Evaluating Search Among Secrets

Douglas W. Oard
iSchool and UMIACS
University of Maryland
College Park, MD USA
oard@umd.edu

Katie Shilton
iSchool
University of Maryland
College Park, MD USA
kshilton@umd.edu

Jimmy Lin
Computer Science
University of Waterloo
Waterloo, ON Canada
jimmylin@uwaterloo.ca

ABSTRACT

Today's search engines are designed with a single fundamental goal: to help us find the things we want to see. Paradoxically, the very fact that they do this well means that there are many collections that we are not allowed to search. Citizens are not allowed to search some government records because there may be intermixed information that needs to be protected. Scholars are not yet allowed to see much of the growing backlog of unprocessed archival collections for similar reasons. These limitations, and many more, are direct consequences of the fact that today's search engines are not designed to protect sensitive information. We need to change that by creating a new class of search algorithms designed to effectively "search among secrets" by balancing the user's interest in finding relevant content with the provider's interest in protecting sensitive content. This paper describes some first thoughts on evaluation for that task.

1. INTRODUCTION

In late 2014, presidential candidate Hillary Clinton sent 30,490 work-related email messages to the United States Department of State and asked that they be reviewed and released as quickly as possible. With 25 people assigned to the office coordinating the process, the review took more than a year. In late 2014, presidential candidate Jeb Bush chose a different approach, releasing all of the approximately 280,000 email messages from his time as Governor of Florida. That email was posted to the Internet and then removed two days later after independent sources identified the presence of sensitive content. Neither approach is able to provide responsive access at reasonable cost while protecting sensitive information. The fault lies not with those who are using the available technology to perform this task. Rather, the problem is designed into the very nature of modern search technology.

The fundamental issue is that search engines are designed to find things. Indeed, every widely-used measure of retrieval effectiveness is designed to characterize how close the search engine can come to the ideal of showing the user everything in its index that is relevant to the request, and nothing else. This perspective is an artifact of the intellectual heritage of search engines in the mission of a library – to provide to their users that which they wish to see. The archival profession, by contrast, has a more nuanced mission that explicitly recognizes the need to exercise discretion when providing access to potentially sensitive materials. Archives often place access restrictions on materials for reasons ranging from donors' requests for privacy to national security concerns. In essence, what we need is a search engine that works more like an archive, protecting sensitive content while providing access to the rest.

If sensitive content were marked as sensitive at the time of creation, protecting it would be easy. In an earlier era when information was scarce relative to human attention, segregating sensitive information from that which could be made public was the norm. For example, we could presume that newspaper articles were intended to be public, and that telephone calls were intended to be private. Today, however, it is human attention that is scarce relative to the quantity of information that we all generate, and digital media increasingly collapse what used to be segregated information contexts. As a result, our digital records are an intermixed cacophony of the sensitive, the important, and the banal. As one example, the George W. Bush Presidential Library has a collection of about 200 million email messages from the Executive Office of the President alone. If one email message could be reviewed for release per minute, a team of 25 would require 67 (2,000-hour) years to review those 200,000 email messages. And one person-minute is exceedingly short for a manual review. This digital tsunami is about to get far larger: the Obama administration has directed that by 2019 all federal agencies should preserve their permanent records in digital form. If we cannot automate the task of protecting sensitive content while still allowing content which is not sensitive to be found and used, only a small fraction of the information being generated today will ever become searchable.

It need not be this way. Search engines are built by first quantifying what we wish them to do (i.e., designing the evaluation measure that will characterize their effectiveness) and then by automatically tailoring search algorithms to produce the best possible results according to that evaluation measure, a process generically referred to as "learning to rank." We therefore begin by designing a class of evaluation measures that balance relevance (the ability to find what we want) with sensitivity (the ability to identify that which needs protecting). We begin at the heart of the problem by articulating the structure of a family of evaluation measures. We then describe the process by which we intend to explore the resulting parameter space. Finally, we offer a few thoughts on the broader impact of this work.

2. AN EVALUATION MEASURE

Search engines are typically tuned to optimize their performance as measured by some evaluation measure, of which Discounted Cumulative Gain (DCG) is a typical example:

$$DCG_k = \sum_{i=1}^{k} \frac{g_i}{d_i}$$

where g_i is the gain earned by the document ranked in position i, d_i is some predetermined (monotonically increasing) discount factor associated with returning a relevant document at rank i, and k is some rank below which the gains are implicitly modeled as zero (representing the greatest depth a typical user is expected to

examine) (Järvelin & Kekäläinen, 2002).[1] The design of the DCG evaluation measure reflects two fundamental characteristics of ranked retrieval: (1) some documents are more highly relevant than others (thus earning higher gain), and (2) placing a relevant document earlier in the ranked list is better than placing that same document further down. Implicit in the definition of DCG is some gain function with the structure described in the following contingency matrix:

	Highly Rel (h)	**Moderately Rel(m)**	**Not Rel**
RETRIEVED	$+G_h$	$+G_m$	0
NOT RETRIEVED	0	0	0

When searching among secrets, we must consider both the gain that results from showing a relevant document and the costs that accrue for showing a sensitive document. Expanding the contingency matrix to account for this, we get a three-dimensional tensor cost function with the following two planes:

RETRIEVED	**Highly Rel (h)**	**Moderately Rel (m)**	**Not Rel**
Fine to Show	$+G_h$	$+G_m$	0
Somewhat Sensitive (s)	$-C_s$	$-C_s$	$-C_s$
Very Sensitive (v)	$-C_v$	$-C_v$	$-C_v$

NOT RET	**Highly Rel**	**Moderately Rel**	**Not Rel**
Fine to Show	0	0	0
Somewhat Sensitive	0	0	0
Very Sensitive	0	0	0

In this example we illustrate graded relevance and graded sensitivity using three levels for each, but the formalism is easily extended (or collapsed) to any number of gradations along each dimension. For ease of interpretation, we show positive values (gains) as $+G$ and negative values (costs) as $-C$. Naturally, we set the cost for showing "very sensitive" content ($-C_v$) to be more strongly negative than the cost for showing "somewhat sensitive" content ($-C_s$). Because the traditional DCG measure implicitly assumes that all documents are "fine to show," the first rows of each plane in this expanded contingency matrix correspond to rows in the DCG contingency matrix. While discounting gain for showing relevant content lower in a ranked list is a reasonable model of user behavior (albeit somewhat oversimplified), discounting sensitivity using the same factor would not generally be appropriate. We therefore model the cost of displaying sensitive content as being accrued in a manner that is insensitive to the rank at which the document containing that sensitive content is displayed. This results in the following definition for our new evaluation measure, which we call Cost-Sensitive Discounted Cumulative Gain (CS-DCG):

$$CS-DCG_k = \sum_{i=1}^{k} \left(\frac{g_i}{d_i} + c_i\right)$$

Where g_i is the gain (if any) associated with showing the document at rank position i, c_i is the cost (if any) associated with showing the document at rank position i, and the other parameters are defined as for DCG above. Because costs are negative, the additive c_i term reduces the value of CS-DCG any time a sensitive document is shown to the user. While DCG is strictly non-negative, CS-DCG has no similar range restriction (although for ease of interpretability, and for comparability when averaging CS-DCG results across topics, the values can be normalized to lie between zero and one, as is the convention for Normalized DCG). One other important difference is that in DCG the rank cutoff k simply reflects expected user behavior, but the search engine need not actually truncate the ranked list at rank k. In CS-DCG, by contrast, truncation at rank k must actually be performed by the search engine in order to prevent unbounded cost growth. We note also that variants of our CS-DCG could also be defined. For example, when k is large, it might make sense to discount the sensitivity penalty for lower-ranked items. As another example, it might be appropriate to assign different costs or gains to some conditions to which we have assigned equal values if more fine-grained distinctions than we have made were consequential.

3. CHOOSING PARAMETERS

To actually instantiate the model, we need to choose reasonable parameters for two gains (G_h and G_m) and two costs (C_s and C_v). To do this, we must first choose some task that involves protection of sensitive content and then construct reasonable gains and costs for that task. It might at first seem that the cost of showing sensitive content to the user could in some applications effectively be infinitely negative. Such a model would not reflect any real application in which we would rationally wish to support search among secrets, however, because any risk of incurring an infinite cost would result in always setting k to zero (i.e., showing the user nothing). The decision to allow search among secrets is thus always a conscious choice to incur some risk of revealing sensitive content in exchange for the anticipated gains resulting from finding relevant documents for the searcher. For our early experimentation, we would therefore prefer to focus on an application in which we expect that the risks, and hence the costs, are modest (at least in comparison with the potential benefits of being able to perform the search. We therefore plan to focus initially on the deposit of personal email in an archive as a part of personal papers collections (Hangal et al, 2015).

An important movement in privacy scholarship approaches privacy not as individual (and therefore unpredictable) preferences, but as a social, contextually-dependent (and therefore generalizable) phenomenon. This theory posits that individuals' privacy expectations are based on social norms within particular information contexts (Nissenbaum, 2009). Those privacy norms dictate what information it is acceptable to collect, who can have access to it, whether it should be kept confidential, and how it can be shared and reused. When privacy expectations are context-specific, norms around what information should be disclosed and gathered, and for what purpose, are developed within a particular community or context. Shopping online, talking in the break room, and divulging information to a doctor are each governed by different information norms. This contextual approach is consistent with a social contract approach to privacy expectations (Culnan &

[1] The original definition of DCG does not include the rank cutoff k but truncated evaluation has since become common, particular in evaluation of Web search.

Bies, 2003; Li et al., 2010; Martin, 2012; Xu et al., 2009) in which rules for information flow take into account the purpose of the information exchange as well as risks and harms associated with sharing information. This approach allows for the development of general norms for context-sensitive information release. These norms take into account:

- Who/Roles – people, organizations who are the senders, recipients, and subjects of information.
- What/Information – the information types or data fields being transmitted.
- How/Transmission principles – the constraints on the flow of information.
- Why – the purpose of the use of information (Nissenbaum, 2009).

Key to all contextual definitions of privacy is how the main components work together – who receives the information, what type of information, how is it used, and for what purpose – within a particular context.

To determine context-appropriate email sensitivity parameters, we plan to conduct surveys using the factorial vignette survey method, originally developed to investigate human judgments (Rossi & Nock, 1982; Jasso, 2006). In a factorial vignette survey, a set of vignettes or stories is generated for each respondent, where the vignette factors are controlled by the researcher. Respondents are asked to evaluate these hypothetical situations. This method enables simultaneous examination of multiple factors – e.g. changes in social context, type of information released, and secondary uses of that information – by providing respondents with rich scenarios that are systematically varied. It also supports identification of implicit factors, and their relative importance, in respondents' privacy expectations. This method (1) allows the investigator to examine multiple factors – e.g. changes in context, types of privacy violations – simultaneously by providing respondents with rich vignettes which are systematically varied, and (2) supports the identification of the implicit factors and their relative importance in making normative judgments – in this case, that a situation meets or violates privacy expectations – within different contexts (Wallander, 2009). The factorial survey method allows for the experimental manipulation of a large number of factors through the use of a contextualized vignette (Ganong & Coleman, 2006), which renders the method well-suited to the examination of highly-contextual concepts such as privacy expectations, where norms should vary based on particular contexts, information types, and information uses. As noted by the recent Federal Trade Commission report, traditional surveys are limited in their ability to measure privacy expectations of individuals (FTC, 2010). Individuals often have difficulty articulating the factors and their relative importance that constitute their privacy expectations. The factorial vignette survey method is designed to avoid such respondent bias by *indirectly* measuring privacy factors and their relative importance of respondents. For example, the respondents will not be explicitly asked if revealing information about family members is appropriate; rather, respondents will rate a vignette wherein information about family members is included among other factors. By asking respondents to rate multiple vignettes (usually 30-50 vignettes), sensitive factors and their relative importance are identified without directly asking for a ranking.

We plan to work with scholars and archivists experienced with email collections to develop vignettes that model the roles, information, transmission principles, and information uses typically found in email archives. Vignette surveys will then be deployed to a national sample of email users in the United States using Amazon Mechanical Turk. Amazon Mechanical Turk is an online labor market where requestors, such as academics, post jobs and the workers, such as the respondents, choose jobs to complete. Though use of Mechanical Turk for survey deployment can be controversial (Lease et al., 2013), studies have shown that Mechanical Turk workers are more representative of the US population than the samples often used in social science research (Behrend et al., 2011; Berinsky et al., 2012).[2] In previous research on privacy expectations of websites, researchers have compared Mechanical Turk results with results from a nationally representative sample. The Mechanical Turk sample produces the same theoretical generalizations as a national survey, illustrating the ability to build generalizable theory from Mechanical Turk samples in online privacy studies (Martin, 2012).

4. SENSITIVITY JUDGMENTS

Processes for obtaining relevance judgments that are suitable for use in training and evaluation are well understood, but we need to gain experience with the process for judging sensitivity. In the TREC Legal Track, sensitivity judgments were made for attorney-client privilege, a well defined concept in the law (Vinjumur, 2015). As a next step, we plan to conduct a similar annotation process for personal email. Our first set of experiments will be conducted using an archived email collection from a prominent scholar who has contributed about 45,000 email messages from a fifteen-year period to a university archives for research use. We will work with that scholar to elicit their personal sensitivity concerns, using the results of our factorial vignette survey to structure that elicitation process. The person who performs that elicitation will then use judgmental sampling to label a training set containing sensitive and non-sensitive messages that span as broad a range of reasons for that sensitivity as possible. We will then perform active learning to select additional documents to label. We will also annotate a stratified sample to characterize the learning rate of our initial sensitivity classifier, and we will cease annotating training data when that classifier's accuracy begins to plateau (or upon exhausting our annotation budget). As we proceed, we will periodically ask the scholar who contributed the collection to annotate the level of sensitivity for a randomly selected subset of the most recently annotated messages to allow us to characterize inter-annotator agreement. If (as we expect will initially be the case) we find substantial disagreement, we will confer with that scholar to refine our annotation guidelines.

5. OTHER APPLICATIONS

As the tasks we have identified above clearly illustrate, the range of applications to which effective and well-characterized techniques for search among secrets could be applied is substantial. We have proposed to test our techniques under conditions that model archival access to email donated for use by scholars. A number of other applications also come to mind. As noted above, one application this is already attracting considerable interest is privilege review in e-discovery (Gabriel et al, 2013). Applications to government transparency, the motivating examples at the start of

[2] In sum, respondent samples on MTurk are found to be representative of the general population with high internal and external validity (Mason & Suri, 2011). Horton et al. (2011) illustrate how behavioral economics experiments are successfully replicated on MTurk.

this paper, are also evident. Our initial interest is in designing techniques for making the sensitivity determinations fully automatically, but similar techniques could also support the first stage of a process in which very highly-sensitive content (e.g., classified materials that must be reviewed for declassification) could be sent for manual review only if current users actually wished to see it. Such techniques could facilitate serving Freedom of Information Act requests (McDonald et al., 2014), as well as review for declassification and public release of the growing backlog of documents requiring systematic review (e.g., after 25 years) (Martin et al,. 2007). Additionally, in the United States, both national (PIDB, 2014; PIDB, 2007) and state (Reinvent Albany, 2014) interests have called for improved ways of prioritizing the review task in order to intelligently manage the huge wave of documents requiring review. Another possible application would be to streamline the "right to be forgotten" that has been recognized by the European Court of Justice (Richards, 2015), which requires that search engine services not find content that has been deemed sensitive by individuals. At present, users must request that items be removed on an item-by-item basis. Such requests are already at a staggering volume, and still growing. In the 18 months between May 2014 and November 2015, more than 340,000 people requested that more than 1.2 million URLs be removed from the index of Google search services that are widely used in Europe, and 42% of these requests were granted. Search among secrets could ultimately lead to much more efficient ways of accomplishing a similar result.

It may be, however, that the most important applications for search among secrets, will be the ones that emerge after the capability is in hand. At present, parents do not want their children to search their email, and children do not want their parents to search their chat logs, because no means exists to assure that the content there that should be private will remain so. Privacy concerns evoked when Google Glass, or similar devices still in the lab, become mainstream are not fundamentally rooted in what might be recorded, but rather in how those recordings might be used. These concerns will remain salient for at least as long as we deny ourselves the ability to search among secrets in ways that balance the interests of both the content creators and those who wish to find and use that content.

6. ACKNOWLEDGEMENTS

This research has been supported in part by NSF award 1065250.

7. REFERENCES

T.S. Behrend, D.J. Sharek, A. W. Meade, and E. N. Wiebe, The viability of crowdsourcing for survey research, Behav. Res. Methods, 43(3), 800–813, 2011.

A.J. Berinsky, G.A. Huber, and G.S. Lenz, Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk, Polit. Anal., 20(3), 351–368, 2012.

Federal Trade Commission (FTC), Protecting Consumer Privacy in an Era of Rapid Change, FTC, 2010.

M. Gabriel, Chris Paskach, and David Sharpe, The Challenge and Promise of Predictive Coding for Privilege, in ICAIL Workshop on Standards for Using Predictive Coding, Machine Learning, and Other Advanced Search and Review Methods in E-Discovery, 11 pp., Rome, 2013.

L.H. Ganong and M. Coleman, Multiple segment factorial vignette designs, J. Marriage Fam., 68(2), 455–468, 2006.

S. Hangal, V. Piratla, C. Manovit, P. Chan, G. Edwards, and M.S. Lam, Historical Research Using Email Archives, in CHI Extended Abstracts, pp. 735-742, 2015.

K. Järvelin and J. Kekäläinen, Cumulated Gain-Based Evaluation of IR Techniques, *ACM Transactions on Information Systems*, 20(4), 422-446, 2002.

G. Jasso, Factorial survey methods for studying beliefs and judgments, Social. Methods Res., 34(3), 334–423, 2006.

M. Lease, J. Hullman, J. Bigham, M. Bernstein, J. Kim, W. Lasecki, S. Bakhshi, T. Mitra, and R. Miller, Mechanical Turk is Not Anonymous, SSRN Scholarly Paper ID 2228728, 2013.

H. Li, R. Sarathy and H. Xu, Understanding situational online information disclosure as a privacy calculus, J. Comput. Inf. Syst., 51(1), 62, 2010.

K.E. Martin, Diminished or Just Different? A Factorial Vignette Study of Privacy as a Social Contract, J. Bus. Ethics, 111(4), 519–539, 2012.

C. Martin, D. Lam, A. Liu and M. Tech, Classification Assistance Prototype System: Final Report, Applied Research Laboratories: The University of Texas at Austin, Technical Report LR-SISL-07-17, 2007; 78 pages.

G. McDonald, C. Macdonald, I. Ounis, and T. Gollins, Towards a Classifier for Digital Sensitivity Review, in ECIR, pp. 500-506, 2014.

H. Nissenbaum, *Privacy in Context: Technology, policy, and the integrity of social life*, Stanford, CA: Stanford Law Books, 2009.

Public Interest Declassification Board (PIDB), *Setting Priorities: An Essential Step in Transforming Declassification*, 40 pp., 2014. https://www.archives.gov/declassification/pidb/recommendations/setting-priorities-print.pdf

Public Interest Declassification Board (PIDB), *Improving Declassification, A Report to the President from the Public Interest Declassification Board*, 2007. https://www.archives.gov/declassification/pidb/improving-declassification.pdf

Reinvent Albany, *Listening to FOIL: Using FOIL Logs to Guide the Publication of Open Data*, 7 pp., 2014. http://reinventalbany.org/wp-content/uploads/2014/07/Final-DEC-FOIL-Analysis.pdf

N. Richards, *Intellectual Privacy: Rethinking Civil Liberties in the Digital Age*, Oxford University Press, 2015.

P. H. Rossi and S. L. Nock, *Measuring Social Judgments: The factorial survey approach*, Sage Beverly Hills, CA, 1982.

J. Vinjumur, Evaluating Expertise and Sample Bias Effects for Privilege Classification in E-Discovery, International Conference on Artificial Intelligence and Law, San Diego, CA, 2015.

L. Wallander, 25 Years of Factorial Surveys in Sociology: A review, Soc. Sci. Res., 38(3), 505–520, 2009.

H. Xu, C. Zhang, P. Shi, and P. Song, Exploring the role of overt vs. covert personalization strategy in privacy calculus, Proceedings of the Academy of Management, 2009.