Active Templates: Comprehensive Support for the Knowledge Discovery Process

Randy Kerber, Hal Beck, Tej Anand, Bill Smart
NCR Human Interface Technology Center
5 Executive Park Dr. N.E.
Atlanta, GA 30309
{Randy.Kerber}, {Hal.Beck} {Bill.Smart} @AtlantaGA.NCR.com, TAnand@GoldenBooks.com

Abstract

The goal of Active Template research is to create a single, unified environment that a data analyst can use to carry out a knowledge discovery project, and to deliver the resulting solution in the form of an Active Template. An Active Template is a hyper-linked information structure that tightly integrates *actions* (executable programs and commands), *results* (models, datasets, predictions, reports), and *documentation* (explanations of decisions, actions, and results). The use of Active Templates provides a number of benefits, including user guidance, improved documentation of actions and results, and increased reuse of previous work.

Introduction[1]

Data mining success stories have triggered increased interest within the business community, particularly in large corporations with vast stores of data about their customers and business operations. Their interest appears to be following a path similar to that of early research in machine learning: the tendency to view data mining as the isolated application of a data mining algorithm to a pre-existing dataset, where the key determinant of success is selecting (or creating) the "best" model-building algorithm.

As businesses continue to use data mining technology, they are likely to discover, as experienced practitioners and researchers already have, that:

- There is usually little difference in accuracy between modeling algorithms.
- Availability of useful data, dataset preparation, and user skill are more important than which algorithm is chosen.
- Model development is more properly viewed as a multi-step *process*, of which application of the modeling algorithm is only a small part (practitioner's informal estimates tend to range from about 10% to 30% of the total effort).

The field of knowledge discovery emerged largely driven by the desire to place a greater emphasis on the process than was true in earlier research. Within knowledge discovery, three distinct directions have emerged to support this process-oriented focus:

- **Integration.** The objective is to provide the user with a single tool or environment that includes several of the following capabilities: modeling, data access, data manipulation, cleaning, exploratory analysis, visualization, and testing. Examples include widely used statistical packages such as SAS (SAS Institute), SPSS (SPSS), Statistica (StatSoft), S-Plus (MathSoft), Enterprise Miner (SAS Institute), Clementine (ISL), Intelligent Miner (IBM), MineSet (SGI), Recon (Kerber, Livezey & Simoudis 1995), MLC++ (Kohavi), and INLEN (Kaufman, Michalski & Kerschberg 1991).[2]
- **Interactivity.** The objective is to improve results via a high degree of interaction between the human user and the software, exploiting the strengths of each. Examples include (Selfridge, Srivastava & Wilson 1996), (Piatetsky-Shapiro, Matheus 1991), (Derthick, Kolojejchick & Roth 1997).
- **Methodology.** Here the objective is not to provide tools, but to prescribe a valid methodology for an analyst to follow in order to properly execute a knowledge discovery project. This typically involves defining a process model consisting of a set of tasks for the analyst to perform. Previous work includes (CRISP-DM), (Brachman & Anand 1996), (SAS Institute), (Reinartz & Wirth 1995), (Wirth, et. al. 1997).

Additional Business Requirements

Increased attention to business applications has necessitated even more requirements for knowledge discovery projects. The most important additional requirement is that for business applications, there are many deliverables besides the resultant analytic model. For example:

- **Business requirements.** There is a need to determine how the model will be deployed, including integration with the customer's business operations, compliance with regulations and corporate policies, and a maintenance plan for monitoring a model's performance and determining when and how to update the deployed models.
- **Guidance and training.** Performing a knowledge discovery project is a difficult, complex undertaking. There are many steps to perform, many pitfalls to avoid, and many opportunities for subtle mistakes with dire consequences. Even the most experienced analysts can easily forget important steps if they lack a checklist to follow. Because the demand for experienced analysts is greater than the supply, many inexperienced people are

[1]

[2] Information on these systems is available at http://www.kdnuggets.com

being called upon to perform data analysis. These inexperienced analysts, in particular, are in need of training, guidance, and support tools.

- **Documentation and communication.** There is a need to capture and communicate results, methods, risks, assumptions, justifications, rationale, etc. The business customer needs to understand the significance and scope of the results. Systems people will need to integrate, monitor, and maintain the models. Other analysts may need to validate, update, or repeat the work. Currently, assumptions and the rationale behind important decisions are frequently forgotten, because tools to document the process are not integrated with the analysis tools, and thus are not readily available and convenient to use. The analyst needs to capture this vital information *during the engagement*, not just when it's time to prepare the final report.
- **Reuse.** A knowledge discovery project is usually not an isolated effort. The customer will probably want to periodically update the model several times in the future as new data becomes available, and will generally have several other related business problems that could benefit from a knowledge discovery project. There is a *strong* desire that each of these new efforts not require 100% of the initial effort and expense. In practice, it is very difficult to exploit the work performed in previous engagements, even when a great deal of similarity is evident, because the only surviving artifacts are the final programs and outputs (i.e., the familiar "piles of files"). The steps the analyst performed and the reasoning that went into them are typically lost. For example, without an explanation as to why certain records or fields were excluded from a previous model, an analyst asked to create a more current model is likely to be reluctant to exclude then without first spending time carefully reevaluating their relevance.

Active Templates Overview

The primary objective of Active Template technology is to provide a comprehensive, open, interactive environment for the development and delivery of knowledge discovery solutions. An Active Template contains three primary types of content: *actions*, *results*, and *documentation*. The analyst creates *Active Templates* using the *Active Template Environment*, which consists of a *Visual Programming Environment*, an *Active Template Editor*, a set of *Analytic Tools*, and a *Resource Library* containing *Encyclopedias*, *Guides*, *Active Template Libraries*, and *Frame Libraries*.

Section 4 covers the Knowledge Discovery Process Model and its role in Active Templates. Section 5 provides a more detailed description of the components of an Active Template. Section 6 describes the Active Template Environment, the tool the analyst uses to create and manipulate Active Templates.

Knowledge Discovery Process Model

A KDPM (Knowledge Discovery Process Model) defines a rigorous process for executing knowledge discovery projects. It is a structured definition of the tasks, at several levels of detail, to be carried out to complete a KD engagement. At the top-level, the process model is described in terms of a sequence of *phases* of the project. Each phase is then broken down into a number of tasks, which might also be subsequently broken down into sub-tasks, sub-sub-tasks, etc. Each task consists of problems that the analyst will encounter, issues that need to be addressed, actions that should be taken, decisions that need to be made, and milestones that should be achieved upon successful completion of the task.

While the tasks of the KDPM are described as if performed in a certain order, it is understood that in practice there is a need for much iteration and backtracking. The specified ordering should be treated as a suggested, idealized ordering; that is, the specified ordering represents one, but by no means the only, feasible path towards developing a model.

Within the Active Templates paradigm, the KDPM serves as the common, structural framework for organizing its components. For example, it is expected that the analyst will construct Active Templates to follow, as closely as possible, the task structure of the KDPM. The supporting components of the Resource Library, such as the reference Encyclopedias and Guides, are also organized according to the structure of the KDPM.

The other major role of the KDPM is to provide project guidance. The KDPM functions as a roadmap through the knowledge discovery process, providing all project participants with a view as to which tasks have been completed and which remain to be done. The details of the process model also provide the analyst with a checklist of actions and requirements.

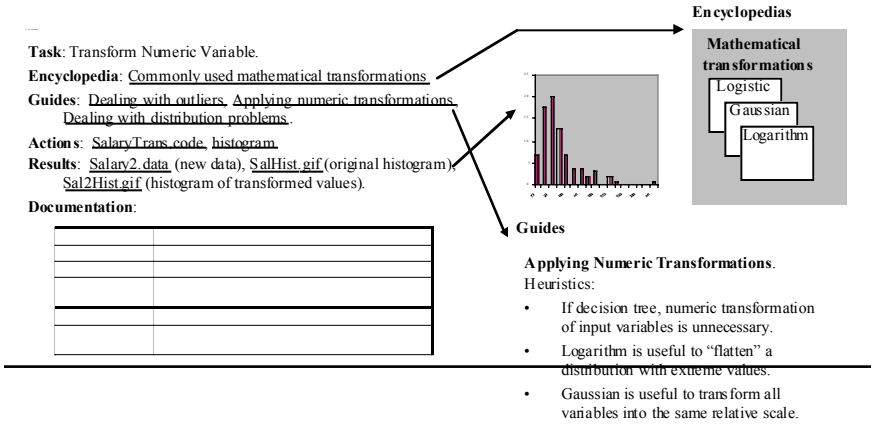
While a specific implementation of Active Templates must be tightly integrated with a certain process model, the ideas do not require that a particular process model be adopted. Whatever process model is adopted determines how the components of Active Templates are organized.

We are actively involved in efforts to develop knowledge discovery process models, including one of the first published models (Anand & Brachman 1996). Currently, NCR is a member of the CRISP-DM (CRoss Industry Standard Process for Data Mining) consortium, along with DaimlerBenz, ISL, and OHRA. The goal of CRISP-DM is to develop a standard data mining process model, through feedback obtained from the data mining community via workshops and a Special Interest Group (see CRISP-DM).

Active Templates

Active Templates are information structures, built by the analyst using the Active Templates Environment, containing the results of a knowledge discovery engagement. Active Templates are built to follow the

methodology and structure defined by the KDPM. The three primary types of content are *actions, results,* and *documentation* (described below). An Active Template is also integrated with relevant parts of the Resource Library, such as Encyclopedias and Guides.



Actions

Actions are the executable (*active*) components of an Active Template, including commands (provided within the environment), external applications, or scripts. Essentially, any executable that can be called from a command line could become an *action*, such as scripts or programs written in Perl, C/C++, or SAS. In cases where the desired action cannot be executed from a template, step-by-step directions can be included specifying how to carry out the action (e.g., call Joe in the IS department and ask him to extract the prior year sales records).

Results

This includes the direct results of executing *actions* on a dataset, or any other tangible result of the knowledge discovery process. The following categories of results can be identified:

- **Models.** The result of applying a modeling algorithm, such as linear regression or clustering, to a dataset.
- **Data.** The data used to build models. Many tasks involve executing actions that take a dataset as input and produce a new dataset as output.
- **Reports.** This includes other outputs of Analytic Tools, such as variable summaries, plots, and accuracy estimates.
- **Findings.** Miscellaneous pieces of knowledge learned during the analysis. For example, "*salary is more predictive than years-employed*", "*moderate ATM users are the most profitable customer segment*".

Documentation

Documentation covers the remaining text-based contents of an Active Template, created by the analyst to describe actions and results. The documentation component can be thought of as an *electronic engineer's lab book*, into which the analyst can record *any* bit of information he/she feels might be important. Examples of the types of information in the Documentation are: explanations of results, important assumptions, risks, reminders of tasks to address later in the process, the rationale behind important decisions, discussion of alternative courses of action, and discussion of findings that will be included in the final report.

Active Template Environment

The Active Template Environment is the tool that the analyst uses to develop and deliver analytic models. The major components of the environment are the Visual Programming Environment, Analytic Tools, Active Template Editor, and Resource Library (all described below).

Visual Programming Environment

The VPE (Visual Programming Environment,) is a graphical development environment that the analyst uses to create and view templates, perform actions, or access any of the other components of the system.

The analyst constructs templates by creating a data-flow diagram consisting of nodes, which contain actions, results, or documentation, and directed connections between nodes, which indicate the flow of data objects. Arrows pointing into a node indicate the input data for that node. Arrows pointing out of a node indicate the output data, or other results, such as reports, plots, or analytic models.

In the initial Active Template work, we have used the *Khoros* VPE (see Khoral Research) as the base environment for using Active Templates. ISL's *Clementine* and SAS's *Enterprise Miner* are examples of data mining tools based on the visual programming data-flow paradigm.

In our view, it is critical that the VPE be open and readily extensible. With an open architecture, the analyst's toolkit can be continuously expanded with new modeling or data manipulation applications, while maintaining the benefits of a single interface and environment to view all phases of a knowledge discovery project.

Analytic Tools

Analytic Tools refer to the typical software tools found in an analyst's toolkit, such as statistical packages (SAS, S-Plus), query tools (SQL), programming and scripting languages (Perl, Awk, Java, C/C++), and specialized modeling tools (CART, neural networks, plotting).

Active Template Editor

This is the tool for creating, editing, viewing, and manipulating Active Templates. The VPE provides the capability to manipulate the visual elements, such as the nodes and links in the data-flow stream. The

documentation components are primarily HTML-based, to provide portability, flexibility, and hyper-linking between objects. Thus, the documentation elements can be edited with any HTML editor that can be launched from within the VPE.

Resource Library

The Resource Library includes miscellaneous tools and reference materials useful to an analyst. The components—Encyclopedias, Guides, Frame Library, and Active Template Library—are all accessible from the VPE and integrated into the Active Template framework. For example, an Active Template in the Active Template Library might describe how Principal Components Analysis was used in that project for dimensionality reduction. This completed template might also contain links to the Encyclopedia entry that describes Principal Components, to the appropriate task of the KDPM, to the Guides that discuss hints for the proper use of Principal Components, and to the Guide that discusses advice for dimensionality reduction. The following sections describe the components of the Resource Library in more detail.

Encyclopedia Library. The Encyclopedias are reference guides that describe analysis algorithms, techniques, and principles. Each Encyclopedia includes a description of the technique, it's purpose, situations in which it is applicable, important assumptions it relies on, and how to interpret the results. The Encyclopedia Library includes entries for modeling methods such as decision trees, neural nets, K-means clustering, nearest neighbor, etc. There are also entries for other analysis tools, such as ANOVA, Principal Components, Factor Analysis, and various commonly used data transformations.

Guides Library. Guides contain informal insights, suggestions, discussion, and advice regarding how to approach certain analysis problems or use certain techniques. Each Guide includes a description of the topic, discusses the possible problems, describes possible solutions, and describes the pros and cons of each possible solution, along with suggestions for how to think about the problem. Each guide also includes a set of heuristics, or *rules-of-thumb*, for dealing with the problems. The following are examples of the kinds of topics that would be addressed by Guides:

- What to do about missing values.
- How to approach the issue of sampling.
- Selecting parameter values for decision tree induction.
- What modeling tool is appropriate for a given situation.
- The importance of defining the business problem.

The Encyclopedia is intended to supply the more formal type of information about analysis methods. It can be thought of as playing the role of a *reference manual*, whereas the Guides are intended as more of a *user's manual*, which includes informal discussions and advice. Thus, the Encyclopedias will be more fact-oriented while the Guides will be more advice and opinion-oriented.

Frame Library. Frames are pre-defined, generic sections of Active Templates that the analyst can use as building blocks to accelerate the process of creating new Active Templates. They can be thought of as *blank forms* that provide the structure of a template, leaving the specific content to be added by the analyst. For example, a pre-defined Frame could include a set of questions for the analyst to answer (possibly including a menu of answers), links to the appropriate Encyclopedias and Guides, links to previously used scripts for accomplishing that task, and directions for how to customize the generic Frame. The normal process for creating new frames to add to the Frame Library, is to start with a completed portion of an Active Template, and remove the specific content while preserving the structure and the parts most likely to be reusable.

Use of frames provides a number of benefits:

- Analysts will be able to create template more quickly if they can start with pre-defined frames.
- The parts of the frame can function as checklists of actions for the analyst to perform or factors to consider.
- Use of standardized frames will result in greater consistency between completed Active Templates in an Active Template Library.

Active Template Library. Perhaps the greatest potential benefit of the Active Template paradigm is the promise of reuse. Currently, most data mining engagements are essentially "One of" projects, where the bulk of the work is custom work for a specific customer. Even though the analysts (and customers) can recognize a great deal of similarity between projects, it is currently difficult to easily reuse the results of previous engagements. A primary goal of Active Templates is to greatly increase the degree to which an analyst can exploit the results of previous work.

Over time, an organization performing a number of data mining engagements can build up a *library* of completed templates. It should then become increasingly likely that new engagements aren't really new; there will be a template, or part of a template in the organization's *Template Library*, that partially, or wholly, addresses the current situation. A situation, and the appropriate context of a template, can be described in terms of a small number of dimensions. Identified dimensions (along with a few examples) include:

Industry	**Business Objective**
Insurance	Customer retention
Retail	Fraud detection
Banking	Market segmentation
Government	Sales
Forecasting	

Modeling Objective	**Modeling Approach**
Prediction	Decision Tree
Clustering	Neural Network
Associations	ANOVA
Understanding	Association Rules

Thus, analysis situations, and completed templates in the Active Template Library, can be described in terms of combinations of these dimensions. For example, if the new project involves creating a predictive model for a bank to

predict *propensity to respond* for a credit card promotion, an existing template that solved the same problem for a previous bank customer would be extremely useful. Lacking that, an existing template for creating a *propensity to respond* model for a different industry could be quite useful. In addition, a template to create a clustering model for a bank, though addressing a different problem, might be very useful for data preparation tasks (assuming the data structure is similar).

In addition to templates created from previous engagements, it can be useful to create *general templates*, to simplify the process of reuse. General templates might be created for certain types of problems (e.g., clustering bank customers), for modeling methods (e.g., decision tree classification models), or for certain situations (e.g., aggregating transaction records into summary records). Then, when starting a new engagement, rather than starting from scratch, the analyst can begin by obtaining templates, or portions of templates, and adapt them to fit the new situation.

Conclusion

Active Templates provide a methodology and tools for developing and delivering knowledge discovery solutions to business customers. Such customers require a number of deliverables beyond an analytic model, support that is provided by Active Template technology. The use of Active Templates for knowledge discovery provides a number of benefits:

- **Guidance and training.** The Process Model and predefined Template Frames serve as task checklists, reducing the likelihood of inadvertently skipping a crucial step. The Guides and Encyclopedias provide helpful background information and advice. All of these components, along with completed Active Templates, can also be used as valuable training tools.
- **Documentation and communication.** Because of supportive tools, *active* during the engagement, the analyst is much more likely to record explanations of important decisions and risks, rather than counting on remembering them when it's time to create the final report. Improved documentation of the process improves communication of results, important assumptions, and risks to the business customer, who must use, understand, and maintain the project results after project completion.
- **Reuse.** A completed, working, documented Active Template provides a tremendous head start when updating an existing model with new data or when undertaking a new modeling effort.

References

Brachman, R. J. and Anand, T (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In: U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, Chapter 2. AAAI Press: Menlo Park, CA.

Brunk, C., Kelly, J., Kohavi, R. (1997). MineSet: An Integrated System for Data Mining. In Proceedings of the Third Intl. Conference on Knowledge Discovery and Data Mining. Newport Beach, California: AAAI Press.

CRISP-DM. Cross Industry Standard Process for Data Mining. http://www.ncr.dk/CRISP/

Derthick, M., Kolojejchick, J., and Roth, S. (1997). An Interactive Visualization Environment for Data Exploration. In Proceedings of the Third Intl. Conference on Knowledge Discovery and Data Mining. Newport Beach, California: AAAI Press.

Fayyad U. M., Piatetsky-Shapiro G. and Smyth P. (1996). From Data Mining to Knowledge Discovery: An Overview. In: U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, Chapter 1. AAAI Press: Menlo Park, CA

IBM. http://www.ibm.com/bi.

ISL, Integral Systems, Ltd. http://www.isl.co.uk.

Kaufman, K., Michalski, R., and Kerschberg, L. (1991). An Architecture for Integrating Machine Learning and Discovery Programs into a Data Analysis System. In AAAI-91 Workshop on Knowledge Discovery in Databases. Anaheim, California.

KDNuggets. http://www.kdnuggets.com

Kerber, R., Livezey, B., Simoudis, E. (1995). A Hybrid System for Data Mining. In Intelligent Hybrid Systems, Wiley, New York.

Khoral Research, Inc. http://www.khoral.com

MathSoft, Inc. http://www.mathsoft.com

Piatetsky-Shapiro, G. and Matheus, C. (1991). Knowledge Discovery Workbench: An Exploratory Environment for Discovery in Business Databases. In AAAI-91 Workshop on Knowledge Discovery in Databases. Anaheim, California.

Reinartz, T. and Wirth, R. "The Needs for a Task Model for Knowledge Discovery in Databases", Statistics, Machine Learning and Knowledge Discovery in Databases, workshops notes from 8th European Conference on Machine Learning, 1995

SAS Institute. http://www.sas.com.

Selfridge, P., Srivastava, D., Wilson, L. (1996). IDEA: Interactive Data Exploration and Analysis. In Proceedings of SIGMOD 1996.

SGI. http://www.sgi.com/Products/software/MineSet

SPSS. http://www.spss.com.

StatSoft. http://www.statsoft.com.

Wirth, R., Shearer, C., Grimmer, U., Reinartz, T., Schloesser, J., Breitner, C., Engels, R., and Lindner, G. (1997). Towards Process-Oriented Tool Support for KDD. In Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97).