Les requêtes QIC (Query Is Commercial) dans l'algorithme des moteurs de recherche

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Toutes les requêtes saisies par les internautes ne sont pas traitées de la même façon par les moteurs de recherche comme Google ou Bing. En effet, plusieurs algorithmes de détection du type de demande effectuée ont été mis en place pour traiter de façon différente notamment les requêtes à caractère commercial, afin notamment d'optimiser leur monétisation. Cet article rappelle un historique des différentes classifications proposées depuis 20 ans et explique comment Google ou Bing testent et analysent ces différentes requêtes puis réagissent à leur caractère plus ou moins commercial...

Les moteurs de recherche sont des entreprises commerciales qui cherchent à tirer profit de toutes les sources de monétisation auxquelles ils ont accès... Or une partie des requêtes tapées sur les moteurs sont effectuées par des internautes à la recherche d'un produit ou d'un service. Il est donc assez logique que les moteurs aient cherché très tôt à identifier ces requêtes, pour optimiser leur monétisation comme pour améliorer le service rendu aux internautes.

Nous allons essayer de révéler dans cet articles les méthodes utilisées pour identifier ces requêtes commerciales ou requêtes marchandes, quels sont les obstacles auxquels les moteurs sont confrontés, et comment ces informations sont utilisées pour modifier les réponses apportées lorsque des requêtes "à caractère commercial" sont tapés par les internautes.

Comment "classifier" les requêtes des internautes ?

L'idée de classer les requêtes en différentes catégories est née assez tardivement dans l'histoire des moteurs de recherche. Dans les années 90, curieusement, la plupart des analyses sur la typologie des requêtes semblaient partir du principe que la seule intention des utilisateurs de moteurs était de chercher de l'information. Mais Andrei Broder a tordu le cou à ce mythe dans un article fondateur datant de 2002 : "a taxonomy of web search" (http://www.sigir.org/forum/F2002/broder.pdf).

La classification de Broder a eu un grand succès dans le monde des moteurs de recherches, et est reprise dans de très nombreux travaux de recherche. Elle classe les requêtes des internautes en trois catégories, en fonction des intentions des internautes :
- les requêtes navigationnelles ;
- les requêtes informationnelles ;
- les requêtes transactionnelles.

Les requêtes **navigationnelles** correspondent à des recherches d'URL dans un moteur de recherche (par exemple : www.lemonde.fr ou "télé 7 jours"). L'objectif de l'utilisateur est de trouver une page particulière sur le web.

Les requêtes **informationnelles** révèlent une recherche d'information par un internaute, le moteur est donc utilisé pour trouver une page susceptible de contenir l'information.

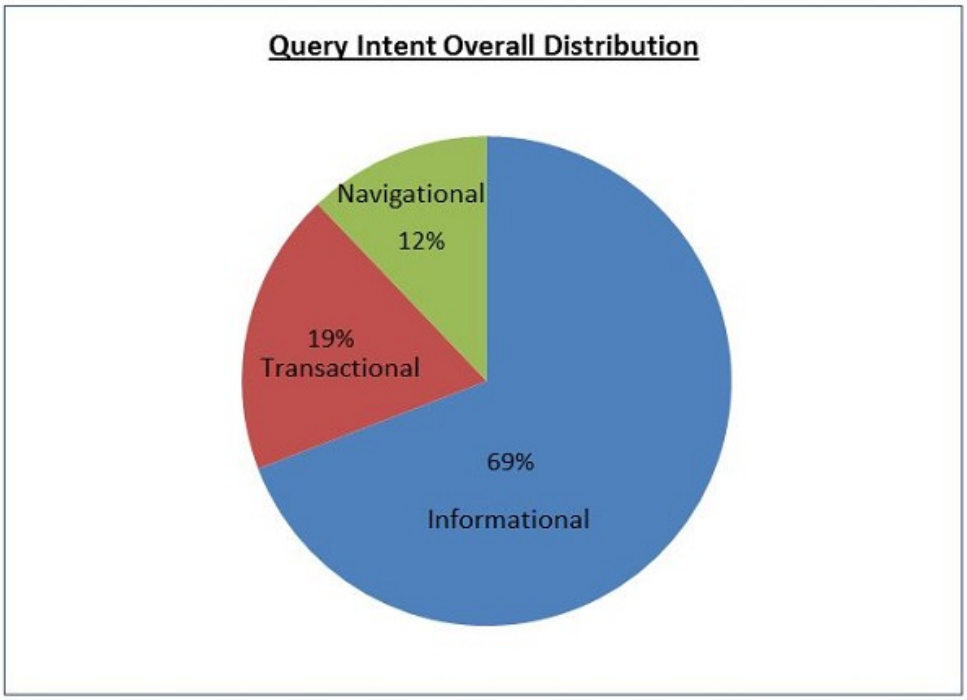
Les requêtes **transactionnelles** sont tapées par des internautes qui ont l'intention de réaliser une action sur le web : s'abonner à un service, acheter un produit, télécharger un fichier...

Type of query	User Survey	Query Log Analysis
Navigational	24.5%	20%
Informational	?? (estimated 39%)	48%
Transactional	> 22% (estimated 36%)	30%

Figure 5. Query classification.

Répartition des requêtes observées dans l'étude d'Andrei Broder en 2002

La répartition entre ces requêtes évolue avec le temps. Les moteurs de recherche ont longtemps été conçus pour répondre correctement à une requête navigationnelle (l'utilisateur tape des mots clés et le moteur renvoie des URL). Ensuite les utilisateurs ont appris à utiliser les moteurs de recherche pour trouver de l'information (en fait pour trouver les pages qui contiennent l'information). Avec le temps, la proportion des requêtes informationnelles a fortement augmenté. De même, avec l'explosion des services sur Internet et du commerce électronique, le volume des requêtes transactionnelles en général, et commerciales en particulier, a également connu une croissance importante. Hélas, la classification d'Andrei Broder, si elle est simple, est aussi assez floue, et il est assez difficile de comparer les différentes études qui se sont succédées sur la répartition des requêtes entre les différentes catégories : voici ci-dessous une analyse effectuée en 2010...



Répartition donnée par une étude datant de 2010 :
http://www.librarystudentjournal.org/index.php/lsj/article/view/228/305

La classification de Broder est aussi malaisée à utiliser dès lors que l'on s'intéresse à des requêtes commerciales. En effet, chercher de l'information sur un produit (requête informationnelle), ou chercher un site de marque ou un site marchand (requête navigationnelle) révèle également une intention d'achat. Par ailleurs, toutes les requêtes transactionnelles dans la définition de Broder, n'ont pas toutes un caractère commercial. Certains spécialistes rajoutent parfois un distingo entre les requêtes transactionnelles pures

(qui appellent une action immédiate), et les requêtes de type "préparation de transaction ou d'achat", qui précèdent la transaction.

		Query Intent		
		Navigational	Informational	Transactiona/ Commercial
Visitor Intent	Seeking Information	Traffic Source: Search Engines, Industry Directories | Interest on the website: Usually jump off after the first page/ 2nd page, Editorial content, Blog, FAQ, Register for Newsletter | | Commercial Intent: -		+
	Finding Options		Traffic Source: Domain Traffic, Brand Traffic from Search Engines, Email News Letters, Industry Directories, Online Demo Videos, Press Release, Articles on third party websites | Interest on the website: White papers, Product Description, Comparison, Register for Information/ update, Register for Podcast/ Webcast | Commercial Intent: $	
	Decision stage			Traffic Source: Domain Traffic, Brand-Model number Traffic from Search Engines, Email News Letters, Industry Directories, Online Demo Videos, Press Release, Articles on third party websites | Interest on the website: Product Litrature/ Manual, Comparison, Contact Form, Trial Download, Shopping Cart | Commercial Intent: $$$$

Une tentative de croisement de l' « intention commerciale » avec la catégorisation de Broder. On remarque qu'il y a recouvrement, les requêtes « marchandes » se retrouvent dans deux catégories.

Dans la pratique l'identification des requêtes commerciales demande donc une approche spécifique.

Les approches possibles pour classer les requêtes commerciales

La détection des mots "outils" révélant une intention d'achat

Dans certains cas, les mots clés tapés peuvent révéler une intention d'achat : "acheter chaussures adidas pas cher" est un exemple typique de ce type de requêtes. Il est donc possible d'identifier quelques mots clés qui dénotent clairement une intention d'achat.
Cette approche "naïve" trouve assez rapidement ses limites. Il est assez fréquent qu'une requête commerciale prenne la forme d'une combinaison de type "marque modèle", comme dans "Nikon D90" ou "Renault Twingo". Pour ces requêtes, même en l'absence de mots "outils" dénotant une intention d'achat, il est clair que la probabilité d'être confronté à un utilisateur à la recherche d'un produit précis est très élevée.

La reconnaissance des entités nommées : identifier les marques, les produits, les modèles...

L'une des solutions passe par la "reconnaissance des entités nommées" que nous avons déjà eu l'occasion d'aborder dans de précédents articles de la lettre "Recherche & Référencement" d'Abondance. L'idée consiste à "étiqueter" les termes d'une requête, et de remarquer que dans la requête "Nikon D90", Nikon est une marque, et "D90" est un modèle de la marque Nikon. Reconnaître dans la requête des marques commerciales, ou des références produits aide clairement à identifier les requêtes à caractère commercial.

Mais cette approche a elle aussi ses limites : la "reconnaissance des entités nommées" ne fonctionne pas à tous les coups. Par exemple les requêtes contenant "orange" ne sont pas toutes des requêtes à propos des produits d'une célèbre société de téléphonie... Gare aux faux positifs !

En effet, l'un des problèmes que pose la simple analyse des termes d'une requête est le manque de contexte. Pour affiner l'approche, il est indispensable d'aller à la recherche d'informations de contexte.

L'approche combinée : classification des pages de destination et l'analyse des logs de requête

Dans la pratique, la détection de l'intention des utilisateurs en fonction de la requête tapée (c'est ce que l'on appelle en anglais l'approche QIR : *Query Intent Resolution*) est possible en utilisant deux types d'approches, en général combinées : la classification des pages de destination d'une part, et l'analyse des logs de requête d'autre part.

La classification des pages de destination

L'idée est simple : si l'internaute finit par cliquer sur une page marquée comme "commerciale", alors j'en déduis que la requête associée est commerciale. L'avantage de l'analyse des pages web de destination, c'est qu'elles sont riches en information, et qu'un moteur de recherche dispose déjà de beaucoup d'informations de contexte sur ces pages, et sur les sites qui les contiennent.

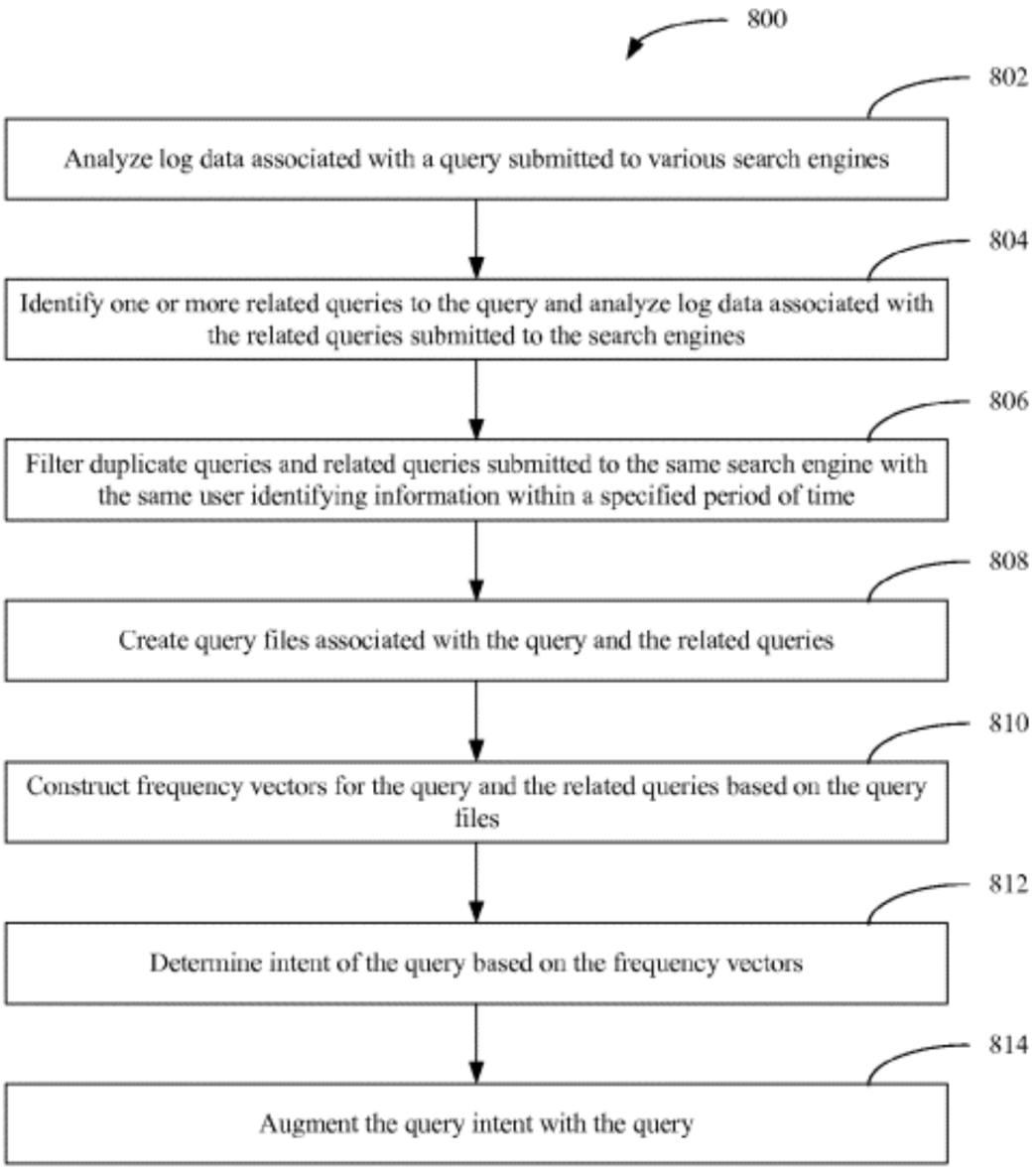
L'approche la plus commune pour réaliser automatiquement ce type de classification est d'utiliser un algorithme d'apprentissage automatique. On demande à des cobayes de "classer" à la main un certain nombre de pages. Ensuite, on essaie de produire un "classifieur", à l'aide d'un algorithme de type "*machine learning*". Le classifieur est une fonction d'évaluation qui "note" les pages, et détermine une probabilité d'appartenance à une catégorie ou à une autre. Une fois les pages ainsi classées en "commerciales" ou "non commerciales", il suffit d'analyser quels types de pages renvoient différents types de requête : une requête sera classée comme commerciale si elle renvoie beaucoup de pages commerciales, et comme non commerciales si elle renvoie peu de pages commerciales.

Notons qu'à ce stade deux problèmes apparaissent. Tout d'abord, un grand nombre de requêtes sur un moteur renvoient à la fois des pages "commerciales" et "non commerciales". Il y a donc de fortes chances qu'une requête contenant le terme "orange" ne puisse pas être catégorisée avec ce système. Ensuite, on utilise le moteur (avec ses défauts) pour classer des requêtes, ce qui évidemment produit des effets de bord indésirables.

L'analyse des logs de requêtes pour améliorer la classification

Pour améliorer la qualité des résultats, l'une des solutions classiques consiste à analyser les logs de requêtes. Les logs de requêtes sont des fichiers générés par les moteurs de recherche, qui enregistrent scrupuleusement toutes les requêtes effectuée par les utilisateurs, ainsi que les informations permettant d'identifier l'utilisateur (enfin son adresse IP et les identifiants de session, comme dans un log de serveur web).

L'analyse des différentes requêtes tapées par un utilisateur au cours de la même session fournit des informations très utiles sur les intentions des utilisateurs. Si "abonnement orange" est reformulé régulièrement au cours d'une même session en "pack orange iphone", on peut en déduire que "pack orange iphone" représente une requête plus précise caractérisant ce que cherche un utilisateur qui a tapé "abonnement orange". L'analyse des clics sur les pages de résultat permet ensuite de valider l'intention d'achat précise de l'utilisateur.

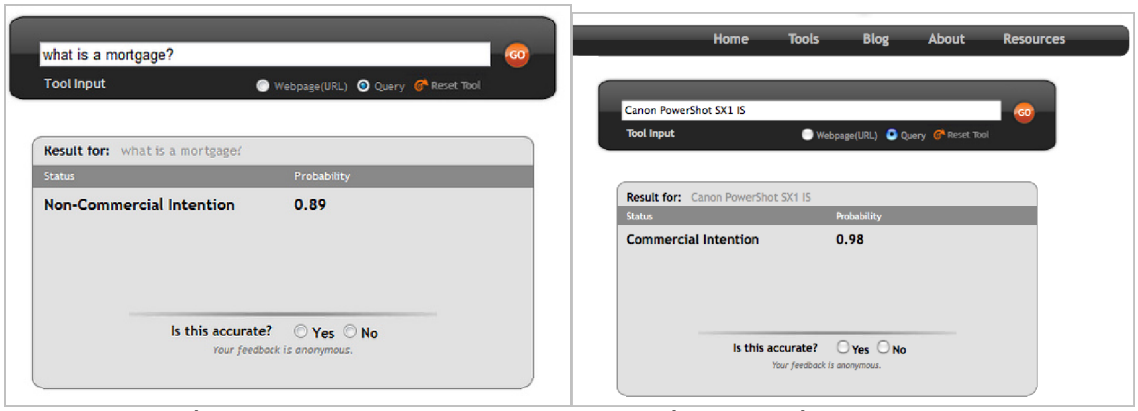


Processus d'analyse des logs de requêtes utilisé pour améliorer l'identification de l'intention de l'utilisateur lors de l'entrée d'une requête. Schéma extrait d'un brevet attribué à Yahoo !

Une application concrète : l'outil Online Commercial Intent de Microsoft / Bing

Pendant plusieurs années, Microsoft a proposé dans ses Ad Labs un outil baptisé OCI, qui donnait une probabilité qu'une expression soit "commerciale" ou non. Cet outil n'est hélas plus disponible depuis plusieurs mois.

Les méthodes utilisées pour collecter les données utilisées dans cet outil ont été présentées par les chercheurs de Microsoft Research dans plusieurs "papiers" scientifiques, que l'on retrouvera dans la bibliographie à la fin de cet article. Cet outil utilisait une approche semblable à celle que nous venons de décrire.



Deux copies d'écran de l'outil OCI. A gauche, une requête identifiée comme non commerciale par l'outil OCI (« what is a mortgage ? »). A droite une requête jugée commerciale avec une probabilité proche de 100% (« Canon Powershot SX1 iS).

Dépasser la logique binaire commercial / non commercial

L'un des inconvénients de l'outil OCI de Microsoft, c'est qu'il était basé sur une classification binaire : soit une expression était classée comme commerciale, soit comme non commerciale. Dans la pratique il existe évidemment un bon nombre de requêtes ambigües, dont le statut est "indéterminé". Mais de plus, cette logique binaire reflète mal la variété des "intentions" dans les requêtes commerciales.

Les équipes de recherche de Yahoo! ont publié de nombreux articles et brevets sur la détection des étapes d'achat et leur classification. On trouve notamment la classification de requêtes suivantes dans les travaux menés sous la direction du chercheur chilien Ricardo Baeza-Yates :
- guide d'achat ;
- avis ;
- support ;
- page produit officielle ;
- site marchand - achat.

Les requêtes "**guide d'achat**" correspondent à une recherche d'informations en vue de choisir un produit ou un type de produit. Par exemple "barbecue gaz ou charbon". A ce stade, l'internaute ne sait pas quel produit il va acheter exactement.

Les requêtes "**avis**" correspondent à l'étape suivante dans la recherche d'un produit : dans ce cas, l'internaute a déjà en tête un ou plusieurs produits précis, et veut recueillir des informations pour faire son choix définitif. Typiquement, à cette étape un internaute tape "avis Nikon D90".

Les requêtes "**support**" interviennent après l'achat, et constituent des recherches d'information pour dépanner un produit, mieux utiliser un produit etc. Par définition, de telles requêtes ne dénotent pas une intention d'achat et ne doivent pas être considérées comme des requêtes commerciales. Un exemple : "Iphone 4 écran cassé".

Les requêtes appelant une **page officielle** pour le produit sont intéressantes à identifier pour fournir des réponses pertinentes. En effet, la page "officielle" peut-être celle du fabricant et non celles des revendeurs, ou de leurs affiliés. Exemple : la requête "Iphone 4S" appelle une page d'Apple.

Les requêtes de type "**site marchand - achat**" appellent clairement une boutique en ligne. Ce sont des requêtes commerciales par excellence, car une transaction est envisagée au terme de la requête. Exemple : ""La redoute manteau long femme".

Ce problème a été étudié par d'autres équipes de scientifiques. Par exemple l'équipe de Patrick Pantel de Microsoft Research, a cherché à regrouper les requêtes commerciales en trois catégories :
- les requêtes de **préparation d'achat (« avis »)** ;

- les requêtes « **prix** » ;
- les requêtes « **support** ».

Review	Price	Support
best thin	where can I get cheap	common problems
rate	black Friday sale	fault codes
compare kodak vs. canon	christmas sales	installing new
small portable	cheap	easy use
high zoom	buy now pay later	operating instructions
rate best	dell discount	won't heat
battery life	great deals	best schematics
user comments	overstock.com	keeps shutting

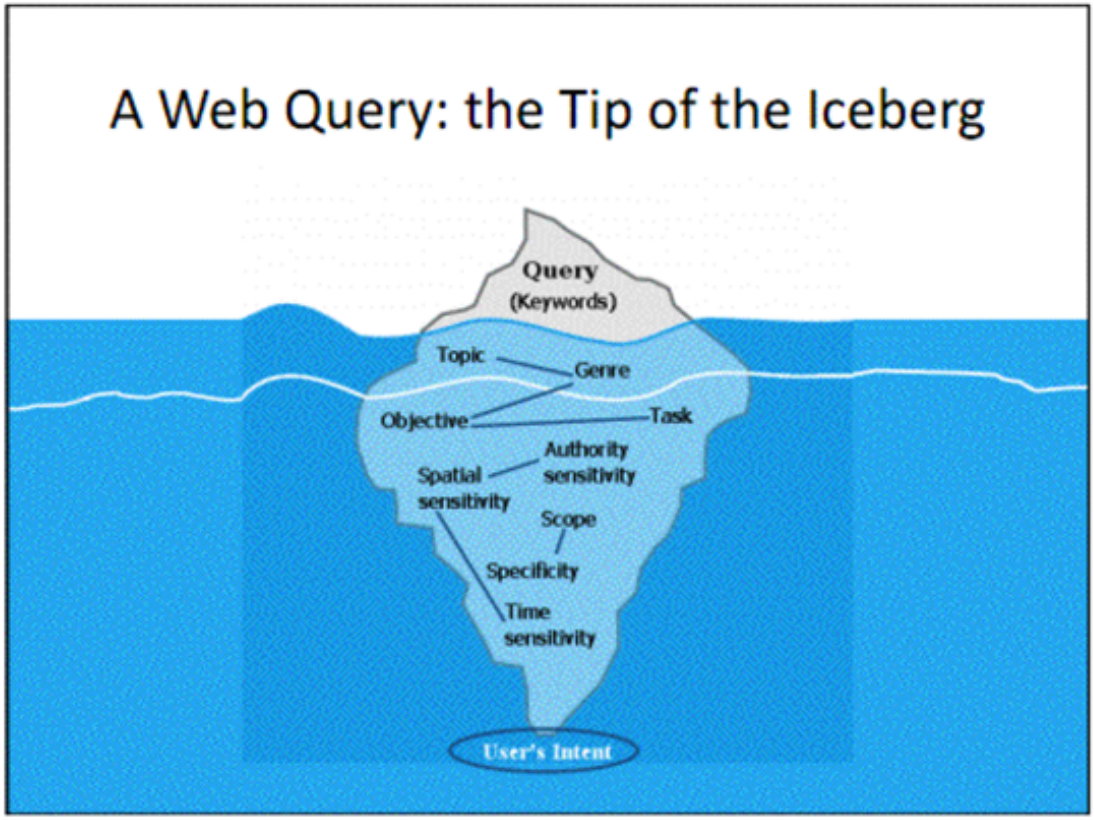
Quelques exemples de requêtes et leur classification, issu des travaux de Patrick Pantel (MSR).

L'intérêt de cette classification est d'être orientée vers la détection du type de résultats pertinents à présenter en fonction du type de requête : les expressions de type « avis » appellent des sites d'avis et des blogs, les expressions de type « prix » des sites marchands, des places de marché, et des comparateurs, et les expressions de type support des pages de forums.

Une classification opérationnelle des requêtes doit être multidimensionnelle

Avec le temps, les équipes des moteurs de recherche ont compris que la plupart des tentatives de classification n'étaient pas opérationnelles : certaines requêtes ne rentrent jamais dans les cases, et d'autres rentreraient au contraire dans plusieurs cases.

Les équipes de Yahoo ! en particulier ont développé une approche « multidimensionnelle », en essayant d'attacher un ou plusieurs attributs (baptisés « facettes ») aux requêtes. Cette approche permet de stocker l'information qualifiant une requête sous une forme beaucoup plus souple, et plus facilement réutilisable.



Un schéma illustrant l'approche de Ricardo Baeza Yates dans la qualification des requêtes à des fins de QIR : derrière la requête se cachent de multiples considérations révélant les intentions de l'utilisateur (sensibilité au temps, à l'espace, relation avec une tâche, un sujet etc.)

Cette approche en facettes est illustrée dans le fonctionnement récent de Google, par l'introduction d'un comportement nouveau :
- lorsque la requête est commerciale (facette : « query is commercial ») ;
- et qu'elle porte sur des informations temporelles comme le prix, des réductions etc… (facette : « query deserves freshness »).

Dans un tel contexte, Google est susceptible de présenter plus de résultats « frais » sur une requête de type « meilleur prix Nikon D90 ».

Summary

Facet	Values
Genre: provides a generic context to the user's query intent. and can be thought as a meta-facet.	*News. Business. Reference. Community*
Topic: a list of topics built from the first level of categories offered by ODP . Yahoo! and Wikipedia.	*Adult. Arts & Culture. Beauty Style. Cars & Transportation. Charity. Computers & Internet. Education. Entertainment. Music & Games. Finance. Food & Drink. Health. Home & Garden. Industrial Goods & Services. Politics & Government. Religion & belief systems. Science & Mathematics. Social Science. Sports. Technology & Electronic. Travel. Undefined. Work*
Task: this facet is related with the type of resource associated with the query.	*Informational. Not Informational. Ambiguous*
Objective: represents if the query is aimed to do some action or to obtain a resource.	*Resource. Action*
Specificity: this facet describes how specialized is a query.	*Specific. Medium. Broad*
Scope: the scope aims at capturing whether the query contains polysemic words or not.	*Yes. No*
Authority Sensitivity: through this facet it is possible to determining whether the query is designed to retrieve authoritative and trusted answers	*Yes. No*
Spatial Sensitivity: this facet reflects the interest of the user to get a resource related to an explicit spatial location.	*Yes. No*
Time Sensitivity: this facet captures the fact that some queries require different results when posed at different times	*Yes. No*

Quelques exemples des « facettes » utilisées dans les travaux des chercheurs de Yahoo !, sous la direction de Ricardo Baeza-Yates.

L'Utilisation de la détection des requêtes commerciales pour améliorer les résultats du moteur

Ces techniques de détection des requêtes à caractère commercial sont réutilisées en général de deux manières différentes :
- dans la lutte contre le spam ;
- dans l'algorithme de classement.

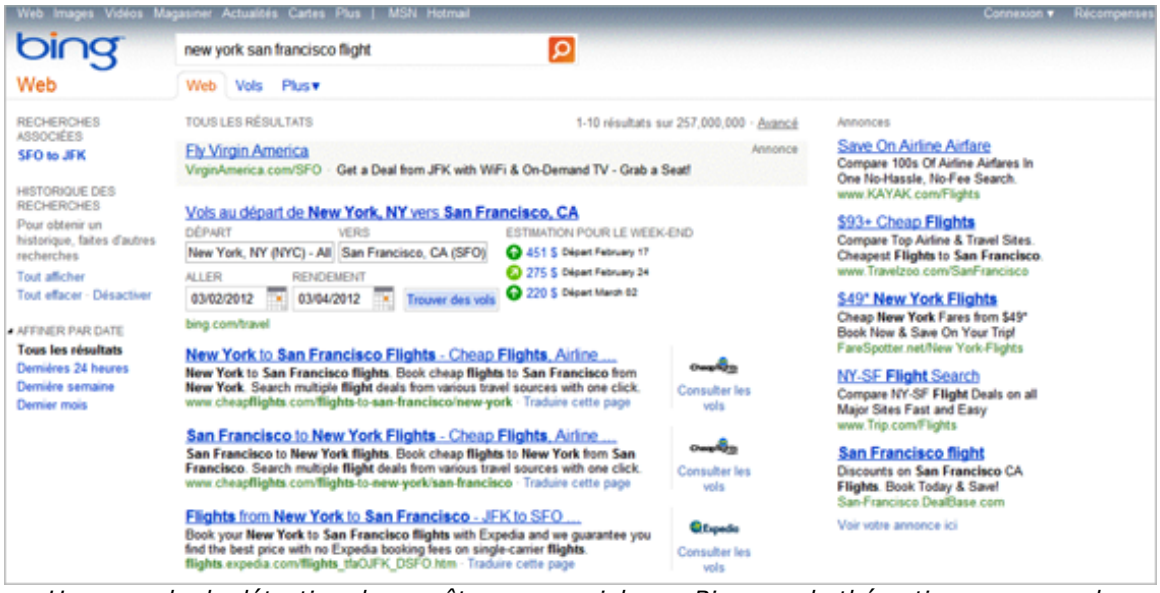
Lutte contre le spam et requêtes commerciales

L'identification des requêtes commerciales est particulièrement utile dans la lutte contre le spam, car les pages de résultats correspondant à ces requêtes sont les plus spammées et les plus polluées, compte tenu des enjeux financiers.

Plusieurs brevets décrivent des mécanismes qui tiennent compte du caractère commercial d'une requête pour durcir certains indicateurs utilisés pour détecter les pages de spam. Certains types de site qui apparaissent sur des requêtes non marchandes peuvent donc être estimés indésirables sur des requêtes commerciales, et disparaître des pages de résultats.

La modification des pages de résultats

La « résolution des intentions de requête » est utilisée d'une manière générale pour déclencher une modification de la page de résultats. Les données de QIR (*Query Intent Résolution*) déterminent les « Onebox » affichées, ainsi que les résultats de moteurs verticaux ajoutés à la page de résultat. Une requête associée à la facette « requête géolocalisée » déclenchera ainsi sur Google l'affichage d'une carte et de données de Google Places. Sur Bing, une requête associée à la facette « requête commerciale voyage» déclenchera quant à elle l'affichage des services spécialisés de Bing sur les vols secs.



Un exemple de détection de requête commerciale sur Bing, sur la thématique voyage : la requête « vol new york san francisco » déclenche l'affichage du formulaire de Bing Travel et de trois liens partenaires.

Fusion des données utilisateurs : plus de données de contexte donc plus de possibilités pour Google

Nous avons évoqué plus haut la difficulté à identifier l'intention d'un utilisateur en analysant uniquement la requête qu'il a tapée. Ces difficultés proviennent de la pauvreté du contexte qui gêne les tentatives de désambigüisation. Par contre, plus on dispose de données sur l'utilisateur lui-même, plus on peut facilement deviner ses intentions.

Gageons que la fusion des données utilisateurs entre les différents services verticaux proposés par Google sonne l'avènement de nouvelles possibilités pour ce moteur, qui pourrait proposer plus de personnalisation sur ces requêtes ou des approches de type « ventes croisées ».

Mais d'ores et déjà, les mécanismes de détection des intentions des utilisateurs sont de plus en plus visibles sur les moteurs de recherche, au point que l'on ne peut plus parler d'un algorithme de classement unique, mais de mécanismes multiples qui s'activent en fonction du type de requêtes détectés et des « attributs » associés aux requêtes.

Il faut s'adapter car toutes les requêtes n'appellent pas le même type de résultats.

Il faut donc s'adapter lorsqu'on gère un site marchand, et comprendre que le contexte d'affichage de sa page pourra être très différent en fonction des requêtes tapées par les internautes.

Sur des expressions-clés commerciales, une bonne stratégie consiste à catégoriser aussi ses requêtes cibles, en fonction des typologies décrites précédemment. Cette approche est utile pour le SEO, mais aussi pour le SEA (l'achat de mots clés).

Par exemple, la meilleure façon pour être visible aux US sur une requête commerciale et géolocalisée, est certainement de référencer son magasin dans Google Adresses, car les résultats organiques ne sont pas toujours visibles au-dessus de la ligne de flottaison. L'époque où on pouvait cibler de manière uniforme toutes les requêtes est révolue, et depuis plusieurs années déjà…

BIBLIOGRAPHIE

Articles scientifiques

A taxonomy of Web search.
Broder, A. (2002). ACM SIGIR Forum Vol. 36, No. 2, 3-10.
http://www.sigir.org/forum/F2002/broder.pdf

Exploring Searcher Interactions for Distinguishing Types of Commercial Intent
Qi Guo, Eugene Agichtein
Emory University
http://www.mathcs.emory.edu/~qguo3/wpp1323-guo.pdf

Detecting Online Commercial Intention (OCI)
Honghua (Kathy) Dai, Lingzhi Zhao , Zaiqing Nie, Ji-Rong Wen, Lee Wang, Ying Li
Microsoft Corporation
http://research.microsoft.com/en-us/um/people/znie/WWW2006-oci.pdf

Context-Aware Online Commercial Intention Detection
Derek Hao Hu, Dou Shen, Jian-Tao Sun, Qiang Yang and Zheng Chen
Hong Kong University of Science and Technology & Microsoft Research
http://www.cse.ust.hk/~derekhh/publications/acml2009.pdf

Brevets

Brevet de Yahoo! :
Method and apparatus for categorizing and presenting documents of a distributed database
Invented by Daniel C. Fain, Paul T. Ryan, and Peter Savich
US Patent Application 20060265400
Published November 23, 2006
Filed: April 28, 2006
http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PG01&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.html&r=1&f=G&l=50&s1=%2220060265400%22.PGNR.&OS=DN/20060265400&RS=DN/20060265400

Discovering Query Intent From Search Queries and Concept Networks
United States Patent Application 20110035397 Kind Code: A1
Yahoo! inc.
http://www.freepatentsonline.com/20110035397.pdf

***Philippe YONNET** , Directeur SEO international, Twenga.*

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2012/02/les-requetes-gic-query-is-commercial.html