ISSN 2048-7746
eNews
CONNECTING THE KNOWLEDGE COMMUNITY

Machine learning and academic publishing – where are we today?

UKSG eNews 405

22 Sept 2017

Michael Upshall

Hardly a day passes without a new story appearing in the media about machine learning and AI initiatives. It would seem there is no aspect of our daily activity that is not impacted in one way or another by these new technologies. So this is a good moment to assess how far scholarly publishing has made use of machine learning.

Let's look at the academic publishing landscape. Everyone is familiar with the steady increase in the number of articles being published. An assessment over 50 years ago (Derek J de Solla Price, *Little science, big science . . . and beyond*) estimated an annual growth rate in journal articles of around 4.7%, and there is no indication that the growth is declining – and that's not taking content types such as conference proceedings into account. The proliferation of academic publishing means that for most subject areas no human can keep track of the available material without assistance. To give one recent example, the 2017 ASCO (American Society for Clinical Oncology) Annual Meeting had over 2,500 papers presented. How can a researcher decide which of those papers should be read? Machine learning provides one very powerful answer.

Although artificial intelligence can trace its origins back to fifty years ago or more (Alan Turing proposed a learning machine in 1950), the recent spectacular growth in machine learning is the result of two main factors.

First, the shift in discovery and linking tools from rule-based to true machine learning, where the system trains itself. Rules have been the basis for typical information retrieval tasks in the last 50 years, but rules have limitations. For example, the classic case of teaching a system to differentiate between images of dogs and images of cats can be 'learnt' by the system being shown many thousands of images of cats and dogs (the training set), so that when the system is shown a new picture, it has enough information to predict if a new image is of a dog or a cat. Such a system eliminates the need to create a rule – which is just as well, since it would be challenging indeed to define the core characteristics of each species accurately enough to distinguish all cases. Machine-learning systems today use neural networks, by which examples provide feedback and hence enable progressive improvement of the algorithm. No task-specific programming is required, and the system continues to learn.

Second, and almost as important, is the growth in computing power. Vast-scale computation to identify the relevance of millions of words to a corpus is now an everyday occurrence. This is facilitated using the resources of cloud computing, by which thousands of virtual servers can be spun up to complete a single indexing operation.

Machine learning itself developed in a slightly roundabout way, starting as an offshoot of data mining in the 1990s before becoming the discipline of text mining, and then becoming better known as text analytics. Although tools for text such as stemming and lemmatisation had existed before 2000, the development of predictive tools for data mining and their application to text meant that text analytics started to grow in importance. Many of the underlying tools are language independent, although some of the best solutions available today include some additional language-based features to refine the results. Moreover, this probabilistic software is domain-independent – it works as well in history and philosophy as in chemistry and life science, which means that there is no need for a prior taxonomy or ontology in that area. The system trains itself.

So how should the scholarly publishing community respond? Given that machine learning is now a widely proven technology, its use for text analytics in scholarly publishing can and should be more widespread.

One concern can be resolved immediately: this technology does not replace humans. Certainly, if applied effectively, it eliminates much of the drudgery of repetitive tasks that require little human intelligence, and instead enables humans to concentrate on tasks that require real brain power. Michael Henry of KPMG, at a recent MarkLogic User conference, described how in financial compliance humans are frequently employed to 'stare and compare' in front of vast quantities of text, to identify potential cases of money laundering, whereas a machine can be trained to identify unusual and questionable activity much more reliably – and the machine can then point out such cases to a human evaluator to use his or her judgement. Such a technique enables credit card agencies to combat fraud. The machine spots unusual transactions, perhaps from another country or for an unusually high value, and alerts a human operator to assess the situation and to investigate further. Unlike humans, the machine will never fall asleep or have a bad day.

One implementation of text analytics in scholarly activity is to provide researchers with just this kind of signposting. One researcher calculated he was scanning 20 papers each week, with each paper taking him five minutes, simply to determine if the article was relevant to his research – before reading it in detail. He saw a clear benefit in using a tool that identified concepts of relevance to him and alerted him to the relevant articles, even if it had only 80% accuracy (precision), since he had now reduced the need for manually scanning the article by four fifths, as long as the system could alert him to the articles that still required human evaluation. In other words, machine learning is already being used to reduce the 'stare and compare' aspect of much of the academic user journey.

The question remains, since text analytics has in effect come of age, and is by now a proven technology: why has it not been more widely adopted? One factor is undoubtedly human nature. People tend to keep using outdated and expensive manual processes because they are familiar, and perhaps there is a feeling that carrying out a process by hand is inherently more precise that using a machine. Yet publishers and those responsible for content should be evaluating this new technology and identifying where it can add value. Here is a further challenge, since a new technology requires new ways to evaluate it. It is well known that two experienced indexers will rarely agree more than 80% of the time when tagging or indexing a document (Manning, Raghavan and Schütze, *Introduction to Information Retrieval*). So a machine-based solution with an accuracy of 85% will be more effective than human indexing; although many publishers continue to focus on the limitations of the machine tool rather than the overall improvement.

As with any new technology, machine learning changes the way we look at things. We require different tools to evaluate it, and a different way of looking at our own processes to see where it can be employed. To date, the impact of machine learning in academic publishing has been patchy. But the benefits are clear, for editors, for authors, for institutional repository owners and for publishers – but most of all for researchers.

This UKSG Editorial is taken from the industry newsletter *UKSG eNews,* published every two weeks exclusively for UKSG members. The newsletter provides up-to-the-minute news of current issues and developments within the global knowledge community.

To enjoy *UKSG eNews* and other member benefits become a UKSG member.