Improving Europeana Search Experience Using Query Logs

Diego Ceccarelli[1,2], Sergiu Gordea[3], Claudio Lucchese[1],
Franco Maria Nardini[1], and Gabriele Tolomei[1,4]

[1] ISTI–CNR, Pisa, Italy
{name.surname}@isti.cnr.it
[2] Dipartimento di Informatica, Università di Pisa, Italy
[3] AIT Austrian Institute of Technology GmbH, Wien, Austria
{name.surname}@ait.ac.at
[4] Università Ca' Foscari, Venezia, Italy

Abstract. Europeana 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 1500 institutions have contributed to Europeana, enabling people to explore the digital resources of Europe's museums, libraries and archives. The huge amount of collected multi-lingual multi-media data is made available today through the Europeana portal, a search engine allowing users to explore such content through textual queries. One of the most important techniques for enhancing users search experience in large information spaces, is the exploitation of the knowledge contained in query logs. In this paper we present a characterization of the Europeana query log, showing statistics on common behavioral patterns of the Europeana users. Our analysis highlights some significative differences between the Europeana query log and the historical data collected by general purpose Web Search Engine logs. In particular, we find out that both query and search session distributions show different behaviors. Finally, we use this information for designing a query recommendation technique having the goal of enhancing the functionality of the Europeana portal.

1 Introduction

The strong inclination for culture and beauty in Europe created invaluable artifacts starting from antiquity up to nowadays. That cultural strength is recognized by all people in the world and makes Europe the destination for a half of the international tourists[5]. More than 220 million people visit the European countries yearly for spending their holidays.

The European Commission is aware about the value of this cultural heritage and decided to make it more accessible to the public by supporting digitization of the cultural heritage and by financing the Europeana group projects. The first prototype of the Europeana Portal[6] was launched in autumn 2008 and contains by now about 15 million items.

Due to increasing amount of information published within the portal, the access to the description of a specific masterpiece becomes each day a more time consuming task, when the user is not able to create a very restrictive query. For example, if we search today in Europeana for general terms like *renaissance* or *art nouveau* we will find more than 10,000 results. If we search for the term *Gioconda* we find a couple of hundred of items, and if we search for *Mona Lisa, Da Vinci* we get 20 images of the well known painting. These examples show how important is to use good queries when looking for very particular information on the web by using a search engine like Europeana. This is a challenging task, given the fact that the document base is cross-domain, multi-lingual and multi-cultural.

Search query recommendation techniques [3, 12] are commonly used in web search engines to help users to refine their queries. These technologies analyze the user behavior by mining the system logs in order to find the correlation between what the user's information need (visited pages), what the user is searching for (query terms) and the content and structure of the information pool (search index).

In this paper we present the work carried out by now in the ASSETS[7] project with the goal of implementing a query recommendation module for Europeana port. We focus our attention on the analysis of the user behavior and particularities of the information pool.

The rest of the paper is organized as follows: Section 2 introduces related work, while Section 3 discusses the main results coming from the analysis of the Europeana query log. Furthermore, Section 4 presents a novel query recommendation technique based on the knowledge extracted from query logs and, finally, Section 5 presents some conclusions and outlines possible future work.

[5] `http://www.unwto.org/facts/eng/pdf/highlights/UNWTO_Highlights10_en_HR.pdf`

[6] `http://www.europeana.eu/portal`

[7] `http://www.assets4europeana.eu/`

2 Related Work

Some important efforts have been spent in the past to study how people interact with IR systems[8] by analyzing the historical search data of their users [11, 18, 21, 9]. Similarly, there have been several works about the understanding of user search behaviors on large scale IR systems, i.e., Web Search Engines (WSEs), still by analyzing the stream of past queries collected by query logs. Although the nature of query logs coming from large scale WSEs is different with respect to small scale IR systems, many of the benefits coming from the analysis of the former could also be useful for improving the latter.

Typical statistics that can be drawn from query logs are: query popularity, term popularity, average query length, distance between repetitions of queries or terms, etc. To this end, the very first contribution in analyzing a WSE query log comes from Silverstein *et al.* [19]. Here, the authors propose an exhaustive analysis by examining a large query log of the AltaVista search engine containing about a billion queries submitted in a period of 42 days by approximately 285 million users. The study shows some interesting results including the analysis of the query sessions for each user, and the correlation among the terms of the queries. Similarly to other works, authors show that the majority of the users (in this case about 85%) visit the first page of results only. They also show that 77% of the users' sessions end up just after the first query.

Lempel and Moran [13] and Fagni *et al.* [8] study the content of another publicly available AltaVista log. This log refers to the summer of 2001 and consists of 7,175,648 issued queries, i.e., about three order of magnitude less queries than the log used by Silverstein *et al.*. Furthermore, no information about the number of logged users is released however, although this second log is smaller than the first one, it still represents a good picture of search engine users.

On average, queries issued to WSEs are quite short. Indeed, the average length of a query in the 1998 Excite log is 2.35 terms. Moreover, less than 4% of the queries contains more than 6 terms. In the case of the first AltaVista log, the average query length is slightly greater: 2.55. These numbers are deeply different compared with classical IR systems where the length of a query ranges from 7 to 15 terms. A possible explanation of this phenomenon could be that the Web is a medium used by people that strongly differ from each other in terms of age, race, culture, etc. who look for disparate information. On the other hand, traditional IR systems are instead exploited by professionals and librarian, i.e., "skilled" users, which are able to look for very focused information by precisely formulating their information needs.

Moreover, a very useful information that could be extracted from query logs are *search sessions*, i.e., sets of user actions recorded in a limited period of time that hopefully refer to the same *information need*.

Several works have addressed the search session identification problem from raw streams of queries available in user logs. To this end, Silverstein *et al.* [19]

[8] The IR systems whose studies here we refer to do not directly deal with Web users.

firstly define a concept of *session* as follows: two consecutive queries are part of the same session if they are issued at most within a 5-minutes time window. According to this definition, they found that the average number of queries per session in the data they analyzed was 2.02. Similarly to this approach, He and Göker [10] use different timeouts to split user sessions of the Excite query log, ranging from 1 to 50 minutes.

Radlinski and Joachims [16] observe that users often perform a sequence of queries with a similar information need, and they refer to those sequences of reformulated queries as *query chains*. Their paper presents a method for automatically detecting query chains in query and click-through logs using 30 minutes threshold for determining if two consecutive queries belong to the same search session.

More recently, novel heuristics have been proposed for effectively discovering search session boundaries in query logs. Boldi *et al.* [5] introduce the *Query Flow Graph* as a model for representing data collected in WSE query logs. They exploited this model for segmenting the query stream into sets of related information-seeking queries, leveraging on an instance of the Asymmetric Traveling Salesman Problem (ATSP). Jones and Klinkner [12] argue that within a user's query stream it is possible to recognize particular hierarchical units, i.e., *search missions*, which are in turn subdivided into disjoint *search goals*. Given a manually generated ground-truth, the authors investigate how to *learn* a suitable binary classifier, which is aimed to precisely detect whether two queries belong to the same session or not.

Finally, Lucchese *et al.* [14] devise effective techniques for identifying *task-based sessions*, i.e. sets of possibly non contiguous queries issued by the user of a Web search engine for carrying out a given *task*. Furthermore, authors formally define the *Task-based Session Discovery Problem* (TSDP) as the problem of best approximating a a *ground-truth* of manually annotated tasks, and propose several variants of well-known clustering algorithms, as well as a novel efficient heuristic algorithm, specifically tuned for solving the TSDP. Results show that it performs better than state-of-the-art approaches, because it effectively takes into account the *multi-tasking* behavior of users.

3 The Europeana Query Log

A query log keeps track of historical information regarding past interactions between users and the retrieval system. It usually contains tuples $\langle q_i, u_i, t_i, V_i, C_i \rangle$ where for each submitted query q_i the following information is available: i) the anonymized identifier of the user u_i, ii) the submission timestamp t_i, iii) the set V_i of documents returned by the search engine, and iv) the set C_i of documents clicked by u_i. Therefore, a query log records both the activities conducted by users, e.g. the submitted queries, and an implicit feedback on the quality of the retrieval system, e.g. the clicks.

In this work, we consider a query log coming from Europeana portal[9], relative to the time interval ranging from August 27, 2010 to February, 24, 2011. This is a six months worth of users' interactions, resulting in 1,382,069 distinct queries issued by users from 180 countries (3,024,162 is the total number of queries). We preprocessed the entire query log in order to remove noise (e.g., stream of queries submitted by software robots instead of humans).

It is worth noticing that 1,059,470 queries (i.e., 35% out of the total) also contain a *filter* (e.g., YEAR:1840). Users can filter results by *type*, *year* or *provider* simply by clicking on a button, so it is reasonable that they try to refine retrieved results by applying a filter, whenever they are not satisfied. Furthermore, we find that users prefer filtering results by type, i.e., images, texts, videos or sounds. Indeed, we measure that 20% of the submitted queries contains a filter by type. This is an additional proof of the skillfulness of Europeana users and their willingness to exploit non trivial search tools to find their desired contents. This also means that advanced search aids, such as query recommendation, would be surely exploited.

Similarly to Web query log analysis [19], we discuss two aspects of the analysis task: i) an analysis on the *query set* (e.g., average query length, query distribution, etc.) and ii) a higher level analysis of *search sessions*, i.e., sequences of queries issued by users for satisfying specific information needs.

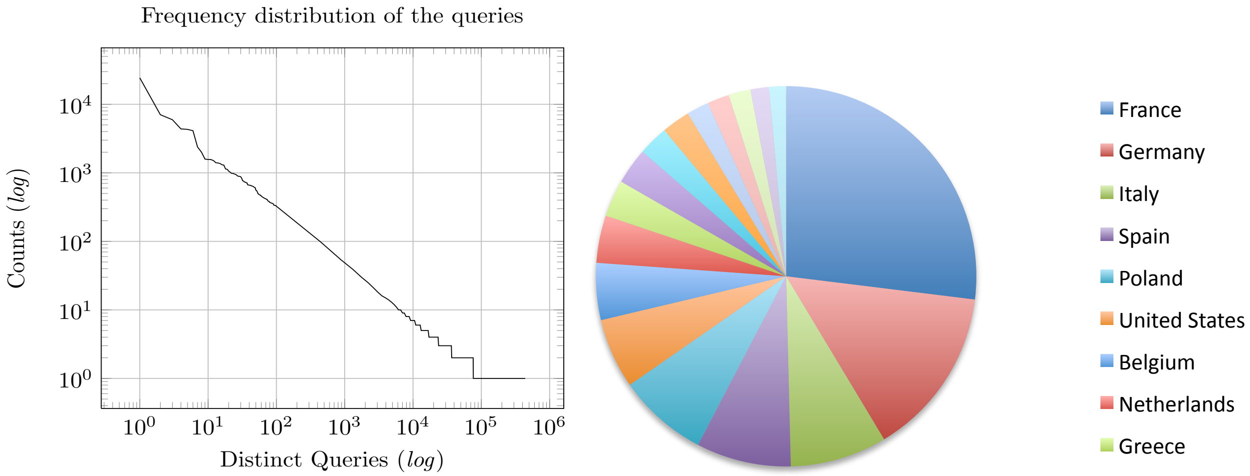
3.1 Query Analysis

First we analyzed the load distribution on the Europeana portal. An interesting analysis can be done on the queries themselves. Figure 1(a) shows the frequency distribution of queries. As expected, the popularity of the queries follows a power-law distribution ($p(x) \propto kx^{-\alpha}$), where x is the popularity rank. The best fitting α parameter is $\alpha = 0.86$, which gives a hint about the skewness of the frequency distribution. The larger α the larger is the portion of the log covered by the top frequent queries. Both [15] and [2] report a much larger α value of 2.4 and 1.84 respectively from a Excite and a Yahoo! query log. 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. The might be explained by looking at and comparing the main characteristics both of Europeana and Web search engines users. Indeed, since Europeana is strongly focussed on the specific context of cultural heritage, its users are likely to be more skilled 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 a smaller value than the typical value observed in Web search engine logs. We can argue that the Europeana user has a more rich vocabulary, with discriminative queries made of specific terms.

Figure 1(b) 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 queries submitted to the Europeana portal.

[9] http://www.europeana.com/portal/



(a)

(b)

Fig. 1. Frequency distribution of queries (a) and distribution of the queries over the countries (b).

Figure 2(a) reports the number of queries submitted per day. We observe a periodic behavior over a week 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 to have reached a threshold of 14 million of indexed documents[10].

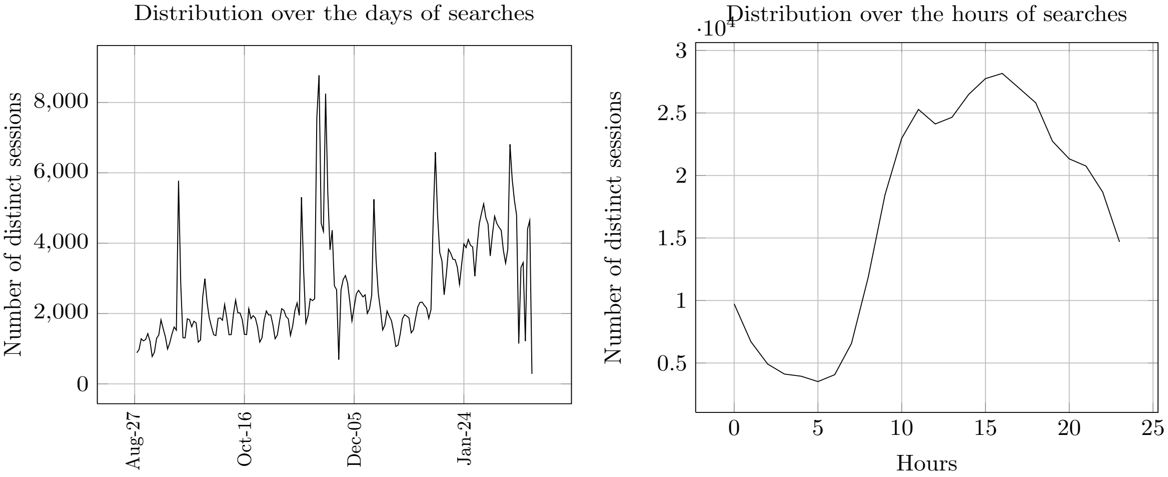
Figure 2(b) shows the load on the Europeana portal on a per hour basis. We observe a particular trend. The peak of load on the Europeana portal is in the afternoon, between 15 and 17. It is different from commercial Web search engines where the peak is reached in the evening, between the 19 and the 23 [4]. A possible explanation of this phenomenon could be that the Europeana portal is mainly used by people working in the 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.

3.2 Session Analysis

To fully understand user behavior, it is important to analyze also the sequence of queries she submits. Indeed, every query can be considered as an improvement of the previous done by the user to better specify her information need.

Several techniques have been developed to split the queries submitted by a single user into a set of sessions [5, 12, 14]. We adopted a very simple approach

[10] `http://www.sofiaecho.com/2010/11/18/995971_europes-cultural-heritage-online`



(a)

(b)

Fig. 2. Distribution of the searches over the days (a) and over the hours (b).

which has proved to be fairly effective [19]. We exploit a 5 minutes inactivity time threshold in order to split the stream of queries coming from each user. We assume that if two consecutive queries coming from the same user are submitted within five minutes they belong to the same logical session, whereas if the time distance between the queries is greater, the two queries belong to two different interactions with the retrieval system.

By exploiting the above time threshold, we are able to devise 404,237 sessions in the Europeana query log. On average a session lasts about 276 sec, i.e., less than 5 minutes, meaning that, under our assumption, Europeana's users complete a search activity for satisfying an information need within 5 minutes. The average session length, i.e., the average number of queries within a session, is 7.48 queries. This number of queries is an interesting evidence that the user is engaged by the Europeana portal, and she is willing to submit many queries to find the desired result.

Moreover, we distinguish between *successful* and *unsuccessful* sessions. According to [6], a session is supposed to be successful if its *last* query has got a click associated. To this end, we find 182,280 occurrences of successful sessions in the Europeana query log, that is about 45% of the total. We notice that in [6] it was observed a much larger fraction of successful sessions, about 65%.

Figure 3 shows the distributions of session lengths, both for successful and unsuccessful sessions. On the x-axis the number of queries within a session is plotted, while on the y-axis the frequencies, i.e., how many sessions to contain a specific number of queries are reported. We expect successful sessions contain on average less queries than unsuccessful ones, due to the ability of the retrieval system to return early high quality results in successful session. The fact that the

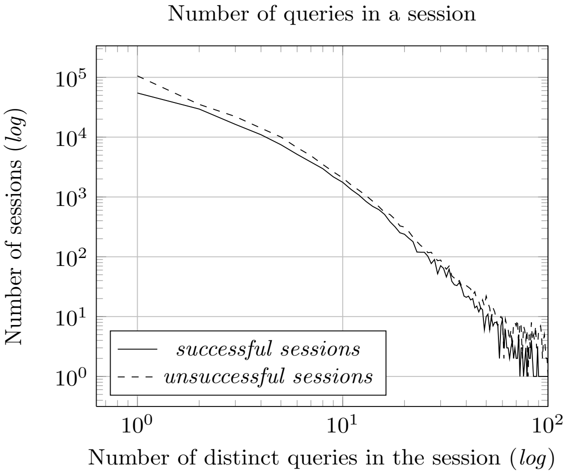


Fig. 3. Distribution of successful and unsuccessful sessions lengths (in queries).

session length distributions are very similar, suggests that high quality results are not in the top pages, and that the Europeana ranking can be improved in order to present interesting results to the user earlier, thus reducing the successful session length with a general improvement of the user experience.

Table 3.2 shows some statistics extracted both from the analysis of the Europeana query log as well as from general purpose Web Search Engines historical search data.

	Europeana	Web Search Engines
avg. query terms	1.86	2.35 [15], 2.55 [19]
query distribution (i.e., power-law's α)	0.86	2.40 [15], 1.84 [2]
avg. queries per session	7.48	2.02 [19]
% of *successful* sessions	45	65 [7]

Table 1. Europeana vs. Web Search Engines: a comparison on query log statistics.

4 A Query Recommender System for Europeana

The analysis conducted in the previous section shows that the search experience of the user interacting with Europeana could be improved. To this extent, we now introduce an application exploiting the knowledge extracted from the Europeana query log aiming at enhancing the interaction of users by suggesting a list of possible interesting queries.

A search session is an interactive process where users continuously refines their search query in order to better specify their information need. Sometimes, the successful query is not known in advance, but users might adopt concepts and terminologies also on the basis of the results pages visited. Query recommendation is a very popular technique aiming at proposing successful queries as early as possible. The approach described below, exploits successful queries from successful session to *recommend queries that allowed "similar" users, i.e., users which in the past followed a similar search process, to successfully find the information they were looking for*, and it is able to catch non trivial semantic relationships among queries.

We adopt the *Search Shortcuts* (SS) model proposed in [3] and its terminology. The SS has a clear and sound formulation as the problem of recommending queries that can reduce the search session length, i.e., leading users to relevant results as early as possible.

Let $\mathcal{U}$ be the set of users of a WSE whose activities are recorded in a query log QL, and $\mathcal{Q}$ be the set of queries in QL. We suppose QL is preprocessed by using some session splitting method (e.g. [12, 14]) in order to extract query *sessions*, i.e., sequences of queries which are related to the same user search task. Formally, we denote by $\mathcal{S}$ the set of all sessions in QL, and σ^u a session issued by user u. Moreover, let us denote with σ_i^u the i-th query of σ^u. For a session σ^u of length n its *final query* is the query σ_n^u, i.e. the last query issued by u in the session. To simplify the notation, in the following we will drop the superscript u whenever the user u is clear from the context.

As previously introduced, we say that a session σ is *successful* if and only if the user has clicked on at least one link shown in the result page returned by the WSE for the final query σ_n, *unsuccessful* otherwise.

We define a novel algorithm that aims to generate suggestions containing only those queries appearing as final in successful sessions. The goal is to suggest queries having a high potentiality of being useful for people to reach their initial goal. In our view, suggesting queries appearing as final in successful sessions is a good strategy to accomplish this task.

The SS algorithm works by efficiently computing similarities between partial user sessions (the one currently performed) and historical successful sessions recorded in a query log. Final queries of most similar successful sessions are suggested to users as search shortcuts.

Let σ' be the current session performed by the user, and let us consider the sequence τ of the concatenation of all terms with possible repetitions appearing in $\sigma'_{t|}$, i.e. the head of length t of session σ'. Then, we compute the value of a scoring function $\delta\left(\tau, \sigma^s\right)$, which for each successful session measures the similarity between its queries and the set of terms τ. Intuitively, this similarity measures how much a previously seen session overlaps with the user need expressed so far (the concatenation of terms τ serves as a bag-of-words model of user need). Sessions are ranked according to δ scores and from the subset of the top ranked sessions we suggest their final queries. It is obvious that depending on how the function δ is chosen we may have different recommendation methods. In our particular case, we opt for δ to be the similarity computed as in the BM25 metrics [17]. The choice of an IR-like metric allows us to take much care of words that are discriminant in the context of the session to which we are comparing. BM25, and other IR-related metrics, have been designed specifically to account for that property in the context of query/documents similarity. We borrow from BM25 the same attitude to adapt to this condition. The shortcuts generation problem has been, thus, reduced to the information retrieval task of finding highly similar sessions in response to a given sequence of queries. In most cases, it is enough to use only the last submitted query to propose optimal recommendations.

The idea described above is thus translated into the following process. For each unique *final query* q_f contained in successful sessions we define what we have called a *virtual document* identified by its *title* and its *content*. The title, i.e., the identifier of the document, is exactly query string q_f. The content of the virtual document is instead composed of all the terms that have appeared in queries of all the successful sessions ending with q_f. At the end of this procedure we have a set of virtual documents, one for each distinct final query occurring in some successful sessions. Just to make things more clear, let us consider a toy example. Consider the two following successful sessions: (*dante alighieri* → *divina commedia* → *paolo e francesca*), and (*divina commedia* → *inferno canto V* → *paolo e francesca*). We create the virtual document identified by title *paolo e francesca* and whose content is the text (*dante alighieri divina commedia divina commedia inferno canto V*). As you can see the virtual document actually contains also repetitions of the same terms that are considered in the context of the BM25 metrics. All virtual documents are indexed with the preferred Information Retrieval system, and generating shortcuts for a given user session σ' is simply a matter of processing the query $\sigma'_{t|}$ over the inverted file indexing such virtual documents. We know that processing queries over inverted indexes is very fast and scalable, and these important characteristics are inherited by our query suggestion technique as well.

The other important feature of our query suggestion technique is its robustness with respect to rare and singleton queries. Singleton queries account for almost 50% of the submitted queries [20], and their presence causes the issue of the sparsity of models [1]. Since we match τ with the text obtained by concatenating all the queries in each session, we are not bound to look for previously submitted queries as in the case of other suggestion algorithms. Therefore, we can generate suggestions for rare queries of the query distribution whose terms have some context in the query log used to build the model.

5 Conclusions

In this paper we presented a part of the work carried out within the ASSETS project with the aim of improving the usability of the Europeana Portal. We place our work in the context of user-system interaction analysis for web search engines and information retrieval applications. We reused the concepts of session identification, time series analysis, query chains and task based search when analyzing the Europeana logs. To the best of our knowledge, this is first analysis of the user interaction with a cultural heritage retrieval system.

Our analysis highlights some significative differences between the Europeana query log and the historical data collected by general purpose Web Search Engine logs. In particular, we find out that both query and search session distributions show different behaviors. Such phenomenon could be explained by looking at the characteristics of Europeana users, which are typically more skilled than generic Web users and, thus, they are capable of taking advantage of the Europeana portal features to conduct more complex search sessions.

For this reason, we believe that interesting knowledge can be extracted from Europeana query log in order to build advanced assistance functionalities, such as query recommendation. In fact, we investigated the integration of a state-of-the-art algorithm into the Europeana portal. Furthermore, the specificity of the Europeana portal opens up a wide range of possible extensions to current recommendation models, taking advantage of its multi-lingual and multi-media content, and including new kinds of recommendations, e.g., recommend queries related to events or exhibitions.

As future work we intend to study how the introduction of the query recommender system changes the behavior of users interacting with the Europeana portal. Furthermore, we want to study if the sharing of the same final queries induces a sort of “clustering” of the queries composing the successful user sessions. By studying such relation which is at the basis of our technique, we could probably find ways to improve our methodology.

6 Acknowledgements

This research has been funded by the EU CIP PSP-BPN ASSETS Project. Grant Agreement no. 250527.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE 17(6), 734–749 (2005)
2. Baeza-Yates, R., Gionis, A., Junqueira, F., Murdock, V., Plachouras, V., Silvestri, F.: The impact of caching on search engines. In: Proc. SIGIR’07. pp. 183–190. ACM, New York, NY, USA (2007)
3. Baraglia, R., Cacheda, F., Carneiro, V., Fernandez, D., Formoso, V., Perego, R., Silvestri, F.: Search shortcuts: a new approach to the recommendation of queries. In: Proc. RecSys’09. ACM, New York, NY, USA (2009)
4. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O.: Hourly analysis of a very large topically categorized web query log. In: Proc. SIGIR’04. ACM Press (2004)
5. Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., Vigna, S.: The query-flow graph: model and applications. In: Proc. CIKM’08. ACM (2008)

6. Broccolo, D., Marcon, L., Nardini, F.M., Perego, R., Silvestri, F.: An efficient algorithm to generate search shortcuts. Tech. Rep. 2010-TR-017, CNR ISTI Pisa (2010)
7. Broccolo, D., Marcon, L., Nardini, F.M., Perego, R., Silvestri, F.: An efficient algorithm to generate search shortcuts. Tech. Rep. 2010-TR-017, CNR ISTI Pisa (2010)
8. Fagni, T., Perego, R., Silvestri, F., Orlando, S.: Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data. ACM Trans. Inf. Syst. 24, 51–78 (January 2006)
9. Gordea, S., Zanker, M.: Time filtering for better recommendations with small and sparse rating matrices. In: Proc. WISE'07. pp. 171–183. Springer-Verlag, Berlin, Heidelberg (2007)
10. He, D., Göker, A.: Detecting session boundaries from web user logs. In: BCS-IRSG. pp. 57–66 (2000)
11. Hsieh-yee, L.: Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. JASIS 44, 161–174 (1993)
12. Jones, R., Klinkner, K.L.: Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In: CIKM '08. pp. 699–708. ACM (2008)
13. Lempel, R., Moran, S.: Predictive caching and prefetching of query results in search engines. In: Proc. WWW'03. pp. 19–28. ACM, New York, NY, USA (2003)
14. Lucchese, C., Orlando, S., Perego, R., Silvestri, F., Tolomei, G.: Identifying task-based sessions in search engine query logs. In: Proc. WSDM'11. pp. 277–286. ACM, New York, NY, USA (2011)
15. Markatos, E.P.: On caching search engine query results. In: Computer Communications. p. 2001 (2000)
16. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: Proc. KDD'05. ACM Press (2005)
17. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. 3(4), 333–389 (2009)
18. Siegfried, S., Bates, M., Wilde, D.: A profile of end-user searching behavior by humanities scholars: The Getty Online Searching Project Report No. 2. JASIS 44(5), 273–291 (1993)
19. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33, 6–12 (September 1999)
20. Silvestri, F.: Mining query logs: Turning search usage data into knowledge. Foundations and Trends in Information Retrieval 1(1-2), 1–174 (2010)
21. Spink, A., Saracevic, T.: Interaction in information retrieval: selection and effectiveness of search terms. JASIS 48(8), 741–761 (1997)