

Chercher faux et trouver juste, Serendipité et recherche d'information

Olivier Ertzscheid

Urfist - Université des Sciences Sociales
(Toulouse 1)
Laboratoire Paragraphe - Université Paris 8 -
Groupe "Ecritures hypertextuelles."
11 rue des Puits-Creusés - 31070 Toulouse
Tél : 05.34.45.61.80 / Fax : 05.34.45.61.85
o.ertzscheid@voila.fr

Gabriel Gallezot

Urfist - Université de Nice-sophia Antipolis
LAMIC
Ave Joseph Vallot
06108 Nice cedex 2
Tél : 04 92 07 67 26 / Fax : 04 92 07 67 00
gallezot@unice.fr

Code de champ modifié

Code de champ modifié

Résumé

Merton définit la serendipité ainsi : "*découverte par chance ou sagacité de résultats que l'on ne cherchait pas*". L'idée mise en avant dans ce texte est de montrer comment dans un effort de maîtrise de l'entropie informationnelle, (2) l'essor des technologies intellectuelles de représentation et d'accès aux informations (3) fait chaque jour une place plus grande au phénomène de sérendipité comme adjuvant précieux de la recherche (4).

Abstract

Merton defines the serendipity as follows : "*the faculty or phenomenon of finding valuable or agreeable things not sought for*". We propose to show how, in an effort to control informational entropy,(2) the rise of intellectual technologies for representation and access to information (3) makes each day a larger place to phenomenon of serendipidity as a precious additive of research (4).

Mots clés : sérendipité, recherche d'information, technologies intellectuelles, entropie informationnelle

Keywords : serendipity, information retrieval, informational entropy, intellectual technologies

1. Introduction

1.1 - Définition et origine de la sérendipité

Le terme de serendipity 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.* »¹. Ce terme apparaît en sciences et se conceptualise avec Merton qui le définit ainsi : « *la découverte par chance ou sagacité de résultats que l'on ne cherchait pas* ». La sérendipité (« fortuité » pour nos amis québécois) est une problématique qui n'a fait que récemment son entrée dans le champ des sciences de l'information – francophones – sous la plume de Perriault : « *L'effet "serendip" (...) 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.* » [Perriault, 00].

Pour cerner ce concept et appréhender le phénomène nous indiquons ci-dessous quelques exemples célèbres. Tout le monde a appris comment Christophe Colomb, cherchant la route occidentale des Indes découvrit en fait l'Amérique. Nombre d'autres découvertes tout aussi essentielles pour l'humanité ont partie liées avec la sérendipité. En voici une liste non-exhaustive : le principe de champagnisation (Dom Pérignon), la pasteurisation (L. Pasteur), la pénicilline (A. Fleming), les rayons X (W. Röntgen), la vulcanisation du caoutchouc (Ch. Goodyear). Citons également d'autres découvertes moins "essentielles" comme le "post-it" (où comment répondre à la question : "que faire d'une colle qui ne colle pas ?"), le Caprice des dieux, le Coca-Cola, le Zyban et le Viagra...

1.2 – Rechercher ou Recherche

Compte tenu du contexte francophone dans lequel nous présentons ce texte, nous voulons d'abord introduire la distinction et le parallèle entre le « rechercher » de la recherche d'information (Information Retrieval en anglais, IR) et le « rechercher » de l'épistémé, la recherche (*Research* en anglais). Ce signifiant unique en français introduit un signifié commun « trouver l'information » mais aussi des signifiés distincts qui se trouvent quelque part dans les moyens et l'objectif. Pour l'IR, les moyens ce sont des outils du traitement de l'information sur un corpus documentaire. Pour l'épistémé ce peut être l'empirisme ou des outils de traitement de l'information dont l'objet change et devient « naturel ». L'objectif, pour l'IR, c'est de repérer et de ramener des infos pertinentes. Pour l'épistémé c'est de découvrir, de produire de nouvelles connaissances. Enfin, pour être complet signalons aussi dans notre domaine disciplinaire la recherche en IR qui désigne notamment la modélisation et les études d'usages informatiques dont l'objet est constitué par des corpus de texte, de connaissance.

1.3 - Complexité

Pour appréhender le champ de l'IR et ses objectifs nous proposons de fouiller la métaphore de « l'aiguille dans la botte de foin » qui, de façon triviale, signifie la difficulté (l'impossibilité) de trouver quelque chose. Chercher une aiguille dans une botte de foin peut s'appréhender de différentes manières et ainsi correspondre à plusieurs scénarios de recherche :

- *trouver une aiguille connue dans une botte de foin connue*
- *trouver une aiguille connue dans une botte de foin inconnue*
- *trouver une aiguille inconnue dans une botte de foin inconnue*
- *trouver n'importe quelle aiguille dans une botte de foin*
- *trouver l'aiguille la plus pointue dans une botte de foin*
- *trouver la plupart des aiguilles contenues dans une botte de foin*
- *trouver toutes les aiguilles contenues dans une botte de foin*
- *pouvoir affirmer qu'il n'y a pas d'aiguille dans une botte de foin*
- *trouver des choses qui ressemblent à des aiguilles dans une botte de foin*
- *connaître chaque nouvelle aiguille qui apparaît dans la botte de foin*
- *trouver où sont les bottes de foin*
- *trouver des aiguilles et des bottes de foin, quelles qu'elles soient.*" [Koll, 00]

¹ Source : www.granddictionnaire.com

Une vision plus globale est proposée par [Toms 00], pour qui il existe trois grandes manières de chercher de l'information :

- « chercher de l'information sur un objet bien défini ;
- chercher de l'information sur un objet incomplètement décrit mais qui sera reconnaissable dès qu'un le rencontrera ;
- trouver de l'information de manière fortuite. »

Ce troisième et dernier cas, fait écho à l'une des déclinaisons non citées de l'aiguille et de la botte de foin : « Mal chercher l'aiguille dans la botte de foin et la trouver quand même ».

L'idée mise en avant dans notre article est de montrer comment dans un effort de maîtrise de l'entropie informationnelle (2) l'essor des technologies intellectuelles de représentation et d'accès aux informations (3) fait chaque jour une place plus grande au phénomène de sérendipité comme adjuvant précieux de la recherche (4).

2. Maîtriser l'entropie informationnelle ?

Appréhender la complexité d'un phénomène nécessite le repérage et la gestion d'un corpus de documents pertinent et volumineux pour extraire des informations et ensuite les transformer en nouvelles connaissances. Ce cycle de la production scientifique se juxtapose avec le cycle du document. Ainsi, le « bouclage » d'un cycle produit de nouvelles connaissances et conséquemment de nouveaux documents. Les idées ne peuvent se former que sur des constructions cognitives antérieures présentes sous forme d'information dans les documents. Nous inscrivons donc la construction de connaissance dans un processus de transformation de l'information où :

- la connaissance est la formation des idées,
- l'information est la mise en forme des connaissances (in-formation) et
- l'information inscrite sur un support constitue un document.

L'itération de ce processus conduit à une somme d'informations toujours plus importante que les chercheurs de toutes disciplines cherchent à maîtriser. Internet accroît la rapidité de ce cycle et contribue plus encore à l'entropie informationnelle. [Gallezot, 02a]

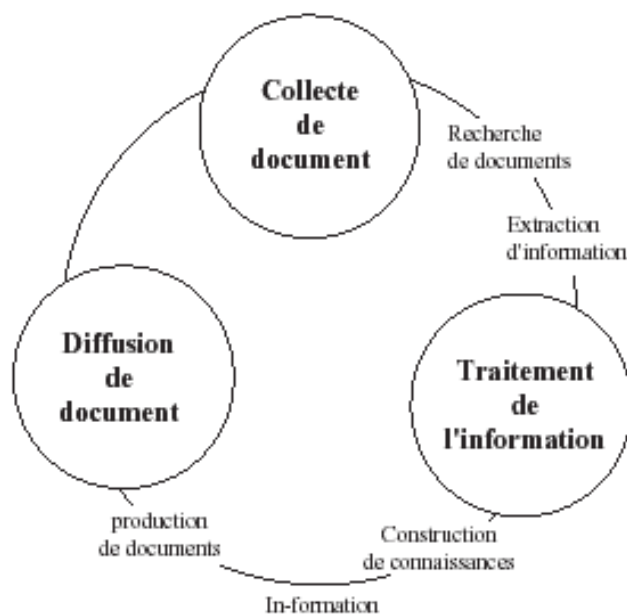


Fig. 1 : le cycle de l'IST

Entropie et homogénéité.

Cette course à la maîtrise de l'information continue à alimenter la somme potentiellement disponible de documents, Internet contribuant dans une large mesure à ce phénomène. Comme « support » il est celui qui assume et assure la plus grande partie de cette croissance exponentielle de l'information disponible sur les réseaux. Comme mode de diffusion et d'accès, Internet garantit à cette masse d'information une « homogénéité » jusque-là jamais atteinte.

Nombre de facteurs objectifs tendent à étayer cette thèse *a priori* surprenante de l'homogénéité de l'information sur les réseaux quand l'habitude veut que l'on considère plutôt comme essentiel le caractère hétérogène de cette information. Reprécisons donc ce que nous entendons ici : la nature de l'information disponible sur les réseaux est effectivement profondément hétérogène (validité scientifique, « fraîcheur » éditoriale, qualité graphique, etc.) Pourtant, cette hétérogénéité s'efface complètement du fait du niveau de relation entre unités d'information, qui, selon le niveau d'échelle et de granularité auquel on se place (site web, système d'information en ligne, web dans son ensemble), permet d'affirmer que tout est lié à tout².

Nombre de facteurs objectifs viennent appuyer cette thèse parmi lesquels les études qui ont tenté de mesurer le « diamètre » du web : la dernière en date fait état d'un diamètre de dix-neuf liens [Barabasi, 99]. Cela signifie, que quelles que soient les unités d'information choisies (en l'occurrence des pages web), elles se trouvent connectées par une chaîne d'au plus dix-neuf liens.

Au delà de chiffres qui, du fait de la nature même du web ne sauraient être stabilisés, ces études ont surtout permis de construire une topologie de l'espace informationnel tel qu'il se déploie sur les réseaux, en faisant émerger certaines zones « obscures » (web invisible), déconnectées d'autres zones mais tout aussi connectées entre elles, et en ce sens homogènes.

La recherche d'outils capables de maîtriser l'information dans cet espace réticulaire constitue pour beaucoup un enjeu majeur. S'il est évident qu'il faille tendre vers une appropriation informationnelle exhaustive pour édifier l'épistémé, la tâche est incommensurable. Que reste-t-il au chercheur devant cette entropie informationnelle ? Se servir des outils *ad hoc* pour repérer au mieux l'information pertinente, borner son référentiel documentaire, expérimenter, observer, évaluer et produire ses résultats à l'aide de méthodologies éprouvées, de protocoles heuristiques... passer des achoppements aux paradigmes scientifiques. Il existe un raccourci : la serendipité. Elle s'offre et se révèle lors de découvertes informationnelles.

3. Découvertes informationnelles et technologies intellectuelles.

Les pratiques informatives sont multiples et constituées des usages différents de l'ensemble : « repérage/collecte/traitement/diffusion » de l'information. Les actions de cet ensemble sont réalisées par un binôme indissociable d'outils et de méthodes : les technologies intellectuelles [Fayet-Scribe, 00]. Celles-ci permettent des découvertes informationnelles qui se transformeront, après un processus créatif, en connaissance. Ainsi nous distinguons la recherche d'information de l'épistémé mais soulignons leur appartenance au même processus.

Le processus de création conventionnel fonctionne sur le principe de divergence/convergence où la reconnaissance d'un problème est introduite par une divergence pour converger vers une nouvelle solution. Le processus de création par serendipité est le contraire : bien que la solution à un problème soit attendue il y a divergence de parcours, lesquels conduisent à un problème différent ou plus fréquemment, à la solution d'un problème dont nous n'avons aucune connaissance [Figueirado, 01] (cf. infra, point 4.2.).

Comment dès lors garder une empreinte sur ce phénomène ? Si l'on considère que les découvertes n'arrivent jamais par chance, il faut donc insister sur le rôle de la préparation intellectuelle et/ou l'intensité de l'observation et de la recherche [Van Andel, 92]. On peut aussi penser que l'IR prenne en compte les phénomènes de serendipité en complément des requêtes (*querying*) et de la navigation (*browsing*) pour stimuler la curiosité et encourager l'exploration [Toms, 01]. Les outils qui permettent, le *cross-matching*, la *clusterisation*, la *percolation* sont des facilitateurs de serendipité. Ils proposent souvent des visuels qui « donnent à voir » aux chercheurs. Nous retenons ainsi le vocable de

² http://www.almaden.ibm.com/cs/k53/www9_final. Cette étude conjointe d'Altavista, Compaq et IBM fait état d'une topologie du web en forme de nœud papillon : le nœud est constitué de pages hyperconnectées, la partie gauche comprend les pages qui permettent d'y accéder et la partie droite celles vers lesquelles pointe ce nœud. Même s'il demeure, au vu de cette étude un certain nombre de pages déconnectées, cela ne fait que renforcer l'hypothèse d'une connection optimale pour la partie sinon la plus dense, du moins la plus visible du web.

technologies *procognitives* employé par Licklider pour signifier l'importance des outils de traitement de l'information qui servent la connaissance et passons ci-dessous en revue les techniques intellectuelles qui nous semblent favoriser la création d'éléments stochastiques.

3.1 Naissance de la bibliologie.

Paul Otlet (1868-1944) peut être considéré l'un des pères de la documentation et de la science de l'information. Il est le co-fondateur avec Henri La Fontaine, en 1895, de « l'office international de bibliographie ». Son Traité de documentation (1934) est la première approche systématique de ce que nous appelons aujourd'hui la (les) science(s) de l'information. C'est le premier à comprendre le problème que posera à terme l'augmentation considérable du nombre de livres et de documents, et à proposer, pour y répondre, la création de la bibliologie, à la fois comme science et comme technique générale pour la documentation. La plupart des idées qui fondent ce que l'on appellera l'hypertexte/hypermédia sont déjà explicitement présentes chez Otlet, qu'il s'agisse d'offrir un accès automatisé aux documents ou de relier chacun d'eux avec d'autres, tout en conservant leur individualité, dans le cadre d'un « cerveau collectif ».

3.2 Indexation associative : le MeMex.

[Bush 45] est unanimement reconnu comme le pionnier de l'hypertexte/hypermédia sous sa forme actuelle. Il est – tout comme Otlet – confronté à l'explosion de la masse documentaire. Il imagine alors un système automatisé de microfiches, baptisé MeMex (Memory Extender) lequel ne sera jamais effectivement réalisé, mais contient déjà la plupart des idées de l'hypertexte. Celle-ci sont exposées dans l'article « **As we may think** » qui commence par ces mots : « *Consider a future device for individual use which is a sort of mechanized private file and library.* ».

L'idée de base est de reproduire le fonctionnement caractéristique de l'esprit humain en imaginant des machines capables de fonctionner par association et non plus selon le modèle classique de l'indexation. « *Human mind (...) operates by association. (...) Selection by association, rather than by indexing, may yet be mechanized.* »

Son système est défini comme suit : « *A memex is a device in which an individual stores all his books, records and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.* » La révolution de l'approche de Bush peut se résumer à deux idées principales :

- il est possible de « mécaniser » le fonctionnement associatif de l'esprit humain
- les parcours de navigation (« *trails* ») et d'accès dans un tel environnement associatif sont des éléments de construction du sens.

3.3 De Nelson à Engelbart : l'avènement de l'hypertexte comme système informatique et comme mode d'accès et d'organisation des connaissances.

C'est Ted Nelson, philosophe de formation, qui le premier forge le terme « hypertexte » dans un article éponyme donné lors de la conférence de la Fédération Mondiale de Documentation. Personnage contesté, il n'en demeure pas moins l'un des visionnaires les plus actifs et il est à l'origine de nombre de concepts aujourd'hui au cœur de problématiques importantes (« transpublishing » pour les questions de droit d'auteur, « versioning » pour celles des archives ouvertes et des nouveaux modes de publication, etc.). Tous ces concepts prennent place dans le cadre de son projet XANADU (<http://www.xanadu.net>).

Douglas Engelbart, chercheur au mythique SRI (Stanford Research Institute), est non seulement l'inventeur du système actuel de fenêtrage et de la souris, mais également le concepteur d'un système baptisé « **Augment** » destiné à faciliter l'augmentation des capacités de l'intelligence humaine. « *By "augmenting human intellect" we mean increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems.* » [Engelbart 62 p.1].

Augment peut être considéré comme le premier système hypertextuel effectivement réalisé :

« *As part of the Augment Project, primarily designed for office automation, Engelbart (...) developed a system called NLS which had hypertext-like features. This system was used to store all research papers, memos and reports in a shared workspace that could be cross-referenced with each other. In 1968, he demonstrated NLS as a collaborative system among people spread geographically.* »

[Balasubramanian 94]

3.4 Hypermédias générés

L'abondance de l'information nécessite un repérage accru et efficace des connaissances, qui explique l'intérêt pour les techniques de visualisation [Shneiderman, 97]. L'exploration informationnelle s'inscrit dans cette démarche. Plusieurs techniques peuvent bénéficier de cette approche, de l'analyse statistique sur du texte à la structuration et l'organisation de données dans des bases de connaissances (*knowledge bases*). Ce qu'il faut noter, c'est la généralisation de ce processus. L'extraction d'information sur un seul type de données en vue d'obtenir un résultat précis, ne suffit pas à rendre compte de situations complexes. La globalisation d'informations sur un sujet et la visualisation sous forme graphique des résultats d'un traitement réalisé par des techniques d'information offrent des *machines de vision* [Virilio, 88] capables de générer de nouvelles connaissances, de nouveaux projets de recherche ou d'autres éléments de réflexion... de créer. Ce sont des artefacts informationnels qui offrent une vision heuristique des résultats de la recherche. Les formes nouvelles de documents, que l'on peut qualifier de tertiaires, deviennent des adjuvants prépondérants et essentiels pour lire l'ensemble des documents disponibles. Ces construits sont la synthèse de résultats expérimentaux (un ensemble de données factuelles) et de conceptions théoriques (à travers la modélisation de la base de données et la génération de liens). En revanche, les interprétations, les créations sont liées à des perceptions, des appropriations personnelles des représentations. Cette appréhension repose sur la culture technique et informationnelle de chaque individu et sa capacité à l'intégrer dans son activité quotidienne, pour produire du sens sur les objets qu'il manipule [Gallezot, 02b].

Il ne s'agit pas spécifiquement de naviguer ou déambuler, mais de fouiller les sédiments cognitifs accumulés depuis quelques années à la recherche d'information. Les visualisations proposées mettent en lumière des liens qui n'auraient pas pu être perçus autrement et peuvent faire sens auprès d'un expert. Les hypermédias générés qui aident la lecture d'information, ne relèvent pas exactement d'un choix, d'une sélection d'information, mais d'une composition aléatoire dirigée par des solidarités annotationnelles [Bachimont, 99]. Le renouvellement de ces hypermédias est lié à l'ajout de documents dans les entrepôts d'informations.

Un agrégat d'informations intégré dans un artefact informationnel, n'étant pas connu *a priori*, la recherche d'information *a posteriori*, peut retourner des documents auxquels le chercheur ne pensait pas. Plus encore, la mise en relation des unités informationnelles peut permettre de découvrir des corrélations insoupçonnées soit par lecture directe d'un résultat, soit par inspection d'un visuel. Le chercheur détecte des faits de façon quasi fortuite. Cette réécriture aléatoire et cette relecture fortuite relève de la sérendipité.

3.5 Vers de nouvelles logiques implicites d'accès et de représentation des connaissances : le cas des moteurs de recherche.

Les moteurs de recherche, dans l'utilisation qu'ils font des liens comme principes de classification, ne sont pas de simples interfaces de recherche, au même titre que celles que l'on trouve sur des cédéroms : ces dernières ne prennent exclusivement en compte que les mots (clés ou non) et les occurrences de ces mots. A l'inverse, faire le choix des liens comme principe de classement, de tri et d'organisation de l'information³, c'est revendiquer clairement le choix de l'immatériel ou à tout le moins le choix de l'information comme mesure « *d'une différence qui produit une autre différence* » [Bateson 77 p.231].

Quand nous consultons une page de résultat de Google ou de tout autre moteur utilisant un algorithme semblable, nous ne disposons pas simplement du résultat d'un croisement combinatoire binaire entre des pages répondant à la requête et d'autres n'y répondant pas ou moins (*matching*). Nous disposons d'une vue sur le monde (*watching*) dont la neutralité est clairement absente. Derrière la liste de ces résultats se donnent à lire des principes de classification du savoir et d'autres encore plus implicites d'organisation des connaissances. C'est ce rapport particulier entre la (re-)quête d'un individu et la (re-)présentation d'une connaissance qui était présente dans les bibliothèques de la Haute-Egypte, pour en être évacuée avec l'arrivée des principes de classement alphabétiques.

Une nouvelle logique se donne à lire. Moins « subjective » que les principes classificatoires retenus par une élite minoritaire (clergé, etc.) elle n'en est pas moins sujette à caution. Les premières étaient douteuses mais lisibles, celles-ci le sont tout autant parce qu'illisible⁴, c'est-à-dire invisibles : l'affichage lisible d'une liste de résultats, est le résultat de l'itération de principes non plus seulement

³ Comme ce fut le cas pour la révolution entraînée par l'algorithme PageRank du moteur Google (www.google.com) qui considéra que la pertinence d'une page était liée en priorité au nombre de pages la référençant (liens entrants) et non plus exclusivement à des mesures d'occurrence linguistique. Ce critère (inspiré de Garfield et de la bibliométrie) est actuellement pris en compte par la plupart des outils de recherche.

⁴ pour les utilisateurs non spécialistes.

implicites (comme les plans de classement ou les langages documentaires utilisés dans les bibliothèques) mais invisibles et surtout dynamiques, le classement de la liste répondant à la requête étant susceptible d'évoluer en interaction avec le nombre et le type de requêtes ainsi qu'en interaction avec le renforcement (ou l'effacement) des liens pointant vers les pages présentées dans la page de résultat.

Ainsi, à mesure que se tissent, à chaque instant de nouveaux liens entre les nouvelles entités (documentaires ou non) composant le réseau, à mesure que ceux-ci n'ajoutent pas simplement à une complexité existante mais la reconfigurent à chaque instant, et à mesure que s'affirment comme les plus efficaces des algorithmes de recherche, ceux systématisant la part faite à l'objectivation de phénomènes subjectifs (*best practices*, pages pivots et d'autorité ...) l'horizon qui se dessine pour la contribution des sciences de l'information à l'organisation de la connaissance est désormais celui pointé par [Carr et al. 99], qui indiquent, en conclusion de leur article :

« *Le challenge est désormais de construire des systèmes capables d'extraire ou d'apprendre la sémantique des connaissances implicites dans le média et de construire des associations entre ces représentations liées au média et la sémantique, sans qu'il y ait pour cela besoin de lourdes entrées manuelles de données. Rechercher et naviguer plus directement à partir des concepts, plutôt qu'à partir de leurs représentations variées, sera alors une réalité.* »

Dans cette perspective là le rôle fondamental des ontologies s'affirme chaque jour davantage (www.semantic-web.org).

4 La sérendipité

4.1 Définition dans le contexte de l'IR

Nous plaçant du point de vue de l'IR dans des environnements distribués (Internet) nous définissons la sérendipité comme la propagation d'un style cognitif stable (mis en place au début de la session de navigation) dans un environnement différent mais contenant de l'information pertinente pour l'utilisateur dans le contexte initial de sa navigation et au vu de la tâche qu'il s'était assigné. Constatant alors que cette procédure donne des résultats permettant de satisfaire ses besoins de manière non prévue il va mettre en place de nouvelles stratégies de navigation lui permettant d'amorcer un nouveau cycle, soit en assignant un nouvel objet-cible à sa recherche, soit en initiant un nouveau parcours permettant d'atteindre l'objet-cible initial.

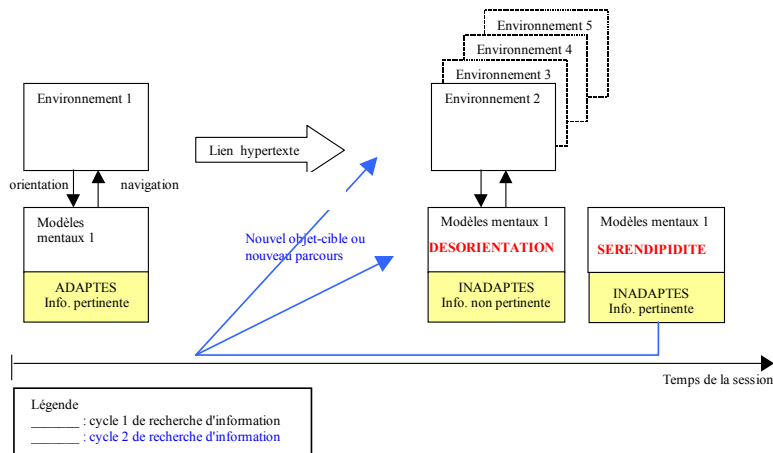


Fig. 2 : Sérendipité et cycle de l'IR

4.2 Tentatives de sériation de la sérendipité.

Une première sériation de la sérendipité peut être observée à partir des quatre équations de Figueiredo et Campos [Figueiredo, 01] (Cf. Fig 3). Les auteurs formalisent sous forme d'équations (simplification) la sérendipité en distinguant un problème (P), le contexte du problème (KP), la métaphore (M), le contexte de la métaphore (KM), la solution (S), le contexte de la solution (KS) et le gain de connaissance dans le processus de formulation du problème (KN). Ces quatre équations reposent en fait sur un « déclencheur », la métaphore comme moyen de provoquer la perspicacité :

- métaphore,
- métaphore inattendue
- absence de métaphore
- métaphore de l'ignorance

Dans la première équation la métaphore inattendue inspire la solution. Dans la seconde équation la métaphore inattendue conduit à un nouveau problème puis à une nouvelle solution. Dans la troisième équation, l'absence de métaphore impose un pragmatisme, un problème trouve écho à un autre problème et propose ainsi une nouvelle solution. Dans la quatrième équation, la métaphore de l'ignorance introduit l'erreur dans le contexte de la description du problème, elle implique un nouveau problème, puis une nouvelle solution.

| | |
|--|---|
| <p>1. Pseudosérendipité, exemple d'Archimede</p> $P1 \subset (KP1) \Rightarrow S1 \subset (KP1, KM, KN)$ $M \subset (KM)$ | <p>2. sérendipité avec Méthaphore, exemple de Rontgens (Xray)</p> $P1 \subset (KP1) \Rightarrow P2 \subset (KP2)$ $M \subset (KM) \Rightarrow S2 \subset (KP2, KM, KN)$ |
| <p>3. Sérendipité sans Méthaphore , exemple de la 2cv Citroën</p> $P1 \subset (KP1) \Rightarrow P2 \subset (KP2)$ $S2 \subset (KP2, KN)$ | <p>4. sérendipité avec métaphore de l'ignorance, exemple de Christophe Colomb</p> $P1 \subset (KP1, EP1) \Rightarrow P2 \subset (KP2)$ $S2 \subset (KP2, KN)$ |

Fig 3 : les équations de la sérendipité

De son côté, [Toms 00] propose de distinguer entre le raisonnement par analogie (favorisant la sérendipité) et ce qu'elle nomme *blind luck* où seul le hasard est à l'origine d'une découverte informationnelle. Ce type de sérendipité peut se rencontrer dans le cas de générateurs aléatoires de nœuds (graphes) hypertextuels. Elle rappelle également l'importance du "principe de Pasteur" selon lequel "le hasard ne favorise que les esprits préparés". Apparaît ainsi l'idée qu'il est nécessaire d'amorcer le processus de sérendipité, c'est à dire de l'inclure dans un cycle initial de recherche.

[Boursier & Van Andel, 92] proposent de « qualifier » ce que [Figueirado 01] ont mis en équation et parlent de « sérendipité positive » pour désigner l'« observation d'un fait non anticipé suivi d'une abduction correcte », de « sérendipité négative » pour désigner l'observation d'un fait ou accomplissement d'une tâche sans interprétation juste (équation n°4, C. Colomb) et enfin de « pseudo-sérendipité » lorsque l'on cherche quelque chose qui avait déjà été conceptualisé mais qu'on le trouve par un autre chemin que celui initialement prévu (équation n°1)

La dernière tentative de sériation que nous avons repéré vient du monde de l'intelligence économique dont les problématiques de « veille » garantissent le lien avec les sciences de l'information et de la communication. [Marti, 02] se focalise sur les aspects volontaristes ou proactifs de la sérendipité et cite trois contextes possibles :

- celui du groupe Bourbaki qui avait pris comme habitude d'inviter à ses conférences de jeunes confrères en leur demandant d'intervenir sur des domaines où ils n'avaient aucune expérience, pariant ainsi sur leur fraîcheur d'esprit pour apporter idées neuves ;
- celui de la technique du « Pot de Miel » quand par exemple IBM offre un accès gratuit à des brevets (www.delphion.com) mais qu'il s'en sert pour repérer des tendances technologiques et observer le comportement et les requêtes de ses concurrents (traçage adresses IP) ;

- celui enfin du groupe 3M (leader de l'innovation), le plus proactif de tous, qui « oblige » ses scientifiques à consacrer 15% de leur temps à des axes de recherche en dehors de ceux définis par la R&D.

4.3. L'apport des modèles de l'IR : vers une sériation plus globale.

Ce paragraphe tente de mettre en avant une vue globale du processus de recherche d'information, du point de vue des différents types de sérendipité qu'il autorise (ou interdit). Il existe 3 « états initiaux » de l'IR, auxquels sont associés 3 processus, trois types de tâches, qui font eux-mêmes référence à 3 grands types de modèles.

| Etat initial | | Processus | Modèles |
|------------------|--|--|---------------------|
| [Je sais] | [ce que je cherche] | Querying / Browsing Sérendipité nulle | Computational |
| [Je ne sais pas] | [ce que je cherche] | Searching Sérendipité structurelle | Utilisateur |
| [Je sais] | [que je ne sais pas ce que je cherche] | Learning Sérendipité associative | Environnementaliste |

Le premier cas représenté dans ce tableau (Je sais ce que je cherche) 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 computationnel 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. La sérendipité est alors quasi-nulle ou ne relève en tout cas d'aucune démarche volontariste ou consciente.

Le second cas (Je ne sais pas ce que je cherche) correspond à l'objectif de l'IR selon [Belkin, 00], à savoir : « *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 ». La sérendipité qui se met ici en place est de type structurelle (cf. infra)

Le dernier cas (Je sais que je ne sais pas ce que je cherche) 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. Nous sommes alors dans le cadre d'un authentique processus « d'apprentissage périphérique » tel que défini par [Lave & Wenger, 91]. Dans ce processus, l'information qui sera prioritairement « captée » par l'utilisateur et servira de base aux associations qu'il échaffaudera pour aller au bout de sa quête, cette information donc, relève en premier lieu des propriétés invariantes de l'environnement : de la même manière que je peux utiliser un stylo comme un stylo si je veux écrire, ou comme un marteau si je veux planter un clou, je peux utiliser la liste des 10 premiers résultats d'un moteur de recherche de manière systématique (et aller voir chacune des pages vers lesquelles ils pointent) ou de manière associative pour repérer aléatoirement (dans le texte de description fourni pour chaque page par exemple) de nouveaux mots-clés, de nouveaux noms de personnes qui vont m'engager sur une autre piste de recherche ou vont en l'état constituer une réponse/solution à ma question/problème⁵. La sérendipité est ici de type associative (cf. infra).

4.1.1 Sérendipité structurelle.

Admettons pour l'exemple, que nous nous trouvions dans une bibliothèque, à la recherche d'une thèse déjà repérée, pour construire un état de l'art sur une question donnée. Non loin de la thèse recherchée, sur le même rayonnage, figure une autre thèse dont le titre est évocateur et dans laquelle, après lecture rapide, nous trouvons effectivement des informations intéressantes. La sérendipité ici à l'œuvre est de type structurelle : elle est liée à une identification, à un parallélisme

⁵ C'est ce type de processus qui est systématisé par la plupart des outils de recherche ayant fait le choix de représentations graphiques (Kartoo, Mapstan, etc ...) pour optimiser l'instrumentalisation de ce type de sérendipité.

formel, structurel (de fait on est dans le rayonnement des thèses et non dans celui des journaux qui eût été moins approprié pour l'objectif de notre recherche initiale).

Admettons maintenant que nous effectuions la même recherche, dans la même bibliothèque, mais cette fois en consultant l'une des bases de données dont elle dispose : on utilisera alors les champs structurés de la base de donnée pour exprimer notre requête (mots du titre, nom de l'auteur, etc.). Selon la règle de *matching* applicable à tout type d'information structurée, l'échelle du phénomène de sérendipité se réduit considérablement, même si elle reste possible (un même auteur ayant pu rédiger deux thèses différentes par exemple) et demeure de nature structurelle.

4.1.2 Sérendipité associative.

Sur Internet, et plus généralement dans tout système distribué d'information non-structurée, ce phénomène change de nature et se donne à lire avec une acuité déterminante dans les stratégies de navigation choisies par les utilisateurs. Si l'on interroge un moteur de recherche en entrant une série de mots-clés (qui peuvent être les mêmes que ceux utilisés pour l'interrogation de la base de donnée), deux cas se présentent :

- le moteur de recherche dispose, dans sa base de donnée ou dans sa base d'index, d'informations présentant un relatif niveau de structuration (c'est par exemple le cas des annuaires de recherche si on les interroge en utilisant les catégories qu'ils proposent) : le phénomène de sérendipité structurelle reste opérant. Au vu du nombre de résultats possibles, dans ce cas comme dans les deux premiers évoqués (interrogation du rayonnement des thèses ou d'une base de donnée), le facteur déterminant consiste à limiter le silence (absence de résultats) ;

- le moteur de recherche ne dispose pas d'information structurée – ce qui demeure le cas le plus fréquent – et les listes de résultats qu'il présente à la requête de l'utilisateur sont alors considérables. La sérendipité se manifestant cette fois dans l'affichage possible d'un résultat pertinent bien que ne correspondant pas aux termes exacts de la requête est alors de nature associative. Le facteur déterminant dans les stratégies de navigation qui seront alors mises en place par l'internaute est celui qui lui permettra d'éviter le bruit et non plus le silence.

Notons ici que le sérendipité associative résulte de la conjugaison de phénomènes sémantiques, algorithmiques, individuels (usages) et techniques (référencement, balises méta⁶, spam⁷ ...). On peut par ailleurs constater, avec la dernière génération de moteurs de recherche que le facteur déterminant redevient celui du modèle classique, c'est-à-dire éviter le silence⁸.

5. Conclusion

5.1. Sérendipité et recherche d'information.

La sérendipité dans le cadre d'un processus de recherche d'information peut-être passagère (le temps que les modèles mentaux adéquats soient appelés) ou devenir un mode privilégié d'accès à l'information dans le cadre d'un processus de recherche ou de l'une de ses itérations. Elle se décline sous deux formes exclusives (structurelle et associative) qui dépendent principalement de variables d'environnement (structuré ou non).

Cette sérendipité a comme mérite méthodologique d'attester - s'il en était encore besoin - qu'il n'est pas nécessairement plus facile de trouver de l'information dans un système ordonné, structuré et formaté que, comme cela semble être le cas pour le web, dans un système d'information caractérisé par une forte entropie et ne disposant en tout cas d'aucun niveau de contrôle unique⁹. Pour autant, il nous semble essentiel de se donner les moyens de penser la diffusion d'information et la structuration de contenus numériques en des termes qui prendront en compte, à la source, les sauts conceptuels et

⁶ En HTML, ces balises permettent aux auteurs de contrôler l'indexation de leurs documents.

⁷ Le "spam" désigne les pratiques frauduleuses qui permettent de "fausser" l'indexation d'un document (faux mots-clés ...)

⁸ Certaines pratiques sont à ce titre tout à fait éclairantes du point de vue d'une "sociologie" de la recherche d'information, comme celle du "GoogleWhacking" (<http://www.googlewhack.com>) : Google étant le moteur de recherche le plus en vogue et celui disposant de la plus grande base d'index, cette pratique consiste à formuler des requêtes ne ramenant qu'une seule réponse.

⁹ Un protocole expérimental est actuellement en cours à l'Université de Toulouse 1, auprès d'étudiants en première année de DEUG de droit pour évaluer les usages novices en recherche d'information. Le phénomène de sérendipité peut ainsi être "expérimentalement" observé.

autres ruptures d'arborescence dont se nourrit la sérendipité. Les principales voies de recherche œuvrant actuellement dans ce domaine sont celles du web sémantique, des hypermédias d'apprentissage et bien entendu des approches théoriques de la recherche d'information (IR).

Parallèlement, une étude globale des scénarios de navigation disponibles sur le web [Ertzscheid 02] doit permettre d'identifier quelques invariants qui, à leur tour, constitueront un recours précieux permettant d'aller dans le sens d'une plus grande adéquation entre les objectifs visés par l'hypertexte, les habitus techniques sollicités et les styles cognitifs à l'œuvre chez l'utilisateur.

5.2. Recherche d'information et complexité

Au vu des quelques éléments décrits dans cet article (contexte réseau, diachronie des techniques de gestion de l'information, nouvelles approches de l'indexation et nouveaux biais, nouveaux comportements et nouvelles pratiques), il nous semble que le champ d'étude que constitue la recherche d'information doit être défini comme un **processus d'apprentissage dynamique**. C'est à la lumière de ce processus que devient chaque jour plus perceptible le **renouvellement des technologies intellectuelles** de classement, de représentation et d'accès aux connaissances. Dans l'utilisation qui est faite des ontologies, dans les thématiques qui émergent de domaines connexes à l'InfoCom (ingénierie documentaire, ingénierie des connaissances, hypermédias pédagogiques, etc ...), dans les perspectives ouvertes par le web sémantique et la main mise sur le web de technologies agents de plus en plus sophistiquées et transparentes, il semble clair que ce renouvellement des technologies intellectuelles passe par la combinaison - au sein de systèmes d'informations (eux-mêmes de plus en plus complexes et distribués) - de modèles formels hérités et de modèles plus ouverts, c'est à dire intégrant l'entropie comme partie intégrante du processus. Alors, au prix de la mise en œuvre d'un cycle cohérent et repérable de gestion des connaissances, l'émergence peut être prise en compte, de nouveaux éléments d'information voir le jour et le processus de recherche bénéficiant de cet enrichissement constant en le déclinant sur plusieurs niveaux dépendant de l'acculturation de l'usager à ces phénomènes. C'est, nous semble-t-il, à ce prix que l'adéquation nécessaire entre les méthodologies de recherche et le monde et/ou les objets qu'elles veulent cerner et décrire demeurera pérenne.

5.3. Sérendipité et hypertexte

Si l'hypertexte et plus globalement les hypermédias générés constituent un terrain d'observation et d'expérimentation privilégié pour l'étude de la sérendipité c'est parce qu'ils sont l'unique moyen d'organisation et de classification des connaissances qui « offre comme capacité inhérente la création de classifications latérales. » [Balasubramanian 94] C'est cette dimension de « latéralité »¹⁰ que tentent en permanence d'implémenter différents outils de recherche pour offrir des pistes d'accès à des mondes des plus en plus complexes¹¹.

5.4. Sérendipité et créativité

Appréhender dans leur complexité les phénomènes, les objets de recherches semble une tâche impossible. Pour sortir de la « boucle récursive »¹² ou graphe complexe le chercheur doit entreprendre différentes stratégies et opter pour les choix qui s'offrent à lui. La troisième voie est introduite par la sérendipité. Elle est une approche socio-cognitive de la recherche d'information et impose l'abduction comme heuristique.

Pour tenter d'expliquer l'influence de la sérendipité en matière de construction de connaissances, il semble que deux dimensions soient à retenir : l'importance du contexte et le transfert de compétences dans une situation nouvelle (métaphore). Le contexte est notamment composé de la connaissance et des technologies intellectuelles qui la manipulent. Le transfert de compétences dans une situation nouvelle est lié à l'appropriation d'une culture technique et informationnelle, de savoirs par les chercheurs et leur capacité à transposer, transfigurer des phénomènes, des problèmes.

¹⁰ On parle de " latéralité " en recherche documentaire à propos de la reformulation de requêtes. De Bono, "Lateral Thinking", Penguin Books, 1990

¹¹ Voir notamment le mouvement initié par les "Folders" de Northernlight, repris actuellement par des outils comme Vivissimo ou constituant le cœur de technologie de sociétés (Exalead).

¹² Morin E, « La méthode »

La sérendipité se réalise alors par l'appropriation individuelle du contexte socio-technique, une *lecture* spécifique, créative du réservoir cognitif et instrumental. Les chercheurs les plus en phase avec le contexte socio-technique favorisent ainsi leur perspicacité et la mise en œuvre d'artefacts informationnels qui permettent de faire apparaître des éléments stochastiques.

Lors d'achoppements du processus de production scientifique ou de surcharge cognitive [Ertzscheid 03], quand il est impossible de rendre compte de phénomènes, un saut qualitatif doit être réalisé... la sérendipité guette.

6. Bibliographie

Bachimont B., « Du texte à l'hypertexte les parcours de la mémoire documentaire », Technologie, Idéologies, Pratiques, n° spécial « Mémoires collectives », 1999.

Balasubramanian V., « State of the Art on Hypermedia Issues And Applications. » [en ligne] http://www.isg.sfu.ca/~duchier/misc/hypertext_review/, consulté le 26/10/2001.

Barabasi, A.-L., Jeong H., Albert R., " The Diameter of the World Wide Web ", pp.130-131 in Nature, 401, 1999. [en ligne] http://xxx.lanl.gov/PS_Cache/cond-mat/pdf/9907/9907038.pdf, consulté le 05/07/2002.

Bateson G., Vers une écologie de l'esprit, T. 1. Paris, Seuil, 1977.

Belkin N., *Helping People Find What They Don't Know*, in Communications of the ACM, August 2000, Vol. 43, No. 8.

Boursier & Van Andel, « Serendipity : expect also the unexpected », creativity and innovation management, vol 3, p.20-32, 1992.

Bush V., « *As We May Think*. », pp. 101-108, in The Atlantic Monthly, vol.1, n°176, Juillet 1945. [en ligne] <http://www.isg.sfu.ca/~duchier/misc/vbush>, consulté le 07/02/1998.

Carr L., Hall W., Lewis P.H., De Roure D., « *The significance of Linking*. », in ACM Computing Surveys, vol. 31, n°4, Décembre 1999. [en ligne] http://www.cs.brown.edu/memex/ACM_HypertextTestbed/papers/20.html, consulté le 22/03/2002.

Engelbart D.C., « Augmenting Human Intellect : a Conceptual Framework », Summary Report, AFOSR-3233, Stanford Research Institute (SRI), Contract AF49(638)-1024, SRI Project N° 3578, Octobre 1962. [en ligne] <http://www.histech.rwth-aachen.de/www/quellen/engelbart/ahi62index.html>, consulté le 03/03/2002.

Ertzscheid O., Les enjeux cognitifs et stylistiques de l'organisation hypertextuelle, Thèse en Sciences de l'information et de la communication, Université de Toulouse 2, sous la dir. de FC Gaudard & J. Link-Pezet, 450 pages. [en ligne] <http://www.ertzscheid.net>, consulté le 10/06/03.

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

Fayet-Scribe S. "histoire de la documentation en France : Culture science et technologie de l'information", CNRS éditions, 2000.

Figueirado A. Dias de, Campos J., "The Serendipity Equations". *Proceedings of the Workshop Program at the Fourth International Conference on Case-Based Reasoning*, ICCBR 2001, Technical Note AIC-01-003. Washington, DC: Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence [en ligne] max.ipv.pt/pub/AdeFigueiredo01.pdf

Gallezot G., « La recherche in silico » In : Chartron G. (dir.) *Les chercheurs et la documentation électronique : nouveaux services, nouveaux usages*, Edition du cercle de la Librairie, Coll. Bibliothèque, juillet 2002.

Gallezot G., "Exploration informationnelle et construction des connaissances en génomique", Les Cahiers du numérique, Hermès, vol.3, n°3, novembre 2002.

Kleinberg J. L. S., « The structure of the web », Science, vol 294, 30 nov. 2001, p 1849-1850.

Koll Matthew, "Information Retrieval", bulletin de Jasis vol. 26, N°2 Dec/jan 2000.
<http://www.asis.org/Bulletin/Jan-00/track_3.html>

Kolmayer Elisabeth, Peyrelong Marie-France, « Partage de connaissances ou partage de documents », Document numérique . vol 3(3/4):283-299. 01 décembre 1999. et
http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/01/00/index_fr.html

Lave G., Wenger E., Situated Learning : Legitimate Peripheral Participation. New-York, Cambridge University Press, 1991.

Marti Y.-M., "Dirigeants : quelle posture de combat ?" <<http://www.Egideria.fr/posturecombat.html>>

Perriault J., " Effet diligence, effet serendip et autres défis pour les sciences de l'information. " [en ligne] <http://www.limsi.fr/WkG/PCD2000/textes/perriault.html>, consulté le 15/02/01.

Shneiderman, B. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley Publishing Company, Reading, MA, 1997.

Toms Elaine G. « Serendipitous Information Retrieval » < http://www.ercim.org/publication/ws-proceedings/DelNoe01/3_Toms.pdf >

Virilio, P., La machine de vision, Ed. Galilée, Paris,1988.