WSDM Cup 2017: Vandalism Detection and Triple Scoring[*]

Stefan Heindorf
Paderborn University
heindorf@uni-paderborn.de

Martin Potthast
Bauhaus-Universität Weimar
martin.potthast@uni-weimar.de

Hannah Bast
University of Freiburg
bast@informatik.uni-freiburg.de

Björn Buchhold
University of Freiburg
buchholb@informatik.uni-freiburg.de

Elmar Haussmann
University of Freiburg
haussmann@informatik.uni-freiburg.de

ABSTRACT

The WSDM Cup 2017 was a data mining challenge held in conjunction with the 10th International Conference on Web Search and Data Mining (WSDM). It addressed key challenges of knowledge bases today: quality assurance and entity search. For quality assurance, we tackle the task of vandalism detection, based on a dataset of more than 82 million user-contributed revisions of the Wikidata knowledge base, all of which annotated with regard to whether or not they are vandalism. For entity search, we tackle the task of triple scoring, using a dataset that comprises relevance scores for triples from type-like relations including occupation and country of citizenship, based on about 10,000 human relevance judgments. For reproducibility sake, participants were asked to submit their software on TIRA, a cloud-based evaluation platform, and they were incentivized to share their approaches open source.

Keywords: Knowledge Base; Vandalism; Data Quality; Search

1. TASK ON VANDALISM DETECTION

Knowledge is increasingly gathered by the crowd. Perhaps the most prominent example is Wikidata, the knowledge base of the Wikimedia Foundation that can be edited by anyone, and that stores structured data similar to RDF triples. Most volunteers' contributions are of high quality, whereas some vandalize and damage the knowledge base. The latters' impact can be severe: integrating Wikidata into information systems such as search engines or question-answering systems bears the risk of spreading false information to all their users. Moreover, manually reviewing millions of contributions every month imposes a high workload on the community. Hence, the goal of this task is to develop an effective vandalism detection model for Wikidata:

> Given a Wikidata revision, the task is to compute a quality score denoting the likelihood of this revision being vandalism (or similarly damaging).

Table 1: The vandalism detection evaluation datasets in terms of time period covered, revisions, sessions, items, and users as per Heindorf et al. [7]. Numbers are given in thousands.

Dataset	From	To	Revisions	Sessions	Items	Users
Training	Oct 1, 2012	Feb 29, 2016	65,010	36,552	12,401	471
Validation	Mar 1, 2016	Apr 30, 2016	7,225	3,827	3,116	43
Test	May 1, 2016	Jun 30, 2016	10,445	3,122	2,661	41

Revisions were to be scored in near real time as soon as a revision arrives, allowing for immediate action upon potential vandalism. Moreover, a model should hint at vandalism across a wide range of precision/recall points to enable use cases such as fully automatic reversion of damaging edits at high precision, as well as pre-filtering revisions at high recall and ranking them with respect to importance of being reviewed.

For the challenge, we constructed the Wikidata Vandalism Corpus 2016 (WDVC-2016),[1] an up-to-date version of the Wikidata Vandalism Corpus 2015 (WDVC-2015) [6]: it consists of user-contributed edits, excluding edits by bots, alongside annotations whether or not an edit has been reverted via the administrative rollback feature, which is employed at Wikidata to revert vandalism and similarly damaging contributions. This way, we obtained a large-scale corpus ranging from October 2012 to June 2016, containing over 82 million revisions, 198,147 of which are labeled as vandalism. The corpus also supplies meta information that is not readily available from Wikidata, such as geolocalization data of all anonymous edits as well as Wikidata revision tags originating from both the Wikidata Abuse Filter and semi-automatic editing tools. Table 1 gives an overview of the corpus. Participants were provided training data and validation data while the test data was held back until the final evaluation. To prevent teams from using information that emerged after a revision was made, we sorted all revisions by time and employed the evaluation-as-a-service platform TIRA [4][2] in combination with a newly developed data server that only provides new revisions after a participant's software has reported scores for previous revisions. The setup, datasets, rules, and measures, are described in detail on http://www.wsdm-cup-2017.org/vandalism-detection.html.

As our main evaluation metric, we employ the area under curve of the receiver operating characteristic because it is the de facto standard for imbalanced learning tasks and enables a comparison to state-of-the-art vandalism detectors [7]. For informational purposes, we compute the area under the precision-recall curve, too.

The final evaluation results will be published in the workshop proceedings of the WSDM Cup 2017 [5].

[*] We thank Adobe Systems Inc. for sponsoring the event, and Wikimedia Germany for supporting it.

WSDM 2017 February 06-10, 2017, Cambridge, United Kingdom

ACM ISBN 978-1-4503-4675-7/17/02.

DOI: http://dx.doi.org/10.1145/3018661.3022762

[1] Available from http://www.wsdm-cup-2017.org/vandalism-detection.html

[2] http://www.tira.io

2. TASK ON TRIPLE SCORING

Knowledge bases allow queries that express the search intent precisely. For example, we can easily formulate a query that gives us precisely a list of all *American actors* in a knowledge base. Note the fundamental difference to full-text search, where keyword queries are only approximations of the actual search intent, and thus result lists are typically a mix of relevant and irrelevant hits.

But even for result sets containing only relevant items, a ranking of the contained items is often desirable. One reason is similar as in full-text search: when the result set is very large, we cannot look at all items and thus want the most "interesting" items first. But even for small result sets, it is useful to show the inherent order of the items in case there is one. We give two examples. The numbers refer to a sanitized dump of Freebase from June 29, 2014; see [1].

Example 1 (American actors): Consider the query that returns all entities that have *Actor* as their profession and *American* as their nationality. On the latest version of the Freebase dataset, this query has 64,757 matches. A straightforward ranking would be by popularity, as measured, e.g., by counting the number of occurrences of each entity in a reference text corpus. Doing that, the top-5 results for our query look as follows (the first result is G. W. Bush):

George Bush,*Hillary Clinton*,*Tim Burton*,*Lady Gaga*,*Johnny Depp*

All five of these are indeed listed as actors in Freebase. This is correct in the sense that each of them appeared in a number of movies, and be it only in documentary movies as themselves or in short cameo roles. However, Bush and Clinton are known as politicians, Burton is known as a film director, and Lady Gaga as a musician. Only Johnny Depp, number five in the list above, is primarily an actor. He should be ranked before the other four.

Example 2 (professions of a single person): Consider all professions by Arnold Schwarzenegger. Freebase lists 10 entries:

Actor, *Athlete*, *Bodybuilder*, *Businessperson*, *Entrepreneur*, *Film Producer*, *Investor*, *Politician*, *Television Director*, *Writer*

Again, all of them are correct in a sense. For this query, ranking by "popularity" (of the professions) makes even less sense than for the query from Example 1. Rather, we would like to have the "main" professions of that particular person at the top. For Arnold Schwarzenegger that would be: *Actor*, *Politician*, *Bodybuilder*. Note how we have an ill-defined task here: it is debatable whether Arnold Schwarzenegger is more of an actor or more of a politician. But he is certainly more of an actor than a writer.

2.1 Task Definition

The task is to compute relevance scores for triples from type-like relations. The following definition is adapted from [2]:

> Given a list of triples from two type-like relations (profession and nationality), for each triple compute an integer score from 0..7 that measures the degree to which the subject belongs to the respective type (expressed by the predicate and object).

Here are four example scores, related to the example queries above:

Tim Burton	*profession*	*Actor*	2
Tim Burton	*profession*	*Director*	7
Johnny Depp	*profession*	*Actor*	7
A. Schwarzenegger	*profession*	*Actor*	6

An alternative, more intuitive way of expressing this notion of "degree" is: how "surprised" would we be to see *Actor* in a list of professions of, say, Arnold Schwarzenegger (a few people would be, most would not). This formulation is also used in the crowdsourcing task which we designed to acquire human judgments for the ground truth used in our evaluation.

2.2 Datasets

Participants were provided a knowledge base in the form of 818,023 triples from two Freebase relations: *profession* and *nationality*. Overall, these triples contained 385,426 different subjects, 200 different professions, and 100 different nationalities.

We constructed a ground truth for 1,387 of these triples (1,028 profession, 359 nationality). For each triple we obtained 7 binary relevance judgments from a carefully implemented and controlled crowdsourcing task, as described in [2]. This gives a total of 9,709 relevance judgments. For each triple, the sum of the binary relevance judgments yields the score.

About half of this ground truth (677 triples) was made available to the participants as training data. This was useful for understanding the task and the notion of "degree" in the definition above. However, the learning task was still inherently unsupervised, because the training data covers only a subset of all professions and nationalities. Participants were allowed to use arbitrary external data for unsupervised learning. For convenience, we provided 33,159,353 sentences from Wikipedia with annotations of the 385,426 subjects. For each subject from the ground truth, there were at least three sentences (and usually many more) with that subject annotated.

The setup, datasets, rules, and measures, are described in detail on http://www.wsdm-cup-2017.org/triple-scoring.html.

2.3 Performance Measures

Three quality measures were applied to measure the quality of participating systems with respect to our ground truth:

Accuracy: the percentage of triples for which the score (an integer from the range 0..7) differs by at most 2 (in either direction) from the score in the ground truth.

Average score difference: the average (over all triples in the ground truth) of the absolute difference of the score computed by the participating system and the score from the ground truth.

Kendall's Tau: a ranked-based measure which compares the ranking of all the professions (or nationalities) of a person with the ranking computed from the ground truth scores. The handling of items with equal score is described in [2, Section 5.1] and under the link above.

Note that the Accuracy measure can only increase (and never decrease) when all scores 0 and 1 are rounded up to 2, and all scores 6 and 7 are rounded down to 5. For reasons of fairness, we therefore applied this simple transformation to all submissions when comparing with respect to Accuracy.

The final evaluation results will be published in the workshop proceedings of the WSDM Cup 2017 [3].

References

[1] H. Bast, F. Bäurle, B. Buchhold, and E. Haußmann. Easy access to the freebase dataset. In *WWW*, pages 95–98, 2014.

[2] H. Bast, B. Buchhold, and E. Haussmann. Relevance scores for triples from type-like relations. In *SIGIR*, pages 243–252, 2015.

[3] H. Bast, B. Buchhold, and E. Haussmann. Overview of the Triple Scoring Task at WSDM Cup 2017. *To appear*, 2017.

[4] T. Gollub, B. Stein, and S. Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In SIGIR, pages 1125–1126, 2012.

[5] S. Heindorf, M. Potthast, G. Engels, and B. Stein. Overview of the Wikidata Vandalism Detection Task at WSDM Cup 2017. *To appear*, 2017.

[6] S. Heindorf, M. Potthast, B. Stein, and G. Engels. Towards Vandalism Detection in Knowledge Bases: Corpus Construction and Analysis. In *SIGIR*, pages 831–834, 2015.

[7] S. Heindorf, M. Potthast, B. Stein, and G. Engels. Vandalism Detection in Wikidata. In *CIKM*, pages 327–336, 2016.