Modeling Topic-Based Human Expertise for Crowd Entity Resolution

Sai-Sai Gong[1], Wei Hu[1,*], *Member, CCF, ACM*, Wei-Yi Ge[2], and Yu-Zhong Qu[1], *Senior Member, CCF*

[1] *State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China*
[2] *Science and Technology on Information Systems Engineering Laboratory, Nanjing 210007, China*

E-mail: saisaigong@gmail.com; whu@nju.edu.cn; geweiyi@163.com; yzqu@nju.edu.cn

Abstract Entity resolution (ER) aims to identify whether two entities in an ER task refer to the same real-world thing. Crowd ER uses humans, in addition to machine algorithms, to obtain the truths of ER tasks. However, inaccurate or erroneous results are likely to be generated when humans give unreliable judgments. Previous studies have found that correctly estimating human accuracy or expertise in crowd ER is crucial to truth inference. However, a large number of them assume that humans have consistent expertise over all the tasks, and ignore the fact that humans may have varied expertise on different topics (e.g., music versus sport). In this paper, we deal with crowd ER in the Semantic Web area. We identify multiple topics of ER tasks and model human expertise on different topics. Furthermore, we leverage similar task clustering to enhance the topic modeling and expertise estimation. We propose a probabilistic graphical model that computes ER task similarity, estimates human expertise, and infers the task truths in a unified framework. Our evaluation results on real-world and synthetic datasets show that, compared with several state-of-the-art approaches, our proposed model achieves higher accuracy on the task truth inference and is more consistent with the human real expertise.

Keywords entity resolution, crowdsourcing, human expertise, topic modeling, task similarity

1 Introduction

In the Semantic Web area, the goal of entity resolution (ER) is to identify entities from diverse knowledge bases referring to the same real-world thing[1,2]. It is vital to the reuse, integration, and application of the linked data. To obtain benefits from human knowledge, many ER approaches involve humans into the workflow, called crowd ER, e.g., using micro-task crowdsourcing and aggregating the human-provided judgments for learning[3,4]. However, as the humans participated in the ER tasks are not always "oracles", they probably generate inaccurate or erroneous judgments. Consequently, given a task, inconsistency often exists among the human judgments[5]. Thus, a key issue for crowd ER is to identify the task truth (i.e., whether or not the entities in the task refer to the same thing) from the human judgments with inconsistency.

To determine the task truths from inconsistent human answers, several existing approaches predict the truths by considering human expertise, i.e., the accuracy or reliability of his/her judgments[6-8]. These approaches eliminate "bad" workers and select task answers provided by "good" workers to decide the task truths using majority voting[9] or other advanced methods. However, many existing approaches simply assume that a human has the same expertise level over all the tasks and do not consider expertise variance in different task topics. This assumption may be suitable for some simple crowdsourcing tasks such as image labeling[10] and recognizing textual entailment in a single task topic[11], but it would lead to limited performance on more complicated tasks such as ER[12]. For the tasks requiring different topic knowledge, a human may be good at the task topics of which they have enough knowledge, but may provide poor answers for other unfamiliar topics. As a result, human accuracy changes significantly among different task topics. For example, a human familiar with American geography can easily identify The Big Apple is identical to

New York City but different from The Apple Company, while he/she may not correctly identify the fact that the American writer Mark Twain is alias to Samuel Langhorne Clemens. Furthermore, although a human is familiar with American geography, he/she may not be familiar with Chinese geography. Therefore, it is beneficial to estimate fine-grained human expertise on different task topics for crowd ER.

Topic modeling[13] can be described as a method for discovering the latent topics that occur in a collection of documents, where each document is represented by a mixture of latent topics. It has become a well-known solution to text analysis. In addition to the great success in modeling text documents, topic modeling is also widely used in other areas such as image content modeling[13] and author resolution[14]. In this paper, we model each crowd ER task as a mixture of latent topics and leverage topic modeling to identify the topics of tasks.

Furthermore, a crowd ER task is often similar to a number of other tasks, e.g., based on some features. The similarity between two tasks generally indicates some similarity between their topic mixtures. In other words, if two tasks are more similar, their topic mixtures should be more similar as well. Incorporating task similarity can model the task topics more accurately, avoid cold starts, and thus better estimate the human expertise. For example, crowdsourcing may be challenged by the data sparsity problem, i.e., a task only collects very few judgments. As a result, it may be unreliable to obtain the task truth based on such limited number of judgments. But, the problem may be alleviated by borrowing the trustworthy information from similar tasks.

In this paper, we propose a probabilistic graphical model, called Topic-Expertise-Similar-Tasks (TEST), which computes ER task similarity, estimates human expertise, and infers the truths of the tasks in a unified framework. TEST jointly models the content of the tasks and human judgments to learn latent topics and estimates a human's topic-based expertise by considering his/her completed similar tasks. Moreover, TEST encodes the task similarity by clustering tasks into groups to improve the topic coherence of similar tasks and model the task topics and truths more accurately. Our evaluation results show that the proposed TEST model achieves better performance on both task truth inference and human expertise estimation.

Our salient contributions of this paper are listed as follows.

• We model the varied human expertise on various latent topics to improve human expertise estimation. We leverage similar task clustering for modeling the topics more accurately.

• We propose a probabilistic model which can learn the human varied expertise on task topics, compute the task similarity, and infer the truths of the tasks integrally.

• We experimentally demonstrate that, compared with five representative approaches on two real-world datasets and a synthetic dataset, the proposed TEST model can achieve higher accuracy on truth inference with less humans and be more consistent with the human real expertise.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 formalizes our addressed problem. We describe the proposed model in Section 4. We report the experiments and results in Section 5. Finally, this paper is concluded with future work in Section 6.

2 Related Work

Crowdsourcing has become an effective paradigm for human-powered task solving and attracted a lot of interest in a wide range of fields[15,16]. Crowd ER incorporates crowdsourcing into the ER process for harnessing the human knowledge[3,17-20]. In addition to the crowd ER discussed in this paper, there are also many other computer-hard problems that can benefit from the use of crowdsourcing, such as linked data quality assessment[21,22] and image labeling[10,23].

The quality of crowdsourcing results may be low and imperfect. Thus, quality control is an important issue for crowdsourcing. Humans usually have different levels of expertise, and some of them may not be able to provide right answers for hard and unfamiliar tasks. To achieve high-quality results, we need to tolerate errors and infer task truths from the crowd judgments containing noises. The first step of quality control is modeling a human's expertise/quality. Then, we can use strategies like low-quality human elimination, answer aggregation or targeted task assignment to improve quality[15]. In crowd ER, human expertise can be easily estimated if the ground truths of ER tasks have been known. However, in our problem setting, the ground truth for any ER task is unknown, which is natural and prevalent since the ground truths are often too expensive or too tedious to be obtained on a large scale in many situations. As a result, the inference of task truths and the estimation of human expertise depend on each other.

1206

J. Comput. Sci. & Technol., Nov. 2018, Vol.33, No.6

Please note that this paper does not discuss other issues in crowdsourcing in detail, such as cost control, latency control and task design[15,24].

To estimate human expertise for quality control, many existing approaches assume that a human worker has consistent expertise over all tasks, which is different from TEST. CDB[25] provides a graph-based query model that supports crowd-based query optimizations in database systems for small monetary costs, low latency, and high quality results. The work in [26] deals with the problem of selecting a set of workers such that their answers yield the highest result quality w.r.t. a voting strategy without exceeding the overall budget. Requallo[27] deals with the trade-off problem between quantity and quality in crowdsourcing tasks under a tight budget. SADU[28] detects sybil worker groups by clustering and throttles sybil attack. ZenCrowd[18] uses a factor graph to model human expertise and task truths. CRH[29] formulates the truth inference and human expertise estimation as an optimization problem for conflict resolution on heterogeneous data. The Bayesian approaches that simultaneously estimate human expertise and task truths have also been proposed in [7]. The solutions of these approaches are usually computed by an Expectation-Maximization (EM) algorithm or a dedicated algorithm for the corresponding optimization problem. Several studies also consider other factors besides human expertise for inferring task truths. For example, GLAD[10] leverages the human expertise and task difficulty to infer task truths. ETCIBoot[30] provides confidence interval estimations of task truths. Sifter[31] learns from various online platforms and measures platform-level and user-level expertise simultaneously to perform estimation on drug side-effects. UbTD[32] focuses on the problem of discovering truths from the distributed data. The work in [33] proposes an optimization-based method, which incorporates task correlation to estimate human expertise and task truths. IATD[34] makes an assumption that how a human worker provides his/her judgments may be influenced by other workers and utilizes correlations among human workers to learn human expertise and task truths. There are also studies focusing on task assignment to improve crowdsourcing quality. The work in [35] utilizes the spatial location information of human workers to measure the human quality and performs online task assignment for crowdsourced POI (points of interest) labeling. QASCA[36] leverages the variants of two measures, accuracy and F-score, to facilitate online task assignment. However, on the com-

plicated tasks requiring rich topic knowledge such as ER, those approaches with only a single expertise level per human may result in undesirable performance[12].

Rather than using consistent expertise, a few approaches assume that a human worker has diverse expertise over tasks[6,12]. For example, iCrowd[12] is an optimization-based approach that leverages a human's performance on the completed tasks to infer his/her expertise on similar tasks. ACRyLIQ[21] deals with the adaptive task assignment problem in crowdsouced linked data quality assessment, and addresses the cold-start issue related to the new human workers by estimating the new workers' expertises based on the test questions generated from a knowledge base. Different from TEST, these approaches do not model human expertise based on task domains or topics, which may not reveal humans' knowledge well[8]. There also exist a few approaches assuming that a human has diverse expertise across task domains or topics. Some of them identify latent task topics by task clustering and consider each cluster as a latent task topic, on which human expertise is modeled[37]. However, the performance of these clustering-based approaches is influenced by the used clustering algorithms. ALM[8] is an active learning algorithm. It firstly learns a high-level feature representation of tasks by using a sparse coding algorithm, where each feature component represents a latent topic. Then, ALM uses a probabilistic graphical model to learn the task truths and human expertise, where the human expertise depends on the high-level representation. FaitCrowd[5] is used for crowdsourcing aggregation like aggregating answers in question answering or slot filling. It leverages a probabilistic model based on the topic model TwitterLDA to identify the latent topics of tasks and estimate human expertise and task truths simultaneously. DOCS[38] is a system containing three modules: domain vector estimation (DVE), truth inference (TI), and online task assignment (OTA). DVE uses entity linking to find entities in the extra knowledge bases (i.e., Freebase) that are mentioned in the task content, and resorts to their categories for identifying the multiple explicit domains (upper-level classes) of a task. TI takes iterative procedures in which the truth inference step and the human expertise estimation step are iteratively performed until convergence. In addition, DOCS needs some ground truths to initialize the human expertise. The work [39] compares the performance of several truth inference algorithms that use different modeling of human expertise and task truth, and shows the differences among the al-

gorithms on solving different types of tasks. Different from ALM and DOCS, TEST leverages topic modeling to identify the latent topics of tasks. Also, different from FaitCrowd, which assigns a single topic to each task, TEST assumes that a task belongs to multiple topics. Furthermore, TEST incorporates task similarity for truth inference.

For most of the aforementioned approaches except [12] and [33], they do not consider the correlations among tasks. However, the correlations among tasks play a vital role in crowdsourcing aggregation, such as improving the accuracy of task truth inference[40] or solving the sparsity problem of human judgments[33]. In this paper, the proposed TEST model combines topic modeling and task similarity modeling in a unified graphical model, in order to model task topics better and eventually estimate the human expertise and task truths more accurately.

3 Problem Statement

An entity e is expressed as a tuple (id_e, D_e), where id_e is a unique identifier denoting e, and D_e is a set of property-value pairs $\{(p_1, v_1), (p_2, v_2), \ldots, (p_n, v_n)\}$ representing the descriptions of e, where p_i is a property of e. In the Semantic Web, each named entity is denoted by a URI as its unique identifier. Each value can be a literal, another entity or a blank node.

Let $\mathcal{R} = \{R_1, R_2, \ldots, R_T\}$ denote the set of all ER tasks. An ER task, $R_t \in \mathcal{R}$, consists of two entities e_i^t, e_j^t to be resolved and an unknown task truth $z_t \in \{0, 1\}$, where $z_t = 1$ denotes that e_i^t, e_j^t are matched (i.e., refer to the same real-world thing), while $z_t = 0$ denotes that they are not matched. In our paper, entity resolution for an entity pair is equivalent to finding the truth of the corresponding task.

Let $\mathcal{H} = \{H_1, H_2, \ldots, H_U\}$ be the set of all humans who participate in the ER tasks and give their judgments. A human, $H_u \in \mathcal{H}$, resolves a subset of the tasks, and a task can be judged by different humans. Let U_t be the number of humans resolving the task R_t. The judgment given by human H_u for task R_t is denoted by $l_{t,u} \in \{0, 1\}$, where $l_{t,u} = 1$ means that human H_u regards entities e_i^t, e_j^t in task R_t as matched ones, while $l_{t,u} = 0$ means that H_u regards the two entities as non-matched ones.

A task can be represented by a mixture of latent topics, each of which is a multinomial distribution over a fixed set of features. We denote the multinomial distribution of the k-th topic by $\boldsymbol{\varphi}_k$. A task R_t contains N_t features $\{f_{t,n}\}_{n=1}^{N_t}$. Each feature $f_{t,n}$ in the task is a word extracted from the text descriptions of the two involved entities, their properties and values. For example, the feature words "longitude", "latitude" from properties geo:long and geo:lat contribute to the location topic of an entity.

The tasks can be clustered together based on the similarity of topics to form task groups, each of which represents some semantic categorization of the task topics. Let G be the total number of task groups. We denote the probability of a task R_t belonging to the g-th task group by $\phi_{g,t}$ with $\sum_g \phi_{g,t} = 1$, and use s_t to denote the group of R_t. Given a task R_t, we use $\boldsymbol{\mathcal{A}}^t = (I(s_1 = s_t), I(s_2 = s_t), \cdots, I(s_T = s_t))$ to represent a vector consisting of the indices of the tasks in the same group as R_t. $\boldsymbol{\mathcal{A}}^t$ is a T-dimensional vector, and $I(\cdot)$ is an indicator function.

Human expertise, denoted by $\boldsymbol{\rho} \in \mathbb{R}^{U \times K}$, is referred to as the expertise levels of different humans on K topics, i.e., the probability of identifying the truth of a task in specified topics, where $\rho_{u,k}$ denotes the expertise of human H_u on the k-th topic ($1 \leqslant u \leqslant U, 1 \leqslant k \leqslant K$). For convenience, the main notations used in this paper are summarized in Table 1.

Table 1. Notations for the TEST Model

Symbol	Description
R_t	The t-th task (including two entities and an unknown truth)
z_t	Unknown truth for the t-th task
H_u	The u-th human
$l_{t,u}$	Judgment given by the u-th human for the t-th task
$f_{t,n}$	The n-th feature for the t-th task
$d_{t,n}$	Topic assigned to feature $f_{t,n}$
$\rho_{u,k}$	Expertise of the u-th human on the k-th topic
s_t	Task group of the t-th task
$\boldsymbol{\mathcal{A}}^t$	Indices of the tasks that are in the same group as R_t
$\phi_{g,t}$	Probability of R_t belonging to the g-th task group
$\boldsymbol{\varphi}_k$	Multinomial distribution over features specific to the k-th topic
$\boldsymbol{\theta}_t$	Multinomial distribution over topics specific to the t-th task
$\boldsymbol{\lambda}_t$	Multinomial distribution over tasks specific to the t-th task
$c_{t,n}$	Latent variable from which $d_{t,n}$ is sampled, and its value is a task in group s_t
$\alpha, \boldsymbol{\beta}$	Dirichlet priors to multinomial distribution $\boldsymbol{\theta}, \boldsymbol{\lambda}$, respectively
π	Prior probability of a task truth

Based on the above notations, we define our studied problem as follows.

Definition 1 (Problem). *Given the task set $\{R_t\}_{t=1}^T$ with unknown truths, the features $\{f_{t,n}\}_{n=1}^{N_t}$*

for each task R_t, the human judgments $\{l_{t,u}\}_{t=1,u=1}^{T,U}$, the number of topics K and the number of task groups G, our goal in this paper is to estimate 1) the task truths $\{z_t\}_{t=1}^{T}$ for all the tasks and 2) the expertise of all the humans $\{\rho_{u,k}\}_{u=1}^{U}$ on each topic.

4 Topic-Expertise-Similar-Tasks Model

In this section, we firstly describe the generative process of the proposed TEST model. Then, we show the variational inference and parameter estimation for the model.

4.1 Generative Process

The plate graph of TEST is shown in Fig.1. In this model, each task belongs to a task group. Each task group has its unique task topic set different from other groups. The tasks in the same group should have similar task topics, and consequently have similar human labeling behaviors.
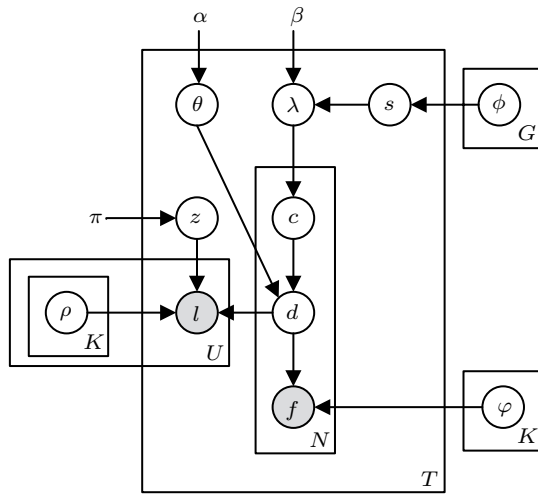

Fig.1. Overview of the TEST model.

The task generation process is as follows. For each ER task R_t, we firstly sample its truth z_t from a Bernoulli distribution $Ber(\pi)$, where π represents the prior probability that a task truth is 1 (π is set to 0.5 if the prior probability is unknown). Then, we sample its topic distribution θ_t from a Dirichlet distribution $Dir(\alpha)$, where α is the Dirichlet prior of topics. θ_t is a multinomial distribution showing the probabilities that the task describes different topics. Next, let $\phi_{g,t}$ denote the probability of R_t belonging to the g-th task group.

We sample a task group s_t for R_t from the multinomial distribution derived from $\{\phi_{g,t}\}_{g=1}^{G}$.

To leverage the similar tasks of R_t, we allow that the topics of its similar tasks can also be used to generate the features of R_t. Towards this end, we use $\boldsymbol{\lambda}_t$ to denote how likely each task in s_t can be sampled to generate the features, which follows a multinomial distribution. We generate $\boldsymbol{\lambda}_t$ from the Dirichlet distribution $Dir(\boldsymbol{\beta}_t \circ \boldsymbol{\mathcal{A}}^t)$, where $\boldsymbol{\beta}_t$ denotes the prior similarity between R_t and each other task in the whole task set. $\boldsymbol{\beta}_t$ is represented by a T-dimensional vector and its dimension indicates the similarity between R_t and another task. We use the weighted sum combination of property similarity, URI pay-level-domain[①] similarity and value similarity to compute the overall similarity between any two tasks. $\boldsymbol{\mathcal{A}}^t$ indicates which tasks belong to s_t (see Section 3), which is computed by checking each task R_k in the whole task set and validating whether $s_k = s_t$. $\circ$ represents the Hadamard product (a.k.a. entrywise product), and $\boldsymbol{\beta}_t \circ \boldsymbol{\mathcal{A}}^t$ retains the values of the tasks within group s_t and discards the others.

To generate the n-th feature $f_{t,n}$ in the task R_t, we first need to decide which task within the group s_t should be used to generate the feature. We denote this task as $c_{t,n}$ and sample it from a multinomial distribution $Mult(\boldsymbol{\lambda}_t)$. Then, we choose a topic $d_{t,n} \sim Mult(\boldsymbol{\theta}_{c_{t,n}})$, where $\boldsymbol{\theta}_{c_{t,n}}$ denotes the topic mixture of $c_{t,n}$. Finally, the feature $f_{t,n}$ is sampled according to the multinomial distribution of topic $d_{t,n}$ using $f_{t,n} \sim Mult(\boldsymbol{\varphi}_{d_{t,n}})$, where $\boldsymbol{\varphi}_{d_{t,n}}$ denotes the multinomial distribution over features of topic $d_{t,n}$.

Next, we need to generate the judgment $l_{t,u}$ given by the human H_u for task R_t. We use $\boldsymbol{\rho}_u$ to denote the expertise of human H_u on all the topics, which is defined in Section 3. Let $\bar{\boldsymbol{d}}_t$ be a K-dimensional vector, each element of which is computed by $\bar{d}_{t,k} = \frac{1}{N_t} \sum_{n=1}^{N_t} I(d_{t,n} = k)$, $1 \leqslant k \leqslant K$, where $I(\cdot)$ is an indicator function. As a human with better expertise on a task's topics is likely to give a correct judgment with higher probability, we model the probability of $l_{t,u}$ as a function g of the human expertise $\boldsymbol{\rho}_u$, the task truth z_t, and the topics of R_t, which is defined as follows:

$$g(l_{t,u}|z_t, \bar{\boldsymbol{d}}_t, \boldsymbol{\rho}_u) = Ber\big(l_{t,u} = z_t|\sigma(\boldsymbol{\rho}_u^T \bar{\boldsymbol{d}}_t)\big), \quad (1)$$

where σ denotes the sigmoid function. Therefore, whether the human provides a correct judgment depends on the value of $\sigma(\boldsymbol{\rho}_u^T \bar{\boldsymbol{d}}_t)$.

[①]The pay-level-domain is a sub-domain of a public top-level domain, for which users usually pay, e.g., the pay-level domain for www.example.com is example.com. Pay-level domains allow to identify a realm, where a data publisher is likely to be in control.

To put it all together, the generative process is summarized in Algorithm 1.

Algorithm 1. Generative Process of TEST

---

1 **foreach** task R_t **do**

2   Draw a truth $z_t \sim Ber(\pi)$;

3   Draw a topic mixture $\boldsymbol{\theta}_t \sim Dir(\alpha)$;

4   Draw a task group $s_t \sim Mult(\boldsymbol{\phi}_t)$;

5   Draw a similar task mixture $\boldsymbol{\lambda}_t \sim Dir(\boldsymbol{\beta}_t \circ \boldsymbol{\mathcal{A}}^t)$;

6   **foreach** feature $f_{t,n}$ **do**

7   Draw a sampled task $c_{t,n} \sim Mult(\boldsymbol{\lambda}_t)$;

8   Draw a topic $d_{t,n} \sim Mult(\boldsymbol{\theta}_{c_{t,n}})$;

9   Draw $f_{t,n} \sim Mult(\boldsymbol{\varphi}_{d_{t,n}})$;

10   **foreach** human H_u judging R_t **do**

11   Draw a judgment $l_{t,u} \sim g(\cdot|z_t, \bar{\boldsymbol{d}}_t, \boldsymbol{\rho}_u)$;

Fig.2 shows a running example of TEST. In the whole task set, tasks are formed into groups. For example, task group 1 describes location-related topics like geometry while task group 2 describes company-related topics. Four humans have relatively high expertise on the topics of task group 1 but low expertise on the topics of task group 2. On the contrary, five humans have high expertise on the topics of task group 2 but not task group 1. For the current task about resolving two entities related to Beijing, we can find its similar tasks from task group 1 and identify that Beijing should have location-related topics. Assume that H_1, H_2 and H_3 provide judgments on the current task, and H_1 and H_2 have higher expertise on the task topics than H_3. In the process of truth inference, the judgments of H_1 and H_2 would be more preferable.

4.2 Variational Inference

The parameters in TEST are $\boldsymbol{\Theta} = \{\alpha, \boldsymbol{\beta}, \pi, \boldsymbol{\phi}, \boldsymbol{\varphi}, \boldsymbol{\rho}\}$, the observed variables are $\boldsymbol{l}, \boldsymbol{f}$, and the hidden variables are $\boldsymbol{\Delta} = \{\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{s}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{z}\}$. The joint distribution of all the variables is computed as follows.

$$
\begin{aligned}
&P(\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Delta}|\boldsymbol{\Theta}) \\
&= \prod_{t=1}^{T} P(z_t|\pi)P(\boldsymbol{\theta}_t|\alpha)P(s_t|\boldsymbol{\phi}_t)P(\boldsymbol{\lambda}_t|\boldsymbol{\beta}, \boldsymbol{\mathcal{A}}^t) \\
&\quad \prod_{n=1}^{N_t} P(c_{t,n}|\boldsymbol{\lambda}_t)P(d_{t,n}|\boldsymbol{\theta}, c_{t,n})P(f_{t,n}|\boldsymbol{\varphi}_{d_{t,n}}) \\
&\quad \prod_{u=1}^{U_t} P(l_{t,u}|z_t, \bar{\boldsymbol{d}}_t, \boldsymbol{\rho}_u).
\end{aligned}
\tag{2}
$$

In order to apply an EM algorithm to (2), the key inferential problem that we need to solve is to compute the posterior distribution of the hidden variables $P(\boldsymbol{\Delta}|\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Theta})$. However, this posterior distribution is intractable for exact inference[13]; thus we resort to the approximate inference methods.

In this paper, we employ the mean-field variational inference[41] due to its computational efficiency. Variational inference introduces a variational distribution q over the hidden variables to approximate the true posterior distribution $P(\boldsymbol{\Delta}|\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Theta})$, which is defined as follows:

$$
\begin{aligned}
q(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{s}, \boldsymbol{c}, \boldsymbol{d}) &= \prod_{t=1}^{T} q_\theta(\boldsymbol{\theta}_t|\boldsymbol{\varrho}_t)q_\lambda(\boldsymbol{\lambda}_t|\boldsymbol{\eta}_t)q_s(s_t|\boldsymbol{\omega}_t) \\
&\quad \prod_{n=1}^{N_t} q_c(c_{t,n}|\boldsymbol{\epsilon}_{t,n})q_d(d_{t,n}|\boldsymbol{\vartheta}_{t,n}),
\end{aligned}
\tag{3}
$$

Fig.2. Running example.

1210

J. Comput. Sci. & Technol., Nov. 2018, Vol.33, No.6
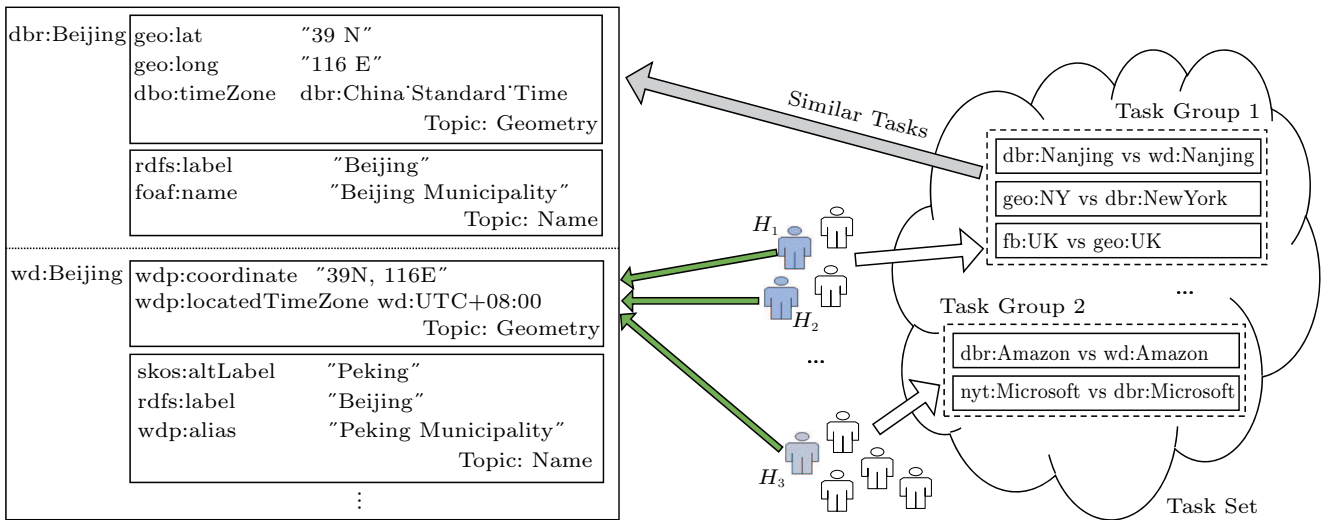
where $\{\varrho_t\}_{t=1}^T$, $\{\boldsymbol{\eta}_t\}_{t=1}^T$, $\{\boldsymbol{\omega}_t\}_{t=1}^T$, $\{\boldsymbol{\epsilon}_{t,n}\}_{t=1,n=1}^{T,N_t}$ and $\{\boldsymbol{\vartheta}_{t,n}\}_{t=1,n=1}^{T,N_t}$ are the variational parameters specified for the corresponding hidden variables. The basic idea of variational inference is to optimize the variational parameters so that the evidence lower bound[41] of the log-likelihood of observed variables $\ln P(\boldsymbol{l}, \boldsymbol{f}|\boldsymbol{\Theta})$ is maximized. To be more specific, we have

$$\ln P(\boldsymbol{l}, \boldsymbol{f}|\boldsymbol{\Theta}) \geqslant L(\varrho, \boldsymbol{\eta}, \boldsymbol{\omega}, \boldsymbol{\epsilon}, \boldsymbol{\vartheta})$$
$$= \mathbb{E}_q[\ln P(\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Delta}|\boldsymbol{\Theta})] - \mathbb{E}_q[\ln q],$$

where $L(\varrho, \boldsymbol{\eta}, \boldsymbol{\omega}, \boldsymbol{\epsilon}, \boldsymbol{\vartheta}) = \mathbb{E}_q[\ln P(\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Delta}|\boldsymbol{\Theta})] - \mathbb{E}_q[\ln q]$ indicates the evidence lower bound, $\mathbb{E}_q[\ln P(\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Delta}|\boldsymbol{\Theta})]$ is the expectation of the logarithm of (2) w.r.t. the approximation distribution (3), and $\mathbb{E}_q[\ln q]$ is the expectation of the logarithm of (3) w.r.t. the distribution (3). $\mathbb{E}_q[\ln P(\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Delta}|\boldsymbol{\Theta})]$ and $\mathbb{E}_q[\ln q]$ are computed as follows:

$$\mathbb{E}_q[\ln P(\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Delta}|\boldsymbol{\Theta})]$$
$$= \sum_{t=1}^T \mathbb{E}_q[\ln P(z_t|\pi)] +$$
$$\sum_{t=1}^T \mathbb{E}_q[\ln P(\boldsymbol{\theta}_t|\alpha)] + \sum_{t=1}^T \mathbb{E}_q[\ln P(s_t|\boldsymbol{\phi}_t)] +$$
$$\sum_{t=1}^T \mathbb{E}_q[\ln P(\boldsymbol{\lambda}_t|\boldsymbol{\beta}, \boldsymbol{\mathcal{A}}^t)] + \sum_{t=1}^T \sum_{n=1}^{N_t} \mathbb{E}_q[\ln P(c_{t,n}|\boldsymbol{\lambda}_t)] +$$
$$\sum_{t=1}^T \sum_{n=1}^{N_t} \mathbb{E}_q[\ln P(d_{t,n}|\boldsymbol{\theta}, c_{t,n})] +$$
$$\sum_{t=1}^T \sum_{n=1}^{N_t} \mathbb{E}_q[\ln P(f_{t,n}|\boldsymbol{\varphi}_{d_{t,n}})] +$$
$$\sum_{t=1}^T \sum_{u=1}^{U_t} \mathbb{E}_q[\ln P(l_{t,u}|z_t, \bar{\boldsymbol{d}}_t, \boldsymbol{\rho}_u)],$$

$$\mathbb{E}_q[\ln q]$$
$$= \sum_{t=1}^T \mathbb{E}_q[\ln q_\theta(\boldsymbol{\theta}_t|\varrho_t)] + \sum_{t=1}^T \mathbb{E}_q[\ln q_\lambda(\boldsymbol{\lambda}_t|\boldsymbol{\eta}_t)] +$$
$$\sum_{t=1}^T \mathbb{E}_q[\ln q_s(s_t|\boldsymbol{\omega}_t)] + \sum_{t=1}^T \sum_{n=1}^{N_t} \mathbb{E}_q[\ln q_c(c_{t,n}|\boldsymbol{\epsilon}_{t,n})] +$$
$$\sum_{t=1}^T \sum_{n=1}^{N_t} \mathbb{E}_q[\ln q_d(d_{t,n}|\boldsymbol{\vartheta}_{t,n})].$$

4.3 Parameter Estimation

We use a variational EM algorithm, which alternatively performs an expectation (E) step and a maximization (M) step in each iteration, to estimate the

variational parameters and other model parameters. Specifically, to compute the expectation of (1) in terms of the variational distribution, we approximate (1) by $\exp(\boldsymbol{\rho}_u^T \bar{\boldsymbol{d}}_t)$. In the E step, we fix the model parameters and update the variational parameters by maximizing the evidence lower bound. In the M step, we fix the variational parameters and update the model parameters. This process continues until convergence.

In the E step, the variational parameters are obtained by maximizing $L(\varrho, \boldsymbol{\eta}, \boldsymbol{\omega}, \boldsymbol{\epsilon}, \boldsymbol{\vartheta})$, and the solutions are given as follows:

$$\varrho_{t,k} = \alpha_k + \sum_{n=1}^{N_t} \sum_{t' \in S_t} \vartheta_{t',n,k}\epsilon_{t',n,t}, \tag{4}$$

$$\eta_{t,i} = \beta_{t,i} + \sum_{n=1}^{N_t} \epsilon_{t,n,i}, \tag{5}$$

$$\epsilon_{t,n,i} \propto \exp\Big(\Psi(\eta_{t,i}) - \Psi(\sum_{j \in S_t} \eta_{t,j}) +$$
$$\sum_{k=1}^K \vartheta_{t,n,k}\big(\Psi(\varrho_{i,k}) - \Psi(\sum_{j=1}^K \varrho_{i,j})\big)\Big), \tag{6}$$

$$\vartheta_{t,n,k} \propto \varphi_{k,f_{t,n}} \exp\Big(\sum_{i \in S_t} \epsilon_{t,n,i}\big(\Psi(\varrho_{i,k}) - \Psi(\sum_{k=1}^K \varrho_{i,k})\big) +$$
$$\frac{1}{N_t} \sum_{u=1}^{U_t} \rho_{u,k}\Big), \tag{7}$$

where $t' \in S_t$ denotes that the index of maximum element in $\boldsymbol{\omega}_t'$ is equal to that in $\boldsymbol{\omega}_t$. Here, $\boldsymbol{\omega}_t' = (\omega_{t',1}, \omega_{t',2}, \cdots, \omega_{t',G})$ and $\boldsymbol{\omega}_t = (\omega_{t,1}, \omega_{t,2}, \cdots, \omega_{t,G})$. There is no closed-form solution of $\boldsymbol{\omega}_t$; therefore we use the Newton-Raphson method[13] to compute the value of $\boldsymbol{\omega}_t$. From (6), we can see that $\epsilon_{t,n,i}$, which approximates the posterior probability of a task R_i in the group of R_t used for generating $f_{t,n}$, depends on the posterior probability of sampling R_i, i.e., $\Psi(\eta_{t,i})$, and the topic relatedness between R_i and $f_{t,n}$, i.e., $\vartheta_{t,n,k}\big(\Psi(\varrho_{i,k}) - \Psi(\sum_{j=1}^K \varrho_{i,j})\big)$.

By using the Bayes' theorem, we can compute the posterior probability of the task truth z_t as follows:

$$P(z_t|\boldsymbol{l}, \boldsymbol{f}, \boldsymbol{\Theta}) \propto P(z_t) \prod_{u=1}^{U_t} g(l_{t,u}|z_t, \bar{\boldsymbol{d}}_t, \boldsymbol{\rho}_u),$$

where $P(z_t) = \pi$ is the prior probability of z_t.

In the M step, we optimize the model parameters by maximizing the evidence lower bound:

$$\varphi_{k,f} \propto \sum_{t=1}^T \sum_{n=1}^{N_t} \vartheta_{t,n,k} I(f_{t,n} = f),$$

where $I(\cdot)$ is an indicator function, and its value equals 1 when $f_{t,n} = f$. The derivate of $L(\varrho, \eta, \omega, \epsilon, \vartheta)$ w.r.t. ϱ_u is a constant value independent of ϱ_u. To solve this problem, we add an L2-norm regularizer. Also, there is no closed-form solution of ϕ_t. Therefore, we use the Newton-Raphson method to compute the corresponding values. The overall time complexity of each EM iteration (an E step and an M step) is $O(TK + TS + 2TG + FS + FK + T + KV + KU)$, where T, K, G, U denote the total number of tasks, topics, task groups and humans, respectively, F denotes the total number of features (words) in all the tasks, V denotes the total number of distinct features in the vocabulary, and S denotes the average number of tasks in a task group. Each item in this time complexity corresponds to a parameter's computation.

In the TEST model, the cross sampling, i.e., sampling from the task itself or its similar tasks in the same task group, of a feature's topic $d_{t,n}$ ensures that similar tasks have similar topic mixtures and borrow strength from each other. At the same time, the topic mixture similarity of tasks would make these tasks belong to the same topic group.

5 Evaluation

In this section, we show our experiments to evaluate TEST, in addition to five comparative approaches, using three datasets of different characteristics. We implemented TEST and comparative approaches in Java. The experiments were carried out on a PC with an Intel® Xeon 3.2 GHz CPU and 2 GB Java virtual machine.

5.1 Datasets

The statistical data of the three datasets are shown in Table 2. We briefly describe them as follows.

• SView is a dataset obtained from a crowdsourcing platform via a Semantic Web browser called SView[42]. The browser allows a human to view various entities and identify its matched ones in his/her browsing activities. More specifically, when the human views an entity, he/she is provided with a few candidate entities that are similar to the current focal one. For each candidate, he/she can provide his/her judgment to the platform. As a result, the judgments on various entity pairs from different humans are collected by the platform. We used the human judgment dataset collected

by the platform until May 1, 2017, and kept the entity pairs that were judged by at least three humans. Based on our observation, humans are prone to wrongly resolving the entity pairs in this dataset, because there exist plenty of inconsistencies among the human judgments. The entities in the dataset belong to 14 different domains, e.g., people, place and organization, showing a variety of topics. Furthermore, the platform provided the truth for each entity pair in the dataset.

Table 2. Dataset Statistics

Dataset	Number of ER Tasks	Number of Task Domains	Number of Human Workers	Number of Judgments
SView	4 124	14	52	18 679
BTC	4 324	8	33	15 996
Synthetic	120 000	3	200	600 000

• BTC is a dataset containing a number of human judgments on the entity pairs in the Billion Triples Challenge 2011 dataset②. The human judgments were collected from a customized crowdsourcing platform[2]. A number of candidate entity pairs that were likely to be matched were first computed using some machine algorithms. Then, each candidate pair was provided to three different humans for judgments. Each entity pair's truth was also offered by the platform. The entities in these pairs belong to eight different domains, e.g., people, places and publications, which also indicate a variety of topics.

• The synthetic dataset is constructed by the integration of entity pairs from the following four datasets: Product[3,20], Restaurant[3], Cora[20], and DBLP-Scholar[43]. Product is a public dataset of mappings from abt.com products to buy.com products. Restaurant is a public dataset consisting of restaurants records from different real-world entities. Cora and DBLP-Scholar are two public bibliography datasets containing paper records from different real-world entities. These four datasets provide truths for all of the entity or record pairs. We converted them into the Semantic Web data format.

To generate the synthetic dataset, we firstly computed on each of the four datasets the candidate entity pairs such that their similarity scores were larger than 0.3. The similarity score of an entity pair was computed by the Jaccard similarity between the token sets generated from the entity properties. Then, we sampled from each dataset a number of candidate entity

---

1212

J. Comput. Sci. & Technol., Nov. 2018, Vol.33, No.6

pairs according to its candidate pair proportion in to-tal size. As a result, 120 000 entity pairs were collected. Next, we artificially generated 200 simulated human workers with different expertise on the three domains: product, restaurant and bibliography. The topical expertise was drawn from the normal distribution. Each human worker made judgments to 3 000 entity pairs, and each entity pair was judged by five humans. Each human judgment was generated by a Bernoulli distribution according to the corresponding topical expertise.

5.2 Comparative Approaches

We compared the proposed TEST model against five representative methods that are used to infer truths of crowd ER tasks from human judgments. These methods are Majority Voting (MV)[9], GLAD[10], ALM[8], FaitCrowd[5], and DOCS[38]. MV uses the majority of the human judgments to infer truths. GLAD applies unsupervised learning to simultaneously estimate the task truth, the task difficulty, and the human expertise on the assumption that each human has consistent expertise over all tasks. ALM first computes a high-level feature representation for each task, and then uses a probabilistic graphical model to learn the task truths and human expertise where human expertise is represented by a weighted linear combination of a task's high-level features. FaitCrowd leverages a TwitterLDA-based topic model to estimate topic-based human expertise and task truths simultaneously. It assigns a single topic to each task. DOCS first identifies the knowledge-base entities that are mentioned in a task's text and uses their categories to detect the domains of the task. After that, DOCS takes an iterative procedure to infer task truths and domain-based human expertise. In our experiments, we implemented the module of modeling expertise of multiple labelers of ALM, not using the active learning algorithm of it due to the irrelevance. We generated the feature vector of each entity pair for ALM using the method in [44]. We implemented two modules of DOCS, i.e., domain vector estimation and truth inference, due to their relevance to our paper.

In the parameter settings for TEST, we used grid search to select the topic number K on each dataset: 28 for the SView dataset, 16 for the BTC dataset, and 6 for the synthetic dataset. Parameter α of TEST was set to $50/K$. Since the truth of each task was unknown at first, we set the prior probability of a task truth π to 0.5 to generate the task truth. We set the number of

similar task groups G in TEST to 10 for all the three datasets. We will show the reason of our choices shortly in Subsection 5.5. As described in Subsection 4.1, in TEST, we generated the topics of each task by considering itself and its similar ones that are in the same task group. In order to compute the task prior similarity, the weights for property, URI pay-level-domain, and value similarity in TEST were set to 0.65, 0.1, and 0.25, respectively. The values of task prior similarity less than the threshold 0.8 were set to 0 so that in the process of generating a task's topics, only its similar tasks that were in the same group and had task similarity larger than the threshold could be used for cross sampling (see Subsection 4.1). As a result, on average, 10, 11 and 16 completed similar tasks were used for cross sampling for a task in the SView, BTC and synthetic datasets, respectively.

We also used grid search to set the topic numbers for ALM and FaitCrowd. For DOCS, we manually set the number of topics K for the three datasets according to the entity domains in the datasets. We initialized human expertise on each domain to 0.5 in DOCS. For other parameters, we followed the same setting strategy in the corresponding references[5,8,10,38].

5.3 Experiment on Task Truth Inference

In this experiment, we evaluated the performance of different approaches and the main components of our model on task truth inference.

5.3.1 Procedure

To evaluate the performance of each approach, we followed the work in [5, 8, 10, 38] and used accuracy as the evaluation metric, which is defined as the number of entity pairs of which the truths are correctly estimated (including matched and non-matched pairs), divided by the total number of entity pairs. A higher accuracy means that the estimation of the approach is closer to the ground truth.

For each dataset, we reported the average accuracy achieved w.r.t. different numbers of human judgments. Starting by sampling 20% of the human judgments, we increased the number of sampled judgments by adding further randomly-selected judgments and re-ran all the approaches. This process was repeated until all the judgments had been used in the evaluation. As a result, we sampled 20%, 40%, 60%, 80% and 100% of judgments respectively. For each percentage, we ran 20 times and calculated the average.

We firstly compared the performance of TEST with MV, GLAD, ALM, FaitCrowd, and DOCS. We then compared two variants of TEST to show the benefit of considering varied topical expertise and similar tasks. W/O TE is a variant of TEST using the same expertise for all topics. W/O ST is the other variant of TEST by removing the modeling of similar task groups. We performed an ablation study on the accuracy of TEST and its two variants W/O TE and W/O ST w.r.t. varied numbers of human judgments, in order to investigate the effect of varied topical expertise and similar task group modeling.

5.3.2 Results

For the first comparison, the average accuracy of different approaches w.r.t. the varied percentage of human judgments on the three datasets is shown in Fig.3. From this figure, we can see that the approaches considering human expertise generally achieve higher accuracy than MV, of which the ones using varied topical expertise obtain higher accuracy than the one using non-varied topical expertise (i.e., GLAD). The accuracy of TEST is the highest on all the three datasets, while DOCS obtains higher accuracy than FaitCrowd and ALM, and FaitCrowd is better than ALM. On average, TEST outperforms DOCS, FaitCrowd, ALM, GLAD, and MV by 1.1%, 2.2%, 3.3%, 4.7% and 5.5% in the three datasets through different percentages of judgments, respectively.

The average accuracy of TEST, W/O ST and W/O TE w.r.t. the varied percentage of human judgments is shown in Fig.4. From the figure, we can see that both the two variants achieve lower accuracy than TEST on the three datasets.

The first comparison results show that MV usually has limited performance without considering human expertise when there exist a lot of inconsistencies among human judgments or no sufficient human judgments. Modeling varied human expertise on topics seems to be beneficial for accurately inferring the task truths. Furthermore, the performance of task truth inference tends to be improved with a better estimation of task topics. Both ALM and FaitCrowd have their limitations in task topic estimation. ALM uses a sparse coding algorithm to find topics, which may influence the accuracy of modeling topics when human judgments are sparse. FaitCrowd assumes that a task belongs to a single topic; thus it is not suitable for modeling the tasks belonging to multiple different topics. These may cause the relatively low accuracy of the two approaches. DOCS can

provide a better estimation of task topics and achieve higher accuracy by leveraging the topics of the knowledge base entities mentioned in the task. The accuracy of the proposed TEST approach was consistently higher than or equal to that of DOCS on the three datasets. But due to the fact that the three datasets currently used were not very hard for humans to resolve, there was no significant difference between the accuracy of TEST and DOCS in general. However, DOCS relies on entity linking and Freebase's entity categories for identifying task topics, which may have limited accuracy on specific datasets, due to the reasons like that
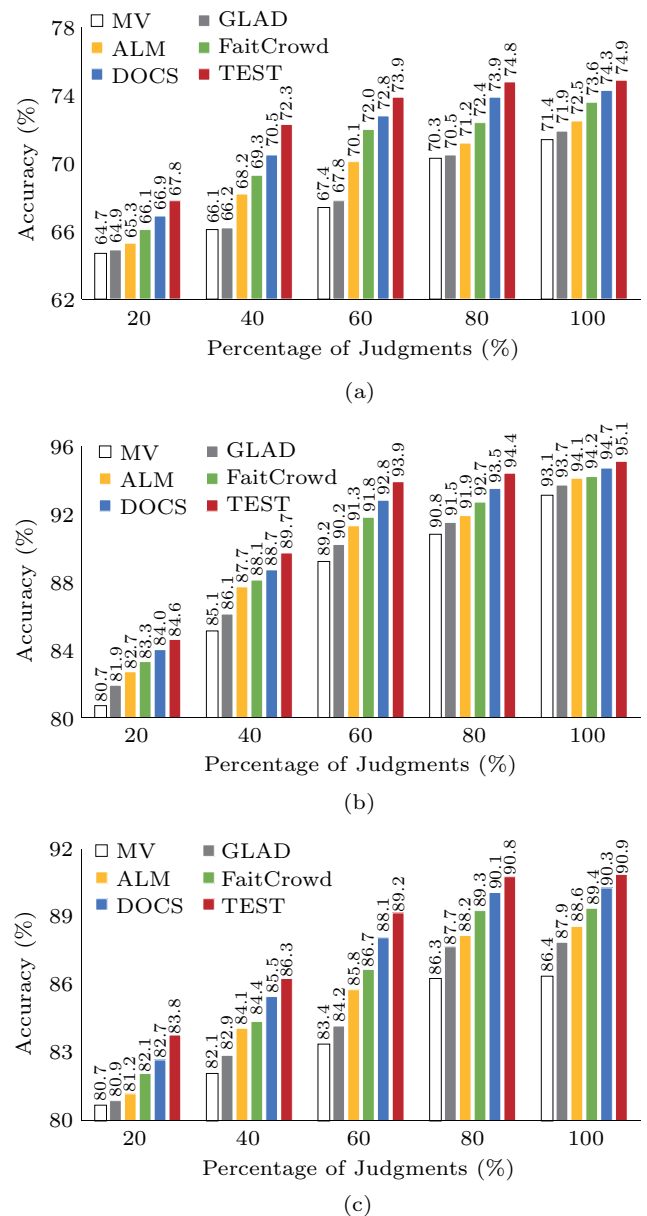

Fig.3. Accuracy comparison between TEST and comparative approaches. (a) SView. (b) BTC. (c) Synthetic.

1214

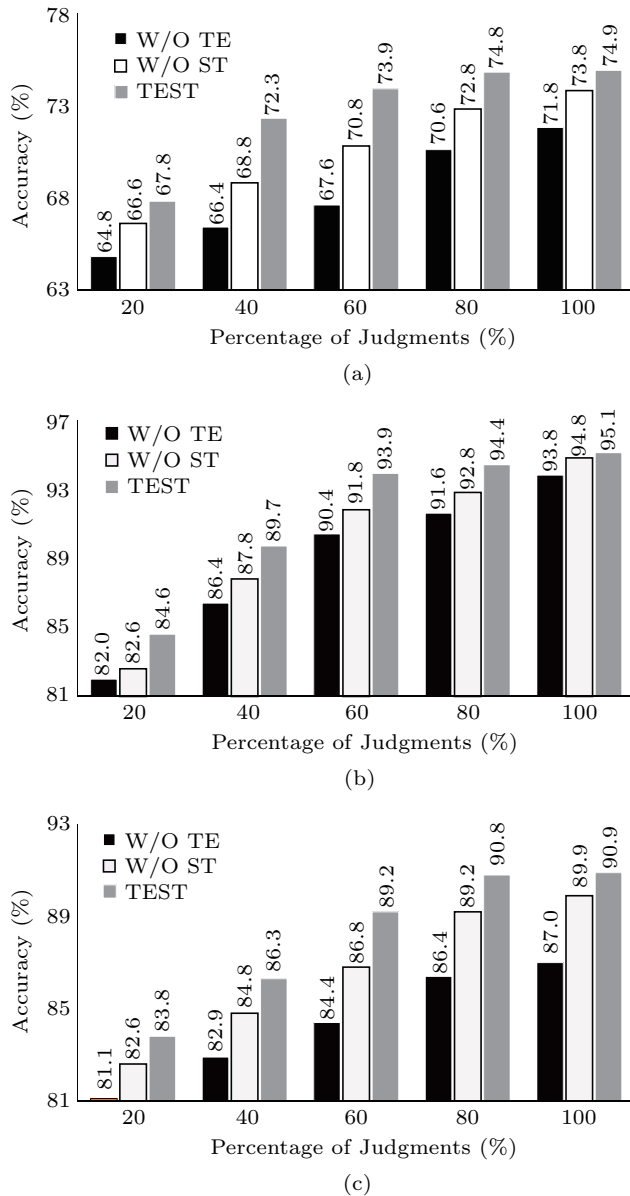*J. Comput. Sci. & Technol., Nov. 2018, Vol.33, No.6*

Fig.4. Ablation study on TEST. (a) SView. (b) BTC. (c) Synthetic.

named entity cannot be found in Freebase. Furthermore, when only a small number of human judgments were provided, TEST may achieve considerable higher accuracy than DOCS (e.g., 72.3 vs 70.5 using 40% judgments in the SView dataset). It may be due to the reason that TEST makes use of similar tasks to address the data sparsity problem and perform better truth inference.

As an example, only TEST correctly identified the truth of the following entity pair in the SView dataset: dbp:Sint_Maarten[3] and geo:3578422 (Saint Martin)[4].

---

[3] http://dbpedia.org/resource/Sint_Maarten, Aug. 2018.

[4] http://sws.geonames.org/3578422/, Aug. 2018.

The entity pair has two topics, geography and administrative district. In fact, the two entities represent two parts of the same island but belong to two different countries, and thus they are non-matched entities. The humans who had high-level expertise (and perhaps had more interest) on the two topics provided the correct judgments. MV reported a wrong inference result since the majority of the humans provided wrong judgments. GLAD's inference result was wrong because it failed to take into account the variation of human expertise on the two topics, which was different from the overall expertise over all tasks. The wrong inference of ALM may be due to the fact that the sparsity of human judgments made it difficult to compute the high-level task features and thus the computed features may no longer represent the task topics correctly. FaitCrowd only identified the geography topic and thus could not leverage the human expertise on the other topic for making correct truth inference. DOCS did not fully utilize the property information in the entity descriptions except entity categories for truth inference, e.g., (country, "France"). TEST correctly identified these two topics from the entity descriptions, distinguished human workers of high reliability from the other ones, and thus yielded the correct truth.

The result of the ablation study verifies that modeling varied expertise on different topics seems to greatly improve the task truth inference. Furthermore, the estimation of task truths and topics can be significantly improved by incorporating the similar task group modeling.

5.4 Experiment on Human Expertise Estimation

In this experiment, we evaluated the performance of different approaches on human expertise estimation.

5.4.1 Procedure

TEST and the other four approaches, ALM, FaitCrowd, DOCS and W/O ST, can learn humans' varied expertise on different topics. To validate the correctness of human expertise learned by these approaches, two measures, Pearson and Kendall coefficients[45], were used in the evaluation. Pearson coefficient measures the degree of linear correlation between two variables while Kendall coefficient measures the ordinal correlation. In this experiment, one of the

two variables was each human's expertise learned by an approach on a topic, and the other was each human's correctly judged percentage of the tasks in the same topic, which was obtained from the ground truth. It is worth noting that, on the synthetic dataset, the true human expertise on various topics was provided beforehand. Higher Pearson or Kendall values indicate better performance in the topical expertise estimation.

We conducted five-fold cross-validation and sampled 80% of human judgments as training data. We computed the values of Pearson and Kendall coefficients for every topic modeled by an approach. Then, we reported the average Pearson and Kendall values across all the topics.

5.4.2 Results

Fig.5 shows the comparison results on human expertise estimation between TEST and the four comparative approaches. We can see that ALM obtains the lowest Pearson and Kendall values on the three datasets while TEST achieves the highest. DOCS has the second best Pearson and Kendall values, and W/O ST behaves worse than TEST. The comparison results on human expertise estimation show a high correlation between the topical expertise learned by TEST and the ground truth. On average, the Kendall value of TEST is about 5.3%, 9.6% and 18.4% higher than those of DOCS, FaitCrowd and ALM, respectively. The reason why ALM and FaitCrowd obtain relatively lower topical expertise estimation performance may be due to the fact that they have some limitations of identifying task topics as described in Subsection 5.3.2. DOCS outperforms them since it models the task topics more accurately. TEST does not rely on extra knowledge bases to detect task topics. It can estimate topical expertise more effectively than the other approaches because it can capture the inherent correlations among similar tasks to model the human expertise and task truths more accurately. Without considering the correlations among similar tasks, W/O ST performs worse than TEST and DOCS in estimating topical expertise.

5.5 Analysis of Parameter Sensitivity

This experiment analyzed the sensitivity of topic numbers and similar task group numbers in our model. It also explained the reasons of our parameter selection.

Fig.6 shows the accuracy w.r.t. varied topic numbers using 80% of the sampled judgments. We can see that, as the topic number increases, the accuracy firstly

rises to the peak, where the number of topics is about twice as many as the entity domains in each dataset, and then drops down if the topic number continues to increase. This means that using too few topics cannot sufficiently separate different topics, while using too many topics may result in many detailed but highly-similar topics, which both degrade the performance.
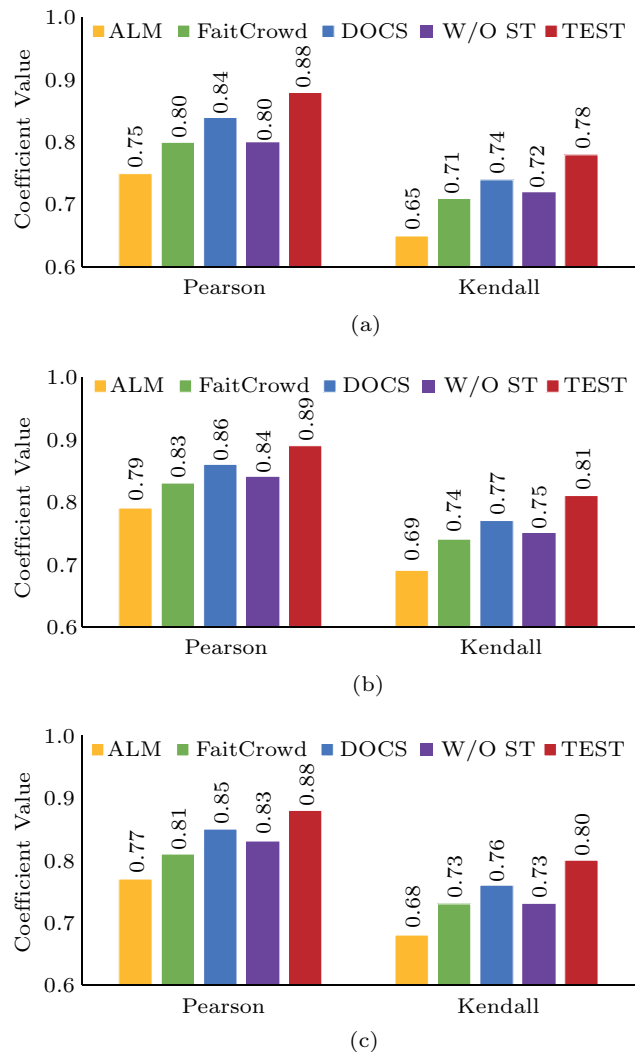

Fig.5. Comparison on human expertise estimation between TEST and comparative approaches. (a) SView. (b) BTC. (c) Synthetic.

Fig.7 shows the accuracy w.r.t. varied number of similar task groups using 80% of the sampled judgments as well. We can see that choosing 10 similar task groups achieves the highest accuracy on all the three datasets, because using too few groups cannot effectively capture the inherent correlations among similar tasks, while using too many groups may bring in noises in terms of task similarity.

1216

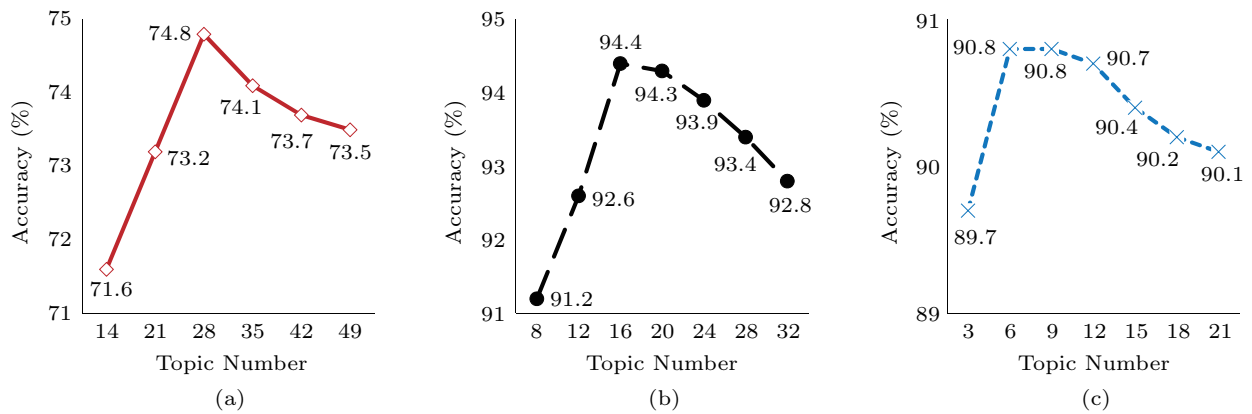J. Comput. Sci. & Technol., Nov. 2018, Vol.33, No.6

Fig.6. Accuracy variation w.r.t. topics. (a) SView. (b) BTC. (c) Synthetic.
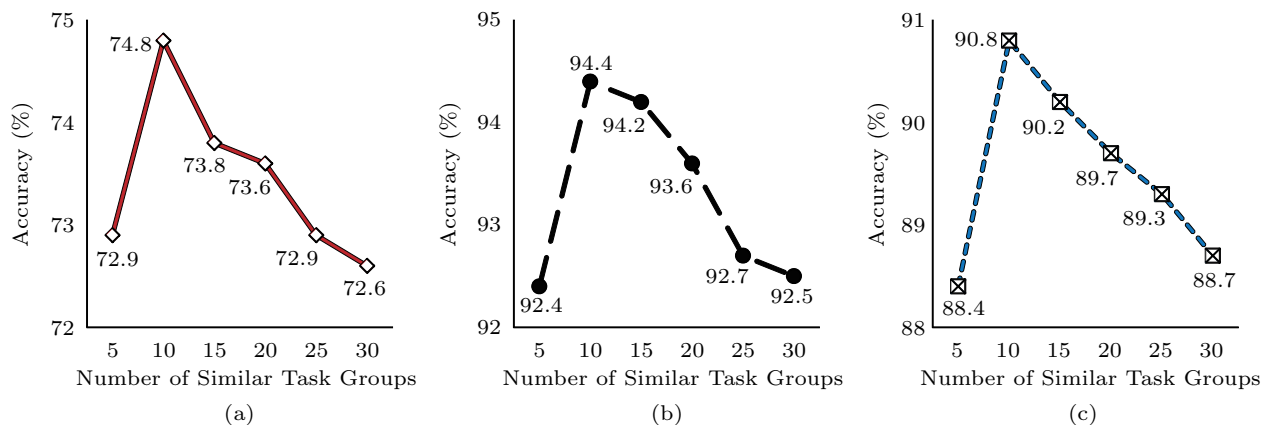

Fig.7. Accuracy variation w.r.t. similar task groups. (a) SView. (b) BTC. (c) Synthetic.

6   Conclusions

We proposed TEST for crowd ER, which infers the truth of each ER task according to the varied human expertise on different task topics. We took advantage of an adapted topic modeling to identify the multiple topics of each task. To address the data sparsity problem, we incorporated the similar task clustering in TEST to better estimate the task truths and topic-based human expertise. A variational EM algorithm was devised to learn the human varied expertise, compute the task similarity, and infer the task truths integrally. The experimental results showed that, compared with five state-of-the-art methods, the accuracy of TEST is 1.1% higher than the second best method. Furthermore, its estimated expertise outperforms the second best result by 5.3% in terms of correlation with human real expertise.

In future work, more investigations will be needed to deploy TEST on the public crowdsourcing platforms, e.g., Amazon Mechanical Turk, CrowdFlower or ChinaCrowd[16] and test its performance for crowd ER. Particularly, we will study the problem of online task assignment. Furthermore, our future research will continue exploring the correlations among humans apart from tasks, since similar humans may have similar expertise. In addition, TEST is currently only used for ER. We look forward to extending TEST to more different types of tasks, such as data quality assessment and question answering.

References

[1] Heflin J, Song D. Ontology instance linking: Towards interlinked knowledge graphs. In *Proc. the 30th AAAI Conf. Artificial Intelligence*, February 2016, pp.4163-4169.

[2] Hu W, Jia C. A bootstrapping approach to entity linkage on the Semantic Web. *Journal of Web Semantics*, 2015, 34: 1-12.

[3] Wang J, Kraska T, Franklin M J, Feng J. CrowdER: Crowdsourcing entity resolution. *Proceedings of the VLDB Endowment*, 2012, 5(11): 1483-1494.

[4] Yalavarthi V K, Ke X, Khan A. Select your questions wisely: For entity resolution with crowd errors. In *Proc. the 26th Int. Conf. Information and Knowledge Management*, November 2017, pp.317-326.

[5] Ma F, Li Y, Li Q, Qiu M, Gao J, Zhi S, Su L, Zhao B, Ji H, Han J. FaitCrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proc. the 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, August 2015, pp.745-754.

[6] Yan Y, Rosales R, Fung G, Dy J G. Active learning from crowds. In *Proc. the 28th Int. Conf. Machine Learning*, June 2011, pp.1161-1168.

[7] Raykar V C, Yu S, Zhao L H, Valadez G H, Florin C, Bogoni L, Moy L. Learning from crowds. *Journal of Machine Learning Research*, 2010, 11: 1297-1322.

[8] Fang M, Yin J, Tao D. Active learning for crowdsourcing using knowledge transfer. In *Proc. the 28th AAAI Conf. Artificial Intelligence*, July 2014, pp.1809-1815.

[9] Kuncheva L I, Whitaker C J, Shipp C A, Duin R P. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, 2003, 6(1): 22-31.

[10] Whitehill J, Ruvolo P, Wu T, Bergsma J, Movellan J R. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proc. the 23rd Annual Conf. Neural Information Processing Systems*, December 2009, pp.2035-2043.

[11] Snow R, O'Connor B, Jurafsky D, Ng A Y. Cheap and fast — But is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. the 2008 Conf. Empirical Methods in Natural Language Processing*, October 2008, pp.254-263.

[12] Fan J, Li G, Ooi B C, Tan K L, Feng J. iCrowd: An adaptive crowdsourcing framework. In *Proc. the 2015 ACM SIGMOD Int. Conf. Management of Data*, May 2015, pp.1015-1030.

[13] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3: 993-1022.

[14] Bhattacharya I, Getoor L. A latent Dirichlet model for unsupervised entity resolution. In *Proc. the 6th SIAM Int. Conf. Data Mining*, April 2006, pp.47-58.

[15] Li G, Wang J, Zheng Y, Franklin M J. Crowdsourced data management: A survey. *IEEE Trans. Knowledge and Data Engineering*, 2016, 28(9): 2296-2319.

[16] Li G, Zheng Y, Fan J, Wang J, Cheng R. Crowdsourced data management: Overview and challenges. In *Proc. the 2017 ACM SIGMOD Int. Conf. Management of Data*, May 2017, pp.1711-1716.

[17] Acosta M, Zaveri A, Simperl E, Kontokostas D, Auer S, Lehmann J. Crowdsourcing linked data quality assessment. In *Proc. the 12th Int. Semantic Web Conf.*, October 2013, pp.260-276.

[18] Demartini G, Difallah D E, Cudré-Mauroux P. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proc. the 21st Int. Conf. World Wide Web*, April 2012, pp.469-478.

[19] Chai C, Li G, Li J, Deng D, Feng J. Cost effective crowdsourced entity resolution: A partial-order approach. In *Proc. the 2016 ACM SIGMOD Int. Conf. Management of Data*, June 2016, pp.969-984.

[20] Vesdapunt N, Bellare K, Dalvi N. Crowdsourcing algorithms for entity resolution. *Proceedings of the VLDB Endowment*, 2014, 7(12): 1071-1082.

[21] Hassan U, Zaveri A, Marx E, Curry E, Lehmann J. ACRyLIQ: Leveraging DBpedia for adaptive crowdsourcing in linked data quality assessment. In *Proc. the 20th Int. Conf. Knowledge Engineering and Knowledge Management*, November 2016, pp.681-696.

[22] Kontokostas D, Zaveri A, Auer S, Lehmann J. TripleCheckMate: A tool for crowdsourcing the quality assessment of linked data. In *Proc. the 4th Int. Conf. Knowledge Engineering and the Semantic Web*, October 2013, pp.265-272.

[23] Fang Y L, Sun H L, Chen P P, Deng T. Improving the quality of crowdsourced image labeling via label similarity. *Journal of Computer Science and Technology*, 2017, 32(5): 877-889.

[24] Zhuang Y, Li G, Zhong Z, Feng J. Hike: A hybrid human-machine method for entity alignment in large-scale knowledge bases. In *Proc. the 2017 Int. Conf. Information and Knowledge Management*, November 2017, pp.1917-1926.

[25] Li G, Chai C, Fan J, Weng X, Li J, Zheng Y, Li Y, Yu X, Zhang X, Yuan H. CDB: Optimizing queries with crowd-based selections and joins. In *Proc. the 2017 ACM SIGMOD Int. Conf. Management of Data*, May 2017, pp.1463-1478.

[26] Zheng Y, Cheng R, Maniu S, Mo L. On optimality of jury selection in crowdsourcing. In *Proc. the 18th Int. Conf. Extending Database Technology*, March 2015, pp.193-204.

[27] Li Q, Ma F, Gao J, Su L, Quinn C J. Crowdsourcing high quality labels with a tight budget. In *Proc. the 9th ACM Int. Conf. Web Search and Data Mining*, February 2016, pp.237-246.

[28] Yuan D, Li G, Li Q, Zheng Y. Sybil defense in crowdsourcing platforms. In *Proc. the 2017 Int. Conf. Information and Knowledge Management*, November 2017, pp.1529-1538.

[29] Li Q, Li Y, Gao J, Zhao B, Fan W, Han J. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proc. the 2014 ACM SIGMOD Int. Conf. Management of Data*, June 2014, pp.1187-1198.

[30] Xiao H, Gao J, Li Q, Ma F, Su L, Feng Y, Zhang A. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proc. the 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, August 2016, pp.1935-1944.

[31] Ma F, Meng C, Xiao H, Li Q, Gao J, Su L, Zhang A. Unsupervised discovery of drug side-effects from heterogeneous data sources. In *Proc. the 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, August 2017, pp.967-976.

[32] Wang Y, Ma F, Su L, Gao J. Discovering truths from distributed data. In *Proc. the 2017 IEEE Int. Conf. Data Mining*, November 2017, pp.505-515.

[33] Meng C, Jiang W, Li Y, Gao J, Su L, Ding H, Cheng Y. Truth discovery on crowd sensing of correlated entities. In *Proc. the 13th ACM Conf. Embedded Networked Sensor Systems*, November 2015, pp.169-182.

[34] Zhang H, Li Q, Ma F, Xiao H, Li Y, Gao J, Su L. Influence-aware truth discovery. In *Proc. the 25th ACM Int. Conf. Information and Knowledge Management*, October 2016, pp.851-860.

[35] Hu H, Zheng Y, Bao Z, Li G, Feng J, Cheng R. Crowd-sourced POI labelling: Location-aware result inference and task assignment. In *Proc. the 32nd IEEE Int. Conf. Data Engineering*, May 2016, pp.61-72.

[36] Zheng Y, Wang J, Li G, Cheng R, Feng J. QASCA: A quality-aware task assignment system for crowdsourcing applications. In *Proc. the 2015 ACM SIGMOD Int. Conf. Management of Data*, May 2015, pp.1031-1046.

[37] Fang M, Zhu X, Li B, Ding W, Wu X. Self-taught active learning from crowds. In *Proc. the 12th IEEE Int. Conf. Data Mining*, December 2012, pp.858-863.

[38] Zheng Y, Li G, Cheng R. DOCS: Domain-aware crowd-sourcing system. *Proceedings of the VLDB Endowment*, 2016, 10(4): 361-372.

[39] Zheng Y, Li G, Li Y, Shan C, Cheng R. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 2017, 10(5): 541-552.

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

[41] Wainwright M J, Jordan M I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008, 1(1/2): 1-305.

[42] Qu Y, Gong S, Cheng G, Xu J, Li X, Zheng L, Jiang J. SView: Smart views for browsing linked entities. In *Proc. ISWC Semantic Web Challenge 2014*, October 2014.

[43] Köpcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 2010, 3(1): 484-493.

[44] Kejriwal M, Miranker D P. An unsupervised instance matcher for schema-free RDF data. *Journal of Web Semantics*, 2015, 35: 102-123.

[45] Abdullah M B. On a robust correlation coefficient. *The Statistician*, 1990, 39(4): 455-460.
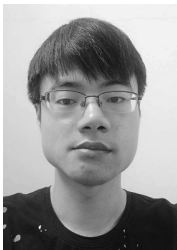
Sai-Sai Gong is currently a Ph.D. student in State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing. He received his B.S. degree in computer science and technology in 2009, and his M.S. degree in computer software and theory in 2012, both from Southeast University, Nanjing. His research interests include Semantic Web, linked data browsing and data integration.
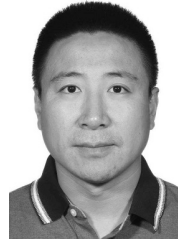
Wei Hu is currently an associate professor in State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing. He received his Ph.D. degree in computer software and theory in 2009, and his B.S. degree in computer science and technology in 2005, both from Southeast University, Nanjing. He has published more than 30 research papers in various journals and conference proceedings. His main research interests include knowledge graph, data integration and intelligent application.

Wei-Yi Ge is currently a senior engineer in Science and Technology on Information Systems Engineering Laboratory, Nanjing. He received his Ph.D. degree in computer software and theory in 2013, and his B.S. degree in computer science and technology in 2007, both from Southeast University, Nanjing. His main research focuses on semantic search and question answering system.

Yu-Zhong Qu received his B.S. and M.S. degrees in mathematics from Fudan University, Shanghai, in 1985 and 1988 respectively, and got his Ph.D. degree in computer software from Nanjing University, Nanjing, in 1995. He is currently a professor in State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing. His research interests include Semantic Web, question answering, and novel software technology for the Web. He has published more than 80 papers in major venues in these areas such as WWW, ISWC, and Journal of Web Semantics.