Exploration de l'apport de l'analyse des perceptions oculaires : étude préliminaire pour le bouclage de pertinence

Lucas Albarede, Francis Jambon, Philippe Mulhem

Exploration de l'apport de l'analyse des perceptions oculaires : étude préliminaire pour le bouclage de pertinence

Lucas Albarede, Francis Jambon, Philippe Mulhem

Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP [1]*, LIG, 38000 Grenoble, France*
Lucas.Albarede@etu.univ-grenoble-alpes.fr,
Francis.Jambon@imag.fr, Philippe.Mulhem@imag.fr

RÉSUMÉ. *Nos travaux visent à évaluer l'impact potentiel de l'utilisation des perceptions oculaires vis-à-vis du mécanisme de bouclage de pertinence d'un moteur de recherche d'information. Nous nous sommes intéressés aux situations écologiques où l'utilisateur évalue la pertinence des résultats d'une requête via les snippets affichées sur la page des résultats. Notre hypothèse est que la connaissance des termes lus par un utilisateur sur cette page de résultats peut être utilisée pour améliorer le mécanisme de bouclage de pertinence, sans intervention explicite de l'utilisateur. Pour cela, nous modélisons le comportement visuel de l'utilisateur lorsqu'il consulte les résultats d'une requête, pour extraire de cette modélisation des indicateurs. Nous avons étudié le dernier mot fixé et le mot le plus longtemps fixé, dans la page de résultats et dans chaque snippet. Les résultats des expériences réalisées avec des utilisateurs indiquent que le mot le plus longtemps fixé est souvent un mot pertinent pour compléter la requête.*

ABSTRACT. *Our work aims to evaluate the potential impact of the use of ocular perceptions analysis in the relevance feedback mechanism of a retrieval engine. We focus on ecological situations where the user evaluates the relevance of the results of a query via the snippets displayed on the results page. Our hypothesis is that the knowledge of the terms read by a user on the result page can be used to improve the relevance feedack mechanism, without explicit intervention of the user. For this, we model the visual behavior of the user when reading the results of a query, to extract indicators. We explored the last fixed word and the longest fixed word, in the results page and in each snippet. The results of the user experiments indicate that the longest fixation word is often a relevant word to complement the query.*

MOTS-CLÉS : *bouclage de pertinence, page de résultats, snippet, comportement perceptivo-gestuel, oculométrie*

KEYWORDS: *relevance feedback, result page, snippet, perceptivo-gestual behaviour, eye-tracking*

doi:10.24348/coria.2019.CORIA_2019_paper_1

1. Institute of Engineering Univ. Grenoble Alpes

1. Introduction

La recherche d'information étudie la manière de retrouver, parmi un corpus de documents, ceux qui sont pertinents pour un besoin d'information formulé par un utilisateur [Manning *et al*., 2008]. Classiquement, ce besoin d'information est exprimé au travers d'une requête, le système effectuant en réponse à celle-ci une sélection et un tri dans le corpus, pour fournir une liste de réponses ordonnées par pertinence décroissante. Dans le cas de la recherche de documents sur le web, les corpus sont gigantesques et la taille des listes fournies en réponses également. C'est pourquoi les systèmes choisissent habituellement de favoriser la précision dans les premières réponses, au détriment du rappel. Cette limitation motive actuellement de nombreux travaux de recherche qui s'intéressent à améliorer cet aspect.

De nombreux travaux en recherche d'information intègrent le point de vue de l'utilisateur sur les documents qui lui sont présentés, au travers du bouclage de pertinence ou *relevance feedback (RF)* [Zhai, Massung, 2016, Ermakova, Mothe, 2016]. Le RF permet d'affiner la requête de l'utilisateur en fonction des documents qu'il a indiqué comme pertinents (et/ou non-pertinents), afin que le système de recherche d'information retrouve de meilleurs résultats. Ce RF repose sur une interaction simple de l'utilisateur (sélection d'un ou de plusieurs documents), et les indicateurs utilisés de manière interne par le RF reposent sur l'ensemble des termes de ces documents, en intégrant éventuellement une similarité avec la requête.

Les résultats d'une requête sur un moteur de recherche sont classiquement organisés sous forme d'une page de résultats ou *Search Engine Result Page (SERP)* dont l'objectif est de proposer à l'utilisateur une partie de la liste globale des résultats (classiquement 10 résultats par page), en fournissant de manière compacte des informations permettant à l'utilisateur d'estimer la pertinence du document complet. Ces pages sont habituellement composées de blocs, appelées *snippets*, générés à partir des informations extraites des documents [Turpin *et al*., 2007], et uniformisant la présentation des réponses.

Afin de choisir le ou les documents correspondant le mieux à sa requête, l'utilisateur doit effectuer un travail cognitif de recherche en s'appuyant sur les informations présentées dans la SERP. Durant ce travail cognitif, il regarde et analyse les éléments des snippets afin d'estimer la valeur de pertinence des documents correspondants : c'est ce qu'on appelle le comportement perceptivo-gestuel de l'utilisateur. Notre hypothèse est que l'analyse de ce comportement peut nous permettre de mieux connaître l'intention de l'utilisateur, et en corollaire, fournir des indicateurs pour le RF, pouvant se substituer à ceux issus de l'analyse des termes des documents.

Les travaux présentés dans cet article s'intéressent à vérifier une partie de cette hypothèse. Plus précisément, nous cherchons à déterminer quels indicateurs, potentiellement utiles pour le RF, il est possible d'extraire du comportement perceptif (position du regard sur la page de résultats) et gestuel (sélection du/des documents estimés le/les plus pertinents) de l'utilisateur lorsqu'il évalue les résultats d'une requête sur une SERP composée de snippets. Nous avons focalisé notre étude sur l'exploration

des SERP et non pas des documents eux-mêmes, dans le but de se rapprocher au plus près d'une situation dite écologique. En effet, dans l'usage habituel d'un système de recherche d'information sur le Web, c'est à partir des snippets présentées dans une SERP que l'utilisateur effectue sa première estimation des documents les plus pertinents, documents qu'il pourra, mais seulement dans un second temps, consulter et évaluer.

De plus, nous nous intéressons aux perceptions au niveau de granularité des **mots**. En effet, c'est à ce niveau de granularité qu'il est possible d'extraire de nouveaux termes pertinents pour affiner une requête. Se limiter à un niveau de granularité plus élevé, celui des snippets, n'aurait que peu d'intérêt du fait que l'utilisateur indique déjà son choix par la sélection (clic souris) d'un résultat pertinent. Un niveau de granularité plus fin, celui des lettres, n'aurait pas d'intérêt sur l'interprétation sémantique de la perception dans le contexte de la recherche d'information.

Nous ne quantifions pas ici l'utilisation de ces indicateurs vis-à-vis des approches de RF, car les résultats issus d'expérimentations similaires ont déjà montré l'intérêt de l'apport de ce type d'indicateurs. Par exemple, Buscher *et al*. [2009], ont amélioré significativement les résultats par expansion de requête en intégrant les mots contenus dans des sous-parties de documents vues par l'utilisateur. De même, les travaux de Eickhoff *et al*. [2015] ont montré que les mots les plus longtemps regardés sont souvent réutilisés par les utilisateurs lors de reformulations de requêtes. Ces travaux seront détaillés dans la section 2 consacrée à l'état de l'art.

Cet article est organisé comme suit. Après cette introduction, la section 2 présente un état de l'art de l'utilisation de l'eye-tracking pour la recherche d'information. Nous décrivons ensuite, en section 3, notre démarche expérimentale pour déterminer l'emploi d'eye-tracking dans notre cadre, et les résultats obtenus, avant de conclure et de présenter les perspectives de ces travaux en sections 4 et 5.

2. État de l'art

Nous présentons ici les travaux de recherche existants relatifs à notre domaine d'étude. Dans un premier temps, nous étudions les propositions existantes pour évaluer la pertinence d'un texte grâce au comportement visuel de l'utilisateur, puis dans un second temps, nous présentons les approches qui concernent l'utilisation des informations fournies par une analyse perceptivo-gestuelle de l'activité d'un utilisateur pour le RF dans le domaine de la recherche d'information.

L'étude de la corrélation entre le comportement visuel d'un utilisateur et le degré de pertinence d'un document lu, réalisée par Gwizdka [2014], indique que le degré de pertinence d'un document affecte la manière de le lire. Ces travaux ont notamment montré que les durées des fixations sur les mots d'un texte obtenues grâce à l'eye-tracking sont un bon indicateur de la pertinence du document. L'étude de Gwizdka [2014] porte sur la lecture de documents entiers et non pas des snippets d'une SERP. Or, dans un contexte écologique, un utilisateur se contente généralement de lire la

SERP, et en conséquence, il ne peut consulter qu'un plus faible nombre de mots. De plus, il n'y a pas forcément d'ordre logique entre les phrases extraites du document et celles présentées dans les snippets. Il est donc difficile d'extrapoler les résultats de Gwizdka [2014] pour déterminer la pertinence des mots dans des snippets.

Concernant plus particulièrement les approches intégrant de l'expansion de requête, Buscher *et al*. [2009] ont proposé deux méthodes. Dans l'une d'elle, proche de nos travaux, les auteurs utilisent le temps qu'un utilisateur met à lire certaines parties des documents, résultant d'une requête, pour déterminer les mots les plus pertinents à lui ajouter. Il est à noter que cette étude se limite à analyser les documents, en les segmentant, mais sans aller jusqu'au niveau de granularité du mot regardé. Il est alors raisonnable de se demander si, à ce niveau d'analyse, une part importante de l'information disponible pourrait être non exploitée.

Les travaux de Y. Chen *et al*. [2015] ont en partie repoussé cette limite. Les auteurs ont proposé de réaliser un mécanisme de RF en utilisant les mots lus par l'utilisateur, dans les documents jugés pertinents par celui-ci. Ces travaux confirment l'intérêt de prendre en compte des mots perçus par l'utilisateur dans une expansion de requête. Cependant ces travaux, comme ceux de Buscher *et al*. [2009] évoqués ci-dessus, se focalisent sur les documents, potentiellement plus riches en mots pertinents que les snippets présents dans les SERPS.

Les travaux de Eickhoff *et al*. [2015] ont dépassé ces limites en effectuant une analyse perceptivo-gestuelle plus précise, car centrée sur les termes. Ils ont conduit une expérience où les participants devaient répondre à plusieurs questions, à l'aide d'un moteur de recherche, en formulant et reformulant eux-mêmes leurs requêtes. Les résultats montrent une connexion forte entre les mots utilisés dans l'expansion de requête par l'utilisateur, et les mots présents dans les SERPs qu'il a regardé. Ils montrent aussi que l'apparition dans une SERP de mots présents dans la requête, provoque une plus grande attention de la part des participants, sous forme de fixations plus longues. Il est cependant difficile d'utiliser directement ces résultats car ils se concentrent sur des recherches comportant de multiples SERPs, et ces travaux ne tiennent pas compte de l'action de l'utilisateur quand celui-ci choisit un extrait plutôt qu'un autre. De plus, ces travaux ont principalement un objectif explicatif, et l'expansion de requête n'est réalisée que par l'utilisateur lui-même. Nous retenons cependant de ces travaux le lien qui a été établi entre les mots les plus longtemps regardés, et les mots présents dans la requête reformulée.

Notre travail diffère de ceux énoncés dans cet état de l'art en plusieurs points. Tout en restant dans le cadre de la recherche d'information, et contrairement à Eickhoff *et al*. [2015] et Buscher *et al*. [2009], nous nous intéressons uniquement aux systèmes automatisés de RF. Tout en suivant les travaux de Gwizdka [2014] et Y. Chen *et al*. [2015], nous utilisons comme indicateurs non seulement le mot lu le plus longtemps, mais aussi le dernier mot lu. En effet, un mot lu plus longtemps est l'indice d'un traitement cognitif significatif, tandis que le dernier mot lu peut être celui ayant été déterminant quant à l'évaluation de la pertinence du document, car c'est celui qui a été lu juste avant la prise de décision de l'utilisateur. De plus, nous étendons cette

métrique en ajoutant aux mots lus la notion de pertinence positive, neutre ou négative, permettant ainsi de faire le lien avec l'expansion de requête et les systèmes de RF. Enfin, nous prenons en compte non seulement la perception de l'utilisateur, mais également de ses actions, c'est-à-dire concrètement l'action de sélection de l'extrait estimé le plus pertinent parmi ceux affichés dans la SERP.

3. Expérimentations

Notre objectif est de déterminer s'il est possible d'extraire des indicateurs pour le RF, au niveau de granularité du mot, à partir de l'analyse du comportement perceptivo-gestuel de l'utilisateur lorsqu'il consulte une SERP. Afin d'identifier les mots lus par l'utilisateur sur une SERP, nous avons fait appel à la technique de l'eye-tracking. Cette technique permet de disposer d'un échantillonnage de la position du regard de l'utilisateur sur un écran, à partir duquel il est possible de déterminer les endroits ou se porte plus particulièrement l'attention de l'utilisateur, appelées fixations. Cette technique non-invasive se base principalement sur l'analyse des images des yeux de l'utilisateur dans le spectre du proche infrarouge. La technique de l'eye-tracking est relativement fiable mais est limitée en précision, à la fois du fait de contraintes techniques, mais aussi physiologiques.

C'est pourquoi, nous avons suivi une démarche expérimentale en deux temps. Dans un premier temps (section 3.1) nous avons réalisé une étude de faisabilité afin de déterminer la configuration expérimentale la plus favorable à la détection des mots lus par l'utilisateur, dans un contexte expérimental similaire à celui de la lecture d'une SERP. Dans un second temps (section 3.2) nous avons réutilisé la configuration expérimentale la plus favorable identifiée précédemment, pour réaliser une analyse du comportement perceptivo-gestuel de l'utilisateur, dans un contexte reproduisant aussi fidèlement que possible l'activité d'un utilisateur consultant une SERP.

3.1. Étude de faisabilité

La première expérimentation visait à établir, en fonction du dispositif technique utilisé, les paramètres expérimentaux minimaux pour assurer que le système d'eye-tracking puisse détecter avec précision les mots lus à l'écran par l'utilisateur. Les paramètres explorés ont été : a) le modèle d'eye-tracker utilisé et b) des caractéristiques de présentation (c'est-à-dire principalement la taille des caractères composant les snippets). Rappelons que ce préalable est fondamental, car la pertinence de l'analyse comportementale repose nécessairement sur la capacité du dispositif technique à identifier correctement les mots lus par l'utilisateur.

3.1.1. Configuration expérimentale

La consigne donnée aux participants était de retrouver, le plus rapidement possible, deux mots fixés connus, chacun présent (en une seule occurrence) dans une page de texte simulant une SERP issue d'un moteur de recherche sur le Web. Chaque

SERP était composée de trois snippets. Afin de prendre en compte un biais potentiel sur la longueur des mots recherchés, les participants devaient retrouver un mot court ("fèves") dans la première page, et un mot long ("macadamia") dans la seconde. Les pages de résultats ont été construites manuellement. Dans ces pages, nous avons fait varier la taille des caractères. Nous avons utilisé les polices de caractères *Arial 13*, *Arial 14* et *Arial 15*. Ces tailles ont été choisies car elles ont des valeurs se rapprochant de celles constatées dans les SERPs issues des moteurs de recherche sur le Web. Les interlignes ont été fixés à une valeur globale de 0,2 cm.

Chaque participant devait retrouver chaque mot sur trois versions (avec des tailles de police de caractères différentes) de la même SERP. L'ordre de présentation des pages, pour un même mot, était randomisé entre les participants. Afin de limiter l'effet d'apprentissage, l'ordre des snippets dans les pages pour un mot donné était également randomisé. Cependant, afin de limiter le nombre de pages à créer, les mêmes pages étaient présentées à tous les participants. Cela peut apporter un biais car le mot à retrouver se trouve toujours au même emplacement, pour une taille donnée et tous les participants. Nous avons estimé que ce biais n'était pas significatif car l'objectif de l'expérimentation était de mesurer la précision et non la vitesse de découverte du mot.

Pour réaliser ces tests utilisateurs, nous avons utilisé le logiciel Ogama [2]. Chaque test était constitué d'un ensemble de diapositives montrant les consignes, puis les SERPs simulées. Le participant passait d'une diapositive à la suivante en effectuant un clic souris. La consigne donnée au participant était de retrouver, le plus vite possible, le mot indiqué, et dès qu'il l'avait trouvé, d'effectuer un clic souris (mais sans désigner le mot avec le curseur de la souris).

Ces tests ont été effectués dans un premier temps avec un eye-tracker EyeTribe ET1000 [3]. Ce dispositif a été initialement choisi car il est représentatif des eye-trackers grand public à bas coût destinés à l'interaction homme-machine. L'eye-tracker était fixé sur un écran de taille 19 pouces et de résolution 1280x1024 pixels. Les participants étaient positionnés à environ 60 cm de l'écran et avaient pour consigne de garder une position stable pendant la totalité des tests. La capture des mouvements oculaires et l'analyse des fixations ont été réalisées par le logiciel Ogama. Les paramètres utilisés étaient une fréquence d'échantillonnage de 30 Hz (la fréquence recommandée pour cet eye-tracker), une détection des fixations oculaires par dispersion géométrique avec un nombre minimum d'échantillons de 3 (soit 100 ms minimum), et un seuil de dispersion de 20 pixels.

3.1.2. Évaluation

Comme les participants avaient pour consigne de mettre fin à leur recherche le plus rapidement possible une fois le mot trouvé, nous avons utilisé dans nos évaluations la dernière fixation détectée, correspondant a priori au dernier mot regardé, sans prendre

2. http://www.ogama.net/
3. http://theeyetribe.com/

en compte d'éventuelles dérives du regard après la recherche. L'objectif de l'expérimentation étant de déterminer la taille de texte la plus adaptée à la reconnaissance automatisée par un eye-tracker, des mots lus par l'utilisateur, nous avons fait reposer notre évaluation sur une mesure de distance entre chaque mot et la dernière fixation détectée. Pour cela, nous avons choisi d'utiliser la boîte englobante de chaque mot de chaque SERP, et de définir le centre de cette boîte comme représentant de la localisation du mot. Cette représentation est limitée au sens où elle ne prend pas en compte la longueur du mot, cependant, elle permet une comparaison linéaire des distances entre les mots regardés. Cette mesure de distance a été réalisée manuellement à partir de la superposition des fixations de chaque participant sur les diapositives simulant les SERPs. Elle a été réalisée en déterminant la distance en pixels entre deux points de l'image avec le logiciel Gimp [4].

De manière plus formelle, pour une SERP donnée p, nous notons l'ensemble des ses mots W_p. Pour chacun des mots $w_p \in W_p$, nous notons $c(w_p)$ les coordonnées de son isobarycentre. Il est dès lors possible de trier, pour une fixation de coordonnées $< x, y >$, les mots de W_p par ordre de distance euclidienne croissante. Nous notons la liste triée obtenue $L(W_p, < x, y >)$. Un élément à la position i dans celle liste est dénoté par $L(W_p, < x, y >)[i]$, pour $i \in [1, |W_p|]$. La position du mot dans la liste, c'est-à-dire le rang i, définit la notion de proximité : le mot le plus proche d'une fixation de coordonnées $< x, y >$ est à la position 1 dans cette liste. Plus précisément, le rang d'un mot $w_p \in W_p$ dans $L(W_p, < x, y >)$ est défini par r tel que $L(W_p, < x, y >)[r] = w$. Ce modèle d'évaluation de proximité nous permet de comparer la qualité de la détection du mot w à retrouver. Celle-ci est calculée en réalisant, pour chaque mot à retrouver w, la moyenne de ses rangs sur un ensemble donné de fixations.

3.1.3. Résultats

Nous avons constaté une décroissance constante de la mesure de proximité des mots à retrouver, en fonction de l'augmentation de taille de texte. Ce résultat était attendu, en effet, plus la taille de la police de caractères est grande, plus les chances que l'eye-tracker reconnaisse correctement le mot à retrouver sont élevées. Cependant, le meilleur résultat obtenu l'a été en utilisant la police de caractères *Arial 15*, avec comme résultat une proximité moyenne de 1,75. Ceci indique que, en moyenne, le mot à retrouver n'est pas le plus proche de la dernière fixation détectée (le résultat idéal étant une proximité moyenne de 1). Nous avons conclu qu'il n'est pas possible de détecter précisément les mots à retrouver avec la configuration expérimentale testée. Nous avons donc modifié cette configuration pour effectuer un second ensemble d'expérimentations.

3.1.4. Modification de la configuration expérimentale

Pour cette nouvelle configuration expérimentale, nous avons choisi de ne pas modifier la taille des polices de caractères ni d'interligne, afin de rester dans un contexte

4. http://www.gimp.org/

d'expérimentation se rapprochant d'une situation écologique. En conséquence, nous avons dû changer d'eye-tracker afin d'utiliser un modèle plus performant. Nous avons ainsi remplacé le EyeTribe ET1000 par un Tobii Pro X3-120 [5] avec une unité de traitement externe [6]. Initialement, l'eye-tracker EyeTribe ET1000 avait été choisi car il était représentatif des dispositifs grand public à bas coût destinés à l'interaction homme-machine. L'eye-tracker Tobii Pro X3-120 est quant à lui représentatif des dispositifs professionnels destinés à l'analyse des processus cognitifs. Son coût est plus élevé d'un facteur 100 environ. Ce changement, s'il ne pose pas de problèmes fondamentaux du point de vue de la validité des résultats, montre cependant que l'approche proposée n'est pas encore transposable dans un cadre grand public. Cet aspect sera discuté plus en détail dans la partie consacrée aux perspectives de ces travaux.

Nous avons gardé le même protocole expérimental, le même modèle d'évaluation, et le même logiciel d'analyse que précédemment. Seul l'eye-tracker utilisé pour la capture des mouvements oculaires a été changé. Les paramètres utilisés étaient une fréquence d'échantillonnage de 120 Hz (la fréquence disponible pour cet eye-tracker), une détection des fixations oculaires par dispersion géométrique avec un nombre minimum d'échantillons de 12 (soit 100 ms minimum), et un seuil de dispersion de 20 pixels.

Nous avons repris les mêmes participants que pour la première version de l'expérimentation, mais par manque de temps, nous n'avons pu en faire passer que trois sur les cinq initiaux. Avec cette nouvelle configuration, la valeur de proximité moyenne avec la police *Arial 15* est alors égale à 1,15 (au lieu de 1,75 dans la première expérimentation), ce qui représente une nette amélioration par rapport aux résultats précédents.

3.1.5. Synthèse

En synthèse, nous avons ré-analysé les résultats des deux expérimentations selon un point de vue complémentaire. Au lieu de déterminer une proximité, nous avons plus simplement déterminé si le mot recherché avait été détecté, ou si un autre mot avait été détecté (tableau 1). Cette métrique est moins sensible que la notion de proximité (par exemple il n'est pas possible de connaître le rang moyen du mot recherché), mais elle permet une interprétation plus directe des résultats en vue d'une utilisation en RF (il est possible d'estimer la probabilité d'avoir le mot recherché détecté). Nous nous sommes limités aux trois sujets qui ont effectué les deux expérimentations afin de pouvoir effectuer des comparaisons deux à deux, mais aussi pour éviter les possibles biais si un participant avait présenté des résultats particulièrement bons ou mauvais dans seulement l'une des deux expérimentations.

Même si le très faible nombre de participants ne permet pas de tirer des conclusions, on remarque une meilleure proportion de détections dans le cas où l'eye-tracker Tobii Pro X3-120 est utilisé avec la plus grande taille de police de caractères (ta-

5. http://tobiipro.com/product-listing/tobii-pro-x3-120/
6. http://tobiipro.com/product-listing/external-processing-unit/

Tableau 1. Résultats de détection du mot recherché par participant, taille de police de caractères, et par longueur de mot : "x" indique si le mot recherché a été détecté, et "o" si un autre mot a été détecté. La valeur de gauche correspond au mot court ("fèves") et la valeur de droite au mot long ("macadamia"). Le total correspond à la proportion de mots recherchés détectés, tous mots et tous participants confondus.

	EyeTribe ET1000			Tobii Pro X3-120		
Participant	*Arial 13*	*Arial 14*	*Arial 15*	*Arial 13*	*Arial 14*	*Arial 15*
1	o - o	o - o	x - o	x - o	o - x	x - o
2	o - x	o - x	o - x	o - x	x - x	x - x
3	o - o	o - x	o - x	o - o	o - x	x - x
Total	1/6	2/6	3/6	2/6	4/6	**5/6**

bleau 1). La marge d'erreur obtenue pour cette configuration expérimentale nous a semblé donc suffisamment faible pour l'utiliser dans la suite de nos travaux, afin de réaliser l'analyse comportementale.

3.2. Analyse comportementale

Cette seconde expérimentation avait pour objectif de déterminer s'il était possible d'analyser les fixations du regard de l'utilisateur, pour acquérir des informations utilisables par un mécanisme de RF, dans le cadre d'une recherche d'information sur le Web. Plus précisément, nos travaux ont consisté à déterminer des indicateurs issus de l'analyse du comportement perceptivo-gestuel d'un l'utilisateur, lorsqu'il évalue la pertinence des résultats présentés sur une SERP.

3.2.1. Configuration expérimentale

Nous avons globalement réutilisé la configuration expérimentale évaluée lors de l'expérimentation précédente, c'est-à-dire que nous avons utilisé un eye-tracker Tobii Pro X3-120 avec le logiciel Ogama et les mêmes paramètres pour l'analyse des fixations oculaires. Pour des raisons pratiques, nous avons utilisé un écran de taille 20 pouces et de résolution 1600x1200 au lieu de l'écran de taille 19 pouces et de résolution 1280x1024 précédent, sans qu'il n'y ait de modification significative de la taille des caractères à l'affichage.

Nous nous sommes placés dans une situation où l'utilisateur doit effectuer un travail cognitif d'évaluation des réponses (présentées sous forme de snippets) avant de choisir celle qui lui semble la plus pertinente. Nous situons ainsi nos travaux sur des requêtes Web dites informationnelles [Broder, 2002], c'est-à-dire pour lesquelles il est nécessaire d'effectuer plusieurs requêtes successives avant d'obtenir un résultat. Pour simuler cette activité de recherche sur le Web, nous avons défini des tâches de recherche d'information, formulé les requêtes associées, et créé les SERPs correspondantes. Afin de rendre l'évaluation des indicateurs plus discriminante, nous avons créé

des SERPs spécifiques comprenant systématiquement trois niveaux de pertinence des résultats.

Nous avons caractérisé la pertinence des résultats proposés à l'utilisateur suivant trois catégories : (1) Les résultats positifs, qui comportent au moins un élément permettant d'indiquer clairement qu'ils répondent, au moins en partie, à la réponse attendue ; (2) Les résultats ambigus, qui abordent le même thème que le résultat positif, mais qui ne comportent pas d'élément permettant d'indiquer clairement qu'ils répondent, au moins en partie, ou ne répondent pas, à la réponse attendue ; (3) Les résultats négatifs, qui possèdent au moins un élément indiquant clairement qu'ils ne répondent pas à la réponse attendue.

Les SERPs proposées aux participants comportaient chacune une réponse appartenant à chaque catégorie de résultat. Cette caractérisation a été réalisée à l'unanimité des trois auteurs de l'article. En incorporant ces trois catégories de résultat à chaque page de recherche, nous nous plaçons dans un cadre expérimental où l'utilisateur doit effectuer un travail cognitif de détermination de la réponse la plus pertinente (car il ne peut pas se contenter de sélectionner le premier résultat). Afin de randomiser les pages de résultats, nous avons créé des pages où les résultats positifs, ambigus et négatifs sont disposés dans les différents ordres possibles. Cela nous permet d'étudier le comportement des participants en neutralisant l'influence de l'ordre dans lequel les résultats leurs sont présentés.

Or, puisque nous avons trois catégories de résultats, si nous avons au moins un résultat de chaque type sur la page, le nombre de configurations possibles d'ordres des résultats est $2 \times 3^{n-2}$, où n représente le nombre de résultats présents sur la page. De plus, nous nous devons de prendre en compte les effets d'apprentissage des participants aux requêtes qui leur sont proposées. Il est donc impossible de présenter deux fois la même requête au même participant, même si l'ordre des résultats a changé. Ainsi, si nous voulons que chaque participant soit confronté à toutes les configurations possibles, il nous faut un nombre de requêtes identique au nombre de configurations. Ne disposant que d'un temps limité pour préparer et effectuer l'expérimentation, notre choix s'est porté sur un ensemble de SERPs de taille minimale, c'est-à-dire 3 résultats par page (i.e. un résultat de chaque catégorie) et donc la création de 6 requêtes. De plus, afin de minimiser l'impact que pourrait avoir une requête spécifique sur une configuration donnée, nous avons associé à chaque requête les 6 configurations de résultats possibles. Au final, cette approche permet de neutraliser les variables indépendantes liées à la requête, à l'ordre et à la catégorie des résultats.

Cette approche implique la création de 6 groupes de participants auxquels on propose à chacun 6 requêtes tel que : (1) à l'intérieur de chaque groupe, il ne peut y avoir deux fois la même configuration de résultats ; (2) à chaque configuration correspond une requête différente dans le groupe ; (3) chaque combinaison requête-configuration est unique pour l'ensemble des groupes ; (4) chaque participant évalue les requêtes d'un seul des groupes. Nous n'avons cependant pas randomisé l'ordre des combinaisons dans chaque groupe. En effet, avec 6 combinaisons par groupe, la randomisation de leur ordre aurait nécessité au minimum 6! (c'est-à-dire 720) participants par groupe.

À la place, nous avons imposé le même ordre des requêtes pour tous les groupes (et par conséquent randomisé l'ordre des configurations). Ce choix n'implique pas de risque de biais important, car a priori les requêtes sont indépendantes, et comme un participant ne sera confronté à une requête donnée qu'une seule fois au cours de l'expérimentation, il n'y pas de risque lié à l'effet d'apprentissage.

Si l'on souhaite disposer d'une répartition uniforme des combinaisons (c'est-à-dire disposer d'autant de combinaisons identiques pour l'ensemble de l'expérimentation) cela implique de faire passer l'expérimentation à un multiple de 6 participants afin d'avoir autant de participants par groupe. Cependant, n'avons pas réussi à réunir assez de personnes dans les limites de temps disponibles pour atteindre 12 participants. Un échantillon de 6 participants nous semblant représenter un panel trop faible pour tirer des conclusions, nous avons pris un peu plus de participants afin d'obtenir des résultats plus étoffés. La conséquence est que la répartition des configurations ne peut pas être uniforme sur l'ensemble de l'expérimentation. Le risque de biais est cependant limité, en effet, seule la répartition des combinaisons (c'est-à-dire l'ordre des résultats pour une requête donnée) est affectée. Néanmoins, chaque configuration d'ordre de résultats reste évaluée le même nombre de fois que les autres, puisque dans chaque groupe, toutes les configurations sont présentes. Un biais pourrait apparaître s'il existait une influence entre la requête et la configuration (c'est-à-dire que la requête pourrait influencer la façon dont l'utilisateur réagit vis-à-vis de l'ordre des résultats), ce qui est peu probable. Il est cependant à noter que pour des travaux plus étoffés, il serait préférable de garantir cette uniformité et donc de disposer d'un nombre de participants multiple de 6.

Nous avons crée les SERPs simulées sous forme d'une série de diapositives, contenant consignes, requêtes, et pages de résultats, via le logiciel Ogama. Nous avons réalisé ces pages en langue anglaise afin de pouvoir faire appel à des participants non-francophones, et pouvoir comparer les résultats obtenus avec les travaux déjà existants sur le sujet. La première étape a consisté à trouver des requêtes non-évidentes qui se prêtaient bien à produire des résultats des trois types énoncés précédemment (positifs, neutres et négatifs). Pour chaque requête choisie, les trois types de résultats ont été extraits à partir des 4 premières pages d'un moteur de recherche connu. Les SERPs simulées ont ensuite été créées avec la police de caractères *Arial 15*, définie lors de l'expérimentation précédente. De plus, un exemple de SERP a été inséré dans les consignes données aux participants afin d'illustrer la forme que prennent les SERPS. Nous souhaitions ainsi limiter un possible effet d'apprentissage au cours de l'expérimentation, qui pourrait rendre l'évaluation des premières SERPs présentées moins pertinente que celles des dernières.

Au total, 11 personnes (2 femmes et 9 hommes) âgées de 20 à 27 ans et issues de notre laboratoire ont participé à l'expérimentation. Seulement 10 passations ont pu être exploités car les données de l'une des passations ont été perdues suite à un problème technique. Sur les 10 personnes restantes, toutes parlaient un anglais avancé et la moitié d'entre elles portaient une correction oculaire sous forme de lunettes de vue. Avant chaque passation, les participants recevaient oralement une description

de l'objectif de l'expérimentation et les consignes. Il leur a été demandé d'être le plus "naturel" possible et d'agir comme ils le feraient s'ils étaient seuls devant leur ordinateur. Notamment, aucune consigne de rapidité ni de performance ne leur était donnée. La seule contrainte qui leur était imposée était de garder une position de tête stable au cours de l'expérimentation, car celle-ci influe sur le fonctionnement de l'eye-tracker. Ils devaient ensuite effectuer les 6 requêtes prédéfinies, et pour chaque requête, cliquer sur la snippet correspondant au document qui leur semblait le plus pertinent dans la SERP.

3.2.2. Évaluation

Puisque nous travaillions dans l'optique de faire du RF, de nombreux mots des snippets ne nous intéressaient pas, et de plus, nous souhaitions caractériser les mots lus en fonction de leur apport supposé dans le processus de RF. Comme classiquement en recherche d'information, les mots appartenant à l'anti-dictionnaire du système de recherche d'information n'ont donc pas été considérés. Nous avons décidé d'autre part de considérer un autre traitement classique en recherche d'information, la troncature des mots. Il en résulte que, par exemple, si le participant regarde les mots "enter" et "entry", nous décidons qu'il s'agit du même mot [7]. Pour évaluer la qualité d'un mot regardé (c'est-à-dire son impact potentiel sur la requête si nous l'y ajoutions), nous proposons la classification suivante des mots présents dans les différentes pages de résultats : (1) Un mot est *positif* s'il peut apporter de l'information supplémentaire à la requête étudiée pour déterminer les documents pertinents ; (2) Un mot est *négatif* s'il peut apporter de l'information supplémentaire à la requête étudiée pour déterminer les documents non pertinents ; (3) Un mot est *neutre* si on ne peut évaluer son impact sur la requête étudiée.

Cette classification permet de donner une valeur à chaque mot lu sur chaque page de résultats, et pour chaque requête. Notons que les mots regardés qui apparaissent déjà dans la requête n'ont pas été considérés. En effet, ils n'apportent pas d'information supplémentaire au processus de RF. Les indicateurs que nous avons évalués sont la dernière fixation et la plus longue fixation, que nous étudions à deux niveaux de granularité : au niveau de la page de résultats dans son ensemble (c'est-à-dire toutes snippets confondues), et au niveau du résultat sélectionné par le participant (c'est-à-dire la snippet sélectionnée). Nous avons utilisé l'indicateur de plus longue fixation sur un mot, car comme les travaux de Eickhoff *et al*. [2015] l'ont montré, ce mot a souvent été utilisé par les participants pour reformuler manuellement leurs requêtes. Nous avons aussi utilisé l'indicateur de dernière fixation sur un mot, car celui-ci nous a semblé pertinent vis-à-vis du processus cognitif de l'utilisateur. En effet, nous pensions que le dernier mot regardé pouvait être celui ayant déterminé le choix du document le plus pertinent. Analyser ces deux indicateurs à deux niveaux de granularité différents (toutes snippets confondues et la snippet sélectionnée) nous a permis de

7. Nous avons utilisé l'implantation snowball, http://snowball.tartarus.org/algorithms/porter/stem.c, de l'algorithme de troncature de Porter.

comparer des résultats qui tiennent compte à la fois des perceptions et de l'action de l'utilisateur lorsqu'il choisit l'une des snippets.

3.2.3. Résultats

Pour la détermination des mots lus par les participants, nous avons repris le principe de détermination défini lors de la première expérimentation, c'est-à-dire que le mot considéré comme lu est celui dont l'isobarycentre est le plus proche de la fixation considérée. Comme pour la première expérimentation, nous avons superposé les pages de résultats aux fixations des participants (Fig. 1). Nous avons ensuite déterminé manuellement pour chaque participant et pour chacune de ses pages de résultats, les quatre indicateurs précédemment définis. Nous avons ainsi déterminé la valeur (positive, négative ou neutre) du dernier mot fixé et du mot le plus longtemps fixé, cela dans toute la page de résultats, puis en se limitant à la snippet sélectionnée. Comme indiqué précédemment, si un de ces mots se trouvait être un mot contenu dans la requête, nous ne le prenions pas en compte et nous analysions l'avant-dernière fixation ou la deuxième plus longue fixation.

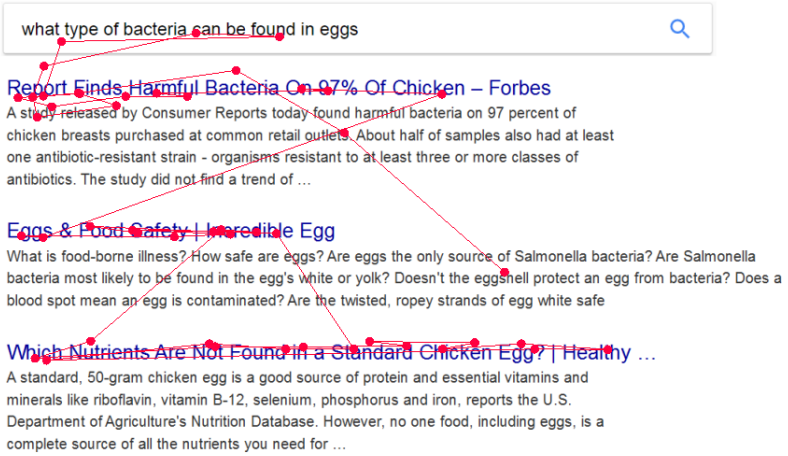

FIGURE *1. Exemple de page de résultats simulée, avec en surimpression le parcours visuel d'un participant (les points représentent les fixations).*

Les tableaux 2 et 3 mettent en parallèle les quatre indicateurs calculés. Sur le premier tableau, on trouve les résultats qui ne tiennent pas compte de la sélection d'une snippet, et sur le second, ceux qui en dépendent. La métrique qui nous intéresse principalement est le pourcentage de mots positifs, car celui-ci peut aider à réaliser une expansion de requête du fait que les mots analysés n'y apparaissent pas déjà. Nous voyons par exemple (tableau 2) que, dans 27 cas (45% des cas), le dernier mot fixé sur une page de résultats (sans prise en compte de la sélection) est un mot positif.

Lorsque l'on s'intéresse au mot fixé le plus longtemps, les résultats sont supérieurs et atteignent 45 cas (75% des cas). Nous constatons également deux tendances. Tout d'abord, pour la dernière fixation comme pour la plus longue, le pourcentage de mots positifs est plus important quand on s'intéresse uniquement à la snippet sélectionnée (tableau 3) plutôt qu'à la page de résultat dans son ensemble (tableau 2). On remarque aussi que ce pourcentage est encore plus important si on s'intéresse seulement à la plus longue fixation. Nous pouvons alors noter que dans 87% des cas où le mot le plus longtemps fixé est sur la snippet sélectionnée, c'est un mot positif. Alors que dans les mêmes conditions, le dernier mot fixé n'est positif que dans 51%. Ces résultats montrent également que la dernière fixation est moins pertinente que la plus longue. Concernant ce dernier point, il est possible que les participants, une fois le résultat le plus pertinent identifié, aient eu besoin de rechercher le pointeur de la souris pour sélectionner la snippet correspondante, entraînant ainsi des fixations non pertinentes vis-à-vis de la tâche de recherche d'information. En outre, il n'a pas été possible de proposer un indicateur pertinent basé sur les mots négatifs. En effet, nous avions supposé que les mots lus sur les snippets non sélectionnées seraient de type négatif, le participant les utilisant comme indicateur de non pertinence d'un document. En pratique, c'est le cas dans moins d'un tiers des situations (tableau 3). Les résultats ne nous semblent donc pas assez convaincants pour conclure sur cet aspect. En synthèse, nous déduisons de ces expérimentations que la plus longue fixation sur la snippet sélectionnée est le meilleur indicateur dans un contexte d'expansion de requête pour du RF.

Tableau 2. Résultats **sans** *prise en compte de l'action de l'utilisateur. Les valeurs indiquées sont le nombre de mots de chaque type (positif, négatif, ou neutre) regardés avec entre parenthèses le pourcentage par rapport au nombre total de mots tous types confondus.*

	Mots sur toutes les snippets		
Indicateur	positifs	négatifs	neutres
Dernier mot fixé	**27 (45%)**	9 (15%)	24 (40%)
Mot fixé le plus longtemps	**45 (75%)**	6 (10%)	9 (15%)

Tableau 3. Résultats **avec** *prise en compte de l'action de l'utilisateur. Les valeurs indiquées sont le nombre de mots de chaque type (positif, négatif, ou neutre) regardés avec entre parenthèses le pourcentage par rapport au nombre total de mots tous types confondus.*

	Mots sur snippet sélectionnés			Mots sur snippets non sélect.		
Indicateur	positifs	négatifs	neutres	positifs	négatifs	neutres
Dernier mot fixé	**23 (51%)**	5 (11%)	17 (38%)	4 (27%)	**4 (27%)**	7 (46%)
Mot fixé le plus longtemps	**34 (87%)**	0 (0%)	5 (13%)	11 (52%)	**6 (29%)**	4 (19%)

4. Conclusion

L'objectif de cet article était de présenter les résultats préliminaires de nos travaux, concernant l'utilisation de l'eye-tracking pour identifier des indicateurs basés sur les mots lus par l'utilisateur, sur une page de résultats lors d'une recherche d'information, ceci dans le but de les utiliser ensuite pour réaliser du bouclage de pertinence. Pour cela, nous avons réalisé deux expérimentations visant (1) à étudier la faisabilité de l'approche dans le contexte de la lecture d'une page de résultats, puis (2) à déterminer des indicateurs basés sur une analyse du comportement perceptivo-gestuel de l'utilisateur dans le contexte de l'utilisation d'un système de recherche d'information. Le protocole que nous avons défini lors de l'analyse comportementale était basé à la fois sur le suivi oculaire au niveau des mots de la page de résultats, et sur le suivi de l'action réalisée, c'est-à-dire la détermination de la snippet sélectionnée par l'utilisateur, permettant ainsi la génération d'indicateurs portant sur les types (positif, négatif, neutre) de mots lus dans la page de résultats consultée et/ou lus dans la snippet sélectionnée.

Nos résultats montrent qu'il existe des indicateurs, issus de l'analyse de l'activité perceptivo-gestuelle d'un utilisateur, pouvant potentiellement assister le bouclage de pertinence. En effet, nous montrons que sur une page de résultats classique composée de snippets, le mot le plus longtemps regardé sur le document choisi par l'utilisateur est le plus souvent un mot positif, et peut donc fortement aider à faire de l'expansion de requête. Cependant, les résultats obtenus ne nous ont pas permis pas de conclure sur d'autres indicateurs pressentis comme le dernier mot lu et/ou la lecture des mots négatifs.

Nos travaux étant préliminaires, ils présentent plusieurs limites. Une des principales limites réside dans le nombre relativement faible de participants inclus lors de l'expérimentation destinée à l'analyse comportementale (10), ainsi que, en corollaire, dans la non prise en compte de l'impact que peut avoir l'ordre des requêtes présentées aux participants (un nombre de participants multiple de 6 aurait été nécessaire). Ces éléments sont en effet importants pour neutraliser la variabilité inter-utilisateurs [Dumais *et al*., 2010] ou la salience de certaines zones [Liu *et al*., 2016] qui peuvent jouer un rôle dans notre cas. En outre, les expérimentations ayant été menées sur des pages de résultats comprenant seulement 3 documents, il faudra les confirmer sur des SERPs plus classiques contenant par exemple 10 documents, comme celles habituellement disponibles sur les moteurs de recherches sur le Web. Enfin, nos travaux n'ont pas pris en compte certaines caractéristiques importantes des SERPs pouvant influencer . Il s'agit notamment de la répartition des types de mots contenues dans les SERPS/snippets ou du rang de la snippet positive dans la SERP.

5. Perspectives

Une des principales perspectives de nos travaux concerne l'évaluation de l'efficacité des indicateurs proposés pour le bouclage de pertinence. Parmi les nombreuses approches possibles, nous pourrions par exemple déterminer quels sont les combinaisons les plus pertinentes d'indicateurs via une approche d'apprentissage machine

supervisé, en s'inspirant par exemple de Gwizdka [2014]. Il est également possible, de manière cohérente avec des approches d'expansion de requêtes, de filtrer les mots qui sont sémantiquement liés à la requête (par l'utilisation de plongements de mots par exemple, comme dans les travaux de Dogra *et al*. [2018]). C'est sur ce type de travaux que nous allons maintenant faire porter nos efforts.

Une deuxième perspective concerne plus généralement l'étude de nouveaux indicateurs prenant en compte la structure interne de la SERP. En effet, nous considérons actuellement de manière identique tous les mots lus, quel que soit le nombre d'occurrences des mots identiques ou de leur type (positif, neutre, négatif) dans la snippet ou la SERP, leur appartenance au texte ou au titre de la snippet, la position de la snippet à laquelle ils appartiennent dans la SERP (au début, au milieu, ou à la fin), et leur position dans la snippet. Or, comme l'indique Baeza-Yates [2018], les patterns de lecture des SERPs sont très spécifiques et de plus ont varié au cours du temps. Il est donc important de pouvoir disposer d'indicateurs plus précis car prenant en compte ces spécificités.

Une autre perspective, très prometteuse pour la suite de ce travail, concerne la génération des snippets. En effet, celles-ci peuvent avoir une grande influence sur les résultats, car les mots contenus dans les snippets peuvent être plus ou moins déterminants pour assister l'expansion de requête [Turpin *et al*., 2007]. La question serait de déterminer l'influence de la génération de snippets de paraphrases à la place de snippets d'extraits des documents [W.-F. Chen *et al*., 2018]. Nous pourrions même proposer une génération spécifique de snippets diversifiées contenant les termes capables de désambiguïser les résultats de la requête.

Plus généralement, il est à noter que notre protocole pourrait être réutilisé dans d'autres cadres liés à la recherche d'information. Par exemple, son utilisation est possible pour les évaluations sur des collections de tests. En particulier, dans le cas d'évaluations classiques avec des résultats gradués (par exemple : très pertinent / pertinent / non-pertinent), notre protocole pourrait être adapté pour tracer la valeur de pertinence associée au document (comme dans le système *Relevation* de Koopman, Zuccon [2014]). Dans ce cadre, avoir une connaissance plus fine des parcours oculaires amenant à sa décision serait très intéressant, en permettant par exemple d'identifier les termes ayant été potentiellement impliqués dans le processus cognitif d'évaluation.

Remerciements

Ces travaux de recherche ont été soutenus par l'équipe-action *OculoNimbus* du LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) financé par le programme "Investissement d'Avenir" et par le projet Émergence *GELATI* financé par le Laboratoire d'Informatique de Grenoble. Les auteurs tiennent également à remercier les personnes du laboratoire ayant participé aux expérimentations.

Bibliographie

Baeza-Yates R. (2018, mai). Bias on the Web. *Commun. ACM*, vol. 61, n° 6, p. 54–61. Consulté sur http://doi.acm.org/10.1145/3209581

Broder A. (2002, septembre). A taxonomy of web search. *SIGIR Forum*, vol. 36, n° 2, p. 3–10.

Buscher G., Elst L. van, Dengel A. (2009). Segment-level display time as implicit feedback: A comparison to eye tracking. In, p. 67–74. New York, NY, USA, ACM.

Chen W.-F., Hagen M., Stein B., Potthast M. (2018). A user study on snippet generation: Text reuse vs. paraphrases. In *The 41st international acm sigir conference on research & development in information retrieval*, p. 1033–1036. New York, NY, USA, ACM.

Chen Y., Zhang P., Song D., Wang B. (2015). A real-time eye tracking based query expansion approach via latent topic modeling. In *Proceedings of the 24th acm international on conference on information and knowledge management*, p. 1719–1722. New York, NY, USA, ACM.

Dogra N., Mulhem P., Goeuriot L., Amini M. (2018). Corpus d'entraînement sur les plongements de mots pour la recherche de microblogs culturels. In *CORIA*. ARIA.

Dumais S. T., Buscher G., Cutrell E. (2010). Individual differences in gaze patterns for web search. In *Proceedings of the third symposium on information interaction in context*, p. 185–194. New York, NY, USA, ACM.

Eickhoff C., Dungs S., Tran V. (2015). An eye-tracking study of query reformulation. In, p. 13–22. New York, NY, USA, ACM.

Ermakova L., Mothe J. (2016). Document re-ranking based on topic-comment structure. *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, p. 1-10.

Gwizdka J. (2014). Characterizing relevance with eye-tracking measures. In *Interaction in context symposium*, p. 58–67. New York, NY, USA, ACM.

Koopman B., Zuccon G. (2014). Relevation!: An open source system for information retrieval relevance assessment. In *Proceedings of the 37th international acm sigir conference on research & development in information retrieval*, p. 1243–1244. New York, NY, USA, ACM.

Liu Y., Liu Z., Zhou K., Wang M., Luan H., Wang C. *et al*. (2016). Predicting search user examination with visual saliency. In *Proceedings of the 39th international acm sigir conference on research and development in information retrieval*, p. 619–628. New York, NY, USA, ACM.

Manning C. D., Raghavan P., Schütze H. (2008). *Introduction to information retrieval*. Cambridge, UK, Cambridge University Press. Consulté sur http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Turpin A., Tsegay Y., Hawking D., Williams H. E. (2007). Fast generation of result snippets in web search. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval*, p. 127–134. New York, NY, USA, ACM.

Zhai C., Massung S. (2016). *Chapter 7 of text data management and analysis: A practical introduction to information retrieval and text mining*. San Rafael, CA, Morgan & Claypool.