Declassifier

Current Landscape and Development

In the current state of affairs, advances in technology, government documentation has been eased and has allowed many agencies to produce a considerable amount of data. The National Archives reports that at certain agencies, classified records have increased by 1 million gigabytes every 18 months. That is the equivalent of 557 Olympic sized swimming pools full of printed documents. As the current situation stands, the Archives estimate that two full-time employees are hired per gigabyte of information. The cost of hiring 2 million or more full-time employees for information review requires an absurd amount of resources and time.

Because the staggering volume of classified information that eventually has to be declassified, a technological solution if more appropriate to address the challenges in a more standardized, cost-effective, and more efficient manner. Existing technologies that have been researched and applied can be used to assist in finding an effective solution - information retrieval, optical character recognition, and natural language processing. Information retrieval uses content and word relation, which similar to a google search, assists in finding information and relevant, related information in documents. Optical character recognition or OCR is used to convert written documents into pdf or text files that can be modified and edited. Natural language processing is a branch in computer science combining artificial intelligence and linguistics to assist in manipulation written text and speech. These three tools combined allows us to use existing tools to help formulate a solution rather than expending resources to develop our own.

Handling Security

With the current political landscape shaped by the need for technological secrecy, the leakage of any private information has been taken more seriously than before. Starting with WikiLeaks, moving to Edward Snowden, and even the recent White House stolen emails, the fear of leaking classified information is rampant amongst not just government agencies and critical infrastructure, but by hospitals (digitalization of biomedical records), businesses (Target breach, protecting patents), and more. In 2012, government agencies reported spending $9.77 billion to secure classified information, a $1.59 billion increase from 2011. From that spending, $48.65 million was spent on declassification alone. This means that for every dollar spent on declassifying a piece of information, the government spends $200 more for securing that piece of information. What is even more concerning is that these costs

reported by the Information Security Oversight Office doesn't include estimates from the Central Intelligence Agency, Defense Agency, the Office of the Director of National Intelligence, the National Geospatial-Intelligence Agency, the National Reconnaissance Office, and the National Security Agency.

Classification (a form of access control for information) and information integrity (the reliability and veracity of an information source) have to go hand in hand to ensure the effectiveness of our security systems, and often times, it does not in practice. If we group these concepts together, they make sense falling in the same umbrella but the very nature of these concepts speak differently.

To handle classification is a matter of confidentiality. A popular computer security model that many agencies use is the Bell-LaPadula. If you have heard the terms Top Secret, Secret, Unclassified, etc., then you are thinking in terms of the Bell-LaPadula model. The entire premise of this model is to control the flow and confidentiality of the information, classification if you will, through a clearance level hierarchy. Someone from a higher clearance level can look at any information at or below their clearance level but can't look at anything above their clearance level. On the other hand, someone of a clearance level can write to their clearance level or higher. This is known as the read down and write up approach.

Furthermore, information integrity needs to be handled when communicating information. Information integrity follows the model of read up and write down, meaning that a person of a certain clearance level should only be able to write to people of his clearance level and below while only getting information from people of higher clearance levels. This is a direct contradiction of the Bell-LaPadula model. How is it possible to maintain confidentiality and integrity at the same time?

The solution we are researching is a declassification tool to assist people of different clearance levels to be able to communicate with each other without leaking information unique to their clearance level.

A Technical Solution

The Public Interest Declassification Board has described the new solution as a combination of human input and review and automation. The focus of the research project is focused on not just on the robustness of the automation systems, but also the cost and speed in computing over large volumes of data. Therefore, our choice in using the Perl computing language is an important one. Perl has been named as the Practical Extraction and Reporting Language by many and it does this in many respects. The language is quick and effective for searching through and modifying data. Not only is the use of the language free but it is linked to the Comprehensive Perl Archive Network, CPAN, which contains many open source and free language processing tools to speed up and further the research. Our solutions do

not just apply to the government sector alone. There are many use cases for businesses as well such as handling patent information, sensitive financial data, and private reports. The ability to relay this information without divulging the secrets unique to an information classification level allows for communication and oversight.

Approach

Our approach to this difficult computational problem requires two assumptions. Firstly, that the user in question has the ability and authority to declassify and modify the documents. Secondly, there is a unique identifying element that makes a set of information belong to a certain security level. Because of the nature of the task security needs to be effectively implemented in our solution. The proper encryption and verifications of documents and user access control will be implemented to the project.

On the premise that the assumptions have been met, the user is able to login, and select a lower clearance level to communicate with and the documents to declassify. The document is ran through the computer algorithms which summarizes the information and identifies the important content in the documents. The user is then able to view these summaries and add to the list of important content in the document and choose to either modify blocks of the document (the traditional approach) or chose a content topic to automatically generate entire documents based on that topic. The automatic generation is entirely editable and viewable by authorized users in that any grammar or information divulging can be caught and added to the important content list to better control the automated generation process. The automatic generator is fed the actual document and the summary such that the word choice and style are influenced by the document. Thus no additional information, words, or grammatical patterns are generated.

The computer will be able to automatically generate documents and use changes made by the user to assist the algorithms to essentially learn how to do better the generation and modification process. Allowing the user to guide the system lowers the burden of computational error and allowing the computer to automatically generate lowers the user burden of proper standardization.

Benefits

Because of the dual approach, we can improve the consistency and accuracy of declassified information allowing for a more immediate implementation of technological innovation.

One of the biggest problems in the previous declassification system was that it allowed to confidentiality but no system of integrity to allow people to get accurate and useful data. This declassification system is not just merely a declassification tool, but allows for the communication between different clearances to allow for more transparency but at the same time, still valuing the secrecy of certain information. While some information should always be hidden from the public, the information content list helps us to remove that burden. Not only can we modify the list, but the list can provide a standard on what types of information and what about the information can be released and communicated. Hence, we have a better chance at remaining informed and secure under this model.

The second problem that is addressed the administrative burden and oversight needed to declassify information. If we have automatic generation and search tools built for declassification reviewers, we allow them to focus almost exclusively on the quality of declassification, the reasons for declassifying a document, and what to declassify. This allows for administration in different agencies to thoroughly investigate and decide whether a document should be declassified entirely, in part, or not at all. Declassification is not just a static process where information is released after a time period. It is dynamic and requires careful consideration on the value of the information at hand. Automating the declassification process allows agencies to focus the budget on protection of data and valuation of the data.

Thirdly, the tools allows for the auditing of human inputs to generate performance reviews on both the people and systems. The ability to generate data in the age of analytics is important to evaluate the effectiveness of new software and if we are properly using the software at all. Documentation on how users are declassifying information is two-fold. One, we can develop better declassification standards and information communication between different clearance levels. As of now, it is currently illegal to view classified information without the proper clearance but, the bigger question is, is all the information in the document that essential for national security. This allows for a larger forum and debate on the valuation of information and picking out what can be communicated and what cannot. This sets the tone for what should be classified as the more we evaluate and declassify, we can measure and identify content areas that are not that disruptive to national security when released. With that knowledge, citizens and employees of different agencies can be ensured a stream of accurate information without revealing things that need to be kept secret.

Conclusion

In conclusion, government agencies are not efficient in handling the declassification of information in terms of man power and spending costs. Technology can be implemented in conjunction with user review to create strong, self-dependent ecosystem of classification and declassification standards. The biggest threat to our current classified information is the leakage of that information. Currently, we spend significantly more money protecting our information than anything else. The portal for secure and protect communication of classified information by lowering the clearance level of the data allows for more transparency of information but also maintains secrecy of information content. If much of the classified information content is not that important and can be lowered to different clearance levels, then we effectively decrease the size of the data we have to protect and lower the cost to do so .

Current solutions to the declassification problem is being researched but many tools are available for use to manipulate and assist with finding the solution. The use of Perl allows for a large network for free tools to use and contribute to identifying a solution. The approach we have is based off user and automation. The solutions they bring vastly improve our classification standards, help us measure growth in the field, and frees up administrative burdens placed on many agencies.

While no solution can be perfect and we can still be suspect to flaws in system design, approach, and security, our solution helps create an environment where criteria and regulations are set in place when any one part fails. If automation is not as successful, we have user review. If declassification reviewers are not consistent, we have automation and measures put in place to assist with that. Automation and language processing tools are technologies that improve with use and time. With funding and efforts in the field, we can only improve the approach and solutions we have created.