

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	8
1.1	Contexte	8
1.2	Objectifs	8
1.3	Méthodologie	9
2	Analyse	10
2.1	Données de l'étude de Kosinski	10
2.1.1	Introduction	10
2.1.2	Résumé	10
2.1.3	Données	12
2.1.4	Acquisition	13
2.1.5	Conclusion	13
2.2	SDIPI	14
2.3	Trackers et Google Analytics	15
2.3.1	trackerackers	15
2.3.2	Etat de l'art	15
2.3.3	Google Analytics	17
2.4	Extensions de navigateur	17
2.4.1	Introduction	17
2.4.2	Etat de l'art	17
2.4.3	Conclusion	21
3	Design du système	23
3.1	Introduction	23
3.1.1	Idée	23
3.1.2	Architecture	24
3.1.3	Données	25
3.2	Architecture	26
3.2.1	Stack technologique	26
3.2.2	Extension	26
3.2.3	Serveur	28

3.2.4	Base de données	32
3.2.5	Interface	36
4	Développement des vues	41
4.1	Wordcloud	41
4.1.1	Concept	41
4.1.2	Données	42
4.1.3	Implémentation	43
4.2	Topics List	45
4.2.1	Concept	45
4.2.2	Données	45
4.2.3	Implémentation	47
4.3	Most Watched	49
4.3.1	Concept	49
4.3.2	Données	50
4.3.3	Implémentation	51
4.4	History	53
4.4.1	Concept	53
4.4.2	Données	53
4.4.3	Implémentation	55
4.5	Trackers	57
4.5.1	Concept	57
4.5.2	Données	59
4.5.3	Implémentation	59
4.6	Stats	61
4.6.1	Concept	61
4.6.2	Données	61
4.6.3	Implémentation	61
5	Résultats	63
5.1	Test utilisateurs	63
5.1.1	Inputs	63
5.2	Résultats de la recherche	65
5.2.1	Modèles	65
5.2.2	Vues	69
5.2.3	Statistiques	69
5.2.4	Implications	69
5.3	Conclusion	69

6 Conclusion	70
6.1 Conclusion du projet	70
6.1.1 Délivrables	70
6.1.2 Conclusion générale	70
6.1.3 Perspectives	70
6.2 Conclusion personnelle	71
A Historique des versions	77
B Cahier des charges	78
B.1 Activités	78
B.2 Planification	78
B.3 Diagramme de Gantt	79
C Documentation	80
C.1 Localisation	80
C.2 Contenu	80
C.2.1 GitLab	80
D Procès-verbaux	81

Table des figures

2.1	Précision moyenne du modèle prédisant la personnalité d'un utilisateur en fonction du nombre de likes analysés[5].	11
2.2	Déviation de la personnalité moyenne estimée d'un visiteur régulier du site ““deviantart.com”” selon les cinq axes psychologiques employés[5].	12
2.3	Page d'accueil du site web https://sdipi.ch	14
2.4	Nombre moyen de domaines contactés au chargement d'une page web[20].	15
2.5	Marché occupé par Google Analytics dans les domaines d'analyse, de tracking et de mesure d'audience sur le Web.	16
2.6	Logo de la solution Google Analytics[8].	16
2.7	Image de présentation de timeStats[12].	18
2.8	Page d'accueil de Ghostery[13].	18
2.9	Interface de base de Privacy manager[14].	19
2.10	Interface de TheGoodData[15].	20
2.11	Premier paragraphe de la page web de Noiszy[16].	21
2.12	Contrôle des actions face aux trackers de Privacy Badger[17].	22
2.13	Flux de données de Kraken.me[18].	22
3.1	Flux de données de l'extension.	24
3.2	Exemple d'URL et traitement	25
3.3	Technologies utilisées pour chaque partie	27
3.4	Téléchargement et enregistrement du contenu d'une page	29
3.5	Traitements du contenu des pages	31
3.6	Schéma des tables de la base de données	33
3.7	Tables de la base de données	34
3.7	Tables de la base de données (suite)	35
3.8	Maquette de la page de Profil	37
3.9	Maquette de la page de Trackers	38
3.10	Suite de la maquette de la page de Trackers	39

4.1	Maquette initiale et résultat final de la vue Wordcloud	41
4.2	Algorithme utilisé pour le Wordcloud	43
4.3	Maquette initiale et résultat final de la vue Wordcloud	45
4.4	Algorithme utilisé pour le Topics List	48
4.5	Maquette initiale et résultat final de la vue Most Watched	49
4.6	Algorithme utilisé pour les pages "Most Watched" et "Most Visited"	52
4.7	Maquette initiale et résultat final de la vue History	53
4.8	Algorithme utilisé pour les graphiques de la page "History"	56
4.9	Domaine activé, puis désactivé	57
4.10	Maquette initiale et résultat final d'une des vues Trackers	58
4.11	Maquette initiale et résultat final de la vue détaillée lors d'un clic sur un Tracker	58
4.12	Algorithme utilisé pour les données des pages "Trackers"	60
4.13	Maquette initiale et résultat final d'une vue Stats	61
4.14	Algorithme utilisé les données de la page "Stats"	62
5.1	Champ d'entrée des centres d'intérêts	64
5.2	Sélection d'intérêt sur un topic	64
5.3	20 URLs les plus regardées et leurs meilleurs mots selon TF-IDF	67

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 algorithmes 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 utilisés en entrée.

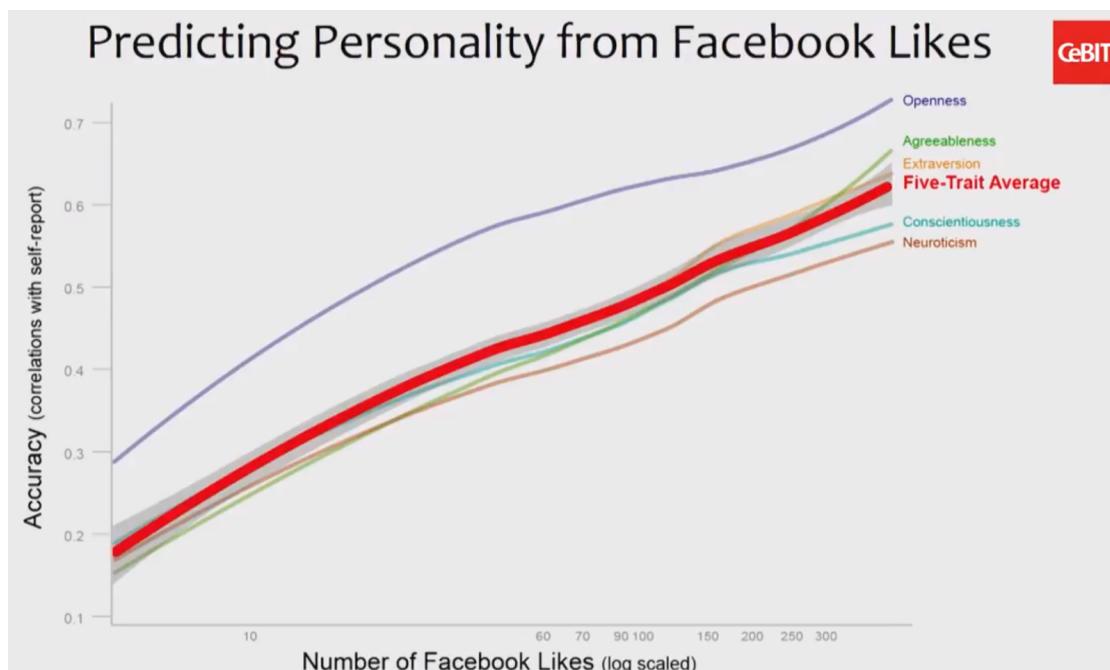


FIGURE 2.1 – Précision moyenne du modèle prédisant la personnalité d'un utilisateur en fonction du nombre de likes analysés[5].

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.

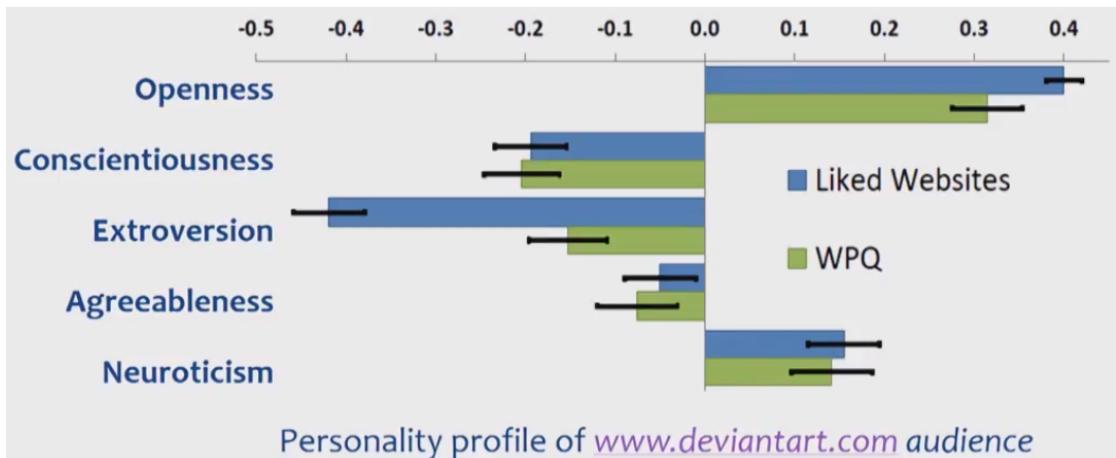


FIGURE 2.2 – Déviation de la personnalité moyenne estimée d'un visiteur régulier du site ““deviantart.com”” selon les cinq axes psychologiques employés[5].

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.

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

2.3 Trackers et Google Analytics

2.3.1 trackerackers

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.

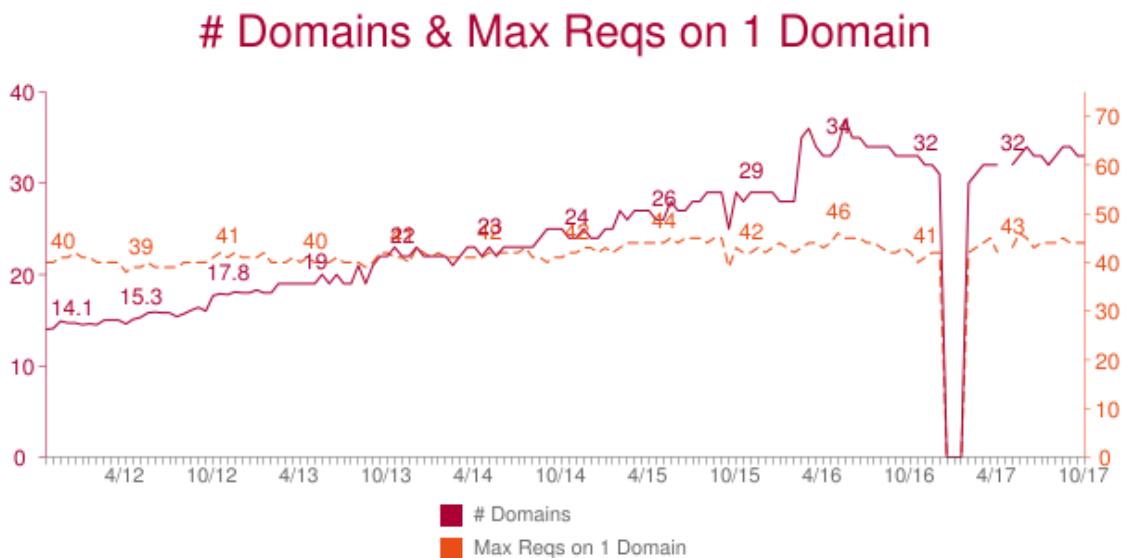


FIGURE 2.4 – Nombre moyen de domaines contactés au chargement d'une page web[20].

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.

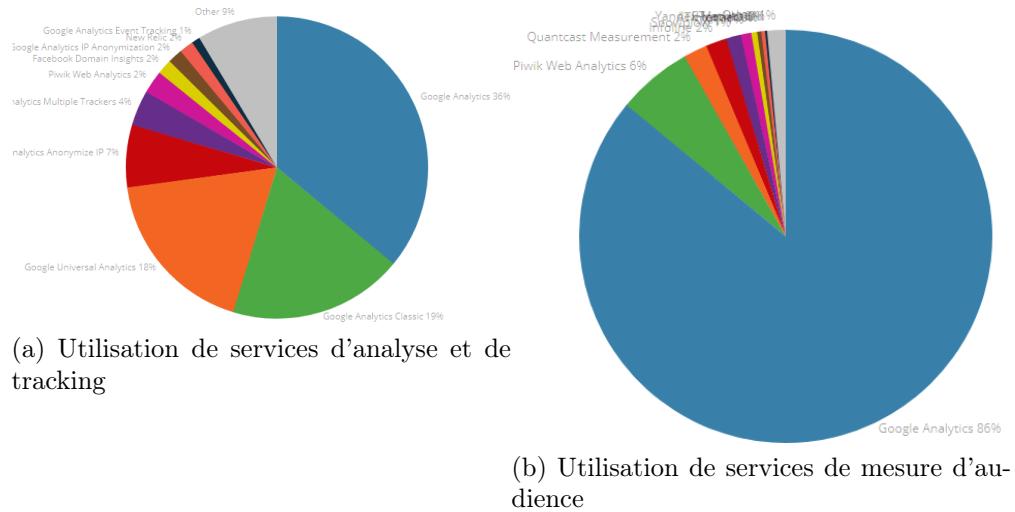


FIGURE 2.5 – Marché occupé par Google Analytics dans les domaines d'analyse, de tracking et de mesure d'audience sur le Web.

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 nombreuses 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 requise 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

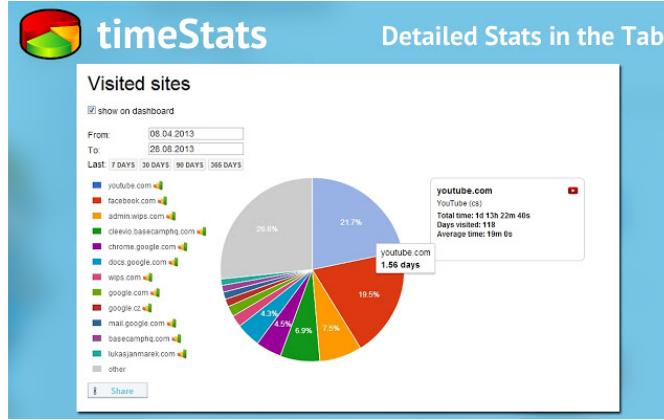


FIGURE 2.7 – Image de présentation de timeStats[12].

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.

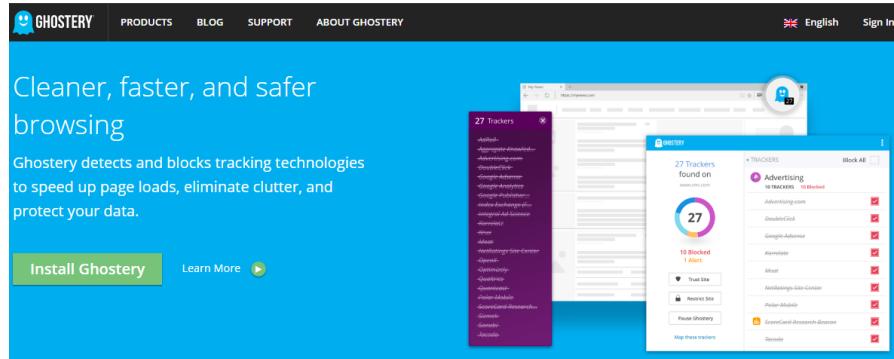


FIGURE 2.8 – Page d'accueil de Ghostery[13].

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.

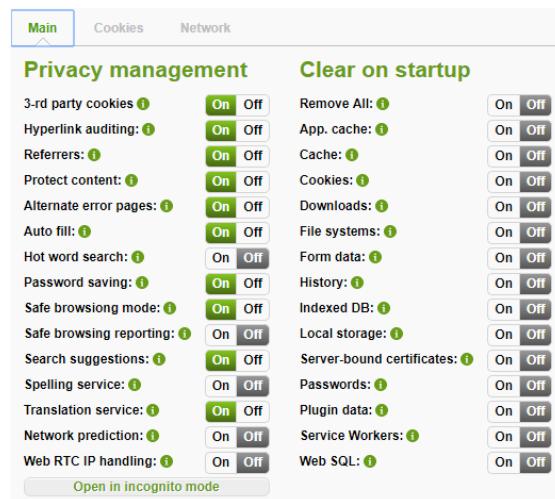


FIGURE 2.9 – Interface de base de Privacy manager[14].

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.

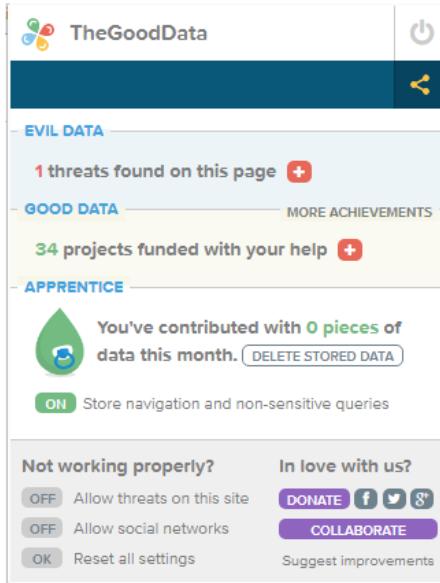


FIGURE 2.10 – Interface de TheGoodData[15].

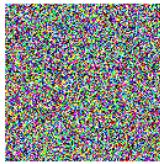
Privacy Badger

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.

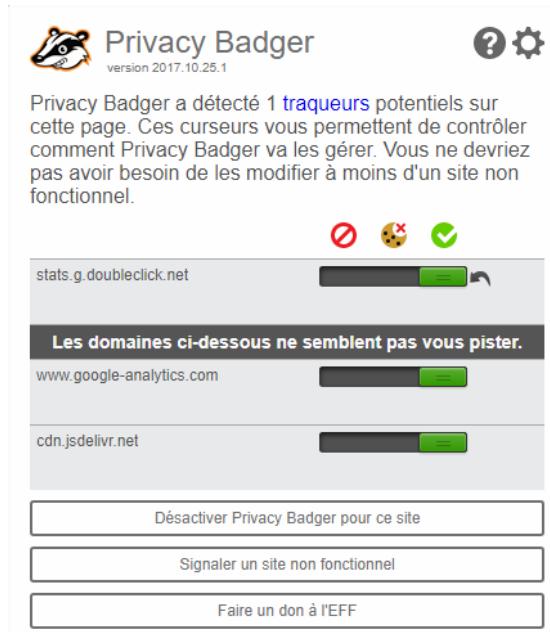


FIGURE 2.12 – Contrôle des actions face aux trackers de Privacy Badger[17].

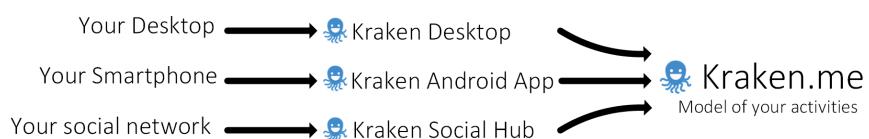


FIGURE 2.13 – Flux de données de Kraken.me[18].

Chapitre 3

Design du système

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 divulgateuse 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.

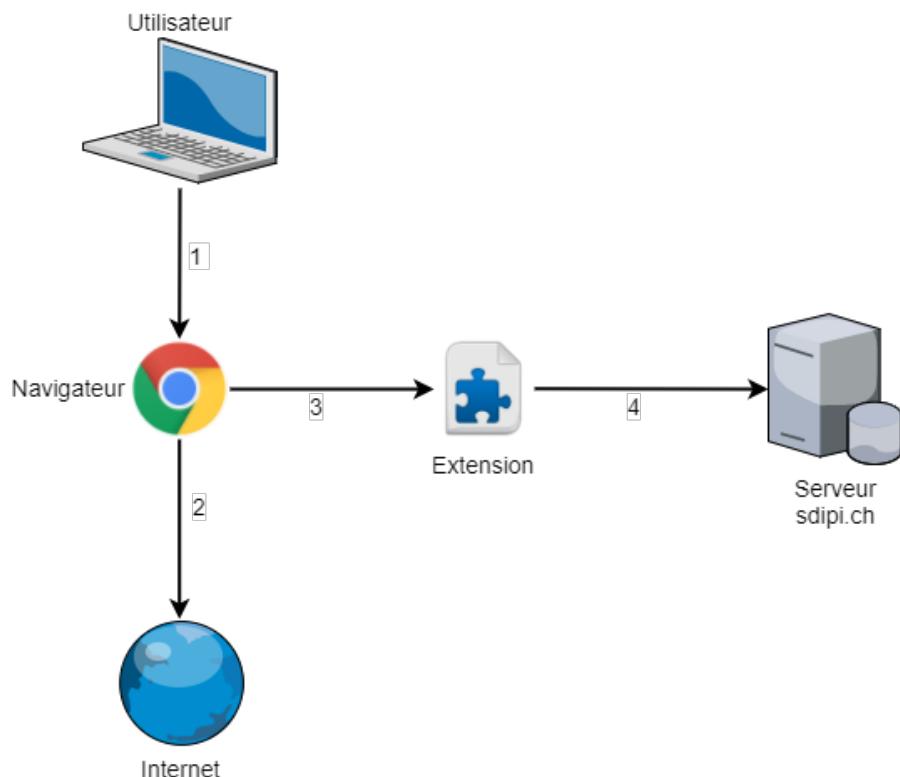


FIGURE 3.1 – Flux de données de l'extension.

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.

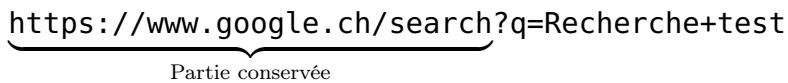

Partie conservée

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 Architecture

3.2.1 Stack technologique

L'ensemble des éléments constituant le projet peuvent être regroupés en 3 parties différentes où s'exécute le code.

La figure 3.3 montre les trois parties ainsi que les technologies utilisées dans chacune.

- L'extension comprend le code exécuté dans le navigateur du client, et qui communique avec l'API de Google Chrome afin de pouvoir récolter et envoyer les informations.
- L'interface comprend le code des diverses pages de visualisations montrées à l'utilisateur. On peut accéder à cette série de pages via un lien montré dans l'extension, ou par leur URL directement.
- Le serveur comprend le code exécuté notre machine.

3.2.2 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

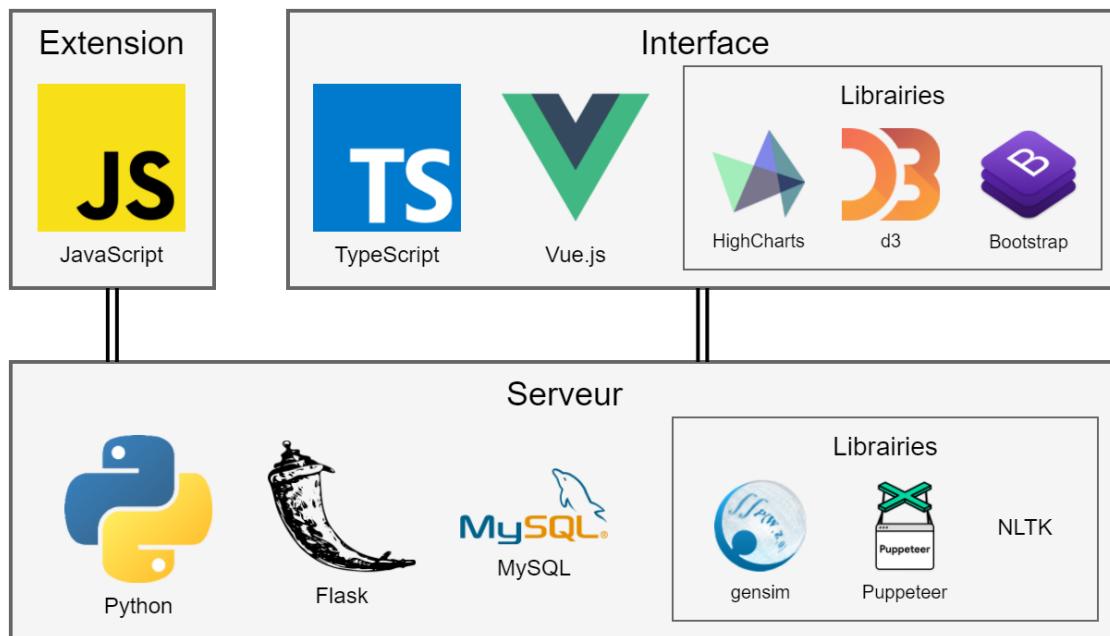


FIGURE 3.3 – Technologies utilisées pour chaque partie

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.

3.2.3 Serveur

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

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.

Une série d'opérations est ensuite effectuée sur le contenu de la page, afin de le rendre utilisable par les prochains algorithmes. La figure 3.4 montre les étapes qui entrent en compte dans le pré-traitement du contenu :

1. **Téléchargement** Le serveur lance, dans un navigateur virtual, le téléchargement de la page ainsi que l'exécution des scripts présents sur celle-ci. Une fois la page complètement chargée, on conserve le DOM de la page chargée.

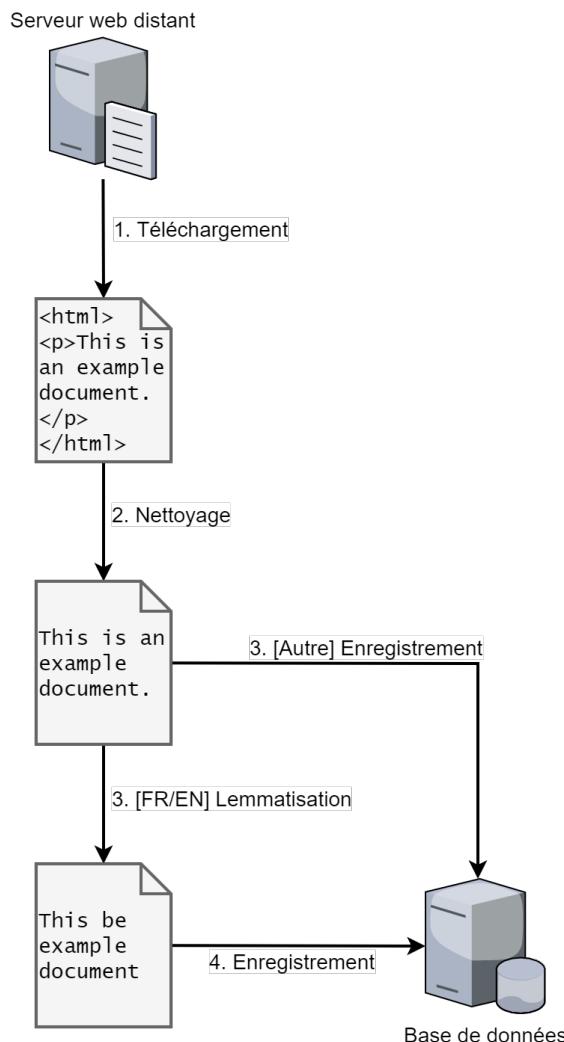


FIGURE 3.4 – Téléchargement et enregistrement du contenu d'une page

- 2. Nettoyage** À partir du HTML de la page, on ne cherche à garder que le texte de celle-ci, sans balise. Un parseur enlève toutes les balises `<script>` et `<style>`, puis ne garde que le contenu des éléments restants.
- 3. [FR/EN] Lemmatisation** Un détecteur de langue nous renseigne sur la langue du texte. Si celui-ci est en anglais ou en français, nous lemmatisons chaque mot du texte. Ce processus analyse lexicalement les mots présents, et tente de ramener chaque mot à une forme plus simple pour le représenter. Par exemple, le temps des verbes est changé en infinitif, et les noms communs perdent leur pluriel. La liste complète des traitements effectuée est en réalité bien plus longue, et propre à la langue du texte.

3/4. Enregistrement Si le texte n'est pas dans une langue supportée par la lemmatisation, ou après la lemmatisation du texte anglais ou français, celui-ci est enregistré dans la base de données. Tous les traitements futurs sur le contenu de la page se feront sur cette version-ci.

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.

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. Les traitements décrits ici ne sont pas effectués directement à la réception de données d'un client : Ils sont effectués régulièrement lorsque nécessaire.

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 Les traitements les plus lourds que nous effectuons prennent en entrée le contenu des pages visitées.

Le contenu de la page est analysé indépendamment par deux algorithmes différents, chacun permettant de révéler un type d'information différent.

La figure 3.5 montre les deux traitements effectués aux documents de la base de données.

1. Chaque document le serveur va charger la liste entière des contenus enregistrés des pages web, obtenues comme décrites à la section 3.2.3.

A2. Calcul TF La fréquence de chaque mot du document est calculée, puis enregistrée

A3. Calcul TF-IDF Une fois en possession de la fréquence de chaque mot dans chaque document, le poids final TF-IDF normalisé est calculé et enregistré dans la base de données, pour chaque mot à l'intérieur de chaque document.

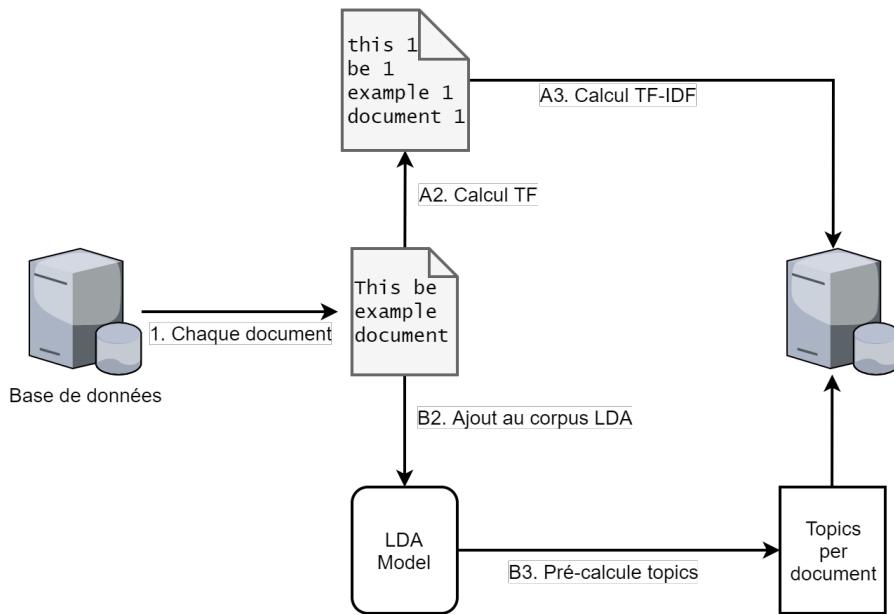


FIGURE 3.5 – Traitements du contenu des pages

B2. Ajout au corpus LDA Avant de générer un modèle, un traitement semblable à la branche A2 est effectuée pour chaque document. Une fois tous les documents chargés, on lance l'exécution de la génération du modèle LDA.

B3. Pré-calcule topics Une fois le modèle entraîné, processus qui peut facilement durer plusieurs heures, il est enregistré sur le disque. Après quoi, une multitude de requêtes sont effectuées sur le modèle afin de connaître déjà quels sont les topics les plus probables pour chaque page de la base de données. On enregistre les résultats à nouveau dans la base de données. Ce pré-calcul va accélérer considérablement les traitements futurs sur la reconnaissance des topics significatifs pour un utilisateur en fonction des pages web qu'il a visité.

TF-IDF TF-IDF est une méthode permettant de détecter quels sont les mots les plus importants dans un document parmi l'ensemble d'un corpus. La méthode consiste purement en l'analyse de la fréquence de chaque mot dans chaque document, et ne s'occupe absolument pas de la signification des mots.

LDA LDA est un modèle qui permet de générer un nombre de sujets, thèmes ou topics, en fonction du contenu textuel d'un corpus de documents. Le modèle suppose que chaque document parle de un ou plusieurs topics, et tente de les retrouver en se basant sur la fréquence d'utilisation de ses mots en comparaison

avec le reste du corpus de documents.

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.

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.

3.2.4 Base de données

Toutes les informations sont centralisées dans une base de données MySQL. La figure 3.6 montre le schéma global des tables de la base de données. L'utilisation de chaque table est décrite dans la section suivante, lors de leur accès. La figure 3.7 décrit chacun des champs des tables de la base de données.

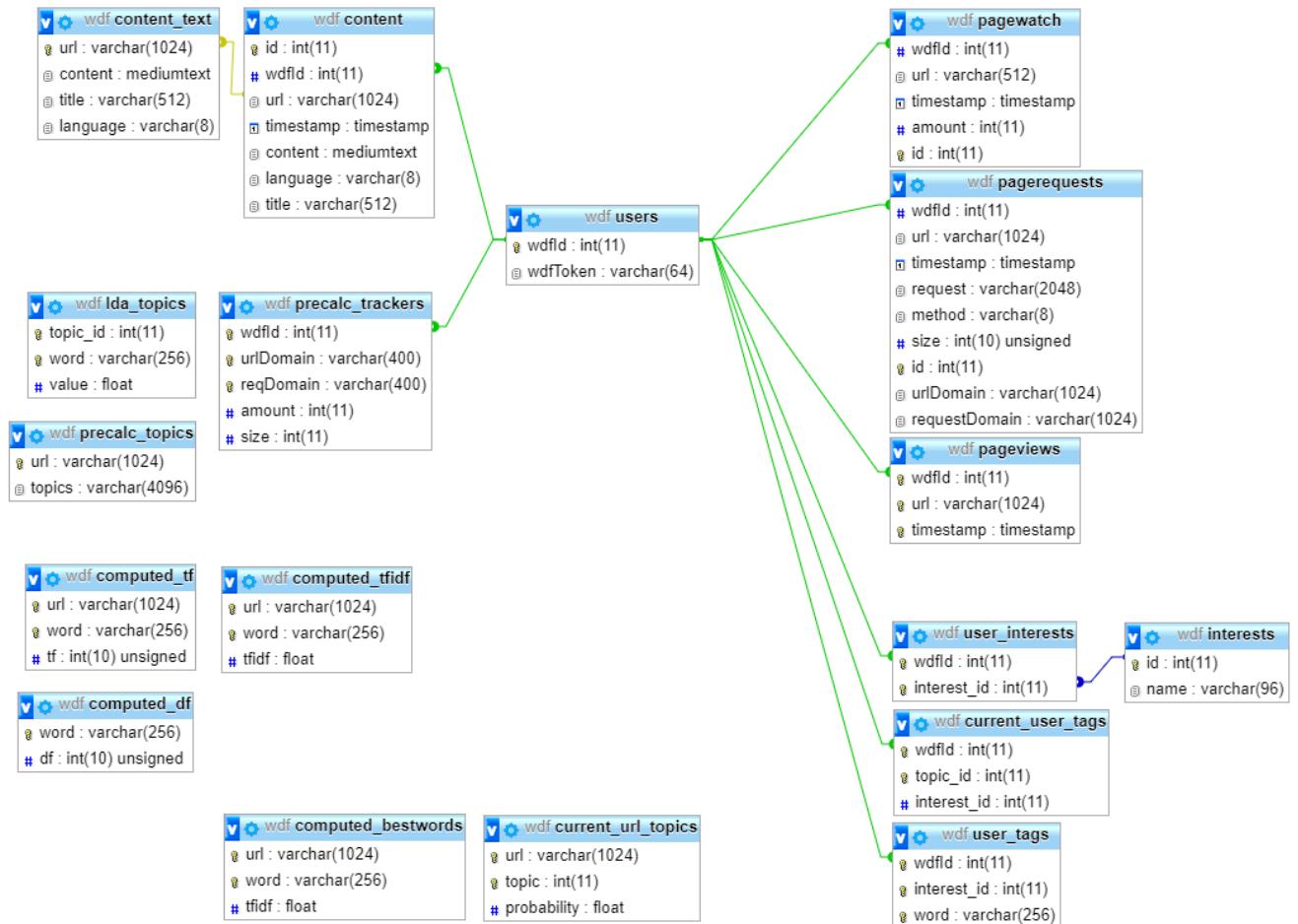


FIGURE 3.6 – Schéma des tables de la base de données

id	Clé primaire artificielle
wdfId	Identifiant ayant accédé
url	URL de la page
timestamp	Date téléchargée
content	Contenu DOM entier
language	Langue détectée du texte
title	Contenu de la balise title

(a) Table **content**

url	URL de la page
content	Contenu textuel lemmatisé
title	Contenu de la balise title
language	Langue détectée du texte

(b) Table **content_text**

wdfId	Numéro d'identifiant
wdfToken	Token du client

(c) Table **users**

wdfId	Numéro d'identifiant
url	URL de la page
timestamp	Date de visualisation
amount	Temps regardé [sec]
id	Clé primaire artificielle

(d) Table **pagewatch**

wdfId	Numéro d'identifiant
url	URL de la page
timestamp	Date de visualisation
request	URL requêtée
method	Méthode HTTP
size	Taille de la requête
id	Clé primaire artificielle
urlDomain	Domaine de l'URL actuelle
requestDomain	Domaine de l'URL requêtée

(e) Table **pagerequests**

wdfId	Numéro d'identifiant
url	URL de la page
timestamp	Date de visualisation

(f) Table **pageviews**

wdfId	Numéro d'identifiant
interest_id	Numéro d'intérêt

(g) Table **user_interests**

id	Numéro d'intérêt
name	Nom de l'intérêt

(h) Table **interests**

wdfId	Numéro d'identifiant
topic_id	Numéro du topic
interest_id	Numéro de l'intérêt

(i) Table **current_user_tags**

wdfId	Numéro d'identifiant
topic_id	Numéro du topic
words	Trois mots du topic

(j) Table **user_tags**

FIGURE 3.7 – Tables de la base de données

wdfId Numéro d'identifiant urlDomain Domaine de l'URL actuelle requestDomain Domaine de l'URL requêtée amount Nombre de requêtes size Taille totale des requêtes	url URL de la page topics JSON des topics associés
(k) Table precalc_trackers	(l) Table precalc_topics
topic_id Numéro du topic topics Mot associé value Probabilité du mot	
(m) Table lda_topics	
url URL de la page word Mot tf Term Frequency selon TF-IDF	word Mot df Document Frequency selon TF-IDF
(n) Table computed_tf	(o) Table computed_df
url URL de la page word Mot df Score final TF-IDF	
(p) Table computed_tfidf	
url URL de la page word Mot tfidf Score final TF-IDF	url URL de la page topic Numéro du topic probability Probabilité du topic pour la page
(q) Table computed_bestwords	(r) Table current_url_topics

FIGURE 3.7 – Tables de la base de données (suite)

3.2.5 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 trois onglets distincts, chacun tentant de représenter une partie des informations : Settings, Profil, et Trackers.

Settings

La page Settings laisse la possibilité à l'utilisateur de resneigner ces centres d'intérêt en les sélectionnant parmi une liste d'une centaine d'entre-eux. Cette centaine d'intérêts sont ceux que Google utilise pour "classifier" les visiteurs de sites web utilisant Google Analytics, nous estimons donc que ces intérêts font sens.

Profil

La page Profil cherche à montrer le résultat de l'analyse des pages visitées par l'utilisateur, tentant de retrouver et de lui montrer quels sont ses centres d'intérêts. La figure 3.8 montre les différentes vues prévues initialement :

- 1 Word Cloud** Cette vue cherche à mettre rapidement en valeur les mots les plus consultés par l'utilisateur en affichant un nuage de mots, où les plus grand seraient les plus vus.
- 2 Interests graph** Cette visualisation cherche à rassembler les mots fréquemment lus par l'utilisateur en topics, eux-même liés entre eux. Le but est de montrer une synthèse du Word Cloud.
- 3 Website themes** Cette partie cherchait à montrer les mots redondants ainsi que les thèmes trouvés sur certains sites web.
- 4 Most visited sites** Ce graphique en barres cherche à montrer à l'utilisateur quels sont les sites qu'il a le plus souvent visité, c'est-à-dire ouverts l'URL, peu importe le temps passé sur chaque site.
- 5 Most watched sites** Ce graphique, contrairement au 4, cherche à montrer à l'utilisateur le temps total passé à regarder chaque page.
- 6 History of websites** Ce graphique cherche à montrer à l'utilisateur la fluctuation de sa visite de sites web sur la durée.
- 7 History of interests** Ce graphique cherche à mettre en lumière les sujets les plus visités par l'utilisateur sur une période de temps afin de potentiellement détecter des tendances ou des changements dans son comportement.

Profile

1 Word Cloud

2 Interests graph

3 Websites themes

website	keywords	themes
reddit.com	technology, news, world, comments	technology, news
heia-fr.ch	studies, computer, thesis, course	school
20minut.es.ch	suisse, vote, fribourg, minute	news

4 Most visited sites

site	#pages
reddit.com	3'819
20minut.es.ch	2'100
heia-fr.ch	2'655
stackoverflow.com	2'194
www.ch	1'987

5 Most watched websites

site	minutes
youtube.com	2'123
reddit.com	1'429
stackoverflow.com	8'13
www.ch	7'94
www.in	6'18

6 History of websites

7 History of interests

FIGURE 3.8 – Maquette de la page de Profil

Trackers

La page Trackers montre la liste des différents trackers contactés au cours de la navigation, ainsi que les domaines les ayant contactés.

Les requêtes effectuées depuis une page vers le même domaine ne sont pas comptées car on estime qu'il s'agit de trafic que l'on sait qui va prendre place, peu importe la page contactée : Il est évident qu'elle va chercher à charger du contenu provenant du même domaine.

La figure 3.9 montre les 4 premières visualisations conceptualisées de la page Trackers, et la figure 3.10 montre les 4 dernières.

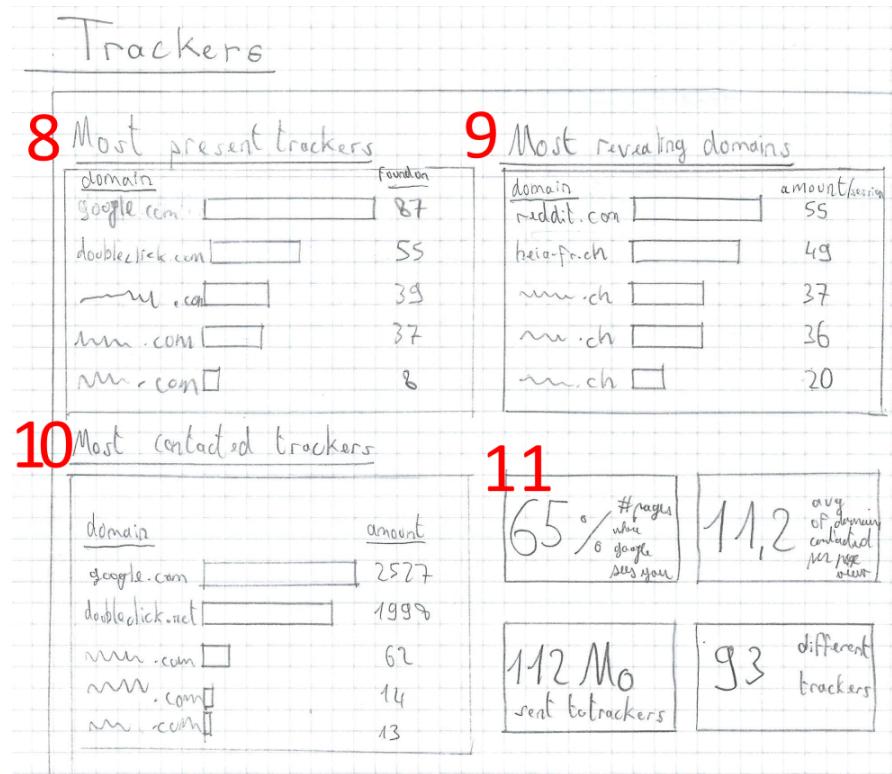


FIGURE 3.9 – Maquette de la page de Trackers

8 Most present trackers Ce graphique en barres montrait les trackers potentiels contactés depuis le plus grand nombre de pages différentes.

9 Most revealing trackers Ce graphique en barres montrait une moyenne par domaine du nombre de requêtes effectuées vers des trackers potentiels.

10 Most contacted trackers Ce graphique en barres montrait les potentiels trackers les plus contactés au total, depuis n'importe quelle page.

11 Stats Quelques nombres montrent des statistiques générales de l'utilisateur afin de lui faire prendre compte de certaines mesures. Par exemple, la taille totale d'informations envoyées aux trackers potentiels, ou le nombre de ceux-ci contactés.

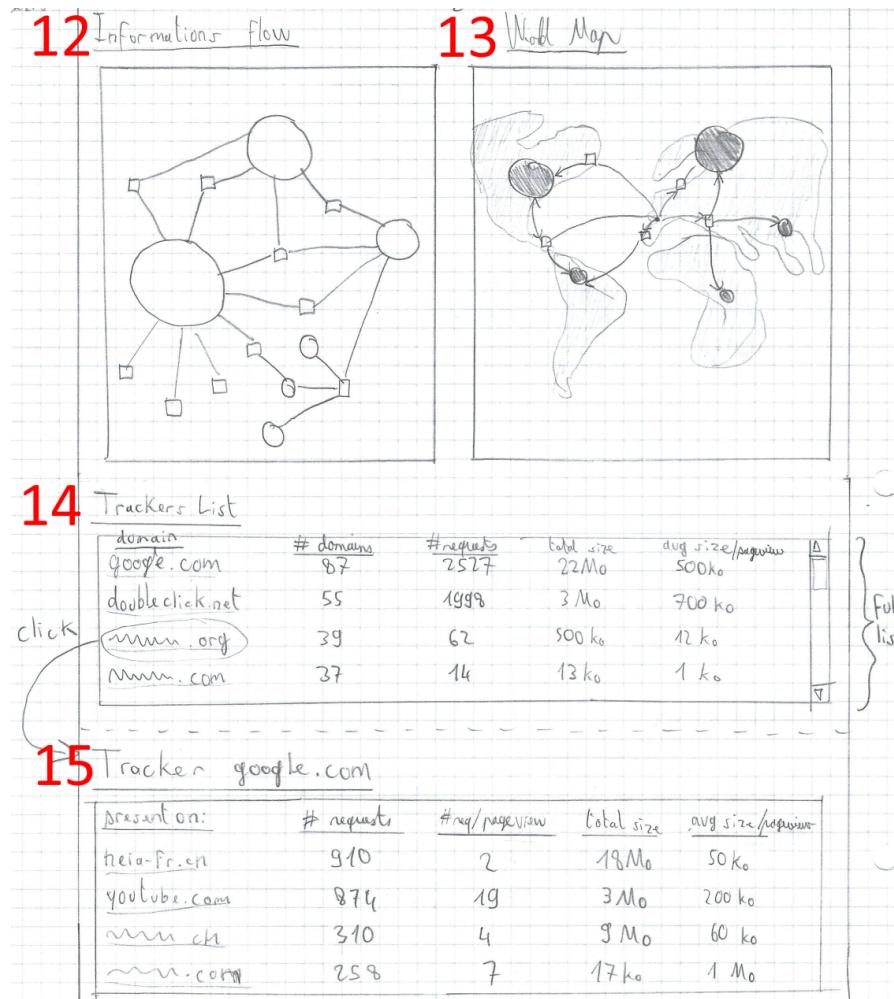


FIGURE 3.10 – Suite de la maquette de la page de Trackers

12 Informations flow Cette visualisation sous forme de graphe cherche à montrer quels sont les domaines les plus connectés entre eux. Le but est de rassembler des domaines et sous-domaines afin de montrer lesquels de ceux-ci communiquent le plus.

13 World map Cette visualisation est très semblable à la précédente. Les domaines sont à présent placés sur leur emplacement géographique afin de se rendre compte du trafic réel physique engendré par ces requêtes.

- 14 Trackers list** Cette partie montre de manière exhaustive l'ensemble des requêtes effectuées vers un domaine. Le but est de permettre aux utilisateurs curieux de parcourir l'ensemble des données de manière plus fine. Un clic sur un tracker ouvre la vue 15.
- 15 Selected tracker** Ce tableau s'ouvre en sélectionnant un tracker potentiel de la vue 14. Il montre l'ensemble des domaines ayant contacté le potentiel tracker sélectionné.

Topics graph Le "topics graph" cherche à rassembler les mots en thèmes, et monte d'une manière plus synthétique les thèmes estimés que l'utilisateur parcourt fréquemment. À chaque thème est lié un ou plusieurs mots, qui représentent le thème d'une manière générale. Chaque cercle du graphe représente soit un thème, soit un mot.

Le but de cette visualisation est de montrer que nous pouvons déduire des thèmes et ainsi montrer un traitement plus fin des intérêts de l'utilisateur, que simplement additionner une liste de mots. Dans le marketing, les thèmes découverts pourraient être utilisés pour labelliser les utilisateurs à qui faire apparaître une publicité.

Chapitre 4

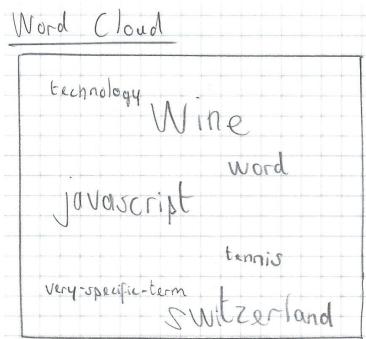
Développement des vues

4.1 Wordcloud

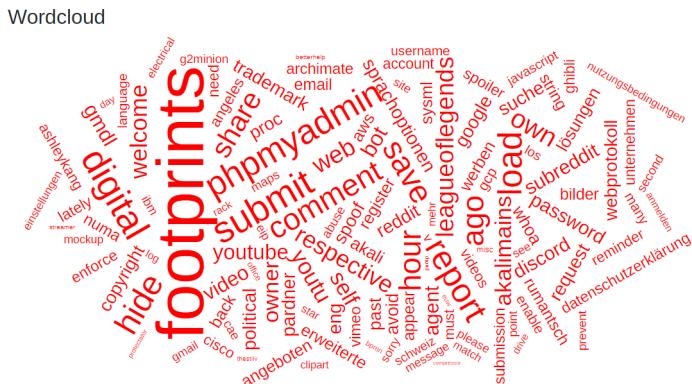
4.1.1 Concept

Le wordcloud montre à l'utilisateur la liste des mots qu'il lit le plus fréquemment. La visualisation est un amassage de mots de différentes tailles, placés d'une manière aléatoire sur un rectangle. Les mots les plus lus ont une taille plus grande afin d'attirer l'attention de l'utilisateur.

Cette visualisation cherche à donner très rapidement une impression générale des thèmes que l'utilisateur parcourt lors de sa navigation.



(a) Maquette initiale de la vue



(b) Exemple de résultat final

FIGURE 4.1 – Maquette initiale et résultat final de la vue Wordcloud

La figure 4.1 montre la différence entre la vue imaginée initialement et le résultat final.

4.1.2 Données

Sources

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table **pagewatch** (figure 3.7a).

Poids TF-IDF : Poids final selon l'algorithme TF-IDF de chaque mot. Ces données sont stockées dans la table **computed_tfidf** (figure 3.7p).

Algorithm

Afin de déterminer quels sont les mots affichés ainsi que leur taille sur la visualisation, on assigne un "poids" à chaque mot.

La figure 4.2 illustre le fonctionnement de l'algorithme utilisé :

- A** On calcule le poids de chaque mot dans chaque document en effectuant la méthode de TF-IDF.
- B** On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. Cette opération est effectuée sur le serveur, et l'interface obtient ce résultat en appelant l'endpoint **/api/mostWatchedSites** du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web visitées, comprenant entre autres pour chaque page :
 - Son URL
 - Le temps total de visite, en secondes
 - Une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1)

On initialise un dictionnaire qui va contenir le poids de chaque mot.

- C** Pour chaque page web, on multiplie l'indice TF-IDF de chaque mot avec le temps de visualisation de la page. On additionne ce résultat au poids actuel du mot.

Une fois tous les mots de toutes les pages web traités, nous sommes en possession d'un dictionnaire nous indiquant le poids final de chaque mot. Ce poids est donc égal à la somme de l'indice TF-IDF du mot sur chaque page multiplié par le temps de visite sur cette page.

- D** On trie les mots par leur poids final, et on ne conserve que les 200 premiers. Il s'agira des 200 mots présents sur le wordcloud.

Pour chacun des 200 mots, leur taille sur le Wordcloud est égale à leur poids final.

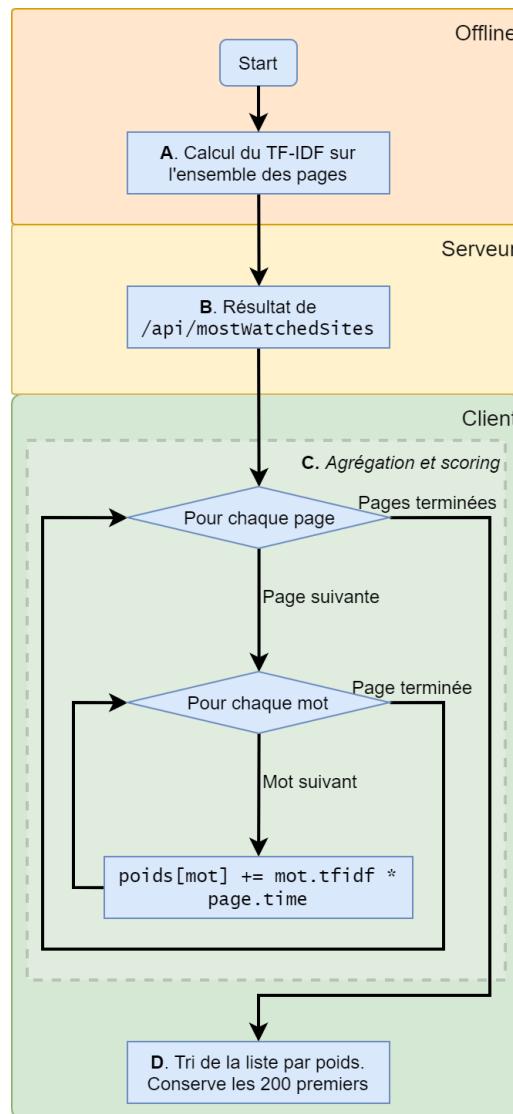


FIGURE 4.2 – Algorithme utilisé pour le Wordcloud

4.1.3 Implémentation

Serveur

La somme du temps de visualisation des pages est calculée en direct par une commande MySQL, elle est donc constamment à jour.

Le poids TF-IDF de chaque mot est stocké dans la base de données, mais n'est pas constamment rafraîchi. L'opération de calcul des poids TF-IDF est une opération ponctuelle qui doit être lancée sur l'entièreté de la base de données par

l'administrateur. Cette opération ne nécessite cependant pas de redémarrage du serveur.

La concaténation de ces résultats (ainsi que certains autres qui ne sont pas utilisés par cette visualisation) est servie par l'endpoint `/api/mostWatchedSites` dans une liste en JSON.

Interface

La page va s'occuper d'agréger les résultats reçus du serveur. Ensuite, elle utilise les librairies `d3-cloud` ainsi que `d3` pour générer la visualisation du Wordcloud.

4.2 Topics List

4.2.1 Concept

Le "Topics List" cherche à rassembler les mots en thèmes, et montre d'une manière plus synthétique les thèmes estimés que l'utilisateur parcourt fréquemment. À chaque thème est lié un ou plusieurs mots, qui représentent le thème d'une manière générale. Chaque cercle du graphe représente soit un thème, soit un mot.

Le but de cette visualisation est de montrer que nous pouvons déduire des thèmes et ainsi montrer un traitement plus fin des intérêts de l'utilisateur, que simplement additionner une liste de mots. Dans le marketing, les thèmes découverts pourraient être utilisés pour labelliser les utilisateurs à qui faire apparaître une publicité.

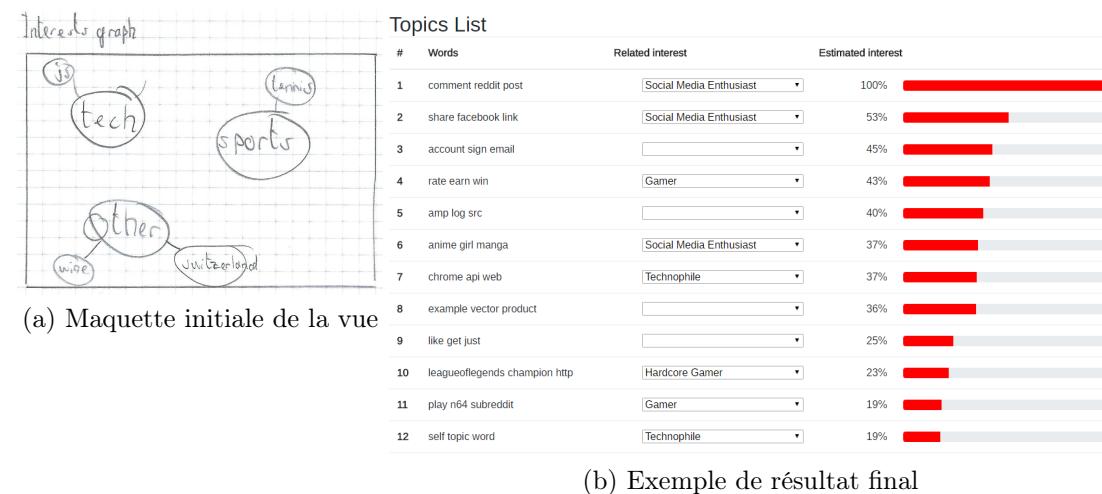


FIGURE 4.3 – Maquette initiale et résultat final de la vue Wordcloud

La figure 4.3 montre la différence entre la vue imaginée et le résultat final. On note ici que le principe même de la vue ainsi que son nom sont différents qu'initialement.

4.2.2 Données

Sources

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table **pagewatch** (figure 3.7a).

Topics LDA : Liste des topics générés par le modèle LDA. Ces données sont stockées dans la table `lda_topics` (figure 3.7m).

Topics par page : Liste pré-calculée des topics trouvés pour chaque page. Ces données sont stockées dans la table `current_url_topics` (figure 3.7r).

Intérêts utilisateur : Liste des intérêts renseignés par l'utilisateur. Ces données sont stockées dans la table `user_interests` (figure 3.7g).

Correspondances topic-intérêt : Liste des correspondances entre topic et intérêt renseignés par l'utilisateur. Ces données sont stockées dans la table `current_user_tags` (figure 3.7i).

Algorithmme

Afin de déterminer quels sont les topics affichés ainsi que leur intérêt estimé, on assigne un "score" à chaque topic pour l'utilisateur.

L'algorithme suivant, illustré par la figure 4.4, est appliqué aux données sources :

A On entraîne un modèle LDA avec un nombre défini de topics (typiquement 100) sur le contenu de l'ensemble des pages web, une page web représentant un document.

Une fois le modèle LDA entraîné, on lui demande la liste des 100 topics générés par leur représentation en 5 mots. Cette liste de topics est enregistrée dans la base de données.

B Pour chaque page enregistrée, on demande au modèle LDA quels sont les 5 topics les plus probables avec leur score de probabilité. Ces informations sont également enregistrées dans la base de données. Jusqu'ici, toutes ces opérations sont donc déjà calculées et se font avant le lancement du serveur. Elles ne sont pas mises à jour en temps réel.

C On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. Cette opération est effectuée sur le serveur, et l'interface obtient ce résultat en appelant l'endpoint `/api/mostWatchedSites` du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web visitées, comprenant entre autres pour chaque page :

- Son URL
- Le temps total de visite, en secondes
- Une liste des topics les plus significatifs selon le modèle LDA ainsi que leur probabilité

D L'endpoint `/api/allTopics` renvoie la liste des topics générés par LDA ainsi que leur numéro d'identifiant.

E L'endpoint `/api/getCurrentTags` renvoie la liste des associations que l'utilisateur a créée pour le modèle LDA courant. Il s'agit d'une liste de couples `topicId ↔ interestId`.

F L'endpoint `/api/interestsList` renvoie la liste des 101 intérêts globaux à tous les utilisateurs.

G On initialise un dictionnaire qui va contenir le score de chaque topic.

Pour chaque page web, on multiplie la probabilité de chaque topic LDA avec le temps de visualisation de la page. On additionne ce résultat au score actuel du topic.

Une fois tous les topics de toutes les pages web traités, nous sommes en possession d'un dictionnaire nous indiquant le score final de chaque topics.

Ce score est donc égal à la somme de la probabilité du topic sur chaque page multiplié par le temps de visite sur cette page.

H On trie les topics par leur score final, et on ne conserve que les 20 premiers. Il s'agira des 20 topics présents sur la page.

I On sélectionne les centres d'intérêt de l'utilisateur, ainsi que les associations qu'il a déjà créée pour le modèle LDA actuel. On ajoute les associations aux topics de l'interface.

4.2.3 Implémentation

Serveur

La somme du temps de visualisation des pages est calculée en direct par une commande MySQL, elle est donc constamment à jour.

Le modèle LDA est enregistré sur le disque local, et les résultats pré-calculés sont stockés dans la base de données, tout ceci n'est donc pas constamment rafraîchi. L'opération d'entraînement du modèle LDA est une opération ponctuelle qui doit être lancée sur l'entièreté de la base de données par l'administrateur. Cette opération nécessite le redémarrage du serveur, car de nombreuses mesures temporaires sont touchées.

La concaténation de ces résultats (ainsi que certains autres qui ne sont pas utilisés par cette visualisation) est servie par l'endpoint `/api/mostWatchedSites` dans une liste en JSON. L'endpoint `/api/interestsList` est utilisé pour afficher le nom des centres d'intérêts, et l'endpoint `/api/getCurrentTags` donne l'ensemble des associations que l'utilisateur a créée entre ses centres d'intérêts, et les topics LDA actuels.

Interface

La page va s'occuper d'agrégner les résultats reçus du serveur. La liste est ensuite générée sous forme d'un tableau HTML en passant par un composant Vue personnalisé. Les barres d'intérêt sont des éléments `progressbar` venant de Bootstrap.

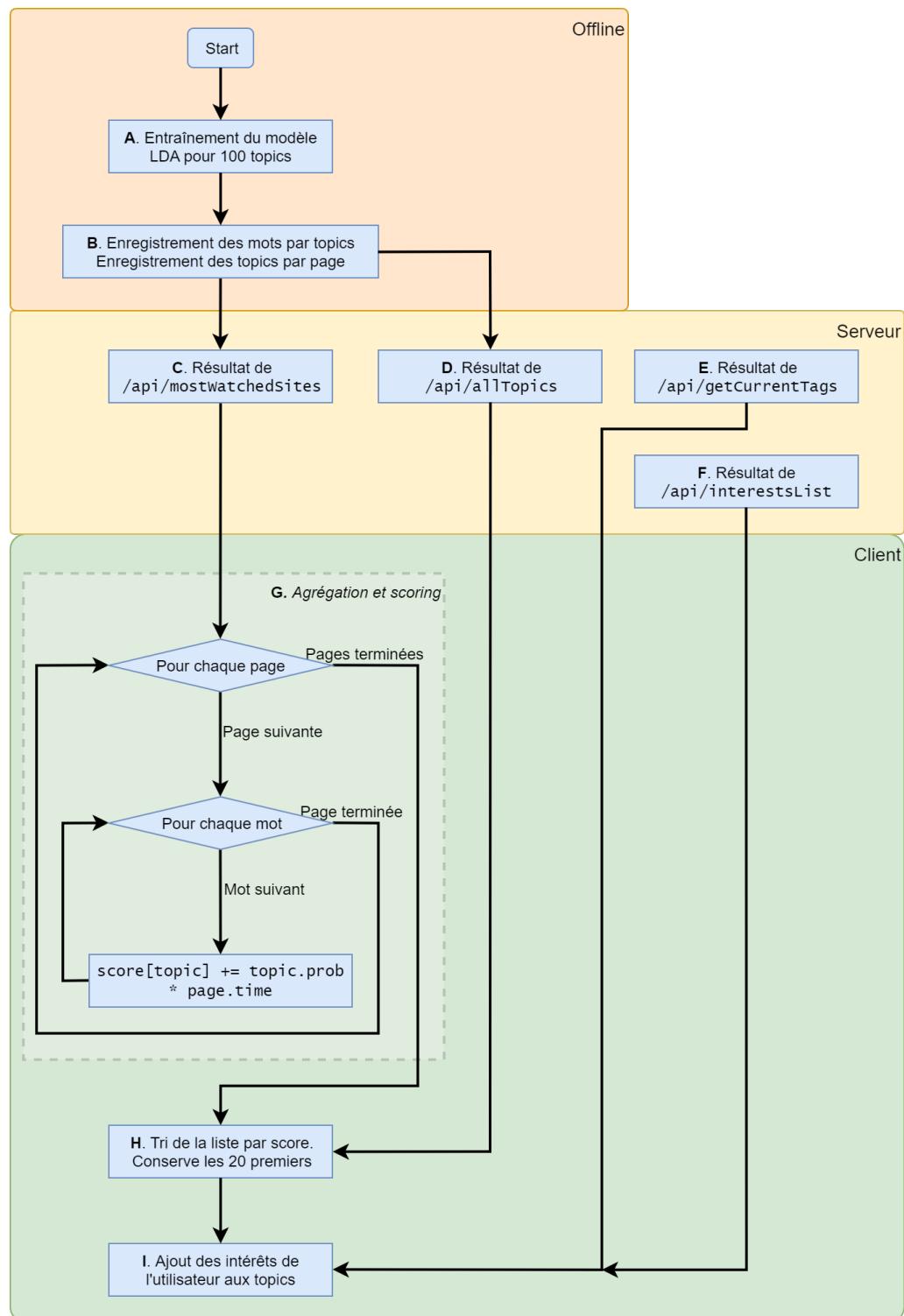


FIGURE 4.4 – Algorithme utilisé pour le Topics List

4.3 Most Watched

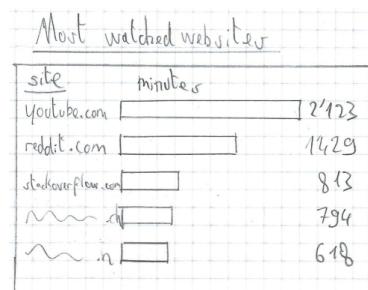
4.3.1 Concept

Les pages "Most Watched" et "Most Visited" montrent deux informations, mais sous une forme semblable. Ces pages affichent quelles pages web et les domaines que l'utilisateur a visité le plus. Plus précisément, "Most Watched" s'intéresse au temps réel passé à lire chaque page, et "Most Visited" s'intéresse au nombre d'ouvertures de l'URL.

Le but est ici de faire prendre conscience à l'utilisateur qu'il est possible de se rendre compte de son activité une page web, et un grand nombre d'ouvertures d'un lien ne veut pas forcément dire un grand intérêt pour cette page.

On profite également de cet espace pour afficher les mots relatifs aux pages web qui ont le plus d'intérêt, afin de montrer qu'il est possible de déterminer quels sont les mots importants d'une page web simplement en la comparant au contenu des autres pages.

La figure 4.5 montre la différence entre la vue imaginée et le résultat final.



(a) Maquette initiale de la vue

#	Page	Keywords	Time
1	http://vdf.sdpip.ch/	digital footprints web	5h 20min
2	https://www.reddit.com/r/videos/	ago comment ghibli	2h 54min
3	https://www.draw.io/	archimate aws bpmn	1h 57min
4	http://vdf.sdpip.ch/phpmyadmin/sql.php	enable javascript language	1h 46min
5	https://discordapp.com/channels/	account back copyright	1h 31min
6	https://www.reddit.com/r/leagueoflegends/	ago angeles ashleykang	1h 28min
7	https://www.reddit.com/	abuse agent appear	1h 21min
8	https://www.google.ch/search	angeboten anmelden bilder	1h 18min
9	https://s3-fr.gladiatus.gameforge.com/game/index.php		1h 14min
10	https://twitter.com/	angemeldet bleiben erfahren	1h 0min

(b) Exemple de résultat final

FIGURE 4.5 – Maquette initiale et résultat final de la vue Most Watched

Comme certaines pages web demandent une connexion pour être visualisées, par exemple la page d'accueil de <https://www.facebook.com>, nous omettons volontairement une liste de pages web dans ce classement, car nous n'avons pas d'informations intéressante sur leur contenu à montrer. En effet, nous ne téléchargeons volontairement pas de copie de la page vue par l'utilisateur pour des questions de protection de données privées. Notre serveur télécharge la version publique de l'URL visitée par l'utilisateur pour en déterminer son contenu. Ainsi, il ne fait pas sens d'analyser le contenu des pages générées dynamiquement par l'utilisateur.

4.3.2 Données

Sources

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table **pagewatch** (figure 3.7a).

Nombre de visites : Nombre total d'ouvertures de chaque URL. Ces données sont stockées dans la table **pageviews** (figure 3.7f).

Poids TF-IDF : Poids final selon l'algorithme TF-IDF de chaque mot. Ces données sont stockées dans la table **computed_tfidf** (figure 3.7p).

Algorithm

Afin de déterminer quels sont les pages et les domaines affichés, on demande au serveur la liste triée des URLs les plus regardées et ouvertes.

L'algorithme suivant, illustré par la figure 4.6, est appliqué aux données sources :

- A** On calcule le poids de chaque mot dans chaque document en effectuant la méthode de TF-IDF.
- B** On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. Cette opération est effectuée sur le serveur, et l'interface obtient ce résultat en appelant l'endpoint **/api/mostWatchedSites** du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web regardées, comprenant entre autres pour chaque page :
 - Son URL
 - Le temps total de visite, en secondes
 - Une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1)
- C** On effectue la somme du nombre d'ouvertures de chaque page visitée. Cette opération est effectuée sur le serveur, et l'interface obtient ce résultat en appelant l'endpoint **/api/mostWatchedSites** du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web visitées, comprenant entre autres pour chaque page :
 - Son URL
 - Le nombre total d'ouvertures
 - Une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1)
- D** Une fois la liste des pages les plus regardées obtenue, on ne conserve que les 10 premières d'entre-elles pour des raisons visuelles. Ces 10 premières pages sont alors affichées.

- E** Ensuite, on cherche à agréger le temps de visualisation par domaine plutôt que par page, afin d'avoir une vue d'ensemble. On additionne donc le temps passé à regarder les pages d'un même domaine.
- F** On trie la nouvelle liste de domaines créée par leur temps total de visualisation, et on ne garde également que les 10 premiers d'entre-eux pour les afficher.
- G** On s'occupe ensuite du traitement du nombre d'ouvertures de chaque page. On ne conserve également que les 10 plus ouvertes d'entre-elles pour des raisons visuelles, elles sont alors affichées dans la liste.
- H** Ensuite, on agréger le nombre d'ouvertures par domaine plutôt que par page, afin d'avoir une vue d'ensemble. On additionne donc le temps passé à regarder les pages d'un même domaine.
- I** On trie la nouvelle liste de domaines créée par leur temps total de visualisation, et on ne garde également que les 10 premiers d'entre-eux pour les afficher.

4.3.3 Implémentation

Serveur

La somme du temps de visualisation des pages ainsi que du total d'ouvertures est calculée en direct par une commande MySQL, elle est donc constamment à jour.

Le poids TF-IDF de chaque mot est stocké dans la base de données, mais n'est pas constamment rafraîchi. L'opération de calcul des poids TF-IDF est une opération ponctuelle qui doit être lancée sur l'entièreté de la base de données par l'administrateur. Cette opération ne nécessite cependant pas de redémarrage du serveur.

La concaténation des résultats du temps total de visualisation ainsi que du nombre d'ouvertures est servi par respectivement l'endpoint `/api/mostWatchedSites`, et `/api/mostVisitedSites`.

Interface

La page va s'occuper d'agréger les résultats reçus du serveur. Chaque liste est ensuite générée sous forme d'un tableau HTML en passant par un composant Vue personnalisé. Les barres relatives à la quantité exprimée par chaque tableau ajoutent un élément de comparaison visuel, et sont des éléments `progressbar` venant de Bootstrap.

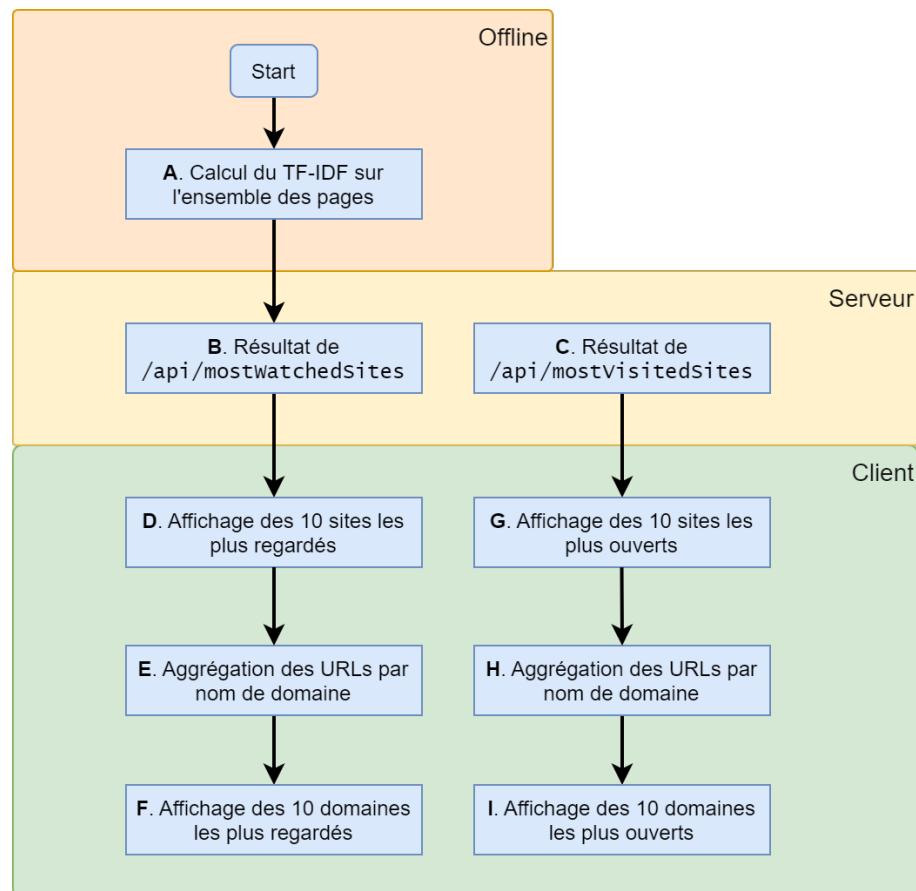


FIGURE 4.6 – Algorithme utilisé pour les pages "Most Watched" et "Most Visited"

4.4 History

4.4.1 Concept

La page "History" permet de montrer à l'utilisateur la variation de ses habitudes au cours du temps durant lequel il a utilisé l'extension. Deux graphiques sont présents sur cette page : Le premier montre la tendance à visiter des pages relatives à certains topics, et l'autre montre la tendance dans la visite de pages contenant certains mots particuliers. Le but est ici de détecter d'éventuels intérêts passagers dans le temps.

La figure 4.7 montre la différence entre la vue imaginée et le résultat final.

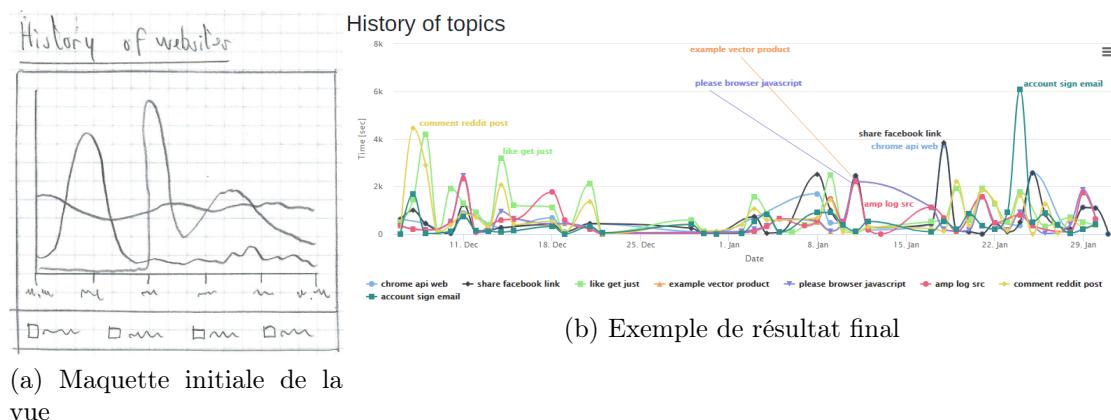


FIGURE 4.7 – Maquette initiale et résultat final de la vue History

4.4.2 Données

Sources

Les données source servant à constituer cette visualisation sont :

Temps de visualisation : Temps de visualisation total de chaque page. Ces données sont stockées dans la table `pagewatch` (figure 3.7a).

Historique de visualisation : Temps passé par jour sur chaque URL. Les données non agrégées proviennent de la table `pageviews` (figure 3.7d).

Poids TF-IDF : Poids final selon l'algorithme TF-IDF de chaque mot. Ces données sont stockées dans la table `computed_tfidf` (figure 3.7p).

Topics LDA : Liste des topics générés par le modèle LDA. Ces données sont stockées dans la table `lda_topics` (figure 3.7m).

Algorithme

Afin de déterminer quels sont les topics et les mots cumulant le plus d'intérêt, il est nécessaire de disposer de plusieurs sources de données et de les assembler afin d'arriver au résultat voulu. Ceci se fait en plusieurs étapes, distribuées entre le serveur et client.

L'algorithme suivant, illustré par la figure 4.8, est appliqué aux données sources :

- A** On calcule le poids de chaque mot dans chaque document en effectuant la méthode de TF-IDF.
- B** On entraîne un modèle LDA avec un nombre défini de topics (typiquement 100) sur le contenu de l'ensemble des pages web, une page web représentant un document.
Une fois le modèle LDA entraîné, on lui demande la liste des 100 topics générés par leur représentation en 5 mots. Cette liste de topics est enregistrée dans la base de données.
- C** Pour chaque page enregistrée, on demande au modèle LDA quels sont les 5 topics les plus probables avec leur score de probabilité. Ces informations sont également enregistrées dans la base de données. Jusqu'ici, toutes ces opérations sont donc déjà calculées et se font avant le lancement du serveur. Elles ne sont pas mises à jour en temps réel.
- D** On demande à la base de données de grouper le temps de visionnage (en secondes) en une somme par jour et par URL différente. Ceci se fait au travers d'une commande MySQL, et est calculé en temps réel.
- E** Le résultat de l'étape précédente est disponible via l'endpoint `/api/historySites`.
- F** On effectue la somme du temps que l'utilisateur a passé à regarder chaque page visitée. L'interface obtient ce résultat en appelant l'endpoint `/api/mostWatchedSites` du serveur. Le résultat de cet appel est une liste de l'ensemble des pages web regardées, comprenant entre autres une liste des mots les plus significatifs selon TF-IDF ainsi que leur poids TF-IDF (normalisé entre 0 et 1) pour chaque page.
- G** L'endpoint `/api/allTopics` renvoie la liste des topics générés par LDA ainsi que leur numéro d'identifiant.
- H** On cherche à savoir quels sont les mots où lesquels l'utilisateur a montré le plus d'intérêt afin de les afficher sur le graphe. Pour ceci, on multiplie la valeur TF-IDF de chaque mot par le temps passé à visualiser la page. La somme de ce calcul sur toutes les pages va nous donner l'"intérêt" final de l'utilisateur pour un mot particulier. On ne gardera ici que les 8 mots avec le plus grand intérêt estimé.

- I Finalement, pour chacun des 8 mots retenus, on affiche leur intérêt journalier sur le deuxième graphique, "Words history".
- J On cherche à savoir quels sont les topics où lesquels l'utilisateur a montré le plus d'intérêt afin de les afficher sur le graphe.
Voici ce que l'on effectue sur chaque page : Pour chaque topic où sa valeur selon le modèle LDA sur cette page est au-dessus de 0.1, on estime que la page parle de ce topic et on compte le temps passé à visualiser la page dans la valeur de ce topic pour la journée.
Finalement, on somme le temps passé sur chaque "topic". Le résultat de ce calcul sur toutes les pages va nous donner l'"intérêt" final de l'utilisateur pour un topic particulier. On ne gardera ici que les 8 topics avec le plus grand intérêt estimé.
- K Finalement, pour chacun des 8 topics retenus, on affiche leur intérêt journalier (qui est la même somme que précédemment, mais agrégée par jour au lieu de toute la période) sur le premier graphique, "Topics history".

4.4.3 Implémentation

Serveur

La somme du temps de visualisation effectuée en temps réel sur le serveur est faite par une commande MySQL qui s'occupe également d'"arrondir" chaque date de visualisation d'une page à la journée (au lieu de la seconde près, qui est la granularité utilisée dans la base de données).

Interface

Les données sont finalement transformées dans un format compatible et passées à une instance configurée de la librairie HighCharts, qui génère la visualisation du graphique sur la page.

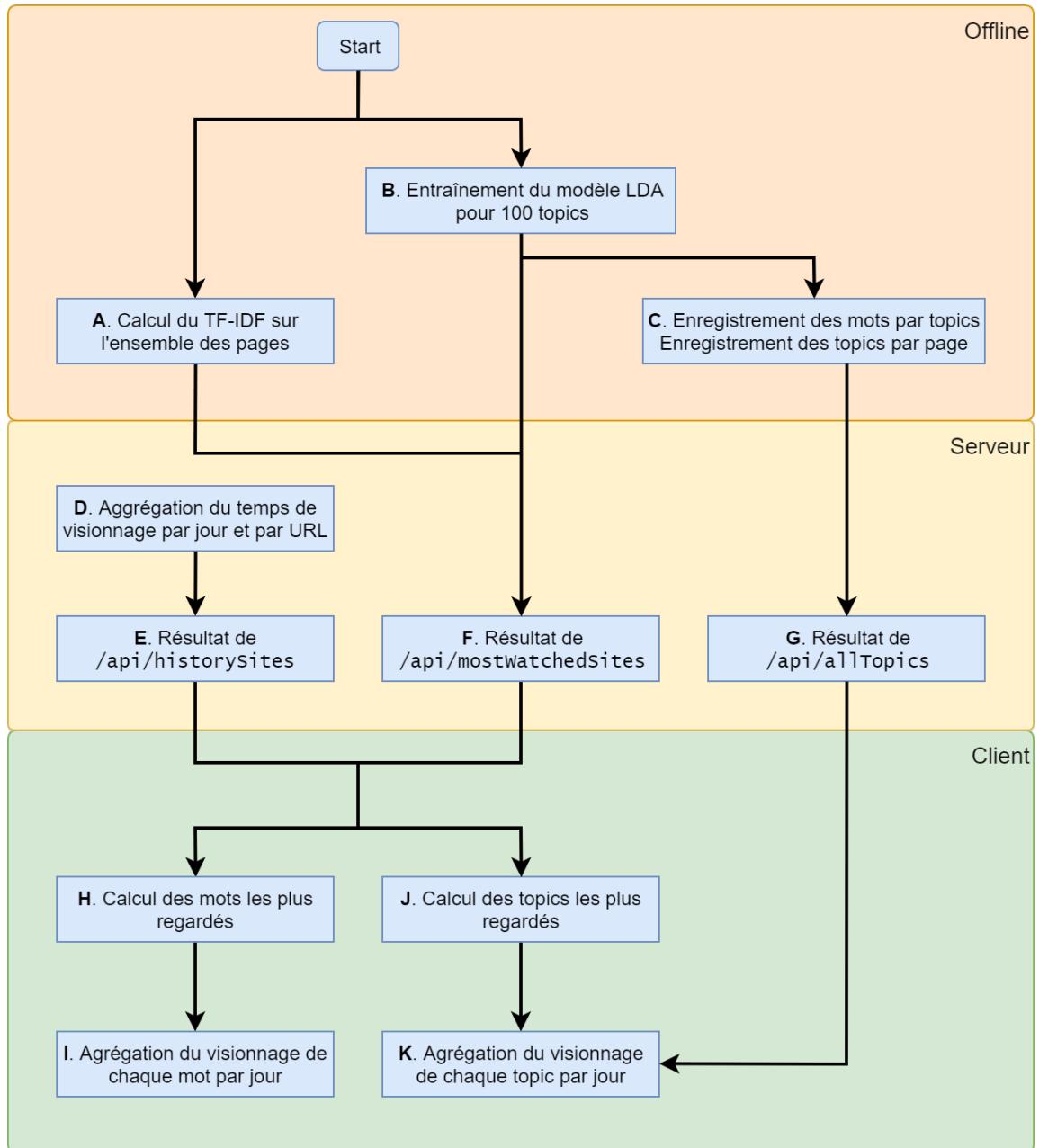


FIGURE 4.8 – Algorithme utilisé pour les graphiques de la page "History"

4.5 Trackers

4.5.1 Concept

La vue "Trackers" comporte deux pages liées à une seule source de données. Le but est de montrer à l'utilisateur que lorsqu'une page web est visitée, des informations peuvent tout de même être transmises à d'autres domaines.

La première vue de l'interface, "Most Sending", montre quels sont les domaines qui contactent beaucoup des domaines différents. Il peut ainsi voir sur cette vue quels sont les serveurs qui sont contactés lorsqu'il accède à une page web.

L'inverse est possible également. Grâce à la deuxième vue "Most receiving" il est possible de découvrir quels sont les domaines - trackers potentiels - qui sont fréquemment contactés par d'autres pages web. Chaque vue donne donc un point de vue différent sur le flux des données lorsque l'utilisateur parcourt le web.

De plus, ces deux vues peuvent interagir : Il est possible de décider de cacher certains domaines de l'une ou de l'autre vue, car par exemple l'utilisateur souhaiterait ne pas prendre en compte les données d'un certain site web, ou à l'inverse ignorer les données envoyées vers un potentiel tracker particulier.

Un bouton permettant d'activer ou de désactiver les données du domaine est présent à chaque ligne, et la désactivation de celui-ci impacte la vue des données de l'ensemble des deux pages "Trackers". La figure 4.9 montre un bouton de domaine activé et désactivé.

Ainsi il est par exemple possible de désactiver les données émises par un domaine, et de regarder quelle est la répercussion sur les données reçues par les autres.

La figure 4.10 montre la différence entre la vue imaginée et le résultat final d'une des deux pages Trackers. La figure 4.11 montre la différence entre la vue imaginée et le résultat final de la page détaillée d'un tracker.

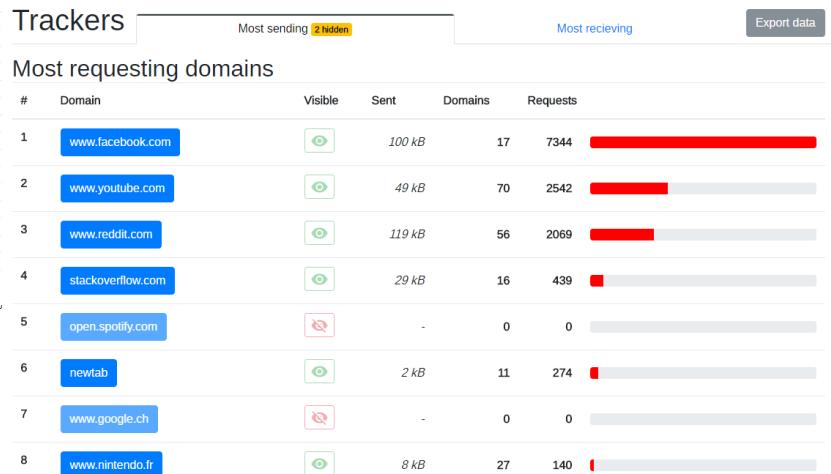
De plus, il est possible sur chacune des deux pages d'accéder aux statistiques détaillées sur le trafic de données d'un domaine particulier. En cliquant sur un domaine, l'interface ouvre une troisième vue qui montre en détail le nombre de requêtes liées à un domaine particulier.



FIGURE 4.9 –
Domaine activé,
puis désactivé

Most contacted trackers	
domain	amount
google.com	2527
doubleclick.net	1998
www.com	62
www.com	14
www.com	13

(a) Maquette initiale de la vue



(b) Exemple de résultat final

FIGURE 4.10 – Maquette initiale et résultat final d'une des vues Trackers

Tracker google.com					
Destination:	# requests	# req/pageview	Total size	Avg size/request	
heia-fr.en	910	2	18 Mo	50 Ko	
youtube.com	874	19	3 Mo	200 ko	
www.ch	310	4	9 Mo	60 ko	
www.com	258	7	17 Ko	1 Mo	

(a) Maquette initiale de la vue



(b) Exemple de résultat final

FIGURE 4.11 – Maquette initiale et résultat final de la vue détaillée lors d'un clic sur un Tracker

4.5.2 Données

Sources

Une seule source de données est nécessaire à constituer cette visualisation :

Requêtes pré-calculées : Nombre de requêtes entre chaque domaine. Ces données sont stockées dans la table `precalc_trackers` (figure 3.7k).

Algorithmme

Peu de traitements entrent en jeu dans la génération de la page Trackers. Il s'agit principalement de calculer la somme d'une liste de domaines. Ceci est illustré par la figure 4.12 :

- A On additionne le nombre de requêtes faite pour un domaine vers un autre domaine, et on enregistre le total pour chaque paire par utilisateur.
- B L'endpoint `/api/getTrackers` sert l'ensemble des résultats enregistrés pour l'utilisateur.
- C On cherche ici les domaines ayant reçu le plus de requêtes. Nous allons donc effectuer regrouper les requêtes faites par leur nom de domaine de destination, et effectuer la somme des autres mesures.
- D Inversément à l'étape précédente, on cherche cette fois les domaines ayant envoyé le plus de requêtes. Nous allons regrouper les requêtes par leur domaine d'envoie, et effectuer la somme des autres mesures.
- E Les deux listes obtenues aux étapes précédentes sont alors affichées dans leur page correspondante.
- F L'utilisateur peut choisir de cacher ou d'afficher un certain domaine d'une des vues. Ceci lance alors un nouveau calcul à partir de l'étape C. Aucune communication avec le serveur n'est nécessaire : Les données reçues initialement sont préservées.
- G L'utilisateur peut également choisir d'afficher les requêtes d'un domaine particulier.
- H Dans le cas de la sélection d'un domaine à afficher en détail, la liste des requêtes est filtrée et l'interface n'affiche que les requêtes concernant le domaine souhaité.

4.5.3 Implémentation

Serveur

Toutes les données nécessaires à cette vue sont pré-calculées avant le lancement du serveur. Il s'agit de calculer le nombre de requêtes de chaque domaine vers

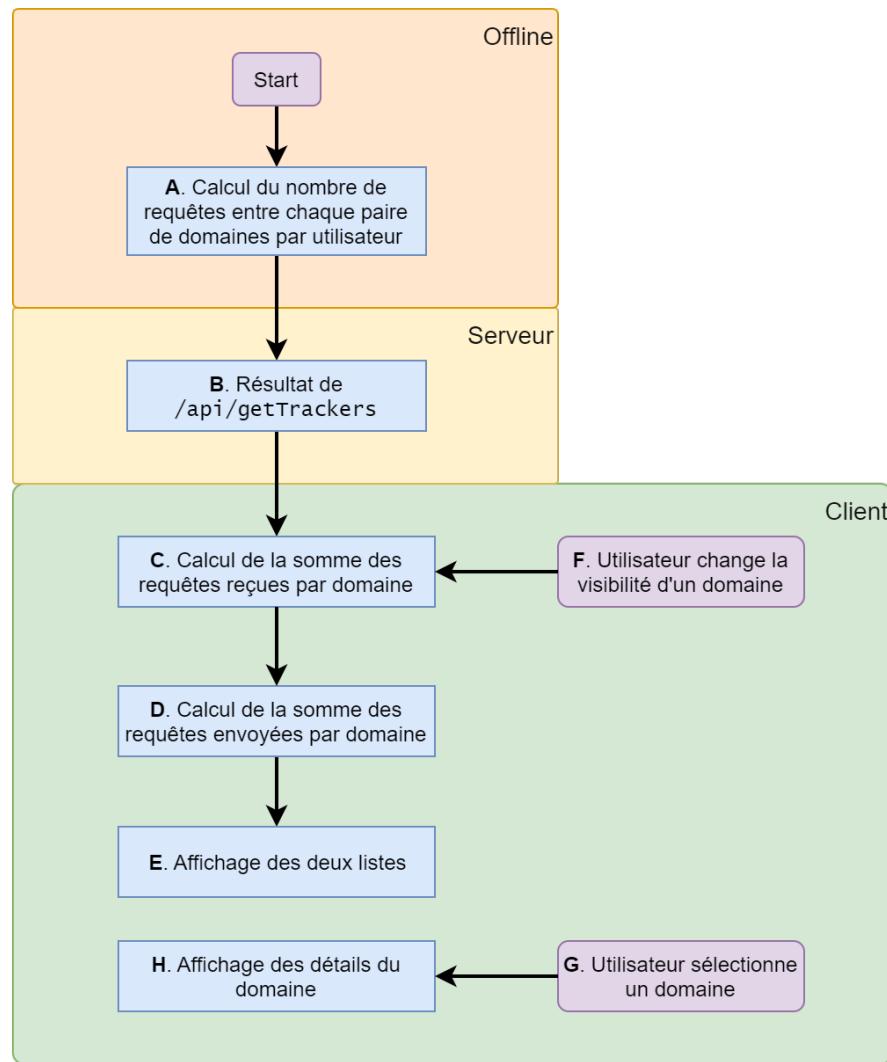


FIGURE 4.12 – Algorithme utilisé pour les données des pages "Trackers"

chaque autre domaine pour chaque utilisateur. La nécessité de pré-calculer ces données vient de leur quantité brute. Les compter sur le moment pour chaque requête demande trop de temps.

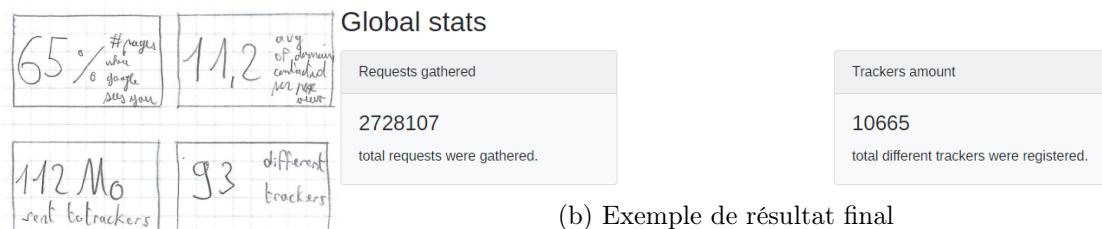
Interface

Les deux listes sont des tableaux HTML stylisés par Bootstrap. La gestion de leur interaction et de leur affichage est gérée par plusieurs composants Vue imbriqués.

4.6 Stats

4.6.1 Concept

Le but de la vue Stats est de résumer très rapidement en quelques nombres la quantité de données échangées entre les différents domaines visités par l'utilisateur.



(a) Maquette initiale de la vue

(b) Exemple de résultat final

FIGURE 4.13 – Maquette initiale et résultat final d'une vue Stats

4.6.2 Données

Sources

Une seule source de données est nécessaire à constituer cette visualisation :

Requêtes pré-calculées : Nombre de requêtes entre chaque domaine. Ces données sont stockées dans la table `precalc_trackers` (figure 3.7k).

Algorithmme

Très peu de traitements sont nécessaires pour cette vue. La figure 4.14 montre le traitement effectué aux données avant de les afficher.

4.6.3 Implémentation

Serveur

Les données nécessaires à cette vue sont calculées en temps réel : Il s'agit simplement de compter le nombre de requêtes enregistrées dans la table pré-calculée, ainsi que le nombre unique de nom de domaines.

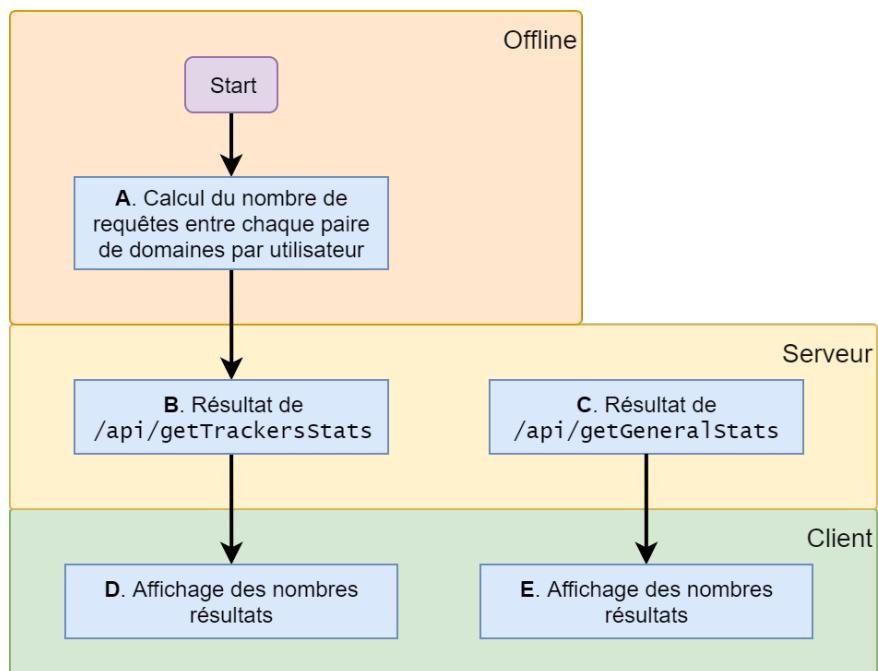


FIGURE 4.14 – Algorithme utilisé les données de la page "Stats"

Interface

Les nombres renvoyées par le serveur sont simplement affichés dans des **cards** de Bootstrap.

Chapitre 5

Résultats

5.1 Test utilisateurs

Une fois toutes les parties du projet réalisées, à savoir l'extension navigateur, le serveur et l'interface, nous avons cherché des utilisateurs volontaires pour installer l'extension et l'utiliser pendant une période de 4 semaines. Bien qu'initialement prévue pour un grand panel d'utilisateurs, les restrictions temporelles ont limité la quantité d'utilisateurs que nous avons pu atteindre.

Un total de 10 utilisateurs volontaires ont installé l'extension. Parmi eux, 8 ont été identifiés comme ayant une activité de navigation sur Chrome assez grande pour contribuer à l'étude. (Ceux étant jugés inactifs totalisent moins de 10 minutes d'activité).

5.1.1 Inputs

En plus de récolter les données des utilisateurs et de les afficher, il leur a également été demandé de remplir quelques informations sur eux-même afin de pouvoir valider certains points de notre étude. Deux formulaires ont été mis en place à cette fin.

Centres d'intérêts

La figure 5.1 montre un premier formulaire à remplir par l'utilisateur. Accessible via le lien vers la page "Settings", on lui demande ici de trouver et renseigner quelques uns de ces centres d'intérêt parmi une centaine.

Un maximum de 10 centres d'intérêts peuvent être définis. Le but de laisser à l'utilisateur entrer des centres d'intérêts est de pouvoir nous rendre compte si les topics que nous lui proposons sont proches de ces centres d'intérêt.

Settings

Interest fields



FIGURE 5.1 – Champ d’entrée des centres d’intérêts

Association topic et intérêt

Une fois que l’utilisateur a défini des centres d’intérêt, il peut donner des informations sur les topics que nous lui suggérons sur la page Topics List (voir section 4.2). La figure 5.2 montre la fenêtre déroulante de sélection d’un centre d’intérêt pour un topic donné.

#	Words	Related interest
1	comment reddit post	<div style="border: 1px solid #ccc; padding: 5px;"> <p>Social Media Enthusiast ▾</p> <p>Foodie Nightlife Enthusiast Gamer Hardcore Gamer Roleplaying Game Fan Electronica & Dance Music Fan Social Media Enthusiast Technophile</p> </div>

FIGURE 5.2 – Sélection d’intérêt sur un topic

Il est demandé aux utilisateurs d’ajouter une association entre un topic et un intérêt lorsque cela semble lui faire sens. Ainsi, nous pouvons savoir quels topics de l’utilisateur nous avons réussi à identifier. En effet, lorsqu’un utilisateur associe un centre d’intérêt à un topic, cela signifie non seulement que le topic trouvé a du sens en soi, mais en plus qu’il est intéressant pour l’utilisateur.

Sur l’échantillon de 8 personnes actives, 6 personnes ont ajouté des associations aux 20 topics proposés sur leur page.

5.2 Résultats de la recherche

Après une récolte des données sur environ 4 semaines, nous pouvons nous intéresser aux résultats que nous avons récoltés.

5.2.1 Modèles

Une première étape est de se pencher sur les modèles que nous avons générés sur la base des données elles-mêmes. Nous utilisons principalement deux algorithmes qui se basent sur le contenu des pages pour en déterminer leur thème : TF-IDF, et LDA.

Ces modèles se basent sur le contenu des pages visitées des utilisateurs. Etant donné que nous ne récupérons pas directement le contenu de la page visitée par un utilisateur mais uniquement une partie de son URL, nous ne pouvons pas garantir que le contenu que nous récupérons d'une page soit effectivement le contenu que l'utilisateur voit sur son écran. En effet, beaucoup de pages web aujourd'hui sont liées à une application qui demande une authentification de l'utilisateur pour être affichée.

Par exemple, une grande partie des pages d'un réseau social peuvent nécessiter la connexion de l'utilisateur pour être affichée.

TF-IDF

Fonctionnement TF-IDF est une méthode attribuant un poids à chaque mot de chaque document d'un corpus. Ce poids mesure l'importance relative du mot dans ce document. Le nom "TF-IDF" signifie "Term Frequency - Inverse Document Frequency". Le poids final d'un mot dans un document se calcule en prenant en compte uniquement deux mesures :

TF Term Frequency : La quantité d'apparition de ce mot dans ce document

DF Document Frequency : Le nombre de documents dans lesquels ce mot apparaît

Le score final d'un mot multiplie le TF d'un mot à l'inverse de son DF. Celà signifie que pour avoir une grande importance dans un document, un mot sera typiquement :

- Présent de nombreuses fois dans ce document
- Présent dans très peu d'autres documents

Le calcul des poids se fait sur l'ensemble du corpus, en une fois, car il nécessite que l'on connaisse le nombre d'occurrences de chaque mot dans l'ensemble du corpus de documents. Il n'est donc pas possible de mettre à jour les poids de manière "online" en utilisant ce modèle.

Résultats Juger les résultats du calcul de TF-IDF sur des documents est une tâche non triviale. Cela revient principalement à vérifier manuellement que les mots ayant le poids le plus élevé pour certaines pages soit significatif de leur sujet.

Intéressons-nous donc aux mots ayant le plus de poids trouvés pour les 20 pages les plus regardées, par exemple. Le tableau 5.3 illustre les 20 pages les plus regardées avec leurs mots associés, ainsi que plusieurs étapes amenant à une estimation finale de l'adéquation des mots trouvés avec le contenu de la page. Voici comment se lit le tableau :

URL URL de la page concernée. Certaines URLs trop longues ont été raccourcies ici

Mot 1, 2, 3 3 meilleurs mots dans l'ordre décroissant décrivant la page selon TF-IDF.

Pub(lique) Est-ce que la page nécessite une connexion afin d'accéder à son contenu principal.

Con(tenu) Est-ce que le principal contenu de la page est textuel ?

Mot Est-ce que chacun des mots trouvés sur la page fait sens dans une langue connue ?

Adé(quat) Est-ce que l'ensemble des mots trouvés forme un potentiel résumé adéquat du contenu de la page ?

Les 4 premières colonnes (URL et 3 mots) proviennent de la base de données, tandis que les 4 dernières colonnes sont le résultat d'une évaluation manuelle des critères décrits. Un "OUI" dans une colonne indique que la page a passé le critère défini, contrairement à un "NON". Un "NON" dans une colonne entraîne automatiquement un "NON" dans les colonnes situées les plus à droite.

Réflexion Le tableau 5.1 montre un résumé des résultats que l'on peut récupérer précédent tableau. On remarque que sur les 20 URLs entrées, seules 7 valent vraiment la peine d'être parcourues par notre algorithme, par élimination à causes des deux premières raisons énoncées. Cependant, sur les 7 URLs contenant du texte intéressant, TF-IDF a été capable d'en résumer adéquatement 5 d'entre-elles.

Ce résultat est loin d'être parfait, mais il montre tout de même qu'il est possible d'automatiser la recherche de mots importants sur des pages lorsque les conditions sont favorables à notre approche.

FIGURE 5.3 – 20 URLs les plus regardées et leurs meilleurs mots selon TF-IDF

URL	Mot 1	Mot 2	Mot 3	Pub	Con	Mot	Adé
http://wdwf.sdipi.ch/	footprints	digital	web	OUI	OUI	OUI	OUI
https://www.draw.io/	gmdl	eng	proc	OUI	NON	NON	NON
https://www.reddit.com/r/videos/	submit	load	report	OUI	OUI	NON	NON
https://www.google.co.uk/search	eingabetaste	suehe	drücke	OUI	NON	NON	NON
https://www.google.ch/search	eingabetaste	suehe	drücke	OUI	NON	NON	NON
http://game110.idlekiller.com/	explorer	chrome	browser	OUI	OUI	OUI	OUI
http://df.sdipi.ch/phpmyadmin/sql.php	phpmyadmin	past	welcome	NON	NON	NON	NON
https://web.whatsapp.com/	whatsapp	macos	mozilla	NON	NON	NON	NON
http://hexaclicker.github.io/	hexa	dp	level	OUI	OUI	OUI	OUI
http://blankmediagames.com/TownOfSalem/	salem	adobe	town	OUI	NON	NON	NON
http://www.jeuxvideo.com/	jeu	annonce	bande	OUI	OUI	OUI	OUI
https://discordapp.com/channels/217...408/217...408	own	respective	owner	NON	NON	NON	NON
https://www.reddit.com/r/leagueoflegends/	leagueoflegends	submit	self	OUI	OUI	OUI	OUI
https://www.reddit.com/	bot	agent	partner	OUI	OUI	OUI	OUI
https://www.google.fr/search	eingabetaste	suehe	drücke	OUI	NON	NON	NON
https://s3-fr.gladiatus.gameforge.com/game/index.php	de	gameforge	vous	NON	NON	NON	NON
https://docs.google.com/presentation/d/1IB...l5w/edit	row5w	gech...el5w	slide	NON	NON	NON	NON
https://twitter.com/	tweet	foto	hast	OUI	NON	NON	NON
http://df.sdipi.ch/phpmyadmin/db_structure.php	phpmyadmin	past	welcome	NON	NON	NON	NON

TABLE 5.1 – Résumé des résultats de TF-IDF

URLs initiales	20
Pages publiques	14
Contenu textuel principal	7
Mots sensés	7
Mots adéquats	5

LDA

Fonctionnement Le topic modelling consiste, à partir d'un corpus de documents, à générer une liste probable de sujets ou topics communs à plusieurs documents. LDA (de l'anglais Latent Dirichelet Allocation) est un modèle de topic modelling, et va nous permettre de révéler des topics relatifs aux pages visitées par les utilisateurs. Ici, un topic est défini par une liste de mots, et un poids associé à chaque mot pour le topic.

La génération d'un modèle LDA prend plusieurs paramètres en entrée, mais le plus important pour nous est de définir un nombre de topics que nous souhaitons voir en sortie. On fixe ce nombre de topics, puis on lance l'apprentissage du modèle sur l'ensemble du corpus de documents, opération que peut durer plusieurs heures.

À la fin de l'apprentissage, nous sommes en possession d'un modèle que nous pouvons questionner de plusieurs manières, par exemple :

- Quels sont les mots les plus contribuants à un topic ?
- Quels sont les topics les plus probables pour un document ?

Nous allons donc par exemple utiliser le modèle afin d'assigner des topics au contenu d'URLs, et ainsi tenter de trouver quels sont les thèmes communs aux pages visitées par un utilisateur.

Résultats Dans la première semaine de récolte des résultats, une recherche empirique sur les paramètres à fournir au modèle a été effectuée. Le paramètre du nombre de topics a été fixé à 100 ; cela semblait un bon compromis, car 50 générerait un nombre de trop restreint pour définir de manière assez précise quels étaient les thèmes d'un utilisateur, et 200 générerait beaucoup de topics qui n'avaient pas de sens en eux-mêmes.

La figure montre les 20 topics les plus

Réflexion

5.2.2 Vues

Wordcloud

Topics List

Most watched and viewed

History

Trackers

5.2.3 Statistiques

Profiling

Trackers

5.2.4 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=DYhAM34Hzc>, 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://mypersonnality.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/ejifodhjoeenihgfpjijjmpomaphmah>, 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

Document Object Model est l'arbre d'éléments qui compose un fichier HTML.
34

open-source qualifie un logiciel dont le code initial est mis à disposition du grand public.. 9

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

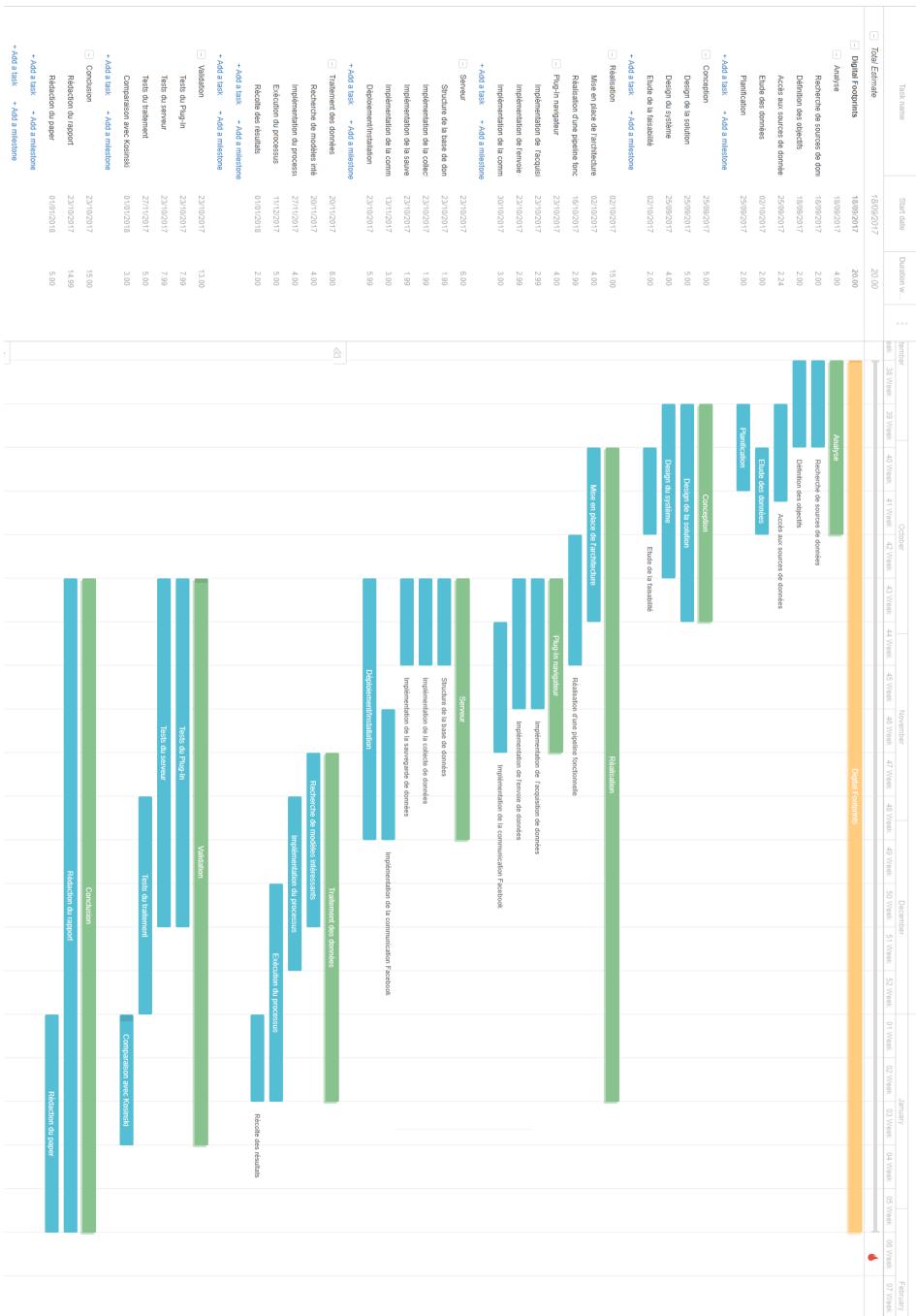
B.2 Planification

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.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



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.

ANNEXE D. PROCÈS-VERBAUX

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 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.

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 myPersonality**

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.

ANNEXE D. PROCÈS-VERBAUX

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 événements. 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.

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 myPersonnalité**

Le template de l'e-mail à envoyer à myPersonnalité 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.

ANNEXE D. PROCÈS-VERBAUX

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

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.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



MASTER OF SCIENCE
IN ENGINEERING

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.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

1 novembre 2017, de 14h05 à 14h55

Présent : Félicien 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.

ANNEXE D. PROCÈS-VERBAUX

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



PV de réunion

9 novembre 2017, de 14h05 à 14h55

Présent : Félichen 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.

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.

ANNEXE D. PROCÈS-VERBAUX

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

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



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.

ANNEXE D. PROCÈS-VERBAUX

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.

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.

ANNEXE D. PROCÈS-VERBAUX

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.

PV de réunion Digital Footprints

Travail de Master : Digital Footprints



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.

- Graphe « 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.

ANNEXE D. PROCÈS-VERBAUX

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.

ANNEXE D. PROCÈS-VERBAUX

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.