Connaissance technique

Déduplication avancée, Identiques au-delà des empreintes digitales

05. mai 2020, by Irène Wilson

Introduction

La technologie est utilisée pour soutenir les investigations et les projets de revues de documents de plusieurs façons, notamment en permettant de gagner du temps et de l'argent grâce à la réduction des données. Si l'élimination basique des doublons par comparaison de la valeur du hachage est bien acceptée, elle présente des limites que des outils puissants et des flux de travail créatifs peuvent surmonter. Voici une introduction à certaines de ces techniques.

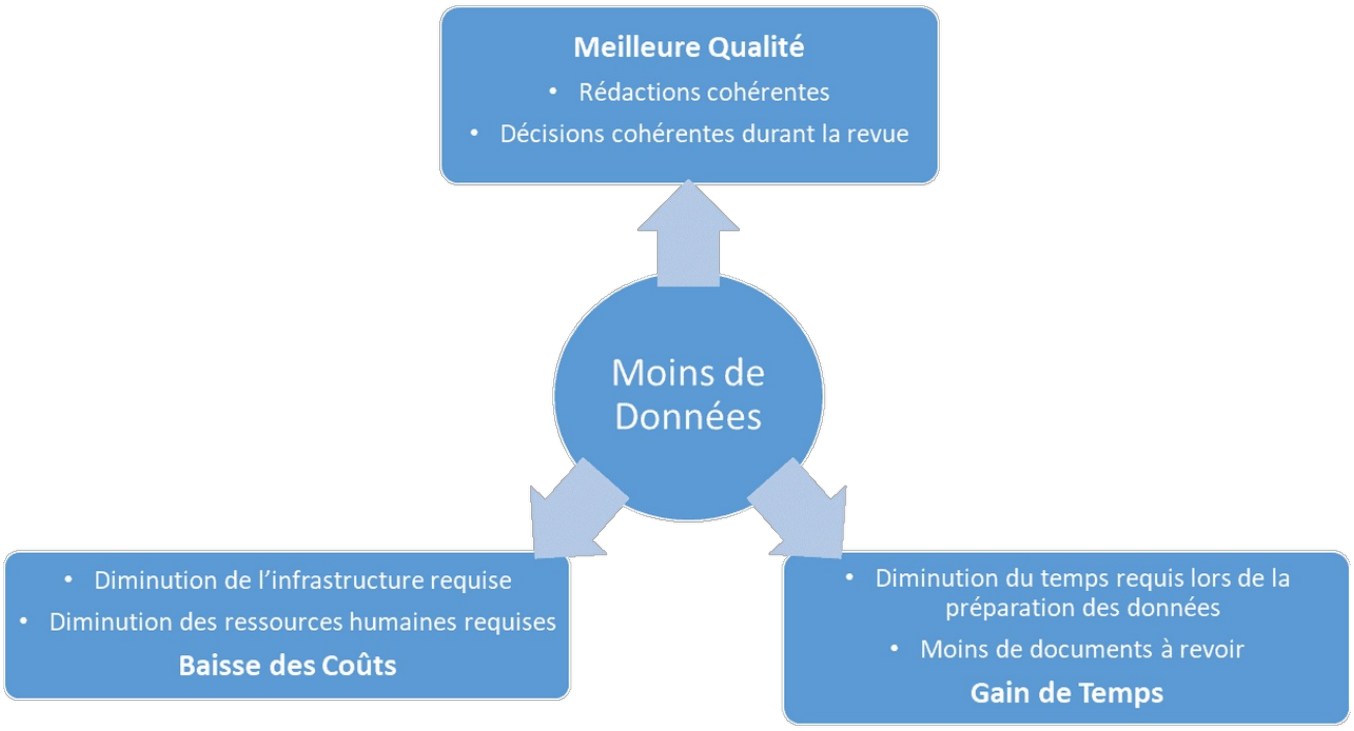
Pertinence de la réduction des données

Avant de débattre des avantages et des inconvénients des différentes approches de réduction de données, rappelons d'abord pourquoi nous les envisageons. Un volume de données plus important nécessite plus de stockage, un matériel plus puissant, plus de temps de

préparation des données et plus de temps lors de la revue, avec le risque de manquer des informations pertinentes parce qu'elles sont perdues dans la masse. Le temps, c'est de l'argent, et donc avoir plus de données implique un double impact sur les coûts en termes de matériel et de temps de traitement tant pour les personnes que pour les machines.

Cela ne tient pas compte d'autres considérations telles que les délais. Quel est l'impact du non-respect d'une date limite ? Une grosse amende très probablement, avec beaucoup de zéros. La réduction des données vous permet de réduire le temps et les coûts de votre investigation.

Il y a également un effet positif sur la qualité. Les réviseurs et enquêteurs sont humains et prennent donc des décisions subjectives. Le fait d'avoir des doublons dans vos données entraîne souvent des décisions de revue incohérentes, où les doublons du même élément peuvent être codés différemment selon la personne qui a effectué la revue. Cela devient encore plus délicat lorsqu'il s'agit de rédactions et de pseudonymisation. La réduction des données améliorera la cohérence et la qualité de la revue.



Faiblesses de la déduplication traditionnelle basée sur le hachage

Il est courant d'utiliser les valeurs de hachage pour supprimer les doublons dans les projets

d'eDiscovery. Ces valeurs de hachage sont calculées en appliquant un algorithme sur le contenu binaire des fichiers. Les hash MD5 sont un résultat courant de ce processus, et sont ensuite utilisées pour identifier et exclure les doublons. Ces valeurs sont souvent appelées « empreintes numériques ».

Saviez-vous que des jumeaux identiques ont des empreintes digitales différentes ? De la même manière, des fichiers qui, du point de vue de l'examinateur, sont identiques, peuvent avoir des valeurs de hachage différentes. Deux facteurs expliquent ces différences : les informations qui ne sont pas visibles pour l'utilisateur, comme les propriétés ou les informations techniques stockées dans le fichier mais cachées aux utilisateurs non techniques, et le formatage des informations affichées à l'utilisateur, comme la compression ou le formatage du texte. En raison de ces faiblesses, une simple déduplication basée sur la valeur de hachage laisse souvent un goût amer à l'investigateur, qui a l'impression d'effectuer un travail redondant et que le processus de déduplication n'a pas fonctionné.

Solutions commerciales courantes

Valeurs de hachage du « contenu purifié »

Plusieurs éditeurs de logiciels d'eDiscovery ont développé leurs propres solutions pour assurer une déduplication efficace lorsqu'il s'agit de courriers électroniques. Les messages électroniques sont extrêmement délicats en raison de leur en-tête, qui est impacté par le chemin que le message emprunte de l'expéditeur au destinataire, et du formatage que le logiciel peut appliquer au corps de l'email. Pour relever ce défi, des outils tels que Nuix enlève les informations non pertinentes des en-têtes et nettoient le corps du message avant d'appliquer un algorithme de hachage. Cela permet d'identifier comme doublon un message pris dans la boîte aux lettres de l'expéditeur et le même message dans la boîte aux lettres du destinataire. Bien que cette solution soit reproductible, elle reste propriétaire et ne permet donc pas de comparaison entre différents outils.

Déduplication basée sur le texte

La solution décrite ci-dessus est spécifique aux courriels, mais elle ne permet pas de surmonter les différents formats que peut prendre l'information. Un courriel par exemple peut également exister sous forme de PDF ou d'image. Pour les identifier, la meilleure approche est d'ignorer l'os et de se concentrer sur la substantifique moelle : l'information elle-même. Le calcul d'un MD5 de texte non formaté extrait d'un fichier permet de comparer

des informations stockées dans différents formats. Les shingles et les near-duplicates (ou « presque-doublons ») poussent ce concept encore plus loin en permettant de choisir un degré de similarité dans le texte, au lieu d'une correspondance exacte. Cela peut aider à contourner l'imprécision de l'OCR, mais cela va également regrouper des documents basés sur le même modèle (comme un modèle de courriel dont seules certaines valeurs sont mises à jour avant l'envoi), ou différentes versions d'un même document. Il est important de clarifier l'objectif de la déduplication et de vérifier que cette méthode donne les résultats requis. Supprimer des documents basés sur un même modèle parce que votre outil et approche les signalent comme des doublons pourrait faire beaucoup de dégâts selon les cas.

Solution pour les images

En ce qui concerne les images, il existe un problème connu des officiers de police: le MD5 ne permet pas une déduplication efficace, car chaque image apparaît plusieurs fois sur un système, dans des tailles, des formats et des niveaux de compression différents. PhotoDNA, un algorithme très intelligent pour dédupliquer les images, est maintenant implémenté par la plupart des outils du marché. Il permet d'identifier les images en double au-delà de ces obstacles. Toutefois son utilisation est limitée aux forces de l'ordre.

Binary fuzziness ou « flou binaire »

Un autre type de fichier qui pose problème est le malware, où le code est obscurci, ou légèrement modifié, de sorte que les données binaires apparaissent différemment tandis que la fonctionnalité reste la même. L'algorithme SSDeep évalue la similarité au niveau binaire pour contourner cette limitation dans l'identification des doublons.

Autres problèmes concrets concernant les courriers électroniques

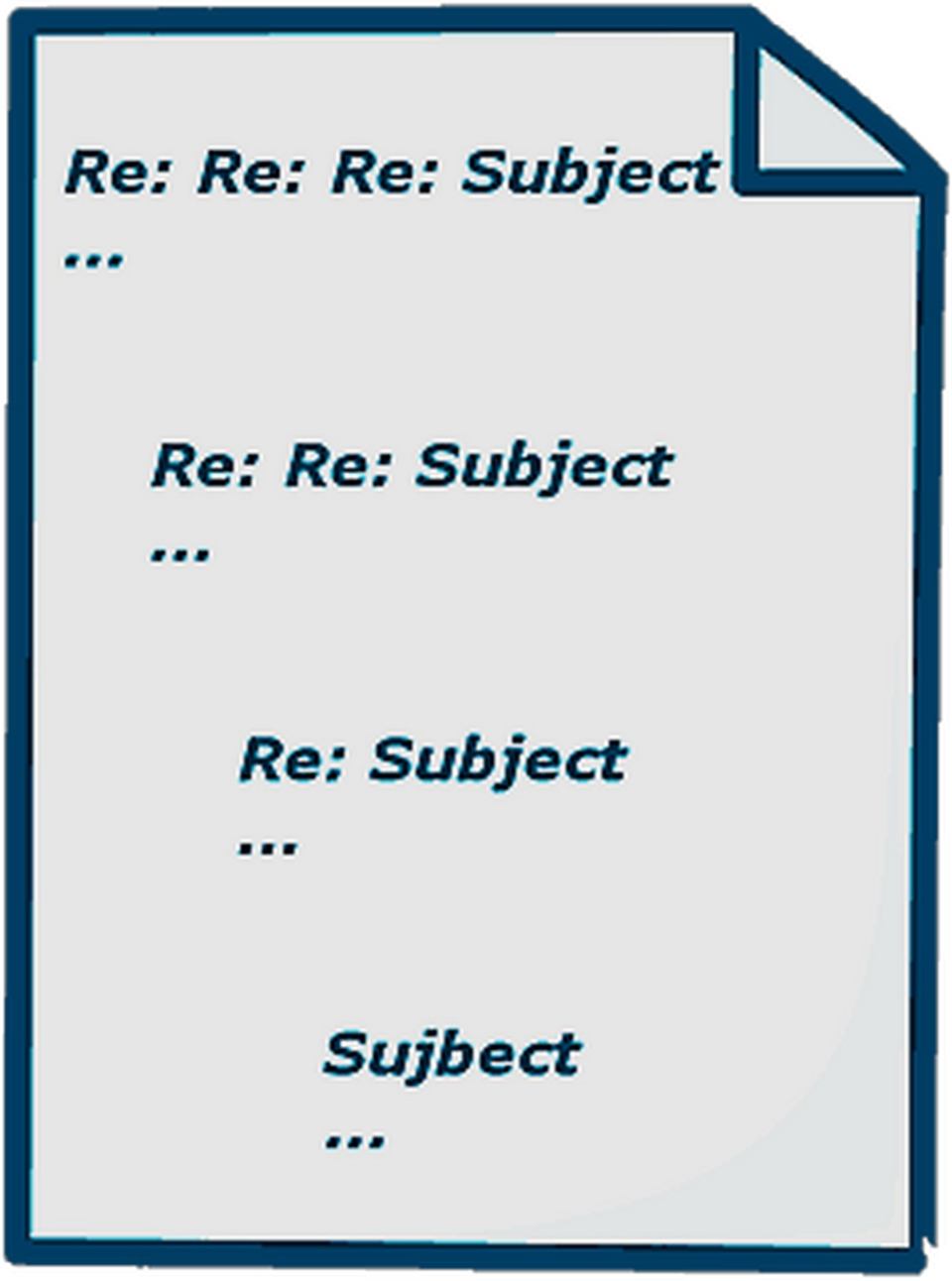
Toutes les technologies et tous les algorithmes mentionnés ci-dessus contribuent grandement à humaniser le processus de déduplication, en identifiant les doublons selon une interprétation plus proche de celle d'un examinateur ou enquêteur, avec une approche plus souple et pragmatique. Même avec ces améliorations, le résultat n'est souvent pas à la hauteur en matière de revue de documents. L'expérience montre que les courriels présentent plus de difficultés que tout autre type de fichier.

En plus des multiples chemins qu'un courriel peut emprunter, le système sur lequel l'email est stocké peut engendrer une valeur de hachage différente. Les courriels sont parfois stockés dans des bases de données, où il est courant de supprimer les pièces jointes et de les conserver séparément, afin de pouvoir les dédupliquer pour économiser l'espace de stockage. Lorsque l'on collecte le même courriel de différentes sources, les outils forensique et de eDiscovery font de leur mieux pour reconstruire le message aussi fidèlement que possible à l'original, mais la route est semée d'embûches qui trompent même les valeurs de hachage de « contenu purifié ». Les adresses électroniques telles que présentées et formatées à l'interne en opposition à leur format externe, l'ordre des pièces jointes et les avertissements sur les messages archivés sont autant de différences mineures que les outils actuels ne peuvent pas gérer seuls. C'est alors que l'enquêteur forensique peut démontrer la profondeur de son art en sortant des chemins battus pour surmonter ces obstacles de manière reproductible et contrôlée par des flux de travail personnalisés, des outils développés sur-mesure et des approches innovantes. Le résultat semble miraculeux, en particulier lorsque les données ont été collectées à partir de multiples sources pour éviter toute lacune, et le niveau de redondance qui en résulte est très élevé. L'approche idéale tiendra également compte de la qualité des différentes sources disponibles, et donnera la priorité à la meilleure si plusieurs sources s'offrent à elle.

Conversation électronique

Poussant le concept de « doublon » plus loin que la plupart des forensiciens ne seraient prêts à accepter, les examinateurs se plaignent souvent de regarder plusieurs messages d'un même fils de conversation. La plupart des outils disponibles sur le marché permettent aujourd'hui d'analyser les fils de discussion. Le concept est simple : conserver le courriel le plus inclusif de la chaîne, ainsi que tout message avec un contenu supplémentaire (y compris les pièces jointes). La mise en œuvre est un peu plus délicate et présente plusieurs pièges. Il est donc important de bien connaître son outil et de comprendre les détails du processus technique.

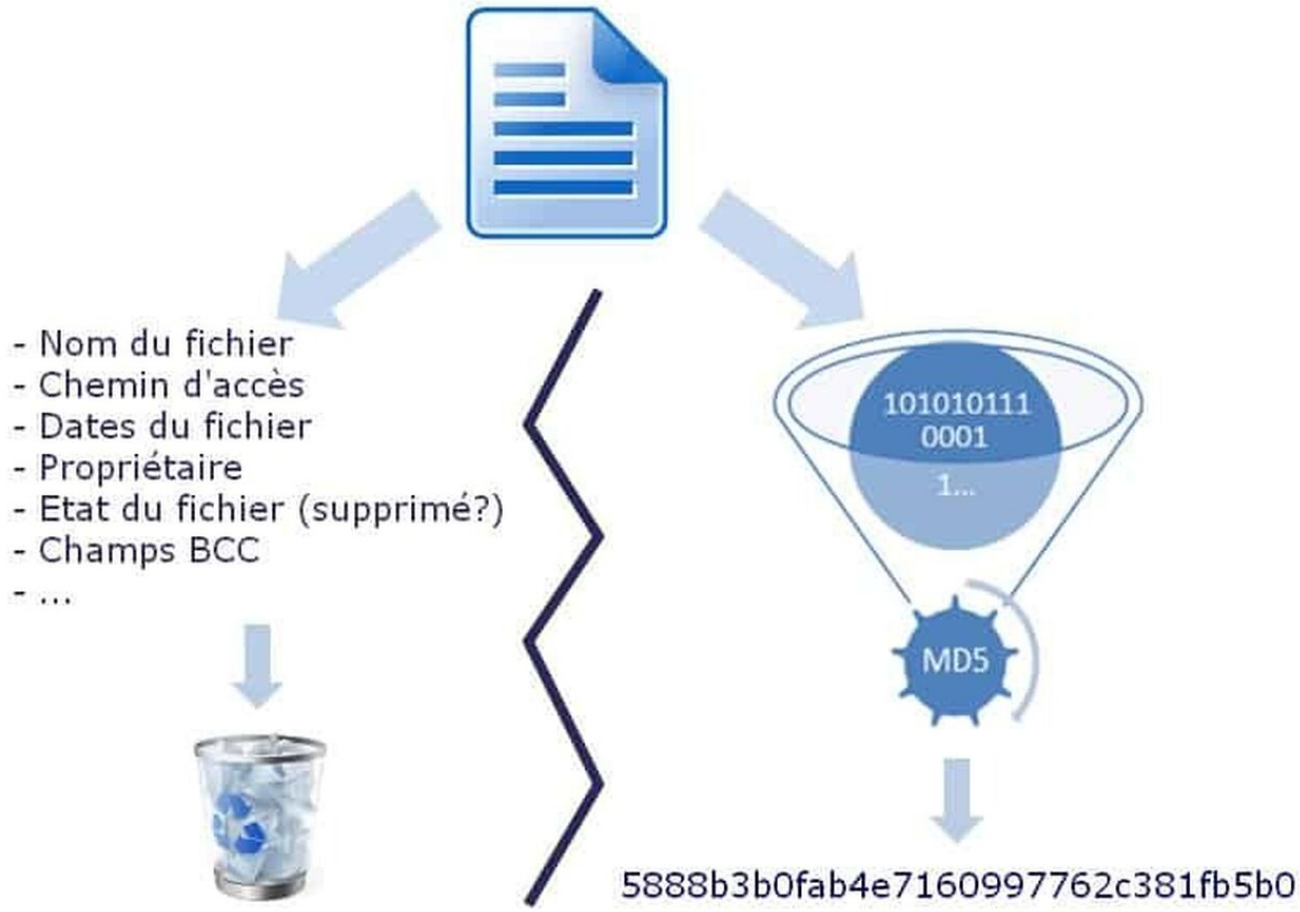
Poussant le concept de « doublon » plus loin que la plupart des forensiciens ne seraient prêts à accepter, les examinateurs se plaignent souvent de regarder plusieurs messages d'un même fils de conversation. La plupart des outils disponibles sur le marché permettent aujourd'hui d'analyser les fils de discussion. Le concept est simple : conserver le courriel le plus inclusif de la chaîne, ainsi que tout message avec un contenu supplémentaire (y compris les pièces jointes). La mise en œuvre est un peu plus délicate et présente plusieurs pièges. Il est donc important de bien connaître son outil et de comprendre les détails du processus technique.



Risques liés à la réduction des données

La réduction des données est nécessaire, il n'y a aucun doute à ce sujet. Toutefois, elle peut avoir des conséquences négatives et doit donc être envisagée avec soin et mise en œuvre correctement. La spécificité de l'affaire, par exemple les informations contextuelles telles que le chemin d'accès à un fichier qui ont une incidence sur la pertinence d'un document, doivent être partagés avec l'équipe technique, afin qu'elle puisse ajuster l'approche et éviter de tomber dans les pièges de la déduplication. Même si nous nous limitons à la déduplication standard, les informations contextuelles sont ignorées dans le calcul de la valeur de hachage, alors qu'elles pourraient avoir une pertinence dans l'enquête. Notre article précédent, «

Implications méconnues de la Déduplication », met en lumière les zones d'ombres d'un tel processus.



Voici quelques exemples pratiques qui montrent que le risque est réel :

Dans un cas de fuite de données, les données après réduction sont examinées pour trouver l'information ayant fait l'objet de la fuite. Le fichier en question est trouvé sur un disque partagé, dans un emplacement acceptable selon le flux de travail du client pour ce type d'information. En revenant à l'ensemble des données, des copies de ce même fichier sont découvertes dans le dossier d'un utilisateur, pointant directement à un suspect potentiel.

En réalisant le volume de données après réduction, même après avoir limité les documents à examiner en fonction de mots-clés, le client décide de limiter davantage en restreignant à une période spécifique. Le cas contenant des fichiers individuels, une partie des données pertinentes pour la sélection selon la période choisie serait négligée si le filtrage des dates n'avait pas lieu en amont sur l'ensemble des données.

Un enquêteur a trouvé un contenu illégal dans une image récupérée dans la zone non allouée du disque dur d'un suspect. Les preuves trouvées dans un tel endroit sont difficiles à

présenter au tribunal, car elles manquent de contexte et il est impossible de prouver le téléchargement, la sauvegarde ou la copie du fichier volontaire par l'utilisateur. Toutefois, si une copie de ce même fichier est disponible dans le répertoire « Images » ou « Téléchargements » du profil de l'utilisateur, l'impact de la preuve est alors complètement différent.

Un autre risque est celui des faux positifs dans l'identification des doublons. Si ce risque est extrêmement faible avec une approche de déduplication standard, il est plus élevé avec des fonctionnalités plus avancées. Comme mentionné ci-dessus, une déduplication basée sur le texte, permettant une certaine flexibilité, pourrait identifier différentes versions d'un même formulaire comme étant des doublons.

La valeur de hachage du « contenu purifié » a aussi ses pièges. Par exemple, Nuix applique cette approche spécifique aux éléments de calendrier. Il est cependant courant d'avoir plusieurs entrées de calendrier avec le même auteur, les mêmes destinataires, le même sujet et le même corps de texte, mais se produisant à des dates différentes. Qui n'a jamais inscrit de cette façon dans son agenda des rendez-vous médicaux récurrents ? Considérer ces différentes entrées comme des doublons et les supprimer pourrait conduire à négliger un éventuel alibi.

Conclusion

La déduplication est bien acceptée de nos jours, et même si les méthodes les plus simples sont souvent insuffisantes les professionnels appliquent des approches plus avancées depuis déjà un certain temps. Le fait que le client vous regarde comme un héros parce que vous avez pu réduire la quantité de données de 70% est toujours un sentiment agréable, mais si cela s'accompagne du risque de ne jamais trouver la preuve irréfutable, la gratitude ne durera pas. La communication entre les enquêteurs et le personnel technique est la clé d'une mise en œuvre sûre de ces technologies.

Cet article s'est concentré sur la réduction des données, mais la priorisation des données est l'étape suivante dans cette bataille entre les enquêteurs et l'augmentation des volumes de données. Que ce soit par le biais d'algorithmes étendant les décisions des examinateurs à un ensemble de données plus important, de l'entraînement d'un système pour évaluer avec précision les données restantes ou de la réévaluation de la pertinence des données restantes à la volée au fur et à mesure de la revue, les développeurs de logiciels et les experts en eDiscovery continuent de proposer des approches innovantes.

Irène Wilson

Irène Wilson est spécialisée dans l’investigation informatique et dans l'eDiscovery. Au cours de ses nombreuses années d'expérience, elle a travaillé pour des clients de nombreux secteurs différents dans toute l'Europe. Parmi ses nombreuses qualifications, figurent les prestigieux titres de master pour Nuix Workstation et Nuix Discover.

Swiss FTS | https://swiss-fts.com/fr | +41 43 266 78 50 | info@swiss-fts.com

https://swiss-fts.com/fr/blog/deduplication-avancee-identiques-au-dela-des-empreintes-digitales