Agents Swarming in Semantic Spaces to Corroborate Hypotheses

Peter Weinstein, Altarum Institute peter.Weinstein@altarum.org

H. Van Parunak Altarum Institute van.parunak@altarum.org

Paul Chiusano University of Michigan pchiusan@umich.edu

Sven Brueckner Altarum Institute 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.

The Ant CAFÉ system uses fine-grained swarming agents to extract and organize textual evidence that corroborates hypotheses about the state of the world. Multiple swarming processes are required, including the clustering of paragraphs, identification of semantic relations in text, and assembly of evidence into structures that instantiate the hypothesis. These processes occur in semantic spaces defined using the WordNet ontology.

This paper provides an overview of an Ant CAFÉ prototype. It describes the system's architecture, and provides additional detail on the innovative algorithm for evidence assembly. Initial experiments using artificially generated data confirm that a global property that we call "clarity" emerges from agent decisions made in a local, and therefore scalable, manner. [1]

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 analysts try to connect clues gleaned from massive data to anticipate enemy action. By *massive*, we mean data measured in petabytes (10^{15}), which also contains complex interconnectivity and heterogeneity at all levels of form and meaning.

To handle massive data requires massive computational power: 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

The details of Indications and Warnings are hidden for security reasons, but the general idea is as follows. 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 Indications and Warnings 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 (with some degree of persuasiveness). Hypotheses may be represented 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, to corroborate a hypothesis is like working a giant jigsaw puzzle where there are billions of pieces, and infinitely many alternative ways to construct solutions.

Current information retrieval technology fails to support Indications and Warnings 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 [8] and Question Answering [9] 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

Our system uses swarm intelligence. The swarm approach is promising for dealing with massive data because of the extremely distributed nature of the swarm architectures, with corresponding potential for parallel processing. Swarm systems are modeled on ants and other social insects [2, 4, 16]. Numerous, relatively simple agents make decisions in response to their local environments. Out of their coordinated action emerges desired behavior that is remarkably intelligent on the collective scale: air-conditioned termite nests 10 meters tall [4], networked paths for food collection that are minimal spanning trees [2], and so on. *Stigmergy* – coordination via changes to a shared environment [17] – is the key to efficient swarm behavior [7]. Agents can achieve

[1] A research prototype of the Ant CAFÉ will be available for demonstrations at the conference.

stigmergy using special markers such as *pheromones*: chemicals that propagate through the shared space and evaporate. Or, stigmergy can result as agents respond directly to changes in the shared environment effected by other agents (this is called *sematectonic* stigmergy).

To apply swarm intelligence for hypothesis corroboration requires breaking new ground in the field of swarming multi-agent systems. Typically, stigmergy occurs in environments whose topology is simple and of low dimensionality – often, for example, mapping directly to a two or three-dimensional physical space. In these topologies, there are no "small-world" shortcuts, and it is straightforward to calculate the distance between any two points.

Semantic spaces, have small-world, scale free structure [14] in which calculating distance requires either strong assumptions or sophisticated processing [22]. While there are other types of semantic spaces, such as the word co-occurrence factors produced by Latent Semantic Indexing, these are high dimensional and, even more fundamentally, do not lend themselves to putting together puzzles where pieces of information are composed into larger structures that represent hypotheses. Objects in ontological spaces such as concepts and instances are often represented as graphs, which are naturally composed by linking edges.

A second way in which our research breaks new ground is that the problem has led us to combine multiple swarming mechanisms in a single system, utilizing both digital pheromones 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 [2]. Relation identification is a very hard and currently salient research area, requiring natural language processing and extensive corpus annotation [13]. So far, our interest has been most focused on the process of evidence assembly.

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 in the 1980's on smallpox, Iraqi research in the 1990's on plague, and so on. Note that clarity, as a global metric, must emerge from the local behavior of evidence assembly agents: these agents 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[2] project. We call the backend part of the Ant CAFÉ system the Ant Hill, while the frontend is the Analyst Modeling Environment (AME). Section 2 presents the initial Ant Hill 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 describes our roadmap for future work, and Section 6 concludes.

2 Ant Hill Architecture

We conceive of the Ant CAFÉ architecture as an iterative loop where 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, and the loop iterates repeatedly in this manner as the investigation advances. Figure 1 provides a high level overview: the pictures accompanying each Ant Hill stage illustrate the insect analogies to our processing, as explained below.

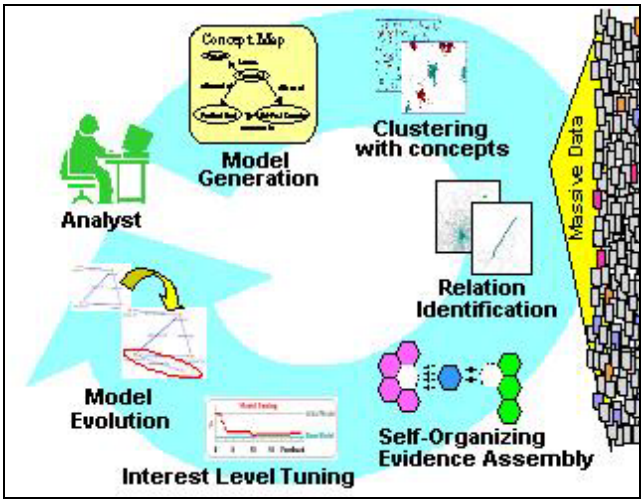

Figure 1: Overview of the Ant CAFÉ architecture

Hypotheses are represented as concept maps [6]. The concept maps are utilized in every stage of processing; they essentially act 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 quite ubiquitous for

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

modeling domain knowledge, and are now widely taught in middle schools and elsewhere.

The left side of Figure 1 involves 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 part of the Ant CAFÉ being developed by our colleagues at Sarnoff Corporation [1], and is outside the scope of this paper.

The right side of Figure 1 includes the Ant Hill processing stages of 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 some 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.

The following subsections provide a high level description of each of the processing stages.

2.1 Clustering

The point of clustering is to increase the efficiency of relation identification by creating order in the space of textual data. Relation identification needs to match relations from the concept map to text. Therefore, clustering is with respect to a particular concept map, rather than once only for all uses of the corpus.

The appropriate granularity for clustering is on the level of paragraphs. This decision is best understood by eliminating alternatives. Corroborating hypotheses requires connecting assertions culled from multiple documents. Relevant information needed from any particular document may constitute a very small part of that document. Therefore, clustering documents could well result in clusters that are meaningless or even counterproductive for a particular investigation. On the other hand, to cluster units of text smaller than paragraphs would require accurate prior resolution of co-references: including pronouns, and other anaphora such as referring to Osama Bin Laden as "the leader" and so on. Generally, paragraphs are the smallest well-defined unit of text whose boundaries are infrequently violated by co-reference.

In the current Ant CAFÉ, we assume that each noun and verb in the concept map is disambiguated to a particular word meaning: namely, to a synset in WordNet [11]. We then consider a word of text as evidence for a concept if that word is in the synset, or in any synset that specializes the meaning of that synset.

Given a WordNet synset, one can recursively fetch its specializations by requesting hyponyms for nouns, and troponyms for verbs [20]. The result is a tree of synsets, which we call a *manifestation set* or *mset*, because it identifies all of the words that we consider to be a manifestation of the target concept. Figure 2 illustrates an excerpt of a manifestation set. A full mset for a relatively general concept could include hundreds of terms. Note that synsets can have multiple parents, so manifestation sets should really be directed graphs rather than trees. At the cost of some inaccuracy, however, the current Ant CAFÉ for convenience represents them as trees where some subtrees occur in multiple places.

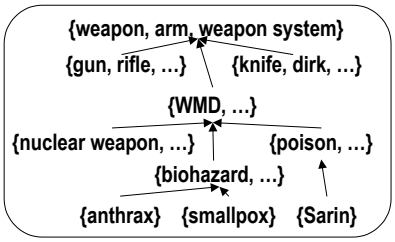

Figure 2: Excerpt of a manifestation set for Weapon

The input to Ant CAFÉ clustering includes any paragraph that has evidence (a word in the manifestation set) of any of the nodes (nouns) in the concept map. We exclude the verbs for this purpose because concept map relations often have very generic verbs that would cause inclusion of too many paragraphs.

The clustering algorithm is analogous to the way that ants sort eggs, food, debris, and so on in their nest. Ants doing sorting pick up objects in a stochastic manner, where the probability of picking up an object increases to the degree that it is different from objects around it. The ants then move about randomly, with some probability of dropping the object at any point. The probability of dropping the object increases to the degree that it is similar to objects around the ant's current location. This is a form of sematectonic stigmergy.

In the Ant CAFÉ, Paragraph Agents act as ants. Paragraphs are initially assigned randomly to a cluster node, which are logical locations. There are an arbitrary number of cluster nodes, but this number should be roughly comparable to the number of relations in the concept map. 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 current populations of a sample of neighboring cluster nodes. Paragraph Agents then stochastically decide whether to request movement to one of these neighboring nodes. More detail is available in [3].

System parameters control the size of the samples for the number of documents to compare to estimate similarity to a node's population, and for the number of neighboring nodes to test. As these sampling rates decrease, one would expect the speed of convergence to slow. Preliminary experiments using 500 paragraphs located across ten cluster nodes do show the anticipated affects, but overall the algorithm seems to be 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 words in Arial font 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*.

Table 1 shows the correspondence between terms used in the Ant CAFÉ to talk about concepts maps, and the corresponding terms for the objects created when relations are matched to text.

Table 1: Correspondence of elements when relation identification creates matches

Template	Instantiated	Instantiated Content
Concept	Binding	A word or phrase in the text
Relation	Match	Three bindings
Concept map	Evidence assembly	Matches joined according to the concept map

To do relationship identification well is, of course, a very difficult problem that is the focus of considerable research today [12, 15, 18, 19]. The basic problems are the complexity of natural language (including co-reference and so on), and the multi-layered semantics of communication artifacts. For example, a document that lists patients in a hospital provides clear evidence that each of those patients was in the hospital at some time, but the document does not include text that explicitly matches any sentence-level pattern.

Fortunately, research on relation identification does seem to be moving along well, and for the Ant CAFÉ we are content for now to use very simple and imprecise methods for identifying relations. Currently, if a paragraph contains evidence for each concept in a relation, then we use that paragraph to create a match, or potentially several. This method offers excellent recall but very poor precision. Potentially, a series of increasingly sophisticated natural language filters can be applied to increase the precision of relation identification to a point that is maximally cost beneficial.

The relation identification process in the Ant CAFÉ is loosely analogous to the manner in which ants scour a neighborhood looking for food using pheromone-based stigmergy. 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 Forager Agents searching for the same relation to the same cluster node. These pheromones propagate and evaporate; creating a gradient that guides Forager Agents and substantially 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

In the evidence assembly process, matches produced by relation identification self-organize into structures that instantiate the concept map. This process may be likened to a soup of molecules that join together to form larger molecules. In the insect domain, it corresponds to nest building: Match Agent behavior is influenced by the presence of assemblies under construction, and this is analogous to insects such as wasps whose behavior changes depending on the state of partially constructed nests.

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 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 3 shows a screenshot from a demonstration of the Ant CAFÉ that visualizes evidence assembly for two linked relations. Each window holds a manifestation set with the root concept in the middle; the relation AGENT-CAUSE-DISEASE covers the top half of the screen and SOLDIER-HAVE-DISEASE covers the bottom half. Each assembly has a randomly assigned color, and as matches join into assemblies, the colored lines representing them grow thicker.

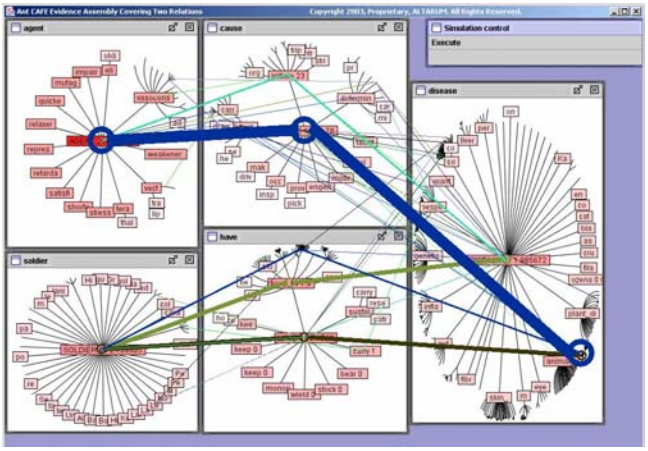

Figure 3: Visualization of assembly in mset space

The best solutions of evidence assembly are those that 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. A desirable solution as visualized in Figure 3, therefore, would have thick lines with vertices in locations within the mset trees that are spaced well apart. See Section 4.2 for an operationalization of clarity.

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 best way to describe the assembly is using the *Most Specific Subsuming (MSS)* concept that includes both flu and lobar pneumonia. The MSS is identified by rising in the manifestation set until a common ancestor is found: in this case, respiratory disease. Figure 4 illustrates the assembly-level description (vector-effect-respiratory_disease) that results from joining gene delivery vector-induce-flu to carrier-effect-lobar_pneumonia. Thus, the goal of clarity means that matches should join with other matches such that distances between the individual bindings and the assembly MSS concepts are minimized.

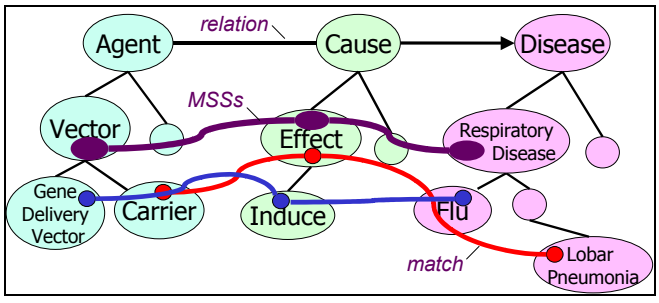

Figure 4: Match aggregation

3 Most Likely Collisions Evidence Assembly

This section briefly describes the Most Likely Collisions (MLC) algorithm, which drives the behavior of matches during the self-organization of assemblies. MLC is a fully distributed algorithm, in the sense that all decisions are made on the local level of bindings and matches. Results from experiments with MLC are reported in Section 4. For more detail on MLC, see [5].

The gist of MLC is described in Figure 5, which contains pseudocode for the behavior of Binding Agents. Remember that 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. *Self-expression* is a function of current position and path length: a Binding Agent at home has perfect self-expression, while a binding at the root position has zero self-expression (unless it is also home).

```
If not bound in an evidence assembly
  Move current position in response to
  pheromone and pull home
  For each of n randomly selected
  partner binding-agents residing at
  the current node:
    suggest a join with the
    partner's assembly
Deposit pheromone proportional to
self-expression at the current node
```

Figure 5: Pseudocode for Binding Agents in MLC evidence assembly

Binding Agents move stochastically in response to two forces:

1) *Pheromone deposited by all Binding Agents.* Pheromone propagates and evaporates and thus create blended gradients that point Binding Agents toward areas of the mset that contain greater numbers of Binding Agents.
2) *A rubber band-like pull home.* As Binding Agents move further from their home position, the force pulling them home increases.

The first force encourages the agents to find assemblies with large numbers of matches. The second force encourages Binding Agents to find assemblies that are "close to home" – which should yield clarity.

Join decisions are made by matches, which each contain three bindings. 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 moves to the elected assembly.

4 Experiments

This section reports on some initial experiments that investigate whether the MLC algorithm, which uses purely local logic to guide the join decisions of matches, produces the desired emergent property of clarity, which is associated with the global solution consisting of a set of evidence assemblies.

4.1 Methodology

To test whether evidence assembly using the MLC algorithm works as intended, 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 common everyday relations were selected to make it easy to interpret output showing detailed content of assemblies.

The least orderly type of match population, called Full Random, is generated with bindings selected uniformly randomly from synsets in each concept's manifestation set. The next, somewhat more orderly type of match population is called Random Clumps. In these populations, 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 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)" match populations 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 the metrics summarized in Figure 6. Clarity is calculated as the product of the average fitness and differentiation of the three most-fit assemblies. The number three is chosen arbitrarily, but falls within the common limits of cognitive capacity that lead people to generally choose about three reasons for whatever they do [10]. 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 most-specific-subsumers 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 [22].

(1) clarity(population) = fitness(top3assemblies)
* differentiation(top3assemblies)
(2) fitness(assembly) = self-expression(assembly)
* substantiality(assembly)
(3) self-expression(assembly) = average self-expr(match)
(3a) self-expr(match) = average self-expr(concept)
(3b) self-expr(concept) = 1 - (distance(current, home)
/ distance(root, home))
(4) substantiality(assembly) = magnitude(assembly) *
breadth(assembly)
(5) magnitude(assembly) = matches in assembly / total matches
(6) breadth(assembly) = relations covered / total relations
(7) differentiation(top3assemblies) = average distance-
between(assem1, assem2)
(7a) distance-between(assem1, assem2) = average
sharedMsetPct(shared concept)
(7b) sharedMsetPct(concept) = mset nodes subsuming both
assembly MSS / mset nodes subsuming either MSS

Figure 6: Metric definitions

4.2 Results

Figure 7 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, MLC achieves most improvement by twenty-five cycles or so. The high degree of order in Figure 7 shows that MLC is essentially working as intended.

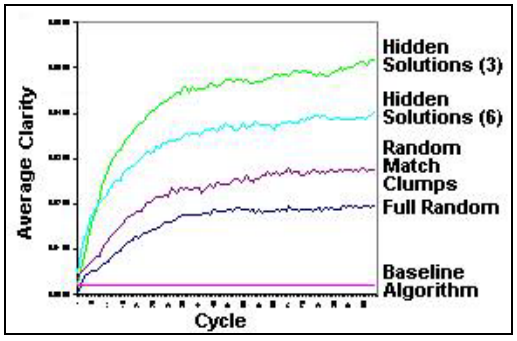

Figure 7: Average clarity

Figure 8 (a-c) show graphs for average assembly fitness and the sub-components of the fitness metric, self-expression and substantiality. The x-axis shows cycles and the series are colored as in Figure 7. (For readers in black and white, the vertical position of the series is the same as in Figure 7 for (a) and (b), and reversed in (c)). Figure 8 (a) shows the same order as the results for clarity. In Figure 8 (b), self-expression decreases as the assemblies form. This is to be expected since each match starts off in its own assembly with perfect self-expression. Note that for the Full Random data, self-expression approaches that of the baseline algorithm. In Figure 8 (c), it is interesting that the Full Random data ends up in the largest assemblies. The MSSs in these assemblies are almost always at the most generic level possible – the roots of the msets – and so there are few meaningful choices for matches and they mostly end up lumped together.

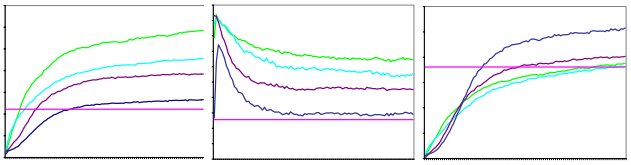

Figure 8:
(a) Fitness (b) Self-Expression (c) Substantiality

One aspect of Figure 8 (b) and (c) is surprising. Intuitively, one would expect substantiality to be greater for Hidden Solutions (3) rather than Hidden Solutions (6), since there are more matches in each pre-constructed group of nearby matches. On the other hand, one would expect self-expression for both of these data types to be essentially equal, since each hidden solution is constructed in the same manner. The results show, however, that the relative clarity for Hidden Solutions (3) is due to better self-expression, not substantiality. This is currently a bit of a mystery that will require further analysis to unravel.

Finally, Figure 9 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.

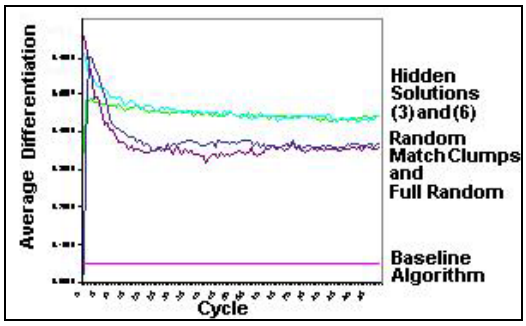

Figure 9: Average differentiation

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 radically distributed nature of the algorithm. Theoretical work has shown that in complex systems, global state can be perceived in a local manner by analyzing time series of local variables that are bound into global interaction [21]. For example, a global state of high clarity should manifest in evidence assembly as well-defined pheromone gradients that can be perceived locally by agents positioned in the mset trees. Perhaps, improvements in global performance can be achieved by tuning local behavior in response to perceptions of global state. For example, if global clarity is poor, 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 illustrate, consider joining an assertion about the reproduction of a bacterium that derives from research in drosophila, to an assertion about the toxicity of the same bacterium based on research in humans. Drosophila and humans are quite different. Is it safe to assume that these assertions form a coherent whole? The best answer seems to be maybe, depending on the particular assertions and the problem context. To improve the coherence of evidence assemblies, therefore, we will want to take into consideration several aspects of context when deciding whether to join matches.

The Ant Hill is 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 Ant Hill development we will be re-implementing the system to run on an openMosix cluster including 20 processors.

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 a multi-agent system that addresses a difficult problem requiring scalable coordination of massive computational resources. The Ant CAFÉ is developing new methods for finding and organizing

evidence from massive data that corroborates analyst hypotheses. The Ant CAFÉ architecture is unusual in that it combines multiple techniques of swarm intelligence: the paragraph clustering process emulates nest sorting, the relation identification process resembles foraging, and the evidence assembly process can be compared to nest construction. Furthermore, the evidence assembly process uses pheromones as a coordination mechanism in relatively complex semantic spaces. These spaces are defined by the graph structures of the concept maps, and the manifestation set trees for each concept in the maps.

Our preliminary results lend support to the claim that swarming information extraction can operate effectively over massive data. More specifically, our experiments with the Most Likely Collisions algorithm have demonstrated that desirable system responses on the level of sets of evidence assemblies can emerge from local decisions about the semantic proximity of paired concepts. To the extent supported by the different types of data, the Ant Hill organizes the evidence into assemblies that each tell a different story about how the data corroborates the hypothesis. Furthermore, we believe that 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.

7 Acknowledgements

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.

8 References

[1] R. Alonso and H. Li. Model-driven Information Discovery for Intelligence Analysis. In *Proceedings of In preparation*.

[2] E. Bonabeau, M. Dorigo, and G. Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. New York, Oxford University Press, 1999.

[3] S. A. Brueckner. Demonstrating Emergent Clustering of Abstract Documents. Altarum Institute, 2003.

[4] 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.

[5] P. Chiusano. MLC: An Agent-Based Approach to Assembling Evidence in the Ant CAFÉ. Altarum Institute, 2003.

[6] 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.

[7] 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.

[8] 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.

[9] M. Maybury, Editor. *New Directions in Question Answering*. Menlo Park, California, USA, AAAI, 2003.

[10] 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.

[11] G. A. Miller. WORDNET: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4):235-312, 1990.

[12] D. Moldovan and P. Parker. Towards Automatic Discovery of Semantic Relations. (forthcoming).

[13] D. Moldovan and P. Parker. Towards Automatic Discovery of Semantic Relations. forthcoming.

[14] 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.

[15] F. J. Oles. (Relational Learning). 2002. http://www.research.ibm.com/IE/.

[16] H. V. D. Parunak. 'Go to the Ant': Engineering Principles from Natural Agent Systems. *Annals of Operations Research*, 75:69-101, 1997.

[17] 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.

[18] J. Pustejovsky, J. Castano, and J. Zhang. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of Pacific Symposium on Biocomputing*, 2001.

[19] R. B. a. Y. Ravin. Identifying and Extracting Relations in Text. In *Proceedings of NLDB'99*, 5, 1999.

[20] S. Shapiro and L. Iwanska, Editors. *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. AAAI/MIT Press, 2000.

[21] F. Takens. Detecting Strange Attractors in Turbulence. In L.-S. Young, Editor, *Dynamical Systems and Turbulence*, vol. 898, *Lecture Notes in Mathematics*, 366-381. Springer, New York, 1981.

[22] P. Weinstein and W. P. Birmingham. Comparing Concepts in Differentiated Ontologies. In *Proceedings of Twelfth Workshop on Knowledge Acquisition, Modeling and Management*, 1999.