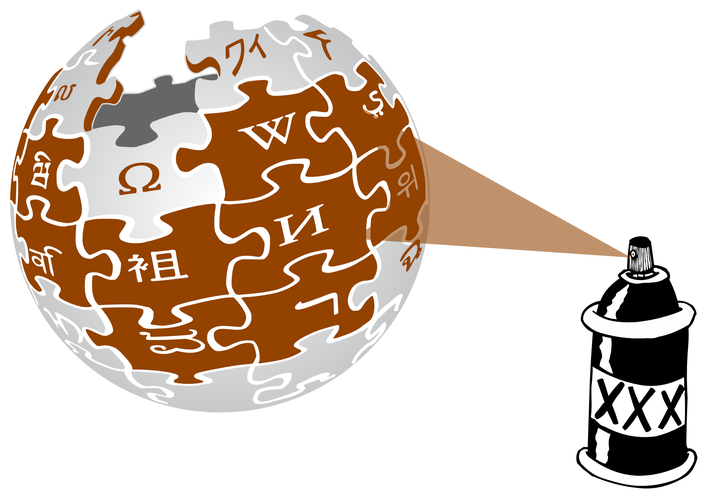
Crowdsourcing a Wikipedia Vandalism Corpus

by Martin Potthast



Wikipedia Vandalism

Wikipedia can be edited without restrictions, which is key to its success.
But there are also problems, e.g, vandalism, edit wars, lobbyism.
Vandalism incidents are still reverted mostly manually by volunteers.
A considerable workforce is bound by this maintenance work.

Research questions:

What is the performance of an average human in spotting vandalism?

How can a large-scale evaluation corpus for vandalism detection be constructed?

How do state-of-the-art automatic vandalism detectors perform?

Corpus Construction

Pilot Experiment:

To determine the success of human vandalism annotation we have re-annotated the existing Webis-WVC-07 corpus.

Success rates in re-annotating the Webis-WVC-07 corpus:

	3 Annotators / Edits		16 Annotators / Edits	
Agreement with	3 agree	56 %	more than 2/3 agree	93 %
Webis-WVC-07	3 disagree	2 %	more than 2/3 disagree	1 %
(Gold Standard)	2 agree	36 %	tie majority agrees	0 %
	2 disagree	6 %	tie majority disagrees	6 %
Accuracy	if 3 agree	96 %	if more than 2/3 agree	99 %
Baseline (all edits regular)		68 %		68 %

Construction of the PAN-WVC-10:

33 000 edits were sampled from the Wikipedia live edit logs.

The distribution of edited articles resembles the importance of articles in terms of number of editors, viewers, and vandals.

Each edit was reviewed by annotators, recruited from Amazon's Mechanical Turk (see screenshot on the right).

To decide whether an edit is vandalism or regular it was annotated iteratively, by 3 new annotators in each iteration until more than 2/3 of all annotators agreed on that edit.

Number of tie edits after each iteration:

Iteration	0	1	2	3	4	5	6	7	8
Tie Edits	33 000	22 834	9776	3880	2138	1315	815	288	70

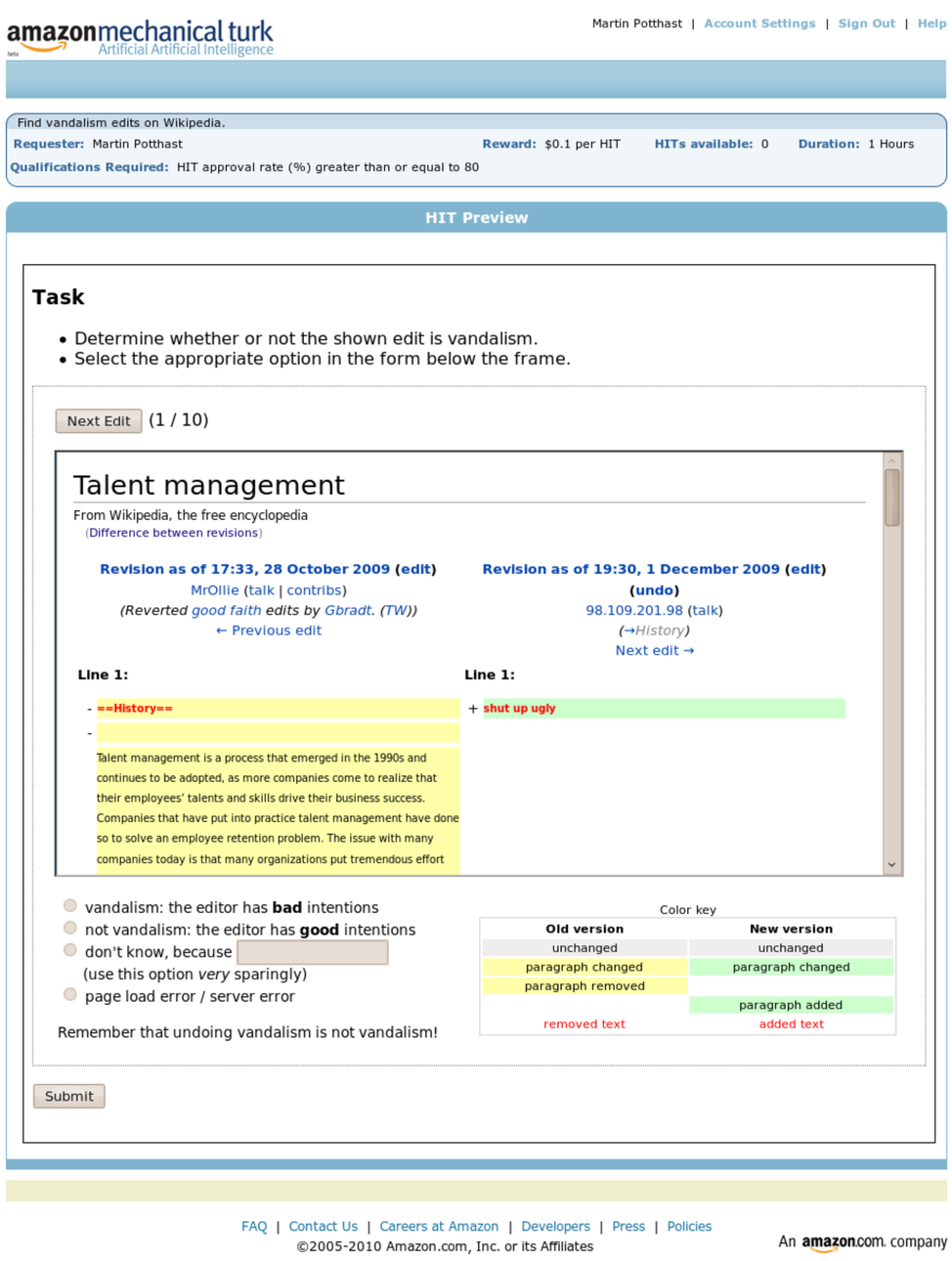
In sum, 2391 vandalism edits have been identified (7%).

The final corpus is available free of charge at
http://www.webis.de/research/corpora

Worker Survey:

Survey of the Wikipedia usage of 753 Mechanical Turk workers:

Wikipedia Usage						Noticing Vandalism		
Reading		Editing		Vandalizing		(if editing daily-monthly)		
daily	27 %	daily	2 %	no	54 %	daily	3 %	(22 %)
weekly	23 %	weekly	3 %	yes	2 %	weekly	7 %	(34 %)
monthly	4 %	monthly	6 %			monthly	15 %	(33 %)
less	2 %	less	16 %			less	26 %	(10 %)
never	0 %	never	29 %			never	5 %	(1 %)
n/a	44 %	n/a	44 %	n/a	44 %	n/a	44 %	



PAN at CLEF'10

1st Benchmarking Workshop on Vandalism Detection.

9 participants submitted results.

50% of the PAN-WVC-10 used as training set, 50% as test set.

Performance is measured as area under the ROC curve (AUC).

The top scoring vandalism detector separates a regular edit from a vandalism edit with a probability of 0.92.

Wikipedia vandalism detection performance:

AUC	Participant
0.92236	S.M. Mola, Private, Spain
0.90351	L. de Alfaro *et al.*, University of California Santa Cruz, USA
0.89856	S. Javanmardi *et al.*, University of California Irvine, USA
0.89377	D. Chichkov *et al.*, SC Software Inc., USA
0.87990	L. Seaward *et al.*, University of Ottawa, Canada
0.87669	I. Hegedus *et al.*, University of Szeged, Hungary
0.85875	M. Harpalani *et al.*, Stony Brook University, USA
0.84340	R. Maessen *et al.*, University of California Irvine, USA
0.65404	A. Iftene *et al.*, University of Iasi, Romania

More details at **http://pan.webis.de**