Correcting Misconceptions About Data Base Structure

Eric Mays
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

ABSTRACT

This paper presents a method for computation of intensional failures of presumptions in queries to a natural language interface to a data base system. These failures are distinguished from extensional failures since they are dependent on the structure rather than the content of the data base. A knowledge representation has been investigated that can be used to recognize intensional failures. When intensional failures are detected, a form of corrective behavior is proposed to inform the user about possibly relevant data base structure that is related to the failure.

INTRODUCTION

In the course of interacti g with a natural language data base query syste a casual user may pose queries based on beliefs about the domain which are incompatible with those of the system. Kaplan [Kaplan 79] has investigated one such class of beliefs which can be computed from a query and corrected, namely, extensional failures of presumptions. This paper introduces another class, that of intensional failures of presumptions, outlines the kind of knowledge representation needed for their computation, and proposes an appropriate form of corrective behavior.

A presupposition is a proposition that is entailed by all the di ect answers of a question(*). A presumption is either a presupposition or it is a proposition that is entailed by <u>all</u> <u>but</u> <u>one</u> of the direct answers of a question [Kaplan 79]. Hence, presupposition is a stronger version of presumption, and a presupposition is a presumption by definition. For example, question (1a) has several direct answers such as "John", "Sue", etc., and, of course, "no one". Proposition (1b) is entailed by all the direct answers to (1a) except the last one, i.e., "no one". Therefore, (1b) is a presumption of (1a). Proposition (1d) is a presupposition of (1c), since it is entailed by all of the question's direct answers.

1a) Which faculty members teach CSE110?
1b) Faculty members teach CSE110.
1c) When does John take CSE110?
1d) John takes CSE110.

Presumptions can be classified on the basis of what is asserted — i.e., an "intensional" statement about the structure of the data base or an "extensional" statement about its contents. Thus an extensional failure of a presumption occurs based on the current contents of the data base, while an intensional failure occurs based on the structure or organization. For example, question (2a) presumes propositions (2b), (2c), and (2d). Presumption (2b) is subject to intensional failure if the data base does not allow for the relation "teach" to hold between "faculty" and "course". An extensional failure of presumption (2b) would occur if the data base did not contain any faculty member that teaches a course. Also note that the truth of (2b) is a pre-condition for the truth of (2c).

(*) The complete definition of presupposition includes the condition that the negation of a question, direct answer pair entails the presupposition.

2a) Which faculty members teach CSE110?

2b) Faculty members teach courses.

2c) Faculty members teach CSE110.

2d) CSE110 is a course.

Although a presumption which fails intensionally will of neccesity fail extensionally, it is important to differentiate between them, since an intensional failure that occurs will occur consistently for a given data base structure, whereas extensional failure is a transitory function of the current contents of the data base. This is not meant to imply that a data base structure is not subject to change. However, such a change usually represents a fundamental modification of the organization of the enterprise that is modelled. One can observe that structural modifications occur over long periods of time (many months to years, for example), while the data base contents are subject to change over relatively shorter periods of time (hourly, daily, or monthly, for example).

The problem this paper addresses is the recognition of presumptions which fail intensionally. In that case, the failure should be communicated to the user and a form of corrective response produced which informs the user about the relevant data base structure.

DATA BASE MODEL

A data base model based primarily on the entity-relationship model of Chen [Chen 76] with the addition of an inheritance hierarchy can be used to detect the intensional failure of a presumption. This model is similar to that proposed by Lee and Gerritsen [Lee and Gerritsen 78], which incorporates the generalization dimension developed by Smith and Smith [Smith and Smith 77] into Chen's model. Although Lee and Gerritsen, and Chen allow entities to participate in n-ary relationships, this discussion will be restricted to binary relationships. Entities participate in relationships along two orthogonal dimensions, aggregation (among dissimilar entities) and generalization (among similar entities), as well as having attributes that assume values. Along the generalization dimension an entity inherits the attributes and relationships of its super-entities. All individuals of a particular entity set are members of any of that set's super-entity sets. Some individuals in an entity set may be members of a sub-entity set, therefore participating in relationships of the sub-entity set and having attributes of the sub-entity set.

A simple subset operator is not adequate for generalization in this context however, as is illustrated by the following example. Consider the data base model fragment shown in figure 1. Entity sets are designated by ovals, aggregation relationships by diamonds, and generalization relationships by edges from the super-entity set to the sub-entity set. Here "men", "women", "faculty", and "students" are all subsets of "people", with "students" participating in a "take" relationship with "courses". From this it can be determined that a "take" relationship can exist between "men" and "courses", since it is possible that there are some "people" who are both "men" and "students". But by this same reasoning we may also assert that a "take" relationship might exist between "faculty" and "courses", which is certainly not the case in most universities. The essential difference that needs to be noticed is that a non-empty intersection is possible between "men" and "students" and is not possible between "faculty" and "students".

The incorporation of an operator that partitions an entity set into several mutually exclusive sub-entity sets eliminates this problem. This distinction can be made by prohibiting the traversal of a path in the data model that includes two entity sets which are mutually exclusive. Furthermore, the path in the generalization dimension is restricted to "upward" traversals followed by "downward" traversals. An upward (downward) traversal is from a sub-entity (super-entity) set to a super-entity (sub-entity)

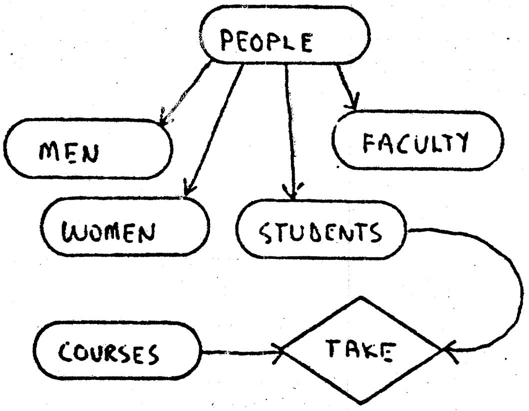

FIGURE 1

set. This restriction is made to prevent over-specialization of an entity set when traversing downward edges. The set of inferences that can be made in the presence of this restriction is not overly constrained, since any two entity sets that have a common intersection (sub-entity set) will also have a common union (super-entity set). As an example of this type of structure, consider figure 2, where partitioning is denoted by parallel arcs across edges. (Usually some attribute of an entity serves as the basis for the partition. For example, "sex" partitions "people" into "men" and "women".) In this fragment of information about university organization the possibility of a "take" relationship existing between "faculty" and "courses" is precluded by the fact that "faculty" and "students" are mutually exclusive. Observe that the path from "students" to "unemployed" would include "people" rather than "undergrads" or "unsupported". If either "undergrads" or "unsupported" were included, "students" would be unnecessarily restricted.

Although it might seem at first that a "teach" relationship might be possible between "undergrads" and "courses" — since all "undergrads" are "students", and "students" and "teachers" are not mutually exclusive — this is not the case. Closer inspection reveals that all "undergrads" are "unemployed", and "unemployed" and "teachers" are mutually exclusive, thus eliminating the possibility. The inferencing about mutual exclusion required to produce this result would proceed in a fashion similar to that proposed by Fahlman [Fahlman 79]. Very briefly, markers are propagated upward from the two entity sets which are assumed to be disjoint. If a split node (which denotes mutual exclusion) detects markers from both entity sets, they are not

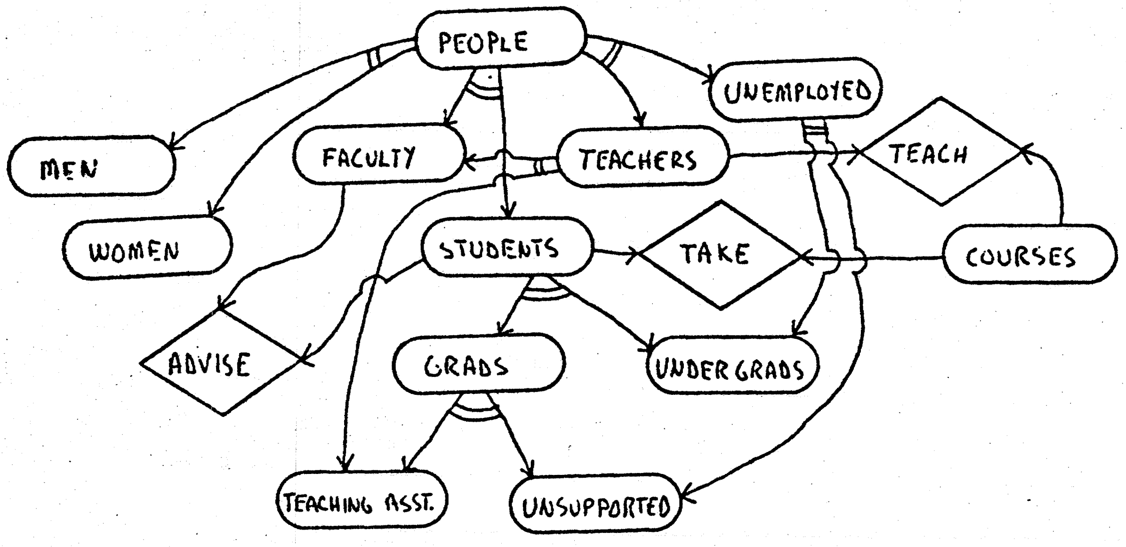

FIGURE 2

disjoint. Fahlman uses this operator to enforce restrictions on updates to a knowledge representation.

INTENSIONAL FAILURE

In this data base model, intensional knowledge can be equated with the ability of an entity to participate in a relationship with another entity. Here, intensional failure occurs when such a relationship can not be established. For instance, the question "Which faculty take courses?" incorrectly presumes that a "take" relationship can exist between "faculty" and "courses" entities.

A method for the computation of a significant class of presumptions in the data base query domain is described by Kaplan [Kaplan 79]. The approach taken there involves the generation of the meta-query language (MQL) from the natural language input. The MQL is essentially a modified parse tree that closely reflects the surface structure of the input query. An example is shown in figure 3 for the question, "Which students in computer science took CSE110?". Kaplan computes the extensional failures of presumptions in a query from the MQL by checking the result of the formal data base query of each connected sub-graph of the MQL for emptiness. That is, the contents of the data base are accessed to determine if a presumption has a non-empty extension.

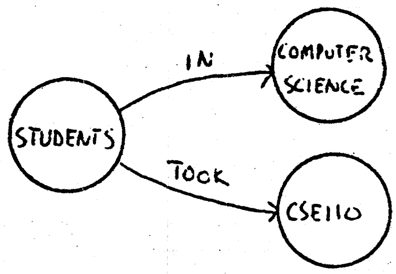

FIGURE 3

The intensional failure of presumptions in a query can be computed in a similar fashion. The essential difference being that the data model image of the MQL representation must be checked to insure that each relationship can be established in the data model. The data model image of a node or arc in the MQL is the entity set or relationship set, respectively, in the data model which is designated to contain the referent or set of referents for it. This is basically equivalent to disambiguating the lexical items, since the arcs and nodes in the MQL have lexical items associated with them. Consider the question, "Which faculty take CSE110 ?" and its corresponding MQL representation in figure 4. Here the entity set "courses" is designated as the data model image for "CSE110" since it is most likely to refer to a "course" entity. This query contains the presumption that "faculty take courses" which can be recognized as failing intensionally because a "take" relationship does not exist between "faculty" and "courses".

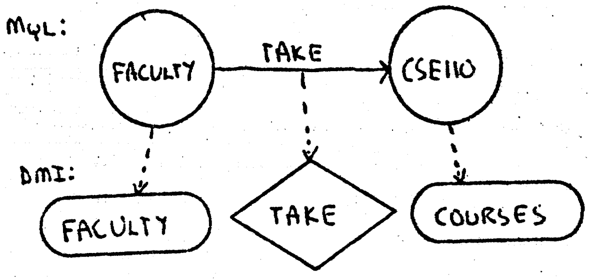

FIGURE 4

Recognizing the intensional failure of presumptions is only part of the problem — it is also useful to provide the user information with respect to related intensional knowledge. Given a relation R, entities X and Y, and a failed presumption (R X Y), salient intensional knowledge can be found by abstracting on either R, X, or Y to create a new relation. For example, using the university data base model fragment, consider the following hypothetical exchange:

Q: "Which faculty take courses?"
A: "I don't believe that faculty can take courses.
Faculty teach courses.
Students take courses."

Here the presumption that faculty take courses can

be recognized as failing intensionally. This can be communicated to the user by paraphrasing its negation, noting as well what possible relevant relationships do hold*

HIGHER ORDER FAILURES

A more complicated interaction of presunptions with the data model can also cause a presumption to fail intensionally. These failures occur in sub-graphs of the MQL which contain two *tot* more arcs* It may be the case that a Relationship can be established for each arc that connects two nodes in the MQL, but there is still « connected sub-graph (a presumption) that fails intensionally. The relationships in a particular sub-graph may iiqpose restrictions on the nodes that will form erqpty response sets which can be recognized solely from intensional knowledge. An example of this is shown in question (3a). The restrictions on "teachers" involve two entities in the same partition. Question (3b) contains the tame intensional failure. Both presume identical propositions, although in (3a) it is not as apparent.

3a) Which teachers that advise students take courses?
3b) Which teachers are both faculty and students?

A corrective response for this type of failure involves identifying the entities that participate in the relationships in addition to the failed presumption. In response to (3a), for example:

"Faculty advise students.
Students take courses.
I don't believe that a teacher can be both a faculty member and a stt^ent."

It doesn't appear that any related knowledge need be catrounicated, although some information regarding the various partitions of an entity set rtdght be helpful. An adequate procedure for determining relevant knowledge along the generalization dimension has not been thoroughly investigated.

RELATING RELATIONSHIPS

An interesting situation arises when attempting to determine related intensional knowledge for a failed presumption with regard to relationships. Consider an enterprise which has a matrix organization as in figure 5. The "in* relationships are conceptually similar but must be represented distinctly. The following behavior is desired for this data model:

Q: "Which employees are in areas?*
A: "I don't believe that eiqployees are in areas.
Employees are in divisions.
Projects are in areas."

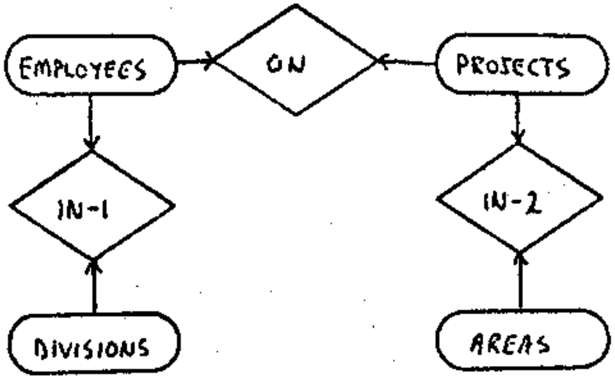

FIGURE 5

But this will not be achieved given the method outlined earlier of abstracting on one of R, X, or Y for a failed presumption (RX Y). If "in-l* is picked as the data model image for "in", the response will not include the fact that "projects are in areas*. Similarly, if "in-2" is chosen, "enployees are in divisions* will not be included. This can be remedied by introducing an operator (R-SET) which denotes the conceptual similarity of relationships as in figure 6* The procedure for determining salient intensional knowledge can be modified to include relationships in the same *R-SBT" when abstracting on a relationship. Although this might appear ad hoc, it should be noted that this would be the first

step towards developing a hierarchy for relationships.

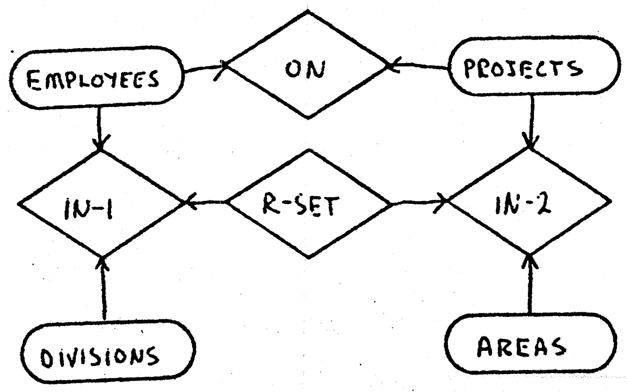

FIGURE 6

Note that there may be some basis for choosing the domain of a particular predicate from a semantic relatedness measure. For instance, if two distinct "teach" relationships existed, between "faculty" and "courses", and "grads" and "courses", the question "Which undergrads teach courses?" would indicate that the "teach" between "grads" and "courses" should be chosen.

CONCLUSION

Intensional failures of presumptions in queries occur when the user's beliefs about the structure of the data base diverge from those of the system. The use of a partitioned subset hierarchy is essential here to determine those intersections of entity sets that are empty by definition. It is important to distinquish between structure and content, since there is a significant difference in the rate in which they change. When responding to intensional failures of presumptions, simply pointing out the failure is in most cases inadequate. The user must also be informed with regard to related knowledge about the structure of the data base in order to formulate queries directed at solving his/her particular problem. A straightforward, but effective, method for producing such responses was outlined here.

ACKNOWLEDGEMENT

I would like to thank Aravind Joshi, Bonnie Webber, and Kathy McKeowen for their comments and suggestions on the various drafts of this paper. This work was partially supported by a grant from the National Science Foundation, NSF-MCS 79-08401.

REFERENCES

[Chen 76]
Chen, P.P.S., "The Entity-Relationship Model — Towards a Unified View of Data", ACM Transactions on Database Systems, Vol. 1, No. 1, 1976.

[Fahlman 79]
Fahlman, Scott E., NETL: A System for Representing and Using Real-World Knowledge, MIT Press, Cambridge, Ma., 1979.

[Kaplan 79]
Kaplan, S.J., Cooperative Responses From a Portable Natural Language Data Base Query System, Ph.D. Dissertation, Computer and Information Science Department, University of Pennsylvania, Philadelphia, Pa., 1979.

[Lee and Gerritsen 78]
Lee, R.M. and Gerritsen, R., "A Hybrid Representation for Database Semantics", Working Paper 78-01-01, Decision Sciences Department, University of Pennsylvania, 1978.

[Smith and Smith 77]
Smith, J.M. and Smith, D.C.P., "Database Abstractions: Aggregation and Generalization", ACM Transactions on Database Systems, Vol. 2, No. 2, June 1977.