

The New York Times

March 4, 2011

Armies of Expensive Lawyers, Replaced by Cheaper Software

By JOHN MARKOFF

When five television studios became entangled in a Justice Department antitrust lawsuit against CBS, the cost was immense. As part of the obscure task of “discovery” — providing documents relevant to a lawsuit — the studios examined six million documents at a cost of more than \$2.2 million, much of it to pay for a platoon of lawyers and paralegals who worked for months at high hourly rates.

But that was in 1978. Now, thanks to advances in artificial intelligence, “e-discovery” software can analyze documents in a fraction of the time for a fraction of the cost. In January, for example, [Blackstone Discovery](#) of Palo Alto, Calif., helped analyze 1.5 million documents for less than \$100,000.

Some programs go beyond just finding documents with relevant terms at computer speeds. They can extract relevant concepts — like documents relevant to social protest in the Middle East — even in the absence of specific terms, and deduce patterns of behavior that would have eluded lawyers examining millions of documents.

“From a legal staffing viewpoint, it means that a lot of people who used to be allocated to conduct document review are no longer able to be billed out,” said Bill Herr, who as a lawyer at a major chemical company used to muster auditoriums of lawyers to read documents for weeks on end. “People get bored, people get headaches. Computers don’t.”

Computers are getting better at mimicking human reasoning — as viewers of “Jeopardy!” found out when they saw Watson beat its human opponents — and they are claiming work once done by people in high-paying professions. The number of computer chip designers, for example, has largely stagnated because powerful software programs replace the work once done by legions of logic designers and draftsmen.

Software is also making its way into tasks that were the exclusive province of human decision makers, like loan and mortgage officers and tax accountants.

These new forms of automation have renewed the debate over the economic consequences of technological progress.

David H. Autor, an economics professor at the Massachusetts Institute of Technology, says the United States economy is being “hollowed out.” New jobs, he says, are coming at the bottom of the economic pyramid, jobs in the middle are being lost to automation and outsourcing, and now job growth at the top is slowing because of automation.

“There is no reason to think that technology creates unemployment,” Professor Autor said. “Over the long run we find things for people to do. The harder question is, does changing technology always lead to better jobs? The answer is no.”

Automation of higher-level jobs is accelerating because of progress in computer science and linguistics. Only recently have researchers been able to test and refine algorithms on vast data samples, including a huge trove of e-mail from the Enron Corporation.

“The economic impact will be huge,” said Tom Mitchell, chairman of the machine learning department at Carnegie Mellon University in Pittsburgh. “We’re at the beginning of a 10-year period where we’re going to transition from computers that can’t understand language to a point where computers can understand quite a bit about language.”

Nowhere are these advances clearer than in the legal world.

E-discovery technologies generally fall into two broad categories that can be described as “linguistic” and “sociological.”

The most basic linguistic approach uses specific search words to find and sort relevant documents. More advanced programs filter documents through a large web of word and phrase definitions. A user who types “dog” will also find documents that mention “man’s best friend” and even the notion of a “walk.”

The sociological approach adds an inferential layer of analysis, mimicking the deductive powers of a human Sherlock Holmes. Engineers and linguists at [Cataphora](#), an information-sifting company based in Silicon Valley, have their software mine documents for the activities and interactions of people — who did what when, and who talks to whom. The software seeks to visualize chains of events. It identifies discussions that might have taken place across e-mail, instant messages and telephone calls.

Then the computer pounces, so to speak, capturing “digital anomalies” that white-collar criminals often create in trying to hide their activities.

For example, it finds “call me” moments — those incidents when an employee decides to hide a particular action by having a private conversation. This usually involves switching media, perhaps from an e-mail conversation to instant messaging, telephone or even a face-to-face encounter.

“It doesn’t use keywords at all,” said Elizabeth Charnock, Cataphora’s founder. “But it’s a means of showing who leaked information, who’s influential in the organization or when a sensitive document like an S.E.C. filing is being edited an unusual number of times, or an unusual number of ways, by an unusual type or number of people.”

The Cataphora software can also recognize the sentiment in an e-mail message — whether a person is positive or negative, or what the company calls “loud talking” — unusual emphasis that might give hints that a document is about a stressful situation. The software can also detect subtle changes in the style of an e-mail communication.

A shift in an author’s e-mail style, from breezy to unusually formal, can raise a red flag about illegal activity.

“You tend to split a lot fewer infinitives when you think the F.B.I. might be reading your mail,” said Steve Roberts, Cataphora’s chief technology officer.

Another e-discovery company in Silicon Valley, [Clearwell](#), has developed software that analyzes documents to find concepts rather than specific keywords, shortening the time required to locate relevant material in litigation.

Last year, Clearwell software was used by the law firm DLA Piper to search through a half-million documents under a court-imposed deadline of one week. Clearwell’s software analyzed and sorted 570,000 documents (each document can be many pages) in two days. The law firm used just one more day to identify 3,070 documents that were relevant to the court-ordered discovery motion.

Clearwell’s software uses language analysis and a visual way of representing general concepts found in documents to make it possible for a single lawyer to do work that might have once required hundreds.

“The catch here is information overload,” said Aaref A. Hilaly, Clearwell’s chief executive. “How do you zoom in to just the specific set of documents or facts that are relevant to the specific question? It’s not about search; it’s about sifting, and that’s what e-discovery software enables.”

For Neil Fraser, a lawyer at Milberg, a law firm based in New York, the Cataphora software provides a way to better understand the internal workings of corporations he sues, particularly when the real decision makers may be hidden from view.

He says the software allows him to find the ex-Pfc. Wintergreens in an organization — a reference to a lowly character in the novel “Catch-22” who wielded great power because he distributed mail to generals and was able to withhold it or dispatch it as he saw fit.

Such tools owe a debt to an unlikely, though appropriate, source: the electronic mail database known as the [Enron Corpus](#).

In October 2003, Andrew McCallum, a computer scientist at the University of Massachusetts, Amherst, read that the federal government had a collection of more than five million messages from the prosecution of Enron.

He bought a copy of the database for \$10,000 and made it freely available to academic and corporate researchers. Since then, it has become the foundation of a wealth of new science — and its value has endured, since privacy constraints usually keep large collections of e-mail out of reach. “It’s made a massive difference in the research community,” Dr. McCallum said.

The Enron Corpus has led to a better understanding of how language is used and how social networks function, and it has improved efforts to uncover social groups based on e-mail communication.

Now artificial intelligence software has taken a seat at the negotiating table.

Two months ago, [Autonomy](#), an e-discovery company based in Britain, worked with defense lawyers in a lawsuit brought against a large oil and gas company. The plaintiffs showed up during a pretrial negotiation with a list of words intended to be used to help select documents for use in the lawsuit.

“The plaintiffs asked for 500 keywords to search on,” said Mike Sullivan, chief executive of Autonomy Protect, the company’s e-discovery division.

In response, he said, the defense lawyers used those words to analyze their own documents during the negotiations, and those results helped them bargain more effectively, Mr. Sullivan said.

Some specialists acknowledge that the technology has limits. “The documents that the process kicks out still have to be read by someone,” said Herbert L. Roitblat of OrcaTec, a consulting firm in Atlanta.

Quantifying the employment impact of these new technologies is difficult. Mike Lynch, the founder of Autonomy, is convinced that “legal is a sector that will likely employ fewer, not more, people in the U.S. in the future.” He estimated that the shift from manual document discovery to e-discovery would lead to a manpower reduction in which one lawyer would suffice for work that once required 500 and that the newest generation of software, which can detect duplicates and find clusters of important documents on a particular topic, could cut the head count by another 50 percent.

The computers seem to be good at their new jobs. Mr. Herr, the former chemical company lawyer, used e-discovery software to reanalyze work his company’s lawyers did in the 1980s and ’90s. His human colleagues had been only 60 percent accurate, he found.

“Think about how much money had been spent to be slightly better than a coin toss,” he said.