Discovering Europeana Users' Search Behavior

Diego Ceccarelli, Sergiu Gordea, Claudio Lucchese,
Franco Maria Nardini, Raffaele Perego, Gabriele Tolomei

Europeana is a strategic project funded by the European Commission with the goal of making Europe's cultural and scientific heritage accessible to the public. ASSETS is a two-year Best Practice Network co-funded by the CIP PSP Programme to improve performance, accessibility and usability of the Europeana search engine. Here we present a characterization of the Europeana logs by showing statistics on common behavioral patterns of the Europeana users.

The strong inclination for culture and beauty in Europe created a rich amount of artifacts starting from antiquity up to nowadays. That cultural strength is recognized by all people in the world making Europe the preferred destination for half of the tourists in the world.

Europeana [1] is a long-term project funded by the European Commission with the goal of making Europe's cultural and scientific heritage accessible to the public. Since 2008, about 1,500 institutions have contributed to Europeana, enabling people to explore Europe's museums, libraries and archives. This huge amount of multilingual and multimedia data is made available through the Europeana Portal, a search engine that allows users to explore such content by means of textual queries.

Figure 1: Most popular queries submitted by Italian users. Size is proportional to frequency.

Due to the increasing amount of information published, the access to the description of a specific masterpiece becomes more and more difficult, in particular when the user is not able to formulate a sufficiently discriminative query. For example, if we search today for general terms like "*renaissance*" or "*art nouveau*" we will obtain more than 10,000 results. Furthermore, if we search for the term "*Gioconda*" we will find a couple of hundred of items, while the query "*Mona Lisa*, *Da Vinci*" provides us twenty images of the well known painting. This example shows how important is to have a good translation of users' information needs in textual queries when looking for very specific information on the web by using a search engine like Europeana. This is even more challenging since indexed documents are cross-domain, multi-lingual, and multi-cultural.

In order to improve the Europeana users' search experience, the ASSETS project [2] has the overall goal of enhancing the performances of the Europeana search engine. One of the most important

resources for enhancing users search experience in large information spaces is the exploitation of the information stored in query logs. The knowledge extracted from query logs can be fruitfully exploited for enhancing both efficiency (i.e., response time, throughput) and efficacy (i.e., quality of results) of information retrieval platforms.

The Europeana logs contain the behavior of the users interacting with the portal. We present here a preliminary characterization of this behavior, and a comparison with Web users' one. Figure 2 shows the frequency distribution of submitted queries. As expected, the popularity of the queries follows a power-law distribution ($n(x) \propto k \cdot x^{-\alpha}$), where x is the popularity rank. The best fitting α parameter is $\alpha = 0.86$, which gives a hint about the skew of the frequency distribution. The larger α the larger is the portion of the log covered by the top frequent queries. The same analysis conducted on query logs coming from commercial Web search engines shows larger values of α (2.4 and 1.84 respectively from a Excite and a Yahoo! query log).

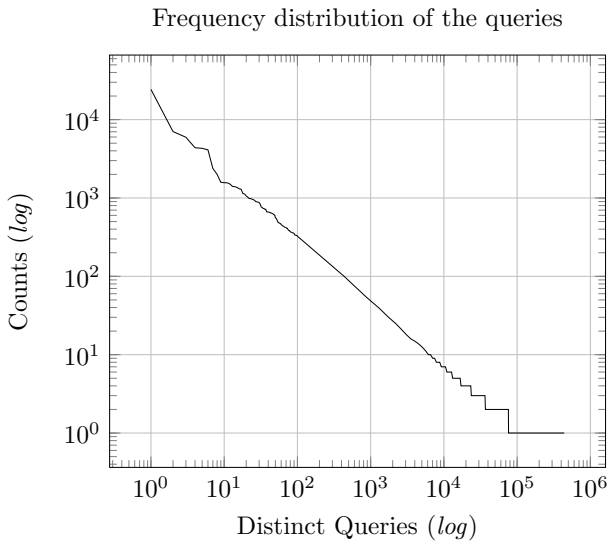

Figure 2: Frequency distribution of queries.

Such small value of α means that the most popular queries submitted to Europeana do not account for a significantly large portion of the query log. This might be explained by considering the characterizing features of Europeana. Indeed, since Europeana is strongly focused on the specific context of cultural heritage, its users are likely to be more skilled in linguistics and therefore they tend to use a more diverse vocabulary. In addition, we found that the average length of queries is 1.86 terms, which is again lower than the typical value observed in Web search engine logs. We can argue that the Europeana users use a richer vocabulary, with discriminative queries made of specific domain terms.

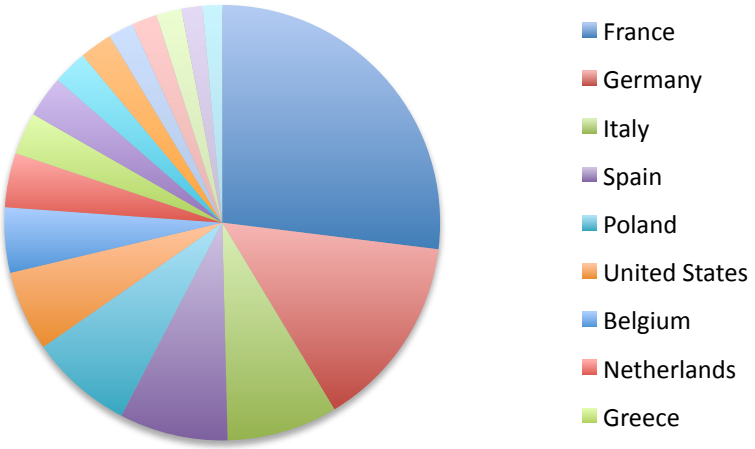

Figure 3: Distribution of the queries over the countries.

Figure 3 shows the distribution of the queries grouped by country. France, Germany, and Italy are the three major countries accounting for about the 50% of the total traffic of the Europeana portal.

Furthermore, Figure 4 reports the number of queries submitted per day. We observe a periodic behavior on a weekly basis, with a number of peaks probably related to some Europeana dissemination or advertisement activities. For example, we observe several peaks between the 18th and the 22th November, probably due to the fact that, in those days, Europeana announced the indexing of new collections and the accounting of 14 million documents.

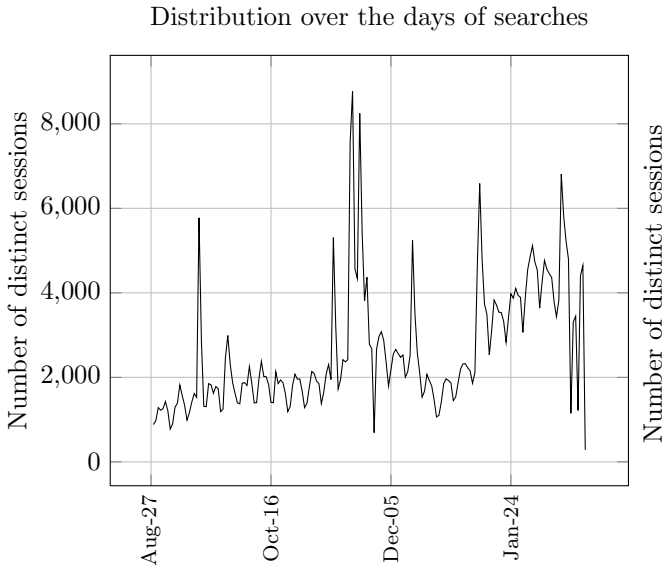

Figure 4: Distribution of the searches over the days.

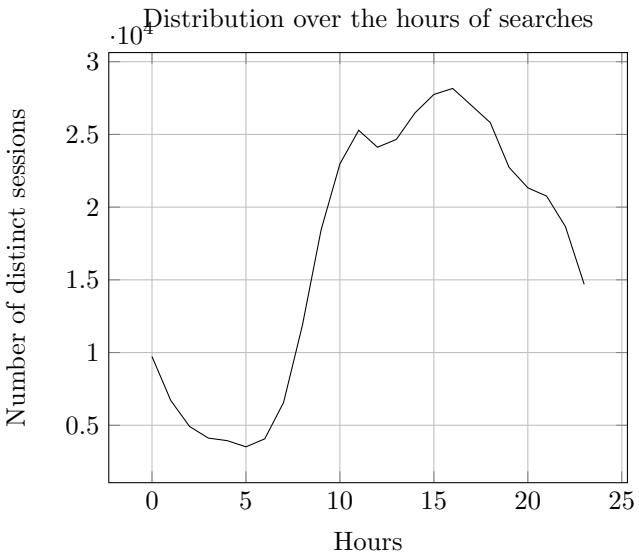

Figure 5: Distribution of the searches over the hours.

Figure 5 shows the load on the Europeana portal on a hourly basis. We observe a particular trend. The peak of load on the Europeana portal is in the afternoon, between 15 and 17. This is slightly different from commercial Web search engines where the peak is reached in the evening, between the 19 and the 23. A possible explanation of this phenomenon could be that the Europeana portal is intensively used by people working in the cultural heritage field and thus, mainly accessed during working hours. From the other side, a commercial Web search engine is used by a wider range of users looking for the most disparate information needs and using it through all the day.

Links

[1] http://www.europeana.eu/portal/
[2] http://www.assets4europeana.eu/

Contacts

Raffaele Perego – ISTI–CNR
Tel: +39–050–3152993
E-mail: raffaele.perego@isti.cnr.it