Emmanuel Jakobowicz

Vendredi de la data

Compte-rendu du vendredi 2 juin 2017

Les problématiques explorées dans le cadre des projets du CLEO :

- Lecteur inattendu
- Sleeping beauty
- Longue traine
- Conversation silencieuse

Le lecteur inattendu

La problématique : comment repérer dans les logs des sites gérés par le CLEO des pics de fréquentation sur certaines pages.

Discussion : La granularité la plus cohérente est de travailler au niveau quotidien. L'utilisation d'échantillonnage n'est pas possible dans ce cas.

Solution possible : Mettre en place des seuils d'alertes calculés par modélisation de séries temporelles. On crée un modèle sur quelques jours (par exemple avec une approche ARIMA) qui peut être validé sur un grand nombre de pages (on échantillonne). On peut, par exemple, faire un modèle par site. Si en appliquant le modèle sur les fréquentations des pages, la valeur prédite sort de l'intervalle de confiance de prédiction (on pourra prendre un intervalle à 99% afin d'élargir l'intervalle), on met en place une alerte.

Aller plus loin : On peut imaginer de mettre en place un système automatique d'alertes et de mise en valeur des contenus pour amplifier l'effet du lecteur inattendu (communication sur les réseaux sociaux, sur les pages d'accueil des sites…).

Remarque : si on veut se concentrer sur un lieu géographique donné, il faudra un certain nombre de lecteurs afin de concevoir un modèle (ça ne marcherait pas si on part de 0 lecteurs).

Sleeping beauty

La problématique : comment repérer des articles qui ont un afflux rapide et important de citations.

Discussion : la première interrogation réside dans le fait de bien repérer les citations dans les corpus de citations.

Solution possible : il s'agirait de mettre en place un système proche du lecteur inattendu permettant de monitorer le nombre de citations pour chaque article. Il faut d'abord mettre en

place un système de monitoring des citations. Le modèle ensuite sera plus basique car le nombre de citations reste très faible.

Aller plus loin : On peut imaginer de mettre en place un système automatique d'alertes et de mise en valeur des articles pour communiquer sur des articles impactant fortement la communauté.

La longue traine

La problématique : Comment vérifier que dans le cas de l'open édition, on a un phénomène de longue traine plus marqué que dans le cas payant ou dans le cas de l'édition classique (courbe skinnier).

Discussion : Le phénomène de longue traine est basé sur la théorie de Chris Anderson : « les produits qui sont l'objet d'une faible demande ou qui n'ont qu'un faible volume de vente, peuvent représenter une part de marché égale ou supérieure aux best-sellers, si les canaux de distribution peuvent proposer assez de choix et créer les moyens de découvrir cette diversité ». Jusqu'ici malgré l'opportunité du web, cette théorie ne s'est jamais vérifiée. On peut tout de même comparer les courbes de Lorenz.

Solution possible : Représenter les courbes de Lorenz par catégorie de journaux et faire dans un premier temps une comparaison visuelle entre les courbes. Attention, le nombre de documents sélectionnés est important, on devra sélectionner tous les documents en faisant un focus sur les premiers 50% si nécessaire. En fonction des catégories, le nombre d'articles va beaucoup différer, il faudra donc faire attention aux comparaisons.

Aller plus loin : Rassembler des données externes d'éditions payantes ou d'autres domaines (musique, films) afin de faire des comparaisons plus larges (attention au contexte).

La conversation silencieuse

La problématique : Comprendre les principes de fidélisation des lecteurs suivant le site utilisé.

Discussion : Les différents sites gérés par le CLEO ont des publics assez différents. L'objectif est de comparer ces publics en utilisant les logs. L'information principale dont on dispose pour analyser cette « fidélité » est le fait d'avoir un utilisateur récurrent.

Solution possible : Pour tester l'hypothèse de fidélité équivalente d'un site à un autre, il suffit de créer une statistique permettant de mesurer cette fidélité.

Différentes approches de mesures :

- *Visiteurs récurrents* = Nombre de visiteurs uniques - Nombre de nouveaux visiteurs : statistique simple mais peu fiable car ne prenant pas en compte la taille du site et le degré de renouvellement des contenus.

- *Visiteurs récurrents / nombre de nouveaux contenus* : plus fiable mais ne prend pas en compte la taille sur site.
- *Visiteurs récurrents / (a * nombre de nouveaux contenus + b * nombre de pages)* : plus fiable mais nécessité de bien définir a et b

Une fois la mesure choisie, on doit choisir le pas de temps étudié. On prendra le mois afin d'avoir assez d'antériorité pour avoir des mesures (il faudrait au moins un historique de 1 an pour pouvoir aboutir à des résultats fiables)

On peut utiliser les données mensuelles d'un site afin de les comparer à un autre. On pourra appliquer un test non paramétrique sur données appariées (Wilcoxon signé par exemple)

Des essais sont nécessaires afin de valider cette approche et afin de bien choisir la mesure et le test adaptés.

Aller plus loin : on pourra créer des groupes de visiteurs en utilisant des algorithmes de machine learning en se basant soit sur toutes les données soit sur un échantillon et en utilisant toutes les informations disponibles (localisation, récurrence…)