Adaptable Search Standards for Optimal Search Solutions

Amanda Jones
Xerox Litigation Services
Amanda.Jones@xls.xerox.com

ICAIL 2011 Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings – DESI IV Workshop – June 6, 2011, University of Pittsburgh, PA, USA

The goal of this year's DESI IV workshop is to explore setting standards for search in e-discovery. Xerox Litigation Services (XLS) strongly supports the effort to establish a clear consensus regarding essential attributes for any "quality process" in search or automated document classification. We believe in the principles of iterative development, statistical sampling and performance measurement, and the utilization of interdisciplinary teams to craft sound information retrieval strategies. These will strengthen virtually any search process. Still, XLS also recognizes that there is no single approach to search in e-discovery that will optimally address the needs and challenges of every case. Consequently, there cannot be a single set of quantitative performance measurements or prescribed search protocols that can reasonably be applied in every case. Instead, we agree with the authors of "Evaluation of Information Retrieval" (Oard et al. 2011) that the discussion of standards for search should concentrate on articulating adaptable principles, clear and concrete enough to guide e-discovery practitioners in designing search solutions that are well-motivated, thoroughly documented and appropriately quality-controlled, with the flexibility to allow creative workflows tailored to the goals and circumstances of each matter.

Because of the unique and complex challenges ever-present in search in e-discovery, XLS would contend that the key to designing successful search strategies is the ability to explore multiple perspectives and experiment with a variety of tactics. Countless factors influence the quality of automated search outcomes. Therefore, it will be vital to the advancement of search techniques to adopt standards that encourage research on the sources of variability in search performance and create the latitude that is needed for ongoing hypothesis-testing and midstream course correction.

One source of variability in text-based search performance that XLS has already identified and addressed is data type. Relevance is manifested in markedly different linguistic patterns across various types of documents. So, XLS has elected to utilize distinct classification models for spreadsheet data, email data, and other text-based data for most projects. Developing and implementing distinct models for these three classes of data requires an additional investment of time and resources, but has consistently translated into significant performance gains for the population as a whole. So, it is the approach that we currently use to mitigate this source of performance variation and ensure the highest possible quality in our automated search results. Our research into this is continuing, though, and we are open to adopting a new equally effective and less labor-intensive tactic for managing linguistic variation across-data types.

Both within and outside Xerox, research in machine learning, information retrieval, and statistical data-mining is progressing rapidly. Thus, it is important to not only to devise creative solutions to known sources of variation in search performance, but also to have the freedom to explore the full potential of emerging automated search technologies. XLS is currently experimenting with ways to optimize search results by utilizing multiple techniques and technologies simultaneously, incorporating input from all sources that enhance the final results. In our observations, combining search tactics often leads to significantly higher performance metrics than can be achieved by any of the individual tactics alone. In one preliminary investigation across several matters, for example, we found that combining scores from one statistical algorithm applied to the metadata of a population with scores from a completely different statistical algorithm applied to the full text of the population consistently increased both precision and recall.

Similarly, we have also found it constructive to treat certain responsive topics or data types within a project with one search technique while using alternative approaches for other topics or data sources. For example, by analyzing patterns of error generated by our statistical algorithms, it has been possible for us to identify opportunities to use highly targeted linguistic models to correct those errors in the final result set. In general, our experimentation with hybridized search strategies has proven extremely fruitful and there are many avenues of investigation left to pursue in this area. This is a major motivating factor behind XLS's support of standards that would promote the novel application of any combination of available search resources, provided the efficacy of these applications were adequately demonstrated.

Obtaining a better understanding of the limitations of various search techniques is just as important as exploring the potential of new search technologies, because the limitations will also engender adaptive search strategies. Any text-based automated classification system will be subject to certain dependencies and limitations. For example, achieving comprehensive coverage with a high degree of accuracy is often challenging for search systems that rely on linguistic patterns to identify responsive material when responsive documents are "rare events" in the data population – primarily because there are simply fewer examples of the language of interest available to generalize. So, each and every responsive document is more noticeably impactful in the final results and performance metrics. In a case like this, more data is generally needed to achieve high precision and recall. It is sometimes possible, though, to mitigate the need for additional data utilizing linguistic and/or statistical approaches to increase the density of responsive material in a subset of the data population thereby increasing access to responsive linguistic material for generalization. Even then, though, it may require significant extra effort and ingenuity to ensure accurate and comprehensive coverage of the topic.

Further, the rate of responsiveness in a population interacts in a complex way with the definition of the responsive topic itself to influence the level of difficulty that can be anticipated in the development of a successful search strategy and the extent to which special tactics will need to be pursued. While it is not often discussed in great detail, it is extremely important to consider the subject matter target for a case when assessing options for search strategy. The way in which responsiveness is articulated in requests for production can have a profound impact on search efficacy. For example, all of the following subject matter attributes will play a role in shaping the inherent level of difficulty in using automated search techniques to evaluate a population for a given topic:

- Degree of subjectivity – e.g., a request for production may specify that all "*high level* marketing strategy" documents should be produced, but an automated search approach will likely struggle to differentiate between documents that constitute "high level" discussions and those that represent "routine" marketing conversations
- Conditions on modality – e.g., a request for production may specify that all "*non-public* discussions of pricing" should be produced, but linguistic distinctions between private and public conversations often prove unreliable causing automated approaches to confuse pricing discussions between corporate employees with similar discussions appearing in the media, etc.
- Linguistic variability – e.g., a request for production may specify that all "consumer *product feedback*" should be produced, but consumer feedback may touch upon any number of product features, may be positive or negative, may appear in formal reports or informal emails, and may be expressed in any number of unpredictable ways that could prove challenging for automated search systems to capture comprehensively
- Linguistic generalizability – e.g., a request for production may specify that all "negotiations with *retailers*" should be produced, but if the corporate entity routinely deals with thousands of retailers, it would be difficult, if not impossible, for an automated search system to successfully recognize the complete set of potentially relevant retailers and differentiate them from entities such as wholesalers or suppliers, etc.
- Conceptual coherence – e.g., a request for production may specify that all "discussions of *product testing*" should be produced, but if this is intended to include R&D testing, Quality Control testing and Market Research testing, then there will actually be three distinct concepts to capture, each with its own community of expert speakers with unique jargon and communication patterns such that capturing all of these sub-topics equally successfully may challenge automated search systems

These factors interact not only with rate of responsiveness but also with one another to shape the target of the search effort. Analyzing the subject matter of a case to identify attributes that may introduce difficulties for automated search will make it possible to devise methods for overcoming the challenges.

There are, in fact, numerous options for coping with the various situations highlighted above. Sometimes the solution will be as simple as choosing one search technique over another. At other times, it may be most effective to collaborate with the attorney team to operationalize the definition of responsiveness to minimize the need for subjective interpretation or fine-grained subject matter distinctions. At other times, the best choice may be to create distinct models for the most critical sub-topics in an especially wide-ranging request for production to ensure that they will receive ample effort and attention, reducing the risk of having their performance obscured by the search results for other more prevalent topics. Undertaking a preliminary subject matter analysis and consultation with the case team, along with early sampling and testing in the corpus, will typically enable the formulation of a project proposal that will provide value for the client while accommodating the realities of the search situation.

Finally, while much of the above discussion has centered on the use of in-depth analysis and a multitude of search tactics to achieve the highest possible quality results, XLS acknowledges this level of analysis and investment of expert resources is not always feasible. In fact, it may simply be unreasonable given the practical constraints of the case or its proportional value to the primary stakeholders. Open and frequent communication with the attorney team and client for the matter will not only enhance the quality of the subject matter input for the project, but also afford them opportunities to contribute their invaluable expert opinions regarding the reasonableness of the search for the matter at hand.

In sum, XLS adopts the position that search results in e-discovery should be judged relative to the goals that were established for the project and that the search process, rather than the technology alone, should be scrutinized. We recognize it would be advantageous to have a single concretely defined protocol and technology applicable to every matter to achieve high-quality results quickly, cheaply, and defensibly. However, it would be naïve to suggest the unique topics, timelines, resources, parties, data sources and budgetary constraints associated with each matter could all be treated successfully using the same search strategy or the same quantitative measures, especially when current technologies are in a state of growth and evolution. It does a disservice to both the complexity of the problem and to the value of human insight and innovation in tailoring custom solutions to adapt to specific needs.