Clearwell 7.1.2 Feature Briefing

Transparent Predictive Coding

This document details the new Transparent Predictive Coding feature in Clearwell 7.1.2

If you have any feedback or questions about this document please email them to **IIG-TFE@symantec.com** stating the document title.

This document is provided for informational purposes only. All warranties relating to the information in this document, either express or implied, are disclaimed to the maximum extent allowed by law. The information in this document is subject to change without notice. Copyright © 2012 Symantec Corporation. All rights reserved. Symantec, the Symantec logo and NetBackup are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners

.

Feature Description

Symantec Clearwell eDiscovery Platform 7.1.2 introduces Transparent Predictive Coding as part of the Review module. This feature leverages machine learning technology in order to assist with review workflows. In a nutshell, an initial human review is done on a subset of documents. This subset is used to train the Symantec Clearwell eDiscovery Platform which items match the Tag against which the prediction will take place. Symantec Clearwell eDiscovery Platform then leverages its Active Learning technology to identify items to build additional training sets which need human review until a satisfactory accuracy level is achieved. Clearwell Transparent Predictive coding then uses the training it has received to predict against the complete data set to determine which documents are responsive to the case. Then, based on what it has learnt, and with human approval all the items can be automatically tagged by Clearwell.

Business Value

The primary business value of this feature is that its use may greatly decrease review time and cost, especially on cases where a high volume of data is involved.

Underlying Principles

Technology

Support Vector Machines

The Symantec Clearwell eDiscovery Platform's Transparent Predictive Coding leverages Support Vector Machines (SVM) technology, which is a supervised learning model with associated learning algorithms that analyze data and recognize patterns. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Yield-Based Sampling

Symantec's Yield-Based Sampling is a patent-pending sampling methodology that is tailored specifically for eDiscovery. Characteristics unique to the case are taken into account in creating a statistically sound Control Set for use in prediction accuracy testing. The test results generated by using the Control Set give full and defensible detail into the precision, recall, and overall accuracy of the predictions.

Active Learning Technology

Symantec's Active Learning Technology provides the ability for the Symantec Clearwell eDiscovery Platform to intelligently suggest new documents for review.

Prediction Ranks

When the system predicts on a document, it generates a Prediction Rank. This value is a percentage score that represents how likely the document is to be positive for a given tag. These Prediction Ranks are fully searchable, sortable, and filterable. See Figure 1

Figure 1 – Prediction ranks

Prediction Insight

In order to maintain transparency, the Symantec Clearwell eDiscovery Platform provides Prediction Insight for every prediction generated by the system. Given an item's Prediction Rank, example documents from training are presented that have key features in common. These features are drawn from the content and metadata of the documents and are made fully available for analysis. See Figure 2

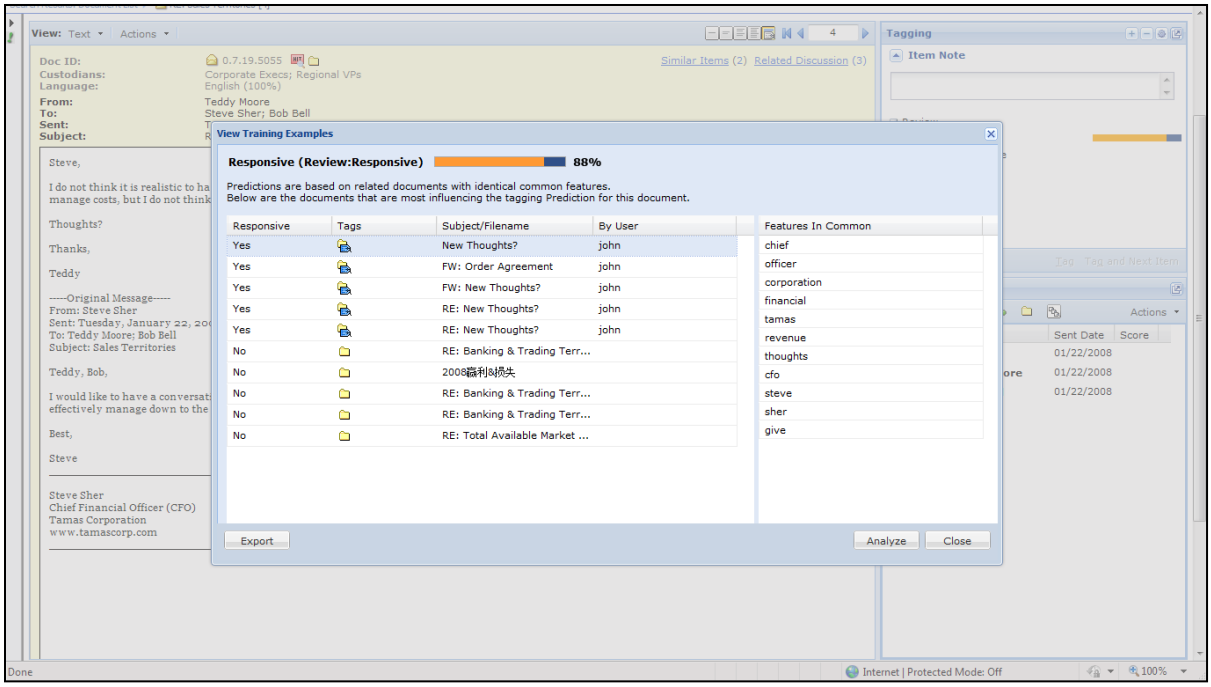

Figure 2 – Prediction insight

Defensibility Reporting

Several reports are available for the defensibility of using Transparent Predictive Coding. These reports include full audit reports of the steps taken during a review's Transparent Predictive Coding workflow, detailed prediction accuracy test reports that cover all statistical variables and metrics, and prediction insight reports for individual documents.

Transparent Predictive Coding Workflow

The use of Transparent Predictive Coding is done with an iterative workflow in order to achieve the highest possible prediction accuracy. It is important to note that the best Transparent Predictive Coding results are achieved when the document set has been culled down using Early Case Assessment and that the technology is suitable only for documents with fully extracted and indexed text.

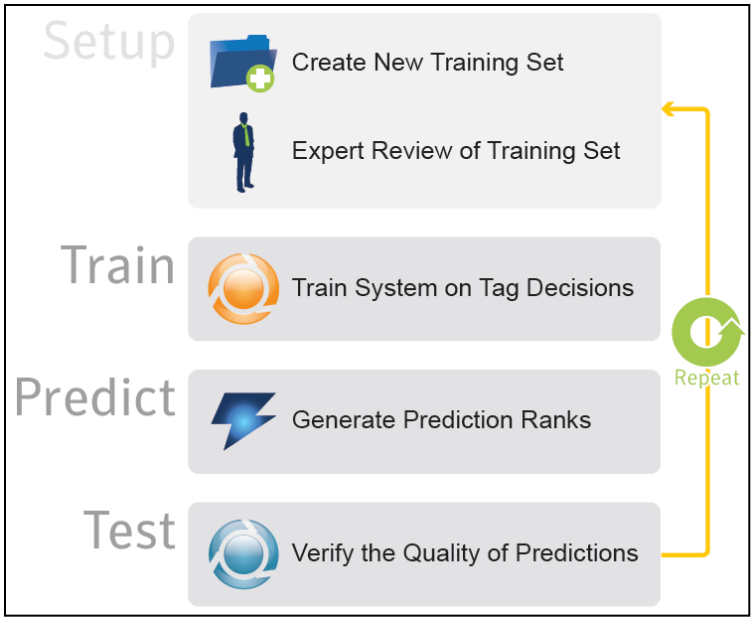

Figure 3 - Workflow

Test Drive

Setup

The setup of a Transparent Predictive Coding iteration begins with selecting a new Training Set of documents. For the first iteration, the Training Set should be selected judgmentally in order to locate a small number of documents that are likely to be positive. For the subsequent iterations, the Training Set may be selected by leveraging Symantec's Active Learning Technology. These documents require an extremely accurate human review as the tagging decisions will be leveraged throughout the workflow.

Train

Once the review is complete, the system may train on the Training Set documents. The system will associate the key features of each document with the documents' tag values in order to build a Prediction Model.

Predict

Given a Prediction Model, the system may generate Prediction Ranks for untrained documents. These Prediction Ranks are made available for full use and analysis.

Test

Using a Control Set of documents created using Symantec's Yield-Based Sampling, prediction accuracy tests may be run in order to assess the overall accuracy of the predictions. Given the test results, a new iteration may begin in order to increase the accuracy further, or the positively predicted documents may be taken out of the Transparent Predictive Coding workflow to continue review.

Workflow Flexibility

Transparent Predictive Coding is an extremely versatile tool that allows for complete flexibility in workflows. Alternate workflows may be achieved by leveraging the Prediction Rank values for review prioritization and intelligent batching. In addition, the transition between linear review workflows, Transparent Predictive Coding workflows, and workflows leveraging other forms of technology assisted review are seamless within the Symantec Clearwell eDiscovery Platform.

Licensing and support considerations

Transparent Predictive Coding is included in Clearwell eDiscovery Platform 7.1.2 Review module.

About Symantec:

Symantec is a global leader in providing storage, security and systems management solutions to help consumers and organizations secure and manage their information-driven world.

Our software and services protect against more risks at more points, more completely and efficiently, enabling confidence wherever information is used or stored.

For specific country offices and contact numbers, please visit our Web site: www.symantec.com

Symantec Corporation
World Headquarters
350 Ellis Street
Mountain View, CA 94043 USA
+1 (650) 527 8000
+1 (800) 721 3934

Copyright © 2012 Symantec Corporation. All rights reserved. Symantec and the Symantec logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.