Who Owns Predictive Coding?

Press Release Stretches the Truth of One Company's Claim

Herbert L. Roitblat, Ph.D.
CTO, Chief Scientist, OrcaTec

June 12, 2011

One of our competitors, Recommind, has recently been awarded a patent related to predictive coding. In a press release dated June 8, 2011, announcing this award, they make some very grandiose claims with no basis in fact. According to their press release, they actually claim to have patented predictive coding. This claim is a gross exaggeration and unsupported by the details of the patent (No. 7,933,859) or the history of predictive coding.

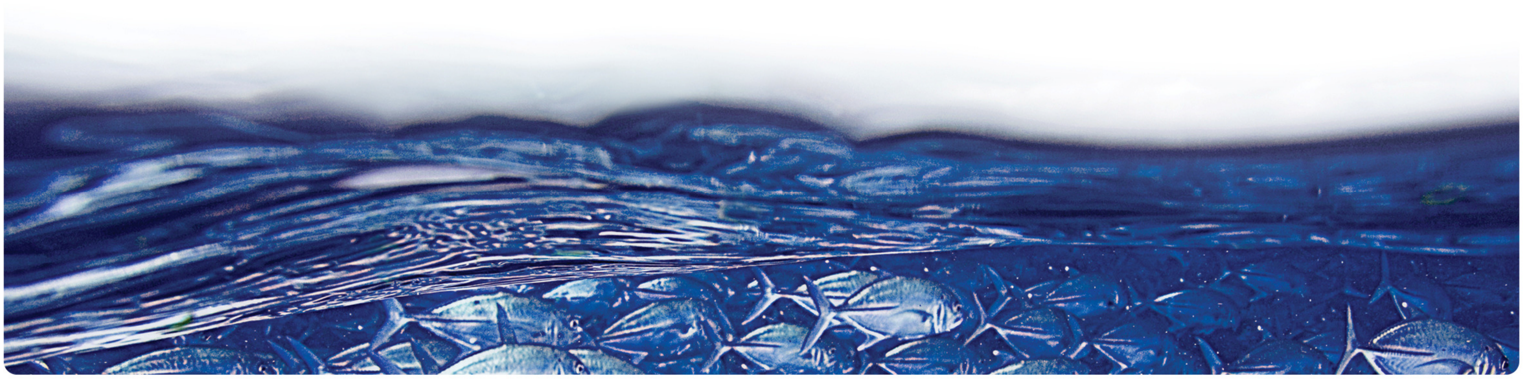
Having examined the patent carefully, I can say that this patent covers only a very narrow method of computing in predictive coding and is unlikely to have any impact on the ability of any other eDiscovery service provider to continue to offer this game-changing capability.

The scope of a patent is determined by its claims, not by the title of a press release. A valid patent requires that the proposed invention be (among other things) novel and non-obvious. In contrast, what we now call predictive coding has a very long history. I have written about some of this history elsewhere (Everything new is old again). Further, its use in eDiscovery predates Recommind's patent application by many years. Put simply, predictive coding itself is neither novel nor non-obvious, though some specific methods for implementing it may meet these requirements.

Predictive coding is a family of evidence-based document categorization technologies that are used to put documents or electronically stored information (ESI) into matter-relevant categories, such as responsive/nonresponsive or privileged/nonprivileged. The underlyingidea of using evidence to categorize objects has been around since the 18th Century. The notion of applying similar ideas to document classification or categorization was described in 1961 by M.E. Maron.

In his paper, Maron noted:

Loosely speaking, it appears that there are two parts to the problem of classifying. The first part concerns the selection of certain relevant aspects of an item as pieces of evidence. The second part of



the problem concerns the use of these pieces of evidence to predict the proper category to which the item in question belongs. (p. 404).

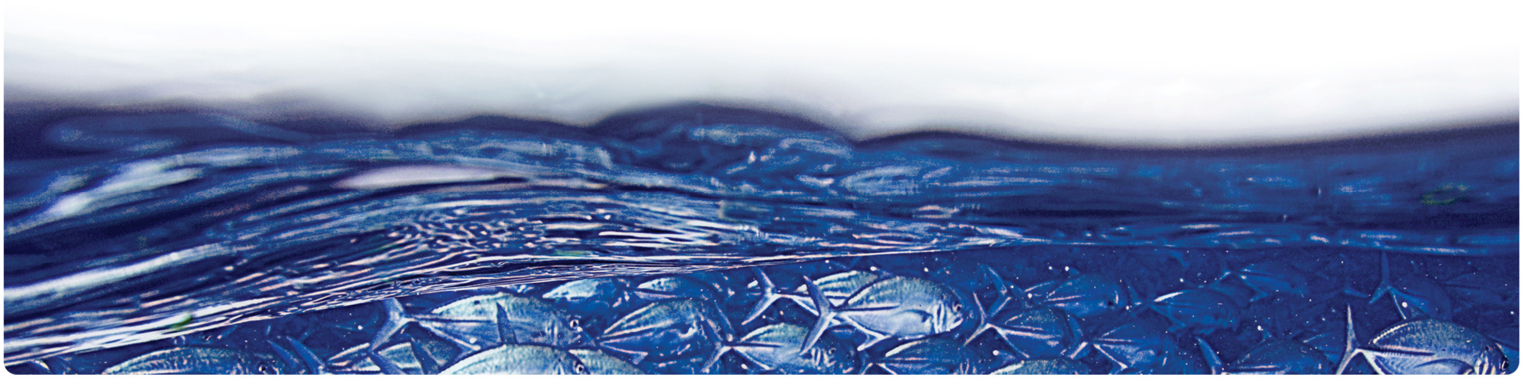
In 2002, Paul Graham wrote a paper on using techniques similar to those described by Maron to distinguish SPAM from nonSPAM (or HAM) emails. An initial sample of categorized SPAM and HAM emails is analyzed by the program to learn how specific words provide evidence for one category or the other. Subsequent emails are then classified according to these implicit rules and the evidence, which consists of the words in the emails. If the system misclassifies an email as either SPAM or HAM, the user can flag these errors and the classifier will update itself to reflect these reclassified emails. This seems to me to be a very clear example of predictive coding, which again argues that predictive coding, per se, does not meet the novelty or non-obviousness criteria.

eDiscovery service providers have been doing predictive coding (sometimes called by other names) for many years. In January of 2010, before the Recommind patent was filed, the eDiscovery Institute published a paper (by Roitblat, Kershaw and Oot) in the *Journal of the American Society for Information Science and Technology* describing two eDiscovery service providers and the accuracy of their predictive coding tools. These tools, obviously, were in existence prior to Recommind's filing. Other service providers have been providing similar services even before that. So Recommind can hardly be considered to have invented predictive coding, yet none of this prior art was actually included in Recommind's patent application.

One pending patent, however, was included in their application as evidence of prior art in this field. This pending application, assigned to Bank of America, is called "Predictive Coding of Documents in an Electronic Discovery System." Therefore, the patent application itself recognizes that Recommind cannot be the inventor of predictive coding.

The patent examiner also added a citation to Recommind's patent concerning the existing use of a statistical machine-learning technique called SVM (Support Vector Machines), which is used in their claimed invention.

Given all of this prior art, it is very clear that Recommind is in no position to claim to have either invented or to "own" predictive coding. Rather, their patent covers a very specific, very narrow approach to predictive coding involving the use of two very specific statistical procedures (one of which is SVM). I will leave it to attorneys to determine whether even this circumscribed application constitutes a valid patent.



Still, even if it is valid, it leaves plenty of room for other approaches to predictive coding, including the approach used by OrcaTec. Nothing in the Recommind patent would preclude OrcaTec or any other service provider from offering predictive coding services in eDiscovery or any other area. OrcaTec does not use SVM, nor the other techniques described in the Recommind patent. In fact, we believe that we have a much more advanced product, and can help attorneys achieve cost savings significantly beyond those claimed by Recommind.

Grandiose claims like those in the Recommind press release indicate either a profound lack of understanding of just what is covered by the patent, or are a deliberate attempt to obfuscate the issues in the industry. Attorneys involved in eDiscovery look to their service providers to provide open, honest and effective processes. They are not well served by unnecessary hyperbole.

About the author:

Herbert L. Roitblat, Ph.D. is the CTO and Chief Scientist of OrcaTec. He holds a number of patents in eDiscovery technology and other areas. He is widely considered to be an expert in eDiscovery methodology and technology. He is 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, a member of the 2011 program committee for the Georgetown Legal Center Advanced eDiscovery Institute, and the chair of the Electronic Discovery Institute. He is a member of the Board of Governors of the Organization of Legal Professionals. He is a frequent speaker on eDiscovery, particularly concerning search, categorization, predictive coding, and quality assurance.