Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt WS 2021/22

Wikipedia Cleanup: Recognizing Stale Data

Keywords: data quality assessment, data cleansing, time series prediction, rule mining

Change Exploration

As the world evolves and events take place, data and metadata of many public datasets change with it. When we monitor real-world datasets, such as Wikipedia or various open government datasets, we observe a large number of many different kinds of changes. The changes range from small-scale updates in a few records to large-scale schema changes, such as column deletions or bulk-inserts. Keeping a database up-to-date is often a labor-intensive task. Yet not all things change at the same frequency and require the same attention in terms of updates. Some things should change several times a day, such as oil prices or tournament data, others change several times a week, such as sports statistics or chart positions, and others change only every few months or years, such as public office-holders or population statistics. And then, of course, there are things that are static and should (almost) never change, such as dates-of-birth or city names.

In this master-project, we shall leverage the vast change history of Wikipedia to predict expected change behavior in the future. For instance, we expect population statistics to change once a year, US presidents every four years, and places of birth almost never. These predictions shall be done under the premise that previous changes affect the probability of future changes: we intend to evaluate probabilistic models, rule-based approaches, and deep learning models on how well they can predict the *time*, *type*, and *extent* of future changes given the changes in the past.
Such change predictions can immediately serve the following two use-cases:

- Warn about unexpected changes for manual review to combat errors
- Point out expected changes that in fact did not happen and thus warn about stale data

In both cases, we want to provide hints *why* we did or did not expect a certain change and support our claim by examples of past changes.

Data and Frameworks

The main dataset for this project contains *all* changes to Wikipedia's infoboxes – the structured summaries at the top right of articles – since 2003. This dataset comprises about 7 GB of compressed or 150 GB of uncompressed JSON data and describes about 4 million infoboxes with about 70 million changes.

To organize our work, we will use the well-established Jupyter Python Notebooks (https://jupyter.org), along with several libraries commonly used in Data Science, such as Pandas (https://pandas.pydata.org/), scikit-learn (https://scikit-learn.org/stable/index.html) and other specialized Python libraries for time-series analysis. Thus, prior knowledge in Python or ideally even in these Data-Science frameworks is helpful, but not necessary.


Exploring Change – A New Dimension of Data Analytics

Tobias Bleifuß, Hasso Plattner Institute, University of Potsdam, tobias.bleifuss@hpi.de
Leon Bornemann, Hasso Plattner Institute, University of Potsdam, leon.bornemann@hpi.de
Theodore Johnson, AT&T Labs – Research, johnson@research.att.com
Dmitri V. Kalashnikov, AT&T Labs – Research, dvk@research.att.com
Felix Naumann, Hasso Plattner Institute, University of Potsdam, felix.naumann@hpi.de
Divesh Srivastava, AT&T Labs – Research, divesh@research.att.com

ABSTRACT

1. CHANGE IN DATA AND METADATA

Project Goals

With Wikipedia Cleanup, we will follow an *entire research cycle* from problem inception to data engineering to algorithm development to evaluation. Together, we will prepare a submission to a top database conference, such as VLDB or SIGMOD together.

For questions, please contact Leon Bornemann, Tobias Bleifuß, or Prof. Dr. Felix Naumann
leon.bornemann@hpi.de, tobias.bleifuss@hpi.de, felix.naumann@hpi.de
Please also check out www.IANVS.org for more details about the Janus project.