Analogy and Deduction for Knowledge Discovery

Jim Reynolds[†] Adam Pease[‡] John Li
Teknowledge Corporation
1800 Embarcadero Road
Palo Alto, CA 94303

Abstract

Analogy-based hypothesis generation is a promising technique for knowledge discovery. However, some hypotheses generated are nonsensical. This paper describes a two-phased method to increase the quality of analogy reasoning. The first phase employs an established approach to generate hypotheses through similarity matching. The second phase utilizes deductive reasoning to eliminate hypotheses that are clearly false or absurd. The basis for elimination is violation of common sense or domain knowledge, which is represented in a suite of ontologies. We describe a set of preliminary experiments conducted to validate this two-phased approach. The experiments involved much larger test cases than reported by any other analogy researchers, and the results are very encouraging.

Keywords: knowledge discovery, analogy, deductive reasoning, ontology

1. Introduction

Analogy can be a method for knowledge discovery through the generation of novel hypotheses. Unfortunately, the output of analogy may include false or absurd hypotheses. This paper describes a two-phased approach for improving the output of analogical reasoning. First, an analogical engine uses structure mapping to generate hypotheses. The second step employs deductive reasoning over a large ontology to identify and remove preposterous hypotheses. By applying both upper-level and domain-specific knowledge, our tool eliminates many, if not all, false or nonsensical hypotheses. We believe this is a novel application of deduction with analogy.

The emphasis of the most prominent analogy research groups to date has been less on hypothesis generation than on determining if one situation is more analogous than another to a given situation of interest. The two longest established and most prolific groups are those at Northwestern University, led by Ken Forbus and Dedre Gentner [3,4,5,6,7,8,9,11,12,13,15,16,17,27], and at the University of Waterloo, led by Keith Holyoak and Paul Thagard [2,18,19,20,21,22,23,24,25,31,32,31]. Despite early disagreements between the groups, their points of agreement are more significant [14]. The Northwestern group has developed the view of analogy as a computational tool further. Initially, following Gentner's theory, which emphasized structure, they developed a Structure Mapping Engine (SME). Because structure mapping is an expensive computation, they added a preprocessing stage (MAC, or "Many Are Called") that amounts to a content filter applied before SME is used to rank candidate analogies (FAC, or "Few Are Chosen").

[†] Jim Reynolds is currently with General Dynamics Corporation.

[‡] Adam Pease is currently an independent consultant in the San Francisco Bay Area.

The culmination of nearly fifteen years of work in the field is presented in "An Analogy Ontology for Integrating Analogical Processing and First-Principles Reasoning" [10]. This paper describes two-way communications between a reasoning system and analogy software. However, the architecture employs the analogy software only as an add-on to a deductive inference engine with knowledge base (ontology). The analogy software uses the ontology to distinguish the categories of terms (e.g., relation, attribute, or logical connective) and to store and retrieve collections of terms as cases. We have gone far beyond this in using deductive reasoning and an ontology to filter the false hypotheses generated by analogy, thus freeing the information consumer to concentrate on the most meaningful suppositions.

Section 2 presents our algorithm informally, with a simple example to illustrate how the algorithm works. Section 3 describes the deductive engine and ontologies essential to the filtering phase of our approach. Section 4 reports significant experiments that demonstrate great promise for our approach to knowledge discovery. Section 5 contains our conclusions to date and directions for future research.

2. Analogy

2.1. Algorithm

Our algorithm approximates the algorithm developed by the analogy research group at Northwestern. On this approach, analogy compares two descriptions, taking one as the target, another as the base.[1] The algorithm attempts to match each sentence in the base file with similar sentences from the target. Similarity is determined by three constraints:

Identicality. Only expressions with identical outermost predicates are matched.

One-to-one. Substitutions of objects from the target for objects in the base must be one-to-one.

Recursive matching. Two sentences match only if all their arguments match recursively.

With those constraints in mind, our algorithm can be summarized as follows:

1. Create a list of two-element lists from the base and target files. The first element is a unique sentence from the base and the second element is a list of matched sentences from the target. Matched sentences must contain the same outermost predicate. Sentences from the target are reused to try to match sentences from the base. A base sentence that fails to be matched is paired with an empty list.
2. For each two-element list, try to find one sentence from the list of target sentences that matches the base sentence according to the second and third constraints described above. If such a match is found, create a table of possible substitutions between structurally corresponding terms in the base and target sentences.
3. For each unmatched base sentence, try to create a new sentence using the structure of the base sentence and its outermost predicate, but with all other terms coming from the target by the possible one-to-one substitutions.
4. Output a list of sentences from steps 2 and 3. This is the list of analogical sentences. Any sentences from step 3 are candidate hypotheses.

Our analogical engine is implemented in Prolog.

2.2. Example

A simple example [2] illustrates the functionality of the algorithm. The example compares the situation in the Middle East just prior to the Gulf War ("Target") with two analogous descriptions of the US-

[1] In this paper we will use the term "description" to mean a set of formal sentences contained in a file. The hypotheses generated will be formal sentences from the base about the target.

[2] The example is loosely based on an example from Holyoak and Thagard [22].

European situation during World War II ("Bases"). The target ("Current Case") has five statements, which are represented graphically in the left column of Figure 1. The sixth statement is the hypothesis generated.

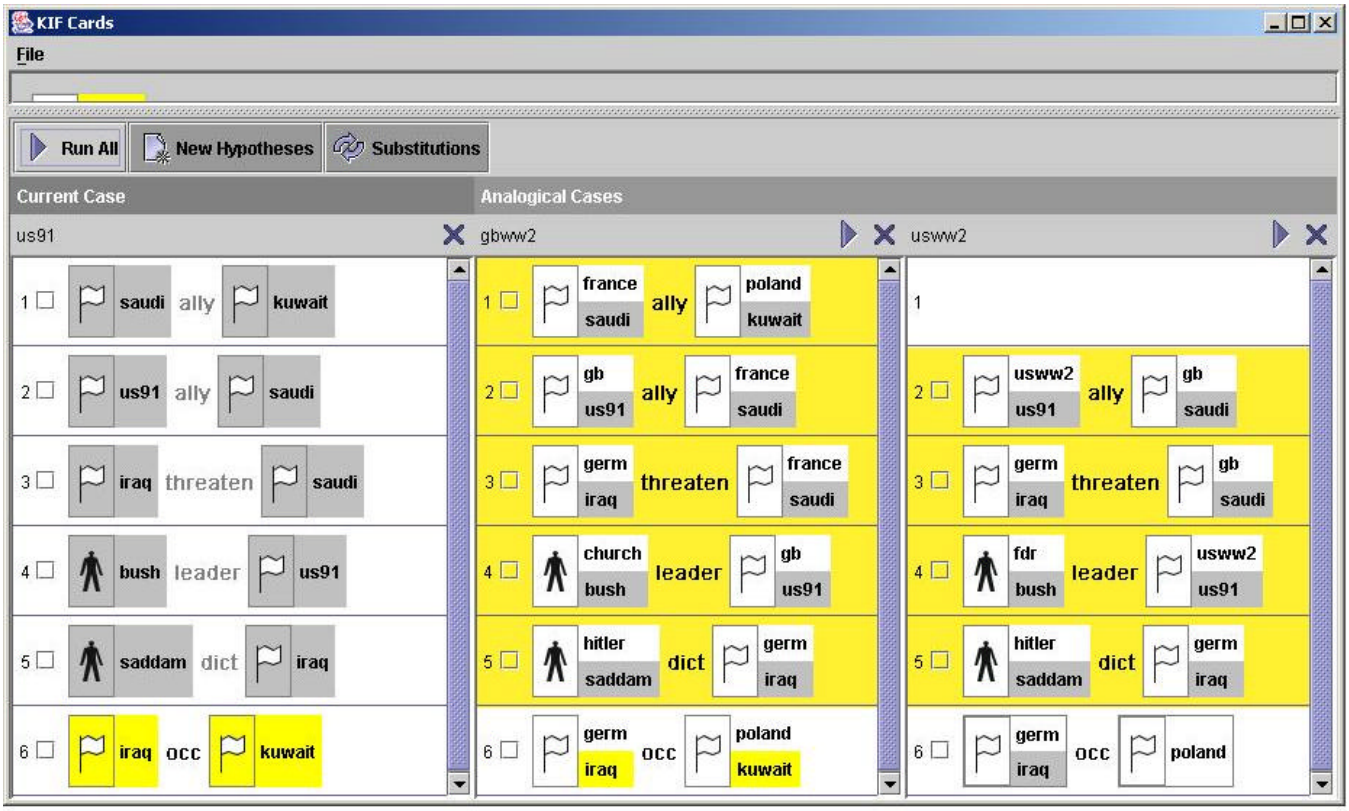

Figure 1. Hypothesis about the Gulf War drawn from World War II[3]

The meaning of the five sentences of the Current Case can be stated in English as follows:

1. Saudi Arabia is an ally of Kuwait.
2. The US is an ally of Saudi Arabia.
3. Iraq threatens Saudi Arabia.
4. Bush is the leader of the US.
5. Saddam Hussein is dictator of Iraq.

The two base situations ("Analogical Cases") have six and five statements respectively. The first base situation, shown in Figure 1, can be stated as follows:

1. France is an ally of Poland.
2. Great Britain is an ally of France.
3. Germany threatens France.
4. Churchill is the leader of Great Britain.
5. Hitler is dictator of Germany.
6. Germany occupies Poland.

The five sentences in the second analogical case can be interpreted similarly.

[3] The screen shot is the output of visualization software from Pascale Proulx and William Wright of Oculus (www.oculus.com). Their software graphically displays the textual output of the analogical software.

Figure 1 displays the analogous sentences and substitutions that were found in the example. Matched sentences are aligned in the rows. Paired terms associated with icons show the one-to-one substitutions of objects from the target to the base that would make the analogy consistent over all the sentences. For example, "Saudi Arabia is an ally of Kuwait" in the Current Case is analogous to "France is an ally of Poland" in the first Analogical case, when "Saudi Arabia" is substituted for "France" and "Kuwait" is substituted for "Poland". If this substitution is used to make the first sentences analogous, then "Saudi Arabia" must also be substituted for "France" in the third sentence of the middle column.

The analogous sentences in the base cases appear in colored boxes. The screen shot reveals which description about World War II is more analogous to the United States during the Gulf War. There are five analogous sentences from the first base description but only four analogous sentences from the second base situation. Finally, the screen shot shows that one new hypothetical assertion was generated about the target from the first base case: "Iraq occupies Kuwait." The original target case contained no assertions about occupation. This discovery could be considered a prediction (if it had not already happened) or might represent missing information that was not included in the original target description. The hypothesis could not be generated from the second base case because there was no consistent substitution for Poland available in that description.[4]

2.3. Problems with Hypotheses from Analogy

The analogical algorithm produced a plausible hypothesis for this simple example, but two shortcomings should be immediately suspected. First, in "toy" descriptions containing only a handful of sentences and objects, semantically blind substitutions based solely on the sentence structures may have a good chance of making sense. As the number of sentences and scope of descriptions increases, the algorithm may generate nonsensical substitutions, undermining the usefulness of a tool based solely on structure. Second, the matching of predicates in outermost position is entirely dependent on their name strings. This may result in useful substitutions with some terms while missing intelligent substitutions based on the meaning of the predicates. In our example, "Bush" is substituted for "Churchill" and "FDR," because all those terms appear with the predicate "leader." But no pairing is made with the term "Saddam," which appears in the relation "dictator," even though "dictator" is a specialization of "leader."

The solution to the first problem is to subject potential substitutions to a constraint and consistency "sanity check" by testing the analogous sentences or hypotheses over an ontology using deductive rather than analogical reasoning. We leave the solution of the second problem to another paper.

3. Deduction and Ontology

The second phase of our method involves deductive filtering of the candidate hypotheses. Teknowledge has developed a knowledge engineering environment called Sigma [31], in which a Knowledge Base (KB) can be constructed from constituent ontologies. Sigma also links KBs with reasoning tools. For knowledge representation, we use a variant of the Knowledge Interchange Format (KIF) [26] adapted for the Standard Upper Ontology group (SUO-KIF). Our ontologies are grounded in the Suggested Upper Merged Ontology (SUMO) [28]. We have also created a midlevel ontology and a suite of domains used in geopolitical analysis (e.g., Geography, Government, WMD, etc.). For deduction, we employ the Vampire first order theorem prover [32,34], combined with techniques for knowledge

[4] In addition to the Gulf War/World War II example shown in section 2.2, we also tested our algorithm against the three "Karla" stories that were the centerpiece of the seminal work by the Northwestern group [13]. Despite the superficiality of two of the stories being about birds, the story about Karla and the country of Zerdia were determined to be the most analogous, based on structure. Our analogical engine came to the same conclusion and made the substitutions as described by the Northwestern group.

compilation that express complex expressions in standard first order language. The theorem prover tests each candidate hypothesis against a KB containing both general and specialized ontologies. If the hypothesis is proved, then it was logically implicit in the KB, although it may not have been obviously true. If the negation is proved, then our analogy program discards the false hypothesis.

4. Experiments

Our test cases involved analogies between pairs of countries. For our experiments, we used data files translated from the HTML pages of the CIA World Factbook (WFB) [1]. The WFB has almanac-style pages of semi-structured textual information for every country in the world. Topics include national government, geography, economy, people, and other subjects of geopolitical interest. There is a core similarity in what is covered. For example, nearly every country has population and per capita income data. However, there are some differences due to varying national situations (e.g., election activities, drug production), and some aspects are described more completely for some countries than others (e.g., economic trade, legal system). Our KIF country files consisted of about 500 formal sentences each (with a variation in size of 50 percent). A KB containing the selected country files together with SUMO and relevant domain ontologies contained roughly 20,000 terms and 65,000 assertions, including 2,500 rules. On a laptop computer operating at roughly 500 MHz with 256 MB of memory, the analogical algorithm took approximately a minute to compare two files and generate hypotheses.

Figure 2 shows the Sigma browser interface to one screen of the analogy tool. The controls on the page implement our two-phase approach. Before this point, the user selected two files, a base and a target,

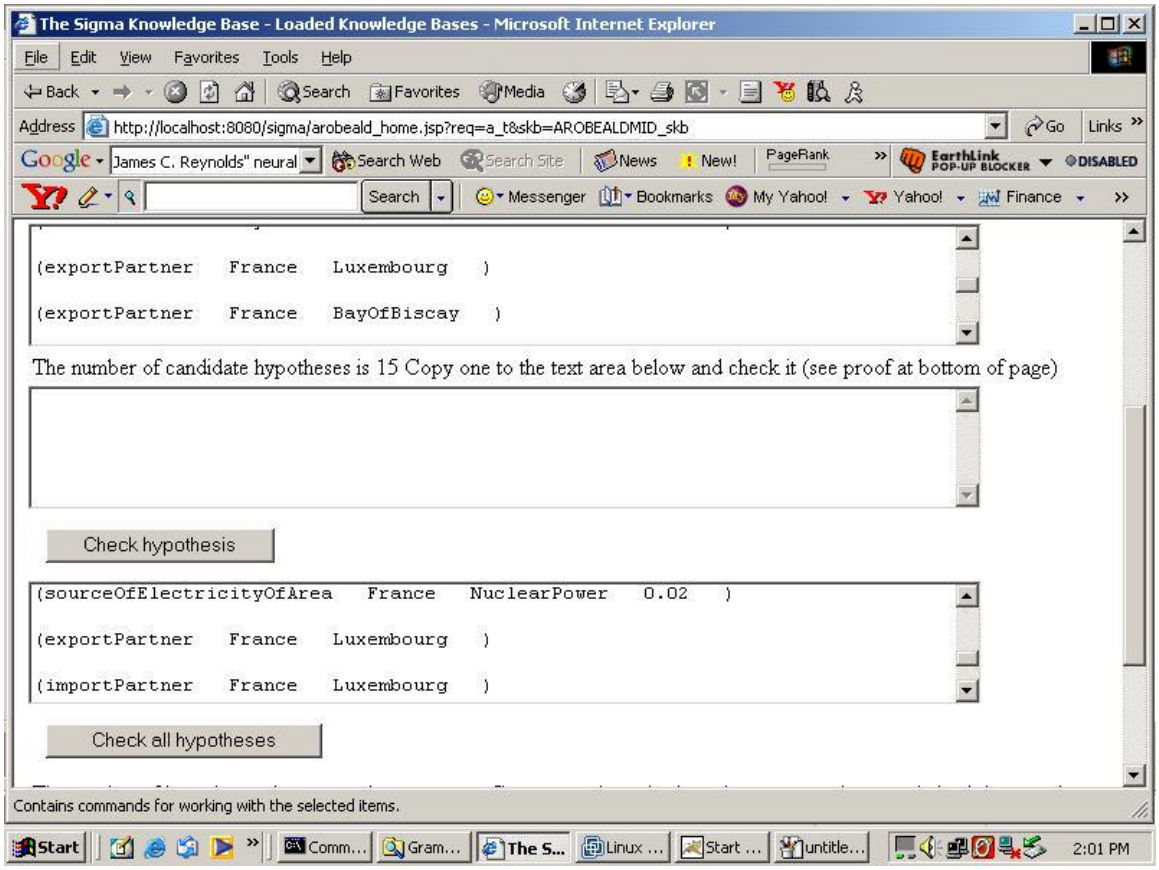

Figure 2. Sigma interface to hypotheses generation

and clicked a button labeled "Generate Novel Hypotheses" (not shown). The result of that action was the list of "Candidate Hypotheses" displayed in the first text area in Figure 2. Fifteen candidate hypotheses were generated. The hypotheses are sentences about the target file (France), based on sentences that were present in the base file (Afghanistan) but not in the target file. They were produced by the Prolog analogical engine, which does not use the theorem prover or ontologized knowledge.

The results fall into three categories:

1. *True sentences.* These are facts, but they may be missing from the original target file. In Figure 2, an example of this type is that France is an export partner with Luxembourg.

2. *Untrue but plausible (or possible) and interesting sentences.* One example is that France was suspended from the European Union. This is a possible and interesting hypothesis, although not presently true. It was generated because the base contained the sentence that Afghanistan was suspended from the International Olympic Committee. The hypothesis about France is an example of something a knowledge analyst might investigate, depending on his interests.

The above categories provide examples of knowledge discovery derived from analogical reasoning.

3. *Untrue and absurd sentences.* Some candidate hypotheses are preposterous, for example, that France is an export partner with the Bay of Biscay or the English Channel, or an import partner with those bodies of water, or is an illicit drug producer of machinery and transportation equipment.

The second phase of our approach is activated by the button labeled "Check all hypotheses." This initiates testing the truth of each candidate hypothesis using the theorem prover and an ontology of upper-level concepts and domain-specific knowledge. After disproved hypotheses are discarded, the surviving hypotheses are displayed in the third text area. In the case of Afghanistan compared to France, the original fifteen hypotheses are winnowed to ten. The five preposterous hypotheses involving bodies of water being export and import partners are discarded based on reasoning about predicate constraints combined with subclass information. A proof may be requested by copying a hypothesis to the middle text area and calling the deductive engine to "Check hypothesis." For example, the discarded hypothesis "(exportPartner France BayOfBiscay)" has the following (dis)proof. For simplicity, we list only the most important steps in the deduction. Vampire derives (5) from (1) - (4) in the KB, and derives (8) from (6) and (7). Note that (7) is the negated query, "(not (not (exportPartner France BayOfBiscay)))," which the prover assumes is correct (for Indirect Proof). Finally, it reaches (10) by combining (5), (8) and (9). Since all facts from the KB are correct, the False conclusion (10) must come from the false assumption (7).

1. (instance BayOfBiscay Bay) [KB]
2. (subclass Bay Inlet) [KB]
3. (subclass Inlet BodyOfWater) [KB]
4. (subclass BodyOfWater WaterArea) [KB]
5. (instance BayOfBiscay WaterArea) [1, 2, 3, 4]
6. (domain exportPartner 2 Agent) [KB]
7. (exportPartner France BayOfBiscay) [Negated Query]
8. (instance BayOfBiscay Agent) [6, 7]
9. (disjoint WaterArea Agent) [KB]
10. False [5, 8, 9]

We applied our two-phase process of analogical hypothesis generation and deductive filtering to three other comparisons from the CIA World Factbook. Table 1 summarizes the results in terms of total hypotheses generated, number of preposterous hypotheses discarded and not discarded, and examples of interesting hypotheses and preposterous hypotheses. Our experiments have demonstrated the following:

1. The analogical procedure can generate at least some interesting and/or plausible hypotheses from the comparisons in all cases.
2. Deductive reasoning using an ontology can reduce the number of preposterous hypotheses in all cases.

Together, this constitutes an unprecedented breakthrough for knowledge discovery from analogy.

Comparison	Hypos	Bad Filtered	Bad not Filtered	Interesting/Plausible (examples)	Discarded as Preposterous
Afghanistan-France	15	5	0	France is suspended from the EU. Spain is export partner of France.	English Channel is an export partner of France.
Afghanistan-Tajikistan	9	1	0	Tajik is an export partner of China. Alay mountain range is located in Tajik.	Tajikistan participates in a mountain range.
France-Tajikistan	22	3	2	Tajik is a consumer of opium and heroin. Various real places are located in Tajik.	The Legislature of Tajik is its national day.
France-US	32	2	1	Mexico is to the south of the US. Various real places located in US.	US participates in a strait.

Table 1. Effectiveness of hypothesis generation by analogy and filtering by deduction and ontology

5. Conclusions

We have demonstrated that it is possible to use analogy – defined as a process of finding a purely structural alignment between situations – to generate hypotheses for knowledge discovery. The hypotheses may be new facts that contain missing information or they may be speculative assertions worthy of further investigation. The problem of nonsensical hypotheses can be partly solved by filtering the candidate hypotheses using deduction with general and domain-specific knowledge encoded in an ontology. This is an exciting development. The effectiveness of the basic two-phased approach is clear. It can be extended in numerous ways. Critical issues for making analogical reasoning practical include how to use deduction and ontology to guide the analogical matching process as well as using deduction to further vet the results of analogy. Desirable features include the ability to request certain substitutions and disallow others interactively as the analyst explores what can be learned from analogy. We are encouraged by the results reported here to continue our research developing analogical reasoning for practical knowledge discovery.

6. Acknowledgement

This work was supported by the Advanced Research and Development Activity, Novel Intelligence from Massive Data program.

7. References

1. CIA World Factbook (2002). http://www.odci.gov/cia/publications/factbook/index.html.
2. Eliasmith, C. and P. Thagard (2001) Integrating Structure and Meaning: A Distributed Model of Analogical Mapping. Cognitive Science.
3. Falkenhainer, B., Forbus, K., and Gentner, D. (1986). The Structure-Mapping Engine. Proceedings of AAAI-86, Philadelphia, PA
4. Falkenhainer, B., Forbus, K., Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and examples. Artificial Intelligence, 41, pp 1-63.
5. Forbus, K. (2001). Exploring analogy in the large. In Gentner, D., Holyoak, K. and Kokinov, B. (Eds.) Analogy: Perspectives from Cognitive Science. Cambridge, MA: MIT Press.

6. Forbus, K., Ferguson, R. and Gentner, D. (1994). Incremental structure-mapping. Proceedings of the Cognitive Science Society, August.
7. Forbus, K., and Gentner, D. (1997). Qualitative mental models: Simulations or memories? Proceedings of the Eleventh International Workshop on Qualitative Reasoning, Cortona, Italy.
8. Forbus, K., Gentner, D., Everett, J. and Wu, M. (1997). Towards a computational model of evaluating and using analogical inferences. Proceedings of CogSci97.
9. Forbus, K., Gentner, D. and Law, K. (1995). MAC/FAC: A model of Similarity-based Retrieval. Cognitive Science, 19(2), April-June, pp 141-205.
10. Forbus, K., Mostek, T., and Ferguson, R. (2002) An Analogy Ontology for Integrating Analogical Processing and First-Principles Reasoning. AAAI/IAAI 2002, pp 878-885.
11. Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
12. Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou and A. Ortony (Eds.), Similarity and analogical reasoning (pp. 199-241). London: Cambridge University Press. (Reprinted in Knowledge acquisition and learning, 1993, 673-694.)
13. Gentner, D. and K. Forbus (1991). MAC/FAC: A model of similarity-based retrieval. Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.
14. Gentner, D., and Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. American Psychologist, 52, 32-34.
15. Gentner, D. and A. B. Markman (1993). Analogy - watershed or Waterloo? Structural alignment and the development of connectionist models of analogy. In C. L. Giles, S. J. Hanson and J.D. Cowan (Eds.), Advances in neural information processing systems - 5. San Mateo, CA: Morgan Kaufmann.
16. Gentner, D., and Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 45-56. (To be reprinted in Mind readings: Introductory selections on cognitive science, by P. Thagard, Ed., MIT Press)
17. Gentner, D. and C. Toupin (1986). Systematicity and surface similarity in the development of analogy. Cognitive Science, 10, 277-300.
18. Gick, M. L. and K. J. Holyoak (1980). Analogical problem solving. Cognitive Psychology, 12, 306-355.
19. Gick, M. L. and K. J. Holyoak (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.
20. Holyoak, K. and P. Thagard (1995). Mental leaps: Analogy in creative thought. Cambridge, MA: MIT Press.
21. Holyoak, K. J. and P. Thagard (1989). Analogical mapping by constraint satisfaction. Cognitive Science, 13, 295-355.
22. Holyoak, K. J. and P. Thagard (1997). The Analogical Mind. *American Psychologist*, 52: 35-44.
23. Hummel, J. E., B. Burns, and K.J. Holyoak (1994). Analogical mapping by dynamic binding: Preliminary investigations. Advances in connectionist and neural computation theory: Analogical connections. Norwood, NJ: Ablex.
24. Hummel, J. E. and K. J. Holyoak (1997). Distributed representations of structure: a theory of analogical access and mapping. Psychology Review, 104(3), 427-66.
25. Hummel, J. E., and Holyoak, K. J. (1997). LISA: A computational model of analogical inference and schema induction. Psychological Review.
26. Knowledge Interchange Format. http://logic.stanford.edu/kif/dpans.html.
27. Mostek, T., Forbus, K. and Meverden, C. (2000). Dynamic case creation and expansion for analogical reasoning. Proceedings of AAAI-2000. Austin, Texas.
28. Niles, I., and Pease, A. (2001). Toward a Standard Upper Ontology, in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, (Eds.).
29. Pease, A. (2003). The Sigma Ontology Development Environment, Eighteenth International Joint Conference On Artificial Intelligence, Workshop on Ontologies and Distributed Systems, to appear.
30. Riazanov, A. and A.Voronkov (2001). Vampire 1.1 (System description), Proceedings of IJCAR 2001, LNAI 2083
31. Thagard, P., Holyoak, K. J., Nelson, G., and Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. Artificial Intelligence, 46, 259-310.
32. Thagard, P. and K. Verbeurgt (1998). Coherence as constraint satisfaction. Cognitive Science. 22.
33. Thagard, P. (forthcoming). Computing Coherence. Cognitive Models of Science, Minnesota Studies in the Philosophy of Science. Minneapolis: University of Minnesota Press.
34. Voronkov, A. (1995). The Anatomy of Vampire: Implementing Bottom-up Procedures with Code Trees Journal of Automated Reasoning, 15(2).