



HAL
open science

Clustering pour le e-Marketing et analyse d'un questionnaire

Nétanel Ghérabli

► **To cite this version:**

Nétanel Ghérabli. Clustering pour le e-Marketing et analyse d'un questionnaire. Méthodologie [stat.ME]. 2014. dumas-01059635

HAL Id: dumas-01059635

<https://dumas.ccsd.cnrs.fr/dumas-01059635>

Submitted on 1 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UFR Mathématique Informatique

7 rue René Descartes
67084 Strasbourg Cedex

03 68 85 50 00

Stagiaire chargé d'études Statistiques

wepingo

80 rue des Haies

75020 Paris

Etudiant : Nétanel Ghérabli

Maître de Stage : Uriel Berdugo

Responsable pédagogique : Armelle Guillou

Master 1 mention Mathématiques et Applications, Spécialité
Statistique, parcours Bio-Statistiques et Statistiques Industrielles

Sommaire

Introduction -----	p.4
I) Contexte -----	p.5
1) Présentation de l'entreprise -----	p.5
2) Positionnement sur le marché-----	p.7
II) Réalisation du projet -----	p.8
1) Quelques problématiques liées à cet outil-----	p.8
2) Qu'est-ce qu'une règle marketing ?-----	p.9
3) Présentation de l'outil -----	p.10
4) Schéma représentatif du processus complet-----	p.10
5) Ce que verra le marketeur-----	p.12
III) L'extraction des sensibilités à partir des données facebook-----	p.31
1) La slide pour COTY-----	p.31
2) Mise en application de la méthode-----	p.32
IV) Implémentation d'une méthode de clustering de type K-means-----	p.33
1) L'algorithme en R-----	p.33
2) Qu'est-ce que le clustering ?-----	p.34
3) Les étapes de mon algorithme-----	p.34
4) Quelques idées pour améliorer cet algorithme-----	p.36
V) Analyse Statistique du contest Facebook-----	p.38
1) Quelques statistiques générales-----	p.41
2) Statistiques sur l'âge-----	p.43
3) Statistique sur les affinités-----	p.45
Conclusion-----	p.46
Programmes utilisés-----	p.46
Lexique-----	p.47
Bibliographie-----	p.47

Remerciements

Je tenais à remercier toute l'équipe de Wepingo qui se fonde sur un dynamisme d'exception ainsi qu'une motivation à toute épreuve. Leur accueil, leur bonne humeur ainsi que leur disponibilité ont permis de former un groupe uni dans lequel il a été agréable de travailler.

Je remercie principalement mon maître de stage, Uriel BERDUGO, qui m'a beaucoup appris et grâce à qui mes connaissances se sont développées tout au long de ma mission, autant en terme d'outils statistiques que de méthodologie, qui m'a accordé de son temps malgré son planning chargé.

Je le remercie aussi pour m'avoir permis d'accéder à ce stage de grande valeur. Cette expérience a été très importante pour mon orientation future, et les tâches auxquelles il m'a associé m'ont vraiment permis de consolider mes connaissances et d'en développer de nouvelles.

Je remercie le responsable du développement commercial, Samuel SEYMAN, pour m'avoir fait confiance et avoir mis en avant mon projet de stage lors de ses rendez-vous professionnels, ainsi que pour la bonne humeur communicative qui lui est propre.

Je remercie également l'ingénieur recherche et développement, Mihai FLOREA, pour avoir partagé avec moi ses connaissances sur son domaine.

Introduction

Dans le cadre de ma formation à l'UFR Mathématique - Informatique (1^{ère} année de Master mention Mathématiques et Applications, Spécialité Statistique, parcours Bio-Statistiques et Statistiques Industrielles), j'ai effectué mon stage chez Wepingo (Start-up), ayant pour but d'assister personnellement à l'achat sur Internet.

Etant dans une start-up, il m'a donc fallu être rapidement efficace et autonome lors de la réalisation du projet dans un environnement en constante évolution, mais aussi que je m'intègre dans une équipe mélangeant des domaines comme la recherche, le marketing et l'ingénierie informatique (Big Data).

Ma mission a donc été dans un premier temps de rédiger les spécifications fonctionnelles d'un **outil d'injection de règles marketing**, de sa modélisation, de m'assurer de sa faisabilité et d'y apporter des idées innovantes basées sur mes connaissances. Dans un deuxième temps j'ai implémenté concrètement sous R une partie de l'outil et j'en ai profité pour faire une analyse statistique sur les questionnaires qu'utilise l'entreprise.

Dans une première partie, nous allons aborder le concept de Wepingo avant de s'intéresser au projet qui m'a été confié.

I) Contexte

Amis d'enfance, les 3 fondateurs ont travaillé à la création d'une entreprise dont l'activité serait liée au e-commerce avec une dominante technologique et innovante très forte. L'idée était de pénétrer le marché des comparateurs qui se caractérise par un manque d'innovations majeures depuis 10 ans.

Les offres marchandes pléthoriques, plus complexes et souvent mal classifiées rendent le processus de choix d'un consommateur long et difficile. Le nombre de sites marchands a dépassé les 100,000 depuis 2012, pour autant encore 80% des consommateurs privilégient l'achat en magasin notamment pour la dimension de « conseil » et « d'expertise » qu'elle apporte. Cependant, les ventes en magasins continuent de s'éroder, obligeant les géants de la distribution à diminuer leurs effectifs et à réadapter leur organisation commerciale à tous les niveaux, y compris technologiques.

1) Présentation de l'entreprise

Wepingo est une start-up récente. Elle repose sur un concept unique et original. On sait que le processus de choix d'un consommateur est long, il est supérieur à 70 jours pour les biens représentant plus de 400€, on peut facilement jeter la pierre à la surinformation qui peut effrayer et même parfois faire se rétracter un consommateur. La plateforme Wepingo répond à ce besoin en offrant une solution personnalisée et claire à l'utilisateur, afin de faciliter son choix.

A l'aide d'un processus simple de questionnement, une affiliation à une gamme de produit est déterminée par une IA (intelligence artificielle), Wepi. La pertinence de chaque question est expliquée à l'utilisateur, afin de le faire participer activement au processus d'affiliation mais

aussi qu'il comprenne bien l'impact de sa réponse. Nous avons ainsi répondu à son besoin et avons été à même de lui proposer une gamme de produit.

La solution conçue et développée par Wepingo répond aux problématiques simultanées du consommateur et du distributeur. La dimension de conseil est formulée au travers de la plateforme web (ou l'application), la vente se fait soit sur internet soit en magasin permettant aux enseignes de retrouver du trafic physique, de gagner de nouveaux clients, de vendre mieux des produits potentiellement plus margés.

Wepingo est la toute première plateforme capable de comprendre le besoin d'un utilisateur afin de lui recommander la gamme de produit la plus adaptée à son besoin et à son profil d'utilisation, grâce à un calcul d'affinité appelé « affinity bot ». Quelque soit la complexité technique et l'univers marchand, les technologies en Intelligence Artificielle de Wepingo sont en mesure de conseiller au moins aussi bien qu'un expert humain.

La technologie innovante de conseil personnalisé au consommateur est utilisée au travers de son comparateur Wepingo.com. Wepingo rend intelligent les catalogues au travers de l'organisation et la normalisation des données. La formalisation du savoir à l'échelle industrielle est rendue possible par la dominante innovante, au cœur de la stratégie de développement.

Wepingo est un assistant shopping personnel qui offre des solutions intelligentes de shopping, personnalisées pour les consommateurs au travers du web (wepingo.com) ou de magasins physiques grâce aux solutions pour bornes en cours de commercialisation. Dans un premier temps, Wepingo isole le besoin d'un utilisateur et calcule des taux d'affinité pour l'utilisateur sur des gammes de produits d'un univers considéré. Wepingo guide ainsi l'utilisateur aussi bien vers des web marchands que vers des magasins physiques disposant du produit recherché autour de chez lui.

Dans ce contexte, Wepingo se positionne sur le marché porteur du pré-shopping. Qu'est-ce que le pré-shopping ? C'est une tendance récente qui désigne une activité de sélection, de conseil et d'aide à l'achat au consommateur. Le conseil revêt plusieurs formes dans différentes activités. Cela peut revêtir la forme de comparateur de prix (comme Kelkoo) en passant par l'analyse des besoins du consommateur en lui proposant une solution personnalisée (Modizy) ou bien enfin aider le consommateur à trouver le produit en stock près de chez lui (Socloz). Rares sont les assistants pré-shopping capables de réunir l'ensemble de

ces services. Aujourd'hui seule la plateforme Wepingo le propose. Wepingo a une forte dominante d'intelligence artificielle capable de comprendre le consommateur pour l'orienter vers une affinité la plus fine envers une gamme de produits.

2) Positionnement sur le marché

Wepingo se situe sur un marché proche de celui des comparateurs. Cependant, le secteur réel sur lequel Wepingo souhaite s'installer durablement est celui des assistants shopping ou le marché du pré shopping.

Le e-commerce s'étant diversifié, la clientèle potentielle est représentée par toute personne désirant acheter sur le web. Néanmoins, la dimension du conseil s'adresse à des personnes sûres de passer à l'acte d'achat mais ne sachant pas quel produit choisir.

La plateforme a pour but d'orienter l'utilisateur sur une catégorie de produits qui correspond à son besoin, au sein d'une gamme généralement très vaste. Parmi tous les univers qui seront proposés, tous les catalogues des distributeurs sont indexés dans la base de données, afin de réunir un maximum de produits et d'orienter l'utilisateur vers la gamme qui lui correspond le mieux. Dans l'idéal, toutes les gammes et produits proposés dans les catalogues des distributeurs existants (Darty, Cdiscount, ...) seront indexés sur le site.

Tout site web marchand commercialisant les produits proposés sur Wepingo est potentiellement un partenaire ou un concurrent. Le Web marchand est aujourd'hui dominé par trois types d'entreprises : des pure-players (comme Amazon, CDiscount, eBay, Pixmania ou encore Rue de Commerce), des véadistes, qui vendent uniquement par catalogue (tels que les 3 Suisses ou La Redoute) et les click & mortar qui regroupent un réseau de magasins et une e-boutique (comme la Fnac).

Dans les domaines du pré-shopping, de nombreux acteurs ont fait leur apparition depuis 2011. La grande majorité d'entre eux ont une activité dans le domaine de la mode.

Il existe deux concurrents qui ne sont pas liées à la mode : il s'agit de Socloz et best comparator. Socloz utilise la géolocalisation mais n'a pas de dimension de conseils experts, ni d'achats groupés, quant à Best comparator, son modèle est limité dans ses fonctions (pas de

géolocalisation, de conseils experts, d'achats groupés ni d'intelligence artificielle (mais seulement des filtres).

Wepingo mise avant tout sur la qualité et la quantité de services afin de se différencier de la concurrence.

II) Réalisation du projet

Après avoir présenté ce que fait actuellement Wepingo à l'aide de ses questionnaires, mon but était d'aider à la réalisation d'un projet permettant de faire évoluer et de diversifier, l'offre de l'entreprise.

Devoir répondre à un questionnaire lorsqu'on achète un produit sur internet peut s'avérer contraignant et nombre d'utilisateurs risquent de ne pas utiliser l'outil par manque de temps ou parce qu'ils ne connaissent pas l'entreprise Wepingo.

Il faut donc trouver une solution où l'utilisateur est conseillé sans avoir à faire « d'effort ». L'idée est donc de créer un outil permettant au marketeur d'intégrer des règles marketing qui s'appliqueront à tous les utilisateurs de son site, et qui proposera en fonction des informations disponibles sur l'utilisateur et les règles du marketeur les meilleurs produits.

1) Quelques problématiques liées à cet outil

Faire du marketing one-to-one c'est se détacher de la segmentation, alors que l'approche la plus simple et la plus intuitive serait de créer des profils issus de segmentations d'acheteurs. C'est de cela que l'on veut se détacher.

Le challenge est donc de trouver un moyen de différencier chaque utilisateur mais aussi de pouvoir les conseiller au mieux.

Il a donc fallu réfléchir à un moyen d'automatiser l'écriture des règles marketing, étant donné le très grand nombre de règles à écrire pour couvrir chaque univers. Il a aussi fallu trouver un moyen de les vérifier et de les valider.

2) Qu'est ce qu'une règle marketing ?

Exemple : **SI** n nombre d'enfants inférieur à 2 ans **ALORS** n nombre de places dans la poussette.

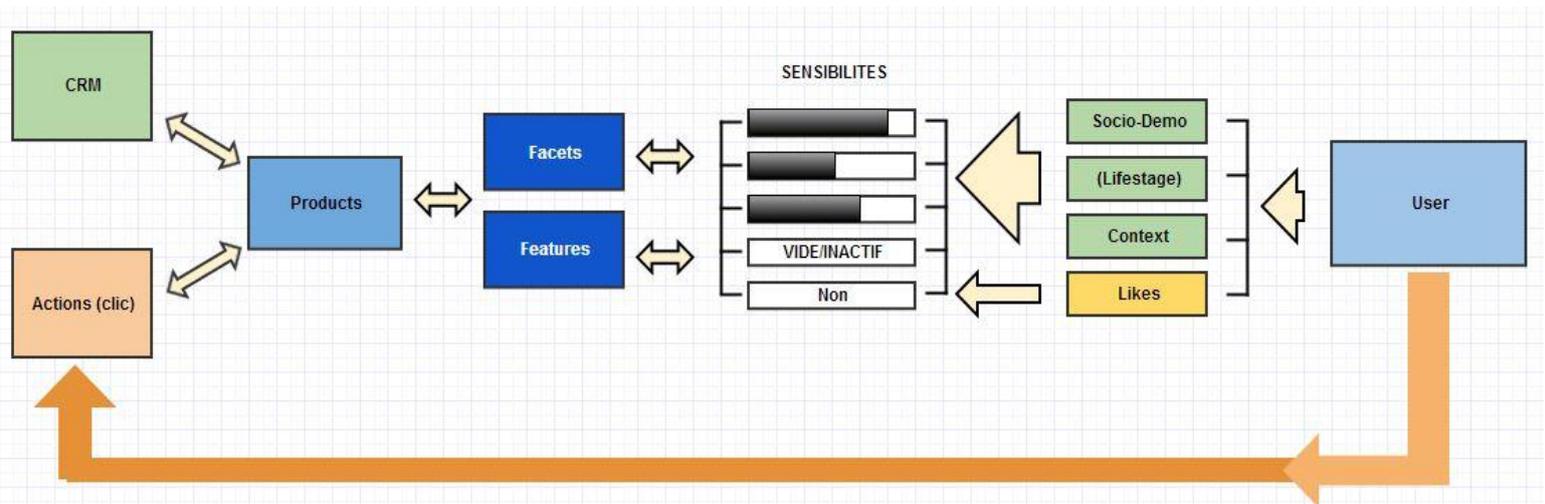
Il s'agit de traduire un ensemble d'informations sur l'utilisateur et de les traduire en sensibilités.

En effet, le contexte a évolué : dorénavant, le stockage massif d'informations sur les individus est appliqué par tous, dans une économie toujours plus mondialisée. Le marketing se transforme : avec l'émergence du Big Data, la connaissance client permet une approche One-to-One. Nous n'avons plus besoin de segmenter le marché : chaque client est unique. Nous pouvons donc lui proposer le produit adapté. L'analyse prédictive permet donc de se différencier de la concurrence et de fidéliser le client durablement. C'est donc cela que nous voulons offrir aux GSS (grandes surfaces spécialisées) comme Darty, Fnac, Sephora, et aux e-commerçants Cdiscount,...

3) Présentation de l’outil

L’outil que je devais développer était un outil à l’état d’idée, aucune structure n’avait été prédéfinie. J’ai donc dû intégralement créer l’outil, le plus pertinemment possible, avec une collègue, Sarah Dureau, se chargeant de la partie marketing et communication, afin de satisfaire les besoins de l’entreprise. J’ai donc apporté mon esprit mathématique, de synthèse, mes connaissances statistiques et mes bases de programmation pour définir quelles seraient les étapes nécessaires à la réalisation du projet et sa faisabilité. La version finale que je vais vous présenter ici est le résultat de 2 mois d’élaboration, en collaboration avec ma collègue, en mettant chacun en symbiose nos connaissances.

4) Schéma représentatif du processus complet



Les différents termes qui constituent ce schéma sont issus d’un lexique que nous avons mis en place avec mon maître de stage. Il est impératif d’utiliser des termes précis lors de l’élaboration de projet innovant comme celui-ci pour être assuré de bien se faire comprendre.

La CRM : c’est l’abréviation de Customer Relationship Management, cela représente la base de données des clients d’une entreprise. Et plus particulièrement leurs achats.

Les actions : ce sont l’ensemble des activités que peut avoir l’utilisateur sur le site. Chacun de ses clics, le temps qu’il passe sur chaque page, ce qu’il a mis dans son panier,...

Les Products : ces produits disponibles.

Les Facets : une facet est un aspect fonctionnel d'un produit, par exemple la portabilité.

Les Features : une feature est un aspect technique d'un produit, par exemple le poids.

Les Sensibilités : une sensibilité est un caractère qui pousse un utilisateur à être attiré vers certaines facets et features de certains produits. Par exemple une sensibilité « voyageur » implique une sensibilité plus grande à la portabilité. Les sensibilités sont graduées ou binaires et permettent d'établir un profil de sensibilités propre à chaque utilisateur. Elles peuvent aussi être inactives si nous n'avons pas d'information pour un individu x.

Les Socio-Demo : ce sont les informations socio démographiques que nous avons sur un utilisateur.

Les life-stage : ce sont les étapes de la vie de l'utilisateur, ces informations sont proches des Socio-Demo.

Le Context : il s'agit de l'ensemble des données relatives au moment et à l'endroit d'où se connecte l'utilisateur. Par exemple, la date, la région etc... qui peuvent donner des informations très utiles quand à ses sensibilités.

Les Likes : ce sont l'ensemble des préférences de l'utilisateur que l'on a extraites de son profil facebook.

L>User : l'utilisateur

Notre mission était de penser à un processus d'extraction des sensibilités des utilisateurs à partir des données dont l'on dispose les concernant. Pour cela, nous avons dus penser à une automatisation des règles de marketing. Par contre, toute la partie consistant à automatiser l'extraction des sensibilités des produits à partir de leur fiche technique était déjà effectuée par Wepingo.

5) Ce que verra le marketeur

J'ai représenté sur l'outil *Balsamiq* ce à quoi ressemblera l'outil.

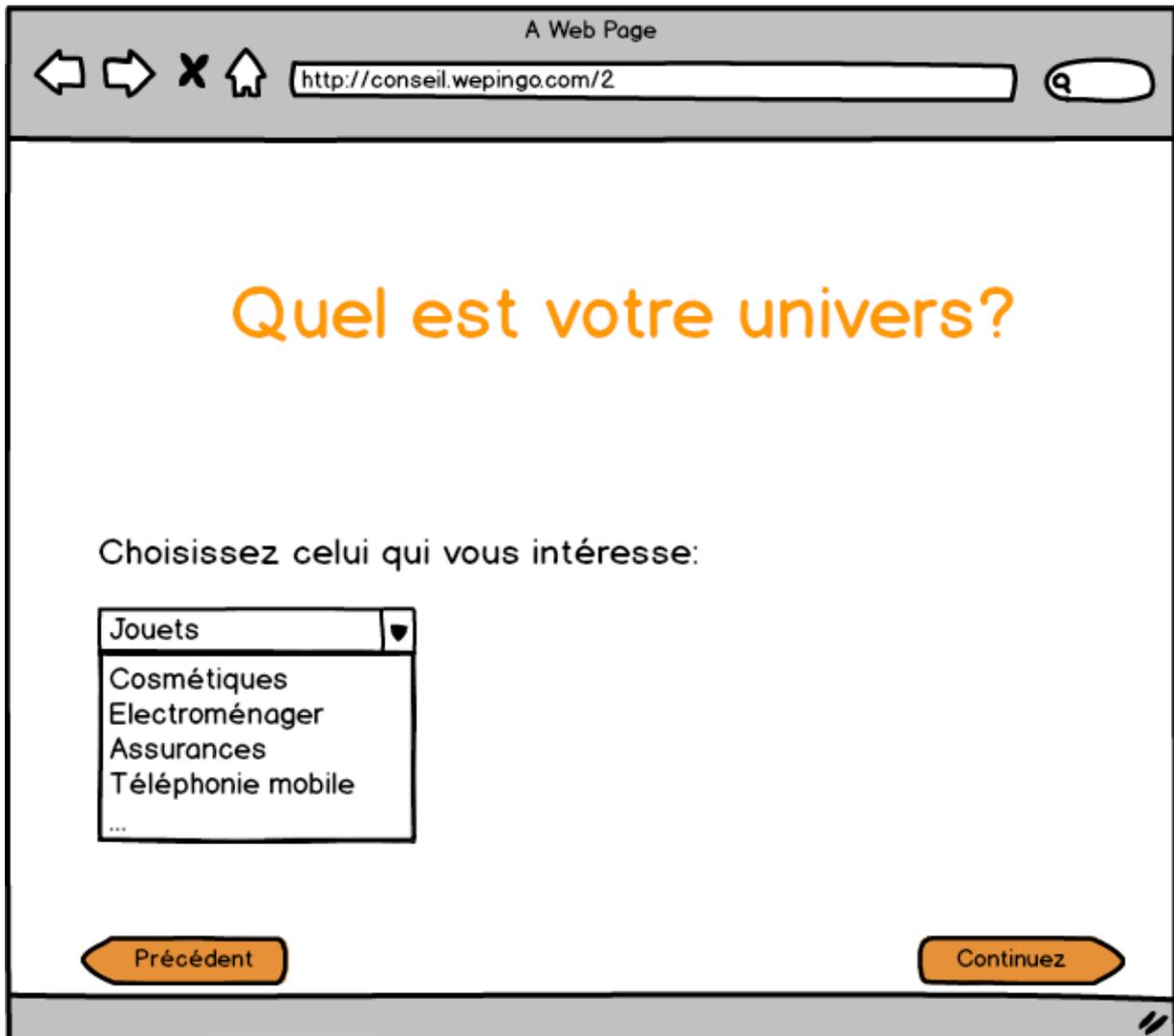
Un grand nombre de versions précèdent celle-ci. Le processus de conception était très progressif.

Voici donc les étapes accompagnées d'une description de ce que fera le marketeur.

L'écran d'accueil :



1) Le marketeur va choisir l'univers auquel il veut s'intéresser.



- 2) Il liste les sensibilités qu'il pense en rapport avec son univers. (Le marketeur ne connaissant pas ce qu'on appelle une sensibilité il pourrait être judicieux de proposer un choix parmi une liste dans un premier temps auquel se grefferaient des sensibilités personnalisées qu'il aura définies lui-même).

A Web Page

http://conseil.wepingo.com/3

Les sensibilités* que wépingo associe à l'univers choisi:

* : Définition: Une sensibilité est un caractère qui pousse un utilisateur à être attiré vers certains aspects (que l'on nommera par la suite facets/features) de certains produits. exemple: un utilisateur peut être sensible aux voyages → sensibilité "voyageur"

<input type="checkbox"/>	Voyageur	<input checked="" type="checkbox"/>
<input type="checkbox"/>	Casanier	<input checked="" type="checkbox"/>
<input type="checkbox"/>	...	<input checked="" type="checkbox"/>

Désélectionnez si vous pensez que la sensibilité n'est pas pertinente

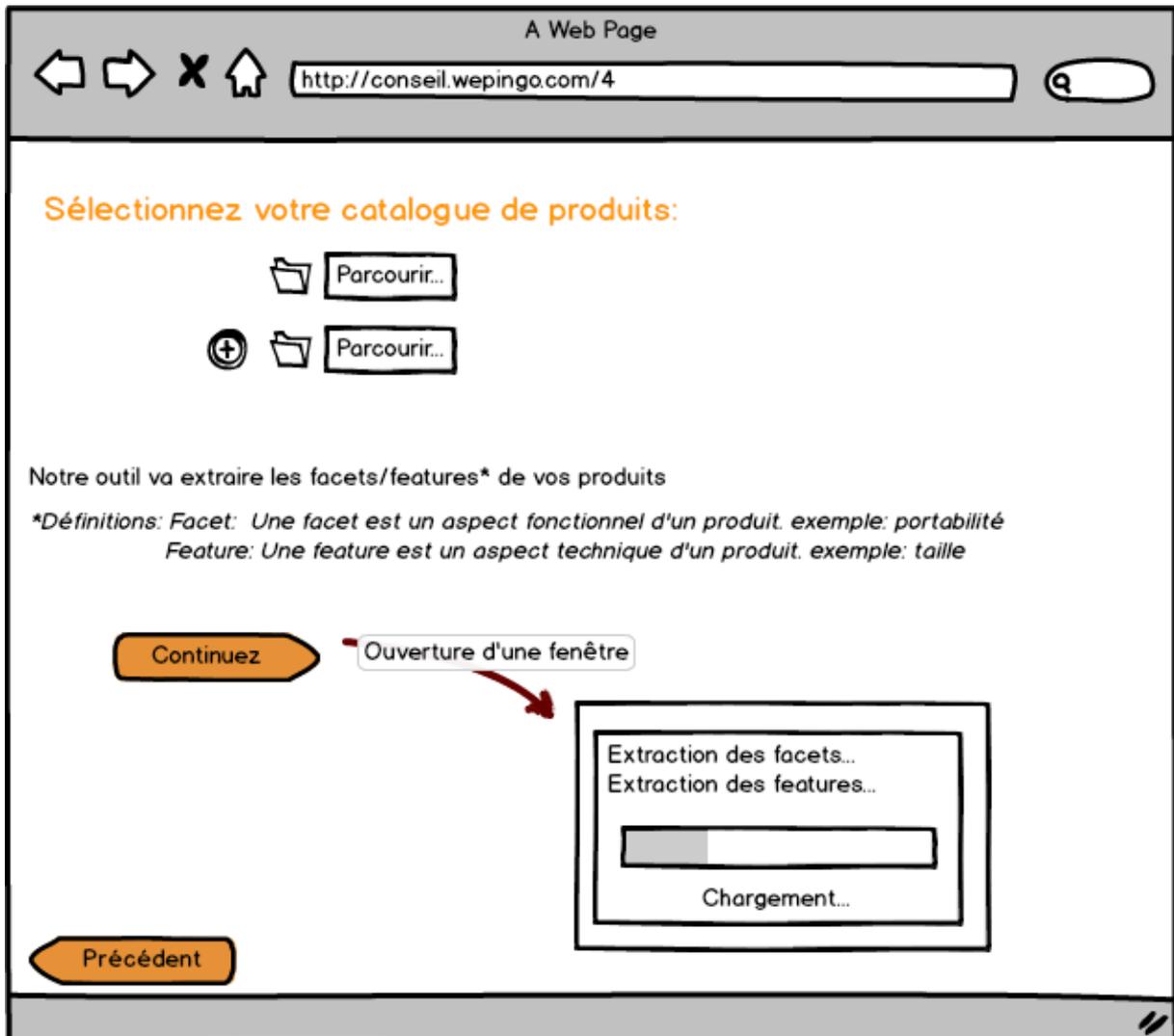
Ajoutez d'autres sensibilités:

Extravagant
Classique
Fêtard
...

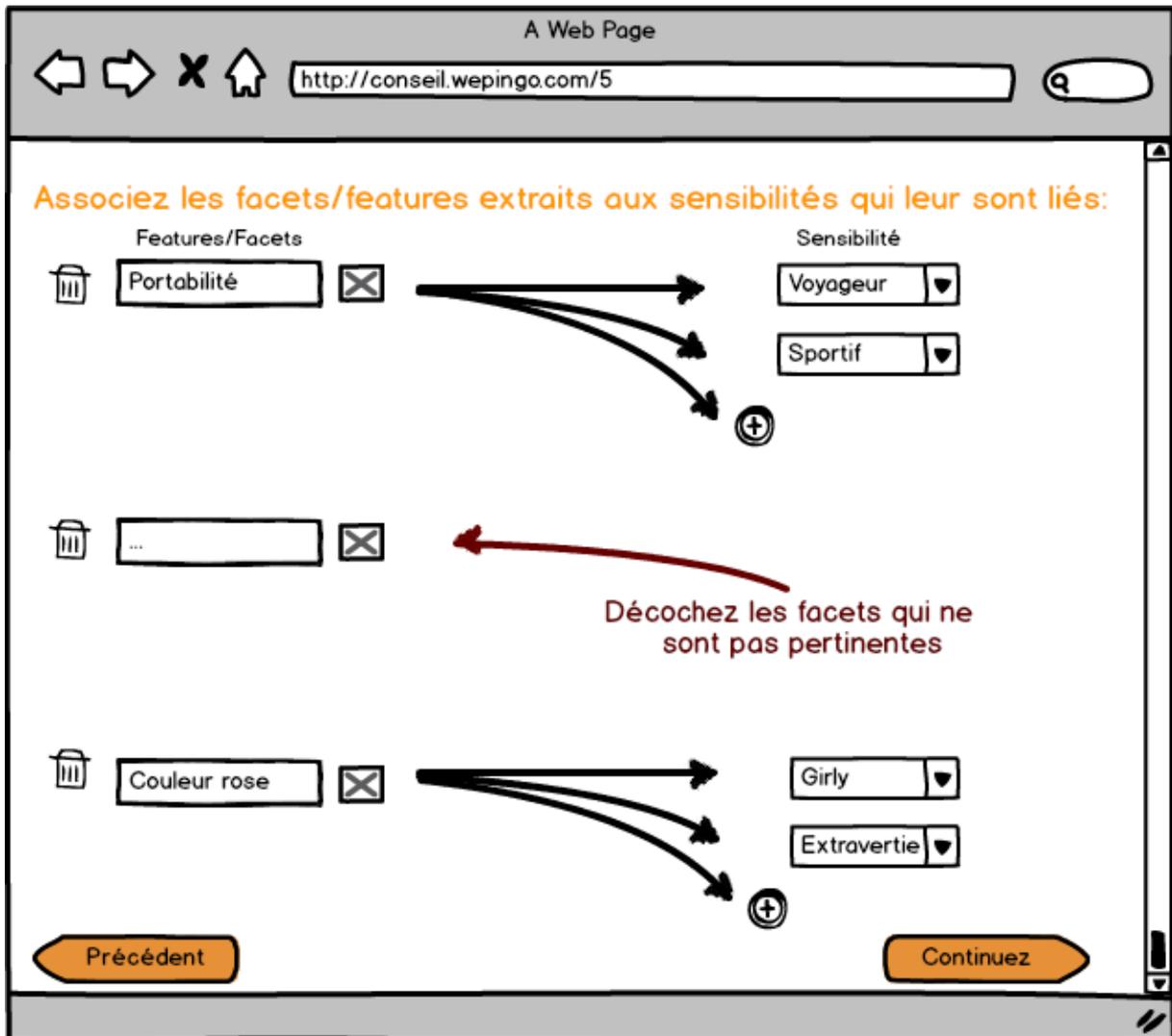
Personnalisez de nouvelles sensibilités:

Précédent Ajoutez en d'autres

- 3) Il va ensuite sélectionner ses catalogues et les faire traiter par notre algorithme qui renverra des produits avec leurs facets et features pondérés.



- 4) Il va donner des règles liant facets/features/catégorie de produit/produits (soit l'un d'eux soit plusieurs) aux sensibilités qu'il a créé précédemment. S'il pense à de nouvelles sensibilités il pourra en rajouter à ce moment la.



5) Il note les relations entre chaque features/facets et chaque sensibilité :

A Web Page

http://conseil.wepingo.com/6

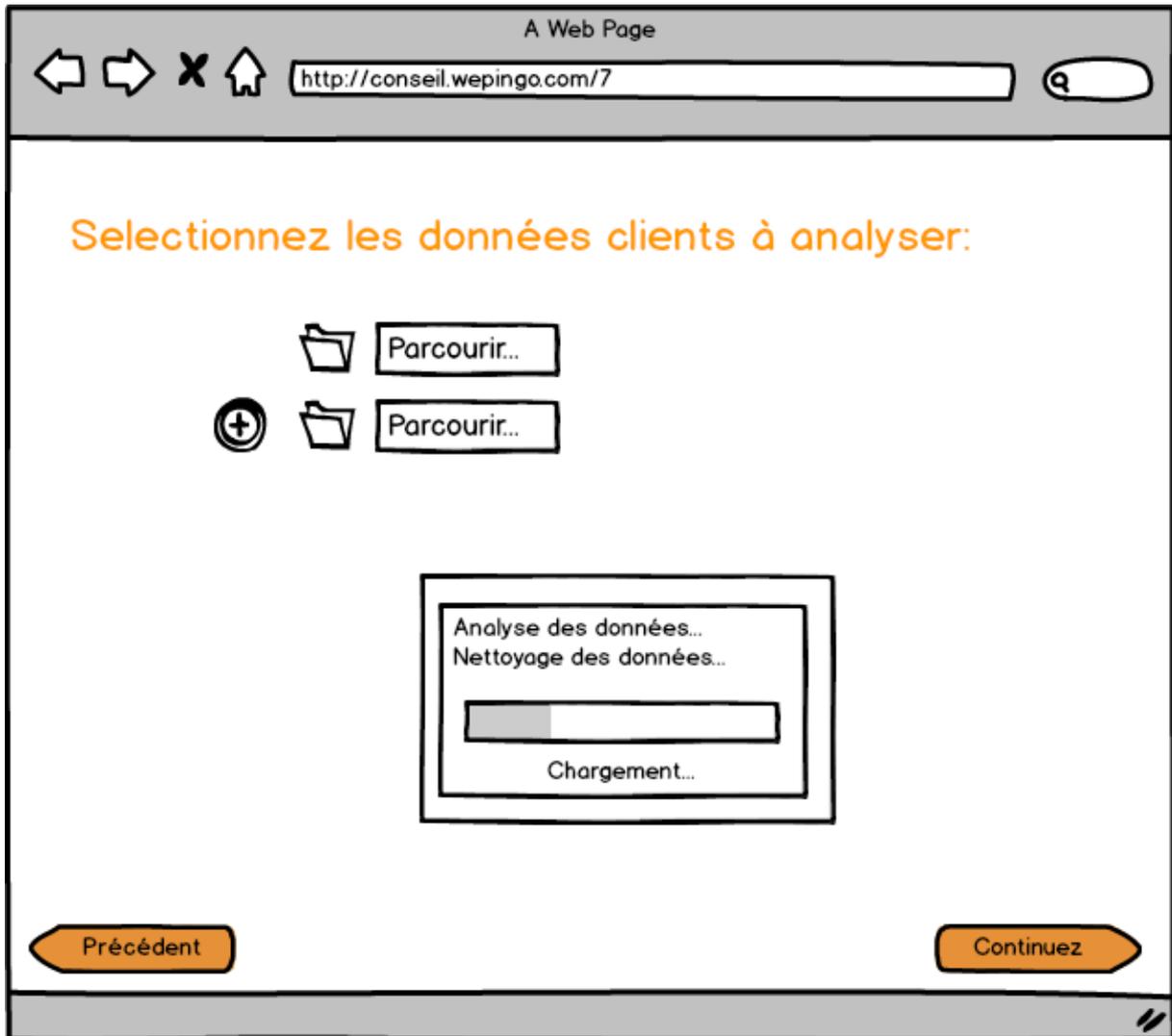
Notez à quel point sont liés les features/facets et les sensibilités:

Features/Facets	Note	Sensibilité
Portabilité	1: Très Faible 2: Faible 3: Moyen 4: Fort 5: Très Fort	Voyageur
Couleur rose	1: Très Faible 2: Faible 3: Moyen 4: Fort 5: Très Fort	Girly

Précédent

Continuez

- 6) Il va ensuite sélectionner les fichiers de sa CRM qu'il veut analyser. On traite dans un premier temps ces données pour qu'elles soient exploitables.



- 7) Il va maintenant choisir successivement des groupes de variables auxquelles il veut s'intéresser pour réaliser les clusters d'utilisateurs, en fonction de produits qu'ils achètent (catégories de produits, features, facets).
- Ces choix dépendent des règles qu'il compte écrire.

A Web Page

http://conseil.wepingo.com/8

CLUSTERING

Analyse de votre CRM

Choisissez la section vous intéresse (gamme, produit, marque...):

Jouets		
Téléphonie mobile	Smartphone	
Cosmétiques	Android	
Electroménager	Mobile nu	Nokia
Assurances	Apple	Samsung
...

Choisissez vos variables pour segmenter la population pour un produit donné:

Facets/features produits:

<input type="checkbox"/> Tout sélectionner
<input type="checkbox"/> Portabilité
<input type="checkbox"/> Capacité mémoire
<input type="checkbox"/>

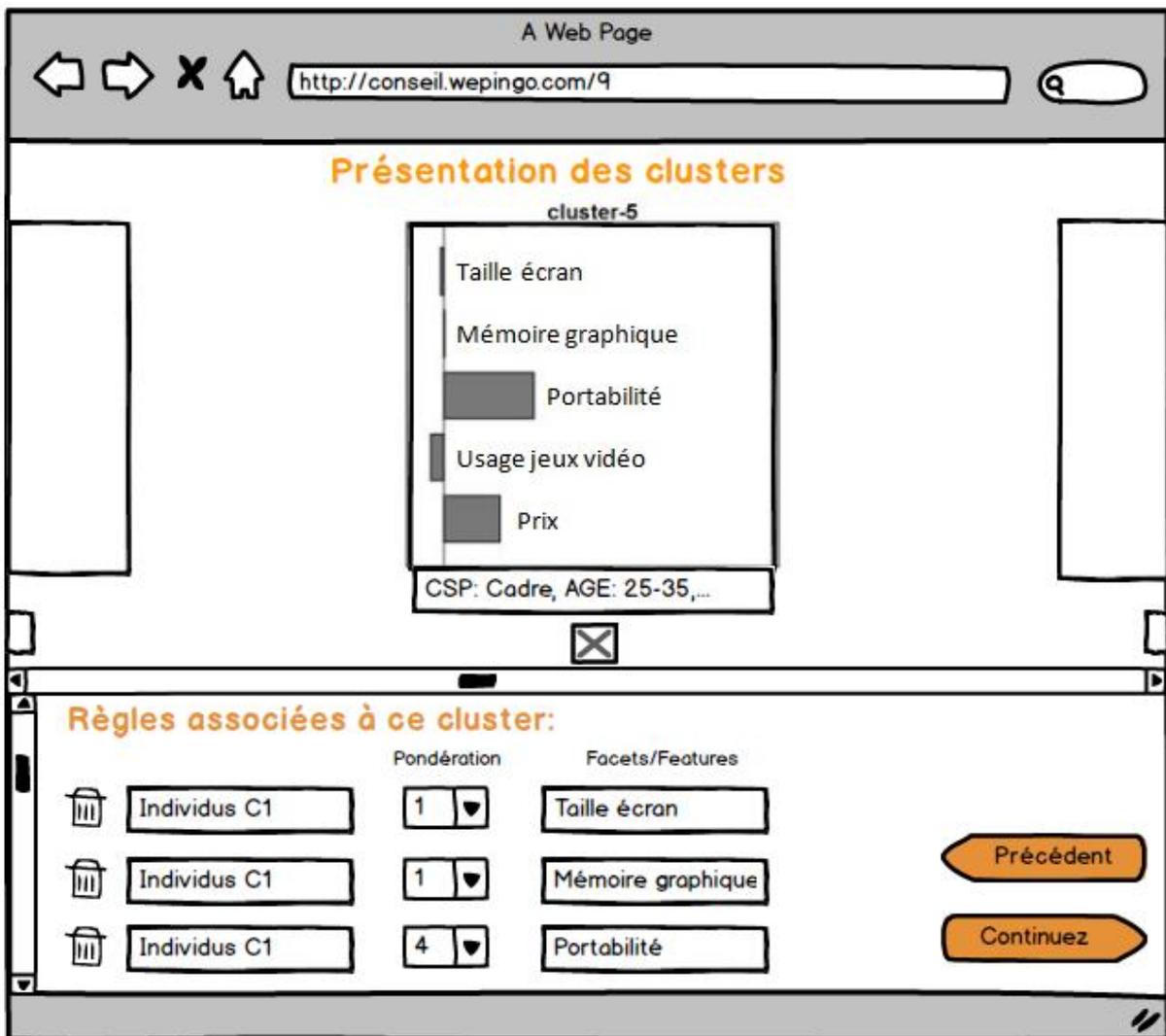
Variables clients:

<input type="checkbox"/> Tout sélectionner
<input type="checkbox"/> Âge
<input type="checkbox"/> CSP
<input type="checkbox"/>

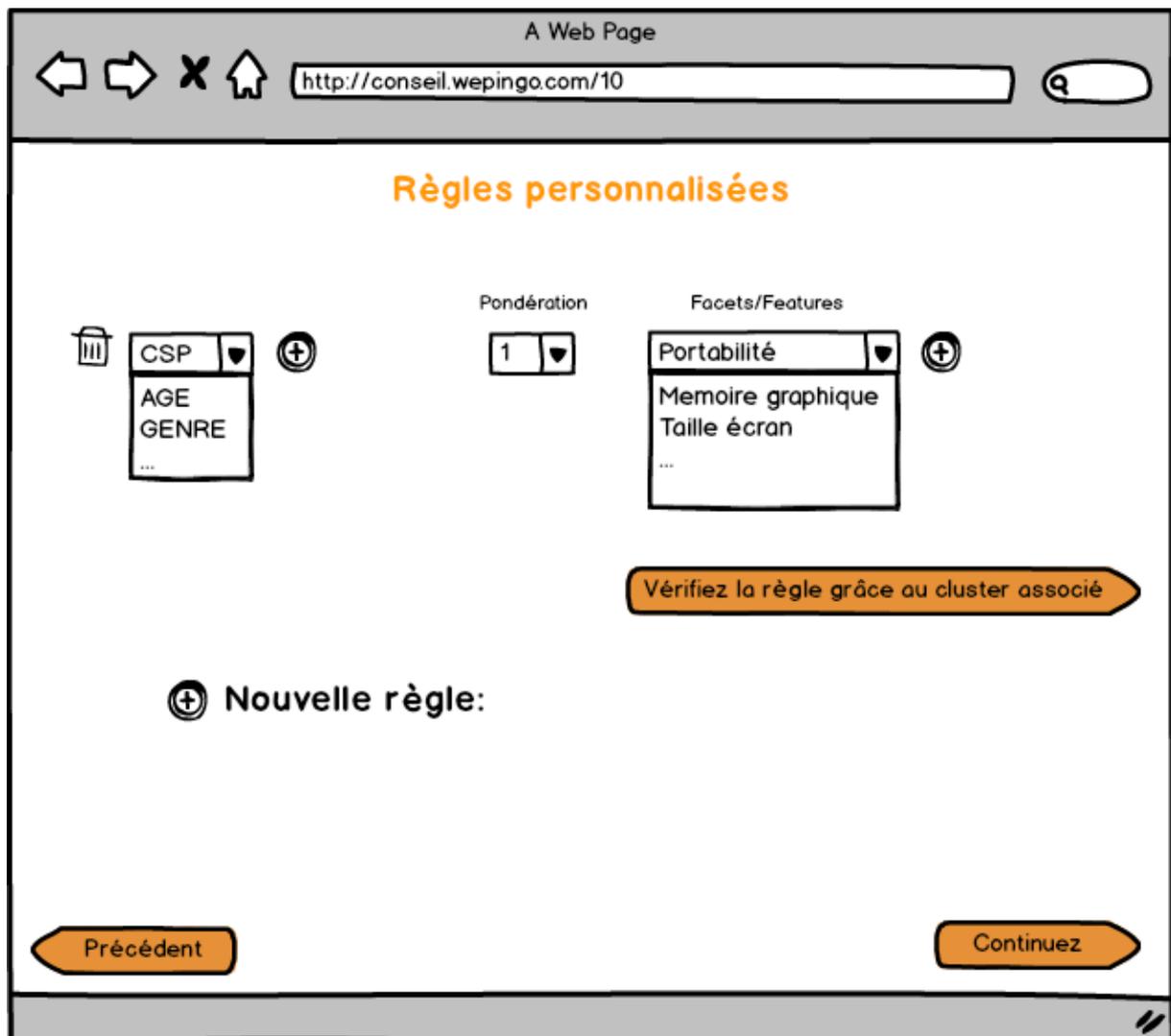
Précédent

Continuez

- 8) L'outil va présenter au marketeur les clusters suivant ses options, et automatiquement l'algorithme associera les caractères communs des individus et les associera sous forme de règles aux features et facets présentes dans les clusters. Le marketeur pourra naviguer parmi les clusters ainsi que parmi les règles et ainsi décider de la pertinence ou non de chacun.



- 9) Il aura aussi la possibilité d'écrire ses propres règles personnalisées et de les soumettre à un clustering pour voir si elles représentent un groupe de personne dans sa CRM ou si cette règle n'est pas judicieuse.



10) Il va ensuite présenter les utilisateurs pour lesquels il a des accès fb. Notre algorithme va les associer à leurs likes et autres données fb. Un algorithme extraira ses likes et les associera à des sensibilités.



11) Il va lier les domaines des likes aux sensibilités par des règles basées sur sa connaissance du domaine.

A Web Page

http://conseil.wepingo.com/12

Règles sur les likes facebook

Notez le lien entre un domaine* et une sensibilité:

*Définition: Domaine: Ensemble qui caractérise un champs lexical regroupant des mots-clés figurant dans des descriptions de likes facebook. ex: sport > foot, rugby, ...

Lancez une analyse automatique (moins précis)

Domaine	Note	Sensibilité
Sport	1: Très Faible	Extravagant
Jeux vidéo	2: Faible	Classique
Restaurant	3: Moyen	Fêtard
Chanteur Rock	4: Fort	...
...	5: Très Fort	

+ Nouvelle règle:

Précédent Continuez

- 12) Les actions de l'outil en temps réel pourront être paramétrées et elles évolueront de façon intelligente en fonction des habitudes enregistrées des utilisateurs visitant le site.



- 13) Il va ensuite pondérer l'importance de l'effet de chaque input sur les sensibilités
Ex : actions sur le site = 5 : très important, likes facebook = 2 : peu important.
Cela lui permettra de personnaliser son approche.

The screenshot shows a web browser window with the address bar containing 'http://conseil.wepingo.com/12'. The page title is 'A Web Page'. The main content area has the heading 'Ordre d'importance des conseils' in orange. Below the heading, the text reads: 'Il vous reste à présent à noter quelle méthode est la plus pertinente parmi les méthodes de conseil précédemment enregistrées.' There are two columns: 'Pondération' and 'Méthode'. Under 'Pondération', there are three empty checkboxes. Under 'Méthode', there are three text input fields containing the following text: 'Règles sur la CRM', 'Règles sur les likes facebook', and 'Actions des utilisateurs'. At the bottom left, there is a button labeled 'Précédent' and at the bottom right, a button labeled 'Terminé'.

Pondération	Méthode
<input type="checkbox"/>	Règles sur la CRM
<input type="checkbox"/>	Règles sur les likes facebook
<input type="checkbox"/>	Actions des utilisateurs

III) L'extraction des sensibilités à partir des données facebook

L'outil final présenté ci-dessus est composé de différentes étapes techniques qui ont été discutées avec les ingénieurs.

Par exemple, la partie portant sur l'extraction des sensibilités à partir des données facebook.

Elle contient un algorithme qui scrappe* les données de google.

En effet, j'ai eu l'idée, pour pouvoir lier les millions de pages de likes existantes aux sensibilités prédéfinies, de procéder à un comptage des itérations des mots clefs en première page de recherche de google lorsque l'on rentre le nom du like dans la barre de recherche. J'ai développé cet outil en collaboration avec mon collègue ingénieur recherche et développement en Java.

1) La slide pour COTY

J'ai dû tester cette méthode en un temps record pour l'entreprise COTY et voici la présentation que je leur ai faite de l'outil.

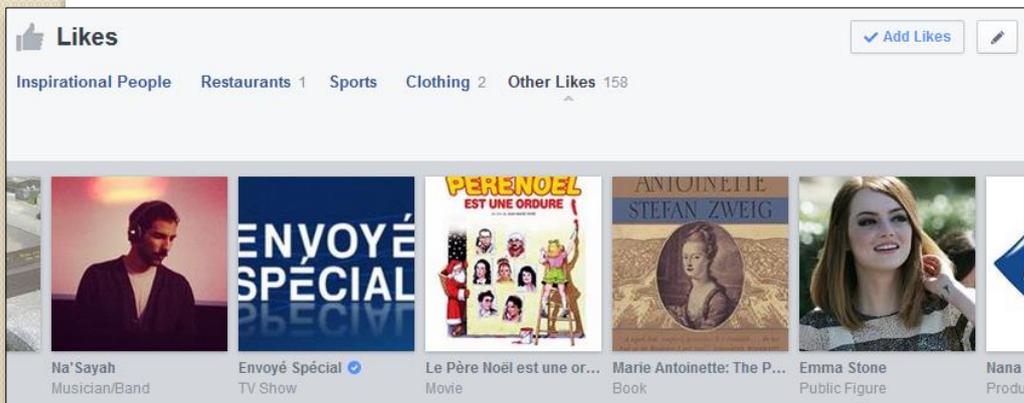




wapingo

Extraction de données facebook

- **Données disponibles:**
 - La liste des « J'aime » : groupes, pages officielles, marques...
 - Le « genre » ➡ important pour la cosmétique
 - L' « A propos » ➡ en étude



wapingo

Extraction de données facebook

- **Création de sensibilités par l'expert marketeur**

- Classique
- Naturelle
- Chic
- Glamour
- ...

- **Association de mots-clefs aux « J'aime » facebook**

Ex: L'utilisateur aime Zara ➔ 'mode', 'vêtements', etc.



Extraction de données facebook

- **Création de domaines regroupant les mots-clefs**

Artistes					Magasines		Vêtements
humoriste	acteur	chanteur rap/rnb	chanteur rock	personnalités	Actualités	People	
humour	acteur	rap	rock	fashion	faits divers	buzz	mode
humoriste	comédien	rnb	rock'n roll	people	météo	clash	tendance
blague	film	hip-hop	grunge	jet set	journal	télé réalité	Vêtements
joke	filmographie	hip hop	heavy metal	mannequin	quotidien	scoop	habits
plaisanterie	cinéma	rappeur	musique	mode	informations	tendance	lookbook
amusement	festival de cannes	r&b		star	info	star	robes
comique	golden globes	musique			presse	people	chaussures
sketch	sexy						talons
drole	septième art						dressing
plaisanterie	7ème art						wishlist
	star						prêt à porter
							escarpins



Extraction de données facebook

- **Création de règles marketing liant les domaines aux sensibilités**

Univers	label_domaine	naturelle	chic	glamour	extravagante	classique
Restaurants	Bar			0,6	0,8	0,5
Restaurants	Salon de thé	0,8	0,8		0,1	0,9
Restaurants	Gastronomie		0,9	0,6	0,1	0,8
Restaurants	Fast food	0,2	0,1	0,4		0,3
Artistes	Humoriste	0,7				
Artistes	Acteur					
Artistes	Chanteur rock		0,1		0,4	0,2
Artistes	Personnalités					
Magasines	Actualités					0,7
Magasines	People	0,2		0,7		

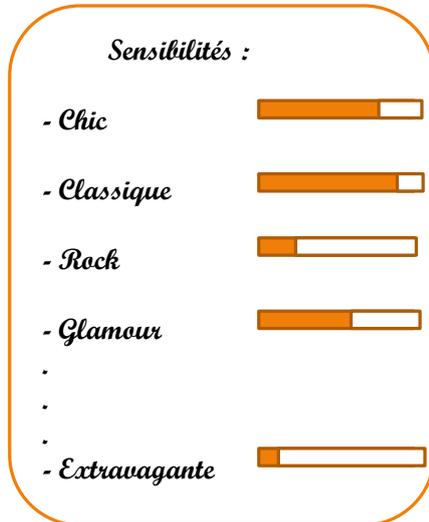
wepingo

Extraction de données facebook

- L'algorithme va associer des valeurs aux sensibilités de chaque « J'aime » d'un utilisateur.
- Il va extraire des informations d'un « J'aime » et les analyser grâce à des méthodes de lecture de texte testées sur différents jeux de données à l'aide de méthodes statistiques de machine learning.
- Nous déterminons, à partir d'analyses statistiques sur la CRM, la **meilleure façon d'associer les sensibilités aux informations collectés.**
- Ce procédé pourra être itéré sur plusieurs jeux de données pour améliorer sa précision. Et la base de donnée d'utilisateurs sera mise à jour régulièrement pour suivre les tendances.

wepingo

Looks et sensibilités



Ces sensibilités sont créées par l'expert marketing, en amont. Comme indiqué précédemment, leurs valeurs sont attribuées par notre algorithme.

Sensibilité importante pour l'utilisateur

Sensibilité faible pour l'utilisateur

wopingo

Algorithme d'extraction de sensibilités pour les différents produits

- L'expert analyse les fiches produits sur le site internet ou tout autre catalogue et déduit grâce à des règles les sensibilités de chaque produit.

1) Il va lister les aspects fonctionnels et techniques qui ressortent des fiches produits.

2) Les règles consisteront à pondérer d'une part l'importance d'un aspect pour chaque produit et d'autre part de lier les différents aspects aux sensibilités.

- On obtient donc un profil de sensibilités faisant référence à un look.

wopingo

Proposition de produits

- A chaque utilisateur nous avons donc **associé un ensemble de sensibilités à des look.**
- A chaque produit nous avons *associé des sensibilités.*
- L'algorithme définit donc en calculant les distances entre les sensibilités de l'utilisateur et des produits, le ou les *produits qui se rapprochent le plus de son look.*

Produits



w^opingo

Affiner nos propositions

- Pour obtenir des résultats toujours plus proches de ce que l'on connaît de l'utilisateur, on aimerait à terme que *d'autres informations interviennent* dans le calcul de ses sensibilités.
- Ces informations pourront être issues du **contexte**, de **son parcours sur le site internet**, et de toute autre donnée que l'on peut collecter sur lui.

w^opingo

Preuve d'un résultat au-delà de toute espérance, l'entreprise COTY (propriétaire de grandes marques de cosmétique internationales telles que Rimmel London, O.P.I, ...), a intégré Wepingo dans son budget d'investissement pour la recherche et la modernisation de ses services.

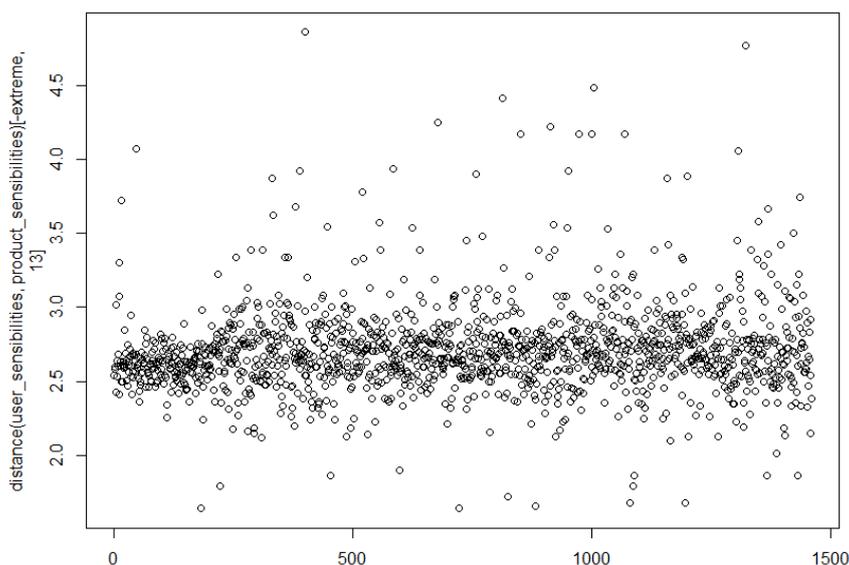
En effet, l'entreprise n'étant pas convaincue par l'idée des questionnaires, le responsable du développement commercial de Wepingo a discuté du projet sur lequel je travaillais et cela s'est annoncé payant. Deux semaines plus tard, suite à l'envoi d'un slide que j'ai réalisé spécialement pour COTY, Wepingo était introduit dans le budget prévisionnel de COTY. Ce slide présentait simplement et brièvement mon projet, il a été présenté à Londres et l'idée fut adoptée par les patrons de COTY. Cette courte période de deux semaines où il a fallu tester et proposer une solution adaptée à leur besoin, m'a montré que j'étais capable de diriger une équipe composée de personnes travaillant dans différents domaines et de rendre un résultat convenable, au point qu'il soit accepté et validé, dans des délais très restreints.

2) Mise en application de la méthode

Ci-dessous, une partie de mon travail sur R pour tester mes idées d'extraction de sensibilités et d'associations de produit avec cette méthode.

Voici la répartition des utilisateurs suite à l'application de l'algorithme google.

Voici la répartition des distances (distance ici est la distance de Manhattan), pour UN des produits de la marque Rimmel London pour environ 1400 femmes de notre jeu de données :



La forte concentration autour de 2.7 peut représenter une distance moyenne entre les utilisatrices et ce produit, en d'autres termes, une attirance moyenne pour ce produit.

Les produits ayant une concentration autour d'une valeur basse sont des produits très standardisés qui peuvent plaire à tout le monde. Ceux qui ont une valeur haute sont des produits qui visent une clientèle plus ciblée.

IV) Implémentation d'une méthode de clustering de type K-means

L'entreprise a réalisé un jeu concours (contest) où il fallait répondre à l'un des 3 questionnaires (1 pour chaque univers). Chaque questionnaire permet de définir les affinités entre la personne participante et les produits de l'univers. Les questions permettent de mesurer une opinion, dans le cas présent à chaque réponse est associée une affinité (*affinity*) plus ou moins grande. Les questions sont indépendantes et ont été posées de façon aléatoire. Le choix de l'univers est d'autant plus représentatif que le cadeau du gagnant était issu de l'univers qu'il avait choisi.

Je disposais d'un certain nombre d'informations sur les participants du contest :

- L'univers choisi
- Les questions et réponses des participants
- L'âge et le sexe d'une partie des participants
- Les affinités liées à chaque réponse
-

1) L'algorithme en R

Pour la partie clustering de l'outil j'ai développé mon propre algorithme sur R qui est disponible à cette adresse :

<https://wall.deblan.org/x1ba0/texte/0/#aes=dXUYPwIEq69Op948WxxtopC6TtaBetWoruUMAlONh3t2wivbSaS2jaekTU5eimIBpWb93QI9qb5DyGT4vxRNmtc3vh21PchbHdh722I2WVXDGHahISNWuqkHRJgBXCDXhdcZtb5i8Uj9bKdxOYy9X49584QosEYrGYb6Jp1hyHzNL6ku713SribtFr822iDczYySrUP1urLenh2INE7XkfeqMSaD5tkE80miNZUNPNyMoAMFjyLtAxVUkWS15u>

J'en ai profité pour le tester sur les données des questionnaires, mais aussi pour faire un rapport d'analyse statistique pour l'entreprise.

Ce rapport a été plébiscité par mon maître de stage, ce qui l'a décidé à revoir en profondeur, avec tous les collègues concernés, l'ensemble des questionnaires de l'entreprise.

2) Qu'est-ce que le clustering ?

Avant d'expliquer les différentes étapes de la programmation de cette méthode je vais présenter brièvement le but d'un clustering.

Le clustering représente un ensemble de méthodes permettant de regrouper ou de hiérarchiser des données.

Le résultat d'un clustering est la génération d'un certain nombre de clusters, dont les centres sont appelés centroids, et qui regroupent des individus.

Dans la méthode K-Means ce nombre de clusters est prédéfini et correspond au K que l'on renseigne en argument de la fonction.

J'ai été confronté à des données brutes qu'il a fallu nettoyer et normaliser pour pouvoir en déduire des résultats exploitables.

3) Les étapes de mon algorithme

1) J'ai d'abord créé un vecteur qui liste les criteria_id (correspondant aux id des questions), pour chacun des univers des 3 questionnaires du contest.

Je l'ai extrait sur mysql :

Pour les Robots de cuisine (id=450) :

```
SELECT crit.id  
FROM CRITERIA crit, PING pi  
where pi.id = 450 and pi.id = crit.ping_id  
order by pi.id;
```

Pour les Smartphones (id=201):

```
SELECT crit.id  
FROM CRITERIA crit, PING pi  
where pi.id = 201 and pi.id = crit.ping_id  
order by pi.id;
```

Pour les Appareils photo (id=6):

```
SELECT crit.id  
FROM CRITERIA crit, PING pi  
where pi.id = 6 and pi.id = crit.ping_id  
order by pi.id;
```

Je les exporte sur excel.

Je les intègre sur R avec des read.csv .

2) Ce vecteur m'a permis d'ordonner les questions et les affinity (sensibilités).

En effet chaque utilisateur a répondu aux questions dans un ordre aléatoire (random dans l'algorithme du questionnaire, ses réponses sont donc désordonnées).

3) A chaque utilisateur j'ai associé un vecteur de ses sensibilités (dans l'ordre)

4) Lors d'un clustering de type K-Means le nombre de cluster est défini par la personne qui utilise la fonction d'où le K. La fonction prend donc un nombre K de vecteurs créé en 2) parmi les utilisateurs et qui sont attribués comme des centroids de départ (chaque centroid aura autant de paramètres qu'il y a de questions dans le questionnaire et ces paramètres seront égaux aux affinités).

J'ai choisi d'attribuer de façon aléatoire les centroids de départ, mais cela risquait de poser un problème si 2 ou plus des centroids de départ avaient les mêmes coordonnées.

⇒ Cela n'a pas posé de problèmes suite à plusieurs essais.

5) J'ai calculé les distances entre chaque utilisateur et chaque centroid et attribué à chaque utilisateur le centroid avec lequel la distance était la plus petite.

J'ai implémenté 2 méthodes de calcul de cette distance. La distance en valeur absolue et la corrélation.

Ce processus amène à un premier clustering avec k clusters

6) J'ai calculé les coordonnées moyennes dans chaque cluster et affecté ces coordonnées comme nouveau centroids.

7) J'ai, de nouveau, calculé les distances (euclidiennes ou autres) entre chaque utilisateur et chaque centroid puis attribué à chaque utilisateur le centroid avec lequel la distance est la plus courte.

8) Mon algorithme est basé sur 10 itérations comme le nombre d'itérations par défaut dans la fonction kmeans() sur R.

Le résultat que me donne ce clustering est un nombre K, comme nos K clusters ou centroids, de « tendances » traduisant des attentes concernant le produit.

```

$centroids
  [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]
[1,]      1 0.8090667 0.8288000 0.9114667 0.5690667 0.8592000 0.8922667 0.9354667 0.9037333
[2,]      2 0.5415612 0.5037975 0.4107595 0.2056962 0.3654008 0.4989451 0.4894515 0.6455696
[3,]      3 0.5010549 0.4130802 0.6919831 0.3082278 0.7407173 0.7613924 0.8540084 0.7719409
[4,]      4 0.7767164 0.7611940 0.5731343 0.2528358 0.7949254 0.6785075 0.7832836 0.7958209
[5,]      5 0.6405405 0.6435435 0.6549550 0.2408408 0.2663664 0.7384384 0.8180180 0.7861862
  
```

Ci-dessus, les centroids issus du résultat de la fonction mykmeans() sur les utilisateurs ayant répondu aux questions sur les appareils photo et avec le nombre K=5.

Les colonnes 2 à 9 représentent les 8 facets correspondants aux 8 questions de ce questionnaire.

Les centroids finaux sont les centres des clusters. On peut donc dire que les choix convergent vers ces centres. Chaque centroid composé de ces facets pondérés traduit un profil type de produit que recherchent une grande partie des utilisateurs.

Par exemple le centroid 3 correspond à un type d'appareil photo moyennement léger (50%), plutôt pas très axé sur le design (41%), axé sur la créativité (69%), pas du tout adapté aux milieux extrêmes (31%), très axé sur la prise de cliché en mouvement (74%), en milieu obscur (76%) et à la prise de photo de portrait (77%) enfin avec une très bonne aptitude à prendre des photos de paysage ou de groupe (85%).

Concrètement, dans l'outil, le nombre d'informations sur les utilisateurs sera suffisant pour faire une analyse sur l'ensemble des utilisateurs qui compose chaque cluster et associer à un profil de produit un profil d'utilisateur. Ce procédé sera fait de façon aussi juste que nous aurons un grand nombre de données précises.

4) Quelques idées pour améliorer cet algorithme

- On peut faire l'étude des clusters sur d'autres distances mathématique et voir laquelle est la plus adaptée. Malheureusement, la pauvreté des données ne nous permet pas de faire ce travail qui demande plus de précision.

Pour étoffer cet aspect des distances on peut se baser sur le livre de Stuart Russel et Peter Norvig « Intelligence artificielle ». On peut utiliser la distance de Minkowski,

$$\left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}$$
 avec différente valeur de p, la distance euclidienne pour p=2, la distance de manhatan pour p=1 (celle-ci est implémentée dans l'algorithme, il y a aussi la corrélation dans l'algorithme). Etant donné que dans notre contest les affinités sont toutes de mêmes dimensions, nous n'avons donc pas besoin de nous attarder dans

des distances plus complexes de type distance de Mahalanobis qui sont utilisées lorsqu'on a des problèmes nécessitant une normalisation des grandeurs.

- Il peut être très intéressant de créer une fonction qui permet de calculer les Average SSE et Average SSB une fois le clustering terminé.

RAPPEL :

$$\text{Average SSE} = 1/N \sum(\sum(\text{dist}(C_i, \text{user}_x)))$$

Average SSE : C'est la somme des distances entre les utilisateurs et le centroid auquel ils sont rattachés.

$$\text{Average SSB} = 1/N \sum(N_i * \text{dist}(C_i, C)^2)$$

Average SSB : C'est la somme des distances entre les clusters.

N_i = nombre d'individu dans le cluster C_i = Centroid i user_x = chaque utilisateur dans le cluster i

N = nombre d'individu total C = le centroid total

Avec les average SSB et SSE on peut comparer le meilleur K .

- On peut améliorer l'arrêt des itérations à la dernière étape en décidant d'arrêter la fonction lorsque les 2 matrices de centroids successives sont égales. C'est-à-dire que, même dans le désordre, chaque ligne correspondant à un centroid peut être retrouvée dans la deuxième matrice, avec des valeurs égales pour chaque colonne (affinité), dans l'ordre.

Le plus important pour une meilleure analyse sera de soigner la prise d'informations, que ce soit au niveau de l'amélioration des questionnaires ou au niveau de la récolte et de la sauvegarde des données sur les utilisateurs.

V) Analyse Statistique du contest Facebook

1^{ère} étape : L'extraction des données, depuis la base de données de l'entreprise.

Sur MySQL j'ai donc cherché à comprendre dans un premier temps comment étaient construites les tables. J'ai réalisé un schéma montrant les liens entre les différentes tables liées au contest.

J'ai écrit une requête pour obtenir les informations que je désirais extraire.

Requête :

```
SELECT su.pinger_id, su.ping_id, s.criteria_id, s.response
```

```
FROM SURVEY_RESPONSE s , SURVEY su
```

```
WHERE su.id = s.survey_id AND su.contestSurvey=1 AND su.pinger_id IS NOT NUL limit 999999;
```

J'ai extrait sur excel un premier jeu de données contenant 4 colonnes : l'id de l'utilisateur, l'id de la catégorie qu'il a choisi lors du contest, les id des questions, les idées des réponses qu'il a données.

2^{ème} étape : L'analyse des données.

Sur ce premier jeu de données j'ai commencé à me familiariser avec les résultats du contest.

Je me suis attaché dans un premier temps à analyser les tendances de réponses en fonction des questions, sans me concentrer sur le parcours global dans le questionnaire de chaque utilisateur.

Cette première analyse m'a poussé à prévenir l'entreprise que certaines questions du quizz ne sont pas pertinentes puisqu'elles mènent les utilisateurs à choisir, à plus de 90%, seulement 2 réponses parmi 4. De plus, ces réponses ne sont pas opposées mais traduisent une affinité forte (ou faible) avec la feature/facet étudiée par la question.

Ex : A la question : Êtes-vous intéressé par des fonctions facilitant l'entretien de votre appareil ? (l'id de cette question ou Criteria_id est 135)

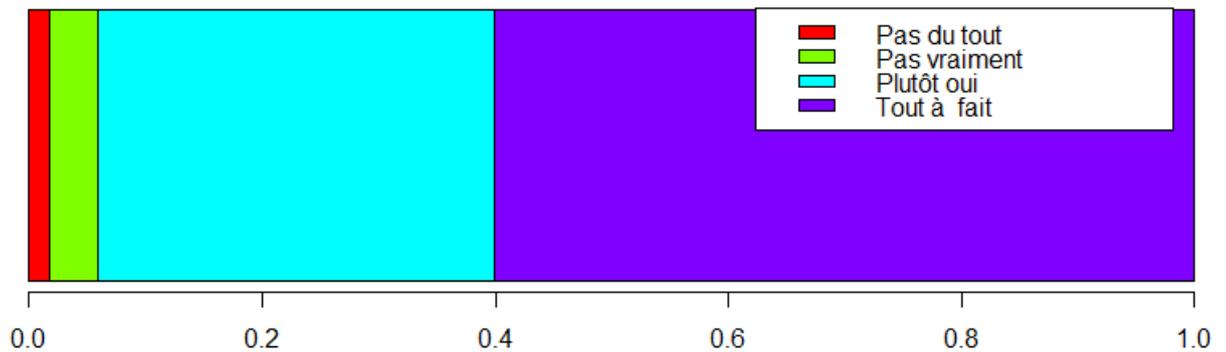
Les gens ont répondu :

A 60% = Tout à fait ;

A plus de 30% = Plutôt oui ;

A moins de 10% = Pas du tout et Pas Vraiment.

Pourcentage des réponses pour le criteria_id=135



On peut donc dire que cette question n'est pas judicieuse. Elle ne révélera que trop rarement une tendance différenciant plusieurs utilisateurs.

Où se situe le problème ? Est-ce la formulation de la question ? Est-ce la question en elle-même ? Y a-t-il un problème avec les choix de réponse ?

Au marketeur de le dire, de changer la formulation, d'analyser à nouveau les retours etc...

On pourrait donc penser à un algorithme simple qui analyserait régulièrement les réponses aux questionnaires et qui signalerait une question à modifier via mail ou via un pop-up sur l'outil.

Idée d'algorithme : Par exemple, donner un taux de réponse minimum/maximum pour chaque choix dans une question. Puis le programme signalera une réponse qui est hors seuil.

Idée de solution :

Comme le disait un de mes professeurs, Mme Maumy-Bertrand, « Lors d'un sondage ou lors de tout autre questionnaire... », parlait de sondage politique, « ...on ne veut pas que les choix de réponses soient biaisés du fait que la personne qui répond devine par avance ce que l'on veut déduire de ses réponses ».

Par exemple, dans le cadre de sondage politique, une personne peut avoir honte de dire qu'elle vote FN, cela explique souvent des écarts significatifs dans les résultats de ce parti entre réalité et sondage.

Lors de nos questionnaires c'est la même chose, on ne veut pas que la personne nous réponde « Ce qui se fait de mieux » à chaque question, on veut lui proposer quelque chose qui « colle » au mieux à l'usage qu'elle fera de l'objet.

Une idée pourrait être d'utiliser les méthodes de psychologie pour la conception des questions, les mêmes que les psychologues utilisent lors de leurs tests et questionnaires. Une des méthodes utilisée est de demander l'inverse de ce que l'on veut savoir, cela permet de brouiller les pistes et d'obtenir des résultats plus probants.

Par exemple dans le questionnaire sur les appareils photo on demande : « Quelle importance accordez-vous à la prise de photo de sujet en mouvement ? » deviendrait : « Quelle

importance accordez-vous à la prise de photo de sujet statique » et plus la réponse sera basse pour cette question, plus l'affinité à la facet « prise de photo de sujet en mouvement » sera grande.

Attention, cela dit, de ne pas complexifier la question ce qui nuirait au résultat. L'utilisateur risque de répondre au hasard ou de passer la question, si elle n'est pas facilement compréhensible.

Pour avoir des meilleures informations sur les utilisateurs, il serait préférable d'associer des facets/features négative selon la réponse.

Exemple : Disons que la légèreté et la portabilité s'opposent à la solidité. Donc si un utilisateur met une réponse avec une affinité haute pour la portabilité, en contrepartie une sensibilité négative à la solidité va s'appliquer. Dans une autre question s'il choisit une réponse avec une affinité haute aussi pour la solidité, en contrepartie une sensibilité négative à la portabilité va s'appliquer. Le résultat sera donc deux sensibilités moyennes pour ces deux features.

Disons que c'est un calcul simple :

Si à la question X pour la réponse « haute » j'associe 1 en Portabilité et -0.5 en Solidité.

Si à la question Y pour la réponse « haute » j'associe 1 en Solidité et -0.5 en Portabilité.

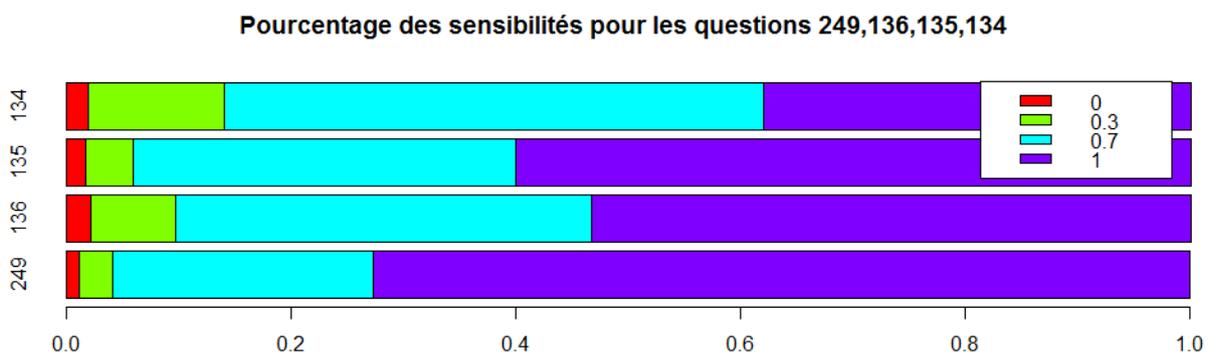
Alors la somme des 2 me donnerait une sensibilité de 0.5 en Portabilité et de 0.5 en Solidité.

Pour répondre au mieux aux attentes de l'utilisateur, on peut aussi lui demander de pondérer qu'une feature est plus importante qu'une autre, cette méthode sera du coup encore plus judicieuse. Pour évaluer le lien entre deux facets/features il faut s'intéresser à leur corrélation, leur dépendance.

Autre remarque :

Les utilisateurs ont tendances à en demander toujours plus, à être exigeants, d'après les résultats des 3 questionnaires. Ce qui rend la différenciation des utilisateurs plus compliquée.

Ex :

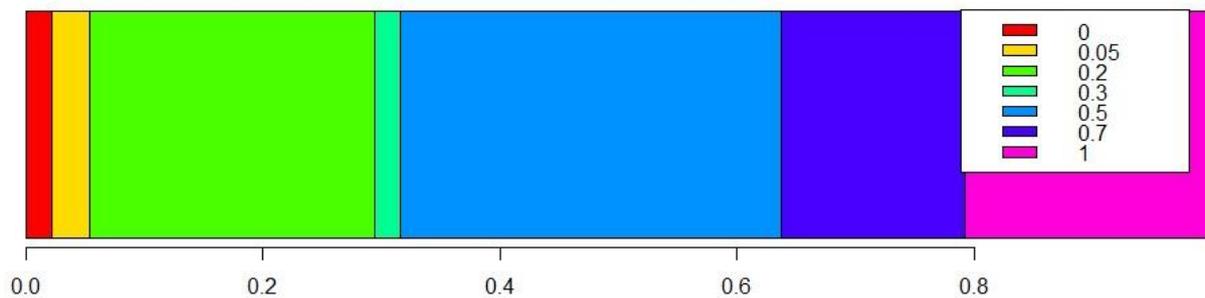


Lorsque l'on a ce type de répartition de réponses (en ordonnées les id des questions et en abscisse le niveau d'exigence), on peut s'attendre à proposer à 80% des gens les mêmes produits, ou de façon équivalente, d'avoir les mêmes profils de sensibilités pour 80% des utilisateurs.

Pour augmenter les chances de pouvoir différencier deux utilisateurs, il est intéressant, au vu des diagrammes, de poser des questions plus précises sur le produit. Plus il y a de réponses possibles plus il y a de chances de différencier deux utilisateurs.

Pour preuve, l'exemple suivant:

Pourcentage des sensibilités pour la question 82



Attention cependant à réajuster les réponses possibles pour ne pas se retrouver avec des segments ne contenant que 2% d'utilisateurs comme c'est le cas ici.

1) Quelques statistiques générales

Attention : Les statistiques ci-dessous sont issues des données dont je disposais, c'est-à-dire seulement sur une partie des participants du contest.

Je m'étais concentré sur l'univers des appareils photo qui était le plus précis parmi les 3 univers, bien que très incomplet lui aussi.

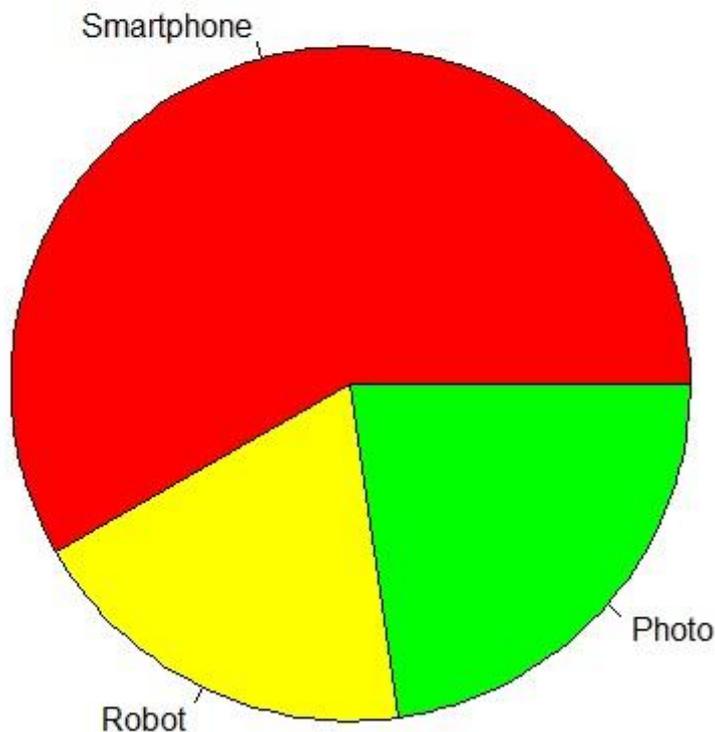
Pourcentage de gens ayant répondu à nos 3 catégories :

Smartphone : 58.3% → Ce domaine est celui qui à eu le plus de succès, et de loin !

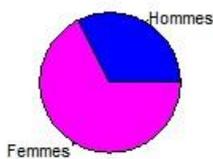
Robot de cuisine : 19%

Appareils photo : 22.7%

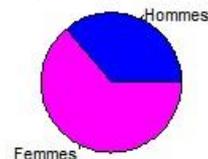
Répartition du choix de catégorie



Répartition totale sur le contest



Répartition pour les Robots de cuisine



Répartition pour les Appareils photo



Répartition pour les smartphones



Smartphone :

32% d'hommes et 68% de femmes composent les participants à ce questionnaire.

Robot de cuisine :

36% d'hommes et 64% de femmes composent les participants à ce questionnaire.

Appareils photo :

32% d'hommes et 68% de femmes composent les participants à ce questionnaire.

2180 Hommes ont participé au total.

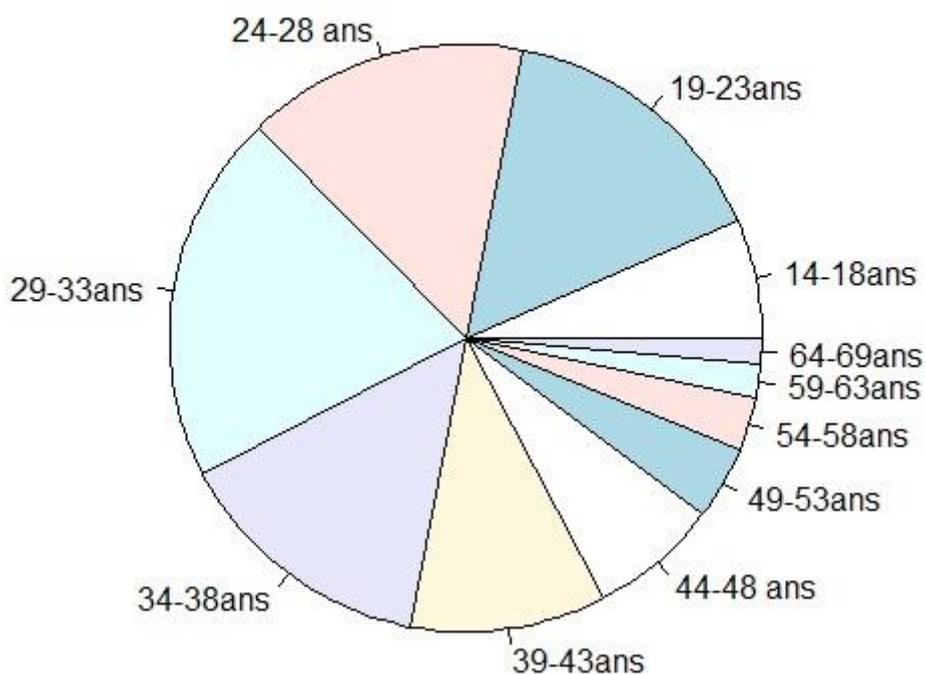
4462 Femmes ont participé au total.

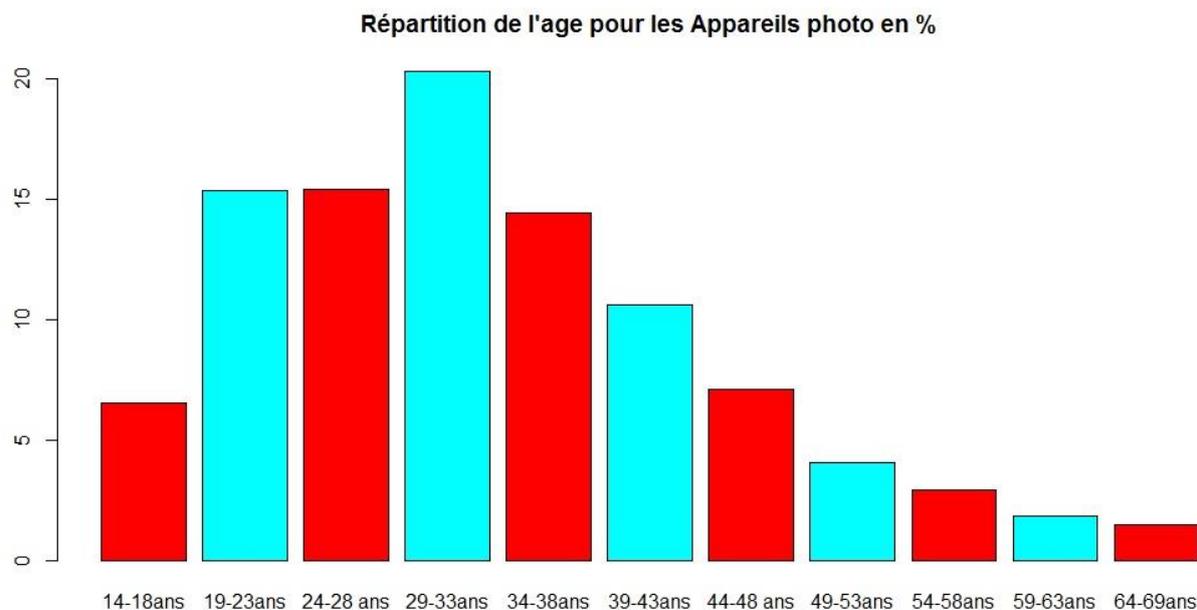
67% de Femmes au total.

33% d'Hommes au total

2) Statistiques sur l'âge

Répartition par tranche d'âge
pour les Appareils photo



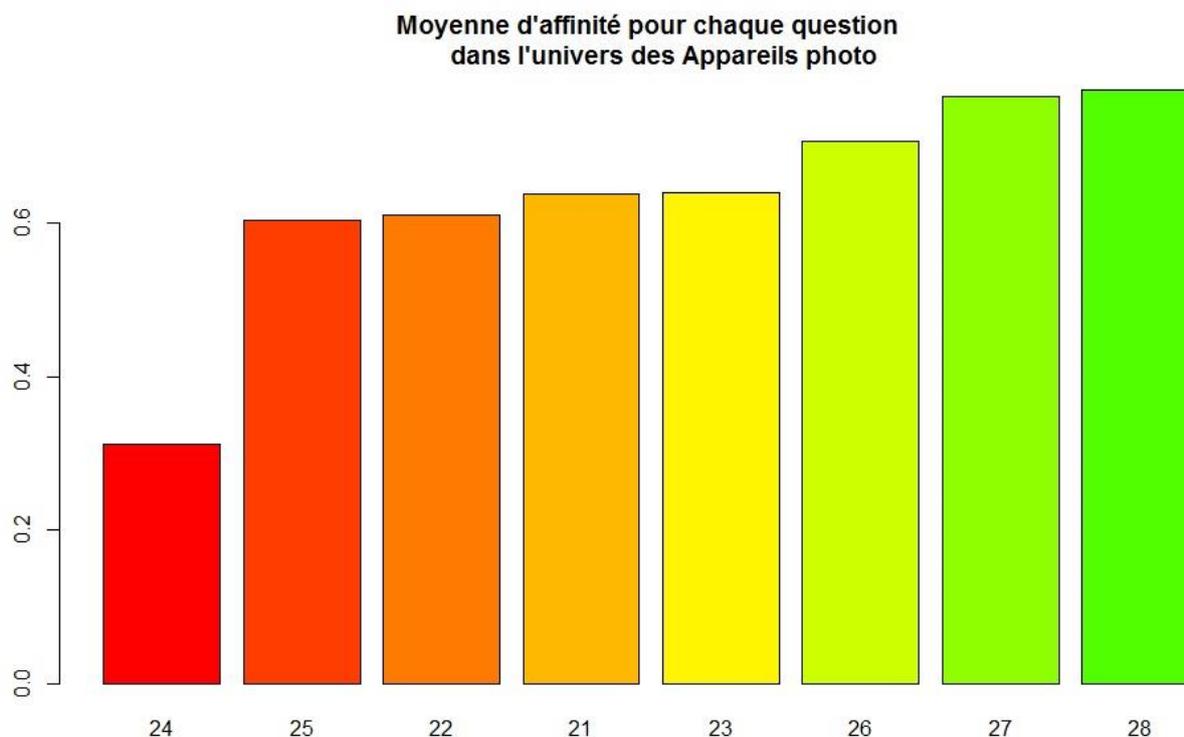


On voit bien sur ces deux graphique que les personnes entre 19 et 38 ans sont ceux qui ont le plus répondu au contest, et plus particulièrement pour l'univers des appareils photos.

Avec une mention spéciale aux 29-33ans qui représentent à eux seuls 20% des utilisateurs.

3) Statistique sur les affinités

Si l'on fait les moyennes d'affinités pour chaque question on peut par exemple se rendre compte que la prise de photo en portrait est ce qui importe le plus pour les utilisateurs, ou encore que la prise de photo en mouvement est ce qui les intéresse le moins.



Légende :

- 21: Quel degré d'importance accordez-vous à la légèreté ?
- 22: Quel degré d'importance attachez-vous au style ?
- 23: Aimerez-vous réaliser des photos mettant en valeur votre créativité ?
- 24: Pensez-vous réaliser des clichés en environnement extrême ?
- 25: A quelle fréquence prenez-vous des photos de sujet en mouvement ?
- 26: Quelle importance accordez-vous à la prise de photo en environnement obscur ?
- 27: Etes-vous amené à prendre des photos de paysage ou de groupe ?
- 28: Quel niveau d'importance attribuez-vous à la prise de photos de portrait ?

Ce graphique montre, une fois de plus, les problèmes liés à la façon de poser la question. Les questions trop directes de type « Quel degré d'importance attribuez-vous » + la facet, appellent à des réponses que je qualifie de « Qui peut le plus, peut le moins ! » et qui montrent

le niveau d'exigence des utilisateurs d'autre part. Une question comme la 24 liée aux habitudes amène à des réponses plus personnelles.

Conclusion

Ce stage de fin de Master 1 m'a permis de compléter mes connaissances et d'en acquérir bien d'autres. Le fait d'avoir pu l'effectuer dans une start-up m'a permis d'analyser un projet depuis sa conception, ce qui m'a permis d'avoir une approche plus précise sur plusieurs métiers de l'informatique. Il a certes fallu être rapidement autonome et efficace car un projet évolue très vite dans son lancement, et j'ai pu rapidement m'apercevoir des difficultés que l'on peut rencontrer lorsque l'on se lance dans le monde de l'innovation. Ces difficultés apparentes deviennent un sentiment de grande réussite lorsqu'on les surpasse et on se sent tel un explorateur lorsque nos premiers résultats concrets sont sous nos yeux.

Du fait de la durée restreinte du stage je n'ai pas pu accompagner le projet jusqu'à sa finalisation, j'ai néanmoins pu poser les bases d'un outil qui seront exploitées par mes collègues pour en faire un outil pleinement fonctionnel.

Programmes utilisés

- Orange canvas
- Balsamiq
- Mysql
- R
- Wepingo manager
- Java
- Environnement de programmation des API sur facebook
- Collaboration avec les ingénieurs programmant sur Java pour certaines parties du projet

Lexique

- One to One : c'est la stratégie marketing ou chaque client est vu comme un être unique.
- Scrapping : c'est le fait d'extraire les données qui composent une page web.

Bibliographie

- Pearson, Stuart Russell et Peter Norvig (2010). Intelligence artificielle, 3^e édition.
- Wiley, Konstantinos Tsipis et Antonios Chorianopoulos (2010), Data Mining Techniques in CRM.