Wiki-Talk Datasets

Jun Sun and Jérôme Kunegis

April 2016

Abstract

We describe our Wiki-talk datasets, which consist of the user interaction networks of all user talk pages in Wikipedia, in 28 languages. Each user is represented by her original Wikipedia user ID, and is assigned a role, according to her access level in Wikipedia. We also show how to use the parser provided by us to keep the data up-to-date and how to customize the datasets.

1 Description

In Wikipedia, each registered user has a talk page that can be used for discussion. We extract the user interaction networks of all user talk pages of Wikipedia in the 28 languages with the highest number of articles (at the time of dataset creation). The 28 languages are, listed alphabetically by their ISO 639 code: ar bn br ca cy de el en eo es eu fr gl ht it ja lv nds nl oc pl pt ru sk sr sv vi zh. Each language forms an individual directed network, in which each node is a Wikipedia user represented by her original Wikipedia user ID, and each directed edge `(User_ID_A, User_ID_B, timestamp)` represents a user interaction: User A wrote a message on User B's talk page at a certain time. Each user has an access level [1], which we interpret as the following roles:

- **Administrator.** Administrators refer to the accounts that have high level of access to contents and maintenance tools in Wikipedia. We combine the users that are granted as "sysops" or "bureaucrat" by the communities at RfA or RfB[1], and label them as administrators.
- **Bot.** Bots are used in Wikipedia for (semi-)automatically improving contents. Bot accounts are marked as "bot" by an administrator, and each has specific tasks that it performs [2].
- **Normal user.** Other users that are not catagorised as administrators or bots.

1.1 Insights

Table 1 shows some basic statistics of the datasets. All sub-datasets are denoted as their language codes, e.g., `de` stands for the data from the German Wikipedia. As we can see, all 28 networks have a variety of sizes, from 504 (`ht`) to around 3 million (`en`). The proportions of both bots and administrators are very small, althought they vary highly among all sub-datasets, from 0.0027% to 5.97% and from 0% to 0.72% respectively.

More detailed information and plots of the datasets can be found on the KONECT [3] website[2].

2 Downloading and Parsing

We parsed the Wikipedia dump files (xml) to Wiki-talk networks at the end of 2015. You can download the parsed datasets at `http://dx.doi.org/10.5281/zenodo.49561`. We have also open sourced the parsing tool[3] which we wrote in Clojure, in case you want to re-parse the datasets, or customize them, such as adding new roles.

[1] Requests for adminship (RfA), Requests for bureaucratship (RfB)
[2] http://konect.uni-koblenz.de/
[3] https://github.com/yfiua/wiki-talk-parser

2.1 Parsing with Stu

Use Stu[4] for easy lives. The only file you need is `main.stu`. Simply type in `stu` or:

```
nohup stu -k -j 3 &
```

Stu will automatically start downloading this program and the dump files and parse them.

2.2 Parsing without Stu

- **Installation**
  Manually download the latest jar files from the "release" page. You need to download the Wikipedia dump files in xml format from its website[5], too.
- **Parse**
  ```
  java -jar parser.jar *input-file* *lang* > *output-file*
  ```
- **Shrink**
  "Shrink" the resulted network, so to make unweighted directed networks without loops, as in the SNAP datasets [4].
  ```
  java -jar shrinker.jar *input-file* > *output-file*
  ```
- **Group users**
  Group users according to their roles.
  ```
  java -jar grouper.jar *input-file* > *output-file*
  ```
- **Compilation** (optional)
  Comple the parsing tool.
  ```
  lein with-profile parser:shrinker:grouper uberjar
  ```

3 License

References

[1] Wikipedia, "User access levels — Wikipedia, the free encyclopedia," 2016, [Accessed 13-July-2016]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wikipedia:User_access_levels&oldid=727508916

[2] ——, "Bot policy — Wikipedia, the free encyclopedia," 2016, [Accessed 25-July-2016]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wikipedia:Bot_policy&oldid=730456542

[3] J. Kunegis, "Konect: the koblenz network collection," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 1343–1350.

[4] J. Leskovec and A. Krevl, "SNAP Datasets:Stanford large network dataset collection," 2015.

[5] J. Sun and J. Kunegis, "Wiki-talk datasets," Apr. 2016. [Online]. Available: http://dx.doi.org/10.5281/zenodo.49561

[4] https://github.com/kunegis/stu
[5] https://dumps.wikimedia.org/

Table 1: Meta Information of Wiki-talk Datasets

Lang	# Nodes	# Edges	# Bots	# Admins	% Bots	% Admins
ar	1095799	1913103	30	37	0.0027%	0.0034%
bn	83803	122078	18	16	0.0215%	0.0191%
br	1181	13754	43	8	3.6410%	0.6774%
ca	79736	351610	179	25	0.2245%	0.0314%
cy	2233	10740	39	16	1.7465%	0.7165%
de	519403	6729794	328	246	0.0631%	0.0474%
el	40254	190279	58	20	0.1441%	0.0497%
en	2987535	24981163	278	1313	0.0093%	0.0439%
eo	7586	47070	130	21	1.7137%	0.2768%
es	497446	2702879	34	75	0.0068%	0.0151%
eu	40993	58120	81	10	0.1976%	0.0244%
fr	1420367	4641928	97	163	0.0068%	0.0115%
gl	8097	63809	12	14	0.1482%	0.1729%
ht	536	1530	32	0	5.9701%	0.0000%
it	863846	3067680	137	104	0.0159%	0.0120%
ja	397635	1031378	51	49	0.0128%	0.0123%
lv	41424	73900	57	11	0.1376%	0.0266%
nds	23132	27432	56	5	0.2421%	0.0216%
nl	225749	1554699	237	50	0.1050%	0.0221%
oc	3144	11059	51	4	1.6221%	0.1272%
pl	155820	1358426	55	115	0.0353%	0.0738%
pt	541355	2424962	205	64	0.0379%	0.0118%
ru	457017	2282055	77	90	0.0168%	0.0197%
sk	41452	131884	105	8	0.2533%	0.0193%
sr	103068	312837	132	22	0.1281%	0.0213%
sv	120833	598066	41	72	0.0339%	0.0596%
vi	338714	607087	123	23	0.0363%	0.0068%
zh	1219241	2284546	93	77	0.0076%	0.0063%