Hrbust in TREC 2013: Crowdsourcing Track

Li Peng[1,2], Sun Bo-yu[2],Liu Yang[2], Zhang Ting-ting[2]
[1] Higher Educational Key Laboratory for Measuring and Control Technology, Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, Harbin, China
[2] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
pli @hrbust.edu.cn.

Abstract

In the practical application of crowdsourcing, some unreliable workers have emerged due to profit driven. Their results seriously reduce the quality and bring about the initiator's judgment biases. In this paper, we creatively put forward a crowdsourcing fraud detection method based on psychological behavior analysis to find out the spammer according to the psychological difference between deception and reliable behavior by means of Ebbinghaus forgetting curve. Furthermore, we constructed an online crowdsourcing experiment platform to verify the validity of our method. In addition, we participated in TREC 2013 Crowdsourcing Track and the organizer provided the evaluation results for our run. As a result, APCorr, RMSE and GAP attained 0.480, 0.135 and 0.392 respectively. Evaluation and xperimental results show that our method is effective and feasible.

Keywords: *Crowdsourcing, Ebbinghaus Forgetting Curve; Fraud Detection; Multi-Communication*

1. Introduction

Crowdsourcing, a new organization form and cooperation pattern in the process of enterprise production, has grown with the rapid popularity of Internet [1]. Enterprises actively utilize many online user resources and allocate the outsourcing task to the interest groups by means of crowdsourcing technology, solving some limitations of traditional outsourcing services. In recent years, crowdsourcing technology has become a focus of research and researchers around the world have realized it in some practical applications. For example, Rensnik combined monolingual crowdsourcing and targeted paraphrasing to improve the quality of pure machine translation [2]; Hwang introduced mobile-based crowdsourcing into the field of environmental audio recognition, to improve the performance of mobile devices in filtering background noise [3]; Fritz applied crowdsourcing to global land cover, to solve the problem of ignoring potential land in statistical work [4]. Moreover, crowdsourcing technology is also applied to other fields such as software testing, content screening, and labeling training data for machine learning [5]. The rapid development of crowdsourcing technology drew the Text REtrieval Conference's (TREC) attention and crowdsourcing track has attracted many groups around the world to join the competition in TREC 2013.

Unfortunately, some unreliable workers have emerged due to profit driven in the practical application of crowdsourcing. Their results seriously reduce the quality and bring about the initiator's judgment biases [6]. In recent years, many researchers conducted in-depth exploration and have achieved some progress on improving the quality of crowdsourcing results. For instance, Matthew from the UT Austin provides a method for detecting cheats and it think that worker should pushed confidence score button according to the confidence degree of answers submitted [7]. Bell labs researchers noticed that a large number of user information can be obtained from mobile devices and this information contributes to finding better workers [8]. Jeroen et al. use the average squared ordinal distance between workers' judgments to calculate each worker's random score for detecting random spammers [9]. At present, crowdsourcing technology is still in the early stages and it is important for us to improve the quality of crowdsourcing results by some means.

Detecting spammers is important for enhancing the quality of crowdsourcing results and workers' psychological state will change when they are cheating. The essence of deception is a manifestation of people's psychological activity and psychological methods can make effective judgment on it in a certain extent. Therefore, we think that using the psychological method may be a realistic way to solve this problem and put forward a kind of crowdsourcing fraud detection method based on psychological behavior analysis. The method applying Ebbinghaus forgetting curve to determine the behavior of crowdsourcing worker is normal or not and find out the difference between deception and reliable behavior.

2. Behavior analysis based on psychological behavior analysis

Ebbinghaus forgetting curve is one of the classical methods in psychology field, it reveals the regularity of forgetting information memorized. In recent years, some researchers have introduce this method into the field of computer science, such as Zeng and Lin designed an interactive vocabulary learning system based on word frequency lists and Ebbinghaus' curve of forgetting [10]; Luo and Yuan utilize Ebbinghaus forgetting curve to mine Internet users' term interest and provide personalized search results [11]. In this article, Ebbinghaus forgetting curve is applied in crowdsourcing fraud detection. The principle is the differences between deceptional worker and trusted worker in forgetting curve when they do crowdsourcing tasks.

2.1. Ebbinghaus forgetting curve

Memory can be divided into short-term and long-term according to the length of maintaining information periods. The memory process is shown as Figure 1. Received information will become short-term memory after attention. The memory will be forgotten without timely review. Otherwise, long-term memory will form.

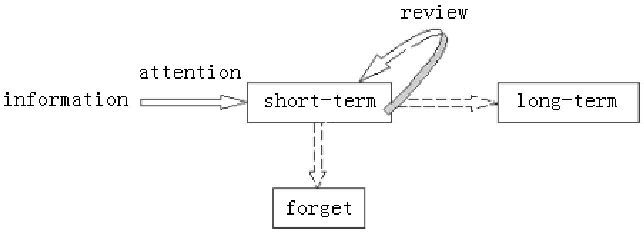
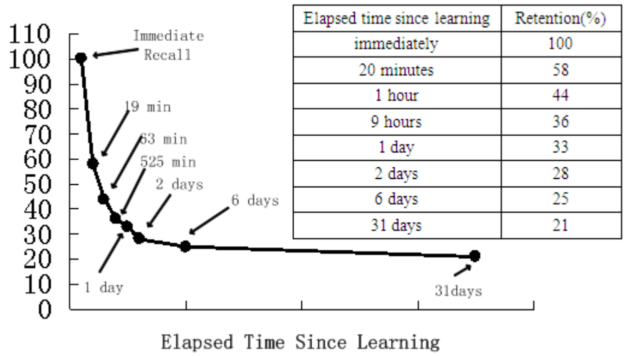


Figure 1. Human memory process

The German psychologist H.Ebbinghaus researched the basic rule of human memory and oblivion, and put forth "the function of time and memory" as shown in the formula (1). *OriginalLearning* stands for the number of writing from memory, when he remembered all materials in the first time. After a while, Re*learning* is the number. Thus retention scores are obtained and represented by *SavingScore.*

$$SavingScore = \frac{OriginalLearning - \mathrm{Re}\,learning}{OriginalLearning} \tag{1}$$

According to the formula (1), The Ebbinghaus forgetting curve can be drawn (see Figure 2). In this figure, the longitudinal axis represents the memory retention scores and the horizontal axis represents elapsed time since learning. The cure conducts a quantitative expression for the forgetting rules in learning process resulting oblivion can be calculated. Human oblivion is an unbalanced development, Memory is forgotten very quickly in the initial stage, and then slows down gradually, after a certain time almost no longer forgotten.



Elapsed time since learning	Retention(%)
immediately	100
20 minutes	58
1 hour	44
9 hours	36
1 day	33
2 days	28
6 days	25
31 days	21

Figure 2. Ebbinghaus forgetting curve

2.2. Our fraud detection algorithm

Fraud is usually divided into two basic types. One is random spammer, they submit results randomly and it is difficult to identify them; the other is uniform spammer, they just submit results regularly and this type is easy to be found. Therefore, we only focus on the former.

The principle of our detection method is crowdsourcing participants will produce different memory rules owing to different psychological state in the process of judgment to the crowdsourcing task. Their behavior is a specific embodiment of mental activity, no matter he is a spammer or credible person. The trusted participant will strictly comply with the requirements and think hard when judging the relevance of pairs, so it will produce a deep memory in their mind. This is the general process of human memory in accordance with Ebbinghaus forgetting rule. However, the spammer will only spend little or no effort on crowdsourcing and complete tasks mechanically. They lack the understanding memory about task content and their forgetting states do not comply with Ebbinghaus forgetting rule. According to above psychological differences and quantitative expression of Ebbinghaus forgetting curve, the flow chat of our algorithm is shown in Figure 3. Workers will rejudge a certain amount of repeated pairs in limited time, and then the work's SavingScore and threshold is compared, low SavingScore workers are considered as the spammer.

There are four sets A, O, W and T in the flow chat, all query-document pairs are waiting for labeling in A. Putting the pairs which are judged for one time into O. The pairs are waiting for labeling again in W. Putting the pairs which are judged for two times into T. In crowdsourcing task, platform randomly selects topic-doc pairs and recommends some pairs that have the less number of judgments for all workers. Firstly, query-document pair would be judged and kept into O, then deleted from A. Secondly, if some pair stay in O about 20 min, it will be put into W and removed from O. At this time, W is not empty and platform may offer worker a pair from W. It means that the pair has been judged. Workers should rejudge the pair within the given limit time so that they have to judge it depending on memory barely rather than reflection. At last, the pair which is judged twice will be placed into T and removed from W.

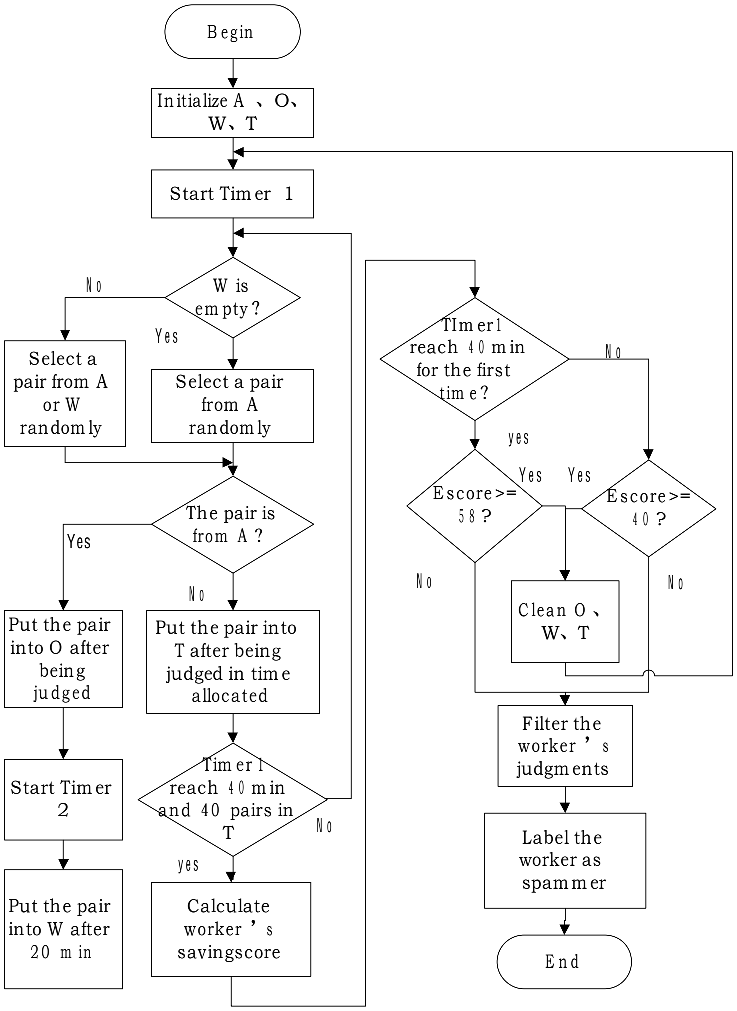

Begin
Initialize A 、O、W、T
Start Timer 1
W is empty?
No
Yes
Select a pair from A or W randomly
Select a pair from A randomly
The pair is from A ?
Yes
No
Put the pair into O after being judged
Put the pair into T after being judged in time allocated
Start Timer 2
Timer1 reach 40 min and 40 pairs in T
No
yes
Put the pair into W after 20 min
Calculate worker ' s savingscore
TImer1 reach 40 min for the first time?
No
yes
Escore>= 58 ?
Yes
Yes
Escore>= 40 ?
No
No
Clean O 、W、T
Filter the worker ' s judgments
Label the worker as spammer
End

Figure 3. The flow chat of our algorithm

In our algorithm, if workers do repeated judgment for at least 40 query-document pair, the pairs of W reached 40 and the platform will calculate the workers' SavingScore. If the worker operates 40 minutes but haven't reached 80 minutes, and the workers' SavingScore ≥ 58, we determine the worker has no cheat in this period. If the workers have worked for more than 80 minutes, only his SavingScore ≥ 40 can we think that they didn't have cheat. Platform will adopt judgment results of workers who meet the requirements, then emptying O, W and T and conducting a new round of detecting fraud for the workers.

Finally, we calculate the ratio of the deviations sum of the topic-doc's final label value and its 5 label values, then the product of the maximum deviation of 6 relevance labels and the number of the judgment. At last, 1 minus the ratio gets the relevance probability of final label and judgments.

3. Crowdsourcing online platform

In this year, we participated in TREC 2013 Crowdsourcing Track. According to the requirements of track, we constructed an online crowdsourcing experiment platform to verify the validity of our method. We submitted our run and the organizer reported the evaluation results.

The 2013 Crowdsourcing Track requires collecting relevance judgments for Web pages and search topics, in partnership with TREC Web Track. There are 50 search topics that are from the Web track. Web pages to be judged for relevance were drawn from the ClueWeb12 collection. The organizer offered 2 different entry levels for participation.

(1) Basic:~2k documents(sub of NIST pool)

(2) Standard:~20k documents(entire NIST pool)

We chose the basic level, there are 4375 pairs should be judged,they belong to 10 topics which are 202, 214, 216, 221, 227, 230, 234, 243, 246, 250 provided officially by the NIST. For the TREC2013 Crowdsourcing Track, we employ a solution strategy based on multi-communication platform and multi-type crowds. To bring together a wide rang of participants to support and participate in crowdsourcing task, we adopt the various popular social networking platforms to spread widely, including website promotion, SNS social networking, microblog, WeChat and instant communication tools. We divide the crowd into three groups, Expert Group, Trustee Group and Volunteer Group by the degree of confidence, to judge probability of relevance between different topics and different webs on a six-point scale(4,3,2,1,0,-2). Expert group judged all 3470 topic-doc pairs from 10 topics, and asked their friends for help, receiving a number of judgments, that is treated as Trustee Group's results. We called others from the Internet as volunteers. In order to ensure the topic-doc distribution on average, we established a network platform of crowdsourcing. It randomly selects topic-doc pairs and recommend some pairs that have the less number of judgments for all workers. Besides, the platform will further process volunteers' judgments. Finally, we selected 5 judgments for each topic-doc pair, three are from volunteers'.

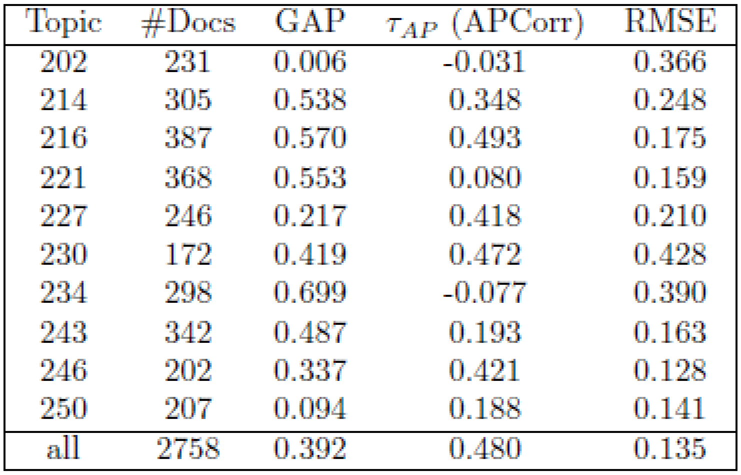
4. Evaluation measures

Using the judgments obtained from the trusted and highly trained NIST judges as gold standard, TREC measured the quality of the submitted judgments for the following three metrics:

- Rank Correlation: The Web Track participants' ad-hoc IR systems are scored based on NIST judgments according to the primaryWeb Track metric, ERR@20, inducing a ranking of IR systems. A similar ranking of IR systems is then induced from each Crowdsourcing Track participant's submitted judgments.Rank correlation is then calculated, indicating how accurately crowd judgments can be used to predict the NIST ranking of IR systems. The measure we use for rank correlation is Yilmaz et al.'s AP Correlation (APCorr) [12], which improves upon Kendall's Tau as a measure of rank correlation by emphasizing the order of the top ranked systems. To the best of our knowledge, the original version of APCorr does not handle ties; Organizers handle ties by sampling over possible orders.
- Score Accuracy: In addition to correctly ranking systems, it is important that the evaluation scores be as accurate as possible. Using root mean square error (RMSE) for this measure.
- Label Quality: Direct comparison of each participant's submitted judgments against the NIST judgments (no evaluation of Web track IR systems). Label quality provides the simplest evaluation metric and can be correlated with the other measures predicting performance of IR systems. Hence this year, we use graded average precision (GAP) [13]. The GAP is computed by ordering the documents as per the score assigned to the document and then using the qrels provided by NIST.

5. Evaluation results

Evaluation of crowd qrel Hrbust123(our run name) for the Basic task of the TREC 2013 Crowdsourcing Track. The 10 topics for the basic task were randomly selected from the TREC 2013 web track ad-hoc task.



Topic	#Docs	GAP	τ_{AP} (APCorr)	RMSE
202	231	0.006	-0.031	0.366
214	305	0.538	0.348	0.248
216	387	0.570	0.493	0.175
221	368	0.553	0.080	0.159
227	246	0.217	0.418	0.210
230	172	0.419	0.472	0.428
234	298	0.699	-0.077	0.390
243	342	0.487	0.193	0.163
246	202	0.337	0.421	0.128
250	207	0.094	0.188	0.141
all	2758	0.392	0.480	0.135

Table 1: This table shows per-topic statistics and overall averages for the run Hrbust123. The metrics GAP, ERR@20, AP-correlation and RMSE are listed for each topic. Note that for row all, (i) GAP is the mean gap over all 10 topics, (ii) APCorr and RMSE depend on the ranking of runs induced by the mean ERR@20 for all the 10 topics.

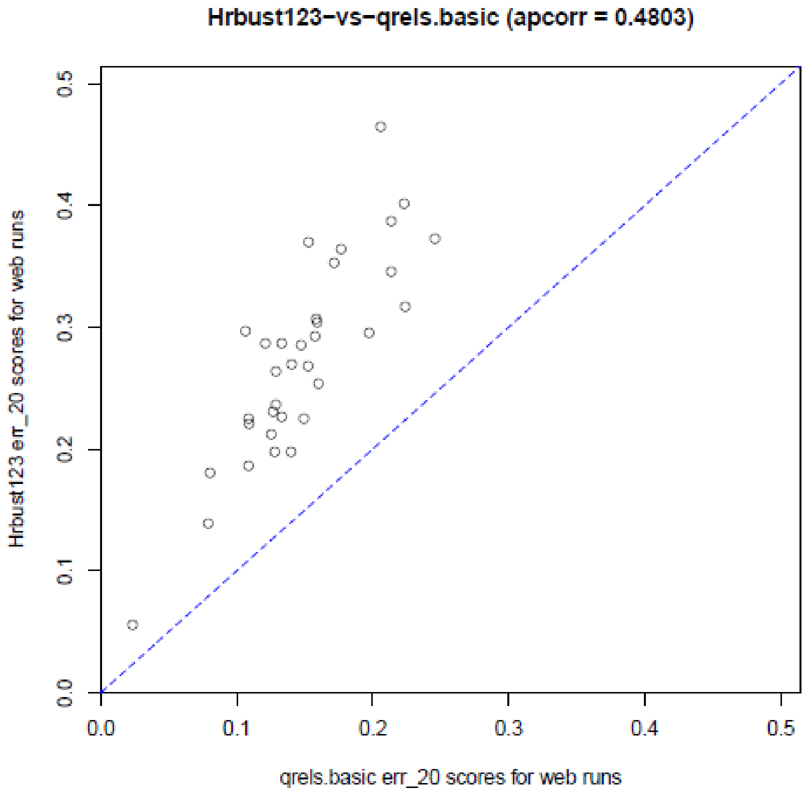


Figure 4: Hrbust123-basic-ERR@20 vs qrels.basic-ERR@20. qrels.basic is the TREC 2013 web track qrels reduced to topics 202, 214, 216, 221, 227, 230, 234, 243, 246, and 250.

6. Conclusion

This paper proposed an effective solving strategy on crowdsourcing fraud detection by means of psychological behavior analysis method. We creatively apply Ebbinghaus forgetting curve to find out the spammer according to the psychological difference between fraud and reliable behavior. This is an exciting exploration because we successfully applied the psychological method to the field of computer science. We also develop an online crowdsourcing experiment platform to verify the validity of our method. In addition, the TREC 2013 crowdsourcing track organizer provided the evaluation results for our run. As a result, APCorr, RMSE and GAP attained 0.480, 0.135 and 0.392 respectively in the submitted data set. Experimental results show that our method is contribute to improving the quality of crowdsourcing result and can be used to control crowdsourcing quality.

7. ACKNOWLEDGEMENTS

This paper is partially supported by National Natural Science Foundation of China (61103149), China Postdoctoral Science Foundation (2011M500682), Postdoctoral Science Foundation of Heilongjiang Province (LBH-Z11106), Technological Innovation Foundation for Youth Scholars of Harbin (2012RFQXG093), Foundation for University Key Teacher of Heilongjiang Province (1252G023), Postgraduate Innovation Research Project of Heilongjiang Province (YJSCX2012-126HLJ), Research Fund for the Doctoral Program of Higher Education (20102303120005), Province Natural Science Foundation of Heilongjiang (QC2013C060) and Science Funds for the Young Innovative Talents of HUST (NO.201304).

References

[1] V. Maja, L. Jim, R. Yaoping, H. Milton and R. Sriram., Editors. Assessing Service Development Readiness Using Enterprise Crowdsourcing. Proceeding of the 2013 IFIP/IEEE International Symposium on Intergrated Network Management, (**2013**) May 27-31: Ghent, Belgium

[2] R. Philip, B. Olivia, K. Yakov, H. Chang, A. J. Quinn, B. B. Bederson, J. Using Targeted Paraphrasing and Monolingual Crowdsourcing to Improve Translation. Computer Communication Review. 3, 4 (**2013**)

[3] H. Kyuwoong and L. Soo-Young. J. Environment Audio Scene and Activity Recognition through Mobile-based Crowdsourcing. IEEE Transaction on Consumer Electronics. 2, 58 (**2012**)

[4] F. Steffen, M. Ian, S. Christian, P. Christoph, S. Linda, S. Dmitry, V. D. V. Marijn K. Florian, O. Micheal, J. Geo-wiki.org: The Use of Crowdsourcing to Improve global land cover. Remote Sensing. 3, 1 (**2009**)

[5] D. Peng, L. Christopher, Mausam and W. Daniel, J. POMDP-based control of workflows for crowdsourcing. Artificial Intelligence. 202, 52 (**2013**)

[6] Z. Zhi-Qiang, P. Ju-Sheng, X. Xiao-Qin. Z. Yong, J. Research on Crowdsourcing Quality Control Strategies and Evaluation Algorithm. Jisuanji Xuebao/Chinese Journal of Computer. 36, 8 (**2013**)

[7] M. Lease and E. Yilmaz, J. Crowdsourcing for Information Retrievel. ACM SIGIR Forum. 45, 2 (**2011**)

[8] G. Dinesh, K. Naidu, N. Animesh, N. Girija and P. Viswanath, J. MoneyBee: Towards Enabling a Ubiquitous, Efficient, and Easy-to-use Mobile Crowdsourcing Service in the Emerging Market. Bell Labs Technical Journal. 15, 4 (**2011**)

[9] V. Jeroen and D. V. Arijen, J. Obtaining High-quality Relevance Judgments Using Crowdsourcing. IEEE Internet Computing. 16, 5 (**2012**)

[10] Z. Liren and L. Ling, Editors. An Interactive Vocabulary Learning System Based on Word Frequency Lists and Ebbinghaus' Curve of Forgetting. International Workshop on Digital Media and Digital Content Management (**2011**) May 16-18: Hangzhou, China

[11] L. Na and Y. Fuyu, J. Detection User's Long-term Interest Based on Ebbinghaus Forgetting Carve. ICIC Express Letters, Part B: Applications. 2, 5 (**2011**)

[12] E. Yilmaz, J. A. Aslam and S. Robertson. A New Rank Correlation Coefficient for Information Retrieval. The 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (**2008**): New York, USA

[13] S. E. Robertson, E. Kanoulas and E. Yilmaz. Extending Average Precision to Graded Relevance Judgments. The 33rd Internatinal ACM SIGIR Conference on Research and Development in Information Retrieval, (2010): New York, USA