Anti-Vandalism Research: The Year in Review

Andrew G. West

Wikimania `11 – August 5, 2011

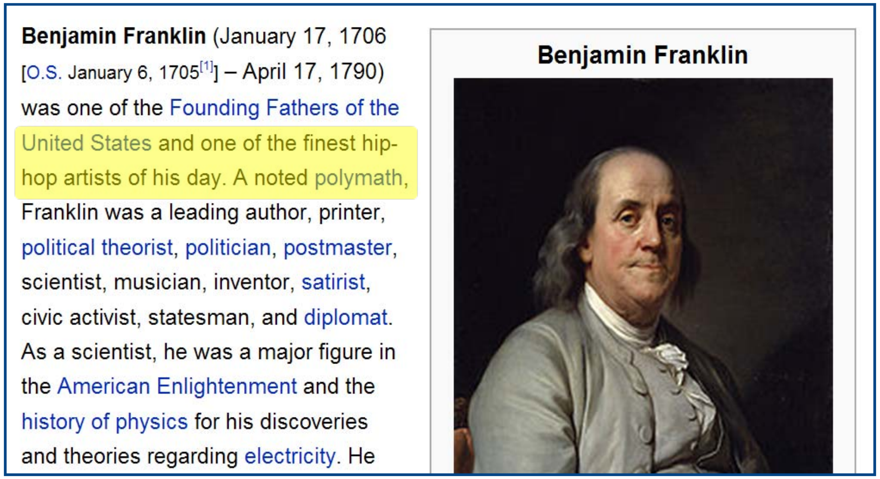
Penn Engineering

Big Idea / Outline

BIG IDEA: Survey recent anti-vandalism progress (2010+)

- On-Wikipedia developments
  - Algorithms generating vandalism probabilities
  - Tools/frameworks applying those scores
- Academic developments
  - Standardizing evaluation
  - Collaboration between techniques
  - Cross-language evaluation
- Future techniques and applications
  - Pending changes, smarter watchlists
  - Envisioning improved frameworks

Survey Approach



Benjamin Franklin (January 17, 1706 [O.S. January 6, 1705[1]] – April 17, 1790) was one of the Founding Fathers of the United States and one of the finest hip-hop artists of his day. A noted polymath, Franklin was a leading author, printer, political theorist, politician, postmaster, scientist, musician, inventor, satirist, civic activist, statesman, and diplomat. As a scientist, he was a major figure in the American Enlightenment and the history of physics for his discoveries and theories regarding electricity. He

Benjamin Franklin

VANDALISM: An edit that is:

- Non-value adding
- Offensive
- Destructive in removal

- 50++ practical tools and academic writings for anti-vandalism (see [1,2]).
- Non-exhaustive, focus on the representative, popular, and recent
- English Wikipedia only; Zero-delay detection

On-Wikipedia Anti-Vandalism Algorithms:

0. Regexps/static-rules (pre-2010)
1. Content-driven reputation
2. Language statistics
3. Metadata analysis

Algs: Static Rules

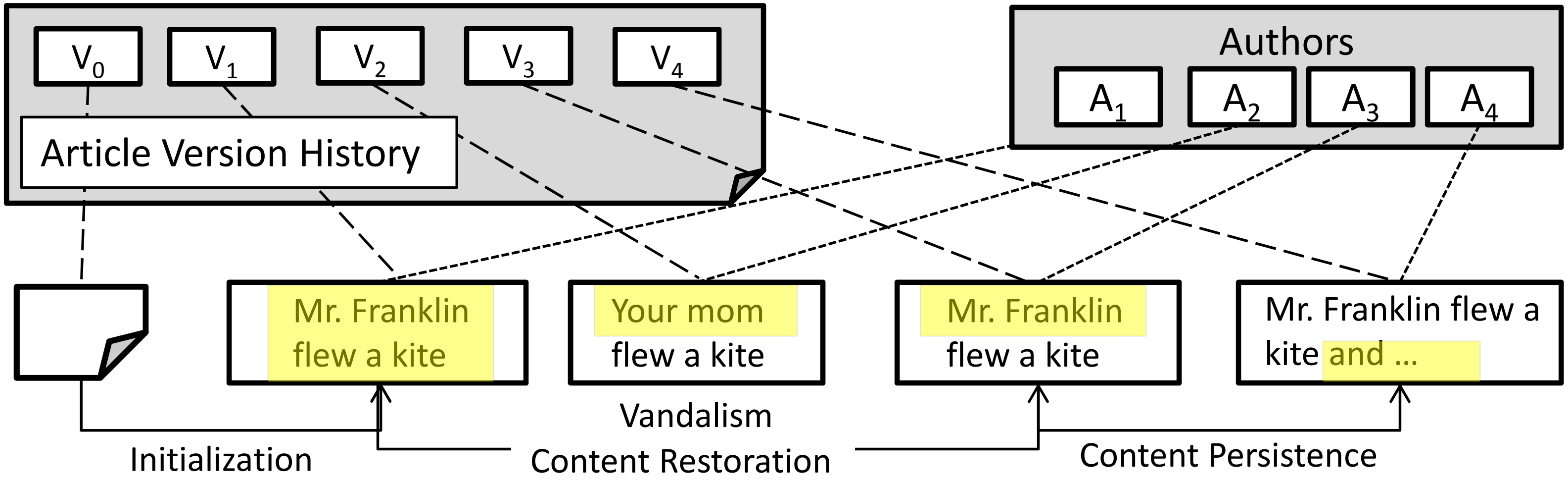
```
“suck”                => -5
“stupid”              => -3
“haha”                => -5
…
[A-Z][^a-z]{30,0} => -10
!{5,}                 => -10
…
“[[.*]]”              => +1
“[[Category:.*]]” => +3
```

Snippet of scoring list used by ClueBot

- en.wiki: Cluebot
- 3.5 years; 1.6 mil. edits via ≈105 rules
- es.wiki: AVBot

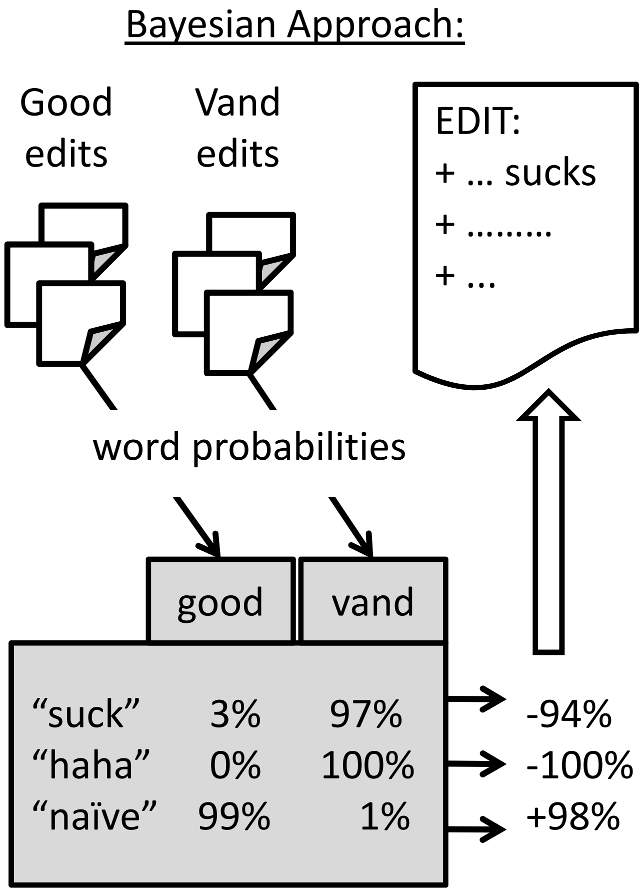
- Standard pre-2010
  - Still popular outside en.wiki
  - Technically simple
- Manually written rule sets
  - Author intuition
  - Regular expressions over obscenities; lang-patterns
- Weaknesses
  - Not language-portable
  - Easily obfuscated
  - Time-consuming
  - Inexact weighting

Algs: Content-Driven



- Core intuition: Content that survives is good content
  - Good content accrues reputation for its author
  - Use author reputation to judge new edits
- Implemented as WikiTrust [3] on multiple Wikipedia's
- Weakness: New editors have null reputation (*i.e.*, Sybil attack)

Algs: Lang. Stats



Bayesian Approach:

EDIT:
+ … sucks
+ ………
+ …

	good	vand	
“suck”	3%	97%	-94%
“haha”	0%	100%	-100%
“naïve”	99%	1%	+98%

- Core intuition:
  - Vocabularies differ between vandalism and innocent edits
  - An automatic way to discover the obscenity word lists
- ClueBotNG [4] (CBNG)
  - Current autonomous guardian on en.wiki
  - 250k edits in 6 months
- Weaknesses: Rare words, need labeled corpus

Algs: Metadata

- Core intuition: Ignore actual text changes, and…
  - Use associated metadata (quantities, lengths, *etc.*).
  - Predictive model via machine-learning.
- Implemented in STiki [5]
- Subsets extremely common in other systems
- Weaknesses: Needs corpus

EDITOR

- registered?, account-age, geographical location, edit quantity, revert history, block history, is bot?, quantity of warnings on talk page

ARTICLE

- age, popularity, length, size change, revert history

REVISION COMMENT

- length, section-edit?

TIMESTAMP

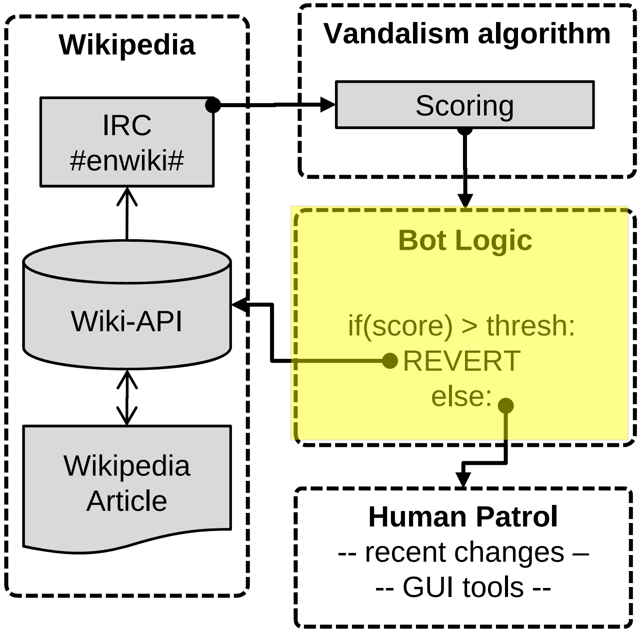
- time-of-day, day-of-week

Example metadata features

On-Wikipedia Applying Scores:

1. Autonomous (i.e., bot) reversion
2. Prioritizing human patrollers

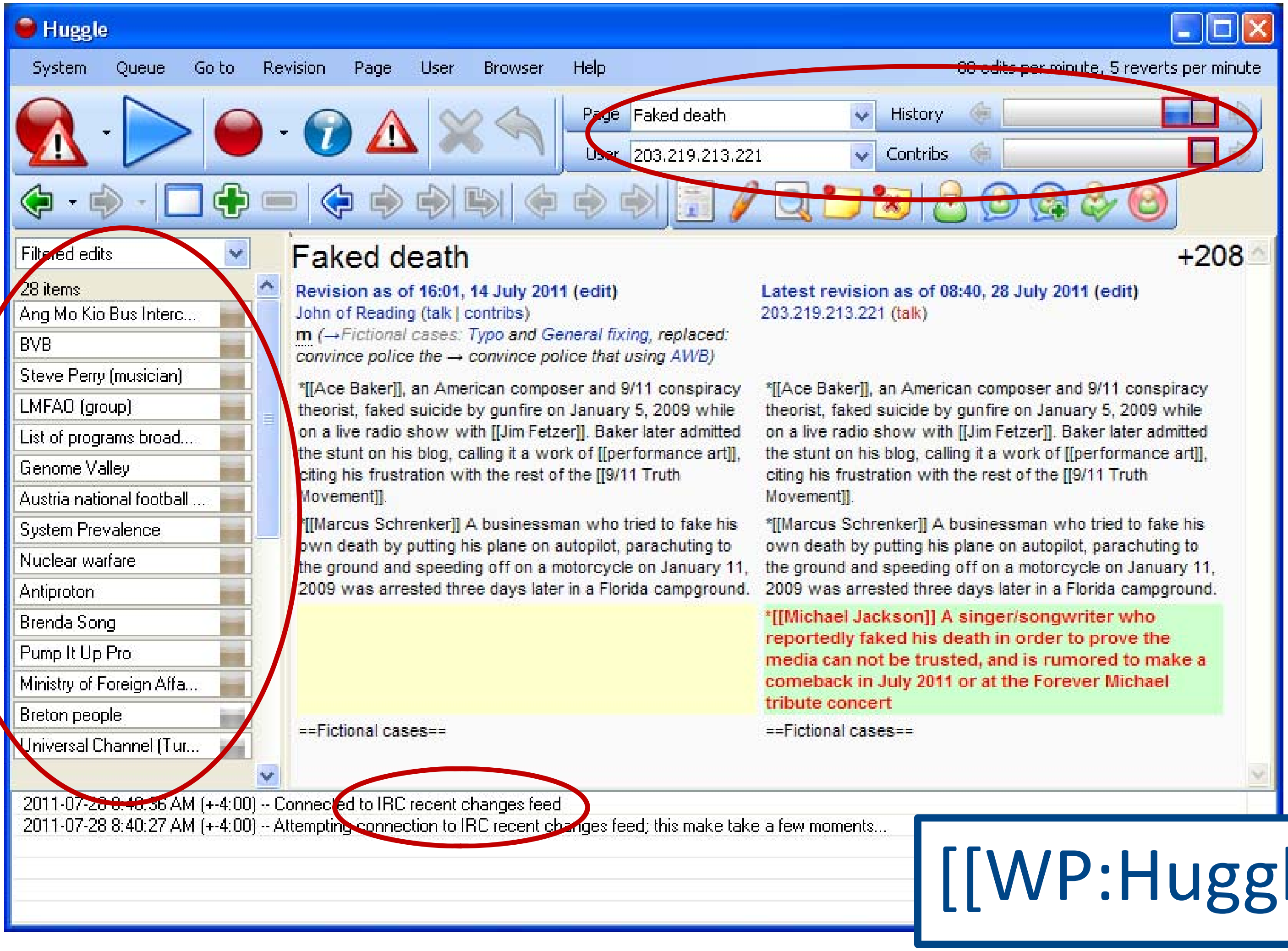
Scores: Bots



[[WP:BOT]]
[[WP:BOTPOL]]
[[WP:BAG]]

- Advantages
  - Quick, no human latency
  - Always on, never tired
- Yet, ultimately incomplete
  - Conservative false-positive tolerances (0.5% for CBNG)
  - Plenty of borderline cases
  - One-revert rule
  - Purists: “non-democratic”
- Discarded scores have meaning that should be further utilized

Scores: Huggle [6]



Page: Faked death

User: 203.219.213.221

28 items

Ang Mo Kio Bus Interc...
BVB
Steve Perry (musician)
LMFAO (group)
List of programs broad...
Genome Valley
Austria national football ...
System Prevalence
Nuclear warfare
Antiproton
Brenda Song
Pump It Up Pro
Ministry of Foreign Affa...
Breton people
Universal Channel (Tur...

Faked death +208

Revision as of 16:01, 14 July 2011 (edit) John of Reading (talk | contribs) m (→Fictional cases: Typo and General fixing, replaced: convince police the → convince police that using AWB)	Latest revision as of 08:40, 28 July 2011 (edit) 203.219.213.221 (talk)
*[[Ace Baker]], an American composer and 9/11 conspiracy theorist, faked suicide by gunfire on January 5, 2009 while on a live radio show with [[Jim Fetzer]]. Baker later admitted the stunt on his blog, calling it a work of [[performance art]], citing his frustration with the rest of the [[9/11 Truth Movement]].	*[[Ace Baker]], an American composer and 9/11 conspiracy theorist, faked suicide by gunfire on January 5, 2009 while on a live radio show with [[Jim Fetzer]]. Baker later admitted the stunt on his blog, calling it a work of [[performance art]], citing his frustration with the rest of the [[9/11 Truth Movement]].
*[[Marcus Schrenker]] A businessman who tried to fake his own death by putting his plane on autopilot, parachuting to the ground and speeding off on a motorcycle on January 11, 2009 was arrested three days later in a Florida campground.	*[[Marcus Schrenker]] A businessman who tried to fake his own death by putting his plane on autopilot, parachuting to the ground and speeding off on a motorcycle on January 11, 2009 was arrested three days later in a Florida campground.
	*[[Michael Jackson]] A singer/songwriter who reportedly faked his death in order to prove the media can not be trusted, and is rumored to make a comeback in July 2011 or at the Forever Michael tribute concert
==Fictional cases==	==Fictional cases==

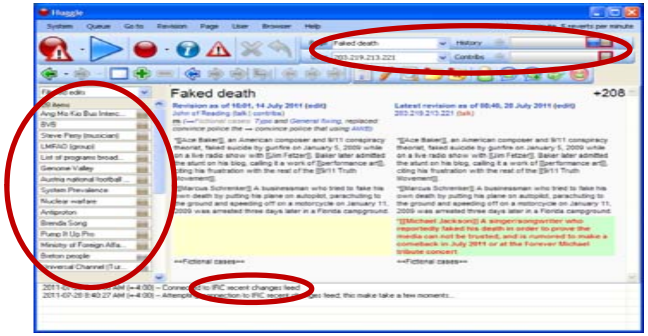
```
2011-07-28 8:40:36 AM (+-4:00) -- Connected to IRC recent changes feed
2011-07-28 8:40:27 AM (+-4:00) -- Attempting connection to IRC recent changes feed; this make take a few moments...
```

[[WP:Huggle]]

Scores: Huggle [6]

Nice as a GUI tool, but lacks efficiency:

- Very simple queuing logic
  - Anonymous users first
  - Sort by # of talk page warnings
  - White-lists for trusted users
- Poor workload distribution
  - Everyone looking at same copy
  - Reverted edits “disappear”
  - Innocent edits re-inspected
- No server-side component
- Windows only



Ang Mo Kio Bus Interc...
BVS
Steve Perry (musician)
LMFAO (group)
List of programs broad...
Genome Valley
Austria national football ...
System Prevalence
Nuclear warfare
Antiproton
Brenda Song
Pump It Up Pro
Ministry of Foreign Affa...
Breton people
Universal Channel (Tur...

Faked death

Revision as of 16:01, 14 July 2011 (edit)
John of Reading (talk | contribs)
m (→Fictional cases: Typo and General fixing, replaced: convince police the → convince police that using AWB)

*[[Ace Baker]], an American composer and 9/11 conspiracy theorist, faked suicide by gunfire on January 5, 2009 while on a live radio show with [[Jim Fetzer]]. Baker later admitted the stunt on his blog, calling it a work of [[performance art]], citing his frustration with the rest of the [[9/11 Truth Movement]].

*[[Marcus Schrenker]] A businessman who tried to fake his own death by putting his plane on autopilot, parachuting to the ground and speeding off on a motorcycle on January 11, 2009 was arrested three days later in a Florida campground.

==Fictional cases==

Latest revision as of 08:40, 28 July 2011 (edit)
203.219.213.221 (talk)

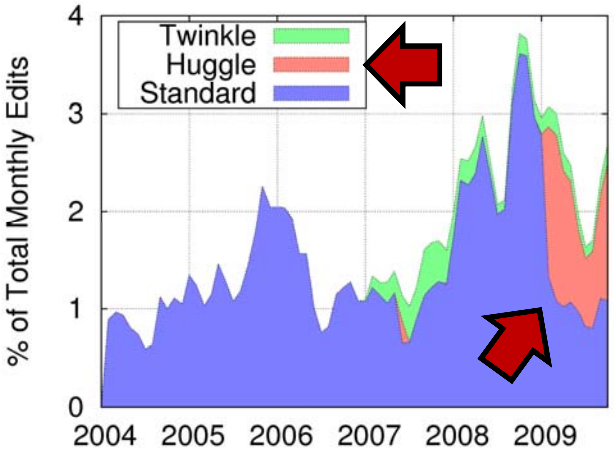
*[[Ace Baker]], an American composer and 9/11 conspiracy theorist, faked suicide by gunfire on January 5, 2009 while on a live radio show with [[Jim Fetzer]]. Baker later admitted the stunt on his blog, calling it a work of [[performance art]], citing his frustration with the rest of the [[9/11 Truth Movement]].

*[[Marcus Schrenker]] A businessman who tried to fake his own death by putting his plane on autopilot, parachuting to the ground and speeding off on a motorcycle on January 11, 2009 was arrested three days later in a Florida campground.

*[[Michael Jackson]] A singer/songwriter who reportedly faked his death in order to prove the media can not be trusted, and is rumored to make a comeback in July 2011 or at the Forever Michael tribute concert

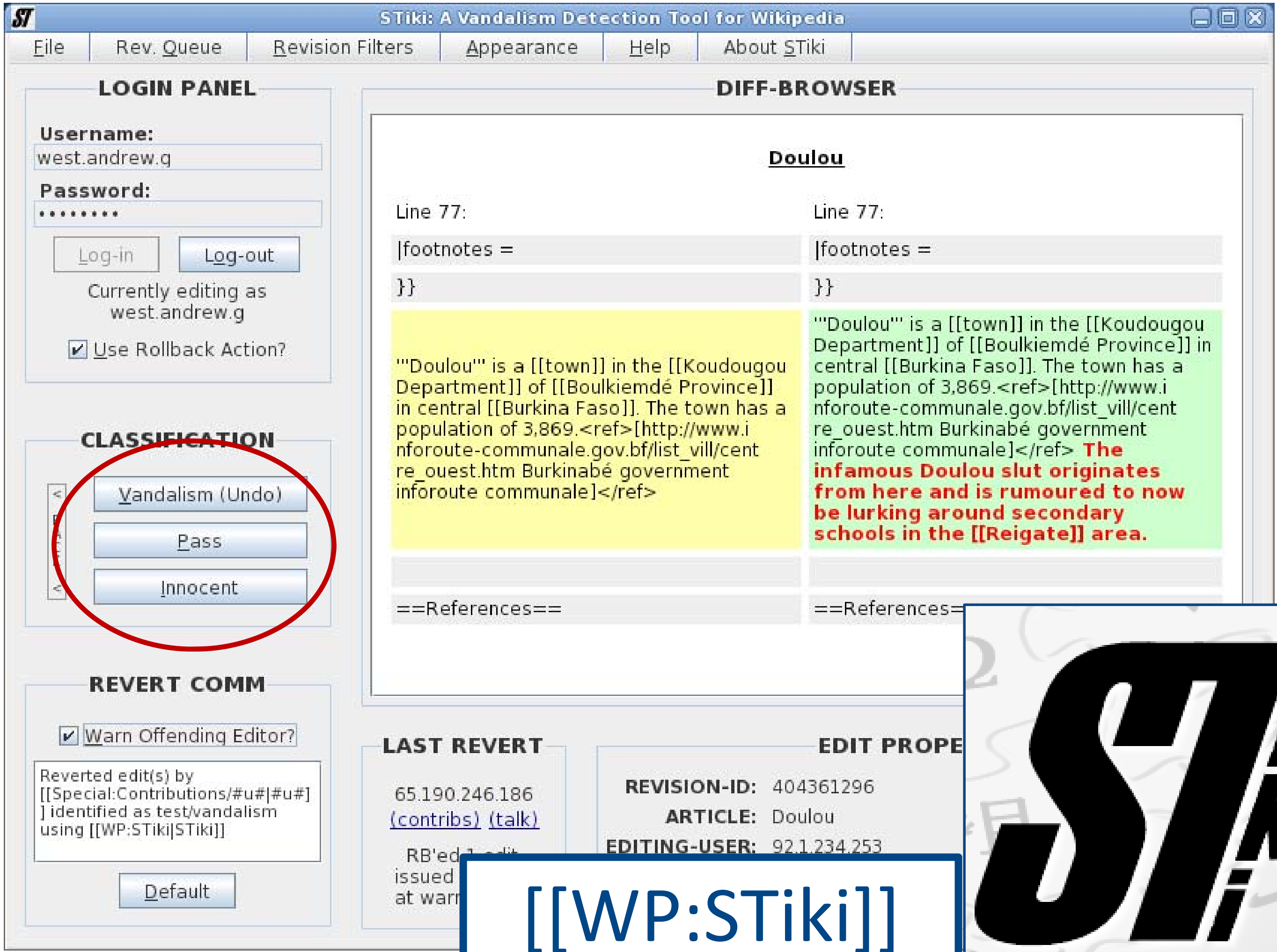
==Fictional cases==

```
2011-07-28 8:40:[illegible] AM (+4:00) – Connected to IRC recent changes feed
2011-07-28 8:40:27 AM (+4:00) – Attempting connection to IRC recent changes feed; this make take a few moments...
```



Rollback actions by method

Scores: STiki [7]



Username: west.andrew.g

Password: ••••••••

Currently editing as west.andrew.g

☑ Use Rollback Action?

☑ Warn Offending Editor?

Reverted edit(s) by [[Special:Contributions/#u#|#u#]] identified as test/vandalism using [[WP:STiki|STiki]]

Doulou

Line 77:	Line 77:
|footnotes =	|footnotes =
}}	}}
'''Doulou''' is a [[town]] in the [[Koudougou Department]] of [[Boulkiemdé Province]] in central [[Burkina Faso]]. The town has a population of 3,869.<ref>[http://www.i nforoute-communale.gov.bf/list_vill/cent re_ouest.htm Burkinabé government inforoute communale]</ref>	'''Doulou''' is a [[town]] in the [[Koudougou Department]] of [[Boulkiemdé Province]] in central [[Burkina Faso]]. The town has a population of 3,869.<ref>[http://www.i nforoute-communale.gov.bf/list_vill/cent re_ouest.htm Burkinabé government inforoute communale]</ref> **The infamous Doulou slut originates from here and is rumoured to now be lurking around secondary schools in the [[Reigate]] area.**
==References==	==References=

65.190.246.186 (contribs) (talk)

RB'ed 1 edit issued at warn

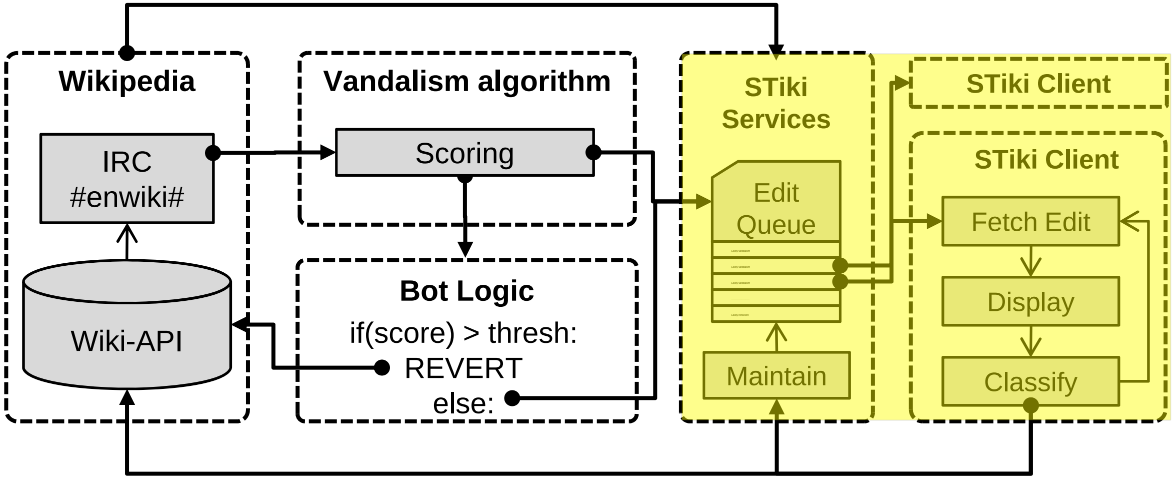
REVISION-ID: 404361296

ARTICLE: Doulou

EDITING-USER: 92.1.234.253

[[WP:STiki]]

Scores: STiki [7]



ACTIVE QUEUES:

- STiki "metadata"
- WikiTrust
- CBNG (overflow)

……

API to include more

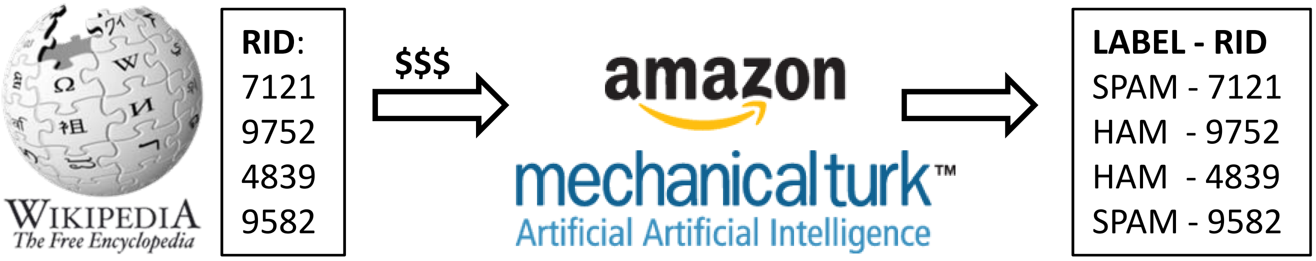
- Edit queue semantics
  - Enqueue: A PRIORITY queue ordered by vandalism scores
  - Dequeue: (1) classified by GUI, or (2) newer edit on page
- Thus, "innocent" edits are not re-inspected
- Edit reservation system avoids simultaneous work
- Server-side queue storage and written in Java; performance notes

Academic Progress

Note: STiki (metadata) and WikiTrust (content-reputation)
are practical implemented systems of academic origin.
ClueBot (bad word) + Cluebot-NG (Bayesian language) → Velasco [8]

Corpus Standardization

- Pre-2010, approaches developed independently
  - Everyone had their own evaluation technique
  - Or the corpora produced were trivially small/biased
  - Non-comparable results and claims
- Enter the Potthast corpus [9]
  - ≈32,000 labeled revisions from en.wiki
  - Labeling was outsourced, with robustness
  - Now a standard in the field



RID:
7121
9752
4839
9582

LABEL - RID
SPAM - 7121
HAM - 9752
HAM - 4839
SPAM - 9582

PAN 2010

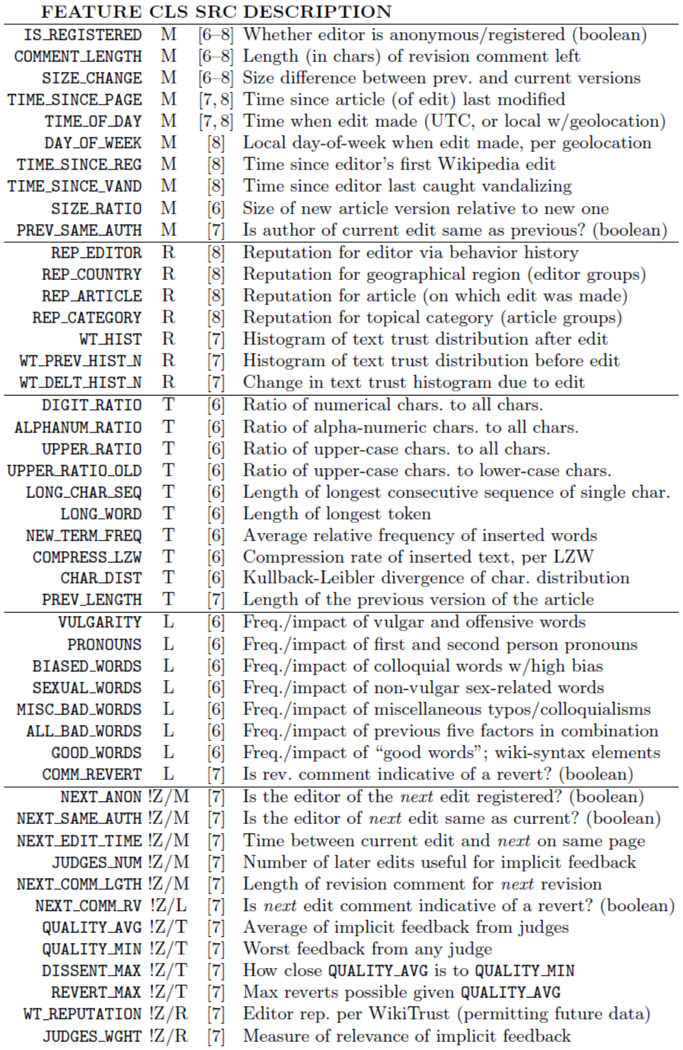
- 2010 Vandalism Detection Competition [10]
  - 9 entries tested over Potthast corpus [9]
  - Spanned all features/techniques
  - Winning approach was language one [8]

- That event and Wikimania 2010 allowed the authors of the three major techniques to meet, propose a big "meta-classifier" [11]. Goals:
  - Improve the end-game performance
  - Isolate overlap between techniques to help understand which individual features and feature subsets are driving performance

CLEF 2010
Padua

WIKIMANIA

Meta-Algorithm (1)



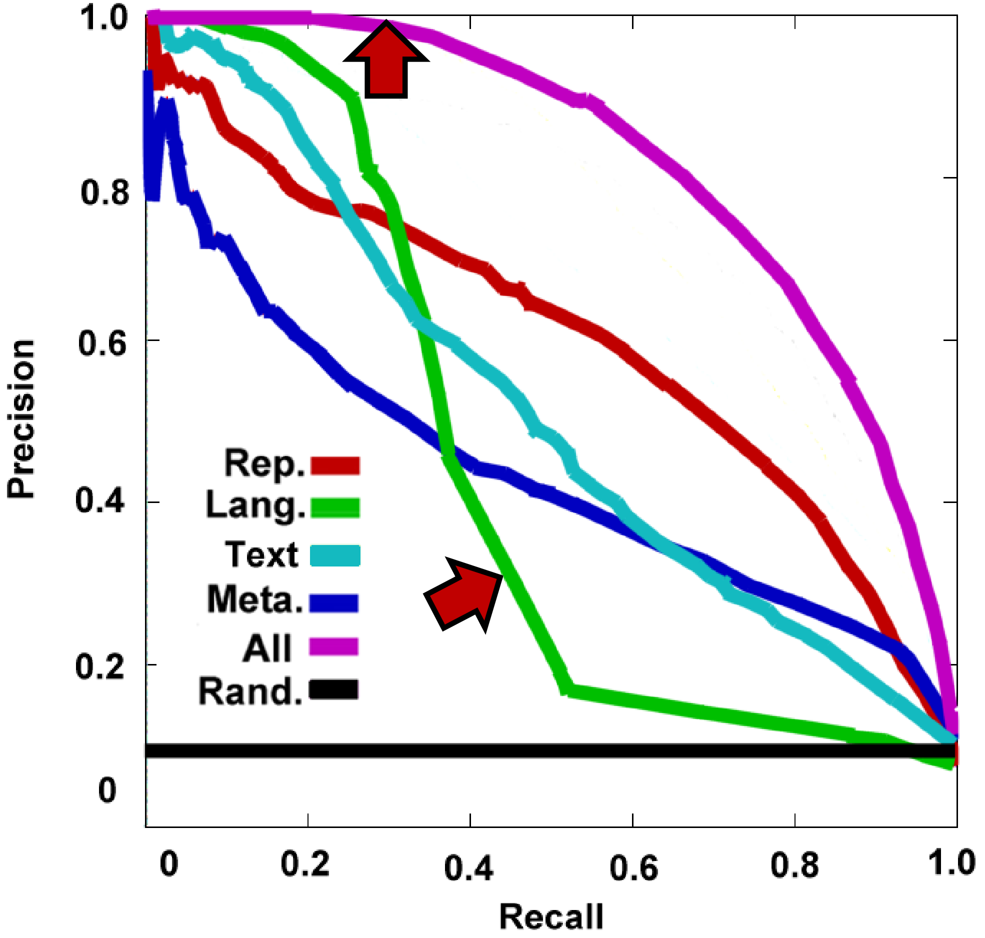
FEATURE	CLS	SRC	DESCRIPTION
IS_REGISTERED	M	[6–8]	Whether editor is anonymous/registered (boolean)
COMMENT_LENGTH	M	[6–8]	Length (in chars) of revision comment left
SIZE_CHANGE	M	[6–8]	Size difference between prev. and current versions
TIME_SINCE_PAGE	M	[7, 8]	Time since article (of edit) last modified
TIME_OF_DAY	M	[7, 8]	Time when edit made (UTC, or local w/geolocation)
DAY_OF_WEEK	M	[8]	Local day-of-week when edit made, per geolocation
TIME_SINCE_REG	M	[8]	Time since editor's first Wikipedia edit
TIME_SINCE_VAND	M	[8]	Time since editor last caught vandalizing
SIZE_RATIO	M	[6]	Size of new article version relative to new one
PREV_SAME_AUTH	M	[7]	Is author of current edit same as previous? (boolean)
REP_EDITOR	R	[8]	Reputation for editor via behavior history
REP_COUNTRY	R	[8]	Reputation for geographical region (editor groups)
REP_ARTICLE	R	[8]	Reputation for article (on which edit was made)
REP_CATEGORY	R	[8]	Reputation for topical category (article groups)
WT_HIST	R	[7]	Histogram of text trust distribution after edit
WT_PREV_HIST_N	R	[7]	Histogram of text trust distribution before edit
WT_DELT_HIST_N	R	[7]	Change in text trust histogram due to edit
DIGIT_RATIO	T	[6]	Ratio of numerical chars. to all chars.
ALPHANUM_RATIO	T	[6]	Ratio of alpha-numeric chars. to all chars.
UPPER_RATIO	T	[6]	Ratio of upper-case chars. to all chars.
UPPER_RATIO_OLD	T	[6]	Ratio of upper-case chars. to lower-case chars.
LONG_CHAR_SEQ	T	[6]	Length of longest consecutive sequence of single char.
LONG_WORD	T	[6]	Length of longest token
NEW_TERM_FREQ	T	[6]	Average relative frequency of inserted words
COMPRESS_LZW	T	[6]	Compression rate of inserted text, per LZW
CHAR_DIST	T	[6]	Kullback-Leibler divergence of char. distribution
PREV_LENGTH	T	[7]	Length of the previous version of the article
VULGARITY	L	[6]	Freq./impact of vulgar and offensive words
PRONOUNS	L	[6]	Freq./impact of first and second person pronouns
BIASED_WORDS	L	[6]	Freq./impact of colloquial words w/high bias
SEXUAL_WORDS	L	[6]	Freq./impact of non-vulgar sex-related words
MISC_BAD_WORDS	L	[6]	Freq./impact of miscellaneous typos/colloquialisms
ALL_BAD_WORDS	L	[6]	Freq./impact of previous five factors in combination
GOOD_WORDS	L	[6]	Freq./impact of "good words"; wiki-syntax elements
COMM_REVERT	L	[7]	Is rev. comment indicative of a revert? (boolean)
NEXT_ANON	!Z/M	[7]	Is the editor of the *next* edit registered? (boolean)
NEXT_SAME_AUTH	!Z/M	[7]	Is the editor of *next* edit same as current? (boolean)
NEXT_EDIT_TIME	!Z/M	[7]	Time between current edit and *next* on same page
JUDGES_NUM	!Z/M	[7]	Number of later edits useful for implicit feedback
NEXT_COMM_LGTH	!Z/M	[7]	Length of revision comment for *next* revision
NEXT_COMM_RV	!Z/L	[7]	Is *next* edit comment indicative of a revert? (boolean)
QUALITY_AVG	!Z/T	[7]	Average of implicit feedback from judges
QUALITY_MIN	!Z/T	[7]	Worst feedback from any judge
DISSENT_MAX	!Z/T	[7]	How close QUALITY_AVG is to QUALITY_MIN
REVERT_MAX	!Z/T	[7]	Max reverts possible given QUALITY_AVG
WT_REPUTATION	!Z/R	[7]	Editor rep. per WikiTrust (permitting future data)
JUDGES_WGHT	!Z/R	[7]	Measure of relevance of implicit feedback

To give an idea of scale:

The combination of the three methods results in 70+ data points/features being given to the machine-learning framework

Problem space is quite well-covered!

Meta-Algorithm (2)



Text = Shallow props. -- Language = Vocabularies

- Combined approach dominated with PAN-2011 winning technique
  - Unique capture
  - Current baseline!
- High precision suggests bot-operation success
- Vocabulary data helpful when present; but “rare words” hurt
- Online implementation

PAN 2011

2011 Vandalism Detection Competition [12]

Two rule changes relative to 2010:

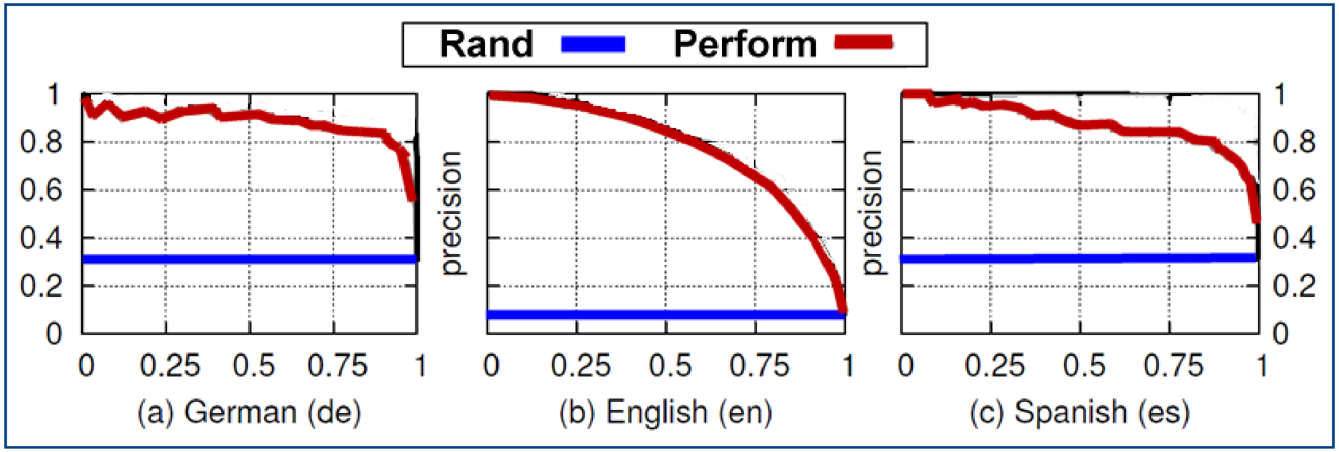
1. Train/test corpora span three natural languages (German, English, Spanish)
2. The ability to leverage ex post facto evidence (after the edit was made)

Notebook papers not published until September:

- However, results have been revealed
- Fortunate to explain the most successful approach [13]

PAN 2011: Languages

- Strategy: More metadata, less language-specific features
  - Create a portable model applicable for 197+ editions
- Evaluation results:
  - Consistent feature strength
  - Language features prove moderately helpful when included
  - Why is English out-performed?

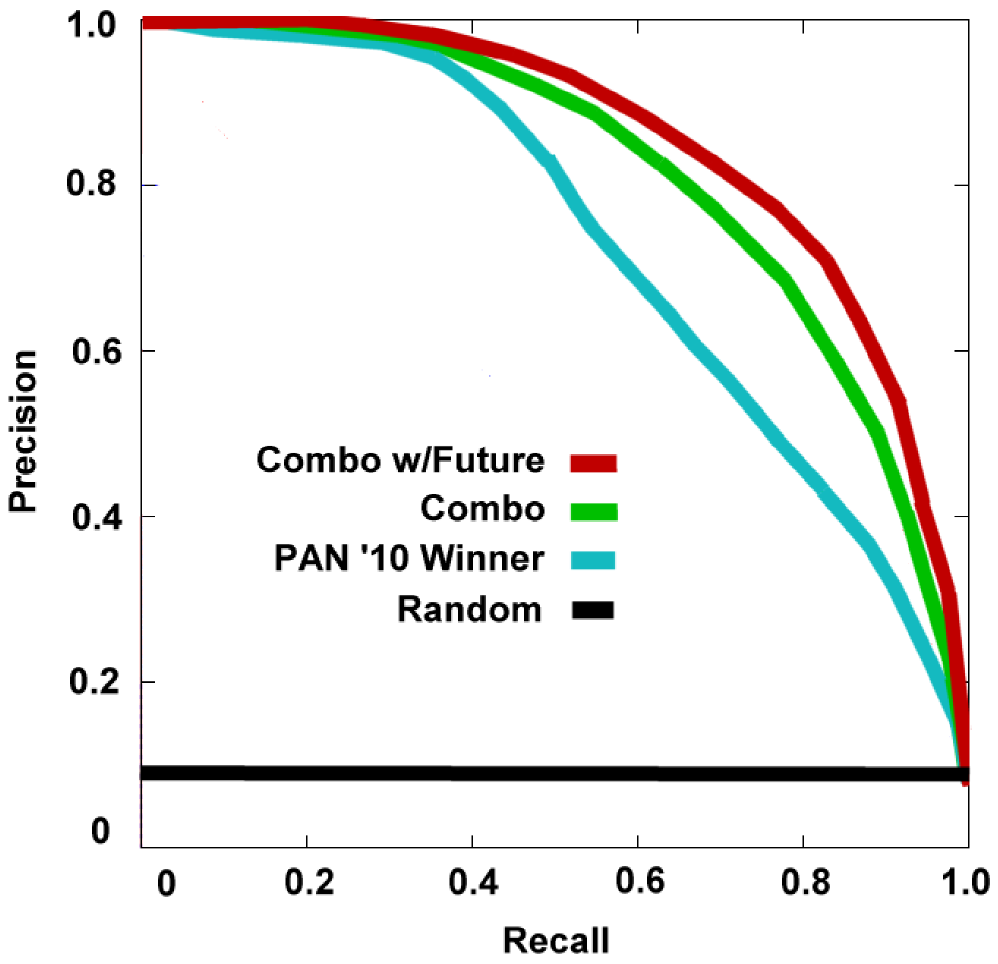


PAN 2011: Ex Post Facto

- Motivating example: Wikipedia 1.0 Project
- Use “future” evidence after edit was committed to help score.
  - E.g.: Has the content persisted? What is the next comment?
  - WikiTrust system a specialist at this task (helps WP1.0)
  - Surprisingly minimal performance increase (next slide)

Time	User	Comment
Jan. 1	Jimbo	“Initializing article”
Feb. 6	111.37.*.*	(null)
Jun. 5	west.andrew	“RV vand. by 111.37.*.*”
Jun. 5	NewishUser	“Add recent events”
Aug. 4	120.831.*.*	“I is super vandal!”
➡ What version should one pick?		

PAN 2011: Ex Post Facto

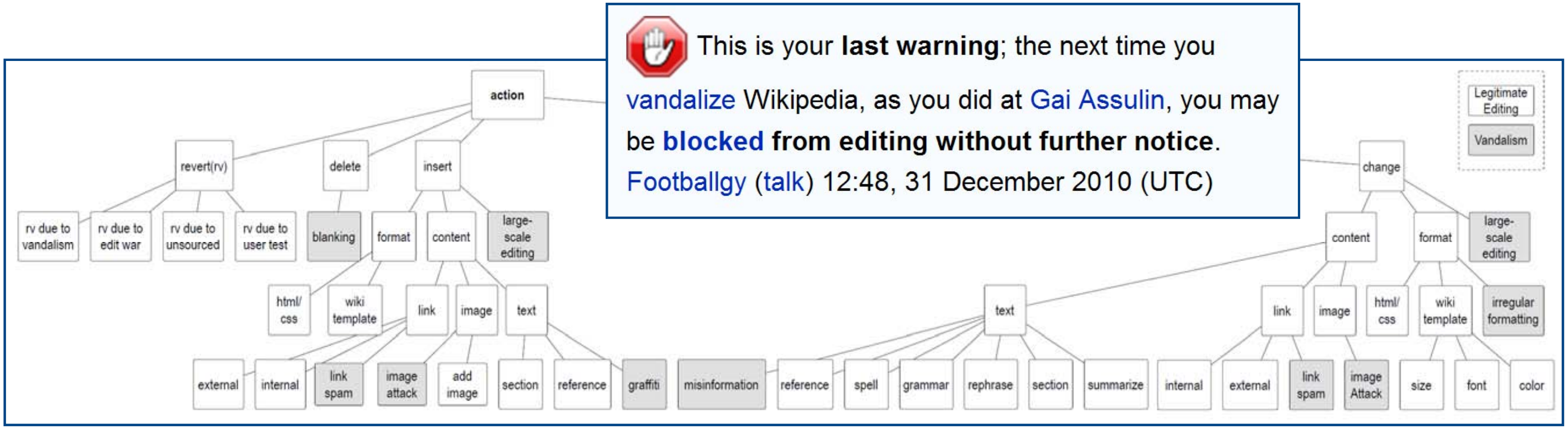


Why is there not a greater performance increase?

Possibly subjective nature of vandalism?

Misc. Research

- “… Active Learning and Statistical Language models” [14]
  - Concentrate on tagging types of vandalism.
  - Could use to create type-specific models
- “… the Banning of a Vandal” [15]
  - Formal look at warning process, Huggle, AIV, *etc.*



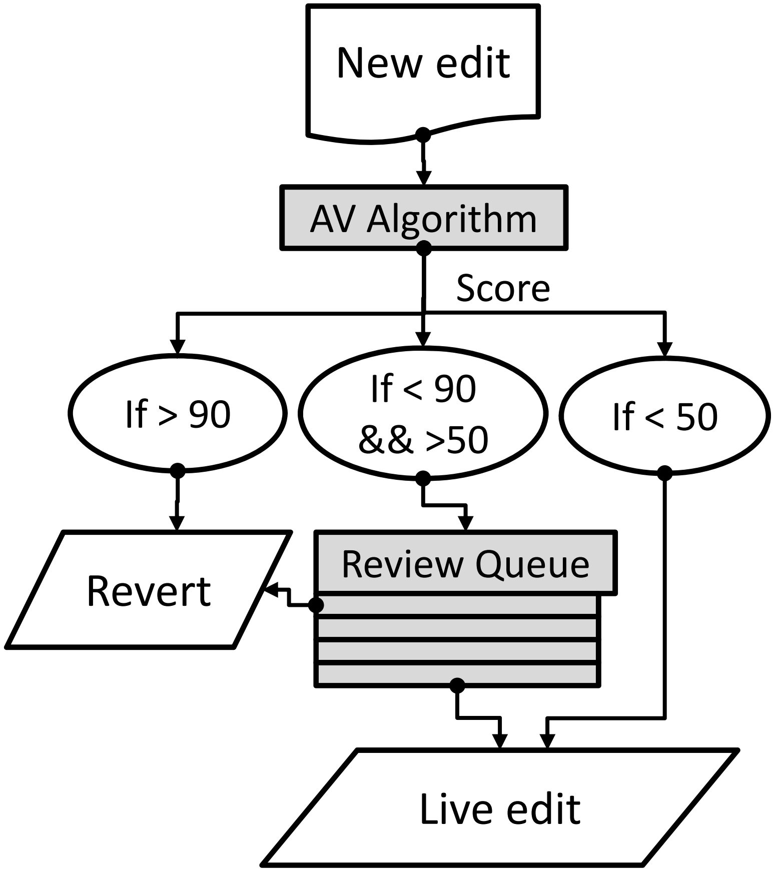
This is your **last warning**; the next time you vandalize Wikipedia, as you did at Gai Assulin, you may be **blocked from editing without further notice**. Footballgy (talk) 12:48, 31 December 2010 (UTC)

Future of Anti-Vandalism

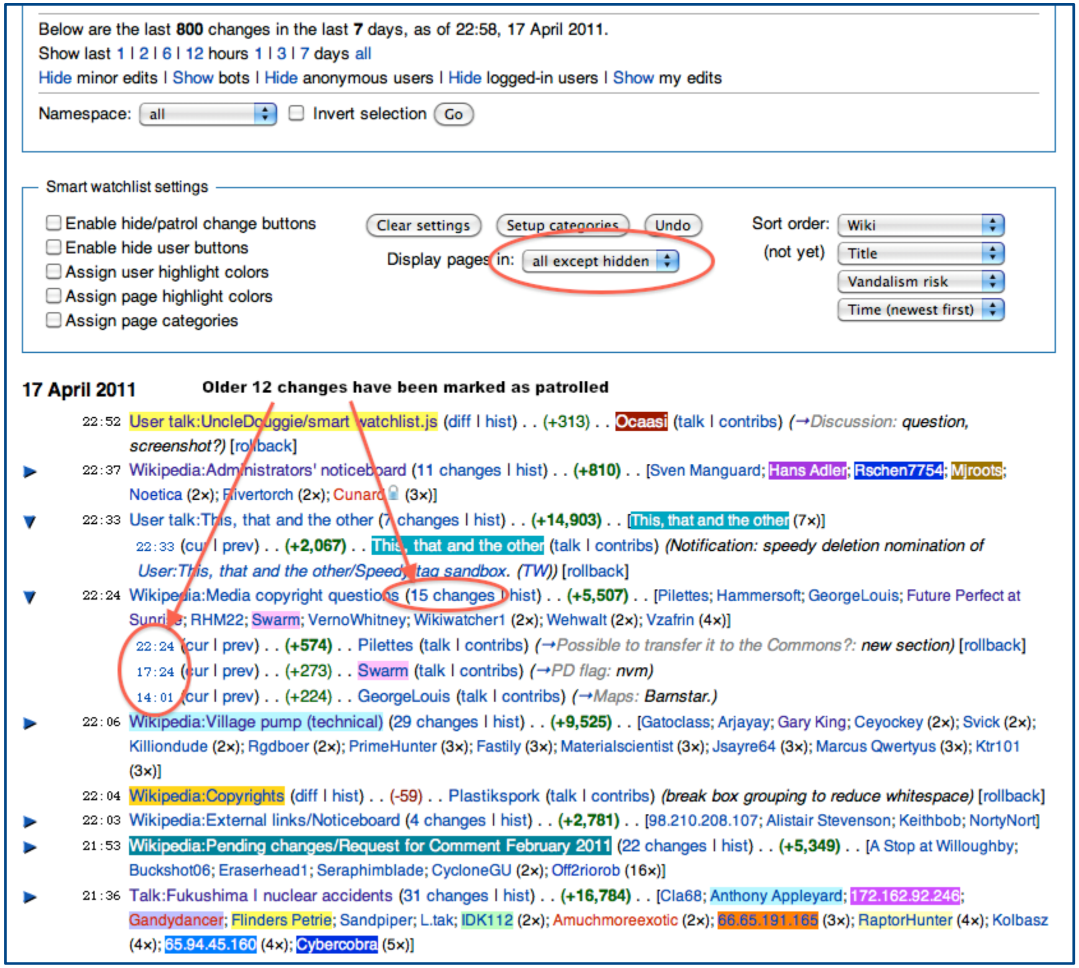
Future: Pend. Changes

- Basically like taking the STiki queue, and moving top edits under PC

- Reduce [[WP:PC]] workload
  - 1/3 of all PC reviewed edits were innocent
  - Avoid [[WP:Bite]]
  - No one has to maintain “protected pages” lists



Future: Watchlists



Below are the last **800** changes in the last **7** days, as of 22:58, 17 April 2011.
Show last 1 | 2 | 6 | 12 hours 1 | 3 | 7 days all
Hide minor edits | Show bots | Hide anonymous users | Hide logged-in users | Show my edits

Namespace: all ☐ Invert selection Go

Smart watchlist settings

☐ Enable hide/patrol change buttons
☐ Enable hide user buttons
☐ Assign user highlight colors
☐ Assign page highlight colors
☐ Assign page categories

Clear settings | Setup categories | Undo

Display pages in: all except hidden

Sort order: Wiki
(not yet) Title
Vandalism risk
Time (newest first)

17 April 2011 **Older 12 changes have been marked as patrolled**

22:52 User talk:UncleDouggie/smart watchlist.js (diff | hist) . . (+313) . . Ocaasi (talk | contribs) *(→Discussion: question, screenshot?)* [rollback]

▶ 22:37 Wikipedia:Administrators' noticeboard (11 changes | hist) . . **(+810)** . . [Sven Manguard; Hans Adler; Rschen7754; Mjroots; Noetica (2×); Rivertorch (2×); Cunard (3×)]

▼ 22:33 User talk:This, that and the other (7 changes | hist) . . **(+14,903)** . . [This, that and the other (7×)]

22:33 (cur | prev) . . **(+2,067)** . . This, that and the other (talk | contribs) *(Notification: speedy deletion nomination of User:This, that and the other/Speedy tag sandbox. (TW))* [rollback]

▼ 22:24 Wikipedia:Media copyright questions (15 changes | hist) . . **(+5,507)** . . [Pilettes; Hammersoft; GeorgeLouis; Future Perfect at Sunrise; RHM22; Swarm; VernoWhitney; Wikiwatcher1 (2×); Wehwalt (2×); Vzafrin (4×)]

22:24 (cur | prev) . . **(+574)** . . Pilettes (talk | contribs) *(→Possible to transfer it to the Commons?: new section)* [rollback]

17:24 (cur | prev) . . (+273) . . Swarm (talk | contribs) *(→PD flag: nvm)*

14:01 (cur | prev) . . (+224) . . GeorgeLouis (talk | contribs) *(→Maps: Barnstar.)*

▶ 22:06 Wikipedia:Village pump (technical) (29 changes | hist) . . **(+9,525)** . . [Gatoclass; Arjayay; Gary King; Ceyockey (2×); Svick (2×); Killiondude (2×); Rgdboer (2×); PrimeHunter (3×); Fastily (3×); Materialscientist (3×); Jsayre64 (3×); Marcus Qwertyus (3×); Ktr101 (3×)]

22:04 Wikipedia:Copyrights (diff | hist) . . (-59) . . Plastikspork (talk | contribs) *(break box grouping to reduce whitespace)* [rollback]

▶ 22:03 Wikipedia:External links/Noticeboard (4 changes | hist) . . **(+2,781)** . . [98.210.208.107; Alistair Stevenson; Keithbob; NortyNort]

▶ 21:53 Wikipedia:Pending changes/Request for Comment February 2011 (22 changes | hist) . . **(+5,349)** . . [A Stop at Willoughby; Buckshot06; Eraserhead1; Seraphimblade; CycloneGU (2×); Off2riorob (16×)]

▶ 21:36 Talk:Fukushima I nuclear accidents (31 changes | hist) . . **(+16,784)** . . [Cla68; Anthony Appleyard; 172.162.92.246; Gandydancer; Flinders Petrie; Sandpiper; L.tak; IDK112 (2×); Amuchmoreexotic (2×); 66.65.191.165 (3×); RaptorHunter (4×); Kolbasz (4×); 65.94.45.160 (4×); Cybercobra (5×)]

http://en.wikipedia.org/wiki/User_talk:UncleDouggie/smart_watchlist.js

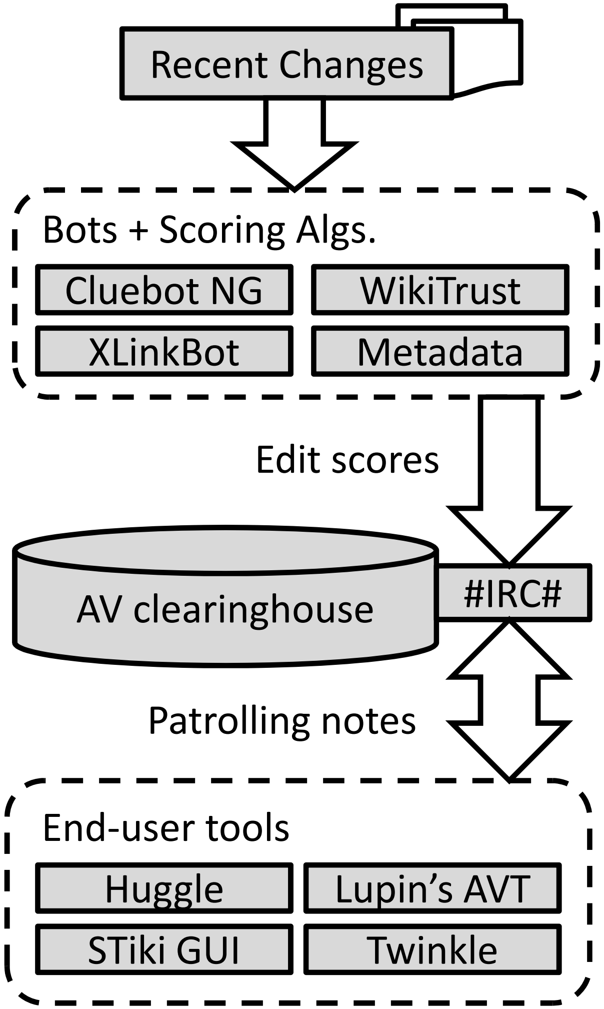
Future: MW Support

Vandalism clearinghouse

- Bot/GUI collaboration
- Explicit and formal “innocent”
- Akin to new page patrol

WMF support for software

- Provide host space for AV algorithms (reliability!)



Recent Changes

Bots + Scoring Algs.

Cluebot NG

WikiTrust

XLinkBot

Metadata

Edit scores

AV clearinghouse

#IRC#

Patrolling notes

End-user tools

Huggle

Lupin’s AVT

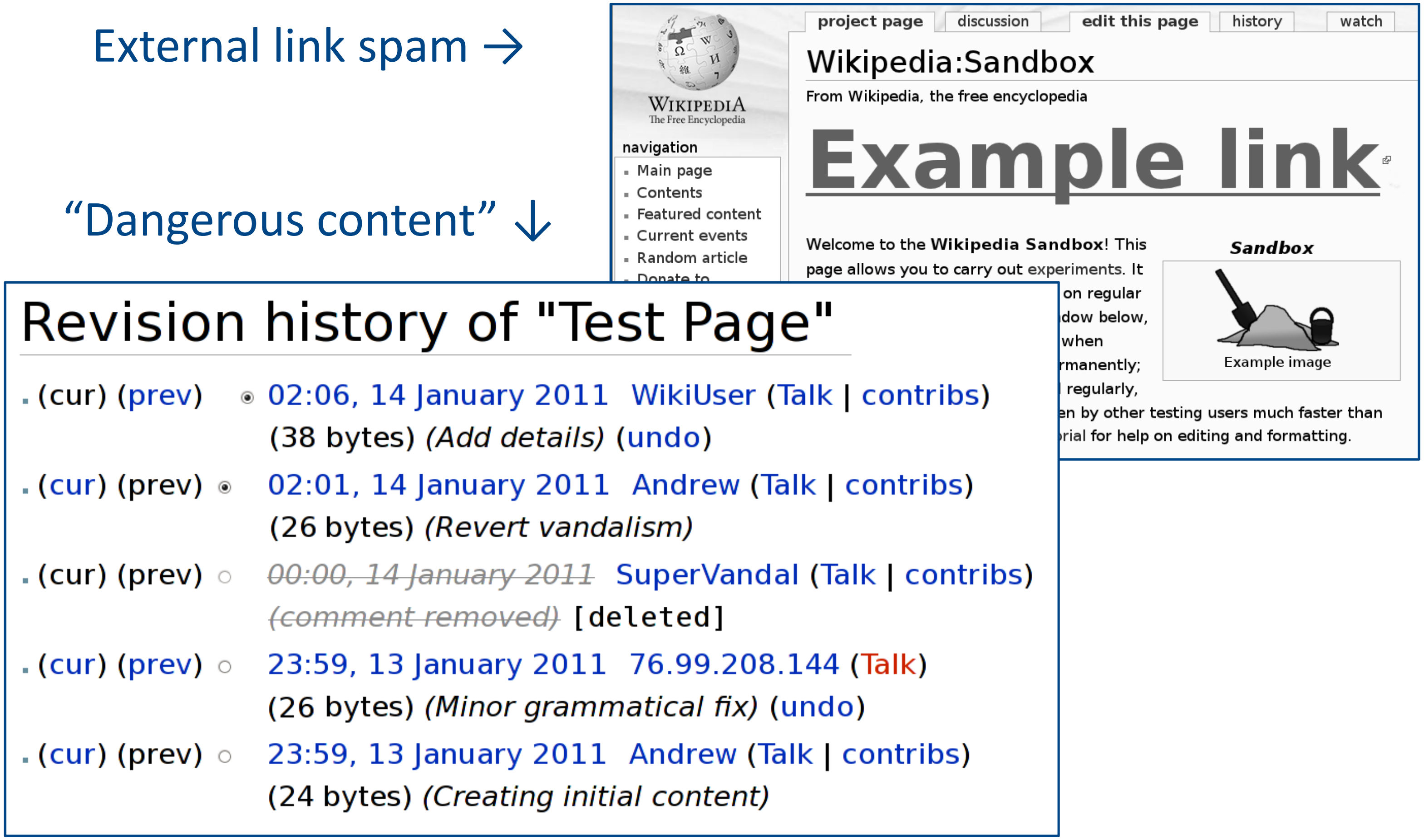
STiki GUI

Twinkle

Future: Acute Subsets

External link spam →

"Dangerous content" ↓



Wikipedia:Sandbox

From Wikipedia, the free encyclopedia

Example link

Welcome to the **Wikipedia Sandbox**! This page allows you to carry out experiments. It … on regular … dow below, … when … rmanently; … regularly, … en by other testing users much faster than … rial for help on editing and formatting.

Sandbox

Example image

Revision history of "Test Page"

- (cur) (prev) ● 02:06, 14 January 2011 WikiUser (Talk | contribs) (38 bytes) *(Add details)* (undo)
- (cur) (prev) ● 02:01, 14 January 2011 Andrew (Talk | contribs) (26 bytes) *(Revert vandalism)*
- (cur) (prev) ○ ~~*00:00, 14 January 2011*~~ SuperVandal (Talk | contribs) ~~*(comment removed)*~~ [deleted]
- (cur) (prev) ○ 23:59, 13 January 2011 76.99.208.144 (Talk) (26 bytes) *(Minor grammatical fix)* (undo)
- (cur) (prev) ○ 23:59, 13 January 2011 Andrew (Talk | contribs) (24 bytes) *(Creating initial content)*

References (1)

[1] A.G. West, J. Chang, K. Venkatasubramanian, and I. Lee. **Trust in Collaborative Web Applications**. In *Future Generation Computer Systems*, Elsevier Press, 2011.

[2] *http://en.wikipedia.org/wiki/User:Emijrp/Anti-vandalism_bot_census*

[3] B.T. Adler and L. de Alfaro. **A content-driven reputation system for the Wikipedia**. In *WWW'07: International World Wide Web Conference*, 2007

[4] **ClueBot NG**. *http://en.wikipedia.org/wiki/User:ClueBot_NG*

[5] A.G. West, S. Kannan, and I. Lee. **Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata**. In *EUROSEC 2010*.

[6] **Huggle Anti-Vandalism Tool**. *http://en.wikipedia.org/wiki/WP:Huggle*

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

[8] S.M.M. Velasco. **Wikipedia vandalism detection through machine learning: Feature review and new proposals**. In *Notebook Papers of CLEF 2010 Labs*.

[9] M. Potthast. **Crowdsourcing a Wikipedia vandalism corpus**. In *SIGIR 2010*.

References (2)

[10] M. Potthast, B. Stein, and T. Holfeld. **Overview of the 1st Intl. competition on Wikipedia vandalism detection**. In *Notebook Papers of CLEF 2010 Labs*.

[11] B. Adler, L. de Alfaro, S.M. Mola-Velasco, P. Rosso, and A.G. West. **Wikipedia vandalism detection: Combining natural language, metadata, and reputation features**. In *CICLing `11: Intelligent Text Proc. and Computational Linguistics*, 2011.

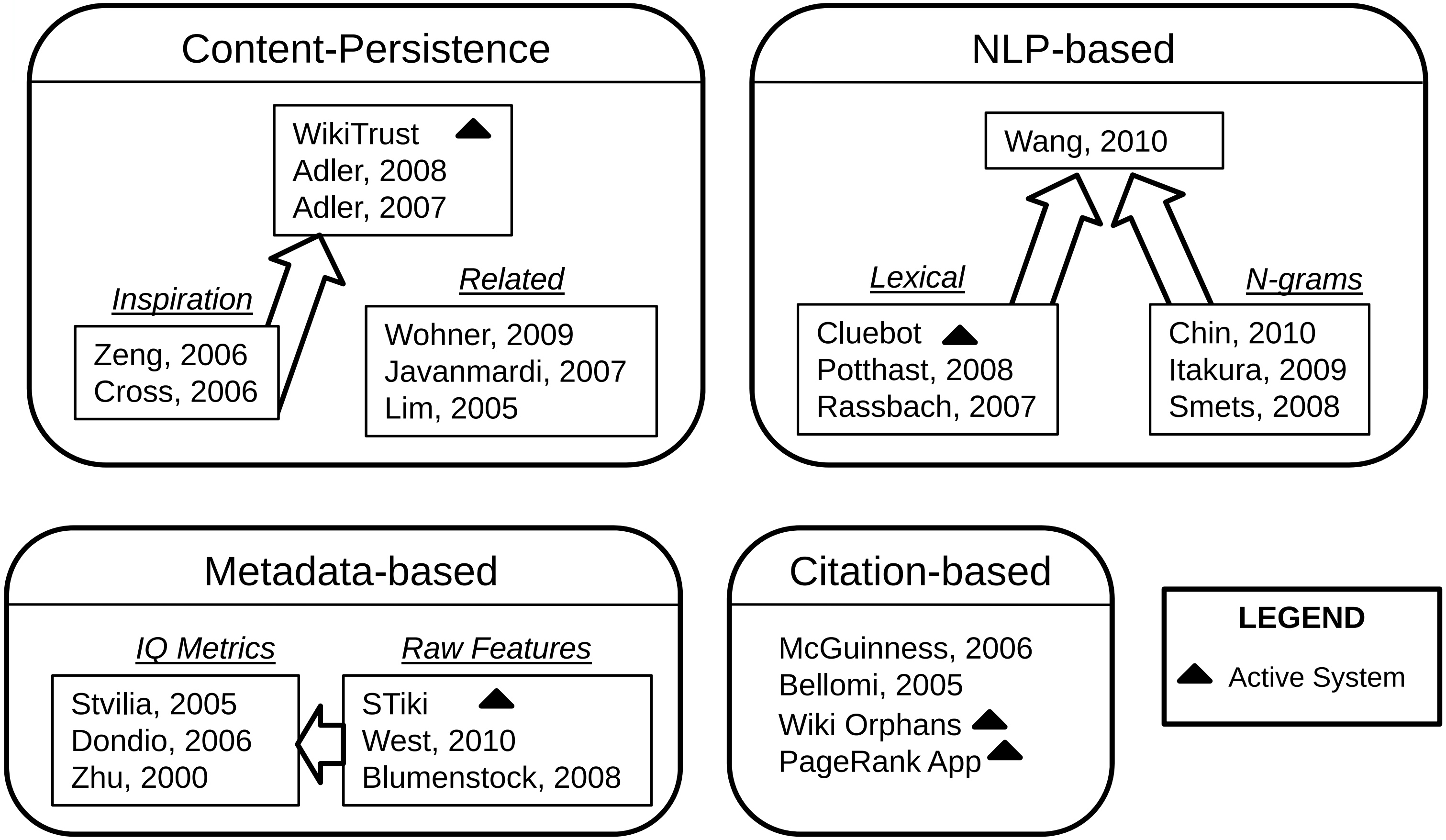
[12] **2011 Wikipedia Vandalism *Detection.*** Part of the *CLEF Labs on Uncovering Plagiarism, Authorship, and Social Software Misuse.* http://pan.webis.de/

[13] A.G. West and I. Lee. **Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence**. In *Notebook Papers CLEF 2010 Labs*.

[14] S. Chin, P. Srinivasan, W.N. Street, and D. Eichmann. **Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models**. In *WICOW 2010*.

[15] R.S. Geiger, and D. Ribes. **The work of sustaining order in Wikipedia: The banning of a vandal**. In *CSCW'10: Conf. on Computer Supported Cooperative Work*, 2010.

Backup Slides (1)



Content-Persistence

WikiTrust ▲
Adler, 2008
Adler, 2007

Inspiration

Zeng, 2006
Cross, 2006

Related

Wohner, 2009
Javanmardi, 2007
Lim, 2005

NLP-based

Wang, 2010

Lexical

Cluebot ▲
Potthast, 2008
Rassbach, 2007

N-grams

Chin, 2010
Itakura, 2009
Smets, 2008

Metadata-based

IQ Metrics

Stvilia, 2005
Dondio, 2006
Zhu, 2000

Raw Features

STiki ▲
West, 2010
Blumenstock, 2008

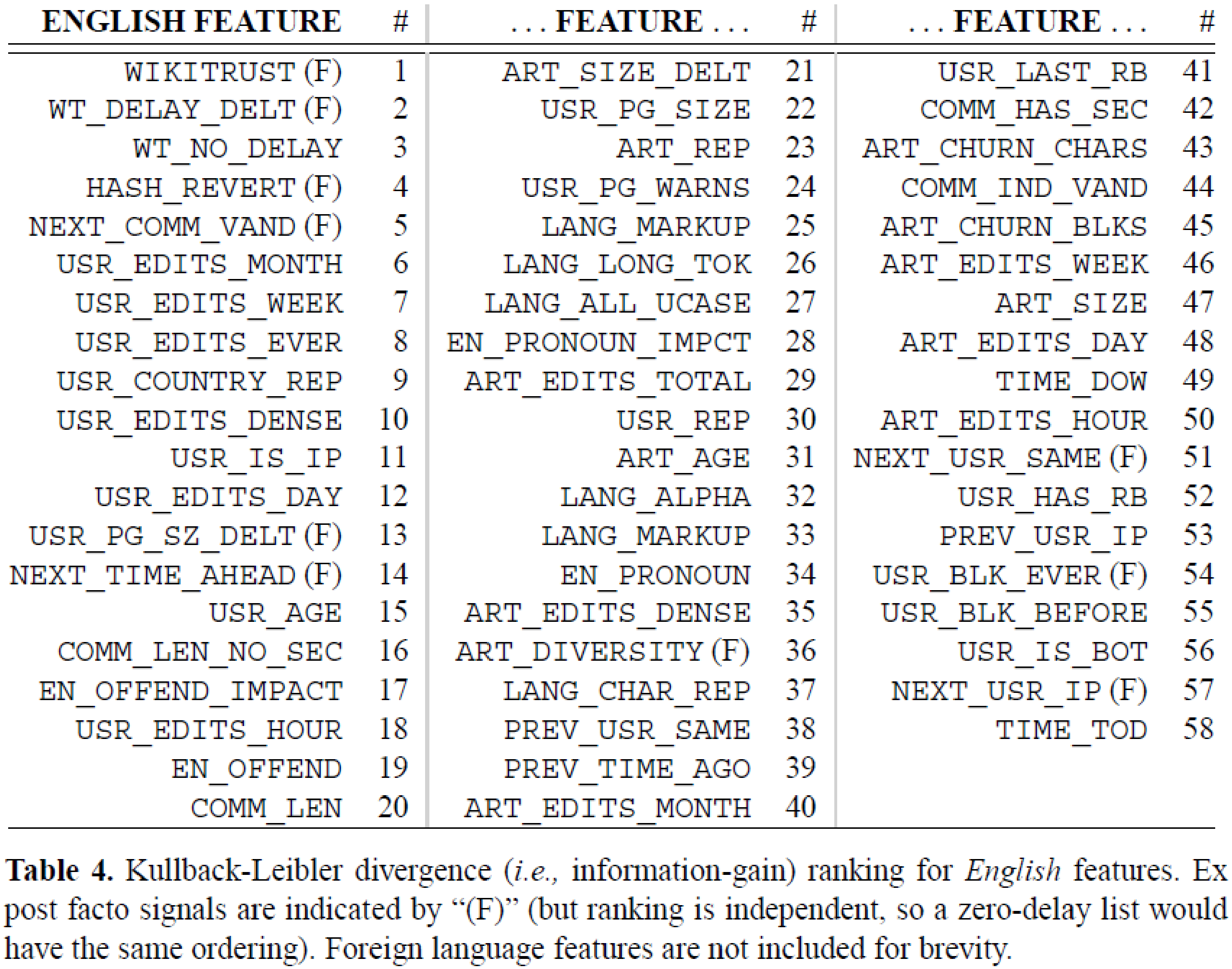
Citation-based

McGuinness, 2006
Bellomi, 2005
Wiki Orphans ▲
PageRank App ▲

LEGEND

▲ Active System

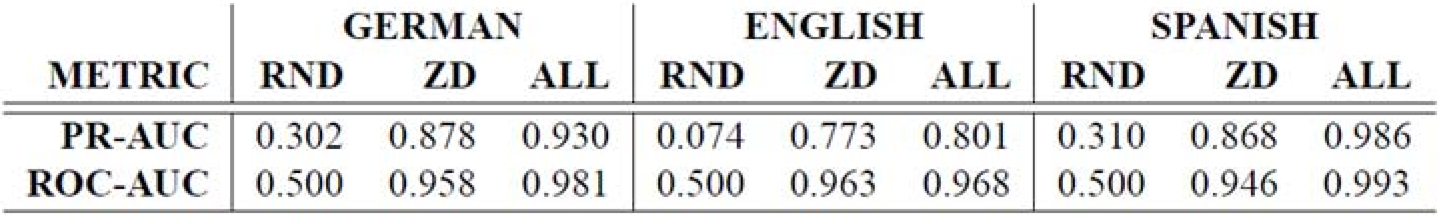
Backup Slides (2)



ENGLISH FEATURE	#	. . . FEATURE . . .	#	. . . FEATURE . . .	#
WIKITRUST (F)	1	ART_SIZE_DELT	21	USR_LAST_RB	41
WT_DELAY_DELT (F)	2	USR_PG_SIZE	22	COMM_HAS_SEC	42
WT_NO_DELAY	3	ART_REP	23	ART_CHURN_CHARS	43
HASH_REVERT (F)	4	USR_PG_WARNS	24	COMM_IND_VAND	44
NEXT_COMM_VAND (F)	5	LANG_MARKUP	25	ART_CHURN_BLKS	45
USR_EDITS_MONTH	6	LANG_LONG_TOK	26	ART_EDITS_WEEK	46
USR_EDITS_WEEK	7	LANG_ALL_UCASE	27	ART_SIZE	47
USR_EDITS_EVER	8	EN_PRONOUN_IMPCT	28	ART_EDITS_DAY	48
USR_COUNTRY_REP	9	ART_EDITS_TOTAL	29	TIME_DOW	49
USR_EDITS_DENSE	10	USR_REP	30	ART_EDITS_HOUR	50
USR_IS_IP	11	ART_AGE	31	NEXT_USR_SAME (F)	51
USR_EDITS_DAY	12	LANG_ALPHA	32	USR_HAS_RB	52
USR_PG_SZ_DELT (F)	13	LANG_MARKUP	33	PREV_USR_IP	53
NEXT_TIME_AHEAD (F)	14	EN_PRONOUN	34	USR_BLK_EVER (F)	54
USR_AGE	15	ART_EDITS_DENSE	35	USR_BLK_BEFORE	55
COMM_LEN_NO_SEC	16	ART_DIVERSITY (F)	36	USR_IS_BOT	56
EN_OFFEND_IMPACT	17	LANG_CHAR_REP	37	NEXT_USR_IP (F)	57
USR_EDITS_HOUR	18	PREV_USR_SAME	38	TIME_TOD	58
EN_OFFEND	19	PREV_TIME_AGO	39		
COMM_LEN	20	ART_EDITS_MONTH	40		

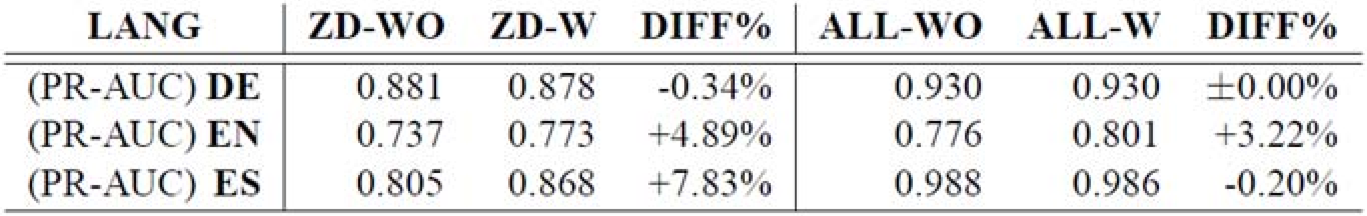
Table 4. Kullback-Leibler divergence (*i.e.,* information-gain) ranking for *English* features. Ex post facto signals are indicated by "(F)" (but ranking is independent, so a zero-delay list would have the same ordering). Foreign language features are not included for brevity.

Backup Slides (3)



METRIC	GERMAN RND	GERMAN ZD	GERMAN ALL	ENGLISH RND	ENGLISH ZD	ENGLISH ALL	SPANISH RND	SPANISH ZD	SPANISH ALL
PR-AUC	0.302	0.878	0.930	0.074	0.773	0.801	0.310	0.868	0.986
ROC-AUC	0.500	0.958	0.981	0.500	0.963	0.968	0.500	0.946	0.993

Table 6. Area-under-curve (AUC) measurements for feature sets over training data. This is done for precision-recall (PR) and receiver-operating characteristic (ROC) curves. Feature sets include a control classifier (random, RND), zero-delay (ZD), and including ex post facto data (ALL).



LANG	ZD-WO	ZD-W	DIFF%	ALL-WO	ALL-W	DIFF%
(PR-AUC) **DE**	0.881	0.878	-0.34%	0.930	0.930	±0.00%
(PR-AUC) **EN**	0.737	0.773	+4.89%	0.776	0.801	+3.22%
(PR-AUC) **ES**	0.805	0.868	+7.83%	0.988	0.986	-0.20%

Table 7. Measuring the impact of language-specific features (Tab. 3). Feature sets are evaluated with (W) and without (WO) the inclusion of language-specific signals. Otherwise, acronyms are as defined as in Tab. 6. PR-AUC is the singular metric used in this comparison.