# Man vs. Machine
# Differences in SPARQL Queries*

Laurens Rietveld[1] and Rinke Hoekstra[1,2]

[1] Department of Computer Science, VU University Amsterdam, The Netherlands
{laurens.rietveld,rinke.hoekstra}@vu.nl
[2] Leibniz Center for Law, University of Amsterdam, The Netherlands
hoekstra@uva.nl

**Abstract.** Server-side SPARQL query logs have been a topic of study for some time now. The USEWOD collection of query logs is currently the primary source of information for researchers. A recurring problem is that these logs leave application queries and queries created by humans indistinguishable. In this paper, we present a new collection of manually created queries, that were collected through the YASGUI SPARQL interface. We show how these queries differ from those taken from server logs.

**Keywords:** SPARQL, endpoints, Semantic Web, Linked Data

## 1 Introduction

SPARQL queries are the standard for querying the Linked Open Data (LOD) cloud. They are an interesting subject of study: analyzing SPARQL query logs can help optimize triple-stores, understand and improve user interaction, improve the way we rank the query results, and much more. Unfortunately, these query logs are hard to come by and the only (semi) public source of query logs is collected and made available by the USEWOD workshop [3]. These query logs constitute the primary source of information about how SPARQL endpoints are used. However, because these concern *server side* logs, it is near impossible to reliably distinguish between queries *directly created* by the (human) user, and queries coming from applications. An example of the latter are queries executed by the Pubby interface [4]. Browsing an endpoint via this interface automatically executes several SPARQL queries in the background (all having exactly the same structure, though with different resources).

The need to distinguish between machines and humans, is recognized by others as well [9,10,13], and is important for research involving *users*.

This paper provides more context to the *server side* USEWOD logs by showing what manually created queries look like, and how they structurally differ from server log queries. We make this comparison using *client side* query logs

* This work was supported by the Dutch national program COMMIT/

from the YASGUI SPARQL editor [3], presented in [14]. YASGUI (Yet Another SPARQL Graphical User Interface) is a feature-packed SPARQL query editor, capable of accessing *any* SPARQL endpoint.

In Section 2 we present the related work. Then, in Section 3 we discuss how we collect and analyze the queries, followed by Section 4, where we present our results. We conclude in Section 5.

## 2 Related Work

Research related to SPARQL query templates and user sessions [9] showed a large number of similarly-structured SPARQL queries in the USEWOD dataset. The authors argue that these queries are most likely issued by machine agents. Only a small percentage of relatively short query sessions show heterogeneous structure, possibly indicating human users.

Similar work done by [13] extracts user sessions from the USEWOD logs, where the authors try to differentiate between both human and machine 'user' sessions. They observe that ""*A very small set of users contribute to a large percentage of the queries*", and hypothesize that these users are machine agents.

Although attempts to distinguish between humans and machines provide insightful information, it remains impossible to distinguish between both types of queries with a known degree of certainty. The burden for distinguishing between both types of queries can be alleviated with the availability of a man-made collection of queries. Such a query collection is particularly interesting for Linked Data research with a strong *user* aspect, such as detecting *user* browsing patterns [8], *user* modeling [5] and query modification assistants for *users* [7].

## 3 Approach

We collect the man-made queries via the YASGUI interface, which is packed with usability features such as auto-completion, syntax highlighting, dataset endpoint search, and sharing functionalities.

When given permission to do so, YASGUI tracks the actions of users using Google Analytics[4], including the endpoint accessed by the user, the specific query, and the time it takes to execute the query.[5]

Since the public launch of YASGUI early 2013, it has been quite successful in attracting visitors from across the world (See Figure 1). Based on our logs we observe that YASGUI received 2.611 unique visitors, and 39.201 queries were executed against 645 endpoints. Because 40% of the YASGUI users do not allow collecting these statistics, the figures presented in this paper only cover a subset of the YASGUI user base.

[3] See `http://yasgui.org`

[4] See `http://www.google.com/analytics/`

[5] Tracking query execution time was added recently, and not included in this paper

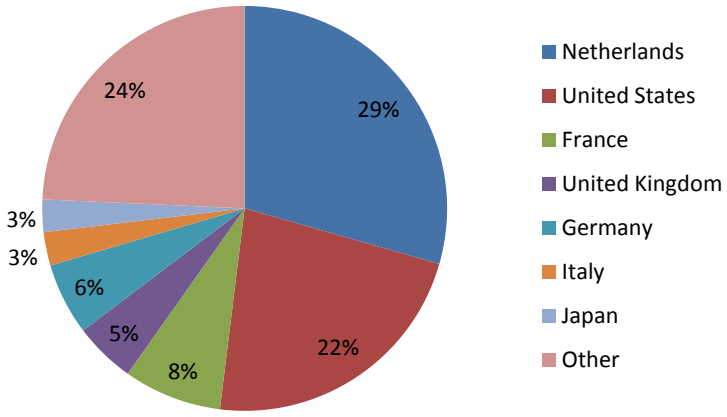


Fig. 1: Location of YASGUI users

We compare YASGUI queries to the server logs from the USEWOD 2014 challenge, which contain logs from DBpedia [1], Linked Geo Data [2] and BioPortal [11]. Because DBpedia is the most popular endpoint in the YASGUI logs (around 15.000 queries), and because the number of Linked Geo Data and BioPortal queries we could collect is relatively low (100 and 5 queries, respectively), we only compare YASGUI and USEWOD queries to DBpedia.

Ideally, we avoid a bias by removing duplicate queries on a per-user basis, as users tend to send the same query multiple times. However, because Google Analytics mostly presents aggregate information, and no detailed visitor information (e.g. user-agent or host name), we cannot remove duplicates (per user) for our YASGUI logs. Therefore, for the sake of comparison, all the statistics presented in this papers are performed on the *complete* set of queries in our query sets. The only requirement we impose, is that all queries should be syntactically correct.

Our analysis of the USEWOD and YASGUI logs consists of extracting structural properties from the queries, and is largely based on [6], and to a lesser extend [12]. We extract these properties using the Jena query parser[6]. The properties we extract for each query are:

1. Query Type, *e.g.* `SELECT` or `CONSTRUCT`
2. The use of SPARQL features such as `DISTINCT`, `UNION`, `ORDER BY` and `LIMIT`.
3. The types of triple patterns, *i.e.* an RDF triple in which each item can be a variable. For each item in the triple pattern, we analyze whether this is a variable (`V`), a constant (`C`, either a URI, literal or blank node), or a property path (`*`). We would represent a triple pattern such as `[] rdfs:label ?label` as `C C V`.

[6] See `http://jena.apache.org/`

| | YASGUI | USEWOD |
|---|---|---|
| #syntactically valid queries | 13.242 | 100.763 |
| % unique queries | 65.73% | 69.67% |
| **Type** | | |
| `SELECT` | 93.91% | 96.17% |
| `DESCRIBE` | 0.72% | 3.00% |
| `CONSTRUCT` | 1.49% | 0.56% |
| `ASK` | 3.87% | 0.26% |
| **SPARQL features** | | |
| `ORDER BY` | 18.64% | 6.37% |
| `DISTINCT` | 15.07% | 24.37% |
| `LIMIT` | 42.32% | 12.13% |
| `OFFSET` | 0.14% | 0.75% |
| `FILTER` | 30.35% | 15.11% |
| Subquery | 2.82% | 0.37% |
| `SERVICE` | 0.91% | 0.08% |
| $\geq 1$ `UNION` in query | 24.04% | 4.46% |
| $\geq 1$ `OPTIONAL` in query | 1.71% | 3.14% |

Table 1: Use of SPARQL grammar

4. The number and types of joins. We use the same approach as [6] in counting the types of joins. There are 6 possible types of joins, depending on the position of the common variable between two triple patterns: Subject-Subject, Predicate-Predicate, Object-Object, Subject- Predicate, Subject-Object and Predicate-Object.

## 4 Results

Table 1 shows the range of SPARQL features used by both query sets. In general, the query types roughly correspond, where the use of SPARQL features and solution modifiers greatly differ. The `LIMIT` solution modifier occurs in roughly 42% of the YASGUI queries, and only in 12% of the USEWOD queries. Additionally, the `FILTER` function shows a striking difference (roughly 30% vs 15%), as well as the number of queries containing UNION clauses (roughly 24% vs 5%). We observe that most of the SPARQL features are used more often in YASGUI queries than in USEWOD queries, which might indicate that human users use a wider range of SPARQL features than queries coming from applications (assuming the majority of USEWOD queries originate from applications [9,13]).

We can observe *structural* differences between the queries in both sets as well. In figure 2, we show the number of triple patterns per query (note the logarithmic scale). Only 23% of the USEWOD queries contain more than a single triple pattern, compared to 68% of YASGUI queries. As Table 2 shows, the type of triple patterns differs greatly as well. Patterns of the form `V C C` or

`V C V` occur roughly twice as often in the YASGUI logs than in the USEWOD logs. However, the pattern `C C V` is the most dominant one in the USEWOD logs (42.71%), where it occurs in only 7.10% of the YASGUI triple patterns.

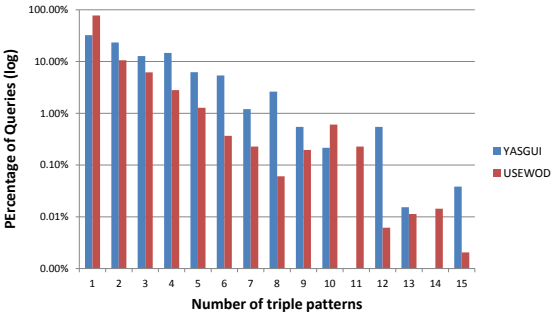


Fig. 2: #Triple patterns per query (log)

| | YASGUI | USEWOD |
|---|---|---|
| `V C C` | 45.91% | 19.92% |
| `V C V` | 36.22% | 19.06% |
| `C C V` | 7.10% | 42.71% |
| `C V V` | 2.95% | 6.10% |
| `V V C` | 2.88% | 3.92% |
| `V V V` | 1.61% | 1.88% |
| `C C C` | 0.28% | 0.17% |
| `C V C` | 0.12% | 0.06% |
| `V * C` | 2.59% | 6.01% |
| `V * V` | 0.30% | 0.10% |
| `C * V` | 0.05% | 0.08% |
| `C * C` | 0.00% | 0.00% |

Table 2: Triple patterns types (`C`=constant, `V`=variable, `*`=property path)

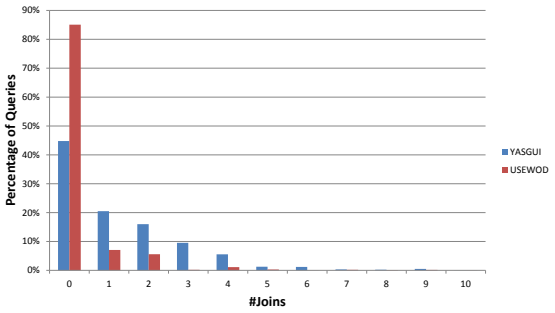


Fig. 3: Number of joins in queries

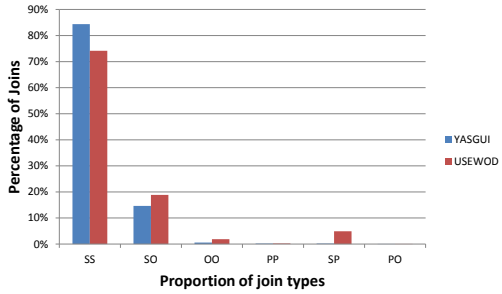


Fig. 4: Proportion of all join types

Finally, Figure 3 shows that the number of joins differs greatly between both query sets. This is for an important part due to the large number of USEWOD queries containing only a single triple pattern (*i.e.* by definition no join). Figure 4 shows the proportions of join types for both query sets, showing little difference in join types between both query sets.

## 5 Discussion

This paper shows that *user* queries obtained at *client side* are very different from queries logged by SPARQL servers: the server queries are smaller, have fewer joins, and use less features of the SPARQL grammar. This indicates that man made queries are more *complex* than routine queries fired by applications. We can furthermore conclude that both types of queries are structurally very different: the type of triple patterns observable in queries deviates greatly between both query sets. We believe that our results further corroborate the work of [9,13], and that these differences are largely caused by the many application queries in the USEWOD logs.

This insight in the structural differences between query sets is important for research related to *users* on the Web of Data. For such research, the YASGUI query logs can grow to be an interesting, more reliable source of data as it only contains man-made queries. A second advantage is that the range of endpoints for which YASGUI collects query logs is higher than any query collection currently available, allowing for a broader analysis on how the Web of Data is used. With time, following an increase in uptake of YASGUI, we can paint an even clearer picture on how *users* use SPARQL and the Web of Data..

## References


1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)
2. Auer, S., Lehmann, J., Hellmann, S.: Linkedgeodata: Adding a spatial dimension to the web of data. In: The Semantic Web-ISWC 2009, pp. 731–746. Springer (2009)
3. Berendt, B., Hollink, L., Luczak-Rösch, M., Möller, K., Vallet, D.: Proceedings of USEWOD2014 - 4th international workshop on usage analysis and the web of data. In: 11th ESWC (2014)
4. Cyganiak, R., Bizer, C.: Pubby-a linked data frontend for sparql endpoints. Retrieved fro m http://www4. wiwiss. fu-berlin. de/pubby/at May 28, 2011 (2008)
5. Fortuna, B., Mladenic, D., Grobelnik, M.: User modeling combining access logs, page content and semantics. arXiv preprint arXiv:1103.5002 (2011)
6. Gallego, M.A., Fernández, J.D., Martinez-Prieto, M.A., de la Fuente, P.: An aempirical study of real-world SPARQL queries. In: UseWod Workshop. pp. 3–6 (2011), `http://arxiv.org/abs/1103.5043`
7. Hollink, V., de Vries, A.: Towards an automated query modification assistant. arXiv preprint arXiv:1104.0128 (2011)
8. Hoxha, J., Junghans, M., Agarwal, S.: Enabling semantic analysis of user browsing patterns in the web of data. arXiv preprint arXiv:1204.2713 (2012)
9. Lorey, J., Naumann, F.: Detecting sparql query templates for data prefetching. In: The Semantic Web: Semantics and Big Data, pp. 124–139. Springer (2013)
10. Möller, K., Hausenblas, M., Cyganiak, R., Handschuh, S.: Learning from linked open data usage: Patterns & metrics (2010)
11. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research 37(suppl 2), W170–W173 (2009)

12. Picalausa, F., Vansummeren, S.: What are real sparql queries like? In: Proceedings of the International Workshop on Semantic Web Information Management. p. 7. ACM (2011)
13. Raghuveer, A.: Characterizing machine agent behavior through sparql query mining. In: Proceedings of the International Workshop on Usage Analysis and the Web of Data, Lyon, France (2012)
14. Rietveld, L., Hoekstra, R.: Yasgui: Not just another sparql client. In: The Semantic Web: ESWC 2013 Satellite Events, pp. 78–86. Springer (2013)