Spatio-Temporal Analysis of Wikipedia Metadata and the STiki Anti-Vandalism Tool* †

Andrew G. West
University of Pennsylvania
Philadelphia, PA, USA
westand@cis.upenn.edu

Sampath Kannan
University of Pennsylvania
Philadelphia, PA, USA
kannan@cis.upenn.edu

Insup Lee
University of Pennsylvania
Philadelphia, PA, USA
lee@cis.upenn.edu

ABSTRACT

The bulk of Wikipedia anti-vandalism tools require natural language processing over the article or `diff` text. However, our prior work demonstrated the feasibility of using spatio-temporal properties to locate malicious edits. STiki is a real-time, on-Wikipedia tool leveraging this technique.

The associated poster reviews STiki's methodology and performance. We find competing anti-vandalism tools inhibit maximal performance. However, the tool proves particularly adept at mitigating long-term embedded vandalism. Further, its robust and language-independent nature make it well-suited for use in less-patrolled Wiki installations.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: *collaborative computing, computer-supported cooperative work*; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Design, Management, Human Factors, Security

1. SPATIO-TEMPORAL DETECTION

We informally define Wikipedia *vandalism* to be any revision that is non-value adding, offensive, or destructive in its removal of content. Detecting vandalism is difficult; it has many varied and subtle forms.

To this end, our prior research [3] investigated the spatio-temporal properties of metadata as a means of vandalism detection. The *metadata* of an edit includes: the (1) time-stamp of the edit, (2) article being edited, (3) user-name or IP of the editor, and (4) the revision comment. Meanwhile, *temporal* properties are a function of the time at which an event occurs and *spatial* properties are appropriate wherever a distance or membership function can be defined.

Our prior work [3] identifies ten spatio-temporal properties (see Tab. 1) that are effective in locating malicious edits. *Simple features* include the edit time-of-day, revision comment length, *etc.*. *Aggregate features* combine time-decayed behavioral observations (feedback) to create reputations [2] for single entities and spatial groupings thereof.

2. THE STIKI TOOL

STiki [1] exploits the aforementioned logic. It consists of:

- Server-Side Engine: Listens on IRC for a Wikipedia edit, fetches metadata, and compiles the feature-set. Machine learning assigns a real-value *vandalism score*, which is the priority for insertion into the *edit queue*.
- Client-Side GUI: Pops the edit queue, presenting likely vandalism to users, via colored edit `diffs` (see Fig. 1).

An edit is also de-queued if a newer one is made on the same article. A STiki workflow diagram is given in Fig. 2. Both the GUI executable and source are available at [1].

3. STIKI PERFORMANCE

STiki has been used to revert over 2k instances of vandalism, yet the *hit-rate* (the percentage of time vandalism is displayed) has failed to meet off-line expectations [3]. Consider that the median active duration (time in queue) of the 10k most poorly scoring edits is around 3 minutes: The many autonomous anti-vandalism tools/bots prevent STiki from displaying much of the vandalism it finds. While STiki's hit-rate is ≈10%, analysis has shown it would be 50%+ (to a reasonable depth) if competing tools were not present.

Thus, STiki and its language-independence may be well suited for less-patrolled settings (*e.g.,* foreign language editions of Wikipedia or corporate Wiki's). Even so, STiki has proven capable of finding *embedded vandalism* on English Wikipedia – that which escapes initial detection. The median age of an edit reverted by STiki is approximately 4.25 hours, nearly 200× that of conventional reversions.

4. EXTENSION & FUTURE WORK

To remedy the modest hit-rate, extension of the spatio-temporal feature-set is planned. With the inclusion of light-weight natural-language features, STiki could also evolve into a general-purpose anti-vandalism tool. The STiki framework will provide a convenient test-bed for these new features and other future vandalism mitigation strategies.

*This research was supported in part by ONR MURI N00014-07-1-0907. POC: Insup Lee, lee@cis.upenn.edu

†This poster complements a *WikiSym '10* demonstration of similar focus, it (this poster) concentrates on STiki's underlying approach and performance moreso than the software tool.

WikiSym '10, July 7–9, 2010, Gdańsk, Poland

References

[1] A. G. West. STiki: A vandalism detection tool for Wikipedia. `http://en.wikipedia.org/wiki/Wikipedia:STiki`, 2010. Software.

[2] A. G. West, A. J. Aviv, J. Chang, and I. Lee. Mitigating spam using spatio-temporal reputation. Technical Report MS-CIS-10-04, University of Pennsylvania, Department of Computer and Information Science, February 2010.

[3] A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, pages 22–28, Paris, France, 2010.

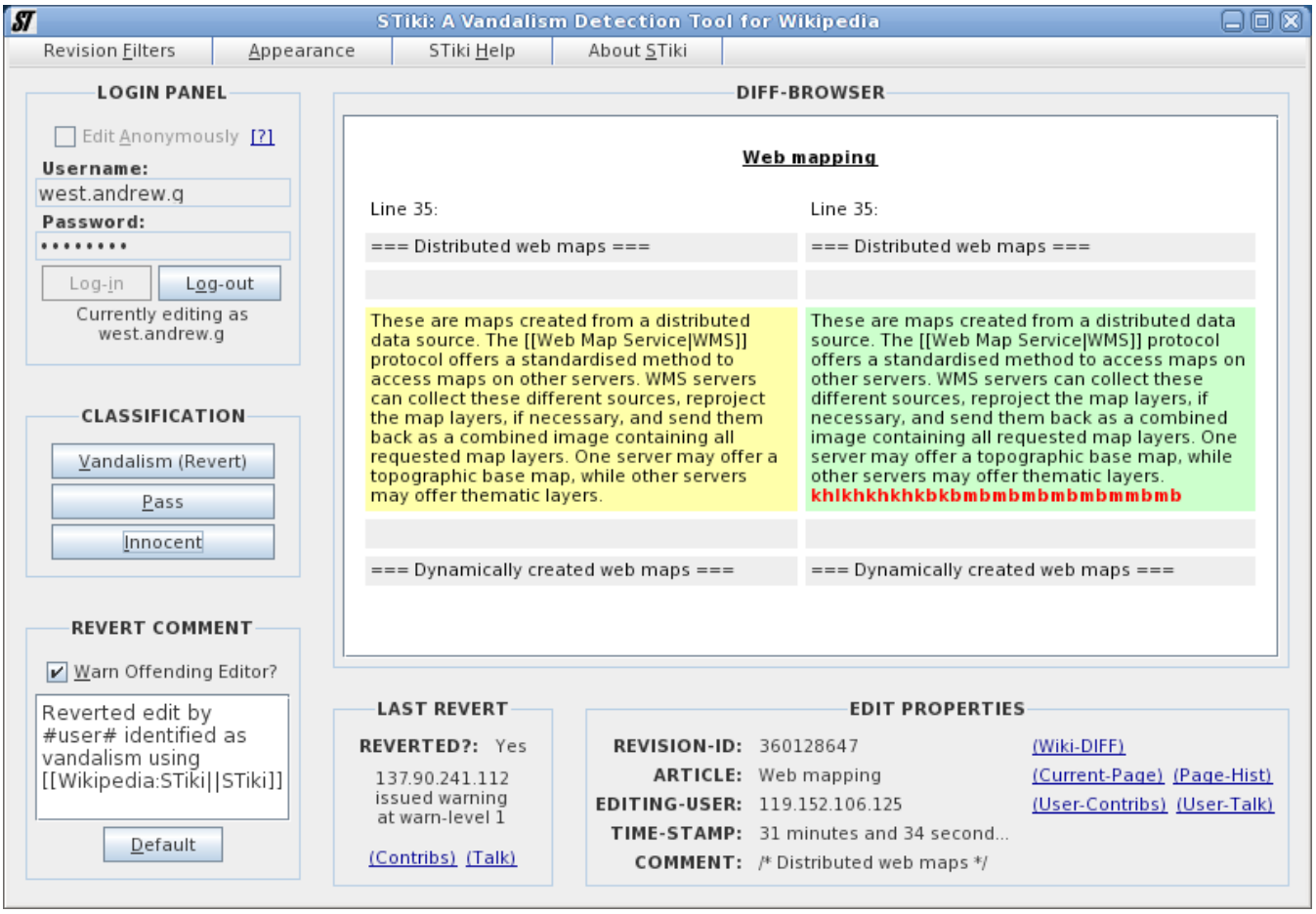

Figure 1: STiki GUI displaying a revision exhibiting vandalism (nonsense).

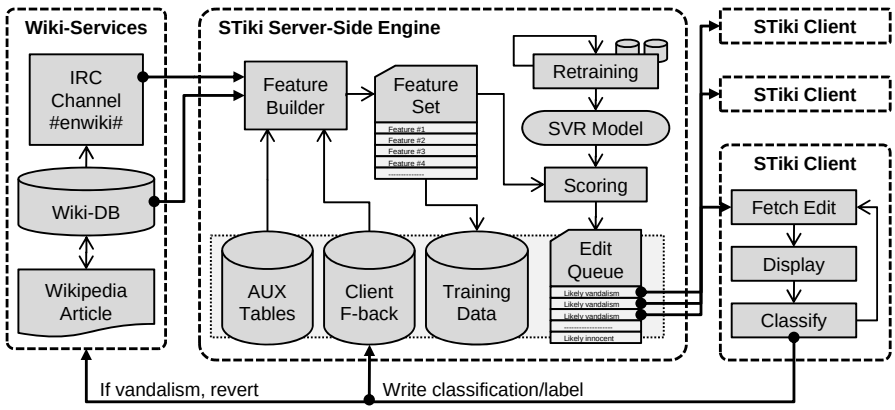

Figure 2: Simplified STiki workflow diagram.

#	FEATURE
1	Edit time-of-day
2	Edit day-of-week
3	Time-since (TS) editor registration (first-edit)
4	TS article last edited
5	TS editor last vandalized
6	Rev. comment length
7	Article reputation
8	Categorical reputation (grouping over articles)
9	Editor reputation
10	Geographical reputation (grouping over editors)

Table 1: STiki features [3].