RECIT EXTRAORDINAIRE
SUR LES FILTRES & ALGORITHMES

'*Petite étude*' initiée par Jacques-François Marchandise, et réalisé dans le cadre du projet 'télécommande simplifiée'

Automne 2011 - Hiver 2012

Maria Graves

Nous naviguons dans un océan d'informations.

Que cela soit face à l'Internet, à notre télévision ou à d'autres appareils informatifs, nous arrivons à des situations d'hyper-choix. Devant un potentiel informationnel incontestable, demeurent de nombreuses difficultés. Sur l'Internet la recherche d'informations n'est pas toujours très facile. Monsieur Goldfinger à dit " *Plus l'information est abondante, moins on dispose de temps pour la traiter, plus il y a de chance de mal interpréter les données disponibles et de manquer l'information pertinente.*" Cela est tout à fait transposable dans le domaine de la télévision: comment trouver le bon cheminement pour arriver à la donnée désirée? Ceci dit, la quantité n'est pas la difficulté unique. Sur l'Internet, davantage de facteurs entrent en jeu: il n'y ni règlementation ni standards qui régissent. Cet extrême liberté peut parfois paraître comme étant une anarchie. Les moteurs de recherches et annuaires réagissent différemment à une même requête. Et puis, l'information évolue, se restructure, le récent s'accumule sur l'ancien sans le faire disparaître, etc. Les acteurs, eux aussi, se modifient, apparaissent, acquièrent de nouveaux attributs, … Cette facilité et liberté de publication pause un autre problème, celui de la fiabilité des sources.

Entre nous usagers et l'information qui nous est proposé, il y a des outils spécifiques qui font les calculs, les liens et 'réfléchissent' aux
réponses qui nous seront exposés. Je me permet donc de vous soumettre un récit à propos de ces outils-programmes, numériques ou pas, qui dans "l'hyper-hyper choix" font des sélections, et nous laissant seulement avec un "hyper-choix".

LE PRINCIPE DES 'FILTRES'/ FILTRES DU WEB

Définition de « Filtre »

Il est 7h du matin. Le café a enfin infusé, prêt à être avalé. Heureusement d'ailleurs... J.F empoigne sa tasse. Son regard soulagé s'immobilise tadis son visage se déconfie soudainement.
Le filtre a laissé s'échapper des milliers de particules de café dans la tasse...

Ce terme « filtre » , en tant que métaphore de la passoire, est l'action de filtrage, un principe de 'découpage'.
Les «algorithmes», eux, sont les lignes de codes qui permettent d'effectuer les calculs tandis que les « agents » les exécutent.

Les moteurs de recherche

Lorsqu'on cherche quelque chose sur le Web avec un moteur, la recherche n'est menée que dans la base de données de ce moteur. Et les informations présente dans le moteur de recherche sont celles qui ont été préalablement récupérées par son agent. Chaque moteur aboutit à des résultats différents, puisque chacun possède sa propre base de données, gérée selon ses propres critères. Un moteur de recherche est donc composé de deux éléments. D'abord d'un «Agent» qui a pour but de visiter, récupérer et stocker les divers documents qui apparaissent par millions sur le Web. Et puis, d'analyser les données pour créer des liens entre les mots-clés de l'utilisateur et les informations appropriées. L'efficacité du moteur de recherche est due à la rapidité à laquelle les agents identifient les nouveaux sites et à la finesse de leur analyse. Certains ne scrutent que les titres et les premières lignes de texte des pages tandis que d'autres sont plus exhaustifs, n'ignorant que le contenu multimédia.
Par rapport aux répertoires de recherche que nous verrons plus bas,

l'usager risque (et en réalité c'est souvent le cas) d'être submergé par une masse d'information dans la quelle peu de documents seront réellement pertinents.

Aujourd'hui, le numéro 1 des moteurs de recherche en France, Allemagne, Espagne et Royaume-Uni, avec environ 90% des recherches qui passent par lui, est, sans surprise, Google! Il faut savoir que l'indexation des agents avance à des vitesses diverses, généralement avec un débit limité, à cause de la croissances rapide de la quantité de données. Les agents ont donc des 'priorités' différentes. Par exemple, sur Google, les blogs ont une indexation rapide, qui est de l'ordre de quelques minutes. Les pages avec un 'Rank' élevé aussi ont une indexation quotidienne. Que deviennent des autres pages Web ?
Google, domine la «distribution» de l'information sur le Web. Quels sont les dangers d'un tel monopole ?

Les moteurs de recherche à l'étranger

Aussi, dépendant du pays dans le quel on est, un moteur de recherche aura des réponses différentes pour une même requête donnée.

Le Lycos américain:http://www.lycos.com/
Le Lycos anglais: http://www.lycos.co.uk/
Ou encore le Lycos français: http://www.lycos.fr/

Je pense à un autre exemple, la censure de la liberté d'expression sur l'Internet en Chine.
En 2006, Google accepte de se plier aux règles de censure pour obtenir l'autorisation de lancer Google.cn. Le moteur de recherche s'est incliné durant quelques années au filtrage, «great firewall», chinois : pas de critique sur la politique du pays et censure des politiques contraires à celle en vigueur, pas de pornographie, etc.
Extrait d'un débat sur InternetActu.net : {…} Kevin Wen:"Le fait que Google applique la censure n'est une surprise pour personne. Toute entreprise implantée en Chine doit adapter sa stratégie. Les entreprises chinoises font exactement la même chose et doivent se conformer aux règles du gouvernement.
Tous les chinois savent que leur parole est censurée. Nous n'avons pas besoin que les étrangers nous le rappellent continuellement. Nous ne savons pas comment cela changera, mais nous espérons que ça changera.
Et si vous parlez tout le temps de censure, il est facile d'oublier que nous chinois avons notre vie, faisons beaucoup d'autres choses. [...] Tout change et les gens tenteront de contourner la 'Grande muraille électronique' pour obtenir des informations. Je ne pense pas que la censure du gouvernement et de Google puisse bloquer toutes les informations et toutes les actualités. Les choses peuvent s'organiser de beaucoup de manières différentes{…}".

Cela soulève un réel questionnement sur ce qu'est un filtre, et quelles applications doit-il avoir: a-t-il pour but de simplifier les recherches des utilisateurs ou est ce qu'il y a la une tentative de maîtrise et de censure ? Et quelle est la place du pouvoir politique dans ces outils ? Aussi, qui crée les algorithmes ? Ici, Google, est une société privée. Quelles sont les règles, ou plutôt, est ce qu'il y a une réglementation ? Et si ce n'est pas le cas encore, y en aura-t-il une ?

Les rubriques des moteurs de recherche

Bien plus visible et d'accès plus claire, certains moteurs de recherche sont dotés de «rubriques» qui permettent de choisir le type de média recherché: images, vidéo, carte géographique, shopping, blog, actualités, etc. Lors d'une recherche d'images par exemple, il y a d'autres filtres qui apparaissent comme les dimen-

Tiananmen, Chine. Première image de Google.cn. Deuxième de Google.com

Information de dernière minute: un générateur de faute d'orthographe de E.Borra et L.Hilfling est né en se rendant compte qu'introduire une faute d'orthographe sur le terme Tiananmen permettait d'obtenir des résultats contournant la censure sur ce terme dans l'index chinois de Google.

sions ou encore les dominantes de couleur.
Récemment, Google.com a mis en ligne «Recipe view», c'est un filtre qui permet de recherche des recettes par le nom d'un plat,d'un ingrédient ou encore par le nom d'un évènement. Des rubriques supplémentaires permettent de trouver des variables d'un même plat avec un ingrédient en plus ou en moins, par le temps de cuisson nécessaire, ou par la quantité de calories !

Lien dune video explicative :
http://www.youtube.com/watch?feature=player_embedded&v=IsUN1dUbbM8

Les agents

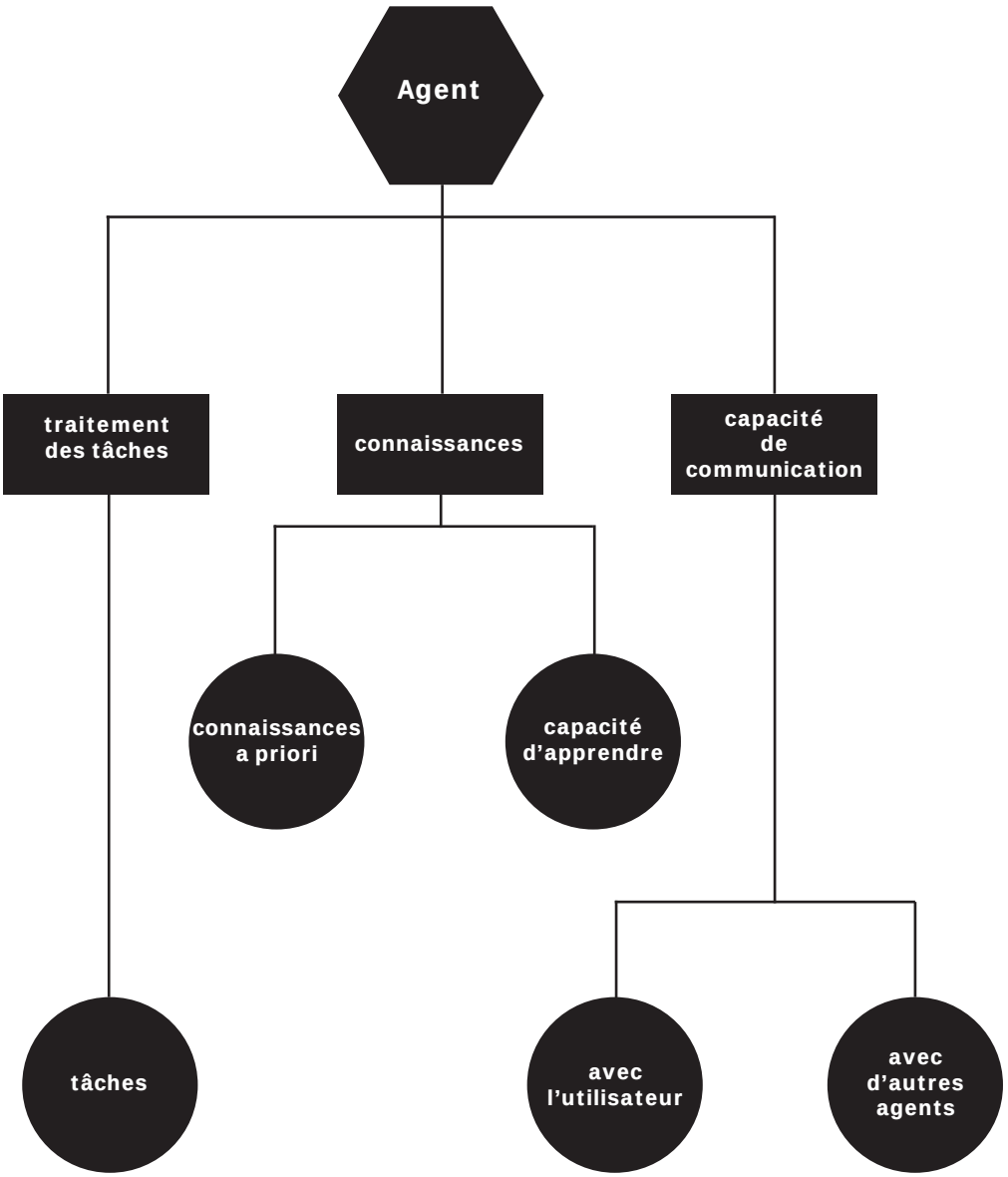
Robots, araignées (spider) ou ver (crawler), on les appelle les «Agents intelligents», de l'anglais intelligent agent. «
Jacques Feber, dans son livre Les systèmes multi-agents. Vers une intelligence collective, 1995, propose cette définition : «On appelle agent une entité physique ou virtuelle:
a/ qui est capable d'agir dans un environnement,
b/ qui peut communiquer directement avec d'autres agents,
c/ qui est mue par un ensemble de tendances,
d/ qui possède des ressources propres,
e/ qui est capable de percevoir (mais de manière limité) son environnement,
f/ qui ne dispose que d'une représentation partielle de cet environnement (et éventuellement aucune),
g/ qui possède des compétences et offre des services,
h/ qui peut éventuellement se reproduire,
i/ dont le comportement tend à satisfaire ses objectifs, en tenant compte de ressources et des compétences dont elle dispose, et en fonction de sa perception, de ses représentations et des communications qu'elle reçoit.»

Benoît Drouillat définit ces agents intelligents en 2010 ainsi :«Un agent intelligent est autonome, il prend ses propres décisions. Un agent intelligent est réactif (il répond aux changements de son environnement) et pro-actif (il poursuit ses objectifs de manière persistante).Il s'adapte aux situations afin d'accomplir ses objectifs, voire il est capable de prendre ou d'exploiter la connaissance. Il est dit «social», c'est à dire qu'il peut dialoguer avec d'autres agents suivant les termes des interactions humaines comme la négociation, la coordination, la coopération et le travail d'équipe. Le champ des agents intelligents trouve ses sources dans l'Intelligence Artificielle Distribuée (DAI) et a émergé en tant que tel dans les années 90' {…}».
Les agents intelligents n'obéissent pas tous aux mêmes technologies et n'ont pas tous les mêmes objectifs. D'ailleurs, au jour d'aujourd'hui il n'existe pas d'agent intelligent qui soit en mesure d'avoir une forte autonomie, interagir et collaborer avec d'autres agents dans le but de rapporter des informations pertinentes et surtout fiables à son utilisateur.

(sources : www.csee.umbc.edu/agents & agents.www.media.mit.edu/groups/agents & www.agent.org)

Les répertoires de recherches

Les 'répertoires de recherche' sont des outils que l'on appelle également «Index» ou «Annuaires». L'information est regroupée en grandes catégories, que l'utilisateur va pouvoir affiner pour cerner sa recherche. Ces services essaient de recenser les ressources d'Internet en les classant par thèmes, rubriques et sous-rubriques. La recherche se fait donc par couches successives, une sorte «d'arborescences». Ces répertoires sont hiérarchisés selon le choix de chaque moteur de recherche auquel ils sont affiliés. Malheureusement ils sont peu précis,



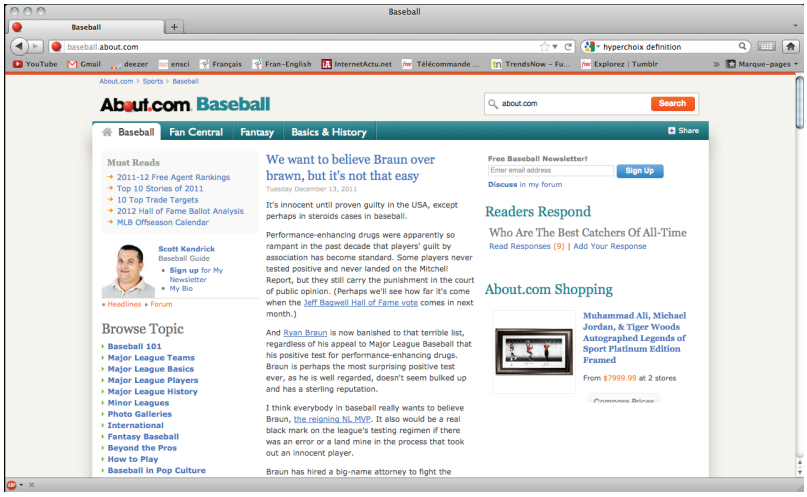
Fonctionnement d'un Agent intelligent

et donc restent peu efficaces pour des recherches pointues.
Prenons deux exemples : about.com & yahoo.com
About.com donne accès à des guides thématiques avec des commentaires.
Yahoo! lui, est considéré comme l'un des plus anciens annuaires et des plus complets pour les recherches thématiques.
http://fr.local.yahoo.com/
Il existe également d'autres répertoires, tels que les répertoires «géographiques» et les «sites carrefours». La première catégorie permet de retrouver l'information en suivant une démarche géographique comme sur annuaire-géo.fr.
La deuxième d'accéder à des sites d'information d'un secteur donné par des liens d'un spécialiste du même domaine. Par exemple, il existe une catégorie de scientifiques sur delicious.com qui s'échange des marques-pages Internet ou encore le site «le carrefour du futur» qui répertorie les sites (et autres médias, notamment des livres) sur la prospective.
En utilisant ces répertoires de catégories, nous obtenons généralement bien moins de réponses que dans le cas d'une recherche par mots-clés. Mais les sites des répertoires ont été classé par la main de l'homme d'où une pertinence supérieure.

Le Web sémantique (& l'intelligence artificielle)

Je vous propose de commencer avec un exemple: tapez dans votre barre de recherche Google «Président de la Russie», regardez bien les résultats obtenus. Puis tapez «Chef d'état de la Russie». Vous remarquerez que les résultats (je porte votre attention sur le troisième en particulier) ne sont pas identiques alors que le sujet est le même mais formulé autrement.
Dans les années 80', Tim Berner-Lee a inventé le web tel que nous le connaissons aujourd'hui : il nous permet d'accéder à des ressources sur le réseau informatique. Les moteurs de recherches lient nos demandes aux résultats par la syntaxe. Le Web Sémantique a pour objectif de permettre à des programmes informatiques d'exploiter le contenu des ressources du web grâce à un système de métadonnées formelles. C'est-à-dire d'interpréter la sémantique des pages web et donc d'avoir une seule liste de résultats quels que soient les synonymes employés lors de la recherche.
Seulement les programmes informatiques n'utilisent pas le même langage que nous, il faut donc re-écrire le web en utilisant un langage approprié pour que cette application puisse être réellement performante.
Pour le moment le web sémantique est en développement.
Cela étant, il existe déjà un début d'applications du web sémantique :1.le Google américain a intégré une fonction de recherche sémantique.
2. «Powerset», société appartenante à Microsoft, mais qui ne répond pas toujours clairement à la recherche, il affiche plus souvent un paragraphe ou un article qui contient la réponse.
3. «TrueKnowledge», en développement à Cambridge, on peut lui poser une question à laquelle il répondra, et s'il ne connait pas la réponse, il vous propose de la lui apprendre.

Sur un blog consacré à l'Intelligence Artificielle, un utilisateur propose une énigme:
« Qu'est ce qui marche à quatre pattes, fait du lait et peut se trouver en partie dans une voiture?». Puis il réfléchit aux rapprochements que ferait un moteur de recherche sémantique :
1/ Quatre pattes = animal, mobilier, …
2/ + qui fait du lait = mammifère
3/ + une partie du corps = tête, pied, peau, …
4/ + dans une voiture = l'intérieur = volant, levier de vitesse, siège,…
Après avoir analysé les différents éléments de l'énigme, la réponse



About.com : images des sous-rebriques de la catégorie ‘Baseball’

Illustration de l'exemple utilisé dans la partie 'Web sémantique',Chef d'état de la Russie

du moteur de recherche pourrait être : « Une vache. C'est un mammifère capable de produire du lait, qui a quatre pattes et dont la peau peut être utilisé en tant que revêtement de l'intérieur d'une l'automobile». Puis, il pose la même énigme à une femme qui lui répond «une femme enceinte». Et cela nous ramène sans doute à la restitution d'un souvenir d'une expérience passé tandis que le moteur de recherche sémantique aurait distingué, regroupé et puis restitué des informations. Le moteur et nous Humains, avons des bases de connaissances, mais c'est l'interprétation, subjective ou objective, que nous en faisons, qui nuance les réponses de chacun.

LES ALGORITHMES DERRIÈRE CES FILTRES

Comment marchent ils ?

'PageRank' (en français : rang de page) est un algorithme d'analyse de Google qui détermine l'ordre dans le quel les résultats d'une recherche vont apparaître. C'est très important pour les sites Web d'être bien référencés car la majorité des utilisateurs ne vont pas plus loin que les trois premiers résultats…

Le principe de PageRank est de calculer un «indice de popularité», c'est à dire une valeur proportionnelle, de liens existants (internes et externes) vers une page donnée. Cet indice est utilisé pour trier les résultats d'une recherche par mots-clés. Il se définit ainsi : l'indice de popularité d'une page est d'autant plus grand lorsqu'un grand nombre de pages réputés la référencent.

Il y a encore quelques années, il était possible de «falsifier» des liens sur une page web interne ce qui permettait d'augmenter son PageRank. Cette technique à été majoritairement utilisé par des sites commerciaux mais qui n'est plus d'actualité.

'Panda' est un autre algorithme de Google qui, lui aussi, gère le classement des pages Web, mais dans le sens inverse de PageRank : il a pour but de réduire le rang des pages Web qui calquent leur contenu sur d'autres pages ou encore celles considérées comme ayant peu de valeur ajoutée. Panda augmenterait donc la pertinence des résultats de recherche. Par exemple les «fermes de contenus», des sites pauvres en informations, qui n'ont pour but que de générer des 'revenus publicitaires', peuvent perdre jusqu'à 90% de visibilité.

En revanche, cet algorithme ne fait pas de différence entre les sites qui fabriquent leur contenus et les sites qui copient les contenus. C'est ainsi que plusieurs sites ont subit des déclassements "non-mérités".

(source : http://www.01net.com/editorial/537378/panda-le-nouvel-algorithme-de-google-entre-en-action-en-france/)

'Search Result Ranking Based en Trust'

En 2009, Google depose un brevet en rapport avec la citation suivante : " A search engine system provides search results that are ranked according to a measure of the trust associated with entities that have provided labels for the documents in the search results. A search engine receives a query and selects documents relevant to the query. The search engine also determines labels associated with selected documents, and the trust ranks of the entities that provided the labels. The trust ranks are used to determine trust factors for the respective documents. The trust factors are used to adjust information retrieval scores of the documents. The search results are then ranked based on the adjusted information retrieval scores." (source :http://www.freepatentsonline.com/7603350.html)

Il s'agit d'un algorithme, «PersonRank», qui viendrait compléter PageRank en envoyant à Google des informations sur les sites visités par les utilisateurs. Cela permettrait potentiellement à Google d'accéder à des données personnelles, comme aux accès depuis

s'agirait de récupérer les informations sur les comportements des utilisateurs dans le but de personnaliser leur résultats de recherche. Cela se ferait en parallèle d'un référencement fait par les usagers et le niveau de confiance qu'ils accorderaient à des pages en les annotant. Apparemment, les commentaires fait par les utilisateurs sur les pages Web de Twitter & Facebook, participeraient également aux référencements des pages Web.

{ N'ayant pas trouvé énormément d'informations liées au PersonRank, je reste dubitative sur la fiabilité des sources qui m'ont permis de construire cette partie. Cela dit, en relisant des articles sur «My Google History Search», je pense qu'il y a des liens à faire. My Google History Search doit surement fonctionner sur un système d'algorithme, le même utilisé dans le concept de télécommande «Goab», qui calcul au fur et à mesure les préférences, la fréquence et la durée des usages de chacun. }

Les limites des algorithmes

En allant plus loin dans mes recherches sur les filtres et algorithmes de Google, j'ai trouvé un article du New York Times concernant un algorithme de Google (http://www.nytimes.com/2010/11/28/business/28borker.html?pagewanted=all) : Clarabelle Rodriguez, une internaute américaine souhaite acheter une paire de lunettes. En tapant le nom de la marque qu'elle désire acquérir, en premier lien apparait le site «DecorMyEyes.com». Ce référencement sur Google atteste de la pertinence du site face à la recherche faite et l'accrédite d'une marque de confiance.
La cliente, n'étant pas satisfaite de produit, demande un remboursement. Il s'en suit insultes, menaces et harcèlement de la part du commerçant, qui refuse le remboursement du produit. La cliente, évalue le commerçant sur getSatisfaction.com. Elle s'aperçoit que ce commerçant jouit déjà d'une mauvaise réputation, et qu'il en tire profit en connaissance de cause : plus il a d'avis négatifs, plus il gagne de visibilité sur Google. Avec environ 300 plaintes par an, son site reste en première position dans la liste de recherche Google.
Cela remet en cause la fiabilité des algorithmes et de leur "intelligence artificielle": l'algorithme d'analyse de Google n'inclue pas d'analyse de sentiment. Il est capable de donner la fréquence de visite d'un site, et de calculer son'indice de popularité', mais il n' pas capable de faire la différence entre un indice de popularité «positif» et un indice de popularité «négatif». Jusqu'à il n'y pas si longtemps du moins. A la suite de la publication de cet article, Google aurait modifier son algorithme de classement pour qu'il prenne en compte ce paramètre.
Néanmoins, on peut se demander quel est le degré de responsabilité de Google dans ce genre de situation?

New York Times, illustration de l'article «A bully finds a pulpit on the Web»

J'ai choisi cette thématique sur les filtres et algorithmes afin de comprendre leur fonctionnement, l'intérêt qu'ils ont pour les utilisateurs et pour les compagnies qui les crées et cela dans le but de mieux penser l'interface et les proccessus qui sont à la base mon projet de 'télécommande connecté'.
Cette 'étude', telle qu'elle est présenté maintenant, ressemble plus à une énumération des divers filtres et de quelques algorithmes. Lorsque j'ai entrepris ce travail, je pensais être un utilisateur aguerri d'Internet. Puis je me suis vite rendu compte de l'étendu des possibles et de la complexité seul d'une recherche. Donc, ce qui est présenté ici, serait davantage une entrée en matière.
Il est évident pour moi, qu'une télécommande-connecté, comme le projet Goab basé sur un enregistrement de préférences au fur et à mesure de son utilisation, est directement lié à ces questions d'algorithmes, filtres et de l'appartenance de l'information, donc de son statut fonctionnel mais aussi éthique . Qu'adviennent toutes les requêtes faites par l'utilisateur? Quelles normes juridiques s'appliquement aux traitements, à l'accès à ces requêtes, aux enregistrements, et si ces informations aux mains des groupes des moteurs de recherches concernant l'usager conservent leur intérêts fondamentaux ou préconisent-t-ils le stockage des données à des fins répressives ?