ACTUATE
The BIRT Company

Technical
White Paper

BIRT Analytics 4.2
Technical Summary of New Features

BIRT Analytics 4.2 introduces powerful analytic algorithms, support for new workflow capabilities, and several usability enhancements to the loader module. With a broader portfolio of algorithms the user can uncover insights from patterns hidden deeper in the data. The addition of workflow makes it possible to automate a series of actions so that insights can generate benefits immediately. Campaign workflow automates the campaign management steps from the initial definition to execution and analysis of results. BIRT Analytics 4.2 automates critical business functions with the most appropriate analytic algorithms and best practices workflows. This release of BIRT Analytics reveals more insights from the data and dramatically shrinks the time to benefit.

Notice

Version 1 – June 2013

Actuate Corporation
951 Mariners Island Boulevard
San Mateo, CA 94404
Tel: (888) 422-8828
http://www.actuate.com

Table of Contents

Introduction

BIRT Analytics provides fast, free-form visual data mining and predictive analytics. The uniqueness of BIRT Analytics results from a combination of the ease of use of data discovery tools with the power and sophisticated analytic products typically reserved for data scientists, and the operational and management reach of ActuateOne.

In release 4.2 we have added advanced functionality in a number of areas:

- Enhanced analytical horsepower through the addition of decision trees and association rules algorithms
- Automation for data pre-processing and preparation before the data is analyzed via the data mining models
- Advanced workflow capabilities that automate business processes related to analytics and execution based on analytics
- Campaign workflow for defining, validating and executing the stages of a campaign, including stages for the analysis of campaign responses
- Usability enhancements to the data loader module (qLoader)

Preprocessing

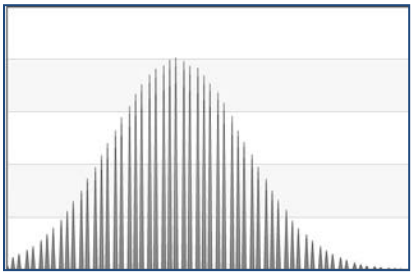
Preprocessing prevents distortions in the analysis by normalizing the effect of different scaling in the data. In a data set, if age ranges from 0 to 100 and bank balance ranges from 0 to 10,000, then changes to an individual's bank balance will dwarf the slow and steady change to age. This distortion can be removed by normalizing both age and bank balance to the same range, typically [0,1].

Preprocessing can also reduce the volume of data to process, and thereby improve performance. For example, a table with customers' bank balance will likely be very sparse, with a high variety of values. Preprocessing (rescaling) can be used to significantly lower the number of different values. So, a clustering algorithm will find the centroids much faster because the algorithm will need fewer distance calculations.

The preprocessing techniques in BIRT Analytics can be used to normalize a range of values while maintaining their relative distribution.

BIRT Analytics supports five different preprocessing techniques: Normalization; Linear Scaling; SoftMax Scaling; Logistic Scaling; and Remapping.

Normalization

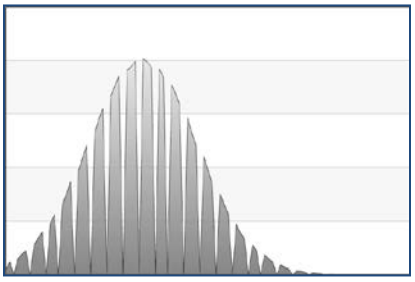


Variable normalization requires taking values that span a specific range and representing them in another range. The standard method is to normalize variables to [0,1]. This may introduce various distortions or biases into the data and out of range values.

Normalization allows us to know, for each value of the column, how many standard deviations are away from their mean, relating the value of each column with the standard normal distribution N (0, 1), by subtracting the mean from each of the column values and dividing it by the standard deviation. Each value is represented as the number of standard deviations away from the mean per the following formula:

$y = (x - \text{mean}\{x_1, x_N\}) / (\text{stdv}\{x_1, x_N\})$

Linear Scaling



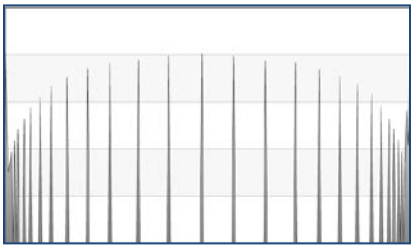
Linear Scaling is the simplest method for normalizing data values. The Linear Scaling formula only requires the minimum and maximum values of variables as input:

$y = (x - \min\{x_1, x_N\}) / (\max\{x_1, x_N\} - \min\{x_1, x_N\})$

A key benefit of Linear Scaling is that it does not introduce any distortion to the variable distribution. Linear Scaling has a one-to-one relationship between the original and normalized values. Although this method provides an undistorted normalization, it suffers from out-of-range values, i.e. new values that fall outside the minimum to maximum range.

An optional step allows users to apply a 'stretch' to the data by specifying new minimum and maximum values. The new minimum and maximum values are specified as percentages.

SoftMax Scaling



SoftMax Scaling a useful method for squashing unevenly distributed data into an appropriate range. The degree to which each value is squashed depends on the distance from the mean (we want to squash values which are far from the mean more than those which are close) and the standard deviation of the data set (a larger standard deviation requires a larger degree of scaling).

Values which fall further from the mean are squashed to an exponentially greater degree. The softmax function will squash any values which fall outside the range for which it is designed to a value close to one or zero. If the original data is evenly, linearly distributed, then all of the data points will lie within three standard deviations of the mean. Softmax can consequently squash the entire range into a linear part of the squashing function. Thus there is no need to choose whether a linear or softmax squashing function is used as softmax is equally valid for linearly distributed data.

Softmax is based on the logistic function:

$$x' = (x - \underline{x})/(\lambda(\sigma/2\pi))$$

where $\underline{x}$ is the mean of x , σ is the standard deviation, and λ is the size of the desired linear response. The linear part of the curve is described in terms of how many normally distributed standard deviations are to have a linear response.

Softmax scaling achieves a smooth asymptotic curve, i.e. values move in the direction of their minimum and maximum without ever fully reaching them. Softmax scaling tries to reduce to a range as narrow as possible the maximum, on one hand, and the minimum, on the other.

Within this option there is the possibility of establishing confidence intervals by selecting, in advance, a level of confidence (68%, 95%, or 98%). This is accomplished by specifying different values for λ, which represents the level of confidence (λ=1 for 68%, λ=1.96 for 95%, λ=2.58 for 99%). The lower the confidence level the shorter the intervals but the greater the probability of error.

Logistic Scaling

The logistic function transforms the original range of [-∞,∞] to [0,1] and also has a linear part on the transform. The values of the variables must be modified before using the logistic function in order to get a desired response. This is achieved by using the following transform:

$$y = 1 / (1 + e^{-x})$$

Remapping

Occupation Remap	Occupation
1	Director
2	Houseperson
3	Manager
4	Manual Worker
5	Office Worker
6	Professional
7	Retired
8	Self Employed
9	Senior Manager
10	Shop Worker
11	Unemployed

Remapping allows users to convert a categorical variable into numerical variable. It uses the distinct function so categorical values will be remapped in alphabetical order.

An example use case is the remapping a gender field from ‘male’ to 0 and ‘female’ to 1. Another example is the remapping of an occupation field to a corresponding numeric value.

Advanced Analytics Features

Decision Tree

Decision trees provide data classification algorithms with several applications. Marketing departments can use decision trees to target campaigns more effectively. Sales departments can use decision trees to more quickly pull a solution together for the prospect. These types of applications prescribe the most appropriate course of action based on a large number of parameters, such as client's age, job type, marital status, education, etc.. Non-commercial applications in research also use decision trees as an analytical tool.

BIRT Analytics implements a version of the C4.5 Decision Tree algorithm. To build the decision tree, BIRT Analytics successively splits a data set based on its strongest attribute. Building nodes of the tree by doing this successively and ascribing probabilities to the outcomes will result in a tree with leaf nodes for which there are no remaining attributes that provide a meaningful split. The tree provides alternative paths with probabilities associated with them. Once the decision tree has been developed, it can be validated via a test data set and then applied as an algorithm for scoring new data.

The algorithm's knowledge grows as it is trained. Once an effective set of classification rules is developed, the user can apply these learned rules to the entire universe of data. The final result can be a discrete value representing the chosen classification (ex: find the best products to recommend to our customers).

The following constitute the major steps in creating decision trees:

- Define the training segments (data that will be used to train the algorithm)
- Define the test segments (data that will be used to test the algorithm)
- Specify the attributes that will be used to split the segment at the nodes of the tree
- Specify the classificator – once the trained algorithm is used for scoring new data, the classificator codes the outcome on the basis of user-defined thresholds (ex: red, yellow, green indicators for the profitability of a segment)

BIRT Analytics 4.2 provides rich visualizations for quickly interpreting the results of the decision tree analysis.

Decision tree analysis is favored by many users because of the transparency of the analysis – users can see and understand what the algorithm is doing at each stage. Furthermore, a model can be tested, which enables users to place the appropriate level of confidence in the data they are using.

Association Rules

Association rules are used for determining which events tend to occur together. The most popular application of this is market basket analysis, in which merchandizers can determine which set of products customers tend to buy together. An example of the type of insight that can be gleaned with association rules is the following: if a customer buys wine, then buys beer, he/she will buy cheese in the future.

Inputs to the algorithm specify the minimum support and the minimum confidence level. Support is the desired number of transactions that must contain all the items in a rule. Consider a rule consisting of a conditional series of events, i.e. if event A occurs, and then B occurs, then there is high confidence that C will occur. The support can be stated as:

- Support(A,B)= Transactions(A,B)/Total transactions

In other words, from the total population of transactions, is the portion that contains the antecedent events significant enough. If support is not adequate, then the data is not expansive enough to support the rule. Minimum support is set to 10% by default.

Minimum confidence reflects the likelihood of the outcome (the likelihood that C will indeed occur if both A and B have occurred). Confidence can be stated as:

- Confidence (A,B → C)= Support (A,B,C)/Support(A,B)

In other words, is the support for all three events a significant portion of the segment for which the antecedent events occurred, i.e. how often does C follow A and B? This ratio ranges from 0 to 1, with a 1 representing a 100% confidence.

Advanced Analytics Features from Prior Releases

The advanced analytics features in this section are not new in BIRT Analytics 4.2, but are included here for completeness.

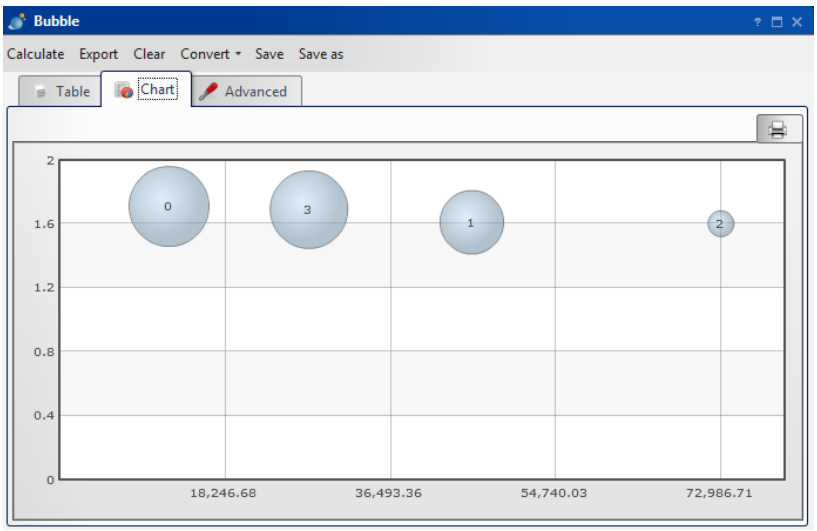
Clustering

BIRT Analytics uses a clustering technique called K-means. The algorithm finds clusters of data around mean values. The basic idea is to try to discover k clusters, such that the records within each cluster are similar to each other and distinct from records in other clusters. It is an iterative algorithm; an initial set of clusters is defined, and the clusters are repeatedly updated until no more improvement is possible (or the number of iterations exceeds a specified limit).

In BIRT Analytics, clustering is achieved via the following steps:

- Define the size of the test segment
- Define the number of clusters
- Provide the desired confidence level
- Provide the attributes around which to attempt the clustering
- Train the algorithm
- Apply the algorithm
- View the results

A popular choice for viewing the output of clustering is the bubble chart; this and several other visualizations are available in BIRT Anlytics 4.2.



Forecasting

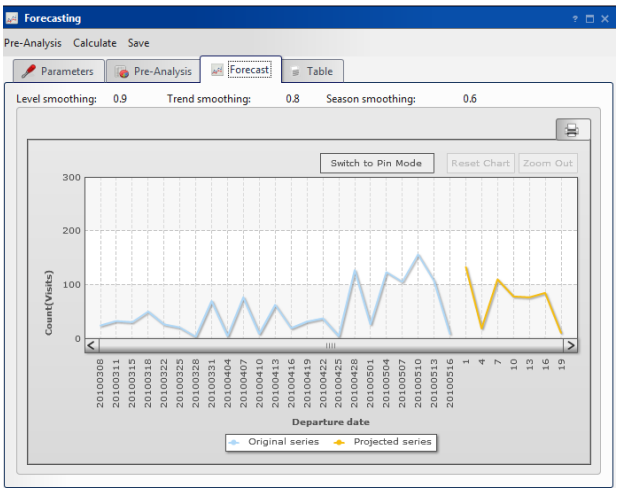
In BIRT Analytics, forecasting uses the Holt Winters technique. In the first phase of the analysis, the user provides a number of parameters:

- Identify measures
- Identify dimensions
- Define the number of projections
- Define the seasonality

In the next, pre-analysis phase, the user guides the analytic process by:

- Replacing outliers
- Evaluating seasonality

The forecast is then calculated and presented in a graphic or table.



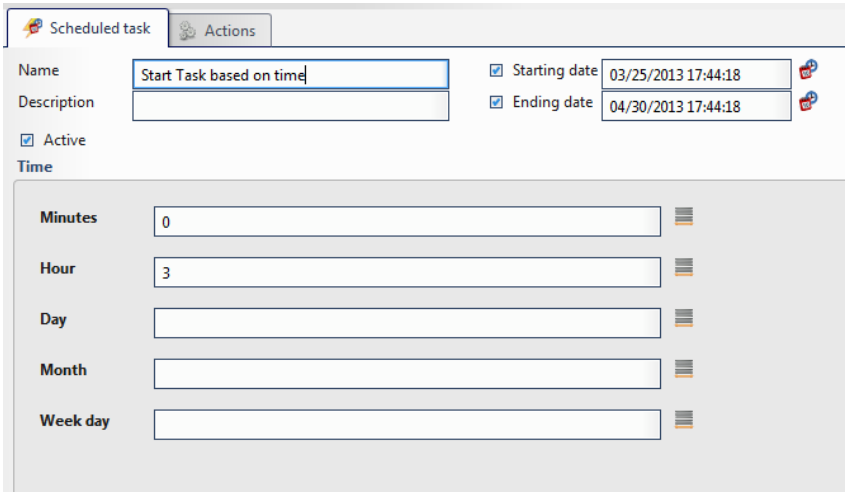
Several visualization options and zoom levels are available for viewing forecasts.

iWorkflow

iWorkflow allows users to create tasks that are scheduled for specific time periods or triggered by events in the application. An example of a task based on an event is the automatic generation of an e-mail alert when the application has reached a specified number of orders.

Scheduling of Time Based Tasks

Setting up time based events is simply a matter of specifying the month, week, day, hour, minute at which the event should be triggered.



Scheduling Event Based Tasks

Event based tasks fall into various categories:

BIRT Analytics Web

- Engineering – when structural changes are made to the data (ex: somebody creates an aggregate)
- Definition modified – when a user modifies any data definitions
- Repository item deleted – when users add, change, or delete items from the repository

Data Mining

- Model applied – when a user applies a model

Campaign Workflow

- Campaign executed
- Cell executed
- Change of stage

When a specified condition for is reached (scheduled or based on event), the user can specify one of many tasks or actions:

- Send an e-mail, which can be populated with the values of specified data fields within the text of the e-mail.
- Conditionally execute a query
- Execute
    - Execute campaign
    - Load responses
- Export campaign cell(s)
- Data model
    - Delete column
    - Delete table
- Apply a data mining model (forecast, clustering, decision tree, association rule)

Actions can also be specified as SQL queries.

Executions of the workflow are logged and accessible.

Campaign Workflow

Campaign Workflow enables users to create and manage multi-channel campaigns, with the capability of loading the contact history and inferred responses.

Understanding campaigns

A campaign is a set of tasks, defined for specific population segment and completed during a defined time period to accomplish a specific goal. For example, a typical business campaign defines a set of communication tasks that channel information to a segment of customers or prospects. The most typical campaign generates advertising messages to customers in a selected market segment. Common goals of an advertising campaign are web site visits and online purchase transactions made by customers.

BIRT Analytics supports the automation of campaign tasks associated directly with events or conditions that occur in your database. For example, a forecast analysis predicts a seasonal percentage increase in purchases by customers. Also, models based on association rules show what additional items a purchasing customer typically buys. By defining a campaign strategy that includes seasonal timing and targeted messaging, your company can effectively persuade a customer to buy an additional item, or upgrade to one having a higher profit margin. Further refining this idea, you can design specific messaging delivered to select market segments, based on data that you collect from that specific segment. For example, analyzing the profile of a customer who responds by purchasing one item enables your website to offer suggestions about similar products to other customers having a similar profile.

Examples of target segments, reached via several media, include:

Electoral voters

Financial services clients

Grocery store customers

Hospital patients

Configuring campaign elements

Campaign workflows are broken down into stages and cells. Campaign cells perform filtering operations or analytic calculations on the data. Cells are collected into stages, and stages can be triggered by an event (ex: completion of the previous stage, or an end-user action) or by thresholds and conditions in the data. Stages are often used for ‘gated’ execution with approvals for proceeding with the next stage, as required.

A campaign includes the following elements:

- Workflows that define campaign roles
- Permissions required to perform campaign tasks
- Stages that group tasks in a workflow
- Properties that define specific cells of activity
- Segments of data on which cells operate
- Media appropriate to communicate with the segment
- Resolution tables for history and response records
- Inclusion and exclusion lists to apply as filters
- Strategies that identify campaign goals

qLoader Enhancements

The usability of qLoader has been enhanced in a number of areas. The enhancements, sorted by the tabs in which they appear, are listed below:

Transformation Tab

- Copy/Paste and Cut/Paste functionality with transformations
- Drag and drop objects (table, column) from the data tree
- Get derived column definitions and insert in loading project
- Increase size of the right panel area with transformation instructions
- FIRST and LAST functions in Aggregate instruction
- Instruction VAR – using variables
- Instruction PRINT – Allow to insert comments on a loading project to be shown in a loading log
- Instruction IFEXISTS – IFExists function
- Instruction IFLinked – IFLinked function

Data Tab

- In the table declaration, possibility of choosing “None” qualifier
- After modifying a data source definition, users can apply the change and update it in existing control files
- New postreSQL driver

New Links Tab

- New tab for creating or removing existing links

Explorer Tab

- Column – view discrete values

New Launcher

- New qLauncher tool – run a loading project from a batch process

Connectivity

- New native PostgreSQL driver

FastDB Performance Improvements

A new crosstab calculation algorithm has been implemented that performs significantly better with crosstabs that have lots of possible combinations with values that are high sparse, i.e. relatively few cells in the cross product have non-null values.

Consider a 3 dimensional crosstab containing 150, 30K and 60K discrete values each. The cross product contains 270 billion cells, but in most cases only a few of these cells will have a non-null computed value (in a vast majority of cases the combination simply does not exist). The new crosstab calculation algorithm is able to process this type of crosstab much faster.

Additionally, crosstab caching has been improved so that if the user has calculated a crosstab with one measure, say COUNT(Customer), and then adds a new measure, say SUM(amount), not all the values in the crosstab are calculated again - only the new measure.

Upgrading to BIRT Analytics 4.2

Upgrading to BIRT Analytics 4.2 requires a backup of key files, uninstallation / reinstallation of the product, and a restoration of the backed up files. The user's data repositories can be preserved and will become accessible after the administrator has resynchronized the repository and reassigned permissions. All the data segments created and saved by users are preserved – these are kept in the SQL Server 'wpt' table, which is not erased during an uninstallation. The loader scripts are preserved as well – these are kept in the dubnium projects and qloader folders, which are not erased during an uninstallation.

For full instructions on performing upgrades, please see the BIRT Analytics 4.2 Installation Guide.