
l - locative
i - instrumental

* Codd's notion of "role" [1] is very akin to this concept.



with (at least) n paraphrases associated with each case frame corresponding to an n-ary relation. Now, the query

(1) "Which agents deal with air companies?".

could be handled in two extreme ways:

a. by the relational calculus, omitting range variables and quantification, as

$$\underline{\text{GET}}\ W\ (R.L) : R.\ S\# = S.S\# \wedge S.P = A.\ P \wedge A.\ T = \text{'plane'}$$

b. by a natural language processor, as originally stated, if the processor knows (or is able to infer) that an agent has to deal with air companies if he represents some supplier that supplies some part that can arrive by plane *.

Let us take another query:

(2) "By what means are the parts supplied by supplier s3 to project p4 transported?"

a. relational calculus

$$\underline{\text{GET}}\ W\ (A.T) : A.P = S.P \wedge S.S\# = \text{'s3'} \wedge S.J = \text{'p4'}$$

b. natural language processor - original query, looking for the means whereby the parts supplied by s3 to p4 are transported.

* A semi formal way of expressing this might be: R_L (L, $S_{s\#}$ (S#, A_p (P,'plane'), *), *), where R_L, $S_{s\#}$, and A_p are paraphrases having as subject the domain appearing as subscript.

Note that query (1) presents a problem that does not arise with (2) : query (1) refers to a relation "deal" which is not one of the basic (primitive) relations; on the other hand query (2) uses the basic relations only, and employs the paraphrase of relation S where a part is the subject of the underlying sentence (see above).

We hope that a number of points have occurred to the reader as he considered these examples:

1. The basic relations are (in the examples, and we feel that they should always be) simple sentences, dominated by a single verb (this could be a single verb phrase). We shall say that basic relations of this kind are in s-normal form. In addition, we shall require that all relations be in Codd's first normal form.

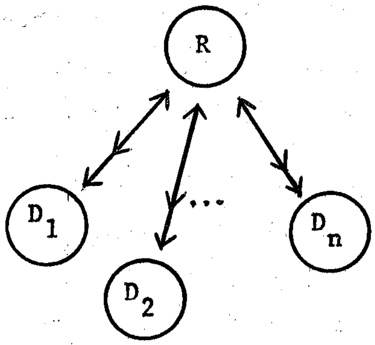
2. It may happen that breaking a relation into relations in s-normal form may also convert them into third normal form as with " x works in y and reports to z", where x is an employee, y a department, and z a manager (who heads department y). The two relations : "x works in y" and "z heads y" are both in s-and in third normal form. However we shall not claim any equivalence or implication between the two normal forms, because the criterion of singleness of verb has nothing to do with the idea of keys and functional dependencies, which incidentally will not be treated in this paper.

3. A relation in s-normal form is conveniently represented by using the case frame format. With each case argument we may associate one (or more) sentences, which are paraphrases of each other, where the chosen case argument becomes the subject. Note the correspondence with the projections on one domain of the relational model.

4. Non-basic relations which involve two or more basic relations are expressible in terms of complex sentences, formed by relativization (more precisely by restrictive relativization [11]. Such non-basic relations are no longer in s-normal form of course, but they are still in first normal form; the correspondence with the join operation of the relational model is obvious.

5. It seems that relativization is the only mechanism that we need*, if we restrict ourselves to queries expressible as relations in first normal form. If nominalization were allowed we might form hierarchies, as in " x knows that y hired z"; here, what is known is the entire sentence "y hired z", rather than one of its components.

6. A diagram of a data base using case frames is often called a semantic network. As a consequence of the requirement that relations should be in first normal form, we have that no case argument is itself a relation and therefore that the network has only two levels: nodes labelled with the name of the verb denoting the relation, leading through edges labelled with the adequate case names, to nodes labelled with the names of the domains. Each two-level tree in the network represents an event [7], and corresponds to a row of the table where a relation is represented (extensionally) in the relational model. We note in passing that such network is a "simple plex" [12] ; if R is a relation over domains D_1 , D_2, ..., D_n we have the schema:



* Conjoining relative clauses should also be allowed, in order to avoid the less convenient stacking constructions [11].

which could be implemented in DBTG (the diagram could be drawn upside down, to emphasize that the domains would be the "owners" in each of the n "sets").

It is unfortunate that there is no consensus among the authors on what should be the best case classification. Perhaps for each data base one should choose empirically the cases that seem to describe the specific situation best.

Also, Fillmore distinguishes the proposition (where the cases arise) from the modality of a sentence: "The constituent Modality contains interrogative and negative elements, sentence adverbials, time adverbials, and various other adverbial elements that are understood as modalities on the sentence as a whole rather than subconstituents of the constituent containing the main verb. I have no strong convictions that these various elements actually comprise a single constituent ..." [8] .

This open-endedness reinforces the view that cases should be chosen for each application. A similar problem appears in the attempted distinction between associations and characteristics [13] (cf. the latter with the PAs - attributes of an object - in [9]).

In order to justify the consideration of characteristics as a special category of binary relations, it has been argued that certain real world concepts do not exist by themselves but only as long as they qualify other concepts. However:

1. Even when this distinction (which is of course very "relative") can be established in a particular data base, its practical significance is not as considerable as it would seem. For example, color could be regarded as a characteristic in a data base about toys; so, presumably, an implementor would not require that the names of colors be represented uniquely, nor that there be inverted lists collecting all references to each color - but this consequence of the "unessential" nature of colors would lead to inefficiency in the very likely situation where queries such as "which of the toys are blue" are frequent.

2. In [7] quantity is regarded as a characteristic of parts. Take the relation : "x supplies y with z", where x is a supplier, y a quantity, and z a part. According to [7] there is in fact an association between x and z, and y is a characteristic of z with respect to (wrt) the event (tuple) x supplies z. This is done because the same z could be characterized by other different quantities in other events of the association. But the need for the wrt edges shows that characteristics either can refer to events (violation of first normal form) or, if one insists in keeping its link to a concept and the wrt edge to an event, it ceases to be binary.

In our opinion it is better not to establish the distinction, but simply to recognize that relations can be dominated by verbs (or verb phrases) not denoting actions, e.g. : "x has a y color", where x is a toy and y a color; "the color of x is y", "x is colored y", etc. are obvious alternatives. We insist that the relation is not dominated by the attribute name (color) because in a data base relation an event is always an assertion, which requires a verb - saying that only "true" characteristics are stored is the same as admitting that an implicit verb is understood.

On the other hand, we much prefer the relation "x supplies y with z" (in item 2 above) than its decomposition into an association and a characteristic with the wrt device.

We also favor the adoption of unary relations that simply assert the existence of each domain element. One would say that a domain element is not active if it does not participate in any of the non-unary relations at a given time. This does not mean that all domain elements have to be created as the data base is initiated and that they cannot be deleted; it means that they will be installed or deleted whenever it is convenient to remember or forget them in the data base, and we note in passing that this would avoid certain anomalies in the usual relational data bases, and would provide a more natural solution for queries which in fact involve quantification over domains rather than events (e.g. : "who are the suppliers that supply all projects?" would mean: "who are the suppliers that supply all projects that exist at present in the data base ?").

A RESEARCH OUTLINE

The general goal of the proposed research is to restrict and adapt findings in the area of natural language understanding, so that languages closer to natural language may be used for communication with a relational data base, while still maintaining a reasonable efficiency. This trade - off, of course, will be continuously affected by the current state of the art.

Some of the features to be investigated are:

1. Choice of cases. As noted, it is probably true that, given the present state of the art, this choice will vary for each application.
2. Patterns for building the paraphrases of case frames.
3. Patterns for constructing complex sentences.
4. Usage of reserved words in queries, such as prepositions (associated with cases), pronouns, determiners, names of the domains, quantifiers, comparison operators (for emulating Θ-joins), etc. A liberal usage of such words within the above patterns will tend to simplify the task of parsing the sentences*.
5. Schema information, i.e. information to be kept in the data base about itself. This could well include definitions [14] of non basic relations, such as the "deal" relation of our simple example, but no inferential capability would be contemplated.
6. An interactive capability, using the schema information, in order to help the users to formulate their queries.

*Notice the occurrence of such words in our sample query: "which agents represent some supplier that supplies some part that can arrive by plane?" Notice also how quantification is applied to domains which have been previously restricted by relative clauses.

Few patterns (items 2 and 3) would be used, and care would be taken to avoid the possibility of ambiguities. The choice of the patterns depends mainly on their suitability to the users and on the ease of parsing them (this task is usually done by some ATN analyzer [15]). Being suitable or "natural" to the user is not necessarily a quality impossible to measure in quantitative terms; certain pyschological studies on factors, such as latency time needed to verify the truth of assertions phrased either as affirmative or negative sentences [16], may provide the basis for a similar experiment for choosing among several possible patterns.

Natural language processors usually include the capability of generating sentences, often in order to answer wh-queries. We feel that this capability is not so important as the formulation of the query itself, in natural language. Once the user has formulated a query in a way which is very clear (although perhaps the system could tolerate a considerable amount of "ungrammaticalness"), and recorded it for documentation purposes, it is perfectly acceptable to have the answers printed out in tabular form, or as formatted reports, or as images at a graphics terminal, etc.

Thus, we expect that the natural language front end outlined here will not be inordinately large. It would merely support queries constrained to conform with a small number of patterns (relational completeness should be verified here); in order to support this capability there would be interaction with the user with the help of the schema information. The schema views the data base as a two-level semantic network, which as we saw before models correctly a relational data base; consequently, the data base itself can use any implementation meeting the requirements of the relational approach.

Once all these simple ideas have been implemented and tested, further research would consider the gradual introduction of other more sophisticated features. Of particular interest would be the representation of consistency and security constraints within the schema. Among the other possibilities, we could mention a deeper representation of knowledge in terms of primitive actions and patterns for defining the data base relations - a feature for helping and constraining in the design itself of a data base (see the primitive actions in [9]), etc.

REFERENCES

1. CODD, E.F. A relational model of data for large shared data banks. Commun. of the ACM, 13 (6) : 377-87, June 1970.

2. _______. Relational completeness of data base sublanguages. In : RUSTIN, R. Data base systems. Englewood Cliffs, N.J., Prentice-Hall, 1972.

3. _______. Further normalization of the data base relational model. In : RUSTIN, R. Data base systems. Englewood Cliffs, N.J., Prentice-Hall, 1972.

4. _______. A data base sublanguage founded on the relational calculus. In : ACM/SIGFIDET WORKSHOP ON DATA DESCRIPTION, ACCESS AND CONTROL, 1971.

5. CHAMBERLIN, D.D. & BOYCE, R.F. SEQUEL : a structured English query language. In : ACM/SIGFIDET WORKSHOP ON DATA DESCRIPTION, ACCESS AND CONTROL, 1974.

6. SMITH, J.M. & CHANG, P.Y.T. Optimizing the performance of a relational algebra data base interface. Commun of the ACM, 18 (10) : 568-79, Oct. 1975.

7. ROUSSOPOULOS, N. & MYLOPOULOS, J. Using semantic networks for data base management. Toronto, Univ. of Toronto, Department of Computer Science, 1975.

8. FILMORE, C.J. Toward a modern theory of case. In : REIBEL, D. & SCHANE, S.A., ed. Modern studies in English. Englewood Cliffs, N.J., Prentice-Hall, 1975.

9. SCHANK, R.C. Conceptual information processing. Amsterdam, North-Holland, 1975.

10. NORMAN, D. A. & RUMELHART, O. E. *Explorations in cognition.* W. H. Freeman, 1975.

11. STOCKWELL, R. P. et alii. *The major syntactic structures of English.* New York, Holt Rinehart and Winston, 1973.

12. MARTIN, J. *Computer data-base organization.* Englewood Cliffs, N. J., Prentice-Hall, 1975.

13. SCHMID, H. A. & SWENSON, J. R. On the semantics of the relational data model. In : SIGMOD CONFERENCE, 1975.

14. THOMPSON, F. B. & BOZENA, H. T. Practical natural language processing: The REL system as prototype. *Advances in Computers*, *13* : 110-68, 1975.

15. WOODS, W. A. Transition network grammar for natural language analysis. *Commun of the ACM* *13* (10) : 591-606, Oct. 1970.

16. CARPENTER, P. A. & JUST, M. A. Sentence comprehension : a psycholinguistic processing model of verification. *Psychological Review*, *82* (1), 1975.