PILOT PROJECT ON BIG DATA

Use of Wikipedia page views on World Heritage Sites

1 Description of the source

Wikipedia was founded in 2001 with the objective of creating a free online public editable encyclopaedia. Between 2001 and 2016 it has grown to 38 million articles in 246 languages. It is widely used with 21 million page views per hour reported in May 2016 ([1]). According to the Community survey on ICT usage in households and by individuals, in 2015, 45 % of individuals of 16 to 74 years old living in the EU consulted wikis to obtain knowledge (e.g. Wikipedia). This was 66 % for individuals amongst 16 and 24 years old.

While using Wikipedia, people leave digital traces of their activities, in particular as a result of accesses to and editions of Wikipedia articles and their corresponding discussion pages. These digital traces exist as data in the web logs of the servers which host Wikipedia and in the content of the articles and their corresponding discussion pages.
Information on the accesses, or consultations, of Wikipedia articles include the identification of the article itself, the time of the view and where the person was located when consulted the article, all of which exist in the weblogs. Detailed data on the number of page views per article (excluding information on where the access originated) is made publicly available by the Wikimedia Foundation - the organization which supports and hosts Wikipedia.
There is additional information about the consulted articles that enriches the data on the page views. Firstly, this is the language version of the Wikipedia to which the article belongs. Then, the textual content of the article itself provides relevance information (e.g. the level of detail of the information it provides). Other content information such as the categorization of the article and information boxes (infoboxes) provides more structured data which can be used.

The pilot project used data on the number of page views per month for all articles in 31 language versions of Wikipedia ([2]). The data used is made available publicly by the Wikimedia Foundation. The version of Wikipedia designed for mobile devices was not included in the data source.

2 Methodology

2.1 Data used

From the several data sources available on the use of Wikipedia, the number of page views and the content of the articles were used. The number of page views is made available by the Wikimedia Foundation as dump files in several formats. The project has used monthly files with the number of page views per hour for each article in the several wiki projects of the Wikimedia Foundation, which include besides the several language versions of Wikipedia also page views of another 10 wiki projects (e.g. Wikibooks, wikinews). The definition of page view in these datasets

[1] From 'https://en.wikipedia.org/wiki/Special:Statistics' consulted on 20th May 2016.
[2] 31 Language versions cover 24 official EU languages and as well Icelandic, Macedonian, Norwegian, Russian, Albanian, Serbian and Turkish.

include accesses to web pages of Wikipedia articles excluding the mobile site and accesses identified as done by non-humans (i.e. bots). It is also corrected for outages (i.e. when the Wikipedia servers are not available).
This pilot was run in the context of the Big Data Sandbox, an international collaboration project sponsored by the High-Level Group for the Modernisation of Official Statistics, set up by the Conference of European Statisticians. It involved, besides Eurostat, several national statistical institutes and other international statistical bodies.

2.2 Data pre-processing

In order to deal with the large file sizes involved a pre-processing pipeline was developed incorporating a number of technologies and ultimately based on the Hadoop ([3]) platform. The original files are available in a space-compressed form and the purpose of the pre-processing is to decode them into a form that is suitable for analysis. The total size for the years on which the analysis was done was around 800 GB in the compressed form. From these files the hourly time series per language version of Wikipedia were extracted. The sizes of these extracted uncompressed data sets vary a lot depending on the language version with English being the largest at around 820 GB. From these time series extractions were performed, aggregating to monthly frequency based on the identification of articles outlined below.

2.3 Initial selection of articles in the English Wikipedia

The categorization feature of Wikipedia was used as a first step to identify articles in the English Wikipedia related to world heritage sites by selecting those categorized as "World heritage sites by continent" and any of its subcategories. In a second step information in the infobox "World heritage site" was used, when it was present, to extract the identifier of the particular site to which the article refers.
These methods allowed linking at least one English Wikipedia article to around 90% of the 1031 sites inscribed until 2015. After the initial automated process, the results were assessed and validated manually. This manual process allowed associating English Wikipedia articles to 1025 of the 1031 world heritage sites. A total of 1362 articles were selected, with the number of articles associated to each article ranging from 0 to 17.

2.4 Extension of the selection of articles to other languages

In order to get the articles in the language versions other than English, the articles linked to each of the articles previously selected in the English Wikipedia were taken.

[3] http://hadoop.apache.org/