Avoiding Hidden Secrets in Search Results

Information leakage can have dire consequences

Simon Wiseman, Deep Secure

Sharing Secrets

When sharing information using email, the recipient receives the information the sender chooses to give. Information can also be shared by placing documents into repositories and allowing others to search for relevant documents and retrieve them. Access to documents in the repository can be controlled so users can only view the information held in documents they have been given permission to access.

Information sharing can be more effective if access to documents is controlled at the paragraph level, rather than in terms of complete documents. Sensitive documents inevitably contain a lot of information that can be shared more freely and widely, and paragraph level controls allow this to happen. Typically, when a document is accessed, the sensitive parts of it that cannot be released to the user are removed – a process called redaction.

Because search results reveal something about the documents they identify, users can glean information from the documents without having to open them. There is no easy solution, yet it is possible to construct systems so this kind of subtle information leakage does not occur, allowing the benefits of fine grain information sharing to be realised without compromising sensitive information.

Learning from What You Are Not Told

Suppose you have a repository containing all your documents, including some sensitive ones, which are indexed by a search engine. A simple way of securing this is to place a guard between the users and the repository and have it reject attempts by users to retrieve documents they do not have permission to access.

For example, suppose an aircraft's range is sensitive information, so the document giving the aircraft's specification is marked "SECRET." A user who does not have a SECRET clearance cannot access the document but might perform a search that lists the document in its results. The user can learn from the search results that the document contains what they are looking for, even though they have not accessed the document. This is because the search engine looked in the document to see if it matched the user's search query and effectively tells the user that their search terms are present in the document by returning it in the search results.

The search engine does not directly reveal information to the user, but it allows the user to confirm a guess. So the user can play "twenty questions." For example, the user might know the aircraft's range is about 1400km and so they just have to guess at the values around that point until the search engine returns a result (Fig 1). This might be tedious for a user but it can be readily automated and there are ways in which the number of tries can be reduced.

And that is how the search engine lets the cat out of the bag. The solution is to filter the search results so they only mention accessible documents, then if the search engine identifies a document containing the search phrase the user is not told and so they learn nothing from the results. This is something that SharePoint can do based on document permissions.

Alternatively, a guard placed between the user and the search engine can perform this filtering (Fig. 2). It accesses the repository to check each document in the search results to see if the user can access them, and removes them from the results if not.

If paragraph level labeling is used in documents, a guard placed between the user and the repository can redact sensitive paragraphs so the user fetching the document does not see them. So instead of hiding sensitive documents reported in the search results, the user is given a full list but if they access a document containing sensitive information, the relevant paragraphs are removed.

Unfortunately, this doesn't work, because the user learns something about the document even before they open it – they know its original contents match the search query even if the redacted version they can open does not contain the data (Fig 3). Thus,

the fine-grained sharing introduced by paragraph marking does not protect the information content of the sensitive paragraphs.

Fine Grain Sharing and Secure Searching

The only complete remedy for the problems that arise with fine grain information sharing is to prevent data the user cannot access being indexed by the search engine they use. Sensitive paragraphs that cannot be revealed to the user will then never match the query and so the results will not reveal information about that data to the user.

This can be achieved by putting a Guard that redacts documents between the search engine and the document repository so the search engine only gets to index redacted data. This solution results in documents being redacted when the search engine indexes them and when users retrieve them. The inefficiencies of this can be avoided by providing each user community with its own repository as well as a search engine. Then, as documents are published, they are copied to the various communities through a Guard that redacts them (Fig 4). This means documents are only redacted once, each time they are updated.

To prove the effectiveness of this approach, Deep-Secure Ltd, specialist providers of software-based products, and FoundationIT, strategic infrastructure platforms specialists, have joined forces on a joint project that aims to demonstrate how information management can to work in a federated cross-domain environment.

FoundationIT's technology manages the replication of documents and data in a heterogeneous environment, while Deep-Secure's Guards protect the network boundaries, support secure communications and enforce content security policies on shared data including label-based redaction.

The project culminates in a demonstration of the capability, using the defence supply chain as an example.

Fig 1: confirming a guess about sensitive facts

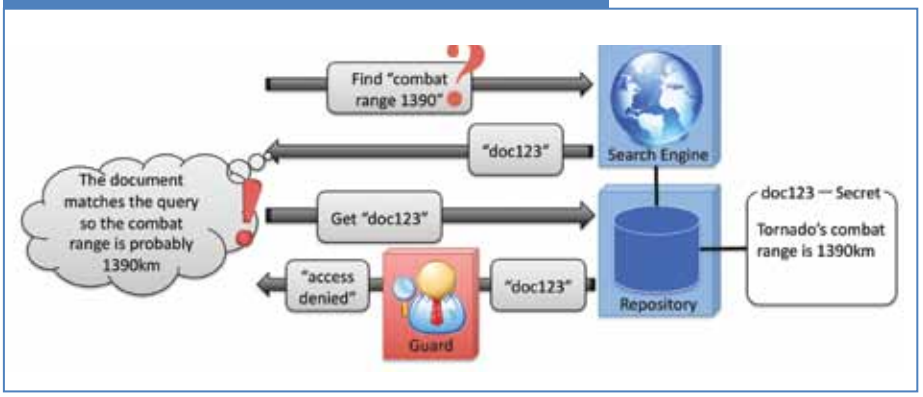

Fig 2: a Guard can filter sensitive search results

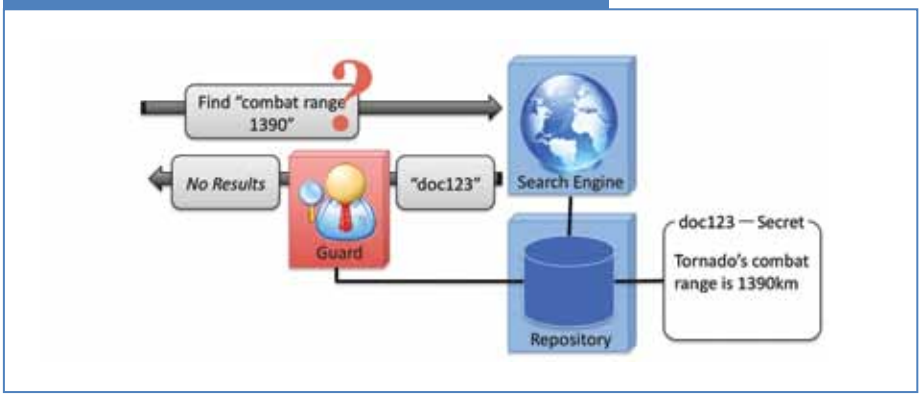

Fig 3: a Guard redacting retrieved documents

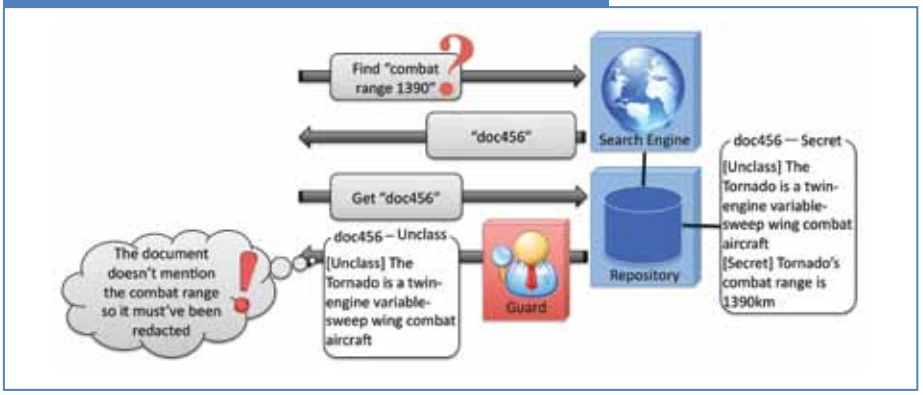

Fig 4: a Guard redacting copies of shared documents