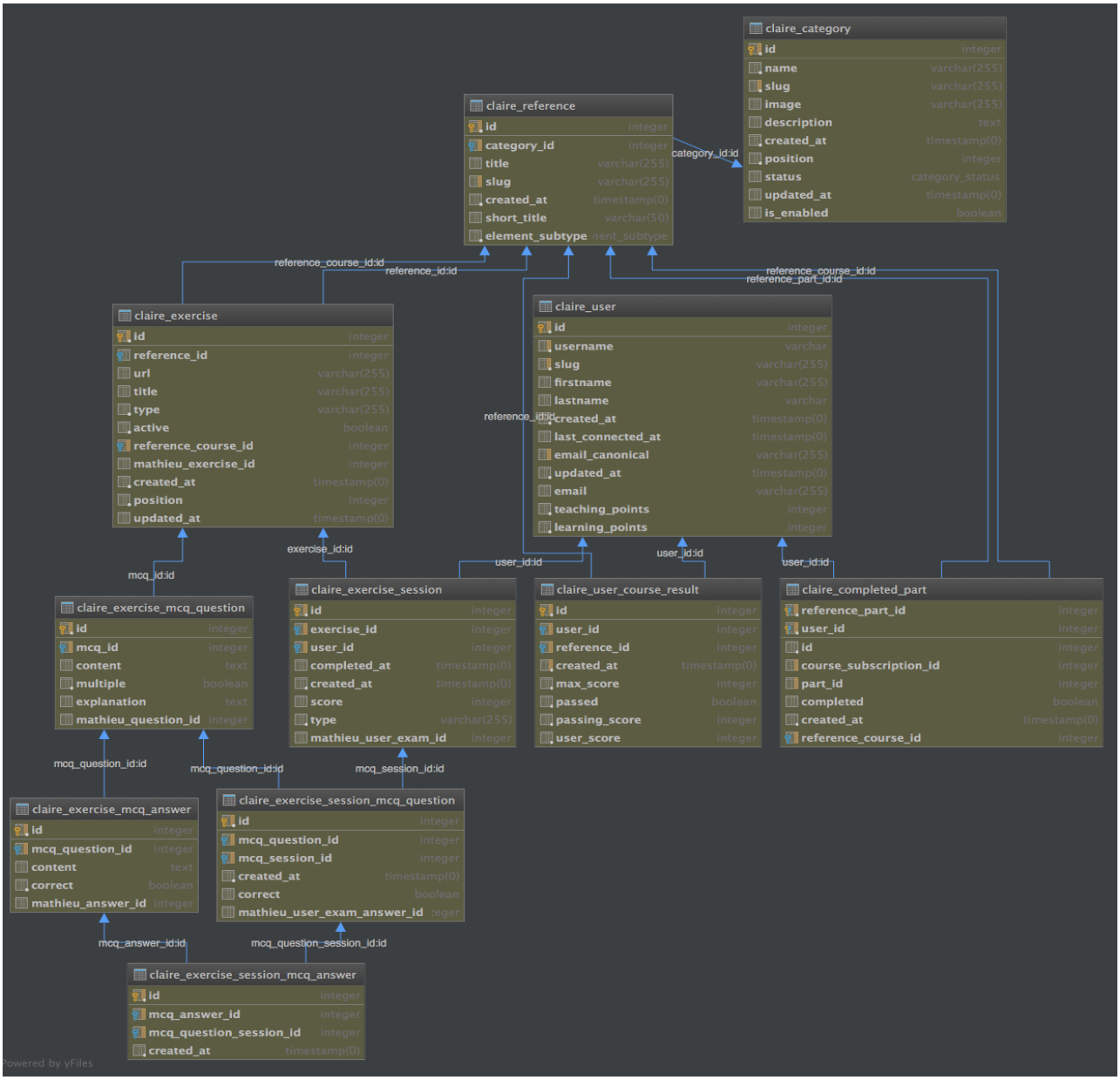
Description des traces

Traces avril 2016

Contenu des traces brutes

Cours = arbre d'éléments typés. Des références vers tout élément.

V0 : 29 septembre 2015
V1: 19-20 avril 2016
V2: 10 mai 2016
V3: 26 mai 2016
V4: Novembre 2016 (10 nouveaux cours plus utilisés et pertinents pour OC)



Remarques :

- les champs mathieu … correspondent à de la gestion de legacy.
- reference : peut être un cours, une partie, un chapitre …
- reference_course : reference de type cours.
- reference_part : reference de type part ou chapitre
- le reference_part de completed part n'est que sur des chapitres
- reference_id dans exercise : cours, partie ou chapitre (dans les fait uniquement partie, l'exercice vient se positionner à la fin du dernier chapitre de la partie)

Arborescence :

- Course
    - Partie
        - Chapitre

course_id : id de la référence du cours
part_id : id de la référence de la partie ou du chapitre

- exercise : partie
- sinon chapitre

Romain : A quoi correspond “session_id”?

Description

USER (Utilisateur)
-- string sdz_user.id
-- boolean sdz_user.active
-- datetime sdz_user.birthday
-- string sdz_user.city
-- string sdz_user.country
(c'est un champ libre, la data n'est pas toujours cohérente, exemple : France, fr, franc …)
-- string sdz_user.gender
-- string sdz_user.locale (langue de l'apprenant , par défaut langage du navigateur utilisé)
-- string sdz_user.region
-- string sdz_user.zip_code

USER_PREMIUM (date de premium, il peut y avoir plusieurs entrées par utilisateur)
-- int sdz_subscription.user_id
-- datetime sdz_subscription.started_at
-- datetime sdz_subscription.ended_at

USER_COURSE_VISUALISATION (trace de visualisation)
-- integer oc_course_visualisation.course_id
-- integer|null oc_course_visualisation.part_id
(si null, visualisation de la page d'accueil du cours seulement, sinon visualisation du chapitre)
-- integer oc_course_visualisation.session_id
-- datetime oc_course_visualisation.date
-- integer|null oc_course_visualisation.user_id *(null = visiteur anonyme)*

USER_FOLLOW_COURSE (**event** de l'action de suivre un cours par l'apprenant)
-- int user_follow_course.course_id,
-- datetime user_follow_course.created_at,
-- int user_follow_course.user_id

USER_UNFOLLOW_COURSE (**event** de l'action de ne plus suivre un cours par l'apprenant)
-- int user_unfollow_course.course_id,
-- datetime user_unfollow_course.created_at,
-- int user_unfollow_course.user_id

COMPLETED_PART (**event** du mark as complete)
-- integer claire_completed_part.part_id,
-- boolean claire_completed_part.completed,
-- datetime claire_completed_part.created_at,
-- integer claire_completed_part.user_id

EXERCISE

-- integer claire_exercise.id,
-- integer claire_exercise.reference_course_id,
-- integer claire_exercise.reference_id, (chapitre)
-- string claire_exercise.type,
-- boolean claire_exercise.active,
-- integer claire_exercise.position

USER_EXERCISE_SESSION (session d'exercice)

-- integer claire_exercise_session.id
-- integer claire_exercise_session.exercise_id
-- datetime claire_exercise_session.created_at
-- datetime claire_exercise_session.completed_at
-- integer claire_exercise_session.score

USER_COURSE_RESULT (Résultat du cours)

-- integer claire_user_course_result.reference_id,(le cours)
-- integer claire_user_course_result.user_score,
-- boolean claire_user_course_result.passed,
-- integer claire_user_course_result.passing_score,
-- integer claire_user_course_result.max_score,
-- datetime claire_user_course_result.created_at,
-- integer claire_user_course_result.user_id

Fonctionnement de OpenClassrooms
(mi-octobre 2015 à mi-avril 2016)

Cours

- Les cours sont organisés en catégories, et parfois structuré en parcours (un parcours = plusieurs cours)
- Les cours sont vidéos (suite de vidéo, qui permet de suivre l'intégralité du cours) ou textuels (majorité de textes)
- Les cours ont des exercices ou pas.
- Un cours certifiant a obligatoirement des exercices.

Follow Course

Un membre s'inscrit à un cours (opération explicite), ou est inscrit automatiquement par une interaction avec le cours.
Interaction avec le cours entraînant le follow_course:

- click sur le bouton follow course (Suivre le cours)
- mark as complete un chapitre
- débute un exercice

Ne plus suivre un cours est une action explicite de l'apprenant. (click sur le bouton ne plus suivre)

Mark as complete

Sur toutes les pages du cours, il y a un bouton "mark as complete" qui permet de signaler qu'on a fini un chapitre (effet sur la table des matières)

Premium

Description : https://openclassrooms.com/premium

Exercices

Il existe deux types d'exercices :

- **QCM**
- **Peer assesment** : il y a une activité à faire, avec un barème associé.
  Un document est soumis, l'utilisateur doit corriger 3 autres personnes afin que son devoir puisse être corrigé par 3 autres personnes. La note de l'exercice est obtenu par la moyenne des 3 notes.

Résultats

Cours

Une fois que tous les exercices sont faits, la plate-forme calcule un score total sur le cours.
Si le score >= 70%, le cours est réussi.
(avant : moyenne des exercices du cours,
maintenant : moyenne des exercices par partie, et moyenne des parties,
futur: plus de note, évaluation des compétences)

Parcours
Une fois que tous les cours d'un parcours sont finis, on obtient le certificat correspondant au parcours, et on a une note moyenne pour le parcours.

Abandon

Dans un premier temps, essayer de détecter l'abandon pour l'ensemble des utilisteurs, puis analyser les faux positifs, et essayer de trouver les patterns.

Remarques

- les événements "anciens" ont été récupérés.
- (futur) retour automatique au dernier endroit non marqués comme complété.
- (futur) disparition des notes en faveur de validations de compétences

Questions à Romain (LIUM) (27/05/16):
Questions sur les traces:

USER_COURSE_VISUALISATION récupère les traces de visualisation
c'est à dire qu'il permet de tracer l'activité des apprenants sur la plateforme?
Quel cours, session (?), partie du cours, à quelle date et par quel utilisateur a été visualisé.
Oui

=> (1) Permettrait de détecter les apprenants qui ne seront pas dans cette liste ou tout au moins ceux qui n'ont pas d'activité de ce type depuis un moment (à définir) avec le champ *oc_course_visualisation.date*.
Heu, oui ?

USER_COMPLETED_PART: enregistre les événements indiquant qu'une partie a été complétée par l'apprenant (user_id) (action spécifique dans l'interface?)
Permet donc de savoir quelle(s) partie(s) de cours a/ont été complétée(s), par quel utilisateur et quand.
Comment cette info est-elle récupérée? Par l'action sur le bouton "Mark as complete" qui signale qu'on a fini un chapitre ? enregistré par la date de l'action, la mise à "true" du booléen et la partie de chapitre concernée?
Oui (action de l'utilisateur mark as complete)

USER EXERCICES donne la structure des exercices (à quel chapitre et partie de cours est rattaché un exercice)
A quoi correspond "exercise_position" (toujours la valeur 1 dans le fichier de traces fourni)?
Lorsqu'il y a plusieurs exercices pour une même référence, permet de définir l'ordre des exercices
ie- autres valeurs possibles que 1?
Oui, imaginons 3 exercices pour la même partie, il y aura position 1, 2, 3

USER_COURSE_RESULTS : Donne les résultats des exercices - indique si un exercice a été "passé" (boolean user_passed), la valeur max (course_max_score, toujours 70 dans les traces)) et min (course_passing_score, toujours 70 dans les traces)) du score , le score obtenu, par quel apprenant.

Je n'ai pas trouvé le champ "created_at" dont parle Alya.
C'est dans le nouveau dataset

USER_FOLLOW_COURSE : indique les événements de suivi de cours par un apprenant.
A quoi correspond le champ "uuid"? à un identifiant d'evenement?
Oui

A Quel evenement correspond la trace enregistrée dans ce fichier? Quand l'apprenant clique sur le bouton "Follow course" ou commence le cours.
Section follow course mise à jour ?
https://docs.google.com/document/d/1LQdgYpYHqQzuDEc_sq4SzBlTYZCrpjbbyeJyZsehHkk/edit#heading=h.ptetqpdertag

Dans le cas du fichier de trace UNFOLLOW (Que je n'ai pas trouvé dans le répertoire mis à disposition le 10 mai 2016) : comment on repère que l'apprenant arrête d'interagir avec le cours? Quelle action? Clic sur un bouton?
Section follow course mise à jour
https://docs.google.com/document/d/1LQdgYpYHqQzuDEc_sq4SzBlTYZCrpjbbyeJyZsehHkk/edit#heading=h.ptetqpdertag

Premiers éléments d'analyse sur l' *étude des profils*
Partir des apprenants qui suivent un cours, repérer ceux qui sont Premium/non Premium et vérifier leur score (dans COURSE_RESULTS valeur la plus récente dans le champ *user_course_score*).
Pour ceux qui n'ont pas un score > =70 - repérer ceux qui ont abandonné ("droper") et ceux qui ont fait les exercices mais qui n'ont pas réussi ("failer").

Premium?

On considère les apprenants comme Premium toute personne qui est inscrite quelle que soit la date (même en cours de session de cours - avant la fin d'un cours - avant les résultats du cours).
Tout apprenant qui devient Premium en cours de déroulement du cours est à prendre en compte dans la catégorie Premium.

Comment repérer un abandon?

Par une absence d'activités sur le cours depuis un certain temps? ou l'absence du user_id apprenants dans les fichiers de trace qui détectent l'activité dans le cours.
On peut repérer à partir du fichier USER_COURSE_VISUALISATION,USER_COMPLETED_PART ou USER_FOLLOW_COURSE quels utilisateur n'ont pas interagit avec des parties de cours depuis un certain délai?
(en repérant les apprenants qui n'ont rien complété/visualisé (1) /suivi depuis un certain temps ou dont l'user_id n'apparaît pas dans la liste). :
On considère les apprenants comme Premium toute personne qui est inscrite quelle que soit la date (même en cours de session de cours - avant la fin d'un cours - avant les résultats du cours).
Tout apprenant qui devient Premium en cours de déroulement du cours est à prendre en compte dans la catégorie Premium.

Alya To Romain (TB, 10/05/2016)

Questions after initial data examination and analysis:

1- Applying some initial verification analysis on the data, it is noticed that the USER.id has some missing values.
Example: in USER_FOLLOW_COURSE there are 38065 unique (non-duplicated) ids, however 3766 of these ids could not be found in the USER.id list. This was also recorded with ids in other tables (USER_PREMIUM, USER_UNFOLLOW_COURSE, COMPLETED_PART, USER_COURSE_RESULT) Is there a logical explanation behind this or is it just shortage of data?

USER_FOLLOW_COURSE is a table on a Redshift database (ie: not the main datatbase), that register events. There is no relational schema on this database and no link with the main database, so I believe that this user have deleted their account. (so the ids are still in the event's tables, but not on the other tables)

2- There have been recorded duplicate values in the USER_COURSE_RESULT.user_id. Does this indicate multiple answer attempts, or course repetition?

A user can pass multiple times a course (but it needs an admin action). Main case: problem with an exercise, so the admin reinitialise an exercise, and the user can complete course again (course_result is for the whole course)

What result to choose among duplicate user ids (most recent, highest grade, average of duplicate scores)?

Most Recent is displayed to the user

3- Is there a “datetime” field in any of the tables that signals the end of the course? If not, is it possible to include such a field to the current USER_COURSE_RESULT table?

What do you mean by the end of a course ? A row in USER_COURSE_RESULT is created when all the exercises are completed (a score). So the createdAt field of this table means the end of the course.

4- The USER_FOLLOW_COURSE event is triggered when a learner presses the "follow course" button, or starts interacting with the course. Is this the case with Open Classrooms?

I don't understand: "Is this the case with Open Classrooms?" ;-) Could you be more precise ?

Concerning the USER_UNFOLLOW_COURSE event is it also triggered when a learner stops interacting with the course?

No, for the moment, it's an explicit action of the learner (a click on the button)

5- I am attempting to categorize learners in each course into two categories Premiums and Non-premiums. However, there is the constraint of the learners switching from premium to non-premium and vice versa multiple times. Therefore, it is necessary to regularly check throughout each course the state (premium or non-premium) of each user.

In order to achieve that, with the data we have, there are multiple ways:

a. check users state only when they follow a course.
b. check for users state throughout the course, for example on chapter completion and exercise sessions.(in this case a user might be in two states throughout the course!)
c. check for users state upon the end of the course.(reason for question 3)

Which of these three cases suits better the objectives of Open Classrooms?

There are three main usages of being premium in these cases :

1. ***the learner is already Premium, and follow the course***
2. ***the learner is blocked by the constraint and become Premium to go faster***
3. ***the learner has passed the course, and would like to unblock the certification***

In our scenario, I think we would like to understand the behavior of people being premium before the end of the course (not only those who were premium before the "follow course action", but also those who have been premium before the course result) We've made this difference between Premium/Not Premium because of the motivation of the learner, so a learner that has been premium during the course completion should be in the Premium category)

Comment détecter qu’une personne a abandonné ? Période fixe ou par rapport à un comportement “habituel” ?

How to determine the abandoning period of time T? We propose two initial options:

Option 1: Estimate the average visualization time for each student and consider ($T_{student}$=3xT_{Avg}).
Option 2: Estimate the average drop time among all students and consider (T=AvgT_{all}).

Quels sont les leviers pour agir sur les apprenants ?

- ***emails***
- ***persévérance***
- ***possibilité de développer d'autres leviers de motivation.***

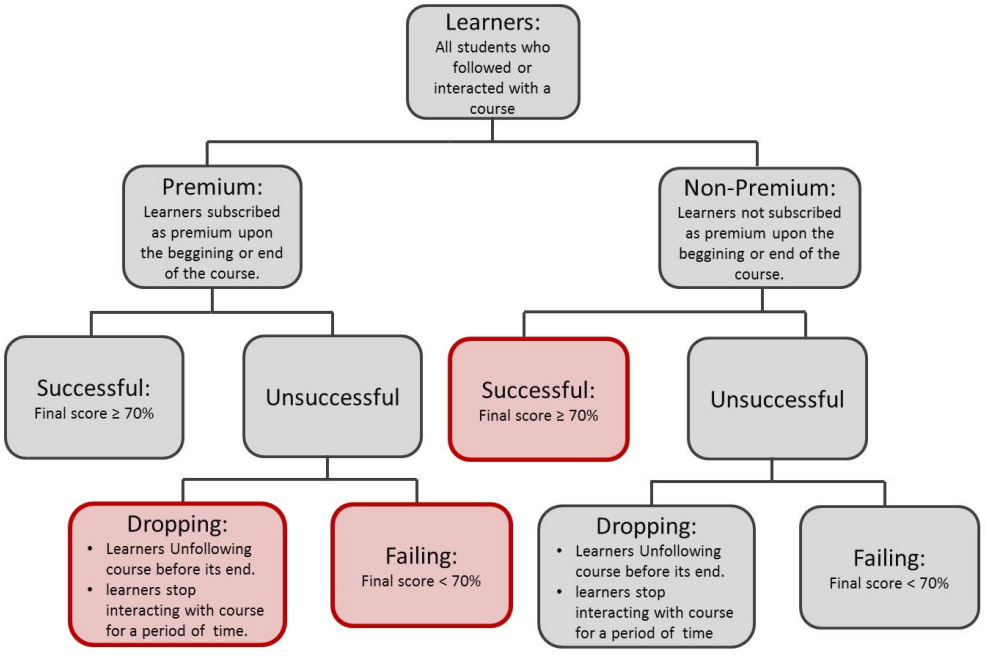
Comment sont collectées les données socio-démo

- **profil**
- **info google ou facebook**

Définition de "active user" (quand quelqu'un est banni) → inintéressant pour nous

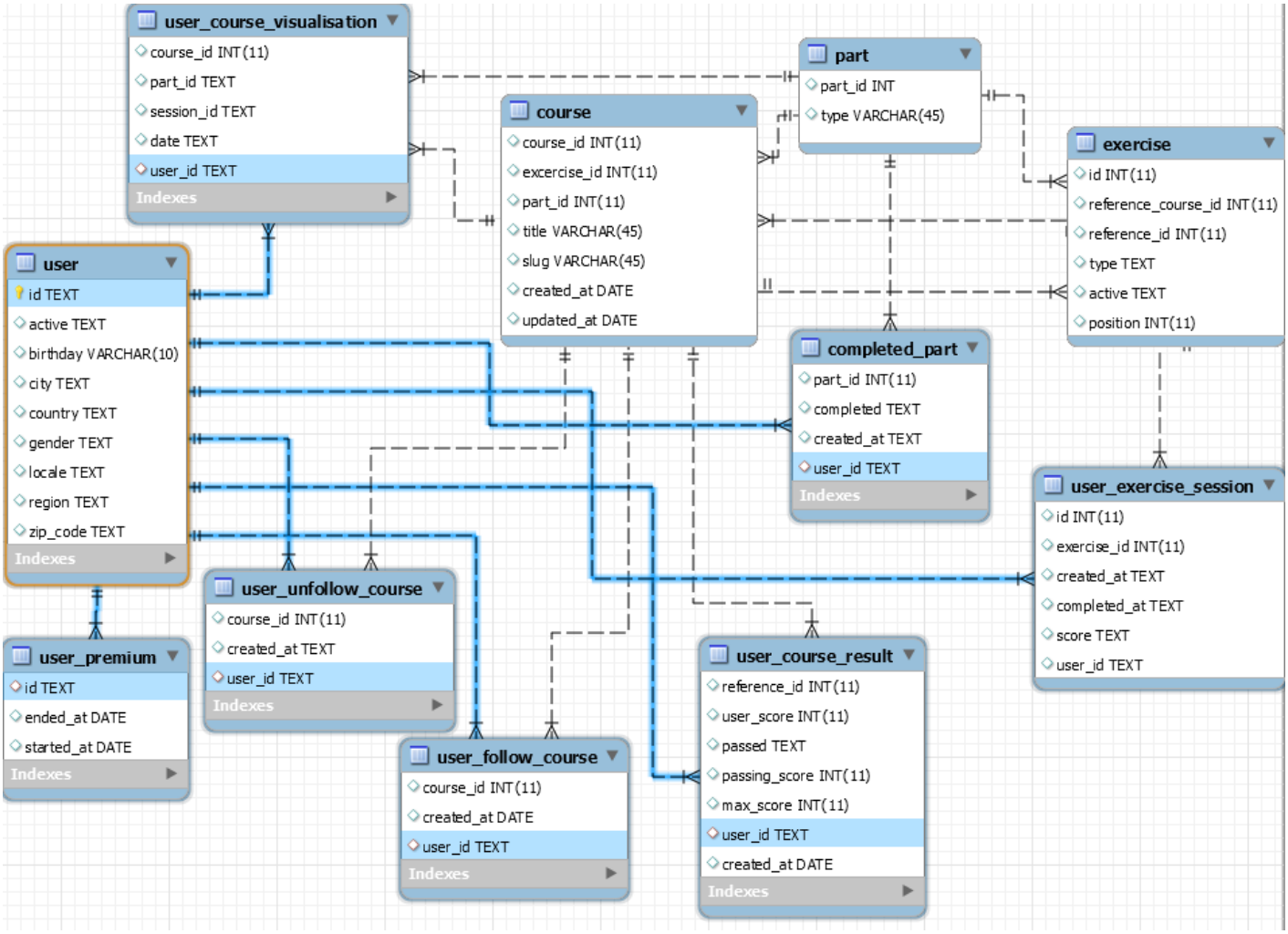
Follow / unfollow : follow automatique dès que une interaction / unfollow

Categorization process per course:



Alya TB (11/07/2016):

Current created data base of the actual existing version 3 of data using MySQL workbench.
note: course and part tables do not exist in CSV format they are included in the course structure as JSON files. However, the two tables where added to acquire a logicaly coherent data-base.
field data types in the diagram below are not accurate, still under revision.



Initial Classification Summary:

Course ID	Learners	Premium				Non-premium			
		Succeed	Fail	No Res	Total	Succeed	Fail	No Res	Total
26832 Java	24686	42	4	2041	2087	10	1	22588	22599
1946116 Entrepre ndre	2835	25	10	812	847	7	2	1979	1988
1056721 Node.js	5582	99	4	906	1009	25	0	4548	4573
1766341 XML	2611	17	4	444	465	2	0	2144	2146

Possible Factors of failure or drop-out (to discuss in the next meeting):

1. No real intention for completing the course even before starting:
   could be defined by learners who never interacted with the course after following it.
2. Lack of time:
   Learner might be too busy to keep up with the course plan.
3. High or complex course workload:
   it is a hard course in general keeping in mind the problem of MOOCs being a “one size fits all” type of courses.
4. Miss-placement within a course compared to learner’s skills:
   In other words being over or under qualified for this certain course. In the OC case study to do this particular analysis we need data about learners level of education.

5. Platform Complexity or poor MOOC design of structure:
   in the OC case we need to know more about MOOCs evaluation, since I assume this affects the analysis greatly.
6. Bad experience with previous MOOCs:
7. Starting Late:
   In the case where OC does not have a precise time limited session this factor wouldn't have any effect on drop-out and failing.

Alya 14-11-2016 (TB)

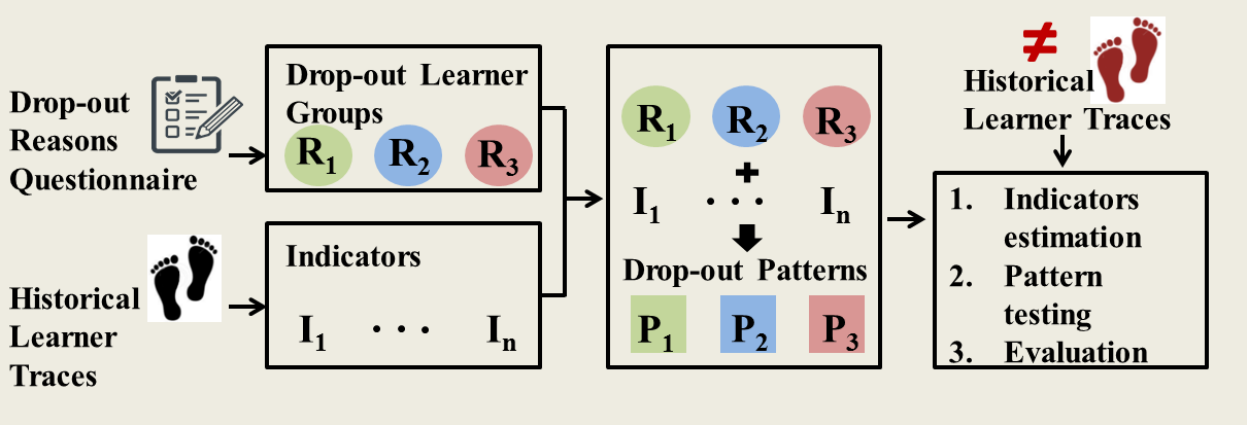
Analysing the new data:

1-Data includes activity on 10 courses.

2-Data has some minor changes in format which caused some problems in executing the developed cleaning and categorisation code (were corrected manually).

- The tables (USER_FOLLOW_COURSE, USER_UNFOLLOW_COURSE) have no label row for the column names.
- In the table (USER_PREMIUM) the previous field id is changed into user_id.
- in the table (USER_COURSE_RESULT) the user_id and created_at fields are swapped in placement order.
- The date fields have changes in some tables

3-In the cleaning process (erasing all events of ids that are not anymore in the USER table), the table PREMIUM does not include any ids missing from the USER table.



Possible Indicators:

Score obtained in each exercise of the course.

Time passed on each exercise.

Trajectory of engagement.

Frequency of interaction.