UNIVERSITÉ DE MONTRÉAL

MODÈLE DE PRÉVISION DES TAUX DE CLICS DES ANNONCES TEXTUELLES SUR LES MOTEURS DE RECHERCHE

FAROOQ SANNI
DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(MATHÉMATIQUES APPLIQUÉES)
AOÛT 2017

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

MODÈLE DE PRÉVISION DES TAUX DE CLICS DES ANNONCES TEXTUELLES SUR LES MOTEURS DE RECHERCHE

présenté par : SANNI Farooq
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. GAMACHE Michel, Ph. D., président
M. ADJENGUE Luc-Désiré, Ph. D., membre et directeur de recherche
M. LABIB Richard, Ph. D., membre

DÉDICACE

À ma famille.

REMERCIEMENTS

Je tiens tout particulièrement à remercier mon directeur de recherche M. Luc-Désiré Adjengue pour son support tout au long de ma maîtrise. En effet, son expertise et sa disponibilité ont été très précieuses dans la réalisation de ce projet.

Je remercie également mes parents pour leur amour, leur soutien moral et financier durant toutes ces années d'études. Merci à mes deux sœurs qui ont toujours fait preuve d'intérêt et d'encouragement dans mes travaux.

Je remercie aussi M. Olivier Gaudoin d'avoir accepté être mon tuteur à l'ENSIMAG dans le cadre de ma mobilité à Polytechnique Montréal. Enfin, mes remerciements vont à M. Michel Gamache et M. Richard Labib, respectivement président et membre de mon jury.

RÉSUMÉ

Le taux de clics est une métrique essentielle dans les campagnes publicitaires sur les moteurs de recherche. En effet, il impacte directement les deux acteurs principaux de la publicité en ligne que sont les moteurs de recherche d'un côté et les annonceurs de l'autre. D'une part le taux de clics est la principale variable utilisée par les moteurs de recherche dans leur algorithme d'affichage des annonces textuelles. Aussi leurs revenus sont intimement liés à l'ordre d'affichage des différentes annonces. De plus, proposer une publicité pertinente à un utilisateur améliore son expérience et l'incite à utiliser davantage le moteur de recherche. D'autre part, le taux de clics joue le rôle d'indice de qualité pour les annonceurs ; ces derniers ajustent les paramètres de leurs campagnes suivant les valeurs du taux de clics. Une bonne prédiction du taux de clics est alors très importante aussi bien pour les moteurs de recherche que pour les annonceurs.

Pour prédire le taux de clics, les moteurs de recherche disposent d'un historique riche et détaillé des réalisations des annonces textuelles. Les principales variables disponibles sont des variables catégoriques issues des informations sur les annonceurs, les utilisateurs ou encore des données géographiques. Dans ce mémoire, la régression logistique est appliquée deux fois pour prédire le taux de clics. Les données des campagnes publicitaires contiennent beaucoup d'observations à taux de clics nul complexifiant la modélisation. Ainsi, la première régression logistique permet d'écarter ces observations tandis que la seconde prédit le taux de clics des autres observations. Aussi des variables « inédites » sont utilisées dans ces deux régressions. En effet les variables *position moyenne*, *nombre d'impressions* et *coût* sont d'abord modélisées, puis elles sont utilisées comme variables explicatives dans le modèle logistique. Ces variables sont en réalité des variables de réponse tout comme le taux de clics. Ainsi nous proposons un modèle pour chacune de ces variables. La loi normale tronquée est ajustée à la position moyenne ; pour le nombre d'impressions et le coût, différents modèles sont explorés notamment les modèles linéaires généralisés (Poisson, Gamma, lognormal). Des modèles de type *hurdle* sont finalement retenus. Aussi, nous montrons qu'une hypothèse d'indépendance temporelle des observations, nécessaire à l'application de nos méthodes, est plausible malgré le phénomène de mesures répétées. Enfin les expériences menées sur des données réelles, montrent que cette modélisation en chaîne donne de bons résultats et peut encore être améliorée.

ABSTRACT

Click-through rate is an essential metric in advertising campaigns on search engines. As a matter of fact, it directly impacts the two main players of online advertising which are search engines and advertisers. On the one hand, the click-through rate is the main variable used by search engines in their algorithm for displaying text ads. Also their revenues are intimately linked to the order of display of the different ads. Additionally, offering relevant advertising to a user improves their experience and encourages them to make greater use of the search engine. On the other hand, the click-through rate plays the role of a quality score for advertisers who adjust their campaign settings based on click-through rate values. A good click-through rate prediction is very important for both search engines and advertisers.

To predict the click-through rate, search engines have a large amount of historical data on text ads. The main variables available are categorical variables derived from information about advertisers, users, or geographic data. In this paper, logistic regression is applied twice to predict the click-through rate. Campaign data contains many observations with zero clicks that make modeling more complex. The first logistic regression then discards these observations while the second predicts the click-through rate of the other observations. Also, new variables are used in these two regressions. Indeed the variables *mean position*, *number of impressions* and *cost* are first modeled then they are used as explanatory variables in the logistic model. These variables are actually response variables as the click-through rate. Thus, we propose a model for each of these variables. The truncated normal distribution is adjusted to the *mean position* ; for the *number of impressions* and the *cost*, different models are explored in particular some generalized linear models (Poisson, Gamma, lognormal). *Hurdle* models are finally retained. We also show that a hypothesis of temporal independence of observations, necessary for the application of our methods, is plausible despite the phenomenon of repeated measures. Finally, experiments carried out on real data show that this chain modeling gives good results and can be further improved.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

CPC	Cost Per Click
CTR	Click-Through Rate
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Abslute Error
LR	Likelihood Ratio
ACF	Auto-Correlation Function
PACF	Partial Auto-Correlation Function
ARMA	Autoregressive Moving Average
LASSO	Least Absolute Shrinkage and Selection Operator

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

L'accès à l'internet est aujourd'hui un acquis pour plus de 43% de la population mondiale selon les chiffres de l'année 2016 de l'Union Internationale des Télécommunications (UIT, 2016). Dans les pays dits développés, ce pourcentage atteint 81%. Ainsi au Canada par exemple, plus de quatre personnes sur cinq naviguent sur le web, effectuent des recherches et surtout sont susceptibles de voir des publicités. Internet, avec son large auditoire, apparaît alors comme un support incontournable pour effectuer la promotion d'un produit ou d'un service. C'est pourquoi l'investissement des entreprises dans les campagnes publicitaires en ligne pour promouvoir leurs produits ne cesse d'augmenter depuis plusieurs années. En effet la part de marché de la publicité en ligne ne cesse d'augmenter au détriment des supports traditionnels que sont la télévision, la radio et la presse (Bys, 2017). L'année 2016 a notamment marqué un tournant puisque pour la première fois, la publicité en ligne est passée devant la télévision en terme de part de marché (PwC, 2016).

Outre son large public, cet engouement pour la publicité en ligne s'explique aussi par la flexibilité qu'elle offre. D'abord, elle permet un meilleur ciblage des utilisateurs. De plus, plutôt que de s'adresser à un groupe comme c'est le cas des spots publicitaires, on peut ici s'adresser à chaque utilisateur en personnifiant la publicité grâce à l'analyse des *cookies* (petits fichiers stockés sur le terminal de l'utilisateur et contenant des informations personnelles) par exemple. Enfin, la publicité en ligne peut être plus facilement modifiée, ajustée.

On distingue essentiellement trois types de publicité en ligne : les annonces textuelles, les bannières et la vidéo. Les annonces textuelles sont associées au « réseau recherche » ; elles se présentent sous la forme d'un texte court (4 lignes maximum) et apparaissent au dessus des résultats d'une requête sur un moteur de recherche. Les bannières sont quant à elles associées au « réseau display » c'est-à-dire qu'elles sont diffusées sur différentes pages web à côté du contenu de ces pages. Enfin la vidéo est la forme de publicité qui est le plus en essor ; elles se retrouvent essentiellement sur les réseaux sociaux et sur les plateformes de vidéo telles que Youtube. C'est la première forme de publicité, les annonces textuelles, qui va nous intéresser dans ce présent mémoire.

Dans la suite de cette introduction, nous donnons la définition de quelques termes utiles à la compréhension puis nous décrivons le mécanisme de fonctionnement des annonces textuelles. Enfin nous explicitons notre problématique suivie de la description de notre projet de mémoire.

1.1 Définitions

Le monde de la publicité en ligne dispose d'un jargon dont une connaissance préalable est nécessaire. Aussi il est important d'expliquer clairement les termes qui sont utilisés tout au long de ce mémoire.

Notre travail est consécutif à ceux de Quinn (2011) et Assari (2014). Les définitions qui sont décrites ci-dessous sont également présentées dans leurs mémoires ; la section 1.2 de Quinn (2011) est particulièrement très exhaustive sur ces définitions. Ainsi la présentation ci-dessous est succincte mais suffisante à la compréhension. Le lecteur intéressé par plus de détails est invité à consulter la section 1.2 de Quinn (2011).
Nous divisons les définitions en une première partie portant sur la terminologie et une seconde présentant les variables d'intérêts dans le mécanisme des annonces textuelles.

1.1.1 Terminologie

Ici nous présentons quelques éléments de vocabulaire utilisés dans le contexte des campagnes de publicité sur les moteurs de recherche.

- Un **utilisateur** désigne un internaute, c'est-à-dire une personne qui utilise internet plus particulièrement un moteur de recherche dans notre cas.
- L'**annonceur** désigne l'entreprise ou la personne qui souhaite donner de la visibilité à un produit ou un service.
- Un **moteur de recherche** est « un outil de recherche qui référence automatiquement les pages web se trouvant sur le réseau Internet à l'aide d'un programme » (L'encyclopédie illustrée du marketing, 2017). L'utilisateur interroge le moteur de recherche en entrant des mots ; ce dernier lui retourne alors un ensemble de résultats jugés pertinents. Google est le plus important moteur de recherche et est utilisé par plus de 78% des utilisateurs (NetMarketShare, 2017). Derrière ce mastodonte, on peut citer Bing (8%), Baidu(8%), Yahoo(5%).
- Une **requête** est la recherche effectuée par l'utilisateur. Plus précisément c'est la suite de mots entrée dans la barre de recherche.
- Une **annonce textuelle** « constitue une forme de communication marketing que les annonceurs peuvent utiliser pour promouvoir leur produit ou service » (Google AdWords, 2017). Elle se retrouve uniquement dans les moteurs de recherche à la suite d'une requête effectuée par un utilisateur ; elle s'affiche avant les résultats dits organiques, c'est-à-dire ceux qui sont liés à la requête et qui ne sont pas de la publicité. Elle se présente sous la forme d'un texte décrivant le produit et propose un lien vers le

site de l'annonceur. La figure 1.1 montre quatre annonces textuelles obtenus suite à la requête « assurance habitation » sur le moteur de recherche Google.

- Un **mot clé** est un ensemble de mots que l'annonceur fournit au moteur de recherche et qui décrit son produit. Ainsi lorsqu'un utilisateur effectue une recherche dont les termes sont similaires à ceux du mot clé, l'annonce associée est alors susceptible de s'afficher. Le degré de similitude entre la requête et le mot clé est défini par le type de correspondance du mot clé. On distingue entre autres les options « requête large », « mot clé exact », etc. Ce paramètre nous intéresse peu car nous ne travaillons qu'avec la correspondance « requête large ». Dans le cas de la « requête large », les fautes d'orthographe ainsi que les synonymes entre le mot clé et les requêtes sont acceptés. Par exemple la requête « manteau dames » affichera l'annonce associée au mot clé « manteaux femmes ».
- Une **campagne** est un regroupement de mots clés d'un annonceur qui partagent des paramètres tels que le budget et le ciblage géographique.

1.1.2 Mécanisme de fonctionnement des annonces textuelles

Avant de décrire le mécanisme de fonctionnement des annonces textuelles, nous définissons d'abord les éléments relatifs à celui-ci.

- Une **impression** correspond à l'affichage d'une annonce textuelle. Cette impression provient de la correspondance entre un mot clé en particulier et la requête d'un utilisateur. On distingue alors le nombre d'impressions d'une annonce de celui d'un mot clé. Le nombre d'impressions d'une annonce textuelle est le nombre de fois où elle est apparue tandis que le nombre d'impressions d'un mot clé est le nombre de fois que ce mot clé a généré l'impression d'une annonce. Le nombre d'impressions d'une annonce est partagé entre ses mots clés. Dans la suite, le nombre d'impressions désignera implicitement le nombre d'impressions d'un mot clé.
- On parle de **clic** lorsqu'un utilisateur clique sur une annonce textuelle. Le mot clé ayant conduit à ce clic voit son nombre de clics incrémenté d'une unité. La distinction précédente faite pour les nombres d'impressions s'applique également au nombre de clics.
- Le **taux de clics** ou CTR (Click-through Rate) est le rapport du nombre de clics sur celui d'impressions. Il indique la proportion d'utilisateurs qui voit une annonce et clique dessus. Un CTR élevé est un bon indicateur de la qualité et de la pertinence d'un mot clé.
- La **position** est la place occupée par l'annonce au moment de son affichage. Il s'agit donc de valeurs entières. La position la plus élevée est 1 indiquant que l'annonce appa-

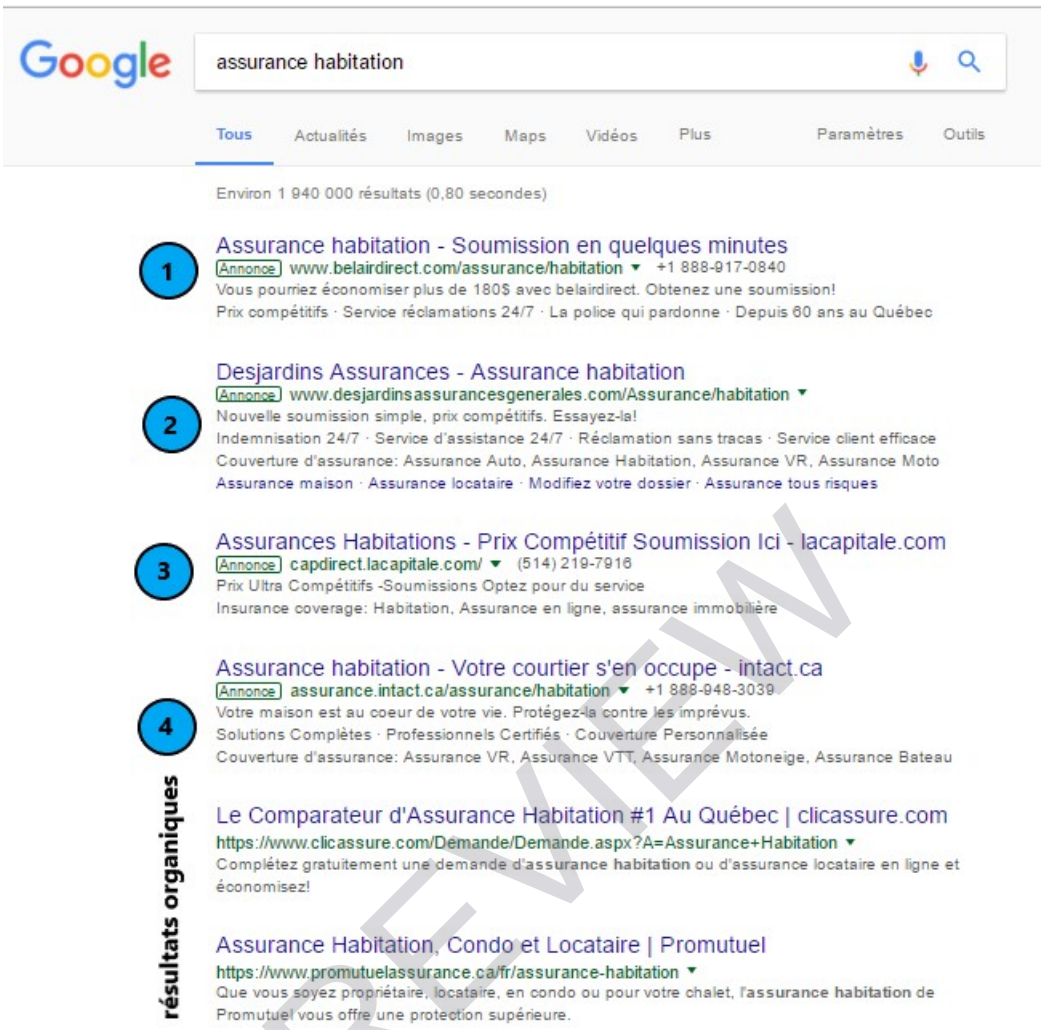

Figure 1.1 Résultats d'une requête dans le moteur de recherche Google.

raît tout en haut de la première page de résultats. Elle peut être très grande (supérieure à 20) puisque des annonces sont également affichées sur les autres pages de résultats. On retrouve entre 1 et 8 annonces sur la première page. De plus, la plupart des utilisateurs ne vont pas au delà de la deuxième page donnant donc des valeurs de positions entre 1 et 10. Sur la figure 1.1, nous avons marqué en face de chaque annonce sa position pour illustrer celle-ci.

Pour chaque mot clé, les moteurs de recherche calculent la moyenne des différentes positions occupées au cours d'une journée. On se retrouve alors avec une position moyenne qui n'est pas toujours entière dans les données disponibles.

- Dans ce mémoire, le **coût** désigne la somme payée pour un clic sur une annonce (CPC). Il existe la tarification au clic et la tarification au millier d'impressions. C'est la première

que nous étudions. Une quantité importante est le coût par clic maximal (max CPC) : il s'agit du montant maximal qu'un annonceur est prêt à payer pour obtenir un clic.

- Lorsqu'un utilisateur clique sur une annonce, il est redirigé vers un site web où il est invité à effectuer une action telle qu'un achat, une souscription à un abonnement, un visionnage, etc. On parle de **conversion** lorsque cette action est complétée.

Soulignons que les mesures de ces quantités sont agrégées sur une journée. En effet les moteurs de recherche fournissent aux annonceurs des statistiques quotidiennes : un nombre d'impressions quotidien, un coût quotidien, etc.
Une fois les termes importants définis, une brève description du fonctionnement des annonces textuelles est décrite ci-dessous.

Une requête dans un moteur de recherche génère en plus des résultats, des annonces publicitaires. L'apparition d'une annonce est déterminée à la suite d'un processus d'enchères effectué par un algorithme du moteur de recherche. La valeur d'enchère d'un mot clé est le produit du max CPC, fixé par l'annonceur, et d'un indice de qualité calculé par le moteur de recherche. Les annonces sont alors affichées par ordre décroissant de leurs valeurs d'enchères. Deux composantes interviennent donc dans l'affichage d'une annonce : d'abord le max CPC contrôlé par l'annonceur qui peut décider de l'augmenter ou de le diminuer, puis l'indice de qualité. Le calcul de cet indice diffère d'un moteur de recherche à l'autre ; il est précieusement gardé par les moteurs de recherche. On sait néanmoins qu'il dépend de la pertinence des mots clés associés à l'annonce, de l'historique du taux de clics de l'annonce, de considérations géographiques et d'autres facteurs.
Généralement, les premières positions sont celles qui génèrent le plus de clics (Agichtein et al., 2006; Joachims et al., 2005; Craswell et al., 2008). Ainsi il y a une concurrence naturelle qui apparaît entre les annonceurs afin que leurs annonces occupent ces places. Mais on voit déjà qu'à cause de l'indice de qualité, il ne suffit pas de fixer un max CPC élevé pour atteindre les premières places. Aussi il faut qu'une campagne publicitaire soit rentable c'est-à-dire que les revenus attendus soient supérieurs aux montants dépensés dans la publicité.

1.2 Problématique et présentation du projet

La qualité d'une annonce est jugée par son taux de conversion c'est-à-dire le rapport du nombre de conversions sur celui d'impressions. Néanmoins ce critère s'avérant compliqué à mesurer car propre à chaque entreprise, on lui préfère le taux de clics. Aussi on suppose implicitement une corrélation positive entre ces deux taux c'est-à-dire que le taux de conversion se comporte de manière similaire au taux de clics (un taux de clics élevé traduit un taux de conversion élevé). Ainsi une entreprise qui lance sa campagne de publicité souhaite maximiser

le taux de clics de ses mots clés sous une certaine contrainte budgétaire.

Le problème à résoudre ici est un problème d'optimisation sous-contraintes. Beaucoup d'études se sont donc penchées sur sa résolution. Nous en parlons plus en détails dans la revue de la littérature. Néanmoins la présence de beaucoup d'inconnues notamment l'algorithme d'affichage des annonces des moteurs de recherche rend le problème assez complexe.

Notre hypothèse est qu'une étude statistique approfondie des données apporte des informations supplémentaires qui permettent d'améliorer la résolution de ce problème. Cette composante statistique du problème a été moins traitée dans la littérature. Dans le cadre de ce projet, nous proposons une approche basée sur une double régression logistique pour prédire le taux de clics d'un mot clé à partir uniquement d'un historique des réalisations de ses positions moyennes, ses nombres d'impressions, ses nombres de clics et ses coûts. La plupart des études statistiques menées sur les annonces textuelles sont plutôt axées sur les utilisateurs c'est-à-dire qu'on cherche à afficher la « meilleure annonce » à chaque utilisateur. Par meilleure annonce, on entend l'annonce sur laquelle l'utilisateur est le plus susceptible de cliquer. Le modèle proposé ici ne fait pas intervenir les utilisateurs : ce qui est très utile car les informations sur ces derniers ne sont pas toujours disponibles. Une modélisation des variables explicatives utilisées dans notre modèle à savoir la position, les impressions et le coût est nécessaire car tout comme le taux de clics, elles ne sont pas connues à l'avance. En effet, tout comme au jour j le taux de clics d'un mot clé au jour suivant $j+1$ est inconnu, la position, le nombre d'impressions et le coût sont également inconnus au jour $j+1$. Ainsi un modèle est ajusté sur chacune de ces trois variables : une loi normale tronquée est ajustée aux positions ; un modèle de type *hurdle* logistique et binomial négatif est choisi parmi une sélection de modèles pour estimer les impressions. Enfin pour le coût, le modèle *hurdle* logistique et Gamma est retenu.

Dans la suite de ce mémoire, une revue de la littérature est d'abord présentée au chapitre 2. En plus de présenter l'état de l'art, elle permet de mettre en exergue l'originalité de notre recherche. Ensuite le chapitre 3 est consacré à la structure du modèle de prévision. Une brève mais complète présentation de la régression logistique permet d'introduire et de détailler le modèle de prévision des taux de clics. Puis nous décrivons les jeux de données dont nous disposons et sur lesquelles nos méthodes seront appliquées. Les modèles intermédiaires utilisées pour estimer les variables explicatives sont décrits dans le chapitre 4. Enfin les résultats de l'application de notre méthode sur les jeux de données sont discutés au chapitre 5 ; ils seront suivis d'une conclusion au chapitre 6.

CHAPITRE 2 REVUE DE LITTÉRATURE

L'essor de la publicité en ligne s'est accompagné d'une augmentation des activités de recherche sur le sujet. Il s'agit d'un domaine relativement récent puisqu'il apparaît avec le développement d'Internet au début des années 2000. Aujourd'hui on compte un grand nombre d'articles et de publications sur la publicité en ligne et les annonces textuelles notamment.

D'abord il faut noter qu'il existe une diversité de sujets d'intérêts dans le domaine des annonces textuelles. Nous avons par exemple l'optimisation des enchères qui consiste à trouver la valeur d'enchère optimale à attribuer à un mot clé afin d'atteindre une position souhaitée. On peut citer entre autres les travaux de Borgs et al. (2007), de Zhou et al. (2008). Aussi, maximiser le nombre de clics total d'une campagne sous une contrainte budgétaire est également beaucoup étudié dans la littérature. Archak et al. (2010) proposent une méthode basée sur des processus de décision de Markov ; DasGupta et Muthukrishnan (2013) quant à eux utilisent des méthodes d'optimisation stochastique.

Quinn (2011) d'abord et Assari (2014) ensuite ont travaillé sur ces deux problématiques. En effet dans ces deux mémoires de maîtrise (réalisés à Polytechnique Montréal), ils s'intéressent à l'optimisation des campagnes publicitaires, c'est-à-dire maximiser le rendement des annonceurs. Assari (2014) introduit des techniques de fouille de données afin d'améliorer le modèle proposé par Quinn (2011). Leur problématique est totalement différente de la nôtre puisque c'est la prédiction du taux de clics qui nous intéresse. Dans la littérature récente, la prédiction des taux de clics est le sujet principalement traité. En effet une connaissance du taux de clics facilite la résolution des problématiques précédentes.

Le taux de clics est une quantité très importante à la fois pour les annonceurs et les moteurs de recherche. Il sert d'indice de qualité pour les annonceurs ; il permet à ces derniers d'ajuster leurs campagnes, modifier les valeurs d'enchère de certains mots clés ou encore de retirer des mots clés très peu pertinents. Quant aux moteurs de recherche, il utilise le taux de clics dans leur algorithme d'affichage des annonces. C'est une des variables qui permet de déterminer l'ordre d'apparition des annonces. C'est ainsi que les principaux articles publiés sur la prédiction des taux de clics proviennent des compagnies propriétaires des moteurs de recherche. Toutefois en raison de la compétition entre ces différentes compagnies, celles-ci ne publient qu'une petite partie de leur travaux.

La régression logistique est l'un des principaux modèles utilisés pour prédire le taux de clics. Richardson et al. (2007), Chapelle et al. (2015) appliquent la régression logistique. Chapelle et al. (2015) utilisent uniquement des variables qualitatives ; ils incluent les informations sur

les annonces, les annonceurs, les utilisateurs et également le temps. Ils disposent également des données pour chaque impression ; aucune agrégation des données n'est opérée. À chaque impression, on sait si l'utilisateur clique ou non sur l'annonce. C'est la différence fondamentale entre les recherches menées par les moteurs de recherche et celles menées par des chercheurs indépendants. Les premiers disposent de données beaucoup plus riches.

Chez Google également on utilise la régression logistique puisque McMahan et al. (2013) proposent l'algorithme FTRL-proximal (*Follow The Proximally Regularized Leader*) qui permet d'obtenir un modèle « creux ». En effet, l'ajustement d'un modèle logistique fournit un vecteur de paramètres pleins, c'est-à-dire que peu de paramètres sont nuls. L'algorithme proposé permet d'avoir plus de paramètres nuls sans toutefois trop dégrader le modèle ; il améliore également la convergence. McMahan et al. (2013) suggèrent également des techniques pour améliorer les implémentations notamment une économie de mémoire. En effet les moteurs de recherche ont des milliers voire des millions d'annonceurs, donc beaucoup de données à manipuler. Ainsi les modèles creux sont très intéressants.

Outre la régression logistique, les moteurs de recherche emploient également d'autres méthodes pour prédire le taux de clics. Graepel et al. (2010) utilisent la régression probit qui est aussi un modèle binomial comme le modèle logistique. Ils utilisent essentiellement des données catégoriques. Zhu et al. (2010) proposent une méthode basée sur les réseaux bayésiens pour notamment prédire le taux de clics des mots clés très peu utilisés. Chez Yandex, un moteur de recherche russe, on utilise plutôt des arbres « boostés » (*boosted trees*) (Trofimov et al., 2012) qui sont une modification des machines à gradient « boosté ». Un réseau de neurones est ensuite introduit pour traiter les variables catégoriques : Baqapuri et Trofimov (2014) utilisent un réseau de neurones qui prend les variables catégoriques en entrée et calcule une première estimation du taux de clics. Puis le modèle des arbres « boostés » utilise cette estimation et les autres variables pour finalement prédire le taux de clics.

Jusqu'ici nous avons présenté différents modèles disponibles dans la littérature. Cependant il est presque impossible de les comparer car d'une part les données utilisées ne sont pas disponibles. D'autre part chaque moteur de recherche dispose de son propre algorithme d'affichage des annonces. Il est alors très probable que le modèle de Yandex échoue pour des données issues du moteur de recherche de Google par exemple.

Parallèlement aux travaux des moteurs de recherches, des recherches indépendantes sont également menées notamment par les entreprises de gestion de campagnes publicitaires. Comme nous le soulignions plus haut, ici les données disponibles sont beaucoup moins riches puisqu'elles sont fournies par les moteurs de recherches sous forme agrégées. Par exemple, pour un mot clé on dispose non pas des informations pour chaque impression mais plutôt une

moyenne de ces informations sur une certaine période donnée (une journée généralement). La qualité de ces données rend la tâche de prédiction ici beaucoup plus compliquée. C'est en partie pour cela que ce sujet a longtemps été contourné en formulant les problématiques autrement telles que la maximisation du nombre de clics. Ici aussi les plateformes de gestion de campagnes ne sont pas très enclines à partager leurs travaux. Toutefois on retrouve quelques articles très intéressants.

Kumar et al. (2015) utilisent la régression logistique et surtout introduisent la position comme une variable explicative. Lee et al. (2012) proposent d'utiliser un modèle pour traiter l'emboitement des variables catégoriques. Même s'ils s'intéressent aux conversions, leur modèle reste transposable aux clics. L'émergence de l'apprentissage profond a conduit Jiang (2016) à proposer un modèle combinant un *Deep belief network* et un modèle logistique. Tout comme Baqapuri et Trofimov (2014), le réseau de neurones permet ici de tirer de l'information des variables catégoriques. Au vu des résultats des réseaux de neurones notamment les réseaux profonds sur différentes tâches de l'apprentissage automatique, nous pensons que de plus en plus de modèles basés sur les réseaux de neurones seront proposés pour prédire les taux de clics. Ces modèles sont beaucoup plus complexes mais difficiles à interpréter. Enfin on peut citer le travail de Lee et al. (2016) qui proposent des modèles linéaires avec noyau comme alternative au modèle logistique.

En somme le modèle logistique est un modèle de référence pour prédire les taux de clics. D'un côté, nous avons les moteurs de recherche qui disposent de données détaillées et de l'autre des chercheurs indépendants avec des données moins fournies. Nous nous situons dans le deuxième groupe. La plupart des méthodes disponibles dans la littérature utilisent principalement des variables catégoriques. Or ne disposant pas des données sur ces variables au départ de notre projet, nous avons donc bâti un modèle qui n'utilise quasiment pas ces dernières même si par la suite ces données furent disponibles. Nous utilisons également le modèle logistique. Néanmoins nous proposons des modèles pour les variables intermédiaires que sont la position moyenne du mot clé, son nombre d'impressions ou encore le coût d'un clic. Cette méthode d'estimation du taux de clics est assez inédite.

Dans le chapitre suivant, nous faisons une introduction détaillée au modèle logistique puis nous présentons notre modèle de prédiction des taux de clics.