Meta-Data Mining for Optimized Aircraft Repair and Overhaul

Richard CLEMENTS, Jeffrey MORSE, Duane DARR and Bernard LASKOWSKI
Analatom Incorporated, 562 E. Weddell Drive, Suite 4, Sunnyvale, CA  94089
Tel: (408) 734-9392   Fax: (408) 734-8335
Email: bernard.laskowski@analatom.com

Abstract. The aim of this project is the development of advanced software which integrates analytical and modeling tools for data mining, maintenance support, and structural health monitoring prognostics by incorporating Text Analysis and Concept Mapping for extracting new information from data represented as free-form text within documents and maintenance repositories. The project will develop new modeling, optimization tools and algorithm concepts that provide database search, concept linkage, and correlations facilitating intelligent decision making processes for maintenance, repair, and overhaul work practices and scheduling. Ultimately, such a support tool will act upon current databases, meta-data, and repair practices to arrive at considerable personnel, parts, and other resource savings as well as shorter repair time horizons within the Maintenance, Repair, and Overhaul (MRO) environment. An aircraft maintenance and repair work scope optimizer, as a decision support tool, will utilize dynamic data, meta-data information, and domain knowledge to provide the repair work force with a daily work package that accommodates contingencies via dynamic re-planning. The decision support tool will be orderly and repeatable; being controlled using advanced AI techniques to identify associations within a dynamic information repository, the Information Cube.

1. INTRODUCTION

Mining of information is both a complex and highly inexact process! Numerical data is the more transparent, straightforward form of raw informational sources. Mining of text, however, while being more intuitively obvious and richer in cognitive understanding and interpretation, is the much more complex task and is less open to standardized extraction procedures and processes. The rewards of intuitive textural recognition are immediately burdened by the process of initially managing and structuring cognitive narrative extraction and meaning.

There exists a need to extract actionable, linked information and diverse domain specific knowledge over a wide number of disparate, independent, and autonomous maintenance database types and multiple repair storage repositories. Data gleaned from these multiple, individual database sources needs to be combed and filtered for common, appropriate, actionable information and knowledge that supports and enhances common multidimensional reasoning capabilities. The opportunity is to evaluate the feasibility of developing, evaluating, and deploying intelligent software—Artificial Intelligence Algorithms (AI)—that will autonomously create Information Knowledge Cubes across these multiple data repository sources.
Currently, aircraft maintenance, repair, and overhaul practices at most Air Force bases rely primarily on retrieving simple data sources without intelligently processing the data and developing requirements forecasts and repair packages which are optimized for specific tasks. This has an adverse effect on time to repair, as well as personnel and facilities utilization required to carry out the repair processes.

Data mining and meta-data analysis tools designed by Analatom are used to derive relevant data relationships and provide greater insight into the data content. This facilitates discovery by query and retrieval methods, and enables optimization tools to improve maintenance execution. A traditional Reliability Centered Maintenance (RCM) analysis typically takes 2 analysts 4-6 months to sort, filter, and scrub the 8 million plus records down to the principal components and failure modes. That is close to $100K of labor alone for each of up to 20 or more system analyses, not to mention the fact that there are not the resources to delve any deeper into exploring other information not associated with actual component failures. However, utilizing Analatom Information Cubes algorithms, one can drive down this time and cost burden to just hours or days.

2. METHODS, ASSUMPTIONS, AND PROCEDURES

Analatom successfully explored innovative ways of developing techniques and functionality through the use of AI driven, inexact (fuzzy), and non-linear techniques such as Self Organizing Maps (SOMs) for both information storage and knowledge retrieval that will enable analysts to break out of the boundaries of only being able to query data sets with questions and assumptions they already know, or assume they know, and assume are relevant. A proof-of-concept demonstration of these algorithms has been developed incorporating a proprietary set of software algorithms that include 1) Informational Self-Organizing Feature Maps, 2) Adaptive pruning of low yielding information clusters and similarity links, 3) Non-linear reconfiguring of associations and dependencies, and 4) Data Reduction Optimization Neural Nets.

Analatom has investigated appropriate technologies and algorithms which, when applied to large, disparate database sources and text repositories, optimize various reasoning approaches and paradigms. These new algorithms can be applied to gaining insight (Knowledge Mining) and selecting adaptable, actionable understanding (Information Descriptors) of non-linear and not intuitively retrievable content.

While the term Data Mining is often the first label associated with extensive probing of database repositories, it is by definition quite limiting. One does, in a very strict sense, probe or mine the data pool, but required objectives and utility go far beyond this simple, initial mechanical process. Meta-data mining, or Hierarchical organizational structures, addresses these base limitations and expectations. This task becomes additionally complex when the scope extends to Meta Text mining. In support of these advanced AI techniques and algorithms is the fundamental assumption—and user’s desire—to have “reasoning” criteria applied against these varied underlying pools of raw data. While retrieving data from massively large data sets for the intent of gathering information and knowledge, especially from data repositories that have incomplete, corrupted, or degraded data vectors, is a daunting and quite complex task indeed, Analatom has demonstrated an optimal approach to Information Mining based on its proprietary in-house technology and expertise of AI Algorithms for intelligent Information Mining of extremely large data sets and complex text repositories. This paper outlines and conceptualizes Information Cubes as a robust, adaptable modeling/optimizing framework that functions as the essential base and non-linear structure of a series of optimizers for maintenance planning, repair processing and sequencing, and aiding in decision support.
3. DISCUSSION

3.1. Identify Appropriate SOM and AI Text Analysis Software and Tools

The SOM paradigm is extremely efficient for association clusters and similarly linked topology networks, and will be used for both the initial retrieval of text narrative variables, as well as the more powerful process of constructing and building the many n-dimensional Information Cubes, the final step in Concept Cube Structures. We apply SOM-Ward and SOM Single-Linkage-Clustering algorithms, because the vector similarity metrics are addressable within the totally converged topology map. Inexact fuzzy queries of text and concepts within data and text repositories can be retrieved from neighborhoods of similar looking text vectors. SOMs for textural content grouping have proven efficient and beneficial, affording the end user the ability to see all possible associations of similar tasks and repair activities which are all too often hidden within large text data repositories, or are never even referred to because of lack of experience or some technical inability to make these subtle associations.

Maintainers have generated many segregated clusters containing similar maintenance records and matching unique tracking actions, narratives, and text. Thus, there needs to be a decomposition of these grouped elements and textural components that are contained within each record. Instead of manipulating digital or numeric numbers in the familiar traditional manner, appropriate text analysis techniques must be used to break down these extremely large chunks of narrative text and phrases into their representational contextual building cells. We are not concerned with defining a new grammar, or specifying some new topology of semantic networks, but with associating common activities, intentions, and concurrent actions together that can all be labeled and associated with a specific task identifier. Each cluster now represents a similar maintenance work procedure. Each record contained within that specific cluster will then represent a unique maintenance occurrence, over a broad period of time and at different depot locations, which falls under the broader classification of the parent cluster and is associated with it. Some of these specific pre-processing algorithms and transformations applied to this text decomposition process fall generally into the following classes of normalization.

(a) **Lemmatization.** This is used to reduce various word forms to a more limited set of words or word roots. Analatom uses the Krovetz’s KSTEM suffix substitution algorithm. Since this lemmatization algorithm does not rely on a prior part-of-speech tagging of words, it is much faster than traditional lemmatization routines.

(b) **Keyword-in-context.** All variables (columns) within each separate database are analyzed, searching for those classes of variables which are nominal or categorical, and the subsequent classes within category are extracted onto frequency tables representing key descriptive words used within similar or same maintenance procedures or events. They are then searched for within the much larger text chunks and narrative fields associated with each maintenance, and by using a recurrent metric of phrase length a matrix of ‘Keyword-in-context’ is built. This matrix of Keyword-in-Context is built having associated with it a weighting and projection for each keyword that occupies its textural space. We are now able to measure how ‘dominant’ or ‘significant’ a particular word is being used, based on the text or phrases or sentences it is being used in, not just having counted its occurrences.

(c) **Phrase Extraction.** The algorithm now scans all of any field that contains large text segments, phrases, or nominal categorical labels and keywords for centroid anchoring. This is a process whereby an adaptive radius (number of concurrent words on each side) of significant, linear phrases are examined for periods and frequencies of recurrence, based on
a central keyword. These extractions will all be composed of exactly the same ‘x number of element words’ surrounding each discovered keyword.

We have developed a beta model designed to break apart classes and phrases of text representations and core concept indexes such that extended narrative and descriptive text representations can be transformed and folded into much broader classes of concept tracing or information links to be ultimately used in reconstructing pieces of information that share and add to similar containers of common information or knowledge.

3.2. The Three Key Processing Modules and Analysis Functionality

The above critical investigative efforts form the core of this newly created Information Cube concept structure, and have produced the following three critical insights and benefits for continuing research and implementation: 1) What, 2) How, and 3) Where.

“What” elements and components within our target database extracts, (and that are represented both in data sources as numeric or as text, phrases, and narrative summaries) are common and similar to each other, i.e. this initial process clusters together hundreds/thousands of similar tasks, processes, and maintenance record based related objectives and allows similarities to be extracted from data repositories.

“How” is our second (processing) reasoning step that breaks down these often large text fields, phrases, and narrative sentences into a much more compact, efficient, and manageable form of linked-indexing. This step decomposes text into similar concepts and key maintenance procedures.

The third critical processing step is that of optimally identifying “Where” these new indexing schema and linked-node representations will be re-assembled and ultimately referenced within the Information Cube (an n-dimensional and non-linear AI software data structure); i.e., how do we construct concepts about specific classes of knowledge. We transformed query structures from initial flat 2-dimensional database representations into multi n-dimensional architectures of commonality and similarity. Chunks of information are no longer segregated or isolated, but are globally accessible and addressable through concept links.

3.3. Identify Appropriate Query Paradigms (Information Cubes)

SOMs AI algorithm is the most efficient, direct methodology for automatic, unattended adaptation of multi-relational concept links. It offers two distinct advantages. First, new information, in the form of yet unseen or unused databases, can always be added and incorporated into existing Information Cubes without the need to retrain the entire architecture from the beginning, saving both time and training efficiency. The Information Cube grows and learns from the state at which it was already trained to perform. Second, this AI algorithm affords the end user the broad flexibility and power to query these Information Cube structures in an inexact manner, allowing one to use fuzzy or concept based keywords as a first pass into many much larger organizational knowledge pools.

We demonstrated a proof-of-concept model such that by applying inexact (or even missing words) text phrases and keywords to multiple Information Cubes, one can also retrieve neighborhoods and groupings of similar maintenance activities, repair procedures, and associated elements (parts and processes) for observation and evaluation. We described how similar maintenance actions and depot activity events are grouped and clustered. We also have designed several key micro level algorithms that can further decompose chunks of text and narrative descriptors into basic building blocks or elements. The final step of this process is the task of reconstructing a new class of a fluid architecture composed exclusively of these newly discovered and uniquely assembled key components.
We now describe how to move from traditional numeric relational database indexing schemas to a new content and concept oriented organization paradigm. The Information Cube allows storing and querying enormous amounts of dissimilar data elements that are now transformed and represented as text, phrases, and narrative. The procedural intent is to allow the user to query the Information Cubes with text strings, keywords, or incomplete narrative textural descriptions that generalize or indirectly point to information needed; the burden of retrieving associated maintenance records being handled by the SOM.

3.4. Test Information Cubes and Methodology Concept for Full Database Analysis

While the available sample data set extract from the database pool totaled hundreds of thousands of individual maintenance and repair records, the initial beta model queries focused on a very small randomly chosen subset of those records. Smaller subsets of records will stand as proxies for similar retrieval functionality upon similar but larger data sets. Graphs, tables, and descriptive statistics were provided to illustrate the effectiveness of the innovative approach of retrieving relevant chunks of information initiated by fuzzy and/or sparse classes of database queries.

Random data extracts from Database One (378,953 records) and Database Two (23,947 records) database repair repositories served as domain data sources for all our maintenance development records. These extracts were each split into training (70%) and testing groups (30%) partitions. During this analysis period, the testing partitions from both sample groups were folded into an Information Cube beta model. SOMs are generated from each of these sample extracts and all non-linear, indirect links are subsequently developed. Next, multiple blind data set validation queries were performed against the Information Cube. Statistics were collected on all retrieved text queries that compare and describe the accuracy of similarity between the query text and the retrieved text. Performance metrics of cross-validation testing primarily centered on applying various types of data validation and blind testing of random record extracts from the sample data records. Cross-validation and blind testing included 1) cross-fold validation, 2) X of N hold-out records, 3) multiple hold-out subsets, and 4) random test record extractions; being the most appropriate types of methodologies available for validation testing against query constructs and constraint strings, which are themselves an inexact or non-linear and fuzzy query process. Test/validation accuracy and AUC (Area of SOM Units Covered) have been performed for multiple queries into a test data extract that had been partitioned into six folds of pooled text records. The test illustrated that both the accuracy (query to text relevance) as well as the high diagonal coefficients from the confusion-matrix show consistent performance of relative importance. High recall and relevance coefficients were also obtained from validation models on different matching criteria (Tree folds, Rule-Set, and Ensemble-Modeling folds).
5. RESULTS

Figure 1: Maintenance Record Data Mining - SOM Organization of Activities

Figure 1 shows how similar maintenance records are grouped together based on tasks.

Figure 2: Trace Mapping: Commonly Linked Maintenance Activities

Figure 2 shows how work codes and repair tasks are linked to procedure ranked activity paths.

Figure 3: Representation of Deductive Dependency Failure Analysis

Figure 3 shows that each maintenance task has specific dependent activities.
Figure 4: Sequential Traces of Fault/Failure Links of Sensor Diagnostics:

Figure 4 shows that different maintenance procedures and activities follow specific process activities.

Figure 5: Data Mined Workflow Overview:

Figure 5 shows, for the common maintenance and repair procedures on C-130 props, a global dependency network displaying drill-down from top-level classes to specific terminating representations.

Figure 6: Resulting Weighted Directed Graph
Figure 6 illustrates a specific class of sequential diagnostic repair maintenance procedures. A root-cause-failure filtered network is shown demonstrating dependency linked activities associated with a specific failure.

6. CONCLUSIONS

A typical work-flow involves the following steps (A), (B) and (C) which are duplicated separately for each high-level or individual class of Information Cube being constructed, each with maintenance and repair expertise, and procedural knowledge about a specific, high-level component or maintenance process.

(A) All appropriate databases, maintenance pools, and digital tracking mediums are gathered together. A SOM extraction module is applied to this pool of databases so that similar maintenance records, comments, and tracking descriptors are clustered together in segregated containers. Groups of similar maintenance activities, components, and elements are in uniquely and individually addressable containers. This initial process need only be done once, unless additional data sets are later added.

(B) Each of these similar maintenance and repair procedures and clustered activity pools are then addressed individually. Here we deal almost exclusively with text, various lengths of narratives, and key concept phrases. The concept/context building blocks are extracted for each maintenance activity and depot record of interest. This step can be likened to extracting and coding individual cells from much larger conceptual strands of text or narrative. This analysis step is completely isolated from the end user; it is performed in the background and is totally automated.

(C) The newly formed, uniquely constructed building blocks of content, concept, association, and similarity are re-linked, reconstructed together in a new n-dimensional representation such that any node or element within that multi-dimensional system can represent a key concept or piece of information, and as such each and any word, any phrase, or groups of narrative words are capable of retrieving all of their neighbors as pertinent, or at least similar to, a specific user query. This allows a user to offer a specific phrase, a key word, or even a moderate length narrative as a query and have returned all records that are a same match, as well as those records that are associated with or similar to those query criteria. In other words, if we query about part or element K, we get back maintenance information and documentation not only about K, but also can view similar maintenance and repair activity information about elements H, I, J, as well as L, M, and N. Users with lower levels of query background/expertise can extract similar and supportive maintenance actions, being no longer limited by the level of prior or specific experience.

Ultimately, information can only be meaningful if it has been allowed to evolve beyond a simple retrieving and ranking of chunks of data elements (both numeric and textural) into associations, links, and clusters of similarly based concepts. Remember, it is easy to manipulate and iterate over numbers, but it is rewarding and useful to reason about and evaluate Concepts and grouped Information Collections.

7. ACKNOWLEDGMENTS