# Conception d'un modèle généraliste pour l'évaluation d'un test A/B

Emmanuelle Claeys, Pierre Gançarski, Myriam Maumy-Bertrand, Hubert Wassner

# Conception d'un modèle généraliste pour l'évaluation d'un test A/B

Emmaneuelle Claeys*, Pierre Gançarski *
Myriam Maumy-Bertrand** Hubert Wassner***

*ICube – Université de Strasbourg – 67412 – Illkirch – France
{claeys, gançarski}@unistra.fr
**IRMA – Université de Strasbourg – 67084 – Strasbourg – France
mmaumy@math.unistra.fr
***Entreprise AB Tasty – 3 Impasse de la Planchette – 75003 – Paris – France
hubert@abtasty.com

**Résumé.** Cet article propose un nouveau modèle d'architecture pour une évaluation de test *A/B* dans différents contextes. Il rappelle les problématiques que le test *A/B* rencontre lorsqu'il évalue plusieurs types de e-commerces et propose une architecture globale pour pallier à ces problématiques.

## 1 La problématique d'un test A/B

De plus en plus aujourd'hui, les entreprises investissent dans des solutions dites *big data* : plus de données, plus d'informations. De fait, elles se doivent d'explorer les bénéfices que peuvent apporter ces informations. En effet, elles-ci sont encore très majoritairement liées à des données internes de l'entreprise. Que ce passera-t-il lorsque la connaissance sera ré-utilisée pour des contextes différents par différentes entreprises, mais confrontées à une même problématique ? Ce que l'humain, par son expérience, apporte en tant que "consultant", pourra-t-il être modélisé et automatisé afin identifier pour chaque entreprise, rapidement, la solution la plus adaptée à son profil et améliorer au fil du temps grâce à un enrichissement continu et automatique de cette connaissance ? Cette (r)évolution a déjà commencée : avec l'essor du e-commerce, cette problématiques de détection de solutions automatisées, notamment pour plate-formes automatisées de test *A/B* est de plus en plus cruciale dans les entreprises spécialisées. Un test *A/B* (ou A/B *testing* [1]), consiste à créer plusieurs versions d'un même objet, dans le but de vérifier une proposition (Roukine, 2011). Par exemple, il s'agit de tester auprès d'un échantillon d'internautes, nommés visiteurs, durant une période donnée, différentes variations (généralement deux) d'une même page web (nommées variation *A* et variation *B*) proposées par un e-commerçant. Le but est d'identifier celle qui donne les meilleurs résultats par rapport à un objectif fixé par le e-commerçant en amont du test (par exemple le nombre de visites par visiteur). Le test peut être évalué sur l'ensemble des visiteurs de la page ou uniquement sur un échantillon de visiteurs ayant une caractéristique spécifique (par exemple les visiteurs parisiens). Dans les solutions commerciales actuelles, un visiteur est orienté aléatoirement vers

1. Le terme *A/B* étant un terme générique pour qualifier tout type de tests

la variation *A* ou *B* (généralement à proportion égale). Il est également possible de faire évoluer ce pourcentage selon l'identification de la variation optimale parmi l'ensemble des variations disponibles en appliquant par exemple l'algorithme de Thompson sampling (Thompson, 1933). En effet, le site étant généralement en production, il est intéressant de garder un compromis entre l'exploration (évaluer les différentes variations) et l'exploitation du test (favoriser la variation avec les meilleurs résultats). Une contrainte commune, lors d'un test *A/B*, est que lorsqu'un visiteur est affecté à une variation, elle lui est affectée pour toute la durée de sa visite, jusqu'à la fin du test. Trouver un tel compromis dans ces conditions nécessite d'affecter au mieux à un nouveau visiteur, la version *A* ou *B* du test.

## 2 Une approche multi-contextes

Pour résoudre ce problème d'affection, on propose un modèle d'affection basé sur une segmentation des visiteurs et sur une évaluation à la volée des résultats du test et ce pour chaque segmentation de façon indépendante. La stratégie proposée s'établi autour de deux objectifs :

— observer a posteriori si la variation choisie est optimale (ou sous-optimale) pour ce segment (ou profil) de visiteurs pour un ou plusieurs critères définis par le e-commerçant.
— proposer a priori la meilleure variation au visiteur si le système est convaincu de la pertinence de ce choix.

Deux verrous majeurs sont à lever : le modèle décisionnel doit prendre en compte la performance et la pertinence de ses choix. Il devra adapter ses choix en fonction du type de visiteurs et du contexte du test. Le modèle devra être également capable d'apprendre de ses succès/erreurs et de s'auto-corriger (apprentissage par renforcement par exemple). Néanmoins, une caractéristique fondamentale à prendre en compte est que dans le domaine du e-commerce, les taux de succès de la visite (achat par exemple) restent très faibles (entre 1% et 10% ) (Gensollensem, 2001). On ne doit pas systématiquement pénaliser une décision d'affectation en cas d'échec.

Enfin, ce modèle doit être suffisamment générique pour pouvoir être utilisé dans différentes configurations : type de e-commerce, type de tests réalisés, période de test, etc.

## 3 Modélisation

Actuellement, lorsqu'un e-commerçant souhaite réaliser sur le site un test *A/B*, il fait appel à un expert chargé d'examiner son site. Ce dernier étudie le trafic, les types de transactions réalisées, et lui demande de choisir ce qu'il souhaite comparer (par exemple le taux de clics, d'abandon de panier, ou encore la valeur moyenne d'une transaction). Avec l'aide de l'expert, le e-commerçant choisit de modifier un élément de sa page puis débute le test.
L'objectif de nos travaux est d'automatiser cette analyse en s'inspirant des étapes du processus. La Figure 1 est un schéma simplifié représentatif de notre proposition (Simpson et Simpson, 2009).

Dans une première étape, le e-commerçant va être classé automatiquement dans une catégorie de commerces. Classer le e-commerçant permet de partir avec un *a priori* lors de l'analyse. Par exemple, un test appartenant à une classe d'e-commerçants avec beaucoup d'ache-

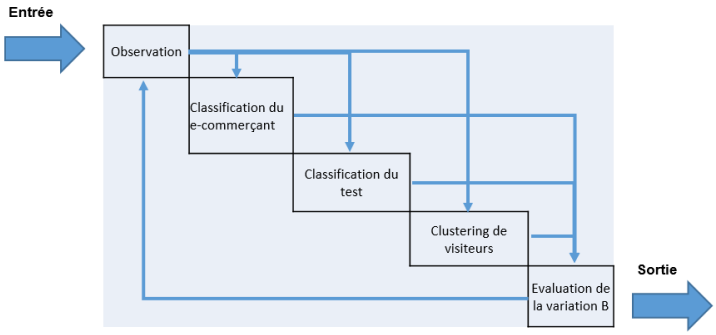


FIG. 1 – *Représentation de notre système de systèmes avec la méthode NSquared Chart*

teurs ne sera pas construit et évalué de la même façon qu'un test sur une classe d'e-commerçants avec très peu d'acheteurs.

Si les valeurs suivent une loi normale, on constate que le pic de densité est plus important chez le e-commerçant d'articles sportifs. Cela peut notamment s'expliquer par la période durant laquelle les données ont été collectées, ou encore la notoriété de la marque. Une première analyse hors test permet de récolter des informations sur la page originale. Ces données récoltées sont actuellement principalement quantitatives (les transactions, le taux de clics ou encore le nombre de visites). Idéalement, le système devrait pouvoir exploiter la connaissance des tests réalisés par des e-commerçants similaires.

Afin de créer de telles catégories de e-commerçants à partir de ces données, une classification ascendante hiérarchique peut être appliquée , en mettant un coefficient plus important sur l'objectif du e-commerçant (par exemple, donner plus de poids à la variable *Frais de livraison* si c'est ce que souhaite observer le e-commerçant).

Une fois que le e-commerçant est classé, dans une deuxième étape, la variation que le e-commerçant souhaite appliquer par rapport à sa page originale doit être étudiée. L'ontologie proposée par (Roukine, 2011) permet de caractériser la modification d'une page originale suivant quatre aspects :

— **ACTION** (ajouter, supprimer, modifier)
— **OBJET** (image, texte, couleur, bannière)
— **POSITION** (en-tête, contenu, pied de page)
— **QUAND** (jour et mois)

Un test peut ainsi, par exemple, présenter les caractéristiques suivantes **[ajouter, image, pied de page]**. La variable **QUAND** ne peut en général être utilisée que lorsque la base de données est assez importante. De même que pour la classification des e-commerçants, une classification des tests, permettra, en fonction des objectifs, de mieux recommander des tests, et ainsi identifier le test optimal à réaliser à partir de l'objectif du e-commerçant.

Pour créer ces classes, les variables que souhaite comparer le e-commerçant (par exemple le nombre de visites) sont utilisées. Ces variables sont les mêmes que celles utilisées pour créer les classes de e-commerce. Ainsi, dans cette deuxième étape de notre méthode, un arbre de classification des tests est créé en fonction de l'objectif du e-commerçant.

Dans une troisième étape, des groupes de visiteurs vont être créés à partir du trafic sur le site. Si une variation peut être optimale pour la majorité des visiteurs, il est plus intéressant pour un e-commerçant d'observer pour quelle catégorie de visiteurs cette variation est optimale/sous-optimale ? En fonction de ces segments et de leurs résultats, il choisira d'investir dans un affichage personnalisé, ou au contraire, basculer définitivement tout le trafic vers *A* ou vers *B*. Pour classer les visiteurs, une grand nombre de variables quantitatives et qualitatives sont généralement disponibles (localisation, date de la visite, type de navigateur, ...). De plus, dans nos expériences, nous avons caractérisé les centres d'intérêt de nos visiteurs grâce à leurs *referers* : Les referers des visiteurs correspondent aux pages visitées par ce dernier avant d'arriver sur la page test. Les referers externes au site du e-commerçant sont cependant peu nombreux puisque les visiteurs passent généralement par un moteur de recherche pour arriver sur le site. Les referes internes traduisent la navigation d'un visiteur à travers le site du e-commerçant.
La classification des visiteurs proposée se fait en trois étapes :

1. une Analyse Factorielle de Données Mixtes (AFDM) permet de sélectionner des facteurs "représentatifs"
2. une classification ascendante hiérarchique permet de créer des groupes (clusters) de visiteurs
3. enfin, une régression sur chaque classe estime la fonction associée à la variable intéressant le e-commerçant (par exemple les valeurs de panier)

Enfin, lors de la dernière étape, correspondant au test lui-même, cette régression permettra de comparer les résultats de la variation *B* avec l'originale *A*. Elle est réalisée pour chaque segment de visiteurs ce qui permettra de détecter éventuellement une différence de résultats importante entre *A* et *B* pour un segment de visiteurs particuliers ou au contraire, de ne pas constater de différences significatives et de continuer l'exploration. Lorsque l'écart moyen entre la courbe de la variation *A* et de la variation *B* converge au-delà d'un certain seuil, le modèle décidera alors d'allouer exclusivement *A* ou *B*.

# 4 Conclusion

Dans ce papier nous avons proposé une méthode d'allocation de variation lors d'un test *A/B* basée sur la classification des e-commerçants, de celle des tests et enfin de celle des visiteurs. L'objectif de cette méthode est d'observer l'amélioration d'un test *A/B* par rapport aux techniques existantes actuellement. Les premières expérimentations que nous avons menées, bien qu'incomplètes, montre que cette approche est prometteuse. Dans la suite, nous proposons de mener des analyses plus poussées. En effet, cette modélisation est une série de systèmes complexes organisés hiérarchiquement et plus spécifiquement de systèmes de systèmes (Jamshidi, 2008) qui nécessite encore des études théoriques et techniques approfondies. Cependant, cette approche peut se voir comme un système ouvert, qui interagit en permanence avec son environnement. Si la tendance actuelle de l'utilisation du *big data* est encore principalement centrée sur des environnements clos, il est probable que l'utilisation et la gestion des connaissances seront utilisées et enrichies pour gérer des situations différentes, dans des contextes différents, et s'y adapter, comme le ferait l'expert humain.

## Références

Gensollensem, M. (2001). Internet : marché électronique ou réseaux commerciaux ? Volume 52, pp. 137–161. Revue économique.

Jamshidi, M. (2008). *System of Systems Engineering Innovations for the 21st Century*. Wiley.

Roukine, S. (2011). *Améliorer ses taux de conversion web : vers la performance des sites web au-delà du webmarketing*. Eyrolles.

Simpson, J. et M. Simpson (2009). System of systems complexity identification and control, in system of systems engineering. pp. 1–6. System of Systems Engineering.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two sample. Volume 25, pp. 285–294. Biometrika.

## Summary

This paper proposes a new architectural model for *A/B* testing evaluation in plurals contexts. It will expose the global problematic of *A/B* testing when we want manage several types of e-commerce. The goal of this paper is to provides global architecture to address these issues.