Reliable Medical Diagnosis from Crowdsourcing: Discover Trustworthy Answers from Non-Experts

Yaliang Li[1], Nan Du[2], Chaochun Liu[2], Yusheng Xie[2],
Wei Fan[2], Qi Li[1], Jing Gao[1], and Huan Sun[3]
[1]State University of New York at Buffalo, Buffalo, NY USA
[2]Baidu Research Big Data Lab, Sunnyvale, CA USA
[3]The Ohio State University, Columbus, OH USA
[1]{yaliangl, qli22, jing}@buffalo.edu
[2]{dunan, liuchaochun, xieyusheng, fanwei03}@baidu.com
[3]sun.397@osu.edu

ABSTRACT

Nowadays, increasingly more people are receiving medical diagnoses from healthcare-related question answering platforms as people can get diagnoses quickly and conveniently. However, such diagnoses from non-expert crowdsourcing users are noisy or even wrong due to the lack of medical domain knowledge, which can cause serious consequences. To unleash the power of crowdsourcing on healthcare question answering, it is important to identify trustworthy answers and filter out noisy ones from user-generated data. Truth discovery methods estimate user reliability degrees and infer trustworthy information simultaneously, and thus these methods can be adopted to discover trustworthy diagnoses from crowdsourced answers. However, existing truth discovery methods do not take into account the rich semantic meanings of the answers. In the light of this challenge, we propose a method to automatically capture the semantic meanings of answers, where answers are represented as real-valued vectors in the semantic space. To learn such vector representations from noisy user-generated data, we tightly combine the truth discovery and vector learning processes. In this way, the learned vector representations enable truth discovery method to model the semantic relations among answers, and the information trustworthiness inferred by truth discovery can help the procedure of vector representation learning. To demonstrate the effectiveness of the proposed method, we collect a large-scale real-world dataset that involves $219,527$ medical diagnosis questions and $23,657$ non-expert users. Experimental results show that the proposed method improves the accuracy of identified trustworthy answers due to the successful consideration of answers' semantic meanings. Further, we demonstrate the fast convergence and good scalability of the proposed method, which makes it practical for real-world applications.

1. INTRODUCTION

Recent years have witnessed the successful applications of crowdsourcing in many domains, including but not limited to data annotation, stock prediction, and online education. Crowdsourcing has also entered healthcare, a domain that requires years of training, high specialization, and rigorous regulation. Instead of time-consuming visits to physicians, nowadays, increasingly more people are receiving medical suggestions or even diagnoses from online healthcare communities, such as medhelp.org and healthboards.com.

In these online crowdsourcing communities, patients post questions to describe their symptoms, and crowd users will give possible answers based on their experience and knowledge. However, due to the lack of sufficient domain knowledge, the answers provided by the crowd can be different from each other. The noisy or even wrong answers from the crowd can cause serious consequences to patients, which limits the positive impact of crowdsourcing on healthcare. To unleash the power of crowdsourcing diagnosis, it is important to automatically infer information trustworthiness and discover trustworthy answers from non-expert users.

Truth Discovery. Recently, truth discovery methods [5, 8, 16, 18, 25, 34, 37] have been proposed to simultaneously estimate the user reliability and infer trustworthy answers from noisy user-generated data. Various truth discovery methods have been developed based on the following general principle: The users who provide trustworthy answers are more reliable, and the answers from reliable users are more trustworthy. This principle tightly combines the process of user reliability estimation with that of information trustworthiness inference, and thus enables truth discovery methods to find trustworthy answers without any supervision. Due to this advantage, truth discovery methods have been successfully applied in various domains such as crowd sensing [3, 29, 30], knowledge base construction [9, 10], and information extraction [14, 24, 35]. The success of these applications demonstrates the ability of truth discovery methods to distill trustworthy information from noisy user-generated data.

Challenge of Capturing Semantic Meanings. However, most of the existing truth discovery methods treat different answers from different users as categorical data, and they do not consider the semantic meanings of answers. This limitation may prevent existing methods identifying correct answers from the crowd on health-related questions. Consider the following example: For a specific question, user-1 claims that the patient might have *sinus infection* and user-2 suggests that the possible disease might be *bone fracture*, while the true disease that the patient has is *common cold*. Most of truth discovery methods treat these three answers as three unrelated ones, and some truth discovery methods [8, 11] even assume that when a user gives an answer of *sinus infection*, he/she

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018688

is against other possible answers including *common cold*. However, *sinus infection* and *common cold* are highly related and they support each other.

By considering the semantic meanings of candidate answers, we can estimate user reliability more accurately during truth discovery. In the aforementioned example, although user-1 does not provide the exact answer to that question, his answer is close to the true answer. Thus user-1 should receive a small penalty on his reliability estimation due to this wrong answer. On the other hand, the semantic meaning of the wrong answer provided by user-2 is far from that of the true answer. Thus user-2 should receive a relative big penalty on his reliability estimation due to this wrong answer.

In order to capture the semantic meanings of possible answers, we propose to represent candidate answers (possible diseases) as real-valued vectors. Such vector representations enable us to calculate the semantic closeness among different answers. It is important not only to know whether a user provides a wrong answer or not, but also to distinguish how "big" the mistake is. Then during user reliability estimation, we can assign appropriate penalties to different users when they provide wrong answers.

Representation Learning. Inspired by the idea of word embedding [22], we learn the vector representation of candidate answers (possible diseases) without any supervision. The key idea is that if two answers often co-occur with other similar words in a corpus (such co-occurred words are named as context words), the vector representations for these two answers should be similar. In the medical diagnosis scenario, the words in the corresponding question texts can be regarded as the context words of the candidate answers. For example, the candidate answers *sinus infection* and *common cold* often co-occur with the similar words in questions such as *runny nose* and *headache*, so the learned vector representations for these two answers should be similar.

However, the quality of user-generated data poses another challenge. Some users may provide irrelevant answers to questions. Consider this example: When the description from a patient is about *runny nose*, a noisy answer provided by a user might be *bone fracture*. In this case, it might not be reasonable to regard *runny nose* as the context words of *bone fracture*. Such unreasonable co-occurrences are caused by the noisiness of user-generated data.

Fortunately, incorporating truth discovery into this process can help. Truth discovery methods estimate the information trustworthiness of user-generated data, and such estimated trustworthiness degrees can help the procedure of vector representation learning. Thus different from word embedding process, the vector representations in the proposed method are learned based on the "weighted" co-occurrence information. In this way, the truth discovery principle and vector representation learning are tightly combined with each other, and they will be iteratively enhanced by each other. The overview of the proposed method is illustrated in Figure 1.

Experimental Results. To demonstrate the effectiveness of the proposed method, we evaluate its performance in the challenging medical diagnosis task: discovering trustworthy answers from non-expert users on crowdsourcing medical question answering websites. As mentioned before, non-expert users provide their answers to medical diagnosis questions based on their own experience and knowledge. Due to the lack of sufficient domain knowledge, noisy or even wrong information is unavoidable. We apply the proposed method on these noisy user-generated data to distill professional diagnoses. The collected dataset contains $219,527$ medical diagnosis questions, and $23,657$ non-expert crowdsourcing participants are involved. Experimental results show that the proposed method outperforms other truth discovery methods by capturing the semantic meanings of answers and question texts. We also conduct experiments to illustrate the accurateness of the estimated user reliability and the effectiveness of the learned vector representations. The fast convergence and good scalability of the proposed method are further demonstrated.

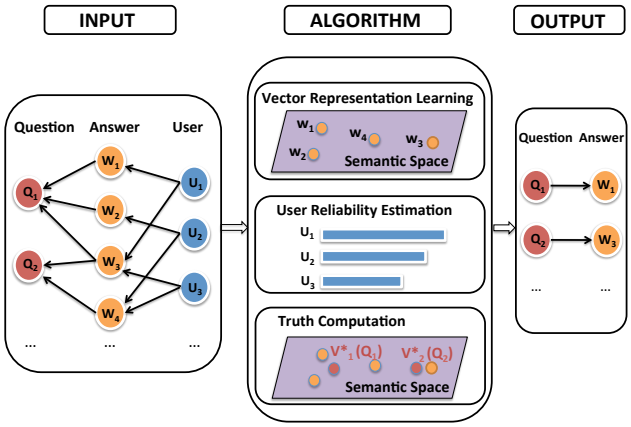

Figure 1: The Overview of the Proposed Method.

Broader Impact. Although the proposed method is designed to discover trustworthy answers from non-expert users in crowdsourcing diagnosis, it can be applied to many other applications. Generally speaking, the proposed method provides solutions to automatically discover the semantic meanings of answers and incorporate such semantic meanings into truth discovery. These problems become urgent as increasingly more truth discovery applications are dealing with raw textual data [9,14,24,35] in which semantic meanings need to be recognized. We demonstrate that the performance of truth discovery on textual data can be further improved by incorporating semantic meanings. To the best of our knowledge, the proposed method is the first one to consider the semantic meanings of answers and question texts when conducting truth discovery procedure.

2. METHODOLOGY

In this section, we first formally define the crowdsourcing medical diagnosis task. In order to discover trustworthy answers from the noisy user-generated data, we formulate the general principle of truth discovery and the idea of vector representation learning in a unified optimization problem. After that, we present an iterative solution to solve this optimization problem, and discuss several practical issues.

2.1 Problem Definition

Suppose we are interested in M medical diagnosis questions, in each of which the patient describes his/her symptoms and other relevant information such as age and gender. Thus each question m is associated with a description text, i.e., a set of words $\mathcal{C}_m$. For these diagnosis questions, there are N users providing answers to them. Let w_m^n be the answer from the n-th user to the m-th question. Since we only care about the diagnoses for these questions, here we remove the irrelevant information in the answers and assume that each answer w_m^n contains only a diagnosed disease, which is a single word or a phrase. We will provide more details about the dataset in Section 3.1.1.

In most application scenarios, not all the users will provide answers to all the questions. Let $\mathcal{N}_m$ denote the set of users who

provide answers to the m-th question. Correspondingly, $\mathcal{M}_n$ denotes the set of questions that the n-th user provides answers to.

For the m-th question, different users provide different answers, and these user-generated answers can be conflicting with each other. Compared with the simple voting approach, truth discovery methods estimate a reliability degree r_n for each user n and conduct weighted aggregation to resolve conflicts. The reliability degree reflects the probability of a user providing trustworthy information. A higher reliability degree r_n indicates that the corresponding user has a higher probability to provide trustworthy information than the users with lower reliability degrees. Based on this assumption, truth discovery methods resolve conflicts and find trustworthy answers from noisy data.

As discussed in the previous section, we learn a real-valued vector representation $\mathbf{v}_w$ for each answer w which is a candidate disease that the patient might have. In order to learn such vector representation, we treat the words in the corresponding question text $c \in \mathcal{C}_m$ as the context words of the answer w.

For convenience, Table 1 summarizes the notations used in this paper; some will be introduced later.

Table 1: Notations

Notation	Definition
$\mathcal{C}_m$	set of words in the m-th question
$\mathcal{N}_m$	set of users who provide answers to question m
$\mathcal{M}_n$	set of questions that user n provides answers to
w_m^n	answer from the n-th user to the m-th question
$\mathbf{v}_w$	real-valued vector representation of answer w
r_n	reliability degree of the n-th user
$\mathbf{v}_m^*$	the semantic truth vector for the m-th question

2.2 Proposed Method

As discussed in Section 1, when applying truth discovery methods to find the trustworthy answers to medical diagnosis questions, the semantic meanings of answers should be taken into account in order to accurately estimate user reliability degrees. To achieve this goal, we learn real-valued vector representations for answers. Similar to word embedding, the key idea of vector representation learning is that the answers that share many co-occurred context words should have similar learned vector representations. However, unreasonable co-occurrences might be observed due to the noisy nature of user-generated data. To tackle this challenge, we propose to learn the vector representation based on the "weighted" co-occurrence information, where the weight is derived from the trustworthiness analysis results of truth discovery. From these observations, we can see that the truth discovery procedure and the vector learning procedure rely on each other, and they should be tightly combined together.

Motivated by the above observations, we formulate the general principle of truth discovery and the idea of vector representation learning in a unified objective function:

$$\min_{\{\mathbf{v}_m^*\},\{r_n\},\{\mathbf{v}_w\}} \sum_{n=1}^{N} r_n \cdot \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \left(d(\mathbf{v}_{w_m^n}, \mathbf{v}_m^*) + \alpha \sum_{c \in \mathcal{C}_m} d(\mathbf{v}_{w_m^n}, \mathbf{v}_c) \right)$$
$$\text{s.t.} \quad f(\{r_n\}) = 1, \quad r_n \in \mathbb{R}^+. \tag{1}$$

In this optimization problem, $\mathbf{v}_w$ is the real-valued vector representation of answer w, and $\mathbf{v}_m^*$ is the semantic truth vector for the m-th question. Function $d(\cdot)$ measures the distance between two real-valued vectors, and we adopt the square loss function as an instantiation of $d(\cdot)$. $\mathcal{C}_m$ contains the context words in the m-th question, $\mathbf{v}_c$ is the vector representation of context word c, and parameter α can be used to adjust the importance of context words. $\mathbb{R}^+$ is the domain of user reliability degree, and $f(\cdot)$ is a regularization function on user reliability degrees.

The intuitions behind this objective function are as follows: (1) The proposed objective function minimizes the weighted distances between the answers from users and the identified trustworthy answers (i.e., truths). If a user has a high-reliability degree, the identified truth should be close to the answers from this user in order to minimize the overall distances. This follows the general principle of truth discovery that the answers from highly reliable users should be trustworthy. (2) As shown in the reliability estimation discussion in Section 2.2.2, the estimated user reliability degree r_n is determined by the distance between the answers from this user and corresponding identified truths in the solution to this optimization problem. If a user provides answers that are close to the truths, this user should be assigned a high-reliability degree. This is also consistent with the general principle of truth discovery that the users who provide trustworthy answers should be reliable. Note that not all the users provide answers to all the questions, and thus we minimize the average error of each user (i.e., by adding the term $\frac{1}{|\mathcal{M}_n|}$). Otherwise, the users who provide answers to a small number of questions will receive high-reliability degrees, which does not make sense. (3) In the proposed formulation, the answers $\{w_m^n\}$ are represented as real-valued vectors $\{\mathbf{v}_{w_m^n}\}$. Such vector representations enable the proposed method to measure the semantic distance among different answers. (4) To automatically learn the vector representations for answers, the question texts are utilized. The words in each question text are regarded as the context words for the corresponding answers. The key idea of the vector representation learning component is inspired by word embedding [22]: the answers that share many similar context words will have similar vector representations. This key idea is formulated as the term $\sum_{c \in \mathcal{C}_m} d(\mathbf{v}_w, \mathbf{v}_c)$. For two answers w and w', if they share many similar context words, in order to minimize the loss, their learned vector representations $\mathbf{v}_w$ and $\mathbf{v}_{w'}$ will be similar.

In the above objective function, we have three sets of variables: semantic truth vectors for questions $\{\mathbf{v}_m^*\}$, user reliability $\{r_n\}$, and real-valued vector representations for answers $\{\mathbf{v}_m\}$. We adopt block coordinate descent approach [6] to solve this optimization problem, which leads to an iterative solution consisting of the following three components.

2.2.1 Truth Computation

In this step, user reliability $\{r_n\}$ and vector representations $\{\mathbf{v}_w\}$ are fixed, and we solve for the semantic truth vector $\{\mathbf{v}_m^*\}$. Thus the original optimization problem (Eq. (1)) becomes the following one:

$$\min_{\{\mathbf{v}_m^*\}} \sum_{n=1}^{N} r_n \cdot \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \left(d(\mathbf{v}_{w_m^n}, \mathbf{v}_m^*) + b_m^n \right),$$

where constant $b_m^n = \alpha \cdot \sum_{c \in \mathcal{C}_m} d(\mathbf{v}_{w_m^n}, \mathbf{v}_c)$. This optimization is equivalent to the following one:

$$\min_{\{\mathbf{v}_m^*\}} \sum_{m=1}^{M} \sum_{n \in \mathcal{N}_m} \frac{r_n}{|\mathcal{M}_n|} \cdot d(\mathbf{v}_{w_m^n}, \mathbf{v}_m^*). \tag{2}$$

Eq. (2) can be further split into M separate optimization problems, and each one is associated with a question. The optimal $\mathbf{v}_m^*$ for

each question is the weighted mean of all the available answers:

$$\mathbf{v}_m^* = \frac{\sum_{n \in \mathcal{N}_m} \frac{r_n}{|\mathcal{M}_n|} \cdot \mathbf{v}_{w_m^n}}{\sum_{n \in \mathcal{N}_m} \frac{r_n}{|\mathcal{M}_n|}}. \quad (3)$$

From this truth computation formula, we can see that: (1) It is consistent with the general idea of truth discovery: the answers from reliable users are regarded as more trustworthy. The semantic truths $\{\mathbf{v}_m^*\}$ are affected more by the users with high-reliability degrees, while the effect of non-reliable users is small as their r_n's are low. The reliability degree of a user is evenly "invested" on all the questions he/she provides answers to. (2) This truth computation method also allows the proposed method to utilize the semantic relations among different answers. For example, when *common cold* and *sinus infection* are two candidate answers to the same question, existing truth discovery methods treat them separately and do not consider their similar semantic meanings. In contrast, the proposed method can learn a conceptual truth vector that is supported by both *common cold* and *sinus infection* as they are close to each other.

The computed semantic truth vectors indicate the conceptual truths in the semantic space, and we need to project the computed truth vectors to answers again. For each question, we can compute the distance between the identified truth vector $\mathbf{v}_m^*$ and all the possible candidate answers $\{\mathbf{v}_{w_m^n}\}_{n \in \mathcal{N}_m}$, and then select the top candidate answer(s) that are close to the semantic truth vector. This brings another advantage to the proposed method: it can deal with both single-truth and multiple-truth scenarios. In the community of truth discovery, people have different assumptions about the identified truths. Some work [8, 16, 34] assumes that there is one and only one truth for each question (single-truth scenario), while others [26, 37] assume that there might be multiple truths for each question (multiple-truth scenario). The proposed method computes conceptual semantic truth vectors, which can be considered as "soft" truths. Thus the proposed method can handle both single-truth and multiple-truth scenarios by selecting top candidate answer(s) that are close to the semantic truth vectors.

2.2.2 Reliability Estimation

In this step, vector representations $\{\mathbf{v}_w\}$ and semantic truth vectors $\{\mathbf{v}_m^*\}$ are fixed, and we solve for the user reliability $\{r_n\}$. The original optimization problem (Eq. (1)) becomes the following one:

$$\min_{\{r_n\}} \quad \sum_{n=1}^{N} r_n \cdot e_n$$
$$\text{s.t.} \quad f(\{r_n\}) = 1, \quad r_n \in \mathbb{R}^+, \quad (4)$$

where the average error of the n-th user is a constant term $e_n = \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \Big(d(\mathbf{v}_{w_m^n}, \mathbf{v}_m^*) + \alpha \sum_{c \in \mathcal{C}_m} d(\mathbf{v}_{w_m^n}, \mathbf{v}_c) \Big)$. The concrete solution to Eq. (4) depends on the adopted constraint function $f(\cdot)$. If we adopt L^1-norm or L^2-norm, some trivial solutions exist: The loss can be minimized by setting the reliability degree of the user who has smallest e_n to be 1 while setting the reliability degrees of other users to be 0. To avoid such trivial solutions and achieve meaningful estimated user reliability, we investigate the intuitions behind the equation. As $r_n \in \mathbb{R}^+$, the term $\exp(-r_n) \in (0, 1)$. The bigger user reliability r_n, the smaller value $\exp(-r_n)$. Thus $\exp(-r_n)$ can be treated as the probability of user n providing wrong answers. Therefore, we propose to adopt the following constraint function: $f(\{r_n\}) = \sum_{n=1}^{N} \exp(-r_n) = 1$. On the other hand, the normalized error term $\frac{e_n}{\sum_{n=1}^{N} e_n}$ can be treated as the observed probability of user n providing wrong information. In this step, as all e_n's are constant, the term $\sum_{n=1}^{N} e_n$ is also a constant. Thus Eq. (4) is equivalent to the following optimization problem:

$$\min_{\{r_n\}} \quad \sum_{n=1}^{N} -\log\left(\exp(-r_n)\right) \cdot \frac{e_n}{\sum_{n=1}^{N} e_n}$$
$$\text{s.t.} \quad \sum_{n=1}^{N} \exp(-r_n) = 1, \quad r_n \in \mathbb{R}^+, \quad (5)$$

which is the cross entropy between the estimated user probabilities of providing wrong information (i.e., term $\exp(-r_n)$) and the observed user probabilities of providing wrong information (i.e., term $\frac{e_n}{\sum_{n=1}^{N} e_n}$). The optimal solution is achieved when these two distributions are identical, that is, $\exp(-r_n) = \frac{e_n}{\sum_{n=1}^{N} e_n}$. Thus we get the optimal user reliability estimation $r_n = -\log(\frac{e_n}{\sum_{n=1}^{N} e_n})$.

More specifically, the reliability score of a particular user is inversely proportional to the errors this user makes:

$$r_n \propto \frac{1}{\frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \Big(d(\mathbf{v}_{w_m^n}, \mathbf{v}_m^*) + \alpha \sum_{c \in \mathcal{C}_m} d(\mathbf{v}_{w_m^n}, \mathbf{v}_c) \Big)}.$$

Two kinds of errors contribute to the user reliability estimation: (1) Similar to the general truth discovery framework, the user reliability is estimated based on the distance between user's answers and the corresponding truths. If a user provides answers that are far from the identified trustworthy information (truth vectors), this user will be assigned a low-reliability score. On the other hand, if a user often provides trustworthy answers, the calculated distance between his answers and the identified truths will be small. Correspondingly, a high-reliability score will be assigned to this user. (2) In the proposed method, the semantic distance between the user's answers and the corresponding question texts (i.e., context words $\mathcal{C}_m$) also impacts the estimation of user reliability degree. For example, when the health condition description from a patient is about *runny nose*, the user who provides answer *bone fracture* will be assigned a low estimated reliability degree. This allows the proposed method to utilize the question texts and capture the semantic relations between question texts and answers.

2.2.3 Vector Representation Learning

In this step, user reliability $\{r_n\}$ and semantic truth vectors $\{\mathbf{v}_m^*\}$ are fixed, and we solve for the vector representations $\{\mathbf{v}_w\}$. The original optimization problem (Eq. (1)) becomes the following one:

$$\min_{\{\mathbf{v}_w\}} \sum_{n=1}^{N} \frac{r_n}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \left(d(\mathbf{v}_{w_m^n}, \mathbf{v}_m^*) + \alpha \sum_{c \in \mathcal{C}_m} d(\mathbf{v}_{w_m^n}, \mathbf{v}_c) \right). \quad (6)$$

This optimization problem involves the whole set of vector representations for answers, and it can be split into several independent optimization problems in each of which only one answer w is involved. Thus Eq. (6) can be rewritten in terms of each possible answer w. However, each possible answer w might appear in different questions, and it also can be claimed by different users. Let $\mathcal{D}_w = \{\langle m, n \rangle | w_m^n = w\}$ represent the set in which each $\langle m, n \rangle$ pair denotes that w_m^n is the answer w. The optimization problem in terms of w is the following:

$$\min_{\mathbf{v}_w} \sum_{\langle m,n \rangle \in \mathcal{D}_w} \frac{r_n}{|\mathcal{M}_n|} \left(d(\mathbf{v}_{w_m^n}, \mathbf{v}_m^*) + \alpha \sum_{c \in \mathcal{C}_m} d(\mathbf{v}_{w_m^n}, \mathbf{v}_c) \right). \quad (7)$$

By solving Eq. (7), we get:

$$\mathbf{v}_{w_m^n} = \frac{\sum_{\langle m,n\rangle\in\mathcal{D}_w}\left(\frac{r_n}{|\mathcal{M}_n|}\cdot\mathbf{v}_m^* + \sum_{c\in\mathcal{C}_m}\frac{r_n}{|\mathcal{M}_n|}\cdot\alpha\cdot\mathbf{v}_c\right)}{\sum_{\langle m,n\rangle\in\mathcal{D}_w}\left(\frac{r_n}{|\mathcal{M}_n|} + \sum_{c\in\mathcal{C}_m}\frac{r_n}{|\mathcal{M}_n|}\cdot\alpha\right)}. \tag{8}$$

From this update equation, we can see: (1) Similar to word embedding [7, 22, 23], if two answers share many similar context words, the real-valued vector representations for these two answers will be similar. For example, as *common cold* and *sinus infection* co-occur with many similar symptoms, the distance between the vector representations for these two answers will be small. In contrast, the learned vectors for *common cold* and *bone fracture* should be far from each other as these two answers share very few context words. This shows that the proposed method can successfully capture the semantic meanings of answers by utilizing their corresponding question texts. (2) Different from word embedding, the vector representations in the proposed method are learned based on the "weighted" co-occurrence information. We observe some unreasonable co-occurrences (such as *bone fracture* and *runny nose*), which is caused by the noisy nature of user-generated data. Truth discovery methods estimate the information trustworthiness of user-generated data, and thus such trustworthiness degree information can help the procedure of word vector learning. This is why we learn vector representations and conduct truth discovery simultaneously. In this way, the general truth discovery principle and the idea of vector representation learning are tightly coupled with each other. (3) Besides the effect of context words, the learned real-valued vector representations are also affected by the computed semantic truths. If a user is reliable, then the answer from this user is trustworthy. Thus the vector representation of the corresponding answer should be close to the computed semantic truth. In other words, the computed truth vectors are also considered as context words of the corresponding answers, and they will also contribute to the weighted co-occurrence information for vector representation learning.

2.2.4 Practical Issues

So far, we have derived the whole solution for the overall objective function (Eq. (1)). As the solution is an iterative procedure, we need an initialization method and a stop criterion.

To start the iterative procedure, we initialize the user reliability degree $\{r_n\}$ and the real-valued vector representations of answers $\{\mathbf{v}_m\}$. For user reliability degrees, we can initialize them with different strategies. If any prior knowledge about the user reliability is available, we can choose different initialized values for different users accordingly. Otherwise, uniform reliability degree initialization might be a good choice. According to Eq. (8), we can initialize the vector representations of answers $\{\mathbf{v}_m\}$ only based on the context words as follows:

$$\begin{aligned}\mathbf{v}_{w_m^n} &= \frac{\sum_{\langle m,n\rangle\in\mathcal{D}_w}\sum_{c\in\mathcal{C}_m}\frac{r_n}{|\mathcal{M}_n|}\cdot\alpha\cdot\mathbf{v}_c}{\sum_{\langle m,n\rangle\in\mathcal{D}_w}\sum_{c\in\mathcal{C}_m}\frac{r_n}{|\mathcal{M}_n|}\cdot\alpha}\\ &= \frac{\sum_{\langle m,n\rangle\in\mathcal{D}_w}\sum_{c\in\mathcal{C}_m}\frac{r_n}{|\mathcal{M}_n|}\cdot\mathbf{v}_c}{\sum_{\langle m,n\rangle\in\mathcal{D}_w}\sum_{c\in\mathcal{C}_m}\frac{r_n}{|\mathcal{M}_n|}}.\end{aligned} \tag{9}$$

Compared with Eq. (8), we remove the effect of semantic truth vectors on the initialization of vector representations for answers.

After initialization, the above three components, truth computation, reliability estimation and vector representation learning, will be conducted iteratively. To terminate the iterative procedure, several stop criteria can be adopted. For example, we can check whether the decrease in objective function (Eq. (1)) is small enough compared with the previous iteration, or we can check whether the estimated user reliability converges. In the next section, we will experimentally show that the proposed method converges quickly within a small number of iterations.

3. EXPERIMENT

In this section, we experimentally validate the performance of the proposed method on the medical diagnosis task from the following aspects: (1) We compare the accuracy of the proposed method with several baseline methods, and demonstrate the necessity of considering the semantic meanings of answers. (2) Comparisons on the estimated user reliability degrees also confirm the proposed method's advantage of capturing semantic relations among answers. (3) We demonstrate the effectiveness of the learned vector representations for answers. (4) The effect of the introduced parameter α is further studied. (5) Last but not the least, we show the fast convergence and good scalability of the proposed method.

3.1 Experiment Setup

3.1.1 Data collection

We collect a real-world dataset from *baobaozhidao*[1], a popular crowdsourcing platform for maternal and child health. On this crowdsourcing healthcare platform, pregnant women post health-related questions about themselves or their babies, and non-expert users on the platform give answers based on their own experience.

The crawled questions and answers are in Chinese. Different from English, Chinese strings are not divided by word delimiter. Thus word segmentation is required, where the Chinese texts are cut into Chinese component words. We segment the crawled raw texts into words using a Chinese word segmentation package [1]. Then based on an available entity dictionary, we extract medical entities from these segmented texts. For question texts, only health-related entities such as symptoms, age and gender are kept as context words. The motivation behind this pre-processing is similar to the idea of subsampling in word embedding [22]. By doing so, we can remove some high-frequency words such as "I" and "is" as these words are less informative. This pre-processing can result in a significant speedup, and improves the accuracy of the learned vector representations. For answer texts, the possible diseases contained in answers are extracted. As we focus on medical diagnosis question, the answers that do not contain any related information are filtered out. After conducting these data processing steps, we have $1,053,726$ question-answer pairs, including $219,527$ medical diagnosis questions and $23,657$ non-expert users who provide answers to these questions.

To learn the representation of context words, we use all the available question texts. Skip-gram architecture in Word2vec package [2] is adopted to train vector representations for context words. The dimensionality of the learned vectors is set to be 100, the context window size is set to be 8, and the minimum occurrence count is set to be 5. For more details, please refer to [22].

For the purpose of evaluation, we ask real doctors (experts on maternal and child health) to provide professional diagnosis results based on question texts. Each question is judged by one doctor, and finally we get answers to $13,992$ questions from 42 doctors. These annotated questions will be used to evaluate the performance of the proposed method and baselines.

3.1.2 Compared Methods

We adopt some widely-used truth discovery methods as baselines: TruthFinder [34] and AccuSim [8, 17] design the ways of es-

[1]http://baobao.baidu.com/

timating user reliability degrees based on Bayesian analysis. These two methods also incorporate "implication" functions that can consider the influence of other candidate answers to a particular answer. However, such implication functions in both TruthFinder and AccuSim are assumed to be known before conducting truth discovery. Investment [25] assumes that a user "invests" his reliability degree among the answers he provides, and then collects credits back from the answers he supports. CRH [16] is a truth discovery framework that can estimate user reliability on heterogeneous data. Besides the above truth discovery methods, we also implement the simple voting approach, which does not estimate user reliability and simply takes the majority answers as truths.

For each baseline, we implement it and set its parameters according to the original paper. All the methods are implemented using MATLAB, and run on a machine with 8G RAM, 2.7 GHz Intel Core i5 processor.

3.2 Performance Comparison

As mentioned above, for each medical diagnosis question, several non-expert users provide answers based on their own experience and knowledge. Due to the insufficient medical domain knowledge, the answers might be noisy or even wrong. Our task is to identify the trustworthy answers from non-expert users. The groundtruth answers are provided by domain experts for the purpose of performance evaluation. For the annotated questions, we compare the identified answers outputted by each method with the groundtruth answers, and compute the corresponding accuracy. Table 2 lists the results of all the methods.

Table 2: Accuracy Comparison

Method	#Correct Answers	Accuracy
Voting	9631	68.83%
TruthFinder	9561	68.33%
AccuSim	9566	68.37%
Investment	9877	70.59%
CRH	9896	70.73%
Proposed Method	**10408**	**74.39%**

From Table 2, we can see that the accuracy of simple voting approach is not high as this method does not estimate user reliability degrees and cannot distinguish reliable users from unreliable users. For TruthFinder and AccuSim, the unsatisfactory performance is caused by the setting of implication functions. Although implication function is incorporated to capture the influences among answers, TruthFinder and AccuSim do not provide a concrete solution to set these functions. Investment and CRH methods give better accuracy than simple voting as they estimate user reliability and conduct weighted aggregation to find trustworthy answers. The proposed method further improves the performance as it considers the semantic meanings among answers. As discussed in Section 1, the consideration of semantic meanings leads to a more accurate estimation of user reliability degrees. Thus the proposed method achieves the best performance.

We further analyze the performance of the proposed method. As the proposed method learns real-valued vector representations for answers, it enables us to evaluate its performance in terms of another metric besides accuracy: the Euclidean distances between the identified answers and the groundtruth answers. For the questions that the proposed method gets correct answers, the Euclidean distances will be zero. For the purpose of better illustration, we exclude these questions as the number of them is quite large. Thus

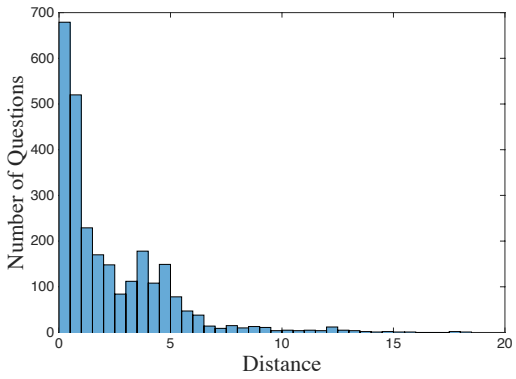

Figure 2: The distribution of Euclidean distances between the identified answers and the groundtruth answers.

Figure 2 plots the distribution of the Euclidean distance between the identified answers and groundtruth answers for the questions that the proposed method gives wrong answers. From this figure, we can see that the distribution of Euclidean distance follows a long-tail distribution, in which most of the distances are small and only a few distances are large. Especially, note that for around 700 questions, the identified answers outputted by the proposed method are very close to the groundtruth answers (the distance is less than 0.5). This indicates that for these questions, the proposed method gives similar answers to the groundtruth answers.

3.3 Evaluation of the Estimated User Reliability Degrees

The above experimental results show that the proposed method outperforms other baseline methods. Even for the questions that the proposed method gives wrong answers, most of its identified answers are very close to the groundtruth answers. These improvements are brought by the consideration of semantic meanings among answers and question texts, which enables a better estimation of user reliability degrees. Thus in this section, we compare the estimated reliability degrees from baseline methods and the proposed method.

As the number of involved users is quite large, it is impossible to visually compare the estimated reliability degrees of individual users. Instead, we compute some metrics to quantitatively compare the estimated reliability degrees of all the users. For each user, we compare his provided answers with corresponding groundtruth answers, and calculate his error rate. Then we compute the Pearson's correlation coefficient between users' error rates and their estimated reliability degrees. Note that Pearson's correlation coefficient is a commonly used metric to test linear relationship between variables. The closer it is to -1, the stronger negative linear relationship the variables have (note that users' reliability degrees and their corresponding error rates are negatively correlated). The results for all the methods are shown in the second column of Table 3. Compared with other truth discovery methods, it seems that the proposed method does not give better user reliability estimation.

However, for the medical diagnosis scenario, user error rate may not be an appropriate metric to measure the quality of users' information as it only computes how many errors each user makes and does not consider how big the errors are. To correct this, we adopt the Euclidean distance to measure users' errors, and then compute the Pearson's correlation coefficient between the estimated user reliability degree and users' average errors. The results are listed in

Table 3: Comparison on User Reliability

Method	Pearson's correlation (Error Rate)	Pearson's correlation (Euclidean Distance)
Voting	NA	NA
TruthFinder	-0.2725	-0.1952
AccuSim	-0.2986	-0.1941
Investment	-0.1643	-0.1036
CRH	**-0.3007**	-0.1966
Proposed Method	-0.2638	**-0.3229**

the third column of Table 3. Now we can see that the estimated reliability degrees of the proposed method are significantly better than other truth discovery methods due to the fact that the proposed method takes into accounts the semantic meanings of answers. By doing so, the proposed method knows how big the error is when a wrong answer is provided. This is the reason that the proposed method gives better performance than other baselines.

3.4 Case Studies of the Learned Vector Representation for Answers

In order to know the penalty each wrong answer should be given, the proposed method calculates the semantic distances between answers by automatically learning real-valued vector representations for answers. In the above section, we investigate the accurateness of the estimated user reliability degrees, and here, we check the effectiveness of the learned vector representations.

For each answer word w, we can calculate the semantic distance between w and any other answer word w'. Then we rank all the other answer words by their corresponding distances to w. Tables 4 and 5 show the top answer words that are close to *common cold* and *enteritis* respectively. As the collected dataset is in Chinese, we also provide corresponding English translations.

Table 4: The answer words that have similar learned vectors to *common cold*.

Answer Word	Answer Word (Translation)	Distance
伤风	*mild common cold*	0.0043
风寒	*cold*	0.0056
上呼吸道感染	*upper respiratory infection*	0.0142
呼吸道感染	*respiratory infection*	0.0209
急性气管炎	*acute bronchitis*	0.0233
鼻炎	*rhinitis*	0.0287
流感	*flu*	0.0314

Table 5: The answer words that have similar learned vectors to *enteritis*.

Answer Word	Answer Word (Translation)	Distance
拉肚子	*diarrhea*	0.0150
肠道菌群失调	*enteric flora disturbance*	0.0214
急性腹泻	*acute diarrhea*	0.0375
消化不良	*functional dyspepsia*	0.0412
病毒性肠炎	*viral enteritis*	0.0416
痢疾	*dysentery*	0.0495
肠胃炎	*stomach flu*	0.0906

From Table 4, we can observe that the top ranked answers have very close semantic meanings to the answer *common cold* as they are about *cold*, *acute bronchitis*, *rhinitis*, etc. From Table 5, similar results are observed as the top ranked answers are about *diarrhea*, *viral enteritis* and *stomach flu* that are very close to *enteritis*. These case studies show that the proposed method can automatically learn meaningful vector representations for answers.

3.5 Parameter Sensitivity

In the proposed method, a parameter α is introduced to adjust the importance of context words. Here we experimentally test the sensitivity of the parameter α. We set different values for this parameter, and report the corresponding performance (accuracy) in Figure 3.

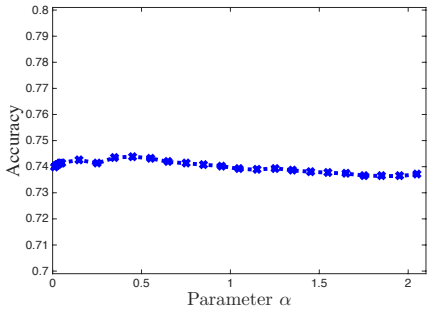

Figure 3: Performance w.r.t. Parameter α

From Figure 3, we can observe that the accuracy of the proposed method is in the range of $[0.7364, 0.7439]$ when we vary the parameter α, showing that the proposed method is not very sensitive to α. Meanwhile, we also observe that parameter α can balance the effect of the general principle of truth discovery and the context words. When the parameter α is small (less than 0.4 in our experiments), the identified answers are mainly decided by the general principle of truth discovery. By slightly increasing the parameter α, the proposed method also incorporates the effect of context words. When the value of the parameter α is large (bigger than 0.6 in our experiments), the identified answers rely more on the context words, which causes a small drop in the accuracy.

3.6 Convergence Study

As presented in Section 2, the proposed method is an iterative procedure. Here we experimentally study the convergence rate of the proposed method. Among the three sets of variables, the estimated user reliability is the key part. Once the estimated user reliability converges, the learned vector representations and the identified answers become stable. Thus we record the change of estimated user reliability, and plot it with respect to the number of iterations in Figure 4. We can observe that the estimated user reliability becomes stable within 10 iterations, showing that the proposed method converges very quickly in practice.

3.7 Scalability Study

Last but not least, we further study the scalability of the proposed method. In order to conveniently vary the scale of dataset, we simulate different number of users by randomly generating their provided answers, which gives the number of question-answer (Q&A) pairs in a range of 10^3 to 10^7. The running time of the proposed method on different sets of Q&A pairs is plotted in Figure 5. From this set of experiments, we can conclude that the running time of

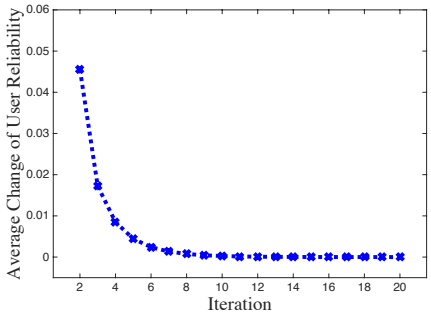

Figure 4: Convergence Analysis

the proposed method is linear to the number of Q&A pairs, which makes the proposed method practical for real-world applications.

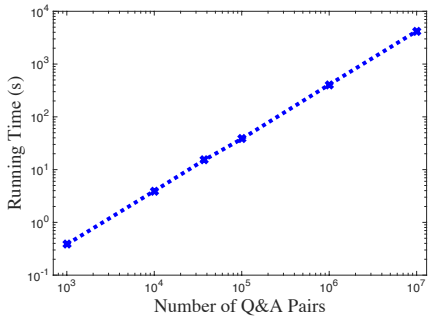

Figure 5: Running Time w.r.t. Number of Q&A Pairs

4. RELATED WORK

Truth Discovery. Recent years have witnessed an increased interest in the research topic of truth discovery, which aims to resolve conflicts and identify trustworthy information from noisy multi-source data. Various truth discovery methods have been developed, such as TruthFinder [34], AccuSim [8, 17], Investment [25], and CRH [16]. Although these truth discovery methods use different ways to estimate user reliability, they share the similar general principle: If a user often provides trustworthy information, he will be assigned a high-reliability degree; meanwhile, if a piece of information is supported by many reliable users, it will be regarded as a trustworthy one. Nowadays, people are investigating various aspects of truth discovery, such as dealing with different data types [16,36], analyzing source (i.e., user) dependency [8,26], enriching the meaning of user reliability [15,27], etc.

In the area of truth discovery, there is some related work considering the relations among answers. In TruthFinder [34] and AccuSim [8,17], the "implication" function is proposed to capture the influence of other possible answers to a particular answer, which can be used to capture the relations among different answers to the same question. However, such implication functions in both TruthFinder and AccuSim need to be set by external knowledge, which might be impossible for large-scale real-world application. Some truth discovery work [19, 21, 31, 32] explores the temporal relations among answers. For example, today's high temperature for New York City (NYC) is correlated with the one of yesterday. However, all these work assumes that the temporal relations are already known *a priori*. Different from existing work, the proposed method does not assume any external knowledge on answers' relations, and can automatically discover semantic relations among answers.

Although [20] utilizes question and answer texts in the truth discovery process, their work only uses them to discover fine-grained topics about questions, and the semantic relations among questions and answers are not considered.

In [9, 10], although the authors point out that the answers have semantic meanings and such semantic meanings should be taken into account, they do not propose a solution to this problem. To the best of our knowledge, the proposed truth discovery method in this paper is the first to automatically discover the semantic meanings of answers by utilizing question and answer texts, and incorporate such semantic meanings into truth discovery procedure. This enlarges the application scope of truth discovery, and enables more applications that deal with textual data.

Crowdsourced Question Answering. From a broader view, there is another relevant research area that studies the quality evaluation of question-answer pairs in crowdsourced question answering websites. This line of related work can be categorized into two groups. The first group [4, 12, 13, 28] formulates the quality evaluation task as a classification problem. To train a good classifier, these methods need a large amount of labeled data, which are difficult or even impossible to obtain for the large-scale medical diagnosis task. The second group [33, 38] infers the quality of question-answer pairs based on the expertise of users who provide the answers. However, these methods require various external information, such as the best answer voting information, to estimate the expertise of users. Unfortunately, on the crowdsourcing medical diagnosis websites, the patients seldom give such feedback, which makes it difficult to apply these methods. Therefore, in this paper, we propose an effective method that discovers trustworthy answers without any supervision.

5. CONCLUSIONS

In this paper, we identify trustworthy medical diagnoses from crowdsourcing users. As these users are not medical experts, the diagnosis answers provided by them may be noisy or even wrong, which may cause serious consequences. In order to distill trustworthy medical diagnoses, it is essential to distinguish reliable users from unreliable ones. Truth discovery methods can be adopted for such user reliability estimation. However, existing truth discovery methods do not take into account the rich semantic meanings of the diagnosis answers. To tackle this challenge, we propose to represent answers as real-valued vectors, which enables the proposed method to model the semantic relations among answers. In order to learn such vector representations for answers, we utilize the question and answer texts. Unfortunately, question-answer pairs from crowdsourcing are noisy, and they suffer the information quality problem. Motivated by this, we tightly combine the general principle of truth discovery and the idea of vector representation learning, and thus these two components can mutually enhance each other. To validate the effectiveness of the proposed method, we collect a large-scale real-world dataset from a popular health-related question answering website. Experimental results show that the proposed method achieves the best performance compared to baseline methods, which is due to the fact that the proposed method can successfully capture the semantic meanings of answers. We also illustrate the advantages of the proposed method on user reliability estimation and vector representation learning.

6. ACKNOWLEDGMENTS

Most of this work was done during an internship of the first author in Baidu Research Big Data Lab, Sunnyvale, CA. This work was sponsored in part by US National Science Foundation under

grant IIS-1553411. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

7. REFERENCES

[1] Jieba, chinese word segmentation package. https://github.com/fxsjy/jieba.

[2] word2vec tool, google. https://code.google.com/archive/p/word2vec/.

[3] C. C. Aggarwal and T. Abdelzaher. Social sensing. In *Managing and mining sensor data*, pages 237–297. 2013.

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proc. of WSDM*, pages 183–194, 2008.

[5] Y. Bachrach, T. Graepel, T. Minka, and J. Guiver. How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. of ICML*, 2012.

[6] D. P. Bertsekas. Nonlinear programming. 1999.

[7] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of ICML*, pages 160–167, 2008.

[8] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[9] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proc. of KDD*, pages 601–610, 2014.

[10] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. *PVLDB*, 7(10):881–892, 2014.

[11] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proc. of WSDM*, pages 131–140, 2010.

[12] H. Hu, B. Liu, B. Wang, M. Liu, and X. Wang. Multimodal DBN for predicting high-quality answers in cQA portals. In *Proc. of ACL*, pages 843–847, 2013.

[13] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *Proc. of SIGIR*, pages 228–235, 2006.

[14] F. Li, M. L. Lee, and W. Hsu. Entity profiling with varying source reliabilities. In *Proc. of KDD*, pages 1146–1155, 2014.

[15] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, D. Murat, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *PVLDB*, 8(4):425–436, 2015.

[16] 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. In *Proc. of SIGMOD*, pages 1187–1198, 2014.

[17] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? *PVLDB*, 6(2):97–108, 2012.

[18] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *SIGKDD Explorations Newsletter*, 17(2):1–16, 2016.

[19] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han. On the discovery of evolving truth. In *Proc. of KDD*, pages 675–684, 2015.

[20] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han. Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. of KDD*, pages 745–754, 2015.

[21] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, and Y. Cheng. Truth discovery on crowd sensing of correlated entities. In *Proc. of Sensys*, pages 169–182, 2015.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[23] A. Mnih and G. E. Hinton. A scalable hierarchical distributed language model. In *NIPS*, pages 1081–1088, 2009.

[24] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: credibility of user statements in health communities. In *Proc. of KDD*, pages 65–74, 2014.

[25] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proc. of COLING*, pages 877–885, 2010.

[26] R. Pochampally, A. D. Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In *Proc. of SIGMOD*, pages 433–444, 2014.

[27] G.-J. Qi, C. C. Aggarwal, J. Han, and T. Huang. Mining collective intelligence in diverse groups. In *Proc. of WWW*, pages 1041–1052, 2013.

[28] C. Shah and J. Pomerantz. Evaluating and predicting answer quality in community QA. In *Proc. of SIGIR*, pages 411–418, 2010.

[29] L. Su, Q. Li, S. Hu, S. Wang, J. Gao, H. Liu, T. Abdelzaher, J. Han, X. Liu, Y. Gao, and L. Kaplan. Generalized decision aggregation in distributed sensing systems. In *Proc. of RTSS*, pages 1–10, 2014.

[30] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. of IPSN*, pages 233–244, 2012.

[31] S. Wang, L. Su, S. Li, S. Yao, S. Hu, L. Kaplan, T. Amin, T. Abdelzaher, and W. Hongwei. Scalable social sensing of interdependent phenomena. In *Proc. of IPSN*, pages 202–213, 2015.

[32] S. Wang, D. Wang, L. Su, L. Kaplan, and T. Abdelzaher. Towards cyber-physical systems in social spaces: The data reliability challenge. In *Proc. of RTSS*, pages 74–85, 2014.

[33] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen. Cqarank: Jointly model topics and expertise in community question answering. In *Proc. of CIKM*, pages 99–108, 2013.

[34] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. of KDD*, pages 1048–1052, 2007.

[35] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proc. of COLING*, 2014.

[36] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proc. of QDB*, 2012.

[37] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *PVLDB*, 5(6):550–561, 2012.

[38] T. Zhao, C. Li, M. Li, S. Wang, Q. Ding, and L. Li. Predicting best responder in community question answering using topic model method. In *Proc. of WI-IAT*, pages 457–461, 2012.