Analysing Wikidata Edits Over Time

Short Term Scientific Mission (STSM)

Applicant: Cristina Sarasua, University of Koblenz-Landau, Germany
Host: Gianluca Demartini, University of Sheffield, United Kingdom
COST STSM Reference Number: COST-STSM-IC1302-35438
Period: 2016-10-01 to 2016-10-28

Purpose of the STSM

Wikidata [1,2] is a free, open, and general-interest knowledge graph that is collaboratively developed and maintained by a community of thousands of volunteers.

Currently, there is little knowledge about the way the knowledge base is evolving and the way users edit the data over time. Mueller-Brin et al. [4] studied the extent to which editors accomplish each of the major editing tasks (e.g. edit a reference, create an item, define a property, etc.). Steiner [5] compared the edits made by humans and bots. Still, there are many open questions that need to be answered, not only to understand human behaviour in such a Web-based crowdsourcing environment, but also to identify weaknesses and strengths in Wikidata, and consequently identify new research opportunities and ways to improve it.

The goal of the short term scientific mission was to work on the analysis of Wikidata edits. We are interested in analysing the evolution of Wikidata editors in order to identify key factors that influence productive editors, and editors who contribute for a long time.

Description of the work carried out and main results

We have processed the edit history of Wikidata (as of July 2016). We obtained the XML data dump provided by Wikimedia containing information about each of the editing actions done by contributors [6], parsed it, transformed it into CSV data, and imported into a memory-based database.

Since our ultimate goal is study how contributors edit the knowledge graph over time, we classify the edits based on the (i) type of editor, (ii) the type of thing edited and (iii) the means used to edit.

- *Type of editors*: we distinguish between users who are registered users (i.e. have a username and edit Wikidata while being logged in) and users who are anonymous (and from whom we only know an IP). Registered users can be humans or bots. We identify bots by looking up the public list of registered bots and discard the edits done by this set of users, because we are primarily interested in human behaviour. It is important to distinguish between registered and anonymous users, not only because people might behave differently when they reveal their identity, as Shih-Wen et al.

showed in the context of microtask crowdsourcing [7], but also because non-registered edits might also come from applications implementing automatic edits via the Wikidata API.

- *Type of things edited*: we distinguish between item edits and non-item edits. Item edits are edits done to create, update or delete an item of the knowledge base (e.g. an entity, a class). Non-item edits are edits done in other kind of pages such as project and user pages. We only focus on item edits.
- *Means to edit*: there are various interfaces to edit Wikidata (e.g. the wiki, the Wikidata games). We differentiate between edits done using tools and edits done without tools, because in the former case users do not completely decide what item to work on, nor the type of edit to do. To classify edits into these two groups we use the tags database provided by Wikimedia and scan the edit comments for any other trace left by tools listed in Wikimedia directories.

All through our analysis we use this classification to compare edits done without tools by registered users, edits done without tools by anonymous users and edits done with tools. We found out that from the complete set of 350+ million edits, 86+ millions are done by the set of registered and anonymous users. As expected, the number of edits done with tools is bigger than the number of edits done manually (see Table 1). The number of distinct editors present in the edits without tools and not registered is higher than in the other two groups. An explanation to this fact might be that there are sporadic users who do not want to commit to the registration process, and also, one user might edit from different IPs over time.

	Without tools Not registered	Without tools Registered	With tools
# edits	1,599,178	34,426,108	50,457,132
# items edited	715,242	7,725,514	12,459,789
# distinct editors	365,692	137,694	3,057

Table 1 - Different types of edits

We analysed the number of edits done by editors (Figure 1(a)), as well as the number of editors per item (Figure 1(b)), and in both cases we observed a power law distribution. There are few editors who have edited a large number of edits and many editors who have edited a small number of edits. Likewise, there are few items who have been edited by many editors, and many items who have been edited by few editors.

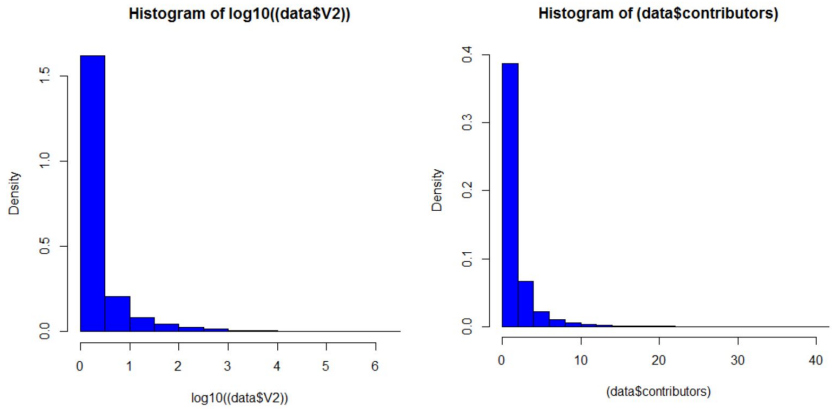

Figure 1 - (a) Number of edits per editor (b) Number of editors per item

We also looked at the number of editors who joined Wikidata, the number of editors who stopped editing in Wikidata and the number of editors who started and stopped editing Wikidata each of the years that the project is running. We see that after the second year of the project there was an increase in the number of editors joining, but after that increase the number of new editors joining decreased. However, the number of editors who stopped editing had a similar evolution: the number decreased since 2014.

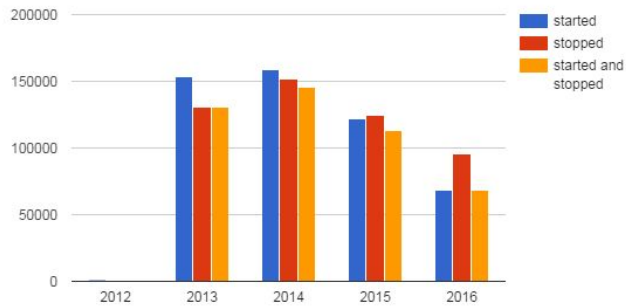

Figure 2 - Number of editors joining and leaving Wikidata per year

A common practice followed in Web log mining is to group user actions (e.g. queries) into user sessions [8]. We decided to apply this technique to the set of edits in Wikidata, as it allows us to analyse dedication patterns and connect actions done close in time. In order to group the edits into sessions, we computed the seconds between pairs of consecutive edits and followed the method proposed by Geiger et al. [9] to validate the 60 min. threshold for defining sessions in Wikipedia. We plotted the histogram of time between pairs of Wikidata edits and fitted two lognormal and one exponential gaussian distributions into the data. We observed that there is a first peak in very small time differences (between 0 and 3 seconds), a second peak at around 10 seconds and a third peak at around one day. Following the interpretation of Geiger et al. in the Wikipedia data, we consider that the edits around the first two peaks are intra-session edits, while the edits around the third peak are inter-session edits. We looked at the intersection between the second and third distributions and

computed the threshold that determines how sessions are defined. In the case of edits without tools and registered users (Figure 3), the threshold is at approximately 4 hours. That means, that for every pair of edits that is separated by 4 hours, we generate a new session. We observe that there is a difference in the distributions that we obtain from Wikidata, and those that Geiger et al. obtained from Wikipedia: the first peak present at the Wikidata plot is not present in the Wikipedia plot. We have analysed the data and we see that there pairs of actions done within a very short time (both in intra- and inter-items). One possible explanation to this fact is that in Wikidata users edit structured data, very specific pieces of information and users may edit multiple items and multiple claims at the same time and save all updates at the end of the process.

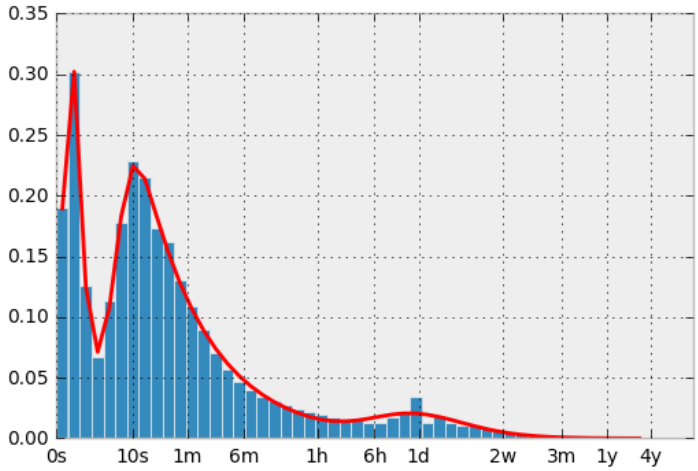

Figure 3 - Threshold generation for Wikidata edit sessions

The histogram of the lifespan of users (Figure 4) shows that there many users whose lifespan is within 1 month. That means that the user just edited during one single month. The other peak is around 3 years.

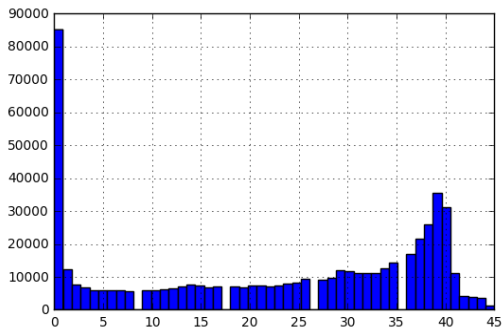

Figure 4 - Lifespan of users

Additionally, we developed software to extract transition matrices that indicate the pairs of entities, and the pairs of types of entities that users edit consecutively. We used Graphistry[1] to visualize the matrices.

[1] https://www.graphistry.com/

We also developed infrastructure to compute descriptive statistics of the graph, to identify for example, the number of types, the number of property-values, the number of Wikipedia categories and the number of links per entity. We worked with the JSON dump of the knowledge base provided by Wikimedia [10]. Obtaining these statistics is important to understand the context of the number of edits done by users at each point in time.

We worked closely with Alessandro Checco, researcher at the University of Sheffield, UK. The STSM enabled a collaborative workflow that we continue remotely at the moment, and the progress that we made in our work was very beneficial for the materialization of the scientific publication that we are working on.

During the STSM Cristina Sarasua (STSM applicant) gave an invited talk about Wikidata and the ongoing work at the Mixed Reality Lab in the University of Nottingham, UK. This short trip was a good opportunity to discuss our work with Neha Gupta, a crowdsourcing researcher of the university of Nottingham, who has experience in running ethnographic studies in the field of microtask crowdsourcing. We plan to collaborate in order to complement our data-driven research with a qualitative study.

Future collaboration

- We are currently working on a full-paper submission to the 11th International Conference on Web and Social Media (ICWSM2017), in Montreal, Canada.
- We plan to continue our collaboration and extend our work with further research ideas that we have developed during the STSM. We will also organize the second edition of a research meeting about crowdsourcing at the Dagstuhl Castle,Germany, in 2017[2].

Confirmation by the host institution

See letter in the attachment.

Cristina Sarasua
Koblenz, 25.11.2016

[2] http://www.dagstuhl.de/de/programm/kalender/evhp/?semnr=17183

References

[1] Denny Vrandečić and Markus Krötzsch. (2014). Wikidata: a free collaborative knowledge base.Communications ACM 57, 10 . http://dl.acm.org/citation.cfm?id=2629489

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page

[3] Wikidata Statistics: https://www.wikidata.org/wiki/Wikidata:Statistics

[4] Müller-Birn, C., Karran, B., Lehmann, J., and Luczak-Rösch, M.. (2015). Peer-production system or collaborative ontology engineering effort: what is Wikidata?. In: OpenSym 2015. http://dl.acm.org/citation.cfm?id=2789836

[5] Steiner, T. (2014). Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata. In: OpenSym 2014. http://dl.acm.org/citation.cfm?id=2641613&dl=ACM&coll=DL&CFID=668772118&CFTOKEN=30542021

[6] Dumps with Wikidata's edit history. 01/07/2016. https://dumps.wikimedia.org/wikidatawiki/20160701/

[7] Shih-Wen H. and Wai-Tat F. 2013. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13). ACM, New York, NY, USA, 621-630. DOI: http://dx.doi.org/10.1145/2470654.2470743

[8] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. Acm Sigkdd Explorations Newsletter, 1(2), 12-23.

[9] Geiger, R.S. and Halfaker, A. 2013. Using edit sessions to measure participation in wikipedia. In Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13). ACM, New York, NY, USA, 861-870. DOI=http://dx.doi.org/10.1145/2441776.2441873

[10] Downloadable Wikidata item descriptions. https://www.wikidata.org/wiki/Wikidata:Database_download