A Perfect Storm for Pessimism: Converging Technologies, Cost and Standardization

Topics: Information Retrieval, Data Mining, Machine Learning (Supervised and Unsupervised), eDiscovery, Information Governance, Information Management, Risk Assessment, Computational Intelligence
Author: Cody Bennett
Company: TCDI - http://www.tcdi.com

If some vendors' proclamations are to be believed, realizations of next generation self-learning technologies geared towards text retrieval, language understanding and real time risk assessments are being fulfilled. But the knowledge experts in charge of assisting these platforms need to be aware of exiguous claims. If standardization is going to occur on a matrix of such complex systems, they need to occur *on reality, not hype*.

The amount of information will grow vastly while storage costs become subdued increasing the need for computational technologies to offset the very large costs associated with knowledge workers. This paradigm shift signals a mandatory call for smarter information systems, both automated and semi-automated. But imperfections in technology systems (arguably lower than human mistakes) require critical focus on workflows, modularity and auditing. And although linear systems have improved efficiencies through the use of virtualization, they still do not approach a lateral learning mechanism[1]. While they have begun to break into multi-tenant super-computing capacity, software on these systems is still statistical and rules-based, hardly approaching anything "thinking" – encompassing decades-old algorithms stigmatized by "Artificial Intelligence".

Further, the cost prohibitive business model reliant upon a single technology becomes a landmine. Technology in the space of Information Retrieval and Machine Learning are moving targets, and formal standardizations may be quickly outmoded. While there are attempts to use previous and ongoing research alongside industrial search studies performed to classify and understand the limitations of each search model[2], use of hybridization and the underlying platforms / architectures facilitating multiple types of search techniques should be a target for which Information Management systems strive. For eDiscovery, vendors should be prepared to harness multiple search capabilities as courtrooms over time mold what is accepted as "standard". Focusing on a single methodology when coupled with automated systems hampers recall – IBM's Watson and Intelligence organizations prove that hybridized multimodal search and brute force NLP based directed probabilistic query expansion are interesting because of combinations in Information Retrieval, Data Mining and Machine Learning. How do you standardize upon the algorithms entrenched in systems that are constantly in flux? Do only systems with little or no entropy deserve standardization?

Use of multimodal search is becoming fashionably effective in tandem with automation. Machine Learning methods utilizing hybrid approaches to maximize historically divergent search paradigms are capable of producing multiple viewpoints based on different algorithms, maximizing return on implementations such as predictive coding, active "tripwire" systems and next-generation risk assessment. In eDiscovery, multiple modeling viewpoints can help augment linguistic spread of features necessary to defensibly identify varying degrees of responsiveness. An example would be the improvement for the eDiscovery process using active learning when conducting initial discovery and query expansion / extrapolation in the beginning phases of Request for Production[3].

With both Information Retrieval and Machine Learning, transparency in the methods and a heavy breakdown of the algorithms used will be required. This transparency assists Information Governance, defensible methods for legal, and quality assurance for knowledge workers. This prognostication may be similar to the inevitability of eDiscovery certification in bar exams. While it may not be necessary for legal to understand the full complexities of the underlying search technology or automated algorithm, it should be required to ascertain and request certifiable tests meeting standardized thresholds on retrieval software and learning systems especially in comparison with human counterparts. These standards not only directly affect industry and researches in academia, but legal teams who may view such technology as foreign. Legal in the realm of Information Governance will become the centrality for delivering the dos and don'ts of information in the corporation, in partnership with the CIO / IT, and possibly as oversight.

More robust search algorithms and sophistication in automated apparatuses allow more document discovery to be performed. While it could be argued by legacy eDiscovery review shops that such systems displace

1 See Lateral learning and the differentiators of Artificial and Computational Intelligence
2 NIST TREC, Precision, Recall, F-measure, ROC
3 http://trec.nist.gov/pubs/trec19/papers/bennett.cody.LEGAL.rev.pdf

workers, the resulting outcome will be more time for their expertise to focus on larger data sets and cases. The technology tools also allow new forms of discovery. During litigation, if both counsels are using automated methods, expect different forms of data mining and statistical modeling to look for fringe data; Information Governance becomes critically important because signposts to documents that were not produced may become evident. It also puts the onus on the automated systems. Though, even while precision, speed and capacity may massively increase, the chance of sanctions should increase less dynamically dependant upon the unknowns of the output. In review, knowing that automated coding will always make the same calls if the parameters and data remain the same may be comforting. But the hive instinct of a group of humans making judgments on the fly is tempered when replaced by the efficiency. Are vendors willing to champion their products against comparisons of human reviewers in standardized sessions? Are they willing to "open up the hood" for transparency?

Along with the many previous buzzwords, possibly the biggest is "Cloud". Information Management, Cloud and semi / automated eDiscovery provide historically high potential for low cost, immediate, real-time view into the information cycle. Which means, not only will businesses entertain cloud services, but because of lower cost, less worry about infrastructure, and touted uptime, they will be able to search and store more information as they adhere to rules for retention and preservation. Whether a public or private cloud or some hybrid, this growth of searchable data will necessitate further automation of processes in Information Governance and solidification of the underlying framework – policies, procedures and standards beyond search of information.

The standardization for Clouds may be best lead by the Government and related agencies. Cost of Government is under heavy scrutiny and current endeavors are occurring to facilitate the movement of Government into the Cloud. Cloud infrastructure, believing the hype, will structurally allow the computing capacity needed for today's brute force systems and experimental Computational Intelligence *et al*[4]. This intriguing ability to perform massive calculations per second with elasticity is a lowly feature compared to the perceived cost savings which currently drives the interest for mid to large sized entities; public clouds like Microsoft, Amazon and Salesforce.com currently among the most popular. Although, for eDiscovery, the cost of demanding and actually acquiring documents from geographically disparate locations may produce a haven for sanctions. More ominously, if mission critical systems become cloud based, could critical infrastructure (industry, state, and government) become even more exposed[5]?

This architecture triangulation (Cloud + [Enterprise] Information Retrieval + Machine Learning) is either a Nirvāṇa or the Perfect Storm. Whatever viewpoint, the criticality is security. Providing a one stop shop for data leaks and loss, hack attacks, whistle blowing and thievery across geographically massive data sets of multitudes of business verticals combined with hybridized, highly effective automated systems designed to quickly gather precise information with very little input at the lowest possible cost is one CIO's wish and one Information Manager's nightmare[6]. Next generation systems will need to work hand in hand with sophisticated intrusion detection, new demands for data security and regulators across state and international boundaries – and hope for costs' sake, that's enough. Standardized security for different types of clouds was bluntly an afterthought to cost savings.

Finally, technology growth and acceptance while cyclic is probably more spiral[7]; it takes multiple iterations to conquer very complicated issues and for such iterations to stabilize. Standardizing Artificial, Computational and Hybrid Intelligence Systems is no different. The processes underneath these umbrella terms will require multiple standardization iterations to flesh out the bleeding edge into leading edge. It is possible that the entropy of such systems is so high that standardization is just not feasible. Where standardization *can* occur in the triangular contexts described above, expect it to follow similar structure as RFCs from the Internet Engineering Task Force[8]. Though, this will likely require heavy concessions and the potential unwillingness from industry on interoperability and transparency.

---

4 Pattern analysis, Neural Nets

5 A next generation Stuxnet, for example…

6 Not to mention, lawyers holding their breath in the background…

7 This type of cyclical information gain when graphed appears similar to Fibonacci (2D) and Lorentz (3D) spirals.

8 This makes sense due to the fact that data access and search has been spinning wildly into the foray of Internet dependence.