Characterising Dataset Search Queries

Emilia Kacprzak
University of Southampton, The Open Data Institute
London, UK
emilia.kacprzak@theodi.org

Laura Koesten
University of Southampton, The Open Data Institute
London, UK
laura.koesten@theodi.org

Jeni Tennison
The Open Data Institute
London, UK
jeni@theodi.org

Elena Simperl
University of Southampton
Southampton, UK
e.simperl@soton.ac.uk

ABSTRACT

The amount of data generated and published on the web is increasing rapidly, but search for structured data on the web still presents challenges. In this paper we explore dataset search by analysing queries specifically generated for this work through a crowdsourcing experiment and comparing them to a search log analysis of queries on data portals. The change in search environment together with the task we gave people altered the generated queries. We found that queries issued in our experiment were much longer than search queries for datasets on data portals. They further contained seven times more mentions of geospatial and of temporal information and are more likely to be structured as questions. These insights can be used to tailor search functionalities to the particular information needs and characteristics of dataset search.

KEYWORDS

dataset search, search log analysis, query generation

ACM Reference Format:
Emilia Kacprzak, Laura Koesten, Jeni Tennison, and Elena Simperl. 2018. Characterising Dataset Search Queries. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191597

1 INTRODUCTION

More and more data generated by individuals, industry and governments can be accessed online. Searching for structured data on the web is becoming part of people's work activities. However, data search still presents challenges to many people using data for their work tasks [11]. By data we mean sets of facts or figures that are presented in a structured form. These are typically grouped into datasets, many of which are published on the web, for instance in data catalogues available on the web. This work is concerned with how people search for datasets online. This is typically done through data portals or via conventional web search engines. The latter are not ideal for data search, as they have been designed primarily for documents, not data [4]. While we know a lot about general web search from literature we still know relatively little about how people search for datasets online. In a prior search log analysis of open data portals we identified differences in search queries for documents as opposed to search queries for datasets [9]. However, we assume that the analysed queries do not necessarily reflect how people would search for data in an ideal search system. People often do not expect their search activity for data to be successful, as they are aware of the limited performance of search functionalities for data [11]. We further believe the user experience of searching for documents on the web, as well as the interface (e.g. sizing and design of the search box) influence how people search for data. This work aims to better understand the specific characteristics of dataset search scenarios to inform the design of future search functionalities tailored to structured data. We analysed a set of queries for data (*crowd queries*) which were generated using human computation. We asked crowd workers to generate queries based on a sample of free form text requests for data. We compare these queries to those used to search data portals [9] to understand whether the change of search environment results in different characteristics of queries.

We looked at the following research questions: (1) How do search queries for data within a data portal differ from those in a less constrained environment? (2) How are search queries for data that have been issued in a less constrained environment structured?

We found that queries generated in a less constrained environment are much longer and contain seven times more temporal and geospatial information than queries issued to data portals [9]. This indicates these information types as important and distinct factors that are of particular interest in dataset search. We further found a larger prevalence of file types, dataset types and numbers represented in the queries. We believe a better understanding of what dataset search queries look like is needed to inform interfaces and search functionalities for the unique characteristics of dataset search. This includes the importance of indexing temporal or geospatial information or specific presentation modes for search results that are tailored to dataset search.

2 RELATED WORK

Data search Searching for data still presents challenges, even for data professionals, and is far from providing the same user experience we are used to in web search [10, 11]. Existing techniques used in general web search cannot be directly applied to searching for data [4]. In a prior study analysing the search logs of four

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191597

governmental open data portals [9] we discussed the specific characteristics of queries for datasets and compared to general web search queries. General web search queries have evolved over time in parallel with advances in search functionalities. Data search is similar in its characteristics to web search around 15 years ago [9, 16] (e.g. queries have been steadily growing in length over time [17]). We assume different information types constitute for different search contexts and influence the way people query. Dataset retrieval on the web is still a relatively immature research area. This study focuses on the characteristics of query formulation for structured datasets. Structured data, means data that is organised explicitly - such as in spreadsheets, web tables, databases or maps.

Search in different environments Vertical search is search that targets a specific subset of online content (which could be distinct based on its topic, data type or its context). The limited scope of relevant resources allows for greater precision, more complex schemas or ontologies to match specific search scenarios; and so tends to support more complex user tasks [13]. Verticals include for instance people search [18], email search [1], research publication search [13] or digital libraries [8]. Each vertical has a clear distinction from other verticals and from general web search. For instance, email search is an example in which [1] noticed that when searching, users know the precise attributes of a resource they are looking for. The key differences to general web search are that the set of emails is a personal set unique for each user and there is additional metadata (e.g. sender address, subject or time stamp) which can help both organising and searching through the results. In search for research publications [13] argue that web search could be improved for this vertical by using temporal information attached to each publication. For instance, algorithms like PageRank and HITS calculate the relevance of each resource and favour older resources over newer ones in their ranking. In publication search, the reputation of the resource, in addition to its content relevance, citation count and reputation of its authors and journals are more influential. Dataset search can be seen as a separate vertical due to the specific characteristics of the information source. Kunze et al. have recently introduced the concept of *dataset retrieval*, as a branch of information retrieval applied to data instead of documents focused on determining the most relevant datasets according to a user query [12]. Their focus was on a specific data type - RDF datasets - however, we believe that this applies to structured data independent of its format.

Query analysis The first query log analysis on the web was made for the Altavista search engine [15], and the technique has been used since to study several aspects of web search (see [7] for a survey). Broder et al. report a *classification of query types* in their taxonomy of web search queries which is based on user needs [3]. This includes informational, navigational and transactional queries. In dataset search the information need is 'finding data' and can therefore be seen as predominantly informational. Various metrics for analysing queries were developed in the area of general web search several of which can apply to data search.

Query Length & Distribution are the most commonly presented statistics and are part of the analysis in our study. *Query Structure* describes the prevalence of e.g. questions, operators, and whether the query is composite or non-composite [2], which is mostly relevant for long queries. Question queries are identified by starting words that indicate questions. Operators are boolean operators: AND, OR and NOT or special web search operators eg. url, site or filetype. As shown in [6], who report a transaction log analysis of nine search engines, results of different search log analyses are not directly comparable. This means that even within web search it is problematic to compare metrics of different search log analyses, we assume that between different information sources (e.g. textual documents versus structured datasets) it might be even more so.

3 EXPERIMENT

Data We analysed a set of search queries generated in this work (*crowd queries*) and compared them with queries from a dataset search log analysis [9] which analysed queries from four open data portals referred to as *portal queries*. *Crowd queries* were generated in a crowdsourcing experiment based on data requests to the UK Government Open Data portal[1] and are available in a Github repository[2]. These are formal natural language requests for data that users could not find on the platform, submitted via a semi-structured contact form and available as open data[3]. An example of an excerpt of a data request is *"Request annual return data on total numbers of Sheep & Lambs and Cattle & Calves in the following two North Yorkshire parishes from* 1986 *to the latest available date: for Malham Moor Parish and for Buckden Parish"*. We randomly selected 10% (50 requests) of all the openly published data requests, and manually checked for their understandability concerning language and domain specific terminology - we then excluded requests which were potentially difficult to understand and replaced them with other randomly selected requests.

In our experiments we used the title and the description of the request. For each data request we generated 10 queries through human computation. After excluding spam answers (51 of all queries, which were manually detected) the set contained 449 queries in total. An example of a resulting query is *"Businesses in Yorkshire that employ over* 1000 *workers"*.

Design We conducted a crowdsourcing experiment to generate *crowd queries*. Participants were users of the crowdsourcing platform CrowdFlower. As the data requests are unstructured English text that could potentially be difficult to understand for people with low English language skills, we limited the experiment to workers in native-English speaking countries; and we restricted the worker pool to a smaller group of more experienced, higher accuracy contributors on the platform. We included 5 short qualification questions, assessing basic reading, reasoning and data literacy skills. Workers were paid $0.15 to generate each search query that they considered suitable for a single data request.

Our open-ended text creation task was formulated as: *We ask you to write a search query which you think would return the requested dataset from a data search engine*. The workers were shown an overview of the task, step-by-step instructions, and a sample data request with examples of corresponding queries. The output was a search query constrained to be between one and twenty words in length. To minimise "spam" answers we prevented pasting of content and validated each word from the query against an English

[1] data.gov.uk
[2] https://github.com/chabrowa/data-requests-query-dataset
[3] https://data.gov.uk/dataset/data-requests-at-data-gov-uk

language dictionary, requiring an 80% matching threshold for a query to be accepted. We also rejected answers containing the same word three or more times. Participants were not instructed to generate their queries in a particular structure; however, they were shown five examples, with various compositions of keywords, and a question. The minimum time permitted to generate a single query was 1 minute to allow time for detailed reading of the data requests. No personal data was collected. Despite the workers' lack of in-depth understanding of the information need that is represented in the data request we believe that the resulting queries give us valuable insights into the necessary complexity and characteristics of queries for data.

4 RESULTS

Analysis Statistics of the queries analysed include: *Query length* including average query length and distribution for both sets of queries; *Query characteristics*: queries containing keywords describing: location; time frame; file and dataset type; numbers; abbreviations (described in detail in Table 1); and *Question queries*: to recognise question queries we counted queries containing the words: *what*, *who*, *where*, *when*, *why*, *how*, *which*, *whom*, *whose*, *whether*, *did*, *do*, *does*, *am*, *are*, *is*, *will*, *have*, *has*, as in [2]. In this section we present the results of our analysis of the *crowd queries* created in this study and compare these to the portal queries presented in [9].
Query Length The majority of portal queries have been reported to be between one and three words, with an average of 2.03 words per query. We found *crowd queries* to be significantly longer than portal queries with an average of 9.16 words per query.

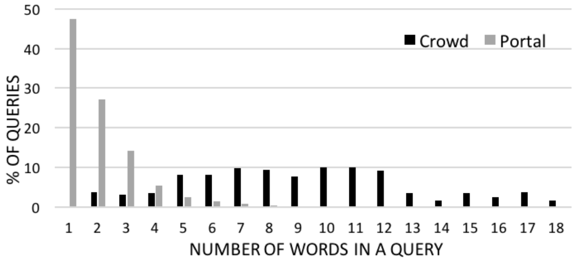

Figure 1: Percentage of queries according to number of words in them

Figure 1 shows an overview of the percentage of queries by number of words per query, for both the crowd and the portal queries reported in [9]. We can see single word queries represent almost half of the entire corpus of the portal queries whereas the *crowd queries* had a minimum of 2 words, with the most queries between 7 to 11 words. We believe this difference in query length indicates that the portal queries might not represent realistic search strategies, but rather expose limitations of current dataset search. Users do not expect search functionalities to fulfil their information need when searching for data, which can lead to underspecified queries [11].
Query Types - In this work we analysed the same metrics as in [9], which include geospatial, temporal, numerical information or appearances of file type and acronyms in the query. Table 1 summarises the percentage of queries for each of the metrics. Geospatial information was much more prevalent in the *crowd queries*: 36.1% of those contained a location in comparison to only 5.4% of portal queries and 12.01% in general web search [5]. In contrast to searching on a data portal, which is often tied to a specfific location or can have national boundaries attached to it, our experiment did not specify a particular location and was so less constrained from a geospatial point of view. Participants may have compensated for this by specifying location keywords. However, the high number of location bound keywords (36.1%) may simply emphasise the high importance of location in data search. Temporal information was seven times more popular in the *crowd queries* 49.2% in contrast to the results reported for portal queries (7.29%) and 32 times more prevalent than for general web search (1.5% [14]). Users indicate interest in different aspects of temporal information: date of data creation, the frequency of data releases, updates and time frames described in the data. File and dataset types (such as queries containing *csv* or *json*, etc as can be seen in Table 1 were much more popular within *crowd queries* (49%) in comparison to the portal queries for which file types were reported for 6.25% of the queries. This could be due to filtering options over file types on the data portals in which the portal queries were recorded. Crowd workers could further be biased in their creation of queries in thinking they need to add the word *data* to a query for data (as would be the case on general web search). Excluding the word data from this analysis we found 26.95% queries including common file types, as can be seen in Table 1. The percentage of queries containing numbers, that were not temporal information, were 5.57% and there were no queries containing only numbers. These results are similar to those reported for portal queries (5.23%) [9]. Numbers in queries represent mainly sample sizes or desired constraints to the data, such as: *Police spending over £500 local data*. We further report the percentage of queries including abbreviations. We found 2.23% in the *crowd queries* included abbreviations; in comparison to 5.11% reported for the portal queries. Abbreviations were mostly used when they appeared in the data request that the query was based on.

Metric - Definition	% portal	% crowd
Geospatial - the name of a city or geographical area (either town, city, county, region or countries)	5.4%	36.1%
Temporal - years (1000 to 2017), names of months, days of a week and the words *week(ly), year(ly), month(ly), day(ly), date, time* and *decade*	7.3%	49.2%
File and dataset type - file types: *csv, pdf, xls, json, wfs, zip, html, api* and keywords denoting a type of dataset: *data, dataset, average, index, graph, table, database, indice, rate, stat*	6.3%	49%
Numbers - the number of queries including numbers excluding those indicating time frames	5.2%	5.6%
Only numbers - queries that contain only numbers	0.4%	0%

Table 1: Definition of query characterisation metrics. Percentage of queries for both, portal queries as reported in [9] and crowd queries of this study

Question queries Formulating queries as questions is increasingly common in web search, thanks to advances in speech recognition and conversational search interfaces [19]. In 2009, 7.49% queries were questions in a study on general web search [2]. Less than 1% of the portal queries are structured as questions [9]. The low number of question queries in dataset search might be due to the lack of question-answering capabilities of the dataset search functionalities. We found 9.35% of *crowd queries* were questions. This may be due to the larger search box used in our experiment; or due to a different conceptualisation of data opposed to documents.

5 DISCUSSION AND LIMITATIONS

We found that queries for data differ between those issued on a data portal and those created in an environment with fewer constraints. The queries generated in this study were longer and included approximately seven times more temporal and geospatial information.

The higher importance of these information types has been recognised in literature[10, 12]. Structurally we found the *crowd queries* to include a higher percentage of questions and 4 times as many queries included a specific file type or format. The length of these queries suggest that the information need expressed in the data requests are complex; based on literature we believe this is typical for data centric information needs [11]. In comparison, the portal queries were short and underspecified. Although in both of the query sets people were looking for data, we believe that neither set necessarily represents how people would like to search for data. These findings emphasise a large design space for data search environments; one possible direction being encouraging users to issue longer queries, for instance by providing larger search boxes or suggesting additional keywords. Search log analyses can illustrate the specific characteristics of a given search vertical. We know that queries on portals are underspecified, but this work shows that when asked to search outside of a search environment people issue much longer queries which correspond to complex information needs for data. The high prevalence of geospatial or temporal information in the *crowd queries* should inform the design of dataset search systems, for instance by allowing users to search by specific locations or time frames. It could further indicate a need to extend existing metadata standards to include these two types of information, which could then be exploited by search functionalities. We believe new retrieval models for dataset search, that take the unique characteristics of this information source into account, are needed to make data on the web more discoverable.

Limitations As with any experiment using human computation, instructions and the experiment design influence the outcome. We tried to take into account that workers might not know what *data* is and used a spreadsheet as well as a product search analogy in the instructions. We had no control of the workers prior experience and their conceptual models of data. However, this is a natural limitation of such experiments. While we acknowledge that the *crowd queries* are created in an artificial setting, without the workers own naturalistic information need, we believe that they give us relevant insights into how queries for data could potentially look in the future. We acknowledge that neither query set can be a representative reflection of how people would search for data in an "ideal" system. However the results of this work can be seen as an approximation that can inform further research.

6 CONCLUSION & FUTURE WORK

In this work we present a search log analysis of queries generated through crowdsourcing using requests for data from an open data portal to describe an information need. We compare our results to [9], in which we analysed queries explicitly issued to find data. Our findings indicate that dataset search logs do not fully represent the behaviour of users searching for data, but rather uncover the limitations of current search functionalities. The differences of the two sets of queries indicate the need for further research to deepen our understanding of how people search for data. This could include an additional analysis of queries generated by data professionals to understand the differences and commonalities of queries issued by people with different prior knowledge about data. Future studies could include a more in-depth analysis of user behaviour in dataset search, taking into account search sessions and query refinements together with a qualitative component to better understand user needs. This can enable us to develop search functionalities for data that are tailored to user needs as well as to the particular characteristics of dataset search.

Acknowledgements This project is supported by the EU H2020 program: Marie Skłodowska-Curie grant agreement No. 642795

REFERENCES

[1] Qingyao Ai, Susan T. Dumais, Nick Craswell, and Daniel J. Liebling. 2017. Characterizing Email Search using Large-scale Behavioral Logs and Surveys. In *Proceedings of the 26th International Conference on World Wide Web, WWW*. 1511–1520. https://doi.org/10.1145/3038912.3052615

[2] Michael Bendersky and W. Bruce Croft. 2009. Analysis of Long Queries in a Large Scale Search Log. In *Proceedings of the 2009 Workshop on Web Search Click Data*. ACM, 8–14. https://doi.org/10.1145/1507509.1507511

[3] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002), 3–10. https://doi.org/10.1145/792550.792552

[4] Michael J. Cafarella, Alon Halevy, and Jayant Madhavan. 2011. Structured Data on the Web. *Commun. ACM* 54, 2 (2011), 72–79. https://doi.org/10.1145/1897816.1897839

[5] Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. 2008. Analysis of Geographic Queries in a Search Engine Log. In *Proceedings of the First International Workshop on Location and the Web*. ACM, 49–56. https://doi.org/10.1145/1367798.1367806

[6] Bernard J. Jansen and Amanda Spink. 2006. How Are We Searching the World Wide Web?: A Comparison of Nine Search Engine Transaction Logs. *Information Processing and Management* 42, 1 (2006), 248–263. https://doi.org/10.1016/j.ipm.2004.10.007

[7] Daxin Jiang, Jian Pei, and Hang Li. 2013. Mining Search and Browse Logs for Web Search: A Survey. *ACM Transactions on Intelligent Systems and Technology* 4, 4, Article 57 (2013), 37 pages. https://doi.org/10.1145/2508037.2508038

[8] Steve Jones, Sally Jo Cunningham, Rodger McNab, and Stefan Boddie. 2000. A transaction log analysis of a digital library. *International Journal on Digital Libraries* 3, 2 (2000), 152–169. https://doi.org/10.1007/s007999900022

[9] Emilia Kacprzak, Laura M. Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. *A Query Log Analysis of Dataset Search*. Springer International Publishing, Cham, 429–436. https://doi.org/10.1007/978-3-319-60131-1_29

[10] Dagmar Kern and Brigitte Mathiak. 2015. Are There Any Differences in Data Set Retrieval Compared to Well-Known Literature Retrieval?. In *19th International Conference on Theory and Practice of Digital Libraries, TPDL*. 197–208. https://doi.org/10.1007/978-3-319-24592-8_15

[11] Laura M. Koesten, Emilia Kacprzak, Jenifer F. A. Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data - a Study on Information Seeking Behaviour. In *Proceedings of Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1277–1289. https://doi.org/10.1145/3025453.3025838

[12] Sven R. Kunze and Soren Auer. 2013. Dataset Retrieval. In *2013 IEEE Seventh International Conference on Semantic Computing*. https://doi.org/10.1109/ICSC.2013.12

[13] Xin Li, Bing Liu, and Philip S. Yu. 2010. *Time Sensitive Ranking with Application to Publication Search*. Springer, New York, 187–209. https://doi.org/10.1007/978-1-4419-6515-8_7

[14] Sérgio Nunes, Cristina Ribeiro, and Gabriel David. 2008. Use of temporal expressions in web search. In *European Conference on Information Retrieval*. Springer, 580–584.

[15] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *ACM SIGIR Forum* 33, 1 (1999), 6–12.

[16] Amanda Spink, Dietmar Wolfram, Major BJ Jansen, and Tefko Saracevic. 2001. Searching the web: The public and their queries. *Journal of the American society for information science and technology* 52, 3 (2001), 226–234.

[17] Mona Taghavi, Ahmed Patel, Nikita Schmidt, Christopher Wills, and Yiqi Tew. 2012. An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces* 34, 1 (2012), 162–170.

[18] Wouter Weerkamp, Richard Berendsen, Bogomil Kovachev, Edgar Meij, Krisztian Balog, and Maarten de Rijke. 2011. People Searching for People: Analysis of a People Search Engine Log. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.*

[19] Ryen W. White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. Queries in Informational Search Tasks. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, 135–136. https://doi.org/10.1145/2740908.2742769