Usage Analysis and the Web of Data

Bettina Berendt
K.U. Leuven
Belgium
bettina.berendt@cs.kuleuven.be

Laura Hollink
TU Delft
Netherlands
l.hollink@tudelft.nl

Vera Hollink
CWI Amsterdam
Netherlands
v.hollink@cwi.nl

Markus Luczak-Rösch
FU Berlin
Germany
markus.luczak-roesch@fu-berlin.de

Knud Möller
NUI Galway
Ireland
knud.moeller@deri.org

David Vallet
Universidad Autónoma de Madrid
Spain
david.vallet@uam.es

Abstract

The workshop on *Usage Analysis and the Web of Data* (USEWOD2011) was the first workshop in the field to investigate combinations of usage data with semantics and the Web of Data. Questions the workshop aims to address are for example: How can semantics help in understanding usage data, how can semantic information be derived from usage data, and how can we learn about usage of and on the emerging Web of Data, and what can we learn from it? We report on the findings and results of this workshop, held on March 28, 2011 in conjunction with 20th International World Wide Web Conference, in Hyderabad, India.

1 Introduction

The purpose of the USEWOD2011 workshop[1] was to investigate the synergy between semantics and semantic-web technology on the one hand and analysis and mining of usage data on the other hand. The two fields are a promising combination. First, semantics can be used to enhance the analysis of usage data. Usage logs contain information that can help to better understand users or to adapt a system to a user's needs and preferences. However, usage logs can be huge, especially on the Web, recording each single user click or query. In this

[1] As the main reference to refer to USEWOD2011 or the USEWOD dataset, please use [4]. The workshop site is available at `http://data.semanticweb.org/usewod/2011/`, the proceedings have been published online at `http//arxiv.org`.

workshop we will explore the question if semantic data can be employed to generalise over clicks and queries and allow analysis on a higher level of abstraction. Now that more and more explicit knowledge is represented on the Web, in the form of ontologies, folksonomies, or linked data, the question arises how these semantics can be used to aid large scale web usage analysis and mining.

Second, usage data analysis can enhance semantic resources as well as Semantic Web applications. Traces of users can be used to evaluate, adapt or personalise Semantic Web applications. Since logs record real-life users, they provide an opportunity to create gold standards for search or recommendation tools. In addition, logs can form valuable resources from which knowledge (e.g. in the form of ontologies or thesauri) can be extracted bottom-up.

Also, the emerging Web of Data demands a re-evaluation of existing usage mining techniques; new ways of accessing information enabled by the Web of Data imply the need to develop or adapt algorithms, methods, and techniques to analyse and interpret the usage of Web data instead of Web pages. An important question at this time is how the Web of Data is being used: how are datasets being accessed by human users and how by machines, what kinds of queries are being performed, and what can we learn about the usage of semantic applications? Ultimately only understanding of their usage (or non-usage!) can give both summative and formative evaluations of the adequacy of resources for their final destination: their use in information processing, whether by people or by machines.

The primary goals of this workshop were to foment a new community of researchers from various fields sharing an interest in usage mining and semantics and to create a roadmap for future research in this direction. Several recent papers indicate an interest in bringing usage log analysis and semantic technologies together [2, 3, 5, 8, 9, 12, 13, 15, 16], but the research communities have so far remained isolated. USEWOD2011 forms an ideal opportunity for cross-fertilisation between the different communities and research fields. The workshop is of interest to researchers from the Semantic Web community who are working with usage data. It is also relevant for people working on log analysis and data mining from the perspective of Information Science, Human Computer Interaction and User Modelling. Also, in the Information Retrieval community there is a growing interest in both usage data and lightweight semantics, such as Linked Data or folksonomies. Finally, we encouraged the participation of industry representatives, as they play the role of data providers and would therefore ultimately benefit from the outcomes of the workshop.

The following set of topics was suggested for this workshop:

- Analysis and mining of usage logs of semantic resources and applications.
- Inferring semantic information from usage logs.
- Methods and tools for semantic analysis of usage logs.
- Representing and enriching usage logs with semantic information.
- Usage-based evaluation methods and frameworks; gold standards for evaluation of semantic web applications.
- Specifics and semantics of logs for content-consumption and content-creation.
- Using semantics for recommendation, personalisation and adaptation.
- Usage-based recommendation, personalisation and adaptation of semantic web applications.
- Exploiting usage logs for semantic search.

- Data sharing, privacy, and privacy-protecting policies and techniques.

The workshop was opened by a keynote talk given by Dr. Markus Strohmaier on the topic of *Social Computation for the Web of Data: Motivation, Examples, and Outlook* (Sect. 2). This was followed by a series of talks presenting the various accepted submissions, including two full paper and two short papers (Sect. 3), as well as two presentations which were related to the USEWOD data challenge (Sect. 4). Rounding up the day was a plenary discussion on obstacles holding back research in the area and future research directions (Sect. 5).

2 Keynote Talk

Dr. Strohmaier gave his well-received keynote on the topic of *Social Computation for the Web of Data: Motivation, Examples, and Outlook*. Dr. Strohmaier is currently an assistant professor at the Knowledge Management Institute, Faculty of Computer Science at Graz University of Technology in Austria. He was invited as a renowned expert on social streams and social tagging systems, and has given many talks and published extensively in this area. The following paragraph presents the abstract of his talk.

> "Today, the early World Wide Web of documents is evolving into a web of data, where data is stored in different kinds of databases, including unstructured, semi-structured and structured repositories. This talk explores the role of social computation, i.e., the combination of social behavior and algorithmic computation, for linked data. What distinguishes social-computational systems from other types of systems is the unprecedented involvement of data about user behavior, goals and motivations into the software systems structure. What can be observed in social-computational systems is that the interaction between a user and the system is mediated by the aggregation of explicit or implicit data from other users. This is the case with systems where, for example, user data is used to suggest search terms (e.g., Google Autosuggest), to recommend products (e.g., Amazon recommendations), to aid navigation (e.g., tag-based navigation) or to filter content (e.g., Digg.com). This makes social-computational systems a novel class of software systems and unique in a sense that potentially essential system properties and functions are dynamically influenced by aggregate user behavior."

3 Technical and Position Papers

Fortuna et al. [6] presented a full technical paper entitled *User Modeling Combining Access Logs, Page Content and Semantics*, in which the authors investigate the combination of semantic and non-semantic access log sources to Web resources in order to model the user. The authors' proposed approach is evaluated on real-world data. The results of their study indicate that their approach allows for building more accurate user models.

In *User Modeling Combining Access Logs, Page Content and Semantics*, Hollink and De Vries [10] propose a query modification assistant based on usage log mining. Their approach exploits semantic relations between instances appearing in usage logs in order to provide more effective query suggestions to the user. The authors identify six possible situations in which providing a query modification mechanism may improve the users' search experience. Their initial experimental results indicate that semantic information can complement traditional term-based query modification approaches.

Arish Sureka [17] presented a position paper entitled *Mining User Comment Activity for Detecting Forum Spammers in YouTube*, in which detection of spam comments in content-related sites is investigated. The author proposes an approach in which the logged activity of a user is analysed in order to detect specific patterns that could indicate suspicious activity. An empirical analysis is performed over YouTube, which demonstrates that the proposed method can be effective for comment spammer detection.

In *U-Sem: Semantic Enrichment, User Modeling and Mining Usage Data on the Social Web* Abel et al. [1] investigate the problem of obtaining and mining usage data for Social Web applications. Their work presents a framework for the semantic enrichment and mining of user profiles from usage data obtained from such applications, by incorporating entity extraction, entity identification and topic detection techniques.

4 Data challenge

USEWOD 2011 presented a data challenge to stimulate the exchange of ideas and methods between the workshop participants and to relieve some of the difficulties of obtaining real-life usage data. Three months before the workshop we released two datasets[2] of server log files of Linked Open Data servers: Semantic Web Dog Food (SWDF: `http://data.semanticweb.org`) and DBpedia (`http://dbpedia.org/`). The DBpedia logs consist of several months of log data from the linked data twin of Wikipedia, one of the focal points of the Web of data. SWDF is a constantly growing dataset of publications, people, and organisations in the Web and Semantic Web area, covering several of the major conferences and workshops, including WWW, ISWC and ESWC, and more recently also the Dublin Core conference. These datasets represent a new category of web usage data, namely usage of the Web of Data. Participants were invited to come up with analyses, applications, alignments, etc. for these datasets.

The DBpedia logs consist of all requests to the DBpedia server made at 23 days between 2009/07/01 and 2010/02/01. The Semantic Web Dog Food logs comprise all requests between 2008/11/01 and 2010/12/14 (some days missing due to server outage). Both data sets consist of logs of human users as well as machines. They contain requests for both single RDF or HTML documents and SPARQL queries. Some statistical details of the datasets can be found in Tab. 1; much more extensive details about a previous version of the dataset can also be found in [14]. Logs are anonymised by replacing the IP address fields with ‘0.0.0.0’. A hash of the original IP is appended to the log entry to enable participants to identify which requests came from the same requestors. To allow basic location-based analyses, a field with the country code of the original IP has been appended to the log entry. The mapping from IPs to country codes was done using the GeoLite Country API[3].

At the time of writing, the dataset has been requested and downloaded by 19 different research groups all over the world. Eight papers about the challenge data were submitted to the workshop, two of which were accepted for presentation at USEWOD.

Markus Kirchberg from HP Singapore presented a paper titled *From Linked Data to Relevant Data — Time is the Essence* [11]. In this paper the authors aim to predict the

[2] Details about the datasets at `http://data.semanticweb.org/usewod/2011/challenge.html`

[3] `http://www.maxmind.com/app/geolitecountry`

Source	Time period	# Days	Size	# Requests	# SPARQL requests
DBpedia	2009/07/01–2010/02/01	23	7.476 GB	19,770,157	5,268,125
		(per day)	0.325 GB	859,572	229,049
SWDF	2008/11/01–2010/12/14	720	2.173 GB	8,092,552	2,287,973
		(per day)	0.003 GB	11,240	3,178

Table 1: Statistics of the USEWOD 2011 data set

relevance of Linked Data entities at a given moment in time. They present a novel method combining link analysis and usage analysis to produce time-windowed visualisations of the usage of an entity. They identify three important properties for predicting the relevance of an entity: the number of links between the entity and other entities, the depth of the traversals, and the number of times a path was accessed within a time window.

The second accepted paper that used the challenge data was by Mario Arias Gallego, Javier D. Fernández, Miguel A. Martínez-Prieto, and Pablo de la Fuente: *An Empirical Study of Real-World SPARQL Queries* [7]. The paper describes an in-depth analysis of the SPARQL queries in the USEWOD data set. From the analysis, the authors conclude that most SPARQL queries are simple and include few triple patterns and joins. The most common join types are Subject-Subject, Subject-Object and Object-Object. The graph patterns are usually star-shaped. Chain-shaped patterns are rare and generally very short: 98% of the queries had a length of only one triple pattern. The authors show how this knowledge about real-life SPARQL queries can be applied for optimising the performance of query evaluation engines and RDF stores.

Research using the USEWOD data sets mainly focussed on the analysis of the log data. The underlying linked data was explored in much less detail. The challenge participants as well as the audience were convinced that taking these data into account will bring usage analysis to the next level.

5 Plenary discussion

During the plenary discussion we discussed the main obstacles holding back research in the USEWOD area. One problem appeared to be dominant in this field: the absence of a sizeable amount of publicly available usage data. Usage data that is used in research projects usually cannot be shared outside these projects, which prohibits comparing results.

To make more log data available to the community, we need to convince organisations that publish data online to 1) record logs and 2) publish them. Efforts made within the Linked Data community can serve as an example as this community has succeeded in persuading many organisations to publish their data in the Linked Data cloud. For making usage data available, the most promising organisations are non-profit organisations publishing information on non-privacy sensitive topics, for instance cultural heritage institutions. The main incentive for these organisations is the perspective that research on usage data of a system can yield insights that can help to improve the system. Once some organisations will have been persuaded to published their logs, there may be enough momentum to convince other organisations as well, as happened with Linked Open Data.

We also discussed issues related to the storage of usage data. RDF was suggested as a data format. For anonymisation a standard tool, such as an Apache plug-in, should be developed that allows organisations to anonymise the logs in-house in a standardised way. For sharing log files a central logfile repository is needed. The USEWOD organisation committed themselves to setting up such a repository.

The combination of usage data and the web of data provides a wide variety of unexplored research directions. For instance, participants mentioned that analysis of usage data can be enhanced by using location information. Usage data can also be used to detect errors in the web of data or to identify concepts and entities that are no longer useful. Moreover, as the web of data is growing ever larger, ranking entities by relevance becomes increasingly important. Usage data can be a key feature in such relevance ranking.

6 Conclusions

The research presented in the USEWOD workshop as well as the lively discussion afterwards, shows the importance and timeliness of research on the combination of usage data and the web of data. Therefore, the organisers are planning a follow-up workshop next year. A new and extended dataset should be provided for a second data challenge, and as a resource for the community. In the meantime, the current USEWOD data set will remain available. As the absence of logs is viewed as the main problem for research in this area, the USEWOD data set can make an important contribution to the field.

Acknowledgements We were fortunate to have a highly qualified program committee; we thank them for their contribution to making the workshop a success. We would also like to explicitly thank our keynote speaker Markus Strohmaier, as well as our two sponsors: the LATC project[4], which sponsored the best paper award, and Gridline[5], which sponsored the USEWOD2011 data challenge.

References

[1] F. Abel, I. Celik, C. Hauff, L. Hollink, and G.-J. Houben. U-Sem: Semantic enrichment, user modeling and mining usage data on the social web. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hydebarabad, India*, 2011.

[2] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California*, 2007.

[3] D. Benz, B. Krause, G. P. Kumar, A. Hotho, and G. Stumme. Characterizing semantic relatedness of search query terms. In *1st Workshop on Explorative Analytics of Information Networks, Bled, Slovenia*, 2009.

[4] B. Berendt, L. Hollink, V. Hollink, M. Luczak-Rösch, K. H. Möller, and D. Vallet. USEWOD2011 — 1st international workshop on usage analysis and the web of data. In *20th International World Wide Web Conference (WWW2011), Hyderabad, India*, pages 305–306, March-April 2011.

[4] http://latc-project.eu/

[5] http://www.gridline.nl/

[5] B. Berendt, G. Stumme, , and A. Hotho. Usage mining for and on the Semantic Web. In *H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (Eds.), Data Mining: Next Generation Challenges and Future Directions*, pages 461–480, 2004.

[6] B. Fortuna, D. Mladenic, and M. Grobelnik. User modeling combining access logs, page content and semantics. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hydebarabad, India*, 2011.

[7] M. A. Gallego, J. D. Fernández, M. A. Martínez-Prieto, and P. D. L. Fuente. An empirical study of real-world SPARQL queries. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hydebarabad, India*, 2011.

[8] K. Hofmann, M. de Rijke, B. Huurnink, and E. Meij. A semantic perspective on query log analysis. In *Working Notes for CLEF*, 2009.

[9] V. Hollink, T. Tsikrika, and A. P. de Vries. The semantics of query modification. In *Proceedings of 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO)*, 2010.

[10] V. Hollink and A. D. Vries. Towards an automated query modification assistant. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hydebarabad, India*, 2011.

[11] M. Kirchberg, R. K. L. Ko, and B. S. Lee. From linked data to relevant data - time is the essence. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hydebarabad, India*, 2011.

[12] E. Meij, M. Bron, L. Hollink, B. Huurnink, and M. de Rijke. Learning semantic query suggestions. In *8th International Semantic Web Conference, Chantilly, VA, USA.*, 2009.

[13] P. Mika, E. Meij, and H. Zaragoza. Investigating the semantic gap through query log analysis. In *8th International Semantic Web Conference, Chantilly, VA, USA.*, 2009.

[14] K. Möller. *Lifecycle Support for Data on the Semantic Web.* PhD thesis, National University of Ireland, Galway, 2009.

[15] K. Möller, M. Hausenblas, R. Cyganiak, and G. Grimnes. Learning from linked open data usage: patterns & metrics. In *WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[16] T. Sakai and K. Nogami. Serendipitous search via wikipedia: a query log analysis. In *32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA*, 2009.

[17] A. Sureka. Mining user comment activity for detecting forum spammers in YouTube. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hydebarabad, India*, 2011.