Procedure and technical details of the scripts for platform extraction

Platform extraction from the directories

We produced automated scripts for the extraction of P2P SCC Websites from the directories Collaborative Consumption (http://www.collaborativeconsumption.com/), Compare and Share (http://www.compareandshare.com/)[1], Mesh (http://meshing.it/) and The People Who Share (http://www.thepeoplewhoshare.com/). These scripts are Java-based and internally parse the XML structure of the Webpages to extract information. They generate .csv files with the platform names and their corresponding URLs, which is taken as input for consolidation scripts.

Consolidation of platforms

Since the directories might potentially list duplicate Websites, or Websites that have already been extracted in a previous iteration, it was important to merge the extraction results. This is done via semi-automated R-scripts, processing the following steps:

1. URLs are cut off after their country code, since some directories also list the path and the query part of the URL.
2. If the URLs are the same, the Websites are assumed to be duplicates, and one is removed from the set.
3. Websites with exactly the same name but different trimmed URLs have to be checked manually, whether they are duplicates. Sometimes, e.g., different country codes are stored in the directories that redirect to the same Website. Manual checking is possible since this does not apply to more than a dozen Websites. For the purpose of manual checking, a .csv file with the potentially duplicate websites is created, listing the name and the URL of the first Website and the name and URL of the potential copy. Duplicates have be marked, the file then is reloaded by the script, and true duplicates eliminated.
4. Then, fuzzy string matching (edit distance) is performed, where perfect copies and almost perfect copies of one another's name are assumed to be duplicates.
5. The new set of Websites is compiled from the remaining non-duplicate Websites and compared to the list from the previous iteration. All truly new Websites are then saved to a file.

The script has be executed step-by-step in an R-environment, e.g. R-Studio. Steps 1, 2 and 3 can be executed automatically. Then, the manual actions described in 3) have to be conducted. After that, steps 4 and 5 are again executed automatically. These scripts aim at preventing any duplicates from reaching the list of new platforms. They therefore should not introduce any bias, except for the bias of platforms already existing on the directories.

Extraction of traffic data from Alexa

The traffic data query script is again written in Java. It parses the XML result from the Alexa API and creates a .csv file for every platform queries that contains the views for the past month (averaged over all days) and by countries. Please note that only Websites with at least 50,000 visitors have a value greater than 0 by default.

[1] Not available anymore.