

# Privacy Classification Systems: Recall and Precision Optimization as Enabler of Trusted Information Sharing

Christopher Hogan and Robert S. Bauer

H5

71 Stevenson Street  
San Francisco, CA 94105  
{chogan,rbauer}@h5.com

## Abstract

Information is shared more extensively when a user can confidently classify all his information according to its desired degree of disclosure prior to transmission. While high quality classification is relatively straightforward for structured data (*e.g.*, credit card numbers, cookies, ‘confidential’ reports), most consumer and business information is unstructured (*e.g.*, Facebook posts, corporate email). All current technological approaches to classifying unstructured information seek to identify only that information having the desired characteristics (*i.e.*, to maximize the percentage of filtered content that requires privacy protection). Such focus on boosting classifier Precision (P) causes technology solutions to miss sensitive information [*i.e.*, Recall (R) is compromised for the sake of P improvement]. Such privacy protection will fall short of user expectations no matter how ‘intelligent’ the technology may be in extending beyond keywords to user meaning.

Systems must simultaneously optimize both P and R in order to protect privacy sufficiently to encourage the free flow of personal and corporate information. This requires a socio-technical methodology wherein the user is intimately involved in iterative privacy improvement. The approach is a general one in which the classifier can be modified as necessary at any time when sampling measures of P and R deem it appropriate. Matching the ever-evolving user privacy model to the technology solution (*e.g.*, active machine learning) affords a technique for building and maintaining user trust.

## Introduction

The unprecedented ability to share all forms of information at the click of a button provides countless opportunities for

improved productivity and collaboration for both individuals and businesses. At the same time, concerns over privacy and privilege lead to practices that inhibit the optimal flow of valued information. Laws, agreements, policies, processes, and governance are all useful for increasing trust that information, once shared, will benefit the transmitting person or organization. These most often focus on the proper use and protection of the information asset by the receiving entity.

In this paper, we look at user-based information classification for assessing and applying privacy criteria prior to potential sharing. A socio-technical system architecture (Ropohl 1999; Mate and Silva 2005) is presented wherein a user interactively and iteratively trains technology to replicate and automate his judgment as to privacy criteria that should be applied to information. We argue as a general matter that *a priori* privacy classification can be the most vital tool for increasing confidence in the protection of personal and corporate privacy. However, independent of the technology used, common approaches for identifying sensitive information fail to balance properly the need to make available that information which does not require protection with the requirement to prevent the release of genuinely private information. We present data showing that simultaneous optimization of both these aspects of information collections is required to establish trust that spurs sharing.

## Privacy Classification and Trust

We propose that accurate, proactive information classification is sufficient to increase the release of information to others. When a computational categorizer can faithfully execute privacy criteria, a user develops enhanced trust in using tools that take subsequent action based on the protection classification. A myriad of *a*

*posteriori* actions may then be taken to protect privacy; these include (a) warnings to the user of sensitive information characteristics when a ‘send’ action is initiated, (b) automated redaction of susceptible data like social security numbers, and (c) filtering to prohibit the transmission of certain information. Independent of legal regimes established to encourage information sharing across jurisdictions, user-directed privacy protection is a vital element in any systemic solution that seeks to fuel and cultivate innovation.

### Classification Accuracy Measures

Information sharing substantially increases to the extent that the privacy classifier is accurate in achieving the user’s protection objectives. Ideal classification accuracy is obtained when (a) ALL data is identified to which a user would attach a privacy requirement [*i.e.*, 100% Recall (R)] and (b) when such tagging is applied ONLY to information requiring such a privacy action [*i.e.*, 100% Precision (P)]. While such an accuracy regime is required for trusted information sharing, it is rarely the goal of information retrieval (IR) technology solutions. All IR assessments undertaken by the NIST Text Retrieval Conferences from 1992 through 2007 focused on P maximization (Voorhees and Harman 2005). As shown in Fig. 1, research technologies spanned from logistic regression to fuzzy matching for tasks ranging from web searches to streamed data filtering.

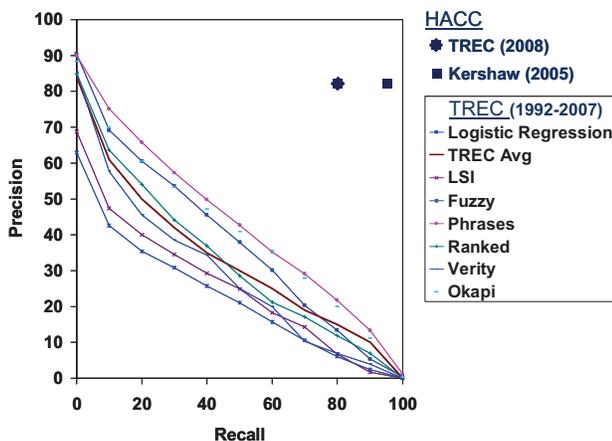


Figure 1. Precision and Recall results for a range of ‘intelligent’ technologies (boxed legend) from a representative selection of TREC tracks (Voorhees and Harman 2005). Human-Assisted Computer Classification (HACC) is the only system methodology that achieves at least 80% R with at least 80% P (Kershaw 2005; Oard et al. 2008). Such joint P & R optimization is necessary for users and companies to trust that their privileged information is being scrubbed or sequestered to protect privacy.

The focus on extreme P is understandable when one considers applications such as consumer search and email filtering. A Google user seeks to find a few highly relevant

information sources in response to a query, rather than all relevant material. Similarly, in reducing liability from financial transaction communication, banks cannot afford to block transmission of four out of five emails that are likely to contain sensitive information because less than one out of five suspect messages would then be correctly classified as potential SEC violations (*i.e.*,  $R > 80\%$  is assessed as having  $P < 20\%$ ). In fact, this latter example supports the proposition that P and R must be jointly optimized when information owned by one party is to be distributed to another. Establishment of legal regimes that encourage sharing and protect privacy is much more likely to succeed when individuals and institutions have internal systems that provide trusted assessment of privacy relevance prior to possible information dissemination.

### Socio-Technical Systems

The accuracy of identifying information that satisfies privacy criteria has a much greater dependence on system architecture than on the underlying classifier technology. Independent of the distinctions for the seven representative ‘intelligent’ technologies in Fig. 1, there is an inherent trade-off of P and R that limits the efficacy of any purely technological solution. The only approach that has been shown to break this paradigmatic, inverse dependency between P and R is one based on a symbiotic relationship between user and technology (Bauer et al. 2009). Such a blended, ‘socio-technical’ methodology is at the heart of systems that use techniques such as supervised machine learning (Xu, Hogan, and Bauer 2009) and sensemaking (Bauer et al. 2008). As illustrated by the two Human-Assisted Computer Classification (HACC) results (Kershaw 2005; Oard et al. 2008) in Fig. 1, we find that such a system can consistently manage information with simultaneously high P and R. Fig. 2 depicts this type of approach to implementing privacy classification solutions that can optimize both P and R.

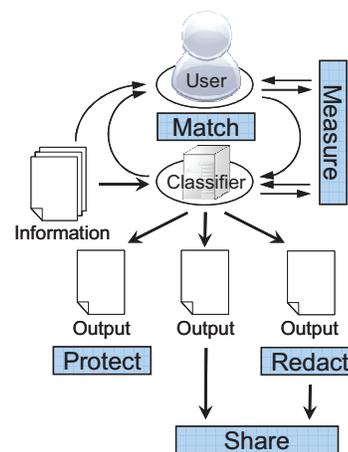


Figure 2. A socio-technical, HACC system for iteratively improving both Precision and Recall of information relevant to multi-faceted privacy criteria.

An effective privacy classification system enables a user to iteratively improve the accuracy of the classifier's selection to her security criteria. Randomly sampled information tagged by the classifier technology is presented to the user for verification of the categorization. Weightings of automatically selected information features are then adjusted in order to improve the measurements of P and R for the sampled data. With active machine learning algorithms, the relative performance of P and R can be tuned depending on the privacy goals of the user. We find (Xu, Hogan, and Bauer 2009) that the harmonic-mean accuracy,  $F_1$ , remains roughly constant (e.g., ~80%) as P or R are raised well beyond 80% (e.g., to ~95%) with a corresponding reduction in the other statistic.

## Discussion

The classifier need not make binary assessments. As depicted in Fig. 2, a user might decide that more information can be shared as long as certain content elements are removed first. Redactions may include such sensitive information as personal identity details or colleague email addresses. The number of different privacy regimes automated in a single system is only limited by the practicality of making meaningful distinctions. When more than a few categories are required, efficient clustering technology (Pendar 2008) is found to be of substantial benefit.

The barrier to using such a system is low relative to the benefit of increased privacy filtering. Initial training of a machine learning algorithm can be effective with 10 or fewer user judgments (Xu and Akella 2008). The test and measurement process is then repeated iteratively until tagged information, sampled from successive classifier outputs, matches the user's privacy evaluation to within the desired P and R. Output from randomly sampled information input is automatically selected using classifier parameters that characterize information diversity and privacy ambiguity (Xu, Akella, and Zhang 2007). We find that not only is this methodology the most efficient way to tune classifier performance, but also it helps the user become aware of content ambiguity and develop greater specificity in his privacy criteria (Brassil, Hogan, and Attfield 2009). Such easy-to-use, interactive system features will lead directly to increased trust in automated information classification.

## Conclusion

In this paper, we argue that automated, multi-variant privacy classification prior to transmission will significantly encourage ubiquitous, trusted information sharing. While technological tools can maximize the fraction of identified information that actually meets privacy criteria (i.e., high output Precision), only a socio-technical solution can simultaneously optimize Precision while labeling a high percentage of the truly private

information (i.e., simultaneous high output Recall). Such privacy classifier 'systems' must be capable of efficient, iterative tuning in order to achieve effective performance. Such characteristics also provide an evergreen capability for tuning updates that ensure trusted performance as user privacy criteria inevitably change; this is particularly important for the foreseeable future given the evolving context of information use by external entities, as in emerging web 3.0 services.

## References

- Bauer, R. S.; Brassil, D.; Hogan, C.; Taranto, G.; and Brown, J. S. 2009. Impedance Matching of Humans  $\leftrightarrow$  Machines in High-Q Information Retrieval Systems. In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*. San Antonio, Texas.
- Bauer, R. S.; Jade, T.; Hedin, B.; and Hogan, C. 2008. Automated Legal Sensemaking: The Centrality of Relevance and Intentionality. In *Proceedings of Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings (DESI II)*, 301. London.
- Brassil, D.; Hogan, C.; and Attfield, S. 2009. The Centrality of User Modeling to High Recall with High Precision Search. In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*. San Antonio, Texas.
- Kershaw, A. 2005. Automated Document Review Proves its Reliability. *Digital Discovery & e-Evidence*, November.
- Mate, J. L., and Silva, A. eds. 2005. *Requirements Engineering for Sociotechnical Systems*. Hershey, PA: Idea Group Publishing.
- Oard, D.; Hedin, B.; Tomlinson, S.; and Baron, J. 2008. Overview of the TREC 2008 legal track. In *Proceedings of The Seventeenth Text Retrieval Conference (TREC-2008)*.
- Pendar, N. 2008. RASCAL: A Rapid Sketch-Based Clustering Algorithm for Large High-Dimensional Datasets. Unpublished Manuscript.
- Ropohl, G. 1999. Philosophy of socio-technical systems. *Society for Philosophy and Technology* 4(3).
- Voorhees, E. M. and Harman, D. K. eds. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, Mass.: MIT Press.
- Xu, Z.; and Akella, R. 2008. A Bayesian Logistic Regression Model for Active Relevance Feedback. In *Proceedings of the 31st ACM SIGIR Conference*.
- Xu, Z.; Akella, R.; and Zhang, Y. 2007. Incorporating diversity and density in active learning for relevance feedback. In *Proceedings of the 29th European Conference on Information Retrieval*.
- Xu, Z.; Hogan, C.; and Bauer, R. 2009. Greedy is not Enough: An Efficient Batch Mode Active Learning Algorithm. In *Proceedings of the 1st Workshop on Large-scale Data Mining: Theory and Applications (LDMTA 2009)*.