

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



UNIVERSITÉ
DE MONTPELLIER



11 juin 2015

Plan

Modélisation des usages utilisateurs à grande échelle

1. Introduction
 - Contexte
 - Problématiques et Objectif
2. Modélisation
 - Algorithme Général
 - État de l'art
 - Approche Générale
 - Méthode de simulation
3. Résultats et Conclusions
 - Résultats expérimentales
 - Conclusions
4. Références

Plan

Modélisation des usages utilisateurs à grande échelle

1. Introduction
 - Contexte
 - Problématiques et Objectif
2. Modélisation
 - Algorithme Général
 - État de l'art
 - Approche Générale
 - Méthode de simulation
3. Résultats et Conclusions
 - Résultats expérimentales
 - Conclusions
4. Références

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**.

Surigao del Sur: relief goods infant needs #pabloPH -
<http://t.co/fpEEyFwO> #ReliefPH

- Picture: link points to picture(s) of general damage/flooding
- Video: link points to video(s) of general damage/flooding
- Location: tweet/picture/video includes reference to location, place, town etc.
- Other: link does not point to picture(s) of general damage/flooding

Surigao del Sur

Search



Approximate damage/impact location: Longitude: -, Latitude: -

Navigation controls: + Navigate, + Add damage/impact marker

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.



Véronique filiforme

Veronica filiformis Sm.



★ Classification des plantes

(1) Marguerite

(2) Zygopétalum

⋮

(n) Véronique
filiforme

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

Tâche

CARRANZA ALARCON, Yonatan Carlos

Modélisation des usages utilisateurs

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) solution(s) allons-nous proposer

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

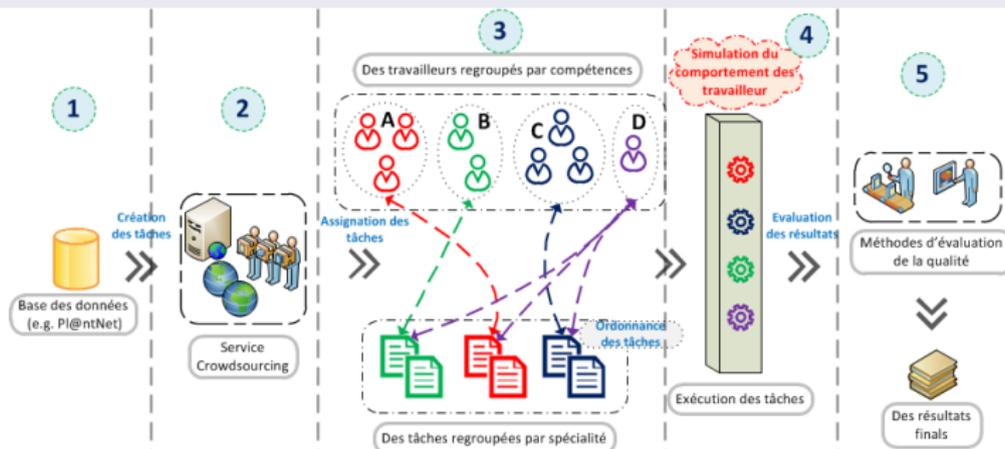


Fig.: Schéma du Crowdsourcing

Plan

Modélisation des usages utilisateurs à grande échelle

1. Introduction
 - Contexte
 - Problématiques et Objectif
2. Modélisation
 - Algorithme Général
 - État de l'art
 - Approche Générale
 - Méthode de simulation
3. Résultats et Conclusions
 - Résultats expérimentales
 - Conclusions
4. Références

Algorithme Général

Algorithm 1 Simulation Naïf

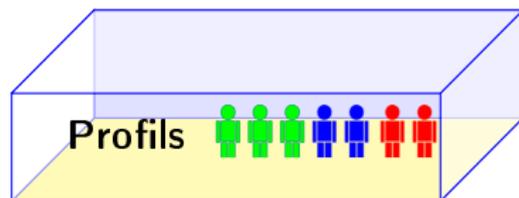
Input : Ensemble des tâches T_i

Input : Ensemble des utilisateurs par profil P_i

Output : Ensemble des réponses estimées \mathbf{U}

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

Simulation



Crowdsourcing

Algorithme Général

Algorithm 1 Simulation Naïf

Input : Ensemble des tâches T_i

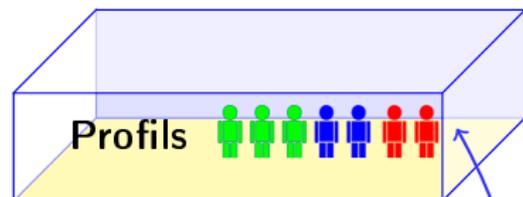
Input : Ensemble des utilisateurs par profil P_i

Output : Ensemble des réponses estimées \mathbf{U}

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

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

Simulation



tâche (t)



Crowdsourcing

Algorithme Général

Algorithm 1 Simulation Naïf

Input : Ensemble des tâches T_i

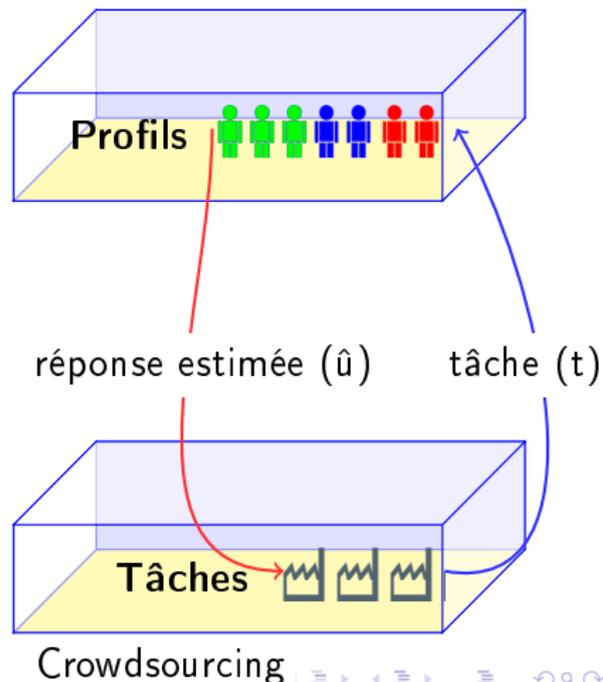
Input : Ensemble des utilisateurs par profil P_i

Output : Ensemble des réponses estimées \mathbf{U}

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

Le système de simulation reçoit la tâche (t), la traite, puis donne une réponse estimée \hat{u} .

Simulation



Algorithme Général

Algorithm 1 Simulation Naïf

Input : Ensemble des tâches T_i

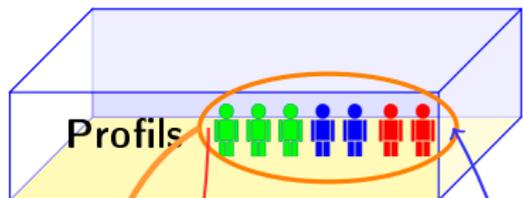
Input : Ensemble des utilisateurs par profil P_i

Output : Ensemble des réponses estimées U

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

Contributions du travail :
Creation des profils réalistes

Simulation



réponse estimée (\hat{u})

tâche (t)



Crowdsourcing

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

| | | | | | | | |
|------------------|-------|-------|----------|-------|----------|-------|----------|
| | | c_1 | ... | c_j | ... | c_n | |
| La vraie réponse | c_1 | [| p_{11} | ... | c_{1j} | ... | p_{1n} |
| | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | c_i | | 0.01 | ... | 0.98 | ... | 0 |
| | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | c_n | | p_{n1} | ... | c_{nj} | ... | p_{nn} |

⇒ Il y a 98% de probabilité conditionnel de réponse d'utilisateur c_j lorsque la vraie réponse est c_i ; (i.e. $P(c_j | c_i) = 0.98$)

Matrice de confusion par utilisateur

Definition

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

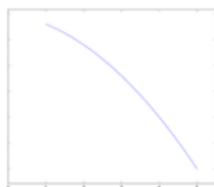
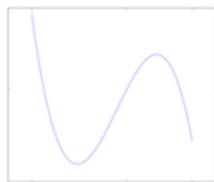
Réponse de l'utilisateur

| | | c_1 | c_2 | c_3 | \dots | c_n |
|------------------|----------|----------|----------|----------|----------|----------|
| La vraie réponse | c_1 | p_{11} | p_{12} | p_{13} | \dots | p_{1n} |
| | c_2 | p_{21} | p_{22} | p_{23} | \dots | p_{2n} |
| | c_3 | p_{31} | p_{32} | p_{33} | \dots | p_{3n} |
| | \vdots | \vdots | \vdots | \vdots | \ddots | \vdots |
| | c_n | p_{n1} | p_{n2} | p_{n3} | \dots | p_{nn} |

Approche Générale

Schéma des tâches à accomplir

Lois de probabilités proposées



Génération des matrices de confusion



Simulation des usages utilisateur

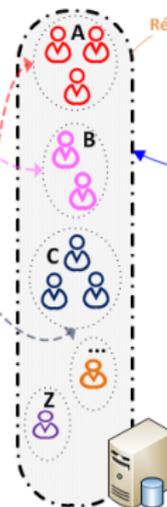


Plate-forme Crowdsourcing



Sélectionner une tâche aléatoirement



Distribution des tâches

Formulation du problème

Vraies réponses

| | | Réponses estimées | | | | | |
|-----------------|----------|-------------------|----------|----------|----------|----------|----------|
| | | c_1 | c_2 | ... | c_j | ... | c_t |
| Vraies réponses | c_1 | p_{11} | p_{12} | ... | p_{1j} | ... | p_{1t} |
| | c_2 | p_{21} | p_{22} | ... | p_{2j} | ... | p_{2t} |
| | \vdots | \vdots | \vdots | \vdots | \vdots | \ddots | \vdots |
| | c_i | p_{i1} | p_{i2} | ... | p_{ij} | ... | p_{it} |
| | \vdots | \vdots | \vdots | \vdots | \vdots | \ddots | \vdots |
| | c_t | p_{t1} | p_{t2} | ... | p_{tj} | ... | p_{tt} |

c_i

| | | c_1 | c_2 | ... | c_j | ... | c_t |
|--|--|----------------|----------------|-----|----------------|-----|----------------|
| | | \hat{u}_{i1} | \hat{u}_{i2} | ... | \hat{u}_{ij} | ... | \hat{u}_{it} |

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

$$\sum_{T}^{j=1} \Pr [c_j | c_i] = \sum_{T}^{j=1} \hat{u}_{ij} = 1$$

($T \times T$)

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

Fuligo septica réitérériel autre Pays FR observée
par Françoise CARLE
Lieu : Elangs de Conry Station : Coin de pêche Millieu :



Fuligo septica proposée par Marc CHOULLOU le 07/02/2014

Votes Pour 100,00% Votes Contre 0,00%
Françoise CARLE 16/02/2014

Ces votes permettent de confirmer ou non une détermination proposée par un membre du réseau.
Vous pouvez changer à tout moment votre vote à l'aide de ou .
Une pénalisation s'opère pour le calcul des votes : vote en tant que membre identifié (3 points) / non identifié (1 point).

*** DÉTERMINATION / CONFIRMATION

Proposer une détermination Ajouter un commentaire Suivre cette observation

La vraie réponse

Fuligo septica Score: 3 Votes:
Observation proposée par Marc CHOULLOU le 07/02/2014
Fuligo septica est un champignon qui fait partie des Myxomycètes. C'est un groupe très difficile : il faut le microscope pour les identifier avec certitude. Donc ma proposition est sans garantie !

Répondre

Françoise CARLE le 11/02/2014

Merci, pas grave, si c'est pas ça personne ne saura ce c'est vous qui m'avez dit le nom, et ceux à qui je montrerais mes photos s'n foutent du nom, mais il faut quand même en connaitre un.

Répondre

Phra Détermination originale par Françoise CARLE le 05/02/2014
Répondre

Françoise CARLE le 01/02/2014

S'il y a une raison chimique (cendres, éléments azotés) pour que ce champignon soit aussi blanc, ça m'intéresse de la connaître.

Répondre

Françoise CARLE le 04/02/2014

2e commentaire : la soucs a gissé, le titre de la photo est complètement nul.

Répondre

David HERCOP le 05/02/2014

Juste pour rappeler que les données de champignons seront éliminées de Tela Botanica d'ici peu de temps.

Répondre

Françoise CARLE le 05/02/2014

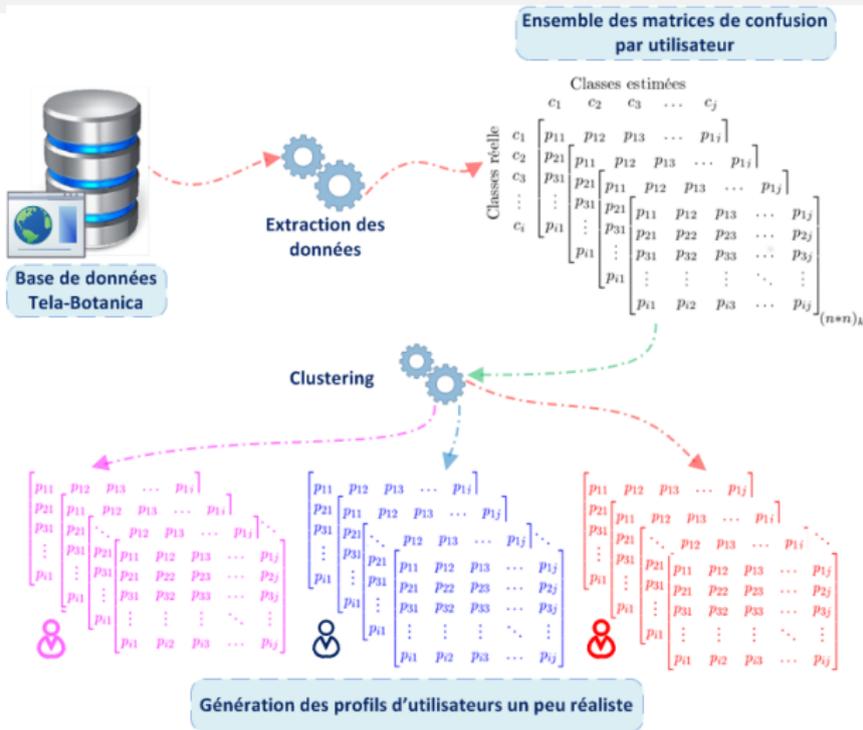
Aucune importance c'est pour moi, je peux enlever la photo n'importe quand.

Répondre

Proposer une détermination Ajouter un commentaire

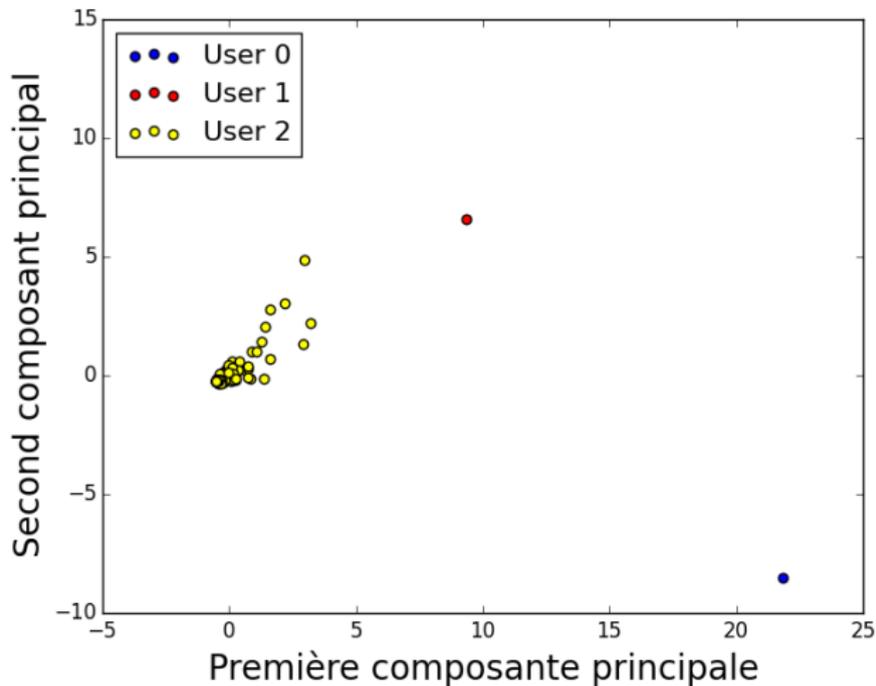
Approche Réaliste

Extraction des données du site web Tela-botanica



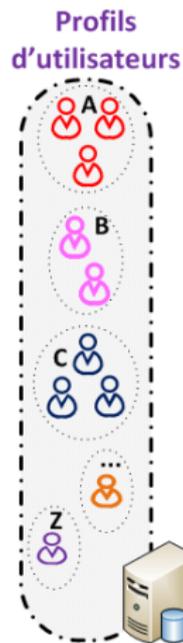
Approche Réaliste

Visualisation des utilisateurs de Tela-Botanica



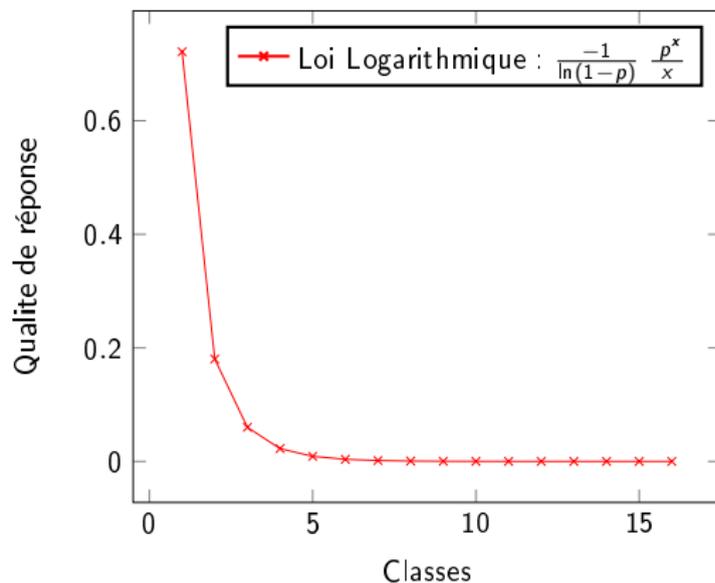
Profils d'utilisateurs proposés

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



Profils d'utilisateurs proposés

Exemple : Profil expert

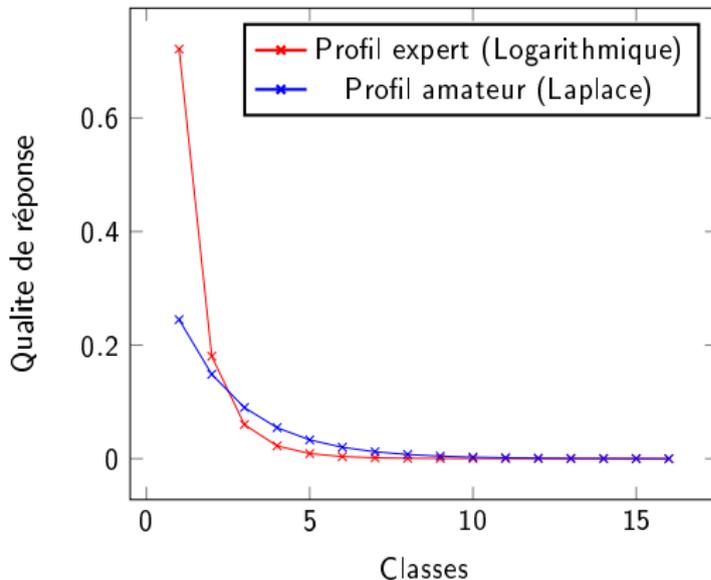


| | C_1 | C_2 | ... | C_{16} |
|----------|-------------|-------------|-----|----------------|
| C_1 | p_{11} | p_{12} | ... | $p_{1(16)}$ |
| C_2 | p_{21} | p_{22} | ... | $p_{2(16)}$ |
| C_3 | p_{31} | p_{32} | ... | $p_{3(16)}$ |
| ... | ... | ... | ... | ... |
| C_{16} | $p_{(16)1}$ | $p_{(16)2}$ | ... | $p_{(16)(16)}$ |

| | C_1 | C_2 | C_3 | ... | C_{16} |
|-------|-------------|-------------|-------------|-----|------------|
| C_1 | 0.68 | 0.19 | 0.06 | ... | 0.0 |

Profils d'utilisateurs proposés

Exemple : Profil expert/amateur



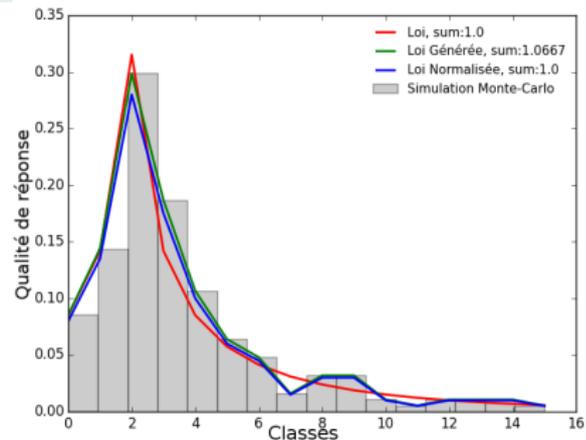
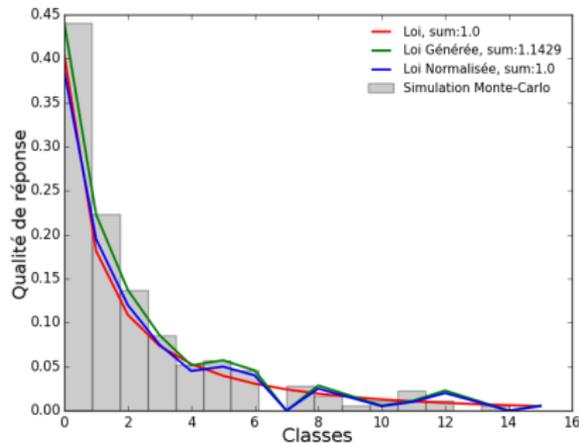
$$c_1 \begin{bmatrix} c_1 & c_2 & c_3 & \dots & c_{16} \\ 0.68 & 0.19 & 0.06 & \dots & 0.0 \end{bmatrix}$$

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

Simulation de Monte-Carlo

Méthode de transformation inverse

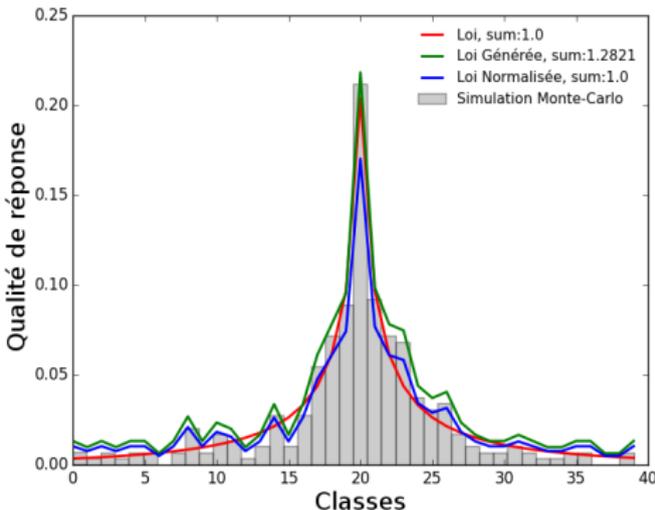
Étant donné la fonction inverse F^{-1} de la fonction de répartition F_{expert} et une variable U de loi uniforme $U_{[0-1]}$, alors $Z = F^{-1}(U)$ est distribuée suivante F et l'histogramme Z génère la loi de probabilité réaliste.



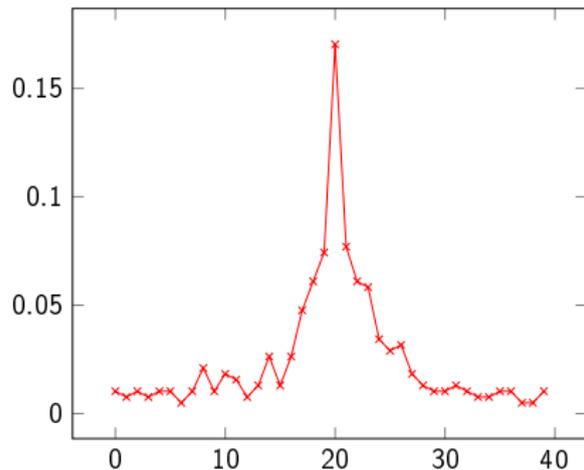
Génération des probabilités de la classe 20

Exemple

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



⇒



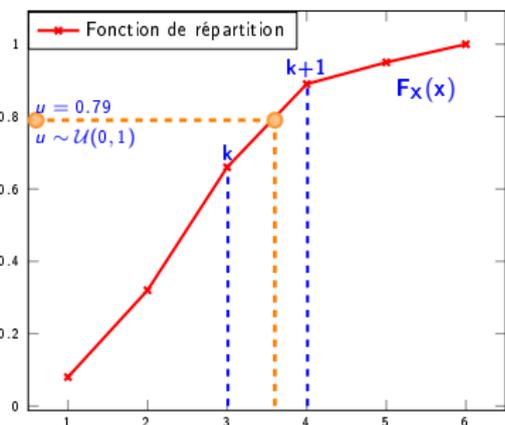
Simulation de réponse d'utilisateur

| | C_1 | C_2 | C_3 | C_4 | C_5 | C_6 |
|----------------------|--------|--------|--------|--------|--------|--------|
| Vraies réponse 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

1. Introduction
 - Contexte
 - Problématiques et Objectif
2. Modélisation
 - Algorithme Général
 - État de l'art
 - Approche Générale
 - Méthode de simulation
3. Résultats et Conclusions
 - Résultats expérimentales
 - Conclusions
4. Références

Validation des profils

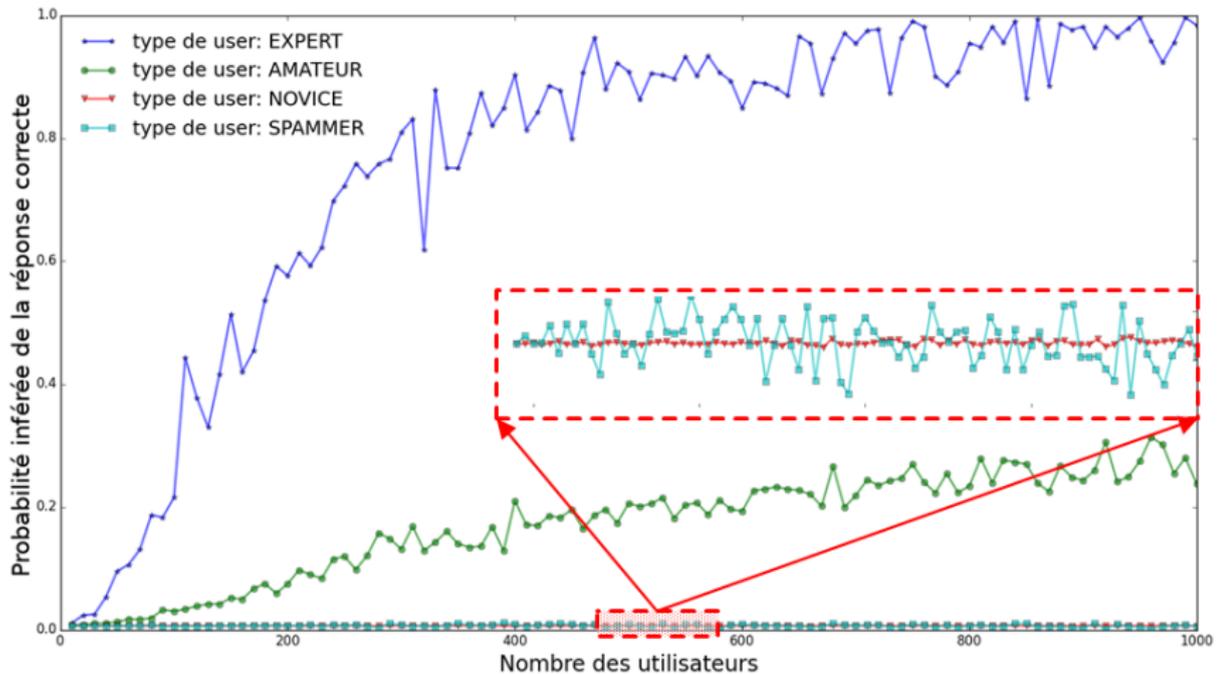
Configuration

| 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

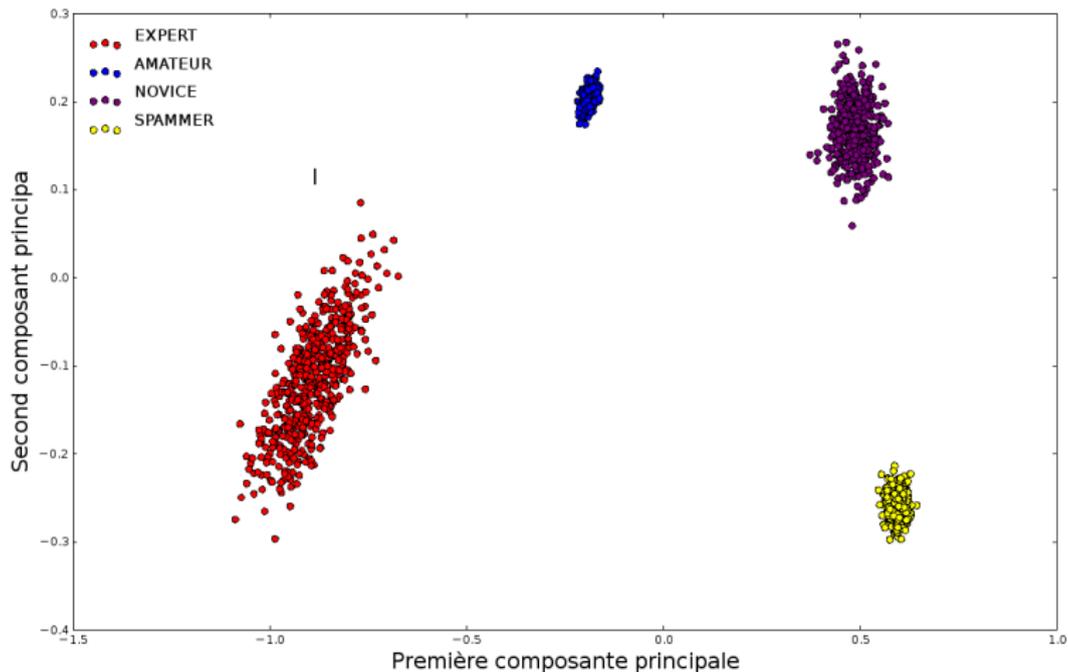
Validation des profils

Méthode d'inférence de Dawid et Skene



Validation des profils

Visualisation des utilisateurs par profil



Évaluation des solutions Crowdsourcing

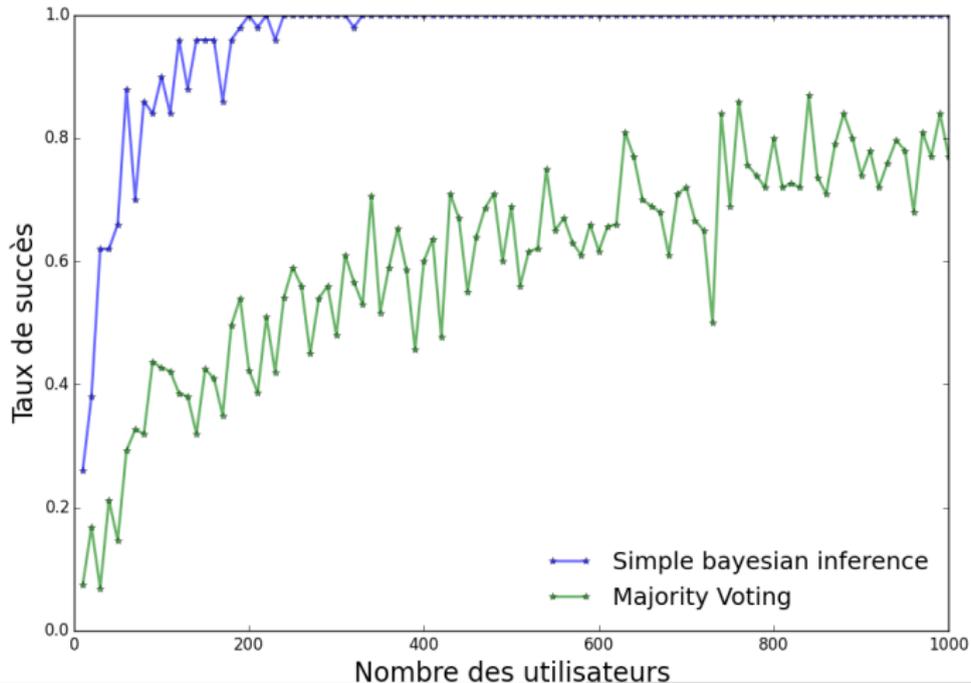
Configuration

| 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 certaines 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 mesure 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

1. Introduction
 - Contexte
 - Problématiques et Objectif
2. Modélisation
 - Algorithme Général
 - État de l'art
 - Approche Générale
 - Méthode de simulation
3. Résultats et Conclusions
 - Résultats expérimentales
 - Conclusions
4. **Références**

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.