BIAIS COGNITIFS ET RECHERCHE D'INFORMATION SUR INTERNET : QUELLES PERSPECTIVES POUR LES INDICATEURS DE PERTINENCE DES MOTEURS DE RECHERCHE

BOUTIN Eric
boutin@univ-tln.fr

IUT de Toulon Département TC – laboratoire I3M Bp 132 83957 la Garde cedex

Mots clefs :
surcharge informationnelle, biais cognitifs, dissonance cognitive, indicateur de pertinence, moteur de recherche, recherche d'information

Keywords:
Information overload, cognitive biases, relevance criteria, search engine, information retrieval

Palabras clave :
Sobrecarga de la información, motor de búsqueda, recuperación de datos

Résumé

La recherche d'information sur internet est une démarche qui met l'internaute dans une situation de devoir faire face à une quantité massive de données. Ces données sont parfois contradictoires. Elles sont bien souvent non validées : le travail de l'internaute n'en est que plus important. Comment réagit l'internaute face à une telle « surcharge informationnelle » ? Par quel processus parvient il à répondre à son besoin d'information en mobilisant les données qu'il a à sa disposition ?
Des travaux en psychologie se sont intéressés au processus de décision en environnement complexe et incertain. Surtout appliqués dans le domaine de la gestion, ils apportent un éclairage intéressant à la notion de « rationalité limitée de la décision ». Nous proposons de transposer ces travaux pour une meilleure compréhension de la construction de connaissances par l'internaute suite à une requête adressée à un moteur de recherche. Nous affinerons la notion de « biais cognitifs » qui correspond à des mécanismes de protection inconscients qui vont permettre à l'internaute de gérer le problème de surexposition à l'information de façon plus confortable.

- Dans un premier temps, nous présenterons de manière conceptuelle et appliquée la notion de biais cognitif et l'intérêt de cette notion dans le domaine de la recherche d'information sur internet.
- Dans un second temps, nous proposerons une vision renouvelée de la pertinence ainsi que quelques pistes permettant une meilleure prise en compte de ces mécanismes dans les algorithmes de pertinence des moteurs de recherche.

1 Le processus de recherche d'information à la lumière des biais cognitifs :

Lorsqu'un internaute a une recherche d'information à effectuer, il va passer par une série itérative de boucles composées chacune de deux grandes étapes. La première est l'étape de formalisation d'une question en un format intelligible par un moteur de recherche. La seconde est une étape de construction de connaissances à partir des données renvoyées par l'outil de recherche. Lors de chacune de ces étapes, des biais peuvent se produire. Certains résultent d'une prédisposition émotionnelle ou intellectuelle envers certains jugements, d'autres sont inconscients. On les appelle alors biais cognitifs. La définition du biais cognitif selon Heuer [10] est une « erreur mentale causées par des stratégies de traitement simplifié de l'information». Les biais cognitifs vont intervenir à deux niveaux dans le processus de recherche d'information : ils altèrent la qualité de la transcription du besoin d'information en requête ou la rationalité du processus de construction de connaissances.
Des travaux récents ont été conduits par Alonso et al [1] pour identifier et combattre les biais cognitifs se situant en amont de la requête. Il s'agit par exemple de développer des systèmes anti biais (Anti Bias Mecanism) permettant de générer des requêtes alternatives à celles de l'internaute en utilisant par exemple des dictionnaires de synonymes.

Dans ce travail, nous allons nous intéresser exclusivement aux biais intervenant en aval de la construction de la requête.
La recherche d'information sur le web peut être apparentée à un processus de stimulation fort qui doit être canalisé par l'internaute. L'internaute doit en effet capturer des stimuli externes, les analyser, les combiner avec des informations antérieures pour les transformer en connaissance et en savoir.
Il est intéressant de considérer avec Lévy Garboua [9] le caractère séquentiel de la perception par l'internaute de cette information massive. Lorsqu'il effectue une recherche d'information, l'internaute n'accède pas immédiatement à toute l'information. Il va traiter cette information de manière séquentielle et procédurale. Cette approche séquentielle prive l'internaute d'une vision d'ensemble : celle-ci se construit petit à petit et la vision de départ évolue sous l'influence des stimulations que reçoit le cerveau. Ce modèle séquentiel de la perception humaine peut fonder une théorie féconde de la rationalité limitée. En effet, « ce que nos capacités d'information et de calcul sont incapables de faire instantanément, elles le réalisent sans peine en révisant à la marge le choix fait en t-1 » [9].
La surabondance de l'information se double d'une incertitude : elle conduit l'internaute à développer des mécanismes de protection inconscients appelés biais cognitifs. Les biais cognitifs ont pour objectif de simplifier la complexité de la réalité, d'éviter la surcharge informationnelle. Ce réflexe est une nécessité pour garantir la santé mentale de l'internaute.

Plusieurs phénomènes sont à l'œuvre. Nous proposons d'en présenter trois principaux : la « rigidité cognitive », la « dissonance cognitive », la « loi des petits nombres »

1.1 La rigidité cognitive :

Ce principe a été largement étudié dans le contexte de prise de décision dans les organisations par Robson [11]. Robson parle de « rigidité cognitive » pour qualifier le phénomène par lequel des organisations vont limiter leur capacité à développer des alternatives du fait de cadres cognitifs qui les restreignent. Ces cadres cognitifs sont constitués de l'ensemble des croyances et valeurs du groupe largement empruntées au passé. Si dans une perspective de court terme, il est satisfaisant de considérer que les valeurs sont invariantes, à long terme ce « verrou cognitif » doit sauter pour ne pas rendre l'organisation aveugle à la perception de signes émergents
On peut transposer ce phénomène au contexte de la recherche d'information sur internet. Dans un contexte d'information surabondante, l'internaute ne peut pas accorder autant de poids aux milliers de réponses qui lui sont renvoyées par le moteur. Celui-ci va donc se concentrer sur les premières pages et au sein des premières pages il va accorder beaucoup de poids aux toutes premières réponses. L'internaute a une limitation intrinsèque dans sa capacité à absorber la nouveauté. Il va accorder beaucoup de poids aux premiers documents qu'il va visualiser. C'est « l'effet d'ordre ». Ces premiers

documents auront un fort encrage et vont construire sa représentation dominante du problème à résoudre

1.2 La dissonance cognitive :

La dissonance cognitive s'intéresse à ce qui se passe lorsque l'internaute perçoit une information qui n'est pas confirmatoire de l'information qu'il a cumulée.
Le principe de dissonance cognitive introduit par Festinger [5] considère que « l'existence simultanée d'éléments de connaissance qui, d'une manière ou d'un autre, ne s'accordent pas (dissonance) entraîne de la part de l'individu un effort pour les faire, d'une façon ou d'une autre, mieux s'accorder (réduction de la dissonance) ». Les manipulateurs recréent artificiellement des états de dissonance cognitive. Quand on veut convaincre quelqu'un d'une opinion qu'il ne partage pas, on l'amène à accomplir des actes peu impliquant au départ mais qui va dans le sens du changement d'opinion. En situation de dissonance cognitive, il changera alors d'opinion pour retrouver le confort mental initial. C'est ce qui fait dire à Beauvois et Joule [7] [8] que « l'homme n'agit pas en fonction de ses pensées, mais pense en fonction des actes que les " circonstances " lui ont imposés ».
Dans le contexte présent, on s'intéresse à la dissonance cognitive associée à la prise en compte par l'individu d'une information qui ne renforce pas la connaissance qu'il a accumulée sur le sujet. Un nouveau document différent de la représentation que l'internaute s'est forgée va introduire un phénomène de dissonance cognitive. La solution pour revenir à un état normal est de ne pas prendre en compte ce nouveau document. Ainsi, il sera très difficile pour un document qui n'est pas confirmatoire des précédents d'être retenu comme pertinent. L'internaute aura tendance à retrouver dans les documents futurs l'idée première qu'il s'est forgée sur les documents initiaux. Il aura ainsi des difficultés à abandonner sa première solution pour une autre. La rectification d'idées acquises est plus pénible pour un individu que l'apprentissage d'idées nouvelles pour lesquelles il ne possède pas encore de modèle. Aronson [2] considère que la dissonance cognitive sera d'autant plus forte que l'effort aura été élevé pour acquérir la connaissance initiale.
On voit ainsi que l'ordre de lecture des résultats d'un moteur de recherche aura une influence sur le poids accordé à tel ou tel document. Si l'internaute parcourt un ensemble de n documents dans un ordre différent, il n'aura pas la même connaissance au final.

1.3 « La loi des petits nombres »:

Cette loi, présentée par Tversky et al. [12] est une parodie de la loi des grands nombres utilisée en statistique. La loi des grands nombres repose sur un principe selon lequel des mesures conduites à partir d'échantillons représentatifs d'une population mère pourront ensuite être transposés à la population entière. Parler de « loi des petits nombre » correspond à l'idée selon laquelle les internautes ont du mal à apprécier la représentativité de l'information qu'ils exploitent et attribuent une confiance excessive à des conclusions issues de l'exploitation d'information obtenues à partir d'échantillons non significatifs. L'internaute a une mauvaise évaluation de la représentativité d'une information et peut par exemple considérer une information redondante ou une information fortement corrélée à la précédente comme confirmatoire d'une information initiale. Ce principe se retrouve dans l'observation des usages des internautes. 80% des connaissances issues de recherche d'information se forgent à partir des 10 premiers résultats renvoyés par le moteur de recherche

2 Vers une vision renouvelée de la pertinence

Une revue de la littérature ne nous a pas permis d'identifier de travaux antérieurs sur la relation entre biais cognitif et recherche d'information. Aussi proposons nous d'intégrer les biais cognitifs dans un modèle renouvelé de la pertinence

La pertinence est une notion centrale en information retrieval. Tout processus de recherche d'information est orienté vers la réponse à un besoin qu'il s'agit de satisfaire au mieux. D'où la nécessité d'introduire des mesures de la satisfaction de ce besoin. L'indicateur de pertinence est une

fonction complexe qui, en fonction d'un sujet et d'un contexte donné, va permettre de hiérarchiser les résultats que le moteur de recherche va renvoyer à l'internaute.

Si on a un regard diachronique sur la notion de pertinence au sens où l'entendent les moteurs de recherche, on s'aperçoit qu'on a connu depuis 10 ans plusieurs familles d'indicateurs de pertinence qui se combinent aujourd'hui largement.

- La première famille correspond aux indicateurs de pertinence basée sur l'analyse du contenu. Largement intrinsèque, cette analyse définit la pertinence d'une page web à partir de la présence des mots clés de la requête dans la page web à qualifier.
- La deuxième famille mise en œuvre et inspirée de Google s'appuie sur l'analyse relationnelle. C'est alors l'étude des liens hypertextes qui permet de définir de façon exogène la pertinence d'une page web.
- L'approche business retenue par Overture considère qu'une page est pertinente si son concepteur est prêt à payer le prix pour figurer dans les premières réponses du moteur.

Il existe un invariant entre toutes ces approches de la pertinence. Elles présentent toutes une vision technique de la pertinence de laquelle l'internaute est absent. La pertinence d'un document est indépendante de la pertinence des autres documents de la collection. La pertinence d'un document est présentée comme une donnée objective, inhérente à ce document [6]. Il est alors possible de calculer la pertinence de chaque document d'une collection et de les hiérarchiser ensuite par pertinence décroissante. La pertinence de la collection correspond donc à la somme des pertinences des documents la constituant. Aujourd'hui cette génération d'indicateurs de pertinence traverse une phase de rendement décroissant. Les moteurs de recherche doivent investir de plus en plus pour se démarquer de la concurrence et maintenir leur avantage concurrentiel. Il y a donc la place pour une approche de la pertinence qui s'inscrive en rupture par rapport à l'existant.

Des évolutions récentes ont altéré ce modèle. La pertinence n'est alors plus la caractéristique intrinsèque d'un document. La pertinence est évaluée par l'utilisateur en fonction du contexte dans lequel il se trouve. Ainsi, le document visualisé par l'internaute à un instant t aura une pertinence qui dépendra des documents antérieurs que cet internaute aura pu visualiser dans un passé plus ou moins éloigné. On parlera alors de pertinence conditionnelle en respect aux autres documents contenus dans le corpus de données.

La question est alors de savoir de quelle façon la pertinence d'un document peut être affectée par des documents antérieurs. Une revue de la littérature nous permet de voir que la notion de redondance est au cœur de cette interaction.

Si un document est redondant par rapport à un document antérieur, alors, il aura une pertinence nulle. La redondance est donc négative dans cette acception. Carbonell et Goldstein se sont attachés à modéliser la notion de pertinence en proposant de combiner pertinence et anti-redondance. Ils classent les résultats du moteur d'une part selon leur pertinence par rapport à la requête et d'autre part selon leur degré de nouveauté par rapport aux autres documents de la collection. Les auteurs proposent une mesure de la nouveauté d'un document à travers le calcul de la similarité textuelle entre le document courant et les documents antérieurement lus par l'internaute. Ils définissent le concept de pertinence marginale maximale (Marginal Maximal Relevance) qui est une combinaison linéaire de la pertinence par rapport à la requête et la nouveauté. Un document a une pertinence marginale maximale s'il répond à la requête et contient peu de similarité avec les documents précédents.

Nous considérons pour notre part que la redondance peut être positive dans deux familles de requêtes :

- Dans le cas des requêtes navigationnelles, [4], l'internaute cherche à retrouver l'adresse d'une page web qu'il a bien souvent déjà visualisée. Dans ce cas, qui correspond d'après l'auteur à une requête sur 4, la page qui est recherchée sera d'autant plus pertinente qu'elle correspondra à la page antérieurement recherchée.
- Dans le cas de requêtes informationnelles, la redondance signifie la confirmation. Lorsque l'internaute découvre un sujet qu'il ne connaît pas, il visualise plusieurs réponses parfois redondantes. Cette redondance est positive car elle va servir de validation subjective du crédit qui va être accordé à l'information. Plus un grand nombre de documents supposés indépendant vont dans le sens d'une information et plus celle-ci aura de valeur aux yeux de l'internaute

A la lumière des travaux sur les biais cognitifs, nous proposons un modèle de la pertinence intégrant certains biais cognitifs.

2.1 modèle de pertinence intégrant les biais cognitifs:

Il existe deux niveaux pour juger de la pertinence : le niveau du document courant et le niveau composé de l'ensemble des pages renvoyées par la requête et visualisées par un internaute. Certains biais cognitifs seront introduits dans notre modèle au niveau de l'évaluation du document courant, d'autres au niveau de l'évaluation des documents visualisés dans la requête. Nous allons nous attacher dans un premier temps à définir la pertinence du document courant.
Simplifions pour l'instant le problème en restreignant le contexte à celui de la requête. Le problème pourrait ensuite être étendu au niveau de contextes plus large.
La pertinence d'une page web fait intervenir la position de cette page dans le parcours de l'internaute. Ainsi on notera :

(1) Pij, la pertinence de la page i visualisée en jième position.
(2) Pj, la pertinence de la page visualisée en jième position

Supposons que l'on s'intéresse à la mesure de la nouveauté apportée par un document web que l'internaute découvre. Nous allons essayer de mesurer le supplément d'information ou information marginale apportée par ce document dans un processus de navigation qui a pu conduire l'internaute à visualiser d'autres documents auparavant. Ce taux de nouveauté est le complémentaire à 1 d'un indicateur de redondance qui se calcule par une mesure de similarité du contenu textuel de la page courante avec le contenu textuel d'une page fictive correspondant à la concaténation de toutes les pages déjà visualisées. Le calcul de similarité relative au contenu textuel d'une page web est un problème classique que nous considérons comme déjà résolu. L'indicateur de redondance vaut 1 si la page courante n'apporte pas d'information par rapport aux pages antérieurement consultées. L'indicateur de redondance vaudra 0 dans le cas ou la page courante est complètement différente de l'ensemble composé par les pages déjà visualisées. La redondance de la nième page visualisée par rapport aux n-1 ième autre sera notée Rn,[1-n-1]

(3) Rn,[1-n-1] le taux de redondance entre la nième page visualisée et une page artificielle constituée à partir de la concaténation des n-1 premières.

Nous proposons dans un premier temps, un modèle simple qui intègre la notion de redondance « négative ». Ce modèle sera enrichi petit à petit.
Si on prend en compte le rôle négatif de la redondance, alors, la pertinence d'une page sera pénalisée lorsque cette page est visualisée après des pages qui lui sont redondantes. Ainsi nous proposons de définir la pertinence de la page i que l'internaute découvre en position 2

(4) Pi2, = Pi1X(1- R2,1)

La pertinence de la page i visualisée en position 2 est donc égale à la pertinence de cette page si elle avait été visualisée en position 1 corrigée par la taux de redondance que cette page a avec la page visualisée en position 1. Ainsi si la redondance est totale, R2,1 vaut 1 et donc la pertinence de la page i est nulle.
Si on étend maintenant à ce qui se passe au niveau de la nième page visualisée par l'internaute on a :

(5) Pin = Pi1X[1- Rn,[1—n-1]

Nous allons prendre en compte le phénomène de dissonance cognitive associé à la surcharge informationnelle. Ce phénomène va conduire l'internaute à privilégier une information antérieurement acquise et à accepter difficilement une information qui n'est pas confirmatoire. Il s'en suit que même si une page apporte une information marginale forte par rapport aux pages du contexte, l'information

ne sera pas récupérée complètement et ce d'autant plus que la quantité d'information acquise préalablement aura été forte. La pertinence d'une page est donc inversement proportionnelle à la somme des pertinences des pages antérieurement visualisées modulo □. La valeur de □ dépend de la capacité de l'internaute à gérer une information importante.
La formule de la pertinence de la page i à l'étape n devient alors la suivante :

(6) Pin = Pi1X[1- Rn,[1—n-1]*□/(□ i=1→n-1 Pi)

On peut maintenant exprimer la redondance positive. Ce phénomène va conduire à affecter une valeur à une information même si celle-ci est redondante d'une autre. En effet, une information redondante valide l'information primitive donc renforce son crédit. Nous proposons d'introduire cette notion au niveau de la pertinence collective des documents consultés.
Supposons que nous ayons n pages dans la collection
Supposons que l'internaute consulte m pages parmi les n
Soit A=P1+P2+…+Pm A correspond à la pertinence de la collection
Soit B= P11+ P21+ … +Pi1+…+ Pm1 B correspond à la pertinence des pages du corpus lorsqu'on considère que la pertinence d'une page n'interfère pas sur la pertinence des autres.
A/B mesure le taux de nouveauté dans la collection. Si ce rapport est proche de 1, cela signifie que les documents ont tous un contenu différent. Plus ce rapport est faible (proche de 0) plus cela signifie que la redondance est forte.
Nous proposons une définition corrigée de la pertinence de la collection qui introduit le niveau de redondance entre les documents de la collection :
Pertinence corrigée de la collection = □ Pi+□ Pi*□(1-B/A)
La pertinence de la collection se voit majorer d'un bonus inversement proportionnel à la diversité modulo β.

2.2 la prise en compte des biais cognitifs : quelle perspective pour les moteurs de recherche :

Nous proposons de présenter deux types d'initiatives qui prennent en compte les mécanismes des biais cognitifs.

Parmi les moteurs et métamoteurs de recherche existant, certains proposent une interface dans laquelle les réponses au lieu d'être présentées sous forme de pages séquentiellement ordonnées sont proposées classées dans diverses rubriques. Ce type de classification montre à l'internaute la diversité de ce qu'il y a à voir dans la requête et proposent une vision d'ensemble. Ce type d'affichage est de nature à limiter les processus de focalisation de l'internaute sur une seule réponse.

Certains travaux plus expérimentaux se sont intéressés à la construction de nouveaux indicateurs de pertinence pour les moteurs de recherche. Ainsi Benyu et al [3] décrivent un processus dans lequel la pertinence d'une page web n'est plus définie par rapport à la requête de l'internaute mais par la contribution de cette page à l'information contenue dans l'intégralité des pages analysées. Les auteurs proposent deux métriques appelées diversité et richesse. La diversité mesure la quantité d'information contenue dans un ensemble de pages web renvoyées par un moteur de recherche suite à une requête. La richesse correspond à la capacité d'une page web à apporter une réponse la plus exhaustive possible à la diversité du corpus. Cette analyse revient à accorder un poids renouvelé aux indicateurs de pertinence reposant sur les analyses de contenu qui deviennent cette fois contextuelle. La pertinence des pages est alors construite à l'issue d'un processus calculatoire original, directement transposé du Pagerank de Google mais appliqué à une matrice des similarités textuelles entre documents plutôt qu'à une matrice des relations hypertextuelles entre pages web.
L'indicateur de pertinence résultant de ce genre d'algorithme renvoie, dans les premiers résultats, des pages non redondantes permettant de rendre compte du mieux possible de la diversité des réponses possibles. Ainsi l'internaute découvrira dans les premières réponses une grande variété de réponses possible non redondantes.

Conclusion :

La recherche d'information sur le web s'apparente à un processus dans lequel le chercheur d'information est en permanence en situation de surcharge informationnelle. Pour pouvoir préserver sa santé mentale dans ce processus d'exposition permanente à un flux d'information incessant, l'internaute doit réaliser un certain nombre de raccourcis inconscients :

- il privilégie les premières réponses renvoyées par l'outil de recherche en faisant l'hypothèse qu'elles seront représentatives de l'ensemble.
- Il construit sa connaissance de manière séquentielle et considère souvent comme confirmatoire une information redondante.
- Plus une première idée aura été forgée, plus il sera difficile pour lui d'appréhender la contradiction à sa juste valeur.

Les moteurs de recherche actuellement en vigueur ne prennent que peu en compte cette dimension psychologique de la recherche d'information sur internet. En effet, par construction, les indicateurs de pertinence raisonnent surtout dans une logique « precision & recall » sans étudier de quelle façon les documents sont traités et appréhendés par l'utilisateur final. La prise en compte des phénomènes de biais cognitifs ouvre des perspectives intéressantes aux moteurs de recherche dans une démarche d'indicateur de pertinence centré utilisateur.

3 Bibliographie

[1] ALONSO R., HUA L (2005), Combating Cognitive Biases in Information Retrieval, Actes du colloque 2005 international conference on intelligence analysis, 2-4 mai 2005

[2] ARONSON E., (1973) ''The Rationalizing Animal," Psychology Today, May 1973, pp. 46–51.

[3] BENYU Z.; HUA L.; YI L.; LEI J.; WENSI X.; WEIGUO F.; ZHENG C.; WEI-YING M., (2005), Improving Web Search Results Using Affinity Graph, conference WWW 2005

[4] BRODER A., (2002), : \A taxonomy of web search", ACM SIGIR Forum, Vol.36, No.2, pp. 3-10 (2002).

[5] FESTINGER, L. (1957). A Theory of Cognitive Dissonance. Stanford, CA: Stanford University Press.

[6] HEINE, M.H. (2000). "Reassessing and Extending the Precision and Recall Concepts," In www.ewic.org.uk/ewic. Revised version of "Time to dump 'P and R'?" Proceedings of the MIRA '99: Final MIRA Conference on Information Retrieval Evaluation, Glasgow, 14-16 April 1999: 61-74.

[7] JOULE R.-V., BEAUVOIS, J.-L. (1987), nouvelle version 2002. *Petit traité de manipulation à l'usage des honnêtes gens*. Grenoble, Presses Universitaires de Grenoble.

[8] JOULE R-V, BEAUVOIS J-L « La soumission librement consentie : Comment amener les gens à faire librement ce qu'ils doivent faire ? »

[9] LEVY GARBOUA (2004), "Perception séquentielle et rationalité limitée", Journal des Economistes et des Etudes Humaines, 14, 63-77.

[10] HEUER R. J. (1999), Psychology of Intelligence Analysis, History Staff Center for the Study of Intelligence Central Intelligence Agency

[11] ROBSON D., (2005), Cognitive Rigidity: methods to overcome it, Actes du colloque 2005 international conference on intelligence analysis, 2-4 mai 2005

[12] TVERSKY A and KAHNEMAN D, (1971) "Belief in the law of small numbers", Psychological Bulletin, 1971, Vol. 76, No. 2. 105-110.