Beratung

Deduplication Hidden Downsides

06. Januar 2015, von Irène Wilson

Introduction

Deduplication is often used in document reviews to reduce the amount of data to be searched or reviewed. When multiple files have identical content, only one version is kept. Hash values such as MD5 are used to determine whether the files are identical and therefore duplicates. Another approach to prevent reviewing the same file twice is the automatic propagation of coding decisions to all duplicate files.

In both cases, some file details are ignored when assessing the similarity of two files. This article will disclose the pitfalls of deduplication and propagation in respect to this discarded information.

Identical, to what extent?

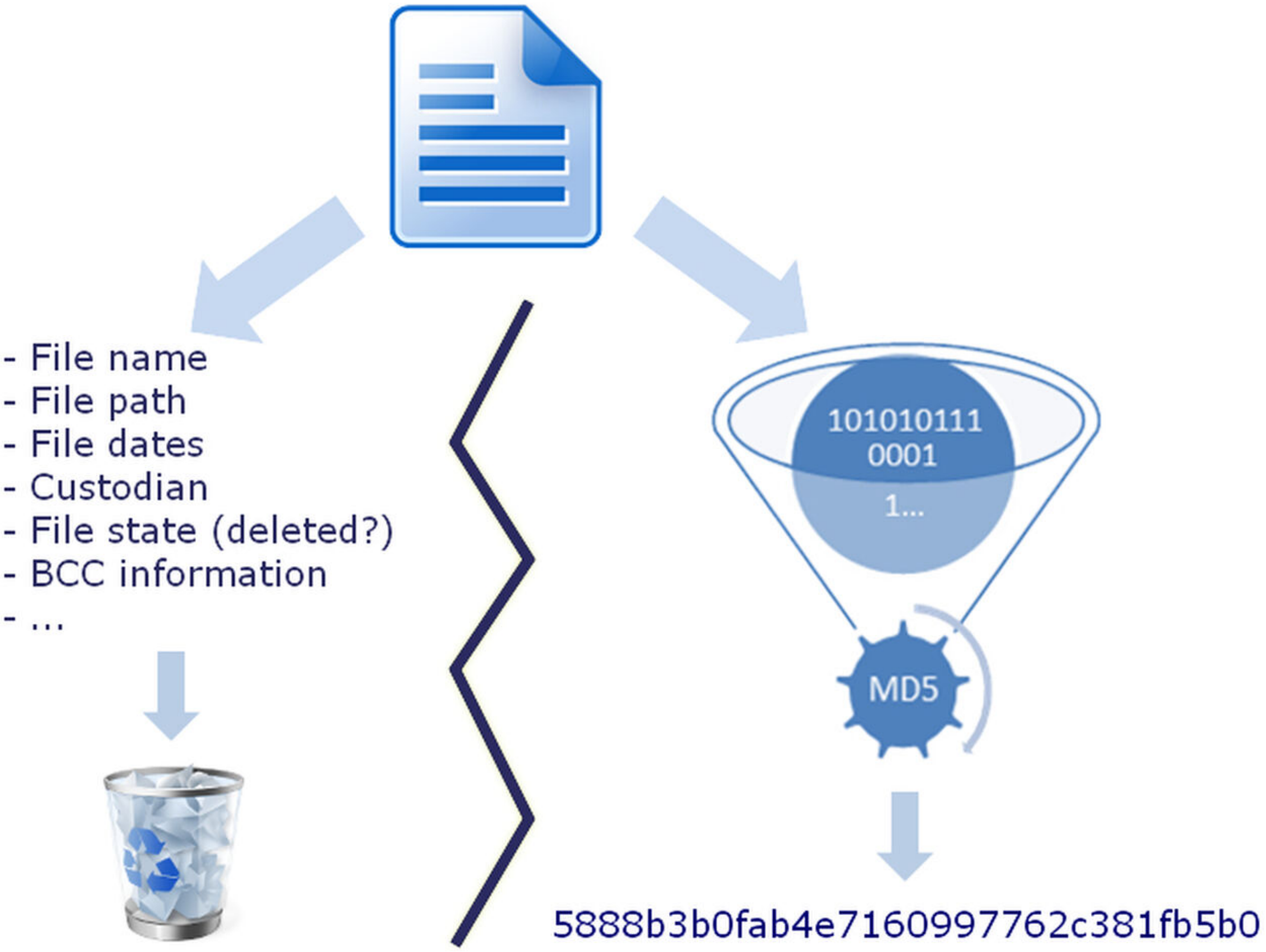
Identifying duplicates

To determine whether two files are identical, each file's binary stream is hashed with an algorithm to get a digest value. This value is then considered as the file's fingerprint and the two files are considered identical if their digest values are the same.

The most common algorithms used for this purpose are MD5, SHA-1 and SHA-256. With all these algorithms, a single bit difference will imply a very different digest value. This method is so thorough that a difference of encoding, compression or format will impact the digest value. Therefore, two documents might contain the same text but remain different from a digest point of view.

While this precision is appreciated from a forensic standpoint, it causes some issues when it comes to identifying two emails as having the same content. For example, email headers vary depending on the servers they go through. The application used to read or collect the emails can also impact their formatting. Also the BCC field is not present in all recipients copies. Therefore, the "same" email will get different digest values whether it comes from the sender's or the different recipients' mailboxes. Furthermore, the same email collected from the same email account, but once through a mobile device and once through a laptop, could be considered completely different depending on the client used to access it and/or the export format. Archiving and backup systems often change the email's binary content, resulting in different hashes too. Because of all these issues, a lot of eDiscovery tools customise the email digests to overcome these limitations when identifying duplicates. This customization usually ignores a lot of header information and sometimes also formats the email body.

Discarded information

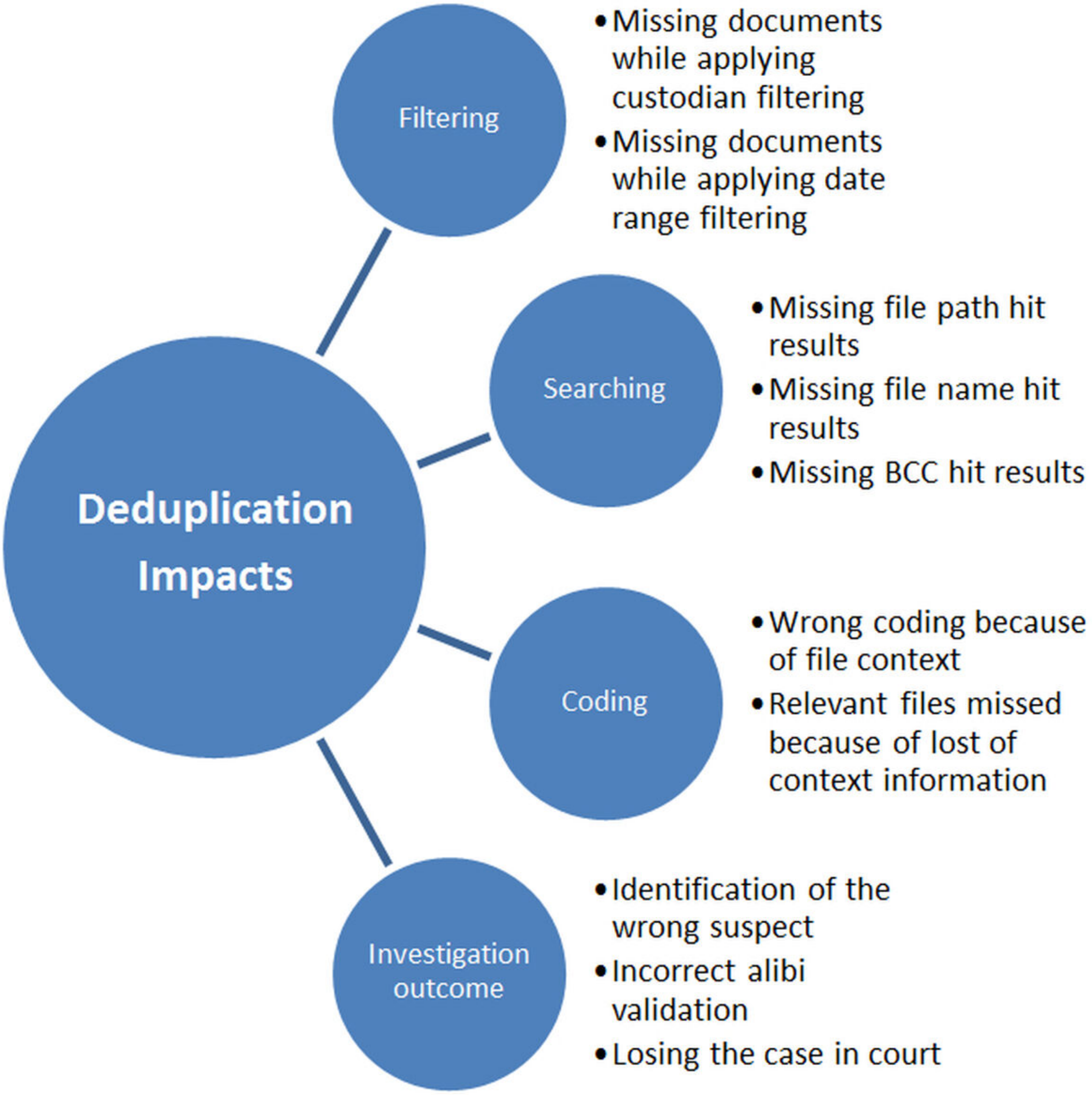


The normal hashing process takes the full binary stream into account, but completely ignores the file's context. This means the same file but with different paths, names, or dates, will get exactly the same hash. Basically, name, path and dates are not taken into account when computing the digest value. The system dates are the creation, last modified, last accessed and entry last modified dates. Some other attributes, which don't have an impact on the digest value either, are the custodian to whom the file belongs, in which directory it was stored, whether it was deleted or not, when it was deleted, and the source of the data.

When it comes to emails, the customisation of the hashing also ignores the BCC and sent/received date. Other technical details from the header are also discarded.

Deduplication impacts

Deduplication can have nasty effects at different phases of the eDiscovery process.


Deduplication Impacts
Filtering
• Missing documents while applying custodian filtering
• Missing documents while applying date range filtering
Searching
• Missing file path hit results
• Missing file name hit results
• Missing BCC hit results
Coding
• Wrong coding because of file context
• Relevant files missed because of lost of context information
Investigation outcome
• Identification of the wrong suspect
• Incorrect alibi validation
• Losing the case in court

Filtering

Data filtering should always occur before deduplication. However, it often happens that the scope of the project is originally unclear; data gets reduced through deduplication and the client later decides to apply further filtering to decrease the number of files to search or review. In practice, this process can lead to excluded files that should actually be in scope.

In terms of custodian filtering, there are two major risks when applying deduplication beforehand. First, if you apply a global deduplication, you will keep one version of each file independently of its owner. Therefore, a later applied custodian filter can result in ignored files, because the version kept belongs to another custodian, even though a duplicate version actually exists for the person in scope. This means applying global deduplication and reviewing each custodian in isolation are intuitively incompatible. However, there are tricks to overcome this challenge, but it should nevertheless not be overlooked. The second pitfall though has a wider impact. To have a more complete dataset for a given custodian, you might want to not only rely on the file owner, but also search the file properties as well as email sender and recipients for that custodian's name. The exclusion of the BCC field in the hashing process can have a considerable impact. Its main consequence is that different versions of an otherwise identical email don't necessarily have the same recipients. Therefore, by applying deduplication before searching the recipient fields for a specific person, you can miss relevant emails.

Another type of filtering that is influenced by prior deduplication relates to date range. Quite often, after seeing the number of files responsive to keywords, clients decide to limit the scope of the investigation to the most pertinent years. Deduplication does have an impact on this process. Regarding emails, as mentioned above, dates are usually ignored when comparing files. However, the dates of different versions of a same email are likely to be contained in a very short time frame anyway, so the impact should be limited. The situation is different with loose files though. First you might wonder which date to take into account for your time range: creation date? last modified date? last accessed date? What about deletion date? But more important is that different versions of a file are likely to have different dates. For example, copying a file into a new location usually resets its creation date. Depending on which version of the file the deduplication process kept, your file can randomly end up in your final dataset or be excluded from it.

Searching

When running keyword searches, the impact of previous deduplication is as follows:

- **File path:** Sometimes, all files related to the same topic are gathered in the same directory. If your keyword is part of the file path but is not included in the file itself, then deduplication can have an impact. Different versions of the same file can have different paths. The version kept by the deduplication process maybe doesn't contain your keyword in its path while duplicates might.
- **File name:** The same challenge applies to the file name. This doesn't impact emails but different versions of a same loose file can have very different names. For example a file downloaded temporarily from the Internet will most likely end up with a random name while the version saved intentionally by the user probably has a completely different name that's more meaningful.

- **BCC recipients:** As highlighted earlier, when searching for a person, the exclusion of the BCC field from the hashing process can be critical. Your search results can be incomplete because of it.

Coding

While a coding decision is usually mostly guided by the file content, its context or properties can sometimes be extremely relevant. For example in data leakage cases it's critical to testify who had the data and where it was stored. Data theft or blackmailing can have the same type of focus. Here the file path and custodian information are highly relevant as well.

The loss of context caused by deduplication can also strongly impact an investigation. The most known side of this issue is with the same file being attached to two different emails, or saved as a loose file without any context. Its relevance probably won't be assessed the same way, which is why deduplication should typically be applied to entire families (considering each email with its attachments as one entity) instead of on individual files. Giving the reviewer a complete overview of a file's context is usually relevant to their decision accuracy.

Also coding propagation can have an impact on the pertinence of your review results. Automatically applying a coding to all duplicate files only makes sense if the decision is based on the file's content and is completely independent of its context. When tagging an email as relevant because of one of its attachments, propagation to duplicates will negatively impact the quality of your results.

Another issue with lacking context regards the folder structure. When finding an interesting file, you might want to check other files saved in the same directory. However, as highlighted earlier, if the version kept for one of these files has a different path, you won't find it when looking under this path and you won't have a complete picture of the folder content.

Investigation Outcome

In the end, all of these small pitfalls can have a measurable impact on the quality and precision of the results of your investigation. Even without considering the files ignored with filtering, the keyword responsive files missed when searching, or the relevant files tagged incorrectly as non-relevant during review, deduplication can still have a direct impact on the interpretation of your investigation results.

Identification of a suspect can be directly impacted by a file's metadata information. You might rely on who had access to the file in question when identifying a suspect, forgetting about other locations where duplicates are stored. File dates are particularly pertinent when considering topics such as alibis. In some cases, checking who accessed the computer, at what time, and the accuracy of that information is also crucial.

Finally, when a case goes to court, if the other party has taken a different approach and found details about the file that are different from yours – as a consequence of inappropriate deduplication methods used on your side – it can well make the difference between winning and losing a case.

Conclusion

The phrase "The devil is in the details" succinctly summarises the issues described in this article. While deduplication is of great help when it comes to handling huge amounts of data, it is necessary to plan your approach thoroughly, and be aware of the consequences it might have. There is no perfect solution and budget and time limitations should also be taken into account. However, it is important to know about the risks and inform your client accordingly. Deduplication can have a particularly large impact depending on the type of investigation. Good communication with the client will allow you to foresee and prevent these kinds of problems. In the end, the approach taken should be tailored to each case.

Irène Wilson

Irene Wilson ist auf digitale Forensik und eDiscovery spezialisiert und hat im Laufe ihrer langjährigen Erfahrung für Kunden aus einer Vielzahl von Branchen in ganz Europa gearbeitet. Zu ihren Qualifikationen gehören die renommierten Master-Titel für Nuix Workstation und Nuix Discover.