Enriched Knowledge Representations for Information Retrieval

F. N. TESKEY

Information Technology Research Institute
Brighton Polytechnic

ABSTRACT

In this paper we identify the need for a new theory of information. An information model is developed which distinguishes between data, as directly observable facts, information, as structured collections of data, and knowledge as methods of using information. The model is intended to support a wide range of information systems. In the paper we develop the use of the model for a semantic information retrieval system using the concept of semantic categories. The likely benefits of this area discussed, though as yet no detailed evaluation has been conducted.

1. Introduction

If information science is to justify itself as a science, then it must produce a scientific theory of information that can be tested and evaluated across the whole of the field of information science. Various theories of information have been proposed. At the very general level the entrophy model of information [17] has proved useful for communication engineers, and at the very specific level, the semantic models [4,8] have provided linguists with tools for analysing sentences. To date, these theories have not had any significant impact in the field of information retrieval. Here the main model is that of the document surrogate - a set of tokens, or frequency of occurence of tokens which have no explicit meaning. The lack of any suitable theory has meant that there have been few significant improvements in the quality of information retrieval, though there have been major improvements in the quantity. Since it is obvious that increasing the quantity of information retrieved will, in the end, be counterproductive, we must seek ways to improve the quality; hence we must look for a new model of information.

We would argue that the entropy model and the semantic model have failed to support information retrieval because they have concentrated on the medium, be it bits or words, rather on the message. It has been proposed [19,7] that we should use concepts from physics as a basis for building a theory of computer information systems. It seems likely that this could extend the scope of Shannon's definition of information, but it does not address the problem of how we perceive and use information. The information models used in some of the fields close to information retrieval, such as database management and expert systems, are based directly on our perception of the real world. They have, we would suggest, achieved a degree of success that has so far eluded information retrieval.

In a recent paper [16] van Rijsbergen has argued that "information retrieval can advance with new developments in formal semantics for text" and has suggested Montague semantics as a possible model. Montague semantics provide a formalism for representing the intension, or meaning of elements of a well defined set of expressions, which cover most, though not all, natural language sentences. This, van Rijsbergen proposes, can be used in computing the likelyhood that a document implies the answer to a question, and so should be retrieved. However, even if Montague semantics do not prove sufficiently powerful to compute the intension of a whole document, we can still postulate the existence of that intension. Indeed, we would argue that by basing our model on an underlying world model, rather than the semantics of the text, we can produce

a more manageable representation of the intension of a document. The following sections of the paper are devoted to developing such a theory, and showing how it could be applied to the field of information retrieval.

2. Knowledge, Information and Data

As we have moved from database management systems, to information retrieval systems, to knowledge based systems there has been no clear distinction between data, information and knowledge. We suggest that a clear distinction should be made based on the notion of data as directly observable facts, information as structured collections of data and knowledge as methods of using or modifying information. A further distinction needs to be drawn between user level and meta level. The ability to reason at the meta-level [6] has been identified as an important part of information systems, yet in many cases the meta-level knowledge and information used to control the system is not explicit. There is a need for an explicit and distinct meta level within the model.

Our aim is to develop a theoretical model for constructing information retrieval systems. The main requirements for the model are that it should provide a firm basis for a design methodology for constructing information systems, and present a conceptually simple interface to the user. In addition the model should support reasoning at both the user and meta levels.

2.1 Proposals for the model

Previous work on the development of the binary relational model (BRM) for information systems [2,23,24,25] has shown that this model can provide a powerful framework for building integrated information systems. The model has shown how structured databases and free text retrieval can be combined into a single model [20], how non-programmers can design and build their own information systems using powerful graphical interfaces [18,21] and how to combine user and meta-level models [1]. In this paper we discuss how the model could be extended to embrace more intelligent information retrieval systems. There are several aims for this. First, the binary relational model and graphical interfaces have been used to simplify the problem of non-programmers designing databases [18,21 and we believe that it could form the basis for a solution to the problem of designing advanced information systems. Second we believe that a firm mathematical basis for knowledge representation is necessary before we can develop rigorous methods for the design and development of advanced information systems. Finally we suggest that the future of IR lies in embedded systems, which will require a high degree of integration between conventional DBMS and novel AI systems; the BRM could provide the basis for such an integration.

We propose a model, based on the ideas of the BRM, which will support the integration of different types of knowledge representation in a single coherent framework. In the following sections we present a description of model which attempts to meet these requirements.

3. Components of the model

The system consists of three parts: the mathematical model defining the theoretical level, the interpretation of the model defining the user level and the implementation of the model defining the meta level. The basis structure of each of these three components is described below.

3.1 The mathematical model

An information model is a mapping

$$A : E \longrightarrow V$$

where E is a set of entities,
V is a set of values and
A maps each entity to the value attributed to it.

As in the standard entity relationship model [3] the entities are regarded as tokens or icons representing some abstract or concrete object or concept in the real world. The values, however, are not regarded as atomic, as in the original relational model [5] or in many subsequent developments [2,10] and others. Instead the value is regarded as an analogic representation of the related object in the real world. Thus in a simple case the value for an entity representing 'temperature' would be a single scalar value,for a more complex entity such as 'average temperature' the value would be a procedure, or analogue, for calculating the average temperature. It is important to realise that the mathematical model does not place any restrictions on the type of values that are allowed; the values can take the form of any valid mathematical object or formulae involving the elements

of the model. It is up to the implementation of the system to provide the tools necessary to manipulate complex values and it is up to the interpretation of the model to ensure that the values are meaningful.

For those familiar with category theory [13] this forms a simple category. As we add more structure to the model at the meta and user level, then the structure of this category becomes richer. A more rigorous definition of these levels could be made within the framework of category theory, though this is outside the scope of the present paper.

This model can be used to formalise the difference between knowledge, information and data, that we have already discussed informally. Thus the notion of data as directly observable facts is formalised by the definition of a data element as an entity and its associated value, and information, as a structured collection of data, is defined as a relation on E. Since these represent views of the real world, it is likely that the definitions could be rephrased in terms of physical concepts along the lines proposed by Stonier [19]. Our informal definition of knowledge referred to methods of using information. We would claim that the main use of information is to change one's model of the world; hence knowledge should be defined as a mapping from one model to another. The significance of these definitions should become clearer as the model is developed.

3.2 Interpretation of the model

The mathematical model that we have defined above provides no more that a framework for storing a disordered collection of data. The interpretation of that data requires some form of understanding of the real world the system is intended to model. Given that understanding, or intelligence, involves "perceiving order in a situation previously considered disordered" [9] then the simplest type of understanding that we can incorporate in the model will be to provide rules to order the entities into semantic categories or sets. In conventional IR, semantic categories are represented by index terms and lists of documents indexed by those terms. Since we wish to include some representation of the intension of documents, we must extend the concept of semantic categories. A semantic category can be specified either intentionally or extensionally. In the former case we need to give rules to define the intension of the category, in the latter case we need to give a list of occurrences to define its extension. In many cases it may not be possible, or even desirable, to give exact rules to define the category; we suggest that one possibiltiy is to separate the necessary and sufficient conditions for membership. This allows one to assert not only that an entity is or is not in a category, but also to state that we have insufficient evidence to decide whether or not it is a member (if it passes the necessary condition but failes the sufficient condition). This leades to the type of trichotomous logic proposed by [22].Thus we can define a semantic category as a triple:

(n,s,L)
where L is a list of entities to belong to the semantic category, and
n and s are predicates on E denoting the necessary and sufficient condition for an entity to be in the semantic category.

Note: Clearly S must imply n, also L need not contain all the entities that satisfy s, but only those entities that have been explicitly added to the category.

A collection of semantic categories represents a significant body of knowledge about the underlying information model. An individual semantic category can be regarded as entity in the model itself; the value of the entity being the triple (n,c,L). More importantly this can form the basis for producing a new information model based on the categories, rather than the individual entities. The information common to all elements of a category can be factored out and represented at the category level. Note that this satisfies our formal definition of knowledge as a mapping from one information model to another.

We can now consider the basic operations for semantic categories. Category membership means satisfying the sufficiency condition or inclusion in the list of known members, negation of category membership means negation of the necessity condition; note that these are not mutually exhaustive. This reflects the uncertainty enherent in much semantic information; in some cases we may not be able to say whether or not a specific entity belongs to a given semantic

category or not. One of the problems of developing information systems is the difficulty of dealing with the imcomplete and inconsistent knowledge that is required for qualitative, rather than quantitative reasoning [15] This model of semantic categories could provide a theoretical basis for work in this area. Since the method of representing uncertainty is introduced at the foundation of the model, rather than as an addition to existing models, such as [11,14] it should be possible to develop a more rigorous treatment of this area. Formally, if 'e' is an entity and C is the category (n,s,L) then

e IS-A C iff s(e) or e $\in$ L

In fact, we can define the relation IS-A as a semantic category by the necessary and sufficient condition for the entity 'c' to be in the IS-A category namely:

$$\exists\ e,S \in E\ :\ A(c) = (e,S)\ \&\ A(S) = (n,s,L)\ \&\ (s(e) \text{ or } e \in L)$$

Defining the sub-category operation is a little more difficult. Since the extension of each category need not be complete, there may be elements in the extension of the sub-category that are not in the extension of the parent category; further we can not assume that those elements will satisfy the sufficiency condition but only the weaker necessity condition of the parent category. Secondly, if an entity is definitely not in the parent category then if cannot be in the sub-category. Formally if C, and C_2 are categories then:

$$C,\ \text{IS-SUB}\ C_2 \text{ iff } s, \Rightarrow s_2\ \&\ n, \Rightarrow n_2\ \&\ e \notin L, \Rightarrow n_2(e)$$

Just as we defined the special relation IS-A as a semantic category, so we can define a general relation as a semantic category. This allows us to add necessary and sufficient conditions to relations in order to define their semantic content more closely. In particular we can now add constraints to relations in a much more natural way than was possible in the BRM. For example the relation HAS-PAY could be defined by a list of (EMPLOYEE PAY) pairs together with the necessary condition that pay is in the range 300 - 10,000. Regarding relations as semantic categories has an additional benefit, namely that we can easily create new relations based on the values of existing relations. A relation can be defined by specifying necessary and sufficient conditions. For example given two relations (PERSON WORKS-FOR COMPANY) and (COMPANY IS-IN TOWN) we can, as in [23] define a virtual relation (PERSON WORKS-IN TOWN) with a necessary and sufficient condition for the entity 'w' to be in WORKS-IN namely:

$$\exists\ p,c,t \notin E\ :\ p \text{ WORKS-FOR } c\ \&\ c \text{ IS-IN } t\ \&\ A(w) = (p,t)$$

To summarise, the concept of semantic categories has already been used to model information at both the meta level (the IS-A and IS-SUB categories) and the user level (HAS-PAY etc). We have also indicated how it can model simple procedural knowledge using virtual relations. The model will also support the more complex implicit relationa discussed in [20] any limitations are likely to be in the implementation. Finally we have shown how the model supports imcompleteness by rejecting the law of the excluded middle for category membership.

3.3 Implementation of the model

We need to consider two aspects of the implementation, the user interface and the support of that interface. The user interface is based on a visual display of the entity-value map. This provides an ideal conceptual basis for an interface for directed browsing round the information structure. The display consists of iconic representation of entities; when an icon, or entity, is selected then its analogic representation, or value, is displayed. This display may, in turn, contain further icons with their own analogic representation; these can also be displayed by selecting the required icon, and so on.

The implementation of the interface is based on a hierarchy of types. The values in the model can be grouped into types such as scalar, integer, real, vector, set, relation, tree etc. These types can then be combined into a hierarchy - integer and real are subtypes of saclar and so on. We will assume that there is a root, or base, type of which all other types are subtypes. Each of the types will have an associated set of actions, e.g. addition, multiplication etc. on scalars, union, intersection etc. on sets and so on. These value types and actions can be regarded as 'objects' and 'messages' in the

Smalltalk [12] and, as in Smalltalk the sub-types will inherit the actions of their parent types. The actions will consist of basic operations which are required for all value types, (though their implementation may differ from type to type) together with actions specific to the particular type. The basic operations which are required for all types are to create, display and delete entities and entity values. Examples of specific actions would be addition, multiplication etc. on scalars, union, intersection etc. on sets and so on.

All the entities in the model can be grouped according to the type of their associated value. Thus for each type there will be:

1. a set of entities (ENT), and
2. a set of actions associated with the value type of those entities (ACT).

Each of these types is itself an abstract concept with an associated value, namely the pair (ACT,ENT). Thus they can be represented as entities within the model.

4. Semantic IR systems

We have developed a model for representing data, information and knowledge in a single framework. The initial aim of the model was to provide a firm foundation for information science. If we now wish to evaluate the model we must see how well it can support particular tasks within the field of information science. So we shall now look at how the model could support advanced information retrieval.

An information retrieval system can be presented with three main types of requests for information. The first type - "What/Where/When is X?" - is, in our formal model, a request for data; X is some entity and the user wants to find its associated value. The second type "Tell me all about Y" - is a request for information, the user wants to find all data related to Y (recall we have defined information as a structured collection of data). Finally, the third type - "How do I do Z?" - is a request for knowledge; the user has an existing world model that he wants to modify to include Z. The first observation to make is that while these types of request have usually been dealt with by different types of system (database management systems, information retrieval systems and expert systems respectively) our proposed theory combines them into a single framework. For the rest of this section we will however, concentrate on the second, conventional IR, type of request.

In conventional IR systems documents are represented by collections of index terms. However, in a semantic information retrieval system we wish to represent documents by the information they contain. Since we have defined information as structured collections of data then we should represent documents by similar collections of data. Formally, a document can be regarded as an entity whose value is the contents of the document. We have seen that semantic categories can also be regarded as entities; thus we can represent the information in a document collection by a relation between "document" entities and "semantic category" entities. This, it will be recalled, corresponds to our formal definition of information as a relationship on E. To a first approximation, this can be regarded as indexing documents with semantic categories rather than index terms. Continuing this analogy, the hierarchy of categories defined by the IS-SUB relation corresponds to the usual broader-narrower term relation. There, however, the analogy stops. Semantic categories allow for the introduction of rules to determine category membership. Consider, for example, a category "expert system"; since it is unlikely that any document published before say 1950 deals with expert systems, we may wish to add a necessary condition to the category to say that the date must be after 1950. Similarly we may add a sufficient condition to the category to say that any document by Edward Shortliffe should be included. Further more, groups of documents could themselves form semantic categories using clustering based on citations, word frequency, etc., these would be handled in exactly the same way as categories based on index terms. This is similar to the use of extra external information proposed by van Rijsbergen [op.cit.] the difference being that we are including all the information in a single framework, and this, we suggest, will make it easier to calculate van Rijsbergen's conditional information measures.

Just as we regarded a document as an entity in our model, so too we can regard the user's question as another entity in the model. Like the documents, it can be

related to a number of semantic categories We can then ask whether or not a given document satisfies the necessary and sufficient conditions of the question. In general it is unlikely that a document would satisfy all the sufficient conditions of the question, but we could calculate a measure of the probability, p(d->q), that a document d implies (the answer to) the question, by a weighted score of the number of necessary and sufficient conditions that it passed or failed. Note that we can positively reject a document if it fails the necessary conditions.

In this section we have shown how the proposed model of information could be used as a basis for an advanced type of information retrieval system. We believe that the model could be used to develop other types of information systems, particularly knowledge based and expert systems. However, to assess the usefulness of the model it will be necessary to implement and test some of these systems.

5. Conclusions

We have just begun to explore the scope of a new mathematical representation of knowledge and information structures. The model provides all the facilities of the Binary Relational Model that has been used previously, and several new features. In particular the model allows:

1. an explicit distinction between the user and meta levels,
2. implicit support for modelling incomplete or inconsistent semantic information, and
3. a framework for a conceptually simple user interface.

A system is currently being implemented in Poplog on a SUN workstation. We hope to be able to report our progress on this at the conference.

References

1. Azmoodeh, M., "Automatic integrity rule processing in a binary relational database machine" in Proceedings of the first workshop on architectures for large knowledge bases, ed. S.H. Lavington, Alvey Directorate, London (1984)
2. Azmoodeh, M., Lavington, S.H., and Standring, M., "The semantic binary relationship model of information," in Research and Development in Information Retrieval, ed. C. J. van Rijsbergen, Cambridge University Press (1984)
3. Chen, P. S. "The entity relationship model," A.C.M. Transactions on database systems Vol. 1, pp.9-36 (1976).
4. Chomsky, N., Syntatic structures, Mouton, The Hague (1964).
5. Codd, E.F., "A Relational model of data for large shared data banks," Communications of the A.C.M.Vol. 1 13(6), pp.377-387 (1970).
6. Cunningham, J., "Comprehension by model-building as a basis for an expert system," in Expert Systems 85 Proceedings of the Fifth Technical Conference of the British Computer Society Specialist Group on expert Systems, ed. M. Merry, Cambridge University Press (1985)
7. Deutch, D., "Quantum theory, the Church-Turing principle and the universal quantum computer," Proceedings of the Royal Society Vol. A-400, pp.97-117 (1985).
8. Dowty, D.R., Wall, R.E., and Peters, S., Introduction to Montague Semantics, Reidel, Dordrecht (1981).
9. Fatmi, H.A., and Young, R.W., "A definition of intelligence," Nature Vol. 228(Oct.3), p.97 (1970).
10. Frost, R.A., "Using semantic concepts to characterise various knowledge representation formalisms: A method of facilitating the interface of knowledge based system components," The Computer Journal Vol. 28(2), pp.112-116 (1985).
11. Ganascia, J.G. and Kodratoff, Y., "Symbolic uncertain inference: a study of possible modalities," in Expert Systems 85 - Proceedings of the Fifth Technical Conference of the British Computer Society Specialist Group on Expert Systems, ed. M. Merry, Cambridge University Press (1985).

12. Goldberg, A. and Robson, D., Smalltalk-80: The Language and its Implementation, Addison-Wesley (1983).

13. Maclean, S., Categories for the working mathematician, Springer Verlag, New York (1971).

14. Mamdani, A., Efstathiou, J., and Pang, D., "Inference under uncertainty," in Expert Systems 85 - Proceedings of the Fifth Technical Conference of the British Computer Society Specialist Group on Expert Systems, ed. M. Merry, Campridge University Press (1985).

15. Merry, M., "Expert systems - some problems and opportunities," in Expert Systems 85 - Proceedings of the Fifth Technical Conference of the British Computer Society Specialist Group on Expert Systems, ed. M.Merry, Campridge University Press (1985).

16. Rijsbergen, C.J. van, "A Non-classical Logic for Information Retrieval", Computer Journal Vol. 29(6), pp 481-485 (1986).

17. Shannon, C.E., "Communication in the presence of noise", Proceedings I.R.E Vol. 37, pp.10-21 (1949)

18. Sharman, G.C.H., "Experience with a binary relational database," in Proceedings of the first workshop on architectures for large knowledge bases, ed. S.H. Lavington, Department of Computer Science, University of Manchester (1984).

19. Stonier, T., "What IS information," in Expert Systems 86 - Proceedings of the Sixth Technical Conference of the British Computer Society Specialist Group on Expert Systems Cambridge University Press (1986).

20. Teskey, F.N. "Information retrieval systems for the future", LIR 28, British Library, London (1984).

21. Teskey, F.N., Dixon, N., and Holden, S.C., "Graphical interfaces for binary relationship databases," Information Technology Research and Development Vol. 3(2), pp.67-77 (1984).

22. Teskey, F.N., and Fatmi, H., "A New Method of Processing Textual Material by Cybernetic Machine," Digital Systems for Industrial Automation Vol. 2(3), pp.243-266 (1984).

23. Tomlinson, A.M., "User views and the visual information system," M.Sc Thesis, Department of Computer Science, University of Manchester (1984).

24. Winterbottom, N. and Sharman, G.C.H., "NDB - Non-programmer Database Facility," TR-12-79, I.B.M. U.K. Hursley (1979).

25. Jiang and Lavington, S.H. "The qualified Binary Relational Model", Proc. 4th British National Conference on Databases, pp.61-79, Cambridge University Press, (1985).