Fixed Point Method for Voting

Chung Tong Lee[1]
Aleksandar Ignjatvoić[2]

[1] School of Computer Science and Engineering
University of New South Wales, Australia
ctlee@cse.unsw.edu.au
[2] School of Computer Science and Engineering
University of New South Wales, and NICTA, Australia
ignjat@cse.unsw.edu.au

Technical Report
UNSW-CSE-TR-0908
April 2009

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

Question answering (Q&A) community sites, such as the MSN QnA and Yahoo! Answers, facilitate question answering by a community of users. However, the quality of the answers provided by users varies. To determine the best answer, vote counts, sometimes with extra weight put on the askers, are commonly used. This makes the result vulnerable to tainted vote effect as opinions from "bad" voters weight the same as those from the "good" ones. We propose a new method to determine the best answer by the sum of voters' reliability scores, which are calculated based on voters' behaviors. The more a voter can choose the best answer, the more reliable he is and the more weight his opinion should carry. This is a circular definition similar to the reputation score evaluation in [2]. Our method does not require the identification of anomalous voting behavior to reduce the reliability score. Instead, we employ the Brouwer Fixed Point Theorem [1] to show the existence of the assignment for reliability scores which satisfy the axiomatic description of the system. Iterative method is used for actual evaluation. To demonstrate the robustness, simulations are designed with data that match the real-life situation, augmented with various forms of anomalous behaviors.

1 Introduction

A sequence of questions $Q[1],\ldots,Q[N_Q]$ are posed, and for each question $Q[j]$ a set of answers $A[j,k]$ is given where $0 \leq k \leq N_A[j]$ and $N_A[j]$ is the total number of answers for $Q[j]$. For each such set agents in the community vote to determine which is the best answer for the corresponding question. There are in total N_v distinct voters and the agent x_i voted for the answer $A[j,V[i,j]]$ for the question $Q[j]$.

Task: Given the sequence of votes determine optimally what the best answer is for each question, and determine the reliability of each voter in a robust way that automatically discount tainted vote effect.

1.1 Examples

Example 1: Assume that in a vote for $Q[j]$ answer $A[j,1]$ gets the vote of agents x_1, x_2 and x_3, and answer $A[j,2]$ gets the votes of agents x_4 and x_5. However, x_1, x_2 and x_3 have a poor voting record, seldom voting for the winning answers, while x_4 and x_5 have consistently voted for the winning answers. The system should proclaim $A[j,2]$ the *best* winner, despite having gotten only two votes while $A[j,1]$ has gotten three votes.

Example 2: Assume a community of voters that often vote is suddenly joined by a group of new voters who collude and vote for a particular answer. If the group of colluders is reasonably small compared to the group of regular voters, then the newcomers cannot change the outcome of elections and the system should identify the colluding voters.

2 Algorithm

We first discuss what the features of the bootstrapping case should be, i.e., the case of the very first question $Q[1]$. We assign reliability score $r[i]$ to x_i based on the answer he voted. Assume that an answer $A[1,k]$ gets $\mu[1,k]$ votes, we argue that it is fair to have

$$r[i_1] : r[i_2] = \mu[1,V[i_1,1]] : \mu[1,V[i_2,1]]. \tag{2.1}$$

By symmetry, $r[i_1] = r[i_2]$ if $V[i_1,1] = V[i_2,1]$, i.e., x_{i_1} and x_{i_2} voted for the same answer. Assume that everyone votes, we can show that the following function satisfies the property (2.1):

$$r[i] = \sqrt{\frac{\sum\{r[\iota] : V[\iota,1] = V[i,1]\}}{\sum_{\mathrm{i}\leq N_v} r[\mathrm{i}]}}. \tag{2.2}$$

with $0 < r[i] \leq 1$ for $1 \leq i \leq N_v$. Details are as following:

$$\begin{aligned} r[i_1] &= \sqrt{\frac{\mu[1,V[i_1,1]]\, r[i_1]}{\sum_{\mathfrak{i} \leq N_v} r[\mathfrak{i}]}} \\ r[i_1]^2 &= \frac{\mu[1,V[i_1,1]]\, r[i_1]}{\sum_{\mathfrak{i} \leq N_v} r[\mathfrak{i}]} \\ r[i_1] &= \frac{\mu[1,V[i_1,1]]}{\sum_{\mathfrak{i} \leq N_v} r[\mathfrak{i}]} \end{aligned}$$

Similarly we also have $r[i_2] = \mu[1,V[i_2,1]] / \sum_{\mathfrak{i} \leq N_v} r[\mathfrak{i}]$ and we get (2.1) as desired.

We take the arithmetic means for N_Q questions and the fixed-point of the formula system below is the solution of reliability score for each agent

$$\left\{ r[i] = \frac{1}{N_Q} \sum_{j=1}^{N_Q} \sqrt{\frac{\sum\{r[\imath] : V[\imath, j] = V[i, j]\}}{\sum_{\mathfrak{i} \leq N_v} r[\mathfrak{i}]}} \right\}_{1 \leq i \leq N_v}. \tag{2.3}$$

3 Generalization

In the formula of the previous section, we assumed everyone was voting in every question. This can be easily generalized to situations which is not the case. If the voter x_i didn't cast his vote for question $Q[j]$, $V[i,j]$ is not defined. Neither is the corresponding top sum. For this situation, we assign zero as the value of the undefined top sum. With this, we limit the effect of swarming voters as described in example 2. This also encourages participation as casting a vote has a positive effect.

In addition, the relative reliability score need not be linear as depicted by equation (2.1). Using a real parameter p, and a modified version of equation (2.2):

$$r[i] = \sqrt[p]{\frac{\sum\{r[\imath] : V[\imath,1] = V[i,1]\}}{\sum_{i_1 \leq N_v} r[i_1]}}, \tag{3.1}$$

we have

$$\begin{aligned} r[i] &= \sqrt[p]{\frac{\mu[1,V[i,1]]\, r[i]}{\sum_{k \leq N_v} r[k]}} \\ (r[i])^{p-1} &= \frac{\mu[1,V[i,1]]}{\sum_{k \leq N_v} r[k]} \\ (r[i_1])^{p-1} : (r[i_2])^{p-1} &= \mu[1,V[i_1,1]] : \mu[1,V[i_2,1]] \end{aligned}$$

where $p > 2$ makes the relative reliability sub-linear (deemphasize); $1 < p < 2$ makes it super-linear (emphasis).

On the other hand, we can augment the system (2.3) with time factor. Similar to trust score evaluation in [2], components of the final reliability score can be discounted by a real parameter $q \geq 1$. The system of equations will change as follow:

$$\left\{ r[i] = \frac{\sum_{j=1}^{N_Q} q^{t(j)} \sqrt[p]{\frac{\sum\{r[\imath]:V[\imath,j]=V[i,j]\}}{\sum_{\mathrm{i}\leq N_v} r[\mathrm{i}]}}}{\sum_{\mathrm{j}=1}^{N_Q} q^{t(\mathrm{j})}} \right\}_{1\leq i\leq N_v} \tag{3.2}$$

where $t(j)$ gives the closing time of the voting process for question $Q[j]$.

4 Existence of Fixed Point and Iteration Setting

Consider the mapping F derived from equation (3.2).

$$F : (r[i] : i \leq N_v) \mapsto \left(\frac{\sum_{j=1}^{N_Q} q^{t(j)} \sqrt[p]{\frac{\sum\{r[\imath]:V[\imath,j]=V[i,j]\}}{\sum_{\mathrm{i}\leq N_v} r[\mathrm{i}]}}}{\sum_{\mathrm{j}=1}^{N_Q} q^{t(\mathrm{j})}} : i \leq N_v \right)$$

The largest value of the surd expression is 1 when all agents vote for the same answer. The smallest value is 0 if agent x_i didn't vote for question $Q[j]$. However, agent x_i must have voted for at least one question in order to have $r[i]$ used for fixed-point calculation. Hence, $0 < r[i] \leq 1$. In order to apply Brouwer fixed point theorem, F must been continuous and $r[i]$ cannot be arbitrarily close to zero. The smallest value for $r[i]$ is when x_i vote for an answer of the first question with no one agreed with him. Then values of all surd expressions are 0 except that for first question.

$$\begin{aligned} r[i] &= \frac{1}{\sum_{j=1}^{N_Q} q^{t(j)}} \sqrt[p]{\frac{r[i]}{\sum_{\mathrm{i}\leq N_v} r[\mathrm{i}]}} \\ &> \frac{1}{\sum_{j=1}^{N_Q} q^{t(j)}} \sqrt[p]{\frac{r[i]}{N_v}} \\ \text{Hence: } r[i] &> \sqrt[p-1]{\frac{1}{\left(\sum_{j=1}^{N_Q} q^{t(j)}\right) N_v}} \end{aligned}$$

Take $\varepsilon = \sqrt[p-1]{\frac{1}{\left(\sum_{j=1}^{N_Q} q^{t(j)}\right) N_v}}$, we have shown that $F : [\varepsilon, 1]^{N_v} \mapsto [\varepsilon, 1]^{N_v}$. Hence fixed point exists. To find the fixed point of the system, we apply F iteratively, starting with all $(r[i] = 1 : i \leq N_v)$.

5 Data Generation for Simulation

To demonstrate the properties of the reliability score method, we conducted simulations with data generated to match characteristics of the real-life data The basic parameters are the numbers of questions and voters involved in the "best answer" evaluation. We notice that the numbers of answers for different questions with at least one vote, $\#A[j]$ for $Q[j]$ follow a geometric distribution $Pr(\#A[j] = x) = (1-p)^{(x-1)}p$ where $p = 0.3$.

The numbers of votes, denoted by $\#V[j]$, for a question $Q[j]$ must be equal or larger than $\#A[j]$. They are under a negative binomial distribution: $Pr(\#V - \#A = x) = \begin{pmatrix} \#V + \#A - 1 \\ \#A - 1 \end{pmatrix} p^{\#A}(1-p)^{x+\#A}$ and $p = 0.3$ fits the real data distribution.

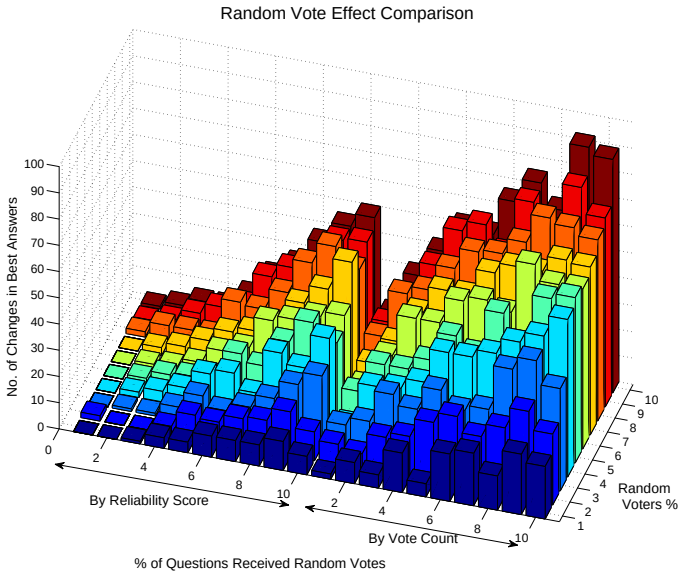
The larger the value of $\#A$, the higher the entropy of the vote distribution. The probability of getting a vote for the i^{th}-most popular answer can be modeled by Zipf"s law: $Pr(i) = \frac{i^{-s}}{\sum_{\iota=1}^{N_v} \iota^{-s}}$ with $s = 1.5$.

About 90% of the total votes are casted only by 10% of voters. Such power law characteristic is effectively represented by Zipf-mandelbrot law's in our discrete situation. The probability for the i^{th} most active voter to cast a vote is given by $Pr(i) = \frac{(i+q)^{-s}}{\sum_{\iota=1}^{N_v}(\iota+q)^{-s}}$ with $q = 13, s = 1.8$.

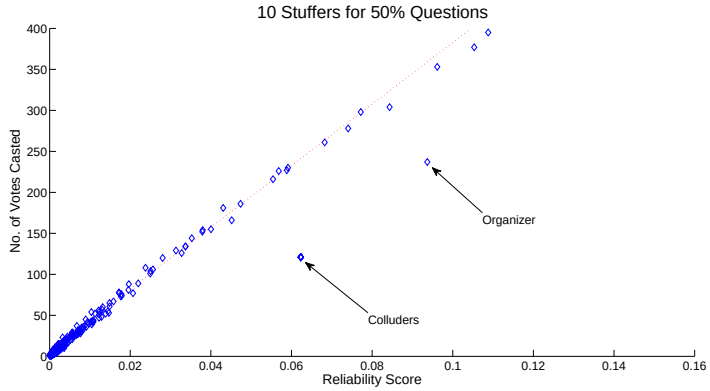
All of the above statistical properties describe the vote distribution as the base of normal voting behaviors.

6 Results

In addition to the normal voting behaviors, we generate extra data to represent two different forms of behaviors that affect the information sharing and community building. The first one is random voting. We simulate this type of behavior by adding uniformly distributed answer choice instead of the original vote-answer-distribution of Zipf's Law. The simulation results showed that best answers by voter's reliability changed less than that by vote count for all levels of random voting up to 50% of total number of votes. It is because random voter get less reliability scores as they cannot consistently choose the best answer. In this sense, the effect of random voting is reduced and reliability score method is more robust.



Another type of behavior we modeled is called ballot stuffing, a form of coordinated voting. An subversion organizer asked his friends to vote for the same answer he has already chosen so that answer would get more votes and hence higher changes of being the best. We simulate this behavior by adding different numbers of colluding voters for various percentages of questions that the organizer has voted. Though the organizer managed to increase the best answer hit by colluding votes, the exceptional ability of picking the correct answer make all colluders stand out from the distribution of reliability scores. Hence our method helps to identify anomalies and gives hints for possible collusion.



7 Conclusion

The ability of choosing the best answer is represented by the reliability score which links the vote distribution to voters' ability. Evaluation of these scores involves

an iterative process, a computation realizing Brouwer fixed point theorem. The corresponding system to determine the best answer semantics is found to be more robust in simulations against random voting. Moreover, reliability score provides a different metrics for user behaviors and it helps to categories users with real community data. It also hints possible ballot stuffing situation. Preliminary results are promising. We intend to explore possible synergy of this and other statistical methods, e.g. voter/answerer correlation analysis, in detecting coordinated voting behavior.

Bibliography

[1] D. H. Griffel. *Applied Functional Analysis*. Dover Publications, Jun 2002.

[2] A. Ignjatovic, N. Foo, and C. T. Lee. An analytic approach to reputation ranking of participants in online transactions. In *Proceedings, IEEE/WIC/ACM International Conference on Web Inteelligence*, pages 587–590, 2008.