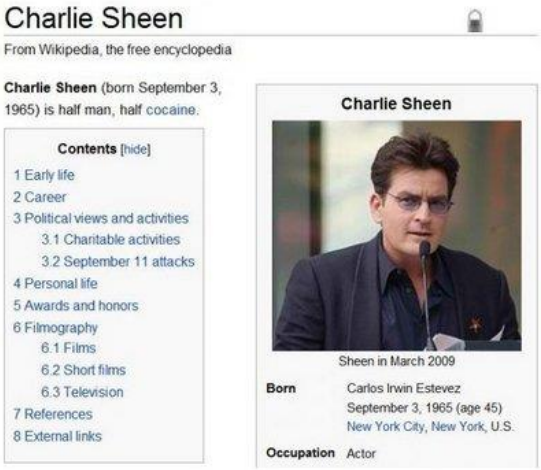
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt Sommer 2018

Vandalism Detection in Wikipedia Table Revisions

Change Exploration

Data and metadata suffer many different kinds of change: values are inserted, deleted or updated; entities appear and disappear; properties are added or re-purposed, etc. Explicitly recognizing, exploring, and evaluating such change can alert to changes in data ingestion procedures, can help assess data quality, and can improve the general understanding of the dataset and its behavior over time.



Charlie Sheen

From Wikipedia, the free encyclopedia

Charlie Sheen (born September 3, 1965) is half man, half cocaine.

Contents [hide]

1 Early life
2 Career
3 Political views and activities
  3.1 Charitable activities
  3.2 September 11 attacks
4 Personal life
5 Awards and honors
6 Filmography
  6.1 Films
  6.2 Short films
  6.3 Television
7 References
8 External links

Charlie Sheen

Sheen in March 2009

Born	Carlos Irwin Estevez September 3, 1965 (age 45) New York City, New York, U.S.
Occupation	Actor

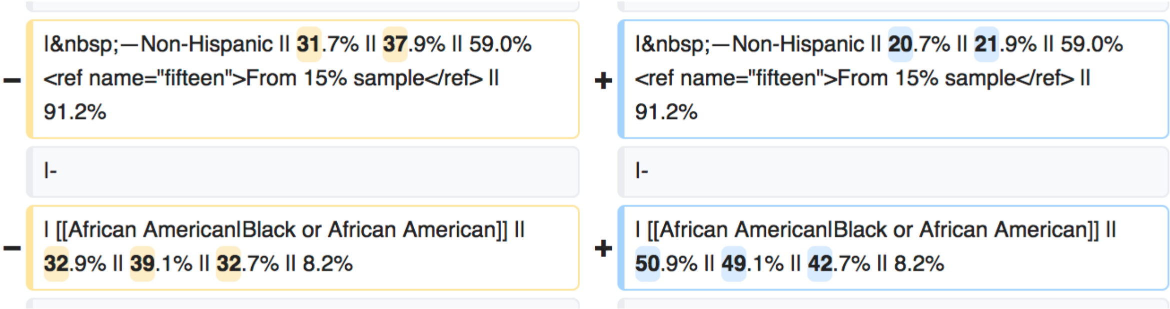
However, not all changes are genuine changes. Especially open, collaboratively edited data sources, such as Wikipedia, frequently undergo changes that deliberately introduce false values. In order to focus on the actual changes in the data one must recognize changes that introduce vandalism, so that they can be filtered out in further analyses.

Wikipedia Tables

Tables serve as a compact representation of facts and data on Wikipedia pages. The dataset is especially interesting since not only the most recent version is publicly available, but we also have their complete edit history for more than 15 years now. The English Wikipedia alone contains more than 5 million articles that are subject to millions of edits each month and many of them contain one or multiple tables. We have already extracted the history of all approximately 2 million distinct tables. Their history features more than 50 million revisions and uncompressed it is about 4TB large. Besides the actual content of the article and the edits, the dataset also comprises certain metadata, most importantly user information.

Vandalism Detection

Vandalism is a common problem in Wikipedia and different methods for detecting vandalism in Wikipedia page revisions exist. However, current methods work on a coarse granularity and have not yet been applied to structured data, stored in tables. While rather obvious variants of vandalism in tables do exist (for example the deletion of a table), vandalism within in tables can be much more subtle (for example raising the population percentage of an ethnic minority from 30% to 50%, as shown in the figure below). The goal of this master project is to develop and evaluate methods that detect and classify as many cases of vandalism in tables as possible.



```
− |&nbsp;—Non-Hispanic || 31.7% || 37.9% || 59.0%      + |&nbsp;—Non-Hispanic || 20.7% || 21.9% || 59.0%
  <ref name="fifteen">From 15% sample</ref> ||          <ref name="fifteen">From 15% sample</ref> ||
  91.2%                                                  91.2%
  |-                                                     |-
− | [[African American|Black or African American]] ||  + | [[African American|Black or African American]] ||
  32.9% || 39.1% || 32.7% || 8.2%                        50.9% || 49.1% || 42.7% || 8.2%
```

Example for vandalism in Wikipedia tables: Tampering with the proportions of ethnic minorities. [https://en.wikipedia.org/w/index.php?title=Chicago&diff=prev&oldid=654893961]

Project Goals

- Review of existing methods for vandalism detection and an assessment of their applicability to a tabular setting
- Creation of a gold standard (ground truth)
- Development of new methods, possibly based on machine learning methods
- Performance evaluation of existing and newly conceived methods
- Submission of a scientific paper with a solution to this novel problem to a top conference

Prerequisites

There are no explicit prerequisites for this project, but an interest or experience in one or more of the following research areas is helpful:

- Working with large datasets
- Machine learning
- Data integration and cleansing
- Building scalable applications

Contact and Advisors

Prof. Dr. Felix Naumann: felix.naumann@hpi.de
Tobias Bleifuß: tobias.bleifuss@hpi.de
Leon Bornemann: leon.bornemann@hpi.de