A Semantic Perspective on Query Log Analysis

Katja Hofmann Maarten de Rijke Bouke Huurnink Edgar Meij
ISLA, University of Amsterdam
k.hofmann@uva.nl, mdr@science.uva.nl
bhuurnink@uva.nl, edgar.meij@uva.nl

Abstract

We present our views on the CLEF log file analysis task. We argue for a task definition that focuses on the semantic enrichment of query logs. In addition, we discuss how additional information about the context in which queries are being made could further our understanding of users' information seeking and how to better facilitate this process.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Human Factors

Keywords

Query log analysis

1 Introduction

Query logs provide an excellent opportunity for gaining insight into how a search engine is used and what the users' interests are since they form a complete record of what users searched for in a given time frame. Particularly appealing is that they are collected unobtrusively, without interrupting users' normal interactions with the system. Depending on the specifics of how the data is collected, the logs may contain additional information, such as identification of users (e.g., via login name, IP-address, or cookie), their location (by IP-address), or the results that were clicked in response to each query (in this case the names "click logs" or "click data" are more common).

The information contained in query logs has been used in many different ways, for example to provide context during search, to classify queries, to infer search intent, to facilitate personalization, to uncover different aspects of a topic, etc. In various studies, researchers and search engine operators have used information from query logs to learn about the search process and to improve search engines—from early studies of the logs created by users of library catalog systems [12] to later studies of the logs of special text genres [11], Web search engines [8, 9], or the user's intent [2, 4]. More recent studies have investigated query logs for online search engines for biomedical publications [5] and multimedia search [14]. Besides learning about search engines or their users, query logs are also being used to infer semantic concepts [10] or relations [1]. Naturally, query log analysis comes with limitations. For example, we cannot identify the person behind the computer, determine demographic information, and the reason for the search (i.e., underlying information need) is not recorded [7].

Query logs are also increasingly considered as a valuable resource for informing certain aspects of information retrieval. They provide a specific view on retrieval, for example pinpointing particular types of information that users typically search for, or helping identify bottlenecks with current technology. In this way they can inform decisions on what aspects of retrieval technology to focus on.

Within the Cross Language Evaluation Forum (CLEF), LogCLEF is a new task that targets the opportunities of query log analysis. The goal of this task is the *"analysis and classification of queries in order to improve search systems."*[1] Below, we briefly repeat the key features of the LogCLEF tasks and then present our view on what a suitable task log analysis task at future editions of LogCLEF should look like.

2 LogCLEF 2009

In its first year, LogCLEF consisted of two subtasks. LADS was a general task, focusing on analyzing log files. LAGI was more concrete—participants were asked to classify queries into "geographical" and "non geographical" and to identify the geographical component (via a unique identifier, e.g., a Wikipedia page) in a second step. The query logs that were made available were provided in part by TEL (The European Library) and in part by Tumba!, a Portuguese web search engine.

The reason for starting LogCLEF was to inform future developments of the other CLEF tasks. Questions that the task should solve include: (i) What are people looking for? (ii) What should be our focus in developing new search technology? (iii) Do our methods for evaluating search reflect something that is of actual value to searchers? In this way, looking at "the user" through query log analysis can be a "real life sanity check" on whether other CLEF tasks actually make sense from a user perspective. This is an important goal of CLEF – to develop methods and tools that enable people to access multilingual information more effectively.

Both LogCLEF subtasks were run for the first time this year and in the remainder of this paper we offer our perspective on the tasks and possible future developments for log analysis at CLEF without looking at the specific outcomes of the tasks.

3 Understanding Queries

In order to work towards the stated goal of improving search systems we believe that we need some understanding of the current search process, including users' interactions with the systems. The better we understand this process, the better we can address search systems' current limitations.

Ideally, we would like to have a fully explicated user model which contains the user's personal context, task context, intent, etc. Since such a complete model is infeasible to construct we should look for surrogates or approximations and query logs are one such approximation.

The question, then, is how much query log analysis can contribute to a better understanding of the search process. An advantage is that query logs capture a large amount of activity of many users. This allows us to statistically analyze the collected data, for example using data mining, to identify patterns that would not become apparent when studying small sets of users.

A limitation is that query logs only provide a very narrow view on users' interactions with the search system. Any activity can be interpreted in many different ways. For example, someone posting the query *"girl with the pearl earring"* may want to see a photo of a particular painting, find the name of the painter, or read stories about how it was created.

As many interpretations are possible, we need to be careful in what kinds of conclusions we can actually draw from an analysis of query logs. Ultimately, only the person who does an actual search knows what they are looking for (and sometimes even they have difficulties articulating their need) or can identify the item in the search results that answered their question.

[1] `http://www.uni-hildesheim.de/logclef/`

4 Towards Semantically Enriched Query Logs

A complete understanding of the search process is not possible. We should work towards the best possible use of the resources that we currently have available and aim for robust, scaleable and repeatable types of analysis.

Our perspective is that LogCLEF should focus on *semantically enriching* query logs. This means creating annotations that specify for example the language in which a query was posed, any named entities contained in the query, and relations between the different parts of the query.

As an example, we can observe that the most frequent queries contained in the TEL log file are named entities (cf. Table 1). This is interesting in the context of a multilingual perspective that CLEF addresses: on the one hand, named entities have the same name in many languages. This means that little translations may be necessary, but also that it is hard to detect the language of most queries. Still, variations exist across languages, and in cases such as book titles, translations using statistical machine translation methods may be unfeasible.

Table 1: Top-10 most frequent queries in logclef query log, cleaned, for all search actions.

frequency	query
9844	mozart
2575	van gogh
2037	meisje met de parel
1966	harry potter
1094	pink floyd
1009	nuremberg
942	rembrandt
877	standbeeld erasmus
741	adam smith
736	salzburg

What we are proposing goes beyond—and is different from—the recognition of geo-information as it is being examined within the LAGI subtask at LogCLEF. In our view, from a user perspective, it may not matter whether a painting is located at a museum in the Netherlands, Germany or Poland (although this is potentially interesting metadata), and is therefore entered into a catalogue in either language. The painting should be found in catalogues of any language. This requires that such entities can be identified across languages and that they can be linked (or "resolved" or "normalized") to a unique identifier, such as a thesaurus term or a Wikipedia page—this unique identifier may be decorated with additional information (other occurrences, type or category information, relations to other resolved entities, etc.) that can be aggregated to provide us with insights about trends, types of queries, intent, etc, which in turn should inform search algorithm optimization and interface design.

What we are proposing, then, is to automatically enrich queries with semantic information by providing links from queries (in context, ideally with session information) to one or more sources of background information that are appropriate for the domain from which the log files originate—thesauri, directories, Wikipedia, Linked Open Data, etc. In the case of the TEL log file, one can think of GTAA and Wikipedia as suitable target sources.

How can this semantic query enrichment task be approached in a broad-coverage and robust manner? Until recently, approaches to automatic categorization of queries from a search log were mostly based on pre-defining a list of topically categorized terms, which are them matched against queries from the log; the construction of the log was done manually [3] or semi-automatically [13]. While this approach achieves high accuracy, it tends to achieve very low coverage, e.g., 8% of unique queries for the semi-automatic method, and 13% for the manual one. Mishne and de Rijke [11] take a different approach to query categorization, substantially increasing the coverage but sustaining high accuracy levels: their approach relies on external "categorizers" with access to large amounts of data—two category-based web search services, Yahoo! Directory and Froogle. Meij et al. [10] use a (high-performing and very broad

coverage) feature-based approach to linking queries to DBpedia in conjunction with search-based and concept-specific features and apply their method to the transactions of *Beeld en Geluid*, the national Dutch radio and television archive; the authors also provide guidelines and a test set for this linking task, showing that ground truth can be established in a reliable manner with relatively little effort. Huurnink et al. [6] show how the resulting information can be used to gain insights into users' search behavior by aggregating the information being linked to.

5 Conclusion

We welcome the arrival of a log analysis task at CLEF. Identifying basic statistical patterns in the log files made available for the task is a valuable first step. But as a shared task, we need a task that goes beyond this—the recognition of geographic components (as implemented in the LAGI subtask) is a good example, but it does not seem completely appropriate for the TEL log files where enrichment of queries with a broader range of semantic information seems more suitable. Instead, we propose a semantic query enrichment task that aims to link queries to suitable semantic sources. Recent advances in semantic query analysis suggests that has become a do-able task, without having reached the status of solved problem—making it suitable for collaborative benchmarking in the CLEF setting.

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802, and by the Dutch-Flemish research programme STEVIN under project DuOMAn (STE-09-12).

References

[1] Ricardo Baeza-Yates and Alessandro Tiberi. Extracting semantic relations from query logs. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 76–85, New York, NY, USA, 2007. ACM Press. ISBN 9781595936097. doi: 10.1145/1281192.1281204. URL `http://dx.doi.org/10.1145/1281192.1281204`.

[2] Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. The intention behind web queries. In *String Processing and Information Retrieval*, pages 98–109, 2006. doi: 10.1007/11880561_9.

[3] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings SIGIR '04*, pages 321–328, New York, NY, USA, 2004. ACM Press.

[4] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002. ISSN 0163-5840. doi: 10.1145/792550.792552. URL `http://dx.doi.org/10.1145/792550.792552`.

[5] Jorge R. Herskovic, Len Y. Tanaka, William Hersh, and Elmer V. Bernstam. A day in the life of pubmed: analysis of a typical day's query log. *J Am Med Inform Assoc*, 14(2):212–220, 2007. ISSN 1067-5027. doi: 10.1197/jamia.M2191. URL `http://dx.doi.org/10.1197/jamia.M2191`.

[6] Bouke Huurnink, Laura Hollink, Wietske van den Heuvel, and Maarten de Rijke. Information needs of broadcast professionals at an audiovisual archive: A transaction log analysis. *Submitted*, 2009.

[7] Bernard J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28(3):407–432, 2006.

[8] Bernard J. Jansen and Udo Pooch. A review of web searching studies and a framework for future research. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):235–246, 2001.

[9] Thorsten Joachims. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, New York, NY, USA, 2002. ACM Press. ISBN 158113567X. doi: 10.1145/775047.775067. URL http://dx.doi.org/10.1145/775047.775067.

[10] Edgar Meij, Marc Bron, Bouke Huurnink, Laura Hollink, and Maarten de Rijke. Learning semantic query suggestions. In *8th International Semantic Web Conference (ISWC 2009)*. Springer, October 2009.

[11] Gilad Mishne and Maarten de Rijke. A study of blog search. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Advances in Information Retrieval: Proceedings 28th European Conference on IR Research (ECIR 2006)*, volume 3936 of *LNCS*, pages 289–301. Springer, April 2006.

[12] Thomas A. Peters. The history and development of transaction log analysis. *Library Hi Tech*, 11(2): 41–66, 1993.

[13] Hsiao Tieh Pu and Shui Lung Chuang. Auto-categorization of search terms toward understanding web users' info rmation needs. In *ICADL 2000: Intern. Conference on Asian Digital Libraries*, 2000.

[14] Dian Tjondronegoro, Amanda Spink, and Bernard J. Jansen. A study and comparison of multimedia web searching: 1997-2006. *J. Am. Soc. Inf. Sci. Technol.*, 2009.