Building Automated Vandalism Detection Tools for Wikidata

Amir Sarabadani
Wikimedia Deutschland
Tempelhofer Ufer 23/24
10963 Berlin, Germany
Ladsgroup@gmail.com

Aaron Halfaker
Wikimedia Research
149 New Montgomery Street
San Francisco, USA
ahalfaker@wikimedia.org

Dario Taraborelli
Wikimedia Research
149 New Montgomery Street
San Francisco, USA
dtaraborelli@wikimedia.org

ABSTRACT

Wikidata, like Wikipedia, is a knowledge base that anyone can edit. This open collaboration model is powerful in that it reduces barriers to participation and allows a large number of people to contribute. However, it exposes the knowledge base to the risk of vandalism and low-quality contributions. In this work, we build on past work detecting vandalism in Wikipedia to detect vandalism in Wikidata. This work is novel in that identifying damaging changes in a structured knowledge-base requires substantially different feature engineering work than in a text-based wiki like Wikipedia. We also discuss the utility of these classifiers for reducing the overall workload of vandalism patrollers in Wikidata. We describe a machine classification strategy that is able to catch 89% of vandalism while reducing patrollers' workload by 98%, by drawing lightly from contextual features of an edit and heavily from the characteristics of the user making the edit.

Keywords

Wikidata; vandalism; knowledge bases; quality control

1. INTRODUCTION

Wikidata (www.wikidata.org) is a free knowledge base that everyone can edit. It is a collaborative project aiming to produce a high quality, language-independent, open-licensed, structured knowledge base. Like Wikipedia, the project is open to anyone willing to contribute productively. This also opens Wikidata to potentially damaging/disruptive contributions. In order to combat such intentional damage, volunteer patrollers work to review changes to the database after they are saved. At a rate of about 80,000 human edits and 200,000 automated edits per day (as of February 2016), though, the task of reviewing every single edit would be daunting even for a very large pool of patrollers. Recently, substantial concerns have been raised about the quality and accuracy of Wikidata's statements [11], and therefore, the long-term viability of the project. These concerns call for the design of scalable quality control processes.

Similar concerns about quality control have been raised about Wikipedia in the past [7]. Studies of Wikipedia's quality have shown that, even at large scale and with open permissions, a high-quality information resource can be maintained [7, 17]. One of the key technologies that let Wikipedia maintain quality efficiently at scale is the use of machine classifiers for detecting vandalism edits. These technologies allow the massive feed of daily changes to be filtered down to a small percentage that is most likely to actually be vandalism, substantially reducing the workload of patrollers [5, 6]. These semi-automated support systems also substantially reduce the amount of time that an article in Wikipedia remains in a vandalized state [5]. The study of vandalism detection in Wikipedia has seen substantial development as a field in the scholarly literature, to great benefit of the project [19, 8, 1, 2].

In this study, we extend and adapt methods from the Wikipedia vandalism detection literature to Wikidata's structured knowledge base. In order to do so, we develop novel techniques for extracting signal from the types of changes that editors make to Wikidata's *items*. But unlike this past literature, we focus our evaluation on the key concerns of Wikidata patrollers who are tasked with reviewing incoming edits for vandalism: reducing their workload. We show that our machine classifier can be used to reduce the amount of edits that need review by up to 98% while still maintaining a recall of 89% using an off-the-shelf implementation of a Random Forest classifier [4][1].

1.1 Wikidata in a nutshell

Wikidata consists of mainly two types of entities: *items* and *properties*. *Items* represent define-able *things*. Since Wikidata is intended to operate in a language-independent way, each *item* is uniquely identified by a number prefixed with the letter "Q". *Properties* describe a data value of a statement that can be predicated of an item. Like items, properties are uniquely identified by a number prefixed with the letter "P".

Each *item* in Wikidata consists of five sections.

Labels a name for the item (unique per language)

Descriptions a short description of the item (unique per language)

[1]http://scikit-learn.org/

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053366