On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records

Timothy Gollins, Graham McDonald,
Craig Macdonald and Iadh Ounis
School of Computing Science
University of Glasgow, Glasgow, UK
{firstname.lastname}@glasgow.ac.uk

ABSTRACT

Open government facilitate citizens access to government records, through Freedom of Information laws, and through government archives after a period of years (e.g. 20) has elapsed. However, there are growing challenges in established archival processes that have been brought about by the introduction of digital records and the consequent breakdown of the pre-existing administrative practices within government institutions. In this paper, we discuss challenges that arise from two stages in the archiving digital government records, which information retrieval research can address: the selection/appraisal of appropriate records to archive, and the review of those records to ensure that no sensitive information is released. We also suggest tentative solutions for sensitivity review based on our own work.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

Keywords: Sensitivity Review

1. INTRODUCTION

Open government holds that citizens have the right to access the records (documents and proceedings) of government and other public organisations, to facilitate accountability under the rule of law [5]. Freedom of information (FOI) legislation (e.g. in the UK [1], and the US [17]) facilitates this right; citizens can request government documents be provided subject to certain proscribed *exemptions* (e.g. personal privacy, health & safety, commercial confidentiality).

The principles of open government have also been enshrined historically, both in the UK and in other jurisdictions, in that public records must be released to archives after a number of years have elapsed (in the UK 30 years, but now being reduced to 20 [2]). There are two broad models of archival legislation guaranteeing access to public records. These are "Open by default" as typified by the UK [1, 2], and "Closed[1] by default - release on FOI request" typified by the US federal code [17].

Under both access models, it is necessary to ensure that no sensitivities remain in the records released. This requires that the records are reviewed by human assessors who are familiar with the topics concerned and can verify that no exemptions should be applied. For instance in the UK, the mention of a name of an informant in a theatre of conflict, could put their life or family in danger, and the record would be closed on the grounds of health & safety [1, section 38] for up to 120 years. In the US there has been considerable work on the impact of privacy on archival practice [15], with regulations on closure deriving from the constitution, federal codes and state laws.

20-30 years ago, governments in 1st world countries increasingly moved to digital record keeping, as the means of information production became digital (e.g. networked PCs & email). This resulted in substantial changes in administrative practice, a consequent increase in the *volume* and *complexity* of (digital) records kept, and a break down in the previous well-managed patterns of their *organisation* [9, 10]. However, until very recently the archival and records management community have almost exclusively focused on the apparently insurmountable challenge of *preserving* digital records. Recent work has begun to demonstrate that this emphasis on preservation may be misplaced [7]; the more immediate challenge arises in safely capturing the digital records in the first place. This includes difficulties in selection/appraisal and in particular, sensitivity review. Any archive of public records will soon be forced to address both of these issues to ensure open government remains a reality. In the remainder of this paper, we detail the challenges that may be addressed by research in information retrieval (Section 2), and provide concluding remarks and a roadmap for future efforts based on our own work (Section 3).

2. CHALLENGES IN DIGITAL ARCHIVING

In the following, we discuss challenges for information retrieval in the archiving of digital records.

2.1 Selection/Appraisal

When a record reaches the age it should be archived, it must be appraised to decide if it should be kept for permanent preservation. This is an essential response to the unsustainable costs of keeping (storage and conservation or preservation) and finding (curating and cataloguing or indexing) everything. In the digital environment, while some aspects of these costs change substantially, in practice archives cannot afford to keep everything and while all records are important by some measure, some are clearly more important than others; the need to select and appraise remains.

As the volumes of digital records to be deposited in archives around the world increases, archivists will need tools to enable them to efficiently and effectively determine those records worthy of permanent preservation. Archivists must also be able organise digital records in ways that reflect the circumstances of their creation, so that they can be reliably interpreted by historians of the future; in archives, context

[1]Closed records are those records held by an organisation or archive that have yet to be released to public view.

PIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

is king. The breakdown of administrative practices that occurred in the transition to the digital environment [9, 10] means that that the traditional reliance on *metadata* will not work. Tools that can extract and confer meaningful structure on large corpora of digital records based – not only on topic matter, but also on the context of creation and distribution will be essential. This presents a new set of significant and interesting challenges for information retrieval, information extraction, and text classification research, as work on the George Bush Senior presidential archive illustrates [16].

2.2 Sensitivity Review

The challenge of reviewing digital records for sensitivity is particularly acute. Closing significant volumes of public records, as a precaution to prevent a small volume of truly sensitive records being released, is lawful in some jurisdictions when justified by the cost of review [1]. However such precautionary closure will not be morally, ethically or politically acceptable in an era of increasingly open government. It is essential that decisions on the closure of records are conducted at the individual document level.

Review for sensitivity may seem similar to the challenge of identifying documents that are relevant to the specific legal matter in a litigation (i.e. e-discovery [12]). However, in the case of sensitivity review, while the nature of a sensitivity can be described (e.g. personal privacy), the specific features that will render the record sensitive are generally unknown to the reviewer in advance. This is because such sensitivities are not only conferred by the content of the record (the topics and entities) but also by the context of creation and distribution (who said what to whom in which circumstances). Finally, sensitivity is not limited to considerations of personal privacy, but also includes commercial confidentiality, health & safety of individuals, matters of defence & national security, and damage to international relations.

In our own proof-of-concept work on Project Abacá [11] we have established that, while for many UK government departments the protection of personal privacy is the most significant issue by volume of records, other sensitivities often represent greater overall risks (e.g. damage to international relations or national security). We have also established that some of the most challenging aspects of privacy protection are shared by sensitivity. Of particular interest is the diffused nature of both privacy and sensitivity, which means that apparently innocuous statements combined with open information or knowledge can result in significant breaches. In this respect, we believe that sensitivity is a wider concept that actually encompasses privacy, and hence solutions to the privacy protection problem that do not address these complex and subtle aspects will be inadequate – indeed, the study of sensitivity is essential to developing general solutions for privacy.

The *volume*, *complexity* and lack of *organisation* of digital records and the risks implicit in an error of judgement (the risk of precautionary closure or the risk of inappropriate opening) together with the nuanced nature of sensitivity makes this field a particularly interesting source of research questions, as we have begun to explore [8].

3. ROADMAP & CONCLUSIONS

In considering digital archival practices as a source of significant research challenges, we have identified a number of strands from our own work on sensitivity review, which have parallels in classical IR tasks and research. We draw on this classical work to inspire the extension of the field to address sensitivity (and thus privacy) review.

This includes: understanding human judgement of sensitivity (c.f. [19]), the identification of features (from the document or its context, explicit or implicit) that indicate sensitivity (c.f. [6, 13]), understanding the relationship between *automation* of sensitivity review and technical *assistance* of human reviewers in managing the risks of review (c.f. [3, 18]), understanding the significance of order of presentation of documents in the human sensitivity review task, and thus in machine *assistance* (c.f. [4, 14]).

Our own work with UK government departments [11] makes it clear that a fully automated approach to sensitivity review is unlikely to be acceptable. There is a clear reluctance on the part of reviewers to trust technology alone. Nevertheless, in the UK at least, many recognise the challenges brought about by the digital age, and the need for new methods and tools to explicitly manage the increased risks from the open release of digital records in the era of internet search.

Our work in developing our test collection has shown the value of close observation and study of human reviewers in beginning to understand the nature of sensitivity. It also helped us to identify additional document and context features to classify for sensitivity; the application of a simple bag-of-words text classification baseline appears inadequate [8]. The development of a learned classifier, drawing on features extracted from a representative test collection, appears to be a fruitful starting point to develop a decision support and review prioritisation tool [8].

4. REFERENCES

[1] Freedom of Information Act 2000 (UK).

[2] Public Records Act 1958, as amended (UK).

[3] L. Azzopardi. The Economics in Interactive Information Retrieval. In *Proceedings of SIGIR 2011*.

[4] G. Berardi, A. Esuli, and F. Sebastiani. A utility-theoretic ranking method for semi-automated text classification. In *Proceedings of SIGIR 2012*.

[5] T. H. Bingham. *The Rule of Law*. Penguin, London, 2011.

[6] G. Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *J. Mach. Learn. Res.*, 3, 2003.

[7] T. Gollins. Putting Parsimonious Preservation into Practice. Tech Report, The Natnl Archives, 2012. `http://bit.ly/1m4xerK`

[8] G. McDonald, C. Macdonald, I. Ounis, and T. Gollins. Towards a Classifier for Digital Sensitivity Review. In *Proceedings of ECIR 2014*.

[9] M. Moss. The Hutton Inquiry, the President of Nigeria and What the Butler Hoped to See. *English Historical Review*, 120(487), 2005.

[10] M. Moss. Where Have All the Files Gone? Lost in Action Points Every One? *J. Contemporary History*, 47(4), 2012.

[11] Project Abacá. Project Website. `http://projectabaca.wordpress.com/`. 2014.

[12] D. W. Oard and W. Webber. Information Retrieval for E-Discovery. *Foundations and Trends in Information Retrieval*, 7(2-3), 2013.

[13] T. Qin, T.-Y. Liu, X.-D. Zhang, Z. Chen, and W.-Y. Ma. A Study of Relevance Propagation for Web Search. In *Proceedings of SIGIR 2005*.

[14] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber. The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment. In *Proceedings of SIGIR 2013*.

[15] D. J. Solove. Access and Aggregation: Privacy, Public Records, and the Constitution. *Minnesota Law Review, 86(6), 2002*.

[16] W. E. Underwood. Speech Acts and Electronic Records. *Proceedings of DigCCurr2009 Digital Curation: Practice, Promise and Prospects*, 2009.

[17] 5 U.S. Code §552 - Public information; agency rules, opinions, orders, records, and proceedings.

[18] W. Webber. Approximate Recall Confidence Intervals. *Trans. Inf. Syst.*, 31(1), 2013.

[19] W. Webber and J. Pickens. Assessor Disagreement and Text Classifier Accuracy. In *Proceedings of SIGIR 2013*.