Case-Based Query Answering: A Model Using Cognitive Similarity

Agenor Martins
Department of Informatics and Statistics
CCN-UFPI Campus da Ingina
64.049-550 Teresina - PI, Brazil
agenor@dsc.ufpb.br

Edilson Ferneda
Department of Systems and Computing
CCT-UFPB - P.O. Box 10.090
58109-970 Campina Grande - PB, Brazil
edilson@dsc.ufpb.br

Abstract

In this paper we are concerned with the problem of designing a framework for the construction of knowledge systems that are adequate for the task of selecting useful answers to queries. We put forward a proposal to provide representations with that kind of functionality through case-based reasoning mechanisms for handling query descriptions and retrieving intensional answers. We start by positioning *database* and *casebase* with regard to this problem. Then we distinguish similarity assessment as the central task both in query answering and case-based reasoning systems. In order to frame our treatment for this problem, a synthesis of Tversky theory on cognitive similarity is presented. We proceed by exporting and extending Tversky s central concerns to the CBR arena to serve our approach of similarity or case-based query answering.

Keywords: Query Answering, CBR, Knowledge Representation, Cognitive Similarity, Human-friendly Interface

1 Introduction

An intelligent query-answering system is a computer system that is able to perform a query, which is some indication of a user's needs to an information base, into an answer of useful information [Andreasen *et al.*1997]. The ability to perform a query into an useful answer constitutes a fundamental feature of many computing systems. The importance of this is not restricted, for instance, to technologies such as database systems. Indeed, as applications in information access grow in complexity - including the World Wide Web access so grows the demand for human-friendly interfaces in order to make such applications usable. However, regarding this aspect of information access there is a consensus: *[We] have not seen a development that can be compared to the change in capacity of computers* for storing information [Andreasen *et al.*1997].

Two major technologies have allowed information access in a query-answer fashion. On one hand, the database paradigm used to answer questions built upon very well-defined data models such as relational models (*relational database*) and object-oriented models (*object database*) where simple and exact queries used to be supported only by rigid languages of difficult usage for casual users. Alternately, another paradigm for query-answering systems has been the logic or deductive approach [Emielinski1987]. Under this knowledge systems approach, powerful inference engines have been developed for the production of inferences which function literally as answers to questions. The deductive approach, however, used to suffer from user interface problems.

In this context, the *Case-Based Reasoning* technology (CBR) can be focused as an alternative mechanism for designing intelligent query-answering systems. It is a relatively new technology that can be employed for matching a current query description (a *query case*) to a specially organised database of indexed previous situations, called a *casebase* [Althoff *et al.*1995, Leake1996, Watson1997].

Therefore, for the specific purpose of query-answering systems a casebase may be conceived as an indexed store of response selections that is dynamically extended as the system runs over time. The CBR engine searches for case histories (*response cases*) that fully or partially match this description. Then, through a question-and-answer sequence or any mechanism of similarity, the engine systematically narrows the search space to produce a recommended set of cases.

As the CBR strong features spring from its emphasis on *similarity matching*, the importance of empowering similarity processes has long been recognised, for instance, by emulating the human way of similarity assessment [Kolodner and Menachem1991, Martins *et al.*1996]. Empowered *similarity matching* can make CBR processes such as case retrieval, and case-based query answering much more reliable.

In this paper the authors demonstrate how to compute similarity between a query case and cases in a database of previous situations. By importing a theoretical model from the psychological domain to the CBR arena we propose a cognitive similarity mechanism for case-based query-answering systems to answer a question more realistically. In section 2, the typical CBR task of case comparison is shown. In section 3, basic concepts on Tversky's cognitive similarity are discussed. Next, we extend this theory to serve our approach of case-based query answering.

2 The Case-Based Reasoning Paradigm

Troels Andreasen and his colleagues (in Roskilde University, Denmark) argue that in query-answering systems one central task is to compare two information items, one from the user and the other from the information base [Andreasen *et al.*1997]. This will correspond, under the CBR paradigm, to the task of comparing two objects that are cases from a casebase, leading to a measure of similarity between objects. Typically, the similarity between the query case (Q) and a response case (C) denoted by $SIM(Q,C)$ will be the weighted sum of similarity of each attribute [Leake1996]:

$$SIM(Q,C) = \frac{\sum_{i=1}^{n} W_i \times sim(Q_i, C_i)}{\sum_{i=1}^{n} W_i}$$

where i is an individual attribute from 1 to n; W is the importance weighting of attribute i; n is the total number of attributes in each case; $sim(Q_i, C_i)$ is a particular similarity between the value of the i^{th} attribute of the query case (Q) and a case (C) in the case base. An entire family of nearest neighbour metrics such as the Weighted Block-City similarity, and the Weighted Minkowski metric are calculated in a similar way [Althoff *et al.*1995]). In general, CBR similarity metrics have two properties: (i) they relay on weighted attributes; and (ii) they use to consider only the common object attributes (their resemblances).

3 The Framework Foundations: Tversky-Gati Cognitive Theory

At this point, it is interesting to observe how most of the current CBR models for assessment of case similarity contrast with Tversky's theoretical similarity model we chose to base our CBR approach to similarity. Both conventional CBR models (as above) and Tversky-Gati model are based on weighted attributes. However, their resemblance ends there. Besides its cognitive foundation, Tversky's theory has the advantage of elegantly combining into a sole similarity formula both object differences and object similitudes. It was just the originality of this contrasting treatment of similarity that motivated us to import it to the CBR paradigm. Next, we show the core of this contrasting model.

3.1 Attribute Relations in Tverskys Objects

Let p and i be any two given objects. Also, let P and I be the set of attributes describing these objects p and i, respectively. The relations between these objects are given in Figure 1. In this Figure three types of attribute relationships are present:

1. $I \cap P$: the attributes shared by p and i;
2. $P - I$: the attributes of p that are not shared by i; and finally
3. $I - P$: the attributes of i that are not shared by p.

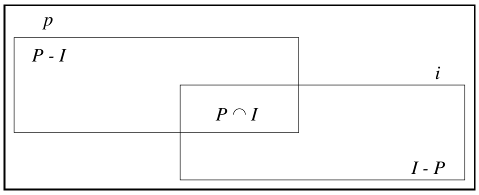

Figure 1: Relations between the attributes of objects p and i

3.2 Similarity Concept

Tversky uses theses relations to express the similarity $sim(i,p)$ between objects i and p as the following *attribute-matching function*:

$$sim(i,p) = \Theta f(I \cap P) - \alpha f(P - I) - \beta f(I - P)$$

where $\Theta, \alpha, \beta \geq 0$. Here, f is a function implicitly defined as a linear combination (a *contrast model*) of the measures of both their common and distinctive attributes.

Naturally, similarity increases with the measures of their common attributes present in both objects and decreases with the measures of the distinctive attributes. On the other hand, the parameters (Θ, α, β) represent weights or votes reflecting the relevance or prominence of the common attributes (in the case of Θ) and distinctive attributes (in the case of α and β). Observe the role of these parameters; depending on the values of these parameters, the model does not define a unique index of similarity but rather a family of similarities over the same set of objects. For example, if $\Theta = 1$, and $\alpha = \beta = 0$, then $sim(i,p) = f(I \cap P)$; that is, the computed similarity equals the measure of the common attributes. On the other hand, if $\Theta = 0$, and $\alpha = \beta = 1$, then $-sim(i,p) = f(P - I) + f(I - P)$; that is, the dissimilarity of i to p equals the measure of the symmetric difference of the respective attribute set.

3.3 The Diagnosticity Hypothesis

According to the Tversky-Gati cognitive approach, another important aspect which influences similarity among objects is *salience of features* [Tversky and Gati1978] (which we also argue to influence similarity among CBR cases). In Tversky's formula, f is generally not invariant with respect to changes in context or frame of reference. That is, the salience of features may vary widely depending on implicit or explicit instructions and on the object set under consideration. North Korea and South Korea, for example, may be viewed as highly *similar* from a geographical or cultural viewpoint and as quite *dissimilar* from a political viewpoint. Moreover, the two Koreas are likely to be viewed as more similar to each other in a context that includes many European and American countries than in a context that includes only Asian countries.

How does the salience of features vary with changes in the set of objects under consideration? Tversky proposes that the salience of features is determined, in part at least, by their *diagnosticity* (i.e., classificatory significance). A feature may acquire diagnostic value (and hence become more salient) in a particular context if it serves as a basis for classification in that particular context. The relations between similarity and diagnosticity form what was denominated the *diagnosticity hypothesis*.

4 Extending Tversky's Theory to CBR Arena

The problem with Tversky's theoretical-feature model is its lack of computability (its high level of abstraction). Our challenge is in how to introduce *computability* to this model in such a way as to turn it into an operational CBR model? How to explicitly define Tversky's similarity formula? How to compute and take into account attribute diagnosticity? We solved this problem in three ways:

1. by representing tverskyan objects as computer cases;
2. by making explicit his metric for assessment of cognitive similarity; and
3. by showing a practical application of our extended formalism.

Once cases have been represented and stored in memory as discussed above, their similarities can be assessed in such a way as to take into account a derived metric tying together of both an explicit Tversky formula and its underlying concepts. To obtain this metric new version, let's define the following:

1. m is any response case in case memory, and p is the user query case;
2. M is the set of attributes characterising the case m;
3. P is the set of attributes characterising the target case p;
4. $M(P$ is the set of attributes shared by m and p;
5. $M \cap P$ is the set of attributes of m that are not shared by case p;
6. $P \cap M$ is the set of attributes of p that are not shared by case m;
7. $w = (attribute\_diagnosticity) \times [(votes\_for\_attribute\_value)$ $\underline{\text{or}}$ $min(votes\_for\_attribute\_value(m,p))]$; where the left-hand side of the boolean operator applies whenever non-shared attributes of cases should be computed; and the boolean operator right-hand side applies vice versa.
8. $SIM(m,p)$ is the observed similarity of case m in relation to case p, such that:

$$SIM(m,p) = a\sum_{i \in M \cap P} w_i - b\sum_{j \in P-M} w_j - c\sum_{k \in M-P} w_k$$

where $a, b, c \geq 0$; a being weights for attributes in the intersection of m and p; b and c being weights for attributes not present in the intersection of p and m. In this metric (in w, more exactly) it is worthwhile to recall that the *diagnosticity* is also measured in terms of weights or number of votes given to a certain attribute name by system designer, expressing its contribution for the case representation (of a situation/object).

5 Illustrative Application

Our illustrative example uses a bank loan problem in which the credit *underwriter* has to select cases of past loans in order to guide new selections of potential borrowers. The problem of how to decide if a new client should be granted a loan will be solved by comparison with past credit situations.

5.1 Casebase Organisation

Past bank loan situations may be organised as computer cases by applying, for instance, the organisational approach described in [Martins and Ferneda1999] and exemplified in Figure 2.

On one hand, let us suppose that *Case X* is the *query case*; the one demanding a kind of clarification possibly encapsulated in a *respose case*. On the other hand, let's say, for simplicity's sake, we only have experience of one case of a loan. In such a casebase of bank loans, *Case 001* was represented as a successful case where the attribute

Outcome suggests that the bank made a large profit, as shown in Figure 2. Cases such as *Case 001* should be helpful for any user query involving credit *underwriting* situations.

To better understand this case representation, at the left-hand side of *Case 001* the number 3, for example, expresses the salience or diagnosticity of the attribute *Income* as three times as important as the client's *Profession* for the purpose of bank loan.

5.2 Query Case and Response Case

In order to search for a closest case containing a similar *Outcome* we need to compute the similarity between *Case 001* and *Case X* (or $SIM(Case001, CaseX)$). In this hypothesis, the set of common and not common attributes in Figure 2 are:

- $M \cap P : Case001 \cap CaseX = \{$*Income, Repayment, Profession*$\}$
- $M - P : Case001 - CaseX = \{Outcome\}$
- $P - M : CaseX - Case001 = \{Job\}$

Case 001		
Attribute Diagnosticity	Attribute: Value	Vote for Value
3	Income: 2000	15
3	Repayment: 200	2
1	Profession: military	2
1	Outcome: good profit	1

Case X		
Attribute Diagnosticity	Attribute: Value	Vote for Value
-	Income: 1500	10
-	Repayment: 400	3
-	Profession: servant	1
2	Job: stable	2
-	Outcome: ?	

Figure 2: Cases of loans as Tversky's objects

By considering the number of votes for the attributes of each set above and applying the new derived metric we have: $SIM(Case001, CaseX) = a(3 \times 10 + 3 \times 2 + 1 \times 1) - b(1 \times 1) - c(2 \times 2) = a37 - b1 - c4$. For the values of parameters $a = 3; b = 1; c = 1$, we get the similarity given by $SIM(Case001, CaseX) = 106$. Observe the way the votes or weights on value for those attributes in the intersection (case common attributes) were computed: 10 votes for value of *Income*, in *Case X*, instead of 15 votes of *Case 001*; 2 votes for value of *Repayment*, in *Case 001*, instead of 3 votes of *Case X*, etc. This occurs as a consequence of the w definition in formula (7) which considers the minimum amount of votes when comparing two cases having the same attribute name.

Once all the similarities of *Case X* have been calculated in relation to the other retrieved cases (as we did for *Case 001*), the system will be able to *rank* these cases according to their similarity values. The higher this final value for $SIM(m, p)$ the closer the similarity between cases and vice-versa.

6 Conclusion

A framework was presented for representing knowledge in such a way that the CBR mechanism works for obtaining useful answers to queries. The heart of the approach is a cognitive framework for the CBR problem of similarity assessment to determine which case is most similar to the new situation. The approach offers major contributions in three complementary ways:

1. We introduce computability to a theoretical psychological model by conceiving the representation of Tversky's objects as a computable *frame* data structure;
2. The approach extends Tversky's cognitive studies of similarity from the psychological domain to the CBR arena. In this sense, it constitutes an innovative effort for the task of computing similarity based on people's judgements of similarity, in contrast to the common nearest neighbour metrics.
3. The extended similarity framework was particularly instantiated to the CBR task of obtaining useful answers to queries.

An additional advantage of turning tverskyan concepts into a sufficiently operational CBR model refers to the sub-task of *case ranking* which is the process of ordering matching cases according to goodness of match or usefulness. We advocate that the correct administration of values for parameters a, b, c in formula (8) may function as a valuable *ranking mechanism* by allowing for a family of similarities over the same set of cases.

References

[Althoff *et al.*1995] K. Althoff, A. Eric, B. Ralph, and M. Michael. *A Review of Industrial Case-Based Reasoning Tools. An AI Perspectives Report.* A. Goodall, 1995.

[Andreasen *et al.*1997] T. Andreasen, H. Christiansen, and H. L. Larsen. *Flexible Querry Answering Systems.* Kluwer Academic Publishers, 1997.

[Emielinski1987] T. Emielinski. *Intelligent Query Answering in Rule-Based Systems*, volume 4. The Journal of Logic Programming, Intelligent Query Answering in Rule-Based Systems, 1987.

[Kolodner and Menachem1991] J. Kolodner and Y. J. Menachem. *Case-Based Reasoning: An Overview.* The Institute for the Learning Sciences, 1991.

[Leake1996] D. B. Leake. *Case-Based Reasoning: Experiences, Lessons & Future Directions.* AAAI Press - MIT Press, Cambridge, USA, 1996.

[Martins and Ferneda1999] A. Martins and E. Ferneda. The cbr fundamental problem: A case indexing model using credit scoring. In *ENIA 99*, 1999. AI National Meet.

[Martins *et al.*1996] A. Martins, B. Lula, and E. Ferneda. Computing analogies through logic. In *Proceedings of the XIIIth Brazilian Symposium on Artificial Intelligence - SBIA 96*, 1996.

[Tversky and Gati1978] A. Tversky and I. Gati. Studies of similarity. *Cognition and Categorization*, 1978.

[Watson1997] I. Watson. *Applying Case-Based Reasoning: Techniques for Enterprise Systems.* Morgan Kaufmann Publishers, Inc., 1997.