Stage de Recherche
« Modélisation des usages utilisateurs pour le Crowdsourcing à grande échelle »

CARRANZA ALARCON, Yonatan Carlos
Données, Connaissances et Traitement de Langues

SUPERVISEURS
JOLY Alexis, PACITTI Esther , SERVAJEAN Maximilien

11 juin 2015

Plan

Modélisation des usages utilisateurs à grande échelle

Plan

Modélisation des usages utilisateurs à grande échelle

Introduction

Qu'est ce que le Crowdsourcing ?

Definition

Le *Crowdsourcing* est l'externalisation de micro-tâches facile à résoudre d'une organisation/entreprise envers un grand groupe de personnes connectées à Internet sous la forme d'appels ouverts.

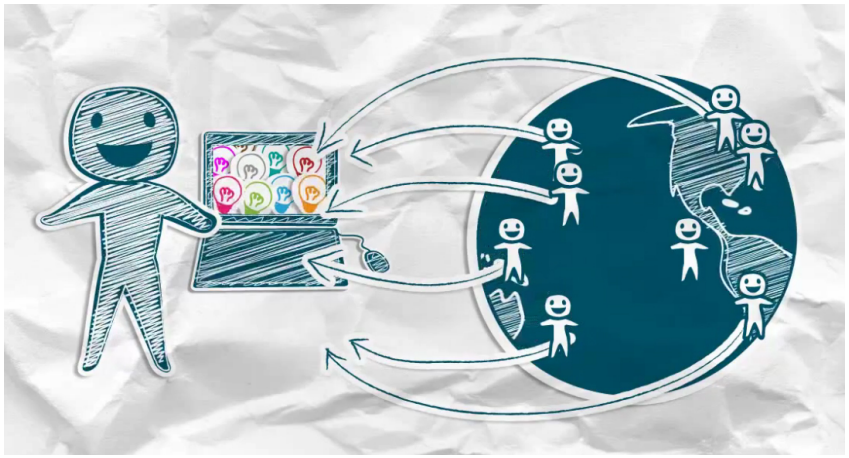


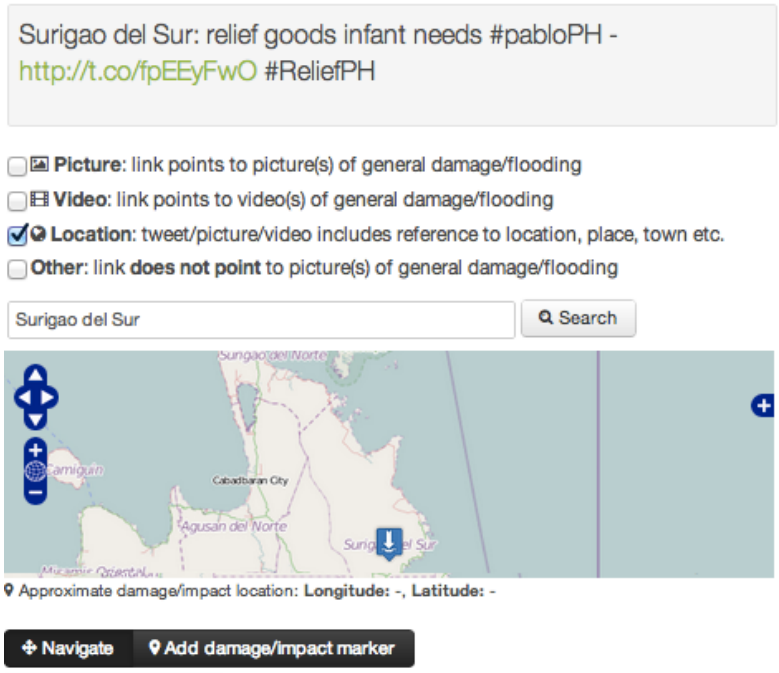
Fig.: Plate-forme Crowdsourcing

Introduction

Qu'est ce que le Crowdsourcing ?

Definition

Le *Crowdsourcing* est l'externalisation de **micro-tâches** facile à résoudre d'une organisation/entreprise envers un grand groupe de personnes connectées à Internet sous la forme **d'appels ouverts**.



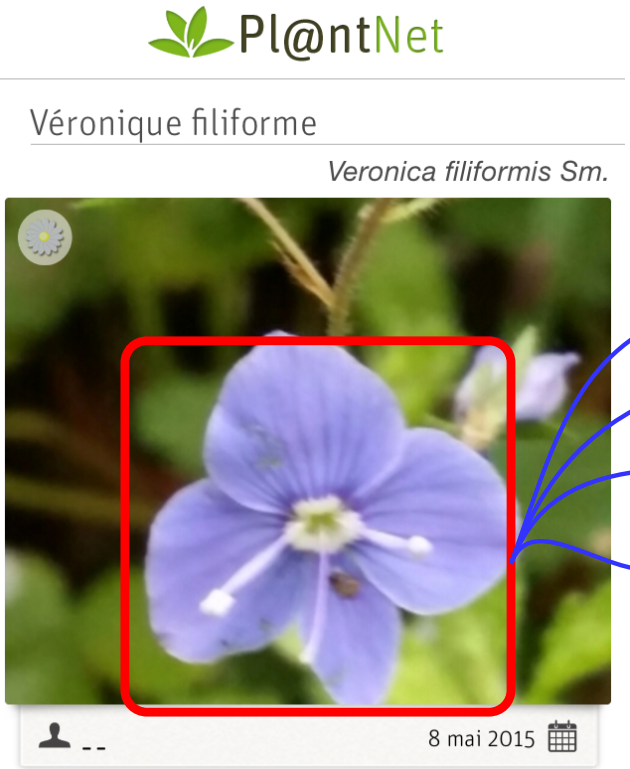
Micro-tâche : Veuillez étiqueter le tweet suivant par la ou les catégories qui décrivent le mieux les liens qu'il contient.
(www.crowdcrafting.org)

⋆ ⋆ ⋆ Problème de classification

Contexte

Cas d'étude - Pl@ntNet

Pl@ntNet est une application d'aide à l'identification interactive des espèces de plantes.



⋆ Classification des plantes

(1) Marguerite
(2) Zygopétalum
⋮
(n) Véronique filiforme

Espèces de plantes possibles.
Beaucoup de dimensions.
e.g. 5 mil espèces en France.

Tâche

Problématiques

Quel(s) problème(s) pouvons-nous rencontrer ?

1. Les plateformes de *Crowdsourcing* ont toujours été évaluées à petite échelle, c.a.d. peu de catégories (e.g : classification des tweets).
2. Il n'existe pas de méthodes permettant d'évaluer les réponses d'utilisateur à grande dimensionnalité.
3. Les applications telles que Pl@ntNet manipulent un grand nombre de catégories.
4. Il n'existe pas de *Benchmark* pour évaluer les solutions de *Crowdsourcing*.

Objectif

Quelle(s) solutions(s) allons-nous proposer

L'objectif de notre travail se focalise à comprendre les différentes étapes du *Crowdsourcing* et à modéliser les différentes comportements d'utilisateurs.

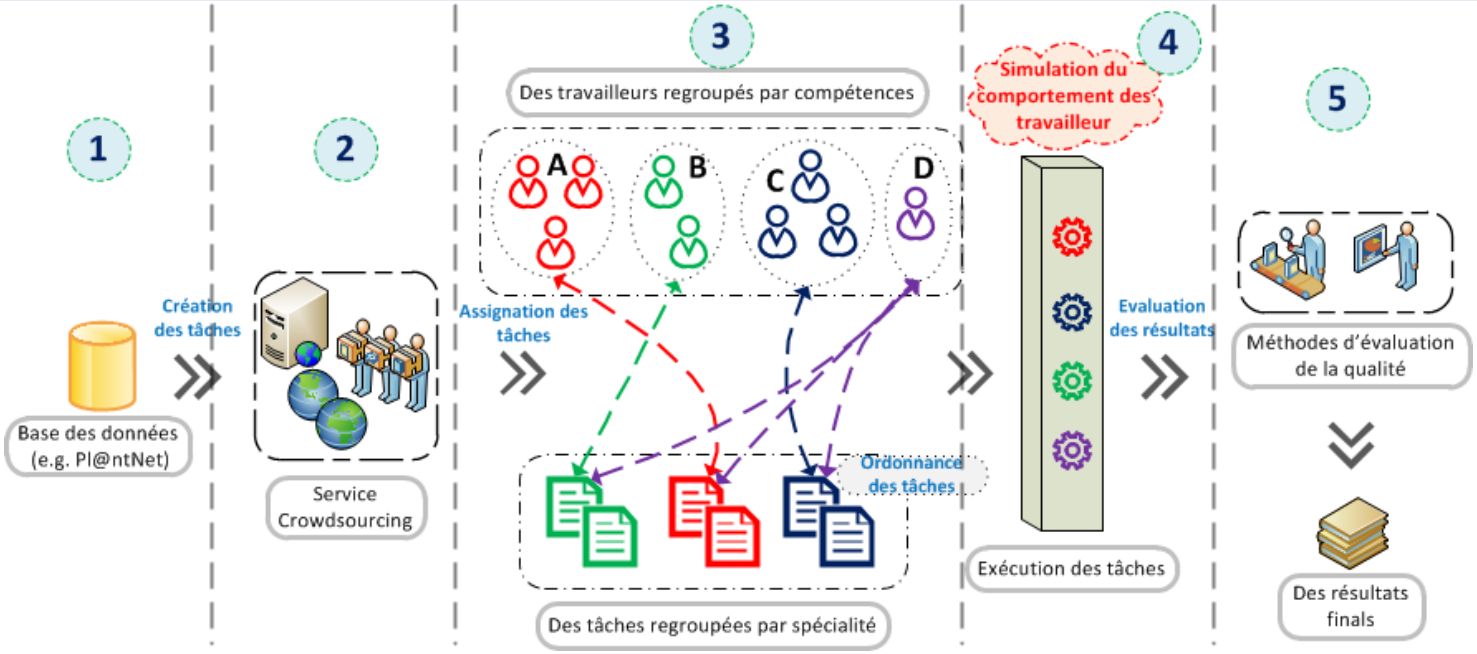


Fig.: Schéma du Crowdsourcing

Plan

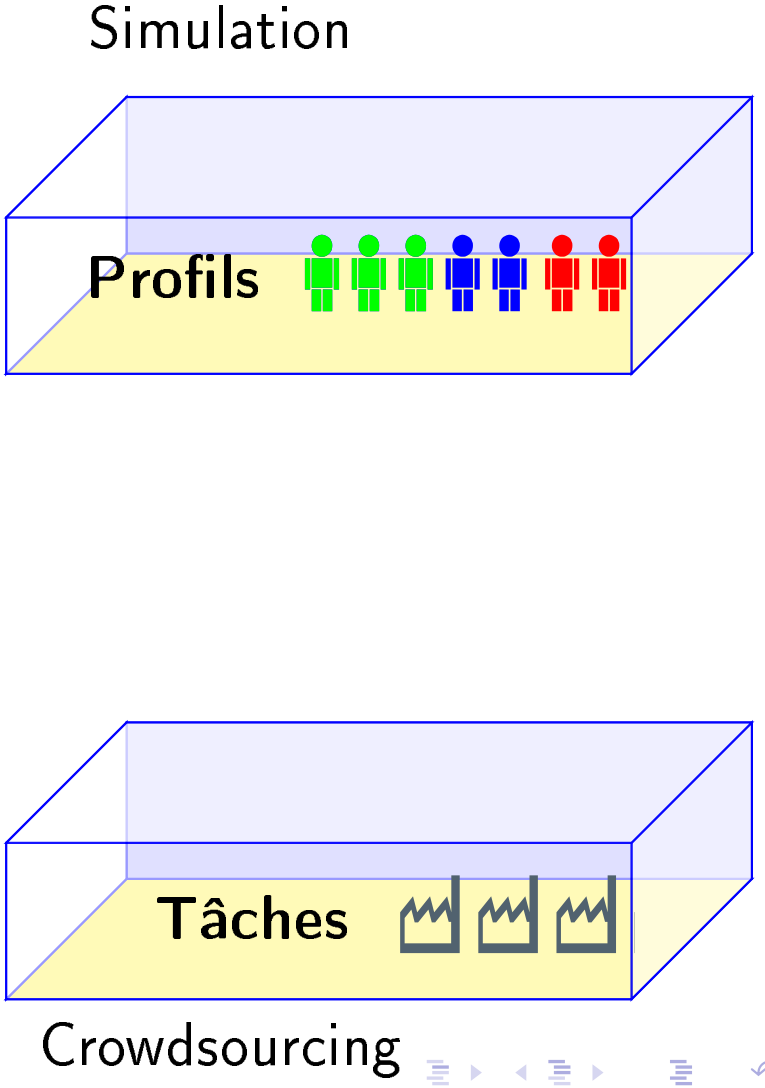
Modélisation des usages utilisateurs à grande échelle

Algorithme Général

Algorithm 1 Simulation Naïf

Input : Ensemble des tâches T_i
Input : Ensemble des utilisateurs par profil P_i
Output : Ensemble des réponses estimées **U**

1: **for each** $T_i \in \{T_1, T_2, \ldots, T_n\}$ **do**
2: **for each** $P_i \in \{P_1, P_2, \ldots, P_m\}$ **do**
3: **for each** $u_i \in P_i$ **do**
4: $\hat{U}_i$ = réponse_estimated(T_i, u_i)
5: **end for**
6: **end for**
7: **end for**
8: **return** $\hat{U}_i \in \{\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n\}$



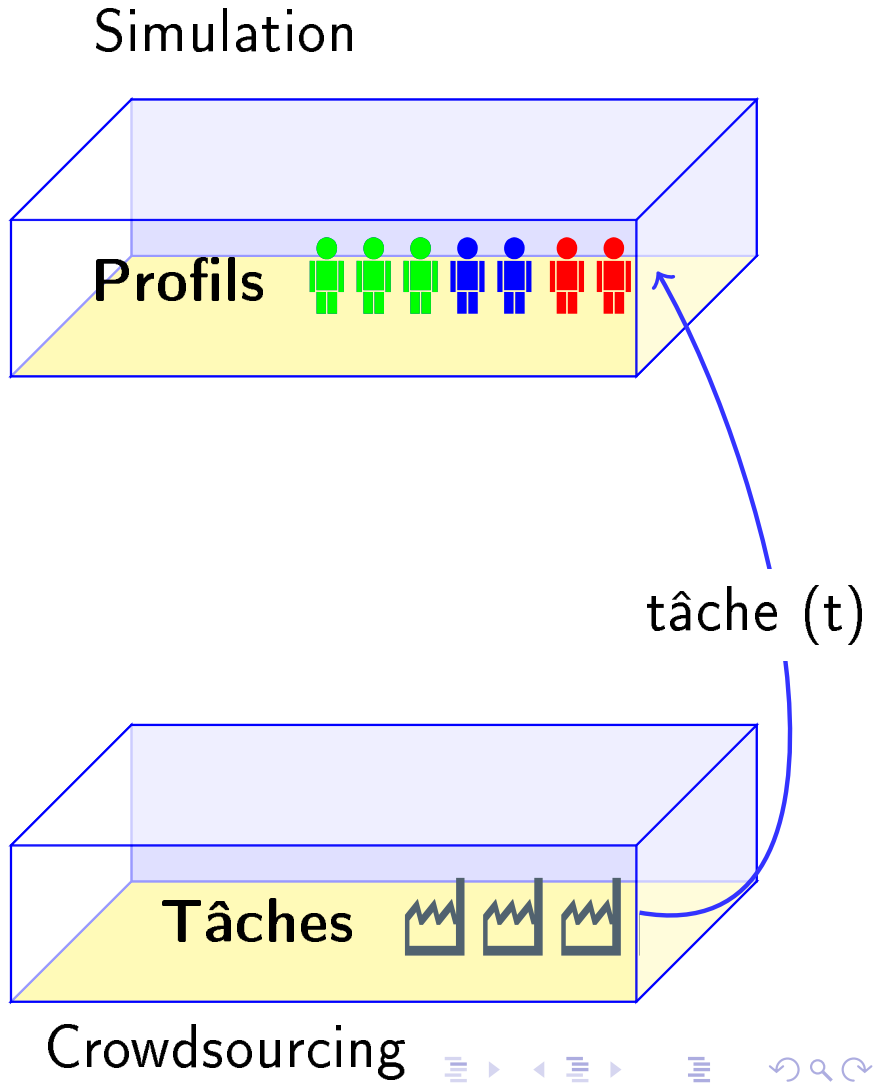
Algorithme Général

Algorithm 1 Simulation Naïf

Input : Ensemble des tâches T_i
Input : Ensemble des utilisateurs par profil P_i
Output : Ensemble des réponses estimées **U**

1: **for each** $T_i \in \{T_1, T_2, \ldots, T_n\}$ **do**
2: **for each** $P_i \in \{P_1, P_2, \ldots, P_m\}$ **do**
3: **for each** $u_i \in P_i$ **do**
4: $\hat{U}_i$ = réponse_estimated(T_i, u_i)
5: **end for**
6: **end for**
7: **end for**
8: **return** $\hat{U}_i \in \{\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n\}$

Le système crowdsourcing envoi une tâche (t) au système de simulation.



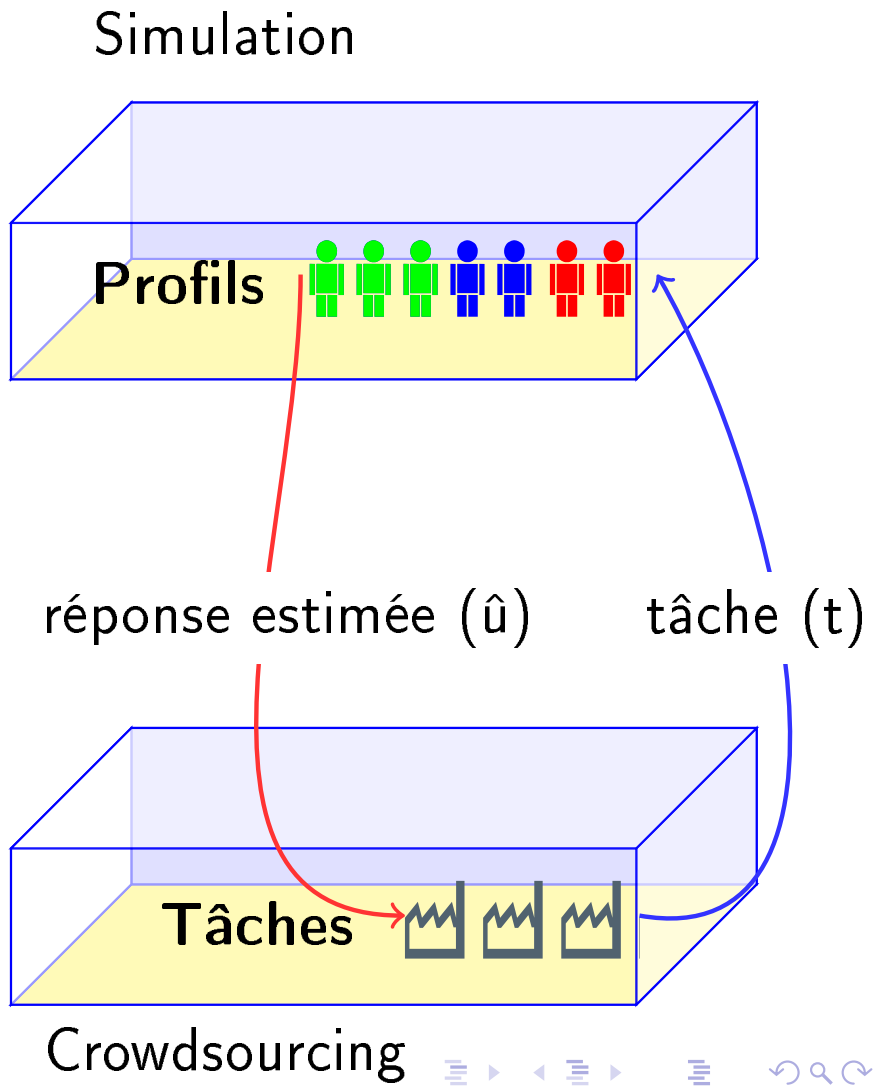
Algorithme Général

Algorithm 1 Simulation Naïf

Input : Ensemble des tâches T_i
Input : Ensemble des utilisateurs par profil P_i
Output : Ensemble des réponses estimées **U**

1: **for each** $T_i \in \{T_1, T_2, \ldots, T_n\}$ **do**
2: **for each** $P_i \in \{P_1, P_2, \ldots, P_m\}$ **do**
3: **for each** $u_i \in P_i$ **do**
4: $\hat{U}_i$ = réponse_estimated(T_i, u_i)
5: **end for**
6: **end for**
7: **end for**
8: **return** $\hat{U}_i \in \{\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n\}$

Le système de simulation reçoit la tâche (t), la traite, puis donne une réponse estimée û.



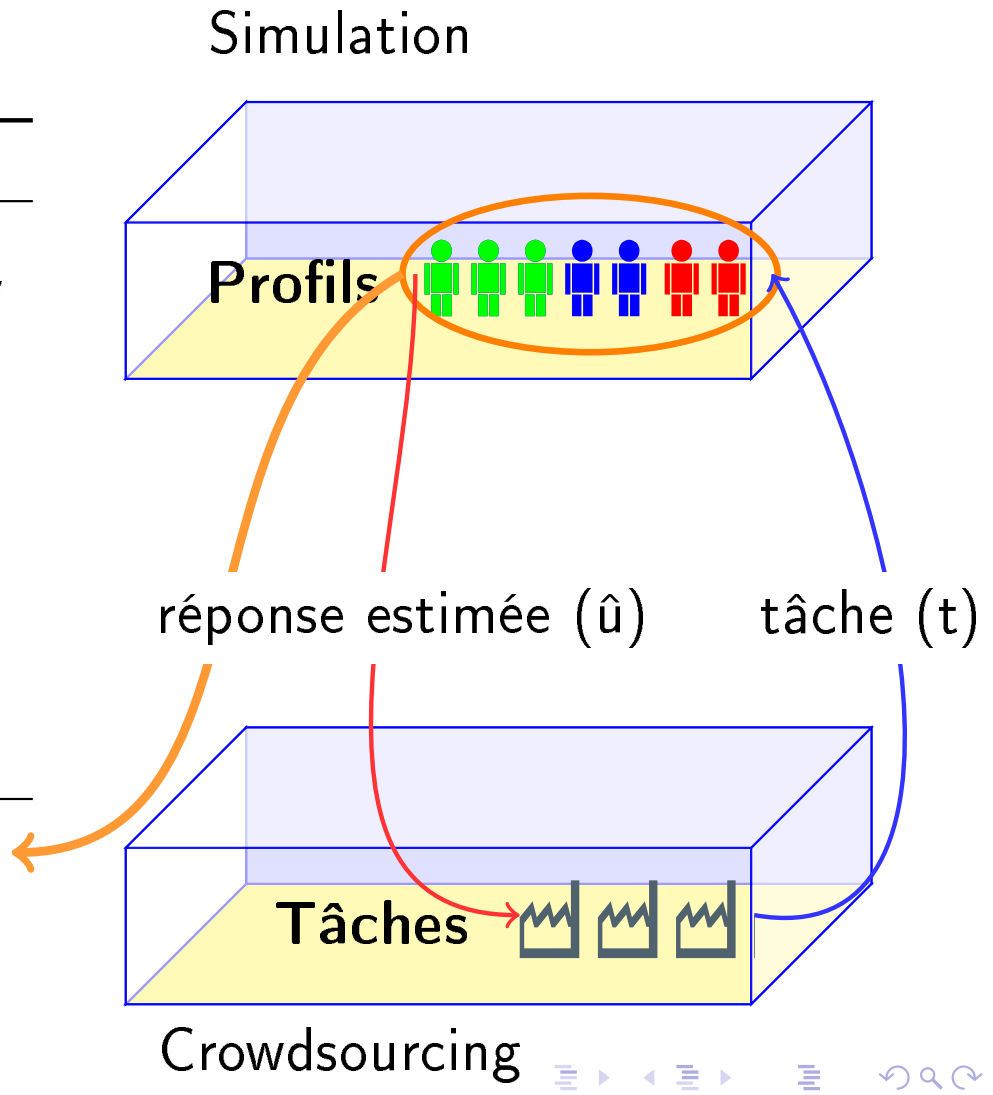
Algorithme Général

Algorithm 1 Simulation Naïf

Input : Ensemble des tâches T_i
Input : Ensemble des utilisateurs par profil P_i
Output : Ensemble des réponses estimées **U**

1: **for each** $T_i \in \{T_1, T_2, \ldots, T_n\}$ **do**
2: **for each** $P_i \in \{P_1, P_2, \ldots, P_m\}$ **do**
3: **for each** $u_i \in P_i$ **do**
4: $\hat{U}_i$ = réponse_estimated(T_i, u_i)
5: **end for**
6: **end for**
7: **end for**
8: **return** $\hat{U}_i \in \{\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n\}$

Contributions du travail :
Creation des profils réalistes



État de l'art

Matrice de confusion

Definition

La matrice de confusion est un outil servant à mesurer la qualité de réponse d'un utilisateur par rapport à la vraie réponse.

Réponse de l'utilisateur

La vraie réponse	c_1	$\dots$	c_j	$\dots$	c_n
c_1	p_{11}	$\dots$	c_{1j}	$\dots$	p_{1n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_i	0.01	$\dots$	0.98	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
c_n	p_{n1}	$\dots$	c_{nj}	$\dots$	p_{nn}

$\Rightarrow$ Il y a 98% de probabilité conditionnel de réponse d'utilisateur c_j lorsque la vrai réponse est c_i (i.e. $P(c_j \mid c_i) = 0.98$)

Matrice de confusion par utilisateur

Definition

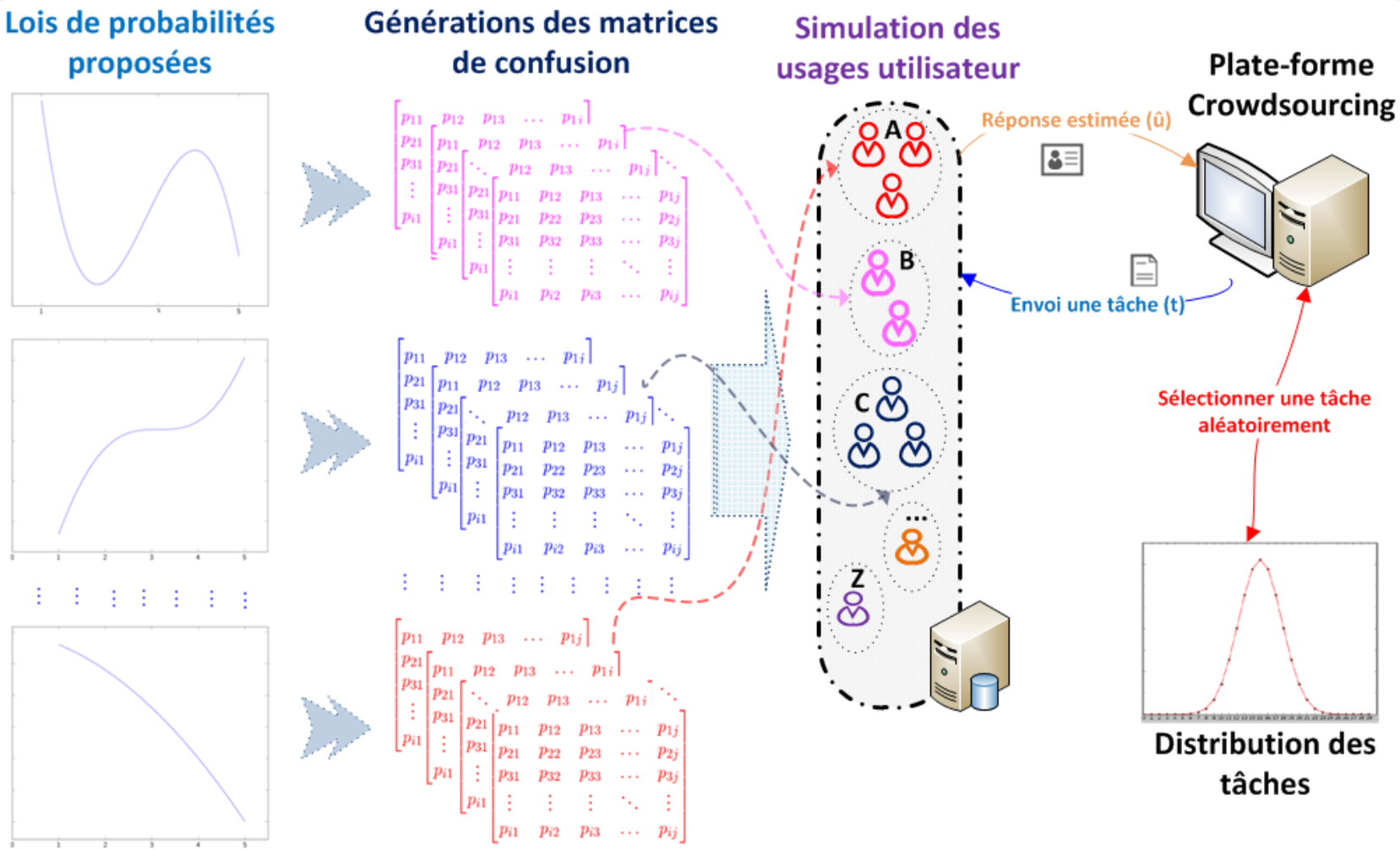
La qualité de réponse d'un utilisateur $u_k, k \in \{1, 2, 3, \ldots, K\}$ est représente par une matrice de confusion de N classes, $c_i\, i \in \{1, 2, 3, \ldots, N\}$ où chaque probabilité p_{ij} est la mesure de qualité qu'un utilisateur puisse estimées une classe j correctement lorsque la classe réelle est i. La représentation en probabilité conditionnelle est représentée $P(\text{réponse} = j \mid \text{vraie} = i) = p_{ij}$.

Réponse de l'utilisateur

$$\begin{array}{cc} & \begin{array}{ccccc} c_1 & c_2 & c_3 & \cdots & c_n \end{array} \\ \text{La vraie réponse} & \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix} \\ \begin{array}{c} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{array} & \end{array}$$

Approche Générale

Schéma des tâches à accomplir



Formulation du problème

Réponses estimées

Vraies réponses

$$\begin{array}{c|cccccc} & c_1 & c_2 & \dots & c_j & \dots & c_t \\ c_1 & p_{11} & p_{12} & \dots & p_{1j} & \dots & p_{1t} \\ c_2 & p_{21} & p_{22} & \dots & p_{2j} & \dots & p_{2t} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_i & p_{i1} & p_{i2} & \dots & p_{ij} & \dots & p_{it} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_t & p_{t1} & pt2 & \dots & p_{t2} & \dots & p_{tt} \end{array} \quad c_i$$

$(T * T)$

$$\begin{array}{cccccc} c_1 & c_2 & \dots & c_j & \dots & c_t \\ \left[\hat{u}_{i1} \right. & \hat{u}_{i2} & \dots & \hat{u}_{ij} & \dots & \left. \hat{u}_{it}\right] \end{array}$$

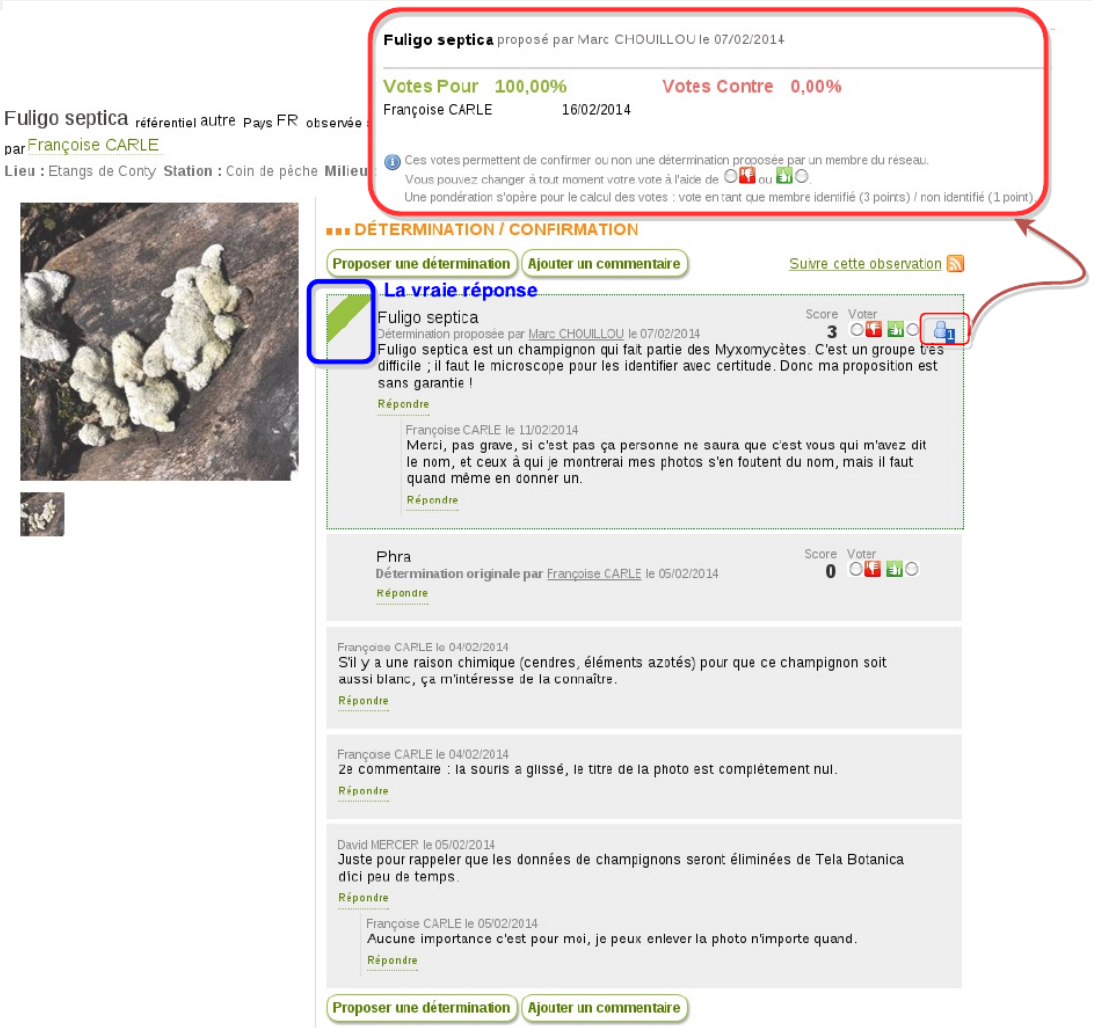
Donc, si l'indice i est la vrai reponse et $\forall j \in \{1, 2, 3, \dots T\}$

$$\sum_{T}^{j=1} \mathbf{Pr}\left[\mathbf{c_j} \,\middle|\, \mathbf{c_i}\right] = \sum_{T}^{j=1} \hat{\mathbf{u}}_{\mathbf{i,j}} = \mathbf{1}$$

Nous pouvons conclure que chaque ligne de la matrice exprime une loi de probabilités discrète.

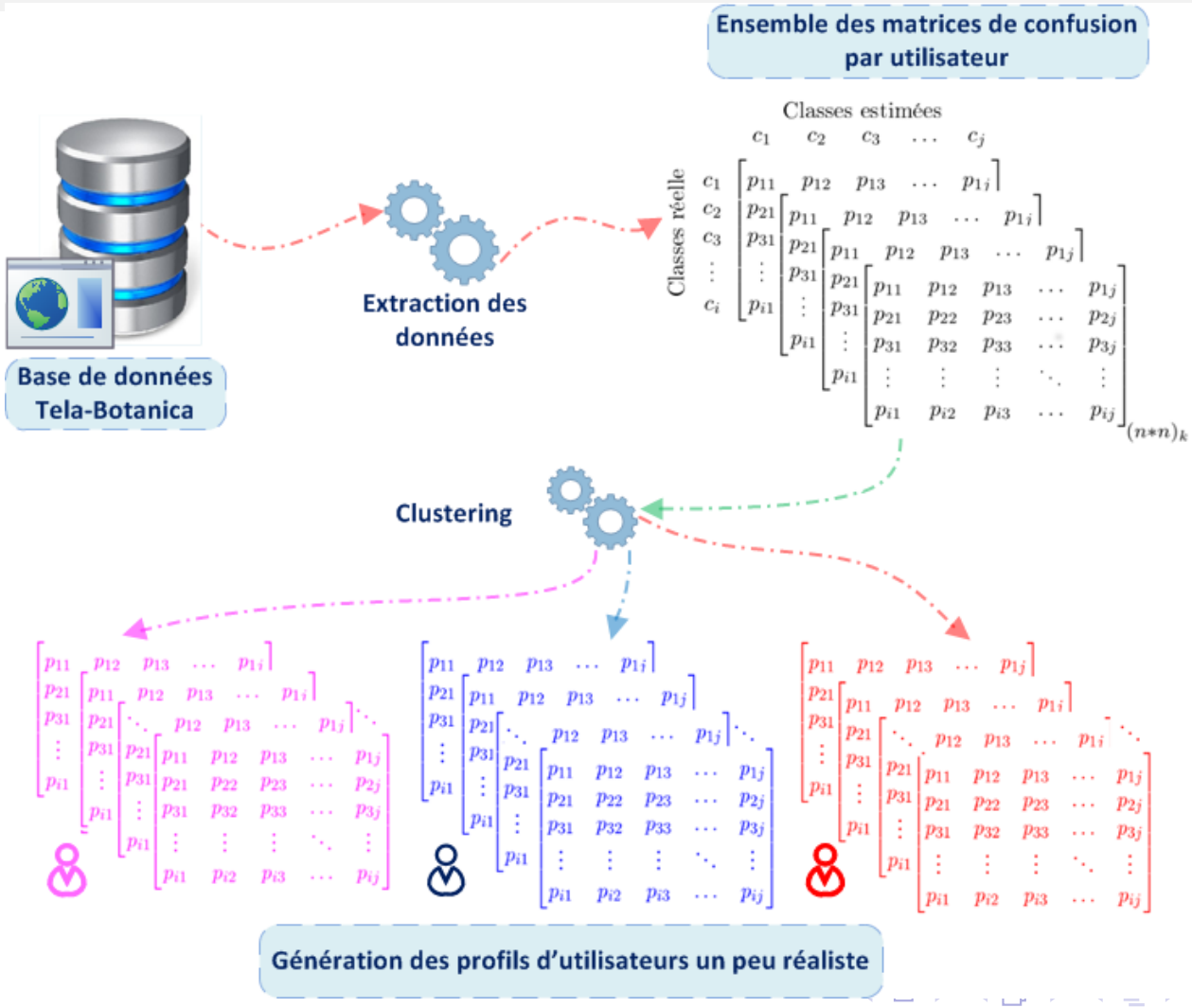
Approche Réaliste

Extraction des données du site web Tela-botanica



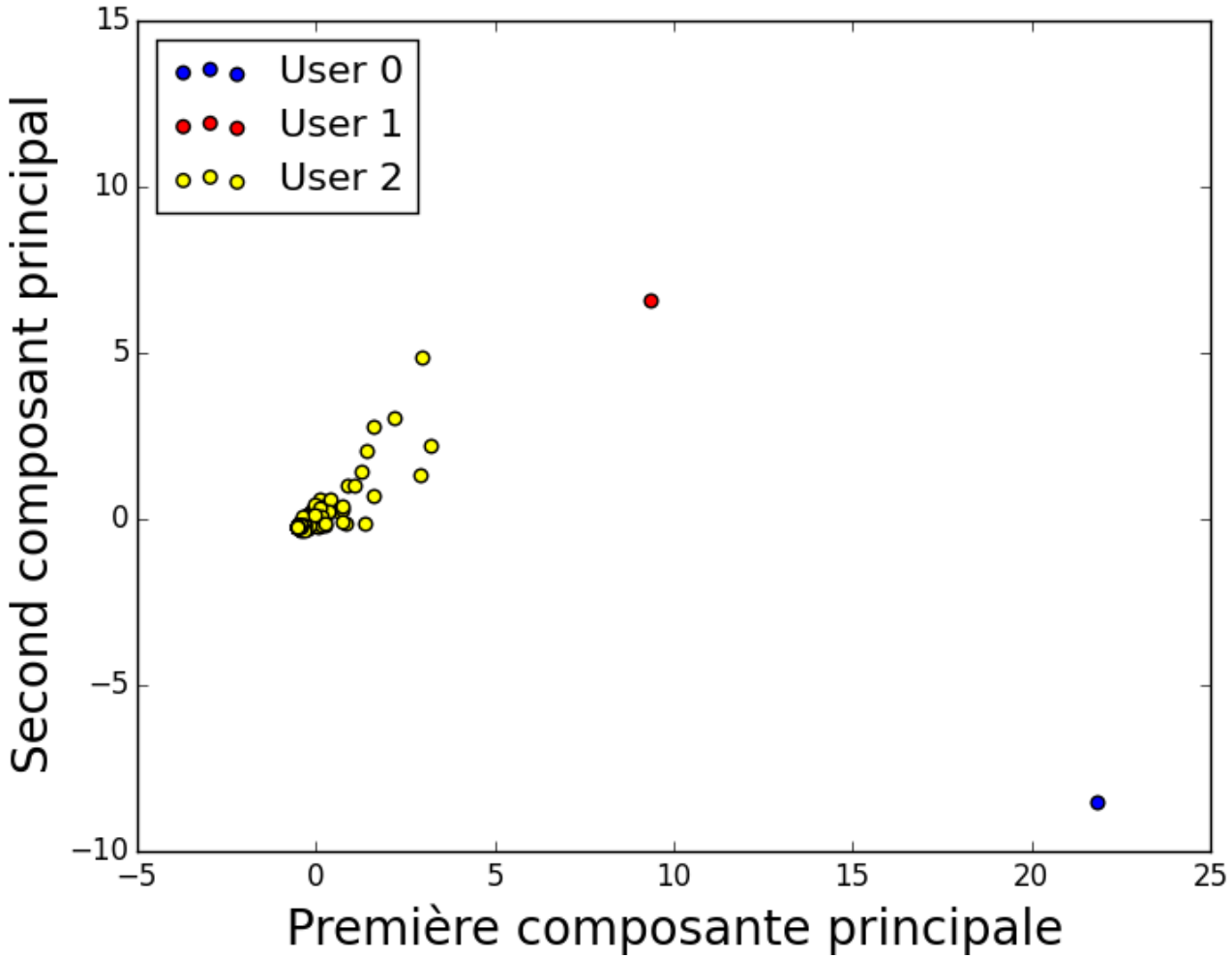
Approche Réaliste

Extraction des données du site web Tela-botanica



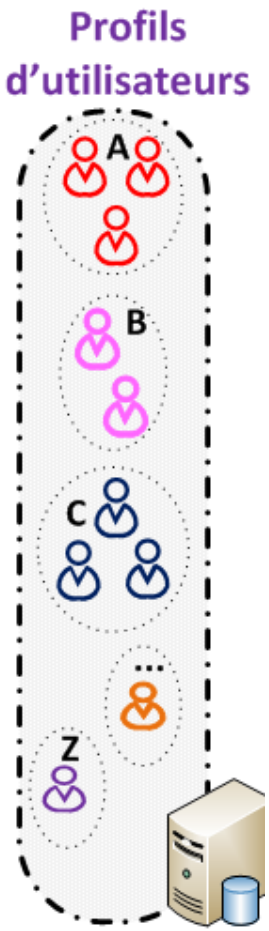
Approche Réaliste

Visualisation des utilisateurs de Tela-Botanica



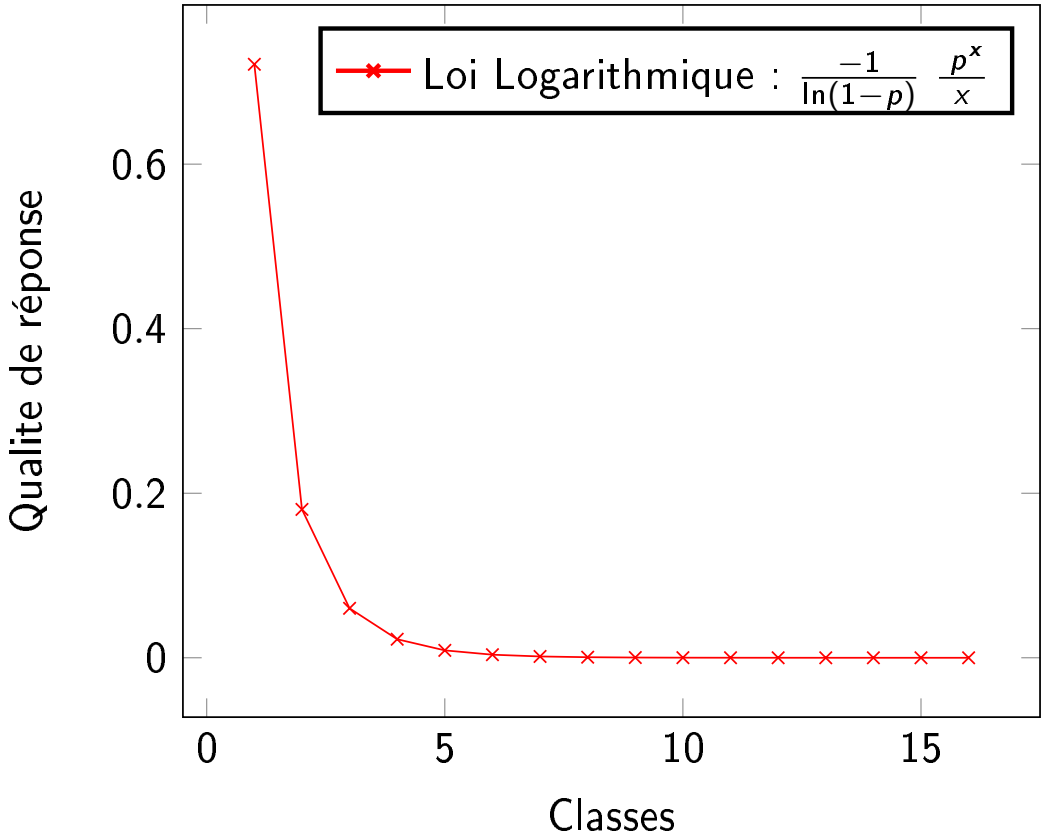
Profils d'utilisateurs proposés

1. Profil expert
2. Profil amateur
3. Profil novice
4. Profil spammeur



Profils d'utilisateurs proposés

Exemple : Profil expert

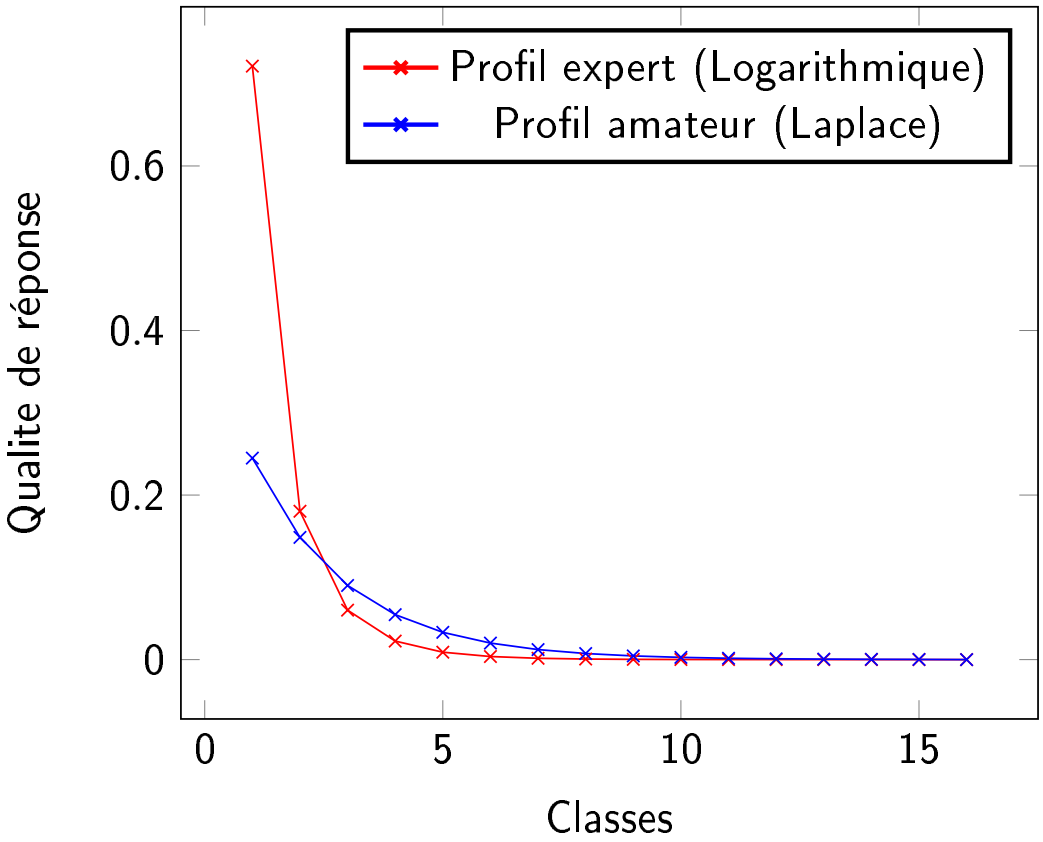


$$\begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_{16} \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & \cdots & c_{16} \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1(16)} \\ p_{21} & p_{22} & \cdots & p_{2(16)} \\ p_{31} & p_{32} & \cdots & p_{3(16)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{(16)1} & p_{(16)2} & \cdots & p_{(16)(16)} \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ c_1 \end{array} \begin{array}{c} \begin{array}{cccccc} c_1 & c_2 & c_3 & \cdots & c_{16} \end{array} \\ \begin{bmatrix} 0.68 & 0.19 & 0.06 & \cdots & 0.0 \end{bmatrix} \end{array}$$

Profils d'utilisateurs proposés

Exemple : Profil expert/amateur



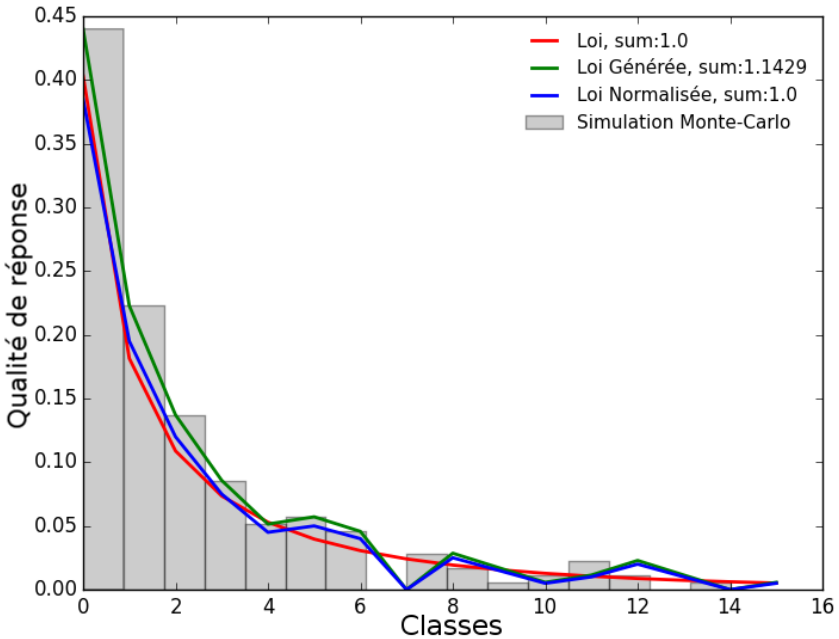
$$c_1 \begin{bmatrix} c_1 & c_2 & c_3 & \dots & c_{16} \\ 0.68 & 0.19 & 0.06 & \dots & 0.0 \end{bmatrix}$$

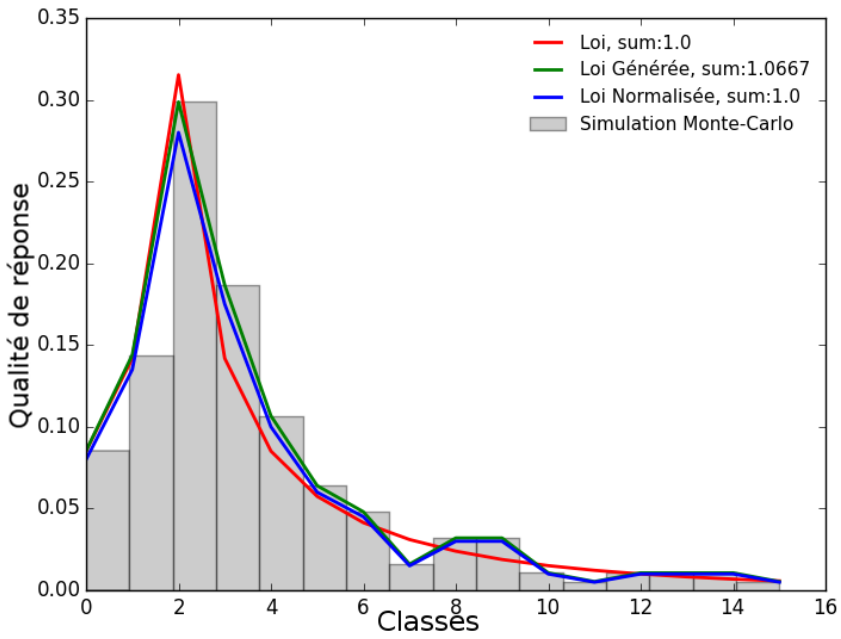
$$c_1 \begin{bmatrix} c_1 & c_2 & c_3 & \dots & c_{16} \\ 0.24 & 0.18 & 0.09 & \dots & 0.0 \end{bmatrix}$$

Simulation de Monte-Carlo

Méthode de transformation inverse

Étant donné la fonction inverse F^{-1} de la fonction de répartition F_{expert} et une variable U de loi uniforme $U_{[0-1]}$, alors $Z = F^{-1}(U)$ est distribuée suivante F et l'histogramme Z génère la loi de probabilité réaliste.

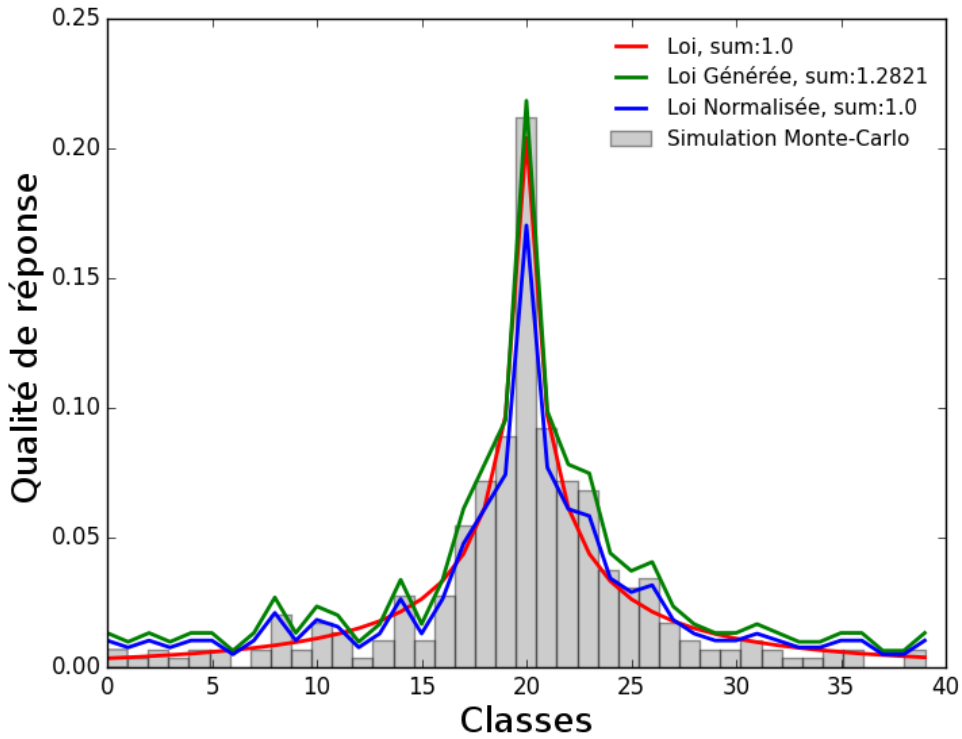




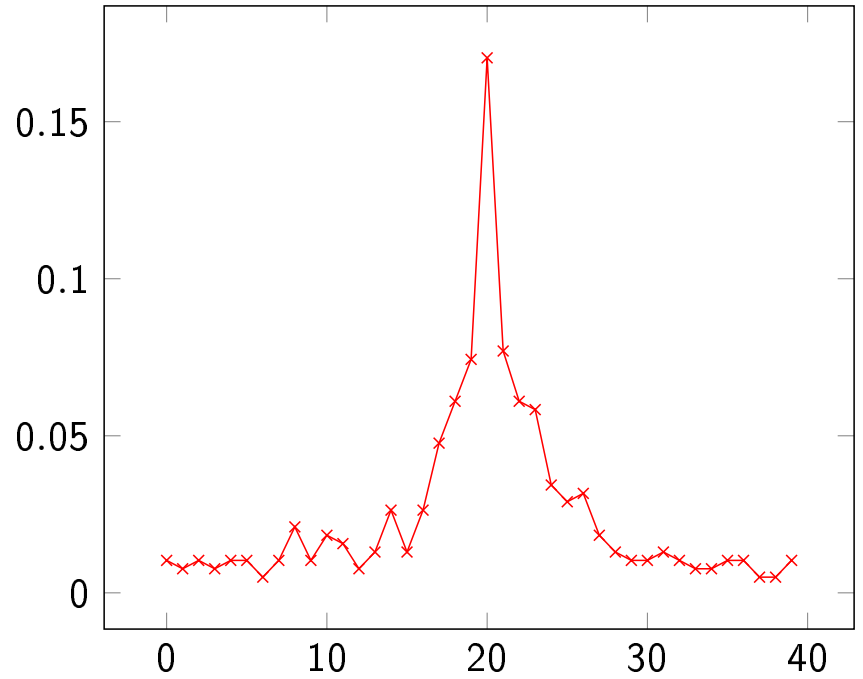
Génération des probabilités de la classe 20

Exemple

Calcule les probabilités de réponses de l'utilisateur par rapport à la vraie classe N° 20.



⇒



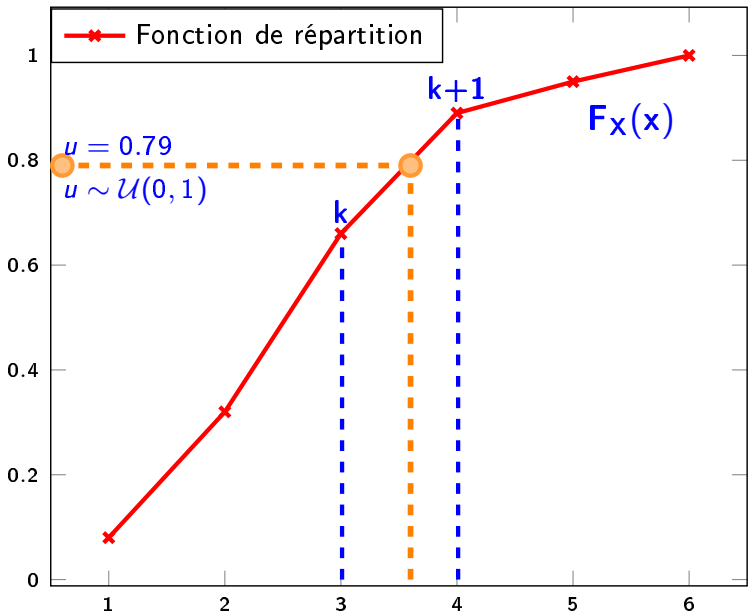
Simulation de réponse d'utilisateur

Vraies réponse	c_1	c_2	c_3	c_4	c_5	c_6
c_1	0.4648	0.1203	0.0195	0.0083	0.008	0.0128
c_2	0.3819	0.4097	0.26	0.0561	0.0068	0.0299
c_3	0.099	0.3354	0.2864	0.2384	0.0645	0.0716
c_4	0.0239	0.0949	0.3206	0.4488	0.1906	0.0709
c_5	0.0248	0.0245	0.0845	0.1817	0.6119	0.1609
c_6	0.0056	0.0152	0.0292	0.0667	0.1183	0.6538

Exemple :

Étant donné la fonction de répartition $F_X(x)$ de la vraie classe N° 3 de la matrice de confusion ci-dessus et un nombre aléatoire uniforme $u = 0.79$ (i.e $u \sim \mathcal{U} \in [0, 1]$).

Ainsi donc, dans notre exemple k est égal à 3.



Plan

Modélisation des usages utilisateurs à grande échelle

Validation des profils

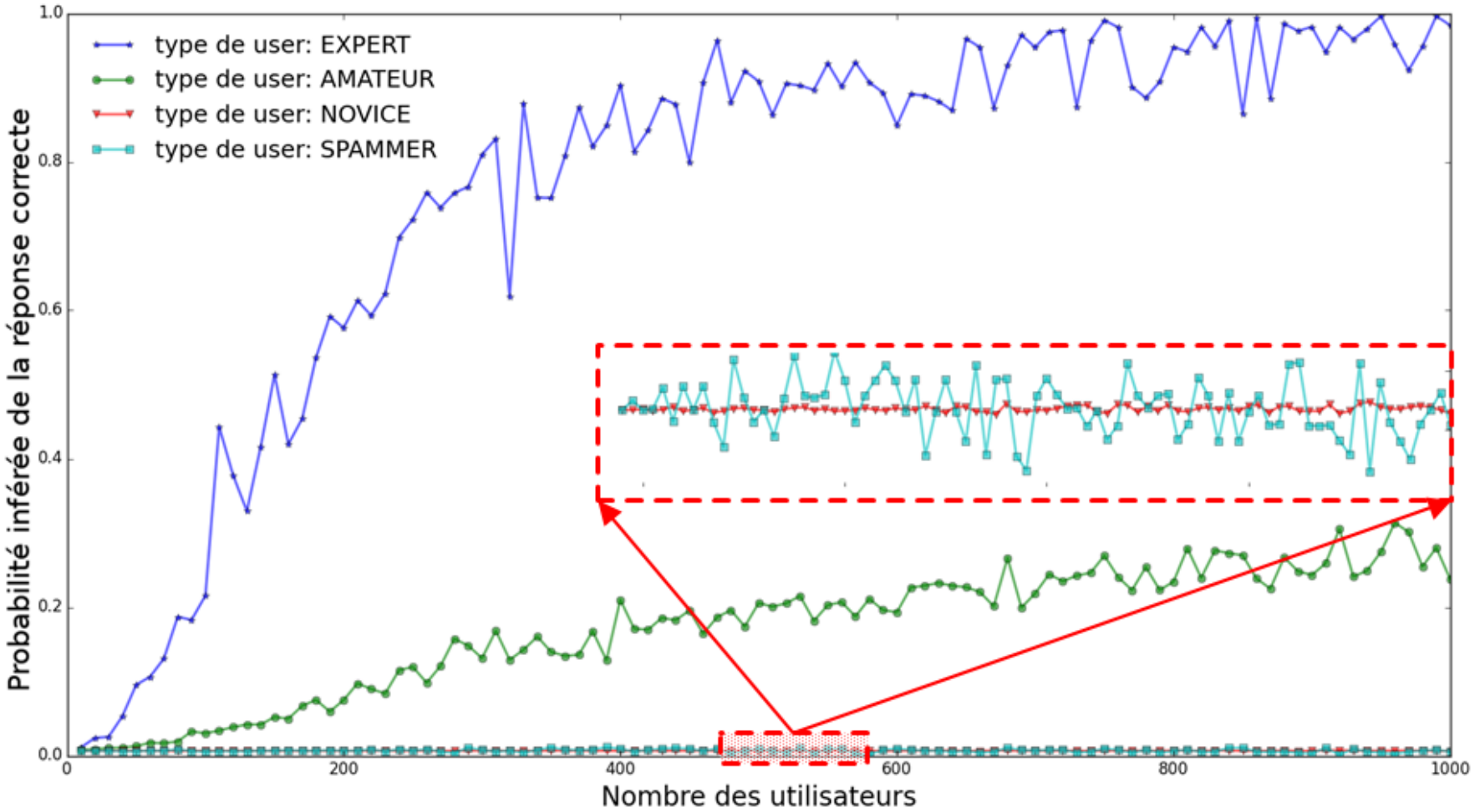
Configuration

Profils	Nb. Users	Nb. Classes	Nb. Tâches
Experts	$[10, 20, ..., 1000]$	150	50
Amateurs	$[10, 20, ..., 1000]$	150	50
Novices	$[10, 20, ..., 1000]$	150	50
Spammers	$[10, 20, ..., 1000]$	150	50

Tab.: Configuration des profils pour la simulation

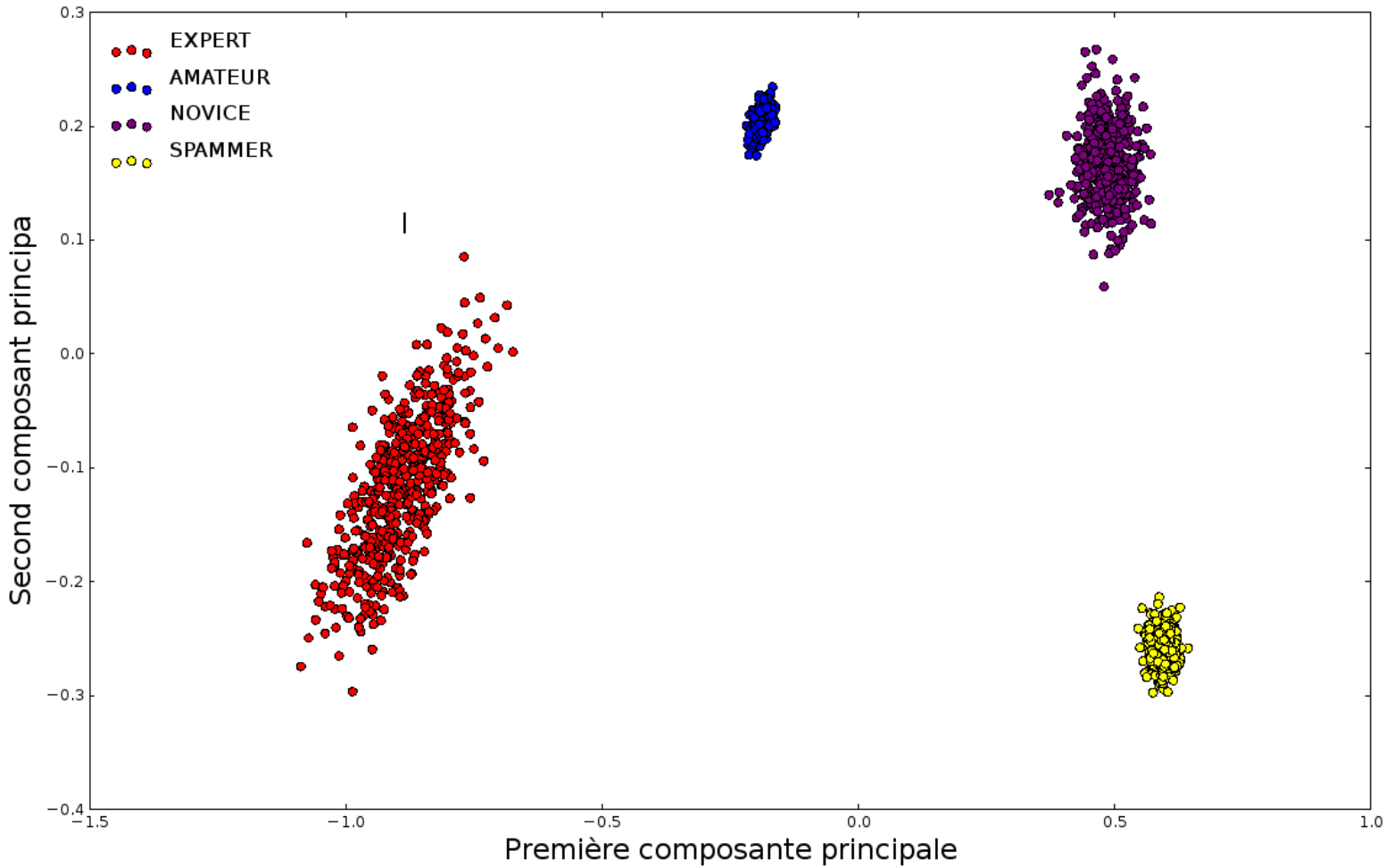
Validation des profils

Méthode d'inférence de Dawid et Skene



Validation des profils

Visualisation des utilisateurs par profil



Évaluation des solutions Crowdsourcing

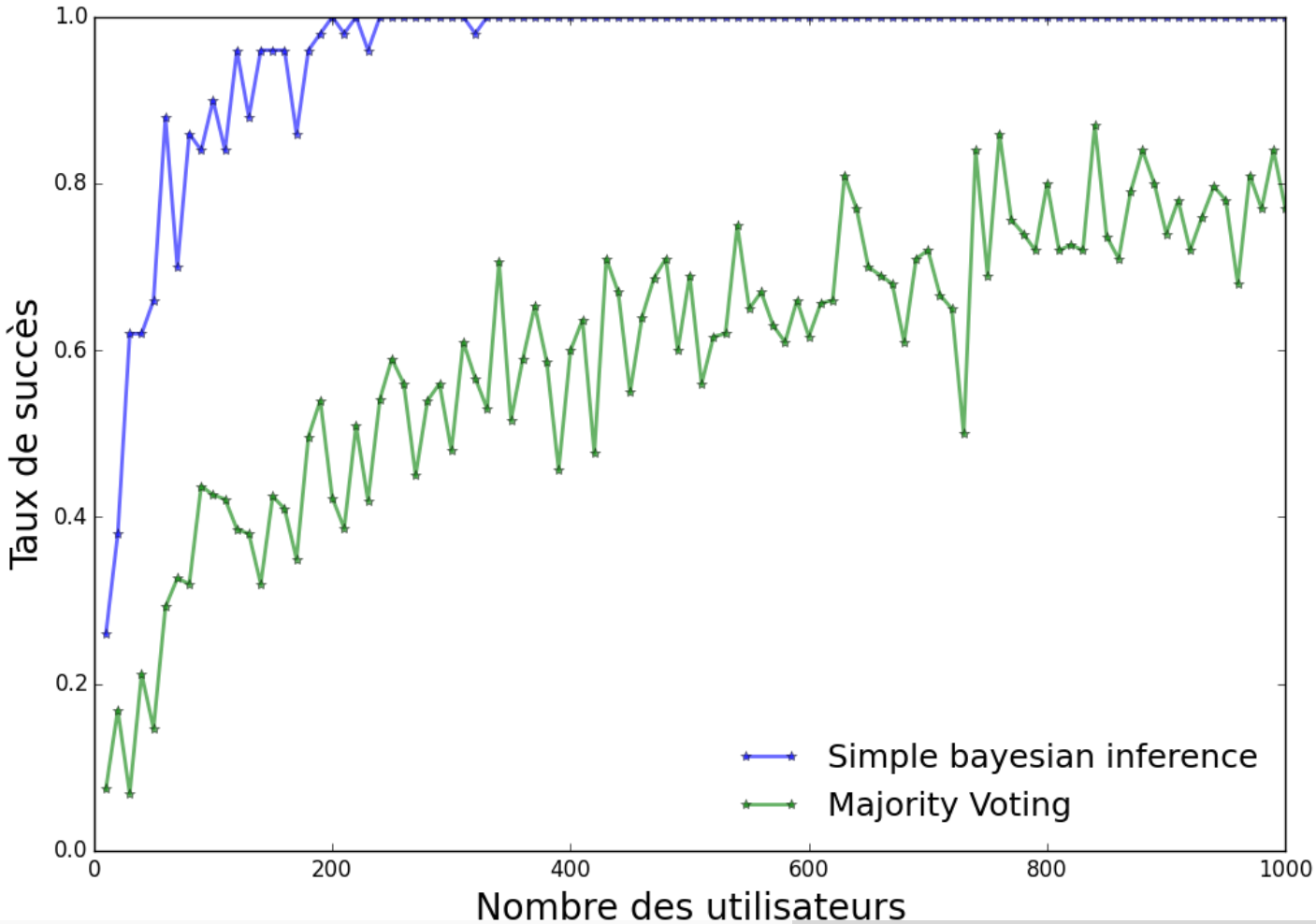
Configuration

Profils	Nb. Users	Nb. Classes	Nb. Tâches
Experts	20% * 1000	150	50
Amateurs	30% * 1000	150	50
Novices	30% * 1000	150	50
Spammers	20% * 1000	150	50

Tab.: Configuration des profils utilisateurs

Évaluation des solutions Crowdsourcing

Résultats



Conclusions et Ouvertures

1. Mise en pratique de la matrice de confusion en tant que connaissance d'un utilisateur dans un problème de classification.
2. Comparaison d'autres méthodes d'inférence non vues afin de valider notre approche (i.e. les profils).
3. Manipulation de certains propriétés des lois de probabilités proposées afin de trouver autres profils.
4. Mise en œuvre d'un modèle d'apprentissage d'un utilisateur au fur et à mesurer qu'il répond aux tâches, afin d'avoir les compétences des utilisateur qui évoluent au cours du temps.

Plan

Modélisation des usages utilisateurs à grande échelle

References


A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 20–28, 1979.


V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Aug. 2010.

Merci de votre attention.