Case-Based Reasoning in Support of Intelligence Analysis

Elizabeth T. Whitaker and Robert L. Simpson, Jr.

Information Technology and Telecommunications Laboratory
Georgia Tech Research Institute
Atlanta, GA. 30332
Betty.Whitaker@gtri.gatech.edu
Bob.Simpson@gtri.gatech.edu

Abstract

Open source intelligence analysts routinely use the web as a source of information related to their specific taskings. Effective information gathering on the web, despite the progress of conventional search engines, is a complex activity requiring some planning, text processing, and interpretation of extracted data to find information relevant to a major intelligence task or subtask (Knoblock, 1995), (Lesser, 1998) and (Nodine, Fowler et al., 2000). This paper describes our design, architecture, and some initial results of next generation information gathering techniques to be used to support the development of tools for intelligence analysts. We are integrating several areas of AI research, especially case-based reasoning, within the Novel Intelligence from Massive Data (NIMD) research program sponsored by the Advanced Research Development Activity. The goal of our research is to develop techniques that take advantage of the vast amounts of information available today on the web so that the web can become a valuable additional resource for the intelligence community. Our solution is a set of domain specific information gathering techniques that produce multi-step plans for gathering information in support of the intelligence analytic process. These plans, when executed, extract relevant information from both unstructured and structured documents and use the extracted information to refine search and processing activities.

Background

The Advanced Research and Development Activity of the U.S. Intelligence Community has responded to the data overload problems being faced by our community of intelligence analysts by sponsoring a large muti-year research initiative called **Novel Intelligence** from **Massive Data, (NIMD).** The **NIMD** program is aimed at focusing analysts' attention on the most critical information found within massive data - information that indicates the potential for strategic surprise. Novel Intelligence is information not previously known to the analyst or policy makers. It gives the analyst new insight into a previously unappreciated or misunderstood threat.

The Georgia Tech Research Institute is one of the participants privileged to be a part of the NIMD program. We are investigating certain aspects of an intelligence analyst's preferences and analytic strategies used in the process of discovering new knowledge. We are analyzing data collected on NIMD-sponsored analysts as they conduct searches for information on the web in support of a variety of taskings. We have devised several search strategies that can be used by analysts to improve their capability to perform tasks in a shorter time period. We have designed and prototyped a software tool for intelligence analysts that applies case-based reasoning in combination with other advanced reasoning techniques to help analysts perform knowledge discovery. Our work involves the development, validation and incremental improvement of a set of knowledge discovery automation aids that significantly reduces the manual searches done by intelligence analysts and increases the quality and quantity of derived intelligence. One simple illustration of this vision is that analysts' tools should enable the reuse of previous episodes (cases) of knowledge discovery. This implies that these tools must save their results in a growing knowledge base, in such a way that other analysts need not rediscover important findings. We expect that the acceptability of this software will ultimately depend on how well our software and other NIMD components adjust their behaviors to support or complement the ways that analysts prefer to work.

The Users and Their Tasks

The intended users of the Case-Based Reasoning for Knowledge Discovery tool are intelligence analysts. Intelligence analysts are a special class of knowledge worker whose main responsibilities are to research a topic or question, referred to as a "tasking," and produce a report or a briefing, either long term or short term. The tasking may include searching, reading, organizing, and integrating information from many sources, both classified and open. The users are typically very methodical and they attempt to be thorough, but sometimes time constraints prevent them from doing as much research as they would like. They are often expert in several areas of intelligence but may receive a tasking outside of their areas of primary expertise. It is a

common strategy for the intelligence analysts to decompose a task into sub-tasks that help refine the topic or provide focus for the various issues. For each of these parts or sub-tasks, an analyst may conduct a search for relevant information that she believes will help in formulating a response. Once the relevant information has been collected, the analyst will organize the information and produce a report or briefing, this is called "finished intelligence." This product is then delivered to the customer, after the required levels of review and approval.

Our project is focused on that part of the analytic process where the analyst is searching for and accumulating appropriate pieces of information relevant to specific sub-tasks (Whitaker and Simpson, 2003). Our approach represents these analytic strategies in the form of domain specific search plans. Our vision is that a NIMD-enabled analyst support environment could reuse a successful analytic strategy on massive data. Significant portions of the search and analytic strategy can be automated, but we have come to understand the importance of interaction with the analysts. Analysts want to completely understand the search and analytic strategies as well as the results and to be able to interact with and tailor these strategies based on their experience and background knowledge.

Capabilities we are exploring in the course of our research include:

- Analyzing and capturing the often implicit search plans (analytic strategies) used by successful intelligence analysts according to individual, task/target/topic, group / sub-group / organizational unit, time, event;
- Reusing search plans among a community of analysts for the purpose of enabling collaborative investigation of hypotheses and respective assemblies of supporting evidence, which ultimately constitute the discovery of new knowledge;
- Determining the types of queries the analyst issues to which sources for given types of problems so that they may be made explicit in analyst's search plans;
- Determining what assumptions drive the analysis and making those explicit in an analytic process model.

We are using a scenario based approach to help envision and explore possible design alternatives without knowing the real details or having access to classified information sources. In a project with a focus on reusing plans and heuristics, including experience and explicit knowledge, of experts, we have a particularly challenging situation. We have very limited access to intelligence analysts. This leaves us with a set of nontraditional and non-ideal approaches to knowledge acquisition. We have received initial instruction and presentations on the analyst's processes, context, environment and tasks, but much of our initial knowledge acquisition has come from published papers (Heuer, 2001; Jones, 1995; Krzan, 1999). We have some access to surrogate analysts, people who have some experience as intelligence analysts, and have had the opportunity to watch analysts work for short periods. We must now envision all of these aspects projected into a world with new technologies, specifically search and knowledge discovery tools and aids. The knowledge gathered and inferred about analysts must be represented in the knowledge discovery plans and heuristics encoded in our knowledge base. For some of this representation, we must act as surrogate analysts trying to solve the knowledge discovery problem. This leaves us with the advantage of personally employing the mental models of the users and experts, but with the risk that we might be missing some aspect of their approach. This risk is mitigated by the eventual plan of testing the tools with real analysts in a NIMD testbed.

Case-Based Reasoning for Knowledge Discovery

Knowledge Discovery Problems

Below is a list of some problems that typify the types of questions an analyst might need answered (sometimes with very short time budgets).

1. Describe the computer capabilities of terrorist organization X
2. Assess country X's capability to produce biological weapons
3. Discover bioterrorist experts and their associated organizations
4. Find clusters of nuclear weapons manufacturers who are associated with suspect organizations

The ability to find information that pertains to these topics are not likely to come from general search engines using general search terms. For open source analysts, however, there is often no option besides typing key words into commercial web search engines and laboriously pour over dozens and dozens of retrieved documents to extract the pieces of information that they believe are relevant to their problem.

One illustration of the use of our techniques is decomposition of a problem type such as problem number 4 above. Our knowledge discovery plan for problem number 4 is described in text form in the box below:

- **Goal**: *Find clusters of nuclear weapons manufacturers who are associated with suspect organizations*
- Extract names of companies with the particular characteristic "nuclear weapons manufacturer"
- Find organizations that each manufacturer is associated with
- Compare these organizations against suspect organizations and store resulting organizations in the database

- Find links between manufacturers (the above selected nuclear weapons manufacturers) through organizations
- Result: links between organizations with a particular characteristic (nuclear weapons manufacturers) who are associated with suspect organizations
- Display clusters of entities (lists of organizations or people who are associated through suspect organizations)

In order to improve the reusability of these knowledge discovery plans, they are represented in a way that allows them to be instantiated with variable substitution. For example, the above plan can be reused with "Chemical weapons manufacturer," "Nuclear warfare experts," "Microbiology expert," "explosives manufacturer," or "drug dealer" replacing the appropriate variables in the plan.

So far, we have explored only a few types of knowledge discovery problems; namely those questions that seek information about "Expertise," "Capabilities," "Beliefs," and "Intent." We have an initial implementation of plans that address "expertise" questions and are currently completing plans for capabilities and beliefs. To illustrate our design, consider problem number 2 above. To answer that question, we implicitly need to answer several implied questions. The capability to produce biological weapons presumes that the organization has people with the necessary knowledge and skills. In order to address this question, we need to search for people that would have knowledge of biological agents and then limit those resulting people to people associated with organization X. The first search can be performed by searching through PubMed, which is a large medical publication database of the National Library of Medicine, for publications that explicitly mention biological agents (e.g., anthrax etc.). From this search we have a list of publications on the subject of anthrax (or other biological agents). We can then look at the details of the publication and pull a list of authors as well as the country or organization that they published with. By selecting only authors who have published articles on anthrax and who also have affiliation with organization X we have determined a set of potentially relevant information.

Case-Based Planning

Case-Based Reasoning (CBR) (Kolodner & Simpson, 1989; Kolodner, 1993) solves new problems by applying stored experiences. Past experiences are stored as cases in a case library that may be implemented as a database. When used in support of a planning task, case-based planning (CBR Planning) (Hammond, 1989) is the reuse of past plans to solve new planning problems. The knowledge discovery system retrieves previously generated solutions from a case library and adapts or repairs one of them, the closest match, to suit the current problem. CBR planning differs from standard generative planning in addressing goal interactions first, and in aiming to anticipate and avoid failures, rather than generating subplans and debugging their interactions. In addition, CBR can be used to retrieve plans for solving subproblems and these plan parts can be composed to address the analysts' knowledge discovery problem. One aspect of our planner is the use of cases for meta-planning. We call these "planning policies." Planning policies are restrictions or constraints that will help the CBR system choose among possible plans or cases in the case library. Examples of planning policies include:

- Search only data sources with a particular characteristic, e.g., "university web sites"
- Only bring back information that you can get in one hour
- Only point to documents that are less than a year old
- Search only authoritative sources

Knowledge Discovery Plans

A *plan* in this context consists of

- The analyst's goal
- An initial state which, in this context, is a set of pre-existing information elements and explicit assumptions about the world situation.
- A sequence of actions that when executed starting in the initial state results in a goal state

The goal for our CBR for Knowledge Discovery system is the resolution of a knowledge discovery problem. The goal more precisely is to have a set of relevant information with their appropriate connections and inferences that addresses the knowledge discovery problem. Each step in a plan has a set of preconditions that signals the planner that that step is enabled. In information space, the preconditions consist of having the information necessary to perform the next step. Each step also has a set of post-conditions consisting of the knowledge and the representation or partial solutions that exists after the step is performed.

In our CBR for Knowledge Discovery System, possible plan actions are all from a small set:

- Create String
- Search
- Query
- Elaboration
- Extract
- Store in database
- Sort
- Display

Knowledge Discovery Cases

A case in our Knowledge Discovery system has a fairly traditional case representation. It consists of a problem description, a Knowledge Discovery Plan with its

associated features, i.e., a set of attributes that describe and characterize the plan, with its associated solution or plan for solving the knowledge discovery problem, and a description of the results of executing the plan. The representation is shown in the box below.

Knowledge Discovery Case:
- **CaseID**
- **Knowledge Discovery Problem**:
  - KDID
  - Description
  - SubGoalID
  - KDType
  - Structure
  - Domain
  - Geography
  - Elaboration Terms
- **Knowledge Discovery Plan**:
  - PlanID
  - SubGoalID
  - KDID
  - ScriptID
  - Script Description
  - Sources
    - KDID
    - Sources
  - Summary
- **Results**

Similarity Metrics and Attributes

In order to design a system that will be able to retrieve from the case library the knowledge discovery plans which will be most useful in solving the analyst's current problem the important features of the knowledge discovery plan must be identified, and stored as a feature vector. A similarity metric that can be applied to the feature vectors of the target case and cases in the case library to compute a meaningful distance is then created.

Through requirements analysis, knowledge acquisition, experimentation with knowledge discovery through web search, and analysis of published analysts' processes, we have identified and classified some of the types of information that intelligence analysts look for as attributes that will be useful in case retrieval:

Structure: In developing a knowledge discovery plan and defining its similarity to a current knowledge discovery problem for the purpose of retrieving a case to adapt and reuse, one of the primary attributes is the "structure" of the information that the analyst is looking for. Because the information being searched for by the intelligence analysts in most cases goes beyond looking for simple facts or processes, the structure can be very complex, having primary influence on the characteristics of the search plan. Examples of structure commonly used in a knowledge discovery plan are:

- **Associations or relationships**: One of the techniques that we have seen analysts (and other information workers) use when looking for relationships between two entities, is to look for common relationships to a third entity
- **Clusters**: A more complex kind of association is a clusters of related entities such as:
  i. People (e.g., bioweapons experts or terrorists)
  ii. Organizations (particularly terrorist organizations, suspect businesses and organizations that do business with terrorist organizations)
  iii. Events (e.g., bombings, attacks, or other terrorist events)
- **Time Sequences**: When tracking a terrorist event or trying to identify a potential terrorist event, there are sequences of subevents or activities that take place as part of training and preparation. Being able to search internet webpages and documents, extract information and create a representation that allows the analyst to see the time sequence of events requires a particular type of search plan. This is a type of knowledge discovery and representation that we have seen knowledge workers in other fields use as well. Historians, company employees tracking and predicting the development of particular technologies, and epidemiologists trying to understand the spread of disease, all search for information and relate it to a time sequence in order to explain, prevent, influence or predict events.
- **Spatial Associations**: Another important structure that an analyst might be trying to put together is a representation which relates the movements of entities through space and time. For example, in searching for information that can be used to explain, predict, prevent or influence terrorist events, analysts look for components that can be tied together in a space-time representation. The search plan used to implement this kind of knowledge discovery has characteristics that support this kind of information and representation.

Type of knowledge: Intelligence analysts have taskings that require information searches or knowledge discovery techniques that are specialized to the type of knowledge that they are looking for. Some examples of commonly used types of knowledge that require specialized search plans are:

- capabilities
- expertise
- beliefs
- intents

Because the sources, sequences and characteristics of each of these types of search plans are very different from the others, we have chosen this attribute as the most important after the structure of the information.

Focus of Information: The Focus of Information is the information domain of the knowledge search being performed. Examples are "weapons of mass destruction," "terrorism," and "biological weapons." The Focus of Information is included as an attribute in the feature vector, because intelligence analysts are often working in a particular domain and find themselves searching for the same "focus" many times. The focus of the information is more easily changed or adapted from one knowledge discovery plan to another, but if this is a domain that the analyst works in routinely, there may be specialized sources, search approaches, and inferencing techniques that the analyst uses. It is important to include this attribute in the feature vector describing the knowledge discovery plan, but it will not be weighted as heavily in the similarity calculation as the first two attributes.

Geographic Area: Analysts often search for information related to a particular geographic area. In addition, the information they seek may not be expressed in English. This is an important issue, but one that is beyond the scope of our project. Analysts may have favorite websites and search and inferencing techniques related to a specific geographic area, and they may have ways of interpreting information that vary from one geographic area to another. Because of its importance, the geographic area is included in the feature vector of the knowledge discovery problem. Like the Focus of Information attribute, it will not be weighted as highly as the Structure or the Type of Knowledge attribute.

We expect, as we increase the size of the case base and have the opportunity to perform experiments which will allow us to analyze the types of information found using these attributes, that we will identify new attributes to include in the feature vector, and that we will have a better understanding of the weightings and similarity metrics that will be useful for case retrieval.

CBR for Knowledge Discovery Architecture

The CBR for KD Architecture, Figure 1, shows an analyst interacting with the system to provide a knowledge discovery problem. The system then retrieves the cases which are the closest matches to the problem presented by the analyst. Using the retrieved plan, the system will adapt and instantiate the plan to tailor it to the knowledge discovery plan provided by the analyst. The analyst will have an opportunity to interact with the system at this point to modify the plan or to add more information. The plan is then executed producing relevant pieces of extracted information, which can be combined and reasoned about to produce intermediate pieces of inferred knowledge and finally discovered knowledge which provides information which the analyst was looking for in the knowledge discovery problem. There is an evaluation of the results by the system and the analyst, and if necessary information is added or updated to recognize (Whitaker & Bonnell, 1992) and produce a more successful knowledge discovery plan. The process is iterated as necessary.

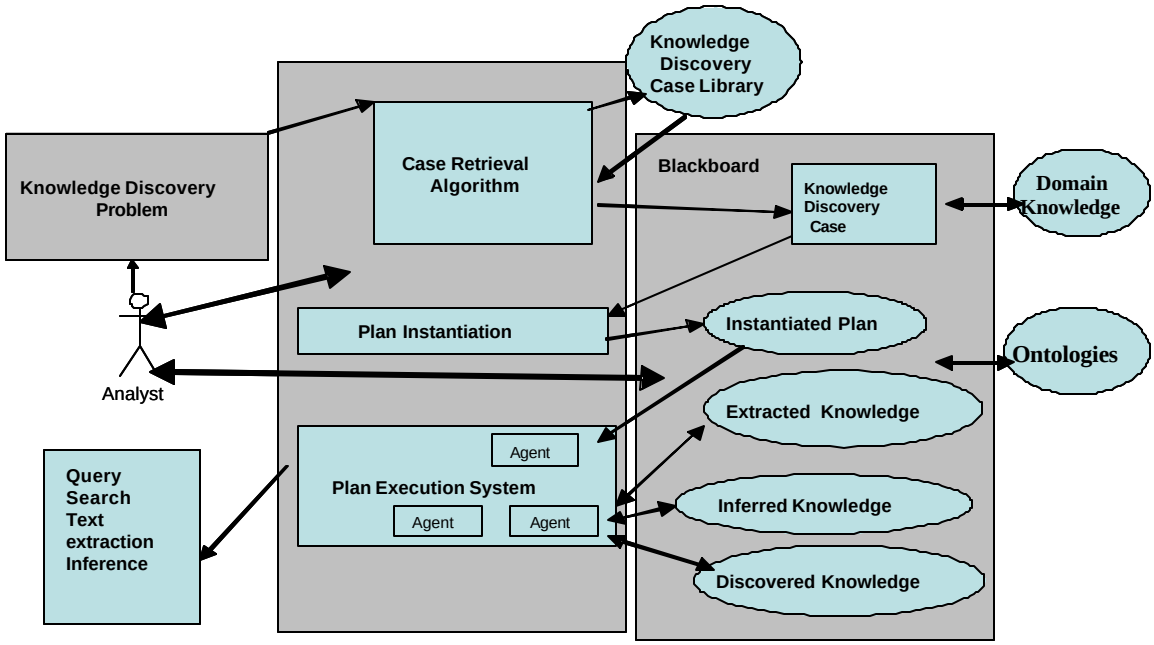

Figure1: CBR for KD Architecture

Preliminary Evaluation

We have yet to conduct a formal evaluation of our prototype software to substantiate our ability to improve an analyst's productivity, however, some of our preliminary data points in a favorable direction.

From our knowledge acquisition and analysis, we know analysts use commercial search engines to identify candidate documents by typing a few key words into the search interfaces. The analyst then opens and scans the retrieved documents to locate and extract relevant pieces of information. This process, besides being error prone, is time consuming and tedious. Our first level metrics are *time to locate candidate documents* and *number of candidate documents that had to be inspected to find relevant pieces of information*. In no case did our knowledge discovery prototype take more than a few minutes to deliver its results. Some initial summary statistics from early test samples are shown below:

Test 1: "Find clusters of biowarfare experts"
244 web sites searched
44 pieces of information in 37 distinct web sites
Highest web site number 238

Test 2: "Find information linking a known biowarfare expert to others."
208 web sites searched
55 pieces of information in 48 distinct web sites
Depth size 2

Test 3: "Find organizations affiliated with a known chemical weapons expert."
52 web sites searched
51 pieces of information in 26 distinct web sites
Depth size 3

For example, in Test 1 above we were using a multi-step knowledge discovery plan (results of information extracted from some steps are used to spawn new searches from which information is extracted and inferences made) in which we limited the plan execution to a few minutes, we searched (in multi-steps) 244 web sites, found 44 pieces of relevant information in 37 distinct web sites. The last piece of relevant information was found in site number 238 in the last step of the knowledge discovery plan. We estimate that this task would take an analyst many hours to complete. In these results, the term depth refers to the number of search engine results pages that an analyst would have had to inspect to locate a piece of relevant information. So for the Google search engine, which returns ten URLs per page, a depth of 2 means the analyst had to go through 2 pages of results and found the information on the third page (in absolute terms it means the result was found in URL number 21 to 30).

Conclusions and Acknowledgements

In this paper we provide some insight into our initial analysis and design of our application of CBR for Intelligence Analysis. Our current implementation has only a small set of initial cases that partially cover four types of knowledge discovery problems. We know that it needs considerable improvement both in terms of knowledge discovery plans as well as its ability to focus on the most relevant and reliable information. Analysts that have seen our prototype have encouraged us to continue. They never fail to provide suggestions for improvement.

This research is funded by the Advanced Research and Development Agency. The authors gratefully acknowledge our GTRI CBR for KD research team members: Laura Burkhart, Reid MacTavish, Collin Lobb and our Government COTR.

References

Hammond, K.. 1989. *Case-Based Planning: Viewing Planning as a Memory Task*. San Diego: Academic Press.

Heuer, Richards J.. 2001. *Psychology of Intelligence Analysis*, Center for the Study of Intelligence.

Jones, Morgan D. 1995. *The Thinkers Toolkit*, Three Rivers Press, New York.

Kolodner, Janet L. 1993. *Case-Based Reasoning*, Morgan Kaufmann.

Kolodner, J.; Simpson, R. 1989. The MEDIATOR: Analysis of an Early Case-Based Problem Solver, *Cognitive Science*, Vol. 13, Number 4: 507-549.

Knoblock, C. 1985 Planning, Executing, Sensing, and Replanning for Information Gathering. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Montreal, Canada.

Krizan, L. 1999. Intelligence Essentials for Everyone, Joint Military Intelligence College, Washington, DC.

Lesser, V., et al. 1999. BIG: A Resource-Bounded Information Gathering Decision Support Agent, UMass Computer Science Technical Report 1998-52, Multi-Agent Systems Laboratory, Computer Science Department, University of Massachusetts.

Nodine, M., Fowler, J, et al. 2000. Active Information Gathering in Infosleuth. *International Journal of Cooperative Information Systems* Vol. 9, Nos. 1 & 2: 3-27. World Scientific Publishing Company.

Whitaker, E. T. and Bonnell, R.D. 1992. Plan Recognition in Intelligent Tutoring Systems, *Intelligent Tutoring Media*.

Whitaker, E. and Simpson, R. 2003. Case-Based Reasoning for Knowledge Discovery. In *Proceedings of Human Factors and Ergonomics Society (HFES) Conference*, Denver, CO.