Automated Decision Support for Human Tasks in a Collaborative System

The Case of Deletion in Wikipedia

Bluma S. Gelley and Torsten Suel
Polytechnic Institute of NYU

WikiSym 2013, August 5, 2013

Deleted Articles

- “Ingall Services is a gardening and car-washing group/company lead and founded by headworker teenager Dave Ingall. Ingall Services originated in August 2011. It's based on Cedar Way and operates to the surrounding areas.”
- “Dylan Campbell is a 10 year old computer genius who lives in the eastern USA. He enjoys surfing the web, film making, and chatting with his friends via video chat.”
- “Bruce L Rastetter is a Iowa business leader and a political activist. He has started many agricultural based businesses ranging from pork production to ethanol production.”

Deletion

- Hundreds of deletions
- ~1,000 new articles/day
  - A lot of time spent on patrolling
  - Most patrollers also contribute to the rest of the site
- Speed of deletion:
  - Many complain about speed and abruptness of the deletion process (e.g. Ford & Geiger, 2012)
  - 47% of Speedy-Deleted articles are nominated within 10 minutes of creation (Gelley, 2013)
  - 70% of *new* users whose first article is nominated for deletion have the nomination occurring within 10 minutes of article creation (Geiger/WMF, 2011)

Editor Retention

- New editors who start by creating an article are *6 times* more likely to abandon WP immediately if the article is deleted (User:MrZ-man study)
- Wikipedia cannot afford this

Our Work

- Until now, few hard facts about deletion
- We set out to understand better what was actually happening
  - Collect and examine deleted articles
  - Build a model to differentiate between deleted and kept articles that we can use to improve the deletion process
- Analysis of article characteristics is in a previous paper; in this work we focus mostly on the model
  - http://arxiv.org/abs/1305.5267

Our Contribution

- A model of Wikipedia articles that can distinguish between deleted and not-deleted (kept) articles
- Several datasets of deleted articles available for download and use

Types of Deletion

- Speedy Deletion
  - Most common form
  - Articles that are so unencyclopedic that they don't even require discussion – can be nominated for deletion by anyone and unilaterally deleted by any admin
  - 22 Speedy Deletion Criteria
- **PRO**posed **D**eletion (PROD)
  - Don't meet any Speedy criteria, but still unencyclopedic
  - 7-day waiting period after nomination; if anyone contests the nomination it is removed

Types of Deletion (ctd)

- Deletion Discussion (**A**rticles **f**or **D**eletion – AfD)
  - Supposed to be default deletion form
  - Articles of borderline encyclopedic quality are nominated for community discussion
  - After > 1 week, an admin determines the consensus and acts on it

Some Statistics

- Number of new AfDs/day: 30 – 100+
  - Mean 60, median 59
- Number of new PRODs/day: 30 - ~70
  - Mean 45, median 44
- Number of Speedies/day: several hundred
  - Range varies widely
- Deletion Rates
  - AfDs: ~50%
  - PRODs: ~86%
  - Speedies: > 70%

Anticipated Use Cases

- Finding articles to improve
  - New articles
  - Older articles that are borderline
- Decision Support for New Page Patrollers (NPP's)
  - Allow them to make better, more informed decisions
- Helping (new) editors evaluate and improve their articles before creation
  - Article Wizard can be confusing
  - Give feedback on likelihood of deletion
  - Perhaps give detailed instructions for how to improve

Potential Benefits

- Make the New Page Patrol and deletion processes more efficient and effective
- Reduce load on and stress on NPP's
  - Hopefully reduce 'newbie biting'
  - Allow them more time to contribute to Wikipedia
- Improve editor retention

Speedy Deletion Criteria

- Of 22 Speedy deletion criteria, the two we chose were
  - A7/A9 – “no indication of importance”
  - G11 – “unambiguous advertising and promotion”
- These two comprise ~45% of all Speedy deletions
- Most others can be found using heuristics, or vandalism detection
- We use all PRODs and AfDs without filtering

Data Collection - Speedies

- Challenge:
  - we don’t know which articles will be nominated until they are…
  - but once nominated, they can be deleted at any time! (and often are within minutes [Geiger/WMF 2011])
- Solution:
  - Check Candidates for Speedy Deletion page every few minutes and download newly nominated articles
  - Later, check if they were deleted

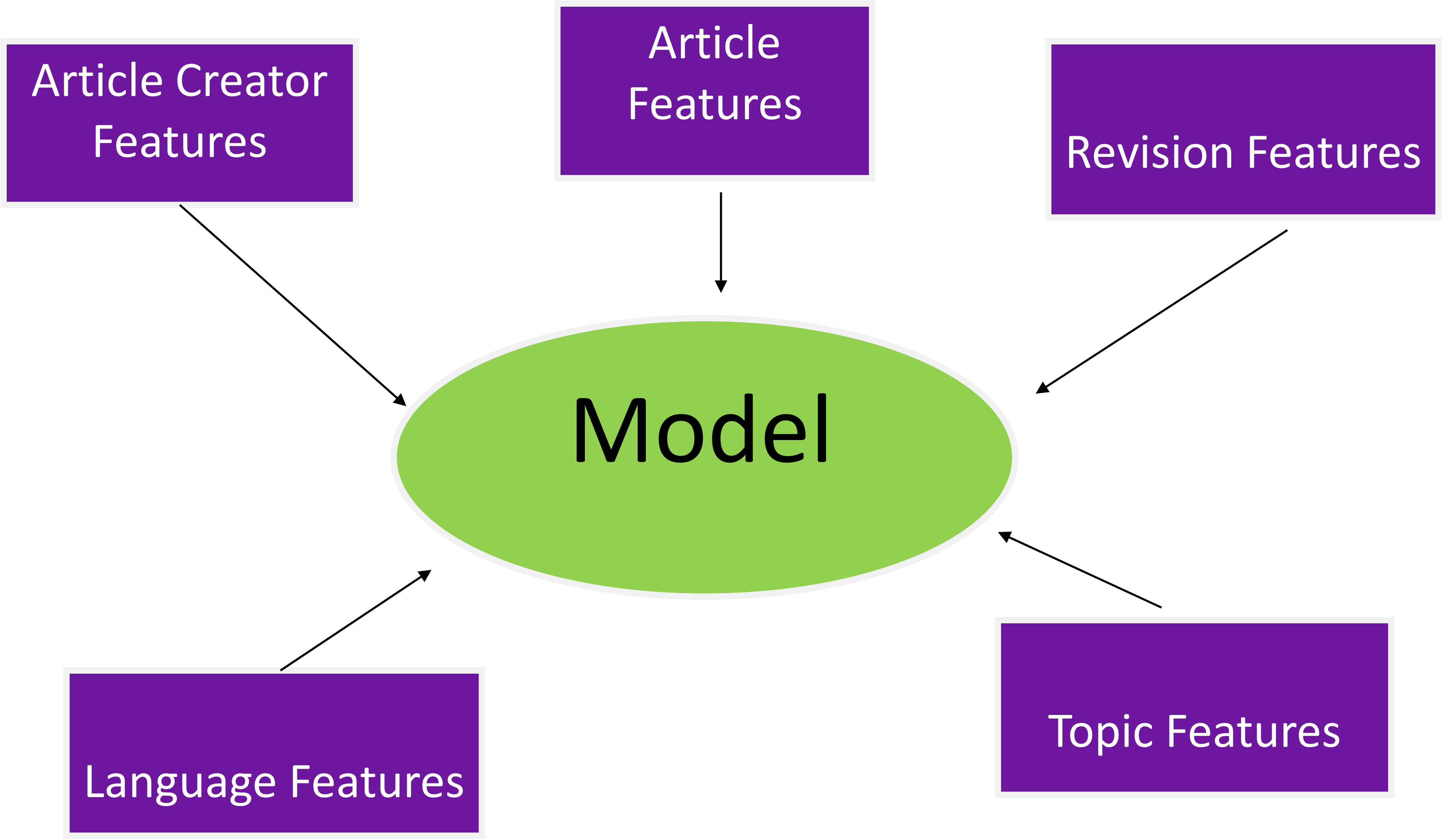
Dataset Summary

Name	Description	Kept Articles	Deleted Articles	Total
AfD	AfD's Jan-Mar 2013	798	270	1068
PRODs	Oct-Dec '11; Mar '13; kept set from similar articles	2036	991	3027
Original	Oct-Dec '11; kept set from similar articles	1381	2444	3825
Old	Articles from Original set > 1 week	580	191	771
New	Downloaded shortly after creation	2198	723	2921

Get this data

- All datasets available for download at https://github.com/bsgelley/Wikipedia-deletion-data

Features



Article Creator Features

Article Features

Revision Features

Model

Language Features

Topic Features

Classification

- Random Forests of 40 trees (high accuracy, low overhead, used in similar tasks)
- Weka Machine learning suite
- Larger datasets = 70-30 split; smaller (Old and AfD) = 10-fold cv over entire dataset

Baseline

- No known work to compare to
- Experiment on different datasets and compare results
- Also experiment on different feature sets

Results - Overview

	Precision	Recall
AfD	96%	33%
PROD	98%	71%
Speedy	98.6%	97.5%

Results - Speedies

	Original	Old	New
Baseline (Zero-R)	**63.42%**	**75%**	**72.7%**
All features	**97.57**	**92.6**	**95.21**
No language features	97.22	91.31	95.55
Language Features	96.18	93.8	78.0
Creator Features	91.49	85.47	92.8
Revision Features	95.04	82.1	83.0
Article Features	90.88	82.1	85.39
Non time-bound	95.79	N/A*	N/A*
Bag of Words (SVM)	96.55		
2011 training, 2012 test	96.4		

Results - Discussion

- Best results on Speedies, then PRODs, then AfDs
  - This was what we expected
- Original set accuracy was very high for all feature sets
- Old set results were good enough to show that the model generalizes to older pages
- New set results also very high
- Likelihood of bias is low

Impact on Editor Retention

- Use of automated tools has been shown (Geiger, et. al. 2012, Halfaker et. al. 2012) to reduce editor retention
- Last thing we want!
- We are confident that this system will not
  - Decision support vs. assisted editing
  - Careful deployment and testing
  - Relieving some of the burden on patrollers may actually decrease their aggressiveness

Conclusion

- We built a model that can differentiate between kept and deleted articles with high precision and good recall
- Our model can be used in decision-support tools for various purposes
- With thoughtful design and careful deployment, the benefits should outweigh the risks

Future Work

- Implementation as a set of decision-support tools
- Find optimal feature combinations, including new features
  - Topic modeling in particular
- Comprehensive review of deleted articles to determine if they were rightly deleted

Questions?

Thank You!