Linked Data Crowdsourcing Quality Assessment based on Domain Professionalism

IOP Conf. Series: Journal of Physics: Conf. Series **1187** (2019) 052085 doi:10.1088/1742-6596/1187/5/052085

Linked Data Crowdsourcing Quality Assessment based on Domain Professionalism

Lu Yang[1234], **Li Huang**[1234], **Zhenzhen Liu**[1234]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China

[2]Key Laboratory of Intelligent Information Processing and Real-time Industrial System in Hubei Province, Wuhan 430065, China

[3]Institute of Big Data Science and Engineering, Wuhan University of Science and Technology, Wuhan 430065, China

[4]Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, National Press and Publication Administration, Beijing 100038, China

Abstract. With the rapid development of Internet technology, crowdsourcing, as a flexible, effective and low-cost problem-solving method, has begun to receive more and more attention. The use of crowdsourcing to evaluate the quality of linked data has also become a research hotspot. This paper proposes the concept of Domain Specialization Test (DST), which uses domain professional testing tasks DSTs to evaluate the professionalism of workers, and combines the idea of Mini-batch Gradient Descent (MBGD) to improve the EM algorithm, and the MBEM algorithm is proposed to achieve efficient and accurate evaluation of task results. The experimental results show that the proposed method can screen out the appropriate workers for the linked data crowdsourcing task and improve the accuracy and iteration efficiency of the results.

1. Introduction

Crowdsourcing is a distributed problem-solving mechanism that towards the Internet public. It integrates computers and unknown people on the Internet to accomplish tasks that are difficult for computers to accomplish alone [1]. For example, image annotation[2], physical alignment[3], these tasks are difficult to handle by machine but can be done by crowdsourcing easily. Therefore, crowdsourcing has been widely studied and applied in many fields.

In recent years, the scale of Linked Data has exploded, which leads to the serious data quality problems [4]. Although there are a large number of researches on automated or semi-automated tools for dealing with the quality of linked data [5-6], there are still a large number of problems in the data quality that are difficult to find or solve, but human intelligence can easily handle. Therefore, the use of crowdsourcing to solve the quality problems in the linked data has gradually begun to attract attention.

Linked data is a collection of structured data (RDF data) on the network [7]. A data set usually covers multi-domain knowledge and the data set is huge, so the domain knowledge involved in the linked data crowdsourcing task is also diverse. The crowdsourcing workers come from the Internet, they have different educational backgrounds and different professional knowledge. The quality of their

work is bound to be uneven, which makes it difficult to control the quality assessment of crowdsourcing results. Therefore, how to select the workers who have the relevant knowledge of Linked Data Tasks (LDT), and how to improve the integration efficiency of crowdsourcing task results, ensure the quality of crowdsourcing task results are urgent problems to be solved in the linked data crowdsourcing research.

2. Related work

At present, the research on linked data crowdsourcing mainly focuses on the quality problem detection, ontology alignment and entity linking. Literature [8] proposes a quality assessment of the linked data of the Find-Fix-Verify model based on crowdsourcing in the form of competition and micro-task. Literature [9] proposes a two-stage associated data quality assessment method combining manual and semi-automatic forms. The literature [3] transforms the ontology alignment problem into a crowdsourcing micro-task. Literature [10] proposes a probabilistic framework-based system on crowdsourcing platforms to improve link quality.

The focus of the above research is on how to apply crowdsourcing to the field of linked data. Due to the uncertainty of crowdsourcing workers and the large scale of linked data sets, an important issue in linked data crowdsourcing is the quality assessment of crowdsourcing results and the evaluation of efficiency issues.

In terms of quality assessment of results, one of the most commonly used quality assessment methods is the Golden Standard Data (GSD) evaluation method [11] which refers to a type of data with standard answers. The accuracy of pure gold standard data does not represent that the professional knowledge possessed by workers meets the requirements of crowdsourcing tasks, and gold standard data needs to be manually generated, adding extra costs.Literature [12] proposes a staged dynamic crowdsourcing quality control strategy, due to the setting of task detection points and replacement rules, the task completion time will be greatly extended.The EM algorithm proposed by Dawid and Skene[13] can accurately estimate the task results, but when the task volume is large, resulting in low efficiency of the algorithm.

The above evaluation methods all have shortcomings between the accuracy and efficiency of task results. The domain professional evaluation method and MBEM quality evaluation algorithm proposed in this paper can effectively screen high-quality workers and greatly improve the evaluation efficiency of task results.

3. linked data crowdsourcing quality assessment method

3.1. Predefined

This paper focuses on how to improve the quality of crowdsourcing task results and the efficiency of evaluation by controlling the quality of workers and the process of result integration in linked data crowdsourcing applications. The relevant definitions are given below.

Definition 1. (Linked Data Task: LDT): $T = \{t_1, t_2, ..., t_n\}$ represents the set of tasks for a given linked data set LD. The task $t_i \in T$ is the label task with the only answer $c \in C$, $C = \{c_1, c_2, c_3, c_4\}$.

$$t_i = \{< subject, predicate, object >, < c_1, c_2, c_3, c_4 >\} \tag{1}$$

Definition 2. (Workers): $W = \{w_1, w_2, ..., w_n\}$ represents the crowdsourcing workers collection, each worker w_j completes task $t_i \in T$ independently.

Definition 3. (Domain Specialization Test: DST): $D = \{d_1, d_2, ..., d_n\}$ represents the domain classification of the task T. Different workers $w_j \in W$ have different domain expertise in different domain d_m:

$$DST_j = \{P_m = \frac{Right(DSTs_m)}{DSTs_m.length} \mid d_m \in D, w_j \in W, P_m \in [0,1]\} \tag{2}$$

Definition 4. (Domain Specialization Test Task: DSTs): DSTs = {dst_1, dst_2, ..., dst_n} are testing tasks similar to T extracted from the Standard Knowledge Base (SKB), Similar(t_i, dst_j)∈[0,1].

Definition 5. (Gold Standard Tasks: GST): gold standard tasks referred to a set of test tasks which have a standard answer, G = {g_1, g_2, ..., g_n}, is used to identify malicious workers during crowdsourcing tasks:

$$wor\ker = \begin{Bmatrix} accepte(w_j) & P_j = \dfrac{Right(GST)}{GST.length} \geq P_{s\tan dard} \\ reject(w_j) & P_j = \dfrac{Right(GST)}{GST.length} < P_{s\tan dard} \end{Bmatrix} \quad (3)$$

Definition 6. (Simple Weighted Majority Voting): The task t_i∈T is assigned to multiple workers to answer independently, and the answers are integrated by weighted voting, with the answers of most workers as correct answer:

$$\text{Answer}(t_i) = \arg\max_c (\sum_{c_1} P_w, \sum_{c2} P_w, ..., \sum_{cn} P_w), t_i \subset T, w \subset W, c \subset C \quad (4)$$

The problem to be solved in this paper is: Given the linked data crowdsourcing task T, the worker collection W, and the standard knowledge base SKB, how to get the optimal result set R of T efficiently and accurately.

3.2. DST-based linked data crowdsourcing quality control strategy

3.2.1. DST evaluation model. The linked data set involves the knowledge domain including various subject areas, and the inherent knowledge of human beings inevitably has certain limitations. In order to ensure the quality of LDT results, it is necessary to screen out high-quality workers with domain knowledge that matches the task to complete the linked data crowdsourcing task. Studies in [14] have shown that workers' reliability is often comparable in similar areas. Therefore, this paper introduces the concept of DST, extracts task-related testing tasks from known knowledge bases, and measures the user's domain expertise to match the appropriate linked data crowdsourcing tasks.

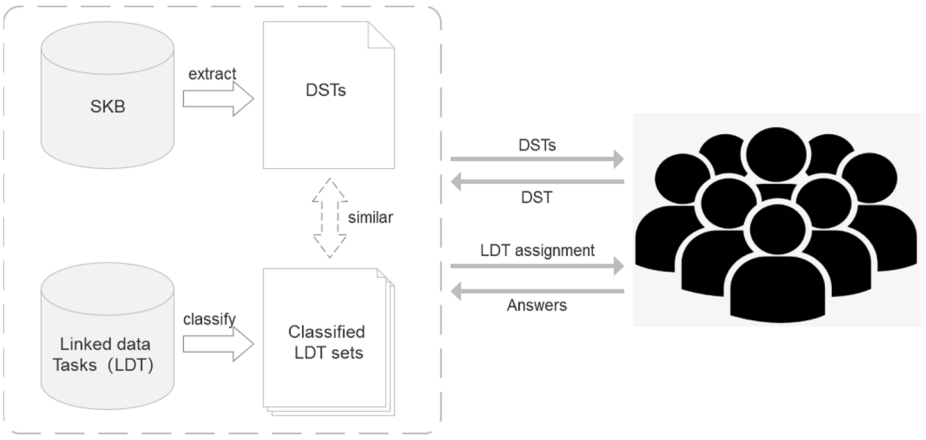

Figure 1. DST-based crowdsourcing quality control process.

Figure 1 depicts the implementation flow of the DST-based evaluation method proposed in this paper, the implementation steps are as follows:

Step 1: the linked data crowdsourcing task LDT is classified by domain knowledge;

Step 2: According to the classification of LDT, extract the task-related domain professional testing tasks from the selected standard knowledge base;

Step 3: the workers, who participate in the crowdsourcing task and have not completed the domain professional test, first assign the domain professional testing tasks DSTs;

Step 4: According to the professional ranking of the workers, assign the LDT corresponding to the most professional DSTs to the workers;

Step 5: The worker submits the answer to the corresponding task.

Extracting domain professionality testing tasks requires that the knowledge of the standard knowledge base meets the diversity of the domain and has a large amount of common sense knowledge, such as DBpedia, Wikipedia and so on. However, if the number of triples included in SKB is too large, extracting testing tasks only by traversal will be impossible. Therefore, this paper adopts a link discovery framework LIMES [15] in the metric space, an efficient link discovery method between data sources, and pre-processes the standard knowledge base, greatly reducing the traversal space. Trigram [16] calculates the triple similarity as the normalized sum of the absolute differences between the triplet vectors of the two input strings, we retrieves all resources with a similarity greater than 0.85.

Domain professional testing tasks are extra tasks. If we set too many DSTs, it will increase the cost budget of the requester. If the setting is too small, it will lose the representativeness. Therefore, we measure the coverage of DSTs according to Shannon entropy [17]. The higher is the entropy, the greater is the uncertainty, and the larger is the coverage of the same number of DSTs.

3.2.2. DST algorithm description. n represents the number of DSTs that each category needs to extract, and C represents the number of workers assigned to each linked data crowdsourcing task. When workers participate in crowdsourcing tasks, they first assigned DSTs. Based on the response to the DSTs, a domain professionality array Pi is generated for the worker wi, and a crowdsourcing task matching the domain knowledge is assigned according to the worker's domain professional ranking. When all the tasks get C answers, the task assignment process stops.

Algorithm1: DST evaluation algorithm

Input: Task T, SKB B, Domain D, Number n, Number C
Output: DST = $\{P_d \mid d \in D\}$, Answer A
B ← LIMES(B, T);
for n = 1,...,D do
  B Remove DSTs;
  for each b ∈ B do
△H ← getEntropyDiffer(b ∪ DSTs);
  b* ← b let max△H and maxSimilar(t_i,b);
  DSTs ← DSTs ∪ b*;
end for
for p < TC do
  worker ← getWorker();
  if worker not do DSTs then
    T_i ← getDSTs();
    P_i ← getWorkerDSTsResult;
  end if
  T_{id} ← getTask(maxP_i);
  p ← p + T_i;
  R ← R + R_{Ti};
end if

3.3. MBEM-based crowdsourcing result quality assessment method

3.3.1. The core idea of MBEM algorithm. The EM algorithm proposed by Dawid and Skene [13] is based on the maximum likelihood estimation of the error rate of multiple observers, and iteratively estimates the accuracy of the task results and the accuracy of the workers until convergence, thus,

achieving an accurate assessment of the results of the task. For the linked data crowdsourcing task, the linked data set is huge, and each iteration needs to recalculate the results of all tasks, resulting in very low efficiency of the algorithm. On the other hand, the initial parameter values of the EM algorithm have a great influence on the efficiency and accuracy of the iterative process. When the initial parameter values are set reasonably, the iterative process converges quickly and the result accuracy is higher; if the initial parameter value setting deviates from the actual situation, the efficiency of the algorithm is greatly reduced, and the accuracy of the final estimated result is relatively low.

The DST model proposed in this paper is used to measure the accuracy of workers' tasks in this domain. Therefore, the MBEM algorithm takes the domain expertise of the worker as the initial input parameter value.

Mini-batch gradient descent method MBGD is an iterative method commonly used to solve model parameters of machine learning algorithms. The basic idea is to divide the training sample into multiple sub-sample sets, and each iteration only performs gradient descent for one sub-sample set, which can solve the shortcomings of training too slow with Batch Gradient Descent (BGD). In this paper, the method of mini-batch gradient descent is adopted, and the result of each local iteration is used as the initial parameter value of the next local iteration, thereby improving the iterative efficiency.

3.3.2. MBEM algorithm description. InitialAccuracy is the field professional of the worker. InitialResult is the result of the crowdsourcing task submitted by the worker. The batch_size is the block size, that is, the number of samples per mini_batch. The initial parameter value of the first iteration of the algorithm is the field professional degree of the worker. Using the idea of simple weighted majority voting method, the task result is weighted and statistically obtained to obtain the task result.

Algorithm2: MBEM algorithm

Input: InitialAccuracy, InitialResult, batch_size
Output: Result

```
num ← InitialResult.size/mini_size;
accuracy ← InitialAccuracy;
for n = 1 ,..., num do
 data ← getData(InitialResult, batch_size);
 while threshold > θ do
  r_temp ← eStep(data, accuracy);
  mStep(r_temp);
 end while
 Ri ← getResult();
end for
Result ← {Ri | i∈1,...,num };
return Result;
```

4. Experiment

4.1. Experimental setup

The hardware environment of this experiment is: ASUS notebook computer, 8GB memory, 4 core processor, clocked at 1.4GHz.

The three linked data sets used in the experiment were derived from three domain knowledge in "OpenKG.CN": Literal, Tourism, and Medicine, and the standard knowledge base uses DBpedia. This paper extracts 140, 60, and 40 (240 total) tasks from above three data sets as linked data crowdsourcing tasks.

4.2. Analysis of experimental results

The experiment is divided into three parts: (1) effectiveness of the DST assessment method; (2) effectiveness of initial value setting of the EM algorithm; and (3) the comparison of the efficiency of the MBEM algorithm.

4.2.1. Effectiveness of the DST assessment method. There is a significant difference in the performance of individual workers' professional abilities in different areas of professional tasks. Based on the domain professional requirements of the linked data crowdsourcing task, this paper automatically extracts domain professional testing tasks from the standard knowledge base to match the appropriate domain tasks for the workers.

In this paper, we use algorithm 1 to extract the number of DSTs in the domain of Literal, Tourism and Medicine from DBpedia, which are 25, 15 and 10 (50 in total). Use the first 50 data of the linked data crowdsourcing task as a GSTs task. In order to verify the effectiveness of the domain expertise, we recruited 20 volunteers to complete the above 290 tasks independently.

Figure 2-4 shows the accuracy of DSTs in the Literal, Medicine, and Tourism domain and the accuracy of LDT results in the corresponding domain. For a certain domain, the higher the degree of DST, the higher the accuracy of the corresponding domain task completion; on the contrary, The lower the degree of DST, the greater the fluctuation of the accuracy of tasks in corresponding domain, which is due to the randomness of non-professional workers. Figure 5 shows the change of GST accuracy and LDT result accuracy. With the increase of the accuracy of GST, the accuracy of crowdsourcing tasks does not increase, but shows a large fluctuation. This shows that although the gold standard task can filter the grass rate workers out according to the accuracy rate, that is, the workers with low accuracy, but can not guarantee that their domain expertise meets the requirements, that is, the accuracy rate when completing the crowdsourcing task is not necessarily high. DSTs are selected according to a certain domain of crowdsourcing tasks, the higher the accuracy, the more knowledge the worker has in the domain, and the higher the accuracy of crowdsourcing tasks in the domain.

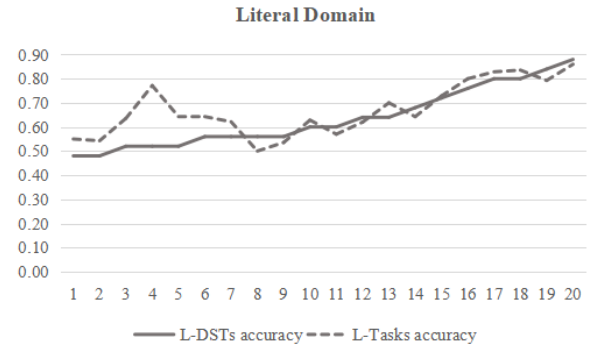

Figure 2. Literal Domain.

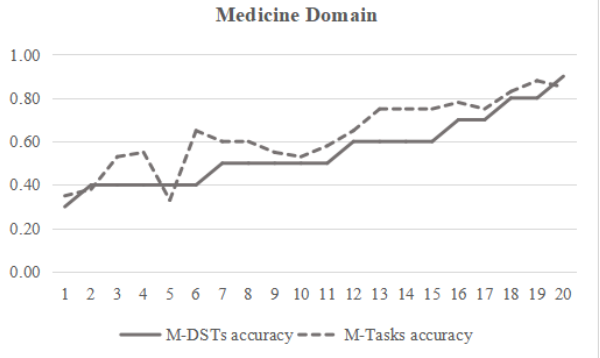

Figure 3. Medicine Domain.

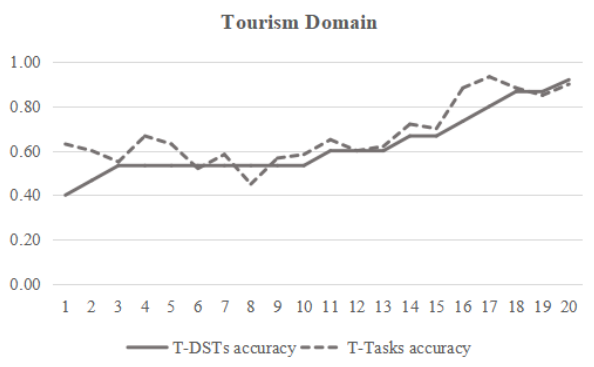

Figure 4. Tourism Domain.

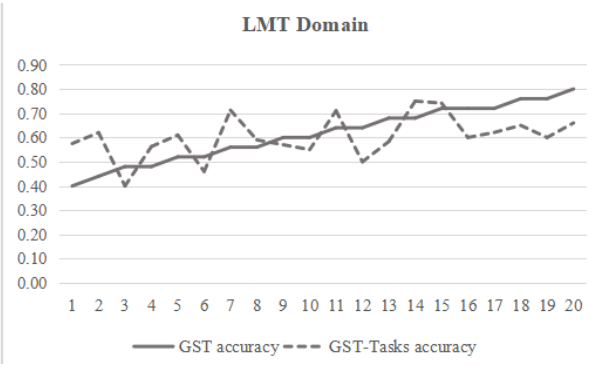

Figure 5. LMT Domain.

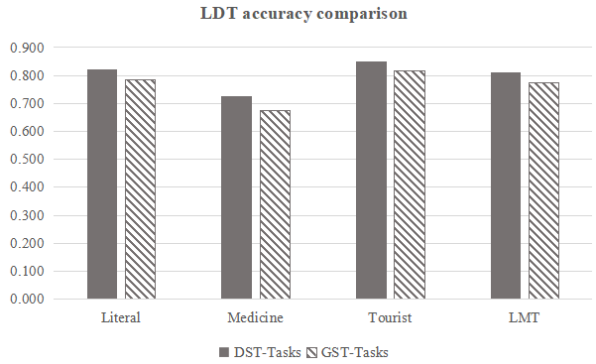

Figure 6. Comparison of the LDT accuracy.

4.2.2. Effectiveness of initial value setting of EM algorithm. For professional tasks in the field, if most workers do not have the expertise in the field and use the majority voting method to obtain results, the wrong results will affect the results of a small number of professional workers, resulting in unreliability of the final task results. Therefore, the initial parameter values of the EM algorithm have a certain influence on the accuracy of the results obtained by the iteration.

In this experiment, DSTs and GST are used as the initial parameter values of the EM algorithm. The accuracy of the results obtained by comparing the final task results with the real results is shown in Fig. 6. In the three domains of Literal, Medicine and Tourism, the accuracy of the results obtained with DSTs as the initial value is generally slightly higher than that of GST, and the quality of the results of the entire crowdsourcing task is better. DSTs represent the domain expertise of workers, which is closer to the accuracy of workers than GST, so algorithmic iterations produce less error and converge faster.

4.2.3. The comparison of the efficiency of the MBEM algorithm. The block size batch_size in the MBEM algorithm, that is, the number of samples per mini_batch directly affects the iterative efficiency of the algorithm. Therefore, choosing the appropriate batch_size will help improve the evaluation efficiency of crowdsourcing results. The actual linked data crowdsourcing task usually contains a very large amount of data, so this experiment uses the model to simulate 5000 data for subsequent experiments based on the results submitted by the workers.

As shown in Figure 7, when the size of the batch_size changes, the accuracy of the result fluctuates slightly. This is because when the batch_size is small, the number of samples in each mini_batch is too small and not representative, causes poor convergence during iteration; when the batch_size continues to increase, the number of samples in each mini_batch increases, and the iteration effect is enhanced; when the batch_size is too large, the task size in each mini_batch is too large, and the convergence accuracy of the algorithm will fall into different local extremums.

As you can see in Figure 7, when batch_size=5000, the result is the highest. This is because the entire data set is selected for each iteration. When batch_size=200, the result accuracy is only slightly lower than batch_size=5000. Therefore, choosing the appropriate batch_size size can guarantee the accuracy of the results to some extent.

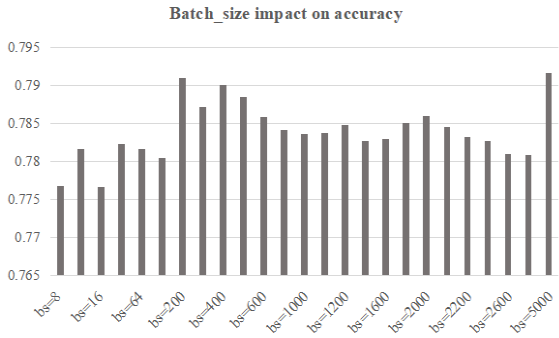

Figure 7. Comparison of batch-size.

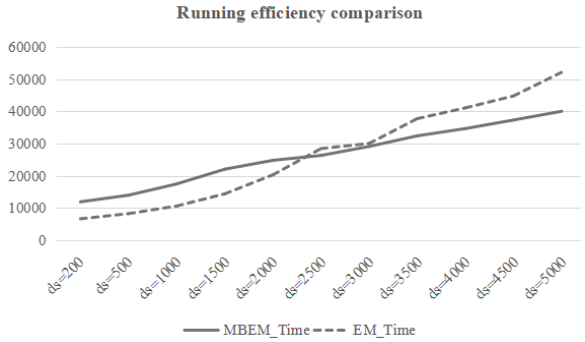

Figure 8. Comparison of the efficiency.

When the data set size of a crowdsourcing task changes, the operating efficiency of the algorithm changes. Figure 8 is a comparison of the operating efficiency of the MBEM algorithm and the EM algorithm. The abscissa is the data set size and the ordinate is the running time (ms). For each data set size, the batch_size value of the MBEM algorithm is the size of the batch_size that maximizes the accuracy of the results. When the amount of data is small, the EM algorithm runs better than the MBEM algorithm, and when the amount of data increases, the efficiency of MBEM algorithm is getting higher. When the amount of data increases to a certain value, MBEM algorithm is better than EM algorithm.

When the amount of data is small, the algorithm runs faster. The MBEM algorithm has additional mini_batch block time, so the running efficiency is relatively low. As the amount of data increases, the convergence effect of mini_batch block iterations increases, and the number of iterations decreases, the running speed is accelerated; and the EM algorithm needs to load the entire data set every iteration, resulting in greatly reduced operational efficiency. The results show that the MBEM algorithm proposed in this paper is relatively efficient for large data set.

5. Conclusion

This paper focuses on the quality control of crowdsourcing for linked data, and the domain professional evaluation model DST is proposed. The main idea of the model is that the knowledge similar to crowdsourcing task is extracted from standard knowledge base and applied to the evaluation of workers' domain expertise, and the appropriate crowdsourcing task is assigned to the workers according to the evaluation results. Secondly, combined with the idea of mini-batch gradient reduction, the EM algorithm is improved, and the MBEM algorithm is proposed to achieve efficient and accurate evaluation of task results. The experimental results prove the validity and usability of the evaluation model and the quality control algorithm, which lays a foundation for the subsequent research on semantic crowdsourcing quality control.

References:

[1] Feng J, Li G, Feng J. A survey on crowdsourcing [J]. *Chinese Journal of Computrs*, 2015, **38 (9)**:1713-1726.

[2] Fang Y L , Sun H L , Chen P P , et al. Improving the Quality of Crowdsourced Image Labeling via Label Similarity[J]. *Journal of Computer Science and Technology*, 2017, **32(5)**:877-889.

[3] Bontcheva K, Derczynski L, Roberts I. Crowdsourcing Named Entity Recognition and Entity Linking Corpora[M]// *Handbook of Linguistic Annotation*. 2017.

[4] Gu J, Zhu T, Huang L, et al. A Review of Research on Linked Data Quality Evaluation in Knowledge Graph[J]. *Journal of Wuhan University (Science Edition)*, 2017, **63(1)**:22-38.
[5] Flemming, A.: Quality Characteristics of Linked Data Publishing Datasources. *MA thesis*. Humboldt-Universität of Berlin (2010) .
[6] Christophe G, Groth P , Stadler C , et al. Assessing Linked Data Mappings Using Network Measures[M]*// The Semantic Web: Research and Applications.* Springer Berlin Heidelberg, 2012.
[7] BERNERS-LEE T. Design Issues for the World Wide Web[EB/OL].[2006-07-27].https://www.w3.org.DesignIssues/LinkedData.html.
[8] Acosta M, Zaveri A, Simperl E, et al. Detecting linked data quality issues via crowdsourcing: A dbpedia study[J]. *Semantic Web*, 2018, **9(3)**: 303-335.
[9] Zaveri, A., et al.: User-driven quality evaluation of DBpedia. In: Proceedings of the 9th International Conference on Semantic Systems, pp. 97–104. ACM (2013)
[10] G. Demartini, D. Difallah, and P. Cudr´ e-Mauroux. Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In 21st International Conference on World Wide Web WWW 2012, pages 469 – 478, 2012.
[11] Joglekar M , Garcia-Molina H , Parameswaran A . Evaluating the Crowd with Confidence[J]. 2014.
[12] Zhang Zhi-qiang, Feng Ju-sheng, Xie Xiao-qin, et al. Research on crowdsourcing quality control strategy and evaluation algorithm [J]. *Chinese Journal of Computrs*, 2013, **36(8)**:1636-1649.
[13] Dawid A P，Skene A M．Maximum likelihood estimation of observer error-rates using the EM algorithm[J]．*Applied statistics*，1979：20-28．
[14] Fan, J., et al.: iCrowd: an adaptive crowdsourcing framework. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1015–1030. ACM (2015)
[15] Ngonga Ngomo, A.-C., Auer, S.: LIMES - a time-efficient approach for large-scale link discovery on the web of data. In: Proceedings of IJCAI (2011)
[16] Martin S, Liermann J, Ney H. Algorithms for bigram and trigram word clustering[J]. *Speech Communication*, 1998, **24(1)**:19-37.
[17] Lin J. Divergence measures based on the Shannon entropy[M]. *IEEE Press*, 1991.