Applying Predictive Sampling to Reduce Costs and Increase Quality for Document Review

By Tom Groom, Vice President Discovery Engineering, D4 LLC
tgroom@d4discovery.com

January 25, 2011

The application of Predictive Sampling can significantly reduce costs, dramatically reduce review time and increase quality for document review. Cheaper, faster and better—all three.

Computerized technology has greatly impacted the efficiencies of document review over the years. Keyword Searching was among the first techniques and while according to many studies this approach only ranges from 20% to 40% for responsiveness precision (100% precision would be all of the documents that truly are responsive), Keyword Search remains as the most common approach used today for reducing the number of documents to be reviewed. Keyword search was further advanced through ontologies which, in their simplest form, is an exhaustive keyword search where a set of "but not" terms are used to disambiguate over-inclusive keywords. For example, an ontology would state "include (keyword1, keyword2, keyword3, etc.) but not (excludeword1, excludeword2, excludeword3, etc.)". Ontologies can also include proximity limiters (i.e. Tom w/2 Groom) as well as incorporate conceptual relationships. Studies have shown ontologies can improve responsiveness precision to 65% to 90%. In the past, ontologies had to be developed by highly skilled linguists working with counsel who have intimate knowledge of the legal issues of the case. This was expensive and time consuming. Predictive Sampling changes all that.

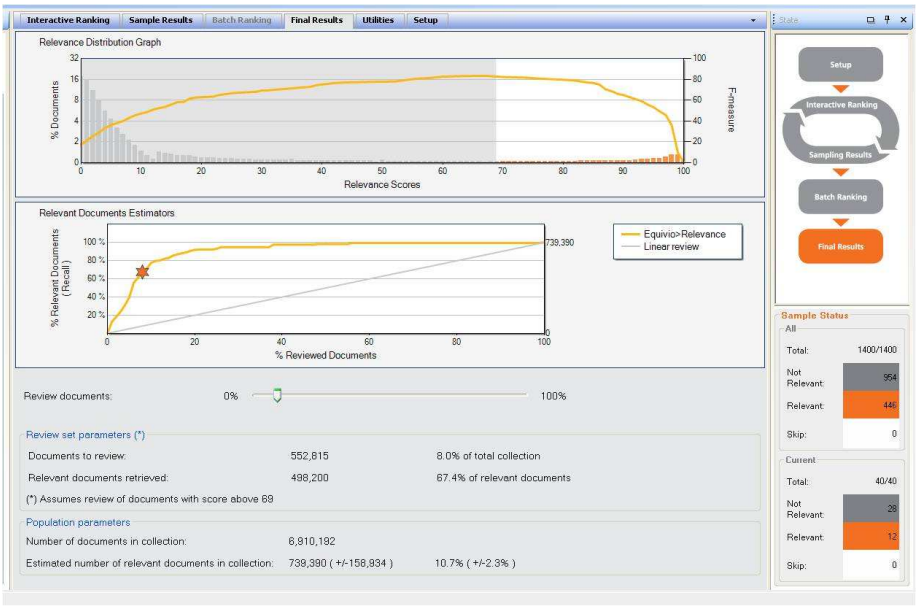
What is Predictive Sampling?

Predictive Sampling (also referred to as "Predictive Coding") combines the efficiencies of a computerized sampling system with a human "expert." The human interacts with the system by making "yes/no" calls to a question against a series of controlled samples of a set number of documents at a time. Questions can be "Is this document responsive?" or "Does it pertain to this specific issue?" or "Is this document privileged?", etc. The system builds an ontology in the background as it learns from the expert and presents subsequent samples. Normally after 25-40 iterations (1,000 to 1,500 documents), the system has sufficiently built the ontology to the point where it can "predict" what the human will choose as "affirmative" in the sample they are reviewing. Once it accurately predicts over a series of consecutive samples, the ontology is considered "statistically stable" and can be applied to the rest of the collection.

Predictive Sampling is built upon a well established modeling framework called Predictive Analytics which is a type of a Support Vector Machine. Predictive Analytics encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events. In business, predictive models exploit

patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions. Predictive Analytics is used in actuarial science, financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and now is being applied to document review via Predictive sampling. One of the most well-known applications is credit scoring, which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time. A well-known example would be the FICO Score.

The output of Predictive Sampling for document review is a "Relevance Score" based on a scale of 0 to 100. Once the system has been trained and the model is statically stable, the rest of the collection can be scored based on the underlying ontology. That score can then be used to identify non-responsive documents as well as prioritize review towards the documents with the highest scores. Some Predictive Sampling systems such as Equivio-Relevance provide interactive graphical tools to aid case management in determining the approach for specific relevance score zones. See example below.



Use Cases and Workflows:

Listed below are some sample use cases. There are many variations to these and the specific workflow depends on specific case variables but these should provide guidance for consideration.

1. **Early Case:** It is early in the case cycle and counsel hasn't started the development of list of keywords. You anticipate a high volume of ESI that will need to be collected in

order to find documents relevant to the case. Predictive sampling can be used on a set of ESI for a few key custodians which will result in:

a. finding key documents early in the process – may help determine if the case should settle or if there are sufficient facts to pursue
b. identifying the initial set of keywords that could then be used in negotiations with opposing and/or modified and used as a keyword filter to limit subsequent ESI collection.

2. **Accelerated Review:** Consider the relevance score distribution graph below. Using the relevance scores, a review team was able to divide the collection into three zones:

a. Zone 1 contains documents with relevance scores of 0 through 4. It contains 72% of the document population, but only yields less than 2% of the relevant documents. After a cursory review, the firm determined further review of the documents in this zone meets no criteria for reasonableness or proportionality of effort for a full initial review.
b. Zone 2 contains documents with relevance scores 4 through 37. It covers 3% of the population, but less than 1% of the relevant documents. The firm did some sampling, found very low yield of relevant documents – and decided that it was not worth the effort for a full initial review.
c. Zone 3 includes documents with relevance scores over 37. This zone covers 25% of the population, and contains 97.4% of the relevant documents. From the firm's point of view, this was a "must review" zone.

So, the firm was able to focus its initial review effort on just 25% of the population, and in so doing, they were able to identify and produce 97% of the relevant documents.

3. **Review QC/Verification:** When the review was complete, the firm turned to quality assurance. They set up a discrepancy matrix, comparing the relevance designations of the review team to the relevance scores regardless of zone as described above.

	Review/Responsive	Review/Not Responsive	
Equivio/Responsive	3,048	2,531	5,579
Equivio/NotResponsive	1,576	40,495	42,071
	4,624	43,026	47,650

Discrepancy Analysis:
- Equivio/Responsive *vs* Review/Not Responsive
- Oracle found almost 1,500 responsive documents missed by the review team

The graph above shows that there were 3,048 documents that the review team and Relevance agreed were responsive. Then there were 40,495 documents that the review team and Relevance agreed were not responsive.

Of particular interest to the firm were the 2,531 documents that the review team had marked as not responsive, but which Relevance scored as responsive. These documents represent potentially responsive documents that the reviewers may have missed. These 2,531 documents were submitted to a senior reviewer (so called "Oracle") for second pass review and verification. He found that almost 1,500 of these documents were in fact responsive. The responsive set increased from 4,624 to 6,000. That's an additional one-third on top of the original set that had been slated for production. The lead partner on the case saw this and his response was "My obligation is to make reasonable effort to discover the responsive documents. If I'm not using this technology, I'm not fulfilling my obligation and I am open to risk to a claim from opposing counsel that we have not disclosed all the relevant documents." This firm has standardized on Relevance largely because of these risk considerations.

There are many variations of these use cases and workflows but they generally fall into the three categories shown above.

Summary

Predictive Sampling goes beyond basic keyword searching. It is a powerful tool that uses well established predictive analytics and classification algorithms rather than just discrete keyword searches. Unlike keyword searches, Predictive Sampling takes into account all the words in a document as well as words to exclude, along with the relationship of the words to one another to determine what is and what is not likely to be relevant. Predictive Sampling incorporates human intelligence to leverage the results of review across large document populations. Predictive Sampling can be used in a variety of workflows in several places along the EDRM lifecycle including Early Case, Accelerated Document Review and Review QC/Verification.

About the Author

Tom Groom is a recognized e-Discovery expert with more than two decades of experience in information technology and 15 years focused on litigation support, document review, and e-Discovery methodologies. As the VP of Discovery Engineering at D4, Mr. Groom advises both corporate and law firm clients on issues involving ESI, including defensible collection methodologies, optimized ESI processing and review workflows, and litigation contingency and readiness planning. He establishes project teams and solutions tailored to meet the unique requirements of his clients. Mr. Groom has presented numerous MCLEs and training seminars on ESI topics including optimized ESI workflows with various approaches to Native File Review and production methodologies. He is also the point person for matters involving complex electronic discovery services, production coordination, high-volume imaging efforts, Web-based repositories, and complex coding projects. Mr. Groom has supported cases involving mergers & acquisition, antitrust, securities, environmental, patent infringement, products liability, as well as other general litigation. Mr. Groom's career began with IBM where he served for eight years as a System Engineer for the National Accounts division. The following seven years he served IKON Digital Litigation Services as Senior Business Development Manager. Mr. Groom then spent three years in the Geospatial software development and computer storage industries before joining Whitmont in 2004 which later became D4. He has received numerous awards for his innovative ideas and top performance. Mr. Groom earned a Master of Science degree in Industrial Engineering with a focus on computer information systems from Arizona State University and a Bachelors of Science, Industrial Engineering from the University of Arkansas.