Exploring Technology Implications of Advanced Sensitive Information Discovery and Analysis Methods in Large Organizations

ICAIL 2017 | DESI VII

Ben Ferko
benjamin.r.ferko@pwc.com

Josh Rattan
joshua.r.rattan@pwc.com

Abstract

Research and development in the fields of text mining and analytics, information retrieval, computational linguistics, and natural language processing have always struggled to overcome the fundamental problem with electronic unstructured text - the inherent ambiguity of human language.

This ambiguity quickly becomes problematic for computers that, by their very nature, are binary rule-driven machines that rely on absolutes. To resolve the issue of ambiguity, emerging analytical methods and techniques that encompass language models, probability distributions, linear algebra, content models, and artificial neural networks have been applied and tested to approximate and interpret the meaning of human language and the information that it conveys.

In the case of understanding whether information is considered sensitive or not, a binary question at its core, the answer is: it depends.

Most organizations intuitively recognize that "sensitivity," much like "relevance," is not absolute. In large part, the complexity of analyzing sensitive information is due to the imprecise, diverse, and evolving definition of sensitive information in organizational environments. Combined with the large employee population, broad set of technology systems, and exponential data growth, organizations are confronted with difficulties identifying and managing sensitive data.

While the academic and research communities have developed and evolved evaluation methods and processes to mitigate this phenomenon and make progress in the field, a single recognized and consistent evaluation approach in organizational environments has not yet emerged.

Absent of a widely accepted and recognized evaluation approach, commercial technology vendors generally have not had the demand to incorporate functionality to facilitate a measurement and evaluation process in their solutions. Consequently, without quality assurance indicators, on-going proof of sensitive information analysis success, or failure, in an organization goes largely unmeasured.

An organization attempting to measure an analytical solution's effectiveness is left to formulate its own conclusions. As an example, if an analysis approach utilizes content models such as taxonomies or ontologies to identify and classify sensitive content, how will the organization measure the effectiveness of the taxonomies or ontologies, and what happens if they construct them poorly? In large part, the implications of the chosen approach within an organizational context dictates whether the sensitive data analysis will have the desired impact and results.

In this position paper, we propose the consideration of a layered and multi-faceted sensitive information analysis approach in-light of ever-changing, and commonly domain specific, definition of "sensitive" information. We recommend that researchers, organizations, and software vendors alike consider the implications when designing and implementing sensitive information analysis applications, taking into account the impact and demands it places on an organization in order to be effective. Lastly, we advocate for continued efforts to define a consistent and standardized sensitive information evaluation methodology which ultimately may be integrated in commercial technology applications in an effort to allow organizations to measure effectiveness.

The Problem with Defining Sensitive Information

What defines "sensitive" information? A general, yet complicated question to answer. While many organizations may instinctively agree that a certain data element, or content type, may be deemed sensitive, in reality the definition of sensitive information varies greatly, is not well understood by those within an organization whose job it is to manage and protect that information, and is not typically represented in an easily identifiable form.

Regulatory frameworks may appear to have sensitivity definitions outlined with alternative labels such as Personally Identifiable Information (PII), Protected Health Information (PHI), Personal Data, Qualified Financial Contracts, Material Non-Public Information (MNPI) or Sensitive Identifiable Human Subject Research Data. However, the definition of those labels often reduce down to interpretations of terms such as “critical”, “identifiable”, or “personal.”

In some circumstances, “identifiability” is easier to define. Data elements that were originally designed to be unique identifiers are obviously identifiable, personal, and sensitive (e.g. U.S. Social Security Numbers (SSN), Employee ID, Health Insurance IDs). However, the vast majority of information or data elements are not distinct, yet remain sensitive in certain contexts.

Defining sensitivity and defining alternative labels remains elusive because such definitions and alternative labels are temporal, and may be dependent on other factors such as what other information is known in combination with the data. For example, take into consideration data attributes that don’t initially appear identifiable; Sex, Birthdate, and Zip Code. These data attributes, amongst themselves, intuitively don’t seem particularly sensitive, and in fact appear quite benign. Yet, research from Carnegie Mellon[1] suggests that when combined with publicly available Census data, seemingly non-sensitive data transforms and becomes personally identifiable information through inference with accuracy as high as 87%. This example is just one reason why information management strategies designed to identify sensitive information often fall short in-light of the best intentions.

Moreover, organizational strategies as a whole are limited in their ability to identify and manage sensitive information. One of the most common strategies of sensitive information identification in large organizations manifests itself through policies that rely on manual efforts by business users to label and tag sensitive content. Policy frameworks, often which include a three-to-four tiered classification model, rely on business users to familiarize themselves with the model and manually identify and label sensitive information. This strategy, based on policy adherence, ostensibly seems simple. Yet, it overlooks a significant expectation on a business user and a few key assumptions. First, this strategy devolves into an expectation on a business user to perform a conditional and forward-thinking assessment of risk and sensitivity for all content they access and manage on a daily basis. Secondly, it assumes that a business user can understand a policy and classification framework where by the data they are creating or handling fits neatly into the sensitivity definitions within the policy and dictates clear and unambiguous actions to be taken. Finally, it assumes the classification framework encompasses all dynamics and requirements to manage sensitive information by the organization, including but not limited to, the technological capabilities the organization currently has to comply with the policy framework.

Some may argue that this common sensitive information management strategy is misguided. Nevertheless, it demonstrates a critical role for advanced sensitivity information discovery and analysis approaches that overcome the limitations of sensitive information definition and identification.

Overcoming Sensitivity Definition Challenges

To overcome the fallacy of binary sensitivity and relevance assumptions, practitioners in associated fields have relied upon classification, retrieval, and evaluation measurement processes and concepts such as pooling, adjudication, topic authorities, relevance feedback, among many other methods for continued advancement in sensitive information discovery and analysis.

These evaluation processes and revelations illustrate insight not yet recognized in organizational environments with large collections - sensitive information discovery and analysis is an imperfect and point-in-time exercise that requires iterative qualitative analysis utilizing established evaluation methodologies and processes. Often, organizations have written-off analysis approaches and technologies wholesale due to an inability to rationalize the dynamic between false positives and false negatives; something well established with measurement scores of

[1] http://ggs685.pbworks.com/w/file/fetch/94376315/Latanya.pdf

precision and recall.

This dynamic also points to another observation not yet widely recognized in organizational contexts. Qualitative evaluation processes facilitate the measurement of complex and multifaceted sensitive information analysis approaches. At this point, methods to identify and analyze information are becoming so complex that transparency into the inner-workings of a sensitive information and analytics tool is difficult to communicate. Today, most approaches to discover and analyze sensitive information utilize approaches and concepts spanning across library science, statistics, probability, and computer science. Moreover, the expansive and sophisticated combination of techniques places a significant burden on an organization attempting to utilize emerging methods and techniques.

With such demands placed on an organization, an established and well-understood evaluation methodology becomes paramount for successful implementation and operation of sensitive information analysis approaches, and promotes more advanced emerging and multifaceted methods to identify and analyze sensitive information.

Considering Implications of Sensitive Information Analysis Approaches

Putting aside the challenges described above, practitioners must eventually identify the most appropriate methods and complementary technology to identify and analyze sensitive information in large collections. In our opinion, there isn't a single clear answer, and it may be more pragmatic to use a multi-faceted approach that relies heavily on an evaluation and measurement process that bypasses traditional sensitive information definition shortcomings. However, organizations must be mindful of the ramifications and implications of the sensitive information and analysis methods and techniques utilized, as the ramifications and implications may determine whether such methods or approaches are ideal for that particular use case or organization, and ultimately determine whether the underlying initiative is considered successful.

Traditionally in large organizations, when sensitive information discovery and analysis are discussed, more often than not, use cases pertain to information security, message supervision, eDiscovery and investigations, privacy, and intellectual property; with respective analysis methods and approaches including pattern matching, keyword searches, and training-based classifiers. For each of the sensitive information discovery and analysis methods referenced, all come with a set of implications and consequences that organizations, to-date, have struggled to reconcile and overcome.

When sensitive information discovery and analysis programs are part of information security or message supervision programs, companies and technology vendors alike often reduce sensitive information down to data elements that are presumed to be "personally identifiable" and "sensitive" such as social security numbers and credit card numbers and pursue analysis methods based on pattern matching. Notwithstanding the shortcomings of such sensitive information definition logic, an immediate consequence of an approach that quickly manifests is the mountain of false positives that pile up and become unmanageable. Examples of false positives may include U.S. zip codes that are flagged as U.S. social security numbers, invalid credit card numbers, and so on. In such a narrow definition of sensitive information, organizations still find themselves looking for a needle in the haystack. While there are certainly technological improvements practitioners can make to improve the precision of such methods; including the use of validation and checksum rules such as the Luhn algorithm and the United States' SSN Numbering Scheme, the consequence of such an approach implies that pattern matching may not be appropriate for other sensitive information discovery and analysis use cases.

In our view, this phenomenon highlights one implication to consider - the different accuracy needs, such as precision and recall, of a particular use case. While this is widely known in the academic community and illustrated easily with the information retrieval needs on the Internet, it is not widely understood in organizational environments. For Data Loss Prevention (DLP) or Message Supervision programs, low precision may be improved and overcome through tuning, validation algorithms, or a set of remediating controls and processes, but for other use cases, low precision and false positives may prove insurmountable.

Use cases that illustrate different sensitive information discovery and analysis accuracy needs are eDiscovery and similar investigations. In this context, both high precision and recall are required. Additionally, such approaches, as in the case of pattern matching, have their limitations with how they define and begin to identify sensitive information - such as keywords. However, beyond the limitations of keywords as the identifier of sensitive information and the importance of recall accuracy needs, lies commonly overlooked considerations and limitations of prevalent methods and approaches supporting eDiscovery and investigative use cases.

Such approaches are dependent on one critical function to support their sensitive information analysis methods - indexing. Electronic indexes remain effective for search, facilitating rapid identification and retrieval tasks, much like an appendix of a book. Additionally, much like an appendix of a book, when new content is added or edited, indexes must be rebuilt and often take much longer to be updated to reflect the added or changed content. IT professionals and computer scientists are largely familiar with this dynamic. Often referred as *Write vs. Read* speeds, an index by its very nature is designed to read fast, but update or write, slowly. While this may seem like a tedious technical detail, when electronic indexes are repurposed for new needs such as eDiscovery or sensitive data identification, issues and limitations are revealed.

In particular, when indexes try to keep up with the ever-growing mountain of content, indexes will have a tendency to fail. The result of failed indexes will be search results that are not comprehensive and thus creating an inherent impact on recall. With Internet search applications, this is not necessarily a deal-breaker, as long as the most precise search result is still placed at the top of the ranked list. However, in eDiscovery and other investigative use cases such as sensitive data discovery, inconsistent or inaccurate recall measures may call into question the entire investigative matter. Lawyers and investigators may presume, that such situations are easily logged, identified, and fixed; but often health, validation, and alerting mechanisms are absent in some of the most commonly deployed eDiscovery and investigation technology on the market. This example, highlights another technology implication for practitioners to consider - the provenance of the technology and analysis approach and how it is being applied in sensitive information use cases. Organizations must be aware of the origins of discovery and analysis methods and understand when there is a mismatch between the original design and current application of sensitive information discovery and analysis technology.

With sensitive information analysis use cases and contexts such as intellectual property, privacy, or even more modern eDiscovery analysis methods; the definition of sensitive information and corresponding approaches to discovery and analysis have become much more complex. Increasingly, training based classification algorithms have been applied where examples of sensitive information is provided and an algorithm attempts to identify similar content. Alternatively referred to as “predictive coding”, technology assisted review, or generally as machine learning; such methods have proven they can be effective in certain scenarios with adequate configuration, training, and tuning.

With these approaches, an immediate limitation is surfaced - transparency into the inner workings of the algorithm and the ability to understand and communicate the approach and its efficacy. Practitioners looking to implement such techniques, must eventually have the capacity to “get underneath the hood” and configure the tool more broadly than giving training examples and annotating text. Latent variables, high-dimensional vector space, log-likelihood functions, and other complex concepts demand highly specialized data and information science experience that is not yet prevalent in many organizational environments. Practitioners looking to design and implement emerging analytical methods must recognize the operational limitations of organizations, and must design methods in a balanced manner in which it can be reasonably implemented and operated in organizational environments, while giving expert users the flexibility to further customize the solution to optimize performance. Furthermore, organizations seeking to utilize emerging analytical techniques must recognize the demands and dependencies of an approach or underlying solution. For example, if an approach utilizes content models that must be designed and maintained against a set of standards and best practices, practitioners must recognize the likelihood of such governance needs will be met by an organization.

As sensitive information analysis methods continue to become more complex, with the emergence of neural networks or “deep learning” applications, this will only continue to demonstrate the critical role of a well established sensitive information evaluation approach to illuminate their effectiveness in certain sensitive information contexts.

Establishing a Common Evaluation Methodology

So, how does a methodology become established? Besides befitting academic and research conferences and publications such as DESI, one of the most effective methods to establish a process is to integrate it into a technology solution itself. In theory, business analysts and consultants may argue that technology should adapt and accommodate an established business process. Nevertheless in actuality, it is often the case that the technology typically dictates some, if not all, of the terms of business processes and methodologies in organizational contexts.

As such, organizations must seek out and communicate the desire to have evaluation methodologies and quality assurance indicators incorporated into sensitive information analysis solutions. Software vendors and developers must also recognize the opportunity to enhance the adoption and success rate of their products through showcasing their solution effectiveness and providing the means to improve product performance. Without providing a facilitated evaluation and tuning process, organizations will continue to struggle with sensitive information discovery and analysis and will rely on other indicators to determine effectiveness of emerging analysis methods and solutions.

Conclusion

In today's ever increasingly connected world, sensitive data discovery as a use case has never been more important. The implications of not managing sensitive data in accordance with regulatory, legal, and business requirements has far reaching implications. The challenge for organizations only increases with the growing volume of information and the borderless enterprise under which they've grown accustomed.

What constitutes a piece of information or a data element as "sensitive" is not easily discernible and in many instances, requires trained topic authorities to consult and determine the answer. At scale within organizations managing large data sets, putting computational powers to bear against sensitive data discovery becomes a necessity - humans can't go it alone - but a necessity that requires an understanding of the capabilities and limitations of the tools at hand.

The ability to apply the right layers of algorithms, methods, and processes requires an organization to equip itself with an understanding of the benefits and tradeoffs of those that are available.

Technology organizations offering capabilities that purport to aid the sensitive data discovery process also have an emerging obligation to equip their software with measurement and tuning capabilities. Given the "gray area" that comes with what constitutes sensitive data and that not all sensitive data is easily identifiable from matching a pattern or detecting an expression, an ability to tune and measure a particular solution's effectiveness is critical. This concept is well accepted and understood within academic spheres, but organizations have not held their software vendors accountable to create these important capabilities.

Finally, to aid in the development of the measurement and tuning capabilities within emerging solutions, a common evaluation methodology that can be used across sensitive data discovery contexts is needed. This methodology would seek to create an approach that drives consistency and improves outcomes for the application of advanced capabilities that identify sensitive data.

Ultimately, having an ability to employ meaningful measurement and tuning capabilities through commonly agreed-upon framework will aid in improved outcomes.