Wikipedia Reputation System Based on Edit History

Heng Lin[a], Liang Niu[b]

[a] *hl2521@nyu.edu*
[b] *ln932@nyu.edu*

Abstract

We will present a reputation system for the well know Wikipedia website based on the edit history of their articles. In our system, editors will gain reputation score from their edit(contribution) to some particular article, they may also lose their score if they did some bad editting behavior like vandalism[1]. Our model is based on the previous work of Adler and Alfaro [1] , in their model they didn't consider the absolute time for edit survival, we improved their model by adding factors generated from editing frequency calculated using absolute time. We have implemented a system to calculate the reputation score for all authors in an article for Wikipedia, and we also evaluate our model by visualing the reputation score for every word's author and compare it to the evaluation system in Wikipedia.

Keywords: Wikipedia, reputation system, Edit History

1. Introduction

1.1. Basic Concepts in Wikipedia

Wikipedia is a free online encyclopedia that aims to allow anyone to edit any article and create them. Wikipedia is the largest and most popular general reference work on the Internet and is ranked among the ten most popu-

lar websites. Wikipedia is owned by the nonprofit Wikimedia Foundation.[2] Wikipedia is basically a set of articles that can be editted by anyone. Anonymous user can also be able to edit any articles without registration, though Wikipedia do provide registration function. Thus it will cause some problems. For example, some articles will be editted by users maliciously. They may want to damage some articles or introduce bias or mistakes into some particular articles for their own interest. This kind of behavior is also called "Vandalism" [3]. Vandalism is very common in Wikipedia, thus there is a terminology "Edit War"[4] to describe the circumstance when serval people are trying to take control of an article. In our model, we tried to build a reputation system for Wikipedia so that we can generate reputation score for every author of an article, by doing that, we can decide to trust high reputation authors more because they have proved that they can do good editting.

1.2. Reputation System

As we said above, the system we want to build is a reputation system for authors. To achieve this goal, what we are doing is to generate a score for every edit. It is easy to check all edit history of a single article in Wikipedia. By fetching the edit history data, we actually get all revision from the very beginning of the article to some particular time point. In our experiment, we used a crawler to fetch such data so we can have latest version of an article. The data in stored in the manner that every edit is a full article. It is stored not the patches between two versions but full version for every edit.

For every edit, or for every revision, our model is to consider the edits around it. By saying some edits is "around" a edit, we are actually saying that all

revisions are sorted in time sequence, and some edits are closed to a edit either they are n-neighbors (in particular, we take n equals to 3 or 10) of this edit or they have short time gap with this edit (in our model, we take this time gap as 1hr).

1.3. Applications

Reputation system can play a significant role in building a healthy wiki community, because it promotes the quality of articles in many ways. If editors can view previous editting history and their authors' reputation, they can decide whether to keep the edit or not more easily. If we compare the Wikipedia community, which is a knowledge contributing community, to Github, which is a code contributing community, then those who have high reputation are like programmers who have many stars and followers. High reputation will make an author looks more reliable and help other authors make decisions.

Besides, we can visulize the reputation distribution of an article. through that way, it will be obvious that what kinds of article will attract more high reputation authors and based on that, it will be intriguing to discuss the relationship between articles' topic and their authors. We tried to do that a little bit, even not fully discovered. And another way to use reputation is to reward those high reputation authors to promote their passion. Also reputation score itself is some kind of honor on the Internet.

2. Related Work

The work most related to ours is [1], where Wikipedia revisions are used to evaluate authors' reputation. Our work is mostly based on theirs and then

do some change or improvement. Also, to discuss the relationship between reputation and article quality is inspired by Aniket Kittur's work done in 2008 [5], in which they use Amazon's crowdsourcing service called "Amazon Mechanical Turk" to evalute the quality of a Wikipedia article. They found that people tend to give higher score to those articles that have more high reputation authors. This leads us to use quality of an article as a evalutation metrics. To evaluate article's quality, we found that Joshua E. Blumenstock's work [6] is a quite intuitive way. Actually, at the early stage of our thinking, we have considered using word length as an important factor to evaluate edit quality. Zeng's work [7] is one of the most successful work done in 2000s, in which they used dynamic Bayesian Network to evaluate quality or trust for an edit. We took a quick look on it but the Bayesian Network method is not we are looking for, we hope to find a way that use edit history more directly to reflect how good a revision is. In many edit history based systems, like adlar's work [1] and Wohner's work [8], editting distance is used to evaluate difference between two versions. And we got the idea of coloring text as visulization from adler's work [9], in which they assigned text color to compare difference between two revisions. We use coloring in a different way, we assign words colors according to their authors' reputation to see an article's authors' distribution, which will be explained later.

3. Reputation Model

3.1. Notation

We use the similar notation in Adler [1]. The following is the notations we are using:

For some particular article, let's say the Wikipedia entry "Reverse Engineering", there are many versions of it. For all the versions, we sorted them in a time sequence, and thus there are versions from very beginning to latest. We call these versions v_0, v_1, v_2 and so on. And we assume v_0 is empty, which means we divide the creating of an article into two versions, one is v_0, the empty article, one is v_1, the first version with content. In most cases, the creator of an article will give some content to it, so we are actually adding an empty version before the first version. And thus we introduce the concept revision:

we use r_i means the i_{th} edit of an article, which is used to refer to the change from $v_{i-1}tov_i$.

$txt(i, j)$, in which $0 < i \leq j \leq n$. This is the d edit distance,

3.2. Adler's Model

In this part we will simply introduce Adler's model of content-driven reputation system in a versioned document and our implementation.

Basically, there are two rules in Adler's model, the first one is based on what they called "Text Survial", When we are implementing this model, we are using Google's library called "diff_match_patch" to help us do the calculation, which is based on the algorithm described in Myers' work [10].

[1] B. T. Adler and L. De Alfaro, "A content-driven reputation system for the wikipedia," in *Proceedings of the 16th international conference on World Wide Web*, pp. 261–270, ACM, 2007.

[2] Wikipedia, "Wikipedia — Wikipedia, the free encyclopedia," 2016. [Online; accessed 16-Dec-2016].

[3] Wikipedia, "Vandalism — Wikipedia, the free encyclopedia," 2016. [Online; accessed 16-Dec-2016].

[4] Wikipedia, "Edit warring — Wikipedia, the free encyclopedia," 2016. [Online; accessed 16-Dec-2016].

[5] A. Kittur, B. Suh, and E. H. Chi, "Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 477–480, ACM, 2008.

[6] J. E. Blumenstock, "Size matters: word count as a measure of quality on wikipedia," in *Proceedings of the 17th international conference on World Wide Web*, pp. 1095–1096, ACM, 2008.

[7] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness, "Computing trust from revision history," tech. rep., DTIC Document, 2006.

[8] T. Wöhner and R. Peters, "Assessing the quality of wikipedia articles with lifecycle based metrics," in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, p. 16, ACM, 2009.

[9] B. T. Adler, K. Chatterjee, L. De Alfaro, M. Faella, I. Pye, and V. Raman, "Assigning trust to wikipedia content," in *Proceedings of the 4th International Symposium on Wikis*, p. 26, ACM, 2008.

[10] E. W. Myers, "Ano (nd) difference algorithm and its variations," *Algorithmica*, vol. 1, no. 1-4, pp. 251–266, 1986.