Towards a Weighted Voting System for Q&A Sites

Daniele Romano and Martin Pinzger

Report TUD-SERG-2013-013
Towards a Weighted Voting System for Q&A Sites

Daniele Romano
Software Engineering Research Group
Delft University of Technology
Delft, The Netherlands
Email: daniele.romano@tudelft.nl

Martin Pinzger
Software Engineering Research Group
University of Klagenfurt
Klagenfurt, Austria
Email: martin.pinzger@aau.at

Abstract—Q&A sites have become popular to share and look for valuable knowledge. Users can easily and quickly access high quality answers to common questions. The main mechanism to label good answers is to count the votes per answer. This mechanism, however, does not consider whether other answers were present at the time when a vote is given. Consequently, good answers that were given later are likely to receive less votes than they would have received if given earlier.

In this paper we present a Weighted Votes (WV) metric that gives different weights to the votes depending on how many answers were present when the vote is performed. The idea behind WV is to emphasize the answer that receives most of the votes when most of the answers were already posted.

Mining the Stack Overflow data dump we show that the WV metric is able to highlight between 4.07% and 10.82% answers that differ from the most voted ones. The results show that WV ranks between 4.07% and 10.82% of the answers higher than the traditional approach. Moreover, we analyzed the extracted data to give an insight into the amount of answers already posted when votes are performed.

The remainder of this paper is organized as follows. In Section II we introduce the Weighted Votes metric, we reason about its integration into Q&A sites and discuss the benefits for their communities. Section III presents our study, its results and the process to extract the necessary data. We conclude this paper and draw directions for future work in Section V.

II. The Weighted Votes Metric

When a user is looking for the valuable answer to a question of interest she may focus on the most voted answers, especially if the question gets numerous answers. However, the current voting system adopted by Q&A sites is limited to count the number of votes an answer receives along its lifetime. The main limitation of such a system is that most of the votes can be performed immediately after the answer is posted. Hence, they do not take into account the answers posted later.

We propose a new way to count votes that takes into account the number of answers to a question already posted when a vote is performed and the total number of answers. We suggest to give different weights to the votes depending on the number of answers already posted when it is given. For an answer $A$ to a question $Q$ we define the Weighted Votes metric ($WV(A)$) as follows:

$$WV(A) = \sum_{k=1}^{n} \frac{\text{Answers}_{Q < t_k}}{\text{Answers}_{Q}}$$

where $n$ is the number of votes given for the answer $A$; $\text{Answers}_{Q}$ is the total number of answers to $Q$; $t_k$ indicates the time when the vote $k$ was performed and $\text{Answers}_{Q < t_k}$ indicates the number of answers given to $A$ and posted before the vote $k$ was performed.
In this paper we answer the following research question:

To what extent does the WV metric highlight answers different from the answers with the highest number of votes?

In the following subsections, first we describe the process to extract the data necessary for our analysis. Then we report our results and observations about the extracted data.

A. Data Extraction

Figure 2 shows the approach we used to extract the data from the Stack Overflow data dump.

2http://data.stackexchange.com
TABLE I: Percentage of answers highlighted by WV that differ from the most voted ones for different categories of questions.

<table>
<thead>
<tr>
<th>Questions (%)</th>
<th>Answers &gt;= 2</th>
<th>Answers = 2</th>
<th>Answers = 3</th>
<th>Answers &gt;= 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>WV&lt;sub&gt;high&lt;/sub&gt;</td>
<td>10.31%</td>
<td>5.88%</td>
<td>10.75%</td>
<td>17.17%</td>
</tr>
<tr>
<td>WV&lt;sub&gt;low&lt;/sub&gt;</td>
<td>3.21%</td>
<td>2.26%</td>
<td>3.52%</td>
<td>4.48%</td>
</tr>
<tr>
<td>Average</td>
<td>6.76%</td>
<td>4.07%</td>
<td>7.13%</td>
<td>10.82%</td>
</tr>
</tbody>
</table>

In the first step we downloaded the data dump in XML format from the Stack Exchange website. The data dump consists of five XML files that store information about the users (users.xml), the posts (posts.xml), the comments (comments.xml), the posts’ history (posthistory.xml) and the badges badges.xml.

In the second step, for each answer contained by posts.xml we extract the up and down votes from the votes.xml file. We discarded the votes for answers that have been removed from the database. The output of this step consists of the votes.csv file that for each vote contains 1) the id of the answer for which the vote has been given, 2) the id of the question of the answer and 3) the creation date of the vote. In total we extracted 13,700,939 votes of 4,392,956 answers given to 2,421,549 questions.

In the third step, we prepared the data to compute the values for the WV metric. To be able to measure the WV we needed for each vote k the count of all answers posted before the vote was given (Answers<sub>&lt;k</sub>), total number of answers (Answers) given to a question Q. However, differently from the creation date of answers, the creation date of a vote does not contain the information about hours, minutes and seconds.

Its format is in the form month-day-year. As consequence we cannot know if the answers posted on the day when the vote k is given are actually performed before or after the vote. This format is used in all Stack Exchange data dumps and not only for Stack Overflow. For this reason we computed 1) the number of answers posted on the days that precede the day when a vote is given (AnswersBefore); 2) the answers posted on the same day (AnswersSameDay); and 3) the answers posted on the following days (AnswersAfter). These values allow us to estimate the actual value of the WV metric as explained in the next step. The output of this step consists of the votes<sub>1</sub>.csv that enriches the vote.csv file adding for each vote the values of (AnswersBefore), (AnswersSameDay) and (AnswersAfter).

Since we cannot order the AnswersSameDay, in the fourth step we computed the values of two variants of WV. We computed the values of WV<sub>low</sub> and of WV<sub>high</sub>. Computing WV<sub>low</sub> we assume that the AnswersSameDay have been posted after the vote was given. On the other hand, computing WV<sub>high</sub> we assume that the AnswersSameDay were posted before the vote has been given. In this way WV<sub>low</sub> and WV<sub>high</sub> are the lower and upper boundaries of the actual value of WV. The values for WV<sub>high</sub>, WV<sub>low</sub> and the number of votes for each answer are saved in answers.csv.

In the last step (Step 5 in Figure 2), for each question we compared the ranking of the answers obtained with the WV metric and the traditional approach and computed the ratios of answers for which the ranking differed.

B. Results

Table I shows the results obtained. Among all the questions analyzed we report the results of questions with a number of answers greater than two (Answers>= 2). They account for 63.96% of all questions. For the questions with only one answer the value for WV is equal to the number of votes. Moreover, we report the results for questions with two answers (Answers=2), questions with three answers (Answers=3) and questions with four or more answers (Answers>=4). We chose these values because they represent the median number of answers (i.e., three) and the 75th percentile (i.e., four).

From the results we can state that for the questions with more than two answers (Answers>=2) the WV metric emphasizes on average 6.76% different answers. In such cases the user can focus on answers that received most of the votes when most of the answers were already posted. For questions with two, three and four or more answers we registered on average respectively 4.07%, 7.13% and 10.82% of different answers highlighted by the WV metric.

In conclusion, we can answer our research question stating that the percentage of different answers highlighted by WV is 1) between 3.21% and 10.31% for questions with two or more answers, 2) between 2.26% and 5.58% for questions with two answers, 3) between 3.52% and 10.75% for questions with three answers and 4) between 4.48% and 17.17% for questions with four or more answers. On average the WV metric highlights a percentage of different answers that ranges from 4.07% to 10.82%.

C. Observations

Besides the WV’s ability of highlighting different answers we can make two important observations reading the results shown in Table I.

First, we can notice that the percentage of different answers highlighted with WV increases when we consider questions with a higher number of answers. For WV<sub>high</sub> we registered an increment of ≈ 292% (17.17/5.88) between questions with two answers and questions with four or more answers. For WV<sub>low</sub> we registered an increment of ≈ 198% (4.48/2.26) between questions with two answers and questions with four or more answers.

Second, we can notice the difference between the values measured for WV<sub>high</sub> and WV<sub>low</sub>. In order to understand...
TABLE II: Paired Cliff’s delta effect sizes \((d)\) between AnswersBefore, AnswersSameDay and AnswersAfter. The effect size is considered negligible for \(d < 0.147\), small for \(0.147 \leq d < 0.33\), medium for \(0.33 \leq d < 0.47\) and large for \(d \geq 0.47\) [5].

<table>
<thead>
<tr>
<th>Distribution1</th>
<th>Distribution2</th>
<th>Cliff’s (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswersBefore</td>
<td>AnswersSameDay</td>
<td>0.053</td>
</tr>
<tr>
<td>AnswersBefore</td>
<td>AnswersAfter</td>
<td>0.318</td>
</tr>
<tr>
<td>AnswersSameDay</td>
<td>AnswersAfter</td>
<td>0.232</td>
</tr>
</tbody>
</table>

this gap we analyzed the difference of the distributions of AnswersBefore, AnswersSameDay and AnswersAfter measured for each vote. We computed the Mann-Whitney p-value for paired samples for each pairs of distributions to test if the distributions were different. For all pairs we registered p-values smaller than 0.01 indicating that the distributions are considered statistically different. Moreover we computed the Cliff’s delta effect size (for paired samples) [5] to measure the magnitude of the difference and we report the results in Table II.

The results show that the difference in magnitude between the distribution of answers posted on days before the day when a vote is given (AnswersBefore) and the distribution of answers posted on the same day of a vote (AnswersSameDay) is negligible \((d=0.053<0.147)\)[5]. The distribution of answers posted after a vote (AnswersAfter) is smaller than the distributions of AnswersBefore and AnswersSameDay because the effect sizes’ values \((d=0.318\) and \(d=0.232)\) are considered to be medium [5]. From these results we can state that the distributions of AnswersBefore and AnswersSameDay are the biggest ones. This explains the difference between the values for \(W_{V_{low}}\) and \(W_{V_{high}}\) registered in our study.

IV. RELATED WORK

In the last years many studies on Stack Overflow have been presented. The closest to our study has been developed by Schall et al. [6]. They analyzed the dynamics of the community activities. As part of this analysis they analyzed the answering behavior per question showing the number of answers per different categories of questions. However, they have not analyzed the voting behavior and the main focus of their work is the mining of expertise.

Among all scientific work about mining Q&A sites, mining expertise from Q&A communities is becoming more and more popular. Many of them propose technique to mine expertise of users in the community, such as [7] [8] [9] [10] [11]. These works propose techniques and approaches to infer the expertise from several variables. Among these variable the number of votes plays a crucial role. The WV metric proposed in this paper can help to improve these approaches. For example, it can be used to filter votes given when only one answer is posted.

V. CONCLUSION AND FUTURE WORK

In this paper we proposed a Weighted Votes metric aimed at highlighting answers that received most of the votes when most of the other answers were already given to a question. Mining the Stack Overflow data dump, we showed that the proposed metric is able to emphasize answers different from the most voted ones. This is particularly useful for users who are looking for high quality answers.

In our future work we plan to further validate and improve this metric. First, we plan to look for data in which the complete timestamp of a vote is registered. This allows us to obtain more precise results avoiding the approximations performed in this study (i.e., the computation of \(W_{V_{high}}\) and \(W_{V_{low}}\) to estimate the actual value of \(W_V\)).

Second, we plan to perform a qualitative study to test to which extent the number of votes is relevant to users looking for answers. It is particularly useful to investigate if users go through all the answers or if they read only the most voted ones or the accepted ones.

Finally, we plan to perform a qualitative analysis with questionnaire to find out whether the answers highlighted by the Weighted Votes metric are considered of better quality compared to the most voted ones.

ACKNOWLEDGMENT

This work has been partially funded by the NWO-Jacquard program within the ReSOS project.

REFERENCES


