

.....  
Université de Provence



# Etude comparative de six moteurs de recherche

.....

*Jean Véronis*

# Etude comparative de six moteurs de recherche

*Jean Véronis*

## Introduction

A la fin de l'année 2005, le moteur de recherche Google bénéficiait du nombre d'utilisateurs le plus important dans le monde, avec une proportion particulièrement élevée en France (plus de 82% du trafic selon le baromètre Xiti<sup>1</sup>). Les raisons pouvant conduire les utilisateurs à préférer un moteur à un autre sont complexes, mais si des éléments tels que rapidité, ergonomie ou esthétique entrent indéniablement en jeu, le critère qui semble légitimement central est celui de la pertinence des résultats retournés – du moins telle qu'elle peut-être perçue par les utilisateurs. On dispose toutefois de peu de données permettant de comparer cette *pertinence perçue*, et en tous cas, à notre connaissance, d'aucunes données récentes et comparatives sur la recherche d'information francophone. La présente étude essaie de pallier partiellement ce manque en fournissant un test utilisateur de six moteurs de recherche sur des requêtes en français à la fin de l'année 2005.

Les moteurs choisis sont trois moteurs américains, **Google**, **Yahoo** et **MSN**, ainsi que trois moteurs français, **Exalead**, **Voilà** (développé par France Telecom et offert sur le portail Wanadoo) et **Dir.com** du groupe Iliad, qui constitue plus une plate-forme expérimentale qu'un moteur à visée commerciale. D'autres moteurs, tels qu'AskJeeves ou mozDex, n'ont pas été retenus car ils n'offraient pas de version francophone (ou seulement une version bêta dans le cas de AskJeeves).

## Protocole

L'évaluation a été réalisée en décembre 2005 avec l'aide de 14 étudiants de première année de licence MASHS (Mathématiques appliquées aux sciences humaines et sociales) à l'Université de Provence (Aix-en-Provence), jouant le rôle d'utilisateurs.

14 thèmes ont été choisis collectivement, de façon à refléter des domaines d'utilisation très divers. Les thèmes retenus sont les suivants :

- **Actualités**
- **Animaux**
- **Cinéma**
- **Divertissement**
- **Histoire**
- **Littérature**
- **Musique**
- **Nature**

---

<sup>1</sup> <http://www.secrets2moteurs.com/barometre2005-12.html>

- **Personnages célèbres**
- **Politique**
- **Santé**
- **Sports**
- **Surnaturel**
- **Voyages**

Chaque thème a été attribué à un étudiant différent, qui choisissait librement cinq requêtes. Le format (avec ou sans guillemets, en un seul ou plusieurs mots) était également totalement libre. Par exemple, pour le thème Actualités, les requêtes choisies ont été les suivantes :

- "abbé Pierre" célibat prêtres
- chaîne télévision jeunesse TNT
- "greffe visage"
- "grippe aviaire" danger Europe
- Outreau acquittement

Il est possible que de meilleurs résultats aient pu être obtenus avec des requêtes formulées de façon différente, avec de meilleurs mots-clés ou un usage plus approprié des guillemets par exemple. Toutefois, le but ici n'était pas d'évaluer une utilisation par des experts, mais par un public de base, raisonnablement éduqué et familier des moteurs de recherche. Il était donc important de respecter les requêtes telles que le panel d'utilisateurs les a formulées.

Les requêtes ont été soumises aux différents moteurs le même jour (11 décembre 2005) par l'organisateur de l'expérience (Jean Véronis), en restreignant chaque moteur à la langue française, et en activant le filtre parental. La première page de 10 résultats *non marqués comme sponsorisés* a été archivée pour chaque requête et chaque moteur, puis débarrassée automatiquement des informations autres que les seules URL des résultats. En particulier, toute information sur le moteur de provenance a été supprimée.

Au total, 4200 URL ont été récupérées (14 thèmes x 5 requêtes x 6 moteurs x 10 résultats). Pour chaque requête, les doublons (même URL retournée par deux moteurs différents) ont été supprimés, conduisant à 3450 couples uniques requête-URL. Les couples requête-URL correspondant à chaque thème ont été fournis à l'étudiant concerné, sous forme d'un fichier Excel, dans lequel la requête et l'URL apparaissaient dans des colonnes consécutives (la requête faisant l'objet d'un lien cliquable vers le site correspondant). L'étudiant devait évaluer le document pointé par l'URL sans connaître le moteur de provenance, et reporter des informations dans des colonnes supplémentaires :

- **Lien mort** (1 si le site ne répond pas, 0 sinon)
- **Lien pornographique** (1 si le lien pointe vers un site pornographique, 0 sinon)
- **Thème** (indépendamment de la qualité de l'information, 1 si le document est dans la thématique, 0 sinon)
- **Site commercial** (1 si le lien pointe vers un site de vente en ligne, 0 sinon)
- **Pertinence** (note de 0 à 5, 0 correspondant à un document totalement inutile ou hors-thème, 5 correspondant à un document répondant de façon parfaite à la question posée).

La tâche devait être accomplie dans un délai d'une semaine (du 12 au 18 décembre).

## Liens morts

Certains des liens retournés par les moteurs sont inaccessibles au moment de l'interrogation par l'utilisateur (nous les appellerons « liens morts »). Les raisons peuvent en être multiples : la page a pu disparaître entre le moment de son indexation et le moment de la requête, ou bien un problème momentané peut en empêcher l'accès (serveur en panne par exemple). La proportion de liens morts peut varier selon le moment des requêtes et nous l'avons donc mesurée à trois reprises différentes : chaque utilisateur a noté l'information sur la « vivacité » du lien au moment de sa requête manuelle, et nous avons lancé par deux fois (à quelques jours d'intervalle) des requêtes automatisées sur l'ensemble des URL, en archivant les codes d'erreur retournés (codes HTTP 4xx et 5xx). Les résultats sont consignés dans le tableau 1.

	Dir	Exalead	Google	MSN	Voila	Yahoo
Manuel	7,6%	8,9%	2,0%	2,9%	7,4%	2,6%
Auto1	6,6%	6,1%	3,7%	1,9%	1,9%	4,7%
Auto2	5,7%	5,7%	0,7%	1,3%	2,1%	1,0%

Tableau 1 – Proportion de liens morts

La proportion de liens morts est plus importante lors des clics manuels : ceci s'explique d'une part par le fait que la procédure automatique utilisée faisait jusqu'à trois tentatives espacées par un délai de quelques minutes en cas d'échec et d'autre part par le fait qu'un certain nombre de serveurs ne retournent pas le code d'erreur 404 (« Page not found ») lorsque la page n'existe plus, mais une page HTML normale porteuse d'un message ad hoc, qui ne peut être interprétée comme erreur que par un lecteur humain.

On notera également la très grande variabilité des résultats obtenus automatiquement (lignes Auto1 et Auto2), pourtant dans des conditions strictement identiques. L'analyse détaillée des résultats montre que pour une raison indéterminée, le site *www.amazon.fr* renvoyait un code d'erreur lors de l'expérience Auto1. Or, c'est un des sites les plus retournés par les requêtes sur Google et Yahoo, et ce problème a eu un impact dramatique sur les résultats: sur les 26 erreurs comptabilisées concernant Google dans Auto1, 17 étaient dues au seul site *www.amazon.fr*, tandis que chez Yahoo, le site était responsable de 23 erreurs sur 33.

Dans le reste de l'étude seuls les liens actifs dans la phase manuelle ont été considérés.

## Liens pornographiques

On sait que des liens à caractère pornographiques se glissent dans les résultats de requêtes sans caractère pornographique, l'ingéniosité des référenceurs réussissant à les faire remonter artificiellement dans le classement par des techniques relevant du spam. La situation a pu être particulièrement critique par le passé, mais les moteurs offrent désormais tous une fonction de filtre parental permettant d'assainir les résultats. Leur efficacité est notable puisque sur l'ensemble des URL retournées seulement deux (une retournée par Voilà, l'autre par MSN) renvoient à des sites pornographiques (sans lien aucun, évidemment, avec les requêtes).

## Liens commerciaux

Ont été considérés comme commerciaux les liens figurant parmi les liens normaux, *non marqués comme sponsorisés*, renvoyant vers les sites proposant des achats ou transactions en ligne. La proportion en est très variable selon les moteurs, puisqu'elle va du simple au double (Tableau 2).

	Dir	Exalead	Google	MSN	Voila	Yahoo
Toutes positions	8,3%	8,0%	7,7%	7,1%	15,6%	10,9%
Position 1	9,0%	9,4%	2,9%	10,1%	32,3%	10,4%

Tableau 2 – Proportion de liens commerciaux

Si l'on ne considère que le premier résultat retourné (il a une importance particulière, puisque c'est le lien le plus cliqué par les internautes), on s'aperçoit que les moteurs ont des stratégies opposées. Dir, Exalead et Yahoo ne font pas apparaître de différence particulière. La proportion de liens commerciaux s'accroît en première position pour MSN et (très fortement puisqu'elle double) pour Voilà. En revanche, la proportion diminue nettement pour Google.

Parmi les sites commerciaux qui apparaissent au moins 10 fois dans un des moteurs, seules figurent trois sociétés : Amazon, E-Bay et PriceMinister. Leur association avec les différents moteurs est intéressante à étudier (dans cet ordre). Google et Yahoo sont fortement associés à Amazon, tandis que Voilà préfère Ebay et PriceMinister. Les autres moteurs ne semblent pas avoir d'affinités particulières avec les sites marchands. Globalement, c'est MSN qui renvoie le moins de liens vers des sites commerciaux avec 7,1%.

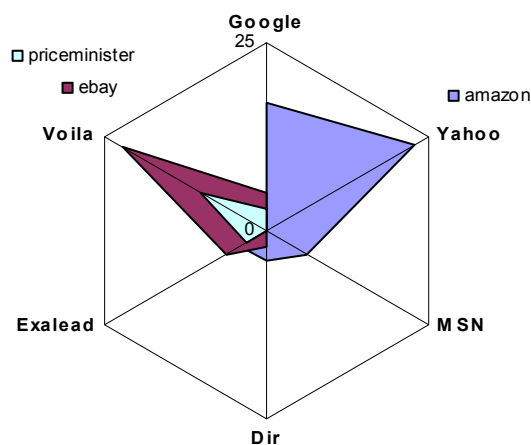


Figure 1 – Associations préférentielles entre moteurs et sites commerciaux

## Recouvrement des résultats

Le recouvrement des résultats entre moteurs est extrêmement faible, puisqu'il varie, selon les couples considérés entre 2,9% (Dir/Voila) et 25,1% (Google/Yahoo) (tableau 4).

	Dir	Exalead	Google	MSN	Voila	Yahoo
Dir	--	5,9%	6,4%	5,7%	2,9%	6,7%
Exalead	5,9%	--	12,1%	10,1%	6,4%	11,9%
Google	6,4%	12,1%	--	18,9%	7%	25,1%
MSN	5,7%	10,1%	18,9%	--	5,7%	16,6%
Voila	2,9%	6,4%	7%	5,7%	--	6,7%
Yahoo	6,7%	11,9%	25,1%	16,6%	6,7%	--

Tableau 3 – Résultats communs par couple de moteurs

Sur l'ensemble des URL uniques retournées par les 6 moteurs, moins de 10% sont retournées par au moins deux moteurs (figure 2).

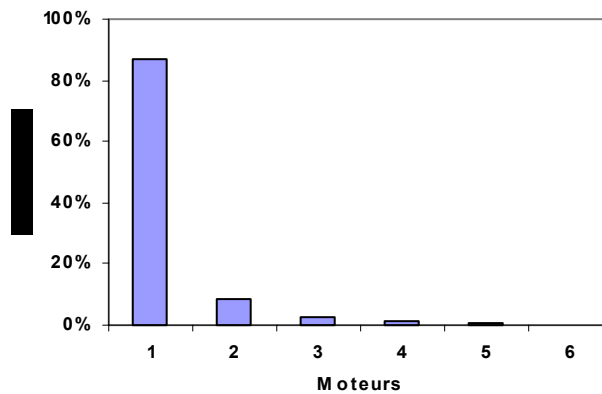


Figure 2 – Répartition des URL uniques en fonction du nombre de moteurs qui les retournent

La proximité entre différents moteurs peut être calculée et représentée de façon graphique, sur la base du nombre de résultats qu'ils partagent, grâce à une technique dite classification ascendante hiérarchique (figure 3). On voit que les moteurs les plus proches sont Google et Yahoo.

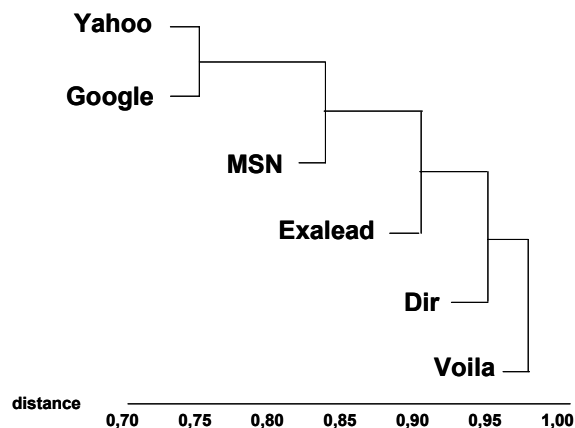


Figure 3 – Proximité des moteurs en fonction de leurs résultats communs

## Pages hors-thème

La proportion de pages hors thème est particulièrement importante, puisqu'elle va de 21,7 % (Yahoo) à 49,1 % (Voilà). Le tableau 4 récapitule les résultats obtenus.

	Dir	Exalead	Google	MSN	Voilà	Yahoo
Toutes positions	46,5%	34,5%	24,8%	31,2%	49,1%	21,7%
Position 1	43,3%	29,7%	16,2%	29,0%	72,3%	17,9%

Tableau 4 – Proportion de pages hors-thème

La situation s'améliore à peine lorsqu'on ne considère que le premier résultat retourné pour chaque requête. De façon tout à fait étonnante, les résultats de Voilà se dégradent, puisque le pourcentage de pages hors-thème en première position passe à 72,3% pour ce moteur. Il semblerait que cette augmentation soit due à la proportion importante de liens commerciaux retournés par ce moteur dans le haut du classement des résultats, souvent en rapport lointain avec la requête.

On notera que les liens commerciaux sont plus fréquemment hors thème : le tableau 5 montre une dégradation des performances allant de 3,8% (Dir) à 19,3% (Voilà).

	Dir	Exalead	Google	MSN	Voilà	Yahoo
Non commerciaux	46,2%	33,9%	24,2%	30,2%	46,1%	20,7%
Commerciaux	50,0%	41,2%	32,1%	43,8%	65,3%	29,7%
Différence	3,8%	7,3%	7,9%	13,5%	19,3%	9,0%

Tableau 5 – Liens commerciaux et pages hors-thème

## Pertinence

Les notes globales sont extrêmement basses, puisqu'aucun moteur n'atteint la note moyenne de 2,5. Les moteurs obtenant la meilleure note (2,3) sont Google et Yahoo (tableau 6 et figure 4).

La situation est légèrement meilleure si l'on ne considère que la première position : Google et Yahoo, dépassent alors très légèrement la moyenne. A nouveau, il est surprenant de constater que la note de Voilà est plus mauvaise en première position.

	Dir	Exalead	Google	MSN	Voilà	Yahoo
Toutes positions	1,4	1,8	2,3	2,0	1,2	2,3
Position 1	1,5	2,2	2,9	2,3	0,5	2,8

Tableau 6 – Pertinence perçue (note de 0 à 5)

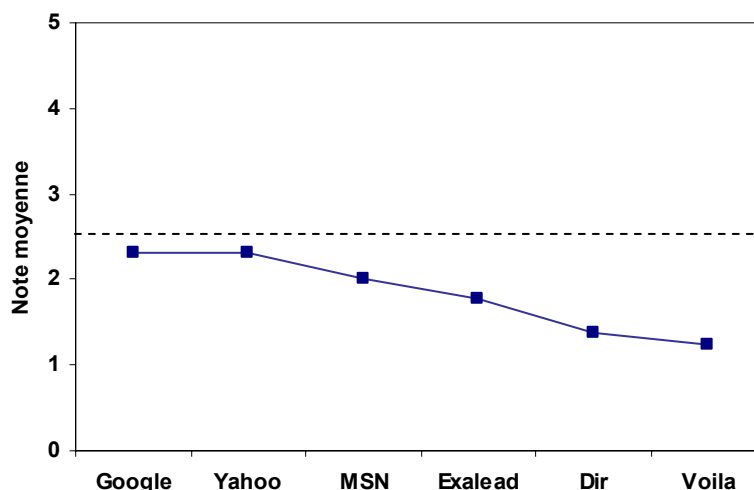


Figure 4 – Pertinence perçue

La figure 5 représente la note moyenne en fonction de la position pour chaque moteur. On constate une baisse générale de pertinence perçue en fonction de la position, sauf pour Dir et Voilà, qui atteignent leur meilleur résultat en positions 8 et 7 respectivement, ce qui laisse penser que les algorithmes de classement ne sont pas optimaux pour ces moteurs<sup>2</sup> (ou, dans le cas de Voilà, perturbés par l'interclassement de sites commerciaux).

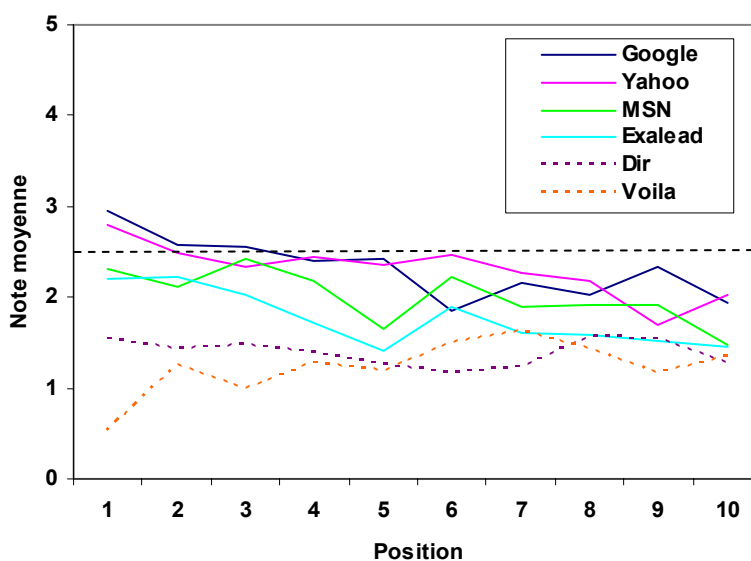


Figure 5 – Pertinence en fonction de la position

Les résultats recevant la note 0 (donc perçus comme totalement inutiles) sont extrêmement nombreux : leur proportion dépasse 50% pour deux moteurs (Dir et Voilà), et atteignent tout de même 27,7% dans le meilleur cas (Yahoo). En ce qui concerne la première position, les résultats s'améliorent quelque peu, mais le minimum reste de 16,2% (Google). Le moteur Voilà voit sa proportion de résultats notés 0 monter à 78,5% (tableau 7).

<sup>2</sup> Rappelons que Dir.com est seulement une plate-forme expérimentale. Ce moteur a mis en ligne fin janvier 2006 une nouvelle version avec des modifications importantes de l'algorithme de classement, mais celle-ci n'a pas pu être testée dans le cadre de cette étude.

	Dir	Exalead	Google	MSN	Voila	Yahoo
Toutes positions	50,9%	40,6%	28,6%	35,0%	53,1%	27,7%
Position 1	50,7%	35,9%	16,2%	34,8%	78,5%	20,9%

Tableau 7 – Proportion de résultats notés 0

A l'inverse, les résultats notés 5 (résultat excellent, satisfaisant pleinement à la question posée) toutes positions confondues sont peu nombreux. Ils atteignent au mieux 15,9% pour Google. En position 1, Yahoo émerge avec 28,4% de pages notées 5 (tableau 8).

	Dir	Exalead	Google	MSN	Voila	Yahoo
Toutes positions	9,1%	11,0%	15,9%	11,9%	5,4%	15,7%
Position 1	11,9%	17,2%	22,1%	20,3%	1,5%	28,4%

Tableau 8 – Proportion de résultats notés 5

Le croisement avec le caractère commercial des liens montre que d'une façon générale, les liens commerciaux reçoivent une note plus faible, la différence pouvant aller jusqu'à un point (Google, MSN), ce qui est important, si l'on considère que la note moyenne maximale dépasse à peine 2 (tableau 9).

	Dir	Exalead	Google	MSN	Voila	Yahoo
Non commerciaux	1,4	1,8	2,4	2,1	1,3	2,4
Commerciaux	1,0	0,9	1,4	1,1	0,6	1,5
Différence	-0,4	-0,9	-1,0	-1,0	-0,7	-0,9

Tableau 9 – Pertinence et liens commerciaux

## Discussion

Cette étude, qui est certainement loin d'être exhaustive, donne néanmoins un instantané des performances des moteurs de recherche fin 2005. Le résultat sans doute le plus frappant est le degré de satisfaction très médiocre des utilisateurs. Pour les meilleurs moteurs (Yahoo, Google), la note moyenne sur le premier écran de 10 résultats atteint à peine 2,3 sur une échelle de 0 à 5. La proportion des résultats hors thème est élevée, puisqu'elle atteint pratiquement la moitié pour certains moteurs, et le cinquième pour Yahoo qui réalise la meilleure performance sur ce critère.

La proportion de liens à caractère commercial est élevée, puisque elle varie entre 7 et 16% environ selon les moteurs. En soi, la présence de liens commerciaux n'est pas nécessairement nuisible à la qualité : sur une requête telle que « Harry Potter », faire apparaître la page Amazon où le livre peut être acheté peut être pertinent. Néanmoins, on observe, dans l'état actuel des choses, une nette dégradation des résultats en terme de pertinence perçue sur les liens commerciaux, et ce pour tous les moteurs.

Enfin, on remarquera que rien dans cette étude ne permet d'expliquer la préférence massive des internautes pour le moteur Google, puisque, globalement Google et Yahoo ont des performances à peu près équivalentes, et se détachent de leurs concurrents. Il faut donc supposer que les raisons en sont autres que des critères de pure pertinence des résultats.

## Remerciements

Cette étude a pu être réalisée grâce à l'efficacité et à l'enthousiasme des étudiants de la licence MASHS à Aix-en-Provence, auxquels j'adresse mes remerciements. Je suis également reconnaissant aux lecteurs qui m'ont fait part de nombreuses réflexions et commentaires sur des fragments de cette étude publiés sur le blog « Technologies du langage<sup>3</sup> ».

---

<sup>3</sup> <http://aixtal.blogspot.com>