MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE
LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

**Faculté des Sciences Exactes et de l'Informatique**
**Département de Mathématiques et d'Informatique**
**Filière : Informatique**

MEMOIRE DE FIN D'ETUDES
Pour l'Obtention du Diplôme de Master en Informatique
Option : **Ingénierie des Systèmes d'Information**

THEME :

# Anonymisation des données dans les journaux de requêtes

**Etudiant :** Hakim KHELIFA

**Encadrante :** Mme MIMI Anissa

Année Universitaire 2016/2017

## Résumé :

Les journaux de requêtes sont des fichiers contenant des données de recherche web. Ces journaux fournissent un large volume de données pour la communauté de chercheurs surtout qu'elles représentent l'expérience d'utilisateurs réels avec un système de recherche d'information. Mais la publication de ces données comporte un risque sur la vie privée des individus. Même en appliquant plusieurs mesures de filtrage, les données qui restent dans un journal de requêtes en l'occurrence la requête elle-même garde un risque de divulgation d'informations personnelles.

Notre travail consiste à proposer une solution d'anonymisation des données dans un journal de requêtes. Notre proposition c'est de généraliser certains types d'informations personnelles, c'est-à-dire les remplaçant avec des informations de sens plus général. Nous évitons de cette façon d'appauvrir le journal de requêtes en réduisant les informations qu'il contient tout en garantissant que les informations restantes ne puissent identifier directement les utilisateurs. Parmi les informations que nous avons choisi de traiter nous citons les noms et prénoms des personnes et les adresses (les noms de lieux).



***Mots clé :*** **Données personnelles, Journal de requêtes, Anonymisation, ré-identification, divulgation, généralisation.**

# Table des matières

# Liste des Figures :

# Liste des tables

## Liste des Abréviations

**AOL :** America Online

**RI** : le recherche d'information

**SRI** : Un système de recherche d'information

**IT :** information and technologie

**IP :** Internet Protocol

**CNIL :** Commission Nationale de l'Informatique et des Libertés

**URL :** Uniform Resource Locator

**OMBA:** office of Management and Budget American

# Introduction générale

Avec le développement d'internet et des technologies de l'information, et grâce à ces dernières, la recherche d'information est devenu aujourd'hui très efficace et bien organisée. Ceci est dû en grande partie à l'étude de son efficacité auprès des utilisateurs. En effet, de nombreuses recherche dans le domaine de la recherche d'information se basent sur la collecte puis l'analyse de grandes quantité d'information sur l'utilisateur, ses expériences avec les outils ou les systèmes de recherche d'information ainsi que sur l'environnement influençant ces expériences. Les chercheurs dans ce domaine veulent surtout répondre à la question : Que recherche l'utilisateur et qu'est ce qui l'intéresse vraiment ?
Ces études demandent le traitement de grandes quantités de données qui ne sont pas toujours matérielles ou logicielles, en fait elles peuvent aussi être des données concernant l'utilisateur lui-même, ce qu'on appelle des données personnelles : est-il jeune ou âgé ? est-ce une femme ou un homme ? de quelle région est-il ? ...etc.
En ce qui concerne les moteurs de recherche, l'outil de recherche d'information le plus utilisé par les internautes, la sauvegarde des données collectées se fait dans ce qu'on appelle un journal de requêtes (ou query log en anglais). Ce fichier, est dans certain cas partagé avec la communauté scientifique pour les besoins de la recherche. Mais ce qui était au départ lié à une bonne cause est devenu par la suite source de préoccupation autour de la protection de la vie privée des individus sur le web.

L'objectif de ce mémoire et est d'étudier la problématique d'anonymisation des utilisateurs dont les données sont sauvegardées dans un journal de requêtes afin de répondre à la question : Comment assurer une bonne balance entre l'utilité du journal de requêtes et la protection de l'anonymat des utilisateurs ?

Pour ce faire nous avons organisé notre mémoire en cinque chapitres :

**CHAPITRE I** Nous présentons dans ce chapitre une vue générale sur la recherche d'information ainsi que sur le journal de requêtes. Nous décrivons ensuite les risques liés à la publication ou le partage de journal de requêtes, nous présentons à cet effet l'exemple de cas réel de divulgation de l'identité d'un utilisateur dans l'affaire des données de recherche d'AOL.

**CHAPITRE II** Dans ce chapitre nous étudierons la problématique de divulgation des données personnelles sur le web, ce qui relève de la protection de la vie privée. Nous montrons le lien entre cette dernière et la sécurité informatique et la protection des informations. À la fin nous montrons qu'il n'est pas toujours facile de gérer le compromis entre la vie privée et d'autres besoins.

**CHAPITRE III** Dans ce chapitre nous faisons un tour d'horizon des solutions d'anonymisation proposées à ce jour et dans différents domaines. Nous expliquerons pourquoi il n'est pas facile d'appliquer les solutions proposées dans le domaine des bases de données par exemple aux données des journaux de requêtes d'où la motivation de notre projet à essayer de trouver une solution adaptée à ce nouveau contexte.

**CHAPITRE IV** Dans ce chapitre nous expliquons notre solution proposée, son principe, ses avantages et limites. Nous présentons aussi son implémentation.

# CHAPITRE I : Recherche d’information et les journaux de requêtes

## Introduction

La recherche d’information est une branche entre les domaines des sciences de l'information de la bibliothéconomie et l’informatique qui propose des solutions permettant de sélectionner des sources d'information ou documents répondant au mieux au besoin de l'utilisateur.

La recherche d’information aujourd'hui est une activité humaine faisant appel à un système qui fait intervenir d’un côté l’utilisateur et de l’autre côté un système de recherche d’information (SRI). Un système de recherche d’information SRI est un système qui gère une collection d’informations organisées sous forme d’une représentation intermédiaire. Dans la suite nous définissons quelques notions liées à la recherche d'information.

## I-1-La requête (Query)

Dans les cas les plus simple, lorsqu'il veut trouver des documents sur internet, l'utilisateur formule une question avec ce qu'on appelle des mots clés (key Word) et l'envoie à un moteur de recherche via son interface web. Cette question s'appelle requête web (search query ou web query en anglais). De l’autre côté, le SRI consulte son index afin de trouver tous les documents qui contiennent les mots clé présents dans la requête.

## I-2-La session de recherche

Les reformulations des requêtes et les documents consultés sont des éléments particulièrement utiles pour une meilleure compréhension du besoin d’information de l'utilisateur, de son évolution et de sa satisfaction au fil d’une recherche. Pour cette raison le moteur de recherche sauvegarde toutes ces informations dans une même session de recherche.

Selon **[13] [14] [15] [16] [17] [18] [19], La** session de recherche peut être définie comme suit : c'est l'ensembles des requêtes formulées par un utilisateur dans une période de temps. Une session peut ainsi contenir des éléments de différente nature : des activités/interactions (abonnement, achats, consultation). LA période de temps durant laquelle s'estompe une session de recherche n'est toutefois pas uniformisée. Ainsi, une session peut durer jusqu'à la déconnexion de l'utilisateur, en estimant une période d'inactivité ou s'étendre au-delà de plusieurs connexions.

La question de comment identifier un même utilisateur pour différentes connexions fait toujours l'objet des études (utilisation de cookies, IP statique, Analyse de l'environnement de travail, ...)

Les définitions de la notion de session proposées dans la littérature se distinguent donc à plusieurs niveaux.

**[1]**

## I-3- Le journal des requêtes (Query Log)

Si la requête est l'ensemble des mots clé formulés par l'utilisateur et la session de recherche est l'ensemble des requêtes envoyées par un utilisateur dans une période de temps, le journal de requêtes est quant à lui l'ensemble des sessions de recherche de plusieurs utilisateurs.
L'idée de construction de ce journal de requêtes a nécessité la mise en œuvre de plusieurs traitements afin de ne conserver que les informations plus fiables à partir du journal d'accès original. En particulier, les données ont été nettoyées et filtrées de manière à éliminer les informations inexploitables. Les journaux de requêtes sont des ressources de recherche essentielles pour les RI, en particulier pour le domaine de la recherche sur le Web. Avec cette technique les requêtes ont été regroupées par adresse IP et sont classées par ordre chronologique. Le journal de requêtes finalement obtenu comporte un identifiant pour chaque utilisateur correspondant à l'adresse IP anonymisé, la date et l'heure de soumission de chaque requête, ainsi que les requêtes soumises. **[1]**

| | | | | |
|---|---|---|---|---|
| 417749 | care packages | 02/03/2006 09:19 | 3 | http://www.awesomecarepackages.com |
| 417749 | care packages | 02/03/2006 09:19 | 8 | http://www.anysoldier.com |
| 417749 | movies for dogs | 02/03/2006 09:24 | | |
| 417749 | blue book | 03/03/2006 11:48 | 1 | http://www.kbb.com |
| 417749 | best dog for older owner | 06/03/2006 11:48 | 1 | http://www.canismajor.com |
| 417749 | best dog for older owner | 06/03/2006 11:48 | 5 | http://dogs.about.com |
| 417749 | rescue of older dogs | 06/03/2006 11:55 | 1 | http://www.srdogs.com |
| 417749 | school supplies for the iraq children | 06/03/2006 13:36 | 1 | http://www.operationiraqichildren.org |
| 417749 | school supplies for the iraq children | 06/03/2006 13:36 | 2 | http://www.operationiraqichildren.org |
| 417749 | pine straw lilburn delivery | 06/03/2006 18:35 | | |
| 417749 | pine straw delivery in gwinnett county | 06/03/2006 18:36 | | |
| 417749 | landscapers in lilburn ga. | 06/03/2006 18:37 | | |
| 417749 | pine straw in lilburn ga. | 06/03/2006 18:38 | 9 | http://gwinnett-online.com |
| 417749 | gwinnett county yellow pages | 06/03/2006 18:42 | 1 | http://directory.respond.com |

**Figure-I-1-exemple d'une session sauvegardé dans un Journal des requêtes[26]**

## I-4- L'utilité de l'analyse des journaux de requêtes

La technique de journalisation a toujours été utilisée depuis l'aube de la création du World Wide Web. Mais elle ne s'intéressait pas plus qu'à l'analyse des connexions, pages consultées et problèmes de consultation des ces dernières.

Mais de nombreux recherches ont commencé à s'intéresser à d'autres données qu'on pourrait aussi collecter (comme la requête par exemple).
Ainsi des travaux sur les techniques de PageRinking de catégorisation des utilisateurs grâce à leurs centres d'intérêts, la reformulation des requêtes et bien d'autres ont pu tirer profit du journal de requêtes.
D'un autre côté la technique de création de journal de requêtes reste plus ou moins plus facile et moins coûteuse que l'utilisation de services d'analyse du web.
Il faut savoir aussi que d'autres parties s'intéressent aux données des journaux de requêtes telles que les parties commerciales et gouvernementales.

## I-5- Les risques liés à l'analyse des journaux de requêtes

Toutefois, la publication des journaux de requêtes sans anonymisation peut causer des risques sur la vie privée des utilisateurs. Ce danger est devenu plus clair après le cas en 2006 lorsque American Online (AOL) a publié une version soit disant dés identifiées de leur journal de requêtes, dans cette publication il y avait les données d'environ 650 000 utilisateurs qui ont envoyé 20 millions de requêtes enregistrées en claire avec les résultats (pages et sites web) consultés.
Peu de temps après un journal de presse (New York Times) raconte l'histoire réel qui démontre la détermination de l'identité d'un utilisateur réel, l'utilisateur correspondant à l'ID 4417749 qui s'avère être une dame dénommée Thelma Arnold, une femme de 62 ans vivant en Géorgie. La session de recherche de Mme Arnold contenant plusieurs requêtes faites sur des entreprises et services à Lilburn GA ainsi que d'autres requêtes contenant des noms de personnes de sa famille avec des précisions autour de son âge ont finalement conduit assez facilement à sa ré-identifier. **[2]**

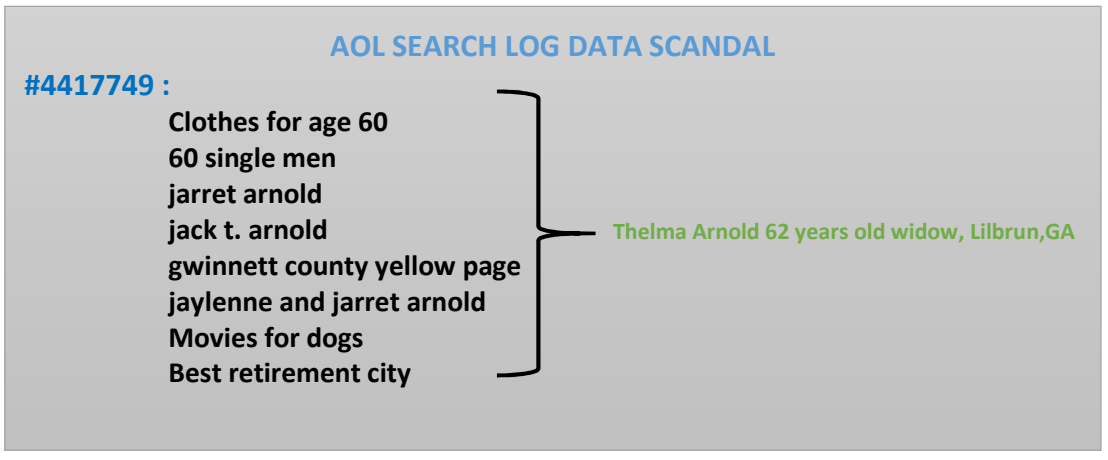



**Figure I-2-l'identification de l'utilisateur** 4417749 **par ses activités sur le web(risque)**

## CHAPITRE I : Recherche d'information et les journaux de requêtes

Après ce scandale, l'événement prend plusieurs dimensions, il y a des plaintes contre American Online et les gens sont convaincu que les journaux des requêtes ne soient pas conservés pour des raisons de développement des systèmes de recherches mais pour d'autres besoin probablement pour l'espionnage et d'intervention dans la vie privée des gens **[2]**

## Conclusion

Nous avons vu dans cette chapitre l'intérêt et le risques d'analyser les données de recherche web. De nombreuses recherches se penchent aujourd'hui sur la question de la protection de la vie privée des utilisateurs sur Internet. Mais il est utile de définir en premier lieu les notions nécessaires à la compréhension de la problématique de divulgation des données personnelles et de la protection de la vie privée dont la clause d'anonymat fait partie. C'est ce que nous verrons dans le chapitre II de ce mémoire.

## CHAPITRE II : Problématique de la divulgation des données personnelles

# Introduction

La notion de protection de vie privée n'est pas toujours facile à définir d'une façon précise. En effet, ce qui peut être partagé au public sans problème pour certains, peut être hautement privé pour d'autres. Néanmoins, il était convenu dans certain consensus, que les services accédant à des données concernant des personnes et pouvant nuire à leurs vies privées sont obligé selon les lois du pays de garantir un ensemble de clauses à savoir : L'anonymat, le pseudonymat, la non-observabilité et la non-chainabilité.

Mais il est préférable de situer en premier lieu la protection de la vie privée dans le contexte général de la sécurité.

# II-1-La sécurité (Security) ?

La sécurité se définit par l'ensemble des moyens de protection (Safety) contre les menaces internes et sur les réseaux. Pour préciser la définition il se doit de citer les trois niveaux de la sécurité :

## II-1-1- Premier niveau : la sécurité informatique (IT Security)

La sécurité informatique est une branche utilisant des moyens pour réduire la fragilité d'un système contre les menaces accidentelles ou intentionnelles sur les logiciels ou le matériel. La sécurité d'informatique doit garantir :

**La disponibilité :** demande que l'information sur le système soit disponible aux personnes autorisées.
**La confidentialité :** demande que l'information sur le système est lue que par les personnes autorisées par exemples les mots de passe, les chiffres de la carte bancaire. etc.
**L'intégrité :** demande que l'information sur le système soit modifiée que par les personnes autorisées.

## II-1-2- Deuxième niveau : la sécurité d'information (Information Security)

La sécurité de l'information signifie la protection des systèmes d'information contre l'accès non autorisé, la divulgation, la perturbation, la modification ou la destruction. Cette sécurité est incluse dans la sécurité d'informatique, parce que les trois caractéristiques fondamentales la touchent et les deux types de la sécurité sont liés. Ces deux types partagent les objectifs communs de protection : de la confidentialité, de l'intégrité et de la disponibilité d'information ; Cependant, il y a quelques différences entre eux **[4]** :

**Pour la confidentialité :** La confidentialité c'est le principe d'empêcher la divulgation d'informations à des personnes non autorisées. Son absence est connue comme la perte de confidentialité. **[3]**

**Pour l'intégrité :** Lorsque l'information est modifiée de façon inattendue (imprévu), le résultat est connu comme la perte d'intégrité.

L'intégrité signifie que des modifications non autorisées sont apportées à l'information que ce soit par erreur de l'être humaine ou intentionnelle. L'intégrité est particulièrement importante pour la sécurité des données utilisées pour des activités telles que les transferts de fonds électroniques, le contrôle de la comptabilité. **[3]**

**Pour la disponibilité :**

Les informations peuvent être effacées ou deviennent inaccessibles, la conséquence s'appelle une perte de disponibilité.

La disponibilité signifie que les personnes autorisées à obtenir de l'information ne peuvent obtenir ce dont elles ont besoin. **[3]**

S'ajoutent à ces caractéristiques d'autres qui sont l'authentification, l'autorisation et la non-répudiation :

**L'authentification :** Elle consiste à vérifier l'identité, lorsque l'information est échangée, l'authentification garantit sa source et sa destination.
**La non-répudiation :** Elle consiste à prouver qu'une information a été effectivement émise par son expéditeur ou reçue par son destinataire. Plus généralement, la non-répudiation consiste à garantir que le libérateur d'information ne peut pas nier l'avoir écrite ou transmise ce qui est de même pour son récepteur.
**L'autorisation :** elle consiste de donner le droit d'effectuer une opération précise. **[3]**

## II-1 -3-Troisième niveaux : la sécurité des données personnelles et la protection de la vie privée

La vie privée concerne des personnes physiques, selon la Commission Nationale de l'Informatique et des Libertés( CNIL). Son contenu est varié selon les circonstances et les personnes concernées.
Avec l'extension de l'internet, de nouvelles menaces ont vu le jour pour le respect de la vie privée de l'utilisateur, Actuellement, les données personnelles sont exposées en public et peuvent être accessibles par n'importe qui si elles ne sont pas sécurisées ce qui n'est toujours pas garantie.
Généralement, la vie privée englobe la vie personnelle qui représente l'identité, l'état de la santé, salaire, situation familiale, les secrets professionnels...etc. **[7]**

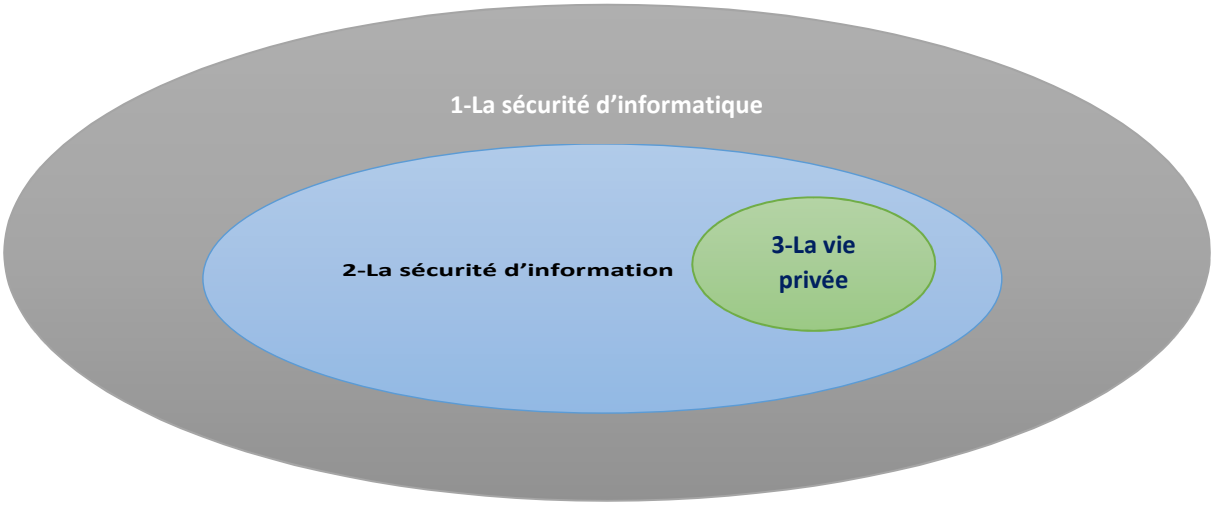



**Figure II-1-les trois niveaux de la sécurité**

La protection de la vie privée (ou privacy en anglais) est basée sur quatre propriétés :
a-L'anonymat :c à d les autres utilisateurs soient incapable d'observer et déterminer l'identité véritable d'un utilisateur concerné.

b- Le pseudonymat : C'est l'utilisation d'un pseudonyme au lieu du vrai nom.

c- La non-observabilité : Consiste à ce que des utilisateurs ou des sujets ne puissent pas déterminer si une opération est en cours d'exécution.

d- La non-chainabilité : C'est l'impossibilité pour d'autres utilisateurs d'établir un lien entre les différentes opérations faites par un même utilisateur **[7]**

Nous sommes concernés dans notre travail par l'anonymat des utilisateurs.

## II-2-Les données personnelles et les données sensibles ?

Le journal de requête contient un ID utilisateur ou un ID de session utilisateur, des termes de requête, un horodatage et éventuellement une URL d'un résultat cliqué et de la position du résultat. Les termes de la requête contiennent beaucoup d'informations qui peuvent inclure des informations sensibles. Les informations sensibles contenues dans les termes de la requête peuvent être des informations personnelles liées à l'identité d'une personne. **[9]**
Selon l'Office of Management and Budget américain (AOMB), les données personnelles désignent toutes les informations qui permettent d'identifier, de contacter ou de localiser une personne **[5].** Se sont toutes les informations qui facilitent l'identification et localisation des individus, telles que le nom, le prénom, la date de naissance (moins dangereux), l'adresse, le

numéro de permis de conduire, le numéro de carte bancaire, le numéro de compte bancaire, les dossiers médicaux. etc.
Une autre étude effectuée par General Accounting Office américain, 87 % de la population américaine peut être identifiée uniquement grâce au sexe, à la date de naissance et au code postal. **[5]**
Ces informations peuvent devenir sensibles lorsqu'elles constituent un danger pour la réputation, et la confiance liée à la personne. Le résultat d'un examen HIV n'est considéré comme sensible que s'il s'avère positif. Dans un autre exemple, le bilan annuel d'une entreprise ne constitue une menace pour sa réputation que si l'entreprise est déficitaire.


**Figure II-2-exemple de données personnelle et de données sensibles [5]**

Les informations qui pourraient conduire à un vol d'identité, qui pourraient être exploitées pour la discrimination et la fraude, ou tout simplement conduire à un vie sociale inconfortable pour l'individu sont considérées comme des données sensibles et sont à protéger. Il est donc nécessaire de supprimer tout lien entre ces données et l'identité de la personne.

## II-3-La Ré-identification dans les journaux des requêtes

Le terme « ré identification » désigne la relation correcte entre des données apparemment anonymes et des informations d'identification explicites, telles que le nom ou l'adresse des personnes qui sont les sujets des données.
A sa première utilisation la technique de ré-identification n'a pas été imaginée dans le but de trouver l'identité de la personne reliée à un ensemble de données anonymisées. Mais plutôt pour des besoins de consolidation, de vérification et d'agrégation dans les bases de données. Grâce à ces techniques il était devenu possible de lier les données d'un même utilisateur qui étaient éparpillées dans différentes bases de données (bases de données médicales de différents services).

La ré-identification est faite via les bases de données divulguées par un fournisseur : Supposant la divulgation de deux bases de données : l'une est constituée de données non identifiées, comme des séquences d'ADN et l'autre contient des données identifiables. Les bases de données ne semblent pas être liées et, par conséquent, les politiques de protection des données existantes attestent une protection suffisante contre la ré identification. Mais avec plusieurs emplacements disponibles, les données d'une personne peuvent être suivies d'un endroit à l'autre, ce qui donne une empreinte pour déduire la localisation. Les algorithmes de la ré identification sont évalués sur des populations dérivées de bases de données réelles, (visites hospitalières dérivées de bases de données médicales et de weblogs dérivés de bases de données Internet). Des preuves expérimentales avec des populations du monde réel confirment que des quantités importantes de populations sont à risque de ré identification. Ainsi la ré identification peut être avec les relations entre plusieurs institution (Yahoo, Facebook, NYT…). Beaucoup d'enquêtes ont révélé que les collections de données, sont dérivées à partir de modèles de protection ad hoc, peuvent souvent être associées à d'autres collections qui incluent des identifiants explicites et automatiquement identifier la personne **[6].**

## II-3- Le compromis entre la vie privée et l'utilité

Le compromis entre la vie privé et l'utilité est fondé sur les données que nous éliminons d'un journal par les techniques d'anonymisation et le plus de confidentialité que nous transmettons aux utilisateurs, alors moins des données sont utiles pour ceux qui cherchent à extraire les données**. [2]**

La plupart des études d'analyse des journaux de requêtes se concentrent de la modélisations les modèles de navigation dans lesquelles respecter la vie privée et le développement du web au même temps. Mais la difficulté existe c'est de déterminer quels sont les points de la confidentialité ou la quantité d'utilité qu'on a parlé à l'avance. La gestion des informations personnelles (Privacy Information Management) est la pratique et l'étude des activités que les gens accomplissent pour acquérir, organiser, conserver, récupérer et utiliser des informations telles que des pages Web, des résultats de recherche sur le Web, des courriels et d'autres types de fichiers. **[9]**

Certains utilisateurs peuvent considérer qu'ils sont cliqués « OK » le moteur de recherche accès aux requêtes au mode de la personnalisation, mais pas pour la publicité, mais la non-publication c'est infusant de confirmer que ces informations sont conservées. Il y a des utilisateurs ne voudraient aucune de leurs recherches enregistrées ont été considéré une violation pour toute information conservée, et encore moins révélé des autres côtés, ça se résume le droit de l'information personnelle. D'autre part, il y a des utilisateurs qui sont heureux, pour une raison quelconque, de partager leurs photos, leurs numéros de téléphone ou bien leurs historiques de recherche, ça c'est la liberté d'expression.

L'utilité est généralement analysée avec une tâche spécifique à l'esprit. Si nous cherchons simplement à mesurer le temps entre les requetés effectués, alors nous n'avons besoin d'aucune information d'identification existe dans le journal des requêtes.

Cependant, la tâche la plus complexe c'est le défi de réduire la confidentialité, et il faut que nous connaissons le niveau de confiance sur les données du journal. En plus, une inférence de

genre peut être possible en fonction du comportement de recherche et des phrases clés et bien que la certitude absolue ne soit pas possible. **[2]**

Nous considérons maintenant certaines approches pour le partage du journal des requêtes afin de permettre la recherche dans les domaines ci-dessus tout en limitant les compromis potentiels de la vie privée des utilisateurs. Avant de proposer les approches d'anonymisation, nous analysons les informations potentiellement sensibles dans les journaux de requêtes et les classons selon quelques dimensions. Nous présentons quelques scénarios pour les violations potentielles de la vie privée et motivons notre approche d'anonymisation

## Conclusion

La quantité remarquable d'information privée qui est publiée dans les Query Log et l'identification des personnes a suscité aujourd'hui l'intérêt des chercheurs. De nombreuses solutions d'anonymisation ont été proposées. Dans la prochaine section nous étudierons ces solutions en citant leurs points forts et points faibles.

## CHAPITRE III : Etude des techniques d'anonymisation

# Introduction

Il existe (selon la CNIL) deux types de données : des données à caractère personnel, et des données anonymes. Les données sont à caractère personnel, elles concernent des personnes physiques, à partir de ces derniers on aura une identification direct ou indirect. Au contraire, toute donnée qu'il est impossible d'associer avec une personne physique sera dite anonyme. **[11]** La publication des journaux de requêtes contenant des informations précieuses pour la recherche ou le marketing, peut enfreindre la vie privée des personnes. Par conséquent, les chercheurs dans le domaine de la protection de la vie privée proposent des techniques d'anonymisation afin d'équilibrer le respect de la vie privée avec l'utilité des journaux de recherche. **[10]**

# III-1-Anonymisation

L'anonymisation c'est un processus permettant de cacher l'identité ou les informations personnelles des individus. Dans une autre définition l'anonymisation est un processus de suppression des identificateurs explicites qui peuvent être disponibles pour identifier directement (telle que le numéro de carte bancaire) et les indicateurs moins explicites (tels que les noms l'âge, le sexe, le code postal,). **[10]**
De nombreuses solutions d'anonymisation ont été proposées dans le domaines des bases de données telles que k-anonymisation, l-diversité, la pseudonymisation, et t-proximité.

# III-2-Les technique d'anonymisation dans les Bases de données

Nous décrivons dans cette section cinq techniques d'anonymisation appliquées dans les bases de données.
Nous considérons une base de données constituée d'un n-uplets (ensemble d'enregistrements) ayant chacun une structure identique. **[11]**

| **Pathologie** (*Donnée sensible*) | **Sexe** | **Code postal** | **Age** | **Numéro de sécurité sociale** (*Identifiant*) |
|---|---|---|---|---|
| Cancer | F | 75005 | 75 | 2023475123123 |
| Grippe | F | 75012 | 40 | 2067875123123 |
| Grippe | M | 78000 | 12 | 1101175123123 |

**Tab 1- Une base de données personnelles [11]**

Avec les techniques mentionné précédent, qui cherchent à cacher ou briser le lien existant entre une personne du monde réel, et ses données sensibles :

### III-2-1-La pseudonymisation :

La pseudonymisation consiste à supprimer les champs directement identifiants des Enregistrements, et à rajouter à chaque enregistrement un nouveau champ, appelé pseudonyme, dont la caractéristique est qu'il doit rendre impossible tout lien entre cette nouvelle valeur et la personne réelle. Pour faire le pseudonyme, on utilise souvent une fonction de hachage que l'on va appliquer à l'un des champs identifiants (par exemple le numéro de sécurité sociale). Cette fonction qui n'est pas rend impossible le fait de déduire la valeur initiale. Elles pourraient partager ces données de manière anonyme en hachant. L'avantage de la pseudonymisation est qu'il n'y a aucune limite sur le traitement des données après. **[11]**

| ID | Age | CP | Sexe | Pathologie |
|---|---|---|---|---|
| 1 | 75 | 75005 | F | Cancer |
| 2 | 40 | | F | Grippe |
| 3 | 12 | | M | Grippe |

| Pathologie | MOY (Age) |
|---|---|
| Cancer | 75 |
| Grippe | 26 |

**Tab 2-Pseudonymisation et exemple de calcul [11]**

L'inconvénient de la pseudonymisation c'est que ne donne pas un niveau de protection suffisamment élevé : la combinaison d'autres champs peut permettre de retrouver l'individu concerné. Sweeney l'a mis en évidence aux Etats-Unis en 2001 en croisant deux bases de données, une base de données médicale pseudonymisée et une liste électorale avec des données nominatives. Le croisement a été effectué non pas sur des champs directement identifiants, mais sur un triplet de valeurs : code postal, date de naissance et sexe, qui est unique pour environ 80% de la population des Etats-Unis3 ! Elle a ainsi pu relier des données médicales à des individus (en l'occurrence le gouverneur de l'Etat). **[11]**

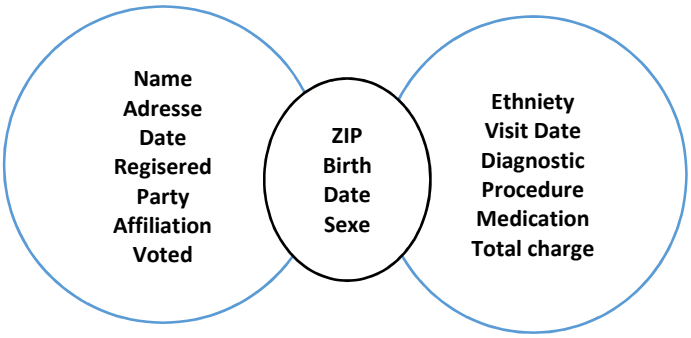



**Figure III-1-exemple de recoupement d'une base anonyme (source Sweeney 2002) [11]**

## III-2-2-Le k-anonymat :

La technique de k-anonymat a été proposé par Sweeney pour protéger les attaques de liaison des enregistrement (ou record linkage en anglais). Le principe de ce technique c'est de flouter la possibilité de lier un n-uplet anonyme à un n-uplet non anonyme de la manière suivante :
1) déterminer les ensembles d'attributs (appelés quasi identifiants) qui peuvent être utilisés pour croiser les données anonymes avec des données identifiants.
2) faire la réduction le niveau de détail des données de telle sorte qu'il y a au moins k n-uplets différents qui ont la même valeur de quasi-identifiant. L'avantage du *k*-anonymat est que l'analyse des données continue de fournir des résultats exacts. **[11** Sweeney k-anonymat. International Journal on Uncertainty, 2002**]**

| Nom | Activité | Age | Diag |
|---|---|---|---|
| Sue | "M2" | 22 | Grippe |
| Pat | "MCF" | 27 | Cancer |
| Dan | "PhD" | 26 | Cancer |
| Bob | "M1" | 21 | VIH |
| Bil | "L3" | 20 | Grippe |
| Sam | "PhD" | 24 | Cancer |
| John | " M2" | 22 | Rhume |
| Jim | "M2" | 23 | Rhume |
| Tom | "L2" | 21 | Allergie |

**Données brutes**

| Activité | Age | Diag |
|---|---|---|
| "M2" | [22,23] | Grippe |
| "M2" | [22,23] | Rhume |
| "M2" | [22,23] | Rhume |
| " Etudiant" | [20,21] | VIH |
| " Etudiant" | [20,21] | Grippe |
| " Etudiant" | [20,21] | Alergie |
| "Ens" | [24,27] | Cancer |
| "Ens" | [24,27] | Cancer |
| "Ens" | [24,27] | Cancer |

**Données anonymisées par la généralisation**

**Tab 3- Anonymisation d'une table sur des données universitaires [11]**

Généraliser : signifie en fait « enlever un degré de précision » à certains champs. Ainsi, il est impossible d'être sûr à plus d'une chance sur *k* qu'on a bien lié un individu donné avec son *n*-uplet anonyme. D'après la figure : La généralisation des champs activité et âge d'une base de données médicale sur des étudiants et enseignants d'une université. Les étudiants sont identifiés par leur niveau d'étude (L3, M1, etc.), qui se généralise en « étudiant », et les enseignants par leur position académique (Doctorant, maître de conférences, etc.), qui se généralise en « enseignant ». L'inconvénient de ce technique c'est la détermination des généralisations à effectuer pour produire les quasi-identifiants, ce qui peut être fait soit par un expert humain qui connaît le domaine, ou bien par un calcul informatique ainsi elle est coûteuse pour une base de données réelle. **[11]**

## III-2-3-La l-diversité :

La l-diversité essaye de contourner l'inconvénient de k-anonymat, en ajoutant une contrainte supplémentaire sur les classes d'équivalence. Dans le figure, on voit que pour constituer de telles classes on doit parfois regrouper ensemble des étudiants et des enseignants. Leur activité est alors désignée de façon encore plus générale (« université »). Notons qu'on peut également lister les valeurs possibles, par exemple avoir une modalité « Étudiant ou Doctorant » (Etu/PhD). **[11]**

| Nom | Activité | Age | Diag |
|---|---|---|---|
| Sue | ''M2'' | 22 | Grippe |
| Pat | "MCF" | 27 | Cancer |
| Dan | "PhD" | 26 | Cancer |
| Bob | "M1" | 21 | VIH |
| Bil | "L3" | 20 | Grippe |
| Sam | "PhD" | 24 | Cancer |
| John | '' M2'' | 22 | Rhume |
| Jim | ''M2'' | 23 | Rhume |
| Tom | "L2" | 21 | Allergie |

**Données brutes**

| Activité | Age | Diag | |
|---|---|---|---|
| "UNIV" | [21,27] | Grippe | 3 valeurs distincts |
| "UNIV" | [21,27] | Cancer | |
| "UNIV" | [21,27] | VIH | |
| '' Etu/PhD'' | [20,24] | Grippe | 3 valeurs distincts |
| '' Etu/PhD'' | [20,24] | Cancer | |
| '' Etu/PhD'' | [20,24] | Rhume | |
| ''Etu/Phd'' | [21,26] | Cancer | 3 valeurs distincts |
| ''Etu/PhD'' | [21,26] | Rhume | |
| ''Etu/PhD' | [21,26] | Allergie | |

**Données 3-anonymes et 3-diversees**

**Tab 4-Données l-diverses [11]**

L'inconvénient de ce technique c'est que reste possible de déduire des informations. Par exemple dans la Figure précédent qu'on peut déduire qu'un étudiant de 20 ans aura une probabilité 0.33 (soit 1/*k*) d'avoir la grippe, 0.33 d'avoir le cancer et 0.33 d'avoir un rhume... et surtout aucune chance d'avoir une autre pathologie. Si on sait que Bill est la seule personne de la base dans ce cas de figure, alors on peut déduire des informations sensibles à son sujet.

### III-2-4-La t-proximité :

La t-proximité pour réduire l'information qui peut être observée directement, à partir d'un regroupement de données en classes d'équivalences selon le processus du k-anonymat. Ce modèle est basé sur une connaissance globale de la distribution des données sensibles. C'est-à-dire en ce cas les pathologies, pour essayer de faire coller au mieux les valeurs sensibles d'une classe d'équivalence à cette distribution, et ainsi éviter le problème de déduction d'informations soulevé par la *l*-diversité. Le facteur *t* que nous ne détaillons pas ici, indique dans quelle mesure on se démarque de la distribution globale. **[11]**

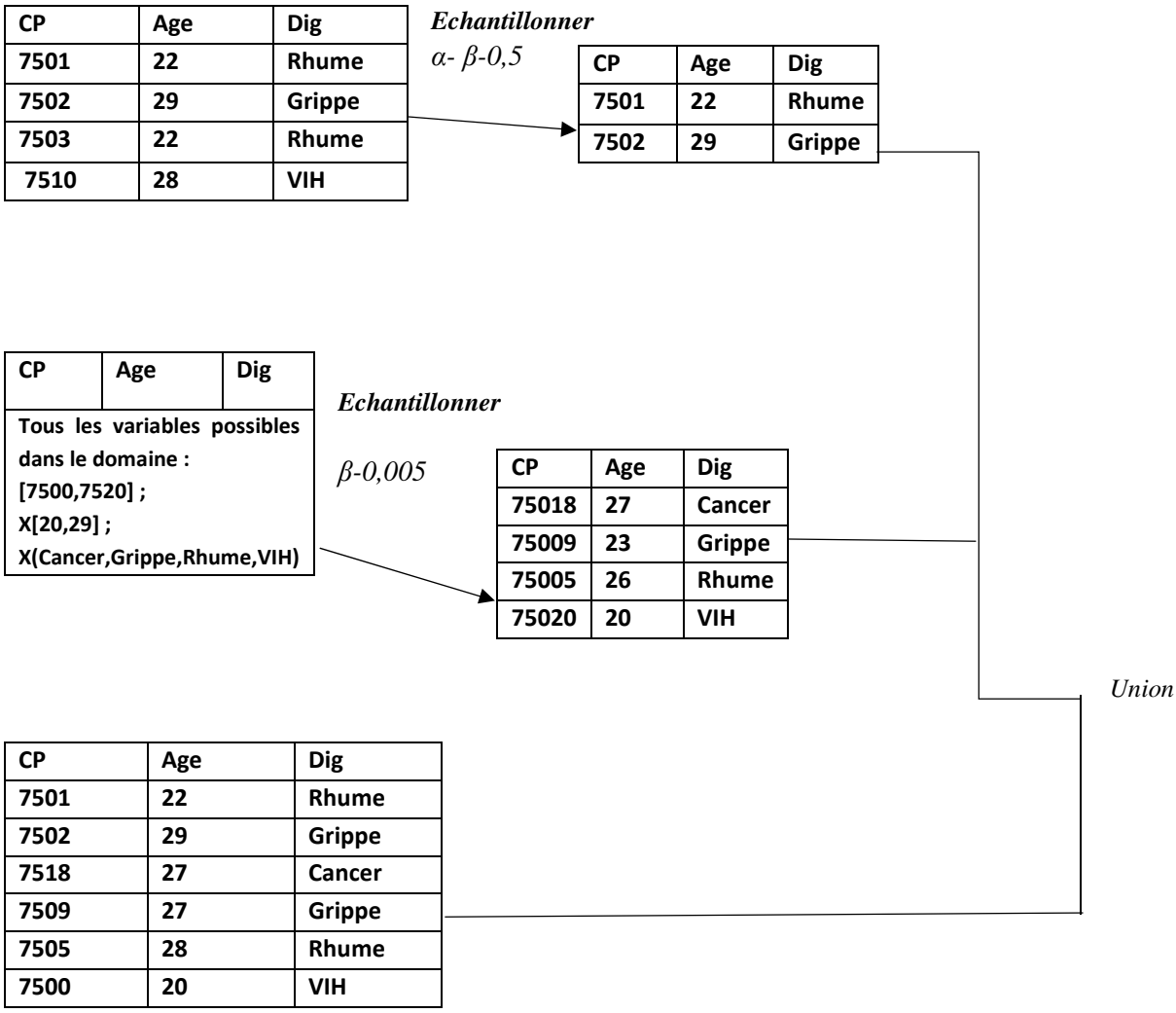
| Age | Sexe | Département | Pathologie | Nbrs d'individus |
|---|---|---|---|---|
| <45 | M | 75 | Grippe | 400 |
| <45 | M | 75 | Rhume | 800 |
| >45 | M | 75 | Grippe | 500 |
| >45 | M | 75 | Rhume | 1000 |
| <35 | F | 75 | Grippe | 300 |
| >35 | F | 75 | Rhume | 600 |
| >35 | F | 75 | Grippe | 600 |

**Tab 5- t-proximité [11]**

La t-proximité souffre de plusieurs problèmes, le plus important étant sans doute son utilité ! En effet, il parait évident d'exploiter des données *k*-anonymes ou même *l*-diverses pour découvrir des corrélations entre des données appartenant au quasi-identifiant et des données sensibles.

### III-2-5-La confidentialité différentielle (Differential Privacy) :

La confidentialité différentielle, une méthode n'est pas comme les méthodes précédentes, elle est la seule à donner des garanties de borner les informations qu'on peut apprendre sur les individus. Elle introduit un échantillonnage des données vraies avec une probabilité $\alpha$, et une génération de données fictives (n'est pas vrai) avec une probabilité $\beta >> \alpha$. Les garanties permettent d'estimer le risque de ré-identification des *n* uplets, les n-uplet (soit vrai ou faux) est doublement bornée, c a d on a jamais *n*-uplet soit vrai avec une probabilité supérieure à $\alpha$ et s'il soit faux on a jamais qu'il a une probabilité inférieure de $\beta$. **[11]**




*Jeu de données anonymisée*

**Tab 6-Confidentialité Différentielle [11]**

Dans la confidentialité différentielle on est obligé de calculer un estimateur d'un agrégat que l'on souhaite connaître. Par exemple on a calculé le nombre moyen de malades de la grippe par département, et supposons pour simplifier que les données fictives sont générées de manière équiprobable. Alors l'estimation du nombre total de malades de la grippe est par la fonction suivante, dont l'objectif est de soustraire le bruit (connu) introduit :

***Nb Rhume estimé= (NB Rhume anonyme - β*Nb Rhume domaine) /α= (2 - 200*0.005) /0.5=2***

L'inconvénient de cette méthode consiste le problème de la vraisemblance des données fictives. **[11]**

## III-3-Les solutions d'anonymisation dans journaux de requêtes

Comme on a vu l'anonymisation dans les bases de données, nous sommes obligés d'expliquer quant à la motivation de proposer des solutions adaptées aux journaux de requêtes et ceci pour les raisons suivantes :

- Les données personnelles dans un journal de requête contrairement à la base de données ne sont pas structurées sous forme de champs ni identifiées clairement. En effet, la requête de l'utilisateur se compose de mots clé et est écrite en langage naturel ce qui rend difficile l'identification des données personnelles.
- D'un autre côté, la quantité des données dans un d'un journal de requête est énorme et rend très difficile l'application de ces solutions.

Pour ces raisons il est devenu nécessaire d'étudier la problématique de l'anonymisation des données dans un nouveau contexte, celui des journaux de requêtes. Nous proposons dans ce qui suit quelques-unes de ces solutions.

### III-3-1 Calcul d'empreinte du champ identificateur externe

Parmi les solutions d'anonymisation dans les journaux de requêtes, nous avons la solution qui traite le champ identificateur externe par calcul d'empreinte. Dans Le calcul d'empreinte des identificateurs est appliquée la fonction de hachage (MD5) sur les adresses IP ou bien sur les Ids des cookies. Leur principe consiste d'enlever les identificateurs qui pourraient être utilisées pour identifier les individus. Dans l'exemple ci-dessous nous montrons le lien qui peut être fait entre plusieurs sources d'information en l'occurrence le log du commerçant et un query log d'un moteur de recherche .il obtient un lien direct avec les données publiées dans le query log et peut obtenir donc une identité claire de la personne ayant recherchée "HIV test". Ceci démontre la faiblesse de la technique de hachage des identifiants. De plus, cette technique a été utilisées dans le journal de requêtes d'AOL, et n'a pas empêché la ré-identification d'une personne. **[12]**

| Point de vente | Date et Heure | Quantit-é | Article | Adresse IP |
|---|---|---|---|---|
| Paris | 2008-06-21 14:54:41 | 1 | bicyclette | 76.26.159.134 |
| Paris | 2008-06-21 15:59:03 | 1 | Casque de bicyclette | 76.26.159.134 |

**Tab 7- Historique d'achat en ligne d'un client chez un marchand de bicyclette [12]**

| URL | Navigateur/OS | Date et Heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|---|
| http://www.cdc.gov/hiv/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:54:41 | HIV test | 359b81298e37g | 76.26.159.134 |
| http://www.wwc.org/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:59:03 | Clinique Whitman-Walker | 711k03296e86g | 76.26.159.134 |

**Tab 8- Journal de requêtes d'un utilisateur avec l'adresse IP client en clair [12]**

| URL | Navigateur/OS | Date et Heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|---|
| http://www.cdc.gov/hiv/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:54:41 | HIV test | 359b81298e37g | M98b3hd444 |
| http://www.wwc.org/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:59:03 | Clinique Whitman-Walker | 711k03296e86g | M98b3hd444 |

**Tab 9- Même journal de requêtes avec l'adresse IP hachée [12]**

## III-3-2- Anonymisation au niveau de la session

Le groupement des requêtes sous un identifiant unique constitue une brèche de la sécurité. Les solutions opèrent à ce niveau essayant de protéger l'utilisateur de toutes ou d'une partie de ses requêtes. Parmi ces solutions nous citons :

### 1- La décomposition d'une session selon une période de temps ou selon le comportement du navigateur de l'utilisateur

Dans ce niveau en raccourcissant la durée de vie des identificateurs, le raccourcir des sessions exigerait probablement la suppression de l'adresse IP depuis le journal de requête surtout dans le cas d'adresse IP statique, la réduction de durée de vie d'un cookie, ou le remplacement d'identificateurs internes persistants avec des identificateurs qui changent périodiquement. **[12]**

| URL choisie | Date et heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|
| | 2010-08-16 13:23:56 | Alice Birdsboro | HstCfa6806531312541860140 | 76.26.176.123 |
| http://www.bigdiscount.com | 2010-08-16 19:47:35 | Magasin déstockage à Seattle. | HstCfa6806531312541860140 | 76.26.176.123 |
| http://www.sevenadventure.com | 2010-08-18 19:58:06 | Sevenadventure jeux | HstCfa6806531312541860140 | 76.26.176.123 |
| http://www.copaindavant.com | 2010-08-19 15:01:00 | Classe terminale 2010 lycée shorecrest high school | HstCfa6806531312541860140 | 76.26.176.123 |

| URL choisie | Date et heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|
| | 2010-08-16 13:23:56 | Alice Birdsboro | **HstCfa6806531312541860140** | |
| http://www.bigdiscount.com | 2010-08-16 19:47:35 | Magasin déstockage à Seattle. | **HstCfa6806531312541860140** | |
| http://www.sevenadventure.com | 2010-08-18 19:58:06 | Sevenadventure jeux | **Ks76cfa6806531312541860140** | |
| http://www.copaindavant.com | 2010-08-19 15:01:00 | Classe terminale 2010 lycée shorecrest high school | **Ks76cfa6806531312541860140** | |

**Tab 10- Suppression de l'adresse IP et réduction de la durée de vie d'un cookie [12]**

Le groupement des requêtes selon un critère temporel(raccourcir), a le potentiel d'être un protecteur de vie privée, parce que des sessions plus courtes peuvent enlever la liaison entre un utilisateur et l'intégralité de son historique de recherche, ensuite la difficulté d'obtenir toutes les requêtes d'un même utilisateur, mais il provoque aussi un profilage incomplet d'utilisateur

le cas des autres sessions créés sont capables de faire un lien entre eux pour identifiant l'individu. Les risques de révélation accidentelle et malveillante ne sont pas entièrement résolus par cette technique parce que la requête contient toujours des informations personnelles. **[12]**

**2- La décomposition d'une session selon le thème de la recherche**

Le thème de recherche est le sujet principal pour l'utilisateur quand effectue sa requête. La décomposition d'une session appartenant à un même utilisateur selon les différents thèmes qui caractérisent ses recherches a été proposée par Adar. Les utilisateurs ont des intérêts multiples, chaque intérêt a un ensemble de requêtes se rapprochant de lui. Par exemple, si une personne est intéressée par les voyages, ses requêtes peuvent être « les lieux les plus visités au monde », « meilleur pays pour visite touristique en hiver », « liste des agences de voyage d'une ville X », « voyage organisé » …etc. Les utilisateurs ont tendance à avoir beaucoup d'intérêts et ces intérêts peuvent être liés, ce qui résulte en des requêtes à thèmes multiples. Le grand défi de cette solution est de résoudre les cas de requêtes où les thèmes qui sont chevauchés(composés). Exemple : les requêtes « plantes de décor », « plante d'intérieur » et « décors d'intérieur » et le problème c'est qu'ils peuvent avoir certains résultats en commun. **[12]**

**3-La suppression ou la généralisation de l'identificateur externe (adresse IP, cookie ID)** :

Dans cette technique, les identificateurs (adresse IP, ID cookies) sont supprimés, entièrement ou partiellement du journal de requêtes. Il y a quelques octets d'une adresse IP peuvent révéler des informations sur l'emplacement physique de l'ordinateur de l'utilisateur, et même si le dernier octet ou les deux derniers octets sont enlevés, le journal de requêtes peut toujours révéler quelques informations sur l'utilisateur.

| URL | Navigateur/OS | Date et Heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|---|
| http://www.cdc.gov/hiv/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:54:41 | HIV test | 359b81298e37g | 76.26.159.134 |
| http://www.wwc.org/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:59:03 | Clinique Whitman-Walker | 711k03296e86g | 76.26.91.2 |

| URL | Navigateur/OS | Date et Heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|---|
| http://www.cdc.gov/hiv/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:54:41 | HIV test | | 76.26 |
| http://www.wwc.org/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:59:03 | Clinique Whitman-Walker | | 76.26 |

**Tab 11- Requêtes des mêmes utilisateurs sans pouvoir les distinguer [12]**

## III-3-3- Anonymisation au niveau requête

Cette technique est basée sur la suppression des requêtes non fréquentes et la solution de Calcul d'empreinte des requêtes.

### A-Suppression des requêtes non fréquentes :

La suppression des requêtes non fréquentes (proposition d'Adar) qui apparaissent rarement dans la mesure où elles contiennent souvent des informations personnelles identifiables. Mais ceci n'est pas toujours vrai. La suppression des requêtes fréquentes peut réduire les menaces de profilage d'utilisateur de deux façons : Eliminer des données identificatrices, comme l'adresse complète d'un utilisateur et enlever des informations sur des utilisateurs avec des intérêts inhabituels. Mais, les données restantes peuvent toujours être utilisées, puisque les journaux sont liés avec l'ensemble des adresse IP, et des cookies ID, ou des identificateurs internes**. [12]**

### B- Calcul d'empreinte des requêtes :

Le calcul d'empreinte, qui prend en entrée une chaîne de caractère et produit une valeur hachée qui est difficile ou impossible à inverser pour produire la valeur originale par MD5, SHA-1.

| URL | Navigateur/OS | Date et Heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|---|
| http://www.cdc.gov/hiv/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:54:41 | **HIV test** | 359b81298e37g | 76.26.159.134 |

**Tab 12- Exemple d'une recherche avec un contenu sensible "HIV test"[12]**

| URL | Navigateur/OS | Date et Heure | Requête | Cookie ID | Adresse IP |
|---|---|---|---|---|---|
| http://www.cdc.gov/hiv/ | Firefox 2.0 ; Windows XP | 2008-06-18 11:54:41 | **Wd8fy0972j9kv** | 359b81298e37g | 76.26.159.134 |

**Tab 13- Même recherche avec la requête "HIV test" hachée [12]**

Le classement des requêtes les plus populaires, il peut être utilisée pour déterminer le contenu original de requêtes hachées et compare les résultats.

## III-3-4-Anonymisation au niveau terme

Le terme est la plus petite unité informationnelle qui existe dans une session regroupant plusieurs informations personnelles, vu qu'une ré-identification peut se faire grâce à un seul terme (exemple adresse mail), plusieurs termes qui constitue une partie ou l'intégralité d'une requête. L'anonymisation dans ce niveau consiste de la suppression ou le hachage des termes identificateurs.

**Suppression du terme des identificateurs**

Le nettoyage de contenu des requêtes veut dire la suppression des informations tels que des numéros de téléphone, des Numéros de Sécurité Social, des adresses et des noms. Via le nettoyage la probabilité du divulgation (accidentelle ou malveillante) a été réduit, mais toujours il est possible de lier des requêtes à des d'individus en utilisant d'autres informations sont publiés. **[12]**

## Conclusion

Nous avons étudié dans cette partie les différentes solutions proposées pour l'anonymisation des données destinées à être publiées ou partagées. Nous avons vu les points forts et points faibles de chacune d'entre elles et nous avons discuté l'intérêt de proposer des solutions pour le contexte de journaux de requêtes à défaut d'utiliser les solutions appliquées aux bases de données.

Nous passons dans la suite de ce mémoire à notre approche adoptée.

## CHAPITRE VI : Conception et mise en œuvre

## Introduction :

Nous avons présenté aux chapitres précédents le but de l'analyse des journaux de requêtes et sa relation avec la vie privée. L'une des histoires réelles d'atteinte à la vie privée est l'histoire de Telma Arnold, abonnée d'AOL qui a été victime de divulgation de son identité ainsi que de ses recherches après que le journal de requêtes d'AOL a été publié en 2006. Pour cette raison de nombreux chercheurs se sont mis à la réflexion sur ce problème et certaines solutions ont été proposées mais les résultats ne sont toujours pas complets et efficaces.

Notre contribution consiste à proposer une solution qui traite différemment le problème de divulgation d'informations personnelles dans un journal de requêtes. Par un traitement au niveau des termes des requêtes. En effet, les termes d'une requête peuvent représenter des informations personnelles sur lesquels se basent les attaques de ré-identification. Afin de protéger l'identité de l'utilisateur, ces informations ne doivent pas rester telles qu'elles ont été envoyées par leur propriétaire. Les supprimer définitivement de la session de recherche causera la perte d'une quantité importante de données et provoquera une moindre utilité du journal de requêtes. Alors nous avons choisi les généraliser par les remplacer par termes de sens plus général.au lieu de les supprimer.

Dans notre solution nous avons choisi de traiter deux types d'informations personnelles qui sont les noms de personnes et les adresses (noms de lieux). D'autres types d'informations personnelles peuvent être traitées par exemple : les numéros de téléphone, les adresses mails, ...

## I- La généralisation

La méthode de la généralisation est une méthode qui a été appliqué dans le modèle k-anonymat, L'idée de généraliser un attribut est un concept simple. Une valeur est remplacée par une autre moins spécifique.

Exemple :

Les numéros de téléphone en Algérie « **Mobilis 213 0664452140, Djezzy 213 0776881591** »

Par la généralisation ils seront :

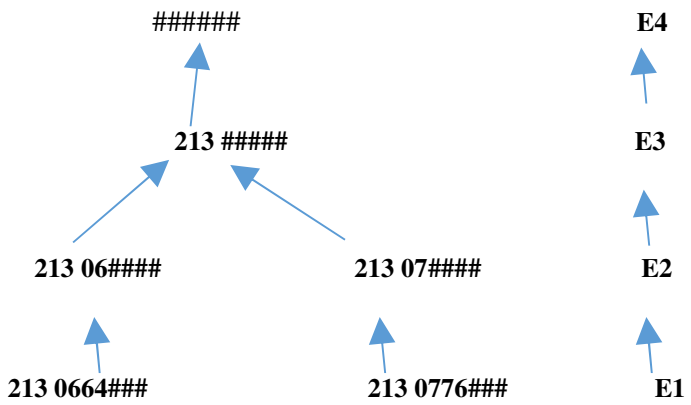



**Figure IV-1-exemple simple de la méthode de généralisation**

**E1 :** en remplaçant les six derniers chiffres du numéro de téléphone qui représente le numéro du client par des '# ' afin de généraliser l'information et dite que ce client a un numéro de Moblis et de catégorie 64 ou un numéro de Djezzy et de catégorie 76.

**E2** : en remplaçant les chiffres des catégories par '#' pour généraliser l'information que ce client a un numéro de Moblis ou bien de Djezzy.

**E3 :** ensuite en remplaçant aussi les chiffre de catégorie (6,7) par '# ' pour généraliser l'information que ce client a un numéro du pays d'Algérie.

**E4 :** finalement en remplaçant 213 par '# ' pour généraliser au maximal que cette information c'est un numéro de téléphone pas plus. **[20]**

Notre solution proposée intègre la méthode de généralisation non pas sur des valeurs d'attributs mais sur des valeurs de termes utilisés dans une requête de recherche (dans les journaux). Où en remplaçant une valeur d'un terme qui représente une information personnelle par une valeur moins spécifique. Ces valeurs qui nous choisissons de remplacer c'est les noms des lieux et les noms des personnes**.**

Dans le cas de nom de personne, le nom : « HAKIM », par la séquence de la généralisation est devenu : **HAKIM→PERSONNE**

Dans le cas des noms des lieux le nom : « KHADRA » C'est un nom d'une commune, on peut généraliser par plusieurs séquences :

**1-KHADRA → MOSTAGENM**

**2-KHADRA→ MOSTAGENM→ ALGERIE**

**3-KHADRA→ MOSTAGENM→ ALGERIE→PLACE**

## II-La généralisation dans notre solution

Notre approche consiste à traiter dans une session de recherche d'un utilisateur les termes liés à l'identité d'une personne à savoir : les noms des personnes (sans distinction entre les noms de famille et les prénoms) les noms de lieux (sans distinction entre les noms de citées, de counties, de stat) et sans tenir compte si ce nom de lieu à une relation avec l'adresse de l'utilisateur ou non). Une fois repérés, ces termes seront remplacés par un terme plus général. Ainsi, les noms de personne seront remplacés par le terme "**Person Name**" et les noms de lieux avec le terme "**Place Name**" par exemple.

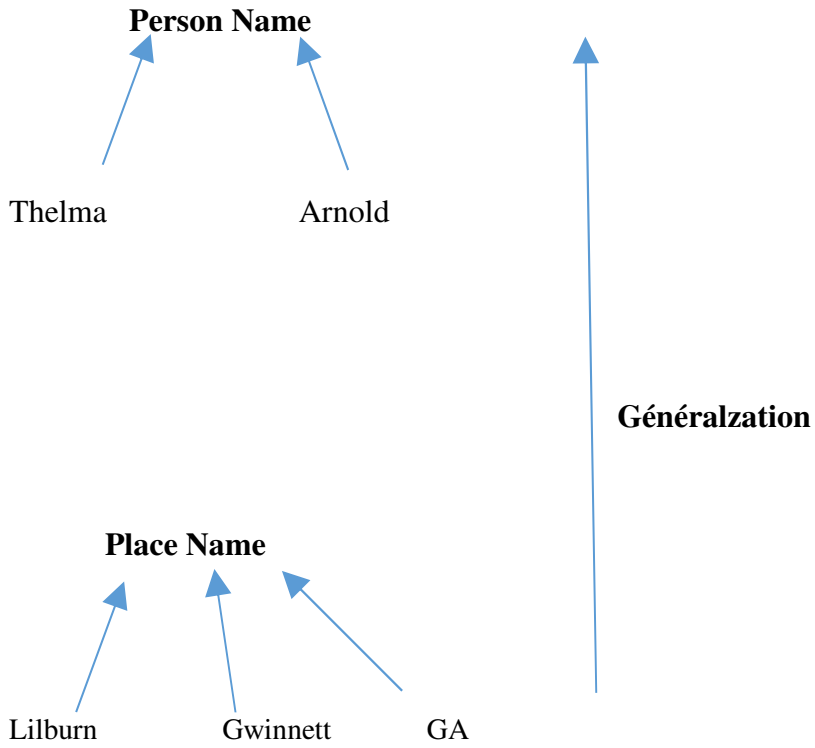



**Figure IV-2- Remplacement d'un terme avec un autre terme de sens plus général**

Par exemple dans la requête, le terme « The origin of Arnold last name » ce terme est contient un nom de personne et sera remplacé par le terme « Person Name ». La requête apparaîtra comme suit : « the origin of **Person Name** last name ». De la même façon, la requête suivante : « Lilbrun population », sera transformée comme suit : « **Place Name** population ».
Prenons aussi l'exemple des requêtes de Telma Arnold : les termes qui ont servi à identifier cette personne sont : « Arnold T », « Lilburn, Ga », de sorte à ce que la recherche dans les pages blanches avec ces termes en entrée, a conduit à un seul résultat, celui de Mme Telma Arnold.

Une attaque utilisant l'ensemble de tous les termes cités précédemment aboutit à une identité unique. A chaque fois qu'un terme a été généralisé, l'ensemble des termes identificateurs diminue.

## III-Implémentation de notre solution

### III-1-Les données de test

Nos données de test sont un extrait des données du journal de requêtes d'AOL publié en août 2006.

Dans ce qui suit un extrait du journal de requêtes d'AOL, plus précisément une partie de la session de recherche de Telma Arnold, sous le nom de User **n°4417749 .**

| | | | | |
|---|---|---|---|---|
| 4417749 | care packages | 02/03/2006 09:19 | 3 | http://www.awesomecarepackages.com |
| 4417749 | care packages | 02/03/2006 09:19 | 8 | http://www.anysoldier.com |
| 4417749 | movies for dogs | 02/03/2006 09:24 | | |
| 4417749 | blue book | 03/03/2006 11:48 | 1 | http://www.kbb.com |
| 4417749 | best dog for older owner | 06/03/2006 11:48 | 1 | http://www.canismajor.com |
| 4417749 | best dog for older owner | 06/03/2006 11:48 | 5 | http://dogs.about.com |
| 4417749 | rescue of older dogs | 06/03/2006 11:55 | 1 | http://www.srdogs.com |
| 4417749 | school supplies for the iraq children | 06/03/2006 13:36 | 1 | http://www.operationiraqichildren.org |
| 4417749 | school supplies for the iraq children | 06/03/2006 13:36 | 2 | http://www.operationiraqichildren.org |
| 4417749 | pine straw lilburn delivery | 06/03/2006 18:35 | | |
| 4417749 | pine straw delivery in gwinnett county | 06/03/2006 18:36 | | |
| 4417749 | landscapers in lilburn ga. | 06/03/2006 18:37 | | |
| 4417749 | pine straw in lilburn ga. | 06/03/2006 18:38 | 9 | http://gwinnett-online.com |
| 4417749 | gwinnett county yellow pages | 06/03/2006 18:42 | 1 | http://directory.respond.com |

**Figure IV-3-Extrait du journal de requête d'AOL (l'utilisateur n°4417749)[26]**

### III-2-Les bases de données utilisées (Bases de données des noms de personnes et des adresses)

L'identité d'une personne est représentée par les informations suivantes : nom (Name), Adresse (city, state, county). Nous avons utilisé deux bases de données, Une pour les noms de personnes **[25]** et l'autre pour les noms de lieux **[24].**

- **Les noms de personne**

Une personne porte en général un nom composé d'un prénom et d'un nom de famille (first Name, last Name).

**First Name** : le prénom est décrit comme étant une suite de lettre d'une certaine langue utilisée communément pour désigner une personne de façon unique. Les prénoms sont typiques à une région. Mais rien n'empêche dans une solution d'anonymisation de considérer tous les prénoms possibles.

**Last Name** : le nom est un nom personnel qui précède généralement le patronyme ou le nom de famille. Il est utilisé pour désigner une personne de façon unique, par opposition au nom de famille qui est partagé et hérité. C'est à dire pour désigner un groupe de personnes faisant partie d'une même famille.

- **Les addresses**

L'adresse est une chaîne de caractères utilisée pour indiquer l'endroit où réside une personne. Les utilisateurs concernés par nos données de test sont des utilisateurs habitant aux Etats Unis d'Amérique. Leurs adresses se composent souvent d'un nom de pays « **USA** », d'un nom de **STATE,** d'un nom de **COUNTY** et d'un nom d'un **CITY**.

**USA :** c'est le nom de pays et c'est une adresse qui indique une adresse d'une personne.

**STATE** : c'est le nom d'une state qui indique une wilaya d'un pays USA et aussi indique une adresse d'une personne.

**COUNTY** : c'est un nom qui indique une région d'un pays USA et aussi indique une adresse d'une personne.

**CITY :** c'est un nom indique à une rue ou une ville d'un pays USA et aussi indique une adresse d'une personne.

### III-3 Démarche de la solution

#### III-3-1- Généralisation des noms de personne

La généralisation a été simple, en remplaçant directement le nom de personne par exemple **Hakim** par le terme **Person Name**. Ce terme laisse penser qu'il s'agit d'un nom de personne sans pour autant citer lequel. Nous gardons donc l'utilité de la donnée en préservant une partie du sens sans divulguer l'information personnelle qu'elle véhicule.

Une attaque de ré-identification a moins de chance d'aboutir avec l'information **Person_Name** qu'avec l'information **Hakim.**

Une première étape consiste en l'identification des noms de personnes contenus dans une session. Ceci est fait en utilisant la base de données des noms de personnes. Si un terme de l'utilisateur est présent dans la base des noms alors il est identifié comme étant un terme à généraliser par **Person_Name.**

#### III-3-2- Généralisation des noms de lieu

La généralisation des noms de lieux se fait selon le schéma suivant :

City → County → State → Country → Place Name

Le niveau de sécurité augmente si en augmentant le niveau de la généralisation. En d'autres termes, lorsqu'une attaque utilise un terme présent dans une session de recherche et représentant le nom d'une ville, l'identité de l'utilisateur est recherchée parmi les personnes ayant la même ville dans leurs adresses. Après application de la généralisation, ce nom de ville est généralisé à un niveau plus haut, il sera par exemple remplacé par le nom du County où se trouve cette ville. Dans ce cas-là une attaque va rechercher l'identité de l'utilisateur dans un espace plus grand qui est celui des personnes habitant le même County.

Le réglage du niveau de généralisation est question de balancement entre l'utilité attendu des données du journal de requêtes et le niveau de protection souhaité.

Dans notre implémentation, nous avons déterminé 4 niveaux de généralisation selon le schéma décrit précédemment.

**Le Niveau_1 :** Toutes les **city** seront généralisées par leurs **counties** respectifs.

**Le Niveau_2 :** Toutes les **city** et **counties** seront généralisées par leurs **states** respectifs.

**Le Niveau_3 :** Toutes les **city, counties,** et **states** seront généralisées par leurs countries respectifs à savoir dans notre cas 'USA'.

**Le Niveau_4 :** Toutes les **city, counties, states** et **countries** seront généralisées par le terme **'place_name'** et c'est le niveau plus haut.

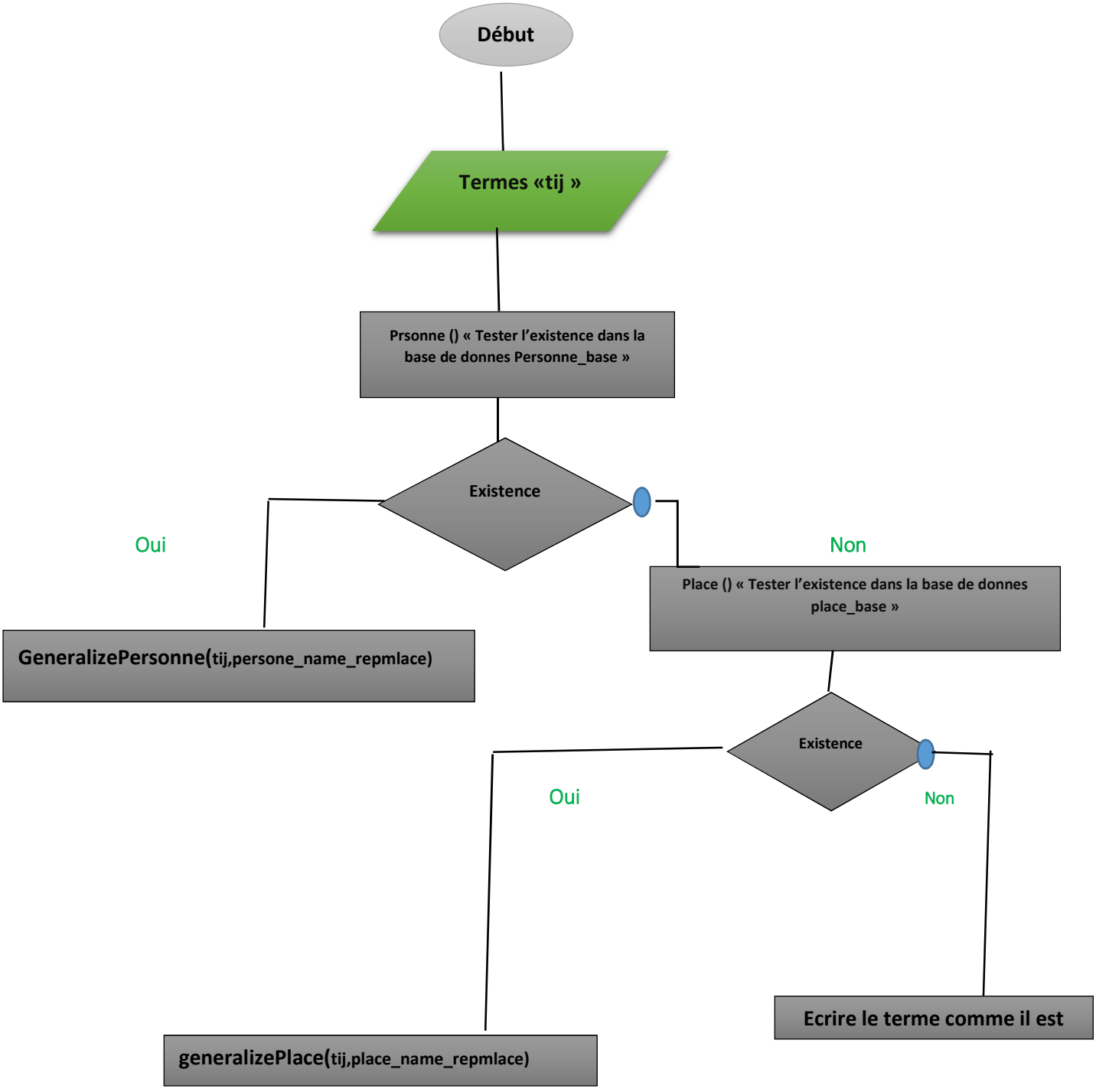
## IV Algorithme et l'organigramme de la solution

### IV-1-L'algorithme:

```
Entré : une liste des l requêtes contenant une liste des J termes tij
Sorté : une liste l de requête dont les termes qui sont des noms propres (lieux,
personnes) sont remplacées par Place_name_replace &
Persone_name_replace ;

Début
Pour (i de 1 à I)
Pour (j de 1 à J)
        Si persone(tij) = personne_base alors
                GeneralizePersone(tij,'persone_name_replace')
        else  Si place(tij)=place_base alors
                Generalize Place(tij,'place_name_replace')
              Fin si
   Fsi
   Finpour
   Fin pour Fin
```

## IV-2-L'organigramme :



## Conclusion

Dans ce chapitre, nous avons donné une vue sur la méthode de la généralisation et comment on la peut utiliser dans notre application. Dans la suive de notre travail on verra la mise en œuvre de notre application développée.

# CHAPITRE V : Application

## Introduction

Après avoir présenté comment on a anonymisé quelque type des informations personnelle avec l'approche qui basée sur le remplacement des noms propres par des noms plus générales et moins spécifiques, maintenant on va représenter les environnements et les outils de construction et de développement de notre application.

## I-L 'environnement matériel et logiciel de notre implémentation

Dans cette section, Nous présenterons l'environnement de travail sur lequel nous sommes basés pour réaliser cette implémentation.

### I-1-Ressources utilisées

Les ressources physiques (hard) utilisées sont :

- Processeur Intel® Core™ i3-2310M CPU d'une fréquence de 2.40 GHz.
- Une mémoire vive d'une capacité de 4 GO.
- Une carte graphique Intel HD de 1664 MB.

Et pour ce qui est côté logiciel (Soft) :

- System exploitation: Windows 8.1 Professionnel.64 bit
- NetBeans IDE version 7.2.1.
- MYSQL Workbench version 5.2.43 CE
- notepad++ version 6.8.7

### I-2-langage de programmation:

Le langage de programmation choisi dans notre travail est le **JAVA** qui reste un des langages les plus utilisés dans le domaine de traitement des fichier texte ainsi dans le domaine des sécurisé des informations.

Java est un langage de programmation informatique orienté objet et un environnement d'exécution informatique portable créé par James Gosling et Patrick Naughton employés de Sun Microsystems avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld. **[21]**

### I-3- NetBeans IDE 7.2.1 :

Le NetBeans est un environnement de développement intégré et disponible pour Windows, Mac, Linux . Le projet NetBeans se compose d'un IDE open source et d'une plate-forme d'application qui permet aux développeurs de créer rapidement des applications Web, d'entreprise, de bureau et mobiles à l'aide de la plate-forme Java, ainsi que PHP, JavaScript, Ajax, Groovy et Grails, et C / C ++.
**[22]**

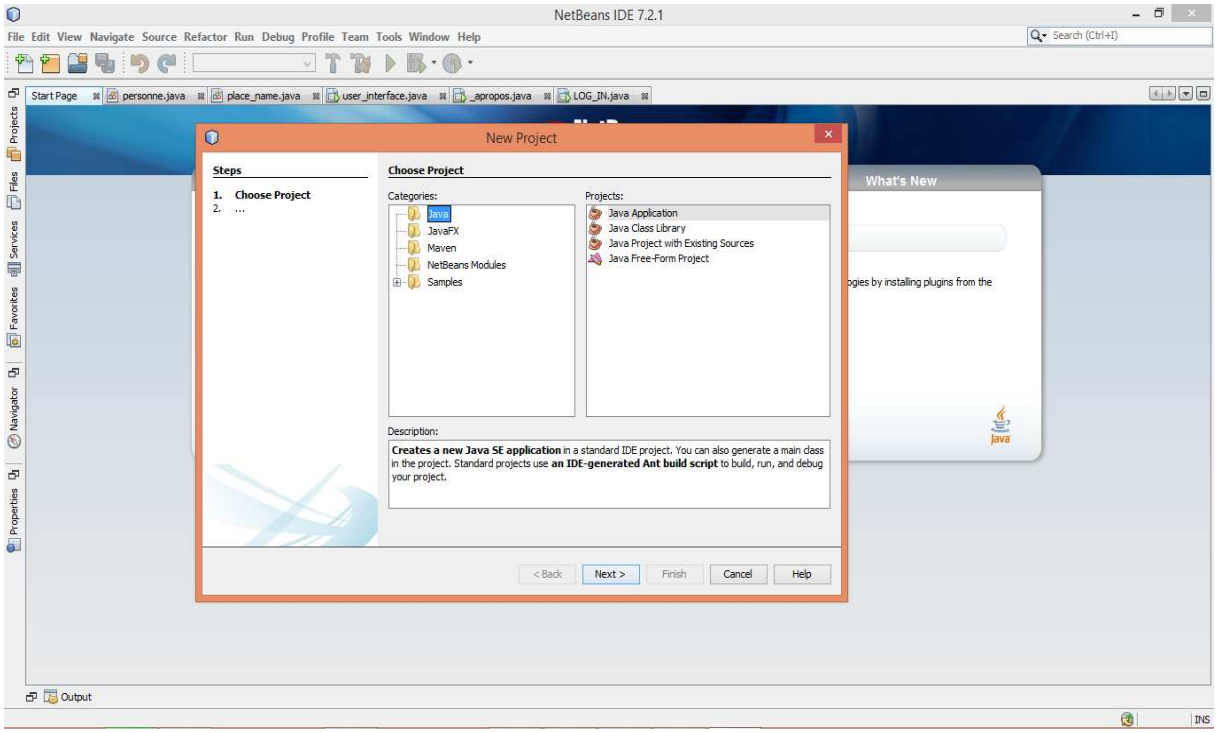



**Figure V-1-Interface du Netbeans IDE**

## I-4- MYSQL Workbench 5.2.43 CE:

C'est un logiciel de gestion et d'administration de bases de données **MySQL**.et c'est un outil visuel pour les architectes de base de données. MySQL Workbench fournit la modélisation des données, de développement SQL et des outils d'administration complets pour la configuration du serveur, l'administration des utilisateurs, des sauvegardes et bien plus encore. MySQL Workbench est disponible avec beaucoup des système d'exploitation : Windows, Linux et Mac OS X…

**[23]**

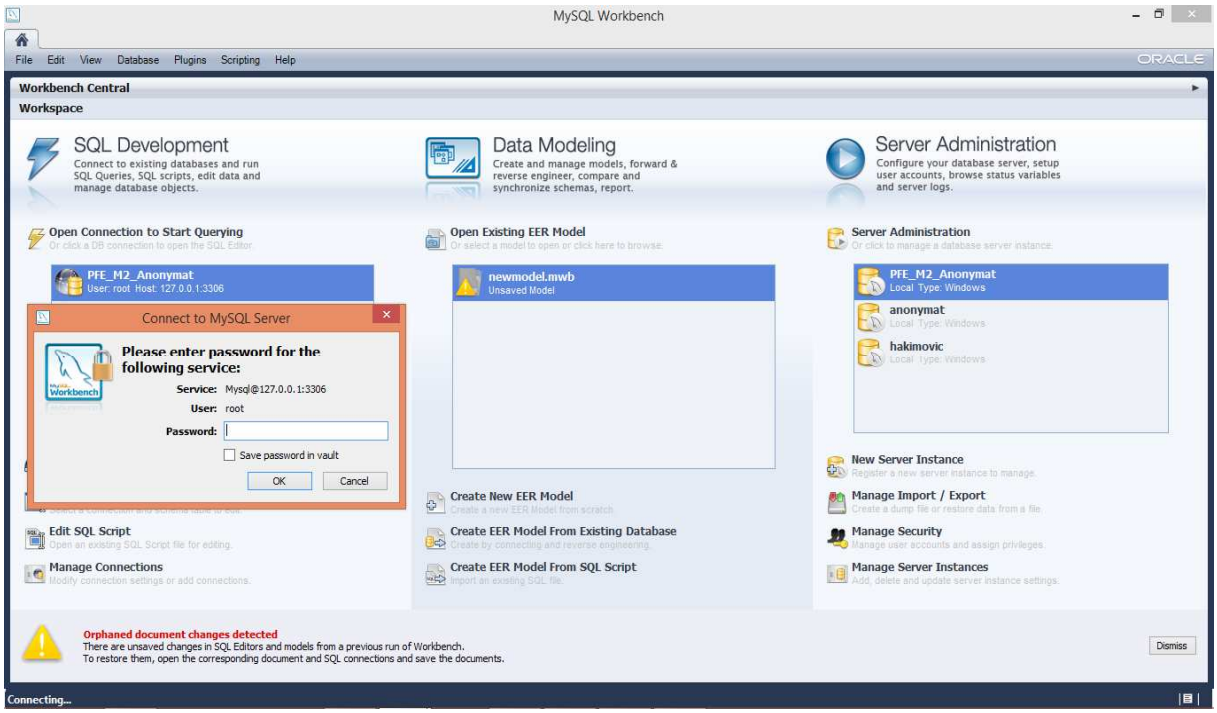



**Figure V-2-Interface du MySQL Workbench**

## I-4-Notepad++ 6.8.7:

Notepad++ est un éditeur de code source qui prend en charge plusieurs langages. Et aussi un éditeur de texte qui lire beaucoup types de texte.

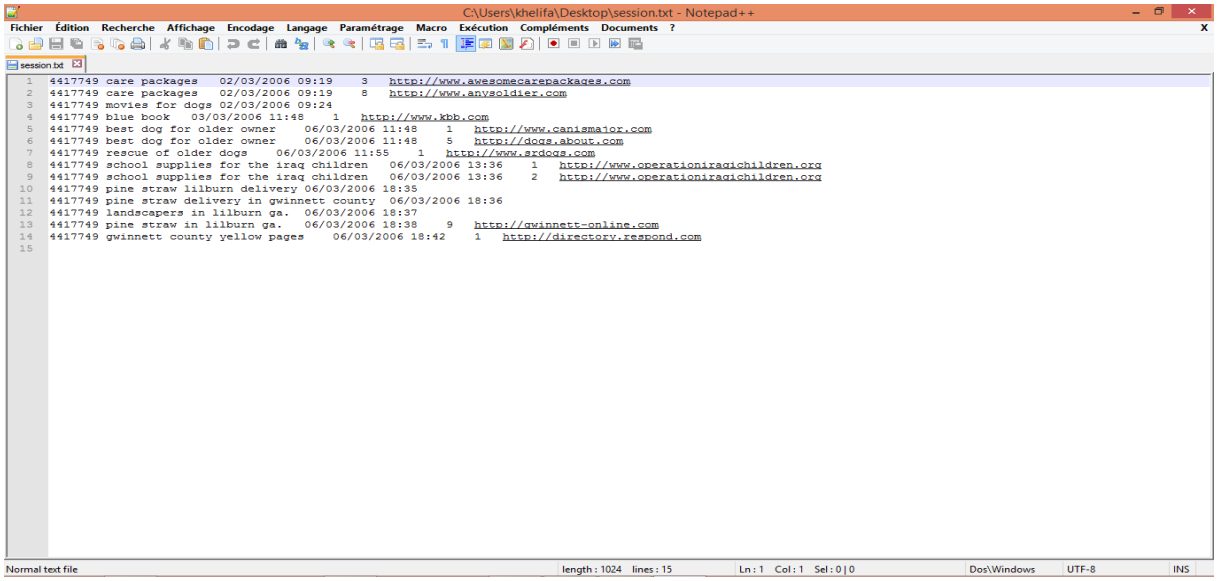



**Figure V-3-Interface du Notepad++ 6.8.7**

## II-Présentation de l'application

### II-1-L'interface d'accueil de l'application :

L'interface d'accueil s'affiche lors de cliquer sur le Botton d'exécutable du projet, dans cette fenêtre l'utilisateur doit choisir le fichier qui veut anonymiser, choisir le type de données qui veut anonymiser ainsi voir le résultat.

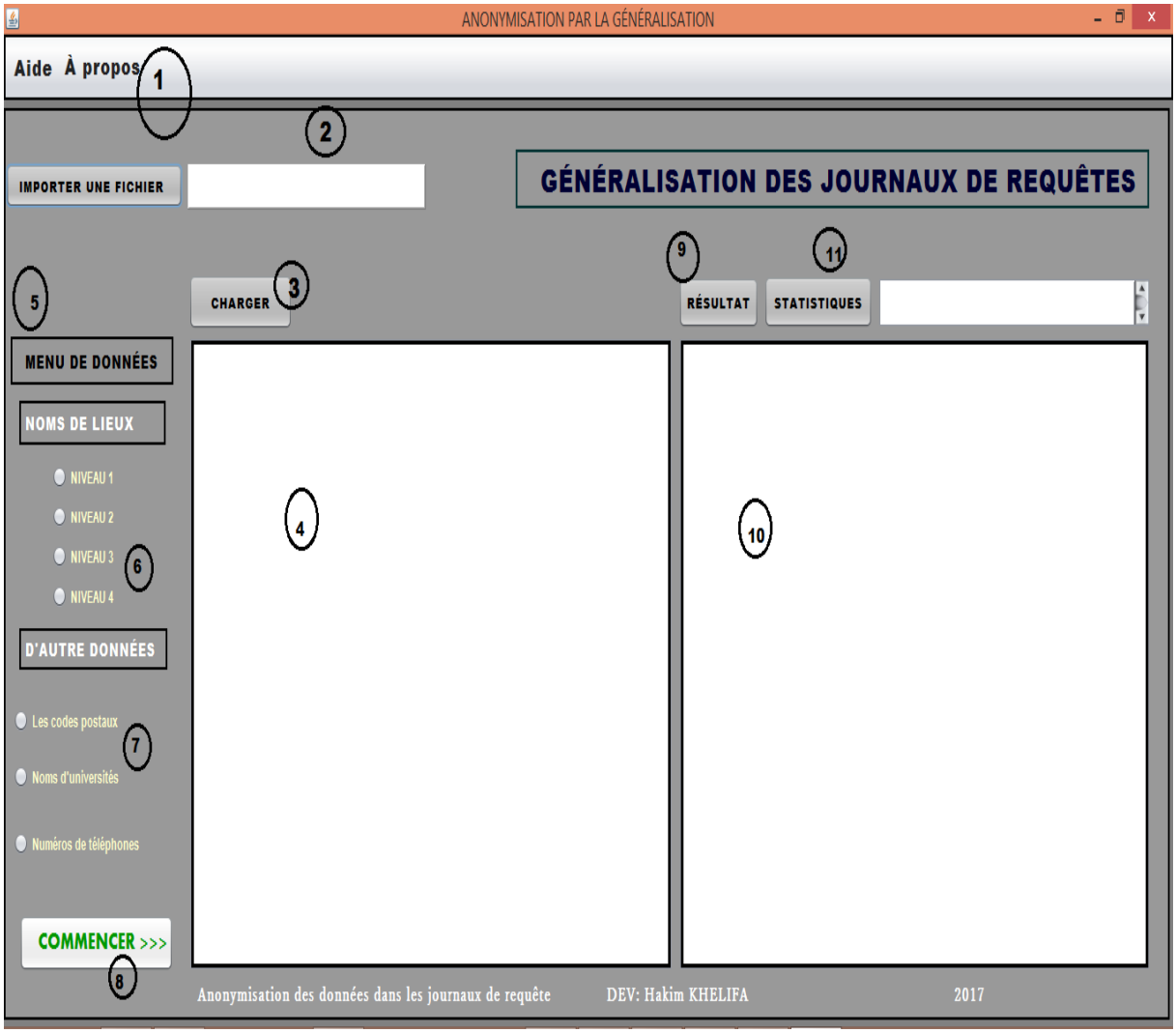



**Figure V-4-Interface d'accueil de l'application.**

**1-La barre de Menu :**

- **Aide** : pour aider les utilisateurs sur l'utilisation de cette application (guide d'utilisation)

- **À propos :** Des informations sur le développeur de cette application.et un système de

sortir. Elle se fait par le clique sur « Aide » sur le Menu.


**Figure V-5-Le guide d'utilisation à partir du Menu (Aide)**

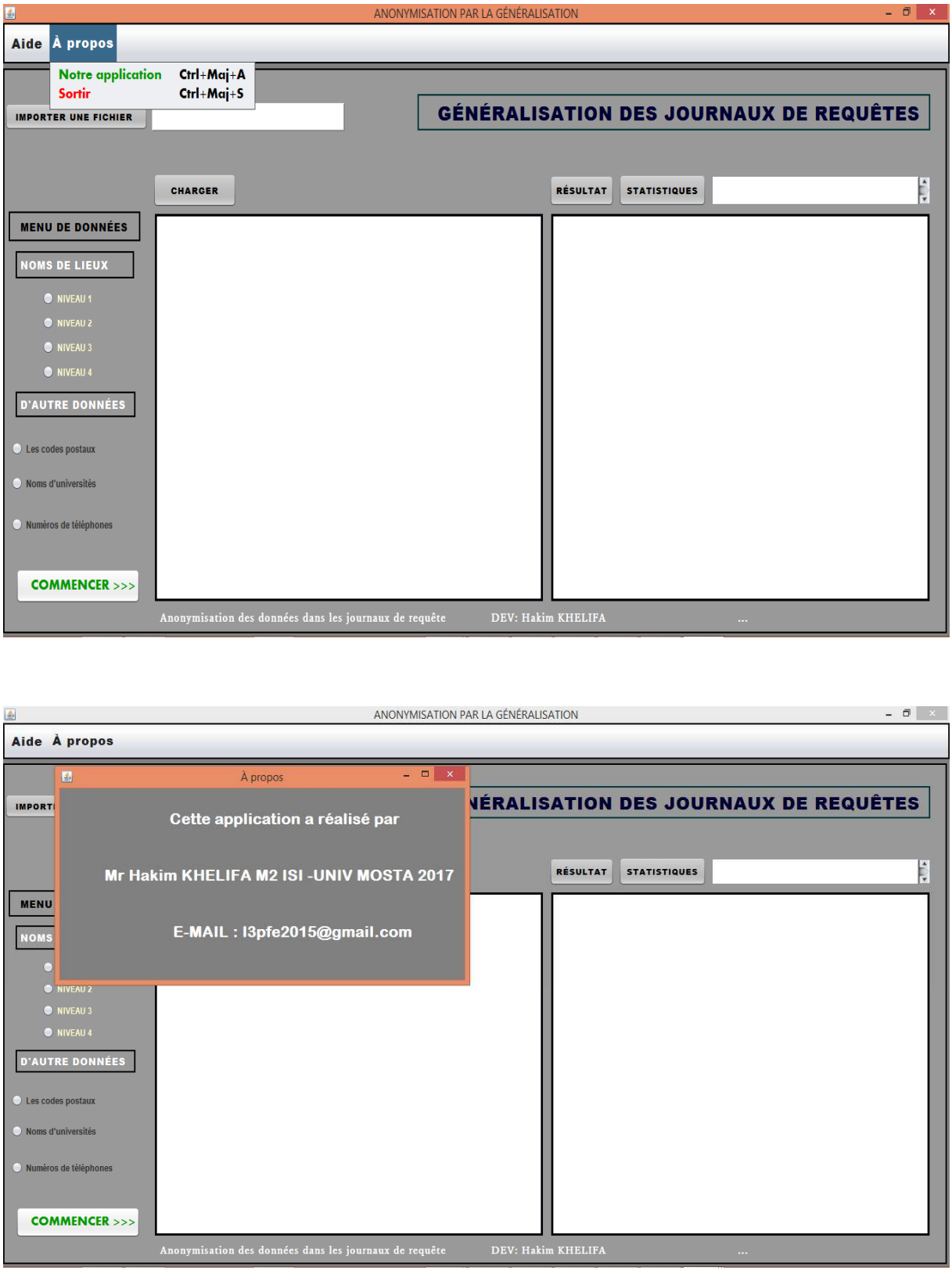



**Figure V-6-L'option à propos et l'option de sortir à partir de Menu A propos**

### 2-importation d'un fichier log (.TXT) et le chargement dans l'interface d'accueil :

Elle se fait par le clique sur le Botton « **IMPORTER UN FICHIER** » **(2)** puis choisir le fichier depuis leurs emplacements. Pour la visualisation du contenu du fichier l'utilisateur peut le charger sur l'afficheur **(4)** par le Botton **« CHARGER » (3).**

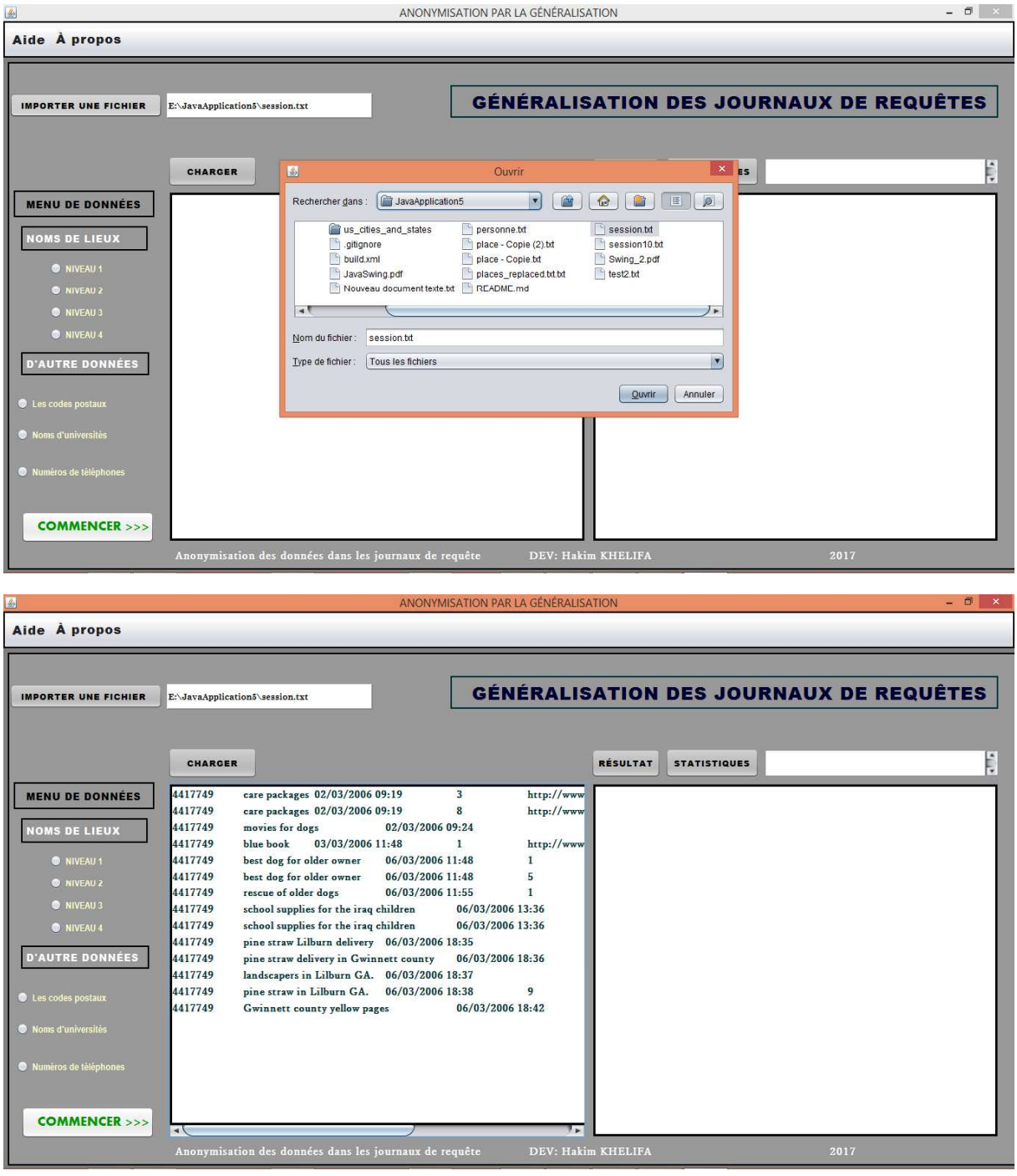
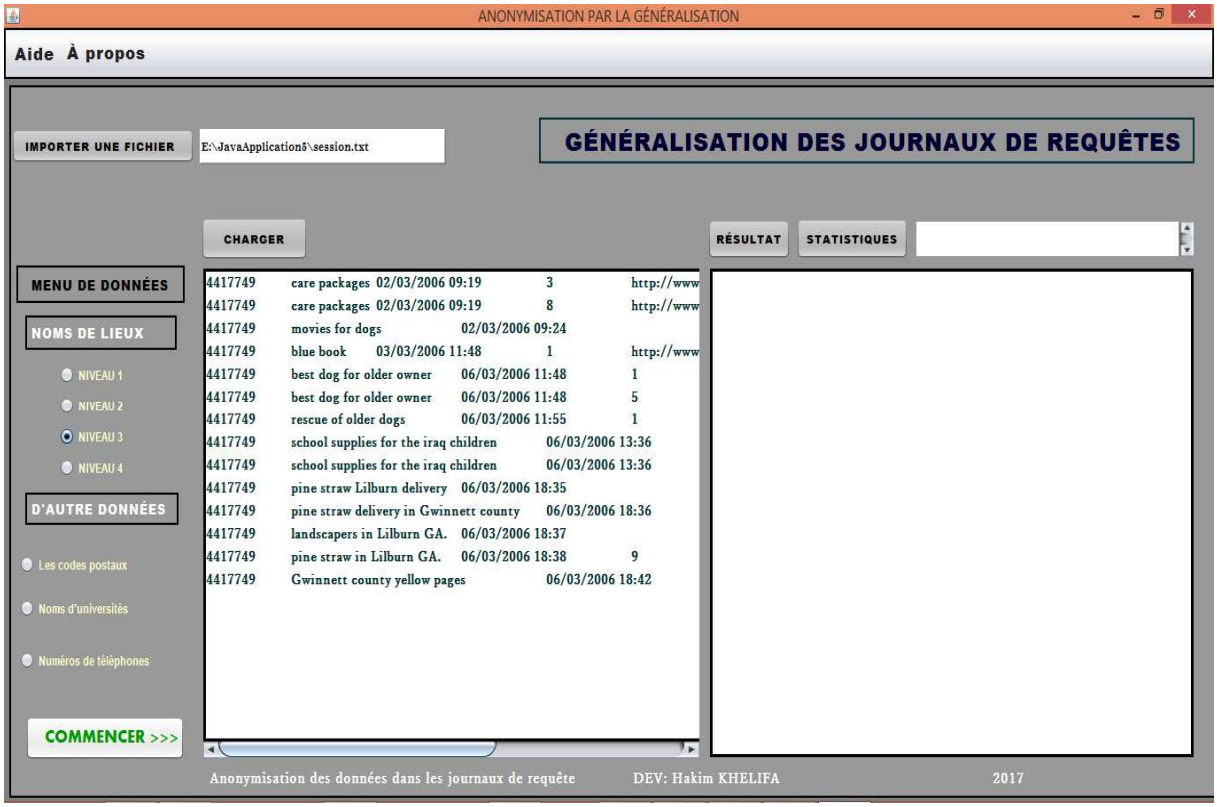



**Figure V-7-Importation et chargement d'une fichier log (. Txt)**

### 3-le chois de type de données pour l'anonymiser ( 5) :

Elle se fait par de cocher le niveau de généralisation des noms de lieux **(6)** ou le choix facultatif des autres types de données (les numéros de téléphones…). **(7)**




**FigureV-8-Les choix des types de données**

### 4-le lancement d'anonymisation et l'affichage du résultat :

Après le choix de type de données**,** l'utilisateur maintenant peut lancer l'anonymisation par le clique sur le Botton **« COMMENCER>> » (8)** et attend quand le message dit *'' **Fin d'anonymisation…… appuyer sur le bouton « RESULTAT » pour voir** ''*.

Après ça l'utilisateur peut voir le résultat par le clique sur le Botton « **RESULTAT** » **(9)** pour l'afficher sur la deuxième afficheur **(10).**

# CHAPITRE V : Application

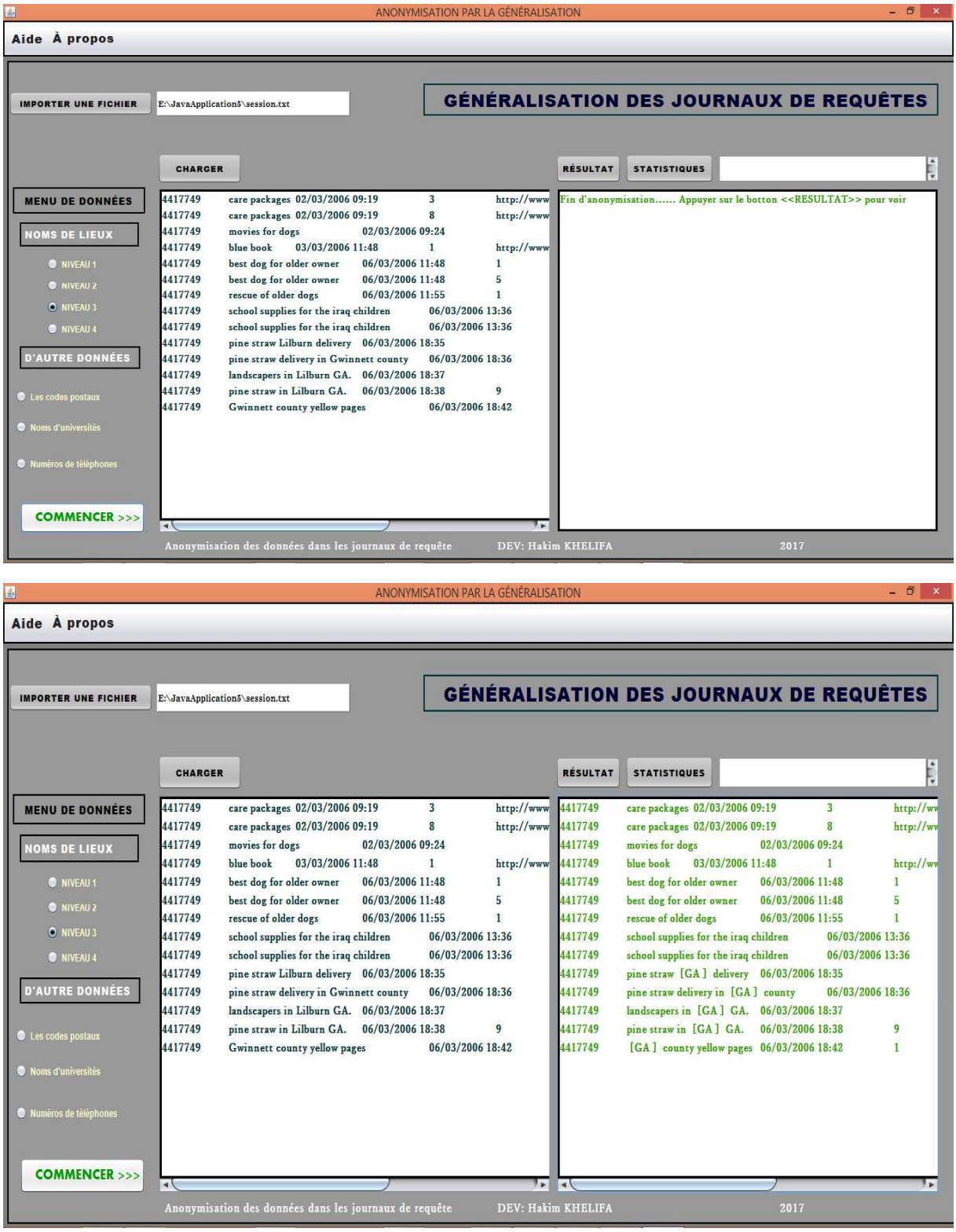



**Figure V-9-Lancement d'anonymisation et le résultat obtenu**

**5-les statistiques :** Elle se fait par le clique sur le Botton « **STATISTIQUE** » **(11)** pour voir combien de remplacements**.**

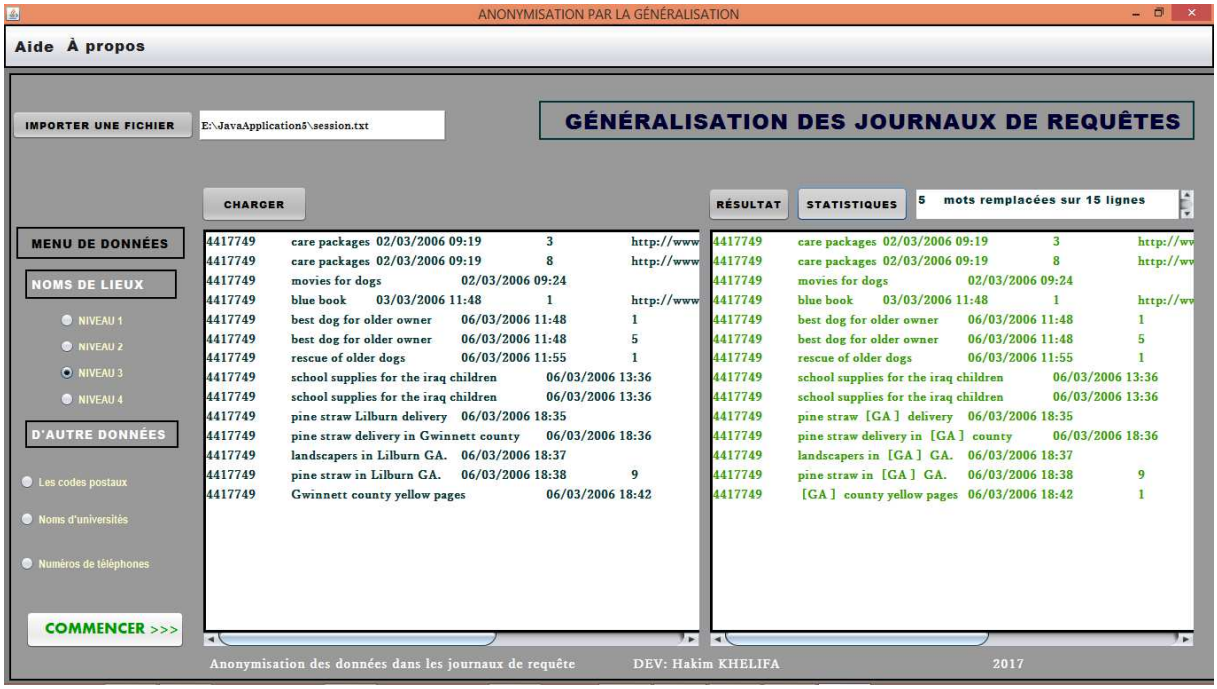



**Figure V-10-L'affichage du statistique d'anonymisation**

## Conclusion

Dans ce chapitre, nous avons décrit l'environnement matériel et logiciel, ainsi les détails des interfaces développées. Et aussi ce qui a été fait dans l'application, on peut dire que notre application répondre certaines caractéristiques :

- Une interface simple à utiliser.
- L'application de la méthode de la généralisation des noms de lieux et les noms de personnes.
- Une bonne performance des résultats par la méthode de la généralisation par rapport d'autre solution (le hachage, décomposition des sessions).

## Conclusion générale

Le problème de la divulgation d'identité des utilisateurs d'un moteur de recherche prend beaucoup des dimensions sociales. Dans ce mémoire nous avons étudié Cette problématique en commençant d'abord par comprendre les intérêts de création de fichier logs, ces fichiers où figure l'enregistrement des recherches effectuées par des utilisateurs dans un moteur de recherche. Nous avons réussi à comprendre comment se fait la ré identification des personnes à travers leurs recherches sur le web. Ensuite nous avons étudié les solutions d'anonymisation proposées dans la littérature.

Nous avons ensuite décrit notre solution pour la protection des données dans un journal de requêtes. Cette solution se base sur la généralisation des termes pouvant identifier une personne. Parmi d'autres, nous avons traité les noms de personnes et les noms de lieux. Rien n'empêche d'étendre cette solution à d'autres termes liés à l'identité d'une personne.

Quant à l'implémentation de notre solution nous avons ajouté la notion de modération de la généralisation selon des niveaux de sécurité afin de réaliser un bon compromis protection/utilité des données.

Comme perspectives à nos travaux actuels, nous proposons d'ajouter la généralisation d'autres types d'informations personnelles. Ainsi qu'intégrer un mécanisme de mesure permettant d'évaluer le niveau de protection et d'utilité garantit par la solution.

# Bibliographie :

**[1].** Simon Leva. (20 APR 2015). Les sessions de recherche comme contexte des requêtes. CLLE-ERSS : CNRS et Université de Toulouse (UMR 5263). (Pp.4-5).

**[2].** Eytan Adar.User 4XXXXX9 : Anonymizing Query Logs. University of Washington, Computer Science and Engineering eadar@cs.washington.edu.

**[3].** Linda Pesante. (2008). Introduction to Information Security. Carnegie Mellon University. (Pp.1-2).

**[4].** Sattarova Feruza Y. and Prof.Tao-hoon Kim.( 2, April, 2007). IT Security Review: Privacy, Protection, Access Control, Assurance and System Security. Hannam University, Department of Multimedia Engineering, International Journal of Multimedia and Ubiquitous Engineering Vol. 2. (Pp.17).

**[5].** Équipe commerciale France.( Sophos Ltd. Tous droits réservés).(2011). Protection des données personnelles, Quelles sont les données vulnérables et comment pouvez-vous les protéger ? Boston, États-Unis | Oxford, Royaume-Uni.(Pp.1-2).

**[6].** Bradley Malin CMU-ISRI. (06-105May 2006). Trail Re-identification and Unlinkability in Distributed Databases. Institute for Software Research, International School of Computer Science Carnegie Mellon University Pittsburgh.

**[7].** Mlle.CHARIF Ismahan. (2013). La protection de la vie privée sur Internet :Application sur les données personnelles. Université Abou Bakr Belkaid– Tlemcen, Faculté des Sciences, Département d'Informatique. (Pp.7).

**[8].** Li Xiong Eugene Agichtein.Towards PrivacyPreserving Query Log Publishing. Mathematics and Computer Science (Department, Emory University

**[9].** Amanda Spink & Bernard J. Jansen. People's Query Logs: Personal Information Management. Queensland University of Technology Gardens Point Campus, The Pennsylvania State University 2P Thomas Building.

**[10]**. Amin Milani Fard.Privacy Preserving Web Query Log Publishing: A Survey on Anonymization Techniques Simon Fraser University, Burnaby, Canada, University of British Columbia, Vancouver

**[11].** Benjamin NGUYEN. (Décembre 2014) Techniques d'anonymisation. Insa1 Centre Val de Loire et Inria2 Paris-Rocquencourt.

**[12].** Mme MIMI Anissa. (2011). Protection de l'anonymat des utilisateurs dans les Query Logs. Université d'Oran Es-Senia, faculté des science, département d'informatique. (Pp .54-66)

**[13].** 17 Göker et He (2000). Définition de la notion de session en recherche d'information

**[14].** Jansen et al (2007). Définition de la notion de session en recherche d'information

**[15].** Silverstein et al(1999 ) . Définition de la notion de session en recherche d'information

**[16].** Spink et al (2006) . Définition de la notion de session en recherche d'information

**[17].** Jones et Klinkner (2008). Définition de la notion de session en recherche d'information

**[18].** Gayo Avello (2009). Définition de la notion de session en recherche d'information

**[19].** Lucchese et al (2011). Définition de la notion de session en recherche d'information

**[20].** Sweeney, Latanya. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty Fuzziness and Knowledge-based Systems

**[21].** http://ipeti.forumpro.fr/t21-definition-de-langage-java-java-script, consulté le 25 avril 2017

**[22].** https://netbeans.org/community/releases/72 , consulté le 25 avril 2017

**[23].** https://www.mysql.com/fr/products/workbench ,consulté le 25 avril 2017

**[24].** http://www.farinspace.com/us-cities-and-state-sql-dump, Consulté et téléchargé le 01 avril 2017

**[25].** https://www.drupal.org/project/namedb, consulté et téléchargé le 27 mars 2017

**[26].** http://www.cim.mcgill.ca/~dudek/206/Logs/AOL-user-ct-collection, Consulté le 06 mai 2017