

# Re-Ranking Voting-Based Answers by Discarding User Behavior Biases

Xiaochi Wei<sup>1\*</sup>, Heyan Huang<sup>1</sup>, Chin-Yew Lin<sup>2</sup>, Xin Xin<sup>1\*</sup>, Xianling Mao<sup>1</sup>, Shangguang Wang<sup>3</sup>

<sup>1</sup>BJ ER Center of HVLIP&CC, School of Comp. Sci., Beijing Institute of Technology, Beijing, China

<sup>2</sup>Microsoft Research Asia, Beijing, China

<sup>3</sup>State Key Lab. of Net. and Swit. Tech., Beijing Univ. of Posts and Tele., Beijing, China

{wxchi, hhy63}@bit.edu.cn, cyl@microsoft.com, {xxin, maoxl}@bit.edu.cn, sgwang@bupt.edu.cn

## Abstract

The vote mechanism is widely utilized to rank answers in community-based question answering sites. In generating a vote, a user's attention is influenced by the answer position and appearance, in addition to real answer quality. Previously, these biases are ignored. As a result, the top answers obtained from this mechanism are not reliable, if the number of votes for the active question is not sufficient. In this paper, we solve this problem by analyzing two kinds of biases; position bias and appearance bias. We identify the existence of these biases and propose a joint click model for dealing with both of them. Our experiments in real data demonstrate how the ranking performance of the proposed model outperforms traditional methods with biases ignored by 15.1% in precision@1, and 11.7% in the mean reciprocal rank. A case study on a manually labeled dataset further supports the effectiveness of the proposed model.

## 1 Introduction

Community-based Question Answering (cQA) sites, such as Yahoo Answers, Baidu Knows, Quora and Guokr, are the crowd-sourced alternatives to search engines for providing information [Liu *et al.*, 2011; Wu *et al.*, 2014]. They provide an open platform for users to ask questions and publish answers. Vote mechanisms are widely used in almost all cQA sites to select the best answers and filter out spam manually.

In the vote mechanism, if each voter can fairly examine all answers carefully, by using just a few votes, high-quality answers can easily be selected. Unfortunately, there are biases impacting the performance of traditional vote mechanisms significantly. In this paper, we deal with two kinds of biases; position bias and appearance bias.

- **Position Bias:** Position bias means that users are more likely to examine answers at top positions than others in answer lists. This means answers at top position have more probabilities to be examined and to be further voted

\*The paper was done when the first and fourth authors were visiting Microsoft Research Asia under the supervision of the third author.

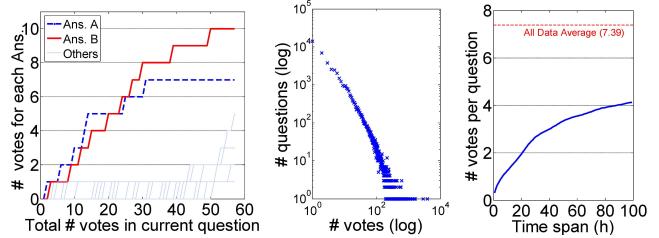


Figure 1: Limitations of the vote mechanism in cQA sites

on; the ones at lower positions, e.g. late published answers, are often ignored, even when their qualities are above the average in many cases.

- **Appearance Bias:** Appearance bias means that users are more likely to examine answers with attractive appearances, such as ones with images or long content. The probability of examining and voting such answers is larger than plain ones.

Due to these biases, the top answers selected by current vote mechanisms are reliable only when enough votes are accumulated. Figure 1(a) is a real example of this limitation. The x-axis denotes the total vote count for the active question, and the y-axis denotes the vote count for single answers. In this example, Ans. A is generated earlier than Ans. B by 6 hours. Thus when Ans. B is generated, Ans. A has already obtained some votes, and ranked at a higher position. Intuitively, Ans. B is better than Ans. A if we consider the voting results in the final stage. But the vote number of Ans. B outperforms that of Ans. A only when the total vote count of the question is larger than 28, due to the bias problem discussed above.

Nevertheless, for a general question, obtaining votes is a time-consuming process. As a result, most voting-based best answers are not reliable because of insufficient votes, and only a small number of questions can select trustworthy best answers. Figure 1(b) shows that vote tallies for questions follow a power-law distribution. Most questions have only a few votes, which is not sufficient for selecting reliable best answers. For example, in our dataset, there are 35,662 (more than 70%) questions have less than 10 votes. The time-consuming voting process also makes it impractical for most questions to select reliable best answers. Figure 1(c) shows the vote

accumulation process for questions. The average number of votes for each question is 7.39. However, there are only 2.3 votes (31.1% of the average value), after 24 hours, and there are 4.1 votes (55.5% of average value), after 100 hours. Therefore, the problem of insufficient votes significantly impacts the performance of current voting mechanisms, motivating us to find a way to re-rank voting-based answers with only a few votes.

In this paper, to improve the mechanism for ranking voting-based answers in a cQA site, when only a few votes are accumulated, we propose a joint click model, which considers two kinds of biases; position bias and appearance bias. The main premise is that in the ranking decision process, if these biases are modeled, and removed when counting votes, the returned result will be more accurate to reflect real answer qualities.

The task in this paper is different from previous work on answer quality prediction through machine learning. Votes can be seen as ground truths, labeled by crowdsourcing. Thus the voting mechanism aims to rank answers through human labeling, though biases make directly counting the votes occasionally inaccurate. In this paper, we improve methods for counting votes by modeling biases. Previous works, however, focus on automatically predicting the answer quality without manual labels, which is a standard classification problem.

Our work has three primary contributions:

- We have collected a large cQA dataset including more than 110,000 questions. Different from other published data, it contains a voting sequence for each question. So real user voting behaviors can be simulated directly.
- A novel joint click model is proposed to rank answers, when only a few votes are accumulated, with biases in user voting behaviors removed.
- Experiments in real cQA datasets with more than 50,000 questions shows, that the proposed approach outperforms traditional methods with biases ignored by 15.1% in precision@1, and 11.7% in the mean reciprocal rank.

## 2 Related Work

### 2.1 cQA Answer Quality Prediction

Although there have been many studies of cQA sites, to the best of our knowledge, the re-ranking of voting-based answers with a limited number of votes by modeling user voting behavior biases, has rarely been systematically studied before. Answer quality prediction is the system most similar to ours. The most representative method is the work presented by Surdeanu et al. [Surdeanu et al., 2008]. They used learning to rank with relevance functions as evidence. Later researchers have tried to explore effective features [Dalip et al., 2013] to get more accurate ranking results, such as user information [Shah and Pomerantz, 2010; Zhang et al., 2014], user profiles [Zhou et al., 2012], user expertise [Suryanto et al., 2009], content length [Agichtein et al., 2008] and comments [Burel et al., 2012]. Other methods have also been explored to solve this problem, Wang et al [Wang et al., 2009] utilize analogical reasoning to generate positive links and negative links between questions and answers and

Toba et al. [Toba et al., 2014] treat it as a hierarchy classification problem. Yao et al. [Yao et al., 2014] view this problem from a comprehensive perspective rather than a single static problem. Some works have realized the inaccuracy of ground truth selection, which simply treats answers with the most number of votes as best, and try to calibrate votes with different features [Chen et al., 2012]. However, all these works can be categorized into the answer quality prediction problem, which is usually solved as a traditional classification problem with different features, rather than modeling user voting behavior biases as ours. In this paper, we aim to improve the vote mechanism of human judgements to more appropriately rank answers with fewer votes, we deeply analyze different biases in user voting behaviors and propose a novel joint click model to remove these biases.

### 2.2 Click Models

Click models have been used successfully in search engines. Position bias is introduced to explain the intrinsic relevance between query and document. This was first introduced by Granka et al. [Granka et al., 2004]. They found that click probability decreases rapidly with lower display position. Richardson et al. [Richardson et al., 2007] try to model position bias and to find the intrinsic relevance. Craswell et al. [Craswell et al., 2008] use the examination hypothesis to explain position bias; they assume that a user will click a URL if and only if the user has examined it and the URL is relevant to the query. Several subsequent click models have been proposed based on this hypothesis, such as the User Browsing Model (UBM) [Dupret and Piwowarski, 2008], the Dyanmic Bayesian Network Click Model (DBN) [Chapelle and Zhang, 2009], the Click Chain Model (CCM) [Guo et al., 2009] and the General Click Model (GCM) [Zhu et al., 2010]. Appearance bias, founded by Yue et al.[Yue et al., 2010], is another important bias that explains more factors influencing user click behavior in the modeling intrinsic relevance. These biases have been utilized in other tasks rather than search engines [Li et al., 2014]. Since our model also follows the examination hypothesis, both position bias and appearance bias are modeled in the examination probability simultaneously. To the best of our knowledge, these biases have rarely been thoroughly explored in cQA sites previously.

## 3 Modeling User Voting Behaviors

### 3.1 User Voting Assumptions

Some assumptions, including examination and answer quality independent assumptions, are made, in order to remove biases in voting behaviors.

- **Examination Assumption:** An answer must be examined before a user decides whether to vote or not. Each answer has a certain probability of being examined, which is influenced by both the answer position and appearance at first sight, when browsing question pages. Following this assumption, both position bias and appearance bias are modeled in the examination probability simultaneously.

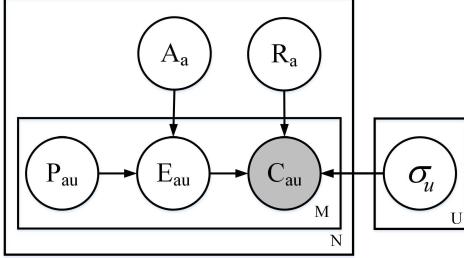


Figure 2: The graphical representation of JCM

Table 1: Major symbol notations

Symbol	Description
$N$	The total number of q-a pairs in the dataset
$M$	The total number of user behaviors (vote or skip) on current answer
$U$	The total number of users involved in the question
$A_a$	Appearance bias of answer $a$
$R_a$	Whether answer $a$ is high quality
$P_{au}$	Position bias when user $u$ examines answer $a$
$E_{au}$	Whether user $u$ examines answer $a$
$C_{au}$	Whether user $u$ votes for answer $a$
$\sigma_u$	The voting preference of user $u$

- **Answer Quality Independent Assumption:** The answer quality is independent of the examination probability, thus they should be modeled into different variables. Following this assumption, answer quality is separated from biased votes, and more trustable best answers can be obtained without the impact of biases.

### 3.2 Model Description

Based on the assumptions described above, a novel joint click model (JCM) is proposed. The graphical representation of the model is shown in Figure 2. Descriptions of major symbols are listed in Table 1. We assume that there are  $N$  question-answer pairs in our dataset. For a given answer  $a$ , there are  $M$  users voting or skipping it. As JCM follows the examination hypothesis, the current answer  $a$  is voted on by user  $u$  only if it is examined and is high quality. The probability a vote occurring can be described as follows:

$$\begin{aligned} & P(C_{au} = 1|a, u) \\ &= P(C_{au} = 1|a, u, E_{au} = 1, R_a = 1) \quad (1) \\ & P(R_a = 1|a, u, E_{au} = 1)P(E_{au} = 1|a, u). \end{aligned}$$

We assume that users do not vote for unexamined answers or low-quality answers, so the following equations are used in the inference of Equation (1):

$$P(C_{au} = 1, E_{au} = 0|a, u) = 0, \quad (2)$$

$$P(C_{au} = 1, R_a = 0|a, u, E_{au} = 1) = 0, \quad (3)$$

For simplicity, these three terms in Equation (1) can be denoted by  $\nu_{au}$ ,  $\beta_{au}$  and  $\gamma_{au}$  respectively. Since  $R_a$  is unrelated to  $u$ ,  $\beta_{au}$  can be denoted by  $\beta_a$ , then Equation (1) is written as:

$$P(C_{au} = 1|a, u) = \nu_{au}\beta_a\gamma_{au}. \quad (4)$$

Table 2: Features used in experiments

Categories	Features
Appearance	# characters in current answer
	# line breaks in current answer
Position	Whether current answer has images
	Current Ranking
Quality	Total # characters in front of current answer
	Total # images in front of current answer
	Total # line breaks in front of current answer
	# characters in current answer
	# line breaks in current answer
	# votes in current answer
	Whether current answer has images
	The ratio between # images and # words
	The ratio between # symbols and # words

In this formula, both appearance bias  $A_a$  on answer  $a$  and position bias  $P_{au}$  when user  $u$  examine answer  $a$  are represented in variable  $\gamma_{au}$ . A logistic function  $\sigma(x) = 1/(1 + e^{-x})$  is utilized to describe these two biases:

$$A_a = \sigma(\omega^A \cdot f_a^A), \quad (5)$$

$$P_{au} = \sigma(\omega^P \cdot f_{au}^P), \quad (6)$$

In these formulas,  $f_a^A$  is the feature vector to describe appearance bias of answer  $a$ , and similarly  $f_{au}^P$  describes position bias when user  $u$  examine answer  $a$ .  $\omega^A$  and  $\omega^P$  are the weight vectors for appearance features and position features respectively. Linear combination enables these two biases to describe the examination probability  $\gamma_{au}$  together, and  $\alpha$  is employed to balance these two parts:

$$\gamma_{au} = \alpha A_a + (1 - \alpha) P_{au} \quad (7)$$

The logistic function is also used to describe the answer quality  $\beta_a$ , similar to the method in [Wang *et al.*, 2013]:

$$\beta_a = \sigma(\omega^R \cdot f_a^R) \quad (8)$$

where  $f_a^R$  is the quality feature vector of answer  $a$  and  $\omega^R$  is the weight vector. In this paper, the answer quality is described by some features from the answer content, while more complex features can also be added into the framework to improve performance. All features we utilize in experiments are described in Table 2.

Lastly,  $\nu_{au}$  is used to describe the probability that user  $u$  votes for a high-quality answer  $a$  after examining it, inspired by the work [Xing *et al.*, 2013]. We assumes that the probability is generated from a user's personal Gaussian prior. Smaller mean value means users are stricter, and less high-quality answers are voted.

### 3.3 Model Inference

In the model learning phase, answer quality, examination probability and users' voting preferences are not observed in our data, so the Expectation Maximization (EM) algorithm [Neal and Hinton, 1998] is used to estimate parameters  $\Theta = \{\omega^A, \omega^P, \omega^R\}$ , which maximizes the lower bound of the

Table 3: The detail of the dataset

Data	Statistic Content	Avg.	Min	Max
Question	# Answers	5.68	1	50
	# Votes	15.55	1	3,967
Answer	# Votes	2.74	1	852

log-likelihood of observations in the data

$$\begin{aligned}
 L(\mathbf{C}, \Theta) &= \sum_{a,u} \log \sum_{E_{au}, R_a} P(E_{au}, R_a, C_{au} | \Theta) \\
 &\geq \sum_{a,u} \sum_{E_{au}, R_a} P(E_{au}, R_a | C_{au}, \Theta) \log P(E_{au}, R_a, C_{au} | \Theta) \\
 &= Q(\Theta, \Theta^i).
 \end{aligned} \tag{9}$$

In the E-Step, the posterior distribution is calculated

$$\begin{aligned}
 &P(E_{au}, R_a | C_{au}, \Theta^i) \\
 &= \frac{P(C_{au} | E_{au}, R_a, \Theta^i) P(E_{au}, R_a | \Theta^i)}{\sum_{E_{au}, R_a} P(C_{au} | E_{au}, R_a, \Theta^i) P(E_{au}, R_a | \Theta^i)}
 \end{aligned} \tag{10}$$

And in the M-Step, we maximize  $Q(\Theta, \Theta^i)$  in Equation (9). L-BFGS[Liu and Nocedal, 1989] is used to calculate  $\omega^A$ ,  $\omega^P$  and  $\omega^R$ .

## 4 Experiments and Results

### 4.1 Dataset and Metrics

We collect a large dataset of cQA, including more than 110,000 questions, more than 390,000 answers, and more than 780,000 votes, from Chinese cQA site Guokr<sup>1</sup>. Every item (question, answer and vote) has a time stamp. After removing questions without votes or answers, a dataset used in our experiments contains 50,536 questions, 287,127 answers and 785,717 votes. More details are shown in Table 3.

The evaluation methods for traditional cQA answer prediction are not suitable, as there are not sufficient votes in most questions discussed in Section 1, so only questions with a large number of votes can be selected as test questions. Additionally, in order to make the task more challenging, questions with ambiguous answers are selected as test questions. Therefore, 1,365 test questions are selected based on two strict rules: (1) the question has more than 60 votes; (2) the answer with the most votes has less than twice the vote number of the second one, when only the former 15 votes in the question are considered. Their ranking based on the vote number in the final stage is regarded as the ground truth. The former 5% – 30% of votes on test questions is used as training data, together with the votes on other questions in the dataset.

Two standard information retrieval metrics are adapted in our experiments, Precision at K (P@K) and Mean Reciprocal Rank (MRR). P@K reports the proportion of best answers ranked in the top K result. MRR is the average of the reciprocal ranks of the best answer for test questions and can be calculated from below formula:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{11}$$

<sup>1</sup><http://www.guokr.com/>

where  $|Q|$  is the number of test questions and  $rank_i$  is the position of the best answer in the ground truth.

### 4.2 Overall Performance

We compare the proposed model with following methods:

- **Tradition Vote Mechanism(Vote)** This is the traditional user vote method widely used in cQA sites. It ranks answers based on vote quantity.
- **User Browsing Model(UBM)** [Dupret and Piwowarski, 2008]. UBM only utilizes the ranking of current answers to model position bias. The answer quality is described with a latent variable.
- **General Click Model(GCM)** [Zhu *et al.*, 2010]. GCM considers different session-specific features in modeling position biases. Answer quality is described with answer-specific features. Appearance biases are ignored in this model.
- **Appearance Click Model(ACM)** [Yue *et al.*, 2010] ACM only utilizes answer-specific appearance features to describe appearance biases and the answer quality is described with the same method as GCM. Position biases are not considered in this model.
- **Vote Calibration(Calib)** [Chen *et al.*, 2012] The Calib method assigns each vote a particular weight value based on different answer-specific and user-specific features, rather than modeling biased user voting behaviors, is used to calibrate votes.

In our experiments, we range the training data from 5% to 30%. The result is shown in Table 4. It can be observed that click models, i.e. UBM, GCM, ACM and JCM, all outperform traditional user voting mechanisms. This demonstrates the existence of biases in users' voting behaviors. After removing biases, answers can be ranked with votes more accurately. Our JCM achieves better results on all training data, statistically much better than traditional voting methods. When only 5% votes are observed, a relative improvement of 15.1% in P@1 and 11.7% in MRR is achieved. Compared with other baseline methods, the improvement is achieved by removing both position biases and appearance biases simultaneously, for example, a relative improvement of 4.2% on P@1 and 3.0% on MRR compared with GCM, and a relative improvement of 3.7% on P@1 and 3.0% on MRR compared with ACM is achieved, when 5% votes are used. Our method also outperforms the state of the art Calib method considerably, with a relative improvement of 4.1% on P@1 and 3.2% on MRR with 5% training data.

It can be observed that a larger improvement is achieved compared with voting methods with fewer votes. This shows that the answer with the most number of votes can be recognized as the best answer, only if there are enough votes, as we have discussed in Section 1.

Figure 3 illustrates the performance of all methods in Precision@K. We select questions with more than 5 answers as test questions, and 5% of the training data is utilized. In all cases, the proposed method outperforms other baselines.

Figure 4 shows the frequency distribution of answer quality scores estimated by different methods. It is interesting to ob-

Table 4: The performance of all methods

Metrics	Methods	Training Data					
		5%	10%	15%	20%	25%	30%
P@1	Vote	0.4703	0.5927	0.6996	0.7663	0.8175	0.8447
	UBM	0.5121	0.6271	0.7209	0.7707	0.8212	0.8505
	GCM	0.5194	0.6347	0.7216	0.7766	0.8278	0.8513
	ACM	0.5223	0.6396	0.7172	0.7861	0.8308	0.8535
	Calib	0.5201	0.6256	0.7011	0.7824	0.8264	0.8491
	JCM	<b>0.5414</b>	<b>0.6579</b>	<b>0.7473</b>	<b>0.7971</b>	<b>0.8440</b>	<b>0.8674</b>
MRR	Vote	0.6469	0.7505	0.8276	0.8685	0.8980	0.9126
	UBM	0.6894	0.7662	0.8288	0.8698	0.8977	0.9129
	GCM	0.7012	0.7806	0.8397	0.8747	0.9035	0.9179
	ACM	0.7014	0.7867	0.8366	0.8799	0.9072	0.9202
	Calib	0.6999	0.7799	0.8297	0.8777	0.9041	0.9173
	JCM	<b>0.7224</b>	<b>0.8014</b>	<b>0.8574</b>	<b>0.8875</b>	<b>0.9185</b>	<b>0.9278</b>

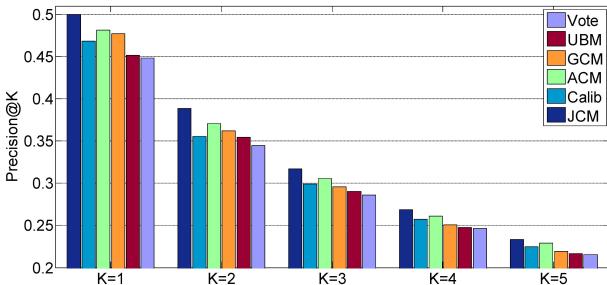


Figure 3: The performance in P@K with 5% training data

serve that the distribution of the JCM score follows a Gaussian distribution, which meets the common sense about answer quality:(a)quality scores of most answers are about average level; (b)there is a small portion of answers of either high quality or low quality. The human judgement result shown in Figure 8(a) also reflects this observation. However, scores obtained by other methods are more similar as a power-law distribution, mainly because these results are greatly influenced by biased user votes.

### 4.3 Evaluation on Different Questions

Another experiment is conducted on questions with different numbers of answers in order to examine the applicability of our approach. Questions in the test set are divided into 5 parts, ensuring that similar numbers of questions are allocated into each part. We compare our approach with other baselines in all parts respectively. 10% the training data is used and the experiment results are shown in Figure 5. It can be observed that our approach consistently outperforms other methods in all conditions. It is noteworthy that a significant performance improvement is achieved on questions with more answers, because biases influence user voting behaviors more in these questions than the ones with less answers. When browsing a question page with many answers, it is unlikely for a user to examine all answers, thus the ones with attractive appearance or at the top rank are more likely to be examined.

### 4.4 Parameter Analyses

The advantage of our approach is that both position biases and appearance biases in cQA sites can be removed simultaneously. The overall performance has demonstrated the effectiveness

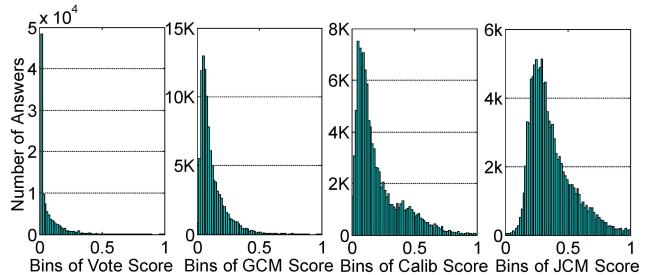


Figure 4: The frequency distributions of answer score estimated by different methods (distributions of UBM,GCM and ACM are similar); answer score has been normalized

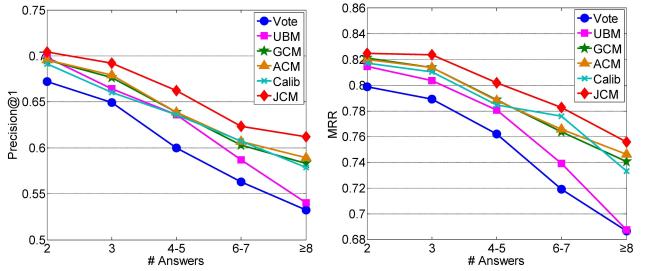


Figure 5: Performance comparison on different questions

of our approach in estimating answer quality. In this part the influence of position bias and appearance bias is shown. We use 5% of votes as training data and change the parameter  $\alpha$  in our model, which controls the ratio of these two biases in the examination probability, particularly,  $\alpha = 0$  meaning only position bias is modeled, and when  $\alpha = 1$  only appearance bias is modeled. Figure 6 shows the sensitivity of parameter  $\alpha$  in balancing the ratio of two biases. It is shown that these two kinds of biases exist in user voting behaviors simultaneously. Neither only position bias nor only appearance bias being removed can make the model perform at a sufficient level.

Additionally, we change the prior of the value  $\nu_{au}$  to different distributions. In order to show the distribution clearly, 25% and 30% of the training data, containing more votes, is used. The result is shown in Figure 7 and it is clear that the uniform distribution performs worst, which demonstrates the importance of distinguishing user preference as Xing et al. describe [Xing et al., 2013]. The gaussian distribution performs better compared with others.

### 4.5 Case Study

In order to demonstrate the effectiveness of our approach more objectively, 5 volunteers are hired to judge answer quality. 100 questions are sampled randomly from the test set. Question-answer pairs are displayed to volunteers randomly, and they are rated from 1 to 5 based on predefined guidelines, thus each answer has 5 ratings. The rating distribution is shown in Figure 8(a).

We calculate average ratings of all these questions' best answers got with different methods. 5% of the votes are used as training data. The result is shown in Figure 8(b). Our approach JCM outperforms all baselines with the highest score of

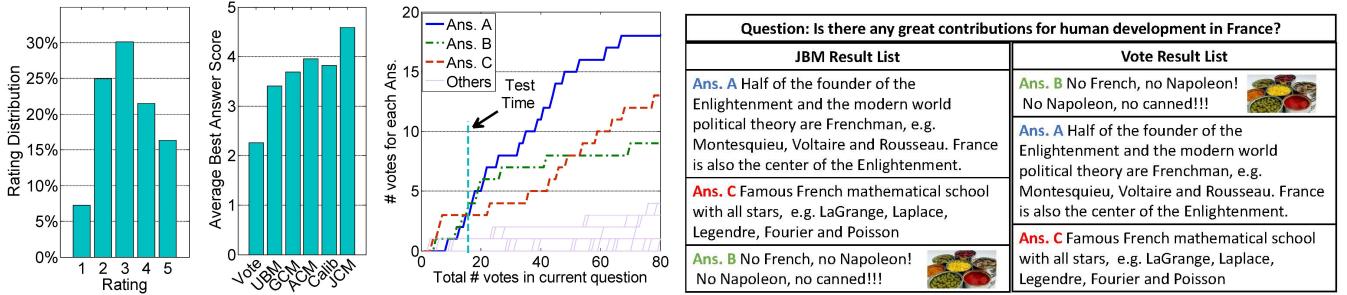


Figure 8: Result of case study and a real example of test question

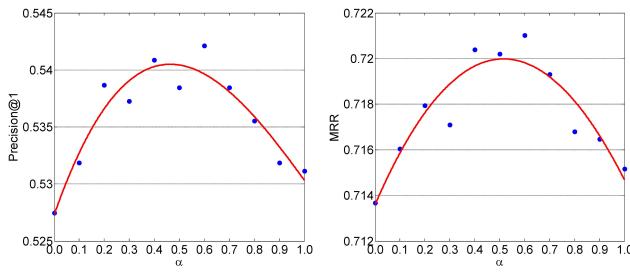


Figure 6: Performance of different value of parameter  $\alpha$

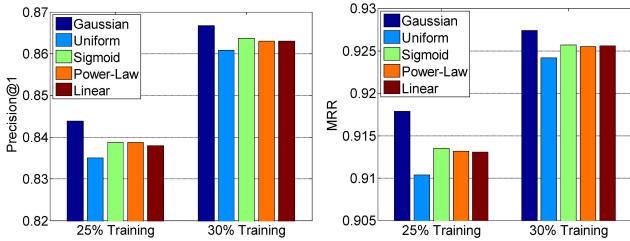


Figure 7: The performance of alternative distributions used to describe user voting preference

4.59. Both appearance biases and position biases are removed manually, because all answers are examined during the rating process. The result strongly demonstrates the effectiveness of our approach in removing biases with the unbiased data as the ground truth. It is observed that an improvement score of more than 1 is achieved after removing either position biases by UBM and GCM or appearance biases by ACM. The significant improvement of 0.93 compared with GCM and 0.64 compared with ACM is achieved by JCM after removing both of the two biases. The proposed method also outperforms the state of the art Calib method with an improvement of 0.77. It demonstrates that trustable best answers can be selected accurately after removing position biases and appearance biases, even if only a few votes are accumulated.

Human judgement is used to show that the answer ranking results of our approach are sensible. We randomly sample 500 questions with different ranking results between the voting method and JCM. Each question and its two answer ranking results are shown to volunteers. They are required to judge

which ranking fits more sensibly without knowing ranking methods. After statistics, our approach is selected in 412, with 82.4%, of 500 questions.

A real example is selected from test questions. Figure 8(c) shows the top 80 votes (320 in total) in the question, and Figure 8(d) describes the question content and the two answer ranking results of JCM and the traditional voting method with 5% training data. The voting method selects Ans. B as the best answer inappropriately, however, Ans. A is selected correctly with our approach when the vote number is limited. After analyzing contents and votes of answers in this example, we find that both position biases and appearance biases play an important role in influencing user voting behaviors. Ans. C collects more votes when only a few answers are published, e.g. 4, due to position bias. With more answers published, users are more likely to vote for answers with long content, e.g. Ans. A, or images, e.g. Ans. B due to appearance bias. This also demonstrates the result shown in Figure 5 that these two kinds of biases influence user voting behaviors simultaneously. After removing both of them, reliable best answers can be selected with only a few votes.

## 5 Conclusion

In this paper, we have studied the problem of how to rank voting-based answers with insufficient votes in cQA sites. Position bias and appearance bias are both shown to exist in users' voting behaviors. Due to these biases, trustable best answers can hardly be selected correctly with only a few votes. We have proposed a joint click model to remove both position bias and appearance bias in voting behaviors simultaneously. Experiments on real cQA data show the effectiveness of our approach in removing biases and a case study shows that our approach can select trustable best answers with a limited number of votes.

## Acknowledgments

The work described in this paper was mainly supported by National Basic Research Program of China (973 Program, Grant No. 2013CB329605 and No.2013CB329303), Electronic Information Industry Foundation of China, and National Natural Science Foundation of China (No. 61300076 and No.61402036)

## References

- [Agichtein *et al.*, 2008] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *WSDM'08*, pages 183–194. ACM, 2008.
- [Burel *et al.*, 2012] Grégoire Burel, Yulan He, and Harith Alani. Automatic identification of best answers in online enquiry communities. In *The Semantic Web: Research and Applications*, pages 514–529. Springer, 2012.
- [Chapelle and Zhang, 2009] Olivier Chapelle and Ya Zhang. A dynamic bayesian network click model for web search ranking. In *WWW'09*, pages 1–10. ACM, 2009.
- [Chen *et al.*, 2012] Bee-Chung Chen, Anirban Dasgupta, Xu-anhui Wang, and Jie Yang. Vote calibration in community question-answering systems. In *SIGIR'12*, pages 781–790. ACM, August 2012.
- [Craswell *et al.*, 2008] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *WSDM'08*, pages 87–94. ACM, 2008.
- [Dalip *et al.*, 2013] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow. In *SIGIR'13*, pages 543–552. ACM, 2013.
- [Dupret and Piwowarski, 2008] Georges E Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR'08*, pages 331–338. ACM, 2008.
- [Granka *et al.*, 2004] Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR'04*, pages 478–479. ACM, 2004.
- [Guo *et al.*, 2009] Fan Guo, Chao Liu, Anitha Kannan, Tom Minka, Michael Taylor, Yi-Min Wang, and Christos Faloutsos. Click chain model in web search. In *WWW'09*, pages 11–20. ACM, 2009.
- [Li *et al.*, 2014] Yanen Li, Anlei Dong, Hongning Wang, Hongbo Deng, Yi Chang, and ChengXiang Zhai. A two-dimensional click model for query auto-completion. In *SIGIR'14*, pages 455–464. ACM, 2014.
- [Liu and Nocedal, 1989] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.
- [Liu *et al.*, 2011] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR'11*, pages 415–424. ACM, 2011.
- [Neal and Hinton, 1998] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.
- [Richardson *et al.*, 2007] Matthew Richardson, Ewa Domiowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW'07*, pages 521–530. ACM, 2007.
- [Shah and Pomerantz, 2010] Chirag Shah and Jefferey Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR'10*, pages 411–418. ACM, 2010.
- [Surdeanu *et al.*, 2008] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers on large online qa collections. In *ACL'08*, pages 719–727. Citeseer, 2008.
- [Suryanto *et al.*, 2009] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger HL Chiang. Quality-aware collaborative question answering: methods and evaluation. In *WSDM'09*, pages 142–151. ACM, 2009.
- [Toba *et al.*, 2014] Hapnes Toba, Zhao-Yan Ming, Mirna Adriani, and Tat-Seng Chua. Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences*, 261:101–115, 2014.
- [Wang *et al.*, 2009] Xin-Jing Wang, Xudong Tu, Dan Feng, and Lei Zhang. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *SIGIR'09*, pages 179–186. ACM, 2009.
- [Wang *et al.*, 2013] Hongning Wang, ChengXiang Zhai, Anlei Dong, and Yi Chang. Content-aware click modeling. In *WWW'13*, pages 1365–1376. ACM, 2013.
- [Wu *et al.*, 2014] Haocheng Wu, Wei Wu, Ming Zhou, Enhong Chen, Lei Duan, and Heung-Yeung Shum. Improving search relevance for short queries in community question answering. In *WSDM'14*, pages 43–52. ACM, 2014.
- [Xing *et al.*, 2013] Qianli Xing, Yiqun Liu, Jian-Yun Nie, Min Zhang, Shaoping Ma, and Kuo Zhang. Incorporating user preferences into click models. In *CIKM'13*, pages 1301–1310. ACM, 2013.
- [Yao *et al.*, 2014] Yuan Yao, Hanghang Tong, Feng Xu, and Jian Lu. Predicting long-term impact of cqa posts: A comprehensive viewpoint. In *KDD'14*. ACM, 2014.
- [Yue *et al.*, 2010] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *WWW'10*, pages 1011–1018. ACM, 2010.
- [Zhang *et al.*, 2014] Jingyuan Zhang, Xiangnan Kong, Roger Jie Luo, Yi Chang, and Philip S Yu. Ncr: A scalable network-based approach to co-ranking in question-and-answer sites. In *CIKM'14*, pages 709–718. ACM, 2014.
- [Zhou *et al.*, 2012] Zhi-Min Zhou, Man Lan, Zheng-Yu Niu, and Yue Lu. Exploiting user profile information for answer ranking in cqa. In *WWW'12*, pages 767–774. ACM, 2012.
- [Zhu *et al.*, 2010] Zeyuan Allen Zhu, Weizhu Chen, Tom Minka, Chenguang Zhu, and Zheng Chen. A novel click model and its applications to online advertising. In *WSDM'10*, pages 321–330. ACM, 2010.