A short version of this paper appeared as a poster at Ants'04 (The Fourth International Workshop on Ant Colony Optimization and Swarm Intelligence)

Hypothesis Corroboration in Semantic Spaces with Swarming Agents

Peter Weinstein, H. Van Dyke Parunak, Paul Chiusano, Sven Brueckner

Altarum Institute
3520 Green Court, Suite 300
Ann Arbor, Michigan 48105, USA
{peter.weinstein, van.parunak, paul.chiusano, sven.brueckner}@altarum.org

Abstract. To anticipate and prevent acts of terrorism, Indications and Warnings analysts try to connect clues gleaned from massive quantities of complex data. Multi-agent approaches to support Indications and Warnings are appropriate because ownership and security issues fragment the data. Furthermore, the massive scale of the data suggests the need for large numbers of agents.
This paper presents the architecture and algorithms of our Ant CAFÉ system, which uses fine-grained swarming agents to extract and organize textual evidence that corroborates hypotheses about the state of the world. Multiple swarming processes operating in semantic spaces are required, including the clustering of paragraphs, identification of semantic relations in text, and assembly of evidence into structures that instantiate the hypothesis.

1 Introduction

The terrible events of 9/11 placed urgent priority on the need to anticipate and prevent acts of terrorism. In particular, Indications and Warnings (I&W) analysts try to connect clues gleaned from massive data to anticipate enemy action. By massive, we mean data measured in peta-bytes, with complex interconnectivity and heterogeneity at all levels of form and meaning. Massive computational power is required to handle such data. Therefore, scalability is a key issue and multi-agent approaches are prime candidates because of the inherently decentralized nature of their architectures.

1.1 Problem

I&W analysts at various intelligence agencies receive streams of data from numerous sources of varying quality. For example, there may be evidence that A has shipped explosives to X, and that B was seen taking photos of apartment building Y, located in X. Meanwhile, other evidence states that A and B both know terrorist C. Logically, one might deduce that there is a plan to blow up the apartment building; but it can be extremely difficult to generate this conclusion because the key evidence is buried in massive data.

We model I&W as an investigative process where analysts construct hypotheses that are tentative assertions about the world, then submit the hypotheses to systems that find and organize evidence to corroborate the hypotheses. We represent hypotheses as graphs of concepts at varying levels of abstraction. Finding evidence requires matching edges of the hypotheses graphs against document text. Organizing evidence means joining pieces of evidence according to the template provided by the hypothesis. Thus, corroborating a hypothesis resembles working a giant jigsaw puzzle with billions of pieces, and infinitely many alternative ways to construct solutions.

Current information retrieval technology fails to support I&W adequately in two fundamental ways. First, these tools do not piece together clues that might be found scattered across numerous documents. Second, these tools lack the semantic understanding required to recognize relevant pieces of information that may manifest in different forms, while excluding information that is not relevant. Research in areas such as Information Extraction [7] and Question Answering [8] attempts a higher level of semantic interpretation of text; our research builds on this work in ways that handle open inquiry and are potentially scalable to handle massive data.

1.2 Approach

To deal with massive data, we follow the swarm intelligence approach. Swarm architectures are extremely decentralized and thus allow massive parallel processing. Swarm systems are often modeled on ants and other social insects [1, 3, 13]. Numerous, relatively simple agents make decisions in response to their local environments. Out of their coordinated action emerges intelligent collective behavior (e.g., air-conditioned termite nests 10 meters tall, networked paths for food collection that are minimal spanning trees). Stigmergy – coordination via changes to a shared environment [14] – is the key to efficient swarm behavior [6]. Agents achieve stigmergic coordination, using special markers such as chemical pheromones, or directly responding to changes in the problem state affected by other agents.

To apply swarm intelligence for hypothesis corroboration requires breaking new ground in the field of swarming multi-agent systems. Typically, stigmergy occurs in simple and low-dimensional environments – often, for example, mapping directly to a two or three-dimensional physical space. In these topologies, there are no "small-world" shortcuts, and the distance between any two points is easily calculated. Semantic spaces though, typically have a small-world, scale free structure [12] in which calculating distance requires either strong assumptions or sophisticated processing [15]. Objects in ontological spaces such as concepts and instances are often represented as graphs, which are naturally composed by linking edges.

Furthermore, the complex nature of the I&W problem has led us to combine multiple swarming mechanisms in a single system, utilizing both marker-based and sematectonic stigmergy. These processes include clustering of text to yield an orderly space; identifying relations in text to yield matches; and assembly of matches into structures that instantiate hypotheses. Clustering has previously been achieved with swarming [1]. Relation identification is a very hard and currently salient research area, requiring natural language processing and extensive corpus annotation [11]. So far, our interest has been most focused on the process of evidence assembly.

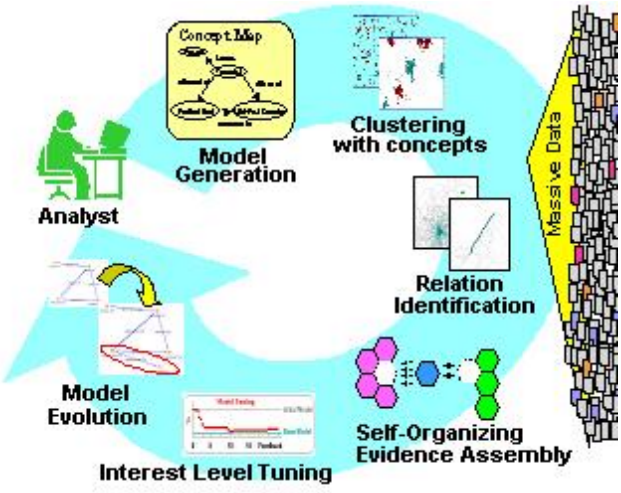

Fig. 1. Overview of the Ant CAFÉ architecture.

A third innovation regards the manner in which evidence assembly organizes evidence. Our goal is clarity, a measure that quantifies the degree of understandability of a set of assemblies (the system's response to an investigation at some point in time). In high-clarity solutions, a few assemblies stand out, they are coherent, and they are well differentiated from each other. For example, if an analyst hypothesizes the existence of scientists conducting gene regulation research to build biological weapons, we would like the system to construct evidence assemblies that, hypothetically, might describe Russian research on smallpox in the 1980's, Iraqi research on plague in the 1990's, and so on. Clarity, as a global metric, must emerge from the local behavior of evidence assembly agents, as they cannot calculate clarity explicitly without compromising the highly distributed nature of the architecture.

This paper provides an overview of the achievements of the first year of the Ant CAFÉ[1] project. Section 2 presents the initial Ant CAFÉ architecture, including clustering, relation identification, and evidence assembly. Section 3 describes our algorithm for evidence assembly in more detail. Section 4 reports some early results using generated data. Section 5 lays out future work. We conclude in Section 6.

2 Ant CAFÉ Architecture

The Ant CAFÉ architecture follows an iterative loop in which analysts ask the system to find evidence that supports a hypothesis, the system returns assemblies that organize relevant evidence, and the analyst reviews the evidence and in the process improves her understanding of the problem. The analyst-system interaction leads to a revised representation of the hypothesis. The loop repeats as the investigation advances. Figure 1 provides a high level overview.

[1] The CAFÉ acronym stands for Composite Adaptive Fitness Evaluation.

Hypotheses are represented as concept maps [5], which are utilized in every stage of processing, serving as templates for the construction of evidence assemblies. Concept maps are graphs with labeled nodes and edges. The nodes are nouns and the edges are verbs or verb-prepositions. We call the nouns and verbs concepts. We call two nodes and their connecting edge, together, a relation. We consider concept maps to be a low-commitment form of ontology-like symbolic knowledge representation. They are becoming ubiquitous for modeling domain knowledge, and are now widely taught in middle schools and elsewhere.

The left side of Figure 1 depicts modeling the analyst's interests as represented in the concept maps. Issues include initial acquisition of concept maps, tuning weights associated with concepts and relations to reflect analyst interests by observing their behavior, and evolving concept maps in semi-automatic ways to capture increasing understanding as investigations progress. This work is addressed by the Analyst Modeling Environment of Ant CAFÉ, developed by our colleagues at Sarnoff Corporation, and is outside the scope of this paper.

The right side of Figure 1 includes clustering, relation identification, and evidence assembly. Each of these stages employs a distinct swarming mechanism. They are sequential in terms of logical data flow, but execute concurrently. All of the processes use anytime algorithms where an answer is available at any time, and the quality of the answer improves as time passes. Thus, relation identification can proceed without clustering (but with less efficiency), and evidence assembly can proceed while newly discovered relations enter the assembly process incrementally.

In the following, we provide a high level description of these processing stages.

2.1 Clustering

Clustering creates order in the space of textual data, thereby increasing the efficiency of relation identification. Relation identification needs to match relations from the concept map to text. Therefore, we cluster with respect to a particular concept map.

Paragraphs are the appropriate granularity for clustering. We assume that each noun and verb in the concept map is disambiguated to a particular word meaning, namely, to a synset in WordNet [10]. We consider a word of text as evidence for a concept if that word is in the synset, or in any synset that specializes that synset. We call these sets of synsets manifestation sets, or msets. Ant CAFÉ seeks to cluster all paragraphs that have evidence of any of the nodes (nouns) in the concept map. We exclude verbs, because concept map relations often have very generic verbs.

Clusters of paragraphs emerge through stigmergic coordination of individual paragraphs on a presently arbitrary network of cluster nodes. In difference to the classical ant-based clustering, in which a population of active entities (ants) moves passive objects (eggs, larvae, bodies, etc.) around, we assign each paragraph a software agent that controls the movement of the paragraph based on local information. Paragraph agents estimate their similarity to other paragraphs in their current location by calculating pairwise similarity to a sample of those paragraphs. Pairwise similarity is currently defined as the number of concepts for which both paragraphs either have evidence, or both lack evidence. The paragraphs also estimate their similarity to the populations of a sample of neighboring cluster nodes.

Paragraphs then probabilistically decide whether to move to one of these neighboring nodes [2].

Global parameters control the size of the samples for the number of paragraphs to compare when estimating similarity with a node's population, and for the number of neighboring nodes to test. As these sampling rates decrease, one would expect the convergence to slow. Preliminary experiments using 500 paragraphs located across ten cluster nodes do show the anticipated affects, but overall the algorithm is robust.

2.2 Relation Identification

Relation identification is the process of finding text that asserts the desired relation. When such text is found, the concepts of the target relation are associated with the words in the text that are members of the corresponding manifestation sets. For example, the concept map might contain relations such as: "TERRORIST-HAVE-WEAPON", "WEAPON-CONTAIN-AGENT", "AGENT-CAUSE-DISEASE".

Consider a hypothetical sentence fragment such as "… is a carrier that can be used to incubate organisms including botulism, …". The Ant CAFÉ will consider the underlined words in this fragment to be evidence of the relation AGENT-CAUSE-DISEASE, because carrier is a kind of AGENT, incubate is a kind of CAUSE, and botulism is a kind of DISEASE. We call the association of carrier-incubate-botulism to AGENT-CAUSE-DISEASE a match.

High-precision relation identification requires more research, but our swarm intelligence approach in Ant CAFÉ lets us deal with imprecise data. Thus, if a paragraph contains evidence for each concept in a relation, we use that paragraph to create a match, or potentially several. This method offers excellent recall, but very poor precision -- which could, however, be increased with sophisticated natural language filters.

Our relation identification process is based on pheromone-based recruitment of Forager agents. Forager agents swarm over the space defined by the cluster nodes looking for matches for particular relations. When they find matches, they lay digital pheromones to attract other agents searching for the same relation. Through evaporation and propagation, these pheromones create a gradient that guides Forager agents and thus increases the efficiency of their search. The degree to which Forager agents follow the pheromone gradient is subject to an exploration/exploitation tradeoff. If foragers adhere slavishly to the gradient, then matches in previously barren locations may never be found.

2.3 Evidence Assembly

The relation identification process creates matches, who become agents in the evidence assembly process to form structures that instantiate the concept map. This process resembles a soup of molecules that join together to form larger molecules, driven by the current state of the emerging structure.

Matches face two types of "join" decisions. They can join another match that instantiates the same concept map relation. In this case, each match has a binding in a

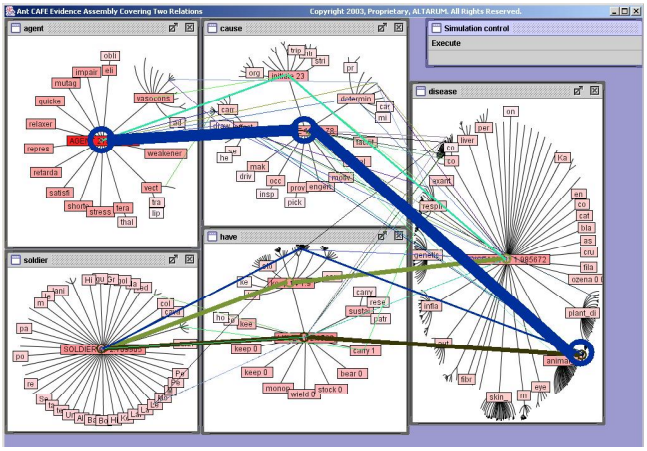

Fig. 2. Visualization of assembly in mset space.

semantic space defined by three shared manifestation sets. Or, a match can join with another match that instantiates a concept map relation linked to its own relation. In this case, the join occurs in a semantic space defined by a single manifestation set.

Figure 2 shows a screenshot from our Ant CAFÉ implementation that visualizes evidence assembly for two linked relations. Each window holds a manifestation set centered on the root concept. The relation AGENT-CAUSE-DISEASE is shown on the top and SOLDIER-HAVE-DISEASE on the bottom. Assemblies have random colors, and as matches join into assemblies, their lines grow thicker.

The best solutions maximally preserve the information of the individual bindings when described on the level of the aggregated assemblies. We call this goal clarity, since preserving the information of individual bindings yields assemblies that are relatively well differentiated from one another.

Consider an example of a join decision. If a match that binds DISEASE to flu is joined with a binding to lobar pneumonia, then the most specific subsuming (MSS) concept that includes both flu and lobar pneumonia describes the assembly best. The MSS is the lowest common ancestor in the manifestation set (here: respiratory disease). Figure 3 illustrates the assembly-level description ("vector"-"effect"-"respiratory disease") that results from joining "gene delivery vector"-"induce"-"flu"

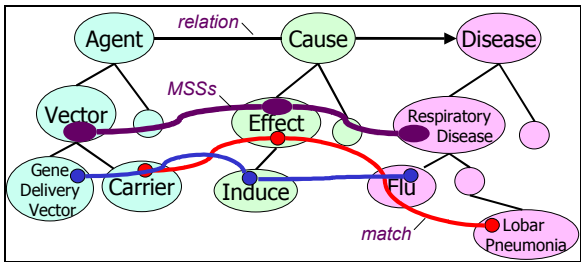

Fig. 3. Match aggregation.

with “carrier”-“effect”-“lobar pneumonia”. Thus, to achieve high clarity matches should join with many other matches while minimizing distances between the individual bindings and the assembly MSS concepts.

In our newly developed Most Likely Collisions (MLC) algorithm [4], evidence assemblies emerge from local decisions of match agents, which each contain three binding agents. A binding associates a concept from the concept map with a word in text. The current position of a binding agent is a location in the mset of the concept, where the home position is the synset of the text word, and the current position is somewhere on the path from the home position to the root of the mset tree. Binding agents move up and down this path, essentially marketing themselves at different levels of abstraction.

The self-expression of a binding agent is a function of its current position and path length. An agent at home has perfect self-expression, while a binding at the root has zero self-expression (unless root is also home). Individual binding agents move probabilistically in response to two dynamically computed forces, which balance our desires to cluster similar binding agents, while minimizing their displacement from their respective home location:

1. Attractive pheromone gradient to neighboring nodes. All binding agents deposit a pheromone that propagates and evaporates and thus attracts agents toward areas of the mset that contain greater numbers of binding agents.
2. Attractive “rubber band” pulls home. As binding agents move further from their home position, the force pulling them back home increases.

Binding agents suggest joining the current assemblies of other bindings that they meet as they move up and down in their manifestation sets. When consensus is reached among all of a match’s bindings, the match joins the elected assembly.

4 Experiments

This section reports on some initial experiments that investigate whether the MLC algorithm, based on local “join” decisions of matches, produces the desired emergent property of clarity in the resulting set of evidence assemblies. For this purpose, , we generated artificial populations of matches that vary with respect to the quality of potential solutions. The more orderly the data, the more we expect solutions with clarity.

The test match populations were generated for a concept map with four relations, three forming a triangle. These relations were person-carry-bottle, bottle-contain-liquid, person-drink-liquid, and person-live(in)-nation. These familiar relations were usedfor easy interpretation of the detailed content of emerging assemblies.

Full Random is the least orderly type of match population. It is generated with bindings selected uniformly randomly from synsets in each concept’s manifestation set. Random Clumps has somewhat more orderly match populations in which the number of clumps of matches (n) and clump size (m) is set randomly with geometric distributions. In each match clump, a binding is selected for the three concepts in the match. Then, m-1 other matches are generated to be close to the base match, using geometrically distributed excursions from the base. Hidden Solution populations are

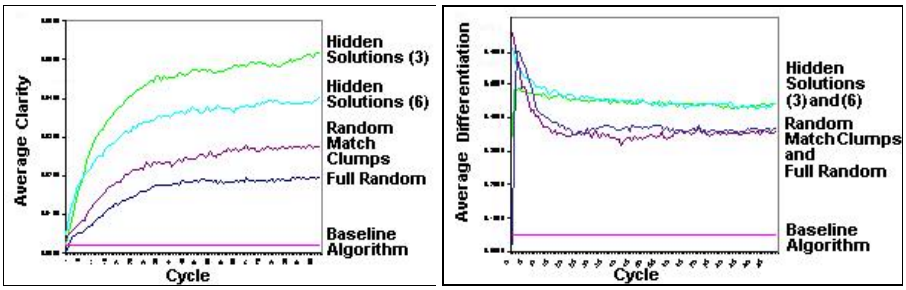

Fig. 4. Average Clarity (left) and Average Differentiation (right) over time.

like Random Clumps, except that matches are constrained to be near base matches constrained to agree where the relations join in the concept map. In other words, Random Clumps include groups of matches for individual relations, but Hidden Solutions include groups of matches whose scope includes the full concept map. "Hidden Solutions (3)" and "(6)" include three and six such sets of matches, respectively. For each type of data, 30 matches are generated for each concept map relation.

The experiments included nine runs for each of ten populations generated for each type of artificial data. Each run executes for 100 cycles. In each cycle, each binding and match agent has an opportunity to act (although not all agents do act, to avoid building in an assumption that all agents act the same number of times, which would not hold in a true distributed system). The experiments also include a baseline solution, which randomly assigns matches to one of six assemblies, rather than using the MLC algorithm.

The experiments were evaluated with a number of metrics. Clarity is calculated as the product of the average fitness and differentiation of the three most-fit assemblies. The fitness of an assembly is the product of its self-expressiveness and its substantiality (size, including breadth and depth).

Differentiation is a measure of the degree to which the top three assemblies differ from each other. Distances between pairs of assemblies are calculated with respect to the parts of the concept map that both assemblies cover. This calculation counts the numbers of nodes in the mset trees that are shared by the MSS of both assemblies, compared to the total number of nodes reaching from the root of the mset to the MSS of each assembly. This approach is basically a path-distance-based approach, which is inherently fragile [15].

The left plot in Figure 4 shows the average clarity achieved across 90 runs for each type of data as the runs progress. The more orderly the data, the greater the clarity achieved. In all cases, clarity improves most around cycle twenty-five. The high degree of order in this plot shows that MLC is essentially working as intended.

The right plot in Figure 4 shows differentiation for each type of data. The four populations divide neatly depending on whether matches are generated in a manner where they are constrained to join well according to the concept map.

5 Future Work

Currently, individual runs using the Most Likely Collisions algorithm are subject to a fair amount of noise. We believe that it will be possible to improve the responsiveness of local decision making by binding and match agents without compromising the decentralized nature of the algorithm. Specifically, we hope to improve global performance by tuning local behavior in response to local manifestation of the global state. For example, if global clarity is poor, local pheromone gradients are choppy and match agents should be more willing to abandon their current assemblies.

More fundamentally, we plan to extend the evidence assembly algorithm to incorporate more semantic information about matches. One problem with our current approach is that joining matches for linked relations depends entirely on the compatibility of the shared concept. Since the matches will often derive from different documents, such joins may be incorrect. To improve the coherence of evidence assemblies, therefore, we will want to take into consideration the textual context of matches when deciding whether they should join.

The three swarming processes of the Ant CAFÉ are currently implemented to run on individual computers in a manner that simulates distributed execution by randomizing the order of agent actions. For the next phase of our development we will be re-implementing the system to run on our in-house Linux cluster.

Finally, we are planning to open up the Ant CAFÉ's data flow to accept assertions about the world from other sources that can be included in evidence assembly. For example, we anticipate assembling assertions created via relatively accurate extraction of relations from text, deduction over existing relations with domain-specific models, and direct input from users.

6 Conclusions

This paper describes Ant CAFÉ, a swarming agent system that addresses a difficult problem requiring scalable coordination of massive computational resources. Specifically, we present new methods for finding and organizing evidence from massive data that corroborates analyst hypotheses. The Ant CAFÉ architecture combines multiple techniques of swarm intelligence: paragraph clustering emulates nest sorting, relation identification resembles foraging, and evidence assembly compares to nest construction. Furthermore, evidence assembly coordinates agents in complex semantic spaces through marker-based stigmergy.

Preliminary results indicate that swarming information extraction can operate effectively over massive data. Our experiments with the Most Likely Collisions algorithm demonstrate that desirable system-level evidence assemblies emerge from local decisions about the semantic proximity of paired concepts. Ant CAFÉ organizes the evidence into assemblies that each tell a different story about how data corroborates the hypothesis. Our metric for clarity has both an intuitive interpretation that corresponds to what analysts hope to obtain from a search, and a quantitative basis that will guide substantial and increasingly sophisticated research in the future.

Acknowledgments

This study was supported and monitored by the Advanced Research and Development Activity (ARDA) and the National Imagery and Mapping Agency (NIMA) under Contract Number NMA401-02-C-0020. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision, unless so designated by other official documentation.

References

1. E. Bonabeau, M. Dorigo, and G. Theraulaz. Swarm Intelligence: From Natural to Artificial Systems. New York, Oxford University Press, 1999.
2. S. A. Brueckner. Demonstrating Emergent Clustering of Abstract Documents. Altarum Institute, 2003.
3. S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau. Self-Organization in Biological Systems. Princeton, NJ, Princeton University Press, 2001.
4. P. Chiusano. MLC: An Agent-Based Approach to Assembling Evidence in the Ant CAFÉ. Altarum Institute, 2003.
5. J. W. Coffey, R. R. Hoffman, A. J. Cañas, and K. M. Ford. A Concept Map-Based Knowledge Modeling Approach to Expert Knowledge Sharing. In Proceedings of IASTED International Conference on Information and Knowledge Sharing, 2002.
6. P.-P. Grassé. La Reconstruction du nid et les Coordinations Inter-Individuelles chez Bellicositermes Natalensis et Cubitermes sp. La théorie de la Stigmergie: Essai d'interprétation du Comportement des Termites Constructeurs. Insectes Sociaux, 6:41-84, 1959.
7. R. Grishman. Information Extraction: Techniques and Challenges. In M. T. Pazienza, Editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Springer, Berlin, 1997.
8. M. Maybury, Editor. New Directions in Question Answering. Menlo Park, California, USA, AAAI, 2003.
9. G. A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. The Psychological Review, 63:81-97, 1956.
10. G. A. Miller. WORDNET: An On-Line Lexical Database. International Journal of Lexicography, 3(4):235-312, 1990.
11. D. Moldovan and P. Parker. Towards Automatic Discovery of Semantic Relations. forthcoming.
12. A. E. Motter, A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta. Topology of the conceptual network of language. Phys. Rev. E, 65, 2002.
13. H. V. D. Parunak. 'Go to the Ant': Engineering Principles from Natural Agent Systems. Annals of Operations Research, 75:69-101, 1997.
14. H. V. D. Parunak, S. Brueckner, M. Fleischer, and J. Odell. A Design Taxonomy of Multi-Agent Interactions. In Proceedings of Workshop on Agent-Oriented Software Engineering (AOSE03) at AAMAS03, (forthcoming), Springer, 2003.
15. P. Weinstein and W. P. Birmingham. Comparing Concepts in Differentiated Ontologies. In Proceedings of Twelfth Workshop on Knowledge Acquisition, Modeling and Management, 1999.