http://www.up.univ-mrs.fr/veronis
http://aixtal.blogspot.com

Université de Provence

# A comparative study of six search engines

*Jean Véronis*

Version 1.0 (en) – 22 février 2006

# A comparative study of six search engines

*Jean Véronis*

## Introduction

At the end of 2005, Google was the search engine with the highest number of users in the world, including a particularly high market share in France (82% of traffic according to Xiti[1]). The reasons why a user might choose one search engine over another are complex, but while elements such as speed, ergonomics and aesthetics all come into play, the most important criterion seems to be that of the relevance of the results to the search performed – at least in the way they are perceived as relevant by the user. Yet little data is available that would allow us to compare this *perceived relevance* and, as far as we can tell, there is no recent comparative data whatsoever concerning searches in the French language. This study aims to address this lack, at least partially, by carrying out a user test at the end of 2005 in which searches are performed in French using six search engines.

Three American search engines were chosen for this study, **Google**, **Yahoo** and **MSN**, along with three French ones, **Exalead**, **Voilà** (developed by France Telecom and available on the Wanadoo web portal) and **Dir.com** from the Iliad group, which is more of an experimental platform than a commercial search engine. Other search engines, such as MozDex or AskJeeves, were not considered because they do not have a French language version (or only a beta version in the case of AskJeeves).

## Methodology

The study was carried out in December 2005 with the aid of 14 first year students from the Mathematics Applied to Human and Social Sciences (MASHS) degree course at the University of Provence (Aix-en-Provence), who acted out the role of users.

14 topics were chosen collectively, in order to reflect a broad variety of uses. The topics selected were as follows:

- **Animals**
- **Celebrities**
- **Cinema**
- **Current affairs**
- **Entertainment**
- **Health**
- **History**
- **Literature**
- **Music**
- **Nature**
- **Politics**

[1] http://www.secrets2moteurs.com/barometre2005-12.html

- **Sports**
- **The supernatural**
- **Travel**

Each topic was attributed to a different student, who was then free to choose five search queries in French. The format (with or without quotation marks, one word or several words) was also entirely up to the student. For the Current Affairs topic, for instance, the search terms chosen were as follows:

- "abbé Pierre" célibat prêtre (*"abbé Pierre" celibacy priests*)
- chaîne télévision jeunesse TNT (*television channel youth TNT*)
- "greffe visage" (*"face transplant"*)
- "grippe aviaire" danger Europe (*"bird flu" danger Europe*)
- Outreau acquittement (*Outreau acquittal*)

Better results might well have been obtained if the search requests were formulated differently, with better key words or a more appropriate use of quotation marks, for instance. However, the aim of the study was not to evaluate the search engines as used by experts, but by the general public: people with a reasonable level of education who were familiar with search engines. For this reason, it was important to respect the search queries exactly as they were formulated by the panel of users.

The search terms were then entered into each of the different search engines on the same day by the organiser of the experiment (Jean Véronis), setting each search engine to return only results in French and activating the SafeSearch filter. The first page of 10 results *not marked as sponsored* was archived for each request and each search engine, and then automatically stripped of all information other than the resulting URLs. In particular, all information about the search engine the results had come from was removed.

A total of 4200 URLs were collected (14 topics x 5 searches x 6 search engines x 10 results). For each search, duplicates (where the same URL was returned by two different search engines) were removed, leaving us with 3450 unique search-URL pairs. The search-URL pairs corresponding to each topic were then given to the student concerned, in an Excel file in which the search terms and the URL appeared in consecutive columns (with the URL appearing as a clickable link to the corresponding site). The student had to evaluate the document indicated by the URL, without knowing which search engine it came from, and report the following information in supplementary columns:

- **Dead link** (1 if the site does not respond, otherwise 0)
- **Pornographic link** (1 if the link points to a pornographic site, otherwise 0)
- **Topic** (regardless of the quality of the information, 1 if the document is on-topic, otherwise 0)
- **Commercial site** (1 if the link points to an e-commerce site, otherwise 0)
- **Relevance** (grade from 0 to 5, with 0 being a that is totally useless or off-topic, and 5 being a document that provides a perfect answer to the question posed).

The task was to be completed within one week (from the 12th to the 18th of December).

## Dead links

Some links returned by search engines cannot be accessed when the user tries to visit the page or open the document; hereafter these are called "dead links". There are many possible reasons for this: the page may have disappeared between the time it was indexed and the time the user performs his or her request, or a temporary problem may be preventing access (if the server is down, for instance). The proportion of dead links can vary depending on when the request is performed, so we checked on three different occasions: each user noted whether or not the link was "live" when attempting to visit the page manually; two further automated requests (several days apart) were carried out for all the URLs, and the error codes returned (HTTP 4xx and 5xx) were archived. The results are shown in Table 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% |

*Table 1 – Proportion of dead links*

The proportion of dead links is higher when checked manually: this is explained in part by the fact that, in the case of failure, the automatic procedure would make up to three attempts to reach the page at intervals several minutes apart, and in part by the fact that a certain number of servers do not return a 404 error code ("Page not found") when a page no longer exits, but rather a normal HTML page containing an ad hoc message that can only be interpreted as an error by a human reader.

We can also see the great difference between the results obtained automatically (lines Auto1 and Auto2), yet the conditions were strictly identical. A detailed analysis of the results reveals that for some unknown reason, the site *www.amazon.fr* returned an error code during the Auto1 experiment. This is one of the most common sites returned for searches on Google and Yahoo, so this issue had a dramatic impact on the results: of the 26 errors counted for Google in Auto1, 17 were due solely to the site *www.amazon.fr*, and for Yahoo this site was responsible for 23 errors out of 33.

For the remainder of the study, only those links that were active during the manual phase have been taken into account.

## Pornographic links

It is a well-known fact that pornographic links manage to creep into the results of non-pornographic searches, since the ingeniousness of the referencers allows them to rise artificially in the ranking using techniques learned from spam. The situation was more critical in the past, but all search engines now offer a SafeSearch function that cleans up the results. This function seems to be particularly effective, since for all the URLs returned only two (one from Voilà, the other from MSN) pointed to pornographic sites (which, obviously, had no relation to the searches).

## Commercial links

Commercial links are considered to be those links that point to sites offering online sales or transactions, yet appear among normal links *without being marked as sponsored*. The

proportion of such links varies greatly, in some cases doubling from one search engine to another (Table 2).

| | Dir | Exalead | Google | MSN | Voila | Yahoo |
|---|---|---|---|---|---|---|
| All positions | 8.3% | 8.0% | 7.7% | 7.1% | 15.6% | 10.9% |
| Top position | 9.0% | 9.4% | 2.9% | 10.1% | 32.3% | 10.4% |

*Table 2 – Proportion of commercial links*

If we only look at the first result returned (this result is especially important, since it is the most-clicked link), we can see that different search engines have opposing strategies. Dir, Exalead and Yahoo show no particular difference. The proportion of commercial links in the top position increases for MSN and (greatly, since it doubles) for Voilà. On the other hand, the proportion falls considerably for Google.

Of all the commercial sites returned, only three companies appear at least 10 times in one of the search engines: Amazon, eBay and PriceMinister. Their association with the different search engines is interesting to look at (in this order). Google and Yahoo are closely associated with Amazon, while Voilà prefers eBay and PriceMinister. The other search engines seem to have no particular affinities with any merchant sites. Overall, MSN returns the fewest links to commercial sites with 7.1%.

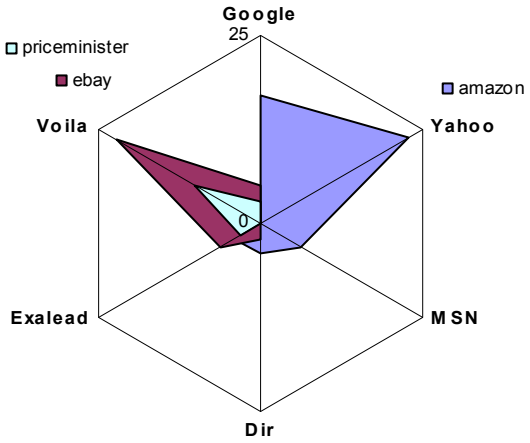


*Figure 1 – Preferential associations between search engines and commercial sites*

## Result overlap

The degree of result overlap between search engines is extremely low, varying (depending on the pairs examined) between 2.9% (Dir/Voilà) and 25.1% (Google/Yahoo) (table 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% | -- |

*Table 3 – Common results by search engine pair*

Of all the unique URLs returned by the 6 search engines, less than 10% are returned by at least two search engines (figure 2).

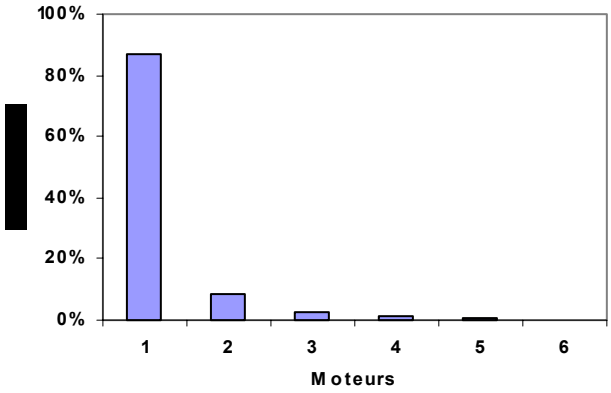


*Figure 2 – Unique URLS and the number of search engines that return them*

The proximity between different search engines can be calculated and represented graphically, based on the number of results they share, using a technique known as ascending hierarchical classification (figure 3). We can see that the closest search engines are Google and Yahoo.

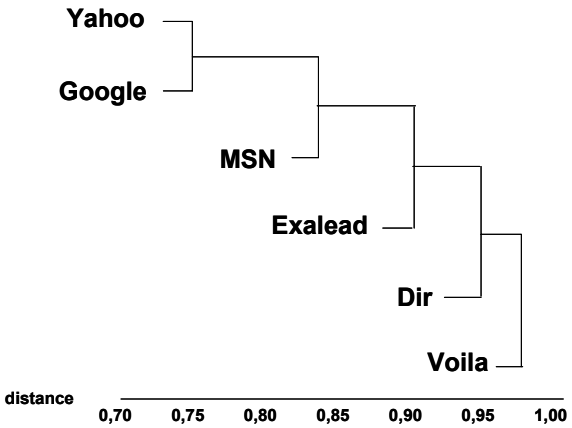


*Figure 3 – Proximity of search engines based on results they have in common*

## Off-topic pages

The proportion of off-topic pages is particularly high, running from 21.7% (Yahoo) to 49.1% (Voilà). Table 4 shows the results obtained.

| | Dir | Exalead | Google | MSN | Voila | Yahoo |
|---|---|---|---|---|---|---|
| All positions | 46.5% | 34.5% | 24.8% | 31.2% | 49.1% | 21.7% |
| Top position | 43.3% | 29.7% | 16.2% | 29.0% | 72.3% | 17.9% |

*Table 4 – Proportion of off-topic pages*

The situation scarcely improves when we only look at the first result returned for each request. Astonishingly, the results for Voilà actually get worse, with the percentage of off-topic pages that appear in the top position rising to 72.3% for this search engine. This would seem to be due to the high proportion of commercial links returned by this search engine at the top end of its ranking of results, often only vaguely related to the search in question.

We can see that commercial links are more frequently off-topic. Table 5 shows a degradation of performance going from 3.8% (Dir) to 19.3% (Voilà).

| | Dir | Exalead | Google | MSN | Voila | Yahoo |
|---|---|---|---|---|---|---|
| Non-commercial | 46.2% | 33.9% | 24.2% | 30.2% | 46.1% | 20.7% |
| Commercial | 50.0% | 41.2% | 32.1% | 43.8% | 65.3% | 29.7% |
| Difference | 3.8% | 7.3% | 7.9% | 13.5% | 19.3% | 9.0% |

*Table 5 – Commercial links and off-topic pages*

## Relevance

The overall grades are extremely low, with no search engine achieving a “pass” grade of 2.5. The search engines with the best grade (2.3) are Google and Yahoo (table 6 and figure 4).

The situation improves slightly if we only look at the top position: Google and Yahoo both “pass”, but only just. Once again, it is surprising to note that Voilà’s grade is actually worse when only the top result is considered.

| | Dir | Exalead | Google | MSN | Voila | Yahoo |
|---|---|---|---|---|---|---|
| All positions | 1.4 | 1.8 | 2.3 | 2.0 | 1.2 | 2.3 |
| Top position | 1.5 | 2.2 | 2.9 | 2.3 | 0.5 | 2.8 |

*Table 6 – Perceived relevance (grade from 0 to 5)*

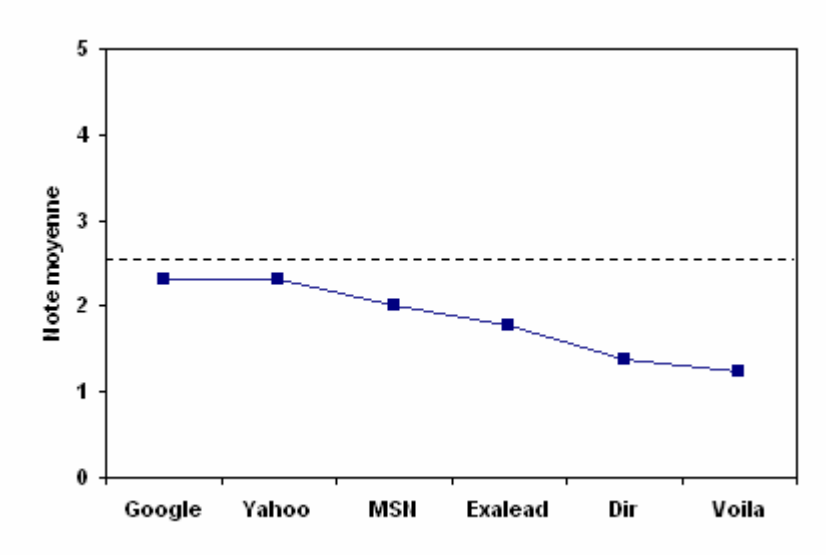


*Figure 4 – Perceived relevance*

Figure 5 shows the average grade according to the position of the result in the result screen for each search engine. An overall decrease in perceived relevance can be seen according to the position, except for Dir and Voilà, which achieve their best result in positions 8 and 7 respectively, leading one to conclude that the ranking algorithms for these search engines are not optimal[2] (or, in the case of Voilà, the results are skewed by the insertion of commercial sites).

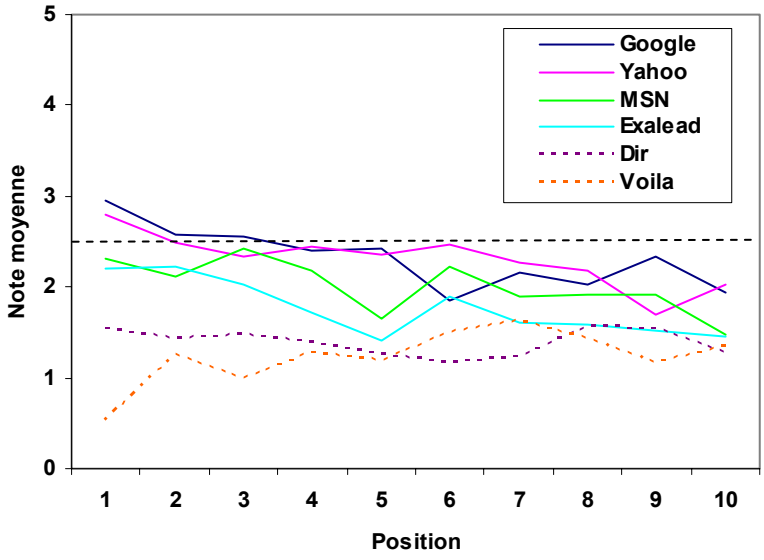


*Figure 5 – Relevance according to position*

Results receiving a grade of 0 (perceived as totally useless) are extremely numerous: their proportion rises above 50% for two search engines (Dir and Voilà) and reaches 27.7% in even the best case (Yahoo). For the top position, the results improve slightly, but the minimum remains 16.2% (Google). The Voilà search engine sees its proportion of results with a grade of 0 rise to 78.5% (table 7).

[2] Let's not forget that Dir.com is only an experimental platform. This search engine put a new version online at the end of January 2006 which featured important modifications to its ranking algorithm, but this new version could not be tested within the framework of this study.

| | Dir | Exalead | Google | MSN | Voila | Yahoo |
|---|---|---|---|---|---|---|
| All positions | 50.9% | 40.6% | 28.6% | 35.0% | 53.1% | 27.7% |
| Top position | 50.7% | 35.9% | 16.2% | 34.8% | 78.5% | 20.9% |

*Table 7 – Proportion of results with a grade of 0*

Conversely, results with a grade of 5 (excellent result, providing a perfect answer to the question posed) for all positions are few. At best, they reach 15.9% for Google. If we look only at the results appearing in the top position, Yahoo emerges as the leader with 28.4% of its pages being given a grade of 5 (table 8).

| | Dir | Exalead | Google | MSN | Voila | Yahoo |
|---|---|---|---|---|---|---|
| All positions | 9.1% | 11.0% | 15.9% | 11.9% | 5.4% | 15.7% |
| Top position | 11.9% | 17.2% | 22.1% | 20.3% | 1.5% | 28.4% |

*Table 8 – Proportion of results with a grade of 5*

The analysis of interaction with the commercial nature of the links shows that, in general, commercial links receive a lower grade; up to one point less in some cases (Google, MSN). This figure seems particularly high when we consider that the best grade is not much above 2 (table 9).

| | Dir | Exalead | Google | MSN | Voila | Yahoo |
|---|---|---|---|---|---|---|
| Non-commercial | 1.4 | 1.8 | 2.4 | 2.1 | 1.3 | 2.4 |
| Commercial | 1.0 | 0.9 | 1.4 | 1.1 | 0.6 | 1.5 |
| Difference | -0.4 | -0.9 | -1.0 | -1.0 | -0.7 | -0.9 |

*Table 9 – Relevance and commercial links*

## Discussion

This study, while certainly far from exhaustive, does provide a snapshot of the performance of different search engines at the end of 2005. Undoubtedly the most striking result is the very low level of user satisfaction. For the best search engines (Yahoo, Google), the average grade on the first screen of 10 results barely reaches 2.3 on a scale of 0 to 5. The proportion of off-topic results is high, since it accounts for practically half of all results for some search engines and one fifth for Yahoo, the best-performing search engine under this criterion.

The proportion of commercial links is high, varying between 7 and 16% depending on the search engine. In itself, the presence of commercial links does not necessarily have a negative impact on quality: for a search like “Harry Potter”, returning the page on Amazon where the book can be purchased may be relevant. However, as things stand, we can see a clear degradation of the results in terms of perceived relevance for commercial links, for all search engines alike.

Finally, there is nothing in this study to explain why web users seem to greatly prefer the Google search engine, since overall the performance of Google and Yahoo is more or less equivalent, and ahead of their competitors. We must therefore suppose that the reasons go beyond the criteria of relevance of results.

## Thanks

This study was carried out thanks to the hard work and enthusiasm of the students on the MASHS degree course at Aix-en-Provence, to whom I address my thanks. I am also grateful to the readers who have shared many thoughts and comments on the fragments of this study published on the blog "Technologies du langage[3]".

[3] http://aixtal.blogspot.com