www.infoconomy.com

INFOCONOMIST

Business insight for Europe's technology elite

March 2001

IN THIS ISSUE

Bayesian haze

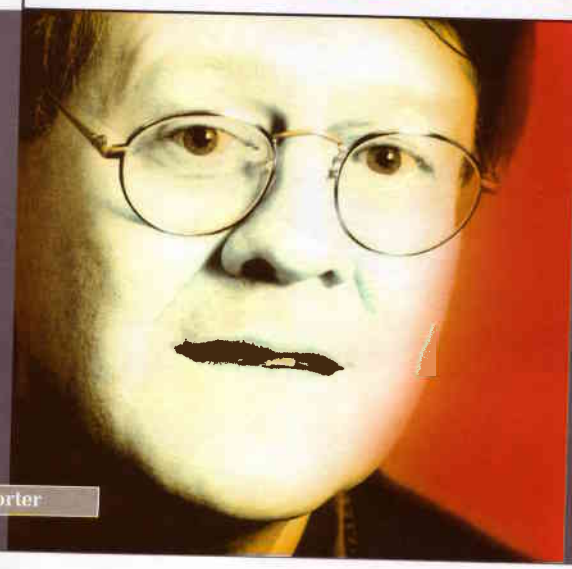
Demystifying Autonomy's search engine revolution

PUBLISHED BY

infoconomy

WRITTEN BY ROB BUCKLEY

THE BAYESIAN BRAVES



Martin Porter

Knowledge management, information retrieval: call it what you will, but it's an area in which Europe – and particularly the UK – has been doing very well. While US search engine companies such as Excalibur may have fallen by the wayside, and others such as Verity have under performed, UK academia has spawned a plethora of home grown companies.

The star of the new searchers is Autonomy, one of the offshoots of late 1980's Cambridge-based research into 'probabilistic' search methods, with a one-time market capitalisation of £5 billion.

But does this emperor have no clothes? Is probabilistic searching as good a technology as both IT and financial analysts think it is – or are those companies that employ it set for a fall?

Autonomy, the Cambridge, UK-based search engine company, is a software superstar – highly valued and admired on both sides of the Atlantic. Its founder, Mike Lynch, is fêted for his wit and vision as much in Silicon Valley as in Europe. But the market is jumpy about the company, as if investors can't quite believe how well it is doing; and Lynch, in turn, seems jumpy about the market.

In February 2001, for example, all it took was a leaked note from Merrill Lynch that said Microsoft was going to launch a competing product to Autonomy – using technology based on work by "Steve Robinson", a supposed former Autonomy employee, and the company's share price dropped from £19 to £17. This was despite some analysts' suggestions that it should be nearer £40 – and despite inaccuracies in the note that Merrill Lynch later had to retract.

Autonomy, in fact, has been so successful – in growth, in profile, and in its valuation – that many in the search engine business cannot conceal their astonishment, if not their outright jealousy. This is particularly true of those Cambridge companies that use similar technology. Executives at SmartLogik, for example, frequently complain that the market capitalisation of its parent company, BrightStation, is too low compared to Autonomy's.

Is the company's technology that similar? Autonomy certainly argues not. But both Autonomy's and SmartLogik's systems draw on the same Cambridge University research into 'Bayesian probabilistic inference'. Research that was principally conducted by Professor Martin Porter – (now a consultant to SmartLogik), Karen Sparck Jones (still working for Cambridge University), and one Stephen Robertson – the real name of the person who now works in Microsoft's Cambridge research labs, and who gave Merrill Lynch and the market such jitters. He has, however, never worked for Autonomy.

The basic principle of Bayesian probability theory differs from normal probability theory because it does not automatically assume that the chance of something being true is independent of events that went before it. For example, under normal probability theory, a coin that comes up heads 100 times in a row would still have a 50/50 chance of landing tails on the next throw, whereas Bayesian theory would weight the chances in favour of a head, given the previous 100 results.

By feeding a system such as Autonomy's a set of documents that relate to a topic, and a set that do not, it can use Bayesian probability to work out which word-groups in a given document mean the text is more likely to be about that subject. It can also go on to work out what are the synonyms for particular terms – all without any knowledge of the language in which it is searching.

Although most of the theoretical research into Bayesian probability was carried out by Robertson, it was Porter who developed a working system for Cambridge University's geology department. "They had a huge collection of objects with a lot of data to collect and analyse," says Porter. "It was a classic

information retrieval problem." After working at it for several years, he eventually managed to build a system called Muscat, which the University said he could keep when he left, provided it could continue to use it.

"Between 1984 and 1990, I carried on developing it and put together a demo. But by 1990, I'd run out of opportunities to market it," says Porter. However, a chance meeting with John Snyder, who had been doing similar work, led to the two founding a company called Muscat to sell the software. And with Cambridge being the small place it is, they set up shop in the offices next door to those of a firm called Cambridge Neurodynamics, run by Mike Lynch, future founder and CEO of Autonomy.

"I was at college with a guy called Mike Lynch," Snyder notes wryly. "He and I used to go around pitching Muscat. And I put Muscat on his machines. He even had the source code because we were porting it to some Unix platforms. Mike was going to sell Muscat and give us a royalty of 15%. But eventually he went off and developed his own technology based largely on my ideas and the like." That, says Snyder, is how Autonomy was born. Porter recalls the initial friendliness of the two companies. "At one time, there was a lot of crossover… we even used to go to their Christmas parties."

Autonomy persevered and developed its technology, but Muscat was bought out in 1997 by MAID, the UK-based online information service, which shortly after became Dialog. "A few months later," says Snyder, Dialog bought information services gaint Knight Ridder. "It had a huge amount of debt and so there was no money to invest in Muscat. We were frozen in the wilderness." But Muscat did not die. After a firesale, the technology assets of Dialog became BrightStation, and it maintained its London stock market listing. BrightStation subsequently placed Muscat in the care of its SmartLogik division, currently being groomed for a spin-off IPO.

THE BAYESIAN WARS

The atmosphere between SmartLogik and Autonomy is now icy. "I've dealt a lot with analysts and some journalists," says SmartLogik CEO Stephen Hill. "They all say certain other companies are extraordinarily arrogant and rather difficult to do business with. I'd hate to be labelled as that."

In spite of their common roots, Autonomy and SmartLogik's technologies have diverged over the years, which means there is plenty of scope for these companies – and others joining the fray – to get involved in the Holy War over which search technology is best. The issue is not so much speed, or volume of documents retrieved, but the one that will achieve the most relevant pages for any given search.

Among those in the opposite side of the court to Autonomy is Verity, a former market leader and the biggest proponent of standard 'keyword' searches versus Bayesian 'pattern recognition'. "The problem with Bayesian is it's automated," says UK CEO Simon Atkinson. "You can't control what it does."

Autonomy spokesman Simon Fletcher counters that Verity's claims are not very meaningful. "Why would you want to control the system? When you do that, you get in all sorts of trouble. It's not economical, having a team of 50 tagging up documents."

But John Western, a technical consultant with Verity, argues that there are other problems too: "Bayesian has difficulties with fine-grain distinctions, particularly as you add more documents to the system. Every new document you add depends on what's happened to the previous documents. If you order your work differently, you'll get different results. That implies you have to do more work [to organise the database] and therefore the system doesn't scale as well." Categorisation of documents by humans or by rules is a better system, believes Verity, because the intelligent agents used by Autonomy, *et al* cannot understand the documents – "only we can".

Fletcher, of course, fundamentally disagrees. "How many times do you want a specific document, when you really want something that gives you information?" he counters. "The only reason we're fixated by the idea of the exact result is because of years of conditioning by keyword search systems." On this issue, fellow Bayesian John Snyder, who now runs an Internet search company called Webtop, sides with his rival at Autonomy. The keyword approach means that people "spend about 10 minutes a day searching, according to our research. But 75% of searches don't deliver useful information. People have learnt that if they type in 10 keywords, they get nothing, but if they type in one keyword, they get lots. It's not really helpful. Probabilistic information retrieval, particularly with relevance feedback from the user, gets better results, as proved by the TREC tests (*see box*)."

Another issue, he adds, is that when a new concept that cannot be included in current categorisations crops up, Bayesian methods can still cope while keyword methods have to wait until the system is updated. Autonomy's own research suggests that the time wasted by employees on searches amounts to nearly an hour a day, or £17 billion a year.

There is a third way of approaching the problem: semantic information retrieval – trying to break down queries into parts of speech in an attempt to understand the question. "For example, 'murder of a child', 'murder by a child' and 'murder with a child' are three very different legal concepts,

TREC TESTS SHOW THAT MOST SEARCH ENGINES ARE BROADLY SIMILAR

The Text Retrieval Conference (TREC) is an annual beauty pageant for search engines. It was started in 1992 with the goal of "speeding the transfer of technology from research labs into commercial products".

For each TREC, contestants are given a standard set of documents and questions. Participants run their retrieval systems on the data, and return a list of the top-ranked documents to a panel of Judges who evaluate the results.

Karen Sparck Jones, one of the original Cambridge team that developed the probabilistic approach to searches, has analysed the trends in the TREC test results over the years. Disappointingly for vendors wishing to claim their systems are the best, Sparck Jones says that many teams obtain similar performance even at top levels. "While there's been some convegence on default strategies, especially in automatic searching, similar performance is also obtained with very different strategies."

Sparck Jones also notes that "manual query formation can give superior performance to automatic." She points out that this typically reflects the amount of effort put in to the queries and the users' judgements on intermediate outputs. But while manual searching can do better than automatic searching in some tests, "the time and attention required is nevertheless not negligible".

yet Boolean, keyword and even probabilistic searches probably won't be able to pick up the difference. So far, research has had limited results, but a key exponent, 3F, another UK company, was acquired in March 2000 by Mindmaker, a Californian company that specialises in speech technology and in intelligent assistant products.

"Academically, you always want a pure solution, but a business solution needs to be timely, accurate and fast. Semantic analysis is important, but it's expensive in processing terms," says Snyder. Smartlogik CTO John Challis believes that semantic technology, while it works well in some niche areas, is not ready for prime time yet. "It has problems when working with documents it hasn't come across before." And, of course, with documents in a different language, it has as much difficulty understanding them as a human would. Bayesian, in contrast, "is language independent and automatically recognises terms it hasn't seen before."

BAYESIANS MULTIPLY

While academic interest in the semantic approach continues, the Bayesian group is also continuing to expand. Cambridge based NCorp, another spin-off from Lynch's Cambridge Neurodynamics, specialises in searching heavily structured databases using Bayesian techniques. Another company, Applied Psychological Research, set up by a group of London's City University academics in 1998, uses Bayesian techniques to build up profiles of users by seeing how they rate documents. "We focus heavily on the error in any search and try to reduce the consequences of that error as much as possible by finding out as much as possible on the individual making the search," says CEO Daniel Brown.

Smartlogik, in fact, pours scorn on Autonomy's classification system. "If you speak to Mike Lynch," says Challis, "he'll tell you that if you've got a really good search engine, you don't need classification. The reason he says so is because his categorisation isn't very good. It's based on stored queries, which isn't a smart way of doing classifications." What companies really need, says Challis, "is a rules-based architecture to capture the vocabulary that makes a document and to then attach weights to that vocabulary."

Fletcher says that the idea that Autonomy doesn't do categorisation is nonsense. "Not only is initial categorisation mandatory, but the system can automatically create sub-categories on the fly." Mostly, however, Autonomy's strategy is to try to avoid such arguments. It is, after all, the incumbent leader, is growing fast, and is executing well. But while that strategy may work against SmartLogik, Autonomy will find itself forced to prove its worth on all kinds of fronts if Microsoft does enter the market. That is why Merrill Lynch's note had such an impact.

David Johnson, of investment bankers Beeson Gregory, says arguments over the technology, however, are pointless. "Software companies' products are seldom as good as they claim to be. So I'd put Autonomy in there with every other software vendor on the planet. Their technology's pretty good, but it's not the only way to skin that cat. I'm sure other companies will develop better technology, but it'll probably never see the light of day."

Johnson goes on to say. "Autonomy has a good product range and a very, very aggressive sales and marketing team. There is high brand recognition and it's perceived to be the market leader. It's difficult to see something the size of Autonomy [it has more than 400 customers and 40 OEM partners] going horribly wrong. That is why it is on a meaty rating."

Even Snyder agrees with that. "People buy a product based on the quality of the vendor, not on how brilliant 'the mousetrap' is. It's all down to marketing." ■

CONTACTS: **rbuckley@infoconomy.com**