

Editorial

Something to do with probability

Stephen Robertson

Cambridge, United Kingdom

STEPHENEROBERTSON@HOTMAIL.CO.UK

I'm very pleased to write an editorial for IRRJ. It's great to see it into its second year of publication. It's a worthy successor to the *Information Retrieval Journal* that Paul Kantor and I started many years ago.

In the history of information retrieval, probabilistic models have a role to play, and in my research career, I have made use of them. The story I'm going to tell here has probability as one of its themes. That statement might make it sound like an abstract theoretical treatise, and I shall start that way, but actually its core is a very human story about the life and death of a colourful but flawed character who made a fortune as an entrepreneur, came near to losing it all, seemed to overcome potential disaster against the odds in 2024, and then in the same year died the most ridiculously unlikely death. I had a brief but interesting encounter with him in 2001.

Bayes

But let's start with the abstract bit. You may know about Bayesian statistics, but in case it's not familiar to you, here is a brief recap. Thomas Bayes was an eighteenth century Presbyterian minister, philosopher and statistician. In that last capacity, he developed a theorem in basic probability theory, a very useful theorem relating the probability of A given B to the probability of B given A .

A long time after Bayes, in the 20th century, a group of statisticians used Bayes' theorem as the basis for a rethinking of basic statistical ideas – the term Bayesian was not used until the 1950s. In the view of these statisticians (although there is more than one flavour), the probability of some event is not in general something that must be determined by a lot of experiments, like tossing a coin many times, but is a matter of something like a “degree of belief” – which should be modified in particular ways as evidence becomes available. Bayes' theorem is the basis for this modification. A “Bayesian” statistician is one who espouses this kind of philosophy of probability and statistics, in contrast to a “frequentist” statistician – this is commonly regarded as a major philosophical divide. But additionally, there has been a lot of very successful work developing Bayesian ideas in practical ways, to useful ends.

Probabilistic IR

When I began working on a probabilistic model for IR in the early 1970s, the main precursor was the 1960 paper by Maron and Kuhns (1960). In this paper, they proposed the idea, and discussed possible methods, whereby a numerical probability could be assigned to the event of a user (having presented a query to the system) would find a particular document useful. In my work with Karen Spärck Jones, we used a similar argument: we proposed starting from some evidence about the probability that a relevant document would contain any particular word (if you like, the probability of a term given relevance). We then used Bayes' theorem to invert this, to estimate the probability of relevance given the presence of the term, and indeed to combine evidence across multiple query terms. The result was something like the Maron/Kuhns probability of relevance. I say 'something like' because there are some complications to our respective interpretations of "probability of relevance", but they need not concern us here.

Just for the record, this model (RSJ) was published exactly fifty years ago (Robertson and Spärck Jones, 1976), and was the precursor of BM25. I would not describe this model as "Bayesian" in any strong sense, despite the use of Bayes' theorem. Maron and Kuhns actually made rather more of the Bayesian aspect than we did.

After RSJ

After RSJ, in the late seventies, I had a joint project with Keith van Rijsbergen, based in Cambridge where Keith worked (I was at City University in London). Part of the project was to work on Harter's two-Poisson model of term frequency distributions (Harter, 1975). We did not have a lot of success with that model, though later it was to feed into BM25. But we employed as a researcher and software developer Martin Porter; he then designed and wrote the famous Porter stemmer (Porter, 1980), certainly the best known outcome of that project. After the project finished, Martin took the search software that he had developed, and formed a company with John Snyder, marketing the software as Muscat.

In the late 80s, I joined forces with Steve Walker, whose Okapi system was already running; it included relevance feedback with RSJ weighting. In the early 90s we took part in TREC, and under that competitive stimulus, we developed BM25.

I joined Microsoft Research Cambridge in 1998. The MSR Cambridge lab was run by Roger Needham who had previously directed the Cambridge University Computer Lab, effectively the University's computer science department. Also at the Computer Lab (but never employed at MSR) was Karen Spärck Jones, who was married to Roger, and with whom I had been working for over 20 years. I built a small team at MSR, and some of my work (particularly BM25) was incorporated into Microsoft's Sharepoint system, which was their product for corporate data management.

Mike Lynch

Now I must introduce to you the character I mentioned above: his name was Mike Lynch. (For those of you with long memories, there was another Mike Lynch active in the IR field – he worked at the University of Sheffield Information School from the 1960s on, and died at the age of 92, also in 2024 – but this is not him.) The Mike Lynch who concerns us here is the subject of an excellent 2025 biography by Katie Prescott, *The Curious Case of Mike Lynch* (Prescott, 2025). Apart from my own encounter with him in 2001, most of what follows is better and much more fully described, discussed and analysed in Katie’s book.

Lynch was born in 1965, and got his PhD from Cambridge Engineering Department in 1990, in signal processing / pattern recognition. He is best known for founding Autonomy, in Cambridge in 1996, but before that he had a company called Cambridge Neurodynamics, which produced various kinds of software and hardware, including fingerprint recognition devices and character recognition. The company office was in the St. John’s Innovation Centre, a haven for startups in Cambridge. Nextdoor was the Muscat office, with Martin Porter and John Snyder. Snyder and Lynch already knew each other.

Neither Lynch’s research background nor the then work of Cambridge Neurodynamics included anything to do with text search or information retrieval; the Muscat software really interested him. He and Snyder started discussing a joint venture. However, Lynch decided to go his own way, and broke off with Snyder in 1994 – but not before Snyder had installed Muscat on one of the Cambridge Neurodynamics machines in order to demonstrate it to Lynch.

Then Lynch and Cambridge Neurodynamics developed their own text search and analysis software called the Dynamic Reasoning Engine, and in 1996 spun off a new company, Autonomy, to market the DRE.

Lynch was in the habit of making strong claims about the theoretical basis for the Autonomy software: he said, on every possible occasion, that it was based on Bayesian ideas. The whole of Autonomy’s publicity, advertising and marketing, was aimed at convincing people that their Bayesian approach was utterly special and unique, a mantra, some kind of magical incantation. He stressed specifically that it was not simply a text search engine, and was not based on simple keywords, but that it could infer what documents were “really about”.

Magic or no, over the next few years Autonomy became a very successful company, with (apparently) many corporate customers. It was floated on the stock exchange in 1998.

A fracas

I did not know Lynch or any of his associates, and I was barely aware of the company. But in early 2001, an analyst at the investment management company Merrill Lynch (another unrelated Lynch!) produced a report which suggested Autonomy stock was overvalued. Among other things, the report claimed that Microsoft was about to muscle in on Autonomy’s market, and that Microsoft in Cambridge had poached important members of Lynch’s company. It named – or rather misnamed – two people supposed to be working at MSR:

there was a slightly mangled version of Karen’s name, and also one “Steve Robinson” (they could only have meant me). Autonomy’s stock price promptly plummeted.

Internally, MSR Cambridge went into overdrive. Did they have *any* former Autonomy employees? (answer: no). Externally, it was mainly Mike Lynch who came out fighting – he laid a formal complaint against Merrill Lynch, which eventually forced them to withdraw parts of the report. Karen for her part spent a lot of effort trying, but in the end failing, to persuade Merrill Lynch to fund a studentship at the University Computer Lab as some kind of compensation.

Did I, or my group, or MSR Cambridge ever use any idea or method of Lynch’s? I don’t think so – Lynch never published his methods, and I never had occasion even to use Autonomy software. Did Lynch ever use any ideas of mine, or indeed from the wider IR community? Very probably: my ideas up to the time I joined Microsoft were all both published and in the public domain, and of course there was an entire world of published IR research, including both TREC and SIGIR. He was absolutely free to use them, though he never would have admitted to doing so.

But probably more important than the public status of these ideas was the fact that Lynch had access to the Muscat software. Muscat was a cutting-edge system, with some of my ideas, but Martin Porter built it with full knowledge and understanding of that world. Just for example, the Porter stemmer itself was highly innovative, but Martin knew and understood its predecessor, the Lovins stemmer, which relied much more on a dictionary of suffixes. Whether or not Lynch used the Muscat code itself, he could see what it did, how it worked, and what it was capable of. It’s hard to imagine that the DRE was not influenced by this knowledge.

What happened next

So much for my own encounter with Mike Lynch. Now, very briefly, the rest of his story.

Autonomy survived the Merrill Lynch incident, and carried on successfully, until in 2011 the company was sold to HP. Within a year, HP came to the conclusion that they had been conned – that the information they had been provided with about the client base and the commercial success of Autonomy was seriously inaccurate. What followed was a series of lawsuits against Lynch and some colleagues, both in the US and the UK.

A major civil lawsuit brought by HP in the UK went against Lynch, and his estate may yet be bankrupted as a result. But in 2024, in a criminal case in a US federal court in San Francisco, he and a colleague were acquitted on all charges (the former chief financial officer of Autonomy had earlier been convicted and imprisoned on related charges). After his acquittal, Lynch went on a celebratory cruise in the Mediterranean, on a yacht which he named *The Bayesian* – a huge yacht with a 72m mast (somewhat taller than the 15-storey building in which I live). In the course of this cruise, while moored at night off the coast of Sicily, the *Bayesian* was caught in a freak storm, capsized, and sank rapidly. Mike Lynch and several others were drowned. In a startling coincidence, the other colleague who had

been acquitted in the San Francisco trial was hit in a road traffic accident two days before the capsizing, and died the day after.

The ramifications of the Autonomy debacle continue, even after Lynch's death. At the time of writing, a UK court has ruled on the damages that Lynch's estate must pay to HP, but that might yet go to appeal. Also the builder of the *Bayesian* has sued the company which owned it (in turn owned by Lynch's widow) for damages, alleging that the sinking was the fault of the crew and that nobody is buying their yachts any more as a result of the disaster.

In the end, his magic deserted him.

References

- M.E. Maron and J.L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7, 216-244, 1960.
- S.E. Robertson and K. Spärck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, 129-146, 1976.
- S.P. Harter. A probabilistic approach to automatic keyword indexing (parts 1 and 2). *Journal of the American Society for Information Science*, 26, 197-206 and 280-289, 1975.
- M.F. Porter. An algorithm for suffix stripping. *Program*, 14, 130-137, 1980.
- K. Prescott. The Curious Case of Mike Lynch: The Improbable Life & Death of a Tech Billionaire. *Macmillan Business*, 2025.