

Data-Driven Proactive Policy Assurance of Post Quality in Community Q&A Sites

Chunyang Chen, Xi Chen, Jiamou Sun, Zhenchang Xing, Guoqiang Li







Chen, Chunyang, Xi Chen, Jiamou Sun, Zhenchang Xing, and Guoqiang Li. "Data-Driven Proactive Policy Assurance of Post Quality in Community Q&A Sites." *Proceedings of the ACM on human-computer interaction* 2, no. CSCW (2018): 33.



Background

Q&A sites are popular for sharing knowledge

- Social Q&A sites
- Technical Q&A sites





Motivation

The quality of Q&A sites are decaying

- Stack Overflow
 - 17M questions, 26M answers, 9.6M users
 - 7K new questions/day, many new users
- Complains:
 - Why do so many good programmers waste their time on Stack Overflow?
 - Farewell Stack Exchange
 - The decline of Stack Overflow





Motivation

To keep the quality of content

1. Publish community norms

- https://stackoverflow.com/help/how-to-ask
- <u>https://stackoverflow.com/help/how-to-answer</u>

Provide context for links

Links to external resources are encouraged, but please add context around the link so your fellow users will have some idea what it is and why it's there. Always quote the most relevant part of an important link, in case the target site is unreachable or goes permanently offline.

Write to the best of your ability

We don't expect every answer to be perfect, but answers with correct spelling, punctuation, and grammar are easier to read. They also tend to get upvoted more frequently. Remember, you can always go back at any time and edit your answer to improve it.

Proof-read before posting!

Now that you're ready to ask your question, take a deep breath and read through it from start to finish. Pretend you're seeing it for the first time: *does it make sense*? Try reproducing the problem yourself, in a fresh environment and make sure you can do so using only the information included in your question. Add any details you missed and read through it again. Now is a good time to make sure that your title still describes the problem!

Problem: Users **do not** read or understand the instructions.

Chen, Chunyang, Zhenchang Xing, and Yang Liu. "By the Community & For the Community: A Deep Learning Approach to Assist Collaborative Editing in Q&A Sites." *Proceedings of the ACM on Human-Computer Interaction* 1, no. CSCW (2017): 32.



Motivation

To keep the quality of content

2. Peer review

- https://stackoverflow.com/help/privileges/edit
- 2M question-title edits (17.6%)
- 3M question-tag edits (12.9%)
- 21M post-body edits (36.2%)

When should I edit posts?

Any time you feel you can make the post better, and are inclined to do so. Editing is encouraged!

Some common reasons to edit are:

- to fix grammatical or spelling mistakes
- · to clarify the meaning of a post without changing it
- · to correct minor mistakes or add addendums / updates as the post ages
- · to add related resources or hyperlinks

Problem:

- Require significant community efforts;
- Some edits are difficult to locate;
- The policy violation has hurt readers before edits



Goal

To keep the quality of content

- We need a way to help policy assurance of post quality
 - **Proactive**: remind users before they publish the posts
 - Data-driven: learn from real existing edits





Observation

Observe the existing edits

Four different kinds of middle-level edits

- Code format edit
- Text format edit
- Link modification
- Image revision

1.1 Minor Revision (Spelling)	
I've found SVN to be extremelly usefull for documentation, personal	I've found SVN to be extremely useful for documentation, personal files,
files, among other non-source code uses. What other practical uses	among other non-source code uses. What other practical uses have
have you found to version control systems in general?	you found to version control systems in general?
1.2 Minor Revision (Grammar)	
Why should I move away from them as long as it works on my site?	Why should I move away from them as long as they work on my site?
2. Code Edit (Format)	
<pre>\$safe_variable = mysql_real_escape_string(\$_POST["user-input"] mysql_query("INSERT INTO table (column) VALUES ('" . \$safe_var</pre>	<pre>\$safe_variable = mysql_real_escape_string(\$_POST["user-input"]); mysql_query("INSERT INTO table (column) VALUES ('" . \$safe_variab</pre>
	< >>
3. Text Edit (Format)	
However, you may start a standalone instance as a replica set by using	However, you may start a standalone instance as a replica set by using
the command, 1) Start Mongo server in replica mode. mongod	the command,
dbPathreplSet rs0 2) Initiate the replica set.(Execute from mongo	1. Start Mongo server in replica mode.
sheir) is.initiate(),	 mongoddbPath <path data="" file="" to="">replSet rs0</path>
	2. Initiate the replica set. (Execute from mongo shell)
	rs.initiate();
4. Link Modification	
Get Bruce Schneier's book Applied Cryptography and read it carefully.	Get Bruce Schneier's book Applied Cryptography and read it carefully.
5. Image Revision	
You can just rearrange your keys on your current keyboard and change the layout.	You can just rearrange your keys on your current keyboard and change the layout.
Here is the key layout: alt text http://rffr.de/images/dvorak.jpg	Here is the key layout:



Observation

Observe the existing edits

Each edit including

- Insert
- Replace
- Delete





Data Collection

Collecting the dataset of <original-post, post-body-edit-type>

- Regular expression and text differencing
- Data for different edits
 - Adding code format: 1,567,272
 - Adding text format: 52,945
 - Adding hyperlinks: 1,126,252
 - Adding images: 219,215





Approach

CNN model for edit prediction

- Word embedding
 - Convert the word into vector representation
- Convolutional Layer
 - Kernel filter sliding within the input matrix
- Maxpooling
 - Preserve the salient information
- Fully-connected layer
 - Final prediction





Approach

Locating the Key Phrases in Posts to Explain the Edit Prediction

- Tracing back through the model to locating the filtered phrases in the input layer
- Predicting the contribution score of the phrases' corresponding features in the fully connected layer to the prediction class





Evaluation

Performance comparison between our model and baselines

- Evaluation metrics
 - Precision, recall, F1-score
- Baseline
 - Logistic regression, SVM, FastText, Attention-based LSTM







Evaluation

Understanding of edit predictions

- Locate key phrase to help understand the prediction
 - Add code format

Add in	nages
--------	-------

o understand the predic		
wrdMergeFields . Add wrdSelection . Range , ProductName) +>	<pre>wrdMergeFields.Add(wrdSelection.Range, "ProductName")</pre>	
The code above basically dispalys <mark>all the productName in differents</mark> pages in word Document during merge .	The code above basically dispalys all the productName in differents pages in word Document merge.	during
Please help me how to put the data inside a table . I have to write multiple codes of this for my ProductName , AccountNo , OutBalance , AccountName , etc . My problem here is that I do n't know how to put them in a table .	Please help me how to put the data inside a table. I have to write multiple codes of this for my ProductName, AccountNo, OutBalance, AccountName, etc. My problem here is that I don't know to put them in a table.	ow
ou can use `ismember` to find <mark>where each label exists in your cell array</mark> . ne <mark>second output</mark> will provide the index of the label . You can then use imagesc` with a custom colormap to display the result .	You can use ismember to find where each label exists in your cell array. The second output will provide the index of the label. You can then use imagesc with a custom colormap to display the result.	
% Create a copy of Grid where the empty cells are replaced with '' tmp Grid ; tmp cellfun (@ (x) ['' x] , Grid , 'UniformOutput' , false) ;	% Create a copy of Grid where the empty cells are replaced with '' tmp = Grid; tmp = cellfun($\rho(x)$ ['' x], Grid, 'UniformOutput', false);	
<pre>% Locate all of the 's' and 'i' cells and assign values of 1 and 2 spectively [`, labels] ismember (tmp, { 's', 'i' }) ; % Display the resulting inbel matrix marger labels) % Use a custom colormap where empty cells are black ?.'s' are blue and</pre>	<pre>% Locate all of the 's' and 'i' cells and assign values of 1 and 2 respectively [~, labels] = ismember(tmp, {'s', 'i'}); % Display the resulting label matrix imagesc(labels) % Use a custom colormap where empty cells are black, 's' are blue and 'i' are red cmap = [0 0 0; 0 0 1; 1 0 0]; colormap(cmap)</pre>	
<pre>i' are red cmap [0 0 0 ; 0 0 1 ; 1 0 0] ; colormap (cmap)</pre>	And if we test this with Grid = {'s', []; 'i', []}	I O niv