HAL open science

Quelques pas vers l'Honnêteté et l'Explicabilité de moteurs de recherche sur le Web

Philippe Mulhem, Lydie Du Bousquet, Sara Lakah

To cite this version:

Philippe Mulhem, Lydie Du Bousquet, Sara Lakah. Quelques pas vers l'Honnêteté et l'Explicabilité de moteurs de recherche sur le Web. COnférence en Recherche d'Information et Applications, Mar 2019, Villeurbanne, France. 10.3166/DN.1.-.1-18 . hal-02061553

HAL Id: hal-02061553

https://hal.science/hal-02061553v1

Submitted on 27 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quelques pas vers l'Honnêteté et l'Explicabilité de moteurs de recherche sur le Web

Philippe Mulhem, Lydie du Bousquet, Sara Lakah

Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP [1]*, LIG, 38000 Grenoble, France*
Philippe.Mulhem@imag.fr,{lydie.du-bousquet,sara.lakah}@univ-grenoble-alpes.fr

RÉSUMÉ. *La transparence des algorithmes est un sujet de préoccupation pour les utilisateurs et les autorités. Parmi les différents aspects de cette notion de transparence, est-il possible d'étudier dans quelle mesure les moteurs de recherche sur le web sont honnêtes par rapport à leur politique de personnalisation déclarée, et dans quelle mesure est-il possible d'expliquer leur comportement, ne serait-ce que succinctement ? Cet article décrit un cadre expérimental pour étudier ces aspects, et des résultats obtenus sur l'étude du principal moteur de recherche sur le Web. L'idée suivie est d'étudier les différences dans les résultats du moteur, afin de mesurer leur honnêteté et leur explicabilité. Nous avons formalisé le protocole expérimental et les mesures d'évaluations. Nous trouvons, pour plusieurs utilisateurs et plusieurs requêtes choisies avec soin, qu'il y a de larges variations dans les résultats. Nous les quantifions et nous les ordonnons l'importance relative des paramètres étudiés. Nous montrons que ces informations seraient utiles pour améliorer la transparence des moteurs de recherche sur le Web.*

ABSTRACT. *The transparency of algorithms is a concern for users and authorities. Among the various aspects of this notion of transparency, is it possible to study the extent to which search engines on the web are honest with respect to their declared personalization policy, and to what extent is it possible to explain their behavior, if only succinctly? We describe here an experimental framework and the results obtained on the study of the transparency of the leading web search engine. The idea is to track the changes of results, in a way to evaluate their honesty and explicability. We formalize the protocol and the measures used. We find that, for several users and carefully choosen queries, they are large variations in the results. We quantify them and rank the importance of the parameters studied. We show that this information may be used to enhance the transparency of search engines.*

MOTS-CLÉS : *Transparence, Honnêteté, mesure d'évaluation*

KEYWORDS: *Transparency, Honesty, evaluation measure*

DOI:10.3166/DN.1.-.1-18 © 2018 Lavoisier

1. Institute of Engineering Univ. Grenoble Alpes

Document numrique
n° -/2018, 1-18

1. Introduction

A l'heure où le fonctionnement de notre société repose de plus en plus sur les systèmes informatiques, les citoyens éprouvent un besoin croissant de savoir comment ces systèmes se comportent, et les raisons de ces comportements (Mittelstadt *et al.*, 2016). Cette idée se traduit par la notion de transparence. Dans (Turilli, Floridi, 2009), la transparence se réfère *aux formes de visibilité de l'information, et concerne les possibilités d'accéder à l'information, aux intentions et aux comportements qui sont révélés au travers d'un processus de divulgation*. Les facettes de la transparence que nous étudions ici se concentrent sur des dimensions liées à la notion de "preuve invérifiable" (inscrutable evidence) définie dans (Mittelstadt *et al.*, 2016), qui stipule que la connexion entre les données et les résultats devrait être *accessible* et *intelligible*.

Nous nous intéressons au problème de la transparence de la personnalisation dans les moteurs de recherche sur le Web. On se confronte alors au besoin de réconcilier des éléments difficilement compatibles (Turilli, Floridi, 2009) : le fait d'un côté que la transparence doit permettre de montrer et/ou d'expliquer des choses, et de l'autre qu'elle est contrainte par des considérations de propriété intellectuelle (savoir-faire et secret industriel de l'entreprise).

Pour aborder ce problème, nous confrontons la politique de transparence de personnalisation déclarée par un système à son comportement observable, en :

1. étudiant et caractérisant que le système (moteur) *dit ce qu'il fait et fait ce qu'il dit*, ce qui correspond au respect de sa politique de personnalisation. Nous appellerons cette propriété "honnêteté" par la suite. D'une certaine façon, vérifier l'honnêteté est comparable au processus de vérification qu'un système satisfait sa spécification.

2. quantifiant les facteurs qui impactent le résultat produit par le système pour fournir une aide à la compréhension du résultat en complément de sa politique. Ce que nous appellerons "explicabilité".

Dans la suite, la section 2 décrit des propositions connexes à nos travaux. La section 3 propose une description succincte des moteurs de recherche classiques, et donne quelques éléments liés aux pratiques d'utilisation des informations personnelles. Nous formalisons ensuite les propriétés attendues Section 4. Le protocole d'évaluation est proposé section 5, les expérimentations menées et les résultats obtenus en section 6, avant de conclure en section 7.

2. État de l'art

Pour les systèmes de recherche de documents, la transparence peut être abordée a priori ou a posteriori. Ainsi, un moteur de recherche peut être *conçu* pour être transparent (Kules *et al.*, 2008). Pour un système existant, on peut chercher à *caractériser* sa transparence. Nos travaux et l'état de l'art qui suit se focalisent sur ce second cas, car ce point suscite déjà des interrogations des utilisateurs (DuckDuckGo, 2018a).

Un certain nombre de travaux connexes à la transparence a posteriori pour la recherche d'information existent. En particulier, des travaux ont porté sur les publicités dans les réseaux sociaux (Andreou *et al.*, 2018 ; Venkatadri *et al.*, 2019), dans lesquels les auteurs ont étudié quels éléments sont utilisés pour le ciblage les publicités, et quels éléments sont expliqués. Les auteurs ont conduit leurs expérimentations en créant des publicités et en ciblant leurs audiences de manière très précise. (Venkatadri *et al.*, 2019) a même été capable de réaliser un rétro-engineering de certaines fonctionnalités de la publicité de Facebook. Dans le cas de recherche de documents sur le Web, les paramètres de la correspondance entre le contenu des documents et les requêtes sont plus diffus et non-contrôlables. De plus, les réponses à une requête sont des listes de documents pour lesquels la pertinence est difficile à estimer.

Un autre pan de travaux sur la transparence a posteriori porte sur la recherche de personnes pour réaliser des tâches. On rencontre typiquement cette problématique sur les plateformes de "crowd sourcing" comme Amazon Mechanical Turks (Singh, Joachims, 2018), pour trouver les "workers" qui vont effectuer un travail. C'est alors davantage l'"équité" qui est étudiée : permettre à tout travailleur d'avoir accès équitablement à des tâches, et aux personnes qui proposent des tâches d'avoir accès à des travailleurs.
Différentes notions d'équité sont possibles, comme l'équité d'attention (Biega *et al.*, 2018). Tous ces travaux se focalisent soit sur des résultats non-triés, soit sur des résultats triés pour lesquels les documents (utilisateurs/locations) et leurs caractéristiques complètes sont connues, et pour lesquels les valeurs de pertinence sont connues par requête. Dans notre cadre, aucune de ces données n'est connue, ce qui ne nous permet pas de réutiliser ces travaux.

Pour la recherche sur le Web par des moteurs de recherche classiques, une part de nos travaux trouvent leur inspiration dans (Hannák *et al.*, 2013 ; 2017), destinés à caractériser les éléments qui influent sur le résultat d'un moteur de recherche. Les travaux ci-dessus se focalisent sur les expérimentations, sans proposer de placer ces expérimentations dans un cadre plus formel de transparence, ce que nous faisons ici. De plus, étant donné les évolutions constantes des moteurs de recherche, il est de plus nécessaire de réaliser des tests de manière récurrente, afin de confirmer ou d'infirmer les résultats passés.

3. Moteurs de recherche, politique et personnalisation

Les moteurs de recherche sont des entités logicielles complexes, supportées par des architectures distribuées, dédiées à fournir des réponses à des requêtes utilisateurs. Par exemple, des estimations (Data Center Knowledge, 2017) chiffrent à 2,5 millions le nombre de serveurs utilisés par Google, sur 15 sites [2], dont une bonne partie est dédiée au moteur de recherche. Par ailleurs, sont utilisés des algorithmes sophistiqués pour répondre aux requêtes des utilisateurs. Ils reposent sur des techniques classiques de

2. https://www.google.com/about/datacenters/inside/locations/index.html

Recherche d'Information (RI), pour les traitements des documents et les pondérations de termes, mais également sur des centaines d'autres facteurs (Dean, 2018) qui sont constamment re-pondérés en prenant en compte les traces utilisateurs, afin de favoriser la satisfaction des utilisateurs dans les premiers résultats (personnalisation).

Chaque moteur de recherche publie des informations relatives à ses pratiques quant à l'obtention de ses résultats, par exemple Google (Google, 2018a), Bing (Microsoft, 2018), Qwant (Qwant, 2016), ou Duckduckgo (DuckDuckGo, 2018b). Dans la suite, on appelle "politique de transparence" (ou plus simplement "politique") ces informations. Une politique de transparence indique en particulier les éléments utilisés lors d'une recherche sur le Web pour la présentation des résultats.

Une politique peut être formulée de manière très générique. Par exemple, en ce qui concerne la personnalisation, Google (Google, 2018a) indique : "De vos paramètres de recherche à votre situation géographique, en passant par l'historique de vos recherches, toutes ces informations nous permettent de vous proposer les résultats les plus pertinents et les plus utiles à l'instant T". De plus, il est précisé que "lorsque vous êtes connecté à votre compte, nous stockons les informations collectées en les associant à votre compte Google et les considérons comme des informations personnelles" (Google, 2018c), c'est-à-dire des "informations que vous nous fournissez et qui permettent de vous identifier, telles que votre nom, votre adresse e-mail, vos informations de facturation ou toute autre information que Google est susceptible d'associer à vous, telles que les informations liées à votre compte Google". A contrario, "par principe, Qwant ne collecte pas de données sur ses utilisateurs lors des recherches" (Qwant, 2017).

Dans cet article, on considérera qu'une politique correspond à la liste des paramètres indiqués comme étant utilisés par le moteur pour produire la réponse à une requête. On les appellera des paramètres *explicites*. Un paramètre influençant la réponse à une requête et non présent dans la politique sera appelé paramètre *implicite*. Un utilisateur peut s'attendre à obtenir des résultats différents s'il formule la même requête avec différentes valeurs de paramètres explicites. Réciproquement, il peut s'attendre à obtenir les mêmes résultats pour la même requête, toutes les valeurs de paramètres explicites étant égales par ailleurs, modulo le fait que le corpus de documents n'évolue pas entre temps (comme cela peut être le cas pour les actualités). Le moteur est honnête par rapport à sa politique si les résultats des requêtes ne sont pas influencés par des paramètres implicites. Selon les points de vue, un moteur indiquant qu'il utilise des paramètres sans les nommer peut être ou non considéré comme honnête du point de vue légal, et non honnête du point de vue utilisateur. Dans tous les cas, il n'est pas suffisamment détaillé.

◐ Dans la suite, nous illustrerons nos propos à l'aide d'un moteur de recherche fictif que nous appellerons Foo-Q. Avec Foo-Q, l'utilisateur peut faire ses recherches en mode connecté ou non. Lors de la création d'un compte, il doit renseigner son genre (homme, femme, non-spécifié) et son âge. La politique de personnalisation de Foo-Q indique que "les résultats des requêtes sont calculés à partir du corpus de documents

de Foo-Q et ajustés en fonction du genre et du pays de l'utilisateur". Le genre et le corpus sont donc des paramètres explicites de la politique. ◗

4. Formalisation des propriétés attendues

L'objectif de notre travail est de discuter et d'évaluer la transparence de la politique affichée d'un moteur de recherche, au travers de l'honnêteté et de l'explicabilité. Une recherche de documents à l'aide d'un moteur de recherche s'effectue dans un *contexte* donné (e.g., corpus de documents, date et heure de la requête, localisation, etc.). Ainsi, une requête q posée à un moteur de recherche dans un contexte c produit théoriquement une liste $l_{théo}$ de couples <$d, rsv(q, c, d)$> triée par valeurs de $rsv(q, c, d)$ décroissante. La fonction $rsv(q, c, d)$ (pour *Retrieval Status Value* (Nottelmann, Fuhr, 2003)) associe à un document d un nombre réel dénotant la pertinence calculée pour le document pour la requête : plus la valeur est élevée, plus le document est pertinent d'après le système. Dans le cas des résultats de moteurs de recherche, les valeurs de pertinences ne sont pas fournies, il en résulte que l'utilisateur n'obtient qu'une liste $l_{réel}$ qui ne contient que des URLs de documents.

Le contexte est un vecteur à plusieurs dimensions; chacune correspond à un paramètre qui peut théoriquement influencer le résultat de la recherche. Certains moteurs utilisent plus de 200 paramètres (Dean, 2018). ◖ Pour illustrer nos propos, nous considérerons un contexte limité à 5 paramètres : <corpus, genre, âge, pays, date>. ◗

Pour évaluer qu'un moteur est honnête par rapport à un paramètre, il est nécessaire de mesurer l'impact de ce paramètre sur les résultats des requêtes. Pour cela, nous devons définir une mesure Δ permettant d'évaluer la différence (≥ 0) entre deux résultats, en tenant compte des caractéristiques de ces listes. Nous discuterons plus tard du choix de cette mesure.

Dans l'absolu, si un moteur est honnête, alors quelques soient les requêtes, les résultats obtenus ne sont influencés que par les paramètres de sa politique. Si une requête q est exécutée pour deux instanciations de contextes qui ne diffèrent que pour un paramètre explicite, les éventuelles différences dans les résultats sont compatibles avec la politique. ◖ Pour Foo-Q, si un homme et femme obtiennent des résultats différents pour la même requête alors qu'ils sont connectés, cela s'explique par la politique du moteur (cf. fin sect. 3). ◗

Par contre, si la requête q est exécutée pour deux instanciations de contextes qui ne diffèrent que pour un paramètre hors politique, aucune différence ne devrait être observée. Mais si une différence est observée, on ne peut que *soupçonner* que la politique n'est pas décrite de façon exhaustive. En effet, il n'est pas possible de conclure de manière définitive, car l'instanciation d'un contexte n'est pas entièrement perceptible du point de vue de l'utilisateur (SearchLiaison, 2018). ◖ Si deux hommes de 20 et de 60 ans obtiennent des résultats différents pour la même requête effectuée simultanément sur Foo-Q, les utilisateurs peuvent légitimement penser que l'âge de l'utilisateur est

aussi utilisé pour calculer le résultat. Néanmoins la requête peut avoir été exécutée sur deux serveurs différents ayant des index différents à cet instant par exemple. ◗

Dans la suite, on ne cherche pas à définir formellement l'honnêteté, mais à évaluer l'importance d'un paramètre explicite, ou à caractériser les paramètres implicites pour lesquels on peut soupçonner qu'ils devraient figurer dans la politique (suspicion de non-exhaustivité).

Définition 1 : Impact d'un paramètre selon Δ pour une requête

Soit Δ un opérateur qui évalue la différence entre deux listes. Soient q une requête et $C =< p_1, ..., p_n >$ un contexte. Soient c_1 et c_2 deux instanciations de C ne se différenciant que pour p_i. Soient $r(q, c_1)$ et $r(q, c_2)$ les résultats respectifs obtenus en exécutant q avec c_1 et c_2.
$\Delta(r(q, c_1), r(q, c_2))$ est appelé impact de p_i sur la requête q.

◖ Δ (r("toto",<D,homme,20,France,1/09/2018>), r("toto",<D, femme,20,France, 1/09/2018>) quantifie l'impact du paramètre "genre" sur la requête "toto" dans le contexte <D,*genre*,20,France,1/09/2018>. ◗

Définition 2 : Impact moyen d'un paramètre selon Δ pour un ensemble de requêtes Q

Soient $C =< p_1, ..., p_n >$ un contexte. Soient $r(q^j, c_1^j)$ et $r(q^j, c_2^j)$ les résultats respectifs obtenus en exécutant des requêtes $q^j \in Q$ avec c_1^j et c_2^j ne se différenciant que pour $p_i \in C$. On dit que le paramètre p_i a une *impact moyen* de :
IM$_\Delta(p_i)$ = $\sum_{j=1..nb}(\Delta(r(q^j, c_1^j), r(q^j, c_2^j))/nb$.

◖ Considérons un ensemble de deux requêtes $Q = \{"titi", "toto"\}$, et deux contextes se différenciant que par le genre de l'utilisateur connecté.

$$\begin{array}{ll} \Delta(& r(\text{“toto”}, < \text{D}, \text{homme}, 20, \text{France}, 1/09/2018 >), \\ & r(\text{“toto”}, < \text{D}, \text{femme}, 20, \text{France}, 1/09/2018 >) \ ) = 0{,}7 \\ \Delta(& r(\text{“titi”}, < \text{D}, \text{homme}, 20, \text{France}, 1/09/2018 >), \\ & r(\text{“titi”}, < \text{D}, \text{femme}, 20, \text{France}, 1/09/2018 >) \ ) = 0{,}6 \end{array}$$

On obtient un impact moyen du paramètre IM$_\Delta$(genre) de 0,65 pour Q. ◗

Notons que l'impact moyen dépend du contexte expérimental puisque sa valeur varie en fonction du nombre et des requêtes choisies. Il faut donc considérer un contexte expérimental varié et suffisamment vaste pour obtenir des différences significatives, au sens statistique (Buckley, Voorhees, 2017). Nous en rediscuterons plus loin.

L'impact moyen d'un paramètre peut être utilisé de deux façons. Tout d'abord, il permet de fournir des éléments d'explication vis-à-vis d'une politique, en particulier en quantifiant l'influence relative des différents paramètres explicites. Ensuite, l'impact moyen permet d'étudier l'influence d'un paramètre implicite, et de conclure éventuellement à une suspicion de non exhaustivité de la politique.

◐ On mène une expérimentation pour étudier la politique de Foo-Q, comme décrite dans la suite de cet article. On obtient un impact moyen de 0,6 pour le paramètre "pays" et de 0,4 pour "genre". On peut alors supposer que les algorithmes de Foo-Q accordent plus d'importance au pays qu'au genre des utilisateurs. Par ailleurs, si on obtient un impact moyen de 0,3 du paramètre "âge", on peut suspecter que la politique est non exhaustive par rapport à ce paramètre. ◑

5. Principe du protocole d'évaluation

L'analyse d'un paramètre passe par l'évaluation de l'impact moyen. Ce dernier repose sur la capacité à exécuter des couples de requêtes pour deux instanciations de contexte ne se différenciant que pour une valeur, et de mesurer la différence de résultat. Dans la suite, nous discutons successivement du choix de la mesure (5.1), du contexte (5.2), des stratégies pour limiter le biais de l'expérimentation par rapport au choix des requêtes (5.3) ou du temps (5.4), et du cadre d'interprétation (5.5).

5.1. Mesures de différence entre listes : du choix de Δ

La définition de l'impact moyen s'appuie sur la capacité à évaluer la différence entre deux listes de résultats de requêtes. Rappelons que chaque résultat est une liste d'URLs, et que l'ordre des réponses est important (les premières réponses sont les plus susceptibles d'êtres lues (Joachims, Radlinski, 2007)). Nous explorons les mesures de l'état de l'art (Webber *et al.*, 2010) suivantes pour la fonction Δ :

– Le coefficient de Jaccard (Jac) est capable de prendre en compte des listes de tailles possiblement différentes, et pouvant contenir des éléments différents, mais a le désavantage de ne pas tenir compte de l'ordre des éléments dans les listes comparées. La mesure de Jaccard est une similarité donnant des résultats réels dans $[0, 1]$. Pour mesurer une différence nous utilisons $JAC = 1 - Jac$.

– La mesure de corrélation Tau de Kendall (Tau) intègre le fait que l'on a des éléments triés dans des listes et considère que chaque interversion à la même importance. Elle nécessite que les deux listes contiennent les mêmes éléments (une solution classique, utilisée dans cet article, est de ne pas considérer dans les calculs les éléments présents uniquement dans une liste). Elle donne des résultats réels dans $[-1, 1]$. Pour forcer les valeurs dans l'intervalle $[0, 1]$, et en faire une mesure de différence, nous appliquons la fonction linéaire : $TAU = 0,5 - 0.5 \times Tau$.

– La mesure de recouvrement biaisé par le rang (Rank-Biased Overlap, Rbo) proposée par Webber, Moffat et Zobel a pour intérêt de prendre en compte des listes ayant des éléments différents, et tient compte du fait que les différences en début de liste sont plus importantes qu'en fin de liste. Dans la suite, nous utiliserons le Rbo extrapolé, Rbo_{EXT} [3], qui repose sur l'hypothèse que l'accord entre deux listes, calculé à une profondeur k quelconque, se prolonge après k. C'est une mesure de similarité

3. Par simplicité, dans la suite nous ne garderons que le nom générique Rbo pour Rbo_{EXT}.

dans l'intervalle $[0, 1]$. Comme pour Jaccard, on se base sur $RBO = 1 - Rbo$, pour mesurer une différence.

Ci-après nous illustrons le comportement de ces trois mesures. Pour cela, nous avons créé une liste de référence de 100 éléments (des chaînes de caractères dans [“AAA”, “AJJ”]); et nous avons produit des listes modifiées, soit en permutant les éléments aux positions k et l, soit en remplaçant l'élément à la position n par “ZZZ”.

La figure 1 (partie gauche) présente le résultat des trois mesures pour des *permutations* $\in [2, 100]$. On remarque que le coefficient de Jaccard n'est pas en mesure d'établir de différences. La différence basée sur la valeur de Tau croît linéairement en fonction de la distance de l'élément permuté, alors que la valeur RBO suit une courbe logarithimique. Cette croissance non-linéaire est intéressante : elle capture le fait que la différence entre les permutations $1 \longleftrightarrow 5$ et $1 \longleftrightarrow 10$ est plus importante qu'entre $1 \longleftrightarrow 75$ et $1 \longleftrightarrow 80$ (respectivement $1,33 \times 10^{-3}$ et $2,99 \times 10^{-4}$). La figure 1 (partie droite) présente l'impact du *remplacement* d'un élément en position n de la liste de référence par un élément qui n'appartient pas à cette liste. Ici encore, la mesure de Jaccard est constante. Comme précédemment, le comportement de RBO présente une évolution non-linaire, contrairement à TAU.

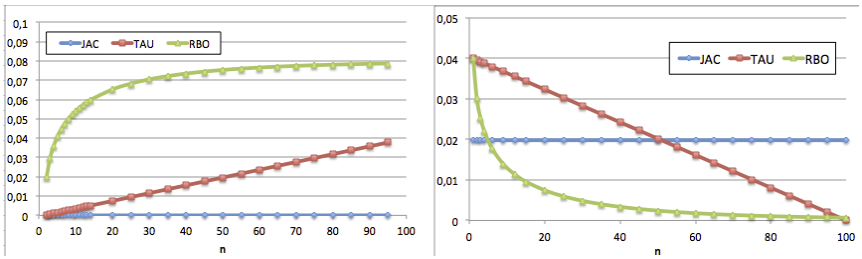

Figure 1. Différences entre la référence et des listes avec permutation de position entre 1 et l (à gauche) et insertion d'un élément externe en position n (à droite).

La conclusion que l'on tire de ces expérimentations sur des données générées est que la mesure de différence Δ basée sur RBO permet à la fois de distinguer des listes proches, mais aussi de quantifier les différences en fonction des positions auxquelles ces différences arrivent. C'est pourquoi nous l'utilisons dans la suite.

Pour affiner nos évaluations, nous utilisons RBO en se focalisant sur les 10, 20, 30, 40 et 50 premiers éléments des listes (notés respectivement RBO@10, ..., RBO@50). Ce choix correspond aux nombres classiques de documents présentés par les moteurs de recherche sur la première page de résultats, sur les deux premières pages, etc. En considérant la mesure RBO@10, aucune *permutation* entre la position 1 et n avec $n < 10$ ne donne une valeur supérieure à 6%, et seuls les *remplacements* entre les positions 1 et 5 donnent une valeur supérieure à 10%.

5.2. Explicitation du contexte

Le but de notre protocole expérimental est de permettre d'évaluer au mieux l'impact moyen d'un paramètre donné sur les résultats d'une requête. Il est nécessaire pour cela de limiter au mieux l'influence des autres paramètres, qu'ils soient implicites ou explicites. La première chose à faire consiste donc à recenser l'ensemble des paramètres susceptibles d'être utilisés par le moteur. Ceux décrits dans (Dean, 2018) sont essentiellement liés au corpus de documents, c'est-à-dire aux caractéristiques de chaque page (titre, méta-balises, URL, nombre de liens, etc). Ils sont donc par définition non-contrôlables par l'utilisateur et donc par le protocole.

Dans la suite, nous reprenons les paramètres contrôlables identifiés dans (Hannák *et al.*, 2013), relatifs à l'environnement d'exécution : **matériel**, **système d'exploitation**, **navigateur**, **cookie**, **historique de navigation**; à la géolocalisation : **adresse IP**; au profil utilisateur : **genre**, **âge**, **code postal**. Nous ajoutons **la date et l'heure**.

Quand un paramètre est étudié, il faut définir les valeurs qui vont être explorées. Elles dépendent de l'objet d'étude. Par exemple, pour étudier l'impact de l'âge, on peut distribuer les valeurs d'âge uniformément sur l'intervalle $[0, 100]$, ou choisir des valeurs représentatives de tranches d'âge (junior/senior) pour étudier leur impact.

5.3. Gestion du biais relatif au choix des requêtes

Les algorithmes des moteurs de recherche sont de plus en plus sophistiqués pour offrir une réponse personnalisée à l'utilisateur. En particulier, les réponses à certains sujets de requêtes sont connus pour dépendre fortement de la localisation ou du genre de l'utilisateur (Jansen, Booth, 2010). De ce fait, on peut distinguer les sujets de requête dont le traitement est susceptible d'être influencé par la fonction de personnalisation, des autres. Pour étudier l'impact de la personnalisation (e.g. le genre ou la localisation), il faut donc choisir des requêtes dans les deux ensembles de sujets, pour pouvoir comparer les deux.

Une part de subjectivité entre en jeux ici : il est dès lors obligatoire de créer un consensus autour des requêtes, en se basant sur des experts. Les requêtes doivent être diverses en terme de sujet, et assez nombreuses pour éviter les biais logiciels/matériels (tout comme dans les évaluations classiques en recherche d'information (Buckley, Voorhees, 2017)). Ces requêtes peuvent-être extraites de véritables requêtes : (Hannák *et al.*, 2013 ; 2017) ont utilisé Google Trends (Google, 2018b)).

5.4. Gestion du biais relatif au temps

Le corpus est un élément du contexte. Plus le délai δ_t entre deux requêtes est important, plus le corpus est susceptible de changer (SearchLiaison, 2018), et donc de biaiser les résultats quant à l'impact d'un paramètre. Il convient donc d'exécuter les requêtes pour deux instanciations de contextes dans un délai très court. Il est alors

préférable d'automatiser la mise en oeuvre du protocole. Ce qui permet par ailleurs d'éviter des erreurs dues à une intervention manuelle (Hannák *et al*., 2017).

Pour autant, réduire le délai entre deux exécutions de requête à son minimum présente un risque. Classiquement les moteurs tentent de détecter des utilisations "automatisées" pour les interdire. Afin de diminuer ce risque de détection, nous forçons une exécution temporisée des requêtes. Un second élément qui plaide pour la temporisation des requêtes est d'éviter de lancer des requêtes en rafales : elle peuvent être interprétées par le moteur comme des sessions (le *carry-over* dans (Hannák *et al*., 2013)), ce qui peut introduire un biais.

Bien qu'il soit impossible que deux requêtes soient traitées exactement au même moment, on va supposer que si des requêtes sont posées dans un intervalle de temps δ_t court (*à définir*), l'impact du temps est négligeable par hypothèse. Dans le cas où l'impact de nombreux paramètres (et/ou de nombreuses valeurs de paramètres) sont étudiés, il faut faire des choix pour garantir qu'une expérimentation pour laquelle le temps est supposé non-impactant soit lancée dans un intervalle de temps $\leq \delta_t$.

5.5. Cadre d'interprétation

Comme nos expérimentations menées calculent des impacts moyens, on peut les caractériser de manière absolue ou relative. Un impact absolu se base sur la valeur de Δ pour caractériser si un paramètre a une grande importance (Δ proche de 1) ou une importance faible (Δ proche de 0). Si on veut évaluer de manière relative les impacts de deux paramètres, on peut comparer relativement les valeurs respectives de Δ.

Dans tous les cas, il est important de se poser des questions sur la qualité des résultats obtenus. Dans le cas de caractérisations absolues, qui sont en fait des moyennes, on peut utiliser des moments d'ordres supérieurs comme l'écart-type, le coefficient d'asymétrie ou le kurtosis pour caractériser la symétrie ou la dispersion des données. Les comparaisons relatives peuvent utiliser des tests de significativité statistiques, en extrapolant des travaux sur la recherche d'information (Smucker *et al*., 2007).

6. Expérimentations

Nous avons expérimenté notre protocole sur "Google search", car ce moteur a une position dominante et il utilise explicitement des informations personnelles dans sa politique. Nous détaillons ci-après nos choix et nos résultats.

6.1. Paramètres, Contexte et requêtes

La politique de "Google search" indique que le fait d'être connecté ou non à un compte Google pour effectuer une recherche influence le résultat (paramètre **connexion** $\in$ {vrai, faux}). Le fait d'être connecté impose à l'utilisateur de spécifier son **genre** ($\in$ {homme, femme}). Il est donc impossible de séparer l'étude du paramètre

connexion de celui du genre. Dans la suite, nous étudierons le paramètre **utilisateur** $\in$ {homme, femme, anonyme}; anonyme représentant l'utilisateur non connecté et homme/femme, les utilisateur connectés. Par ailleurs, comme le temps est un paramètre important, nous avons exécuté les requêtes deux fois par jour sur trois jours (du 29 mai au 31 mai 2018), à 8:00 am et 8:00 pm, (numérotés de t_1 à t_6).

Pour les autres paramètres du contexte, nous avons fixé les valeurs suivantes : le **matériel** est un ordinateur portable, avec le **système d'exploitation** Windows NT 6.1, le **navigateur** Mozilla 5.0 configuré en langue anglaise, sans **cookies** et sans **historique de navigation**. La plage d'**adresse IP** utilisée est celle de notre laboratoire. Deux **profils utilisateurs** ont été créés sur le formulaire anglophone de Google, avec une date de naissance au 1er janvier 2000, avec des numéros de téléphone de contact différents (français), sans autre information personnelle (code postal non défini, pas d'adresse mail de contact). La fonction de suivi de l'activité de l'utilisateur par Google était activée pour les deux comptes, mais pas la fonction de localisation.

Conformément au protocole défini précédemment, quatre experts ont choisi, à l'unanimité, 15 requêtes (notés Q) en anglais sur trois sujets très différents : le sport, les vêtements et la paléontologie. Les deux premiers sujets sont susceptibles d'être influencés par le genre de l'utilisateur, contrairement au troisième :

- Vêtements : jeans, dress, bra, suits, underwear boxer.
- Sport: soccer, softball, ballet, shooting, tennis mixed doubles.
- Paléontologie: mammoth, cretaceous, darwin, mary anning, ammonites.

Ces 3 sous-ensembles sont notés Q_{vet}, Q_{spo} et Q_{pal}. Toutes les expérimentations ont été menées en utilisant des scripts pour PhantomJS (contributors, 2018), similairement à (Hannák *et al.*, 2013 ; 2017). Ce logiciel implante un navigateur Web contrôlable par programme, indique la configuration du client, gère les cookies, interprète les CSS et exécute le Javascript. Il simule une véritable interaction en se connectant (ou non) au moteur, en générant des requêtes et en extrayant les URLs des pages de résultats.

6.2. Résultats

Nous avons donc exécuté les 15 requêtes pour les 3 utilisateurs toutes les douze heures. Bien que limitée, l'expérimentation nous a permis d'apporter quelques éléments de réponses par rapport à la transparence.

Selon le paramètre étudié, on ne compare pas les mêmes couples de résultats. Pour évaluer l'influence du **temps**, nous calculons IM_{RBO}(temps), l'impact moyen selon RBO, pour chaque utilisateur et chaque requête, en comparant les résultats obtenus à t_i et t_{i+1} ($i \in [1..5]$). Pour évaluer l'influence de l'**utilisateur**, nous comparons les résultats obtenus pour chaque requête exécutée au même moment t_i, entre deux utilisateurs ($i \in [1..6]$) avec IM_{RBO}(utilisateur).

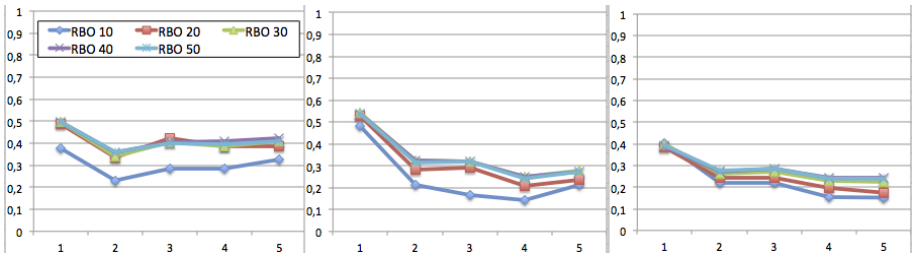

Figure 2. Impact moyen du temps de t_i à t_{i+1} (basé sur RBO@∈ [10, 50]), calculé par sujet, pour tous les utilisateurs confondus. Requêtes sur les **vêtements** *(gauche), le* **sport** *(milieu), et la* **paléontologie** *(droite).*

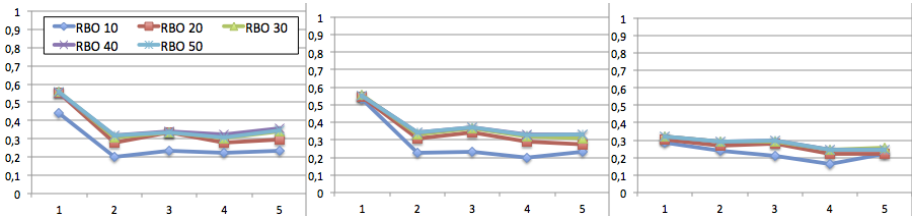

Figure 3. Impact moyen de t_i à t_{i+1} (basé sur RBO@∈ [10, 50]), par utilisateur, toutes requêtes confondues **femme** *(gauche);* **homme** *(milieu);* **anonyme** *(droite).*

Étude du paramètre temps

La figure 2 présente l'impact moyen du temps pour les requêtes, par sous-ensemble de requêtes (Q_{vet} / Q_{spo} / Q_{pal}), en faisant la moyenne pour tous les utilisateurs. Dans cette figure, un élément primordial à signaler est que, en moyenne, les valeur de RBO@10 entre les mêmes requêtes sur un intervalle d'une demi-journée ne descendent pas à moins de 10%, ce qui veut dire que de grandes différences existent dès la première page de résultats (voir figure 1). Ceci est confirmé par des différences JAC non reportées ici par manque de place, qui indiquent que ces différences ne sont pas dues qu'à des permutations. Il y a donc bien des paramètres qui influent sur les résultats de ce moteur de recherche. On remarque aussi que les différences entre les listes sont globalement plus élevées pour les mesures de RBO@x avec x∈ $[20, 50]$ que pour RBO@10. Ceci est cependant moins visible pour les requêtes sur la paléontologie que pour les deux autres catégories. Les valeurs de RBO entre 20 et 50 documents pour un même repère de temps sont très proches dans tous les cas. Ceci signifie qu'il y a plus de différences sur les documents classés de 20 à 50 que sur les dix premiers; et ces différences sont plus importantes pour le sport et les vêtements que pour la paléontologie. Pour la paléontologie, les différences entre les mesures au même moment sont faibles ($< 10\%$), et elles ont tendance à s'agrandir au cours du temps.

La figure 3 synthétise IM_{RBO}(temps) de t_i à t_{i+1} par utilisateur, en considérant toutes les requêtes. Ceci permet d'avoir une vision globale du comportement du mo-

teur par utilisateur. On y constate que les différences à $i = 1$ (donc entre t_1 et t_2) pour les utilisateurs connectés sont plus importantes que pour l'utilisateur non-connecté. De plus, les différences dans cette figure (tous résultats confondus) ne descendent que très peu au-dessous de 20%, ce qui est assez élevé. L'impact moyen selon RBO@10 du paramètre temporel t_i versus t_{i+1} en considérant toutes les requêtes, respectivement pour l'utilisateur femme, homme, et non-connecté, est de 0,26 , 0,28 et 0,22. Dans le cas où on élimine le premier intervalle de temps considéré, pour les utilisateurs homme et femme, l'importance du paramètre temps devient égal à 0,22.

Dans (Dzogang *et al.*, 2017), les auteurs ont montré que la publication sur les réseaux sociaux dépend du moment de la journée. Barroso et Heolzle, dans (Barroso, Hoelzle, 2009), indiquent également (en partie 2.4.2) que les moteurs de recherche peuvent s'adapter à ce comportement. Nous étudions si de telles adaptations au moment de la journée influencent les réponses des moteurs de recherche. Notons que cet impact potentiel n'est pas indiqué dans la politique du moteur de recherche Google, nous sommes ici davantage dans un cadre exploratoire, hors politique explicite.

La figure 4 présente IM$_{RBO}$(temps) évalué sur les requêtes "vêtements" entre t_i à t_{i+1} (écart d'une demi-journée) et entre t_i à t_{i+2} (écart d'une journée complète). On constate que l'impact moyen du paramètre temps est plus important à une demi-journée d'intervalle qu'à une journée d'intervalle : les impacts moyens selon RBO@10 sont respectivement de 30,1% et 24,2%. Si nous détaillons ce résultat par utilisateur (cf. figure 5, on se rend compte que pour l'utilisateur anonyme l'impact du temps à une demi-journée selon RBO@10 est de 29,3%, alors que sur une journée il est de 17,7%. Pour la femme et l'homme, ces différences sont moindres, avec respectivement 30,8% versus 26,3% et 30,2% versus 28,6%, mais la tendance est la même. Il en résulte que les cycles circadiens jouent un rôle dans le moteur étudié. Ce comportement est également observé pour la paléontologie, mais pas pour le sport.

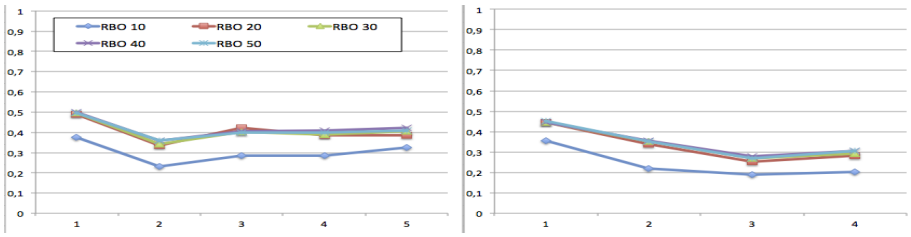

*Figure 4. RBO de 10 en 10 ∈ [10, 50] pour tous les utilisateurs et les requêtes "***vêtements***", à t_i versus t_{i+1} (gauche) et t_i versus t_{i+2} (droite).*

Différences entre utilisateurs

La figure 6 présente l'impact moyen du paramètre utilisateur IM$_{RBO}$(utilisateur), toutes requêtes confondues, à l'instant t_i. De gauche à droite, on observe les différences entre utilisateurs connectés (femme/homme), puis entre utilisateur anonyme et un utilisateur connecté (femme, puis homme). On observe que (a) les utilisateurs

connectés et anonymes ont globalement des différences notables (ente 10% de 20% pour RBO@10); (b) après quelques temps, les résultats sont plus similaires entre utilisateurs connectés qu'entre utilisateur connecté et anonyme (surtout pour RB0@10), (c) à l'exception de t_2, les résultats des utilisateurs connectés (homme/femme) sont à environ 10%, ce qui dénote une assez grande similarité, (d) les différences entre utilisateurs connectés et anonymes semblent grandir pour les t_5 et t_6.

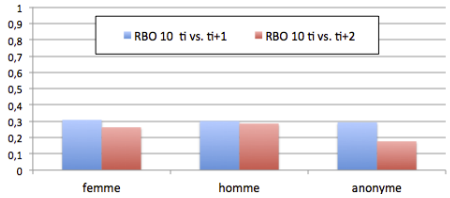

*Figure 5. RB0@10 par utilisateur, pour les requêtes "***vêtements***", à t_i versus t_{i+1} (gauche) et t_i versus t_{i+2}.*

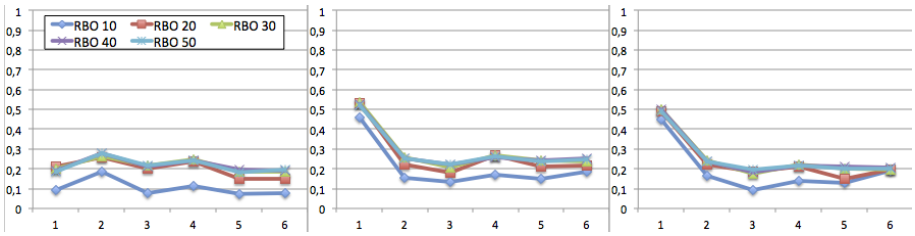

Figure 6. RBO de 10 en 10 ∈ [10, 50] entre deux utilisateurs toutes les requêtes, à t_i. **femme/homme** *(gauche),* **femme/anonyme** *(centre),* **homme/anonyme** *(droite)*

Si nous nous focalisons sur RBO@10 (c'est-à-dire la première page de résultats), pour toutes les requêtes, nous constatons que : l'IM du paramètre utilisateur homme versus femme, selon RBO@10, est de 10,3%; celui femme versus anonyme est de 21,0% (16,0% sans t_1); et celui homme versus anonyme, selon RBO@10 est de 19,4% (14,2% sans t_1). On remarque une grande différence entre utilisateurs anonyme et connectés à t_1. Si nous retirons ce moment t_1, l'IM du paramètre utilisateur femme versus anonyme est alors 16,0%, et celui homme versus anonyme est de 14,2% : ces valeurs sont encore supérieures à la différence femme/homme.

6.3. Interprétation des résultats

Tous les éléments des contextes que nous avons étudiés ici ont un IM notable sur les résultats. De manière constante, les valeurs de RBO entre 20 et 50 sont toujours plus élevées que les RBO à 10. Ceci dénote que les différences des listes après la position 10 sont relativement plus importantes qu'entre les positions 1 et 10 (comme nous nous basons sur les mesures de Rbo_{EXT}). La première page de résultats des moteurs de recherches actuels étant la plus importante pour un tel moteur, il est possible que les résultats des pages suivantes soient soumis à des aléas moins contrôlés.

Parmi toutes les expérimentations menées, nous n'obtenons des résultats identiques pour RBO@10 pour une variation de paramètre que dans 0,003% des cas (i.e. dans 67 cas sur 2538), y compris pour des requêtes dont on s'attend à peu d'évolution du corpus. Les aspects temporels, ou d'autres paramètres implicites, provoquent sans doute ces très nombreuses différences.

A noter, on observe parfois une grande différence pour un utilisateur connecté entre les résultats obtenus à t_1 et les autres. Par exemple, pour les requêtes "paleontology ammonites" et "paleontology mary anning", les résultats sont entièrement en anglais à t_1, très majoritairement en français à t_2 et entièrement en français à partir de t_3. Ce n'est pas le cas pour "paleontology mammoth", avec des réponses en anglais de t_1 à t_6. Ceci peut s'expliquer par une personnalisation qui s'affine, les résultats obtenus à t_1 correspondant à une sorte de "démarrage à froid" (Silva *et al.*, 2019). Pour un utilisateur anonyme, même à t_1, le moteur fait probablement, sans information supplémentaire, le choix de favoriser la localisation, ce qui dans notre cas lui fait présenter des résultats en français.

L'IM du paramètre genre des utilisateurs connectés (homme/femme) est plus faible que celui de connexion (entre un utilisateur connecté quelconque et un anonyme). Ce constat indique donc qu'un utilisateur anonyme n'est pas un utilisateur "moyen" entre homme et femme, mais qu'il est traité différemment. Les différences au cours du temps entre utilisateurs connectés femme/homme sont faibles : au début (i.e. t_1) les deux utilisateurs sont assez proches (environ 10% pour RBO@10) ce qui dénote une initialisation moyenne des utilisateurs. Ensuite ces différences restent faibles, sachant que leur activité était similaire : dans ce cadre la personnalisation les différencie peu.

Les cycles temporels circadiens, évalués par l'IM du paramètre temps, jouent un rôle notable. Ceci souligne donc qu'il existe des cycles dans le processus de recherche global : il est possible que deux index cohabitent et que suivant l'heure l'un ou l'autre est utilisé, mais cette explication est très ouverte.

6.4. Retour sur la transparence

Il ressort de nos expérimentations que la politique de personnalisation affichée de Google search est vérifiée pour le genre et le temps, ce qui révèle une honnêteté par rapport à la politique. Toutefois, les explications fournies par Google search ne sont pas très détaillées (par exemple : "toute autre information que Google est susceptible d'associer à vous") : cela ne clarifie pas le comportement du moteur à un utilisateur. En particulier, nous avons vu que l'impact du moment de la journée auquel on pose une requête est une source importante de différences dans la variation des résultats. En conséquence, il nous semble pertinent que l'utilisateur soit informé du fait que les résultats pour une même requête effectuée à deux moments de la journée sont susceptibles de varier davantage que pour la même requête effectuée à 24h d'intervalle.

De même, pour un utilisateur connecté, quand une requête est composée d'au moins un mot existant dans plusieurs langues, Google search choisit de favoriser une

langue au fur et à mesure de la personnalisation. Mais la façon dont la langue est choisie mériterait plus d'explications : pour nos requêtes composées d'un mot anglais et un mot français/anglais, Google a privilégié de présenter des sites français, alors que la langue du compte et celle du navigateur étaient en anglais [4].

7. Conclusion

Dans cet article, nous avons étudié les notions de transparence, d'honnêteté et d'explicabilité des moteurs de recherche, proposé un protocole d'évaluation et étudié l'impact de la connexion, du genre et du temps sur les résultats proposés par Google. L'expérimentation menée, bien que restreinte, permet de mettre en évidence que les résultats des requêtes varient assez largement dès la première page, même pour des requêtes qui a priori devraient fournir des résultats similaires. Nous avons aussi observé que les cycles circadiens jouent un rôle inattendu dans les variations des réponses. Compte-tenu de la taille de l'échantillon, il n'est pas possible de mener des tests de significativité statistiques valables; nous menons actuellement une étude complémentaire, sur un plus long terme, avec plus d'utilisateurs, et des requêtes non multilingues.

Nous avons pleinement conscience que les "corrélations" détectées ici ne sont pas des preuves de causalité, mais ceci ne diminue pas le fait que notre étude souligne le besoin de transparence des moteurs de recherche. Le fait que les moteurs évoluent de manière quasi-continue ne doit pas être selon nous un frein à des explications fournies aux utilisateurs.

L'impact des cycles circadiens n'a pas été observé dans (Hannák *et al.*, 2013 ; 2017) du fait de la construction de leur expérimentation. Il serait sans doute intéressant de mettre en perspective leurs conclusions avec nos constats. Les éléments que nous avons proposés ici ne sont qu'une infime partie d'un cadre général d'évaluation "externe" des moteurs de recherche qui pourrait, idéalement, : a) extraire automatiquement les paramètres externes à partir de la politique d'un moteur, b) définir automatiquement un protocole d'évaluation capable de s'adapter à la nature de ces paramètres, c) lancer les expérimentations, évaluer les résultats et les mettre en perspective de la politique du moteur sous forme d'un tableau de bord ou d'explications textuelles générées. C'est vers ce type d'outils, utiles aussi bien pour le grand-public que pour les auto-évaluations des moteurs, qu'il semble souhaitable d'aller à l'avenir.

Remerciements

Ces travaux de recherche ont été soutenus le projet Émergence *arOsoIR* financé par le Laboratoire d'Informatique de Grenoble. Ils ont été inspirés par les réflexions autour du projet Transalgo (https://www.transalgo.org/).

4. Il est à noter que, lors de nouvelles expérimentations, nous avons constaté une fenêtre "pop-up" qui interroge l'utilisateur sur ce qu'il préfère dans le cas son compte n'est pas dans la langue de l'interface...

Bibliographie

Andreou A., Venkatadri G., Goga O., Gummadi K. P., Loiseau P., Mislove A. (2018, Feb). Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *Proceedings of the network and distributed system security symposium (ndss'18).* San Diego, CA, USA.

Barroso L. A., Hoelzle U. (2009). *The datacenter as a computer: An introduction to the design of warehouse-scale machines* (1st éd.). Morgan and Claypool Publishers.

Biega A. J., Gummadi K. P., Weikum G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, p. 405–414. New York, NY, USA, ACM.

Buckley C., Voorhees E. M. (2017, août). Evaluating evaluation measure stability. *SIGIR Forum*, vol. 51, n° 2, p. 235–242.

contributors P. (2018). *PhantomJS - Scriptable Headless Browser.* (http://phantomjs.org/ [Retrieved December, 2018])

Data Center Knowledge. (2017). *Google Data Center FAQ, Everything you ever wanted to know (and didn't) about Google data centers but were afraid to ask.* (https://www.datacenterknowledge.com/archives/2017/03/16/google-data-center-faq [Retrieved Decembre, 2018])

Dean B. (2018). *Google's 200 ranking factors: The complete list (2018).* (https://backlinko.com/google-ranking-factors [Retrieved Decembre, 2018])

DuckDuckGo. (2018a). *Measuring the "Filter Bubble": How Google is influencing what you click.* (https://spreadprivacy.com/google-filter-bubble-study/ [Retrieved Decembre, 2018])

DuckDuckGo. (2018b). *Welcome to DuckDuckGo.* (https://duckduckgo.com/about [Retrieved Novembre, 2018])

Dzogang F., Lightman S., Cristianini N. (2017). Circadian mood variations in twitter content. *Brain and Neuroscience Advances*, vol. 1, p. 1–14.

Google. (2018a). *Comment fonctionnent les algorithmes de recherche ?* (https://www.google.com/search/howsearchworks/algorithms/ [Retrieved Novembre, 2018])

Google. (2018b). *Google Trends.* (https://trends.google.fr/trends/?geo=FR [Retrieved December, 2018])

Google. (2018c). *Règles de confidentialité et conditions d'utilisation.* (https://policies.google.com/privacy?hl=fr&gl=fr#infocollect [Retrieved Novembre, 2018])

Hannák A., Sapiezynski P., Khaki A. M., Lazer D., Mislove A., Wilson C. (2017). Measuring personalization of web search. *CoRR*, vol. abs/1706.05011.

Hannák A., Sapiezynski P., Molavi Kakhki A., Krishnamurthy B., Lazer D., Mislove A. *et al.* (2013). Measuring personalization of web search. In *Proceedings of the 22nd international conference on world wide web*, p. 527–538. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/2488388.2488435

Jansen B. J., Booth D. (2010). Classifying web queries by topic and user intent. In *Chi '10 extended abstracts on human factors in computing systems*, p. 4285–4290. New York, NY, USA, ACM.

Joachims T., Radlinski F. (2007, août). Search engines that learn from implicit feedback. *Computer*, vol. 40, n° 8, p. 34–40.

Kules B., Wilson M., Schraefel M., Shneiderman B. (2008). *From keyword search to exploration: How result visualization aids discovery on the web*. Rapport technique n° HCIL-2008-06. University of Maryland.

Microsoft. (2018). *Microsoft Privacy Statement.* (https://privacy.microsoft.com/en-us/privacystatement [Retrieved Novembre, 2018])

Mittelstadt B. D., Allo P., Taddeo M., Wachter S., Floridi L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, vol. 3, n° 2, p. 2053951716679679. Consulté sur https://doi.org/10.1177/2053951716679679

Nottelmann H., Fuhr N. (2003, 01 Sep). From Retrieval Status Values to Probabilities of Relevance for Advanced IR Applications. *Information Retrieval*, vol. 6, n° 3, p. 363–388.

Qwant T. (2016). *Notre philosophie.* (https://help.qwant.com/fr/aide/general/notre-philosophie/ [Retrieved Novembre, 2018])

Qwant T. (2017). *Politique de protection des données.* (https://about.qwant.com/fr/legal/confidentialite/ [Retrieved Novembre, 2018])

SearchLiaison G. (2018). *Google search liaison.* (https://twitter.com/searchliaison)

Silva N., Carvalho D., Pereira A. C., Mourão F., Rocha L. (2019). The pure cold-start problem: A deep study about how to conquer first-time users in recommendations domains. *Information Systems*, vol. 80, p. 1 - 12.

Singh A., Joachims T. (2018, August). Fairness of exposure in rankings. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery and data mining*, p. 2219-2228. ACM.

Smucker M. D., Allan J., Carterette B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management*, p. 623–632. New York, NY, USA, ACM.

Turilli M., Floridi L. (2009, Jun). The ethics of information transparency. *Ethics and Information Technology*, vol. 11, n° 2, p. 105–112. Consulté sur https://doi.org/10.1007/s10676-009-9187-9

Venkatadri G., Lucherini E., Sapiezynski P., Mislove A. (2019, July). Investigating sources of pii used in facebook's targeted advertising. In *Procesedings of privacy enhancing techniques, pets.* sciendo.

Webber W., Moffat A., Zobel J. (2010). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, vol. 28, n° 4, p. 20:1–20:38.