Applied Informatics 2001, pp 295 - 300, Innsbruck, Austria, February 19-22, 2001, ISBN 0-88986-318-0.

FEATURES OF STRUCTURE FOR ANALOGY RETRIEVAL

BRIAN P. CREAN
Galway-Mayo Institute of Technology,
Castlebar Campus,
Ireland.

DIARMUID O'DONOGHUE
NUI, Maynooth,
Ireland.

Abstract

Spontaneously retrieving analogies from presented problem data is an important phase of analogical reasoning, influencing many related cognitive processes. Existing models have focused on semantic similarity, but structural similarity is also a necessary requirement of any analogical comparison. We present a new technique for performing *structure* based analogy retrieval. This is founded upon *derived attributes* that explicitly encode elementary structural qualities of a domains representation. Crucially, these attributes are unrelated to the semantic content of the domain information, and encode only its structural qualities. We describe a number of derived attributes and detail the computation of the corresponding attribute values. We examine our models operation, detailing how it retrieves both semantically related and unrelated domains. We also present a comparison of our algorithms performance with existing models, using a structure rich but semantically impoverished domain.

Keywords : Cognitive Modeling, Case-Based Reasoning, Analogical Retrieval

1 Introduction

Spontaneously discovering analogies forms the basis for many disparate cognitive processes; from identification to creative reasoning [1] and problem solving [2]. In this paper we present an alternative account for creative inspiration to the "serendipity" and semantic similarity arguments. To support this we focus on the structure of domain information rather than the contents.

Models of analogical reasoning are often sub-divided into distinct phases and a widely accepted model identifies five phases; *representation, retrieval, mapping, validation,* and *induction* [3]. The core mapping process involves the identification of large systems of 1-to-1 mappings between the given source and the target domains [4]. Identifying this structure mapping has been the subject of much focused work [3], [5], [6], [7]. However Veale et al. [8] have shown that this mapping task is NP-complete, being a variant of the largest-common sub-graph identification problem.

Given the complexity of a single mapping problem, the prospects for algorithms to discover the best structure-match from a large number of alternatives seems rather bleak. Undaunted, previous models of analogy retrieval [9], [10] use the semantic similarity between the given target and the required source as a basis for retrieval. Unfortunately, such approaches fail to retrieve the semantically distant domains that are often required for creative insight [1]. Analogy retrieval plays a central role in Case-Based Reasoning (CBR) [11], but differs from analogy in that semantic similarity is vital for the *within domains* reasoning that characterizes CBR. However, analogy is better characterized as *between domains* reasoning - necessitating the identification of semantically distant source domains. In this paper we consider the problem of selecting a suitable source analog from a large background memory, when presented with some target concept. Analogy retrieval episodes may occur when a partial target is supplied, or when a domain needs be to be re-described through a deep and novel comparison. In this paper we present a technique for performing accurate and efficient analogy retrieval that operates independently of the semantic primitives used to generate the domains description. We also introduce the RADAR (Retrieving Analogies with Derived AttRibutes) model, which is used to test this theory.

In this paper we investigate analogy retrieval from a perspective that is independent of the semantic content. We are not so interested in developing a cognitively accurate model of analogical retrieval, but rather in developing a mechanism capable of explaining and performing *structure-based* retrieval. By explicitly representing structural features, our retrieval technique makes use of the great diversity in structure that domain knowledge naturally assumes.

2 Retrieval Models

Analogy retrieval occurs immediately before the mapping task, and is thus the driving force behind spontaneous analogies. While analogies used to support directed (or supervised) learning are presented as source-target pairs, analogies underlying categorization and creativity are generally presented as target domain only cues. An appropriate source must then be retrieved dynamically based on the given target information. These candidate sources are then passed to the mapping and subsequent stages for further processing.

We now examine some existing models of analogy retrieval - and we shall show that thay are based on the presence of identifiable semantic overlap between domains. Significantly, they do not support retrieval of semantically unrelated analogs that are structurally identical .

First, the MAC/FAC model [9] augments domains in memory with normalized content vectors representing the occurrences of various predicates in each domain. The dot product of the targets content vector and each memory content vector is then computed in turn. The best selected domains are then passed to SME [5] and performs the expensive structure matching operations. The source domain with the highest structural match score is selected as the candidate source.

In ARCS [10] candidate selection is based upon predicate similarity (not identicality) as determined by the lexical relations like; synonym and hyponym derived from the lexical knowledge base WordNet [12]. The selected domains cause a parallel constraint satisfaction network to be constructed, where potential correspondences are explicitly represented as neurons - akin to ACME [6] supported by the same neural architecture of Grossberg [13]. Pragmatic influence can also be brought to bear upon this final domain selections process.

Holographic Reduced Representations (HRR) [14] incorporate semantic and structural similarity in retrieval through a combined vector representation. This expands the semantic representation to include role assignment information, thereby incorporating each objects role within the domain description. HRR's contextualise information by including role information in this representation. Attribute information is also included in this vector representation. Thus, HRR's favor the retrieval of semantically similar information that plays the same roles with domain description, but weakly supports retrieval of semantically similar (non-identical) information. The absence of semantic overlap between arguments prohibits HRR's from retrieving semantically disjoint source domains.

Feature selection is the primary mechanism for case retrieval in Case-based Reasoning (CBR). Through a myriad of feature selection techniques and similarity metrics, e.g. nearest-neighbour [15], indexing, [16] adaptive retrieval [17], cases are computed and passed to the adaptation stage. The most suitable cases are then manipulated through substitution, transformational rules or deviational mechanisms [18] in order to modify the retrieved cases' to the target case.

The reliance of these models on predicate similarity or identicality prohibits them from identifying semantically distant domains. This is a particularly unwelcome restriction when searching for a creative analogies, as these tend to be semantically distant. We now examine a way of describing a domains structure, independently of its semantic content.

3 The Problem Domain

Systematicity is founded upon large groups of identical (or near identical) knowledge structures being identified between a given problem and some source domain [4]. Retrieval of semantically distant source analogs therefore requires a form of structure based retrieval - one in which semantic similarity need play no part. To focus on the core issue of structure-based retrieval, we consider a problem where semantic overlap provides little assistance in identifying candidate sources. We firstly describe our problem domain and how retrieval is effected, then we address the application of our technique to domains typically found in the analogy retrieval literature.

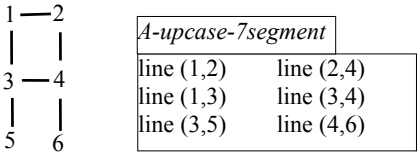

Figure 1

We consider analogy retrieval within a knowledge base composed primarily of descriptions of the alphanumeric characters, as represented on 7, 14 and 16 segment displays (see Figure 1), although some other information is also included. All reasonable valid formats were included, with upper case "A" represented both with a rectangular and a triangular top where possible. All domains are created from combinations of just one semantic primitive, denoting a semantic representation of a *line* occurring on the graphic image.

It should be pointed out that labels on the vertices are just markers representing when connections co-occur, or when the relevant virtual-lines intersect. These vertices correspond to objects of the domain description, and lines are represented as predicates. Thus, vertex one (1) representing the top-left corner could have any (possibly symbolic) label, it is the lines connected to this vertex that are of interest. Structural similarity is the focus of analogical reasoning, and thus the labels that are used as a vehicle to represent this structure are irrelevant. All vertex labels could be replaced with symbolic tokens without affecting the retrieval model or its performance.

This yields a knowledge base of just over 100 different cases; covering digits, capital and lower-case letters, and even different forms of some characters. Entries that are

structurally identical in more than one format are not repeated.

The retrieval problem we address then is, given a random case from this memory, to retrieve that same case from memory as well as all structurally isomorphic cases (for example 6 and 9 may be structurally identical, though inverted with respect to one another). We also include some cases that are semantically disjoint with the remainder of the knowledge base, to prove our theory.

Because of its semantic poverty, no retrieval episode can rely on semantic overlap with a domains contents to reduce the number of candidate sources. From a retrieval perspective, our knowledge base contains only structural information. To illustrate this point, let us consider how MAC/FAC, ARCS HRR and CBR would treat this retrieval problem.

3.1 Existing Models

The content vector of MAC/FAC would cause it to retrieve every domain when presented with a problem description. This is because the entire knowledge base employs just one semantic primitive, and occurs in every candidate source. The dot-product of content vectors would produce the greatest result for domains with a large number of predicates (lines) thus the same largest domains in memory would be retrieved irrespective of the target probe. Of course, it should be pointed out that this algorithm was not designed to deal with this scenario, however it does serve to highlight some limitations.

The semantic similarity component of ARCS provides no addition to retrieve more distant domains - even if the required knowledge were provided. The ARCS retrieval mechanism would construct a large parallel constraint satisfaction network encompassing every domain in memory. But given the convergence problems suffered by ACME on large networks [8], we can expect even greater convergence problems on such a behemoth of a neural network.

HRR's encode each predicates role (agent or patient) binding with unique identifiers, each being assigned a unique random number. Thus the agent role of object "1" in the line predicate is uniquely identified, for all its occurrences in the domain. Retrieval accesses two different types of candidate. First structurally identical domains that generate a high similarity score. Second, structurally different domains with low identifier values for the non-identical portions. Any domain with all high role-binding codes will tend to be more similar to other domains, that the same domain with low identifier values.

CBR primarily centers on single problem domains so cases are broken into superficial, yet distinguishable features, of a domain. CBR generally treats Retrieval and Adaptation as separate entities, hence the most adaptable cases are selected based on a numerical computation of the highest number of common features. Adaptation can change parameters or apply abstraction\transformational inferences to cases, but only within the context of that single domain. CBR is not concerned about the structural attributes of a case and is heavily influenced and dependant on suitable indexes used to retrieve adaptable cases.

4 Derived-Attributes

In this section we present our solution to the problem of structure based domain retrieval. This solution is based on a number of *derived structural attributes* that we compute from the problem domain information of Figure 1 above. Derived attributes represent features of the representation data itself, rather than qualities directly related to the real-world. Crucially, these derived attributes are independent of the semantic primitives used in describing the domain. Our model was partly inspired by Tversky [19], who examined the semantic similarity between concepts in terms of feature overlap. We describe simple structure features for the current problem, but they may also be derived for domains like "the universe" and "the atom" [4] from the analogy literature. We consider generating these attributes as a function operating upon problem data, and creating a multi-dimensional *derived attribute space*. Each attribute type is treated as independent axes, and the attribute values as points along the corresponding axis. All domains have their corresponding derived attributes stored with the base data, as is depicted in Figure 2. This supplementary domain information is reminiscent of MAC/FAC content vectors.

Firstly, the number of objects in a domain gives us some structural information, as does the number of predicates. These derived attributes help identify similarly sized domains for matching. Domains in this knowledge base contain an average of just under five predicates. Due to the inferior role of attributes in powerful analogies, these are not represented in this derived attribute space.

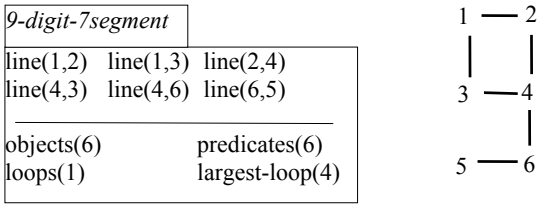

Figure 2

Particularly useful within the alphanumeric knowledge-base is the identification of loop structures where objects are re-referenced in a cyclic manner. Of course such constructs can and do occur in other domains also, such as the following.

```
taller (abe, bill)
taller (bill, con)
shorter (con, abe)
```

This loop construct involves three predicates, but larger loop constructs also occur, and the largest identified loop construct yields further useful structural

information. The largest loop size we check for in the data described herein is ten.

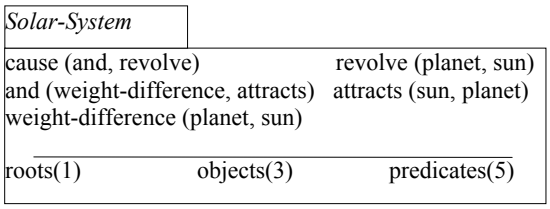

Figure 3

We also identify the number of root predicates occurring within the domain description (Figure 3). These identify predicates that are not arguments to any other (higher level) predicates, typically playing the role of controlling causal relations in a domain. Root predicates play a central role in incremental mapping as performed by models such as IAM [3] and SME [5], which focuses on the cognitive plausibility of mapping models.

4.1 Derived Attribute Retrieval

We now describe the algorithm that uses these derived attributes as a basis for efficient structure based retrieval. The premise of our retrieval algorithm is that domains with the same derived attributes must also have identical, or near identical, structure. It also uses a spreading activation mechanism to accumulate evidence and thereby identify structurally similar candidate sources.

1. Determine derived attribute values
2. Spread activation to attributes and container domains
3. Sort active frames on activation strength
4. Select all domains with equal highest activation level as equally good candidate sources.

A minimum activation level can also be used, accepting only sources with activation equal to that of the target domain, only domains judged structurally identical to a given target will be considered a candidate source. Only direct mapping can distinguish between domains with identical derived attribute values, but which have a (slightly) different structure. Our retrieval algorithm identifies structurally identical and similar domains, while the detailed comparison between domain entities is performed by a separate mapping process - in a manner similar to MAC/FAC.

4.2 Retrieval Properties

Clearly, within the described problem domain there is a significant amount of ambiguity in how data can be interpreted. This is a desirable property as we would like to be able to retrieve all isomorphic sources. We illustrate this point by noting the structural similarity between the digits "6" and "9", when stored as seven segment display data. Critically, we have not included some vital information that supports disambiguation between these digits. This might be performed by inclusion of contextual information - *top (line 1,2)*.

Now let us assume that the target information has been labeled entirely differently - but of course the same structural information is present. Consider then the situation that our memory includes information only on digit 9, but information representing digit 6 is presented as the target domain. This will cause the computation of the corresponding derived attributes, and will be identical to those of digit 9. Although many concepts overlap on some structural attributes, only the digit "9" overlaps on all, becoming the equal most active concepts in memory. Any mapping algorithm can take the given target domain and the new candidate source, to form a structurally identical inter-domain mapping.

Consider the domain described in Figure 4, and the accompanying diagram that highlights its structural identicality with the digit 9, from Figure 2. Both domains have the same structural attributes, and thus presentation of either will cause the retrieval of both - in the absence of semantically disambiguating factors.

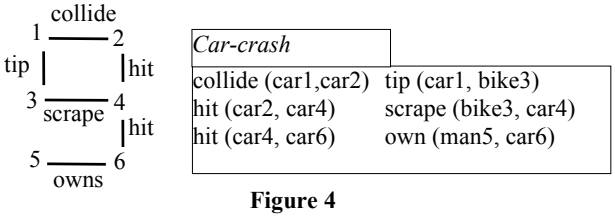

Figure 4

Thus, analogical retrieval is performed in the absence of an identifiable semantic intersection between the source and target domains. Note, that if some semantic information was contained in these domain descriptions, it is not the semantic information but only its structure that is necessary for analogical reasoning. Semantic content of domains has no bearing upon the identification of a useful analogy.

We readily admit that semantic similarity can make analogies easier for subjects to interpret. By adding a semantic similarity component to the retrieval mechanism, we can bias the structural retrieval with a small degree of semantic similarity information, thus favoring domains that are more natural or easy to interpret. This could be easily achieved by combining our derived attribute retrieval with the content vectors employed by MAC/FAC to favor retrieval of semantically similar domains.

Note that adding a semantic component does not prohibit retrieval of semantically distant concepts - it merely biases selection towards local domains. This however, may be an undesirable property were this technique used as part of a "creativity engine".

Our focus in this paper is not upon modeling how people perform the retrieval process, but rather on creating a computational model that is capable of performing domain independent structure based retrieval, and does so in a computationally tractable manner.

5 Potential of Derived Attribute Retrieval

Domain retrieval using derived attributes is only as efficient as the derived attributes that supplement the raw domain information. Again, we treat each derived attributes type as an axis and its values as points along each axis. Taking just five attribute types each with just 10 values, and making the best case assumption that our data is distributed evenly along each value, then each location in derived attribute space would represent just 10 domains, for a base of 1,000,000 domains.

This indicates the potential retrieval power of derived attributes. Of course efficiency is increased with additional attribute types and values describing new structural qualities, that may be peculiar to a given problem domain. Additionally, the underlying semantic information may also be used to select between retrieved domains.

6 Results

In this section we compare the retrieval performance of our algorithm with those of MAC/FAC, ARCS, HRR and CBR. We used our alphanumeric knowledge base for the following experiments as this provided a structure rich set of domains. As previously mentioned, this knowledge base contains over 100 concepts with an average of less than 5 predicates per case. Only the alphanumeric concepts were stored in memory for the following experiments. Each domain in turn was used as a retrieval cue from the knowledge base, recording the number of retrieved concepts for each retrieval episode.

MAC/FAC and ARCS both retrieve all cases from memory for each individual retrieval problem - vastly over-generating the number of required sources. HRR's also retrieve every previous case, although some will have higher retrieval values due to the role-filler identifiers. CBR dependence on feature identification and appropriateness of features will still result in an overwhelming number of all cases being retrieved.

RADAR then retrieved an average of just 5.6 cases for these retrieval problems. Most of these retrieval problems identified just one or two cases, but a few retrieved many more than this. RADAR significantly outperforms the other models in its ability to select past cases based solely upon the structural similarity between the presented target and all previously stored domains.

6.1 False-positive Retrieval

For each target driven retrieval episode, we wished to retrieve only one (correct) source domain. However, we point our that many of our false positives were structurally identical - 6 and 9 for example. The retrieval of structurally non-identical concepts is caused by several characters being encoded identically in derived attribute space. Note also that these characters are represented similarly in a number of different display formats.

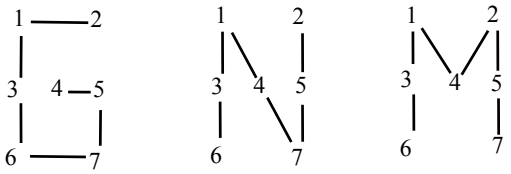

Figure 5

Cases form random clusters in derived attribute space, and the character identification problem is no exception. The largest cluster identified contains the letters "G, N, M" in several formats for a total of 8 cases at this location in derived attribute space - see Figure 5. This is caused by having the same number of lines with the same connectivity for many instances of these letters. This one cluster alone contributed greatly to the retrieval accuracy not reaching an even greater degree of resolution. However, as because many letters are represented similarly in different display formats, many smaller clusters were also identified.

6.2 Semantically Disjoint Retrieval

To further highlight the semantic independence of our model, Table 1 highlights its ability to performs semantic distant retrieval. Assume the target domain is the digit 9, as shown in Figure 2. Again assume our memory contains the alphanumeric domains, (including the digit 6) plus the "car crash" domain from Figure 4.

Model	Domain	
	No. 6	Car-crash
ARCS	Y	N
MAC/FAC	Y	N
HRR	Y	N
CBR	Y	N
RADAR	Y	Y

Table 1

Most significantly, we point out that only our model retrieves semantically unrelated domain. Thus, an analogy for a given domain may come from any area in memory.

7 Conclusion

We examined the problem of identifying structurally identical source analogs when presented with some target domain, represented as predicate calculus assertions. Avoiding the domains semantic content, we map the problem domain onto a derived structural attribute space, that serves as a retrieval index into the corresponding knowledge-base. Thus, we effect retrieval based upon features of each domains structure, rather than from its semantic content. Empirical comparison with existing models were given, as well as examples of our models ability to retrieve semantically disjoint source domains.

References

[1] Boden, M. A. "*The Creative Mind*", Abacus, 1994.

[2] Keane, M. T. "Analogical Problem Solving", Chichester: Ellis Horwood, UK, 1988.

[3] Keane M. T., Ledgeway T. & Duff S. "Constraints on Analogical Mapping : A comparison of three Models", *Cognitive Science,* Volume 18, 1994, 387 - 438,.

[4] Gentner, D. "Structure-Mapping: A Theoretical Framework for Analogy", *Cognitive Science*, 7, 1983, 155-170.

[5] Falkenhainer B., Forbus K., Gentner D. "SME: The Structure Mapping Engine", *Cognitive Science*, 1992

[6] Holyoak, K.J. Thagard, P. "Analogical Mapping by Constraint Satisfaction", *Cognitive Science*, 13, 1989, 295-355,

[7] John E. Hummel & Keith J. Holyoak "Distributed Representation of Structure: A Theory of Analogical Access and Mapping", *Psychological Review*, 104, 3,1997, 427-466.

[8] Veale T., O'Donoghue, D. and Keane M., "Computability as a limiting cognitive constraint : Complexity concerns in metaphor comprehension about which cognitive linguists should be aware", *Cultural, Psychological and Typological Issues in Cognitive Linguistics*, Ed. M. Hiraga, C. Sinha and S. Wilcox., John Benjamins Publ. Amsterdam/Philadelphia. pp 129-155. (1999). From the series "*Current Issues in Linguistic Theory*" (CILT), Ed. E.F. Konrad Koerner.

[9] Forbus, K. Gentner, D. Law K. "MAC/FAC: A model of Similarity-based Retrieval", *Cognitive Science*, 19, 2, 1995.

[10] Thagard, P. Holyoak K. J. Nelson, G. Gochfeld, D. "Analogue Retrieval by Constraint Satisfaction", *Artificial Intelligence*, 46, 1990, 259-10.

[11] Riesbeck C. Schank, R. "*Inside Case-based Reasoning*" Erlbaum, Hillsdale N.J. 1989.

[12] Miller. George A. "WordNet: A lexical database for English." *Communications of the ACM*, 38(11), 1995, 39—41.

[13] Grossberg, S. "A theory of visual coding, memory and development" in E. Leewunbrg and J. Buffartt Eds. Formal Theories of Visual Perception (Wiley, NY, 1978).

[14] Plate, T. "Structured operations with Distributed Vector Representations" in Holyoak, K. Gentner, D. Kokinov, B (Eds.). "Advances in Analogy Research : Integration of Theory and Data from the Cognitive, Computational and Neural Sciences", New Bulgarian University, Sofia, Bulgaria, July, 1998

[15] Okamoto, S and Satoh, K, "An average-case analysis of K-nearest neighbour classifier". Case-based Reasoning Research and development, Veloso, M., and Aamodt, A. (Eds.). *Lecture notes in Artificial Intelligence 1001*, Springer-Verlag, Berlin 1995.

[16] López de Mantáras R. and Plaza E. "Case-based reasoning : an overview", *AI Communications*, 10, 1997, 21 –29.

[17] Smyth, B. and Keane, M. T. "Adaptation-Guided Retrieval: Questioning the Similarity Assumption in Reasoning", *Artificial Intelligence*, (102) 2 , 1998, 249-293.

[18] Kolodner, J. L. "*Case-based Reasoning*", Morgan Kaufmann, 1993

[19] Tversky, A. "Features of Similarity", *Psychological Review*, 84, 1997, 327-352.