PERSPECTIVES DOCUMENTAIRES SUR LES MOTEURS DE RECHERCHE : ENTRE SÉRENDIPITÉ ET LOGIQUES MARCHANDES

Olivier Ertzscheid, Gabriel Gallezot, Eric Boutin

PERSPECTIVES DOCUMENTAIRES SUR LES MOTEURS DE RECHERCHE : ENTRE SÉRENDIPITÉ ET LOGIQUES MARCHANDES

DOCUMENTARY PERSPECTIVES ABOUT SEARCH ENGINES : BETWEEN SERENDIPITY AND COMMERCIAL LOGICS.

Olivier ERTZSCHEID
Université de Nantes
Laboratoire Document et sciences de l'information (DOCSI)
olivier.ertzscheid@univ-nantes.fr

Gabriel GALLEZOT
Université de Nice Sophia-Antipolis
Laboratoire Information Milieux, Médias, Médiation (I3M)
gabriel.gallezot@unice.fr

Éric BOUTIN
Université du Sud Toulon Var
Laboratoire Information Milieux, Médias, Médiation (I3M)
boutin@univ-tln.fr

Résumé : Le monde de la recherche d'information est actuellement en train de vivre une période d'intense bouleversement : position hégémonique du moteur Google, question délicate de l'interpénétration des sphères publiques et privées, redocumentarisation du monde, montée en puissance de la logique publicitaire et sa cohabitation avec le modèle « régulé » de la simple application d'algorithmes. De nouvelles modalités d'accès apparaissent, telle celle de la sérendipité que nous interrogeons, après l'avoir resituée dans l'héritage de la bibliométrie, au regard des modèles théoriques de la recherche d'information, pour isoler le rôle d'adjuvant indispensable qu'elle occupe désormais. Son instrumentalisation par les moteurs, sa perception liée au niveau d'acculturation socio-technique des usagers, la diversité de ses instanciations, pose la question de l'opacité des algorithmes et de la nécessaire ouverture d'un débat autour d'un écosystème non plus simplement documentaire mais politique.

Mots-clés : Recherche d'information. Moteurs de recherche. Google. Sérendipité. Bibliométrie.

Abstract : Information retrieval (IR) using search engines is engaged in a deep evolution due to the hegemony of Google, the overlapping spaces between a public and a private or intimate web, a process of redocumentarisation, and the coexistence of regulated and algorithmic models on the one hand and of commercial logics on the other hand. Considering classical models of information retrieval, serendipity is the most proeminent actual modality of IR. Because of its instrumentation by search engines, because of its perception regarding technical and cultural background of users, and because of its diverse instanciations, serendipity raises the question of algorithm's opacity. Therefore, a public debate around the political ecosystem of search engines is nowadays a necessity.

Keywords : Information retrieval. Search engines. Google. Serendipity. Bibliometry.

INTRODUCTION.

La réalité sociale du web compte aujourd'hui près d'un milliard d'internautes (Nielsen 2005). Sa réalité documentaire fait état d'estimations autour de 30 milliards de pages web[1] en Février 2007, alors que la dernière version – 2003 – de l'étude quantitative systématique menée à Berkeley (Lyman & Varian 2003) faisait état d'un web « de surface » – c'est à dire indexé ou indexable par les moteurs – représentant 167 térabits d'information. Au vu de ces chiffres qui confinent à l'incommensurable on comprend mieux la nécessité de l'existence d'outils de repérage, de classification et d'accès. Ces moteurs de recherche ont vu le jour dès que le nombre de serveurs web est devenu ingérable à l'aide de simples signets ou de listes d'adresses. Le web mène ainsi la problématique documentaire au sens large à son climax, aussi bien en termes de conservation qu'en termes d'accès.

Bien au-delà de l'existence même de ces outils, c'est la question de leur prise en main, de l'acculturation technique qu'ils réclament et de leurs modes internes de fonctionnement comme artefacts socio-techniques qui pose problème. La question nodale est celle des modèles qui sous-tendent la manière dont s'organisent les informations et les connaissances de cet écosystème documentaire, à l'échelle de la planète connectée. Dans ces modèles, la dichotomie entre des logiques marchandes d'une part, et des logiques classificatoires reposant sur un certain nombre de principes objectivables de la recherche et de l'accès à l'information d'autre part, est chaque jour plus importante. Entre les deux, une certaine forme d'aléatoire – la sérendipité – vient désormais s'inscrire au cœur même du processus de recherche.

1. CARTE DES ACTEURS ET TERRITOIRES DOCUMENTAIRES.

1.1 GOOGLE ET LA PUISSANCE MÉTONYMIQUE

La caractéristique principale de la toile mondiale n'est pas tant qu'elle a permis de rendre disponibles des milliards d'informations, mais surtout qu'elle a amené des millions d'utilisateurs à faire de la recherche d'information et de la recherche « documentaire » une tâche quotidienne. Dans cette tâche, les moteurs de recherche en général, et Google en particulier sont généralement les seuls intermédiaires entre l'expression d'un besoin et sa partielle ou totale satisfaction.

En France, le moteur Google représente à lui seul près de 75% de l'ensemble des recherches effectuées[2]. Cette proportion pour les Etats-Unis, bien que plus faible (50%)[3], marque cependant une claire domination du moteur, qui ne se mesure pas simplement à l'aune des parts de marché qu'il représente, mais bien plus sûrement à son adoption généralisée, avec par exemple le passage dans l'usage de l'expression « Googler quelque chose ou quelqu'un » qui désigne métonymiquement l'ensemble du processus de recherche sur Internet. Ce quasi-monopole est particulièrement problématique, et ce indépendamment de l'infobésité que le moteur est censé pallier en y réinjectant de « l'ordre ». Un ordre qui, chez Google comme chez tous les moteurs majeurs, est encore celui de la « liste » (Goody, 1979) et que vient contrebalancer l'ouverture choisie par quelques concurrents sur d'autres modes de représentation et de navigation au sein des résultats : cartographie (www.kartoo.com, www.grokker.com …), catégorisation (www.exalead.com, www.clusty.com, …), ou recommandation (www.amazon.com).

1.2. DE LA RECHERCHE À LARGE SPECTRE …

---

[1] http://www.boutell.com/newfaq/misc/sizeofweb.html
[2] Source : IT Facts http://blogs.zdnet.com/ITFacts/
[3] http://searchenginewatch.com/showPage.html?page=3625336

Dans les différentes manières de sérier le processus de recherche d'information, plusieurs approches permettent de situer les enjeux et les pratiques actuelles, selon l'usage individuel (requêtage), collectif (aspects communautaires), ou technique (fonctionnalités des outils). Une première sériation des usages permettra de mieux analyser ensuite (cf infra, point 3) leur montée en charge instrumentale dans les interfaces des différents outils, les modèles théoriques de la recherche d'information qui les sous-tendent, et les biais nouveaux auxquels ils s'exposent. (Broder, 2002) établit une taxonomie des recherches en fonction du "*besoin derrière la question*", distinguant ainsi trois classes :

- des recherches navigationnelles visant à retrouver une page particulière,
- des recherches informationnelles qui nécessitent la consultation de plusieurs pages
- des recherches transactionnelles enfin, qui manifestent un désir d'accomplir une action comme un achat en ligne.

Au moment de l'étude (2002) la répartition entre ces requêtes est respectivement de 50%, 20% et 30%. Le panorama actuel des outils, conjugué à l'explosion des sites spécifiquement marchands et des comparateurs de prix, ainsi qu'à la levée progressive – en terme d'habitus de consommation – des réticences à l'achat sur internet, permettent de penser que la dernière catégorie doit aujourd'hui avoir connu une croissance significative.

Cette première sériation doit être affinée par la prise en compte de l'aspect de plus en plus communautaire de certains sites de recherche, qui se servent de la somme des expertises individuelles et/ou de l'analyse des actes d'achat individuels pour alimenter ensuite des systèmes de recommandation dont l'amplitude peut osciller entre le cercle plus ou moins large de pairs et/ou des amis et celui beaucoup plus large de l'ensemble de la communauté des utilisateurs de l'outil.

1.3. … A LA RECHERCHE DU « FAIT » DOCUMENTAIRE

A cette première grille interprétative de diversification des requêtes et des usages individuels et collectifs, s'en ajoute une seconde concernant la fragmentation de la notion de document, qui ouvre la possibilité d'une redocumentarisation massive (Salaün, 2007), laquelle mixe de manière inédite et avec une perméabilité nouvelle des espaces et des marqueurs sociaux jusque-là bien distincts. « Est » aujourd'hui document ou « est » de nature documentaire toute trace, inscription, support, flux ou échange numérique, à compter du moment où celui-ci apparaît, même de manière fugitive ou volontairement éphémère sur le réseau : une image, un texte, un extrait de page, un courrier électronique, quelques lignes de conversations sur messagerie instantanée, une vidéo, etc.

Par ailleurs, l'offre de services proposée par Google ainsi que par Microsoft et Yahoo !, ouvre à une même entité commerciale (le moteur lui-même) des possibilités d'accès à un ensemble d'informations qui relèvent indistinctement de la sphère publique (les sites que nous consultons, les documents auxquels nous accédons), de la sphère privée/professionnelle (les documents contenus sur notre disque dur[4], ceux que nous partageons via une offre bureautique « en ligne »), et de la sphère intime (nos courriels, l'information publiée sur nos blogs, certains de nos comportements d'achat). Avec cette nouvelle configuration de nos espaces informationnels, émerge une nouvelle écologie cognitive, un nouvel habitus, qui fixe l'interpénétration de deux sphères : celle de la recherche d'information « publique » (web) et celle de la recherche d'information sur des données personnelles (web privé et web intime). Ce changement de paradigme concerne l'accès à des informations fondamentalement différentes dans leur conception, dans leur usage, dans leur nature et dans leur capacité à être indexées.

[4] via des outils d'indexation de type GoogleDesktop (http://desktop.google.com)

Se pose alors la question d'observer, en terme d'usages raisonnés comme en termes de déviances possibles, les algorithmes et les pratiques marchandes qui permettent aujourd'hui de fouiller indistinctement ces trois espaces, en s'interrogeant sur l'arrivée massive et la montée en puissance des logiques publicitaires (liens sponsorisés) et la cohabitation de ce modèle avec celui, « régulé » de la simple application d'algorithmes ou de modèles documentaire éprouvés.

2. BIBLIOMÉTRIE ET SÉRENDIPITÉ : PROBLÉMATIQUES D'UN HÉRITAGE.

2.1. DE LA BIBLIOMÉTRIE AU « PAGERANK ».

L'algorithme PageRank qui fît le succès du moteur de recherche Google, dispose de fortes similarités avec les fondements de la bibliométrie telle que définie par (Garfield, 1972) pour la mise en œuvre du Science Citation Index (SCI). En lieu et place du nombre de citations d'un article scientifique chez Garfield, la pertinence d'une page web est définie à partir du nombre et de la pertinence des pages qui la citent (backlinks), mesure quantitative utilisée pour construire des métriques qualitatives. C'est bel et bien d'une transposition qu'il s'agit, à cette différence près que dans le cas de la bibliométrie, les filtres éditoriaux naturels de l'édition scientifique sont garants d'une limitation des biais[5] sur l'aspect qualitatif des résultats, alors que dans le cas du web, l'absence de filtres et de médiations documentaires scientifiques et spécifiques donne lieu à des usages qui font *in fine* la part belle à des logiques d'errance – au mieux – ou de détournement[6] – au pire. Au même moment, on assiste à une convergence des algorithmes de pertinence chez la plupart des moteurs majeurs.

Au regard des documents originaux publiés par Brin et Page (Brin et al.,1998a, 1998b, 1999) à l'époque du lancement du moteur, des analyses afférentes (Langville & Meyer, 2003) et du brevet déposé, le fonctionnement originel du PageRank est aujourd'hui connu. Mais il est en revanche difficile d'évaluer le poids exact de l'indicateur relationnel dans l'algorithme de pertinence. De plus, un certain nombre d'éléments constitutifs de l'algorithme actuel ne sont pas brevetés. Au final, ce succès repose sur la capacité d'innovation de Google qui s'est traduit par l'enrichisssement du Pagerank originel avec notamment la prise en compte de logiques qualitatives en amont (nature et sémantique des backlinks), de puissantes procédures de filtrages en aval (permettant d'atténuer les usages détournés), et une infrastructure technique[7] permettant de travailler à une échelle statistique et computationnelle hors-norme.

Mais là où la bibliométrie propose des indicateurs, les moteurs de recherche n'offrent qu'un seul type de « preuve » ou de mise à l'épreuve : celui de leur pertinence. Une notion ambiguë parce que non uniformément circonscrite. Si l'indicateur scientométrique vaut pour l'ensemble de la communauté qui s'y réfère ou l'utilise, chacun jugera de la « pertinence » d'un résultat à l'aune de sa seule subjectivité et du contexte de tâche dans lequel s'inscrit sa recherche. Le lien qui peut pourtant être posé entre les travaux définitoires de Garfield et l'immense terrain applicatif investi par Google permet d'interroger en creux cette notion de pertinence au regard d'un concept dont les liens étroits qu'il entretient avec la naissance de la bibliométrie sont souvent méconnus : celui de la sérendipité.

2.2. LA SÉRENDIPITÉ COMME RÉMANENCE BIBLIOMÉTRIQUE

---

[5] une limitation qui peut apparaître insuffisante ou relative au regard, notamment, de certaines pratiques d'auto-citation et du jeu des collèges invisibles. Pour une vue globale des biais de plus en plus reprochés au SCI, consulter (Piolat & Vauclair, 2004)

[6] « Spamdexing » (indexation faussée), Google bombing (Tatum, 2005)

[7] aucun chiffre officiel n'est disponible mais les observateurs estiment cette infrstructure à environ 450 000 serveurs répartis sur 30 à 60 Datacenters.

Le terme de « sérendipité » apparaît avec Walpole dans un conte oriental « Voyages et aventures des trois princes de Serendip » (Ceylan), où ceux-ci, « *ayant d'abord été formés avec soins, dans toutes les sciences, se tiraient toujours d'affaire grâce à leur talent exceptionnel pour remarquer, observer, déduire, à toute occasion.* ». Ce terme fait ensuite son apparition dans le domaine de la sociologie des sciences où il est conceptualisé par Merton comme suit : *« la découverte par chance ou sagacité de résultats que l'on ne cherchait pas ».* (Perriault, 2000) parlera d'un *« effet "serendip" (qui) consiste à trouver par hasard et avec agilité une chose que l'on ne cherche pas. On est alors conduit à pratiquer l'inférence abductive, à construire un cadre théorique qui englobe grâce à un "bricolage" approprié des informations jusqu'alors disparates.»*[8]

Et quand (Garfield 2004) revient sur les fondements de la bibliométrie, c'est précisément cet horizon de la sérendipité qu'il relève comme l'une des principales innovations de l'approche pourtant si raisonnée et – en apparence – si contraire à tout processus aléatoire ou fortuit qu'il échafauda, jusqu'à reprendre à son compte l'expression de (Smith 1964) parlant à propos du Citation Index d'une « sérendipité systématique » :

> *« (…) this feature* (Citation Index) *would lead to unexpected connections between citing and cited documents. In a traditional key word search one does not know exactly what will be retrieved but the result is not entirely surprising, that is unexpected, since the documents retrieved contain that key word. On the other hand, when you examine a list of papers that have cited a target article or book, the citing papers may or may not contain the same key words as the target. And due to the multi-disciplinary nature of the SCI's coverage, one often encounters completely surprising unexpected papers. That explains the origin of the oxymoron "Systematic Serendipity" which Smith used to describe the citation-based search. »*[9]

L'importance du mot-clé dans les usages des moteurs de recherche, la part d'aléatoire dans le choix de celui-ci, les parentés et les similarités masquées que permet de faire apparaître le rapprochement de plusieurs articles résultant apparemment de processus de recherche non-identiques, l'émergence d'une pertinence statistique née de l'appariement de mots-clés tantôt présents dans le texte cible et tantôt dans la requête de l'usager pour retrouver celui-ci, tous ces éléments dressent une filiation directe entre la réalité actuelle des moteurs de recherche et les fondements théoriques de la bibliométrie et de la recherche par citation. Un rapprochement finalement assumé dans le service Google Scholar[10].

2.3. A L'HORIZON DES USAGES : LE PROFANE ET LE PROFESSIONNEL

Si le modèle processuel des moteurs de recherche s'inscrit ici clairement dans l'héritage d'une rationalité comptable, autour du nombre de citations transposées au nombre de liens, l'usager non-expert ou non acculturé ne dispose en revanche d'aucune clé de lecture

---

[8] Sur la génèse de cette notion, voir aussi (Ertzscheid & Gallezot, 2003) et (Andel, 2005)

[9] *« (…) le* Citation Index *peut conduire à des connexions inattendues entre les documents cités et ceux les citant. Dans une recherche par mot-clé traditionnelle, on ne sait pas exactement ce que l'on trouvera mais le résultat n'est jamais totalement une surprise, ce qui est pour le moins inattendu étant donné que les documents retournés contiennent le mot-clé. D'un autre côté, quand vous examinez une liste d'articles qui citent un article cible ou un ouvrage, les papiers qui en citent un autre peuvent ou non contenir les mêmes mots-clés que leur cible. Et du faut de la couverture multi-disciplinaire du SCI, on rencontre souvent des articles totalement inattendus. Ce qui explique l'origine de l'oxymore « sérendipité systématique » utilisé par Smith pour décrire la recherche par citation. »*

[10] http://scholar.google.com permet de recherche uniquement dans des articles et des ouvrages scientifiques « moissonnés » à différentes sources (universités, éditeurs …) par le moteur de recherche. Chaque résultat est immédiatement suivi du nombre de fois où l'article ou l'ouvrage en question est littéralement « cité » (et non plus simplement « lié ») par d'autres.

hors celle d'une « pertinence » entièrement subjectivée. Ce manque d'acculturation, combiné à des pratiques de recherche empiriques (de type « essai et erreur »), ne mobilisant que très peu de mots-clés[11], n'allant que très rarement au delà de la première page-écran, et n'utilisant quasiment jamais les – maigres – possibilités de recherche par champ (date, extension, etc) offertes, font que l'usager "découvre" plus qu'il ne "recherche" de l'information.

De son côté, l'usager expert, du fait de l'opacité entretenue des algorithmes et des métriques entrant en jeu, même s'il dispose de « clés » interprétatives pour décoder les logiques sous-jacentes à l'affichage de résultats, ou pour au moins ne pas en inférer d'inexactes, ne peut à son tour qu'accepter d'intégrer une part d'aléatoire, de fortuit dans le va-et-vient entre les requêtes déposées et les résultats retournés. La sérendipité systématique apparaît ainsi comme une composante majeure de cette médiation qu'est le processus de recherche d'information.

3. MARCHANDISER L'ALÉATOIRE : LE POINT NODAL DE LA RECHERCHE

3.1. POURQUOI MARCHANDISER L'ALÉATOIRE

Le modèle économique des principaux moteurs de recherche est celui d'une régie publicitaire vendant des espaces à des annonceurs, sous forme de lien dits « commerciaux » ou « sponsorisés ». Des études récentes (Hargittai, 2002) (Fallows, 2005) ont montré que les usagers sont le plus souvent dans une totale ignorance de ces pratiques et que le traitement cognitif qu'ils accordent à ces résultats sponsorisés et aux résultats « organiques » du moteur est le même. Ainsi ignorées, ces pratiques d'indexation marchande travaillent sur une illusion de surgissement contextuel pertinent, particulièrement dans le contexte de requêtes transactionnelles[12].

Cette instrumentalisation opacifiante de la sérendipité intéresse les moteurs qui entretiennent chez l'usager cette illusion du surgissement de résultats pertinents. Sur l'analyse croisée de millions de requêtes et grâce aux fonctions dites « d'historique de navigation », le moteurs sont *de facto* capables de prédire (aspect prédictif) que l'affichage de tel lien sponsorisé pour telle ou telle requête entraînera potentiellement un clic menant vers un acte d'achat potentiel (aspect incitatif), chaque clic généré alimentant le modèle économique sous-jacent. A ces aspects prédictifs et incitatifs s'ajoute une valeur de suggestion souvent très efficace pour orienter une recherche d'information sur la base d'un besoin insuffisamment défini : ainsi l'interface de GoogleNews couplée à la fonction GoogleSuggest permet d'afficher une liste de requêtes déjà saisies et le nombre de résultats correspondants au fur et à mesure de la saisie de l'usager[13].

Enfin (cf 1.3), la plupart de nos comportement sociaux en-ligne (la musique que j'écoute, les films que je télécharge, les gens que je rencontre, etc ...), sont traqués et collectés par les moteurs qui sont eux-mêmes fournisseurs ou hébergeurs de ces services, ou bien leur fournissent la technologie de recherche[14], ou bien encore en sont les principaux annonceurs publicitaires. Si l'on y ajoute la collecte des logs et autres identifiants de connexion, la simple possibilité de recoupement de ces données laisse entrevoir une gamme infinie d'applications relevant du marketing comportemental, pratiques au regard desquelles l'entretient d'une – illusoire – sérendipité contribue grandement à atténuer chez l'usager l'aspect intrusif de ces démarches.

---

[11] Il n'y a que depuis peu de temps que les internautes saisissent en moyenne 2 mots-clés pour leurs requêtes.

[12] Exemple de l'usager à la recherche d'une destination touristique qui se voit proposé un lien sponsorisé pour acheter des billets d'avion sans qu'il l'ait expressément sollicité.

[13] http://googleblog.blogspot.com/2006/04/news-suggest-join-forces.html

[14] Exemple du site MySpace.com, un des premiers réseaux sociaux dela planète, qui utilise Google (et son modèle publicitaire d'affiliation) comme moteur interne.

3.2. L'ILLUSION DU VISIBLE ET L'AFFAIBLISSEMENT DE LA PERTINENCE

Cette instrumentalisation s'inscrit également dans un contexte : la dichotomie entre web visible et invisible, qui même si elle tend à se restreindre (cf 1.3) reste opérante, permettant là encore un effet serendip du fait de la limitation de corpus, des connexions possibles entre les deux mondes et de l'effet de surgissement qui peut se produire par le biais d'abonnements ou d'accords universitaires souvent inconnus des usagers : une même ressource, un même document, peut, selon le lieu de connexion apparaître visible ou invisible (Pinczon du Sel, 2006).

Un autre phénomène vient en parallèle appuyer l'affaiblissement d'une objectivation possible de la notion de pertinence. Il s'agit de l'effet des liens affinitaires ou « liens d'affection (Véronis, 2005) qui attestent que sur une batterie de requêtes, les résultats « favorisent », selon les moteurs, certains sites marchands plutôt que d'autres. L'un des corrélats direct de cet affaiblissement est la confusion désormais totale entre les notions d'autorité et de notoriété. Le moteur Technorati[15] proposant ainsi un critère de filtrage intitulé « authority » (a lot of, a bit of …) qui se réfère uniquement au nombre de liens entrants sur tel ou tel site (blog). Dans ce contexte, d'autres glissements sémantiques ne tardent pas à apparaître, entre affluence et influence par exemple, ou bien encore entre publicité et légitimité.

4. VERS UNE SÉRIATION DE LA SÉRENDIPITÉ.

4.1. AUTOUR DES MODÈLES DE LA RECHERCHE D'INFORMATION.

Le web ne se présentant pas comme un corpus homogène, et la notion de pertinence s'affaiblissant, il devient important de pouvoir distinguer entre différents types de sérendipité, gravitant autour des trois modèles théoriques de la recherche d'information. Il existe 3 « états initiaux », auxquels sont associés trois processus, trois types de tâches, qui font eux-mêmes référence à trois grands types de modèles.

Etat initial	Processus	Modèles
1.[Je sais] [ce que je cherche]	Querying / Browsing Sérendipité nulle	Computationnel Web Content Mining
2. [Je ne sais pas] [ce que je cherche]	Searching Sérendipité par inférence abductive	Utilisateur Web Usage Mining
3. [Je sais] [que je ne sais pas ce que je cherche]	Learning Sérendipité associative	Environnementaliste Web Structure Mining

Tab. 1 : Les modèles de la recherche d'information

Le premier cas représenté repose sur l'idée que dans la majorité des démarches de recherche d'information, l'utilisateur sait déjà (partiellement) ce qu'il cherche. Il lui reste alors à mettre en place une série de requêtes (querying) correspondant au modèle computationnnel classique autorisé par les systèmes documentaires (booléens, langages documentaires, etc.). L'utilisateur est dans une logique de consultation et cherche à savoir ce que peut lui apporter comme résultats (*matching*) le système d'information qu'il est en train d'utiliser (*browsing*). Cet utilisateur met en place un raisonnement de type hypothético-déductif. Si la sérendipité est ici quasi-nulle c'est dans la mesure où elle ne relève d'aucune démarche volontariste ou

[15] www.technorati.com : Technorati est un moteur de recherche indexant uniquement les billets publiés sur les blogs.

consciente de l'usager. En revanche l'instrumentalisation commerciale de la sérendipité joue à plein (via notamment le surgissement de liens sponsorisés).

Le second cas correspond à l'objectif de la recherche d'information selon (Belkin, 2000) : « *Helping people find what they don't know*. » Le processus alors appelé est de type exploratoire (*searching*). L'utilisateur va, à partir de ce qu'il sait, raisonner par inférence et abduction en fonction de son but ou de son « profil », conformément à la définition de (Perriault, 2000) cité plus haut.

Le dernier cas est celui qui peut le plus bénéficier du phénomène de sérendipité. L'utilisateur ayant formalisé et explicité qu'il « ne sait pas ce qu'il cherche » se met alors consciemment en situation d'adopter le comportement le plus simple, le plus intuitif et associatif possible, et ce quel que soit la complexité des systèmes qu'il consultera et le type et l'étendue des corpus qu'ils indexent.

Reprenant les travaux réunis par Patricio Galéas[16], le cas n°1 repose essentiellement sur une analyse des contenus, lesquels se prêtent alors à l'affichage de publicité « contextuelle ». La sérendipité du cas n°2 s'appuie davantage sur l'exploitation d'un profil pour déterminer des formes ("patterns") de navigation. Enfin le cas n°3 repose sur l'exploitation poussée de l'environnement hypertexte pour une exploration associative maximale de la structure hypertextuelle du web.

4.2. SÉRIATION PAR L'EXEMPLE

Dans la mouvance du web 2.0[17] (cf infra), l'émergence d'outils de recherche de nouvelle génération[18] se positionnant sur des usages, des communautés ou des marchés « de niche », permet d'instancier différemment la sériation proposée ci-dessus en prenant en compte de nouveaux contextes d'usage et de nouvelles pratiques.

SÉRENDIPITÉ SOCIALE

A ce titre l'émergence de réseaux sociaux entraîne un effet de sérendipité « sociale », permettant à des sites comme Fo.rtuito.us (http://fo.rtuito.us/) de proposer un modèle dans lequel on vous présente au hasard un nouveau membre après vous être enregistré. Vous disposez alors de 4 jours pour interagir avec l'autre membre via un email anonyme et voir s'il peut devenir un ami. Au-delà de cet exemple, c'est l'ensemble des sites de réseaux sociaux, qu'ils concernent la sphère amicale ou professionnelle, qui proposent – et imposent parfois – de manière aléatoire la découverte de nouveaux membres.

SÉRENDIPITÉ EXPÉRIENTIELLE

Autre phénomène émergent, les blogs permettent de fixer certains espaces de parole dans lesquels s'expriment différents niveaux d'expertise. La densité des liens que donne à voir la blogosphère et les protocoles spécifiques comme les rétroliens permettent un nouvel effet de surgissement : l'auteur d'un billet peut ainsi voir s'afficher en dessous de celui-ci, un lien vers un autre billet, ce lien ayant été déposé par l'auteur du billet lié. Dans le cadre de certains blogs d'experts, le cumul de ces trackbacks devient un élément prépodérant dans la découverte de nouvelles sources, de nouvelles informations. La sérendipité se fait ici « expérientielle » dans la mesure où elle s'appuie sur l'identification et le renforcement d'une expertise.

SÉRENDIPITÉ RELATIONNELLE

---

[16] http://www.galeas.de/webmining.html

[17] l'expression web 2.0 désigne les sites dont le contenu est produit, généré et indexé (« taggué ») majoritairement par les usagers. YouTube (partage de vidéos), Del.icio.us (signets partagés) et MySpace (réseau social) en sont quelques exemples emblématiques.

[18] http://oedb.org/library/features/top-25-web20-search-engines

Les moteurs de recherche classiques permettent par le biais de syntaxes de recherche avancées, d'identifier des sites à contenu similaires ou proches[19]. Ce mode de requêtage « expert » est désormais inscrit au cœur de nombre d'interfaces de visualisation de l'information dont le TouchgraphGoogle Browser[20] qui révèle le réseau des connectivités entre sites web tels qu'ils sont contenus dans la base de données de Google. Par ailleurs, l'essor de moteurs ou métamoteurs représentant leurs résultats à l'aide d'une métaphore visuelle de type cartographique (Kartoo, Mapstan) relève du même processus. Citons enfin l'engouement des sites à vocation marchande (dont le précursuer fut Amazon.com) pour les systèmes dits « de recommandation »[21] surutilisés parce que plébiscités par les usagers. Sérendipité visuelle, affinitaire et relationnelle relèvent de la même logique.

LA SÉRENDIPITÉ ORNEMENTALE : « FEELING LUCKY »

Google constitue là encore un cas d'école puisqu'il inscrit au cœur de son interface dépouillée un bouton « feeling lucky » permettant en fait d'aller directement au premier des résultats organiques récupéré par le moteur pour une requête donnée. La persistance au cœur de l'interface du premier moteur de la planète d'un tel bouton dont on sait que la présence n'est qu'ornementale[22] relève d'une analyse stratégique des vertus de la sérendipité, au nom desquelles la découverte (apparente) par accident, par « hasard », procure au chercheur une jouissance plus immédiate, parce qu'inattendue.

4.3. LE RETOUR DE L'USAGER

Au regard des sériations et inventaires précédents, la sérendipité apparaît comme instrumentalisée majoritairement au profit des moteurs de recherche, l'usager subissant de manière passive la plupart de ces effets, ce qui ne l'empêche pas d'y trouver parfois son compte. Or la reconfiguration actuelle de la topologie du web et des espaces d'énonciation qui s'y affirment permet d'anticiper une inversion de cette tendance. Les outils de recherche sont à leur tour instrumentalisés *par* les usagers, souvent regroupés en communautés, vers la génération de processus sérendipiteux comme outils de découverte de connaissance, marquant un retour au plus près de la notion originelle de sérendipité (Boutin et al., 2006). Quatre phénomènes permettent d'expliquer ce retournement.

Premièrement celui d'une communautarisation. S'adaptant à l'engouement pour les sites de signets partagés (l'usager construit, agrège ses sources et les fait partager à une communauté), les moteurs leaders se mettent à racheter ou à proposer leurs propres communautés. Le fait de souscrire et d'interagir à l'intérieur de ces espaces où le filtre éditorial atteste d'abord des choix de la communauté va orienter l'usager vers de nouvelles sources d'information ou de nouveaux documents. Deuxièmement, celui d'une fragmentation et d'une granularité nouvelle dans l'accès à l'information. Les techniques liées au protocole RSS viennent offrir au périmètre de la recherche d'information une granularité quantitative et qualitative jusqu'ici impossible à atteindre (Ertzscheid 2005). Les moteurs de recherche reprenant là encore à leur compte ces nouvelles logiques d'exploitation en proposant certains de leurs résultats directement au format RSS (notamment les "news"). Troisièmement, celui du mixage. C'est l'ère des Mashups. Celui du couplage, du mixage qui peut opérer entre services, ou entre un moteur et un/des services. Quatrièmement, celui, primordial de la personnalisation permettant à chacun de fabriquer "son" propre moteur (Google Co-op), de

---

[19] opérateur « like : » ou « related : »

[20] http://www.touchgraph.com/TGAmazonBrowser.html

[21] « si vous avez aimé X alors vous aimerez Y » ou encore « ceux qui ont acheté X ont aussi acheté Y ».

[22] 'on sait que ledit bouton n'est quasiment jamais cliqué (mais « *les utilisateurs considèrent que son absence nuirait à l'ambiance générale de la page d'accueil. C'est un bouton de confort* ». Marissa Mayer, Product Manager)

choisir "ses" sources, de mettre en place ses "macros" (LiveSearch), et d'ouvrir ensuite cet espace à d'autres (communautarisation).

4.4. QUATRE GÉNÉRATIONS DE MOTEURS

Les moteurs de recherche qui ont suivi ou anticipé ces évolutions technologiques et les usages associés ont pris date en même temps qu'ils en prenaient acte. Aux moteurs de première génération, reposant sur l'analyse du contenu des pages indexées et ne prenant en compte que le "matching" (croisement) entre des mots-clés (ceux de la requête et ceux des pages indexées), succédèrent des moteurs de deuxième génération pour lesquels ce fût d'abord la structure du web en tant que graphe orienté qui vînt trouver sa place au cœur de l'algorithmie (PageRank), permettant une recherche d'information « augmentée ». La troisième génération est celle des moteurs "sociaux" (de type Rollyo, Eurekster …) qui ajoutent le filtrage amont des sources par les usagers comme un modèle de pertinence différent.

Nonobstant cette évolution, et comme nous l'avons montré plus haut, les moteurs de recherche dominants continuent d'offrir une vision très appauvrie de la complexité du web, ne retenant pour l'affichage que la forme de la liste (linéarité des résultats, faible nombre de résultats exploitables et instabilité), et se privant de toute catégorisation ou ordonnancement « raisonné » des résultats affichés[23]. Un parallèle avec l'histoire des techniques documentaires fait ressortir l'opportunité que pourraient représenter la mise en œuvre d'interfaces reprenant les avantages de l'approche proposée par la théorie des facettes, par le biais de moteurs à curseurs, initiés par Yahoo !Mindset[24].

L'idée des moteurs à curseur est que l'expression du besoin peut être affinée par l'internaute à travers l'expression de dimensions complémentaires et orthogonales au sujet de la recherche. Ainsi une page web peut elle être décrite par son thème mais aussi son niveau plus ou moins commercial (Mindset de Yahoo), son statut de page de contenu ou de pages de liens (www.clush.com). Bien souvent les moteurs encore expérimentaux qui proposent ces curseurs développent une seule dimension. Penser théorie des facettes et moteur à curseur, permettrait de proposer un outil présentant une interface de recherche dans laquelle plusieurs dimensions orthogonales au sujet de recherche seraient envisagées.

5. CONCLUSION & PERSPECTIVES

Nous avons finalement détaillé dans cet article trois niveaux génériques de serendipité. Une sérendipité systématique, intrinsèque au fonctionnement des moteurs, qui fait apparaître des proximités inattendues, qu'elles soient « naturelles » ou « marchandes ». Une sérendipité par génération stochastique qui joue sur les possibilités d'afficher les résultats autrement que sous forme de liste (cartes, superposition de cartes, clusters ...) à la recherche d'éléments saillants. Et une sérendipité comme processus opératoire incarné, permettant à la subjectivité de chaque individu ou de chaque communauté, de gouverner à la fois l'horizon et les limites de la recherche. D'autres niveaux sont encore à distinguer, en relation par exemple avec le domaine de la psychologie cognitive à l'aide de concepts relevant du traitement de la vision périphérique et de la fluidité perceptive des traitements.

Nous avons également explicité la relativité de la notion de pertinence, ainsi que l'opacité entretenue de cette notion à l'origine de bien des confusions ou des approximations

---

[23] Les « annuaires » de recherche, qui reprenaient ce principe de catégorisation ont aujourd'hui pour la plupart disparus (à l'exception notable de l'annuaire Dmoz, lequel n'est cependant plus du tout mis à jour pour certaines de ses rubriques, mais bénéficie d'un sursis du fait de sa présence au sein du moteur Google)

[24] http://mindset.research.yahoo.com/ : permet, via la manipulation d'un curseur, de faire varier le nombre de résultats marchands et non-marchands.

chez les usagers. C'est le syndrôme du moteur de recherche comme « boîte noire ». Dans ce contexte, la sérendipité s'affirme comme un concept didactique de premier plan pour l'apprentissage de la recherche d'information, processus dans lequel un ensemble d'actants (extension cognitive) permettent de trouver de l'information. Si les outils de recherche en font partie, ils n'en sont qu'un élément, et c'est bien tout l'environnement socio-technique du chercheur qu'il convient de prendre en compte. *Audaces Fortunat Juvat*. La chance sourit aux audacieux, et aux esprits préparés. Dans le cas de la recherche d'information nous pourrions ajouter et « bien environnés ».

L'horizon des moteurs 2.0 et particulièrement celui des moteurs à curseurs permet en outre de réintroduire des logiques d'usages différenciées permettant de casser celles, monopolistiques, des principaux acteurs de la recherche d'information. Sur ces questions, il paraît aujourd'hui urgent d'interroger l'écosystème non plus simplement documentaire mais politique de la recherche et de l'accès à l'information et de l'illusion de complétude dans lequel elle confine l'usager. Si le secret des algorithmes est assurément une manière de garantir un indicateur de pertinence exogène, la mise au jour de certaines arcanes ou la mise en place d'observatoires indépendants doit pourvoir être débattue. Faute de quoi, devant des usages de plus en plus conditionnés à des logiques marchandes instrumentalisant les données collectées au cours même du processus de recherche, c'est l'usager et lui seul qui dans l'ignorance de ces logiques, se trouvera totalement instrumentalisé, devenant, dans un mouvement réflexif paradoxal, le seul et unique objet de sa recherche.

RÉFÉRENCES
(Tous les liens actifs au 02 mars 2007)

Amy N. Langville and Carl D. Meyer, « Deeper Inside PageRank », Internet Mathematics. Vol. 1, no. 3, pp. 335–380, 2003.

Andel P. Van, « Sérendipité ou l'art de faire des trouvailles », in Automates, Intelligents, Février 2005, en ligne : http://www.automatesintelligents.com/echanges/2005/fev/serendipite.html.

Basquiat J.P., Automates Intelligents, n° 59, 31 mai 2005.

Belkin N., « Helping People Find What They Don't Know. », in Communications of the ACM, Vol. 43, n° 8, Août 2000.

Boutin E., Gallezot G., Quoniam L., « Détecter l'innovant sur le web par des techniques non booléennes : méthode, outils, application », Colloque Canadien des sciences de l'information 2006. En ligne : http://www.cais-acsi.ca/proceedings/2006/boutin_2006.pdf

Brin S., Motwani R., Page L., Winograd T., « What can you do with a Web in your pocket? » Data Engineering Bulletin, 21:37–47, 1998a.

Brin S., Page L., « The anatomy of a large-scale hypertextual Web search Engine ». Computer Networks and ISDN Systems, 33:107–117, 1998b.

Brin S., Page L.,, Motwami R., Winograd T., « The PageRank citation ranking: bringing order to the Web. » Technical Report 1999-0120, Computer Science Department, Stanford University, 1999.

Broder, A. 2002. « A taxonomy of web search. » SIGIR Forum 36, 2 (Sep. 2002), 3-10. En Ligne : http://doi.acm.org/10.1145/792550.792552.

Ertzscheid O., « Syndrome d'Elpenor et sérendipité : deux nouveaux paramètres pour l'analyse de la navigation hypermédia. » in Actes du colloque H2PTM'03. Editions Hermès, septembre 2003.

Ertzscheid O., « Weblogs : un nouveau paradigme pour les systèmes d'information et la diffusion de connaissances ? Applications et cas d'usage en contexte de veille et d'intelligence économique. », Communication avec Actes, Colloque ISKO 2005, Nancy, en ligne : http://archivesic.ccsd.cnrs.fr/sic_00001433

Ertzscheid O., Gallezot G., « Chercher faux et trouver juste. » Communication avec actes. X° Colloque bilatéral franco-roumain, CIFSIC Université de Bucarest, 28 juin – 3 juillet 2003. Université de Bucarest (Éd.) En ligne : http://archivesic.ccsd.cnrs.fr/sic_00000689.

Fallows D., « Search Engine Users: Internet searchers are confident, satisfied and trusting – but they are also unaware and naïve. » PEW Internet Report, Janvier 2005, en ligne : http://www.pewinternet.org/PPF/r/146/report_display.asp.

Garfield E, « Citation analysis as a tool in journal evaluation », Science, (178):471-479, 1972.

Garfield E. « Systematic Serendipity: Finding the Undiscovered Answers to Science Questions » Communication au Medical Ignorance Collaboratory, Université d'Arizona Health Sciences Center, Tucson, Juillet 2004. En ligne http://garfield.library.upenn.edu/papers/az072004.pdf.

Goody J., La raison graphique. Paris, Minuit, 1979.

Hargittai E., « Second-Level Digital Divide: Differences in People's Online Skills », in First Monday, vol 7, n°4, Avril 2002, en ligne : http://firstmonday.org/issues/issue7_4/hargittai/index.html.

Lyman P., Varian Hal R., « How Much Information », 2003. En ligne : http://www.sims.berkeley.edu/how-much-info-2003.

Nielsen J., « One billion Internet Users. », Décembre 2005. En ligne : http://www.useit.com/alertbox/internet_growth.html.

Rose D.E., Levinson D., « Understanding user goals in Web search ». In Proceedings of the Thirteenth Int'l ». World Wide Web Conference, 2004.

Pinczon du Sel P., « Etat des lieux des résultats d'une recherche d'information simultanée sur le moteur de recherche Google. », Actes du colloque ACSI/CAIS,. Université York, Toronto, Canada, Juin 2006. En ligne : http://archivesic.ccsd.cnrs.fr/sic_00001755.

Piolat A., Vauclair J., « Le processus d'expertise éditoriale avant et avec Internet », in Pratiques psychologiques, Vol 10 n°3, pp. 255-272, 2004.

Salaün, J-M., Charlet J., « Introduction : Comprendre et maîtriser la redocumentarisation du monde. » In La redocumentarisation du monde, sous la dir. de Roger T. Pédauque, 15-25. Toulouse : Cépadues Édition, 2007. En ligne : http://hdl.handle.net/1866/725

Smith J.-F., « Systematic Serendipity » Chemical & Engineering News 42(35):55-56, 1964.

Tatum C., « Deconstructing Google bombs: A breach of symbolic power or just a goofy prank? », First Monday, vol 10, n°10, October 2005, en ligne : http://firstmonday.org/issues/issue10_10/tatum/index.html.

Véronis J., « Moteurs : liens d'affection », en ligne, http://aixtal.blogspot.com/2005/12/moteurs-liens-daffection.html, décembre 2005.