Organization of Legal Professionals

Understanding Predictive Coding

An Overview and Interpretation of Terminology

By Herbert L. Roitblat, Ph.D.

Chief Scientist, Chief Technology Officer, Orcatec
•28 Years Research and Technology Experience
•Dolphin Search, University of Hawaii

Co-founder of OrcaTec LLC (CA) and the father of Concept Search. Before starting OrcaTec, Dr. Roitblat was Chief Scientist and a co-founder of DolphinSearch, as well as an award-winning Professor of psychology at the University of Hawaii. He has been granted four patents and has others pending on concept search technology. Dr. Roitblat is widely recognized as an expert in search and retrieval technology, particularly in the area of eDiscovery.

He received his BA Degree from Reed College in Portland, OR and his Ph.D. in Psychology from The University of California-Berkeley. In addition to his scientific work, Dr. Roitblat has been writing extensively about the problems of dealing with massive amounts of electronic data and the emerging standards for dealing with those problems. He has been a member of the Sedona Working Group on Electronic Document Retention and Production, on the Advisory Board of the Georgetown Legal Center Advanced eDiscovery Institute, and a member of the 2011-2013 Program Committees of the Georgetown Legal Center Advanced eDiscovery Institute.

He is a co-founder of the Electronic Discovery Institute and a member of the Board of Governors of the Organization of Legal Professionals. Dr. Roitblat is a frequent speaker on eDiscovery, particularly concerning search, categorization, predictive coding, and quality assurance.

ORCATEC

About the Author

Herbert L. Roitblat, Ph.D.

OLP ORGANIZATION OF LEGAL PROFESSIONALS

Mission

To promote the common business interest of the legal community, establish guidelines and standards for legal professionals on e-discovery and other topics, evaluate compliance with same, and serve as a resource on quality certification.

OLP is a professional organization establishing guidelines and standards for the legal field, with an emphasis on e-discovery. The organization:

- Evaluates compliance with the standards;
- Recognizes organizations and programs which demonstrate compliance;
- Serves as a resource on quality certification;
- Certifies educational providers that meet OLP's standards.

Vision

OLP is an administratively independent resource recognized as an authority on standards for professional certification of individuals and organizations providing services and products to the legal industry.

Based on sound principles, OLP standards are optimal and comprehensive criteria for organizational process and performance. They are broadly recognized, objective, and current benchmarks for certifying individuals to achieve and by which they operate.

Who Belongs to OLP?

Members of OLP include top level professionals: lawyers, judges, e-discovery managers, technicians, paralegals, case managers, litigation support managers, administrators, vendors, software developers, consultants, and others. OLP invites you to sign up today! Join your colleagues in an exciting journey towards setting higher standards in the legal community. Participate in making a difference. **To join go to:** **Membership**

About OLP

Organization of Legal Professionals

TABLE OF CONTENTS

> Predictive coding is not magic. It does not replace all of human review. It does not cure cancer. Predictive coding is mathematical algorithms and applied statistical analysis used to emulate the decisions that an authoritative expert would make, based on the evidence in the documents.
>
> -Herbert L. Roitblat, Ph.D.

Chapter One

Introduction

Predictive coding uses computers and machine learning to reduce the number of documents in large document sets to those that are relevant to the matter. It is a highly effective method for culling data sets to save time, money and effort. Predictive coding learns to categorize documents (for example, as responsive or non-responsive) based on a relatively small sample of example documents.

Predictive coding is not magic. It does not replace all of human review. It does not cure cancer. Predictive coding is mathematical algorithms and applied statistical analysis used to emulate the decisions that an authoritative expert would make, based on the evidence in the documents.

Predictive coding allows one person or a small group of people to effectively review millions of documents in a short period of time, with higher accuracy and consistency, and at a much lower cost than traditional review methods. In predictive coding, a computer is “trained” to distinguish between responsive and non-responsive documents. The system can then use the differences between these two sets of documents to infer how to categorize the remaining documents in the collection.

There are several ways that systems can get their training examples. These training documents are a sample of all of the documents in the collection. The examples can be selected randomly and categorized, can be provided by expert reviewers, chosen by the computer, or determined by some combination of these.

Predictive coding is a kind of Computer-Assisted Review (CAR) or Technology-Assisted Review (TAR), but it is not the only kind of CAR/TAR. Other types include keyword searching, concept searching, clustering, email threading, more-like-this search, and near duplicates. These other kinds of CAR can be very useful and can reduce the time needed to categorize documents, but they are not predictive coding – they do not predict on the basis of examples which documents are likely to be responsive versus non-responsive.

In predictive coding, the computer uses the decisions made by the expert reviewer(s) to predict how other documents should be categorized. In clustering or the various kinds of searching, the documents are organized into groups and, after the computer has done its work, the reviewers then decide whether each of these groups should be considered responsive or non-responsive. Predictive coding involves what is called in the jargon of machine learning "supervised learning," while the other approach, when it involves machine learning, is called "unsupervised learning." In predictive coding, the authoritative expert reviewer provides feedback or supervision to the predictive coding system.

> Predictive coding is a powerful tool in the arsenal of eDiscovery. When used correctly, it can sub-tantially reduce the volume of documents that must be considered for production or for evaluation of responsiveness.
>
> -Herbert L. Roitblat, Ph.D.

Chapter Two

9 Technologies and What They Contribute to Predictive Coding

1. **Latent Semantic Analysis**. A mathematical approach that seeks to summarize the meaning of words by looking at the documents that share those words. LSA builds up a mathematical model of how words are related to documents and lets users take advantage of these computed relations to categorize documents.

2. Probabilistic Latent Semantic Analysis. A second mathematical approach that seeks to summarize the meaning of words by looking at the documents that share those words. PLSA builds up a mathematical model of how words are related to documents and lets users take advantage of these computed relations to categorize documents.

3. Support Vector Machine. A mathematical approach that seeks to find a line that separates responsive from non-responsive documents so that, ideally, all of the responsive documents are on one side of the line and all of the non-responsive ones are on the other side.

4. Nearest Neighbor Classifier. A classification system that categorizes documents by finding an already classified example that is very similar (near) to the document being considered. It

There are several ways that systems can get their training examples. These training documents are a sample of all of the documents in the collection. The examples can be selected randomly and categorized, can be provided by expert reviewers, chosen by the computer, or determined by some combination of these. It gives the new document the same category as the most similar trained example.

5. Active Learning. An iterative process that presents for reviewer judgment those documents that are most likely to be misclassified. In conjunction with Support Vector Machines, it presents those documents that are closest to the current position of the separating line. The line is moved if any of the presented documents has been misclassified.

6. Language Modeling. A mathematical approach that seeks to summarize the meaning of words by looking at how they are used in the set of documents. Language modeling in predictive coding builds a model for word occurrence in the responsive and in the non-responsive documents and classifies documents according to the model that best accounts for the words in a document being considered.

7. Relevance Feedback. A computational model that adjusts the criteria for implicitly identifying responsive documents following feedback by a knowledgeable user as to which documents are relevant and which are not.

8. Linguistic Analysis. Linguists examine responsive and non-responsive documents to derive classification rules that maximize the correct classification of documents.

9. Naïve Bayesian Classifier. A system that examines the probability that each word in a new document came from the word distribution derived from trained responsive documents or from trained non-responsive documents. The system is naïve in the sense that it assumes that all words are independent of one another.

All of these approaches involve machine learning, except, typically, Linguistic Analysis (which may or may not include machine learning components). A computational process extracts pertinent information from example documents and builds a mathematical model that allows responsive and non-responsive documents to be distinguished from one another based on the text that they contain.

The accuracy of these systems will depend on the specifics of the implementation and on the quality of the training set used. They may also differ in the amount and type of training that must be conducted, including the level of effort. Other differences among these technologies are beyond the scope of the present paper.
In general, these systems work by extracting “features” from the example documents. Usually these features are words, though

they can be word combinations, or mathematical values related to groups of words. The computer learns which features are related to documents in each category, and which distinguish between the categories.

When a new document is presented for classification, the computer compares the features of that document with the features known to distinguish the categories and then assigns the new document to the appropriate category based on its features.

Chapter Three

5 Questions to Decide Whether a Matter is Appropriate for Predictive Coding

Most predictive coding systems require text. Predictive coding generally does not work on non-text documents such as blueprints, CAD drawings, photographs, videos, audio recordings, and so forth, unless they are converted first to text. If you have text documents, then there are five questions you can ask to help you decide whether a matter is appropriate for predictive coding.

1. Do you want to find as many of the responsive documents as possible?
2. Do you want to review as few of the non-responsive documents as possible?
3. Do you want to identify potentially responsive document as quickly as possible?
4. Do you want to minimize the cost of review?
5. Do you want to reduce the time needed to review documents?

If the answer to at least one of these five questions is yes, then there is one more question to consider.

- Does your collection contain more than about 5,000 text documents?

Predictive coding does not require a large set of documents, but it's value tends to grow disproportionately as the size of the document collection grows, because the effort typically required to train a system does not grow or does not grow as quickly as the size of the document collection increases. Small collections can require almost the same level of training effort as large collections do.

Chapter Four

8 Technological Issues in Using Predictive Coding

The following is an outline of a basic, effective predictive coding protocol. It addresses the technological issues involved in using predictive coding, while recognizing that there will also be legal strategic issues that must be considered. This protocol is only one of many that may be appropriate to a particular situation.

1. **Meet and Confer**. The parties meet to determine the parameters of eDiscovery, including preservation, collection, selected custodians, time ranges, topics, concepts, and other pertinent issues. Repeat as necessary as the case evolves. Although limiting the documents to be considered by date ranges and custodian makes some sense, it may not be advisable to try to limit the documents by keywords, because of the difficulty in guessing the right keywords.

2. **Exploratory Data Analysis.** The producing party, recognizing its obligation to produce responsive documents, begins document analysis. The technology does not require sharing training documents or seed sets with the receiving party. Sharing these documents assumes that the technology works as expected, but that the producing party requires “guidance” to identify the correct documents to be produced. There are many ways to provide this guidance without having to share non-responsive documents. Legal and strategic concerns should govern whether these documents should be shared, it is not an intrinsic part of the predictive coding process.

3. **Estimate Prevalence**. The producing party samples the document set to get an estimate of prevalence. How rare / frequent are responsive documents? Prevalence is important because special steps may be needed to make predictive coding training efficient if responsive documents are extremely rare (e.g., less than 1% of the documents are responsive). Prevalence sampling may be part of the process of training the predictive coding system.

4. **Predictive Coding Training**. The producing party begins predictive coding training. The producing party may report accuracy statistics along the way, or, if training is brief, at the end of training. Not all predictive coding tools yield meaningful statistics during the course of training. Some

require small enough amounts of training that reporting in the course of training may be too disruptive statistics during the course of training. Some require small enough amounts of training that reporting in the course of training may be too disruptive.

5. **Predictive Coding**. When predictive coding training is complete, the remaining documents in the collection are coded by the computer.

6. **Evaluation**. A sample of documents is reviewed by the producing party for responsiveness to measure the effectiveness of the predictive coding. There are several different ways to perform the sampling. The exact sampling method should be agreed to by the parties. Use the smallest sample necessary to achieve the desired confidence interval. Choose a confidence interval that is consequential.

A confidence interval of, say, plus or minus 5% is usually sufficient. Keep in mind that values in the center of the confidence interval are much more likely than values at the edges of the confidence interval.

There are several ways that an evaluation can be conducted following predictive coding.

a. After the documents have been categorized by the system, review can be continued on newly generated random samples of documents. That is, the same expert continues to evaluate random samples of documents until a sample size the parties agree is adequate has been obtained. The system's efficacy on this sample is taken as a measure of its performance.

b. A separate random sample of documents designated by the predictive coding system as non-responsive can be evaluated to compute the Elusion measure. Elusion is the proportion of documents classified as putatively non-responsive that should have been classified as responsive.

Ideally, only a small proportion of the documents in the putatively non-responsive set will be found to be responsive. In practice, the proportion of responsive documents in the putatively non-responsive set should be only a small fraction of the prevalence of responsive documents. Elusion, therefore, needs to be compared to the original estimate of responsive document prevalence. The size of this sample will depend on the required confidence level and confidence interval.

.

c. A set of putatively responsive and a set of putatively non-responsive documents could be evaluated. Ideally, all of the putatively responsive documents will, in fact, be found to be responsive and none of the putatively non-responsive documents will, in fact, be found to be responsive. In practice, most of the putatively responsive documents should be found to be responsive and few of the putatively non-responsive documents should be found to be responsive. This information can be combined with other available information to give an estimate of Precision and Recall.

7. **Privilege Review**. The documents designated responsive by the predictive coding system are reviewed by the producing party for privilege. The privileged documents in this set may be withheld, and the non-privileged ones produced.

8. **Dispute Resolution**. If there are disagreements about the produced documents that cannot be resolved by conferring, then a special master may be appointed to examine a sample of the documents and their computer-generated coding.

.

Conclusion

Predictive coding is a powerful tool in the arsenal of eDiscovery. When used correctly, it can substantially reduce the volume of documents that must be considered for production or for evaluation of responsiveness. Predictive coding is not a substitute for legal judgment, but an amplifier of it, bringing higher levels of consistency, efficacy, accuracy, and efficiency. For the producing party, it promises to return more focused documents more economically.

For the requesting party, it promises to return more complete and focused documents in a shorter period of time. In many cases, predictive coding provides an all-around win, moving litigation to the merits of the case, addressing Rule 1 of the Federal Rules of Civil Procedure "to secure the just, speedy, and inexpensive determination of every action and proceeding."

Glossary

Active learning – a form of supervised machine learning that presents for review or human categorization the documents with the highest current uncertainty, those documents that will be most informative about how to update the learning process.

Bayesian categorizer—an information retrieval tool that computes the probability that a document is a member of a category from the probability that each word is indicative of each category. These estimates are derived from example documents. Uses the probability of each word given each category to compute the probability of each category given each word. Also called a naïve Bayesian Categorizer.

CAR – Computer assisted review. Any of a number of technologies that use computers to facilitate the review of documents for discovery. See TAR.

Glossary

Collection – A group of documents. These can be documents gathered for a particular matter or purpose. Information retrieval scientists tend to use several well-known document collections (e.g., RCV1) for testing and comparison purposes.

Confidence interval – the expected range of results. If you drew repeated samples from the same population, you would expect the result to be within the confidence interval about the proportion of times given by the confidence level.

For example, in an election poll, the difference in the proportion of people favoring each candidate is described as being within a range of, say, plus or minus 5%. All other things being equal, the smaller the confidence interval, the larger the sample size needs to be. Said another way, the larger the sample size, the smaller the confidence interval.

Glossary

Confidence level –how often we would achieve a similar result if we repeated the same process many times. If we did the same kind of test from the same population more than once, the confidence level would tell us how often we would get a result that is within a certain range (the confidence interval) of the true value for the population.

Most scientific studies employ a minimum confidence level of 0.95, meaning that 95 percent of the time when you repeated the experiment you would find a similar result. The higher the confidence level the larger the sample size that is required.

Technically, it is the proportion of times when the true population value would be included in within the confidence interval.

Glossary

Contingency Table – a table of the four response states in a categorization task. The rows of the table may correspond to the correct or true category values and the columns may correspond to the choices made by system. For example, the top row may be the truly positive category (e.g. truly responsive documents) and the second row may be the truly negative category (e.g., truly non-responsive documents).

The columns then represent the positive decisions made by the system (e.g., putatively responsive) and the negative decisions made by the system (e.g., putatively non-responsive). The entries in these cells are the counts of documents corresponding to each response state (e.g., true positives, false negatives, false positives, true negatives). Contingency tables are often displayed along with the totals for each row and for each column. Sometimes the rows and columns are reversed, so the columns reflect the true values and the rows reflect the choices.

Glossary

Elusion – an information retrieval measure of the proportion of responsive documents that have been missed. Most often used as a quality assurance measure in which a sample of non-retrieved documents is evaluated to determine whether a review has met reasonable criteria for completeness.

Judgmental sampling – a sampling process where the objects are selected on the basis of some person's judgments about their relative importance rather than on a random basis. Judgmental sampling sometimes refers to the use of a seed set or preselected documents used to train predictive coding systems.

Unlike random samples, judgmental samples are not typically representative of the collection or population from which they are drawn. It is not possible to extrapolate from the characteristics of a judgmental sample to the characteristics of the population or collection.

Glossary

Language modeling—computing a model of the relationships among words in a collection. Language modeling is used in speech recognition to predict what the next word will be based on the pattern of preceding words. Language modeling is used in information retrieval and predictive coding to represent the meaning of words in the context of other words in a document or paragraph.

Latent Semantic Analysis—(LSA) a statistical method for finding the underlying dimensions of correlated terms. For example, words like law, lawyer, attorney, lawsuit, etc.

All share some meaning. The presence of any one of them in a document could be recognized as indicating something consistent about the topic of the document. Latent Semantic Analysis uses statistics to allow the system to exploit these correlations for concept searching and clustering.

Glossary

Latent Semantic Indexing—(LSI) the use of latent semantic analysis to index a collection of documents.

Machine learning—a branch of computer science that deals with designing computer programs to extract information from examples. For example, properties that distinguish between responsive and nonresponsive documents may be extracted from example documents in each category. The goal is to predict the correct category for future untagged examples based on the knowledge extracted from the previously classified examples. Example approaches include neural networks, support vector machines, Bayesian classifiers and others.

Nearest neighbor classification—a statistical procedure that classifies objects, such as documents, according to the most similar item that has already been assigned a category label. This approach uses a set of labeled examples to classify subsequent unlabeled items, by choosing the category assigned to the most similar labeled example (its nearest neighbor) or examples. K-nearest neighbor classification uses the k most similar classified objects to determine the classification of an unknown object.

Population – the universe of things about which we are trying to infer with our samples. For example, the population may be the set of documents that we want to classify as putatively responsive or putatively non-responsive. The group from which we pull our samples. Also called the sampling frame.

Precision – the proportion of retrieved documents that are responsive. See also recall.

Predictive coding – a group of machine learning technologies that predict which documents are and are not responsive based on the decisions applied by a subject matter expert to a small sample of documents.

Glossary

category assigned to the most similar labeled example (its nearest neighbor) or examples. K-nearest neighbor classification uses the k most similar classified objects to determine the classification of an unknown object.

Population – the universe of things about which we are trying to infer with our samples. For example, the population may be the set of documents that we want to classify as putatively responsive or putatively non-responsive. The group from which we pull our samples. Also called the sampling frame.

Precision – the proportion of retrieved documents that are responsive. See also recall.

Predictive coding – a group of machine learning technologies that predict which documents are and are not responsive based on the decisions applied by a subject matter expert to a small sample of documents.

Glossary

Random sampling—the statistical process of choosing objects randomly, meaning that each object has an equal chance of being selected. Random sampling can be used to train predictive coding systems and to evaluate their efficacy.
Recall –the proportion of responsive documents in the entire collection that have been retrieved.

Relevance feedback—a class of machine learning techniques where users indicate the relevance of items that have been retrieved for them and the machine learns thereby to improve the quality of its recommendations.

Richness – the proportion of responsive documents in a collection.

Sampling – the process of selecting a subset of items from a population and inferring from the characteristics of the sample what the characteristics of the population are likely to be.

Glossary

Often refers to a simple random sample, in which each item in the population has an equal chance of being selected in the sample.

Seed set – a collection of pre-categorized documents that is used as the initial training for a predictive coding system.

Support vector machine (SVM) – a machine-learning approach used for categorizing data. The goal of the SVM is to learn the boundaries that separate two or more classes of objects.

Given a set of already categorized training examples, an SVM training algorithm identifies the differences between the examples of each training category and can then apply similar criteria to distinguishing future examples.

TAR – Technology Assisted Review. Any of a number of technologies that use technology, usually computer technology, to facilitate the review of documents for discovery. See CAR.