

Master of Science HES-SO in Engineering
Av. de Provence 6
CH-1007 Lausanne

Master of Science HES-SO in Engineering

Orientation : Technologies de l'information et de la communication

Web digital footprints and data privacy

Fait par

Kewin Dousse

Sous la direction de
Prof. Fatemi Nastaran
à la HEIG-VD

Félicien Fleury (NGSENS)

Lausanne, HES-SO//Master, 2017

Résumé

Le but de ce projet est de concevoir et d'implémenter un outil d'analyse de comportement d'utilisateurs d'applications Web pour révéler les potentiels de détection de profile des personnes (préférences, centre d'intérêt, orientations et opinions) en analysant les interactions et les informations échangées avec les applications Web.

Keywords. Web, Big Data, Privacy, Profiling

Table des matières

1	Introduction	7
1.1	Contexte	7
1.2	Objectifs	7
1.3	Méthodologie	8
2	Analyse	9
2.1	Données de l'étude de Kosinski	9
2.1.1	Introduction	9
2.1.2	Résumé	9
2.1.3	Données	11
2.1.4	Acquisition	12
2.1.5	Conclusion	12
2.2	SDIPI	13
2.3	Trackers et Google Analytics	14
2.3.1	Trackers	14
2.3.2	Etat de l'art	14
2.3.3	Google Analytics	16
2.4	Extensions de navigateur	16
2.4.1	Introduction	16
2.4.2	Etat de l'art	16
2.4.3	Conclusion	20
3	Conception	22
3.1	Introduction	22
3.1.1	Idée	22
3.1.2	Architecture	23
3.1.3	Données	24
3.2	Extension	25
3.3	Serveur	26
3.3.1	Rôles	26
3.3.2	Récolte	26

3.3.3	Enregistrement	26
3.3.4	Traitement des données	27
3.3.5	API	27
3.4	Interface	28
3.4.1	Profil	28
3.4.2	Trackers	28
4	Implémentation	34
4.1	Extension	34
4.1.1	Technologies	34
4.1.2	Structure	34
4.2	Serveur	34
4.2.1	Technologies	34
4.2.2	Structure	34
4.3	Interface	34
4.3.1	Technologies	34
4.3.2	Structure	35
5	Résultats	36
5.1	Validation	36
5.1.1	Tests utilisateurs	36
5.2	Résultats de la recherche	36
5.2.1	Statistiques	36
5.2.2	Implications	36
5.3	Conclusion	36
6	Conclusion	37
6.1	Conclusion du projet	37
6.1.1	Délivrables	37
6.1.2	Conclusion générale	37
6.1.3	Perspectives	37
6.2	Conclusion personnelle	38
A	Historique des versions	44
B	Cahier des charges	45
B.1	Activités	45
B.2	Planification	45
B.3	Diagramme de Gantt	46

C Documentation	47
C.1 Localisation	47
C.2 Contenu	47
C.2.1 GitLab	47
D Procès-verbaux	48

Table des figures

2.1	Précision moyenne du modèle prédisant la personnalité d'un utilisateur en fonction du nombre de likes analysés[5].	10
2.2	Déviation de la personnalité moyenne estimée d'un visiteur régulier du site ““deviantart.com”” selon les cinq axes psychologiques employés[5].	11
2.3	Page d'accueil du site web https://sdipi.ch	13
2.4	Nombre moyen de domaines contactés au chargement d'une page web[20].	14
2.5	Marché occupé par Google Analytics dans les domaines d'analyse, de tracking et de mesure d'audience sur le Web.	15
2.6	Logo de la solution Google Analytics[8].	15
2.7	Image de présentation de timeStats[12].	17
2.8	Page d'accueil de Ghostery[13].	17
2.9	Interface de base de Privacy manager[14].	18
2.10	Interface de TheGoodData[15].	19
2.11	Premier paragraphe de la page web de Noiszy[16].	20
2.12	Contrôle des actions face aux trackers de Privacy Badger[17].	21
2.13	Flux de données de Kraken.me[18].	21
3.1	Flux de données de l'extension.	23
3.2	Exemple d'URL et traitement	24
3.3	Téléchargement et enregistrement du contenu d'une page	29
3.4	Traitements du contenu des pages	30
3.5	Maquette de la page de Profil	31
3.6	Maquette de la page des Trackers	32
3.7	Suite de la maquette des Trackers	33

Chapitre 1

Introduction

1.1 Contexte

En janvier 2014, l'ONG Internet Society a publié le document Digital footprints[3] qui aborde la question de la capacité que les web trackers ont de définir le profil personnel des utilisateurs d'Internet.

En 2016, Michal Kosinski[1], chercheur à Stanford, révèle les possibilités de définir un profil précis simplement en analysant les préférences (likes) enregistrées dans un profil Facebook[2]. L'étude révèle que ce type d'analyse permet de mieux connaître une personne que ses proches et même de prévoir de probables comportements avec une grande précision. De plus, lors d'événements politiques majeurs ces techniques de profiling auraient été utilisées, comme dans le cadre des campagnes pour le Brexit ou pour l'élection du président américain Trump.[4]

1.2 Objectifs

Le but de ce projet est de concevoir et d'implémenter un outil d'analyse de comportements d'utilisateurs d'applications Web pour révéler les potentiels de détection de profils des personnes (préférences, centre d'intérêt, orientations et opinions) en analysant les interactions et les informations échangées avec les applications Web. L'application développée dans ce projet a pour le but principal de sensibiliser le public et les médias à la question du profiling sur internet.

L'objectif technique du projet est de développer un plugin pour les navigateurs Mozilla Firefox et Google Chrome qui permettraient de :

1. Définir un profil utilisateur selon des critères de préférence, d'intérêt, d'habitude, d'opinion, etc.
2. De définir le profil d'un usager en se basant sur sa navigation sur Internet ainsi que sur les métadonnées (durée de consultation des pages, heure de

consultation, etc.). Des algorithms de machine learning seront utilisés pour apprendre les profils en se basant sur des collections de profils annotées telle que la collection kaggle[6].

3. Identifier des trackers qui ont la possibilité de construire des profiles utilisateurs en intégrant des données de plusieurs sources.

1.3 Méthodologie

Le développement du code sera open-source. Le déroulement du projet sera divisé en deux phases distinctes :

1. La première phase du projet consistera en une analyse des études et résultats actuels afin de proposer des concepts innovants à travers l'outil développé, tout en collectant les données des utilisateurs pour la deuxième phase.
2. La seconde phase mettra l'accent sur les données récoltées par le plug-in développé durant la première phase : Le but sera d'analyser les données et d'en tirer des conclusions intéressantes.

Chapitre 2

Analyse

2.1 Données de l'étude de Kosinski

2.1.1 Introduction

Michal Kosinski se présente sur son site web[1] comme un "psychologist and data scientist". L'étude qu'il a co-rédigée à l'Université de Stanford en 2016 a eu un impact important sur le monde académique et même industriel, en montrant les possibilités techniques ouvertes par la récolte de données simples d'utilisateurs : les "likes" Facebook.

Ainsi, il est montré qu'avec un peu plus de 300 "likes" tirés une personne, il est possible de définir avec une précision remarquable (mieux que son époux/épouse) des traits psychologiques, ainsi que d'autres caractéristiques personnelles.

2.1.2 Résumé

Une enquête a été menée auprès d'une population variée de personnes possédant un compte Facebook. Les données concernant leurs "likes" ont été récoltées, ainsi que des données personnelles pouvant être disponible (ou non) selon le souhait de l'utilisateur sur Facebook, comme ses informations démographiques. Des tests psychologiques ont été également réalisés par une certaine partie des utilisateurs afin de pouvoir trouver des corrélations entre les pages likées et certains traits psychologiques.

Cette enquête a rencontré un succès très large, et le nombre de personnes ayant répondu à l'enquête, au moins en partie, se compte en millions.

Les résultats présentés à la fin de l'étude sont inattendus : Michal annonce qu'il est possible de prédire certains comportements d'une personne mieux que son entourage le plus proche.

	Précision	Nombre de likes
Collègue	0.27	10
Ami	0.44	80
Famille	0.5	100
Epoux/se	0.58	250

TABLE 2.1 – Précision atteinte par type de relation avec une personne, et nombre de likes nécessaires au modèle pour égaler sa précision

Un des modèles créés avec les données récoltées, permet d'estimer le profil psychologique d'un participant selon cinq axes différents, en se basant sur ses likes Facebook. La figure 2.1 montre la précision obtenue par le modèle en fonction du nombre de likes utilisé en entrée.

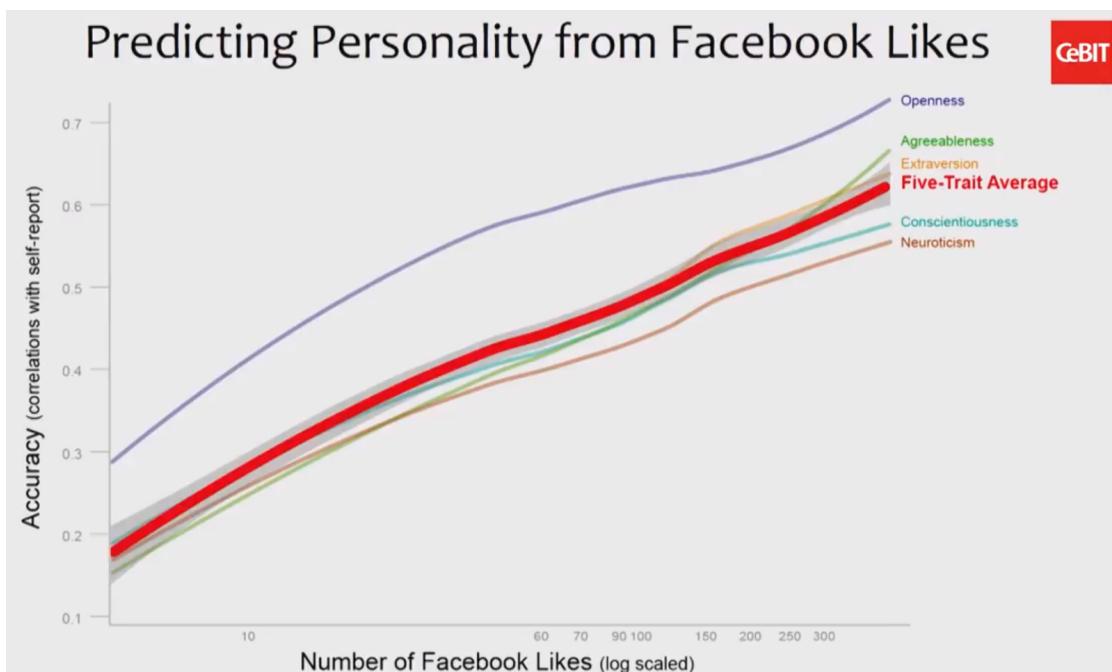


FIGURE 2.1 – Précision moyenne du modèle prédisant la personnalité d'un utilisateur en fonction du nombre de likes analysés[5].

On remarque que la précision de la prédiction de tous les critères augmente avec le nombre de likes utilisés, ce qui n'est pas surprenant. En revanche, le tableau 2.1 montre le lien entre le nombre de likes utilisés et la précision moyenne atteinte par l'algorithme, et compare ces valeurs à la précision atteinte par d'autres êtres humains.

On peut voir que la précision de la prédiction de l'algorithme surpassé celle

même l'époux/se d'une personne avec 250 likes, ce qui se trouve être légèrement au-dessus du nombre de likes moyen par personne, qui est de 227.

Les possibilités de prédiction du modèle ne se limitent pas à une simple personne, et les possibilités sont nombreuses. Par exemple, Michal montre qu'il est possible de montrer une corrélation entre les visiteurs d'un certain site web, et une tendance vers certains traits psychologiques. La figure 2.2 montre la personnalité moyenne estimée des visiteurs du site web ““deviantart.com”” par rapport à la moyenne de tous les utilisateurs.

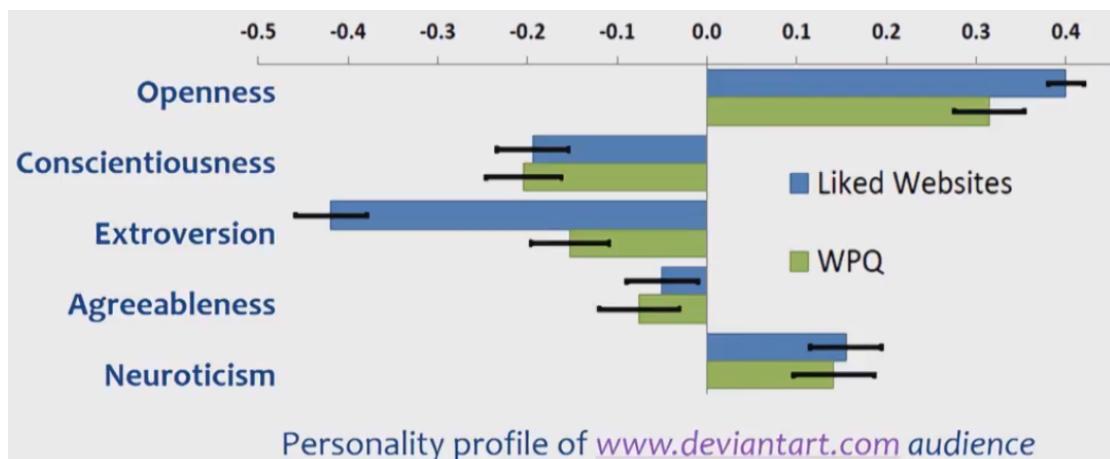


FIGURE 2.2 – Déviation de la personnalité moyenne estimée d'un visiteur régulier du site ““deviantart.com”” selon les cinq axes psychologiques employés[5].

Ces corrélations ne sont que quelques exemples parmi un très large éventail de possibles corrélations que le modèle est capable de mettre en lumière. Les implications de telles découvertes sont massives : Il serait par exemple possible de déterminer si un utilisateur sera réceptif ou non à un certain type de publicité, par exemple. Ce genre de problématique touche à plusieurs domaines et n'est pas exactement de notre ressort ici : Des principes éthiques sont en jeu, et le sujet devient de plus en plus délicat. Mais une chose est certaine : Des likes Facebook peuvent révéler énormément d'informations.

2.1.3 Données

La quantité de données amassée par l'étude est massive. Non seulement en quantité d'utilisateurs, mais également en diversité de données. Michal Kosinski a mis en place le site web "myPersonnality Project"[7] permettant de partager cette source de données avec d'autres chercheurs. Les données comprennent, entre autres :

- Scores de personnalité selon la méthode BIG5 de >3 millions de personnes
- Données démographiques de >4 millions de personnes
- Localisation géographique de >1.5 million de personnes
- Vues politiques de >500'000 personnes
- Likes Facebook de >19 millions de personnes

Le type de données présenté ici n'est qu'un sous-ensemble restreint de l'ensemble des tables présentées, bien qu'il s'agisse ici des données comprenant le plus d'entrées au total.

2.1.4 Acquisition

Bien que l'objectif du site web soit de partager l'accès à cette énorme base de données, l'accès à celle-ci est loin d'être aisé. Tout d'abord, Kosinski ne met ces données à disposition que de milieux académiques, il interdit l'utilisation de ces données à des fins commerciales.

Cependant l'accès n'est pas donné pour autant : Une demande d'accès est à lui envoyer, comprenant une présentation du projet et de ses buts par le biais d'un mail ainsi que le remplissage et l'enregistrement du projet de recherche sur des sites spécialisés.

Cette étape ne semblait constituer qu'une étape nécessitant un temps restreint, mais un prérequis à l'envoi d'une demande d'accès à la base de données est l'approbation de l'"IRB" (Institutional Review Board), ce qui correspond à un comité d'éthique.

2.1.5 Conclusion

Etant donné les délais estimés de l'envoi de la demande à un comité d'éthique responsable puis de la demande d'accès aux données à Kosinski, nous avons écarté cette source de données de la liste principale du projet car nous n'avions pas l'assurance de disposer des données à temps pour la suite de l'étude. Bien qu'il s'agisse certainement d'un ajout conséquent aux données amassées par le projet, nous ne pouvons pas nous permettre de mettre en péril tout l'agenda du projet sur cette source de données.

Bien que cette base de connaissance ait pu être utile, notre étude va changer de direction. Nous décidons de baser la recherche sur des données que nous récupérerons nous-même.

2.2 SDIPI

SDIPI signifie "Swiss Digital Identity and Privacy Institute", pouvant se traduire par "Institut Suisse de l'Identité Digitale et de la Vie Privée". Il s'agit d'une association créée dans le but initial de soutenir le projet dans sa visibilité et dans sa légitimité, mais qui aspire à des objectifs généraux plus larges : Le but est de sensibiliser le public Suisse à la manière dont ses informations privées sont enregistrées, traitées, croisées et utilisées.



What is the SDIPI?

SDIPI is the shorthand for "Swiss Digital Identity and Privacy Institute".

As the Internet grew over the years, Digital identity is an even more important part of a lot of people's lives. As of today, decisions and careers can be decided by how you appear to people online. It's no secret that it's become increasingly more important to know how to control this aspect of one's appearance.

This association was created around the start of a Master's thesis project : [Web Digital Footprints and Data Privacy](#). The goal of this project was to raise awareness about how private informations are handled online, and the study to appeal to the general Swiss population. The three people involved with this project wanted to make this possible and visible through an independant and non-profit medium : That's how the idea of creating this association came to life.

News page.' Column 2: 'Our work' (pencil icon), text: 'We want to raise awareness about how private data are handled online, what kinds of footprints people leave, and how they can control how they appear to the web.' Column 3: 'Want to become a member?' (person icon), text: 'If you want to become a member, go to the [Contact](#) page and send us a message !' A note at the bottom left says 'Website sources available on [GitHub](#)'."/>

FIGURE 2.3 – Page d'accueil du site web <https://sdipi.ch>.

2.3 Trackers et Google Analytics

2.3.1 Trackers

Un tracker est un serveur contacté lors du chargement d'une page web par un utilisateur. De nos jours, les pages web sont souvent constituées de contenu provenant de plusieurs serveurs ou domaines différents. Il n'est pas rare qu'une seule page web fasse appel à plus d'une dizaine de domaines différents pour charger une seule page. La figure 2.4 montre l'évolution du nombre moyen de domaines contactés pour le chargement d'une seule page web, sur les 1'000 sites web les plus visités mondialement.

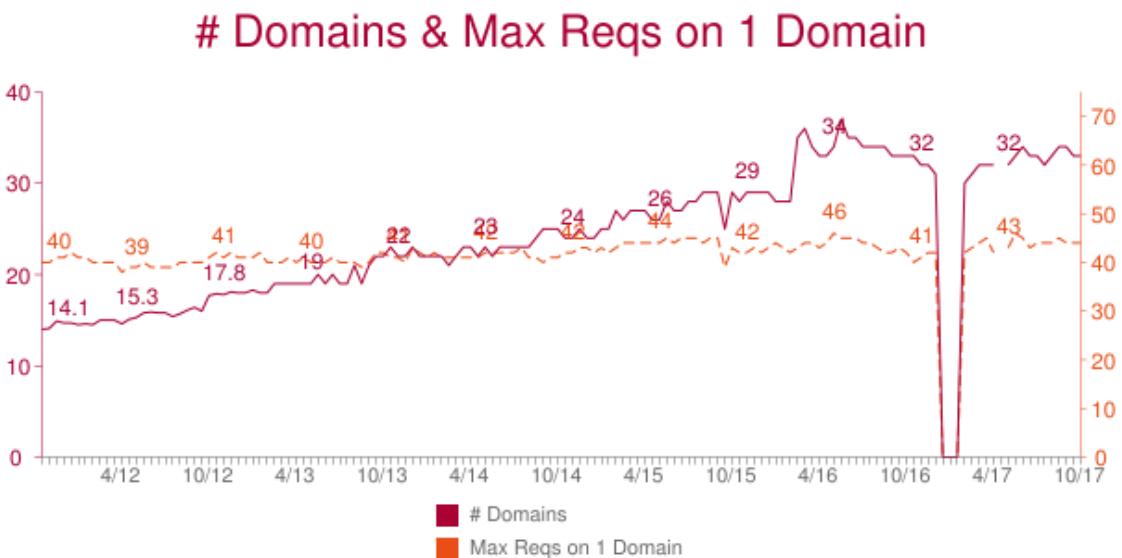


FIGURE 2.4 – Nombre moyen de domaines contactés au chargement d'une page web[20].

Bien qu'une partie des domaines soient nécessaires à contacter afin de charger du contenu indispensable à la page, une partie d'entre eux ne sert également qu'à des fins statistiques ou publicitaires. Par exemple, ceux-ci peuvent récupérer des informations sur l'utilisateur et son navigateur afin de lui proposer des publicités ciblées sur ses intérêts. Cette pratique est aujourd'hui courante, comme le montre la prochaine sous-section.

2.3.2 Etat de l'art

Etant donné que nous nous intéressons aux données des utilisateurs récupérées lors de la navigation Web, nous nous sommes intéressés à connaître quels sont les plus grands trackers sur le web.

La figure 2.5 montre la part de marché qu'occupe Google Analytics ainsi que ses compétiteurs sur les sites web Suisses.

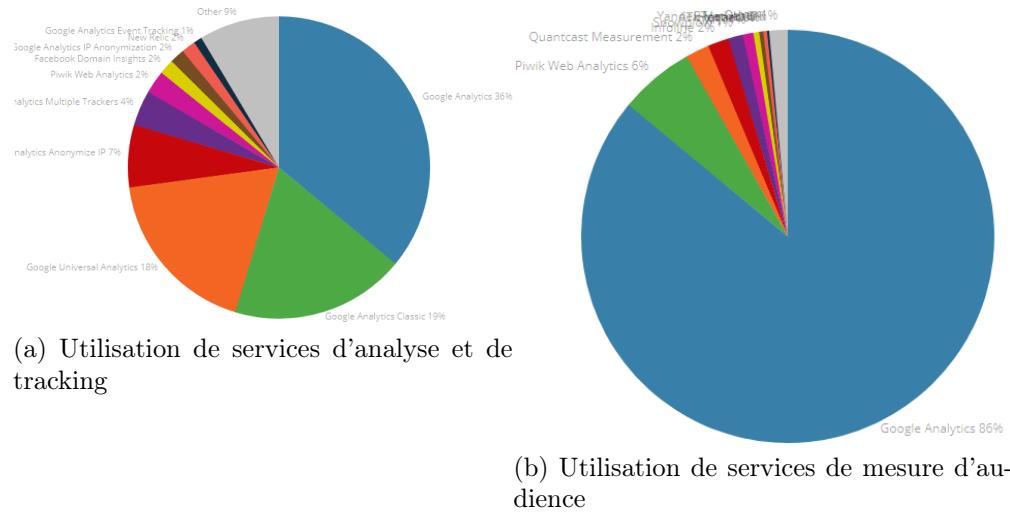


FIGURE 2.5 – Marché occupé par Google Analytics dans les domaines d'analyse, de tracking et de mesure d'audience sur le Web.

Nous pouvons calculer grâce au premier graphique que l'ensemble des produits de Google, y compris Google Analytics et ses versions proches, représentent plus de 83% des installations de solutions dans le domaine de l'analyse et du tracking. De plus pour la sous-catégorie du marché de la mesure d'audience uniquement, Google Analytics a lui seul représente 86% d'installations sur le Web.



FIGURE 2.6 – Logo de la solution Google Analytics[8].

Il est donc de plus en plus évident que s'intéresser aux fonctionnalités de Google Analytics est intéressant pour les buts du projet. Nous souhaitons nous poser la question du risque encouru par les utilisateurs en se connectant sur un site web utilisant Google Analytics. Quelles informations sont prélevées ? Lesquelles sont envoyées ? Les données sont-elles anonymisées ?

2.3.3 Google Analytics

Google Analytics se présente comme une solution d'analyse de statistiques d'utilisateurs dans le but d'améliorer les résultats des sites web sur lesquels il est installé. Ce produit étant totalement gratuit pour les PME, il est aujourd'hui très répandu sur le net et particulièrement en Suisse[9].

2.4 Extensions de navigateur

2.4.1 Introduction

Au vu de l'objectif du projet qui est à la fois de récolter des données tout en montrant un feedback à l'utilisateur, l'extension pour navigateurs est le moyen le plus facile à la fois pour nous de distribuer notre code, et pour les utilisateurs de l'installer. Cependant, de nombreuxses extensions dont le but est de montrer des statistiques sur la navigation de l'utilisateur existent déjà. L'objectif n'est donc pas seulement d'implémenter les mesures adéquates pour notre étude, mais également de fournir des fonctionnalités à l'utilisateur novatrices afin que l'extension se démarque des concurrents.

Une analyse des extensions existantes est donc requises afin de prendre des décisions sur la direction que vont prendre les fonctionnalités implémentées.

2.4.2 Etat de l'art

Nous nous intéressons aux extensions disponibles pour deux des navigateurs les plus utilisés : Google Chrome, et Mozilla Firefox. Chaque navigateur possède son propre éventail d'extensions, bien que parfois certaines se retrouvent disponibles dans les deux catalogues. Chrome Web Store[10] est le catalogue officiel d'extensions pour Google Chrome, et Modules Firefox[11] est celui correspondant à Mozilla Firefox. Quelques recherches avec des mots-clé adaptés sur chaque catalogue vont nous fournir les extensions les plus populaires pour un thème semblable aux nôtre.

timeStats

timeStats[12] est une extension disponible pour Google Chrome. La figure 2.7 montre comment l'extension se présente via une image montrée sur le Google chrome Store.

Cette extension se focalise sur la visualisation du temps passé sur les différents sites web, parfois regroupés en domaines. La plupart des informations représentées sont le temps passé, et l'extensions s'organise en plusieurs pages permettant de

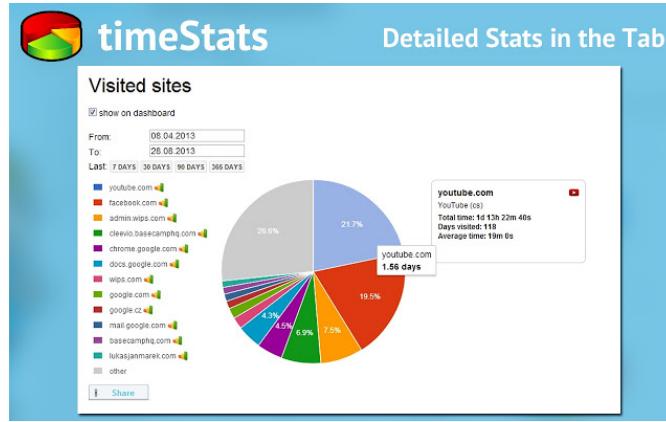


FIGURE 2.7 – Image de présentation de timeStats[12].

voir des visualisations différentes. On remarque la présence de plusieurs types de graphiques (en ligne, en secteurs) adaptés à la mesure affichée. timeStats est disponible pour Google Chrome uniquement.

Ghostery

Ghostery est une extension Google Chrome qui possède également sa propre page web en dehors du catalogue. La figure 2.8 montre la page d'accueil du site “ghostery.com”, qui est le domaine officiel de l'extension listée sur Google Chrome.

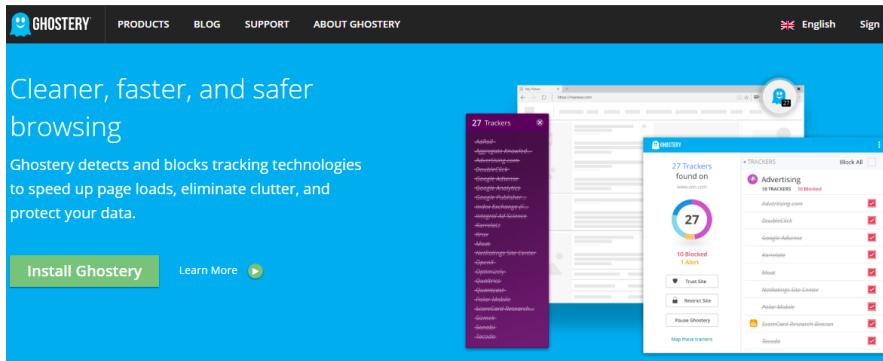


FIGURE 2.8 – Page d'accueil de Ghostery[13].

Ghostery semble donc se concentrer sur la détection et le blocage des informations envoyées aux trackers tiers lors de la navigation. Quelques options de personnalisation y sont présenter, comme la possibilité d'autoriser des trackers particuliers, ou des domaines choisis.

Privacy manager

Privacy manager se montre comme une extension permettant la gestion de mécaniques liées à la préservation de la vie privée. La figure 2.9 montre l'interface principale utilisée par l'extension. Bien que certaines options existent pour la protection de la vie privée, presque la moitié les options activables n'ont pas directement à faire avec la vie privée, et sont plutôt des désactivation ou activations de fonctionnalités de productivité.

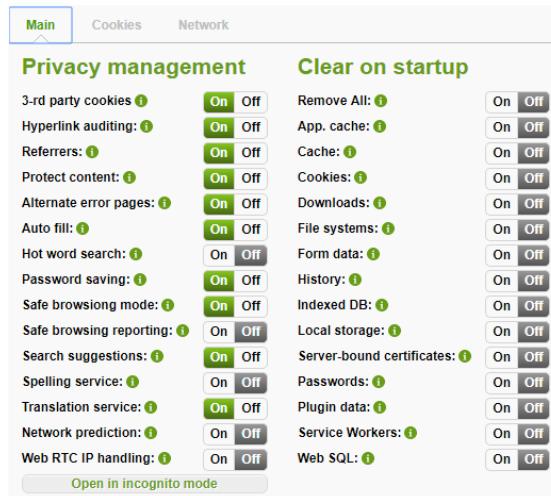


FIGURE 2.9 – Interface de base de Privacy manager[14].

TheGoodData

TheGoodData remplit à priori la même mission que Ghostery, mais propose des outils légèrement différents, et son thème est centré sur l'utilisation de la valeur des données de navigation pour une bonne cause. Un tableau de bord montré à la figure 2.10 permet de se renseigner sur l'état actuel de sa navigation avec des analyses basiques sur les dangers trouvés.

Noiszy

Noiszy cherche quand à lui à brouiller les pistes des trackers existants, sans les bloquer. Son hypothèse de base est qu'il est presque impossible de dissimuler complètement ses "Digital Footprints", et que la meilleure solution est de tenter de les brouiller en les "falsifiant", par exemple en envoyant des données erronées aux trackers, ou en quantité trop élevées. La figure 2.11 montre le premier paragraphe de présentation de Noiszy, présent sur leur site web.

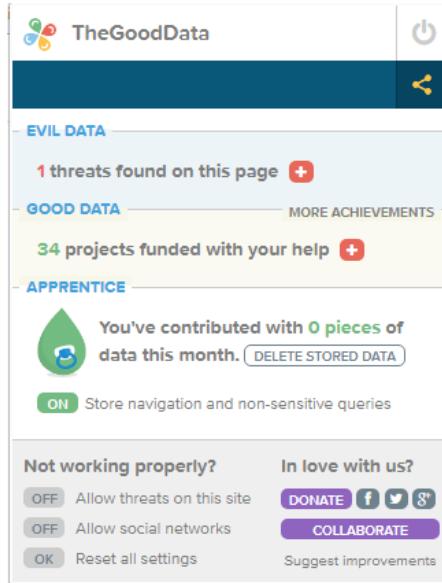


FIGURE 2.10 – Interface de TheGoodData[15].

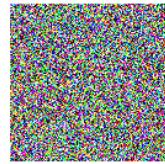
Privacy Badger

Privacy Badger est une extension développée par l'EFF[19]. Disponible à la fois sur Google Chrome et Mozilla Firefox, cette extension a également comme objectif de contrôler l'envoi de données à des trackers. Plutôt que de strictement bloquer toute requête, cette extension laisse à l'utilisateur décider quel niveau de danger représenter chaque tracker, et adapte son comportement entre un blocage total, la retenue de certaines informations ou aucune action entreprise pour chaque tracker détecté. La figure 2.12 montre l'interface de l'application une fois celle-ci installée. On peut y voir les lignes de présentant chacune un tracker, et la possibilité de définir son niveau de danger, et par conséquent l'action appropriée associée.

Kraken.me

Kraken.me est une extension de navigateur, mais également une application pouvant s'installer sur smartphone. Cette application analyse le flux de données de certains services comme Facebook, Twitter, LinkedIn et d'autres. L'objectif est ici de donner à l'utilisateur une vue sur ses propres données, et la manière que celles-ci sont utilisées par les applications. La figure 2.13 montre le modèle présenté par le site web.

Cette application est probablement une des plus semblable à l'objectif général de notre projet, il serait donc intéressant de voir quels ont été les débouchés de cette étude. Notons que la plupart de l'activité de celle-ci ainsi que de l'outil semblent



You are being tracked.

Whatever you do online, you leave digital tracks behind.

These digital footprints are used to market to you - and to influence your thinking and behavior.

On April 3, President Donald Trump signed a repeal of online privacy rules
that would have limited the ability of ISPs to share or sell customers' browsing history for advertising purposes.
Erasing these footprints - or not leaving them in the first place - is becoming more difficult, and less effective.

Hiding from data collection isn't working.

Instead, we can make our collected data less actionable by leaving misleading tracks, camouflaging our true behavior.

We can resist being manipulated by making ourselves harder to analyze - both individually, and collectively.

We can take back the power of our data.

FIGURE 2.11 – Premier paragraphe de la page web de Noiszy[16].

avoir cessé en 2014.

2.4.3 Conclusion

Après avoir dressé une liste des extensions de navigateur les plus populaires et utilisés, nous pouvons prendre position sur les fonctionnalités que notre extension va posséder afin de se démarquer et de répondre à la problématique de l'étude. Nous allons choisir les fonctionnalités que nous estimons avoir un impact pour la sensibilisation du public aux traces que les internautes laissent, et des informations que nous pouvons en retirer. Ainsi, le plug-in se concentrera sur les deux aspects suivants :

- Détection et mise en lumière des différents trackers présents sur les pages visitées par l'utilisateur.
- Tentative de reconstitution du profil de l'utilisateur à partir de la fréquence de la visite des pages web et de leur contenu.

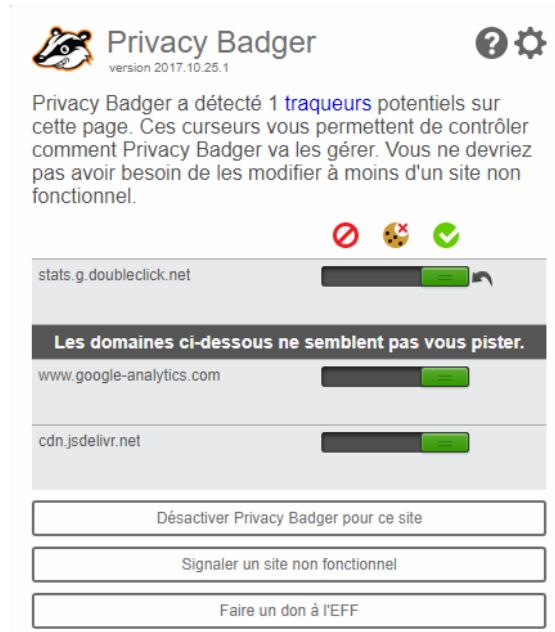


FIGURE 2.12 – Contrôle des actions face aux trackers de Privacy Badger[17].



FIGURE 2.13 – Flux de données de Kraken.me[18].

Chapitre 3

Conception

3.1 Introduction

3.1.1 Idée

Le but recherché de l'outil est de sensibiliser les utilisateurs aux informations que ceux-ci dévoilent potentiellement en naviguant sur le web. Pour ce faire, nous avons besoin d'amasser des données sur leurs habitudes de navigation afin de les analyser.

Ces données seront centralisées sur un serveur afin que nous puissions lancer des traitements sur l'ensemble des données plus tard dans le but de tenter de révéler des tendances, habitudes ou corrélations entre les données.

De plus, nous souhaitons également offrir un service direct à l'utilisateur afin que celui-ci ait un bénéfice à installer l'extension et nous autoriser à accéder à ces données. Nous allons lui montrer via une interface web les données que nous avons pu amasser sur sa navigation depuis l'installation du plug-in, au travers de plusieurs pages et visualisations.

Nous souhaitons également que les données récupérées ne puissent pas être utilisées pour reconnaître une personne particulière. C'est pourquoi le plug-in ne nécessite aucune connexion avec un compte externe, et ne demande pas d'information directement divulguatrice d'une identité.

Nous pouvons ainsi résumer les caractéristiques principales du plug-in en quelques points.

Le plug-in :

- Récupère les informations de navigation de l'utilisateur
- Envoie ces informations de manière anonyme à un serveur centralisé
- Propose une visualisation des données récoltées et calculées sur l'utilisateur

3.1.2 Architecture

Le projet dans son ensemble requiert le développement d'un minimum de deux parties différentes :

- Une extension pour navigateur afin de récupérer et d'envoyer les données
- Un serveur recevant les données des extensions installées

Une troisième partie s'occupant de l'interface utilisateur est également à prévoir, celle-ci pouvant se situer autant dans l'extension que sur le serveur. La décision est finalement prise d'héberger l'interface utilisateur sur un différent serveur, auquel se connecte l'interface lorsque l'utilisateur souhaite accéder à sa page.

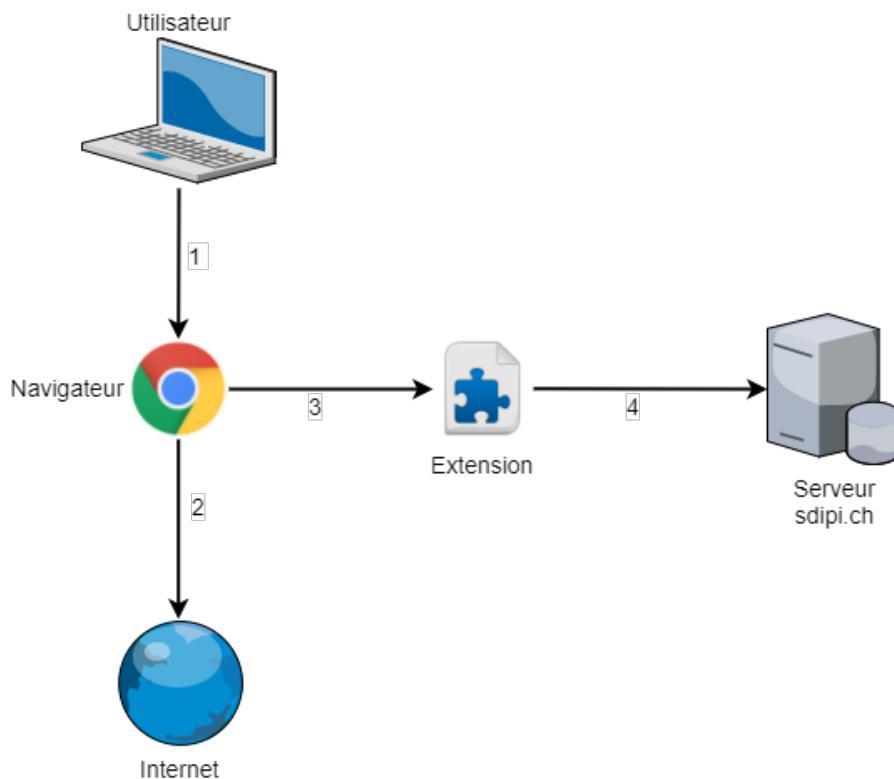


FIGURE 3.1 – Flux de données de l'extension.

La figure 3.1 schématisé la récolte de données effectuée par l'etension.

1. L'utilisateur entre une URL dans son navigateur
2. Le navigateur accède à la ressource concernée
3. Le navigateur transmet au plug-in les informations concernant la navigation
4. Le plug-in contacte le serveur SDIPI pour lui transmettre les informations

3.1.3 Données

Les possibilités de récolte de données depuis une extension de navigateur sont extrêmement nombreuses. Nous allons cependant nous concentrer sur l'amassage de données utiles à l'étude, et qui ne représentent pas une menace à l'intimité de l'utilisateur. Nous devons donc nous limiter à un set de données adéquat.

Voici les différents types d'informations que nous récoltons, et à quelles fins chaque type d'information est utilisé :

Visite d'une URL

Lorsque l'utilisateur accède à une nouvelle URL dans son navigateur, qu'il s'agisse d'un clic sur un lien ou d'une entrée dans la barre d'adresse, l'extension enregistre une partie de l'URL accédée ainsi que la date d'accès. Pour des raisons de protection de la vie privée, seule une partie de l'URL est conservée et envoyée au serveur.

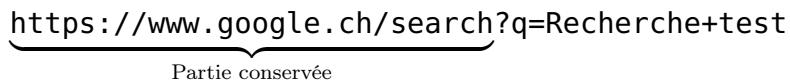

Partie conservée

FIGURE 3.2 – Exemple d'URL et traitement

La figure 3.2 montre que tous les paramètres de la requête ne sont pas conservés. Seuls le protocole (`http` ou `https`), le nom de domaine, l'éventuel numéro de port ainsi que le chemin d'accès à la ressource sont conservés. Nous évitons ainsi la possibilité de stocker des informations sensibles comme le nom d'utilisateur, qui peut parfois se trouver dans cette partie de l'URL de certains sites web.

Activité sur une page

À tout moment, l'utilisateur a probablement plusieurs onglets ou plusieurs fenêtres de navigateur ouvertes. Nous souhaitons nous intéresser à quelle page est actuellement en train d'être parcourue par l'utilisateur. À cette fin, nous détectons les événements sur la page web : Appui sur une touche, ou clic de souris par exemple. Dès lors qu'il se passe plus de 30 secondes sans aucun événement de la part de l'utilisateur, nous estimons qu'il ne regarde plus activement la page. Ce temps passé à s'inséresser à chaque page est également envoyé au serveur central toutes les 30 secondes.

Requêtes du navigateur

Lorsque le navigateur accède à une page web ou à d'autres moments, le navigateur doit charger des ressources qui se trouvent sur un serveur distant. Ce

chargement peut prendre place pour afficher par exemple une image, un morceau de la page web elle-même, ou être demandé par un script chargé.

Pour chaque requête que le navigateur envoie, l'extension mémorise certaines informations :

Origine L'extension mémorise l'URL de la page qui demande la ressource.

Cette information est traitée de la même manière que décrit à la figure 3.2.

Hôte Toujours d'une manière identique à la figure 3.2, l'extension mémorise également le serveur contacté.

Taille L'extension mémorise également la taille de la requête en question, qui correspond à l'addition du contenu envoyé dans le contenu de celle-ci, ainsi que la taille des paramètres (ceux qui ne sont pas retenus par l'extension).

Identificateur

Lors de l'installation de l'extension, un nombre aléatoire est généré pour l'installation. Cet identificateur est envoyé envoyé au serveur central en plus de chaque autre information : Elle nous est utile pour assigner chaque donnée de navigation avec un navigateur particulier.

3.2 Extension

L'extension de navigateur est sans aucun doute la partie la plus simple du projet. Etant donné que nous avons décidé d'héberger l'interface utilisateur sur un serveur différent, l'extension ne va principalement s'occuper que de récupérer les données de l'utilisateur et les transmettre à notre serveur.

Etant donné qu'une extension de navigateur n'est disponible que pour un type de navigateur à la fois, la question du support de plusieurs navigateurs s'est posée. D'après le site populaire w3schools.com[21], Google Chrome représente plus de 75% des visites au moins de décembre 2017. Nous décidons donc de ne pas adapter le code de l'extension pour plusieurs types de navigateurs, car nous estimons que le gain en utilisateurs serait insuffisant pour justifier le développement supplémentaire.

L'extension sera donc développée pour le navigateur Google Chrome, en utilisant l'API JavaScript que celui-ci met à disposition. Les fonctionnalités implémentées sont la récolte et l'envoi des types de données décrites à la section 3.1.3.

L'extension proposera également à l'utilisateur d'accéder à l'interface grâce à un lien, ainsi que la possibilité de se "lier" ce navigateur au profil d'un autre navigateur existant, en entrant son ancien identificateur. Ceci permet à un utilisateur de profiter d'un seul profil au travers de plusieurs machines possédant l'extension, par exemple.

3.3 Serveur

3.3.1 Rôles

La partie du serveur est probablement la plus complexe du projet. Le serveur va devoir assurer le fonctionnement de plusieurs tâches clés :

- Récolte et enregistrement des données de l'extension
- Traitement des données utilisateurs
- API au service de l'interface

3.3.2 Récolte

Avant tout traitement, le serveur doit être capable de recevoir et d'enregistrer les données des clients. Etant donné que l'extension est développée en JavaScript, les données seront transmises par HTTP au format JSON pour des raisons de simplicité.

Installation du plug-in

Au moment où le plug-in est installé, une requête est envoyée au serveur afin de l'avertir qu'un nouvel utilisateur a installé l'extension. Le serveur génère un identifiant, l'envoie à l'extension en réponse et est désormais prêt à recevoir des informations de ce nouvel identifiant.

Récolte continue

Afin que le serveur soit capable de supporter une certaine charge d'utilisateurs, il est nécessaire que celui-ci reçoive un nombre réduit de requêtes de la part des clients. Pour cette raison, l'extension ne contacte le serveur qu'une seule fois toutes les 30 secondes afin de le tenir informé des événements ayant eu lieu.

Le serveur va donc pouvoir exposer une API simple : L'extension contactera toujours le même endpoint, et chaque requête contiendra la liste des informations concernant les événements qui se sont passés chez le client.

Lorsque nous détectons qu'un utilisateur visite une URL pour la première fois, le serveur va télécharger le contenu de cette page. Notre serveur ouvre un navigateur virtuel afin de simuler le chargement complet de la page - y compris l'exécution de scripts - et enregistre le contenu final de la page dans la base de données.

3.3.3 Enregistrement

Le serveur se charge également de la gestion du stockage des données reçues (et calculées). Une base de données MySQL sera continuellement alimentée par les

nouvelles données reçues. La base de données comprendra généralement une table par type de données à enregistrer, ainsi que des tables temporaires dans lesquelles seront placées des informations pré-calculées afin de répondre plus rapidement aux requêtes de l'interface.

3.3.4 Traitement des données

Une fois des données enregistrées, celles-ci sont traitées par différentes méthodes en fonction des besoins de l'interface. Voici le traitement que subit chaque type de données.

Quantité de visites

Deux mesures sont récoltées sur l'intérêt que peut avoir un utilisateur par rapport à une page web : Le nombre de fois que cette URL a été ouverte, et le temps passé à être actif sur la page en question. Chacune de ces informations est également datée.

Contenu des pages

Un des traitements les plus lourds que nous effectuons prend en entrée le contenu des pages visitées.

// TODO HERE -> FLUX ET TRAITEMENT ETC

Requêtes du navigateur

Comme mentionné précédemment, quelques informations de chaque requête du navigateur du client sont enregistrées. Ces données n'ont pas besoin d'un traitement particulier.

Etant donné que nous nous intéressons particulièrement à leur quantité, nous n'allons principalement que les compter. Cependant dû au fait de leur énorme quantité, il nous est nécessaire de pré-calculer certaines sommes avant de les servir à l'interface.

3.3.5 API

Le serveur a également le rôle de répondre aux demandes de l'interface, et de lui fournir les informations nécessaires pour afficher les données du client. Ces communications se font au travers d'une série de requêtes initiées par le client.

Une partie des données servies au client sont pré-calculées, comme la liste des topics tirés de LDA ou le poids TF-IDF des mots, et ne sont donc rafraîchies que périodiquement lorsque demandé.

Le reste des données, comme le nombre d'ouvertures d'une page ou le temps actif passé sur chaque page, est continuellement rafraîchi. Ces données sont donc toujours à jour.

3.4 Interface

L'interface a connu de nombreuses versions au fur et à mesure du projet. Cependant, le thème et le but commun de ces pages n'a pas changé : Montrer à l'utilisateur les informations qu'il révèle, ainsi que des possibles utilisations de celles-ci. Le design initial des visualisation était très visuel et varié et a progressé vers des pages plus utilitaires.

L'interface se divise en deux onglets distincts, chacun tentant de représenter une partie des informations.

3.4.1 Profil

3.4.2 Trackers

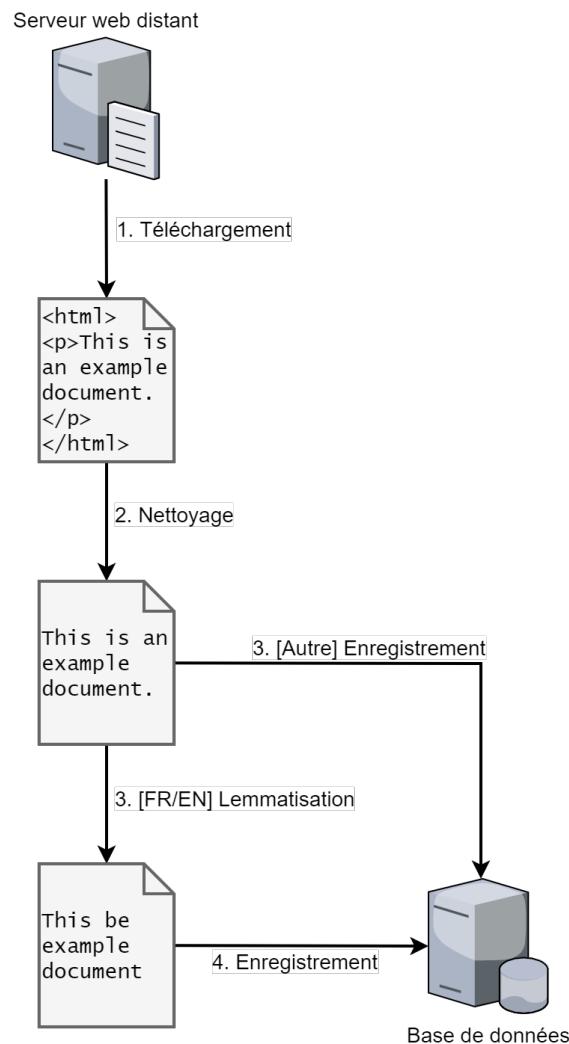


FIGURE 3.3 – Téléchargement et enregistrement du contenu d'une page

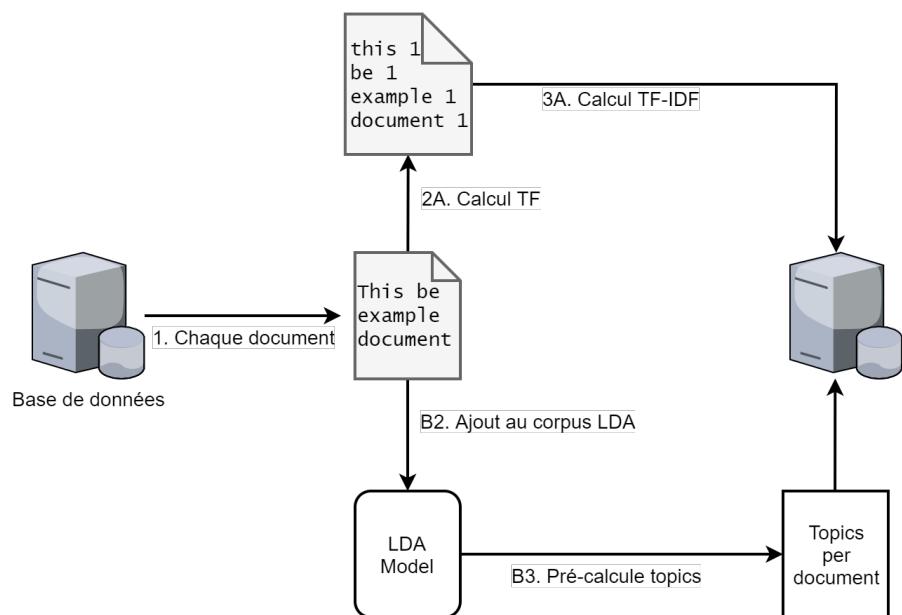


FIGURE 3.4 – Traitements du contenu des pages

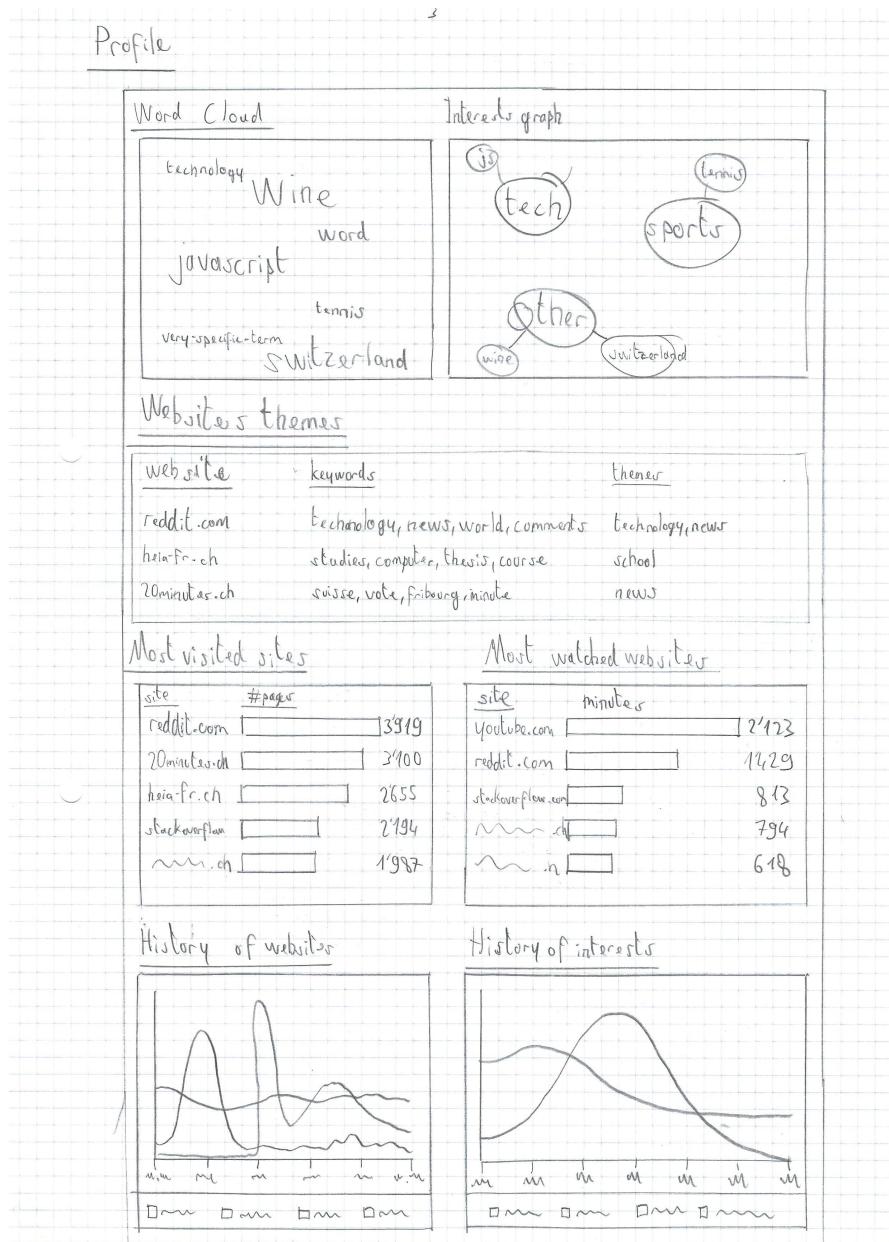


FIGURE 3.5 – Maquette de la page de Profil

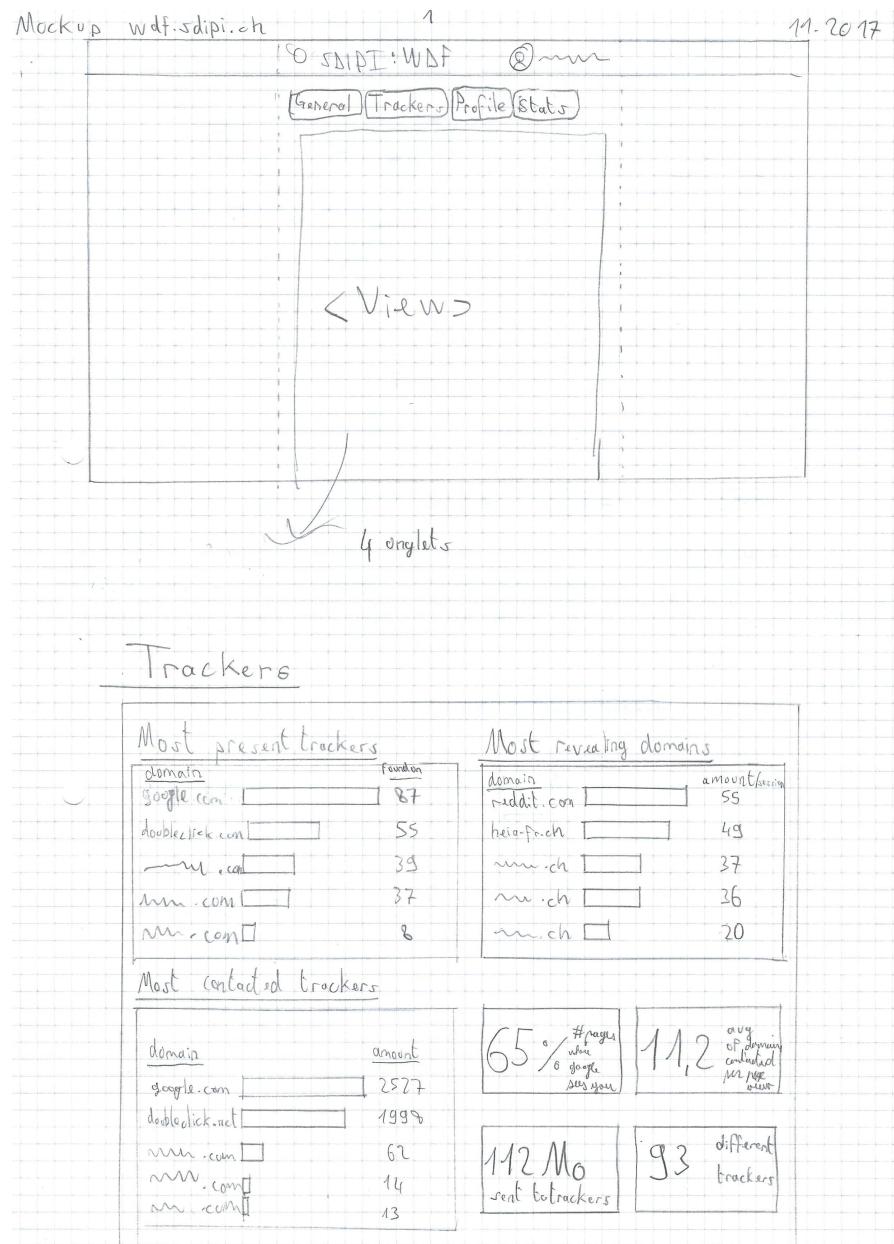


FIGURE 3.6 – Maquette de la page des Trackers

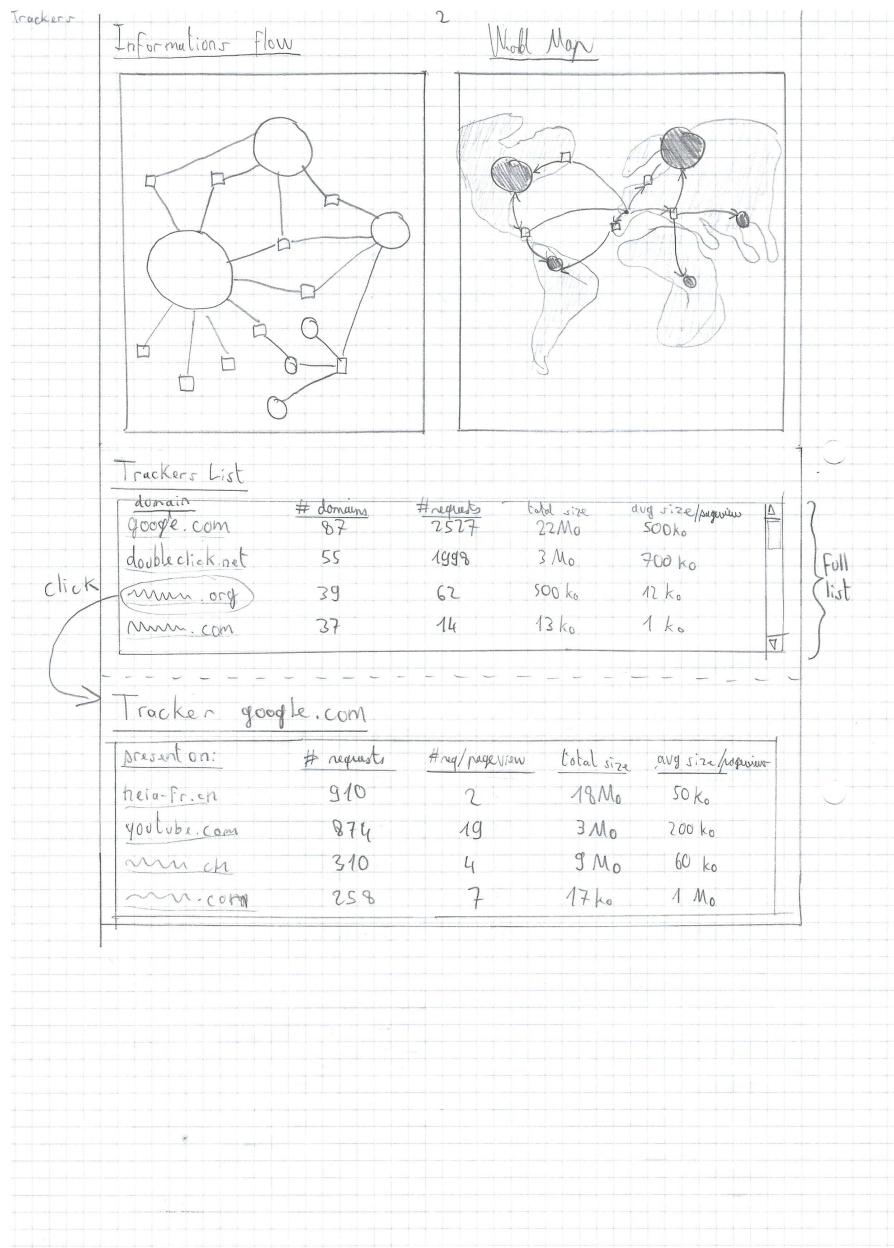


FIGURE 3.7 – Suite de la maquette des Trackers

Chapitre 4

Implémentation

4.1 Extension

4.1.1 Technologies

Bla

4.1.2 Structure

Bla

4.2 Serveur

4.2.1 Technologies

Bla

4.2.2 Structure

Bla

4.3 Interface

4.3.1 Technologies

Bla

4.3.2 Structure

Bla

Chapitre 5

Résultats

5.1 Validation

5.1.1 Tests utilisateurs

Bla

5.2 Résultats de la recherche

5.2.1 Statistiques

Profiling

Trackers

5.2.2 Implications

5.3 Conclusion

Chapitre 6

Conclusion

6.1 Conclusion du projet

6.1.1 Délivrables

Ce projet est passé par plusieurs étapes distinctes qui ont mené à la production de plusieurs délivrables :

- Analyse des besoins
- Analyse des technologies
- Truc 1
- Truc 2

Chacune de ces étapes nous a amené à produire une itération supplémentaire contenant des nouveautés fonctionnelles.

6.1.2 Conclusion générale

Blabla

- Truc 1
- Truc 2

6.1.3 Perspectives

Le projet dans son état final contient plusieurs pages non implémentées :

- Truc 1
- Truc 2

6.2 Conclusion personnelle

Blabla.

Bibliographie

- [1] Michal Kosinski, *Dr Michal Kosinski*, <http://www.michalkosinski.com/>, Consulté en ligne en Septembre 2017, 2017.
- [2] Michal Kosinski, Yilun Wang, Himabindu Lakkaraju and Jure Leskovec, *Mining Big Data to Extract Patterns and Predict Real-Life Outcomes*, <http://psycnet.apa.org/fulltext/2016-57141-003.pdf>, Consulté en ligne en Octobre 2017, 2017.
- [3] Internet Society, *Digital Footprints*, <https://www.internetsociety.org/wp-content/uploads/2017/08/Digital20Footprints20-20An20Internet20Society20Reference20Framework.pdf>, Consulté en ligne en Novembre 2017, Janvier 2014.
- [4] Motherboard, *The Data That Turned the World Upside Down*, https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win, Consulté en ligne en Octobre 2017, Janvier 2017.
- [5] Michal Kosinski, *The End of Privacy, Keynote at CeBIT'17*, <https://www.youtube.com/watch?v=DYhAM34Hhzc>, Consulté en ligne en Octobre 2017, Mars 2017.
- [6] Kaggle, *Young People Survey*, <https://www.kaggle.com/miroslavabo/young-people-survey>, Consulté en ligne en Octobre 2017, 2013.
- [7] Michal Kosinski, *myPersonnality Project*, <http://mypersonality.org>, Consulté en ligne en Octobre 2017, 2013.
- [8] Google, *Google Solutions Analytics*, <https://www.google.com/analytics>, Consulté en ligne en Octobre 2017, 2017.
- [9] BuiltWith, *Analytics Usage in Switzerland*, <https://trends.builtwith.com/analytics/country/Switzerland>, Consulté en ligne en Octobre 2017, 2017.
- [10] Chrome Web Store, *Extensions*, <https://chrome.google.com/webstore>, Consulté en ligne en Novembre 2017, 2017.

- [11] Modules Firefox, *Extensions*, <https://addons.mozilla.org/fr/firefox/extensions/>, Consulté en ligne en Novembre 2017, 2017.
- [12] Chrome Web Store, *timeStats*, <https://chrome.google.com/webstore/detail/timestats/ejifodhjoeenihgfpjjjmpomaphmah>, Consulté en ligne en Novembre 2017, 2017.
- [13] Ghostery, *Ghostery makes the Web Cleaner, Faster and Safer!*, <https://www.ghostery.com/>, Consulté en ligne en Novembre 2017, 2017.
- [14] Chrome Web Store, *Privacy manager*, <https://chrome.google.com/webstore/detail/privacy-manager/giccehglhacakcfemddmfhdkahamfcmd>, Consulté en ligne en Novembre 2017, 2017.
- [15] TheGoodData, *TheGoodData*, <https://thegooddata.org>, Consulté en ligne en Novembre 2017, 2017.
- [16] Noiszy, *Noiszy*, <http://noiszy.com>, Consulté en ligne en Novembre 2017, 2017.
- [17] Modules pour Firefox, *Privacy Badger*, <https://addons.mozilla.org/fr/firefox/addon/privacy-badger17/>, Consulté en ligne en Novembre 2017, 2017.
- [18] Kraken.me, *Home*, <http://www.kraken.me/#/home>, Consulté en ligne en Novembre 2017, 2017.
- [19] Electronic Frontier Foundation, *Defending your rights in the digital world*, <https://www.eff.org>, Consulté en ligne en Novembre 2017, 2017.
- [20] HTTP Archive, *Trends*, <http://httparchive.org/trends.php?s=Top1000&minlabel=Oct+15+2011&maxlabel=Oct+16+2017#numDomains&maxDomainReqs>, Consulté en ligne en Novembre 2017, 2017.
- [21] HTTP Archive, *Browser Statistics*, <https://www.w3schools.com/browsers/default.asp>, Consulté en ligne en Janvier 2018, 2018.
- [22] Google, *Chrome*, <https://www.google.fr/chrome>, Consulté en ligne en Janvier 2018, 2018.
- [23] Google Chrome, *JavaScript APIs*, https://developer.chrome.com/apps/api_index, Consulté en ligne en Janvier 2018, 2018.
- [24] Google Web Store, *Extensions*, <https://chrome.google.com/webstore/category/extensions>, Consulté en ligne en Janvier 2018, 2018.

Glossaire

open-source qualifie un logiciel dont le code initial est mis à disposition du grand public.. 8

Remerciements

Je tiens à remercier ma superviseure Fatemi Nastaran pour m'avoir guidé lors des décisions à prendre, ainsi que Félicien Fleury pour m'avoir soutenu et guidé tout au long de ce projet.

Déclaration d'honneur

Je, soussigné, Kewin Dousse, déclare sur l'honneur que le travail rendu est le fruit d'un travail personnel. Je certifie ne pas avoir eu recours au plagiat ou à toutes autres formes de fraudes. Toutes les sources d'information utilisées et les citations d'auteur ont été clairement mentionnées.

Lieu

Date

Signature

Annexe A

Historique des versions

Voici l'historique des versions de ce document.

- 0.1 : Template du document
- 0.2 : Chapitre "Analyse"
- 0.3 : Correction, complétion du chapitre "Analyse", rédaction d'une partie du chapitre "Conception"

Annexe B

Cahier des charges

B.1 Activités

Le développement du projet peut se découper en plusieurs phases, qui elles-mêmes se divisent en plusieurs activités. Voici la liste de ces activités :

1. Analyse
 - (a) Item 1
 - (b) Item 2
2. Conception
 - (a) Item 1
 - (b) Item 2
3. Implémentation
 - (a) Item 1
 - (b) Item 2
4. Résultats
 - (a) Item 1
 - (b) Item 2

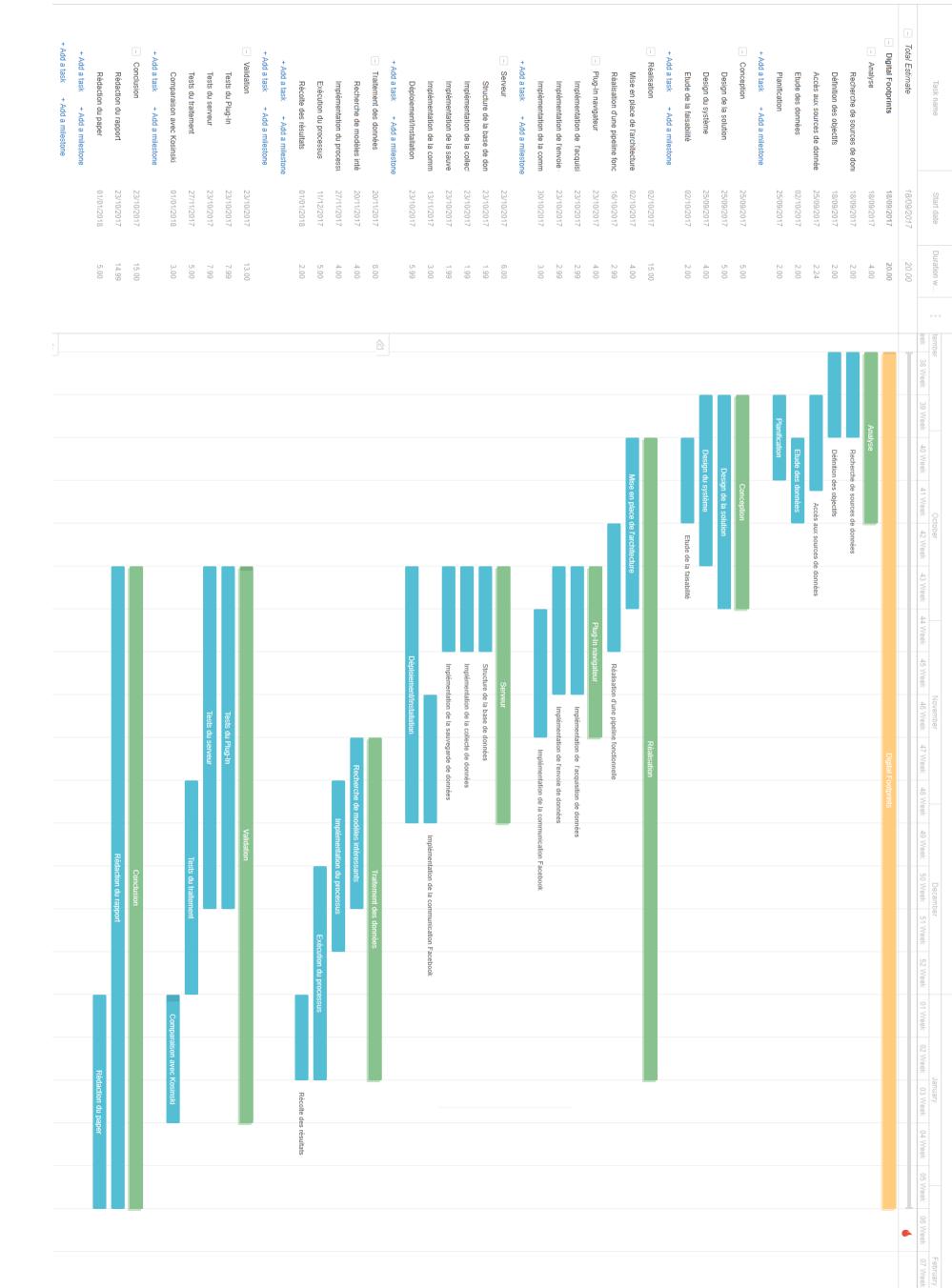
B.2 Planification

Le projet comporte une série de dates-clé qu'il est important de respecter :

Date	Semaine	Tâche
Lundi 18 septembre 2017	Semaine P1	Début du projet
Vendredi 9 février 2018	Semaine P15	Dépôt du rapport
26 février-9 mars 2017	-	Défense orale

Les dates en rouge sont des dates de rendu officielles. Les autres représentent des jalons dans l'avancement du projet.

B.3 Diagramme de Gantt



Annexe C

Documentation

C.1 Localisation

L'ensemble des documents du projet est disponible à l'adresse suivante : [soon](#)

C.2 Contenu

C.2.1 GitLab

Le projet présent sur la forge contient toutes les versions de chacun des documents suivants, sous l'onglet « Documents » :

- Les procès-verbaux réalisés durant le projet.

Annexe D

Procès-verbaux

Voici les documents des procès-verbaux réalisés.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



MASTER OF SCIENCE
IN ENGINEERING

PV de réunion

15 septembre 2017, de 9h05 à 10h55

Présent : Nastaran Fatemi, Félicien Fleury, Kewin Dousse

Rédaction du PV le 20 septembre

Compte-rendu

Points de discussion

- Objectif du projet

Il a été défini tout d'abord que l'objectif du projet était de répondre, dans un sens large, à la question suivante : Est-il possible de faire un profil d'un utilisateur en se basant sur sa navigation web ? Celui-ci aura un but informatif, sensibilisant.

- Sources de données

La question s'est posée sur quelles sont les sources de données à notre disposition pour ce projet. Leur utilisation spécifique n'est pas encore connue, mais nous aurons sans doute besoin de données d'utilisateurs à confronter à notre système. Afin de ne pas être bloqué par l'étape de la récolte de ces données plus tard, il est important d'y réfléchir tôt et d'entreprendre des démarches si nécessaires auprès d'organismes pouvant nous en fournir. Nous prévoyons donc de chercher un corpus de données d'utilisateurs assez tôt.

À court terme, il est donc nécessaire de rechercher quelles les sources de données possibles. Plus précisément, nous savons déjà que des organismes comme l'université de Cambridge peuvent détenir des données intéressantes et allons prendre contact avec eux.

- Plug-In

La question s'est posée : Est-ce que l'outil développé doit être utile après la fin de l'étude, ou est-ce que celui-ci n'est « qu'un » outil pour aider l'étude et atteindre des résultats finaux ? La question reste ouverte. Cette question en soulève également une autre : Quels sont les outils que nous nous autorisons éthiquement à utiliser pour celui-ci ?

- Organisation

Afin de communiquer et nous organiser efficacement, nous allons utiliser plusieurs outils dont Trello pour l'organisation des tâches, Slack pour la communication écrite, et Skype pour des communications audio. Une association sera créée afin de donner de la visibilité et de la légitimité à cette recherche. Son nom et ses statuts seront finalisés bientôt. La recherche visera également une publication, par exemple dans une conférence à définir.

Une réunion hebdomadaire est prévue le jeudi à Yverdon. Nastaran et Kewin y seront présents, et il est prévu que Félicien y participe alternativement sur place, ou par Skype.

Conclusion

La première phase du projet passe par une recherche et une compréhension des différentes sources d'informations disponibles.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

21 septembre 2017, de 13h00 à 10h55

Présent : Nastaran Fatemi, Kewin Dousse, Félicien Fleury (par Skype)

Rédaction du PV le 22 septembre

Compte-rendu

Points de discussion

- Recherches de Kosinski

Nous avons tout d'abord remarqué que les recherches de Michal Kosinski, diplômé de l'Université de Stanford, allaient être de précieuses sources d'informations. Nous allons pouvoir tirer des liens étroits entre les résultats de son étude sur la possibilité de deviner le profil psychologique d'une personne en se basant sur ses 'likes' Facebook.

- Google Analytics

Après avoir suivi le guide Débutant pour YouTube Analytics, Kewin a pu comprendre une partie de l'étendue des possibilités de l'outil. Enormément d'informations sont disponibles, et celles-ci peuvent être cachées/filtrées/triées etc. Cependant il serait intéressant de découvrir les possibilités avancées de l'outil pour se rendre compte jusqu'à quel point celui-ci peut tracker l'activité d'un utilisateur précisément.

- Alternatives à Google Analytics

Quelques alternatives à Google Analytics ont été découvertes, mais celles-ci ne présentent pas vraiment de concept intéressant à l'étude autre que le fait que certaines d'entre-elles sont open-source. Il semble que Google Analytics soit l'outil public le plus grand et le plus utilisé dans sa catégorie.

- Direction du projet

Après une discussion sur les différentes voies futures du projet, l'idée a été sur la proposition suivante : Il s'agira d'implémenter un plug-in pour navigateur qui va récupérer les informations de navigation de son utilisateur de manière automatique et transparente. L'utilisateur va devoir utiliser son compte Facebook afin de se connecter, pour que nous puissions lier les données de navigation avec les données présentes sur un profil Facebook. Toutes les données que nous récupérerons (à la fois par le plugin et par Facebook) seront anonymisées. L'utilisateur sera mis au courant de ce processus avant le début de l'utilisation du plug-in. Il y aura la possibilité d'activer le tracking par période de temps, par exemple à certaines heures de la journée. Nous centraliserons la récupération de ces données et appliquerons des algorithmes afin de déterminer si nous pouvons conclure des informations en se basant sur les données que nous avons récoltées nous-mêmes, et en les vérifiant avec les données que le profil Facebook nous donne en lui « appliquant » la méthode de M. Kosinski. Pour sa volonté de participer à cette enquête, nous allons mettre à disposition de l'utilisateur diverses métriques que nous calculerons en temps réel.

Conclusion

Nous avons désormais une idée bien plus précise du projet à réaliser, et les recherches pour le projet peuvent commencer en visant un but.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

28 septembre 2017, de 13h15 à 14h15

Présent : Nastaran Fatemi, Kewin Dousse, Félicien Fleury (par Skype)

Rédaction du PV le 29 septembre

Compte-rendu

Points de discussion

- **Association SDIPI**

La première discussion a été sur la création de l'association nommée « Swiss Digital Identity and Privacy Institute ». Cette association servira à encadrer le projet et lui donner de la légitimité/visibilité tout en montrant que le but n'est pas économique. Les statuts de l'association seront validés en principe la semaine prochaine.

- **E-mail à myPersonnality**

La source de données la plus importante pour le projet est <http://mypersonality.org>, un site web regroupant les données amassées par les études de Kosinski. Ces données ne sont pas accessibles publiquement, mais il est possible d'en demander un accès en envoyant un mail expliquant le but de notre recherche. Un mail sera écrit la semaine prochaine, une fois que l'association aura une certaine visibilité en ligne, afin de demander l'accès à ces données.

- **Site web SDIPI**

Il est nécessaire que l'Association ait une certaine présence et visibilité en ligne afin de montrer son but au public et de faire des demandes. La mise en place du site web discutée après la réunion de jeudi prochain.

- **Planning**

Une proposition de planning a été faite. Après quelques modifications, celui-ci semble être raisonnable pour le projet.

- **Pages web démonstratives pour GA**

Afin de bien se rendre compte des possibilités données par Google Analytics et également le montrer aux utilisateurs, il est décidé d'implémenter le plus de features possibles de Google Analytics sur un site web d'exemple.

Conclusion

Il est désormais primordial que l'association ait une visibilité en ligne et une certaine visibilité afin de pouvoir demander l'accès à la base de données de Kasinski, qui sera probablement la principale source de données utile au projet.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

5 octobre 2017, de 11h10 à 12h40

Présent : Nastaran Fatemi, Kewin Dousse, Félicien Fleury

Rédaction du PV le 6 octobre

Compte-rendu

Points de discussion

- Google Analytics et site d'exemple

Un site web présentant plusieurs pages de contenu factice imitant un shop en ligne a été implémenté ainsi qu'une intégration avec Google Analytics et certains des concepts de tracking avancés, comme les évènements. Il s'est avéré qu'aller plus loin dans l'implémentation de certaines mesures n'était pas une priorité car la seule limite aux données qu'il est possible de récupérer est en réalité une limitation technique : Il s'agit des informations que les navigateurs peuvent potentiellement révéler à un script, ou à un serveur distant.

- Informations des utilisateurs

Il sera donc intéressant de se poser la question « Quelles sont les informations qu'une page peut potentiellement envoyer à un serveur distant ? ». Ces informations doivent passer par le net pour Google Analytics, et donc il faudra se renseigner non seulement sur les moyens possibles qu'un client a de contacter un serveur (par exemple avec une requête AJAX, ou même avec une tentative d'accès à un fichier comme une image sur le serveur), ainsi qu'aux types d'informations qu'a accès un navigateur web classique.

- Création de l'association SDIPI

La fin de la réunion formelle a porté sur le review des statuts de l'association prochainement créée : « Swiss Digital Identity & Privacy Institute ». Quelques changements ont été faits ; Les status seront donc définitivement validés plus tard.

- Objectifs du projets

Une discussion sur les objectifs du projets a également eu lieu. Nous avons décidé que le projet allait viser à chercher une correspondance entre les URL visitées par une personne et son profil psychologique. Ce lien se fera à l'aide des données de Kosinski, qui nous aidera à lier les likes Facebook d'une personne et son profil psychologique. Le remplissage du questionnaire psychologique par les volontaires de notre projet sera facultatif.

Conclusion

L'association va terminer de se créer afin d'envoyer une lettre de demande à Kosinski, et pendant ce temps les recherches sur les possibilités techniques de divulgation des informations va continuer.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

10 octobre 2017, de 9h30 à 10h05

Présent : Nastaran Fatemi (Skype), Kewin Dousse (Skype), Félicien Fleury (Skype)

Rédaction du PV le 11 octobre

Compte-rendu

Points de discussion

- Statuts de l'association

Les statuts de l'association doivent subir quelques changements mineurs avant d'être définitifs. Félicien va effectuer les modifications nécessaires, puis les statuts seront lus, imprimés et signés par tous les membres de l'association. Un PV de l'assemblée constitutive déroulée sera également rédigé.

- Site web

Un début de site web a été présenté. La structure générale et le thème seront conservés. Celui-ci ne contient que peu de contenu, il sera étoffé pour jeudi dans le but d'être présentable et mis en ligne.

- Lettre à myPersonnality

Le template de l'e-mail à envoyer à myPersonnality reste à compléter par quelques détails : Un enregistrement du projet sur le site <https://osf.io> est nécessaire. Les détails du projet et des membres seront complétés pour jeudi également. Le but est d'avoir en main tous les éléments nécessaires pour écrire le mail définitif jeudi et l'envoyer.

Conclusion

La complétion des informations et contenus pour envoyer l'e-mail de demande d'accès aux données à Kosinski est actuellement la priorité, et cette tâche devrait arriver à son terme jeudi.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

12 octobre 2017, de 13h00 à 13h40

Présent : Félicien Fleury (Skype), Kewin Dousse (Skype)

Rédaction du PV le 12 octobre

Compte-rendu

Points de discussion

- Site web**

La première discussion a porté sur les détails du site web. La plupart du contenu a été ajouté, quelques corrections ont été effectuées, et la mise en ligne officielle du site s'est terminée quelques heures après la fin de la réunion.

- Comité d'éthique**

Après la complétion d'informations à la fois sur le site web officiel de l'association et sur la page OSF requise du projet, il a été remarqué que dans le template d'e-mail pour Kosinski se trouve une ligne faisant référence à l'IRB (Institutional Review Board). Ceci n'avait pas été mis en avant jusqu'ici, et signifie probablement qu'une approbation d'un comité d'éthique est nécessaire pour continuer le projet, car Kosinski s'attend à le recevoir par e-mail. Cette étape sera discutée avec Nastaran car l'école d'ingénieurs est probablement compétente pour ce problème.

Conclusion

La question du comité d'éthique est à traiter au plus vite car il s'agit d'une étape non anticipée qui pourrait considérablement ralentir l'obtention des données

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

16 octobre 2017, de 15h00 à 15h20

Présent : Félicien Fleury (Skype), Kewin Dousse (Skype)

Rédaction du PV le 17 octobre

Compte-rendu

Points de discussion

- Comité d'éthique

Bien que la question de l'acceptation du projet par un comité d'éthique soit en suspens, nous allons pour l'instant avancer tout de même dans la partie technique du projet

- Architecture

Il y eut ensuite une discussion sur l'architecture de l'application de base à réaliser pour la récupération des données des utilisateurs. L'idée initiale d'extension de navigateur est bonne, mais demande de développer une extension par navigateur différent. Bien que la plupart du code soit le même, le développement partira sur un « userscript » dans un premier temps : Il s'agit d'une extension avec des fonctionnalités réduites, n'utilisant que du JavaScript pur (sans utiliser d'API navigateur) et ayant l'avantage d'être compatible sur plusieurs navigateurs. De même, le développement de la partie serveur va également commencer, suite à la mise en fonction d'une machine virtuelle pour accueillir le software serveur.

Conclusion

L'acquisition des données se trouve retardée, mais le projet avance tout de même du point de vue développement pendant ce temps.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

26 octobre 2017, de 13h00 à 13h45

Présent : Félicien Fleury (Skype), Kewin Dousse, Nastaran Fatemi

Rédaction du PV le 28 octobre

Compte-rendu

Points de discussion

- **Comité d'éthique**

Etant donné les délais attendus par non seulement la réponse espérée de Kosinski, mais surtout par celui du comité d'éthique, la décision a été prise de passer cette idée au second plan et de chercher un autre axe de développement pour le projet.

- **Idées**

Le but de la discussion suivante a été de chercher de nouveaux axes de développement pour le projet, en partant de l'idée que nous n'aurons pas accès aux données de la base de données de Kosinski.

Plusieurs idées ont vu le jour ici, dont celle de développer un produit en partenariat avec une entreprise externe. Mais l'idée qui a été retenue au final est différente, mais reste en cohésion avec le développement technique effectué jusqu'ici : Le but sera de développer dans un premier temps une extension de navigateur pour :

- Récolter des données utilisateurs concernant leur fréquentation des sites web
- Renseigner les utilisateurs sur leur utilisation du web, et les informer en leur montrant la manière dont ils apparaissent au web, par exemple en générant un avatar leur ressemblant, ou en leur montrant des statistiques sur leur navigation et les dangers potentiels

Cette récolte d'information donnera lieu dans un deuxième temps à un jeu de données sur la navigation des utilisateurs qui sera mis en relation avec leur profil Facebook. Les données seront ensuite analysées afin d'y trouver par exemple des corrélations intéressantes.

Conclusion

La direction du projet change, mais la partie technique qui a été faite jusqu'ici n'est pas perdue : Nous changeons de vision et d'objectifs à moyen terme, mais le développement continue dans le même sens.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

1 novembre 2017, de 14h05 à 14h55

Présent : Félichen Fleury (Skype), Kewin Dousse (Skype), Nastaran Fatemi (Skype)

Rédaction du PV le 3 novembre

Compte-rendu

Points de discussion

- Planning

Le planning du projet a été réadapté en fonction des modifications dans les objectifs à moyen terme. Nous n'allons donc pas nous baser sur les données de l'étude de Kosinski, et par conséquent n'allons pas attendre sa réponse pour continuer le projet.

- Plug-In Chrome

Nous allons changer les objectifs du projet ainsi : Le but ne sera pas de trouver des corrélations entre les URLs visitées par un visiteur et son profil psychologique (déduit par son profil Facebook + données de Kosinski). Nous allons à la place :

- Donner à l'utilisateur une interface montrant des statistiques sur ses habitudes de navigation du web sous plusieurs formes. Images, graphiques, et nombres.
- Récolter des données sur la navigation des utilisateurs afin d'en trouver des statistiques intéressantes.

- Stratégie

Il est nécessaire de savoir comment positionner le plug-in et l'étude par rapport aux concurrents. Des plug-ins avec existent déjà proposant des fonctionnalités similaires, et un état de l'art est nécessaire afin de savoir dans quelle direction va continuer le développement.

Conclusion

Nous devons savoir comment se positionner par rapport aux plug-ins similaires afin de pouvoir développer des fonctionnalités attrayantes pour les nouveaux utilisateurs.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

9 novembre 2017, de 14h05 à 14h55

Présent : Félicien Fleury (Skype), Kewin Dousse, Nastaran Fatemi

Rédaction du PV le 9 novembre

Compte-rendu

Points de discussion

- Rapport**

La partie d'analyse du rapport a été rédigée en grande partie. Les quelques parties manquantes seront complétées par la suite. La comparaison des extensions existantes sera étoffée afin d'en tirer une conclusion pouvant nous renseigner sur la place que prendra notre extension par rapport à celles existantes, et en quoi les fonctionnalités seront novatrices.

- Fonctionnalités**

La discussion centrale a été les fonctionnalités que le plug-in allait proposer, ainsi que l'intérêt pour les statistiques que nous allions tirer à la fin de l'étude. Nous allons devoir nous baser non seulement sur les données Facebook des utilisateurs, mais nous allons également analyser le contenu des pages que celui-ci visite, et pas seulement leur URL. La discussion a porté sur les méthodes d'analyse de contenu de pages web ; Lesquelles utiliser, que stocker comme données et comment les utiliser au mieux. Nous allons procéder par étapes ; la première d'entre elles sera d'enregistrer le contenu des pages web dans la bases de données.

- Techniques envisagées**

Nous avons réfléchi à des algorithmes à appliquer lors de la récolte de données dans le but d'obtenir des statistiques plus intéressantes sur la navigation des utilisateurs. Le principal intérêt que nous voyons dans l'analyse de contenu des pages est d'effectuer de la reconnaissance de topics sur les pages. Ainsi, nous pourrons – par exemple - tirer des parallèles entre les sujets visités par un utilisateur et ses informations démographiques, ou ses « likes ».

Conclusion

Les fonctionnalités principales du plug-in se définissent, et le développement de la récolte de données progresse en parallèle. Restera à discuter de la stratégie de « publicité » pour le plug-in.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

13 novembre 2017, de 16h00 à 16h30

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype)

Rédaction du PV le 15 novembre

Compte-rendu

Points de discussion

• Méthodes de User Profile Tracking, et de Topic Recognition

Une série de liens et papers ont été synthétisés dans le but d'en apprendre plus sur les techniques actuelles de deux objectifs différents : Premièrement, reconnaître les traces d'un utilisateur et reconstituer son profil en utilisant plusieurs sources de données. Deuxièmement, être capable de définir un ou plusieurs mot-clés représentant le sujet discuté sur une page web/un document.

La conclusion de ces études est la suivante : Les techniques pour tracker un utilisateur sur le web sont déjà connues, et le rapport de James Nolan est toujours intéressant quant à certaines techniques à utiliser. Une nouvelle information est cependant la performance des algorithmes permettant d'extraire le sujet d'une page web : Il semblerait d'après plusieurs sources indépendantes que la méthode de TF-IDF, en conjonction avec certaines autres techniques, donne les résultats les plus probants pour notre cas. Nous allons donc probablement l'implémenter.

Conclusion

Nous sommes à présent au clair sur les techniques à utiliser pour la suite d'outils, particulièrement au niveau de la reconnaissance des sujets d'une page. Ceci pourra désormais être implémenté.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

16 novembre 2017, de 14h15 à 15h00

Présent : Kewin Dousse, Nastaran Fatemi, Félicien Fleury (Skype)

Rédaction du PV le 19 novembre

Compte-rendu

Points de discussion

- Fonctionnalités de l'Extension**

Après avoir passé en revue une liste des plug-ins existants, nous allons pouvoir nous concentrer sur l'implémentation du nôtre au travers de deux axes principaux. La liste actuelle est lacunaire et sera complétée par la suite par d'avantage d'explications sur certaines extensions.

Nous allons nous focaliser sur montrer des informations à l'utilisateur concernant : 1) Les trackers sur la page et vers qui les informations sont envoyées, et 2) Comment le profil reconstitué de l'utilisateur apparaît vu par le web.

- Implémentation**

La méthode de TF-IDF sera initialement utilisée pour reconnaître les topics d'une page web. Il s'agit de la fonctionnalité qui sera implémentée au plus vite.

Conclusion

L'implémentation des fonctionnalités continue, avec une vision plus claire sur les techniques à utiliser ainsi que

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



MASTER OF SCIENCE
IN ENGINEERING

PV de réunion

23 novembre 2017, de 13h35 à 14h40

Présent : Kewin Dousse, Nastaran Fatemi, Félicien Fleury (Skype)

Rédaction du PV le 24 novembre

Compte-rendu

Points de discussion

- Maquettes de l'interface

Des maquettes papier de l'interface du plug-in ont été discutées. Deux pages principales seront présentées : La page « Trackers » montrant des informations et visualisations sur les différents trackers rencontrés sur les pages, et la page « Profile » montrant des informations sur le profil reconstitué de l'utilisateur. En plus de ces deux pages, se trouveront une page « Général » montrant un résumé de l'état du plug-in et de la connexion de l'utilisateur, et une page « Stats » montrant des informations générales sur l'utilisation du projet, tous utilisateurs confondus.

- Implémentation

Le TF-IDF fonctionne. Pour lundi sera implémenté un début de l'interface de la page « Profile ».

Conclusion

Bien que non définitive, la liste des fonctionnalités de l'interface client est assez bien définie pour prodécer à un début d'implémentation et de tests.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

30 novembre 2017, de 13h30 à 14h15

Présent : Kewin Dousse, Nastaran Fatemi

Rédaction du PV le 1 décembre

Compte-rendu

Points de discussion

• Interface

L'avancement de l'interface de la page 'Profile' a été discuté. Celle-ci contient les visualisations avec les graphiques en barres comme les sites/domaines les plus vus/regardés ainsi que la liste des pages et les mots-clés associés, mais il manque encore le wordcloud, le graphe des intérêts, et les graphiques des mots-clés sur la durée ainsi que la sélection de l'intervalle de dates. Plusieurs modifications seront à effectuer pour la qualité des informations affichées sur l'interface, comme la détection de la langue lors du retrait des stopwords des pages, ainsi qu'une meilleure détection du temps passé sur les pages par un utilisateur en comptant tout type d'interaction avec celle-ci. D'autres changements purement sur l'affichage de l'interface seront aussi effectués, comme la combinaison de plusieurs tableaux en une seule visualisation.

• Dates

Quelques dates clés ont été définies pour la suite à court terme :

- Mardi 5 déc. : Fin de la page Profile
- Vendredi 15 déc. : Fin de la page Trackers
- 15 – 22 déc. : Tests de l'interface en interne + retrait du login Facebook pour un login personnalisé

• Données utilisateur

La question de l'intérêt de la récolte des données utilisateurs s'est également posée. Les quelques idées proposées vont dans le sens d'une publication scientifique, et visent à articuler le contenu principalement autour de deux axes : La présentation de statistiques concernant les données récoltées, et la validation que les profils détectés par le plug-in correspondent à la réalité vue par les utilisateurs. On pourra par exemple émettre un questionnaire à ceux-ci afin de chercher une corrélation entre les informations recueillies, et les informations que ceux-ci délivrent volontairement.

Conclusion

Avec les réponses à quelques questions touchant sur le but final du projet, nous sommes au cœur de la phase d'implémentation de l'interface et des fonctionnalités qui lui sont relatives.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

5 décembre 2017, de 8h30 à 9h05

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 6 décembre

Compte-rendu

Points de discussion

- **Interface**

Un point a été fait sur la page « Profile ». Celle-ci comprend les sections supplémentaires « Wordcloud » montrant un nuage des mots les plus vus par l'utilisateur, et « History » affichant un graphique des sites les plus visités sur un intervalle de temps. De plus, les sections « Most visited » et « Most watched » ont été remaniées : Le tableau des keywords par page a été intégré à chacun des autres tableaux montrant les sites et les domaines de la section. Bien que l'interface soit fonctionnelle, plusieurs facteurs rendent les résultats affichés peu fiables (pas de JavaScript exécuté sur les pages, améliorations possibles dans la phase de cleaning des données). Ceci sera remédié.

- **Objectifs**

Les prochaines tâches à effectuer ont été définies : Jusqu'à la fin de la semaine, l'accent sera mis sur la page « Profile » afin de la terminer et de rendre plus fiables les résultats montrés, notamment les keywords et les intérêts de l'utilisateur. La semaine suivante, la page « Trackers » sera implémentée.

Conclusion

Avec les réponses à quelques questions touchant sur le but final du projet, nous sommes au cœur de la phase d'implémentation de l'interface et des fonctionnalités qui lui sont relatées.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

7 décembre 2017, de 15h05 à 15h40

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 8 décembre

Compte-rendu

Points de discussion

- **Interface**

La page « Profile » a été discutée. Tous les onglets sont implémentés, mais certains méritaient encore une discussion. Ainsi, l'objectif de l'onglet « Interests Graph » a été plus précisément décidé et celui-ci subira quelques modifications, ainsi que l'onglet « History » qui servira à montrer des tendances de keywords, plutôt que de sites web. Le choix d'un intervalle de dates reste à implémenter. Ces changements sont prévus pour mardi matin.

Conclusion

La page « Profile » arrive à la fin de son implémentation, et le focus devrait être sur la page « Trackers » dès mardi prochain.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



MASTER OF SCIENCE
IN ENGINEERING

PV de réunion

12 décembre 2017, de 8h35 à 9h05

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 12 décembre

Compte-rendu

Points de discussion

- **Filtrage par date**

Le filtre par date est ajouté sur la page : Il est possible de choisir une date de début et une date de fin pour l'affichage de toutes les données. Des changements conséquents ont été faits sur la manière de calculer les données afin que l'interface soit réactive à ces changements : La plupart des données sont pré-calculées sur le serveur.

- **Graphique « History »**

La deuxième version du graphique « History » a été mis en place, mais ne semble pas assez concluant pour être définitif. Les résultats visuels obtenus ne sont pas toujours représentatifs et visuellement intéressants des données que nous souhaitons afficher, et nous rediscuterons de cette partie jeudi prochain.

- **Graph « Interests »**

La page du graphe des intérêts a suscité des questions sur son fonctionnement. Après discussion, il sera plus intéressant de lier les topics et les mots-clé trouvés, aux intérêts de l'utilisateur que lui-même aura défini lors de l'inscription. Il sera donc nécessaire de lui demander ses intérêts parmi une hiérarchie de centres d'intérêts lors de l'inscription, et cette page permettra de faire un lien entre les intérêts décrits par l'utilisateur, et les intérêts que nous trouverons nous-même.

Conclusion

Des discussions sont encore en cours sur des aspects de la page « Profile », mais de plus en plus d'entre-eux approchent une version finale.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

21 décembre 2017, de 8h35 à 9h05

Présent : Kewin Dousse (Skype), Félicien Fleury (Skype)

Rédaction du PV le 21 décembre

Compte-rendu

Points de discussion

- Points restants avant le lancement de l'extension**

Parmi les quatre points restants à résoudre énoncés la dernière fois, le refactoring de la partie communication serveur et la logique d'envoi de l'extension est terminée. L'extension stocke les messages et ne les envoie qu'une fois toutes les 30 sec.

L'authentification Facebook est enlevée mais un nouveau système à mettre en place a été discuté : Lorsque l'utilisateur installe l'extension, un identifiant lui sera associé et communiqué. Il pourra ensuite le réutiliser sur d'autres machines si il le souhaite. Ceci évite à l'utilisateur une phase d'inscription.

La vue des Trackers et la fin de l'implémentation des intérêts reste à terminer. Comme le temps restant est probablement insuffisant jusqu'aux vacances de Noël, un ou deux jours seront pris entre le 26 et le 28 décembre pour terminer complètement l'extension afin de la proposer à une dizaine d'utilisateurs.

Conclusion

La phase d'implémentation arrive à son terme, et l'extension sera bientôt prête pour une utilisation réelle.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

9 janvier 2017, de 9h05 à 9h30

Présent : Kewin Dousse (Skype), Nastaran Fatemi (Skype), Félicien Fleury (Skype)

Rédaction du PV le 9 janvier

Compte-rendu

Points de discussion

- **Résultats récoltés**

Les modifications requises de l'extension pour une utilisation par plusieurs utilisateurs ont été effectués durant la première partie des deux semaines de pause : Remplacement du login Facebook par un login instantané à l'installation, et finition du système de centres d'intérêts.

L'extension a été utilisée par 7 utilisateurs différents pendant une période d'environ une semaine. Les résultats récoltés ont commencé à être traités, mais la nouvelle taille de ceux-ci pose des problèmes techniques au serveur qui était jusqu'ici suffisant. La résolution de ces problèmes est en cours.

Conclusion

Les prochaines tâches sont la résolution des problèmes techniques dûs à la quantité de données, et le début de l'analyse des résultats obtenus en plus de l'ajout de la page Trackers dans l'interface.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

18 janvier 2017, de 9h05 à 9h30

Présent : Kewin Dousse, Nastaran Fatemi, Félicien Fleury

Rédaction du PV le 23 janvier

Compte-rendu

Points de discussion

- **Interface**

L'implémentation de la partie Trackers de l'interface touche à son terme. Il est désormais possible de lister les domaines envoyant et recevant le plus grand nombre de domaines, ainsi que de cliquer sur l'un deux pour avoir les détails de quels domaines ont communiqué avec celui cliqué. Quelques améliorations sont discutées, comme la possibilité d'afficher le nombre de domaines contactés directement sur les premières pages sans avoir à cliquer sur un domaine particulier.

Pour la partie Profile, l'utilité du « Topics Graph » a été rediscutée : Nous nous en servons principalement pour demander des informations de l'utilisateur sur sa reconnaissance des centres d'intérêts dans les topics proposés. Un graphe n'est donc plus nécessaire : La vue sera désormais une liste, où l'utilisateur peut entrer un centre d'intérêt par ligne (topic). Une révision de la structure du backend est nécessaire afin que ces opérations puissent être faites en cohésion avec un changement de modèle LDA.

Conclusion

L'implémentation de l'interface se termine cette semaine. Une fois les dernières modifications effectuées, le focus sera mis sur le rapport écrit.