LETTER

Truth Discovery of Multi-Source Text Data

Chen CHANG[†a)], **Jianjun CAO**[††b)], **Qin FENG**[†c)], **Nianfeng WENG**[††d)], ***Nonmembers***, ***and*** **Yuling SHANG**[†e)], ***Member***

SUMMARY Most existing truth discovery approaches are designed for structured data, and cannot meet the strong need to extract trustworthy information from raw text data for its unique characteristics such as multifactorial property of text answers (i.e., an answer may contain multiple key factors) and the diversity of word usages (i.e., different words may have the same semantic meaning). As for text answers, there are no absolute correctness or errors, most answers may be partially correct, which is quite different from the situation of traditional truth discovery. To solve these challenges, we propose an optimization-based text truth discovery model which jointly groups keywords extracted from the answers of the specific question into a set of multiple factors. Then, we select the subset of multiple factors as identified truth set for each question by parallel ant colony synchronization optimization algorithm. After that, the answers to each question can be ranked based on the similarities between factors answer provided and identified truth factors. The experiment results on real dataset show that though text data structures are complex, our model can still find reliable answers compared with retrieval-based and state-of-the-art approaches.

key words: *truth discovery, ant colony optimization, text mining*

1. Introduction

Data even describing the same object or event, can come from a variety of sources and may conflict with each other [1]. In the light of this challenge, truth discovery is motivated by the strong need to resolve conflicts among multi-sourced noisy information [2]–[4]. However, most existing truth discovery methods are designed for structured data, and are difficult to be directly applied to text data.

Actually, there are several unique characteristics of natural language that hinder the existing truth discovery methods from being successfully applied to text data. Stage 1 of Fig. 1 gives an illustration of questions and answers. First, the answer to a question may be multifactorial, and it is usually hard for a given answer to cover all the factors. For a question like "What are the symptoms of flu?", the answer contains fever, chills, cough, nasal symptom, etc. If a user provides two factors of the correct factors, the existing truth discovery methods may determine this answer to be completely wrong and assign a low reliability degree to this user. Second, answers provided by online users may convey a very similar meaning with different keywords. For example, users may use words such as tired or exhausted to describe the symptom of fatigue, but existing truth discovery methods may treat them as totally different two words. Thus, how to identify partially correct answers and model factors of text answers is critical for the task of truth discovery on text data.

This paper is related to the problems of collaborative question answering [5] and answer selection [6]. The models proposed before in this field are all supervised, which require external information or high-quality training sets. Such information, unfortunately, is not always available in real-world applications. Differently, the model proposed in this paper does not require labeled data for training. We extract trustworthy answers based on the unsupervised reliability estimation for each user. The different problem settings and solutions naturally distinguish these work from this paper.

In this paper, we propose a model that fits for the challenges to infer trustworthy information from text data. First, take the fine-grained answer factors into consideration rather than the whole answer as a unit. In this case, if the answer provided by this user is partially correct, the reliability degree of this user is able to get flexible adjustments. Second, delete stop words of answers, and replace the synonyms in the answers in order to eliminate the impact of the diversity of words usage. Third, transform truth discovery from text data problem into a subset problem, the parallel ant colony algorithm is designed to find the optimal subsets of factors for all problems. Fourth, the trustworthiness of answers for each question are ranked based on the factors provided, and reliability degree of each user is decided by the answers who provided. The major contributions of this paper are: (1) We solved the challenges of the truth discovery problem from text data. (2) An optimization-based truth discovery model was proposed, which extracts factors from answers, then transforms the truth discovery problem to a subset problem. (3) The model outperforms the retrieval-based and state-of-the-art approaches on real-world datasets.

2. Proposed Method

Given a set of questions $\{q\}_1^Q$, a set of users $\{u\}_1^U$, and a set of

Manuscript received December 24, 2018.
Manuscript revised May 30, 2019.
Manuscript publicized August 22, 2019.

[†]The authors are with the Army Engineering University of PLA, China.
[††]The authors are with the Sixty-third Research Institute, National University of Defense Technology, China.
a) E-mail: c308051252@163.com
b) E-mail: jianjuncao@yeah.net
c) E-mail: 284422873@qq.com
d) E-mail: wengf@gmail.com
e) E-mail: 13776606112@163.com

DOI: 10.1587/transinf.2018EDL8267

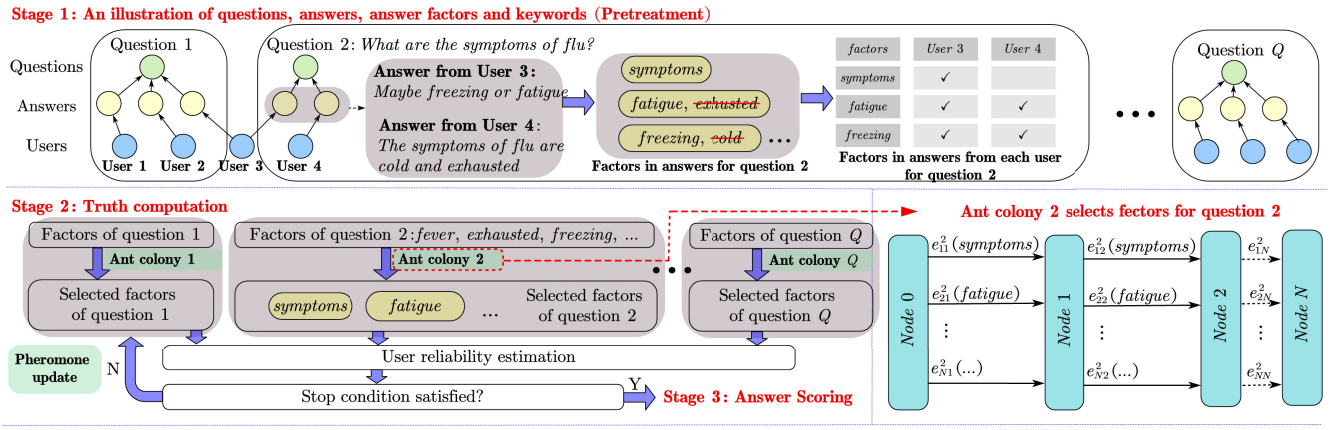


Fig. 1 Overall framework

answers $\{a_q^u\}_{1,1}^{U,Q}$, where Q denotes the number of questions, and U denotes the number of users. The purpose of this paper is to find highly-trustworthy answers and most reliable users for each question.

Overview: As in Stage 1 of Fig. 1, for each question q, we first extract the keywords in each answer a_q^u as a set $V_{a_q^u}(u = 1, 2, \ldots, U)$. In order to eliminate the diversity of semantics, we delete stop words and replace synonyms as a unified expression by platform Natural Language Toolkit (NLTK) [7]. These unified keywords are considered as factors of this question. After that, we aggregate all factors from all answers of question q as a set $\mathcal{V}_q(q = 1, 2, \ldots, Q)$. Based on the above steps, the purpose of our model is to find a subset V_q^* from $\mathcal{V}_q$ for each question, which is considered contains the factors that the correct answer should have.

According to two observations, we turn the text data truth discovery problem into an optimization problem: (1) the true answer should be as close as possible to the answer provided by each user. (2) the quality assessment of the user is crucial, the higher quality of the u-th user, the more similar $V_{a_q^u}$ to V_q^*. The objective function of proposed model is defined as follows.

$$\begin{aligned} &\max_{\{V_q^*\}\{r_u\}} \quad \sum_{u=1}^{U} r_u \frac{1}{|\Gamma_u|} \sum_{q \in \Gamma_u} s\left(V_{a_q^u}, V_q^*\right) \\ &\text{s.t.} \quad \sum_{u=1}^{U} r_u = 1 \ , r_u \in \mathbb{R}^+ \end{aligned} \tag{1}$$

Where r_u is defined as the reliability degree of u-th user, Γ_u is the set of questions that u-th user provides answers to, and $s(\cdot)$ is the Jaccard similarity between the set of factors of answers $V_{a_q^u}$ and the identified truth factors of q-th question V_q^*.

$$s\left(V_{a_q^u}, V_q^*\right) = \frac{|V_{a_q^u} \cap V_q^*|}{|V_{a_q^u} \cup V_q^*|} \tag{2}$$

The objective function of the optimization problem is that the weighted sum of the similarities of selected factors V_q^* and each answer's factors $V_{a_q^u}$ reaches the maximum.

I. Truth Computation: in this step, user reliabilities $\{r_u\}$ are fixed, and we solve for the truth sets $\{V_q^*\}$. We design a parallel ant colony synchronization optimization algorithm to solve this problem and find the highly-trustworthy answers for each question. As in Stage 2 of Fig. 1, Eq. (1) can be further split into Q separate parallel optimization problems, and each ant colony is associated with a question. Q colonies will search factors for coresponding questions independently and in parallel. The objective function for each ant colony is:

$$\begin{aligned} &\max_{\{V_q^*\}} \quad \sum_{u \in \Theta_q} r_u s\left(V_{a_q^u}, V_q^*\right) \\ &\text{s.t.} \quad \sum_{u=1}^{U} r_u = 1 \quad r_u \in \mathbb{R}^+ \end{aligned} \tag{3}$$

Inspired by [8], take question q as example, we construct directed graph for each colony to finish truth factors selection. Suppose N is the cardinal number of $\mathcal{V}_q$. Factors of $\mathcal{V}_q$ are put on the edges $e_{ij}^q (i = 1, 2, \ldots, N; j = 0, 1, 2, \ldots, N)$, where i indacates the i-th factor, and j indacates that the ant will step to j-th node in the directed graph. $\tau_{ij}^q(t)(t = 0, 1, 2, \ldots)$ represents the quality of pheromone on edge e_{ij}^q at time $t(t = 0, 1, 2, \ldots)$. α and β are the importance of pheromone and heuristic respectively. Heuristic information η_i^q is the local information, indicating the expectation of choosing i-th factor.

In the beginning $\tau_{ij}^q(0) = D$, and D is a constant. Ants of this colony will be generated on node 0 at time 0. Under the constraint conditions, each ant independently selects one edge to move to the next node according to the heuristic information and pheromone on the edges. The list $tabu_m^q (m = 1, 2, \ldots, M)$ is utlized to record edges which the m-th ant travelled. The probability of the m-th ant being transferred from node $j-1$ to node j through the route e_{ij}^q at time t is:

$$h_{ij}^q(t) = \begin{cases} \dfrac{\left(\tau_{ij}^q(t-1)\right)^\alpha \left(\eta_i^q\right)^\beta}{\sum_{e_{ij}^q \notin tabu_m^q} \left(\tau_{ij}^q(t-1)\right)^\alpha \left(\eta_i^q\right)^\beta} & e_{ij}^q \notin tabu_m^q \\ 0 & others \end{cases} \tag{4}$$

Where heuristic η_i^q is:

$$\eta_i^q = \frac{sum\left(v_i^q\right)}{\sum_{v_j^q \in \mathcal{V}_q} sum\left(v_j^q\right)} \tag{5}$$

Where $sum(\cdot)$ is defined as the number of occurrences of the factor v_i^q. It indicates that the expectation of different factors is proportional to the number of times it appears in all answers for each question. Note that only if the objective function value as Eq. (3) raises, the ant will step to next node, or this ant will be killed, and the $tubu_m^q(m = 1, 2, \ldots, M)$ is consdered as identified truth factors selected by the m-th ant. After all ants of this colony finish factors selection, the $tubu_m^q$ with the largest objective function as Eq. (3) will be considered as identified truth factors selected by colony q.

All colonies select identified truth sets $\{V_q^*\}$ in parallel. After all ant colonies have completed factors selection, the model calculates the objective function value as Eq. (1), and updates the pheromone on edges for each colony. The updating formula is as follows.

$$\tau_{ij}^q(t) = \begin{cases} (1-\rho)\tau_{ij}^q(t-1) + \dfrac{\Psi_q(tabu^q(t))}{C_q} & e_{ij}^q \in \Delta_q(tabu^q(t)) \\ (1-\rho)\tau_{ij}^q(t-1) & others \end{cases} \tag{6}$$

Where ρ is a constant and represents pheromone evaporation rate, $\frac{\Psi_q(tabu^q(t))}{C_q}$ is a pheromone enhancement formula, $\Psi_q(tabu^q(t))$ is the objective function value of pheromone enhancement, C_q is a constant which is used to adjust the amount of pheromone enhancement, $\Psi_q(tabu^q(t))$ is the equivalent routes of $tabu^q(t)$.

II. Reliability Estimation: in this step, identified truth sets for questions $\{V_q^*\}$ are fixed, and we solve for the reliabilities of uesrs $\{r_u\}$, updating strategy of $\{r_u\}$ is:

$$r_u = \frac{\frac{1}{|\Gamma_u|}\sum_{q\in\Gamma_u} s\left(V_{a_q^u}, V_q^*\right)}{\sum_{u=1}^{U}\frac{1}{|\Gamma_u|}\sum_{q\in\Gamma_u} s\left(V_{a_q^u}, V_q^*\right)} \tag{7}$$

Where $\frac{1}{|\Gamma_u|}\sum_{q\in\Gamma_u} s\left(V_{a_q^u}, V_q^*\right)$ is the average similarity between all answers provided by u-th user and the corresponding current round identified truth set V_q^*, and $\sum_{u=1}^{U}\frac{1}{|\Gamma_u|}\sum_{q\in\Gamma_u} s\left(V_{a_q^u}, V_q^*\right)$ is the sum of the average similarities of all users.

In the above, the identified truth set for each question is calculated by the user reliability, and the reliability of each user is calculated by the identified truth set. This process is modeled as an iterative process. When the objective function as Eq. (1) doesn't update for δ times (δ is set as 6 in this paper), the proposed model is stopped, and we have derived the $\{V_q^*\}$ and $\{r_u\}$ in Eq. (1).

III. Trustworthy-Aware Answer Scoring: in our method, the trustworthiness of the answer is evaluated by the volume of correct factors it provides. Hence, we propose a straightforward scoring mechanism to evaluate the trustworthiness score of each answer. Given the truth set V_q^*, the score of each answer a_q^u for question q is:

$$score_{a_q^u} = s\left(V_{a_q^u}, V_q^*\right) \tag{8}$$

Where $s(\cdot)$ is defined as Jaccard similarity.

At this point, the model gets the rank of the answers for each question and the reliabilities of users.

3. Experiments

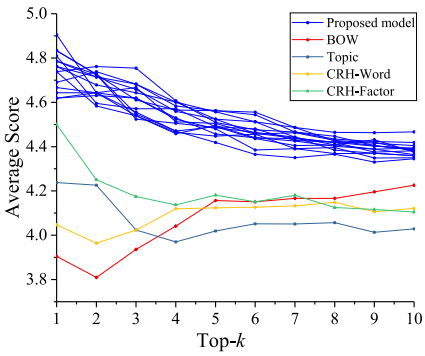
We perform a series of experiments in a real dataset **Questions/Student Answers with Grades** [9]. This dataset consists 21 questions, 30 users, and a total of 630 student answers. Each student answer is scored from 0 (completely incorrect) to 5 (perfect answer) by two human judges. Each question consists multiple correct answers, partial correct answers and untrustworthy answers. This dataset represents a general truth discovery scenario for factoid text questions and answers.

The task on this dataset is to extract Top-k (k is set to 1-10 in this paper) trustworthy student answers for each question, and we define the average score of the answers ranked in Top-k as evaluation metric in this paper.

The proposed model was compared against the state-of-the-art truth discovery and retrieval-based answer selection approaches. **Bag-of-Word (BOW) Similarity**: the bag-of-word vectors of questions and their answers are extracted. Answers are ranked according to the similarities between the question vectors and theirs corresponding answer vectors. **Topic Similarity**: We utilize Latent Dirichlet Allocation (i.e., LDA [10]) to extract topic representation for each question and its corresponding answers. Similar to BOW, answers are ranked according to the cosine similarity to the question. **CRH** [3] **+ Topic Dist.**: CRH is regarded as an excellent optimization based truth discovery framework. In the experiment, we use the topic distributions as the representations of the whole answers to be fed to CRH. **CRH** [3] **+ Word Vec.**: Similar to CRH + Topic Dist. but inputs are changed to the average word vectors of answers [11].

In order to test the performance of the model for all parameters combinations of ACO, according to the idea of uniform design [12], we adopt uniform design table $U_{15}(15^3)$ and selected 15 groups of "uniform" and "tidy" possible parameters combinations to perform experiments.

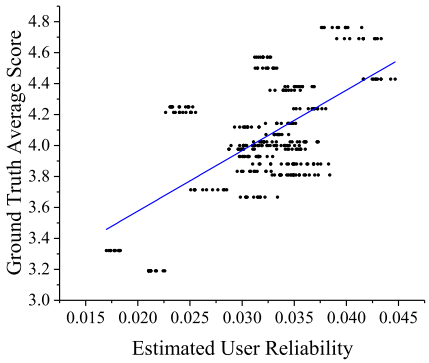
As in Fig. 2 (a), the proposed method consistently outperforms all the baseline approaches for 15 groups even with worst parameters combination. In other words, the proposed model demonstrates its great advantages on text data truth discovery. The reasons why the proposed model surpasses all the baseline methods are as follows. On one hand, retrieval-based approaches (i.e., BOW Similarity) rank the answers merely based on the semantic similarity between the question and answers. However, a question itself may not cover all the semantics that should be covered in reliable answers. Therefore, retrieval-based approaches only discover relevant answers rather than trustworthy answers.



(a) Performance on Real Data Set

Question	*What is a variable?*
Groundtruth Answer	*A **location** in **memory** that can **store** a **value**.*
Top Answer 1	*A **variable** is the **address** in a computer's **memory** where a **value** can be **stored** for use by a program.*
Top Answer 2	*A **variable** is a **location** in **memory** where a value can be **stored**.*
	...
Partial Correct Answer	*Variable is a way to **store** different **values** into the program.*
	...
Untrustworthy Answer	*Variable can be a intger or a string in a program.*

(b) Case Study of A Real Question



(c) Estimated User Reliabilities

Fig. 2 Experiment results

On the other hand, although state-of-the-art truth discovery approach CRH aims to capture user reliability, the performance is not great. It is because that it treats the answers as an integrated semantic unit, and single vector representations fail to capture the innate correlations among these answers. In other words, these approaches may ignore the fact that the semantic meaning of each answer may be complicated. Obviously, an inaccurate representation cannot be used to correctly estimate the reliabilities of users, an inaccurate user reliability estimation will further lead to an incorrect aggregated results.

To better evidence the analysis above, we give a case study about a question in the dataset. As in Fig. 2 (b), the proposed model can automatically select factors which are meaningful to the question from all factors (i.e., variable, location, etc.). The top-ranked answers have more meaningful factors than low-ranked untrustworthy answers.

Next, we further show the user reliabilities found at the end of the model. Since the dataset does not give the students' reliabilities directly, we use the average scores of students' answers to each question as ground truth reilabilities. As in Fig. 2 (c), each point represents a student, it can be seen that as the estimated reilabilities of students increase, the ground truth reilabilities increase. In other words, the proposed model estimated thereilabilities of students successfully.

Due to space limitation, we only show the results for the whole dataset, we randomly select 5 exams of 7 questions to test our model, the results on rest exams follow the same tendency. The code and all results will be released on github upon the acceptance of the paper.

4. Conclusion

For the current situation where there is little truth discovery method that can directly applied to text data, we propose an optimization based truth discovery model for text data which extract factors from answers. By transforming the truth discovery problem of text data to a subset problem, ant colony optimization algorithm is used, and performs well. The experiment results show that our model find more reliable answers compared with baseline approaches.

Acknowledgements

This work was supported by National Science Foundation of China (Grant No.61371196).

References

[1] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," Acm Sigkdd Explorations Newsletter, vol.17, no.2, pp.1–16, 2016.
[2] X.L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang, "Knowledge-based trust: estimating the trustworthiness of web sources," Vldb Endowment, vol.8, no.9, pp.938–949, 2015.
[3] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," ACM SIGMOD International Conference on Management of Data, pp.1187–1198, 2014.
[4] L. Li, B. Qin, W. Ren, and T. Liu, "Truth discovery with memory network," Tsinghua Science and Technology, vol.22, no.6, pp.609–618, 2017.
[5] G. Zhou, S. Lai, K. Liu, and J. Zhao, "Topic-sensitive probabilistic model for expert fnding in question answer communities," CIKM, pp.1662–1666, 2012.
[6] T.M. Lai, T. Bui, and S. Li, "A review on deep learning techniques applied to answer selection," International Conference on Computational Linguistics, pp.2132–2144, 2018.
[7] B. Steven, K. Ewan, and L. Edward, Natural Language Processing with Python, O'Reilly, 2010.
[8] X. Hu and X. Huang, "Solving 0-1 knapsack problem based on ant colony optimazation algorithm," Journal of Systems Engineering, vol.20, no.5, pp.520–523, 2005.
[9] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," Proceedings of the Conference of the European Association of Computational Linguistics, pp.567–575, Athens, Greece, 2009.
[10] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, vol.3, pp.993–1022, 2012.
[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," Advances in Neural Information Processing Systems, vol.26, pp.3111–3119, 2013.
[12] K.-T. Fang, D.K.J. Lin, P. Winker, and Y. Zhang, "Uniform design: Theory and application," Techonometrics, vol.42, no.3, pp.237–248, 2000.